Alejandro Curado Fuentes

Lexical Acquisition in ESP Via Corpus Tools: Two Case Studies

Abstract

This paper aims to illustrate the application of a corpus-based approach in teaching ESP (English for Specific Purposes) by means of two case studies: Group 1 in 2004/2005 and Group 2 the following year. The first group of university students approached lexical material by exploiting micro-skills such as identifying key repetitions, formulating semantic prosody, finding best equivalents in Spanish, etc. The second group used the same academic corpus via electronic resources. According to the overall results, the second group of students outperformed the first one in their use of lexical input and the main findings point to the observation of their improved answers regarding the production of lexis. The tests and questionnaires served as chief control instruments at the end of the courses.

Keywords: academic corpus, glossary, concordance, collocation, lexical acquisition.

1. Introduction

Communicating in university contexts usually implies doing so within one single discipline or domain. There are numerous cases, especially when it comes to presenting one's own research, in which academic communication involves specialized knowledge. In addition, forms of e-communication or digitized media are establishing a type of interactive rapport that is obviously expanding and gaining support over the Internet (e.g., e-forums, discussion groups, virtual platforms etc). In ESP (English for Specific Purposes), the process of learning and using a specialized language is often closely related to the use of information technology (IT) applications (e.g., Rowley-Jolivet & Carter-Thomas, 2005).

Corpus tools or corpus-based approaches are part of this growing amalgamation of technology and language learning for specific purposes (e.g., Gavioli, 2005). Students and faculty actually become active observers of language in use through the development of specific communicative (verbal and non-verbal) skills that are pinpointed, analyzed, and exploited by corpus techniques, as explained in Bowker & Pearson (2002), and Connor & Upton (2004), among others.

In this paper, two case studies are followed. First, a corpus-based academic glossary was used by a group of undergraduates during the 2004/2005 academic year in order to exploit decoding/encoding skills, such as the identification of lexical items based on their significance in the texts, the detection of semantic prosody, findings of best equivalents in Spanish, and so on. Second, a contrastive view with the first group is afforded by the results from a second group of students' use of that academic corpus in an electronic form. In this case, the ESP students relied on electronic concordancers to find key items, repetitions, collocates, etc (tasks carried out during the 2005/2006 academic year). As the overall findings show, the second group of students performed better than the first in terms of the key lexical intake. Test and questionnaire answers corroborate that an important gain was made in the production of lexical items. According to this study, e-corpus management tools may thus have contributed to the beneficial learning and intake.

2. Methodological framework

Corpus development in the classroom may evolve from two chief lines: the bottom-up approach followed by Johns (1991) and Thurstun & Candlin (1998), among others, who demonstrate the usefulness of corpus-driven bottom-up information from concordances. Parallel to this line, and even converging with it at times, is the use of top-down information via corpus applications (e.g., Flowerdew, 2004; Aston, 2001; Hoey, 2005). These authors use "full texts" (Flowerdew, 2004: 15) in specific corpora for full concept terminology clarification and textual collocates and, as Aston (2001:74) notes, knowledge of contextual information on genre and topic facilitates the "use of concordances" and "their interpretation".

In this paper, the methodology is mainly based on the bottom-up approach, yet it also partly derives from the top-down approach. In the first case, lexico-grammatical content items (e.g., nouns, verbs, adjectives, adverbs) interact and behave specifically according to established syntagmatic patterns while with the second approach we look into lexical relationships at the top levels of text and discourse; in agreement with Flowerdew (2000), semantic references are established with text types, genres and subject area/topics. For instance, a collocation with the verb *achieve* (*achieve + success*)

is observed as being primed for use in the type of discourse found in Marketing, particularly in sales/technical reports, where a product sale may either succeed or plummet commercially. In our analysis, it is noted that the lexical item *achieve + success* is textually cohesive for Marketing (i.e., it is common enough to be considered a key lexical construction or item in these texts).

3. The corpus

As Hunston (2002: 16) states, the selection of corpus sources should reflect the communicative exchanges that take place in the target context (i.e., academics at university). This design of the ESP corpus aims at the compilation of both technical/academic writing (e.g., technical reports and journal articles) and 'conversational' material (e.g., Internet forum messages). The hybrid aspect is in fact convenient for special purpose corpora, as Conrad (1996: 302) explains. Academic language is thus closely related to the notion of such words and structures in specific settings. As Thurstun (1996: 3) proposes, academic items are common to different fields of academic learning, associated with specific disciplines from which pivotal words may be compared for corpus study. In this sense, ESP and academic language appear as interrelated notions.

Figure 1 displays the contents of the corpus, including different types of readings as well as subjects in the targeted setting. The first four years of two university degrees, Business Administration and Computer Science, taken by our ESP students are thus represented by the sources selected.

Five subject categories are included in the corpus, corresponding to subjects explored by students during their studies (albeit in different years – e.g., Computer Science people take Statistics in the third year, while Business students do so in their first and second years of study). In this respect, variation can lead to the strengthening of the academic register, as Conrad (1996: 301) states. For instance, in Statistics, a heavier load of textbook reading is involved, whereas in Management Information Systems (MIS) more research articles are used (and, in turn, less e-discussion and textbook material is included).

Sources were mostly retrieved from existing electronic databases where pertinent academic material is available. For instance, related to Business and Information Technology (BIT), many works are provided via the Kluwer search engine (www.kluweronline.com, accessible upon payment of corresponding fees). Also, the free site, "Global Edge", was explored for the identification of files and forums dealing with business topics (www.globaledge.msu.edu).

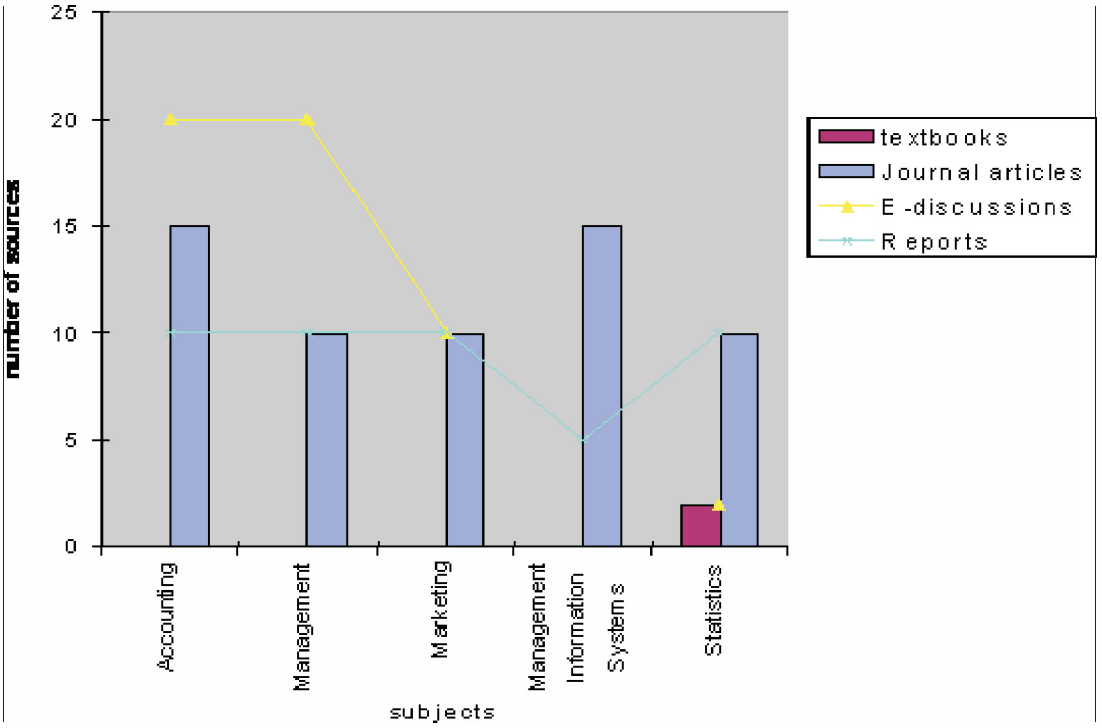


Figure 1: Corpus selection and distribution of sources

The corpus sources are sorted and stored by splitting or adding them in order to produce text files with similar numbers of words (around 2,000). The WordSmith Tools programme (Scott, 1999) allows for this type of file segmentation, aiming to ensure uniformity in the measurement of lexical frequency and dispersion across the corpus. Then, a detailed consistency list (DCL) is generated for the whole corpus by selecting subject categories as word lists (five, as seen in Figure 1). The display of the frequent words distributed across the files, is enabled in the DCL. The top common words, appearing in all five subject categories, are regarded as semi-technical (in agreement with discussions presented in Thurstun & Candlin [1998] or Nation [2001]). These items constitute the core content lexis across the areas, amounting to 1,170 words (after removing grammatical words, e.g., articles, prepositions and so on).

500 content words from these core items in the DCL are taken as basic for the learning context. In statistical terms, this total derives from the proportion between the tokens (overall number of words in the corpus, i.e., 652,034) and types (distinct words in the corpus – 21,963). The result is the STTR (standardized ratio) established by Scott (1999): an average of 37.12 words. A similar result may also be 29.68 words (derived from dividing the tokens into types). In any case, the statistical information tells us that an approximate figure of 30 different new words might be managed by students as the lexical input for each new (2,000 word) text. Since there are five subject categories and four genres/text types in the corpus, the average score (30) is multiplied by 20 (5 subjects times 4 genres/text types). Consequently, 600 may be the total number of words to be managed. However, students tend to have an intermediate level of English

(i.e., according to the "Intermediate Mid" level proposed by the American Council on the Teaching of Foreign Language (Breiner-Sanders et al., 2000: 16), they are capable of producing sentences in short paragraphs, but not wholly connected discourse, for specific functions). As a result, I have determined that some fewer words may be appropriate, and the last 100 items should be removed. The first 500 content words are definitely selected as being core in the aforementioned DCL.

4. Two case studies in corpus-based ESP development

As described, the 500 entries in the corpus-based glossary are organized under the top DCL lemmas. The corpus analysis also considers t-scores for given word combinations, thereby aiming to confirm that lexical co-occurrences are not due to chance. For a detailed description of the analysis of the lemmas in the glossary, see Curado Fuentes (2006).

The lexical resource is made available on-line via a Moodle platform at the University of Extremadura (http://campusvirtual.unex.es). The resource offers the lemmas as headings for the entries and, within such entries, text and hypertext information are available. Figure 2 displays an example of an entry in the glossary. As can be seen, the words within the entry that are also lemmas in the glossary (e.g., the noun information) appear highlighted (i.e., they offer a direct link to their entry). In addition, some sound media is provided in both the headings and bodies of articles. The different word combinations are often (but not always) translated into Spanish. Such translations were provided in agreement with the students (via a comments utility enabled in the glossary resource).

The integration of the glossary into the ESP courses was made possible by complementing regular classes with this explicit lexical material (which accounted for approximately 30 per cent of the course, i.e., one hour and a half per week). Students had to pay attention to word behaviour and use it in their reading and writing activities, and the glossary then served as guidance. In fact, many activities were carried out online (i.e., in the Moodle platform) and, once activated, the glossary could be accessed at any time by clicking words used in the activities that were already stored in the glossary database.

Figure 2: Screenshot of one glossary entry in the Moodle system

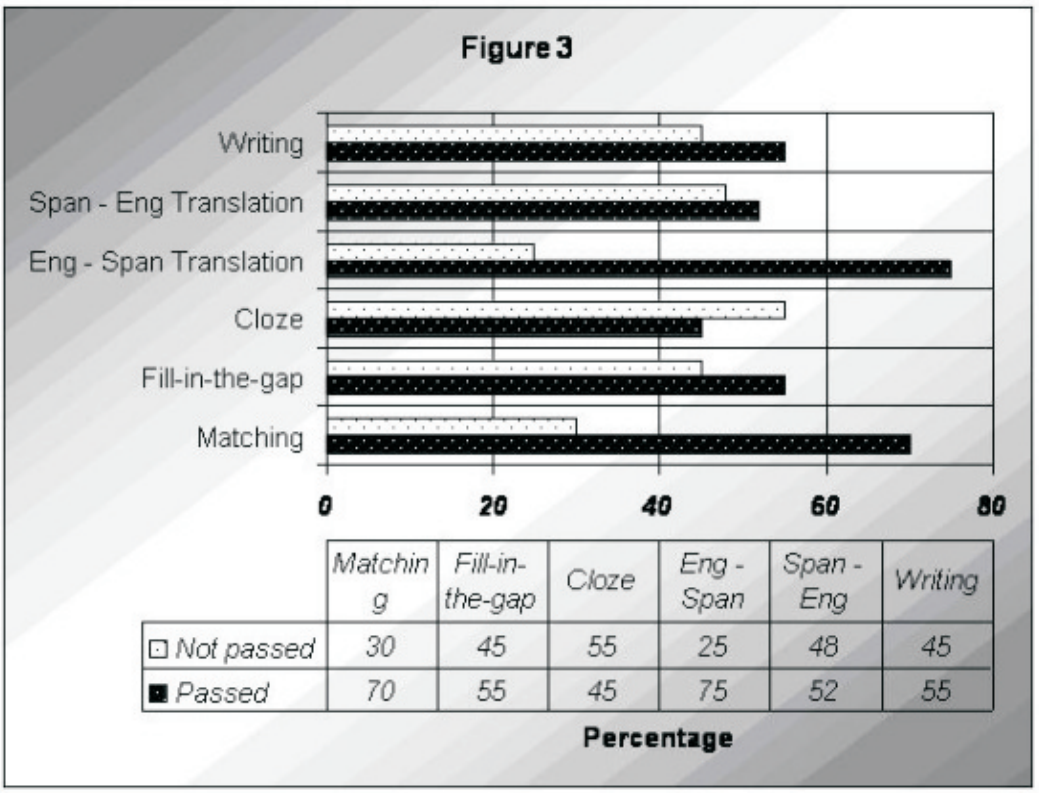
4.1 Case study 1: 2004/2005 group

This group's results are described in detail in Curado Fuentes (2006). As a consequence, I will not document the findings as fully as in the case of the 2005/2006 group below. I shall only briefly explain the overall information obtained from the answers and performance.

As an introductory activity, the students were given word families such as accounting, account, accounted, accounts, derived from the DCL wordlists. They had to use those lists in order to identify the most frequent item and determine collocations (e.g., accounting combines with verbs, nouns, adjectives); then, to translate the identified constructions into Spanish.

In the second stage, the students exploited the glossary entries by working with lexical knowledge (with both previous and acquired content). This direct approach to the glossary seemed to produce more positive feedback as the test results and post-test comments showed.

The test was given to the students at the end of the course. In Curado Fuentes (2006), there is an example of a test where six types of exercises were given (each activity was evaluated from 1 to 10 - 5 considered as a passing grade). The test included a short text where key words are left out (cloze test), a fill-in-the-gap exercise, word matching, two translations (from English to Spanish and vice versa), and a short composition (10-15 sentences) on a familiar socio-technical topic seen in the course (in this case, networked countries vs. developing countries). The activity performed best by students was the English-to-Spanish translation, followed by the matching exercise. The activity carried out worst was the cloze exercise.



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

Figure 3: Passing / non–passing scores by group 1

The results of the tests are in contrast with the post-test questionnaires (http://www.unex.es/lengingles/Alejandro/appendix.pdf). The questionnaire items illustrate how lexical development took place. Overall, the students perceived that the wordlist-based activities are easy but inappropriate for lexical acquisition (questions 4 and 5). In relation to the corpus material from the glossary, the students tended to reveal some doubts and second thoughts (see questions 7 and 8 in the post-task questionnaire). In turn, they liked activities such as lexical skimming/scanning, and inferring from context (questions 6, 9 and 10). They also seemed to prefer academic collocations (question 12), even though they actually scored low for exercises dealing with collocation matching in the tests. Also, they preferred fill-in-the-gap exercises but they did not do too well in these activities.

In the case of writing, the students tended to value this skill in its relationship with lexical activities in class (question 16). The majority of students also favoured the integration of the glossary into the ESP course (question 20), explaining that in some cases it provides clear evidence about words important for academic work. If contrasted with the test results, nonetheless, some students' answers reveal some contradictions. For instance, matching activities were considered less useful but received more passing grades, while the fill-in-the-gap exercise was seen to be interesting yet it was carried out worse. In the translations, most students preferred Spanish-to-English, and yet they tended to perform better with English-to-Spanish translations.

4.2 Case study 2: 2005 / 2006 group

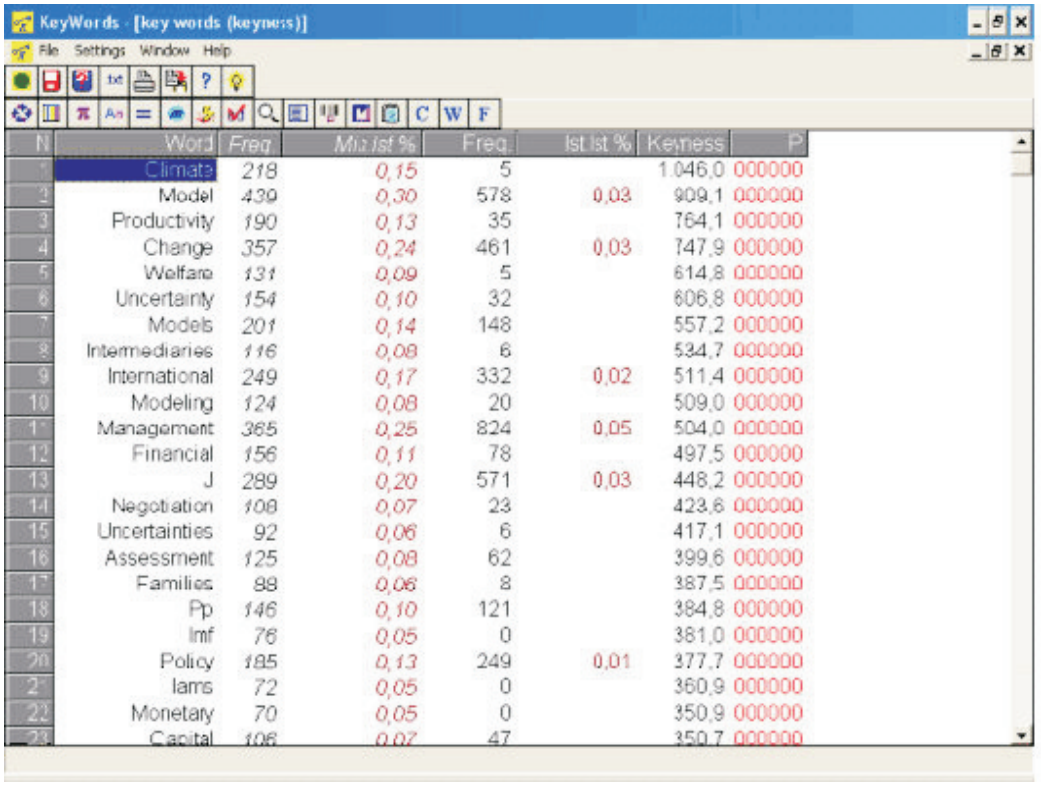
40 other students took my ESP course (for Computer and Business undergraduates) in 2005/2006. In this case, as mentioned, the glossary was not used as a built-in resource based on corpus analysis. Instead, I decided to let the students do their own analysis by means of electronic concordances applied to text sources from the same corpus. The texts were made available by folders (10 sources in each of the five subject folders – i.e., Marketing, M.I.S., Statistics, Management, and Accounting). Again, all the work was supervised online by means of the Moodle platform, enabling the tracking of students' access, development, and production of results. The same test as in the previous (2004/2005) group was given to these students at the end of the course, after they had exploited parts of the electronic corpus provided.

The students used a minimum of 30 digital texts from the corpus (from three subject folders) to answer the activities assigned throughout the course. In contrast with the focus adopted in 2004 (i.e., to explore the 500-word glossary via wordlists and detailed entries), in 2005/2006 the 40 students were instructed to work with their own wordlists and concordance / collocation material.

As the first activity, they had to contrast words in created frequency lists using sub-corpora such as genres or topics. Some students even expanded the sub-categories made to include sections, e.g., article conclusions and abstracts in articles. As the second approach, the students managed the concordance material to find suitable lexical items collocating significantly. Frequency of node-word was thus made a key guiding principle. One example is the noun information combining with other nouns and verbs like technology, processing, provide, flows, as noted down by many students.

The third assignment then asked the students to identify very restricted content items, often considered as technical by the literature (e.g., Thurstun & Candlin, 1998). They

used different corpus selections to exploit subject-based lexical items. Positive keywords were observed as appearing significantly in a given sub-collection of texts, e.g., weather forecasting in Management Information Systems. This particular keyword search is illustrated in Figure 4; for this analysis, only two texts were used (where the topic was discussed).



N	Word	Freq	Mlz list %	Freq	list list %	Keyness	P
1	Climate	218	0,15	5		1.046,0	000000
2	Model	439	0,30	578	0,03	909,1	000000
3	Productivity	190	0,13	35		764,1	000000
4	Change	357	0,24	461	0,03	747,9	000000
5	Welfare	131	0,09	5		614,8	000000
6	Uncertainty	154	0,10	32		606,8	000000
7	Models	201	0,14	148		557,2	000000
8	Intermediaries	116	0,08	6		534,7	000000
9	International	249	0,17	332	0,02	511,4	000000
10	Modeling	124	0,08	20		509,0	000000
11	Management	365	0,25	824	0,05	504,0	000000
12	Financial	156	0,11	78		497,5	000000
13	J	289	0,20	571	0,03	448,2	000000
14	Negotiation	108	0,07	23		423,6	000000
15	Uncertainties	92	0,06	6		417,1	000000
16	Assessment	125	0,08	62		399,6	000000
17	Families	88	0,06	8		387,5	000000
18	Pp	146	0,10	121		384,8	000000
19	Imf	76	0,05	0		381,0	000000
20	Policy	185	0,13	249	0,01	377,7	000000
21	Iams	72	0,05	0		360,9	000000
22	Monetary	70	0,05	0		350,9	000000
23	Capital	106	0,07	47		350,7	000000

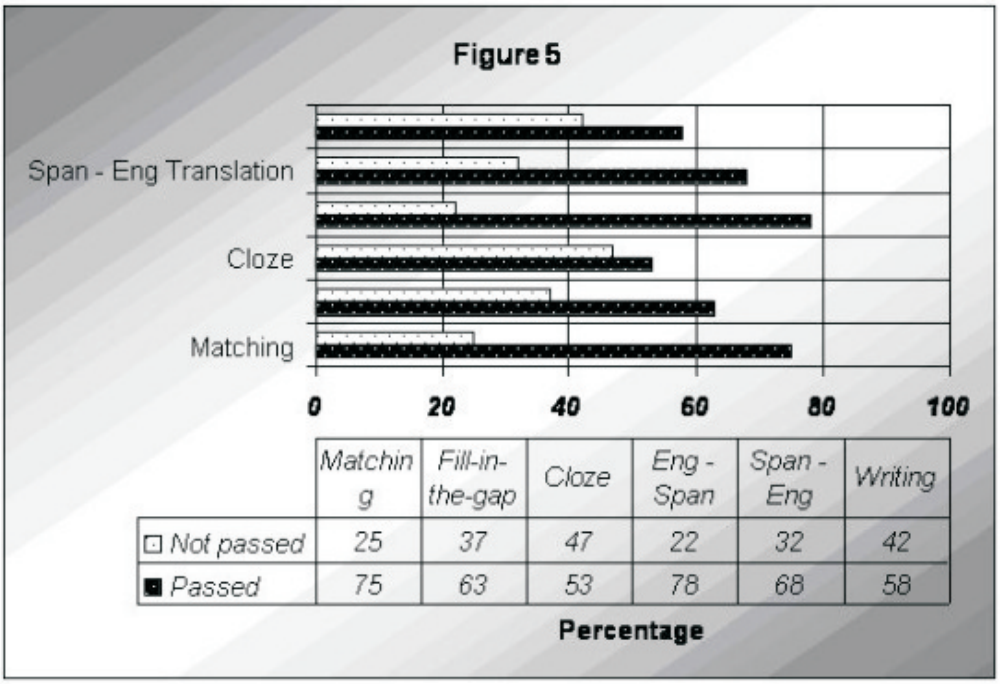
Figure 4: Example of Keyword search for the topic of weather forecasting

The activities were completed successfully by most students. It was generally found that once the students understood what was required and why, they got along fine with the task procedures. In fact, most of the students worked with the tool while having a clear realization of its utility, especially those using it more at home via the Moodle system (i.e., those with more time available to get to know the tool). In general, the Computer Science students became familiarized with the tool faster, but the Business students were equally capable of producing good results.

In the test taken at the end of the course, some positive scores were recorded (Figure 5). The six activities (the same ones that Group 1 did in 2004/2005) showed some improvement, especially in the Spanish-to-English translation, fill-in-the-gap, and cloze exercise (although this last one still scores too low).

The post-test questionnaires (http://www.unex.es/lengingles/Alejandro/appendix.pdf) contribute to providing a contrastive view of the information. The students tended to consider corpus-based activities as easy or average (questions 4, 5, and 6). They also stated their preference for lexical searching, vocabulary organization, and

academic/technical expressions (7 and 8). The lack of Spanish equivalents (e.g., questions 8 and 10) is seen as negative, but their understanding of key academic and technical lexis is important for most students (question 9). In addition, most students are quite aware that working with lexical collocations and concordances is beneficial (question 12).



	Matching	Fill-in-the-gap	Cloze	Eng - Span	Span - Eng	Writing
Not passed	25	37	47	22	32	42
Passed	75	63	53	78	68	58

Figure 5: Passing / non–passing scores by group 2

In relation to the lexical activities, they preferred matching, fill-in-the-gap, and cloze exercises (questions 13, 14 and 15). Writing was regarded as being most likely to improve from lexical work (question 16), followed by speaking and terminology. Many students also acknowledged that text, material and working with special terminology can lead to a positive intake (question 17). In contrast, they tended to view listening skills as being less important (question 18). Some also alluded to the need for collaborative work (question 19). Finally, about 33 per cent referred to the complementary role that electronic activities (question 20) can have for ESP, agreeing that important academic/technical items can easily be approached via these tools.

Overall, the students' responses showed some degrees of satisfaction with lexical acquisition (at least in part or a little, as question 11 shows). In general, their comments tended to corroborate the passing scores in lexical production (e.g., they tended to realize the importance of academic lexical constructions).

Given the scores in each test (for Groups 1 and 2), a statistical correlation can be made to confirm that Group 2 really did better than Group 1, and that such scores were not due to chance. Thus, the mean scores could be a starting point for the calculation of significant difference. The two mean scores result from the 40 scores for each group

in the tests (from 1 to 10): 5.9 (for Group 1) and 6.12 (Group 2). The statistical software (Decision analyst, 1998) also computed the standard deviations of those means (1.782 and 1.471, respectively). With these two values, a t-value can be estimated: .602153 (above .4999 as a significant difference). This score has a probability of 45.12 per cent, i.e., there is a high percentage of difference (almost 50 per cent) between both test performances.

5. Conclusions

The first major conclusion is that the students in Group 2 outperformed those in Group 1 by scoring higher in the lexical and translation/writing activities. It seems that the Group 2 students were better able to capture and process the lexical items. They realized, as the data from questions 7, 9, and 12 may point out in the questionnaires, that they can gain lexical knowledge in the process of electronic corpus-based work.

At the beginning of each course, the two groups took a placement pre-test (corresponding to intermediate levels [Swan & Walter, 1997]), and they obtained similar scores (52 per cent of the students passed in Group 1 and 56 per cent in Group 2). A key difference also shown in the post-tests is that a bigger percentage in Group 2 received a score of 8 or above (eight versus only two in Group 1). Thus, given this linguistic command variation, the test results are not surprising statistically or academically. More students in Group 2 acquired lexical input as the course went along and, in the process, the electronic corpus-based tasks may have helped. Still, as Figure 6 points out this group's rate of lexical recognition, awareness and production is significantly high by comparison. Such a distinction between both groups' attitudes in relation to their lexical performance after the tasks, tests and ESP course cannot be only due to Group 2's slightly higher language command. The type of methodology based on the electronic corpus work must have had something to do with this variation since no other variables were introduced in the course.

The second main conclusion is that the integration of e-corpus tasks in ESP is a means of improving lexical knowledge. Related to this, the increase in Group 2's lexical awareness is perceived and derived from the data in the questionnaires (e.g., item 3). Since this group's members were active observers of language in use by managing their own wordlists, concordance-based items and keywords, they gained more confidence about independent work. They also managed to gain more lexical input in their consistent use of the tool (i.e., over a period of three-to-four weeks). This impression is not only based on their test scores and questionnaire answers, but also on the direct observation of their electronic work (i.e., via the Moodle platform). The time spent on use of the tool led them to appreciate a certain degree of autonomy and

self-esteem as student researchers/analysts.

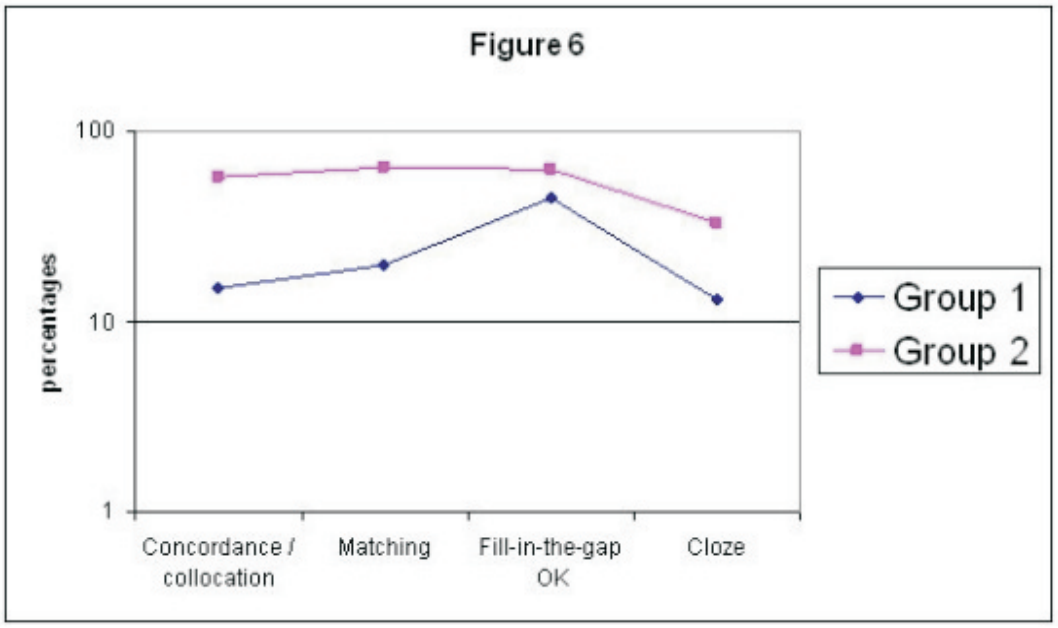


Figure 6: Contrastive display of attitudinal differences between the two groups

The third concluding remark stems from the two groups' work: both tend to show positive views and comments about the lexical focuses used in their respective courses. They realize, for example, that the translation of academic constructions fosters their decoding skills to clarify semantic aspects. Nonetheless, Group 1 did not seem to apply their lexical knowledge as well as could be expected (i.e., as the writing results in the test indicate). In contrast, group 2 did better in this respect, and they fully acknowledged the important place of academic expressions for writing. This group's answers about the importance of subject-based vocabulary, words in context, and technical texts confirm their appreciation of conscious lexical training (items 17 and 20).

Finally, as a fourth conclusion, we should highlight a major negative aspect: the lexical productivity is still far from being optimal as the different answers in the questionnaires suggest (e.g., item 11, where most students - in the two groups - perceived that they had improved their mental lexical database in part or a little, and few of them answer much or very much). This less satisfying side, far from presenting any signs of discouragement or neglect, should kindle the spirit of innovative teaching/research. The dynamic approach to ESP methodology should be always non-conforming and consistently involved, meaning that positive results must be valued and re-assessed for on-going exploration within that path. Since the contrastive study perceived good production feedback, the line of e-corpus work can definitely be used as a complement to foreign language teaching, at least from the experience gathered and contrasted. This claim is in line with previous studies on academic word lists, widely used in ESP, maintained and updated by corpus tools and IT.

References

Aston, G. (2001). Text Categories and Corpus Users: A Response to David Lee. *Language Learning and Technology*, 5(3), 73-76.

Bowker, L. and J. Pearson. (2002). *Working with Specialized Language: A Practical Guide to Using Corpora*. London: Routledge.

Breiner-Sanders, K. E., P. Lowe, J. Miles and E. Swender. (2000). ACTFL Proficiency Guidelines – Speaking (Revised 1999). *Foreign Language Annals*, 33(1), 13-18.

Connor, U. and T.A. Upton (Eds.). (2004). *Discourse in the Professions. Perspectives from Corpus Linguistics*. Amsterdam: John Benjamins.

Conrad, S. (1996). Investigating Academic Texts with Corpus-Based Techniques: An Example from Biology. *Linguistics and Education*, 8, 299-326.

Curado Fuentes, A. (2006). A Corpus-Based Focus on ESP Teaching. Teaching English with Technology. *A Journal for Teachers of English*, 6(4) [online]. Available: http://www.iatefl.org.pl/call/j_esp26.htm (19 December 2006).

Decision Analyst (1998). *Stats, Statistical Software 1.1.* [online]. Available: http://statpages.org/javasta2.html#General (15 January 2005).

Flowerdew, L. (2000). Investigating Referential and Pragmatic Errors in a Learner Corpus. In L. Burnard and T. McEnery (Eds.), *Rethinking language pedagogy from a corpus perspective* (pp. 117-124). Frankfurt am Main: Peter Lang.

Flowerdew, L. (2004). The Argument for Using English Specialized Corpora to Understand Academic and Professional Language. In U. Connor, U and T.A. Upton (Eds.), *Discourse in the Professions. Perspectives from Corpus Linguistics* (pp. 11-36). Amsterdam: John Benjamins.

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

Hoey, M. (2005). Lexical Priming. A New Theory of Words and Language. London: Routledge.

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

Johns, T. (1991). Should you be Persuaded: Two Examples of Data-Driven Learning. Classroom Concordancing. *English Language Research Journal*, 4, 1-16.

Nation, P. (2001). Using Small Corpora to Investigate Learner Needs: Two Vocabulary Research Tools. In M. Ghadessy, A. Henry and R.L. Roseberry (Eds.), *Small Corpus Studies and ELT. Studies in Corpus Linguistics* (pp. 31-46). Amsterdam: John Benjamins.

Rowley-Jolivet, E. and S. Carter-Thomas. (2005). Genre Awareness and Rhetorical Appropriacy: Manipulation of Information Structure by NS and NNS Scientists in the International Conference Setting. *English for Specific Purposes*, 24(1), 41-64.

Scott, M. (1999). *WordSmith Tools 3.0.* Oxford: Oxford University Press.

Swan, M. and C. Walter. (1997). *How English Works*. Oxford: Oxford University Press.

Thurstun, J. (1996). Teaching the Vocabulary of Academic English via Concordances. Paper presented at the *Annual Meeting of the Teachers of English to Speakers of Other Languages*, March, 1996. Chicago.

Thurstun, J. and C.N. Candlin. (1998). Concordancing and the Teaching of the Vocabulary of Academic English. *English for Specific Purposes*, 17(1), 20-34.