

英语专业学生笔语词汇发展研究

刘文慧, 阳志清

(中南林业科技大学语言与教育技术研究所, 湖南长沙, 410006)

摘要: 本研究采用中国学习者英语口语语料库的笔语语料库子库作为样本, 参照兰卡斯特大学的 LANCAWAE 语料库, 从词汇多样性和词汇复杂度两方面分析了中国英语专业学生笔语词汇的发展, 并将这一发展与其口语词汇发展, 和国外学生笔语词汇发展作了比较, 结果发现: 我国英语专业学生笔语词汇发展不是一个持续上升过程, 这个发展轨迹与口语词汇不同; 我国英语专业学生在笔语词汇多样性和词汇复杂度两个指标上, 均低于国外非本族语学生和本族语学生。

关键词: 语料库; 笔语; 词汇多样性; 词汇复杂度

中图分类号: H313

文献标识码: A

文章编号: 1672-3104(2009)02-0296-05

近年来, 国内外学者对二语习得研究越来越广泛深入, 在国内对笔语词汇的研究比较多, 但大多是对语料的共时研究。揭示学习者二语习得的发展规律和发展趋势, 开展学习者语料的历时研究, 已成当务之急。文秋芳教授^[1]以大学英语专业 56 名学生四年跟踪的口语语料对英语专业学生口语词汇进行历时研究, 从流利性、词汇多样性与词频广度分析了学生口语词汇的使用特点和学生口语词汇进步模式变化的趋势。受此启发, 本研究拟以采用中国学习者英语口语语料库(SWECCL)语料库中的书面语语料库子库(WECCL)为依据, 从词汇多样性、词汇复杂度两方面来分析笔语词汇的使用特点, 以揭示英语专业学生在大学 4 年中笔语词汇的变化发展趋势。

一、研究的方法

(一) 测量指标的选择

根据 John Read 的观点, 笔语词汇的丰富度(lexical richness)主要体现在词汇多样性(lexical variation)、词汇复杂度(vocabulary sophistication)、词汇密度(lexical density)和错误数量(number of errors)四个方面^{[2](198-212)}。本研究采用的是没有经过标注的生语料, 所以没有把词汇密度作为测量指标; 另外, 对词汇准确性即词汇错误数量的研究不仅仅局限于词汇拼写错

误, 它涉及错误类型分类、词法、搭配等诸多方面, 由于文章篇幅有限, 词汇准确性不作为测量指标。因此, 本研究对英语专业学生大学四年笔语词汇变化发展的描述主要依据两个指标: 词汇多样性和词汇复杂度。

1. 词汇多样性

词汇多样性也叫形符比(type-token ratio), 是指语料中出现的类符与形符的比率, 形符(type)是指文章中所有不重复的单词, 类符(token)是指文章中所有单词, 包括重复使用过的^{[2](203)}。形符比是研究词汇丰富度的重要指标, 其最基本计算方法是将形符数除以类符数。Wolfe-Quintero 指出这种计算方法过于简单, 文章长短差异过大时得出的数据不科学^[3]。简单的形符比的确存在文章字数相差悬殊时所得出的数据随文章长短变化而变化的问题, 语言学家们在数学转换方面做出各种各样的尝试来校正文章长短不对形符比的影响, 但是 Richards 和 Malvern 认为所有的方法都存在问题^[4]。因此, 为了得到更有意义的形符比, Laufer 提出当文章长短相差悬殊时要对文本进行调整^[5]。本研究采用 Laufer 的方法, 即根据文本中最短文章字数在每篇文章中截取相应字数进行文本采集和计算, 其公式为: 词汇多样性=(形符数/类符数)×100。

2. 词汇复杂度

词汇复杂度也称词汇生僻度(rareness), 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

收稿日期: 2008-06-30

基金项目: 中南林业科技大学青年科学研究基金资助项目(项目编号: 08QY048)

作者简介: 刘文慧(1978-), 女, 辽宁锦州人, 中南林业科技大学语言与教育技术研究所讲师, 主要研究方向: 语料库语言学, 统计语言学, 二语习得。

标^[2]。Ellise 和 Barkguizen 提出过词频广度的概念^{[6](1354-137)}, 词频广度是指口笔语中不同词频等级词形的分布比例^[1]。词汇复杂度与词频广度实际上都是测量词频等级的指标。Laufer 和 Nation 提出词频概貌 (LFP, short for Lexical Frequency Profile) 作为研究词汇丰富度的测量指标, 词频概貌首先在词汇使用频率的基础将词汇由简到繁分成三或四个等级, 之后计算文章中出现的词族 (family) 在每个等级所占的比例^[7]。Laufer 发现 2000 词以外的单词能够更清晰地地区分学习者的词汇表达能力, 而且计算 2000 词以外的单词比例实际上就是计算词汇复杂度^[7], 因此本研究使用 Laufer 和 Nation 开发的 Range 软件中的第三个词表 (baseword list 3) 和 3 个词表以外的单词计算词汇复杂度, 使用的计算单位是词形而不是词族, 计算公式为: 词汇复杂度 = (文章中复杂词形数 / 总词形数) × 100。

(二) 语料的来源及其分析

本研究所用语料来自两个语料库: 中国学习者英语口笔语语料库 (SWECCL) 中的笔语语料库子库 (WECCL) 和兰卡斯特大学的 LANCAWAE 语料库 (Lancaster Corpus of Academic Written English)。

SWECCL 语料库是一个包括笔语和口语语料的大型学习者语料库, 由南京大学与外语教学与研究出版社联合开发。SWECCL 语料库库容为 200 万词, 分为笔语语料库 WECCL 和口语语料库 SECCL 两个子库。WECCL 子库的设计容量为 100 万词, 收集了从国内 9 所不同层次的高校英语专业 1—4 年级学生的英语作文。本研究采用的语料来自 WECCL 生语料库, 选用英语专业一至四年级的议论文、说明文和记叙文。每个年级每种文体依据作文长短随机选择 18-30 篇, 然后以字数最少的四年级记叙文为标准, 从 12 个语料集合中各随机抽取约 6200 字, 总字符数约为 74400 字。之后, 从 WECCL 子库中分别抽取 1、2 年级学生作文各约 5000 字组成约 10000 词的语料, 再抽取 3、4 年级作文各约 5000 字组成约 10000 词的语料。

LANCAWAE 语料库由兰卡斯特大学建立, 收集了一系列非英语本族语学生的作文, 同时还包括一部分本族语学生作文作为参照。该库在 21 世纪初建立, 并一直处于不断增加和完善之中。本研究抽取了早期双语学生 (early bilingual) 作文约 10000 词, 晚期双语学生 (late bilingual) 作文约 10000 词, 另外还有约 10000 词的本族语学生 (native speaker) 作文。

本研究采用 Laufer & Nation 开发的统计软件 Range32 对整理后的所有语料进行分析。之后用 SPSS 中的非参数检验比较英语专业 4 个年级在笔语词汇多样性和词汇复杂性的总体差异, 比较笔语词汇与口语

词汇发展变化, 以及英语专业高年级和低年级与国外非本族语学习者和母语学生之间的差异。

三、研究结果与讨论

(一) 我国英语专业学生大学四年笔语词汇多样性和复杂度的总体变化趋势

表 1 是英语专业 1 至 4 年级学生在记叙文、说明文和议论文中词汇多样性和词汇复杂度的统计数据。

表 1 我国英语专业学生不同年级笔语词汇多样性和复杂度数据表

年级	文体	词汇多样性 (type/token*100)		词汇复杂度 (sophisticated type/total type*100)	
		单项	平均数	单项	平均数
一年级	记叙文	20.19		15.14	
	说明文	17.64	18.65	20.27	18.73
	议论文	18.13		20.77	
二年级	记叙文	21.19		19.38	
	说明文	18.83	20.57	26.81	24.66
	议论文	21.68		27.79	
三年级	记叙文	24.41		23.11	
	说明文	22.97	22.39	31.03	26.97
	议论文	19.80		26.77	
四年级	记叙文	22.98		19.44	
	说明文	20.91	21.40	31.31	25.66
	议论文	20.31		26.24	
卡方检验结果		P = 0.973		P = 0.660	

从表 1 的第 4 列和第 6 列的平均数可以看出, 在 4 年时间里, 学生英语无论在词汇多样性还是复杂度上都呈现先升后降的趋势。用 SPSS 中的非参数卡方进行统计分析这两列平均数, 显示词汇多样性的显著性概率 $P=0.973>0.05$, 词汇复杂度的显著性概率 $P=0.660>0.05$, 两个数据均大于临界值 0.05, 说明词汇多样性和词汇复杂度在英语专业 4 个年级之间不存在显著性差异。

以上结果表明, 英语专业学生在 4 年的学习过程中笔语水平并不一定随着年级的增长不断提高, 而是一个先升后降的过程, 这种趋势与口语在 4 年都保持持续上升的趋势有所不同。相邻两个年级之间的比较数据表明, 1-2 年级和 2-3 年级两个阶段两个测量指标都是负值, 说明学生在 1-3 年级笔语水平呈上升趋势;

3-4 年级出现的正值表明到了 4 年级, 学生笔语水平下降。4 年期间英语笔语水平的变化虽然有波动, 但是上升和下降的幅度不是非常大, 其中的差别不存在统计学意义。

图 1 的(a)和(b)分别展示了 4 年中学生英语词汇多样性和复杂度的年度变化情况。图中词汇多样性在 1-3 年级呈上升趋势, 在 3-4 年级阶段呈下降趋势; 词汇复杂度在 1-2 年级快速上升, 2-3 年级上升的速度放缓, 在 3-4 年级阶段有所下降, 但下降的幅度小于词汇多样性。

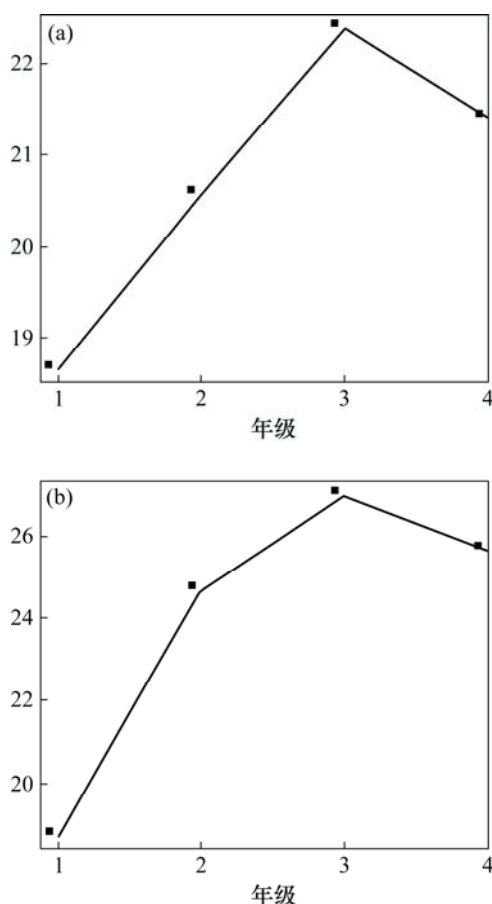


图 1 (a)多样性年度变化曲线; (b)复杂度年度变化曲线

图 1 中的曲线表明 1-2 年级是笔语提高最快的时期, 原因可能在于学生刚刚来到大学, 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, 还有来自周围同学的竞争压力, 在这些因素综合作用下, 学生取得较快进步。2-3 年级学生在专业四级考试的压力下继续努力学习, 但是可能由于教材不适合或教学方法跟不上, 同时学生已经适应大学相对轻松的学习氛围, 课外活动和娱乐时间增多, 进步相对前一阶段减缓。3-4 年级阶段, 学生忙于找工作或其他事务, 学习已经不是最主要的任务, 所以这

个阶段英语笔语水平不升反降。但是由于社会对口语比较看重, 同时口语不需要背诵大量新的难度大的单词, 所以 3-4 年级阶段口语仍然呈上升趋势。这与平时的教学经验也是一致的。

(二) 我国英语专业学生笔语词汇与口语词汇变化趋势的比较

表 2 是我国英语专业学生笔语与口语词汇的比较数据。口语词汇数据来自文秋芳 2006 年的《英语专业学生口语词汇变化的趋势与特点》一文^[1]。因为采用的计算方法不同, 两个词汇多样性的数据有大小差异, 但不影响对发展趋势的比较。前面已经提过, 词汇复杂度与词频广度都是测量词频等级的指标, 所以笔语词汇复杂度与口语词频广度具有可比性。

表 2 我国英语专业学生笔语与口语测量指标配对数据

	笔语词汇多样性	口语词汇多样性	笔语词汇复杂度	口语词频广度
1 年级	18.65	45.10	18.73	7.08
2 年级	20.57	47.80	24.66	7.26
3 年级	22.39	53.24	26.97	10.78
4 年级	21.40	54.68	25.66	11.13

图 2 是根据表 2 中数据在 SPSS 中形成的趋势变化图。

图(a)显示笔语多样性在 1-3 年级快速上升并在 3 年级达到顶点, 之后下降; 图(b)显示口语一直呈上升趋势, 2-3 年级上升速度最快。图(c)显示笔语词汇复杂度曲线在 1-2 年级快速上升, 2-3 年级上升速度减缓, 之后略有下降; 图(d)中口语词频广度曲线一直呈上升趋势, 2-3 年级上升速度非常快, 之后上升速度减缓。

图 2 显示口笔语词汇在大学 4 年的发展变化的轨迹不同。笔语词汇多样性和复杂度的发展轨迹都是呈先升后降, 1-3 年级是词汇多样性发展最快的时期, 而 1-2 年级是词汇复杂度发展的高峰期。由于笔语和口语的不同特点, 与笔语不同, 口语 1-4 年级在流利性、多样性和词频广度三个指标都呈上升趋势, 2-3 年级是进步最显著的时期^[1]。在教学中, 我们应该根据口笔语发展的不同特点有针对性地安排英语课程, 在不同时期有目的地增加或削减笔语或口语练习的强度。

(三) 我国英语学生与国外双语学生笔语词汇的比较

表 3 是我国英语专业学生与国外非本族语学生和

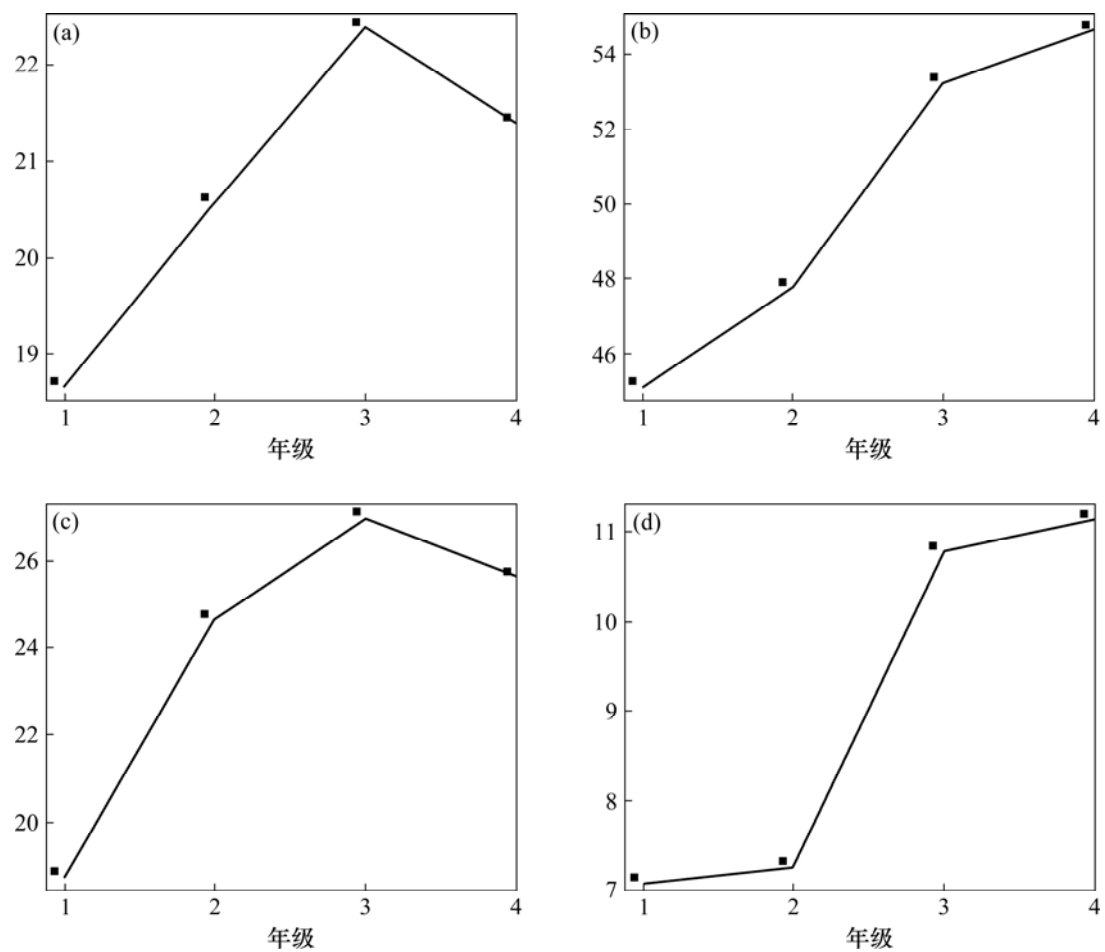


图 2 笔语与口语变化比较: (a) 笔语多样性; (b) 口语多样性; (c) 笔语复杂度; (d) 口语词频广度

表 3 我国学生与国外学生情况比较表

	词汇多样性	词汇复杂度
英语专业低年级	10.48	24.26
国外早期双语学生	14.64	32.13
英语专业高年级	12.87	28.97
国外晚期双语学生	16.57	36.47
本族语学生	17.84	34.81

本族语学生笔语词汇的测量指标数据。使用 SPSS 中非参数双样本检验方法对表中数据进行分组检验, 结果表明我国英语专业低年级学生(1、2 年级)与国外早期双语学生笔语词汇多样性和复杂度 P 值均为 0.000, 远远小于临界值 0.05, 两个样本存在显著性差异; 用同样的检验方法得出我国英语专业高年级学生(3、4 年级)在笔语词汇多样性和复杂度两个指标上与国外晚期双语学生和本族语学生也存在显著性差异。

通过观察可以看出, 在词汇多样性和复杂度两个

指标上, 我国英语专业低年级学生的笔语低于国外早期双语学生, 英语专业高年级学生同样低于与国外晚期双语学生和本族语学生。文秋芳检验英语专业学生口语指标时发现英语专业 4 年级口语词汇的流利性和多样性明显低于本族语大学生, 但词频广度与本族语大学生的差别无统计意义^[1]。我国学生在两个测量指标上均低于本族语学生是预料之中的。样本中的国外双语学生并不是英国本土学生, 他们是到英国留学接受语言进修的留学生, 他们的英语笔语水平明显高于我国学生。我国英语专业学生使用的教材中收集的都是英文原版文章, 词汇难度也比较大, 但是学生掌握的大量词汇和语法不能在写作中熟练运用。这个差距值得我们思考和积极寻求解决方法。

参考文献:

[1] 文秋芳. 英语专业学生口语词汇变化的趋势与特点[J]. 外语教学与研究, 2006, (3): 189-195.
[2] John Read. *Assessing Vocabulary*[M]. Cambridge: Cambridge

- University Press, 2000: 198–212.
- [3] K. Wolfe-quintero, S.Inagaki & H.Y.Kim. *Second Language Development in Writing: Measures of Fluency, accuracy & Complexity*[M]. Hawaii: Hawaii University Press, 1998.
- [4] B.J.Richards, D.D.Malvern. *Quantifying Lexical Diversity in the Study of Language development*[M]. Reading: Faculty of Education and Community studies, 1997.
- [5] B.Laufer. *The Development of L2 Lexis in the Expression of the Advanced Learner*[J]. *Modern Language Journal*, 1991, 75: 440–448.
- [6] E.Ellis, G. Barkhuizen. *Analyzing Learner Language* [M]. Oxford: OUP, 2005.
- [7] B.Laufer, Pole Nation. *Vocabulary Size and Use: Lexical Richness in L2 Written Production*[J]. *Applied Linguistics*, 1995, 19: 255–271.

Lexical developing features in written English of English majors

LIU Wenhui, YANG Zhiqing

(Language and Teaching Technology Institute, Central South University of Forestry and Technology,
Changsha 410006, China)

Abstract: Taking WECCL sub-corpus of SWECCCL corpus as samples, and LANCAWAE corpus as reference, this paper analyzes the written English developing trend of English majors in China. Lexical variation and vocabulary sophistication are taken as measure indexes. Results show that the trend of written English does not go up continually, which is different from spoken English, and the less frequency can be found in Chinese students than that of foreign bilingual and native-spoken students in both indexes.

Key words: corpus; written English; lexical variation; vocabulary sophistication

[编辑: 颜关明]