

**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 3

MEASURING VOCABULARY

3.1. Introduction

Measuring vocabulary is not an easy matter, mainly because the results are conditioned by the type of vocabulary being measured and how it is defined. This chapter focusses on two main points. The first (presented in 3.2 and 3.3), examines how vocabulary, especially productive, has been measured in the literature, with the purpose to determine which are the most appropriate ways to analyse our data. The second point (section 3.4) deals specifically with issues related to vocabulary size, like why it is important or how it has been assessed until nowadays.

We would like to highlight in this introduction an article that Alvar Ellegard published in 1960 entitled 'Estimating Vocabulary Size'. Although it is not a much cited work in the literature, anyone who reads this work today would realise that most of the core issues in vocabulary research that this article deals with are still a challenge in our days, after nearly fifty years. Moreover, this article presents some conceptual and methodological considerations in measuring vocabulary that have been thoroughly debated or turned into standard procedures as the years have gone by.

Taking into account that research in vocabulary acquisition up to that moment had been “largely atheoretical and unsystematic”, had aimed at providing “practical tips for teachers’ avoiding ‘the serious theoretical questions that arise when one moves away from this very basic level” (Meara, 1980:221-222), this article offers some innovative ideas. In order to examine how far we have gone and where we are now as regards vocabulary measurement, several aspects from Ellegard’s seminal work will be drawn on throughout the present chapter, which is organised in three main sections. Each part deals with a particular question: how many different words will there be in a text of a given size?; what do we mean by saying that a writer has a rich vocabulary?; and how shall we estimate the potential vocabulary of an individual?

3.2. Different words in a text sample of a given size

In order to answer this question successfully, there is another one that should be answered first, which is: *what constitutes a word?*. Only once a definition of a ‘word’ has been given will we be able to determine the amount of different words in a text, because depending on what is taken as a word, the count will obviously vary. It is rather difficult to define what a word is. There are even some who do not agree with the use of the term ‘word’, Carter (1987), for instance, states that the variable orthographic, phonological, grammatical and semantic properties of words are best captured by the use of the term ‘lexical item’. Also Sinclair (2004:281), who believes that meaning is related to word patterns and not to individual words, uses the term ‘lexical item’ to refer to ‘one or more words that together make up a unit of meaning’.

Generally, a word is considered to be the linguistic unit which has a space on either side when written, strictly speaking this definition would correspond to what is known as a 'token'. The amount of tokens or 'individual words' in a text is normally put in relation with the number of types (or the number of different words) that a text has. In language acquisition, when talking about words we usually mean 'lemmas'. A lemma consists in the base (defined as the simplest form of a word) and the inflected forms of a word (*waste, wasted* and *wasting* are a lemma). Another widely-used concept is that of a 'word family', which is formed by the base word, all of its inflections and its common derivatives (*light, lights, lighting, lighten* and *enlightenment* constitute a word family).

Ellegard (1960) distinguishes between 'word units' -where inflected forms are counted as separate, hence *estimation, estimate* and *estimating* will be three different word units-, 'lexical units' -where no difference is found between the root (base) and its derivations, hence *estimation, estimate* and *estimating* will be one lexical unit-, and 'semantic units' (similar to the so-called 'morphemes'¹²). However, he considers the lexical unit an "unsuitable basis of operations" (1960:233) especially when analysing SL acquisition, where word units seem to be the most appropriate. This is probably due to the fact that in a SL, the development of grammatical and derivational variants or word forms are informative units of the acquisition process.

It can thus be affirmed that there are many different kinds of vocabulary items or words and this is especially true when nonnative learners "eye their target language

¹² A morpheme is the smallest unit of meaning in a language; e.g. the word 'painter' consists of two morphemes: 'paint' and 'er'.

as linguistic outsiders” (Folse, 2004:1). Also Nation (1990) highlights the fact that the criteria to establish boundaries between words have an important effect on learning to distinguish words, which can be distinguished entirely on their form (*wish* and *wishes* are different words), on their meaning (the *foot* of a person and the *foot* of a bed might be considered one or two words), or with reference to either the learners’ mother tongue or the SL (a *steamboat* is one word in English but *barco de vapor* may be thought to have three words in Spanish).

Folse (2004) classifies vocabulary into single words (*sudden* is one word), set phrases (formed by more than one word that do not usually change, like *all of a sudden*), variable phrases (as *off and on* or *on and off*), phrasal verbs (like *put up with*), and idioms (products that *sell like hotcakes* are ‘marketable’). Thus, apart from the single words mentioned, the rest of his classification would correspond to what Nattinger and DeCarrico (1992) call ‘lexical phrases’, which would account for the presence of multi-word units in language.

3.3. What we mean by saying that a writer has a rich vocabulary

3.3.1. Types of vocabulary

Vocabulary knowledge is complex to define as it can be of different types. First of all, as has been pointed out in the introduction of the present dissertation, one of the most well-known distinctions is the one often made between passive/receptive and active/productive vocabulary. Although Melka (1982, 1997) mentions that it is quite

impossible to find a clear and adequate definition of what is meant by reception and production, and that there have been different attempts to describe these notions especially when applied to vocabulary, we base our work on the definitions by Nattinger (1988:62) and Meara (1990:152-53): as regards receptive vocabulary, it is “the understanding of the meaning of words and storing words in memory”; that is, “you can recognise passive vocabulary when you see it or when you hear it, but you are unable to bring it to your mind without external support”. As regards productive vocabulary, it is the “retrieval of words from memory by using them in appropriate situations”, it is “vocabulary easily accessed from anywhere in the vocabulary network, and in turn it allows easy access to other parts of the system too.”

Knowing a word receptively or productively involves different aspects related to form, meaning and use (Nation, 2001)¹³. From the point of view of receptive knowledge, knowing a word receptively implies knowing what the word means in a particular context, being able to recognise its collocations or being able to identify the written form so that it is recognised when reading. From the point of view of productive knowledge, knowing a word entails, among other aspects, being able to use it in order to suit the degree of formality in a situation or being able to say it with its correct pronunciation.

Secondly, when defining vocabulary knowledge, we may refer to an individual’s overall vocabulary knowledge (quantity of words), and then we talk about ‘vocabulary

¹³ Nation (2001) points out that the difference between kinds of vocabulary knowledge (of form, meaning and use) is crucial as it implies different kinds of learning. This is closely connected to the research on implicit/explicit learning by Ellis (1994b) presented in chapter 2: formal recognition and production would rely on implicit learning while meaning aspects would rely on explicit learning.

breadth' or 'vocabulary size', or we may refer to quality of word knowledge (how well a subject knows the words), which is regarded as 'vocabulary depth'. This distinction has implications for vocabulary learning and instruction and, as we will see in the following sections, it also affects vocabulary testing.

3.3.2. How can 'richness' be assessed

The assessment of vocabulary knowledge and of lexical richness in particular is not less problematic than attempting to define what constitutes a word or what word knowledge is. This lack of common consent makes the situation particularly difficult to tackle:

“On the theoretical level as well as on the practical level, we are confronted with an empty space as far as vocabulary acquisition is concerned. This situation could easily be turned into a vicious circle where everyone is waiting for the others: those who want to define vocabulary knowledge want to be able to measure it, but at the same time test constructors will only be able to develop valid and reliable tools if it is clear what has to be understood by vocabulary knowledge.” (Bogaards, 2000:511).

As regards vocabulary measurement, Meara and Bell (2001) make a distinction between intrinsic and extrinsic measures of vocabulary: in intrinsic measures, the assessment is carried out only in terms of the words that appear in the text, while in extrinsic measures other aspects not included in the text itself are taken into account. In the studies reviewed in chapter 2, different measures were used to assess learners' lexical knowledge. Although there is a variety of lexical measures available nowadays (see for instance Wolfe-Quintero, Inagaki & Kim, 1998), the most typical intrinsic

measure of vocabulary richness in SL studies is Lexical Variation (LV), also called Lexical Diversity or Type-Token Ratio (TTR).

TTR is the number of different words as a ratio of the total number of running words in a text. It is supposed to show how likely it is for a learner to repeat the same words. However, one of the main problems that this measure presents is its sensitivity to text length (Daller, van Hout & Treffers-Daller, 2003; Faerch, Haastrup & Phillipson, 1984, Richards, 1987; Vermeer 2000, 2004), basically because the rate at which new word types appear in a text decreases as the text size increases. There have been some attempts to overcome this problem, such as the use of adapted measures like the Mean Segmental TTR (MSTTR) or the Bilogarithmic TTR (LogTTR) and the Root TTR (also called Guiraud's Index)¹⁴. However, they are just small variations of the TTR and therefore not feasible solutions in most cases and can be specially unstable on short texts. Another possibility has been to fix the length of all the samples we want to analyse so as to keep length constant (Arnaud, 1992), but this also implies that data is lost when cutting the texts¹⁵.

In addition to TTR, Lexical Density (LD) has often been used to assess lexical richness: it consists in the proportion of content words as opposed to function words (Ure, 1971), but it does not seem to be a good measure at low-levels either, due to the fact that some students use telegraphic style, thus they do not make use of much function

¹⁴ MSTTR is defined as the average of TTRs of several consecutive equal-sized samples. Wachal and Spreen (1973) argue that MSTTR would be useful to compare different samples if different researchers used the same segment size and that LogTTR is a more stable ratio than Guiraud's Index. However, Guiraud's index has also been shown to overcompensate for the falling TTR curve.

¹⁵ Malvern and Richards offer one of the best comprehensive bibliographies on lexical diversity on <http://www.personal.rdg.ac.uk/ehsrichb/home2.html>.

words. This would yield higher LD values while it would actually reflect the inability to construct a coherent text (Hyltenstam, 1988)¹⁶.

Recently, a new intrinsic measure, D, has been proposed by Malvern and Richards (1997, 2002) and Malvern et. al (2004). D is an index that measures lexical diversity through a process of curve-fitting¹⁷, which is the general problem of finding equations of approximating curves that fit given sets of data. It is claimed to be more informative than TTR, because, as opposed to the single value of the TTR, it represents how TTR varies over a range of token sizes for each speaker or writer. This measure also has two other advantages. Firstly, because it is not a function of the number of words in the sample, it uses all the data available in the text, so it is not necessary to standardise text length. Secondly, it is claimed to work with short texts (50 tokens are the exact requirement), which is especially relevant when working with low-level learners and oral data, since these learners do not normally produce much.

Jarvis (2002) compares the D formula with other indices of lexical diversity that can also be used in a curve-fitting approach. He concludes that D is accurate for analysing whole texts (with both content and function words), as opposed for instance to U (Uber index: A vocabulary measure: $\log^2 N / (\log N - \log V)$), which seems to be more reliable when carrying out the analysis with just the content words from each text. However, he points out that more evidence is needed to check whether the efficacy of D extends to other types of written and oral texts (he used written narratives in his study)

¹⁶ Ure (1971) also gave very interesting insights into this measure. For instance, for spoken texts, LD is normally under 40% while for the written ones is over this figure. In spoken texts, the measure also varies consistently depending on whether there is feedback for the speaker or not. LD of written texts could also vary according to the personal and social relations between the participants.

¹⁷ Curve-fitting was also presented in Ellegard (1960) as a replacement of range percentages.

produced by a greater variety of learners and NSs. It is difficult to say if D will actually become the most appropriate measure for lexical richness as theoretical and empirical evaluations of *vocd* (the program that computes the D index) have recently been put forward (McCarthy & Jarvis, 2007), suggesting possible weaknesses and ways of improvement. However, it seems to be establishing itself as a quite reliable standard measure (Daller, Milton & Treffers-Daller, 2007).

Extrinsic measures of vocabulary richness are those that classify items according to criteria external to the text itself, they are also claimed to make fairly strong inferences about the total lexical resources that are available to the writer. Some of the traditionally used are Lexical Originality and Lexical Sophistication. The former refers to the percentage of lexical words in a text used by a particular writer and none of the other members of the group¹⁸. The latter is the percentage of lexical ‘sophisticated’ words in a text when they are compared to the words appearing in an external list, which is chosen according to the level of the learner who is tested.

We also find instances in recent research of authors who affirm that the lexical diversity of a text is not fully self-contained and that the contribution that words make to the diversity of a text cannot be determined without considering the word’s role in the language as a whole (Jarvis, 2003) or their frequencies in daily input (Vermeer, 2004).

This idea of using information not present in the text in order to evaluate the performance of the writer/learner is not new, it had already somehow blossomed into a

¹⁸ There are two main problems with this measure that have been acknowledged in the literature. Firstly, the concept of ‘originality’ is difficult to define, especially in SL development, and whether it is a valid concept to assess progression is not clear. Secondly, what might be unique or ‘original’ in one corpus might not be unique to another corpus, thus the learner can have a high Lexical Originality score in one class with a particular group of people and a low one if his/her group of peers change. This fact makes this measure unstable and not very reliable.

formal proposal in Ellegard (1960:240): “Ideally, we should try to ascertain what may be called the ‘vocabulary profile’ of each language user, indicating the percentage of words known within a definite set of frequency ranges”¹⁹. However, vocabulary profiles only become common assessing devices in SL research in the 90s, after the *Lexical Frequency Profile* (LFP) developed by Laufer and Nation (1995). This fall into oblivion was probably due to the difficulty of compiling adequate word lists, the limited use of computer tools and the restricted access to language corpora.

LFP is thought to show the amount and frequency of productive vocabulary available to the learners at a particular stage of their learning. This profile shows the percent of words from four different frequency levels and the calculation is done by the *VocabProfile* program (Nation, 1995a). This program operates on the basis of four word lists (Nation, 1996): *Word List One* (1k) is formed by the 1,000 most frequent words in the language, *Word List Two* (2k) consists of the second 1,000 words, *Word List Three* (3k) is the University Word List -UWL- and, finally, the program classifies automatically as belonging to Level 4 (4k) all the words that do not belong to any of these lists. The program calculates the LFP on the basis of types, tokens and word families. In Waring’s words:

“A vocabulary frequency profile measures the amount of words known at various frequency bands as a snap shot at one point in a learner’s progression to higher levels of language proficiency. It is not intended to provide a size figure, but to generate information to see how a learner’s vocabulary is distributed.” (Waring, 1997:53)

¹⁹ He was obviously well-acquainted with Zipf’s Law (1935), which states that the occurrence of words in a language is regular to some extent. In a corpus of natural language utterances, the frequency of any word is roughly inversely proportional to its rank (first, second...twentieth etc. position) in a frequency list.

The creation of this program brought some advantages (Laufer & Nation, 1995): first, it provided a more detailed picture of the different type of words that learners used. Second, it made a distinction between subjects who used frequent and less frequent vocabulary and not just between those who were or were not able to vary their limited vocabularies. Moreover, LFP is claimed to be stable across administrations, to show positive correlations with other measures of lexical knowledge and to work well with relatively short texts.

LFP has lately been used among researchers for different purposes: for evaluation of the vocabulary presented in language classrooms (Meara, Lightbown & Halter, 1997) or textbooks (Milton & Hales, 1997), for analysis of writing development (Laufer, 1994; Lenko-Szymanska, 2002; Muncie, 2002, Lee & Muncie, 2006), as a predictor of academic and pedagogic performance of TESL trainees (Morris & Cobb, 2004), to study the relationship between active and passive vocabulary knowledge (Laufer, 1998) or to assess lexical richness of spoken productions (Ovtcharov, Cobb & Halter, 2006). Although it is claimed to discriminate between proficiency levels, Horst and Collins (2006) found out that in some cases, LFP did not identify the expected increases in use of less frequent words; therefore they complemented their analysis with other indicators such as the Greco-Latin cognate index, which is also an extrinsic measure inasmuch as words are categorised in terms of their origin (that is, whether they are present in a list of cognates or not).

In spite of its extended use, some shortcomings of the LFP have already been pointed out (Coniam, 1999; Meara, 2005), two of them being that the data it produces

is not easy to work with and the mathematics behind not sophisticated enough. Therefore, an alternative approach was proposed by Meara and Bell (2001): *P_Lex*.

P_Lex is a computational tool which assesses the lexical richness of texts and gives information about how frequent the vocabulary learners use is. Among the advantages of *P_Lex*, Meara and Bell (2001:13-14) highlight especially two: it works better than LFP with shorter texts and the output it produces (lambda values) is easier to work with, although the mathematical process it uses to arrive at a final score is more complex. These authors have also found that: 1) *P_Lex* scores are reliably stable across administrations; 2) there is an overall good correlation with other measures of productive vocabulary (the Vocabulary Levels Test -VLT- by Nation, 1990) and scores for groups of different proficiency levels are reliably different; and 3) *P_Lex* can discriminate with short texts.

When a text is load into *P_Lex*, it splits the text into parts of 10 words each (ignoring punctuation)²⁰ and it counts the number of infrequent words in each segment; that is why it is considered a measure of lexical sophistication. The program is able to do this count because it operates on word lists. The output consists first of a graphic showing the proportion of segments containing 0 difficult words, the proportion of segments containing 1 difficult word... up to the proportion of segments containing 10 difficult words (in easy texts we will find a high probability of having 0 or 1 difficult words, whereas in harder texts the probability of getting a large number of difficult

²⁰ Therefore, *P_Lex* will always give us analyses based on a number of words which will be a multiple of 10. For instance, if the text has 193 words it will give us the score based on 190 words (19 segments of 10 words).

words will be higher). Second, it also outputs the lambda, which is the value that describes the graphic given, and its error.

Therefore, the program works on the assumption that difficult words are infrequent occurrences and thus it uses the *Poisson Distribution*²¹ as its basis. This type of distribution on which *P_Lex* works is used in statistics when the number of trials n is big but, at the same time, the probability of success p is very low, so np has a moderate size. The reason why the Poisson distribution resembles the distribution of data produced by SL learners is that both are usually skewed to the left (the probability of having segments with 0,1,2,3 or even 4 infrequent words is higher than the probability of having 10-word segments with 8, 9 or 10 infrequent words). The function that describes this distribution is: $P_N = (\lambda^N \cdot e^{-\lambda}) / N!$, where λ is the average of occurrences and $e=2.71828...$ the basis of natural logarithms. *P_Lex* calculates the theoretical Poisson curve that matches most closely the data our text has produced. Hence, the lambda describes the shape of the curve produced by our text (once the lambda is known, all the other values of the distribution can be known automatically, so we can describe the data curve *P_Lex* produces by using the lambda value). Of course, the match between the theoretical Poisson curve and the curve produced by the text we have loaded into the program is not always perfect and there is an error nearly always, which indicates how close the match is.

²¹ The Poisson formula was also used by Ellegard (1960) when trying to answer the question of how many different words will there be in a text sample of a given size: he divided the vocabulary into several portions, each consisting of words within a definite range of relative frequency, and then calculated the number of words contributed by each portion by means of this formula.

In this section we have reviewed several measures of lexical richness. However, besides knowing how rich a text is according to several measurements, more information on the learner's proficiency could be obtained from his/her production. One of the most challenging queries, both for diagnostic and research purposes, is to find out how large the vocabulary of the writer/speaker is from the sample of words s/he writes/utters.

3.4. Estimations of vocabulary size

3.4.1. Why vocabulary size is important

There are three main reasons why vocabulary size is considered to be important in a SL. First of all, in English, vocabulary size is related to proficiency: the bigger one's vocabulary is, the more proficient in a language. This might not be the case for other languages, but in English the relationship between vocabulary size and how well one understands, reads, writes and performs on other formal linguistic tasks is close (Kelly, 1991; Henriksen, Albrechtsen & Haastrup, 2004; Zareva, 2005), and it is also related with academic achievement (Saville-Troike, 1984). In brief:

“Tests of vocabulary size have been shown to predict success in reading, writing, and general language proficiency as well as academic achievement [...], whereas other types of vocabulary research as yet have not.” (Laufer & Goldstein, 2004:401-402).

Secondly, learners think that vocabulary is one of the most difficult components to master in a FL (Ishihara, Okada & Matsui, 1999; Laufer, 1986), it is considered to be

a very demanding task, even once they have mastered grammar. The third reason, which is also especially significant in the context of SLA is that, initially, learners' skill in using the language is heavily dependent on the number of words they know. As "without words to express a wide range of meanings, communication in an L2 just cannot happen in a meaningful way" (McCarthy, 1990:viii), the "development for second language learners beginning with an emphasis on vocabulary size [...] [is] an essential prerequisite to the development of skill in language use" (Nation, 1993a:131).

On that account, a fair amount of studies are devoted to roughly calculate threshold vocabularies at different learning stages. To serve as example, West (1936/1953) suggests that a minimally adequate vocabulary must consist in at least 2,000 words for communication. Liu and Nation (1985) consider 3,000 word families known receptively as a crucial threshold. This type of data can be put in relation to the amount of vocabulary required to perform certain tasks in a SL, for instance, Laufer (1988) claims that, at any level, in order to guess successfully from context, a 95% of the text should be understood.

This kind of research on vocabulary estimates is also important because, as Zechmeister et al. (1993) suggest, our metacognitive knowledge about how many words we know or at how many words we aim at in the target language is very limited. This superficial knowledge may also be applied to teachers' perceptions of students' vocabulary. Riley and Whistler (2000) report a study where subjective estimates on the part of the teachers are compared with those obtained in a levels test by Japanese learners of English. In addition to a noticeable disagreement between the teachers' judgements, an underestimation of the students' vocabulary was also observed.

Suggestions by Takala (1985) involve more research on young populations to include lower stages of vocabulary development and on end-of-secondary school students with different ability levels. His studies on the English vocabulary of schoolchildren in Finland give estimates of vocabulary knowledge ranging from 450 words in slow learners to 1,500 in fast learners, which are estimations done after 7 years of studying the language (about 450 hours of exposure).

3.4.2. Difficulties in estimating vocabulary size

Estimating vocabulary size is probably the oldest type of vocabulary research. It has not only dealt with estimations based on NSs but also on SL learners, especially from the 1930s, and we notice in studies from this decade onwards that researchers have been confronted with some of the problems that still persist nowadays. Although some authors affirm that the differences between subjective estimations and objective results are not considerable (Ringeling, 1984), consistent deviation between subjective and objective data has usually been found (Zechmeister et al., 1993). A wide variability in estimations tends to be the norm rather than the exception and there are a number of aspects that are a hindrance when trying to obtain estimates.

According to Seashore (1933), difficulties in estimating vocabulary sizes come from a variety of sources: variation (in spoken, written or recognition vocabularies), differences associated with various criteria of knowledge, use of roots or derivative words as the basis of the count and inclusion or exclusion of special terms such as proper nouns and technical vocabulary. To these difficulties Hartmann (1941) adds the

existence of multiple meanings for identical symbolic forms and the unscientific or non-statistical conventions of dictionary makers and printers. Sampling from dictionaries, for instance, reveals itself as one of the major causes of concern (Cooper, 1997) still present in Nation (1993b), who points out the steps that should be -but rarely are- followed in order to obtain reliable samples from dictionary to make tests. Therefore, selecting the source against which a particular text should be evaluated is not a matter that can be resolved in a straightforward way.

3.4.3. 'Theoretical' vs. 'observed' vocabularies'

It is obvious that we cannot test all the words a language has, so any estimate should have a particular source of English vocabulary. As Ellegard already noted, it is impossible and futile for this source to be the language as a whole. For him, the "theoretical vocabulary" should be the frequency ranges coming from "very extensive and methodologically selected material"(1960:240)²². This would allow us to see the relationship between theoretical vocabulary (based on thoroughly mixed samples) and the observed values (based on consecutive texts, that is, an individual's performance on a sample of words). He remarks that "the observed vocabularies remain fairly constant as measures in percent of the theoretical values" and that "hence these percentages may to some extent be used as a measure of the richness, or variety, of an author's vocabulary independent of the text size" (1960:230). Furthermore, one of the most important issues

²² He obtained a frequency distribution of what he calls "English semantic elements", a notion similar to that of morphemes, in order to infer the vocabulary sizes of writers such as Chaucer or Shakespeare.

to bear in mind when selecting this theoretical vocabulary is the ‘subject matter’: Ellegard notes that a Middle English text by Chaucer talking about religion or love would have a different vocabulary from an Elisabethan English text about jealousy and death by Shakespeare. Therefore, the source to represent English vocabulary will probably depend on the purpose of the estimate.

Since the 60s, numerous materials, especially word frequency counts, have been compiled, most of them having as its basis M. West’s General Service List -GSL- (West, 1936/1953), the aforementioned Nation’s Vocabulary Lists (1996), the lists derived from the British National Corpus by Leech, Rayson & Wilson (2001) or the last Jacet List (Ishikawa et al., 2003), compiled in Japan for pedagogical purposes. Consequently, it would seem that with such range of materials available, the selection of ‘theoretical vocabulary’ should be much easier, the estimation process more standard and estimates of vocabulary more consistent. However, as shown in the next section, this is not actually the case.

3.4.4. Vocabulary size estimates

For the purpose of the present study, we are interested in reviewing the studies that estimate vocabulary sizes of learners of English as a SL, especially in formal settings. An overview of those studies is shown in Table 3.1 (see pages 71-73). However, it should be taken into account that techniques for estimating vocabulary have been also used with NSs and that some results are already available. For instance, Seashore (1933) considered that a junior college student knew about 15,000 non-

technical English root words, about 52,000 derivatives of roots and 3,000 special terms as he found with a four choice recognition type of test whose sample came from a dictionary. More recent studies like Goulden, Nation and Read (1990) suggest that average educated NSs know about 20,000 word families, excluding proper nouns, compound words, abbreviations and foreign words. These results were obtained using a sample from the Webster's dictionary and were similar to estimations of a college student by D'Anna, Zechmeister & Hall (1991), about 16,785 different words. Nation and Waring (1997) have offered some indications on the size of children's vocabulary in English as an L1. They state that a five-year-old knows between 4,000 and 5,000 words, of which 2,000-3,000 are also known productively. In secondary education, Cameron (2002) used the VLT and the Yes/No Test to infer receptive vocabulary sizes of native or near-native students in the UK.

In addition to studies with NSs, studies dealing with vocabulary estimates for learners of English in natural settings can also be found. For example, Qian (2002) used the VLT to estimate the receptive vocabulary of 217 university students and undergraduates (beyond the intermediate level) learning English in Canada. He operationalised the estimate as 'VS' ('Vocabulary Size' Measure), giving a mean result of 59.99%, which is not very informative unless we use the same sort of operationalisation. Zimmerman (2004) investigated whether there was a difference between the vocabulary size scores of newly placed students (new arrivals) and continuing students in the US at three different levels. Using the productive version of the VLT, it was unexpectedly found that new arrivals at any given level had larger vocabularies than continuing students at the same levels, with a difference of at least 377

word families. Also in a natural setting, Mochida and Harrington (2006) used a Yes/No test and VLT to infer the receptive vocabulary of 36 undergraduate and postgraduate students learning English in Australia. Finally, it should be pointed out that there is research as well on estimations with learners of a SL other than English (for instance Hazenberg & Hulstijn, 1996 and Eyckmans, 2004 with Dutch).

Author & Year	Country	Subjects	Type of vocabulary	Test	Estimate Vocabulary Size
Gui, S. (1982)	China	Secondary School	Receptive	Multiple-Choice	1,200 words
Takala, S. (1985)	Finland	Primary school (after 450h of instruction) N=2,415	Receptive & Productive	Translation test (direct/indirect)	About 1,000 words receptively and productively (1,500 for the best students and 450 for poor learners).
Jaatinen, S. & Mankkinen, T. (1993)	Finland	Undergraduate and Graduate N=89 (52 first-year, 37 advanced)	Receptive	Two multiple choice vocabulary tests	·18,100 words ·Advanced students knew 2,400 words more than their peers (19,500 vs. 17,100)
Laufer, B. & Nation, P. (1995)	New Zealand & Israel	University students divided into 3 proficiency groups. N=65	Productive	Lexical Frequency Profile (LFP)	(page 316)
Waring, R. (1997)	Japan	Elementary to Upper Intermediate N=76	Receptive & Productive	Vocabulary Levels Test (VLT): Receptive and Productive versions.	·Low proficiency (34%) ·Middle proficiency (46%) ·High proficiency (52%)
Nurweini; A. & Read, J. (1999)	Indonesia (Sumatra)	University (1st year) N=324	Receptive	Translation task (Based on the GSL -first 2,000 words- and the UWL).	1,226 words (receptive)

Author & Year	Country	Subjects	Type of vocabulary	Test	Estimate Vocabulary Size
Cobb, T. & Horst, M. (1999)	China (Hong Kong)	University: N=21 (1st year) N=28 (2nd year)	Receptive	Vocabulary Levels Test (VLT)	(page 64)
Ichihara, K.; Okada, T. & Matsui, S. (1999)	Japan	University (2nd year) N=362	Receptive & Productive	Recognition test (supply Japanese equivalents for English words) Production test: write English words for six semantic contexts	·Receptive: 2,000-2,500 words ·Productive: scores around 30, they correlate with receptive scores.
Fan, M. (2000)	China	University (1st year) N=138	Receptive & Productive	Receptive: VLT Productive: 9 different versions similar to the VLT.	Receptive: ·62.12% (2,000 wordlist) ·48.16% (3,000 wordlist)
Cobb, T. (2000b)	Canada	University (after 9 years of instruction) N=More than 1,000	Receptive	Vocabulary Levels Test (VLT)	·74% (2,000 wordlist) ·68% (UWL)
Tschirner, E. (2004)	Germany	University: (1st year) N=142	Receptive & Productive	Vocabulary Levels Test (VLT) (Receptive and Productive)	72% of the students do not have a receptive vocabulary of 3,000 words and a 79% fail the productive 2,000 level.
Zareva, A. Schwanenflugel, P., & Nikolova (2005)	US & Bulgaria	Undergraduates at University. ·Intermediate (N=17) ·Advanced (N=17)	Receptive	A type of Vocabulary Knowledge Scale (with words selected by sampling from a dictionary)	VS measure derived from the words known in the scale.

Author & Year	Country	Subjects	Type of vocabulary	Test	Estimate Vocabulary Size
Jiménez Catalán, RM., Ruiz de Zarobe, Y. & Cenoz, J. (2006)	Spain	Primary Education (Grade 6)	Receptive	Vocabulary Levels Test (VLT)	Estimates lower than 1,000 words
Miralpeix, I. (2007)	Spain	University N=93 (Advanced) N=64 (Intermediate)	Receptive	X_Lex and Y_Lex	·5,954 words (Advanced) ·3,950 words (Intermediate)

Table 3.1. Studies on estimations of the vocabulary sizes of learners of English in instructional settings. The squares that present a page number refer to the page of the article where the results can be found, as they consist of means and *sd* for different frequency levels, which cannot be summarised further to include in the table.

As regards the studies included in Table 3.1, it must be acknowledged that some of them have other aims different from vocabulary estimations. For example, Nurweini and Read's study (1999) has two aims, the first being estimating size and the second analysing vocabulary depth. In these cases, only the information on the vocabulary size study is presented, that is, taking as example Nurweini and Read's study, we discuss here the method and results to estimate vocabulary size and not the word associates tests and the interview that they conducted to fulfill the second purpose, which was related to the analysis of vocabulary depth. It is also worth mentioning that studies by Laufer and Nation (2005) and Zareva, Schwanenflugel and Nikolova (2005) have been included in the table because they both present research in instructional settings (Israel in the former and Bulgaria in the latter), although a group of subjects in Laufer and Nation learns English in New Zealand and Zareva, Schwanenflugel and Nikolova include also a group of NSs in the US in their study.

Out of the fourteen studies included in Table 3.1, many involve the use of the VLT, either the receptive or productive version (the latter is better known as the Vocabulary Size Test of Productive Ability). It is usual that in low levels, only some of the sections of the receptive test are used (as in Jiménez Catalán, Ruiz de Zarobe & Cenoz, 2006) and in some recent studies, the revised versions by Schmitt, Schmitt and Clapham (2001) are used instead of the original versions of VLT. Other authors in the studies presented adapt this test for their purposes as it is the case with Fan (2000), who creates nine different versions similar to VLT to assess productive vocabulary at different levels. It is worth noticing that, in some way or another, VLT is used in six of the studies summarised in the table. Apart from this test, other ways of assessment

observed in these studies take as their basis other standard frequency counts different from the Nation's lists; it happens in Nurweini and Read (1999), who create their tests using the GSL as a point of departure and Miralpeix (2007), who uses the Jacet list. On the contrary, Jaatinen and Mankkinen (1993) use 100-word samples from the *Collins Cobuild Dictionary* to design their multiple choice tests, the same procedure is applied by Zareva, Schwanenflugel & Nikolova (2005), who use words selected by sampling from a dictionary to test receptive vocabulary knowledge.

It is precisely receptive vocabulary which is more commonly estimated in the literature: of the studies presented here, eight estimate receptive vocabulary (like Gui, 1982 and Cobb & Horst, 1999), five estimate both receptive and productive vocabularies (as Tschirner, 2004) and just one only productive vocabulary (Laufer & Nation, 1995).

This general information offered in Table 3.1 should be further specified. Concerning receptive vocabulary, estimations carried out from data obtained when supplying L1 equivalents or synonyms to the words tested (Ishihara, Okada & Matsui, 1999; Zareva, Schwanenflugel & Nikolova, 2005) are of a different nature from those obtained when matching words with their definitions or synonyms (as it happens in the VLT used in some studies). As regards the productive vocabulary estimated in these studies, it can also be of a different character. Firstly, there are instances of what might be called 'controlled productive vocabulary', that is, where the subject has a clue to produce the word. This is the case in the productive version of the VLT, where the first letters of a written word are given in a context for the learner to supply the rest of the word. It is also the case with Takala (1985), who uses translations for the estimations of productive vocabulary. Ishihara, Okada and Matsui (1999) and Laufer and Nation (1995)

estimate what might be called “free productive vocabulary”. Ishihara, Okada and Matsui (1999) ask the learner to produce words belonging to a particular semantic field and Laufer and Nation (1995) analyse the learners’ compositions using the LFP.

However, estimates resulting from the LFP or the VLT are given as percentages of words known in each of the frequency bands and thus a collection of figures is necessary. This is why some of the results are not included in Table 3.1 and there appear the pages instead: the results for these particular studies entail a whole range of figures for each frequency list -and most of the time these studies also involve different groups-. Therefore, results were not specified in the table if they could not appear either as one or two percent figures (representing how much vocabulary the learner knows in the test) or as a general estimate.²³

Some authors have contemplated the possibility of assembling into one -usually the mean- all the figures coming from a profile or frequency bands in tests like VLT (for instance in Waring, 1997). Laufer (1995) suggested that the figures representing each band of the profile could be turned into two (bands 1k and 2k on the one hand and 3k and 4k on the other) and have a condensed profile. The ‘beyond 2,000’ measure she proposes is then the percentage of words belonging to bands 3 and 4.

There are some studies not included in Table 3.1 that are also worth mentioning. The first three present tests devised to estimate vocabulary size, either receptive like the *Eurocentres Vocabulary Size Test -EVST-* (Meara & Jones, 1988), productive like

²³ Otherwise, Zareva, Schwanenflugel & Nikolova (2005) use a vocabulary size measure derived from the scale devised for that particular study.

Lex30 (Fitzpatrick & Meara, 2004) or both, like the *Computer Adaptive Test of Size and Strength -CATSS-* (Laufer et al., 2004). Others are descriptive studies that give general estimates of vocabulary according to official ministry reports, without being empirical studies or giving the means by which estimate results are obtained. Hui (2004) is an example of this type of reports. This study argues that most university students fail to meet basic requirements in the English College syllabus in China, as they know less than the 4,200 words expected at the end of secondary education. Finally, studies like Jiménez Catalán and Ojeda (2004) and Jiménez Catalán and Moreno (2005), cited in Jiménez Catalán, Ruiz de Zarobe and Cenoz (2006), give general indications of the amount of words used in a composition by Spanish schoolchildren learning English at different proficiency levels, the aim of their study is descriptive and they find that a whole group of Grade 4 students know about 765 types productively and 866 in Grade 5.

The measures used in the present study are of two kinds: descriptive and inferential. As shown above, the construct of vocabulary knowledge is complex and different measures are needed if students' performance has to be fairly assessed. There is no perfect measure and therefore, "it is necessary to develop a whole range of instruments to address the various purposes for vocabulary assessment" (Read, 2000:149).

Some traditional descriptive intrinsic measures of vocabulary richness are used in chapter 5 to assess vocabulary performance. Nevertheless, our study is also an attempt to provide the so-often requested evidence for or against the adequacy of the D measure for vocabulary acquisition research, compared to other more traditional measures. It

would also be of great interest to use D in age-related studies in particular, given the variety of measures that have been used in previous studies, which makes it difficult to generalise and compare the results obtained in different contexts. As Read (2000:209) points out, the fact that researchers use different statistics or compute them differently makes meaningful comparisons impossible. Two of the extrinsic measures presented above (LFP and lambdas) are also applied and discussed in chapter 6. In addition, a new method to estimate productive vocabulary size in different tasks (both oral and written) is proposed and implemented in chapter 7.