《中国学术期刊网络出版总库》及CNKI系列数据库入选期刊

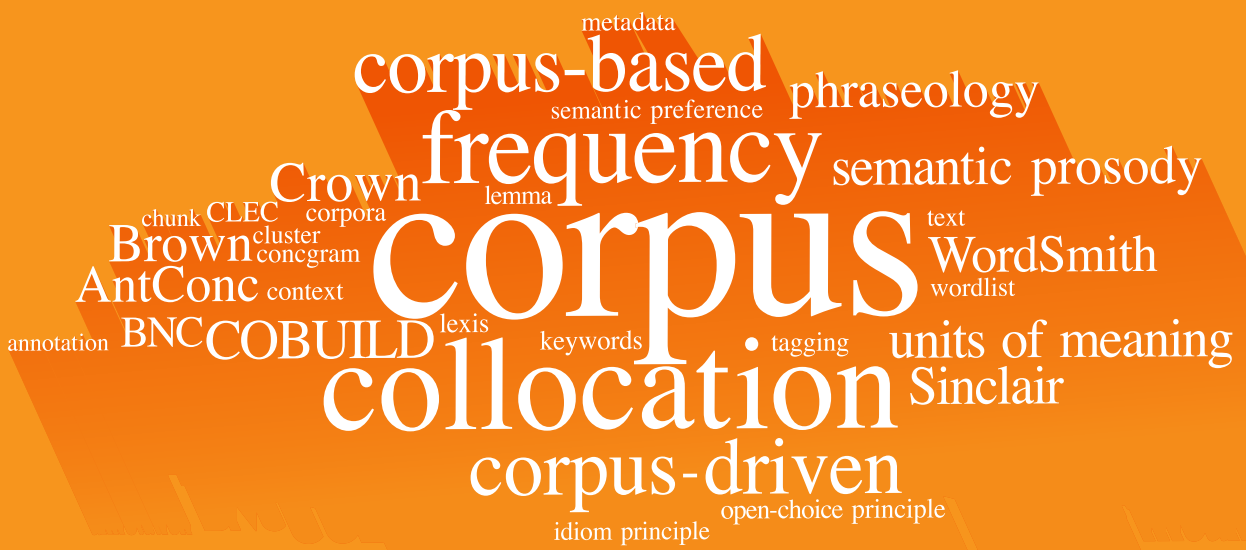
二〇一七 第四卷 第二期

语料库语言学

CORPUS LINGUISTICS

2
Vol. 4 No.2
第4卷 第2期
2017

北京外国语大学中国外语与教育研究中心
中国英汉语比较研究会语料库语言学专业委员会
梁茂成 主编



外语教学与研究出版社
FOREIGN LANGUAGE TEACHING AND RESEARCH PRESS

语料库语言学 外研社

语料库语言学
（半年刊）

Corpus Linguistics
(Biannual)

主　管：中华人民共和国教育部
主　办：北京外国语大学
承　办：中国外语与教育研究中心
　　中国英汉语比较研究会语料库语言学专业委员会
出　版：外语教学与研究出版社

Administered by the Ministry of Education of China
Directed by Beijing Foreign Studies University
Edited at the National Research Centre for Foreign Language Education and Corpus Linguistics Society of China, China Association for Comparative Studies of English and Chinese
Published by Foreign Language Teaching and Research Press

主　编：梁茂成
编　校：解碧琰、李晓雨

Editors: Liang Maocheng
Proofreaders: Xie Biyan and Li Xiaoyu

电　　话：（010）88818687
电子邮箱：bfsucrg@sina.com
投稿网址：http://ylyy.chinajournal.net.cn

本刊地址：北京市西三环北路19号北京外国语大学中国外语与教育研究中心
　　《语料库语言学》编辑部（100089）

《语料库语言学》
2017年　第4卷　第2期

目　录

学者聚焦

研究论文（栏目主编：北京航空航天大学　卫乃兴教授）

研制开发

CORPUS LINGUISTICS

Volume 4, Number 2, 2017

Table of Contents

叩鸣录：杨惠中先生答客问

编者按：

本采访由许家金于2013年策划并列出采访提纲，李文中基于该提纲进行了修改，并负责采访及录音，吴进善、张绵、刘守智将录音转写成文本。在本稿转写并修订过程中，我们尽量保留原话，有意不删除谈话中的口语特征，以体现先生在学术上思想犀利、寓深奥学理于平实言语之中、娓娓道来、返璞归真之神韵；结语部分选用了近两年杨惠中老师与编者在个人通信中对相关问题的深入讨论。本稿标题为编者所加，题意出自《礼记·学记》第十八“善待问者，如撞钟，叩之以小者则小鸣，叩之以大者则大鸣；待其从容，然后尽其声”。实际上，先生待学者殷切率真，学生每有请益，辄小叩大鸣，神思如瀑，逸兴湍飞，其发聩灌顶，酣畅淋漓之处，可使人得意而忘言也。本稿最后经杨惠中先生亲自审定。

问：20世纪80年代初，您就开始尝试把计算机应用到语言和教学研究中，主要着眼点有语料库、自然语言处理、CALL、EST和测试。我们就从80年代初开始谈起，根据文献，当时您开始使用计算机对文本进行处理时，最初的动机是什么，主要受哪些思潮影响？

答：其实最初的想法很朴素，因为80年代初，大概1982年开始，教育部决定制定教学大纲，教学大纲里面很重要的一个问题就是教哪些词汇，教哪些语言现象。当时语言交际、功能这些实验已经开始有了，我也很感兴趣。上海交通大学是参与这个工作的，我也在最初的研究小组里面。因为看到一些有关词频、词表概念的报道。我印象比较深的是Edward Thorndike的《教师一万词手册》、《教师三万词手册》，还有Michael West的《常用词三千》，觉得很有意思。大学四年，英语学习一共才有240多个学时，我们用的那个教材也旧，当时学生高中毕业的时候全国平均水平是大概掌握1,800个词，那么这个有什么问题呢？一共才有240左右的学时，学生进入大学的时候才学1,800个词，那么每个小时如果要学生掌握10个词的话，到大学毕业也只能掌握2,400个词。两个加在一起也不到5,000个词，这些词够用吗？当时我们也进行过大规模的社会调查，编制了一个需求分析调查问卷。在此之前，改革开放以后的教学大纲，只要求学生在1分钟内读17个英语单词，大约需要掌握5,000个词汇量，而这个能力，全国也只有三分之一的大学毕业生能达到，那么在这样的情况下，怎样提高效率？究竟掌握多少词才能够

阅读？掌握哪些词？一小时可以教他多少词？这些都是非常现实的问题。当时在讨论词表的时候，我举了下面这个例子，比如priest这个词要不要？有老师说要，也有的说不要，意见不一致。John Sinclair讲得非常有道理。他说每个人的阅历、经验、爱好、读过的书不一样，不可能取得一致意见。这时我们就需要用证据来讲话。所以当时我们读了上述几本书，知道了Service Wordlist的概念，就是哪些词是学英语首先要掌握的。当时也看了一些中文资料，发现汉语也有常用字，因为词没法区分，所以有常用字的概念。比如到底要掌握多少常用字就可以读《人民日报》了？我当时也看了一些报道，说如果中国的字表制定得好的话，小孩子到小学三年级就可以看小说了，因为他半猜半看。所以我们觉得，确定常用词表，可能是我们很重要的一步。那么怎么确定呢？我们非常不熟悉。因为我看到的资料里面，Thorndike是抄卡片的。当时计算机已经开始有了，上海交通大学最大的一个计算机是王安机，放在新建楼的顶上，占了整个阳台的一半，其实它的功能还远不如现在的一台微型机。当时我们自己用Basic编程序，实际上也是抄卡片，然后叫打字员录入。他打进去的这个文档就被切分了，因为词很容易定义，两个空格之间连续的字母数字串就是一个词。打进去this is a desk四个词，如果再碰到this这个词，this频次加一次。那么怎么校验呢？就是完了以后有些非词，我们就把它去掉。一边打，一边统计。我想打到100万个词就差不多了，因为Thorndike初步的工作就是100万词。就是这样一直干。干了大概三四个月，磁鼓坏了，存的东西全不见了，当时也没有备份。那个时候磁盘还不常见，容量很大的软盘也很少。所以那天很痛苦。我跟黄（人杰）老师俩人白辛苦了半年，已经积累了很多量了。后来就忽然有一个想法：当时的程序是输进去以后就切分，文本没了，为何不把文本放在那里？然后我需要的时候把文本调出来再切分。文本不是也可以拿来用的吗？那个时候实际上就建立了一个文本库。差不多那个时候知道了Brown 语料库。我们发现Brown语料库做的就是这个工作。后来看了Brown的一些资料，发现原来建库还要有原则的。于是我们制定了语料库建库的原则。当时我们考虑为文本标记编制一条很长的码。我跟黄人杰一起商量，后来你都知道，这个码里面有包含文本来源、体裁等信息。对于科技英语而言，我们当时定了大概10个领域。我记得好像还包括船舶工程等，但是没有生物。当时科学技术的发展是提到日程上面来了，重点大学都是理工类的。学科选择必须是上海交通大学也要有的，而且国内用户要广。因此，我们没有考虑文科、医学等领域。文本是理工类大学使用的教学大纲，是为这个目的服务的。每一个连续的文章的选取原则是什么？我记得当时定了几个原则：每个文本不少于500个词；基本上没有图表跟公式；应当是随机地在一本书里去选；书必须是英美出版的；作者必须是母语者；还要考虑出版的年份、体裁等。后来我们规定，前言要选，评语也要选，大概有这几个类型。选材一定要保证随机性，因为当时已经对语料库也有一些概念了。建语料库，第一，必须要有representativeness（代表性）；第二，必须

是random sampling（随机抽样），如果它是varied sample（多类型样本）的话，可以代表整个语言使用的characteristic features（典型特征）；第三，必须要有一定的容量，如果语料库太小了，总共只有1万个词，是没有说服力的。当时看见的一些报道，像Brown语料库，LOB语料库都是100万词。之所以这样，有各方面的原因：第一，当时的计算机能力就是那么大；第二，有人力花费，当时没有机器可读文本，都要打字。如果你这个采样设计的好，它还是有价值的。至少对于描写的有些方面是有价值的。另外，它是个structural corpus（结构性语料库），它的结论是有说服力的。而且当时也看了些报道，说LOB语料库就是Leech跟Svartvik他们跑到美国去，向Brown学了以后，完全按照它的做。

LOB语料库完全是Brown语料库的对应库，是Brown语料库的美国英语语域的一个英国英语对应语域。那么我们想同样搞一个EST counterpart（科技英语对应库）。它必须是原始的，这样才有价值，而且能满足科技英语教学的目的。频率、覆盖率、分布率这些概念当时全都有了。假定根据频率词表选定的5,000个词（types）能够覆盖任何文本99%的内容，对教学来讲的话是非常有效的。当时就是这样的想法，形成了关于frequency（频率）、distribution（分布）的一些概念。而且我们把distribution（分布）分成text或者passage distribution（文本或者文章分布）、specialty distribution（专业分布）这样一些概念。所谓text或passage distribution，就是这篇有的，那篇有吗？specialty distribution（专业分布）是这个领域里面有，另一个领域有吗？区分这样一些概念以后，还有coverage（覆盖率）的概念。我们做了大量统计分析，这样建起来的语料库就是个structural corpus（结构性语料库）。是有意识建设的，不光是为了数字，还可以做很多其他的事情。这个时候我已经主动地开始作语料库语言学的研究了。

问：当时为什么会想到建一个EST语料库？

答：当时为什么要搞EST呢？因为EST在那个时候也是个重要的体裁，我记得在参加一些学术会议的时候，EST这个概念是有很多学者怀疑的。English就是English，哪里有什么EST呢？实际上不同领域里面的专家讲话都有一套术语，不在一个领域就听不懂。那么对EST来讲，是非常需要的，不是可有可无的。当时很多教材也是这样，比如说Henry G. Widdowson就搞了Scientific English的一套教材，里面有很多想法都是很新的。我记得他当时提出了nucleus（核心）的概念，所以我们有一套教材就叫《核心英语》。他说核心这个概念，语言是学不完的，它是一个不可穷尽的开放集，所以要学核心的东西。那么核心的东西是什么呢？词汇有核心的东西，那就采用词表。交际功能也有核心的，交际功能哪一些是必须要用的？我们要通过调查研究把这些核心的东西教给学生。所以理解这些

概念非常重要。差不多在那一段时间，我听了很多英国语言学家讲communicative approach（交际方式）的讲座，有一些在上海，有一些在北京。我记得Henry G. Widdowson、Christopher Candlin、John Sinclair、Tim Johns等很多学者都来了，有些是我们请来的。所以在那个时候，一些概念已经基本上都形成了。比如必须通过调查研究确定学生的交际需要，根据学生的交际需要来确定他需要哪些交际功能，他需要掌握哪些词，哪些常用的结构，哪些微型技巧，后来我们制定的教学大纲里面都有四个附表。这四个附表是常用词表，常用功能意念表，常用语法结构表和微技能表。这些是通过调查研究出来的，不是拍脑袋想出来的。

问：当时是如何基于语料库确定词表的？

答：当然我们做了大量校对工作，后来动员很多人去校对，所以它的可靠性还是比较高的。查出来以后得到的词表里面没有priest这个词的，再讨论的时候就没有问题了。你说priest这个词学生学了以后，在100万词里面1次也没出现，那么学来干嘛？这就是cost-effectiveness（成本效率）嘛。所以在讨论词表的时候，比较容易取得一致的意见。我记得我们讨论词表的时候，譬如说，我确定在1小时内学生可以掌握10个词。1小时就是1个教学时，1节课掌握10个词。你给学生240个学时，他能够掌握2,400词，再加上假定中学里掌握的2,000个词，那么他毕业时可以达到4,400个词的词汇量。那如果我们要教学生3,000个词的话，学生每一节的记忆任务就不是10个词，可能要十三四个词。如果这是不可能的话，那么你这个目标就是不现实的。根据这个原则，我们选词汇时是非常严格地，一个词一个词来看。有一些词在第1个1,000词的时候没有问题，第2个1,000词时可能有一些意见分歧，第3个1,000词时可能意见分歧就很大了。我们作调查研究，把这个词表发给100多个有经验的教师，由他们根据教学经验决定要还是不要。这也算是一种调查方法吧。

问：那就是先从语料库统计出分级词表，然后再去征求意见对吗？

答：频率越高的词，大家越没有意见。频率降低一点，最低就是1次词了，once word。当词的频率比较低的时候，不同的人就有不同的意见了。有的人说要，有的人说不要。我们当时做了两个工作。第一，我大概征求了100多个老师的意见，要或者不要，随便你，这是主观选择的，根据个人经验判断，回来以后我们进行量化分析，这个词在100个人里面，99个人认为要的或者99个人认为不要的，都是量化的结果。因为有经验的老师的直觉也是很重要的，不是可有可无

的。我们搞测试里面很重要的一点就是，你对学生能力的判断准确吗？你听听老师的意见嘛。就我个人的经验，随便哪个班级，我上课一个礼拜，我就知道谁好，谁差，然后我不断调整，调整到一个月以后，我很有把握，这个班里面谁最好，谁最差。如果这个老师不负责任，上完课就走，那你不要去找他，你找有经验的老师来判断，而且是他一个学期以后的判断，我们的小组以前都是这样做的。我要找负责任的老师，有教学经验的老师，这个是学校推荐的。我不要多，100个人就够了，然后不要学生的分数，只排队，名次由高到低排下来，而且你一个月排一次，就是第一个月排，第二个月你再调整，到学期末肯定准确了，然后把我考试的结果跟他的排名对比，在我的排名表中的第一名，考试时是不是第一名，如果他排的前几名在我的测试中排在后面，那只能是我不对，不能是他不对，这个是相关研究。相关研究里面，他排的是第三四名，我的结果变为第四三名，它已经要删去了，但实际上如果要分好中差的话，还是够相关的。所以这样是非常严格的，我们相关系数有0.7，非常高的相关度。这是三跟四和四跟三这种差异都已经要删去了，还有0.7，如果你再分好中差，那就接近于1。我刚才举的是考试的例子，现在我是讲有经验教师的判断是有价值的。那么我现在找有经验的老师判断词汇的取舍，这就是quantitative analysis（定量分析）了。

问：当时是否要求教师建议还有哪些词需要加进来？

答：还有增加的。如果他认为要增加的词是必需的，语料库里肯定有这个词。定量分析作完以后，我们又加了定性分析。我们发现，语料库100万词的库容太小，所以在有些研究词汇学的书里，还有一种讲法，叫availability，那么availability怎么翻译呢？这也是很难的，但是它是什么意思呢？你教了他星期一，教了他星期二，英文不像中文的，中文就星期一、星期二，你一二三四加下来就完了。英文每一个都要学的，Monday、Tuesday、Wednesday一个一个学的，而你六个都有了，Thursday没有，这个可能的吧？那么你说availability纠正了一下，这个词要不要进去啊。这个就让我想到Chomsky讲的skewness（偏态）。为什么呢？因为你随便选什么材料，Sunday可能很容易出现，Friday可能也很容易出现的，有Black Friday对不对？但说不定其中有一天很难出现，所以我常常举这个例子：如果你问Where are you from？他说I’m from New York，这个概念很大，因为纽约人多，你说I’m from High Wycon，我听也没听说过High Wycon！我有一位英国朋友住在英国的一个山里面，叫High Wycon，那个地方很漂亮的，上面的一个村，这个村总共也就200人，所以你说整个语料库里面Where are you from？正好碰到这两个人的概率是很低很低的。所以他说any corpus must be skewed，我觉得不是没有道理的，哪怕现在你有Web as corpus，也很难找到一个I’m from High

Wycon的。那么在这个过程中，我们又得到了启发，就是frequency（频率）跟probability（概率）是两个概念。Frequency是一个词在这个文本里面出现几次，在另外一个文本里面不一定就是那么多，当然几次是绝对的，你讲这个文本里面出现3.5%，换一个文本不一定3.5%，可能3.3%，可能3.7%。所以frequency是在变化的。但是当文本足够大，不是100万词，是1,000万词，是1亿词的时候，这个频率不变了，这个时候就是probability。趋于稳定的频率就是概率。概率是语言的特征之一，所以后来我们就把概率作为我们选词的一个标准，语料库必须要足够大。然后定性分析怎么办呢？我定了几条原则，比如政治的原则，首先是社会学的。其实这就是社会准则，在中国教英语，没有“socialism”这个词不行吧？没有communism这个词不行吧？但是我查的书都是英美原版，所以社会准则又肯定是需要的。

问：当时选取语料时是怎么保证随机抽样的？

答：其实我们当时找材料，首先是一定要严格保证它的随机性。我们确定了10个领域，然后我跑到上海交通大学图书馆，图书馆里面的外文书跟中文书是分开放的，我每十本书取出来一本，一看第十本书是俄文的，因为俄文的什么都放在一起的，或者说法文的，这些都不要，那么第十一本，如果是第十一本的话，以后就全部是第十一本。因为它是按图书分类法分类的。如果说机械的，我从这个架上去找，找第十本，一看英国或美国出版的，作者是讲母语的，这本书就拿来了。如果它是法语的，那么就是第十一本，以后就第二十一本，第三十一本，这样选。每次从图书馆抱了一大摞书回来，回来以后呢，我们发现，这一本书从第10页开始，那么第十页，第六十页，第一百二十页，就这样找，必须要完整的段落，这个段落里边公式很多，不要，要跳开这个，这样保证它的随机性，而且保证它是以文本为主的，不要公式什么的，不然你怎么分析呢？因为我们的目的是查词嘛。这个过程非常困难，当时没有可机读的文本，全是手工打字输进去的，是键盘输入，所以当时打字员我觉得有两三个人，还要找很多青年教师来校对，花很大工夫。后来有人反映了一个情况，引起我的一些思考，他说他学了一段时间以后，看书里面专业的这一部分困难不大了，但是前言部分看不懂。我想这是什么道理？原来这也是科技英语的特点。因为后面的正文部分是对外部世界的描述，符合科技英语的一些特点，比如现在时、被动语态、客观陈述等比较多。前言有很多虚拟语气，否则是不谦虚的。他明明是这样想的，还讲得弯弯绕，学生看不懂。所以后来我们决定，把前言也作为一个体裁，也选一部分。后来又有人反映一个问题，就是看教科书比较容易懂了，但是看杂志上的东西还有困难。可能有时新的信息，这是一个因素吧，还有杂志上的文章发表的时候，也是

跟前言一样的。教科书的内容都是已经公认的事实了，里面虚拟语气比较少。是这样就是这样，不必客气的。杂志上是比较新的东西，作者还没有把握。一个是时新的信息，一个是新术语等等，还有一些语气上的影响因素。所以我们后来决定，杂志也必须选一部分，还有文献综述也要选一部分，要保证有各种体裁，有一定代表性。那么这样选完了以后呢，已经100万词了，那么教学的时候呢，刚才讲到教学的定性准则里面，我记得提了三条原则：一个是社会准则，一个是刚才已经讲了，就这个例子，还有一个是语言教学准则。为什么？因为上课你要用到paraphrase这个词吧？但是paraphrase这个词在这些原版的科技书里面是不可能有的。它怎么可能paraphrase呢？它只可能有definition，不可能有paraphrase。但是我上课是要用到paraphrase这个词，也就是上课用到的一些教学的术语，这个你要加进去的。我们当时也提倡用英语授课，你词表中连这个词也没有，你怎么讲得起来？譬如说idiom这个词重要吧？ expression这个词重要吧？ expression这个科技英语倒可能用到的，它可能频率不高，但是你上课总是要这些词吧？第三个就是语言准则，这个包括哪些呢？比如词的派生能力以及搭配能力。我记得当时还有暗指的能力、联系的能力以及availability，大概有这几条。在判断这个词要不要的时候，我们就根据这些标准来决定取舍。那么在决定词表的时候呢，有一条原则就是只选根词，不要派生词，除非这个派生词学生不学不行。我记得我们举了一个很典型的例子是describe，describe这个词在词表里算一条，但是described不算，describable不算，describability也不算。虽然有人攻击我们说日本人学3万词、4万词、6万词什么的，那都是瞎攻击，因为我们根本就是不算派生词的，我们是按根词计算的，我一个根词describe很容易变成5个嘛，那5,000个词就成2.5万个了。

问：请谈谈您对ESP的看法？

答：我有两个地方观念上有些改变，因为我们采访了一次Christopher Candlin，他是主张ESP的，他提出个观点，说ESP很重要，ESP、EST是怎么发展起来的？就是交际方式，因为原来搞的英语教材，如基础英语，是不分专业的，到交际方式出来以后，首先要关注交际需求。我是医生，你做生意，他搞造船，我们的需求是不一样的。你要教我有用的东西，否则我学了半天，用不上。这个时候就出现了ESP的概念，那么ESP里面有人又分了EST、EOP、EAP等等，都是在这个理论里面发展起来的。我觉得Christopher Candlin对ESP的发展、以及ESP对交际方式的发展，对register（语域）概念的发展，对语言使用和语言教学的发展都是有很大的贡献的，这是ESP的贡献。但是最后他提出很重要的一点，就是ESP的研究与教学、教材编写要防止过分的fragmentation（碎

片化），如果分得过细的话，课就没法上了。比如说医学，医学里面有内科跟外科，内科里面还要分心内科、消化科；讲到心脏科，还有心脏内科跟心脏外科了，还有心血管内科跟心血管外科，那么细都是它的交际需求吧？分到最后上课只有一个人，教材编了只有三个人买，你编不起来了，出版也出版不起来了，所以他认为要防止碎片化，而防止碎片化的方法是什么呢？是common core（共同核心），这个就是我们搞核心英语的出发点，所以后来再看雅思考试，我们发现直到今天它都只分三大类，一个是science and technology（科学技术），一个是medicine and biology（制药和生物），还有一个是management and social sciences（管理和社科），是为了专门目的使用语言，就是把语言作为工具来教与学所要使用的三个大类。

问：其实这句话到现在还是非常有用的，就是两种思潮，一种是宏大的，一统性的描述或者分析，就像以前说English for General Purposes，就是笼统的，一般的用途。具体用途语言其实把这个东西给颠覆了，更着重具体要求，学习者或者交际者的个别要求，但是走了极端就是碎片化了，这是因为任何一个人的交际都有自己具体的交际需求，过分碎片化了可能也不行，中间要取得一种平衡。

答：对。那么在这里我又有一个想法，这些观点后来都逐步形成了，也比较成熟了，我觉得学语言必须要抽象。那么抽象出什么东西呢？一抽象就出现了语法，出现了句法，这个是一种抽象。我教你一条语法，你不会在一个地方用完。我说过去时加ed，你可以用很多地方。当然不规则动词是另外一种情况，你还得学一条。但是学了一个规则你就可以用，这就是一个抽象。抽象概括是语言教学必不可少的。但是概括是有问题的，因为它脱离了语言的使用。那么第二种办法，我概括语言的使用，使用是在一定的语境里面。一个语境就是一个语境，我能不能概括到语境呢？有一套教材叫linguaphone（灵格风英语），就是按语境编排的对话，如at the station，in the barbershop，以及at the post office。这种方法是很好的。但是，它有个问题，你在邮局里面，你一定讲这个情景使用的词吗？比如你可能碰到一个朋友，请他去吃饭，看戏。所以在邮局不一定讲邮局的事情。这是第一个困难。你应该把语境列一个穷尽性的单子出来吧？你列不出来的。你列不出来就是偶然性的例句了，不完全的。另外我在这个语境里我可以讲别的事情。所以这个是有问题的。那么比较好的概括是什么呢？比较好的概括是交际功能，它不是邀请怎么表示，因为那会有很多语言形式。那么他不是往往地表示拒绝吗？拒绝怎么表示，有很多方式了。肯定的可以表示否定的意义，陈述句也可以表示命令。这些发现很是有意义的。那么还有一种抽象是pattern（型式）。所以我还有

pattern drills（型式训练）这么一个理念，就是语言当中的句子是一个无限的集合，但是句型是一个有限的集合，你说I am a student，那么he is a student，这是两个句子，但这是一个句型。这样的话，如果我能够把句型，甚至是型式概括出来的话，我就可以用有限的句型。就是说语言是个无限的集合，句型是个有限的集合。英语900句，我就可以用有限的句型来描写无限的句子，这个效率肯定高。实际上这个型式也是很难定义的，这是另外一个事情。用这个办法来讲，我也统计了下，像《新概念英语》，实际上就500来个型式，你掌握了一个，你的交流灵活性就大一点，你掌握了十个，就更大一点，掌握了一百个更大，你如果掌握了四五百个，那就差不多了，可以常用了。这个也是一种抽象，抽象程度最高的实际上是这个型式的概念应当包括collocation（搭配）和colligation（类连接），所以我认为型式训练还是一个很好的理念，是有价值的。实际上，提出型式训练概念的人是Charles Carpenter Fries, 他是密歇根大学的英语语言学院的第一任主任，第二任主任是Robert Lado，第三任主任是Sillyca，第四任主任是John Swales。后来大概有几个，现在好像是Larsen-Freeman，是她后来提出grammaring（语法技能）的概念，实际上我认为就是型式训练，你不是讲听说读写吗？都把它做成细化技巧——听说读写，其实语法也应当叫语法技能，那么这个就跳到脑袋里去了。这些概念都是很有用的，但是从语言哲学的角度来讲，上升到最高层面的应当是交际功能。

问：这就是说，把语料库的概念或方法运用到语言分析最初的设想是着眼于语言学习和语言教学，或者说是怎么学好英语。再就是受到当时ESP运动以及交际语言教学思潮的影响，当时是如何确定科技英语词汇的呢？

答：我们考虑怎么把这些概念归并到一起，后来黄人杰老师搞了一个公式，这个公式是经验公式，其中有三个要素，第一个是选词指标，第二个是它的频率，第三个是它的分布，有文本分布，还有专业分布，再加上常数。这三个要素怎么调整是根据经验公式。我们发现，某些词在所有文章里都有，而且每一个专业领域都有分布，有些词只出现在若干个领域里面，有的词只出现在一个领域里的几个文本里面，所以当然选那个在各个专业里都有的，所以专业的权重要大，只在某个文本里面的词权重小，根本不出现的等于零，是不会进来的，这样这个词表弄了以后就有现在这个形式的根据，那是有价值的。后来在这个基础上我们又发现了一个规律，这个规律是三类词，纯粹根据统计突出，一类是function words（功能词汇），一类是technical terms（技术词汇），一类词是sub-technical terms（准技术词汇）。这个观念就在这个基础上形成了。所谓技术词汇，它的特点是在一个领域内频率非常高，跨了一个领域后根本不出现的。所以当这个词频

率高得不得了，但文本分布或者叫文章分布，跟专业分布低得不得了的时候，它是技术词汇。那么另外一个极端就是功能词，如 as、of、not、and所有领域都有，非常稳定，它分布得高，频率也高。功能词是一个封闭词集，就那么多。处在二者之间的就是准技术词汇。这些词的特征是什么呢？频率相当高，但是没有功能词汇和技术词汇高，分布又比技术词汇高得多，但没有功能词汇那么高，那么这些就是要教的词。从这个里面我又得到一个结论，技术词汇不是外语教学的任务，其他两个就是外语教学的任务。比如autism就是孤独症，没有旁的解释了。这是在学专业的时候必须学的，但不是教语言的任务。第一，他学这个专业，没有这些词是学不下去的；第二，他如果不知道这个词可以去查字典；第三，这些词很容易记，文章里讲了十遍十五遍，他怎么还能记不住呢？但sub-technical words（准技术词汇）不是这样。比如look学会了，look at学会了，look after学会了吗？ look forward to学会了吗？这个都要一条一条学的，这个才是语言教学的内容。

问：前面谈的是大致思路。当时您是如何确定从JDEST中提取多词技术词语和准技术词汇的？当年您曾经在*Literary Computing*上发表了一篇论文，报道了这些结果。

答：我当时觉得一个词在一个领域频次高，跨了一个领域根本不出现的肯定是术语。这是指单个词，那么对于多词序列，就会产生一个文本切分的问题。所以我当时设计了一个算法，我记得我当时做了electromagnetic force这个研究，如果是两个词以上，我就把技术词汇分为两个，一个是mono-word，single-word terms（单词术语），一个是multi-word terms（多词术语）。当时理念是什么呢？因为有一个term bank（术语库）的问题，我们国家术语译名也是要统一的，这是术语译名统一委员会做的首要工作，要找到术语啊，但术语是在不断地生长，不断地产出。譬如说，拿计算机来讲，本来有hardware这个词，hardware本来不是硬件的意思，是五金，五金店叫hardware嘛。后来，有计算机了以后，它变成硬件的意思了，那时也是一个新词啊，它是词义的变化，产生新的意义嘛。那么它作为术语，作为计算机术语，现在没有五金的意思啊，它就是硬件的意思，它是个术语。在这个基础上，后来出现更重要的是software, software大概是六七十年代以后才出现的，后来发现我们搞教学还有courseware，搞语言翻译还有linguaware，这些词都出现了，后来我还看见一个叫personware，搞计算机不能光做硬件，应当人机一体嘛，要把人这个因素考虑进去才能发挥它的最大的作用嘛。这里讲到的很多问题，对我们国家来讲都是基础性的。

问：到现在我们在讨论ESP的时候，实际上它是有一个区分的，第一种是普通一般的学术性词汇，再往上一层是就是真正深入到专业领域的词汇。

答：专业领域的术语不是我们的教学内容。后来我在英国发表了相关的文章，有不少人写信来说要拿这个搞实用系统。我当时设想，术语是不断增长的，不管你出现linguaware也好，还是personware也好，我有观测的工具，我把每年出现的语料从头捋一遍，就可以把新词找出来。找出来之后，术语库就更新了。所以你怎么验证、怎么定义是另外一个事情了。当然这个不是一次词，它作为术语，频率是非常高的，它有特征了，当一个术语库不断更新时，就可以记录下来。

问：当年您在*Literary Computing*期刊上发文，在结语部分，您有一个预测，说这种提取技术词和准技术词的方法，也可以用来讨论或者是计算一个文本的aboutness（所言之事）和它的topic（主题性），请您谈谈当初的想法。

答：当初的想法就是提取。我下面讲的就是这一点，很多术语是多词的，举个例子吧，electromagnetic force（电磁力），它要作为术语，必然频率很高；第二，这个序列必须出现，你不能说force of electromagnetism，它必须是electromagnetic force这个形式，词序不能变，内部结构也不能变，所以这个怎么提取呢？当时我想了个办法，实际上这个切分啊，我可以给它一些信号，好像叫作stop sign（停止信号）。第一，一个连续的术语，一个连续的多词术语不可能跨越句子界限，不能electromagnetic，force，不可能的吧？所以，我这个切分就是停止信号；第二，它也不可能跨越冠词，一部分在冠词前，一部分在冠词后也不可能；第三，它不可能跨越一个介词，除了of，因为of可能是术语的一部分。还有其他几条。这样一来的话，这个文本做预处理的时候已经被切成一段一段了。把功能词全去掉，留下来的很容易找到多词结构，由这个办法就找到了很多多词的技术术语。像cathode-ray tube display unit就是个术语，至于cathode-ray tube是不是也是个术语，这个也可以讨论。后来更有意思的是，我做了个实验，九个领域都是不同的科学技术的专业，第十个是文学，当时文学在Brown语料库、COBUILD里面都有，COBUILD语料库里面1,000万个词是有关科学技术的，结果多词术语，或者在一个领域非常高在另外一个领域没有的，在文学文本里根本就没有找到，只找到了一个post office。后来我又根据这个想法搞了个自动提取，只提取一两百个词，我把最常用的术语放在第一个，它也可能讨论什么放在第二、第三个。现在很多专业人员，如果一篇文章长得不得了，来不及看，就看摘要好了。当然，这是很粗糙的，现在自动提取除了这个，还要加上很多语义的，所以分析的方法就准确多

了，但是这个思想我觉得是在技术上面不断完善了的。

问：我觉得当时有两点特别了不起，一个是超越频数的概念，不单是看它在整个语料库里边的频数有多少，频率有多少，还要看它的分布，就是一个横向视角，看它的宽度。它可能就是一些词在某一个领域高度的聚集，但是在其他领域不出现或很少出现，这是一种现象，然后利用这种现象开发出一种算法，设计出一种算法来把这些东西提取出来。我想确认一下当年的一个情况，就是您去伯明翰大学，是受A. S. Hornby Educational Trust Scholarships（霍恩比教育基金奖学金）资助的。在那个时候，上海交通大学的科技英语语料库是已经完成了还是正在筹备？

答：语料库已经基本上建成了。我选择到伯明翰大学去，计划是搞语料库研究，再就是对教学指标的制定等等。我之所以能去，当时有一个人很关键，叫Alan Maley。Alan Maley是应用语言学家，是当时英国的English Language Officer（英语语言官员），他在中国大概做了半年。他对中国英语教学的贡献极大。后来他实际提供给我的是A. S. Hornby Educational Trust Scholarships（霍恩比教育基金奖学金）。但它有一个规定，那就是苏联和中国的学者不得享受。为什么呢？因为Hornby是第二次世界大战以后日本和英国政府交换的首位战俘。他本来是印度English Language Institute（英语语言机构）负责的，后来被日本请去了，他对日本英语教学的贡献也很大。第二次世界大战以后他被日本人俘虏了。后来在第二次世界大战末期的时候，英国跟日本提出交换战俘，他是英国人提出名单上的第一名，他就是宝贝呀。是Hornby提出sentence pattern（句型），VP的概念。我第一次接触sentence pattern这个概念，就是看他的书。看的是什么呢？看的是俄文版的，印象非常深刻。他提出句型，比如assume这个词。第一个义项可以用于VP1 · VP5 VP7 VP8 · · 这个就是动词型式。我觉得非常有道理。动词后面可以什么也不跟，是一个型式，后面可以跟个及物动词，可以跟个及物动词或者间接宾语，都是不同的型式。他回到英国以后成立了一个霍恩比教育基金奖学金。他赚了很多钱，都是版税。所以他说这个要取之于民，用之于民，全世界的学者跟学生都可以申请这个奖学金，到英国去深造。但是苏联和中国除外，因为当时两国还没有完善的版权保护法规，他没有从我们这里赚到钱。Alan Maley特别给我申请了一个，作为特例。所以我说觉得非常荣幸。大概1982年底1983年初的时候，我们正要筹建外语系。我正好到北京出差，想去听听许国璋先生的意见。我当时提出的想法呢，就是我们不走文学道路，我们要根据上海交通大学的三个特点，社会需要、个人产出和环境可能。我当时也不知道应用语言学这个术语。我就搞语料库，搞EST。他非常赞成，给我写了封推荐信。可能这个东西到British

Council（英国文化教育协会）那里也起作用了。因为我也看了Sinclair的一些书，也知道他在搞语料库研究。我也想看看他们在研究什么。而且他们给我的条件是很优惠的。我们那时叫洋插队，所以在那里还好，后来跟John Sinclair、Tim Johns都很熟。

问：当时他们在做COBUILD语料库，你怎么看Sinclair所做的工作？

答：我认为你要赶快去挖掘John Sinclair的基本思想。这些思想要发展，这是未来。我认为他比一般的语言学家至少领先十年。很可惜他去世了。

问：Michael Hoey、Mike Scott以及Susan Hunston说过同样的话，Sinclair的思想非常超前。

答：我谈谈我的想法，最好你们把它继续下去。他对于语言本质的观测，比旁人早得多，很多人现在还不知道欣赏他的思想，这是很遗憾的。我要给你重点讲几个概念。一个是我刚才跟你讨论过的，我认为Chomsky讲的competence（语言能力）是不可观测的东西。而我们关心的是performance（语言运用），是可以直接观察到的东西。你同意吧？

问：我是觉得看是怎么理解乔姆斯基的那个观点。就是说如果是不认可他的二元划分，不把语言分成语言能力和语言运用的话，纯粹用语料库语言学视角的话，没有语言能力，只有语言运用，是吧？

答：我们现在不讨论这个问题。我现在讲观测。语言能力不好观察，对吧？那么语料库是东西，具体在哪里。我直接观察嘛！对吧？ John Sinclair不同意这个观点。我先把这一点讲一讲。他不同意，他说语料库提供的是不能直接观察到的东西。为什么不能直接观察到？我举了很多Chomsky的例子，他后来回信了，他说把他的陈述修正一下：首先不谈Chomsky，Chomsky根本不关心观不观察得到的问题，所以可以把他排除；第二，他讲的不能直接观察到的是超越人类的五种观感。请你想想，深入地想一想这个问题。五种感官是什么？是视觉、听觉、嗅觉、触觉、味觉。那么超越人类的五种观感的是什么呢？我想了想就是语感。而是不直接的。一个东西你可以触摸，闻味道，眼见为实看一看。语感是什么？如果没有语料库的东西，语感你要观察也观察不到的。语料库给了你了，你才有可能是非直接的观察。但是看见的又不是语感本身吧？非常深刻啊！他举了一个例

子：一个英语不是母语的参议员在国会讲了Let's forget all this nitty-gritty。nitty-gritty就是细枝末节了。我们不要管这些细枝末节嘛，我们管大事嘛！这样理解是不对的。nitty-gritty是正面的意思，不能反面用的，这个人想得很周到，连nitty-gritty也想到了。那么我不把这个东西拿出来，你的语感能起作用吗？不能起作用的。那么我拿给你了，但还看不到，因为我没有这个语感。所以究竟语料库语言学提供了什么？再反过来想语言是什么？这些问题都值得我们深入思考。我认为你们年轻一代要在这个思路上，赶快去动脑筋，去挖掘，去理解。还有关于full sentence definition（整句释义法），这也是我现在要跟你讨论的，我下面就讲这个问题吧。

问：这是一个很创新的做法，就是在Collins Cobuild Dictionary中应用的。

答：他有很多哲理呀！很多哲学思想在后面啊！你们在这个思路上去挖掘。我现在很多想法呢。当时到了Birmingham以后，很有启发。John Sinclair当时搞那个语料库，他跟我讲好像是花费一百万英镑，那个时候的一百万英镑相当于现在的一千万了，他说这是整个英国历史上文科里面最大的唯一一个科研项目，从来没有过这么大的，而且一般的人都持怀疑态度。当时它有一台主机heliword，还有一台中型机在办公室里面，还有一台小型机，比现在这个微型机还要大一点，搞计算机的人是Jeremy Clear。Jeremy Clear是他们英语系的学生，但是这个人聪明得不得了，我非常佩服。他是英语系的，自学编程，Tim Johns也是自己编程，用汇编语言Assembly Language编的，他自己又研究应用语言学，又学了编程，所以编的东西都很到位。他当时跟我讲，为什么美国的计算机辅助语言教学（computer-assisted language learning）都是多项选择题，因为它都是搞计算机的人编的，他只会这样做，他也不知道究竟语言教学有什么语用方法，有哪些变量。Tim Johns自己编了一个程序，就是关于交际的。例如讲borrowing，你不能上来就讲Can you lend me什么什么，这是不行的，这个取决于你是谁，对方是谁，在什么场合下，借什么，对吧？如果在排队等公共汽车，你正好少了五分钱，5 pence对吧，那你还要想个理由出来吧，又不认识人家。如果是兄弟，Have you got 5 pence?就完了，如果是不认识人家，要讲一大段话的，而且人家可能借给你，也可能不借给你。但是如果你在银行去贷款，5,000英镑，Have you got 5,000 pounds?。这个是不行的，贷不到钱的。所以他就编了个程序，把这些都编写进去了，他的理念是什么呢，就是concordances（索引行）了，我记得那个时候它已经达到两千万的词，两千万个词，每个词啊，包括the，都打索引行出来，统统打出来以后呢，一大堆东西怎么办？当时使用的是microfiche。这样一张A4的上面可以容纳120页了。他把它们放在一个房间里的书架上面，在一个柜子里，一盒一

盒的。然后把它们打印出来。他专门租了一个三层的花园洋房，里面大概有二十几个人，每个人都是语言学家。然后把所有这些都打印出来，你负责A，他负责B，然后一看，assume这个词占好几页，这个义项是第一次碰到，用黄的笔画一画，第二个义项没有碰到过，绿的笔画一画，第三个跟第一个义项是一个，黄的笔画一画，这个工程量很大。然后从这个里面归纳出来：assume第一个义项出现频率最高，作为第一项，然后例句是什么呢？这就是证据。所以他一定强调证据，没有任何一点是拍脑袋的，有一次他有强调一点：说到自然性，说你自己拍脑袋想的例句、语法都对，就是缺少一点自然的感觉。他觉得这个很重要。自然性包括哪些要素？指的是什么？肯定是符合语感的，搭配，类连接都没问题的，否则的话就是桂老讲的虽然有可能性，但却不大可能发生。这样的一种就是缺少自然性的，他没有从这个角度来定义自然性究竟是什么？大量的工作使我非常感动，我觉得这个东西是等于为corpus研究树立了一个榜样，我们当初叫corpus-based lexicography（基于语料库的词汇学）。他呢，后来他说实际上有人，就是他后来的太太嘛，把它定义叫语料库驱动的词汇学，那么现在区别问题在哪里呢？问题在我们国内现在很多也是用计算机编字典，它也叫基于语料库，实际上这是计算机辅助词汇学，他所有的例句都不是从语料库里面来的。他说，我要一个例句，这个例句进了库了，以后我查的时候也很方便，但是那不过是计算机辅助词汇学，跟语料库是没有关系的，他建立的这个只能叫作database（数据库），不是语料库。语料库驱动就是一切要从当时实际使用的语言的证据来决定，我认为他这个是创造了一个先例。后来呢，Jeremy Clear在我离开伯明翰大学之后大概二三年，很快就被Oxford University Press请去了，后来又被美国的IBM请去了。如今很多的大型出版社出版一个词典，如果没有语料库或原始语料库，是不行的，不能东抄西抄的，所以它成了一个榜样了。Jeremy Clear本人也很有意思了，这个小先生很聪明，后来他说他也想去念博士，我跟他开玩笑说“那你肯定是读人工智能、计算机、自然语言处理这样的专业了，”他说“No！一天到晚搞计算机太枯燥了”，我说“你想学什么呢？”。“古代英语”。所以我觉得他这就是生活了，不是为了谋生啊！研究要和兴趣结合起来，对吧，他搞计算机也不是为了谋生，他有兴趣嘛！我觉得教育搞到这个程度就是活了。

问：当年您在伯明翰大学时，有哪些印象较深的经历吗？

答：我讲几个我印象比较深刻的东西啊。有一个是印象比较深刻的：有一段时间，每个星期二下午是学术活动时间，John Sinclair在那个大洋房下面的一个大房间里面搞学术研讨会，很有意思，那里的教师一个顶一个的，没有滥竽充数，你教语言学，你必须本身是语言学家，要拿到这个职位是很不容易的。我记得大

概就是Michael Hoey和搞文本分析的好几个人。

问：应该有Malcolm Coulthard。

答：有Malcolm Coulthard。好多人都来的，来了以后他叫Jeremy Clear做了个程序，这个程序是空白的，“你们大家讲下面一个是什么词？”，那就只能瞎猜了，因为它连第一个词也没有，对吧！随便你猜什么词，必须记录下来。然后“啪”出来第一个词，“all”，好！那么你猜all下面是哪一个词，你的根据是什么？你说它可能是个限定词，那么这个下面可能是名词，为什么呢？因为限定词下面肯定是名词，对吧，那么有人猜，它可能是个代词。那么你猜的根据是什么，是我的句法知识，你就预测嘛，猜完记录下来了，第二个词出来“of”，那么有人猜对了，有人没猜对嘛，对吧，你说“all students”，那么出来了“students”你猜对了，“all”你猜错了吧，那么你讲为什么猜错了，我认为它是个代词，我猜对了，“all of ”嘛，那么“o f ”下面都对了，肯定是个名词，但是不一定啊，也可能是形容词，那么我猜两个，可能是名字，可能是形容词，根据是什么你记录下来，“energy”是名词，但是我猜不出来这个词，那么“energy”出来了以后，下面就可以预测很多了，因为我的discourse（语篇）、text（文本）、context（上下文）、situation（情景）、genre（体裁）、register（语域）出来了，你猜下面是什么？我猜呢，下面很可能是一个动词，对吧！为什么呢？因为肯定是“is produced，is derived，is consumed”，总归是这样的词了，随便你猜，不可能是“sing”、“song”、“dance”、“ballet”，这个预测的依据是什么呢？是词汇预测和语义预测，结构出来，“is a…”都蛮好的，那么is出来以后呢，我马上可以肯定后面是个分词，因为这是句法决定的嘛，那么这个词也是有限的啦，“is produced，is derived，is consumed”总归是这样的词了，那么你说词汇预测，有几个人猜对了，有几个人猜不对，原因都记录下来，因为他们都是语言学家啊，好，“derived”出来没有问题，后面肯定是“from”，对吧，“from”后面什么呢，大概是“All forms of energy is derived from the sun. ”那么这个用到的什么呢，是world knowledge（世界知识），因为所有的能量都来自太阳的嘛，煤也好，石油也好，都是来自太阳嘛。好，再接下来是什么呢，又猜不到了，我告诉你，看吧！再下面呢，“either”什么，“or”什么，“either in the form of coal，or in the form of petroleum.”（哈哈）好！这个东西实际上是语言生成的过程。在语言生成的过程中，预测是非常重要的，我把它归纳起来，共有三个，是句法预测、词汇预测、世界知识预测三个方面。而这三个方面的东西，是语言交际和使用过程当中不可或缺的。Sinclair在那个时候，说他想开发一个叫expectancy grammar（预期语法）的东西，不过后来没搞出来了。但是什么事情你首先要有个思想，这个思想能不能变成现实有很多制

约条件的。后来Sinclair又提出了linear unit grammar（线性单位语法），基本是也是这样一些思想嘛。我认为这个要很好地深入思考，来观察语言究竟是怎么一回事情，人究竟是怎么使用语言的。意义不是机械的、现成的、孤立地附着在词上面。Words are used in isolation with fixed meaning。No! Meaning conveying is very much dependent on skills of interpretation。这些都是非常深刻的思想。由此我就想起毛荣贵先生讲的话了。词、概念和外部世界，这是三角关系啊！词代表的是一个概念；概念是外部世界同一类事物的本质特征的概括，对吧？但是，外部世界在人类头脑里的映像是不一样的，因为经验不一样嘛，就算经验一样，反应也不一样，也可能是认知结构什么的都不一样嘛！所以毛先生的文章里举的例子就是房子，对吧？房子就是一个建筑用来住的，但是房子反映在北京人脑子里是四合院，反映在天津人脑子里是小洋楼，反映在上海人脑子里是亭子间，这个这就产生了词、概念和外部世界所指的一个关系，那么在文章里出现房子这个词，你怎么诠释？每个人不一样吧！所以就产生问题了，一篇文章是不是只有唯一的解？就是从阅读理解的角度来讲，一篇文章是不是只有一个理解方式？恐怕不是的。否则的话同样一篇讲原子能的文章，原子能科学家、大学生、小学生的阅读理解不可能是一样的，对不对？所以像John Sinclair的这些思想，我觉得都是非常深刻的。那么我刚才讨论到直接跟非直接的观察，我跟他举了个我认为是直接观察的例子，我就是说写的文章你可以反复看吧！写下来的文章是语言运用吧，你是不是可以去摸呢，怎么不是直接的呢？那么我说语料库语言学，语料库作为一个工具，有点像望远镜、显微镜，我可以放大放远去看，对吧？我还举了个例子，我说你看遥远的五千光年以外的星球，你用望远镜去看，能不能知道上面有没有铁？我不知道，这个眼睛怎么看得见呢？但是我现在可以分析光谱啊！光谱里面哪一个波长长，上面有铁的，这个观察是直接观察还是间接观察呢？第一，我没有这样的三棱镜，我没有这样的望远镜，五千光年以外的星球看也看不见，如果我看见了这就是直接观察，但是我通过分析知道里面有铁，这是非直接观察，但是两个都是通过观察得到的，我说我写的东西，说的话我可以用录音机录下来反复听嘛，我是直接观察嘛，那记录下来就是个语料库，那么语料库语言学提供的就是直接观察嘛。他不同意，他说我不同意你这个观点。他说语料库语言学提供的就是非直接观察，他说，第一，Chomsky根本不在我们讨论的范围里面，他根本不关心观察不观察的问题，所以这个问题我们不谈；第二，我讲的非直接观察是不能倚靠五种感官来观察的东西。他还有一些很深刻的思想，我后来跟他有讨论，他的COBUILD Dictionary后来搞了双语的，上外出版社引进之后请我写一个序。我写了个序，认为它有五大特点：一个是基于语料库的，这个我说是“为词典学树立了榜样”；第二呢，我说：“所有事都是基于证据的”，这个也是很重要的；第三呢，我说“它的定义方式是独特的”。定义有三个要素嘛，一个要素是被定义的词，一个是它的上位概念，一个是上位概念底下这个词跟其他词的所

差。所以它三个要素就是：一是定义术语，二是它的上位词，三是一般性差异三个要素。举个例子来讲，比如，“a thermometer is an instrument”，“instrument”是“thermometer”的上位概念，那“instrument”多了，有各种各样的东西，“…which measures temperature”这个句子就很好地定义了。但是，COBUILD没有这样的定义。COBUILD怎么做呢？“If you want to say something, you say the…”对吧？所以我说COBUILD Dictionary是没有定义的，它只提供了上下文，而这个上下文是用户非常需要的东西。我这样解释了，我写了封信给他，他回了我一封很长的信。我深受启发，他说我们用的叫full sentence definition（整句释义法），简称FSD。这个东西我以前没有听说过。他后来跟我讲了，说full sentence definition（整句释义法）是基于三个前提：第一，只有人类的自然语言才有的本质特征之一，叫reflexivity（自反性），就是讲自己，只有语言有这个功能——谈论自己。只有人类的语言可以拿来讨论语言自己，非常深刻，人工语言也没有这个能力的。这是第一个特点。第二个特点呢？是paraphrasing（释义），释义是人类语言的特有功能。语言可以用来解释语言自己。这个也是其他书没有的。那么他所有的定义就是释义，对吧？实际上，所讲的正规定义三个要素都含在里面了。它是有superordinate跟differences的，superordinate就是它的上位概念，differences就是它这个词跟其他词的概念的区别，那么这个当然就是定义了，因为我满足定义的所有的要求了嘛，对吧？那么第三点，第三点也是非常重要的，就是语言的通透性。所有的语言都可以用这个来解释语言的现象，他说我提供了这个解释，可能有人感觉到不精确，但是它能解决问题啊。而更重要的是什么呢？他说他提出thesaurus（词林）的概念，我后来想了很多例子，thesaurus的概念是词林，我们国内也有很多词林，中文的词林，譬如说，高兴、欣喜、喜悦、愉快、愉悦等等，但是对我来讲，重要的不是例句一大堆给我讲，而是你要告诉我什么时候用这个什么时候不能用这个。由此我就想到《傅雷谈翻译》这本书，我几次建议你去看这本书，它里面就讲了他有的时候找不到一个词，两天吃不下饭，睡不着觉，想到了就手足舞蹈，为什么？那么你能不能提供出手段，我在这个情况下该怎么讲，你有这个手段的话，那好极了。语音教学、语音学习、词穷的问题就都没有了。讲到father可以是父亲、爸爸、老爸、老子、老爷子。但是很多词有条件，不是什么地方都可以用的。我想到的一个例子，就是patronize（资助），我们后来送了几个研究生去Antoinette那里，我本来想讲Thank you for patronizing the student PhD candidates，后来觉得没把握，就去查了这个词，发现这个词是消极的，谁patronize（资助）谁，他就保护你，而且讲话的腔调里显得我是高于你的，我照顾你，我高你一等，凌驾于上的，是消极的。那么这说明什么概念呢？是这样几个概念：态度的，语用的，跟内涵意义这三个意义。我用full sentence definition可以包含进来，你讲一个定义往往这个没有了，而没有对用户来说是没用了，那么我光理解一堆词更没有用。那么如果这个是对的，Sinclair说为什么其他字典不这样做呢？他说

第一，这个前提必须是基于语料库的，你要有证据才可以嘛。他没有所以他做不到，他就是同意他也做不到。第二，一切事物都是基于证据的，证据是硬碰硬的，他这些思想都是非常非常有用的。我觉得这个是语料库语言学未来的发展趋势，我们讲probabilistic linguistics（概率语言学），expectancy grammar（预期语法），full sentence definition（整句释义法）以及linear unit grammar（线性单位语法）都体现了这样一种思想，而且他在这个思想上搞了个phrase box（短语盒子），你只要定义在什么情况下应当怎么讲，它会告诉你。傅雷在翻译的时候，他想表达这个或那个，应该用什么词，他不用两天吃不下饭了，一分钟就有了。

问：2003年开会的时候，Sinclair在他的主旨演讲里，重新讨论了这些问题。就是您说的第一个关于语言的，用语言讨论语言，是元语言能力、metalanguage。他当时说学生学语言的时候，这也是一个必备的能力之一。就是你要学会怎么用语言来讨论语言，这也是个能力。Paraphrasing（释义）刚才您也讲了，这也是一个非常重要的概念。尤其是我现在这一段时间在研究文本与意义，文本的意义，或者句子内一个词语的意义，它是从哪来的？它是靠其他的词语或文本去解释它，靠释义，它背后有一整套哲学思想体系。

答：对，而且，他还有一个观点也对的，字典的本质是什么，一个是条目，那么有几千个几万个条目，每个条目后面都有这样一个释义的话，它就是个文本，那么一个字典就是个语料库，对于这个语料库，我如果有个手段，就可以互相参考，那我可以从这里查到那里，那么释义的问题也解决了，我需要表达什么从这里也可以解决了，这个是非常重要的。他叫这个为structural semantics（结构语义学）。我相信如果把这些思想用到教学上来，这种教学一定是有效教学，我的有效教学理念就是每一节课教有实效，学有实效，还课堂教学以本来面貌，学习要真真正正地在学习者的大脑中发生。

问：杨老师，最后请您谈谈，语料库以后的发展，你希望哪些事情发生或者有哪些建议？

答：我希望发生的就在这个路子上，把这些很深的思想继续发展而且深入地探索下去，一切要基于事实，一切要基于有效教跟有效学，不是搞花里胡哨的东西，也不是人云亦云，一定要有效，我想了下就是有效测试、有效教学、有效使用。搞语料库语言学也应当积累在这个基础上面。其一，回顾这半个世纪的历程，语料库语言学经历了马鞍形的发展，记得八十年代中期，Sinclair从事语料库语言

学研究时完全处于少数地位，COBUILD是当时整个英国文科领域最大的研究项目，计算机还没有成为文科领域必不可少的研究工具，尤其是用计算机来研究语言，许多人抱着怀疑和等着瞧的态度。因此，说“Sinclair 在学术思想、研究方法及研究发现上独开局面，新创气象”一点不为过，事实上Sinclair 的很多想法都很超前，比如他提出associative thesaurus（联想词林）的想法，据Sinclair 说，柯灵斯出版社未能理解而没有接受，于是只好搁浅，因为找不到竞争出版商，这一方面说明Sinclair超前，另一方面，也说明当时处于少数地位的学者从事语料库语言学研究困难之大；今天，语料库语言学早已成为显学，不但人数多、队伍大，写论文只要从现成的语料库里找一些例句就可以，并以此自称语料库语言学家，也有一些人攻击和批评Sinclair，其实是对Sinclair的学说、对语料库语言学并不甚了了。因此厘清Sinclair 的学术思想及其发展脉络，不仅仅是为Sinclair 的学术贡献进行辩护，更是为了澄清语料库语言学研究的基本思想和观点，这在当前有重要现实意义；Sinclair 的贡献不能抹杀，也不应被忽略。

其二，必须区分语料库与语料库语言学，语料库是研究工具，但语料库语言学不是。语料库实际上就是语言采样的集合，在数字化时代，语言采样并不困难，因此，单纯建立语料库不能成为科研项目的立项依据，这是对的，建立语料库必须说明用来研究什么，通过论证，才能立项。也就是说，把语料库作为一种研究工具，就是基于语料库的语言研究；而语料库语言学则是语言学本体研究，是把语料库作为研究途径，因为语料库提供了观察语言的新的视角和方法，可能成为理论语言学的一个方向，事实上，语言涉及人类生活的方方面面，可以从完全不同的角度进行研究。

其三，我觉得基于语料库的研究与语料库驱动研究两者不应当对立，前者着重在应用研究，后者着重在本体研究，各有自己的活动空间，各自可以做出重要的贡献，在我看来两者都是语料库语言学。我们常说既要重视语料库语言学的应用研究，当前更要重视语料库语言学的理论研究，因为前者做得多一些，而后者则比较薄弱，不加强研究，可能会制约语料库语言学的进一步发展。

其四，扩展意义单位的理论与研究方法奠定了现代语料库语言学的学科基础，这是Sinclair的贡献；Sinclair 继承和发展了弗斯的“语境意义”学说，语言研究不再局限在语言内部，语言作为交际工具当然是社会性的，从语言使用的角度来观察，言语行为必然涉及语言使用者，涉及who、when、where、why、how等诸多因素；patronize 不是简单意义上的“保护，赞助”，而是反映了说话人居高临下的架势，因此If someone patronizes you，说Thank you 是失态的，说“Don’t you patronize me! ”才是得体的。如此看来，“把搭配、语法同现、语义偏好以及语义韵整合起来进行综合分析”才是正确的路子，其研究成果才能更好地用来指导语言实践；“意义-结构”一体化的一元论语言学思想，就目前的认识来讲，正是语

料库语言学本体研究的基础。

其五，Sinclair 语言学的基本特征包括以下几个方面：1）只关注performance（语言运用），不关心competence（语言能力）。记得有一次跟Sinclair通信，谈起Chomsky 的competence（语言运用）/performance（语言运用）两分法，Sinclair非常明确地说Chomsky关心的根本是不同的东西，这个问题可以不予考虑；这不仅是一个理论倾向问题，Sinclair的早期经历，从研究课堂语篇起，到COBUILD词典编纂重大项目的实践，他关注的都是语言运用，而且也只能是语言运用；2）如果关注的是语言运用，研究对象只应当是实际使用的真实文本；3）语言研究必须采集语言数据，采集的数据必须是实际发生的语言运用实例，而且是未经删节的干净文本，生造的例子像Green idea sleeps furiously本身就是理论的一部分，不是证据；4）实证数据只有建成语料库，特别是大容量语料库，才能研究实际使用中的真实语言，这就是大数据原则；5）语料库作为证据源搜集的未经删节的索引行，提供了观察和分析实际使用中的语言的手段；6）完整文本原则使得在上下文中观察和分析语言并通过归纳法得出结论成为可能；7）语感：有一次在通信中我认为“语料库作为工具，为直接观察使用中的语言提供了可能”，还举了电子显微镜、光谱仪等例子，说明有了工具才有可能观察到肉眼观察不到的现象。Sinclair不同意这个说法，他认为使用中的语言的特征是无法直接观察的，语言分析必须依靠“the sixth sense”（第六感），我认为他说的就是语感。的确，并不是一个动词的大量索引行放在一起，就“看见了”这个动词的semantic prosody（语义韵）、semantic preference（语义趋向）以及attitudinal meaning（态度意义）等等，这些东西不能直接看出。归纳出这些特征“不仅需要精准的语言直觉，还需要严谨的技巧训练和反复的实践”，这说得很准确；8）因为是大数据，必须用计算机来处理，因此语料库语言学只可能在计算机时代出现。从上面说的这些来看，Sinclair语言学反映了不少大数据时代的特征，看样子Sinclair 超前不止20 年，可能超前了半个世纪。

其六，由于语言涉及人类社会生活的方方面面，语言现象极其复杂，需要且可以从不同的层面和侧面进行研究，形成许多不同的交叉学科，例如理论语言学、普通语言学、历史比较语言学、心理语言学、社会语言学、人类语言学、认知语言学、应用语言学、语料库语言学、计算语言学、统计语言学、概率语言学等等，还有句法学、词汇学、短语学、语用学、语义学、词典学等等，各个学科应当各自定义自己的研究领域、相互补充，没有必要相互否定；例如，Chomsky关于语言能力的理论，探讨语言能力本质，极具说服力，能够得到神经语言学、发展心理学方面的许多实验结果的支持，尤其在儿童语言习得方面有说服力的实例很多。Chomsky 的理论关注的语言能力，是语言机制的本质，属于个体和心智的范畴；应用语言学关注的是语言运用，是在社会言语社团中实际使用语言的能力，这是

两个不同的范畴，应当相互补充，没有必要相互否定；正像脑外科医生需要研究“布洛卡士失语症”患者为什么不能正确发音、为什么失去语法能力仍能听懂别人说的话，而且其他方面的智力活动没有受到影响等等，这些研究都非常重要，都有巨大的应用价值，但对应用语言学领域的语言教学和研究则没有很大关系，因为本来就是不同的领域，可以各领风骚，各自从不同的侧面为人类认识语言做出贡献，没有必要相互否定；在语料库语言学内部，我认为基于语料库研究与语料库驱动研究之间也不应当划一条鸿沟，在我看来，前者侧重应用研究，后者则侧重理论研究，也就是语言本体研究。

最后，语料库语言学是创造性、开拓性、探索性的研究，要持之以恒。Sinclair 的学术、思想往往领先同时代人超过20年，因此容易受到批评甚至攻击，这是先驱者经常遇到的命运；一方面要开拓研究，不断发展，一方面要从正面进行宣传和论证；同时也要善于等待，坚信真理有时在少数人手里；你说得很对，“但语料库语言学作为一门学科，其发现成果和理论会越来越丰富，其学科辨识度会更加显著”。

短语学视角下的汉英共选型式对等

上海海洋大学 李晓红

提要：本研究以上海交通大学英汉/汉英平行语料库中提取的翻译对等词为起点，在可比语料库中调查各翻译对等词与语义及语用的共选型式，通过对比和考察汉英共选型式的语义韵、语义趋向及搭配的对应情况探讨影响跨语言对等的因素。研究发现，跨语言对等存在于共选型式中。在共选型式中对语义韵常态的偏离会产生新的意义单位，型式的改变产生意义移变单位，目标语中的型式对等也随之改变。研究还发现，共选型式中的固有语义趋向是在语义层面实现型式对等的重要因素，而可选语义趋向对型式对等没有决定性作用。在搭配层面，搭配词的具体语义特征不对应不会影响型式对等，但会产生低频或不适切的用法。

关键词：跨语言对等、语义韵、语义趋向、共选型式、语料库

1. 引言

语料库短语学在近20年间最重要的发现是：语言使用具有短语趋势，语言使用者一次性选择两词及以上的词语组合来表达意义（Sinclair 2004；何安平 2013；卫乃兴 2014）。语言使用的短语趋势将人们对意义单位的理解由单个词语转向词项（lexical item），即Sinclair（2004：24）提出的扩展意义单位（extended unit of meaning），后来称为意义移变单位（meaning shift unit）（Sinclair 2010：44）。随着双语和多语语料库的建设和发展，语料库对比短语学随之产生和兴起（卫乃兴 2014）。对比短语学研究通过观察、描述和对比双语词语在形式、意义和功能层面的特征异同，探索和发现跨语言意义单位（如Berber-Sardinha 2000；Tognini-Bonelli 2002；Xiao & McEnery 2006；Dam-Jensen & Zethsen 2007；Stewart 2009；李晓红、卫乃兴 2012等）。

基于单语语料库的短语学研究一致认为，语言的意义单位不是单词，而应该是短语（如Sinclair 1996；Hunston & Francis 2000；Stubbs 2001；Granger & Paquot 2008等）。短语单位可泛指语言中各种各样的词语组合（何安平 2013：8），在形式上涵盖了习语、固定或半固定搭配、连续性或非连续性多词单位、类联接以及反复出现的词汇-语法型式等。Sinclair（1996）提出的扩展意义单位模型揭示出

意义单位在本质上是词汇、语法、语义和语用各要素的共选（co-selection）结果，包含各共选要素的短语单位是语言基本的表意单位。其中，表达整个短语单位的语用功能和态度意义的语义韵是短语单位的必要组成，体现说话者或作者的交际目的和交际意图（Louw 1993；Sinclair 1996）。

基于双语或多语语料库的对比短语学研究证据发现，跨语言意义单位也主要是短语单位，跨语言对等单位要通过考察双语词语在形式、意义和功能方面的各种对应关系才能确立（Tognini-Bonelli & Manca 2009；卫乃兴 2014）。Stewart（2009：29）在探讨跨语言对等时提出："语义韵应被视为翻译者需要关注的现实，否则会忽略源语文本中的重要因素。"语义韵是短语单位的必要组成和功能指向，是确立跨语言对等的关键因素（Stewart 2009；李晓红、卫乃兴2012）。

然而，短语学视角下的跨语言对等研究还存在亟待解决的问题。首先，语义韵冲突对确立跨语言对等的影响。Morley & Partington（2009：146）指出语义韵冲突是语义韵常态被"关闭、推翻或利用"（switched off, overridden or exploited）（参见Hoey 2005）。Louw（1993）研究发现，故意违反语义韵常态及使用非典型搭配是为了实现特殊的交际目的和交际效果，如bent on self-improvement用于强调，而there's much to be said for failure（Morley & Partington 2009：146）实现讽刺效果。在单语视角下，与某个短语单位共选的语义韵呈现的是语言使用的惯性或趋势，体现出特定语言社团对该短语单位一种已确立的、默认的态度意义。然而，语义韵的默认值可能发生改变，如从积极转变为消极。在跨语言语境中，这种与人们所期待的语义韵相悖的语言使用是否影响以及如何影响目标语中的翻译对等，需要深入讨论。其次，短语单位是词语、语法、语义和功能的共选型式，其中语义趋向体现的是词语和语义类别的共选行为，这些语义特征在局部语境中有助于对语义韵的解读（Partington 2004；Bednarek 2008）。单语和双语视角的研究发现，短语单位核心词可能伴随多个语义趋向（Stubbs 2001；李晓红 2015），在跨语言对比研究中，需要进一步考察双语词语的各语义趋向是否达到一致对应。再次，在识别短语单位的过程中，搭配词往往呈现出多样化和多变性的特征，这些"内部变化"（internal variation）（Sinclair 1996：86）会随着更加抽象的语义特征的概括性描述而趋于消失。在对比研究中，搭配词层面的对应往往受制于语义趋向层面的对应，然而双语词语的搭配词究竟有何种对应特征也是需要深入讨论的问题。

鉴于上述问题，本研究通过描述汉英对等词语在搭配、语义和语用层面的特征异同，讨论汉英共选型式的对应关系，进而探索影响跨语言对等的因素。研究以英汉/汉英双向平行语料库中提取的翻译对等词为出发点，然后在可比语料库中分析翻译对等词在搭配、语义趋向和语义韵层面的对应。我们认为，词语会习惯性地与多种型式共选，每种型式都趋于表达某种态度意义，跨语言对等存在于共选型式中，而不是单纯的词语对应。我们还将指出，每一型式中一个成分的改变

可能会不同程度地影响跨语言对等的确立。

2. 研究设计

2.1 工作定义

在本研究中，与被调查的节点词有规律复现的词语所共同表达的语义特征即语义趋向。语义趋向是节点词与搭配词共同选择的结果，是在搭配词共有的语义基础上更为抽象的语义概括。根据Sinclair（1996）对语义韵的界定，我们将语义韵定义为：由核心词、搭配、类联接、语义趋向各因素相互作用而产生的态度意义，本质上体现说话者或作者的交际目的。语义韵“涉及意义的广义方面，包括态度、策略及语用”（Sinclair 2010：45）。因此，本文中的态度意义即语义韵。本研究中，语义韵的描述包括语义韵极性和具体的交际目的。语义韵极性可通过语义韵力度指数（prosodic strength index，PSI，下同）进行判定，判定公式如下：

$$PSI_{pos} = F_{pos} / F_N$$
$$PSI_{neg} = F_{neg} / F_N$$

F_{pos} 和 F_{neg} 分别代表共选型式表达的积极或消极语义韵，F_N 代表该型式在语料库中的总频数。例如，以动词“夺取”为核心词的共选型式在MCC中出现30次，其中23次表达消极语义韵。根据上述公式，F_{neg}=23，F_N=30，PSI_{neg}=23/30≈0.77，即消极语义韵力度指数为0.77。这意味着该共选型式表达消极语义韵的趋势高于积极语义韵。语义韵力度指数是考察跨语言对等单位对应程度的可行方法。

语义韵常态指核心词与特定语义趋向的共选在语境中惯常表达的态度意义（李晓红、卫乃兴 2012）。语义韵常态是反复出现的惯例语言使用，是语言社团中的“背景语义韵或默认语义韵”（Morley and Partington 2009：148），揭示语言社团的集体心理倾向。与正常预期相悖的态度意义是对语义韵常态的偏离。在实际操作中，语义韵常态的偏离可使用语义韵力度指数进行考察和判定。如果积极的语义韵力度指数高于消极指数，则语义韵常态为积极，表达消极态度意义的用法则被视为对语义韵常态的背离。

在本研究中，共选型式体现的不是词汇与词汇共现，也并非强调词汇与结构共选，而是要突出描述词汇与语义特征及态度意义之间的共选规律。本研究基于扩展意义单位模型（Sinclair 1996）识别核心词同语义趋向和语义韵的共选结果构成的型式，采用如下方式呈现：

([semantic feature]$_{SEMFREF}$) **node** [semantic feature]$_{SEMFREF}$ ==> [attitudinal meaning]$_{SEMPROS}$

在上式中，以节点词node为核心的共选型式包括语义趋向（缩写为SEMPREF，下同）和语义韵（缩写为SEMPROS，下同）两个共选要素。在核心词周围可能伴随多个不同的语义趋向，圆括号代表可被选的语义趋向，即该语义特征并非总伴随在核心词周围；箭头代表语义共选指向特定的态度意义。例如，以“夺取”为核心词的共选型式为：

([intention]$_{SEMPREF}$) 夺取 [victory/success]$_{SEMPREF}$ ==> [desiring to achieve sth. meaningful]$_{SEMPROS}$

在核心词左侧，表达“意图”（intention）的语义特征是可被选的语义趋向，在其右侧语境中，语义趋向“成功/胜利”（victory/success）高频出现在类联接V n中。可选语义趋向和固有语义趋向共同传递“期待实现有意义的目标”（desiring to achieve sth. meaningful）的态度意义。

2.2 语料库

本研究使用两个通用语料库作为可比语料库：英国国家语料库（BNC）和现代汉语语料库（MCC）（参见李晓红、卫乃兴 2012）。本研究主要研究书面语中的语言使用，因此我们选用BNC的书面语部分（http://corpus.byu.edu/bnc/），总词容约为9000万词，涵盖不同的语域和主题。现代汉语语料库的核心部分总词容为2000万字，同样涵盖了多种语域和主题。研究还使用一个平行语料库：上海交通大学平行语料库（SJTU Parallel Corpus, SJTUPC，下同）（参见Wei & Li 2014），我们将在该语料库中提取英汉翻译对等词作为通用语料库的调查对象。

2.3 研究步骤

本研究采用定量和定性结合的方法，具体步骤如下：

首先，从SJTUPC中提取并拟定翻译对等词。SJTUPC的自带数据库能够呈现一对多的翻译对等。例如，“夺取”在SJTUPC中对应8个英语表达：seize，wrest，capture，takeover，accomplish，achieve，win and take。为详尽讨论每组翻译对等在不同层面的对应关系，我们采用相互对应率（mutual correspondence, MC，下同）（Altenberg 1999：254）来选取在语料库中对应度最高的一组（参见李晓红、卫乃兴 2012）。根据相互对应率，我们从前期调查的30对汉英对等词语中选取MC值较高的三组对等词作为调查对象：“助长”vs. *FUEL*（MC=42.9%）、“夺取”vs. *SEIZE*（MC=25%）、“平息”vs. *ASSUAGE*（MC=25%）。

其次，我们以Sinclair（1996）的扩展意义单位模型为调查共选型式的主要工作框架，在通用可比语料库中观察词语在搭配、语义趋向和语义韵层面的对应情况，概括与每个词相关的共选型式，并根据语义韵力度指数判定语义韵常态。

最后，通过考察共选型式在语义韵、语义趋向和搭配层面的对应情况，确立跨语言对等并讨论影响跨语言对等的因素。

3. 共选型式及共选型式对等

3.1 共选型式

本节在可比语料库中观察3组翻译对等词的语义和语用特征并识别其共选型式。通过观察“夺取”与*SEIZE*分别在MCC和BNC中的100行索引，我们发现二者皆与类联接V n共现，根据语义和功能特征，以“夺取”和*SEIZE*为核心词的共选型式概括如下：

表 1. 以“夺取”和*SEIZE*为核心的共选型式及语义韵力度指数（PSI）

1.	([intention]$_{SEMPREF}$) 夺取 [victory/success]$_{SEMPREF}$ ==> [desiring to achieve sth. meaningful]$_{SEMPROS}$ PSI $_{pos}$=1.0
2.	([armed military battle]$_{SEMPREF}$) 夺取 [territories of a sovereign state]$_{SEMPREF}$ ==> [wild ambition or determination]$_{SEMPROS}$ PSI $_{neg}$=0.77
3.	[men of a social class]$_{SEMPREF}$ 夺取 [power]$_{SEMPREF}$ ==> [a tactical plan or a purposeful plot]$_{SEMPROS}$ PSI $_{pos}$=0.63
4.	夺取 [dominant position or competitive advantage]$_{SEMPREF}$ ==> [desirable goal]$_{SEMPROS}$ PSI $_{pos}$=0.6
1.	([armed forces]$_{SEMPREF}$) ***SEIZE*** [places/territories of a sovereign state]$_{SEMPREF}$ ==> [opposition]$_{SEMPROS}$ PSI $_{neg}$=1.0
2.	([intention]$_{SEMPREF}$) ***SEIZE*** [political power]$_{SEMPREF}$（[armed forces] $_{SEMPREF}$）==> [illegal plots]$_{SEMPROS}$ PSI $_{neg}$=1.0
3.	***SEIZE*** [dominant or advantageous position]$_{SEMPREF}$ ==> [desirable goal]$_{SEMPROS}$ PSI $_{neg/pos}$= 0.5

如表1所示，“夺取”被包含在4种不同的共选型式中，且每一型式在语义趋向、语义韵及语义韵力度指数方面均呈现差异。“夺取”常与表达“胜利/成功”的名词共选，如“胜利、全胜”指在军事战斗、体育赛事或商业竞争中取胜，“冠军、金牌、奖牌”特指在体育比赛中取胜，“丰收、高产”表达农耕活动的成功。同时，“为、要、去”等字时常在“夺取”左侧的毗邻语境中出现，指向“意图”的语义趋向。由此，固有语义趋向“胜利/成功”及可选语义趋向“意图”在语境中主要表达下决心或意图实现成功的语义，传递“渴望实现目标”的积极态度意义。在100行索引中，“夺取”与上述语义和语用特征的共选例证共有37例，无一例外地传递积极语义韵，PSI为1.0。

与“夺取”共现的另一组名词搭配词包括：澳门、北京、东北、热河、察哈尔、东方港口、法国的一部分地区、比利时、北欧、波多黎各、关岛和菲律宾

等。其中“夺取土地”出现5次，是MCC中较为高频的搭配。这些共现名词均指向“领土/土地”的语义特征，是“夺取”的固有语义趋向。同时，在“夺取”周围常伴随如“反击、作战、战争、战役、战斗”等词语，均指向“武装军事战斗”的语义趋向。“夺取”与上述两个语义趋向共现时，语境中常见如“英雄、根据地、人民战争、游击战争、打倒军阀、打败侵略者”等表达，体现出夺回失地的坚定决心，在语境中传递积极的态度意义。另一组共现词，如“侵占、吞并、企图、妄图、殖民者、列强、野心、疯狂地”，表达了入侵者的险恶意图，在语境中传递消极的态度意义，且消极PSI为0.77。这说明“夺取”与“领土/土地”及“武装军事战斗”的语义趋向共选时，汉语本族语者更趋于预测消极的态度意义。

在100行索引中，“夺取”与表达“权力”的语义趋向共现18次，实现该语义趋向的名词主要有“政权、统治权、皇权、权力”等。语义韵的解读主要凭借表达动作施为者的名词，这类名词常指代“某一社会阶层的人群”。在特定的历史时刻，代表人民利益通过武力夺取政权、推翻腐朽统治可能是最终的必要手段。然而，为个人利益夺取权力是应受谴责的行为。由表1可知，“夺取”的第3种型式主要表达积极语义韵，PSI为0.63。与“夺取”共现的另一组名词搭配词包括“优势、武器、领先地位、主动权、利润、权益、市场”等，这些词语共同表达“主导/优势”的语义特征，在语境中常传递积极的态度意义，PSI为0.6。

在BNC中，*SEIZE*习惯性地吸引territory/territories, land/lands, town/towns, city/cities等名词，还常与指代不同地区的名词共现，如the town of Guazapa, the port of Batum, islands, the Ukrainian and Silesian territories, the kingdom, Edinburgh，这些表达均指向“领土/土地”的语义趋向。同时，共现表达如guerrillas, Turkish forces, Iran, YUGOSLAV army troops, army rebels, protesters和Fatah guerrillas，暗示局势紧张、时局动荡或非法行动，传递和加强了消极的态度意义。由此，*SEIZE*与“领土/土地”及“武装力量”的语义趋向共现时，无一例外地表达了消极态度意义，PSI为1.0。

在100行索引中，*SEIZE*与power共现17次，是高频的复现搭配。在SEIZE power左侧语境中，planned to, an attempt to 和 attempting to 均表达“意图”的语义趋向。共现语境中另一组名词搭配词表达了“武装力量”的语义特征，如army-backed coup, bloodless coup, coup, armed coup, anti-democratic forces, armed forces, the military, terroristic conspiracy, mercenaries and a group of soldiers。这些语义趋向有助于我们解读出“非法策划夺权”的消极语义韵。由此，我们得到*SEIZE*的第2种共选型式，包含了固有语义趋向“政治权力”、可选语义趋向“意图”和“武装力量”以及消极语义韵。*SEIZE*还与表达“优势/主动”类的名词共选，其积极和消极语义韵力度几乎相等，因此积极和消极PSI都为0.5。

同样，根据共选语义和语用特征，“助长”与*FUEL*的共选型式概括如下：

表 2. 以“助长”和*FUEL*为核心的共选型式及语义韵力度指数

1.	([amplification/negation]$_{SEMPREF}$) 助长 [tendency/habit]$_{SEMPREF}$ ==> [severity]$_{SEMPROS}$ PSI $_{neg}$= 0.8
2.	([possibility]$_{SEMPREF}$) 助长 [social phenomenon or activities]$_{SEMPREF}$ ==> [severity]$_{SEMPROS}$ PSI $_{neg}$= 0.62
3.	([amplification]$_{SEMPREF}$) 助长 [manner/momentum]$_{SEMPREF}$ ==> [severity]$_{SEMPROS}$ PSI $_{neg}$=0.67
4.	助长 [pathological/physical/natural phenomenon]$_{SEMPREF}$ ==> [severity]$_{SEMPROS}$ PSI $_{neg}$=0.63
1.	***FUEL*** [financial or social phenomenon]$_{SEMPREF}$ ==> [severity]$_{SEMPROS}$ PSI $_{neg}$=0.63
2.	***FUEL*** [uncertain/ungrounded belief]$_{SEMPREF}$ ==> [seriousness]$_{SEMPROS}$ PSI $_{neg}$=1.0
3.	***FUEL*** [confrontational ideas/actions]$_{SEMPREF}$ ==> [severity]$_{SEMPROS}$ PSI $_{neg}$=0.84
4.	***FUEL*** [strong emotion/mentality]$_{SEMPREF}$ ==> [severity]$_{SEMPROS}$ PSI $_{neg}$=0.67

表2显示，以“助长”为核心词的共选型式体现三点重要特征。首先，“助长”的4种型式无一例外地传递消极语义韵，强调“恶性发展的严重性”。这说明汉语本族语者趋于将“助长”与消极的态度意义联系起来，并强烈期待该词出现在消极语境中。其次，每种型式中的固有语义趋向由类联接V n中的名词实现，而有些名词强烈依赖前置形容词来限定和丰富其语义。例如，在第3种型式中，“助长”习惯性地与“气焰、势头、气势、声势、威势”等名词共现，这些词语共同表达“使事件进展更迅猛的力量或方式”，语义趋向可概括为“方式/势头”。该语义趋向较为中性，似乎不能提供解读“积极/消极”态度意义的信息，但这些名词强烈依赖其前置修饰语赋予其更多的语义信息，例如：

（1）**法西斯意大利的侵略气焰**

（2）**敌人的嚣张气焰**

（3）**殖民侵略的势头**

（1）—（3）粗体修饰语共同表达“非正义”、“有攻击性”、“势头凶猛”的语义特征。“助长”与这些语义特征共现，暗示这些方式或势头是“严重的、应受谴责的”。

再次，在类联接V adj n界限之外，“助长”的左侧语境中会伴随不同的可选语义趋向。在表2的第1和第3种型式中，“助长”左侧“强化程度”的语义趋向常由“更、更加、足以”等词语实现。在型式2中“助长”常与情态动词“会、可能”共现，表达“可能性”的语义特征。这些可选语义趋向加强了语境中弥漫的消极氛围。

*FUEL*在BNC中的频数为5,154次，在100行随机抽取的索引行中，54%的*FUEL*用于“为汽车或机器提供燃料”之意，剩下46%用于比喻意义。*FUEL*的比喻义在语义上更贴近“助长”，因此我们聚焦*FUEL*的比喻义用法，在BNC中另抽取了100行索引进行观察。表2显示，*FUEL* 与不同的语义趋向共现，组成不同的共选型式，但4种型式均表达“严重性”的消极语义韵。这说明英语本族语者倾向于在消极的词汇-语法环境中期待*FUEL*传递消极的态度意义。另一个需要注意的是，与“助长”的搭配特征不同，*FUEL*的名词搭配词并不依赖前置形容词修饰语来限定或丰富其语义信息。在类联接V n中的名词搭配词可直接揭示隐含的态度意义。例如，在型式2中，与*FUEL*经常共现的名词speculation, rumours, fear/fears, doubt, suspicions, concern/concerns, apprehension共同表达的语义特征为“不确定或无根据的想法”，*FUEL* 与这些名词共选，表达了内心不确定的想法既在蔓延又很强烈，语境中弥漫“严重性”的消极语义韵。在型式3中，与*FUEL*共现的名词controversy, arguments, debate, war/wars, conflict均表达“想法或行动上的对立”的语义特征，核心词与该语义特征在语境中表达“想法或行动上的对立状态持续存在或更加恶化”的语义，暗示说话者或作者认为问题具有“严重性”的态度。

根据搭配、语义趋向和语义韵信息，“平息”和*ASSUAGE*的共选型式概括如下：

表 3. 以“平息”和*ASSUAGE*为核心的共选型式及语义韵力度指数

1.	([by force]$_{SEMPREF}$) 平息 [riot/rebellion against the government]$_{SEMPREF}$ ([devotion]$_{SEMPREF}$) ==> [paying great cost to quell riots]$_{SEMPROS}$ PSI $_{pos}$=1.0
2.	([difficulty]SEMPREF) 平息 [trouble/dispute causing agitation]$_{SEMPREF}$ ==> [making efforts to calm down a disturbing matter]$_{SEMPROS}$ PSI $_{pos}$=0.81
3.	([difficulty]$_{SEMPREF}$) 平息 [mental state of emotional disturbance]$_{SEMPREF}$ ==> [struggling with strong and disturbing emotions]$_{SEMPROS}$ PSI $_{pos}$=0.88
4.	([gradualness]$_{SEMPREF}$) 平息 [disturbing sounds or things]$_{SEMPREF}$ ==> [expecting disturbing noises to subside]$_{SEMPROS}$ PSI $_{pos}$=1.0
1.	([difficulty]$_{SEMPREF}$) ***ASSUAGE*** [strong and unpleasant emotion]$_{SEMPREF}$ ==> [struggling with overwhelming emotions]$_{SEMPROS}$ PSI $_{pos}$=0.88
2.	([difficulty]$_{SEMPREF}$) ***ASSUAGE*** [troubles/problematic situations]$_{SEMPREF}$ ==> [trying to mitigate something unpleasant and serious]$_{SEMPROS}$ PSI $_{pos}$=0.86
3.	***ASSUAGE*** [desire/need]$_{SEMPREF}$ ==> [willing to satisfy needs]$_{SEMPROS}$ PSI $_{pos}$=0.82

如上所示，与“平息”共现的语义趋向共同表达“混乱状态”的语义，“平息”的左侧语境中会伴随否定表达、情态表达或副词，在不同型式中发挥特定的

语义和功能作用。上述4种型式表达的都是积极语义韵，且语义韵力度指数较高。

*ASSUAGE*名词搭配词的语义特征可大致分为三类："强烈的消极情绪"、"困难/麻烦"、"渴望/需要"。当*ASSUAGE* 与表达"困难/麻烦"的语义趋向共选时，其左侧语境中会伴随否定表达、情态表达或副词（下例中的下划线部分），部分例证显示如下：

(4) cussed it with him. My apologies did little to **assuage** the situation and I was informed that

(5) being, a dull, aching pain that could never be **assuaged**. She walked into the flat, and

(6) er: the memory of whose dying he could not **assuage**: Bearing gifts of flowers and sweet nuts

(7) their feelings. Party opinions were partially **assuaged** by a protest meeting at the Carlton Club

(8) nd those pangs which had been temporarily **assuaged** by what it had found in the wall

以*ASSUAGE*为核心词的3种型式传递的语义信息为"尽力缓解令人不愉快且严重的情势"，均表达积极语义韵。

3.2 共选型式对等

在上节中，我们识别了三组翻译对等词的共选型式，本节通过对比语义和语用层面的特征异同来确立跨语言对等。

3.2.1 "夺取"与*SEIZE*的共选型式对等

由表1可知，在类联接V n中，"夺取"和*SEIZE*共同吸引"领土/土地"的语义趋向，构成的共选型式在各自的语境中传递消极语义韵。这说明以"夺取"和*SEIZE*为核心的共选型式在各自语言中表达相似的语义和语用特征，因此可确立为共现型式对等，如表4所示。

表4. "夺取"和*SEIZE*的共选型式对等

共选型式对等1	([armed military battle]$_{SEMPREF}$) 夺取 [territories of a sovereign state]$_{SEMPREF}$ ==> [wild ambition or determination]$_{SEMPROS}$ PSI $_{neg}$= 0.77	([armed forces]$_{SEMPREF}$) ***SEIZE*** [places of a sovereign state]$_{SEMPREF}$ ==> [opposition]$_{SEMPROS}$ PSI $_{neg}$= 1.0
共选型式对等2	[men of a social class]$_{SEMPREF}$ 夺取 [power]$_{SEMPREF}$ ==> [a tactical plan or a purposeful plot]$_{SEMPROS}$ PSI $_{pos}$= 0.61 (PSI $_{neg}$= 0.39)	([intention]$_{SEMPREF}$) ***SEIZE*** [political power]$_{SEMPREF}$（[armed forces]）==> [illegal plots]$_{SEMPROS}$ PSI $_{neg}$= 1.0
共选型式对等3	夺取 [dominant position or competitive advantage]$_{SEMPREF}$==> [desirable goal]$_{SEMPROS}$ PSI $_{pos}$= 0.6	***SEIZE*** [dominant or advantageous position]$_{SEMPREF}$ ==> [desirable goal]$_{SEMPROS}$ PSI $_{pos}$= 0.5

表4显示，“夺取”和*SEIZE*习惯性地与表达“权力”的名词共选，在语义趋向方面实现对应。以*SEIZE*为核心词的共选型式强烈地表达消极语义韵，在同样的语义环境中，“夺取”惯例地表达积极的态度意义，PSI为0.6，这说明以“夺取”为核心词的共选型式有40%的可能性表达消极语义韵，因此以“夺取”和*SEIZE*为核心的共选型式能够在功能上实现对应，可确立为共选型式对等。

对比发现，“夺取”和*SEIZE*都吸引表达优势/主动的名词，如“武器、主动权”等。尽管与“夺取”的共选行为相比，*SEIZE*与该语义趋向的共现频数较低（在BNC中只出现10例），但二者也呈现出相似的语义期待。同时，与二者相关的共选型式均表达积极语义韵，第3组共选型式对等如表4所示。

上述三组已确立的共选型式对等中有两点值得注意的地方。第一，翻译对等词的搭配词呈现语义重叠，但语义趋向往往呈现出复杂对应情况。例如，在类联接V n中，“夺取”和*SEIZE* 共同期待表达“权力”的语义趋向，但“夺取”的左侧语境中会伴随“某个社会阶层的人群”的表达，而在*SEIZE*的左侧语境中会发现“意图”类表达。“夺取”和*SEIZE*的可选语义趋向存在差异，但并未影响其共选型式对等的确立。这一发现说明，与核心词强烈共选的语义趋向是所组成的共选型式的固有成分，其对应程度是考察共选型式对等的必要条件。同时，在核心词周围还伴随可选语义趋向，然而在双语对比环境中其对应程度对判定共选型式对等未必产生决定性作用。第二，源语和目标语中的共选型式对等表达相同的语义韵，但语义韵力度会存在差异。例如，与表达“领土/土地”的语义趋向共现时，以“夺取”为核心词的型式传递消极语义韵，PSI为0.77；以*SEIZE* 为核心词的型式则强烈地表达消极语义韵，PSI为1.0。然而，在共同吸引表达“权力”的语义趋向时，与“夺取”相关的型式表达积极语义韵，PSI为0.61，*SEIZE*的型式则具有1.0的消极语义韵力度指数。说明“夺取”另有40%的可能性期待消极态度意义，在这种情况下，二者可判定为跨语言对等。

3.2.2 “助长”与*FUEL*的共选型式对等

“助长”和*FUEL*都出现在类联接V n中，二者在语义趋向方面呈现一定的相似特征。*FUEL* 强烈吸引enthusiasm和desire以及其他共现频率为一次的名词，如hatreds, hate, anger, annoyance, anti-German feeling, psychological addiction and nationalism。这些名词均指向“强烈情绪或感情”的语义趋向，且大部分表达消极情绪。“助长”常与“趋势/习惯”的语义趋向共现。实现该语义趋向的名词主要表达两类具体语义特征：1、在某一群体中盛行的社会趋势，如“发展、之风、风气、风、倾向”；2、某一个体或群体的思维趋势，如“不满和气愤、主义、思想、心理”。后一语义类别中的名词搭配词与*FUEL*的上述名词搭配词共同表达的语义特征为“情绪、思维/心理”。同时，与“助长”共现的名词常依赖前置修饰

语来表达消极的心理或思想状态，如“个人主义、官僚主义、腐化思想、唯利是图的思想、依赖心理”。这表明“助长”和 *FUEL* 趋于期待相似的语义特征，二者共同表达“增加消极事物强度”的语义，在语义层面十分对应。通常情况下，人们不期望消极情绪或心理的增长，因此以“助长”和 *FUEL* 为核心词的共选型式均表达“严重性”的消极语义韵，可确立为共选型式对等，如表5所示：

表5. “助长”和 *FUEL* 的共选型式对等

共选型式对等1	([amplification/negation]$_{SEMPREF}$) 助长 [unpleasant mental tendency]$_{SEMPREF}$ ==> [severity]$_{SEMPROS}$ PSI $_{neg.}$= 0.8	***FUEL*** [strong unpleasant emotion/mentality]$_{SEMPREF}$ ==> [severity]$_{SEMPROS}$ PSI $_{neg.}$= 0.67
共选型式对等2	([possibility]$_{SEMPREF}$) 助长 [social phenomenon or activities]$_{SEMPREF}$ ==> [severity]$_{SEMPROS}$ PSI $_{neg.}$= 0.62	***FUEL*** [financial or social phenomenon]$_{SEMPREF}$ ==> [severity]$_{SEMPROS}$ PSI $_{neg.}$= 0.63

在另一词汇-语法环境中，“助长”和 *FUEL* 在语义趋向和语义韵层面均呈现相似性。*FUEL* 强烈吸引表达“金融或社会现象”的名词，且大部分名词都指代“金融或社会问题”，如 inflation, growth, uncertainty, unemployment, debts, market for fakes and forgeries, decline, poaching and drugs trade。这些名词的左侧语境中，会伴随 fiery, unsustainable, widespread, massive 等属性形容词，表明形势正在恶化、扩展或蔓延。语义韵可解读为“严重性和严峻性”，PSI 为 0.63。“助长”常吸引“投机活动、盲目生产、重复建设”等表达，其共有的语义特征为“社会现象或活动”。“助长”与该语义趋向以及左侧语境中的强势语，如“往往、足以”，表达了“严重性”的消极语义韵，PSI 为 0.62。由此，以“助长”和 *FUEL* 为核心词的共选型式可确立为跨语言对等，如表5所示。该组型式对等均表达“增加消极社会现象的恶化或蔓延”的语义，暗示说话者或作者认为该行为做法具有“严重性”的态度。

3.2.3 “平息”与 *ASSUAGE* 的型式对等

观察发现，*ASSUAGE* 强烈吸引表达“强烈情绪或心理状态”的名词，如 fears, guilt, loneliness, misery, anxieties, concern, pride。在 *ASSUAGE* 的左侧毗邻语境中表达“困难”的语义趋向较为明显，如否定表达 not really, evidently not, was not, did not, did little to, could not 及半情态表达 aimed at, trying to, aimed to, in an effort to。*ASSUAGE* 与该语义趋向共选，表达“费力地/试图缓解强烈的情绪”。通常情况下，缓解不愉快的情绪是理智的做法，因此语境中暗示出说话者或作者所持有的积极态度，PSI 为 0.88。

另一方面，“平息”吸引表达“强烈情绪或心理状态”的名词，如例（9）—（13）中下划线部分，“万顷波涛、潮水般的、波澜”等词语突出了纷乱的心境或情绪。

（9） 写到这里，姑娘再也无法**平息**感情的万顷波涛，她眼前出现了一组组幻

（10） 我收不起贪婪的眼睛，也**平息**不了激动的感情，作了历史主人的家乡父老，

（11） 略带霉味的空气，想啊，想啊，心潮久久**平息**不下来……马班长同我在电影、小说里看到的老红军

（12） 精沉思了好一会儿，潮水般的思绪才渐次**平息**下来，猛一撂头，发觉张贵喜一直站

（13） 里掀起过狂风巨浪，可她最后还是理智地**平息**了感情的波澜，因为她已经有了丈夫

在“平息”的共现语境中，我们还发现否定表达，如“再也无法平息、平息不下来、平息不了”，以及前置副词修饰语，如“理智地、久久、渐次”，见例（9）—（13）中斜体部分。这些表达均指向“困难”的语义趋向。上述共选行为表达“费力地缓解或克制强烈且纷乱的情绪”的语义。同样地，克制强烈情绪通常是理智且可取的做法，因此由“平息”、语义趋向和语义韵构成的共选型式传递积极的态度意义，PSI为0.88。

对比分析表明，“平息”和*ASSUAGE*在与“情绪/心情”语义趋向共选时，均表达积极语义韵，共选型式对等可确立下表6：

表6. “平息”和*ASSUAGE*的共选型式对等

([difficulty]$_{SEMPREF}$) 平息 [mental state of emotional disturbance]$_{SEMPREF}$==> [struggling with strong and disturbing emotions]$_{SEMPROS}$	([difficulty]$_{SEMPREF}$) ***ASSUAGE*** [strong and unpleasant emotion]$_{SEMPREF}$==> [struggling with overwhelming emotions]$_{SEMPROS}$
PSI $_{pos.}$= 0.88	PSI $_{pos.}$= 0.88

本节采用Sinclair（1996）提出的扩展意义单位模型识别了三组翻译对等词的共选型式，并在双语对比视角下确立了跨语言型式对等。在搭配层面，每组翻译对等词都呈现出具体搭配词选择的多样化。然而，语义相似性并非仅仅体现在搭配词的语义重叠，而是翻译对等词共同期待的语义特征。确立共选型式对等首先要保证语义选择的相似性，值得注意的是，每组型式对等都包括至少一个固有且必要的语义趋向，该语言特征是核心词的必要共选对象，影响共选型式对等的确立。然而，可选语义趋向对判定型式对等没有决定性影响。

在语义韵层面，根据语义韵常态和语义韵力度指数，可以确立不同对应程度的型式对等。共选型式表达相同的态度意义，且语义韵力度指数相近，则可确立为对应程度较高的共选型式对等。如，“平息”和*ASSUAGE*都与表达“强烈

的消极情绪”及“困难”的语义趋向共选，在语义上高度对应；同时，二者在语境中都暗示积极的态度意义，呈现相同的语义韵力度指数，因此以“平息”和*ASSUAGE*为核心词的共选型式是对应程度极高的跨语言对等。需要注意的是，有的型式对等尽管可能表达相同的态度意义，但语义韵力度指数相差较大，下节将讨论共选型式中如果有成分发生改变，共选型式对等是否会发生相应变化的问题。

4. 型式对等与意义移变单位

Sinclair（2010）认为“组成词项的每一个词汇成分在脱离该词项时会产生意义移变”。这里提到的“词项”（lexical item）即由词汇-语法共选结果产生的意义单位，在本研究中我们将其界定为共选型式。Sinclair的论断强调一个意义单位被确立和识别后，其词汇组成成分的语义就已确定，某个成分在该意义单位之外的其他意义单位中会发生语义变化。本节将考察在双语对比视角下共选型式中的某个成分发生变化时是否会影响跨语言对等的确立。

4.1 语义韵的偏离

在扩展意义单位中语义韵是最抽象以及最必不可少的成分（Sinclair 1996）。在本文第3节中，我们发现某个共选型式可能会表达与语义韵常态相悖的态度意义，本节将具体讨论偏离语义韵常态的个例。

由表1可知，“夺取”和*SEIZE*与表达“领土/土地”的语义趋向共现时，都趋于在语境中暗示消极的态度意义，因此可确立为跨语言型式对等。然而，以“夺取”为核心的型式有23%的可能性传递积极语义韵，在MCC中有7例异质用法，我们举其中一例进行讨论。

（14）我们终于打败日本侵略者，夺取了全中国。

例（14）中“夺取”与“全中国”共现时，语境中隐含态度意义的解读要依靠其他语境信息。该句主语we在这里指代抗击日本侵略者的中国人民，因此“夺取全中国”是中国人民的强烈愿望和目标，传递了作者认同和支持的态度。在汉语源语中，该共选行为的交际目的是表达积极认可的态度，如直接把“夺取全中国”译为seize the whole of China。该搭配在英语目标语中可能会被解读成消极反对之意，因为*SEIZE*与“领土/土地”类语义趋向共现时强烈地表达消极语义韵。因此，take over the whole of China可能是更贴切的翻译对等。

我们再以“助长”和*FUEL*为例。表2显示，*FUEL*的第1种型式包括“消极社会现象”的语义趋向和表达“严重性”的消极语义韵，且PSI为0.63。这说明*FUEL*还存在表达积极态度意义的可能性。观察发现，*FUEL*会吸引表达“积极社会现象”类名词，如economic growth, massive development, city boom and

consumption, and the growth in health insurance。语境中暗示了说话者或作者对发展这些社会现象所持有的积极态度，语义韵可解读为“有希望的前景”，这有悖于上述*FUEL*的共选型式表达的消极语义韵。在这种情况下，由核心词*FUEL*、表达“有益社会现象”的语义趋向、表达“有希望的前景”的语义韵组成了一个新的共选型式。如果根据已确立的型式对等，fuel the economic growth对应“助长了经济发展”，然而，该搭配在汉语中较少使用，在这一语境中，“促进了经济发展”是更加贴切的汉语翻译，因为“促进”比“助长”更期待积极的态度意义。

由此可知，与语义韵常态相悖的例子并非是利用语义韵产生修辞效果，语义韵改变的同时，原型式就成为意义移变单位（Sinclair 2010），即不同于原型式的另一型式。在上述例子中，*FUEL*的型式表达消极语义韵，而与“积极的社会现象”共现时传递与之相反的态度意义，从而组成另一型式，在目标语中要求新的型式对等。

同样，“助长”除了与表达消极语义的名词共现之外，还与表达“积极发展”的名词共现，MCC中的例证显示如下：

（15）使其干湿适度，足以**助长**农产物的增加改良，其利益亦

（16）更有深刻之影响，故择交诚为**助长**知识发展之要件也。

（17）它是调动广大职工积极性，**助长**生产力发展的巨大动力。

例（15）—（17）中，与“助长”共现的名词“发展、改良”似乎洗白了其经常携带的消极态度，在语境中表达支持、认同的积极语义韵。由核心词“助长”、表达“积极发展”的语义趋向、积极语义韵组成了另一新的共选型式，但该型式与表2中“助长”的第2种型式在语义韵层面相互背离，也有悖于汉语本族语者对该词使用的正常期待，在汉语本族语的使用中比较低频，是较个别的用法。在这种情况下，*FUEL*不是合适的对应词，因为*FUEL*的共选型式在目标语中会产生消极的语义韵期待，与汉语源语中的态度意义相悖。

综上，与语义韵常态背离的反例并非是为了营造修辞效果的用法，而可能是比较低频的个别用法，语义韵的改变会产生意义移变单位（Sinclair 2010），跨语言对等也随之改变。Morley & Partington（2009）指出，语义韵常态可能会被使用者推翻或利用。我们的证据显示，被推翻或偏离到另一极的语义韵会引起共选型式的改变，跨语言翻译对等也相应改变。根据Hoey（2005）的词汇启动理论，一种语言的使用者对一种型式的态度意义会持有一种固有的预测，这种预测或期待即对该型式的语义韵启动，是在语言社团中长期使用和反复遇见该型式的过程中形成的。因此与该预测相背离的用法是非典型、低频的个别用法，组成的是不同于原型式的新的意义单位，其跨语言对等必然要改变。有悖于语义韵常态的异质个例需要对其共现语境进行充分解读，否则源语中的交际目的在目标语中会被曲解。

4.2　搭配词选择的独特性

3.2.3小节中，我们确立了以“平息”和*ASSUAGE*的共选型式对等，值得注意的是，有个别搭配词可能会打破搭配层面的对应常态，需要进一步讨论。表7显示了“平息”和ASSUAGE的搭配行为。

表7. “平息”和*ASSUAGE*的搭配词对应情况

可译性	ASSUAGE的搭配词	“平息”的搭配词
ASSUAGE＝“平息”	(a) fears, anxieties, discontent, anger	（b）怨恨、不满、不安、怒气、焦躁的心情
ASSUAGE≠“平息”	(a) guilt, loneliness, misery, concern, pride, doubts, grief, conscience	（b）心潮、激动的情感、激动的感情、感情的波澜、感情的万顷波涛、潮水般的思绪

如表7所示，“平息”和*ASSUAGE*的搭配词在语义上最贴切的是（a）组，共同表达的语义特征为：强烈、煎熬、势头凶猛。相比之下，（b）组搭配词在语义上存在差异。与“平息”共现的（b）组搭配词突出情绪或心理的纷乱状态、难以抑制以及如波涛般汹涌的特征，而*ASSUAGE*的搭配词体现出的情绪并非那么猛烈或难以控制。（b）组搭配词在语义上呈现的细微差异可能会影响到搭配层面的对应。例如，根据已确立的型式对等，assuage guilt和assuage loneliness分别对应“平息罪恶感”和“平息孤独感”。然而，这两个搭配在汉语中显得十分蹩脚，原因在于“平息”强烈期待“情绪上的纷乱”和“内心的挣扎”这两个语义特征，其搭配词的选择也局限在该语义特征范围内，而“罪恶感”、“孤独感”并不具备这两个语义特征，因此在搭配层面无法实现高度对应。鉴于此，我们在SJTUPC中再次考察“平息”和*ASSUAGE*的对应例证，发现一则需要个别讨论的例子：

（18）与此密切相关的概念是”和平发展”运动，以　**平息**　外国对中国的军事现代化及其全球化的评议，

is the ‘peaceful development’ campaign to　**assuage**　foreign concerns over China’s military

例（18）中“平息评议”对应assuage concerns 似乎并不妥当。“评议”在该语境中指评判和批评的言论，concerns 主要指对认为重要的事件所持有的关注或忧虑。在汉语源语中，“评议”暗指干涉中国内政的批评性言论，因此assuage criticism是更加贴切的对应表达，而这一搭配在BNC中也出现过。因此，如果目标语中的搭配词在语义上与源语的搭配词不十分贴切，可能会影响语义层面的常

态对应，需要进一步考察才能确立跨语言对等。

“平息”和 *ASSUAGE* 的部分搭配词不对应，因为在“情绪/心情”这一涵盖性语义趋向范围内，“平息”期待语义更加具体的搭配词与之共现，即搭配词选择只能局限于表达“扰乱心绪的强烈情绪”的语义。这一搭配局限对其英语对应词的搭配词选择会产生影响，如果目标语中的搭配词超越了这一语义限制，型式对等在搭配层面将会受到影响，可能会产生如“平息孤独感”这种不恰当的搭配。尽管该搭配在汉语使用中能够传递部分语义信息并且态度意义不发生改变，但该使用十分低频，有悖于汉语本族语者对“平息”用法的惯例期待，会阻碍正常交际。

综上，“平息”和 *ASSUAGE* 与表达“情绪/心情”的语义趋向共现时可确立为共选型式对等。然而，语义趋向不可能实现完全对应。翻译词对可能会呈现相似的语义趋向，但实现语义趋向的搭配词不可能完全重叠。我们发现，型式对等会有部分搭配词实现对应，但一个词语在源语中的期待搭配并非与目标语中对应词的期待搭配完全一致。这一发现与Kübler & Volanschi（2012）的结论一致，他们指出，英语动词 *COMMIT* 及其法语对应词 *COMMETTRE* 都期待相同的语义韵，但却呈现不同的搭配行为特征。在语义趋向不发生变化的情况下，搭配词语义特征的移变并不能改变原型式的语义韵特征，型式对等也不会受到影响，其结果只会产生低频且蹩脚的搭配。

5. 结论与启示

本研究采用Sinclair（1996）提出的扩展意义单位模型，在通用可比语料库中观察3组翻译对等词在语义和语用层面的共选行为，进而识别由核心词、语义趋向及语义韵组成的共选型式，最终确立跨语言对等并分析其影响因素。研究发现，共选型式中最重要的成分语义韵发生改变，就会产生新的共选型式，产生新的意义和功能。换言之，是型式的改变而不是单个词的改变产生了意义移变单位。也正是发现了这一点，Sinclair（2010）将原来提出的“词项”重新界定为“意义移变单位”。其次，语义韵力度指数是描述语义韵极性的有效手段，也是确立跨语言对等的重要参照标准。语义韵力度指数越高，说明在特定语言社团中对这种态度意义的期待程度就越高。背离某一型式的语义韵常态可能是语义和功能均已改变的意义移变单位，形成新的共选型式的低频或个别用法。在翻译过程中，源语中某一型式的语义韵发生改变，目标语中的型式对等也要重新选择，否则会曲解源语中的态度意义。

研究还发现语义趋向是确立跨语言对等的重要因素，然而，共选型式对等在搭配和语义趋向层面无法实现完全对应。共选型式中的固有语义趋向对判定型式对等十分重要，固有语义趋向传递了该共选型式的基本语义信息并引导态度意义

的解读方向。可选语义趋向偶尔伴随核心词出现，与核心词的共选概率较低。尽管可选语义趋向会加强弥漫在语境中的态度意义，但对判定跨语言对等没有起到决定性的重要作用。另一方面，跨语言对等在搭配层面呈现出各种各样的变化，然而核心词会严格限制搭配词的语义特征。研究发现，当跨语言对等呈现语义趋向的相似性时，要保证搭配词具体语义特征的对应，否则会在目标语中产生蹩脚或低频搭配。

研究发现对基于语料库的对比语言学研究和翻译研究以及外语教学具有启示意义。研究揭示了词语层面机械的翻译对等不能准确地表达源语中的意义和功能，只有实现共选型式在语义韵和语义趋向层面的对应才能有助于翻译者和翻译研究者认识跨语言对等的本质，在教学中注重型式对等能够帮助学习者避免翻译中的语用错误，在目标语中传递正确的交际目的和态度意义。本研究是对跨语言型式对等的共时描述，对这一问题还需从历时视角进行探讨，因为语言社团的语义韵启动可能会随时间变化被改变或被颠覆。另一方面，从跨语言视角探讨语义韵的语域特殊性也是从多视角探索语义韵本质的有价值的研究主题。

参考文献

Altenberg, B. 1999. Adverbial connectors in English and Swedish: Semantic and lexical correspondences [A]. In H. Hasselgård & S. Oksefjell (eds.). *Out of Corpora: Studies in Honour of Stig Johansson* [C]. Amsterdam: Rodopi. 249-268.

Bednarek, M. 2008. Semantic preference and semantic prosody re-examined [J]. *Corpus Linguistics and Linguistic Theory*, 4(2): 119-139.

Berber-Sardinha, T. 2000. Semantic prosodies in English and Portuguese: A contrastive study [J]. *Cuadernos de Filología Inglesa*, 9(1): 93-110.

Dam-Jensen, H. & K. K. Zethsen. 2007. Pragmatic patterns and the lexical system: A reassessment of evaluation in language [J]. *Journal of Pragmatics*, 39(9): 1608-1623.

Granger, S. & M. Paquot. 2008. Disentangling the phraseological web [A]. In S. Granger & F. Meunier (eds.). *Phraseology: An Interdisciplinary Perspective* [C]. Amsterdam/Philadelphia: John Benjamins Publishing Company. 27-50.

Hoey, M. 2005. *Lexical Priming: A New Theory of Words and Language* [M]. London/New York: Routledge.

Hunston, S. & G. Francis. 2000. *Pattern Grammar. A Corpus-driven Approach to the Lexical Grammar of English* [M]. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Kübler, N. & A. Volanschi. 2012. Semantic prosody and specialized translation, or how a lexico-grammatical theory of language can help with specialized translation [A]. In A. Boulton, S. Carter-Thomas & E. Rowley-Jolivet (eds.). *Corpus-informed Research and Learning in ESP: Issues and Applications* [C]. Amsterdam/Philadelphia: John Benjamins Publishing Company. 105-135.

Louw, B. 1993. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies [A]. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.). *Text and Technology:*

In Honour of John Sinclair [C]. Amsterdam/Philadelphia: John Benjamins Publishing Company. 157-176.

Morley, J. & A. Partington. 2009. A few frequently asked questions about semantic—or evaluative—prosody [J]. *International Journal of Corpus Linguistics*, 14(2): 139-158.

Partington, A. 2004. "Utterly content in each other's company": Some thoughts on semantic prosody and semantic preference [J]. *International Journal of Corpus Linguistics*, 9(1): 131-156.

Sinclair, J. 1996. The search for units of meaning [J]. *TEXTUS: English Studies in Italy*, 9(1): 75-106.

Sinclair, J. 2004. *Trust the Text* [M]. London/New York: Routledge.

Sinclair, J. 2010. Defining the definiendum [A]. In G.-M. de Schryver (ed.). *A Way with Words: Recent Advances in Lexical Theory and Analysis* [C]. Kampala: Menha Publishers. 37-47.

Stewart, D. 2009. Safeguarding the lexicogrammatical environment: Translating semantic prosody [A]. In B. Allison, P. Rodríguez Inés & P. Sánchez Gijón (eds.). *Corpus Use and Translating: Corpus Use for Learning to Translate and Learning Corpus Use to Translate* [C]. Amsterdam/Philadelphia: John Benjamins Publishing Company. 29-46.

Stubbs, M. 2001. *Words and Phrases* [M]. Oxford/Malden: Blackwell Publishers.

Tognini-Bonelli, E. 2002. Functionally complete units of meaning across English and Italian: Towards a corpus-driven approach [A]. In B. Altenberg & S. Granger (eds.). *Lexis in Contrast: Corpus-based Approaches* [C]. Amsterdam/Philadelphia: John Benjamins Publishing Company. 73-95.

Tognini-Bonelli, E. & E. Manca. 2004. Welcoming children, pets and guests: Towards functional equivalence in the languages of 'Agriturismo' and 'Farmhouse holidays' [A]. In K. Aijmer & B. Altenberg (eds.). *Advances in Corpus Linguistics* [C]. Beijing: Beijing World Publishing Corporation. 371-386.

Wei, N. X. & X. H. Li. 2014. Exploring semantic preference and semantic prosody across English and Chinese: Their roles for cross-linguistic equivalence [J]. *Corpus Linguistics and Linguistic Theory*, 10(1): 103-138.

Xiao, R. & T. McEnery. 2006. Collocation, semantic prosody and near synonymy: A cross-linguistic perspective [J]. *Applied Linguistics*, 27(1): 103-129.

何安平，2013，语料库的短语理念及其教学加工 [M]。广州：广东高等教育出版社。

李晓红，卫乃兴，2012，汉英对应词语单位的语义趋向及语义韵对比研究 [J]，《外语教学与研究》(1)：20-33。

李晓红，2015，双语语料库界面下英汉语义韵对比研究 [M]。上海：上海交通大学出版社。

卫乃兴，2014，对比短语学探索 [M]。北京：外语教学与研究出版社。

通讯地址： 201306 上海市浦东新区沪城环路 999 号上海海洋大学外国语学院

基于语料库的汉语中介语平比句研究

北京外国语大学 华 雨

提要：在汉语作为第二语言习得中，平比句是比较句的一种重要句式，也是语法学习的难点之一。本文旨在通过对HSK动态作文语料库中的平比句进行穷尽性梳理，同时以字数和语体较为匹配的自建中国学生作文语料库为汉语母语参照语料库，梳理并探讨高级汉语学习者的汉语中介语平比句的整体面貌，并从认知的角度讨论其中出现的偏误现象。同时，我们认为，中介语句法类型的研究应当突破偏误研究的范围，着眼于作为独立的中介语语言系统的研究，并从中介语语料的偏误句和非偏误句中发现汉语中介语的特点并揭示其中的语言习得规律。

关键词：语料库、汉语中介语、平比句

1. 研究背景

自《马氏文通》起，汉语比较句大体被分为三类：平比句、差比句和极比句。其中平比句是指“凡象静字以比两端无轩轾而适相等者”，也就是表达两者相似或相等的比较表达形式。自此之后，诸多方家如吕叔湘（1942/1982）、太田辰夫（1958/2003）、贝罗贝（1989）、魏培泉（2001/2009）、谢仁友（2003）、李崇兴（2008）、张赪（2010）、李焱等（2010）、高育花（2016）等都从共时描写和历时梳理等多个方面对平比句的本体研究进行了探讨。而在对外汉语界，针对汉语中介语的平比句研究主要集中附属在比较句的研究框架之下。刘月华等（2001）主要列举了几种高频平比句式，陈珺、周小兵（2005）在比较项的选取和排序研究中涉及了一些平比句教学的取舍问题。肖奚强、郑巧斐（2006）则侧重研究“A跟B(不)一样(X)”中“X”的隐现这一特殊问题。总体而言，大部分比较句研究呈现出差比句繁盛而平比句稍显薄弱的特点，尚未见到以汉语全面描写及分析汉语中介语平比句的研究。

因此，本文基于语料库的范式，旨在建立和标注相关语料信息，穷尽性地梳理汉语中介语平比句的概貌，并以本族语语料库作为参照，发现汉语中介语平比句的语言特点。

2. 研究设计

2.1 研究方法

本研究旨在通过对HSK动态作文语料库中的平比句式进行穷尽性梳理，同时以语体较为匹配的自建中国学生作文语料库为汉语母语参照语料库，梳理并探讨高级汉语学习者的汉语中介语平比句的整体面貌和特征表现，并分析其中出现的偏误现象。

本研究采用中介语对比分析方法(contrastive interlanguage analysis) (Granger 1996；Granger *et al.* 1998/ 2002)。在研究过程中，为避免学习者作文语料库研究过分依赖过度使用（over use）和使用不足（under use）的频率数据的弊端，我们采用基于语料库的自动提取的定量分析和研究者人工甄别的定性分析相结合的研究方法（许家金、许宗瑞 2007）。

2.2 研究问题

本研究试图回答以下几个问题：（1）相对于汉语本族语者，汉语学习者汉语平比句的句法标记类型、句法结构类型的分布是否具有系统性的差异？（2）如果有，汉语中介语平比句的语言表征及成因如何？（3）汉语中介语平比句偏误表现及成因如何？

2.3 语料及工具

本研究选取的汉语学习者语料库为HSK动态作文语料库。该语料库由北京语言大学崔希亮教授主持建设，目前总规模为424万字，收集了1992–2005年母语为非汉语的外国人参加高等汉语水平考试（HSK高等）作文考试的部分作文答卷，作文题目共计10个，作文类型可以分为三类：（1）记叙类，如《我的童年》；（2）议论类，如《我看流行歌曲》；（3）应用文类，如《一封写给父母的信》，每篇作文的篇幅大约在300–500字。

本研究自主建设了中国学生作文语料库（以下简称NATIVE）作为参照语料库。该语料库收集了2006–2016年高考、中考等考场作文、全国四所中学学生平时练习作文、大学生HSK同题作文[1]，共计约320万字。作文题目较为繁杂，但大体也可以分为三类：（1）记叙类，如《和你在一起》；（2）议论类，如《学会宽容》；（3）应用文类，如《写给_____》。作文篇幅大约每篇500–1000字。不同于以往研究中以CCL语料库等通用型语料库或是现当代名家小说等单篇篇幅较长的语料为参照，自建的中国学生作文语料库拥有篇幅相近、语体统一、写作目的

1 南通大学王丽老师提供了她和她的团队所建立的大学生HSK同题作文语料库，特此感谢。

及写作任务类型类似等优点，具有更高的可比性，也更符合提升汉语学习者写作水平的研究目的。

我们使用Power GREP 4.6.3和Edit Pro 7.3.0对自建语料库的文本进行清理，使用UltraEdit-32对其中的平比句框式结构进行甄别和提取。

2.4　平比句的界定与提取

我们认为从语义上看，平比句重在说明比较项之间有无差别，用于比较两种或两种以上具有某种现实联系的不同事物的性质、状态、数量，或者比较不同的行为、动作的程度，旨在权衡对比（李崇兴、丁勇 2008）。从结构上看，平比句主要由比较主体、比较基准、比较标记、比较结果等句子成分要素构成（郑慧仁 2012），如：

我跟你一样高。

在该句中，“我”是比较主体，“你”是比较基准，“跟……一样”是比较标记，“高”是比较结果，其中，比较主体、比较基准、比较标记一般不会省略，比较结果则依据上下文可有可无，如“我跟你一样”在一定语境下可以成立。

据此，本研究在提取平比句时，确定了以下四个步骤：

（1）确定句式标记：我们参考刘月华等（2001），陈珺、周小兵（2005），李焱、孟繁杰（2010），谢白羽（2011），郑慧仁（2012）等研究中列举的平比句式，并联系检索实际，依照比较标记的不同，共得出80种不同的平比句式（详见文末附录），并提取所有语例。

（2）人工去除作为句子成分的平比结构。

（3）人工去除与平比句共享标记类型的比拟句[2]，平比句和比拟句分布具体见表1：

表1.　HSK和NATIVE平比句和比拟句分布表[3]

使用共同标记 总语例数	平比句 语例数	平比句占 总语例数比例	比拟句 语例数	比拟句占 总语例数比例
HSK（2585句）	1204	47.29%	1345	52.71%
NATIVE（3377句）	701	20.76%	2676	79.24%

2 关于汉语平比句和比拟句界定问题的详细论述，请参看高育花、华雨，2016，试论汉语的平比句和比拟句[J],《励耘语言学刊》，2016（2）：69-80。

3 表2.1也体现出汉语中介语使用比较标记（比拟标记）更倾向于表达叙述性的相同、相近或相异的比较，而缺乏修辞性的比拟关系来丰富作文的可读性。

（4）整理语例之后，在HSK中共得到40种平比句式（其中部分是偏误句式，不在附录中），1204个平比句；在NATIVE中共得到36种平比句式，701个平比句。

3. 数据分析及讨论

3.1 平比句标记类型分析

根据李焱、孟繁杰（2010）和语料库浮现的语例实际，将所有平比句依照比较标记的不同划分三类：（1）偏误标记：即HSK动态作文语料库中出现但不合语法的标记类型；（2）单标记：由单一标记构成平比句，句法结构为“比较主体+单标记+比较基准”；（3）双标记：由两个标记构成平比句，句法结构为“比较主体+前标记+比较基准+后标记（+比较结果）”，其中前标记又可分为介词性前标记和动词性前标记，后标记又可分为句法性后标记和词汇性后标记。由上，将平比句句式分为以下类型，具体见表2：

表2. 平比句标记类型和句式一览表

平比句标记类型		平比句句式
偏误标记		比……一样/有差别、差不多、有差别、有差异、那样、跟……比较、与……分别
单标记		等于、不同于、相当于、无异于、像
双标记	双标记A 介词性前标记+句法性后标记	跟/与/和/同……一样/一般
	双标记B 介词性前标记+词汇性后标记	跟/与/和/同……相近/相同/相似/不同/相当/差不多/有差别/有差异/有区别/没两样
	双标记C 动词性前标记+句法性后标记	像/好像/似/如/如同……一样/一般
	双标记D 动词性前标记+词汇性后标记	像/好像/似/如/如同……那样
	双标记E 有……那么	有……那么

再根据表2的分类，对HSK和NATIVE中的平比句标记类型进行统计，具体见表3：

表3. HSK和NATIVE平比句标记类型对比表

平比句比较标记		HSK		NATIVE		Log值	p值[4]
		数量	占总数百分比	数量	占总数百分比		
单标记	单标记A	40	3.32%	60	8.56%	-22.07	0.0000
	单标记B	3	0.25%	33	4.71%	-48.08	0.0000
双标记	双标记A	649	53.90%	207	29.53%	62.42	0.0000
	双标记B	133	11.05%	69	9.84%	0.6116	0.4342
	双标记C	167	13.87%	212	30.25%	-57.09	0.0000
	双标记D	110	9.14%	118	16.83%	-21.08	0.0000
	双标记E	0	0.00%	2	0.29%	-3.99	0.0450
偏误标记		102	8.92%	—	—	—	—

由表3可以看出：

在单标记方面，HSK和NATIVE出现的频率都不高，HSK占3.47%，NATIVE占13.27%，HSK使用单标记的频率显著低于NATIVE，而且HSK绝大部分都是使用单标记“等于”，而NATIVE中不仅不限于“等于”的“于”字单标记，而且“像”类单标记的占比也很高。

在双标记方面，HSK中“双标记A”类型最多（其中“跟……一样”为最高频），占所有平比句的53.90%，远高于其他标记类型。而NATIVE则以“双标记C”类型最多（其中“跟……一样”为最高频），占30.25%，但是紧随其后的“双标记A”类型，也占到29.53%，两种类型几乎不相上下。总体看来，NATIVE平比句的各个双标记类型之间频率差远远小于HSK。也就是说，对于汉语母语者来说，在平比句中，“跟/与/和/同”的介词性前标记和“像/似”等动词性前标记的区别不大。但是，对于汉语学习者来说，他们更倾向于使用“跟/与/和/同”表达

4 这里log值和p值均使用Loglikelihood and Chi-square Calculator 1.0得出取值。该软件由北京外国语大学外语教育与研究中心梁茂成教授开发，主要用于语料库对比研究中对Loglikelihood和 Chi-square的快速计算。本研究取其log值计算并求得p值，以观察两库数值是否具有显著性差异（以$p \leq 0.01$为界）。

平比意义，不太使用“像/似”类标记表达。

那么，回到我们的问题，是不是汉语学习者不太使用“像/似”类动词性前标记呢？我们把“像/似”类标记单独提取出来并作统计，分布如下：

表4. HSK和NATIVE“像/似”类比拟标记分布表

	“像/似”类比拟句标记类型	“像/似”类比拟句数量	“像/似”类比拟句占总比拟句的比例
HSK（1345）	像 像……一样/一般/那么/那样	1325	98.51%
NATIVE（2676）	如 如同 像 如/如同……一样/一般 似……一般 像……一样/一般/那么/那样	1797	67.15%

由表4可知，HSK中高达98.51%的比拟句都是由“像/似”类标记构成的，NATIVE中“像/似”类标记比拟句虽然也占比拟句的67.15%，但还有不少比拟句是由其他句式来表达的。也就是说，汉语学习者并不是不使用“像/似”类标记，而是更倾向于使用“像/似”类标记表示比拟句而非平比句。

对于汉语学习者的这种倾向，我们认为有两方面的原因：

第一，学习者更倾向于根据标记意义来判定句式意义。

但从标记本身来看，语法化程度较高的介词如“跟/与/和/同”，功能上主要就是表达引入对象，在比较句中引入比较基准也十分顺理成章，以之为标记的句子以等比语义为主；而语法化程度较低的“像/似”则含有浓厚的比喻实义（李剑锋 2000），以之为标记的句子以比拟意义为主，也就出现了表3.2中所见的两类标记的差异。

第二，教学大纲设置和教材安排的影响。

我们考察了六份教学语法大纲：《对外汉语教学初级阶段教学大纲（语法大纲部分）》《对外汉语教学语法大纲》《汉语水平等级标准与语法等级大纲》《高等学校外国留学生汉语教学大纲(长期进修)——语法项目表》《高等学校外国留学生汉语言专业教学大纲》《中高级对外汉语教学等级大纲 词汇·语法》和五本经典的通用型汉语教材《初级汉语课本》《桥梁——实用汉语中级教程》《汉语教程》《新实

用汉语课本》《博雅汉语》[5]，发现其中对于平比句的安排设置具备以下特点：（1）所有语法大纲和教材设置的甲级或者初等语法都是“跟……（不）一样”，也就是说“跟……一样”很可能成为学习者最早习得并接受的表达平比意义的句式。（2）所有语法大纲和教材均没有收录“像……（不）一样”的平比句式，仅《汉语水平等级标准与语法等级大纲》将“像……这么/那么”纳入乙级语法。（3）语法大纲将“像”作为表达比拟意义的动词，教材多以“像……似的”作为比拟句式的表达。

因此，在我们的考察范围内，无论是教学大纲还是教材，都在隐约地表现着“跟/与/和/同”等介词性前标记作为平比句式，“像/似”等动词性标记表达作为比拟句式，不免影响到学习者的句式使用分布。但是，由NATIVE反映出这两者在实际语料中表达平比的频次并无明显差异，因此，教学语法也不应忽视这样的语言事实。

最后，除了“跟/与/和/同”等介词性前标记和“像/似”等动词性标记两种主要平比句标记之后，还有双标记E“有……那么/这么”。不过，“有”字平比句在HSK和NATIVE中出现的频率都极低，HSK几乎没有语例，NATIVE也只有2例。作为肯定形式的“有……那么”平比句非常少，但是作为否定形式出现的“没有……那么”差比句虽然不属于差比句的高频句式，但还是多出一些，据统计在HSK中出现42例，NATIVE中出现40例6[**]。有意思的是，前文提及的六种大纲中有5种安排了“有”字句平比句的教学，但只有2种大纲安排了差比句的教学；五本教材中只有2种安排了“有”字句平比句的教学点，另外3种则完全没有设置“有”字句的教学点，具体如下：

表5. 六种教学大纲及五种教材“有”字句教学点安排表

教学大纲/教材	句法结构	“有”字句教学点：平比/差比
1《对外汉语教学初级阶段教学大纲》	A有/没有B（这么/那么）	平比/差比
2《对外汉语教学语法大纲》	A有B（那么）（这么）	平比
3《汉语水平等级标准与语法等级大纲》	有……那么	平比
4《高等学校外国留学生汉语教学大纲》	有……那么	平比

（待续）

5 我们没有选取近期的材料，是因为近期的大纲和教材，不会对当年HSK中的语言材料产生影响。所以我们选择的六种语法教学大纲和五种汉语教材都是在HSK作文库建设之前出版的。

6 根据我们的统计，HSK差比句共计2785句，“没有……那么”句占总数的1.51%，NATIVE差比句共计897句，“没有……那么”句占总数的4.46%。所以，“没有……那么”句肯定不属于差比句的高频句式。

（续表）

	教学大纲/教材	句法结构	“有”字句教学点：平比/差比
5	《高等学校外国留学生汉语言专业教学大纲》	没有……（那么/这么）	差比
6	《中高级对外汉语教学等级大纲 词汇·语法》	无	无
1	《初级汉语课本》	无	无
2	《桥梁——实用汉语中级教程》	无	无
3	《汉语教程》第二册第1课	“有……+形容词”	平比
4	《新实用汉语课本》第28课	“有”“有……这么”	平比
5	《博雅汉语》	无	无

我们可以看到，教学大纲和教材都侧重“有”字句平比句的教学，甚至它们都将其放在大纲的甲级语法或是教材的前几册，但是这样的安排并不符合汉语母语者使用的实际情况。“有”字句平比句，汉语母语者使用得少，学习者使用得更少，而“有”字句差比句具有较高的使用频率却在教材中没有得到体现。所以，大纲和教材可能在句式的选择和侧重上应当考量实际的使用频率。

3.2 平比句句法类型分析

中介语的句法状况是中介语系统性研究的重要观察窗口。朱曼殊、缪小春（1990）就曾指出，句法的发展通常可以从两方面进行评定和分析，一是句子所包含的最基本的意义单位的数量；另一个更为主要的方面是句子结构的多样性和复杂性。

而在本研究中，平比句所包含的意义单位数量基本统一，如果是单标记结构，平比句结构为：比较主体＋单标记＋比较基准。如果是双标记结构，平比句结构为：比较主体＋比较前标记＋比较基准＋比较后标记＋（比较结果）。因此，从句子结构的完整性和复杂性上可以对HSK和NATIVE平比句的句法情况有所发现，而在这其中，平比句的比较主体和比较基准的句法类型较为单一，以光杆名词和并列式名词为主，而比较结果的句法结构则非常丰富，其出现的有无、结构的层次性和内部构成的复杂度均与整个句子的复杂度紧密相关。

我们以句子直接成分分析法将比较结果进行句法分析，并按照句法结构的层级数量分为五大类，具体见表6：

表6. 平比句比较结果的句法类型分类表

比较结果句法分类	比较结果句法结构
第1类：无	无
第2类：光杆结构或并列结构	光杆动词/并列式动词 光杆形容词/并列式形容词 四字成语[7]/并列式四字成语
第3类：两层结构	定中结构/动宾结构/动补结构/状中结构 并列式介宾结构
第4类：三层结构及简单小句	程度副词+动补结构/动宾结构 程度副词+光杆形容词/光杆动词 程度副词+使令兼语结构 介宾结构+光杆形容词/光杆动词 介宾结构+动补结构/动宾结构 连谓结构/使令兼语结构/系表结构 简单谓语小句/简单主谓小句
第5类：复杂小句及复句结构	复杂谓语小句/复杂主谓小句 条件复句/因果复句/转折复句

我们将HSK和NATIVE中可能带有比较结果的双标记平比句提取出来（因为单标记平比句肯定没有比较结果，所以不予分析），其中HSK有1299句，NATIVE有447句，并将它们的句法结构进行了分类统计，见表7：

表7. 比较结果的句法结构类型分布表

比较基准句法分类	HSK		NATIVE		Log值	p值
	数量	占总数百分比	数量	占总数百分比		
第1类：无	784	60.35%	135	30.20%	64.61	0.0000
第2类：单层结构	117	9.01%	61	13.65%	-6.59	0.0103
第3类：两层结构	170	13.09%	73	16.33%	-2.43	0.1192
第4类：三层结构	64	4.93%	110	24.61%	-108.70	0.0000
第5类：四层结构及小句	164	12.63%	68	15.21%	-1.63	0.2022

7我们认为，四字成语在认知加工上通常作为一个构式进行记忆和使用，与光杆结构近似，故并入。

由表7可以看出：

（1）HSK占比最高的情况是不加比较结果，即第1类“无”，占比60.35%，NATIVE占比最高的类型也是不加比较结果，占比30.20%，但是HSK的倾向程度明显更高（$p < 0.01$）。也就是说，学习者倾向于不使用比较结果，这可能是为了降低句法难度，减少偏误。

（2）在出现比较结果的语例中，HSK占比较高的是第3类和第5类句法结构，NATIVE占比较高的是第3类和第4类句法结构。一般来说，从第1类到第5类的整体趋势应该是，句法层数越来越多、句法复杂性也随之越来越高。因此，在带有比较结果的句子中，NATIVE明显倾向使用较为复杂的第4类句法结构，HSK倾向使用较为简单的第3类句法结构比较好理解。但是，HSK使用第5类句法结构的频率仅次于第3类，则需要进一步讨论。

我们认为，这与第5类句法结构的判断有关。如：大家在一起，就像朋友一样，个体人格独立了，“代沟”就不成问题了。这里“个体人格独立了，‘代沟’就不成问题了”，既可以视为平比句的比较结果，也可以认为是在语义上已经脱离比较句，单独成句。也就是说，学习者在输出这个句子时，很有可能是将其作为另一个语块进行单独加工，而不是融合在平比句中，构成一个复杂的句法结构。

总体而言，HSK更倾向于不出现比较结果以及使用简单的句法结构，这显示出HSK的句法结构复杂程度不及NATIVE。

3.3 平比句偏误类型分析

我们结合HSK自身偏误标注系统结合人工筛选，共计发现HSK平比句偏误句181句。我们按照偏误位置将其分为比较标记偏误、比较结果偏误、比较主体偏误、比较基准偏误、比较句副词偏误和比较句语序偏误，频次由高到低排序，具体见表8：

表8. HSK偏误句偏误位置分布表

	偏误类型	语例数	占全部偏误百分比
1	比较标记	111	61.33 %
2	比较结果	25	13.81%
3	比较句语序	22	12.15%
4	比较主体	14	7.73%
5	比较基准	14	7.73%
6	比较句副词	10	5.52%

由于比较结果、比较主体和比较基准的偏误，以自身内部结构性错误为主，与平比句的句法或语义关系不大，而且类型非常分散，所以我们不作单独讨论。我们集中来看三种高频偏误：比较标记偏误、比较语序偏误和比较副词偏误。

3.3.1 平比句比较标记偏误

由表3.7可知，比较标记是偏误率最高的偏误位置，共计111例，具体又可分为三类：比较标记错误、比较标记残缺、比较标记多余，频次由高到低排序，具体见表9：

表9. 平比句比较标记分布及举例：

	偏误类型	语例数	占比较标记偏误百分比	例句
1	比较标记残缺	71	63.96%	我也是{CQ和}那些人一样很喜欢听流行歌曲。
2	比较标记错误	26	23.42%	一开始新生活就比以前不一样了。
3	比较标记多余	14	19.82%	对吸烟者来说，吸烟就是像{CD跟}喝酒一样，一个为离脱生活压力的办法。

由表9可知，比较标记残缺的偏误最多，其次是比较标记错误，再次是比较标记多余。对于比较标记偏误，之前的研究者已有诸多讨论，大多认为比较标记的偏误只是初级阶段的偏误，会随着习得时间的推移而消失（肖小平 2004）。但是我们发现，在作为中高等汉语水平作文库的HSK中，比较标记偏误仍然存在较高的出现频率，甚至所占比例最高。可见，这种偏误似乎并未随着习得而消失，反而极易出现僵化现象。而究其成因，一般从母语负迁移、目的语规则的泛化、教学失误和交际策略等方面进行阐释（姜桂荣 2009），这些都可能成为引起偏误的重要原因，我们不再进行赘述。不过，我们通过仔细观察语料库中析出的偏误语例，发现在偏误数量最高的“比较标记残缺”一类中，以判断句中的比较标记前标记残缺是最常见的偏误类型，而这其中，比较前标记与“是”字判断句的频繁共现引起了我们的关注。如：

我也是{CQ和}那些人一样很喜欢听流行歌曲。

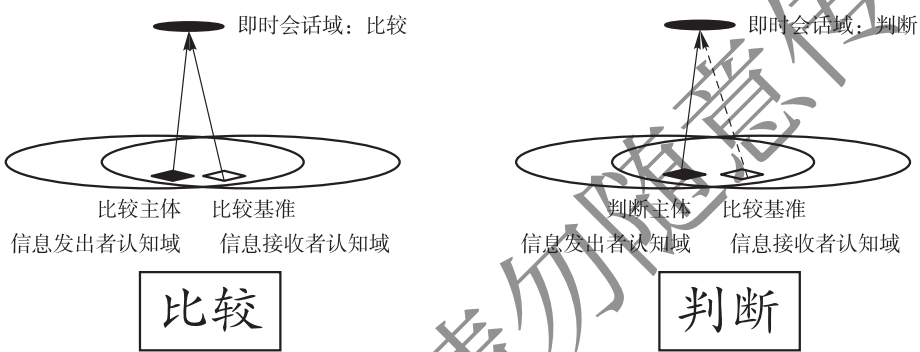
他们愿意别人也是{CQ跟}自己的父母一样，为自己衷心地照顾自己。

这些句子都缺少比较前标记而又同时冗余判断动词“是”[8]，而且两者同时出

8 这里两个例句中的“是”是否一定冗余，根据我们对母语者的询问调查，多数人认为冗余，少数人认为不冗余，没有绝对肯定的答复。不过，如果不是强调或加重语气的需要，母语者一般倾向不加“是”。

现的句子的比较结果都相对复杂，如“很喜欢听流行歌曲”“为自己衷心地照顾自己”，而在这样的句子中，产生比较意义的主要部分“和那些人一样”“跟自己的父母一样”，从句子结构上看都可被认为是状语成分，真正的谓语中心语是“很喜欢听流行歌曲”和“为自己衷心地照顾自己”，所以从句法功能上看并不需要加上判断动词“是”。但是，学习者明显倾向加上“是”，我们认为，语言本身存在的认知概念交叉也可能引起比较标记的偏误。

刘焱（2005）曾指出比较范畴和判断范畴之间边界不明晰、有一定的交合之处。根据刘焱（2005）和吴庸（2015）的描述，我们将比较范畴和判断范畴的认知图示改进并绘制如下：



在交际过程中，信息发出者认知域和信息接收者认知域存在重合领域，比较或者判断都需要在重合领域中进行。其中在比较范畴中，比较主体和比较基准同时被双方激活，如“我比爸爸跑得快”一句中，首先激活了“我”和“爸爸”两个比较项，进入一个临时会话域（或者交际域），从而在某一隐含的比较点“跑步的速度”上产生比较的意义。而在判断范畴中，从显性会话项来看，看似只激活了判断主体一类，如“西施是个很漂亮的美女”一句中，只激活了“西施”，但是实际上在即时会话域中也激活了隐含的比较基准，即“和其他女人比起来，西施很漂亮。”所以，当我们进行判断时，实际上也在进行隐性的比较；而当我们进行比较时，实际上也是一种更加具体的判断。

现在来看HSK中的比较前标记残缺句，当学习者想要表达比较意义时，他们会让其中隐含的判断意义显化，因此倾向于使用习得最早也最为熟悉的“是”字句，而“是”字一旦出现并位于比较主体之后，便占据了本属于比较前标记的位置，而比较后标记“一样”、“一般”等因为“比较”的实义性很强，通常不易丢失省略，于是就产生了不少类似于“我也是{CQ和}那些人一样很喜欢听流行歌曲”这样，冗余判断标记“是”字而缺少比较前标记的比较标记残缺的偏误句。当然，我们也不排除，有时只是学习者并没有真正理解比较成分在比较句中的句法作用，认为必须加上“是”，比较句才拥有了谓语动词，才能保证句子没有错

误，实际上这样反而画蛇添足了。

3.3.2 平比句语序和副词偏误

平比句的语序偏误常常与副词使用偏误重合，故在此合并讨论。语序偏误中，主要以两种副词位置的偏误居多：

（1）否定副词位置错误，如：前几天是父母节了，不{CJX}跟一般的家庭一样，我们和爸妈不一起住，只能早上打了个电话。

（2）程度副词位置错误，如：我突然觉得我还是她的女儿，不放弃自己的意见，这件事完全{CJX}跟她一样。

卢福波（1996）认为否定语义“跟……一样”句中，从语义上说，前后两个比较项已出现了不一致，因此句义往往要求句子在比较的前后进行不同点的说明。但是如果“不”位于“跟”之前，由于“不”的否定阻断作用，使“一样”后不可能再接描述比较点的词语。对于学习者来说，他们习惯由肯定式直接推导否定式，造成了他们目的语知识的过度泛化。程度副词位置的错误也是类似的类推结果导致的偏误。不过，有意思的是，如果否定副词和程度副词同时在平比句里出现，则否定副词“不”可以置于比较标记之前，程度副词则还是要紧邻比较后标记“一样”，如“我们不跟他们完全一样，我们有自己的特点。”我们认为，这时，“不”的否定语义指向不再是“一样”，而是程度副词“完全”，所以“不”并不会起到对比较意义的阻断作用，也就可以像其他句子一样，放在状语中心语之前或者谓语中心语之前了。

4. 数据分析与结论

我们通过对HSK动态作文语料库和自建中国学生作文语料库的梳理与对比，从比较标记、比较结果等方面进行了探讨。首先，从标记使用上看，汉语中介语更倾向于使用语法化程度较高的、介词性句法标记“跟/与/和/同”等为比较标记的平比句，更倾向于使用语法化程度较低的、动词性句法标记“像”为比拟标记的比拟句。其次，从比较结果的有无上看，学习者比本族语者更倾向于不使比较结果出现，因为比较结果偏误率较高，省略它可以在一定程度上回避偏误的产生。从比较结果的句法类型上看，本族语者较多地使用“介宾结构+动宾短语”等第4类的句法结构；而学习者较多地使用“动宾短语”等3类句法结构。同时，学习者也较多地使用“主谓小句”等第5类复杂小句作为比较结果，我们认为这种反常现象并不能证明学习者会使用更高级的句法结构，而是与学习者语言输出时的语块化的认知加工思维有关。同时，我们也考察了HSK中的平比偏误句。从偏误类型上看，共有比较标记、比较结果、语序、比较主体和基准、副词等6种偏误类型，其中以“比较标记偏误”最多。平比句的偏误存在一定的僵化现象，不能

简单地认为这种偏误会随着学习时间的推移而自然消失。

由此，我们可以发现，汉语中介语平比句具有许多自身的特点：首先，从句式选择上看，以“跟……一样”为典型句式，且比在汉语母语者中的典型性更强，其他句式从句法结构和使用语境上都有向典型句式靠近的趋势；平比句内部各成分中以“比较结果”的情况最为灵活，在汉语中介语中，比较结果经常被省略，或是经常可以独立于比较句主体部分，单独成句和表义；平比范畴和判断范畴之间有认知概念的交叉问题，可能从概念上使得汉语中介语的平比句中常常会出现表示判断范畴意义的判断动词“是”。

5. 余论

实际上，在20世纪70年代时，Selinker就指出中介语是一个独立的语言系统，具有自身的系统性。这种观点已经被学界所公认，但是就目前的中介语研究而言，主要还是集中在对一些语音、词汇、语法等方面的偏误研究上，但以大规模语料全面描写中介语各方面特点的系统性的研究还不多见（肖奚强、黄自然 2013），肖奚强（2011）更是指出“时至今日……，中介语的系统性仍然是一种假设而缺乏实证研究的支持。”因此，我们期待借助语料库语言学与第二语言习得研究领域相结合的CIA，通过本族语者语料库和中介语语料库的比较（NL vs.IL），观察和提取出关键语言特征（critical features），从而对中介语的平比句句式做出全面而细致的系统性描写和阐述，而不是仅仅将关注点放在偏误分析上。

另外，在对照大纲和教材平比句教学点之后，我们也发现，习得的难度也不能简单以教材和标准中的先后顺序而定，学习者最后习得的情况也会有所不同。有时会在教材和教学的引导下，形成强烈的使用惯性和使用倾向，如“跟……一样”大量使用于平比句，“像……一样”大量使用于比拟句；有时学习者又完全不“服从”教材的安排，如“有”字平比句，因为在学习了更为典型、使用更便捷的“跟……一样”句之后，即使继续教学“有”字句，学习者也不会经常使用“有”字句了。而且，大纲和教材也应该将知识点在本族语者中的使用频次高低作为编纂的重要依据，“有”字平比句在母语者中使用频率也非常低，因此可能需要推迟它在教学阶段的位置。

诚然，本研究存在诸多不足。例如，对于平比句的描写仅仅选取了句法标记和句法结构类型两个点，不足以涵盖句式描写的各个方面。而其中对于句子的句法复杂度研究只考量了比较结果的句法结构，实际上可能还有待从其他角度着手，对学习者平比句的使用进行多因素分析，从而得到更加系统和完善的考察结果。

参考文献

Granger, S. 1996. From CA to CIA and back: An integrated approach to computerized bilingual and learner corpora [A]. In K. Aijmer *et al*.(eds.). *Languages in Contrast: Text-based Crosslinguistic Studies* [C]. Lund: Lund University Press. 37-51.

Granger, S. (ed.). 1998. *Learner English on Computer* [C]. London: Longman.

Granger, S. 2002. A bird's eye view of learner corpus research [A]. In S. Granger *et al.* (eds.). *Computer Learner Corpora, Second Language Acquisition and Foreign Language Teaching* [C]. Amsterdam/Philadelphia: John Benjamins Publishing Company. 3-33.

Peyraube, A.（贝罗贝）. 1989. History of the comparative construction in Chinese from the 5th century B. C. to the 14th century A. D. [A]. In *Reprinted Proceeding on the Second International Conference on Sinology* [C]. Taipei: Academia Sinca.Vol. 2: 589-612.

陈　珺、周小兵，2005，比较句语法项目的选取和排序[J],《语言教学与研究》(2)：22-33。

高育花，2016，元代汉语中的平比句和比拟句[J],《长江学术》(3)：95-105。

高育花、华　雨，2016，试论汉语平比句和比拟句[J],《励耘语言学刊》(2)：69-80。

耿　直，2012，基于语料库的比较句式“跟”“有”“比”的描写与分析[D]，北京大学博士学位论文。

姜桂荣，2009，基于“HSK动态作文语料库”的“比”字句习得研究[D]，北京语言大学硕士学位论文。

李崇兴、丁　勇，2008，元代汉语的比拟式[J],《汉语学报》(1)：2-10。

李剑锋，2000，“跟X一样”及相关句式考察[J],《汉语学习》(6):72-77。

李　焱、孟繁杰，2010,《汉语平比句的语法化研究》[M]。南京：南京大学出版社。

梁茂成，2009，词性赋码语料库的检索与正则表达式的编写[J],《中国外语教育》(2)：65-81。

梁茂成，2012，语料库语言学研究的两种范式：渊源、分歧及前景[J],《外语教学与研究》(3)：323-335。

刘月华、潘文娱、故　韡，2001,《实用现代汉语语法》[M]。北京：商务印书馆。

刘　焱，2005,《现代汉语比较范畴的语义认知基础》[M]。北京：学林出版社。

卢福波，1996，汉语比较句中肯定式与否定式的不对称现象[A],《国际汉语教学讨论会论文选集》[C]，北京：北京语言大学出版社。

吕叔湘，1942/1982,《中国文法要略》[M]。北京：商务印书馆。

马建忠，1898/1989,《马氏文通》[M]。北京：商务印书馆。

太田辰夫，1987,《中国语历史文法》[M]。北京：北京大学出版社。

魏培泉，2001，中古汉语新兴的一种平比句[J],《台大文史哲学报》(54)：45-66。

魏培泉，2009，中古汉语时期汉文佛典的比拟式[J],《台大文史哲学报》(70)：29-53。

吴　庸，2015，汉语隐性比较构式的认知研究[J],《外语学刊》(2)：29-35。

夏　群，2009，汉语比较句研究综述[J],《汉语学习》(2)：58-64。

肖小平，2004，越南留学生汉语比较句偏误分析及习得顺序考察[D]，广西师范大学硕士学位论文。

肖奚强，2011，汉语中介语研究论略[J],《语言文字应用》(2)：109-115。

肖奚强、黄自然，2013，韩国学生中介语各句长句子定语复杂度发展研究[A],《第二届汉语中介语语料库建设与应用国际学术讨论论文选集》[C]，北京：北京语言大学出版社。

肖奚强、郑巧斐，2006，“A跟B（不）一样（X）”中“X”的隐现及其教学[J],《世界汉语教学》(3)：113-120。

谢白羽，2011，面向对外汉语教学的比较句研究[D]，华东师范大学博士学位论文。

谢仁友，2003，汉语比较句研究[D]，北京大学博士学位论文。

解植永、王　建，2011，韩国留学生习得汉语比较句的偏误分析[J],《云南师范大学学报》(对外汉语教学与研究版)(5)：8-12。

许家金、许宗瑞，2007，中国大学生英语口语中的互动话语词块研究[J],《外语教学与研究》(6)：437-443。

于立昌、夏群，2008，比较句和比拟句试析[J],《语言教学与研究》(1)：14-18。

张　赪，2010,《汉语语序的历史发展》[M]。北京：北京语言大学出版社。

郑慧仁，2012，东北亚语言比较标记的类型学研究[D]，北京大学博士学位论文。

朱曼殊、缪小春，1990,《心理语言学》[M]。上海：华东师范大学出版社。

附录

汉语常见平比句式：

1 跟……(不)一样
2 和……(不)一样
3 与……(不)一样
4 同……(不)一样
5 像……(不)一样
6 如/如同……(不)一样
7 比……(不)一样
8 ……一样
9 跟……(不)一般
10 和……(不)一般
11 与……(不)一般
12 同……(不)一般
13 像……(不)一般
14 如/如同……(不)一般
15 似……(不)一般
16 ……一般
17 跟……差不多
18 和……差不多
19 与……差不多
20 同……差不多
21 跟……相似
22 和……相似
23 与……相似
24 同……相似
25 跟……似的
26 和……似的
27 与……似的
28 同……似的

29 跟……有区别/没区别/没有区别
30 和……有区别/没区别/没有区别
31 与……有区别/没区别/没有区别
32 同……有区别/没区别/没有区别
33 跟……有差别/没差别/没有差别
34 和……有差别/没差别/没有差别
35 与……有差别/没差别/没有差别
36 同……有差别/没差别/没有差别
37 ……有差别/没差别/没有差别
38 跟……有差异/没差异/没有差异
39 和……有差异/没差异/没有差异
40 与……有差异/没差异/没有差异
41 同……有差异/没差异/没有差异
42 比……有差异/没差异/没有差异
43 ……有差异/没差异/没有差异
44 有……那么/这么/那样/这样
45 没有……那么/这么/那样/这样
46 像……那么/这么/那样/这样
47 ……同于……
48 ……不同于……
49 ……异于……
50 ……等于……
51 ……相当于……
52 ……抵得/抵得上/抵得过……
53 ……赶上/赶得上……
54 ……，同样的(地)，……
55 ……像……
56 ……宛然……一般
57 ……宛如……
58 ……好似……
59 ……如/如同……
60 ……没(有)两样

通讯地址： 100089 北京市海淀区西三环北路2号北京外国语大学中国语言文学学院

高中英语写作中定冠词THE的共选特征研究*

扬州大学 陆 军 官丽丽

提要：本文以共选说为理论框架，以英语本族语学生写作语料为参照，考察中国高中生英语写作中定冠词THE的使用特征及其影响因素。数据表明：与英语本族语学生相比，中国高中生作文中定冠词THE主要使用在前置限制、修饰结构中，以THE+N类型为主，但在意义和功能上的误用倾向比较明显。分析表明：1）本族语学生作文中，英语定冠词倾向于与特定的语法范畴共现构成多种类联接，用于表达特定的意义或功能，体现了形式、意义和功能的共选；2）中国高中生懂得定冠词与某些特定语法范畴共现，但是缺乏相应的意义和功能知识；3）中国学习者的定冠词共选特征与其频繁接触THE+N/NP等形式、汉语中缺乏相应的冠词系统以及显性语法教学等因素有密切关联。上述发现对进一步丰富英语冠词研究、促进基础教育阶段的英语功能词教学有一定启示意义。

关键词：英语定冠词、高中英语写作、类联接、意义和功能、共选

1. 引言

英语冠词是使用频率最高的词类，主要包括定冠词THE，不定冠词A/AN和零冠词（Master 2003，2012）。从形式上看，冠词总是位于限定词短语DP的D位置（Abney 1987；Longobardi 1994）；从意义或功能上讲，冠词主要用于表达定指性等信息（Ionin 2003）；从跨语言视角看，英语冠词特别是定冠词THE与很多二语学习者的母语（如汉语）之间的词语对应关系不确定（Chung 2000）。尽管其形式特征看似并不复杂，但所表达的意义和功能却非常复杂，是最难掌握的语法部分，也是最后才能完全习得的部分之一（Master 1990）。再加上冠词在交际中的高频使用，所以一直是二语学习的重点。

由于英语冠词系统非常复杂，相关研究的理论视角或研究体系有很大差异。例如，Bickerton（1981）主要关注“是否特指”（±specific referent）和“是否假定听话人知晓”（±assumed known to the hearer）两组特征。实际使用中，这两组特征并不完全独立，往往会有交叉。例如，在某些情形中，“特指”蕴含着或伴

*本研究为国家社会科学基金项目“语料库驱动的二语隐性、显性知识调查与实验研究”（15BYY067）和扬州大学教改课题“大数据环境下英语自主学习平台研发”的阶段性成果。

随着“听话人知晓”，但孰轻孰重则取决于说话者的交际意图。Gundel, Hedberg & Zacharski（1993）通过已知性层级（Givenness Hierarchy）利用相互关联的六种认知情形反映出具体用法特征（即聚焦→已启动→熟悉的→可以独一无二的确定→指称→可以按类型确定），与自然语篇中的指称语（referring expressions）用法更为切合。不过，这些认知情形可能会出现非常复杂的组合关系，在实际研究中并不容易操作。Master（1990, 2003）将整个英语冠词体系分为两个主要特征，即“分类”（classification）和“定指”（identification），体现出语言使用中冠词（A / AN和THE）在形式与功能之间存在一定的对应关系。相比之下，该体系具有较强的概括性和可操作性。就定冠词THE而言，Master（1990, 2003）指出，当某一名词用于表示“再次提及”（subsequent mention）时，必须要与定冠词THE连用，以表明其不再是新出现词；与等级形容词（ranking adjectives，R-ADJ）共现时，主要分为三类：最高级（superlative），顺序关系（sequential）和唯一性（unique）。其实，最高级和顺序关系可以看作唯一性的特殊类型。例如，*the most beautiful/the next/the only city*中的*city*为特指，主要突出唯一性。根据Master（1990, 2003），当定冠词THE用于表示共享知识（shared knowledge）时，主要包含两种情况：“普遍性”（universal）和“区域性”或“逻辑性”（regional/logical）。例如，使用*the moon/the school/the window*等表达主要为了突出交际双方共同知晓的事物，是已经确定或定指的。在后置修饰用法（post-modification）中，THE可与后置定语共同对目标名词加以限定，使其变得更加确定、凸显，如*The water in this glass is dirty*等。当后置限定“*of*-结构”用来表述中心名词（head noun）时，如*the length of the room*，需使用定冠词THE对中心名词限定使其进一步确定或凸显。

上述文献说明，英语冠词所表达的意义和功能非常复杂。相应地，在二语习得研究中也备受关注。相关研究主要从学习者的语言水平、学习和使用策略、母语影响以及中心名词的语义特征等角度探讨英语冠词的使用特征。研究发现，中国学习者使用英语冠词有很大困难，出现不同类型的错误，与汉语中缺少明确的冠词系统有着密切关联（如蔡金亭、吴一安 2006；李景泉、蔡金亭 2001；闫丽莉 2003；朱叶秋 2003；朱叶秋、文秋芳 2008等）。造成二语冠词难以习得的另一个原因在于其较为复杂的句法、语义特征（如Lardiere 2004, 2009；常辉、赵勇2014；戴炜栋、韦理2008；邵士洋、吴庄2017；周保国2007等）。

尽管EFL学习者冠词使用的缺失或失误与母语影响有着密切的关联，但中国学习者并没有因为汉语中缺少冠词系统而少用定冠词，有时反而倾向于过度使用。例如，朱叶秋（2003）等发现中国英语学习者在汉语使用零冠词的语境中也会使用THE或A/AN。不过，有些研究则认为中国学习者可以完全习得冠词的指称信息（如常辉、赵勇 2014；戴炜栋、韦理 2008；于善志、苏佳佳 2011等）。由此可

见，除了母语影响以外，另有重要因素影响着中国学生的冠词习得。例如，邵士洋、吴庄（2017）指出，学习者除掌握冠词的指称信息外，还需习得相关句法知识。从共选（Sinclair 1991, 1996）视角看，英语冠词总是出现在特定的句法结构中，实现“指称”等特定的意义和功能，应当是形式、意义和功能共选的产物。

此外，研究方法的差异也会对得出的结果产生一定的影响。常用研究方法包括强制选择任务（forced choice）（如朱叶秋 2003等）、语料库数据分析、以及诱导数据分析（如邵士洋、吴庄 2017等）等。其中，强制选择任务往往要求受试者根据语境从THE、A / AN或“空缺”选项中选出最佳答案。此类任务明确要求受试者考虑冠词的用法，倾向于调用冠词的显性语言知识。然而，在真实语言表达中隐性语言知识起主要作用（Ellis 2011；Ellis 2006）。由此可见，强制选择任务能够反映学习者具有哪些冠词知识，但不能代表其在交际中的典型知识。与之相比，语料库数据则有助于反映真实语言使用中的典型语言知识（Barlow 1996；Barnbrook 1996；陆军 2017；陆军、卫乃兴 2014）。再者，学习者频繁使用英语冠词也为探讨冠词习得特征提供了重要证据。学习者语料库数据能够直观地反映出“过多使用”、“错误使用”和“过少使用”三种定量特征（Granger 1998）。其中，“过少使用”可能是由于学习者缺乏相应的语言知识或能力而回避使用相应的语言形式，也可能是由于语料库构成材料的特殊性使得相关语言形式未能充分体现，一定程度上妨碍了相关二语使用特征的研究。为此，一些研究采用诱导数据来克服这一缺陷。不过，诱导数据只是用于反映“可能性”的语言知识（Thráinsson et al. 2007），而语料库数据则是反映“典型”语言知识的首选材料。由此可见，研究方法上的差异对研究结果会产生不同影响。

基于上述分析，二语冠词使用特征及其影响因素等方面仍然有诸多问题迫切需要解决。本研究以共选说为理论框架，基于语料库数据考察中国高中英语学习者使用定冠词THE的特征。主要选题理据如下：首先，语言表达中的词语总是趋向于出现在特定的结构中；而这些结构也总是趋向于与特定的词语共选，实现特定的意义和功能（Sinclair 1991, 1996；Stubbs 1996, 2009）。英语定冠词THE也不例外，总是出现在NP短语中的特定位置上，如THE+N, THE+N1+N2, THE+ADJ+N和THE+R-ADJ+N等（参见Master 1990, 2003）。这些形式都是THE与特定语法范畴的共现，其本质上为类联接（colligation）（参见Hoey 2005；Sinclair 1991，1996；卫乃兴 2002）。THE与这些结构共现表达“指称”、“重复提及”以及“共享知识”等功能。其次，在多数情况下，冠词的具体用法由上下文语义决定，并不能够为现成的规则所囊括（Ellis 1994），二语冠词使用尤其如此。学习者在英语表达中高频使用冠词，高中学生也不例外，相应的语料库数据有助于通过上下文来揭示高中学生的真实冠词使用特征。再者，我国学生在中小学英语学习阶段广泛接触定冠词THE。例如，苏教版《牛津英语》初中教材中定冠词

THE在课文、会话和练习中出现3,200多次。据此估算，从小学到高中，加上课外练习，学习者接触定冠词的次数会非常之多。到高中阶段为止，应已掌握了其丰富的用法知识。一定意义上讲，中小学英语学习是影响冠词习得的关键阶段，而高中学习阶段又是基础教育和高等教育的重要衔接阶段。因此，高中英语学习者的语言使用研究既有助于反映中小学阶段的英语学习效果，同时也能够为后续大学英语教学研究提供参照。此外，定冠词THE与汉语词语对应关系不确定，其习得特征研究对其他英语功能词的教学研究也有启示意义。综上所述，以共选说为理论框架，基于语料库数据研究高中英语学习者的定冠词使用特征及其影响因素具有一定的理论意义和实践价值。

2. 研究设计

2.1　研究问题

本研究尝试回答以下问题：

1）与英语本族语者相比，中国高中生英语写作中的定冠词THE有何类联接特征？

2）与英语本族语者相比，中国高中生英语写作中的定冠词THE在意义、功能上有何特征？

3）影响上述特征的因素有哪些？

2.2　研究工具与对象

本研究以“中国学习者英语语料库”（桂诗春、杨惠中 2002）中的高中学生英语作文子库（ST2）为二语语料库，以英语本族语学生语料库LOCNESS（Granger 1998）为参照语料库，考察中国高中学生英语定冠词THE的使用特征及其影响因素。为了便于描述，下文分别将这两个语料库简称为ST2和LOCNESS。两者皆为学习者语料库，题材比较接近，具有较好的可比性，能为探讨中国高中英语学习者定冠词THE的使用特征以及相应的影响因素提供参照数据。

2.3　研究步骤

定冠词THE所修饰或限定的词语和结构比较丰富，相应的类联接所表达的功能也比较复杂，因此往往需要参照跨越单个甚至多个句子的语境方可确定。经典的KWIC（Key word in context）检索方法（参见Sinclair 1991, 1996, 2004）所提供的上下文语境信息（节点词左右4—5个词的跨距）虽然能够反映共现词语型式，但不便于确定冠词的具体用法特征。为此，我们借鉴Hoey（2005）所采用的考察

手段，以段落作为主要观察单位，必要时参照上下文其他段落，确定目标结构所表达的意义和功能。具体操作步骤如下：

首先，数据提取。分别从ST2和LOCNESS中随机抽取100条符合要求的索引行，同时根据索引行信息定位和获取相应的段落和文章，提取更为丰富的语境信息。其次，类联接、意义和功能标注。逐行观察、分析和概括定冠词THE的类联接特征，同时通过阅读其所在段落（必要时相邻段落）概括其功能，并进行相应的标注。如表1所示，可根据中心词的修饰语位置，将含定冠词THE的结构分为两大类："前置修饰、限定结构"（THE+ PRE-MOD +N）和"后置修饰、限定结构"（THE+NP+ POST-MOD）。如表1所示，前置修饰、限定结构主要包括THE+N, THE+N1+N2, THE+ ADJ +N和THE+R-ADJ+N（其中，R-ADJ表示等级形容词，而ADJ表示除此以外的形容词）等类联接；而后置修饰、限定结构则包括THE+(ADJ+)N +RL（其中RL表示关系从句），THE+(ADJ+)N+*of*, THE+N+PREP(*of* 除外)和THE+NP+ POST-MOD (介词和关系从句除外)等类型。根据定冠词THE所在结构以及上下文语境确定其所贡献的意义或功能，如"再次提及"、"共享知识"、"等级性描述"和"限定性后置修饰"等（参见Master 1990，2003）。

表1. 定冠词THE的类联接类型

分类	类联接	示例
THE+PRE-MOD+N	THE+N	the boy
	THE+N1+N2	the story book
	THE+NUM +N	the two legs
	THE+ADJ +N	the lovely girl
	THE+R-ADJ +N	the next year
THE+NP+POST-MOD	THE+N+RL	the boy who is laughing
	THE+ADJ+N+RL	the young man who is crying
	THE+N+*of*	the window of my house
	THE+ADJ+N+*of*	the beautiful girls of our class
	THE+N+PREP(of除外)	the water in the bottle
	THE+N+POST-MOD (介词和关系词除外)	the book written by Mo Yan

最后，分析讨论。采用中介语对比分析法（Granger 1996），以英语本族语为

参照，并结合相应的母语表达，分析中国高中生使用英语定冠词THE的特征及其影响因素。

3. 定冠词THE的使用特征

3.1　定冠词THE的类联接分布特征

如表2所示，LOCNESS中THE+ PRE-MOD +N和THE+NP+ POST-MOD分别约占60%和40%，而在ST2中则分别占93%和7%左右。相比之下，中国高中生倾向于少用THE+NP+ POST-MOD，而主要使用THE+PRE-MOD +N型式，以THE+N为主，约占全部索引行的53%，如the dog, the door, the fire, the river, the headmaster, the playground和the baby等。而在LOCNESS中，THE+N只占24%左右（如the system, the child, the mother, the author, the government和 the day等），但THE+NP+ POST-MOD 结构频繁出现，包括THE+N+*of*（如the beginning of their articles, the effects of the discrimination和 the subtleties of their opponents' argument等），THE+N+PREP (*of*除外)（如the need for such programs, the rise in crimes 和the great controversy over animal experimentation等）和THE+N+RL（如the censorship that is on public television now, the shows that are aimlessly thrown on the air和 the group that viewed the violent programs等）。与之相比，ST2中THE+NP+ POST-MOD只有少数几例（如the students in our class, the bank on* [1] the lake, the sound of nature, the place where ..., the day which* I went to ... 和the surface of the floor等）。

此外，THE+R-ADJ+N和PREP+THE+N在两个语料库中也都频繁出现。前者在LOCNESS和ST2中分别约占4%和6%左右，与THE+N和THE+ADJ +N在结构上非常相似，所不同的是含有表示“顺序”、“最高级”或“唯一性”的R-ADJ，如first, second, best和most等。与其他类联接不同的是，PREP+THE+N是本研究所讨论的唯一不以THE开头的类联接。在该结构中，THE+N/NP倾向于与特定的介词高频共现形成相对稳定的词语语法共选单位，即通常所说的介词短语，如in the end, on the way 和in the morning等。在这类短语中，THE与其他成分高度融合，形成公式化序列（formulaic sequence）。这类短语单位是二语教学关注的焦点之一，常常作为整体进行教学处理，学习者在表达中也倾向于整体调用，在LOCNESS和ST2中分别约占7%和15%。

1 表示ST2中出现但不符合英语表达习惯的词语组合。

表2. ST2和LOCNESS中定冠词THE的类联接与功能分布

分类	类联接	功能	LOCNESS	ST2
前置修饰、限定	THE+N	再次提及	10%	17%
		共享知识	13%	15%
		误用	1%	21%
	THE+N1+N2	再次提及	2%	1%
		共享知识	1%	3%
	THE+ADJ +N	再次提及	2%	4%
		共享知识	5%	2%
		误用	/	5%
	THE+R-ADJ +N	等级性、唯一性	4%	6%
	PREP+THE+N/NP	公式化构成	7%	15%
	Proper Name	专有名词	14%	/
		专有名词误用	/	2%
	其他组合，如the other, the practical等		1%	2%
后置修饰、限定	THE+(ADJ+) N+RL		/	/
	THE+(ADJ+) N+*of*		40%	7%
	THE+N+PREP（of除外）		/	/

3.2 定冠词THE的功能特征

如表2数据显示，在THE+ PRE-MOD +N类联接中，THE主要用于表达“再次提及”、“共享知识”、“等级性、唯一性”和“公式化构成”等多种功能，ST2数据尤为明显。相比之下，THE+R-ADJ +N和PREP+THE+N与所表达的功能具有较为明确的对应关系。例如，在LOCNESS和ST2中，学习者都倾向于使用THE+R-ADJ +N（如the first+NP, the best+NP和the second+NP等THE与序数词、形容词最高级等词语共现）来强调或凸显“唯一性”或“等级性”。与之相比，PREP+THE+N型词语组合中的THE所贡献的意义或功能已经非常微弱（如in the end, by the way和in the afternoon等），主要表现为与其他词语成分共同构成一个短语单位，即实现“公式化构成”的功能。在其余的前置修饰、限定结构中，同一类联接形式可用于表达两种及以上的功能。LOCNESS数据显示，本族语学生主要

使用THE+N实现“再次提及”和“共享知识”，分别约占10%和13%。与之相比，ST2中THE+N类联接实现“再次提及”和“共享知识”的用法则更为明显，分别约占17%和15%左右。两个语料库中的THE+N1+N2和THE+ADJ +N也都用于实现这些功能。

与LOCNESS相比，ST2中的定冠词THE除了在类联接和功能分布上存在差异以外，还存在大量“误用”现象，约占全部例证的28%。其中，大部分情况是THE与首次出现的NP共现，根据上下文语境判断，作者并没有表达“特指”或“定指”信息的意图。有些英语专有名词或名词短语（如人名、地名和节假日名词）前习惯性地不使用定冠词THE，但中国英语学习者却有使用THE的倾向（如that morning my classmates and I came to the* Peilei Theatre very early和The Spring Festival comes等）。LOCNESS中很少出现类似的误用。此外，LOCNESS中高频使用“THE+N+POST-MOD”结构来实现类似于“共享知识”等类型的具体功能。然而，中国学习者很少使用此类型式。因此，本研究不做具体细分和讨论。

4.1　英语本族语者定冠词THE的使用特征

LOCNESS语料库数据表明，无论是在前置限定、修饰结构中（如THE+N, THE+N1+N2, THE+ADJ+N和THE+R-ADJ+N等），还是在后置限定、修饰结构中（如THE+N+RL, THE+ADJ+N+RL和THE+N+*of*等），在绝大多数情况下（PREP+THE+N/NP除外）定冠词THE都标志着一个NP的开始（即为NP的左边界，是NP短语单位的标志性成分）。在英语本族语中，定冠词THE高频出现，THE+N和THE+ADJ+N+RL为典型的类联接。根据Sinclair（1991，1996）和Hoey（2005），这些结构的大量使用体现了THE与N以及THE与ADJ+N+RL等语法范畴的高频共现，构成THE+N和THE+ADJ+N+RL等类联接。语言使用者反复接触这些类联接，不断建立或加强相应的联系，形成比较牢固的类联接知识。换言之，当使用者接触到THE时，自然而然就会预测到N或ADJ+N+RL等语法范畴的出现，而不是VP等范畴。

数据还表明，这些类联接中的NP主要用于表示中心名词（N）或名词短语（NP）所承载的概念或所传递的信息“被再次提及”、“为说话者或听话者共享”或“NP受到后置限定或修饰”等（参见Master 1990，2003）。由此可见，含有THE的NP短语除了传递具体的命题意义以外，还用于表达上述具体功能，体现了NP短语与相应功能的共现。根据Sinclair（1991, 1996）和Morely & Partington（2009）等，在英语本族语表达中，NP是否与定冠词THE共现取决于说话者的交际目的，即是否为了强调“再次提及”、“共享知识”或“后置限定或修饰”等“定指”或“分类”功能。因此，就英语本族语者而言，THE+N和THE+ADJ+N+RL等类联接知识是NP短语单位与相应功能反复共现后所形成的，

是形式、意义和功能共选的产物。根据Hoey（2005）和Nattinger（1980），由于THE是这些短语单位的标记性开头，语言使用者在反复接触上述共现关系的过程中，不断建立和加强这些NP形式与相应功能的联系。因此，每当他们在交际中接触到THE时就会预测相应的NP或ADJ+N+RL等语法范畴的出现，同时也会自然而然地联系上它们所实现的“再次提及”、“共享知识”或“后置限定、修饰”等功能。而当他们需要表达其中的某（些）功能时，就倾向于借助THE进行表达。

4.2 中国高中学生冠词使用特征

与本族语学习者相比，中国高中生很少使用THE+N/NP+POST-MOD结构，但倾向于高频使用THE+PRE-MOD+N结构，特别是其中的THE+N结构，占主要部分。根据Hoey（2005）和Sinclair（1991, 1996），中国学习者懂得英语定冠词THE总是与NP（特别是N）这一语法范畴共现，即具有相应的类联接知识；当他们在二语表达中需要使用NP表达特定的概念或传递信息时，就倾向于启动定冠词THE并将之置于相应的位置（主要位于N或NP之前）。此外，THE+R-ADJ+N结构也占有相当高的比例，说明他们还知道THE与序数词或形容词最高级等词语共现；PREP+THE+N型式的频繁使用则说明他们具有THE与PREP和N共现的知识。由此可见，中国高中生已经掌握了定冠词THE与特定语法范畴的共现知识，表现为具体的类联接（主要为THE+PRE-MOD+N）知识；但缺少THE +N/NP+ POST-MOD类联接的知识。

上述分布比例分析表明，中国高中生倾向于过度使用THE+N类联接。下文通过其所表达的功能作进一步探讨。数据显示，与英语本族语学习者相似，中国学习者也频繁使用THE+N来表达“再次提及”和“共现知识”等功能。这似乎说明，他们掌握了类联接THE+N以及相应的功能。然而，他们在使用THE+N和THE+ADJ+N型式时普遍存在误用现象，即很多结构中不需要使用定冠词，此类约占全部用法的28%。例如，that is used to prevent the* thief stealing, I know that is the* important time for us to study和you have just move to the* country and was interested in growing some vegetables等。根据上下文语境信息，这些N或NP前无须使用定冠词，既不用于表示任何定指性“共享知识”也不是“再次提及”，像prevent stealing, an important time 和move to the countryside等表达可以更好地传递作者所要表达的意思。这些数据说明，中国学习者虽然掌握了定冠词与特定语法范畴的共现，但是并没有掌握特定类联接与相应意义、功能的共选知识。这一特征阐释了中国学生在母语者使用零冠词的语境中也会使用THE或A/AN的现象（参见朱叶秋2003）。

相比之下，“等级性、唯一性”和“公式化构成”用法分别倾向于与特定的类联接（THE+R-ADJ +N和PREP+THE+N）对应，前者往往带有非常明确的语

法标记，如the first game, the second record和the best month等，有很强的“语法化”倾向，而后者则主要是公式化程度很高的序列，如in the end, in the afternoon和outside the window等，这些表达常常整体存储和调用（参见Becker 1983: 341; Nattinger 1980: 218; Pawley & Syder 1983: 192），有很强的“词语化”倾向。无论是“语法化”还是“词语化”现象，它们的形式与功能高度对应、高度融合。这些数据进一步说明，学习者掌握了相应的共选知识，前者主要是THE与序数词或最高级语法范畴的共选，后者则主要是具体词语（PREP, THE和N）之间的共选，但未必是真正或主动地掌握短语单位与相应功能的共选。上述特征从共选角度解释了邵士洋和吴庄（2017）有关句法-语用接口难以被中国学生完全习得的结论，其原因与他们缺乏形式与意义和功能共选的知识有密切关联。

4.3 原因分析

1）中国学习者使用冠词THE的特征与其频繁接触THE+N/NP等型式有密切关联。上述分析表明，中国学习者具有THE与语法范畴N和NP的共现知识，即具有相应的类联接知识。正如英语本族语数据所示，冠词THE在英语中高频使用，且总是与N或NP共现。这种形式的普遍性对二语学习自然会产生直接影响。如前文所述，仅苏教版《牛津英语》7A-9B（初中英语）教材中，THE就出现数千次，且主要在THE+N/NP结构中。由此可见，即使是在汉语环境下学习英语，中国学生也在频繁接触THE+N/NP结构。也就容易抽象出N或NP与THE在形式上或语法范畴上的共现关系。根据Hoey（2005），在成千上万次接触THE的过程中，学习者反复启动这些共现关系，自然而然地建立了比较牢固的联系。换言之，中国学习者具有比较牢固的类联接知识，以至于在英语表达中需要调用N或NP表达具体概念或意义时，就趋向于启动THE+N/NP等类联接。

2）中国学习者缺乏THE类联接与相应功能的共选知识，与汉语中缺少相应的冠词有密切关联。中国高中生在写作中大量误用THE的数据表明，尽管他们懂得THE与特定语法范畴共选，但并没有掌握THE+N/NP等型式与相应功能的共选知识。产生后者的可能原因包括：汉语中也没有与THE直接对应的表述方式，而是借助于所谓的零冠词或指示词来实现“共享知识”、“再次提及”等功能。因此，尽管定冠词THE高频出现，用于表达多种功能，但中国英语学习者在接触THE的具体使用时，能够感知（听到或看到）到含有THE的词语组合，也能够概括出THE+N或THE+NP等类联接，但是很少能够“掌握”或“运用”这些型式所表达的具体功能与定冠词THE的密切联系。即与英语本族者相比，缺乏相应的形式与意义或功能的共选知识，结果导致他们误用THE+N/NP结构的现象，如“But *the fruit and green vegetables* are the most important.”和“In China, *the women leader* are not better than *the man*.”等。这里的*the fruit and green vegetables*, *the women leader*和*the man*都是首次出现，且没有任何特指或定指的意图。根据

Hoey（2005），中国英语学习者在处理THE+N或THE+NP等类联接型式时，并没有启动相应的功能。

此外，由于英语冠词系统很复杂，再加上汉语中并没有与之直接对应的冠词形式，所以往往通过显性语法教学来帮助学习者获得相应的用法知识。这对中国英语学习者冠词的使用产生一定影响。其中，“特指”和“再次提及”等用法特征往往是强调的中心。例如，《实用语法》（张道真2002：69）对定冠词THE的第一点用法的描述为“和个体名词的单数或者复数连用，表示某个（些）特定的人或者东西”，并使用相关例句说明，如：

（1）Where is (are) *the other girl(s)*? **那个（那几个）姑娘**在哪里？

（2）Put *the parcel(s)* on the table. 把**那个（那些）包裹**放在桌上。

为了凸显这些“特指”用法，往往汉语译文借助于限定词或指示词“这”、“这个”、“这些”、“那”、“那些”和“那个”进行强调。结果，THE+N/NP结构通常都被翻译成“这”、“这个”、“那”和“那个”（Chung 2000；王丽娜2007）。相应地，在英语教学和学习过程中，教师和学生也就会倾向于使用“这”、“这个”、“那”或“那个”与N/NP共现的结构或表述方式来强调THE+N/NP的用法特征。然而，双语词语的对应关系是具有方向性的（卫乃兴、陆军2014），尽管THE+N/NP在很多情况下对应汉语“这/这个/这些/那/那个/那些”+NP，但反之不然，除THE+N/NP之外，它们还可能对应this/that/these/those+N/NP等。因此过度使用限定词来强调定冠词用法等显性语法教学方式也是导致学习者忽视形式与功能共选的因素之一。

5. 结论

本文以共选说为理论框架，以英语本族语学生写作语料为参照，考察中国高中生英语写作中定冠词THE的使用特征及其影响因素。数据表明：与英语本族语学生相比，中国高中生作文中定冠词THE主要使用在前置限制、修饰结构中，以THE+N类型为主，但在意义和功能上的误用倾向比较明显。分析表明：（1）本族语学生作文中，英语定冠词倾向于与特定的语法范畴共现构成多种类联接，用于表达特定的意义或功能，体现了形式、意义和功能的共选；（2）中国高中生懂得定冠词与某些特定语法范畴共现，但是缺乏相应的意义和功能知识；（3）中国学习者的定冠词共选特征与其频繁接触THE+N/NP等形式、汉语中缺乏相应的冠词系统以及显性语法教学等因素有密切关联。

上述发现对深入探索二语语法结构的形式、意义和功能共选，以及相应共选知识的习得研究具有一定理论启示，对二语语法教学也有一定的应用价值。语法结构教学不仅需要强调语法功能，还需要强调其所实现的意义和功能。对于中国

英语学习者而言，定冠词THE属于二语词汇语法项中特有的现象。其形式和用法知识反映了在缺乏或没有母语影响的情况下的二语结构的形式、意义和功能的习得特征，对母语迁移影响分离研究有一定的价值，对于二语教学中具体处理与母语结构有不同对应程度的词汇语法结构具有一定的启示意义。不管母语与目标语词语的对应程度如何，都可以通过所在的短语单位或扩展意义单位（Sinclair 1996）来凸显相应的形式、意义和功能的共选关系。

参考文献

Abney, S. P. 1987. The English noun phrase in its sentential aspect [D]. PhD. dissertation, Massachusetts Institute of Technology.

Barlow, M. 1996. Corpora for theory and practice [J]. *International Journal of Corpus Linguistics* 1(1): 1-37.

Barnbrook, G. 1996. *Language and Computers: A Practical Introduction to the Computer Analysis of Language* [M]. Edinburgh: Edinburgh University Press.

Becker, A. L. 1983. Toward a post-structuralist view of language learning: A short essay [J]. *Language Learning* 33 (S5): 217-220.

Bickerton, D. 1981. *Roots of Language* [M]. Ann Arbor: Karoma Publishers.

Chung, Y. E. 2000. A contrastive analysis of articles and demonstratives in English and modern standard Chinese [D]. Unpublished MA Thesis. California State University.

Ellis, N. C. 1994. *Implicit and Explicit Learning of Languages* [M]. London: Academic Press.

Ellis, N. C. 2006. Modelling learning difficulty and second language proficiency: The differential contributions of implicit and explicit knowledge [J]. *Applied Linguistics* 27(3): 431-463.

Ellis, N. C. 2011. Implicit and explicit SLA and their interface [A]. In C. Sanz & R. Leow (eds.). *Implicit and Explicit Language Learning: Conditions, Processes, and Knowledge in SLA & Bilingualism* [C]. Washington D. C. : Georgetown University Press. 35-47.

Granger, S. 1996. Learner English around the world [A]. In S. Greenbaum (ed.), *Comparing English Worldwide: The International Corpus of English* [C]. Gloucestershire: Clarendon Press. 13-24.

Granger, S. (ed.). 1998. *Learner English on Computer* [C]. London: Addison Wesley Longman Publishing Company.

Gundel, J. K., N. Hedberg & R. Zacharski. 1993. Cognitive status and the form of referring expressions in discourse [J]. *Language* 69 (2): 274-307.

Hoey, M. 2005. *Lexical priming: A New Theory of Words and Language* [M]. London/New York: Routledge.

Ionin, T. R. 2003. Article semantics in second language acquisition [D]. PhD. dissertation, Massachusetts Institute of Technology.

Lardiere, D. 2004. Knowledge of definiteness despite variable article omission in second language acquisition [A]. In A. Brugos, L. Micciulla & C. E. Smith (eds.). *Proceedings of the 28th Annual Boston University Conference on Language Development* [C]. Somerville: Cascadilla Press. 328-339.

Lardiere, D. 2009. Some thoughts on the contrastive analysis of features in second language acquisition [J]. *Second Language Research* 25:171-225.

Longobardi, G. 1994. Reference and proper names: A theory of N-movement in syntax and logical form [J]. *Linguistic Inquiry*, (4): 609-665.

Master, P. 2003. Pedagogical frameworks for learning the English article system [J]. *Applied Linguistics Forum* 24(1): 1-5.

Master, P. 2012. Teaching the English articles as a binary system [J]. *TESOL Quarterly* 24(2): 461-478.

Morley, J. & A. Partington. 2009. A few frequently asked questions about semantic—or evaluative—prosody [J]. *International Journal of Corpus Linguistics*, 14(2): 139-158.

Nattinger, J. R. 1980. A lexical phrase grammar for ESL [J]. *TESOL Quarterly* 14(3): 337-344.

Pawley, A. & F. H. Syder. 1983. Two puzzles for linguistic theory: Nativelike selection and nativelike fluency [A]. In J. C. Richards & R. W. Schmidt (eds.). *Language and Communication* [C]. London: Longman. 191-225.

Sinclair, J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: Oxford University Press.

Sinclair, J. 1996. The search for units of meaning [J]. *TEXTUS: English Studies in Italy 9* (1): 75-106.

Sinclair, J. 2004. *Trust the Text: Lexis, Corpus, Discourse* [M]. London/New York: Routledge.

Stubbs, M. 1996. *Text and Corpus Analysis*. Oxford/Malden: Blackwell Publishers.

Stubbs, M. 2009. John Sinclair (1933-2007): The search for units of meaning: Sinclair on empirical semantics [J]. *Applied Linguistics* 30 (1): 115-137.

Thráinsson, H. *et al*. 2007. The Icelandic (pilot) project in ScanDiaSyn [J]. *Nordlyd* 34 (1): 87-124.

蔡金亭、吴一安，2006，中国大学生英语冠词使用研究[J],《外语教学与研究》(4)：243-250。

常　辉、赵　勇，2014，冠词缺失与中介语句法损伤研究[J],《外语教学理论与实践》(1)：10-16。

戴炜栋、韦　理，2008，中国学习者英语冠词语义特征习得研究[J],《外语教学与研究》(2)：136-142。

桂诗春、杨惠中，2003,《中国学习者英语语料库》[M]。上海：上海外语教育出版社。

李景泉、蔡金亭，2001，中国学生英语写作中的冠词误用现象[J],《解放军外国语学院学报》(6)：58-62。

陆　军，2017，二语隐性、显性搭配知识特征研究—— 一项语料库数据分析与心理语言实验的接口案例[J],《解放军外国语学院学报》(3)：1-10。

陆　军、卫乃兴，2014，短语学视角下的二语词语知识研究[J],《外语教学与研究》(6)：865-878。

邵士洋、吴　庄，2017，语言接口视角下中国学生英语冠词习得研究[J],《现代外语》(4)：552-563。

卫乃兴，2002,《词语搭配的界定与研究体系》[M]。上海：上海交通大学出版社。

卫乃兴、陆　军，2014,《对比短语学探索》[M]。北京：外语教学与研究出版社。

王丽娜，2007，英汉指示词对比研究与翻译[D]，硕士学位论文。上海：上海海事大学。

闫丽莉，2003，中国学生英语冠词习得初探一项基于中国学习者英语语料库的研究[J]，《外语教学与研究》(3)：210-214。

于善志、苏佳佳，2011，There be存在句习得中的定指效应研究[J]，《外语教学与研究》(5)：712-725。

张道真，2002，《张道真实用英语语法（最新版）》[M]。北京：外语教学与研究出版社。

周保国，2007，第二语言习得中英语定冠词过度使用研究[J]，《现代外语》(4)：387-394。

朱叶秋，2003，大学生英语冠词掌握情况调查[J]，《外语教学与研究》(3)：206-209。

朱叶秋、文秋芳，2008，大学生口语中零冠词使用正确性的预测因素研究[J]，《现代外语》(4)：399-405。

通讯地址： 225127 扬州市邗江区华阳西路198号扬州大学外国语学院

理工科研究生论文摘要中it词块的先行和回指特征研究*

上海理工大学 张 乐

提要：本文考察中国理工科研究生毕业论文英文摘要语料库中的it-型高频词块，对比分析中国学习者和本族语学生在先行和回指使用上的短语学特征。研究结果表明：（1）中国理工科研究生大量使用it词块来实现篇章组织和表达态度意义；（2）先行it词块的主要问题存在于共选层面，主要表现为词语与结构之间的错误匹配，具体原因包括过度概括、词汇量制约、论述策略和母语影响；（3）回指it词块的主要中介语特征为代词it以及表达态度意义的词语的过度使用，该特征既凸显了学习者所违反的基本回指原则，也可能与我国研究生毕业论文写作策略有关。

关键词：学术论文摘要、it词块、先行it和回指it

1. 引语

以it为主语（包括先行主语和回指主语）的多词序列在学术英语文本中高频发生，本文称之为it词块。迄今为止，这类词语手段的探索主要集中于发生在外位结构（extraposition）中的“先行it”（anticipatory-it）。先行it本身不包含信息，其语义功能是“预示真正主语将出现于句子的后半部分”（Quirk *et al*. 1985：89）。大多数研究表明，it外位结构在学术文本中的发生概率远高于普通文本（Biber *et al*. 1999；Herriman 2000b），这与学术文本的内在属性有直接关联。一方面，外位结构的使用能在很大程度上削弱作者的个性特征，强调研究的普遍适用性（Herriman 2000a；Hewings & Hewings 2002）。作者在文本中不直接发声，而是通过形式主语、被动语态等“去作者化”手段来缓和表述语气、减轻命题责任、保护同行权威、避免同行批评。另一方面，学术文本是作者和读者发生互动的场所，交际行为极为丰富（Hunston 1993；Hyland 1998）。而it外位结构恰恰能够实现多重情感和态度的表达（Herriman 2000a；Biber *et al*. 1999；Hewings & Hewings 2002；Groom 2005；胡文杰、李晶洁 2015），是文本中实现评价功能的重要词汇-语法综合体（Hunston & Sinclair 2000）。诸多研究也对比调查了本族语和非本族语

* 本研究得到教育部人文社科规划基金项目“基于语料库的中国理工科大学生英语写作教学体系研究”（14YJA740023）和上海理工大学教师教学发展研究项目（CFTD18035Y）的资助。

学生的先行it使用特征（Ädel 2014；Hasselgård 2009；Herriman 2013；Hewings & Hewings 2002；Hyland & Tse 2005；Thompson 2009；Larsson 2016），发现各个层级的学习者在不同程度上发生少用（如Römer 2009）或超用（如Hewings & Hewings 2002）的情况。中国学习者与本族语学生在使用这些表达时有何异同，是本文的一个关注点。

本研究的另一个关注点是it词块的回指功能。这里所述的回指（anaphoric）指的是第三人称代词的所指对象（即先行词，antecedent）出现于第三人称之前的情况（Quirk *et al*. 1985：347）。第三人称代词是实现这一功能的最典型的语言实体之一。然而，作为回指代词的it高频出现于哪些词块，却是一个很少问津的话题。其主要原因是，当it用于指向上文提及的事物时，其回指对象和共现的谓语动词范围广阔，实现形式多变，很难像外位结构一样自成意义单位，似乎不具备短语学所推崇的半固定或固定标准。尽管如此，我们在检查中国理工科研究生毕业论文英文摘要时，发现了相当数量且反复出现的回指性it词块。本文试图探讨，在这些实例中，it的回指对象是否有规律可循，其所在词块在学习者语篇中的形式、意义和功能的特征是什么，与本族语学生摘要相比有何鲜明特点。

毕业论文是研究生攻读学位的最终成果展现，而中英文摘要则凝聚了这一成果的主要理论和方法论特征，是报道研究结论、展现研究价值的窗口。本文利用中国理工科研究生毕业论文英文摘要语料库和自建的参照语料库，对比分析以it为主语（包括先行it和回指it）的高频词块在中国学习者和本族语学生摘要语篇中的形式和功能特征。文章分为5个部分。第2小节介绍本文所使用的语料库，并在总体上概述it词块的形式和功能特征；第3小节报道主要研究结果，包括it词块所实现的5种功能，以及实现这些语篇功能的语言实体在两种语料库中的异同；第4小节总结讨论it词块的中介语问题和特征。最后总结全文。

2. 研究设计

2.1 语料库

本文使用的学习者语料库为上海理工大学在建的中国理工科大学生英语写作语料库（Chinese Science and Engineering Majors Written English Corpus，CSEMWEC）（张乐、刘芹 2017）。其中的学术英语子库（本文简称为Lcorp）主要收录了2010至2014年期间国内12所高等院校的4,400余篇理工科硕博士学位论文英文摘要，库容约为200万词次。

参照语料库为同步建设的理工类本族语研究生学位论文摘要语料库（本文简称为Rcorp）。语料收集于ProQuest学位论文数据库，该数据库主要提供来自欧美国家2,000余所知名大学的优秀硕博士论文。目前，共从中抽取了6,550篇英文摘要。经

筛选加工后，库容同样为200万词次。这些摘要的所属学科、论文完成年份等主要参数与Lcorp基本一致。我们力求两种语料库在规模、性质、构成等方面均满足可比的标准，基于此来对比考察学业（学术）水平相近的研究生的英文摘要写作能力。

此外，我们在讨论环节还使用了上海交通大学建设的JDEST科技英语语料库的一部分数据。与Lcorp和Rcorp不同的是，JDEST的文本作者均为本族语专业人员，且其库容规模更大（约为650万词次）。但由于JDEST不是摘要语料库，故不直接介入对比研究，仅提供必要的辅助验证。

2.2 it词块的分布特征

本研究使用Wordsmith Tools 6.0，在两种语料库中提取以it为主语、至少出现5次的4词词块。提取结果显示，两个语料库的总计词块数量（类符）为188个，总计（形符）3,343词次。词块的基础长度设定为4个单词，这是基于对不同长度词块的观察以及过往研究经验（如Biber *et al.* 1999）的考虑。但之所以称之为“基础长度”，是因为有些词块的特征检查需要我们将短语观察范围扩展至更多单词，比如，若不检查词块it is of great的扩展搭配词（如importance）就很难准确概括其功能。相反，对于一小部分词块来说，观察3个单词就足以揭示其语篇功能，如it is found (that)等诸多主语从句。

先行it所实现的外位结构是学术英语中实现篇章意义的经典表达。本文主要涵盖如下型式（参见Hunston & Francis（2000））：

（1）it v-link v-ed that（如it is believed that …）

（2）it modal be v-ed（如it can be seen that …）

（3）it v-link adj to-inf（如it is important to …）

（4）it v-link v-ed to-inf（如it is expected to …）

（5）it v-link of n（如it is of utmost importance …）

（6）it v that（如it turns out that …）

同时，本文涵盖了回指it词块的如下主要型式：

（7）it modal be v-ed（如it can be used …）

（8）it v (adj) n（如it has a great …）

（9）it modal (adv) v（如it can greatly reduce …）

（10）it v-link (adj) n（如it is an important …）

初步检查后有如下发现：第一，回指it词块在Lcorp中比在Rcorp中明显出现得更为频繁；第二，回指it词块的总出现次数虽不能与先行it词块数量相及，但它们实现了一部分先行it不经常涉及的语篇功能；第三，有些词块在用于不同语境

中时，it的语法属性也会发生变化，甚至难以明确断定，这是学习者文本的常见特征，譬如：

[1] *Four compatilizers have been used to modify WPC material separately and their properties were studied. It showed that the properties of …* (Lcorp)

实例[1]可从3个不同视角去理解。其一，it指代前文的某个名词短语。纵观全句，只有material符合句法标准，但不符合意义标准。其二，it回指前文整个句子，这是最有可能的情形。其三，作者有意使用外位结构，但使用了错误的型式（应为it is/was/has been shown that）。这些可能性需要研究者进一步扩展语境考察和主观判定。

2.3 it词块的功能特征

学术论文摘要中的功能讨论大多围绕语步特征展开，如Salager-Meyer（1990）、Bhatia（1993）、Lores（2004）等人的研究分别确立了摘要文本中具有代表性的语篇行为。本研究中，两种语料库的总计188个it词块主要实现学位论文摘要中不可或缺的5大类功能，对应于摘要语篇的5个主要语步（Swales 2006；Ädel 2014），即背景、目的（意义）、方法、结果、启示（应用）：

（1）文际功能，即“前人尝试了哪些工作，进展和结果如何？”

（2）评价功能，即“本研究所使用的理论、方法、技术、模型等有何种属性？”

（3）指示功能，即“研究中有哪些重要步骤？”

（4）推断功能，即“研究产生了哪些结论，如何解读？”

（5）展望功能，即“研究有哪些理论或实践上的意义？”

必须注意的是，词块的形式和功能并非一一对应关系。“一个形式对应多种功能”是常见现象，典型实例是实施文际功能和推断功能的序列。很明显，in this paper it has been shown that和it has been shown in previous research实现两种不同的语篇功能，在分析it has been shown时需加以甄别。表1列举了实现上述功能的一部分典型实例。

表1. it词块的语篇功能

语篇功能	先行it举例	回指it举例
文际功能	It is well known that ...	-
评价功能	It is difficult to ...	It has the advantages ...
指示功能	It is critical to ...	-
推断功能	It is concluded that ...	-
展望功能	It is expected that ...	It will lead to …

3. 研究结果

3.1 文际功能

实现文际功能的it词块帮助作者回顾前人研究，是文献综述的常见词语手段（如张乐 2017）。在两种语料库中，实现该功能的主要词块如表2所示（括号内为发生次数，下同）：

表2. 语料库中实施文际功能的it词块

Lcorp	Rcorp
it is (well) known that (17); it is hard to (12)	it is (well) known that (27); it has been shown (17); it is believed that (8); it has been demonstrated (6); it is suggested that (6); it is thought that (5); it has been proven (5); it has been observed (5); it has been suggested (5); it is estimated that (5)

一般而言，摘要中并不出现过多的文献引用，更多的是以精练的语言概述该话题的发展脉络、理论成果、技术进步、学科发展等。表2中，Rcorp中相关词块的出现位置均靠近文本的篇首，许多还与in previous publications等短语共现，文际特征显著。中国学习者在摘要中综述研究成果、回顾研究背景的意愿水平低于本族语研究生，尤其是考虑到Lcorp的平均篇幅（约450词/篇）要明显长于Rcorp（约317词/篇）。唯一一个在两个语料库中都高频出现的文际性it词块是it is (well) known。这说明，与综述研究、提炼观点相比，中国学生更倾向于展现学科背景和话语社团的共同认知。此外，在Lcorp中，中国学生常使用it is hard to来评述其他研究所采取的方法，特别是揭示以往研究手段所存在的局限性，这是Rcorp中极少发生的序列。比如：

[2] *This kind of methods is bases on .... <u>It is hard to</u> keep the integrity relationship of sub-problems, the results are hard to guarantee be the globally optimal solution.* (Lcorp)

3. 2 评价功能

在本研究中，it词块的评价功能赋予研究本身或研究中的思想体系、经验认知、理论原则、技术方法、操作手段等一定的价值。所谓价值，实则为作者传达的评判和态度，可粗略划分为积极、消极和中性。这些价值属性构成研究计划提出的动机或研究设想得以实施的前提条件，同时也告知读者该研究计划的可行性和意义，引起读者对研究的兴趣。

评价性it词块在两个语料库中的出现总数达到843次，其中，回指it与先行it分别发生447次和396次。从功能上说，它们主要用于说明一般用途或具体应用（如It can be used to ensure data integrity ...）、表明研究、理论、技术方法的价值或地位（如It has the advantages of small size ...）、表明研究行为的可行性和难易程度（如It is possible to maintain the ...）、表明研究所使用的材料、设备、工艺等其他属性（如It provides a new way to solve such ...）。

根据不同型式，它们在Lcorp和Rcorp中的分布情况如表3所示：

表3. 语料库中实施评价功能的it词块

Lcorp	Rcorp
（1）it is + adj (adj主要为possible, difficult, easy, feasible, the first, of great, significant) + to (185)	（1）it + is/was + adj (adj主要为interesting, of interest, desirable, (not)possible, impossible, difficult, easy, challenging, the first) + to (211)
（2）it + has become/has been/is/was+ a/an/the + adj (adj主要为key, potential, effective, valuable, important, main, hot, common) + n (158)	（2）未发现超过5次的实例
（3）it + could/can/is/has been+ (be) + applied/used/required (151)	（3）it + could/can/is + (be) + applied/used (51)
（4）it has + important/great/many/broad + application/significance/advantages (42)	（4）未发现超过5次的实例
（5）其他：it is different from (5), it is suitable for (5), it plays an important (8), it provides a new (6)	（5）其他：it + is/was +able/capable + to/of (21)

统计结果显示，在表3中的全部843个实例中，出现于Lcorp和Rcorp的次数分别为560和283次，频数差异较为显著。进一步检查后发现，Rcorp的全部实例中，先行it占74.6%。换言之，本族语学生在实施评价行为时，更倾向于使用外位结构。而在Lcorp的全部实例中，回指it占了67%，与Rcorp的情况截然不同。

在Rcorp中，出现不少于5次的不同评价性词块共计24个，Lcorp中的这一数值为48个。相比而言，中国学生所使用的it评价性词块无论形符还是类符数量均大大超过本族语学生。尤其突出的是，他们大量使用词块来描述理论、方法、材料的价值、研究动机、研究意义、应用情况。借助这些具备鲜明态度意义和情感特征的词语（尤其是形容词），中国学习者尽力使读者（如学位论文评阅人）相

信，该研究采纳的理论模型或技术手段与前人相比有显著区别或提升。以下是学习者用于“创建研究空间”和“填补研究空白”的典型实例：

[3] *However, it is still a great challenge to overcome the problems of low photocatalytic activity and low utilization of sunlight*. (Lcorp)

[4] *This paper analyzes the structure of front crotch, and it is the first to mention the X, which was used to represent the fullness of male special feature*. (Lcorp)

3.3 指示功能

实现指示功能的it词块用于提醒和告知读者研究中的注意事项和重要操作步骤。它们在Lcorp和Rcorp中分别出现243和241次，无一例外为先行it结构，并至少包含一个表达重要、关键、必要、建议等意义的词语。两种语料库中的词块分布如表4所示：

表4. 语料库中实施指示功能的it词块

Lcorp	Rcorp
(1) it is (very) + adj + to (adj 主要为 urgent, essential, important, necessary, needed) (226)	(1) it is + adj + to (adj 主要为 critical, essential, imperative, important, necessary, recommended, vital) (185)
(2) it is of great + meaning/significance/importance + to (17)	(2) it is + imperative/important/recommended + that (40)
	(3) 其他：it needs to be, it should be noted (16)

不难看出，本族语学生和中国学生均高频使用型式it v-link adj/v-ed to-inf。本族语学生所使用的不定式动词主要为(to) understand (33)、have (28)、use (20)、determine (16)、develop (14)、find (11)、consider (10)、design (8)、study (7)、predict (7)、reduce (6)、quantify (6)、perform (6)、know (6) 等表达心理或研究行为的词语。而中国学习者使用了更多不同的研究行为动词，如study (36)、improve (16)、use (14)、research (14)、develop (13)、investigate (7)、solve (7)、obtain (6)、consider (6)、achieve (6)、do (6)、establish (5)、reduce (5) take (5)、meet (5)、understand (5) 等。在adj/v-ed方面，学习者的选择范围比本族语者小得多，甚至使用了不正确的词语搭配（如it is needed to do）。

3.4 推断功能

实施推断功能的it词块用于宣布实验结果，提出研究结论，或是根据观察做出解释。推断与总结是摘要中最核心的语篇行为之一，很大程度上决定了能否使读者有兴趣和动机来关注摘要所展示的研究成果。推断性it词块在语料库中极为高频，在Lcorp和Rcorp中总计发生1,724次。根据时态和动词分布，两个语料库中的高频词块可归纳为表5中的如下序列：

表5. 语料库中实施推断功能的it词块

Lcorp	Rcorp
(1) it is v-ed that (v-ed主要为found, shown, concluded, proved, showed, demonstrated, indicated, noted, reported, suggested) (217)	(1) it is (also) v-ed that (v-ed主要为shown, found, observed, concluded, demonstrated, hypothesized, assumed, determined, seen) (359)
(2) it was v-ed that (v-ed主要为found, shown, concluded, indicated, founded) (145)	(2) it was v-ed that (v-ed主要为found, determined, observed, shown, concluded, hypothesized, demonstrated, discovered, seen, noted, noticed, confirmed) (801)
(3) it can be + seen/concluded/found/said (54)	(3) it + will/can be + concluded/seen/shown (35)
(4) 其他：it shows/showed, it turns out/ turned out, it finds/found that, it indicated that, it comes to the conclusion (79)	(4) it appears that (5)
	(5) it is + clear/possible/likely that (29)

从表5中的对比不难发现，在陈述研究结论时，本族语学生所使用的词块在数量上远远超过中国学习者（1,229次对比495次）。我们发现，第一，两种语料库都极其依赖it v-link v-ed that这一型式。区别在于，中国学习者较多使用一般现在时，而本族语学生则更倾向于一般过去时。此外，中国学生超用了proved、indicated、reported、suggested（以及错误使用的showed）等动词，少用了determined、observed、hypothesized、discovered、seen、noticed、assumed、confirmed等动词，显示出同一型式的不同词语选择。

第二，在表达的委婉程度方面，学习者与本族语者均经常使用的型式是it modal v-link v-ed that。It appears that和it is possible/likely that是本族语学生经常使用的实现委婉表述的模糊限制性短语，而中国学生极少使用。

第三，在实施推断行为时，中国学生虽使用类似it appears that的型式（即it v that），但谓语动词与Rcorp截然不同：it turns out/turned out that、it finds that、it found that和it comes to the conclusion that几乎不发生于Rcorp。it indicated that、it shows that和it showed that主要用于回指前文命题内容，句法上与it appears that有本质差异。

3.5 展望功能

展望性it词块的主要目的是预测和期待研究前景，包括研究的应用范围和途径、可预见的社会影响、对学科发展的推动、作者致力于实现的话语社团交际目的等等。此处的“研究前景”与上文已述的“研究价值”有语篇功能上的相近之处。但两者之间存在微妙区别：前者主要预见方法、技术、成果所带来的潜在影响，后者更注重研究热度、理论和方法论价值。展望性it词块在Lcorp和Rcorp中分别出现了143次和31次，如表6所示：

表6. 语料库中实施展望功能的it词块

Lcorp	Rcorp
(1) it is expected that (6)	(1) it is expected/anticipated that (31)
(2) it is expected to (7)	
(3) it can + v (v主要为achieve, get good, greatly produce, improve, meet, provide, reduce) (57)	
(4) it has + adj + n (adj主要为a certain, a good, a great, a large, a wide) (46)	
(5) it will + v (v主要为be a, lead to, promote, reduceresult in) (27)	

就外位结构而言，中国学生和本族语学生均使用了it is expected that，后者还高频使用了it is anticipated that，但总的来说差异不大。二者的主要差异是回指性词块。在Lcorp中，出现不少于5次、实现展望功能的回指性it词块共有20个，总计出现130次。中国学习者高频使用序列（3）和（5）中的表达，主要目的是强调方法、结构、模型、材料等对于提高效率、节约成本、满足标准、扩大影响、增加产能等方面的贡献：

[5] *This process can improve the feasibility and accuracy level, and it can achieve the high efficiency when the large amount of consumer dates are take into account*. (Lcorp)

[6] *It will be a promising project to improve these three methods in order to achieve more efficient communication between IPV4 and IPV6.* (Lcorp)

学习者使用序列（2）和（4）主要用于展望研究成果的应用能力、后续发展、学科前景和影响力，如：

[7] *On the other hand, it has a good extensive application prospect, such as aircraft fleet, formation control of robot and unmanned aerial vehicles.* (Lcorp)

相反，Rcorp中未见发生次数不少于5次、实现展望功能的回指性it词块。很明显，总体而言，中国学生使用it词块来强调研究意义和贡献的意愿要高于本族语学生。

4. 学习者的词语、型式和语篇：问题和特征

从上述分析可以看出，中国理工类高阶段学习者已初步具备使用it词块来构建语篇、表达态度的能力。在实施各个语篇功能时，学习者有意识地利用半固定预制件来创建文本结构，便于读者在最短时间内捕捉块状信息。更重要的是，中国学习者和本族语学生秉持着相同的命题呈述原则，即作者需适时游离于文本之外，使评述和报道尽可能显得真实客观。

但同时，本研究也揭示若干中国学习者特有的it词块使用方式，暴露了程度不一、性质各异的问题。本小节重点讨论学习者在型式使用上的问题和it的过度使用现象。

4.1 先行it：型式和词语的精密选择与冲突

词语和结构存在共选关系，这是新弗斯语料库语言学最重要的发现之一（Sinclair 1991）。语料库研究业已表明（如Hunston & Francis 2000），大多数型式只能容纳有限数量的词语。这些词语作为所在型式的组成部分，往往存在意义上的关联或近似。然而反过来说，并非意义相近的词语都可出现于同一型式之中。中国学生在使用2.2小节中的10个型式时，不同程度上发生了词语与结构的冲突，其潜在原因包括过度概括、词汇量制约、策略因素和母语影响。

首先，词语和结构的共选机制决定了学习者文本中的许多短语问题既可能出现于型式层面，也可能出现于词语层面，甚至二者同时存在。我们在Lcorp、Rcorp和JDEST中分别检索“it * that”，即以it为主语、以单个动词为谓语、以宾语从句结尾的所有实例。至少出现5次的检索结果如下所示（大写单词表示有屈折变化）：

Lcorp: it SHOW that (79); it FIND that (17); it INDICATE that (14); it PROVE that (12)

Rcorp: it APPEAR that (21); It SHOW that (10)

JDEST: it APPEAR that (164); it SEEM that (103); it FOLLOW that (55); it SUGGEST that (22); It SHOW that (17); it ASSUME that (10); it IMPLY (10); it INDICATE that (5)

基于上述数据，我们总结了三种语料库中先行it和回指it在“it * that”中的分布比例。图1显示，中国学习者仅使用了it的回指功能，而本族语学生和本族语专业人员则明显更倾向于使用先行it。中国学生高频使用的it FIND that和it PROVE that很有可能是基于it SHOW that、it INDICATE that等短语的过度归纳，也有可能是在使用FIND和PROVE这两个动词时选择了错误的型式。

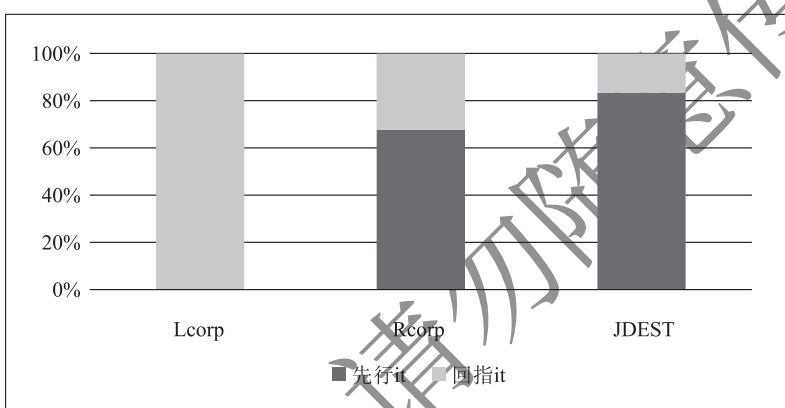


图1 “it * that”中的先行it和回指it（至少发生5次）

第二，有些实例可能与学习者的词汇量受限有关。我们在3.3小节中对比讨论指示功能的实现形式时，发现学习者明显少用了understand和have，过多使用了study、research、improve、solve、obtain等词语。中国学习者过多使用study和research的原因可能是他们试图利用这两个动词所蕴含的“研究”之意来泛指所有研究行为，而不是使用更为具体的examine、investigate、discuss等研究行为动词。

第三，学习者少用understand，且过多使用improve、solve、obtain等动词，似乎意味着中国学习者在辩论和说服策略上更倾向于谈论研究所能实现的目标和成效，而不是针对研究本身或研究中各类问题的理解领会。

第四，学习者语料库中的部分短语与汉语中的相关表达有密不可分的关联。比如，it FIND that和it PROVE that的高频出现很有可能受到了母语中“结果发现”、“结果证明”等习惯性表述的影响。

上述分析说明，中国学习者呈现出来的中介语特征，既非完全关于句法、也非完全关于词语。他们使用大量学术英语常见句型，但对这些句型的核心功能特征和词汇实现方式缺乏清晰的认知。型式作为一种抽象的词汇-语法表征，其具体

实现过程受制于诸多因素，结构和功能一一对应的情况少之又少。难以确定甚至完全忽视词汇填充的制约条件，是Lcorp的常见现象。

4.2 回指it：非策略性和策略性的过度使用

从第3小节的分析可知，中国学习者趋于利用回指it词块来实施评价和展望功能，其使用频率远超Rcorp中的对应数据。回指it的过度使用现象主要可分为两种，第一种与回指的基本原则相关，第二种与语篇策略相关。

统计显示，第三人称代词it在Lcorp和Rcorp中的出现概率分别为3.4次/千词和2.7次/千词。中国学习者比本族语者更加依赖it，特别是回指性it，是一个有趣且值得进一步探索的话题。一般而言，无论近指或是远指，无论指物（如名词短语）或是指事（如整句），it都应有较为明确的回指对象，而学术语言的严谨性更加要求代词与所指对象之间存在易于分辨的关联。然而，通过观察Lcorp中的回指it高频词块，可以发现代词it过度使用、回指对象模糊不清构成了影响语篇理解的障碍。一种情况是，it词块之前存在多个名词短语或小句，但作者未能有效厘清回指关系，造成阅读困难。与例[8]类似的问题在学习者文本中广泛存在：

[8] *When a part is added, the appropriate process is reused after it is classified. So it can achieve the standard and automatic NC programming.* (Lcorp)

另一种情况是，中国学习者对于名词“数”的处理较为随意，语法不一致影响读者对回指对象的判断：

[9] *In addition, the structural characteristics determine its limitations. It can get good control effect only in the simple linear univariate system, but ineffective in the complex systems.* (Lcorp)

最后，回指it的过度使用也在某种程度上反映出学习者不善于使用其他更多的回指性词语手段，比如通过this、such等限定词的使用、所指对象的重述或是近义词、上下义词等手段的使用。This、that、these、those和such这5个词语在Lcorp和Rcorp中分别出现约24,000次和43,000次，差异巨大。

另一方面，在“评估研究价值”这一语篇行为的实施上，中国学习者明显比本族语者更加积极。我们统计了在Lcorp中出现超过5次、实现评价和展望功能的全部it词块，发现这些序列中存在相当数量的如下词语：

（1）难易度：difficult, easy, impossible, challenging

（2）积极意义：important, advantage(s), good, great, key, broad, desirable, effective, feasible, first, interest(ing), large, main, many, most, new, significant(ce), suitable, valuable, wide

（3）性能或用途：used, able, applied, application, capable, meet

（4）现状或趋势：can, will, become, achieve, expected, improve, provide, promote, result

（5）强化词：very, widely, utmost

（6）委婉表达：possible, certain, could

除了第（6）类表达“不确定”的模糊限制语以外，其余词语在不同程度上反映了中国学习者在摘要语篇中采取主动的话语态势，强调研究意义、价值和贡献，凸显应用前景和社会影响力。这是中国研究生在撰写毕业论文时所采取的典型说服性策略。

5. 结语

本研究利用中国和本族语理工科研究生所撰写的学位论文英文摘要，调查了所有以it为主语的高频4词短语。研究生在在读期间研读大量科技文献，接受学术语言训练，对科技英语常见句型已经有了相当程度的认识和积累。从数量上看，中国学习者对第三人称代词it有强烈的使用意愿，展现了去作者化、去主观化的写作技巧，符合理工科学术英语写作的基本特点。从质量上看，学习者在型式和词语的相互选择上与本族语研究生仍有一定差距，在语篇衔接方面未能坚持遵循回指的基本原则，在评价性语言的使用上显示出特有的论述策略和习惯。这些发现对于学术英语写作教学和训练有积极的启示和意义。

参考文献

Ädel, A. 2014. Selecting quantitative data for qualitative analysis: A case study connecting a lexicogrammatical pattern to rhetorical moves [J]. *Journal of English for Academic Purposes* (16)16: 68-80.

Bhatia, V. K. 1993. *Analysing Genre: Language Use in Professional Settings* [M]. London: Longman.

Biber, D. *et al*. 1999. *Longman Grammar of Spoken and Written English* [M]. London: Longman.

Groom, N. 2005. Pattern and meaning across genres and disciplines: An exploratory study [J]. *Journal of English for Academic Purposes* 4(3): 257-277.

Hasselgård, H. 2009. Thematic choice and expressions of stance in English argumentative texts by Norwegian learners [A]. In K. Aijmer (ed.). *Corpora and Language Teaching* [C]. Amsterdam/Philadelphia: John Benjamins Publishing Company. 121-139.

Herriman, J. 2000a. Extraposition in English: A study of the interaction between the matrix predicate and the type of extraposed clause [J]. *English Studies 81*(6): 582-599.

Herriman, J. 2000b. The functions of extraposition in English texts [J]. *Functions of Language* 7(2):203-230.

Herriman, J. 2013. The extraposition of clausal subjects in English and Swedish [A]. In K.

Aijmer, & B. Altenberg (eds.). *Advances in Corpus-based Contrastive Linguistics: Studies in Honour of Stig Johansson* [C]. Amsterdam/Philadelphia: John Benjamins Publishing Company. 233-259.

Hewings, M. & A. Hewings. 2002. "It is interesting to note that...": A comparative study of anticipatory 'it' in student and published writing [J]. *English for Specific Purposes* 21(4): 367-383.

Hunston, S. 1993. Evaluation and ideology in scientific writing [A]. In M. Ghadessy (ed.). *Register Analysis: Theory and Practice* [C]. New York: Pinter Publishers. 57-73.

Hunston, S. & G. Francis. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English* [M]. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hunston, S. & J. Sinclair. 2000. A local grammar of evaluation [A]. In S. Hunston & G. Thompson (eds.). *Evaluation in Text: Authorial Stance and the Construction of Discourse* [C]. Oxford: Oxford University Press. 75-101.

Hyland, K. 1998. Persuasion and context: The pragmatics of academic metadiscourse [J]. *Journal of Pragmatics* 30(4): 437-455.

Hyland, K. & P. Tse. 2005. Hooking the reader: A corpus study of evaluative that in abstracts [J]. *English for Specific Purposes* 24(2): 123-139.

Larsson, T. 2016. The introductory it pattern: Variability explored in learner and expert writing [J]. *Journal of English for Academic Purposes* 22: 64-79.

Lores, R. 2004. On RA abstracts: From rhetorical structure to thematic organization [J]. *English for Specific Purposes* 23(3): 280-302.

Quirk, R. *et al*. 1985. *A Comprehensive Grammar of the English Language* [M]. London: Longman.

Römer, U. 2009. The inseparability of lexis and grammar: Corpus linguistic perspectives [J]. *Annual Review of Cognitive Linguistics* 7(1): 140-162.

Salager-Meyer, F. 1990. Discoursal flaws in Medical English abstracts: A genre analysis per research- and text-type [J]. *Text* 10(4): 365-384.

Sinclair, J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: Oxford University Press.

Swales, J. M. 2006. *Genre Analysis: English in Academic and Research Settings* [M]. Cambridge: Cambridge University Press.

Thompson, P. 2009. Shared disciplinary norms and individual traits in the writing of British undergraduates [A]. In M. Gotti (ed.). *Commonality and Individuality in Academic Discourse* [C]. Bern: Peter Lang. 53-82.

胡文杰、李晶洁，2015，学术语篇中的外置it立场序列——基于语料库的实证研究[J],《东华大学学报》(社会科学版)(2)：61-67。

张 乐，2017，学术英语中文际性句干的隐性评价特征研究[J],《解放军外国语学院学报》(5)：20-27。

张 乐、刘 芹，2017，中国理工科大学生英语写作语料库的设计、构建与前景[J],《当代外语研究》(3)：80-83。

通讯地址： 200093 上海市军工路334号上海理工大学外语学院

线性单位语法框架下的学术英语口语词块研究

复旦大学 张绪华

提要：《线性单位语法：综合口语与笔语》（Sinclair & Mauranen 2006）摆脱传统语法范畴的束缚，突破传统语法研究的上限单位—句子和下限单位—单个词语，关注由词语自然结合形成的词块以及由词块组成的真实话语。本文尝试在线性单位语法框架下，以语音连贯性作为词块整体性的判断标准，即以停顿信息作为词块的边界标识，将学术口语语料切分为词块，并就词块的复现情况、长度信息，以及词块自动切分方法进行讨论。

关键词：线性单位语法、词块、语料库、学术英语口语

1. 引言

《线性单位语法：综合口语与笔语》（*Linear Unit Grammar: Integrating speech and writing*，以下简称LUG）由John Sinclair和Anna Mauranen合著，John Benjamins出版社2006年出版。其目的是弥补大多数传统语法只针对笔语而忽视口语的缺陷，针对传统语法没有给予语言线性特征足够的重视，不能解释会话者如何共同构建会话，不关注语言的线性特征和词块研究等问题。于是提出从语言自然的线性组合关系出发，采用数据驱动方法进行文本切分。经过LUG分析，不同变体的英语文本都可以成为被传统语法接受的标准文本。LUG在口语与语法之间架设桥梁，与传统语法互为补充，可以作为教学语法的一部分。

目前常见的词块研究方法包括基于解读索引行（Sinclair 1991，2004；Stubbs 2001；卫乃兴 2011；甄凤超&王华 2012），基于短语复现统计（Biber & Conrad 1999；Biber *et al.* 2004；Hyland 2008），以及基于语言认知（Divjak & Gries 2008）等方法，从不同的角度对多词单位现象进行了解读。不过，除了在多词单位中占比较小的习语外，其他词块的存在形式和确切边界仍无法证实。《线性单位语法：综合口语与笔语》只是LUG的初始研究，包括文本切块在内的语料分析都依赖于人工处理，其机器自动实现还有待摸索。

以语言流利性为探究目标的学者认为，口语中的停顿和重音等语音连贯性信息，可以作为词块的边界标识（Pawley 1986；Riggenbach 1991；Hickey

1993；Dahlmann & Adolphs 2007）。Dahlmann & Adolphs（2007）以词块复现信息为起点，以口语语料库中高频出现的词块‘I don’t know’和‘I think I’为例，结合语料库中标注的停顿信息，探讨以停顿为标识判断词块边界的可行性。研究发现，停顿是判断词块边界存在的充分非必要条件，即在文本中有理由视为词块边界的地方，存在停顿信息缺省的情况。另一方面，作者认为，停顿为词块边界标识提供极有价值的指示，发生停顿的地方极有可能是词块边界（*ibid*：55）。鉴于该研究所采用的语料规模较小，且来源为非英语本族语者，我们认为有必要在更大规模的英语本族语语料库中，对停顿作为词块边界标识现象进行观察和分析。

迄今为止，国内外仍少见针对该理论的探讨或应用该理论的研究。Mauranen（2009，2016）以LUG为基础，分别探讨了口语的交互性和时间性特质。濮建忠（2009）对LUG做了全面的综述。Huang（2013）应用LUG理论和方法，对英语口语中的话语标记进行了研究。该理论研究和应用匮乏的主要原因在于目前尚无可自动实现文本切块及赋码的可行方法，导致无法利用大规模数据进行研究。有鉴于此，本文第一部分对LUG进行回顾，重点关注其文本切块的思想及方法，第二部分针对大规模文本切块研究的缺失，尝试使用学术英语口语中的停顿信息对语料进行切分，并讨论切分后的词块特征，着重探讨词块复现情况和长度特征。最后，根据对复现词块的功能特征的分析，提出进一步完善文本自动切块的方案。

2. 理论背景

语法研究怎样应对语言的线性机制？会话者如何共建会话？如何解释“不合乎语法”的文本？语言使用者在通过语言交流时怎样表达意义？在读和听时怎样把文本和意义联系起来？这是LUG所要回答的问题。作为Sinclair教授生前最后一部著作，本书包含着对其以往研究思路的反思和对未来研究的探索。书名中的几项信息：线性、词块单位、综合口语与笔语的语法研究，体现了LUG的特点和基本思想。

线性：所有语言共享的基本特性是呈线性排列，但是包括生成语法在内的大多数语法都是针对语言的层级性进行研究，强调语言的聚合性质。而在LUG中，组合关系是最基础的语言项目关系。共选发生于组合层面，是主要的选择因素，语言由组织词块（organizational-oriented units）和信息增量词块（message-incremental units）交替组合而成；聚合层面则是进行微调或次要选择的较低层面。根据该思想，语言的线性属性远比层级属性重要，因此称为线性语法。

单位：在词块单位切分问题上，可切块性（chunkability）是语言文本的基本特性，即文本可以被切分为若干与文本意义相一致的词块。词块单位切分是整个

理论赖以存在的基础，LUG就是建立在这样一个假设之上：语言使用者在感知语言时会自然而然、不可避免地切分词块，因此对任意语言使用者来说，任意文本都可分割为若干词块。尽管对词块起始的划分会有不同意见，但总体上语言使用者针对同一文本所划分的词块非常相似，词块长度不会超过四五个单词的长度，因此推测词块是人类语言中存在的自然单位。对于词块的判断，与典型的Sinclair式数据驱动研究方法不同的是，LUG通过语言使用者对文本进行的逐词分析，凭借语言直觉切分词块。虽然具体方法不同，但作者们始终坚持Sinclair式数据观，对真实语言数据进行研究；保持数据原生态，对“干净文本”进行研究，避免传统语法规则先入为主。

综合口语与笔语：传统语法的分析对象主要是标准、规范、合乎语法的笔语文本，但是人们在日常生活中用以交流的口语中充斥着大量不规范的、不能够通过传统语法分析的语言片段。LUG的首要目标就是弥补口语语法的缺失，把非正式口语，包括任何可以被看作英语实例的文字序列作为研究对象，加强对口语与笔语共性的认识，在“不同的英语变体之间搭建桥梁”（Sinclair & Mauranen 2006：3）。通过对语料样本的分析，作者展示了LUG应用于各种英语变体研究的可行性。

语法研究：以往的语法研究主要有两种类型，针对抽象的语言系统和语言的层级性，以规约性的、自上而下的方式研究语言的抽象语法；针对真实发生的、标准的、符合语法的语言事实和语言的线性特性，避免传统语法框架的先入为主，以自下而上的方式描述语言的数据驱动语法。前者以教学语法为代表，后者以型式语法（Hunston & Francis 2000）为典型，其他各种语法存在于这两种极端类型语法之间。LUG与数据驱动语法较为接近，并且能够弥补型式语法的三项不足：（1）只能分析标准的、合乎语法的语言文本；（2）只以动词、名词、形容词的型式为中心进行研究；（3）无法顾及小句之外的语法特征等缺陷。

本书中阐释的很多思想并非首创，创新点在于选择并糅合这些思想，使之成为一个语法理论和分析系统。始于半世纪前美国结构主义的直接成分分析理论，对于离心结构和向心结构的分析已经颇具词块分析的雏形。Brazil（1995）对口语文本的线性、实时的描述，与LUG最为接近（Sinclair & Mauranen 2006：27），对线性分析和词块研究有一定的启发和影响。作者接受了Brazil的基本观点，即话语意义的产生是在话语产生的过程中而不是话语完成之后，还接受了他的术语“增量”（increment），并将其作为本书所分析的语言结构的一个主要类别。

LUG分析文本的详细步骤如下：

步骤一，面对没有标点的英语文本，研究者通过语言直觉切分词块，每个词块的边界用临时单位边界（provisional unit boundary，简称PUB）标注，两个PUB之间就是一个词块（Sinclair & Mauranen 2006：86）（例1）。

例 1：Mr. Kennedy / now / declares / that / it must be / bold / in its thinking / and / ready to plan / long-term. / Sounding nice / is no longer enough, / he argued.

步骤二，根据不同功能，将词块分为两类。从增加共享经验的观点出发，部分词块被认为是信息增量词块（简称M），另一部分对文本进行管理，被称为组织词块（简称O）。

步骤三，根据组织词块的不同功能，将其分为文本组织词块（OT）和交互组织词块（OI）。这两种组织词块可以与衔接（cohesion）和连贯（coherence）联系起来进行区别：OT可以理解为使用连接词的务实衔接，OI则是不使用连接词的务虚连贯。

步骤四，根据对增进共享经验所做出的不同贡献，信息增量词块又被进一步细分，如例2。

例 2：Mr. Kennedy(M-) / now(OT) / declares(+M-) / that (OT) / it must be(M-) / bold(+M) / in its thinking(MS) / and(OT) / ready to plan(MS) / long-term. (MS) / Sounding nice(M-) / is no longer enough, (+M) / he argued. (M)

最后一个步骤，通过移除交互组织词块以及合并分散的信息增量词块，使原始文本成为连贯，结构合理并且符合传统语法标准的文本。

作者认为LUG可以朝两个方向发展：首先可以作为文本浅层或局部附码器。经过其分析的文本以两种不同的格式呈现：一是简洁、流畅、规则、无标注，容易被读者接受的“干净文本”，即经过LUG第五步分析得出的结果；二是作为信息科学的入口产品，它是包含词块全部标注信息的标注文本，即LUG第四步分析得出的结果。其次从语言的线性关系出发对语言进行描述，揭示以往语法研究所不能观察到的语言事实。

3. 学术英语口语中的词块

本研究以英国学术英语口语语料库（British Academic Spoken English Corpus，简称BASE）（Thompson & Nesi 2001）为语料，基于停顿可以作为边界标记，观察以此区分的词块特征。BASE是由英国华威大学和雷丁大学共同开发的学术英语口语语料库，涵盖艺术与人文、生命与医学科学、自然科学，以及社会研究与科学等四个大的研究领域，共包含160个讲座和40个研讨会音视频的文字转写。语料库建设者在转写语料的过程中，将停顿的位置和时长信息进行了标注。本研究中，我们将BASE语料库158个[1]讲座文本的停顿信息（>0.2秒）视为词块边界

1 BASE语料库中的所有的研讨会文本以及两个讲座文本（lslct035和lslct036）没有停顿信息标注，本研究针对其余158个讲座文本进行研究，库容为1,194,045个单词。

标识，即以语音连贯性作为词块整体性的判断标准，将文本切分为200,223个词块形符，共160,520个词块类符。接下来，我们以LUG为基础，参考Pawley & Syder（2000），Biber *et al.*（2004），以及Hyland（2008）等研究，针对包括词块的复现和长度等基本信息进行分析，并探讨文本切块自动实现的可能方案。

3.1 BASE语料库中词块的复现率

BASE语料库中词块重复率较低，复现两次及以上次数的词块数量较少，有超过77.0%的词块仅出现一次（表1）。总体来看，词块的复现率与其长度成反比。以长度为1至6词的词块为例，仅出现一次的词块占比随着词块长度的增加而增加（10.6%-99.2%）（表1），也就是词块长度越长，复现率越低。表1显示，有10.6%的1词词块仅出现一次，与此形成鲜明对比的是，99.2%的6词词块都只出现一次。需要说明的是，本研究目前仅以停顿信息作为词块边界的标示。在目前的切分基础上，以LUG的文本切块方法，进一步人工切分词块，势必会得到更多的较短词块，人工切分后的词块复现率有待后续研究观察。

表1. BASE语料库词块复现信息

词块长度	形符	类符	仅出现一次的词块数量	仅出现一次的词块/形符
1词	34,299	5,715	3,619	3,619/34,299=0.106
2词	22,378	14,814	12,615	12,615/22,378=0.564
3词	21,644	19,058	17,762	17,762/21,644=0.821
4词	19,631	18,922	18,456	18,456/19,631=0.940
5词	17,170	17,019	16,896	16,896/17,170=0.984
6词	14,975	14,908	14,848	14,848/14,975=0.992
全部词块	200,223	160,514	154,236	154,236/200,223=0.770

以3词和6词词块为例，BASE语料库中仅有17.9%和0.8%的复现3词和6词词块，其中最高频复现的21个3词词块在每百万词中出现10词以上（例如，*in other words*，*and so on*，*in terms of* 和 *and of course*），而绝大多数（93%）复现6词词块在每百万词中仅出现2次（例如，*all the way to the end*，*along the straight line demand curve* 和 *and all that sort of thing*）（表2）。

表2. BASE语料库中高频复现词块

	1词	2词	3词	4词	5词	6词
1	okay	and er	in other words	on the other hand	and at the same time	and so on and so forth
2	and	you know	and so on	thank you very much	and you can see that	at the end of the day
3	so	and then	in terms of	and you can see	can you see a problem	in the direction of the fibres
4	yeah	all right	and of course	at the same time	it's not a trick question	to be marked on the merits
5	the	and so	first of all	lambda at time T	what we want to do	all the way to the end
6	but	the er	if you like	on the one hand	after the Second World War	along the straight line demand curve
7	that	is that	i don't know	that is to say	and if you do that	and all that sort of thing
8	now	which is	so for example	is going to be	and the reason for that	and as we said last week
9	of	i mean	and that is	does that make sense	and what i've got here	and i'll come back to that
10	is	okay so	in order to	or something like that	are there any questions about	and not just to abstract figures

BASE语料库中的词块复现分布趋势与Biber *et al.*（2004）和Hyland（2008）等人基于短语复现统计方法发现大量复现3至6词的词块不同，说明尽管部分词块会在语料中以固定形式大量、反复出现，但并没有在语言使用中，以独立词块的形式出现，或者其内部结构并没有如数字所体现的紧密型，其边界具有一定的模糊性。因此，仅以词语组合的复现频率作为词块单位的判断标准值得商榷。

3.2 BASE语料库中的词块边界

关注复现词块在语料库中出现频率的研究，通常聚焦最高频复现词块的频次

和结构信息，并将此类词块的边界作为真实存在的词块边界，以及词块以独立形式存在的证据。以Biber *et al.*（2004）以及Hyland（2008）的研究为例，‘if you look at’和‘as can be seen’分别出现在两项研究题目中，研究者将其作为典型高频复现词块进行了分析。

然而，本研究中的数据显示，‘if you look at’和‘as can be seen’在BASE中几乎很少以独立词块（分别出现4次和0次）的形式出现。‘if you look at’在BASE中出现122次，其中，仅有4次以独立词块形式出现，另有3次以其结尾，例如‘now if you look at’和‘so if you look at’。其余115个包含‘if you look at’的词块中，7个以‘if you look at the’结尾，49个以‘if you look at’+名词（或名词词组）结尾。尽管‘if you look at’+名词（或名词词组）（例如，*the way*, *the language*, *table two*, *page fifty-nine*, *all the entries*, *the studies*, *er a book*）型式高频出现，不过‘if you look at’与每一个搭配名词（或名词词组）都只出现1次，因此在以复现频率为基础的研究中，通常被研究者所忽略[1]。因此，‘if you look at’可以视为高频复现型式‘if you look at’+名词（或名词词组）中结构完整的组成部分。针对词块‘if you look at’边界的划分，或者说是否应被其视为独立词块需要基于更大规模的语料进一步论证。

3.3 BASE语料库中的词块长度

BASE数据显示，词块的平均长度为5.96个单词，最短词块为1词，最长词块为53词。长度超过10词的词块为31,214个，占总词块的15.6%。显而易见的是，大部分长度超过6词的词块都可以进一步切分为两个或两个以上的词块。例如，根据LUG的词块切分方法，6词词块‘all the way to the end’可以进一步切分为‘all the way’和‘to the end’两个词块。

造成较长词块存在的原因是多方面的：首先，我们仅仅根据停顿信息（unfilled pauses）对文本进行了切分，忽略了其他的填充式停顿（filled pauses）。例如，单词重复和er以及erm等填充词被都研究者视为停顿。本文未针对填充式停顿进行文本切块的原因在于，学界对于填充式停顿的定义和范围尚存不同意见；其次，研究者在对BASE文本的转写过程中，只标注了时长在0.2秒及以上的停顿，这也与Goldman-Eisler（1968）以及Towller *et al.*（1996）的观点相近，即将0.2-0.3秒以上的停顿视为停顿。然而，是否存在更短时长停顿的可能性，即持续时间小于0.2秒的停顿，有待确认。此外，语料转写过程中也可能会遗漏部分停顿信息，造成词块切分出现误差。

Sinclair & Mauranen（2006：6）通过观察语料样本中的词块，认为词块长度

1 不可否认的是，本研究所使用的口语语料库规模有限，在一定程度上导致了搭配词频率较低。

不会超过4、5个单词的长度。Pawley & Syder（2000）提出了one-clause-at-a-time的假设，认为实际交流中所使用的流利单位平均为6个单词。如果根据LUG文本切块的标准以及已有的停顿信息，将长度较长词块进一步切分，不难发现，词块的长度一定会低于6个单词，可能会与Sinclair和Mauranen的估计相近，长度约4、5个单词。

3.4 文本切块的自动实现

LUG思想的几个重要方面都曾经出现在Sinclair不同时期的学术思考中。Sinclair & Coulthard（1975）针对课堂口语语篇的研究，没有局限于传统语法的研究上限单位—句子，而是超越句子，对会话者共同构建的口语语篇进行分析。会话参与者通过竞争与合作，完成构筑话语结构并完成沟通，会话产生的话语基本结构有规律可循。但是在随后的话语分析研究中，研究者更多关注口语与笔语文本之间的差异以及会话参与者之间的竞争关系，而对口语与笔语文本之间的共性以及话语参与者之间通过合作维系会话、保持会话连贯等方面缺乏足够的关注。因此作者希望通过描述会话中的话语合作，以及语言交互作用产生意义，来平衡会话中合作与竞争关系。在LUG分析的第二个步骤中，作者根据词块的不同功能，把组织词块分为OT和交互组织词块OI，其中OI就是会话参与者合作维系会话的产物。此外，在探索语言意义单位的过程中，Sinclair（1991）发现在以往的语言研究中，由于传统语法严格区分语法与词语或者语义研究，重视语言聚合关系的研究，导致词块研究被忽视。因此他倡导语料库驱动的词语研究，关注语言的组合关系、词语的共现、词语搭配以及多词单位，并提出习语原则（语言使用者使用大量预制、半预制短语）以及短语学倾向（Sinclair 1996）（词语聚集在一起通过组合产生意义）等一系列重要的语言学机制。

在延续以上研究思路的基础上，LUG重点讨论文本切块、词块的功能和意义，以及两者之间的关系。文本切块以及词块功能赋码的自动实现是该研究未来的发展方向。Sinclair & Mauranen（2006：158）探讨了机器自动实现文本切块和赋码的可能性，并设想了自动实现的步骤：由人工分析第二步，而不是第一步起始（参见本文第一部分）。其理由是组织词块的组成遵循习语原则，而且数量较少，内部稳定性较高，容易被机器识别，机器很难做到自动切分词块，区分组织词块和信息增量词块却容易很多。移除组织词块的剩余部分就是信息增量词块，对于信息增量词块的识别和切分，可以从α-phrase、质变理论以及离心结构和向心结构等理论中寻求帮助（Sinclair & Mauranen 2006：159）。在应用方面，LUG为外语学习者提供符合传统语法的标准文本，有助于译者关注信息增量词块，加速信息提取。该研究思路的问题在于，组织词块的长度和数量目前难以确定，因此无法用机器对其自动识别。

表2中展示的是在BASE语料库中复现率最高的长度为1至6词的词块。根据LUG的分类标准，其中的1词、2词以及3词词块大多为组织词块（例如，*and*，*so*，*and then*，*okay so*，*in other words*和*and of course*），以及少量信息增量词块（例如，*now*）。我们认为这些长度较短，组织功能或者增加新信息倾向较为明显的词块，可以作为词块自动切分的“锚点”，用以判断词块边界，并以此为依据，进一步加工长度较长、包含多个词块的语言片段。例如，‘and er’和‘now’作为独立词块，在BASE中分别出现273次和535次，具有组织功能和信息增量的功能。此外，两个词块分别在其他词块中出现644次和1506次，例如‘*affected the shapes of the letters and eri'll show you one example of this*’和‘*a carpenter come round the other day now we wanted a cupboard for our bathroom*’。

在后续研究中，可以首先依据停顿信息对文本切块，然后尝试以‘and er’和‘now’为“锚点”，将以上两个例子分别切分为‘affected the shapes of the letters and’，‘and er’和‘i’ll show you one example of this’，以及‘a carpenter come round the other day’，‘now’和‘we wanted a cupboard for our bathroom’等较短的词块。应该先以长度较长的词块作为“锚点”切分词块，然后以长度次之的词块切分，并以此类推。例如，首先以表2中长度为6词的‘and so on and so forth’作为“锚点”，随后依次选择同样出现在其中的‘and so on’以及‘and so’作为“锚点”切分词块。而在整个过程中，需要人工判断、筛选，找到不易造成错误切分的词块作为“锚点”。

以停顿信息寻找词块边界，为词块自动切分提供了另一种可能性：以停顿信息对文本初步切分，识别其中高频组织词块和信息增量词块，并利用这些词块，将其作为“锚点”，对文本进一步切分，最后应用Sinclair & Mauranen（2006:159）所提及的理论对信息增量词块和剩余的组织词块进行辨识。

4. 小结

本文在回顾LUG的基础上，尝试以学术英语口语中的停顿信息对语料进行切分，并分析了切分后词块的复现和长度情况。LUG的核心思想在于：（1）从语言的组合关系出发，以自下而上的方法，把语言当作线性单位进行研究，避免传统语法研究中关于词类、结构等一切传统概念；（2）强调语言的可切块性，认为语言文本由增加信息共享的信息增量词块和对文本进行管理的组织词块交替组合而成；（3）弥补大多数传统语法只针对笔语而忽视口语的缺陷，把线性文本与语法研究相结合；（4）将词块研究与话语分析研究相结合。

LUG针对语言线性的研究视角，其自下而上的研究方法以及综合口语和笔语、将词块研究与话语分析相结合的研究方法都值得我们重视和思考。LUG将词

块视为语言中存在的自然单位，语言使用者依据直觉，将文本切分为词块。这种奠定了整个研究基础的文本切块方法，迥异于以往的多词单位研究方法，为词块研究提供了一个全新视角。

我们发现，根据停顿信息切分的词块的重复率较低，且平均长度小于6个单词。词块的复现率与其长度成反比关系，即词块越短，其被语言使用者用作独立词块的概率越高。此外，停顿信息为定位词块边界，以及词块自动切分提供了新的解决方案。后续研究可以尝试以通过停顿信息识别且长度较短的词块为“锚点”，对长度较长词块切割，并对完整的切块结果做进一步的功能和意义的分析。

参考文献

Biber, D. & S. Conrad. 1999. Lexical bundles in conversation and academic prose [J]. In H. Hasselgard & S. Oksefjell (eds.). *Out of corpora*: *Studies in Honor of Stig Johansson* [C]. Amsterdam: Rodopi. 181-190.

Biber, D., S. Conrad & V. Cortes. 2004. If you look at…: Lexical bundles in university teaching and textbooks [J]. *Applied linguistics* 25(3): 371-405.

Brazil, D. 1995. *A Grammar of Speech* [M]. Oxford: Oxford University Press.

Dahlmann, L. & S. Adolphs. 2007. Pauses as an indicator of psycholinguistically valid multi-word expressions (mwes)? [A]. In *Proceedings of the Workshop on a Broader Perspective on Multiword Expressions* [C]. Oxford: Oxford University Press. 49-56.

Divjak, D. & S. T. Gries. 2008. Clusters in the mind?: Converging evidence from near synonymy in Russian [J]. *Mental Lexicon* 3 (2): 188-213.

Goldman-Eisler, F. 1968. *Psycholinguistics: Experiments in Spontaneous Speech* [M]. Salt Lake City: American Academic Press.

Hickey, T. 1993. Identifying formulas in first language acquisition [J]. *Journal of Child Language* 20 (1): 27-41.

Huang, L. F. 2013. The use of Linear Unit Grammar (LUG) in the investigation of discourse markers in spoken English [J]. *International Journal of Language Studies* 7 (3):139-156.

Hunston, S. & G. Francis. 2000. *Pattern Grammar: A Corpus-driven Approach to the Lexical Grammar of English* [M]. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Hyland, K. 2008. As can be seen: Lexical bundles and disciplinary variation [J]. *English for Specific Purposes* 27(1): 4-21.

Mauranen, A. 2009. Chunking in Elf: Expressions for managing interaction [J]. *Intercultural Pragmatics* 6 (2): 217-233.

Mauranen, A. 2016. Temporality in Speech–Linear Unit Grammar [J]. *English Text Construction* 9(1): 77-98.

Pawley, A. 1986. On speech formulas and linguistic competence [A]. In F. Shehdeh & M. Hoey (eds.). *Kansas Working Papers in Linguisitics* [C]. vol 11:57-87.

Pawley, A. & F. H. Syder. 2000. The one-clause-at-a-time hypothesis [A]. In Riggenbach, H. (ed.). *Perspectives on Fluency* [C]. Ann Arbor: University of Michigan Press. 163-199.

Riggenbach, H. 1991. Toward an understanding of fluency: A microanalysis of nonnative speaker conversations [J]. *Discourse Processes* 14 (4): 423-41.

Sinclair, J. 1991. *Corpus, Concordance, Collocation* [M]. Oxford: Oxford University Press.

Sinclair, J. 1996. The search for units of meaning [J]. *Textus* 9 (1): 75-106.

Sinclair, J. 2004. *How to Use Corpora in Language Teaching* [M]. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Sinclair, J. &M. Coulthard. 1975. *Towards an Analysis of Discourse: The English Used by Teachers and Pupils* [M]. Oxford: Oxford University Press.

Sinclair, J. & A. Mauranen. 2006. *Linear Unit Grammar: Integrating Speech and Writing* [C]. Vol. 25. Amsterdam/Philadelphia: John Benjamins Publishing.

Stubbs, M. 2001. *Words and Phrases: Corpus Studies of Lexical Semantics* [M]. Oxford/Malden: Blackwell Publishers.

Thompson, P. & H. Nesi. 2001. Research in progress, the British academic spoken English (Base) corpus project [J]. *Language Teaching Research* 5(3):263.

Towell, R., R. Hawkins & N. Bazergui. 1996. The development of fluency in advanced learners of French [J]. *Applied Linguistics* 17 (1): 84-119.

濮建忠，2009，意义单位探索——《线性单位语法》评述[J]，《外语教学与研究》(2)：153-55。

卫乃兴，2011，再探经典短语学的要旨和方法：模型，概念与问题[J]，《外语与外语教学》(3)：29-34。

甄凤超、王　华，2012，学习者英语中语法词的短语学特征：以 of 为例[J]，《外语教学与研究》(3)：389-401。

通讯地址： 200433 上海市杨浦区邯郸路220号复旦大学外国语言文学学院

论语料库中的语用标注*

西安外国语大学 姜占好

提要：语料库因数据的真实性和翔实性为研究人员和教学人员所青睐，保障数据真实性的语料库语用标注也逐渐成为研究热点，从而诞生了系列研究成果。本文从语用标注必要性、语用标注对象、挑战和展望三个维度梳理了相关文献，列举了话语标记词、言语行为、韵律、隐含义等传统语用学研究话题的标注方法，指出了未来语用标注研究面临的问题和未来研究的切入点。

关键词：语料库、标注方法、语用标注

1. 引言

语料库标注是对原始的口语或书面语进行语言信息加工，并把表示语言特征的信息赋码添加在相应成分上，以方便计算机识别的行为（Leech 2005；Lu 2014）。标注是原始语料实现机读化的关键步骤，也是语料库语言学研究中的重要课题（Kennedy 1998；崔刚、盛永梅2000；McEnery & Hardie 2011；李文中 2012；Kirk & Andersen 2016）。标注的种类有词性标注、词汇标注、语音标注、语义标注、语用标注、句法标注、语篇标注和语体标注（Garside, et. al. 1997; Leech 2005; Kübler & Zinsmeister 2015），其中语用标注（Kreuz & Riordan 2011; Weisser 2015）随着语用学和二语习得界面研究的增多（Rose & Kasper 2001; Kasper & Rose 2003; Taguchi & Sykes 2013; Ross & Kasper 2013），以及语料库能够提供真实的语料（Rühlemann 2010; Romero-Trillo 2014, 2015, 2016）而受到越来越多的关注，从而诞生了系列研究成果。本文拟从语用标注的必要性、标注对象和方法对这些成果进行梳理和论述，并对语料库的语用标注进行展望。

2. 语用标注的必要性

现有的语料库多收集自然发生的语料，有书面语，也有口语。但是在建库的过程中，一些非文本的信息，如字号、字体、断行、图片和照片没有收录或赋码，

* 本文系国家社科基金项目“英语专业学生语用能力多模态语料库的研发和应用”（16XYY010）的阶段性成果，同时感谢西安外国语大学“应用语言学多维跨域科研创新团队”和2016富布莱特高级访问学者项目资助（GRANTEE ID 68160218）。

而这些信息的缺失削弱了分析者对信息的解读；这在口语语料库中更是如此：转写的口语文本脱离了声音、声调、音质，而成了孤零零的文本，无法为使用者提供丰富动态的语境；即使有的语料库提供了语境信息，也只是参与者性别、年龄和阶层的基本信息（Rühlemann 2010），而不是诸如言语行为（权利、地位和强加度）、合作原则、指示词、礼貌程度和会话含义的系统的语用信息，更没有面部表情、体态语等副语言信息。而想达到语料中语言形式和语用意义的对应或链接，不仅需要传统的标注，如词性赋码，还需要对语言形式及其语境提供解释性的语言，即语用码或语用标注，以达到语言形式和语言功能的对等，使得建库时话语转写更加完善，从而更加客观地呈现原汁原味的话语语境（Kirk 2016）。

3. 语用标注对象和方法

语用标注受到语用学理论的影响，标注的内容聚焦在传统语用学研究的热点上，如话语标记词、言语行为、韵律、暗含义和语境等。

3.1 话语标记词

我们首先描述话语标记词的标注，原因在于在语料库中，标记词在形式和功能上有对应性，方便研究人员析出和提取。常见的标记词如英语中的*well*、*like*、*kind of*、*sort of*、*I mean*等等。

Samy和González-Ledesma（2008）首先根据话语标记词的话语功能，对标记词进行了大体的分类：假设、原因、结果、让步、列举、综合分析、概括、重新描述（reformulation）、话题（topicalization）以及共同争论（co-argumentation）。然后，他们利用PRAGMATEXT，一种可扩展标记语言（Extensible Markup Language-XML），对西班牙语、阿拉伯语和英语平行语料库中的话语标记词进行了标注（详见Samy & González-Ledesma 2008），标注方法是将平行文本中对应的3种语言形式的话语标记词加黑体和链接。

在Schiffrin (1987)，Stenström (1990)，Aijmer (1996；2002)研究的基础上，Kirk (2016)在**SPICE**（**S**ystems of **P**ragmatic annotation in the spoken component of **ICE**-Ireland, 爱尔兰语料库口语语用标注体系）里首先从语音层面（例如*ach*、*och*、*oh*）、词汇层面（例如*actually*、*oh-Lord*、*alright*、*okay*）和句法层面（例如*do you know*）对标记词进行了分类，标注的方法是在标记词的后面加上上标的五角星，例如*well*☆，对于两词及以上的标记词，使用连字符将各词串联，然后在最后一词上标添加五角星，例如*as-you-know*☆。鉴于语音层面和词汇层面的标记词易于理解，这里不再举例；下面引用Kirk使用的例子，说明句法层面标记词的语用标注：

(1) <ICE-NI-DEM-P2A-052$A><rep> 1WE’ve been eating 1potAto bread

for 1yEArs% and *do-you-know*☆ the 1wAy I normally 1cOOk it's% to 1pUt it in the 1pAn% and 1frY it in a 1tIny little bit of 1OIl% </rep>

在这个例子句，*do-you-know*将“we've been eating potato bread for years”和“如何制作土豆面包”的新信息链接起来，起到了语篇连接的作用。

3.2 言语行为

言语行为构成了语料库语用标注的重点内容。早在1989年，Blum-Kulka, House和Kasper在CCSARP（Cross-Cultural Speech Act Realization Project跨文化言语行为实施工程）项目中就开始对诸如“请求”、“道歉”等的言语行为分析，使用了一系列的术语来解析构成言语行为的组成部分。例如首先将言语行为分成两大部分：head act（主要行为）和support move（次要举动），然后对head act进行分类，对support move进行细分；次要举动表现如：“提醒语”（alerters，如“excuse me”），“原因根据”（grounders，“I really need it”），“降低需求”（minimizers，如“just a little”），“消除戒备语”（disarmers，如“I know you are really busy but …”）和“礼貌语”（politeness markers，如“please”）。虽然当时他们没有建立严格意义上的语料库，但是言语行为的分类和细化标签依然对后来的语用标注有借鉴作用。

1992年，心理学家Stiles为了更好地与病人进行交流，设计了8种“言语应答模式”（Verbal Response Mode），并用一个字母赋码相应的言语：Disclose（D）、Edification（E）、Advisement（A）、Confirmation（C）、Question（Q）、Acknowledgement（K）、Interpretation（I）和Reflection（R）。由于作者是心理医生，他的标注体系没有引起语言学界的多少关注（详见Stiles 1992: 17；Archer *et al.* 2008: 622）；后来，大部分标注者遵循了Searle（1969, 1976）的言语行为分类标准，将言语行为以动词为中心，分成了5类[1]：阐述式（representatives）、指令式（directives）、宣言式（declaratives）、陈言式（expressives）和承诺式（commissives），然后使用首字母或字母组合方式标注相应的言语行为（Leech & Weisser 2003; Weisser 2010）。

Leech & Weisser（2003）设计的言语行为标注体系以任务为中心，聚焦于接线员和服务电话中的言语行为，从标注单位分割、句法形式、话题（主题）、情态、和积极（消极）义五个维度出发，设计了近40个言语行为赋码形式，如<decl>（陈述句），<q-yn>（是非句），<q-wh>（特殊问句）<imp>（祈使句）<frag>（碎片结构），<dm>（话语标记词）等。

1 Leech (1983) 增加了“质问咨询式”(rogative)言语行为，例如动词ask、enquire、query和question。

Weisser（2015）在回顾了数据赋码形式后（如XML, SGML, TET）[1]，以实例方式对比了标注言语行为的三种体系：DAMSL、SWBD DAMSL、和DART[2]。例如对于如下的对话：

Do you have to have any special training?

<Laughter>, <Throat_clearing>,

三种标注体系标注方式是不同的：

DART使用 (tag +) act (+ mode)的模式：

q-yn+ reqInfo

<comment type = “laughter”/>, etc.

DAMSL 使用tag附标方式：

Info-Request

n/a

SWDB-DAMSL 也使用tag附标方式，但是更加详细：

YES–NO-QUESTION (qy)

NON-VERBAL (x)

在对比之后，Weisser（2015）特别指出DART标注体系的优点在于简洁明了。例如使用<q-wh>来标注wh-问句，使用*reqInfo*标注询问信息，使用*reqDirect*标注请求指令等等。Weisser为了证明DART语用标注体系的简洁性，通过个案分析，调查了Trainline语料库中英国某订票机构工作人员与客户的互动，用柱状图列举出工作人员使用的言语行为数量和种类，以证明标注体系和人际互动的一致性和简明性（详见Weisser 2015）。

Kirk（2016）遵循COCOA的标注风格，使用尖括号和反斜线对主要言语行为进行了标注，见例（1）中的<rep>…</rep>（阐述式）。其他4类言语行为对应的标注方式分别是：<dir> … </dir>（指令式）、<decl>…</decl>（宣言式）、<com>…</com>（承诺式）和<exp>…</exp>（陈言式）。这种标注的缺点在于其标注的对象聚焦于动词（Archer *et al.* 2008），而有的言语行为可以没有动词的出现。为此，Kirk (2016) 使用<icu>…</icu>来标注无法确定归属的言语行为（例如作为答语的*right*，见例（2）），用<soc>…</soc>标注社会用语（例如招呼语、离别语等，常见的有*hello, hi, how are you? Fine, not too bad*等），用<xpa>…</xpa>

1 XML-Extensible Markup Language; SGML-Standard Generalized Markup Language; TET-Text Encoding Initiative。

2 DAMSL-Dialogue Act Markup in Several Layers; SWBD DAMSL- Switchboard Dialogue Act Markup in Several Layers; DART- Dialogue Annotation and Research Tool。

标注在语用层面上无法分析的话语（例如自然对话里的只言片语，听不清的话语），用<K…>…</K…>来标注字面意义和所指意义不同的话语（见例3：*I'm not even sure exactly when I will send you my bill*，法官说给律师的话，字面上是承诺式，而实际上是指令式，以起到幽默的效果，详见Kirk（2016: 310）。）

（2）<ICE-ROI-TEC-P1A-098$A><#><rep> I'm not even sure 2exActly when I'll 2nEEd somebody from% </rep>
<$B><#><icu> 2Right% </icu>

（3）<ICE-NI-LEC-P2A-061$B><#><dirK> Yeah* <,> I'll 1sEnd you my 2bIll%
</dirK><&> laughter </&>

3.3 韵律

韵律（prosody）包括声调、音质、节奏、音长、重音等。韵律原不是语料库视角下研究"标准"英语的重点。但是，韵律的变化（如声调的起伏）在言外之意和话语立场的表达上起着至关重要的作用，正是通过声调的变化，交际双方能各得其所。不同的多模态语料库的韵律标注体系各异。例如LLC（London-Lund Corpus）和SEC（Spoken English Corpus）使用单词中插入斜线的方法表示元音变化，例如y\es, ye/s, ye/\s, ye\/s。SPICE-Ireland语料库则使用ToBI（Tones and Break Indices音调切分指数）进行韵律标注。该体系被修订为IViE（International Variation in English英语发音的国际变体），成为通行的韵律标注办法（Leech, 2005: 22）。SPICE共有11种不同的韵律标注方式，这里列举两例（详见Kirk 2016: 304）：

韵律1：H*L（降调）

1FIErce fighting in 1GrOzny%<ICE-ROI-BRN-P2B-012$A>

韵律2：L*H（升调）

I 1rEAd the Killachter 2MEAdow% <ICE-NI-CLD-P1B-002$A>

3.4 暗含义/推断义标注

语用学中人们常常利用会话合作原则（Grice 1971）的遵守与否，以及关联理论（Wilson & Sperber 2004）来推断隐含意义。Andersen（2001）和Archer（2002）基于这两种理论，研制出手工标注隐含意义的体系。例如Archer（2002：10）除了使用现有的四个次则（quality、quantity、relation和manner），又增加了coop和

ambi两个标签。Anderson（2001）基于“简化主义理论”（reductionist theory即关联理论），依靠“伦敦青少年对话库”，使用TACTweb软件和标记词标注体系，来判断话语标记词的隐含意义。

3.5 语境标注

作为语用标注体系构成要素，语境标注引起了建库者的极大兴趣。首先，Archer *et al.*（2008）提出6种语境因素：共现文本语境（co-text）、物理语境（如周日的教堂）、个人/社会语境（如交际者相互的关系、地位和权势关系）、认知语境（交际双方的共享知识或背景知识）、文化语境（如价值观、信仰、社会规约）和情景模式语境（语言本身能创造的语境，如小说中的文字创造的语境）；同时，他们根据现有的标注体系将语境标注分为两大类，静态语境标注和动态语境标注。静态的语境信息如：言语者特征、信道特征（口语、书面语）和任务类型，这些信息放在文件的开头，例如BNC语料库，其所遵循的标注体系是文本编码方案（Text Encoding Initiative）；根据EAGLES（Expert Advisory Group for Language Engineering Systems语言工程体系专家咨询小组）的调查，也有将静态的语境信息，单独制作成文档，链接在语料库中（Gibbon *et al.* 1997）。

对于动态的语境信息，比较典型的例子是Archer & Culpeper（2003），Archer（2005）制作的用于社会语用语料库的社会语用标注体系，旨在标识出话语层面上关联言语双方重要的语境信息，涉及语用学、语料库语言学和社会语言学。该体系认为通过标注信息，可以缩短文本和语境之间的距离，呈现动态的交际语境。例如：

> [$ Lord President, $] <u speaker= “s” spid= “s3tchar1001” spex= “m” sprolel= “j” spstatus= “l” spage= “9” addressee= “s” adid= “s3tchar1002” adrolel= “d” adstatus= “O” adage= “9”>**If this be all you will say**, </u>
>
> <u speaker= “s” spid= “s3tchar1001” spsex= “m” sprolel= “j” spstatus= “l” spage= “9” addressee= “m” adid= “x” adrole1= “n” adstatus= “x” adage= “x” >**then, Gentlemen, you that brought the Prisoner hither, take charge of him back again**. </u>
>
> (Archer 2005: 110)

这是一段社会语用语料库中的语料，是Lord President与King Charles I之间的对话，表明“谁在何时何地以何身份向谁讲了什么”，标注的含义分别是speaker= “s”（第一个说话者是单个人，不是群体），spid= “s3tchar1001”（指Lord President），spex= “m”（性别男），sprolel= “j”（法官），spstatus= “l”（地位高），spage= “9”（年龄为9），adrolel= “d”（King Charles I是defendant, 即辩护人），

adstatus= “O”（辩护人是皇室人员），spid= “s3tchar1001”（King Charles I）。</u>表明单个话语（第2个</u>表明法官转向另外一个人开始讲话）。

3.6 混合标注

标注者有时面对的语料需要多种标注方式。例如，Archer（2005）在创建早期现代英语（1640—1760）法庭互动语料库时，拓展了“社会语用标注体系”，使得标注体系包含了语言形式、言外之力、会话互动和语境维度。具体层面有：*initiation*、*response*、*response-initiation*、*report*、*follow-up*、*follow-up initiation*，具体例证如下（Archer *et al.* 2008: 631）：

Recorder: You made the Bed, did not you? [initiation]

Crook: I did. [response]

Recorder: Upon your Oath, what time of Night was it? [follow up-initiation]

Crook: I think it was nearer Eleven than Ten. [response]

[text omitted]

Kings Coun: What time of Night was it that he was making love to you? [initiation]

Crook: I think about Ten a Clock. [response]

Kings Coun: Time passed merrily away with you then. [follow up]

Rich: It was Twelve a Clock. [report]

4. 问题和挑战

文献梳理帮助我们发现已有研究从三个维度来衡量某语用标注体系：研究目的、数据载体（口语/书面语）和标注方法（手动标注还是[半]自动标注），同时我们发现研究中存在的问题：

1）对于话语语篇中丰富的语用信息进行标注，一方面要尽可能细化语用标注方案，这是因为，标注方案的细化有助于语用信息的描写更加具体明确；但另一方面，也要把握细化的度，这是因为过细的标注方案，会让语篇中某种语用标注信息量过少，以至于该标注方案存在的可能性会大打折扣。

2）现有的语用标注体系多以手动、人工操作为主，半自动的标注体系还比较少。（Weisser（2003）研发出SPAACy，能够对言语行为进行半自动化的标注，但是，标注的对象局限于以任务为中心的电话对话，更谈不上完全自动的语用标注体系。）

3）语言类型层面，现有的标注体系以英语为主，虽然不同文化中语用目的实现方式有差异，但修改或调整已有的英语语用标注方案，以用于其他语言，可以

促进我们在语用层面上对语言类型的进一步理解。

4）语用标注能促进语料库语言学和语用学之间的深层关系，但是标注只是阐释语料库中语言信息的一种手段而已，通过语料库，即便是标注详细的多模态语料库，我们也无法全部再现和穷尽原有的语言和语境信息。

5）由于各个研究目的不同，不同的研究人员的标注方案是不同的，所以未来的标注体系若是能够做到兼容，标注方案能够做到融合，无疑会让语用标注经济可行，让标注人员省时省力。

5. 结语

翔实的语料标注不仅是语料机读的前提，而且为语料的分析者提供了真实生动的语境信息。本文从必要性、标注对象、问题和展望三个维度梳理了国内外语用标注文献，发现受传统语用学研究内容的影响，标注对象局限于语用学的传统话题，如话语标记词、言语行为、隐含等；其次，因不同的研究目的，标注方案各异，迄今还没有统一的语用标注体系和方案，更谈不上稍作修改，就可以跨语言使用具有普适性的语用标注方案。这些问题无疑为语料库的标注提供了未来研究的素材，拓宽了语料库标注的研究视阈。

参考文献

Andersen, G. 2001. *Pragmatic Markers and Sociolinguistic Variation: A Relevance-theoretic Approach to the Language of Adolescents* [M]. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Aijmer, K. 1996. *English Conversational Routines* [M]. London: Longman.

Aijmer, K. 2002. *English Discourse Particles* [M]. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Archer, D. & J. Culpeper. 2003. Sociopragmatic annotation: New directions and possibilities in historical corpus linguistics [A]. In A. Wilson, P. Rayson & A. M. McEnery (eds.). *Corpus Linguistics by the Lune: A Festschrift for Geoffrey Leech* [C]. London: Peter Lang. 37-58.

Archer, D. 2002. Can innocent people be guilty? A sociopragmatic analysis of examination transcripts from the salem witchcraft trials [J]. *Journal of Historical Pragmatics*, 3(1): 1-29.

Archer, D. 2005. *Questions and Answers in the English Courtroom (1640-1760): A Sociopragmatic Analysis* [M]. Amsterdam/Philadephia: John Benjamins Publishing Company.

Archer, D., C. Jonathan & D. Matthew. 2008. Pragmatic annotation [A]. In A. Lüdeling & M. Kytö (eds.). *Corpus Linguistics: An International Handbook* [C]. Berlin: Walter de Gruyter.

Blum-Kulka, S., J. House. & G, Kasper. 1989. *Cross-cultural Pragmatics: Requests and Apologies* [M]. Norwood, NJ: Ablex Publishing Corporation.

Garside, R., G. Leech & A. McEnery. 1997. *Corpus Annotation: Linguistic Information from*

Computer Text [M]. London: Longman.

Gibbon, D., R. Moore & R. Winski. 1997. *Handbook of Standards and Resources for Spoken Language Systems* [M]. Berlin: Walter de Gruyter.

Grice, H. P. 1971. Meaning [A]. In D. Steinberg & L. Jakobvits. (eds.). *Semantics: An Interdisciplinary Reader in Philosophy, Linguistics and Psychology* [C]. Cambridge: Cambridge University Press.

Kasper, G. & K. R. Rose. 2003. *Pragmatic Development in a Second Language* [M]. Oxford/ Malden: Blackwell Publishers.

Kennedy, G. 1998. *An Introduction to Corpus Linguistics* [M]. London: Longman.

Kirk, J. M. 2016. The pragmatic annotation scheme of the SPICE-Ireland Corpus [J]. *International Journal of Corpus Linguistics*, 21(3):299-322.

Kirk, J. M. & G. Andersen. 2016. Compilation, transcription, markup and annotation of spoken corpora [J]. *International Journal of Corpus Linguistics*, 21(3): 291-298.

Kreuz, R. J. & M. A. Riordan. 2011. The transcription of face-to-face interaction [A]. In W. Bublitz. & N. Norrick(eds.). *Foundations of Pragmatics* [C]. Berlin: Walter de Gruyter.

Kübler, S. & H. Zinsmeister. 2015. *Corpus Linguistics and Linguistically Annotated Corpora* [M]. London: Bloomsbury Publishing.

Leech, G. 1983. *Principles of Pragmatics* [M]. London: Longman.

Leech, G. 2005. Adding linguistic annotation [A]. In M. Wynne (ed.). *Developing Linguistic Corpora: A Guide to Good Practice* [C]. Oxford: Oxbow Books. 17-29.

Leech, G. & M. Weisser. 2003. Generic speech act annotation for task-oriented dialogues [A]. In D. Archer, P. Rayson, A. Wilson & T. McEnery (eds.). *Proceedings of the Corpus Linguistics 2003 Conference* [C]. Lancaster: UCREL Technical Papers.

Lu, X. F. 2014. *Computational Methods for Corpus Annotation and Analysis* [M]. New York: Springer.

McEnery, T. & A. Hardie. 2011. *Corpus Linguistics: Method, Theory and Practice* [M]. Cambridge: Cambridge University Press.

Romero-Trillo, Jesús. (ed.). 2014. *Yearbook of Corpus Linguistics and Pragmatics 2014: New Empirical and Theoretical Paradigms* [C]. New York: Springer.

Romero-Trillo, Jesús. (ed) 2015. *Yearbook of Corpus Linguistics and Pragmatics 2015: Current Approaches to Discourse and Translation Studies* [C]. New York: Springer.

Romero-Trillo, Jesús. (ed) 2016. *Yearbook of Corpus Linguistics and Pragmatics 2016: Global Implications for Society and Education in the Networked Age* [C]. New York: Springer.

Rose, K. R. & G. Kasper. 2001. *Pragmatics in Language Teaching* [M]. Cambridge: Cambridge University Press.

Ross, S. J. & G. Kasper. 2013. *Assessing Second Language Pragmatics* [M]. New York: Palgrave Macmillan.

Rühlemann, C. 2010. What can a corpus tell us about pragmatics? [A]. In A. O'Keeffe. & M. McCarthy (eds.). *The Routledge Handbook of Corpus Linguistics* [C]. London/New York: Routledge.

Samy, D. & A. González-Ledesma. 2008. Pragmatic annotation of discourse markers in a multilingual parallel corpus (Arabic-Spanish-English) In *Proceedings of the International Conference on Language Resources and Evaluation* [C]. Marrakech.

Searle, J. R. 1969. *Speech Acts* [M]. Cambridge: Cambridge University Press.

Searle, J. R. 1976. A classification of illocutionary speech acts [J]. *Language in Society*, 5(1): 1-23.

Schiffrin, D. 1987. *Discourse Markers* [M]. Cambridge: Cambridge University Press.

Stenström, A.-B. 1990. Lexical items peculiar to spoken discourse [A]. In J. Svartvik (ed.). *The London Corpus of Spoken English: Description and Research* [C]. Lund: Lund University Press. 137-175.

Stiles, W. B. 1992. *Describing Talk: A Taxonomy of Verbal Response Modes* [M]. Newbury Park: Sage Publications.

Taguchi, N. & J. M. Sykes. 2013. *Technology in Interlanguage Pragmatics Research and Teaching* [M]. Amsterdam/Philadelphia: John Benjamins Publishing Company.

Weisser, M. 2010. Annotating Dialogue Corpora Semi-Automatically: A Corpus-Linguistic Approach to Pragmatics [D]. (Unpublished Post-doctoral Dissertation). Bayreuth: University of Bayreuth.

Weisser, M. 2015. Speech act annotation [A]. In K. Aijmer & C. Rühlemann (eds.). *Corpus Pragmatics: A Handbook* [C]. Cambridge: Cambridge University Press. 84-113.

Wilson, D. & D. Sperber. 2004. Relevance theory [A]. In G. Ward. & L. Hord (eds.). *Handbook of Pragmatics* [C]. Oxford/Malden: Blackwell Publishers.

崔　刚、盛永梅，2000，语料库中语料的标注[J]。《清华大学学报》（哲学社会科学版）（1）：89-94。

李文中，2012，语料库标记与标注：以中国英语语料库为例[J]，《外语教学与研究》（3）：336-345。

通讯地址：710128　陕西省西安市西安外国语大学外国语言学及应用语言学研究中心 / 英文学院

MedAca医学学术英语语料库的创建*

福建医科大学外国语学院 冯 欣 吴菁菁 齐 晖
北京外国语大学中国外语与教育研究中心 许家金

提要：本文旨在介绍MedAca医学学术英语语料库（临床医学专业分库）的建库原则、建库目标和实施方案。首先，本文论述了语料取样的代表性，其中涉及医学期刊来源、所收文本年份、期刊文本体裁分布等方面的考量。其次，本文详细说明了MedAca语料库的文件命名、文本清理、元信息标记和词性赋码在内的语料库标注操作等。文章最后探讨了MedAca语料库的发布及其今后在词典编纂等医学学术英语教学与科研领域的应用思路。

关键词：MedAca；医学学术英语；临床医学；语料库

1. 引言

语料库是指按照一定语言学原则，运用抽样方法，收集自然出现的连续语言，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本库（杨惠中 2002：33）。按照所选取语料的特点可分为通用语料库、专用语料库、共时语料库、历时语料库、学习者语料库、单语语料库、平行语料库及双语（多语）语料库等不同类型（梁茂成、李文中、许家金 2010：4-5）。电子语料库的建设始于20世纪60年代，作为第一代电子语料库的代表，Brown语料库和LOB语料库的建立，极大促进了美国与英国的书面英语比较研究。

我国已建成的代表性英语语料库包括上海交通大学科技英语语料库（JDEST）、中国学习者英语语料库（CLEC）及英语专业学习者语料库（SWECCL）等，一系列应用于不同领域的学科专业语料库相继出现，如对外经贸大学的“商务英语语料库”、黑龙江大学的“商务英语语料库”以及大连海事大学的“海事英语语料库”等。它们为深入进行相关专业领域的语言研究提供了必要的基础条件。然而，国内建成的大规模医学英语语料库数量较少，特别是对外开放的医学英语语料库十分有限。相关应用研究取得了一定成效，但仍处于起步阶段，如张文青（2008）提出建立医学英语写作语料库的设想；闵楠（2011）建立了一个较为综合的医学英语语料

* 本文系教育部人文社会科学重点研究基地重大项目（项目编号：17JJD740003）子课题“大数据视野下的外语及外语学习研究”阶段性成果。

库，并提出了建库原则，该库的语料来源于医学书籍、报纸、刊物、应用文等；傅顺等（2012）探讨了教学型医学英语语料库在医学英语词汇教学中的应用并提出了建库原则；王世杰等（2012）在自建语料库中提取、整理并分析了医学英语文献中的高频词块，总结了医学文体的词数、词块比例等特征。此外，已建成的医学学术英语语料库容量普遍较小，难以全面体现医学学术英语的面貌。由于缺乏统一的规范和标准，部分医学学术英语语料库在实际建设时未充分考虑建库原则及语料标注方法，建成的语料库的代表性有限。另外，由于知识产权等方面的原因，大部分医学英语语料库的利用率较低，资源难以共享。

本文以下将介绍MedAca医学学术英语语料库（临床医学专业分库）建库流程。本语料库的英文全称为Medical English Discourse of Academia（Clinical medicine），是DEAP（Database for English for Academic Purposes）学术英语语料库的医学子库，包含1186个文本，共计5,041,631个形符（tokens）、99,765个类符（types）。

2. 建库目标

建设本语料库的目标包括两方面。其一，结合临床医学学科的特点，建成一个覆盖范围广、规模较大、时效性强、能体现医学学术英语语言特征的语料库，探索该库规范的建库原则和标注方法。其二，服务于我国医学学术英语的教学和科研，为教学和科研，如术语提取、教材编写、词典编纂等提供真实的语言数据支持。

3. 语料收集方案

首先，考虑语料的覆盖面。基于临床医学这一一级学科，收集的语料穷尽其所含子学科。其次，医学科学的发展日新月异，为了确保语料的应用和科研价值，时效性是需要考虑的另一个要素。再次，考虑书面文本语言的代表性和语料的质量，我们在数量巨大的医学期刊中甄选影响因子排名靠前的期刊。影响因子是指某一期刊的文章在特定年份或时期被引用的频率，是衡量学术期刊影响力的一个重要指标（Garfield 1955）。目前从全球范围看，期刊的影响因子已经成为期刊影响力的象征，因此，选取各子学科中影响因子排名靠前的期刊能保障语料具有较高的质量。最后，在语料收集时，按照随机抽样的原则，根据各期刊所涵盖的体裁篇数比例采集各期刊中不同体裁的论文样本数。

基于上述语料来源和取样标准，根据教育部《学位授予和人才培养学科目录》医学专业一级学科“临床医学”（学科代号1002）下设的18个二级学科（儿科学、耳鼻喉头颈外科、妇产科学、急诊医学、精神病与精神卫生学、康复学、老年医学、临床检验诊断学、麻醉学、内科学、皮肤性病学、神经病学、疼痛医学、外科学、眼科、影像医学与核医学、运动医学、肿瘤学），分别收集了2012

年至2017年（主要是2015年至2017年）来自各自领域影响因子位列前十名之内的3至5本国际学术期刊的优秀论文全文及其摘要（见表1）。实际操作过程中，根据语料库的建库目的和实际用途，考虑到人员投入等因素，我们分两阶段进行语料收集，2015年收集约为100万词次的语料，建成MedAca V1.0语料库，2017年扩展库容，收集规模约为400余万词次的语料，最终建成规模为500余万词次的MedAca V2.0医学学术英语语料库（临床医学分库）[1]。

表1. MedAca医学学术英语语料库的构成

序号	刊物名称	临床医学子学科
1	New England Journal of Medicine	内科学
2	Lancet	
3	Jama-Journal of the American Medical Association	
4	Annals of Surgery	外科学
5	American Journal of Transplantation	
6	Journal of Neurology Neurosurgery and Psychiatry	
7	Journal of the American Academy of Child Psychiatry	儿科学
8	Pediatrics	
9	JAMA Pediatrics	
10	Neurobiology of Aging	老年医学
11	Aging Research Reviews	
12	Aging Cell	
13	Frontiers in Aging Neuroscience	
14	Lancet Neurology	神经病学
15	Nature Reviews Neurology	
16	Annals of Neurology	
17	Neurology	
18	JAMA Neurology	
19	Molecular Psychiatry	精神病与精神卫生学
20	The American Journal of Psychiatry	
21	JAMA Psychiatry (Chicago,III)	
22	OBSTETRICS & GYNECOLOGY	妇产科学
23	American Journal of Obstetrics & Gynecology	
24	An International Journal of Obstetrics & Gynecology	

（待续）

1 MedAca V2.0语料库库容为504万词次，该数字不含论文References部分的正文文本。若含References部分，库容约650万词次。MedAca V2.0语料库已包含MedAca V1.0语料库的100万词次。

（续表）

序号	刊物名称	临床医学子学科
25	ClinChem Lab Med	临床检验诊断学
26	Clinical Biochemistry	
27	G Ital Dermatol Venereol	皮肤性病学
28	British Journal of Dermatology	
29	Investigative Radiology	影像医学与核医学
30	The Journal of Nuclear Medicine	
31	Radiology	
32	Neuroradiology	
33	European Journal of Radiology	
34	CA: A Cancer Journal for Clinicians	肿瘤学
35	The Lancet Oncology.	
36	Cancer Cell	
37	Neurorehabilitation and Neural Repair	康复学
38	Journal of Fluency Disorders	
39	Journal of Physiotherapy.	
40	Ophthalmology	眼科
41	JAMA Ophthalmology	
42	Progress in Retinal and Eye Research	
43	American Journal of Ophthalmology	
44	Head & Neck	耳鼻喉头颈外科
45	Clinical Otolaryngology	
46	Archives of Otolaryngology-Head & Neck Surgery	
47	Hearing Research	
48	British Journal of Anesthesia	麻醉
49	Anesthesiology	
50	Anesthesia and Analgesia	
51	Annals of Emergency Medicine	急诊医学
52	Internal and Emergency Medicine	
53	Academic Emergency Medicine	
54	Clinical Journal of Pain	疼痛医学
55	Pain Medicine	
56	Regional Anesthesia and Pain Medicine	
57	Exercise and Sport Sciences Reviews	运动医学
58	Medicine and Science in Sports and Exercise	
59	Sports Medicine	

4. 文本命名

采用“子学科名-数量-文体”的顺序对所采集的文件进行分类命名。其中，子学科名使用三个英文字母缩写的形式（见表2）。文本体裁根据医学论文的种类主要分为以下三种：1）研究论文（research paper，缩写为RS）：医学论文中最具代表性的文章，实验研究、临床研究等均属该类，是报道本领域研究成果与实践经验的学术性论文。2）实验报告（report paper，缩写为RP）：包含病例报告、会议纪要、消息动态、汇报论文等。病例报告一般是介绍少量而典型的病例诊治经验；会议纪要类是医学期刊一种常见的报道形式，主要交代会议的基本情况；消息动态常见的内容主要有国内外学术动态、科研简讯、医学新闻、时讯等；汇报论文包含个案汇报，主要围绕某专题或某学科进行系统汇报，介绍医学发展新动向。3）综述动态类文章（review article，缩写为RV）：包含研究综述评论类、述评等，主要反映某一领域或某一专题研究进展或动态的文章。针对国内外某医学研究领域、某学科存在的热点和难点或正在进行的研究专题进行较为广泛而深入的阐述和精辟的评论，并做出初步的评论和建议。例如，文件名“NEU206RS”指的是神经病学这一子学科的第2本期刊所收录的第6篇文章，该文文体为研究论文。

表2 子学科名称表

子学科名	内科学	外科学	儿科学	老年医学	神经病学	精神病与精神卫生学	妇产科学	临床检验诊断学	皮肤性病学
命名	INT	SUR	PED	AGE	NEU	PSY	GYN	LAB	DER
子学科名	影像医学与核医学	肿瘤学	康复学	眼科	耳鼻喉头颈外科	麻醉	急诊医学	疼痛医学	运动医学
命名	IMG	CAN	REH	OPH	ENT	ANS	EMG	PAN	SPO

5. 文本清理

文本整理是语料库建库的关键环节，涉及文本备份、文本提炼、语料元信息标注等。通过网络下载、扫描识别等方法获得的语料可能存在乱码、格式错误等问题，因此必须整理文本，并备份原始文件。收集的大多数医学论文文本为PDF格式文件，使用ABBYY FineReader、Adobe Acrobat Pro等软件将PDF格式文件转换为TXT格式文件，对转换后的文本进行两轮人工核对，清理转换过程中出现的乱码、格式和拼写错误等问题。此外，删除文本中无法体现语言特征的数据，如图表、图表说明等。

6. 语料库标注

6.1 元信息标注

元信息标注可以为语料库检索和分析提供查询条件和依据，主要涉及对文本的语步和内部结构信息的标注，标注的方法为在文中每个语步或内部结构的开始和结尾进行标注，格式统一。医学学术文本是一类较特殊的文章体裁，具有比较突出的体裁特征。研究论文类文章一般包含标题、摘要、关键词、作者及个人信息、前言、材料和方法、结果和讨论、参考文献以及脚注等。而综述动态类文章除了包含标题、摘要、关键词、作者及个人信息之外，一般还包含前言，文献综述以及结论等。语步和内部结构的标注有助于进一步开展与体裁相关的研究，探索学术论文内部不同语步的语言使用情况。下例为对“NEU206RS”这篇文章标题的标注：

<Title>MRI biomarker assessment of neuromuscular disease progression: a prospective observational cohort study</Title>

6.2 词性赋码

MedAca语料库利用了CLAWS自动赋码软件，对语料进行了词性赋码。该词性赋码器选用的是包含有137个赋码的CLAWS软件。Jurafsky & Martin（2000）认为该赋码器对英语本族语者书面语进行自动标注的赋码准确率达到96%-97%。词性赋码为进一步挖掘语料的词汇、语法、搭配、类联接、句法等语言特征提供了保障。

7. 基于MedAca语料库研究和教学应用

建成后的MedAca V2.0医学学术英语语料库（临床医学专业分库）可实现生成词表，计算主题词等多项功能，将发布在语料云网站或CQPweb网站，供医学英语学习者、语言教师以及语言和医学科研人员通过单词、短语或使用正则表达式对词性赋码、类连接等形式进行在线检索，有效服务于医学学术英语教学与研究。

首先，已经开展了一系列的研究。利用语料库语言学中的主题词方法（keyword analysis）筛选相对于通用英语显著多用的词汇，并通过关键主题词方法，制作了医学学术英语词汇表，重点收录介于准术语和通用词汇之间的医学词汇（许家金 2017）。此外，将各子学科的文本制作成子语料库，并利用相同的方法制作了分专业专属词汇表，利用BFSU PowerConc软件，统计制作医学学术英语常用短语表。

其次，计划进行医学学术英语学习词典编纂研究。词典的编纂将基于体裁短语学视角（许家金 2017）。在宏观结构方面，甄选词目，确定词典规模；微观结构方面，利用扩展意义单位的方法，同时考虑临床医学各子学科中医学英语的词汇倾向和语步分布等信息，对词汇和短语进行整句释义并列出例句和常见搭配，并配以图片进行生动直观的演示。考虑到目标读者的英语水平与实际需求，对例句中特别晦涩难懂的词语或插入语等进行适当删减，使所呈现的例句更符合学习者的认知规律。

8. 结语

随着语料库语言学与学术英语的联系日益紧密，亟须建立一批学术英语语料库，开放性MedAca语料库（临床医学专业分库）即应势而建。团队遵循真实性、系统性以及便捷性的原则，分类别、分步骤逐步累积临床医学学科中不同学科领域的高水平论文，严格按照规范对收集文本进行整理、标注、赋码，从语料规模、范围、质量等方面确保该库的代表性及可信度，为教学科研提供具有更高代表性的真实语言材料。

参考文献

Garfield, E. 1955. Citation indexes for science; A new dimension in documentation through association of ideas [J]. *Science* 122(3159): 108-111.

Jurafsky, D. & J. Martin. 2000. *Speech and Language Processing: An Introduction to Natural Language Processing, Computational Linguistics, and Speech Recognition* [M]. New Jersey: Prentice Hall.

傅 顺、罗永胜、杨劲松，2013，教学型医学英语语料库的应用研究[J],《广东医学院学报》(1)：101-103。

梁茂成、李文中、许家金，2010,《语料库应用教程》[M]。北京：外语教学与研究出版社。

闵 楠，2011，关于我校医学英语语料库的建库原则[J],《首都医科大学学报》(社会科学版)：206-207。

王世杰、武永胜、赵玉华，2012，基于语料库的医学英语词块研究及其教学[J],《甘肃中医学院学报》(4)：77-82。

许家金，2017，体裁短语学视角下的医学学术英语词典研编[J],《外语与外语教学》(6)。

杨惠中，2002,《语料库语言学导论》[M]。上海：上海外语教育出版社。

张文青，2008，医学英语写作语料库建设的构想[J],《吉林工程技术师范学院学报》(7)：61-63。

通信地址： 350108 福建省福州市 福建医科大学外国语学院（冯欣、吴菁菁、齐晖）

100089 北京市 北京外国语大学中国外语与教育研究中心（许家金）

English Abstracts

A study of Chinese-English co-selection pattern equivalence through the lens of phraseology

The study takes as a starting point the lexical translation pairs extracted from English-Chinese/ Chinese-English SJTU Parallel Corpus and moves on to examine these pairs in comparable corpora, with their semantic and pragmatic profiles found and co-selection patterns identified. The co-selection patterns are compared in terms of semantic prosody, semantic preference and collocation to explore their respective roles in establishing cross-language equivalence. Results show that cross-language equivalence resides in co-selection patterns, in which the divergence from the prosodic norm leads to a new unit of meaning. This means that it is the altered pattern that generates a meaning-shift unit, which requires new pattern equivalence in the target language. The study also finds that inherent semantic preference is vital in achieving semantic correspondence, while optional semantic preference is not decisive in setting up pattern equivalence. Non-correspondence in terms of specific semantic features of collocates only produces infrequent or inappropriate uses, without affecting co-selection pattern equivalence.

Corpus-based study on equative comparative sentence in Chinese interlanguage

Equative Comparative Sentence is an important part of Comparison Category in Chinese, and it is also one of the difficulties for learning Chinese as a foreign language. The paper attempted to collect all the equative comparative sentences in HSK Dynamic Composition Corpus which is a corpus for Chinese Interlanguage. It described the overall condition of equative comparative sentences in Chinese interlanguage by comparing with the ones in the Corpus of Chinese Compositions. The study revealed that: (1) the comparative results in Chinese Interlanguage are mostly single adjectives and very complicated structures, while the comparative results in Modern Chinese tend to be a continuum of difficulty of grammatical structures, which, we assume, is deeply influenced by the learners' Chinese language proficiency and modularized cognitive processing; (2) the sentences which use the comparative marks with higher degree of grammaticalization, such as "跟gen", "与yu", "和he", "同tong" tend to express the equative meanings, while the sentences which use the comparative marks with lower degree of grammaticalization, like "像xiang" tend to express the figurative meanings. From the perspective

of errors, we found that the largest number of errors are the missing of the first comparative mark and misidentification of equative comparison and comparative comparison. Moreover, this study provided that, Interlanguage, as an independent language system, might develop itself by breaking through the traditional error analysis and we probably could find more language rules in studying both the wrong sentences and the right sentences.

A study of the co-selection features of *the* in high-school EFL writing

The present study aims to examine Chinese high-school EFL learners' use of definite article THE with reference to that used by native students in English writing. The data shows that, compared with native students, Chinese learners primarily use THE in pre-modification structures, particularly in THE+N, and have a strong tendency to misuse the meaning and function. Further analysis suggests that, in native students' English writing, THE tends to co-occur with specific grammatical structures to form certain colligations, expressing particular meanings and functions, which mirrors the co-selection of form, meaning and function. Despite their knowledge of the co-occurrence of THE and corresponding grammatical categories, Chinese high-school learners fail to gain adequate knowledge of their meanings or/and functions. This phenomenon is closely associated with their frequent exposure to THE+N/NP structure, the absence of congruent articles in L1 Chinese, and the practice of explicit grammar teaching. The above findings have some theoretical implications for further research on English articles and some pedagogical implications for improving grammar teaching for beginner learners of English.

Anticipatory and anaphoric features in *it*-chunks in thesis abstracts by postgraduates majoring in science and technology

This paper explores high-frequency *it*-subject lexical chunks in the corpus of English abstracts from theses written by Chinese postgraduates majoring in science and technology, investigating phraseological features of both anticipatory-*it* and anaphoric-*it* across Chinese and native postgraduates. The analysis shows that (1) Chinese postgraduates majoring in science and technology make use of a great number of *it*-chunks to realize the function of textual organization and attitudinal expression, (2) the main problems of anticipatory-*it* occur at the co-selectional level, represented by lexis-structure mismatch and influenced by pattern overgeneralization, vocabulary limitation, use of persuasive strategy, and mother tongue, and (3) the main interlanguage feature of anaphoric-*it* is the overuse of *it* and attitude-laden lexical devices,

due partly to learners' violation of basic anaphoric rules and partly to the writing strategy of postgraduate thesis.

A study of chunks in spoken academic English under the framework of Linear Unit Grammar

"Linear unit grammar: Integrating speech and writing" (Sinclair&Mauranen，2006) is claimed to get rid of the shackles of traditional grammatical categories, break the limit of traditional grammar research units - sentences and single words, and concern about the lexical chunks formed by the natural combination of words and discourse composed of chunks. Withinthis framework, the present study usesspeech consistency as the standard to extract chunks, that is, the pause information is used as the boundary mark to divide the texts in the academic spoken corpus into chunks.The recurrent patterns and the length of the chunks are then analyzed. This paper concludes with a discussion of the automatic segmentation of texts.

Pragmatic annotations in corpus construction

Both researchers and teachers pay more and more attention to corpus due to its data richness and authenticity. Pragmatic annotation, as an approach to keeping language authenticity, has been one of the heated topics in corpus linguistics. This paper, on the basis of literature review, has explored its necessity, its annotated objects, and its challenges as well as its potential research foci.

The construction of the MedAca EAP corpus of clinical medicine

This paper describes the principles, objectives and data processing techniques adopted in the building of the MedAca EAP corpus of clinical medicine. First of all, the representativeness of sampling is explained, which takes account of the sources of medical journals, publication year, and the genre types of the research articles. Secondly, procedures such as the naming of the texts, text clean-up, metadata mark-up and part-of-speech tagging are elaborated. Lastly, the release plan and the pedagogical implications, such medical English dictionary compilation, are discussed.

图书在版编目（CIP）数据

语料库语言学．2017．2 ：汉、英 ／ 梁茂成主编．-- 北京 ：外语教学与研究出版社，2018．7
ISBN 978-7-5213-0331-5

Ⅰ．①语… Ⅱ．①梁… Ⅲ．①语料库－语言学－汉、英 Ⅳ．①H0

中国版本图书馆 CIP 数据核字（2018）第 182366 号

出 版 人　徐建忠
责任编辑　毕　争
责任校对　解碧琰
封面设计　覃一彪　锋尚设计
出版发行　外语教学与研究出版社
社　　址　北京市西三环北路 19 号（100089）
网　　址　http://www.fltrp.com
印　　刷　北京九州迅驰传媒文化有限公司
开　　本　787×1092　1/16
印　　张　7.5
版　　次　2018 年 7 月第 1 版　2018 年 7 月第 1 次印刷
书　　号　ISBN 978-7-5213-0331-5
定　　价　12.00 元

购书咨询：（010）88819926　电子邮箱：club@fltrp.com
外研书店：https://waiyants.tmall.com
凡印刷、装订质量问题，请联系我社印制部
联系电话：（010）61207896　电子邮箱：zhijian@fltrp.com
凡侵权、盗版书籍线索，请联系我社法律事务部
举报电话：（010）88817519　电子邮箱：banquan@fltrp.com
法律顾问：立方律师事务所　刘旭东律师
　　中咨律师事务所　殷　斌律师
物料号：303310001

语料库语言学

CORPUS LINGUISTICS

要　目

责任编辑：毕　争
责任校对：解碧琰
封面设计：覃一彪　锋尚设计

高等英语教育出版分社宗旨：
推动科研·服务教学·坚持创新
外研社·高等英语教育出版分社
FLTRP Higher English Education Publishing
电话：010-88819595
传真：010-88819400
E-mail：ced@fltrp.com
网址：http://heep.unipus.cn

heep 微信公众号　iResearch 微信公众号

记载人类文明
沟通世界文化
www.fltrp.com

ISBN 978-7-5213-0331-5

定价：12.00元