UNIVERSIDAD DE EXTREMADURA
FACULTAD DE POLITÉCNICA
DPTO. DE LENGUA INGLESA

A LEXICAL COMMON CORE IN ENGLISH FOR INFORMATION SCIENCE AND TECHNOLOGY

Alejandro Curado Fuentes

2000

Edita: Universidad de Extremadura
Servicio de Publicaciones
c/ Pizarro, 8
Cáceres 10071
Correo e.: publicac@unex.es
http://www.pcid.es/public.htm

Resumenes en español e inglés

Con este trabajo de tesis doctoral, se apunta a discernir entre los usos léxicos más comunes y especializados en las áreas de las ciencias y tecnologías de la información en lengua inglesa. De ahí que la dirección principal resida, precisamente, en el análisis del núcleo común de palabras y su clasificación según los empleos generales o académicos, técnicos y temáticos.

This PH.D. thesis aims to distinguish common from specialized lexical uses in the areas of Information Science and Technology English. Hence the main orientation is actually placed on the analysis of a lexical common core and its classification according to either general and academic employment or thematic and technical applications.

ACKNOWLEDGEMENTS

I wish to express my most heartfelt gratitude to Professor Patricia Edwards Rokowski, Ph. D. for her always helpful guidance and attention, as well as for the confidence she placed in me at all times.

My sincerest appreciation also goes to all the colleagues and students who have contributed to compiling conceptual perspectives upon lexical knowledge, thus assembling language and subject matter in an effective ESP framework. These are people from my Department of Lengua Inglesa as well as other Departments in Spain and abroad, ranging from friends at the Polytechnic School of Cáceres to physically unknown members of the Corpora electronic mail Discussion list, through Conference organizers and speakers whom I have had the pleasure to meet and exchange information with, such as Mike Scott, Guy Aston, Tim Johns, Peter Howarth, Randi Reppen, Doug Biber and others, frequently cited throughout this thesis, and whose scholarly works have enabled the production of my own humble claims.

For my parents

INDEX

CONCLUSIONS

APPENDICES

BIBLIOGRAPHY AND CORPUS REFERENCES

INDEX 2
SYMBOLS AND ABBREVIATIONS USED

SYMBOLS

SYMBOLS OF COLLOCATIONS AND LEXICAL ITEMS

®.....RESTRICTED COLLOCATIONS
F.....FREE COLLOCATIONS
T.....ITEM INCLUDED IN SOURCE
Ç.....ITEM NOT INCLUDED IN SOURCE
*****.....ITEM REPEATING BELOW THREE TIMES

SYMBOLS OF GENRE SUB-CORPORA

A.....TEXTBOOKS
B.....TECHNICAL REPORTS
C.....RESEARCH ARTICLES

SYMBOLS OF SUBJECTS AND TOPICS

A.....SUBJECT AREA SHARED BY COMPUTER SCIENCE AND TELECOMMUNICATIONS
A.1. / A(1).....History of computers, Hardware, Software
A.2. / A(2).....Computer engineering and architecture, Data; Communications and Client-server architecture

B.....SUBJECT AREA SHARED BY INFORMATION SCIENCE, COMPUTER SCIENCE AND TELECOMMUNICATIONS
B1 / B(1).....Information units management
B2 / B(2).....Online database systems, Computer systems
B3 / B(3).....Automated Knowledge-based systems

C.....SUBJECT AREA SHARED BY INFORMATION SCIENCE AND AUDIO-VISUAL COMMUNICATION
C1 / C(1).....Content analysis
C2 / C(2).....Media documentation
C3 / C(3).....Documentation Legislation

D ..... SUBJECT AREA SHARED BY TELECOMMUNICATIONS AND AUDIO-VISUAL COMMUNICATION
D1 / D(1) ..... Media technology
D2 / D(2 ..... Media theory

E ..... SUBJECT AREA SHARED BY INFORMATION SCIENCE, TELECOMMUNICATIONS, AND AUDIO-VISUAL COMMUNICATION
E1 / E(1) ..... Communication Theory

F ..... SUBJECT AREA SHARED BY ALL DISCIPLINES
F1 / F(1) ..... Perspectives on Information
F2 / F(2) ..... UNIX / Internet
F3 / F(3) ..... HTML, SGML, TEI
F4 / F(4 ..... Hypertext technology
F5 / F(5 ..... Electronic publishing
F6 / F(6) ..... Information infrastructure

OTHER SYMBOLS

ACT ..... ACTIVE VOICE
ADJ ..... ADJECTIVAL FORM
ADV ..... ADVERBIAL FORM
N ..... NOUN FORM
PAS ..... PASSIVE VOICE
PL ..... PLURAL FORM
SING ..... SINGULAR FORM
V ..... VERB FORM

ABBREVIATIONS

(only listed those that may appear unexplained throughout the body of text)

DISCIPLINES and ACADEMIC AREAS

* These five fields are abbreviated differently since they vary from how they appear in the program of studies to their reference in the lexical entries

***A. COM**.....AUDIO-VISUAL COMMUNICATION
ALL.....ALL FOUR DISCIPLINES
***AU**.....AUDIO-VISUAL COMMUNICATION
***AUD**.....AUDIO-VISUAL COMMUNICATION
***COM**.....COMPUTER SCIENCE
***C.S.**..... COMPUTER SCIENCE
DOC.....DOCUMENTATION
I.S......INFORMATION SCIENCE (Library Science and Documentation)
I.S.T......INFORMATION SCIENCE AND TECHNOLOGY (Computer Science, Information Science, Telecommunications, Audio-visual communication)
LIB.....LIBRARY SCIENCE
TEL.....TELECOMMUNICATIONS (Optical / Radio communication)

GENRES

RAs.....RESEARCH ARTICLES
RPs.....TECHNICAL REPORTS
TXs.....TEXTBOOKS

TEXT TYPES

CLASS.....CLASSIFICATIONS
CONCL.....CONCLUSIONS
DEFIN.....DEFINITIONS
DESCR.....DESCRIPTIONS
DISCUSS.....DISCUSSIONS
EXEMPL.....EXEMPLIFICATIONS

OTHER ABBREVIATIONS

ARG…….. ARGUMENTATIVE WORDS
AWL..... ACADEMIC WORD LIST
BBI.....BBI DICTIONARY OF ENGLISH WORD COMBINATIONS

CALL.....COMPUTER ASSISTED LANGUAGE LEARNING
DC.....DICTIONARY OF COMPUTING
DCA……DETAILED CONSISTENCY ANALYSIS
DCL……DETAILED CONSISTENCY LIST
DCTI.....DICCIONARIO COMENTADO DE TERMINOLOGIA INFORMATICA
DIST..……..... DICTIONARY OF INFORMATION SCIENCE AND TECHNOLOGY
DIT.....DICTIONARY OF INFORMATION TECHNOLOGY
DLIM.....DICTIONARY OF LIBRARY AND INFORMATION MANAGEMENT
DM.....DICTIONARY OF MULTIMEDIA
DPCI.....DICTIONARY OF PERSONAL COMPUTING AND THE INTERNET
EAP.....ENGLISH FOR ACADEMIC PURPOSES
EDLSID..... ELSEVIER'S DICTIONARY OF LIBRARY SCIENCE INFORMATION AND DOCUMENTATION
EE……ENGINEERING ENGLISH
EGAP.....ENGLISH FOR GENERAL ACADEMIC PURPOSES
EIS.....……ENGLISH FOR INFORMATION SYSTEMS
EIST.....ENGLISH FOR INFORMATION SCIENCE AND TECHNOLOGY
ESAP.....ENGLISH FOR SPECIFIC ACADEMIC PURPOSES
ESP.....ENGLISH FOR SPECIFIC PURPOSES
EST.....ENGLISH FOR SCIENCE AND TECHNOLOGY
ETI.....ENGLISH FOR THE TELECOMMUNICATIONS INDUSTRY
ETTE.....ENGLISH TEXTS FOR TELECOMMUNICATION ENGINEERING
FOLDOC.....ON-LINE DICTIONARY OF COMPUTING
GAD…… GENERAL ACADEMIC DISCOURSE
GPT.....GLOSARIO PRÁCTICO DE TELECOMUNICACIONES
GRAM / DISC…….GRAMMAR / DISCOURSE WORDS
HKUST.....HONG KONG UNIVERSITY OF SCIENCE AND TECHNOLOGY CORPUS
ICIG.....INTERACTIVE CONNECTIONS INTERNET GLOSSARY
KWIC…… KEY-WORD-IN-CONTEXT
LOB.....LONDON-OSLO-BERGEN CORPUS
LSP……LANGUAGES FOR SPECIFIC PURPOSES
M.I......MUTUAL INFORMATION SCORES
ODC.....OXFORD DICTIONARY OF COMPUTING
OEC.....OXFORD ENGLISH FOR COMPUTING
PRO……PROCEDURAL WORDS
SELMOUS……SPECIAL ENGLISH LANGUAGE MATERIALS FOR OVERSEAS UNIVERSITY STUDENTS

TERMITE.....TERMITE DATABASE OF TELECOMMUNICATION TERMS

VEBD.....VOCABULARIO PARA LOS ESTUDIOS DE BIBLIO-DOCUMENTACIÓN

VOCAB563.....SMALL GLOSSARY OF TERMS: VOCABULARY 563

INTRODUCTION

The study of lexis for specific purposes or in specialized languages is an area of work that has received a great deal of attention over the past three decades. Sorting and systematizing vocabulary according to subject fields and areas has been done in order to identify the frequent use of words in given scientific-technical studies and professions. Parallel to this, the chore of defining functions of languages in specific situations of use has also captured much interest. Thus, vocabulary has generally enjoyed a prominent position at all time periods and trends of language teaching, while it has actually achieved special significance and relevance with the computerized or empirical observation of its employment for the design of teaching priorities and curriculum organization in English for Specific Purposes.

During the 20th century, the focus on lexis for language learning has produced several works that aimed at providing lists of relevant vocabulary for the learner to use in his / her linguistic training. However, not until the introduction of mini and micro-computers in university settings in the 1960s, the first extensive and less intuition-based analysis of word patterns was achieved. This project, begun in the U.S., soon became the model to follow in Europe in 1975 (London, Oslo and Bergen). Since then, many scholars have undertaken less or as ambitious entreprises with the creation of glossaries, dictionaries, teaching and reading materials. The main goal in such a work with lexis has chiefly been the ability to convey novel approaches to vocabulary that might make English registers or varieties of use more transparent and easier to identify.

In the late 1980s and early 1990s, a further convergence in the direction of lexical aims took place by means of two main occurrences: The processing of larger amounts of words and texts than ever before, and the empirical demonstration that vocabulary behavior is measurable and predictable in given types of discourse. A new methodology was thus being fostered, focusing on the way words interact and are used in different fields and modes of language. Because of this strength of lexical analysis, by the end of the decade, awareness has grown in the form of computerized means for the observation of word patterns in given stretches of discourse of any type (e.g. academic articles, lectures, news broadcasts, e-mail messages, etc). Likewise, the realization that the teacher can carry out his / her own research in such a direction has also set in at varying degrees and levels.

In this line of work, the present dissertation sets out to provide lexical feedback in a subject area of great importance for the current time period, when information and technology occupy prevalent positions. At the crossroad of the two millenia, indeed, a demand for knowledge about technology and the English language, especially in relation to the world of Information, is expanding greatly, and the academic environment cannot be isolated from such developments, as the `Informe Bricall´ has recently set forth (*El País*, 27 March 2000, p. 43). According to this report, Spain and the European Union in general must still achieve further competence in educational aspects relating to Information technologies at university level. This suggests, for our purposes, continuous analysis and monitoring of the type of English highly and significantly used in specialized or technical areas of Information technology fields.

Our focus thus humbly originates from the perceived call for intensive and widespread studies on the English register of technology and science in the case of information and documentation studies, computing and the internet. We aim to analyze the written academic

English of Information Science and Technology, in this respect, so that we may foster the construction of a nuclear or common core language in these studies at the tertiary level of education. Our suggested title "A Lexical Common Core in English for Information Science and Technology", aims to bring, in this setting, four main Information-related disciplines under analysis: Computer Science,[1] Information Science (Library Science and Documentation),[2] Telecommunications (Technical Engineering in Telecommunications, Specialization in Optical and Radio Communications)[3] and Audio-visual communication[4]. The primary objective is to regularize lexical occurrences that are repeated across these fields in order to examine their use in the academic texts. The results are meant as support data for teaching material elaboration and syllabi design in English for Information Science and Technology, sifted and structured with the aid of electronic concordancers and parsers. These are software applications suitable and convenient for the management of large quantity of text and lexical data. In this line of work, our research is developed by following current trends of corpus-driven lexical analysis.

Our chief perception springs from the measurement of a common core lexis in the four

[1] Computer Science comprises three different studies in our case: the Bachelor in Computer Science Engineering (`Ingeniería Informática´), Technical Engineering in Computer Technology for Management (`Ingeniero técnico en Informática de Gestión´), and Technical Engineering in Computer Technology for Systems (`Ingeniero técnico en Informática de Sistemas´) --according to the English translation provided by the study programs of *Universidad Politécnica de Valencia*.

[2] The name `Library Science´ is preferred for `Biblioteconomía´; `Librarianship´ is used less in American Universities, while `Information Science´ includes both Library Science and Documentation studies (see chapter 1).

[3] This discipline is referred to as `Telecomunicaciones, especialidad en Sonido e Imagen´ at the *Universidad de Extremadura*. After checking the English equivalent at some Anglo-saxon universities and companies (e.g. University of Pittsburgh in U.S.A. or Multec in U.K.), and Spanish `titulaciones´ translated into English (e.g. *Guía del alumno extranjero* at the *Universidad Politécnica de Valencia* or Department of Signal Theory and Communications at the *Universidad de Vigo*), the name given above is favored.

[4] According to the *Instituto universitario del audio-visual* at the *Universitat Pompeu Fabra*, the term `Audio-visual communication´ is chosen to transfer the meaning of the Spanish `titulación´ `Comunicación audio-visual´. In the Anglo-saxon world, this encompassing term referring to a Bachelor in audio-visual media (TV, radio broadcasts, film analysis, video technology, etc) is not found; the closer equivalents being film and arts studies, radio communication, etc.

mentioned fields. This implies that we may encounter different degrees of lexical use in such a linguistic kernel that must be carefully observed and registered for teaching purposes. In other words, the concept of nuclear vocabulary in Information Science and Technology academic areas is subject to vary according to the frequency and range of the lexical items across representative texts. In this respect, our main analysis should lead to a prevailing statement: that common core word study is described by considering variation in terms of patterning and repetition. Our conclusions to this research, as a result, closely depend on the type of findings encompassed in the management of such vocabulary variables.

The contents of this dissertation are thus conceived and arranged as follows. First, a detailed account of the theoretical background on which this research is based is given in chaptern 1. Attention is mainly paid to linguistic approaches from a corpus-based perspective. In our specialized language study, this involves briefly explaining some Applied Linguistics concepts: in particular, being familiar and dealing with the notions of Register, Discourse and Genre Analysis, all of which exert their particular influence on the conceptual plane of this thesis. Some history and developments in ESP (English for Specific Purposes) are also reviewed, and key sources are pinpointed as useful reference to our work. Then, the relevance of lexical approaches in ESP is examined in order to determine the nature and perspective of the lexical items sought: for example, either technical words with a more specialized sense in the texts, or more general or vernacular meanings of items having a special significance in Information Science and Technology, all accounted for as common core elements.[5] In addition, in a manner of echoing Sinclair's 1991 book,[6] the concepts of corpora, concordance and collocation are analyzed in relation to the study of lexis in academic scientific-technical

[5] In the EST electronic list, e-mails by professors Chandler-Burns (26 April 2000) and Anthony (27 April 2000) address the need of working with both collections of technical and general words, since the former is distinctive of a single field such as Computer Science whereas the latter tends to meet academic aims.

fields. Those are central for our corpus-driven study.

In chapter 2, the aim is to demonstrate the practical applications of the theory laid out in chapter 1 for the elaboration of the framework with which to achieve the end results of our research. In this respect, for the construction of our corpus, register, discourse and genre features serve as contrast / comparison indicators and guiding posts for the types of queries we propose to formulate in the compilation of texts. Likewise, ESP scholarship provides useful orientation when selecting and discriminating texts. Helpful contributions of such reference sources are also considered for establishing a categorization of lexis in our corpus (e.g. the distinction between common core academic and subject items – the former tending to encompass wider textual scopes and senses, whereas the latter may be more technical and restricted). The works dealing with the lexical focus are particularly relevant in this respect.[7] Finally, in the process of managing vocabulary for arrangement and storage, the concepts gathered from Corpus Linguistics (corpora, concordance and collocation) are quite valuable.

Chapter 3 compiles the most significant findings of the analysis of common core lexis across the four disciplines researched. These are examined according to word frequency and range, collocational strength, consistency values and keyword status. A contrastive study is also made according to each discipline or domain in order to describe lexical profiles at distinct levels or planes. The discussion in the following section –chapter 4—is the product of the study of lexis in specific contexts: scientific area and field (register), academic genre and text type. The range of common core lexical items examined in chapter 3 is assessed as serving three main functions in the organization and planning of lexical resources: discriminating subject-related from academic lexis, identifying ESP-based procedural and argumentative vocabulary, and recognizing relevant discourse and grammar features in this

[6] The title of Sinclair's work (1991) is *Corpus, Concordance, Collocation*.

common core list.

Pedagogical implications are also included in chapter 4, described in order to show the direct application of our lexical analysis and results regarding syllabi and material design. The focus is not, however, the display of elaborated resources, since that would be another project in itself. What is contained in this section deals more with implied approaches to language task conception from the viewpoint of lexical patterns: The idea that word combination occupies a very important position in language acquisition and communicative learning, thus deserving prominent attention.[8] A brief examination of how words operate to allow the development of lexical syllabi in English for Information Science and Technology is provided; this occurs at three levels of interaction: words considered in relation to texts and teaching materials, graphics and as rhetorical functions.

In the section dedicated to the conclusions of our research, the notions and significance of the data are reviewed as active components of language description and learning research. Conviction is also placed on the assumption that lexical analyses in innovative technical settings must be made in detail and carefully so that continuous insight and feedback may be gained regarding language and content issues. The bibliography not only includes all the sources cited in the text of this dissertation, but also the references of the corpus analyzed as well. Finally, the appendices supply all the lexical information pointed out in the Results and Discussion sections.

It should be made clear that, as already implied above, the prevailing objective of our research is not the exhaustive compilation of vocabulary or word lists (e.g. for dictionary or glossary-making) as that would entail a more detailed focus on lexemes and word-forms from

[7] Additional sources will be examined in chapter 2, not included in the previous chapter, as long as they point out relevant data for the section.

the lexicographic viewpoint, which is also applied to a certain extent in this thesis. Our lexical analysis is closer to lexicology approaches. The main purpose being, in contrast, to offer a corpus-based analysis of lexis which pursues two specific goals: To gather empirical data that provides insightful feedback on how words are used in academic texts belonging to Information disciplines, and, secondly, to conduct such research for the purposes of developing ESP learning strategies or approaches to the acquisition of lexis in context – mainly through reading, writing and listening / speaking monologue tasks. We propose that these would substantially contribute to the suitable description and organization of materials and course syllabi.

[8] The communicative approach is obviously introduced briefly, since authors like Swan and Widdowson clarify how widely the notion can be framed (Rossner & Bolitho, 1990), and thus, our focus is placed on lexical tasks almost directly.

CHAPTER 1: THEORETICAL BACKGROUND

In the development of this first section, the aim is no other than to expose those Applied and Corpus Linguistics concepts and trends for delimiting and specifying the current idea of the type of analysis carried out in this research. The application of the analysis of our data, as outlined in the following chapter, would be missing much scholarly perspective and support without the inclusion of this first part of scientific justification of the theoretical background.

Some divisions are made to classify the information according to key concepts, those of register, discourse, and genre as introductory scopes in the analysis, then the area of ESP is examined so that the groundwork for operation is established, and the importance of the lexical focus is presented by compiling some relevant works. Each presents a theoretical approach in Applied Linguistics for the analysis of lexis in specialized languages. The academic work that deals with the study of Corpus Linguistics is finally closely followed, as these notions propose detailed feedback for the orientation of our research. The developments are those of corpora, concordance, and collocation, under which we specify our findings.

The main aim of this revision is to provide insight into the study of lexis from different viewpoints and over specific time periods. However, we notice that there are pivotal points in the bibliography accounted for which may overlap or coincide in terms of significant claiming for our analysis. Thus, we take from each source or set of sources within the given categories those statements and contributions that best identify the scope of our study.

I. REGISTER

The notion of register should be the first of the background theory to be outlined, as it is basic to our lexical work. It is a linguistic concept that maintains a close relationship with sub-language and restricted utterances according to language variety. As Jeffrey Ellis states,

> By register itself, a linguistic, not situational category, is meant a division of idiolect, or of what is common to idiolects, distinguished by formal (and possibly substantial) features and correlated with types of situation of utterance... It is a paradigmatic term; the corresponding syntagmatic one is register-features of the utterance (Ellis, 1966: 83).[9]

From this early definition, in fact, three relevant aspects may be extracted regarding register: linguistic category, division of idiolect, and register-features. Let us examine each particular notion more closely as is elaborated in later works:

I.1. LINGUISTIC CATEGORY.

Because it is thus narrowly associated with meaning, Register Analysis is based on linguistic parameters that rely to a great extent on contextual meaning; such a perspective `relates form to situation´(Ellis, 1966: 80). Within this view, Firth's concept of `context of situation´ (1957) plays an outstanding role, and is further elaborated by scholars, e.g. Halliday (1966). It is an essential notion to the understanding of the contribution of register as a linguistic theory in specific language domains. According to this development, register

> can be defined as a configuration of meanings that are typically associated with a particular situational configuration of field, mode, and tenor... the expressions, the lexico-grammatical and phonological features, that typically accompany and realise these meanings... [are] indices in the form of particular words (Halliday and Hasan, 1985: 38).

[9]Ellis and Ure (1969: 251) briefly recall the history of the concept of register, put forward by Reid (1956), applied to translation by Ure (1959, 1963), contrasted and distinguished from other concepts (e.g. style) by Hill (1958), Strang (1962) and Catford (1965), and analyzed in greater depth (e.g. classified) by Dixon (1964), Halliday and McIntosh (1965). The two authors (Ellis & Ure, 1969: 252) also point out that `Firth (1937, 1950, 1959) draws a concept of context of situation´and `restricted languages (characterized partly by patterns of lexical collocation)´in close association with the notion of register. Even the phenomenon of intonation is itself submitted under analysis at this stage as a system to be observed according to patterns (McIntosh, 1963:111).

As Halliday and Hasan go on to explain (1985), in this conception of register as meaningful units of language use, `any instance of living language that is playing some part in a context of situation, we shall call a text´(ibid., 10).[10] Register is described, in fact, as Widdowson puts it, `in terms of language use´(Widdowson, 1983: 10). In this description, issues such as degrees of politeness (i.e. formality), range of topics, actitivity, work, purpose, and mode are dealt with (Finochiaro and Brumfit, 1983: 25). In this respect, register should thus be distinguished from the notion of dialect, since, as Halliday remarks (1978: 185), registers say different things whereas dialects say the same thing with different voices.[11]

Because of this fixed position of language use according to register, in fact, as Biber et al. explain (1998: 35), `it is impossible to generalize from one register to the patterns that will be found in other registers´. There is actually a range of `situational characteristics´(ibid., 157) such as subject matter and methods of presentation that lead register analysts to describe register according to special or restricted use, which leads us on to the second point.

[10] According to this analysis, the text is thus both `product and process´ and `meaning unit´(Halliday and Hasan, 1985: 10). In addition, Hasan goes further to explain that there is `intertextuality´ in association with the notion of register, `in which the classroom is one long text´(ibid., 47).

[11] In other words, `registers are defined according to their situations of use (considering their purpose, topic, setting, interactiveness, mode, etc)...[while] dialects are defined by association with different speaker groups´(Biber et al, 1998: 135). In this respect, a speaker or writer may be able to produce different registers according to contexts of situation, whereas the same person does not have a command of various dialects.

I.2. DIVISION OF IDIOLECT.

This concern with the key features of registers is obviously linked to the study of special languages or sub-languages. Ooi's work (1998), from this viewpoint, defines register as similar to sub-language, and he quotes Sager (1986: 2):

> a language used by a particular community of speakers, say those concerned with a particular subject matter or those engaged in a specialized occupation (Ooi, 1998: 124).

This approach serves the useful purpose for special languages to develop more coherently and empirically. As an example --offered by Halliday (1978: 197)--, the four hundred years that took to develop the English registers of Mathematics, Science and Technology can be examined through diachronic approaches to texts. Synchronically, according to Ahmad and Davies (1997: 159), `differences also manifest themselves across the varieties of English, across registers and so on´.

For English for Specific Purposes (ESP) pedagogy, Register Analysis proves to be highly useful: to teach, as Biber et al. (1994: 174) say, `the markedly different patterns of linguistic form and function that occur across registers´. This is actually the type of work that has often been carried out in ESP ever since scholars such as Ewer and Hughes-Davies (1972: 47) appealed to the importance of `quantitative data ... in the professional training of scientists and technologists´: data, for instance, from the sub-register of `instructional English´ in manuals, handbooks and engineering textbooks (Ewer and Hughes-Davies, 1972: 52).[12]

The application of register studies to ESP then is far from new.[13] Today, as Carter and McCarthy claim (1997 a), what were defined by Sinclair (1966) as `specific habitats (of

[12] The notion of register is often identified as `sub-register´, although it seems to refer to the same concept of sub-language that may be more or less factual, impersonal, etc. (Biber et al., 1998:158).

[13] Another way to put it is worded by Dudley-Evans and St. Johns when they state that register analysis `has come back into something of a prominent position´(1998: 31).

subject area)´ are recurrent fields or scopes in ESP and Genre Analysis, overlapping as key concepts in Applied Linguistics and the Linguistics of Corpus analysis (Carter and McCarthy, 1997 a: 36).

I.3. REGISTER-FEATURES.

Indeed, according to the aforementioned words by Sinclair in 1966, the study of register is mainly based on features of language use. Particularly, he concentrates on `normal collocations´ and `word associations´(Sinclair, 1966: 45). Lexical collocations and multi-word vocabulary are thus crucial elements in the analysis of register. As McCarthy (1990) states concerning this fact,

> there is a close relationship between the idea of field in a register and the schematic vocabulary of a given subject area in that the schematic vocabulary realizes the field recognized by topic(...) A further important point is(...) collocation in characterizing particular fields (...)and multi word items (McCarthy, 1990: 62).

As this author goes on to explain, `computer analysis is a very good way of getting at the vocabulary of a register´ (ibid.,64). From this type of enquiry, we can learn much about `frequency within a particular register´(ibid.) because `words acquire registerial appropriacy only in context´(ibid.,103), and register can thus be best approached by determining how words do actually behave in texts. Register-features are exposed by these at the syntagmatic level.[5]

In the following sections, and throughout the rest of the chapter, we shall explore these ideas more closely –e.g. lexical traits in registers— from other fields in Linguistics, such as the more modern contributions of Corpus Linguistics; in this line of work, the observation of discourse features for lexical analysis is examined next.

---

[13] However, as Ellis states, the identification of register as a paradigmatic concept is also valid (1966: 83).

II. DISCOURSE

This is the second concept[15] that should be revised in its relation and application to our lexical work. The definition of discourse as `a set or collection of connected language units... that together make up a coherent, cohesive text´ (Trimble, 1985: 2) is central to our approach. We refer to the rhetorical features of texts that readers activate by sharing with writers the postulates of scientific discourse. These affect `surface syntax in EST texts so drastically that language and subject cannot be discussed separately´(Selinker and Trimble, 1974: 83). A main aspect to be considered with our lexical focus is thus discourse as macro-structure, and, observed in science and technology, its study is mainly carried out with surface lexico-grammatical features that establish patterns of coherence and cohesion.[16]

II.1. COHERENCE AND COHESION.

The study of textual coherence and cohesion is relevant in that it affects how lexical units relate to each other, `spelling out exactly the relationship holding between two parts of a particular discourse´ (Hoey, 1983: 28). The interaction `writer-reader´ is therefore maintained through the use of coherence and cohesion devices more or less explicitly; the reader is continually asking questions in his / her interpretative process (ibid., 170).[17]

The idea of a `discourse community´ is therefore called upon by scholars. In the

[15] Discourse analysis would be found, according to Munby (1978: 21), near Sociolinguistics in the `linguistic competence constellation´, in close association with Wilkins (1976) and Halliday's (1978) notions of socio-semantic approaches (also Van Elk and the British Council [1975] play relevant roles in this respect).

[16] These devices signal process, instruction, classification and other functions characterizing scientific-technical discourse. There are obviously other characteristics and concerns in discourse studies, such as `involvement of speaker and hearer, their purposes, beliefs; the medium, and its properties; cultural considerations ...´(Ruhl, 1973: 17). A web page is maintained on the internet by Foltz (1999), where the relationship between readers or users on the coherence and comprehension of both linear and hypertextual texts is made explicit (http://www-psych.nmsu.edu/~pfoltz/reprints/).

[17] This is somewhat related in general discourse analysis to what Ross describes as `the pragmatics jungle´ that must be domesticated `to generalize from contextual features in the form of syntactic and semantic patterns´(1969: 2).

context of science and technology, it encompasses `the specialist knowledge of the subject matter and the field of knowledge for which the article was written and the audience for whom it was written´(Winter, 1982: 179).[18] For these readers, coherence is underlined in scientific-technical discourse by the writer's skills: `he states his intents clearly, defines key terms and emphasizes themes related to methods and results´ (Tarantino, 1991: 32).[19] Accompanying this process, cohesion comes into play mainly by the choice of lexical ties (Halliday and Hasan, 1985)[20] or links (Hoey, 1991): `sentence conjunctions, collocations, reference... repetition, paraphrase...´(Hoey, 1991: 83). In this sense, `lexis overlaps both grammar (syntax) form and phonology´(ibid., 208).[21]

II.2. RHETORICAL FEATURES.

As Van Dijk (1985) points out, the importance of both cohesion and coherence is obvious in discourse.[22] In scientific-technical texts, this is usually accomplished by means of rhetorical features such as the functions of description, definition, classification, and so forth, and techniques or patterns of comparison, order of importance, exemplification, etc. (Trimble,

[18] Baunman also makes reference to this proliferation of scientific `exchange of knowledge... in special books, dictionaries...´(1994: 1), which demands an ever-greater presence of communicative descriptions of language resulting from the analysis of `specific features of specialised subject fields´(ibid.).

[19] According to Nattinger (1997: 75), the type of discourse given by the scientific-technical register is categorized as factual information, qualified by different authors as transactional, ideational (Halliday, 1970), referential or descriptive (Lyons, 1970).

[20] As Hasan explains (Halliday and Hasan, 1985: 73), `the nature of this link is semantic: the two terms of any tie are tied together through some meaning relation (...) the basis for cohesion between the messages of a text´.

[21] Through cohesion, in the surface text, according to J.R. Searle (1969), words are `mutually connected within a sequence´(3); in this network, it is shown `how already-used structures and patterns can be re-used, modified or compacted´(ibid., 49).

[22] Widdowson claims that `in all kinds of discourse one can trace propositional development through cohesion and illocutionary development through coherence, and all discourse can be characterized in terms of the relationship between propositions and illocutionary acts´(1978: 30).

1985).[23]

Through the identification of this series of devices, the reader may recognize the `problem solution´ pattern of scientific-technical discourse (Widdowson, 1983: 81), `modifying knowledge in the light of new information´(ibid., 67).[24] In this sense, as Tang states (1992: 180), these rhetorical means are `knowledge structures [Mohan, 1986] that are fundamental across the curriculum... reflected in the macrostructure of text´. They constitute `a conceptual framework for integrating language and content´(Tang, 1992: 180).[25] This is mainly done, according to Mosenthal (1985), through practice in synthesis and analysis:

> (…) to synthesize, both semantically and syntactically, the features and examples in lower specification levels into more generalized features and examples (...) In analysis, one begins with a topic or a set of generalized categories. One then deduces features and examples that further characterize these categories (Mosenthal, 1985: 406).

What learners actually do involves the `combination of schema-based approach and a lexical relations approach to text´(Carter and McCarthy, 1997 b: 205). Thus, reading benefits from lexical work as a `vocabulary-building tool´(ibid.). While instantial relations are recognized, vocabulary items are linked in discourse (ibid., 204). Lexical processing thus accompanies the rhetorical focus on reading scientific-technical texts. In the study of discourse, this is clear according to many authors and from the theoretical perspective of this study on lexis:

---

[23] These signal, as Widdowson (1983) tells us, `schematic constructs to be applied to language events´(57); through them, readers negotiate meaning: they `develop the ability to exploit the resources for meaning´(ibid., 26).

[24] This is, again, related to the reader's capacity for modifying schema, thus negotiating meaning and recognizing coherence and cohesion (Widdowson, 1983: 70). According to schema theory, `a text is an interactive process between the text and the prior background knowledge or memory schemata of the listener or reader´(Carrell, 1982: 482).

[25] A good example is the use of graphic displays by students in science and technology texts `to construct a schematic mental representation of the subject matter which is then elaborated by means of the text´(Schnotz et al., 1993: 196).

> (…) that discourse analysis alone, as presently conceived following the pragmatic work of Sinclair and Coulthard (1975), is inadequate as a theoretical approach for the description of text and must be supplemented by analysis at the level of lexis (Phillips, 1989: 3).

The type of proposal of discourse study underlined by these scholars thus focuses on the observation of lexical items and rhetorical devices in context; those are measureable while the latter is targeted in the form of academic texts for our research purposes.

II.3. SURFACE LEXICO-GRAMMATICAL FEATURES IN ACADEMIC DISCOURSE.

As Phillips claims, the study of macrostructure in academic discourse, following Van Dijk (1973), is based on the textbook chapter because it is felt

> that the particular importance of terminology of this kind of text might mean that it offered the best chance of testing out the theoretical ideas outlined (Phillips, 1989: 11).

The discourse unit is thus the textbook chapter in Phillips's study, a `field of discourse´ in Hallidayan terms (Halliday, 1978, 1985, 1990).[26] The text is integrated by lexical items that make reference to `underlying concepts´ through repetition and synonymy (Yekovich and Walker, 1978: 274). What we are thus experiencing are lexico-grammatical relationships as `a world of things, symbolically fixed to be observed, measured, reasoned about, brought to order´(Halliday, 1990: 21). These words establish patterns and the organization of texts in academic settings, particularly science and technology, contributing to shape the main ideational or transactional type of discourse in the register.[27] To convey such concepts and

[26] According to Halliday, `any instance of language that is playing some part in a context of situation, we shall call a text´ as meaning unit (Halliday and Hasan, 1985: 10). In this context, we may distinguish between field of discourse (what is happening), tenor of discourse (role relationship between participants) and mode of discourse (the part of language --function, channel and rhetoric) (ibid., 12). The field of discourse represents the `world of happening´`in which experience is construed in the form of clauses´(Halliday, 1990: 21).

[27] Vande Kopple sees this same ideational direction in his study of the expository style by examining words and themes (1991). According to him, by investigating how these `progressions´ are produced, `such intertwinings

process in scientific discourse, the compressing of information in specific lexical traits, such as nominal compounds, works effectively in technical texts (Pueyo and Val, 1996: 252).

The given diatype of scientific-technical writing can be approached, in fact, by working with lexical and grammar vocabulary that is common to different disciplines, according to some approaches to discourse.[28] In order to summarize their relevance in this respect, three points may be made regarding the type of discourse studies conducted in relation to lexico-grammatical features and content areas across academic disciplines: II.3.A. for curricula organization aims, II.3.B. construction of nuclear vocabularies, and II.3.C. the semantic analysis of vocabulary.

II.3.A. Organization of the curriculum by focusing on common core lexico-grammar items.

Halliday (1991) serves as the main reference for the notion of the lexico-grammatical component of language, regarded here `as a unified phenomenon, a single level of "wording", of which lexis is the "most delicate" resolution´(Halliday, 1991: 32). In this continuum, lexis would thus be fundamentally `mixed´, i.e. represented by all content or open-ended single and multi-words (nouns, adjectives, adverbs, verbs), as well as grammar items, in the form of closed-ended items such as conjunctions, prepositions, auxiliary verbs, articles, etc. A lexical unit may thus consist of a lexical part and / or grammar component (e.g. multi-word terms and phrasal verbs), described significantly in their combinatory aspects with other words as either collocations (content items) or colligations (grammar elements), as shall be examined in this chapter.[29]

---

might help distinguish genres´(ibid., 331).

[28] In the section dealing with lexical study in ESP, these concepts shall be examined in more detail.

[29] This conception is compatible, according to Hudson, with that which sees the word or lexeme as the most basic unit in linguistics (1994).

This focus is adapted by authors whose main concern is language in use for pedagogical ends: `The specification and organization into a curriculum of the pedagogical methods and techniques, and the discoursal, syntactical and lexical features relevant to learners' needs´(Markee, 1989: 136). This means that `teaching for communicative competence focuses on teaching language as use, without thereby necessarily ignoring considerations of syntactic and lexical correctness´(ibid., 142).[30] Both the analytical and synthetic levels combine to produce `regularities... patterns in observed behavior which are regular´(Gregg, 1989: 22). Through the realization and assimilation of these patterns, the common lexical items of language varieties are used and thus learned.[31] The language program in a given register of language should then integrate these variables according to use in different scientific-technical disciplines (i.e., Social Sciences, Information Science, etc).[32]

II.3.B. The concept of nuclear vocabulary.

In dealing with the nature of academic texts, the study of discourse focuses primarily on the expository style of textbooks as its main field of analysis for subjects across the curriculum. In the sense that `expository texts serve the purpose of updating knowledge, they are used in so-called content areas´(Mosenthal, 1985: 388). Already in secondary schooling, the importance of reading this type of text is noticeable and widespread.[33] For such a task, students must master a common core or nuclear vocabulary that many scholars have strived to

[30] Widdowson's definition of use `covers that which makes evident the extent to which the language user demonstrates his ability to use his knowledge of linguistic rules for effective communication´(1978: 3). Chomsky qualifies the `use of language as an essential feature, no less than the distinctive structural properties of language´(1975: 56).

[31] According to Kyyssönen, `we need to enquire into the way grammar relates to lexis, ... how it functions in the regulation of language use´(1997: 334).

[32] This has been done in working with ESP by various authors (see section dealing with ESP in this chapter).

[33] As Beaugrande states, since this age, the text implies `a textual world... the cognitive correlate in the mind of a text user for the configuration of concepts activated in regard to a text´(1993: 77).

identify.[34] This delimiting chore is not an easy one, since the texts chosen must be truly representative and all English diatypes and registers must be included. As West (1953) puts it, quoted by Stubbs (1986),

> To find the minimum number of words that could operate together in constructions capable of entering into the greatest variety of contexts has therefore been the chief aim of those trying to simplify English for the learner (Stubbs, 1986: 99).[35]

In the academic arena, these words should be `common in technical texts, academic papers and newspaper reports´(Carlson and Tannenhaus, 1984: 23). Very often, they are content words that have a `text-structuring function and signal the logic of the text´(Winter, 1978: 92): Some examples are **cause**, **kind, achieve**, **addition, feature, method, truth**.[36] These words establish a lexical reference to the micro and macrostructure of the expository text,[37] and, since they are repeated across academic texts, they should be included as nuclear lexical items.[38]

[34] According to Stubbs (1986: 99), as early as 1930, Ogden's *Basic English* contains linguistic items as an international auxiliary language; Thorndike and Lorge's *Teacher's Workbook* (1944) is based on lexical items taken from the Bible and the English Classics (three million words). West's (1953) *General Service List*, Kucera and Francis's (1967) *Computational Analysis of Present-Day American English* and others (see section dealing with corpora) have also presented nuclear vocabulary based on large bodies of texts or corpora.

[35] Richards et al.'s *Dictionary of Language Teaching and Applied Linguistics* (1992: 64) offers the following definition of common core language: `(in language teaching) those basic aspects of a language (e.g. vocabulary and grammar) which a learner needs to know whatever his or her purpose is in learning the language´.

[36] For more on this type of common core words, see section dealing with lexical approach.

[37] As Perfetti and Goldman state (1974: 77), `a more general requirement of understanding a discourse is establishing reference´.

[38] There are obviously many other words that must be included. According to Rudzka et al. (1985), for instance, those words with the highest number of collocations; according to Nation and Coady (1997), those that show more variety and range across subject matters, etc. (See more on these principles in the section dealing with the lexical focus).

II.3.C The relationship between words and concepts in EST development.

Across the curriculum, a key academic skill is no doubt that of understanding by bringing one's own textual world to the text (Beaugrande, 1993: 77). In this respect, according to Fillmore (1969), conceptual schemata or frameworks are linked together in the categorisation of verbal actions, which define grammar and language relations.[39] In terms of the word as the central unit of language, lexico-grammatical provisions in a given language, in this manner, name and describe the categories and relations found in schemata (Fillmore, 1969: 365). They do this by holding sense relationships between them: `a matter of interlexical and intralingual relations´(Lyons, 1995: 80).[40] In discovering and coming to terms with these lexico-grammatical relationships, the reader will succeed in understanding the text.[41]

In the case of scientific-technical literature, the members of this discourse community can share concepts and experiences; these are formed in the search for meaning: the `study of the relationship between linguistic objects and the mental states and processes involved in their production and comprehension´ (Winograd, 1976, quoted by Kayser, 1989: 343).[42] In this type of discourse, the writer directs the reader along by means of specific lexico-grammar

[39] As an example, Fillmore (1969: 365) speaks about ergative verbs and their relationship with specific nouns (e.g. **open** and **door**). In such a fashion, as Lehrer says, `the words of a language can be classified into sets which are related to conceptual fields´(1974: 15).

[40] These relationships or bonds constitute, according to Potts (1994: 122), `frames... originally a contribution to the representation of knowledge, rather than of meaning, but have subsequently also been enlisted in accounts of meaning´. This network is often expressed by the `simple linear progression´ in expository writing, in which the rheme of one sentence (basically the sentence's predicate) becomes the theme (subject) of the next, `in a theme-rheme sequence´ (Firbas, 1966: 239). Word frames are thus motivated by knowledge, meaning and sense (Fillmore, 1969).

[41] Lexical relationship management is quite relevant in the learning process, indeed, as, according to Clark (1993: 10), `each semantic field covers a conceptual domain, and the meaning relations that link all the lexical items in the field show how that conceptual domain is represented´.

[42] According to Danes, the writer's `intratextual relations´ in science and technology `belong to the sphere of the author's way of subject-matter presentation´ (Danes, 1989: 236).

items: the `ideational segments of the discourse´ (e.g. through discourse markers such as **because, and,** and **so** --Schiffrin, 1987: 51).[43]

Likewise, `textual structures found in sub-language (...) reflect very closely the structuring of the sub-language's associated conceptual domain´ (McNaught, 1993: 233); due to this fact, the suitability of dealing with academic and subject-based vocabulary that describes this domain is rather convenient. The semantic-conceptual organization of texts (Meijs, 1992: 131) is developed as `words activate, "call up", other words that are related to them in meaning´ (ibid.). This approach to the concept in a specific community (e.g. ESP students and colleagues) should thus `help students to process, utter and write more than single, textbook-bound sentences´ (Robinson, 1980: 53).[44]

Because concepts and words vary with time, as Ullman reminds us (1962: 197-210), context is a key measuring device for updating lexical knowledge.[45] Leech (1981) points out that such a process is considered within the particular category of `conceptual meaning´ as dependent on the notion of contextualism (Firth, 1930). According to Firth, quoted by Leech (1981),

> (…) as we know so little of inner mental states, even by the most careful introspection, the language problem becomes more mysterious the more we try to explain it by referring it to inner mental happenings which are not observable. By regarding words as acts, events, habits, we limit our inquiry to what is objective in the group life of... fellows (Leech, 1981: 61).

Although Leech does not agree with Firth entirely, he proposes this train of thought as a

[43] In fact, according to Wenger and Spyridakis (1993: 347), `information designers need to consider both word- and text-level factors when designing technical information... both factors need to be assessed in the context of the intended audience's familiarity with the material´.

[44] In EST discourse, particularly, the focus on both reading and production actually starts in the `conceptual paragraph´ (see, for instance, Selinker et al., 1976: 284).

[45] This is what leads Summers to rely on the `prevailing view that newly encountered words should only be decoded by means of contextual clues´(Summers, 1997: 112). This is the case of the learner who, `exposed to words in a variety of different contexts,... forms a well-rounded concept of the word's meaning and use´(ibid., 113).

suitable means to regard more or less precisely how conceptual meaning changes with time:

> (…) that context is an undeniably important factor in communication; and let us consider how this semantic role of context can be allowed for within a theory based on conceptual meaning (Leech, 1981: 66).[46]

The meanings of words, in this respect, as McInstosh states,

> (…) are in some direct way associated with our experience of that word in a variety of contexts, our association of that word with other words which have, in our experience, a somewhat similar range (McInstosh, 1961: 333).

This type of meaning is obviously limited to the textual plane in our case in point. It is represented in the notional category of lexico-grammar features that relate to characteristic rhetorical means of science and technology discourse (i.e., definition, comparison, etc) (Candlin, 1976; Trimble, 1985). As Salager says, `lexical (conceptual) rather than syntactic factors are responsible for errors in comprehension´ (Salager, 1987: 24) among science and technology students.

What linguistic devices are used as indicators of rhetorical acts is a key methodological procedure in this respect (Salager, 1987: 24). How lexico-grammatical items behave in texts is thus quite relevant to the perspective of discourse analysis adapted in this dissertation, that of analyzing lexical items used in authentic texts. In agreement with Sinclair, it is `an essentially systemic viewpoint(...) the question of a core vocabulary(...) and the chaining of lexis in text´ (Sinclair, 1988:73).[47] Word use and sense in the explanation, definition, description and classification of concepts can thus be carefully described and worked with for lexicologic, lexicographic and teaching aims.

---

[46] In Munby's taxonomy of language skills in academic/science and technology settings, conceptual meaning would involve understanding contextual clues such as `quantity and amount, comparison, degree, time, location, direction, means, instrument, cause, result, purpose, reason, condition, contrast, definiteness and indefiniteness´(Munby, 1978: 126). In the oral mode of communication, as opposed to the written mode, knowing how to `read´ stressed syllables as lexical prominance also bears a great contextual value (Brazil, 1985).

[47] This exhaustive and systematic study of lexis is possible `because discourse is full of redundancy, anaphora and parallelism (...) clues for understanding new vocabulary´(Nattinger, 1997: 63).

Another important conception linked to the notions and developments explained so far from the scope of register and discourse is the study of genres.

III. GENRE

For the examination of how the study of genre also forms a close bond with our lexical work in specialized languages, our revision extends to three levels in this section: III.1. Genre as the study of corpus, III.2. Genre as a specification of register, and III.3. Genre as the study of lexis.

III.1. GENRE AS THE STUDY OF CORPUS.

As shall be seen in the section dealing with corpora, a corpus or collection of texts `can be designed to serve as a resource for general purposes´(Ooi, 1998: 35), or, on the other hand, `to be representative of a particular sub-language (roughly equivalent to a language genre)´(ibid.). This amassing of texts for corpus design pays special attention to the notion of genre, indeed. Thus, different texts are collected according to such categories as journalistic, popular fiction, `belles lettres´, commercial, etc. This is what, for example, the study of the definite article, according to Varantola (1984: 105), is dependent on as its frequency is contrasted in a scientific genre `as opposed to other genres in the Brown Corpus´.[48] The point, then, to be made, in close agreement with Dudley-Evans and St. Johns (1998: 75) is that genre distinction and autonomy should be considered when selecting various texts for linguistic analysis, since `frequency counts carried out on a corpus containing a number of genres without distinguishing between the genres may produce results of limited value'.

[48] The Brown corpus is one of the earliest corpora that included various genres in its compilation (see section on corpora).

III.2. GENRE AS A SPECIFICATION OF REGISTER.

As Halliday and Hasan state (1985: 61), there are `obligatory elements that define the genre to which a text belongs´. Unlike register, genre study seeks textual distinctions or similarities `on the basis of external criteria relating to the author's or speaker's purpose´(Biber, 1988: 206).[49]

There may be two different genres, according to this notion, and only one register of texts, such as biography and academic prose, which both have the narrative linguistic form.[50] In order to develop a typology of texts, Biber (1988, 1993), in fact, describes the distinctive and common linguistic features of texts in terms of `multi-feature, multi-dimensional analysis´.[51] This type of approach makes the crucial distinction between genre and register dimensions of specialized languages or sub-languages. Particular texts and structures thus `found in sub-language... reflect very closely the structuring of the sub-language's associated conceptual domain´(McNaught, 1993: 233) as well as the respective genre and register.[52]

---

[49] The difference between genre and register is made more explicit by Swales (1981: 10): `such studies (i.e. of genre) differ from traditional register or sub-register analysis in the importance they attach to communicative purposes within a communicative setting´. In fact, as Miller also claims, a genre can be defined as a rhetorical means for `mediating private intentions and social exigence´(Miller, 1984: 162). The aim of its study is to make students `more sensitive to the rhetorical structures that recur in specific genres´(ibid., 163) and `to show that a genre-centered approach offers a workable way of making sense of the myriad communicative events that occur in the contemporary English-speaking academy´(Swales, 1990: 1).

[50] More strictly, as Swales states, a sub-genre distinction should be sought within a given genre. Thus, within administrative writing, for instance, a `good news´ may be distinguished from a `bad news´ letter (Swales, 1990: 53); Bhatia (1993) also claims the existence of different genres in business correspondence. Genre analysis must thus be `a theoretical activity separable from discourse analysis´(Swales, 1990: 61) in order to make possible the distinction of these communicative features; as early as Lowe (1969), in fact, these communicative events may also be distinguished according to speaker-listener interaction, `be such communication in the form of a business letter, a simple conversation, a preface to a scholarly book, or a political harangue´(Lowe, 1969: 401).

[51] Bowker's doctoral dissertation uses this corpus-based multi-dimensional approach to classify terminology: `compiling a balanced corpus, and adding some features to a knowledge-based system which enable it to better handle multidimensionality´(Bowker, 1995: 11).

[52] It is because of this arrangement of information and concept in readers' minds that genre formation is allowed (Swales, 1990: 83). Swales thus alludes to the `concept of schemata´ introduced by Barlett in 1932; depending on the schemata capacity and management of the readers, `experts differ from novices´ when coping with

III.3. GENRE AS THE STUDY OF LEXIS.

In terms of the detailed analysis to which Swales and other authors refer,[53] without the context of words in sentences and phrases, genres could not be identified properly. According to lexical patterns, in fact, `there will be individual variations in the types and strengths of collocation´ (Hoey, 1991: 219). These `mental concordances´of words thus provide `clues of the generic structure´ of texts (ibid., 225).

How these words are presented and distributed across texts is important for the definition of genres. So is the case for finding `subject core vocabulary´ that `will only be expressive of a particular field´ (Carter, 1997: 172). This type of lexis may correspond to a specific genre: `the more specified lexis is involved, the more precise relationship is possible´(ibid., 173). There is thus

> a stable lexical relationship... but... also dynamic... or instantial, that is, they make unique partnerships or combine or associate to produce meanings specific to that individual text (ibid., 177).

These are what Hasan (1984: 183) identifies as `text-specific´ lexical relations that include collocations for grouping and defining words in texts. They allow the reader's creation of coherence; as Carter and McCarthy state (1997 b: 205), this lexis `conditioned by genre... increase the reader's predictive power and ability to create coherence´. In this sense, a genre gives the specialized connotation to the word, and the word, in turn, helps to define the genre dealt with (e.g. collocations such as `land rift´ of the term `rift´ in geology texts [Gavioli, 1997: 87]). The degree of complexity in genres is therefore given by lexical factors such as frequency, distribution, and collocation. This is implied by Hasan´s definition of lexical items in text as `that calibration of values which motivates the elements of its generic structure

---

different genres of texts (Schnotz, 1993: 153).

[53] See, for instance, more recent works on this issue by A. Johns (1997) or Bhatia (1998).

potential´ (Halliday and Hasan, 1985: 104).[54]

IV. ENGLISH FOR SPECIFIC PURPOSES

In addition to the features of register, discourse and genre, the body of this study is to be based on explicit objectives with an English for Specific Purposes (ESP) framework as the setting. Therefore, the concept of languages for specific purposes and specialized languages or sub-languages[55] will be further analysed and examined in order to specifically describe the theoretical grounds of the research. With such a purpose, this section sets out to present those works in LSP[56] that significantly influence the orientation of this thesis. The scholarship to be mentioned and revised contributes to giving the theoretical perspective in the study of lexis in specialized languages or sub-languages within the realm of ESP. Six core statements are thus proposed under which to classify the ESP notions and ideas that are relevant to the study.

IV.1. ESP AS A RESTRICTED LINGUISTIC DOMAIN.

As a first premise, it must be said that `the primary concern in the LSP course is in achieving something outside of language through the medium of language´(Crocker, 1981: 8).

---

[54] Depending on the detailed analysis of texts, furthermore, the distinction is made between the `level of brow´ (Huddlestone, 1971) of specialized texts. Huddlestone focuses on the type of interaction made between writer and reader, which is a successful approach to distinguishing academic genres. For example, the higher brow (more specialized) text would be represented by scholarly journal articles, in which the writer-researcher addresses his/her colleagues.
For other definitions of detailed genre analysis based on complex rhetorical settings (writer, audience, purpose, professional, social, literary identities, and so on) see, for example, Wilkinson (1992), Tang (1992) and Vande Kopple (1991), who apply genre theory to different aspects of language and structure in texts (e.g. graphics). For earlier accounts of genre work, see, for instance, Stanley (1984), Di Pietro (ed.) (1982) and Haviland and Clark (1974), who seek to establish foundations for genre theory by submitting analyses of social, academic and professional discourse (e.g. bureaucracy).

[55] The three terms are used interchangeably in this study.

[56] Although most of the scholars quoted in this section discuss the nature of ESP (English for Specific Purposes), some work on LSP (Languages for specific purposes), and are thus mentioned if theirs is a particularly relevant

The objective is not proficiency in the production of correct utterances, as it may be the case in general English courses (Strevens, 1971). Instead, as Pauline Robinson states, `only relevant vocabulary, grammar and functions are selected. Only suitable themes and topics are chosen´(Robinson, 1980: 12). In this sense, `some linguistic features may remain receptive...that is, the learners will understand them but not necessarily use them in speaking and writing´ (Finochiaro and Brumfit, 1983: 165). Yet, one cannot mean by this that only lists of specialized vocabulary are sought: `one must also teach the contexts and structural relations within which the words have meaning´ (Marshall and Gilmour, 1993: 75). In other words, those of P. Edwards, the function is to `teach language for the subject specialism´ and `teaching tasks based on the specialized content´ (Edwards, 1996: 13).

IV.2. ESP AS AN ADAPTATION TO THE LEARNER'S NEEDS.

The ESP course must be specified and organized `into a curriculum of the pedagogical methods and techniques, and the discoursal, syntactical and lexical features relevant to learners' needs´(Markee, 1989: 136). For examining the former two --methods and techniques-- clear `communicative purposes´ must be kept in mind.[57] As Strevens explains, `we must know what people in those jobs normally do through language´(Strevens, 1977: 157). Secondly, in tackling the specific language features --discourse, syntax and lexis-- the ESP teacher will encounter `the basic components of language´, as Strevens also points out, which do not differ from General English. Those that change are `the statistical properties of the mixture in which these components occur, and the intention, the purpose, behind their selection and use´(Strevens, 1977: 157). Adjusting to these specific variables of language use

---

contribution to this study.

[57] For example, according to Strevens (1977: 158): seeking advice, stating differences and objections, understanding diagrams, which are communicative categories that apply to the technical work-place.

is a crucial step in the formation of the ESP `mentality´.[58]

IV.3. ESP AS TEACHING ENGLISH TO SPECIFIC LEARNERS.

Robinson's proposal of `teaching English to specified people´ (Robinson, 1991: 5) may serve as a suitable premise from which to consider the great importance of the academic community --that of higher eduction-- and the need for these people to work with discourse. In university ESP, as Robinson herself states in 1980, this community, both undergraduate as well as graduate students `process, utter and write more than single, text-book bound, sentences...looking beyond the textbook to...encounter real discourse´ (Robinson, 1980: 53). According to Winter, the ESP focus on discourse emerges `from within the specialist knowledge of the subject matter and the field of knowledge for which something was written and the audience for whom it was written´ (Winter, 1982: 179). With this aim, as Swales points out, `the great majority of ESP programmes are designed to help their customers survive and succeed in an academic environment´ (Swales, 1988: xvi).[59]

---

[57] Hutchinson and Waters (1987: 73) give great importance to this second facet of needs analysis in ESP, not just the target situation but the learner's present needs or learning environment, around which methodology should be devised. Dudley-Evans and St. Johns (1998: 57) also make referecence to this feature of ESP as opposed to General English teaching, and they stress that it is even more important in Business English, where the identification of learners' backgrounds and conditions plays a crucial factor.

[59] It is no surprise, then, that there currently exists a common ground among ESP faculty and researchers found in the notion of EAP (English for Academic Purposes). In a general sense, it refers to the tasks in the academic setting and how these should be carried out on the line of interdisciplinarity: what Baunmann refers to as `an exact exchange of knowledge...according to specific features of specialised subject fields´(Baunmann, 1994: 1), or what Brumfit describes as `an interdisciplinary study in relation to existing practices and institutions´(Brumfit, 1984: 17).
Ever since the British Council coined the term EAP in 1975 to refer to study skills mainly reading related in universities and technical colleges (Hayes, 1998: 149), increasingly more conferences, proceedings and monographs get published dealing with the EAP issue in connection with ESP. See, for instance, Ramón L. Fernández (1990) or Malcolm Hayes (1998), just to name a few.
In addition, Tim Johns's or Peter Howarth's projects with overseas students in Britain are clear proof of the settled perspective of EAP. Johns's current EAP web page which presents strategies that cope with overseas science and technology students' linguistic problems on confronting academic discourse-- http://sun1.bham.ac.uk/johnstf/timeap3.htm --, is innovative and effective while making productive use of computer technology in teaching English.

From these viewpoints, it is thus the specific group's academic environment where the specialized language must be encountered and analysed.

IV.4. ESP AS THE HISTORY OF EST.

The history of ESP is that of EST (English for Science and Technology) from the viewpoint of this study.[60] The main interest is thus the scope of specialized language as the Science and Technology register.

Reference works that should be pointed out in order to give a brief historical account, can be traced back as far as 1962 when Barber writes "Some Measurable Characteristics of Modern Scientific Prose", which displays a first type of frequency analysis, well received under the name of `lexicostatistics´ in Europe (unlike in Britain and USA) (Swales, 1988: 1).[61] Another significant EST study is that by Herbert (1965), which focuses on `semi-scientific or semi-technical words´. In his study of lexis, another author, Higgins also refers to the important status of the `procedural vocabulary´ in Science (Higgins, 1967). In 1972, scholars such as Ewer and Hughes-Davies notice the relevance of `English in the professional training of scientists and technologists´ in South America, and the special position of `instructional English´ as a main `sub-register´ in Science and Technology manuals, handbooks and engineering textbooks.

During this same year, 1972, the `Washington School´ linguists, Lackstrom, Selinker and Trimble, present their research on `rhetorical considerations beyond the bounds of a

---

[60] Although there tends to exist a greater stress placed on the analysis of Business English and EAP nowadays (see Dudley-Evans and St. Johns, 1998, or Jordan, 1997), we focus on surveying the 60s and 70s scholarship that concentrated on the description of EST. The main difference lies in our received legacy of not only register and discourse features analysis, but also Genre and Corpus Linguistics techniques which enrich our perspective.

[61] Swales (1988) collects and adds comments on the sources cited that began the study of EST. The scholar himself wrote a manual in 1971 in which he reflected his realization that a particularly defined EST `trainer´/

single sentence´ in EST. It is a focus that two other authors, Allen and Widdowson (1974), consider as `particularly suitable for the preparation of EST teaching materials´.[62] In fact, as Swales states (1988: 72), `Widdowson and Selinker were the principal architects of the framework of a rhetorical-communicative approach towards EST language learning´, whereas Bates and Dudley-Evans relied on their concept-based analysis for writing the EST Nucleus material (Swales, ibid).

As Swales himself recalls (1990: 131), other works dealing with EST appearing during the remaining 1970s and into the 80s are: Inman (1978) on EST engineering lexis found in research articles, Lackstrom (1978) and Ewer (1979) working with modals in Science, and Trimble (1982) dealing with tense in EST article introductions. Later on, this same author would go on to complete the work which best seems to characterize the rhetorical functional focus of EST, a discourse approach (Trimble, 1985).

In a different article, Ewer (1981) focused on the need to design better EST textbooks that expose `what this language consists of in the first place´ (Khaled, 1984: 118). This is basically the general claim made by scholars in the 80s, who argue that EST work must depend on register analysis. As Robinson states, `very few books are based on any kind of register analysis´ (Robinson, 1980: 23).

The connection between register and sub-language thus mainly emerges with the notion of EST, which is consequently observed to demand a detailed linguistic study. Sager et al. (1980) describe this phenomenon and refer to sub-languages as

> complex semiotic systems based on and derived from general language; their use presupposes special education and is restricted to communication among specialists in the same or closely related fields (Sager et al., 1980: 69).

---

teacher was needed `to cater for the special English needs of special groups´(Swales, 1971: 9), since `the older and more senior English staff have the most difficulty in adapting to an EST situation´(ibid., 11).

[62] The focus on notions and rhetorical functions in EST also yields papers such as Greenhall (1981: 25), which criticizes precedent register views such as Strevens (1971) and `the more unqualified reliance on authenticity´.

Likewise, Lehrberger (1982: 102) refers to `sub-languages´ as having `high frequency of certain constructions´. These and other authors realize how important empirical examinations of language are for defining the EST sub-registers.[63]

A similar line of research dealt with the special purpose design of EST, which, according to Swales, should be distinguished from other types of special Englishes (Swales, 1973). Munby (1978) explictly refers to the specific purpose in EST: `to gain access to the required knowledge that is available´(Munby, 1978: 3). He thus separates the attention paid to EST in the educational plane from that in the occupational level (Munby, 1978: 56-57). For designing EST syllabi, in fact, Munby proves to be a main reference even today.[64]

Currently, EST courses are designed in different ways, and provide, in Orr's words (1995: 3) `research and instructional activity designed to understand and support the effective use of English in scientific and technological fields´. According to this same author,

> (…) this should also include Information on the profession's genres and sub-genres, audiences, purposes, print and electronic formats, high-frequency grammatical forms, high-frequency vocabulary, mechanics, efficient means of writing and revision, dissemination factors, cultural/professional taboo, and so forth (Orr, 1996: 2).

The description of rhetorical features and lexical-grammatical items thus proves to be a main goal in EST, above all, in the academic plane: EAP comprising EST.[65]

---

[63] Zwicky and Zwicky (1982) as well as Varantola (1984) display a special concern about the notion of register in restricted semantic domains. This approach prevails in Eastern and Western Europe and extends to America in the 80s--recently accentuated by works such as Biber et al. (1998).

[64] Obviously, crucial research such as Selinker, Lackstrom and Trimble (1972) cannot be overlooked as it forms the basis from which to draw methodological claims.

[65] Many models in EST syllabus planning are patterned after L. Trimble's detailed description of rhetorical functions and techniques (Trimble, 1985); this work has provided ample ground for the design of EST programs worldwide. In addition, the EAP umbrella has gradually covered the field of science and technology studies at tertiary levels, which means that more emphasis is placed on exploring genres and academic skills (see Johns, 1991; Jordan, 1997; Dudley-Evans and St. Johns, 1998; Seymour, 1999; Van Naerssen and Brennan, 1999).

IV.5. THE CASE OF ENGLISH FOR INFORMATION SCIENCE AND TECHNOLOGY.

What has been previously described with the framework of ESP and EST directly applies to the study of the specialized English under analysis, English for Information Science and Technology (henceforth abbreviated as EIST—see index 2). It is regarded as a sub-register of EST, encompassed by academic and technical English at university level: EAP (see index 2), recorded in fields such as Computer Science, Information Science,[66] Telecommunications (a major in Optical and radio communications) and Audio-visual Communication.

The four fields are chosen as representing the core area of EIST since they share common subjects in some cases or in others, common bibliographical background (e.g. Optical / radio communications and Computer Science). In addition, our study is restricted to English use in this area due to the vital position it occupies as an important communication tool where most of the documentation and reading material is transmitted in the anglosaxon language. A further factor which influenced the specific selection of the four disciplines mentioned is that they can easily lead to interdisciplinary analyses, as they share a great deal of common content in the design of the curricula. Last but not least, these domains are relatively new at the University of Extremadura and thus form the primary nucleus of Information Science and Technology studies that will likely have even greater relevance and an increase in pupils in the newly inaugurated century.[67]

---

[65] In Spain, unlike in USA or Britain, the distinction is made between Information Science --which includes university studies such as Journalism and Audio-visual Communication-- and Documentation --closely related to Library Science or Library science. In this thesis, the term `Information Science´ denotes the anglosaxon sense: Information Science as including both Library science and Documentation studies (i.e. those main studies at the Faculty of Library science and Documentation of the University of Extremadura).

[67] Computer Science courses were first offered in 1986 but did not become a major until 1995 (Computer Engineering). Information Science also began in 1995 (as Library Science; the first year of registration for the major, Documentation, occurred in 1997). Technical Engineering in Telecommunication, Specialty in Optical / radio communications began at the University of Extremadura in 1998. Finally, Audio-visual Communication began as a major at the Faculty of Library science and Documentation in 1999.

EIST is considered, as has already been mentioned, as a sub-register or special language from the viewpoint analyzed. Its definition fits that given by Sager et al. regarding special languages:

> (…) a number of sub-divisions of natural languages used for communication between specialists in different subject fields which overlap to varying degrees with each other and with general language (Sager et al., 1980: 182).

According to this view, as many revised authors state, EST is distinguished as a special language (Herbert, 1965; Lackstrom et at., 1972, etc). Yet, within the vast area of EST, several authors have also referred to what has been termed `sub-registers´or `sub-systems´. Sager et al. explain what is meant by this concept:

> They are in essence extensions of general language achieved primarily by superimposing on it an additional lexis which then becomes the most obvious distinguishing characteristic of special languages (Sager el al., 1980: 230).

In addition to lexis, the importance of rhetorical features is a crucial trait in the description of a sub-system such as EIST.[68]

Further evidence of the existence of sub-registers in EST is given by various works that set out to explore linguistic variation provided by lexical choice in subject specialisms.[69] One such researcher is Salager (1983), who takes text samples from medical fields (Pediatrics, Cardiology, Surgery, etc.) and characterizes Medical English according to lexical occurrences. Another scholar, Inman (1978), focuses on different fields of Engineering (Chemistry and Chemical, Civil, Electrical, etc.) in order to gather empirical data (lexical and rhetorical features) for developing a pedagogic rationale in EST classes for engineers. At the Chinese Petroleum University, Zhu (1989) designed an English course based on collected

---

[68] That special languages consists of more than simply general languages supplemented by special lexis can be seen and demonstrated by previously referred works such as Lackstrom et al (1972), Allen and Widdowson (1974), Ewer (1979), etc.

[69] What has been cited in dealing with specific Englishes is by no means an exhaustive list. As has been mentioned, works are only included if they are relevant or provide orientation to our study.

data from texts studied in sub-fields such as Petroleum Geology, Refinery, Drilling, etc.

Other works dealing with EST vocabulary in specific subject specialisms include Alekseev et al. (1973) in the field of Electronics; Hoffmann in Medicine (1970), Physics (1970), Chemistry (1973), Mathematics (1976), and Animal Husbandry and Veterinary medicine (1978). Varantola (1984) deals with Engineering English,[70] and Yang (1986) whose extensive work with Scientific English encompasses ten different domains of Science and Technology;[71] Baker (1988) also deals with medical lexis, and Flowerdew (1991) works with lectures and readings in Biology.

Additional, more recent, EST lexical work that is based on corpus material includes James (1994) who researches Computer Science English vocabulary; also in this field, Aguado de Cea deals with computer terminology as extracted from popular and academic texts (1996). Farrell (1990) focuses on Electronics lexis, providing lexical distinctions in the sub-domain of Electronics, whereas Lozano Palacios supplies extensive work on Information Science terminology (1999). The latter can also be accessed via Internet (see bibliography references and addresses). Other works that may be operated electronically at different locations also include data-driven material (e.g. corpora, dictionaries, thesauri, etc.) for specific academic needs.[72]

---

[70] In this case, Varantola is more concerned with noun phrase structures in Engineering English, but since this is highly related to specialized lexical association, his work is deemed an important reference.

[71] These technical fields are Aircraft Manufacturing, Atomic Energy, Chemical Engineering, Civil Engineering, Computer Science, Electrical Engineering, Mechanical Engineering, Metallurgy, Naval Architecture and Physics.

[72] There are several web pages devoted to materials such as on-line dictionaries and glossaries of Computing and Internet terms, such as http://www.uwasa.fi/comm/termino/collect and http://pcwebopedia.com but their main drawback is that they almost exclusively focus on technical terminology for concept explanation, not linguistic analysis. In addition, they do not cover wide areas such as Telecommunication engineering or Audio-visual communication (Viel, 1999). On-line corpora that include one or more disciplines are useful for linguistic study. Among them are the renowned Brown Corpus, the British National Corpus and the CoBuild project . Detailed information on current specific corpora is supplied in the next section, dealing with lexical analysis. Also, see bibliographic references for URLs of web pages.

Finally, Robinson's valuable contributions to ESP on specialized language and vocabulary provide several references to research in this area. For instance, her note on Phillips et al.´s study of Agriculture textbooks (Robinson, 1980: 71) or her brief overview of the study of vocabulary in ESP, with many examples of research into the vocabulary of scientific writing (Robinson, 1991: 27-32).

IV.6. ESP AS A COMMON CORE SUBJECT-BASED APPROACH.

With this last statement, the notion aimed at is that proposed by Hutchinson and Waters (1987: 163) when they suggest that the ESP teacher should take an active interest in the subject as an interested student of the subject matter. Given this positive attitude on the teacher's part, the two authors ask the pertinent question: what about the not unusual situation where the ESP teacher is faced with a class of learners with different subject specialisms? (ibid, 165). In this case, the answer may be that suggested by Kennedy and Bolitho:

> Texts of a semi-technical nature may be chosen, perhaps drawing on topics (such as pollution, population, conservation, etc)...which would provide practice in a set of skills, structures, functions and semi-technical vocabulary which the students will meet in their specialist studies (Kennedy and Bolitho, 1984: 50).

What these authors have proposed is the elaboration of common core ESP courses. They recommend not only common syntactic features of subject specialisms but also common lexical distinctions in differing disciplines. In the case of our project, the study of EIST in the four fields of Computer Science, Information Science, Telecommunications and Audio-visual communication, the principles underlined by common core theory are fundamental. Likewise, to a greater or lesser degree, approaches to technical vocabulary that appears as characteristic of single subjects or disciplines may also be regarded as relevant for common core lexis. In our view, we agree with Hudson (1972: 77) that `the structure of the lexicon (...) will vary a

lot from subject to subject´. In this respect, recent scholarship cited in our research as key (e.g. Pueyo and Val, 1996; Liu and Nesi, 1999) deal with the importance of this subject-bound lexis. As shall be seen, as a result, academic and subject words are distinguished in our common core analysis. Much of the scholarship developed by all these authors, in fact, exerts a substantial influence on the aim and development of our dissertation.

V. LEXICAL FOCUS

Based on Kennedy and Bolitho, we subscribe to the more focused analysis on the `semi-technical nature´ (Kennedy and Bolitho, 1984: 50) of EIST. Nonetheless, regarding the category of the more technical or restricted lexis –e.g. more specifically subject-based--, we are open to include, among the common core words in such a list, those specialized items in context, as shall be examined in the following chapters. As has been stated in the previous section, the aim is to build a common core English in the four disciplines of Information Science and Technology, varying in degrees of centrality according to style and subject matter.[73]

The main focus is thus on semi-technical or sub-technical vocabulary, either adapting more general or vernacular senses across a wide range of texts, or, in contrast, being more restrictive within subjects and topics. To these notions, many different writers in ESP have referred. As McNaught claims (1993), indeed, sub-language and the lexical focus are highly related:

1. Sub-language is strongly based in terms of lexis;
2. Sub-language texts focus on content;
3. Lexical selection is syntactified in sub-languages;

[73] We shall examine the nature of these topics and the criteria used for their selection in chapter 2.

4. Surface collocation plays a major role in sub-languages;

5. Sub-languages demonstrate elaborate lexical cohesion.

The following is thus a diachronic view of significant works that help to define the nature of this lexis and how it relates to common core language.[74] Three notions or categories are supplied for arranging the information provided by scholars.

V.1. A DEFINITION OF SUB-TECHNICAL VOCABULARY AS COMMON CORE.

J.R. Firth (1957) is often regarded as a founding father of modern lexical research in Linguistics, and although he was especially concerned with general English words in collocations and associations, he also noticed `the pivotal or key words of (…) restricted language´ (Firth, 1957: 12) frequently appearing in texts. C.L. Barber (1962) may be one of the first to explicitly probe the notion of a general vocabulary of Science, defining it as `high frequency words not contained in West's *General Service List of English Words* (1953)´(Sager et al., 1980: 242). Barber's list of words include **equation, method, approach, accurate**, and the like.

Johansson (1975: 15-20) provides a different list of words contrasting with the Brown Corpus Category J (dealing with Science and Technology) and registers some common core words: **information, process, analysis, surface, description, section**.

M. Inman (1978: 247) notices that most words in scientific and technical contexts are sub-technical: between 70% and 80%; some examples in this author's work include **form, function, temperature, structure, process, result**. Kennedy and Bolitho refer to this lexis as `words which are not specific to a subject specialty but which occur regularly´(Kennedy and Bolitho, 1984: 57). J. Ewer, in fact, sees this `number of language items which are common

[74] As before, the references are not meant to be comprehensive, but only relevant to the analysis provided in this research.

to the subjects´ as the `core language´(Ewer, 1983: 10). Louis Trimble gives a standard definition as he identifies this special vocabulary as `context-independent words which occur with high frequency across disciplines´ (Trimble, 1985:unit 9). He adds that these terms receive `an extended meaning in technical contexts´, which is usually not the standard meaning in general English (ibid.).

Yang (1986: 98) refers to `notions general to all or most of the subject areas´and calls them `sub-technical´words. Farrell (1990: 11) also defines them very explicitly:

> Formal, context-independent words with a high frequency and/or wide range of occurrence across scientific disciplines, not usually found in basic general English courses; words with high frequency across scientific disciplines.

According to more recent works, such as A.P. Cowie's, one may refer to this vocabulary as `common core´ because it shows `semantic variability... depending on specific textual clues´(Cowie, 1997: 130). It contrasts sharply with technical terms, which lack `exact synonyms and multiple meaning... as a category resistant to change´(ibid: 129). Carter (1997: 174), differentiates concepts by claiming that sub-technical terms of restricted languages refer to the `subject core´, whereas technical terminology involves `non-core´.

Della Summers (1997: 115) describes these words as `expected commonly but with many shifts of meaning´.

Furthermore, although Varantola (1984) focuses on noun phrases (NPs) in English for Engineering, he also takes the common core nature of lexis in `multi-word terms´ as `part of everyday engineering vocabulary´(Varantola, 1984: 30). He deems `the headwords of NPs´ as `keywords (general / semi-technical)´ which often contain latinate roots such as in **application, design, configurations**, (ibid., 97). Sager et al. (1980) recognize these NPs as `extended terms´, which include lexical units in opposition to other more technical units (**compression moulding, injection moulding**) (Sager et al., 1980: 233).

V.2. CLASSIFICATION OF SUB-TECHNICAL TERMS AS PROCEDURAL AND ACADEMIC.

Higgins (1967: 32) actually points out the existence of some semi-technical vocabulary as `frame words...which cause difficulty...which although not technical terms, are frequently used in technical writings´.[75] Richards (1974) identifies these as `words which help define other words´ (74), such as **mean, imply, consist of**, etc. E.D. Winter (1977) overtly specifies that this type of terminology --`vocabulary 3´-- is a class of open system lexical items which facilitate a reader's interpretation of a discourse. Such items

> have similar semantic properties to closed system items in sentence connection (...) [and] constitute a special vocabulary of context for the clause relations of English; they are words which can function as special signposts of what a word means in sequence with its adjoining sentences (Francis, 1986: 64).[76]

This type of words is also named by P.J. Roe as `synoptic vocabulary 3´ which is non-technical (Roe, 1977: 68). It includes words like **contrast, add, provide**, etc.

In relation to these terms, Hutchinson and Waters (1980: 3) call upon the notion of `procedural vocabulary´, basic for identifying `coherence and cohesion in texts´. Its function is `structuring the text´(Hutchinson and Waters, 1981: 65). These are words like **type, made up, use, keep**, etc.

Widdowson (1983) develops the concept further and defines procedural vocabulary in EST as `not schematically bound´ and with

> high aggregate frequency across a wide range of texts...especially useful for... defining terms which relate to particular frames of reference... such words function as a procedural vocabulary for establishing the terms which characterise different schemata (Widdowson, 1983: 92).

---

[75] However, Higgins' list of words, like Barber's above, seems to be based on intuition: **increase, record, contain, fill**. Only **obtain** and **accurate** occur in both EST lists.

[76] Thus, Winter refers to nouns such as **contrast** in Vocabulary 3 which equals the subordinator **whereas** (as in vocabulary 1) and the sentence connector **however** (in vocabulary 2).

Widdowson thus identifies common core words as procedural vocabulary that serves to define technical terms in EST (Widdowson, 1983: 93).

Hoey (1983) also establishes a link between this kind of semi-technical words and the structure of scientific discourse. He refers to `specialised lexical signalling´ that `exists to indicate to the reader / listener when he or she is encountering such a pattern´ (Hoey, 1983: 178), e.g. the EST pattern of problem-solution.[77]

Gil Francis (1986: 9) focuses on a class of nouns that appear as `sub-technical vocabulary which cuts across the field-dominated varieties of English´ and are `meta-discursive nouns´. This author calls them A-nouns because of their anaphoric reference. These are `highly informative nouns...contributing to provide the reader with a conceptual framework for understanding the writer's plan´ (ibid., 39).[78]

Procedural vocabulary is thus highly descriptive of the structure of texts; as Martin Phillips states (1989: 3), they supplement pragmatic work for discourse analysis purposes.

Farrell (1990) contributes by identifying this terminology with semi-technical vocabulary (71). McCarthy (1990) refers to it as `core vocabulary´ --in connection with `key procedural vocabulary´-- and defines its function as `to talk about other words and conceptualize items´, to `negotiate meaning... coherence and cohesion´(McCarthy, 1990: 51).[79] He claims that `the signal words are the abstract vocabulary of argument´(words like **issue, matter, affair, raise, outcome, come up with**, etc) (ibid., 123). He thus regards

[77] Swales (1990: 132) explains that the `problem-solution´ pattern is very common in scientific texts. The `IMRD´pattern of academic articles (Introduction, Method, Results and Discussion) is also quite elaborated, according to his work, each section containing its own linguistic characteristics, such as prevailing verb tense and voice.

[78] As Gil Francis states, when coping with these sub-technical words (e.g. **claim, statement, belief**, etc), `the reader draws on two kinds of knowledge: TEXT-PRESENTED KNOWLEDGE...and STORED KNOWLEDGE OF THE WORLD´(Francis, 1986: 39); the author bases his claim on R. de Beaugrande and W.V. Dressler's notions on text linguistics (1981) in this case.

[79] This author also explains that the `problem-solution pattern is very common in academic and technical texts´(McCarthy, 1990: 58).

`procedural vocabulary as fundamental to the strategies of definition and paraphrase´ (ibid.).

Marshall and Gilmour (1993) comment on how important this vocabulary is, since `it is used to modify or to express the relations that exist between the key concepts of the discipline´ (Marshall and Gilmour, 1993: 71).

Carter and McCarthy (1997 b: 208) also call upon the notion of `signposts´, those words that bring the `value of one sentence or clause in relation to another´ in discourse.

From all that has been previously stated, it seems, then, that this first category of sub-technical words is highly related to the meta-discursive plane. The second category, academic vocabulary, has more to do with the level of complexity:

I.S.P. Nation (1990: 65) notes the existence of `high-frequency words across texts´ and more complex academic terms to be learned. Phillips (1989: 32) also alludes to the `pedagogic concerns of EAP language teaching´ and its `restriction on the context of situation´. Burgmeier et al. qualify these `high-frequency words that are broadly applicable in university-level courses´(Burgmeier et al., 1991: viii) as `academic´ since they allow the learning of facts about different subjects.

Thurstun (1996: 3) refers to

> words which are common to all fields of academic learning, not attempting to include specialized or technical vocabulary items associated with specific disciplines.

Stotsky (1983: 438) refers to `words that contribute to cohesive ties in academic discourse... usually the content words generated by authors writing on similar topics´. These are related words `through their association with the topic of the text´(ibid.). They are thus also common core words that offer greater difficulty to non-native or overseas students because they are `often abstract and / or complex´(ibid.). Tim Johns (EAP web page) focuses on them by devoting more time and effort to dealing with lexical contrasts such as **enable** vs. **facilitate**,

compile vs. **compose**, and the like.[80]

Generally speaking, we may therefore assert that there exist two main distinctions in common core lexis from the perspectives examined: procedural items that convey information and express signposting in discourse arrangement, and the set of broadly defined academic words appearing significantly across a given body of academic texts.

V.3. SUB-TECHNICAL VOCABULARY AND THE ESP COURSE.

As a small sampling of worldwide study, conferences on LSP and EST such as our own in Spain of the Latin American countries,[81] have shown a steady interest in searching for sub-technical vocabulary for informative as well as pedagogical purposes in ESP classes.[82]

The issue has been approached for some time now. More than three decades ago, indeed, A.J. Herbert (1965: 18) postulated that

> much more difficult are the semi-scientific or semi-technical words, which have a whole range of meanings and are frequently used idiomatically(...) [and] look harmless but can cause a lot of trouble to the student.

Inman (1978: 246), on her part, concludes from her analysis of 100,000 words in EST that it is `sub-technical vocabulary which should be focused on in teaching EST´. Her study, in fact, shows that up to 80% of the words in the texts are sub-technical.

P. Skehan (1981: 109) proposes that `an effective vocabulary component for an ESP course´ is wordlists of general, technical and sub-technical vocabulary. Wallace (1982)

---

[80] See Johns' EAP web page (bibliography: hypertextual references).

[81] For instance, the Congreso Luso-hispano de Lenguas Aplicadas a la Ciencia y Tecnología (Badajoz, 1990-Portalegre, 2000) can already be counted as a well established one in Spain and Portugal. In addition, since 1984, a ESP / LSP Latin American Symposium is held every two years in different countries (e.g. Chile, Brazil, Mexico and Argentina). In 2000, the location will be Baranquilla, Colombia.

[82] Some, among other conference lecturers in Spanish LSP Congresses who show this interest in lexis for ESP analysis, are Belda Medina (1999), Henderson Osborne and Feito Higueruela (1999), García Núñez (1999), Lerchundi and Moreno (1999), Blanco Gómez and Henderson Osborne (1997), Palacios Pablos (1997), Romero Guillén (1997), Soláns García (1997), Curado Fuentes (1997, 1995), Alejos Juez (1995), Alejo González (1991), Carpintero and Vila (1991), Edwards Rokowski (1991), Vila and Sevilla (1990), etc.

tackles the problem very explicitly, stating that

> the serious problem ... is probably not technical language as such, but the language framework in which the technical expressions are placed... sub-technical words and expressions typical of academic discourse (that is, words such as **ratio, approximate, hence**) which the subject specialist may assume that the student should already know (Wallace, 1982: 17).

Carter (1987 a: 152) agrees that `concrete words are generally learned first... abstract words may be more difficult´, and recalls Richards' similar claim (1974: 75). A core vocabulary thus tends to be set up in general language learning, involving the arrangement of items according to concrete and abstract conceptions (Carter, 1987 b).

Swales acknowledges that ever since Barber (1962), EST practitioners have held the view `that teaching specialized technical terms falls neither within the responsibility nor the competence of the English teacher´(Swales, 1985: 17). The important teaching instrument, according to Swales, lies in the `vocabulary which is generally useful to students of Science and Technology´(ibid., 18), not the technical or highly subject-restricted terms. These are usually encountered and defined by students in their major courses, even though as McDonough (1984: 103) claims, among others, `specialist vocabulary is often a no-man's land, perceived by subject-teachers to be part of language teaching, and by language teachers to be outside their competence´.

However, other, more recent, scholars offer contrastive insight to this view. Liu and Nesi (1999) propose a step beyond in their work with sub-technical and technical vocabulary in the case of postgraduate students at the University of Warwick in Malaysia and Hong Kong. In their claim, the learning of terminology, not just sub-technical lexis, is favored. Nonetheless, as they also state, these terms `probably have greater relevance for postgraduates than for undergraduates, because most postgraduate students already have a certain amount of experience of reading academic texts in English´ (Liu and Nesi, 1999: 146). As a concluding

remark, then, we believe that technical words like **compiler**, **debugging**, etc can be within the ESP teacher's grasp as long as they co-occur frequently enough and with suitable context (i.e. wide contextual, informational, range).

We consider that due to their usually greater significance in terms of frequency and range, as a result, semi-technical items should be prioritized, whereas technical words should mainly be treated as secondary in our analysis, even though, at times, as shall be seen, they may occupy prevailing positions in the arrangement of the lexical information and profiles.

The scholarly documentation examined thus far bring us to the assumption that the most important vocabulary to be worked with and learned in EST is a common core lexis. Like Summers (1996), we believe in the highly demanded empirical analysis of this type of sub-technical vocabulary in authentic texts.[83] The `general core of the language´ should thus be revealed by a `true´corpus of texts: `representative, to a reliable degree, of a broad range of document types´ (Summers, 1996: 262). This takes us to an important notion and means of study in specialized language lexical analysis: the use of corpora.

VI. CORPORA

As Summers (1996) states, in fact, a `true corpus´ will present the core features of a language. The definition of what is generally meant by corpus in ESP and an overview of its particular varieties should thus be included in this section dealing with the specific approach to scientific-technical lexis.

Because of the greater availability of electronic texts (e.g. via Internet), studying specific purpose corpora is now made easier and more effective. Software is available to users via internet to process texts and access data or databases (cf. Barlow, 1996). As Vincent B. Y.

[83] As Pauline Robinson reminds us, this authentic data is the starting point for both needs analysis and materials preparation (Robinson, 1980: 35-37).

Ooi points out, `greater sharing and portability of lexical and textual Information is increasingly enhanced as we move into the 21st century´(Ooi, 1998: 119). This, combined with the fact that computers are becoming more sophisticated and powerful, leads corpus linguists to believe in `reconsidering the question of how automatic discovery procedures and intuition can be combined´(Stubbs, 1995: 48). From this viewpoint, several scholars identify the corpus approach as the counter-research to Chomsky's top-down or deductive approach: `data-based or corpus-based methods are called bottom-up or inductive´(Krishnamurthy, 1997: 34), although this, due to the importance of context, does not imply that deduction techniques are discarded. It includes both lexicographic studies and grammar-based work.[84]

VI.1. THE RELEVANCE OF REPRESENTATIVE CORPORA IN ESP.

What is actually a corpus is a question that must be further explored in order to gain greater insight into its relevance to ESP studies.

According to Francis (1992: 167), a corpus is `a closed set of texts established according to specific structural criteria´. For Francis, it is the result of an empirical process of linguistic data collection combined with the author's own capacity of elicitation and introspection.

The more recent tendency to define corpora, however, is exclusively one which relies on the empirical part. The type of corpus-based study thus

> focuses on the E-language (in the sense that it looks at observable data as the primary type of data) and so any lexical study which derives Information from a corpus might have a theoretical status `hard to justify´, in Chomskyan terms. Nevertheless, the last few years have witnessed a resurgence of interest in the so-called E-language (Ooi, 1998: 4).

[84] In the case of areas where the terminology is changing fast, such as in Business or Telecommunications, the corpus-based lexical approach is particularly useful and is applied worldwide in the making of specialized language dictionaries (see, for instance, Belmonte [1997]). It is also important to look for `universals of language´ in `an age in which science has a high status´(Sager et al, 1980: xv).

In this sense, the data contained in a corpus must be authentic so that reliance on externalized utterances reflects natural language. E-language or empirical language is quite valuable and pertinent to sub-language study.

Yet, from our viewpoint, I-language or intuition / introspection language analysis also plays a considerable role in the form of the researcher's own experience and knowledge: `a rich resource of authentic data containing structures, patterns and predictable features that are waiting to be `unlocked´ by the human intelligence´ (Wichmann et al., 1997: 3). We thus agree with Leech's assertion (1991: 74) that what is important in linguistic studies is the combination of `corpus plus intuition´. Thus, E-language should appear not in opposition to but in combination with I-language.[85]

The method, so closely bound to linguistic authenticity, can be highly fruitful in its application to ESP and should be given more attention in this area, since, as Wichmann et al. state (1997), `the history of LSP has not favoured the corpora approach (...) to find out about characteristics of language varieties´(ibid., 18).

For this reason, Biber et al. (1998) also support the prevailing position that the study of significant linguistic association patterns in a corpus holds for LSP:

> Association patterns represent quantitative relations, measuring the extent to which features and variants are associated with contextual factors (Biber et al., 1998: 4).

Thus a corpus is considered `a large and principled collection of natural texts´ as the basis for the analysis (ibid.).[86]

---

[85] In the context of Computational Linguistics, which deals more with computing and programming notions for language analysis – in addition to `research on automatic translation, electronic production of artificial speech (SPEECH SYNTHESIS) and the automatic recognition of human speech´(Richards et al., 1992: 72)--, the notion of E-language is highly important in order to simulate reality as closely as possible. In Computational Linguistics, the management of computerized language is crucial: texts are `constructed in such a form (...)as to form a `fair cross-section´of authentic usage for the language as a whole or for some particular genre´(Sampson, 1991: 181).

[86] In this sense, the utility of a corpus is still viewed today as what Sinclair and others devised more than three

As a specialized language corpus, it must also constitute `a strict selection of texts´ (Sinclair, 1991: 40) since, as this same author claims, `many words and phrases are rare in a general sample of texts but very frequent indeed in certain specific texts´ (ibid., 45). This is an obvious case in which a corpus must be representative, as Renouf tells us: `a body of text which is intended to be representative of one or more aspects of the language´ (Renouf, 1997: 258).[87]

More recent work with representative corpora differs from earlier corpus analysis (e.g. word frequency counts in the works of `lexico-statitians´--Barber, Herbert—in general) in that it `can provide many additional kinds of information about language use´(Biber el al., 1998: 5). The notion of representativeness is thus the foremost central issue to address before building a corpus:

> (…) a corpus seeks to represent a language or some part of a language. The appropriate design for a corpus therefore depends upon what it is meant to represent. The representativeness of the corpus, in turn, determines the kinds of research questions that can be addressed and the generalizability of the results of the research (Biber et al., 1998: 246).

In an earlier work, Biber (1993: 243) defines `representativeness´ more explicitly: `the extent to which a sample [text] includes the full range of variability in a population´; while Summers (1993: 186) asserts that `unless a corpus is representative, it is ipso facto unreliable as a means of acquiring lexical knowledge´. Representativeness is thus defined in relation to sub-language (register / genre); in Sebba's words (1991), this means that

> the more narrowly defined the domain of a corpus and the more precisely(...) the more representative it will be and the more reliable will be the statements based on it (Sebba, 1991: 23).

---

decades ago; a very useful tool for representing meaningful language: `Any stretch of language has meaning only as a sample of an enormously large body of text´(Sinclair, 1965: 76).

[87] In contrast with text linguistics, which looks within the text, corpus linguistics, according to Renouf's definition, `cuts across texts´ by analyzing language use common to some, doing an exhaustive study of chosen data and looking at actual language use `otherwise inaccessible to the conscious mind´(Renouf, 1997: 258).

The selection of specific texts in certain linguistic domains and / or sub-domains is a crucial step at the time of making the corpus representative; according to Demarty-Warée (1998: 100), `le professeur doit donc rassembler un certain nombre de textes issus d'un domaine spécifique ... afin de constituer un corpus´. The choice of texts depends on such things as the type of knowledge discourse sought, degree of specialization, target audience/users, etc. (ibid.). A key yardstick for this type of textual selection may be formed by `curriculum surveys and guides, recommended reading lists, and various other sources of information´ in academic environments (see Breland et al., 1994: 1, in the dictionary and corpus references of the bibliography).

Since corpus derived data `does not illustrate characteristics of the language in general, but rather of the particular language of the corpus examined´(Gavioli, 1997: 84), it may be concluded that this approach in LSP should be accurate and relevant to reflect the behavior of real language in a particular area such as in Information Science and Technology and at a particular time such as the current period, the threshold of the 21st century.

VI.2. USE OF CORPORA ANALYSIS IN LSP FOR LEXICAL STUDIES.

What is important to view more closely next is whether the use of corpora is a suitable means of describing lexis in a given area of language. Some authors claim that it is, and give several examples from their studies carried out with specialized languages.

Varantola (1984: 49) focuses on noun phrases in Engineering English (EE) and determines that `a corpus of journals ...contains a large number of complex NP constructions´ due to the fact that this genre of reading in EE concentrates on `giving information about new products, processes and developments´(ibid.). He then goes on to compare this extracted vocabulary with a reference corpus, `a `norm´that is used as a yardstick´(ibid., 51), in this

case, the Brown corpus.[88]

It is, in fact, about the time of Varantola's publication (1984) when corpus linguistics begins to be better defined as a distinct conception of linguistic work, e.g. Aarts & Meijs (1984).[89] Nowadays, the popularity of contemporary corpus compilation has been enhanced by the increased availability of Computer Technology for the storage and processing of the lexical data. In this area, corpus may be very similar to the concept of database, which refers more to the method of storing the lexical information (see Svartvik, 1992).[90]

---

[88] The Brown corpus was the first machine-readable corpus of text extracts (Kucera & Francis, 1967) to be designed; it encompassed different registers of American English. Johansson's work (1978) was its British parallel, the London-Oslo / Bergen (LOB) corpus (Longman edition). The London-Lund Corpus was the first to include spoken data (scripts) (Svartvik and Sager, 1975).
The three belong to an earlier stage or first generation corpora (as Svartvik calls them --1992: 18); a more sophisticated level can be represented by larger size corpora such as the International Corpus of English (10 million words-- Greenbaum 1991), the British National Corpus (100 million words-- Quirk 1992-- Oxford edition) and the Birmingham Corpus (20 million words-- Sinclair 1987—Collins-CoBuild). These sources contain large sub-corpora, such as the ICE-GB in the first mentioned, containing one million words of spoken and written adult British English texts (Survey of English Usage, 1998).
It may be said that there is nowadays a further stage or generation --since the early 90's-- as the mentioned corpora and others (e.g. Closing Survey-- 1995-- Cambridge editors, the CIC –Cambridge International Corpus [1992]) have grown beyond recognition (for instance, the Birmingham corpus or Bank of English currently holds more than 327 million words, or the LOB corpus has undergone a much more sophisticated processing stage called the SUSANNE project [explicitly addressing four different genres-- 1992, 1997]). One other famous massive corpus, such as the Longman Corpus, utilized by Biber et al. (1999) to produce a solely corpus-based grammar of spoken and written English, is thus employed under the notion of register variation analysis. CRATER (McEnery, 1994) or Multext (Ide and Véronir, 1994) deal with bilingual and multilingual data. In addition, there is the more recent tendency to process spoken data via voice recognition systems that enable works such as the Santa Barbara Corpus of Spoken American English to be released (2000) (http://humanitas.ucsb.edu/depts/linguistics/research/csae), or the MICASE Corpus of Academic Spoken English at Michigan University, where in May 2000, 200,000 words from oral academic genres (lectures through service encounters) have been tagged (Simpson, 2000).

[89] In the mid-80s, other scholars such as Quirk et al. (1985) also develop the concept of Corpus Linguistics quite extensively, but they mainly focus on preparing corpora for grammar analysis. There grew so much work in Corpus Linguistics after 1984 that by 1992, Leech defines it not just as `a newly emerging methodology for studying language, but... a new philosophical approach to the subject´(Leech, 1992: 106). Nowadays, many agree with Hoey's claim (1997) that Corpus Linguistics is not just a branch, but rather a route into Linguistics. A statement that may be parelleled by Marcos Marín' words (1994) regarding large corpora for language management and standardization: `(…) podemos crear instrumentos gigantescos de almacenamiento lingüístico (…) que nos permiten hoy lograr (…) la regularización de la terminología´(Marcos, Marín, 1994: 57).

[90] A similar term that may be confused with the notion of corpus is `archive´. However, unlike a corpus, an archive is not an established set of representative texts according to specific criteria, but rather an unstructured, unselected store of texts `collected more or less opportunistically, according to what sources of data can be made available and what chances for collection arise´(Leech, 1991: 10). Atkins et al. (1992: 1) also define archive as `a repository of readable electronic texts not linked in any coordinated way´. By way of summarizing, as Ooi (1998: 48) states, `although both [corpus and archive] are instances of observed data, only the corpus is systematically gathered for a particular purpose´; also, as Mason puts it (1998), the corpus is `collected

Before the era of Computers, various attempts were made at working with vocabulary frequencies and glossaries. Kennedy (1992) cites several authors whose pre-Technology projects aimed at vocabulary counts and identification for academic (teaching) or occupational (training) purposes.[91] The interest in compiling corpora for lexical study is thus not a new concern even in LSP. Yet, developments in corpus linguistics suggest new ways of observation and processing of data.

Since the publication of McEnery's and Wilson's *Corpus Linguistics* (1996), computational techniques and methodology have been well-established for the analysis of texts.[92] In the making of dictionaries, these have proven to be essential since the making of the Collins-COBUILD project –with about 330 million words nowadays (329 million on 20 July 1998), developed by the University of Birmingham in association with Collins editors since 1980; before this date, see Sinclair et al. (1969), Sinclair & Coulthard (1975), Roe (1977) for small-scale corpora projects at the University of Birmingham. Renouf (1984, 1988) and Sinclair (1987, 1988) explain the methodology behind the publication of the first edition

---

according to a certain set of criteria´, and in his contribution to the Corpora list (Mason, 28 Jan 2000), he adds that `a corpus tends to be a lump of real language´. Renown examples of archives, electronically accessible, are the OTA (Oxford Text Archive) and ICAME (International Computer Archive of Modern English). Also, see the LDC catalog for entire publications lists of corpora and texts for research (http://www.ldc.upenn.edu/Catalog).

[91] Such works include Kaeding (1897), who worked with an 11 million word corpus to compile German words for stenographers, Eldridge (1911) who collected newspaper articles for non-native speakers of English working in a factory, Horn (1926) and Horn and Ashbaugh (1928) whose more than five million word corpus was aimed at children for vocabulary learning. Thorndike and Lorge (1944) published the most extensive one, encompassing 20 million words that served to release a 30,000 word lemmatized list of most frequent English vocabulary that varied according to range. Finally, West (1953) compiled the most effective list of words for ESL students, his *General Service List of English Words*, based on a corpus of some five million words (for a fuller account of West's work, see McArthur [1978] and Howatt [1985]).
Kennedy (1992: 338) also explains that these scholars meant to outline `the principle that items with a likely high frequency of occurrence in texts should be learned first to avoid memory overload and confusion and to lighten the learning burden´.

[92] Some Ph. D. studies applying computational methods for analyzing language since 1996 are Berri (1996), Rundblad (1997) and Lindberg (1999). They all regard context and phraseology as key items for exploring how language truly operates.

of the CoBuild dictionary (1987).[93] Regarding growing scholarly concern in this field of research, ICAME (the International Computer Archive of Modern English) publishes a journal at least once a year since 1987,[94] and the TALC conferences (Conference on Teaching and Language Corpora) have been held since 1994 –with Wilson and McEnery (Lancaster University) as pioneering editors-- as further evidence of the increased interest in corpora for research and teaching purposes.[95]

Closely related to corpus-based tasks is the aim to set out a range of principles for acquiring texts according to genre --`a situationally defined text category´ (Ooi, 1998: 53)-- and to text type, `linguistically defined´ (ibid.).[96] This should provide `experimental evidence of word uses´ and `word associations´ (ibid., 217), allowing lexical variation `to be observed and described´(Sinclair, 1997: 37). In this sense, the process, as Aston suggests, `implies the use of corpora which are representative of the genre or genres to be taught, so that typical target uses can be identified´ (Aston, 1997 b: 55). In a forthcoming article –to be published in PALC 99 (Proceedings of Practical Applications in Language Corpora)-- Aston adds that

[93] Other works since then include the OALD (Oxford Advanced Learner's Dictionary -- Cowie [1989] and Crowther [1995]) and the LDOCE (Longman Dictionary of Contemporary English -- Summers [1987, 1995]). These `encode such semantic notions as register, semantic type of object and cross-references´ (Ooi, 1998: 33). A Ph. D. study presents how these lexical resources (particularly LDOCE) can be parsed automatically as machine readable texts (McHale, 1995), while Fontanelle's dissertation (1995) deals with how a dictionary can be converted into a lexical database to investigate the structure of the English lexicon by focusing on the relationship between collocations, lexical functions and metaphors / sense extensions.

[94] ICAME was founded in 1980 by Leech and others for `collecting and distributing information on language material available for computer processing´(Svartvik, 1996: 7), a task that Leech and Johansson had already been carrying out manually with advertisements since 1960. The organization archive mentioned in the name is currently at the Norwegian Computing Centre for the Humanities (NCCH) in Bergen.

[95] In our own country, corpora methods are highly developed, for example, at the IULA (Institut Universitari de Lingüística Aplicada) in Barcelona, carrying out effective computerized lexicographic work with LSP multilingual corpora (texts selected from Law, Economics, Medicine, Computer Science and Environment according to criteria of Register and Genre analysis).

[96] Glässer (1998: 130) gives an explicit definition of both genres and text types to distinguish their features: `Text types are understood as basic cognitive operations which are manifest in text segments and speech acts – for example, description, narration, exposition, argumentation,and instruction (cf. Werlich 1976)´(…) `these text types constitute traditional text forms, known as genres and subgenres´ (…)

aiming at genre and text type depiction when building a corpus should be the first priority in LSP regardless of the corpus size. His example of the construction of a small but teaching-effective corpus in the genre of personal ads with descriptive text types is quite illustrative.[97]

When we work with a given genre represented by a corpus, `sentences from the text become very relevant because a limited sample allows the development of the specialized vocabulary´ (Zernik, 1991: 12). The body of texts thus yields the appropriate lexical results according to our viewpoint. Corpora become suitable tools in establishing frameworks for the selection and distribution of lexis in specialized languages, as they are `full of multi-word units, collocations, idioms, repetitions, ellipses, abbreviated styles, "grammaticalities"´ (Calzolari, 1997: 79). These are made plain in a computerized corpus for us to observe and analyze. As Aston asserts, `corpora can play a useful role in the acquisition and restructuring of the schematic knowledge underlying communicative competence´ (Aston, 1997 b: 51).

As a summary, the advantages of using a corpus for lexical study are many as well as objective in comparison with an expert's knowledge or intuition. Let us see four of them as classified by Krishnamurthy (1997: 37):

A corpus can be more comprehensive and balanced.
A corpus can identify what is common and typical.
A corpus can supply fairly accurate statistics.
A corpus can provide countless real examples.

What should be examined hereafter is the means by which lexical research can be better carried out on a corpus. This leads to the notion of concordances for lexical analysis.

VII. CONCORDANCE

The linguistic Information obtained from a corpus through `concordancing and

[97] Defining the methodology is also crucial in combining language teaching and corpus use, according to Aston. As his chart shows (Aston, 1997 b: 61), the paradigmatic line of language (genre and register conventions of situation, topic and genre) must merge with the syntagmatic line (collocation, colligation and connotation of lexico-grammatical items).

frequency tables´ is highly important, `particularly when the corpus is specific´ (Gavioli, 1997: 88). According to current linguistic approaches, the `concordancing software package´ or `concordancer´ is thus exploited as a means of learning about the foreign language´ (ibid., 96). Its main linguistic targets are lexical factors of study such as frequency (word count in texts), range (word occurrence across texts), availability (word occurrence in certain contexts) and coverage (word use as replacement of other words --frequency of collocation). These are given by the `lexico-grammatical patterning of certain semi-technical words related to various functions used in academic writing´ (Dudley-Evans, 1998: 207). Their use varies as either "research" or "classroom" concordancing (Higgins, 1991: 92); as shall be seen, the former are more appropriate for our lexical study.

VII.1. CONCORDANCING AND SPECIALIZED LEXIS.

A concordance listing can be defined as

> a collection of the occurrences of word-form, each in its own textual environment. In its simplest form, it is an index. Each word-form is indexed, and a reference is given to the place of each occurrence in a text (Sinclair, 1991: 32).[98]

The user and linguist can benefit from this type of Information in terms of gaining insight into actual word use and meaning:

> in examining a concordance listing, however casually, the reader is actually constructing various frames of expectations regarding the behaviour of the word. But the reader can do more than just render an informal ocular scan: (...)a linguistic framework of expectations (...) for the systematic sifting of corpus lexical evidence (Ooi, 1998: 19).

It thus provides valuable information for lexical research; as Amsler (1982: 661) states,

> (…) Applying(...) Information Science techniques (cluster analysis, cooccurrence relationships, frequency counts) to gather and present the raw

[98] As Sinclair himself points out in the Preface to his 1987 Collins CoBuild dictionary, these `concordancing procedures´ provide `patterns of words... related to the meanings and uses of words´(xvii).

> text to human lexicographers or experts for their assimilation and restructuring into formal analyses of terminology.

The concordancer may thus be `the basic tool of lexicographers and linguists alike for word evidence´ (Ooi, 1998: 47). The list of words generated by such means, `a concordance listing´, `is necessary in order to show systematically the use of the lexical item´(ibid., 129).

> Given these words and their environments, it is often revealed, in a technical text,
>
> the fact that the realisation of their constituent lemmata are embedded in the text in an environment of non-technical words, some of which will appear frequently enough (Phillips, 1989: 54).

A lexical item such as, for example, `conceptual space´, according to Phillips's analysis, would appear often and thus be included as a core item for the given technical text(s). In the process of lexical systemizing, because of its relationship with `conceptual´, the sub-technical word `space´ receives a more specialized denotation. This may be undertaken through `concordancing to display the contexts of occurrence of particular forms´(Aston, 1997 a: 257).

VII.2. CONCORDANCING FEATURES FOR LEXICAL ANALYSIS.

The main features that concordancers display for lexical study are `Key-Word-In-Context (KWIC) and word-frequency profiles´(Ooi, 1998: 82).

The first item (KWIC) provides a `whole line of text where the word-form occurs´(Sinclair, 1991: 33), and this is regarded as very useful and practical: `any instance of language depends on its surrounding context´(ibid., 5).[99] Thus, in Business English, for instance, where highly specific corpora are designed to be contrasted by means of a KWIC analysis, the collocation `said account´ may be noticeable and significant in a given corpus of

[99] As Carnine et al. (1984: 190) explain, `context reveals the meaning of an unknown word to the extent that "a clear connection between that word and those which illuminate it [is provided]" [(Deighton, 1959, p. 2)]´.

Business correspondence, the PROCOMPARE corpus; however, the same occurrence appears very seldom in a different Business body of texts, the PROLEX corpus (see Ooi, 1998: 135).[100] The user must apply the KWIC feature of analysis to come to realize these lexical differences.

For the identification of these lexical patterns, the analysis is based on the notion of `span´, which is the linguistic context or `co-text´[101] --that is, the surrounding text-- of the node word-form. Long ago, Aborn et al. (1959: 179) asserted that this co-text should be restricted to `ten words... The length at which context attains maximum effectiveness lies between five and ten words´. Sinclair (1991), more recently, reduces this number to eight:[102]

> Sinclair and his associates pointed out that determination of optimum span size was a major objective... that the influence of a node does not extend appreciably beyond a span of four orthographic words on either side (Phillips, 1989: 27).

The importance of co-text is obvious to achieve meaning through the information provided by KWIC (Nation and Coady, 1997: 103). As Summers (1997: 117) reports, `putting individual words into a range of typical contexts and appropriate phrases´ is allowed by observation of spans (i.e., examples in dictionaries are given through this procedure). The linguistic data provided by KWIC is thus `edited and interpreted´ so that generalizations can be made (Gavioli, 1997: 84).

For the second feature mentioned above, word-frequency profiles, what are sought are `words that occur with a high frequency´ (Ooi, 1998: 144). These are significant in terms of not only absolute frequency but, to an even greater extent, co-occurrence frequency with

---

[100] This, as was already seen, has to do with degrees of representativeness (marked by register, discourse and genre features) in corpora..

[101] `Co-text´ should be preferred, according to Sinclair (1997: 34), `because context often has a wider meaning than the surrounding text´.

[102] Although this span number should be flexible, as Berry-Rogghe observes (1970): it depends on the type of grammatical structure and text we are dealing with (e.g. dramatic texts vs. conversation) (Phillips, 1989: 16).

certain words. In the lexical association sense,

> a standard statistical measure which can be used to indicate the significance of the frequency of co-occurrence of the collocates with the selected wordform in question (ibid., 143-144).

This scope, referred to as `quantitative analysis´ (Biber et al., 1998: 8), involves attention to both `frequency of occurrence´ and `co-occurrences frequency´ (ibid.). It is `highly important for processing effort´ (Nattinger and DeCarrico, 1997: 20).[103] Thus, what must be taken into account according to this type of research are both `word probabilities´ and `joint probabilities´ statistics.[104] The former refers to the frequency of the word in relation to the number of total corpus words, whereas the latter is the relationship between the frequency of a given lexical association and the frequency of each word separately (ibid., 22). There are then, as Sinclair proposes (1966), `predictive values´ according to this kind of statistics, about lexical co-occurrence --e.g., that the noun `omen´ is more likely to be preceded by `good´ than `good´ followed by the word `omen´ since `good´ has other various combinations (Phillips, 1989: 44).

---

[103] As Firth's memorable claim goes (1957), `You shall know a word by the company it keeps´(Nattinger and DeCarrico, 1997: 20), word association frequency is especially relevant to lexical analysis to find out about lexical associations (ibid.): as Candlin (1997: 25) puts it, `not only the frequency of individual lexical items but also, and crucially, `the company they keep´. In this view the priority traditionally given to the single word should give way to a much higher regard for multi-word units´(Candlin, 1997: 25).

[103] These statistics are thus measured by means of Mutual Information (M.I.) analysis (Church et al., 1990, 1991) and / or the Z-score (Barnbrook, 1996; Oakes, 1998). The former refers to the degree of expectation maintained by the fact that a given word will co-occur with another (Collocation statistics in a sample of texts); the latter deals with percentages of collocational strength based on a formula which takes absolute frequency of the collocate and length of text as key factors of mesurement (available with some concordance software such as *TACT* or with the CoBuild corpus). In these statistics, co-occurrences appearing less than three times are typically excluded (except for clusters –i.e. sets of words or phrasal items co-occurring in such a manner that their use can be predicted in a given corpus –cf. Sinclair, 1966: 417). In addition there exists the so-called T-score statistical measurement, which also predicts significance of low-frequency co-occurrences to be accounted for (e.g. the verbal phrase **coin** + ***the word***, or ***the term*** co-occurs with a low frequency in our scientific-technical corpus, but in its statistical measurement of proportion with the overall running words **coin** and **term** or **word** in the texts, the collocation is significantly high). For a discussion of the T-score in collocational analysis, the CoBuild page may help: http://titania.cobuild.collins.co.uk/javahelp/help.html. Also, Jem Clear gives a brief but useful account in the Corpora list (Sun, 12 Dec 1999).

This lexical frequency information is quite important for vocabulary studies.[105] As Richards claims (1974),

> the relationship of frequency to information is an important factor in evaluating the role of word frequency in vocabulary selection(...) also the embracing range of contexts where it is found (Richards, 1974: 72).

For both factors, range and coverage of word use, frequency approaches to lexis are therefore effective.[106] The application of these variables to lexical study, as Nattinger and DeCarrico (1997: 182) state, `is only now beginning to appear possible with the power and storage capacity of modern computers´.[107] `The World Wide Web´, as Ooi points out (1998:175), is also enabling this type of computerized work: it `is becoming indispensable for lexical research´(ibid.); moreover, an increase of Ph. D. studies on language learning serves as evidence as they focus on hypertextual documents as a source of linguistic feedback (cf. Wang, 1995).

Two other main reasons for the importance given to this electronic medium are highly practical from the pedagogic plane: Novelty of use and information scanning. In the first case it is true that not only Information technology related issues soon fall obsolete in this day and age, but that these texts entered via the WWW are increasingly being accessed by students. This implies that updatedness contributes to motivating their reading and that the process triggers a series of microskills rather distinctive of the processing of hypertext material –i.e. by means of highlighted links and subject keywords.[108]

---

[105] The observation of frequency also leads to register distinction, above all in terms of the statistical measurement for determining the ratio of `types´ (distinct words in a text) vs. `tokens´ (all words) in a given text –e.g. 15 / 30 in conversation, 30 / 30 in academic prose, and 21 / 30 in lectures (see Ure, 1971).

[106] As Stubbs claims (1995: 51), these techniques provide `a cause for confidence that lexical descriptions in the future will provide more accurate and exhaustive documentation about words´.

[107] As Leech states (1992: 106), the use of computers can be seen, in this sense, `as an essential means to a new kind of knowledge, and as an `open sesame´to a new way of thinking about language´.

[108] See chapters 3 and 4 for developments on these issues.

VII.3. CRITERIA FOR ESP LEXICAL SELECTION THROUGH CONCORDANCING.

Lexical variables, described as early as Mackey (1965), constitute the main factor to consider for selecting a lexical item as significant or common core across texts. Given the advanced situation or status of analysis that concordancing technology allows, however, these criteria may offer `new´ features. Frequency, the first, allows for `the distinction between the possible and the typical´ utterances (Willis, 1990:40): `how the word is typically used, rather than how it might possibly be used´(ibid.). The more frequently a given lexical co-occurrence is repeated in a sub-language --e.g., Business English-- `the more this data (...) is found to characterise Business English´ and `the more it can claim to be a sub-language corpus´(Ooi, 1998: 131).

Noticing the range of word use, a second yardstick, is also decisive in specialized lexis selection. Because, in fact, `word and word senses have quite different distributions across registers´(Biber et al., 1994: 174), concordanced lines must be used to establish patterns that are likely or bound to occur: `identifying the various senses of a given word by this use´(ibid., 176). This proves to be quite valuable in LSP for forming a `text-based model as empirical basis´ (Bergenholtz and Tarp, 1995: 23): `No doubt, a concordanced text(...) constitutes a ready-access and technically reliable tool for this job´(ibid.).

In terms of lexical range analysis –i.e. chances that words will co-occur together (cf. Sinclair, 1966: 426)--, concordances replace any type of linguistic intuition.[109] Word range constitutes a reference point that in ESP leads to specific lexical variability accounts; thus, the treatment should be empirically made (Farrell, 1990: 26-28).

In terms of availability and coverage, the other two criteria for lexical selection, the

[109] The linguist's intuition should only be `brought at a late stage in the process of compiling information´ (Sinclair, 1991: 39), or, as Quirk claims (1997: viii), at the beginning level of dealing with a `pre-established theory or speculative hypothesis´.

retrieval of lexis should be based on two main principles: academic availability, that is, making the texts used in courses `our starting point for the bank´ (Brekke et al., 1994: 104), and encompassing the relevant genres and registers most familiar to students.[110]

The aim, in this respect, is to work with lexical items that are available and covered in class through the material, lectures, readings, etc.[111] By means of the concordancer as `new equipment for classroom and home use´ (Rico Pérez, 1994: 34), there is a focus on this data for lexical activities design:

> the development of strategies for guessing unknown words from contextual clues: the multiple contexts offered by a concordance gives the opportunity for the hypotheses generated by one context to be tested against other contexts... with the role of the teacher restricted to suggesting points at which it may help to solve learning difficulties (Johns, 1986: 160-161).

As Higgins and Johns himself claim a bit earlier (1984), these approaches through concordancing activities make, in fact, the teacher a `manager´, having the responsibility of `initiating and giving feedback in the standard classroom interaction´(Higgins and Johns, 1984: 7):[112]

> The computer can take the drudgery out of teaching by doing all the boring, repetitive work, leaving to the human teacher the more creative aspects of the job... not seen as a replacement... but as a supplement (ibid., 9).

The formation of core vocabularies for teaching purposes that are available through concordancing allows decisions to be made `on the vocabulary to be taught in relation to a

[110] In this second step, it is important to `allow then linguistic characterization of any register, comparisons between any two registers´(Biber et al., 1998: 151), and the contrast of, for instance, a conversational genre such as e-mail with `the often dry, mostly descriptive and not very up-to-date textbooks´(Conrad and Rautenhaus, 1994: 2).

[111] These terms dealt with in class should include `core or primary meanings of words... academic lexis... specialized (technical) vocabulary... discourse markers´ (Renouf, 1997: 259).

[112] In examing vocabulary, in fact, questions that should be asked by the teacher on using this technology, according to Higgins and Johns, deal with frequency, range, availability and coverage: `how often does each word occur? In how many of the different texts in the corpus does each of the different words occur? How many of the total number of words are accounted for by a particular sub-set of the word list? What words tend to occur in the environment of a word? What are the contexts in which a particular word occurs?´(Higgins and Johns, 1984: 89).

particular semantic field´ (ibid., 90). There is thus a coverage of `the effective vocabulary component for an ESP course´, according to `the needs and expectations of many students´(Skehan, 1981: 109) in the sense delimited by Thurstun (1996: 5):

> (…) that they [students] become familiar with the use and meaning of the key words, with the patterns of language in which they are used and the kinds of words that are frequently found before and after these key words. Concordancing has enabled us to present the student with multiple examples of the vocabulary items in context.

VII.4. THE TYPE OF CONCORDANCER APPROPRIATE FOR ESP LEXICAL STUDY.

A suitable concordancing tool to work with is Mike Scott's *WordSmith* (Oxford U.P., 1996), as Geoffrey Williams states in reply to Carlson, a newcomer on the Corpora Discussion Electronic List (1 Dec 1998). Yet, there is many other software packages available for concordancing and analyzing corpora. These should therefore be briefly revised in order to specify the inclination in favor of Scott's application.[113]

WordCruncher (Johnston & Company) is a rather useful concordancer for sorting frequency lists and collocations; yet, `it has limitations. It will often find only some of the desired examples´(Jones, 1997: 155). For Applied Linguistics and teaching purposes, this may not be enough and may lead to a lack of supporting lexical data. For other types of study, in contrast, such as a diachronic view of child language, this tool may provide enough of a guarantee (Kettermann, 1997: 193).[114] Or, in the case of the diachronic Helsinki or ARCHER corpora, which present English texts from various periods of time (Biber et al., 1998: 204),

[113] These offer statistical displays of word frequency and collocations, which is crucial to our study. Yet, the choice of Scott's concordancer obeys, as shall be seen, clarity and organizational aims. There are other tools that deal with lexical sense disambiguation or homograph separation, but are not so interesting for the present work dealing with ESP lexis; as Ooi states (1998: 85), `since there should be less lexical ambiguity in specialised or domain-restricted texts, such sense disambiguation tools are probably more necessary for general language texts´.

[114] A PH.D. thesis working in this sense is Olds (1995), who investigates the infant's first lexicon by isolating meaning-based chunks from ambient speech and observing computational evidence (e.g. frequency) of repeated utterances. Another dissertation employing concordancing methods diachronically is Mills (1992), who analyzes meaning of lexical items from middle to late Cornish.

this tool may prove to be effective, since it offers the function of marking a block of text to be searched, and statistics on contrastive distribution (Lukes, 4 Jan 1994).[115]

TACT (University of Toronto, 1990) and *MonoConc* (Michael Barlow [Rice University], 1995) assist in textual analysis by allowing for `such facilities as concordancing, frequencies, indexing and displaying the results in graphs, lists, and tables´(Ooi, 1998: 143). *TACT* includes the Z-score resource and is highly sophisticated, ready to operate with large banks of texts and for literary analysis purposes (Marcos Marín, 1994, 1996). Its main advantages are its `searching capabilities´(Lukes, 4 Jan 1994), and that it creates one's own textual base and defines needed parameters for such an aim (Marcos Marín, 1996: 24).

System Quirk (Holmes-Higgins et al., 1993) is an extraordinary toolkit for building term bases and carrying out multilingual studies (e.g. translations and dictionaries). It is defined as `a system for teaching and learning terminology, lexicography, translation and text linguistics´ (Holmes-Higgins et al., University of Surrey, 1998).[116] However, we feel that it exceeds our concrete aims of contrasting lexis across samples of genres and text types.

MicroConcord and the *OCP* (*Oxford Concordance Program*) (Oxford U.P., 1993, 1988) provide no indexing, and the concordances are not too extensive (Lukes, 1994; Ooi, 1998), while they are too slow with large texts (Bergenholtz & Tarp, 1995: 36). They are mainly resourceful in documenting `frequency of use of particular items in the corpus´(Thurstun, 1996: 4) and work with their `own bodies of texts´(Inkster, 1997: 268). *MicroConcord*, for instance, comes with five sub-corpora of journalistic writing, academic texts, etc., although this `Corpus Collection B comprises "more or less academic texts" (Manual, p. 53)´ (Witton, 1993: OnCall vol. 8 on-line). Output is also limited to 1,600 entries.

Longman's *MiniConcordancer* (1993) is recommended for beginners and classroom

[115] Lukes's critical view can be found at http://anes235-1.ff.cuni.cz/CAAL/0010.html.

work (Witton, 1993), since its main drawback is that it cannot handle corpora much larger than approximately 50,000 words (McEnery & Wilson, 1996: 16). In contrast with *MicroConcord* above, the *MiniConcordancer* is an `in-memory´ indexing concordancer –it records the texts in memory for each new concordance (Tribble and Jones, 1990: 14). Finally, as Witton also asserts (1993: 32), due to the sometimes high cost of these programs, teachers may `come across a very inexpensive alternative in the form of a shareware program´, although the fact is that most software downloadable through the Internet, such as *WordSmith*, has gradually become more affordable. For example, *Word Pilot*, about £ 20 (Milton, Corpora list, 2 Apr. 1999), is `a good vocabulary tool for language learners... accessed via Word97´. Yet, its search options are not too extensive. Alternatively, a freely available on-line concordancer that works as a powerful search engine in Internet has recently been provided by the Research and Development Unit for English Studies (University of Liverpool, 2000). This tool allows users not only to check concordance lines, but also to locate the web page addresses where the documents are located. For language learning using internet technology, this utility proves quite convenient. Yet, because it operates with databases in the web, such as Yahoo or Altavista, it lacks the possibility to read a user's customized files from other locations.[117]

WordSmith tools is thus our choice, since it provides adequate features that analyze one's own arranged texts in order to specify lexis that is both common and key to them. This is crucial to our study based on common core lexical items in a specialized area of ESP such as EIST.

The *WordSmith* software package includes an integrated suite of programs to observe

---

[116] The software address is http://www.mcs.surrey.ac.uk/SystemQ/indexV20.htm.

[117] The address is http://webcorp.connect.org.uk. One other useful web-based concordance is located at http//www.mario.uklinux.net/concordance.

the behavior of words in texts. Two tools (**Concord** and **Keywords**) provide co-occurrences that are significant in a given body of texts. In **Concord**, access to this information is possible through features like `Collocates´, `Patterns´ and `Clusters´. There is also a `Plot´ button to show where words crop up in the texts. The **Keywords** program provide a useful way to characterize a text according to genre / register notions. It is a key utility in *WordSmith*, as it serves to mark subject / topic difference between texts, a crucial resource for our needs to distinguish word categories, as shall be seen in our lexical analysis. Their statistics are formulated by means of a chi-square test. `Key Keywords´ can also be supplied. These show the keywords that most frequently occur over a number of files (e.g. the most typical in a set formed by two or more subjects).

By means of the **Wordlist** tool, a series of frequency-ordered / alphabetized words can be arranged; then, the results can be compared with another list from a given body of texts in order to search for the specific contexts where the items appear as well as their number of occurrences (merging two lists). The minor problem with this function is that, unlike **concordance**, it cannot read files that are not in Rich Text Format (RTF) or Hypertext Mark up (htm) --e.g. Notepad or Word documents (clp. and doc. formats respectively).[118] Its feature `Detailed Consistency analysis´ is quite useful, in contrast, to specify locations where items show up. In addition, through the `Stop-lists´ feature, function words, for instance, may be excluded from a given list of key words. Finally, a convenient facility provided is a search device for Parts of Speech (POS) tags in large corpora (e.g. the BNC Corpus), which allows for string searches of, for instance, progressive verb forms to be conducted. This enables the recording of structures and patterns according to tags (e.g. “V*G>ING” can yield results of all verbs used in progressive forms).

[118] The solution is then to convert those unsuitable files to RTF, which any word processing application can do (e.g. Wordpad).

These facilities and other issues may have led Ooi to assert (1998: 82) that `Mike Scott's *WordSmith Tools´* is `one of the best currently available´, an evaluation with which we fully agree.[119]

VIII. COLLOCATION

In the conception of specialized languages or sub-languages, McNaught (1993: 233) notes that the concordance analysis of lexical patterns or `surface collocation plays a major role´. Indeed, `the sub-language may be described, in part at least, by the choice of the collocations used´ (Ooi, 1998: 161). In order to get at the specific meaning of a particular word-form, from this collocational information, `it is then possible to identify the different meanings associated with a word´ (Biber et al., 1998: 26).[120] Simultaneously, as Halliday points out back in 1966, `knowledge of collocation is knowledge of what words are most likely to occur together´(McCarthy, 1990: 14).[121] This awareness `also enables us to identify

[119] This assessment does not suggest that there are no improvements that can be made in the development of the tools and features of this and other software –the **wordlist** function inconvenience mentioned above, for instance, or the fact that there is a 16,000 word limit for detailed consistency analyses (although the figure is often high enough). In addition, our employment of the programs for Applied Linguistics and teaching purposes often calls for the need of having the concordancing function available along the displayed lines of the `Collocates´ or the `Clusters´ feature, since this would be found quite useful when seeking the actual co-texts of those co-occurrences without having to key in the data. As far as we know, this facility is not provided by any concordancing software package.

[120] These associations would fall under the category `syntagmatic´ as defined by Meara (1982: 29): `responses which form an obvious sequential link with the stimulus word´(e.g. **library--book** –also called a `situational set´[Dudley-Evans and St. Johns, 1998: 84]) and as `responses which are from the same grammatical form class as the stimulus word... paradigmatic´(e.g. **car--vehicle** --defined as semantic set [Dudley-Evans and St. Johns, ibid.]). Collocations would then be `defined along a syntagmatic, or horizontal, dimension, and a paradigmatic, or vertical, dimension´(Nattinger and DeCarrico, 1997: 20). Mel´Čuk (1998: 24) defines collocations, in this sense, as the `lion's share of the phraseme inventory´, and gives four types: delexical verbs + nouns, non-literary adj. + nouns, intensifiers + adj., and technical collocates (ibid., 31). For L2 learners and children alike, as two Ph.D. studies show (Heimbach, 1993; Jarvis, 1997), new lexical items are acquired by negotiating collocational aspects, among other factors. Hoey states, in fact, that when these linguistic dimensions are realized as lexical ties `like beads on a chain... lexical cohesion can be seen to outweigh by far all other kinds of cohesion´(Hoey, 1988: 155).

[121] This is also noted in Halliday et al. (1965), although it is he (Halliday, 1966) and Sinclair (1966) who pioneer the work, after Firth (1957), into lexical collocation from a register perspective. A decade later, `Richards (1976) maintains that knowing a word includes knowing the probability of encountering the word (collocation), along

multi-word items and further justifies their treatment as single items of the vocabulary´(ibid., 15) that acquire specialized language denotation.[122]

VIII.1. DEFINING COLLOCATION IN RELATION TO ESP LEXIS.

To refer to an updated notion of what collocation is, the definition given by Stubbs (1995) may be quoted:

> By collocation I mean a relationship of habitual co-occurrence between words (lemmas or word-forms). A node word may be observed to co-occur with various collocates within a certain span or window, say 4:4, i.e. four words to left or right... A lemma is a dictionary head-word, which is realized by various word-forms... For example, the lemma CAUSE has the forms cause, causes, caused, causing (Stubbs, 1995: 23-24).[123]

A span (`the co-text within which the collocates are said to occur´-- Martin et al., 1983: 65)[124]

---

with knowing its register characteristics, syntactic behavior, underlying form and derivations, association with other words, and semantic value´(quoted in McKay, 1980: 17). From this language learning view, the language user will thus `tend to produce the most probable utterance for a given meaning on the basis of frequencies of previous utterance-representations´(Bod, 1998: 8). An interesting Ph.D. thesis on how collocational knowledge evolves, playing a crucial role in language learning is Gitsaki (1996), who, among other claims, concludes that grammatical collocations are easier to acquire than lexical ones.

[122] For example, the phrase `announce his candidacy´ in the journalist reports genre is considered a lexical unit because of its high degree of predictability (McKay, 1980: 19). Denotation, in these cases, would be the same as referential meaning, in opposition to what Widdowson describes (1983) as the concept of connotation – conventionally defined as `the personal aspect of lexical meaning - often, the emotional associations which a lexeme incidentally brings to mind´(Crystal, 1997: 172)--; there is an attempt to avoid connotation in Science and Technology vocabulary (Widdowson, 1983: 65).

[123] To go back to early definitions of collocation, the year 1951 would be a good start. It was then when the notion of `collocability´as defined and explained by Firth in his paper "Modes of Meaning" (1951) first appeared; thus, the term collocation was coined. Firth also included the importance of `context of situation´ to achieve meaning at different levels.
McIntosh (1961: 331) also stressed this factor in `the assessment of a collocation´since it involves `all other lexical items in the context´. There are thus, according to McIntosh, `ranges´of collocability.
Halliday called lexis `most delicate grammar´in his 1966 paper because of his realization that`collocational patterns are dependent on the structural relations into which the items enter´(Kryshnamurthy, 1997:35). Later in 1976, with Hasan, Halliday would bring forth the notion of `collocational cohesion´ to refer to `the sense of connected discourse created by the close co-occurrence of relatively low frequency words that tend to appear in similar contexts´(Stotsky, 1983: 432).
Finally, the also well-known linguists Jones and Sinclair (1974: 16) claimed that `precise definition of whether any word or group of words constitutes a lexical item will depend on an examination of its pattern of collocation, or regular co-occurrence with other items´.

[124] In this sense, as Sinclair states (1966: 415), those items `on each side of the node that we consider relevant to that node´.

is generally set at 4:4, as is the case in Stubbs (1995) or Sinclair's (1991) definitions,[125] but there are other claims, such as Martin et al. (1983) and Aborn et al. (1959), who propose 5:5. The span position of a collocate is the number indicating the distance of the collocate from the node. Collocates may be defined as `any lexical item which co-occurs with the node within the specified co-text´ (Martin et al., 1983: 85). A node, finally, is `the lexical item whose collocational pattern we are looking for (ibid., 84).

For ESP lexical studies such as our own, these collocational aspects provide insight into the special use of words and their characteristics in specialized language contexts. These are `realized in a text, which is a collection of instances´ (Sinclair, 1991: 102), and `the process by which such sets are identified is entirely objective´ (Phillips, 1989: 5). Collocational information thus enables the observation of an `ample array of multi-word patterns making up the text´ (ibid., 108). These often involve those `lexico / grammatical problems that the students themselves appreciate need attention (e.g. adverb – verb collocations)´ (Fanning, 1993: 163). In a corpus of academic texts, for instance, there may be more than one way of looking at patterns --i.e. not just by text unit--, as repetition is a crucial feature and therefore implies extended revision of patterns across large amounts of texts (e.g. different chapters, articles, etc) (Hoey, 1991: 134, 188-189). Collocations are, consequently, in ESP a `vital element in the reader's understanding of text´ (Phillips, 1989: 5).[126]

A co-occurrent grammatical feature may also be characteristic of certain lexical items in sub-languages; in such a case, the term collocation is no longer used and the concept of `colligation´ is referred to in turn:

> Colligations involve the use of word-classes to name the collocational class. Colligational labels underline the necessary admixture of `formal´ and

[125] See section on concordance.

[126] This variation does not only apply to the notion of text and reading. Patterns are also found and measureable in the rest of the skills for language learning purposes: writing, listening and speaking.

`functional´ uses (Mitchell, 1971: 53).

Common colligations in technical texts are given by prepositions, as in the example of the node **run**, frequently co-occurring with **by** or **out of**, as in *run by management* or *run out of funds* (Bäcklund, 1981: 27).[127]

VIII.2. COLLOCATIONAL STRENGTH ACCORDING TO REGISTER AND GENRE.

The degree of strength in lexical co-occurrences --i.e. how typical or possible such collocations are in a given register or genre-- varies according to frequency and idiomacity. The former has already been examined in the previous section; the latter is an important factor that was first dealt with by Sinclair (1987) in relation to collocations:

> Sinclair later developed his ideas on the *open choice principle* and the *idiom principle*, two key procedures underlying the organization of language, not mutually exclusive principles, but permanently available to the speaker. By the *open choice principle*... [we refer to] things which occur physically together, or concepts in the same philosophical area, often mentioned together. We also make an initial choice of register... But by the *idiom principle*, language users have available a large number of semi-preconstructed phrases which constitute a single choice (Krishnamurthy, 1997: 37).

Language use would lie between these two principles, and thus collocational strength would increase `where words tend to go together and make meanings by their combinations´ (Sinclair et al., 1995: 8). In agreement with Firth's dictum `you shall know a word by the company it keeps´ (1957: 11), 80% of all expressions including syntagmatic choices stand between the two extremes of either free combinations (open choice) or fixed expressions

[127] There may be two other types of colligational occurrences in interactional discourse (not so frequent in written scientific form). These are phrasal verbs (`motive verb + directional particle´, e.g. **tear up, race along**...-- Mitchell, 1971: 53) and `lexical phrases´: `conventionalized form-function composites that occur more frequently and have more idiomatically determined meaning than language that is put together each time´(Nattinger and DeCarrico, 1997: 12). Lexical phrases may thus be `collocations... that have been assigned pragmatic functions´(ibid., 36). They include meaning and form and are `less susceptible to analysis... they are perceived as units´(Peters, 1983: 11).

(idiom principle).[128]

It is a middle ground, then, as Willis implies, where lexical co-occurrence contributes to meaning: `to use a particular word in any number of different ways´(Willis, 1990: 40). In fact, the more idiomatic the expressions are, the more `relatively small number of distinct patterns´can be found (ibid.). `The points in the cline where patterns and collocations are (...) `familiar´ or `semi-restricted´ (Alexander, 1984: 56) provide, in fact, linguistic instances with a greater relevance to ESP lexis analysis: Neither fixed or idiomatic nor utterly new or uncommon constructions can be accurately measured in a corpus to check variation or uniformity of employment.

How strong these lexical combinations are depends on the recurrent use of them in specialized contexts; thus, they may not be idiomatic in general English, but acquire special status in a given register (e.g. academic prose) or genre (e.g. research article). Collocation, in this respect,

> does not exist in language independent of its grounding in text... depending on the texts we encounter, there will be individual variations in the types and strengths of collocation we each recognize (Hoey, 1991: 219).

They are particularized to an area of use and style. Their collocational degree of strength may be measured in terms of range accordingly, as they may appear to a greater or lesser degree across genres (e.g. technical reports, instruction manuals, textbooks, research articles, etc) or in a given register vs. other register(s).[129]

To measure this extent of lexical operation, statistical techniques are proposed in

[128] As Crystal documents in his Encyclopedia (1997: 162), `it is important to distinguish between collocations and associative responses. A lexeme might bring to mind all kinds of `free associations´. Such knowledge of the behavior of words can be amassed by means of concordances, as has been seen. The study of bilingual collocation concordancers are, in this plane, also applied by research (e.g. Kita and Ogata, 1997). The evaluation of monolingual and bilingual dictionaries for L2 learners is also proposed from such a collocational framework (e.g. Hunt, 1997).

[129] This was already hinted at by McInstosh (1961: 337): `the power of each item in informational terms, i.e. the relative degree of restriction of their accepted collocational range´.

corpus linguistics. Biber et al. (1998: 273) refer to the chi-squared test, `the simplest of these techniques´ via concordance listing, `used to analyze the significance and strength of these associations´(ibid.). Word counts are made of text to identify register and genre: contrastive analysis of characteristics such as past tense frequency, use of pronouns, etc, are surveyed. A cut-off score is supplied in order to `eliminate features that do not contribute meaningfully to the interpretation of a factor´ (ibid., 279). These statistics `provide an important tool for assessing whether observed patterns are meaningful´(ibid., 275).

In our case, nevertheless, dealing with common core lexical items and use, the application of this type of statistics should mainly serve the purpose of checking the influence of the register encompassed (i.e. technical text) or distinction of genres involved (i.e. research article vs. textbook). As Jette Pedersen explains (1995: 64), in dealing with this type of information, `in no way should exclusion be based on frequency´; thus, in ESP lexical study, `collocation will therefore be understood as word combinations with lemmata which occur in texts with a frequency of $\geq$1´(ibid., 65).[130] For example, on revising verb use via concordance listings, we need to find out which verbs collocating with which nouns `may be categorised as non-specific and thus uninteresting´ to our ESP study (ibid., 67). In this respect, all restricted or specific uses need be covered, and lexical co-occurrence is either considered banal or specific; collocations are thus `the products of definitional knowledge, the elements being concept bound´(Martin, 1993: 24).[131]

[130] Jones and Sinclair (1974: 27) already discuss this nature of collocations in scientific text: `very little in common with those in the spoken text´. They propose that significant lexical collocates are usually placed to the immediate right of the node and that they `vary according to the subject matter of the text under examination´(ibid.). The two authors conclude that `satisfying lexical sets could be formed based on collocational information´of this type (ibid., 42).

[131] This restricted use of collocations was already observed by Kjellmer (1984: 35) in the Brown Corpus, mainly through the study of distribution across texts: `the more texts a sequence is distributed across, the more distinctive it is likely to be´. Cowie (1981: 223) also recognized these specialized language lexical co-occurrences, especially `from verb to noun... it may enable the dictionary user to predict the possible occurrence of **dividends freeze** or **pensions freeze**´. In this sense, as Varantola (1984: 8) realizes, what make a language special are `the vocabulary, the terms of a special field... also features that show the place of a language on the

Because, as Pedersen reminds us (1995: 70), `the composition of a specialised text corpus will typically be considerably more homogeneous than general language corpora´, there should not be much trouble in picking word combinations `with a narrow distribution´ (ibid., 71) that are informative of the specialized text involved. In this sense, it may be the case that the longer the sequence, the more the informative degree it holds: e.g. `*heat resistance* addressed to the lemma resistance is less informative than *superior heat resistance*, which is less informative than *provide superior heat resistance to´*(ibid., 72). These are multi-word units that may collocate in a given restricted domain without a wide range (e.g. Electrical Engineering textbooks). Their appearance is thus register and genre -restricted, and provide a great deal of information about the nature of the specialized language.[132]

VIII.3. COPING WITH COLLOCATIONS IN ESP TEACHING.

It may be said that learning collocations belongs to advanced language teaching, when students `can infer the meaning from the context´(Brown, 1974: 3).[133] However, as Brown herself observes, this is especially the case with general English students that work with more creative language and literature (ibid., 10). Through the use of collocation approaches and devices in more restricted domains, in fact, students can be made aware of the linguistic environment for their content studies. In Judd's opinion (1978: 73),

> (…) if the students' knowledge of a specific syntactic structure is only at a receptive level, then the students should only be expected to know vocabulary

---

scale of special languages´.

[132] Leech (1980) – quoted in Ackerman and Goldberg (1996: 28) -- defines a categorizing function for these multi-word terms and structures: When it premodifies the head of the noun phrase, the `premodification structure classifies type whereas the form with postmodification only describes certain attributes in common´. In this type of structures, adjectival past participles are quite common as premodifying markers of the head noun; according to Ackerman and Goldberg (1996: 27), their degree of informativeness increases when they contrast `with semantically more specific predicates (i.e. troponyms)´, e.g. ***killed*** *e-mail messages*.

[133] Brown, indeed, recalls that `it has been shown that advanced high proficiency learners rely on associated meanings to register vocabulary in memory (Henning, 1973)´(Brown, 1974:2)

items which co-occur with that pattern on a receptive level.

This kind of interaction with collocations thus involves decoding activities, for which

> the most important kind of information is no doubt denotative meaning. To a lesser extent, indications of connotative meaning and language variety (`register´) are also helpful (Béjoint, 1981: 210).[134]

As long as the lexical information which these word combinations provide is significant in terms of characterizing the given register, collocation should be at least recognized by ESP students at any level. Learners may be encouraged to work with language through learning-centered tasks varying according to the degree of `acceptability´ (Cowie, 1978: 38) that these forms present (i.e. depending on their frequency and range of use):

> (…) that the fully acceptable must be learnt (according to general principles of co-occurrence, if relevant), though it is less certain whether the learner should be allowed (or encouraged) to produce potentially acceptable collocations (ibid., 38-39).[135]

As a result, in English for the field of Law, some collocations that Cowie (1978: 39) provides, *`legal limit, legal weight, legal word, legal term´* serve as samples through which `the learner should be made aware of the mutual dependence of each meaning and each set of collocates´(ibid., 39). An activity such as attempting to establish three out of four referential or denotative meanings of a word by observing collocations of that node can be drawn is this case: legal referring to 1. speed of driving; 2. measurements; 3. judicial statements, etc.

These are common core meanings in a given body of specialized texts that can be examined; these semi-technical uses show `stability´ for learning purposes, as Cowie explains (1997: 130). They are special language lexical sets that provide a degree of `institutionalization´ through collocates (two words to the right and left of the node, such as

[134] Denotative is, in fact, the meaning which interests ESP lexical study --also called referential or cognitive: `the objective (dictionary) relationship between a word and the reality to which it refers´(Crystal, 1997: 451).

[135] There may thus be different levels of learning; some collocational groups, as Nattinger says (1997: 75), may have to be memorized: `by memorizing collocational groups, students will already be somewhat aware of certain lexical restrictions´.

submit a document and **register a document**, possibly fixed, i.e. institutionalized, in the language of bureucracy --Benson and Greaves, 1981: 49).[136] A dictionary that first appears in 1988*, The BBI Combinatory Dictionary of English*, in fact, reveals this growing concern for exposing the most common collocations of headwords (e.g. **run**: **run *by* [management]**). The dictionary proves to be best for encoding activities such as essay writing among EFL students who seek institutionalized word combinations (Benson and Benson, 1988: 344). Collocational information may thus satisfy not only advanced knowledge of linguistic features, but also less advanced learners' needs as well as their production and reception skills demands – e.g. in coping with scanning typical constructions in readings. The lexical path thus goes from receptive to productive knowledge (Bird, 1994). In this respect, our common core lexical items in Information- related studies are framed within the study of their degree and span of collocation, patterning, typicality, idiomaticity, discoursive marking and signposting.[137]

Having concluded the theoretical justifications underlying collocation, so does the lexical focus of our study. Likewise, we have laid the groundwork of our research with the perspectives taken from ESP, Register, Discourse and Genre studies. At this time, the specific applications to our work arriving from the points of departure underlined, will be dealt with in the next chapter.

---

[129] Teaching implications are correspondingly sought in the ESP scope, as shall be seen in chapter 4, mainly involving academic micro-skills, such as summarizing and finding main ideas; lexical knowledge is thus regarded as belonging to declarative, long term memory learning, and, through its exploitation by means of academic tasks, to the procedural ability `concerning things we know how to do but which are not consciously known´ (Richards et al., 1992: 97).

[136] Rudzka et al. (1985) attribute to structure and collocation the basis for lexico-grammatical analysis, and supply exercises based on these contextual clues of lexical institutionalizing. Michael Lewis (1993: 93) also reflects on the nature of institutionalized lexis; in fact, he calls a `lexical item(...) a minimal unit for certain syntactical purposes(...) each is a social institution´.

CHAPTER 2: APPLICATION

The lexical analysis of our corpus is to be conducted after the directions for such work have been given and justified. These guiding points are based on the principles underlined by the scholarship examined in the previous chapter. The focus is thus placed on applying analytical methods to achieve three main tasks: the identification of register, discourse and genre trends as observed in relation to our corpus so that empirical validity may be addressed. This will contribute to the assessment of the selected texts and their consideration as suitable representative material for our research purposes.[138] Secondly, the study of linguistic traits in the textual body will anticipate relevant lexical items to be considered for the next chapter. This preview allows the design of a framework to categorize or distribute the lexis according to main priorities of use. Finally, in the process of doing so, our classification and storage techniques will be described in order to provide the specific ways or approaches with which to gather our lexical data and consider their significance.[139]

The structure of this section, as a result, is presented as three sub-sections or sub-units planned according to the parameters introduced above: Corpus design, Word taxonomy, and Lexical storage.

I. CORPUS DESIGN

For the compilation of our corpus, empirical evidence is a key factor that determines whether such textual collections present enough suitable linguistic data to meet our research needs. In this sense, the aim is to check that the selection of texts be made by paying attention

[138] As Biber explains, `empirical research should be carried out on (…) [the] corpus to confirm or modify the various design parameters (…) there should also be discrete stages of extensive empirical investigation and revision of the corpus design´ (Biber, 1993: 256). The aim is, among others, to check that `linguistic variations correlate with variations in the type of subject-matter´(Ellis & Ure, 1969: 253).

[139] This lexical significance is valued in a vein similar to that underlined by Ingria et al. (1992: 361): by means of `frequency based elicitation of word distribution patterns, concordance-driven definition of context and word behavior, extracting and representing word collocations, acquisition of lexical semantics of verbs (…)´. The

to the concept of corpus representativeness, and that assembling these samples constitutes a procedure of preparing answer material for the types of queries that we shall formulate.[140]

I.1. SELECTION OF TEXTS.

Before moving on to the development of the study on corpus representativeness, however, the basic criteria for the choice of electronic sources that form our corpus must be examined. There are three main requirements that we consider in the selection of sources: Topic relevance and updatedness, Course syllabi, and Availability.

I.1.A. Topic relevance and updatedness.

The basis for the compilation of the corpus lies in the fact that there are subject-related texts provided as encouraged or suggested reading in courses offered in one or more of the disciplines encompassed: Computer Science, Information Science, Telecommunications (Optical / radio communications) and Audio-visual Communication. The sources do not, however, constitute the exclusive list of readings to be undertaken in the subject areas. As a consequence, our aim is not to collect them in detail or exhaustively.[141] Instead, the chief concern is that, due to the rapidly evolving and expanding nature of Information Science and Technology issues, the texts be quite recent –i.e. having been published in the last five or six years (1994-present)--, and serve the main practical purpose of increasing the student's perception and understanding of the topics and concepts which s/he must cope with in the learning environment.

---

degree of specialization as given by subject matter is also accounted for, since word behavior is determined by context of use.

[140] See the notion of representative corpora in the Corpus sub-section of chapter 1.

[141] In this respect, our research differs from lexicographic work devoted to extracting lexis and grammar for the compilation of dictionaries or vocabularies in given specialized areas, and which takes readings included in the courses as primary sources for their corpora, such as G. James (1994) or A. Lozano (1999).

Our emphasis is thus placed on the adequation of subject-matter and language needs in terms of having learners get involved with –e.g. read and work with—updated text types and genres which will activate their academic skills (e.g. note-taking, summarizing, etc).[142] For successful learning tasks and activities to take place, in fact, the development of the students' relationship with texts of the field-specific type which they will have to work with is crucial. Academic writing, in this respect, is becoming a main focus of Corpus Linguistics studies, as authors search and compare learners' errors and styles within genres like essays, reports, summaries, etc. The path followed closely parallels recent development on English for Academic Purposes (EAP), such as Lockett (1999) and Montemayer-Borsinger (1999); the aim is to generate awareness among learners about `key dimensions along which spoken and written genres vary´(Lockett, 1999: 51).[143]

Two main factors are therefore produced and must be considered in the type of orientation underlined: Novelty and authenticity.[144] The former, which refers to the aspect of keeping abreast with ideas and developments, plays an important role in the selection of significant readings in EIST (English for Information Science and Technology) so that both relevance of topic and motivation for the learner can be met. How these are judged implies

[142] We fully agree with Durán Escribano (1998: 62), as a result, when she asserts that reading texts should serve the purposes of understanding genres and learning to use other skills (e.g. writing).

[143] This is the assertion of recent views in Genre and Discourse Analysis, as has been seen in chapter 1: the notion of genre as defined by the actual academic community's needs (experts and non-experts alike) –i.e. from being accepted as a contributor to a reputed journal to becoming a competent reader of scientific texts (P. Thompson, 1998: 69). This same author also focuses on the spoken genre of academic lectures as the exploitation of listening / speaking monologues in academic skills (Jordan, 1997). His recorded speech corpus is sheer evidence of the constant need to look at evolving academic genres such as this, in which both students and professors are active participants. See Thompson's work at http://www.rdg.ac.uk/AcaDepts/cl/CALS/corpus.html. Another earlier work on lectures from a socio-linguistic standpoint is Collot & Belmore (1993), analyzing 200,000 words from nine different conferences. Granger (1998), in turn, works with academic writing by developing with others the ICLE (International Corpus of Learner English) in Louvain.

[144] These two key elements in the selection of readings are based on Dudley-Evans and St. Johns' chart (1998: 99) that divides material according to carrier content and real content; the former includes conceptual level, novelty, value and interest of texts as important, while in the latter, significance, relation to objectives,

the observation of subject contents in the academic programs and syllabi in order to register the concern of a given course with these topics. As Dudley-Evans and St. Johns state (1998: 174), the ESP teacher thus functions as a provider of the material, especially in an area like Computer Science, where `carrier content dates rapidly´ (ibid.). In addition, in the case of research articles, attention is also paid to the impact rate of journals --SSCI journal citation reports-- to see whether they are widely read and distributed in the academic arena, since this genre mainly challenges the advanced knowledge of students in research projects, postgraduates and professors at tertiary settings.

The second point, authentic use, is applied in the sense of selecting and providing reading input which covers different stages of learning and genres. These are authentic as long as they are actually read and used at different periods of study in high education level institutions. Degrees of complexity, in this sense, obviously vary according to the academic year, but they are also based on and influenced by whether learners are doing their minor or major studies, taking more specialized or less specialized courses, etc. In our institution, the distinction is made between `diplomatura´ (a 3 year degree) and `licenciatura´ (a 5 year degree) students, or between `ingenieros técnicos´ and `ingenieros superiores´, the former generally demanding less abstract or in-depth knowledge than the latter at Polytechnic schools.[145]

In addition, the `level of brow´ of texts, as defined by Huddlestone (1971)[146] is also a feature to consider in our specification of authenticity. This author distinguishes three categories based on the differing relationship between the writer and audience: `High-brow´,

---

exploitability, clarity and accessibility are regarded as relevant. In our case, all of them are pertinent; yet, conceptual level, novelty, significance, relation to objectives and exploitability are especially useful.

[145] As was indicated in the Introduction, Computer Science includes both types of degrees (engineering and technical engineering) at our institution.

[146] See footnote 53 in chapter 1.

which refers to the written medium developed by the researcher to communicate his / her findings to other researchers or colleagues –i.e. expert audience--; `mid-brow´, where the writer addresses readers who are neither expert nor general audience –e.g. undergraduate students--; and `low-brow´, writing for popular / commercial purposes –e.g. general audiences. In our case, the first two are especially significant. Research articles belong to the `high brow´ division of texts, while textbooks are essentially considered `mid-brow´.[147]

Textbooks tend to instruct, and, as Myers states, are `backed by [academic] consensus´ (1992: 5).[148] In contrast with reports, the textbook is the chief reading source in this respect, as A. Johns claims (1997), and she quotes T. Kuhn's observation (1963: 362) about the importance of textbooks in science and technology:

> (…) scientific education (…) is conducted through textbooks, works written especially for students. Until he is ready (…) to begin his dissertation, the student of [science and technology] is seldom asked to attempt trial research projects or exposed to the immediate products of research done by others (…) (Johns, 1997: 46)

The profile of students, for whom textbook reading is more important than any other skill, in our case, as has been suggested above, is mainly shaped by `diplomatura´ undergraduates.[149] This is mainly due to the usually less complex nature of the discourse of this genre.

The other genre to be examined, reports, is also largely found to be `mid-brow´; yet, the information is addressed by a semi-expert author to a semi-expert audience, as Bergenholtz and Tarp state (1995: 19). In addition, they are mainly written in an

[147] In contrast, popular science texts such as magazine articles tend to be in the third category.

[148] Myers goes on to observe that too much concentration on the genre of textbooks is not good for the student, quoting Menard (1971): the drawback is that textbooks are `too settled for rapid advance or rapid careers´(Myers, 1992: 5). The alternative is to combine readings and `keep several genres in mind´(ibid., 9) to guarantee novelty and updatedness of knowledge.

informational tenor. Reports, however, require more specialization on particular topics while they are not as selective as research articles. Their authors may range from advanced students whose research allows them to describe different issues, to professional researchers. Reports are found to be located between the other two types of texts, textbooks and research articles, in terms of carrier content sophistication (i.e. conceptual level and novelty –Dudley-Evans and St. Johns, 1998: 99). Their audience will be formed by undergraduate learners who already have some background knowledge about a given issue in the subject-matter and whose interest in specific developments within broader topics encourages them to use additional description of issues, concepts, applications, etc.[150] From this viewpoint, and in agreement with Myers (1992: 9) and Conrad (1996: 302), different genres must be coped with by learners in order to widen both their knowledge and linguistic competence.[151]

In the case of research articles, in addition, because of their degree of complexity and highly technical content, only researchers –e.g. advanced undergraduate (`licenciatura´) students, postgraduates and professors—make up the academic community to which they are addressed. In the case of the first group, moreover, only those pupils with research needs (e.g. final project presentations) will have to focus on this third academic genre. For these learners, as Brennan and Van Naerssen (1989: 202) state, `journals can be a valuable source

[149] Listening to monologue is also quite important at this stage, as shall be seen in chapter 4. According to A. Johns (1997: 46), `reading and listening are the most important academic skills except in advanced graduate classes´.

[150] As Martin notes (1985: 7), `descriptions are, in fact, examples of reports´, and serve to generalize about a given experience.

[151] Conrad (1996: 302) refers to Jolliffe and Brier's (1988) suggestion that `students cannot learn to write in their disciplines simply by reading textbooks because they will not be exposed to the characteristics of valued texts in the discipline´.

of reading material´, since they serve to `develop such skills as taking notes, using table of content, skimming, scanning´.[152]

As an illustration of how a given text can be chosen according to the two mentioned parameters, novelty and authenticity, let us say that a research article on Artificial Intelligence is considered. It may be meaningful to our corpus as it deals with the latest approaches to the subject; this is measured by noting its publication date and origin, 1999 and the scientifically rigorous journal *JAIR* (*Journal of Artificial Intelligence Research*, ranked in second position with an impact factor of 00.864 in 1997). Being quite technical, the article offers a type of discourse addressed to expert readers, in this case, advanced Computer Science students, and thus, its inclusion in the corpus is made with this type of learner in mind. Its analysis should serve to provide the linguistic and rhetorical devices that characterize such a piece of writing for subsequent pedagogical work --e.g. so that learners may gain insight into lexical and phrasal features in order to become `trained´ users of these structures when exploiting similar texts. As shall be seen below, in the design of our corpus, these and other considerations about genre, register and discourse features play important factors.

I.1.B. Course syllabi.

For most lexicologic and lexicographic work carried out with academic texts,[153] the attention to readings done during the year occupies an important position in the process of text selection. James (1994: 21), for instance, states that `the texts chosen for inclusion in our corpus were taken from the prescribed and recommended textbook lists for undergraduates reading Computer Science at the [university]´. He thus chooses 494 sample text files from 164 textbooks and two sets of course notes published at his institution in Hong Kong. Careful

[152] Posteguillo (1997) finds a similar categorization in Computer Science texts: textbooks for lower undergraduate level and research genres (abstracts, articles) for advanced students.

[153] See, for example, Farrell (1990), James (1994) and other works cited in the sub-section entitled `The case of English for Information Science and Technology´ in chapter 1.

observation is also given to the subjects in order to encompass all the possible lessons and topics in the studies: computer engineering, machine architecture, the human / machine interface, data organization, etc. (James, 1994: 33).[154] The aim of the Computer Science program of studies, `to provide not only a theoretical and practical foundation for contemporary computer applications, but (…) also potential developmental progress´(ibid.), is regarded as a key factor in the construction of the corpus. In addition, the Hong Kong Polytechnic corpus has the objective of presenting `an empirical examination of the actual texts such students are exposed to in order to provide information for the writing of teaching and learning materials for language classes and self-access use´ (Cheng-yu, 1993: 74).

In our case, that of the four Information Science and Technology disciplines chosen, course syllabi, as has been implied above, do not provide the exclusive directions for text selection, given the large area of Information-related disciplines selected. Instead, we find that the main function of the study programs should be employed to offer guidance for building a common core body of sources. These should then aid in the location of common subjects and topics shared by Information Science and Technology studies; for instance, by at least two areas such as Computer Science and Telecommunications (Optical / radio communications) in their curricula. Thus, in the construction of our corpus, the course syllabi constitute one feature to consider, but are not the sole basis or goal for selection.

These programs, offered at our university, illustrate the current subjects taken and descriptions of topics of interest in the four disciplines included. In the case of the two most recent ones being integrated at our institution, Telecommunications (Optical / radio communications) (1998-99) and Audio-visual communication (1999-00), other programs from different colleges are also checked in order to gather additional data for academic years

[154] The classification of Computer Science texts developed by Rodríguez and Vivaldi is also quite extensive, encompassing a wide range of subjects and topics for multilingual corpora (1998). See their web page at

not yet studied `in situ´. In addition, existing subjects taken in tertiary settings such as American universities summarize and pinpoint specific topics of the whole area of Information Science and Technology. These are also valued and taken into account at the time of text compilation according to conceptual domain and novelty. As far as we know, there are no encompassing subjects under the name of `ciencia de la información y tecnología´ in Spanish universities; the U.S. curriculum is thus taken as a main reference in this respect.[155]

The following chart (Figure 1) displays the range of subjects arranged according to the disciplines which share them. The asterisks denote that the given area also includes that subject or its topics in the program of studies.[156] The numbers refer to the academic year in which that subject is taken. It must also be said that all the subjects are compulsory in the columns where they appear written under the discipline (e.g. `Information units management´ in Documentation). Where they are taught as electives, this is indicated by the letter E.

---

http://www.iula.upf.es/corpus/acinfca.htm.

[155] Watters (1992: vii) refers to these emerging Information Science and Technology courses as `the study of many facets of information phenomenon´and as `an interdisciplinary study´(...) with `frequent adoption of terms from several specialized subject areas´.

[156] The names of the subjects are taken from the Spanish degree programs, then contrasted with the denominations given by English (American and British) tertiary education institutions and organizations (e.g. for the first listed, `Information units management´[`Gestión de recursos informativos´], the subject `Library and Information Units management´, included in the Hawkins university curricula [Ireland], serves as a model; or, in the case of `Automated Knowledge-based systems´and `Expert systems´, various conferences on cognitive systems list these as fundamental topics (see http://www.emst.pnl.gov:2080). In addition, on-line terminological databases and dictionaries about Information technology technical concepts also contribute to locate the appropriate transfer (see, for instance, http://reimari.uwasa.fi/comm/termino/diction.html and http://www.clres.com/dict.html#onto, where an electronic dictionary such as *Foldoc* [Computing and the

Library science / Documentation (Information Science)	**Computer Science**	**Optical / radio communications (Telecomm.)**	**Audio-visual Com.**
Information units management (4)	* (2 / 3)	* (2)	
Expert systems and Information retrieval (1/2)	* (2)	* (2)	
Automated Knowledge-based systems (4)	* (3)		
Computer systems (4/5)	* (1/2)	* (2)	
Content Analysis (2)			* (2)
Communication theory and standardization (1)		* (2) E	* (1)
Information policy and legislation (2) * (4) E			* (4 / 5)
* (2 / 3)E	Hypertext systems (2)	* (2) E	* (2/3) E
	Software engineering (5)	* (3) E	
	Computer architecture (4)	* (2 / 3)	
	Database systems (3 / 4)	* (2 / 3) E	
		* (2) E	Audio-visual Communication Theory (2)
		* (2) E	Audio-visual Technologies and equipment (2 / 3)
Audio-visual media and reference documentation (4)			* (4 / 5) E

Figure 1: Arrangement of Information-related subjects according to multi-disciplinary parameters.

Internet] can confirm the use of the technical terms). Finally, the English bibliography included in the syllabi at out university also specify some names of these subjects and topics.

As can be observed, only `Hypertext systems´ is taught in all the disciplines. However, according to the actual course syllabi contents, the topic is included in several lessons within other courses, such as `Information units management´, `Expert systems and information retrieval´, or `Audio-visual technologies and equipment´, which implies that more knowledge about techniques and processes is shared by the areas than is explicitly reflected by the programs of studies. In addition, much research conducted in specific areas, such as information / documentation legislation, can be applied to the use of technology in all the disciplines, e.g. computer and audio-visual communication systems policies.[157]

In the cases where two or three majors are offered within a field of study, such as `software engineering´ by Computer Science and Telecommunications, or ``communication theory´ by Library science, Optical / radio communications, and Audio-visual communication, the reason lies chiefly in the more specific or specialized nature of the subject, not finding its practical application in the remaining disciplines. In such a case, the distinction will be made in the lexical analysis, and, as shall be seen, lexis which tends to be more subject-oriented will be specified as more characteristic of only two or three areas in Information Science and Technology.[158]

As was mentioned above, other syllabi in Spanish and American universities have also contributed to depicting the interelationships shown in Figure 1. For instance, the Universidad Politécnica de Madrid, which offers the technical specialty in Optical / radio communications since 1971 (as Radio) and 1992 (as Optical / Radio), has provided feedback on the bibliography included in second and third year course syllabi, since these are currently being studied or have yet to be studied at the Universidad de Extremadura. An example is

[157] This type of implicit multidisciplinarity that is not shown by the programs of studies is taken into account at the time of developing the lexical analysis in order to establish links between texts and subject areas (see chapter 3).

[158] The discussion offered in chapter 4 deals with this issue in the description of the core lexis.

`software engineering´, taken as an elective in the last year of studies, with shared topics on object-oriented design, programming languages, etc.[159]

Another institution in Madrid, the college of Information Science at the Universidad Complutense, has provided us with guidance concerning academic focuses and requirements in Audio-visual communication studies. An example is the observation of the importance given (two full years) to international policies on copyright and ownership law in the fourth and fifth years, which produces a rather large number of legislation sources. In addition, the technical aspect of using audio-visual technology such as radio, television, video, etc. demands acquaintance with literature that is also included in the Optical / radio communications program.[160]

Several universities in the United States integrate at least two of the Information disciplines in schools. Such is the School of Information Science at the University of Pittsburgh, for instance, which contains four departments that closely parallel our four areas: Department of Information Science, Telecommunications, Computer Science, and Communication.[161] These often share common subjects that deal with topics and concepts in all fields. An example is given by the reading list of the course syllabus `Information Science and Technology´, taught in the Spring semester (1998). It offers the following units:

[159] There are actually other subjects which contain similar Computer Science topics on object-oriented software design, such as `Programming´ and `Digital systems´. In addition, `Operating systems´ is also shared by both programs; yet, this subject is not included in the Universidad de Extremadura curriculum for Technical engineering in Telecommunication, Specialty in Optical / radio communications.

[160] Given the similarity in nomenclature, at first glance, it would seem that Optical / radio communications contains a lot of subjects in common with Audio-visual communication; however, this is not the case, and the use of similar bibliography is only reflected in a couple of courses. The college of Information Science in Madrid also provides some directions in the study program for Documentation in order to contrast or add any feedback to the syllabi examined at the college of Library science and Documentation in Extremadura. For `Information policy and legislation´, particularly, not included originally (i.e. in the 1997 program), the subject `Information and Documentation policies´ is a suitable alternative to checking the contents involved.

[161] The Department of Electrical Engineering is also included in this School of Information Science.

1. INFORMATION SCIENCE: Library and Information Sciences,
2. INFORMATION SCIENCE, COMPUTER SCIENCE and TELECOMMUNICATIONS: History of computers, Hardware, Software,
3. ALL DISCIPLINES: Perspectives on Information, UNIX / Internet, HTML, SGML, TEI,
4. TELECOMMUNICATIONS: Telecommunications,
5. INFORMATION SCIENCE, COMPUTER SCIENCE and TELECOMMUNICATIONS: Information Management, Online database systems,
6. ALL DISCIPLINES: Electronic publishing, Information infrastructure,
7. COMPUTER SCIENCE and TELECOMMUNICATIONS: Data communications and client-server architecture.

As can be seen, all items mentioned in Fig. 1, except the more specialized topics (e.g. Computer engineering and architecture, or Automated Knowledge-based systems), are included in this syllabus. `Audio-visual communication theory´ and `Content analysis´ are not listed either.[162] `Information policy´ is dealt with in point 6, and `Hypertext systems´ in point 3. These are common to all disciplines. Lines 2 and 7 are related to Computer Science and Telecommunications; number 2 can also be included under Information Science (`Computer systems´in Fig. 1). The same applies to item 5, where subject `Information management´ is in Information Science and Computer Science, whereas `Online data systems´ can be placed under Computer Science and Telecommunications. Points 1 and 4 are obviously found only in Information Science and Telecommunications respectively. Like Computer Science and Audio-visual communication, they are thus treated as separate disciplines encompassing their own range of topics on Radars and Optical fibers in Telecommunications, Virtual libraries in Library science, etc.

For the purposes of subsequent common core analysis and reference, given the data obtained from universities and other institutions, both denoted in figure 1 and the syllabi from abroad, the subjects and topics shared by the different disciplines are labelled as follows (see also Appendix 1):

A =COMPUTER SCIENCE and TELECOMMUNICATIONS
--History of computers, Hardware, Software [A**1**],
--Computer engineering and architecture, Data communications and Client-server architecture [A**2**]

B=INFORMATION SCIENCE, COMPUTER SCIENCE and TELECOMMUNICATIONS
--Information units management [B**1**],
--Online database systems, Computer systems [B**2**],
--Automated Knowledge-based systems [B**3**]

C=INFORMATION SCIENCE and AUDIO-VISUAL COMMUNICATION
--Content analysis[C**1**],
--Media documentation [C**2**],
--Documentation Legislation [C**3**]

D=TELECOMMUNICATIONS and AUDIO-VISUAL COMMUNICATIONS
--Media technology [D**1**],
--Media theory [D**2**]

E=INFORMATION SCIENCE, TELECOMMUNICATIONS, and AUDIO-VISUAL-COMMUNICATION
--Communication Theory [E**1**],

F=ALL DISCIPLINES
--Perspectives on Information[F**1**],
--UNIX / Internet[F**2**],
--HTML, SGML, TEI, [F**3**],
--Hypertext technology [F**4**],
--Electronic publishing [F**5**],
--Information infrastructure [F**6**]).

These headings (six – **A** –**F** categories) serve as major references to gather information on subjects and topics, mainly to detect those areas of common academic and professional issues. The items included under each section correspond to those themes coped with in the specified disciplines – e.g. **A1** is the label placed for lexical analysis purposes to those texts dealing with the subject of the history of computers, hardware and software, seen in both Computer Science and Telecommunications areas. These findings from different syllabi in the Information Science and Technology areas are regarded as positive data that support our

[162] This is mainly due to the fact that the Department is referred to as `Communication´ Dept., not more specifically as Audio-visual communication.

aim to work with a multidisciplinary corpus. The search for a lexical common core in these disciplines, which is the central focus of our dissertation, is thus justified and based on academic grounds: the documentation found in actual programs and current syllabi. The choice of texts for our corpus thus originates from these sources, but, as has been explained, does not consider them the sole source of selection.[163]

I.1.C. Availability.

The notion of the `teacher-researcher´ to which T. Johns (1993) refers in his work on data-driven learning specifies that teaching English `trickles up to research´ (Wichmann, 1997: 2).[164] The ease of accessibility to the corpus is an important condition so that texts can be easily exploited in those two facets, learning and investigating language.

The reading material should be readily available, in this sense, not only to facilitate the central role of the `teacher-researcher´, but also for learners to use as independent work. The computer has a central role in this context, as the tradition of CALL (Computer Assisted Language Learning) activities prove --originating shortly after SELMOUS (Special English Language Materials for Overseas University Students) began to operate in Birmingham in 1974 (cf. Crocker, 1981; Jordan, 1997). Nowadays, the position of CALL in language learning is well known and established, as the numerous CALL sites in Internet may prove.[165]

Yet, as Foucou and Kübler (1998: 2) also find, `human interaction improves the selection and preparation of the learning activities´. From this perspective, both teacher and learner are

[163] This information will be contrasted and compared in this chapter.

[164] As Brumfit puts it (1997: 36), `what is needed is research for people whose prime concerns are practical´, and `research needs to be interdisciplinary´(ibid., 37). This recent demand may be contrasted with Strevens (1977) --although his English for educational purposes (EEP) acronym has not been thereafter applied to these cases-- when he affirms that subject specialists collaborate to a great extent with language teachers: `there is even a trickle of subject specialists entering the language-teaching profession´(Strevens, 1977: 59).

[165] See summary of resources at http://www.cltr.uq.edu.au/oncall/callink.htm, or the on-line CALL journal (http://www.cltr.uq.edu.au/oncall/vol8ndx.htm).

free to choose which texts provide the suitable balance according to needs and motivational factors. The main advantage, as Dudley-Evans and St. Johns (1998: 99) claim, is that

> Learners own the texts and are involved and committed to them. These texts may be allotted classtime or self-study time according to whether they represent group or individual needs and interests.

We thus find, given these priorities of text availability, that the most appropriate source of access and management is Internet, where electronic publications are not hard to identify according to topics and degree of complexity.[166] The World Wide Web, in fact, provides users with various means of searching for documentation, mainly by using `Net Search´ options in any browser to locate the site by keying in the relevant word(s). In this respect, as White claims (1999: 1), `the Web is certainly one such tool which has immense power´. As a consequence, according to Tribble (1997: 116),

> (…) the corpus is no longer the preserve of the university or commercial research team. Teachers and students are beginning to have access to corpus resources and are beginning to work with them in interesting and creative ways.[167]

This fact illustrates the difference between doing scholarly research –i.e. Corpus Linguistics generalizations about `the language´-- and `a text-oriented Corpus Lingusitics where the idea is to find about a text or set of texts´, as M. Scott explains in the Corpora discussion list on Internet (9 Oct. 1998).[168] The idea of or belief in `building your own corpus´(G. Williams, 23 Oct 1998) is actually quite extended among the participants of such a forum for debate on Corpus analysis. In this respect, `corpus linguistics is concerned with a core of principles and techniques to analyse language scientifically by using the computer´(Berber Sardinha, 12 Oct 1998). Electronic texts in any format –e.g. full text, Rich

[166] See concordancing features for lexical analysis in chapter 1.

[167] As Pearson asserts (1998: 87), however, learners should not be left to their own devices when choosing texts as some teacher's supervision, mainly through the presentation of `guidelines (…) might help students to choose their texts more wisely´.

Text Format, Hypertext, etc--, in an academic, journalistic, or popular writing register,[169] serve the practical purpose of learning and researching as they are downloaded and exploited from the WWW.[170]

As a result, for the choice of our corpus sources, the third feature, namely, availability is highly considered, given the reasons explained. The other factors examined, updatedness and authenticity, and the use of academic programs as a reference, are also taken into account to a great extent. The Web is thus browsed and searched as the point of departure in our selection, while the key topics are identified according to the needs analysis derived from course syllabi requirements and concerns. In addition, we subscribe to the importance that the texts present challenging and motivating content in relation to the learners' interests and needs at school in this study.[171]

I.2. CORPUS DESIGN ASSESSMENT.

The second sub-division in this section on corpus design is, as was said at the beginning, to present linguistic data collected and structured according to the texts selected.

---

[168] The list server can be reached at corpora@hd.uib.no.

[169] As an example, textbook language such as computer explanations found at http://www.pcwebopedia.com and magazine reviews of products in http://www.compulang.com differ substantially.

[170] This new status in which students become their own teachers of language through the use of WWW resources can be seen reflected in some works published at the proceedings from the international conference *The Major Varieties of English* (MAVEN 97) at Vaxjo University, Sweden, 20-22 November, or those from the *Teaching and Language Corpora Conference* (1998) –e.g.Wyatt, Simpson, Warren, Pearson, Osborne, etc. These are studies of on-line discourse –e.g. e-mail discussion lists or chat lines--, researching how the educational discourse of students operates. Recently, Leech has introduced `the first book devoted to the idea of collecting a corpus of the language produced by foreign language learners´ (Leech, 2000: xiv), referring to Granger's compilation of learner electronic writing. Other works explore on-line language learning through activities based on CALL exercises (see chapter 4 for references).

[171] Some addresses and references of servers which distribute electronic texts and allow their access and management can be found at the following locations, among others: http://www.wkap.nl/kaphtml.htm/ON-LINE, http://xxx.lanl.gov, The On-line Books Page, and The Electronic Text Center of the University of Virginia. These provide access to research on science and technology in English. Some are freely available, like the second address given, or they might require that users subscribe or register for their service, as in the case of the last reference.

The linguistic claims made in this respect will serve to evaluate corpus representativeness in terms of genre and text type concerns. In other words, the analysis undertaken in this part will be used as both a previewing and testing phase in which the aim is to sift through the content of a given range of texts to value the response data needed for lexical queries (e.g. distinguishing word functions and types); we thus aim to check that the corpus is representative in that respect.[172]

The analysis offered below is basically an observation of word frequency according to Mutual Information (M.I.) parameters, a statistical measurement to assess the importance of probable co-occurrence among words. For highlighting significant collocates, as we intend to do, this type of statistics is found to be quite useful, since we seek significantly frequent co-occurrences, that is, lexical items that appear recurrently enough in given stretches of text (e.g. from a sub-section such as a text type, through the genre as a sub-corpus, to the entire body of Information Science and Technology texts). Those words are included as relevant items in the framework by which texts are assessed according to register, discourse and genre values or factors: the academic textual setting thus determines characteristic lexical use, and the employment of certain linguistic features may be explored to revise such a context.[173] The concordancer applied, *WordSmith*, enables such a type of findings and statistics.

We thus agree with Tribble (1998) when he states that `it is possible to use WordSmith Tools to identify criterially significant differences between a genre and a reference corpus´. In our methodology, small corpora of academic genres (textbooks, reports

[172] O. Mason asserts, `it is impossible to define what `representative´ is in any definitive way´ (Corpora discussion list, 9 Oct 1998), mainly due to the fact that language is not an abstract system, and thus, there will always be users who do something different with it. Our own concept of representativeness, based on the literature to a great extent --see corpora sub-section in chapter 1--, is thus also flawed from this absolute viewpoint. Relatively, as a result, our search for this notion chiefly follows the practical applications that the study of common core lexis entails (e.g. common core academic vs. subject items).

[173] As Alejo (1998: 228) states, approaches to lexis should not lack `the implications of their analysis on discourse or discourse patterns´: To summon register and discourse evidence by means of genre contrasts.

and research articles) serve as contrast sources for the examination of key linguistic items. As Johansson (1991: 305-306) states, `smaller sample corpora´ are especially useful for learner language, and thus, our concordancing procedure, like Bernadini's (1998: 24), suggests a `shift of perspective from `corpora as reference tools´ (…) to `corpora as sources of language learning activities´´. The aim is thus to examine the semantic prosody of given items in order to provide the basis for our research project.[174]

The approach taken is similar to that of various authors when using concordancing procedures.[175] Altenberg (1993), for instance, examines the uses of **such** as a multifunctional word which presents `fuzzy´ elements in its functionality (e.g. its use as a vague intensifier). In order to determine where and under which conditions this word is employed (e.g. the `construction with *so* + adjective, optionally followed by a resultive *that*-clause´--Altenberg, 1993: 232), corpora or sub-corpora are analyzed and contrasted. The main goal is to check register varieties of English by means of observation of these semantic prosody changes.[176] In addition, as shall be examined in chapter 3, the study of lexis cannot be entirely treated separately from the observation of grammar features.[177] This view influences our analysis in the sense that we shall look into function words in order to describe the colligations and clusters that commonly characterize our sub-corpora. The contrast is provided among the

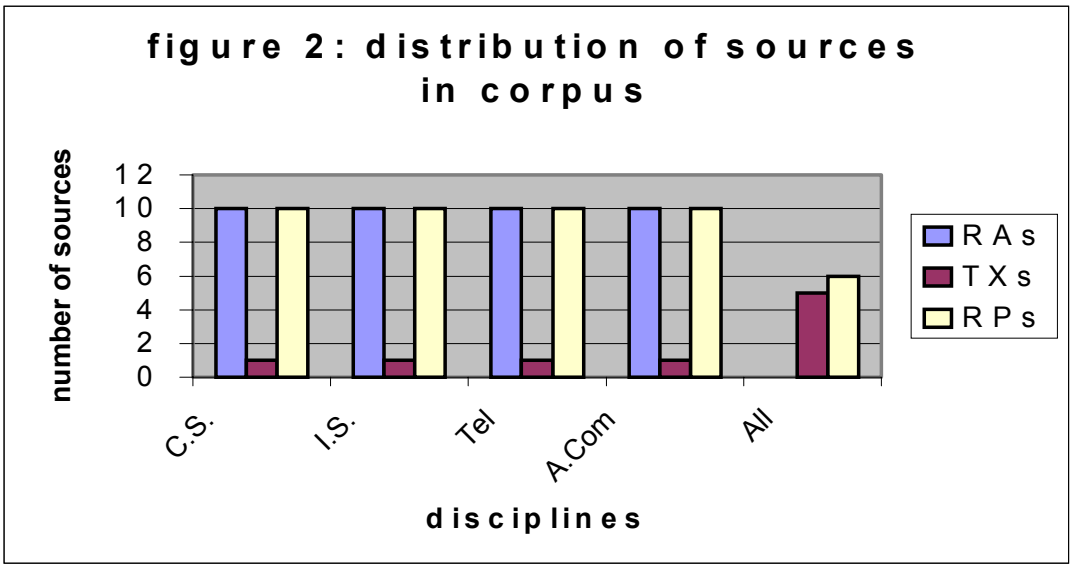
[174] Semantic prosody is `used by Louw (1993) to refer to a type of linguistic relationship which has not previously been clearly documented. *Prosody* has its Firthian sense of a phenomenon which extends over more than one linguistic unit, in this case over a span of words´(Stubbs, 1997: 246). This author then goes on to analyze the senses of the verbs **cause** and **provide** from this concordancing perspective.

[175] See, for instance, Stubbs (1996, 1997), Sinclair (1997), Hoey (1997), Howarth (1996, 1998), and others in concordancing features for lexical analysis in chapter 1.

[176] As Widdowson (1979: 19) states, `the argument here is that the formal linguistic characteristics of these varieties of registers of English are only of interest as realisations of underlying concepts and communicative operations´. Altenberg (1989) also asserts that these negotiations of meaning are thus likely to differ according to register variation; yet, he says that Biber's work (1988) lacks lexical insight.

[177] Sinclair (1966) and Halliday's (1966) vigorous claim of this approach is currently gaining increased recognition in the case of, for instance, assessing grammar textbooks (e.g. Collins, 1998).

Figure 2



three genres: textbooks, reports and research articles (sub-corpora A, B and C correspondingly).[178]

As shown in the Corpus distribution of texts (figure 2 above), in the case of Textbooks, there is one whole textbook for each Information discipline (Computer Science, Information Science, Telecommunications and Audio-visual Communication). In addition, there are five textbooks classified as common to all four disciplines. Regarding this genre, only relevant content chapters have been included, discarding such sections as prefaces, tables of content and bibliographies. As far as Technical Report samples are concerned, ten has been the number of sources selected for each of the four disciplines mentioned. In addition, six reports are added which represent all four technical areas of Information Science and Technology. Finally, in the case of

[178] We refer to the notion of sub-corpora, in this way, as that defined by Marcos Marín (1994: 84): `un sub-grupo dentro de los corpus que (…) funciona como (…) su selección dinámica durante determinado análisis´.

Research Articles, we have worked with 10 texts in each of the fields (see figure 2, where this distribution is represented graphically).

In figure 3, the three genres or sub-corpora are, in turn, compared according to the number of words. As can be observed, the three figures closely parallel each other; in other words, a balance is maintained in terms of frequency and range. The textbooks sub-corpus

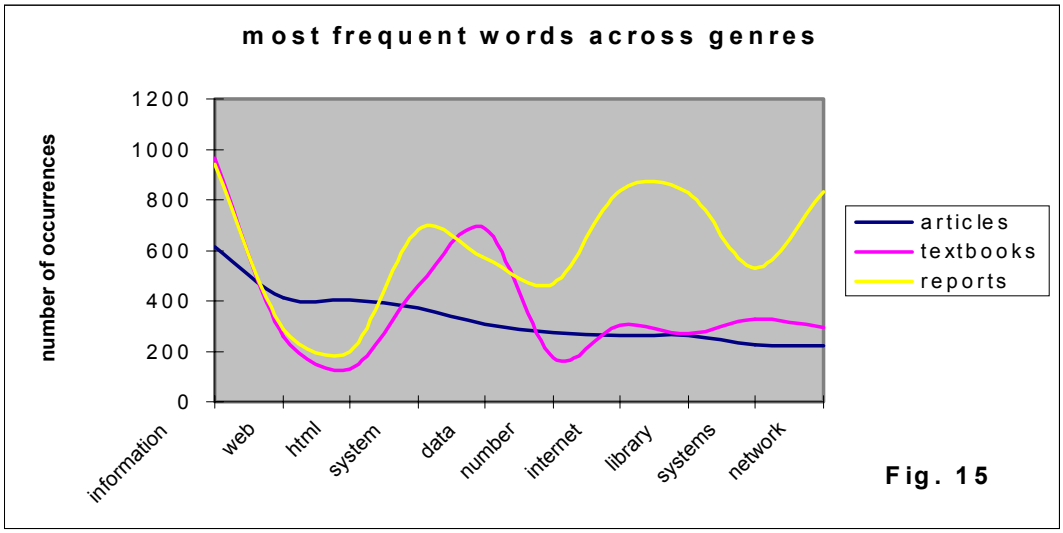


Fig. 15

contains 320,385 running words (tokens) and 14,961 distinct words or types, while reports have 280,549 words and 13,829 types. The research articles sub-corpus presents 256,438 tokens and 11,304 types. This set of texts has a lower standardized ratio than TXs and RPs (34.59%) –this rate indicates the number of repeated items per 1,000 words. Such a lower level of repetition suggests a more lexically developed scientific / academic discourse in the case of Research articles. The overall statistics of the corpus is 857,372 words and 24,833 different types (standardized ratio of 37,05 %).

On this corpus, as a result, we should perform next a testing analysis through which to check the type of academic / scientific-technical register encompassed. The four items to be examined are based, regarding choice, on Genre Analysis.[179] Bhatia (1998: 18) introduces the

[179] As Ventola (1987: 61) claims, in order to get at the register and discourse levels –i.e. to expose features and `the possible combinations of choices from the register networks of field, mode and tenor in texts´, genres must be examined.

observation of first person pronominals --**we--**, modality, cohesive devices and verb tense as highly useful factors for distinguishing text types and genres. Regarding the first feature, he states that the said pronoun is used in journal articles to `set out the claims made in the context of earlier work (…) whereas in textbooks, (…) the use of `we´is often inclusive´(ibid.). About modals or hedges (as they are called in Applied Linguistics), he asserts that there tends to be a lower number of them in textbooks, since in this genre, `claims are accredited facts´(ibid., 19).[180] We thus embrace these statements and set out to test the use of `we-constructions´ and modals in our corpus. In addition, because there tends to be much emphasis placed on aim and purpose in Computer Science research (Fortanet Gómez et al., 1998: 70), we explore the preposition **to** as a colligation of verbs, and the frequency of connectors or cohesive devices such as **so that**, **in order to** and **due to** for the function of purpose. Finally, instead of verb tense, given the predominance of the passive voice, active is discerned from such a voice in our contrastive study.[181]

In the first case, the use of `we-statements´, it can be generally said that in sub-corpus A (textbooks), these constructions reveal a less formal approach than in reports and journal articles. This last genre, indeed, denotes the highest degree of formality (cf. Biber et al. [1998]) by means of the writer's detachment or distance in his /her choice of verbs co-occurring with **we**. In addition, research articles contain a larger number of instances (738) than either textbooks (506) or reports (547). Figure 4 displays the most frequent verbal uses with the first person plural pronoun.

[180] Latour & Woolgar (1986: 77) state a similar claim: textbooks are considered `the prototype of scientific assertion´.

[181] This last choice mainly arises from personal interest in checking empirical evidence of passive use with modals alone; Dudley-Evans and St. Johns (1998: 76) contribute to triggering this type of query.

Research articles (RAs)	**Textbooks (TXs)**	**Reports (RPs)**
____ have (57) described called found used utilized discussed created designed added	_____ have (29) seen found discussed decided to arrived reached built up	____ have (15) built provided found
____ will (18) use describe develop * as ____ shall see (3)	_____ will (16) use define	_____ will (40) use describe provide * as ____ shall see (6)
_____ use (10) see describe provide market state define find (3)	____want (31) to (16) (therefore) include* recognize (that) * need (to) * allow require intend use expect make seek to develop * wish (to) * begin with define think consider (4)	_____ use (31) recommend believe (that) * need (the) * provide present call define suggest want to * write 've expect (3)

Figure 4: Verb forms collocating with the pronoun **we**.

The horizontal lines represent the key word analyzed (**we**), and the numbers refer to the frequency of the words which precede them –e.g. in the first set under RAs (research articles), *we have* appears 57 times as present perfect auxiliary followed by the participles listed by frequency. Verbs that co-occur a minimum of three times are listed so that clusters (signalled by the asterisks) may be accounted for; the *WordSmith* function **Clusters**, in fact, takes three as the lowest frequency in order to delimitate the relevance of these phraseological

devices.[182] Some clusters contain words in parentheses; this means that the given item does not colligate in all the occurrences of the verb –e.g., in TXs (textbooks), *we want to* occurs 16 times out of the 31 instances found of *we want*; *we seek to develop*, in contrast, forms a uniformed entity, as *to develop* always co-occurs with *we seek* in TXs (six times).

The tenses encompassed in Fig. 2 are present perfect, future and present. There is another one used in both sub-corpora A and B (TXs and RPs), yet, it is non-existent with we-statements in the case of research articles. This is the present continuous. There are 28 instances, for example, of the string *we are* + ing-forms in textbooks (followed by examples like **considering**, **doing**, **referring** and **defining** –in order of frequency).[183] Another feature that is not present in sub-corpus C (research articles -- RAs) is *we are* + -ed adjectives, such as **interested** (5 occurrences), **challenged**, **disappointed** and **disciplined** (3)(in RPs) or **interested** (5 times), **concerned** and **experienced** (3)(TXs). There is, in fact, a complete absence of adjectives with we-statements in RAs, whereas in reports, the cluster *we are all familiar* occurs (5 times), *less familiar* also appears (3), and other qualifiers are present but do not occur more than three times: **grateful**, and **happy**.[184]

These and other verb uses (Fig. 2) such as **want**, **intend**, **wish**, **think**, **believe**, **'ve** and **expect** in sub-corpora A and B (TXs and RPs) lead us to postulate that the discourse interaction between the writer and the reader is more personal and less constrained by academic formality than in research articles, where none of these lexical items appear with we-statements. As Bhatia thus claims (1998: 18), this shows a more procedural approach, on

[182] Some lexicographic work also takes this number as the Mutual Information yardstick to register clusters (see footnote 103 in chapter 1).

[183] In reports, the first person plural present with gerunds is employed much less; only **talking** and **developing** are worth mention.

[184] Sub-corpus A does not include this type of adjective either.

the part of RAs, to refer to what has been, is and will be done in the body of text by the author(s), a registerial assertion framed as a result of analyses such as this.[185]

In this genre of research articles, as Swales states (1990), the pattern of Introduction, Method, Results and Discussion (IMRD) is commonplace. We have also identified in ten cases out of the 32 articles which make up sub-corpus C, the inclusion of Conclusions. Ten abstracts placed at the beginning of these texts are also included separately.[186] We thus focus on the specific sections or text types of research articles where the `we-statements´ co-occur with relevant lexical items in order to check variation across different sections within a genre. The aim is to generate awareness among learners about `key dimensions along which (…) genres vary´(Lockett, 1999: 51). This allows the specification of clusters that characterize a given context in RAs; the following table presents the most frequent cases:

[185] Genre analysis thus takes us to observe variation in register features (dimension of science and technology [Biber et al., 1998]), given these claims (see sub-section on genre `Genre as a specification of register´in chapter 1.

[186] Abstracts constitute, however, a completely different genre (see, for instance, Posteguillo's detailed study on Computer Science abstracts [1995]). Yet, the ten cases included in our sub-corpus are treated as sub-sections of Ra due to the chief aim of observing and contrasting co-occurrences of lexical features.

Abstracts	Introduct.	Method	Results	Discussions	Conclusions
..paper __ (^4) ..study __ __ report findings .section _ __ review (24)	In this paper ___ (^6) ___ describe ___ will describe __need to __ must face __ present (87)	_can (^24) _have _define _will _use _describe _chose _ con-ducted _ provide _develop (242)	In this section __ (^11) __ use the __ focus on as __ have __ can state that section ___ outline ___begin by ___can find __ describe (200)	__ found that (^14) what __ have __ already discussed __ apprea-ciated the input __ believe our study __ draw attention to __ found (194)	___ have drawn the (^3) (10)

Figure 5: **We** collocations in Research Article sections.

The numbers at the top indicate the frequency of the first item (e.g. *paper we* in abstracts is repeated four times as either the cluster *In this paper we* or *in the paper we*); these are the highest co-occurring counts in each column.[187] The bottom quantities refer to the absolute instances found of `we-statements´for each section.

The data in Fig. 5 illustrates how the sub-parts of RAs can be identified according to key words and clusters. This is quite relevant to our lexical study, since we can thus check that context determines the lexical items to be used. By means of concordancing procedures, this fact can be verified, as significant co-occurrences are pointed out.[188] An example is the

[187] As in Fig. 2 above, the minimum number of repeated intances to be accounted for is three (the lower items at the table).

[188] KWIC analysis plays a crucial role in determining context through co-text observation (see footnote 98 and Concordancing features for lexical analysis in chapter 1).

reference to the process of what is being done (or can be done) presented in the articles by using various procedural verbal items in the methods sections (e.g. *we define*, *we describe*, *we use*, *we will*, etc -- see figure 5). This can be contrasted with the discussions, where the writer's use of reviewing or concluding remarks in the form of past tense expressions prevails with `we-statements´ (e.g. *we found that*, *we already discussed*, etc – see `Discussions´ in figure 5). A further exemplification is offered by identifying the presence of such utterances as *in this paper, we..., we describe*, or *we present* in the `Introductions´ (see figure 5) to establish an introductory or presentation note for the work to be exposed. In all cases, the lexical elements characterize the textual sample or type being represented, or, in other words, such small sub-corpora determine the items and range of word use to be encompassed.

Other features co-occurring with **we** in the rest of the genres also allow us to note down genre distinction. This time, the main difference is held by sub-corpus B, the set of reports. The elements under examination are the meta-discoursive markers of cohesion that refer to parts or sections of the text by which the author denotes a summarizing technique. In articles, for instance, the cluster *In this section we* is found 12 times. Other examples, *here we*, *then we*, *we* ____ *above* (where the gap is filled by past tense verbs such as *stated, mentioned,* and *described* – see figure 5)[189] are also identified in RAs. However, in TXs, elements such as *hence we* (9) and *in this section we* (7) are frequent, but this is not the case of *we have stated*, *we have described*, etc, according to the data measured in our corpus.[190] In reports, by contrast, these cohesive resources are not employed at all. This fact may contribute

[189] This type of clusters may be classified as anaphoric constructions (see Gil Francis [1986] in Lexical Focus of chapter 1).

[190] In the case of textbooks, in fact, and in opposition to Bhatia's claim that fewer of these elements are contained in such a genre (1998: 18), several cohesive devices are found. Other items of this sort located with we-statements in textbooks, according to our data, are *we therefore* (16 instances) and *however we* (9). Interestingly enough, in reports, the number of these cohesive elements is lower in their co-occurrence with **we**, as our measurement attests: *however we* (8) and *in addition we* (4).

to drawing assumptions about the type of academic discourse in such a genre, where authors do not seem to regard their text as an arrangement of sections or sub-parts, but as a sort of forum by which to expose their claims and findings in regard to a given issue; this is, in fact, a distinct way of performing discourse community roles, according to some scholars.[191]

Last but not least, in relation to our lexical selection, as shall be seen in the next chapter, semi-technical words can be considered in their co-occurrence with `we-statements´ so that their classification and arrangement may be facilitated. This is the case of verbs such as **use** and **have found** (Fig. 2), registered in all genres under the observed semantic influence of `we-constructions´.

As a second concern, the function of purpose is submitted to concordancing techniques as another main means of checking the evaluation of representative traits of style in our corpus. Firstly, the connectors **so that**, **in order to** and **due to** are contrasted across texts. In the following figure (fig. 6), the counts and common co-occurrences of these items are given:

Sub-corpus A (Textbooks)	Sub-corpus B (Reports)	Sub-corpus C (Research articles)
In order to (70) ..make (6) ..transmit ..facilitate So that (78) ..they (11) can ..it ..users Due to (6) the (3)	In order to (78) ..maintain (7) ..provide So that (65) ..they (9) ..the underlying Due to (41) the (16) ..the cost of (3) ..the lack of ..the network	1. **Introductions** In order to (2) So that (2) Due to (1) 2. **Methods** In order to (14) ..allow (3) So that (16) the (6) + pl. N they Due to the (17) 3. **Results**

[191] In this case, the lack of a type of cohesion signals the type of interaction with the discourse community (see `Coherence and cohesion´ in the Discourse section of chapter 1).

		In order to (13) So that (13) users (5) ..we ..it Due to (15) ..the fact (10) ..the activity 4. **Discussions** In order to (25) So that (10) ..information(4) ..use Due to (6) + Pl / Sing N 5. **Conclusions** In order to (1) Due to the (8) + Sing N

Figure 6: Purpose items across genres. Pl.N= plural noun / Sing.N = singular noun.

Research articles, as can be seen, are divided into their sections and these are analyzed accordingly. Collocates are given under each node and arranged, as in the other tables, according to frequency. When no co-occurring items are specified, as in the case of **in order to** in the Discussions sections, it means that there is no individual verb appearing more than twice[192] (e.g. **build** and **provide** occur twice in this case; several others, like **avoid**, **investigate**, etc, are found only once near the node).[193]

The verbs **make** and **maintain** are most often repeated across the textbook and report genres as the clusters *in order to make* and *in order to maintain* respectively. Research articles add *in order to allow* and *in order to provide* as key constructions. In the case of the second connector, *so that they* is actually the most common string identified in textbooks and reports; yet, as shown by RAs, specifically in three key sections, Methods, Results and

[192] Likewise, where the analysis results are not provided, as in the case of **so that** in Conclusions, or in Abstracts, not shown at all, it means that not a single occurrence of the words was located.

Discussions, three other subjects are revealed after the node: the article **the** followed by plural nouns (pl.), **users** and **information,** according to the sub-parts examined in Fig. 4. In the instances checked with **due to**, the prevailing form is *due to the* preceding a singular noun (the symbol `sing´ is used to refer to the singular form). Only the Discussions sections produce a minor change, reflecting an inclination towards employing plural (pl.) or singular (sing.) nouns without articles after the key word (e.g. *due to copyright restrictions* or *due to random selection*).

The second type of approach to the description of purpose from the viewpoint subscribed, is broadened to include the colligation **to** + verbs. Careful attention must be paid to collocates placed both on the left and right of the node in order to check that the rhetorical intention of purpose is truly produced. We thus first select sub-corpus C to view similarities or differences in this use within a genre, or, in other words, to record lexical items in those contexts of research articles which register the mentioned function.

The preposition **to** appears as the top third word in the frequency order of such key sections as Methods, Results and Discussions (1,138 – 1,684 and 1,618 instances respectively). It is fourth in both Introductions and Conclusions (513 and 136 occurrences), while it occupies the sixth place in Abstracts (70). The following table displays the relevant verbs that follow the proposed colligation; they are again arranged by frequency, the first number in the brackets indicating amounts of instances signalling purpose, while the second cipher is the total number of occurrences:

[193] These lexical items, which co-occur only once, are called `hapax legomena´--see, for instance, James (1994: 23). As will be shown in the next chapters, should they offer contrastive feedback on semantic study, these occurrences cannot be entirely discarded.

Abstracts	Introduct.	Methods	Results	Discussion	Conclus.
__solve (3, 3) ___ store (3, 3)	__provide (7, 9) __ identify (3, 4)	__provide (12, 18) __ allow __ create __ ensure __ facilitate __ arrange (3, 3)	__ determine (9, 10) __ deal with __ see __ provide __ allow __ get __ create __ find __ achieve __ generate __ check (3, 6)	__ make (11, 20) __ create __ provide __ support __ facilitate __ indicate __ allow __ take ad-vantage of __ see (4, 10)	___ +vbs (5)

Figure 7: Purpose **to-** verbs across research article (RAs) sections.

The Conclusions category tabulates `hapax legomena´ (verbs occurring only once) which are not relevant in this case, but nonetheless contribute to making us aware of the presence of purpose on five occasions. The verbs written at the top of each category owe their placement to the first number alone, and not to the proportion made between both digits. Thus, for instance, **provide** in the Results column is found more often (13 times) than **determine**, but it presents one purpose occurrence less (eight instances).[194] The two semi-technical verbs registering a wider range of use in this sense, are therefore **provide** and **allow**. These include the semantic prosody of purpose, as used in key sections of research, such as introductions, methods, results and discussions.[195]

By observing the most frequent collocates to the left of the node, in addition, we can determine the lexical items that cause or motivate purpose in these texts. These are verbs like **use, design, add**, **run**, or **select**, and nouns like **aim**, **goal** and **purpose**.[196] Let us look at the

[194] The same exact occurrence happens in the Discussions column with **provide**.

[195] As exhibited in figure 6, *in order to provide* and *in order to allow* are also common; the former in textbooks and reports, while the latter in the Methods sections.

[196] Some remaining collocates that are discarded in the function of purpose include **appears**, **seen**, **need**, **enable, signal, ability, easy**, **difficult, likely, required, tend, inclined, allow, first, is, served, decision, way,**

first two in order to sample lexical associations in this case. We thus mainly seek to develop a framework of clusters:

	Abstr. (15)	Introductions (85)	Methods (276)	Results (263)	Discussions (297)
USE	.used__ (2) *	.used ____ (10) use __ ___	.used ___(30) .be used __ .are used __ .was used__ .can be used _ used __create using __ __ _ used__ obtain used_ arrange	.used __(32) was used __ used __ create used__ generate	.used__ (32) be used __ are used__ was used__ be used__ create
DESIGN	.design ___ (2)*		.designed__ (7) .designed __ allow	.designed __ (12) .designed__ _deal with _determine	.designed_ (13) .be designed __ .designed _facilitate _take ad-vantage of _support .a system designed__

Figure 8: **Use** and **design** as purpose triggering verbs in RAs, followed by most frequent forms.

The noted numbers in each division title refer to the total number of instances that entering all possible word-forms of the two nodes yields in the concordancer (e.g. **use/using/used/design/designed/designing**).[197] In the case of conclusions, the only located forms were *use to* (one instance) and *used to* (one example). In abstracts, the asterisk * is added in order to mark the low number of these occurrences (below the set amount of three). As observed in three of the categories, the expression *used to create* denotes the greatest margin of use. Clusters including more than one gap contain other words, such as *using* __ __

---

opportunities, enough, how, free, led, try, introduction, asked, compared, applied, expected, found, limited, decided, links, necessary, order, and **able**.

__ in the Methods column, referring to five instances found of expressions like *using the Web to*.

In textbooks and reports, such lexical data in the function of purpose can also be contrasted and examined so that the registerial and discoursive values of the sub-corpora may be accounted for in our study on representativeness. We thus obtain a small but supportive sample of word patterning for the aim of developing a lexical framework according to genre distinction and similarity. The verbs chosen are five, three of them quite characteristic of purpose in RA (see Figs. 5 and 6): **provide**, and **create,** while two are used with a lower range to express this function: **support** and **find;** finally, **make** is not used at all in this sense.

The first, **to** + **provide**, often receives this semantic prosody focus in textbooks. Out of the 105 instances found, 34 reflect such an employment –e.g. *the purpose is to provide*. In reports, the treatment as purpose of *to provide* is much lower in the 120 lines found–just a few cases, like the string *the goal is to provide--*. More common instead are phraseological constructions such as *are required to provide* (6 cases) and *is intended to provide* (5). In this sub-corpus B (Reports), the node **to create**, in turn, receives a greater emphasis of use from the perspective sought: 40 instances are identified among 55 occurrences of the word in which purpose is expressed. *Used to create* co-occurs four times and *in order to create* three, while there are other constructions such as *purpose is to create*, *proposals to create* and *stimulus to create* repeated twice. In textbooks (sub-corpus A), however, there are fewer cases in the 50 appearances of *to create* that denote the writer's search for aim: 15 lines, among which *used to create* appears.

The sense of purpose of **support**, mainly located in the Discussions parts of sub-corpus C, finds ample application in both TXs and RP. In 85 lines, the former registers such

[197] This is made possible in *WordSmith* by adding asterisks to the words (e.g. **design***). In the case of **use**, in contrast, using slashes aligned to the right with all the possible word-forms is more desirable, as unwanted

statements as *provide a mechanism to support* on nine occasions, *be provided to support* on five, and *interface to support* (4). Other structures are *technology to support*, *protocol to support, used to support*, *maintain __ to support*, *purpose is to support*, *goal is to support* and *To support…*(three each). In turn, the 66 rows of concordanced text displayed for RPs reflect a lower number of clusters, but they are enough to reveal significance in the function under study: *costs to support*, *research to support, aim to support*, *provide _ _ to support*, and *contain __ __ to support* (repeated 3 times each).

Given the co-texts of the other low-range item, **find** –encountered in five occasions out of nine as a purpose-verb in the Results section of RAs--, a comparison based on its absence as such in the rest of research articles parts, may be made with RPs and TXs. None of these genres provides any significant instances (three at the most in the case of textbooks). Other phrases prevail instead, such as *Adj. + to find*. Finally, the case of the total absence of a verb in RAs as a key item in the sense underlined, serves as contributing evidence for the assertion that different genres exert a varying influence on the word-forms employed. Thus, **make**, with 107 occurrences in sub-corpus A (Textbooks), is widely utilized in statements signalling aim: *in order to make* (six times), *work to make*, *discussion to make*, *using __ to make* and *goal is to make* (four each). Reports, like RAs, in contrast, with 98 lines concordanced, barely offer any cases (merely three instances).

Given the findings on these lexical items, as a result, the following table may be drawn, indicating the presence (+) or absence (-) of purpose in the verbs according to the genre where they are used in the infinitive form:

---

results such as **users** may originate otherwise.

	Sub-corpus C (Research articles)	Sub-corpus A (Text-books)	Sub-corpus B (Reports)
To provide	+	+/ -	-
To create	+	+/ -	+
To support	+ (Discussion)	+	+
To find	+ (Results)	-	-
To make	-	+	-

Figure 9: Representation of purpose verbs in genres.

To support is the only example more recurrently employed to express purpose in all genres; yet, as has been shown, in sub-corpus C, its frequency is only significant in the Discussions sections. **To create,** in turn, is also significantly used in purpose statements across our corpus. Only in textbooks, its distribution is not as high. **To provide**, in turn, is only strong as a purpose indicator in the research articles, and only `half weight´ in textbooks, while it is entirely missed with such a function in RPs.

These findings suggest that the contrasted features contribute to make genre generalizations and text type formalities for the sake of establishing a testing analysis of corpora content for lexical work. Such explorations into the nature of lexical variation across corpora should make the ground of operation readily accessible for pinpointing distinction of lexical patterns according to stylistic variable profiles. Descriptions and characterization

according to lexical employment in Information Science and Technology reading material may thus be adequately sifted.[198]

The next and last issue to be described is the range of combinations which can be collected from modality in our sub-corpora. The observation of this feature wields the determination of degrees of hedging in genres, although that is not our main concern from the viewpoint explored. As above, we seek to provide a contrast with the three reduced bodies of texts that enables the arrangement of significant lexical items used as traits of genre distinction. The environment of use is thus considered representative in terms of key word behavior, which appears in the texts under the influence of context. The table below (Fig. 8) gathers in rows and columns six modals widely used in our corpus. In addition, another factor for assessing genre identity, voice, is also considered with modality. The letters `act´ or `pas´ at the top of each set indicate the prevailing voice which is used in the clusters, while its percentage is offered in brackets. The first numbers refer to the total frequency of instances found.

MODALS	Textbooks	Reports	Research articles
CAN	(1029) **pas** (82 %) that (52) __ be used (63) to (30) __ be viewed as __ be applied to __ be implemented in __ be made __ be shared among __ be achieved __ call to __ accept	(1225) **pas** (81 %) that (37) __ be used (77) to (36) __ be reused __ be made the author ___ __ be useful to you __ also __ have __ be defined __ be invoked __ be read __ be transmitted __ be easily	(884) **pas** (78%) __ be seen (34) __ be used __ be made __ lead to __ be found __ be viewed as __ be defined __ be applied __ be chosen __ be created __ have __ perform __ present

[198] The results can be tested further –in favor or against-- by checking larger bodies of text and other verbs in purpose statements (in greater depth and detail).

	__ experience __ be found __ be performed __ be exchanged (5)	determined __ be made available __ be created __ be found __ be linked with problems __ arise (4)	__ capture __ deal with __ display __ detect (3)
COULD	(328) **pas** (53%) it __ be (19) __ be used (17) to (6) __ be seen as __ be accomplished __ be useful __ be a message __ be a model __ be a process level __ cause (3)	(271) **pas** (77%) __ be used (13) to (9) __ be accomodated __ be decreased dramatically __ be altered __ be spaced __ communicate with __ provide __ decode __ put __operate __ sail (3)	(226) **pas** (51%) __ be seen (11) as (10) __ could have been + p.p. __ proceed by __ be found __ be made __ be applied to __ choose __ consist of __ decide __ think __ end up __ limit __ mean __ prove (3)

MAY	(925) **pas** (52%) it __ be (63) __ be used (61) to (36) __ be specified __ be useful __ specify __ need to __ wish to __ be necessary __ be appropriate __ be desirable __ be interested in __ be available to __ include __ want __ be defined __ be found __ be tagged __ lead to __ choose to __ be difficult to (4)	(674) **act** (55%) __not be (29) __ or __ not + V / be +Adj (28) __ not have + p.p. __ be used __ recover __ be delivered user ___ have a __ arrive out of __ continue to __ have completed __ be able to __ have been filmed __ be provided __ be reproduced __ include the __ use this __ be controlled by (3)	(336) **act** (66%) __ not be + adj. (9) __ also be + adj. __ include links to __ exist between __ have little __ occur __ be important __ be obtained from __ be called __ be involved __ be limited __ seem (3)
MIGHT	(217) **act** (53%) it __be (7) __ need to (5) __ have been as it ___ seem __ be divided __ look like __ be tagged (3)	(195) **act** (78%) __ happen (6) __ be in __ arise (as a) __ be rendered completely __ contain __ delay __ prevent (3)	(99) **pas** (54%) __ be used (5) __ be + N the changes __ be + p.p. __ not have + p.p. (3)

MUST	(238) **pas** (53%) it __must be (12) digital archives __ __ be able to __ be made __ be related to __ be considered __ be given __ be specified __ address __ receive __ be supplied __ be remebered (3)	(393) **act** (56%) there __ be (18) __ have + p.p. (17) __ ensure that __ determine __ be called __ know the contractor __ __ have been __ be available __ be at least # __ instantiate __ be aware __ be considered __ be made __ be able to __ be a superset __ be implemented to __ be included in __ guarantee __ be known by __ be supported as __ support __ pass __ be taken __ equal __ be passed __ make __ equal (3)	(106) **pas** (60%) __ be noted (6) __ include reference to __ be employed __ be used __ be evaluated (3)

SHOULD	(562) **act** (61%) __ the interface (114) provide (51) __ be used __ be optional __ be able to __ make __ there be __ document __ consider the __ be defined in __ be collected __ be represented __ be supported __ be addressed __ define any __ define mapping (3)	(268) **pas** (58%) __ be used (14) __ have been __ be movile __ include __ be avoided __ be routed that ___ have + p.p. __ be amended in __ be communicated by __ be noted that __ recognized by __ constitute __ see __ pass (3)	(199) **act** (66 %) __ be able to (13) __ be + Adj future research __ +V data __ be + p.p. __ be designed to __ be spent in __ degrade __ mind the __ consider the __ search the (3)

Figure 10: Modality in our corpus. Act = active / Pas = passive.

Where N, p.p., Adj or V are included after a + sign, noun phrases, past participles, adjectives or verbs respectively follow the given node. The words are not explicitly listed due to their low individual frequencies –below three. For instance, in the case of **may** in reports, the structure *may or may not involve* occurs twice –represented by `+ V´. In addtion, a phrase like *may or may not be difficult* is also included here (`+ be + adj´).

The modal verbal data allow remarks to be made on the lexical comparisons and contrasts emerging among genres.[199] An example in identifying similarity is given by the verb **use**, which rather frequently appears in the passive voice with nodes such as **can** in all genres, **could, may** and **should** in TXs and RPs, and **might** in RAs**.** Observations based on difference, in contrast, can also be garnered. Within a same genre --reports, for instance--, the key word **provide** is often found in the active with **could**; however, the tendency is for

[199] Additional modal verbs and their significant lexical collocates can be checked, such as the use of **will** and **would**. They corroborate the fact that articles contain a more impersonal and formal style. For instance, *it would be* –most frequent in the genre mentioned—contrasts with *you will be* in reports. Other collocations are: *will be*

provide to be employed as a passive form when co-occurring with **may** in sub-corpus B: *could provide* vs. *may be provided*. Another type of examination is applied to how two nodes are similar or different according to the lexical company they are likely to keep. In this sense, **must** differs from **should** in textbooks**,** for example, in terms of the absence or presence of the passive feature with particular verbs such as **make**, **consider**, **give**, **provide**, **supply**: e.g. *must be made* and *must be considered* vs. *should make* and *should consider*, or *should provide* vs. *must be provided.*

In the case of another modal (**can**), the verbs appear much more often in passive statements (see figure 10). In contrast, **should** is employed actively in quite frequent circumstances, such as in questioning throughout the textbook on Optical / radio communication with the cluster *should the interface*. Within this same genre of textbooks, however, assertions on register uniformity resulting from emerging common patterns, may also be drawn: the example of **must** and **should** with *must be able to* and *should be able to* serve as cases in point.

Finally, to conclude this section on lexical features that describe our sub-corpora, modality and voice can be analyzed within sub-corpus C (RAs) alone to determine the influence, if any, of different sub-sections on lexis. Just to give a brief illustration of this fact, only the last modal offered, **should**, is given this focus:

---

able to, *will need to*, *would have to* and *would like to thank* in articles. *This issue will be used* and *would be able to* in books. Finally, *will need to, will be provided* and *would have been* are significant in reports.

Introduct. (32) **act**. (77%)	Method (52) **pas** (66%)	Results (56) **act** (86%)	Discussions (52) **pas** (54%)	Conclus. (7) **act** (57%)
___ degrade (3)	___ be +p.p. (15) ___ be possible (4)	__ be able to (11) __ consider __ mind the __ search (3)	Future research __ + V (6) __ be designed to __ be easy to __ be spent in (3)	___ + V (4) __ be + p.p. (3)

Figure 11: **Should** across RA sections.

The `Abstract´sections contain no instances of the node-word. Conclusions, only seven. Methods, in contrast, is the context of RAs where more sentences are constructed in the passive with **should**, according to our sub-corpus: **updated**, **captured**, **displayed** are three verbs placed after *should be* in this section (only one occurrence each according to our data). This may contribute to signalling the more informative tenor of these sections – methods-- in developing and presenting proposals for study, but not in reaching assumptions and conclusions more firmly in the research. In contrast, a higher degree of the active voice is effected in the Discussion and Results sections. In the former, in fact, the expression *future research should* is followed by verbs such as *aim at*, *try to,* and *be concerned with* (each appearing twice), indicating a distinctive move in this section for the inclusion of prospective research and references. Furthermore, under Results, the pronoun **they** is actively employed, preceding *should be able to* on five occasions; other subjects, like **workstation** and **machine** are repeated twice before this form. In addition, **consider**, **mind** and **search** are also examples of significant items co-occurring with **should** in the active voice (see figure 11). They all seem to convey the sense of engagement characterizing the development of results in research papers.

Given thus these figures in the tables, we postulate with an introductory note to the more detailed lexical study offered in chapter 3: Collocations and other lexico-grammatical

combinations serve to identify key features of register dimensions, text types and genres.[200] These are therefore underlying markers that enable us to assess whether given stretches of language may be representative for the kind of analysis that we should elaborate from a stylistic plane. In this respect, the evaluation of `we-statements´, purpose, modality and passive voice has allowed the recognition of key vocabulary that recurs to a certain degree across or within sets of text samples. For our purposes of lexical selection, the corpus is thus representative, as the similarities and differences of use for a given item provide us with feedback for lexical description from differing ranges of use.

In the next section, the focus shifts to our main concern of dealing with common core lexis according to various levels or stages in the analysis. The next step is to define the criteria for the manner by which the lexis will be categorized and assessed according to the texts obtained. Then, a detailed explanation of the particular methodology utilized to sort and make the data available must be provided in order to understand our conception of lexical significance.

II. WORD TAXONOMY

Two main goals are pursued in this section: to apply academic definitions and classifications of lexical items in EST for our lexical analysis purposes, and, secondly, to review the texts of our corpus in order to identify those where certain aspects of word use must be foreseen and noted prior to the analysis.[201]

[200] See Register and Discourse sections in chapter 1 (e.g. Sinclair, 1966 [Register-features], and Selinker & Trimble, 1974 [Discourse features]).

[201] The focus is thus placed on coordinating co-text and context effectively (cf. Tribble, 1998; Aston, 1999; Scott, 1999, and others above). This means that word distinction is solely based on how the nodes combine or associate with other terms according to the different textual settings –e.g. text types, genres and subjects (see chapter 3).

II.1. CLASSIFICATION BASED ON LEARNING.

The arrangement and distribution of our lexical items will be made according to four main categories of semi-technical or sub-technical vocabulary –as determined by ESP and lexicology / lexicography experts such as those cited in chapter 1: Academic, subject, argumentative and procedural lexis.[202]

In the first option, academic, we include word patterns which are common to all disciplines, as defined by traditional sub-technical EST lexis.[203] These are core words of high frequency, containing an evenness of range and coverage of text.[204] In this case, we refer to multi-word terms that may be more or less restricted, depending on their use `in specialized senses´ (Howarth, 1996: 68). Thus, acknowledgement is made regarding the fact that, within this semi-technical set of academic collocations, there may be free or restricted patterns, according to how the elements are recurrently used.[205] The difference, as Howarth (1996: 38) states, is that

> [in] free collocations (…) the meaning of each free collocation as a whole is quite transparent, easily derivable from the juxtaposition of the elements in a recognizable syntactic pattern [while] restricted collocations do not form single semantic units and their total meaning is derivable, though often indirectly, from the meanings of the component parts (…)[206]

---

[202] Some authors, listed under each one of the four lexical divisions made, are: Academic: Stotsky (1983), Thurstun (1996); subject: Cowie (1997), Carter (1997); argumentative: Hutchinson & Waters (1980), Francis (1986); procedural: Higgins (1967), Winter (1977). (See Lexical Focus in chapter 1).

[203] See, for instance, Inman (1978), Ewer (1983), Yang (1986). (ESP and Lexical Focus in chapter 1).

[204] As Mackey and Savard state (1967: 83), their measurement is made to check these factors of the high number of occurrences and distribution over texts.

[205] The preoccupation has often interested scholars at the level where patterns and collocations are not fixed (e.g. idioms) but rather free and semi-restricted; for the foreign language learner, this is where the greatest problems arise (e.g. ***long distance*** vs. ***large distance***, ***little*** vs. ***small light***, etc): see Alexander (1984), or Moon (1984, 1997, 1998 a). The latter author also asserts that 40 % of the expressions are semi-restricted in the Hector corpus (18 million words) (Moon, 1998 b: 92). Other works on the collocation of these academic lexical items are Howarth (1995) and Ozkaya (1996), whereas Gitsaki (1996), Fielding (1997), De Cock (1998) and De Cock et al. (1998), analyze the use made of these combinations by non-native students of English.

[206] Howarth's exemplification (1998: 164) of collocation types is quite illustrative: *blow a trumpet* (free), *blow a fuse* (restricted), *blow your own trumpet* (figurative idiom), *blow the gaff* (pure idiom).

In our case, these degrees of restriction are examined by checking lexical strings according to both frequency and range. The higher the frequency and the lower the range of texts in which they appear, the more restricted these collocations tend to be. Thus, for instance, the multi-word unit *manage information systems*, is considered more restricted than, for instance, *exchange information with*. The former includes `a specialized noun determined by verbs (limited set)´ (Howarth, 1996: 43), while the latter holds `possible substitutes of both elements´ (ibid.).

Both are encountered in our corpus quite frequently. This has to do with the fact that our search is centered on semi-technical lexis (words that offer a greater degree of difficulty in language acquisition), and thus, expressions such as the ones seen above tend to contain fewer highly technical items. In addition, as Aisenstadt claims (1981: 56), such free collocations often encompass `productive and widespread verbal patterns´ such as V+Prep+N, which hold colligations as main elements of combination –e.g. *have a significant effect on*, or *jump to a conclusion*.[207] Other examples of this type in our corpus are: *requirements from using the*, *writing for the web*, *spend time working on,* etc.

As a result, the contrast between restricted and free academic lexical items substantially depends on the position of both frequency and range parameters, by which words may receive either specified or more applicable senses across texts (see the word list in appendix 4). In this sense, we come to set a cut-off or definition point by which restricted are distinguished from free lexical combinations. Such a limiting score is established at a 0.3 estimate. This means that every ten concordance lines, there should be at least three texts involved (3 / 10 = 0.3). Such a figure or higher makes a given collocation or cluster a free combination. In contrast, lower numbers lead us to classify those items as restricted to certain

[207] See also Mitchell (1971) (Collocation section and footnote 125 in chapter 1).

texts (e.g. 2 / 10 – ten concordance lines from two sources--, 4 / 20 – twenty lines from four texts, and so on). In this way, a multi-word unit such as *network object runtime*, repeating 48 times in our corpus, is considered restricted since it appears in only one source, a Telecommunications report, whereas the collocation *network administrator*, for instance, occurring only twice, is viewed as free due to its appearance in two different texts: a Computer Science research article and a Documentation technical report (see appendix 5). Furthermore, attention is likewise paid to the fact that a frequent item occurring widely enough (beyond a set number of 20 sources) works as a free combination, since our corpus contains only 96 texts –20 would be measured as a significant amount in our corpus. In this sense, we judge that more than 100 lines of text for a given item within 20 sources or more make such a lexical element a free construction.

The second set of words, subject lexis, unlike the first one mentioned, has a lower range of occurrences in terms of both the subjects and topics in which they appear. Thus, their narrower context of use is mainly caused by specialization and application of subject matter. Significance of patterning can, nevertheless, be achieved too: it `can be established and described within quite small texts´ (Howarth, 1996: 72). In this respect, the `language community's relatively stable shared phraseological competence´ plays a crucial role (ibid.). It can be referred to as the characteristic lexis of a given `sub-discipline´, as Stubbs calls them (1996: 19). In these specialized fields, such as Audio-visual technologies and equipment --the **D1** category in the arrangement of core subjects above--, for instance, the observation of specific contexts yields `stylistic expectations of the academic community´(Cowie, 1998: 12). In this sense, as a result, the fact that these words encompass two or more disciplines in the given subject takes on importance so that they can also be considered common core lexical items according to a single context of use. Some examples in our corpus include *delivered by*

broadcast technology and *broadcast over the network* for the subject mentioned –Audio-visual technology—across reports in both Optical / radio communications and Audio-visual communication.[208]

Regarding the third type of vocabulary, argumentative lexis, the focus is on those lexical items which capture the essence of the discourse flow, i.e. which words are used to signal the author's strategy to condense information and how (e.g. `A-nouns´ –Francis, 1986).[209] Some of our examples range from clusters such as *this issue will*, *a device that*, and *an entity reference* to delexical or indexical verbal items like *make comparisons between* and *take advantage of*. By these, the writer chooses to indicate what is done by means of research in sub-corpus C texts –e.g. compare two concepts, profit by applying a given software tool, etc. The first set displayed, in contrast, gathers in one word (e.g. **device**) the whole reference to the product being described, subject analyzed, or concept defined in textbooks (the `themes´, in Discourse Analysis).

A similar group of lexis characterizes the fourth category distinguished in our analysis. The procedural lexical items are here specified as those open or content words which substitute closed system elements (see Winter, 1977, 1978, 1982; Roe, 1977). Like argumentative items above, these signposts mark coherence and cohesion in the texts to facilitate discourse flow comprehension. Unlike them, they provide this function in an overt and explicit manner. In sub-corpus B (RPs), this group is represented by, among many others, *a further simplification*, *add additional methods*, and *also provides a*, which are expressions operating as markers for the addition of information in the text. They serve as rich input for teaching English discourse devices in terms of prosody, lexis and syntax. As shall be seen in

[208] See more about subject lexis in the second part of this section.

[209] These word distribution patterns are often thematic, affected by the choice and arrangement of grammatical structures (Tarantino, 1991: 31) –see Coherence and Cohesion in chapter 1.

chapter 3, their treatment is thus examined within the restriction of text type samples, represented by discourse rhetorical functions and moves, which provide the writer's strategy in signalling direction in EST discourse (cf. Trimble, 1985).[210]

The last two sets of lexical items, argumentative and procedural, as can be noted, differ substantially from the first two, academic and subject. This is done on purpose to progressively include common core lexis that is closer in its use to discourse and grammar features.[211] In fact, a fifth compilation of words is discussed in chapter 3 comprising recurrent colligations that either serve as connectors for discourse cohesion or function to collocate with key academic vocabulary (see first set of words above). The distinction will thus be made, for example, in the case of the conjunction **as** in composite units of the sort of *as well as* and *as part of*, the former signalling comparison as a rhetorical function of discourse, while the latter contains the node **as** in a phraseological frame which typically occurs in the body of texts chosen.

A final comment must be made on the range of texts where each division of lexical items is based. As has been mentioned, this is done across all genres in the case of the first set of academic items, since these common core multi-word terms and clusters should receive an ample treatment to represent, as widely as possible, the academic area of Information Science and Technology in textbooks, reports and research articles. Secondly, subject lexis is more narrowly focused on given fields or topics of interest highlighted in the curricula, as has also been explained. Hence, there may be only one genre encompassed, or, on the other hand, more, depending on which texts in our corpus deal with that specific subject. What must

[210] This fact will be examined in greater depth in chapter 4. Anping (1997) shows in his Ph. D. study that corpus-based research (in his case, a 91,802 sub-corpus from the London-Lund corpus) is highly useful as a teaching aid to acquire competence on discourse devices application as encouraged by simultaneous speech.

[211] We hereby agree with lexical analysts like James (1994: 56) when he states that a corpus of specialized readings is characterized by `the occurrence not simply of technical or sub-technical lexis, but even of some of the commonest function words in the language (…) affected by the style of writing´.

always be present is a coincidence of a minimum number of two Information studies included for the property of common coreness to be maintained.

The other two lexical categories, argumentative and procedural, however, require, from our viewpoint, that only one genre at a time be selected so that these words point out particular traits in single communicative events and stylistic settings (Swales, 1990: 58). Needless to say, for these two levels of lexical analysis, the aim is to find co-occurrences most characteristic of a given text type, and yet, the findings are also assessed as common core clusters, as the texts are always applicable to more than one Information Science and Technology discipline. The same is done in the case of grammar and discourse words, since our purpose is to offer a view of how function words are differently employed in colligations with key items (e.g. **about** with **information** in *information about* as a grammar combination, or **also** with **and** in *and also,* as a discourse marking colligation).

As a result, not only genres are distinguished but also types according to stylistic concerns. The latter include sections within research articles, but other samples such as classification or definition excerpts are collected for such aims.

II.2. WORD DISCRIMINATION IN COMMON CORE ANALYSIS.

The range of subjects and topics presented and developed by the different texts used must be considered so that keywords signalling the `aboutness´ of texts are specified (Scott, 1997: 234). From this persepctive, the aim is to distinguish lexical groups according to subject fields and topics, and even the occurrence of words which provide little significance in terms of common coreness on a given topic. Thus, we can determine not only their restricted use, but also point out the markedly thematic lexis that is non-core in the sub-corpora. These are words with a high frequency in a text but insignificant lexical range in the

corpus; according to Liu and Nesi (1999: 142), `those words on our list that had a low distribution and a frequency of five or more occurrences´. As a result, some examples are given in order to reflect that a lexical list based on single text frequency does not meet our criteria.

We can discriminate words, in this respect, by using the **Keyword** function in *WordSmith tools*. The procedure is simple enough: to compare the given source –textbook, report or article (target text)—where the occurrence appears, with the bulk of texts in the larger, general (reference) corpus. The analysis yields degrees of `keyness´. Keywords in the text sample are thus calculated by means of statistical tests provided by WordSmith tools. These are the chi-square test of significance and the log likelihood test, which cross-tabulate the data obtained from the following: the frequency of the item in the small corpus word list, the number of running words in the small word list, the word frequency in the reference corpus, and the number of running words in the reference corpus (Scott, 1996). In such a procedure,

> a word will get into the keyword listing if it is unusually frequent (or unusually infrequent) in comparison with what one would expect on the basis of the larger wordlist. Unusually infrequent words are called "negative keywords" and appear at the very end of your listing, in a different colour (Scott, 1996).

Our concern obviously involves the occurrence of positive keywords in this respect: words that appear as distinctive in the sample corpus, indicating the extent to which such a word constitutes a key element in a given text. For instance, the node **genres** appears 65 times in the research article "Designing Genres for the New Media", being the first in the keyword list, given a score of 217.3 by the WordSmith **Keyword** function. This degree of `key-ness´ is `outstanding´ in Scott's terms (1996), since it is much higher than 25.0, the set value at which a lexical element maintains a minimum percentage in the statistical tests to be considered a keyword (Scott, 1996). These high score words tend to indicate the `aboutness´ of the text,

what it is mainly about (in this case **genres** denotes the main theme). Below that point of 25 in the keyness scale, the item would be regarded as negative in that text sample.

Thus, for instance, in the same reading, **commodities** is used eight times, and yet also considered a keyword, occupying the last position in the list with a percentage of 26.7 in the keyness measurement; in this respect, it is also representative in the text, but much less so than the previous item analyzed, **genres**, in the source "Designing Genres for the New Media".

Such a way of differentiating thematic lexis, in the case under study, equally allows us to account for lexical irrelevance in terms of their low range. From this perspective, we can discriminate key words that are proper names, abbreviations or acronyms, and latin root words. The first set includes those references to persons or places (toponyms) often cited in a given document. An example is **Brownrigg**, appearing 16 times in the textbook *Preserving Digital Information*. Brownrigg is a pioneering researcher in the area of networking library documentation according to the source, but just a name for our linguistic concerns. In turn, an example of a recurring word naming a place, is **bones** in the textbook *Writing for the Web*, referring to an Internet address quite often repeated in this text --*bones.med.ohio*—where medical information is supplied.

There are also several instances of words that stand for concepts defined in a single text, such as **M.4**, repeated 117 times in the report "Introduction to M.4 and M.4 VB", concerning powerful software to create information knowledge-based systems. Another one is **O.D.A.**, short for *Office Document Architecture*, recurring with **T.E.I.** (Text Encoding Initiative) in the textbook *Electronic Text*.

With acronyms, however, we must be careful, since in technical settings such as our own, these may function as meaningful lexical items denoting important concepts and

developments in Information Science and Technology. We pay attention to this fact by relying on our observation of the item in the defined corpora of use. For instance, in the case of the two mentioned, **O.D.A.** and **T.E.I.** above, while the latter is used more extensively throughout our corpus as a subject / topic keyword – see chapter 3--, the first acronym referred, **O**.**D**.**A**. is particularized in only one source, thereby being discarded as a core item in our analysis.

In the last division made, a large number of occurrences registering specific use determined by topic in a single text are found to have latin roots, quite closely developed as Spanish cognates. In one article, an example is *quantized coefficients*, often repeated in "Lossless Image Compression Using the Discrete Cosine Transform". As can be seen, these lexical units are also highly technical and characteristic in such a type of textual environment; other examples are **polynomials**, **polymorphism**, **psychometric**, etc. Multi-word terms such as *radiolocation radars*, *metereological radars*, or *radionavigation services* are easily recognized in these settings; yet, their restricted use to single contexts (e.g. in this report, "Procedures for Analyzing Compatibility using Radiodetermination and Metereological Radars") makes them less applicable or relevant for our common core purposes.

Yet, regardless of the observed appearance and use of the items, a final intuition-based decision should be made in terms of teaching implications, as there may be technical elements that appear with a low frequency but offer important feedback concept-wise – e.g. hapax legomena which appear in all three genres.

III. LEXICAL STORAGE

The procedure for dealing with word selection for the arrangement and placement of lexical items under the categories given above is described in this section. The aim is thus two-fold according to the needs of common core lexis distribution and availability for language learning and vocabulary acquisition: first, to establish a framework by which lexis can be seen as relevant or not for selection and analysis, and secondly, to analyze how the five categories of words –academic, subject, argumentative, procedural and discourse / grammar—keep different organizations of the lexical collocates, colligations, patterns and clusters.

III.1. LEXICAL FRAMEWORK IN WORD DESCRIPTION.

The notion underlying our choice of lexical items is the same as the generally accepted premise of vocabulary learning by many scholars: that `in language learning situations people build up a passive vocabulary before they can actively use it´ (Aitchison, 1987: 192).[212] Thus, both phonology and the written medium play important roles at this stage. The goal is to work hence towards a `compromise between the requirements of production and those of comprehension´ (ibid., 193). Word storage, i.e. organization of lexis, is essential in the process as a suitable means of arrangement and disposal of the input. In dictionary making, it is the lexicographic work that combines, as Battenburg states (1991: 7), `art´ --freedom to describe language and arrange entry content-- and `science´ –dealing with the empirical study of form and structure. It is also a method, as Stubbs puts it (1997: 252),

> which can be used to compare and contrast the characteristic uses of (…) words (…) how such comparative profiles are best presented will depend on the purpose: say, a dictionary for learners themselves, information for writers of teaching materials, or comparative information for a stylistic analysis.

[212] Other relevant authors who look at the key of activating vocabulary through semantic networks are Schmitt & McCarthy (1998), Lewis (1997, 1993), Willis (1990), Sinclair & Renouf (1988), etc.

In this respect, handling and managing lexical items involves, from our viewpoint, not just a fixed dictionary `with a set amount of information about each word, but an active system in which new links are perpetually being formed´ to meet learning needs (ibid., 162).[213] In EIST, indeed, as technology and concepts rapidly evolve, new lexical items are swiftly modified and re-introduced. In addition, the notion of purpose, as Marcos Marín (1994: 662) points out, must be taken into account to signal the specific usage of the multi-word terms:

> Los términos se construyen para establecer una distinción que va más allá de las que permite el uso general de la lengua. La ventaja evidente es que las expresiones lingüísticas de cualquier clase pueden pasar a convertirse en términos cuando sea necesario, basta aplicarles esa finalidad determinada o, como también se dice, situarlas en un concreto campo de conocimiento (…)

From this perspective, lexical items may contain a great deal of specialization,[214] in which concepts are developed by words and relations (ibid., 670). For example, in Internet, the transfer of information, i.e. getting a document into one's computer, is expressed by adding prepositions before verbs, seldom done before in standard British English; yet, who can resist nowadays the invitation to adopting such elements as *download the information* or *upload the page*?

These types of multi-word units lead us to seek significance of lexical co-occurrence as based on Corpus Linguistics and ESP scholarship such as that described in sub-sections 2 and 3 of `Collocation´in chapter 1. In addition, awareness of prompting word elicitation

[213] This view has actually been a constant concern in lexicography; see, for instance, Samuel Johnson's statement in the preface to his renowned dictionary of 1755: `No dictionary of a living tongue can ever be perfect, since while it is hastening to publication, some words are budding, and some fading away´. Or similar recent views, such as Országh's (1969), as explained in Battenburg (1991: 68): `It is not a static two-dimensional listing of words that is needed but a dynamic work presenting the vocabulary in motion´.

[214] In terminology studies, technical items are those `special lexical items in a particular discipline or subject matter´(Richards et al., 1992: 212). In our work, highly technical lexis is considered within the realm of terminology; its use as common core vocabulary is thus rare and hardly relevant, as has been mentioned above and will be observed in the following chapters.

becomes essential from this perspective. The lexicon is thus seen as `a sort of connected graph, with lexical items at the nodes with paths from one item to the other´(Fodor, 1983: 80). Considering subject matter and topic is relevant, in this approach, so that words relating to the same issue are stored together. Learners can consequently benefit from this sort of arrangement of the range of word patterns and clusters that are likely to appear according to the text subjects that they read and work with. Attention to text genres and types is also taken into account in this sense. In agreement with Roldán Riejos (1998: 37), as a result, our students will thus be able to apply their understanding of language mechanisms (e.g. lexical item associations) to different tasks requiring their expanding competence: `writing, reading, or even translating reports, letters, publications, or manuals´. In this respect, for lexical entries, we are more interested in revising sense and meaning through direct sampling of key words in context than with actually translating: `Quotations are as important as definitions´ (Battenburg, 1991: 13), or as Bod (1998: 1) illustrates by using Wittgenstein' claim, `Es gibt nur die Beispiele´(*there are only examples*).[215]

Therefore, lexical items, as understood in our study from viewpoints as consolidated as Sinclair's and Halliday's (1966),[216] must be stored by considering the following parameters:

1. The collocational strength of words as measured by frequency and range.
2. The use of possible lexical combinations as determined by context.

[215] Content is thus to be `put back into the (mere) language skills´, as Kramsch puts it (1997: 48) according to the CBI (Content-Based Instruction) perspective. Like hers, Widdowson (1993) and Swales' (1997) views do not make `any simple division between language and content´(Swales, 1997: 216). The student's success in learning has much to do with `the restriction to a single genre (…) allow[ing] exposure to texts (…) to be a coherent rather than disruptive set of activities´(ibid., 227).

[216] To review these already discussed ideas and notions about what lexis is and how it should be dealt with, Sinclair (1966) is a suitable example: Lexis, `running parallel to grammar (…) describes the tendencies of items to collocate with each other´ (Sinclair, 1966: 411). A lexical item is `a formal item (at least one morpheme long) whose pattern of occurrence can be described in terms of a uniquely ordered series of other lexical items occurring in its environment´(ibid., 412). (…)`environment is the extent of text which is relevant in the

3. The learner's needs in vocabulary acquisition for his / her studies.

4. The particular description of EIST in our institution.

As can be seen, while the first two are more narrowly related to Corpus Linguistics approaches, items 3 and 4 reflect the concerns of target and learning situation needs analysis respectively. These latter two are more widely dealt with in chapter 4, whereas points 1 and 2 may be checked more closely in the analysis of the results (chapter 3).

Our corpus is taken as a reference for measurement in statements 1 and 2. This means that our focus is, as Leech asserts, on lexical significance as given not by the size of the corpus –a bit more than 850,000 words--, but by its representativeness as being `domain-specific´ (Leech, 1991: 11).[217] M.I. (Mutual Information) scores are thus applied here, but as McEnery and Wilson warn (1996: 72), not as the sole key to determine which words must be stored and how. As Stubbs states (1996: 6), some word meanings `are brought to the text by readers or listeners according to their specialist knowledge´, so, the relationship of lexical items and concepts developed in the given stretch of language or context tends to be accounted for according to language and conceptual needs.[218]

Thus, for instance, a cluster such as *html writing tools*, found only twice in sub-corpus A (Textbooks), appears in two section headings of a textbook on electronic cataloging. It is then never repeated again in the body of text, but its importance as a restricted multi-word term is derived from its contextual position as a significant topic of discussion in the field of Library science. The texts form, from the perspective given, the basis of this collocation selection in both absolute and relative terms: frequency, range and `familiarity with the

---

description of an item´(ibid.) (…) `grammatical information can be used in lexical description´(ibid., 425), etc. In this sense, even today, as Stubbs says (1996: 47), `a full lexico-grammar of English has yet to be written´.

[217] In addition, we also agree with Owen (1993: 185) when he says that `the grammarian and language teacher need the corpus as servant, not as master´.

particular register of the LSP variety' (Bergenholtz & Tarp, 1995: 124).[219] Lemmata are stored under the premises of being `LSP terms, expressions regularly frequent and with a minimum frequency in LSP´ (ibid., 103). A word is considered a lemma when any of those conditions are present.

Thus, in the aforementioned example, the noun **tools** (figure 12) is searched as the lexeme / lemma, due to its tendency to be used in the plural –rather than in the singular form--, given the frequency, range and coverage of such an item in the texts. In the final arrangement of the word, as a result, it appears in bold –this typographic choice serves to point out that the key word is, indeed, a lemma (Bergenholtz & Tarp, 1995: 226). In the article describing the item, lexical information is supplied in italics in the case of clusters –which should be organized, as Gregory states (1975: 1), `in terms of their frequency of co-occurrence within the span centred around [the node]´--, whereas the word-forms of the lemma (occurring less frequently) are underlined. The following is an example of the article of **tools** (figure 12), and how it can be registered and described in our corpus as an academic lexical item:

TOOLS.- 43 / 78 / 154 TXs RPs RAs
___ *and techniques* (5) / *software* ___ / *html writing* ___ (2) * Lib. TX ___ *such as* (4) /*analysis* ___ *to* + V / *as daily* ___ / ___ *used to* / *internet* ___ /
TOOL.- 29 / 43 / 62 ® *Browsing* ___ (4) / *powerful software* ___ ₣ *Technology is a* ____ (4) / *research* ___

Figure 12: Sample entry of academic item.

[218] This specific context – or genre—for specialized languages thus meets the stylistic expectations of the academic community (Cowie, 1998: 12)

[219] As these same authors state, subject field must thus be considered independently of culture (i.e. general use of terms) (Bergenholtz & Tarp, 1995: 70).

This is thus a sample of how `lemmata are arranged and presented in clusters´(Bergenholtz & Tarp, 1995: 194). In corpora for specific purposes such as our own, in fact, the order of items under a particular lemma follows the priority of frequency of occurrence (Cheng-yu, 1993: 77), and the lemma is the lexeme that is used most frequently – i.e. the highest frequency equals the first meaning and word-form (Summers, 1996: 263). In this sense, lexical storage, as conceived in our work, is closer to the area of lexicology, as defined by Richards et al.(1992: 212), than to lexicography: While the former encompasses `the study of vocabulary items (lexemes) of a language, including their meanings and relations´, the latter refers to `the compiling of dictionaries´ (ibid.), where items are organized in alphabetical order. Bauer (1983) and Lyons (1977) also serve as models in the description of the word forms and lemmata relationships in our corpus.

The numbers given immediately after the entry / sub-entry refer to the total number of instances of the given key word in the genres of textbooks (TXs), reports (RPs) and research articles (RAs) respectively. The lexeme in bold type appears as the lemma in this lexicologic view. As Carter states (1998: 7),

> The term lexeme also embraces items which consist of more than one word-form. Into the category come lexical items such as multi-word [units] (…) [these are] vocabulary items or simply items, a fairly neutral hold-all term which captures and, to some extent, helps to overcome instabilities in the term *word* (…)

All the multi-word terms and clusters are presented in italics; the collocations can be examined and framed within the ESP perspective, as they may become more or less relevant when dealing with issues and language in our technical settings. A double-space separation is made between restricted (®) and free combinations (₣) for each entry. These are included in the second row of lexical items, while the top one comprises restricted expressions.[220] Thus, in the case of the sub-entry tool, for example, *browsing tool* is considered restricted due to its

[220] See Howarth (1996, 1998) above for definitions of these concepts of restricted vs. free collocates.

specialized use in hypertextual searching, while *research tool* tends to be less technical, employed across a broader range of texts and topics. As mentioned in the previous section, this distinction is more objectively measured in terms of the range of texts where the given form occurs. In the first case, it tends to appear highly frequently in a low number of texts, while free lexical combinations may be less frequent but managed across a higher number of different texts. Finally, the asterisk * is employed to indicate the low frequency of a given collocation that is nevertheless valued and included due to its importance in a specific subject field (here, in Library science textbooks).

III.2. LEXICAL ORGANIZATION ACCORDING TO CATEGORIZATION.

Specifying the thematic context in which words interrelate and behave is dealt with from the standpoint of special subject / topic identification, as is the case of subject-based lexis. This type of contextualizing can be done in relation to genre, but, more importantly, according to its position as key in a specific sub-corpus organized in terms of shared conceptual domain. To fully describe common core lexis by means of subject-based lexical items, we must consider `the place the word occupies and the contrasts´ as a characteristic item in subject areas (Carter, 1998: 60).

From this focal point, `collocational and colligational patterns are meaning-creating´ (ibid., 62): the interdependence among grammar, lexis and semantics is of prime importance to define a particular context of use (ibid.). These items work together `in the making of meaning´ (Clear et al., 1996: 313). In this sense, our lexicologic approach cannot be conceived independently from the `conceptual network which has to be recognized and interpreted´(Carter, 1998: 104) so that the specialized sense of the word may be realized. The

analysis thus seeks to develop a display of keywords within each subject specified for our large Information Science and Technology area.

Furthermore, `lexical choices can vary relative to text type or genre of writing´ (Carter, 1998: 107). We therefore deem pertinent that additional information on type of text according to rhetorical moves be encompassed when dealing with other categories, such as procedural or argumentative lexis. These should be regarded, from this perspective, in the context of both text type and genre, given our need to describe stylistic choices.[221]

Let us then examine four instances of lexical entries in our corpus that are organized according to these demands, with labelling to specify the interrelation between textual and lexical units. Such a correlation is something which teachers dealing with lexis should be conscientious of (cf. Hill, 1996). The attempt is to offer a distribution of lexical items so that the entries are not `impoverished when viewed in isolation´ (Boguraev, 1991: 229), and `their interpretation in context´ (ibid., 233) can be facilitated and made accessible.

The approach followed, especially in the case of argumentative, procedural and grammar / discourse items, also considers, to some extent, the schema theory parameters –see sub-sections 2 and 3 of `Discourse´in chapter 1—of instantiating the reader's schemata:

> The data that are needed to *instantiate*, or fill out, the schemata become available through bottom-up processing; top-down processing facilitates their assimilation if they are anticipated by or consistent with the listener / reader's conceptual expectations (Carrell & Eisterhold, 1983: 557).

Thus, for instance, where there are `organized patterns of *comparison*, *causation*, and *problem / solution (...)* the recall of specific ideas from a text´ can be achieved more effectively (Carrell, 1987: 463). This is the case, for instance, of procedural phraseology and

[221] In this sense, by no means do we intend to produce a vocabulary as detailed as, for instance, that suggested by Makkai's socio-linguistic approach to lexicography (1980), where natural semantic nests are included according to frequency of usage, exact range of dialectal habitat, the speaker's sociological status, etc. Our work, in contrast, tends to humbly be modelled after Clear et al.'s (1996) and Sinclair's (1996) description of entries by means of `words enter[ing] into meaningful relations with other words around them´(Sinclair, 1996: 76), the

its employment in `conceptual paragraphs´(Selinker et al., 1976: 284). Their explicit signposting, in this respect, is associated with ideas and notions being discussed in specific sections of the text. As Aston asserts (1997 a: 267), in fact, given phraseologies can be characteristic of a particular part of the text because of their concentration in such a place.[222]

The aim is, from this perspective, to reflect lexicologic data as communicative feedback and a teaching aid,`the word work[ing] indexically to key in with and complement a context of shared knowledge´ (Widdowson, 1993: 309). In this respect, `background information on the topic and (…) knowledge of how discourse is organized´ (Celce-Murcia, 1997: 363) are demanded specifications for subject and procedural lexical items respectively. Thus producing vocabulary for communication (comprehension and production of L2) (Atkins, 1985: 15)–see chapter 4 —is a key objective in our study. The pedagogical concern serves as the end for the development of our methodology in lexical storage; this implies allowing successful learning by having learners cope with `the demands of the mental-model construction (…)´(Schnotz et al., 1993: 191). The combination of quantitative corpus techniques and qualitative interpretation, in this respect, `ensure that we understand the communicative functions that the linguistic features are serving´(Conrad, 1996: 301).

The first example to be examined leads to regarding subject categories in the Information Science and Technology curriculum classified above, as the main reference point for viewing the lexical contexts of operation. The starting point implies, then, selecting texts belonging to a particular subject matter, such as `Content analysis´ (C1 category) for instance, studied in both Library science / Documentation and Audio-visual communication majors. As with word discrimination according to the topics above, the choice of sources is made in

---

pedagogical goal being, as mentioned above, activating passively recognized vocabulary by means of semantic networks (see Fodor [1983], and other references listed throughout this chapter).

[222] The **Dispersion Plot** function of *WordSmith* can be quite useful for this task of observing where words crop up in the texts.

order to compare two word lists, one drawn from the subject texts and the other from the whole corpus.[223] Key words are thus identified and checked in the concordance to see their contexts of use. The lexical units must have a significant range and frequency in the texts chosen, so that they may be considered as common core lexis in terms of subject-driven input. This is a key characteristic of subject lexis; as Tribble (1998) states, the best way to view this distribution is by comparing texts by means of the **Keywords** function in *WordSmith*: the contrastive examination of lexical data in both the smaller sample of subject-based texts and the large reference corpus (our own). Thus, Biber (1988), Hoey (1997) and Scott (1997) complement each other in strengthening the lexical analysis of subject / register variation across texts (cf. Tribble, 1998), since the analysis of keywords according to thematic concerns tends to encompass the three views: checking variation and / or uniformity of use (Biber), regarding lexical association within conceptual rungs as descriptive material (Hoey) and examining pivotal words that concentrate distinctive subject denotation (Scott).

In our analysis, the subject lemma **media** (figure 13), in this respect, occurs in the fourth position on the list of keywords displayed for the `Content analysis´ sources classified under Section **C (1)** in our `Course syllabi´ description above.[224] The `keyness score´ (measured by the **Keywords** function of *WordSmith*) for the node is 254.1, which means, according to Scott's scale for this appreciation (1996), a high degree of relevance of the given word as a keyword (as aforementioned, the percentage indicated by WordSmith so that a lexical item be considered a keyword is 25 --the cut-off point). The noun **media** in this case is rather characteristic of the texts chosen for the subject category **C (1)**. The noun **media** is repeated 74 times in the **C (1)** texts, in contrast with its 478 occurrences found in the whole corpus. Figure 13, taken from appendix 6, collects such subject-based lexical information:

[223] As Tribble and Jones (1997: 36) claim, the starting point should always be `frequency sorted lists´.

MEDIA.— 74 (C1)	**C (1) Content Analysis**
Mass ___ (22) / *communications* ___ / ___ *and information* / *Of mass* ___ / ___ *in the information* / *of communications* ___ / ___ *and information technologies* / ___ *economics* / *the mass* ___ / *mass* ___ *gives* (1) * Aud TX	

Figure 13: Sample entry of subject item.

The number of instances indicated near the node corresponds to the frequency of the word in the specified subject `Content analysis´ (C1). Unlike academic entry articles above, the subject-based entries, as may be inferred, do not include the distinction between restricted and free collocates. The reason being that the academic lexis category, as already explained, is considered to encompass those elements with a large co-occurring range in the whole corpus, and not mainly a few specified subject-based sources, as is the case that occupies our current concern (see corpus references to check the sources comprised in each subject category).

A final comment should be made about the use of the **Keywords** function in *WordSmith*. As stated in chapter 1, this facility and the **Detailed Consistency List** that groups all items according to specified corpora, are central factors for choosing this specific software for our lexical analysis. As can be seen, they both serve as starting points for selecting and working with words. Yet, **Keywords** is mainly employed to specify subject lexis that appears most frequently across a chosen set of sources, with direct access to viewing the concordance lines of such items in context. For our purposes, as long as the level of keyness is above 25.0 –which equals a significant number of occurrences (about five) in the text-- the lexical items of this subject list should be explored. Other examples in the `Content analysis´ category above are **images** (eighth position in the keyword list), **GIF** (10th) and **mass** (11th) (see chapter 3).

[224] See chapter 3 (Subject results).

The remaining examples of stored entries are three: the cases of argumentative, procedural and discourse / grammar items. As stated above, the condition that is hereby sought is that the lexical items be only identified in one particular genre and in specific contexts or parts of texts that characterize them so that uniformity and / or variation may be tested within the chosen environment. This is achieved by using both the **Detailed consistency analysis** and the **Dispersion Plot** features of *WordSmith*.

In following this process, the first function allows us to see that **passage** (figure 14) is a lemma that appears 16 times in two textbooks and five excerpts from these, covering the subjects of hypertext technology and content / semiotic analysis. Thus, it is considered to comprise all the disciplines (categories **F (4)** [Hypertext Technology] and **C (1)** [Content Analysis] in course syllabi shared topics):

PASSAGE.- 16 (TX)	(5) **F [4]** , **C [1]**
The ____ of the text (4) / *the ____ in the text* /	

Figure 14: Sample entry of argumentative word.

This argumentative node refers to the writing process od the academic genre by co-occurring in clusters such as the ones presented. Despite their low frequency, such phraseological devices are to be recorded, since their specific use along the text marks the author's sense of direction in the work, which can therefore serve as a productive reference for academic writing skills development. To observe the particular parts of the texts where these lexical items appear, the dispersion plot statistics (figure 15) proves useful:

N	File	Words	Hits per 1,000	%
2	Book2b.	11,802	8	0.68
3	Book2a.	4,204	2	0.48
4	Book2e.	8,754	4	0.46
5	Book2g.	7,736	1	0.13
6	Audbook2.	21,122	1	0.05

Figure 15: Dispersion plot statistics for the concordanced lexical item **passage**.

The first six sections belong to the book on hypertextual technology, while only one part of the Audio-visual communications textbook contains one instance of the lemma. The statistics in figure 15 are standardized to show the ratio of items per 1,000 words of text (third column), which means that those words may either crop up in a given part –such as case number 3 (with two lexical instances)--, or be spread out more regularly throughout the text – as happens in the top row of eight occurrences. The dispersion plot graphic (figure 16) visualizes this lexical gathering:

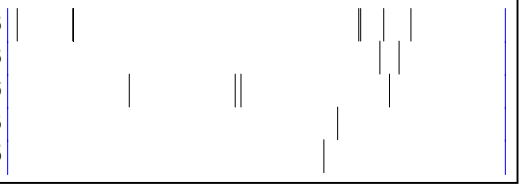
N	File	Words	Hits	per 1,000	Plot
2	book2b.doc	11.802	8	0.68	
3	book2a.doc	4.204	2	0.48	
4	book2e.doc	8.756	4	0.46	
5	book2g.doc	7.736	1	0.13	
6	audbook2.txt	21.122	1	0.05	

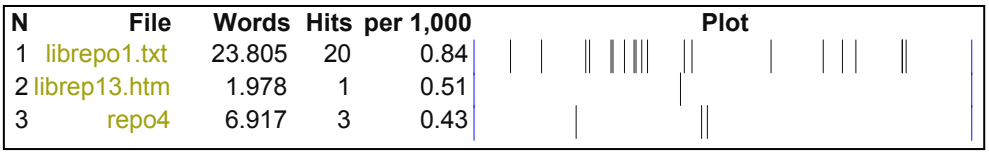
Figure 17: Graphical representation of lexical dispersion in texts.

A further specification of this type of approach can be illustrated by procedural lexis: for instance, the noun **process** (figure 17) is used across genres; yet, the lemma is more frequent in three reports dealing with both content analysis and virtual documentation respectively (Library Science) – **C(1)** – Content Analysis--, and **C(2)** –Media Documentation--, and one covering hypertext topics in Internet (All topics): **F [4]** –Hypertext Technology--category:

PROCESS.- 30 (RPs)	(3) **C [1]**, **C [2]**, **F [4]**
In the ___ of (4) / *with the ___ of* / *the development ___* * Lib. RP	

Figure 17: Sample entry of procedural word.

Once again, the three texts can be scanned by means of the dispersion plot (figure 18) to view the distribution of usage:



N	File	Words	Hits	per 1,000	Plot
1	librepo1.txt	23.805	20	0.84	
2	librep13.htm	1.978	1	0.51	
3	repo4	6.917	3	0.43	

Figure 18: Dispersion plot of concordanced lexical item.

As shown, the node is more evenly employed in the first Library Science report (`librepo1.txt´ -- 20 hits).

A similar type of methodology is followed with the last set of words, the grammar / discourse markers. In this case, equally driven by the need to contextualize the use of these words in the texts of a given genre, our interest extends to viewing samples in, for example, research articles. As with argumentative and procedural vocabulary, text types such as rhetorical functions may be incorporated and labelled in the entry articles, as we shall see in chapter 3. In this respect, the colligating node **rather** (figure 19) can be checked in the context of Discussions as offered by research articles:

RATHER.- 13 (RA)	(1) **F [2]**
___ *than* (5) / ___ *than to* + verb (1) * **Discuss**.	

Figure 19: Sample entry of grammar / discourse item.

The abbreviation **Discuss** (after the clusters) stands for samples of texts selected as representative of discussions in the genre of RAs (which is, in turn, pointed out near the frequency of the item in such a selection – 13 instances—in the lemma slot). The subject area is also signalled in the next space above (F2 = Unix / Internet). This is mainly done as a manner of combining both matters of style and conceptual level in this third class of lexis (text type-based) – while genre-based academic items would only deal with the stylistic concern, and subject-based words would do the same in the thematic sphere.

The dispersion plot can then be used to check whether the lexical item is evenly distributed, but in the text chosen, we see that such words combining with **rather** tend to crop up in the first half (figure 20):

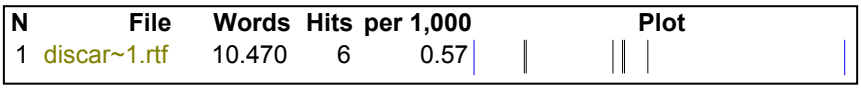


Figure 20: Dispersion plot of word being analyzed.

We thus conclude this second chapter of our study after having developed the methodology of the approach to be followed in the lexical analysis. This is based on both applying Corpus Linguistics techniques and consideration of genre, subject and discourse focal points in the sources of the corpus. This process should provide us with valid material for our research discussion and teaching purposes.

CHAPTER 3: RESULTS

The aim of the third part of our research is to offer the details gathered from the lexical analysis of the texts researched. In such a collection, lexical items are arranged, as has been examined in the previous section, according to the notions of word frequency and range across text types and genres, or within given subject fields and topics in the Information science and technology disciplines. How strong these lexical bonds are, based on statistical relevance –e.g. M.I. (Mutual Information) measurement—and textual significance[225] is also quite pertinent to our study.

The degree of collocation is thus assessed in the light of common coreness: That these patterns are more or less consistent in our corpus is, indeed, a key characteristic to value so that a reference with the total number of texts and running words can be established. As the findings will show, there exist, in addition, representative lexical items for a limited or reduced number of texts --e.g. subject-based vocabulary; hence, the identification of keywords must be pursued not only according to shared themes, but also independently within each discipline. This is mainly done to organize content words according to established categories of Information Science and Technology.

The environment of text and discourse is of prime importance for the situation of lexical items in the scope of academic linguistic competence. A first setting, text types, is approached in relation to how text is organized as regards coherence and cohesion for the display of rhetorical functions, while the genre space is geared towards registering the writer's inclination and intentionality to produce discourse for a community (e.g. academic).[226] There is a third parameter on which the distribution of the lexical items of our corpus is based: subject / topic. In this case, a framework based on content is provided, and the findings yield core lexis according to thematic / conceptual fields. Common coreness, in

[225] See footnote 103 (chapter 1).

this sense, is valued with a reliance on such variables. Distinct degrees of vocabulary centrality or importance may thus be induced for language teaching priorities in our specific setting.

I. TEXT TYPES

The distribution of results is drawn according to three approaches introduced and analyzed in previous chapters: word frequency / range, collocations and keywords. The focus is placed on both the text and the subject-matter. This is essentially done to follow the priority of working with language and content from the ESP perspective, as has been mentioned throughout this study. As a consequence, a categorization is made regarding a specified kind of context --in this first instance, text type.

I.1. WORD FREQUENCY AND RANGE.

In this first division, the most frequent text type words are provided according to how recurrent they appear across six sets of ten texts. These are grouped as follows:

1. Classifications.
2. Conclusions.
3. Definitions.
4. Descriptions.
5. Discussions.
6. Exemplifications.

The samples are taken randomly to represent the rhetorical functions and sub-sections

[226] We judge the distinction of text genre and type to be that claimed by Glässer (1998) –see footnote 95 (chapter 1).

of genres with which the learner must cope and come to grips.[227] As stated in chapter 2, the relevant vocabulary analyzed from this perspective is classified as argumentative, procedural and discourse / grammar items, examined in selective domains such as distinctive subject fields and genres.[228]

The graphic in figure 1 displays the distribution of the texts in the entire corpus sampling. Our immediate concern thus lies in having all the interelationships among the subject fields represented visually in order to make the selection of text samples accordingly: The figures refer to the number of sources belonging to the specified disciplines –pointed out by the abbreviations.[229]

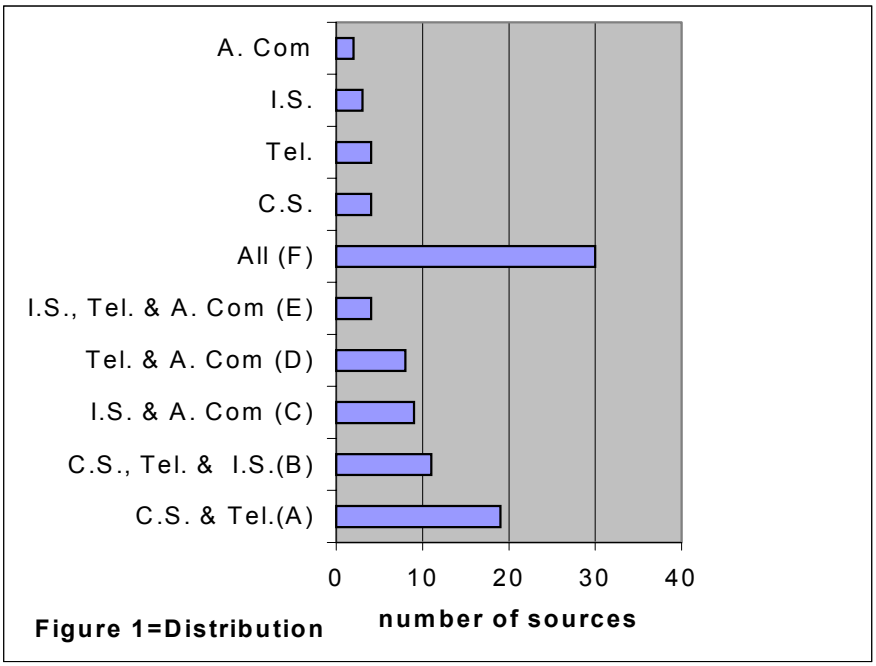


Figure 1=Distribution

[227] Only these six types are chosen due to the fact that others, such as the discourse function signalling contrast is contained by Discussions in five occasions, whereas illustrations are coped with in Exemplifications. In turn, the two sections of research articles included –Discussions and Conclusions—are given priority over Abstract, Introduction, Method and Results since these are already selected for the compilation of Descriptions, Classifications, Definitions and Exemplifications to a greater or lesser degree (see distribution of the text type sub-corpus below. See also text type sub-corpus references after bibliography for checking the sources from which the samples are chosen).

[228] How certain genres and types can be characterized by core or subject-core lexis is described in chapter 1 – sub-sections on `Genre´ [Genre as the study of lexis] and `Lexical focus´[Sub-technical vocabulary and the ESP course] –see Carter (1997, 1988).

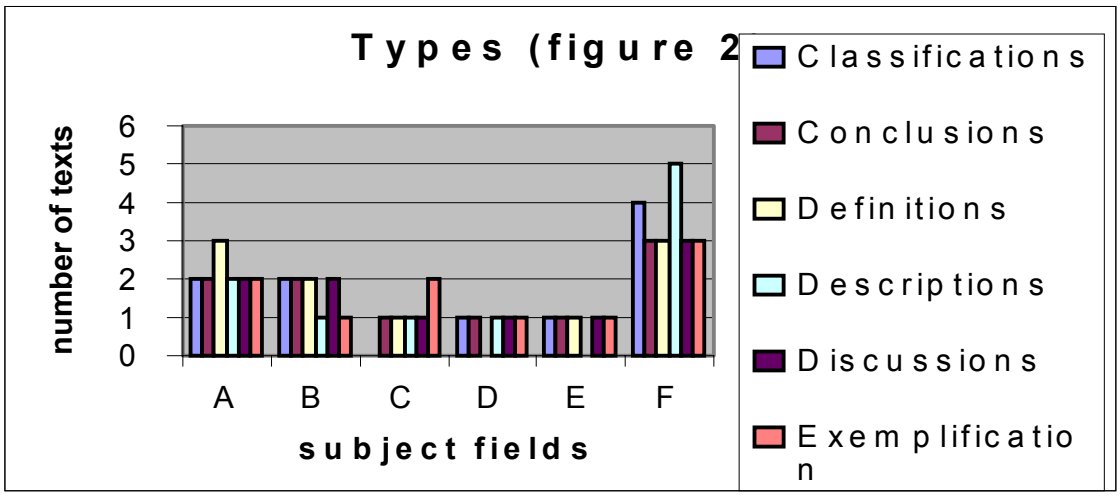
[229] `A. Com´ stands for Computer Science; `I.S.´ for Information Science, `Tel.´ for Telecommunication, `C.S´ for Computer Science (see index 2).

The capital letters refer to the codes used for the subjects / topics within disciplines, as given by course syllabi (chapter 2) and as shown in the Appendices (Appendix 1). As can be observed, in addition to the labels (A – F), each single subject field is also represented individually by some texts (not shared with other studies).

As can be observed, there are more texts in the `F´category, which all four scientific areas share; in contrast, the subject `Communication Theory´, included in the Audio-visual communication, Information Science and Telecommunication programs of studies, is conformed by only four sources. In turn, Audio-visual communication is the discipline with the smallest number of readings involved –only one text for each genre.[230]

Types are then selected from the texts according to the categories, as figure 2 shows. The selection is made by measuring yardsticks of the overall distribution and length of these in the corpus: As a result, if there are up to five descriptions (out of 10 possible ones) included in the F or `All disciplines´ category, this is due to the fact that

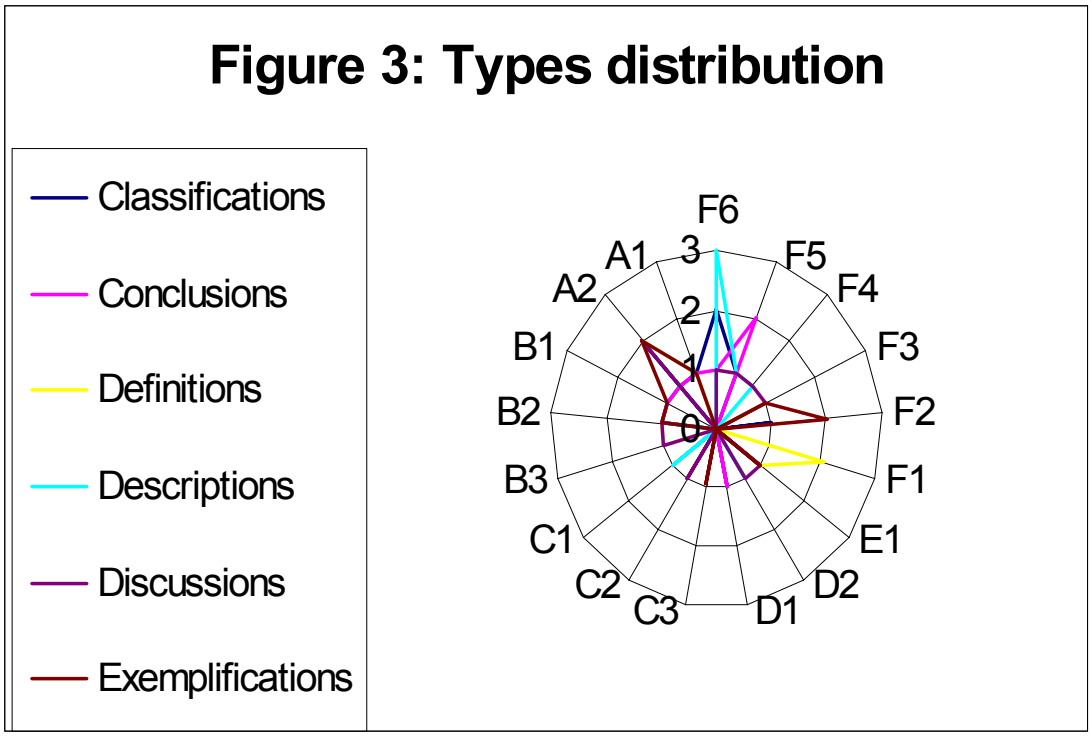
SUBJECT FIELDS
A = C.S.+Tel. / **B** = C.S.+Tel.+I.S. / **C** = I.S.+A.Com / **D** = Tel. + A.Com / **E** = I.S.+Tel.+A.Com / **F** = All disciplines (see index 2)



[230] The correspondence of number of texts and disciplines follows the aim of assembling core language and subject matter: the lower the measure of samples, the more subject-specific the texts tend to be.

the passages are quite common in these readings. In addition, the five samples are not as long as other types, such as definitions in the F division. Finally, that the balance in relation to the entire corpus (figure 1) be kept, is, as has been pointed out, a chief consideration.

Text type choice, in addition to being based on frequency and range, should be made with reference to the subjects and topics comprised. This is mainly done in order to account for differences in the development of topics and concepts depending on



A = C.S.+Tel. / **B** = C.S.+Tel.+I.S. / **C** = I.S.+A.Com / **D** = Tel. + A.Com / **E** = I.S.+Tel.+A.Com /
F = All disciplines (see index 2)

subject matter. Figure 3 provides the distribution of text type sources within each sub-category or label.

The maximum number of texts encompassed is, as pointed out, three in the case of Descriptions on the topic of `Information infrastructure´ (F6). This distinction reflects both the larger amount of readings existent in the corpus dealing with issues of this kind and the recurrence of this type of rhetorical function employed in sub-division F6. In contrast, where

no samples are contained within a given sub-category, the reason is that the model was either less developed or not included at all in the content of the text (e.g. Conclusions in `Perspectives on information´ [F1], `Media theory´ [D2], `Media documentation´ [C2], `Automated Knowledge-based systems´ [B3], etc).

A final comment must be made regarding the importance of keeping a balance with the representation of the three genres in the construction of the text type sub-corpus. Thus, the table in figure 4 displays the proportion of each heading in this respect:

Genre	**Descrip-Tions**	**Classifica-tions**	**Definitions**	**Exempli-Fications**	**Dis-Cussions**	**Conclu-Sions**
TX	F5	F2, E1, A1	F3, E1, A1	F2	E1, A2	C1
RP	F6, F4 F6, D2, A2	F6, D1, B2 B1, A2	F1, B2, A2, A2	F2, E1, C3, B2, A1	F6, D2, B3	F6, A1, B1
RA	F6, C1 B2, A2	F5, F6	**F1, C2, B3**	F3, B1, A2, A2	F3, F4, B2, A2, C2	F5, F5, E1, D1, B2, A2

Figure 4: Types in genres
TX = Textbooks / **RP** = Reports / **RA** = Research articles
A = C.S.+Tel. / **B** = C.S.+Tel.+I.S. / **C** = I.S.+A.Com / **D** = Tel. + A.Com / **E** = I.S.+Tel.+A.Com /
F = All disciplines (see index 2)

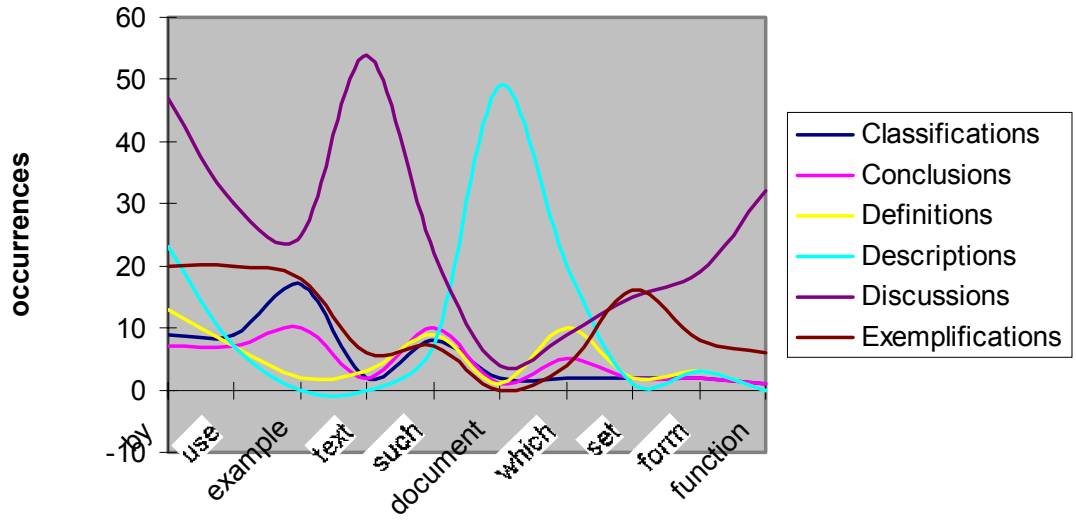
The intent of this arrangement is to offer a solid basis for text selection and analysis. Since the goal of such an organizing procedure is to provide adequate grounds for lexical sifting, these text type sub-corpora should incorporate as many different language and content settings –i.e. contextual factors—as possible. In this sense, some text units that are characteristic of one single genre, as is the case of research article discussions and conclusions, can be likewise located in the other two genres (e.g. a discussion appearing in a

textbook on Communication theory [E1] or a conclusion taken from a report on Software programming [A1], as figure 4 indicates).

Having predetermined these parameters of selection, we apply concordancing methods for the drawing of significant data. Based on each genre according to figure 4 above, the types of lexical items sought are argumentative, procedural and grammar / discourse vocabulary; these, as explained in chapter 2, should be examined within single communicative events of academic writing (Swales, 1990: 58).

Figure 5 gathers this type of lexis as gleaned from sub-corpus C of journal articles. The graphic presents the top ten words in terms of both frequency and range –i.e. given the number of occurrences and of texts in which these words appear.[231]

Figure 5: Text type words in RAs



Due to the fact that Discussions contain a high proportion of text in our sub-corpus,[232] the frequencies are higher in that type, as the graph shows --except for the cases of **document**, **which** and **set**. The ten words displayed are ranked, from left to right, according

[231] For the rest of the research article text type data, see Appendix 2.1.

[232] This is also true in the case of other corpora for academic / teaching purposes, such as Howarth (1998), or Granger (1998).

to both frequency and range. For the determination of this second factor --range-- a file-based concordance is performed, which facilitates sorting co-texts according to the six categories in use (classifications, conclusions, definitions, descriptions, discussions, exemplifications). Thus, the higher the count of text files found for each text type heading, the greater the justification for the word to be considered common core.

As stated, Appendix 2.1 reproduces argumentative, procedural and grammar / discourse vocabulary as collected from the content of these texts in terms of frequency and range. The list stops at a point in which the range was assessed as too low --below three texts. In the illustration --figure 5 above--, one word was omitted due to its occurrence in a mere sample of three texts. The argumentative noun **page** is actually repeated 45 times, of which most were found in a hypertext technology article (F4 = all disciplines).[233]

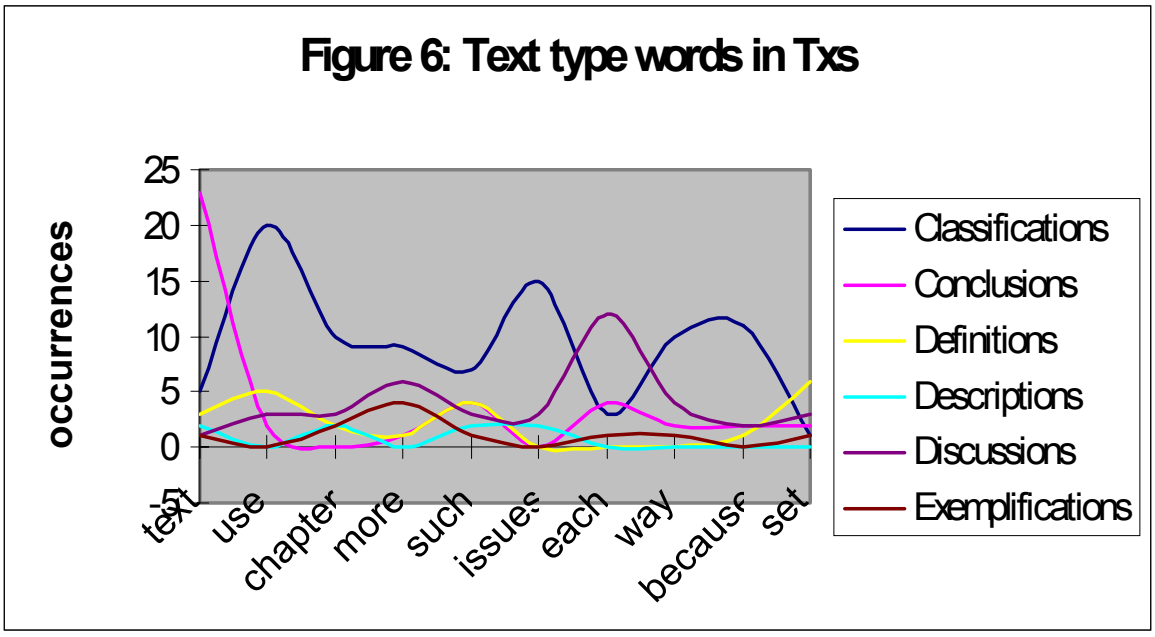
According to the criteria stated in chapter 2 (`Word taxonomy´), these lexical items occur in co-texts where reference is made to the discourse form of creating cohesion and coherence in the given subject field by means of lexical signposting. In this respect, **use, text, document**, and **set** are considered argumentative elements in the process of bringing the exposition together, whereas **example**, **form** and **function** can be analyzed as procedural lexis – key items in offering a rhetorical move. The remaining three, **by**, **such** and **which**, include relevant colligations of grammar / discourse words with content vocabulary.

In order to check that these are meaningful senses in the type context, however, a suitable measure is to preview collocations and clusters with the given nodes. The observation must focus on pinpointing the use of lexical items and phraseology for developing concepts so that themes flow in the subject-matter discourse. This is done in the next sub-section, `Collocations´.

[233] Its title is "The Hyper-Text Markup Language (HTML) and the World-Wide Web: Raising ASCII Text to a New Level of Usability" --see text type corpus references after bibliography.

Before the aforementioned task is carried out, nonetheless, word frequency and range across text types is also examined in the other genres: textbooks and reports. Genre and text type must be regarded separately so that specifications based on the employment of lexis may be drawn in terms of both single environment and text function.

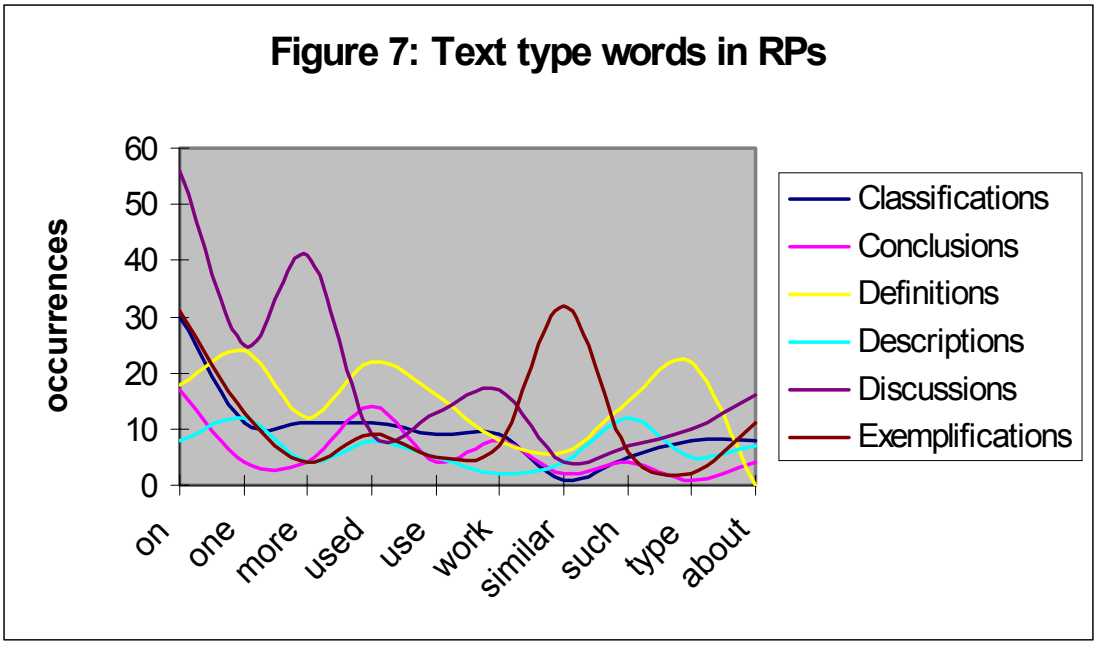
The data in figure 6 illustrates the relationship of words with text types in the genre of textbooks from the joint planes of frequency and range. As in the case of research articles above, the focus is on the ten most repeated vocables belonging to any of the three classes -- argumentative, procedural and grammar / discourse lexis-- (the remaining data are displayed in Appendix 2.2).



Classifications, in this case, contain more words than any other text type (11,313 tokens vs. 2,207 in Definitions, for instance). Yet, in this latter set of texts or in Discussions, the standardized ratio is also lower, which means that there are fewer words repeated and thus a high number of types --e.g. 795 in Discussions vs. 1,370 types in Classifications. The argumentative item **text,** with 23 occurrences in Conclusions is the first form in terms of

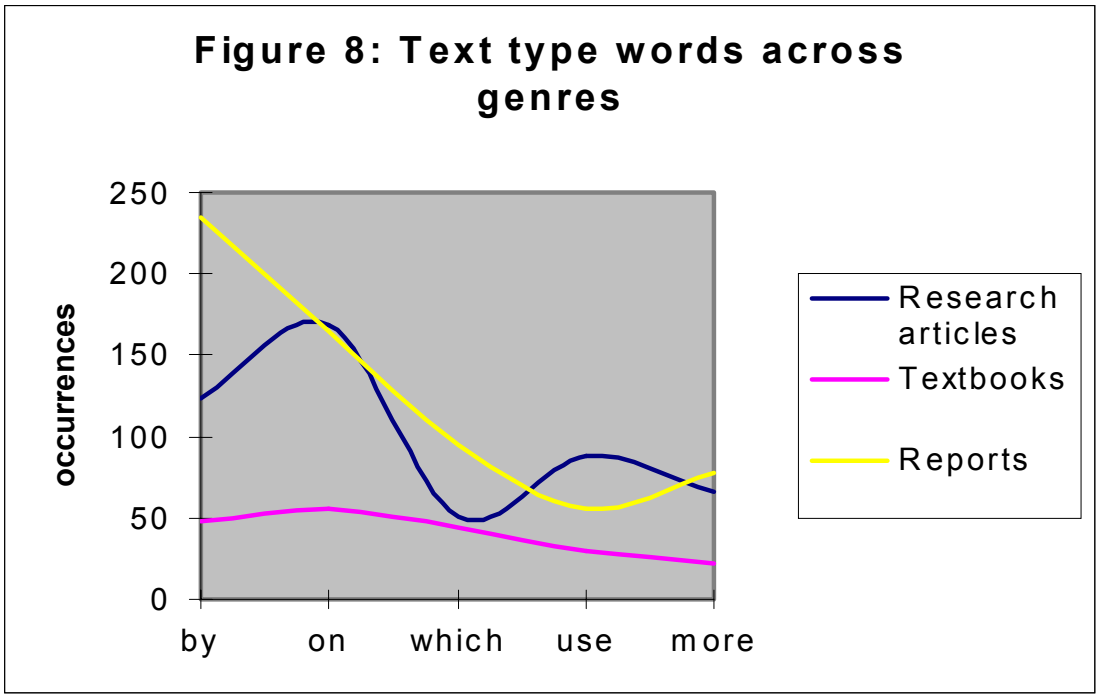
frequency and range, as can be observed. Other members of the argumentative lexis group are **use**, **chapter, issues**, and **set**. **More**, **each** and **way** can be conceived as procedural, given their functions, as co-texts may reveal in the next sub-unit. Finally, **such** and **because** belong to the grammar / discourse positions of lexis.

The other genre encompassed --reports-- also serves as the setting in which we trace the profile of the top ten words from the line of work depicted (figure 7). In this case, there are fewer argumentative items than above -- **used**, **use**, **work** and **type**. The procedural words, in turn, are **one**, **more**, and **similar,** whereas **on**, **such** and **about** are chosen as representative grammar / discourse items occurring among the first 70 function and content items on the list --see Appendix 2.3 for the remaining entries.



As can be observed in figure 7, range is visualized by the decreasing use –from left to right-- of the terms in the chart. Textbooks (figure 6) depict a similar cline, while articles show it less steadily in figure 5. In reports, the Conclusions type may be regarded as the textual setting presenting the most consistent distribution of the ten words (having neither higher than 13 instances nor lower than 3).

A concluding step to be taken regarding word frequency and range in the text types, is the arrangement of the overall lexical data according to each genre. This is



done for the main purpose of perceiving more clearly the text type words that are more evenly distributed across genres.[234] The findings in figure 8 are five occurrences identified among the top 50 items of the Detailed Consistency List (DCL), yielding frequency counts and lexical distribution on texts.[235] In figure 8, in addition, that these five items function as argumentative, procedural and grammar / discourse markers significantly in all the types is highly valued. The property of common coreness is thus assessed by highlighting both factors of frequency and range.

There seems to be a tendency for the grammar / discourse devices to occupy the top positions in the DLC, as the curves show in figure 8, being the genre of technical reports the

[234] Nonetheless, as explained, it is relevant to our contextual analysis of text type vocabulary that the examination of co-occurrences be conducted within each separate genre – to check both variation and similarity of use.

[235] See concordancing functions in chapter 1.

one that introduces a greater amount. This employment is also high in research articles, but, here, the argumentative vocable **use** is more widely expressed, in contrast with the other two genres. Textbooks tend to reflect fewer of these text type-based co-occurrences instead. In the next sub-section, the lexical behavior of three of these forms and some other examples from figures 5, 6 and 7 are checked.

I.2. COLLOCATIONS.

The five words displayed in figure 8 above constitute a suitable primary sample for the analysis of collocates and clusters in text type settings. This is due to the fact that those are the top text type items in terms of frequency and range across types and genres, thus being liable to show, in a broader perspective, their detailed varying or similar behavior in the texts. Appendix 3 adds twenty subsequent entries of collocations arranged from the same viewpoint (frequency and range over genre settings).

The following are the articles of the grammar / discourse item chosen, **by**, the first in the list (figure 8). Co-texts are demarcated and organized in figure 9 according to the priorities and principles underlined in chapter 2 (`Lexical storage´); these are recalled by means of footnotes at the bottom of the graph below (figure 9). In appendix 3, the entries are presented in the same manner and according to the same parameters of organization.

BY -- 48 [1] (TXs) [2]	(10) [3] A1, A2, C1, E1, F2, F3, F5 [4]
Past part. + ___ [5] + agent (12) [6] / ___ *making available* (2) * [7] **Class**. [8]	
Claims made ___ *the text* (1) * **Concl**. [9]	
___ *defining* (1) / *cannot be run* ___ *itself* * **Defin**. [10]	
Defined ___ *procedures* (3) / ___ *providing procedures* / *sent* ___ *telephone lines* / ___ *the end of* (1) * **Discuss**. [11]	
Past part. + ___ + agent (5) **Exempl.** [12]	

BY -- 235 [1] (RPs) [13]	(25) [3] A1, A2, B1, B2, B3, C3, D1, D2, E1, F1, F2, F4, F6 [4]
Past part. + ___ + agent (6) / past part. + ___ + V-ing / *Technology ___ itself* (1) * **Class.** [8] *___ using* (2) / *used ___ Multics* / *___ default* (1) * **Concl.** [9] *Sent ___ reference* (6) / *communicated ___ reference or ___ copy*ing / *used ___ the transport interface* / *assumed ___ the* / *communicated ___ marshaling them* / *___ the designers* / *invented ___* / *needed ___ applications* / *___ means of hyperlinks* (1) * **Defin.** [10] *Discussed ___ the steering committee* (2) / *separated ___ :* / *set ___ certain commands* / *interpreted ___ the browser* (1) * **Descr.** [14] *Requires additional actions ___* (3) / *defined ___* + agent / *Should not be worked ___ a single society* / *Being developed jointly ___ the association* / *___ continuing* / *___ the IEEE* / *get more information ___ visiting* / *report ___ the* / *sponsored ___* * **Discuss.** [11] *___* + agent (15) / *___* + V-ing (7) **Exempl.** [12]	
BY -- 124 [1] (RAs) [15]	(24) [3] A2, B1, B2, B3, C1, C2, C3, D1, F1, F3, F4, F5, F6 [4]
Used ___ location (3) / *___ author, title or subject* (2) * **Class.** [8] *Increasingly replaced ___* (2) / *caused ___* / *Attempts ___ the* (1) * **Concl.** [9] *Influenced ___ the* (2) / *the paper ___ describing* (1) * **Defin.** [10] *Characterized ___* (2) / *created ___* * **Descr.** [14] *Recommended ___ the HTML specifications* (3) / *___ broadening the range* / *___ now,* (2) * **Discuss.** [11] Past part. + ___ + agent (11) / *___* + V-ing / *such as approximating a sum ___ an integral* (1) * **Exempl.** [12]	

Figure 9: Text type grammar / discourse entries.

Footnotes: [1] Number of instances in genre analyzed / [2] Genre of Textbooks / [3] Number of texts containing the lemma / [4] Subject categories where texts are included / [5] Gaps indicating position of lemma / [6] Number in brackets indicating amount of occurrences of the most frequent construction in a text type / [7] Asterisk signalling the least frequent lexical item (two instances or less) / [8] Text type of Classifications where instances are observed / [9] Conclusions / [10] Definitions / [11] Discussions / [12] Exemplifications / [13] Genre of Reports / [14] Descriptions / [15] Genre of Research articles (see index 2)

Text type lexical combinations are thus relevant not only depending on frequency and degree of repetition, as may be inferred, but also on their position in the structure of the text -- cf. Pedersen, 1995 (see chapter 1 `Collocation´). In this respect, a cluster such as *claims made by the text*, located only once in the Conclusions sample of textbooks, is noted down due to its salient condition in the discourse setting: amounting to a remark for collecting thoughts upon notions previously developed. In other cases, the low frequency items are considered as important as the higher frequency ones because of their rhetorical signalling; for instance, *by means of hyperlinks* in the Definitions section of reports, establishing a key procedural statement with the preposition **by**.

As may be contrasted, the lexical information provided by one genre may provide enough data in our analysis. This is mainly due to the fact that it does not differ much from one genre to another. In figure 9, this is made explicit by Exemplifications, as outlined by the three genres. The textual environment of **by** does not change much in this respect, as can be regarded. In a similar fashion, for the Classifications types, functional aspects of the prepositional node resemble each other in textbooks and reports: mainly passive and intrumentation uses with gerunds. In all three genres, in addition, the sections of Discussions tend to produce large sums of lexical combinations. This may be due to the extensive textual elaboration identified for this type.

It is, nevertheless, in the third genre category, RAs, where the lexical contour varies, turning up characteristic constructions that surface as key elements in given types – e.g. Classifications, where *used by location*, *by author, title or subject* seems to be a significant expression in this context of use. Other such distinctive elements can be seen in the employment of clusters such as *characterized by* and *by describing* in Descriptions and

Definitions respectively, where they hold a key status as markers of the text types in this genre of RAs.

The articles for **use**, the first argumentative word in figure 8, are disposed in the same manner: according to genre and text type (figure 10). Collocates and colligations thus supply lexical data which becomes eminent in the elicitation of such argumentative uses – the model and arrangement of the items equally follow the principles underlined in chapter 2 and recalled in figure 9 above. Figure 10 therefore accounts for such collocational data classified as argumentative in our common core scope:

USE- 30 (TXs)	(6) A1, A2, C1, E1, F3, F5
Acceptable ___ (4) / *unacceptable* ___ / *Acceptable policy* ___ / *Formulate their own*___ (1) * **Class**. *Make* ___ *of* (1) * **Concl**. *Make* ___ *of* (3) / *future* ___ / *internal* ___ *within* / *Authorized the* ___ / *Companies commonly* ___ (1) * **Defin**. *They will* ___ (2) / *The effects of that* ___ (1) * **Discuss**.	
USE -- 56 (RPs)	(7) A2, B2, C3, D2, E1, F1, F2
___ *of* (4) / *easy-to-*___ *interface* / *technologies to* ___ (1) * **Class**. *Ease of* ___ *of* (1) / *compiler chose to* ___ *only* * **Concl**. *Heavy* ___ *of* (4) / *because we* ___ / *pickles* ___ *the same* / *Conditions for its* ___ * **Defin**. *We* ___ *it here to* (1) / *plan to* ___ * **Descr**. *our* ___ *of* (3) / *setting organizations to* ___ / *their* ___ *will* (2) * **Discuss**. *The* ___ *of* (2) / *information on the* ___ *of* * **Exempl**.	
USE — 88 (RAs)	(10) A2, B1, B3, D1, F1, F4, F5, F6

The ___ of + device (3) **Class**.

The ___ of forms (2) / *tended to ___ forms* /
Profit from ___ of (1) * **Concl**.

The ___ of new technology (1) / *___ of programming tools* * **Defin**.

The ___ of hypertext (7) / *of system ___* / *research into hypertext ___* /
___ the paragraph (3) * **Discuss**.

Possible to ___ (3) / *tool to ___* / *involved in the ___ of* /
the ___ of an HTML form (1) * **Exempl**.

Figure 10: Text type-based argumentative entries.
* Where there are fewer than the six text type divisions in a given article, this is due to the absence of **use** in the type sample – in this case, both Descriptions and Exemplifications do not contain the node in the textbooks genre.

Reference to how the argumentation of the content is woven tends to be noted by the syntagmatic devices. This is especially the case of Conclusions, Definitions and Discussions across settings. Clusters such as *make use of*, *our use of*, or *use the paragraph* seem to convey information to the reader on the process of developing the exposition. Both **use** as a noun and verb can thus function alike from this persepctive.

Genre distinction, in both figures 9 and 10 above, can be pinpointed through specific phrases. An example is the contrast between Classifications in textbooks and in the other genres. The former exhibits a more concise manner of clause design with items like *acceptable use* or *unacceptable use*, appearing as graded traits in the text, whereas technical reports seem to gradually shift to longer forms: *the use of*, *easy-to-use interface*. Articles also present a greater focus on prepositional expressions such as: *the use of software*, *the use of a scanner*, etc.

The entries for the procedural node, **more**, are identified by checking the factor of adding information along a stretch of discourse. Figure 11 lists the items that denote such a function according to type and genre. As may be surmised, several items act as discourse

markers at strategic positions in the flow of speech or writing, especially those that indicate a pause or transition in the texts, appearing separated by commas, such as *and, more importantly,* (figure 11).

MORE- 22 (TXs)	(5) F[2], E[1], C[1], A[2]
Or ___ likely (2) / *and, perhaps, ___ importantly* * **Class**. *Some carry ___ weight* (1) * **Concl**. *One or ___ of the additional tag sets* (1) * **Defin**. *From one or ___ queues* (3) / *become ___ and ___ important* (2) * **Discuss**. *But, ___ important,* (2) / *___ importantly,* * **Exempl**.	
MORE -- 78 (RPs)	(15) F[6], F[4], F[2], F[1], E[1], D[2], D[1], C[3], B[2], B[1], A[2]
Become ___ and ___ important (3) / *and, ___ specifically,* / *Therefore ___ acceptable* (1) * **Class**. *Slightly ___* (1) * **Concl**. *Or ___ appropriately,* (3) / *slightly ___flexibility* / *___ likely,* (1) * **Defin**. *One or___ of the above information* (1) / *other, ___ specific* * **Descr**. *You will be hearing ___ about* (4) / *get ___ information by visiting* / *For___ information about* / *___ than ever before,* (1) * **Discuss**. *Introduce even ___ overhead* (1) * **Exempl**.	
MORE -- 66 (RAs)	(9) A2, B2, C1, D1, F1, F4, F6

For ___ focussed discussion, (1) * **Class**.

Thereby making the site feel ___ like a real bookstore (1) * **Concl.**

Defined ___ or less broadly (1) / *or the ___ abstract view* /
but they include much ___ * **Defin**.

In a ___ meaningful way (2) / *___ than just access* * **Descr**.

Learning ___ about (2) / *a ___ easily understood lecture* /
Requires ___ than technical knowledge * **Discuss**.

For ___ information, (6) / *understand ___ about* / *requires ___ learning* * **Exempl**.

Figure 11: Text type procedural entries.

Several elements combined in figure 11 include adjectival collocations – e.g. *more meaningful* -- and some others are adverbial – e.g. *more likely*. There are also a few colligations –e.g. *more about*. Verbal and noun collocates are less common – e.g. *requires more* and *more overhead*. All the forms, however, seem to share the key procedural aspect of signalling addition of information in the expository discourse.

This is achieved through varying moves, such as those encompassed by the Definitions in the genres. The tendency is to offer here **more** within coordinating clauses denoting choice, as in the connector *or, more appropriately*. Only in research articles, the discourse function of contrasting also involves the use of **more**: *but they include much more*. Other types, like Discussions, display a more uniform employment of the word in gradation and comparison phrases – e.g. *become more and more important* or *requires more than technical knowledge*. Finally, Classifications may be regarded as bearings for rhetorical standpoints of specification, for example, *or more likely*, *and more specifically*, and *for more focussed discussion*.

More examples --from figures 5, 6 and 7-- should also be dealt with briefly. These are representative type words in their distribution within only one genre, as mentioned. In this respect, for instance, a grammatical / discourse node, **such**, is selected from the RAs (figure

5), as these seem to present a greater quantity of such a lexical category – see figure 8. In contrast, **text** is analyzed as an argumentative instance from the textbook type-based items (figure 6), which render such a word as the first in the enumeration. Technical reports, in turn, register the use of the procedural pronoun **one** as often employed in the text type sub-corpus. In this manner, the three divisions of items are contrasted in figure 12, each developed within a different genre.

SUCH -- 58 (RAs)	(9) F[6], F[4], E[1], D[1], C[2], B[2], A[2]
pl. Noun + ___ *as* + items (3) **Class**. ___ *goals* (2) / ___ *funding* / ___ *motivation* / ___ *an assured basis* (1) * **Concl**. Sing. N + ___ *as* + item (2) / ___ *understanding* (1) * **Defin**. pl. Noun +___ *as* + items (4) / *many* ___ *cut-set instantiations* / ___ *a factorization* (1) * **Descr**. pl. Noun +___ *as* + pl. Noun (6) / ___ *as* + noun -ing / *the advantage of* ___ *a* / ___ *a more context-rich approach* (2) * **Discuss**. pl. Noun +___ *as* + items (4) / ___ *books* (2) * **Exempl**.	
TEXT -- 35 (TXs)	(5) F[5], F[2], E[1], C[1]
in the quoted ___ (3) / ___ *editors* * **Class**. *Does the* ___ + Verb (6) / *locate in the* ___ / *single* ___ / *shaped the* ___ / *made by the* ___ / *to which the* ___ *belongs* (1) * **Concl**. ___ *structure* (2) / *any* ___ *type* (1) * **Defin**.	
ONE -- 84 (RPs)	(10) A1, A2, B1, D2, F5, F6

___ *is that* (2) / *only* ___ *technology* (1) * **Class**.

___ *that* (2) / *the* ___ *case* * **Concl**.

From ___ *program to the other* (4) / *from* ___ *program to another* / ___ *of whose* (1) * **Defin**.

___ *or more of the above* (1) * **Descr**.

As ___ *that* (4) / ___ *key need is* / ___ *significant action* (2) * **Discuss**.

___ *using bit fields and* ___ *without* (1) / ___ *outcome is* / *justify each* ___ (1) * **Exempl**.

Figure 12: Text type samples.

The recorded constructions become relevant and useful in the types and genres where they appear. Thus, in textbooks, for example, the noun **text** is combined with elements whose composition transmits an argumentative gist – e.g. *in the quoted text* refers to previously displayed information, or *locate in the text* signals interaction with the source. In such a respect, as has been implied, the vocables belong to a given lexical category as long as they convey the directions intended. The lexeme **text**, to use the same example, may also function as a technical item in other contexts – e.g. multi-word terms such as *full-text*, *plain text* or *text-based* --, or simply without positing any discoursive characteristics –e.g. as a thematic element in explanations of *text and image*.

In reports, in contrast, regarding the appearance of the pronoun **one** in context, the main procedural technique seems to be conveyed by the possibility to offer alternatives between two elements, as in the clusters *from one program to the other*, *one or more of the above*, *one using bit fields and one without* (Definitions, Descriptions and Exemplifications). Nonetheless, in RAs, uniformity is not complete in the case of the Conclusions text type, as exposed by the use of the colligation **such**. In most cases being within exemplifying

statements *such as* + items, it mainly functions to allude to previously mentioned things in the conclusions of research articles, e.g. *such goals* or *such funding*.

The study of text type words is thus conducted by avowing to these positions of rhetorical functions in discourse. Their inclusion as common core items in our analysis is as applicable to the teaching of semi-technical lexis across disciplines as is the academic and subject-based vocabulary. Argumentative, procedural and grammar / discourse elements are therefore observed in representative contexts of use. To highlight, in this sense, the prevailing words according to the type headings, the next sub-section deals with their status as keywords in the given texts.

I.3. KEYWORDS.

The third and final level of lexical observation with text types undertaken is the exploration of the most characteristic words in a specified sample of texts. For such a purpose, each text type is selected separately across genres and contrasted with the whole corpus --used as a reference corpus-- so that the particular key items in the categories may be pinpointed.[236] Needless to say, the **Keywords** function in *WordSmith* is found to be quite practical for our aim.

The next tables (figures 13 and 14) introduce key argumentative, procedural and grammar / discourse lexemes located according to the six text types.[237] The set of statistics in figure 13 contains the only two keywords identified in the Conclusions category. The first type, Classifications, does not provide any distinctive keywords.

[236] The sources in the reference corpus include all the texts in our corpus

[237] As stated in chapter 2 --`Lexical organization according to text categorization´-- a score of 20 is regarded as the minimum indicator of keyness by which the lexical selection is measured.

N – Cat	Word	Freq. I	Concl.lst %	Freq. II	Overall. lst %	Keyness
30 — ARG	Reason-ing	11	0,10	17	0	58,0
41 -- Gram / Disc.	You	33	0,31	830	0,10	27,3

Figure 13: Conclusion type keywords. **N** = numbers referring to the positions which the given nodes occupy in the keyword list. **Cat** = word category. **Freq I** = word frequency in type. **Concl %** = percentage of word frequency in type. **Freq II** = word frequency in reference corpus. **Overall %** = percentage of word frequency in reference corpus. **Keyness** = degree of keyword position in list.

As the enumeration of items indicates, the lower the quantity in the sorting (**N**), the greater the degree of keyness to the far right of the table (**Keyness**). The **N** number is then followed by the word category assessed. Thus, the first text type element encountered is the argumentative keyword **reasoning**, with a proportionally high frequency in the Conclusions rating of 11 instances (= 0.34 % in such a sample of texts), vs. 17 occurrences in the overall reference corpus (= 0 %).[238] The resulting keyness score measured has a value of 58. This means that we are dealing with a lexical element that more than doubles the minimum value for a word to be considered key in a given sample of texts. Scott (1996) sets this lowest keyword point at 25, as the **Keyword** function shows in WordSmith.[239]

The remaining keyword batches (figure 14) illustrate how and which keywords are identified within each text type heading. As in figure 13 above, the data is collected from the text samples across all the genres and individually in each text type grouping accordingly – in agreement with figure 4's disposition.

[238] If no percentage is inserted, this means that the **Keywords** function deems this measurement as insignificant due to the zero or low rate of instances identified in the reference corpus.

[239] Academic and subject words as well as the most frequent grammar / function items (e.g. **the**, **of**, etc) are discarded in the text type keyword list.

N -- Cat	Word	Freq. I	Defin.lst %	Freq. II	Overall. Lst %	Keyness
	DEFINITIONS					
10 -- PRO	Marshaling	30	0,20	97	0,01	105,4
35 -- ARG	Genre	15	0,10	63		46,3
50 -- PRO	Types	30	0,20	445	0,06	33,2
54 -- PRO	Subtypes	10	0,07	43		30,5
58 -- PRO	Unspecified	7	0,05	14		29,9
62 -- PRO	Relation-ships	15	0,10	128	0,02	29,0
66 -- Gram. / Disc.	Whose	16	0,11	158	0,02	27,3
68 -- PRO	Component	13	0,09	106	0,01	26,1
71 -- PRO	Constituent	6	0,04	14		24,2
N -- Cat	**Word**	**Freq. I**	**Descr.lst %**	**Freq. II**	**Overall. Lst %**	**Keyness**
	DESCRIPTIONS					
33 -- ARG	Images	42	0,41	678	0,09	63,8
50 -- ARG	Handouts	5	0,05	0		43,9
56 -- ARG	Item	13	0,13	82	0,01	39,7
66 -- PRO	Inference	9	0,09	38		33,6
83 -- ARG	Image	27	0,26	658	0,08	25,0
84 -- ARG	Graph	7	0,07	33		24,8
N -- Cat	**Word**	**Freq. I**	**Disc.lst %**	**Freq. II**	**Overall. Lst %**	**Keyness**
	DISCUSSIONS					
25 -- ARG	Tag	65	0,23	267	0,03	128,5
28 – Gram. / Disc.	Our	113	0,40	872	0,11	120,9
39 -- PRO	Level	62	0,22	503	0,06	62,2
46 -- ARG	Rules	29	0,10	119	0,01	57,3
49 -- ARG	Facts	19	0,07	44		54,0
55 -- ARG	Measures	18	0,06	56		43,1
56 -- ARG	Comments	25	0,09	122	0,02	43,0
60 -- ARG	Procedures	28	0,10	161	0,02	41,4
61 -- ARG	Tags	28	0,10	162	0,02	41.2
62 -- ARG	Activity	24	0,08	119	0,01	40,7
69 -- ARG	Heading	14	0,05	38		36,4
85 -- ARG	Body	18	0,06	94	0,01	29,1
90 -- ARG	Proposal	12	0,04	38		28,4
94 -- ARG	Rule	18	0,06	105	0,01	26,2

N -- Cat	Word	Freq. I	Exempl. lst %	Freq. II	Overall. Lst. %	Keyness
EXEMPLIFICATIONS						
33 -- PRO	Similar	37	0,18	245	0,03	66,4
34 -- ARG	Heading	18	0,09	38		64,4
73 -- Gram./ Disc.	You	49	0,24	830	0,10	25,6

Figure 14: Text type-based keywords.

From the review of the samples in figure 14, a generalization may be inferred concerning these distinctive sets of keywords in the text types of our corpus. There seems to be an inclination for each grouping to manage a certain class of items uniquely. In other words, a collection such as the Discussion sections contains a characteristic set of argumentative devices that appear as key items in the arrangement of the content. Something similar occurs in the case of Descriptions. In contrast, the linguistics resources prevailing in Definitions are procedural words. Exemplifications, like Conclusions in figure 13, yield fewer items.

Some examples of this variation in use for each one of the salient types mentioned – Definitions, Descriptions and Discussions – may be afforded by examining the most representative words --i.e. having the highest degree of keyness. Thus, the form appearing in Definitions, **marshaling** functions as a procedural lever in clusters such as *our marshaling code* (3 instances), *marshaling support for* (2), and *invented for marshaling purposes* (1). It is also closely used as a conceptual element in the texts –mainly in a report on object-oriented design (Computer engineering [A2]); yet, the reference provided to the arrangement of the technical developments in the text makes the keyword occur as procedural, a resource employed by the author to align concepts.

Images and **tag**, viewed as argumentative in Descriptions and Discussions respectively (see figure 14), form other types of combinations, such as *paging images below* (4), *a number of images* (2), and *three images were produced* (1) in the case of the former. The latter keyword co-occurs in contexts like *heading # tag* (6), *use the paragraph tag* (3), and *this tag indicates* (2). We may thus glean here a choice of these items as both text type specific and carriers of the information. In other words, these and the data presented in figures 13 and 14 can be regarded as bearing a greater load of specificity as text type-based vocabulary, functioning as keywords of specific textual units, i.e. text type-determined content.

This study enables a rich perception of lexis in a more reduced context, and it thus provides a perception of the different degrees of common core lexis that are dealt with in our research. The scope should therefore take the important notions of word frequency, range and textual dimension into account. Throughout the remaining results, such a conception of common coreness, in which corpora coverage plays a crucial role in determining the type of lexical items yielded, prevails and recurs for each section (genre and subject – based words).

II. GENRES

In this next scope of lexical work --the genre--, the three sub-corpora examined in chapter 2 --research articles, textbooks and reports-- are taken both independently and jointly to serve as the framework in which to view academic and subject lexis. In reviewing frequency and range lists, for example, each genre is autonomous and comparable in relation to the rest, whereas collocations are sought across the whole collection of texts. Keywords, like frequency, should also be checked in the three individual sub-domains so that

characteristic lexis may be inferred.[240]

II.1. WORD FREQUENCY AND RANGE.

A valuable tool for the easy and quick preview of these two parameters is the Detailed Consistency analysis (DCA), providing counts of each item within the three given sub-groups. In our search for the most common words, according to the DCA procedure, one sub-group is utilized as first reference for the sorting of items. This means that *WordSmith* then takes the lexis appearing in this genre as the guiding exponent for the arrangement of entries. In this respect, research articles (RAs) are chosen, since we deem that their academic position as `high-brow´ settings makes them sufficiently characteristic of disciplines and subject fields.

From such a line of work, for instance, the first item appearing below (figure 15) --**information**-- is the number one sub-technical content word in terms of frequency and range in the whole corpus --position 20. Its frequency in research articles (614) directs this count; yet, it is also the top lexeme in the other two settings from the perspective underlined: 967 and 942 instances in books and reports respectively. Its lexical range is thus also significant, since before the high frequency detected in each genre, the overall or total count is taken as first priority of selection.

Figure 15 and 16 illustrate the contrast between the ten most and least frequent lemmas in our corpus respectively, according to the detailed consistency test. As can be noticed, the lowest possible number of instances is one in the research articles so that range can be considered across all genres. These low frequency items are called `hapax legomena´(James, 1994: 23), and should be often taken into account, namely in terms of

[240] As can be deduced, requirements for the specification of textual environment are constant in that the genre setting determines the parameters of selection.

academic application – e.g. utilization in a specific context of use.

The genre-based frequency list of sub-technical vocabulary included in Appendix 4 ends above the one-count position: to be more exact, at the ten-texts mark – any item appearing in a lower number than ten sources within a given genre is thus discarded, regardless of its frequency in a different genre. This is done in order to encompass only those words more evenly distributed across all three genres. We deem that ten is a high enough figure for our selection of common core sub-tecnical items. The genre-based frequency list in appendix 4 includes, in addition, reference to the number of texts where the words are found

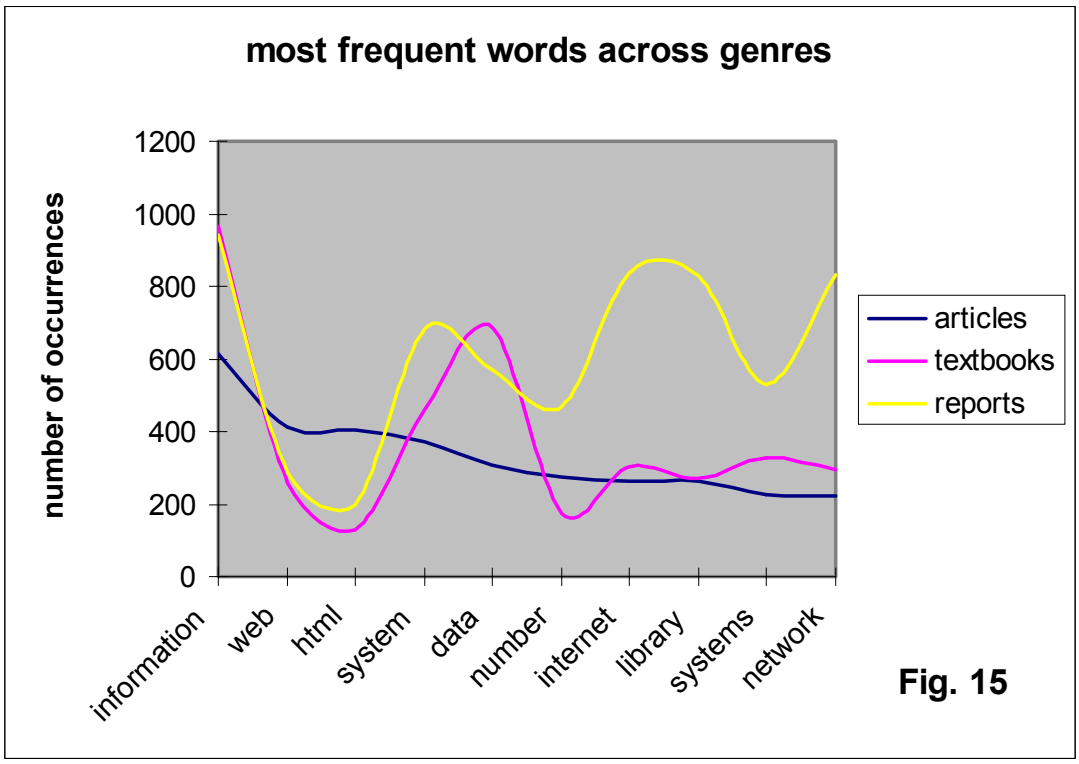


Fig. 15

(indicated under **NS** – meaning `number of sources´).

The amount of occurrences that regulate or lead the lexical lists are the ones found in research articles, as the two graphs (figures 15 and 16) show. In the first case, the curve goes from a top limit of 614 instances for the item **information**, and decreases thereon. In figure

16, in contrast, the low line is steady, indicating that there is only one appearance of the terms registered in the sub-corpus C (Research articles) setting.[241]

Lexical range can thus be attested by examining both the amount of items across the three genres and the quantity of sources that contain the specified words. These are mainly academic words, as subject and text type entries are more specific in narrower sets of texts distributed according to either subject / topic (see Appendix 1) or functional parameters. [242]

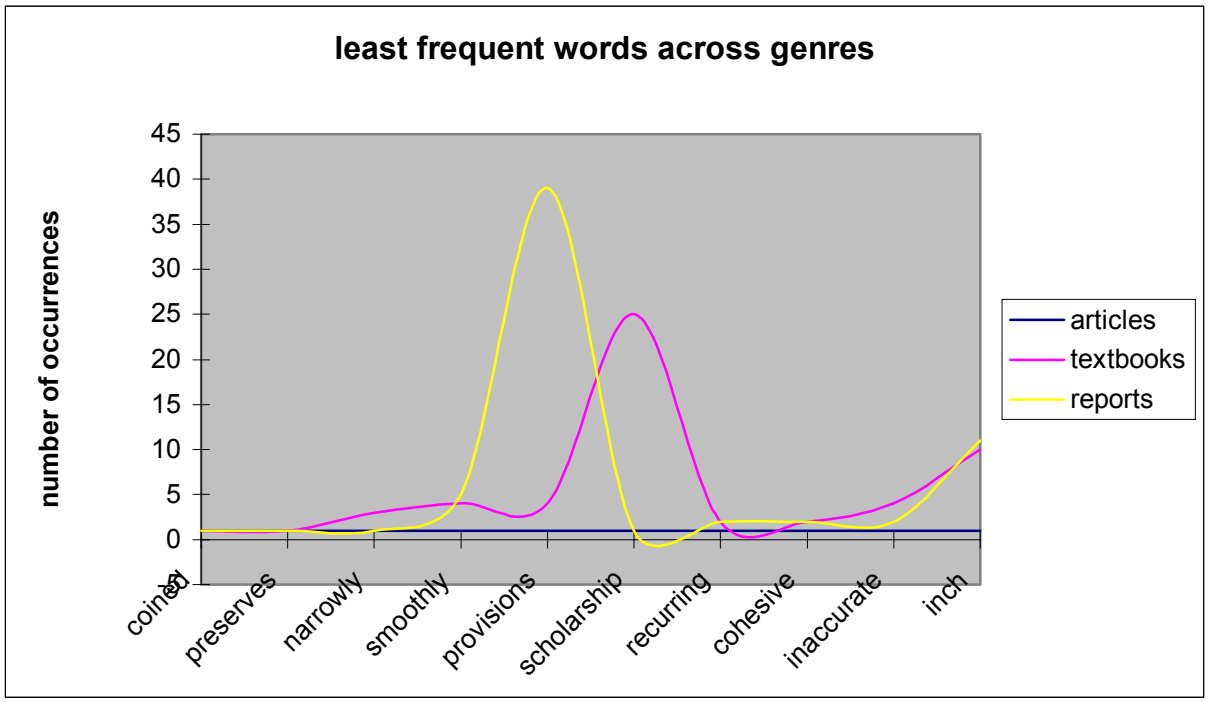


Figure 16

[241] These low frequency words are not, nonetheless, the sole hapax legomena occurring in our corpus, since others appear with only one instance in each genre; these 1 / 1 / 1 registers are the following: **endorsement**, **morphological**, **competent**, **interactively**, **guard**, **glance**, **intermediaries**, **hitherto**, **instantaneous**, **etiquette**, **lean**, **bend**, **leisure**, **notations, hindered**, **manifestations**, **capitalize**, **comply**, **Montreal**, **crafting**, **Ameritech**, **neighborhoods**, **newcomer**, **beats**, **clicks**, **netcom**, **buzzword**, **inject, donate, Philadelphia, flawed, accumulating, concise, onerous, elusive, cognizant, millisecond, motor, coined, Croatia, believing, fundraising, assuring, breath, attracting, circulated, equivalently, orange, catalyst, Nw, algorithmic, abstracted, diminish, assuring, crafting** and **neighborhoods**.

[242] Subject-based vocabulary is seen in more detail in the next section, although their listing is conducted together with academic vocables in Appendix 4 due to their same criteria of selection across genres.

Two examples of lists representing academic genres at university level are published in Xue and Nation (1984) and Coxhead (1998). Due to its updatedness and significance, the latter may replace the former in terms of common coreness (Coxhead, 1998: 1). In fact, several key items in this work, Coxhead's *Academic Word List* (AWL -- 1998), are also detected in our academic word list (appendix 4) as lemmas – i.e. most frequent in their lexical families. Coxhead thus compiles, for instance, the plural noun **issues** as more widely used than the singular **issue** among academic texts. The same occurs in our corpus, where **issues** is ranked number 122 and **issue** is placed in 177th position -- see appendix 4.

Some other similar examples originating from such a contrastive perusal include the following items – located in their corresponding slots in our list according to frequency and range of use: **data** (# 36), **access** (#60), **design** (#92), **file** (#108), **environment** (#183), **approach** (# 203), **features** (#207), **display** (#335), **identified** (#431), **equipment** (#478), **comments** (#516), **ensure** (#601), **infrastructure** (#619), **items** (#645), **corresponding** (#658), **authority** (#791), **label** (#831), **achieve** (#840), **assigned** (#873), **enable** (#937), **emphasis** (#1267), **brief** (#1300), **aid** (#1352), **enhanced** (#1624), **assembly** (#2120), **highlighted** (#2141), **definite** (#2366), **assurance** (#2936), **conceived** (#3471). They all belong in both Coxhead's work and our Information Science and Technology corpus as head-words in terms of frequency.

As may be checked, most are nouns – **data**, **access**, **design**--, also occurring quite often as infinitive verbs – **access**, **design**. In our corpus, this happens in several cases, although subject-based items tend to contain the highest amount of nouns. Among academic lexis, in turn, as may be gleaned from the contrast above, in addition, there are also many lemmas presenting past forms, such as **identified**, **assigned** and **highlighted**.

II.2. COLLOCATIONS

The range of academic nodes and their collocates are analyzed within this scope. Most uses, as shall be examined, are those of content words collocating with other content items and function markers. They are listed according to the genre distribution and regarded as academic. In this respect, common core vocables receive academic tagging when viewed in large contexts in the genres and not just in reduced sub-settings such as topics, subjects or text types.

The following ten samples (# 1 – 10) correspond to the five most and least frequent words as shown in figures 15 and 16 above. For examining the lexical arrangement of 20 other entries, see Appendix 5.

Under the first lemma –**information** (sample # 1) --, the amount of multi-word units and clusters are compiled according to the criteria specified in chapter 2: The top sub-section displays what we have denominated restricted associations (symbolized by ®) whereas the lower part contains free word combinations (₣). As mentioned in the previous chapter, this distinction is mainly assessed by observing the number of texts where the collocation occurs. In the case of free combinations, these tend to appear across a higher number of texts and topics. In addition, we use an asterisk again to indicate the low frequency (below three instances) of a given collocation that is nonetheless relevant in the text (discipline abbreviated in brackets).

Finally, word-forms of the lexemes shown in bold are also included as long as they are employed and substantially collocate across the three genres. Within **information**, for instance, the adjective informational and the verb forms informed and inform are attached;

this is done in a manner resembling Lyons' account of entry relations (Lyons, 1977: 516), or Bauer's conception of word forms (Bauer, 1983: 198).[243]

INFORMATION.-
TXs RPs RAs
967 / 942 / 614
[244]

®
networked ___ *services* (11) / ___ *object* / *hypertext* ___ /
___ *on the web* / ___ *content* / ___ *dimensions* /
___ *networks* / *reference* ___ / ___ *society* /
___ *structures* / *department of* ___ / *reel* ___ /
___ *search* / *browsing the* ___ / ___ *world* /
___ *sharing* / ___ *products* / *format the* ___ /
___ *contact* / *access* ___ *via* / ___ *administration* /
___ *relating to* / ___ *stored* / ___ *created* /
___ *creators* / *geographic* ___ / *scholarly* ___ /
___ *transfer* / ___ *center* / *descriptive* ___ /
___ *handling* / ___ *accessible* / ___ *server* /
deliver ___ / *link* ___ / *managing the* ___ /
retrieve ___ / ___ *servers* / *share* ___ /
encoding the ___ / ___ *age* / *dissemination of* ___ /
review the ___ / *browse the* ___ / *target* ___ *resources* /
___ *commodities* / *disseminate* ___ / *fomatting* ___ /
inaccurate ___ / ___ *market* / *recorded* ___ /
reporting ___ / ___ *modeling* / ___ *rules* /
___ *storage* / *textual* ___ / *joint* ___ /
audio ___ / *exchanging* ___ / *exchanges* ___ /
___ *professionals* / *repository of* ___ / *evolving* ___ /
personal ___ (2) * All Tx / ___ *published* (2) * All Tx /
___*filtering* (2) *Tel. RA /

Ŧ
*Digital*___ (181) / ___ *systems* / ___ *system* /
electronic ___ / ___ *technology* / ___ *services* /
___ *structure* / ___ *span* / ___ *objects* /
___ *resources* / ___ *sources* / *provide* ___ /
___ *about* / *more*___ / ___*from* /
network ___ / ___ *available* / ___ *science* /
library and ___ *science* / ___ *gathering* / *preserving the* ___ /
searching ___ / *using* ___ / *knowledge and* ___ /
specific ___ / *additional* ___ / *distribution of* ___ /

[243] As can be noticed, these less frequent words are underlined so that a distinction from the lemma (in bold) may be drawn.

[244] In entries, as opposed to the frequency detailed listing, the order of the genres is, as shown above and in chapter 2, that of textbooks first, then reports and last articles.

___ *regarding* /	*key* ___ /	___ *over* /
relevant ___ /	___ *needs* /	___ *studies* /
___ *retrieval* /	___ *service* /	___ *providers* /
___ *within* /	___ *technologies* /	___ *provided* /
exchange + ___ /	*report +*___ /	___*found* /
___ *highway* /	___ *required* /	___ *related* /
finding ___ /	___ *processing* /	*witness* ___ /
___ *superhighway* /	___ *environment* /	*find* ___ /
card ___ /	*external* ___ /	*further* ___ /
useful ___ /	*amount of* ___ /	___ *contained* /
send ___ /	*topics on* ___ /	*issues on* ___ /
bibliographic ___ /	*detailed* ___ /	*following* ___ /
online ___ /	*loss of* ___ /	*identifying* ___ /
certain ___ /	___ *concerning* /	*business* ___ /
___ *seeking* /	*specify +*___ /	*supports* ___ /
considering ___ /	*helpful* ___ /	*ordering* ___ /
organizing ___ /	___ *presented* /	*reveals* ___ /
___ *travels* /	*financial* ___ /	*sort +* ___ /

INFORMATIONAL.- 2 / 12 / 6

___ *aspects* (4) / *functional and* ___ / ___ *content* /
___ *resources* / ___ *model* (2) * Lib. RP /

INFORMED.- 10 / 10 / 3

Were overall ___ (3) / *better* ___ / ___ *with regard to* /
make ___ *decisions* (2) * All TX

INFORM.- 2 / 6 / 8

to ___ *the reader of* (3) / *to* ___ *the browser the* (2) *Lib. TX /

Sample # 1

As can be observed, the distinction between free and restricted collocations is not always made. In a vast body of lexical items, such as those posted under the noun **information**, in fact, a term like *information rules* seems to play a more specialized semantic role than, for instance, the colligation *information about*, since the former co-occurs in a more thematically constrained environment. In contrast, under the word forms – informational,

informed, and inform --, all the combinations tend to behave as restricted collocates, in bound textual settings (e.g. a Library science report in the case of the adjective informational).

Let us examine next, in opposition, the entry of one of the the least frequent genre-based academic items – considered hapax legomena (James, 1994). The one chosen is **coined** (sample # 2), which has no enclosed higher-frequency word-forms, and yet, it occurs significantly in a couple of restricted collocations such as the two identified in a Documentation report and Telecommunication article:

COINED.- TXs RPs RAs 1 / 1 / 1
___ *the term* (1) Tel. RA / ___ *the word* * Doc. RA

Sample # 2

The remaining eight articles (samples # 3 – 10) are also contrasted according to the priority of frequency / range. As can be noted, some of the items, such as preserves below (# 4), belong as word-forms in their lemma articles: In the case of the one mentioned, its noun heading, **preservation**, includes up to four verb forms that do not occur too often but still appear rather distributed across the sub-corpora.

WEB. - 258 / 286 / 414 TXs RPs RAs
® ___ *writing* (17) / *writing for the* ___ / *information is available on the* ___ / ___*-to-part* / ___ *document* / ___ *site* / ___ *weaving* / ___ *page* / ___ *space* / ___ *interface* / ___ *browsing* / ___ *access* / *sites on the* ___ / ___*-based applications* / ___ *proxy servers* / *available via the* ___ / ___ *resources* (2) * Lib. TX /

Ŧ
world wide ___ (126) / *available on the ___* /*___ browser* /
___ browsers / *___ server* / *___ servers* /
___ documents / *___ sites* / *___ pages* /
___ accessibility / *using the ___* / *library ___* /
___ authors / *___ client* / *___ content* /
___ based / *use of the ___* / *___ software* /
through the ___ /

Sample # 3

PRESERVATION.-
TXs RPs RAs
153 / 22 / 35

®
___ and access (16) / *digital ___* / *commission on ___* /
___ of digital information / *video ___* / *the ___ community* /
documentation and ___ / *film ___* # / *moving image ___* /
term ___ / *___ principles* (2) * All TX

Ŧ

the ___ of (17) / *for the ___ of* / *___ as* /

PRESERVE.- 42 / 5 / 8

®
___ the integrity of (7) / *___ digital information* / *efforts to ___* /
___ sharing between /

Ŧ

to ___ (12) / *attempt to___* /

PRESERVING.- 58 / 9 / 1

___ digital information (21) / *___ the integrity* / *problems of ___* /
responsibility for ___ (2) * All TX

PRESERVED.- 15 / 6 / 3

®

Catalogued and ___ (1) * A. Com RA/

Ŧ

To be ___ (3) / ___*for future use* /

PRESERVES.- 1 / 1 / 1

Generates, ___ *and improves access* * All TX /
___*, maintains and services records* * All RP /
___ *the label* * Com. RA

Sample # 4

HTML.-
129 / 199 / 406
TXs RPs RAs

®

a ___ *document* (21) / ___ *page* / ___ *tools* /
___ *tools editors* / ___ *writing tools* / *markup language* ___ /
___ *versions of* / ___*files* / *the* ___ # /
composing good ___ /

Ŧ

___ *provides* (6) / ___ *documents* /

Sample # 5

NARROW.-

13 / 4 / 6

The interface is reasonably ___ (3) / ___ *the search* (2) * Doc. RA /
___ *down* (2) A. Com. TX * / ___ *domain* (2) * Tel. RP

NARROWLY.- 3 / 1 / 1

___ *standardized as* (1) * All TX / ___ *speaking* * A. Com. RA /
___ *prespecified* * Doc. RP

Sample # 6

SYSTEM.-
TXs RPs RAs
459 / 682 / 373

®
writing ___ (31) / *writing ___ declaration* / *telephone ___* /
management ___ / *handle server ___* / *knowledge ___* /
object ___ / *distributed programming ___* / *TEI ___* /
filing ___ / *___ design* / *___ of constraints* /
___ variables / *___ designer* / *optic ___* /
communication ___ / *rendering ___* / *document management ___* /
___ entities / *computer ___* / *scanning ___* /
___ tools / *a mixed ___* / *entering the ___* /
technology ___ / *___ interfaces* / *messaging ___* /
a realtime ___ / *___ issue* / *transport ___* /
monthly payment ___ / *search ___ presents* / *a hypertext ___* /
a ___ meets the conditions / *telecommunication ___* / *a ___ entity* /
___ -specific / *___ whose architecture* /

Ŧ
information in the ___ (61) / *our ___* / *information ___* /
operating ___ / *___ use* / *the entire ___* /
users in the ___ / *expert ___* / *library ___* /
distributed ___ / *aim of ___* / *___ designed to* /
___ combined / *___ involved* / *creating a ___* /
___ using / *___ development* / *___ is providing* /
___ described / *___ then* / *the ___ uses* /

SYSTEMS.- 328 / 531 / 225
®
___ communication (15) / *based ___* / *knowledge-based ___* /
distributed ___ communication / *knowledge ___* / *realtime distributed ___* /
intelligent transportation ___ / *hypertext ___* / *imaging ___* /
___ engineering / *management ___* / *heterogeneous ___* /
computer ___ / *computer ___ review* / *electronic ___* /
coupled ___ / *loosely coupled ___* / *tightly coupled ___* /
retrieval ___ / *operating ___ principles* / *___ design* / *radar ___* /
diolocation ___ / *communication ___* / *storage ___* /
file ___ / *mapping ___* / *fiber-optic ___* /
indexing ___ / *publishing ___* / *communications ___* /
traveler information ___ / *___ that support* / *___ that conform* /
Ŧ

information ___ (99) / *distributed ___* / *expert ___* /
database ___ / *___for design* / *access computer ___* /
___ are built / *operating ___* / *___ hold* /
___ require / *available in ___* / *library ___* /
our ___ research / *such ___ work* / *___ have evolved* /

Sample # 7

SMOOTH. –
TXs RPs RAs
2 / 7 / 6

___ *and successful coordination* (6) / ___ *transition* (2) * All RP /
___ *image* (1) * A. Com. Rp / ___ *texture* (1) * A. Com.. RP /

SMOOTHLY.- 4 / 5 / 1
Transmitted ___ (3) / *running* ___ (1) * All TX /
Progressing ___ (1) * All RP

Sample # 8

DATA.-
TXs RPs RAs
685 / 570 / 309

®
___ *collection instruments* (13) / ___ *transfer operations* /
___ *representation services* / ___ *structures* / *broadcast* ___ *transfers* /
performance ___ / ___ *type* / *an abstract* ___ *type* /
___ *flow* / ___ *users* / *processing of* ___ /
modulate video ___ / ___ *modelling* / *recovery of the* ___ /
___ *are recovered* / *support for* ___ *representation* / ___ *service* /
blocks of ___ / ___ *storage* / ___ *transfer procedures* /
collected ___ / ___ *encoding techniques* / *encryption* ___ /
unit of ___ / *units of* ___ / *amounts of* ___ /
source of ___ / *network* ___ / *abstract* ___ *source* /
abstract ___ *sink* / ___ *types* / ___ *channels* /
buffered ___ / ___*fields* / ___ *stream* /
draft standard ___ / ___ *is copied over* / *copies* ___*from* /
discards any ___ / *for reading* ___ / *for writing* ___ /
outstanding ___ / *range of* ___ / *sends the* ___ /
interface ___ / *electronic* ___ *interchange* / *marshal the* ___ /
generic ___ / *a* ___ *record* / *bulk* ___ *transfer* /
view of ___ /

Ŧ
of the ___ (75) / ___ *transfer* / ___ *collection* /
___ *representation* / *exchange of* ___ / *exchange* ___ /
change in ___ / *amount of* ___ / ___ *is transferred* /
transmission of ___ / *loss of* ___ / ___ *bases* /
gathered observational ___ (2) * All RP & Tel. TX

Sample # 9

PROVIDE.-
TXs RPs RAs
492 / 205 / 201

®
___ *a mechanism* (88) / ___ *access to* / ___ *support for* /
implementation to ___ / *to* ___ *access* / *to* ___ *information* /
___ *a function that* / *constraints* ___ *criteria for judging* /
___ *explicit support for* / *interface* ___ *support* / *interface to*___ /
plates to ___ / ___ *a delay* / ___ *a function* /
___ *guidelines for* / ___ *information structures* / ___ *performance data* /
___ *the basis* / ___ *the original* / ___ *the interface to* /
to ___ *performance* / *to* ___ *unmodulated light in* / ___ *an indication* /
___ *an overview* / ___ *written instructions* / ___*facilities* /
___ *a broadcast* / ___ *a capability* / ___ *a default* /
___ *a framework* / ___ *a means* / ___ *a network* /
___ *a prioritization* / ___ *a thread* / ___ *an identifier* /
___ *an inquiry* / ___ *an interface* / ___ *direct support* /
___ *equal access* / ___ *explicit visibility* / ___*for endpoint* /
___*for event* / ___*functionality* / ___ *inquiry functions* /
___ *mappings that* / ___ *new implementations* / ___ *operations for* /
___ *queueing behavior* / *customers who* ___ / *disciplines to*___ /
___ *reliable distributed systems* / ___ *separate requirements* / ___ *services to* /
___ *the threads* / ___ *the training* / ___ *the transmission* /
___ *type checking* / ___ *alternative ways* / ___ *basic capabilities* /
___*for communication* / ___ *the above information* / ___ *the battlefield for* /
___*you with* /

Ŧ
shall ___ (165) / ___ *the ability to* / *in order to* ___ /
required to ___ / *able to* ___ / *used to* ___ /
___ *the capability* / *to* ___*for* / *ability to* ___ /
library will ___ / *need to* ___ / ___ *the necessary* /
first ___ / *intended to* ___ / *designed to* ___ /
___ *information about* / ___ *the necessary background* / *we first* ___ /
we ___ *the* / *who* ___ *the* / *can* ___ *a* /
___ *a more* / ___ *the best* / *should we* ___ /
was to ___ / *appropriate to* ___ / *is expected to* ___ /
___ *a common* / ___ *a way* / ___ *an excellent* /
___ *users with* / *to* ___ *additional* / *want to* ___ /
___ *an ability* / ___ *approximations to* / ___ *the reader with* /
___*us with* /

PROVIDES.- 158 / 152 / 83

®
___ *reference #* (18) / ___ *an overview* / *interface* ___ *operations* /
proposed method ___ / ___ *performance* / ___ *access to* /
___ *a mechanism* / ___ *support for* / *the binding* ___ /
html ___ *the* / *interface* ___ *routines to* / ___ *a model* /
___ *a table* / ___ *for filtering* / ___ *functionality* /
program ___ / ___ *a summary* / ___ *a way* /
___ *separate procedures* / ___ *strong type checking* / *runtime* ___ /
sgml ___ / *registration* ___ / ___ *threads* /
___ *details on* / ___ *discussion* / ___ *information about* /

Ŧ
also ___ (16) / ___*for* / ___ *information* /
___ *the ability to* / ___ *us with* / *this paper* ___ /
this section ___ /

PROVIDING.- 89 / 87 / 52

®
___*public access* (4) / *in* ___ *access* / *of* ___ *access* /
___*functionality* / ___ *support for* / *costs of* ___ /
system is ___ /

Ŧ
___ *access to* (18) / *in* ___ / *by* ___ /
for ___ / *in addition to* ___ / *for* ___ *a* /
such as ___ / *at* ___ / *emphasis on* ___ (2) * Tel RA & Lib. TX

PROVIDED.- 68 / 141 / 72

®
Will be ___ *by an implementer* (24) / ___ *by the agent* / *principle* ___ /
___ *solution to* / *services are* ___ / *standard* ___ /
___ *in support* / ___ *special marshaling* / *pipes* ___ /
Documentation ___ *by* / ___ *a list* / *table* ___ *by* /
name is ___ / ___ *to you* /

Ŧ
___ *by the* (31) / *is* ___ *in* / *can be* ___ /
are ___ *to* / *may be* ___ / ___ *as* /
should be ___ / *also* ___*for* / *has been* ___ /
therefore ___ / ___ *to them* /

PROVIDER.- 112 / 37 / 11

®
Communications ___ (18) / *underlying service* ___ /
Communications service ___ / ___ *shall attempt* /
request the service ___ *to reclaim resources* /
request the service ___ *to deallocate resources* /
used by the service ___ *to identify* / *used by the service*___ *to delete* /

₣
Service ___ (54) / *internet service* ___ /

PROVIDERS.- 41 / 52 / 22

®
network service ___ *(5)* / data among service ___ / ____ would require /
Online service ___ /

₣
Internet service ___ (14) / *by service* ___ /

PROVISION.- 5 / 21 / 28

®

___ *of information structures* (16) / ___ *of organizational links* / *special* ___*for* /

₣

the ___ *of* (30) / ___ *in the* /

PROVISIONS.- 4 / 39 / 1

®

___ *of section #* (8) / ___ *of title* / *penalty* ___ /
license ___ / *service* ___ (1) * All TX

₣

notice ___ (2) * All RP & Tel. RP

Sample # 10

As may be checked, samples # 4, 6, 8 and 10 –the least frequent occurrences—are all parts of more highly employed lemmas – **preservation**, **narrow**, **smooth**, and **provide** respectively. Especially in the case of this last lexeme, the word form provisions follows all

the derivations of a main word -- **provide**.

From the collected data, we might tend to surmise that the more specialized multi-word units are placed in the first section of restricted collocates, whereas the less subject-specific clusters belong to the subsequent set of free combinations. In fact, by contrasting, for instance, the entry for provided (sample # 10), we can notice that function or grammar constructions appear in the F box, while restricted multi-word units occupy the top level. This, however, is not the generality in our approach. A clear way of viewing exceptions to this assumption is by regarding the same element in either the singular or plural form. Thus, *amount of data*, for example, in sample # 9, is labelled as free, whereas *amounts of data* is considered restricted. This is exclusively due to the location of the first item in a broader context: two Telecommunication articles, one Audio-visual article and one textbook on perspectives on Information. *Amounts of data*, in contrast, is only utilized in a Telecommunication report on Fiber Optics technology (D1). Due to the fact that the percentage for determining the difference between restriction and freedom of employment in our corpus has been set in chapter 2 at a minimum of 0.3% (every ten instances, at least three texts must be involved), such a requirement is fulfilled in the first case (in five lines, four texts are encompassed – 0.8%), but not in the second circumstance (four co-texts from the same source – 0.25%).

In this respect, even lower frequency items, such as the expression *gathered observational data* (sample # 9) can be identified as free combinations. In this case, this is due to the fact that the lexical item appears twice: once in an Internet report (F6 subject division) and the other in a textbook on Communication theory (E1) – a percentage of 1.0% in our cut-off point scale. As a result, contextual information offers a way to classify the item

as either more or less restricted in its use.[245]

II.3. KEYWORDS

The inspection of genre-based keywords is included for the main reason of collecting those words that, unnoticed in the detailed consistency sift (DCL), become prominent in any single genre. The reference corpus – which includes the whole body of texts in our corpus –, serves as the object of comparison to elucidate these items.

The following (figure 17) is thus the list of keywords which encompasses the key academic vocables in the setting of textbooks.

N	Word	Freq. I	Books	Freq. II	Refer.	Keyness
3	Requirement	385	0.12	428	0.05	149.5
6	Message	489	0.15	683	0.08	114.5
7	Element	384	0.12	514	0.06	100.3
9	Codes	247	0.08	273	0.03	96.9
10	Texts	278	0.09	336	0.04	91.5
16	Signs	176	0.05	179	0.02	79.8
17	Archives	209	0.07	237	0.03	78.2
19	Application	465	0.15	730	0.09	77.0
21	Attribute	207	0.06	237	0.03	76.0
22	Sign	171	0.05	179	0.02	73.9
23	Digital	671	0.21	1,184	0.14	71.2
24	Language	446	0.14	715	0.08	68.7
29	Group	342	0.11	520	0.06	62.7
31	Elements	359	0.11	559	0.07	61.0
33	Transfer	188	0.06	233	0.03	58.7
36	Ability	231	0.07	321	0.04	54.8
40	Discussion	269	0.08	400	0.05	52.8
41	Model	244	0.08	351	0.04	52.7
43	Ibid	114	0.04	118	0.01	50.2
44	Entity	139	0.04	162	0.02	49.3
47	Encoding	164	0.05	214	0.02	45.5
49	Messages	185	0.06	255	0.03	44.8
50	Apparatus	95	0.03	98	0.01	42.1

[245] In addition, introspection, a much needed instrument in Corpus Linguistics analysis –see chapter 1 (`Corpora´)--, may serve, in some cases, to predict that some clusters are used across a wide range of texts and topics – e.g. grammar constructions such as those exposed in sample # 10.

51	Nov	98	0.03	103	0.01	42.1
54	Working	253	0.08	406	0.05	38.8
55	Provide	485	0.15	908	0.11	38.8
56	People	367	0.11	651	0.08	38.0
61	Social	158	0.05	223	0.03	36.0
62	Mechanism	166	0.05	241	0.03	34.9
63	Preservation	150	0.05	210	0.02	34.9
66	Witness	76	0.02	77		34.7
68	Net	201	0.06	316	0.04	33.1
71	Receive	119	0.04	158	0.02	31.7
74	Operation	160	0.05	239	0.03	31.0
76	Witnesses	65	0.02	65		30.3
81	Meaning	113	0.04	152	0.02	29.1
82	Send	126	0.04	177	0.02	29.0
87	Document	474	0.15	934	0.11	28.4
88	Declaration	90	0.03	111	0.01	28.4
92	Choice	120	0.04	169	0.02	27.5
93	Provider	113	0.04	156	0.02	27.3
96	Binding	91	0.03	116	0.01	26.7
98	Marked	89	0.03	113	0.01	26.3
99	Scholars	78	0.02	94	0.01	25.8
100	Specify	104	0.03	142	0.02	25.8
102	Costs	174	0.05	283	0.03	25.5
103	Scheme	157	0.05	248	0.03	25.4
103	Protocol	181	0.06	298	0.03	25.3
105	Recommended	96	0.03	128	0.01	25.3

Figure 17: Genre-based Keywords in textbooks. **N** = ranked position in list / **Freq. I** = instances in textbooks / **Books** = percentage in textbooks / **Freq. II** = occurrences in reference corpus / **Refer.** = percentage in reference corpus.

The results are valued in terms of both keyness score and range of texts where the words co-occur. In fact, in the list, subject-based words may also show up, but their pairing with limited sets of texts leaves them out in this search. Such is the circumstance of, for instance, words like **interface**, **endpoint**, and **semiotics**, originally positioned among the top six keywords in figure 17. Their prevailing occurrence in texts dealing with the same subject postpones their examination to the next section on Subject words: while the first and second nouns mainly appear in a Telecommunication textbook on data communications (A2), the other node, **semiotics**, co-occurs 100% of the time along an Audio-visual communication

source on Content Analysis (C1). The focus is thus placed on the more broadly developed semi-technical academic lexis.

Figure 18 describes the group of technical reports keywords:

N	Word	Freq.	Reports %	Freq.	Refer. %	Keyness
4	Library	727	0.26	1,206	0.14	159.2
8	Call	309	0.11	397	0.05	122.5
10	Network	699	0.25	1,308	0.15	103.7
12	Libraries	408	0.15	662	0.08	95.0
13	Type	509	0.18	904	0.11	90.4
14	Transport	156	0.06	162	0.02	87.9
15	Copyright	238	0.08	318	0.04	87.4
22	Owner	142	0.05	151	0.02	77.3
23	Address	317	0.11	508	0.06	76.3
26	Handle	170	0.06	210	0.02	72.4
27	Public	432	0.15	784	0.09	71.4
31	File	342	0.12	590	0.07	66.6
33	Exception	138	0.05	162	0.02	64.2
37	Space	239	0.09	377	0.04	60.0
43	Server	296	0.11	518	0.06	55.1
44	Internet	540	0.19	1,112	0.13	53.9
46	Method	295	0.11	525	0.06	52.0
48	IEEE	156	0.06	221	0.03	50.6
49	Connec-tions	132	0.05	174	0.02	49.8
50	Procedure	181	0.06	277	0.03	49.1
51	Calls	135	0.05	183	0.02	48.1
55	Reader	143	0.05	204	0.02	45.7
57	Remote	156	0.06	232	0.03	45.3
58	Functional	156	0.06	232	0.03	45.3
66	Protection	81	0.03	87	0.01	43.4
69	Raises	89	0.03	104	0.01	41.7
70	Services	322	0.11	627	0.07	41.0
71	C	500	0.18	1,074	0.13	40.5
73	Disk	145	0.05	221	0.03	39.7
74	Streams	83	0.03	96	0.01	39.6
76	Concrete	83	0.03	99	0.01	37.6
78	Failure	84	0.03	101	0.01	37.5
79	Registered	77	0.03	88	0.01	37.5
83	Throws	67	0.02	70		37.4
86	Argument	100	0.04	133	0.02	37.0
87	Naming	84	0.03	102	0.01	36.9
90	Standards	278	0.10	538	0.06	36.2

94	Light	113	0.04	163	0.02	35.2
95	Applica-tions	262	0.09	504	0.06	35.0
98	Returns	88	0.03	114	0.01	34.3
100	Training	128	0.05	198	0.02	33.8
106	Connec-tion	234	0.08	446	0.05	32.4
107	Librarians	116	0.04	177	0.02	31.7
109	Directory	132	0.05	213	0.02	31.2
110	Agent	74	0.03	92	0.01	31.2
111	Office	105	0.04	155	0.02	31.0
113	Http	159	0.06	275	0.03	30.7
116	Evaluation	123	0.04	195	0.02	30.5
118	Systems	462	0.16	1,031	0.12	30.3
119	Register	114	0.04	176	0.02	30.2
120	Dirty	61	0.02	69		30.2
122	Table	175	0.06	316	0.04	29.4
123	Labora-tories	50	0.02	51		28.9
126	Procedures	109	0.04	169	0.02	28.6
131	Www	152	0.05	269	0.03	27.3
139	Error	112	0.04	181	0.02	26.4
140	Campus	62	0.02	77		26.2
145	Directories	58	0.02	70		25.7
146	Congress	92	0.03	140	0.02	25.3
147	Collection	156	0.06	285	0.03	25.2
148	Normal	79	0.03	113	0.01	25.1
149	Provided	181	0.06	345	0.04	25.0

Figure 18: Genre-based keywords in technical reports.

According to the lexical data covered in figure 18, technical reports also abound with items that are pivotal in their academic setting. How narrowly determining they are can be chiefly realized, in fact, by scanning the frequencies (**Freq. I** and **Freq. II**), and observing the close counts. For instance, the noun **protection** (ranked 66th) has nearly as many occurrences in the sub-corpus B of technical reports as it does in the general corpus – 81 vs. 87 instances. A form such as **laboratories** (position number 123), in addition, gets ever closer: 50 times repeating in RPs and only once more in the reference source (51).

The third and final body of significant vocables is located by comparing sub-corpus C (RAs) and our main corpus (figure 19).

N	Word	Freq.	Article.lst %	Freq.	Refer.lst %	Keyness
5	Project	660	0.26	1,120	0.13	177.7
6	Design	479	0.19	707	0.08	177.1
10	Problem	391	0.15	588	0.07	139.1
12	Problems	349	0.14	517	0.06	128.1
13	Search	297	0.12	411	0.05	124.5
16	Tools	285	0.11	400	0.05	116.3
22	Tasks	176	0.07	226	0.03	84.2
23	Video	243	0.09	373	0.04	82.7
25	Test	249	0.10	399	0.05	77.0
26	Study	233	0.09	366	0.04	75.4
27	Number	502	0.20	1,032	0.12	75.1
28	Solving	117	0.05	127	0.01	72.3
30	Subjects	123	0.05	139	0.02	71.8
34	Solution	152	0.06	200	0.02	69.8
36	Team	172	0.07	244	0.03	68.8
38	Program	438	0.17	899	0.10	65.9
40	Bounds	99	0.04	104	0.01	63.9
41	Findings	120	0.05	145	0.02	63.4
42	Members	194	0.08	306	0.04	62.2
45	Figure	173	0.07	266	0.03	58.7
48	E	515	0.20	1,142	0.13	56.6
50	Task	248	0.10	463	0.05	51.2
51	Teams	80	0.03	85		50.9
55	Browse	80	0.03	89	0.01	47.8
58	Criterion	76	0.03	83		46.6
60	Queries	75	0.03	82		45.9
64	Users	332	0.13	699	0.08	45.3
70	Variables	104	0.04	143	0.02	44.1
71	Fig	64	0.02	64		44.0
72	Results	191	0.07	344	0.04	43.8
73	Assign-ment	76	0.03	87	0.01	43.6
75	Condition	84	0.03	104	0.01	42.7
78	Graph	66	0.03	74		39.0
84	Label	68	0.03	80		37.4
85	Symbols	72	0.03	88	0.01	37.3
90	Conflicts	54	0.02	56		35.5
91	Large	199	0.08	390	0.05	35.2
92	Managers	64	0.02	76		34.7
96	Stage	70	0.03	90	0.01	33.4
99	Expertise	74	0.03	100	0.01	32.4
101	Perform-ance	293	0.11	650	0.08	32.1
103	Exact	68	0.03	89	0.01	31.5

106	Hierarchy	71	0.03	96	0.01	31.0
110	User	276	0.11	610	0.07	30.8
111	Size	157	0.06	299	0.03	30.5
116	Shown	77	0.03	112	0.01	29.3
118	Sources	113	0.04	195	0.02	29.1
120	Url	70	0.03	98	0.01	28.7
121	Html	258	0.10	572	0.07	28.4
127	Case	248	0.10	547	0.06	27.9
134	Approxi-mate	43	0.02	46		27.1
135	Know-ledge	290	0.11	666	0.08	27.0
137	Random	62	0.02	85		26.4
138	Dance	43	0.02	47		26.3
143	Searching	93	0.04	156	0.02	25.7
145	Organiza-tion	133	0.05	254	0.03	25.6
147	Order	232	0.09	515	0.06	25.4
148	Studies	101	0.04	176	0.02	25.3
151	Cases	131	0.05	250	0.03	25.3
155	Play	76	0.03	120	0.01	24.3
157	Likely	103	0.04	184	0.02	24.2
158	Estimates	42	0.02	48		24.1
159	Right	22	0.02	194		24.0
160	Failure	6	0.01	101		24.3
161	Connec-tions	18	0.02	174		24.5
164	Meaning	14	0.02	152		24.6

Figure 19: Genre-based academic keywords in research articles.

Despite their low frequency, as can be gleaned from figure 19, the last three items in the list – **failure**, **connections** and **meaning** – are recorded due to their occurrence across a diverse number of sources. The first noun, for example, is identified in the introductions, methods, results and discussions of various articles belonging to all the disciplines and dealing with different subjects – Computer architecture (A2), Client-server communications (A2), SGML (F3), Information infrastructure (F6), Audio-visual media technology (D1), and Hypertext technology (F4). In these environments, the co-occurrences *failure report* (2 times) and *failure Boolean* (2) play characteristic roles, as they are not located in any other genre. A similar situation is presented by the employment of the two other words, encompassing an

even wider range of sources and subjects or topics.

A useful approach in the work with academic keywords can be taken regarding their characteristic co-occurrences according to genre settings. From such a viewpoint, collocates and clusters are noted down provided they are exclusive of the given sub-corpus (TXs, RPs or RAs). In other words, an entry can be framed by checking that a node, combining with significant elements in a genre – e.g. textbooks-- does not co-occur with these items in reports and articles.

In this respect, figure 20 provides a series of three entry articles of the most frequent keywords in each category (figures 17, 18 and 19). Although, as stated in the case of academic lexis, the main goal is to achieve a description of items including the three genres, the type of proposal offered in figure 20 can become quite practical for delimiting this sort of specification within a genre. This need may be expressly felt when coping with production skills, so that the task be conducted by regarding characteristic employment in the genre. For example, the forms *requirement 1*, or *the following requirement*, in the **requirement** entry (figure 20 below), are found singularly in textbooks, exposing instructions on a given issue. The former is restricted to a single source, whereas the other cluster – *the following requirement*—is repeated through a higher number of texts. For research articles, in turn, some noteworthy elements functioning as key may be *project milestones*, *project deadlines, the project progressed* and any other clusters including reliable informants on the notion of project development and research, distinctive in the academic sphere of RAs. Such a type of deduction, however, is not always feasible, as some items may correspond to multi-word names lacking semantic effectiveness, such as *Cable Book library* or *Apple library*, found in technical reports (figure 20).

REQUIREMENT.
TXs
- 385K

®
___ # (166) / *basic* ___ / *connection* ___ /
addressed by ___ # / *general* ___ *concerns* / *in* ___ # /
the above ___ / *level* ___ *deals* / ___*for asynchronous* /
___*for support* / ___ *is refined* / ___ *is specified* /
___ *places constraints* / *under* ___ # / ___*for closing* /
___ *in this* / ___ *is for* / ___ *is further* /
___ *is important* / ___ *is included* / ___ *is necessary* /
___ *recognizes that* / ___ *should be* / ___ *specifies the* /
___ *would place* /

Ŧ
the first ___ (15) / *the following* ___ / *a* ___ *that* /
be a ___ / ___ *deals with* / ___ *relates to* /
___ *is* / *a* ___ *would* / *another* ___ /

LIBRARY.-
RPs
727K

®
Cable Book ___ (6) / ___ *program* / *the* ___*'s clientele* /
Networking the ___ / *the Kernel* ___ / *of the* ___ *'s* /
Access to ___ / *Apple* ___ *of* / *Cornell digital* ___ /
Digital ___ *systems* / *electronic doorway* ___ / *inter-*___ *lending* /
*Inter-*___ *loan* / ___*file system* / ___*functions included* /
___ *loan services* / ___ *project manager* / *objects programmers* ___ /
shared ___ *systems* / *the state* ___ / *the local* ___ /
the ___ *requires* /

Ŧ

by the ___ (16) / *at the* ___ / *the* ___ *is* /
___ *as a* / ___ *will provide* / *into the* ___ /
on the ___ / *with the* ___ / *the* ___ *can* /
the ___ *may* / *the* ___ *uses* / *the* ___ *of tomorrow* /
___ *plans to* / ___ *uses the* / *since the* ___ /
the ___ *building* / ___ *catalog* /

PROJECT. -
RAs
660K

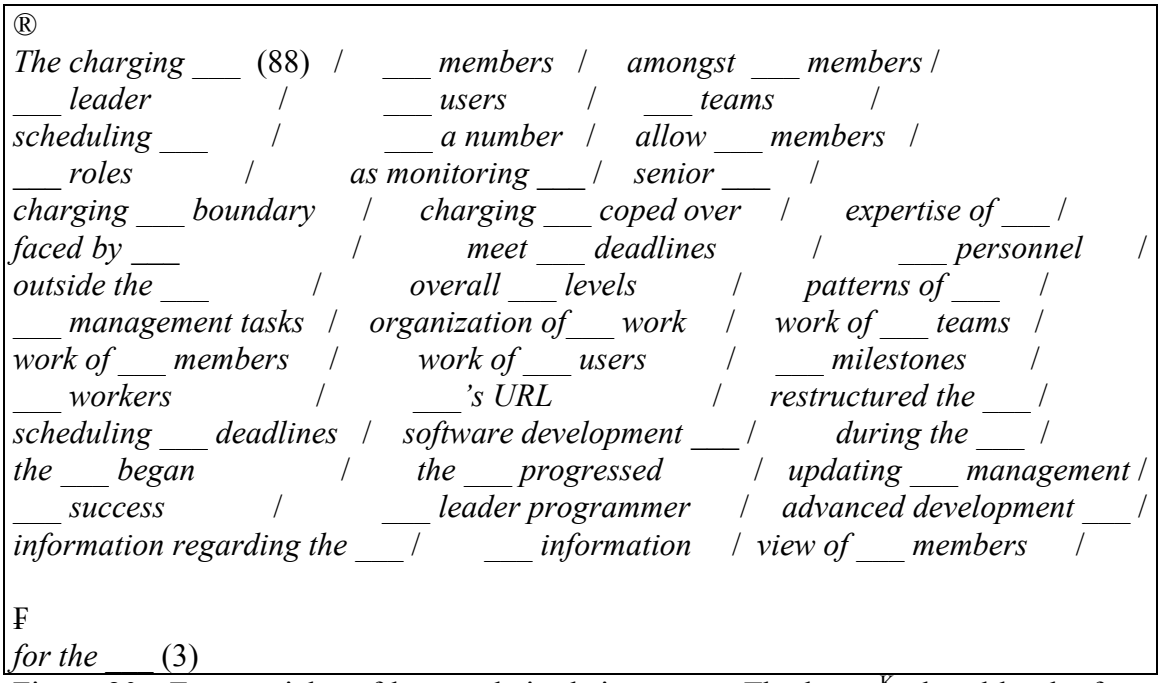


Figure 20: Entry articles of keywords in their genres. The letter [K] placed by the frequency count of the node refers to the position of that item as a keyword, thereby assessed in only one genre.

Perhaps due to its high frequency, the lemma **library** is more widely employed in other genres, and thus, its presence in research articles, for instance, yields some co-occurrences that also appear in technical reports. Examples of these are *in the library*, *of the library*, *Library of Congress*, *American Library Association*, *electronic library*, *digital library*, *library staff*, and *for the library*. In textbooks, in addition, the items *the public library*, *Library of Congress*, *in the library*, *of the library*, and *national library* are shared with sub-corpus B of technical reports. As a result, and stated above, these contructions are discarded in the entries of one-single-genre keywords.

In this respect, in the case of **project**, considered within research articles, the collocations found in technical reports, *of the project*, *project manager*, and *throughout the project* lead us to dispense with them for the analsyis of keyword co-texts. In textbooks, the cluster *of the project* is also spotted. For the key lemma in sub-corpus A (TXs), in turn, the expression *for the requirement* is repeated in technical reports and thus ignored in our

treatment of lexical collocates for genre-based keywords.

Thus, we come to value keyword employment within a contextualized unit such as the genre from stylistic concerns that may be quite convenient for meeting academic needs. As mentioned, academic common core lexis should be framed in the three sub-corpora or genre contexts of use – except in the case of keywords. The next division, subject-based lexis, deals with this same preoccupation; yet, the focus shifts to the perspective of more specified instances according to subject matter.

III. SUBJECTS / TOPICS

The starting point for exploring subject-related lexical items is through the analysis of keywords. These are examined within each subject or topic context category[246] – i.e. the capital letters and numbers distributed according to Appendix 1. The results provided below are thus segmented so that each heading may be described or characterized by the subject words which appear as keywords, most of them nouns.[247]

In addition, for these features to become even more idiosyncratic of the texts collected according to the subject-matter shared by disciplines, a key keywords approach can be carried out. These are databases of the most representative subject words that mark the `aboutness of the texts´ (Scott, 1997: 234). In our case, the aim is to check the occurrence of nouns repeated across each division where more than one subject or topic is encompassed. Finally, to complete the current section on subject vocabulary, a revision of frequency and range in addition to describing collocates is likewise performed.

[246] Topic is distinguished from subject in our corpus in the case of the F, or All disciplines category, where concepts like Hyptertext technology and Information infrastructure are regarded as topics dealt with to a greater or lesser extent in all areas (see chapter 2). Instead, something like Information units management or Software engineering, for instance, are considered subjects, shared by three or two disciplines.

III.1. KEYWORDS.

The scope of this examination is not established by following the objective of identifying topics and subjects in Information Science and Technology by means of specific lexis. For such a detailed measurement, there exists the broad area of terminology which, among other things, focuses on the recording of keywords according to large databases of subject texts.

Quite the opposite, our aim lies in the observation of those words bearing a subject-based semantic load determined by conceptual rungs. In such a way of proceeding, the path to follow is that of seeking out those words most peculiar to a group of texts classified according to subject matter (see Appendix 1 for divisions). In this respect, the approach is qualitative since it is based on detailed observation of items in restricted subjects, in comparison with a more quantitative view of data set up in the overall corpus (i.e. the case of genre-based items across all three subcorpora, TX, RPs and RAs above).

Thus, the following 17 sets of keywords (samples # 11 – 27) show the top ten items in each subject / topic heading. We deem these first ten words in each set to represent each subject better, since they have higher degrees of keyness. The vocables are obtained by comparing two wordlists: a smaller one encompassing the sources in the given category –e.g. A1, `History of Computers, Hardware and Software´-- and the larger reference corpus embodying the whole range of texts in all subjects and disciplines, therefore providing the background data for reference comparison.

[247] This agrees with Scott's assertion in the starting window of *WordSmith* (1996) that `the vast majority of keywords are nouns´.

N	Word	Freq. I	%	Freq. II	%	Keyness
1	Multics	121	0,31	122	0,02	395,1
2	Segment	133	0,34	178	0,02	384,6
3	Minix	104	0,26	105	0,01	339,4
4	Bit	121	0,31	226	0,03	294,1
5	Ring	89	0,23	92	0,01	287,7
6	Segments	81	0,21	123	0,02	220,2
7	Wmf	65	0,16	65		212,8
8	Atm	63	0,16	63		206,3
9	Raid	59	0,15	60		191,9
10	File	130	0,33	555	0,08	165,3

Sample # 11: History of computers, Hardware and Software (A1). From left to right: Ranked position on list / Word / Frequency in subject category / percentage in subject category / Frequency in reference corpus / percentage in reference corpus / Keyness score / [248]

N	Word	Freq. I	%	Freq. II	%	Keyness
1	Design	262	0,36	645	0,09	291,3
2	Tcl	128	0,18	131	0,02	281,0
3	Subjects	120	0,17	138	0,02	246,5
5	Task	180	0,25	406	0,06	219,2
6	Constraints	118	0,16	173	0,02	207,0
7	**Tk**	**92**	**0,13**	**95**	**0,01**	**201,0**
8	Test	155	0,21	331	0,05	199,2
9	Handle	117	0,16	204	0,03	179,9
10	Tasks	119	0,16	213	0,03	179,0
11	Magnitude	84	0,12	99	0,01	170,0

Sample # 12: Computer engineering and architecture; Data communications, Client-server architecture (A2).

N	Word	Freq. I	%	Freq. II	%	Keyness
2	Network	74	1,58	1.245	0,17	194,6
3	Snmp	26	0,55	29		187,2
5	Tools	34	0,72	350	0,05	118,5
6	Tool	21	0,45	130	0,02	92,2
7	Automation	12	0,26	35		68,3
8	Etherpeek	9	0,19	9		66,1
9	Intermapper	7	0,15	7		51,4
10	Helps	9	0,19	41		44,3
11	Ip	12	0,26	110	0,02	44,3
12	Administrator	8	0,17	26		44,1

Sample # 13: Information units management (B1).

[248] This labelling is the same for all the samples (11-31).

N	Word	Freq. I	%	Freq. II	%	Keyness
1	Object	509	0,72	829	0,11	848,5
2	Type	378	0,54	736	0,10	544,7
4	Surrogate	226	0,32	234	0,03	503,1
5	Netobj	197	0,28	197	0,03	446,8
6	Method	248	0,35	477	0,07	360,9
7	Call	201	0,28	326	0,04	335,6
8	Stub	147	0,21	150	0,02	329,7
10	Transport	136	0,19	148	0,02	294,4
11	Thread	135	0,19	163	0,02	275,1
12	Owner	131	0,19	151	0,02	274,4

Sample # 14: Online database systems and Computer systems (B2).

N	Word	Freq. I	%	Freq. II	%	Keyness
1	Project	593	1,26	1.099	0,15	1.278,0
2	**Knowledge**	**422**	**0,90**	**651**	**0,09**	**1.013,7**
3	Program	364	0,78	836	0,11	675,8
4	Members	172	0,37	293	0,04	389,3
5	Team	153	0,33	239	0,03	364,3
8	Senior	102	0,22	119	0,02	282,3
10	M	147	0,31	369	0,05	254,5
11	Work	219	0,47	871	0,12	244,4
12	Tools	135	0,29	350	0,05	227,8
13	Manage-ment	127	0,27	332	0,05	212,8

Sample # 15: Automated knowledge-based systems (B3).

N	Word	Freq. I	%	Freq. II	%	Keyness
1	**Images**	**97**	**1,16**	**605**	**0,08**	**287,8**
3	Img-name	40	0,53	40		257,0
4	Media	74	0,99	478	0,07	254,1
5	Gif	38	0,51	65		214,4
6	Mass	35	0,47	98	0,01	169,9
7	Font	33	0,44	138	0,02	137,9
8	Pornography	21	0,28	23		132,2
10	Width	25	0,33	56		130,4
11	Border	19	0,25	20		120,7
12	Communi-cations	29	0,39	194	0,03	97,7

Sample # 16: Content analysis (C1).

N	Word	Freq. I	%	Freq. II	%	Keyness
1	**Genres**	**59**	**0,52**	**85**	**0,01**	**302,0**
2	Media	75	0,67	478	0,07	204,1
3	Genre	34	0,30	80	0,01	148,3
4	Languages	35	0,31	156	0,02	116,0
6	Activities	32	0,28	194	0,03	89,6
7	Character	34	0,30	233	0,03	88,3
8	Costs	36	0,32	275	0,04	87,1
9	Communities	25	0,22	123	0,02	78,7
10	Multi	21	0,19	73	0,01	78,3
11	Exchange	20	0,18	70		74,3

Sample # 17: Media documentation (C2).

N	Word	Freq. I	%	Freq. II	%	Keyness
2	Copyright	163	0,80	308	0,04	586,1
3	Contractor	84	0,41	105	0,01	352,3
4	Microfilm	91	0,45	153	0,02	342,8
5	Images	143	0,70	605	0,08	335,5
6	Image	140	0,69	628	0,09	315,7
7	Reel	62	0,30	63		277,5
8	Film	81	0,40	228	0,03	241,5
9	Printed	74	0,36	182	0,02	236,1
10	Library	154	0,76	1.184	0,16	221,2
11	Congress	65	0,32	140	0,02	220,6

Sample # 18: Documentation legislation (C3).

N	Word	Freq. I	%	Freq. II	%	Keyness
1	Video	172	0,53	360	0,05	449,7
2	Fiber	102	0,32	139	0,02	328,8
3	Laser	81	0,25	130	0,02	242,5
4	Fibers	63	0,20	64		227,9
5	Light	79	0,24	140	0,02	225,4
7	Optic	56	0,17	65		192,7
8	Glass	52	0,16	54		186,6
10	Labo-ratories	47	0,15	51		165,9
11	Tele-phone	63	0,20	148	0,02	153,9
12	Dance	40	0,12	47		137,0

Sample # 19: Audio-visual media technology (D1).

N	Word	Freq. I	%	Freq. II	%	Keyness
1	Standards	237	1,71	506	0,07	980,2
2	Ieee	151	1,09	161	0,02	778,2
3	Sa	51	0,37	52		265,6
4	Scc	42	0,30	42		219,8
5	Board	47	0,34	114	0,02	184,4
6	Edifact	31	0,22	31		162,2
7	Edi	32	0,23	38		159,9
8	Transportation	29	0,21	31		149,1
9	Chair	33	0,24	55		148,6
11	Com-mittee	31	0,22	66		127,9

Sample # 20: Audio-visual media theory (D2).

N	Word	Freq. I	%	Freq. II	%	Keyness
1	Digital	610	0,87	1.162	0,16	904,3
2	Semio-tics	203	0,29	208	0,03	457,1
3	Codes	213	0,30	272	0,04	422,1
4	Archives	192	0,27	235	0,03	390,5
5	Signs	173	0,25	178	0,02	388,6
6	Semiotic	167	0,24	178	0,02	368,0
7	Sign	156	0,22	179	0,02	329,7
9	Analysis	173	0,25	437	0,06	195,1
10	Social	119	0,17	215	0,03	183,8
11	Meaning	99	0,14	140	0,02	183,5

Sample # 21: Communication theory (E1).

N	Word	Freq. I	%	Freq. II	%	Keyness
1	Basis	63	0,47	157	0,02	249,2
2	Techlib	42	0,32	42		223,0
3	Microsoft	43	0,32	77	0,01	192,2
4	Nt	30	0,23	47		140,1
5	Iis	26	0,20	26		138,0
8	Server	57	0,43	492	0,07	110,4
9	Respondents	20	0,15	23		102,3
10	Active	32	0,24	132	0,02	100,3
11	Dimensions	21	0,16	42		90,3
12	Reports	30	0,23	134	0,02	90,1

Sample # 22: Perspectives on Information (F1).

N	Word	Freq. I	%	Freq. II	%	Keyness
1	Net	216	0,38	311	0,04	464,1
2	Internet	356	0,62	1.029	0,14	440,6
3	Access	349	0,61	1.135	0,16	379,8
4	Usenet	151	0,26	165	0,02	377,6
5	People	254	0,44	627	0,09	366,0
6	Government	139	0,24	234	0,03	270,1
8	Id	114	0,20	185	0,03	227,0
9	Date	112	0,19	217	0,03	196,9
10	Message	140	0,24	360	0,05	194,1
11	Public	201	0,35	771	0,11	177,8

Sample # 23: Unix / Internet (F2).

N	Word	Freq. I	%	Freq. II	%	Keyness
1	Element	427	0,56	484	0,07	847,9
2	Elements	363	0,47	538	0,07	601,6
3	Text	521	0,68	1.346	0,18	511,1
4	TEI	261	0,34	324	0,04	489,1
5	Attribute	222	0,29	228	0,03	466,7
6	Docu-ment	364	0,47	855	0,12	397,9
7	Tag	216	0,28	289	0,04	384,9
8	Sgml	220	0,29	335	0,05	357,0
9	Html	272	0,35	568	0,08	335,8
10	Chapter	174	0,23	259	0,04	287,1

Sample # 24: Html, Sgml, Tei (F3).

N	Word	Freq. I	%	Freq. II	%	Keyness
2	Web	86	0,97	847	0,12	207,5
3	Streaming	27	0,30	35		154,7
5	Html	52	0,59	568	0,08	116,4
6	News	30	0,34	153	0,02	105,7
7	Audio	27	0,30	148	0,02	91,8
8	Sgml	35	0,39	335	0,05	86,0
10	Real-audio	12	0,14	16		68,2
11	Radio	18	0,20	98	0,01	61,4
12	Spanish	11	0,12	16		61,1
13	Structural	16	0,18	75	0,01	58,6

Sample # 25: Hypertext technology (F4).

N	Word	Freq. I	%	Freq. II	%	Keyness
1	Am	138	0,22	181	0,02	295,0
2	Electronic	234	0,38	639	0,09	279,9
3	Text	323	0,52	1.346	0,18	223,4
4	Web	239	0,38	847	0,12	209,6
6	Image	172	0,28	628	0,09	144,8
8	Publishers	71	0,11	112	0,02	134,9
9	Scholars	65	0,10	93	0,01	131,7
11	Images	157	0,25	605	0,08	122,3
12	Sites	110	0,18	326	0,04	120,2
13	Scanning	78	0,13	166	0,02	118,0

Sample # 26: Electronic publishing (F5).

N	Word	Freq. I	%	Freq. II	%	Keyness
1	Library	320	0,65	1.184	0,16	363,3
2	Public	213	0,43	771	0,11	247,6
3	Training	112	0,23	194	0,03	241,0
4	Internet	245	0,49	1.029	0,14	238,7
5	Participants	100	0,20	179	0,02	210,7
6	Libraries	177	0,36	645	0,09	204,2
7	Evaluation	93	0,19	179	0,02	186,8
8	Rural	77	0,16	114	0,02	181,9
9	Nyser-net	64	0,13	68		178,9
10	Community	120	0,24	393	0,05	155,1

Sample # 27: Information infrastructure (F6).

As mentioned above, the word sets illustrate the lexical approach towards conceptual and thematic fields in a pratical manner. This means that these key elements are primarily viewed as building blocks with which to frame the perspective of subject-matter in relation to vocabulary and their surrounding structures. Neither are the lists meant to be all-encompassing nor do they provide an exclusive vision of our subject approach -- in fact, as more and differently selected texts are entered, relevant lexis may be introduced to change or occupy different positions in the lists. The important notion is that these keywords determine the specification of the sources being handled so that learners can have access to them in order to work out patterns and comparisons with their readings.

The ranked positions in the lists depict the importance of these items. Where the order of the ranking is altered, as in sample # 25, for instance, this means that irrelevant keywords for subject-based analysis have been left out. In this case, the word **Virginia**, occupies position # 1 as a keyword in the subject of Hypertext Technology, but, being the name of a state, it is treated as an irrelevant toponym and therefore discarded, according to the parameters described in chapter 2 (`Word discrimination in common core analysis´).

A third point that may be noticed on taking these lists into account is the presence of items that are also considered as academic vocabulary – e.g. **library** and **libraries** in sample # 27. This is inevitable in some cases, since the texts taken for subject-based lexis analysis are partially the same ones that served for drawing academic results. A similar outcome is produced if some of the texts chosen match the text types: Words like **text** and **document** (sample # 24) can thus arise as a result.

The difference, however, is that for our common core analysis, we are chiefly concerned with the lexical patterns of these words according to segmented contexts. The same item can thus function as either subject, academic or text type word, depending on its use in a given set of texts. In this respect, for example, the noun **library** can be checked across genre settings in the academic view, and multi-word units such as *Digital library systems* or *electronic doorway library* be regarded as key academic elements in our corpus of readings. In contrast, the same node can be examined in the subject sub-group where it is key, and more thematically bound collocations come up instead, such as *American Library Association* and *public library community.* The important function to keep in mind is that these elements represent a given stretch of discourse encompassing a conceptual domain shared by two or more disciplines in our Information Science and Technology corpus.[249]

[249] In this respect, they are regarded, like academic and text type words above, as common core lexis (see Chapter 2 – Word Taxonomy).

The other type of analysis carried out with these items is, as pointed out above, the examination of key keywords or words which are `key in more than one of a number of related texts´ (Scott, 1996). As a result, the context contrasted in this case is that of the whole subject area category common to two or more disciplines – e.g. Field `A´, shared by Computer Science and Telecommunication. The divisions which only contain one sub-heading are discarded in this respect, since there must be at least two or more different sets of keywords –e.g. `A´, in which Computer Science and Telecommunication include History of computers (A1) and Computer engineering and architecture (A2) as two shared subjects. Heading `E´ (Communication Theory), with only one subject (E1), is thus excluded.

The procedure is quite simple: A comparison of two wordlists that represent the sub-divisions –e.g. A1 and A2— is made to draw a list of keywords; this one is then contrasted with the entire corpus as a reference corpus. The results (samples # 28-32) yield a thematic outline of the most representative lexis according to shared subject matter in Information Science and Technology from the perspective of our corpus: They expose, in Scott's words, the `aboutness´ of the sections.

Sample # 28: Computer Science and Telecommunication (A):
Left to right, the order is given according to frequency.[250]

1. Bits	2. Hardware	3. Directory	4. Ip	5. Software	6. Pc
7. Cc	8. Computer	9. Tests	10. Pointer	11. Mouse	12. Display
13.Unix	14. Features	15. Multiple	16. Adapter	17. Instruction	18. Versions
19.Optimization	20. Array	21. Operating	22. Tcp		

Sample # 29: Information Science, Computer Science and Telecommunication (B):

1. Tool	2. Tools	3. Automation	4. Ip	5. Documentation	6. Staff
7. Packets					

Sample # 30: Information Science and Audio-visual Communication (C):

1. Mass	2. Media	3. Communications	4. Width
5. Internet	6. Images	7. Copyright	

[250] The same applies to the rest of the key keywords sets.

Sample # 31: Audio-visual Communication and Telecommunication (D):

1. Message 2. Board 3. Standards

Sample # 32: The four disciplines (F):

1. Library 2. Project 3. Libraries 4. Training 5. Librarians 6. Cd
7. Evaluation 8. Study 9. Internet 10. Staff 11. Community
12. Communities 13. Library's 14. Site 15. Connectivity 16. Resources
17. Success 18. Text 19. Equipment 20. E-mail 21. Resolution
22. Connection 23. Learning 24. Elements 25. Links 26. SGML 27. Web
28. TEI 29. HTML

As can be checked, samples #28 and #32 yield a greater number of key keywords. This may be due both to their larger quantity of text and idiosyncratic nature in terms of topics encompassed. Indeed, the F category (sample # 32), predominantly dealing with Internet issues and electronic information management, contains a vast collection of references to library and information resources on such media, since these centres of information are taken as models and described thoroughly to deal with elecronically formatted data. Hence, library**,** libraries **and** librarians **occupy first, third and fifth positions respectively. In the case of the A division (sample # 28), computing terms tend to prevail across texts, as these are mainly about programming, hardware analysis, computer engineering, etc. In this sense, the terms are more peculiar of their key keyword division, perhaps implying that these two disciplines –Computer Science and Telecommunication—have a lot of technical background in common, and that they are thus dramatically marked in relation to the rest of the subject areas.**

III.2. WORD FREQUENCY AND RANGE.

The determination of subject lexical items as keywords should be accompanied by the sorting of these items according to the proportion of their frequency and range across the sources encompassed. This is due to the fact that a given keyword may not endure a

significant range level throughout the genres, which are key contextual yardsticks in the compiling of lexical elements. In such a case, it is convenient to regard the detailed consistency analysis based on the three genres, TXs, RPs, RAs, as a useful measure to determine which subject vocabulary is repeated across the entire corpus (appendix 4). As previously indicated, the more widespread across genres a word is found to be, the more productive the item tends to be in the academic task-wise learning process. In that sense, lists of academic words function to provide common core lexis across a given number of disciplines (appendix 4).

III.3. COLLOCATES.

With the keywords from the lists above (samples # 11 – 27), the recording of collocates is performed to learn from the behavior of these lexical items in the subject context. Unlike academic words above, these are more recurrent within the thematically bound sets of texts. Their use is thus somewhat more restricted. Only their behavior in such texts is appreciated, and the counts in these settings are pointed out by the node.

Let us then examine one entry in each of the 17 categories (samples # 11 – 27). The chosen words (samples # 33 – 49) correspond to the most representative ones in terms of presenting the highest keyness score. In appendix 6, subject entries are included to check the three most frequent uses in each heading, totalling 61 subject-based lemmas. The clusters and multi-word units associated with these keywords are taken from the context of the subject / topic headings alone, rather than the whole corpus.

MULTICS.- 121 (A1) [1]	**A (1) History of Computers, Hardware, Software**
___ *virtual memory* (14) / ___ *virtual memory design* / *work on* ___ / *reflections on* ___ /	

Sample # 33: First entry in A1 category. [1] Frequency of item in the subject-bound texts (the same applies in all the samples).

DESIGN.- 262 (A2)	**A (2) Computer Engineering and Architecture, Data Communications and Client-Server Architecture**
___ process (30) / *systems for ___* / *___ problem solving* / *during the ___* / *architectural ___ problem* / *___ task* / *architectural ___ process* / *depending on ___* / *on ___ phase* / *___ context* / *a ___ problem* / *across ___ disciplines* / *detailed ___* / *instructional ___* / *mechanical ___* / *non-___* / *applied___* / *architectural ___ tasks* / *background on ___* / *complexity of ___* / *___ constraints* / *___ depends upon* / *___ information systems* / *___ involves* / *___ problems throughout* / *___ process forms* / *___ proposal through* / *___ task is* / *___ task may* / *___ task situation* / *___ was introduced* /	

Sample # 34: First entry in A2.

NETWORK.- 74 (B1)	**B (1) Information Units Management**
on the ___ (5) / *of your ___* / *helps the ___* / *keep a ___* / *___ manager* /	

Sample # 35: First entry in B1.

OBJECT.- 509 (B2)	**B (2) Online Database Systems, Computer Systems**
network ___ (78) / *network ___ runtime* / *network ___ type* / *surrogate ___* / *a concrete ___* / *an ___ whose* / *network ___ argument* / *network ___ result* / *network ___ system* / *remote ___ invocation* / *valid network ___* / *___ owned by* / *___ whose methods* / *a remote ___* / *local ___* / *___ in question* / *___ invocation for* / *___ is called* / *___ is invoked* / *___ types* / *___-oriented buffered streams* / *retained pointer protects the ___* / *real ___* / *the common ___* / *___ adaptor of* / *___ argument* / *___ is sent* / *___ marshaled* / *allocates an ___* / *an ___failed* / *an ___ table* / *become ___-oriented* / *closure ___ contains* / *communicating ___ types* / *declares ___ types* / *given an ___* / *given network ___* / *remote ___* / *lying ___ runtime#* / *missing ___ indicates that* / *network ___ agent* / *network ___ method* / *network ___ stored* / *null network ___* / *___ agent#* / *___ argument fs#* / *___ can coexist* / *___ contains* / *___ exported ___* / *___f plays* / *___ generates* / *___ invocation* / *___ is constructed* / *___ is encountered* / *___ is imported* / *___ is legal* / *___ is local* / *___ itself as* / *___ method invocation* / *___ methods getkey* / *___ methods notify___* / *___ methods openname* / *___ overrides* / *get ___ runtime* / *___ runtime allows* / *___ runtime error* / *___ runtime negotiates* / *___ runtime system* / *___ runtime uses* / *___ solution avoids* / *___ stored in* / *___ subtyping together* / *___ system itself* / *___ that creates* / *___ that manages* / *___ to request* / *___ whose Id* / *___ whose method* / *___-oriented programming* /	

Sample # 36: First entry in B2.

PROJECT.- 593 (B3)	**B (3) Automated Knowledge-Based Systems**
Within the ___ (45) / the charging ___ / ___ manager / ___ members / throughout the ___ / amongst ___ members / knowledge of ___ / ___ leader / of ___ users / ___ including the / ___ section # / the ___ teams / scheduling ___ / charging ___fits / overall ___ / ___ roles / ___ users / updating ___ / monitoring ___ / charging ___ boundary / ___ began in / ___ boundary spanning / ___ coped over / ___ could have / ___ deadlines meant / ___ deadlines / ___ during the / ___felt that / ___fits the / ___ information / ___ management tasks / ___ manager acceptance / ___ manager estimated / ___ manager located / ___ managers need / ___ members felt / ___ members sought / ___ members utilized / ___ milestones / ___ organization # / ___ personnel / ___ users found / ___ was integrating / ___ was maintained / ___ was nearing / ___ was organized / ___ workers / readjustments amongst ___ members / regarding the ___ / reorganizing the ___ /	

Sample # 37: First entry in B3.

IMAGES.- 97 (C1)	**C (1) Content Analysis**
RC ___ (21) / *___ may be / ___ on the internet / post ___ / of # ___ / pornographic ___ /*	

Sample # 38: First entry in C1.

GENRES.- 59 (C2)	**C (2) Media Documentation**
___ of communication (6) / *media and ___ / the economics of ___ /*	

Sample # 39: First entry in C2.

COPYRIGHT.- 163 (C3)	**C (3) Documentation Legislation**
the ___ office (35) / *___ law / ___ acquisitions division / ___ card catalog / ___ arbitration royalty / ___facts / owner of ___ / register of ___ / ___ system /*	

Sample # 40: First entry in C3.

VIDEO.- 172 (D1)	**D (1) Media Technology**
___ and audio (8) / *film and ___ / the incoming ___ / the ___ data / ___ preservation / the j# ___ / family of ___ adapters / incoming ___ stream / ___ signals / broadcast ___ signals over / television and ___ / use of ___ /*	

Sample # 41: First entry in D1.

STANDARDS.-237 (D2)	**D (2) Media Theory**
___ board (15) / *body of ___* / *the IEEE ___* / *appropriate ___* / *board ___ coordinating committee* / *ieee ___ activities* / *IEEE ___ association* / *provisional ___* / *ieee ___ computerization* / *a ___ perspective* / *provide ___* / *managing ___ development* / *___ involvement* / *___ activity* / *___ and ensuring* / *___ are strategic* / *___ being developed* / *___ board chair* / *___ board itself* / *___ board takes* / *___ by vendors* / *___ development activities* / *___ development activity* / *___ development meetings* / *___ development organizations* / *___ perspective* / *___ program manager* / *___ provide* / *___ setting IEEE* / *___ testing program* / *___ were published* / *___-development system* / *___-related outputs* / *___-setting organizations* / *status of ___* / *strategic ___ goals* / *strategic ___ objectives* / *systems ___ program* /	

Sample # 42: First entry in D2.

DIGITAL.- 610 (E1)	**E (1) Communication Theory**
___ information (76) / *___ information objects* / *___ archives* / *preserving ___ information* / *the ___ environment* / *in ___form* / *integrity of ___ information* / *archiving of ___ information* / *___ objects* / *certified ___ archives* / *___ preservation* / *___ materials* / *access to ___ information* / *collections of ___ archives* / *refreshing ___ information* / *the ___ landscape* / *the ___ world* / *a ___ library* / *archiving ___ information* / *development of ___ archives* / *information in ___form* / *longevity of ___ information* / *migration of ___ information* / *challenge of archiving ___ information* / *___ library systems* / *___ scanning* / *___ storage* / *distributed ___ archives* / *operating environment of ___ information* / *means of ___ archiving* / *mechanisms for ___ archiving* / *___ documents* / *___ technology* / *providing access to ___ information* / *repositories of ___ documents* /	

Sample # 43: First entry in E1.

BASIS.- 63 (F1)	**F(1) Perspectives on Information**
The ___ techlib (12) / *___ webserver gateway* / *___ Techlib application* / *___ client suite* / *___ document manager* / *___ intranet solution* / *___ Techlib database* / *___ Techlib uses* /	

Sample # 44: First entry in F1.

NET.- 216 (F2)	**F (2) UNIX / Internet**
of the ___ (42) / *to the ___* / *on the ___* / *access to the ___* / *the ___ as* / *the ___ has* / *future of the ___* / *development of the ___* / *have access to the ___* / *the internal ___* / *by the ___* / *to connect to the ___* / *___ access* / *free-___* / *program to connect to the ___* / *available on the ___* / *information on the ___* /	

Sample # 45: First entry in F2.

ELEMENT.- 427 (F3)	**F (3) HTML, SGML, TEI**
Of the ___ (32) / *the ___ may* / *the ___ may be used to* / *Using the ___* / *___ described in the TEI* / *___ used to document* / *A particular ___* / *a single ___* / *comparable ___* / *Documentation ___* / *equivalent or comparable ___* / *on the ___* / *The containing ___* / *the ___ concerned* / *the ___ described* / *An ___ class* / *___ defined in* / *the base ___* / *within an ___* /	

Sample # 46: First entry in F3.

WEB.- 86 (F4)	**F (4) Hypertext Technology**
on the ___ (8) / *of the ___* / *the ___ is* / *the ___ in* / *the ___ to* / *in the ___* / *Over the ___* / *the ___ has* / *using the ___* / *broadcast over the ___* / *Available on the ___* / *used in the ___* /	

Sample # 47: First entry in F4.

AM.- 138 (F5)	**F (5) Electronic Publishing**
___'s experience with (5) / *___ has been* / *___ is being used* / *of ___* / *that ___ has* / *___ staff* / *purpose of ___* (2) * All TX	

Sample # 48: First entry in F5.

LIBRARY.- 320 (F6)	**F (6) Information Infrastructure**
The public ___ (40) / *___ of congress* / *the ___ has* / *___ association* / *___ staff* / *public ___ community* / *role of the ___ staff* / *digital ___ program* / *___'s clientele* / *networking the ___* / *the ___ uses* / *the local ___* / *virtual ___ staff* / *Apple ___* / *connectivity on the ___* / *electronic doorway ___* / *___ and information* / *___ of tomorrow* / *state ___* / *digital ___* /	

Sample # 49: First entry in F6.

Sample entries of this sort generally supply lexical feedback for our inquiry into common core lexis. The main difference with academic words is that this is done across disciplines from the subject-based approach of keywords. The data analysis, however, may sometimes yield similar results, as in sample # 37, for instance, where items like *the charging project*, *project manager* and *project leader* are also accounted for in the academic lexis treatment (see figure 20 above). The major difference, in contrast, is given by the specification of co-occurrences that the node presents in the selection of texts. In this manner, other clusters in sample # 37, such as *project manager acceptance*, *project manager estimated*, *project manager located*, and *project managers need* spell out a description of more specific multi-word items, identified across sources dealing with automated knowledge-based systems (B3). Items of this sort tend to go unnoticed when working with a broader scope of use across the entire corpus.

Having encompassed the three main textual settings for our analysis --text types, genres and subjects / topics-- a subsequent step to be taken is that of a contrastive study. The next section thus presents the assessment of the significance of the results by means of a comparison with lexicographic sources.

IV. CONTRASTIVE STUDY

The aim of this section is to provide contrastive feedback on the results of the previous part by means of lexicographic sources in Information Science and Technology. The information thus obtained should contribute to fostering our common core study in terms of lexical significance and value in the academic arena. In other words, the implications of the results examination are addressed as a means of useful linguistic feedback. As we shall see in chapter 4, the end is the framework of common core pedagogic applications in English studies for Computer Science, Information Science, Telecommunication (Optical / radio

communications) and Audio-visual communication.

From this line of work, in the approach to the lexical entries, the lexicographic plane is also conceived as relevant. With this concept, reference is made to sources such as dictionaries, glossaries and corpora word lists which may serve as models of comparison for our content words.

Given our corpora frequency and alphabetical lists, the procedure followed is that of offering contrastive samples according to different stages: First, our corpus detailed consistency list is balanced with some entries from two sources dealing with Information Science and Technology. In the subsequent levels, each discipline is in turn contrasted with published reference material in the same subject areas: dictionaries, glossaries, vocabularies and word lists. The approach taken is mainly aimed at checking the appearance or absence of our most significant data in related lexicographic work.

Therefore, a detailed or exhaustive examination of the lexical items is not proposed. That would entail a different project in itself, since the amount of items is so vast. In addition, it is not the intended focus of this section. Instead, the attention is placed on probing the major findings from the previous three segments in this chapter. In such a way, the search for lexical items not surfacing from our corpus is not part of the `modus operandi´, although it would most likely prove interesting from the lexicographic viewpoint, and will be left as a prospective avenue for research at some future time.

The following sample (# 50) is a list of two columns of terms that are interchangeably contrasted to show which words are shared by the sources.

Our corpus	***Dictionary of Information Technology*** **(Peter Collin, 1997)**
A □	
A's	
Aa	
Aaas	
Aaas's	
Aachen	
Aal	
Aap	
Aarhus	
Aashto	

Aat
Aau
Ab
Abandon
Abandoned □
Abandoning
Abandonment
Abandonned
Abbate
Abbe
Abbreviate
Abbreviated
Abbreviating
Abbreviation □
Abbreviations
Abby
Abc's
Abcl
Abduction
Abe
Abelsson
Aber
Aberrant
Aberystwyth
Abii
Abil
Abilities
Ability
Able
Ablex
Abnormal
Abnormally
Abolished
Abolishing
Abort
Abound
Abounds
About □

About̃

Sample # 50: Contrastive view of first 48 words ordered alphabetically in our corpus.
In bold = common core words repeating across the three genres in our corpus / □̣= collected from
the *Dictionary of Information Technology* / ̃ = word identified in our corpus.

The reference dictionary being contrasted comprises `all aspects of information technology, including computer hardware and software, world wide web, the internet, telecommunications, multimedia, video, radio, television, film and print media´(Collin, 1997:

preface). Consequently, our aimed sources in this section should focus on specific areas or fields of Information technology by regarding subject and theme as primary elements. Less demarcated works that cope with Information or Communication in general are thus discarded (e.g. Weiner, 1999).

As can be seen in sample 50, given the first 13 items arranged in alphabetical order –from **A** to **About** – in such a source, seven words from the *Dictionary of Information Technology* are compiled in our overall corpus –indicated by a check mark ().

That type of analysis, however, as indicated above, cannot be extended throughout this chapter on results.[251] Instead, the lexicographic authorities should mainly contribute to providing guidance in the observation of the meaningful position of the common core lexis in our corpus and sub-corpora. For instance, a feature that could be considered is the pinpointing of common core words –appearing in all genres and highlighted in bold in our corpus column above-- which do show up in the dictionary. This could be done with the most frequent words, as surmised from the preceding section.

Some examples may be assessed by contrasting, for instance, the phrase *the code had to be abandoned*, repeated four times in our corpus. The reference work cited, in contrast, provides the following sentence for the lemma **abandon**: *once you have abandoned your spreadsheet, you cannot retrieve it again.* Similarities can thus be drawn in terms of context of communication –e.g. genre--, as both occurrences of the item are produced in a computer context –in our case, of a research article. Likewise, the grammatical form may be compared. The word-form employed is, in fact, the past participle in both instances –more recurrent, indeed, in our corpus: the condition of **abandoned** is that of a common core lemma, whereas that of abandon or abandoning is not.

[251] That kind of study would be labelled as contrastive lexicographic analysis, in which the main aim is to assess linguistic or literary sources by means of lexical feedback (see, for instance, Krishnamurthy, 1997).

Throughout the six sub-sections below (Information Science and Technology, Computer Science, Information Science, Telecommunications, Audio-visual Communication, and Internet), the most frequent words in our corpus and sub-corpora are assessed in the light of lexicographic works dealing with similar subject fields. Thus, it is important, from this view, to both unify and separate the Information Science and Technology disciplines so that the descriptive accounts be formed from a more detailed perspective.

IV.1 INFORMATION SCIENCE AND TECHNOLOGY

Under this heading, the point of departure is the detailed consistency list of the genre-based words, as shown in Appendix 4. These are common core content words classified as either academic or subject lexis.[252] The pivoting means of selection of the items for contrastive purposes is marked by the lemmas used, which should be the results of `reducing the words in a corpus to its lexemes´, as Esteves da Rocha claims (Corpora discussion list, 3 Nov 1999). This is also done after the word list model proposed by Flowerdew (1993 b: 233), in which the items `in a specialized corpus should be reduced by some 55%´, to a figure of about 2000 words.[253]

This involves presenting the lemma as the word with the highest frequency and range –higher up in the list-- and discarding all other forms positioned lower in the table –e.g. **library** is ranked 48th in the detailed consistency list, whereas **libraries** is number 98; as a result, the former is regarded as the lemma, and the latter is not used. The specification of the slot which the items occupy in the detailed consistency list is provided in Appendix 7, where

[252] The distinction is specified by checking the keywords of the subject categories, as has been done previously.

[253] We thus carry out a lemmatization task in the choice of lexemes which represent given word-forms. For such a development, there exist automatic lemmatizers such as Berber Sardinha's lemmatization list that can accompany *WordSmith Tools*. In such a table, however, infinitive forms tend to be considered as lemmas (e.g. **abandon** for abandoned, whereas the opposite may occur in our corpus).

the academic and subject vocabulary appearing among the top 500 vocables are recorded.

In such a manner, we intend to compare lemmas based on our frequency and range work with entries organized somewhat differently – e.g. according to the relevance of such sub-technical terms across a number of sources (textbooks, articles in newspapers, magazines, journals, etc). As Martínez states (1994: 81), the identification of these semi-tecnhical words is at times difficult, especially when these items receive a more restricted consideration in specific subjects or topics. This aspect, in addition, `is not a sufficient condition for successful reading of specialised material´(Martínez, ibid). Due to pedagogic concerns, the most interesting lexis to examine is thus the one most often utilized across genres, above all, as in the case of Martínez´s list of engineering lexis, the sub-technical items that do not have Spanish cognates --e.g. **able, accurate, achieve, achieved, actual (...) weight, work, working, world** (Martínez, 1994: 89).

In this way, general items are omitted,[254] the result being found in an intermediate plane of sub-technical lexis: In our case, academic and subject-based words organized according to the contexts of the three genres (Textbooks, Technical Reports and Research Articles).

The data displayed in sample # 51 is therefore selected according to the mentioned parameters. The enumeration follows the total number of texts where the items appear as a first priority, and, as a second requirement, the number of instances computed in the genre of research articles (appendix 4). Two dictionaries are then employed to level the overall use of the academic and subject vocabulary across our corpus: *Dictionary of Information Technology* (1997) --hereafter abbreviated as DIT--, and *Dictionary of Information Science and Technology* (1992) –DIST.[255]

[254] Delexical or indexical verbs are thus also left out –e.g. **have** and **make**.

As in the lexical comparisons below, attention has been paid to choosing monolingual works that furnish a useful amount of collocations. Only two dictionaries, DCTI (Aguado de Cea, 1996) and VEBD (Lozano Palacios, 1999), are bilingual, but they equally provide relevant multi-word terms and clusters (samples # 53 and 55), being our main interest in this work.[256]

The following are the items found among the top 100 words in our overall corpus of Information Science and Technology (IST).[257] When contrasted with the sources, the check mark □ means that the word is included in that work, while the symbol □ refers to the fact that the given item does not show up in the source.

IST	DIT	DIST
Information	□ □	
Web	□	□
Html	□	□
System	□	□
Data	□	□
Number	□	□
Internet	□	□
Library	□	□
Network	□	□
Access	□	□
Text	□	□
Document	□	□
Use	□	□
Digital	□	□
Interface	□	□
Object	□	□

[255] On-line sources covering wider fields not only in science and technology but academic, journalistic and popular writing are discarded due to the need to focus on Information Science and Technology English. In this respect, a very useful database such as *Wordnet* (http://www.cogsci.princeton.edu/~wn) is not consulted, although it provides extensive coverage of word meanings and relationships in General and specific English – e.g. the word **access** is specified in its fifth sense as the multi-word term *memory access* for Computer Science. Neither are the scientific-technical sub-corpora of the Brown and LOB Corpus examined due to their broad scope in science and technology, lacking demarcated application for our purpose (both can be available at http://www.clres.com/dict.html#onto).

[256] The VEBD can also be accessed via Internet, including a IFLA Glossary (Library Science) and a Thesaurus (http://eubd1.ugr.es/tony/risweb.isa.

[257] See Appendix 7 for the contrastive enumeration of these and other words among the first 500 items in our corpus.

Page	□	□
Available	□	□
Images	□	□
Application	□	□
Language	□	□
Design	□	□
Message	□	□
Computer	□	□
Knowledge	□	□
Electronic	□	□
People	□	□
Performance	□	□
Different	□□	
Based	□□	
File	□	□
Form	□	□
End	□	□
Development	□	□
Particular	□	□

Sample # 51: Contrastive view of the top sub-technical items in our corpus (IST).
DIT = *Dictionary of Information Technology* / DIST = *Dictionary of Information Science and Technology*.

As expressed above, since the purpose of this contrastive view is not to assess the absence of data in our corpus or in the lexicographic sources, a step must be taken to examine the shared clusters and word combinations. In this respect, for the top 10 items from above, taking the entries in the `Genres´ section and Appendix 5 as references, the DIT and the DIST contain some collocates that also appear in our work (sample # 52).

DIT

Information content / / information networks / / information processing /
Information providers / / information retrieval / / ***information storage*** */ /*
Information system / / information technology / / information transfer / /

Expert system / / operating system / / system design / / system variables / /

Data channels / / data collection / / data flow / /
(electronic) data interchange (format) / / data service / / (abstract) data sink / /
(abstract) data source / / data stream / / data structures / /
data type / / data transfer / /

Number crunching */ /*

Internet relay chat / / Internet service provider / /

Local library / / ***library catalogue*** */ / library program / /*

Network adapter / / network administrator / / network analysis / /
Network control program / / ***network drive*** */ / network interface (card) / /*
Network layer / / network hardware / / network management / /
Network operating system / / ***network protocol*** */ /* ***network server*** */ /*
Network software */ / network structure / / local area network / /*
Wide area network / /

have access to / / access control / / ***access mechanism*** */ /*
access method / / access permission / / access point / /
access provider */ / access rights / /* ***access unit*** */ /*
direct memory access / / public access terminal / /

text editor / / text file / / text manipulation (facilities) / /
text processing / / text retrieval / /

document delivery (service) / / document image management / /
document interchange (architecture) /

DIST

Information age / / information content / / information retrieval / / Information society / /

Data structures / / data (switching) exchange / / data type / /

Network model */ / network node / /* ***network performance*** */ /*

Digital video (interactive)

Object-oriented (database)

Computer conferencing

Knowledge base / / knowledge engineer / /

Electronic funds transfer / / electronic mail / / electronic publishing / /

Sample # 52: Collocates registered by the sources appearing in our IST corpus.
In bold = low frequency collocations (2 instances or less) – the same applies to the rest of the samples (54, 56, 58, 60 and 62).

In the second dictionary –DIST--, only a total of eight entries are revised, since the remaining ones do not contain any collocates, or if they do, these fail to surface in our corpus –e.g. *text fragment* for **text**.[258] In turn, those combinations recorded in both bold and italics –e.g. ***information storage*** in DIT—have an occurrence lower than three instances in our corpus –i.e. one or two. The words in brackets, in addition, are provided by the given authority but do not appear in our work.

As can be surmised, most forms are subject-based multi-word terms, bound to appear in specific fields, such as *object-oriented database* – recurrent in On-line database systems and Computer systems texts. Their combinations tend to be restricted, sometimes giving origin to acronyms, such as **LAN** (*Local Area Network*) and **EFT** (*Electronic Funds*

[258] The exclusion of these lexical items is thus owed to the fact that the given source does not provide any collocates for that lemma (e.g. **html**, **Internet**, etc). Word combinations are arranged alphabetically –as displayed by the dictionaries--, in opposition to the entries in the results from the previous part and from appendices 3, 5 and 6, where they follow the main criteria of frequeny and range status.

Transfer). In our corpus –IST--, both forms, abbreviated and full, co-exist.

The two dictionaries consulted can thus furnish a suitable layout for comparison with our overall corpus in terms of corraborating these multi-word units appearing across texts, distributed according to the subject parameters exposed in the previous section. In order to take a closer introspection into the type of co-occurrences most frequent and characteristic of each separate discipline, however, the method should be extended in the process of viewing specialized lexical material for contrastive purposes.

IV.2. COMPUTER SCIENCE

In the case of this broad subject field –as in those of Telecommunication and Internet below (D and F respectively)--, many sources are found in an electronic format. This is mainly due to the changing nature of devices and developments, updated relatively fast, fallen soon obsolete, etc. In fact, in the context of computing, the relationship between language and social concept is of prime importance, as is shown in Cony-Beare's Master dissertation (1986), its `computerese´ register or jargon now regarded as somewhat old-fashioned, after nearly 15 years. Such changes are denoted as characteristic of computer terminology in Aguado de Cea (1996). She observes these by the rapid adaptation and borrowing of newly coined words from English into Spanish: `algunos neologismos que aparecieron en la primera época de los ordenadores (...) se han quedado prácticamente arrumbados debido al descubrimiento de nuevos inventos´(Aguado de Cea, 1996: 7).

Because of this concern, several Internet resources are thus devoted to compiling technical glossaries and terminology that serve the purpose of explaining computer notions and advances; this involves updating linguistic forms and utterances. From this perspective, the hypertextual body of texts created at http://www.pcwebopedia.com deal with a large

amount of computing terms by means of linked interrelationships. The possibility of searching and identifying their explanation is thus very productive and enriching for reading purposes in English for Computer Science; yet, few lexical combinations can be singled out.

A much more suitable electronic lexical resource is the *On-line Dictionary of Computing* (FOLDOC), by Denis Howe, accessible at, among other locations, http://www.aanet.aanekoski.fi/foldoc/foldoc.html. It serves as a database of entries on Computer studies that correspond to the main concepts and developments in the field. This work, in addition to other paper form dictionaries and lists, make up the elements of comparison for our Computer Science sub-corpus.

For this and the remaining four disciplines sub-corpora (Information Science, Telecommunications, Audio-visual Communication and Internet), the procedure has been that of identifying those texts included in a given subject, e.g. Computer architecture (A1) which the curricula of Computer Science and Telecommunications (Optical / radio communication) share. Then, the contents of the source are examined in order to determine from which discipline it would most likely originate; this is mainly done by scanning titles, keywords, prefaces, goals, sources and any other pointers to the overall subject field.

For example, in the so-called category A1 (History of Computers, Hardware, Software), the article "Using information systems while performing complex tasks : an example from architectural design" is published in the journal *Computer Studies* vol 46 (1997: 31-54)). The name of the publication mainly contributes in this case to help us label the article within the sub-corpus list of sources for Computer Science. On other occasions, however, this type of resolution is not so clear, as is the case of a report found with this same heading –A1. Dealing with the Minix operating system, its introductory remarks signal that it

belongs to the area of Computer Science: `MINIX (...) is well suited to people who want to run a UNIX-like system on their personal computer and learn about how such systems work inside´. In addition, such a topic is found in the course syllabus for `Introducción a los computadores´, a subject taught in Computer Science at our institution --`History of computers´ is the name given in our subject categories.[259] In this manner, all such information and hints from both texts and programs of studies contribute to the decision process of managing our corpus from this discipline-based perspective.

Sample # 53 collects the data taken from the top 100 items in our Computer Science word frequency list (CS) and their comparison, as mentioned above, in six works: the *FOLDOC dictionary* (1999), the *Oxford Dictionary of Computing* (1996) –ODC--, the *Dictionary of Computing* (1999) –DC--, the *Diccionario comentado de terminología informática* (1996) –DCTI--, the *Dictionary of Personal Computing and the Internet* (1998) –DPCI--, , and the *English in Computer Science* corpus of the Hong Kong University of Science and Technology (1994) --HKUST. Such sources are aimed at `any user who wants to understand the complex terms that are used to dercribe PCs´, as Collin states in the DPCI (1998: Preface).

CS	**FOLDOC**	**ODC**	**DC**	**DCTI**	**DPCI**	**HKUST**
23.- Information	□	□	□	□	□	□(64)
26.- Project	□	□	□	□	□	□(165)
30.- Program	□	□	□	□	□	□(29)
32.- Number	□	□	□	□	□	□(45)
34.- System	□	□	□	□	□	□(27)
37.- Used	□	□	□	□	□	□(38)
42.- Design	□	□	□	□	□	□(67)
43.- Problem	□	□	□	□	□	□(106)
44.- Knowledge	□	□	□	□	□	□(250)
61.- C	□	□	□	□	□	□(86)

[259] In the Corpus References after the Bibliography, both the subject category and the source discipline are labelled –e.g. A1 / CS for these two works.

62.- Software	□	□	□	□	□	□(83)
63.- Network	□	□	□	□	□	□(179)
67.- Data	□	□	□	□	□	□ (21)
69.- Performance	□	□	□	□	□	□(338)
70.- Test	□	□	□	□	□	□(204)
79.- Task	□	□	□	□	□	□(267)
80.- Digital	□	□	□	□	□	□(1517)
83.- Tools	□	□	□	□	□	□(467)
89.- Case	□	□	□	□	□	□(115)
94.- Code	□	□	□	□	□	□(120)
98.- Set	□	□	□	□	□	□ (74)

Sample # 53: Contrastive view of our most frequent Computer Science sub-corpus (CS) items. FOLDOC = *Online Dictionary of Computing* / ODC = *Oxford Dictionary of Computing* / DC = *Dictionary of Computing* / DCTI = *Diccionario comentado de terminología informática* / DPCI = *Dictionary of Personal Computing and the Internet* / HKUST = *Hong Kong University of Science and Technology Corpus of Computer English*.

The numbers on the CS and HKUST columns refer to the positions which the lexemes occupy in their respective frequency lists. Such values are contrasted if a given corpus presents such a utility.[260] As can be checked, the closest numeric equivalent is given by the word **program** (ranked 30th in our CS corpus and 29th in the HKUST), while the adjective **digital** registers the broadest difference (80 vs. 1517 occurrences).

For the observation of collocates (sample # 54), the co-occurrences are seen first in the CS sub-corpus. These are listed as long as they also co-occur in the five dictionaries above. The HKUST corpus is the only one not providing any word combinations.

[260] This is also done in the contrastive approach offered by Information Science sources (sample # 55).

FOLDOC

Information highway / / **information management** / / information processing / /
Information system / / ***information superhighway***

Operating system

Design process

Knowledge base / / *knowledge representation* / /

C language / / *C++* / /

Embedded software

Local area network / / *wide area network* / / *ethernet network* / /

Data processing

Digital data

Source code / / *Machine code* / /

DPCI

System software

C++

Computer network / / *local area network* / / *network adapter* / /
Network adapter / / *network address* / / ***network computer*** / /
Network operating system / / ***network server*** / / ***network time*** (*protocol*) / /

Data compression / / ***data rate*** / / ***data set ready*** / / ***data terminal ready*** / /

Digital signature / / **digital video** (interactive) / /

ODC
Information processing / / *information retrieval* / / *information structure* / / *information technology* / /

System development / / ***system software*** / / ***system support*** / /

Design proposal / / ***design stage*** / /

Knowledge base / / ***knowledge-based system*** / / *knowledge engineer* / /

C compiler / / *C program* / /

Software development / / *software engineering* / / *software tool* / /

Network layer

Data collection / / data compression / / data link / / data management / /
Data path / / *data processing* / / ***data rate*** / / ***data set ready*** / / *data structure* / /
Data terminal ready / / *data transfer* / / *data type* / /

Case tools

Character code

Data set

DC

Information processing / / *information retrieval* / /
Information system / / *information technology* / /

user program / / ***program register*** / /
program specification / / *program testing* / /

computer system / / *expert system* / / *information system* / /
operating system / / ***system control*** (***panel***) / / ***system software*** / /
system support / / ***system variable*** / /

solving a problem

knowledge-based system

C++

System software / / *software development* / / *software engineering* / /
Software package / / ***software system*** / / *software tool* / /

Computer network / / *local area network* / / *network adapter* / /
Network administrator / / *network control program* / /
Network layer / / *network management* / / ***network operating system*** / /
Network server / / *network structure* / / *wide area network* / / ***network computer*** / /

Data access / / ***data area*** / / ***data collection*** / / ***data level*** / / ***data link*** / /
Data management / / ***data manipulation technology*** / / ***data path*** / /
Data processing / / ***data rate*** / / ***data set ready*** / / *data structure* / /
Data terminal ready / / *data transfer* / / *data type* / /

System performance / / ***high performance*** / /

Multitasking workstation

Digital data / / *digital signature* / / ***digital video*** (*interactive*) / /

Machine code / / **object code** / / source code / / **bar code** / /

DCTI
Application program / / *program specification* / /

Expert system / / ***knowledge-based system*** / / ***on-line system*** / /
Operating system / / ***real-time system*** / / *system specification* / /

Computer-aided design / /

Knowledge-based system / /

Software engineering / / *software maintenance* / / *software package* / /
Proprietary software / / *software tools* / /

data buffer / / *data bus* / / ***data file*** / /
data processing / / *data structure* / /

software tools / /

instruction code / /

instruction set / /

Sample # 54: Collocates shared by our CS sub-corpus and the Computer English dictionaries (**FOLDOC**, **ODC**, **DC, DPCI** and **DCTI**).

As pointed out above, the frequency factor is not taken as the primary condition for the arrangement of lexical items, but their mostly alphabetized disposition in the dictionary entries. Those constructions shown in bold and italics are the least frequent ones –below three instances in our Computer Science sub-corpus. As above, the reason for not having some of the items present in these collocational charts –e.g. **project** in the FOLDOC dictionary--, is that the word is either presented as mere definition in the dictionary or exemplified in word combinations which do not occur in our sub-corpus.

Providing feedback in the form of collocates, these works can be taken as suitable models or reference to check the kind of lexical items detected. This information points to what extent a given type of collocate may or may not occur –e.g. less restricted, such as *network adapter* and *expert system* vs. more text-bound combinations, exemplified by *data access* and *computer system*, for instance. The latter type tends to be less frequent in our body

of texts, whereas, in general terms, multi-word terms usually register a greater amount of repetition in a text –e.g. *network adapter* is defined in a textbook on Computer Science concepts and in an article -- "Building Tcl-Tk GUIs for HRT-HOOD Systems". The lexical item *data access*, in contrast, refers only once to the process of exploring documents; this occurs in the article on data structures, "Mutual Search".[261]

Other examples of this sort are embodied by contrastive pairs such as *expert system* vs. *computer system*, as the observation of their co-texts may reveal in figures 21 and 22:

1	verification of the (design of a) database or **expert system.** The current
2	in 1984. Since then, it has been used to develop and field over 7500 **expert system** applications. Commercial and Industrial ARCO,
3	in that the CFP for EUROVAV'99 (The principal international **expert system** V&V biannual symposium) includes databa
4	much longer than another, either by a human judgment or using an **expert system**. For instance, task (a) is much shorter than any of t
5	presentation of knowledge (elicited from an expert in the case of an **expert system**) using rules, facts, procedures, and objects. In additio
6	Knowledge Representation. M.4's Complete **expert system** language includes backward and forward chaini
7	of the KB component. Although these methods are currently used in **expert system** development, it is not yet clear how they would
8	rule sets of a similar nature to that found in **expert system** rule bases. In theory, such hybrid
9	into software applications. A knowledge system (also known as an **expert system**) uses human knowledge and experience to solve probl
10	Support Teknowledge has been a leading **expert system** product and services company for over ten years. With
11	solution. The Embeddable **Expert System** Solution

Figure 21: Co-texts of *expert system* across a diversified number of texts.

1	Memory caches and disk caches are used to speed up **computer system** performance, the details of which are outlined in
2	SCSI (Small **Computer System** Interface) Pronounced "scuzzy." SCSI

Figure 22: Co-texts of *computer system* across a more reduced textual space.

[261] As can be examined in the Corpus references, the texts have been published recently (1999), a factor which contributes to highlighting the degree of updatedness of the concepts and developments being expressed. The DC is also quite recent, and that makes the contrast more reliable.

IV.3. INFORMATION SCIENCE

For this subject area embracing both Library Science and Documentation, the sources examined are three: *Elsevier's Dictionary of Library Science Information and Documentation* (1992) –EDLSID--, the *Dictionary of Library and Information Management* (1997) –DLIM--, and the *Vocabulario para los estudios de Biblio-documentación* (1999) –VEBD--. Our Information Science (IS) list is developed, as in previous cases, by identifying the most frequent content academic and subject vocabulary found among the top 100 items (Sample # 55).[262] In the case of VEBD, the numbers correspond to the positions of the lexemes in the source's frequency list.

IS	EDLSID	DLIM	VEBD
16.- Library	□	□	□(10)
17.- Information	□	□	□(12)
21.- Web	□	□	□(5,915)
22.- Internet	□	□	□ (29,496)
27.- Html	□	□	□(none given)
28.- Use	□	□	□(38)
29.- Public	□	□	□(39)
36.- Access	□	□	□(91)
43.- Electronic	□	□	□(223)
44.- Project	□	□	□(156)
49.- Document	□	□	□ (224)
50.- Services	□	□	□(36)
57.- Sites	□	□	□
58.- Copyright	□	□	□(696)
59.- Text	□	□	□(296)
60.- Knowledge	□	□	□(267)
63.- Network	□	□	□(165)
66.- Resources	□	□	□(214)
74.- Data	□	□	□(31)
81.- Support	□	□	□(290)
83.- Available	□	□	□(112)
85.- Training	□	□	□(153)
87.- System	□	□	□(34)
93.- Community	□	□	□ (120)
99.- Computer	□	□	□(99)

[262] As above, these are the lexemes based on the highest frequency; thus, a word-form such as **libraries** – occurring on 25th place-- is not included, since **library** is already counted (16th).

Sample # 55: Contrastive view of our most frequent Information Science (IS) sub-corpus lexis with lexicographic sources. EDLSID = *Elsevier's Dictionary of Library Science Information and Documentation* / DLIM = *Dictionary of Library and Information Management* / VEBD = *Vocabulario para los estudios de Biblio-documentación.*

VEBD provides the greatest amount of both lemmas and multi-word forms. As displayed, some examples of words that are numerically closer in the frequency lists of both our IS set and the VEBD corpus, are the items **library**, **information**, **use** and **public**. However, as with HKUST in sample # 53, our sub-corpus is minimal by comparison, thus providing a much lower significance rate of accounts.

In the observation of collocates, as a result, the data in sample # 56 reflects the key lexical elements shared by our source with the lexicographic authorities. The three sources provide significant collocations and clusters, as can be gleaned. VEBD offers the widest scope, registering rather useful constructions, such as *convey information*, *make use of, gain access* or *provide support.* For pedagogical purposes, these are rather productive semi-restricted combinations making up the general academic discourse of information-related areas. The use of such a type of clusters is a main discriminating factor for the identification of vocabulary registers, as will be seen in chapter 4.

A few sample concordance lines may reveal this contrast among the uses of verbal constructions,[263] such as the pattern *provide* + *access* (figure 23), co-occurring across different sources in our IS sub-corpus.

1	accessing other types of resources. [26] The University of Virginia **provides access** to many collections of resources using clients
2	described one dilemma of the historian as a tight-rope act between **providing access** to evidence and creating an organizing argument
3	uter resource. In this sense, the net extends the ability of a library to **provide access** to information in a way that most people are
4	t a particular site. Hypertext links are also used to highlight footnotes, **provide access** to support materials, and to define words or
5	ers and limit the use of in-line images as much as possible. When **providing**

[263] As Hornby (1974: 206) states, these verb patterns are regarded as the most important element in sentence patterns.

access to large images, provide the user with a
6 sources in networked information services, e.g., cost per session in **providing access** to remote users of an online catalog, or average
7 formats and interchange. **Providing access** and retrieval capabilities to electronic repositories
8 ML defines glossaries and glossary entries but, in order to **provide access** to an individual glossary entry from a hypertext link
9 .lib.virginia.edu/\par etext/scanner.html.)\ Using the Web to **provide access** to training materials has many strengths. It gives
10 er. Smaller\par glossaries cause few problems, but this makes **providing access** to individual "glossary" entries in a document such
11 uld become either 2,500 files (if the administrator wished to **provide access** at the volume level) or 65,000 files (if access to
12 plexity is often closely linked to the inability of Web servers to **provide access** to complex representations of documents. This
13 OPAC for Intranets. - OPAC for Intranets **provides access** to the library catalog via the open infrastructure of the
14 physical library facilities. What will be more likely is that libraries will be **providing access** to 3-D models of anatomic parts. The practicality of
15 nce of electronic information and the role public libraries can play in **providing access** to this information. Librarians have a history of
16 ing with, and relying on, on-line reserve readings. In addition, IUPUI is **providing access** to these documents from all workstations in the
17 t cable television will increasingly incorporate the World Wide Web and **provide access** for many currently under-represented groups, such
18 to meet the rapidly changing requirements of the digital library and to **provide access** to electronic documents regardless of format or
19 new, for example, that Cornell wanted to explore new methods of **providing access** to its materials. At that time, the on-line card catalog
20 collection need not resort to fragmenting files to make it possible to **provide access** to the component parts of a collection. As
21 Key Concepts This approach to using the Web to provide access to complex textual resources involves many tools

Figure 23: Collocational span of the cluster *provide* + *access* on the IS sub-corpus.

A more general inference may be drawn on observing these co-texts in figure 23. The assertion on the fact that the colligation with the preposition **to** follows the verbal pattern may be well-founded, as the majority of the lines show. In addition, except for the cases of rows 2, 6, 11, and 17, it is equally clear that the construction *provide access to* generally evokes, in

our Information Science context, admission into some sort of collected information. Thus, the semantic marker [+ information] may be added to the lexical item.

Examples of this and other patterns may be sought in the lexicographic sources (sample # 56). The aim is to check that more or less restricted patterns of use are likely to be encountered and explored by means of dictionaries such as the ones applied. In some cases, as in that of specialized subject-based references to concepts on Library Science and Documentation studies – e.g. *information needs*, *data element* or *computer-aided design*--, explanations and examples of usage become rather helpful. In other circumstances, the employment of some lexical items is easier to predict: for instance, *public library*, *electronic document* or *library collection.*

At a middle position, still, we may find those words that, like *data collection*, co-occur with other items unnoted by the lexicographic sources. Such an item, in fact, is often followed by **instruments**, **methods** and **techniques** in our set of texts, and these collocates contribute substantially to situate the item's contextual meaning. In contrast, other words, like *electronic data*, may seem easy to comprehend, and yet, their special associations, as in *electronic data interchange*, refer to a specific concept. This item, followed by *processing*, like the multi-word term *document image processing*, is not recorded in our sub-group of IS sources. As a result, this fact is indicated by means of brackets surrounding the word unmarked in our work – e.g. *electronic data* (*processing*).

EDLSID
Library administration / / ***library management*** / / *library association* / /
Library board / / *library building* / / *Library of Congress* / / ***library personnel*** / /
Library staff / / ***library school*** / / ***library science*** / / ***library profession*** / /

Information center / / *information centre* / / *information content* / /
Information processing / / *information retrieval* / / *information science* / /
Information space / /

Public library

Electronic data (***processing***)

Copyright notice

Data bank / / ***data base*** / / ***data center*** / / ***data compilation*** / / ***data element*** / /
Data gathering / /

System management

DLIM
Academic library / / *local library* / / *public library* / / *special library* / /
Library and information science (*abstracts*) / / *library association* / /
Library of Congress catalog / / ***Library of Congress Classification system*** / /
Library card / / ***library user*** / /

Information centre / / *information content* / / *information management* / /
Information manager / / *information needs* / / *information network* / /
Information retrieval / / *information science* / / *information services* / /
Information system / / *information technology* / /

Web browser / / *web page* / / *web site* / /

Internet service provider

In use / / *be of use* / /

Public domain / / *public library* / / ***public record*** (***office***) / / ***public sector*** / /
Public service / /

Open access

Electronic journal / / *electronic library* / / *electronic mail* / / *electronic publishing* / /

Project manager / / ***project team*** / /

Document image (***processing***)

Copyright infringement / / *copyright law* / /

Text retrieval

Knowledge-base

Data collection / / ***data analysis*** / / ***data bank*** / / ***data entry*** / /
Data management / /

Training materials

Operating system / / *system design* / /

Computer-aided design / / ***computer conferencing*** / / ***computer literate*** / /

VEBD

Academic library *//* *American Library Association //*
British library // *central library //* *city library //* ***corporate library*** *//*
County library *//* *electronic library //* *inter-library loan //* *local library //*
Main library *//* *national library //* *public library //*
Regional library // *special library //* *state library //*
University library *//* ***urban library*** *//*
Library and Information Science // *Library and Information Studies //*
Library association // *library building //* ***library card*** *//* *library catalog //*
Library collection // *library community //* ***library management*** *//*
Library materials // ***library network*** *//* *Library of Congress //*
Library of Congress Classification *//* ***library operations*** *//* ***library patron*** *//*
Library personnel *//* ***library profession*** *//* *library research skills //*
Library resources *//* *library service //* ***library's mission*** *//* ***library software*** *//*
Library staff // *library system //* ***library use*** *//* ***Library user*** *//*

Background information *//* *bibliographic information //* ***business information***
Computerised information *//* ***convey information*** *//* ***descriptive information***
Electronic information service // ***exchange of information*** *//* ***full-text information***
furnish information *//* *geographic information systems*
Government information // *Library and Information Science*
Library and Information studies // *national information*
Online information *//* *public information //* ***price information***
Provide information // ***recorded information*** *//*
Source of information // ***unit of information*** *//* ***information broker***
information content // ***information exchange*** *//* ***Information format***
Information highway *//* *information management //* *information needs*
Information poor *//* *information processing //* *information product*
Information profession *//* ***information professional*** *//* *information provider*
information provision *//* ***information seeker*** *//* ***information seeking***
information skills *//* *Information services //* *the information society*
Information sources // *Information specialists //* ***information superhighway***
information system *//* *information resources //* *information retrieval*
information technology // ***information world***

world wide web // *web browser //* *web page //* *web site //*

MIME (Multipurpose Internet Mail Extension) //
SLIP (***Serial Line Internet Protocol***) *//*
IETF (Internet Engineering Task Force) //

be of use // *fair use //* *in use //* *library use //* ***make use of*** *//*
put to use *//* *use for //*

make + public *//* *general public //* *public access //* *public library //*
public policy *//* ***public transport*** *//*

direct access / / ***gain access*** / / ***online public access catalogue*** / /
open access / / *provide access* / / *public access* / / ***random access*** / / ***remote access*** / / ***subject access*** / / ***access capabilities*** / /

electronic age / / *electronic book* / / *electronic document* / /
electronic format / / *electronic journal* / / *the electronic library* / /
electronic mail / / ***electronic material*** / / *electronic media* / /
electronic publishing / /

ARPAnet / / *pilot project* / /

Electronic document / / ***historical document*** / / ***image-based document*** / /
original document / / ***parent document*** / / ***request + document*** / /
Source document / / ***document content*** / / ***document image*** (***processing***) / /
document management software / / *document management system* / / ***document number*** / / ***document profile*** / /

adult services / / (***information and***) ***referral services*** / / *library services* / /
(***library***) ***support services*** / / (*value added*) *network services* / /

the Copyright Act / / ***copyright infringement*** / / *copyright law* / /
copyright office / /

full text / / (*stand out*) *in the text* / / *text-based* / / *text file* / /
text information / / ***text-oriented*** / / *text retrieval* / / *full text retrieval* / /
text retrieval system / /

acquire + knowledge / / ***area of knowledge*** / / ***gain + knowledge*** / /
intimate knowledge / / ***technical knowledge*** / / ***knowledge base*** / /
knowledge management / /

computer network / / ***cooperative network*** / / (*value added*) *network services* / /

data resources (***Inc.***) / / *information resources* / / ***library resources*** / /

bibliographic data / / ***capture + data*** / / ***collect + data*** / /
numerical data / / ***data base management*** (***system***) / / ***data capture*** / /
data collection / / *data collection instrument* / / ***data element*** / / ***data file*** / / ***data items*** / / ***data loss*** / / ***data resources*** (***Inc.***) / /

financial support / / ***provide + support*** / / ***service support*** / / *support staff* / /
support structure / /

be + available / / *be + available from* / / *make + available to* / /

staff training / / ***training facility*** / / *training program* / /

automated system / / ***bulletin board system*** / / ***circulation system*** / /
class system / / *computer system* / / ***database*** (***management***) ***system*** / /
educational system / / ***information system*** / / ***integrated system*** / /
library system / / *management system* / / ***markup system*** / /
operating system / / *reference system* / / ***retrieval system*** / / ***software system*** / /
transport system / /

academic community / / ***business community*** / / *library community* / /
rural community / / ***community college*** / / ***community education*** / /
community library / /

notebook computer / / ***computer-aided design*** / /
computer-based / / *computer center* / / ***computer conference*** / /
computer conferencing / / ***computer-generated*** / /
online computer library center / / ***computer language*** / /
computer literacy / / ***computer network*** / / *computer science* / /
computer-supported / / *computer system* / / ***computer technology*** / /

Sample # 56: IS- based co-occurrences in our corpus found in the dictionaries (**EDLSID**, **DLIM**, and **VEBD**).

IV.4. TELECOMMUNICATIONS

The reference material for this sub-corpus primarily comes from on-line glossaries and databases of specific terminology. In such a respect, the contrasted data is highly restricted, as the technical formations in sample # 58 may prove. Although our sub-group chiefly contains documents dealing with optical and radio communication –the specialty studied at our institution--, the sources tend to encompass all areas of Telecommunications Engineering.

The dictionaries and glossaries handled are the *Glosario práctico de telecomunicaciones* (1986) –GPT--, *Small Glossary of Terms: Vocabulary 563* (1997) –Vocab563--, and *the Termite Database* (1999) –TERMITE. Our frequency list of the Telecommunications texts (Tel) contains the works labelled at the Corpus References by means of such an abbreviation. Sample # 57 displays the annotation of our top 100 content entries according to the lexical resources:

Tel	GPT	Vocab563	TERMITE
Object	□	□	□
Network	□	□	□
Type	□	□	□
C	□	□	□
System	□	□	□
Data	□	□	□
Method	□	□	□
Standards	□	□	□
Call	□	□	□
Address	□	□	□
Surrogate	□	□	□
Use	□	□	□
Program	□	□	□
Client	□	□	□
Interface	□	□	□
Information	□	□	□
Space	□	□	□
Applications	□	□	□
Code	□	□	□
Example	□	□	□
File	□	□	□
Procedure	□	□	□
Work	□	□	□
End	□	□	□
Server	□	□	□
IEEE	□	□	□

Sample # 57: Contrastive view of our Telecommunication (Tel) frequency list with lexical resources. GPT = *Glosario práctico de telecomunicaciones* , Vocab563 = *Small Glossary of Terms: Vocabulary 563* , TERMITE = *Termite Database*.

Let us next see the collocates of the words appearing in both our body of texts and the entry articles of the dictionaries. As above, where no co-occurrences are shown, the given lexicographic work does not present any matches. In the case of Vocab563, none of its few terms registers any of our collocations. TERMITE, in contrast, contains a wide scope of specialized multi-word units, some of which are collected in Tel, enumerated in sample # 58. The arrangement of the items follows, as above, the order given by the authors – although, in this case, it is not alphabetical.

GPT

Local area network / / ***heterogeneous network*** / /

Message type

Centralized system / / (***service***) ***management system*** / / *distributed system* / / *Telephone system* / / *operating system* / /

(*basic*) *access method*

source program

user interface

line code / / ***redundant code*** / /

TERMITE

Space object / / object-oriented programming / / class object / /
Null network object / / object type / / ***object request*** */ /*

U-network */ / network services / / network failure / /*
Communication network / / ***network system*** */ /* ***network information*** */ /*
Network services protocol / / ***area network*** */ /* ***network operating system*** */ /*
Network service / / ***network interface*** */ /* ***network configuration*** */ /*
Access network */ /* ***network management*** */ /*

Module type */ / message type code / / type test / / data type / /*
Object type / / message type / /

Support system / / communication system / / system design / /
Programming system / / ***multiprogramming system*** */ /* ***cable system*** */ /*
Network system */ / system technology / / operating system / /*
Telephone system / / ***communications system*** */ / telecommunication system / /*
Video display system */ /* ***system data*** */ /* ***radio system*** */ /*
System development / / fiber optic system / / system performance / /
Storage system / /

Data flow / / data conversion / / no data / / data analysis / /
Data transmission / / ***data available*** */ / data collection / /* ***data signal*** */ /*
Data processing */ /* ***digital data*** */ / data transfer / /* ***data entry*** */ /*
Image data / / ***data record*** */ / voice data / /*

Access method / / method of / / current method / / null method / /

(*American*) *Standards Association / /* ***national standards*** */ /*
television standards */ / standards development / /*

call failed / / procedure call / /

address start */ / memory address / / destination address / /*
mail address / / ***address information*** */ / starting address / /*
address user / / ***base address*** */ / single address / / address provided / /*
address requests */ /*

in use / / for the use of / / use of communication / /

application program / / ***program status*** / / ***program structure*** / /
program interface / / ***program specific*** / /

application interface / / *interface requirements* / / *standard interface* / /
network interface / / ***digital interface*** / / *user interface* / /
interface specification / /

information highway / / *information processing* / / *information handling* / /
information system / / *information science* / / ***background information*** / /
information provider / / *information transfer* / / ***network information*** / /
information services / / ***information retrieval*** / / *global information* / /
digital information / / *information technology* / / ***information service*** / /
address information / / ***information storage*** / / ***community information*** / /
federal information processing / / *information infrastructure* / /
specific information / / ***date information*** / / ***color information*** / /
system information / / *routing information* / /

space communication / / *space object* / / *free space* / /

small applications

machine code / / *code division* / / ***oriented code*** / /
identification code / / ***redundant code*** / / *line code* / /
type code / /

file data / / ***file format*** / / ***file management*** / /

test procedure / / *remote procedure call* / / ***analysis procedure*** / /
linking procedure / / ***registration procedure*** / / *procedure for* / /

(***part***) ***time work*** / / *development work* / /

end of data / / *end of message* / / *end user* / / ***end of file*** / /

proxy server / / ***name server*** / / ***internet server*** / /

IEEE computer / / *Institute of Electrical and Electronic Engineers* / /

Sample # 58: Collocates shared by our Tel sub-corpus and the resources on Telecommunication engineering (**GPT**, **Vocab563**, and **TERMITE**).

There seems to exist a greater use of technical multi-word terms, such as those listed in the TERMITE database. However, they tend to be mostly represented by semi-technical contructions in our sub-corpus, combining both the specialized item and an easy-to-understand reference to the concept. Thus, the combination *space object*, for instance, is part of the longer cluster *space object tracking radars* in a radio communications report, whereas *free space* is employed within the phrase *bytes of free space in the buffer*. As we shall see in the next chapter, this type of co-text demarcation often proves to be rather useful in dealing with the clarification of meaning.

IV.5. AUDIO-VISUAL COMMUNICATION

For this and the next category, the focus is placed on audio-visual and internet technology respectively. As in the case of the *Dictionary of Personal Computing and the Internet* (DPCI -- 1998) above (sample # 53), the study of terminology involving equipment and resources must also encompass computer devices --`the desktop revolution´, as Gonzalo (1999: 51) states. Hence, the work with technical issues tends to bring together computing and technology.

Regarding audio-visual media, in this sense, the *Dictionary of Multimedia* –1997— (DM) turns out to be rather helpful as a reference source. It is the only one being contrasted with our Audio-visual sub-corpus frequency list (Au) below (sample # 59). A different authority, the *Dictionary of New Media: Film, Television, Print, Digital, Internet, Multimedia* (1999) is actually more comprehensive, but offers a similar specialization on the given area of information technology, not adding any significant collocates. Instead, it expands on the

subjects of film and mass media, which are seldom dealt with by the texts in our set `Au´.[264]

For our purposes, as a result, the aforementioned work (DM) provides enough contrastive data.

Au	**DM**
Digital	□
Information	□
Data	□
Object	□
Use	□
Network	□
Text	□
System	□
Image	□
C	□
Type	□
Library	□
E	□
Example	□
Standards	□
Work	□
Access	□
Number	□
Media	□
Method	□
Electronic	□

Sample # 59: Contrastive view of our Au sub-corpus frequency list with a dictionary on Audio-visual media. DM = *Dictionary of Multimedia* .

The terms in our collection that the dictionary compiles are reflected in sample # 60.

[264] Only the article "Computer-Assisted Film and Drama Studies" reports on film and theater work; yet, its focus is on the audio-visual technology available in those areas – thus, rather, it belongs in the D1 category (media technology).

DM

Digital audio tape / / ***digital camera*** / / *digital data* / / *Digital video* / /

Data interchange (***format***)

Linked object

Text file

Image storage space

Media server

Electronic data interchange / / *electronic mail* / / *Electronic publishing* / /

Sample # 60: Multi-word units available in DM that also appear in our Au sources.

As may be surmised, the entries mainly make reference to network and internet instruments through which to access, store, and produce audio-visual information. Some co-texts which can reflect these activities are, for instance, those of *linked object*, *digital camera* and *media server* in sentences such as those seen in figure 24.

A click of a mouse key then enables one to move quickly from one **linked object** to another.

Individuals with access to a scanner or a **digital camera** can scan images

For most, it is necessary to save the audio or video file in a proprietary format and then upload it to a Web server or to a proprietary **media server**.

Figure 24: Examples of audio-visual technology terms in use.

IV.6. INTERNET COMMUNICATION

This final division or sub-group of texts made corresponds to those samples not fitting into a specific area of knowledge concisely, but, quite the opposite, regarded as relevant to all the disciplines. In this respect, the sources correspond, to some extent, to the `F´ or All encompassing category (see Corpus references). In addition, the texts deal with Internet applications and issues as a domain closely related to the Information science and technology environment – i.e. applicable in the four disciplines across a significant array of themes and topics (e.g. Hypertext technology, SGML, etc). This fact can also be assessed by observing the dictionaries so far consulted along the four disciplines above (Computer Science, Information Science – Library Science and Documentation--, Telecommunications – Optic and Radio Communication--, and Audio-visual Communication), and by checking in them the high occurrence of Internet – related terms – e.g. *network layer* in Computing (sample # 54), *eletronic mail* in Information Science (sample # 56), or *network services protocol* in sample # 58 (Telecommunications).

In such a respect, for our frequency list of Internet content words (Internet – sample # 61) among the top 100 items, electronically available glossaries can function as contrastive data. In this case, the lexical facilities are very numerous, and yet, rather limited in terms of providing large amounts of vocabulary –unlike the TERMITE database above, for instance. As a result, the most representative ones are chosen among the English terminology collections of Computing and Internet.[265] Others, like the *Ciber-léxico comparativo* (1999), or the *Glosario básico Inglés-español para usuarios de internet* (1999) are equally valid, but their focus is constrained in terms of collocational data, and concentrate on explanations and translation instead –obviously not the priority in our scope. In addition, Internet site

[265] This resource is located at http://reimari.uwasa.fi/comm/termino/collect/special/computing.html.

identification is also conducted by some works, like the Bucknell database.[266] Yet, its purpose is not lexicographic, but rather thematic or concept-bound.

The works thus compared are the *Webguest* (1999), the *Interactive Connections Internet Glossary* (1999) –ICIG--, and the *Dictionary of Personal Computing and the Internet* (1998) –DPCI.

Internet	**Webguest**	**ICIG**	**DPCI**
Text	□	□	□
Information	□	□	□
Use	□	□	□
Element	□	□	□
Access	□	□	□
Document	□	□	□
Example	□	□	□
People	□	□	□
Internet	□	□	□
Page	□	□	□
Network	□	□	□
Line	□	□	□
Electronic	□	□	□
Public	□	□	□
Sgml	□	□	□
Available	□	□	□
Set	□	□	□
Net	□	□	□
System	□	□	□
Time	□	□	□
Computer	□	□	□
Attribute	□	□	□
Chapter	□	□	□
University	□	□	□

Sample # 61: Contrastive view of Internet words in our sub-group (Internet) and the sources. ICIG = *Interactive Connections Internet Glossary* , DPCI = *Dictionary of Personal Computing and the Internet.*

[266] This Web facility is found at http://www.facstaff.bucknell/edu.

Some arrangement of items is provided by the listing tools in sample # 62. Most of the occurrences, as can be contrasted, have already been detected in other sources, except for the more specialized units dealing with Internet issues and terms – e.g. *Internet Address Naming Authority (IANA)* or *home page*.

Webguest *Information superhighway* *Internet Explorer* / / *internet network information center* *Network information center* / / ***network news transfer protocol*** *Electronic mail*
IGIC *Information superhighway* ***Internet address*** *Home page* ***Backbone network*** / / *distributed network* / / *mid-level network* / / ***Network address*** / / ***Dedicated line*** *Electronic mail* *Standard Generalized Markup Language (SGML)* ***Mainframe computer***

DPCI

Text file

Access provider

Internet address / / *IAB* / /
Internet Address Naming Authority (IANA) / /
Internet Engineering Task Force (IETF) / / *Internet Explorer* / /
Internet message (***access protocol***) / / ***internet protocol*** / /
Internet Relay Chat / / ***internet service provider*** / /

Page break / / ***page design*** / / ***page printer*** / /

Network address / / ***Network News Transfer Protocol*** / /

Electronic mail

Public domain

Standard Generalized Markup Language (SGML)

Sample # 62: Internet sub-corpus co-occurrences listed by the dictionaries.

Several instances are specialized noun compounds, as can be perceived. Their occurrences in the sources often work within sentence patterns, as the text fragments in figures 25 and 26 below reveal. The texts dealing with the world of Internet technology in our corpus often tend to introduce concepts by means of such a cohesive device: a structural theme / rheme move where noun compounds are explained. For example, the introduction to SGML coding that a textbook gives, accounts for this procedure (figure 25).

The encoding scheme defined by these Guidelines is formulated as an application of a system *known as the* ***Standard Generalized Markup Language (SGML)***. ***SGML*** *is an international standard for* the definition of device-independent, system-independent methods of representing texts in electronic form.

Figure 25: Example of using lexical cohesion on the definition of an Internet concept.

This is also the strategy in the case of the context of *network address* as offered by a technical report on internet infrastructure (figure 26).

> Every organization connecting to the *network must have a unique identifier*. *This identifier is known as the campus IP* ***network address***.

Figure 26: Example of using lexical cohesion on the description of Internet terms.

Detailed examples of these and other lexical and rhetorical uses across texts are, in fact, reflected on and accounted for in the next chapter, which thus discusses the development of our results, core linguistic items based on type, genre and subject distinctions. In this respect, academic lexical items may be distinguished from subject words, or procedural utilization specified according to contextual dispositions. Such a treatment of words may enable the design of activities taking our corpus data as main elements for fulfilling language requirements and needs in our ESP setting.

CHAPTER 4: DISCUSSION

CHAPTER 4: DISCUSSION

The aim is two-fold throughout this section. The main concern with defining the lexical bulk of common coreness in our corpus is, on the one part, displayed by describing the range of lexical items as employed both recurrently and in isolation. On the other hand, their exploitability in learning English for Specific Purposes is proposed by means of suggested approaches for the acquisition of lexis in context.

In this respect, we distinguish between a first section devoted to the objective inspection of the results, and a second part where the method is also the discussion of the data, but in this case, with a focus on key issues springing from teaching implications.

I. COMMON CORE LEXICAL ITEMS

Utilized as reference, the data gathered not only throughout chapter 3 but also in Appendices 2 – 7 serve as target elements for discussing focal points in our lexical analysis. These are noun compounds, multi-word units, phraseological devices such as clusters, free and restricted collocations, cohesive markers and sentence patterns. They are all registered, for our purposes, as common core lexical items. They are selected as either peculiar within or formal across subject-matter settings. In this regard, their degree of common coreness may vary depending on whether they are more or less widespread throughout the corpus researched.

I.1. TEXT-TYPE ITEMS

The text type-based vocabulary, including argumentative, procedural and grammar / discourse lexis, can be examined in the first sub-section of the previous chapter and in

appendices 2 and 3. This category is chiefly recognized within narrowly specified domains corresponding to discourse functions and textual divisions such as the ones made: classifications, conclusions, definitions, descriptions, discussions, and exemplifications. Nonetheless, the use of some items can be extended across all the genres encompassing different text types; such are the cases of **by**, **between**, **thus**, **your**, **however**, **then**, **such**, **during**, **rather**, **include**, **indicate**, **example**, **more**, **add**, **procedure**, **use**, **document**, and **content**, among others, as chapter 3 and appendix 2 may reveal.

Some of the above items are actually listed as relevant words in the J category -- `Learned and Scientific English´-- of the renown LOB Corpus (Johansson, 1978). Function words such as by**,** between**,** rather**,** then**,** your **and** thus **are, indeed, recorded by the authority (Johansson, 1985: 34). In fact, in the case of the preposition** by**, the one-million words division J of the LOB Corpus registers it as the first grammar word in terms of frequency, exactly as befalls in our text type samples. Procedural and argumentative signalling is also key and distinctive in scientific-technical discourse: verbs like** indicate**,** add**, and** include **may prove this (appendix 2). Yet, as Johansson states (1985: 31), there tends to be a high `degree of "nouniness"´ in texts such as these in comparison with, for example, fictional writing.[267] As a result, in addition to the verbs offered, we may look at recurrent procedural and argumentative nouns such as** definition**,** description**,** procedure**,** comparison**,** addition**,** document**, and** content**, also persisting throughout appendix 2.[268]**

The approach thus undertaken could be conceived as all-inclusive, seeking to determine those items that are characteristic across our text types in the scientific-technical

[267] Hudson (1994: 333) claims that the exact figures of noun phrases in scientific-technical English is 35 % in the case of the Brown corpus and 33.3 % in the LOB (J categories).

[268] The informative tenor of scientific discourse is substantially conveyed by this nominal frequency (Johansson, 1985: 30).

discourse of Information-related disciplines. This is, in fact, a main means of issuing from one lexical group to the next in the discussion of results. However, an additional step can also become pedagogically enriching in the case of text type lexical items: that of discerning those words that co-occur in a predetermined textual set in opposition to their much less frequent or zero occurrences in other sub-corpora. Such is the case of the vocabulary restricted to one genre appearing in appendix 2.1, 2.2 and 2.3. Some examples of this distinctive lexis are constituted by **which**, **comments, findings**, and **solution** in research articles; **explicitly**, **chapter**, **channel** and **general** in textbooks; and **we**, **naming**, **authority**, and **via** in technical reports. Variation in their use across genres does not indicate that they are exclusive of only one writing type, but serves, quite differently, to establish a reference framework for devising academic tasks that focus on working with a predetermined genre or type (e.g. defining in reports, etc.).

Appendix 3, in addition, includes the characterization of text type items by means of significant combinations in given stretches of discourse. Thus, the orientation within only the one genre where that given node co-occurs most frequently is given to set or activate the mentioned scheme of genre and type language work. For instance, in terms of seeking core constructions, the colligation of **about** in research articles brings forth a common noun like **information** --*information about*--, appearing in descriptions, discussions and exemplifications. The same could be said regarding the cluster *on the other hand* in reports, where the procedural word **other** belongs to such a connector signalling contrast in the classification, definition and exemplification samples.

In contrast, equally valid is the approach that aims to pinpoint those forms acting meaningfully in one text type and not in the rest; for example, the argumentative noun **document** in the discussions of research articles. Its collocation with the preposition **within**,

in fact, seems to express that the prepositional phrase *within the document* tends to be harbored in such a text type environment of discussions -- i.e., perhaps due to the fact that in the discussions sections, the writer is likely to refer to a previously mentioned source with a higher frequency. A parallel point may be offered by the grammatical item **how** in textbooks. Here, the exclusive use of this particle in direct questions is observed only in conclusions – e.g. *how does the text...* or *how do people...*, whereas all other instances in different types reflect employment in indirect statements such as *documented how* or *decide how*. A notion to consider for teaching thereof may reside in questioning with **how** as a manner of concluding a paper or summary. The textual setting of textbook chapters serves to disclose such an assumption.

A more reduced context, as stated in the previous chapter, is detailed in the management of text type keywords, i.e. those items which most narrowly characterize the established samples of classifications, conclusions, definitions, descriptions, discussions and exemplifications (section C in `Text types´ -- chapter 3). The examination of these lexical items in their narrow domains can also offer some productive insights, e.g. the cluster *order of magnitude reasoning* in the conclusion of a computer science article, where the argumentative node **reasoning** embodies or sums up the main thread or topic of the paper.

I.2. GENRES: ACADEMIC WORDS

In the assessment of the results as reliable common core items, the second group – academic words--, posits a series of considerations having mainly to do with word classification peculiarities.

The first example is conformed by verbal patterns. Similar to the data provided in the J category of scientific-technical English (Johansson, 1985: 27), our verbal items are

manifested in all forms and tenses. Of these, more numerous are the cases of past and gerunds, as also seems to be the circumstance conformed by Johansson's set of distinctive verbs in the J sub-corpus.

Some examples shared with the LOB corpus in this respect are the following (taken from our appendix 4): **Defined**, occupying the 167th position in our DCL (detailed consistency list), **determine** (# 350), **assigned** (# 873), **corresponding** (# 658), **obtained** (# 948), **occurs** (# 1342), **calculated** (# 1581), **measuring** (# 1731), **executed** (# 1991), **assuming** (# 2019), and **emphasized** (# 2505). These are key or distinctive lexis in comparison with other registers of discourse, e.g. fictional writing (Johansson, 1985: 28). In our analysis, appendix 4 lists these and other verbal occurrences as common core in the sense that they are repeated across a range of at least ten texts. Those words appearing below that level or those absent in any of the three genre sub-corpora are removed from such a list.[269]

Among the adverbial elements, those considered in our academic lexis account are items which meet two main requirements: That they do not occur as discourse connectors or markers – since these are registered as text type vocabulary (in appendix 2)--, and for them to offer significant collocations which are distinctive and typical in a substantial number of academic texts.

The first specification is stated due to the fact that such adverbs as **thus** and **therefore** – characteristic of scientific English (Johansson, 1985: 29)—are recorded as sole grammar / discourse items in appendix 2.1, 2.2 and 2.3. They are viewed, in this way, as discourse markers, `making explicit the underlying relation between two spans of text´ (Oates, Corpora List: 16 Jul 1999). The second prerequisite stated about significance of collocations, refers to the relevant position in academic English of phraseological items such as *generally accepted*,

[269] This discarding of items, however, does not mean that they are not considered for other purposes (e.g. dealing with them in academic tasks or syllabus design – see sections ahead in this chapter).

approximately the same, and *much more effectively* – three typical adverbs that co-occur in the J category of the LOB corpus; these are the constructions considered as core academic elements in the genre-based adverbial group, appearing significantly in our corpus (6 times each).

In this respect, the words cited – **generally**, **approximately**, **effectively** – and other adverbs of this kind, behave with content and function words in patterns which are common to various disciplines, as defined in chapter 2 (`Classification based on learning´). These are thus core words of high frequency and evenness of range and coverage. **Generally**, for instance, is identified in 22 of our sources (# 647 in appendix 4), **approximately** in 14 texts (# 899 in the list), and **effectively** in 21 texts (# 1201). They are then considered as chiefly academic lexis in the sense that their use is extended across our textual material, a quite important notion when concordancing for developing a core vocabulary in academic terms (cf. Thurstun and Candlin, 1998).

As a different word class, adjectives are items likewise identified in the large LOB sub-section of scientific English (Johansson, 1985: 28). Some of Johansson's examples also appear in our work with a prevailing range of use: some are **linear** (ranked number 1238), **conceptual** (1993), **numerical** (number 2701), and **negative** (1141) –see appendix 4.
Other academic adjectives tend to occur as low-frequency items in our texts, as is the case of **narrow** in the expression *narrow domain* (2 instances in a Telecommunications report) or **smooth** in *smooth texture* (one instance in an Audio-visual communication report), according to the results of academic entries in chapter 3. These `hapax legomena´ – i.e. items that, albeit registered in all genres, show a rather low frequency— become significant, given their collocations found as relevant in terms of semantic interest and accurate conceptualizing as well as descriptive charge. Academic items thus convey a relevant degree of practical

usefulness and applicability for learning purposes and academic task observation, as shall be explored below.

I.3. GENRES: SUBJECT WORDS.

As already stated, academic and subject vocabulary may merge as the same group of items – e.g. they are gathered in appendix 4 across genres. Yet, depending on their context – e.g. across or within demarcated subjects--, the lexical behavior of the lemmas may vary and specify certain denotations, thus recorded as either academic or subject-bound accordingly.

The other important set of content words is that of nouns – left out on purpose in point I.2 above, mainly because these are items quite often regarded as subject-based lexis. This consideration is chiefly due to the fact that, given the data obtained in chapter 3 and appendix 6 on subject-restricted vocabulary, most subject keywords seem to be nouns – in agreement with Scott's (1996) assertion that in fact `the vast majority of keywords are nouns´.

Contemplating, as a result, these purely scientific-technical words – (Johansson, 1985: 26)--, subject nouns that are rather common in science and technology, according to Johansson, are **equation** (ranked # 2395 in appendix 4), **diagram** (# 2201) and **constraint** (appearing in position # 6 in the A2 subject category – chapter 3). From this line of work, other nouns follow, such as those analyzed in appendix 6.

Likewise, the word list offered in appendix 7, encompassing items in two specialized dictionaries of Information technology, also yields a large number of significant nouns – out of 177 items, an overwhelming 134 are nouns. Some examples are **development**, **management**, and **implementation**. Others, such as **change**, **handle** and **hand**, can also be employed as verbs, yet more often behaving as subject-based nouns in our texts. For instance, **handle** as a noun typically co-occurs in the multi-word term *the handle server system*,

common in Computer and Telecommunications reports of the A2 subject sub-division.

In the next three sub-sections, a more in-depth focus is placed, in turn, from the perspective of drawing a rationale for differentiating among types of common core lexical items –e.g. academic, subject and text type words.

I.4. ACADEMIC vs. SUBJECT-RELATED LEXIS

As Johns and Dudley-Evans (1980) report, there may exist subject-bound lexis, broadly used semi-technical words, and even colloquial items in given specialized areas of ESP (e.g. English for Transportation and Plant Biology). Such variation among words should be analyzed, from our perspective, as a common foreground of lexical knowledge for students majoring in different disciplines. In this way, their use should favor the development of academic and subject-based tasks. In this sense, Kelly (1991: 136), for instance, claims that `lexical ignorance is the main obstacle to listening comprehension´, especially among advanced learners, since these engage in both top-down and bottom-up processing levels to decipher meaning at the syntagmatic plane. For the purpose of gaining competence to use subject-bound and general or specific academic words, these, indeed, must be learned in context, as `vocabulary learning (...) is a never-ending process´ (Kelly, 1991: 138).

The description of lexical items, in this respect, plays a crucial role in distinguishing relevant keywords and common patterns. As Flowerdew (1993 b, 1994) asserts, concordancing allows this type of query to be approached. In terms of delivering and comprehending subject-matter content, for example, the sorting of words works as a chief element of organization and course design (Flowerdew, 1994: 19).[270]

The main element of accounting for a more subject-oriented aspect of lexical items as

[270] The goal is thus to achieve necessary competence for micro-skill elaboration – e.g. recognize discourse markers, key lexical items of topic, deduce meaning in context, etc.

opposed to a general academic register, therefore, seems to be the manner of dealing with such words in the environment of the subject matter. An illustration of this may be offered by the lexical behavior of the node **internet** as academic in our corpus, i.e. across a wide range of academic texts, and in opposition, within a more restrained setting. The constructions are taken from appendix 5 and 6 respectively:

Clusters such as *connecting to the internet* and *have access to the internet*, are found as academic items in one and three sources respectively (appendix 5). The former expression, *connecting to the internet*, is considered restricted as a result, whereas the latter entails a freer combination treatment. The noun **internet** as a subject-based word, in turn, also gathers these expressions in the F2 main topic category – about Internet and Unix--, but is not recorded as such in the F6 heading – dealing with Information infrastructure (appendix 6). Instead, a variation is produced, since *connected to the internet* and *access to the internet* appear without being preceded by **have**, as included in the F6 entry.

The F2 subject heading, in addition, supplies other forms not found in the more comprehensive plane of academic lexical entries; these are *usenet and internet*, *internet installation phase*, *internet access needed*, *internet request*, and *participating in the internet*, which seem to underline or evoke the subject developed in these texts: the description of Internet and computing facilities on-line. The F6 topic of Information infrastructures also offers some constructions not shown by any of the other three entries of **internet**. Some examples include: *can be used over the internet*, *utilizing the internet*, *available via the internet*, and *internet tools*. The focus thus seems to be placed, in this case, on the manner of using the information available through the internet, and the elements or components that make this utilization possible, as the central or common theme of the texts in the F6 topic category suggests – i.e. issues on information infrastructure.

Other findings suggest a correspondence between academic items that are not used very widely, and those identified in the two subject or topic categories. For example, *using the internet* and *via the internet* are located across a reduced number of sources which include the texts in the F2 and F6 topics. In addition, more broadly employed collocates such as those listed under the sub-section of free combinations in academic entries, can be pinpointed in these subjects too; such are namely the colligations *to the internet*, *on the internet*, and *of the internet*, or multi-word terms like *internet service* and *internet service providers*.

This contrasted data thus suggests two main readings when dealing with issues regarding academic and subject lexical behavior: first, a specification in terms of the type of text topics encompassed when dealing with the general subject –e.g. Internet in this case--, through which we may come to regard, in some instances of use, restricted academic lexical items as parallel to subject-based examples. And, second, the possibility of movement or transposition of lexical elements from given topics to other subject categories –e.g. from the F category (All four disciplines) to the A subject heading (Computer Science and Telecommunications). As a result, corpus-driven lexical information such as this can be either viewed as subject, academic or both. For the comprehension of academic discourse – in written or oral form--, awareness of such feedback proves quite helpful in the sense that it prepares learners to use items in correct contexts or settings.

This entails, for our analysis, that academic lexical behavior be observed in contrast with subject-based designation in order to anticipate the use of certain lexical constructions along points of academic discourse – e.g. in academic lectures.[271] The macrostructure of subject-specific discourse, in this sense, is characterized by `special features´ that must be realized to make students aware of patterns of language in specialized environments (Dudley-

[271] In the case of `point-driven understanding´ of lectures (Olsen and Huckin, 1990), in addition, a `disciplinary "cultural" conditioning tied to methods´ is of prime importance (Oslen and Huckin, 1990: 44).

Evans, 1994: 147).[272] In this respect, this is a more narrowly focused approach on specific content determined by subject matter. Most ESP learners may thus `rely on background knowledge of their subject´ to understand the language (Mason, 1994: 199).

The work with key keywords,[273] in this respect, also presents a predominant condition in the exploitation of subject-related lexis. The noun being checked – **internet**--, for instance, occupies the ninth position in the list of key keywords in sample 32 (chapter 3). The F category (All four disciplines) where it is identified serves as the context of use, and, needless to say, it yields all the items examined in the F2 (Unix / Internet) and F6 (Information Infrastructure) topic headings. However, other elements such as *internet environment* or *internet mail*, appearing in fields not belonging to the F category, such as B2 (On-line database systems), are absent in the F division. Such subject-driven data can, in fact, operate as suitable linguistic content in areas of specific purposes language teaching, such as ESAP (English for Specific Academic Purposes), given their more narrowly focused scope in subjects, topics and sub-topics within Information Science and Technology.

Parallelly, more widely encompassing common core academic items, viewed in different subject contexts, are highly valuable for the development of academic signposting. Structures such as *to have access to the internet, these internet documents*, and *information on the internet*, among many others, serve multidisciplinary or English for General Academic Purposes (EGAP) approaches, where, in fact, a broader range of lexical items –verbs, nouns, adverbs, etc-- is favored (cf. Davies, 1988; Flowerdew, 1993 a; Jordan, 1997). From our view, as shall be seen in the second section of this chapter, both types of lexical study – academic and subject-based—are applicable and necessary in the exploitation of language

[272] Comprehension ability may be transferred to the production of written or oral subject matter discourse, since, depending on the developed schemata by the learners, `subjects who speak or write more coherently and with greater command of the language will appear to have greater comprehension and recall´(Rost, 1994: 97).

[273] See `Keywords´section of `Subjects / Topics´in chapter 3: samples 28-32.

tasks.

As an exemplification, the collocations *code elimination optimization*, and *optimization techniques* are described as key constructions in the body of texts shared by Telecommunications and Computer Science subjects (the A group), and, yet, across a larger bulk of sources, the item *code elimination optimization* also appears, denoting that its use is applicable in other contexts. Yet, the verbal phrase *make this optimization* is not specified in the **optimization** subject entry for the A set texts, which suggests that such a cluster may adequately be labelled as academic, serving broader purposes across common core lexical matters. Among many others, some more instances of this kind may be supplied by contrasting clusters that either appear exclusively in subject entries or occur in academic entries; in the latter, lexical elements, categorized as either restricted or free, constitute items pinpointed in more than one subject domain. These are compounds such as *HTML tag* (more subject-restricted) and *HTML page* (more generally used), *knowledge management terminology* (bound to specific subject) vs. *application knowledge* (less bound), and *Aggroup tools set* (thematically restricted) vs. *tools for enhancing* (freer in its use) (see appendices 5 and 6).

I.5. ARGUMENTATIVE vs. PROCEDURAL VOCABULARY

A further lexical need in General Academic Discourse (GAD) – i.e. across various disciplines—is the acquisition of `procedural knowledge´(Lockett, 1999: 55). The exposure to pivotal means of directing scientific-technical discourse seems to be the all-important strategy of proceeding accordingly (Lockett, 1999; and other authors),[274] in order to bring about

[274] Halliday (1979) and Widdowson (1979) appeal to `the actual use of language in concrete situations´(Halliday, 1979: 4) where the `principal aim is to promote a knowledge of the language system, to develop the learner's competence´ (Widdowson, 1979: 118). For such goals, the exploitation of procedural words in context makes a great deal of sense, as in the `use of diagrams to write definitions or generalizations´(Allen and Widdowson, 1979: 131) for guided paragraph writing (ibid., 139).

performance from competence. By working accurately with this kind of information, lexis is employed to conform to academic aims – e.g. defining, classifying, etc.[275]

In such a process, knowledge of the topic is not enough, as Williams (1984: 7) asserts; we must also `include familiarity with different text types´. The cohesive texture of the message being delivered thus necessarily contains procedural and argumentative lexical ingredients – cf. Widdowson (1978: 22). Identifying these two types of lexical items thus bears a significant status for the overall development of text interpretation and understanding, which is surmised as a fundamental chore. These text type words are examined as those common core lexical elements to which students of different specialisms would be able to contribute and which would provide practice in a set of skills, structures, functions and semi-technical vocabulary which the students will meet in their specialist studies (Kennedy and Bolitho, 1984: 50).[276]

The employment of argumentative and procedural items as linking devices throughout discourse is a main function, yet, their distinction is mainly based on the type of signposting which they infer. Thus, given the functions of reference, substitution, conjunction, and patterning lexis (Pincas, 1982: 55), argumentative features may be rendered apart from procedural usage. In fact, the former –argumentative-- tends to fulfill the first two linking implementations – reference and substitution--, according to our data, whereas procedural words tend to emphasize conjunction and patterning modes of discourse instrumentation.

The lemma **example**, a procedural noun, and **document**, an argumentative one, may

[275] Bygate (1987: 74) introduces the set of organizing activitites to be conducted with procedural knowledge (e.g. compare, categorize, detect differences, etc). The British Council (1985: x) also notes the demand to practice language in relation to the discourse and concepts encompassed.

[276] In contrast, subject-related words would be exploited in specialist studies or sub-groups that share the same subject-matter – they would thus be considered core sub-technical words that are subject-related (Kennedy and Bolitho, 1984: 56). The concept of `instrumental motivation´is also key for these two authors (Kennedy and Bolitho, 1984: 17). Through this, procedural lexis is attained by attempting `alternative ways of expressing the

illustrate this variation in use according to such functions –see appendix 3. The first item combines with the preposition **for** at strategic points of discourse conjunction: *for example* thus starts a new sentence in the classifications and definitions of research articles – e.g. *For example, the top level HTML page is linked...* In addition, *for example* is placed in the middle of clauses in the discussions and exemplifications of the mentioned genre of research articles – e.g. *a variety of cognitive and organizational issues, for example concerning the communication and coordination of knowledge.* These are thus instances of the function of conjunction in procedural words.

Secondly, the patterning factor is also located in procedural usage. In the same entry of the noun **example**, this effect can be demonstrated by noting the weight of the repetition. **Example** is, indeed, quite frequently followed by a colon and an explanatory statement in the discussions of research articles – e.g. *Example: Text describing the...* A similar case occurs with the collocation *example* # in exemplifications, where such a way of repeating is encountered up to 10 times – e.g. *in example 1 and 2, the...*

In terms of argumentative potency, the noun **document** records a value denoting reference by means of clusters such as *information describing the HTML document* or *structure and notational conventions of this document*. These are constructions taken from discussions and exemplifications in research articles. Their main operational concern is that of signalling or pointing to a given source: They therefore establish a frame of reference.

In turn, lexical substitution may be clearly reflected in conclusion samples where the expression *the publication is intended as a companion document* displays the substituting role held by the noun **document** (i.e. replacing **publication**, **source**, **work**, etc). The phrase *the resulting document is created*, in this respect, also evidences such a mode of lexical

---

same concept´(ibid., 36). Procedural thus become functional vocabulary to some extent (ibid., 20), e.g. transforming sentences by focusing on past participles such as **produced**, **formed**, etc (ibid., 40).

replacement -- in this case, **document** serves as the embodiment of the outcome of the research presented in the paper.

Therefore, by basing our lexical division on both empirical data and scholarship such as Pincas (1982) or Francis (1986), we may be able to discern text type collocational reference and value for effective purposes in language teaching.

I.6. DISCOURSE vs. GRAMMAR FEATURES

In contrast to the argumentative and procedural words above, discourse / grammar items are mainly function elements, not nouns, verbs, or adjectives. However, adverbs may be treated as discourse words when they provide reference to rhetorical functions – e.g. **however** signalling contrast. Pronouns, in addition, may work as discourse devices if their use is intended to bring forth writer – reader interactivity – e.g. **you** addressing the audience.[277]

To distinguish, nevertheless, grammar from discourse properties in such a set of words, the prerrogative is placed on the contextual direction.

For one thing, grammar items are more numerous; they tend to fill the top slots in frequency lists such as those found in appendix 2.[278] Examples include **by**, **such** or **on**.[279] These are colligational items that usually occur with a high frequency in scientific-technical corpora – e.g. in our corpus, as in James' HKUST (1994) or Lozano Palacios (1999), the most repeated grammar word is the article **the** (55,147 instances in our IST texts – Information Science and Technology), followed by the preposition **of**, with 30,828

[277] Some of these key aspects have been utilized as linguistic yardsticks for reflecting on corpus representativeness (chapter 2).

[278] See `Text Types´ in chapter 3.

[279] As may be observed, the presence of recurrent prepositions such as **by**, **on**, **in**, etc are, in part, omitted in the word lists of appendix 2, since these are grammar words that typically occur with a high frequency in all text types and genres. The aim is to conduct a search based on main differences and patterns for teaching purposes,

occurrences.[280]

Despite their high ranking, not all grammar words are worth considering as relevant lexical items for our common core purposes. In this sense, we agree with Edwards (1985: v) when she states that the intermediate level learner of English may `know main structures but need to check points of grammar´. Some grammatical aspects thus need to be focused upon while others are more obvious. For instance, prepositions analyzed as colligations accompanying verbal forms seem to present a wide-ranging employment in our texts. These are found to be significant in terms of structuring the text, in which forms such as *by copying*, *based on*, or *used to support* supply new information as rhemes in subordinate clauses. Some other examples can be checked in figure 9 (chapter 3), where **by** is also common in passive clauses – e.g. *recommended by the HTML specifications*.

In addition, prepositions that co-occur with certain nouns in academic and subject words are also deemed as useful grammar constructions. In appendix 3, the lemmas **about** and **between** can illustrate this point by means of recurrent forms such as *information about* and *relationships between*.

Subject-verb agreement can also be an object of inquiry and observation. This may be checked by observing the use of the singular and the plural in ambiguous nouns such as **data**. Grammatical feedback is thus derived from contrasting, for instance, the phrase *data is copied* with *data are discussed*. The former occurs six times in our text type corpus, and the latter twice. There seems to be a tendency to refer to this item in the singular form, since it appears as more broadly employed and common core.[281] This type of grammatical description, as

---

thus, the use of these colligations, tending to be rather uniform, does not supply significant linguistic contrast for our text type-related vocabulary study of lexical distinction.

[280] The HKUST Corpus (James, 1994) contains 78,574 instances of **the** and 33,875 of the preposition **of**; in the VEBD (Lozano Palacios, 1999), the count is 255,095 in the case of the article and 173,214 occurrences for **of**.

Badger (1999: 85) suggests, may become quite valid for detecting specialized discourse traits from a grammar scope.

At the junction of discourse analysis, in addition, features such as conjunctions and prepositions may, in fact, work as connectors in clauses. This leads to their treatment as discourse items. Examples are **as** in *as well as* or *such as*, and **because** in *because of the*. They function as semantic markers, pointing to `the writer's main points, what he is driving at´(Wallace, 1991: 53). Like **however**, or **nevertheless**, these words fulfill a specific role in bringing to the learner an awareness of the importance of certain facets in written or oral texts. This aspect determines, to a great extent, the student's ability to make a note of main points in a lecture, for instance (Wallace, 1991: 57).

Discourse features are thus viewed as markers of `the underlying relation between two spans of text´ (Oates, Corpora List: 16 Jul 1999). Their application is chiefly regarded in stylistic terms from this viewpoint (Jackson and Bilton, 1994). This means that they significantly aid descriptive approaches to academic discourse, since these are `words, phrases or sentences which highlight particular parts of the discourse as important or unimportant, or as being examples, glosses, corrections, reformulations, reminders and repetitions´ (Wijasuriya, 1971: quoted by Jackson and Bilton, 1994: 62). In agreement with Flowerdew (1992) and Burnett (1988), consequently, the focus on such a type of signposting offers `useful input to language teaching´ (Jackson and Bilton, 1994: 63). For oral delivery tasks, for example, `the function of this type is to make sure that the audience understands terms which are introduced as the lecture progresses´(Flowerdew, 1992: 209). This may lead to `a more appropriate selection of classroom activities´(Jackson and Bilton, 1994: 77), due to

[281] Our overall corpus produces a similar proportion: the singular form **is** is used 63 times with **data** whereas the plural **are** only 24. Johns' EAP web page of `kibbitzers´reviews this employment of **data** on 16th June 1996 and on 3rd March 1997; his conclusions are the following: In educated everyday usage as represented by the Guardian newspaper, it is nowadays most often used as a singular (...) the (more recent) meaning "digital information stored or manipulated by a computer" is most often singular´.

the fact that authentic listening and note-taking activities make up a quite pertinent object of study in EAP approaches, as shall be examined in the next section.

The example of **however**, appearing in appendix 3, serves as a demonstration of the emphasis placed on discourse items at certain points on the texts. In a definition of genres in an Audio-visual communication article, for instance, the node condenses the contrasting effect of a specification with a generalization. The phrase *The genre does not, however*, *fully constrain* denotes a shift in the definition, by which a pause and the semantic marker suggest furthering in-depth details regarding the concept defined. Likewise, in a discussion on computer design, after two tasks are explained as sharing common attributes, the placement of **however** follows, indicating that the author of the article goes on to explain that there are important differences between them, employing *however, whereas in…* to delimit distinctions. In a similar manner, this writer registers a specification of what he means by a term previously used; the phrase *I emphasize "inherent", however, because*… is placed after such a concept -- `inherent design´. Yet, because a clarification, segmenting a bit the flow of the discussion, seems to be felt as needed, signalling such an insection is made explicit by means of the discourse marker.

In the second part of this chapter, nonetheless, illustrations of lexis functioning in context to establish a marking function through the flow of discourse are given by examples other than these adverbs. For instance, using conjunctions and prepositions in patterns of repetition and clarification in oral devlivery, such as the phrase *according to this figure,* is expounded within the framework of teaching text type vocabulary in academic common core settings of science and technology.

II. SEARCHING FOR WORDS IN ACADEMIC TASKS

As Hutchinson and Waters (1987: 37) claim, `describing a language for the purposes of linguistic analysis does not necessarily carry any implications for language learning´. However, because of possible comprehension and production problems that academic, subject, argumentative, procedural, grammar and discourse items entail in our ESP courses, teaching them should be justified in needs analysis terms – e.g. target situation and learning demands, wants and lacks. Learners should thus be exposed in order to be made aware of certain relevant lexical items that posit a conundrum of structures and signalling devices, embodied by miscomprehension and misconception in contexts where content and language interact in specific ways.[282]

From the viewpoint of our research, an effective manner of dealing with such issues is to base the explanation and classification of our corpus-based lexical data on the design of tasks and activities. These should be seen as the `ways of producing appropriate situations for language use´ (Brumfit, 1978: 39). The exploitation of our lexis is thus framed within different detailed academic micro-skills of all four language skills – reading, writing, listening and speaking. This is the learning scenario where, according to Nunan (1989: 50), tasks can be carried out with a communicative purpose: where `techniques of inference, linking, skimming, anticipating´ (Nunan, 1989: 73) may be applied; thus, the goal of communicative development is closely followed in `information gap´ centered work (problem analysis, vocabulary exploitation, discussion, note-taking, etc – ibid., 122). Academic purposes thus parallel technical studies (EAP envelops EST, according to various authors – e.g. Jordan, 1997).

[282] In this sense, as Brumfit (1978: 38) states, language must be described also in terms of the types of messages to be conveyed, related to the situations in which it is anticipated that the language will be used´.

It is, in fact, in such a range of cited techniques of linguistic development where the applicability of key lexical devices stands a great deal to gain. An example is the production of phrasing on writing coherent paragraphs, where `learners have to be exposed to samples of the required type of writing if they are themselves going to produce coherent texts´(Kennedy and Bolitho, 1984: 85). In listening for keywords in lectures, in addition, the aim is likewise to draw from language-content schemata to distinguish `formal from informal registers´(ibid, 110). As Flowerdew thus claims (1993 b), the focus, being placed on language or `product-based´ perspectives, equally favors a `task-based´ approach (Nunan, 1989), since `the language presented in the lead-up to the tasks, experience demonstrated, was useful in the performance of the tasks´ (Flowerdew, 1993 b: 242). In addition, as Jordan (1997: 64) asserts, the communicative level of the language learning process (i.e. the integration of function, situation, topic, content, skills, method) leads naturally to academic tasks based on lexis development.

An example may be induced by the recurrent appearance of the noun **contractor** in technical reports dealing with documentation policies (subject category C3 – appendix 6). The item can be checked within clauses such as *the contractor shall assign*, *the contractor shall be liable for*, or *the contractor shall be responsible for*, which clearly carry a tenor related to the subject field of the law. The register is thus forestalled by such a configuration of meanings, `typically associated with a particular situational configuration of field, mode, and tenor´ (Halliday and Hasan, 1985: 38).[283]

In these cases, the emphasis is placed on `the two *process* goals of accuracy and fluency´ (Brumfit, 1978: 39). In the process of ESP practice, indeed, one of our stages is to determine the types of academic exercises for the manipulation of common core lexical items

[283] See chapter 1 (section on `Register´).

in Information-related studies. This step naturally follows quantity analysis (i.e. our lexical analysis), and that of class or course factors (i.e. a target situation and learning needs analysis) -- Kennedy and Bolitho (1984: 137). What are thus being considered are language features and skills / micro-skills demanded in the learning process. In the discussion of our lexical data applicability, these two aspects – language and skills for task development-- function as pivoting principles. In this respect, two sub-divisions are given in the current section: A first one dealing with the types of lexical collocates and clusters and how they should be approached in the task-based approach, and a second part where more explicit examples of activities are suggested for our ESP courses.

II.1. TYPES OF LEXICAL PATTERNS AND ACADEMIC TASKS INVOLVED.

Word behavior awareness and lexical patterning acquisition are two main goals in the process of resolving means for teaching our common core items. In this sense, our central claim lies in the fact that to achieve outstanding performance in a given register of English as a foreign language, both anticipating linguistic constructs and being able to transmit content according to academic standards play crucial roles. The mechanism or engine for the activation of this lexical work may be illustrated by the recording of significant items or samples which capture the essence of linguistic performance when conveying a particular concept.

From this perspective, our lexical selection principles may reflect those explained by Palmer (1922) and his `ergonic system´ school of thought. According to Howatt (1985: 238), Palmer would lay the foundations for the `Behaviorist or Direct Method´ of language learning when he asserted that `limited sets of words´ amount to primary matter samples for infinite combinations of sentence / pattern units (Howatt, 1985: 238). Secondary matter in language

acquisition, in this respect, evolves as a result of the adequate exploitation of such primary elements. Noting down relevant or core linguistic devices such as verbal or noun patterns would thus serve to recall the patterning of a given word in specified contexts of usage.

However, mainly due to the qualitative difference established by the Communicative school of language learning in the 1970s, our proposal for the development of tasks does not mirror Palmer's and Bloomfield's ideas transferred to the Audio-lingual / structuralist methods. In fact, for the productive exploitation of patterns, as Mackey (1978: 84) claims, `structural linguistics did not take context into account and often ignored the existence of the meaning of words´. As a result, the fundamental function of context of use, being left out, was not analyzed as a strong factor of influence for the choice of items, whereas this premise is appraised as axiomatic in our research.[284]

From this scope, as Gutschow (1978: 55) states, The teacher has to know about structure, of course, but it is the distinction between *structure* and *pattern* which helps him to decide how much of his linguistic knowledge he should bring to bear upon his teaching.

The focus is therefore lain upon the `task to provide [learners] with opportunities to acquire `chunks of language´ (and discover their conditions of use) that can easily be reassembled and rearranged according to context´ (Gutschow, 1978: 56): pattern study in the form of the `collocational range of the material, the situational context, and other relevant factors´ thus become essential (ibid., 59). [285] The aim is to fit a contextual plane accordingly, neither falling in the error of `under-contextualization´, nor in the opposite circumstance --

[284] As identified throughout the theoretical background (chapter 1), Firth (1951, 1957) is, in contrast, seen as our first main reference for the communicative approach to lexis, since he established `meaning by collocation´ as one of his modes of meaning (Mackin, 1978: 151).

[285] That `this context in turn depends on the society or the speech community in which our speech takes place´is also pertinent (Mackey, 1978: 97). The questions `what´, `how´and `why´the communication is established are key conditions (ibid., 90).

`over-contextualizing´ (Mackey, 1978: 99).[286]

Co-occurrences, in this respect, can be divided into three main headings or classes of collocations for teaching purposes: verbal, nominal and adjectival.

II.1.A. Verbal clusters.

Given a highly frequent lemma such as **design**,[287] its past participle form designed triggers specific verbal actions in our corpus, based on both frequency and range of employment. Co-occurring with such a word across a broad number of texts, in fact, are other verbs like **provide** and **facilitate**.[288] In addition, within a smaller sample of sources, the clusters *designed to deal with* and *designed to determine* also bear a noteworthy character. These free and restricted constructions should thus be traced as important in the utilization of the meanings intended: for expressing purpose, in fact, a range of similar clusters may be practiced, according to these, which are the most prevailing; *a handbook designed to provide accurate information*, *programs designed to provide skills and training*, *a system designed to facilitate learning*, etc.

Effective patterns thus bear contextual clues on the learning environment; in this respect, Hornby (1974: 206) observes that `verb patterns are the most important element in sentence patterns´. Strevens (1978: 114) points out that items like these constitute `practical techniques´ in language learning for achieving specific goals. Hill (1991: 25) distinguishes the importance of verbal pattern use as an identifier of English as a Second Language (ESL)

[286] With the first term –under-contextualization--, a reference is made to `the application of structural linguistics and generative grammar´, which did not make context explicit; by `over-contextualization´, in contrast, Mackey means the process of producing `utterances which were idiomatic but inappropriate when used outside the context in which they were learned´(Mackey, 1978: 99).

[287] See entry in appendix 5.

[288] See, also, chapter 2 (`Corpus design assessment´), where these forms are examined as common in the communication of purpose.

speakers vs. natives.

Cowie (1978: 129), in turn, approaches the process of written composition by following Hornby's idea that `the foreign learner needs to compose, not to pull to pieces´. Verbal patterns thus serve to flesh out a set of abstract constructions (...) indicating the syntactic distribution of words in their various senses, throwing light on the meaning of words (...) and encouraging the learner to compose sentences which are lexically, as well as syntactically new (Cowie, 1978: 129).

We may thus elaborate further and add a corpus-focused perspective to the study of verb patterns by following Cowie (1978, 1998) as well as other recent work by Hunston and Francis (1998, 2000). This entails forming groups of verb clusters according to their observed behavior in our texts. In this respect, as Sinclair points out (1991: 53), sense and pattern tend to be associated with each other. For learning purposes, thus, rather than filling in words when given a predetermined grammatical chart (as the traditional approach had it), the point is to `to use the patterns as analysis´ (Hunston and Francis, 1998: 62), and to foster `awareness of the importance of collocation to language learning, and an awareness of multi-word lexical chunks (...) without the need for internal processing´ (ibid., 63). The learner should then be able to `associate meaning with pattern and so remember which verbs have which patterns´ (ibid., 55). The concern therefore lies in using `the pattern as a basis for vocabulary building´ (ibid., 70).[289]

In our corpus-based categories of lexical items, academic entries – appearing in appendix 5—may contribute to illustrate the use of the discussed verbal patterns. Cowie (1978: 132) lists, among others, examples such as *check a bill*, *seek the objective*, *consider*

[289] Francis, Hunston and Manning (1998) also propose three levels of activities with word patterns for grammar learning: verbs, nouns and adjectives, offered in the Collins CoBuild series of English teaching.

my suggestion, *the theory is canvassed*, *argue rather less vehemently*, etc.[290] Our verbal sets are also found quite frequently and more or less evenly across academic texts; some examples noted in chapter 3 and appendix 5, as well as others not registered in our work, can be assessed. As Hindmarsh (1978: 141) observes, these verbal forms should be stored by the learner for assimilation aims, while she or he encounters them in reading procedures.

The following are some verbal patterns distributed according to their bearing or behavior in our corpus. A first set of items may be made up by those verbs that prompt specific information-related objects; for instance, *preserve digital information*, *narrow the search*, *developing networked information*, *providing access to*, *broadcast messages*, etc. Similarly, a second group could be devised by detecting subjects which typically accompany verbs in information technology, such as *the interface provides routines to*, *message containing the*, etc.

A third type of verbal patterns seems to be the one framed by passive statements where the agents and subjects are related to the Information world: *messages are exchanged*, *the analysis cannot be applied*, *the lack of information has to be overcome*, *used by the service provider*, etc. Fourthly, there are various forms of verbal co-occurrence that could be conceived as less restricted to the Information area, and, in contrast, more general in academic registers. Examples are *provide support*, *make informed decisions*, *coined the term*, *make available to*, *provide the necessary background*, *overwhelmed with requests*, *receives a growing attention*, etc.

Finally, a fifth collection of verbal combinations may be represented by characteristic colligational occurrences – prepositions, conjunctions, adverbs or any other syntactic components. These can be either representative in Information technology or not; some are

[290] In a later paper, Cowie (1981) divides these specialized verb-noun collocates into three groups:`free phrases´, such as *run a business*, `figurative senses´ (e.g. *canvass someone's opinion*), and `idiomatic or specialized´, such

can be accessed through, *involved in*, *efficiently compute*, *understand how*, etc.

This sorting of verbal collocations is thus produced as a means of recording or writing down lexical feedback in an organized manner. As stated above, this type of practice becomes a suitable task as long as it is done in context. This entails a constant revision of new or different data as extracted from our range of textual material. In such a line of work, the *BBI Dictionary of English Word Combinations* in its third edition (Benson et al., 1997) may be taken as an exemplary source for documenting collocations, where `a special effort has been made to identify and incorporate collocations that have entered the English language in recent years since the interlinking of computers´ (Benson et al., 1997: vii). Thus, clusters such as *to browse the web*, *to go online*, *to access data*, *to send e-mail*, etc. are updated in this dictionary of collocations.

The *BBI Dictionary* (1997: xv-xxxiv) thus presents an in-depth arranged view of grammatical and lexical collocations – up to eight categories in the case of grammatical co-occurrences, and seven headings for lexical combinations.[291] Verbal patterns are found in divisions G8, L1, L2, L4 and L7 – the first one belonging to the range of grammatical collocates, whereas the L sets describe lexical groups.[292] In our corpus, all these types of verbal collocations can be identified to some extent, as the following data may reveal.

For example, as illustrative of the G8 section – grammatical collocations or colligations with verbs--, our body of texts yields forms similar to the ones itemized in the *BBI Dictionary*. Some are *the test consisted of* (18 occurrences), or *this consists in measuring*

---

as *foot the bill*.

[291] Such a type of organizational approach to lexical collocations, as Bahns (1993: 57) claims, is highly demanded in corpus linguistics.

[292] The G8 set is likewise divided into 19 sub-divisions (A-R) which depict syntactic patterns mainly based on the transitivity or intransitivity of verbs – e.g. heading A contains dative movement transformation samples such as *he sent his brother the book*, or group J lists direct object items following the verb in *–ing*, such as *he kept me*

(12 instances). Other examples are *we decided to* + verb (8 times), *they wanted to* (7 items), *served as a* (4 elements), *can be accessed through a* (7 forms), *run by the* (7 items), *messages based on* (9 occurrences), and, of course, phrasal verbs such as *made up of* (18), *keep up with* (17), *come up with* (14), *set up a* (13), *to write down* (18), *shut down* (12), *drank it off* (4), etc. In the case of the lexical collocation categories of verbal co-occurrences (L1, L2, L4 and L7), the instances located in our corpus are equally vast; therefore, only some examples of verbs appearing in the *BBI Dictionary* are given.

The L1 collocations consist of `a verb denoting *creation* and / or *activation* and a noun / pronoun´ (Benson et al., 1997: xxx). In this sense, some of our own occurrences are the following: *provide written instructions*, *make a distinction*, *composing good HTML*, *reaching agreement*, *launched initiatives*, *prompted researchers to find*, *produced inconsistent results*, *request was issued*, etc.[293]

The L2 list displays verbs `meaning essentially *eradication* and / or *nullification* and a noun´ (Benson et al., 1997: xxxii). Some of these combinations detected in our work are *lack of information has to be overcome*, *connection refused*, *destroy the connection*, *denied permission*, etc. L4, in contrast, provides clusters in which a noun precedes a verb: `the verb names an action characteristic of the person or thing designated by the noun´ (Benson et al., 1997: xxxiii). Some instances are *the interface provides*, *buffer is allocated*, *application to determine*, *HTML provides*, etc. Finally, under the L7 heading, collocations consist of a verb and an adverb; some of our items are *efficiently compute*, *thoroughly revised*, *perfectly matched*, *likewise argue*, *integrate technology effectively*, etc.

---

waiting two hours (Benson et al., 1997). In a previous work, Benson (1985: 61) calls these grammatical collocations `close constructions´.

[293] These verbal uses indicate the tendency in `satellite-framed languages´ such as English (Slobin, 1998) to contain much detail with regard to action in construction or motion verbs – e.g. **launch**, **prompt**,etc. Their translation into `verb-framed languages´ such as Spanish, indeed, points to a lack of these particularities (e.g. ***plantear iniciativas***, ***ejercer influencia***, etc).

Having learners record such structures in an organized way can thus become a practical and productive exercise. Under the instructor's supervision, students may achieve the aim of realizing to which groups or categories the given items belong. In this way, the offered samples above – both from our own verbal arrangement and that based on the *BBI Dictionary*-- should grant an appropriate level of competence as long as they are studied in the appropriate textual environment of the technical studies. As Hindmarsh (1978: 146) remarks,

Every item must have a role to play in communication, and thus be worth the student spending time and effort on, for receptive command first, and moving to a degree of productive command later.[294]

Verbal clusters thus seem to advance a befitting range of linguistic input for learners at different learning levels –intermediate and advanced-- in ESP courses, where academic skills play significant roles in accessing lexical knowledge.[295] The tools employed, as Hindmarsh (1978: 143) states, are those that `organize and use the items and features of the language we need most´. As a result -- this author also claims--, `we need to put a lot of energy into mastering the vocabulary of a foreign language, especially at the intermediate level, and to do so in a systematic way´ (ibid).

In fact, working effectively with patterns such as verbal items, may permit overcoming difficulties in the form of common language errors. Some of them, compiled via the British National Corpus and Longman Learners' Corpus in the *Dictionary of Common*

[294] As Cowie (1988: 682) asserts, in fact, in the work with grammar and lexical collocations, the learner will slowly move from decoding skills to encoding abilites; effective learning approaches, including a reliable source of collocation feedback, play important roles: `the introduction of a system of syntactic patterns, (...) and the use of a controlled vocabulary´ (Cowie, 1988: 688).

[295] In fact, Aghbar's and Tang's (1991: 15) study shows that verb-noun collocations such as `gain admission´ should be a constant focus of attention for testing linguistic performance; as Howarth (1998) shows, the more advanced the learner seems to be, the wider array of prefabricated language he or she may be able to handle and

Errors (Turton and Heaton, 1997), are blunders of the type of **give information of the system*, **provided computers CD-drives*, **make a test*, **discussed about the role*, **appeared on our file*, **appears in the screen*, **this paper talks about*, etc.[296] Directing students to identify the right way to reformulate such verbal phrases, in contrast, may serve as a constructive exercise when dealing with our corpus material in class. This chore suggests a cumbersome but rewarding process of registering and differentiating syntagmatic uses – e.g. *provide information about* instead of *information of*, *provided computers with CD-drives*, etc.[297]

II.1.B. Noun collocations.

The other great patterning environment, in addition to verbal clusters, analyzed by scholars is the one associated with nouns. In specialized English domains, such as technical fields, noun compounds are rather distinctive in this sense.[298] A plausible lexical activity, given the extensive amount of noun collocations occurring in our corpus, focuses on challenging learners to cope with correct word order when designating specific concepts by means of nominal compounds. This involves the identification of the core element being defined, described or explained by means of noun compounds and contextual clues.

---

manage. Remarkably, according to Kennedy (1990), the student's capacity to interpret and use collocations does not mean that he or she can intelligibly discern their significance as fixed expressions.

[296] The Cambridge publishing house, in this respect, has created a resource on the web that allows users to access common learners' errors based on collocation and grammar deviations (the address is http://www.cup.cam.ac.uk/elt/reference/clc.htm.

[297] This examination of collocation-based errors, as Lennon (1991: 34) states, is ever useful in English as a Foreign language (EFL) classes, in spite of many years of classroom instruction and ongoing exposure. Indeed, as Farghal and Obiedat (1995) comment, collocations should be stressed further in EFL courses, since they tend to be a neglected variable.

[298] As Bauer (1983: 202) states, `the vast majority of compounds in English are nouns´. See also Varantola (1984), Pedersen (1995), and others in `Lexical focus´ -- chapter 1.

Notions, for instance, related to the world of Internet, are often expressed by combining such nominal items – e.g. *internet service provider*. The aim is to deduce which item functions as the keyword in the cluster, being described by its surrounding textual environment. In this case, **provider**, identified as the central noun, carries the action of providing access to the internet. Other collocations may not be so simple to discern.

For instance, the concept of *base tag set* explained in a textbook chapter on hypertextual composition must be closely examined in context for a more accurate perception: **base** qualifies the nucleus *tag set*, referring to a tagging system; it is thus a modifier of the other two words. The noun **set**, in turn, may be viewed as a keyword used in compounds, according to the *Dictionary of Keywords* (Edwards, 1985). Its semantic position in this case tightly depends on the compound *tag set*. In this associating mode, the noun **tag** would almost act as an adjective with **set**, having a high probability in our corpus to co-occur together, given Mutual Information and T-scores infoming on the dependence of **tag** on its collocation with **set** (see chapter 1, `Concordance´ section).

In fact, noun collocations, as Pedersen (1995: 72) states, are often described in greater detail when there are more elements juxtaposed. For instance, `*heat resistance* addressed to the lemma resistance is less informative than *superior heat resistance*, which is less informative than *provide superior heat resistance to´*(Pedersen, 1995: 72).[299]

This presence of context becomes a tool for the learner to employ in order to explore and devise definitions and meaning in clauses: `Vocabulary mastery´ is, in fact, conditioned by working with vocabulary in context (Allen, 1983: 114).

Furthermore, as corpus-driven explorations into noun and pattern prove (Francis, 1986, 1993), nouns may be labelled within grammatical categories of use according to their

[299] See chapter 1 (`Collocation´).

behavior – e.g. *assumption that*, *assertion that*, *conviction that*, *hypothesis that*, etc (Francis, 1993). In this respect, given our data, multi-word units and clusters are common devices in the bulk of nouns within Information Science and Technology.

The following are some examples of nominal multi-word items as exhibited in appendix 6, distributed according to their appearance as either compounds or clusters with prepositions or conjunctions (included according to their relevance in subject categories A1 – F6).

Compounds:

access mode bit *token ring clients*

task completion time *software tools*

library automation service *network object type*

surrogate stream *remote method call*

knowledge management *program requirements*

mass media *contractor project manager*

microfilm frames *fiber-optic communication*

laser action *standards board* *codes modes*

Microsoft internet information server *internet service providers*

Usenet newsgroups *text elements* *TEI guidelines query*

HTML markup *news broadcasts* *web documents*

Staff training *internet community*

Noun clusters with prepositions or conjunctions:

File and segment *a bit from drive* *ring number of* *task and performance* *set of constraints* *object of type*

Method of transports knowledge and technology

development of the program amongst project members

media and information GIF and JPEG files

of mass media framework for media design

next item on the microfilm sets of images

feet of fiber laser and fiber optics loss in fibers

merging of SCC IEEE board of directors

Semiotics as a method of codes of transmission

system of digital archives

Windows NT and Microsoft internet information server

access to the internet guarantee of access

usenet and the internet use of the elements span of text

news and information text and graphics sites on the web

range of training programs information on the internet

Specifying these multi-word items may be helpful for teaching aims –e.g. noting how certain types of prepositions are used with specific nouns (*access to the internet vs. sites on the internet*). The purpose of such a procedure should remain pedagogic in the sense that learners be afforded enough feedback to draw their own contrastive data by making lexical sets according to use. Nominal clusters having certain colligations (prepositions or conjunctions) tend to provide a productive exercise in this respect, as Hunston and Francis (1998: 51) observe.

In the case of noun compounds, a useful approach seems to be the contrastive observation of different textual excerpts where the specific noun compound is explained. This chore illustrates how specialized meanings may be drawn from `functional shifts´ of words

operating in scientific-technical contexts (Allen, 1983: 88-89). The following samples in figure 1 offer such a procedure for the combination *document management system*, as viewed in Information Science subjects:[300]

1 the document profile information attached to documents in a *document management system*, and the classification information in a
2 document interchange formats. The current Xerox image-based *document management system* encodes information about a scanned document
3 ASIS from Information Dimensions is the most widely used Enterprise *Document Management system* in the world and additional
4 Indianapolis (IUPUI) entered into negotiations to acquire a Xerox image *document management system* and its prototype
5 Java application that was integrated with a FileNET *document management system* so that they could deliver important documents

Figure 1: Using context as a form of describing nominal compounds.

Only from these samples, a better understanding of what this concept refers to may already be inferred. As a result, the core noun –**system—** may be described as serving specific functions --*encoding information* about documents (co-text # 2 in figure 1), *offering profiles* of such documents (co-text # 1), and *contributing to deliver* these documents with Java applications (co-text # 5). Then some examples of these types of systems may be asked for, and the learners should thus pinpoint names such as **Xerox** (co-texts # 2 and 4), **ASIS** from Information Dimensions (# 3), or **FileNET** (#5). The definition of the item *document management system* is thus made explicit by contrasting different contexts, a micro-skill quite valid for academic reading performance.

[300] See chapter 3 (`Contrastive study´-- Information Science category).

II.1.C. Adjectival collocates.

A third major type of collocational device that also recurs throughout our data – after verbal phrases and nominal compounds – is the co-occurrence of adjectives with nouns. In this case, an academic task that may be set up is the learner's drilling exercise with the annotation of different attributes for a given device or product of Information technology. These could be derived from textual descriptions themselves or the induced analysis of data from reading and listening chores such as note making and note taking. The wider the range of texts and situations the specified item is found to encompass, the more likely the results will turn out to be useful.

For instance, the adjective **effective** appears in 14 sources of our corpus (see appendix 4), and in nine of them, it co-occurs with **cost** in the compound *cost-effective*. This collocation may be considered a `restrictive adjective-noun collocation in English´ (Bäcklund, 1981: 5) in the sense that it is `functioning as a single lexical item´ (ibid.). In this respect, it accepts, in turn, a limited set of nouns, such as **method**, **manner**, and **strategy** in our corpus.[301] Thus, clusters such as *a cost-effective method*, *in a cost-effective manner*, and *the cost-effective strategy*, are typical adjectival collocations, chiefly idiomatic. In them, **effective** triggers the employment of **cost**, and this combination, in turn, activates the other forms used.[302]

In this way, for a profitable application of such lexical structures, the learner should be encouraged to find ways of describing issues specifically. For example, he or she should be led to mind the important aspect of economic reliability in networking facilities; by stressing the utility of this guarantee, several forms or adjectives may be evoked, from the more general

[301] The reference example offered by Bäcklund is the compound *blithering idiot* being modified by the modifier **perfect** in *perfect blithering idiot*.

[302] In this sense, Bäcklund (1981: 6) recalls the example of predictiveness that the noun **omen** offers in its co-occurrence with the adjectives **good** or **bad** (see `Concordance´ in chapter 1).

– **cheap**, **worthy**, **economic** – to the academic items registered **–reliable**, *cost-effective*. In addition to guessing, students should work with textual clues in the process of recalling such modifying markers (Kelly, 1990). In this respect, the more restrictive the given adjective + noun collocation is, the more productive the search will likely become for the learner (Bäcklund, 1981).

Similar cases are given by predictable co-occurrences of adjectival past participles with nouns, as in the compound *desired outcome*. Both components occur in the same number of sources (13 texts according to appendix 4); among those texts, five of them register such a collocation. To the same extent, then, both **desired** and **outcome** exhibit a similar degree of Mutual Information – M.I. score—since they always co-occur together in the same number of sources (see chapter 1 [`Concordance section´]).

A similar phenomenon arises from the use of adjectival present participle forms, such as **existing** preceding the noun **program**. The level of co-occurrence is inferior, in this case, for the noun, since **program** tends to collocate with a much wider range of words. **Existing**, in contrast, is an adjective that combines with this noun more frequently than with any other word, according to the collocates chart (12 co-occurrences). The expression *existing program* thus turns out to be restrictive in this sense; consequently, its use in exercises should be prompted – e.g. together with alternative related constructions such as *the current program*, *the actual program*, *the program used*, etc.

Nevertheless, finding qualifying suitors for given items is not always a feasible matter. Among these adjectival collocations, an illustration of such an endeavor may be afforded by contrasting **previous** and **marked** in their intended `marriage´ to the noun **section**. Whereas the first adjective tends to appear with other nouns as well – e.g. **chapter**, **item**, etc--, **marked** remains `loyal´ in keeping company with **section** alone. Nonetheless, the collocation

previous section discloses a higher Mutual Information score, i.e. it is found across a greater number of lines and texts. It is thus more predictable than *marked section*, even though the form **marked** remains bound to the noun **section** in this fictitious `wedlock´.

Still, for our task purpose, the aim is not only to find common collocates, given a designated item. In contrast, we seek to establish optional ways of expressing a specified fact. This agrees with the notion in Corpus Linguistics propounded by scholars in many cases: that showing word use does demand some level of introspection on the part of both learner and teacher in order to assess results: `This seems to be the most favorable point for the operation of instrospection – in evaluating evidence rather than creating it´ (Sinclair, 1985: 82).[303]

In the case of *marked section*, in this way, the construction should be attempted amidst a variety of similar forms – e.g. *the underlined part*, *the highlighted content*, *the marked section*, etc. This production of equivalent patterns is enhanced to articulate correct and coherent phrases. The emphasis is consequently placed on learning not just forms or items, but their use as `chunks of language´ which are introduced in context when the need arises (Norris, 1983: 16). This actually conforms a suitable way of making language mistakes more explicit to the learner, since they get to examine options in a context (ibid., 22): e.g. *the marked section should be included in the document*, or *the marked section may contain these items*.

---

[303] The CoBuild electronic corpus resources thus offer online activities based on both empirical observation and introspection to form meaning and usage. Learners are challenged to cope with concordanced data in order to puzzle out senses and definitions (e.g. in games and competing tasks – see web page at http://www.cobuild.collins.co.uk).
Background theory, in this respect, is given in Rossner (1985: 100): `that the descriptive task would serve as a reminder or as a means of increasing awareness´. In this sense, according to Underhill (1985: 112), `the dictionary entry is not a substitute for the learner's experience of a word in action´ (although it may be a good start). In fact, at times, given this degree of learning inspection, the meaning of a word may be included or excluded `for any number of reasons other than frequency of occurrence´(Hill, 1985: 120). In this sense, as Ilson asserts (1985: 3) lexical demands in the foreign learner may be compared with a native child's need of words: the process of assimilation is through `semantically related senses´, and this can only be accomplished through various stages of exposure, observation, comparing and finally inducing or introspecting. See also sections of `Discourse´ and `Concordance´ in chapter 1, for other scholarly references to the process of introspection or the so-called I-language in corpus-based analysis and learning.

Finally, some adjectives in our corpus collocate in certain environments as `fixed´ or `lexicalized´ items (Bauer, 1983: 48). We may say that these are Information technology-related items already receiving what Bauer calls `society's stamp of approval´ (Bauer, 1983: 43). Some of these `institutionalized´ examples are offered by the adjective **electronic** before nouns such as **mail**, **bulletin**, **text**, **form**, **media**, **information**, **newsgroup**, etc. The fact that abbreviated forms such as **e-mail**, **e-document**, or **e-message** may be widespread and conventionally referred to, seems to further attest to this process of institutionalization, by which previously separated words tend to be processed together as time and events make technological advances more accessible and common in every day use (as was the case of the compound **type-writer** in the XIX century – Bauer, 1983: 57).

These adjectival compounds, becoming less restricted, and, in turn, more familiar to the layman, can be recognized and acquired fairly easily by the learner. This is actually an advantage in ESP task development, since the terms are handled with more ease from the start. For instance, the qualifier **optical** tends to be placed rather successfully before words like **fiber**, **system**, **signal**, **character**, **disk**, etc, in an activity based on the description of Telecommunication lines and media. In the learning process, such words are, in addition, used as a means of `exploring a subject of interest´ (Nolasco & Arthur, 1988: 119), recognizable among students, and thus useful for productive skills in peer-activities.[304] The implications for course syllabus design, as shall be seen below, chiefly focus on this facility to indentify typical lexis defining subjects or topics.

[304] This approach turns out to be a suitable task where there are too many students in one class (Nolasco & Arthur, 1988: 119).

II.2. CORPUS-DRIVEN ACTIVITIES

As stated above, Hutchinson and Waters explain (1987: 24-37) that language descriptions such as structural and register analyses (in the 1940s and 1960s respectively) may not provide a suitable teaching methodology in ESP courses. Instead, a task-based approach focused on the learning process is favored (Hutchinson and Waters, 1987: 107), a model that integrates language and content. As a result, in this learning stage, a narrowly devised perspective from subject-specific task design is superseded by wider scopes across `subject specialisms´ (ibid., 166).

Undoubtedly, we agree in our work with the claims stated by Hutchinson and Waters (1987); and, yet, following other scholars – e.g. Flowerdew (1993 a; 1993 b), Jordan (1997), Dudley-Evans and St. Johns (1998), etc--, further specification can be contrasted in terms of reconciling product with process approaches, subject with common core scopes, and technical with semi-technical lexis.

In the next four segments, a range of task-influenced applications derived from the linguistic data in our corpus is thus accounted for as follows: First, presenting constructive work with academic word lists; second, inducing learners to establish lexical bonds; third, directing students appropriately for lexical sifting, and finally, preparing the learner to anticipate given stretches of discourse. Each of the four headings is in turn viewed within the competence of each linguistic macro-skill – reading, writing, listening and speaking respectively. These should be continuously fostered in the classroom, whereas they prove to be rather becoming for lexical acquisition; several works aim to show such a learning process through academic language tasks via electronic media.[305]

[305] There are increasingly more works that present challenging language activities on-line, via web-based resources or CALL exercises. For example, Leloup and Ponteiro (1999) deal with reading and writing on-line, Laufer (1999) also exploits these skills by paying attention to low frequency lexis in texts and how these words can be learned and recalled by focusing on CALL dictionaries and glosaries; Kronenburg (1995) regards electronic mail as a profitable tool for broadening students' writing fluency, while Heap (1993) applies such a

II.2.A. Academic word lists and reading.

A main scope of Information technology students' language skills in Spanish universities is primarily the understanding and ability to discern senses in texts, according to several sources (Aguado de Cea, 1996; Posteguillo, 1997; Rico, 1999). The learner's reading performance may thus benefit productively from having corpus-driven word lists as an ally. However, the aim is not merely the placement of lexical data for the learner to be able to refer to; the intention should further include the application of corpus-driven lexical categories so that a perspective in context may be afforded. As Yap states (1979: 58), `causal links probably do exist between vocabulary and comprehension, and vocabulary is likely to be the predominant causal factor´; in this respect, reading skills and lexical competence are closely related. The more direct and intensive the approach to pivotal senses of words in texts, thus, the more profitable the tasks may become for specific purposes.

This means that the words in the frequency lists should be as common core as possible, since the broader the range of sources and subjects encompassed in this case, the more likely those words will make up the bulk of academic lexis which the learner must constantly cope with through the course of his or her studies. In this sense, as it occurs in Coxhead's *Academic Word List* (AWL – 1998), `the more frequent the words in the list, the more likely they are to occur in academic texts (...) Words with higher frequency cover more of the corpus´ (Coxhead, 1998: 3). Hers, unlike our academic word list (Appendix 4), covers many more subject areas in Faculty sections such as Arts, Commerce, Law and Science. Thus, common coreness must be based on high figures of frequency and range: `over 100

---

line of work in the area of word processing. In this sense, McCarthy (1993) sums some useful CALL utilities for enhancing language learning tasks. Two other recent sources exploring this world of language and technology for the improvement of student performance are Hogan-Brun and Yung (1999) and Levy (2000). In addition, in the fields of ESP and EST, Viel's electronic handbook *Science Teaching Reconsidered* gives guidance regarding the selection of web resources for enhancing student learning (http://www.nap.edu/readingroom/books/str/7.html).

times in the 3,500,000 word Academic Corpus in order to be considered for inclusion in the list´ (Coxhead, 1998: 2). In turn, Mascull's key words in science and technology (1997) also present broad scopes and frequencies across subject areas, which correspond to pivotal issues in science and technology – e.g. information technology, medicine, physics, genetics, the environment, etc.

The chief aspect that should be valued in such types of lexical arrangements, from a task-based perspective, is that the words listed serve to guide learners in `puzzling out meaning´ while reading (Jordan, 1997: 147). Thus, the vocabulary should be put to use not only `from the point of view of the reading context but also look[ing] at ways of activating it´ (Jordan, 1997: 149). Academic lexical items come in the intermediate phase between comprehension and production, according to this view – also supported by Nattinger (1988) (cited in Jordan [1997: 149]).

The academic word list provides core items, in this respect, for the purpose of both learning (receptive) and acquiring (productive) vocabulary (Jordan, 1997: 153). The figure estimated to be suitable for operative approximations to these words is about 2,000 items, according to several sources (Xue and Nation, 1984; Nation, 1990; Flowerdew, 1993 b; Jordan, 1997; Coxhead, 1998). In our case (appendix 4), there are 3,887 word forms, but if reduced to their head-words or lemmas, the number is also close to the amount propounded by scholars – e.g. **defined**, **define**, **defining**, **defines**, **definition** are represented by the lemma **defined** (the most frequent form).

The *Academic Word List* (AWL, 1998), which replaces other less recent or updated general lists in higher education (Coxhead, 1998: 1),[306] serves as a good example for illustrating the focus on lexical common coreness across subject areas. These tend to be semi-

[306] Some examples of other academic word lists used in universities, previous to Coxhead's AWL (1998), are Campion and Elley (1971), Praninskas (1972), Xue and Nation (1984), and Burgmeier et al. (1991).

technical words that relate to the world of investigation, analysis, evaluation (Martin, 1976), or to the rhetoric and structure of discourse (Baker, 1988; King, 1989). Coxhead divides them into ten sub-lists, the first one containing the most frequent elements, and the last one (number 10), the least often used academic lexis across texts. For instance, in the AWL, nouns such as **approach** and **data** or verbs like **involve** and **vary** appear in sub-list 1, whereas a conjunction like **albeit** or an adverb such as **nonetheless** are included in the last collection of words (# 10). There thus seems to exist a higher occurrence of content items[307] dealing with the research or work carried out at tertiary levels, and a less frequent position, in comparison, of discourse markers and rhetorical elements – e.g. the first item of this kind, **hence**, does not appear until the listing of words in sub-list 4 (Coxhead, 1998).

Our list in appendix 4 does not contain discourse words, since these are dealt with in appendix 2 (together with argumentative and procedural items, as they direct the reader through discourse focal points and strategies). Instead, our academic list encompasses content words -- verbs, nouns, and to a lesser degree, adjectives and adverbs – corresponding, in many cases, to Coxhead's AWL most frequent groups of `investigation, analysis, evaluation´ (Martin, 1976). An example is the lexical family introduced above by the verb form **defined** (ranked number 167 in appendix 4).

Through the development and use of these word lists, emphasis is placed on the `need to develop vocabulary in a systematic way rather than by incidental learning´ (Jordan, 1997: 162). Relationships are thus triggered as word associations in context, whereby students infer and induce meaning in the readings, since `stretches of language are read for a purpose´ (Jordan, 1997: 158). Learning these relationships between word forms and meaning is a key procedure in the process of understanding and making sense in the text. As Horst et al. (1998:

[307] These content words are mainly verbs and nouns, although some adjectives (e.g. **available** and **significant** in sub-list 1) also appear.

208) claim, the process of reading is actually a matter, in many cases, of adequate mental sifting of words – i.e. choosing which and why some are more useful than others.

The appropriate use of word lists as reference or checklists for examining the most frequent occurrences can be rewarding, as learners can then approach the reading by pinpointing where and for what purpose given lexical items occur.[308] The TALO (Text As Linguistic Object) and TAVI (Text As Vehicle of Information) schemes can rightly be met through this contrastive focus: Both vocabulary and task are exploited in working with the key language that is used to deal with academic facets. For instance, the form **defined** is highly employed to introduce the description or definition of a concept; yet, a verb like **called**, appearing in position 157 in our academic word list (appendix 4), is advocated more and across a wider range of texts (36 sources) than the former, **defined** (ranked # 167 and appearing in 25 texts). We may consequently infer that defining by using phrases such as *this is called a third party transfer* or *a method which we call* turns out to be rather common in our academic corpus, even more so than examples like *every element is defined as*, or *we define a point to be.*

Having, as a result, this type of information available in the form of frequency charts of common core academic items aids reading procedures and techniques considerably – e.g. skimming and scanning for specific occurrences of conventionally employed words, predicting and anticipating such uses, associating key words to descriptions, definitions, classifications, etc. A useful activity, in this sense, can be provided by a pre-reading exercise where learners are encouraged to elaborate semantic sets or fields that include frequent words from the academic list. These should be related or associated in context and counted as items belonging to specific semantic categories, listed in frequency order. For instance, the two

[308] Since our academic list contains almost 3,900 items, sub-lists should be handled in the process of lexical work with the readings. For instance, only a 100-word extract may be viewed at a time.

verbs introduced above, **call** and **define**, may be registered within the same group, being given a collectively agreed upon functional label – e.g. `definition´. **Call** would be the head-word in such a set due to its higher frequency and wider range. Other verbs might follow, such as **named**, **designated**, etc, occurring with a much lower frequency in appendix 4. Special attention should also be paid to homographs -- words with the same spelling but different meanings or senses-- such as, in the semantic set of **call**, the word **name** (ranked # 143 in appendix 4). It almost exclusively functions as a noun, and its verb form equivalent (**named**) does not appear until position 1,156. This implies that learners should be made aware and reminded of how the same word can be used quite differently (i.e. with different frequency and ranges). In this way, the ability to predict how words tend to behave in certain cicumstances, from an earlier activity with them, may be exploited by means of lexical links – e.g. making families or connections across `the web of words´ (Carter and Long, 1987: 82).

In the same line of work, subject-focused words may be contrasted with the use of academic items across a larger number of subjects or topics. For example, the verb **organize** (appearing in 12 texts, in position # 2,225 in appendix 4) may be checked in comparison with **arrange** (10 sources and ranked number 3,370). The former covers several subjects with relatively few occurrences (24) -- categories A1 (History of Computers, Hardware, Software), B2 (Online Database Systems, Computer Systems), B3 (Automated Knowledge-based Systems), D1 (Media Technology), E1 (Communication Theory), F3 (HTML, SGML, TEI), F5 (Electronic Publishing) and F6 (Information Infrastructure). The other verb, **arrange**, crops up, instead, in two main subject samples – A2 (Computer Engineering, Data Communications and Client-Server Communication) and B2 (Online Database Systems, Computer Systems) --, with restricted collocations such as *arrange nodes* (in a Computer science article) and *arrange lines* (in a Library Science article). The first verb, in contrast,

offers a broader span of lexical combinations in different academic contexts: **organize** thus collocates with **lessons**, **knowledge**, **documents**, **content**, **information**, and more nouns of this type.

In this respect, `awareness of certain lexical fields or other textual features which are likely to signal the presence of the information we are looking for´ seems to be the product of an effective scanning exercise (Haarman et al., 1988: viii). Such procedures in reading can thus benefit from the use of significant academic word lists.[309]

II.2.B. Lexical bonds and writing techniques.

Establishing mental associations among words in given stretches of discourse seems to be a constructive task for genre awareness-raising (Hoey, 1983; 1991; also Jordan, 1997: 169).[310] Appropriate readings during the academic year may thus trigger the learner's recognition of certain lexical strategies that help `to assimilate conventions´ of genre and register (Jordan, 1997: 169).

The learning process involved in this sort of work entails the development of techniques in acquiring vocabulary for productive use. According to Aitchison (1987: 162), this expansion demands `an active system in which new links are perpetually being formed´. For this, becoming proficient in reading skills contributes to easing the process, since `bonds between co-ordinates and collocational links are found to be particularly strong´ (ibid., 83). For producing lexical behavior in tasks, in fact, two key faculties in conveying meaning – organization of the text and thematization – play key roles (Grellet, 1981: 93, 127). These deal with the capacity to interpret how the information is arranged by focusing on specific

[309] Haarman et al. (1988: 103) also introduce a list of `lexis common to academic prose´, though these are less than 200 and are mostly taken from the Social Sciences.

[310] See sections on `Discourse´ and `Genre´ in chapter 1 for theoretical studies by scholarship on discourse cohesion and genre description through key lexis.

words (Grellet, 1981: 127). Language is thus learnt `from a text through intensive practice on one vocabulary area´ (Williams, 1984: 45).

Again, both product and process orientations may tally in the management of lexical items for productive purposes. Regarded as product, the goal would be coming to terms with the genre and register conventions implied by the subject matter. In addition, rhetorical features of description, funtionality, operation, etc must be accounted for as product-based models of writing (Robinson, 1985: 99). Concerning process, the learner's task also comprises the composition work of reading, revising, drafting, summarizing and paraphrasing (Jordan, 1997: 169). In both estimations, as Pueyo and Val (1996: 255) explain, `the degree of expertise (...) will depend upon the knowledge of the lexis, taxonomies, and activity sequences it contains´. Since subject fields may be recognized by looking at the lexis, what should be underscored, as a result, is the manner or means of lexical arrangement.

As Pueyo and Val (1996: 265) observe, in addition, in many cases of technical writing,

the information is clearly organized into given and new, thus giving rise to a chain-like effect pattern in which the given information in each sentence topic refers anaphorically to the new information in the last occurring comment.

For example, as was examined in the last section of chapter 3, the context of the nominal construction *network address* (figure 2 below), as offered by a technical report on Internet infrastructure, is distributed according to the given / new or theme / rheme parameters:

Every organization connecting to the *network must have a unique identifier*. *This identifier is known as the campus IP* ***network address***.

Figure 2: Defining a noun compound by means of lexical cohesion.

Getting acquainted with this type of structure is important, and, in addition, the concepts which are described or explained in such sentences and paragraphs may also be taken as objects of inquiry and observation. As may be surmised, these are often nominal compounds packing `complex processes and events´ (Pueyo and Val, 1996: 262). [311] At times, as they turn `happenings into things which can be technicalized´ (Pueyo and Val, 1996: 263), the occurrence of these noun compounds is analyzed as a `grammatical metaphor´ (Halliday, 1985; Eggins et al., 1987; Salager, 1990; Love, 1993; Partington, 1998; Peignan, 1999). Through such a notion, nominal collocates act as entities condensing the reference to actions, events and processes: `the packing and compressing of complex information into a compact unit ´ (Pueyo and Val, 1996: 258). For geology textbooks, for instance, Love (1993: 203) recalls `subjects of processes´ such as the nominalization **removal** or **mass-movements**.

In our corpus, the more numerous and clear examples of these occurrences tend to be registered in subject-bound contexts such as those listed in appendix 6, where the vast majority of words are nouns. For instance, in the B3 subject heading -- `Automated knowledge-based systems´-- the compound *knowledge management terminology* is used to refer to *activities -- important and pervasive as they are. Most of the evolving terminology of knowledge management springs from a growing awareness that knowledge management must become a formal business activity.* Due to the contextual clues which the given passages may afford, the `unpacking´ of the encapsulated input in the nominal cluster may be enabled. In

311 As Firth (1930) explained – quoted by Leech (1981: 61), `by regarding words as acts, events, habits, we limit our inquiry to what is objective in the group life of... fellows´(see `Discourse´ section).

this case, *knowledge management terminology* involves the following lexical devices: **activities**, **important**, **evolving**, **awareness**, **formal**, *business activity*. If viewed in a different context, other words may be gathered. For instance, in the excerpt (figure 3), the underlined items could also contribute to gloss the concept.

> Knowledge management terminology, a new language for new leverage: the terminology of knowledge management relates to the proliferation of terminology associated with knowledge management. This is understandable, if nonetheless daunting. White-collar workers spend most of their days in knowledge-related activities: one-on-one conversations, formal meetings, reading, writing, processing information.

Figure 3: Lexical bonds helping to describe the noun compound term.

The main goal is to establish lexical bonds by which the senses embedded in the technical item *knowledge management terminology* may be inferred. Subject-based nouns, in this case, elicit other nouns (a phenomenon occurring, according to Aitchison (1987: 100), around 80 per cent of the time). Thus, the `grammatical metaphor´ contained in the noun compound should become more explicit by means of unfolding its contextual nominal associations.

In the sample examined (figure 3), other properties of *knowledge management terminology* can be detected by focusing on the new information (rheme) following given data (theme): the new **leverage** involves **proliferation** of terminology, which is **understandable** and **daunting**, since there are many activities: **conversations**, **meetings**, etc. In the thread of this pattern, the structure signalled by `cause and effect´ seems to be common in technical writing (Pueyo and Val, 1996: 259-260).

Teaching implications are developed around the notion of acquiring academic writing

micro-skills. These should be fostered through the association of lexis in subject-related or technicalized definitions, descriptions and explanations. Some such techniques may be located in guided writing material such as the British Council's (1980), Pincas (1982), or Riley (1999). In them, conventional ways of expressing academic concepts may be exploited by means of pre-writing exercises, such as the listing of properties in a table or chart. Rosenthal and Rowland (1986: 2) add the important strategy of communicating derived lexical components of a given key word; for instance, in the item described above (figure 3), we can ask that the noun **management** be divided into its constituent parts -- **manage** + **ment** -- to encourage further suffixing (figure 4).

Manage / management / manager / managing / managed

Figure 4: Suffix classification for the stem **manage**, according to our data (appendix 4).

Relationships of this sort enable categorizations to be made, in which a grid may be filled that prompts the learner to identify specified forms. Figure 5 asks for the activation of word-forms derived from use – see more examples of this activity in appendix 8.1.

NOUN	VERB	PERSON	ADJ.	PAST
Management				

Figure 5: Activating word-forms commonly used in our corpus.

Another type of exercise embracing the challenge of producing the right items is the word formation of nouns, based on a similar activity proposed by Riley (1999: 5). The aim is to choose the appropriate form. An example from our data includes the selection of the

correct item (figure 6 below – more examples in appendix 8.1.).

MANAGE +

a. **SHIP**
b. **MENT**
c. **NESS**

Figure 6: Choosing the correct specialized noun form.

These activities should be carried out, as stated above, previously to an integrated writing task, and after having read both intensively and extensively on the subject accounted for (e.g. on knowledge-based systems – B3 category).

The work with lexical bonds from above, in this manner, may then be extended to other kinds of exercises in which the main goal is to reflect knowledge of web-like associations in the text. In this case, as examined above, links are made to establish relationships between words, either based on syntagmatic processing – e.g. collocability--, or on paradigmatic equivalents – e.g. through synonymy, antonymy, hyponymy. Leaverton and Zymantas (1999) offer an activity closely resembling our own (figure 7), in which partnerships or collocations are sought. The purpose of the elaboration, in this case, is to reflect success in the procedure of acquiring key subject nominal clusters.

_____ management _____
_____ of _____ management

Figure 7: Seeking subject-based collocates in written production. See appendix 8.1. for further exemplification.

A variation of this kind of lexical development is displayed in figure 8 (inspired by Riley, 1999: 1), where four words collocate with a main one on both sides – to the left and

right of the node. They serve as feedback for guessing the item wanted.

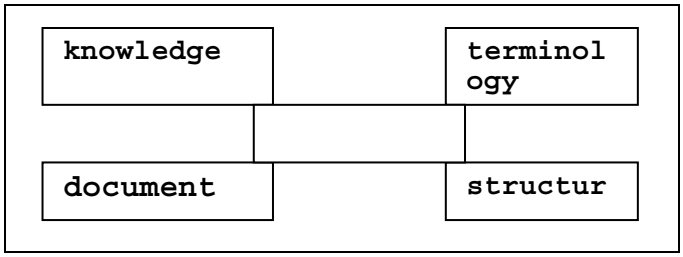
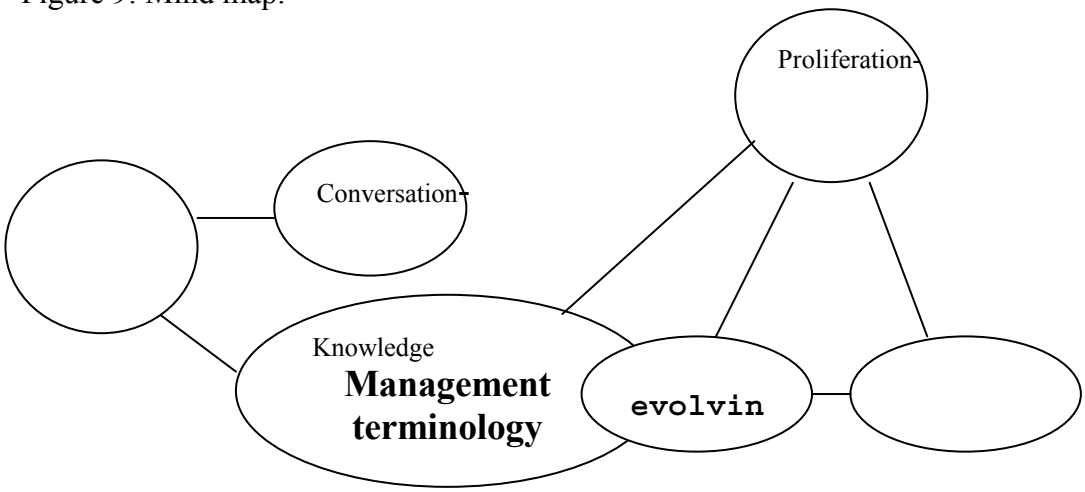


Figure 8: Guessing the central collocate. See appendix 8.1. for further examples.

However, they do not have to co-occur at the same time, as is the case of *knowledge management*, *document management*, and *management structure*. **Terminology**, in contrast, always collocates with *knowledge management* in our corpus.

The other type of lexical connections, based on hierarchical processing, may be defined as the linking of semantic networks – e.g. sets like **home-room-bedroom-kitchen-backyard-pool**, etc (Levy, 1999: 12; Aitchison, 1987: 83). These are spurred by means of charts such as mind maps – influenced by Riley (1999: 10) (figure 9)--, where the key concept *knowledge management terminology* may be broken down into features of description, such as **proliferation** or **activities** – according to the text in figure 3 above. Hence, semantic associations can be traced among words: **conversations** being evoked as an example of **activities**, or **proliferation** triggering synonyms like **evolving** and **growing**. In figure 9 and in appendix 8.1., mind maps may be made either more or less open: some words may be inserted or all the circles can be left empty – the results vary in terms of whether the exercise is intended and designed to be more or less precise according to the text being read.

Figure 9: Mind map.



A different way of having these nodes interrelated is by the application of tables, according to the British Council (1980: 13). These are probably most useful in the induction of features and functions. The plausible disadvantage is that lexical tying is thus conformed in a more austere manner (figure 10), since these ties are represented in columns and rows. While column **A** presents the main features described, column **B** displays the functions of such features. However, this arrangement exercise probably challenges learners to more effectively and more successfully achieve the interpretation of nominal compounds. For such an endeavor, correct placement of nouns into either features or functions, proves to be an evaluating technique. As can be observed in figure 10, again, most lexical bonds are elicited by nouns in technical contexts – i.e. the underlined items in the text below.

In figure 10, the table is based on the contexts that precede it, taken from the F1 topic sub-corpus on perspectives on Information. The issue being dealt with is an application for automated library systems, the *BASIS Techlib application*.

UNIX is used for the database server, a single NT machine can be *the BASIS Techlib application* server and the librarian's workstation.
For *the BASIS Techlib application* there are some additional system administration tasks required
Update privileges (e. g., add, update, delete). In addition, *the BASIS Techlib application* separately authenticates user privileges

A	B
NT machine	Server Workstation
Tasks	System Administration
Privileges	User

Figure 10: Text and its derived table of features (A) and functions (B).

Finally, using lexical relationships may be profitable for the writing process of formulating phrases and paragraphs based on concept or event development. In this respect, a rewriting technique may be appropriate for ensuring the active recognition of compounds in context. For example, verbal phrases or clauses should be transformed into noun-focused sentences, or we may request that certain noun compounds be unfolded – i.e. verbalized; such reconstructing should be made as implicit as possible, so that learners have to bring both their knowledge from the texts and faculty of inducement by contrasting the two clusters shown (figure 11).

New users interact with the Internet	
In the area of human – computerInteraction	

Figure 11: Rewriting by changing verb into noun compounds and viceversa. Other examples are shown in appendix 8.1.

An expansion of this activity tends to be `free writing´ in comparison, by which learners aim to demonstrate their `ability to write freely what has been taught´ (Pincas, 1982: 110). Paragraph composing by elaborating descriptions, for instance, which crowd technical reports (Martin, 1985: 7), may be regarded as part of the task. The main goal is that students form their own lexical and syntactic patterns by relying on previously read material, whereas no hints are provided in terms of contrastive data. Based on the Library Science report from which the co-texts in figure 10 were taken, for example, students are asked to enumerate the features of the *BASIS Techlib application* in the following chart (figure 12):

	Features of *BASIS Techlib*		
1		be the server for	
2			
3			for users

Figure 12: Free writing based on previously read data.

As may be deduced, the major thread for composing seems to be woven by both acts of observation of lexical behavior and working with these relationships actively. The techniques enumerated are merely samples of how production practice should be dealt with from the perspective of lexical bonds in conventional academic settings.

II.2.C. Lexical sifting and listening for the right word.

The closing stage at free writing above is closely related to another task in the academic environment: `note-making´, which involves `summarizing, and paraphrasing´ (Dudley-Evans and St. Johns, 1998: 104), two micro-skills highly related to procedural competence, as was stated in chapter 1, and typical in language teaching as strategies included

in `action research´ (Rea-Dickins, 1993: 84). In short, `getting information down´ is a key device for academic performance (Dudley-Evans and St. Johns, 1998: 104). These two authors also contrast this elaboration with the process of `note-taking´, in which processing language, relating new information to existing schemata and recording new related information are cardinal chores. Note-taking thus differs greatly from written production in that the major goal is to set up `mental maps´ based on previous linguistic and content knowledge (i.e. built schemata) (Jordan, 1997: 189).[312]

In the case of technical input, the task of listening while taking notes is mainly that of `listening to monologues´-- e.g. `to listen to and understand instructions´ (Dudley-Evans and St. Johns, 1998: 102). In the performance, being able to identify pivotal lexical devices such as `micro-markers´ and `macro-markers´ of discourse is analyzed as a chief procedure (Jordan, 1997: 184). The former types of connectors refer to items `structured at the micro-level´, such as words connecting clauses or sentences – e.g. **so**, **and**, **but**, **now** (Flowerdew and Miller, 1997: 33). Macro-markers posit, in turn, major lexical shifts in the discourse in the form of chunks of language that summarize, conclude, recall data, etc. (Nattinger and DeCarrico, 1988). Some examples are **in summary**, **to conclude**, etc.

However, as Flowerdew and Miller claim (1997: 28), students need to develop specific micro-skills in listening to monologues that may not coincide with those stressed in conventional English for Academic Purposes (EAP) material. Because of the lack of authenticity in the technical lectures included in many textbooks, these authors, Flowerdew and Miller, propose working with listening abilities that focus on more interactive features of spoken monologue. These are often embedded in the choice of micro and macro markers that do not correspond to the traditional segmenting items **therefore**, **thus**, **however**, **as a result**,

[312] Schema theory is explained in chapter 1 (e.g. `Rhetorical features´ in `Discourse´ section).

etc. Instead, lecturers seem to favor repetition, redundancy, checking, and integrating visual aids (Flowerdew, 1997: 43), which involve phrasing and clusters such as *according to this figure, er, in this chart…*, *let me put it in a different way…*, *let me put this slide up…*, etc. These expressions would often contain key elements that must be correctly decoded; the goal is for learners `to use their language knowledge and other available or previously acquired information to predict or anticipate what will be said (Kelly, 1991: 138). Actually, among advanced learners, the measurement of their command of language skills can be often conducted through checking these multiword units competence (cf. Cowie, 1978; Kelly, 1991; Hunston and Francis, 1998).

Our corpus-driven academic phraseology may, in some cases, aid to prepare learners for the distinction of such types of occurrences in spoken discourse. A useful exercise is the development of lexical discrimination by which the key elements in phrases are adequately sieved. The procedure should serve to couple these devices as micro and macro connectors in the speech. For example, the manner of starting a new point in the lecture may be signalled by such phrases as *we come now to the next issue*, *this takes us to the following point*, or *now, the other topic that we should examine…* (cf. Flowerdew and Tauroza, 1995; Chan, 1995). In most cases, the items chosen are general academic words – **topic**, **issue**, etc— and procedural expressions like *the following point*.

These belong to semi-technical lexical sets of general vocabulary that is repeated and distinctive across academic texts (Baker, 1988: 95-96). In addition, since `vocabulary learning (...) is a never-ending process´ (Kelly, 1991: 138), these subject-bound and general or specific academic words must be learned in context.[313] Other examples from our corpus are transitional or signalling expressions such as *as we shall see*, *we define this as*, *we can call*

[313] In EAP, these elements are considered characteristic of EGAP (English for General Academic Purposes) (Jordan, 1997: 249-250).

this, etc.

The teaching implications that we may garner from the exposed issues center on ensuring that learners can become aware of such shifting devices in oral discourse: To keep them `involved and remind them of how the information in the lecture can be linked´ (Flowerdew and Miller, 1997: 38). The lecturer's strategies should be understood by appropriately sifting the lexis employed. Thus, working with comprehension checklists that ask learners to reflect their knowledge of the speech structure can be rather convenient: For example, to identify main issues and discriminate secondary ideas, locate an exemplification offered, distinguish the use of contrastive information, etc. The use of graphical aids in such types of questioning may be suitable so that learners may take notes by visualizing lecture organization.

Figure 13 contains a table where students must rearrange some given information based on the listening. Figure 14, in turn, represents a more complex procedure where the aim is to produce the data in context, marked by the lexical items or structures noted down to the left and right of the table. The more advanced form of this drill is depicted in figure 15, since major and minor points from the lecture are asked in a given order, while these are unnumbered in the oral excerpt. In figure 15, according to the features described in the lecture, the learner must distinguish how and when the characteristics are dealt with, noticing their introduction, definition, comparison, etc. For all three models, the focus should be placed on the ability to recognize the development of the lecture segments by means of semi-technical lexis – e.g. words like **kind**, **elements**, **defined**, etc.

The following text is a lecture extract based on a report contained in our corpus, delivered in one of our technical English classes.[314] This is done to follow Flowerdew's and Miller's suggestions (1997: 44) that `language instructors would do better to reduce the text to note form and then present the lecture extract themselves (or have a colleague do it), speaking from these notes´. On such a spoken paragraph, delivered spontaneously from written notes and accompanied visually by an outline of the lecture, the exercises in figure 13, 14 and 15 are based.

Today's topic deals with the fundamentals of all visual communication … These are basic elements, [pointing to the slide] … these are the compositional source of all kinds of visual materials, …for example, the messages, the objects and … the experiences as well… In this way, …we have that the most basic element is the dot, …which can be defined as a pointer, a marker … a marker of space … the other element is the line… This is an articulator of form, …that is, a design item for making a technical plan, … so it designs the form intended, …ok; …another element that we can think of is the shape, which is the basic outline, … for example, we can think of a circle, a triangle, or a square, and these are all elements of shape… OK, then we have direction, which means that there is movement, and this has a lot to do with the recognition of shapes for example… well, then the most basic of all the elements is value, … this is the way of either having or not having light, the marking of light or the … the deletion of light… then, well, …Two more items for the communication of images… You can have hue and saturation, … the first one refers to color—coordination, this, the color-coordination of value, the addition of more light … this also has the given name of chroma… Another feature is texture, texture is another feature,

[314] As the ESP instructor, our role as a content lecturer is not authentic; yet, our only having notes and a slide where an outline of the speech was written, could be identified as natural enough – according to Flowerdew and Miller (1997), this is better for authenticity purposes in ESP and EAP than recorded transcripts, read by native speakers and listened to from the cassette tape. In our case, students could see us physically, body movements,

and it can be optical and also tactile, … We can also define texture as the surface … the external layer of the visual thing … while then, we can see the property of scales… scale is another characteristic, you see… scale is the size of the image, how big it can be, or how small… we can measure the image then… this, the use of scales can do it… OK, only two more now, … you can have dimension of course, and for the movement of things, of course, motion… a characteristic very much used in films, for example, isn't it? … So you have all these elements that are visual, images … based on images… they are the materials used in the field of visual intelligence… here in this field we can see some levels, different levels in the elements… because in these elements the spectator,… the observer can read the picture, that is, the visual element… It is the way we see things… the elements help us see them and distinguish things… This is the visual syntax… how the spectator can read the image... and from this concept we also have the concept,… the idea of what we can call visual literacy … This, the theory of visual literacy is based on the visual elements that we have described … all the elements, value, dimension, all these … [signalling the slide] … OK, so you see that the concept … visual literacy is the ability for being able to know images, to understand the elements, the components …This is the main idea of … the issue … the idea of visual components to let you distinguish elements.

The line is the articulator of form
The fundamentals of all visual communication are its basic elements
make up of color—coordination of value with added component of chroma
The most basic of visual elements, the dot
Visual literacy is the ability, through knowledge of the basic visual elements
with the understanding of these elements that a viewer can come to understand visual syntax

Figure 13: Re-arranging phrases according to their order in the lecture.

gestures, hesitations, eye-contact, etc., which is also quite beneficial for comprehension (presencial interactivity

All kinds of Visual materials:		
The most basic element:		Defined as a pointer or a marker

Figure 14: Locating key words and structures in the lecture, given the context of their placement.

1st element:	**Definition:**
3rd element:	**Examples:**
4th element:	**Example:**
5th element:	**Contrast:**
6th element:	**Reference:**
8th element:	**Classification:**
11th element:	**Exemplification:**
General Field of elements	**General function of elements in field**
Concept of understanding elements	**Example**

Figure 15: Responding by note-taking different focal points along the lecture.

with them) (Flowerdew, 1994: 17).

II.2.D. Anticipating lexical input in oral deliveries.

A convenient number of words for the acquisition of lexical competence in a specialized corpus should be a figure of approximately 2,000 items (Flowerdew, 1993 b: 233). This premise is `based around communicative activities´ as proposed by Hutchinson and Waters (1987) (Flowerdew, 1993 b: 232). The task of giving a short oral presentation before an audience – e.g. the class-- can thus be satisfactorily achieved by the proper use of such a range of words; this assays the learning process in the expansive production of right word combination in context.

Oral deliveries in EST settings should not exceed the seven-minute limit, according to McMurrey (1999: 2), or be between seven and twelve minutes long (Peterson, 1998: 31). The macro-structure tends to be that of presenting a problem and a recommended solution; thus, a well planned introduction is rather important in order to `set an objective framework in which the audience will accept the information as accurate and as significant´ (Laster and Pickett, 1996: 442). However, as Flowerdew and Miller observe (1997), there are several variations and exceptions to the rule. Instead, for example, contrasting data may be a pattern, followed by a discussion or concluding remarks on such figures. Or, in other cases, a product may be described, and advantages and disadvantages may ensue from such a description. In all instances, there seems to be a modus operandi in which the prevailing note is paced by `the preparation and delivery of critical subject matter in a logical and condensed form, leading to effective communication´ (Morrisey and Sechrest, 1987: 2). The inclusion of the following features in successful scientific-technical presentations therefore bear equivalent values:

- Objectives.
- Planning.
- Organization of material.

- Visual aids.
- Delivery and practice.
- Audience.

The adequate measurement and estimation of these elements may lead to an effective manner of presenting and dealing with content in an oral delivery. All the parameters should be made clear and concise. Visual aids –e.g. transparencies, multimedia applications, etc--, in addition, should not be more than twelve, so that learners may not become overwhelmed by the data (Peterson, 1998: 33). Taking the factors of delivery and audience into account is highly important, as well as is that of keeping a proper pace –i.e. neither too fast nor too slow with regard to presenter – audience interactivity (e.g. eye contact, voice projection, etc), these can be central points for facilitating interest and comprehension in the lecture (Peterson, 1998: 35).

The main goal of these tasks is, as a result, effectiveness in oral deliveries from the viewpoint of communicative skills. In scientific-technical settings, this involves the development of strategies for coping with subject-based linguistic content (Rosenburg, 1983: 217). Although learners should be left the freedom to organize their report in specific ways, guidance as to which presentation techniques can be included may be offered by the teacher. For instance, a clear introduction where personal greetings and content preview are stated should be encouraged. Then, other aspects of oral deliveries may be recommended through exposure to previous works (including the teacher's own demonstration) where items such as illustrations, concept and key word explanations, sequencing or listing, introducing a new topic, giving examples, instructing, giving opinions, and concluding are developed (Bennet et al., 1998).

Linguistic devices may be employed significantly and effectively through such organization items. These should be assessed in the light of their usefulness in given sections of the speech (e.g. using passive or *we-statements* to present findings: *we have found the following data*, or *the following data is found*) (Krzanowski, 1999). The fact is that, as our own experience reveals in class, the vast majority of students feel that such a method in oral reporting helps them primarily to gain both content and linguistic knowledge, and within the latter, they consider gained vocabulary competence as the factor most widely achieved --95 %-- (Curado, 1999).[315]

In addition, technical words are not analyzed as more important than academic or semi-technical items by learners. Quite the opposite, during the course of their oral presentations, students tend to feel the need to have a greater command of common core items such as verbal and noun constructions of the type of ***cope with***, ***provide with***, **submit**, *system performance*, *the development of ideas*, etc. In fact, some common mistakes registered among our learners are lexico-grammatical instances such as colligational and collocational blunders –e.g. **discuss about*, **another reasons are*, **comment some issues*, **have errors*, etc., and syntactic errors like **a hacker don't need*, **the major are three*, **we can see what is a virus*, **the program finish*, etc. They tend to be quite frequent in oral presentations, more so as learners read less from their notes and focus, instead, with more intensity on developing the main ideas of their projects in their own words.

In this kind of approaches, as Bygate (1987: 74) claims, organizing content properly is the underlying means for comparing, categorizing, differentiating, and developing other rhetorical devices in a suitable manner for comprehension purposes. Evaluation of presentations carried out by peers in class seems to function as an appropriate mechanism of

[315] It may also be worth mentioning the fact that 100 % of the learners regarded the aid of visual material as quite beneficial for their speech, whereas only 40 % felt positive about their actual delivery.

enabling and allowing learners to realize problems with their work in terms of linguistic performance.[316] Syntax and lexis are focal points of assessment in this sense, in agreement with Jordan (1997: 203-205). In fact, students seem to notice that the balanced semi-technical use of words in their befitting contexts plays a crucial role in the ability to convey subject matter clearly and concisely.[317]

In speaking academic monologues, as a result, as Dudley-Evans and St. Johns state (1998: 112), this lexical `signposting´ or `signalling´ of the input in visuals offer useful aid for the refinement of text type structuring or moves along the speech – e.g. in a report presenting positive evidence for using a certain software program, classifying the advantages should be the focus of the procedure. Available practice should thus be put into specific ways of addressing the topic or subject encompassed. Students may thus `make use of the language from the input and vocabulary and grammar sections (and which have been derived from the corpus) in negotiating the production of the authentic piece of English´ (Flowerdew, 1993 b: 241). Indeed, as this scholar claims, `experience has shown (...) that students at all levels make some use of the language presented and hence benefit from the use of the concordancing from which this language is derived´ (Flowerdew, 1993 b: 241).

The application of corpus-driven typical co-texts, applied in the context of text types and genres, may, from this persepctive, be regarded as an effective means of facilitating learners' familiarization with the lexico-grammar to be used. This can be framed as a task-based exercise that `coaches´ learning delivery of specific co-occurrences, processed as linguistic chunks. In this respect, a given adverb may be actively applied in restricted academic domains, such as the utterance of **perfectly** in the clause *it is perfectly possible to*

[316] Another convenient method involves video-taping learners' presentations and having them watch themselves afterwards, encouraging them to notice mistakes and weaknesses in their reports (García and Curado, 2000).

[317] This is also noted in Hewings (1988), cited by Jordan (1997: 205).

followed by a verb, or the production of the discourse word **then** within the construction *can then be used*. These are just two of the cases in which, in opposition to the traditional language learning methods, lexico-grammar should be absorbed from corpus-determined contexts, not worked with according to established rules of discourse or grammar -- e.g. **then** should not be solely taught as a connector that starts a sentence, since there are more cases where it is embedded within the actual clauses in our corpus.

As a result, students should be left the freedom and choice to realize the actual predominant uses of the lexical items; this can be done through their reliance on concordancing procedures. The assignment could then serve as a suitable preparation exercise for their oral project design. Alternatively, as a post-operative drill, their linguistic errors made during the speech may be contrasted with the authentic data from the corpus; this can, in fact, yield rewarding results, as learners are often eager to discover the legitimate way of phrasing concepts and ideas. In addition, they generally favor correcting themselves on their own and not being corrected in class before their peers, their pride being often hurt as a result of such teacher-based discipline.

Some examples of the application of this technique in class can be documented in the learners' coping with collocations and colligations missed during their oral performance. The cluster *we discuss the*, for instance, appearing quite frequently throughout our texts, may often contribute to prevent a common mistake such as **I discuss about* from being thickened. The verb **comment**, in contrast, should be realized with the colligation **on** according to our data; thus, learners may oppose the utterance *the analyst commented on a* followed by a noun, to customary blunders such as **I comment the*.

For the purposes of exploiting speaking skills, in this respect, language anticipation is encouraged in the form of beholding lexical combinations based on academic employment.

The appropriate concordancing application of key items should enable learners to foretell how lexis is to be managed in the process of pivotal oral delivery techniques. A common design facet such as that of introducing a topic in the speech may thus be correctly done if the chosen verbs –e.g. **comment** or **discuss—**are used appropriately in sentences such as *I will comment on the* or *I will discuss the.*

In summary, our corpus-based lexical management may thus assist in the task-based approach to scientific-technical English in academic settings. The next section is a brief account, in turn, of how our collection and inducement of this type of activities may foster the design and undertaking of corpus-based ESP course syllabi.

III. THE LEXICAL SYLLABUS DESIGN

A lexical approach in course design is favored from the perspective of our research. This should be a `pattern-orientated approach [that] reflects the working of language within a certain domain´ (Gutschow, 1978: 60). The ESP instructor is thus `left with the task of organizing his course in a way that facilitates sentence production and enables the learner to take part in an acceptable form of communication´ (ibid.). Pattern-based methodology in the communicative school (Hornby, 1974) is a first main influence from our standpoint. According to this view, `every item must have a role to play in communication, and thus be worth the student spending time and effort on, for receptive command first, and moving to a degree of productive command later´ (Hindmarsh, 1978: 146).

A second major direction is furnished by the target / learning needs analysis (Hutchinson and Waters, 1987), by which product is contrasted with process aims in task schemes. In this sense, Kennedy and Bolitho (1984: 137) hope that `increasingly, satisfactory and realistic rationales for ESP programmes can be worked out, to the benefit of learners and

teachers alike´.

Thirdly, the notions and ideas cultivated by lexically focused proposals (Willis, 1990; Lewis, 1997) are highly relevant. Last, but not least, are the register and genre dimensions of discourse (Biber, 1988; Swales, 1990; Bhatia, 1993; Biber et al., 1998). These equally afford significant guidance as to how lexical knowledge should be arranged and managed for effective processing.

Because the first two routes – communicative and learning paths – have previously been commented on in the present chapter, we focus on describing the relationship implied by the other two – lexical syllabi and genre / register analysis. This is solely viewed in the application of our corpus-based data for framing and building an ESP program; in this syllabus, lexical behavior, being observed in relation to register and genre, dictates choice and gradation of items.

Because not only user needs, but also the whole academic `eco-system´ -- e.g. students, teacher, course planners, etc—should be addressed (Holliday and Cooke, 1982), target language demands are thus constrained by the learning situation and setting. For instance, in our case, working with readings involving actual subjects and topics dealt with in the degree is important, since most of the material is neither translated nor does it have any Spanish equivalency, while this bibliography takes up a prevalent position in Spanish university course syllabi and research. In addition, understanding concepts, ideas, notions, processes and definitions is appointed as a common activity in all courses. Becoming actively acquainted with special word senses in semi-technical items occurring in the texts favors such a task in this respect. By extension, being able to associate these analyzed meanings in the context of the text types and genres facilitates the learning process in the development of ESP competence for the studies in question.

All macro-skills by focusing on specific micro-techniques, as has been examined, can be exploited from this perspective. What occupies our interest in the design of learning priorities and policies from this lexical axis, seems to be the distinction of three main modules or phases: First, the identification of relevant lexical items, arranged according to their importance in genres and text types; second, the application of these linguistic devices on task-based exercises, and, third, keeping information up-to-date regarding the development of our corpus, as topics and subjects encountered, academic and research interests in the area, and academic or professional challenges, may vary with time and availability of resources (Robinson, 1999).[318]

As a result, in agreement with Willis (1997: 2), in the process of syllabus design, `we need to concern ourselves not only with sentence patterns but also with units like lexical phrases (...) particularly common in academic discourse´ and scientific-technical English –e.g. medical research; these are useful `not only for specifialized language but basic communicative competence´ (ibid.). This is founded upon the ESP teaching connection with `the analysis of the genre´ and the creation of a `pedagogic corpus´(ibid.). Language in use is therefore represented `in high frequency words, grammar and research words, forming the context of use´ of our subject areas (ibid: 3). More specifically in our case, these are common core elements such as semi-technical academic and subject items, argumentative, procedural and grammar / discourse vocabulary.

The study of collocations is thus considered the basic underlying framework for our proposed course design. This is mainly due to pedagogic concerns, since, as Gitsaki (1996: 3),

[318] Robinson (1999) explains that in her case, designing SELMOUS (Special English Language Materials for Overseas University Students), she has come to notice over the past four or five years that learners are eager to explore linguistic information related to genres and functions that they read in their content courses. The use of the concordancer is thus positioned as a highly valued resource in this sense. In our own circumstances, something similar may be regarded in the spread use of the Internet, which means that not only can a corpus be framed by teachers alone, but, if fitting objectives accordingly, the students' choice of texts may also be

citing Brown (1974), reveals, we believe that `an increase of the students' knowledge of collocation will result in an improvement of their oral and listening comprehension and their reading speed´. Lexis becomes the central organizing principle of the syllabus in this respect (Lewis, 1993; 1997). The other major criterion, genre or text type, is essential in the fact that it chiefly provides the necessary degree of authenticity for the design of `in-house teaching material´ (Robinson, 1991: 56),[319] and in that it enables the combination of `linguistic, rhetorical and methodological investigations´ -- e.g. textbook excerpts (Swales, 1995: 6).[320]

The following are three samples of how lexical approaches to syllabus implementation may be edified, according to our corpus-driven itemization. First, the relationship between words and the corpus texts is made explicit; second, developing lexical strategies is exemplified in the use of graphics, and third, working with lexical items as rhetorical keys is constituted.

---

considered. In addition, since ESP courses may be also offered to learners coming from different disciplines in some cases, common coreness parameters may be a constant object of inquiry and discussion.

[319] In this respect, in fact, it is worth mentioning the increased production and availability of CALL (Computer Assisted Language Learning) approaches to ESP material from a corpus-based perspective. Flowerdew (1995), for instance, deals with writing skills through CALL courseware, aimed at `making the correct lexical choice to fit a particular collocation´, among other activities (Flowerdew, 1995: 31). During the CALL conference in 1999, in addition, several papers were indeed presented on the use of small corpora available for research and learning, involving authentic e-mail messages, chat-rooms, MOOs, etc. (Krishnamurthy, Corpora list: 23 Febr 2000). Pennington (1996) also provides a comprehensive account of CALL tasks based on computerized linguistic feedback.

[320] The influence of genre analysis can go much further, as has been indicated throughout our study, since the degree of scientificity, for instance, may be analyzed within research articles `in terms of reformulation of rammtical metaphor, discourse signalling and posture´ (Gledhill, 1995: 2).

III.1. WORDS, TEXTS AND OTHER MATERIALS.

Because `itemising and organising a syllabus´ should be based on the `analysis of actual language´ (Willis, 1990: v), priority in lexical selection depends on frequency and range of use. The course organizer's role, in this respect, lies in her / his ability to report and convey corpus-driven findings in an organized way. Then, to guide learners by means of exposure and checking of items in context -–e.g. reading, listening to texts containing forms and through pattern practice (Willis, 1990: 72). Grammar, in addition, should be organized `almost entirely lexically (...) The value of organising things under words is that words are immediately recognisable´ (ibid., 81). In addition, due to the fact that most ESP learners begin university courses with a tendency to recognize the so-called `rules´ of intermediate English grammar, as our classes increasingly show (cf. Cortés, 1991), focusing on the lexical level seems to yield positive results in terms of providing a suitable alternative to the method previously embedded. Linking language and conceptual / thematic rung (e.g. identifying keywords in texts, searching for main topics, etc) tends to enable a broader scope for communicative lexical interaction.

In this way, the main point in this distribution of words according to their use in the texts is that of `viewing language as lexis, rather than a combination of grammar and vocabulary´ (Lewis, 1997: 13). This notion can thus re-direct learners in the shift from the traditionally grammar-based course syllabi: from required processing of grammar rules to `grammaticalised lexis (...) recalled, prefabricated from memory´ (ibid., 102). The aim is to make `principles more accessible and closer to the learners' everyday experience than grammar rules which were often abstract and remote from the experience of using the language´ (ibid., 67).

In this regard, academic lexis may function as applicable items for the determination

of sections and stages in the elaboration of the course general outline. According to the data listed in appendix 4, as a result, a comprehensive plan of an English for Information Science and Technology (EIST) course may be roughly drawn. In fact, given the top 100 words, the basic direction of the topic-based arrangement of the material may be inferred by means of lexis such as **Information / Web / Html / System / Data / Number / Internet / Library / Network / Access / Digital / C / Interface / Object / Document / Page / Available / Images / Application / Language / Design / Message / Computer / Knowledge / Electronic**. These may serve to determine the texts to be dealt with, where the given word most highly co-occurs. Co-occurrences, in turn, can thus be made relevant as notions, included as key forms under each heading. Figure 16 displays an example of a syllabus entry disposed in such a manner.

INFORMATION
SUBTOPICS ***Digital information / information systems / electronic information** ***Information technology / information services / information society** ***Information Science / Library and Information Science /**
NOTIONS ***provide information / preserving the information / searching information** ***information gathering /** ***information about / information available / information on the web**

Figure 16: Example of unit for lexical syllabus design in EIST courses.

The main characteristic in this type of organization is the identification of such items as recurrent in the texts to be read at a certain point. Figure 16 presents linguistic data that typically appear in introductory material to Computer, Information Science, Telecommunication or Audio-visual communication English for our classes – more examples are offered in appendix 8.2. The arrangement is made so as to dispose of the main sub-topics first – in the top portion of the entry--, while encouraging awareness and familiarization with the most frequent collocations in the bottom segment of the box – differentiating among verbal, nominal and other collocations (adjectival, adverb, and colligations) in pre-established rows (# 1 = verbal, # 2 = noun, # 3= rest).

In this way, as learners should advance with the unit, further items may be recorded according to the two listed components in the lexical syllabus division: topics and notions. For instance, emerging constructions such as *deliver information*, *link information*, *managing the information*, or *retrieve the information* may be conveniently labeled as relevant notions, appearing in concepts like **elecronic information** or **information technology**; their degree of use depends on whether the text focuses more or less on such issues, since they occur exclusively in specific environments. The learner is guided in the recording task while reading the material, and yet, he or she is left the freedom to realize the importance of specified utterances – e.g. through concordancing and frequency word list relevance.[321]

In pinpointing academic lexical items for the design of our syllabi, however, we may disregard the convenience of accounting for certain types of words, as was stated in `Word discrimination in common core analysis´ (chapter 2). According to these criteria, as a result, for the inclusion of lexical data in the word list of appendix 4, such words as the names of the months, for instance, are discarded – e.g. **November** and **February** are the two most frequently cited (abbreviated as **Nov** and **Feb**). Expressions having a latin origin are also left

out of such a list for teaching purposes (e.g. **i.e.** [173 occurrences], **ibid** [118 instances], or the 54 examples of derived nouns with the prefix **poly**, such as **polynomial**, **polymorphic** and **polymorphism**).[322] Proper names are likewise not counted (e.g. **Washington** [48 occurrences] or **Paul** [33 times]).

In contrast, in the configuration of syllabus components from this lexical point of view, subject-based items may be incorporated if given the need from both language and content demands interpretation. This is done by checking whether the texts chosen to be exploited in class do, in fact, present the significant development of such an item through their body of discussion. Thus, regardless of their appearance or absence in the frequency and range list (appendix 4), subject or topic-restricted elements may be introduced as objects of inquiry in our ESP course.

For example, the word **fibers** offers a relevant range of combinations in the D1 subject group, regarding audio-visual media technology (see appendix 6). Although, as can be seen in appendix 6, **fibers** does not occur in research articles at all, and only once in textbooks, the reports in which these constructions are employed may also be applicable to other areas, since themes such as the media in education and other institutions tend to cover a wide field of notions and concerns. Consequently, as shown in figure 17, a subject-circumscribed word may supply enough feedback for making up an independent syllabus unit on its own (other examples displayed in appendix 8.2. are **semiotics** [subject E1 – communication theory] and **adapter** [A1, A2 – history of computing and engineering]).

---

[321] See `Academic word lists and reading procedures´ above.

[322] This is done due to the generally little problem offered by these technical items in the Spanish academic setting; instead, anglo-saxon teaching situations seem to favor a closer focus upon these latin-based words (cf. Liu and Nesi, 1999).

FIBERS
SUBTOPICS
***Optical fibers / glass fibers / lasers and optical fibers /**
NOTIONS
***optical fibers could be / optical fibers offered / asserted that fibers /**
fibers can carry / fibers emerge / optical fibers form /
***best fibers / broken optical fibers / existing glass fibers /**
***loss in fibers / such fibers can /**

Figure 17: Example of subject-based item represented as a syllabus unit.

In addition to the texts handled in our courses, a useful means of pondering the selection of items in our syllabi is the attention to specialized lexical sources such as the dictionaries and glossaries analyzed in chapter 3. However, their use should never be considered as separate from our textual data, but rather as a complementary aiding instrument.

Thus, for instance, an important item like **operating**, which may lead to the collection of linguistic information relating to the subtopic of *operating systems* in our courses (see appendix 8.2.), may not be offered as a priority item by some dictionaries (e.g. the *Dictionary of Information Science and Technology* – DIST-- or the *Dictionary of Information Technology* –DIT--, which offers **operation** instead – see appendix 7).

In addition, functioning as a relevant notion in our syllabus section, a highly recurrent form such as *the answer can be found* may be missed due to the fact that the past participle **found** is not registered in the two lexicographical sources mentioned (appendix 7). Finally,

the fact that some words do appear in the dictionaries may not be explicit enough to lead us to their consideration as important elements in the organization of course content. For instance, the item **mail** merely appears as the compound **mailbox** in the DIST, whereas its opportune use in other structures, such as *e-mail address* and *e-mail message* is left out (see appendix 8.2.).[323]

In a third position, following corpus texts and dictionaries, the consideration of ESP, EST and EAP textbooks is equally germane at the time of framing course programming. As with the lexicographic sources above, the application of this kind of material for our design of priorities in assembling language and subject matter, lies in the condition of whether these works satisfy the range of corpus-driven lexical data to be tackled.

Thus, for instance, in this respect, Burgmeier et al. (1991) present the study of academic lexis from a common core perspective that may occasionally parallel our own view. For example, in their development of comprehension exercises for both Computer and Audio-visual topics (history of computing and motion pictures – units 4 and 5), Burgmeier et al. (1991) examine occurrences such as **market** and **enhance**. These may be analyzed as core words in our study (see appendix 8.2.). Another relevant source, in this regard, is Mascull's *Collins-Cobuild Keywords in Science and Technology* (1997), which also deals with a wide range of topics, including Information technology related units where key elements such as **network**, **information highway**, or **internet** are exploited.

For logical reasons of befitting ESP aims from more narrowly defined domains, the consultation of subject-related textbooks such as Hick's *English for Information Systems* (1991), Boeckner and Brown's *Oxford English for Computing* (1993), or Comfort et al's *English for the Telecommunications Industry* (1994), is also desirable. In fact, according to

[323] The fact seems to be that whereas the DIT encompasses a wider range of words in Information technology, the DIST, in turn, is more narrowly focused on technical items referring to specific concepts of Information

such topics as *computer software* or *computer networks* (Boeckner and Brown, 1993), our corpus lexical items may yield positive feedback with phrases like *software support* and *the global computer network*, which are also unfolded in the source cited. In these cases, the content afforded by the teaching material may be assessed positively and equally introduced in the summany of our syllbus arrangement.[324]

Finally, regarding grammar points, our course content is not organized primarily according to these parameters, although grammatical priorities can be drawn from the study of corpora, as the outstanding work by Biber et al., *The Longman Grammar of Spoken and Written English* (1999), proves. In this source, grammatical assertions are driven by register study and effectively made on account of empirical data – e.g. the marking of future time is more commonly done by the expression *we shall* in academic registers, whereas *is going to* tends to be more frequently employed in conversations (Biber et al., 1999: 456). Another type of finding in this respect is that a lexical bundle such as *in order to* characterizes academic discourse more than it does converstional registers, where an instance such as *I don't know* is found more typically (Biber et al., 1999: 994).[325]

Some considerations should be therefore formulated from this perspective. From our viewpoint –in agreement with scholars (e.g. Halliday, 1991; Lewis, 1997, etc)--, grammatical issues should be conformed as a secondary or complementary scope in the domain of lexical analysis. For instance, because the co-occurrence of modality in our corpus is rather elevated, inducing the senses of **can**, **may** or **should** in common clusters and sentences seems to be

---

Science and Technology (see appendix 7).

[324] Yet, this may not always be the norm, since, for instance, of the three textbooks mentioned, Hick (1991) provides textual data that tends to become obsolete in the light of new technological advances and topics. For instance, in its treatment of area networks, even though this topic is widely dealt with in our classes, the texts supplied by Hick refer to abandoned or replaced notions in the subject of networking. In such cases, the material is disregarded for teaching purposes.

appropriate at early stages of the course outline. In addition, our belief is that such elements, always recurring in texts, should be equally developed and reviewed in more advanced classes – the difference mainly being signalled by the level of difficulty in the exercises.

The suitability of text-driven data is thus enhanced by the inclusion of these other teaching resources as design aiding tools in course syllabi. The next sub-division goes further to allow for another influential feature in the observation of content for planning a language for specific purposes agenda: graphical feedback.

III.2. LEXICAL ITEMS AND GRAPHICS.

The majority of students in our ESP courses tend to judge the inclusion of graphical information in the texts as helpful and useful for comprehension purposes. In fact, as they themselves produce a presentation on a given topic, all of the learners favor such a type of strategy. Answering the question posed by Schnotz et al. (1993: 181), in fact, `whether a text with pictures produces a higher learning effect than a text without pictures´, we may say yes from our own personal experience. Successful language learners in our classes tend to `use a graphic more intensively´ (Schnotz et al., 1993: 191).

This reckoning leads to our reflection on course syllabi elaboration that non-linear texts such as graphs, tables, charts or diagrams must be presented, discussed and written about (Heap, 1993). This is done so that productive skills –written or oral—may be exploited accordingly (ibid.). Like traditional communicative language programs in the notional-functional school, the goal is to `give students an understanding of both form and meaning while engaging in meaningful communicative activities´ (Andrews, 1991: 2). This can be performed by learners while they work to clarify meaning in texts and manipulate information

[325] In both cases, expressing future aspects with *we shall* as well as producing lexical clusters such as *in order to* as purpose items are common linguistic devices in our academic corpus too (as the data analyzed in chapter 2

arranged in a certain way in graphics.[326]

The language teacher, elucidating the type of topics and notions to be assembled in the ESP course, should thus convey the necessity to work with graphics and figures by denoting the lexical importance of core words that serve to explain or describe graphical information.[327] Ultimately, as Tomalin indicates (1986: 11), `successful language learning depends on the teacher's effective use of each aid´, and, thus, from the observation of language in use, signposting structures and functional data based on text type words should be recorded for syllabus operation purposes. Some examples are *this can be seen as*, *which we can see*, *this diagram shows*, *figure # shows*, *explicitly defined in*, *clearly defined in*, etc. In each unit of the course, as a result, this type of linguistic feedback should be taken into account and be exploited through appropriate tasks (written and oral).

III.3. COMMON CORE ITEMS AS RHETORICAL WORDS IN THE LEXICAL SYLLABUS.

Closely connected to this notion of text type lexical elements for the management of graphical information in texts, is the assertion that, to answer questions and instructions in written examinations, procedural vocabulary is highly important (e.g. **define**, **explain**, **discuss**, **describe**, **compare**, **list**, **calculate**, **provide**) (Kennedy and Bolitho, 1984: 131). As

---

seems to reveal).

[326] In this respect, reading material distributed via the Internet, for instance, may happen to be quite useful in the sense that a large amount of visual aid is offered, providing an expansion of the learning environment (Phillips, 1985: 116), but equally leading to distress if learning material is not managed adequately (Mathies and Nelson, 1995: 22).

[327] For ensuring cohesiveness in writing tasks, for instance, this is a key procedure (Hill and Chohan, 1978: iii). Denominated as `guided writing´, this activity focuses on the `ability to interpret the significance of data presented visually. The aim here is to show how such information can be interpreted and organised in paragraphs´ (Fowler, 1989: 58). In Pincas (1982: 94-99), a less complex chore is `controlled writing´, where the end of the process is to construct sentences from patterns in paragraphs by using plans and outlines. Another plausible academic task is that of comparing visual and verbal messages (King, 1994: 219) by relying on keywords (most distinctive across texts).

a result, these key devices should be given the importance that they deserve in the evaluation of language data in ESP syllabi. However, as some sources have proved through research, there may be too much of a narrow focus placed on the development of these rhetorical cues in ESP-EST courses, whereas, as Dunkel and Davis (1994: 55-56) state, these elements do `not have a significant influence on the number of information units or the total number of words´. Consequently, from our viewpoint, they should be integrated in the overall framework of the syllabi, but not overemphasized or given a pedagogic prevalence over other types of words (e.g. academic or subject).

We thus reach the end of our discussion of the results from both organizational and teaching scopes. Our concluding remarks to this work are offered in the next section.

CONCLUSIONS

A commmon core perspective in our lexical analysis for the ESP area of Information Science and Technology has been the overt object of study in this dissertation. The focus of our research is thus placed on the observation of extended word behavior in specific domains of English, such as Information-related disciplines covering issues of Computer Science, Information Science (Library Science and Documentation), Telecommunications (Specialization in Optic / Radio communication), and Audio-visual communication. For the elaboration of a lexical framework that serves to establish linguistic priorities of use and learning, the central concern has been to produce significant lexical data that can be adequately arranged. In this process of making informed decisions on the disposal of word forms and lexemes, a corpus-based approach proves rather convenient. The notions derived from Corpus Linguistics become highly relevant, whereas the study of context brings forth the importance of word behavior according to text type and genre. In such a framework, we establish theoretical conclusions in the next sub-section. Furthermore, our assumptions and inferences based on the methodology proposed, the empirical data obtained, projected research, and general contributions, are also described in the following pages.

I. THEORETICAL CONCLUSIONS

The notions afforded by the theoretical background enable conclusions to be drawn regarding the applicability and effectiveness of using corpus-driven data for lexical purposes. Furthermore, due to the relevance that approaches to lexis from register, discourse and genre scopes entail, some claims may be drafted in terms of the novelty and usefulness that these analyses allow in the development of our lexical framework. Thus, we may surmise that even though studies of lexis in context have been ever addressed, it seems that not until narrower observations conceived from empirical language domains have blossomed (e.g. the study of

register dimensions of discourse – Biber et al., 1998), a more detailed examination of word use and carriage has been made available.

In this respect, the type of lexical work carried out in ESP has constantly stressed the position of contextualization and quantitative analysis (e.g. Ewer and Hughes-Davies [1972: 47]). Based on these principles, our research follows the aim of dealing with both content and language from a corpus-driven approach focused on the importance of noting linguistic change according to subject area and style. With the emergence of fast and powerful micro-computers, we, ESP teachers, thus begin to become aware of our possible role as researchers (cf. Johns, 1993). Likewise, because of the resourceful electronic utilities designed for concordancing and statistical analysis of lexis –e.g. Scott's *WordSmith Tools* (1996)--, the end of fulfilling empirical criteria may be met. It is then when large amounts of text and linguistic feedback can be amassed and analyzed accurately for teaching purposes. The language of a given genre, in this respect, may be distinguished and characterized according to the patterns of lexical data gathered and sifted – e.g. `instructional English´ may be differentiated from a more `informational tenor´ in technical reports. The notion of `context of situation´ --i.e. specific academic conditions in a technical environment-- (Firth, 1957) can then be carefully observed and explored in great detail.

This type of study is thus mainly the one which has been followed. It is based on the lexical features of word associations (cf. Sinclair, 1966). Collocations and multi-word terms become key elements in our analysis, since the focus on specialized areas of English involves `schematic vocabulary of a given subject area´ in that this lexis `realizes the field recognized´ (McCarthy, 1990: 62).

Because, in this sense, subject and language cannot be discussed separately in ESP settings (Selinker and Trimble, 1974), a further assumption to be considered with our lexical

focus is the idea of a `discourse community´. In our case, this public is embodied by the learners of the academic environment in technical studies relating to Information Science and Technology disciplines. As students are faced with textual material and information written and presented in English in their majors, both specialist knowledge of the subject matter and language employed must be decoded. For these learners, the functions of description, definition, classification, and so forth, as well as the techniques or patterns of comparison, order of importance, exemplification, etc. (Trimble, 1985), play fundamental roles for the negotiation of meaning in technical settings. From these premises of discourse analysis, the application to our lexical analysis of common core traits involves those factors that formalize text types in given sets of texts: classifications, conclusions, definitions, descriptions, discussions, exemplifications.

Common core lexical elements are also viewed as items denoting `a world of things, symbolically fixed to be observed, measured, reasoned about, brought to order´(Halliday, 1990: 21). These concepts are chiefly represented by subject-based and academic words that establish patterns, arranged and organized in texts. Such a range of sources should be disposed according to the encompassing `sub-register´or `sub-system´ of Information Science and Technology (Sager et al., 1980). In this respect, we may draw the distinction between free combinations of academic words, highly recurrent across disciplines and genres (Howarth, 1996), and restricted collocations, tending to forge subject-bound lexical ties across a more reduced number of texts. Semi-technical or sub-technical items, being most prominent throughout our corpus, systematize the proposal launched in the work with collocational information. These, however, may be broadly framed to include both general academic and procedural words as well as specialized or technical items used extensively across the texts. In this sense, such lexical facts involve multi-word units and clusters, all regarded as meaningful

material for our purposes seeking suggestive patterning in common core scopes.

To sum up these theoretical conclusions gathered in our approach to lexical matter, we may state that our main concern with laying a sound background theory springs from the need to determine how best to carry out the ESP instructor's duty of providing suitable tools for the study and observation of words in context, a process which we deem as pivotal in the design and elaboration of ESP material (e.g. activities, tasks, and course syllabi).

II. METHODOLOGICAL CONCLUSIONS

In the application of the above-mentioned linguistic analysis devices to our lexical focus, a method is induced for constructing a framework of operation with the relevant textual bodies or corpora. As mentioned above, such a procedural task seeks to enfold subject matter and academic content material. The selection of sources is made by judging as representative data those readings to be undertaken in the subject areas, and, in addition, by examining them not only as thematic contexts of situation, but also as stylistic variables in the form of target genres and type samples in shared fields of Information Science and Technology. Academic writings are thus regarded as valuable in so far as the purpose is to enable learners to cope with different styles within distinct types of writing. The generation of awareness about such lexical variation according to both concept and phraseology in our technical settings has therefore become a central concern from this perspective (cf. Lockett, 1999).

In this way, texts are selected with two main criteria in mind: novelty and authenticity. Relevance of the topics covered is a quite distinctive factor in terms of the existing technology and its rapidly evolving nature within Information-related areas. Issues on networking, for instance, develop almost as quickly as we speak in our daily lives. Computer terminology and concepts also fluctuate according to the advances introduced technology-wise. In this sense,

the more updated the text is found to be, the more profitable it is for students to learn from this subject-matter parameter.

In the academic environment which concerns these studies, the property of textual material to offer authentic communication is also valued. This involves levels or brows of texts (cf. Huddlestone, 1971) according to our academic community's needs: Research papers are included to satisfy the higher stage of inquiry into specialized topics (cf. Willis, 1997), whereas textbooks are taken as primary mid-brow texts for undergraduate students. In contrast, technical reports, also mid-brow, are mainly characterized as descriptive material or supplementary information regarding specialization in a given topic. Our corpus design has thus been balanced accordingly to contain representation of the three genres in order to be academically authentic.

Setting up the corpus on which to base our analysis is not, however, a matter of only including sources recommended or demanded in the students' courses, since these are often too specialized within a discipline, while we seek to establish a common core bulk of readings for the four Information Science and Technology areas – Computer Science, Information Science, Telecommunication, and Audio-visual communication. Still, the main goal or function of such textual material is to offer guidance by aiding in the location of common subjects and topics shared by Information Science and Technology studies. For instance, a subject such as `Software engineering´ encompasses both Computer Science and Telecommunication, while `Hypertext systems´ is dealt with in all four disciplines. By comparing university programs –both national and international-- and pinpointing fundamental topics in course syllabi, key criteria for text selection have been considered.

The method is thus based on corpus design as the product where to locate relevant word behavior. In this respect, the World Wide Web supplies a productive environment as

long as sources are sifted according to subject-based and academic principles. Students may make efficient use, if directed or guided appropriately, of some of these corpus resources. How readily accessible these become, is a feature to value, so that ease and availability may be afforded in interactive ESP teaching scenarios. Both language learning and research activities have thus tended to merge and complement each other in our study.

By sifting through the content of a proposed corpus, we may draw conclusions regarding data needed for lexical queries (e.g. distinguishing word functions and types). However, as Mason (1998) asserts, it is impossible to define what corpus representativeness means in a definitive way, thus, for this Corpus Linguistics concept, we surmise that practical applications are the best form for checking the feasibility of our selected material. As a result, our method encompasses sheer observation of word frequency according to Mutual Information (M.I.) parameters, by which word behavior has been measured and regarded as significant. These should be determined in settings such as text type and genre corpora. These bodies of text should therefore be carefully contrived and specified within an ESP concern, that of providing suitable material from which to draw empirical data and with which to work in our classes.

For the goal of developing adequate means of study, we come to the conclusion that computerized tools must aid the process of lexical analysis and regulation. After examining different utilities for managing language features, Scott's *WordSmith Tools* (1996) is found to be a highly effective electronic software tool, since it enables that small and medium-sized corpora be fittingly handled for the exploration of varying word uses according to specified settings of ESP. In this sense, as Tribble (1998) states, using *WordSmith Tools* is quite valid to identify `differences between a genre and a reference corpus´, or among texts with differing themes (Scott, 1997).

In our methodology, small corpora of academic genres (textbooks, reports and research articles) have served as contrast sources for the examination of key linguistic items. Noting how and which certain lexical items convey the notion of purpose across texts in a sub-corpus, for instance, allows feedback to be drawn regarding register variation and uniformity according to academic genres and types. In this respect, some elements, such as the clusters *in order to make* and *in order to maintain*, or *so that they*, are recorded as more characteristic in textbooks and technical reports, in contrast with the items found in research articles, where they do not show up at all. In contrast, other items are more widely marked in research articles, leading us to consider a distinct assumption, for instance, namely that the form *to provide* for signalling purpose is more elaborated in research articles, whereas it is much less often used in the other two genres examined, textbooks and reports. Research articles, in addition, are also distinguished by clauses such as *in order to allow* and *in order to provide* as key constructions, which are not emphasized in textbooks and reports.

Within the same sub-corpus, such as that of the latter mentioned –research articles--, similarity or variation of use can also be pinpointed. For example, different items collocate with the pronoun **we** in research articles according to this context of use. Thus, in introductions, for instance, an element of discourse cohesion to consider in this respect is *in this paper we*, occurring with a high significance, in contrast with a past tense statement like *we found that*, which works as a descriptive feature of the `Discussion´ section as a result.

These postulates on linguistic change and uniformity are enabled by operating with the **Keyword** function in WordSmith Tools. Such a feature of lexical analysis allows for significant vocabulary to appear as relevant due to the computation of the so-called `key-ness score´, which is the result of comparing frequencies and percentages from two corpora, the smaller targeted sample and the larger overall corpus taken as reference work. Both a chi-

square and log-likelihood test are calculated by cross-tabulating the lexical data from the two lists. The resulting figures are key in the sense that over a 25.0 score, according to Scott (1996), the words show a positive degree of `key-ness´ in the sample setting. Below that number, the items are considered negative keywords and thus show insignificant levels as distinctive vocabulary.

Further concluding remarks may be garnered in the methodology employed by observing a parallel pattern of variation / similarity on dealing with modality and passive / active voice in research article sections. In fact, in this case, presuppositions are fostered in terms of the type of lexical collocation with modal verbs; for example, with **should**, which is less highly employed in the active in the `Methods´ sections, but much more often throughout `Results´, clusters such as *should be updated* or *should be displayed*, abundant in `Methods´, contrast, in this sense, with *the machine should be able to* or *should search* (from `Results´). We may then deduce that presenting a research framework as a methodology in journal articles is signalled by such passive elements in the case of `Methods´, whereas active clauses propound a more direct engagement with the data in the `Results´ sections.

Voice and modality can thus be studied in combination with genres or stylistic concerns: Lexical items may be contrasted in the different use of, for example, **must** and **should**; the former activates significant passive constructions, such as *must be made* and *must be considered* in textbooks, and in this same genre, **should**, in contrast, operates more critically in the active voice (e.g. *should make*).

Methodological deductions, in this respect, approximate empirical observations, setting the groundwork for further contrastive work on lexical use throughout the pre-defined corpora. Only through data analysis and measurement can we come to reach a scheme of undertaking that yields meaningful lexical traits for our purposes. Word behavior has thus

been examined and sorted according to its conjunction with textual environments or contexts of use. Had the importance of text types not been discerned by means of realizing that variation occurs within the same genre – e.g. research article introductions and discussions--, argumentative and procedural lexis might not have received core positions in our domain.

Our word categorization is consequently the direct upshot of working with corpus sources from these register, discourse and genre variation accounts. Depending on the range of sources in the corpus where the lexical items are counted, academic words can thus be set up for distinction: either functioning as restricted or free combinations. They are also differentiated as keywords in specified subject-bound domains. Likewise, considering whether lexis is either argumentative or procedural has been expressed as the product of word behavior in the determined context of text types (e.g. classifications, conclusions, discussions, etc.). From this scope, context of situation clearly influences or determines the type of word behavior identified.

Two functions or facilities in the WordSmith concordancer, **Wordlist** and **Keywords** serve, in this respect, to both amass and discard lexical feedback within each of these textual sets. For example, *manage information systems* and *exchange information with* may be treated as academic occurrences, although the first one appears across a lower number of sources (i.e. it is thus restricted) than the second cluster (i.e. a free combination). Alternatively, a proper name such as **Brownrigg** is not counted as part of our common core lexis, since such a word exhibits an unrelevant status, being found as a toponym in only one text. Other examples of such occurrences in single sources include latin-root words and acronyms, although, if occurring frequently and broadly enough, they can be weighed and valued accordingly (e.g. the acronym **T.E.I.**, standing for *Text Encoding Initiative*).

Applying a corpus-based analysis to monitor word association and co-occurrence is

established in order to cope with differences in lexical meaning and use. This involves at least dealing with two or more disciplines where the lexical items are found so that the concept of common coreness is maintained. Words may thus combine in all the genres, or only one –i.e. whether they are obtained as one type of items or another, e.g. academic vs. argumentative vocabulary--, but the key is that a minimum of two different subject areas in Information Science and Technology contain the elements. Such a requirement has been previously defined in the design of our corpus: the sources, distributed according to subjects or topics, are shared by two or more disciplines under study (see appendix 1) to be considered common-core.

Lexical profiles are devised in a comparative manner, after elaborating the results obtained from the contexts of use: environment is thus constantly taken into account (Stubbs, 1997). In this respect, handling and managing lexical items involves `an active system in which new links are perpetually being formed´(Stubbs, 1997: 162). This is very pertinent to English for Information Science and Technology (EIST) courses, indeed, as technology evolves rapidly, inducing the swift modification and introduction of new texts, and consequently, the assessment of new lexical items.

The method for storing lexical entries follows a lexicologic view of word behavior. After the quantitative exploitation of the lexical devices is effected, associations of words are chiefly concerned, under the premise of providing a qualitative perspective of such elements arranged according to frequency and range of exposure. The study is conducted, in this respect, with underlying scholarly work in mind –i.e. Sinclair's and Halliday's (1966) notions for the identification of lexical behavior as a closely related phenomenon to grammar. In addition, a cut-off point is settled to distinguish between those items that co-occur symptomatically across sources and those whose range is low by comparison: A 0.3 score is

applied, meaning that for every ten lines of concordanced text, there must be at least three texts involved so that the co-occurrence be considered a free collocation, whereas below that number, the expression is regarded as restricted. Only when instances of a given combination exceeds 100 lines and appears in 20 sources or more, the proportion is discarded, due to the already noteworthy character of such a construction as a free lexical item.

Four chief parameters are therefore considered in the study of words: 1. collocational strength as measured by both frequency and range, 2. the use of possible lexical combinations as determined by context, 3. the learner's needs in vocabulary acquisition for his / her studies, and 4. the particular description of EIST linguistic needs and demands at our academic institution.

Entries must therefore contain, according to points 1 and 2, not only lexical elements and linguistic feedback, but also contextual information regarding genre, subject or type of discourse sample involved. In the case of statements 3 and 4 above, these lexical items should be directed to learners and their demand of meeting both language and concept in their context of use, as a correspondence with their studies. For instance, the cluster *HTML writing tools* is to be interpreted or decoded as a restricted academic item in Information Science, within the topics of Internet and HTML language, when dealing with such themes in the texts. Both the frequency and significance of co-occurrences in this specialized setting are regarded as the useful tools with which to achieve not only phraseological competence but the communicative aims of all language teaching concerns (cf. Pedersen, 1995; Willis, 1997).

In addition, discourse type has also been determined as a crucial criterion for the examination of some lexical items, e.g. argumentative and procedural words. These are examined in only one genre -- unlike academic words, mostly pinpointed across genres. These are studied in the excerpts or textual samples that represent rhetorical functions (definitions,

descriptions, discussions, etc). Schema theory is particularly relevant in this scope, since the communicative competence demanded in these cases is closely dependent on the lexical feedback that allows for the understanding and formulation of cohesion and coherence within the rhetoric types available. Organization and making sense of the text are two operations achieved by procedural knowledge: not only in the phraseological domain but also in the employment of such lexical mechanisms in `conceptual paragraphs´(Selinker et al., 1976: 284). These tend to concentrate the relevant information, whereas they are included within the particular part of the text to be decoded (e.g. text type) (Aston, 1998).

All lexical entries include reference to the conceptual or stylistic setting, as a result. This is indicated near the lemmata in the upper section of our entries. These lemmas are defined, in our case, as the most frequent word forms in their lexical families – e.g. **performance** among perform, performed, performing, performs (see appendix 4). The three figures at the top of the lexical articles correspond to the three genre sub-corpora in the case of academic words, while the amount given in the subject-based items indicates the frequency of such a node-word in the specified thematic environment alone. In the case of argumentative, procedural and grammar / discourse items, these quantities reflect the frequency of the word in the one genre encompassed.

The standardization of lemmas and the rest of word-forms in that lexeme's lexical family is indicated by the following typology: the lemma position is highlighted in bold, while the other items –word-forms-- are underlined; clusters and collocations are, in turn, presented in italics. In contrast, the sets of words and numbers contained in some entries (e.g. A1, A2...) comprise references to the subject categories where the lemma is found (see appendix 1), since the items, in such a case, are restricted within concept-framed domains, denoting the scopes or subject areas encompassed. Academic lexis lacks such data, and, instead, drives a

contrast between more restricted co-occurrences (within a reduced number of texts) and freer combinations (across a broader range of textual samples), as measured according to the aforementioned cut-off point of 0.3 correspondence between number of lines and sources. In addition, where abbreviations of disciplines and genres are displayed (e.g. Lib RA), this means that the lexical combination where such a label is placed, appears with a frequency lower than three occurrences in a text that belongs to the given discipline and genre (this is done in agreement with the minimal amount for computing clusters in WordSmith, actually three repetitions).

Therefore, in this type of approach to the common core lexical elements shared by at least two disciplines relating to Information Science and Technology, corpus and word categorization are exploited. The desing of this framework is conducted with a reference to both scholarly work and the demands perceived for language learning objectives set up in the studies of Information-related disciplines – Computer Science, Information Science (Library Science and Documentation), Telecommunication (Optic and Radio communication) and Audio-visual communication. Seeking to fulfill such linguistic needs may, in fact, work out as a suitable means for building upon adequate conceptual knowledge and linguistic competence in the subject areas. Textual material is thus arranged according to relevant subject matter and appropriate linguistic register. Lexical variance and diversity is thus sought and disclosed as an activity that follows the adjustment to research queries formulated in both of these respects: subject-matter and specific language scopes.

III. EMPIRICAL CONCLUSIONS

Lexical analysis relies heavily on the aforementioned parameters of common coreness in our EIST approach. This involves the study of collocational devices in differing textual scopes, defined by representative samples in our academic environment, where different word categories are assessed.

III.1. EMPIRICAL CONCLUSIONS REGARDING TEXT-TYPE ITEMS.

The focus on small text type corpora, such as classifications, conclusions, definitions, descriptions, discussions and exemplifications, embodies the approach aiming to rate the significance of lexical items to express argumentative, procedural and discourse / grammar relations among ideas, notions, and developments in the texts. A balanced number of texts, subjects and topics are considered so that these regards are explored and accounted for in a variety of contexts of situation. The main reason being that, in such a manner, word behavior can be checked and agreed upon across a series of sources, categorized according to the disciplines, subjects and genres encompassed. Thus, the purpose is sought by working with 60 excerpts containing classifications, conclusions, definitions, descriptions, discussions and exemplifications, and fitted correspondingly under textbooks, reports and research articles, as well as in each of the subject headings shared by the disciplines (A1, A2, A3, etc – see appendix 1).

From the review of samples, a key generalization is inferred concerning distinctive sets of keywords in the text types of our corpus. There seems to be a tendency for each grouping to manage a certain class of items uniquely. In other words, a section such as the Discussions text type contains a characteristic set of argumentative devices that appear as key items in the arrangement of the content. Something similar occurs in the case of Descriptions. Definitions,

instead, point to a nucleus of characteristic procedural words. Exemplifications, like Conclusions, produce fewer distinctive items.

Examples of this variation in use have to do with significant lexical clustering or multi-word units appearing in the corpus of text types. Thus, a procedural form appearing in Definitions, **marshaling**, for instance, denotes such a functional property with items like *our marshaling code*, *marshaling support for*, and *invented for marshaling purposes*. Its meaning as an attribute is quite technical or restricted, being measured in terms of frequency and range: The node has several occurrences, but these mainly occur in a report on object-oriented design (within the subject category of `Computer engineering´ [A2]).

By pinpointing how these key elements are employed in a reduced corpus such as our text type samples (see Corpus References), we may glean their specialized sense as closely combining with sub-technical uses. Another example, the noun **tag**, is viewed as argumentative in Discussions, but its contexts reveal different degrees of technicality: *heading # tag*, *use the paragraph tag*, and *this tag indicates*. This study thus provides a rich perception of lexis in reduced contexts, and it introduces the importance of considering varying levels of common coreness in our research. Such a perspective largely depends on the relationship between word frequency, range and textual sample coverage.

We come to determine that in such a type of background, the more widely distributed text types are across genres and subjects, the more extended such common core items are considered. For example, a highly co-occurring colligation such as **by**, examined in the context of research articles, attests to the fact that different senses and functions can be derived from observing the use of this grammar word in the representative samples of text: from expressions in passive statements (e.g. *increasingly replaced by*), through those signalling instrumentation or means (e.g. *by describing*), to the ones indicating location (e.g.

by author, title or subject). These are critical uses in our corpora, coinciding, in fact, with those perceived in a large textual collection like the LOB corpus, in its J category of `Learned and Scientific English´ (Johansson, 1978). Function words such as **by**, **between**, **rather**, **then**, **your** and **thus** are, indeed, recorded by the authority (Johansson, 1985: 34). In fact, in the case of the preposition **by**, the one-million words division J of the LOB Corpus registers it as the first grammar word in terms of frequency, exactly as befalls in our text type samples.

Furthermore, in some cases, pointing out in which rhetorical environment the constructions were identified (e.g. conclusions, definitions and classifications for the three examples of **by** presented above), may become useful in terms of the display of contextual information to interpret the utterance. Thus, in the instance *by author, title or subject*, knowing that this item is found in Classifications can be helpful, as it refers to the manner by which material is arranged in a library. In this respect, the article from which the excerpt was taken belongs to the F5 subject heading (`Electronic publishing´), pertaining to all four disciplines involved. These factors contribute to driving at the estimation of such lexical elements as relevant and productive for our common core purposes in the areas of Information Science and Technology due to their key use according to certain settings.

However, importance is equally given to assumptions aforementioned, that the most representative cases in terms of common coreness need not be the most frequent in absolute terms, and that items covered significantly in one sample of texts, may show varying degrees of employment. For example, the mentioned highly co-occurring preposition **by** is checked in research articles, as this setting enables suitable lexical feedback to be drawn. In a language learning environment related to science and technology, this appropriateness may be tested through teaching in class, and, also, by means of comparisons with other experiences and empirical approaches to specialized corpora. Due to its comprehensiveness and relevance,

Johansson's account (1978, 1985) of the LOB J category of technical writings may provide, in fact, an optimal reference, as in that work, the preposition **by** is taken to be of fundamental use, signalling instrumentation and method, two functions confirmed in our corpus.

Variation may thus prevail in the contrastive data that has been exhibited as a means of prompting awareness of deviation according to genre. An example is the contrast of the argumentative node **use** between Classifications in textbooks and in the other genres. The employment of compounds such as *acceptable use*, seems to be favored in the case of the first genre, whereas technical reports and research articles tend to present longer forms (e.g. *the use of*) according to our data.

In this way, in our goal to bring forth lexis that is relevant to common core purposes in English for Information Science and Technology, the employment of items across text types has chiefly been realized as semi-technical vocabulary. This means that they serve to express key notions and procedures, pinpointing distinctive traits in relation to context. Yet, a greater notch of specificity is not discarded, as in the case of bound lexical items –e.g. **marshaling** and **tag** above. They should be included to highlight the need of encompassing both a more restricted –i.e. technical—and less circumscribed lexical behavior –i.e. semi-technical.

Procedural and argumentative signalling thus becomes key and distinctive in scientific-technical discourse: verbs like **indicate**, **add**, and **include** prove this by means of their observed behavior in the samples (appendix 2). In addition, there exists a high `degree of "nouniness"´ in these technical texts in comparison with, for example, fictional writing, as Johansson (1985) asserts. Nouns tend to be more frequent than verbs in these divisions – e.g. **definition**, **description**, **procedure**, **comparison**, **addition**, **document**, **content** vs. **define**, **describe**, **proceed**, **compare**, **add**, **document**, **contain**, which have lower frequencies and ranges of use.

Variation in text-type lexical employment is consequently checked by observing how such elements operate across predetermined contexts of use (i.e. devised small corpora). A useful starting procedure is that of pinpointing which text-type words are exclusively key in their settings. For instance, the noun **comments** in research articles, or **chapter** in textbooks. Further development of these traits includes identifying distinctive clusters within a specified environment alone; for example, the adjective **other** is characteristic in **on the other hand**, a construction signalling contrast, commonly found in our classifications, definitions and exemplifications of research articles. This analytical procedure proves to work out as a useful means of testing the meaningful behavior of certain forms in one restricted condition and not in others. In this scope, an expression co-occurring with the particle **how**, for example, within the cluster *how does the text*, provides insight into the manner of concluding written work, since asking direct questions with **how** in textbooks is found in conclusion statements, whereas indirect questions with this word are formulated in the concluding remarks of research articles and reports, according to the contrasted results. Likewise, within another narrow domain like conclusions in research articles, the keyword **reasoning** serves to summarize a main development or notion in a study, as the cluster *order of magnitude reasoning* reveals in a Computer science article.

In addition to our claims regarding variation of use in text type samples, the arrangement of the lexical information equally shows the recognition of another parameter: Semantic distinction among vocabulary items; for example, between procedural and argumentative values. In this regard, text type words can be viewed as linking devices that offer different degrees of signposting. This postulate is both surmised as a result of the empirical data analyzed as well as the encompassed scholarship (e.g. Francis [1986]). An example is that a noun such as **document** is recorded as argumentative because its sampled

co-texts present such an item in functions of reference, or substitution. These are mainly characteristic of argumentative clusters (e.g. *the HTML document* refers to a frequently referred item through the text). Procedural words, instead, tend to be more clearly recognized as markers of conjunction and broader repetition across texts (an example is the patterning lexis of the cluster *Example #* in exemplification texts of research articles).

Another type of differentiation can be drawn between grammar words and discourse items. In this case, colligations exert a critical function, presenting key lexical combinations in this distinction. In other words, function or grammar elements correspond to main structures or points of grammar arising in the texts, as in the case of the mentioned preposition **by** followed by gerunds, indicating the often encountered function of denoting method. Discourse lexis, in contrast, is mainly represented by prepositions, conjuntions and adverbs functioning as discourse micro- and macro-connectors in key sections of texts. The conventional **however**, **therefore**, **in other words**, **in addition**, etc appear in our corpus, and yet, other clusters are assessed as highly important and co-occurring frequently enough --e.g. *as we shall see*, *as described above*, *as described below*, etc—which act as signalling statements in the transitional process of discourse development.

III.2. EMPIRICAL CONCLUSIONS REGARDING GENRE-BASED LEXIS

The main proposition drawn in the inspection of the data gathered in all three academic genres is that common core academic words reflect fluctuating planes of behavior according to both the number of instances and texts encompassed. The items are investigated in the overall corpus, embracing all disciplines and subjects (appendix 4), and thus, a minimum number of ten sources is pondered and set as the yardstick to decide the inclusion of the words in such a relation of genre-based items. Such a decision follows the purpose of

establishing both significance of frequency and range as essential common core factors in the sifting of academic lexical items. Alternatively, although excluded from appendix 4, some elements with low rates of appearance, even the so-called `hapax legomena´, with only one occurrence, can produce lexical feedback worth noting as common core academic expressions; their sole condition being, in terms of count, that they occur in all three genres.

A contrastive manner of checking that items in our list do, in fact, incorporate this sought after common core academic notch is, like in the case of the LOB corpus above, that of bringing other academic word lists into perspective, such as Coxhead's (1998) relation of items that significantly recur across a large body of different academic texts and disciplines. Other recent lexicographic material –e.g. scientific-technical dictionaries— can also confirm the validity of our results (see Dictionaries, Corpora and Other Lexicographic Material in the Bibliography).

Some examples originating from such a contrastive analysis with one source (Coxhead, 1998) include common core elements that occupy high positions in our work --e.g. **data** (# 36), **access** (#60), **design** (#92), **file** (#108), **environment** (#183), **approach** (# 203), **features** (#207), **display** (#335), **identified** (#431), **equipment** (#478), etc. They all belong in both Coxhead's work and our Information Science and Technology corpus as common core head-words in terms of frequency and range.

The key notion is that the more widespread across genres words such as these are found to be, the more productive the items tend to be in the academic task-wise learning process. In that sense, lists of academic words function to provide common core lexis across a given number of disciplines (appendix 4). For pedagogical purposes, these are rather productive semi-restricted combinations making up the general academic discourse of information-related areas. The use of such a type of clusters is a main discriminating factor for

the identification of vocabulary registers, which are analyzed as descriptive linguistic sets making up the target academic environment.

In the collected data, we surmise, however, as a second main principle, that these items do, in fact, mainly function as semi-technical multi-word units, viewed as either restricted collocates or free combinations. We notice that several function or grammar constructions appear as less bound to particular text samples (indicated under the symbol ₣ in the lexical boxes), while restricted multi-word units occupy different positions in the corpus (less frequent across texts).

Nonetheless, in various cases, the generality is not what has been mentioned above, that the free associations are characterized by their colligational level. Indeed, lexical evidence for this exception in use is given by observing a cluster occurring as either restricted or free only by changing from the singular to the plural form. Thus, *amount of data*, for example, is labelled as free, whereas *amounts of data* is considered restricted. This is exclusively due to the location of the first item in a broader context: two Telecommunication articles, one Audio-visual article and one textbook on perspectives on Information. *Amounts of data*, in contrast, is only utilized in a Telecommunication report on Fiber Optics technology (D1). As a result, we can expect to have similar lexical patterns in either the restricted or free categories, depending on their level of significance of use in the corpora.

The ratio to measure this degree of co-occurrence and distinguishing between restriction and freedom of employment in our corpus is set at a minimum of 0.3% (every ten instances, at least three texts must be involved). Such a requirement is fulfilled in all cases of academic lexis observation. This is done so that empirical criteria as well as contextual information offer a way to classify the item as either more or less restricted in its use.

The items are analyzed as pivotal in their academic setting. How narrowly determining

they are can be chiefly realized, in fact, by scanning the frequencies in different corpora. For instance, in only one genre –e.g. reports—vs. the overall corpus. The noun **protection** is thus quite distinctive in this set of texts, since its count is nearly the same in RPs (81 occurrences) as that of the general corpus (87 instances). This turns out to be a useful approach in measuring keywords significantly employed in one setting but, in contrast, less characteristic according to other genre samples. From such a viewpoint, an entry can be described contextually by checking that a node, combining with significant elements in a genre – e.g. textbooks-- does not co-occur with these items in the other two -- reports and articles.

The need to define lexical items as keywords in one genre is especially felt in the pedagogic plane of production skills. Characteristic employment in the genre triggers a focused attention to genre-based lexical use. For example, the form *the following requirement*, is commonly found in our textbooks, exposing instructions on a given issue. For research articles, instead, some noteworthy elements functioning as key are *project milestones*, *project deadlines, the project progressed* and any other clusters including reliable informants on the notion of project development and research. Such elements tend to conform contextual values in approaching genre conventions in the ESP class.

Concerning our number of academic lexical items produced more widely across genres –genre-based items --, as in the case of the J category in the LOB corpus, verbal elements are recorded in all forms and tenses, among which past and –ing verbs are most frequently and broadly registered. Adverbial forms are, in turn, compiled in such a common core list according to their function as collocations with content words, not grammar items. Thus, forms such as **therefore** and **on the contrary**, for instance, appear in text type samples, marking cohesion features in the texts. In contrast, in the broader context of the genres, the expressions *generally accepted*, *approximately the same*, and *much more effectively* are

garnered as key adverbial devices, having the main trait of collocational instrumentation. Adjectives occupy an equally significant position as widely employed collocates in our collection of texts, keeping company with nouns in different contexts. Their importance as genre-based academic items is also measured in terms of frequency and range.

The examples demonstrate that keeping a lexical distribution is the chief factor, even more than frequency, to consider in the analysis of common core units.

III.3. EMPIRICAL CONCLUSIONS REGARDING SUBJECT-BASED LEXIS.

A central assumption formed as a result of exploiting the division of subject items in specific textual settings such as those formed by fields and conceptual rungs, is that the vast majority of these keywords are nouns. This agrees, in fact, with other lexical analyses results, such as Scott (1997), or specialized dictionaries on Information Technology, such as *the Dictionary of Information Technology* (1997) (appendix 7). The main difference, however, is that for our common core analysis, we are chiefly concerned with the lexical patterns of these words according to segmented contexts. The same word can thus function as a subject entry, but may also appear as either academic or text type, depending on its range of collocations and clusters. For example, the noun **library** can be examined in the subject sub-group where it is key, yielding thematically bound collocations, such as *American Library Association* and *public library community.* The important factor to keep in mind is that these elements represent syntagmatic data in given textual stretches encompassing a conceptual domain shared by two or more disciplines in our Information Science and Technology corpus.

These subject-based nouns expose, in Scott's words (1997), the `aboutness´ of the sections. For instance, in the case of Computer and Telecommunications subjects like Software engineering or the History of Computers, computing terms tend to prevail, as these

are mainly about programming, hardware analysis, computer engineering, etc. The fact that a key keyword approach can be conducted by means of the **Keyword** function in WordSmith Tools guarantees this perception, evidencing by means of lexical data that substantial technical background is shared between these two disciplines – Computer Science and Telecommunications-- thus dramatically marked in relation to the rest of the subject areas.

Unlike academic or genre-based words above, these subject items are thus pivotal within thematically bound sets of texts. Their use is thus somewhat more restricted. In our entries (appendix 6), designed by focusing on subject and topic reliance, subject words receive a restricted denotation, since they are viewed in reduced sub-corpora, drawn according to the subject or subjects where items are analyzed.

Subject lexical traits obviously configure a tighter specification based on subject domains. However, this premise has not been intended as an indicator of the presence of these words as precise defining devices for giving detailed and exhaustive views of the topics and concepts encountered in Information Science and Technology. Quite the opposite, our conception has been forged under the presupposition that these lists are open and admit or reject current lexical members depending on the changing nature of the themes and notions in our specialized areas. A cardinal facet of this lexis, therefore, is that sources dealing with specific subjects or topics tend to underline use that appears as less relevant when dealing with a broader scope across the entire corpus.

The data suggests that, on dealing with academic and subject words, a clear-cut segmentation may not be so feasible at times, since, in terms of the type of text topics encompassed when dealing with some subjects or issues –e.g. Internet in the F category--, we may come to regard, in some instances of use, restricted academic lexical items as parallel to subject-based examples. Simultaneouly, some lexical elements in subject entries may be also

wielded in a different category, which makes them less restricted in the subject entry described. In this respect, textual settings may overlap in terms of shared collocational content, treated as either subject, academic or both in some cases. However, our chore has emphasized the need to interact further with these circumstances in order to pinpoint lexical variation, delimiting and clarifying the use of common core lexis across and within subject domains. Regardless of such a taxonomical concern, moreover, for our linguistic exploitation purposes, task work that challenges learners to satisfactorily profit from lexical knowledge in subject and academic textual settings prove quite helpful in the sense that it prepares them for achieving lexical competence.

Academic and subject words have also been compared in other sources, such as specialized dictionaries covering distinct areas of Information Science and Technology. The type of work conducted in this respect has stressed the importance of integrating contrastive outlooks by revising that critical and weighty lexicographic works contain key lexical items being handled and exploited in our study. The words searched tend to be verbs and nouns that occur as lemmas in either academic or subject-based contexts. The sources yielding the most productive results, in this sense, are *the Dictionary of Information Technology* (1997), *the Dictionary of Computing* (1999), the *Diccionario comentado de terminología informática* (1996), *Vocabulario para los estudios de biblio-documentación* (1999), *the TERMITE database* (1999), *the Dictionary of Multimedia* (1997), and *the Dictionary of Internet and Personal Computing*. These are set off as models for comparison in each one of the discipline sub-corpora devised within our overall Information Science and Technology corpus: Computer Science, Information Science, Telecommunication, Audio-visual communication and Internet. Needless to say, our texts, not intended as comprehensive or fully descriptive of the lexical material in each division, are reviewed against these authoritative sources. In this

sense, subject-based combinations can be identified in most cases –especially in works such as *the TERMITE database*, the *Vocabulario para los estudios de biblio-documentación*, the *Diccionario comentado de terminología informática*, or *the Dictionary of Internet and Personal Computing*. In addition, verbal, noun and adjective constructions belonging to broader academic use have also been singled out, as in *Vocabulario para los estudios de biblio-documentación*. Generally, we feel that the focus on both overall corpus and single discipline texts for the contrastive view with specialized dictionaries enables an enriching experience for both lexical arrangement and teaching purposes, on account of the addressed need to deal with authentic and updated language.

In the contrast with lexicographic resources, such as those mentioned above, the main implication is that there are core lexical collocations which are central to this study in terms of teaching usefulness, as a result. These are very often verbal + noun combinations that provide significant feedback as key academic phraseology. An example is **convey** + **information** (*Vocabulario para los estudios de biblio-documentación*), or **computer** + **aided** (*Diccionario comentado de terminología informática*). In addition, we can also deem the importance in our corpus, given the contrastive data from these sources, that some colligational devices trigger productive collocations, as is the case of the preposition **to** followed by forms of **provide**, indicating the presence of purpose. Another observed pattern is that the verbal cluster **provide** + **access to** is quite common. This, in our Information Science sub-corpus, points to the semantic marker predictable with this type of clause (+ **information** – e.g. *provide access to electronic repositories*). This is a characteristic employment in such a corpus of discipline-based textual restriction.

In other cases, for example in the Telecommunications collection of texts, as signalled by the TERMITE database of Telecommunications terminology, the tendency is for words

with a higher level of technicality to prevail. Yet, on surveying such elements in a longer co-textual stretch, meaning generally becomes clearer, and specialized terminology turns semi-technical. For instance, *space object* can be regarded this way in the more widely expanded noun compound *space object tracking radars.* As Pedersen (1995) realizes, in technical English, this type of word increase is usually rather convenient for the clarification of meaning. Equally relevant is the contextual information provided for the Internet corpus-driven nominal associations, where the concept is usually defined or described by means of thematic and rhematic moves. A notion of this sort is *Standard Generalized Markup Language*, explained in the lines: *…known as the Standard Generalized Markup Language (SGML). SGML is an international standard for…* as an example of using lexical cohesion on the definition of an Internet concept. Another case is given by the collocation *network address*, which is identified within the following co-textual setting: *…must have a unique identifier. This identifier is known as the campus IP network address*.

In contrast, on examining other contexts of use –e.g. Audio-visual communication technology texts in our sub-corpus, in comparison with one lexicographic source, the *Dictionary of Multimedia*--, the chief conclusion is that the concepts alluded to by the collocations tend to be network and internet tools for managing information in image and sound formats. For instance, *digital camera* or *media server* occur frequently and widely enough to be noticeable.

The arrangement and design of subject-based and discipline-based common core items in English for Information Science and Technology is consequently closely associated with the prevailing position of verb and noun clusters –e.g. the word **handle** most likely to occur as a noun instead of verb in Computer science and Telecommunications reports, as in the example *the handle server system*. This is regarded as a chief element of reference in the

organization of our course syllabi, where stress on noun compounds is evidently placed in the case of certain contexts of use. In this respect, to achieve necessary competence for micro-skill elaboration in tasks, recognizing and utilizing these key lexical items of subject area and topic development are shown as crucial points.

III.4. CONCLUSIONS ON COLLOCATIONAL DATA DISTRIBUTION.

As mentioned above, to make students aware of and ripe for commanding patterns of language in specific environments is our chief concern in the arrangement and distribution of our collocations and clusters. For such a goal, further analysis is made regarding the employment of such data in context. It should thus prepare learners to use items in correct settings and expectations of utilization. In this sense, during the process of acquiring procedural knowledge, for instance, lexis is conformed to academic aims – e.g. expressing typical definition or classification features accordingly. Differentiating among functional templates such as procedural vs. argumentative devices, explained above, is also important as a result, since when referring to a concept, for instance, key argumentative lexis should be conveniently exploited in clusters –e.g. *the HTML document* replacing a specific object or entity.

Further evidence for our claim regarding the objective of adequately sifting the lexical data is the specification of discourse and grammar words, and how these items are employed to highlight functions of common coreness in our ESP classses. Hence, grammar items, being quite numerous and occurring with a high frequency, should be especially valued in key collocations occurring widely, e.g. the preposition **about** in *information about*. Parallelly, discourse markers like *as we shall see* are assessed as noteworthy in their utilization as connectors in important tasks, such as those that challenge the student's ability to interpret

cohesion by means of signposting in lectures.

Given this chief statement drafted in terms of common core language teaching postulates, and drawn as a result of the data gathered and described, uttered as the recognition of lexical patterns that roughly correspond to pedagogic strategies in ESP, namely via a task-based approach, collocational matter is disposed according to level of significance. Verbal clusters are identified as primary elements for the broadly framed field of the academic exploitation intended. They are classified according to both their observed behavior in our corpus and to the fulfillment of such collocational functions as filed and explained in the *BBI Dictionary of English Word Combinations* (1997) – e.g. denoting, among several other senses, activation, such as in the cluster *launched initiatives* or *prompted researchers to*. Employment of these constructions is in their correct contextualization, neither overrating their applicability nor overlooking them in productive exercises requiring their motioning.

As common a pattern, examined through the texts, is that represented by noun compounds, abundant in all technical settings, especially subject-restricted corpora. In contrast with verbal combinations in the previous stage, these tend to be more highly employed in restricted domains, such as in texts referring to a common topic. For instance, the item *document management system* may be decoded and analyzed by working with its co-texts, where the concept is defined and illustrated. In a similar fashion, adjectival collocations are outlined as restrictive formations demanding a more focused attention to the manner of associating with other words. An example is **effective**, coming typically after the noun **cost**, or **marked**, commonly preceding the noun **section** in our sources. An additional division is made up by those adjectival expressions that are fostered by institutionalized manifestations of the sort of *electronic mail* or *optical disk*.

The lexical information gathered, as a result, should not be viewed as static or fixed,

nor should it be left unattended in regard to the updating of textual feedback. For example, due to the demands of space of our dissertation, many other lexical items in entries are derived from our corpus that are not included in this research. The aim in such a type of lexical analysis thus hinges upon the notion that data is ever modifiable and open to additional input. Nonetheless, the establishment of such textual and lexical reference in our research serves as useful guidance for our teaching goals.

IV. PROJECTION OF RESEARCH

As the basic aim for our research is the undertaking of lexical common core parameters in four related areas of Information Science and Technology for dealing with ESP concerns, much effort is placed on defining and curbing significant data in terms of importance in both texts (language) and subject matter (content), two basic supports in learning languages for specific purposes. In the final sections of our work, in fact, postulates are actually raised concerning the intended nature of teaching that working with such empirical information infers. Describing linguistic features does not necessarily mean that these find direct applicability on pedagogy (Hutchinson and Waters, 1987); yet, if an equilibrium is brought forth in which exploitability of language resources follows both our common core proposal and learners' demands and needs in the target situation, suitable pedagogic objectives can be met. We have come to conclude that a becoming approach in this respect is to be favored by the conjunction of lexical performance and academic tasks. The procedure tends to combine, at different planes, both courses of operation in learning linguistic knowledge as product as well as process.

An example is offered by the use of academic word lists such as Coxhead's (1998), or our own (appendix 4) to trigger the learner's awareness of word arrangement and activation.

By realizing certain common core lexical families (e.g. **define**, **call**, etc) and fields (**organize** + **lessons**, **knowledge**, etc) likely to point to the information and semantic values sought, learners can, indeed, manage an effective scanning exercise that requires them to establish statistical and conceptual bonds. Registering verbal constructions in order to assimilate their employment, in this manner, can be done during reading procedures. The purpose is the realization of collocational input significance, as in the example **provide** + **access**, which can be classified as a quite recurrent verbal cluster, belonging to primary sets reflected in our corpus as well as other lexicographic material (e.g. *the BBI Dictionary*). The emphasis is placed on the need to develop vocabulary sets in a systematic way, in which word association in context is triggered, whereby students infer and induce meaning in the readings.

In a parallel manner, noun compounds serve to express technical developments in the subject areas which can result in positive lexical feedback for the process of developing writing skills. The performance of relationships in the form of mental associations based on word behavior makes up another type of task-driven strategy to coping with subject-based items. These, viewed in a large number of cases as multi-word terms spelling out concepts and processes in Information Science and Technology, can be framed within their defining context in order to enact lexical traits that may be subsequently handled for productive goals. Noun compounds can therefore become more intelligible and explicit by means of unfolding their contextual nominal association.

A sequential range of exercises (appendix 8.1) has thus been put forward through which learners may come to check their competence with lexical bond mobilization. The aim is to acquire academic writing micro-skills by means of lexical competence to define, describe or classify concepts and issues in Information Science and Technology. This is done through guided writing exercises, where the construction of nominal compounds in context is

adequately directed and encouraged, e.g. in mind maps, where semantic associations are formed, or by means of organizational charts, which compile lexical devices according to contextual information on a given technical item, such as *the BASIS Techlib application* in the topic of Perspectives on Information.

The issue is to explore these keywords within appropriate clusters, as is the case of *document management system* in syntagmatic relations that explain the function of the compound –e.g. *image-based document management system encodes information*.

In terms of coping with adjectival collocations, writing techniques can also benefit from appropriately utilizing such elements in exercises. For instance, to find ways of describing issues specifically, devising proper combinations between adjectives and nouns is viewed as a useful chore in order to produce forms which occur substantially –e.g. *cost-effective* before **manner** and **method**. In this respect, we deduce that the more restrictive an adjective + noun combination is shown to be, the more productive the itemization tends to be for learning purposes, since such specific expressions concentrate both conceptual and stylistic intake in a relevant manner. From this idiomatic scope, the process of institutionalization endured by some adjectival collocations in our corpus can serve as manifest evidence –e.g. *electronic mail* or *electronic bulletin*. As they are more broadly referred to, these Information-related forms become more familiar to the layman.

Academic goals in our classes are thus determined by both written reception and production skills in terms of constructive work with academic lexical sets and lists as well as the establishment of lexical bonds. In addition, oral tasks are conceived by means of lexical sifting to anticipate listening input and encode speaking output.

Note-taking techniques, for example, are profitable as academic tasks in our setting. They involve listening for the right lexical item, such as micro- and macro- markers of

discourse segmentation –e.g. noticing a summarization device in the use of *in summary*, or *to conclude this*, in concluding remarks. Differentiating among these key elements to understand speech transitional input carries meaningful teaching implications, since it is quite convenient to perform appropriate operations in context, such as to identify main issues and discriminate secondary ideas, locate an exemplification offered, etc. Visual aids should also be managed adequately in the process of the lecture reception in order to enable the correlation between the key lexis employed and the concepts described (i.e. semantic associations).

We deduce that knowledge of semi-technical and technical word behavior entails the capacity to function effectively in decoding and encoding signalling mechanisms throughout academic discourse. These features involve macro- and micro-markers that range from conventional adverbs to interactive clauses evoking a greater degree of involvement between speaker and listener. To materialize the corroboration of correctly interpreting the message conveyed, a series of activities is proposed, in which the central purpose is to bind linguistic expression and conceptual rung by means of distinguishing relevant from irrelevant data.

In addition, we find that the task of giving a short oral presentation before an audience is greatly benefited from the study of collocational itemization. For instance, to cope with subject-based elements that explain a given issue and display its possible solution stage is more successfully formulated by means of patterns that arise in the communicative context. For example, the structures *we cope with* + ideas or developments may serve as a suitable strategy to initiate a topic. These prepare the stage for learners as a working tool for the delimitation of word use, resulting in a more successful mode of learning semi-technical structures. This may be achieved through both concordancing data and video recordings of lectures in which students are made aware of trends and forms of uttering notions as well as of making common mistakes.

The proposed pattern-based methodology follows the aim of developing receptive command first, and then moving to a higher degree of productive exploitation (cf. Hindmarsh, 1978). Furthermore, target and learning needs must be taken into account to determine how relevant our collocations are for the ESP class (cf. Hutchinson and Waters, 1987). For the design of course syllabi, in this respect, such detected lexical behavior plays a fundamental role. This is mainly carried out according to three premises: identifying relevant lexical items in subjects and genres, applying them on task-based exercises, and keeping this information updated by means of corpora revision.

The corpus thus turns pedagogic in the sense that meaningful lexical items operating in the texts are viewed as collocational bricks for laying the foundations of course content. Genre and text type, in addition to subject-based samples, provide the necessary degree of authenticity for the design and construction of in-house material. Common core principles give insight into the kind of data to include in such ESP curricula. The process of working with language for programming the material is thus intimately linked with the scope of the themes and subjects involved. The wider such a coverage on the texts is measured to be, the higher the level of pedagogical implications is felt to be in correspondence with our gradation of lessons and scheduling of class content. In this respect, chunks of lexical feedback are employed to organize the units, presented consequently according to topics and sub-topics, as pinpointed by the verbal, nominal or any other clusters encompassed. The notion of lexico-grammar becomes crucial as a result, since items are distributed and prioritized regardless of their form or class, and chiefly according to their significant use and appearance in demarcated academic textual settings (i.e. genres). In such a modus operandi, an adjective (**operating**), for instance, may be treated as the core item for introducing a lesson (as it co-occurs with **system**, a relevant common core concept in Information Science and Technology)

(see appendix 8.2).

In addition to this application of corpus-based text knowledge for the structuring of lexical syllabi in our area of ESP, the inclusion of items derived from the use of graphics and rhetorical functions throughout the course should also be noted. These are the words that learners may rely on for solving key procedural chores when dealing with the communicative ability for conveying subject matter.

Lexical items consequently emerge as both product and process statements of the learning experience in our daily struggles and joys stemming from work with language for specific purposes. Research thus evolves as a need for defining and applying measurements of specialized language in relation to its context, whereas, as a reciprocal and concomittant sphere, teaching receives a gained insight if such investigations are performed effectively (i.e. developed according to learning demands).

V. CONTRIBUTIONS IN THE GENERAL FRAMEWORK OF ESP

The unfolding of queries and remarks concerning the status of our lexical common core for different disciplines relating to Information Science and Technology, has been addressed in order to answer both linguistic and pedagogic issues. Defining features and functions of common core words, according to concepts of Corpus studies, has been developed through the discussion of examples and scholarly points. In turn, the observation of teaching implications, made by means of sample exercises and the revision of learning approaches, has led us to consider how our research can be projected for practical purposes.

In conclusion, the analysis of common core lexical items thus presented and segmented in our body of texts are deemed as serving or contributing to a main concern: that of determining valid working tools for delivering reasoned and structured course programs in

ESP settings, where language, style and subject area are intimately related. The data and material thus collected, being the product of two consecutive years of research and teaching, should constantly be extended and renewed during the process of our work in the rapidly evolving sphere of Information Science and Technology. The features of lexical collocation and typicality (i.e. + or – idiomaticity) across defined sets of texts, and factors such as variation, distinctiveness and keyness in discourse and genres thus function as important measurement keys for the study of significant lexis in ESP contexts. These should be conceived in agreement with contextual information in order to undergo the sign of the times, in which English is the medium for communication in the Information world, and where Information-related lexical items are massively produced to convey concepts, issues, topics, subjects and developments of Information Science and Technology. [328]

[328] Among many other texts dealing with Internet and the Information technologies in the press, two articles in the newspaper *El Mundo* (18th November 1999 and 6th December 1999), are quite significant in this respect, since they depict how both the educational environment and workplace are increasingly influenced by the use of the web and information networks (Castro, Friedman, Gonzalo). Quality information is thus sought where `the best educational installations and Internet services are provided´ (Friedman, 1999).

APPENDICES

Appendix 1: Subject categories in Information-related disciplines

A = COMPUTER SCIENCE and TELECOMMUNICATIONS

A1 = History of computers, Hardware, Software

A2 = Computer engineering and architecture, Data communications and Client-server architecture

B= INFORMATION SCIENCE, COMPUTER SCIENCE and TELECOMMUNICATIONS

B1 = Information units management

B2 = Online database systems, Computer systems

B3 = Automated Knowledge-based systems

C= INFORMATION SCIENCE and AUDIO-VISUAL COMMUNICATION

C1 = Content analysis

C2 = Media documentation

C3 = Documentation Legislation

D= TELECOMMUNICATIONS and AUDIO-VISUAL COMMUNICATIONS

D1 = Media technology

D2 = Media theory

E= INFORMATION SCIENCE, TELECOMMUNICATIONS, and AUDIO-VISUAL-COMMUNICATION

E1 = Communication Theory

F= ALL DISCIPLINES

F1 = Perspectives on Information

F2 = UNIX / Internet

F3 = HTML, SGML, TEI

F4 = Hypertext technology

F5 = Electronic publishing

F6 = Information infrastructure

APPENDIX 2.1 TYPE-BASED WORDS IN RAs
(Argumentative, Procedural and Grammar / Discourse)
Selected items from the research articles type-based frequency list
* Words included appear in a minimum sample of three texts.

Word	Freq.	%	Word	Freq.	%
By	124	0.38	Range	20	0.06
Use	88	0.27	Include	19	0.06
Example	72	0.22	Characteristic	18	0.06
Text	67	0.21	Either	18	0.06
Such	64	0.20	First	18	0.06
Document	57	0.18	Reasoning	18	0.06
Which	51	0.16	Another	17	0.05
Set	38	0.12	Often	17	0.05
Form	37	0.11	Related	17	0.05
Function	36	0.12	Second	17	0.05
Also	36	0.11	Approach	16	0.05
Most	36	0.11	Query	16	0.05
Particular	34	0.11	Sets	16	0.05
Time	34	0.11	During	15	0.05
Because	32	0.10	Issue	15	0.05
However	32	0.10	Paragraph	15	0.05
Within	32	0.10	Results	15	0.05
So	30	0.09	While	15	0.05
Each	29	0.09	Content	14	0.04
Specific	29	0.09	Following	14	0.04
Study	29	0.09	He	14	0.04
Comments	26	0.08	Indicates	14	0.04
Even	26	0.08	Learning	14	0.04
How	26	0.08	Likely	14	0.04
Our	26	0.08	Result	14	0.04
What	26	0.08	Still	14	0.04
Context	25	0.08	Term	14	0.04
List	25	0.08	Variety	14	0.04
Process	25	0.08	Major	13	0.04
Concept	23	0.07	Rather	13	0.04
Different	23	0.07	Sample	13	0.04
Features	23	0.07	Although	12	0.04
Heading	23	0.07	Definition	12	0.04
Part	23	0.07	Including	12	0.04
Your	23	0.07	Indicate	12	0.04
Type	22	0.07	Point	12	0.04
Who	22	0.07	Presentation	12	0.04
Both	21	0.07	Yet	12	0.04
Forms	21	0.07	Approximately	11	0.03
Relationships	21	0.07	Examples	11	0.03
Way	21	0.07	Focus	11	0.03

Items	11	0.03
Same	11	0.03
Solving	11	0.03
Texts	11	0.03
Then	11	0.03
Concepts	10	0.03
Findings	10	0.03
Issues	10	0.03
Like	10	0.03
Methods	10	0.03
Specification	10	0.03
Types	10	0.03
Ways	10	0.03
Analysis	9	0.03
Defined	9	0.03
Describing	9	0.03
Here	9	0.03
Means	9	0.03
Procedure	9	0.03
Reference	9	0.03
Therefore	9	0.03
Thus	9	0.03
Uses	9	0.03
Contents	8	0.02
Figure	8	0.02
Graph	8	0.02
Graphic	8	0.02
Graphics	8	0.02
Headings	8	0.02
Individual	8	0.02
Instead	8	0.02
Others	8	0.02
Similar	8	0.02
Solution	8	0.02
Thinking	8	0.02
Us	8	0.02
Above	7	0.02
Addition	7	0.02
Additional	7	0.02
Conclusion	7	0.02
Containing	7	0.02
Documentation	7	0.02
Goal	6	0.02
Instance	6	0.02
Last	6	0.02
Parts	6	0.02
Processes	6	0.02
Sort	6	0.02
Specifications	6	0.02
Utilization	6	0.02
Whereas	6	0.02
Approaches	4	0.01
Approximate	4	0.01
Architectural	4	0.01

APPENDIX 2.2 : TYPE-BASED WORDS IN TXs

(Argumentative, Procedural and Grammar / Discourse)
Selected items from the textbooks type-based frequency list
* Words included appear in a minimum sample of three texts.

Word	Freq.	%
Text	35	0.16
Use	30	0.14
Chapter	22	0.10
More	22	0.10
Such	21	0.10
Issues	20	0.09
Each	19	0.09
Way	17	0.08
Because	15	0.07
Set	14	0.06
Book	12	0.05
Message	12	0.05
Analysis	11	0.05
Features	11	0.05
Forms	11	0.05
Most	11	0.05
Process	11	0.05
Your	11	0.05
Over	10	0.05
Thus	10	0.05
Within	10	0.05
Another	9	0.04
Between	9	0.04
Channel	9	0.04
Defined	9	0.04
Documents	9	0.04
First	9	0.04
Operations	9	0.04
Procedures	9	0.04
Reference	9	0.04
Same	9	0.04
Where	9	0.04
Even	8	0.04
Including	8	0.04
Less	8	0.04
Mode	8	0.04
Type	8	0.04
Author	7	0.03
Elements	7	0.03

Word	Freq.	%
General	7	0.03
However	7	0.03
Introductory	7	0.03
Messages	7	0.03
Specific	7	0.03
Things	7	0.03
Unique	7	0.03
After	6	0.03
Articles	6	0.03
Books	6	0.03
Both	6	0.03
Either	6	0.03
Explicitly	6	0.03
Function	6	0.03
Instead	6	0.03
Kind	6	0.03
Level	6	0.03
Others	6	0.03
Paragraph	6	0.03
Particular	6	0.03
Section	6	0.03
Specified	6	0.03
Then	6	0.03
Time	6	0.03
Title	6	0.03
Who	6	0.03
Below	5	0.02
Description	5	0.02
Fact	5	0.02
Few	5	0.02
Include	5	0.02
Kinds	5	0.02
Often	5	0.02
Procedure	5	0.02
Several	5	0.02
Though	5	0.02
Already	4	0.02
Among	4	0.02

APPENDIX 2.3 : TYPE-BASED WORDS IN RPs

(Argumentative, Procedural and Grammar / Discourse)

Selected items from the reports type-based frequency list

* Words included appear in a minimum sample of three texts.

Word	Freq.	%	Word	Freq.	%
On	165	0.33	Where	26	0.05
One	84	0.17	Functions	25	0.05
More	78	0.15	Document	24	0.05
Used	77	0.15	Following	24	0.05
Use	56	0.11	Most	24	0.05
Work	55	0.11	Part	24	0.05
Similar	55	0.11	Procedures	24	0.05
Such	51	0.10	Processing	24	0.05
Type	51	0.10	Both	23	0.05
About	49	0.10	Components	23	0.05
Types	48	0.10	Class	22	0.04
We	48	0.10	Procedure	22	0.04
When	48	0.10	Defined	21	0.04
Images	47	0.09	While	21	0.04
Using	47	0.09	Component	20	0.04
Each	45	0.09	Features	20	0.04
Example	42	0.08	Item	20	0.04
First	42	0.08	Us	20	0.04
Naming	41	0.08	Facts	19	0.04
Row	41	0.08	Over	19	0.04
Message	39	0.08	Whether	19	0.04
Then	39	0.08	Addition	18	0.04
Set	37	0.07	Different	18	0.04
Same	36	0.07	Examples	18	0.04
How	34	0.07	Goals	18	0.04
Your	34	0.07	Here	18	0.04
Authority	33	0.07	Include	18	0.04
Documents	32	0.06	Including	18	0.04
Than	32	0.06	Specific	18	0.04
However	30	0.06	Via	18	0.04
Process	30	0.06	Way	18	0.04
Because	29	0.06	Another	17	0.03
List	29	0.06	Like	17	0.03
Particular	29	0.06	Related	17	0.03
So	29	0.06	Text	17	0.03
Who	27	0.05	Too	17	0.03
Between	26	0.05	Whose	17	0.03
Even	26	0.05	Without	17	0.03
Image	26	0.05	Among	16	0.03

Now	16	0.03
Form	15	0.03
Formed	15	0.03
Index	15	0.03
Section	15	0.03
Note	14	0.03
Often	14	0.03
Since	14	0.03
Through	14	0.03
Together	14	0.03
Although	13	0.03
Around	13	0.03
Below	13	0.03
During	13	0.03
Methods	13	0.03
Operations	13	0.03
Reference	13	0.03
Additional	12	0.02
Author	12	0.02
Either	12	0.02
Entire	12	0.02
Especially	12	0.02
ssues	12	0.02
Likely	12	0.02
Report	12	0.02
Shall	12	0.02
Specified	12	0.02
Title	12	0.02
Various	12	0.02
Within	12	0.02
Written	12	0.02
Added	11	0.02
Certain	11	0.02
Devices	11	0.02
Dispersed	11	0.02
Formats	11	0.02
He	11	0.02
Includes	11	0.02
Instead	11	0.02
Others	11	0.02
Though	11	0.02
Analysis	10	0.02
Authorities	10	0.02
Collection	10	0.02
Comparison	10	0.02
Described	10	0.02
Distributed	10	0.02
Further	10	0.02
Instances	10	0.02
Series	10	0.02
Subtypes	10	0.02
Therefore	10	0.02
Contain	9	0.02
Contains	9	0.02
Corresponding	9	0.02
Fact	9	0.02
Heading	9	0.02
Inference	9	0.02
Method	9	0.02
Present	9	0.02
Rather	9	0.02
Reported	9	0.02
Term	9	0.02
Under	9	0.02
Add	8	0.02
Before	8	0.02
Conclusions	8	0.02
Decisions	8	0.02
Definitions	8	0.02
Essential	8	0.02
Establish	8	0.02
Ever	8	0.02
Factors	8	0.02
Focus	8	0.02
Identify	8	0.02
Involving	8	0.02
Kind	8	0.02
Modules	8	0.02
Once	8	0.02
Operating	8	0.02
Parts	8	0.02
Presented	8	0.02
Processes	8	0.02
Published	8	0.02
Result	8	0.02
Still	8	0.02
Thus	8	0.02
Ways	8	0.02
Amount	7	0.01
Article	7	0.01
Book	7	0.01
Content	7	0.01
Description	7	0.01
Finally	7	0.01
Function	7	0.01
Identified	7	0.01

Involve	7	0.01	Currently	6	0.01
Larger	7	0.01	Due	6	0.01
Last	7	0.01	Exist	6	0.01
Normally	7	0.01	Included	6	0.01
Ones	7	0.01	Indicate	6	0.01
Overall	7	0.01	Manner	6	0.01
Placed	7	0.01	Overview	6	0.01
Results	7	0.01	Paragraph	6	0.01
Separate	7	0.01	Quotes	6	0.01
Subsequent	7	0.01	Relevant	6	0.01
Again	6	0.01	Steps	6	0.01
Amounts	6	0.01	Study	6	0.01
Beyond	6	0.01	Things	6	0.01

APPENDIX 3 : TEXT TYPE-BASED LEXICAL ITEMS

20 ENTRIES – Argumentative, Procedural and Grammar / Discourse
Selected among the top 250 items of the Detailed Consistency analysis results.

OTHER -- 101 (RPs)	**(19) A1, A2, B1, B2, C3, D1, D2, E1, F2, F4, F6**
On the ___ hand, (3) / *in ___ words,* / *for purposes ___ than those* (1) * **Class** *Faced by ___ types of* (1) * **Concl**. *Reference to ___ metadata* (3) / *of ___ objects* / *on the ___ hand* / *From one program to the ___* / *they relate to each ___* (1) * **Defin**. *With ___ , more specific handouts* (1) / *or ___ information* * **Descr.** *But the ___ members* (3) / *___, more revealing* / *every ___ year* * **Discuss**. *___ departments* (2) / *on the ___ hand* (1) * **Exempl**.	

ALSO -- 66 (RPs)	(20) **A1, A2, B1, B2, C3, D1, D2, E1, F1, F2, F4, F6**
You must ___ consider (1) / *it ___ provides* * **Class** *Consider ___ , the* (1) * **Concl**. *It is ___ possible that* (2) / *may ___* / *but ___, other* (1) * **Defin**. *May ___* (4) / *can ___* / *and ___* (2) * **Descr**. *Can ___ be* + Past part. (4) / *may ___ be* + Past part. / *you can ___ get* / *and ___* / *___, the* (1) * **Discuss**. *But ___* (2) / *and ___* / *images shall ___ be coherently organized* (1) * **Exempl**.	

EXAMPLE -- 72 (RAs)	**(12) A2, B1, B3, C2, D1, E1, F3, F4, F5, F6**
, for ___, (2) * **Class** *the most notorious ___ of this* (1) * **Concl**. *For ___,* (10) * **Defin**. *___:* + sentence (9) / *see for ___,* (2) * **Discuss**. *___* # (10) / *for ___* / *of ___* # / *in ___* # (2) * **Exempl**.	

ABOUT -- 50 (RAs)	**(13) A2, B1, B3, C2, D1, E1, F3, F4, F5, F6**
Questions ___ (2) / *learn ___ the internet* / *details ___* / *Cost ___ $ 300,000* (1) * **Class** *Ratio of ___* # (3) / *reasoning ___* / *worry ___* (1) * **Concl**. *Thinking ___* (2) / *chat ___* / *bring ___ changes* (1) * **Defin**. *Information ___* (1) / *concern ___* (1) * **Descr**. *Information ___* (5) / *worrying ___* / *decisions ___* (1) * **Discuss**. *Information ___* (5) / *reasoning ___* / *know ___* / *Understand more ___* (1) * **Exempl**.	

EACH -- 45 (RPs)	(17) **A1, A2, B1, B2, C3, D1, D2, E1, F1, F2, F4, F6**
___ server (2) / *___ tool* / *distinguishing in ___* / *served by ___* (1) * **Class** *___ of* (2) / *___ such name* / *___ such digital object* / *talk to ___ other* / *relate to ___ other* (1) * **Defin**. *Give ___ of* (1) / *for ___* / *by ___* (1) * **Descr**. *So that ___* (1) / *then ___* * **Discuss**. *For ___ of the* (4) / *at ___ level* / *but ___ of the* / *___ task* (2) * **Exempl**.	

DOCUMENT -- 58 (RAs)	(10) **A2, B1, B2, C2, D1, E1, F3, F4, F5, F6**

Intended as a companion ___ (1) * **Concl**.

Not a single ___ *but* (1) * **Defin**.

The resulting ___ *is created* (1) * **Descr**.

The HTML ___ (4) / *within the HTML* ___ / *within the* ___ / *Describing the* ___ / *access the* ___ (1) * **Discuss**.

*Conventions of this*___ (1) / *browse the* ___ (1) * **Exempl**.

OUR -- 108 (RPs)	(11) **A1, A2, B2, C3, D1, D2, F2, F4, F6**

Let's move ___*focus to* (1) / *To* ___ *view* (1) * **Class**

___ *compiler* (2) / *Multi-segment files (described above) were* ___*first attempt to* (1) * **Concl**.

While with ___ *system* (4) / ___ *interface* / *to*___ *knowledge* / *we now define* ___ *terminology* / ___ *whole design* / ___ *marshaling code* (2) * **Defin**.

___ *standards* (4) / *Increasing* ___ *continuous processing procedures* / *to do* ___ *best* / ___ *overall performance* / ___ *use of* / *conduct* ___ *affairs* / *doing* ___*job* / ___ *search engine* / ___ *web site* (2) * **Discuss**.

Extend ___ *businesses* (1) * **Exempl**.

BETWEEN -- 39 (RAs)	(8) **A2, B3, C2, F3, F4, F6**
Relationships ___ (3) / *___ the way in which* / *___ particular individuals* / *___ people and media* (1) * **Defin**. *Relationships ___ concepts* (2) / *links ___ these concepts* / *Values ___ 40 and 78* / *space ___ terms* (1) * **Descr**. *Interactions ___* (4) / *the relationship ___* / *the study of the relation ___* / *exchanged ___* / *discriminate ___* (1) * **Discuss**. *The ratio ___* (3) / *the distance ___* / *transactions ___ devices* (1) * **Exempl**.	

MOST -- 36 (RAs)	(13) **A2, B2, C1, C2, F1, F3, F4, F5, F6**
The ___ important factor is (1) * **Class** *___ of these forms* (1) / *the ___ appropriate source* / *as with ___ remedies* (1) * **Concl**. *But ___ communities* (1) * **Defin**. *___ importantly,* (2) * **Descr**. *Yet ___ research* (2) / *will ___ likely* / *___ of the documents discussed* (1) * **Discuss**. *For ___ of us* (2) / *___ uses of* / *___ users* / *the ___ recent version* / *in ___ cases,* (1) * **Exempl**.	

HOW -- 38 (TXs)	(4) **A1, A2, C1, F2**
Guidelines for ___ to (2) / *documented ___* (1) * **Class** *___ does the text relate to* (2) / *___ do people* / *___ else would you* / *___far does this* (1) * **Concl**. *Decide ___* (1) / *understand ___* * **Discuss**. *Examples of ___* (2) / *know not ___* * **Exempl**.	

FIRST -- 42 (RPs)	(12) **A2, B1, B2, B3, C3, D1, F1, F2, F6**
The instruction comes ___ (1) / *analysis of this* ___ *videoconference* * **Class**. *Our* ___ *attempt to* (1) / ___ , * **Concl**. *Were* ___ *invented* (2) * **Defin**. *At* ___, (1) / ___, / *must* ___ *be mapped* * **Descr**. *This is the* ___ *year in which* (3) / *for the* ___ *time,* / *Proposal is* ___ *sent* (1) * **Discuss**. ___, (1) / *it was* ___ *described* / *the* ___ *task* * **Exempl**.	

BECAUSE -- 32 (RAs)	(9) **A2, B1, C1, D1, F3, F4, F5**
___ *of metadata* (1) / ___ *information* * **Concl**. *Precisely*___ *they* (2) / *also* ___ *of* (1) * **Defin**. *Precisely* ___ *they* (1) / *Additionally,* ___ (1) * **Descr**. *In part* ___ *of the* (2) / *however,* ___ *they* (1) * **Discuss**. ___ *we* (1) * **Exempl**.	

BOTH -- 23 (RPs)	(7) **A1, A2, C3, D2, F6**
___ *the* + Noun *and the* + Noun (2) / *see* ___ *of these products* * **Class**. *Be used* ___ *as* (1) * **Concl**. *Since* ___ *allow* (3) / *support for* ___ *VBScript and JScript* (1) * **Defin**. *For* ___ *kinds of mappings* (1) * **Descr**. *By* ___ (3) / ___ *have the* (1) * **Discuss**.	

DIFFERENT -- 23 (RAs)	(9) **A2, B2, C1, C2, D1, E1, F4, F5**
In a number of ___ ways (1) * **Class** *Sizes are quite ___from* (1) * **Concl**. *Addressed to several ___ purposes* (1) * **Defin**. *For ___fonts* (1) / *for ___ shapes* / *freely use ___ segments* / *In ___formats* * **Descr**. *A variety of ___ ways* (1) / *for ___ types of documents* / *A ___ approach* (1) * **Discuss**. *Infinitely ___ scale* (1) * **Exempl**.	

HOWEVER -- 32 (RAs)	(7) **A2, C1, C2, D1, F4, F6**
, ___, (1) * **Concl**. *The genre does not, ___, fully constrain* (1) * **Defin**. *. ___,* (3) / *Unfortunately, ___, / If, ___,* (1) * **Descr**. *___, whereas* (2) / *___, because* / *was, ___, designed* (2) * **Discuss**. *. ___,* (2) * **Exempl**.	

FORM-- 37 (RAs)	(7) **A2, C1, C2, F3, F4**
They ___ the boundary (1) * **Defin**. *In any ___* (1) * **Descr**. *In digital ___* (3) / *script ___* / *in published ___* / *in the ___ of* (1) * **Discuss**. *Use of an HTML ___* (2) / *has the ___* (1) * **Exempl**.	

OVER-- 19 (RPs)	(8) **B2, B3, C3, D1, F1, F6**
___ *time,* (1) * **Class**. *Provided* ___ *the NII* (1) * **Concl**. *Copied* ___ *the network* (3) / *run* ___ *the network* (1) * **Defin**. *Receive messages* ___ *the network* (1) / *there are* ___ *100 messages* * **Descr**. ___ *the previous 12 months* (2) * **Discuss**. *Refined* ___ *time* (1) * **Exempl**.	

PARTICULAR-- 34 (RAs)	(8) **A2, B1, B2, C2, F1, F4, F5**
In ___, (2) / *in* ___ *situations* / *in* ___ *ways* / *Between* ___ *individuals* / ___ *kinds of media* / ___ *other communities* * **Defin**. *In* ___, (4) / *the* ___ *choice of* (1) * **Descr**. *In a* ___ *context* (4) / *information exchange in* ___ (1) * **Discuss**. *A* ___ *schedule* (1) / *a* ___ *computer* * **Exempl**.	

CASE-- 19 (RPs)	(10) **B1, B2, C2, D1, F1, F4, F6**
The one ___ *where* (1) * **Concl**. *In the* ___ *of* (5) / *In this* ___ (2) * **Defin**. *In this* ___ (5) / *in the worst* ___ / *especially in the* ___ *when* (1) * **Descr**. *In the* ___ *of* (2) * **Discuss**. *The* ___ *in practice* (1) * **Exempl**.	

FOLLOWING-- 24 (RPs)	(11) **A2, B1, B2, B3, C3, D2, F1, F4, F6**
We have the ___ disadvantages (1) * **Class**. *Includes the ___ modules:* (1) / *the ___ Java example* (1) * **Defin**. *The ___ is used* (1) * **Descr**. *the ___ information* (3) / *in the ___ section* / *The ___ is an example* / *the ___ rule* (1) * **Discuss**. *Include the ___* (1) / *like the ___* * **Exempl**.	

APPENDIX 4 : GENRE-BASED LEXIS

(Academic and Subject)

Listed items from the Detailed Consistency List (DCL) arranged according to frequency and range
* Words included appear in the three genres (**TXs**, **RPs** and **RAs**), functioning mainly as academic and subject vocables.
* **NS** = number of sources in which words appear.
* **N** = ranked position of word in DCL.

N	Word	RAs	TXs	RPs	NS
20	Information	614	967	942	74
27	Web	414	258	286	61
32	Html	406	129	199	62
35	System	373	459	682	68
36	Data	309	685	570	67
38	Number	293	185	491	58
43	Internet	271	301	830	67
48	Library	262	269	727	59
58	Systems	236	345	581	57
59	Network	235	291	699	63
60	Access	230	528	408	51
63	Digital	205	671	338	56
64	C	351	223	500	38
66	Interface	85	694	262	32
73	Object	39	166	745	23

74	Document	155	474	305	46
81	**Page**	**163**	**406**	**306**	**45**
82	Available	217	314	309	56
88	Images	192	250	292	44
89	Image	174	732	234	44
90	Application	117	730	465	48
91	Language	125	715	446	44
92	Design	479	79	149	48
94	Message	15	489	179	36
96	Computer	230	290	147	41
97	Knowledge	290	82	294	43
98	Libraries	125	662	408	35
100	Electronic	186	287	180	45
101	People	192	367	92	42
102	Performance	293	202	155	39
104	Based	221	193	208	34
108	File	95	153	342	56
109	Form	197	277	116	45
110	Need	155	263	170	53
111	End	142	209	235	26
112	Development	231	174	173	32
113	Provide	201	492	205	43
114	Particular	194	249	132	42
115	Documents	153	209	205	26

116	Elements	87	359	113	27
117	Objects	42	556	199	25
118	Case	248	169	130	32
119	Following	108	291	148	33
121	Management	145	227	164	34
122	Issues	227	176	132	38
123	Code	134	163	237	22
124	Part	150	258	125	35
125	Level	163	195	173	24
126	Important	194	182	154	48
127	Method	139	91	295	42
128	Group	83	342	95	23
129	Provides	83	158	152	23
130	Order	232	133	150	28
131	Element	35	384	95	28
132	Address	63	128	317	29
134	Implementation	100	249	157	20
135	Communication	97	254	154	43
136	Applications	128	114	262	34
139	Analysis	160	267	66	31
140	Media	113	327	46	18
141	Include	98	250	135	39
142	Connection	40	172	234	28
143	Name	60	198	188	33

144	M	157	39	247	14
145	Content	133	187	107	36
146	Community	132	115	158	15
147	Discussion	50	269	81	32
148	Call	25	63	309	28
149	List	105	89	107	17
150	Format	105	154	124	17
151	Local	100	94	184	16
153	Material	107	177	84	26
154	Methods	132	78	152	19
155	Means	89	165	99	28
156	Model	70	244	37	13
157	Called	70	128	152	36
158	Paper	133	106	110	24
159	Client	51	60	237	14
165	High	111	117	112	29
166	Files	79	106	153	29
167	Defined	50	202	82	25
168	Ansi	117	114	98	12
169	Described	84	129	113	25
170	Including	90	116	115	28
171	Ability	38	231	52	26
172	Current	80	109	131	28
174	Copyright	31	49	238	15

176	Net	46	201	75	15
177	Issue	80	175	61	39
178	Function	122	109	84	25
181	Lt	199	26	88	13
182	National	60	131	121	17
183	Environment	73	145	90	25
184	Common	75	144	89	38
185	Members	194	62	50	19
186	Font	117	96	91	19
187	Materials	93	134	76	26
189	Pages	69	70	161	27
190	Distributed	41	151	102	35
191	Cost	66	129	98	29
192	Necessary	64	145	82	34
193	Book	101	85	105	24
194	Non	105	117	68	27
195	Own	72	118	97	43
196	Database	58	102	125	23
197	Collection	63	66	156	20
198	Hardware	81	100	104	18
199	Costs	56	174	53	19
200	**Found**	**140**	**84**	**56**	**29**
201	Memory	72	82	126	18
203	Approach	118	75	86	33

207	Features	68	119	90	38
209	Http	80	36	159	17
210	Codes	20	273	247	26
212	Associated	52	132	88	27
213	Find	129	73	67	38
215	Needed	71	105	93	33
216	Allows	61	110	98	44
217	Networks	71	123	75	15
218	Class	53	110	105	11
219	Needs	77	88	103	39
220	Figure	173	21	72	28
223	Multiple	56	128	81	24
224	Original	79	83	103	35
225	Basic	38	93	130	38
227	Appropriate	59	106	94	38
228	Get	65	70	124	47
230	Operations	47	149	62	28
231	Cd	121	92	44	16
233	Free	73	112	71	15
235	Languages	93	100	63	20
236	Messages	6	185	64	16
237	Context	101	99	55	21
238	Organization	133	48	73	18
241	J	99	32	119	15

242	Character	54	155	41	22
243	Cases	131	74	45	15
244	Good	81	90	78	23
245	Major	105	77	66	25
246	Base	23	125	100	18
249	Communications	35	152	58	20
250	Developed	67	88	89	26
251	Allow	69	101	72	25
252	Bit	38	101	102	14
253	Mechanism	19	166	56	16
254	Main	93	46	101	32
255	Operation	43	160	36	15
256	Create	69	88	82	30
257	Change	71	101	66	30
258	Government	17	159	62	12
263	Changes	80	80	74	24
265	Functional	32	44	156	13
266	Help	80	70	81	32
267	Operating	52	72	107	32
269	Links	97	62	71	32
270	Fact	91	68	71	31
271	Individual	85	86	59	17
272	Machine	46	85	99	19
273	Forms	81	104	43	24

274	Key	59	78	91	19
275	Become	71	82	74	34
277	Commercial	98	79	46	16
278	Date	35	129	58	18
279	Open	42	107	73	21
280	Included	56	112	53	32
285	Note	42	119	58	29
289	Mail	72	85	58	21
290	Experience	75	88	51	19
291	Encoding	35	164	15	18
294	Directory	26	55	132	22
295	Course	46	97	69	20
296	Handle	9	31	170	23
297	Known	47	77	84	25
298	Introduction	100	62	45	14
300	Involved	112	52	43	28
301	Activities	75	34	98	15
304	Created	45	74	85	26
307	Limited	92	41	69	19
308	Books	54	90	56	16
309	Contains	44	102	54	26
311	Contents	25	116	58	14
313	Computers	64	89	44	14
315	Evaluation	45	27	123	15

316	Existing	48	78	67	23
318	Area	85	63	44	14
320	Names	15	76	100	16
321	Complex	83	64	44	15
322	Discussed	37	124	29	26
323	Components	78	56	56	14
324	Designed	71	56	63	33
325	Domain	73	56	59	16
327	Individuals	110	48	30	15
329	Online	60	82	45	16
330	Feature	45	86	55	23
331	Hand	44	71	71	18
334	Know	47	81	57	26
335	Display	48	69	68	25
336	Conference	56	94	34	22
337	Link	57	49	77	21
340	Calls	18	30	135	18
341	Collections	35	44	103	14
344	Description	41	75	65	15
345	Groups	78	57	46	21
346	Error	31	38	112	16
348	Characters	41	95	45	18
349	Participants	26	48	106	14
350	Determine	38	86	56	23

351	Interest	59	76	45	21
353	Functions	26	65	88	19
354	Define	36	96	46	17
356	Basis	43	44	91	14
357	Librarians	22	39	116	10
362	Hypertext	91	46	39	15
364	Connections	18	24	132	16
365	Location	19	62	93	14
367	Business	59	48	65	14
370	Mode	59	72	40	12
371	Consider	64	57	50	18
373	Compiler	75	32	63	14
375	Includes	44	56	69	16
376	Difficult	58	56	55	23
377	Choice	31	120	18	23
378	Building	65	32	71	23
380	Author	44	32	91	14
386	Indicates	26	72	68	15
389	Copy	21	79	65	12
390	Education	16	52	97	11
391	Develop	34	66	64	18
392	Distribution	83	60	20	16
396	Perform	27	78	57	15
401	Formats	45	69	47	12

402	Definition	40	97	23	15
403	Easy	46	56	58	32
404	Directly	39	51	69	23
405	Concept	60	73	25	22
407	Appear	35	80	43	33
408	Global	10	60	87	14
410	Highly	77	53	27	15
411	News	49	91	16	14
412	Follows	50	70	36	18
413	Currently	41	58	57	18
414	Left	38	57	61	17
415	Office	16	34	105	12
418	Company	58	29	68	14
420	Great	58	63	33	32
421	Bits	44	41	69	11
422	Organizations	61	39	53	14
423	International	39	57	57	15
424	Developing	27	75	50	13
425	Meaning	14	113	25	12
426	Critical	28	62	61	18
427	Owner	4	5	142	12
428	Close	26	50	75	28
429	Optical	61	27	62	10
430	External	84	44	22	14

431	Identified	52	46	52	27
432	Contain	29	70	50	21
433	Added	31	44	74	29
435	Instance	16	72	61	31
436	Oriented	25	71	53	14
437	Perhaps	41	66	42	32
439	Audio	74	51	23	14
440	Default	21	40	86	17
441	Factors	60	49	37	16
442	Architecture	48	27	70	16
443	Findings	120	9	16	19
444	Complete	45	61	39	21
445	Look	40	69	36	34
446	American	31	70	44	21
447	Actual	42	41	61	21
448	Higher	42	51	51	19
449	Amount	50	63	31	24
450	Output	40	41	63	17
451	Early	38	66	39	20
452	Give	47	35	61	30
453	Done	40	67	36	36
454	Focus	44	44	55	26
455	Fonts	18	40	84	21
456	Networked	35	36	71	22

457	Come	43	51	48	34
459	Guidelines	51	63	27	22
460	Active	21	47	73	22
463	Congress	14	34	92	13
464	Levels	49	39	51	14
465	Conversion	55	66	18	10
466	Article	57	38	43	12
467	Fiber	25	10	102	11
468	Notes	21	93	23	15
469	Implemented	43	23	71	12
470	Importance	57	50	29	24
471	Archival	15	89	32	12
472	Input	44	37	54	15
473	Learning	73	23	39	15
474	Clear	45	44	45	14
475	Ftp	32	36	66	17
476	Efficient	39	35	60	16
477	Nature	49	57	28	17
478	Equipment	18	33	83	18
479	Index	19	34	81	15
480	Creating	34	60	39	16
481	Color	34	26	73	18
482	Argument	9	24	100	13
483	Direct	34	55	44	14

484	Low	41	55	37	14
485	Dependent	26	88	19	12
487	Networking	54	30	49	12
488	Beginning	26	60	46	14
489	Delivery	45	21	66	18
490	Let	75	17	40	25
491	Dynamic	24	75	33	16
492	Lower	64	33	35	14
493	Activity	49	34	49	15
494	Identifier	3	97	32	10
495	Events	39	78	15	13
496	Kinds	28	85	18	12
497	Documentation	47	72	12	13
498	Authors	54	43	34	13
499	Aspects	62	28	41	21
500	Especially	25	40	66	27
501	Length	38	39	54	23
502	Exist	46	55	29	23
503	Deal	41	54	35	26
504	Implementatio ns	12	70	48	12
505	Finding	70	25	35	15
506	Finally	63	30	36	38
507	Life	30	64	35	21
508	Formal	17	88	24	12

510	Built	46	39	44	23
512	Overall	57	28	43	26
513	Lists	19	58	51	21
514	Enough	48	39	41	22
515	Advantage	58	49	21	18
516	Comments	47	60	21	29
518	Details	27	54	47	25
519	Expected	54	48	26	22
521	Ideas	30	55	42	14
522	Multimedia	30	58	39	15
523	Effective	50	35	42	14
524	Market	61	46	20	12
525	Event	32	90	5	11
526	Internal	62	26	39	15
527	Browser	33	40	53	18
528	Characteristic	75	38	13	16
529	Far	33	57	36	24
530	Bound	45	68	13	16
531	Noted	19	72	35	15
532	Particularly	64	42	20	17
533	Numbers	25	24	77	15
534	Communities	47	28	50	13
535	Easily	48	28	49	18
536	Gain	27	23	75	18

537	Laser	30	3	92	12
538	Indicate	33	42	50	17
539	Capabilities	40	38	46	20
541	Initial	39	34	51	16
542	Identify	31	59	34	14
543	Later	42	31	51	15
545	Effect	22	49	52	16
546	Computing	31	51	41	19
547	Impact	59	29	35	12
549	Mean	40	64	18	16
550	Old	15	41	66	20
551	Journal	56	48	17	22
553	Mechanisms	16	61	44	12
554	Describe	42	26	53	18
555	Integer	25	36	60	13
556	Add	31	50	39	16
557	Human	35	48	37	19
558	Greater	47	38	35	20
559	Frame	26	28	66	14
560	Considered	51	34	35	23
561	Board	18	41	61	19
562	Ip	32	44	43	22
563	Bandwidth	66	43	10	13
565	Involves	52	54	12	15

567	Day	44	33	41	23
568	Email	24	34	59	16
569	According	43	57	17	17
570	Distance	58	20	39	21
571	Industry	32	57	28	15
572	Clients	13	14	90	12
573	Parts	44	45	28	14
574	Home	27	49	40	16
575	Component	44	24	48	17
577	Center	33	35	48	14
578	Phase	79	7	30	13
579	Personal	29	61	26	11
580	Moving	66	22	27	17
581	Offer	33	31	51	19
583	Larger	49	36	30	17
584	Organizational	76	26	13	13
585	History	32	61	22	12
586	Effort	25	41	49	15
587	Manner	16	83	15	10
590	Normal	15	19	79	14
591	Entire	44	28	41	14
592	Marked	12	89	12	12
593	Degree	44	36	32	14
595	Hard	33	52	27	21

596	Fixed	43	40	28	16
598	Connected	43	28	40	15
599	Cultural	23	74	14	11
600	Abstract	50	22	39	13
601	Ensure	18	38	54	14
602	Intended	20	51	39	13
603	Block	49	5	56	13
607	Association	27	47	35	12
608	Go	33	50	26	19
609	Databases	31	38	40	12
610	Department	50	19	40	12
611	Patterns	53	19	37	11
612	Detailed	40	45	24	14
619	Infrastructure	17	41	50	12
620	Obtain	33	34	40	19
621	Java	11	71	25	14
622	Capture	26	40	41	18
623	Performed	35	31	40	15
625	Graphics	46	51	9	11
626	Papers	18	62	26	12
631	Generated	29	29	47	14
632	Earlier	38	32	34	12
634	Explicit	42	42	20	12
635	Final	35	31	38	15

636	Cable	22	15	67	14
640	Due	59	7	38	16
641	Longer	39	37	27	17
643	Goal	38	40	24	21
644	Naming	11	7	84	15
645	Items	32	40	29	21
647	Generally	32	28	41	22
648	Failure	6	11	84	15
649	Cache	8	35	58	12
651	Increase	35	31	34	23
652	Measurement	53	40	7	12
653	Expertise	74	10	16	14
654	Legal	19	50	31	13
655	Matter	24	46	30	18
656	Count	62	10	28	18
657	Offers	25	47	28	14
658	Corresponding	40	16	43	17
659	Keep	33	28	38	22
660	Loop	44	2	53	14
661	Concrete	12	4	83	12
662	Modes	33	46	20	12
663	Detail	18	37	44	15
664	Describes	12	41	46	14
667	Exchange	30	54	14	15

668	Addressed	36	47	15	13
670	Manager	47	12	39	15
671	Options	26	41	31	19
673	Blocks	17	2	78	13
674	Linked	25	27	45	16
675	Interaction	34	52	11	11
677	Accuracy	43	32	22	14
679	Clearly	36	39	22	17
680	Contrast	24	36	37	18
682	Hierarchy	71	17	8	14
683	English	42	37	17	15
685	Becomes	26	26	44	21
688	Node	60	2	34	18
689	Concern	25	55	16	18
691	Overview	31	35	30	15
692	Log	31	12	53	17
693	Fully	24	45	27	16
695	Entry	13	36	46	18
697	Interesting	36	31	28	17
698	Attention	36	34	25	19
700	Conditions	40	28	27	15
701	Categories	47	28	20	18
702	Frequency	42	7	46	12
704	Begin	34	33	27	15

705	Interfaces	16	17	61	14
706	Appears	37	37	20	16
707	Lead	29	41	24	18
708	Apply	23	57	13	15
710	Package	27	12	54	15
712	Appendix	25	46	22	14
713	Going	41	34	18	16
716	Believe	26	32	34	16
717	Move	14	36	42	12
718	Exists	35	41	16	14
719	Criteria	37	24	31	16
720	Agent	11	7	74	12
721	Browsers	24	17	50	13
723	False	50	16	25	13
725	Perspective	14	42	35	14
726	Follow	16	34	41	15
728	Growth	27	31	33	14
729	Academic	53	27	11	13
730	Actually	16	33	41	18
731	Continue	23	26	41	21
736	Browse	80	8	1	13
737	Addresses	14	42	33	14
739	Bibliographic	18	38	33	16
740	Contained	15	45	29	11

741	Exact	68	15	6	11
743	Meet	20	31	37	13
744	Gives	33	23	32	24
745	Explicitly	21	50	17	11
746	Devices	19	33	36	12
747	Cross	36	26	26	17
748	Containing	18	26	44	22
750	Paragraph	16	8	64	13
751	Outside	30	38	20	14
756	Applied	34	35	18	19
757	Interested	15	43	29	12
758	Compile	54	16	17	15
759	Came	20	33	34	14
760	Numerous	16	52	18	12
761	Concerned	25	47	14	13
762	Figures	60	12	14	12
763	Decisions	37	27	22	15
764	Item	12	24	50	16
765	Easier	29	30	27	18
766	Asked	33	21	32	23
767	Browsing	62	19	5	12
768	Framework	32	39	15	14
769	Frequently	36	26	24	15
770	Maintain	15	35	36	16

771	Concerning	13	58	15	13
772	Changing	27	40	19	24
773	Face	22	36	28	21
775	Exactly	47	20	18	19
776	Learn	23	40	22	14
777	Desired	28	36	21	13
779	Aspect	25	33	27	16
782	Attempt	13	42	30	17
783	Opportunity	25	22	38	16
784	Machines	26	22	37	14
785	Avoid	27	25	33	23
786	Classes	25	23	37	12
787	Established	34	25	26	24
788	Engineering	31	30	23	18
791	Authority	9	13	62	11
792	Parameters	30	26	28	14
793	Customers	57	4	23	12
794	Institute	27	30	27	11
796	Period	18	39	27	23
797	Action	13	28	43	15
799	Allowing	21	32	30	19
800	Allowed	30	27	26	21
801	Creation	19	37	27	20
802	Multi	33	20	30	14

803	Library's	11	8	64	11
804	Essential	25	22	36	22
806	Pattern	14	13	55	12
807	Meeting	23	40	19	19
809	Past	32	30	20	16
810	Double	8	40	34	14
811	Core	25	40	17	11
812	Inc	24	29	29	13
813	Economic	14	56	12	11
814	Discuss	39	19	24	15
815	Implement	24	18	40	14
816	Effects	18	39	25	13
817	Maps	43	8	31	14
818	Accessible	21	34	26	19
819	Build	23	24	34	23
820	Maximum	19	38	24	14
821	Encoded	16	54	11	10
822	Path	30	14	37	18
823	Parameter	16	40	25	12
824	Followed	16	43	22	15
825	Idea	16	31	33	17
826	Operator	23	5	52	12
827	Compared	52	17	11	23
828	Interests	17	42	21	12

829	Cause	30	34	16	24
830	Map	27	18	35	12
831	Label	68	8	4	11
835	Implications	30	43	6	12
836	Errors	18	27	34	11
838	Expressed	20	38	21	18
839	Introduced	31	35	13	15
840	Achieve	40	16	23	21
841	Answer	35	24	20	19
842	Offered	10	35	34	18
843	Difference	16	23	40	18
845	Developers	30	27	21	16
850	Law	5	18	55	12
851	Check	25	16	37	18
852	Capability	26	35	17	14
853	Host	20	13	45	17
855	Determined	32	10	35	14
857	Necessarily	18	49	10	11
858	Command	51	16	10	12
861	Changed	9	38	30	14
862	House	32	7	38	15
866	Noise	63	1	12	12
867	Expert	22	1	53	16
868	Evidence	47	25	4	15

869	Direction	6	14	56	12
873	Assigned	17	14	44	21
874	Concepts	29	31	15	13
875	Account	23	33	19	18
876	Instruction	30	8	37	15
877	Concerns	29	37	9	14
878	Involve	25	32	18	17
879	Advantages	19	28	28	18
881	Entries	20	29	25	16
882	Consists	25	28	21	22
883	Device	16	37	21	15
885	Graph	66	2	6	12
886	Facilitate	25	24	25	22
887	Division	20	28	26	21
888	Generate	14	30	30	14
889	Cover	19	18	37	16
890	Pc	15	34	25	15
891	Combination	48	14	12	12
892	Communicate	17	17	40	16
893	Led	40	21	13	14
894	Funding	18	26	29	15
895	Catalog	15	12	46	12
896	Fast	14	16	43	14
897	Complexity	44	16	13	15

898	Ever	20	27	26	14
899	Approximately	43	23	7	14
900	Empty	11	21	41	12
901	Background	23	17	33	18
902	Comparison	32	20	21	13
904	Act	14	26	32	12
905	Everyone	17	37	18	12
906	Permit	17	24	31	13
909	Maintained	17	24	31	16
910	Differences	20	29	23	17
911	Looking	19	29	24	21
912	Mentioned	31	18	23	25
913	Foundation	20	37	15	12
914	Copying	9	23	40	15
915	Archive	26	23	23	16
916	Depends	31	29	12	16
918	Decision	33	16	22	17
921	Interpreted	15	45	11	15
922	Drive	15	41	15	12
923	Fundamental	9	40	22	10
924	Instructions	10	16	45	14
925	Million	20	31	20	16
927	Inter	16	24	31	16
930	Adding	34	22	14	15

931	Generation	17	29	24	12
932	Increasing	23	20	27	22
933	Central	24	21	25	24
934	Benefit	10	29	31	16
935	Became	17	23	30	19
937	Enable	17	39	14	18
939	Guide	18	23	29	14
940	Overhead	28	9	33	14
941	Mapping	30	21	19	13
942	Organized	31	17	22	16
944	Contact	11	17	42	15
946	Demonstrated	9	44	16	14
947	Commands	52	6	11	14
948	Obtained	34	11	24	22
949	Ascii	23	36	10	12
950	Involving	28	20	21	19
951	Benefits	16	24	29	16
952	Optic	9	4	56	12
953	Doing	16	23	30	18
954	Compressed	26	20	23	17
955	Advanced	24	25	20	18
959	Connecting	3	12	54	21
960	Existence	10	44	15	13
961	Approaches	21	31	17	18

962	Extensions	19	24	26	13
963	Equal	19	35	15	12
965	Expression	12	15	41	16
966	Lot	25	33	10	17
967	Connect	10	26	32	17
968	Comes	16	40	12	21
969	Arguments	5	9	54	11
970	Efforts	18	29	21	12
971	Improve	25	15	28	19
972	Normally	24	30	14	17
974	Interchange	5	43	20	12
975	Job	8	16	44	10
976	Manage	20	16	32	21
977	Boolean	19	15	34	13
979	Maintaining	17	20	30	22
981	Employed	30	24	13	21
982	Obvious	18	33	16	15
983	Extent	18	41	8	11
984	Except	19	29	19	15
985	Articles	26	25	16	14
986	Analog	26	39	2	13
987	Percent	15	36	16	11
989	Lack	24	24	19	23
992	Graphic	27	21	19	21

994	Extended	11	28	27	16
995	Began	13	19	34	21
998	Alternative	14	36	16	14
999	Instances	16	17	33	13
1000	Collected	8	33	25	13
1003	Frames	2	13	51	12
1004	Challenge	17	33	16	14
1006	Availability	22	30	13	18
1007	Carry	9	20	36	17
1008	Listed	14	30	21	14
1009	Movement	27	17	21	13
1010	Administration	16	17	32	14
1011	Bytes	10	21	34	12
1013	Depending	37	12	16	16
1015	Largely	35	14	16	19
1016	Loss	13	24	28	21
1017	Modern	14	33	18	15
1018	Head	12	19	34	15
1021	Correct	26	19	20	16
1022	Occur	24	26	15	13
1023	Extremely	18	32	14	14
1031	Calling	3	26	35	12
1032	Indicated	20	21	23	17
1033	Interactive	23	20	21	15

1034	Channel	17	40	7	10
1035	Channels	24	26	14	13
1037	Indexing	22	21	21	12
1036	Clean	2	24	38	12
1042	Phone	8	22	33	12
1043	Audience	28	28	7	11
1044	Fax	6	13	44	12
1045	Achieved	17	26	20	18
1048	Held	11	30	22	16
1049	Felt	20	34	9	13
1050	Integration	21	20	22	19
1051	Experiment	15	33	15	14
1052	Optimization	27	5	31	13
1053	Choose	18	25	20	26
1055	Literature	22	20	21	16
1056	Option	18	25	20	23
1058	Journals	21	28	14	22
1059	Logic	14	29	19	13
1060	Influence	42	16	4	16
1061	Dimension	10	23	29	14
1062	Permission	8	11	43	11
1063	Hold	11	31	20	18
1064	Measure	29	13	20	23
1065	Colors	13	13	36	14

1068	Growing	21	24	17	22
1069	Latter	21	30	11	21
1070	Front	6	20	36	13
1071	Increasingly	14	31	16	12
1072	Flow	33	16	12	14
1073	Ones	11	14	36	16
1074	Gateway	39	12	10	11
1075	Cpu	9	21	31	14
1076	Modules	25	14	22	15
1077	Integrated	14	25	22	17
1078	Linking	22	25	14	18
1079	Average	28	19	14	15
1080	Acceptable	17	23	21	19
1081	Arts	13	38	10	10
1082	Match	28	10	23	24
1084	Extensive	30	15	16	22
1085	Goals	19	11	31	20
1086	Expensive	16	22	23	15
1087	Connectivity	1	15	45	10
1088	Manuscript	5	27	28	11
1093	Automatically	12	28	20	18
1094	Corporate	24	9	27	16
1095	Completed	21	22	17	19
1096	Observed	29	24	7	17

1099	Discussions	17	31	12	13
1100	Module	16	17	27	11
1101	Paths	32	15	13	11
1103	Broad	11	24	24	15
1104	Carried	31	15	13	23
1105	Derived	19	27	13	21
1106	Originally	8	28	23	13
1107	Argued	6	51	2	10
1108	Modified	11	17	31	14
1109	Allocated	10	25	24	12
1111	Backbone	3	39	17	11
1112	Alone	21	18	20	14
1113	Capacity	11	21	27	14
1114	Invoked	12	16	31	16
1117	Configuration	23	23	12	18
1118	Describing	23	18	17	21
1120	Familiar	8	33	17	12
1121	Completion	24	30	4	11
1123	Mind	19	23	16	21
1124	Facilities	21	18	19	23
1125	Netscape	7	6	45	11
1126	Chosen	37	14	7	16
1127	Orders	29	9	20	15
1130	Encode	4	33	21	12

1131	Factor	30	12	16	16
1132	Located	22	21	14	18
1133	Constant	22	20	15	16
1136	Eventually	16	20	21	19
1138	Construction	12	28	17	16
1139	Displayed	35	9	13	21
1140	Feel	9	22	26	18
1141	Negative	27	9	21	21
1144	Aware	25	17	15	22
1145	Fail	4	32	21	12
1148	Disks	4	29	24	10
1151	Music	19	20	18	10
1152	Dealing	29	18	10	18
1153	Educational	16	21	19	16
1154	Energy	22	3	31	16
1156	Named	3	17	36	11
1157	Librarian	14	17	25	12
1158	Edition	12	34	10	12
1161	Establish	13	23	20	28
1163	Lost	14	28	14	17
1164	Meaningful	23	22	11	16
1165	Assignments	52	2	2	10
1166	Designing	43	4	9	12
1167	Measures	27	2	27	23

1168	Developments	17	26	13	21
1169	Money	17	18	21	14
1170	Everything	14	24	18	29
1172	Age	7	34	15	11
1173	Member	20	21	15	17
1174	Defining	15	27	13	15
1175	Demand	21	25	9	14
1176	Behavior	27	19	9	11
1178	Distinct	15	26	14	16
1179	Crucial	24	20	11	18
1180	Days	23	19	13	14
1181	Accurate	37	10	8	14
1182	Consistent	21	16	18	18
1185	Abstraction	29	16	10	13
1186	Embedded	12	27	16	12
1187	Examined	35	13	6	18
1188	Caused	34	6	14	23
1189	Focused	17	17	20	22
1190	Handles	5	1	48	11
1191	Computational	28	14	12	12
1194	Operate	18	14	22	21
1195	Minimum	14	21	19	19
1197	Equivalent	7	32	15	15
1198	Pair	12	12	30	15

1199	Ordering	22	7	25	22
1200	Gopher	14	12	28	13
1201	Effectively	25	12	17	21
1202	Mailing	8	29	17	12
1203	Meant	17	20	17	21
1204	Expect	17	17	20	24
1205	Performing	16	23	14	21
1206	Handling	11	19	23	23
1207	Card	5	24	24	11
1209	Pay	17	27	9	14
1210	Faster	12	19	22	19
1213	Half	12	19	22	16
1214	Agreement	29	17	7	12
1215	Increased	18	13	22	23
1217	Cut	13	29	11	12
1218	Experts	29	8	16	16
1219	Middle	23	12	18	19
1220	Converted	12	24	17	16
1222	Informal	34	6	13	17
1223	Facility	9	7	36	18
1225	Anyone	11	27	14	15
1226	Byte	6	14	32	12
1228	Bottom	27	14	11	15
1229	Becoming	14	19	19	17

1232	Classification	26	20	6	15
1233	Opposed	6	39	7	11
1235	Editor	6	32	14	12
1238	Linear	35	3	13	11
1242	Convenient	4	19	28	12
1243	Partial	25	14	12	16
1244	Difficulty	24	10	17	17
1245	Completely	12	24	15	24
1246	Optimizations	36	2	13	10
1247	Manual	9	22	20	12
1249	Accept	3	42	6	10
1250	Nothing	11	31	9	11
1251	Improvements	16	8	27	14
1254	Impossible	11	20	20	16
1255	Navigation	30	7	14	12
1256	Institution	10	15	26	12
1257	Category	15	25	11	17
1258	Disciplines	27	15	9	12
1260	Coverage	21	22	7	16
1261	Ibm	5	21	24	11
1263	Involvement	31	9	10	13
1266	Majority	28	15	7	12
1267	Emphasis	11	25	14	15
1268	Live	22	18	10	16

1270	Bring	20	22	8	19
1271	Coded	20	19	11	17
1272	Combined	20	14	16	21
1274	Intensive	36	3	11	16
1276	None	15	21	14	14
1278	Customer	22	6	21	14
1282	Fit	22	10	17	19
1283	Meetings	23	11	15	13
1284	Objectives	24	3	22	18
1285	Opportunities	31	6	12	19
1286	Descriptive	17	26	6	12
1287	Improved	16	10	23	16
1289	Cataloging	26	11	12	12
1290	Extend	12	19	18	17
1292	Locations	9	15	25	11
1293	Packet	14	21	14	12
1294	Applying	18	17	14	15
1295	Late	11	26	12	12
1297	Allocation	19	13	17	12
1299	Assumptions	10	16	23	14
1300	Brief	9	26	14	10
1302	Got	8	25	15	12
1303	Benchmark	22	3	23	11
1304	Binary	7	31	10	10

1309	Anything	15	22	11	16
1311	Limitations	17	18	13	11
1312	Labels	43	2	3	12
1314	Advance	6	20	22	13
1315	Improvement	38	5	5	15
1316	Depth	4	14	30	11
1317	Affect	16	21	11	20
1318	Ago	11	26	11	13
1323	Assume	7	23	17	13
1324	Break	7	8	32	12
1325	Helpful	10	27	10	15
1326	Laboratory	12	11	24	14
1330	Graphical	32	8	7	12
1331	Answers	15	14	18	23
1332	Adequate	23	8	16	19
1333	Efficiently	15	4	28	18
1334	Getting	10	19	18	14
1336	Pcs	10	19	18	12
1337	Concurrent	11	20	16	11
1338	Learned	5	23	19	11
1340	Macintosh	8	18	21	10
1341	Demands	29	4	14	26
1342	Occurs	9	18	20	23
1343	Automated	14	11	22	14

1348	Facts	12	7	27	19
1349	Big	19	15	12	21
1350	Debate	7	32	7	10
1352	Aid	13	8	25	15
1355	Maintenance	10	10	26	21
1356	Locate	8	12	26	21
1360	Certainly	9	22	15	12
1362	Experiences	10	20	16	14
1364	Ask	13	20	13	23
1365	Corporation	6	17	23	12
1366	Phases	37	5	4	12
1368	Magazines	29	10	7	13
1369	Helps	25	15	6	21
1371	Defines	6	21	19	14
1372	Brought	16	22	8	16
1373	Explained	8	34	4	14
1374	Flexibility	19	5	22	10
1376	Packets	21	9	15	12
1377	Citizens	2	27	16	11
1378	Giving	11	15	19	20
1380	Charge	9	17	19	18
1382	Delay	39	2	4	10
1383	Intermediate	20	9	16	12
1384	City	13	19	13	11

1387	Depend	17	19	9	12
1389	Immediately	15	12	18	19
1390	Iso	6	31	8	10
1391	Bug	5	2	38	10
1392	Interactions	15	21	9	11
1393	Compiled	7	16	22	12
1394	Dissemination	4	30	11	10
1396	Leading	7	17	21	18
1397	Accounts	15	12	17	19
1399	Managing	11	18	15	27
1401	Magazine	25	12	7	11
1403	Methodology	30	4	10	12
1404	Experiments	5	13	26	14
1407	Identical	4	22	18	14
1408	Commonly	16	17	11	19
1409	Desktop	13	22	9	12
1410	Drives	12	28	4	10
1411	Excellent	21	13	10	16
1416	Occurred	25	9	10	13
1417	Computation	10	14	20	14
1418	Mechanical	11	30	3	12
1422	Digits	10	4	30	18
1423	Adopted	16	24	4	15
1426	Council	10	20	14	13

1427	Formed	13	6	25	22
1428	Exploration	33	7	3	15
1429	Missing	1	6	36	12
1431	Objective	4	20	19	18
1435	Begins	14	23	6	12
1436	Fine	8	16	19	14
1442	Capable	8	22	13	13
1443	Deliver	14	18	11	21
1444	Creates	9	12	22	21
1447	Participation	7	15	21	14
1449	Conclusions	14	10	19	16
1450	Marketing	30	6	7	11
1451	Measured	25	1	17	18
1453	Delete	10	19	14	14
1454	Associate	2	29	12	12
1455	Considerable	13	18	12	19
1457	Holds	19	8	16	14
1461	Edge	18	17	7	12
1463	Participate	4	17	21	14
1466	Mainly	30	4	8	24
1469	Carrying	22	6	14	28
1470	Offering	17	7	18	23
1471	Compilation	24	9	9	11
1472	Ordinary	9	10	23	14

1474	Implies	12	21	9	17
1478	Differ	11	19	12	15
1481	Captured	15	12	15	18
1484	Illustrate	10	19	13	14
1486	Matching	15	7	20	18
1488	Entirely	12	22	7	10
1489	Estimate	31	4	6	13
1491	Compare	15	12	14	17
1493	Boundaries	12	20	9	12
1495	Financial	24	15	2	11
1496	Adds	5	19	17	18
1497	Closely	10	15	16	22
1498	Continuing	9	20	12	22
1499	Designated	18	5	18	13
1500	Collaborative	21	10	10	14
1501	Master	1	13	27	13
1502	Persistent	6	17	18	16
1503	Convert	14	18	9	23
1507	Minimal	13	18	10	15
1508	Installed	13	10	18	24
1509	Divided	14	15	12	18
1511	Hope	7	15	19	14
1513	Execution	14	21	6	11
1514	Distances	20	1	20	16

1516	Layout	6	12	23	14
1517	Identifying	8	23	10	19
1519	Chip	25	11	5	11
1520	Apparent	17	13	11	23
1521	Looks	8	12	21	20
1522	Copied	10	12	19	11
1524	Locally	12	14	14	11
1525	Colleagues	7	18	15	17
1526	Fairly	8	23	9	16
1527	Dos	5	13	22	13
1528	Analogous	11	16	13	15
1529	Carrier	12	6	22	13
1531	Choices	24	11	5	21
1532	Decide	21	11	8	18
1539	Amounts	17	9	14	12
1540	Consist	14	15	11	16
1541	Guaranteed	9	20	11	17
1543	Delivered	12	7	21	19
1544	Automatic	12	13	15	18
1545	Accepted	18	13	9	21
1546	Interpreter	7	27	6	14
1547	Checking	9	15	16	13
1548	Complicated	13	12	15	21
1549	Letter	6	20	14	16

1550	Decided	18	16	6	18
1551	Hands	7	20	12	18
1552	Faced	26	10	3	21
1554	Occurrence	1	24	14	14
1556	Editing	7	29	3	12
1557	Advertising	17	19	3	23
1558	Photograph	1	28	10	11
1559	Challenges	23	12	4	19
1560	Awareness	14	8	17	26
1561	Mit	5	14	20	12
1562	Fault	2	8	29	13
1563	Environments	16	15	8	19
1565	Older	8	16	15	22
1566	Commission	3	27	9	12
1567	Compatibility	9	14	16	15
1568	Claims	4	14	21	24
1569	Greatly	8	13	18	26
1570	Pass	8	7	24	23
1571	Introduce	20	10	9	21
1572	Assumed	6	17	16	25
1573	Believes	2	33	4	11
1574	Continued	2	10	27	16
1575	Incoming	11	4	24	17
1578	One's	10	27	2	12

1579	Distinguish	3	23	13	15
1580	Exploring	21	6	12	12
1581	Calculated	28	5	6	16
1583	Heading	26	1	11	23
1584	Increases	16	4	18	26
1585	Contribute	12	23	3	19
1589	Cgi	24	12	2	12
1590	Party	10	4	24	13
1591	Implementing	16	2	20	21
1592	Discovery	10	2	26	12
1593	Considering	16	15	7	16
1594	Automation	6	6	26	13
1595	Performs	5	15	18	21
1597	Chance	16	14	8	14
1598	Argue	11	26	1	14
1599	Determining	11	12	15	16
1600	Encountered	2	16	20	18
1601	Compatible	4	12	22	15
1602	Determines	17	15	6	16
1603	Demonstrate	5	11	22	19
1605	Clause	22	9	7	10
1606	Closed	6	27	5	12
1607	Intelligent	3	14	21	12
1608	Applicable	2	26	10	10

1609	Align	7	21	10	10
1610	Obtaining	13	12	13	19
1612	Fall	13	11	14	12
1615	Keeping	8	4	26	17
1617	Assess	14	4	20	29
1623	Evaluate	19	5	13	24
1624	Enhanced	8	8	21	12
1625	Explore	15	12	10	16
1626	Lan	11	22	4	10
1627	Passed	10	11	16	23
1628	Implicit	3	22	12	13
1630	Evaluated	21	1	15	16
1631	Leads	6	19	12	14
1632	Country	6	18	13	12
1633	Discusses	9	18	10	21
1635	Engaged	14	11	12	16
1637	Perceived	15	12	10	22
1639	Dedicated	9	8	20	14
1640	Examine	22	7	8	15
1641	Coming	9	11	17	23
1643	Nested	4	19	14	13
1647	Aim	15	3	19	25
1648	Elsewhere	9	21	7	14
1649	Films	10	12	14	12

1652	Notice	5	8	23	12
1653	Menu	13	4	19	12
1656	Installation	12	1	23	15
1657	Efficiency	7	13	16	12
1658	Conclusion	17	10	9	15
1665	Conduct	15	14	7	13
1666	Broader	9	19	8	14
1668	Analyze	15	12	9	11
1669	Conducted	17	8	11	16
1671	Distinguished	8	18	10	12
1672	Equally	12	14	10	18
1673	Math	18	3	15	13
1676	Hours	6	10	20	12
1677	Assign	6	11	19	18
1678	Empirical	25	7	4	12
1679	Notation	6	12	17	11
1680	Monitoring	10	10	15	12
1683	Constitute	3	24	8	12
1684	Indicating	8	18	9	14
1685	British	17	15	3	12
1689	Commented	9	11	15	14
1690	Illustrations	3	11	21	14
1691	Execute	5	3	27	13
1692	Dial	17	4	14	16

1695	Newsgroup	9	17	9	12
1697	Apple	2	17	16	11
1698	Dimensional	17	2	16	14
1699	Heavily	6	12	17	17
1701	Click	10	15	10	16
1702	Applies	6	13	16	16
1703	Collaboration	12	5	18	17
1706	Circuit	19	12	4	12
1707	Comparing	16	8	11	22
1711	Gets	11	13	11	20
1712	Goes	12	18	5	23
1713	Joint	11	13	11	18
1714	America	2	19	14	11
1715	Highest	10	12	13	23
1716	Descriptions	6	13	15	16
1718	Newer	7	23	4	10
1720	Cooperative	5	22	7	15
1722	Establishing	9	16	9	14
1723	Indexed	16	3	15	12
1725	Composed	7	14	13	16
1726	Mid	2	16	16	15
1727	Instructional	21	7	6	12
1728	Designer	22	8	4	12
1729	Collect	8	13	13	15

1730	Acknowledgement	7	8	19	13
1731	Measuring	20	2	12	15
1732	Closer	18	7	9	11
1733	Operators	18	3	13	11
1734	Discrete	16	14	4	13
1735	Experienced	10	13	11	14
1736	Ease	9	11	14	16
1737	Essentially	13	14	7	19
1739	Initiated	1	17	16	13
1741	Deposit	3	7	24	12
1742	Packages	14	7	13	13
1743	Engine	14	4	16	14
1745	Advice	10	7	17	19
1748	Guarantee	5	8	21	12
1749	Impacts	3	2	29	13
1752	Evolving	8	8	18	16
1753	Actions	9	11	13	14
1754	Emerging	10	11	12	17
1755	Permits	4	13	16	17
1756	Clock	4	20	9	10
1757	Agency	4	13	16	12
1759	Claim	2	13	18	14
1760	Examining	13	3	17	15
1761	Commitment	9	8	16	14

1762	Freely	7	8	18	13
1763	Controlled	13	11	9	13
1764	Dynamically	5	11	17	14
1767	Imaging	5	23	5	11
1770	Annual	11	11	11	12
1774	Employ	10	9	14	23
1775	Competition	7	15	11	16
1776	Box	16	4	13	12
1777	Knows	5	19	9	11
1778	Electronically	9	12	12	13
1780	Months	8	10	15	14
1782	Percentage	27	3	3	12
1783	Independently	9	14	10	18
1787	Exciting	14	12	7	12
1788	Displays	10	10	13	17
1789	Download	8	15	10	14
1791	Bus	12	19	2	11
1792	Attached	2	22	9	10
1793	Null	5	2	26	12
1794	Expressions	11	5	17	14
1797	Addressing	11	2	20	25
1798	Cooperation	3	10	20	14
1799	Forum	6	18	9	11
1800	Personnel	21	2	9	16

1801	Comment	5	17	10	14
1802	Constructed	8	18	6	15
1804	Director	5	12	15	12
1808	Afford	6	22	4	13
1809	Diverse	17	9	6	18
1811	Comparable	14	6	12	15
1812	Accessing	9	2	21	17
1814	Moved	11	14	7	15
1815	Failed	4	8	20	18
1816	Communicating	9	7	16	17
1819	Answering	25	3	4	13
1820	Incorporate	13	12	7	16
1821	Estimated	18	11	3	14
1829	Integrate	12	7	13	15
1832	Illustration	3	15	14	14
1834	Integers	9	4	19	11
1835	Draw	10	13	9	24
1837	Inside	4	14	14	21
1840	Balance	7	10	15	19
1842	Balanced	28	3	1	12
1843	Attempts	7	11	14	16
1844	Existed	19	3	10	20
1845	Charges	6	11	14	12
1848	Discipline	6	20	5	10

1853	Grant	10	5	16	15
1855	Assessment	8	4	19	17
1856	Managed	6	6	19	14
1857	Ensuring	12	10	9	16
1858	Consistency	7	7	17	12
1859	Inherent	8	12	11	13
1860	Daily	10	9	12	14
1861	Desire	8	15	8	13
1863	Continuous	9	9	13	19
1864	Initiative	13	12	6	15
1867	Organizing	14	7	10	19
1868	Cycles	4	9	18	13
1869	Participating	2	6	23	16
1870	Assumption	6	21	4	11
1871	Keyboard	13	9	9	12
1878	Monitor	14	13	4	13
1879	Investment	6	14	11	14
1880	Listing	1	12	17	19
1881	Express	11	13	6	20
1885	Observation	16	9	5	12
1887	Careful	9	7	14	12
1889	Evaluating	12	4	14	14
1890	Leave	3	6	21	16
1893	Identifies	8	14	8	14

1894	Digitized	16	6	8	12
1895	Faq	1	15	14	10
1896	Drawing	8	13	9	14
1897	Agreed	3	15	12	16
1898	Gave	5	15	10	17
1899	Lives	13	11	6	13
1900	Flexible	10	7	13	21
1901	Largest	19	8	3	18
1911	Briefly	10	6	14	16
1915	Carefully	6	8	16	17
1916	Devoted	16	6	8	18
1917	Extends	5	5	20	12
1918	Correspondence	7	6	17	14
1919	Invoke	4	5	21	14
1920	Executing	3	2	25	16
1922	Encourage	12	15	3	13
1924	Matters	10	9	11	16
1925	Analyses	10	17	3	11
1926	Moment	5	18	7	12
1927	Keywords	8	13	9	10
1928	Assist	5	6	18	12
1929	Hundred	6	15	8	13
1930	Expand	7	11	11	12
1931	Debugging	16	6	7	11

1934	Believed	8	17	4	14
1935	Downloaded	6	9	14	13
1936	Budget	5	8	16	14
1937	Faculty	10	6	13	13
1938	Met	12	11	6	16
1939	Month	7	8	14	12
1942	Patron	6	7	16	14
1943	Committees	2	7	20	12
1947	Intel	1	16	12	10
1948	Parties	9	11	9	12
1949	Fourth	5	5	19	13
1951	Perception	8	14	7	12
1952	Load	3	8	18	13
1953	Mission	5	6	18	11
1954	Discovered	14	7	8	10
1955	Anonymous	11	3	15	14
1957	Appeared	9	13	7	14
1959	Catalogue	5	9	15	12
1962	Helped	7	16	6	13
1964	Dictionary	12	12	5	11
1965	Enables	11	10	8	14
1968	Hosts	17	6	6	11
1970	Construct	9	13	6	12
1972	Agencies	7	4	17	11

1973	Considerably	12	7	9	14
1974	Invoking	3	4	21	13
1976	Accomplished	10	11	7	18
1977	Laws	8	10	10	13
1979	Explanation	7	9	12	14
1983	Marks	2	18	8	12
1984	Lay	5	3	20	16
1985	Outlined	4	12	12	15
1986	Immediate	8	8	12	21
1987	Mb	10	1	17	12
1988	Ground	8	14	6	12
1989	Acquire	7	12	9	12
1990	Jobs	8	9	11	10
1991	Executed	7	3	18	12
1993	Conceptual	9	6	13	14
1994	Interpret	5	14	9	10
1995	Granted	1	16	11	10
1996	Enormous	15	10	3	12
1998	Implements	3	3	22	14
2000	Permanent	6	11	11	12
2002	Contributed	6	6	16	12
2004	Driven	9	10	9	12
2005	Demonstration	8	12	8	11
2007	Concluded	7	18	3	10

2008	Bind	2	15	11	12
2009	Optimal	19	6	3	13
2010	Acted	24	1	2	13
2012	Finds	9	8	10	16
2016	Operational	8	9	10	14
2017	Headings	14	8	5	12
2018	Catalogs	7	1	19	11
2019	Assuming	6	18	3	12
2020	Feedback	3	17	7	14
2021	Leaves	7	3	17	16
2022	Accessed	14	3	10	23
2027	Mac	7	8	12	11
2028	Limits	10	11	6	13
2030	Iteration	14	2	11	13
2031	Emerged	9	12	6	12
2032	Ends	5	18	4	13
2036	Deployment	7	2	18	10
2037	Inclusion	12	11	4	12
2038	Alternatives	12	10	5	13
2039	Owned	4	7	16	14
2040	Continues	7	14	6	11
2041	Cycle	11	6	10	10
2042	Passive	12	11	4	10
2043	Consuming	10	10	7	12

2048	Absolute	7	15	5	11
2050	Costly	8	11	8	12
2052	Conform	6	15	6	11
2053	Backup	1	23	3	10
2057	Namely	7	19	1	12
2063	Foreign	8	5	13	12
2064	Fewer	14	6	6	12
2065	Knew	8	7	11	14
2067	Fair	8	13	5	13
2069	Actively	9	9	8	13
2071	Install	10	2	14	15
2074	Inheritance	1	4	21	12
2076	Loaded	4	7	15	15
2078	Finance	18	1	7	12
2080	Collective	13	8	5	11
2082	Businesses	11	6	9	13
2085	Choosing	16	4	6	14
2087	Painting	4	9	13	12
2089	Counting	16	1	9	13
2090	Fee	5	9	12	13
2091	Difficulties	8	12	6	15
2093	Happens	2	11	13	16
2095	Illustrates	7	8	11	14
2098	Modems	2	16	8	10

2099	Decade	6	7	13	12
2101	Deep	3	14	9	12
2102	Meta	2	11	13	11
2103	Expanded	9	11	6	13
2105	Gateways	12	7	7	10
2107	Encounters	1	9	16	14
2109	Acting	4	8	14	13
2110	Evolve	7	13	6	11
2112	Centre	6	4	16	14
2113	Begun	3	14	9	10
2115	Grown	7	7	12	12
2116	Combine	12	8	6	13
2118	Contribution	6	13	6	14
2119	Correspond	10	6	9	13
2120	Assembly	3	14	8	10
2122	Gathered	7	3	15	11
2123	Assistance	7	3	15	12
2124	Cited	5	18	2	11
2127	Improving	5	7	13	14
2130	Ongoing	7	8	10	14
2131	Causing	13	6	6	14
2132	Moves	4	17	4	13
2133	Administrator	15	1	9	10
2134	Ad	7	13	5	11

2135	Competitive	7	10	8	11
2136	Fiction	8	10	7	10
2138	Pairs	7	6	12	11
2139	Advertisement	2	19	4	10
2141	Highlighted	5	13	7	12
2142	Authorized	2	5	18	12
2145	Circulation	5	8	12	11
2146	Asking	9	8	8	12
2147	Huge	4	10	11	12
2148	Analyzing	9	5	11	11
2152	Frequent	16	5	4	12
2153	Branch	2	9	14	10
2157	Electrical	11	5	9	12
2160	Inch	1	10	14	12
2162	Centralized	3	4	18	12
2163	Outstanding	4	15	6	13
2165	Funds	6	8	11	11
2168	Inform	17	2	6	14
2169	Cards	7	12	6	11
2170	Informed	5	10	10	13
2172	Departments	12	3	10	12
2174	Arise	8	5	12	15
2175	Finished	2	7	16	16
2178	Derive	8	11	6	12

2179	Conjunction	8	6	11	10
2180	Paid	11	10	4	11
2181	Isolation	3	7	15	12
2183	Eliminate	3	9	12	10
2185	Integrating	11	2	11	15
2186	Dramatically	4	4	16	12
2187	Guarantees	4	7	13	14
2188	Designs	7	1	16	18
2192	Enhance	5	5	14	16
2193	Identification	6	11	7	13
2195	Evolved	14	3	7	14
2196	Acquired	13	7	4	13
2197	Constructs	2	15	7	10
2198	Emphasize	2	12	10	13
2199	Death	3	12	9	10
2201	Diagram	14	9	1	10
2203	Initiate	1	12	11	10
2204	Observations	13	6	5	11
2207	Architectures	9	6	9	10
2209	Obviously	5	11	8	14
2212	Explorer	5	10	9	10
2213	Broadly	5	15	4	11
2214	Adopt	11	9	4	12
2215	Bulletin	7	4	13	12

2217	Latest	7	9	8	11
2218	Motivation	11	3	10	13
2219	Horizontal	9	5	10	12
2221	Flat	4	5	15	13
2222	Forces	9	8	7	12
2223	Directed	8	4	12	12
2224	Indexes	4	6	14	11
2225	Organize	6	8	10	12
2226	Differing	10	10	4	14
2227	Adult	6	5	13	11
2229	Optimize	11	2	11	13
2230	Centers	6	13	5	11
2232	Investigate	8	2	13	12
2233	Converts	2	8	13	10
2235	Curve	11	2	10	14
2238	Courses	3	8	12	10
2240	Notify	1	3	19	11
2244	Newly	6	4	13	12
2245	Exercise	6	12	5	11
2247	Formatting	4	15	4	10
2249	Generalized	10	9	4	11
2250	Barriers	5	9	9	10
2252	Establishment	3	11	9	10
2253	Offices	6	12	5	10

2256	Compress	10	2	11	13
2258	Arrangement	7	8	8	10
2259	Characteristic	9	8	6	10
2263	Composition	3	6	14	10
2267	Couple	10	9	4	10
2269	Enabling	13	5	5	10
2270	Highway	4	11	8	11
2272	Executable	8	3	12	12
2273	Hour	6	12	5	10
2274	Confusion	7	9	7	10
2275	Aids	7	3	13	12
2276	Combining	11	5	7	10
2277	Marketplace	7	6	10	13
2280	Consultation	14	1	8	11
2281	Deciding	14	5	4	13
2289	Indication	10	8	5	12
2294	Inaccessible	2	2	18	10
2297	Enter	9	8	5	12
2299	Avoided	5	4	13	11
2300	Holding	9	4	9	10
2307	Alternate	5	5	12	10
2309	Duplicate	2	4	16	10
2310	Mhz	10	6	6	10
2317	Consisting	5	8	9	12

2319	Assessing	9	4	9	11
2320	Introducing	12	7	3	11
2323	Knowing	7	9	6	11
2324	Industries	5	13	4	10
2325	Discover	10	6	6	10
2326	Chain	3	12	7	10
2330	Consumer	7	11	4	10
2332	Friendly	10	5	7	13
2333	Forced	7	7	8	12
2340	Handled	3	6	13	11
2341	Passing	3	5	14	12
2344	Hundreds	6	6	10	12
2347	Incorporated	9	3	10	11
2348	Italics	8	9	5	12
2349	Opposite	6	10	6	11
2350	Ordered	9	5	8	12
2352	Implied	3	8	11	10
2355	Bringing	6	10	5	13
2357	Explored	8	5	8	12
2366	Definite	8	4	9	12
2369	Communicated	7	1	13	12
2371	Chose	10	2	9	13
2372	Occurring	9	8	4	10
2374	Issued	3	10	8	10

2376	Attempted	6	7	8	11
2382	Exploit	15	1	5	10
2384	Expressive	9	10	2	10
2385	Degradation	6	10	5	10
2387	Decrease	12	4	5	12
2388	Advances	3	8	10	10
2394	Informational	7	2	12	10
2395	Equation	7	4	10	11
2396	Conflict	11	7	3	11
2397	Highlights	3	11	7	10
2399	Covered	10	7	4	11
2400	Elaborate	8	21	8	5
2406	Greatest	6	11	4	11
2408	Fails	6	7	8	13
2409	Perfectly	5	8	8	11
2412	Listen	4	3	14	10
2414	Governments	5	7	9	10
2415	Enhancements	5	4	12	10
2416	Industrial	5	10	6	12
2417	Ahead	4	3	14	13
2419	Frequencies	2	7	12	10
2421	Orientation	11	3	7	11
2423	Diversity	6	12	3	11
2430	Degrade	6	7	8	11

2432	Counts	3	8	10	12
2433	Covers	8	3	10	13
2441	Charged	7	9	4	12
2442	Filled	9	6	5	13
2443	Imply	3	7	10	12
2448	Demonstrates	10	4	6	11
2449	Accomplish	4	10	6	12
2450	Circumstances	8	9	3	11
2451	Lose	4	9	7	12
2453	Extending	10	7	3	11
2455	Engines	10	7	3	12
2459	Button	4	4	12	13
2461	Interact	7	7	6	12
2462	Brings	5	9	6	12
2464	Enthusiasm	7	4	9	13
2466	Adapted	8	8	4	12
2467	Checked	5	3	12	13
2468	Aims	10	8	2	11
2470	Demonstrating	7	6	7	12
2471	Intelligence	7	4	9	13
2476	Massive	8	6	6	11
2480	Insert	1	8	11	14
2481	Measurements	6	6	8	12
2485	Highlight	5	12	3	10

2487	Helping	5	10	5	11
2493	Overcome	10	9	1	10
2494	Panel	12	2	6	11
2495	Detect	3	2	15	12
2498	Delays	15	3	2	12
2499	Attitude	3	2	15	12
2505	Emphasized	3	12	5	11
2508	Buy	9	6	5	10
2510	Imagine	10	8	2	10
2512	Facilitated	12	6	1	10
2517	Anywhere	6	7	6	12
2520	Employs	6	3	10	14
2524	Differently	4	5	10	12
2527	Heavy	6	4	9	11
2528	Incremental	7	7	5	12
2533	Controls	7	4	8	12
2534	Analyzed	11	3	5	11
2536	Gained	5	8	6	12
2538	Invokes	5	7	7	13
2540	Attempting	5	11	3	12
2541	Blank	4	6	9	10
2543	Acquisition	10	4	5	12
2546	Exchanged	6	9	4	14
2547	Grow	2	5	12	12

2548	Delivering	6	5	8	11
2551	Influenced	8	9	2	12
2552	Dropped	7	8	4	10
2559	Degrees	4	6	9	11
2560	Permitted	9	3	7	11
2562	Mention	6	8	5	12
2564	Individually	9	4	6	11
2565	Feasible	7	7	5	13
2567	Appearance	7	8	4	11
2573	Computed	12	3	4	11
2577	Perspectives	9	7	3	11
2579	Achievement	9	2	8	12
2580	Looked	7	6	6	13
2581	Mostly	8	8	3	12
2584	Coupled	6	9	4	10
2585	Importantly	7	2	10	15
2590	Outcome	12	3	4	13
2594	Branches	7	4	8	10
2598	Grows	12	3	4	10
2601	Distinguishing	2	8	9	11
2605	Correctly	5	7	6	13
2608	Capturing	10	5	3	11
2611	Associates	6	2	10	12
2616	Losses	12	3	3	10

2617	Marking	1	9	8	11
2618	Assisted	10	6	2	10
2621	Encouraged	5	10	3	10
2622	Leaving	5	6	7	11
2625	Cope	10	6	2	12
2626	Kept	5	5	8	13
2631	Encouraging	9	4	5	10
2633	Employing	11	4	3	11
2637	Bibliography	4	3	11	10
2642	Evolution	5	5	8	11
2644	Corresponds	6	9	3	10
2647	Conclude	10	5	3	10
2651	Favor	3	10	5	10
2653	Encounter	8	4	6	10
2658	Mediated	2	10	6	10
2660	Declares	3	8	7	12
2661	Extract	4	9	5	10
2668	Downloading	5	4	9	11
2671	Inexpensive	6	7	5	12
2673	Minor	7	4	7	13
2675	Notions	3	7	8	14
2676	Impose	12	4	2	10
2677	Bulk	8	4	6	12
2679	Join	3	10	5	13

2681	Membership	6	6	6	11
2683	Accompanied	5	8	5	12
2686	Irrelevant	15	1	2	11
2687	County	3	8	7	10
2694	Ai	8	5	5	10
2696	Incomplete	8	8	2	11
2697	Batch	3	5	9	10
2701	Numerical	9	2	6	11
2703	Decline	9	7	1	10
2712	Consulting	6	1	10	11
2717	Determination	3	4	10	10
2722	Installing	2	2	13	10
2725	Lies	3	8	6	10
2731	Outline	9	1	7	10
2740	Accepts	4	7	6	11
2744	Catalogues	4	7	6	10
2754	Digitization	7	3	7	10
2756	Attract	10	4	3	11
2757	Layered	6	5	6	10
2761	Appearing	1	5	11	12
2762	Operates	1	7	9	10
2768	Happen	3	8	6	11
2770	Laid	3	4	10	11
2772	Administrative	4	5	8	10

2774	Brand	10	5	2	10
2775	Exhibit	12	1	4	11
2782	Mix	9	5	3	10
2789	Guides	7	5	5	10
2793	Explain	2	5	10	10
2801	Arises	4	7	6	10
2803	Periodically	6	2	8	11
2811	Grants	1	5	10	10
2819	Constantly	2	5	9	11
2821	Enabled	3	7	6	10
2824	Navigate	7	3	6	10
2825	Dealt	11	3	2	10
2826	Naturally	10	4	2	11
2832	Detected	5	3	8	11
2834	Customized	2	5	9	10
2843	Exclusive	3	5	8	12
2844	Invented	1	7	8	10
2845	Passes	5	5	6	10
2851	Engage	2	8	6	11
2852	Lacks	5	9	2	10
2857	Adjusted	2	9	5	10
2862	Emerge	4	8	4	10
2877	Focuses	4	8	4	10
2880	Evident	11	3	2	12

2882	Disadvantages	2	5	9	10
2889	Partially	9	3	4	10
2891	Burden	5	7	4	10
2897	Accommodate	9	3	4	10
2908	Doubt	5	3	7	10
2912	Disseminate	5	6	4	11
2917	Manages	3	3	9	11
2922	Dispersed	1	5	9	10
2927	Firm	6	5	4	10
2928	Collecting	3	8	4	11
2936	Assurance	10	2	3	10
2937	Exploited	9	2	4	10
2938	Password	9	4	2	10
2942	Gather	7	5	3	11
2947	Achieving	9	1	5	12
2954	Founded	7	6	2	10
2960	Configurations	7	3	5	10
2964	Ignored	8	4	3	11
2967	Compete	8	3	4	10
2968	Investigated	9	3	3	10
2970	Acquiring	4	5	6	12
2973	Mathematics	5	7	3	10
2974	Fundamentally	4	6	5	10
2982	Assets	3	5	7	10

2995	Impression	5	3	7	10
3001	Meets	5	4	6	12
3016	Directions	6	4	5	11
3034	Bibliographies	4	6	5	10
3038	Highlighting	5	5	4	10
3046	Locating	5	2	7	10
3062	Loads	3	5	6	11
3074	Ordinarily	2	2	10	10
3091	Aimed	3	6	5	10
3106	Outcomes	11	2	1	10
3115	Arranged	4	3	7	10
3129	Chooses	7	2	5	10
3130	Manipulate	2	5	7	11
3135	Corrected	4	3	7	11
3136	Fix	3	6	5	10
3141	Approved	3	4	7	10
3145	Develops	2	7	5	10
3147	Fill	4	1	9	11
3148	Accesses	4	5	5	10
3164	Knowledgeabl e	3	5	6	10
3165	Matches	6	1	7	10
3171	Judged	7	4	2	11
3184	Deleting	6	3	4	10
3189	Caught	2	5	6	10

3192	Assure	1	5	7	10
3199	Distinguishes	2	7	4	10
3203	Fashion	7	2	4	10
3204	Modest	2	2	9	11
3205	Explores	7	3	3	10
3211	Incompatible	1	4	8	10
3217	Occasions	8	2	3	10
3226	Loan	1	2	10	11
3236	Ensures	2	5	6	10
3248	Inadequate	3	1	9	11
3253	Enhancing	6	1	6	10
3270	Merge	8	4	1	10
3274	Happening	3	6	4	11
3282	Inevitably	3	4	6	10
3284	Hopes	6	2	5	10
3285	Committed	3	5	5	11
3289	Consistently	4	6	3	10
3294	Denote	4	1	8	10
3295	Lengthy	4	5	4	10
3297	Originating	11	1	1	10
3306	Ended	5	4	4	10
3310	Combines	4	4	5	10
3319	Exceed	3	4	5	10
3321	Digitize	4	5	3	10

3324	Credibility	4	2	6	10
3326	Facing	4	3	5	10
3327	Abandoned	4	5	3	10
3328	Generalize	4	5	3	10
3339	Globally	2	2	8	10
3347	Basically	3	4	5	10
3354	Familiarity	2	6	4	10
3370	Arrange	6	1	5	10
3381	Expressing	7	3	2	11
3383	Lesson	6	2	4	10
3401	Insights	7	2	3	10
3404	Budgets	6	1	5	10
3443	Fits	5	6	1	10
3460	Accurately	2	5	5	11
3464	Inserting	2	5	5	10
3465	Origin	2	8	2	10
3471	Conceived	2	8	2	11
3475	Builds	3	3	6	10
3480	Failing	3	2	7	10
3498	Decreasing	3	2	6	10
3499	Logically	2	4	5	10
3525	Expresses	7	3	1	10
3579	Entirety	5	1	5	10
3618	Accounting	3	3	5	11

3634	Launched	4	2	5	10
3663	Award	4	1	6	10
3667	Compromise	4	4	3	10
3698	Cool	1	3	6	10
3717	Improves	8	1	1	10
3730	Nowadays	6	3	1	10
3736	Convey	2	4	4	10
3738	Indirectly	7	2	1	10
3756	Harder	5	4	1	10
3809	Differentiate	6	3	1	10
3887	Assessed	3	2	5	10

APPENDIX 5 : GENRE-BASED LEXICAL ITEMS

20 ENTRIES – Academic words and restricted (®) / free collocates (₣).
Selected among both the top 100 and bottom 100 items of the Detailed Consistency results.

NUMBER.—
185 / 491/ 293
TXs RPs RAs

®

a ___ of conflicts (36) / *estimation of the ___* / *estimate the ___ of* / *___ of constraints* / *the actual ___ of* / *printed page ___* / *___ of variables* / *mean ___ of* / *___ of figures and arrows* / *a ___ of studies* / *a ___ of new issues* / *___ of endpoints* / *___ of lines* / *shows the ___ of* / *the maximum ___ of* / *the minimum ___ of* / *there is no ___ of* / *if the ___* / *___ of bytes* / *increasing ___ of messages* / *___ of seconds* / *page ___for* / *a limited ___* / *evaluating the ___ of* / *limit the ___ to* / *___ of diseases* / *segment ___* / *a unique ___* / *count the ___ of* / *counting the ___ of* / *depending on the ___ of* / *___ of clauses* / *___ of colors* / *___ of errors* / *___ of purposes* / *___ of visits* / *causing the ___ to* / *control ___* / *___ of customers* / *___ of expectations* / *___ of implications* / *___ of steps* / *___ of interactions* / *the ___ shall increment* / *that the mean ___* / *growing ___* / *___ of images* / *___ of projects* / *___ of researchers* / *___ of symbols* / *___ of values* / *___ and size* / *___ of areas* / *___ of calls* / *___ of coefficients* / *___ of components* / *___ of data* / *___ of edges* / *___ of hits* / *___ of levels* / *___ of organizations* / *___ of paths* / *___ of results* / *___ that distinguishes* / *___ that identifies* /

₣

a ___ of (288) / *a ___ of figures* / *of the ___* / *a large ___ of* / *___ of sections* / *a ___ of users* / *a small ___ of* / *the total ___ of* / *in the ___ of* / *___ of people* / *on the ___* / *___ of issues* / *___ of elements* / *___ of topics* / *by taking the ___ of* / *___ of bits* / *___ of changes* / *given the ___ of* / *___ of cases* / *using the ___ of* / *port ___* / *additional ___* / *average ___ of* / *___ of objectives* / *___ of sites* / *the same ___* / *a great ___ of* / *page ___* / *specify the ___ of* / *___ of characters* /

NUMBERS.-- 24 / 77 / 25
®
a set of serial ___ (7) / assign a set of serial ___ /
___ are not tracked / ___ are tracked / carrying large ___ of /
floating point ___ / the ___ given in table # / limited ___ of /
more revealing ___ / ___for clients / ___ in network /
___ of telephone calls / sequence ___ / transmitting limited ___ /
complex ___ / line ___ /

₣

page ___ (16) / printed page ___ / large ___ of / huge ___ / high ___ /
phone ___ / fax ___ /

ENDORSEMENT.—
1 / 1 / 1
TXs RPs RAs

Support and ___ of (1) * All RP

INTERNET.—
301 / 830 / 271
TXs RPs RAs

®
connecting to the ___ (26) / using the ___ / the ___ community /
via the ___ / an ___ connection / about the ___ /
connected to the ___ / ___ and usenet / the ___ connection /
the use of ___ / use the ___for / from the ___ /
growth of the ___ / images on the ___ / commercialization of the ___ /
___ information server / public access to the ___ / through the ___ /
uses of the ___ / your ___ service / be on the ___ /
by the ___ / into the ___ / Microsoft ___ explorer /
the ___ should / public ___ / ___ resources / ___ access /
___ relay chat / full ___ access / ___ environment /
___ and intranets / ___ documents / ___ installation phase /
___ mail / ___ mailing lists /

₣

to the ___ (119) / on the ___ / of the ___ / access to the ___ /
over the ___ / ___ service / ___ service provider /
___ service providers / available on the ___ /
as the ___ / in the ___ / have access to the ___ /
information on the ___ / the ___ will /

GLANCE.-- 1 / 1 / 1
TXs RPs RAs

At first ___ (1) * All RP

NETWORK.—
291 / 699 / 235
TXs RPs RAs

®
___ *object runtime* (48) / *the* ___ *condition* / *relational* ___ /
campus ___ / *valid* ___ *object* / *across the* ___ /
___ *object argument* / ___ *object result* / ___ *object system* /
our ___ *objects* / ___ *connectivity* / *create a* ___ /
using the ___ / *global computer* ___ / *browsing the* ___ /
copied over the ___ / *data is copied over the* ___ /
___ *object agent* / ___ *structures for browsing* /
surrogate ___ *objects* / *evolved over the* ___ / *structure* ___ /
heterogeneous ___ / ___ *environment* / *semantic* ___ /
communication ___ / *information* ___ / *dependent* ___ *support* /
___ *data* / ___*failure* / ___ *information services* /
___ *interfaces change* / ___ *node offers* / ___ *object marshaling* /
___ *object method* / ___ *object stored* / ___ *objects available* /
___ *objects clients* / ___ *objects generalize* /
___ *objects programmers* / ___ *objects subsume* /
___ *operating systems* / ___ *partition* / ___ *services* /
___ *support for* / ___*-accessible storage* / ___*-mediated scholarship* /

₣

___ *object* (80) / *of the* ___ / *on the* ___ / *in the* ___ /
to the ___ / ___ *object type* / ___ *objects* / ___ *structure* /
by the ___ / *over the* ___ / *computer* ___ /
___ *connection* / *access + to the* ___ / ___ *of information* /
___ *resources* / *local area* ___ / ___ *service* /
___ *service providers* / *the* ___ *is expected to* / ___ *nodes* /
___ *objects provide* / *support* ___ / ___ *references* /
underlying ___ / ___ *links* / ___ *access* / ___ *use* /
___ *literacy* / ___ *manager* / ___ *news transfer protocol* /
___ *administrator* (2) * Com RA & Doc RP

NETWORKS.— 123 / 75 / 71

®
Education ___ (9) / *research and education* ___ /
Community ___ *conference* / *communications* ___ /
Over electronic ___ / *semantic* ___ / *neural* ___ /

₣
Computer ___ (38) / *use of* ___ / *community* ___ /
Information ___ / *communication* ___ / ___ *and roles* /
Social ___ /

NETWORKED.-- 36 / 71 / 35

®
___ *information services* (11) / *has been* ___ *to allow* /
___*-based services* / *developing* ___ *information* (2) * Doc RP

₣
___ *environment* (22) / *in a* ___ *environment* / ___ *information* /
___ *information resources* / *into the* ___ *environment* /
had not been ___ / *access to* +___ *information* /
___ *information service* / *use* ___ *information* /

NETWORKING.—30 / 49 / 54

®
Dial-up ___ *will use* (6) / ___ *connection* /
___ *moving images* / ___ *the library* / ___ *cultural heritage* /
___ *scripting support* /

₣
Dial-up ___ (36) / *community* ___ / *computer* ___ /
___ *uses* / ___ *supports* / ___*follows* /

DIMINISH.-- 1 / 1 / 1
TXs RPs RAs

___ *that ability* (1) * All TX

objects ___ *in apparent size* (1) * Aud RP

execution times ___ (1) * Com RA

ACCESS.—528 / 408 / 230
TXs RPs RAs

®
Preservation and ___ (16) / *___ and task aspects* / *gain ___ to the net* /
___ sites / *a protocol used to provide ___* / *___ to digital information* /
___ computer systems / *repository ___* / *___ and availability* /
___ and opportunity / *___ and tasks* / *___ aspect* / *___ aspects* /
___ control lists / *___ costs* / *___ to project* /
public-___ computer / *shift in ___* / *___ side* /
time and ___ / *___ permission modes* /

Ŧ
___ to the (110) / *___ to the internet* / *have ___ to* / *internet ___* /
providing ___ to / *provide ___ to* / *storage and ___* /
___ to electronic information / *gain ___ to* /
public ___ / *___ permission* / *direct ___* / *for ___ to* /
information ___ / *open ___* / *___ allowed* / *universal ___* /
information systems ___ / *equal ___* / *free ___* / *network ___* /

ACCESSING.-- 2 / 21 / 9

®
For ___ the table (3) / *means of ___ the* (2) * Doc RA

Ŧ
For ___ (4) / *___ the internet* /

ACCESSED.-- 3 / 10 / 14

®
Are ___ through (3)

Ŧ
can be ___ (11) / *can be ___ through* / *can be ___ by* /
shared and ___ /

ACCESSES.-- 5 / 5 / 4

Excessive disk ___ (2) * Com TX / *the user ___ information* (2) All TX

OVERWHELMED.--
1 / 1 / 2
TXs RPs RAs

___ by (2) * All TX & Doc RA / *___ with requests* (1) * Lib RP

DIGITAL.— 671 / 338 / 205
TXs RPs RAs

®
___ *object* (67) / ___ *objects* / *integrity of* ___ *information* /
archiving of ___ *information* / *certified* ___ *archives* /
___ *object whose* / *stored* ___ *object* / *microfilm to* ___*format* /
registered ___ *object* / *collections of* ___ *images* / *composite* ___ *objects* /
___ *technical journal* / *refreshing* ___ *information* / ___ *landscape* /
archiving ___ *information* / ___ *audio recording devices* /
composite ___ *object* / *development of* ___ *services* /
___ *equipment corporation* / ___ *image form* /
longevity of ___ *information* / ___ *data* / *provide + access to* ___ *resources* /
converted into ___*form* / *cornell* ___ *library* / *valuable* ___ *information* /
___ *images shall* / ___ *library community* / ___ *library systems* /
___ *object's data* / ___ *scanning* / ___ *storage* / ___ *storage costs* /
___ *systems research* / *distributed* ___ *archives* /

®
of ___ *information* (79) / *in* ___*form* / ___ *information objects* /
of ___ *archives* / *the* ___ *archives* / ___ *environment* /
preserving ___ *information* / ___ *preservation* / ___ *library* /
___ *materials* / *the* ___ *library* / *the national* ___ *library* /
access to ___ *information* / ___ *technology* /
preservation of ___ *information* / ___ *images* / *the* ___ *world* /
___ *library program* / *use of* ___ *information* / ___ *resources* /
___ *libraries* /

DIGITIZED.-- 6 / 8 / 16

®
___ *material* (2) * Aud RA / ___ *video data* (2) * Tel RA

Ŧ
___ *or animated images* (3) / ___ *images* / *can be* ___ (2) * Tel RA & Lib RA

DIGITIZE.-- 5 / 3 / 4
®

And ___ (3) / *permits users to* ___ (1) * All TX / ___ *documents* (1) * Doc RP /
___ *the content of* (1) * Aud RP

Ŧ

To ___ (7) / *to* ___ *it and* (2) * Tel RA & Aud RA

DIGITIZING.-- 6 / 3 / 2

___ *program* (3) / *for* ___ (2) * Aud RP / *preserving and* ___ (1) * Aud RA /
___ *process* (1) * All TX / ___ *sample images* * Aud RP

ATTENTION.— 34 / 25 / 36
TXs RPs RAs

®
A growing ___ (7) */ give almost no ___ / ___ to content /*
___ to five main issues / receives a growing ___ /
would recommend further ___ to /

Ŧ
___ to the (15) */ no ___ to / draw ___ to /*
pay more ___ to / we draw ___ to /

ATTENDANCE.-- 1 / 2 / 2

®
___ not required (1) * Tel RP / *program ___* (1) * Lib RA

Ŧ
___ at (3) */ ___ at the workshop* (1) * All TX /

ATTENDING.-- 2 / 1 / 1

___ the conferences (2) * All TX & Lib RA /
everyone ___ the workshop (1) * All TX

C.-- 223 / 500 / 351
TXs RPs RAs

®
Procedure ___ (27) */ connection ___ / section ___ / ___ library /*
program ___ / Code ___ / optical ___ / digital ___ / ___ image /
___ system / objective ___ / ___ location / ___ shell / visual ___ /
___ information systems / reader ___ / Writer ___ / attachment # ___ /
from ___ to / image requirement ___ / Object argument ___ /
object result ___ / ___ constraint / ___ calling conventions /
___ language bindings / ___-like syntax /

Ŧ
___ v (43) */ ___ and / for ___ / D ___ / such as ___ / written in ___ /*
___ ++ / programming in ___ / ___ language / Copyright ___ /
appendix ___ / ___ compiler / ___ compilers / ___ programs / ___ contains /
___ program / ___ indicates that / ___ is a / ___ languages / point ___ /
___ returns / defined by ___ / Microsoft visual ___++ / ___ ++ compiler /
___++ compilers / ___ ++ programs / visual ___ ++ interface /

OVERCOME.-- 9 / 1 / 10
TXs RPs RAs

®
lack of information has to be ___ (4)

Ŧ

to be ___ by (7) / *to ___ the problems* (2) * Lib RP & All TX

OVERCOMING.-- 2 / 1 / 1

For ___ (2) * Lib RP & All TX /
vital to / for ___ (2) * Lib RP & Doc RA /
vital to ___ obstacles (1) * Lib RP /
___ various deficiencies (1) * All TX /

PAGE.—406 / 306 / 163
TXs RPs RAs

®
glossary ___ (17) / *___ preface* / *longer than one ___* /
web ___ may include links / *___ number structure* /
bibliographic record print-___ number / *a particular ___* / *___ containing* /
architectural overview ___ / *base documents ___* / *within the ___* /
services ___ / *repeating ___* /

Ŧ
on ___ # (173) / *___ number* / *___ numbers* / *printed ___* /
contents ___ / *___ reproduced* / *of the ___* / *on the ___* /
___ by ___ / *actual printed ___* / *web ___* / *___ documents #* /
___ images / *___ navigation* / *home ___* / *___ table* /
book ___ images / *costs per ___* / *html ___* / *title ___* /
main title ___ / *last ___* /

PAGES.-- 70 / 161 / 69

®
Active server ___ (6) / *white ___* / *global white ___* /
Page or ___ contains / *yellow ___* / *between text ___* /

Ŧ
Web ___ (23) / *text ___* / *all ___* / *as many ___ as* /
most of these ___ / *___ in the book* / *about # ___* /
our ___ / *WWW ___* /

DOCUMENT.—
474 / 305 / 155
TXs RPs RAs

®
___ type definition (26) / *___ management systems* /
___ instance / *___ type declaration* / *TEI ___ type* /
based ___ management / *___ the consensus* /
numbered ___ / *___ using* / *___ management industry* /
___-like objects / *___-type definitions* /
information about the ___ / *the content of a ___* /
the ___ addresses / *originating ___* / *your ___* /
paper ___ / *evolving ___* / *technical ___* /
___formats / *___ structure plans* / *implementation shall ___* /
implementation to ___ / *destination ___* /

Ŧ
___ type (56) / *of the ___* / *in this ___* / *html ___* /
the same ___ / *current ___* / *to ___ the* /
within the ___ / *original ___* / *portable ___format* /
web ___ / *___ has been* / *___ may be* /
___ management / *electronic ___* / *single ___* /
SGML ___ / *an SGML ___* / *___ serves to* /
used to ___ / *___ #* / *entire ___* / *___ structure* /

DOCUMENTS.—209 / 205 / 153

®
base ___ (21) / *your ___* / *___ to be integrated with* /
all TEI ___ / *base ___ provide* / *___ in folder* /
___for sale / *___ marked up* / *ensure that ___* /
point our ___ / *unnumbered ___* /

Ŧ
html ___ (33) / *web ___* / *electronic ___* /
___ on the WWW / *___ can be* / *all of the ___* /
types of ___ / *___ will be* / *page ___ #* /
___ on the web / *other ___* / *set of ___* / *original ___* /

DOCUMENTATION.—72 / 12 / 47

®
implementation ___ (9) / *___ associated with* /
___ and preservation / *apply to ___* /
be aware of the ___ / *conformance ___* /
___for consideration / *___ of interface* /
implementation ___ required / *goal in developing ___* /
development of ___ /

Ŧ
___ provided (6) / *for the ___* / *of the ___* /
___ shall /

DOCUMENTED.-- 32 / 4 / 6
®
Entity not otherwise ___ (4) / *we must be ___* / *___ element* /

Ŧ
should be ___ (7) / *be ___ as* / *be ___ in* / *___ as such* /

DOCUMENT'S.-- 1 / 2 / 1

The ___ header (2) * All RP & All TX / *a ___ author* (1) * Doc RA

AVAILABLE.—
314 / 309 / 217
TXs RPs RAs

®
___ in ASCII format (12) / *___ to the interface* /
usually only ___ / *___ evidence* / *exploring ___ information* /
searching through ___ information / *___free of charge* /
___ hardware / *___ through newsgroups* / *___ to others* /
elements ___ / *writing is ___* / *applications are ___* /
frequently ___ / *___ # hours* / *___for marshaling* /
___ in electronic form / *___ in print* / *___ in reference* /
___ in systems / *___ on internet* / *___ to assist* /
___ to clients / *___ to most* / *___ to people* /
___ via world wide web / *characters not ___* /

Ŧ
___ on (81) / *___ on the web* / *___ to* /
information is ___ / *not ___* / *___ in* / *made ___ to* /
___ by ftp / *___ via* / *___ at* / *___for* /
___from / *may be ___* / *___ information* /
will be ___ / *would be ___* / *images ___* /
___ to anyone / *___ to users* / *readily ___* /
resources ___ to / *systems are ___* /

AVAILABILITY.— 30 / 13 / 22

®
in terms of the ___ (6) / ___ *and importance* /

Ŧ
the ___ *of* (19) / *access and* ___ /

MERGE.— 4 / 1 / 8
TXs RPs RAs

___*-find* (4) / *image* ___ (2) * All TX /
to ___ *virtual library services* (1) * Lib RA

MERGING.— 3 / 6 / 2

the ___ *of* (3) / *packet* ___ (2) * Tel RP / ___ *of systems* (2) * All TX /
___ *elements* (1) * Com RP

MERGED.-- 1 / 2 / 1

Are ___ *to* (2) * All TX & All RP / ___*file cache* (1) * Tel RP /
Images are ___ (1) * All TX

APPLICATION.—
730 / 465 / 117
TXs RPs RAs

®
mechanism for an ___ (27) / client ___ / server ___ /
___ task / ___ domain knowledge / ___ to accept /
___ to answer / an ___ to initiate / an ___ to specify /
the ___ translates into / between an ___ program and /
reduction ___ / should the ___ / ___ process /
the ___ shall / ___ specific commands / Basis Techlib ___ /
maximum ___ size / ___ register / distributed ___ /
___ needs /

Ŧ
an ___ to (85) / for an ___ / to the ___ /
allows an ___ / ___ program / ___ may /
___ domain / ___ to determine / ___ can /
___ of the / ___ interface / ___ areas /
___ knowledge / ___ relating to /
___ would be / intended ___ / ___ developers /
___ level / particular ___ / ___ must /

APPLICATIONS.— 114 / 262 / 128

®
range of ___ (7) / ___for sending / sensor ___ /
unmodified UNIX ___ / unaltered UNIX ___ /
supercomputing ___ / a broad range of ___ /
distributed ___ / ___ are available /
___ use multitasking /

Ŧ
LAN ___ (12) / the use of ___ / used by ___ /
of the ___ / ___ such as / ___ used /
communication ___ / typical ___ / case ___ /
___ written in / other ___ / design ___ /

APPLY.— 57 / 13 / 23

®
requirements that ___ (12) / ___ to multicast / that ___ to broadcast /
___ to similar projects / ___ to determine / ___ to documentation /
___ to programs / ___ to realtime / criteria to ___ /

Ŧ
that ___ to (18) / would ___ to /

APPLIED.-- 35 / 18 / 34

®
analysis cannot be ___ to (6) / *___ to design* / *___ to situations* /
basic and ___ /

Ŧ
be ___ to (38) / *can be ___* / *could be ___* / *___ viewpoint* / *___ research* /
___ to them / *has been ___* / *operation ___* /
to be ___ / *when ___ to* /

APPLYING.—17 / 14 / 18

___ the algorithm (7) / *in ___for* / *interested in ___* /

APPLICABLE.—26 / 10 / 2

®
requirements that are ___ to (6) / *___ to realtime* /
___ to shared distributed systems / *general model ___* /

Ŧ
___ to the (6) / *when ___* / *is ___ to* /

APPLIES.—13 / 16 / 6

___ to the conduct of (6) / *___ beautifully to* / *constraint ___ to* /
really ___ to / *semantics ___ to* / *structure ___ to* /

___ to the (12) / *___ to use* /

ENHANCED.— 8 / 21 / 8
TXs RPs RAs

®
___ by developments (3) / *innovations ___* /
software ___ with (2) * Tel RA

Ŧ
___ by (7) / *can be ___* /

ENHANCE.— 5 / 14 / 5

®
to ___ the presentation (3) / *___ total capacity* (2) * Com TX

Ŧ
to ___ (15) / *to ___ the* / *used to ___* /

ENHANCEMENTS.-- 4 / 12 / 5

®
___ *could be made* (4) / *synergistic* ___ / ___ *in our standards* /

Ŧ
that ___ (6) / ___ *in* /

ENHANCING.— 1 / 6 / 6

®
tools for ___ (3) / ___ *library workshop* (2) * Lib RA /
___ *the illusion of* (2) * Aud RP

Ŧ
for ___ *the* (4) /

ENHANCES.— 2 / 2 / 1

___ *the texture of* (2) * Aud RP / *HTML* ___ *ASCII files* * Doc RA /
the use of images ___ *the visual presentation* (1) * Lib TX

DESIGN.— 79 / 149 / 479
TXs RPs RAs

®
information systems for ___ (31) / ___ *and implementation* /
___ *and coding* / ___ *phase* / ___ *processes* /
first ___ / *second* ___ / *complexity of* ___ / *operating systems* ___ /
___ *information systems* / ___ *context* / *instructional* ___ /
aspects in the ___ / *associated with* ___ / *background on* ___
___ *and development* / ___ *components* / ___ *constraints* /
___ *depends upon* / ___ *involves* / ___ *proposal* /
detailed ___ / *environment of* ___ / *experience in* ___ /
experience with ___ / ___ *information* /
recommendations for ___ / *framework for media* ___ /
___ *a house* / *complicated* ___ *space* / ___ *employs* /
___ *makes heavy use of* / ___ *relies on* / *regarding the* ___ *of* /
divide a ___ *problem* /

Ŧ
of the ___ (69) / *the* ___ *process* / *the* ___ *of* / *in the* ___ /
for ___ / ___ *problem* / ___ *problems* / *during the* ___ /
architectural ___ / ___ *process* / *first* ___ / ___ *task* /
___ # / *regarding the* ___ / *in* ___ /
experimental ___ / *parts of the* ___ / *structure of the* ___ / ___ *disciplines* /
mechanical ___ / *non-*___ / *applied to* ___ / ___ *tasks* /
that ___ / *media* ___ / ___ *applications* / *our* ___ /

DESIGNED.— 56 / 63 / 71

®
the ___ object (12) / *___ to take advantage of* /
the study was ___ to / *___ and established* / *___ to deal with* / *___ to determine* /
mostly ___for the / *were ___ to* /
___ and implemented / *___for ease* /

₣
is ___ to (22) / *___ to be* / *was ___ to* / *are ___ to* / *be ___ to* /
___ to facilitate / *should be ___* / *system ___* / *___for* /
specifically ___ to / *___ as* / *___ to provide* /
___ to measure / *___ to support* / *___ to carry* /
___ to allow / *___ to handle* (2) * Com TX & Lib RP

DESIGNING.-- 4 / 9 / 43

®
___ child play areas (12) / *involved in ___* / *___ the organization* /
domain of ___ / *experience in ___* / *into ___ the* /
put into ___ / *___ can* /

₣
when ___ (8) / *that ___* / *in ___* / *about ___* / *___ is* /

DESIGNS.— 1 / 16 / 7

®
based on ___ (3) / *___ and fiber preparation* / *improved ___* /

₣
our ___ (4) /

MISSING.— 6 / 36 / 1
TXs RPs RAs

®
___ pages (3) / *___ documents* (2) * Aud RP / *___frame* (1) * All TX

₣
___ or unscannable (6) / *___ name* / *___ issues* / *___ images* /

MISSED.— 2 / 2 / 1

frames were *___ (2) * All TX* / we ___ a major opportunity *(1) * Com RP*

APPENDIX 6 : SUBJECT-BASED LEXICAL ITEMS

61 ENTRIES – Selected words among the top five items in each of the 17 subject categories.

SEGMENT.-- 133 (A1)	**A (1) History of Computers, Hardware, Software**
the same ___ (11) / *of a ___* / *___ number* / *bits ___* / *file and ___* / *in a ___* / *multi-___files* / *page of a ___* / *descriptor ___* / *energized ___ causes* (2) * Com RP	

BIT.-- 121 (A1)	**A (1) History of Computers, Hardware, Software**
#-___ (56) / *access mode ___* / *a ___from drive* / *8-___* / *___ unused in* / *privileged ___* / *7-___* / *32-___* / *16-___* / *realtime variable ___ rate* / *___ array* / *___ ASCII* / *___ data* / *___fields* / *___format* / *___ gate* / *___ data must be encoded* / *unencacheable ___ gate* / *32-___ version* / *with a ___ in* /	

RING.-- 89 (A1)	**A (1) History of Computers, Hardware, Software**
___ # (31) / *in ___ #* / *token ___* / *___ # code* / *code in ___* / *___ # is* / *executing in ___ #* / *token ___ clients* / *caller's ___* / *___ number of* / *___ brackets* / *the same ___ number* / *token ___frames* (2) * Com TX	

SUBJECTS.-- 120 (A2)	**A (2) Computer engineering and architecture, Data communications and Client-server architecture**
___ were asked (18) / *___ were asked to* / *the ___ in this condition* / *___ were told* / *all ___ completed* / *___ searched* / *as ___ in* / *___factors* / *encourages ___ to* / *for which ___* / *___ had to answer* / *letting ___free* / *___ are likely to* / *___ in our study* / *___ indicated the* / *___ participated in* / *asked ___ to browse* / *___ understood the* / *___ were free to* / *___ were given* / *___ were instructed on* / *___ were invited to* / *___ were supposed to* / *___ completed a questionnaire* /	

TASK.-- 180 (A2)	**A (2) Computer engineering and architecture, Data communications and Client-server architecture**
of the ___ (18) / *calling ___* / *___ completion time* / *calling ___* / *during ___ performance* / *___ environment* / *control ___* / *design ___* / *particular ___* / *application ___* / *___ aspects* / *problem solving ___* / *___ and structure* / *___ can accept* / *the context of a ___* / *main ___* / *browse ___* / *real ___* / *___ context* / *design ___* / *experimental ___* / *___ assignments* / *speed ___ completion* / *___ and performance* / *___ and type* / *___ characteristics* / *___ depending on* / *___ experiences* / *___ situations* / *___ that encourages* / *___ to meet the* / *___ was introduced* / *___ was used* / *___ will accept* /	

CONSTRAINTS.-- 118 (A2)	**A (2) Computer engineering and architecture, Data communications and Client-server architecture**
Set of ___ (20) / *___ of the* / *number of ___* / *System of ___* / *additional number of ___* / *Large number of ___* / *certain ___* / *collection of ___* / *___ are satisfied by* / *___ as attributes* / *___ considered by* / *___ on learning* / *___ provide criteria* / *___ will be* /	

TOOLS.-- 34 (B1)	**B (1) Information units management**
These ___ (5) / *basic* ___ / *software* ___ / ___ *that are available* /	

TOOL.-- 21 (B1)	**B (1) Information units management**
This is a ___ *that* (3) / *the Aggroup* ___ *set* / *a great* ___ *to use* (2) * Com RA / *a good backup* ___ (2) * Com RA	

AUTOMATION.-- 12 (B1)	**B (1) Information units management**
For ___ (6) / ___ *needs* (2) * Lib RP / *planning for* ___ (2) * Lib RP / *library* ___ *server* (1) * Com RA	

TYPE.-- 378 (B2)	**B (2) Online database systems and Computer systems**
Network object ___ (24) / *the* ___ *of the* / *object of* ___ / *determine its* ___ / *Concrete* ___ / *a sub-*___ *of* / *opaque* ___ / *surrogate of* ___ / *Runtime* ___ / *element* ___ / *be the* ___ / *check the* ___ / *Corresponding to the* ___ / *given a* ___ / *principal* ___ / *procedure of* ___ / *Runtime* ___ *test* / *runtime* ___ *support* / *the* ___ *whose* / ___*-code* / ___ *procedure* / *abstract data* ___ / *abstract* ___ / *distributed* ___*-checking* / ___ *dispatcher* / ___*fingerprints* / ___ *netobj* / ___ *safety* / ___ *declaration* / *by* ___*-code* / ___*-safe* /	

SURROGATE.-- 226 (B2)	**B (2) Online database systems and Computer systems**

The narrowest ___ rule (25) / *___ stream* / *___ object* / *___ will have* / *___ writer* / *___ reader* / *closes the ___* / *holds a ___* / *space holds a ___* / *marshaled as ___* / *get a ___* / *client ___ type* / *create the ___* / *___ allows* / *___ places* / *___ type available* / *___ type declaration* /

METHOD.-- 248 (B2)	**B (2) Online database systems and Computer systems**

Remote ___ call (12) / *___ overrides* / *remote ___ invocation* / *___ suite* / *___ number* / *servicecall ___* / *___ argument* / *___ being invoked* / *___ can raise* / *___ declaration indicates* / *___ declaration specifies* / *___for filling* / *___for writing* / *___ generates connections* / *___ is invoked* / *___ of communicating* / *___ of monitoring* / *___ of transports* / *___ processes the* / *___ recurses on* / *___ should therefore* / *___ waits for* / *___ specification* / *network object ___* / *new ___ called* / *new ___ generates* /

KNOWLEDGE.— 422 (B3)	**B (3) Automated Knowledge-based systems**

___ and technology (30) / *___ and expertise* / *types of ___* / *___ management* / *expertise and ___* / *___ relating to* / *the ___ base* / *___ of project* / *___ within* / *sharing of ___* / *communication of ___* / *distribution of ___* / *domain ___* / *extensive ___* / *___ and information* / *___ possessed by* / *___ engineer* / *___ system* / *___ and experience* / *___ sharing* / *___ systems* / *terminology of ___* / *___ distribution* / *___ regarding* / *programming ___* / *background ___* / *engineering ___ sources* / *extensive ___* / *___ amongst project members* / *___ cuts across boundaries* / *___ described earlier* / *___ distribution within* / *___ management terminology* / *___ use* / *___ was given* / *___ with regard to* / *___-based learning* /

PROGRAM.— 364 (B3)	**B (3) Automated Knowledge-based systems**

Of the ___ (149) / *development of the ___* / *regarding the ___* /
Final ___ / *on the ___* / *requirements of the ___* /
A large ___ / *build the ___* / *large scale ___* /
___ being built / *___ design components* / *the existing ___* /
the overall ___ / *within the ___* / *___ testing* /
the writing of the ___ / *___ module* / *complex ___* /
comprehensive ___ / *complete the ___* / *constructing the ___* /
in determining the overall ___ / *___ involves* / *___ already exists* /
___ analysis / *___ and questions* / *___ and storage* /
taking the ___ into account / *___ meant that* / *___ met* /
___ neared implementation / *___ requirements* /
___ specification / *___ was dependent on* / *___ was enormous* /
___ was intended to / *___ was likely to* / *___ was thinly spread* /
___ was widely acknowledged / *___ would have been* /

MEMBERS.— 172 (B3)	**B (3) Automated Knowledge-based systems**

Of project ___ (26) / *amongst project ___* /
Other project ___ / *resentment amongst project ___* /
Team ___ / *knowledge of project ___* / *all project ___* /
Faced by project ___ / *group ___ designed* /
Knowledge amongst project ___ / *___ acted as* / *___ allowed the* /
___ felt that / *___ had to* / *___ possessed knowledge* /
___ sought advice / *___ used* / *___ utilized* /
___ were concerned / *junior ___* / *most ___* /

MEDIA.— 74 (C1)	**C (1) Content analysis**

Mass ___ (22) / *communications ___* / *___ and information* /
Of mass ___ / *___ in the information* / *of communications ___* /
___ and information technologies / *___ economics* / *the mass ___* /
mass ___ gives (1) * Aud TX

GIF.— 38 (C1)	**C (1) Content analysis**
___ *border* (9) / *images / order _ u.*___ (6) / ___ *and JPEG files* (1) * Aud RA	

MASS.— 35 (C1)	**C (1) Content analysis**
___ *media* (21) / ___ *media and* / *of* ___ *media* / ___ *communication* / ___ *media and information technologies* / *the* ___ *media* /	

MEDIA.— 75 (C2)	**C (2) Media documentation**
New ___ (11) / *for new* ___ / *of new* ___ / ___ *design* / *Framework for* ___ *design* / ___ *and genres* / *to the* ___ / *In other* ___ / *multi-*___ /	

GENRE.— 34 (C2)	**C (2) Media documentation**
Each ___ (4) / ___ *might* / *a given* ___ / *new* ___ *might* (2) * Aud RA / *Small* ___ (1) * Aud RA /	

LANGUAGES.— 35 (C2)	**C (2) Media documentation**
Multiple ___ (8) / *in multiple ___* / *in most other ___* / *the JACKPHY ___* / *written ___* / *the world's written ___* (2) * Aud RA / *speakers of all ___* (1) * Aud RA /	

CONTRACTOR.— 84 (C3)	**C (3) Documentation Legislation**
The ___ shall (32) / *scanned ___'s option* / *the ___ must* / *By the ___* / *the ___ encounters* / *If the ___ encounters missing or unscannable film frames* / *The ___ shall assign* / *other ___ staff* / *___ must determine* / *___ project manager* / *___ must determine this number* / *the ___ may* / *the ___ will* / *___ must recognize this feature* / *the ___ shall be* / *the ___'s project* / *to the ___* /	

MICROFILM.— 91 (C3)	**C (3) Documentation Legislation**
On the ___ (16) / *of the ___* / *in the ___* / *Appear on the ___* / *___ images* / *___frames* / *The ___ materials* / *the ___ roll* / *next item on the ___* / *Frames on the ___* (2) * Aud RP / *no splices in the ___ roll* (2) * Aud RP /	

IMAGES.— 143 (C3)	**C (3) Documentation Legislation**
Digital ___ (25) / *film ___* / *all of the ___* / *___for* / *Film ___ are identified* / *if repeating film ___ are identified* / *___ shall be* / *___ to be* / *___ may be* / *___ that reproduce* / *___ when printed* / *book page ___* / *sets of ___* / *delivered ___* (2) * Aud RP /	

FIBER.— 102 (D1)	**D (1) Media Technology**
___ optic (42) / *___ optics* / *___-optic communication* / *___-optic cable* / *first transatlantic ___* / *___-optic systems* / *a typical ___ optic* / *all ___-optic cables* / *and effective ___ optic cable* / *___ preparation* / *___-optic cables are laid across* / *batch of optical ___s* / *developments in ___ optics* / *distance ___ cables* / *feet of ___* / *___ cable itself* / *___ preparation processes* / *___-optic amplifiers* / *___-optic installations* / *___-optic revolution* / *___-optics systems* / *___-optics trials* / *good ___ to* / *laser and ___ optics* / *crews installed two ___-optic cables* /	

LASER.— 81 (D1)	**D (1) Media Technology**
___ action (14) / *___ light* / *___ consisting of* / *___ could* / *___ using* / *semiconductor ___* / *___ action while using* / *___ beams* / *incoming ___* / *build a ___* / *confining the ___* / *detect an incoming ___* / *efficient communication ___* / *emit ___ light* / *focused ___ beams* / *ruby ___* / *illuminated by ___* / *initiate ___ action* / *___ action using* / *___ light bounces off* / *___ technology* / *___ that operates* / *___ to recreate* / *___ and fiber optics* / *layered semiconductor ___* / *wavelength ___ light* /	

FIBERS.— 63 (D1)	**D (1) Media Technology**
Optical ___ (30) / glass ___ / optical ___ could be / lasers and optical ___ / Optical ___ offered / asserted that ___ / best ___ / broken optical ___ / Existing glass ___ / ___ can carry / ___ emerge / optical ___ form / ___ gained increasingly in transparency / ___ that allow / ___ that meet / loss in ___ / such ___ can /	

IEEE.— 151 (D2)	**D (2) Media Theory**
The ___-SA (23) / ___-SA standards / of the ___ / the ___ standards / ___ draft standard / ___ standards association / ___-SA standards board / dictionary ___ std / ___ bod / ___ standards activities / other ___ entities / standards association ___ / advisor for ___ / association ___-SA / all aspects of ___ / ___ chair / development of ___ dictionaries / ___ board of directors / ___ intelligent transportation / ___ is mentioned / ___ serves the / ___ society / ___ standards computerization / ___ standards provide / ___ supports the / ___ web sites / ___-SA are / ___-SA member / ___-SA participation / proposed ___ standards /	

SCC.— 42 (D2)	**D (2) Media Theory**
The proposed ___ (6) / ___ number / the ___ coordinator / Active ___ meeting / appointed by the ___ coordinator / ___ activity / Interests in ___ / merging of ___ / newly renamed ___ / ___ # / ___ # is / ___ # will / the ___ chair /	

BOARD.-- 47 (D2)	**D (2) Media Theory**
Standards ___ (25) / *IEEE-SA standards ___* / *the standards ___* / *___ web site* / *activities ___* / *as the ___* / *___ chair* / *___ of governors* / *___ standards coordinating* / *___ takes action* / *___ will establish* / *IEEE ___ of directors* / *regional activities ___* / *___ vice-chair* (2) * Tel RP /	

SEMIOTICS.-- 203 (E1)	**E (1) Communication Theory**
___ for beginners (35) / *___ links* / *social ___* / *___ has* / *___ can help us to* / *concerns of ___* / *structuralist ___* / *___ has been* / *___ as a method of* (2) * Aud TX /	

CODES.-- 213 (E1)	**E (1) Communication Theory**
Metonymy ___ (15) / *___ modes* / *metonymy ___ modes* / *___ modes of address* / *elaborated ___* / *___ are described* / *___ of recognition* / *___ of transmission* / *___ such as* / *of such ___* / *signs and ___* / *stylistic ___* / *textual ___* / *character ___* (2) * Aud TX / *semiotic ___* (1) * Aud TX /	

ARCHIVES.-- 192 (E1)	**E (1) Communication Theory**
The digital ___ (26) / *certified digital ___* / *Operation of digital ___* / *system of digital ___* / *___ must* / *The operating environment of digital ___* / *libraries and ___* / *National ___* / *___ and records* / *digital ___ for* / *Digital ___ in* / *digital ___ of* / *digital ___ will* / *Distributed digital ___* / *requirements for digital ___* (2) * All TX / *manage ___ and backups* (1) * Doc RP /	

TECHLIB.-- 42 (F1)	**F (1) Perspectives on information**
The Basis ___ (11) / *the Basis ___ application* / *Basis ___ uses* / *Of basis ___* / *___ is a* / *Basis ___ is a framework designed to* (1) * Lib RP /	

MICROSOFT.-- 43 (F1)	**F (1) Perspectives on information**
___ active platform (8) / *www ___* / *___ internet information server* / *___ information* / *___ windows* (2) * Lib RP / *___ visual* (2) * Lib RP /	

NT.-- 30 (F1)	**F (1) Perspectives on information**
Windows ___ (19) / *windows ___ server* / *the windows ___ server* / *Windows ___ and Microsoft internet information server* /	

INTERNET.-- 356 (F2)	**F (2) Unix / Internet**
To the ___ (57) / *of the ___* / *connecting to the ___* / *On the ___* / *access to the ___* / *___ and usenet* / *Impact of usenet and the ___* / *___ connection* / *the ___ community* / *About the ___* / *be on the ___* / *connection to the ___* / *___ service providers* / *from the ___* / *in the ___* / *___ access needed* / *the ___ network* / *backbone of the ___* / *by the ___* / *development of the ___* / *for the ___* / *have access to the ___* / *have ___ access* / *___ documents* / *___ installation phase* / *___ network information* / *___ request for* / *planning pre-___ installation* / *the ___ should* / *the ___ was* / *the ___ will* / *the public ___* / *through the ___* / *using the ___* / *via the ___* / *is on the ___* / *participating in the ___* (2) * All RP	

ACCESS.— 349 (F2)	**F (2) Unix / Internet**
___ to the (40) / *___ to the internet* / *___ to the net* / *have ___ to* / *public ___ sites* / *service and open ___* / *for universal ___* / *that universal ___* / *___ and availability* / *equal ___ to* / *___ would be* / *internet ___ needed* / *of public ___* / *providing ___ to* / *thoughts on public ___* / *to ___ the* / *universal ___ to* / *___ can be* / *gain ___ to* / *guarantee of ___* / *have internet ___* / *make ___ available* / *provide equal ___* / *with full ___* / *with network ___* /	

USENET.— 151 (F2)	**F (2) Unix / Internet**
___ and the internet (11) / *internet and ___* / *history and impact of ___* / *___ and mailing lists* / *___ software* / *a ___ newsgroup* / *connect to ___* / *___ newsgroups and mailing lists* / *newsgroups on ___* / *on ___ newsgroups* / *professionals versus amateurs on ___* /	

ELEMENTS.— 363 (F3)	**F (3) HTML, SGML, TEI**
___ can be (11) / *___ described in* / *the following ___* / *a series of ___* / *one or more of these ___* / *all the ___* / *___ may be* / *___ unique to* / *the other ___* / *___ are used to* / *___ available in* / *___ defined in* / *use of the ___* / *by the ___* / *___ must be* / *text ___* / *___ needed to* / *___ nested within* / *___ should be* / *___ such as* / *___ to be* / *___ which appear* / *___ within a* / *more identifiable ___* / *the ___ described* / *___ to be synchronized* / *relevant text ___* (2) * All TX /	

TEXT.— 521 (F3)	**F (3) HTML, SGML, TEI**
Of the ___ (40) / *base ___* / *in the ___* / *of a ___* / *A spoken ___* / *default ___ structure* / *span of ___* / *The electronic ___* / *the same ___* / *___ encoding initiative* / *The enclosed ___* / *electronic ___ encoding* / *___ describing the* / *___ elements* / *preformatted ___ tag* / *in a sea of ___* / *the main ___* / *the source ___* / *the ___ itself* / *within the ___* /	

TEI.— 261 (F3)	**F (3) HTML, SGML, TEI**
Of the ___ (48) / *the ___ header* / *the ___ guidelines* / *in the ___* / *The ___ scheme* / *___ document type definition* / *modifying the ___* / *Obtaining ___ wsds* / *all ___ documents* / *available in all ___ documents* / *Chapter appeared in ___* / *encoding initiative ___* / *___ encoding scheme* / *___ guidelines query* / *___-interchange-format* / *the entity ___* /	

STREAMING.— 27 (F4)	**F (4) Hypertext technology**
___formats (6) / *___ audio* / *in ___format* / *video ___* / *___ technologies* (2) * Aud RA /	

HTML.— 52 (F4)	**F (4) Hypertext technology**
___ markup (3) / *___ tags* (2) * Lib RA / *___files* (2) * Aud RA /	

NEWS.— 30 (F4)	**F (4) Hypertext technology**
___ *broadcasts* (4) / *music and* ___ / ___ *and information* / ___*from* /	

ELECTRONIC.— 234 (F5)	**F (5) Electronic publishing**
___ *texts* (33) / ___ *publishing* / ___ *resources* / *the use of* ___ *texts* / *the* ___ *systems* / *use of* ___ *resources* / *in* ___*form* / *on* ___ *texts* / ___*format* / *workshop on* ___ *texts* / ___ *systems hold* / ___ *resources in the humanities* / *other* ___ *resources* / ___ *images* / *preparing* ___ *texts* / *archival* ___ *resources* / *center for* ___ *texts* / ___ *images constitute* / *into* ___*form* / *of* ___ *media* / *text in* ___*form* / *approach to preparing* ___ *texts* /	

TEXT.— 323 (F5)	**F (5) Electronic publishing**
Full ___ (21) / ___ *conversion* / *of the* ___ / *in the* ___ / ___ *and image* / *Model of* ___ / *one's model of a* ___ / *overview of* ___ / *Agricultural* ___ *digitizing* / *bodies of* ___ / ___ *in electronic form* / *Represent* ___ *in* / *sites use* ___ / *tagged* ___ / ___ *and graphics* / ___ *digitization program* / ___ *digitizing program* / ___ *encoding initiative* / *the copy* ___ / ___ *on a page* (2) * All TX / ___ *on the computer* (1) * All TX /	

WEB.— 239 (F5)	**F (5) Electronic publishing**
World wide ___ (48) / *writing for the* ___ / *the world wide* ___ / *Using the* ___ / *on the* ___ / ___ *documents* / *a* ___ *document* / *use of the* ___ / *on the world wide* ___ / ___ *site* / ___ *browser* / *world wide* ___ *sites* / *a* ___ *page* / *allows* ___ *authors to* / *in the* ___ / *sites on the* ___ / *world wide* ___ *organization* / *in the* ___ *environment* (2) * Lib TX /	

PUBLIC.— 213 (F6)	**F (6) Information infrastructure**
the ___ library (42) / *rural ___ libraries* / *___ access to* / *___ libraries in* / *for ___ librarians* / *___ library community* / *on ___ libraries* / *providing ___ access to* / *___ performance right* / *role of the ___ library* / *that ___ librarians* / *connecting rural ___ libraries to the internet* / *in providing ___ access* / *___ display right* / *___ librarians must* / *___ libraries cannot* / *the general ___* /	

TRAINING.— 112 (F6)	**F (6) Information infrastructure**
___ session (10) / *staff ___* / *___ and support* / *range of ___* / *___ and assistance* / *___ evaluation form* / *range of ___ programs* (2) * Lib RP /	

INTERNET.— 245 (F6)	**F (6) Information infrastructure**
to the ___ (37) / *of the ___* / *on the ___* / *access to the ___* / *the ___ community* / *using the ___* / *the ___ connection* / *about the ___* / *via the ___* / *uses of the ___* / *connected to the ___* / *of ___ connectivity* / *over the ___* / *available on the ___* / *available via the ___* / *can be used over the ___* / *get connected to the ___* / *in the ___* / *information on the ___* / *___ service providers* / *suite of ___* / *utilizing the ___* / *suite of ___ tools* (2) * Lib RP	

APPENDIX 7 : Contrastive view of Top 500 Academic / Subject Words

Listed items from the first 500 items in the Detailed Consistency List found in other sources.
* Sources being contrasted include the *Dictionary of Information Technology* (DIT) and the *Dictionary of Information Science and Technology* (DIST).
* **N** = ranked position of word on DCL / √ = item shows up in source /

N	Word	DIT	DIST
20	Information	√	√
32	Html	√	
35	System	√	
36	Data	√	√
38	Number	√	
43	Internet	√	√
48	Library	√	
59	Network	√	√
60	Access	√	
63	Digital	√	√
65	Text	√	√
66	Interface	√	√
67	Use	√	√
72	Document	√	√
73	Object	√	√
81	Page	√	
82	Available	√	
88	Images	√	
90	Application	√	
91	Language	√	√
92	Design	√	
94	Message	√	√
96	Computer	√	√
97	Knowledge	√	√
100	Electronic	√	√
102	Performance	√	
103	Different	√	
108	File	√	
109	Form	√	
111	End	√	
112	Development	√	
118	Case	√	√
121	Management	√	√
123	Code	√	√
124	Part	√	
125	Level	√	
128	Group	√	√
130	Order	√	√
131	Element	√	
132	Address	√	
134	Implementation	√	
135	Communication	√	√
139	Analysis	√	
140	Media	√	√
142	Connection	√	
143	Name	√	
144	M	√	
145	Content	√	
146	Community	√	
148	Call	√	
149	List	√	
150	Format	√	
151	Local	√	√
153	Material	√	
155	Means		√
156	Model	√	
158	Paper	√	
159	Client	√	
165	High	√	√
168	Ansi	√	√
172	Current	√	√
174	Copyright	√	
176	Net	√	
178	Function	√	
182	National	√	
183	Environment	√	
184	Common	√	
186	Font	√	
190	Distributed	√	√
193	Book	√	√
194	Non	√	√
196	Database	√	√
198	Hardware	√	
201	Memory	√	√

209	Http	√	
212	Associated	√	
213	Find	√	
218	Class	√	
220	Figure	√	
223	Multiple	√	√
224	Original	√	
225	Basic	√	
228	Get	√	
231	Cd	√	√
233	Free	√	
237	Context	√	√
238	Organization	√	
242	Character	√	
246	Base	√	√
252	Bit	√	√
253	Mechanism	√	
254	Main	√	
255	Operation	√	
256	Create	√	
257	Change	√	
258	Government	√	
266	Help	√	
270	Fact	√	
271	Individual	√	
272	Machine	√	√
274	Key	√	√
277	Commercial	√	
278	Date	√	
279	Open	√	√
285	Note	√	
289	Mail	√	√
291	Encoding	√	
294	Directory	√	√
295	Course	√	
296	Handle	√	
307	Limited	√	
315	Evaluation	√	
318	Area	√	
321	Complex	√	
325	Domain	√	√
329	Online	√	
330	Feature	√	√
331	Hand	√	
335	Display	√	
336	Conference	√	
337	Link	√	√
344	Description	√	
346	Error	√	
350	Determine	√	
354	Define	√	
356	Basis	√	
362	Hypertext	√	√
365	Location	√	
367	Business	√	
370	Mode	√	
373	Compiler	√	
378	Building	√	
380	Author	√	
389	Copy	√	
404	Directly	√	
408	Global	√	
411	News	√	
414	Left	√	
415	Office	√	√
423	International	√	√
426	Critical	√	√
428	Close	√	
429	Optical	√	√
430	External	√	√
432	Contain	√	
433	Added	√	
435	Instance	√	
439	Audio	√	
440	Default	√	√
442	Architecture	√	√
444	Complete	√	√
445	Look	√	
446	American	√	√
447	Actual	√	
449	Amount	√	
450	Output	√	
451	Early	√	
454	Focus	√	
460	Active	√	
465	Conversion	√	
467	Fiber	√	√
471	Archival	√	
472	Input	√	
473	Learning		√
474	Clear	√	
475	Ftp	√	
476	Efficient	√	
478	Equipment	√	
479	Index	√	√
482	Argument	√	

484	Low	√	
485	Dependent	√	
488	Beginning	√	
489	Delivery	√	
491	Dynamic	√	√
493	Activity	√	√
494	Identifier	√	

APPENDIX 8.1 : SAMPLE ACTIVITIES FOR ACADEMIC WRITING

Items belong as common core academic words in significant positions of our corpus (appendix 4).

I. Activating word forms commonly used in our corpus:

NOUN	VERB	PERSON	ADJ.	PAST
Application				
Knowledge				
Performance				
Development				
Implementation				

II. Choosing the correct noun form:

PERFORM + **KNOW +...?**

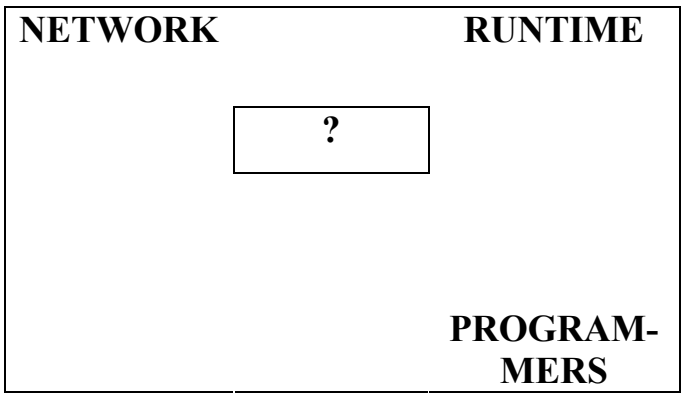
a.**LEDGE**
b.**ANCE**
c.**NESS**

III. Seeking restricted and subject-based collocates in written production.

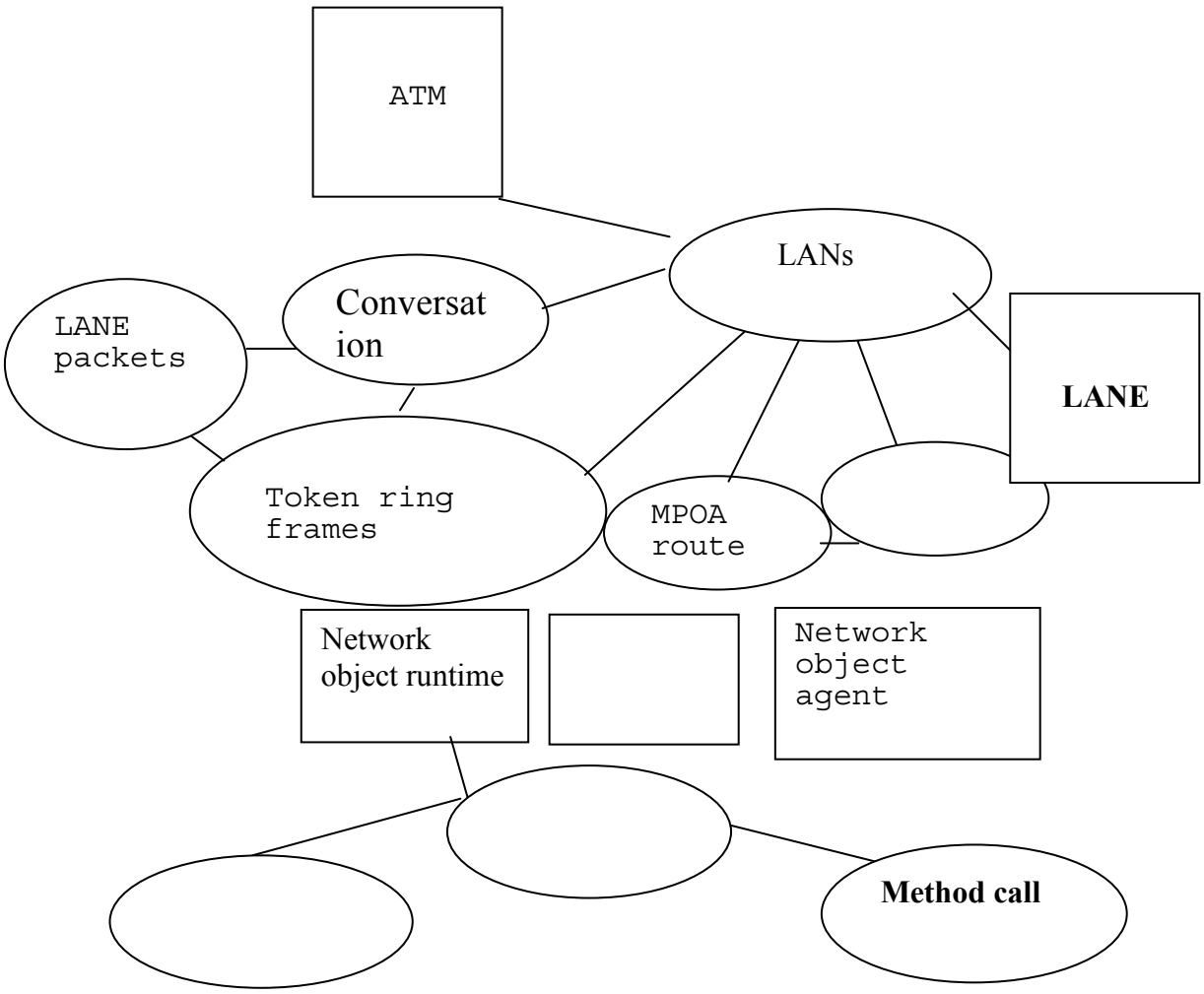
_____ object programmers _____ _____ runtime
_____ of constraints increasing _____ of
Token _____ clients _____ ring _____
_____ solving _____ speed _____ completion

IV. Guessing the central collocate:

TOKEN	**FRAMES**
?	
	CLIENTS


NETWORK
RUNTIME
?
PROGRAM-
MERS

V. Mind maps:


ATM
LANs
LANE
packets
Conversat
ion
LANE
Token ring
frames
MPOA
route
Network
object runtime
Network
object
agent
Method call

VI. Working with verbal and noun clusters by using samples as reference in pre-writing tasks:

Analysis of the performance of the system	
Type checking is performed	

A multiplicity of DTDs develops	
The development of the Net	

Objects that are implemented by	
This project's implementation objectives	

The application of SGML	
Algorithm is applied to	

APPENDIX 8.2 : LEXICAL SYLLABUS UNIT SAMPLES IN EIST

The first four items belong as common core academic and the last two units derive from subject word information

OPERATING

SUBTOPICS

* OPERATING SYSTEMS / HARDWARE AND OPERATING SYSTEMS / DESIGN AND OPERATING SYSTEM TOOLS

* RADARS OPERATING /

*APPLICATION OPERATING ON MESSAGES /

NOTIONS

***Systems operating in** / **Operated by library personnel** / **engine operates**

***Operating environment** / **operation of the Web** /

*** in operation by** / **such an operation** / **to the operating system** /

MAIL

SUBTOPICS

* E-MAIL / MAIL PROGRAMS

* MAIL SERVERS /

NOTIONS

* **Send the e-mail message / using e-mail /**

* **mail address / mail gateway / mail ordering /**

* **the following e-mail / via e-mail /**

MARKET

SUBTOPICS

* FREE MARKET / MARKET SHARE /

* MARKET TECHNOLOGY / MARKETING LIBRARIES /

NOTIONS

* **Gaining market share / market the library / market to exploit**

* **market capitalization / market observers /**

* **on the market / market-driven /**

ENHANCE

SUBTOPICS

* ENHANCEMENT SOFTWARE TOOLS / ENHANCE CAPACITY /

NOTIONS

*** Enhanced by developments / enhance the presentation / used to enhance /**

*** Enhancements / Enhancing library workshop /**

*** enhanced by / for enhancing / enhancements in /**

SEMIOTICS

SUBTOPICS

* SEMIOTICS FOR BEGINNERS / SEMIOTICS AND THE WEB

* SOCIAL SEMIOTICS /

* STRUCTURALIST SEMIOTICS /

NOTIONS

*** Semiotics can / Semiotics has become /**

*** Semiotics links / Concerns of Semiotics /**

*** Semiotics for / Semiotics as a method of /**

ADAPTER

SUBTOPICS

* NETWORK ADAPTER / HOST ADAPTER /

* VIDEO ADAPTER /

NOTIONS

*** adapter plugs into / adapter converts /**

*** terminal adapter / display adapter /**

*** host adapter as /**

BIBLIOGRAPHY AND CORPUS REFERENCES

BIBLIOGRAPHY AND CORPUS REFERENCES

I. BOOKS, ARTICLES, ESSAYS CITED

Aarts, J. and W. Meijs (eds.) (1984) Corpus Linguistics: Recent Developments in the Use of Computer Corpora in English Language Research. Amsterdam: Rodopi.

Aborn, M. et al. (1959) "Sources of Contextual Constraint Upon Words in Sentences". Journal of Experimental Psychology 57 (3): 171-180.

Ackerman, F. and A.E. Goldberg (1996) "Constraints on Adjectival Past Participles". In A.E. Goldberg (eds.) Conceptual Structure, Discourse and Language. Stanford: CSLI Publications.

Aghbar, A.A. and H. Tang (1991) "Partial Credit Scoring of Cloze-Type Items". Evaluative Report 142. ERIC DRS.

Ahmad, K. and A. Davies (1997) "The Role of Corpora in Studying and Promoting Welsh". In Wichmann, A. et al. (eds.) Teaching and Language Corpora. London: Longman.

Aisenstadt, B. (1981) "Restricted Collocations in English Lexicology". International Review of Applied Linguistics 53: 53-61.

Aitchison, J. (1987) "Reproductive Furniture and Extinguished Professors". In R. Steele and T. Threadgold (eds.) Language Topics: Essays in Honor of Michael Halliday. Amsterdam: Benjamins.

Alejo González, R. (1991) "Selección de vocabulario en el `Business English´". In M. McGinity, R. Alejo and S. Gómez (eds.) Lenguas aplicadas a las ciencias y tecnología II. Badajoz: Universidad de Extremadura.

Alejo González, R. (1998) "Code-Switching in Spanish Economic Discourse: Rhetorical Strategies in Research Articles". In I. Fortanet, S. Posteguillo, J.C. Palmer and J.F. Coll (eds.)

Genre Studies in English for Academic Purposes. Castellón: Publicacions de la Universitat Jaume I.

Alejos Juez, M.T. (1995) "The Acquisition of Lexis: Its Pedagogical Implications in ESP". Actas VI Congresso Luso-espanhol de Linguas para Fins Especificos. Viseu: Instituto politecnico superior.

Alekseev, P.M., W.M. Kalinin and R.G. Piotrowski (1973) Sprachstatistik. Berlin: Akademie Verlag.

Alexander, R.J. (1984) Idiomaticity and Other Related Problems. ELR. Birmingham: University of Birmingham.

Allen, V.F. (1983) Techniques in Teaching Vocabulary. New York: Oxford University Press.

Allen, J.P.B. and H.G. Widdowson (1974) English in Focus. English in Physical Science. London: Oxford University Press.

Allen, J.P.B. and H.G. Widdowson (1979) "Teaching the Communicative Use of English". In C.J. Brumfit and R. Johnson (eds.) The Communicative Approach to Language Teaching. Oxford: Oxford University Press.

Altenberg, B. (1989) "Corpora Analsyis". Studia Lingüística 43 (2): 167-174.

Altenberg, B. (1993) "On the Functions of such in Spoken and Written English". In N. Oostdijk and P. de Haan (eds.) Corpus-based Research into Language. Amsterdam: Rodopi.

Amsler, R.D. (1982) "Machine-readable Dictionaries and Lexical Information on Verbs". In H. Tommola el al. (eds.) Euralex' 92 Proceedings (Parts I & II). Finland: University of Tampere.

Andrews, T. (1991) "Instructional Design, a Communicative Language Learning Approach and Hypertext". On-CALL Journal 5 (3).

Anping, H. (1997) A Corpus-Based Analysis of Simultaneous Speech in English Conversation. Unpub. PH D Thesis. Victoria University of Wellington, New Zealand.

Aston, G. (1997 a) "Corpora in Language Pedagogy: Matching Theory and Practice". In G. Cook and B. Seidlhofer (eds.) Principle and Practice in Applied Linguistics. Oxford: Oxford University Press.

Aston, G. (1997 b) "Enriching the Learning Environment: Corpora in ELT". In Wichmann, A. et al. (eds.) Teaching and Language Corpora. London: Longman.

Aston, G. (1997 c) "Small and Large Corpora in Language Learning". In B. Lewandowska-Tomaszczyk and P.J. Melia (eds.) Practical Applications in Language Corpora. Lodz: Lodz University Press.

Aston, G. (1999) "Corpora and Communicative Language Learning: Cementing an Alliance". In B. Lewondowska-Tomaszczyk and P.J. Melia (eds.) Practical Applications in Language Corpora. Lodz: University of Lodz.

Atkins, B.T.S. (1985) "Monolingual and Bilingual Dictionaries: A Comparison". In R. Ilson (ed.) Dictionaries, Lexicography and Language Learning. Oxford: Pergamon Press.

Atkins, B.T.S., J. Clear and N. Ostler (1992) "Corpus Design Criteria". Literary and Linguistic Computing, Journal of the Association for Literary and Linguistic Computing 7 (1): 1-16.

Bäcklund, U. (1981) Restrictive Adjective-Noun Collocations in English. Umea: Acta Universitatis Umensis.

Badger, R. (1999) "The Student as Linguist: A Technique for Helping Students Produce Genre Specific Grammatical Descriptions". In H. Bool and P. Luford (eds.) Academic Standards and Expectations. The Role of EAP. Nottingham: Nottingham University Press.

Bahns, J. and M. Elbaw (1993) "Should We Teach EFL Students Collocations?" System 21 (1): 101-114.

Baker, M. (1988) "Sub-technical Vocabulary and the ESP Teacher. An Analysis of Some Rhetorical Items in Medical Journal Articles". Reading in a Foreign Language 4: 91-105.

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

Barlett, F.C. (1932) Remembering. Cambridge: Cambridge University Press.

Barlow, M. (1995) MonoConc. Rice University. http://www.ruf.rice.edu

Barlow, M. (1996) "Corpora for Theory and Practice". In C. Chapella (ed.) Computers and Applied Linguistics. Rice University.

Barnbrook, G. (1996) Language and Computers: A Practical Introduction to the Computer Analysis of Language. Edinburgh: Edinburgh University Press.

Battenburg, J.D. (1991) English Monolingual Learners' Dictionaries. Tübingen: Max Niemeyer Verlag.

Bauer, L. (1983) English Word Formation. Cambridge: Cambridge University Press.

Baunman, K.D. (1994) "The Significance of the Statistitical Method for an Interdisciplinary Analysis of Professionalism of Texts". In M. Brekke et al. (eds.) Applications and

Implications of Current LSP Research I. Bergen: University of Bergen, Norwegian School of Economics and Business Administration.

Beaugrande, R.D. (1993) Text, Discourse and Process. Norwood, NJ: ABLEX.

Beaugrande, R.D. and W. Dressler (1981) Introduction to Text Linguistics. London: Longman.

Béjoint, H. (1981) "The Foreign Student's Use of Monolingual Dictionaries: A Study of Language Needs and Reference Skills". Applied Linguistics 2 (3): 207-222.

Belda Medina, J.R. (1999) "La terminología informática del futuro". In A. Bocanegra, M.C. Lario de Oñate and P. López Zurita (eds.) Enfoques teóricos y prácticos de las lenguas aplicadas a las ciencias y a la tecnología. Cádiz: Tesitex.

Belmonte, G. (1997) "The Making Process of Bilingual Specialist Dictionaries". In R. Marín and A. Romero (eds.) Lenguas aplicadas a las ciencias y la tecnología. Aproximaciones. Cáceres: Universidad de Extremadura.

Bennet, J. et al. (1998) Doing Effective Presentations in an Intercultural Setting. Wien: Ueberreuter, Community of European Management Schools.

Benson, M. (1985) "A Combinatory Dictionary of English". Dictionaries 7: 59-70.

Benson, J.D. and W.S. Greaves (1981) "Field of Discourse: Theory and Application". Applied Linguistics 2 (1): 45-55.

Benson, M. and E. Benson (1988) "Trying out a New Dictionary". TESOL Quarterly 22 (2): 340-345.

Benson, M., E. Benson and R. Ilson (1997) The BBI Dictionary of English Word Combinations. Amsterdam: John Benjamins.

Bergenholtz, H. and S. Tarp (eds.) (1995) Manual of Specialised Lexicography. Amsterdam: John Benjamins.

Bernadini, S. (1998) "Systematising Serenpidity: Proposals for Large Corpora Concordancing with Language Learners". In E. Wichmann et al. (eds.) Teaching and Language Corpora. TALC' 98. Oxford: Keble College, Humanities Computing Unit.

Berri, A.J. (1996) Contribution a la mèthode d'exploration contextuelle: Applications au résumé automatique et aux représentations temporelles. Réalisation informatique du système SERAPHIN. Unpub. PH D Thesis. ISHA, Paris-Sorbonne. http://www.ifi.unizh.ch/staff/berri.html

Berry-Rogghe, G.L.M.G.A. (1970) Collocations: Their Computation and Semantic Significance. PH. D. Thesis. Manchester: University of Manchester.

Bhatia, V.K. (1993) Analysing Genre: Language Use in Professional Settings. London: Longman.

Bhatia, V.K. (1998) "Generic Conflicts in Academic Discourse". In I. Fortanet, S. Posteguillo, J.C. Palmer and J.F. Coll (eds.) Genre Studies in English for Academic Purposes. Castellón: Publicacions de la Universitat Jaume I.

Biber, D. (1988) Variation across Speech and Writing. Cambridge: Cambridge University Press.

Biber, D. (1993) "Representativeness in Corpus Design". Literary and Linguistic Computing 8: 243-257.

Biber, D., S. Conrad and R. Reppen (1994) "Corpus-based Approaches to Issues in Applied Linguistics". Applied Linguistics 15 (2): 169-189.

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

Biber, D., et al. (1999) The Longman Grammar of Spoken and Written English. London: Longman.

Bird, N. (1994) "Investigating Lexis Beyond the Most Frequent Words – Part 2". Language and Learning. Annual International Language in Education Conference. Hong Kong.

Blanco Gómez, M.L. and R. Henderson Osborne (1997) "La importancia de la formación de palabras en IFE". In R. Marín and A. Romero (eds.) Lenguas aplicadas a las ciencias y la tecnología. Aproximaciones. Cáceres: Universidad de Extremadura.

Bod, R. (1998) Beyond Grammar: An Experience-Based Theory of Language. Stanford, CA: CSLI.

Boecker, K. And P.C. Brown (1993) Oxford English for Computing. Oxford University Press.

Boguraev, B. (1991) "Building a Lexicon: The Contribution of Computers". International Journal of Lexicography 4 (3): 227-260.

Bowker, L. (1995) A Multidimensional Approach to Classification in Terminology. Working within a Computational Framework. Unpub. PH D Thesis. University of Manchester Institute of Science and Technology.

Brazil, D. (1985) The Communicative Value of Intonation in English. ELR. Birmingham: University of Birmingham.

Brekke et al. (eds.) (1994) Applications and Implications of Current LSP Research. Bergen: University of Bergen.

Brennan, M. and M. Van Naerssen (1989) "Language and Content in ESP". ELT Journal 43 (3): 196-205.

British Council, The (1980) Reading and Thinking in English. Oxford: Oxford University Press.

British Council, The (1985) Reading and Thinking in English. Oxford: Oxford University Press.

Brown, D. (1974) "Advanced Vocabulary Teaching: The Problem of Collocation". RELC Journal 5 (2): 1-11.

Brumfit, C. (1978) "`Communicative´ Language Teaching: An Assessment". In Honour of A.S. Hornby. Oxford: Oxford University Press.

Brumfit, C.J. (1984) Communicative Methodology in Language Teaching. Cambridge: Cambridge University Press.

Brumfit, C.J. (1997) "Teacher Professionalism and Research". In G. Cook and B. Seidlhofer (eds.) Principle and Practice in Applied Linguistics. Oxford: Oxford University Press.

Burgmeier, A., G. Elfred and C.B. Zimerman (1991) Lexis. Academic Vocabulary Study. Englewood Cliffs: Prentice Hall.

Burnett, F.D. (1988) Towards a Taxonomy of Items for Listening Comprehension in English for Academic Purposes. PH D Thesis. University of Exeter.

Bygate, M. (1987) Speaking. Oxford: Oxford University Press.

Calzolari, N. (1997) "Lexicon and Corpus: A Multi-faceted Interaction". Cicle de Conferènces 95-96. Lèxic, Corpus i Dictionaris. Barcelona: Institut Universitari de Lingüística Aplicada.

Candlin, C.N. (1976) "Communicative Language Teaching and the Debt to Pragmatics". In C. Rameh (ed.) Semantics: Theory and Application. Washington D.C.: Georgetown University Press.

Candlin, C.N. (1997) "Preface". In R. Carter and M. McCarthy (eds.) Vocabulary and Language Teaching. London: Longman.

Carlson, G.N. and M.K. Tannenhaus (1984) "Lexical Meanings and Concepts". In D. Testen et al. (eds.) Papers from the Parasession on Lexical Semantics. Chicago: Chicago Linguistic Society.

Carnine, D. et al. (1984) "Utilization of Contextual Information in Determining the Meaning of Unfamiliar Words". Reading Research Quarterly 19 (2): 188-294.

Carpintero, N. and M.P. Vila (1991) "Términos técnicos comparados en ingeniería nuclear". In M. McGinity, R. Alejo and S. Gómez (eds.) Lenguas aplicadas a las ciencias y tecnología II. Badajoz: Universidad de Extremadura.

Carrell, P.L. (1982) "Cohesion is not Coherence". TESOL Quarterly 16: 478-488.
Carrell, P.L. (1987) "Content and Formal Schemata in ESL Reading". TESOL Quarterly 21 (3): 461-481.

Carrell, P.L. and J.C. Eisterhold (1983) "Schema Theory and ESL Reading Pedagogy". TESOL Quarterly 17 (4): 553-573.

Carter, R.A. (1987 a) Vocabulary: Applied Linguistics Perspectives. London: Allen and Unwin.

Carter, R.A. (1987 b) "Is there a Core Vocabulary? Some Implications for Language Teaching". Applied Linguistics 8 (2): 178-193.

Carter, R.A. (1997) "Vocabulary, Cloze and Discourse. An Applied Linguistic View". In R. Carter and M. McCarthy (eds.) Vocabulary and Language Teaching. London: Longman.

Carter, R. (1998) Vocabulary. Applied Linguistic Perspectives. London: Routledge.

Carter, R. and M. N. Long (1987) The Web of Words. Cambridge: Cambridge University Press.

Carter, R.A. and M. McCarthy (eds.) (1988, 1997) Vocabulary and Language Teaching. London: Longman.

Carter, R. and M. McCarthy (1997 a) "Study of Vocabulary for Teaching". In R. Carter and M. McCarthy (eds.) Vocabulary and Language Teaching. London: Longman.

Carter, R. and M. McCarthy (1997 b) "Lexis and Discourse: Vocabulary in Use". In R. Carter and M. McCarthy (eds.) Vocabulary and Language Teaching. London: Longman.

Castro, A. (1999) "La educación del siglo XXI". El Mundo 18 nov. 1999. P. 76.

Catford, J.C. (1965) A Linguistic Theory of Translation. An Essay in Applied Linguistics. Oxford: Oxford University Press.

Celce-Murcia, M. (1997) "Discourse Analysis and the Teaching of Listening". In G. Cook and B. Seidlhofer (eds.) Principle and Practice in Applied Linguistics. Oxford: Oxford University Press.

Chan, Y.P. (1995) "Language in Lecturing: A Study of Discourse Markers in Computer Science and Information Systems Lectures". Unp. M.A. Thesis. City University of Hong Kong.

Chandler, B. (1989, 1993) MiniConcordancer. Harlow: Longman Group Ltd.

Cheng-yu, F. (1993) “Building a Corpus of the English of Computer Science”. In J. Aarts, P. de Haan and N. Oostdijk (eds.) English Language Corpora: Design, Analysis and Exploitation. Amsterdam: Rodopi.

Chomsky, N. (1975) Reflections on Language. London: Temple Smith.

Church, K. and P. Hanks (1990) “Word Association Norms, Mutual Information, and Lexicography”. Computational Linguistics 16 (1): 22-29.

Clark, E.V. (1993) The Lexicon in Acquisition. Cambridge: Cambridge University Press.

Clear, J. et al. (1996) “CoBuild: The State of the Art”. International Journal of Corpus Linguistics 1 (2): 305-316.

Collin, S. (1998) “Preface”. The Dictionary of Personal Computing and the Internet. London: HarperCollins.

Collins, P. (1998) “ESL Grammars: An Evaluation”. In P. White (ed.) Language Teaching and Research. CLTR. University of Queensland.

Collot, M. and N. Belmore (1993) “Electronic Language: A New Variety of English”. In J. Aarts, P. de Haan and N. Oostdijk (eds.) English Language Corpora: Design, Analysis and Exploitation. Amsterdam: Rodopi.

Comfort, J. et al (1994) English for the Telecommunications Industry. Oxford: Oxford University Press.

Conrad, S. (1996) “Investigating Academic Texts with Corpus-based Techniques: An Example from Biology”. Linguistics and Education 8: 299-326.

Conrad, B. and H. Rautenhaus (1994) "Innovations in Teachers' Education: Using the Concordancer as a Means for Students at University and at School Level". Meeting of EUROCALL. Karlsruhe, Germany.

Cony-Beare, M.A. (1986) "A Database-aided exploration of some Characteristics of Computer Terminology". M.A. Dissertation. Lancaster: University of Lancaster.

Cortés Gómez, E. (1991) "Investigación operativa sobre inglés para fines específicos". In M. McGinity, R. Alejo and S. Gómez (eds.) Lenguas aplicadas a las ciencias y tecnología II. Badajoz: Universidad de Extremadura.

Cowan, J.R. (1974) "Lexical and Syntactic Resources for the Description of EFL Reading Materials". TESOL Quarterly 8: 389-399.

Cowie, A.P. (1978) "The Place of Illustrative Material and Collocations in the Design of a Learner's Dictionary". In Honour of A.S. Hornby. Oxford: Oxford University Press.

Cowie, A.P. (1981) "The Treatment of Collocations and Idioms in a Learner's Dictionary". Applied Linguistics 2 (3): 223-235.

Cowie, A.P. (1988) "Vocabulary Exercises within an Individualised Study Programme". In R. Carter and M. McCarthy (eds.) Vocabulary and Language Teaching. Harlow: Longman.

Cowie, A.P. (1989) "Language as Words: Lexicography". Special Aspects of Language. Oxford: Pergamon Press.

Cowie, A.P. (1997) "Stable and Creative Aspects of Vocabulary Use". In R. Carter and M. McCarthy (eds.) Vocabulary and Language Teaching. London: Longman.

Cowie, A.P. (1998) "Introduction". In A.P. Cowie (ed.) Phraseology. Theory, Analysis and Applications. Oxford: Clarendon Press.

Crocker, T. (1981) “Scenes of Endless Science: ESP and Education”. The ESP Teacher: Role, Development and Prospects. ELT Documents 112. London: The British Council.

Crystal, D. (1997) The Cambridge Encyclopedia of the English Language. Cambridge: Cambridge University Press.

Curado Fuentes, A. (1995) “An Analysis of Content and Function Words in English for Documentation”. Actas VI Congresso Luso-Espanhol de Linguas para Fins Especificos. Viseu: Instituto Superior Politecnico.

Curado Fuentes, A. (1997) “Inducing Special Terminology through the new Technologies in Computer and Information technology”. In R. Marín and A. Romero (eds.) Lenguas aplicadas a las ciencias y la tecnología. Aproximaciones. Cáceres: Universidad de Extremadura.

Curado Fuentes, A. (1999) “Using Video in Academic Settings to Prepare Students for Oral Technical Reports in International Businesses”. In R. Bell (ed.) International Negotiation: Communication across Business Cultures. Barcelona: ESADE.

Danes, F. (1989) “Report of Roger G. Van de Velde’s Paper `Man, Verbal Text, Inferencing, and Coherence´”. In W. Heydrich et al. (eds.) Connexity and Coherence. Analysis of Text and Discourse. New York: Walter de Gruyter.

Davies, F. (1988) “Designing a Writing Syllabus in EAP: Process and Product”. In P.C. Robinson (ed.) Academic Writing: Process and Product. ELT Documents 129. Birmingham: University of Birmingham.

De Cock, S. (1998) “A Recurrent Word Combination Approach to the Study of Formulae in the Speech of Native and Non-native Speakers of English”. International Journal of Corpus Linguistics 3 (1): 59-80.

De Cock, S., et al. (1998) “An Automated Approach to the Phrasicon of EFL Learners”. In S. Granger (ed.) Learner English on Computer. London: Addison Wesley Longman.

Deighton, L. (1959) Vocabulary Development in the Classroom. New York: Bureau of Publications, Teachers College, Columbia University.

Demarty-Warée, J. (1998) "Le Travail sur Corpus: Une Entrée par les Paradigmes désignationnels". I Jornades Catalanes sobre Llengües per a Finalitats Especifiques. Barcelona: Universitat de Barcelona.

Di Pietro, R.J. (ed.) (1982) Linguistics and the Professions. Proceedings of the 2nd Annual Delaware Symposium on Language Studies. Norwood, NJ: ABLEX.

Dixon, R.M.W. (1964) "On Formal and Contextual Meaning". Acta Linguistica 14: 23.

Dudley-Evans, T. (1994) "Variations in the Discourse Patterns favoured by Different Disciplines and their Pedagogical Implications". In J. Flowerdew (ed.) Academic Listening. Research Perspectives. Cambridge: Cambridge University Press.

Dudley-Evans, T. (1998) "Introduction". In In I. Fortanet, S. Posteguillo, J.C. Palmer and J.F. Coll (eds.) Genre Studies in English for Academic Purposes. Castellón: Publicacions de la Universitat Jaume I.

Dudley-Evans, A. and T.F. Johns (1980) "An Experiment in Team-Teaching of Overseas Postgraduate Students of Transportation and Plant Biology". The Teaching of Listening Comprehension. ELT Documents. 106. ETIC. London: The British Council.

Dudley-Evans, T. and M.J. St. Johns (1998) Developments in ESP. A Multidisciplinary Approach. Cambridge: Cambridge University Press.

Dunkel, P.A. and J.N. Davis (1994) "The Effects of Rhetorical Signaling Cues on the Recall of English Lecture Information by Speakers of English as a Native or Second Language". In J. Flowerdew (ed.) Academic Listening. Research Perspectives. Cambridge: Cambridge University Press.

Duque García, M.M. and A. Ibáñez Ramos (1994) English Texts for Telecommunication Engineering. Madrid: Servicio de Publicaciones ETSI.

Durán Escribano, P. (1998) "Teaching Writing through Reading: A Text-centred Approach". Ibérica 1: 55-62.

Edwards, M. (1985) Dictionary of Keywords. London: MacMillan.

Edwards, P. (1991) "L1, General English and the Rhetoric of ESP". In M. McGinity, R. Alejo and S. Gómez (eds.) Lenguas aplicadas a las ciencias y tecnología II. Badajoz: Universidad de Extremadura.

Edwards, P. (1996) "The LSP Teacher: To Be or not to Be? That is the Question". AELFE 0: 9-25.

Eggins, S., P. Wignell and V.R. Martin (1987) "The Discourse of History: Distancing the Recoverable Past". Working Papers in Linguistics 5: 25-65.

Eldridge, R.C. (1911) Six Thousand Common English Words: Their Comparative Frequency and what can be Done with Them. Buffalo, NY: Clement Press.

Ellis, J. (1966) "On Contextual Meaning". In C.E. Bazell et al. (eds.) In Memory of J.R. Firth. London: Longman.

Ellis, J. and J.N. Ure (1969) "Language Varieties: Register". In A.R. Meeltham (ed.) Encyclopaedia of Linguistics, Information and Control. London: Pergamon International Ed.

Ewer, J.R. (1979) The Modals in Formal Scientific Discourse: Function, Meaning and Use. Santiago, Chile: Research Report, Dept. of English.

Ewer, J.R. (1981) "Formal Written and Oral Scientific English: Main Microacts (functions / notions) and their Indicators (Principal Exponents)". Mimeo. Santiago: University of Chile.

Ewer, J. (1983) “Teacher Training for EST: Problems and Methods”. The ESP Journal 2: 9-31.

Ewer, J.R. and E. Hughes-Davies (1972) “Further Notes on Developing an English Programme for Students of Science and Technology”. ELT 26: 1 & 3.

Fanning, P. (1993) “Broadening the ESP Umbrella”. ESP 12: 159-170.

Farghal, M. and H. Obiedat (1995) “Collocations: A Neglected Variable in EFL”. IRAL 33 (4): 315-331.

Fernández, R.L. (1990) “ESP, Interdisciplinarity and their Application to the Syllabus Designed for the Students of E.U.I.T. Forestal”. In M. McGinity, R. Alejo and S. Gómez (eds.) Lenguas aplicadas a las ciencias y tecnología I. Badajoz: Universidad de Extremadura.

Fielding, R. (1997) “Students’ Use of Lexical Phrases in their Written Work as Indicators of Degrees of Exposure to the Target Language”. Unpub. M.A. Thesis. Language Studies Unit. Birmingham: Aston University.

Fillmore, C.J. (1969) “Toward a Modern Theory of Case”. In D.A. Reibel and S.A. Shane (eds.) Modern Studies in English. Englewood Cliffs, NJ: Prentice Hall.

Finochiaro, M. and C. Brumfit (1983) Functional-Notional Approach. From Theory to Practice. Oxford: Oxford University Press.

Firbas, J. (1966) “Non-Thematic Subjects in Contemporary English”. Travaux Linguistiques de Prague 2: 239-256.

Firth, J.R. (1930) Speech and the Tongues of Men. London: Oxford University Press.

Firth, J.R. (1951) “Modes of Meaning”. Papers in Linguistics 3: 190-215. Oxford: Oxford University Press.

Firth, J.R. (1957) "A Synopsis of Linguistic Theory. 1930-1955". In J.R. Firth (ed.) Studies in Linguistic Analysis. Oxford: Basil Blackwell.

Flowerdew, J. (1991) "Corpus-based Course Design". In J.C. Milton and K.S.T. Tong (eds.) Text Analysis in Computer Assisted Language Learning Applications, Qualifications and Developments. Hong Kong: Language Centre, Hong Kong University of Science and Technology.

Flowerdew, J. (1992) "Definitions in Science Lectures". Applied Linguistics 13: 202-221.

Flowerdew, J. (1993 a) "Content-Based Language Instruction in a Tertiary Setting" ESP 12: 121-138.

Flowerdew, J. (1993 b) "Concordancing as a Tool in Course Design". System 21 (2): 231-244.

Flowerdew, J. (1994) "Research of Relevance to L2 Lecture Comprehension". In J. Flowerdew (ed.) Academic Listening. Research Perspectives. Cambridge: Cambridge University Press.

Flowerdew, J. and L. Miller (1997) "The Teaching of Academic Listening Comprehension and the Question of Authenticity". English for Specific Purposes 16 (1): 27-46.

Flowerdew, J. and S. Tauroza (1995) "The Effect of Discourse Markers on Second Language Lecture Comprehension". Studies in Second Language Acquisition 17: 435-458.

Flowerdew, L. (1995) "Designing CALL Courseware for an ESP Situation: A Report on a Case Study". ESP 14 (1): 19-33.

Fodor, J.A. (1983) The Modularity of Mind. Cambridge: M.I.T. Press.

Foltz, P.W. (1999) "Comprehension, Coherence and Strategy in Hypertext and Linear Text". http://www-psych.nmsu.edu/~pfoltz/reprints/ht-Cognition.html#Heading10.

Fontanelle, T. (1995) Turning a Bilingual Dictionary into a Lexical-Semantic Database. Unpub. Ph D Thesis. University of Liège, Belgium.

Fortanet, I., J.C. Palmer and S. Posteguillo (1998) "Netvertising: Content-Based Subgeneric Variations in a Digital Genre". In R.H. Sprague (ed.) Annual International Conference on Systems Science. VI. IEEE.

Foucou, P.Y. and N. Kübler (1998) "Building a Web-Based Environment for Teaching Technical English to Non-Native Speakers". In E. Wichmann et al. (eds.) Teaching and Language Corpora. TALC' 98. Oxford: Keble College, Humanities Computing Unit.

Fowler, W.S. (1989) Progressive Writing Skills. Hong Kong: Nelson.

Francis, G. (1986) Anaphoric Nouns. ELR. Birmingham: University of Birmingham.

Francis, W.N. (1992) "Language Corpora B.C.". In J. Svartvik (ed.) Directions in Corpus Linguistics. Proceedings of the Nobel Symposium 82. Berlin: Mouton de Gruyter.

Francis, G. (1993) "A Corpus-Driven Approach to Grammar: Principles, Methods and Examples". In M. Baker et al. (eds.) Text and Technology: In Honour of John Sinclair. Amsterdam: Benjamins.

Francis, G., S. Hunston and E. Manning (1998) Collins CoBuild Grammar Patterns 1: Verbs. London: HarperCollins.

Friedman, T.L. (1999) "Y ahora, la e-enseñanza". El Mundo. 18 nov 1999. P. 76.

García Berzosa, M.J. and A. Curado Fuentes (2000 – forthcoming) "Degrees of Effectiveness in ESP Presentations given by University Students in Spain". In A. Del Campo-Cuk et al.

(eds.) One Community and Many Identities. On the Crossroads of a New Europe. 9th Annual SIETAR Europa Congress 1999. Trieste: SIETAR Europa, the Society for Intercultural Education, Training and Research.

García Núñez, J.M. (1999) "Oposiciones léxico-semánticas de carácter deóntico y adverbios normativos del inglés jurídico". In A. Bocanegra, M.C. Lario de Oñate and P. López Zurita (eds.) Enfoques teóricos y prácticos de las lenguas aplicadas a las ciencias y a la tecnología. Cádiz: Tesitex.

Gavioli, L. (1997) "Exploring Texts through the Concordancer: Guiding the Learner". In Wichmann, A. et al. (eds.) (1997) Teaching and Language Corpora. London: Longman.

Gitsaki, C. (1996) The Development of ESL Collocational Knowledge. PH D Thesis. University of Queensland. http://www.cltr.uq.oz.au:8000/users/christina.gitsaki

Glässer, R. (1998) "The Stylistic Potential of Phraseological Units in the Light of Genres Analysis". In A.P. Cowie (ed.) Phraseology. Theory, Analysis and Applications. Oxford: Clarendon Press.

Gledhill, C.J. (1995) Scientific Innovation and the Phraseology of Rhetoric Posture, Reformulation and Collocation in Cancer Research Articles. PH D Thesis. Language Studies Unit. Birmingham: Aston University.

Gonzalo, J. (1999) "Del Desktop al Webtop". El Mundo. 6 Dic 1999. P. 51.

Granger, S. (1998) "Prefabricated Patterns in Advanced EFL Writing: Collocations and Formulae". In A.P. Cowie (ed.) Phraseology. Theory, Analysis and Applications. Oxford: Clarendon Press.

Greenhall, G.M. (1981) "The EST Teacher: A Negative View". ELT Documents 112.

Greenbaum, S. (1991) "ICE: The International Corpus of English". English Today 28: 3-7.

Gregg, K.R. (1989) “Second Language Acquisition Theory”. In S.M. Gass and J. Schachter Linguistic Perspectives on Second Language Acquisition. Cambridge: Cambridge University Press.

Gregory, M. (1975) The Description of English. Toronto: University of Toronto.

Grellet, F. (1981) Developing Reading Skills. New Directions in Language Teaching. Cambridge: Cambridge University Press.

Gutschow, H. (1978) “Pattern-orientated Work in ELT”. In Honour of A.S. Hornby. Oxford: Oxford University Press.

Haarman, L., P. Leech and J. Murray (1988) Reading Skills for the Social Sciences. Oxford: Oxford University Press.

Halliday, M.A.K. (1966) “Lexis as a Linguistic Level”. In C.E. Bazell et al. (eds.) In Memory of J.R. Firth. London: Longman.

Halliday, M.A.K. (1970) A Course in Spoken English: Intonation. London: Oxford University Press.

Halliday, M.A.K. (1978) Language as Social Semiotic. London: Edward Arnold.

Halliday, M.A.K. (1979) “Towards a Sociological Semantics”. In C.J. Brumfit and R. Johnson (eds.) The Communicative Approach to Language Teaching. Oxford: Oxford University Press.

Halliday, M.A.K. (1985) An Introduction to Functional Grammar. London: Edward Arnold.

Halliday, M.A.K. (1990) “A World of Meaning: A Challenge to Applied Linguistics”. Journal of Applied Linguistics 6: 7-36.

Halliday, M.A.K. (1991) "Corpus Studies and Probabilistic Grammar". In K. Aijmer and B. Altenberg (eds.) English Corpus Linguistics. London: Longman.

Halliday, M.A.K., A. McIntosh and P.D. Strevens (1965) The Linguistic Sciences and Language Teaching. London: Longmans.

Halliday, M.A.K. and R. Hasan (1976) Cohesion in English. London: Longman.

Halliday, M.A.K. and R. Hasan (1985) Language, Context, and Text: Aspects of Language in a Social-semiotic perspective. Geelong, Victoria: Deakin University.

Harvey, P. (1998) "Self-access in EAP". In G. Langley (ed.) BALEAP PIM – Professional Interest Meetings. London: South Bank University.

Hasan, R. (1984) "Coherence and Cohesive Harmony". In J. Flood (ed.) Understanding Reading Comprehension. Newark, Delaware: International Reading Association.

Haviland, S.E. and H.H. Clark (1974) "What's New? Acquiring New Information as a Process in Comprehension". Journal of Verbal Learning and Verbal Behavior 13: 512-521.

Hayes, M. (1998) "ESP Course Design: Teaching Medical Students EAP instead of EOP". I Jornades Catalanes sobre Llengües per a Finalitats Especifiques. Barcelona: Universitat de Barcelona.

Heap, S. (1993) "Word Processing for EAP Students: Starting from Scratch". On-CALL Online Journal 7 (2).

Heimbach, R. (1993) The Effect of Interaction on the Comprehension and Acquisition of New Lexical Items by Kindergarden ESL Learners. PH D Thesis. Temple University. http://www.tuj.ac.jp/cite/cv-rh.html

Henderson Osborne, R. and R. Feito Higueruela (1999) "Grammar Instruction in the ESP Classroom". In A. Bocanegra, M.C. Lario de Oñate and P. López Zurita (eds.) Enfoques teóricos y prácticos de las lenguas aplicadas a las ciencias y a la tecnología. Cádiz: Tesitex.

Henning, G.H. (1973) "Remembering Foreign Language Vocabulary: Acoustic and Semantic Parameters". Language Learning 23: 185-196.

Herbert, A.J. (1965) The Structure of Technical English. London: Longman.

Hewings, M. (1988) "The Individualization of Pronunciation Improvement". In A. Brookes and P. Grundy (eds.) Individualization and Autonomy in Language Learning. ELT Documents 131.

Hick, S. (1991) English for Information Systems. Englewood Cliffs, NJ: Prentice Hall.

Higgins, J. (1967) "Hard Facts (Notes on Teaching English to Science Students". English Language Teaching 21 (1): 1-35.

Higgins, J. (1991) "Which Concordancer? A Comparative Review of MS-DOS Software". System 19 (1 / 2): 91-100.

Higgins, J. and T. Johns (1984) Computers in Language Learning. London: Collins ELT.

Hill, A.A. (1958) Introduction to Linguistic Structures: From Sound to Sentence in English. New York: Harcourt, Brace and World.

Hill, C.P. (1985) "Alternatives to Dictionaries". In R. Ilson (ed.) Dictionaries, Lexicography and Language Learning. Oxford: Pergamon Press.

Hill, D.J. (1991) "Interlanguage Lexis: An Investigation of Verb Choice". Edinburgh Working Papers in Linguistics 2: 24-36.

Hill, M.M. (1996) "What's in a Word? Enhancing English Vocabulary Development by Increasing Learner Awareness". In P. Storey, V. Berry, D. Bunton and P. Hoare (eds.) Issues in Language in Education. Hong Kong: Hong Kong Institute of Education.

Hill, L.A. and S. Chohan (1978) From Comprehension to Composition. London: Evans Brothers Ltd.

Hindmarsh, R.X. (1978) "Some Observations on the Construction and Use of an Intermediate Level Lexicon of English". In Honour of A.S. Hornby. Oxford: Oxford University Press.

Hoey, M. (1983) On the Surface of Discourse. London: George Allen & Unwin.

Hoey, M. (1988) "The Clustering of Lexical Cohesion in Non-Narrative Text". Trondheim Papers in Applied Linguistics 4: 154-180.

Hoey, M. (1991) Patterns of Lexis in Text. Oxford: Oxford University Press.

Hoey, M. (1997) "From Concordance to Text Structure: New Uses for Computer Corpora". In B. Lewandowska-Tomaszczyk and P.J. Melia (eds.) Practical Applications in Language Corpora. Lodz: Lodz University Press.

Hoey, M. (1998) "From Word to Text Structure: A Preliminary Investigation into an Unusual Text from a Corpus Perspective". In E. Wichmann et al. (eds.) Teaching and Language Corpora. TALC' 98. Oxford: Keble College, Humanities Computing Unit.

Hogan-Brun, G. and U. Jung (eds.) (1999) Media, Multimedia, Omnimedia: Selected Papers from the CETaLL Symposium on the Occasion of the 11th AILA World Congress in Jyväskyla (Finland) and the Vth Man and the Media Symposium in Nancy (France). Frankfurt am Main, Berlin, Bern, New York, Paris, Wien: Peter Lang GmbH, Europäischer Verlag der Wissenschaften.

Holliday, A. and T. Cooke (1982) “An Ecological Approach to ESP”. Issues in ESP. Lancaster Practical Papers in English Language Education 5: 124-143.

Holmes-Higgins, P. et al. (1993, 1998) System Quirk. Language Engineering Workbench 2.0. http://www.mcs.surrey.ac.uk/SystemQ/indexv20.html

Hornby, A.S. (1974) “Verb Patterns”. In H. Gutschow (ed.) Englisch-Didaktik, Methodik, Sprache, Landeskunde. Berlin: Cornelsen-Velhagen & Klasing.

Horst, M., T. Cobb and P. Meara (1998) “Beyond a Clockwork Orange. Acquiring Second Language Vocabulary through Reading”. Reading in a Foreign Language 11 (2): 207-223.

Howarth, P. (1995) A Computer-Assisted Study of Collocations in Academic Prose, with Special Reference to Grammatical Structure and Stylistic Value. Unpub. PH D Thesis. University of Leeds.

Howarth, P..A. (1996) Phraseology in English Academic Writing. Series Maior 75. Lexicographica. Tübingen: Max Niemeyer Verlag.

Howarth, P. (1998) “Phraseology and Second Language Proficiency”. Applied Linguistics 19 (1): 24-44.

Howatt, A.P.R. (1985) A History of English Language Teaching. Oxford: Oxford University Press.

Huddlestone, R.D. (1971) The Sentence in Written English. Cambridge: Cambridge University Press.

Hudson, R.A. (1972) “Words and Thesaurus”. Journal of Documentation 28: 76-77.

Hudson, R.A. (1994) “About 37% of Word-Tokens are Nouns”. Language 70: 331-339.

Hunston, S. and G. Francis (1998) “Verbs Observed: A Corpus-Driven Pedagogic Grammar”. Applied Liguistics 19 (1): 45-72.

Hunston, S. and G. Francis (2000) A Corpus-Driven Approach to the Lexical Grammar of English Studies in Corpus Linguistics 4. Amsterdam: John Benjamins.

Hunt, A. (1997) “Evaluating Bilingual and Monolingual Dictionaries for L2 Learners”. Kansar Gadai University Journal of Inquiry and Research 65: 15-27.

Hutchinson, T. and A. Waters (1980) “ESP at the Crossroads”. ESP 36.

Hutchinson, T. and A. Waters (1981) “Performance and Competence in ESP”. Applied Linguistics 2 (1): 56-69.

Hutchinson, T. and A. Waters (1987) English for Specific Purposes. A Learning-Centred Approach. Cambridge: Cambridge University Press.

Ilson, R. (1985) Dictionaries, Lexicography and Language Learning. ELT Documents 120. Oxford: Pergamon Press.

“Informe `Universidad 2000´”. El País. 27 March 2000. P. 43

Ingria, R., B. Boguraev and J. Pustejovski (1992) “Dictionary / Lexicon”. In W. Shapiro (ed.) Encyclopedia of Artificial Intelligence. New York: John Wiley.

Inkster, G. (1997) “First Catch your Corpus: Building a French Undergraduate Corpus from Readily Available Textual Resources”. In Wichmann, A. et al. (eds.) Teaching and Language Corpora. London: Longman.

Inman, M. (1978) “Lexical Analysis of Scientific and Technical Prose”. In T. Trimble, L. Trimble and K. Drobnic (eds.) English for Specific Purposes: Science and Technology. EL Institute. Oregon State University.

Jackson, J. and L. Bilton (1994) "Stylistic Variations in Science Lectures: Teaching Vocabulary". ESP 13 (1): 61-80.

Jarvis, S.H. (1997) The Role of L1-Based Concepts in L2 Lexical Reference. Unpub. PH D. Indiana University.

Johansson, S. (1975) Some Aspects of the Vocabulary of Learned and Scientific English. Gothenburg Studies in English 42.

Johansson, S. (1978) Some Aspects of the Vocabulary of Learned and Scientific English. Gothenburg: Acta Universitatis Gothoburgensis.

Johansson, S. (1985) "Word Frequency and Text Type: Some Observations Based on the LOB Corpus of British English Texts". Computers and the Humanities 19: 23-36.

Johansson, S. and A. Stenström (eds.) (1991) English Computer Corpora: Selected Papers and Bibliography. Berlin: Mouton de Gruyter.

Johns, A.M. (1997) Text, Role and Context. Cambridge: Cambridge University Press.

Johns, T. (1986) "Micro-Concord: A Language Learner's Research Tool". System 14 (2): 151-162.

Johns, T. (1991) "Should you be Persuaded: Two Examples of data-driven Learning". Classroom Concordancing. English Language Research Journal 4: 1-16.

Johns, T. (1993) "Data-driven Learning: An Update". TELL & CALL 3: 23-32.

Johns, T. (1997) "Contexts: The Background, Development and Trialling of a Concordance-Based CALL Program". In Wichmann, A. et al. (eds.) (1997) Teaching and Language Corpora. London: Longman.

Johns, T. and M. Scott (1993) MicroConcord Concordancing Software and Manual. Oxford: Oxford University Press.

Jolliffe, D. and E. Brier (1988) "Studying Writer's Knowledge in Academic Disciplines". In D.A. Jolliffe (ed.) Advances in Writing Research v.2: Writing in Academic Disciplines. Norwood, NJ: ABLEX.

Jones, S. and J. Sinclair (1974) "English Lexical Collocations". Cahiers de Lexicologie 24: 15-61.

Jones, R.L (1997) "Creating and Using a Corpus of Spoken German". In Wichmann, A. et al. (eds.) (1997) Teaching and Language Corpora. London: Longman.

Jordan, R.R. (1997) English for Academic Purposes. Cambridge: Cambridge University Press.

Judd, E.L. (1978) "Vocabulary Teaching and TESOL: A Need for Reevaluation of Existing Assumptions". TESOL Quarterly 12 (1): 71-76.

Kaeding, J.W. (1897) Häufigkeitswörterbuch der Deutschen Sprache. Steglitz: Der Herausgeber.

Kayser, H. (1989) "Some Aspects of Language Understanding, Language Production, and Intercomprehension in Verbal Interaction". In W. Heydrich et al. (eds.) Connexity and Coherence. Analysis of Text and Discourse. New York: Walter de Gruyter.

Kelly, P. (1990) "Guessing No Substitute for Systematic Learning of Lexis". System 18 (2): 199-207.

Kelly, P. (1991) "Lexical Ignorance: The Main Obstacle to Listening Comprehension with Advanced Foeign Language Learners". IRAL XXIX (2): 135-149.

Kennedy, G. (1992) "Preferred Ways of Putting Things with Implications for Language Teaching". In J. Svartvik (ed.) Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82. Berlin: Mouton de Gruyter.

Kennedy, C. and R. Bolitho (1984) English for Specific Purposes. ELTS. London: MacMillan.

Kettermann, B. (1997) "Using a Corpus to Evaluate Theories of Child Language Acquisition". In Wichmann, A. et al. (eds.) (1997) Teaching and Language Corpora. London: Longman.

Khaled, T.B. (1984) "Linguistic Analysis as an Input to Course Design". Proceedings of the Annual Linguistics Conference. Jordan University.

King, P. (1989) "The Uncommon Core. Some Discourse Features of Student Writing". System 17 (1): 14-28.

King, P. (1994) "Visual and Verbal Messages in the Engineering Lecture: Notetaking by Postgraduate L2 Students". In J. Flowerdew (ed.) Academic Listening. Research Perspectives. Cambridge: Cambridge University Press.

Kita, K. and H. Ogata (1997) "Collocations in Language Learning. Corpus-Based Automatic Compilation of Collocations and Bilingual Collocation Concordancer". Computer Assisted Language Learning 10 (3): 229-238.

Kjellmer, G. (1984) "Some Thoughts on Collocational Distinctiveness". In J. Aarts and W. Meijs (eds.) Corpus Linguistics. Amsterdam: John Benjamins.

Kramsch, C. (1997) "The Applied Linguist and the Foreign Language Teacher: Can they Talk to each other?". In G. Cook and B. Seidlhofer (eds.) Principle and Practice in Applied Linguistics. Oxford: Oxford University Press.

Krishnamurthy, R. (1997) “Keeping Good Company: Collocation, Corpus and Dictionaries”. Cicle de Conferències 95-96. Lèxic, Corpus i Dictionaris. Barcelona: Institut Universitari de Lingüística Aplicada.

Kroonenburg, N. (1995) “Developing Communicative and Thinking Skills via Electronic Mail”. TESOL Journal Winter: 24-27.

Krzanowski, M. (1999) “Teaching Business and Academic Presentation Skills: Training, Assessment and Feedback”. ENCODE 99 Conference. Barcelona: ESADE.

Kuhn, T. (1963) “The Function of Dogma in Scientific Research”. In A.C. Crombie (ed.) Scientific Change. London: Heinemann.

Kyyssönen, H. (1997) “Grammar and Lexis in Communicative Competence”. In G. Cook and B. Seidlhofer (eds.) Principle and Practice in Applied Linguistics. Oxford: Oxford University Press.

Lackstrom, J. (1978) “Teaching Modals in EST Discourse”. In M.T. Tribble, L. Tribble and K. Drobnic (eds.) ESP: Science and Technology. Corwallis: Oregon State University.

Lackstrom, J.E., L. Selinker and L.P. Trimble (1972) “Grammar and Technical English”. English Teaching Forum X: 3-14.

Laster, A.A. and N.A. Pickett (1996) Technical English: Writing, Reading and Speaking. New York: HarperCollins.

Latour, B. and S. Woolgar (1986) Laboratory Life. The Social Construction of Scientific Facts. Princeton, NJ: Princeton University Press.

Laufer, B. (1999) “What Lexical Information do L2 Learners Select in a CALL Dictionary and How does it Affect Word Retention”. Language Learning and Technology 3 (2): 58-76.

Leaverton, A. and J. Zymantas (1999) "Taking a Lexical Look at Teaching English for Business". ENCODE 99 Conference. Barcelona: ESADE.

Leech, G. (1981) Semantics. Harmondsworth, Middlesex: Penguin. The Chaucer Press.

Leech, G. (1991) "The State of the Art in Corpus Linguistics". In K. Aijmer and B. Altenberg (eds.) English Corpus Linguistics. London: Longman.

Leech, G.N. (1992) "Corpora and Theories of Linguistic Performance". In J. Svartvik (ed.) Directions in Corpus Linguistics. Proceedings of Nobel Symposium. Berlin: Mouton de Gruyter.

Leech, G. (2000) "Preface". In S. Granger (ed.) Learner English on Computer. Austin, TX: Addison Wesley Longman.

Lehrberger, J. (1982) "Automatic Translation and the Concept of Sublanguage". In Kittredge and Lehrberger (eds.) 81-106.

Lehrer, A. (1974) Semantic Fields and Lexical Structure. Amsterdam: North Holland Pub. Co. London: Elsevier.

Leloup, J.W. and R. Ponteiro (1999) "Cooperative Learning Activities for the Foreign Language Classroom". Language Learning and Technology 3 (2): 3-5.

Lennon, P. (1991) "Error and the Very Advanced Learner". IRAL 29 (1): 31-44.

Lerchundi, M.A. and P. Moreno (1999) "Los falsos amigos en los textos técnicos". In A. Bocanegra, M.C. Lario de Oñate and P. López Zurita (eds.) Enfoques teóricos y prácticos de las lenguas aplicadas a las ciencias y a la tecnología. Cádiz: Tesitex.

Levy, M. (1999) "Helping Learners Learn and Remember New Words". EL Projects. London: The British Council.

Levy, M. (2000) Computer-Assisted Language Learning. Oxford: Clarendon Press.

Lewis, M. (1993) The Lexical Approach. London: LTP.

Lewis, M. (1997) Implementing the Lexical Approach. London: Language Teaching Pub.

Lindberg, J. (1999) Handling Lexicalised Phrases for Natural Language Processing. Unpub. PH D Thesis. Stockholm University.

Liu, J. and H. Nesi (1999) "Are We Teaching the Right Words? A Study of Students' Receptive Knowledge of Two Types of Vocabulary: `Subtechnical´ and `Technical´". In H. Bool and P. Luford (eds.) Academic Standards and Expectations. The Role of EAP. Nottingham: Nottingham University Press.

Lockett, A. (1999) "From the General to the Specific: What the EAP Tutor should know about Academic Discourse". In H. Bool and P. Luford (eds.) Academic Standards and Expectations. The Role of EAP. Nottingham: Nottingham University Press.

Louw, W. (1993) "Irony in the Text or Insincerety in the Writer? The Diagnostic Potential of Semantic Prosodies". In M. Baker et al. (eds.) Text and Technology. Amsterdam: John Benjamins.

Love, A.M. (1993) "Lexico-Grammatical Features of Geology Textbooks". English for Specific Purposes 12: 197-218.

Lowe, I. (1969) "An Algebraic Theory of English Pronominal Reference (part 1)". Semiotica 1 (4): 397-421.

Lyons, J. (1977) Semantics 1. Cambridge: Cambridge University Press.

Lyons, J. (1970) Chomsky. London: Fontane / Collins.

Lyons, J. (1995) Linguistic Semantics. Cambridge: Cambridge University Press.

Mackey, W.F. (1965) Language Teaching Analysis. London: Longman.

Mackey, W.F. (1978) "Divorcing Language from Life; Non-contextual Linguistics in Language Teaching". In Honour of A.S. Hornby. Oxford: Oxford University Press.

Mackey, W.F. and J.G. Savard (1967) "The Indices of Coverage". International Review of Applied Linguistics 5 (2/3): 71-121.

Mackin, R. (1978) "On Collocations: Words shall be Known by the Company they Keep". In P. Strevens (ed.) In Honour of A.S. Hornby. Oxford: Oxford University Press.

Makkai, A. (1980) "Theoretical and Practical Aspects of an Associative Lexicon for Twentieth Century English". In L. Zgusta (ed.) Theory and Method in Lexicography. Columbia, SC: Hornbeam Press.

Marcos Marín, F.A. (1994) Informática y Humanidades. Madrid: Gredos.

Marcos Marín, F.A. (1996) El Comentario filológico con apoyo informático. Madrid: Síntesis.

Markee, N. (1989) "ESP within a New Descriptive Framework". World Englishes 8 (2): 133-146.

Marshall, S. and M. Gilmour (1993) "Lexical Knowledge and Reading Comprehension in Papua New Guinea". ESP 12: 69-81.

Martin, A.V. (1976) "Teaching Academic Vocabulary to Foreign Graduate Students". TESOL Quarterly 10 (1): 91-97.

Martin, J.R. (1985) Factual Writing: Exploring and Challenging Social Reality. Victoria: Deakin University Press.

Martin, W. (1993) "Remarks on Collocations in Sublanguages". Commission des Communautés europénnes: Terminologie et Traduction. 2 / 3. Luxembourg: Office des publications officielles del communautés europénnes.

Martin, W. et al. (1983) "On the Processing of a Text Corpus: From Textual Data to Lexicographical Information". In R.R.K. Hartmann (ed.) Lexicography: Principles and Practice. London: Academic Press.

Martínez, M.S. (1994) "Spanish-English Cognates in the Subtechnical Vocabulary found in Engineering Magazine Texts". ESP 13: 81-92.

Mascull, B. (1997) Keywords in Science and Technology. The CoBuild Series from the Bank of English. London: HarperCollins.

Mason, A. (1994) "By Dint of: Student and Lecturer Perceptions of Lecture Comprehension Strategies in First-Term Graduate Study". In J. Flowerdew (ed.) Academic Listening. Research Perspectives. Cambridge: Cambridge University Press.

Mathies, B.K. and B. Nelson (1995) "Developing Communicative and Thinking Skills via Electronic Mail". Techtrends 40 (1): 20-28.

McArthur, T. (1978) "The Vocabulary Control Movement in the English Language, 1844-1953". Indian Journal of Applied Linguistics 4: 47-68.

McCarthy, B. (1993) "Developing CALL Materials for the Foreign Language Classroom: Ideals and Practicalities". On-CALL Online Journal 7 (2).

McCarthy, M. (1990) Vocabulary. Oxford: Oxford University Press.

McDonough, J. (1984) ESP in Perspective: A Practical Guide. London: Collins ELT.

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

McHale, M.L. (1995) Combining Machine Readable Lexical Resources with a Principle Based Parser. Unpub. PH D Thesis. Syracuse University.

McIntosh, A. (1961) "Patterns and Ranges". Language 37: 325-337.

McIntosh, A. (1963) "Language and Style". Durham University Journal 55 (3): 116-123.

McKay, S. (1980) "Teaching the Syntactic, Semantic and Pragmatic Dimensions of Verbs". TESOL Quarterly 14 (1): 17-26.

McMurrey, D. (1999) "Online Technical Writing: Oral Presentations". Online Technical Writing Course. http://www.io.com/(hcexres/tcm1603/acchtml/oral.html

McNaught, J. (1993) "User Needs for Textual Corpora in Natural Language Processing". Literary and Linguistic Computing 8: 227-234.

Meara, P.M. (1982) "Word Associations in a Foreign Language". Nottingham Linguistic Circular 11 (2): 29-38.

Meara, P.M. (1983) "International Communication and the Concept of Nuclear English". Style and Communication in the English Language 2: 37-53.

Meijs, W. (1992) "Inferences and Lexical Relations". In G. Leitner (ed.) New Directions in English Language Corpora: Methodology, Results, Software Developments. Berlin: Mouton de Gruyter.

Mel´Čuk, I. (1998) "Collocations and Lexical Functions". In A.P. Cowie (ed.) Phraseology. Theory, Analysis and Applications. Oxford: Clarendon Press.

Miller, C.R. (1984) "Genre as a Social Action". Quarterly Journal of Speech 70: 151-167.

Mills, J. (1992) Computational Analysis of Meaning from Computer Concordances of Middle to Late Cornish. Unpub. PH D Thesis. University of Exeter.

Mitchell, T. (1971) "Linguistics `going on´: Collocations and Other Lexical Matters Arising on the Syntactic Record". Archivum Linguisticum 2: 35-69.

Montemayor-Borsinger, A. (1999) "Research Articles in Science: A Comparative Study of Linguistic Choices". In H. Bool and P. Luford (eds.) Academic Standards and Expectations. The Role of EAP. Nottingham: Nottingham University Press.

Moon, R. (1984) Monosemous Words and the Dictionary. ELR. Birmingham: University of Birmingham.

Moon, R. (1997) "Vocabulary Connections: Multi-Word Items in English". In N. Schmitt and M. McCarthy (eds.) Vocabulary: Description, Acquisition and Pedagogy. Cambridge: Cambridge University Press.

Moon, R. (1998 a) Fixed Expressions and Idioms in English: A Corpus-based Approach. Oxford: Oxford University Press.

Moon, R. (1998 b) "Phrasal Lexemes in English". In A.P. Cowie (ed.) Phraseology. Theory, Analysis and Applications. Oxford: Clarendon Press.

Morrisey, G.L. and T.L. Sechrest (1987) Effective Business and Technical Presentations. Massachussetts: Addison-Wesley.

Mosenthal, P.B. (1985) "Defining the Expository Discourse Continuum: Towards a Taxonomy of Expository Text Types". Poetics 14: 387-414.

Ide, N. and J. Véronir (1994) "Multext: Multilingual Text Tools and Corpora". COLING 94: Proceedings of the 15th International Conference on Computational Linguistics 1. Paris.

Munby, J. (1978) Communicative Syllabus Design. Cambridge: Cambridge University Press.

Myers, G. (1992) "Textbooks and the Sociology of Scientific Knowledge". English for Specific Purposes 11: 3-17.

Nation, I.S.P. (1982) "Beginning to Learn Foreign Vocabulary. A Review of the Research". RELC Journal 13 (1): 14-34.

Nation, I.S.P. (1990) Teaching and Learning Vocabulary. Boston: Heinle & Heinle Pub.

Nation, P. and J. Coady (1997) "Vocabulary and Reading". In R. Carter and M. McCarthy (eds.) Vocabulary and Language Teaching. London: Longman.

Nattinger, J. (1988) "Some Current Trends in Vocabulary Teaching". In R. Carter and M. McCarthy (eds.) Vocabulary and Language Teaching. Harlow: Longman.

Nattinger, J.R. (1997) "Some Current Trends in Vocabulary Teaching". In R. Carter and M. McCarthy (eds.) Vocabulary and Language Teaching. London: Longman.

Nattinger, J.R. and J.S. DeCarrico (1988) "Lexical Phrases for the Comprehension of Academic Lectures". ESP 7: 91-102.

Nattinger, J.R. and J.S. DeCarrico (1997) Lexical Phrases and Language Teaching. Oxford: Oxford University Press.

Nolasco, R. and L. Arthur (1988) Large Classes. ELTS. London: MacMillan.

Norris, J. (1983) Language Learners and their Errors. ELTS. London: MacMillan.

Nunan, D. (1989) Designing Tasks for the Communicative Classroom. Cambridge: Cambridge University Press.

Oakes, M.P. (1998) Statistics for Corpus Linguistics. Edinburgh: Edinburgh University Press.

Olds, E. (1995) Computational Evidence for the Use of Frequency Information in Discovery of the Infant's First Lexicon. PH D Thesis. New York: City University of New York, GSUC.

Olsen, L.A. and T.N. Huckin (1990) "Point-Driven Understanding in Engineering Lecture Comprehension". ESP 9: 33-47.

Ooi, V.B.Y. (1998) Computer Corpus Lexicography. Edinburgh: Edinburgh University Press.

Orr, T. (1996) "Designing English Writing Instructions for Students in the Science and Technologies: Research Results, and Applications". Annual Meeting of the Japan Association of College English Teachers. Kyoto: Center for Language Research.

Orr, T. (1995) Models of Professional Writing Practices within the Field of Computer Science. PH D Thesis. Dissertation Abstracts International, 56 A, 4754.

Országh, L. (1969) "Wanted: Better English Dictionaries". ELT Journal 23: 216-222.

Osborne, J. (1998) "What can Students Learn from a Corpus? Building Bridges between Data and Explanation". In E. Wichmann et al. (eds.) Teaching and Language Corpora. TALC' 98. Oxford: Keble College, Humanities Computing Unit.

Owen, C. (1993) "Corpus-based Grammar and the Heineken Effect: Lexico-grammatical Description for Language Learners". Applied Linguistics 14 (2): 167-187.

Oxford Concordance Program (OCP) (1988) Oxford: O.U.P. Electronic Publications.

Ozkaya, S. (1996) "The Role of Lexical Phrases in the Macro-Structuring of Written Discourse: A Comparative Study of Medical Research Articles by Native and Non-native Speakers of English". Unpub. MSc Diss. Language Studies Unit. Birmingham: Aston University.

Palacios Pablos, A. (1997) "Las asociaciones léxicas en las lenguas aplicadas". In R. Marín and A. Romero (eds.) Lenguas aplicadas a las ciencias y la tecnología. Aproximaciones. Cáceres: Universidad de Extremadura.

Palmer, H.E. (1922) The Principles of Language Study. London: Oxford University Press.

Partington, A. (1998) Patterns and Meanings – Using Corpora for English Language Research and Teaching. Studies in Corpus Linguistics 2. Amsterdam: John Benjamins.

Pearson, J. (1998) "Teaching Students to Choose their Texts Wisely". In E. Wichmann et al. (eds.) Teaching and Language Corpora. TALC' 98. Oxford: Keble College, Humanities Computing Unit.

Pedersen, J. (1995) "The Identification and Selection of Collocations in Technical Dictionaries". Lexicographia 11: 60-73.

Peignan, A. (1999) "Corpus-Based Research into Metaphor". In L. Camero and G. Low (eds.) Researching and Applying Metaphor. Cambridge: Cambridge University Press.

Pennington, M.C. (ed.) (1996) The Power of CALL. Houston: Athelstan.

Perfetti, C. and S.R. Goldman (1974) "Thematization and Sentence Retrieval". Journal of Verbal Learning and Language Behavior 13: 71.

Peters, A. (1983) The Units of Language Acquisition. Cambridge: Cambridge University Press.

Peterson, M. (1998) "Scientific Presentation". JAIST 4. http://www.jaist.ac.jp/(mark/Article.html

Phillips, M. (1985) Aspects of Text Structure. North-Holland: Elseviers.

Phillips, M. (1989) Lexical Structure of Text. ELR. Discourse Analysis Monograph 12. Birmingham: University of Birmingham.

Pincas, A. (1982) Teaching English Writing. ELTS. London: MacMillan.

Posteguillo, S. (1995) Genre Analysis in English for Computer Science. Unpub. PH D Thesis. Castellón: Universitat Jaume I.

Posteguillo, S. (1997) Network English: English in the World of Computers. Castellón: Publicacions de la Universitat Jaume I.

Potts, T.C. (1994) Structures and Categories for the Representation of Meaning. Cambridge: Cambridge University Press.

Pueyo, I. and S. Val (1996) "The Construction of Technicality in the Field of Plastics: A Functional Approach Towards Teaching Technical Terminology". ESP 15 (4): 251-278.

Quirk, R., S. Greenbaum, G. Leech and J. Svartvik (1985) A Comprehensive Grammar of the English Language. London: Longman.

Quirk, R. (1997) Grammatical and Lexical Variance in English. London: Longman.

Rea-Dickins, P. and K. Germaine (1993) Evaluation. Oxford: Oxford University Press.

Reid, T.B.S. (1956) "Linguistics, Structuralism and Philology". Archivum Linguisticum 8: 32-45.

Renouf, A. (1984) Corpus Linguistics. Recent Advances in the Use of Computer Corpora in English Language Research. Amsterdam: Rodopi.

Renouf, A. (1988) "Corpus Development". In Sinclair, J. (ed.) Looking Up: An Account of the CoBuild Project in Lexical Computing. London: Collins.

Renouf, A. (1997) "Teaching Corpus Linguistics to Teachers of English". In Wichmann, A. et al. (eds.) (1997) Teaching and Language Corpora. London: Longman.

Richards, J.C. (1974) "Word Lists: Problems and Prospects". RELC Journal 5 (2): 69-84.

Richards, J.C. (1976) "The Role of Vocabulary Teaching". TESOL Quarterly 10 (1): 77-89.

Rico Pérez, C. (1994) "Integrating CALL Activities within the Business English Classroom: Teaching Collocation with the Help of Computers". In V. Guillén Nieto, J. Mateo Martínez and B. Moudy Boyle (eds.) English for Specific Purposes: Approaches and Strategies. Working Papers # 3. Alicante: Universidad de Alicante.

Rico García, M. (1999) "Detección de errores en la comprensión lectora mediante el análisis de la lectura en voz alta". In A. Bocanegra, M.C. Lario de Oñate and P. López Zurita (eds.) Enfoques teóricos y prácticos de las lenguas aplicadas a las ciencias y a la tecnología. Cádiz: Tesitex.

Riley, D. (1999) Vocabulary for Computing. London: Peter Collin Pub.

Robinson, P. (1980) ESP (English for Specific Purposes). Oxford: Pergamon Press.

Robinson, P. (1985) Fundamentals in Technical Writing. Boston: Houghton Mifflin Co.

Robinson, P. (1991) ESP Today: A Practitioner's Guide. London: Prentice Hall International.

Robinson, P. (1999) "The Cycle of Materials Production for EAP". Lecture at the I Jornadas sobre la enseñanza del inglés: nuevas perspectivas en investigación y didáctica. 11-13 noviembre 1999. Cádiz, Universidad de Cádiz.

Roe, P.J. (1977) Scientific Text. ELR. Monograph 4. Birmingham: University of Birmingham.

Roe, P.J. (1977) The Notion of Difficulty in Scientific Text. PH. D. Thesis. Bigmingham: University of Birmingham.

Roldán Riejos, A.M. (1998) "Applications of Cognitive Theory to Interdisciplinary Work in Languages for Specific Purposes". Ibérica 1: 29-39.

Romero Guillén, M.D. (1997) "Técnicas para un aprendizaje más eficaz del inglés científico en un contexto veterinario". In R. Marín and A. Romero (eds.) Lenguas aplicadas a las ciencias y la tecnología. Aproximaciones. Cáceres: Universidad de Extremadura.

Rosenburg, R.C. (1983) "The Engineering Presentation – Some Ideas on How to Approach and Present it". In D.F. Beer (ed.) Writing and Speaking in the Technology Professions. New York: IEEE Press.

Rosenthal, L. and S.B. Rowland (1986) Academic Reading and Study Skills for International Students. Englewood Cliffs, NJ: Prentice Hall Regents.

Ross, J. R. (1969) "Auxiliaries as Main Verbs". In W. Todd (ed.) Studies in Philosophical Linguistics 1. Evanston, Ill: Northwestern University Press.

Rossner, R. (1985) "The Learner as Lexicographer: Using Dictionaries in Second Language Learning". In R. Ilson (ed.) Dictionaries, Lexicography and Language Learning. Oxford: Oxford Pergamon Press.

Rossner, R. and R. Bolitho (eds.) (1990) Currents of Change in English Language Teaching. Oxford: Oxford University Press.

Rost, M. (1994) "On-line Summaries as Representations of Lecture Understanding". In J. Flowerdew (ed.) Academic Listening. Research Perspectives. Cambridge: Cambridge University Press.

Rudzka, B. et al. (1985) The Words You Need. London: MacMillan.

Ruhl, C. (1973) "Prerequisites for a Linguistic Description of Coherence". Language Science 25: 15.

Rundblad, G. (1997) Shallow Brooks and Rivers Wide – A Study of Lexical and Semantic Change in English Nouns Denoting `Watercourse´. Unpub. PH D Thesis. Stockholm: Department of English, Universitet Stockholm. http://www.engelska.su.se/gabriell.html

Sager, J.C., D. Dungworth and P.M. McDonald (1980) English Special Languages. Principles and Practice in Science and Technology. Wiesbaden: Brandstetter Verlag.

Sager, N. (1986) "Sublanguage: Linguistic Phenomenon, Computational Tool". In R. Grishman (ed.) Computational Linguistics: An Introduction. Cambridge: Cambridge University Press.

Salager-Meyer, F. (1983) "The Lexis of Fundamental Medical English: Classificatory Framework and Rhetorical Function (a Statistical Approach)". Reading in a Foreign Language 1 (1): 54-64.

Salager-Meyer, F. (1987) "Infinitive Clauses in Medical English Literature: A Rhetorical Grammatical Approach". EMP Newsletter 4 (1): 23-33.

Salager, F. (1990) "Metaphor in Medical English Prose. A Comparative Study with French and Spanish". ESP 9: 145-159.

Sampson, G. (1991) “Analysed Corpora of English: A Consumer Guide”. In M.C. Pennington and V. Stevens (eds.) Computers in Applied Linguistics. Philadelphia: Multilingual Matters.

Schiffrin, D. (1987) Discourse Markers. Cambridge: Cambridge University Press.

Schnotz, W., E. Picard and A. Hron (1993) “How do Successful and Unsuccessful Learners Use Texts and Graphics”. Learning and Instruction 3 : 181-199.

Scott, M. (1996) WordSmith. Oxford: Oxford University Press.

Scott, M. (1997) “PC Analysis of Key Words and Key Key Words”. System 25 (1): 1-13.

Scott, M. (1999) “Reverberations of Echoes”. In B. Lewondowska-Tomaszczyk and P.J. Melia (eds.) Practical Applications in Language Corpora. Lodz: University of Lodz.
Searle, J.R. (1969) Speech Acts. Cambridge: Cambridge University Press.

Sebba, M. (1991) “The Adequacy of Corpora in Machine Translation”. In A. McEnery (ed.) Applied Computer Translation 1 (1): 15-28.

Selinker, L., L. Trimble and R. Vroman (1972) “Report No 5, Presupposition and Technical Rhetoric”. Working Papers in English for Science and Technology. Office of Engineering Research and Dept. of Humanistic-Social Studies. Washington: University of Washington.

Selinker, L. and L. Trimble (1974) “Formal Written Communication and ESL”. Journal of Technical Writing and Communication 4: 81-90.

Selinker, L. et al. (1976) “Presuppositional Rhetorical Information in EST Discourse”. TESOL Quarterly 10: 281-290.

Simpson, R. (1998) “Methodological Challenges of Planning a Corpus with Pedagogical Outcomes”. In E. Wichmann et al. (eds.) Teaching and Language Corpora. TALC’ 98. Oxford: Keble College, Humanities Computing Unit.

Sinclair, J.M. (1965) “When is a Poem like a Sunset?” A Review of English Literature 6 (2): 76-91.

Sinclair, J. (1966) “Beginning the Study of Lexis”. In C.E. Bazell, M.A.K. Halliday and R. Robins (eds.) In Memory of J.R. Firth. Harlow: Longman.

Sinclair, J. (1985) “Lexicographic Evidence”. In R. Ilson (ed.) Dictionaries, Lexicography and Language Learning. Oxford: Oxford Pergamon Press.

Sinclair, J. (1987) “The Nature of Evidence”. In Sinclair, J. (ed.) Looking Up: An Account of the CoBuild Project in Lexical Computing. London: Collins.

Sinclair, J. (1988) “Sense and Structure in Lexis”. In J.D. Benson, M.J. Cummings and W.S. Greaves (eds.) Linguistics in a Systemic Perspective. 39. Current Issues in Linguistic Theories. Amsterdam: John Benjamins.

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

Sinclair, J. (1996) “Units of Meaning” Textus 9: 75-106.

Sinclair, J. (1997) “Corpus Evidence in Language Description”. In Wichmann, A. et al. (eds.) (1997) Teaching and Language Corpora. London: Longman.

Sinclair, J.M., S. Jones and R. Daley (1969) English Lexical Studies. University of Birmingham for the Office for Scientific and Technical Information.

Sinclair, J. and M. Coulthard (1975) Towards an Analysis of Discourse. Oxford: Oxford University Press.

Sinclair, J. and A. Renouf (1988) “A Lexical Syllabus for Language Learning”. In R. Carter and M. McCarthy (eds.) Vocabulary and Language Teaching. London: Longman.

Skehan, P. (1981) “ESP Teachers, Computers and Research”. The ESP Teacher: Role, Development and Prospects. ELT Documents 112. London: The British Council.

Slobin, D.I. (1998) “Two Ways to Travel: Verbs of Motion in English and Spanish”. In M. Shibatani and S. Thompson (eds.) Essays in Semantics. Oxford: Oxford University Press.

Soláns García, M.A. (1997) “El inglés a través de la anatomía: Algunas consideraciones sobre la enseñanza del vocabulario en un curso de inglés médico”. In R. Marín and A. Romero (eds.) Lenguas aplicadas a las ciencias y la tecnología. Aproximaciones. Cáceres: Universidad de Extremadura.

Stanley, R.M. (1984) “The Recognition of Macrostructure: A Pilot Study”. Reading in a Foreign Language 2: 156-168.

Stotsky, S. (1983) “Types of Lexical Cohesion in Expository Writing: Implications for Developing the Vocabulary of Academic Discourse”. College Composition and Communication 34 (4): 430-446.

Strang, B.M.W. (1962) Modern English Structure. London: Edward Arnold.

Strevens, P. (1971) “Alternatives to Daffodils”. Science and Technology in a Special Language. CILT Reports and Papers 7. CILT.

Strevens, P. (1977) “Special Purpose Learning: A Perspective”. Language Teaching and Linguistics: Abstracts 10 (3): 145-163.

Strevens, P. (1978) “The English Language Made Plain for the Teacher of English: The Descriptivist and the Linguistic Traditions”. In Honour of A.S. Hornby. Oxford: Oxford University Press.

Stubbs, M. (1986) “Language Development, Lexical Competence and Nuclear Vocabulary”. Educational Linguistics. Oxford: Basil Blackwell.

Stubbs, M. (1995) “Collocations and Semantic Profiles. On the cause of the Trouble with Quantitative Studies”. Functions of Language 2: 23-55.

Stubbs, M. (1996) Text and Corpus Analysis. Oxford: Blackwell Pub.

Stubbs, M. (1997) “Corpus Evidence for Norms of Lexical Collocation”. In G. Cook and B. Seidlhofer (eds.) Principle and Practice in Applied Linguistics. Oxford: Oxford University Press.

Summers, D. (1993) “Longman Lancaster English Language Corpus. Criteria and Design”. International Journal of Lexicography 6 (3): 181-208.

Summers, D. (1996) “Computer Lexicography”. In J. Thomas and M. Short (eds.) Using Corpora for Language Research. London: Longman.

Summers, D. (1997) “The Role of Dictionaries in Language Learning”. In R. Carter and M. McCarthy (eds.) Vocabulary and Language Teaching. London: Longman.

Svartvik, J. (1992) “Preface”. In Svartvik, J. (ed.) Directions in Corpus Linguistics. Proceedings of Nobel Symposium 82. Berlin: Mouton de Gruyter.

Svartvik, J. (1996) “Corpora are becoming Mainstream”. In J. Thomas and M. Short (eds.) Using Corpora for Language Research. London: Longman.

Svartvik, J. and O. Sager (1975) Modern engelsk grammatik. Stockholm: Esselte.

Swales, J.M. (1971) Writing Scientific English. Sunbury, UK: Thomas Nelson.

Swales, J. (1973) Introducing Teachers to English for Science and Technology. ELT Documents n. 6.

Swales, J. (1981) "Aspects of Article Introductions". Aston ESP Research Reports. Language Studies Unit. Birmingham: The University of Aston in Birmingham.

Swales, J.M. (1985, 1988) Episodes in ESP. New York: Prentice Hall.

Swales, J.M. (1990) Genre Analysis. English in Academic and Research Settings. Cambridge: Cambridge University Press.

Swales, J. (1995) "The Role of the Textbook in EAP Writing Research". TESOL Quarterly 14: 3-18.

Swan, M. (1990) "A Critical Look at the Communicative Approach". In R. Rossner and R. Bolitho (eds.) Currents of Change in English Language Teaching. Oxford: Oxford University Press.

Tang, G. (1992) "The Effect of Graphic Representation of Knowledge Structures on ESL Reading Comprehension". Studies in Second Language Acquisition 14: 177-196.

Tarantino, M. (1991) "EST: Dispassionate Discourse and the Discreet Presence". UNESCO-ALSED LSP Newsletter 13 (4): 28-34.

Thompson, P. (1998) "Current Issues in EAP Writing Instruction". In G. Langley (ed.) BALEAP PIM – Professional Interest Meetings. London: South Bank University.

Thurstun, J. (1996) "Teaching the Vocabulary of Academic English via Concordances". Annual Meeting of the Teachers of English to Speakers of Other Languages. Chicago: Chicago University Press.

Thurstun, J. and C.N. Candlin (1998) "Concordancing and the Teaching of the Vocabulary of Academic English". ESP 17: 20-34.

Tomalin, B. (1986) Video, TV and Radio in the English Class. London: MacMillan Pub.

Tribble, C. (1997) "Improvising Corpora for ELT: Quick and Dirty Ways of Developing Corpora for Language Teaching". In B. Lewandowska-Tomaszczyk and P.J. Melia (eds.) Practical Applications in Language Corpora. Lodz: Lodz University Press.

Tribble, C. (1998) "Genres, Keywords, Teaching: Towards a Pedagogic Account of the Language of Project Proposals". In E. Wichmann et al. (eds.) Teaching and Language Corpora. TALC' 98. Oxford: Keble College, Humanities Computing Unit.

Tribble, C.T. and G. Jones (1990) Concordances in the Classroom. Harlow: Longman.

Tribble, C.T. and G. Jones (1997) Concordancing in the Classroom. Houston: Athelstan.

Trimble, L. (1985) English for Science and Technology: A Discourse Approach. Cambridge: Cambridge University Press.

Trimble, M.T. and L. Trimble (1982) "Rhetorical-grammatical Features of Scientific and Technical Texts as a Major Factor in Written ESP Communication". In J. Høedt et al. (eds.) Pragmatics and LSP Proceedings of the 3rd European Symposium on LSP. Copenhagen: LSP Centre, School of Economics.

Ullman, S. (1962) Semantics. Oxford: Oxford University Press.

Underhill, A. (1985) "Working with the Monolingual Learners' Dictionary". In R. Ilson (ed.) Dictionaries, Lexicography and Language Learning. Oxford: Pergamon Press.

Ure, J.N. (1959, 1963) Types of Translation and Translatability. Proceedings of the IIIrd Congress of the International Federation of Translators, 1959. Oxford.

Ure, J. (1971) "Lexical Density and Register Differentiation". In G.E. Perren and J.L.M. Trim (eds.) Applications of Linguistics. Cambridge: Cambridge University Press.

Van Dijk, T.A. (1973) "Text Grammar and Text Logic". In J.S. Petofi and H. Rieser (eds.) Studies in Text Grammar. Boston: Reidel Publications.

Van Dijk, T.A. (1985) "Semantic Discourse Analysis". Handbook of Discourse Analysis. 2. New York: Academic Press.

Van Elk, J.A. (1975) The Threshold Level for Modern Language Learning in Schools. London: The Council of Europe.

Vande Kopple, W.J. (1991) "Themes, thematic progressions, and some Implications for Understanding Discourse". Written Communication 8: 311-347.

Varantola, K. (1984) On Noun Phrase Structures in Engineering English. Turku: Annales Universitatis Turkuensis.

Ventola, E. (1987) The Structure of Social Interaction: A Systemic Approach to the Semiotics of Service Encounters. London: Francis Pinter.

Viel, J.C. (1999) Science Teaching Reconsidered, A Handbook. http://www.nap.edu/readingroom/books/str/7.html

Vila, M.P. and J. Sevilla (1990) "Propuesta sobre la didáctica de la terminología aplicada a la traducción". In M. McGinity, R. Alejo and S. Gómez (eds.) Lenguas aplicadas a las ciencias y tecnología I. Badajoz: Universidad de Extremadura.

Wallace, M.J. (1982) Teaching Vocabulary. London: Heinemann.

Wallace, M.J. (1991) Study Skills in English. Cambridge: Cambridge University Press.

Wang, H. (1995) Interactivity in ESL Text with a Focus on ESL Hypertext. Unpub. PH D Thesis. Flagstaff, AZ: Northern Arizona University.

Warren, M. (1998) "Exploiting the Hong Kong Corpus of Spoken English for Research, Learning and Teaching Purposes". In E. Wichmann et al. (eds.) Teaching and Language Corpora. TALC' 98. Oxford: Keble College, Humanities Computing Unit.

Wenger, M.J. and J.H. Spyridakis (1993) "Reduced Text Structure at Two Text Levels: Impact on the Performance of Technical Readers". Journal of Technical Writing and Communication 23 (4): 333-352.

Werlich, E. (1976) A Text Grammar of English. Heildelberg: Quelle & Meyer.

White, P. (1999) "Language Teaching and the Web: Tips, Examples and Caveats". Proceedings of the Applied Linguistics Association of Australia. Sydney: University of Sydney.

Wichmann, A. et al. (eds.) (1997) Teaching and Language Corpora. London: Longman.

Widdowson. H.G. (1978) Teaching Language as Communication. Oxford: Oxford University Press.

Widdowson, H.G. (1979 a) "Directions in the Teaching of Discourse". In C.J. Brumfit and R. Johnson (eds.) The Communicative Approach to Language Teaching. Oxford: Oxford University Press.

Widdowson, H.G. (1979 b) Explorations in Applied Linguistics 1. Oxford: Oxford University Press.

Widdowson, H.G. (1983) Learning Purpose and Language Use. Oxford: Oxford University Press.

Widdowson, H.G. (1990) "Against Dogma: A Reply to Michael Swan". In R. Rossner and R. Bolitho (eds.) Currents of Change in English Language Teaching. Oxford: Oxford University Press.

Widdowson, H.G. (1993) "Communication, Community and the Problem of Appropriate Use". In J.E. Alatis (ed.) Language, Communication and Social Meaning. Washington D.C.: Georgetown University Press.

Wijasuriya, B.A. (1971) "The Occurrence of Discourse Markers and Intersentence Connectives in University Lectures and their Place in the Testing and Teaching of Listening Comprehension in English as a Foreign Language". M. Ed. Thesis. University of Manchester.

Wilkins, D.A. (1976) Notional Syllabuses. Oxford: Oxford University Press.

Wilkinson, A.M. (1992) "Jargon and the Passive Voice: Prescriptions and Proscriptions for Scientific Writing". Journal of Technical Writing and Communication 22 (3): 319-325.

Williams, E. (1984) Reading in the Language Classroom. ELTS. London: MacMillan.

Willis, D. (1990) The Lexical Syllabus. A New Approach to Language Teaching. London: HarperCollins.

Willis, D. (1997) "Lexical Phrases in Syllabus and Materials Design". In TFL Online Editor. Centre of English Language Studies. Birmingham: University of Birmingham. http://langue.hyper.chubu.ac.jp/jalt/pub/tlt/97/sep/willis.html

Willis, D. (1999) "Syllabus Design and the Pedagogic Corpus". In R. Budd (ed.) Vocabulary Learning in a Foreign Language. Paris: ENS Editions Fontenay / Saint-Cloud.

Winter, E. (1977) "A Clause-Relational Approach to English Texts: A Study of Some Predictive Lexical Items in Written Discourse". Instructional Science 6, 1.

Winter, E. (1978) "A Look at the Role of Certain Words in Information Structure". Informatics 3 (1): 85-97.

Winter, E. (1982) Towards a Contextual Grammar of English. London: George Allen and Unwin.

Witton, N. (1993) "Software Review: MicroConcord – Presented, Reviewed and Compared with the Mini-concordancer". ON-CALL 8 (2).

Wyatt, B. (1998) "Corpus-Based Language Description in ESP Course Design". In E. Wichmann et al. (eds.) Teaching and Language Corpora. TALC' 98. Oxford: Keble College, Humanities Computing Unit.

Yang, H.Z. (1986) "A New Technique for Identifying Scientific / Technical Terms and Describing Science Texts". Literary and Linguistic Computing 1 (2): 93-103.

Yap, K.O. (1979) "Vocabulary – Building Blocks of Comprehension?" Journal of Reading Behavior 1: 49-59.

Yekovich, F.R. and C.H. Walker (1978) "Identifying and Using Referents in Sentence Comprehension". Journal of Verbal Learning and Language Behavior 17.

Zernik, U. (1991) "Machine Readable Dictionaries". In U. Zernik (ed.) Lexical Acquisition. Exploiting On-line Resources to Build a Lexicon. Hillsdale, NJ: Lawrence Erlbaum Associates.

Zhu, Q. (1989) "A Quantitative Look at the Guangzhou Petroleum English Corpus". ICAME Journal 13: 28-38.

Zwicky, A.M. and A.D. Zwicky (1982) "Register as a Dimension of Linguistic Variation". In R. Kittredge and J. Lehrberger (eds.) Sublanguage. Studies of Language in Restricted Semantic Domains. Berlin: Mouton.

II. DICTIONARIES, CORPORA AND OTHER LEXICOGRAPHIC MATERIAL
(Sources contrasted in our research or at least mentioned as references)

Aguado de Cea, G. (1996) Diccionario comentado de terminología informática. Madrid: Paraninfo.

Breland, H.M., R.J. Jones and L. Jenkins (1994) "The College Board Vocabulary Study". College Board Report No 94-4. New York: College Board Publications.

British National Corpus (1997). Oxford University Press: http://info.ox.ac.uk/bnc/corpora/html

Brown Corpus (1995). Computer Language Resources. Brown University. Lancaster Unit for Computer Language Studies:
http://www.comp.lancs.ac.uk/computing/users/paul/ucrel/corpora.html
http://www.clres.com/dict.html#onto

Campion, M. and W. Elley (1971) An Academic Word List. Wellington: New Zealand Council for Educational Research.

Cambridge International Dictionary of English (1995) http://www.cup.org/titles/48/0521482364.html

Cambridge International Corpus (1997) English Language Teaching Resources. Cambridge University. http://www.cup.cam.ac.uk/elt

Ciber-léxico comparativo inglés-castellano (1999) Ramsay, S. and R. Lozano-Hemmer: http://www.telefonica.es/fat/lex.html

Clason, W.E. and S. Salem (1992) Elsevier's Dictionary of Library Science Information and Documentation. New York: Elsevier Science.

Collin, S. (1997) Dictionary of Information Technology. HarperCollins.

Collin, S. (1997) Dictionary of Multimedia. HarperCollins.

Collin, S. (1998) Dictionary of Personal Computing and the Internet. HarperCollins.

Collin, S. (1999) Dictionary of Computing (1999) London: HaperCollins.

Coxhead, A. (1998) An Academic Word List. English Language Institute Occasional Publication No 18. Victoria University of Wellington.

CRATER Multilingual Corpus (1994). Lancaster UCREL.

Crowther, J. (ed.) (1995) Oxford Advanced Learner's Dictionary. Oxford: Oxford University Press.

De Prada Poole, J. (1986) Glosario práctico de telecomunicaciones. Madrid: IDG Communications.

Dictionary of New Media: Film, Television, Print, Digital, Internet, Multimedia (1999). New York: Readfilm. http://www.readfilm.com/DNM.html

DuBois, J. et al. (2000) Santa Barbara Corpus of Spoken American English. University of California, Linguistic Data Consortium: http://humanitas.ucsb.edu/depts/linguistics/research/csae

Farrell, P. (1990) A Lexical Analysis of the English of Electronics and a Study of Semi-technical Vocabulary. Dublin: Trinity College.

Fernández Calvo, R. (1999) Glosario básico Inglés-español para usuarios de internet. Barcelona: Asociación de Técnicos de Informática.

Hoffmann, L. (1970 a) Fachwortschatz Medizin, Häufigkeitswörterbuch Russisch, Englisch, Französisch. Leipzig: Verlag Enzyklopädie.

Hoffmann, L. (1970 b) Fachwortschatz Physik, Häufigkeitswörterbuch Russisch, Englisch, Französisch. Leipzig: Verlag Enzyklopädie.

Hoffmann, L. (1973) Fachwortschatz Chemie, Häufigkeitswörterbuch Russisch, Englisch, Französisch. Leipzig: Verlag Enzyklopädie.

Hoffmann, L. (1976) Fachwortschatz Mathematik, Häufigkeitswörterbuch Russisch, Englisch, Französisch. Leipzig: Verlag Enzyklopädie.

Hoffmann, L. (1978) Fachwortschatz Tierproduktion und Veterinärmedizin, Häufigkeitswörterbuch Russisch, Englisch, Französisch. Leipzig: Verlag Enzyklopädie.

Horn, E. (1926) A Basic Writing Vocabulary, 10,000 Words Most Commonly Used in Writing. University of Iowa, College of Education.

Horn, E. and E.J. Ashbaugh (1928) The Horn-Ashbaugh Fundamentals of Spelling, Grades One to Eight, Incorporating the Findings of a Tabular Analysis of 5,100,000 Words of Ordinary Writing. Philadelphia: Lippincott.

Howe, D. (1998) On-line Dictionary of Computing (FOLDOC): http://www.aanet.aanekoski.fi/foldoc/foldoc.html

ICE-GB. International Corpus of English. Great Britain (1999) University College London. The Survey of English Usage: http://www.ucl.ac.uk/english-usage/

James, G. (1994) English in Computer Science. A Corpus-Based Lexical Analysis. Hong Kong: Longman.

Johansson, S., G.N. Leech and H. Goodluck (1978) Manual of Information to Accompany the Lancaster – Oslo / Bergen Corpus of British English, for Use with Digital Computers. Dpt. Of English, University of Oslo.

Johnson, S. (1755) A Dictionary of the English Language. London.

Kucera, H. and W.N. Francis (1967) Computational Analysis of Present-Day American English. Providence: Brown University Press.

Lozano Palacios, A. (1999) Vocabulario para los estudios de Biblio-documentación. Granada: Servicio de publicaciones, Universidad de Granada, Facultad de Biblioteconomía y Documentación.

MicroConcord Corpus Collection A & B (1993) Oxford: Oxford University Press.

Moon, R. (1998) The Hector Corpus. London: http://cogsi.berkeley.edu/pub/CorpusSurvey.ascii

Ogden, C.K. (1930) The Basic Words. A Detailed Study of their Uses. London: Orthological Institute.

Oxford Dictionary of Computing (1996). Oxford University Press.

Richards, J., J. Platt and H. Platt (1992) Longman Dictionary of Language Teaching and Applied Linguistics. London: Longman.

Praninskas, J. (1972) American University Word List. Harlow: Longman.

Quirk, R. (1992) The British National Corpus. Oxford University Press.

Rodríguez, H. and J. Vivaldi (1998). Computer Science Texts. Barcelona: IULA http://www.iula.upf.es/corpus/acinfca.htm.

Sinclair, J. et al. (1995) The Collins CoBuild English Dictionary. London. http://titania.cobuild.collins.co.uk/cob2flash.html

Sinclair, J. et al. (1987) Collins-CoBuild Language Dictionary. London: Collins.

Small Glossary of Terms: Vocabulary 563 (1997). Ball State University: http://www.tcom.bsu.edu

Stevenson, J. (1997) Dictionary of Library and Information Management. HarperCollins.

Summers, D. (1995) Longman Dictionary of Contemporary English. London.

SUSANNE project (1998). Survey of English Usage. http://www.ucl.ac.uk/english-usage/

Simpson, R. (2000) The Michigan Corpus of Academic Spoken English. English Language Institute. University of Michigan. http://www.lsa.umich.edu/eli/micase/micase.htm

Termite Database (1999). ITU Global Directory Telecommunication Terminology. http://www.itu.int/search

Thorndike, E.L. and I. Lorge (1944) The Teacher's Work Book of 30,000 Words. New York: Bureu of Publications, Teachers College, Columbia University.

Turton, N.D. and J.B. Heaton (1997) Longman Dictionary of Common Errors. Harlow: Longman Corpus Network.

Watters, C. (1992) Dictionary of Information Science and Technology. New York: Academic Press.

Webguest glossary (1999) http://www.webguest.com/glossary

Weiner, R. (1999) Communications Jargon. http://www.all-links.com/dictionary/index.html

West, M. (1953) A General Service List of English Words. London: Longman.
Wordnet (1997). Princeton University: http://www.cogsci.princeton.edu/~wn

Xue, G. and I.S.P. Nation (1984) "A University Word List". Language Learning and Communication 3: 215-229.

III. HYPERTEXTUAL REFERENCES

(References on the World Wide Web and e-mail lists; http's, ftp's, and e-mail addresses)

Anthony, L. (27 april 2000) EST-Electronic List: EST-L@lists.asu.edu

Berber Sardinha, T. (12 Oct 1998) Corpora Electronic Discussion List: http://www.hd.uib.no/corpora

CALL Conference Proceedings (1999) http://www.intellect-net.com/elm/index.htm

CALL On-line + CALL-EJ (2000) Australia and Japan: http://www.lerc.ritsumei.ac.jp/callej/index.html

CALL On-Line Journal (1991-1999) Australia: http://www.cltr.uq.edu.au/oncall and new

Cambridge University Press Electronic References: http://www.cup.cam.ac.uk/elt/reference/clc.htm

Chandler, R.M. (26 april 2000) EST Electronic List: EST-L@lists.asu.edu

Clear, J. (12 Dec 1999) Corpora Electronic Discussion List: http://www.hd.uib.no/corpora

Collection of Conferences on Information Technology (1998) http://www.emst.pnl.gov:2080

Collins CoBuild Corpus and Concordance (1999): http://titania.cobuild.collins.co.uk/javahelp/help.html.

Computing Services for Language Learning Tools (1999): http://www.compulang.com

Dictionary Collection of Specialized Terminology (1997): http://reimari.uwasa.fi/comm/termino/collect/special/computing.html.

Electronic Texts (1997) Washington D.C. http://www.wkap.nl/kaphtml.htm/ON-LINE

Electronic Texts (1997) Washington D.C. http://xxx.lanl.gov

Esteves da Rocha, M.A. (3 Nov 1999) Corpora Electronic Discussion List: http://www.hd.uib.no/corpora

ICAME, International Computer Archive of Modern English association (1997): http://nora.hd.uib.no/pub.icame

Information Science and Technology program (1998) Prf. Padmini Srinivasan: University of Iowa; University of Pittsburgh:

Information Science and Telecommunications studies program (1998) School of Information Science at the University of Pittsburgh http://www.tele.pitt.edu

Instituto universitario del audio-visual -- recursos (IULA) (1998) Universitat Pompeu Fabra: http://www.iula.upf.es/corpus/corpusuk.htm

Interactive Connections Internet Glossary (1999): http://www.instantweb.com

Johns, T. (1998) EAP web page of `kibbitzers´. Birmingham: http://sun1.bham.ac.uk/johnstf/timeap3.htm

Krishnamurthy, R. (23 Febr 2000) Corpora Electronic Discussion List: http://www.hd.uib.no/corpora

Language Learning and Technology Journal (1997-2000): http://polyglot.cal.msu.edu/llt/

LDC Catalogue of Corpora and Text Archives (1996): http://www.ldc.upenn.edu/Catalog).

Lukes, T. (4 Jan 1994) "Programmes for analysis of texts". In http://anes235-1.ff.cuni.cz/CAAL/0010.html

Mason, O. (9 Oct 1998; 28 Jan 2000) Corpora Electronic Discussion List: http://www.hd.uib.no/corpora

Milton, J. (2 Apr. 1999) Corpora Electronic Discussion List: http://www.hd.uib.no/corpora

Oates, H. (16 Jul 1999) Corpora Electronic Discussion List: http://www.hd.uib.no/corpora

On-line Dictionaries and Glossaries (1997): http://www.clres.com/dict.html#onto

OTA (Oxford Text Archive) ftp://ota.ox.ac.uk or http://ota.ahds.ac.uk

PC Webopedia search engine (1998) http://pcwebopedia.com

Sara Concordance Tools (1999) Lancaster University. http://comp.lancs.ac.uk/computing/research/ucrel/tools.html#sara

Scott, M. (9 Oct. 1998) Corpora Electronic Discussion List: http://www.hd.uib.no/corpora

Specialized On-line Dictionaries (1998) http://www.facstaff.bucknell/edu

Spoken Academic English tagged data (1998) R. Thompson. Reading University. http://www.rdg.ac.uk/AcaDepts/cl/CALS/corpus.html.

TACT Concordance tools (1990) University of Toronto. http://www.students.uiuc.edu

The Major Varieties of English (MAVEN 97) (1998) H. Lindquist et al. (eds.) Vaxjo University, Sweden: http//linguistlist.org/issues/10/10-56.html#1

Universidad Politécnica de Valencia programs of study (1999) Guía del alumno. http://www.fiv.upv.es/cas/lnglnf.html

Webcorp search engine (2000). University of Liverpool, Research and Development Unit for English Studies http://webcorp.connect.org.uk

Williams, G. (1 Dec 1998; 23 Oct 1998) Corpora Electronic Discussion List: http://www.hd.uib.no/corpora

Word Pilot (2000) http://www.compulang.com

WordCruncher (1993) Johnston & Company: http://www.students.uiuc.edu

Wordnet (199) Princeton University: http://www.cogsci.princeton.edu/~wn

IV. CORPÙS REFERENCES

1. References according to Genre and Discipline Sub-corpora:

I. Research articles:

I.A. Computer Science articles

1.-- Buhrman, H., et al (1999) “Mutual Search”. ACM Data structures and Algorithms 19: 18-42.
Subject: A2
Tokens: 2,736 / Types: 867

2.-- Davis, E. (1999) “Order of Magnitude: Comparisons of Distance”. Journal of Artificial Intelligence Research 10: 1-38. New York University. Email: davise@cs.nyu.edu
Subject: A2
Tokens: 12,833 / Types: 1,332

3.-- Díaz Martín, J.C., I. Irala Veloso and J.M. Rodríguez García (1999) “Building TLC-TK GUIs for HRT-HOOD Systems”. Distributed Systems. Topics. Dpto. Informática, Universidad de Extremadura.
Subject: A2
Tokens: 10,099 / Types: 1,099

4.-- De Vries, E. (1996) “Using information systems while performing complex tasks : an example from architectural design”. Vol. 40, No. 4 - IBM ASIC design and testing. Faculty of Technology Management , Eindho University of Technology , The Netherlands,
Subject: A2
Tokens: 32,003 / Types: 1,348

5.-- Hogg, T. (1999) “Solving Highly Constrained Search Problems with Quantum Computers”. Journal of Artificial Intelligence Research 10: 39-66. AI Access Foundation and Morgan Kaufmann Publishers.

hogg@parc.xerox.com

Tokens: 8,569 / Types: 1,022

6.-- Jaakkola, T.S. (1999) “Variational Probabilistic Inference

and the QMR-DT Network”. Journal of Artificial Intelligence Research 10 (1999), pp. 291-322. Submitted 10/98; published 5/99.

Tokens: 2,917 Types: 848

7.-- Lapier, C. (1999) “Network Management with SNMP”. IBM features implementation. 19 (3): 24-38.

Subject: B1

Tokens: 3,304 / Types: 864

8.-- Rotithor, H.G., K.W. Harris and M.W. Davis (1999) “Measurement and Analysis of C and C++ Performance”. Digital Technical Journal 10, 1: 44-66.

Subject: A1

Tokens: 10,018 / Types: 1,684

9.-- Waterson, P.E., C.W. Clegg and C.M. Axtell (1996) “The dynamics of work organization , knowledge and technology during software development”. Software Development 4 (2): 13-32. Institute of Work Psychology , University of Sheffield

Subject: B3

Tokens: 38,222 / Types: 1,540

10.-- Wilkinson, P. (1998) “OPAC for Intranets”. Information Dimensions 6: 23-31.

http://www.InformationDimensions.com

Subject: F1

Tokens: 3,870 / Types: 1.040

I.B. Information Science articles

1.-- Barry, J. (1994) "The HyperText Markup Language (HTML) and the World-Wide Web: Raising ASCII Text to a New Level of Usability". The Public-Access Computer Systems Review 5 (5): 5-62. To retrieve this file, send the following e-mail message to listserv@uhupvm1.uh.edu: GET BARRY PRV5N5 F=MAIL
Subject: F3
Tokens: 13,035 / Types: 1,796

2.-- Chau, M.Y. (1998) "Computer Supported Concept Maps: Excellent Tools for Enhancing Library Workshop Presentations". LIBRES: Library and Information Science Research Electronic Journal 8 (2): 34-56.
may.chau@orst.edu
Subject: B2
Tokens: 6,782 / Types: 1,695

3.-- Collier, B. (1998) "Marketing Today's Libraries". Library and Internet 4: 23-25.
Tokens: 3,701 / Types: 1,108

4.-- Hepworth, M. (1998) "Investigating Methods for Understanding User Requirements for Information Products". Information Research 4 (2): 36-48.
http://www.shef.ac.uk/~is/publications/infres/isic/hepworth.html
Subject: F1
Tokens: 3,608 / Types: 972

5.-- Kipling, P. and T.D. Wilson (1999) "Publishing, Bookselling and the World Wide Web". Information Research 4 (4): 34-57. t.d.wilson@shef.ac.uk>
Department of Information Studies
University of Sheffield, UK. Location: http://www.shef.ac.uk/~is/publications/infres/paper63.html © the authors, 1999.
Subject: F5
Tokens: 5,198 / Types: 1,107

6.-- Price-Wilkin, J. (1994) "A Gateway Between the World-Wide Web andb PAT: Exploiting SGML Through the Web." The Public-Access Computer Systems Review 5, no. 7: 5-27. To retrieve this file, use the following URL: gopher://info.lib.uh.edu:70/
Subject: F3
Tokens: 6,885 / Types1,448

7.-- Price-Wilkin, J. (1994) "Using the World-Wide Web to Deliver Complex Electronic Documents: Implications for Libraries". The Public-Access Computer Systems Review 5 (3): 5-21. To retrieve this file, send the following e-mail message to listserv@uhupvm1.uh.edu: GET PRICEWIL PRV5N3 F=MAIL
Subject: F4
Tokens: 5,736 / Types: 1,389

8.-- Tennant, R. (1995) "The Virtual Library Foundation: Staff Training and Support". Digital Library Research & Development 3: 34-45.
http://www.lib.berkeley.edu/
Subject: F6
Tokens: 2,605 / Types: 851

9.-- Thomas, C.F. and L.S. Griffin (1996) "Metadata in Information Science". Information Science 4: 24-43. Chuck@seal.lib.lsu.edu notlsg@unix1.sncc.lsu.edu
Subject: B2
Tokens: 3,282 / Types: 1,006

10.-- Wilson, T. (1997) "Electronic Publishing and the Future of the Book". Book Science 4: 43-64. Department of Information Studies. University of Sheffield.
t.d.wilson@shef.ac.uk
Subject: F5
Tokens: 3,897 / Types: 1,122

I.C. Telecommunications articles

1.-- Ahmed, N., G. Mandyam and N. Magotra (1997) "Lossless Image Compression Using the Discrete Cosine Transform". Virtual Code 8 (1): 21-26.
Subject: A2
Tokens: 2,278 / Types: 289

2.-- Armstrong, C.J. (1997) "Metadata, Pics and Quality". Ariadne 18 (May): 15-23. Centre for Information Quality Management. lisqual@cix.compulink.co.uk
Subject: A2
Tokens: 3,293 / Types: 1,015

3.-- Charnetski, J.R. (1998) "Avoid Disaster: Use Firewalls for Inter-Intranet Security". Internet Librarian 3: 23-29.
Subject: F2
Tokens: 2,775 / Types: 938

4.-- Correll, K.W. and R.A. Ulichney (1997) "The J300 Family of Video and Audio Adapters: Architecture and Hardware Design". Telecommucation Type 5: 34-49.
Subject: D1
Tokens: 7,728 / Types: 1,435

5.-- Dyer, T. (1999) "Web Proxy Servers". Telecommunications Electronic Review 6 (3): 12-29. Prentice Hall, NJ.
Subject: F6
Tokens: 4,510 / Types: 1,320

6.-- Ellen, D. (1998) "The Role of Telecentres in the Provision of Community Access to Electronic Information". Information Research 4 (2): 21-34.
http://www.shef.ac.uk/~is/publications/infres/isic/ellen.html
Subject: F6
Tokens: 3,597 / Types: 896

7.-- Gupta, G.C., P.J. Legg, I. Andonovic and D. Uttamchandi (1997) "Noise Analysis in Coherence-Multiplexed Optical Fiber Communication Systems". Optical Fiber Technology 3: 21-27.
Tokens: 6,819 / Types: 1,606

8.-- Miller, S. (1994) "Liberals and Libertarians in Cyberspace". The Network Observer 1 (8): 3-10.
Subject: F1
Tokens: 7,162 / Types: 2,039

9.-- Rekers, J. and A. Schürr (1997) "Defining and Parsing Visual Languages with Layered Graph Grammars". The Network Observer 3 (2): 43-47.
Tokens: 1,517 / Types: 523

10.-- Ubois, J. (1998) "Casting an Information Net". Visual Insights 4: 12-23. http://www.lucent.com/visualinsights
Subject: F6
Tokens: 4,339 / Types: 1,374

I.D. Audio-visual Communication articles

1.-- Alsup, M. (1998) "The Voice of the Document Management Industry. Riding the Colors of Light". Inform 6: 3-8. aiim@aiim.org
Subject: E1
Tokens: 1,957 / Types: 669

2.-- Borgman, C.L. (1997) "Multi-Media, Multi-Cultural, and Multi-Lingual Digital Libraries". D-Lib Magazine June: 23-43. Department of Library and Information Science.
Subject: C2
Tokens: 4,921 / Types: 1,177

3.-- Duffy, C. (1998) "Networking Moving Images". Computers and Texts 16 / 17: 23-32. CTI Textual Studies.
cath@pads.ahds.ac.uk
Subject. C1
Tokens: 1,736 / Types: 604

4.-- Farmer, L.S.J. (1997) "Networking CD-ROMs to Distribute Information". Librarian and technology 4: 12-21.
Subject: F6
Tokens: 2,444 / Types: 795

5.-- Fraser, M. and S. Porter (1997) "Computer-Assisted Film and Drama Studies". Computers and Texts 14: 2-12. CTI Textual Studies.
ctitext@oucs.ox.ac.uk
Tokens: 3,645 / Types: 1,160

6.-- Godwin-Jones, B. (1997) "Emerging Technologies: Real-Time Audio and Video Playback on the Web". Language Learning and Technology 1 (1): 5-8.
Subject: F4
Tokens: 2,393 / Types: 795

7.-- Green, D. (1995) "Beyond Word and Image". Computers and Texts 12: 38-54.
david@cni.org
Subject: D1
Tokens: 4,575 / Types: 1,295

8.-- Mehta, M.D. and D.E. Plaza (1997) "Content Analysis of Pornographic Images Available on the Internet". The Information Society 13 (2): 153-162.
Subject: C1
Tokens: 4,428 / Types: 916.

9.-- Reed, R. (1995) "Designing Genres for New Media: Social, Economic and Political Contexts". The Network Observer 2 (11): 6-24.
Subject: C2
Tokens: 8,675 / Types: 2,050

10.-- Society of IM Europe (1998) "A Study for the European Commission's Information Technology Programme". Public Libraries and the Information Society 2: 13-31.
Tokens: 5,509 / Types: 1,138

II. Textbooks:

II.A. Audio-visual Communication textbook

1.-- Chandler, D. (1998) Semiotics for Beginners. Wadsworth Pub.
Subject: C1
Tokens: 42,769 / Types: 5,138

II.B. All-disciplines textbooks

1.-- Commission on Preservation and Access (1996) Preserving Digital Information. The Research Libraries Group.
Subject: E1
Tokens: 24,767 / Types: 3,266

2.-- Daly, J. (1995) Electronic Texts. Washington DC: The Library of Congress.
Subject: F5
Tokens: 46,619 / Types: 4,853

3.-- Hauben, M. and R. Hauben (1995) Netizens: On the History and Impact of Usenet and the Internet. New York City College.
hauben@columbia.edu
Subject: F2
Tokens: 45,031 / Types: 5,455

4.-- Sperberg-McQueen, C.M. and L. Burnard (1995) Guidelines for Electronic Text Encoding and Interchange. Chicago: ACH.
Subject: F3
Tokens: 49,168 / Types: 4,891

5.-- Straubhaar, J. (1997) Communications Media in the Information Society. Wadsworth Pub.
Subject: E1
Tokens: 32,987 / Types: 5,768

II.C. Computer Science Textbook

1.-- Computer Science Encyclopedia (1999) The Computer language Co.
Subject: A1
Tokens: 18,833 / Types: 2,882

II.D. Information Science Textbook

1.-- Schnell, E.H. (1996) Writing for the Web: A Primer for Librarians. Library Book Inc.
Subject: F5
Tokens: 7,997 / Types: 1,525

II.E. Telecommunications Textbook

1.-- IEEE (1995) Distributed Systems Communication. Helvetica-Narrow.
Subject: A2
Tokens: 51,074 / Types: 2,683

III. Technical Reports:

III.A. Audio-visual Communication Reports

1.-- Brennam, J.A. (1997) "Telemedicine". FACEP, New York.
Subject: D1
Tokens: 773 / Types: 365

2.-- Brinson, J.D. and M.F. Radcliffe (1996) "An Intellectual Property Law Primer for Multimedia and Web Developers". Pomona College, CA.
Subject: F6
Tokens: 1,844 / Types: 619

3.-- Chaiken, D., M. Hayter, J. Kistler and D. Redell (1998) "The Virtual Book". SRC Research Report 157.
Subject: F5
Tokens: 1,225 / Types: 314

4.-- De Luca Pretto, N. (1999) "Media Technologies and Education". Goldmiths College, London.
Subject: D1
Tokens: 2,653 / Types: 863

5.-- Fleischhauer, C. (1998) "Digital Formats for Content Reproductions". National Digital Library program, Library of Congress.
Subject: F6
Tokens: 5,153 / Types: 1,183

6.-- "Liability and Audio-visual material" (1997) Copyright Law of the United States, Washington D.C.
Subject: C2
Tokens: 15,329 / 848

7.-- Library of Congress, The (1996) "National Digital Library Program". Washington D.C.
Subject: C3
Tokens: 7,487 / Types: 1,337

8.-- Library of Congress, The (1996) "Microfilm Collections". Washington D.C.
Subject: C3
Tokens: 8.876 / Types: 1,204

9.-- "On-line Visual Literacy project" (1998) Pomona College, Claremont, CA.
Tokens: 7,927 / Types: 1,705

10.-- "Unathorized Fixation and Trafficking in Sound recordings" (1996) Copyright Law of the United States, Washington D.C.
Subject: C2
Tokens: 16,627 / Types: 924

III.B. Computer Science Reports

1.-- Bench-Capon, T., et al. (1997) "Report on the 1st International Workshop on Validation, Verification and Integrity Issues of Expert and Database Systems". University of Liverpool.
Subject: B2
Tokens: 4,739 / Types: 1,039

2.-- "Dial-Up Scripting Command Language for Dial-Up Networking Scripting Support" (1995) Microsoft Corp.
Subject: F4
Tokens: 3,703 / Types: 479

3.-- Grant, C. (1995) "Clients / Browsers; What does the Future Hold?" ACCESS'95. University of New Brunswick.
Subject: A2
Tokens: 1,800 / Types: 620

4.-- Green, P. (1999) "Multics Virtual Memory". Multics Virtual Memory Design, Department of Computer Science, NMSU, Las Cruces, NM.
Subject: A1
Tokens: 10,754 / Types: 1,789

5.-- Hutton, G. (1999) "Comp.lang Functional". University of Nottingham.
Tokens: 7,764 / Types: 1,645

6.-- Kahn, R. (1995) "A Framework for Distributed Digital Object Services". Corporation for National Research Initiatives, University of California at Berkeley.
Subject: A2
Tokens: 18,175 / Types: 1,596

7.-- "Introduction to M.4 and M.4VB". Dept. of Information Technology, University of Pittsburgh.
Subject: B3
Tokens: 4,562 / Types: 1,041

8.-- Moulton, L.W. (1997) "Starting a New Database". Comstow Information Services.
Subject: B2
Tokens: 2,316 / 813

9.-- Older, S. (1998) "A Simple Introduction to Pattern Matching in Caml". Computer and Information Systems, University of Syracusse.
Tokens: 1,927 / Types: 410

10.-- Tanenbaum, A. (1996) "Minix Information sheet". ast@cs.vu.nl
Subject: A1
Tokens: 2,937 / Types: 1,083

III.C. Information Science Reports

1.-- Bates, M.J. (1999) "Final report to the ALCTS/CCS Subject Analysis Committee: Subject Access in Online Catalogs, A Design Model". American Library Association. http://www.ala.org/alcts/organization/cs/sac/rpt97rev.html
Tokens: 2,020 / Types: 674

2.-- Bauwens, M. (1995) "Cybrarians article". CDT reports, University of Cincinnati, Library Archives.
Tokens: 1,948 / Types: 724

3.-- Böok, M., K.M. Coop and S. Sucksdorff (1997) “Networking the Library”. Helsinki City Library.
Subject: F6
Tokens: 874 / Types: 396

4.-- Crocca, W.T. and W.L. Anderson (1995) “Delivering Technology for Digital Libraries: Experiences as Vendors”. Xerox Corporation, Fairport, NY.
Subject: E1
Tokens: 5,915 / Types: 1,428

5.-- Crocker, D. (1995) “To Be on the Internet”. Network Working Group. Brandenburg Co.
Subject: F2
Tokens: 1,174 / Types: 415

6.-- Copyright Office, The (1996) “A brief History of Copyright and Overview”. Library of Congress, Washington D.C.
Subject: C3
Tokens: 3,993 / Types: 989

7.-- McClure, C.R. et al (1995) “The Project Gain report: Connecting Rural Public Libraries to the Internet”. NYSERnet Inc., Liverpool.
Subject: F6
Tokens: 23,851 / Types: 3,003

8.-- Mouton, L. (1997) “Guide to Planning for Automation in a Library”. Comstow Information Services Inc.
Subject: B1
Tokens: 1,037 / Types: 406

9.-- Murray, P.C. (1998) "Core Concepts of Knowledge Management". KM Metazine, London.
Subject: B3
Tokens: 4,158 / Types: 1,246

10.-- Najork, M.A. and M.H. Brown (1994) "SRC Research Report 128 a. A Library for Visualizing Combinatorial Structures". Library of Congress.
Subject: B2
Tokens: 2,200 / Types: 738

III.D. All-disciplines Reports

1.-- Atkinson, R., G. Spafford and M. Moraes (1994) "Network Info". Netannounce, University of Virginia.
Subject: F2
Tokens: 1,084 / Types: 375

2.-- Copyright Office, The (1996) "Internet and Copyright". Project Looking Forward.
Tokens: 34,504 / Types: 2,773

3.-- "Do's and Don't's of Technology Planning". Florida Educational Technology Conference.
Subject: F6
Tokens: 1,296 / Types: 436

4.-- Hallgreen, M. (1996) "Connecting to the Internet". ACM SIGUCCS, Network Working Group.
Subject: F2
Tokens: 7,062 / Types: 1,551

5.-- Tilton, J.E. (1998) “Composing Good HTML”.
http://www.cs.cmu.edu/~tilt/cgh/
Subject: F3
Tokens: 14,437 / Types: 2,344

6.-- Von Tetzchner, J.S. (1997) “Frame 2 HTML Filter”. Norwegian Telecom Research.
Subject: A1
Tokens: 1,328 / Types: 487

III.E. Telecommunications Reports

1.-- Birrell, A., G. Nelson, S. Owicki and E. Wobber (1995) “SRC Research Report 115: Network Objects”.
Subject: A2
Tokens: 1,134 / Types: 534

2.-- Clark, J.M. (1998) “Cellular Options for Mobile Data Connectivity”. CARL Corporation. http://www.carl.org:1080/
Tokens: 3,270 / Types: 863

3.-- Engler, D.R., et al (1997) “Application Performance and Flexibility on Exokernel Systems”. M.I.T. Laboratory, Cambridge, MA.
http://www.pdos.lcs.mit.edu/
Tokens: 15,870 / Types: 2,479

4.-- Keith, L. (1997) “FWTS Fact sheet”. Federal Wireless Telecommunications services.
Subject: D2
Tokens: 16,739 / Types: 2,639

5.-- International Telecommunication Union (1999) “Radio Communication study”. Document US 8B/25, Kansas State University.
Subject: D2
Tokens: 2,396 / Types: 696

6.-- ITU-R Fact sheet (1999) “Procedures for Analyzing Compatibility with Radiodetermination and Metereological Radars”. Working Party: 8B, Document US8B-22, Kansas State University.
Subject: A2
Tokens: 6,332 / Types: 698

7.-- ITU-R XXX (1998) “Characteristics of and Protection Criteria for Systems Operating in the Radiolocation Service in the Frequency Band 420-450 MHz”. Document US8B-18, Kansas State University.
Subject: A2
Tokens: 6,731 / Types: 732

8.-- Loughry, D.C. (1995) “IEEE-SA and Institute Interactions”. IEEE Board of Directors, Connecticut.
Subject: D2
Tokens: 10,585 / Types: 1,152

9.-- National Academy of Sciences (1997) “Modern Communication: The Laser and Fiber-Optic Revolution”. Beyond Discovery research.
Subject: D1
Tokens: 16,962 / Types: 1,729

10.-- Orost, J. (1995) “The Bench++ Benchmark suite”. AT&T Labs.
Subject: A1
Tokens: 5,453 / Types: 1,008

2. References according to Genre and Text type sub-corpora:

I. Research article excerpts:

I.A. Classification samples (Tokens: 2,335 / Types: 613)

1.-- Ellen, D. (1998) "The Role of Telecentres in the Provision of Community Access to Electronic Information". Information Research 4 (2): 21-34.
http://www.shef.ac.uk/~is/publications/infres/isic/ellen.html
Subject: F6

2.-- Kipling, P. and T.D. Wilson (1999) "Publishing, Bookselling and the World Wide Web". Information Research 4 (4): 34-57. t.d.wilson@shef.ac.uk>
Department of Information Studies
University of Sheffield, UK. Location: http://www.shef.ac.uk/~is/publications/infres/paper63.html © the authors, 1999.
Subject: F5

I.B. Conclusion samples (Tokens: 2,931 / Types: 962)

1.-- Alsup, M. (1998) "The Voice of the Document Management Industry. Riding the Colors of Light". Inform 6: 3-8. aiim@aiim.org
Subject: E1

2.-- Correll, K.W. and R.A. Ulichney (1997) "The J300 Family of Video and Audio Adapters: Architecture and Hardware Design". Telecommucation Type 5: 34-49.
Subject: D1

3.-- Davis, E. (1999) "Order of Magnitude: Comparisons of Distance". Journal of Artificial Intelligence Research 10: 1-38. New York University. Email: davise@cs.nyu.edu
Subject: A2

4.-- Kipling, P. and T.D. Wilson (1999) "Publishing, Bookselling and the World Wide Web". Information Research 4 (4): 34-57. t.d.wilson@shef.ac.uk>
Department of Information Studies
University of Sheffield, UK. Location: http://www.shef.ac.uk/~is/publications/infres/paper63.html © the authors, 1999.
Subject: F5

5.-- Thomas, C.F. and L.S. Griffin (1996) "Metadata in Information Science". Information Science 4: 24-43. Chuck@seal.lib.lsu.edu notlsg@unix1.sncc.lsu.edu
Subject: B2

6.-- Wilson, T. (1997) "Electronic Publishing and the Future of the Book". Book Science 4: 43-64. Department of Information Studies. University of Sheffield.
t.d.wilson@shef.ac.uk
Subject: F5

I.C. Definition samples (Tokens: 3,863 / Types: 1,066)

1.-- Hepworth, M. (1998) "Investigating Methods for Understanding User Requirements for Information Products". Information Research 4 (2): 36-48.
http://www.shef.ac.uk/~is/publications/infres/isic/hepworth.html
Subject: F1

2.-- Reed, R. (1995) "Designing Genres for New Media: Social, Economic and Political Contexts". The Network Observer 2 (11): 6-24.
Subject: C2

3.-- Waterson, P.E., C.W. Clegg and C.M. Axtell (1996) "The dynamics of work organization , knowledge and technology during software development". Software Development 4 (2): 13-32. Institute of Work Psychology , University of Sheffield
Subject: B3

I.D. Description samples (Tokens: 3,480 / Types: 1,041)

1.-- Chau, M.Y. (1998) "Computer Supported Concept Maps: Excellent Tools for Enhancing Library Workshop Presentations". LIBRES: Library and Information Science Research Electronic Journal 8 (2): 34-56.
may.chau@orst.edu
Subject: B2

2.-- Jaakkola, T.S. (1999) "Variational Probabilistic Inference
and the QMR-DT Network". Journal of Artificial Intelligence Research 10 (1999), pp. 291-322. Submitted 10/98; published 5/99.

Tokens: 2,917 Types: 848

3.-- Mehta, M.D. and D.E. Plaza (1997) "Content Analysis of Pornographic Images Available on the Internet". The Information Society 13 (2): 153-162.
Subject: C1

4.-- Tennant, R. (1995) "The Virtual Library Foundation: Staff Training and Support". Digital Library Research & Development 3: 34-45.
http://www.lib.berkeley.edu/
Subject: F6

I.E. Discussion samples (Tokens: 10,527 / Types: 2,129)

1.-- Alsup, M. (1998) "The Voice of the Document Management Industry. Riding the Colors of Light". Inform 6: 3-8. aiim@aiim.org
Subject: E1

2.-- Barry, J. (1994) "The HyperText Markup Language (HTML) and the World-Wide Web: Raising ASCII Text to a New Level of Usability". The Public-Access Computer Systems Review 5 (5): 5-62. To retrieve this file, send the following e-mail message to listserv@uhupvm1.uh.edu: GET BARRY PRV5N5 F=MAIL
Subject: F3

3.-- Correll, K.W. and R.A. Ulichney (1997) "The J300 Family of Video and Audio Adapters: Architecture and Hardware Design". Telecommucation Type 5: 34-49.
Subject: D1

4.-- Davis, E. (1999) "Order of Magnitude: Comparisons of Distance". Journal of Artificial Intelligence Research 10: 1-38. New York University. Email: davise@cs.nyu.edu
Subject: A2

5.-- Thomas, C.F. and L.S. Griffin (1996) "Metadata in Information Science". Information Science 4: 24-43. Chuck@seal.lib.lsu.edu notlsg@unix1.sncc.lsu.edu
Subject: B2

6.-- Wilson, T. (1997) "Electronic Publishing and the Future of the Book". Book Science 4: 43-64. Department of Information Studies. University of Sheffield.
t.d.wilson@shef.ac.uk
Subject: F5

I.F. Exemplification samples (Tokens: 9,118 / Types: 1,510)

1.-- Davis, E. (1999) “Order of Magnitude: Comparisons of Distance”. Journal of Artificial Intelligence Research 10: 1-38. New York University. Email: davise@cs.nyu.edu
Subject: A2

2.-- Díaz Martín, J.C., I. Irala Veloso and J.M. Rodríguez García (1999) “Building TLC-TK GUIs for HRT-HOOD Systems”. Distributed Systems. Topics. Dpto. Informática, Universidad de Extremadura.
Subject: A2

3.-- Lapier, C. (1999) “Network Management with SNMP”. IBM features implementation. 19 (3): 24-38.
Subject: B1

4.-- Price-Wilkin, J. (1994) "A Gateway Between the World-Wide Web andb PAT: Exploiting SGML Through the Web." The Public-Access Computer Systems Review 5, no. 7: 5-27. To retrieve this file, use the following URL: gopher://info.lib.uh.edu:70/
Subject: F3

II. Textbook excerpts:

II.A. Classification samples (Tokens: 11,313 / Types: 1,570)

1.-- Computer Science Encyclopedia (1999) The Computer language Co.
Subject: A1

2.-- Hauben, M. and R. Hauben (1995) Netizens: On the History and Impact of Usenet and the Internet. New York City College.
hauben@columbia.edu
Subject: F2

3.-- Straubhaar, J. (1997) Communications Media in the Information Society. Wadsworth Pub.
Subject: E1

II.B. Conclusion samples (Tokens: 4,123 / Types: 491)

1.-- Chandler, D. (1998) Semiotics for Beginners. Wadsworth Pub.
Subject: C1

II.C. Definition samples (Tokens: 2,207 / Types: 595)

1.-- Commission on Preservation and Access (1996) Preserving Digital Information. The Research Libraries Group.
Subject: E1

2.-- Computer Science Encyclopedia (1999) The Computer language Co.
Subject: A1

3.-- Sperberg-McQueen, C.M. and L. Burnard (1995) Guidelines for Electronic Text Encoding and Interchange. Chicago: ACH.
Subject: F3

II.D. Description samples (Tokens: 534 / Types: 159)

1.-- Schnell, E.H. (1996) Writing for the Web: A Primer for Librarians. Library Book Inc.
Subject: F5

II.E. Discussion samples (Tokens: 2,741 / Types: 728)

1.-- IEEE (1995) Distributed Systems Communication. Helvetica-Narrow.
Subject: A2

2.-- Straubhaar, J. (1997) Communications Media in the Information Society. Wadsworth Pub.
Subject: E1

II.F. Exemplification samples (Tokens: 1,109 / Types: 499)

1.-- Hauben, M. and R. Hauben (1995) Netizens: On the History and Impact of Usenet and the Internet. New York City College.
hauben@columbia.edu
Subject: F2

III. Technical Report excerpts:

III.A. Classification samples (Tokens: 6,573 / Types: 1,153)

1.-- Bench-Capon, T., et al. (1997) "Report on the 1st International Workshop on Validation, Verification and Integrity Issues of Expert and Database Systems". University of Liverpool.
Subject: B2

2.-- De Luca Pretto, N. (1999) "Media Technologies and Education". Goldmiths College, London.
Subject: D1

3.-- Fleischhauer, C. (1998) "Digital Formats for Content Reproductions". National Digital Library program, Library of Congress.
Subject: F6

4.-- Grant, C. (1995) "Clients / Browsers; What does the Future Hold?" ACCESS'95. University of New Brunswick.
Subject: A2

5.-- Mouton, L. (1997) "Guide to Planning for Automation in a Library". Comstow Information Services Inc.
Subject: B1

III.B. Conclusion samples (Tokens: 3,679 / Types: 1,004)

1.-- Green, P. (1999) "Multics Virtual Memory". Multics Virtual Memory Design, Department of Computer Science, NMSU, Las Cruces, NM.
Subject: A1

2.-- McClure, C.R. et al (1995) "The Project Gain report: Connecting Rural Public Libraries to the Internet". NYSERnet Inc., Liverpool.
Subject: F6

3.-- Mouton, L. (1997) "Guide to Planning for Automation in a Library". Comstow Information Services Inc.
Subject: B1

III.C. Definition samples (Tokens: 8,637 / Types: 1,457)

1.-- Birrell, A., G. Nelson, S. Owicki and E. Wobber (1995) "SRC Research Report 115: Network Objects".
Subject: A2

2.-- Kahn, R. (1995) "A Framework for Distributed Digital Object Services". Corporation for National Research Initiatives, University of California at Berkeley.
Subject: A2

3.-- Keith, L. (1997) "FWTS Fact sheet". Federal Wireless Telecommunications services.
Subject: D2

4.-- Wilkinson, P. (1998) "OPAC for Intranets". Information Dimensions 6: 23-31. http://www.InformationDimensions.com

Subject: F1

III.D. Description samples (Tokens: 6,261 / Types: 1,240)

1.-- Brinson, J.D. and M.F. Radcliffe (1996) "An Intellectual Property Law Primer for Multimedia and Web Developers". Pomona College, CA.
Subject: F6

2.-- "Dial-Up Scripting Command Language for Dial-Up Networking Scripting Support" (1995) Microsoft Corp.
Subject: F4

3.-- Fleischhauer, C. (1998) "Digital Formats for Content Reproductions". National Digital Library program, Library of Congress.
Subject: F6

4.-- International Telecommunication Union (1999) "Radio Communication study". Document US 8B/25, Kansas State University.
Subject: D2

5.-- Kahn, R. (1995) "A Framework for Distributed Digital Object Services". Corporation for National Research Initiatives, University of California at Berkeley.
Subject: A2

III.E. Discussion samples (Tokens: 15,127 / Types: 1,899)

1.-- Böok, M., K.M. Coop and S. Sucksdorff (1997) "Networking the Library". Helsinki City Library.
Subject: F6

2.-- "Introduction to M.4 and M.4VB". Dept. of Information Technology, University of Pittsburgh.
Subject: B3

3.-- Loughry, D.C. (1995) "IEEE-SA and Institute Interactions". IEEE Board of Directors, Connecticut.
Subject: D2

III.F. Exemplification samples (Tokens: 10,138 / Types: 1,453)

1.-- Crocca, W.T. and W.L. Anderson (1995) "Delivering Technology for Digital Libraries: Experiences as Vendors". Xerox Corporation, Fairport, NY.
Subject: E1

2.-- Crocker, D. (1995) “To Be on the Internet”. Network Working Group. Brandenburg Co.

Subject: F2

3.-- Library of Congress, The (1996) “National Digital Library Program”. Washington D.C.

Subject: C3

4.-- Moulton, L.W. (1997) “Starting a New Database”. Comstow Information Services.

Subject: B2

5.-- Orost, J. (1995) “The Bench++ Benchmark suite”. AT&T Labs.

Subject: A1