CHAPTER 7

Vocabulary acquisition and the learning curve

Michael Daller, John Turlik and Ian Weir

Many studies in a variety of educational contexts show that learning curves are non-linear (e.g. Freedman, 1987 for the development of story telling skills in the first language, DeKeyser, 1997 for the acquisition of morphosyntactic rules of an artificial second language or Brooks and Meltzoff, 2007 for the development of vocabulary in two-year-old infants), but there is no agreement on the best non-linear model which may vary between different contexts. Although there are strong arguments, both on empirical and on theoretical grounds, that a power curve is appropriate in most educational settings (Newell & Rosenbloom, 1981; Ninio 2007) other models have also been proposed (Van de gaer et al., 2009; Verhoeven & Van Leeuwe, 2009). However, little is known about the long-term patterns of vocabulary learning in a foreign language. In the present study we analyse the vocabulary used in 294 essays by 42 students written at regular intervals over a period of two years. We use several measures that focus on vocabulary richness as well as ratings from trained IELTS teachers. Our analysis is supported with structural equation modelling, where a latent learning curve, based on the power law, can be identified. The present study is relevant for the discussion on methodological approaches in the measurement of vocabulary knowledge but also has pedagogical implications, as it allows teachers to identify when a certain plateau has been reached and when further vocabulary learning is only effective with additional pedagogical intervention.

1. Introduction

Learning curves[1] can give valuable insights into how learning takes place, but there is no general agreement on how these curves can be modelled. The idea of modelling learning curves goes back to the early 20th century (Thurstone, 1919) although earlier experimental studies on learning and memory can be traced

1. The term "learning curve" is generally used to describe learning whereas the term "growth curve" is used for other natural phenomena also, such as bacterial growth, cancer cell growth etc. However, this distinction is not made by every author, and the term "growth curve" is also used for learning by some of them. We use the term "learning curve" when we refer to mathematical models but use also the term "vocabulary growth" when we refer to increasing vocabulary size in general.

back to the 19th century, when attempts were made to model the rate of forgetting (Ebbinghaus, 1885/1913). An overview of these earlier studies can be found in Atkinson, Bower and Crothers (1965: 19–28). The main questions that occupy researchers is whether we can find a general learning curve that applies to a large range of learning contexts and how this curve can be modelled. There are various attempts to use curve models for learning in the behavioural sciences. Some of these studies use models that are specific only for certain contexts, some studies claim that their learning curve models apply in many contexts or even apply universally in learning. In many cases the analysis of learning in general is based on experimental studies. Fitts and Posner (1967) discuss a couple of earlier studies of learning not related to language, where practice learning (the learning of new skills through repeated exercise) was involved. They come to the overall conclusion that a power function is the best way of describing this type of learning where "the rate of improvement is reduced as practice continues" (1967: 18; a detailed discussion on the Power Law is given in Section 2.2). They also discuss a number of experimental studies involving language learning but they do not reach clear conclusions about a model for a general language learning curve.

One important aspect of modelling learning curves is the number of measurement points that are available. There are studies on cognitive growth that necessarily need to restrict themselves to linear models, since they have only two points of measurement over time (e.g. Primi, Ferrão and Almeida, 2010 on the relation between the development of fluid intelligence and math scores in secondary school pupils). However, if there are more than two measurement points available, other learning curve models can be considered (e.g. to model a cubic curve, four measurement points are necessary). Choi, Harring and Hancock (2009) assume that an S-shaped learning curve is appropriate in many educational settings and suggest logistic functions as the best models. They refer to earlier research by Hull et al. (1940) who suggested a forward S-shaped curve to model learning in general. The characteristics of such a curve are slow growth at the beginning, followed by an acceleration of the learning process, and then a flattening of the curve towards a plateau. Such a curve starts at an asymptotic lower bound and ends at an asymptotic upper bound. Although a linear model might account for the central portion of the learning process, "it would not be adequate to explain the entire trajectory range of the underlying process itself" (Choi et al., 2009: p. 621). A non-linear model is therefore seen as more appropriate. A widely used sigmoidal function that produces an S-shaped curve (the logistic function) had already been introduced by Verhulst in 1845 for the modelling of population growth.

There has been some criticism of modelling learning curves for samples. "Averaging group data has its limitations" (Larsen-Freeman 2009: 585) and the learning patterns that are found on the basis of group data may not reflect any individual learner (see also Ellis & Larsen-Freeman, 2009; Rast & Zimprich, 2009). Heider and Frensch (2002: p. 404) show that "it is possible to find that power functions fits are perfect, suggesting that learning is continuous ..., when, in fact, the underlying process is discontinuous". Some researchers point out that even if the curves for individual learners are best described by exponential functions, the use of averages will point toward a power function as more appropriate (Myung, Kim & Pitt, 2000), and they argue against using averages. However, in an educational setting, there are always decisions to be made that concern a whole group, such as a class; therefore, generalised findings, such as averaged learning curves, can give valuable information in these settings. In the present study, we use a range of statistical tools to analyse our data from different angles in order to draw a fine-grained picture of the learning processes involved. This includes regression analyses, visual analyses with Loess curves (locally weighted scatterplot smoothing), and structural equation modelling.

2. Language learning curves

2.1 An overview of different learning curves

Although there seems to be no general agreement between the studies discussed so far, we assume, with Newell and Rosenbloom (1981), that there are universals in learning and that we can make some generalisations on appropriate models in this respect. We suggest therefore, that a comparison between different contexts is useful to give us further insight into these universals. In the following section, we give an overview of approaches to model learning in the first (L1) and second language (L2). Since these studies focus on learning curves, they are necessarily based on longitudinal data. Most of these studies are searching for the best non-linear model of learning. This is in line with Larsen-Freeman's argument (2009: 584) that language development is nonlinear and that "It is certainly plausible that there is a nonlinear relationship between a learner's repeating a task and that same learner's showing improvement from a target-language perspective". If this is the case, we assume that vocabulary learning, as an important part of language development, can also be best described by a non-linear curve.

A study that proposes a cubic model for growth in L1 is McDermott et al. (2009). They investigate various aspects of cognitive growth, including math, vocabulary knowledge and listening comprehension. Their participants are more

than 3,000 pupils (age range at the beginning of the study between 33 and 69 months) who are monitored over a period of almost two years. McDermott et al. suggest a cubic model as most appropriate for the learning curve of these children, as it allows the modelling of a plateau effect that is typical for less accelerated growth, due to their summer break (2009: 360). Brooks and Meltzoff (2008), who investigate the vocabulary growth in infants (N = 32) from the age of 10 months to two years, come to the conclusion that a quadratic learning curve is the best model for this group. Brooks and Meltzoff identify three variables (maternal education, infant and adult joint attention in gaze, and pointing) as the best predictors for this stage of vocabulary growth in L1. There are, however, some theoretical problems with quadratic learning curves. Jones et al. (2005) use the Rey Auditory Verbal Learning Test (AVLT) (Lezak, 1995; Rey, 1964) with 169 participants aged 65 and above. They had to recall a list of words that was read to them in five "learning trials". The researchers used linear, quadratic and logarithmic learning curves. Although a quadratic curve leads to an acceptable fit, the authors reject this model since it would imply a gain in the learning process followed by a decline, which seems to be difficult to interpret. They then argue for a logarithmic model because "diminishing gains reaching an asymptote (are) a better conceptual fit to the observed data" (Jones et al., 2005: 303). They therefore use an approximately logarithmic model with "the fifth time-step freely estimated" (Jones et al., 2005: p. 304). This approach tries to find a balance between a purely quantitative, statistical model and a conceptual, qualitative interpretation. Interestingly, Jones et al. argue that vocabulary recall, as measured with the AVLT, is uncorrelated to vocabulary learning as such, and that these are two different processes.

Further support for the use of logarithmic curves comes from a recent study by Verhoeven and van Leeuwe (2009). Their sample consists of 2,819 children in the Netherlands whose reading proficiency in Dutch was tested twice a year during a period of five years, from Grade 1 to 6. Several aspects of word decoding skills were measured and overall a logarithmic model yielded a better fit than quadratic, cubic or linear models. Dale and Spivey (2006) found in research on L1 acquisition that the occurrence of syntactical structures (n-grams), in both the utterances of the children and of the caretakers, follows a Zipf-like distribution. This could be an argument for a function that is based on the power law, a function that can be expressed as $y = ax^k$, where values for $k < 1$ lead to a learning curve that flattens out over time (see also Figure 1 and 2 in Section 2.2.). This means that learning at the beginning of, say, a language course, makes a much larger contribution to the overall vocabulary size than learning towards the end of the course.

Overall, there is no agreement on the most appropriate learning curve in L1 and it might be the case that the most appropriate model differs in different contexts. However, there are strong arguments that the learning curve is non-linear in most contexts. Larsen-Freeman (2009: 584) argues that we "must think longitudinally and nonlinearly" when we investigate learning. This is exactly the approach that is taken in the present study.

Also, there seems to be no agreement between researchers in the analysis of L2 learning. A study with a clear focus on vocabulary learning in L2 is Crossley, Salsbury and McNamara (2009). They carried out a longitudinal study on different aspects of the lexical development of L2 learners (the growth of hypernymic relations and of lexical diversity). Based on the analysis of six learners with 6 data points over a period of one year, they came to the conclusion that the lexical development of the group can be modelled by a linear curve estimation. In addition they analyse the individual developments and come to the conclusion that, in five out of the six cases, there were significant linear trends (Crossley et al., 2009: 320). Ellis and Ferreira-Junior (2009) investigate the acquisition of verb-argument construction by seven second language learners (234 sessions with native speakers were recorded). They come to the conclusion that the distribution of structures in the input of the learners follows a Zipfian pattern and that the acquisition of these structures is affected by the frequency distribution of the input. The verb distribution produces a straight line against the logarithm of the ranks. This is an indication of a function that is based on the Power Law.

It should be noted that one mathematical characteristic of a power curve is the fact that a log-log graph, e.g. with log time on the x-axis and log score on the y-axis, always yields a straight line.

We also have some insight into L2 learning from experimental studies and from an analysis of the possible input in L2. Fan (2006) wants to model growth rates in the possible input for foreign language learners and draws samples from the British National Corpus to illustrate that input. He comes to the conclusion that below 60,000 cumulative word tokens, a cubic curve is the best model for the input, but above that borderline, a power curve would be most appropriate. As it is extremely unlikely that learners would be confronted with very long texts, the conclusion must be that a cubic curve is more appropriate for the input of L2 learners. DeKeyser (1997), on the other hand, argues in favour of power learning curves, in a study of automatisation in comprehension and production. In this study, 61 participants were explicitly taught words and morphosyntactic rules of a fictional L2 over 11 weeks. After they had been taught these words and rules, they were tested in 15 practice sessions on production and comprehension of the fictional language. The aim of this study was to show that proficiency in this language becomes automatized and that, as a result, the reaction time in computer-controlled comprehension and

production tasks gets shorter over time. DeKeyser shows that the decrease in reaction time is best modelled with a power learning curve ($y = x^n$) rather than an exponential function ($y = e^x$). He shows that for his data, the correlation between the reaction time (log scale) and testing sessions (log scale) is almost – 1 (–.932 to –.943), which is typical for data that fit a power function (DeKeyser, 1997: p. 209). This is, of course, a study where a learning curve is falling since reaction time is reduced. To model vocabulary growth, a rising learning curve is necessary, with values for n that are positive.

Many of the studies discussed so far use empirical evidence for the most appropriate learning curve. In the next section we summarise some theoretical arguments for models of learning curves, especially for those based on Power Law functions.

2.2 Learning curves and the Power Law

Every learning process has its upper limitations because of limited human processing capacities (see also Pienemann, Keβler, & Itani-Adams, 2011). The power law, which implies an asymptotic upper limit of learning in various contexts, can be seen as a mathematical approach to formalize this limit or learning "constraint" (Rosenbloom, 2006: p. 47).

A single learning curve is probably counter-intuitive, as learning might be as variable as any other aspect of human performance (Speelman & Kirsner, 2006: p. 54). There are, however, many empirical and theoretical arguments that favour the modelling of learning based on a power law function, and "... the presence of power functions in human learning data is so ubiquitous that the power law of practice has almost become an accepted fact in psychology" (Speelman & Kirsner, 2006: p. 52, see also Lacroix & Cousineau, 2006). If the power law is formulated for situations where, through practice, the time needed to perform a task is reduced, a falling curve with a negative exponent is appropriate. If the power law is used for learning contexts where more and more items are learned but where an upper asymptote is reached, then the exponent is positive but smaller than 1, as illustrated in Figure 1 and Figure 2 ($y = x^{0.5}$ and $y = x^{0.3}$). This results in a steep rise at the beginning of the curve and a constant decrease of the slope over time.[2]

This makes learning curves based on the power law distinct from other curves where there is either now upper asymptote (e.g. quadratic learning curves) or where there is a plateau effect in the middle of the curve with an accelerating growth rate thereafter (e.g. cubic learning curves). Figure 3 and 4 show typical quadratic and cubic learning curves.

2. Figures 1–4 are based on fictional data points

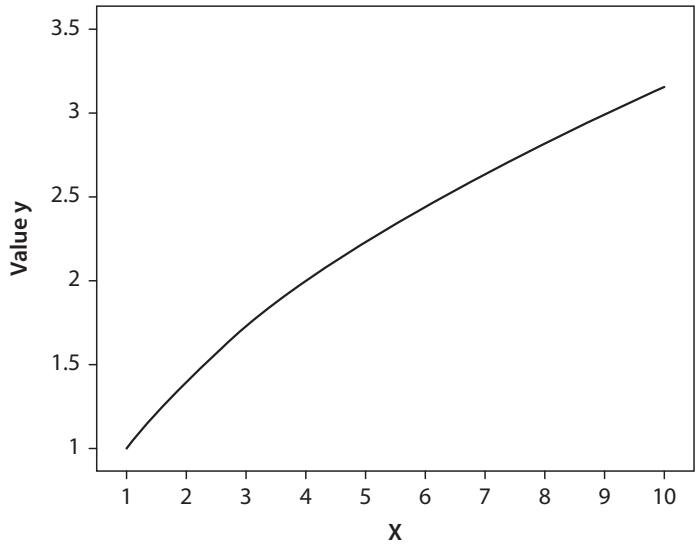

Figure 1. A power law curve for $y = x^{0.5}$

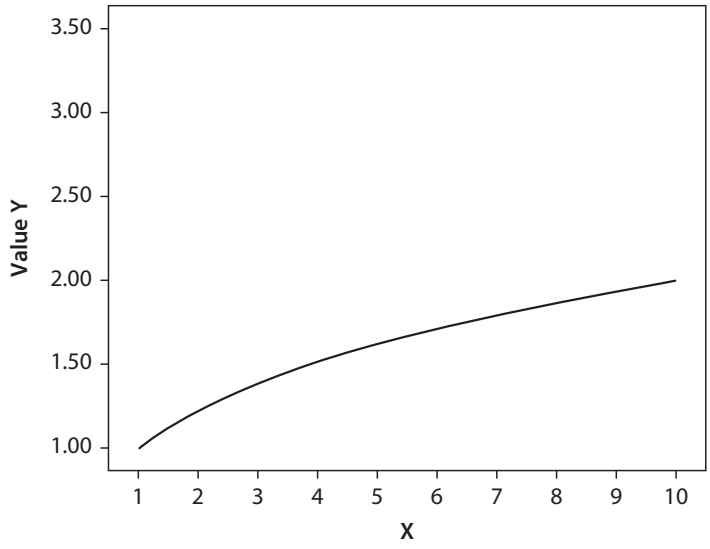

Figure 2. A power law curve for $y = x^{0.3}$

Newell and Rosenbloom (1981) re-analyse data from a wide range of experimental studies going back as far as Snoddy (1926). They come to the overall conclusion that practice learning is best described by power functions and that this "law is ubiquitous over all types of mental behaviour (possibly even more widely)" (1981: 34). Although the power law was first formulated with regard to practice learning, there are strong arguments that it is relevant for learning in general. Newell and

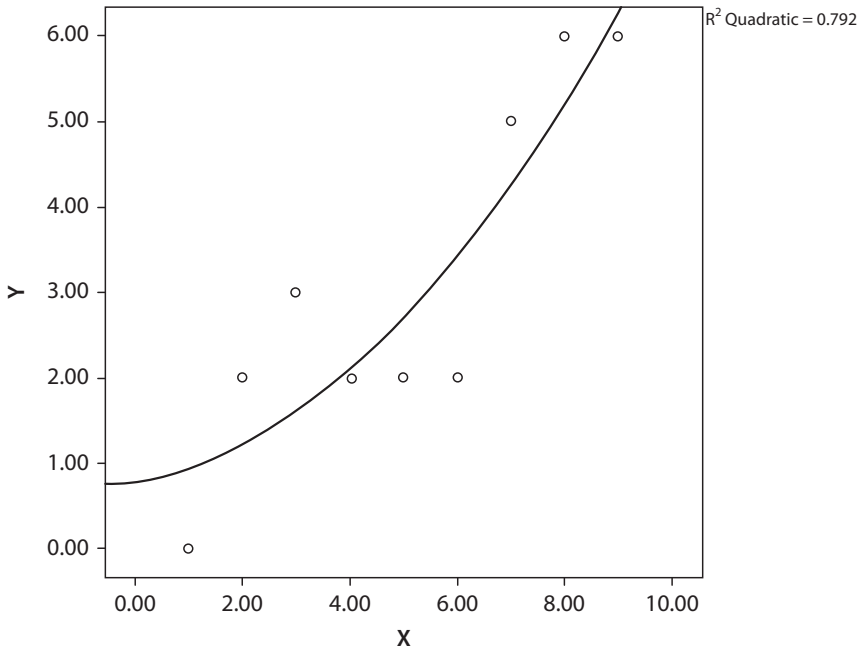

Figure 3. A quadratic learning curve

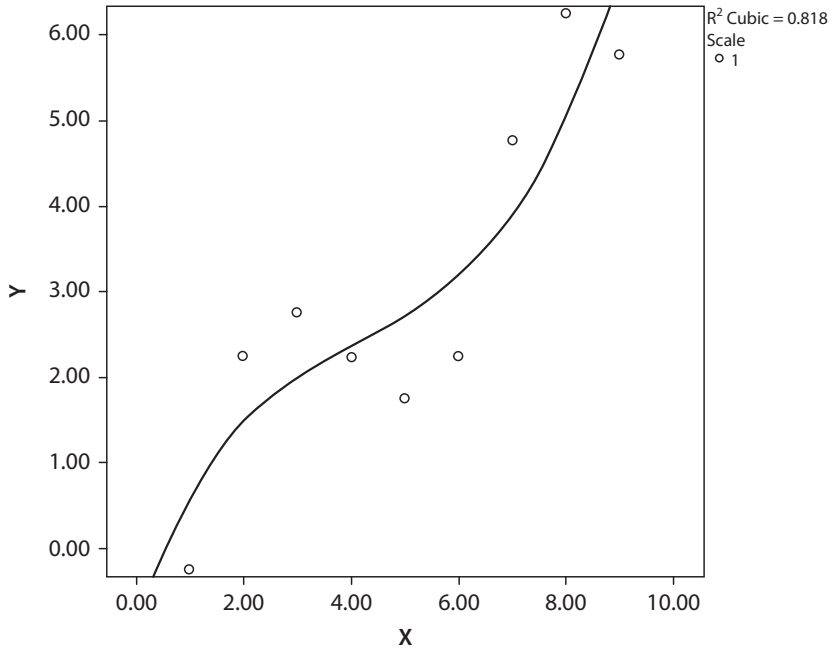

Figure 4. A cubic learning curve

Rosenbloom come to this conclusion on empirical grounds and on the basis of theoretical considerations. They discuss the concept of "exhaustion of exponential learning" (1981: 40 ff.) which states that in learning generally, improvements are harder to find as the learning process advances, and improvements in later learning may be less effective than in earlier learning because later improvements are more specific, and can only be applied in a restricted number of contexts. Newell and Rosenbloom also relate their theoretical discussion to the "chunking theory of learning" (see Miller, 1956). Learning in general starts with the learning of small chunks of information which then become more complex but "the probability of recurrence of an environmental pattern decreases as pattern size increases" (Newell & Rosenbloom, 1981: p. 42; see also Rosenbloom & Newell, 1987; DeKeyser, 2011). Simple chunks are more frequent in the input than complex chunks and are therefore learned easier, and learned more quickly, at the beginning of every learning process, whereas complex chunks are more difficult and their acquisition slows down the learning process. This general argument can be applied to vocabulary learning where the acquisition of more complex chunks, or formulaic sequences, is typical for later stages in the learning of a second language (see Wray, 2008). Overall, Newell and Rosenbloom strongly argue for a power function as the best model for learning in general.

In the same vein, Ritter and Schooler (2002) argue that the power law of practice is ubiquitous for learning and "an important regularity of human behaviour that all theories of learning must address" (2002: p. 8604). They discuss two different theoretical reasons for this. With reference to the concept of hierarchical learning (Newell, 1990; Nerb, Ritter & Krems, 1999) a negatively accelerating learning curve, e.g. a power curve, may be the best model for learning in general, since initially, learning takes place at "low level actions that are very common and thus useful" (Ritter & Schooler, 2002: p. 8604) whereas at later stages, the focus is on larger, infrequent patterns, e.g. complex chunks, and the learning of these new items is slowed down. Again this argument can be adapted to vocabulary learning. At the beginning, learners acquire more general basic words that can be useful in a wide range of contexts. Later in the process, more specific infrequent words are acquired which do not contribute towards an increase in proficiency in the same way as the basic words. Another, related argument would be that the "cognitive system is adapted to the statistical structures of the environment" (Anderson & Schooler, 1991). Learning would then be best described by a power law if the environment (e.g. the distribution of the vocabulary in the input), were also to be modelled by this law (see the discussion below on Zipf's Law).

In line with research on general learning discussed earlier (Fitts & Posner, 1967; Newell & Rosenbloom, 1981) Ninio (2007) argues for a power-law function for language learning both with empirical findings and on theoretical grounds.

Although she is mainly interested in syntactic development in L1, her findings are also relevant for vocabulary learning. Ninio states that "The Power Law of Practice is one of the great success stories of cognitive psychology". She claims that for virtually all learning tasks, the shape of the learning curve "... (is) best described by the mathematical power-law function" (2007: p. 39). Ninio acknowledges that other non-linear functions have been proposed, such as an exponential function or logical sigmoid, but that the power-law function is most appropriate "at least **for the first part** of the learning curve" (2007: p. 43, emphasis added). As the main focus of Ninio's study is on early learning and the acquisition of syntax in L1 by young children, she consequently applies the power-law function and shows that it is the best model for a learning curve in this area. She shows that this model can explain almost all variation in the dependent variable, e.g. the acquisition of Subject Verb Object patterns ($R^2 = 0.97$) (2007: p. 47). Apart from this empirical evidence, Ninio also argues for the power-law function on theoretical grounds, with reference to complexity theory and language as a complex system that can be described by this theory.

Different types of complex systems have been explored over the years. One such system is based on graph theory in which the system is conceived as a network with interconnected nodes. One study that applies this approach to language is Schur (2007) who focuses on word associations in L1 and L2. The power law curve manifests itself in this work in the many words with few connections and the very few words with many connections. In a similar vein, Ninio (2007: p. 125) visualises two types of networks, each of which is characterized by the power law curve. The first is the acquisition of verb units by children and the second is a network of language speakers and words. In both Schur's and Ninio's approach, the distribution of the nodes according to their interconnections cannot be modelled by a normal distribution but must be modelled by a power-law function, with a few nodes with high interconnectivity and many nodes with low interconnectivity (Schur, 2007: 184; Ninio, 2007: p. 122; for further discussion see Watts & Strogatz, 1998; and Barabási & Albert, 1999). Ninio also refers to Zipf's Law, which applies to many natural phenomena (e.g. the amplitude of earthquakes and the population of cities) but is best known for the relation between the frequency of words and their rank (Zipf, 1965 [1935]; Mandelbrot,1966). Again the best model is a power-law function for the relation between the rank of a word (where the most common word in a language has Rank 1 and less common words follow at lower ranks) and the frequency of the words, with some of a higher rank being very frequent and many of a lower rank being very infrequent. If the power law curve is such a prevalent phenomenon in language (be it in the distribution of nodes in a network according to complexity theory, or be it vocabulary as stated in Zipf's

Law), then the power-law function is also a good candidate for an appropriate model for learning a language, at least a first language.

The theoretical arguments for a power-law function for L1 learning made by Ninio (2007) and its empirical underpinnings are convincing but it is unclear whether they can be applied to L2 learning as well. There are fundamental differences between the two contexts because L2 learning is influenced by many other factors, such as repetition, interaction and teacher feedback (see Hall & Verplaatse, 2000) and input mediated by the teacher and the textbook. We therefore do not know whether the power law also applies to L2 learning. Speelman and Kirsner (2006) argue that every task includes new and old components and that a single learning curve for both might be a simplification. "... all human performance is the product of the operation of many component processes" (Speelman& Kirsner, 2006: p. 63). This is certainly also the case for vocabulary learning where many different aspects are involved (for a detailed discussion see Daller, Milton, & Treffers-Daller, 2007). Arguments for other learning curves in L2 acquisition have been made. Meara (1997: p. 115) suggests, on theoretical grounds for the acquisition of L2 vocabulary, that an S-shaped learning curve for incidental vocabulary uptake may be a good starting point for the development of more complex models. We can assume that there is an interaction between the learner's vocabulary size and the uptake of new vocabulary. Beginning learners with a small vocabulary will encounter many new words but if their vocabulary grows, the chance of encountering new words decreases and the vocabulary of the learner reaches an asymptotic upper limit. There might of course be other factors involved and incidental learning might be a "much more random process than we have generally assumed it to be" (Meara, 1997: p. 121), but the assumption of an S-shaped learning curve might be a good starting point.

As mentioned earlier, there has been some criticism on the use of averaged data. In a meta study Heathcote, Brown & Mewhort (2000) come to the conclusion that "The exponential function fit better than the power function in all unaveraged data sets." (Heathcote et al., 2000: p. 185). However, this might not be appropriate for vocabulary learning as an exponential function (exp(x)) has an accelerating slope with no upper asymptote. This makes it implausible as a learning curve for vocabulary since the vocabulary size of an individual cannot be unlimited. Some studies suggest a vocabulary size of around 17,000–20,000 word families for an educated native speaker (Goulden, Nation, & Read, 1990; Zechmeister et al., 1995) whereas others are more conservative and estimate that the receptive vocabulary of native speaker students in the UK is less than 10,000 word families (Treffers-Daller & Milton, 2011). In any case, there is certainly an upper limit for vocabulary knowledge, both in L1 and L2, and a growth curve that has no upper limit is not appropriate.

Overall, we conclude that there is no general agreement on the appropriate curve for learning in general or for vocabulary learning specifically, neither in L1 nor in L2. However, possible models seem to be a learning curve based on the power law (as it is supported in various empirical studies and with theoretical arguments) or a sigmoidal (S-shaped) curve, as there is some theoretical support for this as well. As far as we are aware, there are no longitudinal studies available for the acquisition of L2 vocabulary in a classroom setting. The present study aims to fill this gap in our knowledge.

3. Lexical richness and its measurement

Lexical richness as a cover term comprises several different aspects of vocabulary knowledge (see Read, 2000), such as lexical diversity, lexical sophistication and lexical density. Lexical diversity is the variety of vocabulary that a speaker has at his/her disposal. If the vocabulary is very small, words will be repeated often, which is an indication of low lexical diversity. Lexical sophistication is the knowledge and use of infrequent or rare words and lexical density is the ratio of content and function words in a text. We will deal in this study only with lexical diversity and with lexical sophistication. There are various ways of defining what it means to know a word (see Richards, 1976; Nation, 2001: p. 27; Daller, Milton & Treffers-Daller, 2007) but in the given context, we simply regard every word that is used by the participants in their essays as known, and analyse them with a variety of measures of lexical richness.

A classical measure of lexical diversity is the Type-Token Ratio (TTR), which is the ratio of the number of different words (types) to all words in a text. The more words are repeated by a writer/ speaker, the lower the TTR will be. However, this measure has severe drawbacks. Since the number of words a speaker/ writer knows is always finite, they will be repeated in a text and the longer the text, the more repetition. This means that longer texts systematically get lower scores for the TTR than shorter ones and a comparison between them is seriously flawed (see Malvern, Richards, Chipere, & Durán, 2004).

There have been various approaches to overcome this problem but we restrict ourselves in the discussion to measures that we use in the present study (for an overview of other measures see Turlik 2008). One approach to overcome the problems with the systematically falling TTR is the use of Guiraud's Index (see Guiraud, 1954). Guiraud's work was not focused on learner language but on the discovery of statistical laws in language comparable to Zipf's law (Zipf, 1965 [1935]). Guiraud shows empirically that the ratio between types and the square root of tokens is constant over text lengths between 1k and 100k tokens. He suggested two

measures that are constant over various text lengths: a) types/ √ tokens and b) types/ √ 2 x tokens. Measure (a) should be used when all types are included, measure (b) when function words and very frequent content words – "mots de signification très large" are excluded (Guiraud, 1954: 62). Baayen (2001) shows that Guiraud's index and other mathematical compensations for the falling TTR curve are dependent on sample size and are therefore not suitable to compare texts of different lengths. However, Baayen draws this conclusion on the basis of large texts, e.g. "Alice in Wonderland" with 26505 tokens. Learner texts are much shorter, typically 250–500 tokens and several empirical studies have shown that in the context of learner texts, Guiraud's index is a valid measure to distinguish between different levels of proficiency (Daller & Xue, 2007; Daller & Phelan, 2007; Tidball & Treffers-Daller, 2007; Van Hout & Vermeer, 2007; Housen et al., 2008).

A further approach based on this index, "Guiraud Advanced", has been suggested by Daller, van Hout and Treffers-Daller (2003). Instead of defining function words and frequent content words and removing them from the equation, all words in the first 2k frequency band are taken out and the remaining words are regarded as advanced words. Two adaptations of this measure have been used in a recent study where advanced types were defined by their frequency in two different ways, as being beyond 1k or beyond 2k. The index where all words beyond 1k are counted as advanced shows the highest correlation with tests of verbal intelligence (Mollet et al. 2010). The rationale for taking out the first thousand most frequent words in the analysis is that these words are known by almost all learners and therefore make no contribution to the analysis. It has been argued (Daller, van Hout, Treffers-Daller, 2003) that "Guiraud Advanced" is also a measure of lexical sophistication because it takes the frequency of types into account. Another measure of lexical richness that has been introduced recently is the measure "D" (Malvern et al., 2004). Instead of compensating for the decreasing TTR curve, this measure models the falling curve with a function that uses only one single parameter ("D"). Speakers/ writers with a larger lexicon will have a falling TTR curve with a less steep decrease than speakers/ writers with a smaller lexicon. The former will get a higher value for the parameter "D" than the latter and this parameter is therefore an indication of the size of a person's vocabulary. The measure "D" has been successfully used in a variety of contexts (Stroemqvist et al., 2002; Malvern & Richards, 2002; Yu 2010) although recently, questions have been raised whether it would be necessary to model the falling TTR curve with "D" since alternative approaches are available (McCarthy & Jarvis, 2007; and McCarthy & Jarvis, 2010). For the present study we restrict ourselves to "D" and the Index of Guiraud as measures of lexical diversity because both of them have been successfully used in second language research (see references in the previous paragraph).

4. Hypotheses

The literature reviewed in the preceding sections motivates the following hypotheses.

1. A non-linear curve is the most appropriate model for vocabulary learning in the present context for productive vocabulary growth over a two year span.
2. The non-linear learning curve can be modelled by a power function.
3. Lexical diversity and lexical sophistication show different patterns in their respective learning curves.
4. There is a huge variation between learners and not all follow the average learning curve.

5. Methodology

5.1 Participants

The 42 participants were all female students, United Arab Emirati nationals who entered higher education after successfully completing high school, having chosen to come to a university in the United Arab Emirates that offers bilingual degree courses. The students mostly came from government schools, where the medium of instruction is Arabic, as opposed to private or international schools, where the medium of instruction is English. After an entry test they attended a two-year (maximum) foundation English programme before they studied their chosen subject. At the end of the foundation programme aggregated IELTS scores of 5.0 or above are required for successful completion of the programme.

5. 2 Measures and procedure

The foundation programme has 80 teaching weeks, 20 hours per week. Essays were written after every ten weeks of teaching. However, because the sample size dropped towards the end of the programme we focused only on essay 1 to 7 in our analysis, and in total we had 294 texts from 42 participants. The essay titles were very general to allow the students to use as much of their productive vocabulary as possible[3]. Ideally, the different essay titles would have been counterbalanced to eliminate the possibility that the topic had an effect on the vocabulary. This was,

3. The following essay titles and instructions were used with slight modifications according to different groups of students:
Level 1: What do you think makes a 'good' school?
Level 2: Look at the picture. Describe the woman. Describe her day. What is her daily routine every morning, afternoon and evening?

however, not possible in the given educational setting. The students were given one hour and 15 minutes for each essay, and the handwritten essays were transcribed, verbatim, by a specialist, academic secretarial agency into computer-readable texts. These texts were then edited according to a set of procedures, to ensure consistency. Spelling was corrected, as the view was taken that spelling is part of the learning process and a word used in the correct context but spelt incorrectly should be acceptable.[4] The essays were then transcribed into CHAT format (MacWhinney, 2000a and 2000b) and analysed with the help of CLAN tools, where the command *vocd* can be used to compute values for "D". We also used the programme RANGE (Lists used: GSL and AWL; see Nation URL) to analyse the frequency of the words used in the essays and to determine which words can be classified as advanced. It was also necessary to correct the spelling because otherwise these programmes would identify words that are misspelled as advanced words since they are not in the basic word list. We use Guiraud Advanced as a measure of lexical sophistication, as discussed earlier, and the number of types that are in the Academic Word List (Coxhead, 2000). In addition we asked two experienced teachers who are trained IELTS raters to judge[5] the essays according to the band descriptors of IELTS (International English Language Testing System) in two ways: a holistic rating and a lexical rating based on the range and accuracy of the vocabulary used. IELTS is a widely used English Language Test (see IELTS URL) with individual scores ranging from 0 to 9.

The texts were analysed with the following measures:

Level 3: Use the pictures on the next page. Write five sentences in the past tense about what happened to Eiman yesterday.
Level 4: Imagine that you took a trip last year to collect money to help poor people in your country. Describe a journey.
Level 5: Write a cause and effect essay about damage to the environment.
Level 6: In the lecture, you were given information about several employment-related results from globalisation. Choose *one* problem and discuss its causes and possible solutions.
Level 7: What are some of the causes of internet addiction? What are some of the effects or problems that people with internet addiction have? Choose – 'What are some solutions to these problems?' **or** 'What is the best solution for helping people with internet addiction?'

4. It was beyond the scope of the present study to analyse whether the words were used correctly or not. We welcome the comment of one reviewer of this chapter that this might be a focus for future research.

5. We would like to thank the International English Language Testing System for their permission for the two teachers to rate the essays.

Table 1. Measures of lexical richness

Aspect of vocabulary	Measure	Formula
Lexical diversity	Index of Guiraud (G)	Types ÷ √ Tokens
Lexical diversity	"D"	D/N $[(1 + 2N/D)^{.05}–1]$[6]
Lexical sophistication	Number of advanced types (percentage of all types)	As defined by the word lists included in the RANGE programme
Lexical sophistication	Guiraud Advanced	Advanced Types ÷ √ All Tokens

Table 2. IELTS ratings

Aspect of EFL proficiency	Measure	Scores
Holistic rating	Mean score from two raters	From 0–9
Lexical rating	Mean score from two raters	From 0–9

6. Results

6.1 Text length

In addition to the measures of lexical richness listed in Table 1 we analysed the text length of the essays at the seven points of measurement. Since the time given for each essay was the same at each measurement point (one hour and 15 minutes), an increase in text length can be seen as an indication of learning (e.g. lexical learning or essay writing skills). We used two measures, the number of tokens (all words) and the number of types (all different words). Figure 5 and Figure 6 show the median and the spread (interquartile range and outliers) for these measures in the 7 sets of essays[7].

The boxplots show the mean and the interquartile range (shaded box) for the variables "types" and "tokens". It is apparent that students produce longer texts towards the end of the course. This is in line with general expectations. A first visual interpretation of the medians in each essay set suggests that the increase is not linear, with a steeper increase at the beginning and a flattening out towards the later essays. Apart from this, the boxplots show that there is a huge spread of the number of types and tokens in all essays. Student No 41 produced as few as ten types in the first essay whereas student No 35 produced 96 types. There are clear outliers in this data set but we do not have any evidence that they have a specific learning history or other factors that would justify excluding them.

6. For a detailed discussion of this formula see Malvern et al. (2004: 47 ff.)

7. All computations were carried out with SPSS 19.

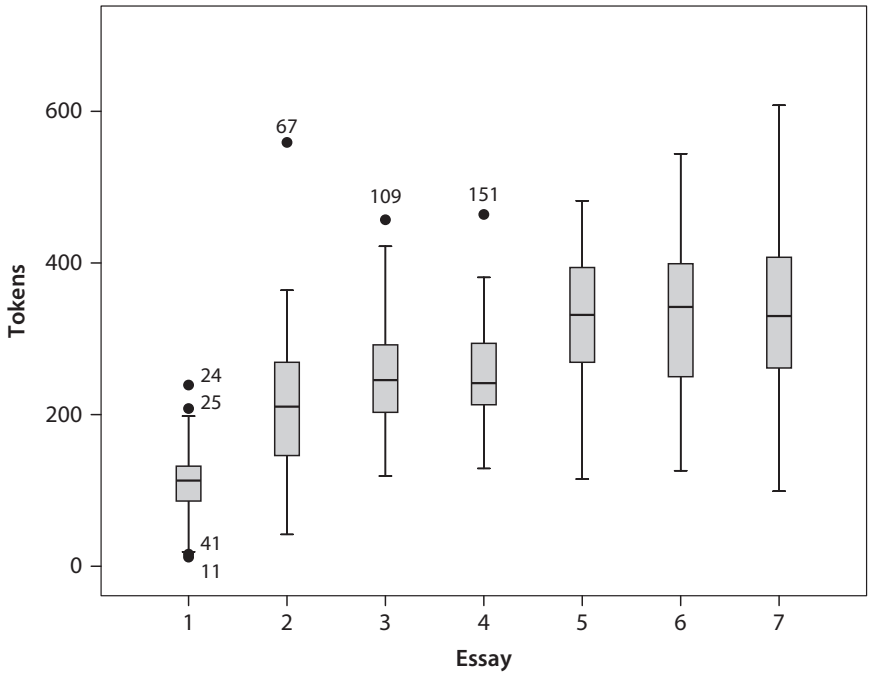

Figure 5. Tokens

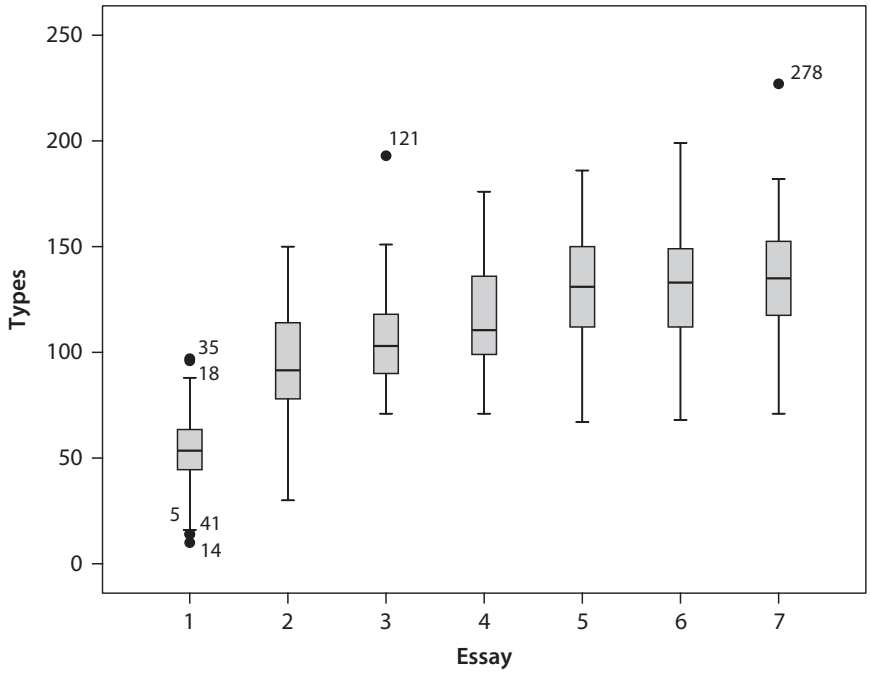

Figure 6. Types

In order to find out whether the apparent increase in scores is significant, we carried out an omnibus ANOVA (repeated measures, Bonferroni correction) and applied the Greenhouse-Geisser correction, since the assumption of sphericity was violated, as shown by Mauchly's W (for tokens: W = .315, df = 20, p < .01; for types: W = .328, df = 20, p < .01). After the correction, the differences between the number of tokens in the seven sets of essays collected at the seven different times are still highly significant (F = 54.069, df = 4.328, p < .001, η_p^2 = .60) as are the differences between the number of types (F = 57.484, df = 4.345, p < .001, η_p^2 = .615). The fact that the essay lengths increase significantly during the course is certainly in line with the general expectations and can be explained by an improvement of a number of subskills, such as planning and organisation of writing, but is certainly also due to an increase in vocabulary knowledge. This will be analysed in the following section.

6.2 Measures of lexical richness

We produced similar boxplots for the four measures of lexical richness and computed omnibus ANOVAs for repeated measurements to test whether the differences between the tests are significant. The results for the two measures of lexical diversity are given in Figures 7 and 8.

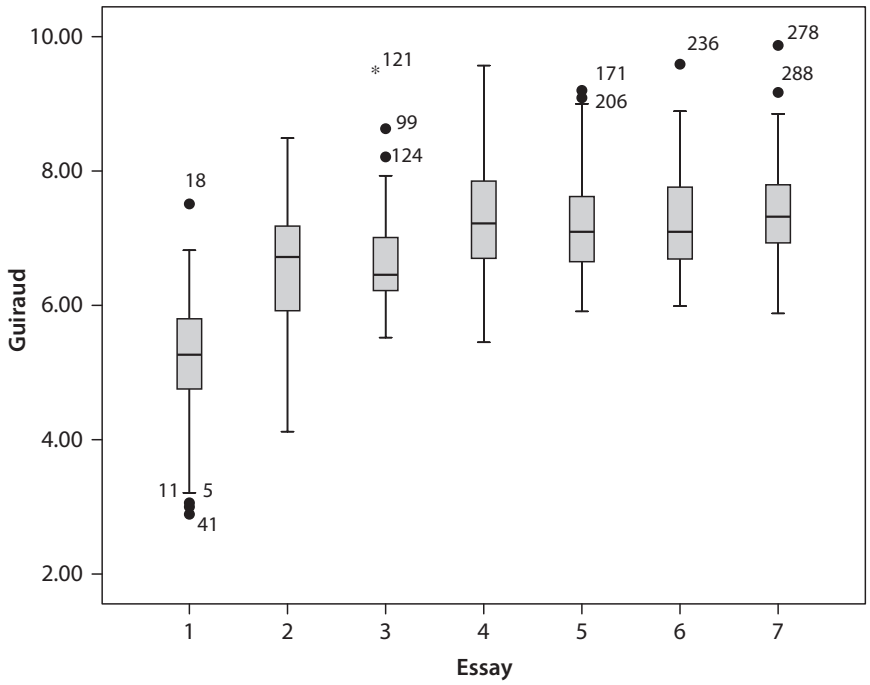

Figure 7. Values for "Guiraud's index"

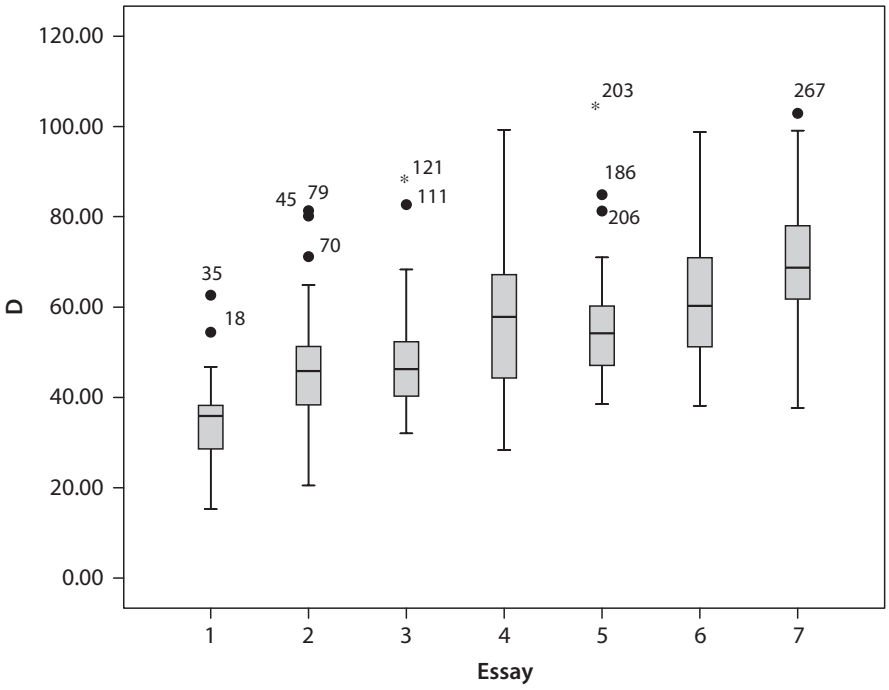

Figure 8. Values for "D"

A visual inspection of these two figures indicates that there is no linear growth, especially for Guiraud, where there seems to be a steep increase in the beginning and a flattening off towards later essays. In order to find out whether the apparent increase in scores is significant, we carried out an omnibus ANOVA (repeated measures, Bonferroni correction) and applied the Greenhouse-Geisser correction since the assumption of sphericity was violated, as shown by Mauchly's W (for Guiraud: W = .308, df = 20, p < .01; for D: W = .216, df = 20, p < .01). After the correction, the differences between the scores for Guiraud in the seven essays are still highly significant (F = 36.191, df = 4.504, p < .001, η_p^2 = .501) as are the differences between the scores for D (F = 31.399, df = 4.571, p < . 001, η_p^2 = .520).

The results for the measures of lexical sophistication are given in Figures 9 and 10.

There seems to be some increase for advanced types and Guiraud Advanced, as the median scores are higher for the later essays. However, this increase seems to be less pronounced for these measures than for the measures of lexical diversity, Guiraud and D. Again we have huge variability with quite a few outliers. We carried out an omnibus ANOVA (repeated measures, Bonferroni correction) and applied the Greenhouse-Geisser correction, since the assumption of sphericity was violated, as shown by Mauchly's W (for Advanced Types: W = .202, df = 20, p < .001; for

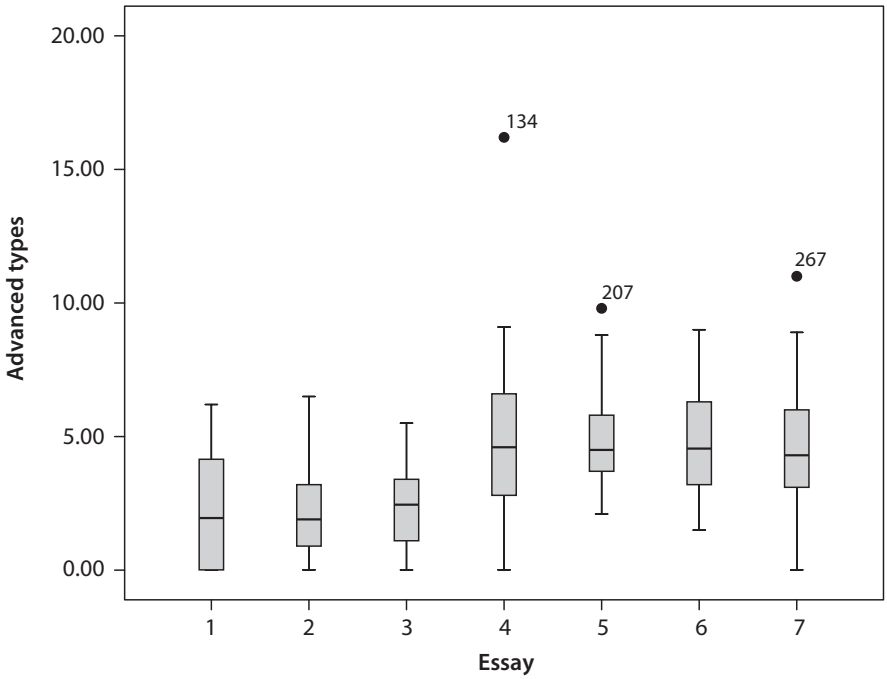

Figure 9. Advanced types (percent of all types)

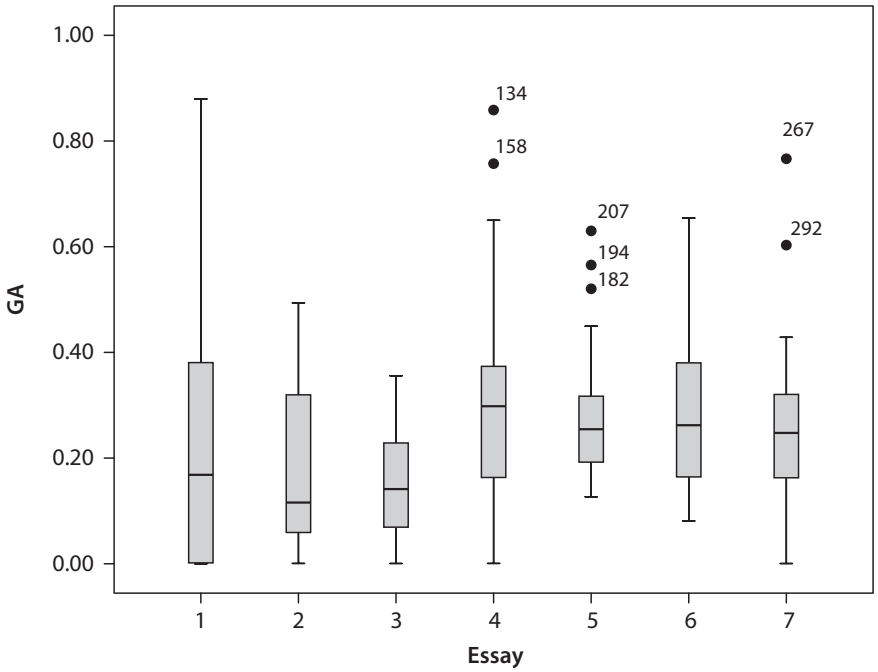

Figure 10. Values for "Guiraud Advanced"

Guiraud Advanced: W = .147, df = 20, p < .001). After the correction, the differences between the scores for Advanced Types in the seven sets of essays are significant (F = 16.783, df = 4.052, p < .001, η_p^2 = .306) as are the differences between the scores for Guiraud Advanced (F = 22.248, df = 3.876, p < .001, η_p^2 = .376).

6.3 Ratings by teachers based on IELTS band descriptors

Two experienced teachers were asked to give a holistic rating of the essays and a lexical rating based on IELTS band descriptors. The correlation between the ratings of the two teachers was only modest but significant for both, holistic (r = .477, p < .001) and lexical rating (r < .588 p < .001). The results for the holistic rating are shown in Figure 11.

There seems to be a non-linear increase of scores. In order to find out whether the apparent increase in rating scores is significant, we carried out an omnibus ANOVA (repeated measures, Bonferroni correction) and applied the Greenhouse-Geisser correction, since the assumption of sphericity was violated, as shown by Mauchly's W for the holistic rating (W = .370, df = 20, p < .029). After the correction, the differences between the holistic ratings of the essays are significant (F = 39.461, df = 4.629, p < .001, η_p^2 = .523). Figure 12 shows the lexical ratings of the essays.

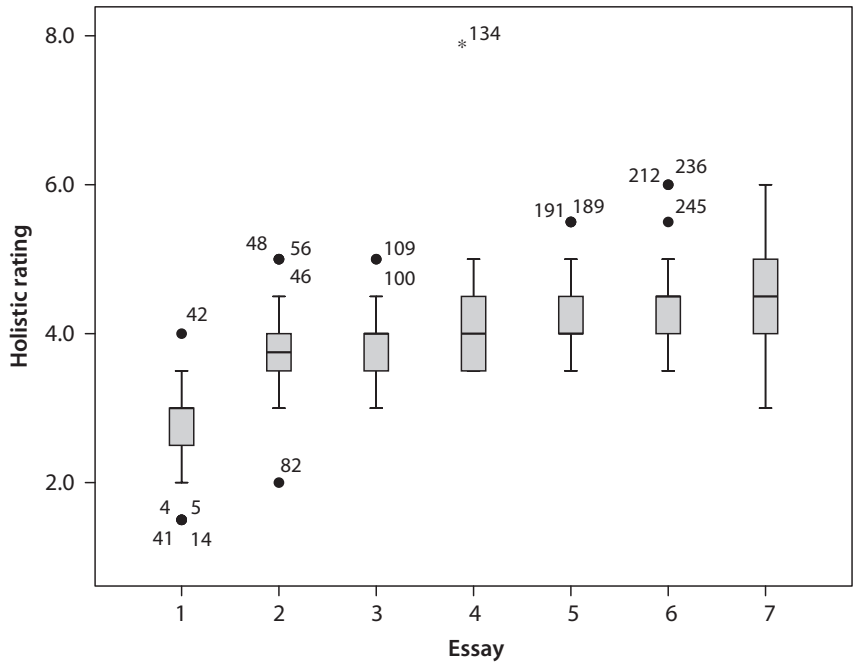

Figure 11. Holistic Rating

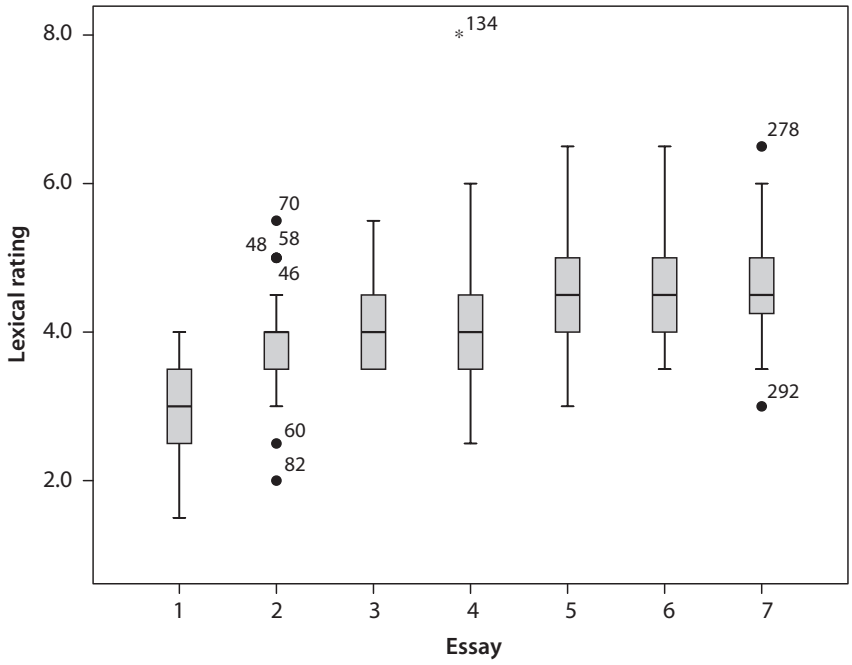

Figure 12. Lexical Rating.

Again there seems to be a non-linear increase of the scores. For the lexical ratings, Mauchly's W is not significant and we can use the unadjusted values for a repeated-measures ANOVA which shows that overall the differences are significant ($F = 35.461$, $df = 6$, $p < .001$, $\eta_p^2 = .496$).

6.4 Linear fit lines and Loess curves

As suggested by Larson-Hall and Herrington (2009) we used Loess curves (locally weighted scatterplot smoothing) in SPSS in addition to linear regression lines to investigate visually whether a departure from a linear model for the different measures is plausible. As there are many outliers, which is normal for real-life data, we produced the following graphs based on the raw data and on data with outliers removed.

Four measures show a picture for the Loess line that suggests a non-linear curve for vocabulary growth, with a steeper increase at the beginning of the course and a flattening out towards the end (number of Tokens, number of Types, Guiraud and the Holistic ratings). Three measures (Advanced types, Guiraud advanced and lexical rating) show a Loess line where there is no increase at the beginning, followed by an increase in the middle of the course and a flattening out towards the end. In all cases, a non-linear curve is plausible. The only exception seems to be "D" where there is not much departure from a linear curve. All measures that

include a definition of advanced types show a very low increase where R^2 is small, in the case of Guiraud advanced, it is .035 with outliers and .028 without outliers. This means that hardly any variance of Guiraud advanced can be predicted. A possible explanation is that the definition of advanced types is not suitable for the specific classroom situation of our participants. In such a setting, the input is mediated through the teacher and the textbook, and therefore our definition of advanced words, based on Nation's range programme (Nation URL) might not be suitable. The visual inspection of the linear and Loess lines indicates that in most cases, a non-linear curve might be the best model. The best candidate for a non-linear curve with a steeper increase at the beginning and a flattening out towards the end seems to be the holistic rating. In our case, the removal of outliers does not alter the curved pattern of any measure. Outliers are therefore included in the following computations, especially since we have no information on which outlier might be due to exceptional circumstances of the individual student.

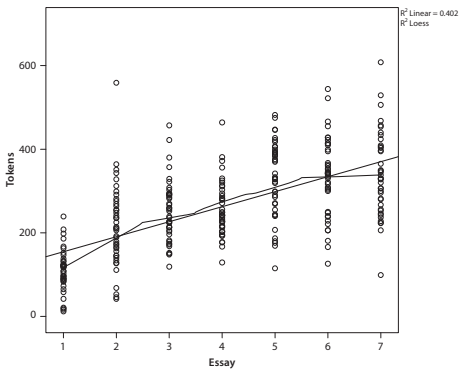

Figure 13. "Tokens"

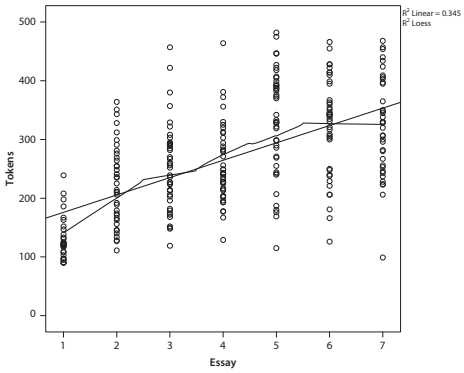

Figure 14. "Tokens" outliers removed

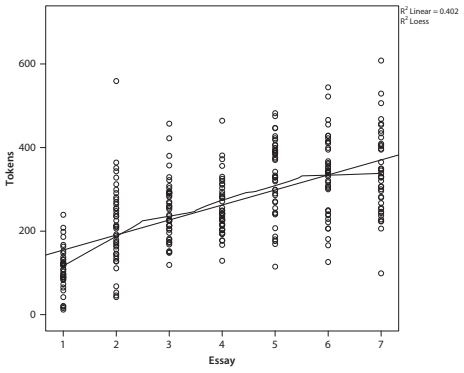

Figure 15. "Types"

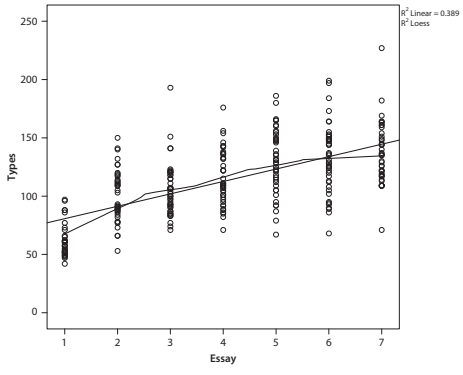

Figure 16. "Types" outliers removed

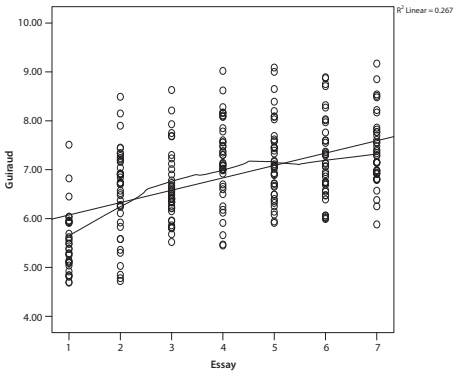

Figure 17. "Guiraud's index"

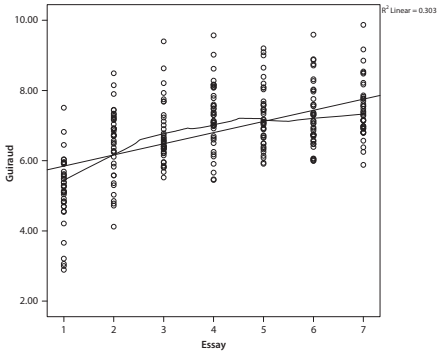

Figure 18. "Guiraud's index" outliers removed

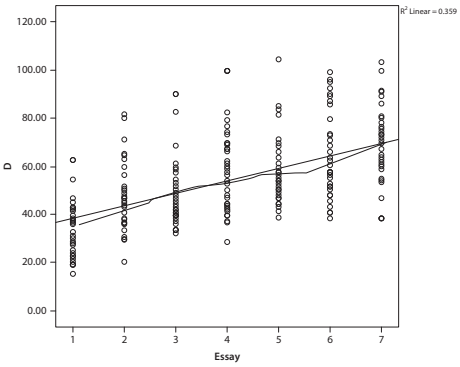

Figure 19. "D"

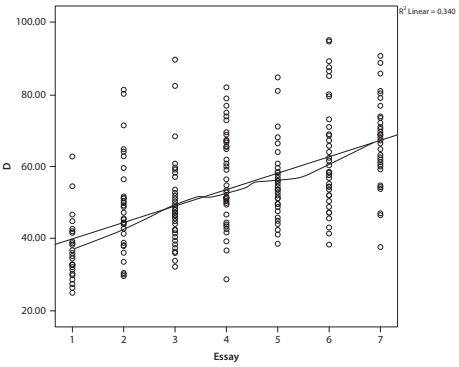

Figure 20. "D" outliers removed

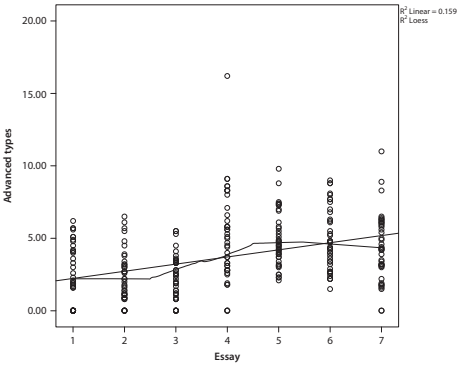

Figure 21. "Advanced types"

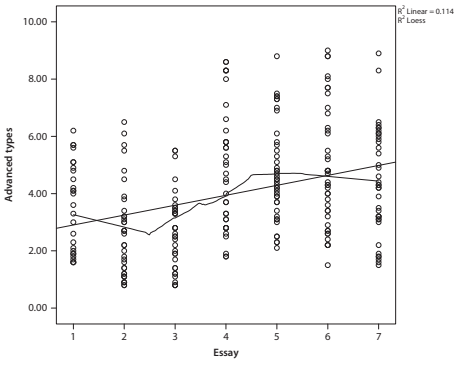

Figure 22. "Advanced types" outliers removed

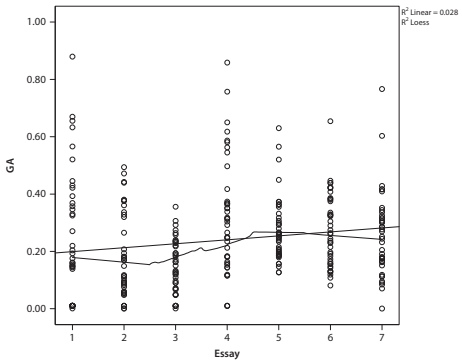

Figure 23. "Guiraud advanced"

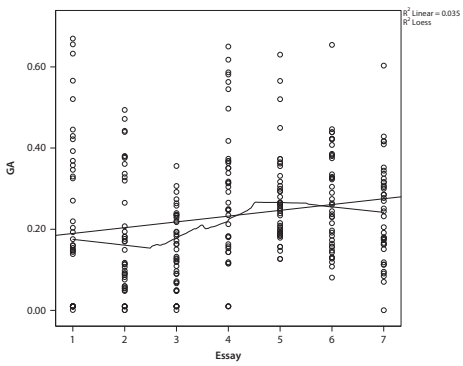

Figure 24. "Guiraud advanced" outliers removed

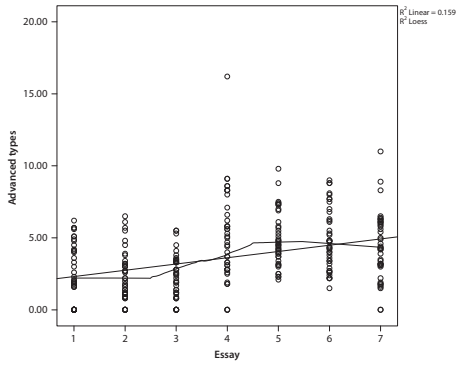

Figure 25. "Lexical rating"

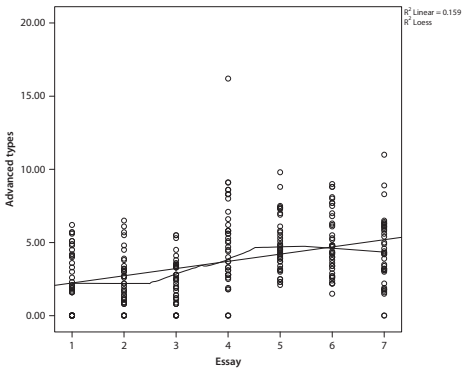

Figure 26. "Lexical rating" outliers removed

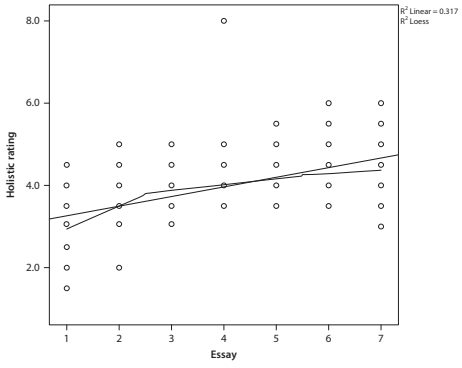

Figure 27. "Holistic rating"

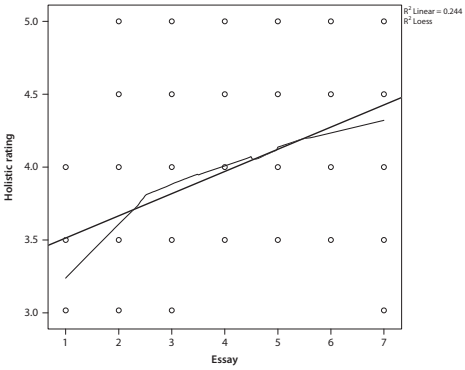

Figure 28. "Holistic rating" outliers removed

6.5 Fit lines

In order to find out which curve would fit the data best we used the curve fitting tool under the regression command of SPSS. A regression, as such, is not possible here as we have repeated measures but finding lines with the best fit is still possible, treating the data as a time series. In comparing linear, quadratic and power fit lines, R^2 was compu ted as an indicator of the fit of the line. Table 3 gives the R^2 value for three different lines in the six measurements of our data set.

From these data, it is obvious that there is a lot of variation or background noise, but nevertheless, a power line yields the highest and a linear line the lowest value for R^2 in all but one of the cases. The best fit can be achieved for the measures of lexical diversity – "Guiraud" and "D" – and for the ratings by judges. The two measures that are based on advanced types produce the highest amount of background noise and are less robust than the other measures, at least in the given context.

A classical way of testing the appropriateness of the power curve is a log transformation, where the correlation between the log of time (in our case essay number) and the log of the measurement should be a straight line, with ideally a correlation close to 1. Table 4 gives the correlations between log time (essay number) and the log for the six measures.

Although all correlations are significant, the correlations for the measures that are based on advanced types (Advanced Types and Guiraud Advanced) are far removed from 1 and we draw the conclusion that a power law curve is certainly not an appropriate model. The correlation between the log of time and the log of the other four measures indicates that a power law curve might be an appropriate model, although the amount of background noise is very high and the model can only explain part of the development. The next section will probe further into the appropriateness of the model based on the power law.

Table 3. R^2 for fit lines

	Advanced Types	Guiraud Advanced	Guiraud	D	Lexical rating	Holistic rating
Linear	.159	.028	.303	.359	.317	.307
Quadratic	.176	.028	.380	.361	.389	.368
Power	.156	.086	.395	.417	.422	.395

Table 4. Log correlations between time and the six measurements

	Advanced Types (log)	Guiraud Adv. (log)	Guiraud (log)	D (log)	Lexical rating (log)	Holistic rating (log)
Log time	.333**	.126*	.628**	.645**	.629**	.650**

* significant at $p < .05$, ** significant at $p < .01$

6.6 The latent growth curve

So far we have dealt with empirical learning curves directly observed from the data and all our measures are directly measured (manifest) variables. It is, however, possible to investigate the latent learning curve that underlies these manifest variables, using structural equation modelling. Due to the huge variability in our data, it is not possible to model this underlying learning curve for all our measurements but it is possible to find a structural equation model based on the holistic ratings. This model is given in Figure 29.

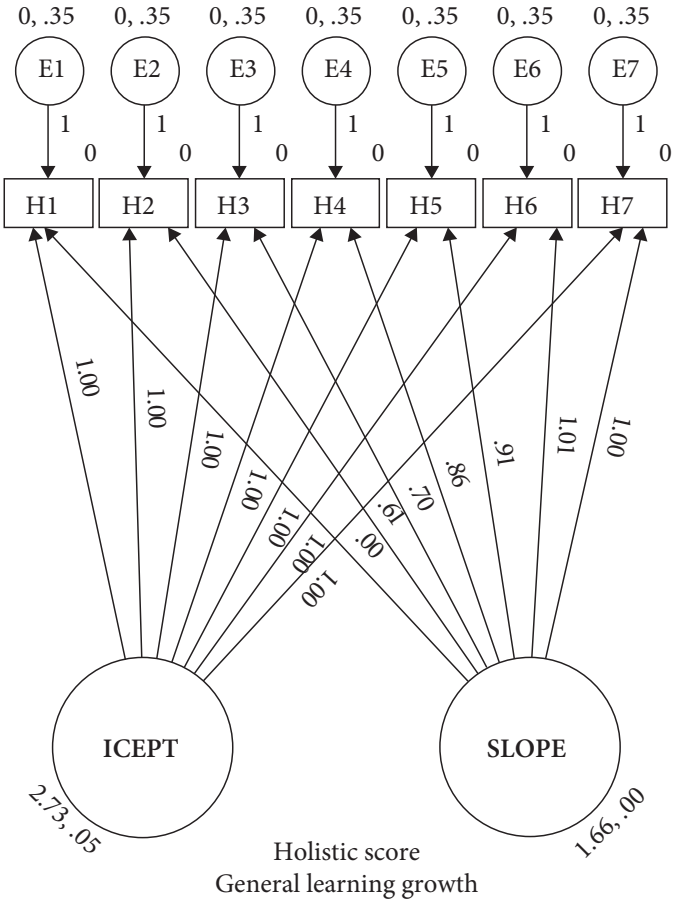

Figure 29. Learning curve based on holistic ratings

In this model, H1 to H7 stands for the holistic ratings of the essays, and E1 – E7 for an error term that has been kept constant. The non-significant p-value indicates that the model is a good fit as it does not differ significantly from the data. The non-zero (2.73) variability of the intercept parameter indicates that there are different starting abilities, which is in line with the general expectations and with the findings from the previous sections. It is possible to identify the latent growth curve with this model by looking at the values for the slope parameter. This parameter is set to zero for Essay 1, the point at which learning starts (0%), and is set to 1.0 for Essay 7, the point at which all learning has been completed (100%). Learning before Essay 1 and after Essay 7 is not analysed in the present study. The increments between 0 and 1 of the slope variable show the shape of the latent learning curve. This shape is given in Figure 30 together with possible power functions with an exponent ranging from .2 to .5.

In Figure 30 the 95% confidence intervals for the shape of the coefficients are given and the figure shows that the learning curve can best be modelled with a power function with an exponent of .3. Since this figure is derived from a structural equation model that does not significantly deviate from the data (see Figure 29) we come to the conclusion that a learning curve based on the power law is an appropriate model for the holistic ratings.

7. Conclusions

The present study was based on real life, longitudinal data from 42 students, who collectively wrote 294 essays over a period of two years (80 teaching weeks). In our

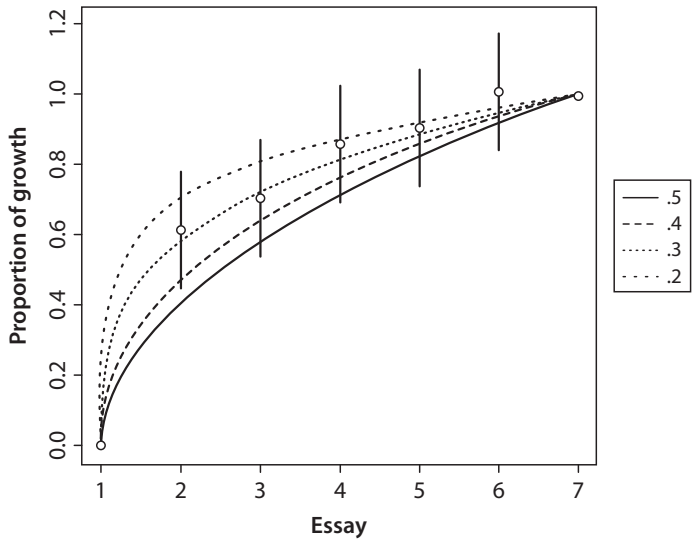

Figure 30. Latent learning curve models

opinion, this data set is unique in covering such a wide range of essays over such a long time period. Despite the fact that such free-production data inevitably contain a substantial amount of variation or background noise, it is possible to identify a latent learning curve, based on the holistic ratings of human judges, with structural equation modelling. This learning curve can be modelled with a power function.

The findings from the directly observed vocabulary measures in the present study are supported by a more in-depth analysis with structural equation modelling, in which we base our analysis on holistic teacher ratings relying on the IELTS band descriptors. As usual with structural equation models, we seek a model that is not in conflict with the data, and our non-significant value for the χ^2 shows no such conflict and therefore the model is appropriate. The model that we obtain with this method suggests a learning curve based on the power law, with a suggested value for the exponent of .3. The fact that structural equation modelling was only possible on the basis of a variable that represents teacher ratings may be an indication that our automated measures, as elaborate as they are, still cannot entirely replace human judgements. One has, however, to bear in mind that the holistic ratings were a measure of proficiency in general, whereas all the other measures dealt more specifically with lexis. This might indicate that the problem is not about whether automated measures are used or not, but about what is measured.

Nevertheless, the findings from the different analyses in the present study support each other and strongly suggest that vocabulary learning does not follow a linear pattern. Although a learning curve based on the power law could only be identified for the holistic ratings, such a curve is still a strong candidate for vocabulary learning in general. This is an important finding since vocabulary growth shows the same characteristics as learning in many other settings, where a steep increase at the beginning is followed by a flattening out of the learning curve in later stages. This has pedagogical consequences for the design of language courses. Once the learners reach a certain plateau, only additional input, for example in the form of extra classes, can improve further vocabulary learning. It is of great interest for practitioners to find out when this plateau has been reached in order to make informed decisions about further teaching.

References

Anderson, J.R., & Schooler, L.J. (1991). Reflections of the environment in memory. *Psychological Science*, *2*, 396–408.

Atkinson, R.C., Bower, G.H. & Crothers, E.J. (1965). *An introduction to mathematic learning theory*. New York, NY: Wiley.

Barabási, A., & Albert, R. (1999). Emergence of scaling in random networks. *Science, 286*, 509–512.

Baayen, R.H. (2001). *Word frequency distributions*. Dordrecht: Kluwer.

Brooks, R., & Meltzoff, A.N. (2008). Infant gaze following and pointing predict accelerated vocabulary growth through two years of age: A longitudinal, growth curve modelling study. *Journal of Child Language*, *35*, 207–220.

Choi, J., Harring, J.R., & Hancock, G.R. (2009). Latent growth modeling for logistic response functions. *Multivariate Behavioral Research*, *44,* 620–645.

Coxhead, A. (2000). A new academic word list. *TESOL Quarterly*, *34*(2), 213–238.

Crossley, S., Salsbury, T., & McNamara, D. (2009). Measuring L2 lexical growth using hypernymic relationships. *Language Learning*, *59*(2), 307–334.

Dale, R., & Spivey, M. (2006). Unraveling the dyad: Using recurrence analysis to explore patterns of syntactic coordination between children and caregivers in conversation. *Language Learning*, *56*, 391–430.

Daller H., van Hout R., & Treffers-Daller J. (2003). Lexical richness in spontaneous speech of bilinguals. *Applied Linguistics, 24*(2), 197–222.

Daller, H., Milton, J., & Treffers-Daller (2007). Editors introduction. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 1–32). Cambridge: CUP.

Daller H., Milton J., & Treffers-Daller, J. (Eds.) (2007). *Testing and modelling lexical knowledge*. Cambridge: CUP.

Daller, H., & Phelan, D. (2007). What is in a teachers' mind? In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 234–244). Cambridge: CUP.

Daller, H., & Xue J. (2007). Lexical richness and the oral proficiency of Chinese EFL students. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.). *Testing and modelling lexical knowledge* (pp. 150–164). Cambridge: CUP.

DeKeyser, R.M. (1997). Beyond explicit rule learning. Automized second language morphosyntax. *Studies in Second Language Acquisition*, *19*, 195–221.

DeKeyser, R.M. (2001). Automaticity and automatization. In P. Robinson (Ed.). *Cognition and second language instruction* (pp. 125–151). Cambridge: CUP.

Ebbinghaus, H. (1885/1913). *Memory: A contribution to experimental psychology*. New York, NY: Teachers College, Columbia University.

Ellis, N., & Collins, L. (2009). Input and second language acquisition: The roles of frequency, form, and function. Introduction to the special issue. *Modern Language Journal*, *93*(3), 329–335.

Ellis, N.C., & Ferreira-Junior, F. (2009). Constructions and their acquisition: Islands and the distinctiveness of their occupancy. *Annual Review of Cognitive Linguistics*, *7*, 188–221.

Ellis, N.C., & Larsen-Freeman, D. (2009). Constructing a second language: Analyses and computational simulations of the emergence of linguistic constructions from usage. *Language Learning* 59(Suppl. 1): 90–125.

Fan, F. (2006). A corpus-based empirical study on inter-textual vocabulary growth. *Journal of Quantitative Linguistics*, *13*(1), 111–127.

Fitts, P.M., & Posner, M.I. (1967). *Human Performance*. London: Prentice/Hall International.

Freedman, A. (1987). Development in story writing. *Applied Psycholinguistics*, 8(2), 153–170.

Goulden, R., Nation, P., & Read, J. (1990). How large can a receptive vocabulary be? *Applied Linguistics*, *11*(4), 341–363.

Guiraud, P. (1954). *Les charactères statistiques du vocabulaire. Essai de méthodologie*. Paris: Presses Universitaires de France.

Hall, J.K., & Verplaetse, L.S. (Eds.). (2000) *Second and foreign language learning through classroom interaction*. Mahwah, NJ: Lawrence Erlbaum Associates.

Heathcote, A., Brown, S., & Mewhort, D.J.K. (2000). The power law repealed: The case for an exponential law of practice. *Psychonomic Bulletin and Review*, *7*(2), 185–207.

Haider, H., & Frensch, P.A. (2002). Why aggregated learning follows the power law of practice when individual learning does not: Comment on Rickard (1997, 1999), Delaney et al. (1998), and Palmeri (1999). *Journal of Experimental Psychology: Learning, Memory and Cognition, 28*(2), 392–406.

Housen, A., Bulté, B., Pierrard, M., & van Daele, S. (2008). Analysing lexical richness in French learner language. In J. Treffers-Daller et M4 group (Eds.), *Journal of French Language Studie*s, *18*(3), 277–298.

Hull, C.L., Hovland, C.I., Ross, R.T., Hall, J.,Perkins, D.T., & Fitch, R.B. (1940). *Mathematico-deductive theory of rote learning*. New Haven, CT: Yale University Press.

IELTS (International English Language Testing System): <http://www.ielts.org/default.aspx> (31 November 2010).

Jones, R.N., Rosenberg, A.L., Morris, J.N., Allaire, J.C. McCoy, K.J.M., Marsiske, M., et al. (2005). A growth curve model of learning acquisition among cognetively normal older adults. *Experimental Aging Research*, *31*, 291–312.

Lacroix, G., & Cousineau, D. (2006). The Introduction to the Special Issue on "$RT(N) = a + bN^{-c}$": The power law of learning 25 years later. *Tutorials in Quantitative Methods for Psychology*, 2(2), 38–42.

Larsen-Freeman, D. (2009). Adjusting expectations: The study of complexity, accuracy, and fluency in second language acquisition. *Applied Linguistics, 30*(4), 579–589.

Larson-Hall, J., & Herrington, R. (2009). Improving data analysis in second language acquisition by utilizing modern developments in applied statistics. *Applied Linguistics, 31*(3), 368–390.

Lezak, M. (1995). *Neuropsychological assessment* (3rd ed.). Oxford: OUP.

MacWhinney, B. (2000a). The CHILDES project: Tools for analyzing talk (3rd ed., Vol. 1: Transcription format and programs). Mahwah, NJ: Lawrence Erlbaum Associates.

MacWhinney, B. (2000b). The CHILDES project: Tools for analyzing talk (3rd ed., Vol. 2: The database). Mahwah, NJ: Lawrence Erlbaum Associates.

Malvern, D.D., & Richards, B.J. (2002). Investigating accommodation in language proficiency interviews using a new measure of lexical diversity. *Language Testing*, *19*(1), 85–104.

Malvern, D., Richards, B., Chipere, N., & Duran, P. (2004). *Lexical diversity and language development: quantification and assessment.* New York, NY: Palgrave Macmillan.

McCarthy, P.M., & Jarvis, S. (2007). Vocd: A theoretical and empirical evaluation. *Language Testing, 24*(4), 459–88.

McCarthy, P.M., & Jarvis, S. (2010). MTLD, vocd-D, and HD-D: A validation study of sophisticated approaches to lexical diversity assessment. *Behavior Research Methods*, *42*(2), 381–392.

Mandelbrot, B. (1966). Information theory and psycholinguistics: A theory of word frequencies. In P.F. Lazarsfeld & N.W. Hendry (Eds.). *Reading in mathematical social sciences* (pp. 350–368). Cambridge, MA: The MIT Press.

McDermott, P.A., Fantuzzo, J.W., Waterman, C., Angelo, L.E., Warley, H.P., Gadsden, V.L., & Zhang, X. (2009). Measuring preschool cognitive growth while it's still happening: The learning express. *Journal of School Psychology*, *47*, 337–366.

Meara, P. (1997). Towards a new approach to modelling vocabulary acquisition. In N. Schmitt & McCarth (Eds.), *Vocabulary: Description, acquisition and pedagogy* (pp. 109–121). Cambridge: CUP.

Miller, G.A. (1956). The magic number seven plus or minus two: Some limits on our capacity for processing information. *Psychological Review*, 63, 81–97.

Mollet, E., Wray, A., Fitzpatrick, T., Wray, N., & Wright M. (2010). Choosing the best tools for comparative analyses of texts. International Journal of Corpus Linguistics *15*(4): 429–473

Myung, L.J., Kim, C., & Pitt, M.A. (2000). Toward an explanation of the power law artifact: Insights from response surface analysis. *Memory and Cognition*, *28*(5), 832–840.

Nation, I.S.P. (2001). *Learning vocabulary in another language*. Cambridge: CUP.

Nation, I.S.P. <http://www.victoria.ac.nz/lals/resources/range.aspx>

Nerb, J., Ritter, F.E., & Krems, J. (1999). Knowledge level learning and the power law: A Soar model of skill acquisition. *Kognitionswissenschaft*, 8(1), 20–29.

Newell, A., & Rosenbloom, R.S. (1981). Mechanisms of skill acquisition and the law of practice. In J.R. Anderson (Ed.). *Cognitive skills and their acquisition* (pp. 1–55). Hillsdale, NJ: Lawrence Erlbaum Associates.

Newell, A. (1990). *Unified theories of cognition*. Cambridge, MA: Harvard University Press.

Ninio, A. (2007). *Language and the learning curve*. Oxford: OUP.

Pienemann, M., Keβler, J.-U., & Itani-Adams. (2011). Comparing levels of processability across languages. In M.H. Daller, D.D. Malvern, P. Meara, J. Milton, B. Richards, & J. Treffers-Daller (Eds.), Measurement of bilingual proficiency. *The International Journal of Bilingualism, 15*(2), 128–146.

Primi, R., Ferrão, M.E., & Almeida, L.S. (2010). Fluid intelligence as a predictor of learning: A longitudinal multilevel approach applied to math. *Learning and Individual Differences*, *20*, 446–451.

Rast, P., & Zimprich, D. (2009). Individual differences in a positional learning task across the adult lifespan. *Learning and Individual Differences*, *20*(1), 1–7

Read, J. (2000). *Assessing Vocabulary*. Cambridge: CUP.

Rey, A. (1964). *L'examen clinique en psychogie.* Paris: Presses Universitaires de France.

Richards, J. (1976). The role of vocabulary teaching. *TESOL Quarterly, 10*(1), 77–89.

Ritter, F.E., & Schooler, L.J. (2002). The learning curve. In *International encyclopedia of the social and behavioral sciences* (pp. 8602–8605). Amsterdam: Pergamon.

Rosenbloom, P., & Newell, A. (1987). Learning by chunking: A production system model of practice. In D. Klahr, P. Langley, & R. Neches (Eds.), *Production system models* (pp. 221–286). Cambridge, MA: The MIT Press.

Rosenbloom, P. S. (2006). A cognitive odyssey: From the power law of practice to a general learning mechanism and beyond. *Tutorials in Quantitative Methods for Psychology*, *2*(2), 43–51.

Schur, E. (2007). Insights into the structure of L1 and L2 vocabulary networks: Intimidations of small worlds. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 182–203). Cambridge: CUP.

Snoddy, G.S. (1926). Learning and stability. *Journal of Applied Psychology*, *10*, 1–36.

Speelman, C.P., & Kirsner, K. (2006). Transfer of training and its effect on learning curves. *Tutorials in Quantitative Methods for Psychology*, *2*(2), 52–65.

Stromqvist, S., Johansson, V., Kriz, S., Ragnarsdottir, H. Aisenman, R. & Ravid, D. 2002. Toward a crosslinguistic comparison of lexical quanta in speech and writing. *Written Language and Literacy, 5*, 45–67.

Tidball, F., & Treffers-Daller, J. (2007). Exploring measures of vocabulary richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 133–149). Cambridge: CUP.

Treffers-Daller, J., & Milton, J. (2011). Vocabulary size revisited: How large are students' receptive vocabularies? (paper presented at the annual conference of the British Association for Applied Linguistics, 1–3 September, University of the West of England, Bristol)

Thurstone, L.L. (1919). *The learning curve equation*. Psychological Monographs, *26*, 1–51.

Turlik, J. (2008). *A longitudinal study of vocabulary in L2 academic English writing of Arabic first-language students: Development and measurement*. Unpublished PhD dissertation, University of the West of England, Bristol.

Van Hout, R., & Vermeer, A. (2007). Comparing measures of lexical richness. In H. Daller, J. Milton, & J. Treffers-Daller (Eds.), *Modelling and assessing vocabulary knowledge* (pp. 93–115). Cambridge: CUP.

Van de gaer, E., Prustjes, H., Van Damme, J., & De Munter, A. (2009). School engagement and language achievement. *Merrill-Palmer Quarterly*, *55*(4), 373–405.

Verhoeven, L., & Van Leeuwe, J. (2009). Modeling the growth of word-decoding skills: Evidence from Dutch. *Scientific Studies of Reading 13*(3), 205–223.

Watts, D.J., & Strogatz, S.H. (1998). Collective dynamics of 'small-world networks'. *Nature, 393*, 440–442.

Wray, A. (2008). *Formulaic language: Pushing the boundaries*. Oxford: OUP.

Yu, G. (2010). Lexical diversity in writing and speaking task performances. *Applied Linguistics*, *31*, 236–259.

Zechmeister E.B., Chronis A., Cull, W.L., D'Anna, C.A., & Healy, N. (1995). Growth of a functionally important lexicon. *Journal of Reading Behavior, 27*(2), 201–217.

Zipf, G.K. (1935/1965). *Psycho-biology of language: An introduction to dynamic philology*. Cambridge, MA: The MIT Press.