THE INFLUENCE OF AGE ON VOCABULARY ACQUISITION IN ENGLISH AS A FOREIGN LANGUAGE

Tesi doctoral presentada per

Immaculada Miralpeix Pujol

com a requeriment per a l'obtenció del títol de

Doctora en Filologia Anglesa

Programa de Doctorat: *Lingüística Aplicada*
(Bienni 2000-2002)
Departament de Filologia Anglesa i Alemanya

Directors: **Dra. Carme Muñoz Lahoz i Dr. Paul M. Meara**

Universitat de Barcelona

2008

CHAPTER 8

DISCUSSION AND CONCLUSIONS

8.1. Introduction

The results of the study carried out will be discussed taking as a point of departure the four main research questions announced in chapter 4. The present chapter is divided into three blocks and a conclusion section. The first discusses the issue of productive vocabulary in relation to age and exposure, the second section deals with the use of particular intrinsic and extrinsic vocabulary measures in the light of the results obtained. Finally, the last part focusses on vocabulary size estimations. Possible explanations of the results as well as limitations of the study are presented and questions for further research are proposed.

8.2. Age, exposure and vocabulary acquisition

This section discusses the results obtained regarding these two research questions, which have to do with productive vocabulary, age and exposure in a formal setting:

- After having received the same amount of exposure, who will have richer productive vocabularies: a group of ES who started learning English at 8 (Grade 3) or a group of LS who started at 11 (Grade 6)?
- Will ES with more exposure than LS have better productive vocabularies at the end of secondary education (Grade 12)?

Although vocabulary has been usually neglected in age studies, it is a domain that cannot be left aside. If age has a role in the process of acquiring a SL, age-effects should be investigated on each language component, not just with the objective to find a ‘Critical Period’ for language acquisition or for different language constituents, but also with the aim to explore which is the most adequate starting time to learn the language in a formal context or the amount of hours needed for students to have a good command of the language. As has already been pointed out at the beginning of the present dissertation, the issue of age is intrinsically related to a question of time, which can be understood not only as starting time but also as a condition to learn the language. This sort of research is pertinent nowadays in the light of the changes introduced by different European governments by which the AO to start the FL in formal settings has been brought down.

Furthermore, although we are nowadays far away from the notion that learning a language means learning a collection of words, the lexical component is still thought to be core in the process of language acquisition. That is why vocabulary should also be included as object of study in research on age effects and language acquisition. Up to now, most studies have showed no concern with vocabulary and many of those that have taken it into account present several problems, such as a very lax control of exposure or the mixing of groups who have started the FL at different AOs.

Results from this study, which has tried to eradicate these flaws, indicate that an early AO in formal contexts does not systematically entail having a richer oral or written productive vocabulary (either free or controlled), as measured by different tasks, in the long run. It is difficult to define and measure ultimate attainment in instructional settings, as it entails longer periods of time than in natural settings: small amounts of time are stretched over several years in school instruction as opposed to the larger amounts of time in natural immersion contexts. We took 'long run' as the end of secondary education, after several years of formal instruction (7 for B3, 8-9 for A3 and 10 for A4), which is 'the longest run' that can be found in our context as regards compulsory education.

Therefore, as regards the long-term comparison, it has been found that, given the same amount of exposure (726 hours), LS tend to be better than ES. That is, if groups A3-B3 are compared, when significant differences emerge, they are normally in favour of B3. These results would be in the same line as those obtained in other instructional settings which pointed out that ES did not surpass LS (Burstall et al., 1974; Cenoz, 2002; Griffin, 1993; Oller & Nagato, 1974; Singleton, 1999). What is more, if in conditions of equal exposure LS tended to be better, ES did not surpass LS either even after the ES group had received one extra year of formal exposure to the language.

As introduced in chapter 6, these results could also be accounted for following Dekeyser (2000), that is, an advantage for LS could be justified in terms of the relationship between age and formal (explicit) learning as opposed to naturalistic (implicit) language acquisition. According to this interpretation, young learners would have an advantage only if the language is learned implicitly in a naturalistic setting, but not if use of explicit learning mechanisms is needed, as it is in school, which then favours the older learners.

Studies like Jia and Aaronson (2003) have showed that children's better vocabulary proficiency in the L2 can be a consequence of a learning experience that is available to them only due to their less developed L1 proficiency. That is, when children acquire the L1, there is a direct mapping between words and concepts. If the L2 is introduced early in life, children learn new words for new concepts too, similar to what happens in L1 acquisition. That is, direct mapping also in the L2 allows them to acquire lexical items efficiently. However, if our learners are introduced to English at the age of 8, although their L1 is not as proficient as LS' L1, a huge part of the vocabulary of the L1 has already been acquired, which means that there will be little room for direct mapping. This might be one of the reasons why the supposed advantage for ES to assimilate new words 'like sponges' is not shown, neither in the short nor in the long-terms in this study either.

The present results are different from the ones obtained in natural contexts where some initial potential advantages that LS might have is progressively lost in favour of ES, as Snow (1983) and Snow and Hoefnagel-Höhle (1978) have proved: although adolescents were better in storytelling at the beginning of the year in a natural immersion setting, older children's eventual attainment by the end of the year was superior to that of their peers. However, it must also be taken into account that even in a natural context, there was a test (the Peabody Picture Vocabulary Test) where adolescents still kept their advantage by the time ES had outperformed them in other areas.

The results of the analysis of long-term achievement were complemented in the present study with those obtained in the short term (Times 1 and 2), both longitudinally and cross-sectionally. The longitudinal study revealed an expected significant effect for time and significant group effects in the storytelling, the composition and the cloze were found to be in favour of the LS. In the cross-sectional study, whenever exposure was kept constant, there were significant differences (if not in all measures) in favour of the LS group at each time. That is, those that were cognitively more mature were superior in spite of having started later. Therefore, the results are in a similar vein to those found, for instance, by authors such as Stankowski Gratton (1980) in Italy and different from the ones obtained by Yamada et al. (1980), whose study, as shown in chapter 2, attributed a superiority for young learners in the short run that was probably due to the nature of the task used.

From the cross-sectional study in the different data collection times, there emerge three other important points that deserve closer attention. The first one is that where the groups diverge most considerably is at Time 2 (with 416 hours of instruction for each group), when the mean AT for A2 is 12.9 and for B2 is 15. Therefore, although LS tended to be superior already at Time 1, the differences are not as striking as they are at Time 2, when both groups have been provided with some more exposure (216 hours).

Secondly, when the development of both groups for free productive vocabulary is not parallel, the rise from T2 to T3 is more noticeable for the ES, LS' gains in productive vocabulary knowledge seems to be a bit better between T1 and T2. The same holds for controlled productive vocabulary as measured by the cloze. Hence, it seems that the most noticeable development, which was found to be quite consistent, is to be found from age 12.9 onwards.

Thirdly, a similarity was found in the behaviour of the groups B1 and A2, B2 and A3, B3 and A4. That is, a group of ES at a particular time seems to perform similarly to a LS in a previous collection time. Between these group pairs, either no significant differences are found or the LS group outperforms the ES group. The pattern that emerges is the following: B1 significantly outperforms A2 in 12 measures (6 in the interview, 2 in the roleplay and 4 in the composition), B2 significantly outperforms A3 in 4 measures (1 in the storytelling and 3 in the roleplay), while B3 outperforms A4 in 2 measures (both in the storytelling). This could be taken as an indication that, in order to perform similarly to LS, ES need some more extra exposure and that the amount of hours of instruction needed would be lower as their cognitive maturity (AT) increases: at Time 4 a difference of 74 hours gives significant differences in favour of B3 in just two measures, while in previous times a difference of 200 or 300 hours gave rise to more differences.

It can also be inferred from the comparison of the groups B3 and A4 that, even if allowances are made for the younger group and more exposure is given to A3, it does not surpass B3 a year later either. However, the number of subjects in A4 is very limited and we should be careful not to generalise the observed behaviour. The tendency shows that A4 students have improved in free productive written vocabulary (maybe due to the emphasis in the University Entrance Examination) between the last two data collection periods. However, they are not clearly superior in spite of having started English earlier and having had an extra course of formal exposure (74 hours more). The vocabularies of these groups do not present crucial differences concerning productive vocabulary in the tasks performed. Therefore, in the light of these results, an early start cannot be considered an advantage or a handicap in itself. What is worth noticing, though, is that ES with 800 hours of exposure perform similarly to LS with less exposure. It thus seems that there are some time periods at which each group of learners tends to progress more as far as productive vocabulary is concerned. This fact raises questions about the most appropriate AO and amount of exposure to be successful in language learning and, above all, the way in which this exposure should be distributed over time in formal contexts for the study of the FL. As an example, we can refer to studies conducted in Quebec (Collins et al., 1999) or in Catalonia (Serrano, 2007), which analyse the effect of time distribution in intensive courses. In relation to lexical acquisition in particular, there is also research that examines the effects of spaced and massed presentations of vocabulary (Dempster, 1987).

It is important to highlight that the results are consistent with others found by the BAF group in different areas such as grammar (Muñoz, 2006b), writing (Navés, Torras & Celaya, 2003) or oral fluency (Mora, 2006). Results in phonetics, from Fullana (2005), were not consistent either in favour of the ES or the LS group, as the differences found were not systematic and depended on the sound under study. Tragant (2006) also argued that an early start does not significantly alter the level of motivation of the students: motivation is related as well to the hours of instruction received and the biological age of the learners, who seemed to be more motivated in secondary than in primary education.

Leaving now aside the differences that exist between the groups, the results obtained in chapter 7, which assign to the learners a productive vocabulary between 1,000 and 1500 words in four different tasks, show that these learners may just probably be in command of a ‘Little Language’. A ‘Little Language’[59] has been claimed to be composed of the first 1,000 words in a language and can be defined as a system that, in spite of being formed by a small amount of words, is sufficient to the learner to make himself /herself understood and allows him/her to perform a series of basic tasks.

The term was first used in relation to L1 acquisition, to refer to the ‘languages’ that children used, which were conceived as languages in themselves, not just part of a language, as a child’s vocabulary is “not random [but] it works in an efficient way: each item has a purpose, has a function, and produces a wanted effect. The range of the vocabulary covers every area of the child’s needs and responses to persons, activities and things which concern him/her” (Nation, 1996: i).

However, as Nation (1996) also acknowledges, the English of a native child and that of a person who has only had school lessons may be different and the latter may not be a language in the way of the former. The results obtained in this study make us wonder if school courses do actually give learners control over a ‘Little Language’, a fair amount of words to express themselves appropriately. Even though the production of the learners analysed indicates that they have learned about 1,000 productive words, it is remarkable that after more than 726 hours of instruction they are not able to be in command of a wider productive vocabulary in basic tasks that cover the basic functions

[59] The term ‘Little Language’ is the translation of ‘Det Lille Sprog’, the title of a book by Aage Salling.

also included in the list of things that the ‘Little Language’ should enable the learner to do, such as making demands, talk about time and relations, describe persons, activities and things or refer to concepts and relations between them. As shown in chapter 5, it is revealing that there were many participants who could not produce 50 words at Times 1 and 2 (after 200 and 416 hours of FL instruction, respectively) and at Time 3 they did not produce much either, even with tasks in which they had to talk about themselves. Furthermore, a quick count suggests that, if evenly distributed, taking as a basis the results displayed in Table 7.27 (about 1,100 words), A3 would have learned 137 words per year along 8 years, A4 122 along 9 and B3 157 along 7 years. These rates are of course very different from the picture that emerges from studies in naturalistic settings, especially from those that have tried to quantify the amount of lexical knowledge gained over a period of time, for instance during a stay abroad. Milton and Meara (1995) maintain that, although some of the subjects did not benefit from these stays as much as others, university exchange students appear to be gaining vocabulary at an average rate of over 2,500 words per year. Obviously, they were older students (university courses) but it can be assumed that the picture would not be that different if we took students four or five years younger still in secondary education, which has not been much researched yet.

That the deficiencies in oral and written production are evident has also been highlighted by teachers conducting research in secondary school classes. Martorell (2006) can be an example: he states that in our present school context there is a “lack of correspondence between the amount of time devoted to the FL teaching and the results obtained, which are poor in relation to other European countries” (2006:38).

Some other studies carried out in Europe, in spite of showing a somewhat better production in the FL, are not optimistic either. The study carried out by Tschirner (2004) can serve as an example: the English vocabulary breadth of 142 learners was measured at the end of secondary education in Germany, where participants had been learning English between 5 and 11 years with an average of 8 years. A 87% of the students had had 8 years of instruction or more. As explained in chapter 3, Nation's Level test was used at the receptive and productive levels and a 72% of the students did not have a receptive vocabulary of 3,000 words and a 79% failed the 2,000 productive word level. Tschirner (2004: 37) concludes that "even extended sequences of English instruction of up to eight years do not necessarily enable students to meet vocabulary thresholds for academic purposes. [...] Particularly the productive goals are missed by a wide margin".

Undeniably, these results may depend on (and have obvious implications for) language learning situations and teaching programmes. As pointed out by Tschirner (2004), very few students use English productively in speaking or writing outside the class - and she found a significant correlation between oral communication opportunities outside of class and the size of the productive vocabulary-.

Kojic-Sabo and Lightbown (1999:190) also claim that "learner initiative and independence, along with the extracurricular time spent on language (and vocabulary) learning, are seen as two crucial factors related to higher levels of achievement". In class, it might be assumed that most of the time learners do not participate because the classes are teacher-centred and students are not asked for an extended participation or are not 'pushed' in their output (Swain, 1985).

Actually, production and practice have an utmost importance for language acquisition and for vocabulary in particular. Kirsner states that practice is essential in the L1 for lexical acquisition:

> "The role of practice in first language vocabulary acquisition has been vastly under-estimated. [...] The practice-counts for early words may be 100 or more times greater than 'comparable' words encountered at or after maturity. This body of highly practised lexical procedures, for use in pattern recognition and production, provide a basic pool of 'automated' examples" (Kirsner, 1994:308).

However, also in the L2 "a basic pool of 'automated' examples" is needed and therefore practice is essential especially in the first stages of learning a language. To practise production is indeed a great challenge posed to the learner (Waring, 1997) and therefore several ways to stimulate that a new word becomes productive have been suggested, especially in the first stages of learning a language, so as to have a vocabulary that allows students to cover the basic communication needs. Lee and Muncie (2006) recommend interactive elicitation of vocabulary on the part of the teacher as well as multimode exposure to target vocabulary. Nation (1995b) acknowledges that speaking tasks (like interviews, roleplays, split information tasks, retellings, etc.) are not usually thought of as having vocabulary learning goals because "it seems difficult to plan vocabulary learning as a part of a syllabus using activities that are largely productive, unpredictable, and subject to the whims of the people who happen to be in the discussion group" (1995b:11). Nonetheless, Nation states that speaking tasks very often need a written input that can be positive for vocabulary learning, as three crucial aspects are present: noticing (that may also mean negotiation of meaning between the learners), retrieval (receptive or productive) and generation (which is usual in the retelling or restructuring of what was given as input).

As regards the lack of vocabulary found especially in the students at primary school (where the use of the L1 was considerable as shown in chapter 4), a comment on the students' course books needs to be made. It is common for low level books to have a large amount of infrequent vocabulary, which stems from the fact that usually young students learn English through stories. Milton and Vassiliu (2000) compared different books for Greek learners of English and concluded that there was a very small amount of common vocabulary to all books and that although they tended to include Nation's Level 1 words (the most frequent vocabulary), young learners were exposed to high amounts of infrequent vocabulary at the outset of their learning. They also noticed that there was a gap between the most frequent and the least frequent words, which coincided with Nation's level 2 (or the second 1,000 most frequent words), which was scarcely represented in the language books they analysed.

In our case, the fact that the vocabulary learners knew could be highly idiosyncratic to the book they used -and could not be transferred to the tasks they were asked to perform- might be one of the reasons of the poor vocabulary production in the short-term, as their experience with language could be reduced to a couple of English textbooks. In any case, this is a question that remains for further research as an analysis of the input the learners had received was not carried out.

However, it is evident in our study that most of the words produced by any of the groups belong to the first band, as shown in the profiles in chapters 6 and 7, and very few to the other bands. This might be due to the course books used, which might be biased in terms of vocabulary representativeness, or to the fact that the infrequent vocabulary they might have learned in the first stages cannot be applied years later, either because the vocabulary of the task is self-constrained, or because they may not have it available productively as it may have not been recycled (as Harwood, 2002 notices, recycling is not usually a standard feature of ELT materials).

Finally, another reason for the poor productive vocabulary could be the small size of learners' receptive vocabularies: with a small receptive vocabulary, the probabilities of communicating effectively are very scarce. Tschirner (2004) states that the main cause for the small receptive vocabularies is that learners do not read on a regular basis. Although the present study has focussed on productive vocabulary, further research can investigate and compare the development and levels of achievement of receptive vocabularies at these levels of language proficiency.

All things considered, the results obtained in this study on vocabulary, age and exposure make one think that, when planning a FL course, it is crucial to bear in mind the following appreciation by Lightbown and Spada on the decision about when to introduce SL instruction:

> "When the goal is basic communicative ability for all students in a school setting, and when it is assumed that the child's native language will remain the primary language, it may be more efficient to begin language teaching later. In research on school learners receiving a few hours of instruction per week, learners who start later (for example, at age 10,11, or 12) catch up very quickly with those who began earlier. Any school program should be based on realistic estimates of how long it takes to learn a SL. One or two hours a week -even for seven or eight years- will not produce very advanced second language speakers" (Lightbown & Spada, 1993:50)

This would suggest that, as far as productive vocabulary is concerned (controlled or free, oral or written), it is not always true that the earlier a SL is introduced in school, the greater the success in learning; which is not the same as saying that en early introduction to the language is not a good choice. As they also point out, "when the objective of SL learning is native-like mastery of the target language, it is usually desirable for the learner to be *completely surrounded by the language* as earlier as possible" (Lightbown & Spada, 1993:49). It is precisely the creation of '*a surrounding with as much English as possible*' that the school should strive for.

8.3. Vocabulary measures

The main purpose of the present dissertation was to analyse issues of age and exposure in relation to productive vocabularies. Therefore, it was decisive to carry out the analysis with the measures that could better gauge vocabulary development and inform both the researcher and the teacher of learners' strengths and needs. Our third research question was to explore how different intrinsic and extrinsic measures described the productive vocabulary of our informants. As Richards and Malvern have noted,

> "the choice of measures needs to be theoretically motivated to have good construct validity in relation to the contexts, purposes, and research questions to which they are applied." (Richards & Malvern, 2007:92)

There are different measures of vocabulary development, especially of vocabulary richness. However, most of them have been showed to have flaws and the search for reliable measures, especially in speech production, has been described as 'a quest for the Holy Grail' (Tidball & Treffers-Daller, 2007). As the measures in this study need to be applied to a very varied sort of productions (from 11 year-olds in Grade 3 to 18 year-olds in Grade 12, and both to oral and written data), some intrinsic measures were chosen to describe the lexical production of all groups. These measures were mainly the traditional ones found in the literature (types, tokens, word families, standardised TTR etc.). Special emphasis was put on the text length of the texts to which the measures were applied, as some measures, like TTR, have been shown to be sensitive to it.

The composition and the interview showed length effects, as they give different results depending on whether the length is kept constant or not. Results from the roleplay and the storytelling were not so dependent on length. Nevertheless, keeping length constant, as has been done in the literature to overcome these problems, has been shown in chapter 5 to be misleading in productions by low level students, due to the inability of these less proficient learners to construct a coherent text.

The D index was also chosen to be used in this analysis, as it has been claimed to be more reliable than any other available measure for texts of short length. It had been taken as well to effectively describe the writing development of children of different ages during the school years, showing a continuous trend between levels (Malvern et al., 2004). As very few studies had used this measure, which seemed to overcome some of the problems the other measures had, especially as regards text length, *D_Tools*[60] was

[60] A new version of *D_Tools* and a new version of *V_Size* (versions 2.0) are now available at http://www.lognostics.co.uk/tools/index.htm. Most of the processes carried out with the first versions (the ones used in the present dissertation, see Appendices D and F) have been automatised in the new versions,

created to compute Ds instead of *vocd,* which could only analyse data transcribed into CHILDES. Therefore, the creation of the program was also a way of extending the use of this measure among other researchers whose data was not coded in CHILDES and of making it available to the scientific community, as only when it will have been applied to large amounts of data reliable conclusions will be arrived at in relation to its behaviour.

Among the results obtained for this measure, the following should be remarked. First of all, it correlates with types, word families and standardised TTR but no consistent correlation was found between D and the amount of tokens in the tasks. This would mean that, contrary to TTR, it is not affected by the length of the texts. Similar results were obtained also by Daller and Phelan (2007): all the measures they used to analyse written texts of foreign university students displayed a positive correlation with length but D presented the weakest. In the present study, D also showed that there had been gains at stages where they were not very noticeable (as with the LS group in the composition between Times 2 and 3) or where they were not properly detected by standardised measures, as it happened in the interview for both groups when measures applied to standardised texts gave better results for ES (see Figure 5.8).

It is also worth mentioning that, following the indications given by Malvern & Richards as regards the values that D could have at the different stages of learning a language, we see that there is a potential for D to be related with the vocabulary size of our learners. As they give mean D values for children learning their L1 (included in

the computations have also become faster and the general functioning more user-friendly.

Table 5.16 in this study), it can be seen that none of our groups has higher Ds in the FL tasks than a 3 year old child speaking in his/her mother tongue (D=47.83), and most of the values calculated for FL learners are around the one obtained for a 2-year-old child (D=27.44). If children are thought to learn about 1,000 words per year in the L1, it could be inferred that the participants in the present study would not have in any case larger productive vocabularies than 3,000 words. This is confirmed in chapter 7, where all the estimates obtained are well below this size. However, as it has been seen throughout this study, inferring a total vocabulary size is risky and not well-founded, and estimates for the vocabulary 'as a whole' in the SL should be treated with due caution.

As far as the relationship between D and TTR is concerned, it should be said that although the authors of this measure have been critical with TTR, TTR is still the essence of the D measure. D is theoretically more valid and it is not affected by text length as TTR is, but we should probably look for other ways of assessing lexical diversity in the near future. Jarvis (2003), for instance, proposes six properties for lexical diversity that should be taken into account: variegation (the number of types that can be found in a text corpus), mass (the size of the sample in which diversity is being considered), balance (how evenly balanced the number of tokens is for each type), dispersion (clustering of types) and disparity (use of infrequent words), which would probably be integrated in a single software tool (MTLD) which is still being tested. Meanwhile, D seems to be more reliable than other measures and it will probably be used in many other studies, with different data, from which further conclusions of its behaviour will be drawn. As van Hout and Vermeer have well advised, "pure mathematical definitions of lexical richness do not suffice and it brings to us the conclusion that we need very large corpora [...] to better inform our study of the small ones" (2007:115).

In addition to using these intrinsic measures, other extrinsic evaluation of the data was carried out. Multiple assessment was thought to be essential to come to reasonable conclusions as regards the issue of age and vocabulary in the FL. In addition, it has also been acknowledged that the frequency factor could be very helpful in improving lexical richness measures (Daller, van Hout & Treffers-Daller, 2003; Van Hout & Vermeer, 2007) and therefore, the vocabulary profiles and the lambda values for each student and task were computed, as well as the cognate and Anglo-Saxon indices.

The vocabulary profiles and the lambdas were computed individually for each task and were also calculated for each corpora of tasks in each group. Hence, possible bias of results due to the very few tokens that the learners produce could be discarded. The results were actually very similar either if they were obtained from the profile of each learner or from the corpora of the task in each group. The profiles for A3, B3 and A4 were remarkably alike and most of the learners' production in all tasks -oral and written- was found in band 1k, that is, among the 1,000 most frequent English words. However, as the cognate and Anglo-Saxon index show, the young ES group (A3) resort to the use of cognates more often. This would be coherent with what was found in chapter 4 in the pilot study with the storytelling, which could also be interpreted as an initial advantage for LS: they spontaneously produced more tokens in the target language and used the L1 less often. What is important with profiles, as they cannot always gauge particular changes in the use of productive vocabulary, is to use other measures such as the aforementioned indices to complement the profile information, as proposed by Horst and Collins (2006).

The lambdas calculated by *P_Lex* did not show any significant differences between the groups either, but two comments need to be made about the lambda values obtained in this study. Firstly, as pointed out by Meara (2001) written texts tend to produce higher lambda values, as the ones resulting from the compositions are higher than those from the oral tasks. Secondly, nearly all the lambdas in this study are below what is supposed to be normal. Usually, a lambda should range between .5 and 4.5 and only in the composition did two of the groups reach means of .446 and .476. Therefore, we could wonder if values lower than .5 would not be common for oral texts at low levels: Daller and Xue (2007) also obtained means of .16 and .23 for Chinese learners of English at University in the oral description of two comic strips; Read's (2005) lambda means ranged between .83 and 1.10, as found in the IELTS Speaking Test of students all over the world (the maximum lambda was never higher than 1.5).

Overall, lambdas can be considered a good reflection of the high and low-frequency vocabulary use; the ones in this study would show that the texts contain a big amount of high-frequency words, as this is also corroborated by the profiles and the estimations. There is, however, something that should be taken into account when interpreting the results from *P_Lex* and it is that, as noted by Malvern et al. (2004), it works on the sampling of tokens rather than types, and this means that it could be possible to have high lambdas for texts than contain a few rare words that are frequently repeated in the text. In any case, it will also be essential that the lists the program works on are selected or adjusted according to the purpose of the study.

The next section deals specifically with the productive vocabulary estimates proposed in the present study as possible extrinsic measures. This proposal constitutes the answer to our fourth research question, i.e. how can the learners' oral and written productive vocabulary sizes for some tasks be estimated.

8.4. Estimating productive vocabulary size

There is a strong need for research in SLA to have reliable tools to estimate the size of learners' vocabulary. As described in the previous chapters, a distinction is usually made between receptive and productive vocabulary size. There are some tests to estimate receptive vocabulary, which results can be used as valuable indicators of what the learner can understand when s/he reads in or listens to a SL. The amount of vocabulary the learner knows receptively can also be a good way of determining not only the sort of input s/he will need for the interlanguage to develop, but also the use s/he will probably make of the input received.

In addition to having an index of receptive vocabulary size, it would be also very practical to know the size of productive vocabularies because this information would help us to predict what the learner will be able to do in a SL, which tasks s/he will be able to perform or up to what extent s/he will be able to make himself/herself understood. As Read (2000) notes, there are studies that try to set threshold levels for reading comprehension but there is no similar basis for estimating the minimum vocabulary required for language use. There are very few studies in the literature that try to estimate the size of productive vocabularies, most of the estimations come from results in different types of tests, which share the trait that the words tested have been selected from different frequency levels. What we have presented in this dissertation is a way to estimate productive vocabularies taking as a point of departure the learner's production in the SL, either oral or written. The fact that frequency profiles can be obtained for any text offers the possibility of describing the text by means of a curve, which then can be used to infer which original vocabulary size could have given rise to it. In this way, an indication of the size of the vocabulary for a particular task can be provided. This estimate of vocabulary size could be a valuable indicator of productive lexical knowledge and could complement the present measures of lexical richness that are used to analyse learners' production.

Although this is one of the first attempts to compute productive vocabulary size using this procedure, and further research will probably be needed to adjust the process, we believe that the methodology and the mathematical process used by the program are well-grounded, as presented in section 8.4.1, and that it could offer a number of possibilities to explore productive vocabulary development in the future.

The proposed model takes advantage of a certain inherent order present in all languages and in the language acquisition process to grossly infer, using a mathematical model, the productive vocabulary available to the learner for certain tasks. As language also displays a certain degree of randomness in the rules it tends to follow, an exact amount of words known productively by the learner cannot be computed. However, deviations from the general rules could be taken as indications of where to adjust the model to make the estimations more reliable.

8.4.1. Towards a model

Language is a complex system, as such it follows certain patterns and shows some order but it also contains a certain amount of entropy or randomness. According to Balasubrahmanyan and Naranan (1996), language exhibits the characteristics of complex adaptive systems, which tend to be found between totally ordered and totally disordered systems.

8.4.1.1. Language as an ordered and as a random system

There are certain rules that language, as a system, tends to follow. One of them is Zipf's Law, which is a universal property of world's languages. This law states that there are certain words that occur more often than others, more specifically, that the frequency of any word is roughly inversely proportional to its rank. Empirical studies have found that the approximations of this experimental law closely resemble what happens in the real world, both in the written and in the oral language. It has normally been applied to written corpora, but oral language also follows this law. For instance, Ridley (1982) concludes that the fact that the law holds for speech samples is more significant than the fact that Zipf found this pattern in written language, as written compositions are subject to numerous trials and revisions before taking their final form, which is not the case for spontaneous oral language. Ridley & Gonzales (1994) showed that in some samples of adult speech production it was impossible to distinguish between individuals on the basis of Zipf's Law deviations, thus demonstrating the tendency for this law to be obeyed in all occasions.

Nevertheless, language also exhibits a certain degree of entropy. This randomness, more than challenging or questioning the general rules that language tends to obey, could be taken as an indication of where the general rule does not apply. The study of entropy could be an effective way of gaining theoretical knowledge about the variations of the general laws, in this case in particular, of Zipf's Law.

Several studies have criticised the use of lexical frequency counts. Gardner (2007), for instance, claims that this type of counts are distorted due to the fact that they do not take meaning into account. He exemplifies his claim with the fact that these lists are often lemmatised and within the same lemma different frequency levels can be found (for instance, *climbed* could appear more often than *climb*). He gives further evidence of the neglect of meaning by stating that the commonest words in a language tend to be polysemous (the 100 most frequent phrasal verbs in the BNC have 559 potential meaning senses, or an average of 5.6 per phrasal verb) and he argues that

> "this line of reasoning brings into serious question the validity of computerised counts of individual word forms or computer-generated lists of individual forms for investigative or instructional purposes, especially if those words are of higher general frequency in a language" (Gardner, 2007:252).

Nonetheless, there are other studies that take into account that words have meaning, for instance, Ferrer i Cancho (2005). In his study, he proposes a model that, in addition to assigning an important role to meaning, tries to limit the variation offered by Zipf's law. Thus, in a certain way, it aims at 'ordering some of the randomness' the law does not predict. Ferrer i Cancho proposes a model based on the fact that words in the language follow a power function (Zipf's Law). As a function, he states that its exponent is known (its value is around 2), and therefore the function has a low and an upper bound. He further expounds that this exponent contains information about the ability to balance, on the one hand, the goal of communication (i.e. maximising the information transfer) and, on the other hand, the cost of communication (imposed by the limitations of the human brain). This would explain why there is a variation in the exponent and would also put some limits to this variation, because on the basis of the goal and cost of communication only a particular range of exponents should be found in human speakers. He shows that a big exponent can be found in fragmented discourse typical of schizophrenia, which is characterised by the absence of a consistent subject and the presence of multiple topics and with a varied lexis; the maximum finite value predicted would be between 2.11 and 2.42. On the contrary, low exponents (around 1.6) would be typical of the speech of young children and of very advanced forms of schizophrenia, where texts are filled mainly with words and word combination related to the patients' obsessional topic; the variety of lexical units employed here is restricted and it includes many repetitions.

Therefore, the criticism that frequency counts do not take meaning into account does not necessarily undermine Zipf's Law validity. First of all, because the law is observed independently of meaning and, secondly, because meaning can be used in different models to predict how the function would work in different types of discourse (by conferring different values to the exponent).

It has also been acknowledged that the Power Law might not apply to the most frequent words (Ridley, 1982) and to the least frequent words in a language. This is not actually a recent objection, as this supposed imperfect adjustment is already acknowledged in Zipf's work and mentioned, for instance, in Crystal's Encyclopaedia (1941/1987). However, Milton (2007) points out that knowledge of high frequency words follow the expected pattern more regularly. New computer tools and the development of corpus studies have allowed to determine in a more systematic way the scope of the variations. Consequently, the improvements in these two fields (technology and corpora studies) have resulted in new data and outcomes that can be used to overcome the 'randomness' problems that the application of this law may entail.

Regarding the possibility that the Law does not properly describe the distribution of very infrequent words, a study by Fuks & Phipps (2006) deserves closer attention. In this study Fuks & Phipps work with language corpora using a network paradigm[61]. By studying the subgraphs generated by the most frequent words in a language, they have shown that "the clustering coefficient of the subgraphs reaches a minumum in the same place where they find an inflection point in the rank-frequency plot" (2006:263). This coincidence is shown to be related to a change in the general structure of the language and the alteration takes place when vocabulary size reaches about 3,000-4,000 words. The authors consider this a threshold that corresponds to the transition from Zipf's Law to a non-Zipfian distribution in the rank-frequency plot.[62]

[61] Complex networks are formed by a large number of components that interact with one another, hence they are useful to explain many phenomena of which the functioning of language is just one example.

[62] This threshold that they proposed would not be in disagreement with one of our findings: in the explorations carried out with *V_Size*, some words in the least frequent bands were needed to be included in the profile in order to obtain vocabularies bigger than 3,000 words. Further research could explore if this transition proposed by Fuks and Phipps is confirmed in other corpora and if vocabularies bigger than this size would normally exhibit a high number of non-frequent words.

Therefore, if there are words that could probably be used less frequently than what the law might have predicted, there may be a need to make certain adjustments, like adding some new parameters because, as Edwards and Collins (2007) point out, this would always improve the fit of the curves, as the final estimates are sensitive to the details of the model. They cite the study by Kanter and Kessler (1995), which shows that a Markov process[63] model would fit the frequency distribution data better than the straight line produced by Zipf's Law. However, in spite of the fact that there would be an improvement of the fit at the lower end of the curve (for less frequent words) other modifications like an increase of 1k words (the most frequent ones) would automatically take place as a result. Thus, special care must be taken with the addition of new parameters.

While Fuks and Phipps (2006) and Kanter and Kessler (cited in Edwards & Collins) present studies that give information on the infrequent words distribution, there are other studies that analyse more specifically what happens with the distribution of the most frequent words, that is, what takes place at the upper bound of the Zipfian curve. One of these studies has been carried out by Ninio (2006), who investigated the statistical features of Hebrew motherese. He focussed on the rank-frequency distribution of sentences containing a verb or an adjective followed by the indirect object marker and the indirect object. Results display a clear power-law distribution but the ten most frequent verbs in this construction had a different power-law exponent from the remainder. He called these verbs the 'kernel vocabulary' of Hebrew ditransitive verbs.

[63] A Markov process is a random procedure in which the probability of a future event is not affected by the past history of events.

The fact that these very frequent Hebrew verbs do not follow a Zipfian distribution at the upper end of the curve is something that might as well happen with other languages other than Hebrew and with word categories other than this particular type of verbs, which is an issue that remains to be further explored.[64]

8.4.1.2. Language acquisition as an ordered and a random process

Language was defined in the previous section as an ordered system that contains a certain amount of entropy as well. Something similar can be said about the language acquisition process. Whether we deal with L1 or L2 acquisition, with natural or with formal settings, there is some agreement on the fact that the process of acquiring a language usually follows some general tendencies that have been shown to be quite stable.

One of the general patterns that (S)LA conforms to has to do with the frequency of the items in the input that a learner receives: it has been shown that there is a positive correlation between the frequency of a form in the target language input and the order in which learners produce it in their output. Ellis (2002) offers a review of frequency effects in language acquisition and processing and Kirsner (1994) focusses on vocabulary in particular.

That the frequency of forms or meanings in the input determines the order in which they will appear in the output is not the same as saying that the frequency of

[64] It has recently been seen that there might be a possibility that the Power Law did not apply to the multi-word phrases frequency distribution (Egghe,1999).

elements in the language in general determines the order of acquisition. This is an important distinction to take into account because the program used to estimate productive vocabulary sizes (*V_Size*) works on frequency lists. As Edwards and Collins (2007) have noted, estimations assume that one learns words in their order of frequency in language in general and that they are not based upon a theory of how frequency arises in natural language and therefore there is no definite evidence why it must be correct. However, the lists chosen to inform *V_Size* were compiled taking as a basis the materials used by students learning a SL in formal contexts, which means that the words included will surely be part of the input that learners in such contexts receive. Therefore, words in the highest ranks of these lists will probably be found in the learners' output before those belonging to the least frequent ranks.

Nevertheless, the randomness in this general tendency could be found in the fact that students also learn very infrequent words in the first stages of learning a language, especially in a FL (see Milton & Vassiliu, 2000)[65]. Hence, even though the frequency lists taken as the basis of the program contain words in decreasing order of frequency in the 'theoretical' input that learners receive, there will always be a certain amount of items that would not follow this trend. That the vocabulary in each frequency band does not increase in parallel or that there are certain bands that behave differently than the others is not a novelty. What is crucial is to control the patterns that these frequency bands could offer. By understanding their behaviour we would not only understand

[65] That is why Edwards and Collins (2007) advocate the need of a probabilistic model for estimations of vocabulary size to be reliable, or a more in-depth modification of the model to take into account the learning process.

better how vocabulary develops but it would also be a great help to obtain more reliable estimates.

With respect to the variability that the frequency bands offer, results from the present study can be interpreted in the light of what was found by Meara (2005) and Edwards and Collins (2007). The proportions of words for each band change as vocabulary grows. Meara (2005) found out that the variation in the 2k band was not plainly evident across the range of vocabulary sizes. Edwards and Collins (2007) saw that as the proportion of 1k and 2k bands decreases, the proportion of 3k increases. Actually, 1k had 10 times as many words as 2k, so if 2k declines in the same proportion as 1k, the amount of 2k words does not vary much. They showed that as vocabulary size increases, the decrease in 1k words is counterbalanced by gains in 3k, and this seems to be a common trait in all profiles[66]. Although Edwards and Collins worked only with three frequency levels (1k, 2k and 3k) and we worked with 5, our results also showed that the variation in bands 2, 3 and 4 were not as important nor did they influence the final estimate as much as bands 1 and 5 did (and vocabularies bigger than 7,200 words were obtained by *V_Size* only if band 5 was higher than the previous ones). The results of the two studies share the finding that the highest and lowest bands have more prominence in the computation of the final estimate. Besides, if Edwards and Collins show that the 1,000 most frequent words have a lot of weight in determining the profile shape and in computing an estimate, we would add that the first 500 most frequent words are the ones that have a more determinant role, as our analyses were based on

[66] They actually propose that just by the proportion of 1k words in a text, a fairly reliable estimate could be obtained.

bands of 500 words and still the results were similar to what Edwards and Collins found.

It may also be probable that information about the word classes in the language in general and in the learners' production in particular could make the estimates more accurate. It is known that there are certain parts of speech that are more difficult to learn than others, for instance, nouns and adjectives are more learnable than verbs and adverbs because the former are more imageable. Furthermore, closed-class words (which contain a smaller proportion of low-frequency words) do not seem to adhere to Zipf's law as rigorously as content-words do (Ridley, 1982), which should be further investigated. However, although at first sight the behaviour of function words might represent a threat to the validity of estimations, it might probably not affect them, as a sample of speech or writing with just function words would not be common.

In sum, there are some general rules in language and in language acquisition. Obviously, there is also some randomness both in the language and in the process of acquiring one, but some of the general tendencies (like the Zipf's Law) could be taken as a way to obtain information about the learners' productive vocabularies. Therefore, *V_Size* is based on a theory of language that, in spite of not being faultless, could reasonably be expected to be used to obtain reliable estimates. As McNamara has acknowledged:

> "Every test is vulnerable to good questions, about language and language use, about measurement, about test procedures, and about the uses to which the information in tests is to be put. In particular, a language test is only as good as the theory of language on which it is based, and it is within this area of theoretical inquiry into the essential nature of language and communication that we need to develop our ability to ask the next question. And the next."(McNamara, 2000:86).

8.4.2. Pros and cons of the model

Edwards and Collins have suggested that the logarithmic implementation of the Zipf's Law is an adequate way to roughly predict word frequencies in a text. We agree with their argument that one of the main advantages of working from a mathematical modelling approach is that good predictions could be obtained if the underlying model works and that if this was not the case, that is, if the predictions were not seem to be reasonable, the model could be adjusted whenever real data deviates from it. Thus, it is crucial to work on the inferences computed by the program and the weighted selection process it involves with more real data with the purpose to make the predictions as accurate as possible. In addition, any mathematical model should be applied to and be tested against as much real data as possible.

While it is evident that the Law can work as a good predictor of word frequencies in a text, caution must be taken if it is used in texts produced by learners, especially as regards the following aspects: the curve-fitting process used for the estimations; the length of the texts produced; and the tasks for which the estimate is obtained.

8.4.2.1. The curve-fitting procedure

Two very important considerations in relation to the process of curve-fitting should be made. First of all, not all the students' profiles are adequate to compute estimates in this way. Curve-fitting will be a valid method whenever the learners' profiles resemble those produced by the logarithmic implementation of the Zipf's Law, i.e, a big amount of frequent words and lower amounts of infrequent words (the curve of the profile goes progressively down from left to right). Even though this is the general tendency in learners' productions, it may not probably be always the case.

There are receptive profiles that Meara and Milton have called 'profiles with structural deficit' (Meara & Milton, 2003). Case A and B in Figure 8.25 could serve as an example. Profile A may belong to a scholar who knows a high amount of technical vocabulary in bands 4 and 5 and has a very reduced 1k vocabulary, while profile B may be from a learner who has taught him/herself the language (Meara, 2006a). Although they are a representation of receptive vocabulary knowledge and as such they are not governed by Zipf's Law, these profiles of receptive knowledge might generate unusual curves in productive vocabulary that would produce very big errors in the estimations through a curve fitting procedure. *V_Size* cannot be claimed to compute estimates for profiles that do not exhibit the typical curve, since where there was considerable deviation from Zipf's Law, errors started being very big from 3,000 words onwards, as seen in chapter 7.

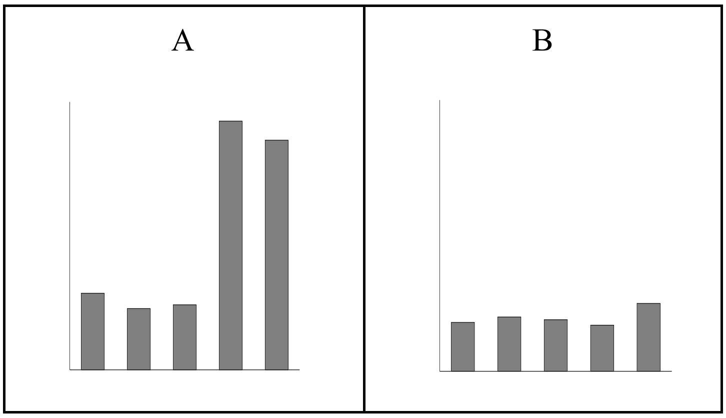

Figure 8.25.Unusual receptive vocabulary profiles in language learners.

Secondly, curve fitting is not an exact method, because there is always an approximation involved, as a way of neutralising the divergence between empirical and theoretical curves. Therefore, the more points we have in the profile, the more faithfully the empirical curve will be defined and the more precise the adjustment with the theoretical curve will be. It is because of this reason that *V_Size* uses 5 points in bands of 500 words each. Bands of 500 words have not normally been used in the profiles literature; however, as at the first stages of learning a language big vocabularies are an exception rather than a rule, bands of 1,000 words would make all these early vocabularies too similar to one another and predictions less precise.

8.4.2.2. The estimation process and text length

It is important to mention that the present study has shown that *V_Size* gives similar estimations for individual tasks (performed by each learner) and for task corpora produced by learners at the same proficiency level. Edwards and Collins pointed out that variability decreases a text size increases -an increase of the amount of words brings about a decrease in the standard variation-. In our case, variability was minimal. We also believe that the larger the corpora, the more exact the estimation will be. However, on an individual basis, with less words estimates for individual tasks do not differ much from the mean of the group if the estimate is computed from the corpora of tasks for that particular group, at least as regards the data analysed for groups A3, B3 and A4. Probably the main reason for the similarity is that the groups are homogeneous. Again, though, further research should look at how different the two groups have to be to find differences between them[67]. It is also worth mentioning that the sample size, i.e. the number of times a trial is made as part of the estimation process in *V_Size*, affects the result very slightly. In addition, moving one word between bands could produce small variations in the results. Therefore, to neutralise this possible effect, a wise indication would be to compute estimates for long texts (or group corpora) or to obtain more than one estimate for different tasks for the same individual, which would avoid possible distortions in the results.

8.4.2.3. Estimations for tasks

Having estimates for different tasks is probably one of the ways in which we could get closer to a general 'vocabulary estimate' for a particular learner, as the idea of computing a 'total productive vocabulary estimation' is more a chimera than an attainable objective. The operationalization, though, could be thought of in terms of estimates for different tasks, as Meara suggests:

> "There are no plausible methodologies that can accurately assess how many words learners know productively in the abstract. Each testing tool that we might use requires the testees to produce words in a particular context, and elicits only a small proportion of the total of words they know. Although it is tempting to use these small samples as a basis for extrapolating to larger 'productive vocabulary' scores, it is far from clear that these extrapolations are sensible. The best that we can do, probably, is to claim that experimental subjects appear to have a productive vocabulary of X words for a specific task". (Meara, 2006b:285)

[67] Meara (2005) has argued that it is complicated to find differences between groups that have big vocabularies. Groups with 2,500-3,000 words are distinguished 90% of the time, while those with 7,500-8,000 are distinguished 12% of the time.

As there were no systematic differences between the estimates in the tasks that the students performed for this study, it is possible that the mean, together with the standard deviations of these estimates, could be a good index of productive vocabulary size. The fact that there are no differences between the oral and the written tasks could be due to the fact that the vocabulary in the interview and in the composition (where participants talked about themselves) is very similar, but the estimates did not differ that much either in tasks like the roleplay or the narration, where the words used are also task-dependent. Estimating vocabulary for certain tasks is also more appropriate than estimating a total vocabulary because, in a collection of texts that share the same topic, the lexis employed by the learners will be used with the same meanings and the target vocabulary will be more fixed or more predictable, i.e. when learners perform the same task "there is less likelihood of semantic disparity, [...] because words and their meanings tend to be closely tied to the themes or topics of those texts" (Gardner, 2007:253). Also Laufer and Shahaf (1995) have pointed out that in spontaneous production learners may use familiar items that may not actually show the vocabulary knowledge they would exhibit in more controlled tasks. A set of tasks, then, would seem to constitute a good way to elicit a sort of learners' vocabulary that reflects to a larger extent, or is a bit close to, their real knowledge.

It is also possible that by building large corpora for particular standard tasks, we could produce frequency lists against which to judge learners' production using *V_Size*. Whether this would result in more accurate estimates than those obtained using the Jacet List as the basis is an issue that remains to be explored.

8.4.3. Validity and reliability: Reasonable expectations

Other limitations of the study carried out with *V_Size* have to do with issues of validity and reliability. Although some conclusions can be drawn from the present study, further research should deal with the validity of the estimation method proposed in a more systematic way.

The results obtained in this study could be taken as a guide to plan further research on the program validity. For instance, as regards concurrent validity, i.e, the operationalisation's ability to distinguish between groups that it should be theoretically able to distinguish between, *V_Size* gives different estimates for NS university teachers and for intermediate SL learners. The profile produced does not seem to be affected by age either, although the final estimate will be different (older native-speaking children have bigger estimates than the young native child). Nevertheless, larger amounts of data would need to confirm this initial suspicion. There is a strong need of studies where the program is applied to the data produced by groups that prove to be different to various extents. It would also be necessary to check its discriminant validity, i.e. the degree to which the operationalisation diverges from other operationalisations that it theoretically should not be similar to. However, the fact that the results yielded by the estimations do not enter in contradiction with those obtained in chapters 5 and 6, where no differences in the long term between A3, B3 and A4 emerge, would seem not to compromise convergent validity, that is, the degree to which the operationalisation is similar to other operationalisations that it theoretically should be similar too.

Regarding face validity, it should be emphasised that 'on its face', and in the light of what has been argued in the previous section, it seems that this estimation process is a reasonable translation of the construct: we would have also speculated that, at the end of secondary education, the learners will not have vocabularies bigger than 2,000 words. However, in assessing face validity we would have the same problem we would find in assessing content validity[68], that is, both types of validity assume that we have a good detailed description of the content domain, which is not completely true in this case: there is still a good deal of controversy about what 'productive vocabulary' is. Therefore, it might seem appropriate to take as 'working definitions' those proposed by Nattinger (1988) and Meara (1990) expounded in chapters 1 and 3, to conduct further research on these two types of validity.

As far as predictive validity is concerned, where the operationalisation's ability to predict something that it should theoretically be able to predict is analysed, *V_Size* could prove to be very useful for predicting what a learner would be able to do in the SL, as his/her success at a particular language level could be determined to a large extent by the lexical resources available for particular tasks.

8.5. Final conclusion

In this final section, some considerations are made in relation to the purposes of the present work, the methodology that has been implemented and the results we have

[68] In assessing content validity, the operationalisation is checked against the relevant content domain of the construct.

obtained. It becomes necessary at this point to make a final review of what has been analysed, what the results suggest and why. Likewise, it is important that we highlight the limitations of the study and give some directions for further research.

This dissertation has analysed the effects of age on vocabulary acquisition in English as a FL, which is an area on which very few studies have focussed. In particular, it has examined productive vocabulary both oral (free productive) and written (controlled and free). Lexical knowledge is an essential aspect of language proficiency at which students should make efforts to be competent for the purpose of communication. However, there has not been much research on whether productive vocabulary may or may not be favoured by an early or late start in FL contexts. This lack of research on these fields is especially remarkable if we take into account our current school curricula, where the age at which FLs are first introduced in compulsory education has been progressively brought down by most of European governments.

One of the main aims of the present study was to find whether there were significant differences at the end of secondary education in productive vocabulary between a group of ES (who started learning English at the age of 8) and one of LS (that started at 11). It has been shown that when exposure was kept constant (726 hours) and significant differences were found, those were usually in favour of the LS group, who were a year older than ES (16.3 vs. 17.9). Furthermore, differences between the two groups were also found in the short and mid term in favour of LS when exposure was kept constant (200 hours at Time 1 and 416 hours at Time 2).

When the comparison in the long term was carried out between ES who had received 800 hours of instruction and LS who had received 726 hours, and AT was kept constant, significant differences were not usually found, A4 does not outperform B3 in spite of having started earlier and received more exposure. As this result might seem surprising, especially in the light of findings in naturalistic settings and the belief that 'the younger the better' when learning SLs, we thought about several reasons to which this lack of significance in the sub-studies performed between A4 and B3 could be attributed to. However, the possibilities that the results could be accounted for by these reasons, as argued below, are actually remote.

The first factor would be that the tasks have not been properly designed or are not suitable for the type of analyses they have been submitted to. However, they have been shown to discriminate between the different levels and data collection times. Although some of the measures did not show significant differences at Times 3 or 4, this does not mean that they are not good enough to gauge progress or proficiency in the FL. The significant differences found in Times 1 and 2 between the groups have made evident that the development of productive vocabularies (both free and controlled) in the two groups is not parallel and that a more considerable improvement of their productive lexical resources is seen from age 12.9 onwards. Furthermore, it is also clear from the analyses performed that the significant results obtained in the different sub-studies are not likely to have been obtained by chance, given the amount of data analysed, the type of analysis performed and the levels of significance involved.

The second factor that might have accounted for the non-significant differences in the long term in spite of larger exposure by ES could have been that the spread of the scores in the different measures was so large that this wide range did not allow to find systematic differences. In statistics, the index that best describes the spread of the scores of a variable is the standard deviation. There were indeed some high standard deviations in some of the measures, which was something we took into consideration. The highest standard deviations were found in measures where length had not been controlled and the spread of the scores was a result of the fact that some subjects produced a lot while others produced much less. As it has been acknowledged at different points throughout the present work, special care was taken to control length, both for comparison purposes (standardised measures and measures applied to whole texts), but also for reliability purposes. Besides, in measures that are not affected, or at least not largely affected, by length such as D, significant results are not related to the value of the standard deviation, i.e. the spread of the scores. In the measures computed from standardised texts, even if the standard deviation was low, we do not find significant differences systematically either.

However, the results concerning the A4 group should be treated with caution, as the amount of participants involved is not high. This is one of the limitations faced in the present study. There was a low number of participants in two particular groups: A4 and the longitudinal group. This means that results from the sub-studies in these two groups must be treated as a gross indication and studies with larger populations would be necessary. Despite the low number of participants, though, data from the longitudinal group has been proved to be characteristic of the behaviour that both LS and ES groups exhibit in the cross-sectional studies. In particular, the longitudinal group confirms a pattern among the groups presented in which LS at both Times 1 and 2 are ahead ES in most measures of productive lexical knowledge.

We believe that the validity of the results in the present study becomes more evident when they are taken as a part of the 'bigger picture', that is, when interpreted in relation to the main findings of the BAF project in other areas. Findings from this dissertation are in the same line as those carried out within the same project. For instance, Torras et al. (2006) regard age 12 as a turning point for the development of grammatical and lexical complexity in writing. This age seems also to be crucial for the rapid development of morphosyntax (Muñoz, 2006b), narrative skills (Álvarez, 2006) and interactional skills (Grañena, 2006). In addition, most of the BAF results show that ES do not surpass LS in the long term (Muñoz, 2006a). In Torras et al.'s study, the significant differences found in lexical complexity measures such as noun, verb, adjective or adverb types at Times 1 and 2 were not significant at T3. Also Álvarez observes that although LS progress further after 416 hours of exposure, differences are not always significant after 726 hours. This fact, which could be taken as an indication that differences are progressively eroded and that in the long term ES will outperform LS, is accounted for by a different prediction in Muñoz (2006b). Muñoz argues that, for some time, LS show an advantage over ES thanks to their initial superior rate of learning but that, when both groups are similar in age and exposure, no differences should be found. That is, "if the older learners' advantage is mainly due to their superior cognitive development, no differences in proficiency are to be expected when differences in cognitive development also disappear with age" (2006b:34).

The results in the present study also suggest that apart from an early start, there are quite possibly other variables at work that might not have been emphasised enough in instructional settings such as quality of input or intensity and amount of exposure

(Muñoz, 2008). The fact that group A2 behaves similarly to B1, A3 to B2 or A4 to B3 might also point out that exposure could be an important factor to take into account when planning language programs at schools, as LS with less exposure get results as good as (and sometimes better than) ES, even after having received considerable amounts of exposure. It might also be taken into account when planning language policies that from around the age of 13 gains in productive vocabulary are more noticeable and that this rise seems also to take place in other areas such as writing development. This does not necessarily mean that we argue against an early start in FL settings, which could possibly be positive if the adequate conditions are met.

One of the main problems of studying productive vocabulary at low stages of development is the very few words the participants produce. The problem was more serious in the present study with the data from the lower grades, especially after 200 hours of exposure. However, in order to draw meaningful comparisons, some measures were needed to be applied to the production of all groups and the measures chosen were applied to both oral and written samples of all proficiency levels. Consequently, some subjects that did not produce enough had to be left aside in certain analyses. We consider, though, that the fact of having information about the lexical development in early stages outweighs the inconvenient of losing participants in some of the analyses, as the adequate statistical test could be equally performed or, in any case, descriptive information could be made available.

It should also be noted that D revealed itself as an index able to show progress where other measures fail to do so in short productions (as for instance the number of word families in the standardised interview). Further research should be conducted using

D as a measure to assess proficiency in lexical production. It has been seen that it could be useful to obtain information on lexical richness that cannot be derived from TTR, even if the texts are standardised, because in some cases keeping length constant in short productions could be deceptive, as has been illustrated with two short compositions in chapter 5. Besides, D may have also the potential of becoming a measure that correlates with vocabulary size, further studies may use it and try to provide yardsticks against which results from other samples could be compared.

There were other measures that could only be applied to groups that were more proficient and therefore their productions were longer and their productive lexical knowledge could be further explored. Nevertheless, in spite of the longer productions of the upper grades, some of the tools used to analyse the data were acknowledged to be more reliable with even longer productions. Consequently, some of the measures such as those given by *VocabProfile*, *P_Lex* and *V_Size* were applied both to individual productions and to whole group productions (by creating corpora with the productions for each group and task). The purpose of doing so was that shorter individual productions might have led to inconclusive results, as the values given for each separate production might have been distorted.

Whether the results from the present dissertation could have been affected by the instructional context (i.e. variables such as teachers, books or methodology) was an aspect that could not be investigated in this study. Vocabulary, as other linguistic aspects, is most probably influenced by the input received (from the teacher, the school or the books). Nevertheless, the influence of these variables was minimised by choosing students from different schools, taught by different teachers and acquainted with different textbooks. By relying on a sample from 21 different schools it is hoped to moderate a possible teacher effect. Hence, the wide range of the input sources those learners had used was thought to neutralise the influence of any possible important effects that might have been produced by these variables. However, a study on the types of input that young and older learners receive in instructional settings may help to elucidate its influence on both their rate of development and the final levels attained.

Finally, as pointed out throughout this dissertation, both *P_Lex* and *V_Size* are exploratory tools that have shown to give quite coherent results in comparison with the standard measures that have also been used. First, the lambda values corroborate the high amount of frequent words that were also given by the profiles applied. Second, the estimations of 1,000 productive words for the tasks analysed seem to be consistent with the results in chapter 6, as the profiles showed that most of the words used were from the first 1,000 band. These profile results can also indicate that the fact that the program gives similar estimates for both groups does not mean that the tool is unable to discriminate, but that the productions are alike, as has been shown in chapter 6.

In the present work, attempts have been made to ground the vocabulary size estimation procedure on a mathematical process based on the Zipf's law. Similarly, evidence has been given that it could possibly be a good way to approach the measurement of productive vocabulary for certain tasks in the future.

However, as indicated in the previous section, further research is necessary to make of this tool a standard measurement procedure and we must be extremely cautious with the results presented. The results in the present work can be taken as a sign of its potential validity, but as Fitzpatrick (2007:116) properly advises, we must be extremely careful in taking preliminary results from a testing tool as a “conclusive proof”.

First of all, it almost goes without saying that, as described in section 8.4.3, issues of validity deserve a much closer attention and reliability issues need to be addressed in depth, which will be the next step in our research. Secondly, the base lists on which *V_Size* works can also be improved to better detect any developments from the students learning English in formal contexts. As has also been shown in chapter 7, the idea of computing a total vocabulary size might probably be out of place if we do not actually deal with vocabularies for certain tasks. Ideally, these lists should be derived from corpora compiled from NSs or very advanced learners’ productions when performing the same tasks so that these instruments can be fine-tuned for different purposes. We think that groundbreaking research can be carried out in the area of productive vocabularies, and that further investigations will inevitably have to deal with it in an immediate future.