Language Learning & Technology
http://llt.msu.edu/vol5num3/curado/

September 2001, Vol. 5, Num. 3
pp. 106-129

LEXICAL BEHAVIOUR IN ACADEMIC AND TECHNICAL CORPORA: IMPLICATIONS FOR ESP DEVELOPMENT

Alejandro Curado Fuentes
University of Extremadura, Spain

ABSTRACT

Lexical approaches to Academic and Technical English have been well documented by scholars from as early as Cowie (1978). More recent work demonstrates how computer technology can assist in the effective analysis of corpus-based data (Cowie, 1998; Pedersen, 1995; Scott, 2000). For teaching purposes, this recent research has shown that the distinction between common coreness and diversity is a crucial issue. This paper outlines a way of dealing with vocabulary in English for Academic Purposes (EAP) instruction in the light of insights provided by empirical observation. Focusing mainly on collocation in the context of English for Specific Purposes (ESP), and, more precisely, within English for Information Science and Technology, we show how the results of the contrastive study of lexical items in small specific corpora can become the basis for teaching / learning ESP at the tertiary level. In the process of this study, an account is given of the functions of academic and technical lexis, aspects of keywords and word frequency are defined, and the value of corpus-derived collocation information is demonstrated for the specific textual environment.

INTRODUCTION

The areas of English for Specific Purposes (ESP) and corpus-based lexical studies seem to converge in the study of terminology (cf. Pedersen, 1995). The main aim in terminology studies is to create specialised dictionaries that reflect knowledge fields and concepts where these are related to the property of lexical use restriction.[1] In the textual collections, collocations play an essential role in the description of this specific language usage (Pedersen, 1995, p. 61). In this sense, word combinations work as building blocks that increase the learner's potential to command special languages.

However, the results of technical collocation studies have little to offer students for academic performance and achievement: that is, they do not help learners meet the "stylistic expectations of the academic community" (Cowie, 1998, p. 12). This is because of the fact that in addition to the specialised terminology, there are other types of combinations that greatly influence the ESP learning context: for example, *seek the objective, consider my suggestion, the theory is canvassed, argue rather less vehemently,* and many other examples of academic discourse (Cowie, 1978, p. 132).

Our approach is precisely based on the distinction between technical and academic word behaviour. We are influenced by lexicography where this this double perspective is exploited (e.g., Lozano Palacios, 1999) according to whom general academic vocabulary is distinguished from more specific word use.

Lexical levels or categories are fostered and described through the application of corpus-based studies. The design of a fit corpus is of prime importance so that lexical profiles can be developed effectively. This means that aspects such as size, type, balance, and integration of texts must be defined from the scope of ESP. In this line of work, small representative corpora are favoured for specific purposes (Tribble, 1997, p. 116).

In addition, an electronic concordancer such as *WordSmith Tools* (Scott, 1996) is rather useful to handle reduced text collections (Tribble, 2000). This includes dealing with differences between one given genre and the reference corpus, or between one specific theme and the overall body of subject texts (Scott, 1997). The results obtained are Keywords, which signal the "aboutness" of the texts (Scott, 2000), and thus receive primary observation in restricted language measurement. General word usage, in contrast, is derived from lexical surveys across subject boundaries. These are examined through critical concordance data, also known as KWIC -- Key Word In Context.

With these notions in mind, particular subject areas are represented by specific corpora. The size and type of the sources can vary, depending on how similar or different the topics are. For instance, related disciplines within the broad domain of Health Sciences can be grouped together (e.g., Nursing, Occupational Therapy, Medicine), because they share knowledge fields. Yet, their organisation and distribution in a specific corpus may present thematic variables due to emphasis on a given branch alone (e.g., Sports Medicine).

These selection principles are conceived according to interests and priorities in university programs and syllabi. In this respect, the domain selected in our research includes some current Information Science and Technology areas, such as Computer Science and Engineering, Optical and Radio Communications, Librarianship and Information Management, and Audio-visual Communications. These degrees are the main headings of our subject area sub-corpora; they are also majors that have been recently incorporated at our university (1995 - 2001).[2]

Due to the fact that changes take place very rapidly in these disciplines, the texts in the corpus should be regularly updated. A five-year time margin is recommended by some of our colleagues as a suitable renewal period. This suggests that we select, for instance, academic textbooks and research articles that have been published recently. In addition, information obtained from the Internet is favoured, since such feedback also tends to be up-to-date. This technical material is assessed conveniently, not only for university studies, but also for future careers where instructions are mostly read in English. As a result, the selection of the sources is made according to two chief principles: the importance of academic readings for tertiary level education and the consideration of technical material for both college and work situations.

The principal objective of this paper is the classification of different lexical categories in English for Information Science and Technology. In this respect, the basis or point of departure is a lexical common core, described in contrast with the diversity of word use. Keywords and word frequency constitute the basic tools for working with this language variation. Collocation information is the main means for observing these linguistic traits in our context. The notion of collocation pervades this analysis of technical and academic constructions in ESP development.

METHODOLOGICAL ISSUES

From our viewpoint, the examination of lexical data in small corpora is related to the analysis of specific purpose languages. This relationship motivates our selection and arrangement of sources according to two main factors: ESP focal points (internal) and contextual conditions (external).

Under the first parameter, texts are updated in terms of subject matter. Dudley-Evans and St. Johns (1998, p. 99) claim that this search for novelty is crucial in ESP; the aim is that language reflects current issues in Science and Technology, where the tendency is for "carrier content" to "date rapidly" (p. 174).

A second internal factor is that material be authentic. This means that texts should be required or recommended in university courses (James, 1994), and that different genres should be included (Conrad, 1996). For instance, a primary or introductory stage involves textbook discourse -- aimed at fulfilling learning demands in first and second year university studies (Johns, 1997, p. 46). Then, technical writing

in reports contains appropriate language for intermediate levels (Bergenholtz & Tarp, 1995, p. 19). Likewise, research papers and articles tend to meet the advanced needs of research students (Brennan & van Naerssen, 1989, p. 202).

The third internal point in our methodology considers textual availability. A priority is that texts are managed electronically. Documentation in electronic format is needed for concordance procedures. Therefore, with the increased production of texts in this manner, students can become genuine users of corpus resources for learning purposes (Johns, 1993). ESP instructors can then work as supervisors of learner-centred reading tasks. The design of the corpus can be carried out by both instructor and learner alike; the former directs operations according to language interests, and the latter contributes special interest topics in the subject area.

Criteria outside the ESP learning context are also applied. These external conditions influence corpus selection because study programs and syllabi must be accounted for as relevant to subject matter. In our institution and similar centres in Spain, they offer guidance for the arrangement of the sources. A contrastive examination of university curricula is encouraged to identify common subjects, taught in more than one of the four disciplines mentioned: Computer Science and Engineering, Optical and Radio Communications, Librarianship and Information Management, and Audio-visual Communication. Shared fields are labeled as subject categories in Table 1, according to the data derived from Information Science and Technology study programs.[3]

Table 1. Subjects Shared by Disciplines

A	Computer Science/Engineering and Optical/Radio Communications
A**1**	History of computers, Hardware, Software
A**2**	Computer engineering and architecture, Data communications and Client-server architecture
B	Librarianship/Information Management, Computer Science/Engineering and Optical/Radio Communications
B**1**	Information units management
B**2**	Online database systems, Computer systems
B**3**	Automated Knowledge-based systems
C	Librarianship/Information Management and Audio-Visual Communication
C**1**	Content analysis
C**2**	Media documentation
C**3**	Documentation Legislation
D	Optical/Radio Communications and Audio-Visual Communication
D**1**	Media technology
D**2**	Media theory
E	Librarianship/Information Management, Optical/Radio Communications, and Audio-Visual
E**1**	Communication
	Communication Theory
F	All Four Disciplines
F**1**	Perspectives on Information
F**2**	UNIX / Internet
F**3**	HTML, SGML, TEI
F**4**	Hypertext technology
F**5**	Electronic publishing
F**6**	Information infrastructure

Texts are selected according to their relevance in the subjects -- A1 to F6 (Table 1). The reading material is either offered in the courses, or recommended by content instructors. For example, textbook chapters on the history of computers, hardware, and software (label A1) are part of the book *Computer Language*

(Díaz & Jones, 1999), suggested as further reading in the named introductory course for Optical/Radio Communication students. In contrast, a research article like "The Audience as Reader" (Callev, 2000), belongs as reference material in Content Analysis (subject C1); it provides technical reading that is helpful for project reports in both Audio-Visual Communication and Librarianship/Information Management studies.

The inclusion of different academic genres balances the corpus. The goal is to provide a representative collection of the subject areas in the learning context of our institution, where specific language competence is mainly demanded for reading and writing. Thus, questions about lexical features are addressed by using the right corpus (Biber, Conrad, & Reppen, 1998). Figure 1 illustrates how our sources can be balanced according to genre and subject area synchronisation.

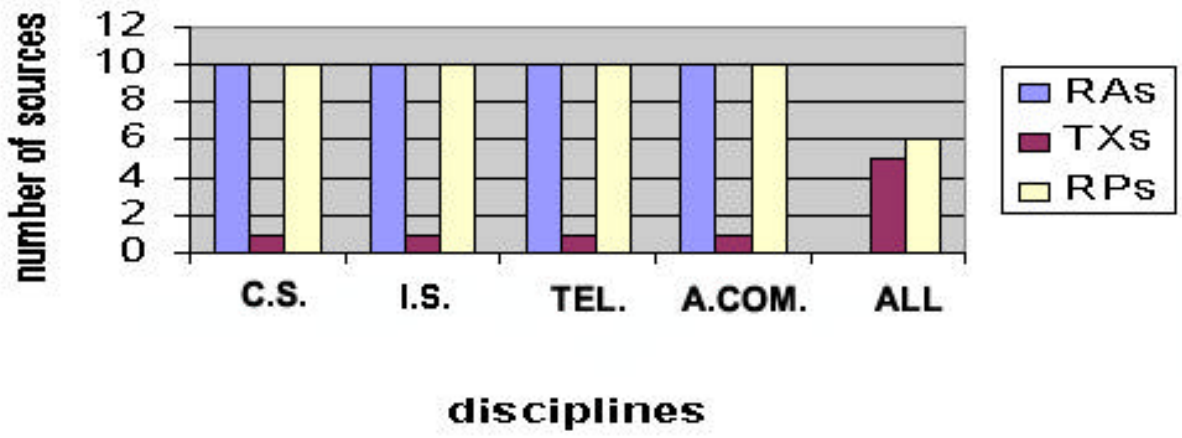

Figure 1. Distribution of sources in corpus

C.S. = Computer Science / Engineering	RAs = Research Articles
I.S. = Information Science (Librarianship / Information management)	TXs = Textbooks
Tel = Telecommunications (Optical / Radio Communications)	RPs = Technical Reports
A.Com = Audio-visual Communication	
All = All four disciplines	

The disciplines serve as reference for textual selection. According to this notion, each of the four areas includes an equal number of sources in each genre. Ten research articles, for example, deal with Computer Science / Engineering topics, drawn from bibliography lists in this discipline. However, the concepts are also examined in other study programs, such as Optic / Radio Communications. The same applies to the other cases, where university curricula provide feedback about reading requirements; these are double-checked by following programs and consulting colleagues in the subject areas.[4]

Figure 1 presents an additional set of sources: five textbook excerpts and six technical reports. These deal with the field of Business Technology, which appears as common core in all the subject areas. It cannot be distinguished as predominant within one single domain, but quite the opposite, it is a complementary part of all the different areas. Its importance is derived from not only study programs, but also the current Spanish job market. For instance, a report entitled "The Do's and Don'ts of Technology Planning" (*FECT & NECC Conference*, 1999) summarises Information Infrastructure issues (category F6, Table 1), which are commonplace in careers related to Information Science and Technology.[5]

The overall corpus does not exceed one million words. The purpose of this limit is to attain specificity. In this sense, a reduced size demands a precise representation of the specialised language. Figure 2 shows

the number of running words (tokens) and distinctive items (types) for each of our genre sub-corpora. Standardised ratios (types per 1,000 words) are also contrasted.

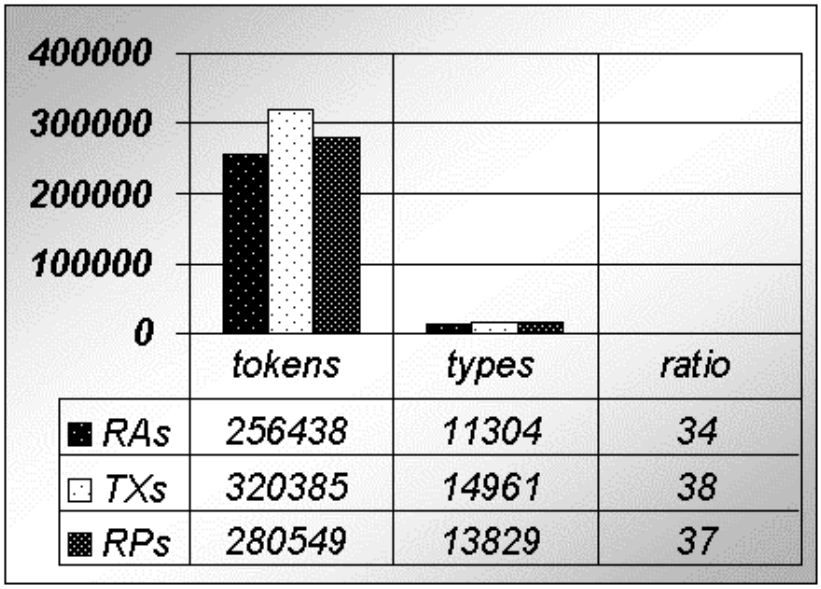

Figure 2. Word distribution

WordSmith Tools provides the basic functions -- Keywords and Collocates -- which perform likelihood tests and Mutual Information measurements. These are made on the corpus to generate a quantitative view of lexical behaviour (cf. Ooi, 1998). Wordlists, another main feature, constitute the cornerstone on which to start the gathering of data. By cross-tabulating wordlists, keywords are obtained. A given sub-corpus (e.g., a subject category in Table 1) is contrasted with the overall reference corpus. The resulting group of words tends to be rather descriptive of the context aimed at. In this respect, the relationship between lexical items and text seems to be bi-directional, as words serve to identify context, and this, in turn, influences the particular bonding of elements.

The results derived from this type of analysis are offered in the following section. The measurement of the data is carried out to observe lexical patterns, and, thus, a convenient classification of words can be made. Then, in the discussion, the significance of the data is assessed for ESP development.

LEXICAL RESULTS

Lexical findings are examined in context. This means that linguistic input is obtained by observing word combinations that are meaningful in the subject and genre domains.

We use concordances to reflect the significance of lexical patterns in specific contexts (Firth, 1957; Halliday, 1966); this implementation constitutes the basis of our work. The contrastive view of the data provides the necessary conditions to check lexical diversity and uniformity in the corpus. The aim is to describe genre and subject matter variables. For this analysis, the operations are ordered as follows: observing, measuring, and classifying lexical data.

To illustrate this analytical procedure, an example is provided with the cluster, *provide access to*, used extensively throughout the corpus. We observe this presence in all the genres and in several subjects, and thus measure its frequency and dispersion in the whole corpus. This is done to make sure that it occurs significantly as a general expression. In this respect, the assertion made about its classification is based on empirical factors.

Unbiased data are those based on lexical behaviour in context (KWIC), which can reveal how common a given expression really is. We determine that the relationship between concordance lines and the number of different sources contained in those lines informs about the type of lexical item. In this sense, the example *provide access to* is analysed according to a 0.3% cut-off point, meaning that, above that level, its occurrence is considered common core in our texts, and, consequently, free and general. This margin higher than 0.3 % refers to the number of sources in the concordance lines: For every 10 lines of concordance text, at least three different sources must be involved.

In addition, we consider that the three genres should be present in the total concordance, and that at least six different subjects must be included. These numbers are considered reliable, since our corpus is not very large. We believe, in fact, that 95 texts, 17 subjects, and three genres are low numbers in comparison with bigger corpora, and that 0.3 % is an appropriate measurement as a result.

As an example of this computation, the collocation *directory service operations* is observed. It is recorded as key within subject domain A -- belonging to the areas of Computer Science and Engineering and Optical/Radio Communication. It appears 70 times but only in three texts, yielding a 0.04% contextual margin. We thus regard this lexical manifestation as specific and restricted, in contrast to the free and general case of *provide access to. Directory service operations* actually behaves as a specialised collocation, in agreement with Pedersen (1995), and, as such, tends to form complex nominal compounds (see also Varantola, 1984).

Finally, lexical elements that have a high frequency in the corpus, but are predominant within one single genre, also deserve attention. They tend to operate as restricted word combinations, but do not denote technical or specialised meanings. Instead, they form compounds of a semi-technical type. An example is *your program directory,* appearing in subjects A1, A2, B1, D1, F1, and F4 (see Table 1), and in 14 different sources. However, only the genre of technical reports contains these instances. This specificity makes these constructions genre-based.

Three main lexical sets thus constitute the object of our study in the results: general elements, specific items in defined contexts, and genre-based constructions.

General Elements

Detailed Consistency Lists (DCL) are made available through the concordancer. These are wordlists arranged according to the contrast of frequencies in different domains (e.g., in genres). For the listing of general academic items, they prove to be rather useful. In our corpus of Information Science and Technology sources, the DCL is considered an academic word list. It is similar to Coxhead's (1998), since it presents input that can become quite relevant for English for General Academic Purposes (EGAP).

Most of the lexical data in the DCL includes verbs and nouns, followed, to a lesser degree, by adjectives and adverbs. An important feature of academic language is that there are more verbs in the past tense and past participle (e.g., *defined, conceived, designed)* than there are present or gerund forms. The same happens, in fact, in Coxhead (1998). In the case of nouns, many correspond to common scientific-technical instances, such as *information, data, Web, HTML,* and *computer*.

Free word combinations result from examining the DCL. For example, the forms associated with the noun *information* are widespread throughout the corpus. They are analysed as free collocations, appearing in contexts that vary significantly in terms of subject matter. They are thus considered semi-technical elements. Some examples are *information system, information technology, digital information,* and *information about.*

In addition to these frequent items, lexical elements found at the bottom of the DCL, are likewise important. Despite having a low frequency, they exhibit contextual significance in their behaviour. An example is the inflected form *coined.* All six occurrences of this item denote academic use. The pattern of

Verb + Noun in the expression *coin + the term*, surfaces in this sense. It is declared academic because of its high degree of dispersion, as it shows up in different genres and subjects. These contexts are a textbook and a research article on information management units, another article on perspectives on information, and two technical reports on information infrastructure.

The term *coined* is judged as important in the DCL, as a result, not only due to its wide range of use, but also to its collocational strength -- denoting a great degree of idiomaticity (Stubbs, 1995). The diversity of contexts in which it is included makes it idiomatic. In addition, it is the only form in its lexical family appearing in all three genres and in more than one subject area. From this perspective, general academic terms can be either frequent or sparse, but they must always present a noteworthy dispersion.

Other examples of low frequency items are the following: *where this technology excels, imported into, select / edit / paste, cable hooked into the, was instrumental in* + verb-ing, *first and foremost, diskette drive, compounded by the, relative autonomy,* and *ticket booth.*

They all share the property of general academic vocabulary. A string like *was instrumental in* + verb-ing, or the cluster *compounded by the,* function as common core in our setting. The same occurs to a noun collocation such as *diskette drive* or *ticket booth.* They receive the same treatment, in this respect, as frequent word combinations, and match in importance the ones mentioned above, for example, *information system, information technology, digital information,* and so forth.

Among the examples of low frequency words, however, a distinct type of item emerges, and an alternative approach is inferred in its classification within the general academic vocabulary. These are the so-called lexical phrases -- for example, *first and foremost.* They tend to behave as procedural items in our context, being closely related to academic use (Stotsky, 1983; Thurstun & Candlin, 1998).[6] They also appear in a wide variety of contexts, functioning as grammar and discoursive markers. Their procedural status derives from the effect of signposting which they demonstrate in the texts. This characteristic is analysed as a rhetorical marking of functions and techniques. For example, they may indicate interaction with the reader, a reference to the text itself or to the investigation carried out.

How these items manifest different rhetorical uses can be checked, for instance, with the behaviour of the preposition *by*. Its variation is made plain by a contrastive view at the corpus. The preposition is seen to denote conventional agent utilisation in many passive clauses, for example, *claims made by the text,* but it can also serve as a highly frequent instrumentalisation device, for example, *by means of.* In addition, it is commonly used in classification statements such as *used by location* and *used by subject.* Finally, it is often included in descriptive phrases like *characterized by* and *defined by.*

This wide range of rhetorical expressions also affects content words. Nouns, for instance, are used in common clauses like *make use of.* These noun expressions appear in all three genres. Adverbs can also function in this way. Some examples are *more likely* and *more appropriately,* extensively produced in our corpus. Despite this inclusion of content words, grammar items such as the combinations mentioned above with *by,* prove to be the most extended type of rhetorical devices.

Specific Items in Defined Contexts

The function of Keywords in the electronic concordancer provides the means needed to describe terminology according to specific textual segments. This procedure is carried out with a given group of sources selected by subject. For example, topic A1 (History of Computers, Hardware and Software; see Table 1) is compared with the entire corpus, and word frequencies in both subject and reference collections are cross-tabulated. The resulting keywords are relevant not only in terms of frequency, but also textual dispersion. Thus, items like *Multics, segment, Minix, bit, ring, segments,* and *ATM,* appeal to the thematic essence of category A1.

This list of keywords demonstrates the importance of "key-ness" scores in *WordSmith Tools*. The percentages in this measurement indicate positive and negative keywords. A "key-ness" level above 25%, in this sense, contains words that are pivotal as subject items ("positive keywords" [Scott, 1997]). Their identification is made in defined contexts such as thematic sub-corpora, as these terms concentrate the "aboutness of the texts" (Scott, 2000). Most of these words are nouns, combining as compounds that weigh heavily on field specialism, and they operate in restricted domains as subject descriptors.

Reference to specific notions is observed in *system project manager revision*, *automation project manager acceptance*, or *project manager report.* They are examples of key combinations derived from the collocation *project manager,* which has a high "key-ness" score in the topic of Automated Knowledge-Based Systems (heading B3 in Table 1). There are several instances of these long nominal compounds in our subject texts, which leads us to consider that the longer the noun compound, the more restrictively it tends to operate in the subject area (Pedersen, 1995).

Specific lexical structures thus reflect technical use, although this is not clear in all cases. For instance, the noun *library* is the top keyword in texts about Librarianship. It collocates with nouns that do not present any semantic complexity, as the instances *virtual library staff*, *connectivity on the library,* and *public library community* prove. Within subject F6 on Information Infrastructure, in fact, these elements specify the procedure of electronic information organisation, but do not offer much comprehension difficulty.

The generality is that most keywords tend to sum up the thematic content of the texts. Several are actually quite descriptive at first sight, such as *images* and *media* in the area of Document Content Analysis (C1), or *copyright* and *contractor* in Document Legislation (C3).

Keywords can also be obtained in the contrastive analysis of two or more disciplines. In this case, they originate from two lists of subjects, for example, A1 (History of Computers, Hardware and Software) and A2 (Computer Engineering and Architecture, Data Communications, and Client-server Architecture). The findings are then identified as broader in scope (for example, applicable to both Computer Science and Optical / Radio Communications), presenting, as a result, a less restricted subject-based pattern. Some examples in this thematic group (A1, A2) include *bits*, *hardware*, *directory*, *IP*, *software*, and *PC.* These items result from contrasting the smaller, theme-restricted context with the overall corpus, as pointed out above.

The data prove to be crucial for the constitution of lexical profiles in the texts. A similar deduction is made by working with the four separate areas, that is,, Computer Science / Engineering, Optical / Radio Communications, Librarianship / Information Management, and Audio-visual Communication. The lexical information analysed in this case can be valuable as a guide to specialised language, much like dictionaries and other lexicographic material are (e.g., Collin's *Dictionary of Computing*, 1999, or the *TERMITE Database of Telecommunications*, 1999).

In this respect, our data can be contrasted with authoritative sources to check similarity / variation features. For instance, in the case of the form *abandoned,* examined in Collin's *Dictionary of Information Technology* (1997), the phrase *abandoned the spreadsheet* is given. In our sources, the clause *the code had to be abandoned,* is similar to that example. This contrastive view is highly recommended from our perspective, due to the fact that the dictionaries and glossaries handled are recent. They therefore provide updated material for linguistic analysis.

Table 2 displays the top three words surveyed in this way. The feedback conveys meaningful discipline-based content, but also diversity of lexical use.

Table 2. Main Items in Discipline-Based Sub-Corpora

Computer Science	Librarianship/Information Management	Optical/Radio Communications	Audio-Visual Communication
Program Data System	Library Information Use	Network System Data	Digital Media Server

Two large corpora are included in this cross-examination: The first two columns (from left to right) correspond to James (1994) and Lozano Palacios (1999). The two sources offer ranked positions of words, which is quite useful for academic purposes, since, as in Coxhead (1998) above, the most frequent verbs and nouns combine critically.

For instance, Lozano Palacios (1999) reports on Verb + Noun and Noun + Noun patterns. The items are deemed as essential academic data. The clusters *provide + access to,* and *data collection + techniques, instruments, methods* are two relevant examples. The former is considered a general academic expression in our corpus, according to the description of the section "General Elements." The latter is specific and subject-based, appearing more frequently within the F domain. However, the compound *data collection +* Noun is not evaluated as strictly technical, mainly due to the common coreness aspect that characterises setting F (see Table 1).

The two main aspects revised, academic and subject-based language, thus seem to merge in discipline-driven vocabulary. The effect produced by such words seems neither common nor restricted. These items would be found at a middle position between general and specialised vocabulary in our study.

Genre-Based Constructions

Constructions that occur across different subjects, but only in one single genre, are also accounted for. These elements are rather frequent in various texts, much like the discipline-based items examined above. Nevertheless, these genre-based combinations are namely treated as specific academic language. Some examples common in technical reports include *information object, networked information services,* and *information on the Web.*

As mentioned in General Elements, the DCL forms the bulk of contrasted vocabulary. The three genres are compared, and their word frequencies serve to establish measurement references. The items identified as relevant in these word lists have a high frequency in one genre alone, and, in contrast, very low usage in the other two contexts. In addition, a Keyword analysis is carried out on the top words of the genre lists in order to check that the lexical items are actually distinctive in their genre categories. The results are classified from most to least typical in terms of genre description; the former operate as positive keywords, and the latter as negative.

An example of a highly positive keyword in the genre of textbooks is *requirement.* Another is *Semiotics.* They represent the two chief types of keywords in this environment: Widely extended across subject areas in the first case (for example, *the following requirement*), and restricted to particular subjects in the case of *Semiotics* within Audio-Visual Communication.

In a different genre, technical reports, the top keyword *library* appears quite frequently. It combines within clusters and compounds more or less familiar, as observed in units like *Cable Book Library, the library's clientele, library program, networking the library, inter-library lending, inter-library loan,* and so forth. In contrast, the noun *protection* illustrates low frequency items in these reports. Despite its fewer occurrences, important grammatical forms such as *protection from* and *protection for,* can be pinpointed. Other significant lexical items which occur with *protection* include *fire protection, copyright protection, protection criteria,* and *protection levels.*

In research articles, the top item is *project,* apparently uncharacteristic and yet, intimately linked to the research activity. Some frequent collocates show endemic traits in this respect: *project deadline, project work, project milestone, project manager, project revision,* and so on.

With the analysis of this data, effective samples are drawn to back up our claims for the next section. Lexical information is then assessed, and implications for ESP development are reflected upon.

DISCUSSION

In the survey of results, three main divisions of lexical behaviour have been found: General academic vocabulary that occurs widely across the corpus, with presence in a diversity of topics; elements drawn from subject matter scrutiny, considered specialised or technical; and genre-driven findings occurring restrictively, thus characteristic of one single genre.

In this section, our main aim is to assess this lexical co-occurrence in its context to determine validity for teaching ESP. The process of language acquisition, in this respect, should be evaluated according to the contextual variables analysed. We approach the description of academic and technical constructions in our environment by evaluating their use in either a great or small number of texts. In both cases, lexical units are influenced by subject matter and academic discourse.

Word combination significance is mainly determined through language task application. This means that effectiveness of data is judged by specific language instruction: "To teach language for the subject specialism," and "teaching tasks based on the specialized content" (Edwards, 1996, p. 13). This evaluation leads to categorising lexical data as priority items for ESP and EAP courses (Jordan, 1997).

Eight types of lexical units are consequently devised. They result from the detailed revision of the data in the previous section, and from how such items serve to fulfil specific language learning demands in our context. They are classified as follows: common core collocations, rhetorical academic elements, technical collocations, thematic combinations, area-based general words, area-based specific words, genre-based academic vocabulary, and genre-based thematic words.

Common Core Collocations

A main group of lexical elements is first inferred by focusing on those words that occur commonly. This level is measured across subject areas, which constitute a common core foreground where constructions are used by "authors writing on similar topics" (Stotsky, 1983, p. 438). The items receive a semi-technical treatment primarily because they are content words conceived, in agreement with Ewer (1983, p. 10), as a "number of language items which are common to the subjects," or as the "core language."

In our scientific-technical context, this semi-technical degree derives from word behaviour registered at a general academic stage. The elements become core combinations related to the academic context. In this sense, they are viewed as

> 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. (Farrell, 1990, p. 11)

Academic combinations function as lexical extensions of General English vocabulary in our specialised corpus. In other words, their meaning is familiar in academic discourse and common in the Information Science and Technology domain, since the expressions denote events and concepts that characterise this area. This language is more general than specific because it describes notions and ideas that are customary in the whole corpus.

As described in the Results above, common core academic elements have high frequency and dispersion rates across sources, such as in the case of the collocations *information technology* and *digital*

information. In contrast, lexical items can show low frequencies, but the number of texts included then is also high by comparison, and the offer of topics is diverse, for example, the aforementioned combination *coined the term.* The cut-off point that distinguishes both planes is 0.3 %, as mentioned previously (three texts for every 10 concordance lines).

Table 3 is an example of a general academic entry. The lemma is *address* (representing both verb and noun), and its derived forms are *addressed* and *addresses*, which are also considered common academic words in our corpus. The number of instances is provided near the entry and divided into the three genre sub-corpus frequencies. Table 3 is organised according to frequency, from highest repetition to least; the most repeated combination is labelled with the times it occurs (shown in brackets).

Table 3. Example of General Academic Entry in our Corpus

	TXs	RPs	RAs
ADDRESS	128	317	63
	___ space (137)	*the same ___*	
	whose ___	*must ___*	
	does not ___	*the network ___*	
	to ___ this	*___ the issue*	
ADDRESSED	47	15	36
	to be ___ (14)	*___ in*	*is ___ by*
	should be ___	*___ here*	
	has ___	*can be ___*	
ADDRESSES	42	33	14
	IP ___ (17)		

TXs = Textbooks; RPs = Technical reports; RAs = Research articles
BOLD = lemma (most frequent item in its lexical family)
UNDERLINE = word forms (less frequent) derived from the lemma

As can be deduced, the collocations in Table 3 are based on verb forms (e.g., *address + the issue)* and nouns (e.g., *address + space).* These are content word associations, similar to the ones that the *BBI Dictionary of English Word Combinations* (Benson, Benson, & Ilson, 1997) describes. This source actually includes common academic items from the world of Information Technology: *access data, browse the web,* and so forth (Benson, Benson, & Ilson, 1997, p. vii).

The combination of grammar items and verb forms is also evaluated at this level of general academic use. Some examples are shown in the section of *addressed* (Table 3) -- for example, *addressed in,* and *addressed by.* These are regarded as general collocations, as a result, since they are found in many different texts. Grammar constructions are academic collocations in this respect. The *BBI Dictionary of English Word Combinations* (Benson et al., 1997) distinguishes grammar from content collocations, but, in our case, this is not so. In our analysis, grammar combinations work at the same common core plane as academic collocations.

We deduce our claim from the management of word lists as academic input for ESP development. Learners are encouraged to carry out lexical profiles when coping with academic reading. Such a chore implies, as a matter of fact, coming to terms with the DCL in the different genres, pinpointing constructions that are common core and typical in the texts. The collocations examined in Table 3 are classified in the form of lexical charts, where the most frequent items are contrasted with less common constructions.

In this line of work, most learners do not differentiate grammar from lexical combinations. This turns out positively in our context of science and technology, since undergraduates are not used to making syntactic

observations. Actually, students tend to carry out an integrated approach, in which content items function as main units that may or may not keep company with grammar words.

For example, given a common collocation such as *address the issue* (Table 3), the lack of a preposition is learned by building collocation charts. These are exploited from both the readings and the DCL, taking the whole corpus as reference. In a related exercise, synonyms are explored. For instance, the academic verb *cope with,* is examined in combination with the noun *the issue.* In such a comparison, students value the collocational strength of the preposition *with* for the construction, in opposition to *address,* which does not demand the colligation.

This drill is performed similarly with low frequency items. The main difference is that patterns are recognised by working with a small amount of lexical information. Useful combinations are then easily perceived, as different subjects are encompassed by that occurrence. A previous example, *coined the term* illustrates this case. Its free distribution is observed when we can detect that the phrase actually occurs in various texts. In addition, its common coreness is reinforced by contrasting synonyms such as *built the term* or *constructed the term,* since these are used in fewer contexts. Students can then be guided to value the more fixed meaning of the verb *coin,* given the fact that synonymous combinations show a lower dispersion rate.

Rhetorical Academic Elements

Rhetorical items also demonstrate common core relevance due to their high frequencies and distribution. They are used as markers of cohesion in the texts, according to the Results section. They tend to convey procedural usage, a feature that relates them to academic elements. Some of the examples mentioned are *by means of,* indicating instrumentalization, and *more likely,* operating as a token of clarification in the sentences.

These constructions are classified at the same level as general academic language. Their procedural status defines them as common core, in agreement with McCarthy (1990, p. 51), and with Hutchinson and Waters (1981, p, 65): They serve as instruments of coherence and cohesion throughout discourse. Some procedural nouns functioning this way are *the use of* and *the device which.*

This language is analysed under the EAP umbrella, which includes EST (English for Science and Technology). In this learning framework, comprehension activities are favoured, as they challenge learners to cope with markers of discourse structure (Flowerdew & Miller, 1997). For academic lectures, in fact, main and secondary ideas are discerned in the texts by exploiting these markers appropriately. For Science and Technology discourse, learners demonstrate their comprehension of content by conceiving appropriate rhetorical boxes (e.g., classifications, explanations, descriptions; Bygate, 1987). These are often built as a result of the adequate interpretation of rhetorical elements.

A suitable exercise is based on the search for lexical formations containing a common grammar word. This type of work allows for the exploitation of procedural language in our corpus. It aims to identify vocabulary that co-occurs typically at the general academic plane. For example, measuring the occurrences of the preposition *by,* as examined in the Results above, provides different semantic features of scientific-technical discourse, for example, denoting functions as agent, instrument, and so forth.

Figure 3 demonstrates another example with the preposition *within.* The word is analysed in different contexts so that learners can contrast its different meanings. This task of inducing sense depends on the main contextual conditions found; some authors refer to this activity as semantic prosody analysis (Stubbs, 1995). It results from the qualitative observation of common core collocations. In this case, the expressions and word combinations convey a strong procedural meaning, since they signal the type of discourse function being used.

WITHIN	*Procedural elements in use*
ELEMENTS	SEMANTIC PROSODY
___ *information*	WITHIN + CATEGORY
___ *software* / ___ *the project* / ___ *a commercial context*	WITHIN + LOCATION
___ (+ a scheduled time)	WITHIN + TIME
__ *headings* / ___ (+ *document*)	WITHIN (INSIDE)

Figure 3. Inferring procedural meaning in discourse

Technical Collocations

The level of technicality in word behaviour is closely related to subject domain. The salient condition is that elements function uniquely in their corresponding field, describing the restricted setting. An example is the range of specific combinations identified with the noun *network* in *U-network, access network, local area network*, and so forth. This is examined within the subject of Client-Server Communication (category A2 in Table 1). The items thus allude to concepts and developments in specialised areas, and their interpretation demands conceptual knowledge. In addition, abbreviations are often key in this context, which is also evidence of the specific understanding that is required at this learning stage, for example, *bit ASCII, LAN distribution*, and *GIF and JPEG files.*

Conceptually restrained, technical vocabulary is formed by collocations that introduce specialised knowledge in ESP. The identification of this special language is made by inferring idiomatic constructions from concordance samples. The aim is to perceive the fixation of long compounds, and to appreciate the value of this lexical restriction in the subjects.

Figure 4 displays the technical collocations of *object*, a critical noun in the setting of Data Communications (category A2 in Table 1). An important collocation like *object-oriented* is first underlined by focusing on the most frequent word that goes with *object* in this context. Then, *object-oriented features* is also marked as important, given its high co-occurrence probability. Finally, according to the Mutual Information scores, the phrase *use of object-oriented features* is recorded in this technical scope.

N	Concordance
1	0 parameters is a class object with three int me
2	s, 5 stores, and 1 class object copy. H
3	he "a" suffix is written in object-oriented C++ styl
4	d P13 where the class object with 3-ints is cop
5	g++ in optimizing object oriented code, tho
6	the appropriate use of object oriented features.
7	into C++, making use of object oriented features
8	at generating code for object oriented construct
9	ck and then unpack an object.
10	a 10-component class object. Ratio: 0.52

Figure 4. Concordance sample for the noun *object* in a technical setting

We can guide learners to explore technical terminology by encouraging this data classification. An example is the relation of the collocations hierarchically. This can be achieved as follows:

OBJECT (subject: Data Communications)
Object
Object-oriented
Object-oriented features
Use of object-oriented features

In order to determine which combinations are productive, students operate with restrictive lexical charts. For instance, a useful type of activity is a filling-in-the-gap exercise where the ability to specify technical items is fostered. Learners pinpoint the restricted elements of subject texts by building tables such as the one shown in Figure 5. In this case, coming to terms with the central collocate in four different combinations demonstrates technical language command.

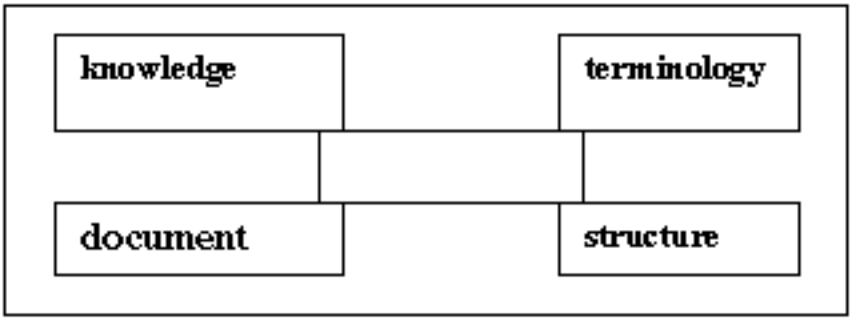

Figure 5. Fill-in-the-gap exercise with technical collocations

In Figure 5, learners take subject B3 as reference for lexical work (texts about Automated Knowledge-Based Systems). The answer to the central node can be found by working with the language of these sources. This means that students must revise concordance material and context as indicated in Figure 4. The word in the blank -- *management* -- can be realised after key technical input is correctly sifted.

Thematic Combinations

Semantic features are examined in technical words by inspecting the subject context. However, exploring the field of knowledge does not always lead to the description of specialised combinations, according to our data. There are forms of lexical behaviour, in fact, which occur critically in the thematic environment but do not classify as technical collocations. These are content words with a less complex level of comprehension, namely due to their greater familiarity in the world of Information Science and Technology. Some examples given in the Results are *virtual library staff, connectivity on the library,* and *public library community,* included in subject F6 (Information Infrastructure).

Other examples reflect the register of a subject in a clear way. For instance, the legislation language of category C3 is clearly revealed through key clauses such as *the contractor shall* and *copyright law* (see Results). The constructions are either specific clusters or multi-word units that identify the subject under analysis. Other elements can be located in the area of Content Analysis (C1), where *mass media* and *of the mass media* operate as typical constructions within their thematic setting. In addition, like technical collocations, these items are almost exclusive of their domain and thus seldom found in any other part of our corpus.

At this level of thematic combinations, we also find lexical data that is characteristic of a related group of subjects, that is, within a major heading from A to F in Table 1. As a result, the language items described in this case are not as precise as technical collocations. For example, key combinations are analysed in the space where Computer Science and Optical/Radio Communications meet (category A). The results refer

to computer and network issues, but do not posit much technical difficulty. Some of these items are *computer program, hardware and software, bits per second,* and *the interface shall provide.*

The analysis is therefore based on language that is segmented according to main subject categories. The keywords that emerge from this task are evaluated in terms of technical use (as was done in the section Technical Collocations). In this observation, we notice that restrictive collocates are not detected. In contrast, a wider possibility of combinations is offered. For example, the synonym *computer application* is found alongside *computer program* in subject domain A, although the former is used less frequently.

This aspect of thematic combinations suggests activities where learners can make lexical decisions. These are based on choices by which synonymous thematic word combinations are explored. Students are given the freedom to decide which structures best fit the topic areas. Some possibilities are offered in Table 4 for the shared background of Computer Science and Optical/Radio Communications (letter A in Table 1).

Table 4. Investigating Synonymous Thematic Combinations

COMPUTER & TELECOMMUNICATIONS WORDS	
Computer program	**=**
The interface shall provide	**=**
A string of bits	**=**
Possible links	**=**
Multiple processes	**=**
Piece of software	**=**

The method presented in Table 4 is a constrastive view with synonyms located anywhere in the corpus. Thematic combinations are thus distinguished from common core items, such as *computer program* versus *computer application* or *the interface shall provide* versus *the interface provides.* Lexical evaluation is then possible by contrasting thematic constructions with general use. In this manner, the level of specification of the former can be appreciated.

The same is applied to other textual segments, for example, to subject items that are not technical. The F6 category examples, for example, *virtual library staff* or *connectivity on the library,* are assessed in this manner, being replaced by common core options like *library personnel* and *connecting virtual libraries.*

The purpose of this work with thematic vocabulary is to value concordance data in different textual positions. In other words, the goal is to train learners in the aspect of lexical variation, which encourages operation according to context; this is a consistent position from our viewpoint at all levels.

Area-Based General Words

The goal at this stage is to describe how language develops within a single discipline of Information Science and Technology. The items are familiar in all four areas, but expound a characteristic tone in a particular one. An example is provided by the cluster *provide + access to.* This structure appears freely throughout the whole corpus, but it receives greater emphasis in Librarianship and Information Management, where its semantic prosody is revealed as *provide + access to + documentation;* this behaviour is confirmed in Lozano Palacios (1999).

Area-based lexical features contribute to enhancing academic word usage. As described in the section Common Core Collocations, common core academic elements are exploited in EGAP (English for General Academic Purposes). The same can be done in the case of area-based general constructions, since this input may be similarly used in EGAP courses. Such similarities at both common core and area-based levels are contrasted by means of specialised dictionaries, glossaries, and corpora about the academic

disciplines. These sources can supply linguistic information that allows learners to pinpoint similarities and differences.

Managing and making sense of dispersion plots, available through *WordSmith Tools* (Figure 6), can also be enriching for learners. The plots signal where certain items crop up in the texts, thus challenging students to cope with visual data. For example, the noun *access* manifests a high concentration in sources dealing with the area of Librarianship. Figure 6 shows this lexical clustering in rows 1, 2, and 3 -- corresponding to text files of reports (row 1) and textbooks (rows 2 and 3). The concordancer can then disclose whether *access* is, in fact, used as a noun in these contexts.

This activity should enable students to check, on their own, the high frequency of *provide access to* in our corpus. The dispersion plots help them to clarify that it is actually emphasised in Librarianship; concordance feedback do the same by allowing learners to examine the semantic prosody + *documentation.* The expression is consequently conceived as a general academic expression due to its common coreness; yet, students notice that it is more heavily used in the context of Librarianship Studies, denoting a special meaning.

The DCL (Detailed Consistency List) of the four disciplines included in our corpus also makes area-focused lexical use easy to perceive. Through this list, the frequency of **access** is seen as higher in Librarianship / Information Management texts (see Table 5).

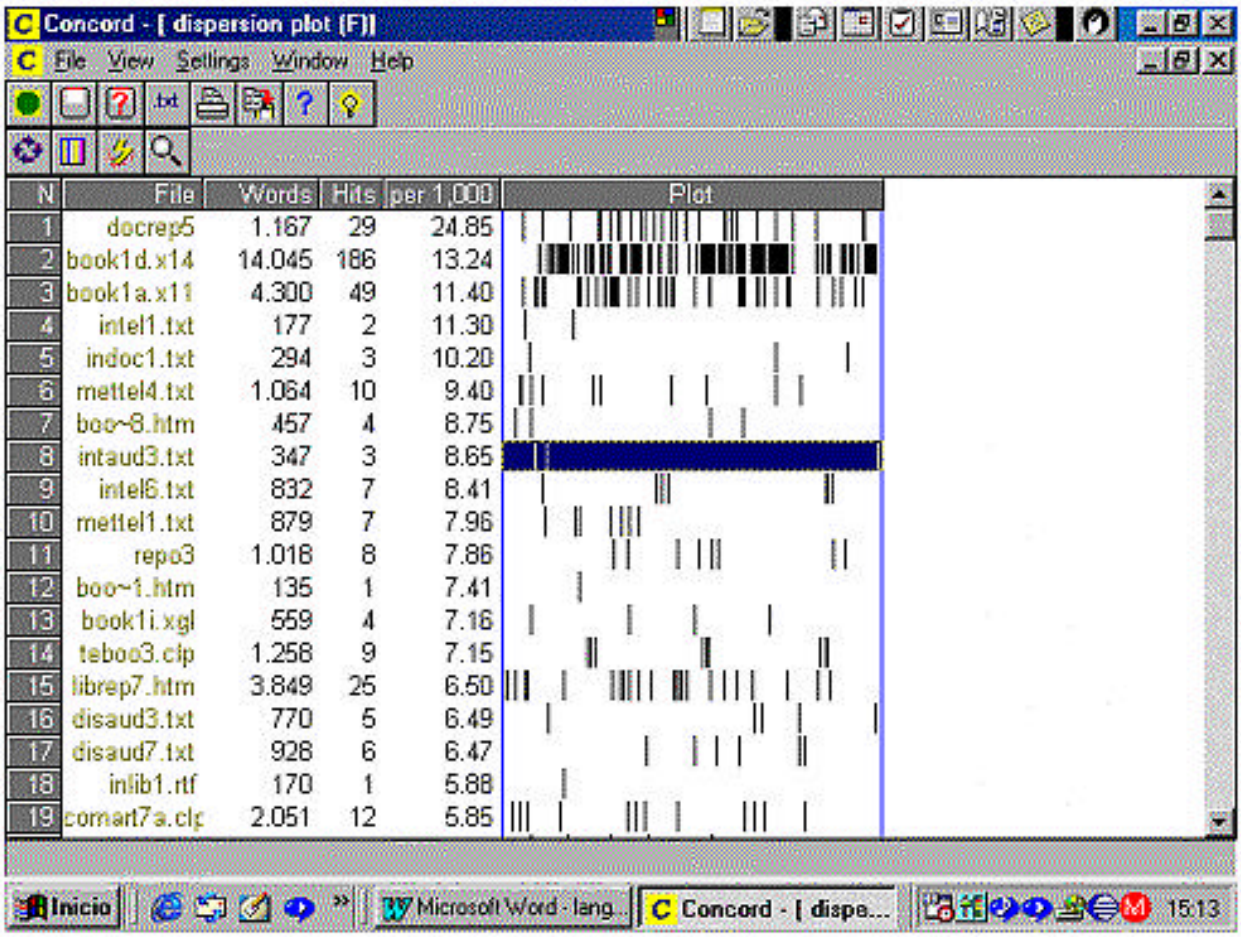

Figure 6. Use of dispersion plots by learners for visualisation of lexical concentration

Table 5. Availability of Frequencies in Discipline-Based DCL

N	Word	Files	Audio	Computer	Library	Telecom
68	Access	4	158	145	415	345

N = position occupied by word in DCL according to frequency and distribution in files

Area-Based Specific Words

The approach described in the section Area-Based General Words is geared towards practical work at the academic level of ESP. The goal is to foster guidance through particular lexical fields in area-based language. This turns out to be particularly useful in EGAP, where getting familiarised with reading material, for instance, becomes a powerful resource. Inspection of word use at this level also demands knowledge of specific concepts in the areas. In this sense, our study includes attention to particular lexical groups in the areas, where learners should thus specialise. An example of these items, examined in the Results, is the unit *data collection*, co-occurring with specific nouns such as *techniques*, *instruments*, and *methods*. These prove to be fixed word combinations, given their high co-occurrence rate.

The focus is then placed on English for Specific Academic Purposes (ESAP) learning (cf. Jordan, 1997). For ESP development, constructions involve a view into concept from this perspective. The approach is made as a response to specific queries regarding a subject area. For example, *data collection techniques* refers to the standard means of gathering data in Librarianship.

The underlying fact is that we investigate context, in this respect, to explore concepts. The activity demands learners examine conceptual paragraphs (cf. Trimble, 1985) that explain notions and clarify technicalities alluded to by the terminology. This contextual information can be exploited for task development; it constitutes support material, for example, for research preparation, that is, doing project reports in English.

Figure 7 presents a set of conceptual paragraphs taken from our corpus. Learners may use them for task-based research. The excerpts are assessed according to specific learning needs. They can then serve as complementary or illustrative material for project reports (e.g., as examples/passages to give in oral presentations).

A range of methods were employed to analyze data from the various **data collection instruments**. Quantitative data from the questionnaires, logs, training assessment, etc. were coded and entered in a spreadsheet for analysis. The techniques used to analyze these data relied primarily on computing averages and frequencies.
Develop and test a range of **data collection instruments** related to measuring the impact of Internet connectivity. Ultimately, the evaluation aspect of the project became the means by which a final report was developed for use by other public librarians and policymakers.
Phase 2 was intended to direct the development and administration of the various **data collection instruments**: What is the value of network connectivity for rural libraries? How does the installation and use of a network connection have impact on library staff, organization, and service provision? What groups in the user community benefit from the network connection?

Figure 7. Instances of specific concept development in an area

Genre-Based Academic Vocabulary

This group is determined from academic discourse study. However, unlike general words in Common Core Collocations, or area-related elements in Area-Based Specific Words, the identity of this level is based on the conception of genre. Awareness of genre features, in this respect (cf. Jordan, 1997), is the

prerogative. This is confirmed, for instance, in the case of advanced learners who are required to perform well in writing assignments.

An example of a lexical item in the genre focus is *project*, as mentioned in Results. This noun is often used in research papers to refer to the investigation being described. In a Computer Science setting, for instance, ***project*** is stressed in *project deadline*, *project manager*, *project work*, and so forth (see Results). The items become common in the genre of research articles. A relevant activity is to contrast genre-based instances such as these with general academic elements. The comparison aims to refine the view into genre-focused words, while general academic items are explored in the whole corpus. Table 6 provides an illustration of this comparative task in research articles.

GENRE-BASED (ARTICLES)	**GENERAL ACADEMIC**
project members	***for the project***
design process	***the design of***
search test	***of the search***

Table 6. Contrastive View of Research Article Items with Common Core Elements in Our Corpus

Lexical variation is thus visualised in the genre. The recognition is beneficial for learners aiming to develop effective writing skills. In particular, technical composition within the genre is enhanced by means of specific genre features. Coping with this should enable the ability to adapt to the conventions of an area like Information Science and Technology. In this regard, we favour an ESAP methodology for genre-based academic items, while the focus is also placed on scientific-technical writing.

Genre-Based Thematic Words

This last category also considers genre awareness as the main scope. The procedure by which this set of items is established falls under the ESAP application. This means that specific language is exploited in tasks designed to make genre features familiar. In addition, thematic influence fortifies the genre-based lexical focus on academic and technical purposes.

An example mentioned in the Results is the noun *Semiotics.* It surfaced in textbook chapters about Content Analysis (heading C1). Academic lectures on this subject offer language greatly influenced by theme. A course in our institution integrates these lessons on Semiotics in Audio-visual Communication and Librarianship/Information Management studies. The lectures encourage the elaboration of summaries and reports, for which familiarisation with typical collocations and structures in the setting becomes beneficial.

Learners apply their note-taking skills to listening and writing activities derived from the lectures. Figure 8 reproduces a short extract of a lecture on semiotic elements, given by an American visiting professor at our school in 1997. Content comprehension is then tested in activities (Figure 9).

> 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, ….

Figure 8. Lecture excerpt on semiotic elements

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 9. Example of an activity with specific subject lecture

The textbook and lecture genres are thus exploited in this course of second year Audio-visual Communication and Librarianship students. Key lexical items are pinpointed as traits of genre-based thematic language. Learners have the option to experience this language by both textbook reading and lecture note-taking. Some examples are *visual elements, Semiotics components, signs and codes,* and *the basic element* (see Figure 9).

We must clarify that these items are not restricted to one given genre. In other words, not only general elements but also specific items may appear in other genres. However, the words are more descriptive of the context being dealt with. For instance, the data explored in the course (Figures 8 and 9) reflects the typical language of the Semiotics subject, expounded through lectures and textbooks. The concepts needed in that content lead to seeking these specific items, belonging to the mentioned topic of Semiotics and to no other.

The integration of genre and subject fosters content-based instruction, an important point in ESP learning. The approach focuses on corpus material, developed with different educational levels in mind. An example is that of learners in second year courses of Librarianship and Audio-visual Communication having to cope with the mentioned genres of textbooks and lectures.

The assessment of our data is proposed as a practical view of ESP from an academic and subject scope. In this sense, it is not an exhaustive view of word behaviour, but an applied one for subject area courses. In the following section, we revise this and other relevant claims made.

CONCLUSIONS

The principal aim in this paper has been to provide evidence that supports the distinction of common core language from restricted lexical behaviour. A central assumption is that two separate levels exist in our sources: academic and technical. Nevertheless, inferred from lexical classification in our specific corpus, both planes are divided into further categories of word use. These encompass regions of lexical use where academic and technical elements apparently coalesce (however, just in appearance, as has been observed, since specialised use is finely specified).

In a corpus that is representative of both academic and technical material in our selected areas, seeking lexical behaviour patterns is primarily done according to contextual parameters. This is achieved by applying genre and subject variables. The aid of study programs and university curricula becomes

essential in this respect, while applying ESP principles is required for consistency. The chief purpose is to collect texts that meet language and content demands in our setting.

Our approach to the data includes empirical observation, classification, and assessment of lexical patterns. In this process, measurement is carried out quantitatively, that is, in the form of absolute and relative word frequencies. These are essential reference statistics used for contrastive analysis: They serve as point of departure in the contextual study. Keywords then play a decisive role for lexical profiles, which demand a qualitative treatment of the data. This means classification of patterns based on frequency and dispersion.

As a result, in the analysis data is assessed as either occurring broadly across texts, or more narrowly within certain sources. The results propose three main types of lexical behaviour based on this: Common core elements in the whole collection, specific words in themes and topics, and elements that are characteristic of only one genre. The three are surveyed through analytical steps related to ESP notions: Settings are defined and described according to specific learning needs.

In the evaluation of lexical information, academic and technical word behaviour is discussed. Eight categories are induced by investigating the relationship between concordance data and context. The way in which these language peculiarities are developed affects our approach to ESP courses.

Common core elements are divided into general academic items and procedural words. Both demonstrate a widespread distribution throughout the corpus, and are subject-independent. This makes us consider them as semi-technical vocabulary. They include content and grammar items that have either a high or low frequency in the corpus. Their function is inferred for EGAP (English for General Academic Purposes) teaching, mainly through the application of academic tasks, for example, using wordlists to point out lexical data in readings. EAP (English for Academic Purposes) thus motivates our work with EST (English for Science and Technology).

Procedural items are common core constructions that mark cohesion in academic discourse. This is a main characteristic in general academic writing as well as in lectures. Their organisation in discourse facilitates comprehension. In contrast with general academic collocations, procedural elements include grammar combinations that have a semantic prosody related to the organisation of discourse.

Regarding subject-based formations, the degree of restriction in the collocations influences the lexical divisions made. In the case of technical vocabulary, combinations are quite fixed. The elements are highly restricted in their behaviour, meaning that they exhibit a consolidated use in the thematic setting where they are identified as key. Through detailed revision of concordance lines, technical compounds are examined within longer phrases. This description is done in a manner resembling specialised dictionary-making, where key constructions function as descriptors for the subject area.

Concordance observation is also useful for underlining thematic influence on those collocations that are not strictly technical. These are valued as significant feedback in the subject area, but denote a less fixed behaviour. This means that they can be replaced by synonymous expressions without making a significant change. However, their use is characteristic in certain subjects, and not in others. In this sense, even though they tend to be easy to understand, they are also considered specific of the subject area.

Discipline-based elements are also distinctive in the subject area. They would be found in the middle ground between general academic expressions and specific language. In this respect, they are treated as common lexical items, identified in different areas, but prevailing in only one. They are conventional within the discipline, referring to aspects that are frequent and widespread. In EGAP development, work with these elements enhances the use of academic language for particular areas.

A different case is the lexical data that refers to concepts exclusive of only one discipline. In that situation, ESAP is favoured: Tasks challenge learners to cope with specific content in their studies.

Knowledge of technical issues is fostered through activities that demand exploitation of conceptual paragraphs, for example, by elaborating oral reports.

Finally, lexical features are analysed in the genre context. In addition, as emphasis is placed on subjects, the genre setting includes thematic items. Both subject and academic elements raise genre awareness in this context. This is especially useful for ESAP writing performance. Genre-based items can present restricted patterns of lexical behaviour, developed within one single genre, or even topic. The elements behave as descriptive items, but the difference is that they may do so in the overall genre, regardless of topic influence, or in a specific subject conveyed through particular genre conventions.

The information obtained and described in this article is therefore assessed for ESP development. However, it is not intended as theory on lexical behaviour in academic and technical contexts. On the contrary, its validity highly depends on practical factors which lead to the design of specific corpora. Large textual collections can serve as reference for the analysis of our data, but do not meet specific learning demands as fittingly as one's own corpus can. In fact, we believe that none but representative material in the teaching environment can really fulfil specific language requirements.

NOTES

1. The sources may either include one major discipline, such as the *Dictionary of Computing* (Collin, 1999), or more than one area, as is the case with the *Dictionary of New Media: Film, Television, Print, Digital, Internet, Multimedia* (1999).
2. The Spanish titles are *"Informática técnica"* and *"Ingeniería Informática," "Sonido e Imagen," "Biblioteconomía y Documentación,"* and *"Comunicación Audio-visual"* (see University of Extremadura Web page at http://www.unex.es/).
3. The university curricula consulted in Spain (in addition to our own institution) are as follows: For Computer Science and Optical and Radio Communications, *Facultad de Informática, Universidad Politécnica (Madrid), Universidad Politécnica de Valencia*, and *Universidad de Vigo (Departamento de Teoría de la Señal y Comunicaciones).* For Librarianship and Information Management, *Facultad de Biblioteconomía y Documentación (Universidad de Granada).* For Audio-visual Communication, *Instituto Universitario del Audio-visual (Universitat Pompeu Fabra, Barcelona)* and *Facultad de Ciencias de la Información (Universidad Complutense de Madrid).*
4. Guidance offered by content instructors is highly valued in the process of textual selection. In addition, as mentioned above, advanced learner's knowledge can produce positive results. Internal (ESP) approaches can thus benefit from these external factors provided by the institution.
5. In fact, the elaboration of a broader corpus that incorporates business texts leads us in such a direction: to integrate material that is generally useful for information technology majors as well as business students, as they cope with common issues and concepts.
6. Stotsky (1983, p. 438) refers to "words that contribute to cohesive ties in academic discourse ... usually the content words generated by authors writing on similar topics." These words are also common core, offering greater difficulty to non-native or overseas students because they are "often abstract and / or complex."

ABOUT THE AUTHOR

Alejandro Curado teaches English for Computer Science and Telecommunications at the Polytechnic School at University of Extremadura (Spain). His doctoral thesis (2000) presents lexical findings according to genre and subject in specific settings. His research aims to integrate both discourse and corpus-based lexical approaches to teaching ESP.

E-mail: acurado@unex.es

REFERENCES

Benson, M., Benson, E., & Ilson, R. (1997). *The BBI dictionary of English word combinations.* Amsterdam: John Benjamins.

Bergenholtz, H., & Tarp, S. (1995). *Manual of specialised lexicography.* Amsterdam: John Benjamins.

Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus linguistics. Investigating language structure and use.* Cambridge, UK: Cambridge University Press.

Brennan, M., & van Naerssen, M. (1989). Language and content in ESP. ELT Journal, 43 (3), 196-205.

Bygate, M. (1987). *Speaking.* Oxford, UK: Oxford University Press.

Callev, H. (2000). The stream of consciousness. *Film-Philosophy, 4*(11). Retrieved August 15, 2001, from the World Wide Web: http://www.film-philosophy.com/vol4-2000/n11callev.

Collin, S. (1997). *Dictionary of information technology.* London: HarperCollins.

Collin, S. (1999). *Dictionary of computing.* London: HarperCollins.

Conrad, S. (1996). Investigating academic texts with corpus-based techniques: An Example From Biology. *Linguistics and Education, 8,* 299-326.

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, UK: Oxford University Press.

Cowie, A. P. (1998). Introduction. In A. P. Cowie (Ed.), *Phraseology: Theory, analysis and applications* (pp. 1-38). Oxford, UK: Clarendon Press.

Coxhead, A. (1998). *An academic word list.* English Language Institute Occasional Publication No 18. New Zealand: Victoria University of Wellington.

Díaz, J. C., & Jones, M. (1999). Computer language. Madrid: UNED.

Dictionary of new media: Film, television, print, digital, Internet, multimedia. (1999). New York: Readfilm.

Dudley-Evans, T., & St. Johns, M. J. (1998). *Developments in ESP: A multidisciplinary approach.* Cambridge, UK: Cambridge University Press.

Edwards, P. (1996). The LSP teacher: To be or not to be? That is the question. *AELFE (Asociación española de lenguas para fines específicos),* 9-25.

Ewer, J. (1983). Teacher training for EST: Problems and methods. *The ESP Journal, 2,* 9-31.

Farrell, P. (1990). *A lexical analysis of the English of electronics and a study of semi-technical vocabulary*. Dublin: Trinity College.

FECT & NECC Conference (1999) Excerpts of Paper "The Do's and Dont's of Technology Planning" Retrieved August 15, 2001 from the World Wide Web: http://fetc.state.fl.us/.

Firth, J. R. (1957). A synopsis of linguistic theory. 1930-1955. In J. R. Firth (Ed.), *Studies in linguistic analysis* (pp. 1-55). Oxford, UK: Basil Blackwell.

Flowerdew, J., & Miller, L. (1997). The teaching of academic listening comprehension and the question of authenticity. *English for Specific Purposes, 16*(1), 27-46.

Halliday, M. A. K. (1966). Lexis as a linguistic level. In C. E. Bazell, J. C. Catford, M. A. K. Halliday, & R. H. Robins (Eds.), *In memory of J. R. Firth* (pp. 148-162). London: Longman.

Hutchinson, T., & Waters, A. (1981). Performance and competence in ESP. *Applied Linguistics, 2*(1), 56-69.

James, G. (1994). *English in computer science. A corpus-based lexical analysis.* Hong Kong: Longman.

Johns, A. M. (1997). *Text, role and context.* Cambridge, UK: Cambridge University Press.

Johns, T. (1993). Data-driven learning: An update. *TELL & CALL, 3,* 23-32.

Jordan, R. R. (1997). *English for academic purposes.* Cambridge, UK: Cambridge University Press.

Lozano Palacios, A. (1999). *Vocabulario para los estudios de Biblio-documentación* [Vocabulary for library science and documentation studies]. Granada: Servicio de publicaciones, Universidad de Granada, Facultad de Biblioteconomía y Documentación.

McCarthy, M. (1990). *Vocabulary.* Oxford, UK: Oxford University Press.

Ooi, V. B. Y. (1998). *Computer corpus lexicography.* Edinburgh: Edinburgh University Press.

Pedersen, J. (1995). The identification and selection of collocations in technical dictionaries. *Lexicographia, 11,* 60-73.

Scott, M. (1996). *WordSmith.* Oxford, UK: Oxford University Press.

Scott, M. (1997). PC analysis of key words and key key words. *System, 25*(1), 1-13.

Scott, M. (2000). Reverberations of an echo. In B. Lewondowska-Tomaszczyk & P. J. Melia (Eds.), *Practical applications in language corpora.* Frankfurt: Peter Lang.

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.

Stubbs, M. (1995). Collocations and semantic profiles: On the cause of the trouble with quantitative studies. Functions of Language, 2, 23-55.

Termite Database (1999). ITU Global Directory Telecommunication Terminology.

Thurstun, J., & Candlin, C. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes, 17*(3), 267-280.

Tribble, C. (1997). Improvising corpora for ELT: Quick and dirty ways of developing corpora for language teaching. In B. Lewandowska-Tomaszczyk & P. J. Melia (Eds.), *Practical applications in language corpora* (pp. 106-118). Lodz, Poland: Lodz University Press.

Tribble, C. (2000). Genres, keywords, teaching: towards a pedagogic account of the language of project proposals. In L. Burnard & T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 75-90). Frankfurt: Peter Lang. Retrieved August 15, 2001 from the World Wide Web: http://ourworld.compuserve.com/homepages/Christopher_Tribble/Genre.htm.

Trimble, L. (1985). *English for science and technology: A discourse approach.* Cambridge, UK: Cambridge University Press.

Varantola, K. (1984). *On noun phrase structures in engineering English.* Turku: Annales Universitatis Turkuensis.