VOCABULARY PROFILES IN ENGLISH AS A FOREIGN LANGUAGE AT THE END OF SPANISH PRIMARY AND SECONDARY EDUCATION:

PERFILES DE VOCABULARIO EN INGLÉS COMO LENGUA EXTRANJERA AL FINAL DE EDUCACIÓN PRIMARIA Y SECUNDARIA EN ESPAÑA

ROSA Mª JIMÉNEZ CATALÁN
University of La Rioja, Spain
rosa.jimenez@unirioja.es

ABSTRACT

In this study we identify the number and the characteristics of the words used in letters by Primary and Secondary school EFL learners at the end of Spanish compulsory educational stages. The compositions were submitted to WordSmith Tools (Version 6.0) so as to classify the words according to alphabetic order, word frequency, statistical lists and collocation patterns in each of the courses under examination. We completed this quantitative analysis by means of a qualitative analysis of the top 50 word types for each course. Results revealed an increase in the means of words used in each educational stage, although the increase was lower than expected and practically based on substantives rather than on verbs, adjectives or adverbs. From a qualitatively perspective, we found similarities in the vocabulary used in the two courses. Nevertheless, we also observed differences that suggest the existence of idiosyncratic features at the end of each educational stage.

Keywords: Vocabulary, Learner Corpus, EFL learners, Primary School Education, Secondary School Education.

RESUMEN

En este estudio identificamos el número y características de las palabras que utilizan los aprendices de inglés como lengua extranjera en la carta como composición escrita al final.

“This study is part of the Research Project “Factores individuales y contextuales en la adquisición y desarrollo de la competencia léxica en inglés como lengua extranjera” funded by the Spanish Ministry of Science and Innovation (FFI 2010-19334).
de las etapas educativas de Educación Primaria y Secundaria Obligatoria en España. Las composiciones fueron analizadas mediante el programa WordSmith Tools (Versión 6.0) con objeto de obtener los listados alfabéticos y de frecuencias, estadística descriptiva y patrones de colocaciones de las palabras utilizadas en cada uno de los dos cursos analizados. Completamos este análisis cuantitativo mediante un análisis de corte cualitativo de las 50 palabras (tipos) más frecuentes en cada curso. Los resultados mostraron un incremento en la media de palabras que se utilizan al final de la etapa de secundaria, si bien dicho incremento resultó ser más bajo de lo esperado y fundamentalmente centrado en substantivos más que en verbos, adjetivos o adverbios. Desde el punto de vista cualitativo encontramos similitudes entre los dos cursos fin de etapa. Sin embargo, también se observaron diferencias que parecen señalar la existencia de rasgos idiosincráticos en cada curso final de etapa educativa.

**Palabras clave**: Vocabulario, Learner corpus, Aprendices de inglés como lengua extranjera, Educación Primaria, Educación Secundaria.

**Recibido**: 12.05.2015. **Aceptado**: 28.10.2015.

### 1. INTRODUCTION

This study looks at the words used in compositions by Spanish students who are learners of English as a foreign language (EFL) at 6th and 10th form. In Spain, each course stands for the end of an educational stage, primary education (age 12) and compulsory secondary education (age 16) respectively. In turn, each course comprises two age groups to which, for the sake of convenience, we will refer to as children and adolescents.

The assembling of corpora from EFL learners at 6th and 10th form is necessary in order to be informed about the language used by children and adolescents in English written tasks at the end of educational stages. It is also a valuable resource to investigate how vocabulary develops throughout the courses, the words learners resort to in specific writing tasks and the characteristics of the vocabulary used by EFL learners. Consequently, it provides researchers, teachers and course designers with a rich source of information as well as with an objective instrument to identify foreign language learners’ needs (Granger, 1998). However, to our knowledge, no study has undertaken the comparative description of the words used in compositions by EFL learners at the end of primary and secondary school education. The present study is a first step to narrow this gap. By means of a corpus-driven study we aim at providing information on the following research questions:
RQ1: How many words do Spanish EFL learners use in a letter-writing task in 6th form and how many in 10th form?
RQ2: Does the number of word types and word tokens increase as age/course increases?
RQ3: Are there similarities or dissimilarities in the characteristics of the words used by children and adolescent EFL learners?

The remaining part of this paper is structured as follows: in the first place, we review research on lexical issues within learner corpus and L2 young learners’ vocabulary studies. Then, we move on to explain the methodology adopted in this study, and finally, report the results obtained by the two groups.

2.A. BACKGROUND

2.A.1. Vocabulary in learner corpus and vocabulary studies

At present, it is possible to find up to ten learner corpora aimed at different purposes (Pravec, 2002). Among them, it is important to highlight the pioneering “International Corpus of Learner English (ICLE)” (Granger, 1993). This comprises a computerized textual database aimed at comparing the inter-languages of language learners of fourteen mother tongues. Within ICLE, there are well-known studies related to lexis such as the frequency analyses applied to the words used by EFL learner natives of French (Granger, 1998), Swedish (Källkvist, 1993), Norwegian (Hasselgren, 1994), and EFL learners of seven European mother tongues and English native speakers (Ringbom, 1998).

Outside ICLE and as far as learner corpora and vocabulary are concerned, we find the study conducted by Shirato and Stapleton (2007) on Japanese EFL learners. These researchers compared the words used in a collection of oral texts by Japanese EFL learners (43,651 words) with oral texts produced by native speakers. Their results indicated differences between the two groups regarding the use of adjectives, discourse markers, modal verbs and word strategies. Likewise, their results revealed EFL learners’ overuse of adjectives and auxiliary verbs, in comparison with native speakers. In the same line, Liu and Shaw (2001) focuses on the frequency analysis of the use of ‘make’ in written essays by native and non-native speakers, the latter group made up of Chinese university students, learners of English as a foreign language at the upper intermediate level. Their results show EFL learners’ overuse of ‘make’ compared to English native speakers.

Within the Spanish context, several studies have examined learners’ vocabulary in written compositions. Jiménez Catalán (1992) looked at word production and lexical errors in descriptive, narrative and argumentative essays by 350
EFL secondary education students from three schools located in three Spanish cities. This research provided evidence of the existence of variation in learners’ performance according to individual factors such as motivation and gender, and instructional factors as the type of composition. More recently, other studies have corroborated the same pattern of variation in texts written by primary school children (Jiménez Catalán and Ojeda Alba, 2008; Agustín Llach and Jiménez Catalán, 2007; Ojeda Alba and Jiménez Catalán, 2007; Jiménez Catalán, Ruiz de Zarobe and Cenoz, 2006) as well as texts by EFL learners in secondary schools (Moreno Espinosa, 2011; Jiménez Catalán and Ojeda Alba, 2010; Ojeda Alba and Jiménez Catalán, 2010).

Overall, the findings point to an increase in the number of tokens and types as age and course increases. Likewise, they point to an increase in the use of other word classes that are different to nouns as well as to a decrease in spelling errors as age and course increases. Regarding the effect of individual and instructional variables on learners’ lexical performance, gender differences have been observed in favor of female students, who, on average, write longer compositions and make use of a richer vocabulary than male students. Females also make fewer errors in compositions than males. However, the mechanisms of lexical creation seem to be similar in both sexes, as no differences were found in the type of lexical errors committed by the two groups. As to the effect of instructional variables, the most relevant were the type of test as well as the topic of composition on learners’ lexical performance. Likewise, the effect of the type of instruction was confirmed on the use of lexical cohesion ties and lexical phrases: EFL learners in a Content and Language Integrated Learning (CLIL) program used a significant higher number of those features than EFL learners in a regular program. However, no comparative studies have been conducted on the vocabulary characteristics of EFL learners at the end of compulsory educational stages.

The research on Spanish children and adolescents reported in this article fills a gap in learner corpus and in vocabulary studies as there is hardly any research on EFL learners at primary or secondary education. In order to draw sound conclusions, we need to conduct longitudinal analyses as well as cross-sectional comparisons on the vocabulary performance of learners at different ages and courses. Particularly important for establishing comparisons with EFL learners in other countries is the compilation and description of the vocabulary used in compositions by EFL learners in Spanish speaking countries at the end of educational stages.

2.B. METHODOLOGY

The informants in the present study were 184 EFL learners equally distributed into two groups: 6th form at primary education and 10th form at secondary education.
In the case of the primary school EFL learners, the sample was drawn at random out of a larger sample of over 300 students from our longitudinal database. They belong to four primary schools located in middle class areas of a city in the north of Spain. As to the 10th form sub-sample, it comprises three instant classes from a secondary school in the same city and area as to the primary school sample.

The study design is cross-sectional as we compare the vocabulary profiles of two groups of learners of two educational stages. Both make up a homogeneous sample concerning the social background and English language teaching approach: the two groups belong to schools located in middle class areas of similar social characteristics and both were learning English as a subject on the basis of three to four hours a week after a communicative approach, in which emphasis was put on the use of English for communication.

Students were required to write a letter to a fictitious host English family in Oxford (with whom they were supposed to spend a month in order to learn English). They were asked to introduce themselves, speak about their school, city, hobbies, family and relations, and daily routines. Instructions were given in Spanish so as to facilitate total comprehension of the purpose of the letter. The time given to fulfil the task was 30 minutes, and no dictionaries or other aids were allowed.

At the moment of data collection, 6th primary school EFL learners had received approximately 629 hours of tuition in English; on its part, 10th secondary school EFL learners had received 1,149 hours. The two collection times were done in different years (2006 in primary schools and 2011 in secondary schools); however, the data gathering conditions such as administration and the amount of time allotted to the fulfilment of the task were kept exactly the same.

The compositions were collected, read, and the scripts keyed into electronic files; each entry was coded for the student’s identification: number, sex, school and course. In addition, a background questionnaire was administered to each learner as to gather information regarding knowledge of other languages, nationality, marks gained in English in previous courses, motivation and effort towards English language, among other variables.

Each composition was edited according to the following protocols: spelling mistakes were corrected, Spanish words and unintelligible words were deleted, words which are read the same in English and Spanish (as for instance, “Africa” or “America”) were kept if the context provided evidence of learners’ knowledge of other related names as in the sentence, “Africa and Europe are continents”. However, they were discarded if the text suggested that learners could be transferring words from Spanish. Likewise, proper names given in the instructions such as Edwards, Peter and Helen, were deleted. Finally, well-known English brand names such as “PlayStation” or “Burger King” were also deleted from the transcription. Our reason was that rather than showing evidence of vocabulary knowledge, they
showed evidence of shared common cultural knowledge in English and Spanish, which, although certainly of great interest, it is not under the scope of the present study.

Each composition was submitted to quantitative analyses by means of WordSmith Tools (Version 6.0). This program provides researchers with different kind of analyses, among them, the arrangement of words into alphabetical and frequency lists, the provision of a statistical report on the number of tokens and types in the texts, as well as word length, collocation patterns and keywords. Due to space limits, this study focuses exclusively on the number of types and tokens and word frequency. We complement this descriptive analysis by means of a manual classification of the words into grammatical and content words, and these in turn, into word classes such as nouns, adjectives, lexical verbs and adverbs.

3. RESULTS

3.1. Word production and vocabulary growth

Our corpus consists of 27,554 word tokens. Out of this figure, 11,887 were elicited from 6th form’s compositions and 15,667 from 10th forms’. Table I shows the means, maximum and minimum of tokens and types per composition obtained by each group.

<table>
<thead>
<tr>
<th></th>
<th>Tokens</th>
<th></th>
<th>Types</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Mean</td>
<td>Max</td>
<td>Min</td>
<td>Mean</td>
</tr>
<tr>
<td>6th Form</td>
<td>128.44</td>
<td>338</td>
<td>133</td>
<td>60.98</td>
</tr>
<tr>
<td>10th Form</td>
<td>170.29</td>
<td>361</td>
<td>164</td>
<td>85.57</td>
</tr>
</tbody>
</table>

If we compare the results attained by each group, we observe clear differences in favour of the older learners. A close inspection of the figures reveals an increase in the overall results of 10th form compared to 6th form, but also an increase of 41.85 tokens and 24.59 types per composition. This indicates that, on average, the compositions by 10th EFL learners are longer than the ones by the 6th EFL learners; they also contain a higher degree of lexical richness as shown by the greater number of word types or different words per composition.
3.2. Characteristics of the words

In this section we aim at ascertain whether there would be differences or similarities in the words used by 6th and 10th EFL learners. In order to answer this question, we will look at two aspects of their vocabulary profiles: word frequency and word class.

3.2.1. Word frequency

Following Richards and Schmidt (2002: 213), we understand the term frequency as “the number of occurrences of a linguistic item in a text or corpus”. As explained in the methodology section, WordSmith Tools were used to obtain the alphabetical and frequency lists of word types (different words) for 6th and 10th EFL learners. The results are described in the following paragraphs.

As mentioned earlier, the total number of word tokens for the whole corpus of edited compositions was 27,554. Out of this figure, the number of types or different words identified in the whole alphabetical and frequency lists was 791 for 6th and 1,053 for 10th EFL learners. These figures comprise the total of different words contained in each corpus and give us a clue about lexical richness in each group. However, it is important to remark that a higher proportion of word types occurred less than 10 times and that 300 word types in 6th form and 423 in 10th form occur only once. In order to warrant representativeness of the words used in texts, it is usual to focus on the most frequent words appearing in them. Following the convention, Tables II y III include the top 50 word types for each group.

A close look at this table reveals the existence of high similarities in the words most frequently used by 6th and 10th EFL learners. Not only do we observe a higher proportion of shared words by both groups (34 out of 50) but also the existence of similarities in the positions of the shared words in the two frequency lists.

Table II. The 50 most frequent words in 6th EFL learners’ compositions.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>%</th>
<th>Rank</th>
<th>Word</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>my</td>
<td>8.26</td>
<td>26</td>
<td>football</td>
<td>0.7</td>
</tr>
<tr>
<td>2</td>
<td>I</td>
<td>6.66</td>
<td>27</td>
<td>old</td>
<td>0.70</td>
</tr>
<tr>
<td>3</td>
<td>is</td>
<td>6.71</td>
<td>28</td>
<td>big</td>
<td>0.66</td>
</tr>
<tr>
<td>4</td>
<td>and</td>
<td>6.06</td>
<td>29</td>
<td>hello</td>
<td>0.64</td>
</tr>
<tr>
<td>5</td>
<td>the</td>
<td>3.23</td>
<td>30</td>
<td>but</td>
<td>0.60</td>
</tr>
<tr>
<td>6</td>
<td>in</td>
<td>2.70</td>
<td>31</td>
<td>brother</td>
<td>0.42</td>
</tr>
</tbody>
</table>
Table III. The 50 most frequent words in 10th EFL learners’ compositions.

<table>
<thead>
<tr>
<th>Rank</th>
<th>Word</th>
<th>%</th>
<th>Rank</th>
<th>Word</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>I</td>
<td>7.71</td>
<td>26</td>
<td>friend</td>
<td>0.06</td>
</tr>
<tr>
<td>2</td>
<td>and</td>
<td>4.53</td>
<td>27</td>
<td>for</td>
<td>0.64</td>
</tr>
<tr>
<td>3</td>
<td>my</td>
<td>3.80</td>
<td>28</td>
<td>name</td>
<td>0.62</td>
</tr>
<tr>
<td>4</td>
<td>is</td>
<td>3.64</td>
<td>29</td>
<td>not</td>
<td>0.64</td>
</tr>
<tr>
<td>5</td>
<td>in</td>
<td>3.43</td>
<td>30</td>
<td>old</td>
<td>0.62</td>
</tr>
<tr>
<td>6</td>
<td>the</td>
<td>3.38</td>
<td>31</td>
<td>do</td>
<td>0.56</td>
</tr>
<tr>
<td>7</td>
<td>a</td>
<td>2.96</td>
<td>32</td>
<td>big</td>
<td>0.58</td>
</tr>
<tr>
<td>8</td>
<td>to</td>
<td>2.45</td>
<td>33</td>
<td>We</td>
<td>0.58</td>
</tr>
<tr>
<td>9</td>
<td>am</td>
<td>1.66</td>
<td>34</td>
<td>love</td>
<td>0.51</td>
</tr>
<tr>
<td>10</td>
<td>very</td>
<td>1.53</td>
<td>35</td>
<td>your</td>
<td>0.40</td>
</tr>
<tr>
<td>11</td>
<td>like</td>
<td>1.26</td>
<td>36</td>
<td>village</td>
<td>0.46</td>
</tr>
</tbody>
</table>
However, we also note some remarkable differences on two regards: the disparity of ranking positions of some of those shared words and the presence of exclusive words to each group. Table IV shows the main differences between the two groups.

**Table IV. Main differences in the 50 top words between 10th Form (Adolescents) and 6th Form (Children).**

<table>
<thead>
<tr>
<th>Position in shared words</th>
<th>10th Form (Adolescents)</th>
<th>6th Form (Children)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Exclusive words</td>
<td>we, your, village, me, cities, music, that, study, going, because, sport, Mr, Mrs, lot, this, for.</td>
<td>Hello, brother, house, beautiful, teacher, family, sister, two, one, class, mother, favorite, father, basketball, dog, good.</td>
</tr>
<tr>
<td>Positions in shared words</td>
<td>got (37th), play (25th), am (9th), friend (26th), football (38th), you (18th), love (34th), with (17th), there (24th), it (15th).</td>
<td>got (15th), play (17th), am (18th), friend (19th), football (26th), you (40th), love (44th), with (25th), there (49th), it (50th).</td>
</tr>
</tbody>
</table>

3.2.2. **Word class**

The 50 most frequently used words by each group consist of 31 content words and 19 grammatical words for 10th EFL learners compared to 39 content words.
and 11 grammatical words for 6th EFL learners. As shown in Table V, nouns make up for the category most frequently used by the two groups, followed by lexical verbs, adjectives and adverbs.

**Table V. Distribution of top 50 content words according to word class.**

<table>
<thead>
<tr>
<th></th>
<th>Nouns</th>
<th>Adjectives</th>
<th>Verbs</th>
<th>Adverbs</th>
</tr>
</thead>
<tbody>
<tr>
<td>10th Form</td>
<td>13</td>
<td>2</td>
<td>8</td>
<td>2</td>
</tr>
<tr>
<td>6th Form</td>
<td>19</td>
<td>5</td>
<td>7</td>
<td>3</td>
</tr>
</tbody>
</table>

4. DISCUSSION

Most studies on the effect of the age factor on second/foreign language acquisition report the supremacy of older learners over young learners. Their findings are justified on the basis of a higher cognitive maturity and wider world experience on the part of older learners. The results obtained in the present study point to the advantage of 10th form (adolescents) over 6th form (children), at least as far as means of word tokens and word types in compositions are concerned. However, we could argue whether the older learner’s advantage is due to the higher number of hours of instruction received by 10th EFL learners (454 hours more) compared to 6th EFL learners. This hypothesis is backed up by research on the effect of exposure, which claims significant differences in favor of those learners who received a greater number of hours of instruction (Turnbull, Lapkin, Hart and Swain, 1998; MacLaughlin, Osterhout and Kim, 2004).

The present study points to similarities and differences in learners’ vocabulary performance. Similarities were found in the great number of shared words (37 out of 50) among the top 50 words as well as in the similar distribution of content words for each group: nouns predominate over verbs, adjectives and adverbs; and regarding the other word categories, there is an increase of only one type in the lexical verbs category; on the contrary, adjectives and adverbs decreased in 10th EFL learners’ production. This finding does not allow us to claim the existence of growth in verbs, adjectives and adverbs as educational stage increases; at least as far as the 50 most frequent words are concerned. We need to look at the overall frequency lists in order to testify some growth in this regard. If we expand our analysis to the 100 most frequent words in each group, we obtain the following results for 10th EFL learners: 32 grammatical words and 68 content words, which classified into word categories yields 30 nouns, 20 verbs, 11 adjectives and 7 adverbs. In comparison, 6th EFL learners as a group produced a total of 25 grammatical words and 75 content words, out of which, 44 were nouns, 14 verbs,
11 adjectives and 5 adverbs. It is clear then, that nouns decrease but verbs and adverbs grow as educational stage increases. Although this pattern seems to be straightforward, care should be taken in the interpretation of these findings as, on the one hand, no tests for significance were applied to the data, and on the other, the number of occurrences for each type is considerably lower in the top 100 than in the top 50 in the frequency lists.

The present study corroborates the findings obtained by Myers-Pease, Berko-Gleason and Alexander-Pan (1993) on English as a mother tongue (L1). They report the predominance of nouns over other categories in early stages of language acquisition. This tendency was also observed by Justicia (1985) in Spanish as L1. His longitudinal study on the vocabulary acquisition and development of Spanish primary school children proved that nouns surpass other categories.

Nouns also prevail over other word classes in most studies with similar samples of EFL learners. The evidence observed in vocabulary production in English and Spanish as L1 and L2 points to universal tendencies in vocabulary acquisition. The fact that 6th and 10th EFL learners resort to nouns rather than to other word categories suggests that both groups are still in the first stages of the acquisition of English; likewise, it also seems to suggest that world categorization throughout nouns also occurs at the first stages of foreign language learning.

Regarding word frequency, two aspects are worth highlighting here. Firstly, the great resemblance of the lists in Tables II and III to the word frequencies found in Collins Cobuild Student’s Dictionary and Grammar (1994) where the 10 most frequent words are grammatical words: the, of, and, to, a, in, that, I, it, was (in descending order). As can be observed, the grammatical words shared by 6th and 10th EFL learners together with positions in the ranking are practically identical to the ones appearing in the above corpus.

Secondly, the worlds projected by the shared content words are very similar in younger learners and adolescents: in both groups we find the personal pronoun I, the verb to be (in its different realizations), and the possessive my and the word name in top positions. These words may be used by learners to express their identities. The presence of my, I, and like among the first positions in the frequency lists of both groups suggests a total correspondence between age and vocabulary; in this study, the informants are young learners and adolescents in the process of constructing their identities as human beings as well as English language learners. Although of different age, both groups share an egocentric view of the world; the first person pronoun I prevails over we or you, which represent the others.

Nevertheless, subtle differences also appear in the exclusive words used by each group: brother, house, teacher, family, sister, class, mother, father, or dog are found within the children’s top 50 but not in adolescents’, where we find words such as village, cities, music, study, or sport. These non-shared words speak of different mental worlds: family represented by their nuclear members (mother, father,
brother and sister, dog included), as well as the school represented by the teacher seems to conform to the universe of the younger group; in comparison, the exclusive words found in the adolescent’s group seem to go beyond the confines of the nuclear family to the world of the village, the city, music and sport: words and worlds that seem to be part of the identity of adolescents regardless of whether they express themselves in their mother tongue or in a foreign language.

5. CONCLUSION

In this study, we looked at the words used in compositions by a group of 6th primary school children and a group of adolescents at 10th form, each course end of Spanish primary and compulsory education. We aimed at finding out the number and the type of words used by each group in an identical writing task. According to expectations, 10th EFL learners’ compositions were longer and lexically richer than 6th EFL learners’. This result is not surprising but it is disappointing if we consider that the older group had received 450 hours more of instruction than the younger group, and that the resulting increase was only 41.85 tokens and 24.59 types per composition. Further research is needed to ascertain whether the better results obtained by the older group are due to the effect of the higher number of hours of instruction; if this is proved so, then it would be necessary to investigate whether the letter as a data gathering task has a top ceiling effect on learners’ vocabulary. If learners could write longer and richer compositions in 30 minutes, we might wonder whether the results obtained by 10th EFL learners at the end of compulsory education were as good as they could be.

The quantitative analysis of the characteristics of the word types used by 6th and 10th EFL learners showed more similarities than differences. This result may be due to the low degree of lexical richness in the two groups but also to task and time effect. Further research is required in this regard too. We know by research that genre and type of composition affect learners’ performance (Jiménez Catalán, 2010), but we do not know whether this effect is constant in different ages, courses and educational stages.

Finally, research is also needed on the exclusive vocabulary used by children and adolescents, and their relationship to age and identity. The preliminary exploration conducted in this study has pointed to the existence of word differences but also to different worlds in children and adolescents that might provide a starting point in the analysis of issues related to vocabulary and identity in foreign language learners.
ACKNOWLEDGEMENT

We would like to acknowledge the learners and teachers who participated in this study. We also acknowledge the proofread our manuscript by Ian Patterson.

REFERENCES


