载《当代外语研究》（国际版），2011 年第 12 期

海事英语的篇际词汇增幅率

李晶洁　　冯志伟

Inter-textual Vocabulary Growth Patterns for Marine Engineering English*

LI Jing-jie

Donghua University, Shanghai, China

E-mail: lijingjie@dhu.edu.cn

FENG Zhi-wei

Institute of Applied Linguistics, The Ministry of Education, Beijing, China

Dalian Maritime University, Dalian, China

E-mail: zwfengde2010@hotmail.com

ABSTRACT

This paper explores two fundamental issues concerning the inter-textual vocabulary growth patterns for Marine Engineering English, viz. vocabulary growth models and newly occurring vocabulary distributions in cumulative texts. Four mathematical models (Brunet's, Guiraud's, Tuldava's, and Herdan's models) are tested against the empirical vocabulary growth curve for Marine Engineering English. A new growth model based on the logarithmic function and the power

* Supported by "the Fundamental Research Funds for the Central Universities" (11D11402 and 11D11412).

law is presented. The theoretical mean vocabulary size and the 95% upper and lower bound values are calculated and plotted as functions of the sample size. For the purpose of this research the DMMEE (Dalian Maritime University Marine Engineering English) corpus was used.
Keywords: Inter-textual Vocabulary Growth Model; Newly Occurring Vocabulary Distributions; Logarithmic Function; Power Law

1. Introduction

In 1957, Firth first put forward the theory of collocation, which separated the study of lexis from the traditional grammar and semantics. With the theory of collocation it became evident that the choice of vocabulary did not depend only on the grammatical structure. Neo-Firthians built on the lexis-oriented research approach and further developed Firth's lexical theory. Halliday (1966: 148-162) presupposed the operation of lexis was at an independent linguistic level and was more than just filling in grammatical slots in a clause.

The vast amount of experimental studies and pedagogical publications has demonstrated, beyond all doubt, that the field of vocabulary studies is now anything but a neglected area (Schmitt 2002). The study of lexis has had a major influence on foreign language teaching, viz. as Smith (1999:50) once commented on the significance of vocabulary, "the lexicon is of central importance and is even described as being potentially the locus of all variation between languages, so that apart from differences in the lexicon, there is only one human language." Thus effective vocabulary acquisition is of particular importance for EFL learners (Ellis 2004) who frequently acquire impoverished lexicons despite years of formal study.

For effective vocabulary acquisition it is important that lexical items are acquired in certain quantities at certain rates. Thus, it is essential for both teachers and learners to be aware of the characteristics of vocabulary growth patterns such as: vocabulary growth models and word frequency distributions. These characteristics are of practical significance not only for the resolution of a series of problems in vocabulary acquisition but also for the theoretical explanation of some of the important issues in quantitative lexical studies.

At present, quantitative lexical studies show characteristics of a multilayer development supported by computer technology and corpus evidence. Specifically, there have been many statistical approaches to the study of lexis acquisition.

Meara (1997) proposed a lexical uptake model (Equation 1.1) for incidental teaching and learning. A learner's acquisition of vocabulary is expressed by a complex interaction between three variables: accumulated capacity of text input (N), number of new words provided in input texts ($V(N)_i$), and likelihood of the learner picking up any individual word from those texts (p).

$$V(N) = p \cdot \sum_{i=1}^{n} V(N)_i \tag{1.1}$$

$V(N)$: size of acquired vocabulary

$V(N)_i$: the new vocabulary the i^{th} input text provides

N: total number of input texts

p is an empirical parameter and its actual value is affected by teacher's and learner's experience, thus model 1.1 is factually determined by two factors, i.e. accumulated capacity of text

input (N) and number of new words provided in input texts ($V(N)_i$). These two factors constrain each other by some complex function relationship and define the vocabulary size of a learner. In other words, as long as we manage to simulate the function relationship of *N* and *V(N)* (also called 'vocabulary growth model' in this paper), we can calculate the acquired vocabulary size and the vocabulary growth rate of a learner. In fact, many attempts have been made to construct an appropriate mathematical model that would express the dependence of the vocabulary size *V(N)* on the sample size *N*. Some models describe the vocabulary growth pattern for General English with sufficient precision, but none of them have ever been tested against the vocabulary development for Marine Engineering English, henceforth MEE. In this paper, four existing models (Brunet's, Guiraud's, Tuldava's and Herdan's models) will be tested against the empirical growth curve for MEE. Furthermore, a new growth model will be proposed and evaluated on the basis of Dalian Maritime University Marine Engineering English Corpus, henceforth DMMEE.

Some basic concepts and notation need to be delimited before making further analysis of vocabulary growth. To be consistent with the examination of current models, the vocabulary size *V(N)* in this paper is defined as the set of all lemmas in a given text, excluding Arabic numerals, non-word strings, and personal and place names, since the inclusion of these would gravely distort the vocabulary growth curves. However, the size *N* of a text or sample refers to the total number of any character cluster or a single character, including words, names, Arabic numerals, and non-word strings in the text, but excluding punctuation marks.

2. Existing Vocabulary Growth Models

For much of the 20th century there has been a tradition of quantitative lexical studies with statistical information for various purposes. In the early 1930s, G. K. Zipf (1935) set out the rank-frequency law which held that the relationship between the frequency of the use of a word in a text and the rank order of that word in a frequency list is a constant. In the 1960s, Carroll (1967; 1969) put into words for the first time that word frequency distributions are characterized by large numbers of extremely low-probability words. This property sets quantitative lexical studies apart from most other studies in statistics and leads to particular statistical phenomena such as vocabulary growth patterns that systematically keep changing as the sample size is increased. In the 1980s and 1990s, special statistical techniques were developed to describe lexical frequency distributions accurately, and a number of vocabulary growth models were proposed. Of particular interest are Brunet's, Guiraud's, Tuldava's, and Herdan's models.

Expression 2.1 is known as Brunet's model (1978), with $log_w V(N)$ being the dependent variable and $log_w N$ the independent variable.

$$V(N) = (\log_N W)^{-\frac{1}{\alpha}} \tag{2.1}$$

Hence,

$$\log_w V(N) = \frac{1}{\alpha} \log_w (\log_w N)$$

W: Brunet's constant serving as the base of the logarithmic function

Parameter *W* has no probabilistic interpretation; it varies with the size of the sample. Parameter α is usually fixed at 0.17, a heuristic value that has been found to produce the desired result of producing a roughly constant relation between $log_w V(N)$ and $log_w N$ (Baayen 2001).

Expression 2.2 is known as Guiraud's model (1990), where the square root of the sample size replaces the sample size in what is known as the type-token ratio, *V(N)/N*, as follows:

$$R = \frac{V(N)}{\sqrt{N}}$$

Hence,

$$V(N) = R\sqrt{N} \qquad (2.2)$$

R: Guiraud's constant serving as the coefficient of the formula

The vocabulary size *V(N)* in Guiraud's model reduces to a simple square function of the sample size *N*, but the parameter *R* changes systematically with the sample size.

According to the hypothesis that the relation between *V(N)/N* and *N* is approximated by the power function:

$$\frac{V(N)}{N} = \alpha N^{\beta},$$

Tuldava (1990) constructed a vocabulary growth model (Expression 2.3) by applying logarithmization to both variables of *V(N)* and *N*.

$$V(N) = Ne^{-\alpha(\ln N)^{\beta}} \qquad (2.3)$$

e: natural base, e = 2.17828…

Parameters α and β have no probabilistic interpretation; they are the coefficients of variety that are supposed to be correlated with the probabilistic process of choosing "new" (unused) and "old" (already used in the text) words at each stage of text processing.

G. Herdan (1964) proposed a power model (Expression 2.4) based on the observation that the growth curve of the vocabulary appears as a roughly straight line in the double logarithmic plane.

$$\log V(N) = \log \alpha + \beta \log N = \log\left(\alpha N^{\beta}\right)$$

Hence,

$$V(N) = \alpha N^{\beta} \qquad (2.4)$$

α and β in Herdan's model are empirical parameters; they have no sensible interpretation.

The current vocabulary growth models are of particular importance for the theory of quantitative statistics, but they are seldom used to solve practical lexico-statistical problems, either because their complex expressions are computationally rather intractable, or the model parameters keep changing systematically with the sample size. Moreover, thus far, none of the current growth models have been tested against the vocabulary development for MEE in the light of corpus evidence.

3. DMMEE and Methodology

Dalian Maritime University Marine Engineering English Corpus (henceforth DMMEE) serves as the database of this study. It was constructed in the year of 2005 under the leadership of Professor Fan Fengxiang. DMMEE was designed to represent the contemporary MEE; it consists of a little more than one million word tokens. Altogether 959 text samples were collected in the corpus; the length of each text varies from 349 to 2070 word tokens. Descriptive data of DMMEE are presented in Table 3.1.

Table 3.1: Detailed information on the size of DMMEE corpus.

NAME	*TOKENS*	*TYPES*	*NUMBER OF TEXTS*	*RANGE OF TEXT SIZE*	*AVERAGE TEXT SIZE*
DMMEE	1,030,522	18,766	959	349 — 2070	1,075

To achieve representativeness and balance in text sample selection, Summers (1991) outlined a number of possible approaches including: an “elitist” approach based on literary or academic merit or “influentialness”; random selection; “currency”; subjective judgment of “typicalness”; availability of text in archives; demographic sampling of reading habits; and empirical adjustment of text selection to meet linguistic specifications. By using a combination of theses approaches, 959 representative texts were selected from today’s most influential MEE materials that are produced or published in either Britain or the USA between 1987 and 2004. The titles of the publications are listed in Appendix A.

DMMEE is organized according to general genre categories such as papers and books. The quantity of text in each category complies roughly with the proportion of each material genre displayed in Appendix A. In DMMEE, about 85% of the text samples are taken from papers and books, 10% from lectures and discussions, and the remainder from such sources as brochures and manual instructions. Table 3.2 presents the general genre categories and their corresponding proportions.

Table 3.2: Overall structure and general genre categories of DMMEE corpus.

GENERE CATEGORIES	*PROPORTIONS (%)*
Papers	65%
Textbooks & Books	20%
Lectures & Discussions	10%
Brochures, Manual instructions, etc.	5%

The text samples of DMMEE were selected from various constituent subject fields of MEE. Thus in the corpus one may find text samples on marine diesel engines, marine power plants, electrical installations, steering gears, marine refrigeration, gas exchange, lubricating systems, propelling systems, maintenance and repair, and etc.

With the aid of software SPSS and data-managing system FoxPro, and focussing on MEE vocabulary growth, this paper aims:

A. To verify the descriptive power of current growth models from fitting the empirical growth curves of EST.

B. To construct a new growth model based on the analysis of current mathematical models.

C. To make a comparison of the new growth model with other known models in terms of parameter estimation, goodness of each model fit, and analysis of residuals.

D. To calculate the theoretical mean vocabulary sizes and their 95% two-sided tolerance intervals using the new model.

E. To plot the distributions of newly occurring vocabulary in cumulative texts

Specifically, data processing was conducted in the following steps:

First, the text samples were randomly extracted from published texts in DMMEE and grouped into two sets, the Sample Set and the Test Set, each of which contained about 500,000 tokens. The Sample Set served as database for testing the four mathematical models and constructing a new vocabulary growth model. The Test Set was used to verify the descriptive power of the new model against MEE.

The number of text samples in each set is determined by two factors (Fan 2006: 115):

- The number of texts: the number of texts in the samples should be sufficiently large. Insufficient number of samples cannot capture true inter-textual lexical characteristics. In MEE, nearly 1000 texts (including Sample Set and Test Set) can basically capture true inter-textual lexical characteristics.
- The cumulative volume of texts: the total cumulative volume of texts in the samples should be large enough to produce sufficient vocabulary size of native speakers. Nation (1990) and Nagy (1997) place the vocabulary size of educated native speakers at 20,000. To be general, an EFL learner's MEE vocabulary cannot be lower than that estimate. A test run revealed that cumulative texts of 500,000 word tokens can roughly cover the MEE vocabulary size of native speakers. Information on the two sets is given in Table 3.3.

Table 3.3: Descriptive data for the Sample Set and the Test Set of DMMEE.

NAME	*TOKENS*	*TYPES*	*NUMBER OF TEXTS*	*AVERAGE TEXT SIZE*
Sample Set	513,658	15,583	479	1,072
Test Set	516,864	15,436	480	1,077

The second step was to explore the newly occurring vocabulary distributions. The newly occurring vocabulary size $V(N)_{new}$ was calculated and plotted as a function of the sample size N for both the Sample Set and the Test Set. Two respective FoxPro programs were designed to perform random selection of the text samples, and tokenization and lemmatization of each text sample. SPSS was used to plot the inter-textual vocabulary growth curves and the scatter-grams of $V(N)_{new}$ against N for the two sets separately.

4 Vocabulary Growth Models for MEE

To keep the integrity of the lexical structure of each individual text, the vocabulary growth curves plotted in this section are measured in units of text, not necessarily at equally spaced intervals. Figure 4.1 plots the dependence of the vocabulary size *V(N)* on the sample size *N* for the Sample Set and the Test Set of DMMEE.

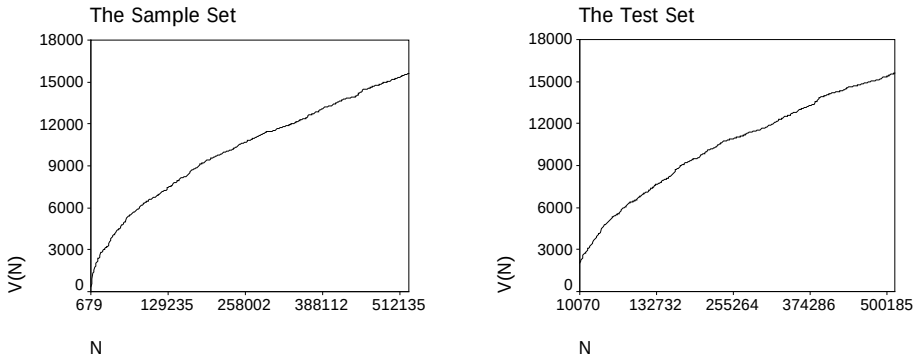
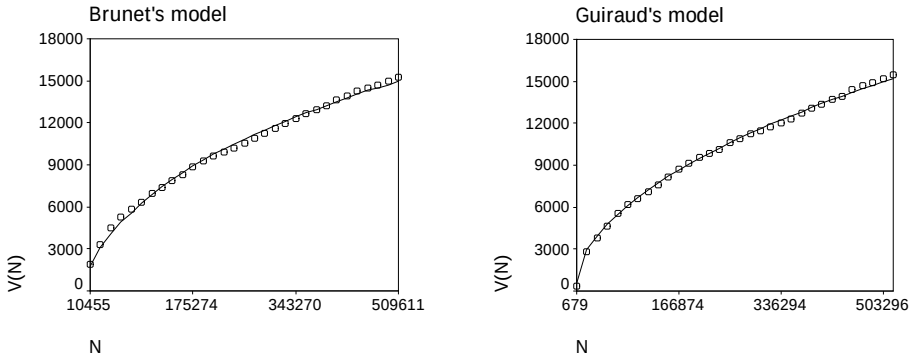


Figure 4.1: Vocabulary size *V(N)* as a function of sample size *N* for the cumulative texts in the Sample Set (left panel) and the Test Set (right panel) of DMMEE.

The vocabulary growth curves for the Sample Set and the Test Set exhibit similar patterns. For example, take the left panel into consideration; the growth curve of *V(N)* is not a linear function of *N*. Initially, *V(N)* increases quickly, but the growth rate decreases as *N* is increased. By the end of the Sample Set the curve has not flattened out to a horizontal line, that is, the vocabulary size keeps increasing even when the sample size has reached 500,000 word tokens.

4. 1 Test of the Existing Vocabulary Growth Models

Figure 4.2 plots the empirical vocabulary growth curve (tiny circles) for the Sample Set, as well as the corresponding expectations (solid line) using Brunet's, Guiraud's, Tuldava's and Herdan's models.



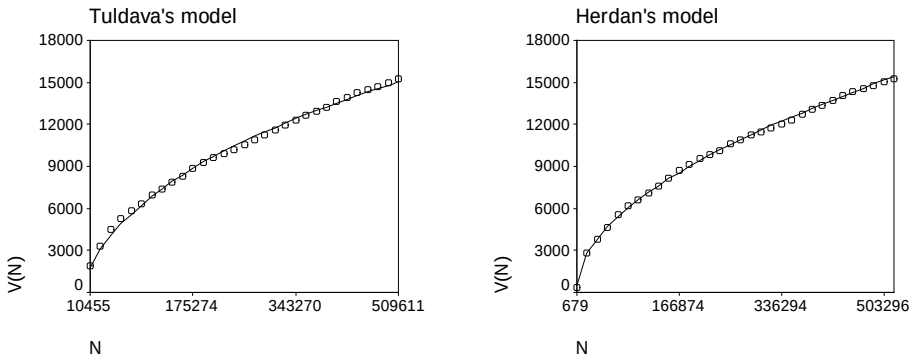


Figure 4.2: Observed (tiny circles) and expected (solid line) vocabulary growth curves for the Sample Set. The upper left panel shows the fit of Brunet's model to the empirical values, the upper right panel the fit of Guiraud's model, the lower left panel the fit of Tuldava's model, and the lower right panel the fit of Herdan's model.

Brunet's model (upper left panel) fits the empirical growth curve well. The value of *R* square (the coefficient of determination) reaches 99.876%. However, the upper left panel shows that Brunet's model underestimates the observed vocabulary sizes both at the beginning and towards the end of the growth curve. The predicted values are presented in the third column of Appendix B (*Pred_B*).

Guiraud's model (upper right panel) does not fit the empirical growth curve quite well. The R square is 99.812%, the lowest value among the four models. The upper right panel shows that Guiraud's model overestimates the observed vocabulary sizes from the beginning of the growth curve to nearly the middle. The predicted values are listed in the fifth column of Appendix B (*Pred_G*).

The fit of Tuldava's model (lower left panel) is close to the empirical growth curve, with the *R* square reaching 99.910%. But Tuldava's model tends to underestimate the observed vocabulary sizes at the beginning of the curve. The expected values are presented in the seventh column of Appendix B (*Pred_T*).

Though similar in expression to Guiraud's model, Herdan's model (lower right panel) describes the empirical growth curve best of the four models. The *R* square reaches the highest value 99.942%. But the panel shows that Herdan's model tends to overestimate the observed vocabulary sizes. The expected values are listed in the ninth column of Appendix B (*Pred_H*).

4. 2 A New Vocabulary Growth Model for MEE

Based on the lexico-statistical study of the Sample Set, a new mathematical model (Expression 4.1) is suggested to fit the vocabulary growth curve for MEE.

$$V(N) = \alpha \times \log N \times N^{\beta} \tag{4.1}$$

The new model is constructed by multiplying the logarithmic function and the power law. Parameter α is the coefficient of the whole expression; parameter β is the exponential of the power

function part of the model. Parameters α and β do not have fixed values; they have to change slightly with specific sample sizes to realize sufficient goodness-of-fit.

The theoretical considerations for proposing the new model are as follows:

- Brunet's model is in essence a complex logarithmic function, with $\log_W N$ being the independent variable and $\log_W V(N)$ the dependent variable. It tends to *underestimate* the observed vocabulary sizes both at the beginning and towards the end of the empirical growth curve.
- Herdan's model is a generalized power function. It tends to *overestimate* the observed values towards the end of the vocabulary growth curve.
- The mathematical combination of the logarithmic function and the power law may provide good fit to the empirical growth curve for MEE.

Figure 4.3 plots the dependence of *V(N)* on *N* for the Sample Set and the expectations using the new growth model.

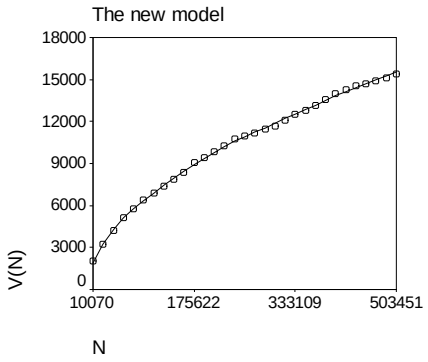


Figure 4.3: Vocabulary size *V(N)* as a function of the sample size *N*, measured in units of text. The tiny circles represent the empirical growth curve for the Sample Set, and the solid line the expected growth curve derived from the new mathematical model.

The new mathematical model provides very good fit to the empirical growth curve. The value of *R* square reaches 99.945%, higher than any other vocabulary growth models, particularly Brunet's model and Herdan's model. The respective values of parameters α and β are 3.529696 and 0.442790 for the Sample Set.

4. 3 Goodness-of-Fit Test for the New Vocabulary Growth Model

The Test Set of DMMEE is used to evaluate the descriptive power of the new vocabulary growth model. Figure 4.4 plots the empirical growth curve for the Test Set (tiny circles) and the corresponding expectations using the new mathematical model (solid line), with α=3.529696 and β=0.442790.

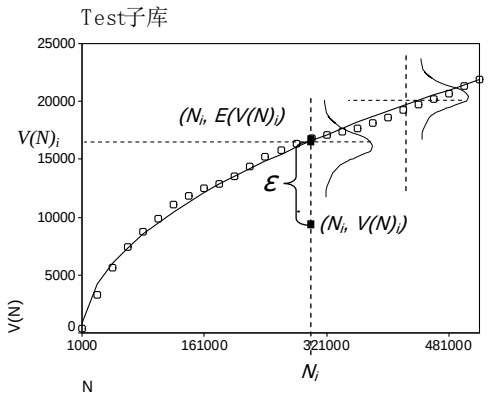


Figure 4.4: Dependence of the vocabulary size *V(N)* on the sample size *N*, measured in units of text. The tiny circles represent the empirical growth curve for the Test Set, and the solid line represents the expected growth curve derived from the new mathematical model.

The Sample Set has similar corpus size with the Test Set, and they both represent the contemporary MEE. Thus, the observed values of parameters α and β derived from the Sample Set also apply to the Test Set. Figure 4.4 shows that the new model provides good fit to the growth curve for the Test Set. The value of *R* square reaches 99.917%.

The new growth model is probabilistic in nature, so the appropriate generalization to this new model assumes that the expected value of *V(N)* is the function of *N*, but that for fixed sample size N_i, the variable $V(N)_i$ differs from its expected value in a random deviation ε (Expression 4.2).

$$V(N)_i = \alpha \times \log N_i \times N_i^{\beta} + \varepsilon \qquad (4.2)$$

The quantity ε in Expression 4.2 is a random variable, assumed to be normally distributed with $E(\varepsilon) = 0$ and $Variance(\varepsilon) = s^2$. The inclusion of ε allows the observed point $(N_i, V(N)_i)$ to fall either above the theoretical model curve (when $\varepsilon > 0$) or below the curve (when $\varepsilon < 0$). For any fixed N_i on the growth curve, $V(N)_i$ is the sum of a constant $\alpha \times \log N_i \times N_i^{\beta}$ and a normally distributed random deviation ε; so itself has a normal distribution, with the expected value $E(V(N)_i)$ from fitting the new model being the theoretical mean. These properties are illustrated in Figure 4.4. The upper and lower bound vocabularies of given point $(N_i, V(N)_i)$ are calculated for a confidence level of 95%, using the tolerance interval formula as follows:

$$V(N)_{upper} = E(V(N)_i) + (tolerance\ critical\ value) \times s$$
$$V(N)_{lower} = E(V(N)_i) - (tolerance\ critical\ value) \times s$$

The standard deviation *s* determines the extent to which each normal curve of vocabulary spreads out about its mean value $E(V(N)_i)$. (Figure 4.4) When *s* is small, the observed point $(N_i, V(N)_i)$ probably falls quite close to the theoretical growth curve; whereas observations may deviate considerably from their expected vocabulary sizes when *s* is large (Devore, 2000). In order to calculate the standard deviation value for each point *(N, V(N))* in units of text, the text samples of DMMEE are randomly selected and rearranged into ten sets of samples, based on which ten sets of vocabulary sizes and their corresponding sample sizes are obtained and displayed in Appendix C.

Devore proposed that when $n_{sample} = 10$ (number of samples), the tolerance critical value is 3.379 for capturing at least 95% of the observed values in the normal population distribution.

SPSS calculates the two-sided tolerance interval for each point $(N_i, V(N)_i)$ on the expected growth curve with a confidence level of 95%. Table 4.1 displays part of the statistics, including the standard deviation *s* and the corresponding upper and lower bound vocabulary sizes.

Table 4.1: Sample size *N*, theoretical vocabulary mean *E(V(N))*, standard deviation *s*, upper bound vocabulary size $V(N)_{max}$ and lower bound vocabulary size $V(N)_{min}$ for each point *(N, V(N))* on the expected growth curve derived from the new model.

N	*E(V(N))*	*s*	$V(N)_{max}$	$V(N)_{min}$
30000	3494	209.73	4203	2786
60000	5069	285.65	6034	4104
90000	6290	310.85	7340	5239
120000	7324	351.62	8512	6136
150000	8239	328.94	9351	7128
180000	9068	338.26	10211	7925
210000	9833	336.09	10968	8697
240000	10545	357.29	11753	9338
270000	11215	359.24	12429	10002
300000	11850	367.31	13091	10609
330000	12454	377.35	13729	11179
360000	13032	383.84	14329	11735
390000	13587	380.82	14874	12300
420000	14121	387.78	15431	12810
450000	14636	393.75	15967	13306
480000	15135	400.59	16489	13782

Figure 4.5 plots the empirical growth curve for the Test Set and the upper and lower bound values using the new mathematical model. The dashed lines show that the observed vocabulary sizes fall within the 95% two-sided tolerance bounds, without any exceptions.

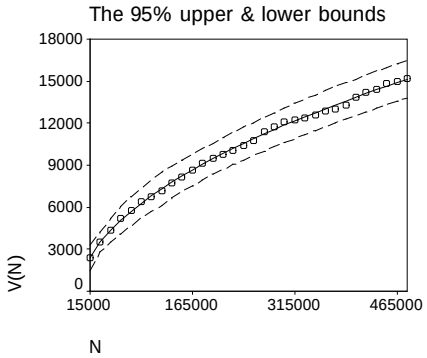


Figure 4.5: *V(N)* as functions of *N*, measured in units of text. The tiny circles represent the empirical growth curve for the Test Set, the solid line the theoretical expectations using the new model, and the dashed lines the upper and lower bound curves for a confidence level of 95%.

4.4 Analysis of Residuals for Vocabulary Growth Models

Vocabulary sizes that did not fit the empirical growth curve, hereafter residuals-from-fit, will be analysed for their variance and the degree to which there are coherent patterns in the residuals.

The variance of residuals shows how accurately the growth models predict the vocabulary values. A smaller variance implies a better fit of the model to the empirical growth curve and lower statistical uncertainty of the model components. SPSS calculates the mean square (a measure of variance) of residuals-from-fit for each vocabulary growth model. (Table 4.2)

Table 4.2: Mean square values (variance) of residuals-from-fit for each of the five vocabulary growth models in the Test Set.

Brunet's model	*Guiraud's model*	*Tuldava's model*	*Herdan's model*	*New model*
19484	22185	13371	11303	10512

The new mathematical model fits the empirical growth curve best, with the mean square value of residuals being the lowest. Herdan's model, Tuldava's model, Brunet's model and Guiraud's model follow in a sequence of decreasing goodness-of-fit.

In view of Scheaffer (1973), if the model makes correct predictions of the observed values, there should be no coherent patterns in the distribution of residuals. The greater the coherent structure of the residuals, the greater the chance that the observations deviate from their expected values. Figure 4.6 shows the residual plot against the sample size the five vocabulary growth models.

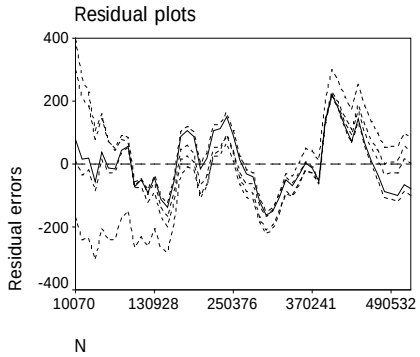


Figure 4.6: Residuals as functions of the sample size *N*, measured in units of text. The solid line represents the residuals from the new model, the dotted lines the residuals from the other four growth models, and the dashed line the expected value for each residual error.

The residuals from the new model exhibit no coherent patterns (solid line). The residuals are randomly distributed about their expected values 0 (dashed line) according to a normal distribution. All but a very few residuals lie between –160 and 160.

The residuals from the other four growth models (dotted lines) exhibit coherent patterns, such as a curved pattern for Brunet's model and an increasing function for Guiraud's model. The respective residuals fall within the interval of [–223, 397] for Brunet's model, [–305, 299] for

Guiraud's model, [–209, 300] for Tuldava's model, and [–150, 220] for Herdan's model.

The new model accounts for the empirical growth curve best of the five models for both the Sample Set and the Test Set. Therefore, it is reasonable to conclude that the new model gives the most accurate predictions of vocabulary growth for MEE. Tables 4.3 and 4.4 present the detailed information and basic statistics for the goodness of each model fit.

Table 4.3: Expressions and parameters of the vocabulary growth models for the Sample Set and the Test Set.

MODELS	*EXPRESSIONS*	*PARAMETERS*
Brunet's model	$\log_w V(N) = \frac{1}{\alpha}\log_w(\log_w N)$	α=18.21789; β= 0.15653
Tuldava's model	$V(N) = Ne^{-\alpha(\ln N)^{\beta}}$	α=0.02584; β=1.90470
Guiraud's model	$V(N) = \alpha\sqrt{N}$	α=21.64422
Herdan's model	$V(N) = \alpha N^{\beta}$	α=15.75641; β=0.52506
New model	$V(N) = \alpha \times \log N \times N^{\beta}$	α=3.529696; β=0.442790

Table 4.4: Goodness-of-fit statistics (*R* square and mean square values) for the five models from fitting the empirical growth curves for the Sample Set and the Test Set.

MODELS	*THE SAMPLE SET*		*THE TEST SET*	
	R SQUARE	*MEAN SQUARE*	*R SQUARE*	*MEAN SQUARE*
Brunet's model	99.876%	17,970	99.847%	19,484
Tuldava's model	99.910%	13,076	99.895%	13,371
Guiraud's model	99.812%	27,273	99.826%	22,185
Herdan's model	99.940%	8,395	99.911%	11,303
New model	99.945%	7,845	99.917%	10,512

5. Newly Occurring Vocabulary Distributions in Cumulative Texts

Feng (1988, 1996) proposed the "law of decreasing new vocabulary growth" in his study of terminology[1]. He noticed that with the increase of terminology entries, the number of high frequency words grows correspondently, whereas the probability of the occurrence of new words decreases. At this point, although the number of terminology entries keeps growing, the growth rate of the total number of new words slows down. The repeated occurrences of high frequency words indicate a tendency of decreasing new vocabulary growth. "The law of decreasing new vocabulary growth" lends itself not only to terminology system but also the process of reading written texts. When starting to read a text written in a less familiar language, one may encounter a body of new words. As more texts are read, the growth rate of new words is gradually reduced. If the reader can master the new words he comes across, reading will be much easier. There exists a

function relationship between the number of new words (*W*) and text capacity (*T*) as follows.

$$W = \Phi(T)$$

With the increase of text capacity (*T*), the growth rate of the number of new words (*W*) begins to reduce and the function curve becomes smoother in a form of convex parabola in the rectangular coordinate system. This function curve indicates the process of vocabulary growth in the reading of written texts. It is a mathematical account of the law of vocabulary change in a reading process.

The corpus evidence proves Feng's theory and suggests that as the sample size *N* increases, the new vocabulary $V(N)_{new}$ that an additional input text produces decreases. At any given N_i the new vocabulary $V(N)_{new}$ of a sample with *N* word tokens can be estimated with Expression 5.1.

$$V(N)_{new} = V(N_i + N) - V(N_i) \tag{5.1}$$

Figure 5.1 plots the newly occurring vocabulary size $V(N)_{new}$ as a function of the sample size *N* for the Sample Set and the Test Set.

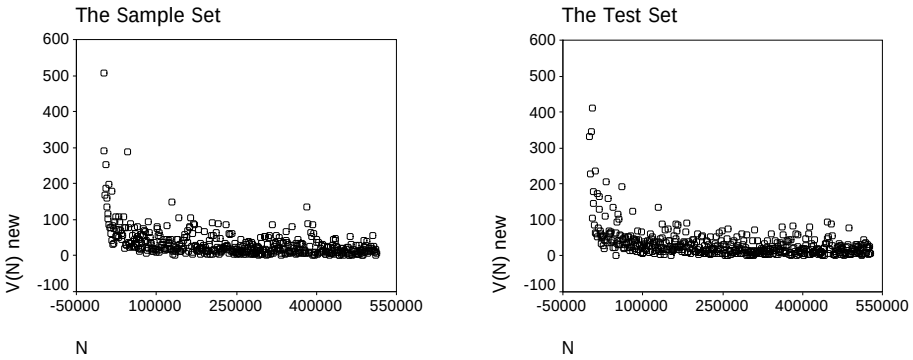


Figure 5.1: Scatter-grams of $V(N)_{new}$ against *N* for the Sample Set (left panel) and the Test Set (right panel).

The left panel for the Sample Set and the right panel for the Test Set reveal surprisingly similar distributions of newly occurring vocabulary sizes. The new vocabulary $V(N)_{new}$ is a rapidly decreasing function of the sample size *N* with a long tail of the low values of $V(N)_{new}$. Initially, $V(N)_{new}$ decreases sharply to some steady-state value that decreases slowly and smoothly. Then as *N* increases, the empirical values of $V(N)_{new}$ are distributed about the stable value in a wide dispersion till the end of the scatter plot. Take the left panel into consideration; the first input text produces more than 500 new word types. Then $V(N)_{new}$ decreases sharply as *N* increases, viz. when the sample size reaches 62,305 word tokens, the value of $V(N)_{new}$ is reduced to less than 150. But from *N* = 62,305 onwards till the end of the scatter plot, the newly occurring vocabulary size decreases very slowly and the observed values of $V(N)_{new}$ are distributed in a fluctuant way. The last input text produces 6 new word types, which proves that new vocabulary still occurs even when the sample size has reached 500,000 word tokens.

6. Examples of Pedagogical Applications

This research has significance in explicit EFL teaching and learning. With the assistance of the new growth model, teachers and students can make reliable estimates not only on the vocabulary size and its intervals for a given textbook but also on the volume of texts that are needed to produce a particular vocabulary size. For example, suppose that a textbook of MEE consists of 50 input texts, with the size of each text varying between 500 and 2,000, and totalling 50,000 word tokens. The vocabulary size of the whole textbook is estimated by using the new growth model, with α= 3.5297 and β= 0.4428:

$$\begin{aligned} E\left[V(N)\right] &= \alpha \times \log N \times N^{\beta} \\ &= 3.5297 \times \ln 50000 \times 50000^{0.4428} \\ &= 4599 \end{aligned}$$

The 95% tolerance interval is calculated to estimate the possible vocabulary sizes, with *critical value* = 3.379 and *s*= 118.1479:

$$\begin{aligned} E\left[V(N)\right] - (critical\ value) \times s \le V(N) \le E\left[V(N)\right] + (critical\ value) \times s \\ 4599 - 3.379 \times 118.1479 \le V(N) \le 4599 + 3.379 \times 118.1479 \\ 4199 \le V(N) \le 4998 \end{aligned}$$

Thus we are "95% certain" that the vocabulary size of the whole textbook lies between 4199 and 4998 word types. If 5% of the total vocabulary input fails to be acquired, then the two-sided tolerance interval for the acquired vocabulary is [4199—4199×5%, 4998—4998×5%], that is, [3989, 4748]. If a new text with 1000 tokens is added into the textbook, the newly occurring types that this input text produces are estimated according to Expression 5.1:

$$\begin{aligned} E\left[V(N)_{new}\right] &= E\left[V(N)_{51000}\right] - E\left[V(N)_{50000}\right] \\ &= \alpha \times \log N_{51000} \times {N_{51000}}^{\beta} - \alpha \times \log N_{50000} \times {N_{50000}}^{\beta} \\ &= 3.5297 \times \ln 51000 \times 51000^{0.4428} - 3.5297 \times \ln 50000 \times 50000^{0.4428} \\ &= 48 \end{aligned}$$

The 95% tolerance interval is calculated for the newly occurring vocabulary, with *critical value* = 3.379 and *s*= 10.1751:

$$\begin{aligned} E\left[V(N)_{new}\right] - (critical\ value) \times s \le V(N)_{new} \le E\left[V(N)_{new}\right] + (critical\ value) \times s \\ 48 - 3.379 \times 10.1751 \le V(N)_{new} \le 48 + 3.379 \times 10.1751 \\ 14 \le V(N)_{new} \le 83 \end{aligned}$$

Another example is to estimate the sample size for a given vocabulary size. For example, The Engineering English Competence Examination has a fairly high requirement on vocabulary proficiency with the minimal limit of 4000 word types. To ensure sufficient amount of vocabulary, the textbook of MEE has to reach a certain size. According to the new growth model, the expected size of the textbook is calculated as follows:

$$V(N) = \alpha \times \log E[N] \times E[N]^{\beta}$$
$$4000 = 3.5297 \times \ln E[N] \times E[N]^{0.4428}$$
$$E[N] = 38550$$

With *critical value* = 3.379 and *s*= 2231.342, the 95% tolerance interval is:

$$E[N] - (critical\ value) \times s \le N \le E[N] + (critical\ value) \times s$$
$$38550 - 3.379 \times 2231.342 \le N \le 38550 + 3.379 \times 2231.342$$
$$31010 \le N \le 46090$$

Thus we are at least "95% confident" that the size of the textbook of MEE has to be between 31010 and 46090 tokens, in order to produce 4,000 word types.

7. Conclusion

This paper explores two fundamental issues concerning the inter-textual vocabulary growth patterns for MEE: the vocabulary growth models and newly occurring vocabulary distributions in cumulative texts.

The paper first explores the MEE vocabulary growth. The growth curve of vocabulary *V(N)* is not a linear function of the sample size *N*. Initially, *V(N)* increases quickly, but the growth rate decreases as *N* is increased. Four mathematical models (Brunet's, Guiraud's, Tuldava's and Herdan's models) are tested against the empirical growth curve for MEE. A new growth model is constructed by multiplying the logarithmic function and the power law. Parameter α is the coefficient of the whole expression; parameter β is the exponential of the power function part of the model.

$$V(N) = \alpha \times \log N \times N^{\beta}$$

The new model describes the empirical growth curve best of the five models for MEE, with the *R* square reaching the highest value and the mean square being the lowest. The upper and lower tolerance bounds are calculated to capture at least 95% of the possible vocabulary sizes in the normal population distribution. The residuals from the five growth models are compared and analyzed; the respective values fall within the interval of [–223, 397] for Brunet's model, [–305, 299] for Guiraud's model, [–209, 300] for Tuldava's model, [–150, 220] for Herdan's model, and [–160,160] for the new model.

The second issue this paper explores is the distribution of newly occurring vocabulary sizes in cumulative texts. The new vocabulary $V(N)_{new}$ is a rapidly decreasing function of the sample size *N* with a long tail of low values of $V(N)_{new}$. Initially, $V(N)_{new}$ decreases sharply to a certain point. Then as *N* increases, $V(N)_{new}$ decreases very slowly and its observed values are distributed in a fairly wide dispersion.

The present study has been exploratory in nature, and some difficult issues have not yet been tackled adequately. For example, the new vocabulary growth model was constructed on the basis

of studies of the Sample Set, which consists of 480 individual texts, totalling about 500,000 word tokens. The model has been verified to provide good fit to the empirical vocabulary sizes within the boundary of the sample size of not more than 500,000 word tokens. However, the possibilities of extrapolation of this new growth model in the direction of larger than 500,000 tokens need further consideration and verification.

References

Baayen, R. H. (2001). *Word Frequency Distributions.* Dordrecht: Kluwer Academic Publishers. ISBN 0-7923-7027-1

Brunet, E. (1978). *Le Vocabulaire de Jean Giraudoux*. TLQ, Volume 1, Slatkine, Geneve.

Carroll, J. B. (1967). *Computational Analysis of Present-Day American English.* Providence: Brown University Press.

Carroll, J. B. (1969). *A Rationale for an Asymptotic Lognormal Form of Word Frequency distributions*. Princeton: Research Bulletin, Educational Testing Service.

Devore, J. (2000). *Probability and Statistics*. Pacific Grove: Brooks/Cole.

Ellis, Rod. (2004). *Second Language Acquisition (5th ed.).* Shanghai: Foreign Language Teaching Press.

Fan, Fengxiang. (2006). A Corpus-Based Empirical Study on Inter-textual Vocabulary Growth. *Journal of Quantitative Linguistics*. Volume 13, Number 1: 111-127. DOI : 10.1080/09296170500500603.

Feng, Zhiwei. (1988). FEL Formula -- Economical Law in the Formation of Terms. *Social Sciences in China (English version)*, No 4. , Beijing.

Feng，Zhiwei. (1996). Introduction of Modern Terminology, The Language Publishing House, Beijing.

Feng, Zhiwei. (2006). Evolution and present situation of corpus research in China. *International Journal of Corpus Linguistics* 11:2, 173-207, Amsterdam.

Guiraud, H. (1990). *Les Caracteres Statistiques du Vocabulaire*. Paris: Presses Universitaires de France.

Herdan, G. (1964). *Quantititative Linguistics.* Buttersworths, London.

Kennedy, G. (2000). *An Introduction to Corpus Linguistics.* Beijing: Foreign Language Teaching and Research Press.

Schmitt, N. & Meara, P. (1997). *Researching vocabulary through a word knowledge framework: Word associations and verbal suffixes.* Studies in Second Language Acquisition, 19, 17-36.

Schmitt, N. & McCarthy, M. (2002). *Vocabulary: Description, Acquisition and Pedagogy*. Shanghai: Shanghai Foreign Language Education Press.

Sinclaire, J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Smith, Neil. (1999). *Chomsky: Ideas and Ideals*. Cambridge: Cambridge University Press

Summers, D. (1991). *Longman/ Lancaster English Language Corpus: Criteria and Design.* Harlow: Longman.

Tuldava, J. (1996). *The Frequency Spectrum of Text and Vocabulary.* Journal of Quantitiative Linguistics, 3, 38-50.

Zipf, G. K. (1935). *The Psycho-Biology of Language.* Boston: Houghton Mifflin.

Zipf, G. K. (1949). *An Introduction to Human Ecology.* New York.

Appendix A

The titles of the publications from which text samples were selected for DMMEE

PUBLICATIONS
Boat-owner's Guide Radar
British Crown Copyright
British Maritime Technology
Diesel & Gasturbine Worldwide
Diesel Progress Engines & Drives
Electricity Applied to Marine Engineering
Freeze Fitting Rudder Bearings Brian Tarrant
Fuel-Injection Pump Construction
Gas Turbine
Heat Transfer Engineering
How Does A Marine Diesel Engine Work?
IMO News
Journal of Marine Engineering and Technology
Journal of Marine Design and Operations
Journal of Ship Research March
Landmobile and Marine Radio Technical Handbook
Marine Auxiliary Machinery
Marine Diesel Engines
Marine Engine and Propulsion System
Marine Engine Net Page January
Marine Engineer International
Marine Engineering Page
Marine Engineering Practice
Marine Engineering Review
Marine Engineers Review
Marine Engineers Review

Marine Engines and Propulsion System
Marine Engines Fuels and Lubricants
Marine Propulsion International
Marine Technology October
Materials For Marine Machinery
Modern Marine Electricity and Electronics
Modern Marine Engineer's Manual
Naval Engineers Journal
North East Coast Institute of Engineer & Shipbuilders
Offshore Marine Technology
Paper Read on 04hv University of Warwick
Propulsion
Ship Repair
Shipping World and Shipbuilder
Ships Monthly
Technology Turbochargers
The Motor Ship
The 53rd Andrew Laing Lecture
The Institute of Marine Engineers
The Institution of Engineers and Shipbuilders in Scotland
The Maintenance and Repairing of Marine Diesel Engines
The Marine Engineers
The Marine Engineers Review
The Me Engine
The Motor Ship
The Naval Architect August
The Running and Maintenance of The Marine Diesel Engine
Warship Technology
Water Jet

Appendix B

Part of Statistics for the Sample Set Using Brunet's Model, Guiraud's Model, Tuldava's Model and Herdan's Model

Tokens	*Types*	*Pred_B*	*Resi_B*	*Pred_G*	*Resi_G*	*Pred_T*	*Resi_T*	*Pred_H*	*Resi_H*
25202	3169	2947	222	3436	-267	3001	168	3225	-56
50471	4658	4500	158	4863	-205	4513	145	4644	14

75420	5719	5679	40	5944	-225	5670	49	5734	-15
100929	6611	6688	-77	6876	-265	6667	-56	6682	-71
125251	7408	7531	-123	7660	-252	7502	-94	7484	-76
150861	8151	8326	-175	8407	-256	8295	-144	8252	-101
175622	9065	9028	37	9071	-6	8996	69	8937	128
200998	9594	9692	-98	9704	-110	9662	-68	9593	1
225305	10258	10286	-28	10274	-16	10258	0	10186	72
250376	10871	10862	9	10830	41	10838	33	10766	105
275847	11303	11414	-111	11368	-65	11395	-92	11328	-25
300330	11695	11918	-223	11862	-167	11904	-209	11845	-150
325548	12301	12413	-112	12349	-48	12405	-104	12357	-56
350479	12817	12882	-65	12814	3	12879	-62	12846	-29
375305	13315	13329	-14	13260	55	13333	-18	13316	-1
400817	14002	13772	230	13703	299	13782	220	13883	119
425830	14306	14190	116	14124	182	14206	100	14278	28
450526	14712	14589	123	14528	184	14612	100	14768	-56
475815	15015	14984	31	14930	85	15014	1	15082	-67
500185	15363	15354	9	15308	55	15390	-27	15483	-120

Appendix C

Statistical Data for the Ten Sets of Text Samples

N	*V(N)*									
	Set_1	*Set_2*	*Set_3*	*Set_4*	*Set_5*	*Set_6*	*Set_7*	*Set_8*	*Set_9*	*Set_10*
15000	2073	2699	2262	2189	2338	2288	2842	2182	1590	1837
30000	3200	3422	3170	3375	3056	3606	3597	3030	2583	3427
45000	3764	4323	4148	4040	4274	4374	4393	3896	4007	4308
60000	5003	4885	4676	4720	4861	5058	4961	4948	5266	5362
75000	5410	5541	5097	5628	5343	6115	5472	5474	5855	6033
90000	5792	6076	5800	6208	5640	7025	6004	5939	6322	6526
105000	6137	6421	6317	6580	5997	7552	6393	6306	6749	6841
120000	6715	6788	6880	7021	6686	7812	6790	6636	7046	7226
135000	7091	7326	7316	7798	7467	8168	7457	7025	7596	7602
150000	7481	7757	7654	8140	8038	8387	7688	7422	8271	7874
165000	8095	8227	7930	8430	8829	8700	7960	8103	8829	8293
180000	8400	8840	8410	8775	9189	9188	8390	8628	9247	8927
195000	8864	9394	8951	9265	9564	9732	8893	8970	9553	9376

210000	9442	9718	9521	9632	9914	10256	9134	9234	9849	9840
225000	9832	10122	10163	9933	10296	10670	9608	9470	10060	10108
240000	10270	10427	10638	10168	10561	10913	10051	9885	10373	10325
255000	10538	11031	10885	10371	11012	11308	10487	10187	10815	10611
270000	10703	11537	11242	10661	11643	11592	11084	10437	11075	10888
285000	10936	11748	11528	10942	11881	11817	11489	10631	11327	11233
300000	11129	12174	11775	11146	12049	12006	11652	10970	11581	11486
315000	11546	12500	11933	11573	12302	12362	11995	11226	11928	11955
330000	11860	12680	12262	12029	12566	12530	12206	11521	12148	12187
345000	12075	12931	12769	12242	12827	12910	12837	11832	12565	12661
360000	12521	13147	13201	12491	13323	13200	13196	12069	12959	13078
375000	12744	13374	13348	12830	13524	13458	13706	12303	13125	13300
390000	13067	13608	13618	13273	13596	13646	13935	12613	13320	13511
405000	13471	14081	13880	13795	13969	13851	14135	12958	13632	13845
420000	13716	14480	14004	14168	14262	14155	14312	13363	13802	14010
435000	13893	14690	14153	14321	14438	14360	14544	13593	14038	14355
450000	14196	15003	14377	14450	14653	14662	14727	13994	14263	14733
465000	14400	15252	14605	14812	14874	14856	15032	14327	14679	15075
480000	14638	15373	14841	15239	15226	15041	15376	14553	15021	15282
495000	15074	15472	15159	15631	15481	15248	15621	15044	15258	15581
510000	15368	15678	15406	15762	15699	15545	15810	15243	15635	15741

[1] Terminology denotes a discipline which systematically studies the labelling or designating of concepts particular to one or more subject fields or domains of human activity, through research and analysis of terms in context, for the purpose of documenting and promoting consistent usage.