基于语料库的中国英语学习者口笔语中词性分布比例研究

陈建生,刘晓翠

(天津科技大学 外国语学院,天津 300222)

[摘 要]基于英语学习者和英语本族语者语料库的研究,采用中介语对比分析的方法,对比分析了中国英语学习者与本族语者口笔语中词性比例的差异。研究结果表明,中国英语学习者口笔语中的词性比例与本族语者相比整体分布趋势一致,但有些词性的比例与本族语者差异较大,且口语、笔语语体特征不明显。

[关键词]词性比例;语体特征;中介语对比分析

[中图分类号]G426 [文献标识码]A [文章编号]1673-9477(2010)04-0090-04

Palmer(1936)认为语言总体来说有三项功能:它表示说话者的思想、感情等;它影响说话者的行为,即引起有效的功能;它也是一个符号体系,把所指的"事物"符号化。从语言的功能可以看出,语言存在口语和书面语两种形式,而从语言的发展来看,任何一种语言都是先有口头交流,后有文字。口语和书面语由于各自使用的场合和使用目的的不同,经过长时间的演变,逐步形成了较为明显的差异。中国英语学习者经过多年的英语学习,积累了一定的词汇量,语法也比较熟悉,但在用口语交流思想或探讨问题的时候,往往会出现一些沟通上的障碍,或者写出的文章总显得不够地道。这其中的原因之一就是英语学习者缺乏对英语口语和书面语差异的了解,混淆了两者的语体特征。本文将使用中介语对比分析的方法,通过对本族语者口语语料库和学习者口语语料库的观察,探讨学习者的口语和书面语中的词性分布比例与本族语者口笔语中的词性分布比例是否存在差异,以了解学习者的语体特征。

一、文献回顾

语料库是指按照一定的语言学原则,运用随机抽样方法,收集自然出现的连续的语言运用文本或话语片段而建成的具有一定容量的大型电子文库(杨惠中,2002:33)。这种运用计算机软件和大量真实语言资料来对语言学理论、结构以及应用进行研究已经成为当前语言研究的一大趋势。Granger把这种将学习者语料库中的语言特征与本族语者语料库中的语言特征进行的对比分析称为"中介语对比分析"(Contrastive Interlanguage Analysis)(Granger,1998)。中介语对比分析法是研究学习者语料的常用方法,如:Ringbom(1998)通过对母语不同的欧洲英语学习者语料进行研究后发现,学习者有过度使用某一类高频词的现象;Aijmer(2002)也基于瑞典英语专业学生语料分析得出学习者有过度使用常见情态动词的情况;马广惠(2002)通过对中美大学生英语作文的语言特征对比后发现,中国学生的作文具有较强的信息性和正式性,而美国学生的作文表现出更强的口语体修辞特征。目前针对英语学习者语料的研究多集中在书面语语料上,少数口语语体特征研究也多是定性研究,语料数量少,且多用于举例佐证。随着现代英语口语语料库的建立和检索与统计软件的问世,对大量语料的定量分析已经成为可能。例如:桂诗春、杨惠中(2003)研究得出学习者书面表达具有一些口语的特征,犹如"写话";文秋芳等(2003)从读者、作者显现度和不同词频等级的词的分布情况这两个角度探讨了中国英语学习者书面语中的确有较强的口语化倾向;文秋芳等(2004)在研究我国英语专业学生频率副词的使用特点后发现学习者口语中过多使用带有书面语特征的频率副词,书面语中也存在过多使用带有口语特征的频率副词的现象,认为这从一个侧面说明中国学习者使用的是混合语体。文秋芳(2006)在研究我国英语专业学生使用口语笔语词汇差异时曾对学习者和本族语者口语及书面语中词性比例做了简单的分析,但其选用的样本偏小,且口语语料来源为学生计时独白,笔语语料为命题作文,均不是典型的口语和笔语。本文将使用《中国学生英语口笔语语料库2.0》为研究对象,并选择口语子库中的"对话"部分语料来突出口语的典型性;对比语料库将使用美国当代英语语料库中的口语和学术文章部分,其词汇量大且语料来源广泛,提高了研究的准确度。

二、数据的采集与分析

(一)语料库

在学习者语料库的选择上,应尽量使用由中等以上水平的中国英语学习者在学习英语过程中的书面语和口语语料所构成的学习者语料库,因为它能够提供较为全面的中国英语学习者中介语发展的有关信息。本研究所使用的学习者语料库是由文秋芳等编辑出版的《中国学生英语口笔语语料库2.0》(SWECCL 2.0)。其中的口语子库部分,其全部语料来源于2003-2007年间的全国英语专业英语四级考试口试部分(TEM4-Oral)。为了突出口语交际中"即兴"的特点,本文只选择了四级口试中第三项对话部分的语料,总词数约为36万(详见表1)。笔语子库是由全国不同区域,不同类型、等级的高校学生作文组成,总词数约120万。

对比语料库部分使用的是美国当代英语语料库(Corpus of Contemporary American English,简称COCA)中的口语和学术文章两部分。这个语料库是由Brigham Young University的Mark Davies教授开发的美国最新当代英语语料库,数据来源为1990-2010年美国境内多个领域的语料,该语料库是开放性的语料库,其规模还在不断扩大,目前的规模已经超过4亿词。

[收稿日期]2010-10-26

[作者简介]陈建生(1953-),男,上海人,教授,研究方向:语料库语言学。

其中口语语料的来源为多个谈话性电视节目或广播节目中的对话转写,总词数将近 9 千万。学术文章语料涵盖了多个学术领域,总词数约为 8 千万。

本文使用的学习者语料库采用 Lancaster 大学设计的 CLAWS 词性赋码器将所有文本自动赋码。

表 1　学习者语料库与本族语者语料库容量

名称	中国学生英语口笔语语料库 2.0(SWECCL 2.0)		美国当代英语语料库(COCA)	
分类	口语(SECCL-3)[2]	笔语(SWCCL)	口语(Spoken)	笔语(Academic)
容量	361,169	1,248,476	87,116,763	82,914,544

(二) 词性分布比例

词性分布比例是指样本中不同词性的形符占总形符的比例,是考察学习者语体特征的一个指标。Biber 等人(1999) 从实词和虚词两个类别方向对本族语者词性分布比例研究后得出各类词性在不同语体中出现的频数均有显著的差异。如:名词在新闻中的频数最高,学术文章居中,口语中最低;形容词在学术文章中频数最高,口语中最低;实义动词和副词在口语和小说中的频数高于新闻和学术文章;代词在口语中频数很高,但在新闻和学术文章中很低;介词和限定词在新闻和学术文章中频数很高,但在口语中很低;情态动词和基本助动词均在口语中更常见;相对于其它虚词,连词的语体特征不明显。为方便同 Biber 等人的结论相比较,本文亦分别从实词和虚词两个词性类别对本族语者以及学习者语料库中名词、实义动词 (如 say、run 等)、形容词、副词、代词、介词、连词、限定词、情态动词和基本助动词(be、have、do) 的分布比例进行对比研究。

(三) 数据分析

1. 各类词性的频数统计

通过在线对本族语者语料库和使用语料库检索工具 WordSmith 对学习者语料库进行检索并统计了本族语者和学习者口语和笔语语料库中各类词性分布的相关数据,如下表所示:

表 2　本族语者口笔语中词性分布比例统计数据

		Spoken		Academic		x^2
词类	词性	频数	PM	频数	PM	
实词	名词	14,619,348	167,813	21,640,430	260,997	p<0.01 差异均显著
	实义动词	9,816,209	112,679	7,893,850	95,205	
	形容词	4,681,429	53,737	7,990,993	96,376	
	副词	6,424,479	73,746	3,847,104	46,398	
	介词	7,869,794	90,336	10,603,669	127,887	
	代词	8,112,459	93,122	2,010,030	24,242	
	情态动词	1,317,321	15,121	859,041	10,361	
虚词	限定词	9,716,683	111,536	9,597,908	115,757	
	基本助动词	6,961,025	79,905	3,584,443	43,231	
	连词	5,319,167	61,058	5,017,913	60,519	p>0.05 差异不显著

表 3　学习者口笔语中词性分布比例统计数据

		SECCL-3		SWCCL		x^2
词类	词性	频数	PM	频数	PM	
实词	名词	54,853	158,800	290,433	232,630	p<0.01 差异均显著
	实义动词	41,933	121,397	149,175	119,486	
	形容词	18,035	52,212	89,192	71,441	
	副词	24,457	70,803	95,882	76,799	
	介词	24,688	71,472	109,334	87,574	
	代词	48,701	140,990	96,363	77,185	
	情态动词	12,422	35,962	40,069	32,094	
虚词	限定词	35,540	102,889	145,121	116,239	
	基本助动词	1,680	4,864	11,268	9,025	
	连词	22,976	66,516	79,876	63,979	p>0.05 差异不显著

从表 2 和表 3 中可以看出,各类词性在本族语者口语和笔语中出现的次数经卡方检验均有显著差异。如名词、形容词、介词及限定词在笔语中所占比例较高,而动词类、代词、副词和连词则在口语中出现比例较高,相比较于其他词性,连词的语体特征差异度最低,这都与 Biber 等人的研究结果基本一致。而学习者方面除连词外,其它词性比例也均有显著的差异,这点与本族语者是基本相同的。图 1 是本族语者与学习者口笔语中各类实词出现频数的标准化(每百万词中出现的频数) 比较,我们可以较为直观的看出学习者与本族语者口笔语中词性比例差异情况。

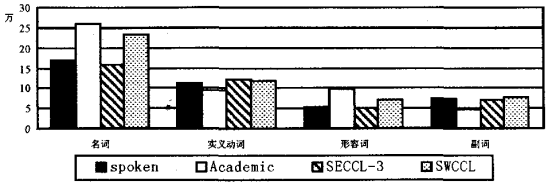


图 1　本族语者和学习者口笔语中各类实词标准化频数比较

从图 1 中我们可以看出,除副词外,学习者口笔语中各类实词的分布比例与本族语者整体趋势基本相同。名词在笔语中的频数均高于口语,可见笔语更侧重于信息的传递。实义动词在学习者口语中的频数虽然高于笔语,但本族语者相比,其频数差异较小,且笔语中的实义动词频数偏高。Biber 等人认为口语中子句普遍简短且数量较多,每一个子句至少需要一个动词,所以实义动词更多出现在口语中。而学习者笔语中高比例的实义动词从某种程度上反应出学习者书面语造句较为繁冗、逻辑性较差,如:

As long as we can look it in our hearts. Almost everyone is different. But someone have their points about it. Let me introduce my points about it. First, …(SWCCL)

例句中的四个句子存在较明显的逻辑联系,可简化的表述为"Everyone may have different points of view on this and mine is firstly…"

另外,Biber 等人提出名词和实义动词的比值也可反应出信息密度和短语、子句的复杂程度。本族语者口语和笔语中名词与实义动词频数比值分别为 1.49 和 2.47,而学习者为 1.31 和 1.95。可见,学习者笔语中句子的复杂程度较本族语者偏低。

形容词主要是用于修饰名词。Biber 等人认为名词比例较高的语体,其形容词的比例也会较高。从图 1 中可以看出,本族语者笔语中名词的频数显著高于其口语,笔语中形容词的频数也的确比其实义动词和副词的频数高。而在学习者笔语中,名词的频数虽然也高于其口语,但形容词的频数是相对最低的,这与本族语者完全不同。此外,学习者笔语中副词的使用频数要高于其口语,且显著高于本族语者笔语中频数。这一方面与学习者笔语中实义动词频数较高有关,因副词多用于修饰动词;另一方面在对学习者笔语语料库检索后发现,学生集中使用常见副词 so、just、only、very 以及 always 等。据统计,这五种副词的频数比例占总副词数的 16.5%,而在本族语者笔语中仅为 7.7%。可见,学生对副词的习得不充分,过于依赖常见副词。

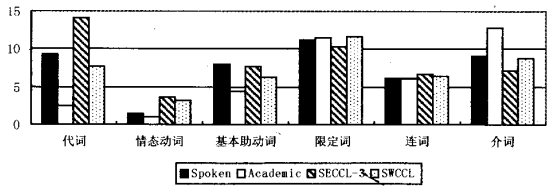


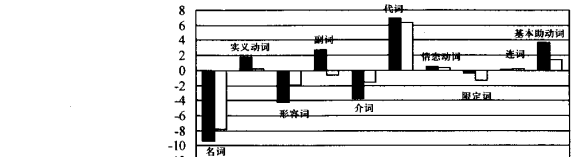
图2 本族语者和学习者口笔语中各类虚词标准化频数比较(单位:万)

图2是本族语者与学习者口笔语中各类虚词出现频数的标准化比较。Biber等人认为,虚词的分布是与实词分布紧密相关的,如口语中较高的代词密度弥补了名词频数偏低的现象;作为名词的延伸和限定成分,介词和限定词也更常出现在名词密度较高的笔语中。学习者口笔语中代词的频数均显著高于本族语者。且统计后得出,学生口语和笔语中出现第一、二人称代词的比例分别为总代词数的82.3%和57.8%;而本族语者分别为56.8%和32.4%。可见,学生无论在会话还是写作中都凸显较强的主观性。

此外,学生口笔语中的情态动词频数也显著高于本族语者。其中在学习者口笔语中频数最高的情态动词均为can、will和should。对本族语者语料库检索后发现,其口笔语中最常出现的情态动词为would、can和will。would和will被Biber等人归类为表示决断或预测的情态动词,而would相对于will表达更为委婉。学习者口语中较多出现的should表示责任和必要,有较强的主观意愿,本族语者较少使用,尤其是在其口语中。由此可见,学习者口笔语表达中忽略了礼貌原则,用词较直接,而本族语者则更倾向于委婉的表达方式。

2. 口笔语中词性比例差异统计

为了深入比较,我们又分别统计了本族语者和学习者口笔语中不同词性所占比例的差值。如图3所示,除副词外,学习者口笔语中各类词性所占比例差值的整体趋势与本族语者大致相同,如名词在本族语者口笔语中所占比例的差值为-9.32,在学习者口笔语中的差值是-7.38,两个差值均为负数,即名词在口语中的比例要低于笔语。进一步比较后可以看出,虽然词性比例差值的整体趋势相同,但学习者各类词性比例差的绝对值与本族语者相比多数较小,尤其体现在形容词、副词、介词和基本助动词上,本族语者这五类词的绝对差值分别为4.26、2.73、3.76和3.67,而学习者的差值仅为1.92、0.60、1.61及1.39。可见,中国英语学习者口笔语特征区分不明确,在某些词性的运用上,口语有些像笔语,笔语又有些像口语。



	名词	实义动词	形容词	副词	介词	代词	情态动词	限定词	连词	基本助动词
■本族语者口笔语中各词性百分比差值	-9.32	1.75	-4.26	2.73	-3.76	6.89	0.48	-0.42	0.05	3.67
□学习者口笔语中各词性百分比差值	-7.83	0.19	-1.92	-0.6	-1.61	6.38	0.39	-1.33	0.25	1.39

图3 本族语者和学习者口笔语词性比例差异对比

三、结语

本文基于中国英语学习者语料库和本族语者语料库,从词性分布比例的角度,探讨英语学习者的语体特征。研究表明,除连词外,各类词性在中国英语学习者口语和笔语中都有显著的差异,这与本族语者基本相同。但与本族语者相比,学习者对各类词性有多度使用和使用不足的现象。如在笔语中过度使用实义动词、副词等,使得句子构成略显简单;而对代词以及情态动词的过度体现出学习者的口笔语表达过于主观、直接。产生这些现象的原因一方面是学习者对某些词性的习得不够充分,过于集中使用其中的某几个词;另一方面,受到母语的影响,或不熟悉西方文化背景而形成了不地道的表达。

另外,各类词性在学习者口笔语中的差值与本族语者相比偏小,这说明学习者在口语和笔语的词汇使用上没有形成显著差异,有混合两者语体特征的现象。从一方面证实了文秋芳(2004)提出的假设:学习者语域是个特殊的语域。口语不完全像本族语者的口语,笔语也不完全像本族语者的笔语;口语中带有笔语特征,笔语中带有口语特征。尽管出于不同场合和不同交流目的,口语和笔语所特有的结构会出现重叠的现象,但从本族语者语料库中的词性分布来看,本族语者口笔语中词汇使用的差异仍然是显著的,了解口语和笔语的特征与差异,有助于学习者针对不同的场合、目的以及不同的交流对象,选择更恰当的表达方式。

[参考文献]

[1] Aijmer, K. Modality in advanced Swedish learners' written interlanguage[A]. In Granger, S.. Hung, J. & Petch - Tyson (eds.). Computer learner Corpora, Second Language Acquisitions and Foreign Language Teaching[C]. Amsterdan: John Benjamin Publishing company, 2002.

[2] Biber, D., Johansson, S., Leech, G., Conrad, S. & Finegan, E. Longman Grammar of Spoken and Written English[M]. Beijing: Foreign Language Teaching and Research Press, 1999.

[3] Granger, S. The computer learner corpus: A versatile new source of data for SLA research [A]. In Granger, S. (ed.). Learner English on Compute[C]. New York: Longman, 1998.

[4] Ringbom,H. Vocabulary frequencies in advanced learner English[A]. In Granger, S. (ed.). Learner English on Computer[C]. New York: Longman, 1998.

[5] 桂诗春,杨惠中. 中国学习者英语语料库[M]. 上海:上海外语教育出版社,2002.

[6] 李荣等译. 语言学概论[M]. 北京:商务印书馆,1984.

[7] 马广惠. 中美大学生英语作文的语言特征分析[J]. 外语教学与研究,2002,(5):345.

[8] 文秋芳,丁言仁,王文宇. 中国大学生英语书面语中的口语化倾向—高水平英语学习者语料对比分析[J]. 外语教学与研究,2003,(4):268.

[9] 文秋芳,丁言仁. 中国英语专业学生使用频率副词的特点[J]. 现代外语,2004,(2):150.

[10] 文秋芳. 英语专业学生使用口语 - 笔语词汇的差异[J]. 外语与外语教学,2006,(7):9.

[11] 文秋芳,梁茂成,晏小琴. 中国学生英语口笔语语料库[M]. 北京:外语教学与研究出版社,2008.

[12] 卫乃兴. 中国学习者英语口语语料库初始研究[J]. 现代外语,2004,27(2):140.

[13] 卫乃兴. 中国学生英语口语的短语学特征研究 - COLSEC语料库的词块证据分析[J]. 现代外语,2007,30(3):280.

[14] 杨惠中. 语料库语言学导论[M]. 上海:上海外语教育出版社,2002.

[责任编辑:王云江]

(下转第109页)

上的指导并提出初步评价意见。针对论文写作中语法错误较多、用词不当等问题，指导教师在初稿阶段没有必要逐字逐句检查，否则形成了学生的依赖心理；在论文中期和后期的修改过程中，在确保论文结构、章节的衔接和逻辑性没有问题的条件下，鼓励组内的学生自我修改论文中的语法错误、拼写错误、用词不当等问题，同时，教师作相应的指导；最后定稿之前，指导教师再作整体上的把关。总之，教师在严格把关学生论文质量的同时，还要加强和培养学生的自主学习能力。

第二，注重学术能力培养。在论文的指导过程中，教师应对如何使用文献资料、如何引用原文作者的资料、引文标示和文献参考部分应作明确的讲解和严格的要求。教师应强调论文的学术性和严谨性的重要性，培养学生严谨的研究态度。

第三，认真作好选题工作。论文选题必须具有一定的学术研究价值和意义，能充分反映作者的专业水平、写作能力和独到的见解。可以从一下几个方面入手，其一，教师把毕业论文教学环节提前可起到事半功倍的效果，对三年级学生，以专业课程（如文学、翻译、语言学等课程）论文的形式培养他们的兴趣点，鼓励学生多阅读，多思考，这样不至于在论文写作时不知如何选题的问题。其二，学生选题要慎重，要选自己熟悉并感兴趣的、相关的选题。学生在选题时应该听取指导教师的意见，把自己的观点和想法与老师交流、沟通，但不能过分依赖老师的意见，否则写作时缺乏自身的观点、逻辑思路，以至于影响论文的质量。此外，教师在指导学生选题时，也要做一些先期调查，论证其可行性，不能只凭自己的经验与理论推测，否则学生选题不成熟，研究路线不清晰，造成中途换题或无法正常完成论文工作。

第四，提高论文的深度和创新性。如没有前期资料的积累和思考，学生的论文很难有一定的深度和创新。可从早期入手，如鼓励二、三年级学生参与教师的科研课题或申报校级学生科技创新课题，辅助教师从事基本的研究工作；对于三年级的专业课程，授课教师应对学生进行学术研究的入门指导，介绍论文写作的基本方法，资料的收集，要求学生写课程论文，从主干课程中提高学生的创新能力。有些同学对创新有畏难心理，总觉得不知如何下手。其实创新可以是观点的创新，也可以是材料、角度、方法的创新。不同观点的比较、不同角度的切入、不同方法的运用、不同学科的交叉等等都可以使自己的论文与众不同。此外，指导教师的作用不容忽视，对学生不能过分放手，要引导和启发学生，还应扮演好“导演”的角色，指导学生对所读文献进行分析和思考，找出有无薄弱点或值得进一步研究的地方，以及给予相应的研究方法的指导。

四、结语

毕业论文写作工作的完成需要教师和学生的共同努力和配合。一方面，在平时的教学中，教师要加强对学生基础写作能力的训练，扩充其各方面的知识面，注重其学术能力的培养，培养其独立思考精神、创新思维与实践能力。教师在指导的过程中，从宏观上对学生论文进行指导，启发学生深入思考，引导学生新视角、新方法的创新，同时发挥检查监督的作用。另一方面，学生必须具备扎实的基础写作技能，在毕业论文的写作中尽量避免出现语法错误、用词不当、时态等等错误；同时，学生也必须掌握必要的论文写作知识，在论文写作中发挥自主学习、研究学习的能力，形成严谨的研究学习氛围，出色的完成毕业论文写作任务。

[参考文献]

[1] 戴炜栋. 对我国英语专业本科教学的反思[J]. 外语界，2007(4)：23.
[2] 穆诗雄. 英语专业毕业论文写作[M]. 北京：外语教学与研究出版社，2002.
[3] 张春芳. 近五年英语专业毕业论文质量调查与思考[J]. 重庆交通大学学报，2009(4)：65.

[责任编辑：王云江]

The problems in english major thesis writing and its strategies

DING Yan - wen
(Hua Zhong Agricultural University, Wuhan 430070, China)

Abstract: English major thesisi writing is one of the most important tasks for every undergraduate before graduation. This paper presents some problems with the students' thesis writing and its possible causes. Then it proposes several ways of improving students' thesis writing for the English majors.
Key words: thesis; problems; strategies

（上接第92页）

A corpus - based study of distribution of parts of speech in Chinese learners' spoken and written English

CHEN Jian - sheng, LIU Xiao - cui
(Tianjin University of Science and Technology, Tianjin 300222, China)

Abstract: This study attempts to compare the two corpora of the native speakers and the Chinese learners of English to find and analyze the differences of distributions of parts of speech in their spoken and written English productions. The results show that the main trend of the distribution of some parts of speech in learners' spoken and written English is similar to the native speaker's. But there still exist differences and the register feature of spoken and written English of learners' is not distinct.
Key words: distribution of parts of speech; register features; contrastive interlanguage analysis (CIA)