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 5

DESCRIPTIVE ANALYSES

5.1. Introduction

Chapter 5 presents the results obtained in the following tasks: an interview, a storytelling, a roleplay, a composition and an English cloze. Using different quantitative lexical measures, the roles of Age of Onset (AO), Cognitive Maturity (operationalised as Age at Testing -AT-) and Exposure are analysed in the long run. We want to see if those students who started learning English earlier (ES) will have better productive vocabularies than those who started later (LS), both after having received the same amount of exposure (726 hours) or right at the end of secondary education for both groups, that is, after the ES have received 74 hours of extra exposure.

This chapter aims at not only describing the state of these learners' vocabularies towards the end of the secondary education (section 5.3.1), but also at showing how these vocabularies developed from the first years of English instruction in the two groups, with a longitudinal study in section 5.3.2. and a cross-sectional one in section 5.3.3. Finally, in the light of the results obtained, a thorough evaluation of the D index of lexical diversity is carried out in relation to the other measures used (5.3.4).

5.2. Methodology: Measures used in the analyses

5.2.1. Traditional measures

Once the items for analysis in each task had been selected as shown in section 4.3.4.1 (Chapter 4), the number of tokens, types and word families were obtained using *VocabProfile* (Nation, 1995a). TTRs for each task were also computed. All these measures were calculated with and without standardising text length in all tasks. The reason for keeping length constant in one of the analysis is to control for a possible length effect in the results, especially in the TTR. Length was set at 50 tokens for the standardised tasks because it was the minimum number of tokens needed to calculate the D index and most of the tasks in our study were at least 50 tokens long. For the storytelling, roleplay and composition, the first 50 tokens of each task were chosen, for the interview we left out the first 20% of learners' production and counted the next 50 tokens, as the openings of the interviews were remarkably similar in all cases (name, grade, age...).

5.2.2. D

In addition to those widely-used measures of lexical richness, the D index was also computed for each task. This index has already been presented in chapter 3 and was originally put forward by Malvern and Richards (1997). For the computation of D a new instrument was created: *D_Tools* (Meara & Miralpeix, 2004). We followed McKee, Malvern and Richards (2000), who devised software (*vocd)* to calculate this index.

Contrary to *vocd,* it is not a requirement for *D_Tools* to have the data coded in CHAT system (MacWhinney, 1995) in order to calculate Ds. Our program does not offer the wide range of possibilities that *vocd* gives, such as morphemicisation, but it accepts data in plain text format and therefore the preparation of the transcripts can be less time-consuming. We believe it is a considerable advantage, especially when there is not a clear consensus about the reliability of this index and a wider implementation is needed among researchers: facilities in computation often lead to more systematic testing.

Furthermore, we also think that a careful study of the results of this index in the few contexts in which it has been applied can help us to interpret the results we obtain. Therefore, we will compare the results with the ones given by several studies. Obviously, the tasks will not be the same and participants can differ in age or proficiency, but bringing results together and classifying them according to different variables that can have an influence on the results (such as task, proficiency or age) can help us to see if results given by this measure tend to be generally consistent or not. This comparison is important especially in cases where the measures implemented, such as D, have not been used extensively in research. Additional results, those that will be obtained when more studies make use of the measure, will probably help to establish a range of values for D at different proficiency levels (ideally both for oral and written language), this would make interpretation of results more meaningful. A set of standard values against which lexical richness could be assessed (both in the L1 or L2) is necessary to overcome the problems researchers have traditionally had with TTR: as different studies usually standardise texts at different lengths, any sort of comparison involving lexical richness with this measure has been shown to be rather pointless.

5.2.2.1. D_Tools: Creation and validation of the program

D_Tools comprises two programs: D_0 and D_1. D is computed by selecting samples from the text of different token size (from 35 to 50 tokens). The program then calculates and averages TTRs at each point and matches the curve produced by our text with a theoretical curve produced by Malvern and Richards' formula: TTR= D/N $[(1+2N/D)^{1/2} -1]$, where N is the number of tokens and D is the value which represents the best match between the two curves and which is calculated using a least-square algorithm (see Appendix D on how to use the program). In order to validate the program, the following steps were followed:

- First of all, in order to confirm that the program was operating properly, results obtained with the program were also computed manually with the help of an *Excel* file for the longest calculations. 40 tasks were chosen at random to avoid any possible task effect (10 interviews, 10 storytellings, 10 roleplays and 10 compositions) and D was calculated in both ways. The results of the Ds computed manually and the Ds computed with the program all show a correlation of 1 [$r = 1$, $N = 10$, $p \leq .01$].

- Secondly, as the D index is the result of a curve-fitting procedure, there is always a small error when computing Ds. Therefore, it was necessary to check that the different results given for the same task in different trials did not significantly differ from one trial to another. For each of the 40 tasks above, D was computed 10 times with *vocd* and 10 with *D_Tools*. The difference between the highest and

the lowest D value computed with *D_Tools* tended to be about 0.5, which is a normal fluctuation, similar also to that given by *vocd*.

- The final step was to see if both programs were giving similar results in just one trial for different tasks. Therefore, the Ds of 16 sets of 4 tasks each (a total of 64 tasks) were calculated with *D_Tools* and *vocd*. As above, 16 were random interviews, 16 storytellings, 16 roleplays and 16 compositions to avoid any task effect. Pearson product-moment correlations were performed, as the variables had a normal distribution, and they showed that there is always a very strong correlation between the results obtained with both programs: D interview [r= .998, N= 16, $p \leq .01$]; D storytelling [r= .999, N= 16, $p \leq .01$]; D roleplay [r= 1, N= 16, $p \leq .01$], D composition [r= .997, N= 16, $p \leq .01$]. Appendix E contains an example of this validation process and the summary table from which these correlations are obtained.

5.3. Results

5.3.1. A long term comparison: The role of AO, AT and Exposure

Vocabulary results in the long-term will be compared first, that is, the results of students after having received at least 726 hours of exposure and having been learning English at school for at least 7 years (from Time 3 onwards). The three groups presented here (A3, B3 and A4) belong to Time 3 and Time 4 data collections (see chapter 4). The interesting point for the comparison of these groups is in relation to the three variables:

AO, AT and Exposure: as shown in Table 5.6, each pair of groups shares at least one of these variables while the other two differ: groups A3 and B3 have both received 726 hours of exposure but A3 started learning English at 8 and B3 at 11, also students in B3 were older than A3 when they were tested. A4 and B3 share the same age at testing (about 17.8 years old) but they had different AO and A4 had received more exposure than B3. Finally, A3 and A4 started English at school when they were 8 but the students in A4 are a year older than A3 and had received more exposure as well.

	AO	AT	Exposure
A3-B3	✘ (8 vs. 11 years old)	✘ (16.3 vs. 17.9 years old)	✓ (726 h)
A4-B3	✘ (8 vs. 11 years old)	✓ (17.7-17.9 years old)	✘ (800 vs. 726 h)
A3-A4	✓ (8 years old)	✘ (16.3 vs. 17.7 years old)	✘ (726 vs. 800 h)

Table 5.6. Common (✓) and different (✗) variables in the groups compared.

Tables 5.7 to 5.10 show the results for each group in all of the tasks. The first four columns as well as the column for D in the tables correspond to the means of the measures computed for each group without standardising task length. The last three columns present the results when length is set at 50 tokens. The reason for a lower number of subjects in the standardised tasks is that some subjects did not produce more than 50 tokens, which was necessary to standardise the tasks and to compute D (as proposed by Malvern et al., 2004). Figures 5.1 to 5.5 summarise this information graphically below. Both in the tables and figures below WF stands for 'word families'.

	Non-standardised						Standardised 50 tokens		
	Tokens	Types	WF	TTR	D		Types	WF	TTR
A3 N=57	145.54 (76.37)	70.79 (25.78)	56.42 (20.07)	.53 (.11)	40.35 (9.81)	A3 N=50	31.96 (3.40)	26.80 (3.26)	.64 (.07)
B3 N=41	207.85 (80.46)	91.83 (23.57)	72.46 (17.59)	.46 (.07)	44.53 (11.62)	B3 N=41	32.88 (2.70)	28.05 (2.76)	.66 (.05)
A4 N=16	173.44 (99.83)	81.25 (32.24)	64.31 (24.11)	.52 (.12)	44.85 (11.60)	A4 N=14	33.14 (4.80)	28.43 (3.92)	.66 (.10)

Table 5.7. Interview: long term.

	Non-standardised						Standardised 50 tokens		
	Tokens	Types	WF	TTR	D		Types	WF	TTR
A3 N=57	89.79 (37.56)	38.54 (13.05)	35.23 (11.64)	.46 (.10)	18.90 (6.26)	A3 N=47	26.81 (4.18)	24.72 (3.95)	.54 (.08)
B3 N=41	110.78 (35.91)	49.46 (11.90)	43.90 (10.77)	.46 (.06)	23.86 (5.91)	B3 N=40	29.23 (3.01)	26.35 (3.07)	.58 (.06)
A4 N=16	91.56 (40.74)	39.31 (11.90)	35.44 (10.17)	.46 (.09)	20.49 (7.40)	A4 N=14	28 (4.37)	25.64 (3.89)	.56 (.08)

Table 5.8. Storytelling: long term.

	Non-standardised						Standardised 50 tokens		
	Tokens	Types	WF	TTR	D		Types	WF	TTR
A3 N=54	65.06 (42.28)	35.61 (15.36)	31.06 (13.47)	.61 (.14)	33.77 (13.49)	A3 N=33	31.27 (4.049)	27.64 (4.29)	.62 (.08)
B3 N=41	70.56 (39.57)	40 (13.82)	35.88 (11.50)	.62 (.11)	38.85 (12.71)	B3 N=28	32.89 (3.20)	29.96 (3.25)	.66 (.06)
A4 N=12	68.17 (34.10)	40.08 (16.77)	35.08 (15.34)	.61 (.09)	40.79 (15.39)	A4 N=9	33.22 (5.63)	29.56 (5.05)	.66 (.11)

Table 5.9. Roleplay: long term.

	Non-standardised						Standardised 50 tokens		
	Tokens	Types	WF	TTR	D		Types	WF	TTR
A3 N=56	93.50 (41.16)	53.48 (18.62)	41.30 (15.47)	.60 (.08)	40.80 (10.63)	A3 N=50	33.20 (3.162)	25.56 (3.48)	.66 (.06)
B3 N=35	96.54 (44.10)	54.34 (20.23)	45.69 (16.99)	.59 (.08)	43.97 (11.84)	B3 N=28	34.86 (3.26)	29.04 (2.77)	.70 (.06)
A4 N=15	118.50 (57.13)	63.31 (23.47)	51.19 (19.41)	.57 (.11)	43.71 (11.63)	A4 N=14	34.86 (2.44)	28.71 (2.84)	.70 (.05)

Table 5.10. Composition: long term.

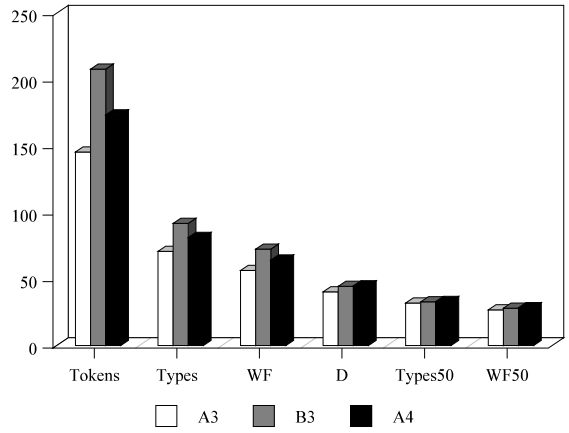

Figure 5.1. Interview: long term.

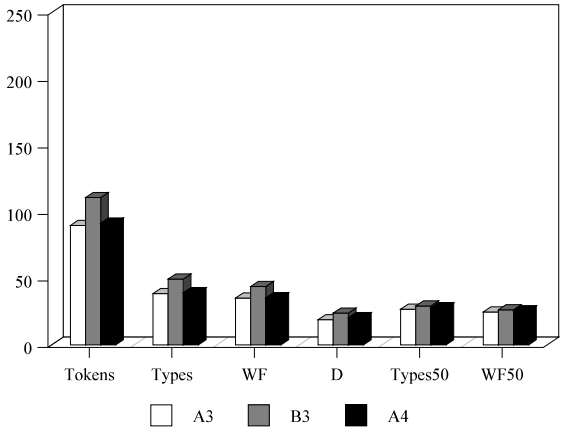

Figure 5.2. Storytelling: long term.

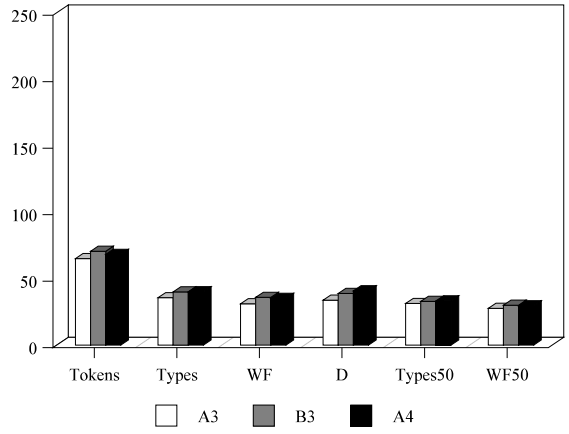

Figure 5.3. Roleplay: long term.

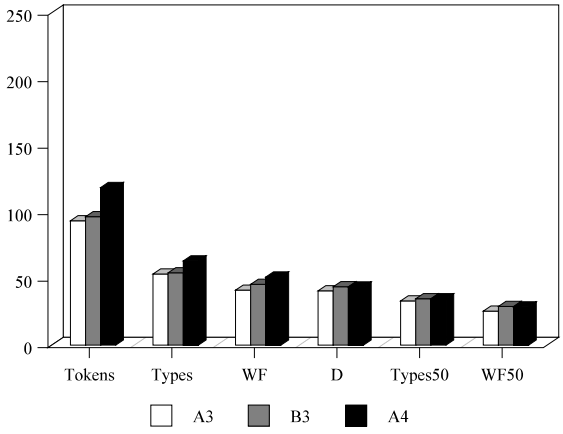

Figure 5.4. Composition: long term.

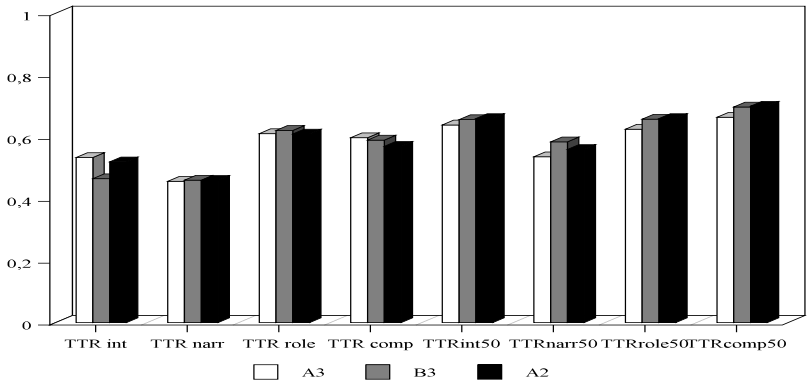

Figure 5.5. TTRs with and without standardising length: long term.

5.3.1.1. A3 and B3: Groups with the same Exposure but different AO and AT

Learners in A3 started English at age 8 (ES) and were tested at a mean age of 16.31, when they had received 726 hours of exposure, the same amount of instruction as the learners in B3. However, the latter are LS (they started at 11) and were older when they were tested (17.9). An independent-samples t-test was conducted to see if differences between groups were statistically significant. Alpha was set at .01 as the analysis involved multiple comparisons. Table 5.11 summarises the results, the shaded squares indicate significant differences.

	Non-standardised					Standardised (50 tokens)		
Tests	Tokens	Types	WF	TTR	D	Types	WF	TTR
Interview	t(96)=3.896 p=.000	t(96)=4.128 p=.000	t(96)=4.107 p=.000	t(95.4)=3.708 p=.000	t(92)=1.888 p=.062	t(89)=1.403 p=.164	t(89)=1.943 p=.055	t(89)=1.403 p=.164*
Story	t(96)=2.779 p=.007	t(96)=4.236 p=.000	t(96)=3.75 p=.000	t(92.8)=.228 p=.820	t(84)=3.766 p=.000	t(82.8)=3.123 p=.002	t(85)=2.117 p=.037	t(82.8)=3.123 p=.002
Roleplay	t(93)=.646 p=.520*	t(93)=1.440 p=.153*	t(93)=1.839 p=.069	t(93)=.381 p=.704	t(59)=1.503 p=.138*	t(59)=1.711 p=.092	t(59)=2.353 p=.022	t(59)=1.711 p=.092
Composition	t(89)=.334 p=.739	t(89)=.207 p=.836	t(89)=1.266 p=.209	t(89)=.424 p=.672	t(74)=1.199 p=.234	t(76)=2.195 p=.031	t(76)=4.541 p=.000	t(76)=2.195 p=.031

Table 5.11. Results of t-test analyses between A3 and B3, when $p \leq .01$.

B3 tend to obtain higher results in most of the tasks, although both groups perform similarly in the roleplay. When there are significant differences between the two groups, they are in favour of B3, except for the non-standardised TTR in the interview. Results from the interview show that B3 outperform significantly A3 as regards the number of tokens, types and word families when length is not standardised. In the storytelling, we find significant differences between the groups in all measures except when TTR is calculated without standardising length. In the roleplay, A3 and B3

perform very similarly, there are no significant differences except for word families when length is kept equal, in which B3 significantly outperform A3. Finally, significant differences between ES and LS are found in the composition only when length is standardised in the amount of types, word families and TTR, as shown above. In the controlled productive measure, the English cloze, the distribution was not normal in any of the two groups. Therefore, a Mann-Whitney analysis was conducted and it was shown that B3 outperformed A3: A3 (M=58.53, SD=27.59); B3 [M=82.52 , SD= 10.46; z= 4.389, p=.000].

As the measures marked with an asterisk in Table 5.11 do not have a normal distribution, either in both or in one of the groups, Mann-Whitney analyses were also conducted. It was thus confirmed that, although the specific significance value could vary, the same type of differences (i.e. significant or non-significant) were found in the same measures and tasks.[35]

In summary, significant differences between the two groups are always in favour of the LS (both in free and controlled productive vocabulary), except from the TTR of the interview, which is higher for ES when length is not fixed. While standardising length does not imply big differences in the significance of the results in the storytelling and the roleplay, results vary in the interview and composition depending on whether length is kept equal in all texts.

[35] For the non-standardised measures, the results of the Mann-Whitney analysis with an alpha level of .01 were the following: Tokens role (z=.613, p=.540); Types role (z=1.376, p=.169); D role (z=1.940, p=.052). For the standardised measures in the interview: TTR (z=1.517, p=.129).

5.3.1.2. A4 and B3: Groups with the same AT but different AO and Exposure

Independent samples t-tests were also conducted between A4 and B3, alpha was set also at .01 as multiple comparisons were involved. These groups share the same AT (about 17.8) but A4 are ES and B3 are LS; A4 has also received more exposure than B3. Significant differences were found in just two measures in the storytelling task: types [t(55)=2.894, p=.005] and word families [t(55)=2.705, p=.009]. As regards the English cloze, whose distribution was not normal in any of the groups, a Mann-Whitney test revealed that there were no significant differences either between B3 (M=82.52, SD=10.45) and A4 [M=73.23, SD= 22.71; z= 1.564, p=.118]. Consequently, the two groups can be said to have a similar behaviour as significant differences in favour of B3 are found in just two measures.[36]

5.3.1.3. A3 and A4: Groups with the same AO but different AT and Exposure

A paired-samples t-tests was conducted with the longitudinal subjects from groups A3 and A4 (N=9). They had started English at 8 but at Time 3 they were a year younger. Significant differences in favour of A4 were found in three measures in the composition: the amount of types in this task increased significantly from Time 3 (M=52.11 , SD=16.89) to Time 4 [(M=59.89, SD=23.76), t(7)=2.38, p=.049], also the word families used: A3 (M=40.89, SD=13.16); A4 [(M=49.33, SD=20.31), t(7)=2.673,

[36] For the variables that did not have a normal distribution, a Mann-Whitney analysis revealed non-significant results as well: Tokens role (z=.213, p=.832); Types role (z=.064, p=.949), D role (z=.602, p=.547), standardised TTR interview (z=.429, p=.668).

p=.032] and the standardised TTR: A3 (M=.6375, SD=.0391); A4[(M=.6875, SD=.0613), t(7)=2.38, p=.049].[37]

5.3.2. A short and mid-term comparison of longitudinal data from early and late starting school learners

Whereas the first study consisted in a long-term comparison of ES and LS' vocabularies, the next two (5.3.2 and 5.3.3) aim at examining learners' lexical performance in the same tasks but after 200 and 416 hours of exposure respectively. Are LS consistently better also in the first stages of learning a language? Do the groups follow a parallel evolution as regards lexical development?

First of all, two groups (ES and LS) of longitudinal subjects were followed after the aforementioned 200 and 416 hours of school instruction. There are thus four groups in this study: ES with 200h (A1) and 416 hours (A2), LS with 200 h (B1) and 416 (B2). The sample is small (see the *Ns* in Table 5.12) as longitudinal data is more difficult to obtain than cross-sectional data, especially when very strict criteria are set, as shown in chapter 4. However, evidence from longitudinal data was considered crucial in reaching any sound judgement on the evolution of these learners' lexical competence.

Only the measures computed for full-length tasks could be taken into account in this analysis, as groups A1 and B1 did not reach a minimum of 50 words in all tasks.

[37] Some statisticians claim that with a low number of subjects, even if the distributions are normal, results should be considered significant just if they are below .01. Therefore, as the significance in this analysis is borderline (p=.032 and p=.049), if we consider that only the ones below .01 are significant, we do not have any significant difference between the groups.

Figure 5.6 shows the means for each measure in each group and it can be appreciated that the LS groups (B1 and B2) are superior to ES groups (A1 and A2) at each time. In order to explore whether there was an impact of Time (200 or 416 hours) and/or Group (ES or LS) on the lexical variables studied, a two-way mixed design within-subjects Anova was conducted with an alpha level set at .01. There was a statistically significant main effect for Time in all the measures and tasks (except in the roleplay if we consider that significant effects are those below .01 due to the few subjects in the analysis, see note 37). There was also a significant effect for Group in the storytelling, the word families in the composition and the cloze (in bold in Table 5.12), whereas Group had a non-significant effect in the interview and in the roleplay. It can also be seen that there was no interaction between time and group except in the cloze.

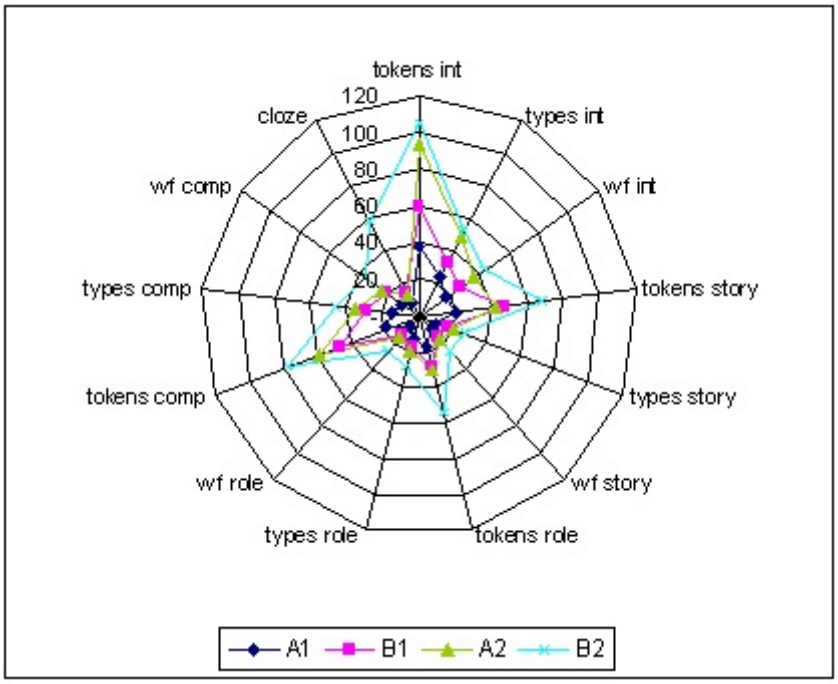

Figure 5.6. Means obtained for A1, B1, A2 and B2 in each measure and task.

	Measures	Test of Within-Subj Contrast		Test of Between-Subj Effects	
Interview N= 9 (ES), 9 (LS)	Tokens	Time	**.000**	Group	.553
		Time*Group	.641		
	Types	Time	**.000**	Group	.525
		Time*Group	.495		
	Word families	Time	**.000**	Group	.310
		Time*Group	.770		
Storytelling N= 9 (ES), 9 (LS)	Tokens	Time	**.001**	Group	**.004**
		Time*Group	.831		
	Types	Time	**.000**	Group	**.007**
		Time*Group	.782		
	Word families	Time	**.000**	Group	**.002**
		Time*Group	.646		
Roleplay N= 9 (ES), 5 (LS)	Tokens	Time	.039	Group	.129
		Time*Group	.434		
	Types	Time	.040	Group	.222
		Time*Group	.682		
	Word families	Time	**.007**	Group	.117
		Time*Group	.371		
Composition N= 9 (ES), 7 (LS)	Tokens	Time	**.000**	Group	**.**020
		Time*Group	.747		
	Types	Time	**.000**	Group	.019
		Time*Group	.689		
	Word families	Time	**.000**	Group	**.008**
		Time*Group	.432		
Cloze N= 9 (ES), 9 (LS)	Cloze	Time	**.000**	Group	**.000**
		Time*Group	**.000**		

Table 5.12. Summary of the effects of the Two-Way Repeated-Measures Anova**.**

In spite of the significant findings, the number of subjects involved is small, and this means that the results should be treated with considerable caution. They should only be taken as a gross indication of the subjects' behaviour as regards lexical performance. In order to have a more complete and reliable view, a cross-sectional study with a higher number of subjects follows.

5.3.3. A short and mid-term comparison of cross-sectional data from early and late starting school learners

This study presents cross-sectional data of the same groups in 5.3.2: A1 (N=31), B1 (N=29), A2 (N=47) and B2 (N=22). The descriptive information of the results

obtained for these groups can be found in Table 5.13. As was also the case in the study 5.3.2, standardised data of some tasks could not be obtained as subjects did not produce more than 50 tokens (shaded squares). A diagram with the means of each group can also be seen in Figure 5.7, where, for the sake of comparison and to obtain a general idea of all the levels, groups A3, A4 and B3 presented in the first study have also been included.

		Non-standardised					Standardised		
		Tokens	Types	WF		D	Types	WF	TTR
I	A1 (N=31)	32.97 (26.26)	21.87 (11.88)	15.84 (9.19)	A1 (N=5)	24.29 (12.02)	30.20 (3.71)	24.60 (5.94)	.60 (.07)
	B1 (N=29)	51.59 (34.59)	28.97 (14.02)	23.07 (10.85)	B1 (N=12)	28.18 (6.62)	26.83 (4.73)	22.17 (3.51)	.54 (.09)
	A2 (N=47)	82.79 (35.91)	45.57 (13.84)	34.79 (11.41)	A2 (N=40)	36.27 (16.46)	30.70 (4.47)	24.30 (3.91)	.61 (.09)
	B2 (N=22)	163.86 (95.89)	72.64 (30.89)	56.55 (22.87)	B2 (N=19)	37.85 (7.67)	32.37 (3.06)	28.05 (2.97)	.65 (.06)
S	A1 (N=31)	28.90 (23.36)	12.45 (7.52)	11.23 (6.46)	A1 (N=5)	9.36 (4.45)	21.80 (2.49)	19.20 (1.64)	.44 (.05)
	B1 (N=29)	43.52 (22.71)	17.45 (6.34)	16.48 (5.89)	B1 (N=8)	8.37 (5.52)	18.63 (5.26)	18 (4.90)	.37 (.10)
	A2 (N=47)	48.32 (26.61)	19.02 (7.39)	16.79 (6.06)	A2 (N=17)	7.67 (4.48)	18.59 (4.14)	16.76 (3.31)	.37 (.08)
	B2 (N=22)	77.14 (30.29)	32.73 (12)	29.05 (10.13)	B2 (N=18)	15.72 (5.02)	26.06 (3.13)	23.72 (2.82)	.52 (.06)
R	A1 (N=31)	14.74 (10.67)	10.77 (7.21)	7.94 (5.02)					
	B1 (N=29)	21.34 (15.67)	14.17 (8.28)	11.24 (6.5)					
	A2 (N=47)	28.64 (19.94)	18.94 (10.92)	15.04 (8.13)					
	B2 (N=22)	41.41 (33.41)	23.77 (15.36)	20.59 (13.07)	B2 (N=6)	22.11 (9.98)	27.83 (2.37)	25.17 (2.40)	.56 (.05)
C	A1 (N=31)	19.74 (13.54)	14.52 (8.19)	10.87 (5.09)					
	B1 (N=29)	40.69 (19.30)	25.62 (9.96)	19.48 (8.70)	B1 (N=8)	30.74 (6.41)	31.38 (2.56)	25.38 (2.67)	.63 (.05)
	A2 (N=47)	63.09 (22.56)	36.64 (10.40)	26.15 (7.91)	A2 (N=35)	30.05 (12.16)	31.03 (3.24)	22.31 (3.53)	.62 (.06)
	B2 (N=22)	89.05 (38.49)	51.05 (18.41)	39.64 (15.47)	B2 (N=19)	39.98 (9.66)	27.06 (2.86)	39.98 (9.66)	.70 (.06)

Table 5.13. Subjects in each group together with descriptive data.

As can be observed in Figure 5.7, a general comparison of the groups with different onset ages confirms the tendencies found in the previous two studies: LS groups are better than ES at each data collection time, B3 obtain better results than the rest of groups (except in some measures in the composition where A4 is better, though not significantly). Therefore, with very few exceptions, the groups' performance would be related as follows: A1<B1<A2<B2<A3<A4<B3. It can also be noticed that B2 and A3 present a very similar behaviour.

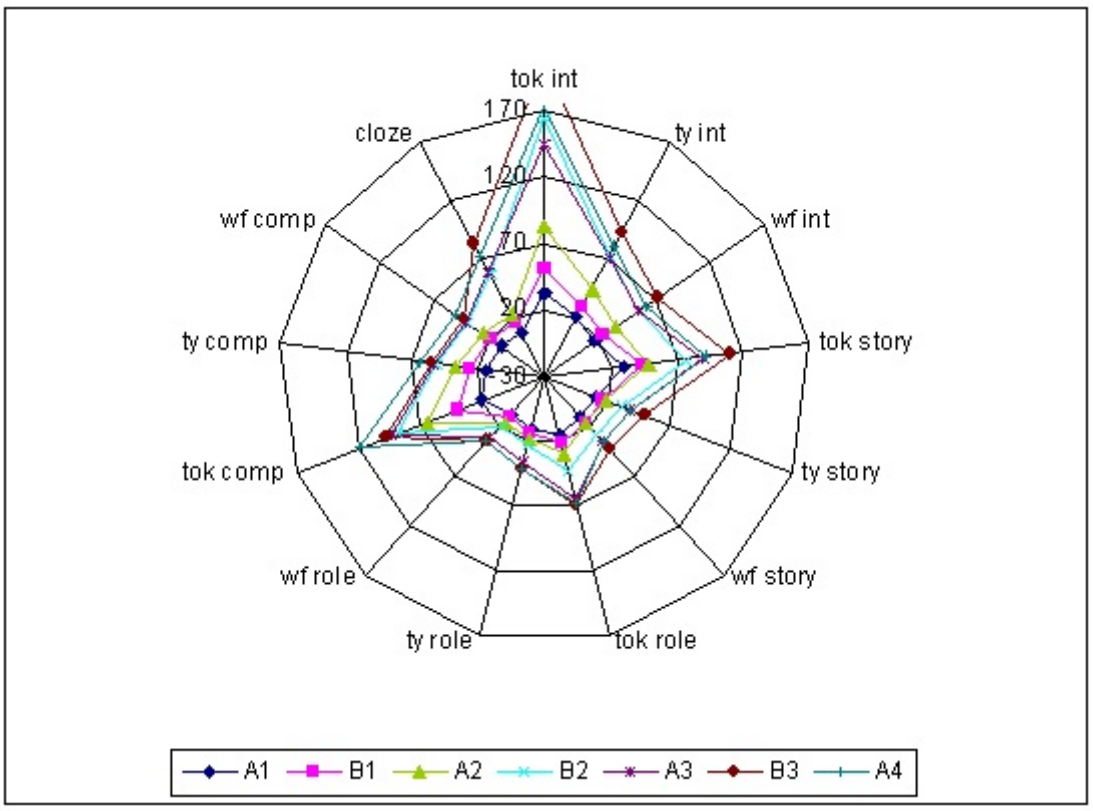

Figure 5.7. Means obtained for all groups in each measure and task.

Figures 5.8 to 5.11 show the evolution of both groups across times in each task until Time 3. As it also happened in the longitudinal study, very few subjects from the

lowest grades reached a minimum of 50 words in the tasks. This is the reason why we do not have diagrams for standard measures in the roleplay and composition (Figures 5.10 and 5.11): only a 16% of the subjects in A1 produced 50 tokens in the interview and just one subject reached this minimal amount of tokens in the composition. Therefore, standardised measures could not be computed for A1 in some tasks and the results obtained could not be statistically analysed either. Something similar happens in the compositions of group B1: a 27.6% of the subjects produced more than 50 tokens, but statistical analysis could be performed as the homogeneity of variances condition was fulfilled. Finally, none of the measures obtained in the roleplay for the groups in Times 1 and 2 (A1, B1, A2 and B2) were computed with 50 tokens, as producing more than 50 tokens in the roleplay was not common. Notice in Table 5.13 that the *N* of the groups for which standardised measures were computed varies and the specific *N is* given in the last four columns. Shaded squares in this table correspond to the groups whose data could not be statistically analysed due to the very few subjects that produced more than 50 tokens.

As regards the measures used, as it has been seen in the first study that TTR with full-length texts was not a reliable index (ES outscored LS while all other measures where pointing at the opposite way), only TTR with standard length was calculated in this study.

Notice in Figures 5.8 to 5.11 that there is some affinity in all tasks. First of all, the biggest difference between groups is normally found at T2 (416h), when A2 is 12.9 and B2 is 15. In addition, when the development of both groups 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 more obvious between T1 and T2. The same holds for the cloze, as can be seen in Figure 5.12.

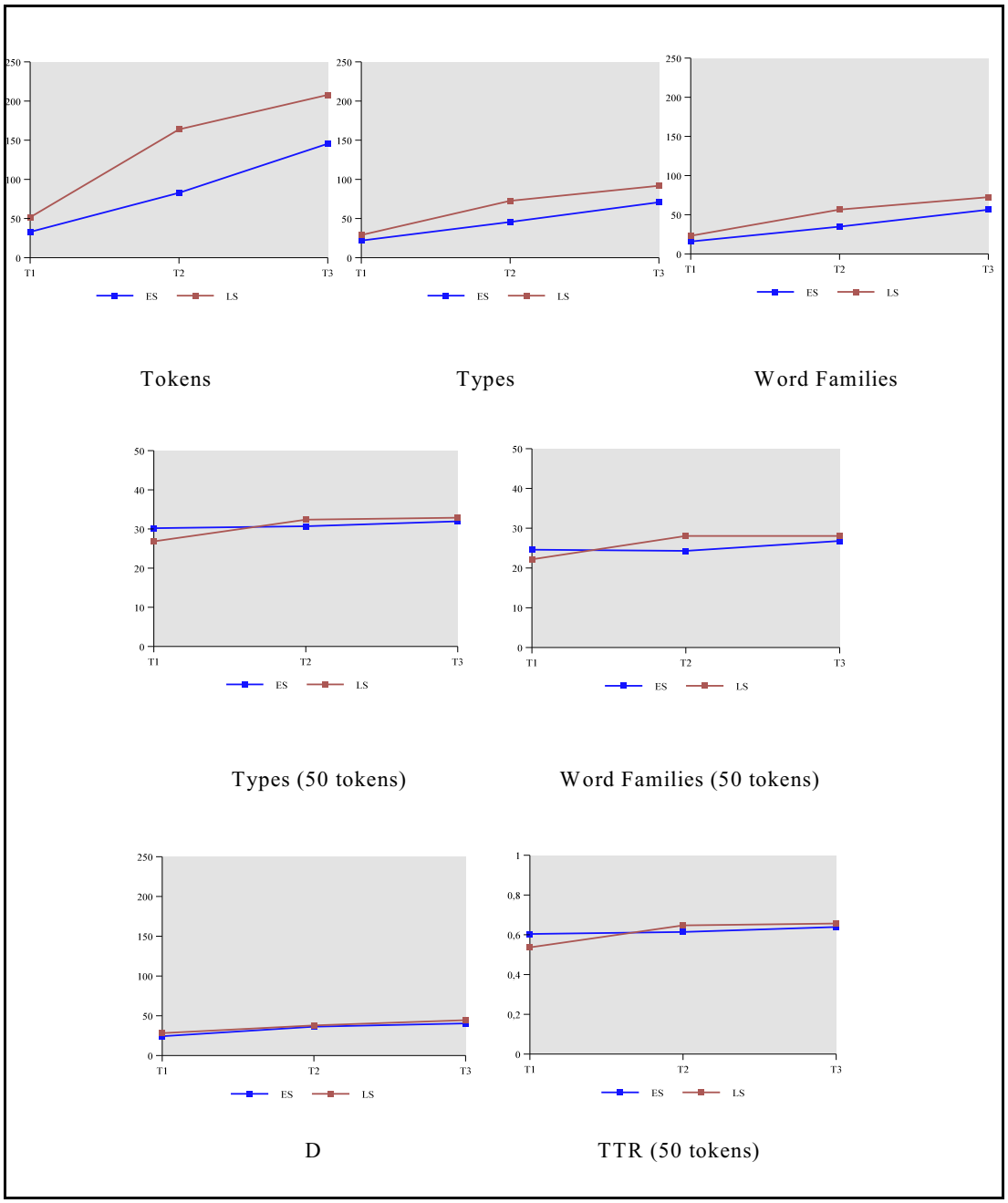

Figure 5.8. Interview: Times 1,2 and 3.

Concerning the interview and storytelling (Figures 5.8 and 5.9), it was found that after 200 hours, ES are better than LS in the standard measures (number of types and

word families as well as the TTR), this does not happen with D and the measures where length has not been standardised. In the roleplay, LS have higher groups means than ES in all measures, as shown in Figure 5.10.

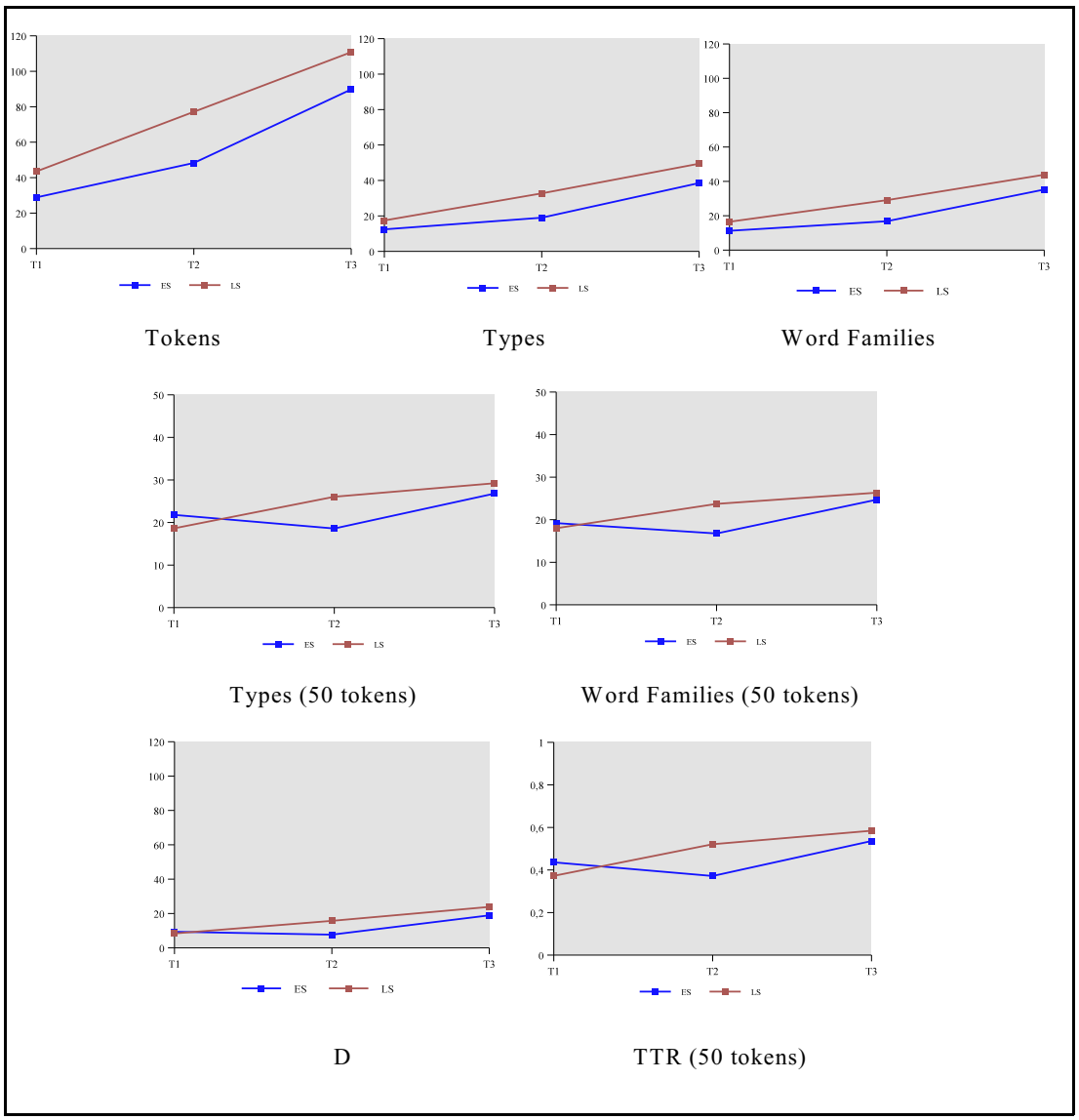

Figure 5.9. Storytelling: Times 1,2 and 3.

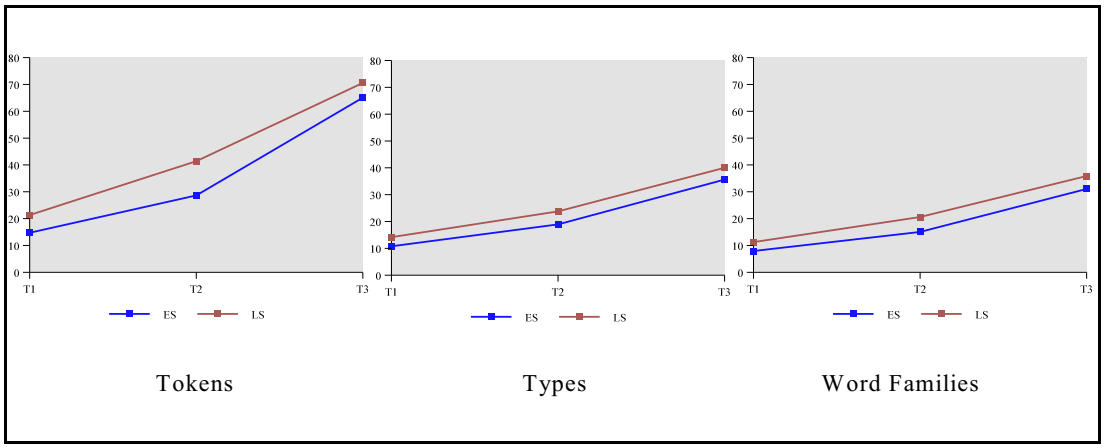

Figure 5.10. Roleplay: Times 1,2 and 3

The composition was the only task in which ES caught up with LS at Time 3 after 726 hours of instruction, even if the number of tokens was not kept constant. At Times 1 and 2, ES' performance was poorer: very few ES could produce 50 words at Time 1 and yet, at Time 3 their results were equal to those of LS. That is, after 416 hours of instruction, LS written vocabularies might not develop in the same way it was shown in the other tasks, while ES' vocabularies continued growing at a rate similar to that between Times 1 and 2.

The minor improvement of the LS group between Times 2 and 3 could be attributed to the scarcity of tokens of the LS group at Time 3 (i.e. short compositions), that is, their short productions prevent us from seeing any actual development in their written productive lexicon, although this lack of production itself is also indicative of their poor development. As the descriptive data shows, LS improve (as becomes evident, for instance, from the D index) but the growth is not as noticeable as the one achieved by ES.[38]

[38] An analysis of the compositions of three groups of adult learners with the same amount of exposure of the groups outlined above (200, 416 and 726h) also revealed that production between Times 2 and 3 remained quite stable (as the LS group): at T3 their compositions were not longer than T2, but an

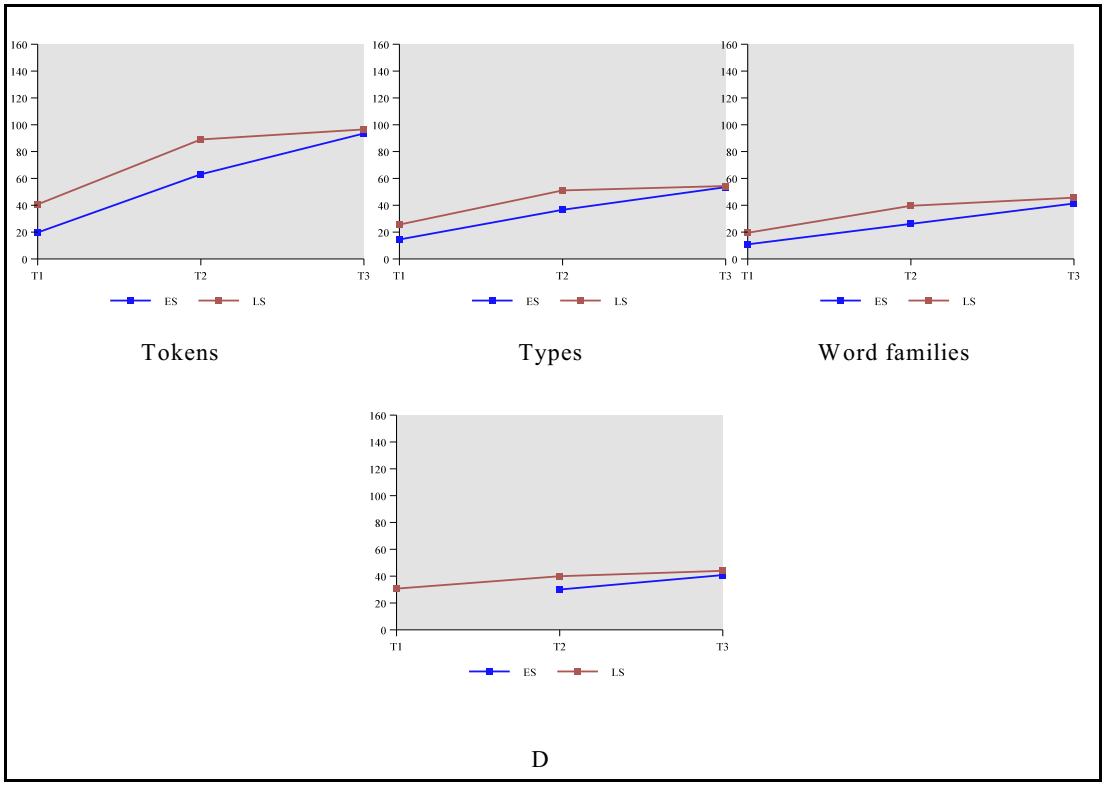

Figure 5.11. Composition: Times 1,2 and 3.

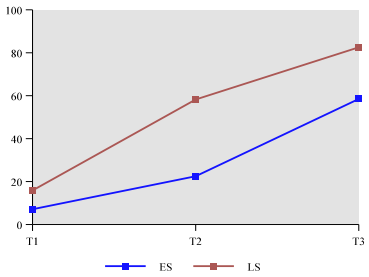

Figure 5.12. Cloze: Times 1,2 and 3.

Independent sample t-tests were conducted to draw meaningful comparisons according to age and linguistic formal exposure of the groups. The t-tests performed (alpha level .01 for multiple comparisons) reinforce what had been shown in 5.3.2 (see Table 5.14): There is a significant difference for Time between Time 1 and Time 2, as

improvement was detected in the D index.

there are significant differences between A1 and A2 and between B1 and B2 in most of the measures. With this analysis we also see that differences are significant between Times 2 and 3 for the ES (A2 and A3). However, this is not so for LS, as differences in the composition are not found between B2 and B3.

As regards Group, significant differences in favour of LS are found between groups that have the same amount of exposure; that is, between A1 and B1 on the one hand (200 hours) and A2 and B2 on the other (416 hours), and as we saw in 5.3.1 there were also significant differences between A3 and B3 in most tasks. It is also worth noticing that there are no systematic differences in the roleplay, while significant differences found in the storytelling are consistent.

Finally, the last rows of Table 5.14 show that significant differences between A2 and B1 and between B2 and A3 are more difficult to find than between other groups. Although there are differences in some measures for all tasks between A2 and B1, we only find some in the roleplay (and in the number of word families in the storytelling) between B2 and A3. As far as the cloze is concerned, significant differences are found between all groups except between A2 and B1 [(n=47,29) t(74)=1.978, p=.052][39] and between B2 and A3 [(n=21,53) t(72)=.040, p=.968].

[39] Tokens in the roleplay did not have a normal distribution in B1 and A2 and neither did the cloze in B1 and B2. A Mann-Whitney test confirmed the results of the t-tests for these variables that did not exhibit a normal distribution (that is, significance did not vary for the same variables and groups), except for the cloze when comparing A2 and B1 (z=2.32, p=.020).

		Non-standardised				Standardised		
		Tokens	Types	WF	D	Types	WF	TTR
A1 (N=31) **B1** (N=29)	I	t(58)=2.358 p=.022	t(58)=2.119 p=.038	t(58)=2.792 p=.007				
	S	t(57.91)=2.457 p=.017	t(58)=2.771 p=.007	t(58)=3.286 p=.002				
	R	t(58)=1.919 p=.060	t(58)=1.697 p=.095	t(58)=2.179 p=.033				
	C	t(58)=4.893 p=.000	t(58)=4.729 p=.000	t(44.53)=4.64 p=.000				
A2 (N=47) **B2** (N=22)	I	t(23.80)=3.842 p=.001	t(25.03)=3.929 p=.001	t(25.78)=4.233 p=.000	t(56.67)=.502 p=.618 (N=39,20)	t(57)=1.469 p=.147 (N=40,19)	t(57)=3.7 p=.000 (N=40,19)	t(57)=1.469 p=.147 (N=40,19)
	S	t(67)=4.010 p=.000	t(28.69)=4.934 p=.000	t(26.28)=5.253 p=.000	t(31)=4.850 p=.000 (N=16,17)	t(33)=6.039 p=.000 (N=17,18)	t(33)=6.705 p=.000 (N=17,18)	t(33)=6.039 p=.000 (N=17,18)
	R	t(67)=1.981 p=.052	t(67)=.1500 p=.138	t(28.86)=1.832 p=.077				
	C	t(27.97)=2.936 p=.007	t(27.46)=3.423 p=.002	t(26.27)=3.860 p=.001	t(52)=3.069 p=.003 (N=35,19)	t(51)=4.009 p=.000 (N=35,18)	t(51)=4.913 p=.000 (N=35,18)	t(51)=4.009 p=.000 (N=35,18)
A1 (N=27) **A2** (N=42)	I	t(67)=7.002 p=.000	t(67)=7.950 p=.000	t(67)=8.192 p=.000				
	S	t(60.01)=3.041 p=.003	t(67)=3.468 p=.001	t(67)=3.497 p=.001				
	R	t(65.32)=3.774 p=.000	t(66.59)=3.691 p=.000	t(67)=3.985 p=.000				
	C	t(67)=8.839 p=.000	t(67)=9.335 p=.000	t(67)=9.027 p=.000				
B1 (N=27) **B2** (N=22)	I	t(23,69)=5.486 p=.000	t(25.79)=6.449 p=.000	t(26.08)=6.654 p=.000	t(26)=3.425 p=.002 (N=8,20)	t(28)=3.661 p=.001 (N=11,19)	t(28)=4.715 p=.000 (N=11,19)	t(28)=3.661 p=.001 (N=11,19)
	S	t(47)=4.302 p=.000	t(30.27)=5.341 p=.000	t(31.99)=5.131 p=.000	t(26)=3.638 p=.001 (N=11,17)	t(24)=4.510 p=.000 (N=8,18)	t(9.14)=3.084 p=.013 (N=8,18)	t(24)=4.510 p=.000 (N=8,18)
	R	t(28.4)=2.597 p=.015	t(30.62)=2.636 p=.013	t(29.39)=3.042 p=.005				
	C	t(30.10)=5.353 p=.000	t(31.30)=5.871 p=.000	t(47)=5.789 p=.000	t(26)=2.6 p=.015 (N=9,19)	t(24)=2.651 p=.014 (N=8,18)	t(24)=1.410 p=.171 (N=8,18)	t(24)=2.651 p=.014 (N=8,18)
A2 (N=42) **A3** (N=40)	I	t(55.32)=4.64 p=.000	t(58.42)=5.33 p=.000	t(59.03)=5.91 p=.000	t(52.81)=1.240 p=.220 (N=35,37)	t(69)=.970 p=.335 (N=36,35)	t(69)=2.609 p=.011 (N=36,35)	t(69)=.970 p=.335 (N=36,35)
	S	t(69.93)=5.351 p=.000	t(59.3)=7.969 p=.000	t(56.73)=8.533 p=.000	t(41.75)=8.265 p=.000 (N=14,31)	t(45)=.7138 p=.000 (N=15,32)	t(45)=7.5 p=.000 (N=15,32)	t(45)=7.138 p=.000 (N=15,32)
	R	t(50.14)=4.965 p=.000 (N=42,38)	t(78)=5.855 p=.000 (N=42,38)	t(59.62)=6.544 p=.000 (N=42,38)				
	C	t(56.53)=4.515 p=.000 (N=42,39)	t(57.24)=5.487 p=.000 (N=42,39)	t(54.23)=5.788 p=.000 (N=42,39)	t(54.72)=4.137 p=.000 (N=31,35)	t(65)=2.917 p=.005 (N=31,36)	t(65)=3.315 p=.001 (N=31,36)	t(65)=2.917 p=.005 (N=31,36)

		Non-standardised				Standardised		
		Tokens	Types	WF	D	Types	WF	TTR
B2 (N=22) **B3** (N=41)	I	t(61)=1.934 p=.058	t(61)=2.759 p=.008	t(61)=3.078 p=.003	t(59)=2.331 p=.023 (N=20,41)	t(58)=.651 p=.517 (N=19,41)	t(58)=.005 p=.996 (N=19,41)	t(58)=.651 p=.517 (N=19,41)
	S	t(61)=3.735 p=.000	t(61)=5.304 p=.000	t(61)=5.324 p=.000	t(55)=4.962 p=.000 (N=17,40)	t(56)=3.665 p=.001 (N=18,40)	t(56)=3.090 p=.003 (N=18,40)	t(56)=3.665 p=.001 (N=18,40)
	R	t(61)=2.936 p=.005	t(61)=4.274 p=.000	t(61)=4.796 p=.000				
	C	t(55)=.655 p=.515 (N=22,35)	t(55)=.620 p=.538 (N=22,35)	t(55)=1.354 p=.181 (N=22,35)	t(45)=1.216 p=.230 (N=19,28)	t(44)=.081 p=.936 (N=18,28)	t(44)=2.338 p=.024 (N=18,28)	t(37)=.082 p=.935 (N=18,28)
A2 (N=47) **B1** (N=29)	I	t(74)=3.731 p=.000	t(74)=5.058 p=.000	t(74)=4.498 p=.000	t(32.99)=2.354 p=.025 (N=39,9)	t(50)=2.396 p=.012 (N=40,12)	t(50)=1.694 p=.097 (N=40,12)	t(50)=2.596 p=.012 (N=40,12)
	S	t(74)=.807 p=.422	t(74)=.950 p=.345	t(74)=.215 p=.830	t(25)=.365 p=.718 (N=16,11)	t(23)=.019 p=.985 (N=17,8)	t(23)=.746 p=.463 (N=17,8)	t(23)=.019 p=.985 (N=17,8)
	R	t(74)=1.675 p=.098	t(70.74)=2.151 p=.035	t(74)=2.129 p=.037				
	C	t(74)=4.435 p=.000	t(74)=4.558 p=.000	t(74)=3.435 p=.001	t(42)=.163 p=.871 (N=35,9)	t(41)=.282 p=.779 (N=35,8)	t(41)=2.294 p=.027 (N=35,8)	t(41)=.282 p=.779 (N=35,8)
B2 (N=22) **A3** (N=57)	I	t(77)=.888 p=.377	t(77)=.270 p=.788	t(77)=.024 p=.981	t(71)=1.029 .307 (N=20,53)	t(68)=.369 .713 (N=19,51)	t(68)=1.444 .153 (N=19,51)	t(68)=.369 .713 (N=19,51)
	S	t(77)=1.411 p=.162	t(77)=1.814 p=.074	t(77)=2.190 p=.032	t(61)=1.876 p=.065 (N=17,46)	t(63)=.691 p=.492 (N=18,47)	t(63)=.982 p=.330 (N=18,47)	t(63)=.691 p=.492 (N=18,47)
	R	t(74)=2.339 p=.022 (N=22,54)	t(74)=3.048 p=.003 (N=22,54)	t(74)=3.097 p=.003 (N=22,54)				
	C	t(76)=.438 p=.663 (N=22,56)	t(76)=.522 p=.603 (N=22,56)	t(76)=.428 p=.670 (N=22,56)	t(65)=.292 p=.771 (N=19,48)	t(66)=.1811 p=.075 (N=18,50)	t(66)=.1635 p=.107 (N=18,50)	t(66)=.1811 p=.075 (N=18,50)

Table 5.14. t-tests results when $p \leq .01$.

A one-way Manova was also conducted as it is recommended when comparing mean scores of more than one variable in more than two groups. It is considered that a series of one-way Anovas or t-tests might inflate Type I error (i.e. believing that there are significant differences where actually there are not). The Manova results confirmed those obtained in the t-tests in groups B1-B2, B2-B3, A2-B2, A2-A3, A3-B2. Regarding A1-B1, A1-A2 and A2-B1, differences that were significant in the t-tests did not reach significance when conducting the Manova test, as it was expected.

5.3.4. D and other measures

Pearson correlations were performed with a twofold purpose: to check how D was related to other measures of lexical richness and especially to see if D correlated with the number of tokens in each task and was therefore affected by text length. The correlations were calculated with all groups so as to have the maximum range. The results of these correlations are displayed in Table 5.15.

		Non-standardised					**Standardised**		
		Tokens	Types	WF	TTR		Types	WF	TTR
D int	N=182	.166* p=.025	.368** p=.000	.374** p=.000	.365** p=.000	N=178	.657** p=.000	.548** p=.000	.657** p=.000
D story	N=148	.410** p=.000	.796** p=000	.784** p=.000	.745** p=.000	N=145	.919** p=.000	.884** p=.000	.919** p=.000
D role	N=83	-.016** p=889	.351** p=.001	.409** p=.000	.605** p=.000	N=81	.814** p=.000	.754** p=.000	.814** p=.000
D comp	N=153	.262** p=.001	.503** p=.000	.500** p=.000	.513** p=.000	N=151	.750** p=.000	.525** p=.000	.750** p=.000

Table 5.15. Correlations between D and other measures in each of the tasks.
* Significant correlation at the .05 level
** Significant correlation at the .01 level

The correlations between D and the other measures in each task have a similar pattern. Positive moderate-strong correlations are found between D and the number of types, word families and also TTR, especially when length is standardised, e.g. in the storytelling [r=.919, N=145, p<.01]. The behaviour of D and TTR, the latter computed with standardised length, can be appreciated in Figure 5.13. Very small correlations are found between D and the number of types and word families in the interview and role when length is not standardised. There is no correlation either between D and the

number of tokens produced in any of the tasks, except from a small one in the storytelling [r=.410, N=148, p<.01].

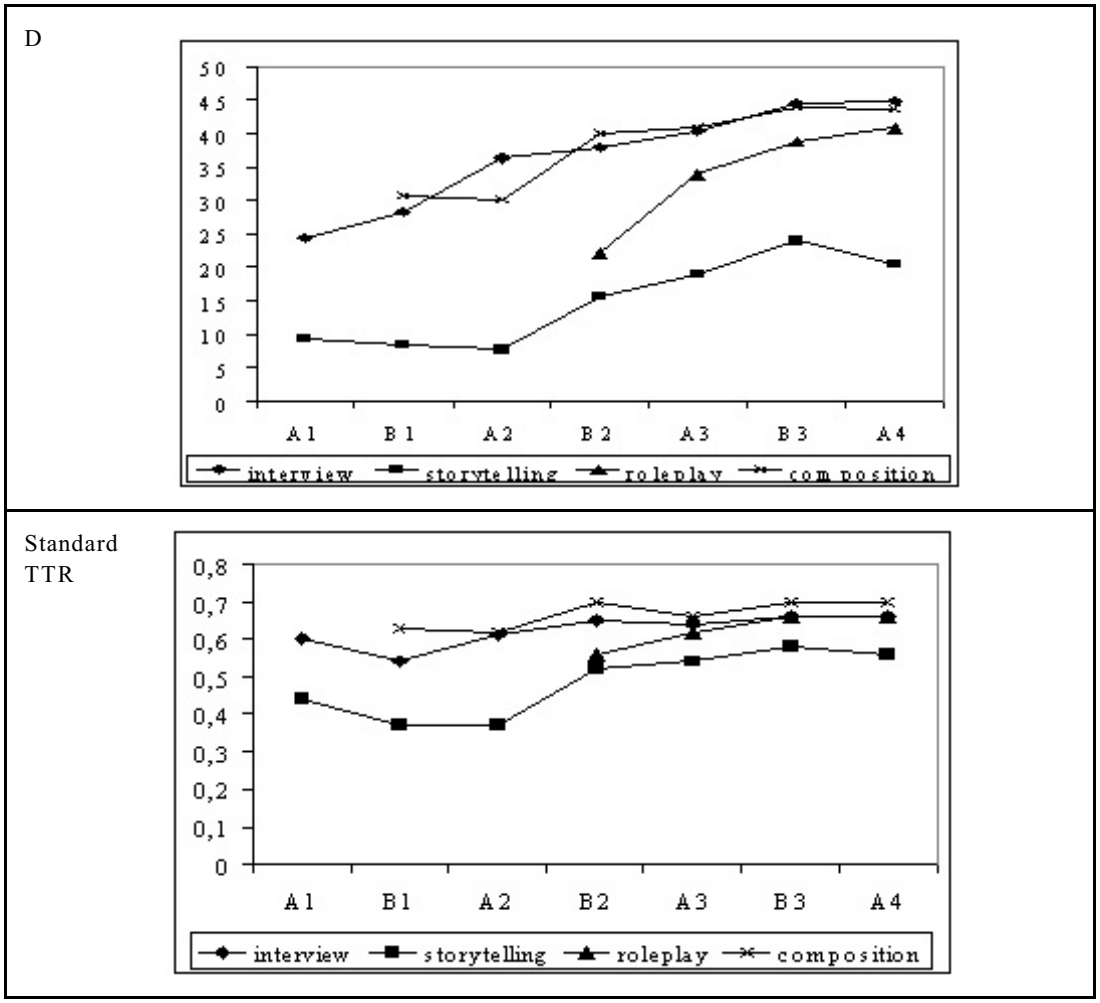

Figure 5.13. A comparison of Ds and standard TTRs (50 tokens) in each task and group.

In Table 5.16 a comparison is drawn between D results obtained in this study for each group and task (shown in the 'author' column by 'M' -Miralpeix- and in grey in the table) and the indications given by very recent studies that have used D as a measure. Those groups in our study that do not appear in the table did not produce enough words for D to be computed, as described above. The majority of the indications

presented come from Malvern et al. (2004), whose results were obtained from the analysis of different corpora of children acquiring English as an L1, from two studies with SL learners and from another with adults (for academic writing). The other studies correspond to Read (2005) -'R' in the table-, who analysed the speaking part of the IELTS test from students with different L1s; to Daller and Xue (2007) -'D&X' in the table-, who recorded the performance of Chinese learners of English at university in picture descriptions; to van Hout and Vermeer (2007) -'VH&V' in the table-, who analysed oral data from native Dutch speakers and from learners of Dutch and finally to Jarvis (2002) -'J' in the table'-, who compared the written performance of English NS and Finnish and Swedish learners of this language. If no indication is given next to the language, it means that it refers to English. In the case of the study by Read (2005), the L1 of the speakers is not acknowledged in the table as examinees come from a variety of countries and backgrounds.

Age (years)	N	Mean	*sd*	Min	Max	L1/L2	Type of Language	Author
12.9 (A2)	17	**7.67**	4.48	1.83	19.24	L2 (L1: Sp/Cat)	oral (storytelling)	M
13 (B1)	8	**8.37**	5.52	2.89	20.25	L2 (L1: Sp/Cat)	oral (storytelling)	M
10.9 (A1)	5	**9.36**	4.45	6.44	15.85	L2 (L1: Sp/Cat)	oral (storytelling)	M
15	18	**14.80**	10.31	1.48	36.99	L1	oral	D et al.
15 (B2)	18	**15.72**	5.02	7.40	24.70	L2 (L1: Sp/Cat)	oral (storytelling)	M
13	35	**18.55**	-	-	-	L2 (L1: Swedish)	written (narrative)	J
16.3 (A3)	47	**18.90**	6.26	5.50	33.50	L2 (L1: Sp/Cat)	oral (storytelling)	M
17.7 (A4)	14	**20.49**	7.40	7.37	30.36	L2 (L1: Sp/Cat)	oral (storytelling)	M
17	20	**21.49**	16.70	2.60	67.24	L1	oral	D et al.
11	35	**22.02**	-	-	-	L2 (L1:Finnish)	written (narrative)	J
15 (B2)	6	**22.11**	9.98	13.79	42.22	L2 (L1: Sp/Cat)	oral (roleplay)	M
15	35	**23.02**	-	-	-	L2 (L1:Finnish)	written (narrative)	J
17.9(B3)	40	**23.86**	5.91	13.30	38	L2 (L1: Sp/Cat)	oral (storytelling)	M
10.9 (A1)	5	**24.29**	12.02	10.30	37.37	L2 (L1: Sp/Cat)	oral (interview)	M
12	22	**24.69**	-	-	-	L1	written (narrative)	J
13	35	**25.64**	-	-	-	L2 (L1:Finnish)	written (narrative)	J
15	35	**26.62**	-	-	-	L2 (L1:Swedish)	written (narrative)	J
2	28	**27.44**	20.52	2.50	84.64	L1	oral	D et al.
10	22	**27.68**	-	-	-	L1	written (narrative)	J
13 (B1)	12	**28.18**	5.52	15.85	35.88	L2 (L1: Sp/Cat)	oral (interview)	M
24	24	**28.59**	5.57	-	-	L2 (L1: Chinese)	oral (picture description)	D&X
14	22	**29.36**	-	-	-	L1	written (narrative)	J
15	35	**29.36**	-	-	-	L2 (L1:Finnish)	written (narrative)	J
12.9 (A2)	35	**30.05**	12.16	11.35	64.77	L2 (L1: Sp/Cat)	written (composition)	M

Age (years)	N	Mean	*sd*	Min	Max	L1/L2	Type of Language	Author
13 (B1)	8	**30.74**	6.41	19.26	39.20	L2 (L1: Sp/Cat)	written (composition)	M
16.3 (A3)	33	**33.77**	13.49	9.90	73	L2 (L1: Sp/Cat)	oral (roleplay)	M
2.2	29	**34.77**	17.70	7.48	65.76	L1	oral	D et al.
23.3	26	**36.22**	7.59	-	-	L2 (L1:Chinese)	oral (picture description)	D&X
12.9 (A2)	40	**36.27**	16.46	13.69	89.99	L2 (L1: Sp/Cat)	oral (interview)	M
15 (B2)	19	**37.85**	7.67	17.96	55.12	L2 (L1: Sp/Cat)	oral (interview)	M
17.9 (B3)	28	**38.85**	12.71	21.90	88.70	L2 (L1: Sp/Cat)	oral (roleplay)	M
25	38	**39.51**	14.12	4.69	63.37	L1	oral	D et al.
15 (B2)	19	**39.98**	9.66	21.57	58.43	L2 (L1: Sp/Cat)	written (composition)	M
16.3 (A3)	50	**40.35**	9.81	22.10	70.60	L2 (L1: Sp/Cat)	oral (interview)	M
17.7 (A4)	9	**40.79**	15.39	17.39	62.87	L2 (L1: Sp/Cat)	oral (roleplay)	M
16.3 (A3)	50	**40.80**	10.63	15.70	78.50	L2 (L1: Sp/Cat)	written (composition)	M
25	29	**41.53**	16.93	4.05	69.67	L1	oral	D et al.
27	29	**43.67**	15.45	10.38	77.88	L1	oral	D et al.
17.7 (A4)	14	**43.71**	11.63	26.49	71.11	L2 (L1: Sp/Cat)	written (composition)	M
17.9 (B3)	28	**43.97**	11.84	28.90	68.50	L2 (L1: Sp/Cat)	written (composition)	M
17.9 (B3)	41	**44.53**	11.62	21.10	68.90	L2 (L1: Sp/Cat)	oral (interview)	M
17.7 (A4)	14	**44.85**	11.60	18.64	69.34	L2 (L1: Sp/Cat)	oral (interview)	M
3	29	**47.83**	13.97	13.26	69.95	L1	oral	D et al.
32	30	**49.48**	15.41	11.22	80.78	L1	oral	D et al.
35	29	**53.12**	13.55	10.57	73.54	L1	oral	D et al.
16	27	**56.28**	14.87	29.74	87.35	L2 (L1:French)	oral (GCSE, conversation)	D et al.
18-30	32	**56.58**	12.10	35.78	91.99	L2 (L1:French)	oral (decision making task)	D et al.
Adults	14	**60.70**	11.40	37.50	76.10	L2	oral (IELTS speaking test)	R
Adults	21	**63.40**	11.30	39.50	86.70	L2	oral (IELTS speaking test)	R

Age (years)	N	Mean	*sd*	Min	Max	L1/L2	Type of Language	Author
5	15	**64.02**	8.46	50.83	83.30	L1	oral	D et al.
Adults	18	**67.20**	16	57	81.40	L2	oral (IELTS speaking test)	R
Adults	17	**71.80**	18.20	61.20	89.50	L2	oral (IELTS speaking test)	R
7.11	16	**72.84**	11.48	-	-	L2 (Dutch)	oral (definition task)	VH&V
7.11	16	**75.17**	12.04	-	-	L1 (Dutch)	oral (definition task)	VH&V
Adults	11	**79**	4.90	87.50	72	L2	oral (IELTS speaking test)	R
Adults	23	**90.59**	10.79	69.74	119.20	L1	written (academic)	D et al.

Table 5.16. D results obtained by different researchers compared to the ones obtained in the present study.
Key 'Author': D&X=Daller & Xue (2007), D et al.=Durán et al. (2004), J=Jarvis (2002), M=Miralpeix (this study), R=Read (2005), VH&V=Van Hout & Vermeer (2007).

5.4. Discussion

In this sub-study we set out to explore if, towards the end of secondary education, students who had been learning English from the age of 8 in a formal context outperformed the students who had started it at 11, as regards oral and written productive vocabulary (Times 3 and 4). We also wanted to investigate the lexical development during the previous years in primary and secondary school (Times 1 and 2).

Results show that, as regards free productive vocabulary, after 726 hours of formal exposure, learners who started learning English earlier did not outperform those who started three years later: LS tend to obtain higher results in most of the tasks. We have instances in which the differences do not reach significance, although LS

productive vocabulary is shown to be a bit more diverse. There are two tasks where the advantage of the older group is just shown in some of the measures (i.e. the interview and the composition), in the roleplay the two groups perform similarly and in the storytelling the differences in favour of LS are evident. The same holds for controlled productive vocabulary: LS obtain significantly better results in the cloze test. Therefore, these findings are consistent with previous research in other formal contexts (Burstall *et al*. 1974; Oller & Nagato, 1974; Singleton, 1999), which do not provide evidence in favour of the ES either. Interestingly enough, with the exception of two measures, ES do not surpass LS either even if the former are given one year more of exposure (there are no main differences in favour of A4 if it is compared with B3). We will return to these results in the final discussion chapter of the present dissertation.

An advantage of LS over ES in lexical knowledge seems to be present since the first stages of learning the FL as pointed out in the analyses in 5.3.2. and 5.3.3. LS lexical gains after 200 and 416 hours of school instruction are superior to those of ES after the same number of hours of exposure. It is worth noticing that the point in time when one group diverges more from the other is at Time 2: after having received 416 hours of exposure, when A2 is 12.9 years old and B2 is 15. In addition, the fact that A2 lexical competence is similar to B1 and that B2 performs similarly to A3 (as very few significant differences are found between these groups) indicates that AO is not as important in determining the students' performance as one might think: at about 13 years of age (no matter if AO is 8 or 11) and at 15-16 (again regardless of the AO) productive vocabulary knowledge seems to be at the same stage. LS with 416 hours are not far away from ES with 726 hours. LS' gains in productive vocabulary knowledge

(both free and controlled) seem to be a bit larger between T1 and T2 (that is, between 13 and 15 years of age) while for ES they are found between T2 and T3 (between 13 and 16). Therefore, this would point out that from 13 onwards, the lexical development is a bit more consistent in both groups, independently of the AO and the amount of exposure that has been received. It is as if cognitive maturity outweighed an early AO and the hours of instruction received in the formal context. These results are then an indication that, given similar opportunities, efficiency in SL learning increases with age as regards productive vocabulary knowledge as well, which is in the same line with the findings of most studies on the age factor and morphology or syntax (Harley, 1986; Snow and Hoefnagel-Höhle, 1978).

Other findings from the BAF project corroborate the results obtained above: ES ask for assistance in the oral tests more often than LS do (see Grañena, 2006), which may also be seen as an indicator of their poorer lexical competence (i.e. asking for help in vocabulary they do not know). In the written data, Navés, Torras and Celaya (2003) also show that LS do normally produce more and that significant differences in favour of the ES are very rarely found (see also Torras et al., 2006).

Our results are also in line with other long-term studies carried out in the Basque Country. As pointed out in chapter 2, in a similar context to ours -where Basque/Spanish bilinguals learn English as an L3- after 6 years of EFL instruction, LS (starting at 11 and being 16 when they were tested) significatively outperformed ES (starting at 8 and being 13 when they were tested).

It can also be seen in our Time 3 data that the interview and the composition show length effects. In the interview, differences are found between the groups in the

amount of tokens, types and word families when length is not set at 50 tokens, but there are no differences when length is kept equal. This might be due to the nature of the task: the interviewer introduces new topics, which may act as a trigger for more lexical variety in this oral data, something which cannot be shown when length is set at 50 tokens. With the help of the interviewer, who asks and provides topics to talk about, it is easier for students to talk more, which does not happen in a task where the interviewer does not intervene or does not have such an active role, as in the storytelling. Therefore, LS have more lexical variety that can only be shown in longer texts. It should also be taken into account that, apart from triggering oral production, the interviewer may also be considered a 'levelling' factor, as s/he can contribute, for instance, to put on a higher level a learner that might have difficulties in oral production, and hence to make the learner's contribution better. This would be a possible explanation for the results obtained in 5.3.2, where we find significant differences in the tasks where there is no interaction with an interviewer (the story, the composition and the cloze).

The results from the storytelling and roleplays are not so dependent on length as the ones obtained from the interview and the composition, as most of the measures do not give different results when length is not kept constant. Of all the oral tasks, the storytelling may be the most adequate to assess productive vocabulary knowledge. Firstly, the results do not vary much depending on length. Secondly, it elicits more words from the students than the roleplay. Probably, the reason why the two groups performed in such a similar way in the roleplay is that it is a dual task and they have to take into account the limitations not only of their own lexical resources but also those

of their partners'. Therefore, more proficient learners might adapt to the demands of a low-proficient partner (use of a less varied vocabulary, asking or answering using very short utterances...). As they are not performing the task alone and they do not have any planning time, like in the composition or the storytelling, it is also more difficult for them to think just after their partner's turn of what they are going to say, which might also be the reason why it is the task where they produce fewer tokens. Probably, the roleplay might not be an adequate task to assess students' production in this study as age might have a direct influence on the learners' behaviour. That is, the argumentative abilities of a ten-year-old are different from a seventeen-year-old: it will be more difficult for the first one to negotiate on how to celebrate a party than to answer the questions of an interviewer or to tell a story. For a teenager the negotiation might be easier and, as s/he has more argumentative resources, the production will increase. However this does not mean that the young child does not know any vocabulary either, but that s/he has less strategies to convince the partner and therefore less opportunities to show his/her lexical repertoire.

As regards TTRs, there is variation in the results when length is not kept constant. Sensitivity to text length is clearly seen in the interviews, where LS produce more types and tokens than ES (and where we find the biggest difference between the group means in these measures). However, it is ES who have a higher TTR. This finding is in line with what has been repeatedly shown in the literature, that is, the dependency of TTR on the number of tokens (Lenko-Symanska, 2002; Richards, 1987; Vermeer, 2000). The obvious solution is to compute TTR with a fixed length, as the majority of studies in the literature do.

If length is standardised to compute measures such as TTRs, amount of types or word families, the results can be misleading with low-level learners. In the interview and storytelling, when length is kept constant, ES are better than LS after 200 hours of exposure (see the standard measures in Figures 5.8 and 5.9) . However, this is not the case if we have a closer look to the whole data: LS produce more tokens and their vocabularies are more varied or at least equally varied.

As a way to illustrate this, we can compare data from two representative learners, one in the ES group and one LS. Table 5.17 shows that if we examine the figures of these two learners, the first three rows seem to point out that the ES subject is better than the LS:

Measures	**ES**	**LS**
Number of types in 50 tokens	26	19
Number of wf in 50 tokens	21	19
TTR with 50 tokens	.52	.38
Total number of tokens	54	79
Total number of types	27	26
Total number of word families	22	24

Table 5.17. Results for two learners (after receiving 200h of instruction) in different measures.

However, by studying the following rows it can be noticed that the LS produces more tokens and that the total amount of types is nearly the same as the ES. The explanation lies in the telegraphic style of the ES, which makes the amount of types raise considerably if the length of the task is standardised. The first 20 words of the story by the ES and LS learner, respectively, are the following:

ES

"The girl and boy have got the bread jam and sandwich. The mum. The mum is the children is looking at the children. They are going. The girl and boy are mountain. The dog jumping of the basket in xxx [...] in the basket no there are sandwich [...]. Sandwiches the dog. The dog eating sandwich [...] sandwiches."

LS

"The mother takes a tea and the xxx and brother and sister and yes her the mother speak with sister and brother and the dog see the eat. The mother speak goodbye with sister and brother and sister and brother are in the xxx [...] in the mountain and the with. The brother and sister they look the dog in the basket and the [...] and they see the eat not [...] xxx they see the eat not. The dog he is eat the sandwich."

Hyltenstam (1988) already pointed out that high values for LV (Lexical Variation: the TTR for content words) do not always reflect a real effective lexical variation, but may in fact be the result of not constructing a coherent text. Here, as the ES uses few function words and does not repeat lexical words to make a more coherent discourse, the number of types used is 21, while we find only 19 in the LS' production. Therefore, although standardising length is necessary when computing TTRs, we should be aware of the fact that with very low level learners, keeping length constant for the calculation of some measures, such as the number of types or the TTR can be deceptive.

In the compositions, even if length is not standardised, both groups have similar results at Time 3; it even seems that LS do not progress between Times 2 and 3. However, it is not that LS are not developing their lexical competence, but it is possible that, as they produce less than in the other tasks, this prevents us from seeing this development. The vocabulary used in the interview and composition is quite similar (the former is oral and the latter is written). Students talk about their hobbies, family, habits...

Nevertheless, the interviewer gives the prompt so that the learner can speak, while in the composition there is no prompt and they write less than they talk. It can be said, then, that in the composition there is a ceiling effect in the use of vocabulary that prevents us to see whether there are any advances. However, if Ds are computed, a slight improvement can be observed between 416 and 726 hours. What can be said therefore about D in relation to the other measures?

We notice a resemblance between the results given by the TTR with length standardised and D values. This might seem to contradict Richards and Malvern (2000) results, because they found no correlation between D and TTR and they claimed to be expecting this lack of correlation. Our correlations between D and TTR without keeping length constant are moderate, but they get stronger if we correlate D and the TTRs computed after standardising length. This close relationship can be appreciated in Figure 5.13, although the scales for the two measures are different, the pattern presented by both measures is very similar. Actually, the result should not come as a surprise, as TTR is contained within the formula for D, so they really should correlate. This fact makes us wonder about the necessity of D and of a curve-fitting approach if the results given by TTRs and Ds are correlated. Actually, D itself might be also considered a variation of the TTR -at some point we do take into account the relationship between types and tokens- (Vermeer, 2004; McCarthy & Jarvis, 2007). Nevertheless it is theoretically more valid and, in addition, much more appropriate for practical purposes as there is no need to standardise text length and discard part of the data. In addition, as noted above, it can be more useful than TTR to analyse low-level learners' production (notice the different results for groups A1 and B1 in these measures in the interview in Figure 5.8: while

other measures, D among them, point at B1's superiority, TTR provides a different result in favour of A1). Apart from this, the positive correlations between D and the number of types confirms its potential as a measure of lexical variation, maybe more useful at these elementary stages than LD. The lack of correlation between D and amount of tokens shows that this measure is not affected by text length as TTR is.

Finally, if the Ds in these studies are compared with those obtained by other researchers in recent studies, it can be appreciated that in no task do these learners score higher than a 3-year-old child in his/her L1. It is also worth noticing that the GCSE conversation, performed by learners of French of about 16 years old, produced a mean D of 56.28. Of the results reported by Durán et al. (2004), this is the one that most resembles the task performed by our learners (the interview) and whose age group is most similar to ours (16 vs. 16.3). While the mean D for the French learners is of 56.28, the mean for the oral interview in A3 is 40.35. The oral decision making task, which is similar to our roleplay, has a mean D of 56.58 while our oldest students score 40.79. While all these comparisons do not necessarily mean that the learners in this study are better/worse than the others or that their SL is poorer than a 3 year-old child's L1 (the tasks performed are not the same either although it is oral language), it is worth noting that if a standard measure like D was adopted instead of TTRs, it would be much easier to establish connections between students' levels in different settings.

Furthermore, two other observations need to be made in relation to the results displayed in Table 5.16. First, D seems to discriminate between different levels: our learners' mean D (in any of the tasks) is always much lower than the mean D for students taking the GCSE or the IELTS test. Second, it might be that D results

systematically vary depending on the language in which it is computed: an oral definition task performed by children in Dutch (either if it is their L1 or their L2), always give very similar results (around 70) to those obtained by adults in their English L1. These are issues that should be further explored. Obviously, in order to elaborate a more reliable standard scale of the sort presented in Table 5.16, more studies should make use of this measure with different learners and in a variety of contexts and outputs (for instance, there is only one reference to academic writing but not L2 writing), and *D_Tools* could contribute to that end: Read (2005) can be an example, he used this program to compute D in 88 IELTS speaking tests from examinees in 14 countries around the world, thus obtaining several D values for each of the bands score levels.

5.5. Conclusion

In summary, we have seen that, regarding productive vocabulary, ES did not obtain better results than LS in spite of their earlier exposure to the FL. Whether there are advantages that these measures could not reveal remains to be explored. For example, differences in other lexical abilities (receptive vocabulary, speed of retrieval...) may be shown by further research. From a pragmatic perspective, more information could be obtained that might also offer an account of lexical performance; i.e. speed at answering, silences or, as suggested in Lorenzo-Dus and Meara (2005), the time spent on the topics selected for conversation in the oral tasks.

In this study, we have used intrinsic measures, that is, the assessment has been carried out in terms of the words that appear in the text itself. This has given us a gross

indication of learners' levels, which can be further explored by using extrinsic measures of vocabulary (Daller, van Hout & Treffers-Daller, 2003; Laufer & Nation, 1995; Meara & Bell, 2001; Vermeer, 2004), i.e. classifying items according to criteria external to the text itself would help us to study the vocabularies of these two groups of learners, which will be carried out in chapter 6.

It was also our purpose to analyse how the D measure was related to other vocabulary measures and to check if it was useful to describe the vocabularies of our learners, who are not at an advanced level and produce short texts. As productive vocabulary knowledge is not always easy to assess, especially when we deal with oral production from low-level students, it is worth exploring new ways to analyse the data. D is a measure that has not been widely used, but the fact that it can become a standard index which can work in a variety of contexts and languages suggests that it can be a more adequate measure of lexical diversity than TTR. Thus, curve-fitting approaches can solve some of the problems that traditional lexical richness measures have shown to have. We are also well aware of the limitations of a curve-fitting approach (Jarvis, 2003), and we acknowledge the fact that the reliability of the D index may not be the optimal one (McCarthy & Jarvis, 2007), but it can be regarded as a good start to explore how lexical diversity might be approached in the future.