EXAMINING ENGLISH FOR ACADEMIC PURPOSES STUDENTS' VOCABULARY OUTPUT: CORPUS-AIDED ANALYSIS AND LEARNER CORPORA

PASCUAL PÉREZ-PAREDES
PURIFICACIÓN SÁNCHEZ HERNÁNDEZ
Universidad de Murcia

ABSTRACT. *Vocabulary is the most prominent linguistic component in the characterisation of reality (Alcaraz 2000), influencing the acquisition of an L2 to a great extent. Consequently, the acquisition of lexical items is of maximal importance in the learning process. This paper deals with the use of corpus linguistics to promote reading and enhance storage vocabulary for production in a specific field. As part of the investigation, we compiled a corpus from the Journal of Psychotherapy. This corpus was at the disposal of students of English for Academic Purposes in the field of psychology. Learners were asked to consult the corpus as often as they wanted and to produce a specialised text. With the students' writings we built a learner corpus and thus established lexical relationships between the input and the output copora. Our results confirm previous corpus-based research on learner interlanguage. Students overused highly technical and general vocabulary in their writing. These learners are more overtly "present" within their discourse than the expert writers, that is, those who contributed the specialised corpus.*

KEYWORDS: *English for Academic Purposes, specialized languages, Corpus Linguistics, vocabulary learning and teaching.*

RESUMEN. *La adquisición del léxico de una lengua extranjera es de suma importancia en el proceso de aprendizaje de la misma. No en vano, el vocabulario de un idioma se configura como uno de los elementos más sustantivos en la caracterización y representación de lo fenomenológico (Alcaraz 2000). El presente artículo se vale de los procedimientos de trabajo propios de la lingüística del corpus con una doble finalidad: favorecer las destrezas lectoras de los estudiantes y, a la vez, mejorar la capacidad de aprendizaje del léxico propio de un lenguaje especializado. Como parte de nuestra investigación, recopilamos un corpus del* Journal of Psychotherapy. *Este corpus se puso a disposición de los aprendices de inglés para fines específicos de la rama de Psicología. Asimismo se les pidió que redactasen un texto sobre la sub-especialidad en cuestión. Con estas redacciones recopilamos un corpus de aprendices de inglés que, posteriormente, usamos para comparar la utilización del léxico con el corpus anteriormente citado. Los* *resultados de nuestro trabajo confirman las conclusiones de investigaciones previas en el campo de la lingüística del corpus. Los estudiantes utilizaron en exceso el vocabulario muy técnico y el vocabulario general, delatando así su "presencia" como autores en mayor medida que los expertos en la lengua de especialidad.*

PALABRAS CLAVE: *Inglés para Fines Académicos, lenguajes especializados, lingüística del corpus, aprendizaje y enseñanza del vocabulario.*

1. Introduction

1.1. Categorizing L2 vocabulary

The development of students' vocabulary is not a specific study skill, but is related to all language learning and is of concern to all four language skills (Jordan 1997). In this paper, vocabulary will be treated as a link between reading and writing as there is a transfer from one to the other as has been expressed by Nattinger (Jordan 1997: 149)

> *Comprehension of vocabulary relies on strategies that permit one to understand words and store them, to commit them to memory, that is, while production concerns strategies that activate one's storage by retrieving these words from memory, and by using them in appropriate situations. The priority this distinction assigns to comprehension is one of many reasons why a growing number of researchers believe that comprehension should precede production in language teaching.*

Students often express a need to expand their vocabulary. In fact, vocabulary seems to be the cause of most difficulties for the students (Jordan 1981). The same author (1997) proposes that, with students of different language levels, background and specific subjects attending English for Academic Purposes (EAP) courses, an understandable emphasis may be placed on indirect learning.

The first question emerging from these considerations is related to the kind of vocabulary that should be taught/learned, and how it should be taught/learned in our learning programs. According to Carter (1987), the vocabulary appropriate for students following EAP courses should clearly be more advanced than the core 2,000-3,000 words that provide the basis of about 80 per cent of the words likely to be encountered in a general language course. Carter (Jordan 1997: 151), in arguing about core vocabulary in discourse, points out that:

> *At least two broad distinctions have to be drawn. There is a level of core vocabulary which is "core" as far as the organisation of the lexicon as a whole is concerned; and there is a level of core vocabulary which is core to a particular field or subject. Subject-specific vocabulary will always be non-core as far as the language as a whole is concerned. This is because it is not neutral in field and is immediately associated with a specialised topic.*

The same author also looks at discourse-genres that apply to writing in different subjects. His initial research suggests that the presence of core, subject-core and non-core lexical items can be connected with particular discourse-genres. The following correlations between lexical coreness and genre have been observed:

Discourse genres	Lexical coreness
Summary	Core
Explanation	
Argumentation	Non-core
Narrative	
Description	
Instruction	Subject core
Report	
Recount	

Table 1. Correlations between lexical coreness and discourse genre according to Carter.

The above seems to present a rationale for determining vocabulary types distribution in L1 texts. Other authors (Kennedy and Bolitho 1991; Dudley-Evans and St John 1998) make a different classification: they speak about 1) highly technical vocabulary the first and technical the second and 2) subtechnical vocabulary. Alcaraz adds a third category to the preceding ones: 3) the general vocabulary of frequent use in a speciality, which in the case of Dudley-Evans and St. John is included in the subtechnical vocabulary.

Every academic subject has its own set of highly technical terms which are an intrinsic part of the learning of the discipline itself, and is formed by lexical units called "terms". The main difference between these "terms" and the lexical units of the general language is that the former are monosemic, whereas the latter are polisemic (Alcaraz 2000). Terminology is the vocabulary that presents fewer difficulties for foreign EAP students since it is monosemic and precise in meaning. Subtechnical vocabulary, on the contrary, consists of those words which are not specific to a subject speciality, but which occur regularly in one field of knowledge. Sager *et al.* (1980) call them "re-designated general language items". It is polisemic vocabulary formed in most cases by extension of the meaning through the process of analogy. And finally, in the third group we include words of general use that, without losing their own meaning, are in the "neighbourhood" of the speciality. These words are non technical *sensu strict* because they keep their original meaning. Due to their high rate of presence in a speciality, they are at least as essential in an academic field speciality as those belonging to the two previous groups. We will use this classification in this study.

Whether the teaching of highly technical vocabulary is the duty of a language teacher is an open question. While most of the authors agree that in general it should not be the responsibility of the ESAP teacher to teach technical vocabulary, it may be his duty to check that the students have understood the lexicon when asked to perform any lexical or grammatical activity (Kennedy and Bolitho 1991; Dudley-Evans and St. John 1998). It follows, then, that the two other categories should be given priority in the teaching of an ESAP course. However, how can teachers evaluate whether these statements hold true for their students? Furthermore, how can they gain any sort of insight into their students' actual use of L2 vocabulary? This is the scope of the following sections.

1.2. Vocabulary learning based on corpora

The teaching of vocabulary in English for Academic Purposes (EAP) follows similar principles to those in English for General Purposes (EGP) (Dudley-Evans and St. John 1998). Notwithstanding, a distinction should be made between vocabulary needed for comprehension and that needed for production. In comprehension, deducing the meaning of vocabulary from context is the most important method of learning new vocabulary. For production, storage is essential. Nattinger (1988) suggests various techniques for storing vocabulary: the use of word association, mnemonic devices and *loci* that is the use of visual images to help remember a word.

One of the most innovative techniques is that of the use of corpora together with situation and textual analysis. The development of corpora of specific texts has provided an invaluable research and teaching tool for vocabulary study and acquisition. Among other facilities corpora provide us with the opportunity to draw up lists of key lexical items, either in general texts or in specific disciplines. Specialised texts of any sort, whether written or spoken, exhibit various characteristic lexical features. These can be isolated, analysed and used as subjects for useful exercises for students (Kennedy and Bolitho 1991).

Granger's work (1998) stresses three areas where learner corpora may be useful in foreign language teaching. One of them is contrastive analysis. Exploring students' output can be thus functional in different ways: teachers may wish to contrast inter-group productions, intra-group productions or L1 and L2 speakers' productions. Altenberg and Tapper (1998) are good examples of such an approach.

Within this general framework, one of the domains where corpora are extremely practical is that of the study of learners' active vocabulary. The computational aspect of corpus linguistics makes the analysis of the aforementioned cross-skill area both feasible and convenient. With different purposes, researchers have used a wide range of approaches to the issue. Dagneaux *et al.* (1998) advocate the use of computer-aided error analysis to scrutinize students' production in L2 learning. Ringbom (1998) relies on descriptive statistical approaches to determine the extent to which a learner corpus displays inherent linguistic features. Other studies using L1 corpora propose concordancing to teach academic English (Thurstun and Candlin 1998), detailed analyses of corpora instances to enhance language learning for specific purposes (Beeching 1997), the convenience of corpus resources for the teaching and learning of L2 vocabulary (Murphy 1996) and, just to exemplify a further approach, the development of specific vocabulary corpora to extend the knowledge of students on less-frequently used academic lexicon (Rance-Roney 1995). One of our concerns in this investigation is to explore our students' production of technical and subtechnical vocabulary using methods pertaining to corpus linguistics.

2. Investigating students' vocabulary output

We programmed a pilot experience in our classroom with students of the 5th year of the Degree of Psychology. We were aware that, as mentioned above, that specific vocabulary can be used for both comprehension and for production purposes. As far as the first one is concerned, we had checked that, in general terms, our students showed no difficulties in deducing the meaning of a word from the context when dealing with specialised texts. This situation is probably influenced by the fact that the roots of many English words used in the speciality come from Latin, and Spanish is a romance language. Also the intermediate level of our students played a significant part in that process. During the course, we exploited the techniques of situation and textual analysis, as well as the ones referred to collocation and the use of corpora.

2.1. Methods

To carry out our experiment, we (1) gathered a mini-corpus of around 50,000 words from texts of the speciality published in English in the *Journal of Psychotherapy Practice and Research* and (2) used a text on Psychotherapy (Gibert Maceda 1991). Both *corpora*[1] were at the disposal of the students. They were asked to read carefully the text on Psychotherapy, on the one hand, and to consult the *Journal of Psychotherapy* mini-corpus as often as they considered it necessary. Subsequently, they were encouraged to produce a small text on the topic of Psychotherapy to be delivered to the teacher four weeks later. Directions were provided on the scope of the task to ensure uniformity. With the writings that we collected, a learner-corpus was built. We wanted to establish a relationship between the input and the students' output, as far as the acquisition of vocabulary is concerned. We will refer to the corpus from the *Journal of Psychotherapy* as Corpus 1, to the text on Psychotherapy as Corpus 2, and the collection of texts produced by the students as Corpus 3.

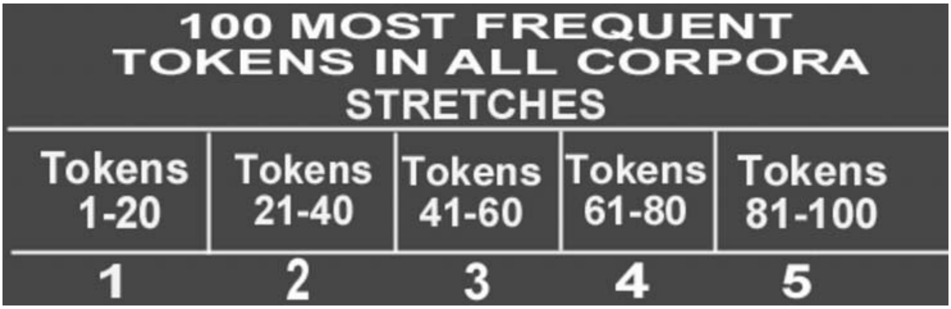
Twenty three students contributed the corpus and 3431 tokens were totalled. In a search for functionality and ease of interpretation, the most frequent one hundred tokens were subject to scrutinising. These tokens were isolated in every corpus and, subsequently, analysed and computed, first in terms of technical/ non-technical adscription and then, those fitting in the first category, in terms of their relationship with the very specialist topic under consideration. Vocabulary items which did not fit in any of the three categories were not considered. Technical vocabulary was broken down in further classifications. We estimated this to be a reliable way to scrutinize the EAP learner vocabulary output, as it covered every aspect of vocabulary learning in specialised language courses. In order to gain a deeper understanding of the lexical frequency distribution of this vocabulary, further splits were performed on all corpora and, as a result, five new stages of analysis emerged. Every frequency list was divided in 20-word layers, or stretches, giving us the chance to identify cumulative frequency distribution up to token 100. Three ideas underlie this approach: to test whether the frequency criterion is of any use to assess students' lexical output; to test the scope of our frequency analysis in terms of foreign language teaching implementation and, finally, to test whether significant differences emerged in the different stages and layers of analysis.

3. Results

3.1. Descriptive statistics

All corpora were subject to analysis with OUP Wordsmith Tools 3.0 and Minitab 13.31. Four wordlists containing the one hundred most frequently used tokens were produced. Lemmas were not considered in this exploratory study, as our primary concern was to gain insight into the nature of student's L2 production and, for that purpose, we believed it necessary not to alter any morphosyntactic features in the corpora. In doing so, we pursued a better contrastive analysis between native and non-native corpora[2] and an in-depth look into students' interlanguage lexical patterns.

We divided these wordlists in five stretches, containing 20 tokens each. Figure 1 illustrates this.



100 MOST FREQUENT TOKENS IN ALL CORPORA STRETCHES				
Tokens 1-20	Tokens 21-40	Tokens 41-60	Tokens 61-80	Tokens 81-100
1	2	3	4	5

Figure 1. 100 most frequent tokens in all corpora.

After that, we classified the tokens included in the first stretch as (1) highly technical vocabulary, (2) subtechnical vocabulary and (3) general words of frequent use in the speciality. With the remaining four stretches in the lists the same procedure was followed. Figures 2 and 3 show the descriptive results obtained after analysing the data in the corpora.

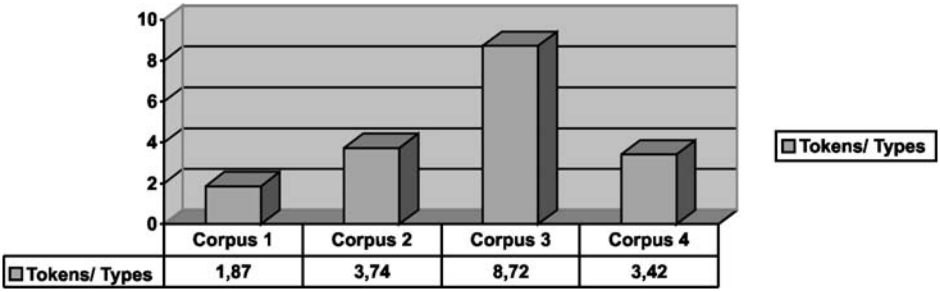


Figure 2. Token/ Types relationship.

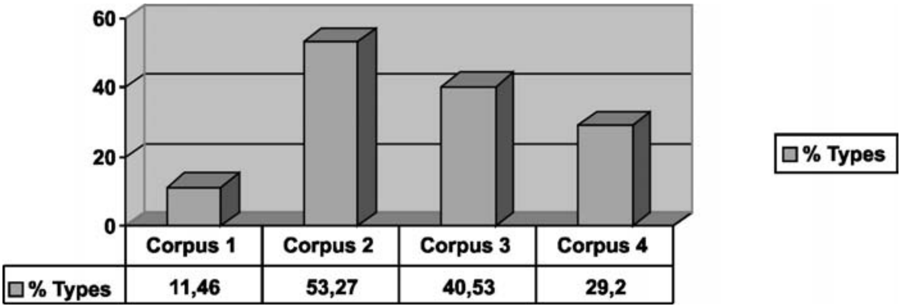


Figure 3. Types percentage.

Corpus 3 yielded a 3.74 Token/Type ratio and 26.73 standardised Type/ Token ratio. Corpus 1 yielded a 8.72 Token/Type ratio and 11.46 standardised Type/ Token ratio. As expected, corpus size plays a major role in determining the magnitude of these figures. In order to increase the validity of our study, we compiled a control sub-corpus from Corpus 1. This new corpus, Corpus 4, consists of exactly the same amount of tokens as Corpus 3, that is, 3431. The texts which contributed Corpus 4 were selected at random to ensure proportional representativeness. Interestingly, texts conforming Corpus 4 had the same token average length, that is, 149 tokens. It yielded a 3.42 Token/Type ratio and 29.10 standardised Type/ Token ratio. Size, as already stated, is a key issue when comparing corpora. Cantos (2000: 73) states that "the reliability of the token-type and type-token ratio as quantitative indicators of lexical diversity or lexical density are constrained because of their dependence on text size"[3].

The Token/ Type ratio is considerably larger in Corpus 3, where 40.53% of the tokens are types at the same time. It is interesting to note here that Corpus 4 rendered a 29.2% figure, significantly lower than Corpus 3 especially if we remember that both have the same number of tokens. Corpus 3 informants seem to rely heavily on vocabulary diversification as text-building strategy. In a similar way, we can point out how Corpus 1 presents a more canonical approach to text-building. Professional, academic writers probably are more concise in their expositions and do not depend so strongly on high lexical density to ensure coherence and cohesion.

Also, the data speak volumes about the uniformed practices of a professional group of widely-read researchers versus a non-uniformed group of would-be psychologists who are still acquiring training. The percentage of Corpus 2 confirms how text/ corpus size can determine findings on lexical density and how careful the statements on these aspects must be.

3.2. Inferential statistics

In order to determine the significance of the data produced by the vocabulary typology in the corpora we decided (1) to calculate a confidence interval for the difference between the two proportions which Corpus 3 and 4 presented, and (2) to carry out a significance test on a difference between those two proportions. In essence, a statistical test is a procedure for deciding whether a hypothesis on a quantitative feature of a population is true or false. A hypothesis of this sort is performed by drawing a random sample from a population and calculating an appropriate statistic. If we obtained a value of the statistic that would hardly ever occur when the hypothesis is true, we would have reason to reject the hypothesis. Following this procedure, it is usual to reject the hypothesis tested when the statistic has a value that is among those that, in theory, would be expected to occur no more than 5 out of every 100 times that a random sample (of the same size) is drawn from the population in question when the hypothesis is, in fact, true. Finally, it is noteworthy that the appropriate conduct of any statistical test invariably requires many careful decisions. It is, for example, always necessary to decide what statistic to use, what sample size to employ and what criteria to establish for rejection of the hypothesis tested.

3.2.1. *Inference on the difference between two proportions*

Table 2 shows how the 100 most frequent tokens in the five stretches studied are distributed according to the vocabulary typology proposed in this work. The figures represent percentages of the total token figure in each corpus.

	S1 G1	S1 G2	S1 G3	S2 G1	S2 G2	S2 G3	S3 G1	S3 G2	S3 G3	S4 G1	S4 G2	S4 G3	S5 G1	S5 G2	S5 G3
Corpus 1	0.7	1.7	0.5	0	0	0.6	0	0.8	0.4	0	0.4	0.3	0.1	0.3	0.6
Corpus 2	0	0	3.2	0.6	1.3	1.9	0	1	0.6	0	1	1	0	0.3	0.6
Corpus 3	2.7	0	1.4	0.5	0	0.6	0.6	0.6	1.9	0	0	0.4	0	0.2	0.3
Corpus 4	0.03	0.08	0.02	0	0.03	0.08	0	0	0.02	0.02	0.02	0.02	0	0.02	0

Table 2. Percentages of the 100 most frequent token in all 5 stretches of analysis.

With the data above we set out to calculate a confidence interval for the difference between the two proportions which shared the same number of components, that is, Corpus 3 and 4. Our ultimate aim was to carry out a significance test on the difference between those two corpora proportions and, in doing so, to check whether a null hypothesis (Corpus 3 proportion 0 Corpus 4 proportion) held true or not. For each stretch, three comparisons had to be set up, one for each group of vocabulary. As significance cut-off, =0.05 was established which implied a new, and more demanding p value of 0.017[4]. Accordingly, figures >0.017 meant that the null hypothesis would be accepted. Table 3 shows p values for every stretch and vocabulary group slot.

	Group 1	Group 2	Group 3
Stretch 1	E p-value 0.000	E p-value 0.655	E p-value 0.000
Stretch 2	E p-value 0.000	E p-value 0.317	E p-value 0.000
Stretch 3	E p-value 0.000	E p-value 0.000	E p-value 0.000
Stretch 4	E p-value 0.317	E p-value 0.317	E p-value 0.000
Stretch 5	E -	E p-value 0.059	E p-value 0.001

Table 3. Behaviour of the proportions analysed: Corpus 3 and Corpus 4.

It is noteworthy that vocabulary Group 3 presents different behaviour in all five stretches. As a way of contrast, vocabulary Group 2 presents similar behaviour in terms of proportions in four of the five stretches analysed. Vocabulary Group 1 behaviour is divergent in the first three stretches, those which encompass most frequent tokens one to sixty.

4. Discussion and conclusions

The discussion on the results is established on the facts presented in Table 3 where the behaviour of the proportions analysed is shown. These data are directly related to the nature of student's L2 production, which was our main concern when we planned this study. Firstly, the p-values exhibited in vocabulary Group 3, that is, the words of general use that without losing their own meaning are in the "neighbourhood" of the speciality, demonstrate that this type of vocabulary has been overused by our students in the five stretches. The figures corresponding to Group 1, that is, highly technical vocabulary, manifest as well an overuse of technical terms in the first 3 stretches. On the contrary, the subtechnical vocabulary, Group 2, displays p-values above 0.017 which means that the use students have made of this type of terms is similar to that made by the expert in Corpus 1 and 2. In none of the three groups an infra use of technical, subtechnical of general vocabulary has been detected, which, in the end, reveals that our students considered the two input corpora in their output one and, going a bit further, that our vocabulary learning methodology gave good results.

If we try to go deep into the reasons why our students have used technical and general vocabulary in excess, we can venture some reasons: on the one hand, technical vocabulary in the field of Psychology, in general, and in the sub-area of Psychotherapy in particular, has its origin in Latin. Spanish is a romance language, consequently we can justify this over-use on the basis of the similar inter-language lexical patterns. In addition, as already hinted, technical terminology is the range of vocabulary that presents fewer difficulties for foreign EAP students, since it is monosemic and more specific in meaning. With respect to the general vocabulary items, we justify their abuse on the basis of the very nature of this type of vocabulary as it is made by the lexical items that students are most used to recognising and so to employing. The most difficult range of terms to acquire and use are the subtechnical. However, according to the findings presented here, the ratio obtained between the reading and the writing production is highly satisfactory.

Our results confirm previous statements on language learners. Petch-Tyson (1998) has made use of corpora to analyse reader and writer visibility. She believes that the presence of the participants in the discourse is encoded more or less overtly. In her research, she showed that the non-native speakers group used more of the features which identify visibility, such as first person reference, fuzziness or imperatives (p. 111). In our study, our purpose was different but we have arrived at the same range of conclusions. Our students overused highly technical and general vocabulary in their writing. These learners are more overtly present within their discourse than the expert writers, that is, those who contributed Corpus 1. We may assume that this overt presence is clear in other language areas such as organizational features or syntactic distributional patterns. However, this was not analysed in this work.

Curado-Fuentes (2001) focused his study on collocation in the context of English for Specific Purposes (ESP), and, more precisely, within English for Information Science and Technology. He showed how the results of the contrastive study of lexical items in small specific corpora can become the basis for teaching and learning ESP. We believe that the corpus-based approach we propose here is optimal to learn EAP vocabulary in its context. *Profiling* (Crystal 1991) the learners' lexical use gave us the tools to assess their actual usage of this crucial aspect of the learning experience, while work with corpora offered students the chance to, apparently, increase their competence in the English language.

5. Notas

1. We will refer to this text from Gibert Maceda (1991) as a corpus as, computationally, it was treated with the same range of analytical tools as the rest of corpora in this investigation.
2. Granger (1998)
3. Biber (1993) has pointed out how cumulative tokens are distributed linearly while the cumulative types are distributed curvinearly.
4. α'=0.05/3=0.017

6. References

Aijmer, K. and B. Altenberg, eds.1991. *English Corpus Linguistics*. London: Longman.

Alcaraz, E. 2000. *El inglés profesional y académico*. Madrid: Alianza Editorial.

Altenberg, B. and M. Tapper, 1998. "The use of adverbial connectors in advanced Swedish learners' written English". *Learner English on Computer.* Ed. S. Granger. Harlow: Longman.

Armstrong, A. ed. 1994. *Using Large Corpora.* Cambridge, Mas.: MIT.

Beeching, K. 1997. "French for Specific Purposes: The Case for Spoken Corpora". *Applied Linguistics*, 18:3, 374-94.

Biber, D. 1994. "Using Register-Diversified Corpora for General English Studies". *Using Large Corpora.* Ed. Susan Armstrong Cambridge, Mas.:MIT.

Carter, R. 1987. "Vocabulary and second/foreign language teaching". *Language Teaching*, 20(1).

Crystal, D. 1991. *Sylistic profiling*. In Aijmer and Altenberg (eds.).

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

Dagneaux, E., S. Denness and S. Granger. 1998. "Computer-aided Error Analysis". *System: An International Journal of Educational Technology and Applied Linguistics* 26(2): 163-174.

Dudley-Evans, T. and M. St. John. 1998. *Developments in English for Specific Purposes*. Cambridge: Cambridge University Press.

Gibert Maceda, T. 1991. *Inglés para Universitarios* (Psicología). Madrid: UNED.

Granger, S. ed. 1998. *Learner English on Computer*. Harlow: Longman.

Jordan, R. R. 1981. "Comment Améliorer l'Anglais Écrit de l'Etudiant Étranger en Université Britannique". Paris: *Études de Linguistique Appliquée*, Vol. 43.

Jordan, R. R. 1997. *English for Academic Purposes. A guide and resource book for teachers*. Cambridge: Cambridge University Press.

Kennedy, C. and R. Bolitho. 1991. *English for Specific Purposes*. London: Macmillan Press Ltd.

Lorenz, G. 1998. "Overstatement in advanced learners' writing: stylistic aspects of adjective intensification". *Learner English on Computer*. Ed. S. Granger. Harlow: Longman.

Murphy, B. 1996. "Computer, corpora and vocabulary study". *Language Learning Journal*, 14, 53-57.

Petch-Tyson, S. 1998. "Reader/Writer Visibility in EFL Persuasive Writing". Ed. S. Granger.

Rance-Roney, J. 1995. "Transitioning Adult ESL Learners to Academic Programs". Washington, DC: National Center for ESL Literacy Education.

Ringbom, H. 1998. Vocabulary frequencies in advanced learner English: a cross-linguistic approach. *Learner English on Computer.* Ed. S. Granger. Harlow: Longman.

Sager, J.C. et al. 1980. *English Special Languages*. Wiesbaden: Brandstetter Verlag KG.

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