ANALYSING UNDERGRADUATE STUDENTS' L2 PRODUCTIVE LEXICAL PROFILE THROUGH *WORDSMITH TOOLS*

SORAYA MORENO ESPINOSA
UNIVERSITY OF LA RIOJA

1. Introduction

WordSmith Tools (SCOTT 1997) is a technological instrument, which provides insight on how words behave in texts. It is a legitimate tool for teaching, learning and research, which has been used either as an assessing instrument (LI 2000, SAGASTA ERRASTI 2000, NAVES NOGUÉS 2001) or as a corpus analyser (COBB 2000, ALTENBERG & GRANGER 2001, WEBER 2001). Nevertheless, as far as we know, there has not been employed, so as to provide a new insight into the productive lexical profile of a sample of Spanish undergraduate learners of English as an L2.

In this paper, we will present the preliminary results of a study that is being carried out with an homogeneous group of learners of English as L2 at *University of La Rioja*. The purpose of this study is to portray undergraduate students' productive vocabulary in English as an L2, by making use of this electronic tool.

Our presentation will be structured as follows: first, we will review the state of the art of studies that have employed WordSmith Tools; secondly, we will describe our sample of informants and our instrument of assessment; and finally, we will focus on a discussion of the following aspects: (a) assessment of our sample of subjects' embedded productive vocabulary in English as an L2 through this technological instrument; and (b) description of test takers' productive lexical profile.

2. Review of studies that have employed Wordsmith Tools

As previously noted, in this section, we will examine different studies, which have employed this electronic analyser, either as an assessing instrument and/or as a corpus comparison tool.

Amongst the researches that have employed it as a corpus comparison tool, we would like to highlight the following ones:

- COBB (2000): In this article, Cobb reviews three corpus comparison studies, and he replies to them by handling different computational tools, being one of them *WordSmith.*
- ALTENBERG & GRANGER (2001): These scholars use *WordSmith Tools* to analyse the collocability of *make* in large corpora.
- WEBER (2001): This paper highlights the fact that an instrument such as *WordSmith* can be used by undergraduate students to investigate different aspects of language use such as concordances; issue which amongst other things, can help them to raise their own awareness of particular areas of difficulty, as well as it promotes learner autonomy. The final stage of this project deals with encouraging learners to use a selection of the lexical items and expressions they have learnt in their work on concordances, in an L2 written essay.

WordSmith Tools has also been used as an assessing instrument. Thus, some investigations have related its outcome with other manual assessing instruments. Amongst others, we would like to refer to the following ones:

- LI (2000): This scholar examines the relationship between: (a) objective computerised text analysis by making use of *WordSmith Tools* amongst others; and (b) subjective evaluation performed by human raters.

- NAVES NOGUÉS (2001): This author explores: (a) whether two computerised tagged text-analyses of linguistic features of L2 writing (being one of them retrieved by

handling *WordSmith Tools*) correlate; and (b) whether there is any correlation between those text analysers, with manually calculated writing measures.

- SAGASTA ERRASTI (2000): She uses *WordSmith Tools* in order to assess the lexical complexity of a sample of written texts in English, Spanish and Euskara.

Despite having observed that *WordSmith Tools* is a legitimate tool for teaching, learning and research; we have not come across any study that has made use of it, so as to portray undergraduate Spanish students' productive vocabulary in English as an L2, by making use of this electronic tool, that is why, we consider that our study is necessary so as to fill in a gap in that respect.

3. Methodology

3.1. Subjects

Low-intermediate learners of business English as a foreign language at *University of La Rioja* are our sample of informants: 19 students –which comprise females (n = 13) and males (n = 6)– representing subjects with an homogenous single mother tongue (Spanish) and cultural background, whose age ranges from 20 to 24 years old (17 of them) and 25-29 years old (2 of them). All our informants had attended business English lectures during the first term for a period of 60 hours, as well as they have been attending four-hour-weekly-business English lectures during the second term, for a period of 60 hours. The percentage of students according to sex is shown in figure 1.

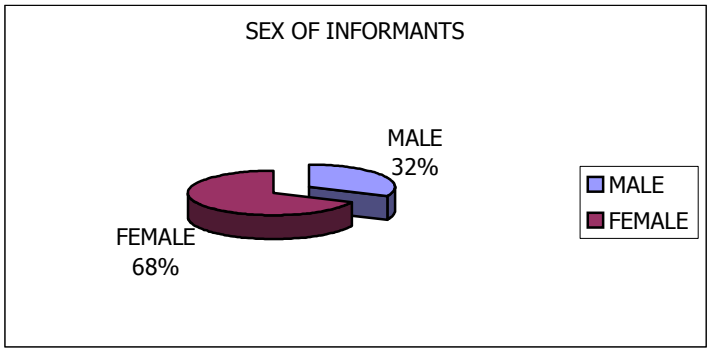


Figure 1. Distribution of students according to sex

3.2. Instruments and procedures

Our data gathering instrument was a written composition task. Clear general instructions were presented orally and in writing, before being undertaken as part of a normal class, early on the second term. Time allotment (30 minutes) was specified in the written instructions. The topic of the composition -chosen by taking as a point of reference the syllabus of Business English I- was identical to all students, so that all learners would be familiar to the topic.

Once the compositions were gathered, we proceeded with the conversion of all the data into a machine-readable format. Since no hand-writing recognition device was available at the time, the researcher herself had to decipher the handwriting and type all the texts without lemmatisation. Typed texts were saved as files in ASCII format (i.e. text format with line breaks). In addition to this, they were blinded by assigning a code number. Subsequently, the vocabulary in written compositions was analysed by making use WordSmith Tools[1].

[1] Its trait definition entails a tool, which assesses embedded, comprehensive and context dependent vocabulary.

This software package[2] enabled us to create word lists (in both alphabetical and frequency order), retrieve concordance output, and get collocation information; data which will put forward a thorough description of testees' embedded lexis. Nevertheless, we would like to stress the fact that, this software program does not draw conclusions in itself, but it may help teachers and researchers to spot lexical patterns, so that they will able to provide their value judgements on the basis of objective data, in order to know approximately what stage of vocabulary development students are at. Therefore, the quantitative data retrieved through *WordSmith Tools* and the linguist's intuition will be complementary, rather than antagonistic.

4. Results

In this section, we will analyse testees' embedded vocabulary by handling *WordSmith Tools*, so as to carry out an empirical research based on quantifying and classifying our sample of informants' lexis, through our instrument of analysis.

We will aim at describing our sample of undergraduate students' productive vocabulary in English as an L2 by taking into account different lexical measures: (a) length of the written composition -i.e. number of tokens-; (b) lexical variation -i.e.type/token ratio-; (c) frequency of words; and (d) collocability –i.e. the tendency of two or more words to co-occur in discourse).

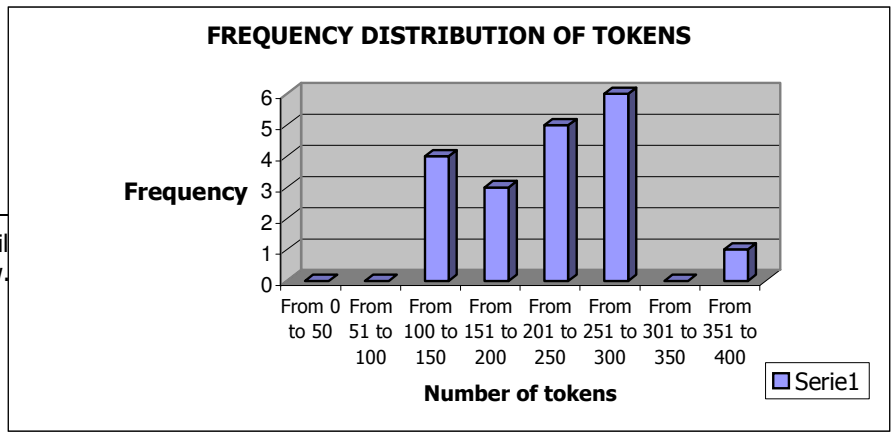
Thus, if we classify the retrieved embedded vocabulary on the basis of the number of tokens put forward by informants in their written compositions, we observe that informants seemed to have produced a great variety of texts according to its length, which ranges from 130 words to 379.

In table 1, we can see the frequency distribution of the number of tokens. If we analyse the results displayed, we can observe that 32% of testees produced from 251 to 300 tokens per composition; this was followed by 26 % of subjects who put forward from 201 to 250 word; 21 % of test takers retrieved the lowest number of tokens which ranged from 100 to 150; 16 % of students recalled from 151 to 200 words; and finally it was only 5 % of informants that was able to put forward the maximum number of tokens (from 351 to 400).

TOKENS	FREQUENCY
From 0 to 50	0
From 51 to 100	0
From 101 to 150	4
From 151 to 200	3
From 201 to 250	5
From 251 to 300	6
From 301 to 350	0
From 351 to 400	1

Table 1. Tokens frequency distribution

The frequency figure (see figure 2) allows us to state that the task was feasible to all students. It represents a negative skewed distribution, which indicates that the task was appropriate to their level, since they were able to recall different vocabulary in order to reach communication.



[2] It is avail web page. http://www.

Figure 2. Tokens frequency distribution

In order to analyse the lexical variation of our sample of written texts, we will identify the number of types put forward by testees. The types range from 80 to 191, which shows a great dispersion of results, since the informant who provides the maximum number of types doubles the figure displayed by the subject who recalls the minimum ones. It should be noted that it is the same test taker, the one that produced the highest figures with regard to types and tokens.

If we analyse the types retrieved by paying attention to its frequency distribution (see table 2), we can see that the great majority of students (74 %) produced between 100 to 150 types; this was followed by 16 % of testees who elicited the minimum number of types (from 51 to 100); and it was only 11% of informants that recalled between 151 to 200 types.

TYPES	FREQUENCY
From 0 to 50	0
From 51 to 100	3
From 101 to 150	14
From 151 to 200	2

Table 2. Frequency distribution of types

In figure 3, we can see that the frequency distribution of types puts forward a normal distribution, since the majority number of types fall at the central point of the scale. Therefore, we can observe that lexical variation across texts seems to be rather stable.

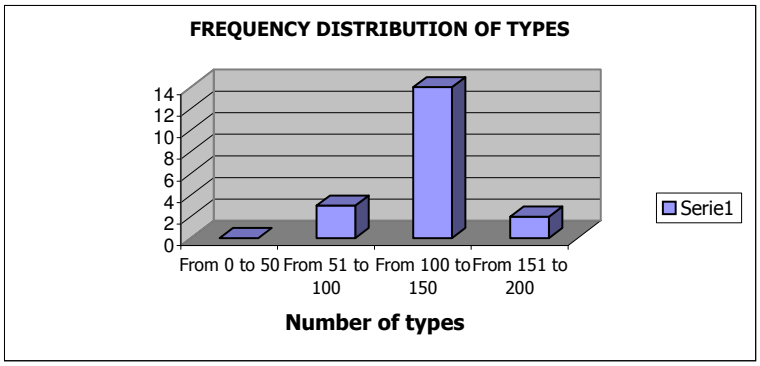


Figure 3. Frequency distribution of types

In table 3, we can see that a lexical variation index between 51 to 60 is shown by 58 % of students; 26 % of informants put forward an index between 61 to 70; and finally, it was only 16 % of test takers that showed a ratio between 41 to 50. It should be noted that, figure 4 displays a normal distribution pattern on the basis of lexical variation.

TYPE/TOKEN RATIO	FREQUENCY
From 41 to 50	3
From 51 to 60	11
From 61 to 70	5

Table 3. Frequency distribution of lexical variation

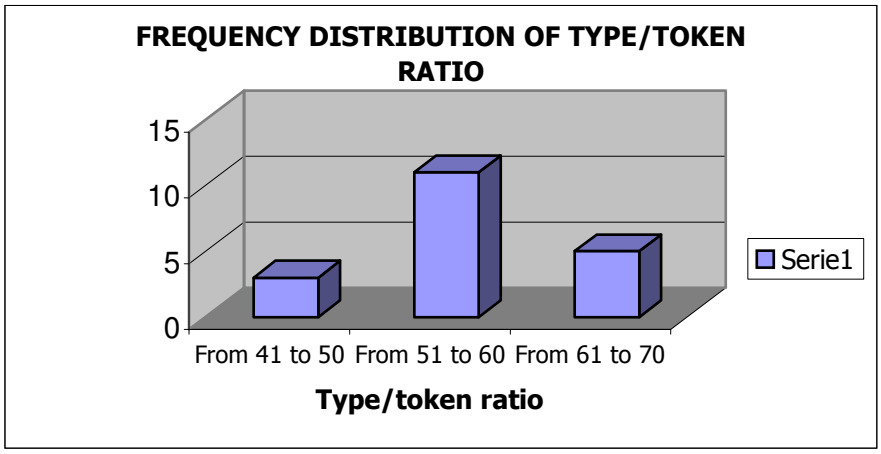


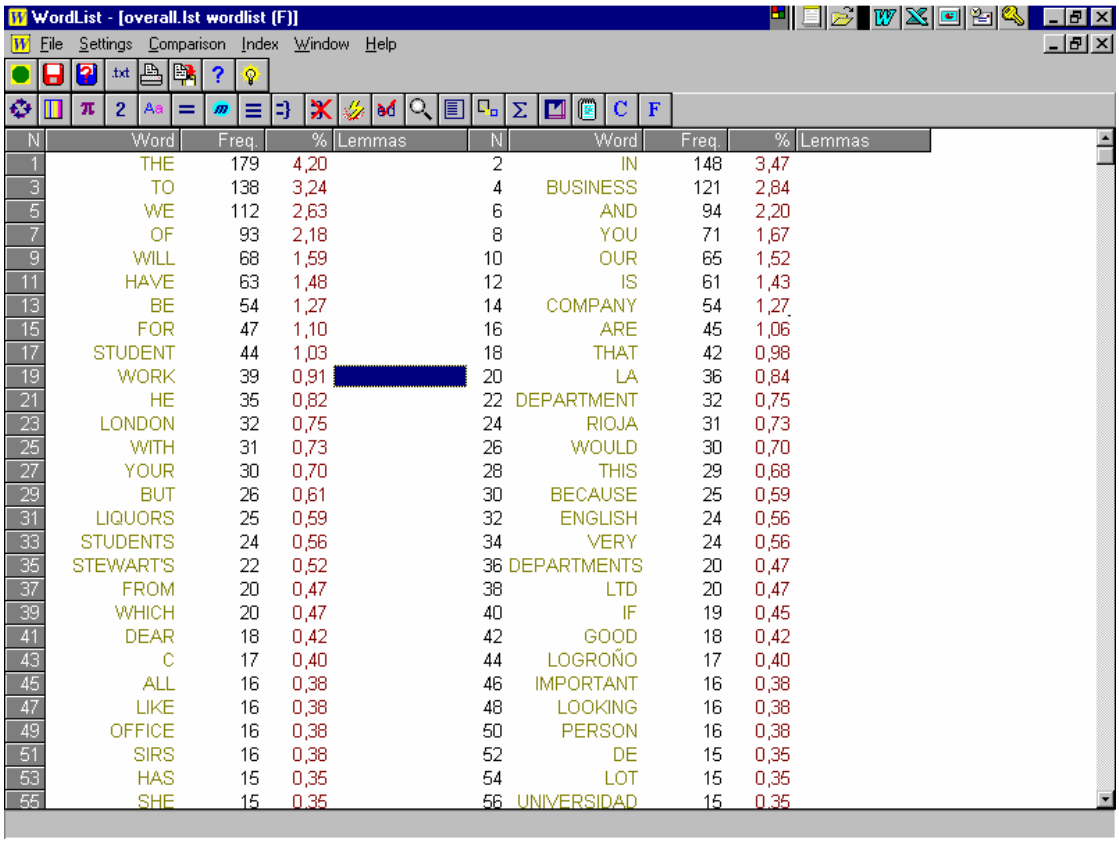
Figure 4. Frequency distribution of type/token ratio

WordSmith Tools also enables us to retrieve the frequency of words elicited by informants. We consider that extracting the whole matrix of words will not be potentially useful. Therefore, we have opted for putting forward an example of the way data is sorted (figure 5). The results obtained can be summarised in the following way:

- The most frequent words retrieved by testees are grammatical words.
- Amongst the most frequent lexical words within the whole set of texts, we can highlight[3]: *business* (2.84 %), *company* (1.27 %), *student* (1.03%), *work* (0.91 %), *department* (0.75 %), *London* (0.75 %), *office* (0.38 %), *Manchester* (0.28 %), job (0.26 %), *branches* (0.23 %), *university* (0.23 %), *marketing* (0.21 %), *pay* (0.21 %), *Spain* (0.21 %), *director* (0.19 %), *Edinburgh* (0.19 %), *learn* (0.19 %), *Liverpool* (0.19 %), *sales* (0.19 %), *country* (0.16 %), *house* (0.16 %), *interested* (0.16 %), *placement* (0.16 %), *cities* (0.14 %), *employees* (0.14 %), *England* (0.14 %), *foreign* (0.14 %), *manager* (0.14 %), *salary* (0.14 %), *Spanish* (0.14 %), *car* (0.12 %), *experience* (0.12 %), *inform* (0.12 %), *production* (0.12 %), *world* (0.12 %), *account* (0.09 %), *bus* (0.09 %), *Europe* (0.09 %), *Glasgow* (0.09 %), *hotel* (0.09 %), *language* (0.09 %), *level* (0.09 %), *transport* (0.09 %), *career* (0.07 %), *export* (0.07 %), *firm* (0.07 %), *friendly* (0.07 %), *future* (0.07 %), *Germany* (0.07 %), *improve* (0.07 %), *opportunity* (0.07 %), *Oxford* (0.07 %), *profile* (0.07 %), *profit* (0.07 %), *resources* (0.07 %), *Scotland* (0.07 %), *airport* (0.05 %), *Britain* (0.05 %), *Cambridge* (0.05 %), *chance* (0.05 %), *computer* (0.05 %), *curriculum* (0.05 %), *customers* (0.05 %), *Denmark* (0.05 %), *development* (0.05 %), *distributed* (0.05 %), *divided* (0.05 %), *earn* (0.05 %), *finance* (0.05 %), *flats* (0.05 %), *France* (0.05 %), *headquarters* (0.05 %), *Italy* (0.05 %), *Japan* (0.05 %), *Leeds* (0.05 %), *Madrid* (0.05 %), *money* (0.05 %), *offer* (0.05 %), *Paris* (0.05 %), *problems* (0.05 %), *quantity* (0.05 %), *research* (0.05 %), *staff* (0.05 %), *vacant* (0.05 %), *wage* (0.05 %).

The lexical words retrieved can be classified into different word fields such as the ones related to the business world, geography, means of transport, income; issue which enables us to claim that testees were able to recall a great variety of vocabulary in the written composition task. We would also like to draw attention to the fact that amongst the different world fields, there are a wide range of words which are related to British geography and also to different world-wide places; we consider this is a relevant feature, since it denotes cultural awareness on the side of informants, a really important element within second and/or foreign language teaching and learning.

[3] It should be noted that verbs such as *to be* and *to have* are highly frequent, however we have considered that it would be more relevant to identify really frequent keywords within *WordSmith*'s output.

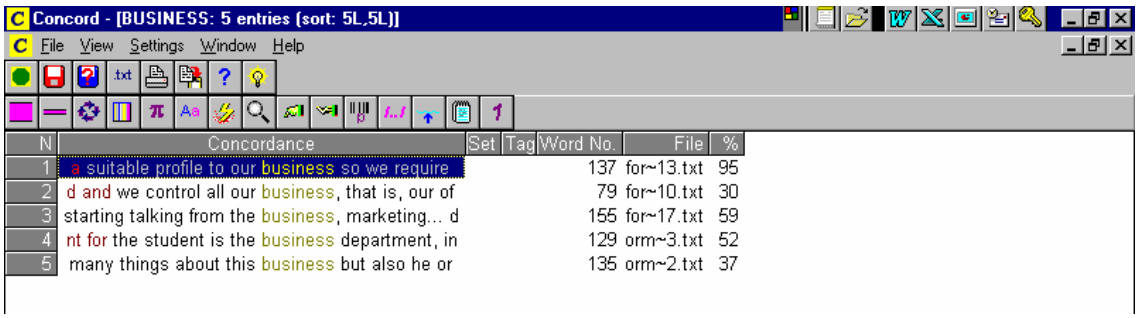


N	Word	Freq.	%	Lemmas	N	Word	Freq.	%	Lemmas
1	THE	179	4,20		2	IN	148	3,47	
3	TO	138	3,24		4	BUSINESS	121	2,84	
5	WE	112	2,63		6	AND	94	2,20	
7	OF	93	2,18		8	YOU	71	1,67	
9	WILL	68	1,59		10	OUR	65	1,52	
11	HAVE	63	1,48		12	IS	61	1,43	
13	BE	54	1,27		14	COMPANY	54	1,27	
15	FOR	47	1,10		16	ARE	45	1,06	
17	STUDENT	44	1,03		18	THAT	42	0,98	
19	WORK	39	0,91		20	LA	36	0,84	
21	HE	35	0,82		22	DEPARTMENT	32	0,75	
23	LONDON	32	0,75		24	RIOJA	31	0,73	
25	WITH	31	0,73		26	WOULD	30	0,70	
27	YOUR	30	0,70		28	THIS	29	0,68	
29	BUT	26	0,61		30	BECAUSE	25	0,59	
31	LIQUORS	25	0,59		32	ENGLISH	24	0,56	
33	STUDENTS	24	0,56		34	VERY	24	0,56	
35	STEWART'S	22	0,52		36	DEPARTMENTS	20	0,47	
37	FROM	20	0,47		38	LTD	20	0,47	
39	WHICH	20	0,47		40	IF	19	0,45	
41	DEAR	18	0,42		42	GOOD	18	0,42	
43	C	17	0,40		44	LOGROÑO	17	0,40	
45	ALL	16	0,38		46	IMPORTANT	16	0,38	
47	LIKE	16	0,38		48	LOOKING	16	0,38	
49	OFFICE	16	0,38		50	PERSON	16	0,38	
51	SIRS	16	0,38		52	DE	15	0,35	
53	HAS	15	0,35		54	LOT	15	0,35	
55	SHE	15	0,35		56	UNIVERSIDAD	15	0,35	

Figure 5. Frequency list retrieved by making use of *WordSmith Tools*

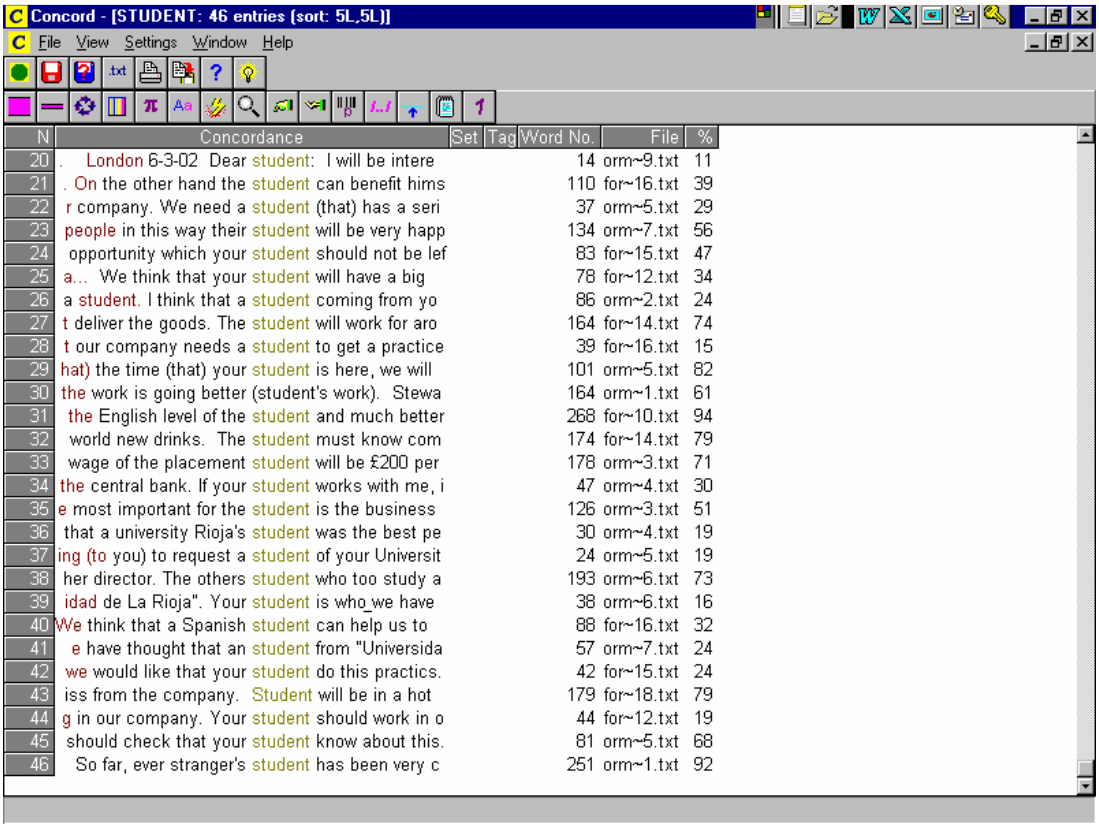
With regard to collocability, we will put forward different collocations that learners seem to retrieve on the basis of their most frequent words. It should be noted that we will not deal with every single collocate students have provided since that will go beyond space constraints. From the different words, we have just chosen three at random, in order to provide an overview of the collocational patterns employed by our sample of informants..

Thus, in figure 6, we can observe the company a word such as *business* keeps, within our sample of texts produced by undergraduate students, which seems to be frequently modified by the indefinite article. On the other hand, testees tend to modify a word such as *student* by making use of: (a) an adjective; (b) an indefinite article; and/or (c) a possessive adjective (see figure 7).



N	Concordance	Set	Tag	Word No.	File	%
1	a suitable profile to our business so we require			137	for~13.txt	95
2	d and we control all our business, that is, our of			79	for~10.txt	30
3	starting talking from the business, marketing... d			155	for~17.txt	59
4	nt for the student is the business department, in			129	orm~3.txt	52
5	many things about this business but also he or			135	orm~2.txt	37

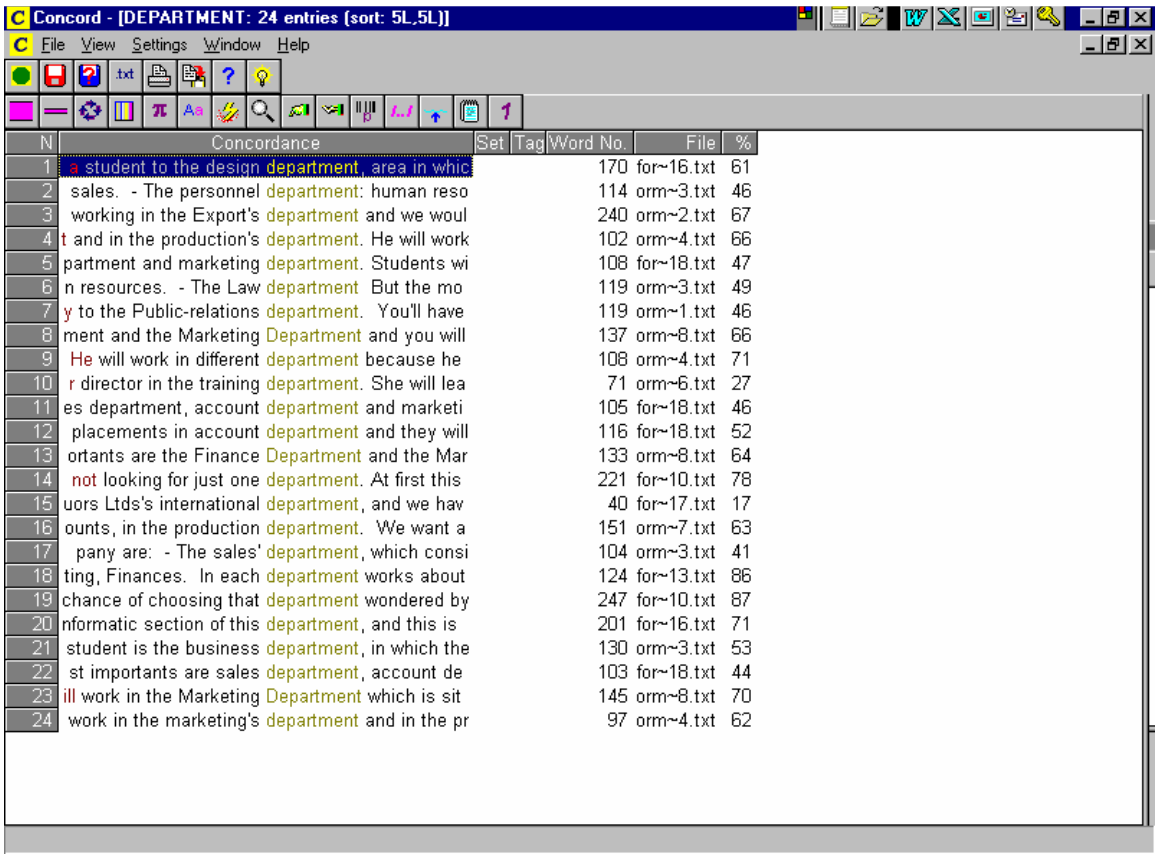
Figure 6. *Business* collocates



N	Concordance	Set	Tag	Word No.	File	%
20	. London 6-3-02 Dear student: I will be intere			14	orm~9.txt	11
21	. On the other hand the student can benefit hims			110	for~16.txt	39
22	r company. We need a student (that) has a seri			37	orm~5.txt	29
23	people in this way their student will be very happ			134	orm~7.txt	56
24	opportunity which your student should not be lef			83	for~15.txt	47
25	a... We think that your student will have a big			78	for~12.txt	34
26	a student. I think that a student coming from yo			86	orm~2.txt	24
27	t deliver the goods. The student will work for aro			164	for~14.txt	74
28	t our company needs a student to get a practice			39	for~16.txt	15
29	hat) the time (that) your student is here, we will			101	orm~5.txt	82
30	the work is going better (student's work). Stewa			164	orm~1.txt	61
31	the English level of the student and much better			268	for~10.txt	94
32	world new drinks. The student must know com			174	for~14.txt	79
33	wage of the placement student will be £200 per			178	orm~3.txt	71
34	the central bank. If your student works with me, i			47	orm~4.txt	30
35	e most important for the student is the business			126	orm~3.txt	51
36	that a university Rioja's student was the best pe			30	orm~4.txt	19
37	ing (to you) to request a student of your Universit			24	orm~5.txt	19
38	her director. The others student who too study a			193	orm~6.txt	73
39	idad de La Rioja". Your student is who we have			38	orm~6.txt	16
40	We think that a Spanish student can help us to			88	for~16.txt	32
41	e have thought that an student from "Universida			57	orm~7.txt	24
42	we would like that your student do this practics.			42	for~15.txt	24
43	iss from the company. Student will be in a hot			179	for~18.txt	79
44	g in our company. Your student should work in o			44	for~12.txt	19
45	should check that your student know about this.			81	orm~5.txt	68
46	So far, ever stranger's student has been very c			251	orm~1.txt	92

Figure 7. *Student* collocates

Finally, we would like to draw attention to figure 8, in which the collocates of *department* are displayed. By having a look at the sorted data, we can see that the great majority of words that modify *department* are nouns, which function as adjectives, in order to describe the different departments.



N	Concordance	Set	Tag	Word No.	File	%
1	a student to the design department, area in whic			170	for~16.txt	61
2	sales. - The personnel department: human reso			114	orm~3.txt	46
3	working in the Export's department and we woul			240	orm~2.txt	67
4	t and in the production's department. He will work			102	orm~4.txt	66
5	partment and marketing department. Students wi			108	for~18.txt	47
6	n resources. - The Law department But the mo			119	orm~3.txt	49
7	y to the Public-relations department. You'll have			119	orm~1.txt	46
8	ment and the Marketing Department and you will			137	orm~8.txt	66
9	He will work in different department because he			108	orm~4.txt	71
10	r director in the training department. She will lea			71	orm~6.txt	27
11	es department, account department and marketi			105	for~18.txt	46
12	placements in account department and they will			116	for~18.txt	52
13	ortants are the Finance Department and the Mar			133	orm~8.txt	64
14	not looking for just one department. At first this			221	for~10.txt	78
15	uors Ltds's international department, and we hav			40	for~17.txt	17
16	ounts, in the production department. We want a			151	orm~7.txt	63
17	pany are: - The sales' department, which consi			104	orm~3.txt	41
18	ting, Finances. In each department works about			124	for~13.txt	86
19	chance of choosing that department wondered by			247	for~10.txt	87
20	nformatic section of this department, and this is			201	for~16.txt	71
21	student is the business department, in which the			130	orm~3.txt	53
22	st importants are sales department, account de			103	for~18.txt	44
23	ill work in the Marketing Department which is sit			145	orm~8.txt	70
24	work in the marketing's department and in the pr			97	orm~4.txt	62

Figure 8. *Department* collocates

Summing up, we can say that we have examine testees' embedded vocabulary in written compositions by making use of *WordSmith Tools*, which has enabled us to classify informants' productive vocabulary on the basis of different lexical measures.

5. Conclusion

On the whole, we would like to state that our main goal has already been achieved by portraying the productive vocabulary of Spanish undergraduate learners of English as an L2 through WordSmith Tools. Thus, we have analysed our sample of informants' embedded productive vocabulary in English as an L2 by dealing with the following lexical measures:

- Length of written compositions: We have identified a great diversity of texts on the basis of the number of tokens; issue which according to scholars such as ARNAUD (1984) may be related to testees' proficiency.
- Lexical variation (i.e. type/token ratio): Despite the great variability of written compositions on the basis of tokens, we have observed that the tokens frequency pattern showed a negative skewed distribution, which indicated that the task was feasible to all students. On the other hand, *WordSmith tools* displayed a normal frequency distribution pattern with regard to the type/token ratio. From our point of view, this stability was due to the fact that the topic of the composition had enabled students to recall the same type of vocabulary, which was based on their knowledge of business English rather than on general English.
- Frequency of words: The quantitative data retrieved has allowed us to infer that the most frequent words were grammatical. On the other hand, we have been able to give account of a great range of words put forward by testees in their compositions, which were

gathered around different word field such as the business word, geography and means of transport, amongst others.

- Collocability: In our quantitative analysis, we have also displayed the collocations of some lexical words, chosen randomly amongst the more frequent testees have used. Thus, we were able to see that their modifiers were mainly indefinite articles, possessive adjectives and nouns functioning as adjectives in order to determine the appropriate meaning of words.

Thus, by drawing attention to the lexical measures already identified, we have been able to describe the undergraduate learners of English as L2's productive lexical profile.

Through this descriptive empirical research, we can conclude that despite our informants seem to be at different vocabulary development stages, they share a common knowledge of English for Specific Purposes, which is influenced by the general English ability. Nevertheless, they are able to recall a great range of vocabulary from different word fields in order communicate, apart from showing some kind of cultural awareness, one of the main factors to avoid misunderstanding within communication.

We consider that more research within this field should be done involving: (a) a larger sample of informants; (b) students enrolled in different courses, so as to outline the lexical competence of students across different levels.

Furthermore, extensive research should be conducted on the basis of electronic tools such as WordSmith Tools. We believe that it can be a useful instrument in order to complement teachers' judgement with regard to the assessment of written compositions on the basis of objective data.

Bibliographic references

ALTENBERG, BENGT & GRANGER, SILVIANE, "The Grammatical and Lexical Patterning of Make in Native and Non-native Student Writing", *Applied Linguistics*, 22, 2, 2001, 173-194.

ARNAUD, PIERRE J. L. , "The lexical richness of L2 written production and the validity of vocabulary tests", Occasional Papers, University of Essex, Department of Language and Linguistics, 29, 1984, 14-28.

COBB, TOM, "An Introduction to Learner Corpus Analysis", Paper presented at AAAL 2000, Vancouber. It can be retrieved from the following Internet URL http://www.er.uqam.ca/nobel/r21270/cv/lc.htm

LI, YI, "Assessing Second Language Writing: The Relationship between Computerized Analysis and Rater Evaluation", *Assessing Second Language Writing*, 127-128, 2000, 37-51.

NAVES NOGUÉS, Mª TERESA, "To what extent do computerised POS tagging text-analysis programs correlate with manually calculated analytical writing measures", Paper presented at A.E.D.E.A.N., Granada, 13-15 December 2001.

SAGASTA ERRASTI, Mª PILAR, *La producción escrita en euskara, castellano e inglés en el modelo D y en el modelo de inmersión*, Tesis doctoral Universidad del País Vasco, 2000.

SCOTT, MIKE, *WordSmith Tools*, Oxford, Oxford University Press, 1997.

WEBER, JEAN-JACQUES, "A concordance- and genre- informed approach to ESP essay writing", *ELT Journal*, 55, 1, 2001.