载《术语标准化与信息技术》2009 年第 2 期。

篇际英语词汇增幅率研究[1]

上海交通大学 李晶洁[2]
教育部语言文字应用研究所 冯志伟[3]

摘要

本文基于 Brown 语料库的 100 万词的数据，考察科技英语的篇际词汇增长模型，以篇章为计量单位，描述科技英语文本中词汇量 *V(N)*与累积文本容量 *N* 之间的函数关系。研究表明，国外现有的词汇增长模型（Brunet 模型，Guiraud 模型，Tuldava 模型和 Herdan 模型）都不能精确地描述科技英语的词汇增长曲线。本文通过对幂函数和对数函数的比较分析，构建了新的词汇增长模型，并应用此模型推导出科技英语的理论词汇增长曲线及其 95%双向置信区间。

本研究属于计算语言学的基础理论研究，反映了语言的经济律，有深刻的理论价值；另外，本文对英语作为外语的教学（English as Foreign Language，简称 EFL），特别是词汇教学和教材编写，也有一定的指导意义。使用这个新的词汇增长模型，可以用来预测任意累积语篇对应的词汇总量（已知 *N* 求 *E(V(N))*），以及生成任意词汇量需要的累积语篇数（已知 *V(N)*求 *E(N)*）。

关键词：篇际词汇增长模型；对数函数；幂函数；拟合度检验

Inter-textual English Vocabulary Growth

Abstract: This paper explores the English inter-textual vocabulary growth model. Four mathematical models (Brunet's model, Guiraud's model, Tuldava's model, and Herdan's model) were tested against the empirical growth curve. A new vocabulary

[1]本文是国家社会科学基金资助项目的研究成果之一，项目编号：03BYY019。
[2] E-mail: cindy_ljj@sjtu.edu.cn Tel: 15902188933
[3] E-mail: zwfengde@gmail.com Tel: 15911066112

growth model was constructed by multiplying the logarithmic function and the power law. The goodness-of-fit Test was made to evaluate the new model. The theoretical mean vocabulary size and the 95% upper and lower bound values were calculated and plotted as functions of the sample size. The research was carried out entirely on the basis of Brown Corpus including one million words.

This research has application in explicit EFL teaching and learning. The new growth model can make reliable estimates not only on the vocabulary size and its 95% confidence intervals for a given textbook, but also on the volume of individual texts that are needed to produce a particular vocabulary size.

Keywords: Inter-textual vocabulary growth model; Logarithmic function; Power law; Goodness-of-fit test

1. 引言

弗斯语言学（Firthian linguistics）认为，词汇是语言描述的中心。1957 年，Firth 首先提出了搭配和类连接理论，在某种程度上将词汇内容从语法和语义学中分离出来。之后数十年，新弗斯学者坚持以词汇研究为中心，强调词汇与语法的辩证关系，深入发展了 Firth 的词汇理论。Halliday（1966: 148-162）提出词汇不是用来填充语法确定的一套空位（slots），而是一个独立的语言学层面；词汇研究可以作为对语法理论的补充，却不是语法理论的一部分。近些年来，语料库证据支持的词汇学研究蓬勃发展。越来越多的实证研究表明，词汇和语法在实现意义时是交织一起的，必须整合描述。词汇是话语实现的主要载体，语法则起到管理意义、组合成份和构筑词项的作用，而不是更多。Smith（1999:50）甚至认为“词汇是语言间所有差异的潜在所在。排除词汇差异这一因素，人类的语言只有一种。”

目前，词汇研究在计算机技术和语料库数据支持下呈现出多元化的发展；对于词汇的描述也不再局限于定性研究而是更多地引入统计分析。

我国学者冯志伟（1988）在研究术语时，曾经提出“术语形成经济律”，揭示了术语系统中词组型术语总是比单词型术语占优势的内在原因和数学机理，用 FEL 公式来描述“术语形成经济律”，并在此基础上，进一步提出了更加具有概括性的“生词增幅递减律”（1996），

他指出：在一个术语系统中，每个单词的绝对频度是不同的，经常使用的单词是高频词，不经常使用的单词是低频词，随着术语条目的增加，高频词的数目也相应地增加，而生词出现的可能性越来越小，这时，尽管术语的条数还继续增加，生词总数增加的速率却越来越慢，而高频词则反复地出现，生词的增幅有递减的趋势。“生词增幅递减律”不仅适用于术语系统，也适用于阅读书面文本的过程，人们在阅读一种用自己不熟悉的语言写的文本时，开始总有大量不认识的生词，随着阅读数量的增加，生词增加的幅度会逐渐减少，如果阅读者能够掌握好阅读过的生词，阅读将会变得越来越容易。在生词数W与文本容量T之间存在着如下的函数关系：

$$W = \Phi(T) \qquad (1.1)$$

随着文本容量T的增大，生词数目W的增幅逐渐减少，反映这种函数关系的曲线也就越来越平滑，整个曲线在直角坐标系内呈现上凸的抛物线形状。这条函数曲线同时反映了阅读书面语时词汇增加的过程，它实际上就是人们阅读过程中词汇变化规律的数学描述。

Meara（1997）在她构建的词汇习得模型中提出（式 1.2），学生的习得词汇量由三个因素决定——累积输入的文本容量 *N*，输入文本提供的新增词汇量 *V(N)*，以及学生习得新增词汇的概率 *p*。

$$V(N) = p \cdot \sum_{i=1}^{n} V(N)_i \qquad (1.2)$$

$V(N)$: 学生的习得词汇量

$V(N)_i$: 累积输入文本提供的新增词汇量

p: 学生习得新词汇的概率

因为 *p* 是一个经验参数，其具体取值由老师或学生根据经验确定，所以模型 1.2 实际上是由累积文本容量 *N* 和新增词汇量 *V(N)*两个因素决定的；而这两个因素以某种复杂的函数

关系相互制约，构成了学生现有的词汇量。换言之，只要模拟出 N 和 $V(N)$的函数关系（本文称之为词汇增长模型），我们就可以计算出学生习得的词汇量和词汇习得的增长率。同时，构建词汇增长模型对二语习得研究、教材编写、大纲设计以及外语教学实践等也都有一定的理论意义和应用价值。

2．对现有词汇增长模型的描述与评估

重视对词汇的定量分析是语言统计研究的传统。早在 20 世纪 30 年代，G. K. Zipf（1935）就提出了著名的 Zipf 定律，即一个词在语篇中的等级序号（该词在按出现次数排列的词表中的位置）与该词的出现次数的乘积是一个常数。20 世纪 60 年代，J.B. Carroll（1967; 1969）首次论证了词频分布不同于其他统计现象的特质性，即词谱中大量存在的超低频词，并推导出新增词汇量与文本容量之间的理论函数模型。20 世纪 70 年代后期，大型机读语料库相继建成，为计量词汇研究提供了强大的数据支持，多种经验和理论词汇增长模型被构建，其中 Brunet 模型，Guiraud 模型，Tuldava 模型和 Herdan 模型影响至今，它们分别从不同的角度描述了文本容量 N 与新增词汇量 $V(N)$之间的函数关系。

2.1　四种现有词汇增长模型的函数表达式

（I）　式 2.1 为 Brunet 模型表达式（1978）：

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

推导得出,

$$\log_w V(N) = \frac{1}{\alpha}\log_w(\log_w N) \tag{2.2}$$

W: Brunet 常量，作为对数函数的底

式 2.2 是一个复杂的对数函数关系式，以 $\log_w V(N)$为自变量，$\log_w N$ 为因变量，W 为对数函数的底数，α 为表达式的参数。W 虽然被称作 Brunet 常量，却并不是一个常数，其取值随

文本容量 N 的变化而变化；系数 α 通常默认取值 0.17，这是一个经验值，不存在理论解释，目的是确保 $log_w V(N)$ 和 $log_w N$ 的常量函数关系（Baayen 2001）。

（II） 式 2.3 为 Guiraud 模型表达式（1954）：

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

所以,

$$V(N) = R\sqrt{N} \tag{2.3}$$

R: Guiraud 常量，作为表达式的系数

Guiraud 模型由类符-形符比（type-token ratio）, $V(N)/N$, 演变生成。系数 R 虽然被称作 Guiraud 常量，却并不是一个常数，其取值会随文本容量的增加呈现出系统性的变化（Baayen 2001）。

（III） Tuldava（1990）在研究中发现类符-形符比 $V(N)/N$ 与文本容量 N 之间呈近似幂函数关系，即：

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

将变量 N 和 $V(N)$ 分别取对数，得到：

$$V(N) = Ne^{-\alpha(\ln N)^{\beta}} \tag{2.4}$$

e: 自然底数, 2.71828

式 2.4 为 Tuldava 模型表达式。参数 α 和 β 为经验系数，不存在概率解释，其取值与所选文本的容量，语体，类型等因素有关。

（IV） Herdan （1964）研究发现词汇增长曲线在双对数平面上呈近似线状，因此推论 $logV(N)$ 和 $logN$ 之间存在线形关系，即：

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

所以,

$$V(N) = \alpha N^{\beta} \tag{2.5}$$

式 2.5 为 Herdan 模型表达式。参数 α 和 β 同样为经验系数，不存在概率解释。

2.2 对现有词汇增长模型的拟合度检验

本节选取 Brown 语料库的 Sample 子库为文本语料，逐一检验上述四种模型对于篇际英语词汇增长的描述能力。 Sample 子库的文本容量约为 500,000 形符，词汇总量为 22,464 类符，虽然不是大型语料库，却能够基本反映出英语词汇的增长规律。图 2.1 为上述四种数学模型模拟 Sample 子库的词汇增长曲线。X 轴为累积文本容量，Y 轴为累积词汇量。圆圈代表词汇增长的实际观察值，实线代表模型预测的理论期望值。

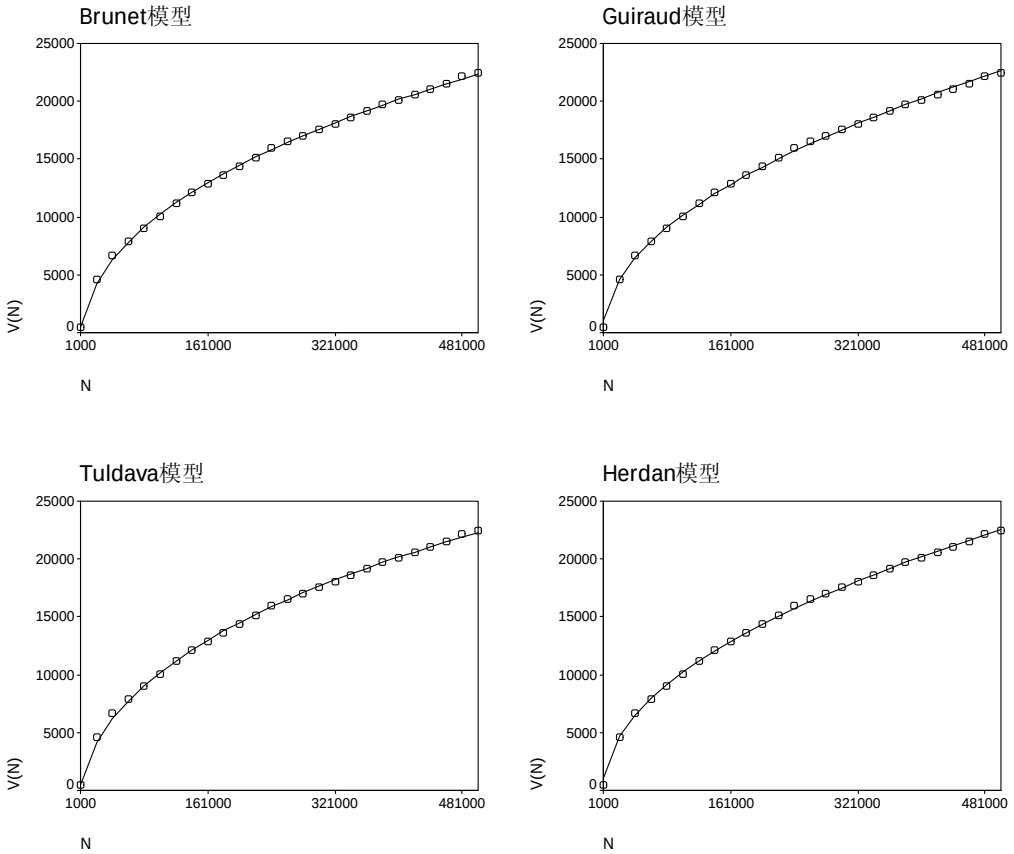


图2.1：Sample子库的经验词汇增长曲线以及Brunet模型（左上图），Guiraud模型（右上图），Tuldava模型（左下图）和Herdan模型（右下图）预测的理论词汇增长曲线。

如图2.1（左上图）所示，Brunet模型较为精确的反映出Sample子库的词汇增长规律，决定系数R^2高达（coefficient of determination）99.931%。回归分析中经常使用R^2来衡量回归模型与观察数据的拟合度，其值越大说明拟合度越高。但统计数据显示Brunet模型在增长曲线

开始部分对词汇的发展模式估算过低，导致词汇量的理论值明显低于实际观察值。详细数据见附录1第三列。

如图2.1（右上图）所示，Guiraud模型不能适当地反映出Sample子库的词汇发展模式。虽然R^2达到99.936%，但是Guiraud模型在增长曲线的后半部分持续低估词汇量的实际观察值。这说明Guiraud模型不能够有效地描述篇际英语词汇增长。详细数据见附录1第四列。

Tuldava 模型虽然能够基本反映出 Sample 子库的词汇增长规律，R^2 达到 99.899%。但是图 2.1（左下图）显示 Tuldava 模型在增长曲线开始和接近结束时过低估计词汇量的实际观察值；而且模型表达形式过于复杂，不适合进行逆推等其他数学运算。数据详见附录 1 第五列。

虽然表达式形式与 Guiraud 模型相近，Herdan 模型对英语词汇增长的描述却是现有模型中最为精确的。R^2 达到 99.940%，超过了上述三种模型。但是如图 2.1（右下图）所示，Herdan 模型也存在问题：在增长曲线开始和接近结束时高估词汇量的实际观察值。数据详见附录 1 第六列。

3. 研究方法与步骤

本文基于 Brown 语料库的证据，研究篇际英语的词汇增幅率。 数据的处理与分析按照以下四个步骤进行：

首先，将 Brown 语料库的所有文本随机分成两个子库——Sample 子库和 Test 子库；每个子库基本统计信息如表 3.1 所示。为确保文本语料抽取的随机性，笔者设计了一组 FoxPro 程序，详见附录 2。

表 3.1：Sample 子库和 Test 子库的总体统计信息。

名称	子库容量	词汇量	平均单个文本容量
Sample	约 500,000	22,464	2,000
Test	约 500,000	21,329	2,000

第二步，以单个独立文本为计量单位，对 Sample 子库和 Test 子库分别进行词干化，计算出累积文本容量和对应的词汇量，并画出篇际英语词汇增长曲线图。

第三步，基于 Sample 子语料库证据，检验四种现有模型对英语词汇增长的描述能力，再通过对幂函数和对数函数的比较分析，构建出新的词汇增长模型。

第四步，基于 Test 子语料库证据，对新模型进行拟合度检验和参数估算，并应用此模型推导理论英语词汇增长曲线及其 95%双向置信区间。

4．新的英语词汇增长模型

为了确保文本内部词汇结构的完整性，本文将以单个独立文本（individual text）为计量单位，对Brown语料库及其子库（Sample子库和Test子库）进行篇章之间的词汇增幅率研究。图4.1为Sample子库（左图）和Test子库（右图）的篇际词汇增长曲线。X轴代表累积文本的单词总数*N*，Y轴代表对应的词汇量*V(N)*。

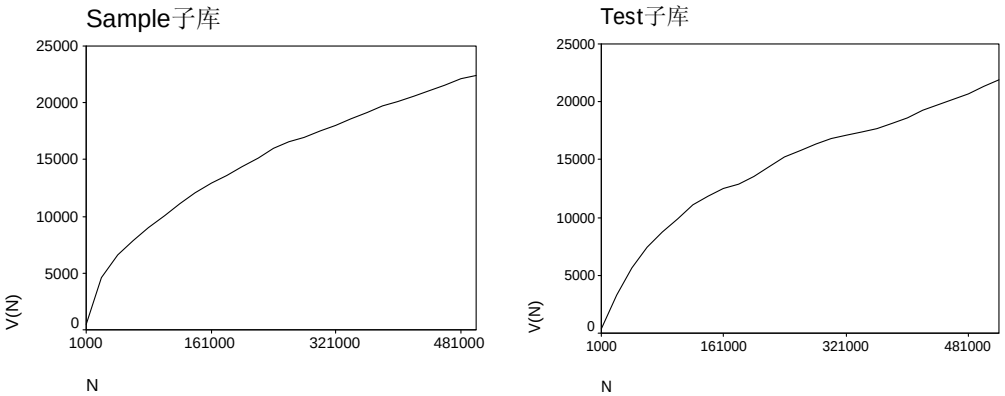


图4.1：Sample子库（左图）和Test子库（右图）的词汇增长曲线。

Sample 子库和 Test 子库的词汇增长呈现出高度的相似性。以左图为例，词汇量 *V(N)*与

累积文本容量 N 之间成曲线递增关系。起初，$V(N)$增长很快，但是随着 N 值的增大，$V(N)$的增长率逐渐减小。当到达平面末端时，增长曲线仍然保持递增状态。这说明，当累积文本容量达到 500,000 形符时，仍然有新的词汇出现。

4.1 新模型的构建思路

在现有词汇增长模型中，Brunet模型和Herdan模型分别代表不同的函数关系——对数函数和幂函数。这两类函数虽然能够基本反映出英语词汇的增长规律，却各自存在着缺陷。

式4.1为Brunet模型的推导式：

$$\log_w V(N) = \frac{1}{\alpha}\log_w(\log_w N) \tag{4.1}$$

式4.2为对数函数的一般表达式：

$$y = \alpha \log_\beta x \tag{4.2}$$

设 $x = \log_w N$，$y = \log_w V(N)$，式4.1可变为：

$$y = \frac{1}{\alpha}\log_w x \tag{4.3}$$

比较式4.2和式4.3，我们可以发现Brunet模型实质上是一个复杂的对数函数关系式，以$\log_w N$为自变量（$x = \log_w N$），$\log_w V(N)$为因变量（$y = \log_w V(N)$）。虽然在表达形式上有所不同，但Brunet模型作为对数函数的一个变体，必然具备对数函数的某些性质，而对于词汇增长的描述也必然受到对数函数的影响。图4.2为对数函数的一般表达式（式4.2）模拟Sample子库的词汇增长曲线。X轴为累积文本容量，Y轴为累积词汇量，圆圈代表实际观察值，实线代表理论预测值。

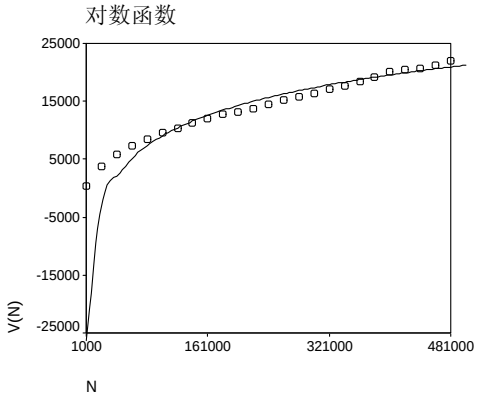


图4.2：对数函数模拟Sample子库的词汇增长曲线，以单个独立文本为统计单位。

如图4.2所示，对数函数无法有效地描述英语词汇增幅率，尤其是在增长曲线开始和接近结束时过低估计词汇量的实际发展，导致词汇量的理论值明显低于实际观察值。Brunet模型虽然在表达式上对对数函数做出了适当的调整和改进，但统计数据显示它不能够从根本上解决对数函数自身存在的问题（图2.1—左上图）。

式4.4为Herdan模型的表达式：

$$V(N) = \alpha N^{\beta} \tag{4.4}$$

式4.5为幂函数的一般式：

$$y = \alpha x^{\beta} \tag{4.5}$$

比较式4.4和式4.5，我们可以发现Herdan模型本身就是幂函数的一般表达式，以N为自变量（$x = N$），$V(N)$为因变量（$y = V(N)$），它虽然比较精确地反映出英语的词汇增长规律，但问题在于：Herdan模型在增长曲线开始和接近结束时过高估计词汇量的实际观察值（图2.1—右下图）。

通过对Brunet模型和Herdan模型的比较分析，我们发现：虽然对数函数和幂函数都可以用来描述英语词汇的变化规律，但却各有利弊。对数函数的主要问题在于：在增长曲线开始和接近结束时**低估**词汇量的实际观察值；而幂函数的问题则是：在增长曲线开始和接近结束时**高估**词汇量的实际值。因此我们提出，对数函数和幂函数的数学结合式更为适合篇际英语

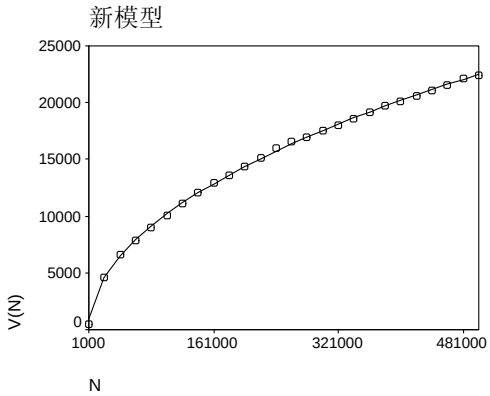
词汇增长的描述，并以此为据构建新的词汇增长模型。

4.2 新模型的函数表达式

新的英语词汇增长模型属于经验模型，以Brown语料库作为数据支持，其表达形式如下：

$$V(N) = \alpha \times \log N \times N^{\beta} \tag{4.6}$$

新模型是对数函数与幂函数的乘式，以*N*为自变量，*V(N)*为因变量。参数α和β均是经验系数，不存在概率解释，其取值会随文本容量的变化呈现出细微差别。图4.3为新模型模拟Sample子库的词汇增长曲线。X轴为累积文本容量，Y轴为对应的词汇量。圆圈表示实际观察值，实线表示理论期望值。



4.3：新模型模拟Sample子库的词汇增长曲线，以单个独立文本为计量单位。

如图4.3所示，新的词汇增长模型能够精确地描述Sample子库的词汇发展模式，尤其是在增长曲线开始和接近结束时准确地预测出词汇量的实际观察值。决定系数达到99.955%，高于四种现有的数学模型——Brunet模型（99.931%），Guiraud模型（99.936 %），Tuldava模型（99.899%）和Herdan模型（99.940%）。参数α与β的取值分别如下：

α= 17.521733040　　β= 0.412847789

4.3 对新模型的拟合度检验

Test子库用来评估新模型对于篇际英语词汇增长的描述能力。图4.4为新模型模拟Test子

库的词汇增长曲线。X轴为累积文本容量，Y轴为累积词汇量。圆圈表示实际观察值，实线表示模型预测值。参数α与β分别沿用图4.3得出的理论调整值（α=17.521733040，β=0.412847789）；原因在于Sample子库和Test子库选自同一个Brown英语语料库，且库容相近，均为500,000形符左右，所以由Sample子库作为数据支持推导出的参数值同样适合Test子库的词汇增长曲线。

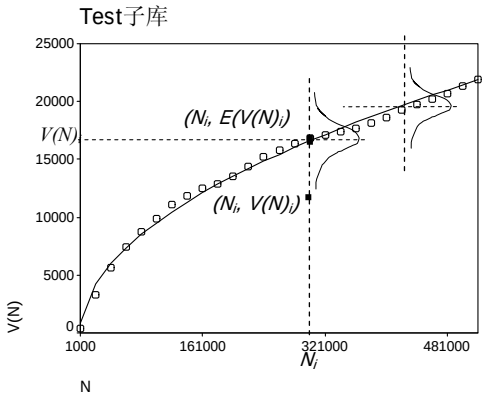


图4.4：新模型模拟Test子库的词汇增长曲线，以单个独立文本为计量单位。

沿用图4.3的参数调整值，新模型依然精确地反映出Test子库的篇际词汇增长模式（图4.4），决定系数达到99.682%，高于四种现有的数学模型——Brunet模型（99.679%），Guiraud模型（=99.659%），Tuldava模型（99.634%）和Herdan模型（99.665%）。因此，新模型最为适合篇际英语词汇增长的描述。表4.1为五种模型模拟英语词汇增长的拟合度数据。

表4.1：五种模型模拟Sample子库和Test子库的词汇增长曲线的拟合度数据。

模型名称	模型表达式	决定系数R^2	
		Sample子库	Test子库
Brunet模型	$V(N)=(\log_N W)^{-\frac{1}{\alpha}}$	99.931%	99.679%
Guiraud模型	$V(N)=\alpha\sqrt{N}$	97.936%	99.659%
Tuldava模型	$V(N)=Ne^{-\alpha(\ln N)^{\beta}}$	99.899%	99.634%

Herdan模型	$V(N)=\alpha N^{\beta}$	99.940%	99.665%
新模型	$V(N)=\alpha\times\log N\times N^{\beta}$	99.955%	99.682%

4.4 理论英语词汇增长曲线的95%置信区间估计

新的词汇增长模型是概率模型，所以由新模型预测的理论增长曲线也具有概率性质，即对于任意给定的累积文本容量*N*，累积词汇量*V(N)*会以不同的程度偏离其理论预测值*E(V(N))*，且偏离的程度是随机的、相互独立的。换言之，对于任意文本容量*N*，其对应的实际词汇增长点*(N,V(N))*可能精确地落在理论模型曲线上，也可能落在模型曲线以外的平面上，详见图4.4。Devore（2000）证明当样本数量足够大时，累积词汇量的实际观察值*V(N)*会以其预测值*E(V(N))*为均值成正态分布状态（图4.4）。因此，根据正态分布的置信区间计算公式（式4.7），我们可以求出词汇增长曲线上任意点*(N, V(N))*的95%双向置信区间。

$$\begin{aligned} V(N)_{upper} &= E(V(N))+(critical\ value)\times s \\ V(N)_{lower} &= E(V(N))-(critical\ value)\times s \end{aligned} \qquad (4.7)$$

在式4.7中，$V(N)_{upper}$和$V(N)_{lower}$分别表示累积词汇量的置信上限和置信下限，*E(V(N))*为词汇量的理论预测值，*critical value*为临界值，*s*为样本标准差。

标准差*s*是测量样本离散度的重要指标，它能够综合反映词汇量正态分布曲线的离散程度（图4.4），即实际观察值*V(N)*以其预测值*E(V(N))*为均值的概率分布以及观察值个体之间的差异程度。简而言之，*s*值越大，*V(N)*偏离*E(V(N))*的程度越大；*s*值越小，*V(N)*相对于*E(V(N))*的离散趋势越小。从理论上讲，增长曲线上有无穷多个词汇增长点，我们无法穷尽每一点的标准差计算。因此，为确保文本内部词汇结构的完整性，本文将以单个独立文本为统计单位选取词汇增长点并计算其对应的正态分布的标准差。具体步骤如下：

首先，对Brown语料库的所有文本进行随机抽取再任意排序，反复10次，得到10个文本集合（$N\geq 500{,}000$）。

然后，以单个独立文本为计量单位，对每一个文本集合进行词干化，计算其累积文本容量和对应的词汇量。

最后，使用SPSS统计软件计算出每一个增长点对应的正态分布的标准差。

根据Devore（2000）理论，当样本数量为10时，95%双向置信区间的临界值为3.379。我们将标准差s、临界值*critical value*和新模型预测的理论词汇量$E(V(N))$代入公式4.7中，可以求出词汇增长曲线上每一个增长点的95%置信上限$V(N)_{upper}$和置信下限$V(N)_{lower}$。表4.2为部分词汇增长点的95%置信区间的统计数据，包括N，$E(V(N))$，s，$V(N)_{upper}$和$V(N)_{lower}$。

表4.2：英语词汇增长曲线上部分增长点的95%置信区间的统计数据，包括N，$E(V(N))$，s，$V(N)_{upper}$和$V(N)_{lower}$。

N	$E(V(N))$	s	$V(N)_{max}$	$V(N)_{min}$
31000	5626	606	7673	3579
61000	7927	968	11199	4655
91000	9690	1056	13258	6121
121000	11171	1144	15038	7304
151000	12472	1121	16260	8685
181000	13646	1181	17635	9657
211000	14722	1125	18523	10920
241000	15721	1142	19579	11863
271000	16657	1089	20338	12977
301000	17541	1090	21225	13858
331000	18380	963	21634	15126
361000	19181	987	22516	15845
391000	19947	1023	23403	16491
421000	20683	1026	24150	17216
451000	21392	1119	25172	17613
481000	22078	1116	25849	18306

图4.5为新模型预测的篇际英语词汇增长曲线的95%置信区间。X轴为累积文本容量，Y轴为累积词汇量。圆圈代表Test子库的经验词汇增长曲线，实线代表新模型预测的理论词汇增长曲线，虚线代表新模型推导的95%双向临界增长曲线。

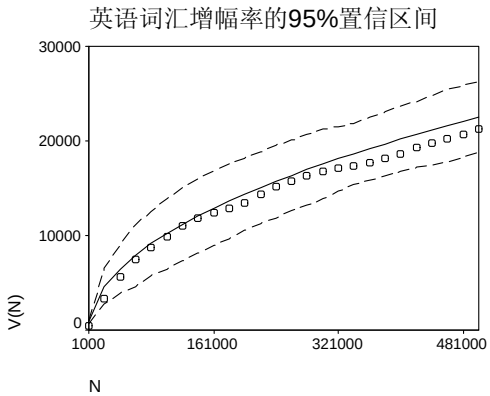


图4.5：新模型预测的英语词汇增长曲线的95%置信区间。

如图4.5所示，经验增长曲线全部落在95%上下临界曲线之间的片面上，无一例外。这说明新模型能够对英语词汇增长曲线做出适当的置信区间估计，也再次证明了新模型适合篇际英语词汇增长的描述。

5. 模型应用示例

新模型能够精确地反映出英语词汇增长规律，因此对教材编写、大纲设计、二语习得研究以及外语教学实践等都有一定的意义。它不仅可以用来预测任意给定语篇的词汇量，即已知 N 求 $E(V(N))$ 及其 95%置信区间，也可以用于估计任意累积词汇量对应的语篇数，即已知 $V(N)$ 求 $E(N)$ 及其 95%置信区间。

示例 1：假设一本英语教材有 30 篇独立文本，每一篇的文本容量在 500 至 2,000 形符之间，共计 30,000 形符。那么根据新的词汇增长模型，我们可以求出该本教材的理论词汇总量。参数 α= 17.5217，β= 0.4128。

$$
\begin{aligned}
E[V(N)] &= \alpha \times \log N \times N^{\beta} \\
&= 17.5217 \times \log 30000 \times 30000^{0.4128} \\
&= 5530
\end{aligned}
$$

根据置信区间计算公式，我们可以求出该本教材理论词汇量的 95%双向置信区间。临界值 *critical value*=3.379，标准差 *s*= 473.32：

$$
\begin{gathered}
E[V(N)] - (critical\ value) \times s \le V(N) \le E[V(N)] + (critical\ value) \times s \\
5530 - 3.379 \times 473.32 \le V(N) \le 5530 + 3.379 \times 473.32 \\
3931 \le V(N) \le 7129
\end{gathered}
$$

所以，我们有95%的把握得出结论：该本英语教材的词汇总量在3,931至7,129类符之间。如果学生在学完教材后可以掌握其中95%的词汇，那么根据Meara提出的词汇习得模型（式1.1），p= 0.95，$V(N)_{i\ min}$= 3931，$V(N)_{i\ max}$=7129：

$$
V(N)_{\min} = p \cdot \sum_{i=1}^{n} V(N)_{i\min} = 0.95 \times 3931 = 3734
$$

$$
V(N)_{\max} = p \cdot \sum_{i=1}^{n} V(N)_{i\max} = 0.95 \times 7129 = 6772
$$

我们可以求出学生的习得词汇量为3,734至6,772类符。

示例2：假设英语考试大纲要求考生的词汇量在4,000类符以上，那么针对该考试编写一本英语教材，其输入文本的理论单词总量应为（α= 17.5217，β= 0.4128）：

$$
\begin{aligned}
&V(N) = \alpha \times \log E[N] \times E[N]^{\beta} \\
&4000 = 17.5217 \times \log E[N] \times E[N]^{0.4128} \\
&E[N] \approx 16000
\end{aligned}
$$

根据置信区间计算公式，理论单词总量的 95%双向置信区间为（*critical value*=3.379，*s*=345.12）：

$$
\begin{gathered}
E[N] - (critical\ value) \times s \le N \le E[N] + (critical\ value) \times s \\
16000 - 3.379 \times 345.12 \le N \le 16000 + 3.379 \times 345.12 \\
14834 \le N \le 17166
\end{gathered}
$$

所以，我们有 95%的把握得出结论：编写该本英语教材需要累积输入 14,834 至 17,166 形符的单词总量才能提供 4,000 类符的词汇量。

6．结论

本文基于 Brown 语料库，从四个角度考察了篇际英语词汇增长规律：（1）对四种现有模型的描述与评估；（2）新模型的提出、拟合度检验及参数估算；（3）理论词汇增幅率的预测及其 95%置信区间的推导；（4）新模型的教学应用示例。

研究表明，现有的词汇增长模型（Brunet 模型，Guiraud 模型，Tuldava 模型及 Herdan 模型）在描述篇际英语词汇增长时各有缺陷。本文通过对幂函数和对数函数的比较分析，构建了新的词汇增长模型，并应用此模型推导出理论英语词汇增长曲线及其 95%双向置信区间。模型表达式如下：

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

新模型是对数函数与幂函数的乘式，以N为自变量，$V(N)$为因变量。参数α和β均为经验系数，不存在概率解释，其取值会随文本容量的变化呈现出细微差别。经多次拟合度检验证明，新模型能够精确地描述篇际英语词汇增幅率，尤其是在增长曲线开始和接近结束时准确地预测出词汇量的实际观察值，其决定系数高于四个现有的数学模型，因此最为适合篇际英语词汇增长的描述。

参考文献

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

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

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

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

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

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

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

8. Herdan, G. 1964. *Quantititative Linguistics.* Buttersworths, London.

9. Meara, P. 1997. Towards a New Approach to Modeling Vocabulary Acquisition. In Schmitt, N., McCarth, M., (eds), *Vocabulary: Description, Acquisition and Pedagogy.* Cambridge: Cambridge University Press, pp.109-21.

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

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

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

13. 冯志伟. 1996.《现代术语学引论》，语文出版社. 北京。

14. 张敏，张秀芬. 2007. 《科技英语阅读教程》. 外语教学与研究出版社. 北京。

附录 1： Brunet 模型，Guiraud 模型，Tuldava 模型和 Herdan 模型分别模拟 Sample 子库词汇增长曲线的部分数据

TOKENS	*TYPES*	*PRED_B*	*PRED_G*	*PRED_T*	*PRED_H*
1000	469	476.40	1010.06	500.94	1035.02
11000	3278	2843.07	3349.98	2739.88	3397.85
21000	4584	4253.02	4628.66	4126.66	4681.88
31000	5666	5353.55	5623.76	5226.46	5678.98
41000	6629	6281.96	6467.52	6162.26	6523.25
51000	7255	7096.65	7213.24	6987.66	7268.64
61000	7871	7829.04	7888.79	7732.05	7943.31
71000	8478	8498.34	8510.89	8413.64	8564.16
81000	8993	9117.32	9090.51	9044.69	9142.27
91000	9472	9694.98	9635.32	9633.93	9685.38
101000	10065	10237.92	10150.94	10187.82	10199.14
111000	10685	10751.20	10641.60	10711.33	10687.83
121000	11141	11238.75	11110.62	11208.36	11154.78
131000	11668	11703.70	11560.62	11682.07	11602.64
141000	12107	12148.63	11993.75	12135.01	12033.57
151000	12500	12575.66	12411.78	12569.32	12449.35
161000	12889	12986.54	12816.18	12986.80	12851.45
171000	13258	13382.80	13208.20	13388.97	13241.15
181000	13611	13765.71	13588.92	13777.15	13619.52
191000	13961	14136.38	13959.26	14152.48	13987.48
201000	14405	14495.79	14320.02	14515.95	14345.85
211000	14672	14844.78	14671.92	14868.43	14695.34
221000	15111	15184.10	15015.57	15210.70	15036.57
231000	15671	15514.42	15351.53	15543.45	15370.10
241000	15970	15836.32	15680.29	15867.29	15696.43
251000	16297	16150.35	16002.30	16182.78	16015.99
261000	16560	16456.98	16317.96	16490.41	16329.20
271000	16810	16756.65	16627.63	16790.64	16636.41
281000	16995	17049.74	16931.63	17083.89	16937.96
291000	17230	17336.62	17230.27	17370.52	17234.14
301000	17532	17617.62	17523.82	17650.88	17525.24
311000	17798	17893.03	17812.54	17925.29	17811.49
321000	18008	18163.14	18096.65	18194.04	18093.14
331000	18297	18428.19	18376.36	18457.40	18370.40
341000	18609	18688.43	18651.89	18715.61	18643.47
351000	18922	18944.06	18923.40	18968.91	18912.53

361000	19147	19195.30	19191.07	19217.51	19177.75
371000	19487	19442.34	19455.06	19461.61	19439.29
381000	19773	19685.34	19715.51	19701.40	19697.30
391000	19978	19924.47	19972.57	19937.05	19951.92
401000	20136	20159.88	20226.36	20168.72	20203.27
411000	20370	20391.73	20477.01	20396.57	20451.48
421000	20618	20620.14	20724.62	20620.74	20696.67
431000	20898	20845.24	20969.31	20841.37	20938.93
441000	21081	21067.15	21211.18	21058.58	21178.38
451000	21265	21285.99	21450.32	21272.49	21415.10
461000	21526	21501.85	21686.83	21483.22	21649.19
471000	21900	21714.85	21920.78	21690.88	21880.74
481000	22136	21925.07	22152.26	21895.56	22109.82
491000	22260	22132.60	22381.35	22097.37	22336.51
501000	22447	22337.53	22608.12	22296.39	22560.88

附录 2：文本随机抽取程序

```
set defau to d:\jingjie\english\corpora\brown
set safe off
CREATE TABLE mtext (text m(4))
CREATE TABLE word1 (word c(50))
CLOSE DATABASES
USE mtext
APPEND BLANK
APPEND MEMO text FROM t1.txt sdf
repla all text with strtran(text,chr(13),' ')
repla all text with strtran(text,chr(10),'')
repl all text with strtran(text,' ',chr(13))
if chr(13)+chr(13)+chr(13)$text
    repla all text with strtran(text,chr(13)+chr(13)+chr(13),chr(13))
endif
close data
USE mtext
COPY MEMO text TO ntext.txt sdf
CLOSE DATABASES
USE word1
APPEND field word from ntext sdf
repla all word with strtr(word,chr(13),' ')
repla all word with strtr(word,chr(10),'')
dele all for len(alltrim(word))=0
PACK
FOR i=1 TO RECCOUNT()
    LOCATE FOR ALLTRIM(word)='~'
    a=RECNO()
    CONTINUE
    IF NOT EOF()
        b=RECNO()
        c='T'+ALLTRIM(STR(i))
        COPY TO &c FOR RECNO()>=a AND RECNO()<b
        DELETE ALL FOR RECNO()>=a AND RECNO()<b
        PACK
    ELSE
    EXIT
    ENDIF
ENDFOR
GO top
LOCATE FOR ALLTRIM(word)='~'
a=RECNO()
```

```
CONTINUE
IF .T.
b=RECNO()
c='passage'+ALLTRIM(STR(i))
COPY word1 to passage1 for RECNO()>=a and RECNO()<b
SELECT 3
USE &c
COPY TO
DELETE ALL FOR RECNO()>=a and RECNO()<b
PACK
IF not.T.
EXIT
ENDIF
ENDFOR
```