A CORPUS-BASED FOCUS ON ESP TEACHING

by **Alejandro Curado Fuentes**

University of Extremadura

Cáceres, Spain

acurado @ unex.es

Abstract

The conjunction of lexical analysis and information technology has often led to the design of specialized material. In my Computer and Business English courses at tertiary level, this scope has enabled a certain degree of experimentation with corpus-based lexical information. This paper describes the main results derived from one year of teaching ESP with a focus on electronic corpora. The main conclusions point to the observation of positive and negative factors in terms of language acquisition, leading to the planning and design of corpus technology priorities.

Introduction

The need to communicate in specialized contexts or domains, such as academic and scientific disciplines at university, is greatly emphasized at the European level. The emphasis is often placed on an effective linguistic development for research purposes (Bricall report, 2000). In addition, as electronic communication and the digital era expand, it is obvious that new lines of work open up for linguists and foreign language researchers (e.g., study of cyber-genres, sociolinguistic analysis of web sites, etc). An example is Giménez (2000), analyzing professional communication and use of e-mail; another one is Pérez Paredes (2001), who underscores the need to integrate real situations derived from Internet use, or Posteguillo (2002), who distinguishes a double focus in the study of network discourse: that of computational linguistics and sociolinguistics. The new venues and scopes suggested for ESP (English for Specific Purposes) point to the importance of conveniently valuing and assessing the development of specialized languages in new and relevant areas (e.g., working with technology).

This paper focuses on academic language for university disciplines like Computer Science and Business Administration as a common area, not as separate fields for lexicographical study (e.g., Collin et al, 2004). Such a holistic approach takes the current key language of subjects shared by the two disciplines and derived from digital resources on economic, financial, technological, socio-technical, and informational topics in Business and Information Technology (BIT). The aim in this sense is academic and research-based; the material collected and designed varies from textbooks to most specialized sources in the form of digital journal articles and research projects.

By handling and contrasting different university syllabi and curricula, a significant conceptual nexus can unfold. In addition, similar topics and interests are revealed in the management of Internet-based databases for such seemingly divergent fields as Business and Computer Science. All this processing of academic goals and contents in the study programs is especially attractive for the analysis of Academic and Technical English at university. Thus, the main objective is to identify inter-disciplinary grounds for the exploitation of common lexical cores. In the process of searching, as mentioned, the lessons and lines of work established in the different subject syllabi at university are followed, determining the inclusion of the various texts.

Similar to Coxhead (1998), the chief lexical scope integrates constructions found across different academic and scientific texts, following previous classifications made of semi-technical and technical words (Curado, 2001). However, in this case, the purpose is not to build different word sets, but quite the opposite: to form a glossary of useful specific expressions for their detailed application to ESP courses. In fact, lexical variation is not accounted for, and, in contrast, the focus is made on linguistic blocks made up of frequent and dispersed expressions found evenly across a corpus. In this regard, special attention is paid to corpus-based lexical frequency, dispersion, concordance, collocational strength, and lexical behavior (Ooi, 1998). Such factors are essentially instruments for the evaluation of word formations in relation to electronic text typology and corpus sources.

This study includes a lexicographical approach by providing the undergraduate university students with a glossary of key academic constructions that should motivate their decoding and encoding skills. The common objective is to offer a framework for activity and task exploitation often dealing with corpus technology (e.g., identifying key repetitions, formulating semantic prosody, finding best equivalents in Spanish, etc). In the following section, a brief description of the corpus is included, from which the most common 500 words are retrieved for the courses. Then, the integration of the glossary framework in the ESP curriculum is described and evaluated within the teaching context. Finally, conclusions based on the results obtained during one year of classroom observation are reviewed.

The corpus-based glossary

The factor of `representativeness´ (Biber et al., 1998, p. 246) leads to the design of a glossary based on characteristic data of the domain / area to be represented, Business and Information Technology (BIT). The electronic sources to be selected and edited must adapt to the objectives of the teaching situation. In Business English, for instance, the low-intermediate level of English that students tend to search for clear equivalents in English and Spanish as well as clarity of concepts (e.g., management software for analyzing sales data, on-line tax software for businesses, etc). In Computer English, in turn, the tendency to have both a higher language level

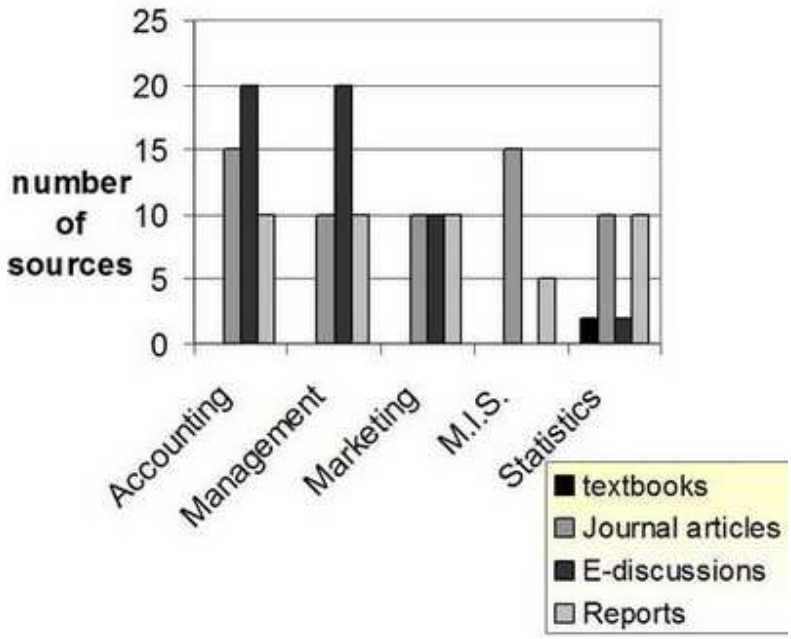
and greater technical knowledge encourages the inclusion of more complex text types and topics.

Actually, most glossary resources dealing with technical English (i.e., Computer and Business) available either freely (on the web) or commercially, do not seem to suit these learners' heterogeneous characterization (i.e., in terms of language level and subject knowledge). For instance, in relation to linguistic concerns, core lexical items consisting of content and grammatical words (e.g., *information available on* + *electronic medium*) tend to be excluded from the resources, whereas, regarding subject matter, the glossaries examined do not give enough lexical information (e.g., http://www.globalbusinessresources.net/spanish.shtml), or are too specific (e.g., http://money.cnn.com/services/glossary/a.html); most only deal with one discipline (e.g., Collin, Piqué-Angordans, Posteguillo & Melcion, 2004). Thus, the aforementioned need of 'representativeness' leads to the development of a specific resource for the target setting.

In order to attempt to cater for such different needs in the learning context, a balance of textual material should be intended in the corpus. This balance implies the follow-up of objective criteria for the selection of sources in the corpus. In terms of language command, common core lexical knowledge is accounted for. In terms of subject matter, both business and information technology issues are explored. Thus, a concrete number of academic and technical lexical items should be balanced (i.e., a lexical bulk that is neither too large nor stays at a basic level of linguistic / conceptual knowledge).

In addition, as Hunston (2002, p. 16) observes, the selection of sources should reflect the communicative exchanges that take place in the target context of research and work. In the same corpus, we may have from formal writing (e.g., technical reports and instruction manuals) to informal / conversational material on the web (e.g., Internet forum messages). Learning practices and perspectives can benefit from this hybrid nature of the collection (Conrad, 1996, p. 302) and academic text heterogeneity (Swales, 2003, p. 4). Figure 1 displays the contents of the corpus, including different types of readings as well as subjects shared by Business and Computer Science students from the first to the fourth year of study.

Figure 1: Corpus



The types of texts in the corpus correspond to three different levels of complexity for the learning situation. The textbook is placed in the first category of introductory and informative types, in agreement with Johns (1997, p. 46). At a middle plane, we have reports and e-discussions (i.e., electronic sets of messages in academic forums), also categorized as descriptive types by Henry (2000). At the third (higher) stadium, we identify journal articles discussing research results and thus presenting a more specialized academic discourse type (Conrad, 1996). In relation to subject matter, as can be checked in Figure 1, five main subjects are explored, seen by students along their respective majors (albeit in different years – e.g., Computer Science people learning Statistics in third year, and Business students in first and second years). Contrary to being an inconvenience, such slight variations can contribute to the inclusion of different genres and text types for each separate subject.

To select the various sources, advises from co-workers and graduate students can orientate our search. Some items, such as Díaz Martín, Veloso & Rodríguez García (1999), written by colleagues in Computer Science, were pertinent, accessible via the professors' websites (at University of Extremadura, the Moodle platform http://campusvirtual.unex.es). Also, electronic databases allows for the identification and classification of the types of texts sought. One example is the Kluwer engine for academic material related to BIT (www.kluweronline.com) and the site `Global Edge´ for electronic files and digital forums (www.globaledge.com / http://globaledge.msu.edu/index.asp), where the corpus designer may examine documents by topic and text type.

These corpus sources in Figure 1 were selected and stored in the year 2000. As a result, many are no longer available on the Web or electronically (only as electronic corpus sources in University of Extremadura Moodle platform mentioned above). In these sources, the textual and visual elements that do not contribute significant lexical information were discarded. For instance, in the e-discussions, retrieved from an academic forum on the web site 'Global Edge', such items as addressers, addresses, and proper names (e.g., university, cities, etc) were omitted. Thus, only plain text is compiled, at times with just a few lines per message, as in this e-discussion example for the topic of Accounting:

> I'm trying to figure out how you compute the distance to the origin, and the contributions of the rows and the cols in the correspondence analysis. Anyone knows how you exactly do that ? I'm trying to reproduce the results of a page I found at this address: ( ) Any help would be greatly appreciated.

Other e-discussion messages were as long as 700-800 words. A key factor in the corpus distribution and design was to keep in mind the need for balance during text retrieval and the compilation of the corpus. This meant that, in the overall corpus design, a similar number of words (ranging between 30,000 and 45,000 words) should be met for each subject category displayed in Figure 1 (e.g., from 32,265 words in M.I.S. to 48,340 in Statistics). This distribution is done so that no particular group of texts in one subject, when fed into the wordlist tool of the concordance software (*WordSmith Tools*, Scott, 2000), may yield a much lower or higher amount of words than the rest. The DCL (Detailed Consistency List) that takes the five subject-driven wordlists and contrasts items in them should, in fact, contain similar numerical data for lexical analysis.

The lexical analysis is performed by following the DCL for the whole BIT corpus. As a result, the main goal is to display the top items that most frequently co-occur across the subject categories. The lexical core would represent the highest degree of general or common academic items shared by the two disciplines.

The academic elements, regarded as semi-technical in many cases (e.g., Thurstun and Candlin, 1998; Nation, 2001) tend to contain a significant semantic basis. The total amount of such items should be around 500 node words – a basic-to-intermediate lexical knowledge to be expected for our learning context. To obtain this estimate, the total number of words or tokens in the corpus (652,034) is divided into the total number of types (distinct words in the corpus – 21,963). The result is then multiplied by the standardized token-to-type ratio established by Scott (2000): the number of distinct types per 1000 tokens (i.e., 37.12 in the corpus). The resulting figure is then divided in half due to the need to adapt to a *medium* level of learning (i.e., intermediate; should it have been lower, the amount would have been divided into three, whereas for a high or high/intermediate level, the core academic vocabulary should be the whole figure, about 1000 words according to this computation). The result (551) is rounded up as the

amount mentioned above. Some scholars (e.g., Nation, 2001; Flowerdew, 2001) also describe concrete lexical amounts according to such different language commands.

The 500 words selected are listed in the detailed consistency list, which specifies word repetition according to each of the five subjects of the corpus. By examining high frequency and dispersion of the items on the list, common use of such elements in the corpus is made a primary factor. An example is the use of the verb **lead** as a common core word (in agreement with other academic lists – e.g., Coxhead, 1998), since it appears in all five subjects; however, a synonym like the verb **cause**, also considered common core academic by Coxhead (1998), would not be included in our case, since it does not appear in all five categories made.

The corpus-based 500-word glossary is organized by listing key word combinations and expressions found across various text categories. The arrangement of the collocational items is based on frequency and dispersion by using their t-score, which indicates that the given combination is not due to chance. This statistical score should be above required measurements (see, for instance, Church, Gale, Hanks & Hindle, 1991) to consider the lexical elements as common core. For instance, to take the same example as above, the verb *lead to* combines with the adjectives *better*, *improved*, *greater*, and *different*. In all cases, given their co-occurrence frequencies, the resulting t-scores are higher than 2 (minimum required). As a result, in the entry for the verb *lead*, the glossary should contain different examples of constructions found with those adjectives (e.g., *lead to a better environment*, *lead to a greater understanding*, etc). In contrast, a combination like *lead to + higher* appears with a t-score below the minimum value. Thus, the phrase *can lead to a higher turnover* would be discarded for the glossary. The same thing applies to collocates preceding the node word: for instance, *expected to lead to* is included (t-score above 2.0), while other forms like *necessarily lead to* and *automatically lead to* are excluded (t-scores below 2).

ESP teaching and glossary development

The glossary and activities based on it were made available on the web (at our university links, e.g., http://epcc.unex.es, www.unex.es/lengingles/ALejandto.htm, and the Moodle platform mentioned above). The undergraduate university students could thus access the alphabetical and frequency lists of academic words with which to answer various on-line questions on word order and formation. For example, given a set of words such as *accounting*, *account*, *accounted*, *accounts*, they have to find the most frequent item and determine its collocations (with verbs, nouns, adjectives, and so forth); then, to translate such constructions into Spanish. The aim of these weekly activities is to familiarize students with the glossary, but not to build lexical knowledge. In fact, according to most students' answers in the post-task questionnaire (see Appendix) given out at the end of the course, the wordlist-based activities are fairly easy but do not contribute to their assimilation of lexical items (questions 4 and 5).

Instead, what seems to qualify as worthwhile lexical development has been a direct approach to the glossary, even though students are divided concerning the type of lexical material explored and how it was organized (see questions 7 and 8 in the post-task questionnaire). They seem to particularly favour the use of lexical skimming and scanning in order to figure out expressions and write them in different exercises, especially translation (see questions 6, 9 and 10).

The most common type of activities based on the corpus tends to involve the use of academic collocates (see Appendix for the test given to learners, where matching, fill-in-the-gap, cloze, translation, and writing exercises are included). Learners tend to value as highly positive to be able to decode academic collocations from Spanish into English (question 12). In addition, related to this is the notion of having to pair up lexical items derived from common corpus-driven expressions (e.g. *exchange* + *information with*). Collocations were examined extensively in the glossary. For instance, matching exercises were developed as a need to 'visualize' the company that node words in the glossary keep. Both commonly used and restrained combinations were explored in this way. In the Appendix (in the test activity A), see, for instance, a more general expression like *may increasingly be* in contrast with *do a significance test*, more typical within Statistics texts. In this test, nonetheless, some types of activities involving collocation work were evaluated as more convenient than others, according to the students. For instance, the activity of matching collocates is seen as less productive for lexical development.

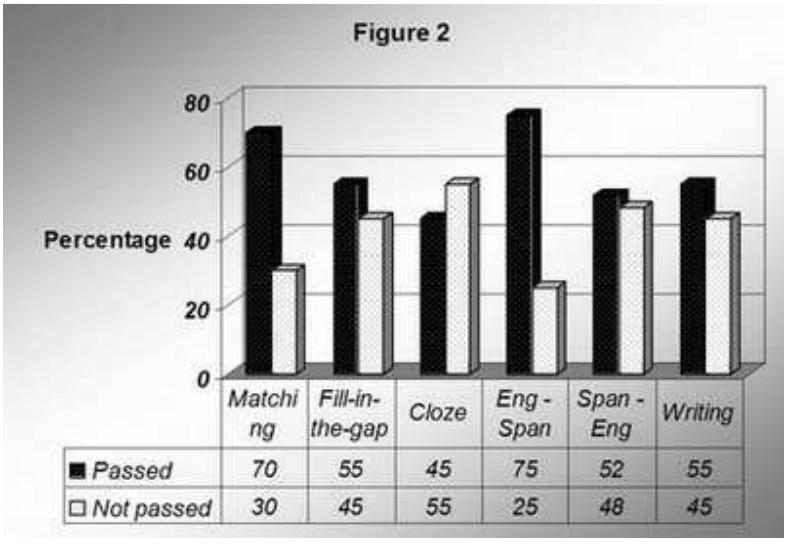
In turn, Fill-in-the-gap exercises (e.g., activity B in test) derive from corpus analysis focusing on lexical development within co-texts or concordance lines. The basis for this type of activity is given by key concordance lines where explicit use of collocations is shown. This type of exercise is valued more highly by students in the questionnaires. Its development corresponds to the need for actual co-texts where node words are found. The selection of these concordance lines is made in agreement with semantic units unfolded for each expression; for instance, for the item *run on*, the key structure in this corpus is *run on* + computerized device. Students seem to favour this kind of help in the exercise, as it apparently improves their recognition of the expressions.

Thirdly, cloze exercises (exercise C in test) also originate as a consequence of corpus-based approaches. In this case, textual chunks containing words that typically co-occur with a given node may be easily spotted and provided to students (e.g., by conducting concordance searches of words in context). This type of lexical gapping is similar to Coxhead's AWL exercises on the web (http://gpoulard.tripod.com/index.htm), where core items from the lists can be automatically removed so that learners can work with them in the paragraphs. Nonetheless, the Cloze activity with whole paragraphs is not evaluated as important by students – see question 15 in Appendix – even though more context for the lexical items is provided.

The other three activities given in the test are two translations of paragraphs (direct and reversed translations) and a short written composition. According to the questionnaires, in terms of communicative skills, students judge that it was their writing which benefited most (question 16 in the Appendix). Then, according to their answers given for question 17, their favourite topics to write about are socio-technical (e.g., giving opinions about advantages and disadvantages on the use of technology in society). Regarding the least improved skill, most answers refer to listening (questions 18) because in many cases (question 19), students claim that there should have been more audios and videos in class or on the Web (see activities on the webpages referred above). Finally, the majority of the students (more than 50 percent) favour the integration of the glossary in the ESP class (question 20), reasoning in some cases that it provides clear evidence about the important language that they should know for their academic work.

Given the percentages of the 40 surveyed students (see Appendix), in order to either corroborate or contradict the stated ideas and opinions, the post-test described above was evaluated by the teacher. This evaluation consisted of the six types of exercises named, focusing on glossary-based lexical items; each activity was evaluated from 1 to 10 (a 5 as a passing grade). In the test, the activity done best by students is the English-to-Spanish translation, as can be examined in Figure 2, followed by the Matching exercise. The one done with the most mistakes is the Cloze exercise.

Figure 2



	Matching	Fill-in-the-gap	Cloze	Eng - Span	Span - Eng	Writing
■ Passed	70	55	45	75	52	55
□ Not passed	30	45	55	25	48	45

Some results in Figure 2 can be contrasted with ideas and impressions received from the questionnaires. For example, according to question 13, the matching activities are considered less interesting, and yet, this type of work was performed well in the post-test. In turn, the Fill-in-the-gap task is regarded as relevant and useful, but it was done more poorly in the test. In the translations, even though learners tend to view Spanish-to-English translation as more relevant and productive, it is the English-to-Spanish translating task that they do better.

Conclusions

Based on the observation of class work, questionnaire answers, and post-test results, some conclusions regarding the nature of corpus-based applications in ESP may be considered. By following the surveyed percentages (see Appendix), some inferences can be drawn in relation to the type of learning context developed.

As answers to questions 1 and 2 demonstrate, most students have recently had English classes and, in most cases, ESP courses at university. For these learners, knowledge of vocabulary was not the highest concern, as responses in question 3 show. In fact, the majority felt less confident about speaking / pronunciation skills. As a result, at this initial stage, we would feel that this class has an average / intermediate level of English and, as typically occurs with Spanish EFL learners, they demand more oral practice. Then, given their impressions to the material assigned and exploited in this class, and by assessing their scores achieved in the test, a different picture is obtained to some degree. The use of the glossary for lexical activities offered little difficulty according to learners, and writing as well as translation skills seemed to benefit most from this material. However, the results in the test point to too many mistakes (below passing) for the types of activities actually claimed by students as most profitable (Spanish-to-English translating, Fill-in-the-gap, and writing on a known socio-technical topic).

Such apparent contradictions may lead to some assertions about ESP development: students appreciate the focus on actual lexical items that are significant and common core in their studies; for instance, they realise that translating academic lexical constructions fosters their decoding skills to clarify semantic aspects in the expressions. Still, learners do not apply this knowledge to encoding activities as well as should be expected in the test on their linguistic intake. In this respect, the answers given to question 11 in the Appendix provide a more realistic outcome of their course work. Most students perceive that they have improved their mental lexical database in part or a little, and few of them answer much or hardly.

As a result of the above, it would seem that there is a need to redirect the corpus-based approach to a learning process that may enhance the learners' progress regarding word behaviour assimilation. A glossary of academic items such as the BIT common core resource should probably have to enable greater active participation; for instance, a more active elaboration of the entries by having students work with the corpus in electronic form, as some authors have recently pointed out for the EFL and ESP learning contexts (e.g., Connor and Upton, 2005; Gavioli, 2005). In fact, as the answers given by students to question 20 may suggest, the exploitation of common core glossaries in ESP such as the BIT resource can be extended to other (more advanced) courses. Some fair reasons include the importance that knowing such academic / common core language implies for students' future jobs (which will involve much handling of information (in English) related to business technology).

In addition, oral work is perceived as a highly demanded task by our L2 learners. The course integrates the obligation to perform various oral discussions and a presentation about a topic chosen from the class syllabus. Most presentations usually demonstrate a fairly suitable use of academic vocabulary and expressions, many of them collected in the glossary (e.g., *information exchange*, *available on the Web*, *management tools*, *face + competition*, etc). Nonetheless, a significant number of the students surveyed (45 per cent) felt quite nervous and intimidated by having to speak in public about such matter, and 30 per cent of them skipped the presentations alleging their lack of competence for the task. In this sense, the corpus-based approach should also encourage oral work (e.g., by fostering confidence through the investigation of word relations to confirm and / or contrast knowledge).

The new scope offered in this ESP course has included the significance of suitably assessing the development of a lexical focus. The academic / research language exploited is perceived as crucial by students, but a relevant percentage of these fail to reflect its production and assimilation along the work done in class. Therefore, the goal of providing students with effective communicative means in the form of key academic language partially succeeds. The corpus-based glossary seems to motivate learners' decoding skills. Activities such as identifying key repetitions, formulating semantic prosody, finding best equivalents in Spanish, and so forth, are generally regarded as positive by students, but do not lead to optimal results in the tests.

The integration of the glossary framework in the ESP curriculum may thus be evaluated as having an acceptable reach in the learning process, but should be further explored for future courses. A possibility may involve the availability of a wider range of on-line tools and applications that enable students to become actual decision-making agents in the design of the corpus-based language. In other words, the glossary may be explored not so much as a product available to learners, but as a means. Its construction may be posed as a challenge for students so that they may work with lexical information to build their own lexical entries. Future research may then query whether a more active involvement on the part of students (e.g., in the building of the BIT glossary entries) can produce better results. This line of work may parallel similar recent studies, such as Paquot (2005), focusing on productively oriented academic word lists. The subjects included in such an experiment would thus become an experimental group whose performance can be contrasted with the 40 students tested in this paper (a control group). Such a research focus may further investigate the influence of the corpus approach on our ESP courses at an intermediate stage of L2 learning.

References

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

Bricall report. (2000, March 20). Informe `Universidad 2000´. *El País*, 14.

Church, K.W., Gale, W., Hanks, P.W., & Hindle, D. (1991). Using statistics in lexical analysis. In U. Zernik (Ed.), *Lexical Acquisition: Using On-line Resources to Build a Lexicon*, 1-27. Hillsdale: Lawrence Erlbaum.

Collin, S. , Piqué-Angordans, J. , Posteguillo, S. & Melcion L. (2004). *Diccionario Bilingüe de Informática*. London: Bloomsbury.

Connor, U., & Upton, T.A. (Eds.). (2005). *Applied Corpus Linguistics. A Multidimensional Perspective*. Amsterdam: Rodopi.

Conrad, S. (1996). Investigating academic texts with corpus-based techniques: An example from biology. *Linguistics and Education 8* (3), 299-326.

Cowie, A.P. (1998). *Phraseology. Theory, Analysis and Applications*. Oxford: Clarendon Press.

Coxhead, A. (1998). *An Academic Word List*. Victoria: University of Wellington.

Curado Fuentes, A. (2001). Lexical behaviour in academic and technical corpora: Implications for ESP development. *Language Learning and Technology 5* (3), 106-129.

Díaz Martín, J.C., Veloso, I., & Rodríguez García, J.M. (1999). Building TLC-TK GUIs for HRT-HOOD systems. In P. García & J. Díaz (Eds.), *Distributed Systems. Topics*, 16-35. Cáceres: Universidad de Extremadura.

Flowerdew, J. (2001). Concordancing as a tool in course design. In M. Ghadessy, A. Henry & R.L. Roseberry (Eds.), *Small Corpus Studies and ELT. Studies in Corpus Linguistics*, 71-92. Amsterdam: John Benjamins.

Gavioli, L. (2005). *Exploring Corpora for ESP Learning*. Amsterdam: John Benjamins.

Giménez, J.C. (2000). Business e-mail communication: Some emerging tendencies in register. *English for Specific Purposes 19* (2), 237-251.

Henry, J. (2000). *Writing Workplace Cultures: An Archeology of Professional Writing*. Chicago: Southern Illinois University.

Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.

Johns, A.M. (1997). *Text, Role and Context*. Cambridge: Cambridge University Press.

Nation, P. (2001). Using small corpora to investigate learner needs: Two vocabulary research tools. In M. Ghadessy, A. Henry & R.L. Roseberry (Eds.), *Small Corpus Studies and ELT. Studies in Corpus Linguistics*, 31-46. Amsterdam: John Benjamins.

Ooi, V.B.Y. (1998). *Computer Corpus Lexicography*. Edinburgh: Edinburgh University Press.

Paquot, M. (2005). Towards a productively oriented academic word list. Retrieved December 12, 2005, from http://palc.ia.uni.lodz.pl/abstract/index.php.

Pérez Paredes, P. (2001). From rooms to environments: Techno-short-sightedness and language laboratories. *International Journal of English Studies 2* (1), 59-80.

Posteguillo, S. (2002). Netlinguistics and English for Internet purposes. *Ibérica 4* (1), 21-38.

Scott, M. (2000). *WordSmith Tools 3.0*. Oxford: Oxford University Press.

Swales, J.M. (2003). Corpus linguistics and spoken English for academic purposes. In E. Arnó & A. Soler (Eds.), *VI Conference on LSP and the Role of Information Technology. Abstracts*, 5-6. Barcelona: Universitat Politécnica.

Thurstun, J., & Candlin, C. (1998). Concordancing and the teaching of the vocabulary of academic English. *English for Specific Purposes 17* (2), 267-280.

Click for Appendix