基于笔语语料库的大学英语学生词汇发展研究*

栾 岚
(上海外国语大学,上海 200083;哈尔滨工程大学,哈尔滨 150001)

提 要: 作为二语习得研究的重要组成部分,词汇评估不仅有助于了解学习者第二语言的掌握程度和学习者的共有特点,而且对促进二语教学具有重要意义。本文以非英语专业(大学英语)学生的定时作文为语料,通过自建小型历时语料库和 Complete Lexical Tutor 提供的 Vocabprofile 功能来分析语料库中的数据,结果表明:词汇发展总体来说是随着年级的增加呈现上升趋势,但上升幅度不显著。高频词中学术词汇(AWL)的变化基本上呈线性增长;词汇的多样性不断提高,学习者用词范围变宽;比起学习者词汇多样性和复杂性的变化幅度,词汇密度的变化呈现出缓慢上升趋势。

关键词: 纵向发展;词汇多样性;词汇复杂性;词汇密度;非英语专业学生
中图分类号: H319.34 **文献标识码:** A **文章编号:** 1000-0100(2013)02-0126-5

Lexical Development of Non-English Major Students
— A Study Based on Written English Corpus

Luan Lan
(Shanghai International Studies University, Shanghai 200083, China;
Harbin Engineering University, Harbin 150001, China)

Lexical evaluation has undoubtedly made much contribution to the study of second language learning. It not only casts some light on the process of second language learning, but also helps us gain some insights into the common characteristics shared by language learners, which is believed to be of great help to language teachers and language learners alike. Based on the corpus of non-English major students' timed compositions, this paper tends to analyze the data in the mini self-built diachronic corpus with the help of "Vocabprofile" provided by Complete Lexical Tutor. The findings are as follows: lexical development in general increases along with the grade, but the increase is not significant. Academic word lists show a linear increase, the lexical variation is also rising and learners have a wider range of words to choose from. Compared with the lexical variation and lexical sophistication, lexical density shows a slower increase.

Key words: longitudinal development; lexical variation; lexical sophistication; lexical density; non-English major student

1 引言

自 20 世纪 80 年代以来,在 Meara 等人的大力倡导和推动下,词汇习得逐渐成为二语习得研究的热点。但根据目前的研究状况,针对词汇能力各个维度的发展的研究才刚刚开始(Laufer 1998, Qian 2004, Schmitt & Meara 1997, 李雪 2012),尚未得出比较全面的结论。本文以非英语专业(大学英语)本科生为研究对象,考察学习者在 1-2 年级共 4 个学期笔语产出性词汇量的发展变化规律。进行本研究的目的是:(1)利用自建的历时语料库,从产出角度考察学习者的词汇量变化情况;(2)从纵向角度探索学习者词汇在各个阶段的发展变化规律,揭示词汇能力发展的动态特点。

目前,词汇评估的方法主要有两类。一类是以目标词汇为中心设计词汇测试。它包括接受性词汇和产出性词汇测试,测试词汇知识的类型涉及词汇量和词汇知识深度。典型的测试方式是 Wesche

* 本文系教育部人文社科研究项目"基于历时语料库的中国英语学习者写作能力发展研究"(11YJA740122)的阶段性成果。

& Paribakht 设计的词汇知识量表(the vocabulary knowledge scale,简称 the VKS) (Wesche & Paribakht 1996) 和词语联想测试 (Schmitt & Meara 1997) 。另一类词汇评估将词汇作为完成一项交际任务(口语产出和书面作文) 不可或缺的一部分,考察学习者在完成交际任务过程中是如何使用词汇的(Hyltenstam 1998,Laufer 1991,Engber 1995) 。这类评估克服了第一类词汇测试脱离或削弱语境作用以及忽视语言交际功能的缺陷。

国内已有很多学者通过语料库研究学习者词汇发展模式(林峻 2008,刘文慧 阳志清 2009,童淑华 2009,郭红霞 2011,王改燕 万霖 2011) 。但是这些研究存在以下不足: (1) 未严格界定词汇发展的维度,使研究结果难以比较; (2) 在研究学习者词汇发展规律时较少涉及产出性词汇; (3) 研究对象基本上限于两个不同组别之间或同一组两个不同阶段之间的比较,难以揭示学习者词汇丰富性发展的态势。正是这些局限使得我们不能全面解释决定中介语发展变化的原因。因此,Tarone 呼吁进行更多的纵向研究,以"揭示中介语某一发展阶段特有的变化规律"(Tarone 1988: 137) 。为了深入了解中介语的发展轨迹,Cobb 认为,'理想的研究方法是采用同一组学习者历时几年的大量语料"(Cobb 2003: 401) 。在中国,最能体现学习者语言水平的方式是写作,写作需要语法、词汇、目的语文化知识等支撑,特别是词汇知识的支撑。因此,本文利用自建的历时语料库,以词汇多样性、词汇复杂性和词汇密度 3 个方面的数据为测量标准,调查我国非英语专业(大学英语) 本科生在 1 – 2 年级期间笔语词汇的发展变化规律。

2　研究设计

2.1 《非英语专业学生笔语历时语料库》的设计理据

本文的研究数据来源于笔者自建的《非英语专业学生笔语历时语料库》,笔语语料来自哈尔滨工程大学 2008 级 5 个班级的学生 1 – 2 年级的 4 次大学英语课程期末考试作文(全部为议论文) 文本,这 5 个班级的学生分别来自该校的人文学院、自动化学院、信息与通信工程学院、机电学院和核学院。语料库规模为 72,243 形符,笔语语料库基本信息详见表$_1$。

笔语历时语料库中的语料均为非英语专业学生期末考试试卷中的作文部分,考试形式与大学英语四级考试中的作文形式相同,时间为 30 分钟,长度要求为 120 – 150 词。由于在文体和写作条件上有所控制,本语料库为研究笔语词汇发展变化提供严格的变量控制,能够更为细致、准确地反映大学英语学生笔语词汇的发展变化规律。

为了更准确地反映中介语笔语的发展变化规律,《非英语专业学生笔语历时语料库》进行了词性赋码和失误标注,语料以两种形式存储: (1) 生语料(raw corpus) ; (2) POS 词类赋码语料。

表$_1$　《非英语专业学生笔语历时语料库》基本信息表

学期	题目	人数	平均字数(篇)	总形符数
第 1 学期	How Should We Spend Our College Time?	115	155.3	17,856
第 2 学期	Where to Learn a Language?	115	171.0	19,670
第 3 学期	It is Necessary to Develop Tourism	115	145.0	16,673
第 4 学期	Global Warming — A Serious Problem in the World	115	157.0	18,044
总　计				72,243

根据一定标准和要求收集的未经处理的电子版学习者语料为"生语料"。生语料提供的信息有限,为了使语料库"增值",研究者通常要对生语料进行赋码或标注处理,因为从赋码后的语料库中可以提取远远超过从生语料中所能提取的信息(Meunier 1998: 45) 。对生语料的常规赋码计算机程序为削尾程序(Lemmatizers,又称词目还原) 、词类赋码(POS-taggers) 、句法切分器(parsers) 和失误标注(error tagging) 。WECCL 对生语料用兰卡斯特大学开发的 CLWAS4 赋码软件进行自动词性赋码。WECCL(2.0) 版还提供一个经过词目还原(Lemmatized) 的文本,文本所有词的屈折变化都已经还原为原形(如 drives/drove/driven/driving 都还原为 drive, am/are/is/was/were 都还原为 be) 。词目还原后的文本适用于词汇分析,如词汇密度等,可以最大程度排除英语词汇屈折变化对密度计算的干扰。

Lemmatizers 和 TreeTagger 等均为处理本族人语料而开发,虽然在本族人语料中的使用准确性很高,但学习者语料中的应用还处于尝试阶段。经 Lemmatizers 处理后的语料对研究词汇密度非常有利,但在中介语学习者语料中使用时会遇到

一些问题。只有正确标准地使用的词汇，才能实现正确的词目还原，如 lose/loses/losing/lost 都可以还原成词目 lose，但如果学生使用错误形式，如 loose/looses/loosing/loosed，程序将无法操作。而这类错误在学习者语料库中非常常见。另外，还有拼写错误，如 addickted(addicted) 等; 构词失误，如 studyed(studied) 和 photoes(photos) 等。这些都会影响削尾程序运作的准确性。而在接近 10 万词的语料中，人工校对削尾错误无疑工作量巨大，因此本研究不对语料做此项处理。

2.2 词汇信息的获取

本研究利用由加拿大魁北克大学开发的免费语料库网站 Complete Lexical Tutor 提供的 Vocabprofile 功能计算出词汇复杂性数据。将文本输入后，网站会计算出文本中的最常用 1000 词(1K) 、次常用的 1000 词(2K) 和 570 个学术词汇(AWL) 的数量和百分比，不在以上 3 个范围内的列入 Off List 中。同时，这个网站还可以计算出文本的形符(token) 、类符(type) 、词汇密度(lexical density) 等信息。

2.3 研究问题

本研究的主要问题是受试者在 1 －2 年级的 4 个学期中的词汇变化情况: 词汇复杂性、词汇多样性和词汇密度。

2.4 研究方法

John Read 认为，“笔语词汇的丰富度主要体现在词汇多样性(lexical variation，简称 LV) 、词汇复杂度(lexical sophistication，简称 LS) 和词汇密度(lexical density，简称 LD) 3 个方面”(Read 2000: 200) 。本研究使用的是未经赋码的生语料，对非英语专业大学生 1 －2 年级期间笔语词汇丰富度发展的描述主要依据词汇多样性、词汇复杂性和词汇密度 3 个指标。这 3 种测量指标的计算方式如表$_2$ 所示。

表$_2$　本研究的测量指标及计算公式

测量指标	内容	计算公式
词汇多样性	1K、2K、AWL 词在文本中所占的比例	1K/2K/AWL ÷ 词汇总数 ×100%
词汇复杂性	形、类符比(type-token ratio，TTR)	类符数 2 ÷ 形符数
词汇密度	实词在总词数中所占的比例	实词数 ÷ 词汇总数 ×100%

词汇多样性(TTR) 是指“语料中出现的类符与形符的比率”(Read 2000: 203) 。本研究使用的计算方式为 TTR = 类符数 2/形符数，虽然这种 TTR 不能像简单的类符数/形符数 ×100 那样比较直观地表示出百单词中的类符数，但却能够一定程度地克服样本长度的干扰，国内已有学者研究使用这种 TTR.

词汇复杂性是研究文章中除普通日常词汇以外相对复杂词汇比例的测量指标。Laufer 和 Nation(1995) 设计词汇频率概貌(lexical frequency profile，简称 LEP) ，该测试系统拥有一个词频表，由英语前 2000 个高频词族(分别以 1K 和 2K 表示) 和 570 个学术词汇(academic word list，简称 AWL) 组成。具体做法是先把文本中所有词汇与英语中最常用的 1000 个词比较，然后与次常用 1000 个词比较，所得比率是作文中使用的词汇在这两类高频词中出现的百分比以及英语最常用的 2000 个词(具有一定难度的相对低频词汇) 之外的词汇比率。

词汇密度测量实词在整个文本中的比率，考察文本的信息含量。Ure(1971) 提出词汇密度的计算公式为实词数除以词汇总数得到的百分比: 词汇密度 = 实词数/词汇总数 ×100%，这是本研究采用的计算工具。

3　研究的结果与讨论

3.1 词汇多样性

本研究对受试者在 1 －4 学期中的词汇多样性进行了差异比较，详见表$_3$。

表$_3$　1 －4 学期词汇多样性差异比较

学期	平均值	人数	标准差	平均数差异	显著性
第一学期/第二学期	45.98	115	8.64	－0.247	0.805
	6.20	115	10.20		
第二学期/第三学期	46.20	115	10.20	－0.683	0.496
	46.92	115	10.33		
第三学期/第四学期	46.92	115	10.33	－7.082	0.000
	54.40	115	8.99		
第一学期/第四学期	45.98	115	8.64	－7.636	0.000
	54.40	115	8.99		

表$_3$ 中的统计数据表明，在 1 －4 学期中，学生的词汇多样性有所提高，虽然前 3 个学期的均值很接近，几乎看不出发展迹象，但第 3 和第 4 学期的均值却呈现出激增趋势。对第 1 和第 2、第 2 和第 3 学期的数据配对样本分析的结果是提高幅

度不具有显著性(t 值分别为 -0.247、0.683，p > 0.05)，而对第 3 和第 4，第 1 和第 4 学期的数据配对样本分析的结果是提高幅度具有显著性(t 值分别为 -7.082、-7.636，p < 0.001)。这表明随着年级的增加，学生作文中词汇的使用范围在拓展，重复用词的现象有所减少。

3.2 词汇复杂性

表$_4$ 显示笔语历时语料库中受试者1 -4学期写作中词汇的复杂性变化。

表$_4$　1 -4 学期词汇频率概貌数据表
(平均值)

项目	学期	平均值	人数	标准差	平均数差异	显著性
2000 高频词	第一学期/	96.18%	115	2.17%	10.26	0.000
	第二学期	92.84%	115	2.74%		
	第二学期/	92.84%	115	2.74%	7.27	0.000
	第三学期	89.35%	115	4.66%		
	第三学期/	89.35%	115	4.66%	-7.66	0.000
	第四学期	92.79%	115	2.56%		
	第一学期/	96.18%	115	2.17%	11.46	0.000
	第四学期	92.79%	115	2.56%		
AWL 词汇	第一学期/	1.56%	115	1.23%	-7.55	0.000
	第二学期	2.52%	115	1.33%		
	第二学期/	2.52%	115	1.33%	-9.71	0.000
	第三学期	4.66%	115	2.41%		
	第三学期/	4.66%	115	2.41%	0.19	0.850
	第四学期	4.62%	115	1.95%		
	第一学期/	1.56%	115	1.23%	-15.14	0.000
	第四学期	4.62%	115	1.95%		

表$_4$ 为 2000 高频词和学术词汇(AWL) 在第 1 至第 4 学期的平均值及其差异比较。表$_4$ 中的数据表明，在 1 -4 学期中，学习者的笔语产出词汇成分有一定变化。首先是最常用的 2000 词变化幅度较大，在第 1 至第 3 学期中呈现出直线下降的趋势，且降幅明显，而第 3 和第 4 学期则又反弹至与第 1 学期相差不多的水平上。我们认为，学习者在 1 -4 学期中的 2000 高频词的变化有很强的显著性。其次，非英语专业学生在 4 个学期中的学术性词汇(AWL) 的变化基本上呈线性增长，在第 4 学期有小幅度下降，下降比例为 0.04%，第 2 和第 3 学期学术性词汇在学生作文中所占的比例越来越大，第 3 学期达到顶点 4.66%。因此，可以认为: 学生写作中的词汇复杂度随着年级的增长不断提高。学术性词汇在 1 -3 学期提高最快，可能是因为学生刚刚进入大学，高中时的学习劲头仍然存在，同时学生接触的教材词汇难度高于高中时期，以及来自周围同学的竞争压力。在这些因素的综合作用下，学生取得的进步较快，并且愿意在考试中使用学到的专业词汇。在 3 -4 学期，由于学生在校学习状态的改变，英语学习已经不是最主要的任务，所以这一阶段学术性词汇的比例不升反降。

3.3 词汇密度研究

一般说来，口语的词汇密度大概在 0.4 以下，而书面语的词汇密度则在 0.4 以上。从表$_5$ 中的词汇密度数据可知，学生在 4 个学期的词汇密度都在 0.4 以上，符合书面语的要求，而且词汇密度的发展，除第 1 和第 4 学期以外，均呈现上升趋势，配对样本检验的结果是第 1 和第 2 学期、第 2 和第 3 学期、第 3 和第 4 学期、第 1 和第 4 学期的提高幅度具有显著性(t 值分别为 -2.511，-8.014，2.506，6.275; p = 0.013，0.000，0.014，0.000，分别都小于 0.05)。尽管第 4 学期学生写作中的词汇密度较第 3 学期有小幅下降，但总体上讲，学生在 4 个学期中的词汇密度还是呈上升趋势。这说明随着年级增加，学生的实词词汇量增加，作文的信息含量提高。

表$_5$　1 -4 学期间词汇密度差异比较

学期	平均值	人数	标准差	平均数差异	显著性
第一学期/ 第二学期	0.49	115	0.03	2.511	0.013
	0.50	115	0.04		
第二学期/ 第三学期	0.50	115	0.04	-8.014	0.000
	0.53	115	0.03		
第三学期/ 第四学期	0.53	115	0.03	2.506	0.014
	0.52	115	0.04		
第一学期/ 第四学期	0.49	115	0.03	-6.275	0.000
	0.52	115	0.04		

4　结束语

本研究通过对笔者自建历时语料库中的非英语专业学生 1 -4 学期的大学英语课程期末考试作文的数据处理及分析，探索学习者词汇的复杂性、多样性、词汇密度方面的变化模式。通过综合平均值和 T 检验的数据分析，得出结论: 词汇发展总体来说是随着年级的增加呈现上升趋势，但上升的幅度不是很大。高频词中只有 1K 词呈逐年下降趋势，学术词汇变化基本上呈线性增长，在第 4 学期有小幅度下降，第 2 和第 3 学期学术性词汇在学生的写作中所占的比例越来越大，第 3 学期达到顶点，提高显著; 同时，词汇的多样性也

不断提高，尤其是第 3 和第 4 学期的提高幅度最大，学生用词范围变宽。比起学习者词汇多样性和复杂性的变化幅度，词汇密度的变化虽然呈现出上升趋势，但幅度小，并没有出现在某个学期急剧上升的现象，而是以缓慢的速度逐渐递增。作为我们正在进行的中国英语学习者词汇能力发展研究的一部分，本文只考察词汇发展的多样性、复杂性和词汇密度的 3 个方面。本研究未来的发展应在理论构建、研究对象、实验设计等方面加以改进，努力揭示我国英语学习者词汇能力发展的特点与规律，使英语词汇教学更具系统性、目的性与实效性。

参考文献

郭红霞. 二语词汇习得中跨语言迁移的语言类型分析[J]. 外语学刊，2011(2).

李　雪. 概念隐喻、概念转喻与词汇研究[J]. 外语学刊，2012(4).

林　峻. 我国二语学习者记叙文写作中产出性词汇的发展状况：一项基于语料库的研究[J]. 扬州教育学院学报，2008(4).

刘文慧 阳志清. 英语专业学生笔语词汇发展研究[J]. 中南大学学报(社会科学版)，2009(5).

童淑华. 英语专业学生口语产出性词汇发展的实验研究[J]. 外语学刊，2009(5).

王改燕 万　霖. 二语阅读中语境线索水平对词义推测的影响[J]. 外语学刊，2011(6).

Cobb，T. Analyzing Late Interlanguage with Learner Corpora：Quebec Replications of Three European Studies[J]. *The Canadian Modern Language Review*，2003(3).

Engber，C. A. The Relationship of Lexical Proficiency to the Quality of L2 Compositions[J]. *Journal of Second LanguageWriting*，1995(4).

Hyltenstam，K. Lexical Characteristics of Near-native Second-language Learners of Swedish[J]. *Journal of Multilingual and Multicultural Development*，1998(9).

Laufer，B. The Development of L2 Lexis in the Expression of the Advanced Learner[J]. *The Modern Language Journl*，1991(75).

Laufer，B & Nation，P. Vocabulary Size and Use：Lexical Richness in Written Production[J]. *Applied Linguistics*，1995(16).

Laufer，B. The Development of Passive and Active Vocabulary in Second Language：Same or Different[J]. *Applied Linguistics*，1998(19).

Meunier，F. Computer Tools for Interlanguage Analysis：A Critical Approach[A]. In Granger，S.(ed.). *Learner English on Computer*[C]. London and New York：Addison Wesley Longman，1998.

Qian，D. D. Evaluation of an In-depth Vocabulary Knowledge Measure for Assessing Reading Performance[J]. *Language Testing*，2004(21).

Read，J. *Assessing Vocabulary*[M]. Cambridge：Cambridge University Press，2000.

Schmitt，N. & Meara，P. Researching Vocabulary Through a Word Knowledge Framework：Word Associations and Suffixes[J]. *Studies in Second Language Acquisition*，1997(19).

Tarone，E. *Variation in Interlangauge*[M]. London：Edward Arnold，1988.

Ure，J. Lexical Density and Register Differentiation[A]. In G. E. Perren & I. L. M. Trim(eds). *Application of Linguistics*[C]. Cambridge：Cambridge University Press，1971.

Wesche，M. & Paribakht，T. S. Assessing Second Language Vocabulary Knowledge：Depth Versus Breadth[J]. *Canadian Modern Language Review*，1996(53).

收稿日期：2012－11－15　　**【责任编辑　孙　颖】**