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 7

VOCABULARY SIZE ESTIMATES

7.1. Introduction

This chapter contains three main sections. In the first, a description of *V_Size* is offered; *V_Size* is a program that estimates vocabulary sizes. In the second, different explorations with this program are carried out as a way of evaluating this tool. In the last section, we use *V_Size* to estimate the productive vocabulary that three groups of learners -A3, B3 and A4- present in four different tasks. The chapter closes with a discussion of the results obtained and with some considerations on *V_Size* as an estimation tool.

7.2. *V_Size*

As it has been advanced in chapter 3, vocabulary estimates have been obtained in the past using different procedures: from multiple choices or yes/no tests to translation tests. The common technique in all these estimation methods, the majority of which deal with receptive vocabulary, is to establish different bands according to a set criteria (frequency, difficulty, etc.) and to choose a representative sample of words in each band

to test. Then, depending on the number of words 'known' in each band an estimate is calculated. The process has been refined along the years and it has been standardised in tests like the Eurocentres Vocabulary Size Tests -EVST- (Meara & Jones, 1988) or the Vocabulary Levels Test -VLT- (Nation, 1990). As it has been shown (see Table 3.1), extrapolations of the vocabulary size on the basis of the VLT have been conducted very often.

As regards productive vocabulary, the LFP has been used to identify the proportions of frequent vs. unfrequent words that learners choose to use in their writings. As described in chapter 5, it gives an idea of vocabulary use over several levels (Laufer & Nation, 1995). Recently, the potential of LFP to inform estimates of productive vocabulary has been acknowledged by different researchers (Meara, 2005; Edwards & Collins, 2007). It is in this framework that *V_Size* was created to go a step further; that is, bearing in mind that profiles could be obtained from the learners' production and that there are some laws that language tends to obey, the program uses a mathematical model to infer vocabulary sizes, as it is presented in the next section.

7.2.1. What the program does

The program performs two main operations: it can estimate the vocabulary size for a particular vocabulary profile that is obtained from a text produced by the learner and it can also give an estimate profile for a particular vocabulary size. As it has been seen in chapter 6, a vocabulary profile is composed by a set of data points (five in *V_Size*), which are used by the program to estimate the vocabulary size of the learner for

a particular task. It does so by comparing the curve of the learner's profile with the curves of theoretical ideal profiles generated by the logarithmic function, as will be shown below. The comparison, which aims at predicting how many words the learner knows productively, is carried out through a process of curve-fitting (with the Least Squares method).

7.2.2. How the program works

In order to understand how the program works, it is necessary to define what the Power Law is and how the method of the Least Squares works.[46]

The Power Law is a scientific rule that can be observed in many kinds of phenomena. It is defined as a relationship between two variables such that one is proportional to a power of the other. Put in other words, it is when a particular fact takes place more often than another, when there are a few entities that get a lot of something (or score very high) while a medium number get little and a huge number score very low. The Power Law has been exploited for research in many fields, there are many examples of collections that approximately obey this law, for instance the size of earthquakes, income distribution, city populations or website use. If it is applied to language, it is best known as Zipf's Law (Zipf, 1935), as an American philologist with this surname was the first to demonstrate that the frequency of a word is roughly inversely proportional to its rank in its frequency table, which can be mathematically expressed

[46] It is the same process of curve-fitting used in chapter 5 when computing the D index. For more specific information about the process see, for instance, Spiegel and Stephens, 1999 (chapter 13).

as (r . f = C)[47]. As an example, an extract from the rank frequency list for the BNC (Leech, Rayson & Wilson, 2001:120), is shown in Table 7.22. In the table, the 30th most frequent word in the corpus ('from'), the 32nd most frequent word ('that'), the 34th most frequent word ('or') etc. are presented. When the rank of each word is multiplied by the word frequency, the values obtained are similar, they tend to be constant (around 120,000 in the example).

r	*x*	*f*	=	*C*
30th	from	4,134		30 x 4,134= 124,020
32nd	that	3,797		32 x 3,797= 121,344
34th	or	3,707		34 x 3,707= 126,038
36th	's	3,490		36 x 3,490= 125,640
38th	n't	3,328		38 x 3,328= 126,464
40th	as	3,006		40 x 3,006= 120,240

Table 7.22. Zipf's law exemplified with an extract from the BNC.

It is interesting to note that Zipf's Law scale is invariant and it can manifest itself in texts of any length. Zipf's Law holds true for all texts and for texts written by any author. There have also been studies that confirm that this law holds true for random texts, i.e., monkey-typing texts also inevitably exhibit a Zipf's-law-like frequency distribution as oral language does as well: Ridley and Gonzales (1994) noted that corroborations of Zipf's Law had been derived from large samples of written speech (Zipf's proposed a length of about 5,000 words to demonstrate the law). They took

[47] Where (r) is the Rank number, (f) is Frequency and (C) is approximately a Constant. With this formula it can be seen that there are words that appear very often in any language while others appear just sometimes and many are almost never used.

written and speech samples of about 400 words and found that the law was also confirmed in these short texts. However, their second aim, which was to find authors' identities by finding systematic deviations from the law in each individual, was not fulfilled, they claim that "the relationship between the frequencies of the words used and the number of different words at those frequencies [i.e. Zipf's Law] appears to be so robust that it is extremely difficult, if not impossible, to find any distinguishing characteristics of speakers, as defined in terms of deviations from Zipf's Law" (Ridley & Gonzales, 1994:154).

Zipf's Law can be considered the cornerstone of *V_Size*, as this tool assumes that we can model text production by weighting each word according to its frequency and selecting at random from a weighted list. This program 'observes' this law in the following way: it uses the Logarithmic Randomisation Function [Ln(rankfreq)*1000] to generate a set of 25,000 word idealised profiles, based on different values of vocabulary size. That is, profiles are generated by choosing words at random using the logarithmic transformation of frequency. For instance, it generates profiles produced by vocabularies of 1,000 words, of 2,000 words, of 3,000 words...[48]

Then, it finds the profile that best matches the profile of the learner we want to estimate the vocabulary size of. It does so by a process of curve-fitting: it compares the empirical curve of the learner's profile to the set of theoretical profile lines that the program has generated. The Least Squares or Ordinary Least Squares (OLS) technique

[48] E.g: if we deal with a vocabulary size of 3,000 words (whose [Ln(rankfreq)*1000] is 8,006), we select a random number between 1 and 8,006. As [Ln(rankfreq)*1000] of 1,000 is 6,907, if the random number selected is lower than 6,907 it is a band 1 'word'. As [Ln(rankfreq)*1000] of 2,000 is 7,600, if the random number falls between 6,907 and 7,600 it is a band 2 'word'... and so on.

is the method used for this process of curve fitting. OLS is a mathematical optimisation technique which, given a series of measured data, attempts to find a function which closely approximates the data (a 'best fit'). It minimises the sum of the squares of the differences (called 'residuals') between the points generated by the function (in this case the logarithmic function) and the corresponding points in the data. In order to obtain a perfect fit, the sum of the squares should be 0. Nevertheless, this is rarely possible when making estimations, and residuals -or errors- are always present. It is required that we make these residuals (the sum of the vertical distances from the data points to the theoretical line) be as small as possible. The program outputs the theoretical profile that best matches the empirical one as well as the estimate computed and the error value. The error value given by the program is the indicator of the goodness-of-fit and it summarises the discrepancy between observed values and the values expected under the model in question.

Figure 7.17 displays an empirical profile from a learner's production and four theoretical profiles that will serve as a reference (actually the program compares the empirical profile with hundreds of theoretical profiles but we present just four for the sake of simplicity). Of the four theoretical profiles, two of them (number 2 and number 3) seem to be close to the empirical one. The program would compute the squared difference between the empirical and the best candidate and output the values of the most similar candidate together with the error of fit.

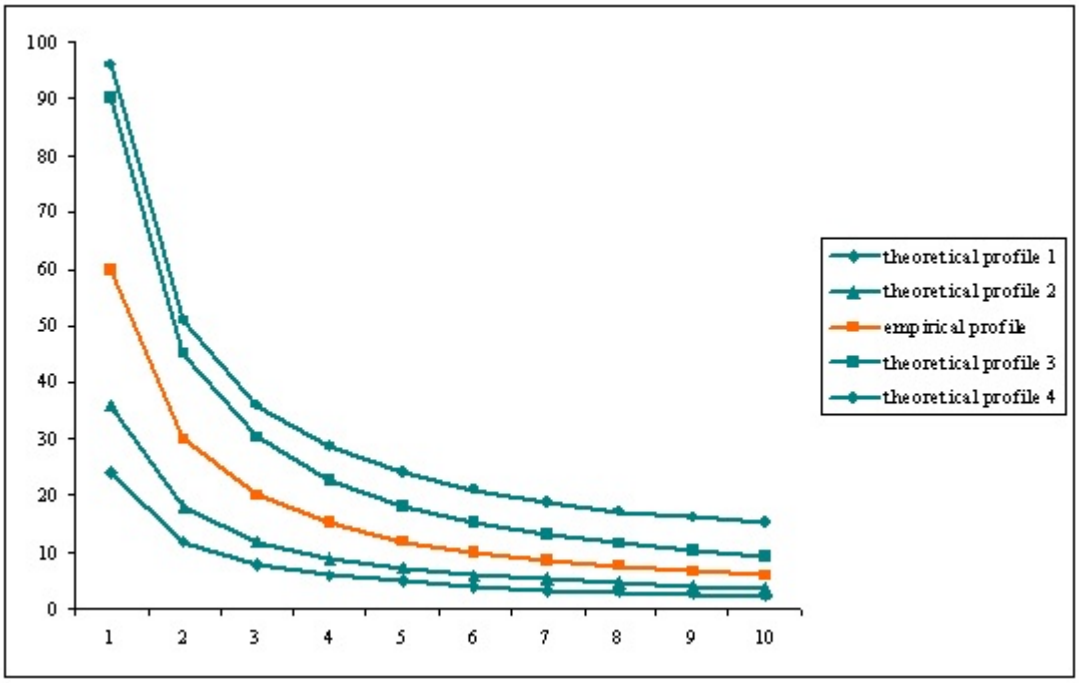

Figure 7.17. The empirical curve of the profile from the learner's production should match the theoretical profiles generated by the program.

Figure 7.18 summarises the process that the program carries out to estimate vocabulary sizes and that has been presented in this section. Appendix F contains the Manual of the program.

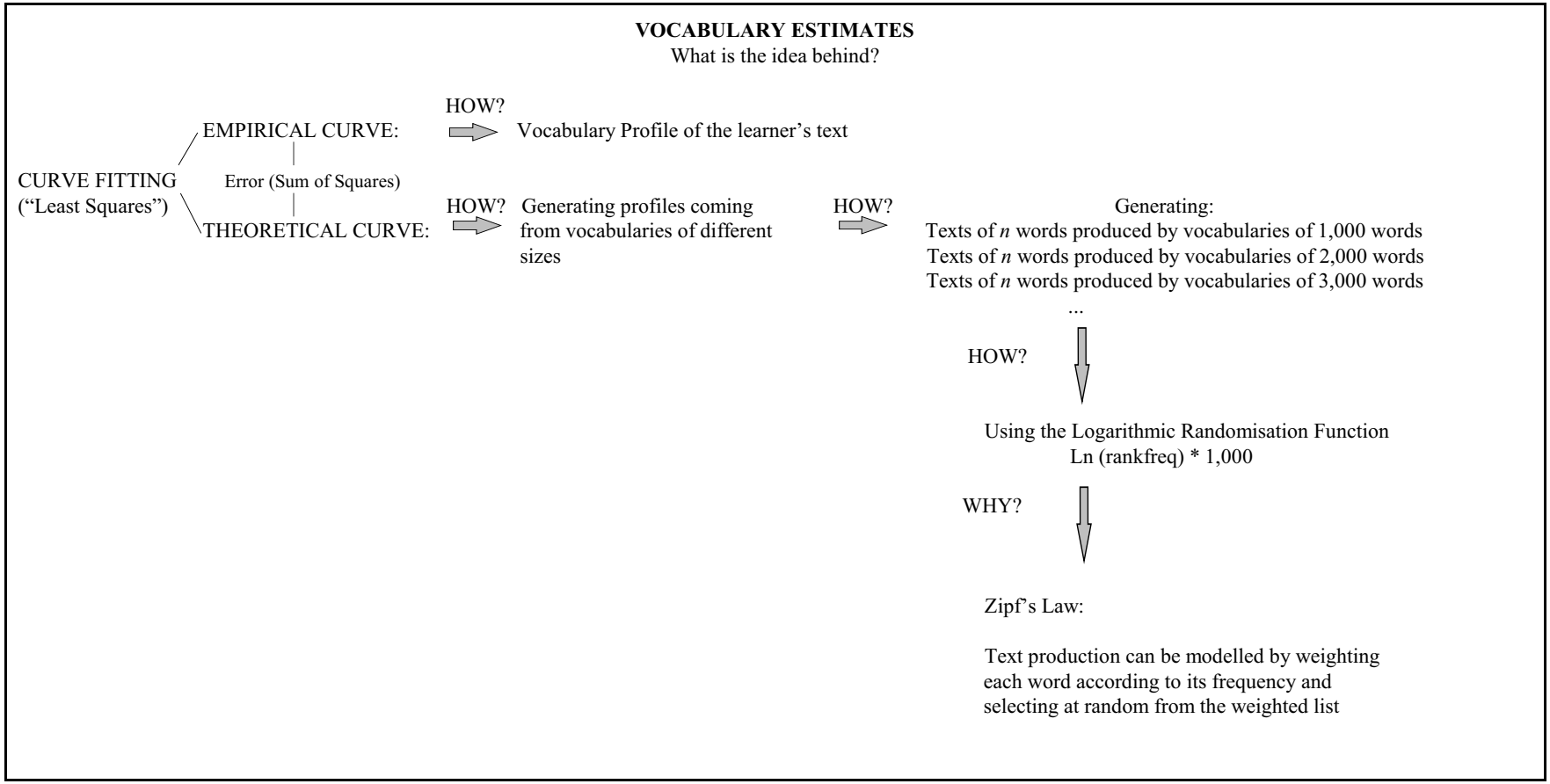

Figure 7.18. How V_Size estimates vocabulary.

7.2.3. The main advantages of the program

Among the advantages that the program offers the following can be highlighted: first of all, one of the innovations the program presents is that the existing tools that compute profiles (*WordClassifier*, Goethals 2005; *VocabProfile*, Nation 1995a) classify words into lists. They show the proportions of words that learners use at different levels of lexical development and therefore they are descriptive tools. In contrast, *V_Size* is inferential: it uses empirical profiles obtained from the learners' productions to infer the vocabulary size that generate these profiles and in doing so it uses five bands while other programs do not use more than four, which means that profiles are more accurate. Besides, it makes the inference based on a reliable mathematical model which, in any case, could be adjusted after experimenting with its application.

The second aspect that should be considered is the selection process to obtain the theoretical profiles, it chooses words at random taking into account that some words will appear more often than others (according to the logarithmic function), therefore the selection is 'weighted'. Finally, the manual comparison of the empirical data with the theoretical curve would be a long tedious process (check the differences, sum of the squares...), while *V_Size* computes the best fit between curves and gives the best size value and its error in a few seconds.

7.3. Experimental work with *V_Size*

As the program is currently under evaluation, in this section an account of the explorations carried out with *V_Size* is given. In order to check how the program worked, we focussed on two points: namely, which profiles it gave when we introduced different vocabulary sizes (section 7.3.1) and which sizes and errors the program computed when different profiles were introduced (section 7.3.2). Therefore, the important data to look at will be: the estimated vocabulary size, the error -sum of squares- (which should be as small as possible so as to find a good fit between the curves) and the profile (given in five bands: 1k, 2k, 3k, 4k and 5k).

7.3.1. From vocabulary sizes to profiles

In this section we want to see, first of all, which profiles are obtained with different vocabulary sizes: from a vocabulary of 500 words to one of 10,000 with increases of 100 words (a total of 96 different vocabulary sizes). For all these sizes, a default sample of 1,000 words was chosen[49]. Secondly, we want to analyse in which way the size of the sample could affects the results (if it does). Therefore, the profiles for 20 of the vocabulary sizes were estimated for different samples: 500; 1,000; 3,000; 10,000; and 50,000.

[49] 'Vocabulary size' is not the same as 'sample size'. The former refers to the total amount of words that the learner is thought to have, the latter means the number of times we make a trial as part of the size estimation process that has been described in section 7.2.2 and illustrated in Figure 7.18.

The complete results can be found in Table 1 in Appendix G, a summary of these results and their implications is presented below. As can be observed in Table 7.23, where three examples of vocabulary sizes are given, there is always a big difference between band 1 and the rest, a sharp fall between bands 1 and 2 is always present. The variation between the rest of the bands is subtle, the maximum difference is 10. These differences are due to the nature of the logarithmic function itself and actually we would expect band 1 to dominate the output as a consequence of Zipf's Law. Note that this is also what happens in learners' profiles: a broad difference between band 1 and the rest can always be found (see for instance Laufer & Nation, 1995:316; Muncie, 2002:229).

Voc.size	LnV*1000	Sample	1k	2k	3k	4k	5k
2,500	7,824	500	79	8	7	3	3
		1,000	79	9	7	3	2
		3,000	78	9	6	4	3
		10,000	79	9	5	4	3
		50,000	79	9	5	4	3
5,000	8,517	500	71	9	5	3	12
		1,000	73	8	5	3	11
		3,000	73	8	5	3	11
		10,000	73	8	5	3	11
		50,000	73	8	5	3	11
10,000	9,210	500	67	8	4	3	18
		1,000	67	8	4	3	18
		3,000	67	8	4	3	18
		10,000	68	7	4	3	18
		50,000	67	7	5	3	18

Table 7.23. Profiles given by the program for some vocabulary sizes and different samples.

Up to a vocabulary of 2,900 words, the curve is progressively going down from bands 1 to 5, while from 3,000 words onwards it falls down until band 4, then it rises until 5 (see Figure 7.19 and 7.20). This would mean that when learners have vocabularies bigger than 2,900, they should produce at least some words that have a low frequency in the language.

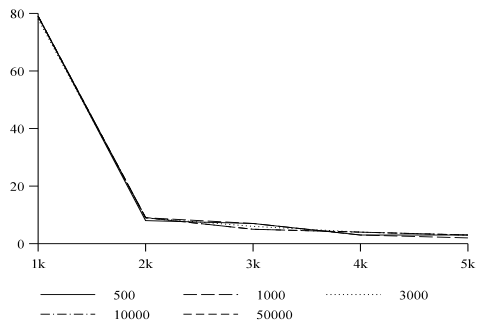

Figure 7.19. Idealised Vocabulary Profile of a vocabulary size of 2,500 words. The different lines correspond to the different sample sizes

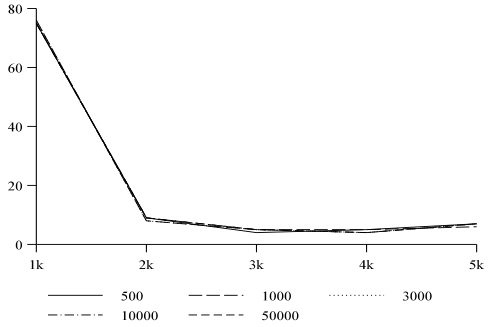

Figure 7.20. Idealised Vocabulary Profile of a vocabulary size of 3,500 words. The different lines correspond to the different sample sizes.

Changing the size of the sample does not make any big difference to the profile as shown in Table 7.23. That is, the number of times a sample is made as part of the vocabulary size estimation process (500 times, 10,000 times etc) does not have a big impact on the results, although it might modify them very slightly. In spite of the fact that one or two words move around, the profiles are nearly the same. The different lines in each of the Figures above (7.19 and 7.20) correspond to the profiles with different samples (500; 1,000; 3,000; 5,000; 10,000). As it can be seen, the shapes are nearly the same.

Finally, it is worth noting that different vocabulary sizes may produce the same profile, but when this happens the vocabulary sizes are similar and the difference is never superior to 500 words (for instance, sizes of 5,700 and 5,800 give a profile of 71-8-4-4-13 and 9,600 and 9,900 give one of 67-8-4-3-18).

7.3.2. From profiles to vocabulary sizes

In this section two issues will be explored. First we want to check which estimates are obtained with different profiles. Therefore, all the 44 possible profiles (all with different shapes) were introduced into *V_Size* to generate vocabulary estimates[50]. The bands of the profile added together have to be 100, as the program makes the calculation using percentages. The profiles studied here are representative of all the possible combinations, from just two bands (e.g. 90-10)[51] to 5 bands (e.g. 70-15-7-5-3), following this rule:

$$\text{band 1} \geq \text{band 2} \geq \text{band 3} \geq \text{band 4} \geq \text{band 5}$$

This rule was followed because it implies that the shape of a normal profile would be one in which the learner would know less (or an equal amount of) words in the higher bands than in the lower ones.

[50] 44 are the total number of possible band combinations that a profile can have to be fed into the program.

[51] The point in working with just two bands is implicit in Laufer's 'beyond 2000' measure, i.e. that just two figures (one for the most frequent words and one for the least frequent) would be enough to characterise productive lexical development. She adds up bands 1 and 2 of the LFP on the one hand and bands 3 and 4 on the other (Laufer, 1995).

The second aim of this section is to examine the effect of moving just one word from one band to another on the total vocabulary size estimated. In order to carry out this analysis, several profiles and all their variations between bands in just one word were taken into account, always following the aforementioned rule by which lower bands contained more words than higher bands.

Regarding the first issue, some important aspects should be emphasised concerning the results. First of all, the highest vocabulary estimate obtained is 7,200 (see the complete results in Table 2 in Appendix G). If we take these results together with those of the previous section, we realise that we would only obtain vocabularies higher than 7,200 if band 5 was higher than the previous ones (except for band 1). For instance, a profile of 68-7-5-3-17 gives an estimate of 9,500 words and a profile of 67-7-5-3-18 one of 10,000.[52]

Secondly, the variation in bands 2, 3 and 4 does not seem to be as important as in bands 1 and especially 5 when computing a vocabulary size estimate. This can be appreciated in the following example:

80	15	3	2	0	→2,200
80	8	7	5	0	→2,200

Both profiles give an estimate of 2,200 words (by moving 7 words from band 2 to bands 3 and 4).

[52] An estimate of 8,700 words would be obtained with profiles such as 40-60, 30-70 or 20-80. However, a profile of this kind would be extremely peculiar as a text and we can completely discard this set of possibilities. In addition, the errors would be really big: 3,778; 5,578 and 7,778 respectively. Profiles where band 2 is higher than band 1 like 30-35-15-12-8 would be rather unusual as well.

70	30	0	0	0	→3,300
70	20	10	0	0	→3,300

Both profiles give an estimate of 3,300 words (by adding 10 words in band 3 from band 2)

90	10	0	0	0	→1,000
80	20	0	0	0	→1,300

The estimate varies from 1,000 words to 1,300 words (by adding 10 words in band 2 from band 1)

65	15	10	8	2	→3,500
65	15	10	5	5	→4,200
65	10	9	8	8	→7,200

The estimate varies in 700 words by adding 3 words in band 5 from band 4 and it increases in 3,000 words by adding a few words in bands 4 and 5 from bands 2 and 3).

Thirdly, it can also be observed that errors start becoming big in vocabularies of more than 3,000 words when the difference between band 1 and band 2 is not so noticeable (Figure 7.21), or when there is not any word from bands 3, 4 and 5 (Figure 7.22), which can be accounted for because these shapes suppose a wide deviation from the typical shape the Power Law gives rise to (as shown in Figure 7.17).

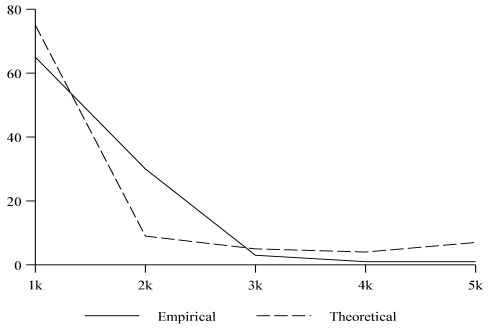

Figure 7.21. Empirical and theoretical profiles for a vocabulary of 3,500 words (Error=590).

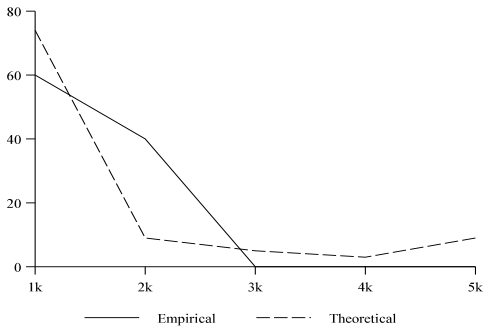

Figure 7.22. Empirical and theoretical profiles for a vocabulary of 4,200 words (Error=1,272).

Concerning the second issue, i.e. whether there were variations when one word was swapped in a band for a word in another, very small variations were found in the results. The biggest differences were around 500 words above or below the original estimated size (and this was the exception rather than the rule). As a rule of thumb, variations -if found- were of about 100 words. We show one of the examples below in Table 7.24: In the profile 85-8-4-2-1, which gives an estimate of 1,800 words and an error of 4[53], a word between different bands was systematically changed. The rows in white correspond to the profiles that we introduced to the program, the rows in grey show the results obtained: the estimate, the error, and the theoretical profile derived from the estimate. The only changes were found when a word was moved between bands 1 and 4, where the estimates were of 1,900 and 1,400 words instead of 1,800.

Change a word between...	Profile					Estimate	Error
bands 1-2	**84**	**9**	4	2	1		
	84	9	5	2	0	1,800	2
	86	**7**	4	2	1		
	84	9	5	2	0	1,800	10
bands 1-3	**84**	8	**5**	2	1		
	84	9	5	2	0	1,800	6
	86	8	**3**	2	1		
	84	9	5	2	0	1,800	10
bands 1-4	**84**	8	4	**3**	1		
	83	9	5	3	0	1,900	4
	86	8	4	**1**	1		
	86	9	5	0	0	1,400	4

[53] The size of the error between the theoretical and the empirical curve, i.e. the difference between the known vocabulary size and the estimated size.

Change a word between...	Profile					Estimate	Error
bands 1-5	**84**	8	4	2	**2**		
	84	9	5	2	0	1,800	6
	86	8	4	2	**0**		
	84	9	5	2	0	1,800	6
bands 2-3	85	**7**	**5**	2	1		
	84	9	5	2	0	1,800	6
	85	**9**	**3**	2	1		
	84	9	5	2	0	1,800	6
bands 2-4	85	**7**	4	**3**	1		
	84	9	5	2	0	1,800	8
	85	**9**	4	**1**	1		
	85	9	4	1	1	1,800	4
bands 2-5	85	**7**	4	2	**2**		
	84	9	5	2	0	1,800	10
	85	**9**	4	2	**0**		
	84	9	5	2	0	1,800	2
bands 3-4	85	8	**3**	**3**	1		
	84	9	5	2	0	1,800	8
	85	8	**5**	**1**	1		
	84	9	5	2	0	1,800	4
bands 3-5	85	8	**3**	2	**2**		
	84	9	5	2	0	1,800	10
	85	8	**5**	2	**0**		
	84	9	5	2	0	1,800	2
bands 4-5	85	8	4	**1**	**2**		
	84	9	5	2	0	1,800	8
	85	8	4	**3**	**0**		
	84	9	5	2	0	1,800	4

Table 7.24. Example of the effects of moving one word between bands.

Therefore, very small variations in the profile may produce different results (this can also be observed in Table 1 in Appendix G)[54], although this does not mean that the program is not useful to obtain rough estimates. A number of estimates would give us a mean estimate and this mean and the spread of the estimates may probably be used to get a close idea of the true value.

7.4. *V_Size* applied to our data

Our purpose in this section is twofold: first of all, we will infer vocabulary sizes from the profiles obtained in different tasks -oral and written- performed by a sample of our three groups of oldest learners: A3, B3 and A4. We want to know if there are differences in the amount of words that these learners use for different tasks, i.e. if an earlier AO (A3 and A4) to the FL entails having bigger vocabularies than those who start the instruction later in life (B3), or, in any case, if an earlier start and more exposure (A4) benefit students in the new curricula in opposition to those in the old one (B3).

As in chapter 6, the analysis will be carried out for each subject and task and four vocabulary estimates will be obtained for each student. In addition, a group estimate will also be computed for each task. There are two reasons for doing this: as has been seen in chapter 6, Laufer and Nation (1995) recommended having profiles of 200 words for

[54] It should also be noted that, in a parallel study, very slight deviations were noticed if band 5 was higher than band 4 (e.g. 81-7-5-4-3 gave an estimate of 2,400 words and 81-7-5-3-4 one of 2,600, which was the same estimate as the one obtained for 81-7-5-2-5). However, 81-7-5-0-7 gave an estimate of 3,600 words.

them to be stable and a 54% of the tasks analysed here do not reach 200 words (see Table 7.25). Therefore, computing estimates for groups of students will help to overcome this obstacle. Furthermore, the indications by Meara (2005) and Edwards and Collins (2007) should also be taken into account. They point out that LFP may not be sensitive enough to estimate vocabularies of particular learners if they do not produce enough words, even if they have a high proficiency. However, given a large amount of tokens (which could be obtained if the production of whole groups is analysed together), variability would decrease -as text size increases- and a rough estimate could be calculated for the group.

Secondly, we will estimate the vocabulary sizes of 6 native speakers (NSs) of English from the profiles obtained for one task: the storytelling. The main reason for computing these estimates in NSs is a theoretical one. There are studies suggesting that educated NSs know about 20,000 word families (Goulden, Nation & Read, 1990; Nation & Waring, 1997) and researchers normally assume that, as a rule of thumb, a NS adds 1,000 word families to his vocabulary (therefore a child of 10 would know about 10,000 and a child of 6 about 6,000). However, estimating vocabulary as a whole is extremely problematic, not just because of the sampling procedure to construct the test but also because of the notion of a 'total vocabulary' in itself. It may be more appropriate to consider a vocabulary for certain tasks (writing a letter to apply for a job is different from talking about the weather), rather than trying to quantify a 'total general vocabulary'. We would then think that different learners will have different vocabulary sizes for different tasks and the same will happen with NSs. In addition, if a task is (lexically) demanding, the estimates for NS will be higher than the estimates for low

level learners. In this case, as the NSs perform the same task as the FL learners, the results will be comparable.

7.4.1. Estimations for EFL learners

As regards the vocabulary estimates for EFL learners, the production of 24 students of English was analysed in this specific sub-study: 8 from A3, 8 from B3 and 8 from A4, who were the same learners as in A3 but tested a year later. Table 7.25 indicates the number of tokens each learner produces for each task as well as the mean and the totals for each task and group.

group	subject code	tokens int	tokens story	tokens role	tokens comp
A3	A3_1	204	143	121	144
	A3_2	154	143	170	119
	A3_3	192	101	251	155
	A3_4	281	117	97	163
	A3_5	244	131	91	119
	A3_6	173	120	81	122
	A3_7	251	101	149	130
	A3_8	339	103	125	122
	TOTAL A3	1,838	959	1,085	1,074
	MEAN A3	229.75	119.88	135.63	134.25
B3	B3_1	178	104	91	103
	B3_2	120	170	66	152
	B3_3	370	114	58	123
	B3_4	254	124	54	130
	B3_5	116	131	56	130

group	subject code	tokens int	tokens story	tokens role	tokens comp
B3	B3_6	172	168	184	102
	B3_7	168	151	53	122
	B3_8	274	167	138	103
	TOTAL B3	1,652	1,129	700	965
	MEAN B3	206.50	141.13	87.50	120.63
A4	A4_1	419	190	62	200
	A4_2	243	92	33	71
	A4_3	277	89	128	137
	A4_4	127	61	52	71
	A4_5	184	66	116	177
	A4_6	283	99	95	-
	A4_7	129	78	58	149
	A4_8	132	112	64	50
	TOTAL A4	1,794	786	608	855
	MEAN A4	224.25	98.25	76	122.14

Table 7.25. Tokens produced in each task by each subject together with the means and the total amount of tokens for each task and group.

The profile for each task was computed based on an adapted version of the Jacet List. Several points were considered for the selection of the list that was going to be used, as we believe that the lists chosen may probably affect the profile giving as a result different shapes of the curve. As the data being analysed comes from learners of English as a FL, it was considered that the list used to compute the profiles had to be compiled for SL learning purposes and should be representative of the data we were analysing, that is, a list compiled for methodological purposes was used rather than lists of general corpora as for instance the British National Corpus (BNC). Of the two up-to-date lists available that fulfilled these criteria (Nations' List and Jacet 8000), the Jacet List was

chosen, although it is more recent than Nations's Lists and therefore it has not been widely used in research yet. Several reasons back up this decision as shown in Table 7.26. First of all, as argued in Ishikawa et al. (2003), in the compilation of the Jacet List an educational viewpoint was incorporated in the scientific process of data processing, that is, the vocabulary of the textbooks, teachers and students was taken into account. Secondly, although the list has been compiled in Japan[55], the list has also been extensively revised (Murata, 2003), compared with and adjusted to other lists (Mochizuki, 2003) and it includes almost all the important words in other major vocabulary lists. Besides, both Nation's and Jacet Lists share very similar rules for lemmatisation. As shown in Murata (2003), Jacet 8000 is also based on a word family system.

As regards its implementation in the present study, the Jacet was considered to be more adequate because, although it is formed by eight frequency bands, in each band the exact frequencies of each word are given. With this information, additional points in the profile could be easily and reliably computed (we would need profiles of at least 5 bands for this study). Nation's lists should have been further divided as they consist of three main lists. In addition, structural words belonged to different bands in the Jacet, which was not the case in Nation's Lists, i.e. in Nation's List, all the structural words count as band 1 words (in the first 1,000 most words in English). In the Jacet, these words are distributed in different bands. For instance, the conjunctions 'but' and 'although' belong to band 1 in Nation's. In the Jacet, 'but' has a frequency rank of 22

[55] Actually, it is the List of the Japanese Association of College English Teachers and it is the fourth edition of the list.

and ‘although’ one of 335. We believe that this reflects the reality of our learners more accurately, as usually ‘although’ is not acquired until connectors such as ‘however’ or ‘in spite of’ are introduced, later than having learned ‘but’. Furthermore, there was a large number of words in our data not present in Nation’s list that would have to be placed to one band or another according to the general principles stated in the introduction of the lists booklet. This would have allowed a degree of subjectivity that could be avoided with the use of Jacet. Finally, the presence of cognates was examined in both lists as there were quite a few in the data analysed. Nation already pointed out that these lists are not suitable enough for speakers whose L1 is a Romance language. It was found out that cognates were present at different levels of the Jacet, while in the Nation’s lists tend to appear in the third thousand list, which was more academically oriented.

Nation's List +	-
· Widely used in research. · Not just frequency (defining power, regular syntax...).	· Not suitable for speakers of Romance languages. · Structural and very general words in band 1. · A large proportion of my words are not in the lists.
Jacet's List +	-
· Compiled for methodological purposes. · Exact frequencies (additional points in the profile). · Checked against other lists: BNC, Nation's, Kilgarriff... · Most of the words in the present study are in the list (representativeness). · Structural words in different bands. · Cognates at all levels (most from the 2,000 band onwards).	· In Japan, for Japanese learners of English.
+/- Similar rules for lemmatisation	

Table 7.26. Pros and cons of using the Nation's Lists or the Jacet List.

V_tools was used in order to compare each band of the Jacet to the words of the task we were analysing. We manually redistributed the words the program could not recognise (plurals, inflected forms of verbs, words found both in band 1 and in Jacet 250 Plus[56]...). Four profiles were computed for each subject (therefore we had a total amount of 95 profiles, 24 from each task, with the exception of a subject in A4 who did not write the composition). According to the Jacet List then, our profiles had twelve bands: 1) 1-500, 2) 500-1,000; 3) 1,000-1,500; 4) 1,500-2,000; 5) 2,000-2,500; 6) 2,500-3,000; 7) 3,000-3,500; 8) 3,500-4,000; 9) 4,000-8,000; 10) Jacet Plus -mainly proper nouns-;

[56] This list contains irregular forms that correspond to very frequent words (e.g. common participles), very frequent auxiliaries and proper nouns, as well as numerals.

11) personal nouns and numbers and 12) not-in-the-list words. We reduced them to 5 bands so that we could introduce the data in *V_Size*.[57]

After that, a vocabulary estimate for each of the profiles was obtained, the results are displayed in Table 7.27. As can be seen, our learners have vocabularies between 1,000 and 1,500 words. This seems quite plausible and would confirm our expectations, as we have the impression that most of our learners have problems in reaching 2,000 words, especially in speaking. There are no outstanding differences between the groups. There are just two learners that have the same estimated vocabulary for each task (A3_6, and A4_7 have all estimates of 1,000 words). All the other subjects differ in the estimates for each task, although there is just a slight variation. In most of the cases, the differences are not bigger than 300 words. However, there are a few cases, which could be considered outliers, in which the estimated vocabularies are of 1,800; 2,400; 2,800 or even 3,600 words (in bold in Table 7.27).

[57] Proper nouns and numbers were added to band 1, as well as Jacet Plus words; bands 5 to 9 (2,000 to 8,000) were added together and constituted band 2k+.

Subject Code	int	narr	role	comp	mean
A3_1	1,100 (2)	1,000 (114)	1,100 (36)	**2,800** (10)	1,500
A3_2	1,100 (50)	1,000 (84)	1,200 (47)	**2,400** (40)	1,425
A3_3	1,200 (34)	**2,400** (28)	**1,800** (80)	1,100 (64)	1,625
A3_4	1,100 (88)	1,000 (86)	1,000 (102)	1,400 (30)	1,125
A3_5	1,100 (62)	1,000 (64)	1,000 (62)	1,000 (110)	1,025
A3_6	1,000 (90)	1,000 (102)	1,000 (50)	1,000 (34)	1,000
A3_7	1,400 (26)	1,400 (10)	1,100 (70)	1,000 (74)	1,225
A3_8	1,400 (54)	1,100 (88)	1,400 (20)	1,200 (58)	1,275
B3_1	1,400 (94)	1,100 (8)	1,000 (114)	1,200 (144)	1,175
B3_2	1,000 (68)	**1,800** (46)	1,000 (126)	1,100 (24)	1,225
B3_3	1,000 (90)	1,400 (22)	1,000 (98)	1,400 (46)	1,200
B3_4	1,100 (32)	1,100 (18)	1,000 (88)	1,000 (38)	1,050
B3_5	**2,400** (22)	1,400 (54)	1,400 (78)	1,000 (88)	1,550
B3_6	1,000 (86)	1,400 (112)	1,000 (166)	1,100 (86)	1,125
B3_7	1,000 (70)	1,100 (102)	1,400 (46)	1,100 (64)	1,150
B3_8	1,100 (6)	1,100 (46)	1,000 (126)	1,100 (22)	1,075
A4_1	1,100 (0)	1,100 (54)	1,200 (18)	1,000 (50)	1,100
A4_2	1,200 (20)	1,200 (34)	1,000 (200)	1,000 (108)	1,100

Subject Code	int	narr	role	comp	mean
A4_3	1,100 (64)	1,400 (90)	1,000 (98)	1,200 (34)	1,175
A4_4	**3,600** (38)	1,200 (88)	1,400 (70)	1,000 (128)	1,800
A4_5	1,100 (92)	1,200 (24)	1,000 (152)	**3,600** (64)	1,725
A4_6	1,800 (54)	1,000 (86)	1,400 (34)	-	1,400
A4_7	1,000 (24)	1,000 (142)	1,000 (98)	1,000 (90)	1,000
A4_8	1,100 (14)	1,100 (46)	1,400 (62)	1,000 (72)	1,175

Table 7.27. Estimated vocabulary for our learners. The error for each estimate is the figure within brackets. In bold, the estimates that differ considerably from the estimates for other tasks performed by the same subject.

In the light of a Wilcoxon Signed Ranks Test, as this data presents a distribution that calls for non-parametric tests, we can say that there are no systematic differences in the vocabulary estimates depending on the tasks[58]. That is, for instance, task X is not producing larger estimates than task Y. The main implication of this finding is that the mean (and probably the standard deviations) for each task can give us a rough estimate of the student's vocabularies. There were no significant correlations between the estimates for the tasks either, except for one between the estimates for the roleplay and the estimates for the interview significant at the .01 level ($r=.55$, $N=22$, $p<.008$). The fact that estimates for one task do not necessarily correlate with estimates for another points out that estimating sizes by just one particular sort of task might be dangerous.

[58] If 'E' means 'Estimates', the results were the following (all non-significant): Eint- Estory= .812; Eint-Erole=.408; Eint-Ecomp=.825; Estory-Erole=.167; Estory-Ecomp=.858; Erole-Ecomp=.461. The degree of significance of the results did not change when the analysis was performed with or without outliers.

It is interesting, though, that the correlation is found between the two tasks in which there are two interlocutors. It is possible that tasks that share traits give similar estimates.

Bigger sizes shown in the composition (as it happens with three subjects) could be explained by the nature of the task: they have time to think about which vocabulary they should use and thus it may be more elaborated; in addition, it is written language, which is the kind of language to which students are more exposed at school, especially regarding the FL class. However, other subjects have bigger estimated vocabularies for the interview and the storytelling. This would be another indication of the necessity to assess vocabulary in different kinds of tasks in order to obtain a more reliable estimate rather than using just one.

Similarly to what has been done in chapter 6, as the amount of tokens produced for some tasks was not large (see the means in Table 7.25), group estimates were also computed for each task. In this way we wanted to check if estimates inferred from profiles obtained from texts that were not actually long diverged from profiles obtained from larger corpora (built up with texts from students in the same group).

Table 7.28 displays the group means for each task when estimates are computed individually. Table 7.29 shows the estimates and errors of fit when the whole group vocabulary for the particular task is considered.

	interview		storytelling		roleplay		composition	
	mean	***sd***	**mean**	***sd***	**mean**	***sd***	**mean**	***sd***
A3 (N=8)	1,175	52.61	1,237.50	173,14	1,200	98.20	1,487.50	250.31
B3 (N=8)	1,250	171.13	1,300	88.64	1,100	65.46	1,125	45.32
A4 (N=8)	1,500	312.82	1,150	46.29	1,175	70.08	1,400	367.75

Table 7.28. Group means and standard deviations when estimates are computed individually for each student. N=7 in group A4 composition.

	interview		storytelling		roleplay		composition	
	mean	***sd***	**mean**	***sd***	**mean**	***sd***	**mean**	***sd***
A3 (N=8)	1,100	36	1,100	48	1,100	48	1,200	44
B3 (N=8)	1,100	46	1,200	44	1,000	86	1,000	48
A4 (N=8)	1,200	44	1,100	62	1,000	84	1,100	54

Table 7.29. Estimates and corresponding errors when estimations are carried out for the tasks in each group. N=7 in group A4 composition.

The comparison between the means for each group in each task can be seen in Figure 7.23. When estimates are computed individually for each task and student the group mean is higher than when a task estimate is computed for the whole group. However, the difference is never bigger than 300 words and the proportion between groups is kept; i.e. the group that has a high estimate in a particular task keeps it high independently of how it has been calculated and the same happens with the group with a low estimate (each pair of lines is parallel in Figure 7.23).

With the estimates computed on an individual basis (as shown in Table 7.28), two Mann-Whitney analyses were performed between groups A3 and B3 and between A4 and B3. A Wilcoxon Signed Rank Test was performed between A3 and A4 as the subjects were longitudinal. Results showed that the difference between the estimates was

not statistically significant in any case. However, how would these estimates for FL learners differ from those of NSs performing one of these tasks? Would the program give higher estimates for NSs ?

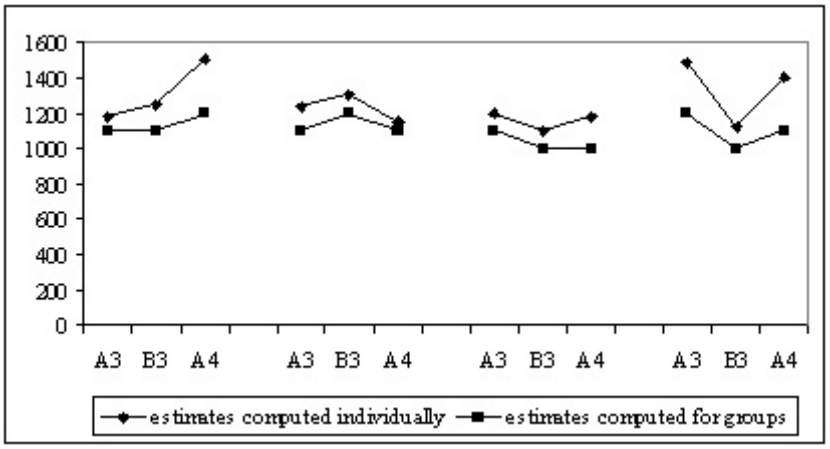

Figure 7.23. Estimate vocabularies for each task computed in to ways: mean of the estimates computed individually and estimates computed for whole groups.

7.4.2. Estimations for NSs

As regards vocabulary estimates for NSs, 3 university teachers (NS1, NS2 and NS3) and 3 children (NS4, NS5 and NS6) were recorded while telling the same story used with the FL learners and the data was transcribed and analysed following the same procedure that has been previously described, as displayed in Table 7.30. The university teachers' ages ranged between 25 and 40. Two of the children (NS4 and NS5) were aged 9 and one (NS6) 6. Including 3 adults and 3 children in the sample was not arbitrary, as we wanted to see the effect of age when performing this type of task. Table 7.30 also shows the amount of tokens they produced and the estimates and error for the

storytelling task. The three last rows summarise the learners' results as a way of comparison. Were we estimating total vocabulary size, following the general rule that NSs add 1,000 words per year to their lexicons until having 20,000 words, NS1, NS2 and NS3 should have about 20,000 words while NS4, NS5 should have about 9,000 and NS6 6,000.

Subject Code	**Age**	**Tokens narration**	**Estimated vocabulary and error**
NS1	40	373	4,800 (16)
NS2	26.5	269	5,300 (16)
NS3	25	143	4,300 (46)
NS4	9	58	4,300 (80)
NS5	9	46	4,300 (158)
NS6	6	72	1,800 (80)
Group	**Mean age**	**Mean tokens narration**	**Mean estimated vocabulary**
A3	16.3	119.88	1,237.50
B3	17.9	141.13	1,300
A4	17.7	97.62	1,150

Table 7.30. NSs'ages and tokens produced, vocabulary estimates and errors (in brackets) compared to learners' performance in the storytelling task.

As the estimates for NSs are generally higher than those of the learners, this makes us predict that the program will probably discriminate between NSs and intermediate learners. In addition, for this task, none of the NSs surpassed the estimate of 5,300 words. Therefore, a learner with an estimate of 5,000 words, for instance, would not be that far from performing in a similar way as a NS as regards vocabulary choice. The fact which is most outstanding is that the estimates for the two nine-year-

olds are very similar to those of university teachers (about 4,300 words). That would indicate that age is not a crucial factor in the performance of this task (remember it is a story about two children having a picnic and the dog eating their food) and it is good evidence that, when assessing vocabulary, results may be task dependent. In this particular case, vocabulary choice is limited by the visual images. It is thus made clear that estimations can be done for particular tasks and it might be dangerous to extrapolate from one of these estimates a total vocabulary size (the estimates for NSs would be around 5,000 and not 20,000 words for this task), as it is quite unthinkable that either a NS or a learner will employ all of his/her productive vocabulary in a particular situation.

Another important aspect is that, although the amount of tokens is different between adults and children, the estimates do not differ much or are the same (see NS3, NS4 and NS5). This shows that regardless of the amount of tokens used, the distribution of the words produced is similar: the shape of the profile does not vary much and therefore the estimate is similar as well, as it is calculated from the shape of the profile.

It should be mentioned, however, that all these results should be interpreted with caution as the number of subjects studied is small. We should treat them as a gross indication of what the program does when applied to real data and how we could use it in order to obtain more reliable results. In addition to observing general trends in our learners' lexical performance, these sub-studies with the data presented (either with learners or with NSs) are aimed at devising ways of improving vocabulary assessment and they need to be further explored (see section 8.4.3).

7.5. Discussion and conclusions

This chapter on vocabulary estimations allows us to throw light on different aspects as well as to point out some directions for future research. We consider that using the logarithmic function and the Least Squares method to infer vocabulary size is a step forwards in the analysis of learner's interlanguage. Making use of inferential procedures to describe how a vocabulary develops can be regarded as a methodological improvement and, in this sense, *V_Size* constitutes one of the first ways to do so. Apart from all the improvements brought about by the program referred to in section 7.2.3, the most important achievement that this tool represents is that the mathematical process that it uses changes the conception of estimating vocabulary, as it is the first serious attempt to make a vocabulary estimation from a piece of writing or speech.

However, the use of this method to estimate vocabulary may have two inherent dangers as well: one is the use of the logarithmic function as the basis for this procedure and the other is the baselist that we use to obtain the profile.

Concerning the use of the logarithmic function, the possibility exists that it is not the most adequate function for estimating vocabulary size with this procedure. As has been found in the second section, for instance, big vocabularies are mainly an effect of band 5 (infrequent words). Moreover, it has been pointed out in the literature that the relationship $r \cdot f = C$ does not always hold for words of the highest and lowest frequencies. A good example of this is given by Crystal (1941/1987): the most frequent word in the London-Lund Corpus, (I) occurs 5,920 times ($r \cdot f$=5,920), and the 100th (he's) occurs 363 times ($r \cdot f$ = 36,300). The difference between 5,929 and 36,300 shows

that, although the result of rank per frequency should be more or less constant, this might not happen if results in the highest and lowest ranks are compared, given the disparity of these two figures.

These weaknesses, rather than disproving the whole process altogether, could be minimised with some adjustments. One of the modifications could be made in relation to the place of the distribution where we apply the function. That is, if Figure 7.24 represents the words in a language, where A is formed by the most common words, B the common and C the least common, Zipf's Law asserts that if we choose a word at random, it is more probable that we take it from A than from B or from C. If we choose a word at random which is not from A, it is more probable that we take it from B than from C. If the relationship $r \,.\, f = C$ does not always hold for words of the highest and lowest frequencies, we should then, for example, apply the function only to B.

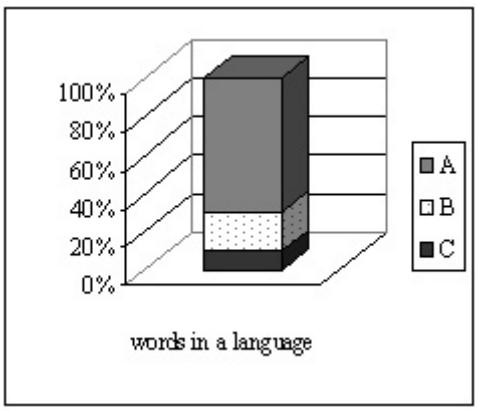

Figure 7.24. An idealistic representation of the frequency of words in any language.

Another shortcoming of estimating size using these calculations could be that although there are precise predictions of the amounts of words that account for parts of

texts, short samples may present variations from these proportions. For instance, according to statistical predictions, the first 15 words of a text account for 25% of the text, the first 100 words for 60%, the first 1,000 for 85%, the first 4,000 for 97,5%. Therefore, in spite of the fact that Zipf's law could be observed in texts of any length, further investigations should be conducted to determine the length at which the estimations would be most reliable with the logarithmic function.

Apart from the possible flaws and the potential solutions suggested in this estimation method, we believe that the logarithmic function is much more satisfactory than any other so far. Moreover, if we are (ever) to achieve a fair way to estimate vocabulary size, even if it is just for certain tasks, there is little doubt that a function that takes probability into account should be used.

It is also interesting to note that, as can be found in Table 1 (Appendix G), the two bands that have more weight in the computation of estimates are band 1 and band 5. These results are in the same line as those obtained by Edwards and Collins (2007), as they also acknowledge the impact of their bands 1 and 3 (in their study, band 1 consists in the first 1,000 words and band 3 in words not included in the first most frequent 2,000). As our bands consist of 500 words, we could further argue that it is possible that the first 500 words are those determining the estimate to a larger extent (our 1k band), together with those not included in the first 2,000 (that is, Edwards & Collins' 3k band and our 5k band).

We should also be conscious of the fact that the profiles are sensitive to the lists on which they are based and probably to the language these lists are in. For this reason, it is very important that, when making estimations, we make explicit the type of lists we

work with (lemmatised or unlemmatised) and what we define as a word. Further research is also needed with estimations in different languages, as making predictions in a language that makes a wide use of agglutination will certainly be different than making them in an isolating language.

All these aspects can bring about an improvement in the precision with which *V_Size* estimates vocabulary. Having a tool that accurately describes the size of productive vocabulary in different tasks is essential to have a better picture of vocabulary development and becomes a pre-requisite to analyse the up-to-now elusive relationship between receptive and productive vocabularies.

As regards the estimates computed for our learners, there are no striking differences in the vocabulary estimations for the three groups. Towards the end of secondary education it is not common that any of the two groups gets more than 1,500 words in the tasks performed. This means that starting the instruction in the FL two years earlier, even if it is accompanied with a few more hours of exposure, does not bring about an advantage as regards the amount of productive vocabulary for these four representative tasks, either oral nor written. What is more, as Table 7.28 shows, A3 exhibits a larger productive vocabulary in the storytelling than A4 (1,237.50 vs. 1,150). Nevertheless, the estimates computed for whole groups (see Table 7.29) give the same result (1,100 words) for these groups in this particular task. Given the experimental nature of the tool used for the estimations, we should treat the estimates with prudence, that is, we should consider them rough approximations that suggest that these groups do not present serious differences in productive vocabulary size for the tasks under study.

Although in previous chapters significant differences had been found in the long term, especially between A3 and B3 in favour of the LS, the vocabulary size estimations for the tasks analysed do not differ systematically between the groups. A possible explanation is that the differences between the groups are not large at this particular point and the program may not sensitive enough to detect subtle differences of vocabulary size.

Finally, the estimations for the storytelling in the NS data show that the concept of a 'total vocabulary' can be misleading. An estimate of 2,000 words for a learner in a particular task does not mean that this learner is far from being proficient because a NS would know 20,000 words: a native vocabulary for that task could be of 5,000, as happened with the storytelling task. A further point would be if it is fair to assess learners taking into account what NSs do, but this is not a point under consideration here. Overall, the estimations obtained for learners and NSs in a particular task reinforce the idea that to establish reliable and meaningful comparisons, estimates should be obtained for different tasks without considering vocabulary as a whole.