《中国学术期刊网络出版总库》及CNKI系列数据库入选期刊

二〇一六 第三卷 第一期

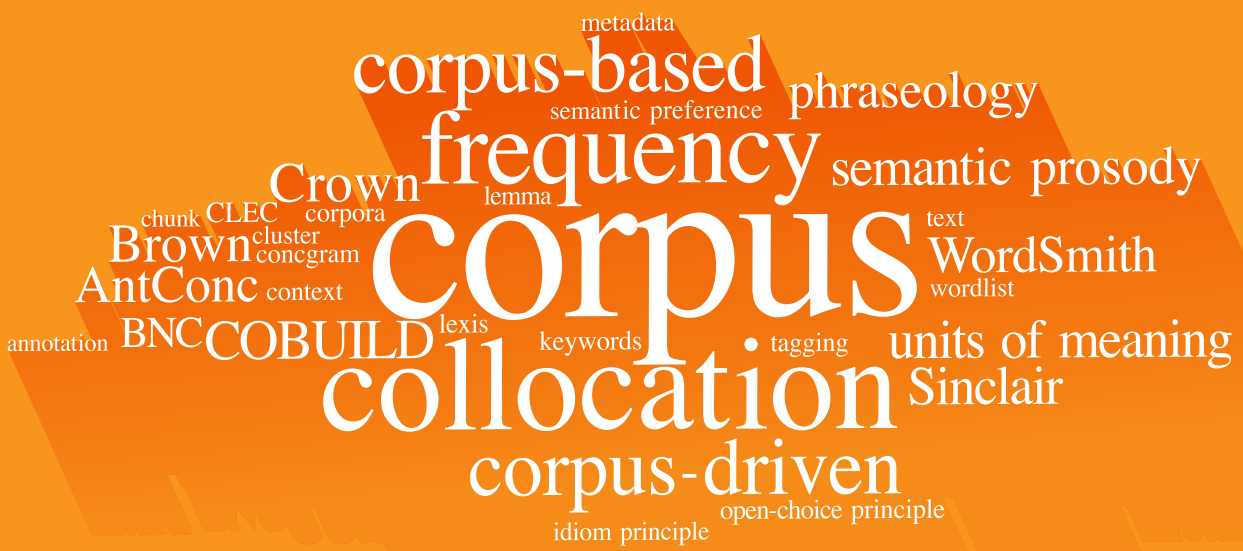
语料库语言学

CORPUS LINGUISTICS

1
Vol. 3 No.1
第3卷 第1期
2016

北京外国语大学中国外语教育研究中心
梁茂成 许家金 主编

语料库语言学



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

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：中国外语教育研究中心
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
Published by Foreign Language Teaching and Research Press

主　编：梁茂成、许家金
编　校：华　雨、徐秀玲

Editors: Liang Maocheng and Xu Jiajin
Proofreaders: Hua Yu and Xu Xiuling

电　　话：（010）88816828
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教育研究中心
《语料库语言学》编辑部（100089）

《语料库语言学》

2016年　第3卷　第1期

目　录

研究论文

研制开发

书刊评介

CORPUS LINGUISTICS

Volume 3, Number 1, 2016

Table of Contents

从推理类话语标记的演化看翻译与现代汉语的互动*

曲阜师范大学 秦洪武 刘丹丹 杜肖颖

提要：本文以汉语推理类话语标记为例，基于汉语历时类比语料，考察翻译在汉语话语标记历时变化中所起的作用及其作用方式。研究发现，较文言文和旧白话文，现代汉语使用话语标记的频率更高，组织手段也更多样，这一变化与翻译，尤其是20世纪初的白话文翻译部分相关。翻译影响汉语话语标记使用的方式多样，但总的说来汉语总是有选择地接受翻译语言的影响，汉语语言手段被翻译调用、发挥并最终影响汉语的发展。

关键词：推理标记、类比语料、现代汉语、翻译语言

1. 引言

我们组织言语表达时通常要明示言语的进程（开始、结束或者过渡）、话题、上下文关系、个人态度等等，这时需要调用各种手段，而最常使用的是语言手段，如词、短语、小句等，这类语言手段通常称作话语标记（discourse markers）。话语标记具有程序意义，有助于形成语篇的连贯性与条理性，并起到一定的指示与提示作用。

根据秦洪武、王玉（2014），现代汉语的话语标记在二十世纪发生了明显的变化，甚至有一些古代汉语和旧白话里没有使用过的话语标记形式频繁出现在现代汉语里，而且这种变化和同一时期汉语翻译语言的变化同时发生。这一变化对于现代汉语交际功能的完善有重要意义，但长期以来，学界对它关注很少，更缺少系统的分析。本文基于历时对比语料库，通过考察汉语推理类话语标记，分析话语标记语在汉语中使用的变化过程，并探讨翻译在其中所起的作用。

2. 推理类话语标记

Fraser（1999）将话语标记分为两类：关联语段信息的话语标记和关联话题的话语标记。有些关联语段的话语标记归纳总结上文信息，推测原因或者引出结论，

* 本研究为国家社科基金重大项目“大规模英汉平行语料库的建立与加工”（10&ZD127）的阶段性成果。

如“因此”、“综上所述”、“显然”等，我们称之为推理类话语标记（inferential discourse markers，IDM）。在英语里，这类标记可以使用词（therefore、hence）、词组（in conclusion、in total）或者小句（to sum up、whatever happens）。

现代汉语频繁使用推理类话语标记，这种篇章组织手段的使用频率和组织方式与英语很相似。这一变化既是语言自身发展的结果，又有诸多外部力量推动，翻译可能是其中的一个重要因素，这也是本文探讨的问题。

3. 翻译语言与目标语言的互动：编码复制理论

现代汉语早期发展包含两个重要阶段：晚清白话文运动和五四时期的白话文学。在五四白话文学时期，汉语变化最为明显，大批支持新文化运动的学者提倡通过翻译“创造出许多新的字眼，新的句法”以改造汉语。傅斯年（1918）欧化主张最为明确、激进，他提出改造汉语的三条途径：（1）读西洋文学和从西洋文译过来的文本；（2）自己直译获得新的表达方式；（3）坚持欧化，做文章时，运用读、译西文所得仿造西文。

怎么仿造呢？在汉语里直接植入西洋语法（如屈折变化）会导致系统混乱，全部移植就更不可能。可行的做法是按汉语的语言规范局部移入。Johanson提出的“编码复制理论”就涉及局部引入问题。该理论认为，翻译作为一种间接语言接触，会引发复制行为（copying），即基本码（basic code，即目标语）使用者复制模型码（model code，即源语）中的成分（Johanson 2008：62）。汉语欧化就是模型码欧洲语言（主要是英语[1]）的表达形式通过翻译被复制进基本码汉语（秦洪武、王玉 2014）。但这种复制不是简单的拷贝，也不是更替原有代码，而是将源语中的代码成分嵌入目标语代码，是一个顺应过程，形成两种代码间的互动（Johanson 2008：62）。从这个角度说，模仿是有选择的，而模仿中语言性质的变化是语法复制，这一变化也就是Johanson（2008：62）所说的“选择性语法复制”（selective grammatical coping），包含词形复制（material copying）、语义复制（semantic copying）、搭配复制（combinational copying）和频率复制（frequential copying）（董元兴、赵秋荣 2012）。从理论上说，话语组织方式上的复制也应如此（秦洪武 2010）。

翻译一般只是复制源语中的一个或多个特征，而且，新引入的表达形式都要经过严格筛选以适应汉语自身的规律，这样的互动过程才会最终引发形态-句法变化，或者目标语话语组织方式上的变化（秦洪武、王克非 2009；秦洪武 2010）。本文以推理类话语标记为例，从编码复制角度探讨翻译和汉语话语标记演化之间的关系。

4. 研究方法

4.1　类比语料数据

五四运动以来，推理标记的使用伴随着该形式在翻译语言中的使用而发生变化，那么，我们就可以通过类比分析的方法，比较汉语原创文本和汉语翻译文本中的话语标记成分，找出那些受翻译语言影响的话语标记使用现象。为此，本研究使用了现代汉语历时类比语料库，其构成如下：

表 1.　现代汉语历时类比语料库（使用 AntConc 统计）

类比语料库	子库名称	库容（词）	语料构成
汉语原创语料库	1911 年前汉语原创子库	1,229,168	文学：1,136,056（92%）；新闻报刊：93112（8%）。晚清白话文运动：主体为白话小说，部分白话报刊，无科技文献。
	1919-1930 年代汉语原创子库	1,236,273	文学：1,029,263（83%）；非文学：207,010（17%）。白话文用于非文学明显少于文学，文学文体丰富。
	1990 年代 - 当代汉语原创（LCMC）	834,007	改革开放后：语料构成平衡。
汉语翻译语料库	1919-1930 年代英汉翻译子库	2,713,469	文学：1,193,695（44%）；非文学：1,519,774（56%）。语料构成较为平衡。
	1980 年代 - 英汉翻译子库	2,123,097	文学：1,006,694（47%）；非文学：1,116,403（53%）。文学、非文学语料构成较为平衡。
类比库库容合计		8,136,014	

表 1 所列的汉语翻译和汉语原创类比语料都是历时的，时间横跨一个世纪；每个阶段的间隔至少为 20 年，即一个语言代际。另外，为反映话语标记使用的实际特点，本项研究使用的语料包含非文学文本，文本类型多样。

4.2　数据提取

在这些语料库里，我们发现话语标记一般出现在句子起始位置，或位于句中某个小句的开头。这一使用特征为我们提取数据提供了方便，我们据此使用定位检

索，只提取由1-4词构成且独立使用的句首成分（检索时附加标点‘,’），即使用正则表达式提取语块[2]，然后进行人工识别和分类，最终筛选出推理类话语标记。

基于提取的数据，我们先对比白话文初期和当代汉语，找到当代汉语不同于早期白话的语言特征，接下来观察这些特征在各个时段的表现，以此考察翻译在这一历时变化中所起的作用。

5. 发现和讨论

5.1 汉语推理类话语标记使用频率的历时变化

为便于观察，我们先将数据标准化，按10万词为单位计算各标记的使用频率，标准化的数据见表2。

表2. 句首话语标记使用频率的历时变化（按句首独立的1–4词检索）

句首话语标记	汉语原创			汉语翻译	
	1911年前汉语原创	1919-1930年代汉语原创	1990年代-当代汉语原创（LCMC）	1919-1930年代英汉翻译	1980年后英汉翻译
推理标记标准化频率（十万词）	27.9	29.3	114	77.6	67.5
汉语句首话语标记总频率（十万词）	65.55	84.6	322.72	190.51	255.18

表2显示，在20世纪，汉语句首话语标记的使用频率逐渐增高，汉语翻译文本中相应的句首标记语也呈现相同的走势，只是1930年代以后的汉语翻译文本中的推理标记使用频率略有下降，这主要是受30年代以后“文白论战”中“白话文反思运动”[3]的影响。另外，1980年后的汉语原创中的推理类标记在频率上远高于同时期的汉语翻译文本，这与文本的性质有关。1980年后英汉翻译子库中多为文学文本，而汉语推理类话语标记多用于论证性的非文学文本。

由于各个子库之间库容差异很大，绝对频次不能直接表现话语标记使用频率在各个时期的差异，为此，我们使用似然率分析观察话语标记使用的历时差异的显著性。见表3、表4。

表3.　五四前后白话文中推理类话语标记使用差异的似然率分析

	1911年前 汉语原创（频次）	1919-1930年代汉语 原创（频次）	对数似然率	显著性
推理类标记	343	337	24.48	0.000

表4.　旧白话和当代汉语中推理类话语标记使用差异的似然率分析

	1911年前 汉语原创（频次）	1990年代-当代汉语原创 （LCMC）（频次）	对数似然率	显著性
推理类标记	343	950	44.54	0.000

与1911年前的白话文相比，现代汉语中推理标记的历时使用频率逐步增高，这一历时变化和汉语翻译语言中话语标记的变化趋向基本重合，这提示我们需要通过类比来分析汉语翻译语言对汉语发展的影响。

5.2　汉语原创推理类话语标记的历时变化

一般来说，汉语话语标记的变化有三种情况：一是话语标记的使用频率发生变化；二是出现功能相对稳定和独立的话语标记；三是出现全新的话语标记。这三种情形是否适用于推理标记尚不可知。我们就以1911前汉语原创子库为参照，观察这类标记的使用特征。

表5.　汉语原创推理类话语标记的历时变化

句首推理类话语标记	1911年前汉语原创	1919-1930年代汉语原创	1990年代-当代汉语原创（LCMC）
类符	34	32	56
形符	343	337	950

从表5和附录中看出，汉语原创推理类话语标记的使用总体看来变化明显，类符和形符均出现明显变化。这种变化在九十年代以后尤为显著。值得注意的是，1919-1930年代汉语原创文本中推理类话语标记的种类和频率有些微减少，这主要是受到了“白话文反思运动”的影响。另外，五四以来受欧化影响比较重的是汉语的书面语而非口语，欧化文体的使用有语体限制，一般都限于书面语，较少进

入口语。而本研究所用1930年代以后的汉语原创语料开始出现一定比例的非文学文本，口语体比例相对减小，受欧化影响的推理标记的使用就相应减少。

表6. 推理类话语标记的历时变化（详见附录）

1911年前汉语原创		1919-1930年代汉语原创		1990年代-当代汉语原创（LCMC）	
总（而言）之	27	总（而言）之	56	总的讲/总的说/总起来看/总前所述/综上所述/	46
因此	29	因此/因之	32	因此/因而/因为	295
所以	11	所以	12	所以	77
于是	1	于是	13	于是	106

表7. 推理类句首话语标记的历时变化（例示，详见附录）

1911年前汉语原创	1919-1930年代汉语原创	1990年代-当代汉语原创（LCMC）
就这样/要能这样/由这样看来/这样、这等……、这么说、这一来	既然这样/若不是这样/照这样看来/照这样做/这样……/这样一来、看这情形、从这里看来	就这样/再这样下去/这样（……）、这么说、这还不说、这是因为、这一来、如此这般
如此/若果如此/事已如此/虽然如此/原来如此、为此、因此、由此看来、据此	既然如此/如此（……）来、继此、如是、因此	如此这般/虽然如此/长期如此、故此、如是、为此、因此、由此看来
总（而言）之	总（而言）之、笼统言之、	综上所述、总的讲、总的说、总起来看、总前所述、总之

表6、表7和附录中的例证说明，推理类话语标记在1919年以后才大量出现，这类标记在1919年以前要么不用或者罕用，要么不能作为独立的句首成分使用。这些话语标记在1919年后经历了从出现到频繁使用的变化过程。另外，推理类话语标记在历时变化中也新出现了很多变式。因此，我们可以将推理类话语标记的历时变化特点总结为三方面：1. 形式更加丰富；2. 组合方式更加多样；3. 使用频率变高。这些剧烈的变化显然不能简单归结为语言自身演化的结果，应该与外部因素如翻译有关。

5.3　翻译与推理标记的语用化进程

表8显示，现代汉语（总体）中推理标记的使用特征和汉语翻译语言基本重合，汉语翻译语言中出现的推理类话语标记在源语英语中有对应的表达形式（如表9所示），且其句法位置和语用功能相同，说明这一变化包含代码复制过程，我们可以据此假定这和翻译有关。

表8.　推理标记在汉语原创和翻译子库中的使用

	汉语原创	1919-1930年代翻译	当代翻译
推理标记	这样（……）、总的……、一句话、显然（……）、简单地说、……归纳、很……	这样（……）、显然（……）、很……	这样（……）、总的……、显然、很……

表9.　推理标记的模型码英语对应词

汉语中的标记	主要英语对应词	其他英语对应词
显然	obvious(ly)、evident(ly)	clear(ly)、apparent(ly)
这样	so、thus、so that	so that

代码复制到复制品成为话语标记可能需要经历一个从概念意义到程序意义的语用化过程。但需要指出的是，代码复制本身无法复制源语的语用化过程，但能加快这一过程。比如，“显然”本来有实在意义，指显著、显扬、显赫，一般用于描述事物容易观察和理解的性质，如：“事迹显然，无可臻惑”；一般不会作用于一个命题信息，也就不会有置于句首充当话语标记的用法。语料检索显示，1911年前的汉语原创和CCL语料库古代部分里就只有词汇意义，没有语用标记这类用法。如：

（1）a. 省疏，并见周氏遗迹，真言显然，符验前诰，二三明白……（概念意义）

b. 就终不回，私与恭疏曰：“大人率厉炖煌，忠义显然，岂以就在困危之中而替之哉？”（概念意义）

但在1919-1930年代汉语原创中已有12例开始出现程序意义，作用于命题内容。如：

（2）a. 爱情很快被销蚀了——这显然不是使那产业车轮运转着的原动力。（概念意义+程序意义）

b. 两个肩头很有力，显然是做惯了苦工的缘故。（概念意义+程序意义）

可以看出，“显然”的概念意义已经减弱，开始用于标记说话者对所述信息的逻辑关系所作的推断，具备了独立用作语用标记的可能。

然而，同时期该话语标记语在翻译语言中也大量使用。在其对应的模型码（英语）中，表达式obvious(ly)、evident(ly)、clear(ly)、apparent(ly)等的典型位置也是句首且独立使用。翻译这些表达形式时，“显然”的语用潜力就得以充分挖掘，成为可独立使用的语用标记。如：

（3）a. *Obviously*, individuals who share this judgment will regard the legal mechanisms under discussion here as incomprehensible at best and perhaps perverse at worst.

显然，持有这种看法的人至多会认为这种法律机制不可思议，甚至会认为它是邪恶的。

b. *Clearly* the Secretary could not contract away his statutory authority.

显然，内政部长不能因为订立了合同就不再行使其法律上的职权。

上面的例证说明，启用目标语中现成的表达形式并不是简单的调用，而是一种改造。“显然”本有的词义改变了，由评价性表达（一般放在被评价成分后边）转变为话语标记（置于句首位置）。也正是从1919年后，“显然”的功能开始变化：它不仅能引出要陈述的内容，还可以独立使用，用于标记后面的推断性命题。鉴于我们很难找到其他可能的原因，现代汉语中句首推理标记的高频使用应是受翻译语言影响的结果，也就是间接受模型码英语影响的结果。汉语翻译语言挖掘了汉语语言中既有表达式的表达潜力，同时又加速了该表达式的语用化进程。表10的统计也可以说明这一点。

表10. “显然”标记的使用与翻译的互动

	1911年前汉语原创	1919-1930年代汉语原创	1990年代-当代汉语原创（LCMC）	1919-1930年代英汉翻译	1980年代-英汉翻译
句首	0	12	57	54	94
句首/独立	0	1	23	18	31

5.4 翻译与现代汉语的互动

从上面的描述和分析中可以看到，汉语句首话语标记的使用频率在短时间内明显增高，翻译在外部起到了重要的推动作用。但翻译语言以什么方式参与了这一历时变化过程呢？本文认为可以从以下两个方面观察。

5.4.1 翻译丰富了现代汉语话语标记的表达形式

比如，表示“总结”意义的话语标记在1919年前（CCL）白话文里只有“总（而言）之”；在1919-1930年代的白话文中，还有“笼统言之”；而在现代汉语

中它的组合方式丰富、灵活，可以检索到更多高频使用的变式，如“综上所述”、“总的讲/说”、“总起来看”、“总前所述”等，这些组合方式大都可以在汉语翻译文本中找到，其中的相互影响是显而易见的。

5.4.2　现代汉语有选择性地接受翻译语言的影响

任何语言的变化都是为了维系既有信息交流系统顺畅运行而进行局部修补或者优化，不会允许整个系统的改换。所以，汉语接受由翻译推动的语言变化时，会按自身的特点有选择地吸收或接受来自翻译语言的影响。就推理标记来说，在翻译时，基本码汉语复制模型码英语对应成分的组合方式及其语篇和句法位置。但这种复制不是将语法特征全部复制（4a中的关系代词that就无法复制），而是保留其语篇位置，但使用汉语的组合方式，如：

（4）a. It is *obvious* that this development is one which could not have taken place, had not circumstances favored the development of a caste of priests.

那是显然的，这个发展，如果情形是不适于一个僧侣阶级的发展，便不会发生出来。

b. They’re the first thing to disappear from bathrooms, *apparently*.

很显然，它们是最先从浴室里消失的东西。

这说明，代码复制受基本码本身形态句法性质的制约，并非单纯的形式拷贝。翻译语言本身也没有力量改变目标语，如果汉语中有翻译语言的某些特征，那只是汉语按照自己的需要接受了翻译语言的影响。因此，翻译语言和目标语之间是互动关系，这种互动既丰富了双语转换中对等成分的使用，又促进了目标语自身的发展。

6. 结语

翻译语言的某些特征以异乎寻常的方式进入汉语和时代有关，20世纪初的白话文运动有意识地推动翻译语言中多少带有“异味”的语言表达方式进入汉语，其中就包含话语标记语。汉语翻译语言引入话语标记语的主要方式是选择性语法复制，这种复制产生高度同构的结构，使得代码在双语间互译性更强，同时也丰富了汉语篇章组织的语言手段，总体上起到了积极的作用。

从现代汉语推理标记使用的历时变化上看，翻译中的语法复制实际上是充分挖掘汉语既有的言语表达资源，或是使用现成但不常用的语言表达形式，或是加速某些表达形式的语用化进程。这说明，汉语接受外来语言的影响有一个前提：一切改变均是为了维系汉语既有语言系统运行顺畅。这意味着，汉语翻译语言也受到汉语本身的约束，与之相关的语法复制只能是局部的，不可能涉及语言系统的改变。

注释

1. 根据Kubler（1985：25），在20世纪初大量译入中国的作品中，译自英国和美国的作品占到全部翻译的62%，来自英语的最多，也最具影响。

2. 如下面的正则表达式：‘<s>(\b[^\x00-\xff]+/[a-z]+\b\s){1,4},’（检索实例：‘<s> “/w 依/v 我/r 想/v ,’）；‘<s>(\b[^\x00-\xff]+/[a-z]+\b\s){1,4},’（检索实例：‘<s>很/d 明显/a ，/w’）

3. 白话文反思运动：1930年代以后，林纾、章士钊、陈寅恪、钱穆等学者对“白话文运动”提出了反对和质疑。继而出现了白话文反思阶段，反对过度欧化。

参考文献

Baumgarten, N. & D. Özçetin. 2008. Linguistic variation through language contact in translation [A]. In P. Siemund & N. Kintana (eds.). *Language Contact and Contact Languages* [C]. Amsterdam: John Benjamins. 293-316.

Fraser, B. 1998. Contrastive discourse markers in English [A]. In A. Jucker & Y. Ziv (eds.). *Discourse Markers: Descriptions and Theory* [C]. Amsterdam: John Benjamins. 301-326.

Fraser, B. 1999. What are discourse markers? [J]. *Journal of Pragmatics* 31(7): 931-952.

Johanson, L. 2008. Remodeling grammar: Copying, conventionalization, grammaticalization [A]. In P. Siemund & N. Kintana (eds.). *Language Contact and Contact Languages* [C]. Amsterdam: John Benjamins. 61-81.

Kubler, C. 1985. *The Development of Mandarin in Taiwan: A Case Study of Language Contact* [M]. Taipei: Student Publishing.

Urgelles-Coll, M. 2010. *The Syntax and Semantics of Discourse Markers: Continuum Studies in Theoretical Linguistics* [M]. London: Continuum.

董元兴、赵秋荣，2012，编码复制框架视角下翻译对现代汉语发展变化的影响——以被动语态为例 [J],《中国地质大学学报（社会科学版）》(3)：129-133。

傅斯年，1918，怎样做白话文 [J],《新潮》(2)：171-184。

贺　阳，2008,《现代汉语欧化语法现象研究》[M]。北京：商务印书馆。

李秀明，2011,《汉语元话语标记语研究》[M]。北京：中国社会科学出版社。

秦洪武、王克非，2009，基于对应语料库的英译汉语言特征分析 [J],《外语教学与研究》(2)：131-136。

秦洪武、王　玉，2014，从详述类话语标记看翻译与现代汉语话语组织的发展 [J],《外语教学与研究》(4)：521-530。

秦洪武,2010，英译汉翻译语言的结构容量：基于多译本语料库的研究 [J],《外国语》(4)：73-80。

冉永平，2003，话语标记语well的语用功能 [J],《外国语》(3)：58-64。

谢世坚，2009，话语标记语研究综述 [J],《山东外语教学》(5)：15-21。

赵秋荣、王克非，2013，英译汉翻译语言的阶段性特点——基于历时类比语料库的考察 [J],《中国翻译》(3)：15-19。

朱一凡,2011,《翻译与现代汉语的变迁（1905-1936）》[M]。北京：外语教学与研究出版社。

附录：

推理类话语标记使用的历时变化

1911年前汉语原创		1919-1930年代汉语原创		LCMC当代汉语原创	
既……	97	不管……	2	不管……	15
就这样	2	不论如何	2	而且	30
据此……	6	不然(……)	32	概而言之	1
那么说	3	从这里看来	2	故此	1
那么着	6	大概言之	2	归根到底	1
如此……	15	既然 如此/这样	5	归纳……	3
如果……	2	继此	1	果然	8
如今看来	2	继上说来	1	很……	9
如若……	7	简单的一句	1	既然……	4
若	9	简单地说	1	假如	4
若果如此	4	结果	4	结果	10
若是不然	17	看这情形	1	究其原因	1
事已如此	1	笼统言之	1	就这样	20
虽然如此	3	那并不	1	举例来说	1
所以	11	那末	44	看来/得出	17
倘/倘若……	13	那么	84	看样子	1
为此	2	如此（……）来	5	可见	15
无论如何	16	……不然	3	明显得很	1
要能这样	1	若不是这样	1	那么	66
要不然	5	如是	3	那末	5
因此	29	所以	12	然则	1
由现在/这样看起来	4	无论……	19	如是	7
由此看来	1	以上……	4	如此这般	2
由此可知	2	因之	1	如果(……)	12
于是	1	因此	31	……表明	11
原来如此	7	应当说	1	说到底	1
再怎么说	1	于是	13	虽/虽然（如此）	2
照……	12	照这样看来	1	所以	77

（待续）

（续表）

1911年前汉语原创		1919-1930年代汉语原创		LCMC当代汉语原创	
这等……	15	照这样做	1	推而广之	1
这么说	3	这样(……)	47	为此	45
这样	4	这样一来	5	无论怎样说	1
这样……	24	总(而言)之	52	显然	20
这一来	2			一句话	6
总(而言)之	27			因此	266
				因而	15
				因为	14
				由此	7
				由此看来	17
				于是	106
				再进一步	1
				再这样下去	1
				长期如此	1
				这还不说	1
				这是因为	6
				这么说	4
				这一来	1
				针对……	21
				正因为……	9
				正是我国春秋时代	1
				综上所述	4
				总的讲	1
				总的说	1
				总起来看	1
				总前所述	1
				总之	38
				这样(……)	61

通讯地址：273165　山东省曲阜市曲阜师范大学外国语学院

语料库驱动的机器词典构建关键问题探讨

解放军外国语学院 曹 蓉
浙江工商大学 濮建忠
解放军外国语学院 黄金柱

提要：语料库驱动的语言研究试图最大限度地摆脱已有理论束缚，力求发现新的反映语言使用本质的事实。在这一理念指引下，我们重新审视和探讨了机器词典构建的几大关键问题（如：语言描述的核心、基本单位、释义模式等），并在此基础上提出一种新的机器词典构建理念，即：以意义为描述核心、以语言使用（或文本）为获取意义的本源、以集词汇、语法、语义、语用于一体的“扩展意义单位”为基本描述单位，采用列举与归纳相结合的释义架构和正则表达式的表示方法。

关键词：机器词典、语料库、扩展意义单位

语料库自诞生之初即与词典构建结下不解之缘。早在1898年，德国学者Kaeding就通过统计单词在大量文本语料中的出现频率编写了《德语频率词典》，这被认为是最早的语料库及语料库在词典编纂上的应用。尽管Kaeding所使用的语料并非机器可读，然而这种从大量真实文本语料出发构建词典的理念可以说是开创性的。1959年，英国伦敦大学的Randolph Quirk发起了“英语用法调查”（SEU）语料库项目，并利用该语料库编写出版了《现代英语语法》。1980年，由英国伯明翰大学的John Sinclair负责的“柯林斯伯明翰大学国际语言资料库”（COBUILD）项目启动，并在此基础上先后出版了多部词典、语法书和用法指南等。2008年，欧洲辞书学会创始人Sue Atkins和国际著名的语料库词典学家Michael Rundell合作撰写了《牛津应用词典学指南》，对基于语料库的单语和双语词典编纂过程进行了详细介绍。在我国，诸如《常用汉字登记表》、《普通话三千常用词表》、《现代汉字综合使用频度表》、《现代汉语用字频度表》等在语料统计基础上构建的词典先后出现，为中文信息处理相关标准的建立提供了科学的基础数据。

可以说，语料库的出现给词典编纂注入了新思路，其在词典构建中的应用也日益受到重视，然而上述应用终归只是将语料库作为词典编纂的一种研究方法，甚至只是一种数据生成的手段或操作工具，而真正由语料库驱动所得的一些重要

研究成果还尚未转化到机器词典的编撰上。鉴于此，我们试图从语料库驱动理念出发，探寻与机器词典构建相关的几个核心问题，如词典应描述什么、词典以什么为词条、词典采用什么释义模式等，以期在为语言信息处理提供更加可靠且适用的语言知识源的道路上做出一些有益的尝试。

1. 词典应描述什么?

机器词典作为语言信息处理的重要知识源，是对语言使用本质的结构化描述。因此，要回答机器词典应描述什么的问题，实际上就是回答语言使用本质是什么的问题。对于这一问题，不同的理论学派持有不同的观点，我们将这些观点分为两大类：

一类我们称之为“编码观”，偏重语言的形式特性。“编码观”认为，语言只是概念信息的符号编码，文本的构成单位如单词、短语等均有既定的意义（即词典中给定的释义），而文本的构建实际上就是根据所要传递的概念信息在语法规则的框架下对单词、短语等单位进行选择和排列的过程（即空位填充式）。这就意味着，在语言符号之外似乎还存在着一些已由词典定义好的、恒久不变的、来自于对外部客观世界反映的概念信息。于是，当需要传递某个概念信息时，使用者只需用语言符号对符合该表达需求的组件进行编码即可。由于这是一个编码过程，而编码过程通常是可逆的，从而认为在给定某个文本时读者或听者总能推断出确定、单一的概念信息；反之，在给定概念信息的情况下，说话者或写作者总能找到确定的、单一的文本来进行编码。

另一类我们称之为“交际观”，偏重语言的意义功能。“交际观”认为语言的基本功能是交际，而交际本质上是一种意义的交互，文本的生成是多个因素（包括词汇、语法、语义、语用等）共同选择的结果，并最终服务于特定的交际目的。“无论是语词还是语句，其功能不在于所指向的外部对象或事实，而在于交流过程中所起的作用”（维特根斯坦 2004：96）。因此，“交际观”十分强调语言使用对意义的决定性作用，因为意义的获取主要取决于语言使用者的解读与阐释。所以，在不同的交际场合下即使是相同的文本也有可能传递出很不一样的概念信息，在这一视角下，语言交际的过程就是语言社团成员对意义的协商过程，当意义得到绝大多数语言社团成员认可时，该意义可视为达到了相对稳定状态，而一旦新的解读进入语言社团的协商中来，则原来的意义就开始发生变化。

显然，上述两种观点之间存在本质的差别。“编码观”强调符号与概念间一一对应的稳定关系，强调语言的使用是对一套既定规则的遵循，重视词典对语言使用的规范作用。按照这一观点，只要掌握了编码和解码规则，语言使用者就能正确地产出和理解话语。然而，这显然与我们语言使用的实际情况不符。一方面，即便是熟练的语言使用者产出话语时也会有辞不达意的时候；另一方面，“一千个

读者眼中有一千个哈姆雷特”，同一词条在不同词典中的释义往往各有不同。因此，与“编码观”不同的是，“交际观”更强调语言的动态性和多样性，将语言视为一个开放的使用实例集合，强调实际使用对意义的决定性作用，重视词典对语言使用实际的客观描述。按照这一理念，无论是面向人的还是面向机器的词典构建均应以意义为核心。而意义源于何处？不是外在的客观世界，也不是人类心智中先天的某种规则或机制，而是语言社团成员实实在在使用过的语言实例（也称为文本或话语）。我们只有回归到这些真实的存在才能真正明白什么是语言使用的本质，即语言的本质。而这正是语料库驱动所遵循的语言观，因此我们认为词典的构建应以意义为描述核心，以文本或话语为探讨意义的出发点和立足点。

2. 词典以什么为词条？

确立了以“意义为描述核心、以文本作为意义的参照本源后”，接下来应解决的则是词条的确定问题。词条是词典最基本的构成单位，也是词典编纂者所定义的语言的基本单位。那么，语言的基本单位是什么呢？“编码观”遵循语言的形式特征，依照书写规范（如符号与符号之间是否有空格、符号与符号之间是否有连接等）将语言的基本单位界定为单个的、孤立的符号（如英语中的单词、汉语中的字或词等）。而由于英语存在大量屈折形态变化，一些单词由于形态的不同通常会有不同的使用并表达不同的意思，因而又有学者提出比单词更小的语言单位，即语素（morpheme），并在此基础上提出了诸如IA（Item-Arrangement）、IP（Item-Process）、WP（Word-Paradigm）以及IC（Immediate Constituent Grammar）等模型与理论（Sinclair 1996：24-25）。而另一方面，对于汉语这种书写中不包含空格的语言而言，由于有些字并不独立表达意义，通常需要与其他字组合共同表达某种意义（如“葡”和“萄”等），于是分词（从另一个角度看也是合词）成为了语言研究尤其是语言信息处理的一个基本的、核心的问题。

可以看出，上述语言基本单位的界定均侧重从语言形式角度出发，并未充分考虑意义本身，最终可能导致的问题就是：一方面，无法解释在语言使用实际中大量存在的复合词、成语、俗语、固定短语、行话等语言现象的意义生成问题；另一方面，无法解释在以单个词为词条的传统词典中大量存在的一词多义问题，以及由此带来的困扰语言学研究者以及语言信息处理研究者们的语言歧义问题。

对于第一个问题，若以语素或单词为语言基本单位，则复合词等多词单位的意义似乎应该可以通过单词意义的加和来实现。然而，我们从真实文本语料库中发现，语言中大量多词组合的整体意义不是构成该整体的单词意义的简单相加。Sinclair（1996）列举出了三种单个词与多词表达之间的意义关系：（1）每个单词意义均不与整体义直接相关，如bear on（表“涉及”义时）；（2）部分单词意义与整体义相关，部分则不相关，如to beat someone up；（3）整体义似乎与每个单词

意义均相关，如the rain beats down。汉语中也不乏类似现象，如“打酱油”（网络用语）、“甩手掌柜”、“吃食堂”等。这些组合的意义均无法从单个词语的意义中简单地推断出来。

对于第二个问题，当前的绝大多数词典以词为意义的首要单位，而同时绝大多数词在词典中的释义却超过一种。这一情况充分说明，作为语言单位的词易有多重的歧义，从而导致由这些词组成的语句或其他语言片段的释义数量显著增多。以英语中最常用的how do you do为例，按照柯林斯词典对how、do、you的单词释义，该句的意义理论上可达到13*32*2*32=26624种，而显然表述者所要传达的意义是单一且明确的。

实际上，上述两个问题对于语言使用者来说并未构成明显的理解障碍。在语言交际过程中，交际双方似乎并不需要实行上述组合运算便可实现顺畅的沟通与交流。这让我们不得不反思，人类对语言的理解是否真的是以词为单位的呢？实际上，一些学者已经在该问题上提出了很多极具启示意义的见解，如语言学伦敦学派创始人、“语境意义论”的提出者Firth就始终强调You shall know a word by the company it keeps（观其伴，知其性）（Palmer 1968：182）；Louw（1993）则认为词义与词义之间存在一种“触染”（contagion）（Sinclair 1996：150）。COBUILD项目负责人Sinclair（2004）在观察分析大量真实语料的基础上明确提出：“大多数词语在使用中靠与其他词语搭配确立意义”（李文中 2010：38）。这些理念与论断启示我们，将意义的分割放在一个扩展的环境下去考虑，不仅有助于我们区别同一个词在不同环境下的作用（也就是单个词义项的划分），也有助于我们找到一个单义的、相对稳定的语言意义单位。

那么，如何才能找到这一最小的、单义的、无歧义的、相对稳定的意义单位呢？ Sinclair（1996）提出了构筑意义单位的五个核心要素，即意核（core）、搭配（collocation）、类联接（colligation）、语义倾向（semantic preference）、语义韵（semantic prosody）。其中，后四个要素均围绕意核展开，依次体现了词语与词语、词语与语法、词语与语义、词语与语用之间的共现或互选。利用这五个核心要素，一些语言学研究者进行了更为具体的研究和探索，例如：Sinclair（1996，1998，2004）以英语的naked eye、true feelings、brook、place、budge、efforts等为意核，卫乃兴（2012）以汉语的“展现”为意核，濮建忠（2014）以汉语的“发展”为意核进行了单语种中意义单位的探索；另有卫乃兴（2011）、陆军、卫乃兴（2012，2013）等对双语意义单位对比进行了有益的尝试。这些研究均证明了“扩展意义单位”在确立单义的、相对稳定的语言单位上的有效性。

为进一步验证该模式的有效性，我们尝试从汉语单词“共同”出发探索对应的意义单位。通过对CQPweb（http://111.200.194.212/cqp/）（许家金、吴良平 2014）提供的The UCLA Corpus of Written Chinese (2nd edition)[1]语料库进行关键词

检索，共得到181条包含“共同”一词的索引行。经过初步考察我们发现，该181条索引由于共现规律的不同大致可以分为两大类：一类为“‘共同’+动词”，侧重于“合作”义；一类为“‘共同’+名词”，侧重于“相同”义。在此，我们仅以前者（共计126条索引结果）作为考察对象进行意义单位的探索，依次观察“共同”左右两侧搭配成分的词汇、语法、语义、语用等特征。首先，“共同”之后所接词语词性绝大部分为动词（占98.4%，仅两例例外[2]），典型的搭配词包括：

努力（9）[3]、发展（7）、奋斗（5）、繁荣（4）、面对（4）、推动（3）、关心（3）、生活（3）、作出（贡献或努力）（3）、携手（2）、参与（2）、参加（2）、富裕（2）

而在“共同”的左侧则存在大量诸如“需要”、“希望”、“愿”、“必须”、“欢迎”、“要求”、“有必要”、“想”等传递正面态度的词，于是我们将该意义单位的语义韵归纳为“希望”。继续观察“共同”的左侧搭配我们发现，典型的搭配词类集中于表数量或衔接的语素或词，前者如：“两”等数字词（22）、“们”（16）、“各”（13）、“全”（6）等；后者如：“与”（18）、“和”（15）等，意即参与动作的主体为两个或两个以上。由此，我们试图将该意义单位归纳为：

（1）意核为“共同”；

（2）搭配词为“们”、“各”、“全”、“与”、“和”、数字等；

（3）类联接为“动词”；

（4）表“合作做某事”的语义倾向；

（5）表“希望”的语义韵。

确定为上述意义单位模式后，对于诸如盗窃、舞弊等我们通常持负面态度的事件而言，我们就不会将其表述为“共同盗窃”、“共同舞弊”了，而是“合伙盗窃”、“合谋舞弊”等，而“共同心愿”（类联接为“‘共同’+名词”）、“共同特征”等表述，则由于不符合类联接规律而应划入其他义项。

按照上述相同的方法，我们对另一类共现索引进行观察分析得到以“共同”为意核，以“们”、“各”、“全/所有”、数字等词为搭配，以“名词”为类联接，以“相同的拥有”为语义倾向，以“认可”（如：“凝聚力量建设……共同目标”）为语义韵的另一个扩展意义单位。

综上我们认为，扩展后的意义单位模式更适合作为词典中的词条：一方面，它使得使用者（无论是语言学习者还是计算机）能通过识别这五个主要构成要素实现对语言的无歧义理解；另一方面，在存在明确意核和语义韵的情况下，该扩展意义单位还允许一定程度的变异，尤其是在语义倾向和语义韵层面，对词汇、语法等不作严格的限制和区分，仅从语义、态度的角度考虑，这充分遵循了语言动态性、灵活性、多样性的本质特点。

3. 词典采用什么释义模式?

现有的应用于语言教学、语言翻译或语言信息处理的各类词典（或语言知识库），无论是采用传统的解释型释义模式还是概念体系型知识库模式，均发挥了其自身在语言描述方面的特长并带来了一些实际的效益。然而，诸如“一词多义”、“多词同义”、“惯用表达”等问题却仍未得到有效的解决。随着信息时代的飞速发展，利用计算机来开展语言信息的处理成为了当前语言学研究的一个越来越重大的课题。如何充分利用计算机的特长来对其开展“语言教学”以帮助它克服上述障碍，更好地处理和生成自然语言，这是机器词典构建需要考虑的又一关键问题。对于计算机而言，其特长无外乎计数、记忆、匹配和存储。因此，对于供计算机来“学习”语言的机器词典而言，无疑应该以计算机能“理解”且又方便其开展计算、匹配等处理的方式呈现。

首先，语料库驱动的机器词典以“扩展意义单位”为基本释义单位，将传统模式中的词条、词性、词义以及例句等整合于一体。具体而言，对于具有显性词汇表现的成分如搭配、类联接等可采用列举模式呈现，便于机器进行匹配和识别；对于在显性词汇表现基础上提炼出的语义倾向、语义韵则可采用归纳式标注，以便机器进行记忆和存储。

其次，采用正则表达式来对释义内容进行表示便于计算机识别。所谓正则表达式，是用来表示匹配（或不匹配）某个字符串的特征模板。一方面，它可以匹配一个或多个给定的字符，也可以匹配所有除了给定的一个或多个给定字符以外的内容。正则表达式在匹配、检索方面的这些优势很好地满足了“扩展意义单位”灵活性和允许变异的特点。另一方面，正则表达式被广泛地应用到各种计算机语言或应用软件中，有助于实现机器词典在不同平台之间的移植与共享。

我们以上文中以“共同”为意核的扩展意义单位为例，采用正则表达式可分别表示为：

$word=~/([们|各|全|与|和|多|两|双|二|三]) (.+)(共同)(.+?)\s(.+?)\/v/)→希望多个主体合作做某事

其中“→”表示整个扩展意义单位的释义，此外还包括各种正则表达式操作符。我们通过编辑程序将上述匹配模板应用于对随机选定的一部分已经词性标注的语料进行意义单位的提取，得到结果如下：

D:\perl>uom2.pl

以“共同”为意核的意义单位识别结果：

共同+U（发展）->希望多个主体合作　发展　某件事

共同+U（奋斗）->希望多个主体合作　奋斗　某件事

共同+U（努力）->希望多个主体合作　努力　某件事

4. 结语

综上，我们在语料库驱动理念的指引下，探讨了机器词典构建的几大关键性问题，包括机器词典的语言描述核心、机器词典的基本描述单位以及机器词典的释义架构和表示方式等。随着电子文本数量不断增加，对大量语言事实的观察不仅成为可能，也成为探索语言本质和语言使用规律的必要手段。原本由于观察手段的局限性而总是被忽略的各种共现、互选现象（包括词汇与词汇的共选、词汇与语法的共选、词汇与语义的共选以及词汇与语用的共选）开始大量地呈现在语言观察者面前，并为语言的研究、语言的应用带来了前所未有的启示。鉴于这些启示，本文尝试提出一种新的机器词典构建模式，即：以意义为描述核心、以语言使用（或文本）为获取意义的本源、以集词汇、语法、语义、语用于一体的“扩展意义单位”为基本描述单位，采用列举与归纳相结合的释义架构和正则表达式的表示方法。然而，这还只是万里长征中带有试探性的第一步，未来还需在验证、实现以及拓展方面做出更多有意义的探索。

注释

1. 该语料库共计1,097,113词，由15个不同的文类构成。

2. 两例例外分别为“共同为世界的和平与进步作出贡献……”和“共同或经协商单独召集有关部门、机构、企业以及……”，实际上此处的“共同”同样指向后接的动词“作出贡献”和“召集”。

3. 括号中数字表示该词与“共同”的共现频次，未特殊标注的即为1，下同。

参考文献

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: John Benjamins. 157-176.

Palmer, F. (ed.). 1968. *Selected Papers of J. R. Firth 1952-1959* [C]. Bloomington: Indiana University Press.

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

Sinclair, J. 1998. The lexical item [A]. In E. Weigand (ed.). *Contrastive Lexical Semantics* [C]. Amsterdam: John Benjamins. 1-24.

Sinclair, J. 2004. *Trust the Text: Language, Corpus and Discourse* [M]. London: Routledge.

Teubert, W. & A. Cermáková. 2009. *Corpus Linguistics: A Short Introduction* [M]。北京：世界图书出版公司。

李文中，2010，语料库语言学的研究视野 [J],《解放军外国语学院学报》(2)：37-42。

陆　军、卫乃兴，2012，扩展意义单位模型下的英汉翻译对等型式构成研究 [J],《外语教学与研究》(3)：424-438。

陆　军、卫乃兴，2013，共选视域下的二语知识研究——一项语料库驱动的使役态共选特征多重比较 [J],《外语界》(3)：2-11。

濮建忠，2014，语料库驱动的翻译研究：意义单位、翻译单位和对应单位 [J],《解放军外国语学院学报》(1)：53-65。

卫乃兴，2011，基于语料库的对比短语学研究 [J],《外国语》(4)：32-42。

卫乃兴，2012，共选理论与语料库驱动的短语单位研究 [J],《解放军外国语学院学报》(1)：1-6。

维特根斯坦，2004,《哲学研究》[M]。上海：上海译文出版社。

许家金、吴良平，2014，基于网络的第四代语料库分析工具CQPweb及应用实例 [J],《外语电化教学》(5)：10-15。

通讯地址：471003　河南省洛阳市解放军外国语学院语言工程系（曹蓉、黄金柱）

310018　浙江省杭州市浙江工商大学外国语学院（濮建忠）

再谈汉语中介语语料库的建设标准*

北京语言大学 张宝林

提要：目前汉语中介语语料库正迎来一个繁荣发展的重要时期，同时也存在影响语料库建设的种种问题。语料库建设标准研究对解决这些问题具有重要意义，但尚未引起学界的广泛关注，更未展开深入的研究，建库实践的随意性依然存在并继续影响着语料库建设水平的提高。为此，本文提出了语料库建设标准的研究设想，重点讨论建设标准的研究内容和研究方法。期待学界尽早认识到该研究的重要意义，开展广泛的研究与讨论，以促进语料库建设的规范化、科学化和系统化，更好地为全世界的汉语教学与研究服务。

关键词：汉语中介语语料库、建设标准、研究内容、研究方法

1. 引言

1.1 研究现状

自1995年第一个汉语中介语语料库“汉语中介语语料库系统”建成以来，基于语料库的汉语偏误分析与习得研究逐渐成为对外汉语教学研究的重要内容。进入21世纪以来，汉语中介语语料库在汉语国际教育研究中的作用日益凸显，其建设渐成高潮，“成为语料库研究中的热点”（谭晓平 2014），并积极影响和推动了国内少数民族语言中介语语料库、特殊教育领域的中介语语料库、国外汉语中介语语料库的建设。今天，以“HSK动态作文语料库”为代表的一批汉语中介语语料库在国内外汉语学界已形成广泛的学术影响，建设队伍不断壮大，语料库类型日益丰富，笔语语料库之外，口语语料库、多模态语料库、语音语料库、汉字语料库纷纷投入建设，并已有成果问世。事实表明，汉语中介语语料库建设正在跨入一个繁荣发展的重要时期。

另一方面，汉语中介语语料库建设至今没有统一标准，不论哪一种语料库，不论是已建成的还是在建的，都是根据建设者的主观认识和研究经验进行建设，建库实践中存在很大的随意性。这种随意性表现在语料收集的原则与类型、规模

* 本研究得到北京市社会科学基金项目重点项目（15WYA017）、教育部哲学社会科学研究重大课题攻关项目（12JZD018）和国家社会科学基金项目（11BYY054）的资助。

和方式、背景信息的项目与内容、语料标注的范围、内容、方法与代码的设置、语料及相关背景信息检索与呈现的内容与方式、语料库建成后是否开放、资源能否共享等诸多方面，其后果是使语料库在规模、功能、质量、用法等方面存在很大局限，尚不能满足用户的多方面使用需求（参见张宝林、崔希亮 2015）。可以说，汉语中介语语料库建设中存在的随意性，已经成为制约语料库建设发展的关键问题。这个问题不解决，语料库的建设水平就无法提高，汉语教学与研究对语料库的迫切需求就无法全面满足。而破解之道，就是制定语料库建设标准。

随着20世纪80年代以来语料库语言学的复苏与发展、自然语言信息处理技术的进步，国内外母语语料库的建设得到极大发展。建设项目剧增，规模成十倍、百倍扩充，标注内容不断丰富、深度不断扩展，已从字、词扩展到句，再扩展到话语篇章、语义、语用。可扩展的置标语言（XML）已成为国内外学界普遍采用的通用标准。国内先后出台国家标准《信息处理用现代汉语分词规范》(GB/T13715-92)、《信息处理用现代汉语词类标记集规范》(GB/T 20532-2006)；其他相关研究则有《北京大学现代汉语语料库基本加工规范》(2002)、《973当代汉语文本语料库分词、词性标注加工规范》(山西大学 2003)、《资讯处理用中文分词规范》(台湾省 1998）等成果问世。然而，在母语语料库建设领域，全面对语料库建设进行规范的建设标准尚属未见，无法给汉语中介语语料库建设标准的研制提供参考与借鉴。

当前，建设标准已成为制约汉语中介语语料库建设水平与发展的瓶颈，20余年的建库实践则提供了进行建设标准研究的坚实基础。因此，进行汉语中介语语料库建设标准研究的时机已经成熟，亟需开展专项研究，以促进语料库建设的进一步发展和建设水平的提高，更好地为全世界的汉语教学与研究服务。

1.2　研究意义

建设标准是汉语中介语语料库建设经验与教训的总结，凝聚着学界对语料库建设的理论思考，标志着语料库的建设水平，对语料库建设具有重要指导意义。

*促进语料库建设水平的提高和规范化。*研究将为正在繁荣发展而又问题丛生的汉语中介语语料库建设制定标准，以促进语料库建设的系统化、规范化、科学化，提升其建设水平，推动其进一步发展，并最终决定着语料库的功能和实用价值。

*为汉语教学与研究提供更好的优质资源，促进资源共享，更好地为汉语国际教育事业服务。*建设标准研究是汉语中介语语料库建设中带有全局性的重大问题，不仅关系到语料库建设本身，对基于语料库的汉语教学与相关研究同样有重大影响。语料库建设水平的提升，在全世界范围内的免费开放，将为全世界的汉语教学与研究提供优质资源，更好地为汉语国际教育事业服务，为落实国家的语言政策服务。

*为母语语料库建设标准的研究提供参考和借鉴。*汉语中介语语料库是在国内外语料库语言学及母语语料库、特别是汉语母语语料库的研究、建设与发展的影响与推动下开始建设并发展起来的，但如上文所述，目前在汉语母语语料库建设领域尚未提出建设标准，也缺乏相关研究。汉语中介语语料库建设标准的探讨与研制将反哺汉语母语语料库、乃至其他语言的母语语料库的建设，为其建设标准的研制提供参考与借鉴。

2. 研究内容

2.1 基本认识

语料库建设本身是一项复杂的系统工程，语料库建设标准研究同样是一项非常复杂的系统工程，涉及语料库建设的整个过程，并要充分考虑到汉语教学与研究领域对语料库的多方面需求。根据以往的建设经验和相关研究，我们认为可以把语料库建设过程细化为彼此相关的11个方面的工作，分别对这11个方面进行研究，得出11个方面的标准，进而整合为汉语中介语语料库建设的总标准。

2.2 主要内容

*语料库建设流程研究。*研究语料库的建设过程，归纳建库的必有环节与程序，设计建库标准流程，可以使当前与今后的建库工作能沿着正确的途径，按部就班地进行，而无需每建一库都从头开始，重复前人走过的弯路，从而避免低水平重复，提高建设速度。

进行建设流程研究，首先需要考察汉语中介语语料语料库的建库实践，对建库的实际过程进行总结，归纳其环节与步骤；进而分析各个环节的必要性，将不可或缺的环节纳入标准流程，而舍弃不必要的环节，合并可有可无的环节。例如我们曾经根据“HSK动态作文语料库”的建设经验，认为建库环节应包括提出建库任务、进行总体设计、落实经费、组建团队、落实语料来源、收集与整理语料、语料录入与校对、标注规范的设计、试标、研讨、修改、定稿、语料标注的实施及检查修改、各种数据的统计与表格编制、网络版上网试运行及修改、正式发布、单机版开发等18个环节。经过进一步分析与删减合并，将其简化为10个步骤，即：a）提出建库任务，进行总体设计；b）语料的收集与整理；c）语料相关背景信息的收集与整理；d）语料的录入与校对；e）制定标注规范与实施语料标注；f）开发人工辅助标注工具；g）各种数据的统计与表格编制；h）语料库管理软件与检索系统的开发研制；i）语料库集成与上网试运行；j）语料库发布与开放（参见张宝林、崔希亮 2015）。显而易见，经过简化的建库标准流程更加简明扼要，便于把握，而且有充分的理据。例如提出建库任务和进行总体设计是紧密相连的两个步骤，有些内容需要同时考虑，甚至某些总体设计的内容在提出建库任

务之前就要考虑清楚，因而完全可以合并；落实经费与组建团队是建库前提，非常重要，但属建库之前就必须解决的问题，无需纳入建库的具体流程。新流程删除了单机版语料库的开发，因为在当今这个以互联网、云计算、大数据为显著特征的时代，网络已经普及，做单机版的意义不大；如果以实时统计的方式处理数据，则数据统计与表格编制也是可以舍弃的。

语料收集标准研究。研究收集什么样的语料、怎样收集的问题。谈及这个问题，首先涉及语料的平衡性，学界对此多有讨论，例如，任海波（2010），施春宏、张瑞朋（2013），张宝林（2012）、张宝林等（2004），张宝林、崔希亮（2015）等。究竟怎样科学地理解这一概念以达到真正的平衡性？是追求“理想化的绝对平衡”，还是接受在分层抽样基础上达到的“实事求是的平衡”，学界意见并不一致，还有待深入的研究，并在此基础上形成平衡性的具体指标。其他具体问题还有许多，例如写作语料是只要有录入版即可，还是同时提供扫描版？原始语料可能不止一页，那么两页或几页原始语料在扫描时要不要合并成一个图片文件？怎样合并？是只要作文语料，还是也可收集造句语料？对这些问题，学界的看法与做法并不一致，而对这些问题的不同处理方式则体现了语料库建设者的不同认识、原则与理念。例如“HSK动态作文语料库”只收集学习者自主产出的成篇语料，而不收集造句语料；不但有录入版语料，也有原始语料的扫描版。该库建设者之所以这样做，是出于对语料真实性的考虑与重视：学习者参加标准化考试即时写作产出的成篇语料是其汉语书面语水平的最真实表现，具有最充分的真实性，而造句语料则完全可能是对某些句子的模仿，其真实性较差；扫描形式的原始语料不但可供用户审核语料的真实性，而且为中介语汉字研究提供了学习者书写的汉字材料，从而方便了对学习者汉字的考察与研究。在收集纵向语料时数据收集的频率和时间间隔也是一个需要仔细考虑的十分重要的问题（曹贤文 2013）。至于汉语中介语口语语料库尚不多见，多模态语料库尚无建成者，口语语料和多模态语料的收集非常缺乏实践经验，研究也很少，尤应作为本项研究的重点。

语料背景信息收集标准研究。语料是语料库建设的前提，没有语料就无法建设语料库。但是，如果只有语料而没有与之相关的背景信息，语料就失去了分析的基础，也是没有价值的。为了充分而有效地使用这些语料，就要从基于语料库的中介语研究的角度出发，研究应该收集哪些背景信息、怎样收集的问题。从目前语料库建设的实际情况看，一般都能提供语料作者的某些背景信息和语料自身的某些信息，例如作者国籍、语料性质（考试答卷还是平时作业）等，但是一般没有作者的母语或第一语言的信息，而这对分析学习者中介语状态的形成原因是非常重要的。在收集方法方面，调查问卷是收集学习者信息的一种很好的方法。可能遇到的问题是，有些学生会出于保护个人隐私的考虑，不愿配合；通过学校的学籍管理系统和教务系统进行收集，又欠缺某些重要信息，例如学习者的母语

信息。如何使学习者乐于配合从而有效收集到这些信息，亟需进行研究并找到切实可行的办法。

语料录入标准研究。研究怎样把书面语料录入电脑并确保其真实性。这个问题看似简单，实则包含着很大的难题，例如对语料中错字的处理便是其中之一。错字电脑是打不出来的，录入时如何处理？用造字程序仿造的字失真度太大，无法体现出学习者书写的汉字的真实面貌；直接录入正确字，更是完全失去了“中介汉字”的特点；用照相的办法嵌入错字，效果非常理想，能完全真实地体现中介汉字的原貌，但程序繁琐，处理速度很慢；还可以附带扫描版语料来保留中介汉字原貌，但录入版和扫描版两版语料中的错字如何定位也是不易解决的难题，应用起来也不是很方便。究竟如何处理，亟需研究。又如语料中行款格式方面的偏误录入时如何处理？也需要进行研究并确定其标准。

语料转写标准研究。研究怎样把口语和多模态语料录入电脑并确保其真实性。这方面的问题更多，例如转写时如何处理口语语料中长短不一的不正常停顿？如何处理声、韵、调的语音偏误？如何处理视频语料中的体态语？等等。口语语料中的转写和语音层面的标注是什么关系？转写与语音标注应合为一次完成，还是分为两次完成？均有待研究，并应有统一标准。语料转写的真实原则、标注完整原则、判断标注内容的准确原则，虽然针对的是英语学习者口语语料库的建设（卫乃兴等 2007），但对汉语中介语口语语料库的建设而言，也是完全适用的。

语料标注原则研究。语料标注是建库中最为重要的环节之一。语料标注研究主要是研究标注模式，包括标注原则、标注内容、标注方式、标注代码、标注流程等问题（张宝林 2013），其本身的内容十分复杂，不是在建设标准研究中能够完全解决的，而是需要进行专门研究，并需通过标注模式研究最终形成标注规范以解决与标注相关的全部问题。然而标注的基本原则应在建设标准中予以规定，例如标注的科学性、系统性、规范性，通用型语料库标注的全面性，语料库建成开放时向用户公布标注的错误率等，皆应成为语料标注的重要原则。语料标注的全面性不仅关系到语料库的建设水平，更决定着语料库的功能和实用价值，可谓非常重要，然而学界对此原则的理解尚存在不同意见。有研究认为，全面性指“应在字、词、短语、句、篇、语体、语义、语用、标点符号等各个层面上对相关的语言现象进行标注，这样才能保证语料库功能的全面”（张宝林 2013）。也有研究认为，“汉语中介语语料标注的全面性应该从标注的广度、深度、角度和准确度四个维度来思考”（肖奚强、周文华 2014）。还有研究认为，“面临对语言现象标注的种种问题……具体对策是：先标注成熟项目；逐步增加标注项目”（张瑞朋 2012）。究竟如何理解与看待这一问题，尚需进一步深入研究与讨论。在汉语中介语语料库建设中是否需要进行分词和词类标注也还存在着完全相反的意见（肖奚强、周文华 2014），因而仍然需要探讨。

语料呈现标准研究。在用户需求分析的基础上，研究语料的检索方式、呈现内容、呈现方式和下载方式，以便根据用户使用电脑的习惯，以用户方便的方式来满足其对语料库的使用需求。具体来说，在语料检索和呈现方面应力求使用户“在检索语料时感到简单方便，在获取语料时感到足量快捷，在解读语料时感到清楚易懂”（任海波 2010）。例如语料应可以按单一条件或多重条件进行检索，多重条件检索功能强大，有助于用户查询使用。语料可以按单句形式呈现，也应可以按复句形式呈现，还应可以按段落形式呈现。不论以何种方式呈现，都应可以在查询到的语料后面显示单项、多项乃至全部背景信息。查询到的语料除了可以在线观察使用，还应可以下载到本地电脑，为用户提供更多、更大的使用方便。口语和多模态语料库则应可以分别以音频、视频和文字形式呈现，也可以音频、视频、文字等3种形式同时呈现。

语料库使用标准研究。研究语料库建成后的使用方式。例如语料库是否向学界乃至社会开放？部分开放还是全部开放？有偿开放还是免费开放？其实质是语料库的资源共享问题，在目前的汉语中介语语料库，乃至母语语料库的使用方面，开放的少，免费开放的尤其少，是一个非常突出而敏感的现实问题。我们从“HSK动态作文语料库”建设之始就主张、宣传和呼吁语料库应向学界免费开放，并在语料库建成之后率先垂范，从上网运行的第一天（即2006年12月24日）起，就向全世界开放该语料库，任何人都可以免费登录浏览使用。但是十多年来这种宣传与呼吁的成效与影响十分有限，至今只有北京语言大学、中山大学、暨南大学华文学院三家的语料库向社会开放。由此可见，只靠宣传与呼吁很难实现资源共享的目标，因而迫切需要在语料库建设标准的研究中制定明确的规则，采取切实有效的办法，真正促进语料库的开放与资源共享。

作者隐私的保护。语料库在服务教学与相关研究的同时，必须保护语料作者的个人隐私，遵守相关法律。例如作者的姓名、语料中显示的人名等个人信息如何处理？可否在语料库中直接披露？多模态语料中的人物形象可否显示？怎样和相关学习者解决其隐私权和肖像权问题？这些问题非常敏感，处理不好将直接引起法律纠纷。这是我们所不愿意看到的情况，应在建设标准的研究中进行深入研究，彻底解决这一问题。

语料库管理程序和检索系统标准研究。研究语料库软件系统的研发标准。例如检索系统是否简洁友好与语料库的使用密切相关，决定着能否最大程度地发挥语料库的效能和使用价值，标志着语料库为汉语教师、学习者、研究人员的服务是否到位，必须予以充分重视。电脑辅助标注工具所使用的语言，可扩展的置标语言XML早已成为国内外母语语料库建设的普遍标准，而在汉语中介语语料库的建设中几乎尚未得到应用，这种情况不利于语料库的资源共享，应当尽快改变。这需要计算机软件开发人员的努力，更需要建设标准的促进与推动。

语料库建设质量标准研究。研究从哪些角度、采用何种方法对语料库的建设质量进行控制。这个问题是目前语料库建设与研究中的薄弱环节，学界尚无相关讨论，更缺乏具体的研究成果。这与语料库的资源共享状况密切相关：免费开放的语料库本来就少，研究者有语料库可用已属难得，遑论其质量如何？然而，质量不佳的语料库实际上并不能给研究带来有实际意义的帮助，甚至会使研究者得出错误的结论，从而误导学界。因此，亟需立即开展这方面的研究。

3. 研究的思路与方法

3.1　研究思路

（1）调查研究，了解情况

需求分析：通过问卷调查、访谈与座谈，了解教师和研究人员的教学与研究需求，了解他们对语料库的意见、要求与期望；

文献研究：通过中国知网（CNKI）、读秀知识库等资源库中的专业论文和学术著作，了解国内外中介语语料库和母语语料库、少数民族语料库、外语语料库的建设与研究情况，了解国内外与语料库建设与应用相关的标准与规范；

实地考察：通过登录实际使用国内外开放的语料库，或走访语料库建设单位，了解现有与在建汉语中介语语料库的具体情况与特点，并考察国内外有影响的母语语料库、少数民族语料库、外语语料库，吸取其成功经验，作为参照与借鉴。

（2）整理归纳，对比分析

对经调查与考察而了解到的上述三方面情况进行整理、分类、统计，通过对比分析，确认现有汉语中介语语料库的优点与不足、成就与差距、可以继承发扬之处和应予以改进之处。

（3）研发设计，形成草案

根据现有语料库的长短优劣，并针对用户教学与研究的具体需求，总结语料库建设与研究的经验和规律，设计改进方案，同时参照现有的一些相关标准和规范，草拟语料库建设标准（草案）。

（4）专家咨询，实验检验

请语料库建设与应用、中文信息处理、汉语教学与研究领域的专家、学者、教师，对标准（草案）进行评审与咨询；

用已建成和在建的语料库对标准（草案）的各项内容进行比较验证，以确认其有效性和可行性。

（5）反复修改，确定标准

在实验研究中发现建设标准的问题，加以改进，再实验，再改进，直至最终形成标准定稿。

上述研究思路可以概括为图1。

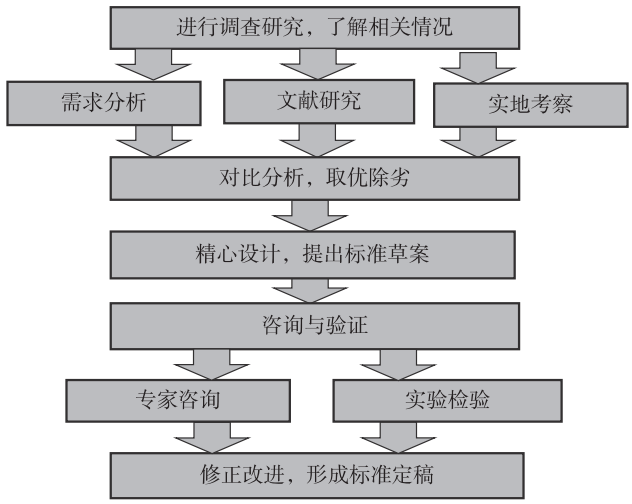


图1. 语料库建设标准研究流程图

3.2 研究方法

问卷调查法、座谈访谈法：这些研究方法是传统的，也是有效的，可以用于调查研究，进行需求分析。

文献研究法：用于调查研究，了解国内外中介语语料库和母语语料库、少数民族语料库、外语语料库的建设与研究情况，以及国内外语料库建设与应用方面的各种标准与规范。需要特别指出的是，由于网络技术的发展，以及中国知网（CNKI）、读秀知识库等文献资源库的建成与开放，研究者除可以十分方便地获取学术论文等研究资料外，还可以得到有助于研究的多种帮助，例如按照文献、期刊、会议等对某类论文进行查询；对查询到的论文还可以进一步按发表年份、作者、机构进行查询；可以查到每篇论文的下载频次和被引频次；可以查到某研究领域的学术趋势、学术研究热点等。资源库的这些功能对检索文献、提高文献研究效率具有重大作用。

实地考察法：用于调查研究，了解国内外已上网开放的汉语中介语语料库的具体情况与特点，以及国内外有影响的母语语料库、少数民族语料库、外语语料库的建设与使用情况。

对比分析法：用于考察与甄别各类、各个语料库的优点与问题，寻找原因，并研究解决办法。

专家咨询法：对研究中发现的难以解决的问题、提出的标准草案等，请学界相关领域的专家学者提供意见，进行指导，以解决遇到的重大问题，使研究工作得以顺利进行。

实验研究法：建设标准草案除请专家把关提供意见之外，还需经过实验研究的检验以证实其有效性与可行性。例如用拟定的标准进行衡量，应能发现已建成或已完成设计的语料库的不足之处。假定把附有外国汉语学习者产出的原始书面语料作为笔语语料库的一条建设标准，那么如果一个汉语中介语语料库只有电子录入版的标注语料而没有收入原始语料，就是一个不合标准的语料库。

4. 结语

汉语中介语语料库经过20余年的探索与发展，已经从当初的筚路蓝缕形成今天蓬勃发展的良好势头。为了使其发展得更好，确实需要进行一些回顾、反省与总结。正如“一个人，一个组织，一个国家，在永恒不断流逝的时间的长河中，到了一定的时候，应该回头看一看，看看走过的历程中自己走得是否都完全正确，正确的要坚持，不正确的要扬弃。这是十分必要的”（季羡林 2008）。研究并制定“汉语中介语语料库建设标准”就是要回头看一看20年来走过的历程，并在调查研究、总结经验教训的基础上，规范语料库建设流程，促进语料库建设的规范化、科学化、系统化，提高其建设水平，促进其更好地发展，最终更好地服务于对外汉语的教学与研究。

在以往的语料库建设与研究中，我们认识到，汉语中介语语料库建设中存在的主要问题之一是，语料库建设没有统一标准，建库实践带有很强的随意性，因而影响着语料库的建设水平与功能（张宝林，2010）。在“全球汉语中介语语料库建设与研究”项目中我们提出了语料库建设标准这一研究任务（崔希亮、张宝林 2011；张宝林、崔希亮 2013）。然而时至今日，专题探讨这一问题的论文可谓少之又少，这一问题尚未引起学界的关注。而这一问题不真正解决，汉语中介语语料库建设的研究就难以深入，其建设水平也就无法提高。

本文在梳理汉语中介语语料库建设与使用中诸多问题的基础上，提出了进行语料库建设标准研究的具体设想，包括建设标准的研究内容和研究方法。这只是我们的主观设想，是否恰当可行还需要研究实践的检验。我们期待学界对这一问题有更多的关注与探讨，欢迎大家的批评和指教。我们相信在学界同仁的共同努力下，语料库建设的本体研究将得到长足发展，语料库建设的理论思考将得以不

断深化，而最终的结果一定是汉语中介语语料库建设水平的极大提高，质量高、功能强的语料库的更多涌现。

参考文献

曹贤文，2013，留学生汉语中介语纵向语料库建设的若干问题 [J],《语言文字应用》(2): 127-134。

崔希亮、张宝林，2011，“全球汉语学习者语料库”建设方案 [J],《语言文字应用》(2): 101-108。

季羡林，2008,《忆往述怀》[M]。西安：陕西师范大学出版社。

任海波，2010，关于中介语语料库建设的几点思考——以“HSK 动态作文语料库”为例 [J],《语言教学与研究》(6): 8-15。

施春宏、张瑞朋，2013，论中介语语料库的平衡性问题 [J],《语言文字应用》(2): 118-126。

谭晓平，2014，近十年汉语语料库建设研究综述 [R]，第七届北京地区对外汉语教学研究生论坛论文，北京。

卫乃兴、李文中、濮建忠,2007,COLSEC语料库的设计原则与标注方法 [J],《当代语言学》(3): 235- 246。

肖奚强、周文华，2014，汉语中介语语料库标注的全面性及类别问题 [J],《世界汉语教学》(3): 368-377。

张宝林，2010，汉语中介语语料库建设的现状与对策 [J],《语言文字应用》(3): 129-138。

张宝林，2012，关于汉语中介语语料库建设的若干重要问题 [J],《数字化汉语教学——2012》[C]。北京：清华大学出版社。

张宝林,2013，关于通用型汉语中介语语料库标注模式的再认识 [J],《世界汉语教学》(1): 128-140。

张宝林、崔希亮，2013，“全球汉语中介语语料库建设与研究”的设计理念 [J],《语言教学与研究》(5): 27-34。

张宝林、崔希亮，2015，谈汉语中介语语料库的建设标准 [J],《语言文字应用》(2): 125-134。

张宝林、崔希亮、任　杰，2004，关于“HSK 动态作文语料库”建设构想 [A],《第三届全国语言文字应用学术研讨会论文集》[C]。 香港：香港科技联合出版社。

张瑞朋，2012，留学生汉语中介语语料库建设若干问题探讨——以中山大学汉字偏误中介语语料库为例 [J],《语言文字应用》(2): 131-136。

通讯地址：100083　北京市北京语言大学语言科学院

语料库语言学与文献计量学的交汇和互补

河海大学 周红英
解放军国际关系学院 李德俊

提要：语料库语言学与文献计量学是不同学科领域的定量研究方法：前者以使用中的语言为基础，基于数据特征和相关语言学理论进行语言意义/功能的阐释；后者基于期刊文献数据库及相关指标测定进行学术研究及技术情报等方面的信息分析。二者就基本分析单位、数据要求以及学术语篇这类研究对象上存在共同点。从研究方法的跨学科视角看，语料库方法与文献计量学方法在具体方法上可以相互借鉴和补充，既可有效拓展语言学研究的视域和维度，又可解决网络环境下传统文献计量学分析面临的某些难题。

关键词：语料库语言学、文献计量学、交汇、互补、跨学科视角

语料库语言学与文献计量学是分属两门不同学科的研究方法：前者是以使用中的语言为基础的定量研究方法，主要用于语言研究，比如，语言结构描写、语言变异研究、词典编撰、词汇搭配研究及语言教学等；后者是基于期刊文献数据库的图书馆情报学研究方法，多用于提取相关学科科技文献各层面的信息，如相关论文数量、作者及年代分布、学科分布、期刊分布、引文量等，或在某一领域中进行学科分析，描述学科结构，其早期应用领域主要是自然科学和医学，后来扩展到人文社会科学领域及其他学科。作为计量实证研究方法，二者都依靠数学工具和统计技术的支持。随着统计技术的发展，文献计量学的数据统计功能已能够在学术、技术情报分析中发挥重要作用，展现出重要的学术价值，比如揭示热点、重点研究领域、新兴研究领域和研究的动态发展，预测未来研究趋势等。作为相互独立的学科研究工具和方法，二者又各具其功能与优势。在具体研究中将二者结合起来使用，一方面可使其各自的功能与优势互为助益，扬长避短，提高研究结果的准确性、完整性和可靠性，另一方面又可促进学科研究方法的交融、学科研究范围与深度的拓展。

1. 语料库语言学：应用及测量指标

语料库语言学滥觞于20世纪60年代初。为了突破当时（以Chomsky生成语言学为主的）语言研究的内省思辨方法和思路，展开语言的实证研究，美国布朗大

学的两位语言学家W. Nelson Francis与Henry Kučera建成了世界首个机读英语语料库——布朗语料库（Brown Corpus）。该语料库收集了100万词次的书面语文本。在此之前，以计算机的使用为界，存在依靠手工收集大量相似文本的传统语料库时期，计算机问世后，其海量存储和处理功能使得传统语料库时期跨入制作和使用机读电子文本库的现代语料库时期。20世纪80年代，随着计算机软、硬件技术的进一步成熟，现代机读语料库成为语言结构描写和语言理论研究的得力工具，语料库语言学作为语言学研究的方法论和实证方法在语言学领域正式立足。

语料库存储语言的使用材料，可以为语言研究提供数据来源，但语料库的功用并不局限于语言研究。电子语料库的发展极大地提高了存储、查询文献和引用来源的便利度和效率。20世纪70年代初，语料库的关键词检索功能激发了图书馆学和信息学领域研究者的灵感：关键词检索后来取代了传统目录索引卡片，也成为自动主题分析的一种有效方法（Hines *et al.* 1970，转引自McCarthy & O'Keeffe 2009：4）。

从布朗语料库的建立至今，现代语料库的种类日趋丰富。从性质和用途上看，有通用语料库、专门语料库；从介质上看，有文字语料库、声音语料库、多模态语料库；从语体上看，有书面语语料库、口语语料库；从时间上看，有历时语料库、共时语料库；从语种看，有单语、双语、多语、平行、非平行、母语、外语学习者语料库（杨惠中 2002：42）。这些语料库已经和正在满足不同的语言研究目的和需要。

1.1　应用

语料库作为一种语言研究工具，广泛运用于语言研究的各个领域，如词典学、理论及应用语言学、话语分析、社会语言学，以及文(体)学、翻译等领域。从分析单元上看，语料库可用以进行词、短语（搭配）、句子结构和语篇等不同层次的研究。比如，进行基于语料库的词义、用法研究，以此作为词典编撰的基础；发现语言结构和规律，对语言研究中提出的假设包括语法理论进行验证，以及在此基础上对语言假说的可证伪性、完备性、简约性、客观性和解释力度进行验证，力图使语言研究达到科学研究的标准（Leech 1992：112-113）；揭示话语层次的某些语言特征，比如节点词的篇章结构和话语功能与特征，来观察语言的语类变异（在不同语境或交际场合下使用特征的变化），等等。在宏观语言研究方面，语料库可以用来监控语言变化、研究语言的历时演变（如柯林斯-伯明翰大学国际语料库COBUILD）。

1.2　主要测量指标

语料库自身并不能提供任何关于其语言数据的有意义的信息。对数据信息的

揭示依赖于语料库软件对相关指标的检索和测量。这些基本测量指标包括：

词频：语料库检索软件如WordSmith、AntConc、Xaira等均可对语料中的所有未经削尾的原词进行频数统计，并将词频列表呈现出来，还可以自动提取包含节点词的所有行，称为索引行。索引行是节点词的上下文语境所在，可用于观察节点词的搭配和类联接倾向、句子结构上的典型用法及其他规律性特征。借助词频统计，通用型的大型参照语料库（如BoE、BNC等）能够揭示（英语）语言中最常用的基本词汇；而由某个特定领域的文本所构成的专门语料库中，所测得的高频词（功能词除外）往往与文本主题内容密切相关，因而也是该专门领域的主题词汇。词频指标是词典编撰及词汇教学的重要依据。

关键词及关键性：语料库语言学意义上的关键词在理论上指的是“在一篇或一组文本中具有这样的质性的词：能够揭示文本的内容，体现词在文本中的重要地位”（Scott & Tribble 2006：73）。从统计上看，关键词在语料库中的使用频率跟某一标准相比而显著偏高。关键词的测定需要基于某个大型的通用语料库，如英语研究通常用大型通用语料库BNC作为参照语料库。在此意义上，关键词与高频词是两个不同的概念：前者是比照参照库而进行统计所确定的，相较于参照库的标准，这些词在目标库中的出现频数达到了统计显著性，其使用情况一方面揭示了相关语料的形式风格特征（频繁使用或避免某些词汇）乃至语料作为某一语类的语言特征；另一方面，某些关键词无疑能够反映相关语料的主题内容，从而凸显语料的内涵特征。而与关键词相比，高频词不仅在作为研究对象的目标语料中，而且在日常语言中也可能频繁使用。关键词检索中，参照库的作用主要就是设定阈值，以过滤这一部分高频词。值得指出的是，虽然大多数研究倾向于把关键词定义为“频率显著偏高的词”并从这个角度对之加以关注，但检索软件在语料库中实际还检索到另一部分关键词——频率显著偏低的词。这部分低频关键词也能提供有意义的信息，因为与高频关键词所代表的高关注度和核心性相比，它们的弱势地位不仅反映了某一类语篇的语言使用特征（Evison 2010：128），还可能揭示那些处于边缘地带、未得到足够关注的概念或内容。

词汇搭配：词汇搭配研究是语料库语言学研究的一项重要内容，考察语料库中具有统计显著性的、且有意义的词汇同现关系。从表层形式看，词汇搭配反映的是语言使用中词汇的组合关系和共现倾向。另一方面，节点词在长期的使用中受到共现词的语义特征或评价色彩的感染而可能反映出某种特定的价值观、文化意义、性别刻板印象等（Moon 2010：208）。搭配的意义或显著性的大小就是搭配力。在统计上，搭配力的运算（如互信息值、T值、Z值）基于词频统计值，用以判断相关节点词的共现几率大小。

在重视实证研究的大环境下，基于语料库的语言研究方法日益成为语言学等人文社会科学领域的重要研究方法。

2. 文献计量学：应用及测量指标

2.1 应用

文献计量学（Bibliometrics）是图书馆学及情报学领域的一个分支。文献计量研究以引文数据库或其他数据源作为数据基础进行学科分析，描述学科的内部结构，或反映相关学科科技文献各个层面的信息。面向学科的文献计量学分析可借以进行学科、学术评估，帮助科研管理部门遴选学科发展的重点领域和优先支持领域，制定科研发展战略，进行科研管理与决策；也可用于比较国内外同领域机构的研究状况（蒋颖 2013：285）。近年来，随着统计技术的发展，文献计量学在科研情报领域展现出重要的方法论价值：科研情报机构大量收集某一领域的有效文献，借助文献计量学方法对相关指标进行测定，用以分析、提取相关领域在研究热点、重点及发展现状和未来趋势等方面的情报信息。

2.2 主要测量指标

文献计量学的计量功能借助于计算机软件的文本处理技术和数学统计功能。在这两大功能的辅助下，传统的文献计量学可以进行多种测度指标的统计，实现不同的研究目的。

关键词统计：关键词是文献计量学考察文献内部特征的基础。文献的内部特征主要指研究主题、内容、涉及的领域分支等。对一篇文献而言，其关键词往往反映出该文献涉及的主要内容和学界的关注点。统计一段时期内某个特定学科或研究领域大量研究成果的有效关键词，可以揭示相关研究的总体内容特征、各研究分支或子领域之间的内在联系、研究的动态发展脉络及发展方向等。关键词的词频统计可以初步揭示研究领域与研究方向上的研究热度（受关注度）和深度；词频的阶段性变化统计可用以分析研究领域与研究方向的稳定、变迁以及新兴领域和发展趋势。

理想状况下，关键词包含的意义是：首先，关键词是反映特定研究领域的学科专业术语；其次,同一篇文献使用的不同关键词术语之间有一定的联系；第三，如果有足够多的研究文献使用了某一对关键词术语,那么这对术语的关系在这些文献所关注的研究领域中具有特别的意义；第四，经过培训的标引者为文献选用的关键词是相关科学概念中可以信赖的一个指标。可靠、有效的文献计量学分析以上述4个假设的全部成立为前提（He 1999）。

共词分析：共词分析（co-word analysis/co-term analysis）是文献计量学的另一重要统计指标。“共词”指的是关键词两两共同出现在同一篇文献中。关键词两两在同一环境中共同出现的频次体现了相关关键词的共现强度。最常用的共现强度指标是共词强度，即通过共词分析测定两个关键词在同一环境中同现的频次。共词强度分析揭示共现词或概念所代表的研究领域、分支之间的亲疏关系。共词分

析最早是在20世纪70年代中后期由法国文献计量学家提出，其思想来源于文献计量学中的引文耦合[1]与共被引[2]的概念。共词分析法通过统计一组关键词或主题词两两在同一篇文献中出现的次数，以此为基础进行关键词聚类分析，从而得出这些词之间的亲疏远近的关系，进而分析出这些关键词所代表的学科或者主题的结构变化情况（蒋颖 2013：260）。共词分析的算法基础是：首先确定和统计作为分析对象的关键词，再两两统计关键词的出现频次，构建共词矩阵，最后根据研究主题来确定多元统计方法。

网络中心度：文献计量学研究常常选用社会网络研究[3]的分析软件UCINET作为分析工具，分析关键词的共词矩阵，对关键词共词关系网络中各个节点之间的相互联系进行可视化呈现。网络中心度（degree centrality）是社会网络理论中最重要和常用的概念工具之一，用以测度网络节点在网络中的位置或优势的差异。在文献计量学分析中，把要考察的对象学科整体看作一个网络，学科内的各个领域分支看成是构成网络的各节点，测度节点在整个网络中的位置，在此基础上判断学科中各个领域的研究状况和地位（见图1）。网络节点由表征学科研究领域的关键词充当，中心度反映了网络节点也即研究领域的重要程度，这些测定数据可用于期刊、论文、作者及研究现状的评价。网络中心度分化为两个常用指标：1）点度中心度（point centrality），用于测度网络中一个节点与其他节点直接连接数的总和，从而反映一节点同其他各节点的关系及关系的疏密程度；2）间距中心度（betweenness centrality）：网络图中某一节点与其他节点之间间隔距离的大小，表示一个节点（研究领域分支）在多大程度上与其他节点（研究领域分支）相关联（同上：173）。

图1为外语教学中自主学习研究的社会网络图谱（转引自唐进 2012）。

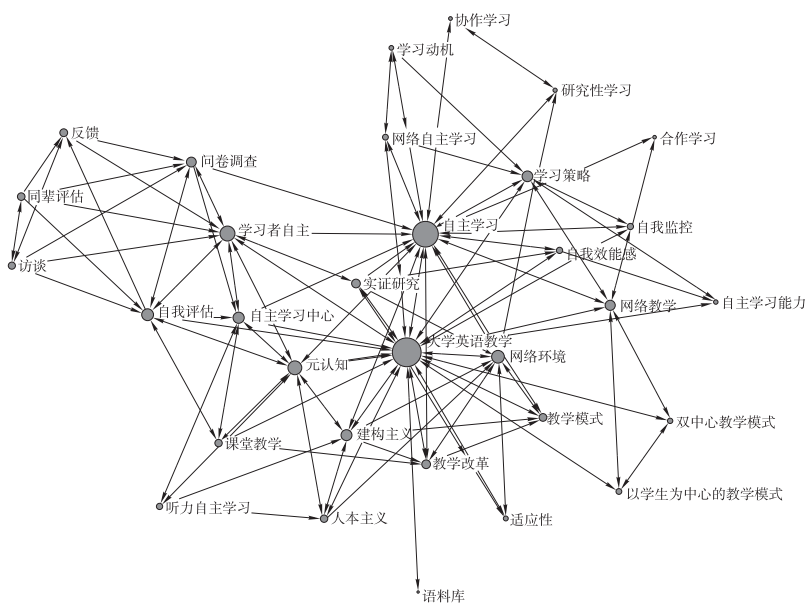


图1.　外语教学研究领域社会网络图

此外，还有一些用以考察文献外部特征、进行期刊评定的指标，包括期刊载文量、总被引频次、影响因子、即年指标、被引半衰期、自引率、自被引率、他引率等。文献计量学最常用的学术评定功能也是通过统计这些指标进行的。

文献计量学研究的可靠性取决于数据是否充足。鉴于文献计量学的定量研究可以相对客观、全面地描述学科对象的研究状况，其适用范围已经从传统的图书馆管理和科研管理与评价扩展至科技情报研究以及其他自然科学和人文社会科学领域。

3. 语料库语言学与文献计量学的交汇点

首先，关键词是语料库语言学所依赖的重要测定单位，也是文献计量学分析的基本分析单位，所不同的是，前者依靠软件的定量统计自动提取关键词；而后者的关键词主要是依靠定性手段从文献的题目、提要、层次标题和正文中提取的，但这些词往往也是文献中出现频次较高的核心词。就关键词与文本主要内容与信息的关系看,语料库语言学的关键词和文献计量学中的关键词在一定程度上具有重合性。在很大程度上，关键词既是文本的语言特征，又反映了文本的关键内容或者说主要信息。对此，杨惠中（2002：160）举例说明："……假如detective一词在参照语料库中出现的频率为1/10000,而该词在与之相比的文本中的出现频率却达到或超过了1/100，则有理由相信该文本跟detective很有关系，很可能是个侦探故事……"。

其次，从关键词统计的数据基础看，语料库语言学与文献计量学具有如下共同点：需要大量的、覆盖面广且有代表性的源数据，以保证计量分析结果的准确性、可靠性。

再次，从应用研究看，文献计量学侧重内容分析，通过对特定时期、特定领域的所有相关文献资料的基于关键词的统计来揭示该领域的研究热点、重点、新兴领域及发展趋势。从语言学的话语研究角度看，这组文献资料实质上就是特定语域的文本或话语，其构成的语料库属于学术语篇语类的专门语料库。二者在这一数据特征上存在一个交点。事实上，语言学研究领域也不乏基于学术语篇这一特定语类的专门语料库研究，包括对学术语篇中动词的类型、指称语的特征、语篇模式、评价意义的表达方式等的研究，这些研究多聚焦于学术语言的本体特征。从学科交叉的角度看，语言学研究可以从文献计量学研究获得一个启示：如果能够借鉴文献计量学的方法来延伸语言学研究，将语言特征的研究深入到内容分析，无疑可以在语言学研究的方法上实现又一个突破。

4. 语料库语言学与文献计量学方法的互补

4.1　语料库方法对文献计量学方法的补充

文献计量学研究的一个主要分析步骤是基于已经摘取的关键词进行词频统计和共词分析。这就使其分析必须面对以下问题：

一方面，文献计量学分析的对象常常是正式发表的文献资料，这些传统的引文数据最明显的局限性就是缺乏时效性，仅能提供基于正式学术交流渠道的研究信息，而无法反映非正式交流渠道或非学术交流的情况。随着网络环境下信息技术的发展，“文献”概念的外延也被扩展，网络搜索引擎和自动索引的引文数据库等都可以视为“文献”。新的检索方法和工具不仅极大地丰富了数据来源，来自网络的数据也比传统引文索引更能有效地探测出研究前沿等相关重要信息，包括最新的研究成果。

另一方面，网络资源的加入也使得传统文献计量学测量面临难题，这就是来自网络的其他非传统文献资源的关键词标引。显然，依靠人工手段对海量数据进行关键词标引[4]这一方法既不现实，也缺乏效率；而作为共词分析的分析单元，关键词的质量对共词分析的结果起着举足轻重的作用。基于主题分析的关键词标引往往存在“标引者效应”（indexer effect）（Healey *et al.* 1986，转引自Hahm *et al.* 2013）所引起的不足：标引者所使用的标引词落后于研究的现状（过时）；标引词过于泛化，不够具体，不能贴近真实的研究；标引者还可能不当甚或错误地摘取关键词，等等。这些缺陷可能产生于标引者缺乏精深的相关学科专业知识，或者是对文献具体内容的忽略。Whittaker（1989）指出，标引的结果往往不能反映研究者本人的理解。

针对网络文献关键词的标引问题，一个有效的解决办法就是关键词标引的自动化。如第一节所述，语料库自带检索软件最基本和重要的文本处理功能之一就是关键词提取。语料库方法对关键词的检索与统计，是在参照库的比照下对库中所存文本的机械化、数字化处理，得到的关键词具有统计上的显著性，有较大的准确性和客观性，因而也更具可靠性。

对英文文献而言，由于语料库检索软件一般只对单个词进行自动提取，考虑到单个的词在意义上往往并不明确，在数据库中提取关键词应经过二次检索：第一次检索得到的节点词，如果在相关研究语境下意义不准确，需要进一步考虑节点词的词串或词丛（指由两个或两个以上的词形构筑的连续的序列）或n联词（n-gram，n个连续单词组成的序列）。首次检索完毕后，根据检索结果，补齐某些节点词的其他形式，再进行二次检索，提取词串/词丛或n联词，以使分析更精准。

有两点值得注意：首先，语料库手段检索到的关键词在文本（文献）中的分

布并不均等，单纯依靠绝对频数为依据来判断研究热点和进行共词分析，结果可能不准确。为分析的科学性起见，确定关键词应该使用频数的标准值，即最终关键词标准频率的计算应为f(s)=F/N，其中，f(s)为标准化的频率，F为节点词在语料中的总频率，N为包含节点词的文本（文献）总数。其次，就频数与内容的关系看，高频关键词是主题内容的必要条件却不是充分条件。由于语料库语言学意义上的关键词提取完全以量化统计为依据，这无法确保每一个关键词都与主题内容相关。对这一部分关键词，应参照主题内容加以过滤。

文献计量学的关键词统计在一定程度上受益于语言学研究。事实上，在美国及欧洲国家的文献计量学专家中，有的就具有深厚的语言学研究背景，如苏联的纳利莫夫是语言学家，英国的布鲁克斯获得过语言学的博士学位（蒋颖2013：1）。

4.2 文献计量学方法对语料库方法的拓展

文献计量学主要是一种内容分析方法和数据挖掘方法。作为其关键测定指标的共词分析以关键词为分析单元。在文献分析中，共词分析可以揭示的信息是：当一对能够表征某一学科领域研究主题或研究方向的专业术语（一般为主题词或关键词）在一篇文献中同时出现时，这两个词之间就必定存在一定的关系；且它们越是在不同文献中有高共现频次，其关系越密切、距离越近。一般来说，关键词的高频共现意味着这两个关键词表征了具有密切联系的主要研究领域和热点问题；相反，如果关键词共现的频次较低或根本没有共现，往往意味着两个关键词无关，相关（交叉）领域的研究比较缺乏，所受关注度不够或属于研究薄弱领域。

统计一组关键词两两之间在同一篇文献中出现的频次，便可形成共词网络。共词网络的可视化图谱中，频次最高的关键词代表最核心的研究热点，通常居于网络中心，节点中心性较高，同其他关键词的共现频次也较高；而频次较低的关键词则分布于网络边缘；网络节点之间的远近反映出主题或者说研究领域之间的亲疏关系。进一步运用现代统计技术如因子分析、聚类分析和多维尺度分析等多元分析方法，可以将一个学科内的重要主题词或关键词再次归类，从而可归纳出该学科的结构、主题及研究热点。

精确的共词分析要求其分析单元——关键词——满足一些基本前提条件，这在2.2节已提及，此处不再赘述。

近40年来,共词分析法在人工智能、科学计量学、信息科学和信息系统、信息检索等领域得到了广泛应用。近年来，共词分析法也为语言研究者所瞩目，并开始应用到语言学研究中，如王立非、李琳（2014）运用共词分析和可视化技术考察了国外商务英语学术研究的现状和发展趋势；王露杨、顾明月（2014）通过共词分析展现了2000-2010年间我国语言学科学研究发展情况；刘浩（2013）运用共

词分析和可视化技术考察了国外2006-2011年间二语习得领域的研究热点和前沿；唐进（2012）运用共词分析和社会网络分析揭示了我国外语教学中自主学习的研究现状与未来发展趋势。

今天的语言学研究已远非语言本体的研究，随着人们对语言功能的认识进一步深化，语言学的研究范围和深度在不断地拓展。深度挖掘语言背后的事实，拓展语言学研究的范围，从顶层把握语言学的现状和发展趋势，需要突破语言学研究的传统思维和方法。这无疑需要依靠跨越学科界限来实现。语言学研究借鉴文献计量学的思路、方法和成果，可以有效拓展语言学传统的研究视域和维度。

5. 结论

语料库语言学研究以语言特征为基础，基于从数据中提取的特征和相关语言学理论进行意义/功能的阐释。相对于此，文献计量学方法的共词分析显示了强大的内容分析和数据挖掘功能：共词分析对包括研究热点等在内的学科研究结构信息更能作全面、精确的揭示。二者在功能设计上的特点基于各自的研究任务、特点和基本理论依据。从跨学科的视角看，在具体的研究中，可以根据需要将二者结合起来，使其各司其职，各显其能。

注释

1. 耦合（coupling）：两篇文献引用同一篇文献，这两篇文献之间具有耦合关系；两篇文献共同引用的文献越多，耦强度越大，其联系越强（蒋颖 2013：259）。

2. 同被引（co-citation）：两篇文献都被同一篇文献引用，这两篇文献具有同被引关系；两篇文献被引用的频率越高，同被引强度越大，它们之间的联系也越强（蒋颖 2013：259）。

3. 20世纪30年代，人类学家布朗（R. Brown）提出了社会网络的思想。在心理学和人类学的社会网络分析中，社会网络可以简单地看作行动者之间连接而成的关系结构。社会网络中的很多概念、方法，如中心度、声望、子群、小世界现象等都在文献计量学中被借鉴使用，甚至有些社会网络研究的常用分析软件如UCINET、Pajek等也是网络计量学和文献计量学研究中经常使用的工具。

4. 所谓标引，系指对文献和某些具有检索意义的特征如研究对象、处理方法和实验设备等进行主题分析，并利用主题词表给出主题检索标识的过程。

参考文献

Evison, J. 2010. What are the basics of analyzing a corpus? [A]. In A. O'Keeffe & M. McCarthy (eds.). *The Routledge Handbook of Corpus Linguistics* [C]. London: Routledge. 122-135.

Hahm, J., S. Kim, M. Kim & M. Song. 2013. Investigation into the existence of the indexer effect in Key phrase extraction [OL]. *Information Research* 18(4). http://www.informationr.net/ir/18-4/paper594.html (accessed 09/10/2015)

He, Q. 1999. Knowledge discovery through co-word analysis [J]. *Library Trends* 48(1): 133-159.

Leech, G. 1992. Corpora and theories of linguistic performance [A]. In J. Svartvik (ed.). *Directions in Corpus Linguistics* [C]. Berlin: Mouton de Gruyter. 125-148.

McCarthy, M. & A. O'Keeffe. 2009. What are corpora and how have they evolved?[A]. In A. O'Keeffe & M. McCarthy (eds.). *The Routledge Handbook of Corpus Linguistics* [C]. London: Routledge. 3-13.

Moon, R. 2010. What can a corpus tell us about lexis? [A]. In A. O'Keeffe & M. McCarthy (eds.). *The Routledge Handbook of Corpus Linguistics* [C]. London: Routledge. 197-211.

Scott, M. & C. Tribble. 2006. *Textual Patterns* [M]. Amsterdam: John Benjamins.

Whittaker, J. 1989. Creativity and conformity in science: Titles, keywords and co-word analysis [J]. *Social Studies of Science* 19(3): 473-496.

蒋　颖，2013，《人文社会科学领域文献计量学研究》[M]。北京：社会科学文献出版社。

刘　浩，2013，2006-2011二语习得研究热点与前沿的可视化分析 [J]，《当代外语研究》（3）：29-33。

唐　进，2012，我国外语教学中自主学习研究综述 [J]，《现代教育技术》（1）：64-69。

王立非、李　琳，2014，基于可视化技术的国外商务英语研究进展考察（2002-2012）[J]，《中国外语》（2）：88-96。

王露杨、顾明月，2014，我国语言学研究热点知识图谱分析——基于CSSCI（2000 – 2011年）[J]，《西南民族大学学报》（人文社会科学版）（6）：234-240。

杨惠中，2002，《语料库语言学导论》[M]。上海：上海外语教育出版社。

通讯地址：210098　江苏省南京市河海大学外国语学院英语系（周红英）

210039　江苏省南京市解放军国际关系学院二系（李德俊）

基于共词分析的语料库语言学研究现状分析（1971–2015）

国防科学技术大学 马晓雷 陈颖芳

提要：本研究以LLBA（Linguistics and Language Behavior Abstracts）数据库收录的23,078篇文献记录为分析对象，利用基于共词分析的聚类分析、多维尺度分析和战略坐标分析等方法，对1971年至2015年间国际语料库语言学的主要研究方向和发展状态进行分析。研究结果表明，涉及语料库的研究方向主要包括语言本体研究、自然语言处理研究、话语分析研究、历史语言学研究、社会语言学研究、词典学研究、儿童语言习得研究、二语习得研究、翻译对比研究、语篇分析研究等。其中，基于语料库的历史语言学研究、社会语言学研究、词典学研究、自然语言处理研究、语言本体研究和话语分析研究已经相对比较成熟。相比之下，语料库在儿童语言习得、二语习得、翻译对比研究和语篇分析等领域的应用还有待进一步加强。

关键词：语料库、共词分析、聚类分析、多维尺度分析、战略坐标

1. 引言

自20世纪八九十年代兴起以来（桂诗春 2014），语料库语言学已经展现出了强大的生命力。作为大规模真实语料的集合，语料库为语言研究者提供了一种新的视角和方法。经过二十多年的发展，语料库语言学研究在深度和广度上都已经有了巨大的进步。在服务语言研究的基础上，语料库在邻近学科也得到了前所未有的拓展（梁茂成 2014）。了解和把握该领域的研究现状，对于更好地推动语料库语言学的深入发展具有重要的意义。

以往国内有关语料库语言学发展现状的分析，大致可以分为三类。一类是由研究者根据经验来梳理和总结语料库语言学的历史和现状（如卫乃兴等 2014；梁茂成 2012；黄立波、王克非 2011；何中清、彭宣维 2011；李文中 1999）。另一类是由分析者根据学术期刊发表的语料库相关论文，通过人工分类来总结和归纳该领域的主要研究方向、发展态势和不足之处（如宋红波、王雪利 2013；杨梅、白楠 2010）。还有一类是借助文献计量学分析工具对期刊论文进行可视化分析（如刘霞等 2014），并在此基础上诊测语料库语言学的主要领域结构和知识热点。以上研究有助于了解语料库语言学的内涵和外延，但同时也存在一定的问题。首先，领域专家可能会受个人专长和兴趣的影响，分析得不够客观和全面。其次，虽然

以期刊论文为分析对象可以在一定程度上解决这个问题，但以往研究大都是以国内期刊为主，较少关注国际学术界的期刊论文，广度不足。再者，以往研究所涉及的期刊数量也相对有限。

针对以上问题，本研究将以Linguistics and Language Behavior Abstracts（LLBA）数据库为基础，检索并收集过去45年间与语料库相关的文献记录，并借助共词分析、聚类分析、多维尺度分析和战略坐标分析等方法，分析语料库语言学的主要研究方向和发展态势。相比以往研究，本研究的优点在于：一是关注国际学术界的语料库研究；二是收集分析的研究数据规模较大；三是借助了文献计量学常用的共词分析法，并结合可视化分析法，使分析的结果更加客观。

2. 基于共词分析的学科知识可视化

共词分析法诞生于20世纪70年代后期，由Law和Rip等法国文献计量学家提出（钟伟金、李佳 2008）。该方法以统计词与词之间的共现频次为基础，来分析词与词的亲疏远近关系，并在此基础上研判某一领域的知识结构。其基本假设是，如果两个词同时被同一篇文献使用，那么这两个词之间就可能存在着某种内在联系。如果这种联系在足够多的文献中都有体现，那么可以认为这两个词共属于同一个研究方向。借助数学、统计和信息可视化技术，就可以描述出某一领域的知识结构、研究热点和发展态势。共词分析常用的可视化方法主要包括包容图、邻近图、因子分析、聚类分析、多维尺度分析、战略坐标分析、社会网络分析等（杨颖、崔雷 2011）。

共词分析的基本流程是：（1）确定研究主题，通常是某一学科领域；（2）围绕主题收集相关文献；（3）从文献中提取关键词表，通常是由关键词、主题词或者从标题、摘要中提取出的词语组成；（4）选取一定频次以上的关键词作为共现统计的对象；（5）统计任意两个关键词在文献集合中的共现频次，并在此基础上生成共词矩阵、相关矩阵或相异矩阵；（6）利用因子分析、聚类分析、多维尺度分析、战略坐标分析、社会网络分析等手段分析关键词之间的亲疏远近关系；（7）根据可视化结果，对学科领域的主要方向、研究重点和发展态势等作出判断。

经过30多年的发展，共词分析已经成为一种分析领域知识结构的重要方法，并在自然语言处理等领域发挥着重要的作用。相比文献计量学中另一种常用的分析方法——共被引分析法，共词分析具有统计对象来源灵活、不受时间滞后影响等优点，但同时也存在着词语歧义、索引者效应（indexer effect）等问题（马晓雷 2011）。

3. 研究方法

3.1　研究数据

本研究的数据来自于ProQuest数据库中的语言学与语言行为文摘（Linguistics and Language Behavior Abstracts，简称LLBA）子库。该文摘数据库集中收录语言学及语言研究相关学科的期刊论文、著作、学位论文等相关信息。LLBA共包括1,500多种出版物，涉及语言研究的方方面面，并且每年以1万4千多条的速度增长。不同于Web of Science等数据库，LLBA主要关注语言研究的各个相关领域，因此检索出的数据对于语言研究者而言更加聚焦和全面。

本研究的检索策略是“corpus OR corpora”，时间范围不限。共得到23,078条记录，时间跨度为1971年至2015年。其中，期刊论文20,305条，著作1,482条，学位论文1,261条，会议论文30条。图1显示了检索论文在45年间的年度分布情况。需要说明的是，尽管corpus一词有多种含义，但由于LLBA数据库收录的文献均与语言研究相关，因此基本可以确定23,078条记录应全部与语料库研究相关，无需进一步筛选。

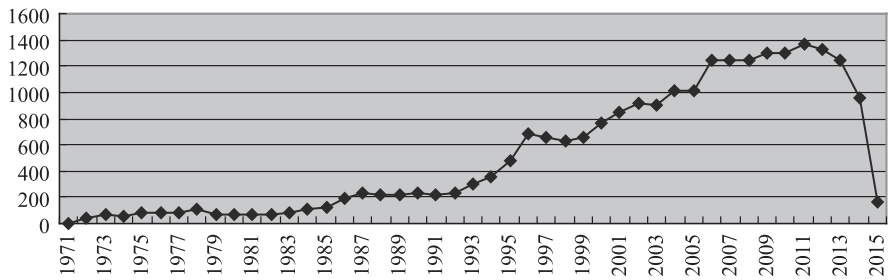


图1.　语料库研究文献的年度分布情况（1971年–2015年）

每一条记录包括的主要信息有：摘要、主题词、标题、标识符、作者、出版物名称、出版年份、文档类型等。本研究主要以主题词作为共词分析的数据来源。

3.2　研究过程

本研究主要分为以下步骤：

（1）提取23,078条记录的主题词信息。将每一条记录表征为主题词的集合。

（2）统计各主题词的频次，并从中挑选出70个频次最高的主题词作为共词分析的对象（见表1）。

由于检索数据都与语料库相关，因此虽然Corpus Linguistics、Corpus Analysis、Computerized Corpora的频次都比较高，但考虑到它们对分析语料库领域的知识结构贡献不大，我们没有将其包括在70个高频词中。同时，English、French、

Spanish、German等词的频次也比较高，但本研究也没有将它们纳入分析范围。

表1. 语料库研究文献中的70个高频主题词

频次	主题词	频次	主题词
1846	Language usage	481	Scientific technical language
1488	Discourse analysis	479	Language acquisition
1325	Text analysis	458	English as a second language learning
1206	Computational linguistics	455	Word order
1167	Language history	451	Pronouns
1072	Computer generated language analysis	446	Language change
960	Verbs	434	Discourse markers
877	Pragmatics	422	Metaphors
842	Syntactic structures	412	Meaning
834	Discourse/Text genres	404	Interpersonal communication
827	Diachronic linguistics	382	Internet
813	Lexicon	376	Syntactic analysis
803	Natural language processing	372	Syntax semantics relationship
790	Historical text analysis	371	Language culture relationship
789	Translation	365	Reference
751	Comparative linguistics	354	Child language
692	Semantic analysis	353	Error analysis
686	Statistical analysis	350	Discourse functions
651	Lexicography	349	Text structure
649	Discourse strategies	348	History of linguistics
614	Regional dialects	337	Nouns
584	Second language learning	332	Conversation analysis
581	Sociolinguistics	329	Tense
580	Semantics	329	Grammaticalization
575	Terminology	323	Voice recognition
569	Collocations	322	Lexicology
566	Oral language	322	English as a second language instruction

（待续）

（续表）

频次	主题词	频次	主题词
557	Word frequency	318	Speech acts
532	Borrowing	315	Language contact
524	Newspapers	313	Lexical semantics
520	Word meaning	313	Conversation
519	Syntax	310	Grammatical analysis
509	Stylistics	306	Machine translation
490	Computer applications	299	Registers
484	Dictionaries	293	Social factors

（3）利用自编Python程序两两统计任意两个主题词在23,078条文献记录中的共现频次，生成一个70 × 70的共词矩阵。限于篇幅，表2仅显示该矩阵的一部分。

表2.　主题词共现矩阵（部分）

	Borrowing	Child language	Collocations	Comparative linguistics	Computational linguistics
Borrowing	532	0	4	6	5
Child language	0	354	0	7	2
Collocations	4	0	569	12	35
Comparative linguistics	6	7	12	751	9
Computational linguistics	5	2	35	9	1206

该矩阵中，每一个单元格代表两个关键词的共现频次。例如Borrowing和Child language的共现频次为0，Borrowing与Collocations的共现频次为4。对角线上的数字代表某一关键词的总频次。例如Borrowing共出现了532次，Child language共出现了354次。

（4）由于聚类分析和多维尺度分析等多元统计方法对矩阵的数据形式有特殊要求（张勤、马费成 2007），因此我们将共词相关矩阵转换为相似矩阵和相异矩

阵。由相关矩阵转换为相似矩阵使用的算法是Ochiai系数，其计算公式为：

$$Ochiai\text{系数}=\frac{\text{A、B两词的共现频次}}{\sqrt{\text{A词出现的频次}}\times\sqrt{\text{B词出现的频次}}}$$

Ochiai系数介于0和1之间，0代表完全不相关，1代表完全相关。表3显示了所生成相似矩阵的一部分。

表3. 主题词相似矩阵（部分）

	Borrowing	Child language	Collocations	Comparative linguistics	Computational linguistics
Borrowing	1	0	0.01	0.009	0.006
Child language	0	1	0	0.014	0.003
Collocations	0.01	0	1	0.018	0.042
Comparative linguistics	0.01	0.014	0.02	1	0.009
Computational linguistics	0.01	0.003	0.04	0.009	1

在相似矩阵的基础上，用1与全部相关系数相减，得到表示任意两个主题词之间相异程度的相异矩阵（见表4）。

表4. 主题词相异矩阵（部分）

	Borrowing	Child language	Collocations	Comparative linguistics	Computational linguistics
Borrowing	0.00	1.00	0.99	0.99	0.99
Child language	1.00	0.00	1.00	0.99	1.00
Collocations	0.99	1.00	0.00	0.98	0.96
Comparative linguistics	0.99	0.99	0.98	0.00	0.99
Computational linguistics	0.99	1.00	0.96	0.99	0.00

（5）在相似矩阵的基础上，利用SPSS19.0的Hierarchical Cluster功能，选择Between Group Linkage算法进行聚类分析。

（6）在相异矩阵的基础上，利用SPSS19.0的Multidimensional Scaling（ALSCAL）功能，选择Euclidean Distance模型进行多维尺度分析。

（7）在聚类分析和多维尺度分析的基础上，计算各个类团内部的紧密程度和与其他类团的联系程度，并以向心度和密度为坐标轴生成语料库语言学领域的战略坐标图。

密度和相似度有多种计算方式（马费成等 2010）。本研究中，密度计算的是每个聚类类团内各个关键词之间共现频次和的平均值。向心度计算的是每个聚类类团内各个关键词与其他类团各关键词共现频次和的平均值。战略坐标的横轴（X轴）和纵轴（Y轴）分别代表向心度和密度，其原点是各聚类类团密度和向心度的平均值。各个聚类类团参照原点，分布在由X轴和Y轴区分的四个象限内。

4. 研究结果

4.1　聚类分析结果

聚类分析是一种无监督的机器学习方法，其目的在于将一组分析对象分成若干个称为类团（cluster）的子集，尽量保证每个类团内的元素之间具有尽可能大的相似性，而类与类之间具有尽可能小的相似性。图2显示了本研究聚类分析的部分结果。从图中可以较分明地看到两个类团，其中一个与话语分析相关，另一个与计算语言学相关。

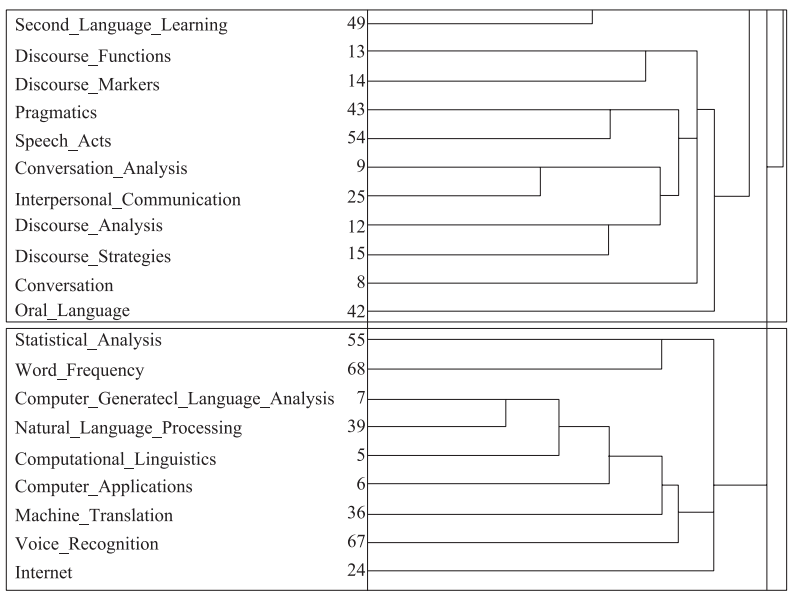


图2.　主题词聚类分析结果（部分）

通过梳理各个聚类类团内部的关键词，可以大致总结出以下涉及语料库研究的主要研究方向。按照从大到小的顺序，依次是：

（1）语言本体研究[1]。该聚类类团的关键词包括意义（meaning）、语义学（semantics）、句法学（syntax）、语义分析（semantic analysis）、句法分析（syntactic analysis）、句法结构（syntactic structures）、句法语义关系（syntax semantics relationship）、动词（verbs）、词序（word order）、时态（tense）、语法分析（grammatical analysis）、语言文化关系（language culture relationship）、隐喻（metaphor）、词汇语义学（lexical semantics）、词语意义（word meaning）、搭配（collocations）、词汇（lexicon）、词汇学（lexicology）、名词（nouns）。

（2）话语分析研究。该聚类类团的关键词包括话语功能（discourse functions）、话语标记语（discourse markers）、语用学（pragmatics）、言语行为（speech acts）、会话分析（conversation analysis）、人际交际（interpersonal communication）、话语分析（discourse analysis）、话语策略（discourse strategies）、会话（conversation）、口语（oral language）。

（3）历史语言学研究。该聚类类团的关键词包括借词（borrowing）、语言接触（language contact）、历史文本分析（historical text analysis）、语言历史（language history）、历时语言学（diachronic linguistics）、语法化（grammaticalization）、语言演变（language change）。

（4）自然语言处理研究。该聚类类团的关键词包括自然语言处理（natural language processing）、计算语言学（computational linguistics）、计算机应用（computer applications）、机器翻译（machine translation）、语音识别（voice recognition）、互联网（internet）。

（5）语篇研究。该聚类类团的关键词包括文本分析（text analysis）、文本结构（text structure）、话语文本体裁（discourse text genres）、文体学（stylistics）、新闻（newspapers）。

（6）二语习得研究。该聚类类团的关键词包括英语作为二语教学（English as a second language instruction）、英语作为二语学习（English as a second language learning）、错误分析（error analysis）、二语学习（second language learning）。

（7）方言研究。该聚类类团的关键词包括代词（pronouns）、指代（reference）、语言使用（language usage）、区域方言（regional dialects）。

（8）词典学研究。该聚类类团的关键词包括词典（dictionaries）、词典编纂（lexicography）、语言学历史（history of linguistics）。

（9）翻译、对比研究。该聚类类团的关键词包括翻译（translation）、对比语言学（comparative linguistics）。

（10）词频统计研究。该聚类类团的关键词包括统计分析（statistical analysis）和词频（word frequency）。

（11）儿童语言习得研究。该聚类类团的关键词包括儿童语言（child language）、语言习得（language acquisition）。

（12）科技术语研究。该聚类类团的关键词包括科技语言（scientific technical language）、术语学（terminology）。

（13）社会语言学研究。该聚类类团的关键词包括语域（registers）、社会语言学（sociolinguistics）。

总的来看，13个研究方向之间的结构关系可以通过图3表现出来。13个聚类既相对独立，又彼此联系。例如，从树形图上来看，翻译、对比研究与语篇研究之间的联系较为紧密，自然语言处理研究与词频统计研究关系紧密。此外，各个聚类内部也存在着一定的层次结构，例如语言本体研究又大致可以分为词汇、句法、语义三个层次。

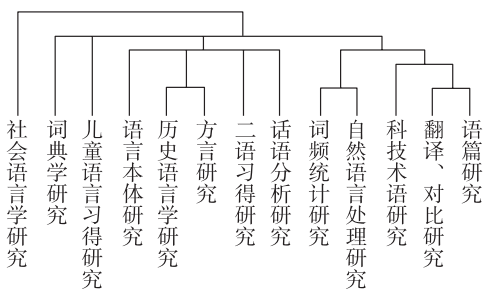


图3.　聚类类团树形图

需要说明的是，作为一种探索性的分析方法，聚类分析的结果也存在着缺乏中心概念、个别聚类难以解释等问题（杨颖、崔雷 2011）。例如，在图3中，仅从聚类间的关联程度来看，社会语言学研究与方言研究、话语分析研究之间并不直接相关，这一点就不是很容易解释。此外，聚类分析结果的解释会受研究者本人知识结构的影响，不同学者对待同样关键词组合的解读可能会有差异。例如，聚类10中包含统计分析（statistical analysis）和词频（word frequency）两个关键词。两者之间具有一定的联系，但是否需要将其单独命名为词频统计研究，可能不同的学者会有不同的看法。

4.2　多维尺度分析结果

多维尺度分析是一种根据研究对象之间的相似或相异程度，通过某种非线性变换，在二维或三维空间中展示各研究对象之间相互关系的方法。在根据共词矩

阵生成的多维尺度图谱中，词与词之间的距离代表相似性，聚集在一起的词代表某一研究方向；词在图谱中的位置代表重要性，越接近图谱中心，在学科中的地位越重要（杨颖、崔雷 2011）。图4显示了本研究多维尺度分析的结果。

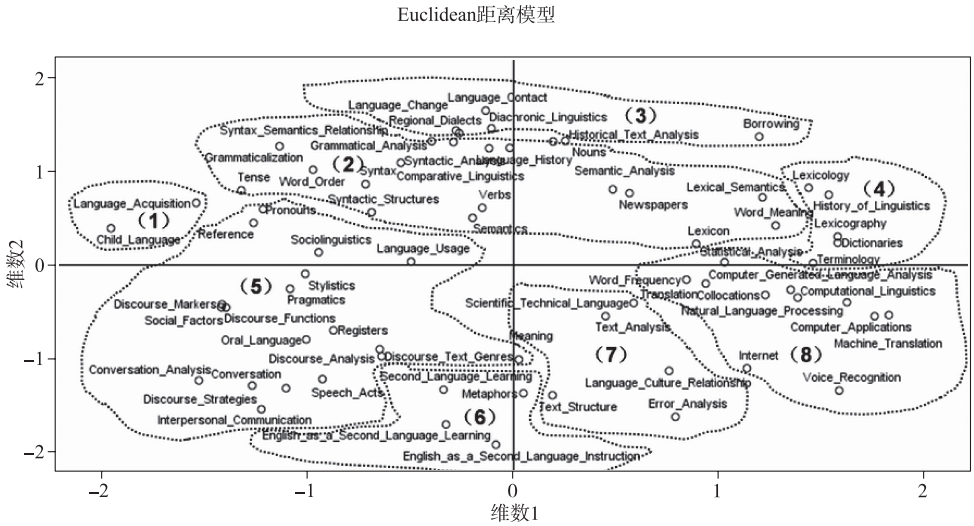


图4. 主题词多维尺度分析图谱

图4中的主题词按照分布位置，大致可以分为8个大小不等的区域，分别代表8个相对独立又彼此联系的研究方向：（1）儿童语言习得研究；（2）历史语言学研究；（3）语言本体研究；（4）词典学研究；（5）社会语言学研究；（6）二语习得研究；（7）语篇分析研究；（8）自然语言处理研究。同聚类分析的结果一样，以上区域的划分具有一定的主观性，不同研究者可能会有不同的归类方式，但总的来说，关键词之间的亲疏远近关系是有一定的合理性的。

从各个研究方向的分布位置来看，“语言本体研究”、“社会语用研究”、“语篇分析研究”和“自然语言处理研究”更接近语料库语言学领域的中心，“儿童语言习得研究”、“历史语言学研究”、“词典学研究”和“二语习得研究”相对处于学科边缘。

总的来看，多维尺度分析与聚类分析的结果是一致的，尽管两者之间也有一定的差异。相比聚类分析，多维尺度分析的结果能够反映出关键词之间的亲疏远近关系（杨颖、崔雷 2011），因此可以辅助解释聚类分析中一些不容易解读的结果。例如，在聚类分析结果中，术语学（terminology）这一主题词并没有直接与词典（dictionaries）、词典学（lexicography）等主题词直接相关，但是在多维尺度分析图谱（图4）中，以上关键词都大致分布在同一区域（图谱右上角）。再如，“语言本体研究”包含词汇、句法、语义三个层次，而其中涉及词汇研究的若干个主题词与词典学研究分布的位置非常接近，这一点也比较有说服力。

4.3　战略坐标分析结果

战略坐标分析是通过将各研究主题投射到以向心度为横坐标（X轴）、以密度为纵坐标（Y轴）的二维坐标系中，来描述各研究领域的结构和发展状态。我们主要以聚类分析的结果为基础，根据多维尺度分析的结果对各类团的关键词进行了个别调整，最终确定以下10个方向作为战略坐标分析的对象。

（1）词汇、句法、语义研究

（2）话语分析研究

（3）历史语言学研究

（4）自然语言处理研究

（5）语篇研究

（6）二语习得研究

（7）词典学研究

（8）翻译、对比研究

（9）儿童语言习得研究

（10）社会语言学研究

表5显示了各研究领域的密度、向心度，以及与战略坐标X轴和Y轴的对应位置。

表5.　各研究方向的向心度和密度

聚类名称	各聚类向心度	向心度X	各聚类密度	密度Y
词汇、句法、语义研究	710.05	-156.15	439.37	117.79
话语分析研究	671.90	-194.29	383.82	62.24
历史语言学研究	989	122.80	360.67	39.09
自然语言处理研究	756.44	-109.76	576.22	254.64
语篇研究	1035.17	168.96	304	-17.58
二语习得研究	508.75	-357.45	143.50	-178.08
词典学研究	958	91.80	334.50	12.92
翻译、对比研究	1458.50	592.30	150	-171.58
儿童语言习得研究	493.50	-372.70	158	-163.58
社会语言学研究	1080.71	214.51	365.71	44.16
均值	866.20		321.58	

图5显示的是10个研究方向的战略坐标分布图。

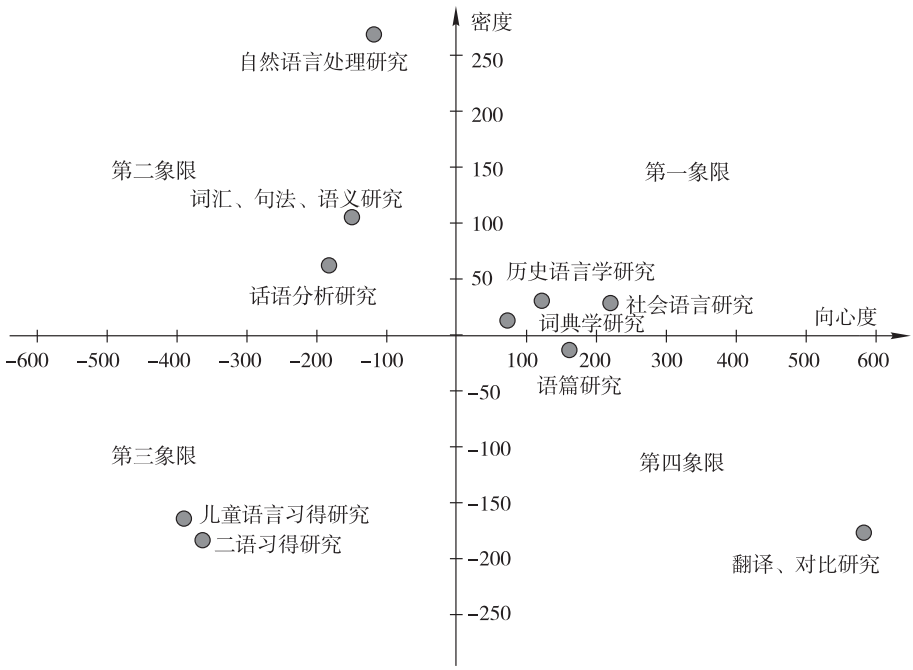


图5. 语料库研究各方向战略坐标分布图

从图5可以看出，位于第一象限的有3个研究方向，分别是历史语言学研究、社会语言学研究和词典学研究。这3个研究方向的密度和向心度都比较高，表明它们均已获得较高的关注度，同时与其他研究方向之间的联系比较紧密。处于第二象限的也有3个研究方向，分别是自然语言处理研究、词汇句法语义研究和话语分析研究。这3个研究方向具有较高的密度，表明它们已经得到较高的关注度，并且相关研究开展得比较充分。但是它们的向心度相对较低，表明这3个研究方向相对比较孤立，自成一体。位于第三象限的有两个研究方向，分别是儿童语言习得研究和二语习得研究。这两个聚类的密度和向心度都比较低，表明它们处于整个研究领域的边缘，研究相对不够成熟。位于第四象限的有两个研究方向，分别是语篇研究和翻译对比研究。这两个方向的密度较低，但向心度较高，表明它们与其他研究方向联系比较紧密，但总的来说还不是特别成熟。

需要指出的是，战略坐标分析需要以聚类分析和多维尺度分析的结果为基础。如前所述，聚类分析和多维尺度分析本身可能存在分析者主观因素造成的问题，因此对战略坐标结果的解读也可能受到影响。这也意味着有关核心度和成熟度的判断需要谨慎对待。

5. 结语

本研究以1971年至2015年的23,078条文献记录为数据来源，通过聚类分析、多维尺度分析和战略坐标分析研究了语料库研究领域的主要研究方向、各研究方向间的关系以及各研究方向的发展状态。研究结果表明，语料库在多个研究领域都有广泛的应用。在语言本体研究、自然语言处理研究、话语分析研究、历史语言学研究、社会语言学研究、词典学研究、儿童语言习得研究、二语习得研究、翻译对比研究等诸多领域，语料库都已成为了数据分析的重要来源。这充分证明了语料库的研究价值。从发展状态来看，在45年的发展过程中，语料库在历史语言学、社会语言学和词典学中的应用已经比较成熟。在自然语言处理、话语分析、语言本体研究中也已形成了较为成熟的研究范式。相比之下，语料库在儿童语言习得、二语习得等领域的应用还没有引起足够的重视，研究的力度有待于进一步加强。语料库在翻译对比研究中的应用前景广阔，但总的来看，相比其他研究方向得到的关注还比较有限。

本研究收集的数据横跨45年，但我们并没有将其划分为不同的时间段分别进行分析，将来可考虑在现有数据基础上开展历时对比研究，以期更好地反映语料库语言学的发展脉络和规律。此外还需要指出的是，聚类分析、多维尺度分析、战略坐标分析可以以可视化的方式直观地反映领域发展态势，但如果能够结合该领域专家的反思和解读，所得到的结果将更加令人信服。

注释

1. 准确来讲，该研究方向应为“涉及语料库的语言本体研究”，而不是泛指所有的语言本体研究。下同。

参考文献

桂诗春，2014，语料库语言学答客问 [J],《语料库语言学》(1)：1-15。

何中清、彭宣维，2011，英语语料库研究综述：回顾、现状与展望 [J],《外语教学》(1)：6-10。

黄立波、王克非，2011，语料库翻译学：课题与进展 [J],《外语教学与研究》(6)：911-923。

李文中，1999，语料库、学习者语料库与外语教学 [J],《外语界》(1)：51-55。

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。

梁茂成，2014，语料库、平义原则和美国法律中的诉讼证据 [J],《语料库语言学》(1)：1-15。

刘　霞、许家金、刘　磊，2014，基于CiteSpace的国内语料库语言学研究概述（1998-2013）[J],《语料库语言学》(1)：69-77。

马费成、望俊成、张于涛，2010，国内生命周期理论研究知识图谱绘制——基于战略坐标图和概念网络分析法 [J],《情报科学》(4)：481-487。

马晓雷，2011,《被引内容分析——探究领域知识结构的新方法尝试》[M]。北京：外语教学与研究出版社。

宋红波、王雪利，2013，近十年国内语料库语言学研究综述 [J],《山东外语教学》(3)：41-47。

卫乃兴、李文中、濮建忠、梁茂成、何安平，2014，变化中的语料库语言学 [J],《解放军外国语学院学报》(1)：1-9。

杨　梅、白　楠，2010，国内语料库翻译研究现状调查——基于国内学术期刊的数据分析（1993-2009）[J],《中国翻译》(6)：46-50。

杨　颖、崔　雷，2011，基于共词分析的学科结构可视化表达方法的探讨 [J],《现代情报》(1)：91-96。

张　勤、马费成，2007，国外知识管理研究范式——以共词分析为方法 [J],《管理科学学报》(6)：65-75。

钟伟金、李　佳，2008，共词分析法研究（一）——共词分析的过程与方式 [J],《情报杂志》(5)：70-72。

通讯地址：410074　湖南省长沙市国防科学技术大学人文与社会科学学院国防语言系（马晓雷）

410074　湖南省长沙市国防科学技术大学人文与社会科学学院语言文化研究所（陈颖芳）

共选视阈下的二语语用知识研究——以中国学生英语状态转变系动词为例*

扬州大学 朱 芸 陆 军

提要：本研究以共选理论为框架，以英国国家语料库BNC为参照，用基于数据的方法考察中国学习者英语中状态转变系动词BECOME、GET和GO的语用知识特征。研究发现：1）英语本族语者和中国英语学习者在使用状态转变系动词时都显现出特定的语用知识特征；2）中国学习者英语的状态转变系动词在类联接上与英语本族语趋于一致，但在语用特征上有较大偏差。分析表明：1）在共选要素中，语义韵制约着词汇和语法的选择，是高度抽象的语用知识；2）学习者不易察觉这些知识，在二语交际中倾向于协同启动母语语用知识及目的语词汇和语法知识是造成语用偏差的重要原因。

关键词：共选、状态转变系动词、语义韵、语用知识

1. 引言

二语知识涉及词汇、语法、语义和语用等方面（Larsen-Freeman 1991；Bachman & Palmer 1996）。其中，词汇和语法知识可以被直接观察和描述，一直是二语研究的热点（如Richards 1976；Nation 1990；Rieder 2003；Ellis 2004，2005；Sonbul & Schmitt 2013；Dąbrowska 2014等）。相比而言，语用特征较抽象，不易被精确描述或界定，一直是研究的难点。近年来，越来越多的研究注意到语用知识在二语能力中起着重要作用（如Bachman & Palmer 1996；Kasper 1997，2001；Ellis 1999；Bardovi-Harlig 2001；Rose 2005等）。

语用知识的重要作用与其所涵盖的内容密切相关。Leech（1983）将语用学分为语用语言学（pragmalinguisitics）和社交语用学（sociopragmatics），分别关注语法、语义的选择和词汇、语境的选择。Thomas（1995）则将语用研究抽象性地概括为交际中的意义研究。Crystal（1997：31）赋予语用学较为具体的定义，即主

*本文是2015年国家社科基金一般项目“语料库驱动的二语隐性、显性知识调查与实验研究”（15BYY067）和江苏省社科基金项目“语料库驱动的隐性、显性知识接口研究”（13YYB006）的阶段研究成果。

要从语言使用者角度研究语言，尤其关注交际中语言的选择、使用限制以及对交际参与者所产生的影响等。Bachman（1990）从语言能力测试角度进行了更为通俗的描述，将语用知识视为语言使用者知道词语或话语在具体情境中的意思及如何反映说话者的意图知识。由此说明，“语用知识”本质上就是知道如何选择合适的词汇和语法来实现特定的交际目的的知识（Bachman 1990；Bardovi-Harlig 2013），既包括语言本身的要素（如词汇和语法等），也包括语言之外的要素（如语言交际目的和效果等）。

由于语用知识的抽象性和复杂性等因素，二语语用知识研究要比词汇、语法和语音等知识的研究更加难以开展（Rose & Kasper 2001：121）。近年来，二语语用知识开始受到重视（如Bardovi-Harlig 2001；Kasper 1997，2001；Liu 2006；Rose 2005；Taguchi 2011；姜占好 2003；戴炜栋、杨仙菊 2005等）。Kasper（1996）将二语语用知识定义为“非母语的第二语言操作者在使用和习得第二语言行为时的语用体系知识和合理使用的知识”。这里的语用体系知识包括语言形式、功能意义和相关的语境特征。合理使用的知识主要指在特定语境中如何正确选择恰当的语言形式以及学习者的语言表达方式如何与本族语者所用表达方式相近的知识。由此，二语语用知识研究主要关注二语学习者在特定语境中如何合理使用目的语形式实施特定交际功能的知识（Bachman 1990；Bachman & Palmer 1996；Purpura 2004）。迄今为止，相关研究主要关注二语语用教学在二语习得中的作用（Bardovi-Harlig 2001；Rose & Kasper 2001；Kasper & Rose 2002）和二语语用能力的发展（Barron 2003；Kinginger & Belz 2005；Schauer 2006；卢加伟 2013）。这些研究倾向于讨论不同类型言语行为的语言表现和语用意图，聚焦于考察语用特征较为明显的语言形式，如话语标记等。然而，根据Leech（1983）和Crystal（1997）等对语用的界定，交际中任何语言形式的选择和组合都会涉及语用目的，词汇和语法等语言形式与语用功能的关系都是语用知识研究不可忽略的内容，甚至考察高频使用的语言形式（如词汇、语法以及互相之间的组合等）的语用知识及其影响因素具有更为重要的理论价值。

随着语料库语言学的兴起，基于大量语料证据考察语用特征已成为现实（如Hoey 2005；Adolphs 2008等）。以Sinclair为代表的语料库语言学家基于大量语料证据发现：在交际中，词汇、语法、语义和语用趋于共选（co-selection）（Sinclair 1991，1996，2004）。基于此发现，他提出了扩展意义单位模型，由词语搭配、类联接、语义选择趋向和语义韵等共选要素共同界定。其中，语义韵属“语义—语用”连续体的语用一侧（Sinclair 1996：87），主要表明说话者在特定语用情景中的态度（Louw 2000），体现了整个短语单位与语用功能的共选（Sinclair 1996；Stubbs 2009）。换言之，语义韵反映了说话者如何根据交际目的共选词汇和语法项（Morley & Partington 2009），其本质上与Crystal（1997）等对语用学的界定是一致的。

基于语料库的研究方法能系统、全面地发现语言的总体状态和特征。扩展意义单位模型集词汇、语法、语义和语用于一体（Stubbs 2009；Stewart 2010），是考察语用知识的实用操作模型，有助于揭示语言使用者如何根据交际目的选择并组合词汇和语法项。英语系动词高频出现于各类文本，在贯穿句子结构、连接主表（补）和说明主语的状况、性质、特征等方面起重要作用，是体现二语语用特征的重要材料。英语系动词可分为静态和动态两大类（Quirk *et al*. 1985），相比而言，动态系动词所实现的语用特征更为明显，更便于考察。本文拟以中国学习者英语动态系动词中的状态转变类系动词为对象，基于扩展意义单位模型，进而考察二语学习者的语用知识特征及其影响因素。

2. 研究设计

2.1 研究问题

本研究试图回答以下问题：

（1）英语本族语者使用状态转变系动词时表现出哪些语用知识特征？

（2）参照本族语者，中国学习者的英语状态转变系动词在使用中表现出哪些语用知识特征？

（3）影响学习者二语语用知识特征的因素有哪些？

2.2 语料来源

本研究以“中国英语学习者口语语料库”（COLSEC）（杨惠中、卫乃兴 2005）为考察对象，以英国国家语料库BNC的口语子库（http://corpus.byu.edu/bnc/）为参照，通过观察学习者英语中状态转变系动词所实现的语义韵来揭示二语语用知识特征。COLSEC和BNC口语子库分别代表中国英语学习者和英语本族语者的真实、即时使用的语言，旨在反映真实语言交际中的语言知识特征。

2.3 研究步骤

本文采用基于语料库数据的方法（参见卫乃兴 2002；陆军 2014），通过考察索引行语境信息确立节点词所在的意义单位的构成要素。具体步骤如下：

首先，选择节点词。本研究主要选取BECOME、GET和GO作为节点词。常用状态转变系动词包括BECOME、GET、GO、TURN、GROW等，但由于TURN、GROW在两库中的频数都太低（在COLSEC中分别为2和19，在BNC中分别为53和52），不足以揭示规律，因此本研究选择频数较高的BECOME、GET和GO（这三个系动词在COLSEC中的频数分别为295、163和21，在BNC中分别为548、2,170和870）。

其次，提取数据。采用随机抽样手段分别以BECOME、GET和GO为节点词从相应的语料库中各提取150条符合要求的索引行（提取COLSEC中所有GO的索引行），逐行观察并确定类联接，参照类联接检查和概括搭配词语义特征，根据类联接和搭配词语义特征归纳出态度意义（赞成、中立或反对）。

最后，分析讨论。以英语本族语为参照，根据类联接、语义选择趋向和语义韵特征讨论学习者语言的语用知识特征和影响因素。

3. 对比分析

3.1 类联接

索引行数据显示，系动词BECOME和GO有三种类联接，分别是：V + AP、V + NP、V + V-EN，但主要用于前两种类联接内，而GET则主要用于V + AP、V + V-EN两种类联接内（见表1）。

表1. 节点词在语料库中的类联接

节点词	类联接	BNC百分比（%）	COLSEC百分比（%）
BECOME	BECOME + AP	61	50
	BECOME + NP	30	49
	BECOME + V-EN	9	1
GET	GET + AP	50	67
	GET + V-EN	50	33
GO	GO + AP	81	88
	GO + NP	16	12
	GO + V-EN	3	0

在BNC中，BECOME + AP、BECOME + NP和BECOME + V-EN分别占61%、30%和9%，与之相比，在COLSEC中，上述型式分别约占50%、49%和1%；在BNC中，GET + AP和GET + V-EN分别占50%左右，而在COLSEC中分别约占67%和33%；在BNC中，GO + AP、GO + NP和GO + V-EN分别约占81%、16%和3%，在COLSEC中则分别约占88%、12%和0%。

以上数据说明：不同系动词的类联接分布比例差异明显；学习者英语类联接的种类与本族语偏差不大，但各类联接的具体使用比例有偏差。

3.2 语义选择趋向和语义韵特征

3.2.1 BECOME的比较

BNC中，在BECOME+AP中，节点词多与表示积极语义的词项共现，如stronger、rich、available（参见例1-3）、clear、necessary、easier、apparent、interested、pregnant、independent、important等，约占67%。BECOME与这类词共现，趋于表达“希望变得强大、变得富有和变得可用”的语义，上下文中的creative things、support and help、like you wish、improve our efficiency and effectiveness等语境信息显示BECOME具有强烈的积极语义选择趋向，实现“赞成”的态度意义。在BECOME + NP中，节点词多与中性词项搭配，如part、member、Christians等（参见例4-6），约占63%。BECOME与这类词共现，趋于表达“变成联盟中的一部分、成为一名爱尔兰共和军军人、成为基督徒”的语义，上下文语境信息显示表达“中立”的态度意义。

（1）there’re some creative things that we can do out of depression, and people can *become stronger*, they can tap into their resources inside with support and help.

（2）like you wish that you’ll *become rich* or something

（3）obviously if systems *become available,* they can improve our efficiency and effectiveness

（4）no apprentices were to *become part* of the union...

（5）they work themselves into the IRA and *become members* and then they feed the information back to British

（6）there’re a lot of new churches being built and a lot of people are *becoming Christians*

COLSEC中，在BECOME + AP的类联接内，AP主要带有积极语义特征，如例（7）-（9）中的convenient、successful、active、excited，还有important、better、useful、beautiful、modern、independent等（约占64%），BECOME与这类词共现，趋于表达“希望变得便利、成功、积极兴奋”等语义，语境信息get the information you want、my dream、love this feeling等表明BECOME具有积极语义选择趋向，BECOME + AP趋于表达“赞成”态度。在BECOME + NP中，NP主要表示美好的职业或关系，带有积极语义特征，如例（10）-（12）中的an interpreter、good friends、a famous writer，类似的还有a university student、the center、a doctor、an interpreter、a proper person、a good man、a successful person、an engineer等（约占74%），BECOME与这类词共现，趋于表达“希望成为一名翻译、好朋友、著名的作家”等语义，I hope、I’m glad、I’m dreaming of等语境信息表明，BECOME + AP倾向于表达“赞成”态度。

（7）It will *become most convenient*, whenever you start the computer, you can get the information you want.

（8）my dream is to *become successful* in my career.

（9）when I'm chatting with them, I *become more active and excited,* I love this feeling.

（10）I hope I can *become an interpreter* in the future.

（11）Finally I'm glad that we can *become good friends* after the test.

（12）I'm dreaming of *becoming a famous writer* some day.

由以上描述可以得出，BECOME + AP在两库中的语义选择趋向和语义韵特征趋同；然而，BECOME + NP在COLSEC中更倾向于积极的语义选择趋向，表达“赞成”态度，与BNC中的“中立”态度略有差异（见表2）。

表2. BECOME语义选择趋向和语义韵分布特征

语料库	类联接	语义选择趋向	百分比（%）	语义韵	百分比（%）
BNC	BECOME + AP	积极	57	赞成	67
		中性	22	中立	5
		消极	21	反对	28
	BECOME + NP	积极	25	赞成	25
		中性	64	中立	63
		消极	11	反对	12
COLSEC	BECOME + AP	积极	62	赞成	64
		中性	14	中立	9
		消极	25	反对	28
	BECOME + NP	积极	62	赞成	74
		中性	22	中立	10
		消极	16	反对	16

注：由于BECOME + V-EN在BNC中只检索到49例，在COLSEC中只检索到2例，不足以说明其语义韵，故未列于表中。对于中性的搭配词项，笔者逐项观察其在语境中的意义，明确其“赞成”、“中立”抑或“反对”的态度，表3、表4亦同。

3.2.2 GET的比较

GET在BNC和COLSEC中都大量出现。在BNC中，无论是在GET + AP中还是在GET + V-EN中，GET都倾向于与消极词项搭配。其中AP有：worse、bored、stuck（参见例13-15）、older、drunk、confused、pissed、depressed、annoyed、frustrated等，约占88%；V-EN有：killed、sacked、caught（参见例16-18）、thrown、kicked、fined、blown、bogged、flooded、trapped等，约占81%。GET与这些词搭配，表示“变得糟糕或陷入某种困境”等，像I’m afraid、tears、damn、bloody awful day、nightmare、it’s a shame、dangerous等语境信息表明GET在GET + AP和GET + V-EN中有消极的语义选择趋向，趋于表达强烈的“反对”态度。表示积极意义的词项不多，主要是固定搭配：ready、married、paid。

（13）that sort of problem will only *get worse*, I’m afraid so

（14）I *get bored* to tears with doing this every damn day!

（15）Why, cos I *got stuck* there, it was a bloody awful day if you ask.

（16）The Duke of Kent *got killed* in a plane crash. What a nightmare!

（17）I mean it’s a shame somebody’s *getting sacked* because of him.

（18）the main roads are little bit more dangerous as far as *getting caught* is concerned

在COLSEC中，GET多数与AP共现，其右搭配词多数为积极语义，有：better、skilled、popular（参见例19-21）、used、ready、higher、familiar、closer、more important、successful、rich、big等，约占80%。GET与这类词搭配，表示美好的趋势或结果，显示“赞成”态度的语义韵。在GET+V-EN中，高频右搭配词有两个：married、prepared（例22、23），表示“结婚”、“准备好”的语义，语境信息I think、a better way表明GET趋于表达“赞成”态度。其余搭配词频数低于2，且多数仍为积极词项，如：obtained（例24）、succeeded、passed、well-paid、improved、increased、started等（约占76%），表达“赞成”态度，也有个别消极搭配，如：unemployed、lost、drunk、polluted、punished等，但都只出现了一次。

（19）if he works hard, he’ll *get better.*

（20）the more they teach, the more they *get skilled*.

（21）I think it will *get more and more popular* in the near future.

（22）before I *get married*, I think I will live with my parents.

（23）I think recreation is a better way to *get prepared* for my study.

（24）We can *get obtained* too much nutrition from food.

上述表明，GET的语义韵冲突明显，中国学习者对于其消极的语义选择趋向

和“反对”的语义韵基本没有察觉，反而将GET与积极语义高频共现，表达“赞成”态度（见表3）。

表3. GET语义选择趋向和语义韵分布特征

语料库	类联接	语义选择趋向	百分比（%）	语义韵	百分比（%）
BNC	GET +AP	积极	10	赞成	10
		中性	4	中立	2
		消极	86	反对	88
	GET + V-EN	积极	15	赞成	15
		中性	10	中立	4
		消极	75	反对	81
COLSEC	GET + AP	积极	80	赞成	80
		中性	1	中立	0
		消极	19	反对	20
	GET + V-EN	积极	70	赞成	76
		中性	7	中立	0
		消极	23	反对	24

3.2.3 GO的比较

GO在BNC中大量出现，无论是在GO + AP中还是在GO + NP中，都趋于与消极词项搭配，如：hungry、mad、deaf、bust、bang、ape（参见例25-30），高频出现的AP和NP如wrong、cold、missing、bankrupt、bad、slow、white、red、bald、berserk等，分别约占83%和84%。GO与这类词共现，趋于表达“变得饥饿、疯狂、聋了、破产、巨响、傻瓜”之意，poverty、bad、fucking、pain、inadequate funding、don’t know why、ugly等语境信息表明GO显示极强的消极语义选择趋向，传递强烈的“反对”态度。

（25）Barnardos says that poverty in Britain is so bad that parents are *going hungry*.

（26）I’m gonna fucking *go mad.* What’s this thing for?

（27）For just a minute I *gone deaf.* All this pain inside...

（28）last week, the National Association of Private Residential Homes were saying that they will be *going bust* as a result of inadequate funding.

（29）I don't know why, it just *goes bang* and shouting so much.

（30）you *went ape* on me. You deserved it, you are ugly.

在COLSEC中仅检索到21例GO的例句，但有效例句仅16条，其中14句为GO + AP的类联接，除了mad、wrong、online（例31）为消极语义选择趋向，传递“反对”态度外，其余全都是积极语义趋向，有：deep（例32）、higher（例33）、online（例34）、deeper、hand-in-hand，GO与这类词共现，趋于表达“变得深入、职位更高、可以上网”等语义，specialize in、with education、develop、explore等语境信息显示GO带有积极语义趋向，表达“赞成”态度。GO + NP的类联接仅两句（例35、36），且都与the middle way搭配。I choose、my opinion这样的语境信息表明GO显示积极语义选择趋向，传递“赞成”态度。

（31）Many people just *go online* to chat with each other, not to get some useful information.

（32）You tend to specialize in that and would *go deep* in that

（33）we can know that with education female graduates have *gone higher*.

（34）Computer is developed more and more, we can always *go online* and explore in the internet.

（35）I choose to *go the middle way* to spend every penny in their pocket.

（36）My opinion is just *go the middle way* between those two opinions.

上述表明，中国学习者对于GO的正确使用仅局限于固定搭配（go mad、go wrong），其余则表现出特有的二语共选特征，对于其在本族语中的消极语义选择趋向和“反对”态度的语义韵没有察觉，更没有运用（见表4）。

表4. GO语义选择趋向和语义韵分布特征

语料库	类联接	语义选择趋向	百分比（%）	语义韵	百分比（%）
BNC	GO + AP	积极	14	赞成	14
		中性	5	中立	3
		消极	81	反对	83
	GO + NP	积极	13	赞成	13
		中性	10	中立	3
		消极	77	反对	84

（待续）

（续表）

语料库	类联接	语义选择趋向	百分比（%）	语义韵	百分比（%）
COLSEC	GO + AP（仅14例）	积极	72	赞成	79
		中性	14	中立	0
		消极	14	反对	21
	GO + NP（仅2例）	积极	0	赞成	100
		中性	100	中立	0
		消极	0	反对	0

4. 讨论

4.1 英语本族语中状态转变系动词的语用知识特征

BNC数据表明，英语本族语中状态转变系动词BECOME、GET和GO都倾向于与特定语法范畴中具有特定语义特征的搭配词共选，实现不同的语用功能。但不同的状态转变系动词在使用频率、类联接分布、语义选择趋向和语义韵上都存在差异。例如，三个系动词中，GET使用频率最高；V + AP型式都高频出现，但所占比例差异明显，BECOME + AP、GET + AP和GO + AP分别占相应索引行的61%、50%和81%；V + V-EN也是三个系动词共有的型式，但只有GET + V-EN高频出现（约占50%），BECOME + V-EN和GO + V-EN仅分别约占9%和3%。不同系动词在相同类联接型式中的语义选择趋向和语义韵差异也很明显，BECOME + AP是积极的语义选择趋向，表达“赞成”态度，而GET + AP和GO + AP则趋于消极，表达“反对”态度。同一节点词在不同类联接型式中的语义趋向和语义韵也不尽相同，如BECOME + AP趋于实现“赞成”的态度意义，BECOME + NP则趋于“中立”的态度意义。

由此可见，尽管从表面上看BECOME、GET和GO都是系动词，也具有相似的语法型式，表示“变得、成为”之义，但用于实现不同程度的“赞成”、“中立”和“反对”态度，英语本族语者能够根据交际目的选择不同状态转变系动词在不同型式中实现不同的语用功能。

4.2 学习者英语状态转变系动词的语用知识特征

COLSEC数据表明，学习者英语与英语本族语在语法型式上偏差不大，但在分布比例上有差异。V + AP都高频出现，BECOME + AP、GET + AP和GO + AP分别占相应索引行的50%、67%和88%，学习者使用BECOME + AP的比例高于本族语，而使用GET + AP的比例低于本族语。V + V-EN是三个系动词共有的型式，

但只有GET + V-EN少量出现（约33%），且少于本族语（约50%），而学习者几乎不会选择使用BECOME + V-EN和GO + V-EN的型式（仅约1%和0%）。

在学习者英语中，状态转变系动词也倾向于与具有特定语法范畴、有特定语义特征的搭配词共选，实现特定的语用功能，但与英语本族语在语用特征上差异明显。差异主要体现在以下几点：（1）不同系动词在相同类联接型式中的语义选择趋向和语义韵趋于一致，即BECOME + AP、GET + AP和GO + AP都是积极的语义选择趋向，表达“赞成”态度，BECOME + NP和GO + NP也都是积极的语义选择趋向，表达“赞成”态度；（2）同一节点词的不同类联接型式在语义趋向和语义韵上的差异也不如本族语明显，即BECOME + AP和BECOME + NP，GET + AP和GET + V-EN，GO + AP和GO + NP全都趋于积极，实现“赞成”的态度意义；（3）学习者英语与本族语的差异因节点而异，其中，BECOME的差异相对较小，GO的显著差异在于使用频率极低，GET的语用偏差最大。进一步探究发现学习者趋于受get ready、get married这样的搭配影响，在二语交际中，大量使用类似积极共现（如get successful、get improved等），呈现显著的“赞成”态度的语用特征。

由此可见：（1）除少数相对固定的词语组合外，中国学习者趋于将这三个系动词视为可以替换的同义词，没有区分出其不同的语用特征；（2）中国学习者只孤立地掌握系动词的语义，并不懂得将其放在整个扩展意义单位中根据类联接的不同区分不同的语用功能；（3）当母语词汇和英语词汇在语义、语法、语用上完全对等时，如BECOME，不易产生语用偏差，反之，当英语词汇对应多个语义，与母语不完全对等时，如GET，则容易产生偏差。对于这类词，学习者趋于参照业已习得的词汇的语用知识在同一范畴内进行扩展、延伸，形成众多类似的语用特征。

4.3 学习者二语语用知识的形成原因

第一，语义韵是高度抽象的语用知识，在二语交际中制约着词汇和语法的选择。本族语中，同一节点词可能有多种类联接，且在具体类联接中各节点词都倾向于与具有特定语义特征的搭配词共现构筑具体的语义韵（Sinclair 1996, 2004；卫乃兴 2002；陆军 2012），实现不同的语用功能。中国学习者在使用BECOME、GET和GO时，在类联接型式和状态转变语义上与英语本族语者并无偏差，但在语义韵上偏差较大。由此推断，学习者注意到这些状态转变系动词的语法型式和语义特征，但并没有注意到它们在语用层面上的差异。根据Schmidt（1993）的观点，话语的语用功能及其相关语境因素非常复杂，对于学习者来说并不十分凸显，即使接触很长时间也不一定会被意识到。节点词所在的整个扩展意义单位中，搭配和类联接属于相对具体的语义、语法知识，而语义选择趋向和语义韵属于语用知识，尤其语义韵是高度抽象的语用知识，不易被学习者察觉。

交际中语言型式的选择总是从话题及对话题的态度出发（Morley & Partington

2009）。语义韵是话语选择的依据，在共选和限制关系中起决定性作用。“说话者或作者在选择一个多词单位时，不仅要考虑一个词相邻位置的词语和语法共选关系，还要考虑到更远的语义选择趋向和相应的语用关系即语义韵”（Tognini-Bonelli 2001：111）。本研究中，本族语者的状态转变系动词的语用知识特征表明本族语者之所以选择某些状态转变系动词型式而不选择其他，主要由所要表达的语用意图决定。语义韵表达态度意义，是实现语用意图的重要方式，制约着状态转变系动词及其搭配词和语法型式的选择。

第二，二语交际中，学习者趋于协同启动母语语用知识及目的语词汇和语法知识，母语语用知识对二语交际中词汇和语法选择产生干扰，从而造成语用偏差。交际过程中，通常是先有交际意图，若此意图经由单一的语用知识传递，很容易达到成功传递的目的。然而，正如Hoey所讲，“在没有浸透式（immersion）学习英语的情况下，二语词汇的启动（priming）必然附加到母语词汇的启动上”（2005：183），这样才能完成交际任务。国外关于语际语用知识的研究已证明，二语学习者语际语用知识与能力的发展常常以母语的语用知识为基础（Kasper & Schmidt 1996；Kasper & Rose 1999）。目的语的语义、语法知识和母语的语用知识并存，二者都被调动起来协同启动，反复实践，逐渐形成了学习者特有的语用知识。若母语的语义韵与目的语的语义韵一致，例如在BECOME+AP的类联接中，则不易产生语用偏差。但是，当母语的语义韵与目的语的语义韵不一致，例如在BECOME + NP的类联接中，或是甚至相反时，例如在GET + AP、GET + V-EN、GO + AP和GO + NP中，则会产生语义韵冲突，造成语用偏差。GET所反映出的学习者趋于参照业已习得的词汇的语用知识在同一范畴内扩展、延伸并形成众多类似的语用特征，是学习者“在启动词汇时潜意识地期待或创造相同的语境”（Hoey 2005：11），其实质是母语中的惯常语用表达与已习得的目的语词汇的语义、语法知识的协同启动。系动词的语用偏差足以证明，母语语用知识对二语交际中词汇和语法的选择产生干扰是造成语用偏差的重要因素。

4.4 对二语教学的启示

系动词的语用偏差证明：在二语教学中，容易习得的语义和型式的传授一直是课堂教学的主要内容，而与语义韵相关的语用知识则被长期忽视。国内的英语教学过度重视单词数量的积累，对词汇的认知也总是强调其音形义的简单记忆，疏漏了语用知识的拓展。本研究所揭示的二语语用知识特征及形成原因表明：语用知识习得具有重要意义，应与语义、语法的习得享有同等地位；交际中的语言是扩展意义单位模型中各要素协同启动的结果，涉及多个层面的共选，因此，二语教学应探讨如何让学习者浸入真实语境，反复接触并体会词语的语义、语法和语用的共选关系，从而掌握词语全面、地道的用法。

5. 结语

本研究以共选理论为框架，在建立类联接的基础上，基于数据对比分析了状态转变系动词BECOME、GET和GO的语用知识特征。研究发现：中国学习者在系动词的使用上有语用偏差，语义韵是高度抽象的语用知识，制约着词汇和语法的选择，学习者协同启动母语语用知识和目的语词汇、语法知识是导致偏差的重要原因。本研究基于共选理论，并利用语义韵探究二语语用知识特征，是一次全新的尝试，所揭示的形成原因为二语语用知识研究提供了崭新的视角。同时对二语教学亦有启示意义：教师在关注学习者词汇数量不断扩大的同时，除了重视搭配、类联接等语义、语法知识，还应重视语义选择趋向，更重要的是，在传授语义韵等语用知识时，应引导学生将词汇置于整个扩展意义单位中，通过大量范例的输入，发现词汇的语义韵，从而提高学习者语用能力以符合交际能力发展的需要。

参考文献

Adolphs, S. 2008. *Corpus and Context: Investigating Pragmatic Functions in Spoken Discourse* [M]. Amsterdam: John Benjamins.

Bachman, L. & A. Palmer. 1996. *Language Testing in Practice* [M]. New York: OUP.

Bachman, L. 1990. *Fundamental Considerations in Language Testing* [M]. Oxford: OUP.

Bardovi-Harlig, K. 2001. Evaluating the empirical evidence: Grounds for instruction in pragmatics [A]. In K. Rose & G. Kasper (eds.). *Pragmatics in Language Teaching* [C]. Cambridge: CUP. 13-32.

Bardovi-Harlig, K. 2013. Developing L2 pragmatics [J]. *Language Learning* 63(S1): 68-86.

Barron, A. 2003. *Acquisition in Interlanguage Pragmatics: Learning How to Do Things with Words in a Study Abroad Context* [M]. Amsterdam: John Benjamins.

Crystal, D. 1997. *A Dictionary of Linguistics and Phonetics (4th edition)* [M]. Cambridge, MA.: Blackwell.

Dąbrowska, E. 2014. Implicit lexical knowledge [J]. *Linguistics* 52(1): 205-223.

Ellis, R. 1999. *The Study of Second Language Acquisition* [M]. Shanghai: Shanghai Foreign Language Education Press.

Ellis, R. 2004. The definition and measurement of L2 explicit knowledge [J]. *Language Learning* 54(2): 227-275.

Ellis, R. 2005. Measuring implicit and explicit knowledge of a second language: A psychometric study [J]. *Studies in Second Language Acquisition* 54(2):141-172.

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

Kasper, G. 1996. Interlanguage pragmatics in SLA [J]. *Studies in Second Language Acquisition* 18(2):145-148.

Kasper, G. 1997. The role of pragmatics in language education [A]. In K. Bardovi-Harlig & B. Hartford (eds.). *Beyond Methods: Components of L2 Education* [C]. New York: McGraw-Hill. 113-136.

Kasper, G. 2001. Four perspectives on L2 pragmatic development [J]. *Applied Linguistics* 22(4): 502-530.

Kasper, G. & K. Rose. 1999. Pragmatics and SLA [J]. *Annual Review of Applied Linguistics* 19: 81-104.

Kasper, G. & K. Rose. 2002. *Pragmatic Development in a Second Language* [M]. Oxford: Blackwell.

Kasper, G. & R. Schmidt. 1996. Developmental issues in interlanguage pragmatics [J]. *Studies in Second Language Acquisition* 18(2): 149-169.

Kinginger, C. & J. Belz. 2005. Sociocultural perspectives on pragmatic development in foreign language learning: Microgenetic case studies from telecollaboration and residence broad [J]. *Intercultural Pragmatics* 2(4): 369-422.

Larsen-Freeman, D. 1991. *An Introduction to Second Language Acquisition Research* [M]. London: Longman.

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

Liu, J. 2006. *Measuring Interlanguage Pragmatic Knowledge of EFL Learners* [M]. Frankfurt am Main: Peter Lang.

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.

Nation, I. 1990. *Teaching and Learning Vocabulary* [M]. New York: Newbury House.

Purpura, J. 2004. *Assessing Grammar* [M]. Cambridge: CUP.

Quirk, R., S. Greenbaum, G. Leech & J. Svartvik. 1985. *A Comprehensive Grammar of the English Language* [M]. London: Longman.

Reider, A. 2003. Implicit and explicit learning in incidental vocabulary acquisition [J]. *VIEWS* 12(1): 24-39.

Richards, J. 1976. The role of vocabulary teaching [J]. *TESOL Quarterly* 10(1):77-89.

Rose, K. 2005. On the effects of instruction in second language pragmatics [J]. *System* 33(3): 385-399.

Rose, K. & G. Kasper. 2001. (eds.). *Pragmatics in Language Teaching* [C]. Cambridge: CUP.

Schauer, G. 2006. The development of ESL learners' pragmatic competence: A longitudinal investigation of awareness and production [A]. In K. Bardovi-Harlig, C. Félix-Brasdefer & A. Omar (eds.). *Pragmatics and Language Learning* [C]. Honolulu: University of Hawaii Press. 135-163.

Schmidt, R. 1993. Consciousness, learning and interlanguage pragmatics [A]. In G. Kasper & S. Blum-Kulka (eds.). *Interlanguage Pragmatics* [C]. New York: OUP. 21-42.

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

Sinclair, J. 1996. The search for units of meaning [J]. *Textus IX*: 75-106.

Sinclair, J. 2004. *Trust The Text* [M]. London: Routledge.

Sonbul, S. & N. Schmitt. 2013. Explicit and implicit lexical knowledge: Acquisition of collocations under different input conditions [J]. *Language Learning* 63(1): 121-159.

Stewart, D. 2010. *Semantic Prosody: A Critical Evaluation* [M]. London: Routledge.

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

Taguchi, N. 2011. Teaching pragmatics: Trends and issues [J]. *Annual Review of Applied Linguistics* 31(6): 289-310.

Thomas, J. 1995. *Meaning in Interaction: An Introduction to Pragmatics* [M]. London: Longman.

Tognini-Bonelli, E. 2001. *Corpus Linguistics at Work* [M]. Amsterdam: John Benjamins.

戴炜栋、杨仙菊，2005，第二语言语用习得的课堂教学模式 [J],《外语界》(1)：2-8。

姜占好，2003，中介语语用学研究及其对提高学生语用能力的启示 [J],《山东外语教学》(2)：64-67。

卢加伟，2013，认知框架下的课堂语用教学对学习者二语语用能力发展的作用 [J],《解放军外国语学院学报》(1)：67-71。

陆　军，2012，共选理论视角下的学习者英语型式构成特征研究 [J],《现代外语》(1)：70-78。

陆　军，2014，语义韵研究的理论、方法与应用 [J],《语料库语言学》(1)：58-68。

卫乃兴，2002，语义韵研究的一般方法 [J],《外语教学与研究》(4)：300-307。

杨惠中、卫乃兴，2005,《中国学习者英语口语语料库建设与研究》[M]。上海外语教育出版社。

通讯地址：225000　江苏省扬州市扬州大学外国语学院

学习者语法错误自动检查研究述评*

对外经济贸易大学 陈 功

提要：对学习者英语中语法错误的自动检测是计算语言学研究领域的一个重要课题，其进一步研究既需要理论层面的发展，也需要研究方法上的突破。本文对学习者语法错误自动检查研究领域的发展与现状进行了全面综述，旨在为中国英语学习者语法检查研究提供新的视角和思考。文章首先阐述了学习者语法错误的特殊性，然后对已有研究中专门针对学习者语法错误的自动语法检查研究进行了回顾和评述，总结了其中可资借鉴之处，同时也指出了研究中存在的问题。最后，本文概括了四点启示，希望对今后的学习者错误自动检查研究有所启发。

关键词：英语学习者、语法错误、自动语法检查

1. 引言

不论是本族语者，还是学习者，写作中出现错误都在所难免，这就为语言校对和语法检查工具带来了大批用户。不过，随着研究的进一步深入，一些问题也随之出现。其中一个主要问题就是：现有语法检查工具大多是以本族语者为目标用户设计而成的（Gamon, *et al*. 2009），在很大程度上无法满足学习者的需求。例如，我们最常用的Microsoft Word语法检查功能，最初就是为英语本族语者设计的（Gamon, *et al*. 2009；Leacock, *et al*. 2010：7），若将其用于学习者语法检查，结果必然不理想，因为学习者语法错误在很大程度上不同于本族语者错误。因此，要想开发一个适合学习者的语法检查系统，首先需要了解学习者语法错误的特殊性，并对相关的错误检查研究做一个全面的回顾和总结，力图发现问题，为今后的研究提供启示。

2. 学习者语法错误及其自动检查研究

2.1 学习者语法错误的特殊性

已有研究表明，学习者的语法错误在很大程度上不同于本族语者。

* 本研究为教育部人文社科青年项目（14YJC740006）、教育部人文社科重点研究基地重大研究项目（11JJD740011）、对外经济贸易大学校级科研课题（12QD16）的阶段性成果。

首先，英语学习者和英语本族语者所表现出来的错误模式非常不同。Connors & Lunsford（1988）对美国大学生语法错误的类型进行了调查。通过对3,000多篇作文的分析，他们整理出了美国大学生作文中出现频率最高的20种错误。之后，Donahue（2001）（转引自Leacock, *et al*. 2010：15）采用Connors & Lunsford（1988）的错误分类对英语学习者作文进行了错误分析。他们随机抽取了200篇学习者英语测试作文进行分析，发现英语学习者和本族语者作文所表现出来的错误模式非常不同（见表1）。尽管这两个研究所采用的错误分类还有待进一步商榷，但是研究所反映出来的现象却值得关注。

对上述研究中的前十种错误类型进行观察（见表1），我们可以发现，本族语者所犯错误较为简单，多为机械性错误，如，标点符号错误或某些具体词汇的错误；而学习者除了会犯本族语者常犯的错误之外，还可能受英语水平或母语迁移的影响而出现其他方面的问题，如，语法规则错误、词汇用法错误等。Chan（2010）对香港学生写作错误的调查也证实了这一点，即学习者写作中的错误形形色色，可能出现在词汇层面、句法层面，甚至语篇层面。

表1　本族语者错误和学习者错误的比较

	Connors & Lunsford（1988）	**Donahue（2001）**
	本族语者	**英语学习者**
1	No comma after introductory element	Comma splice: Two sentences joined by a comma instead of a conjunction
2	Vague pronoun reference	Wrong word
3	No comma in compound sentence	Missing words
4	Wrong word	Wrong tense or verb form
5	No comma in nonrestrictive element	Wrong or missing preposition
6	Wrong or missing inflected end	Wrong or missing inflected ends
7	Wrong or missing preposition	Sentence fragment
8	Comma splice: Two sentences joined by a comma instead of a conjunction	Run on, fused sentences
9	Possessive apostrophe error	Capitalization
10	Tense shift	Wrong verb form

其次，学习者语法错误的情况较为复杂。Kann（2002）认为，学习者的错误往往是多个错误的集合体，同一句中可能出现多个语法错误。学习者语言中不仅

错误类型多，而且可能发生于语言的各个层面，有些错误甚至连人工都难以判定；而本族语者的语法错误往往是孤立的（isolated），就像大海中偶见的岛屿一样，是可以预知的。尽管Kann的比喻略有夸张，但却真实反映了学习者写作中存在的问题，以及现有语法检查系统与实际需求之间的差距，同时也为学习者语法检查系统的研究提出了更高的要求。

最后，在学习者语法错误中，开放性词类错误较多，对语言质量的影响较大。Leacock, *et al.*（2010：17）指出，在“剑桥大学出版社学习者语料库”（Cambridge University Press Learners Corpus，简称CLC）所标注的所有错误中，除拼写错误以外，最常见的错误就是开放性词类（名词、动词、形容词和副词）的错误使用。Leacock & Chodorow（2003）的研究也证明，学习者开放性词类错误不仅数量多，而且严重程度高。研究还发现，在系统检测出的所有错误当中，对作文成绩预测力最强的错误多是开放性词类错误，如主谓一致错误、不定式小句错误（...able *to began/begin a family），以及分词错误（their parents *are expect/ expecting good grades）等。另外，根据Chan（2010）的错误调查可以发现，在中国香港学生的作文中，至少有50%以上的语法错误为开放性词类错误。而英语本族语者由于对其母语具有非常好的语言直觉（Tschichold 2003），通常只会犯一些不太严重的操作性失误。

2.2 学习者语法错误自动检查研究

考虑到学习者语言错误的特殊性及其语言学习的需求，不少研究者对学习者语法错误自动检查进行了专门探讨，并开发出了专门针对学习者的语法检查系统[1]。根据各研究声称的所能检查语法错误的覆盖范围，本研究将现有的学习者语法检查工具[2]大致分为两类：通用型和专用型。通用型语法检查系统的目标是，查找学习者作文中多种类型[3]的语法错误；而专用型语法检查系统指的是，专门用于查找某一类语法错误的系统，例如，冠词错误、介词错误、动词形式错误等。本文将从目标错误、目标用户、实现方法、查错性能等方面对已有研究进行总结回顾[4]，旨在对学习者语法检查研究现状有一个较为全面的认识。

2.2.1 通用型语法检查系统

一直以来，研究者们都在不断尝试，并试图研制能够检查多个，甚至全部目标错误的语法检查系统，并取得了较大的进展。下文将对部分研究进行评述。

Huang, *et al.*（2011）所设计的EdIt被作者称作是“一个普查型的语法检查系统”（a broad-coverage grammar checker）。尽管作者声称该系统对目标错误的覆盖范围较广，但在文中仅给出了三种错误类型的检查示例，而且基本集中在动词形式错误上。另外，值得关注的是，该系统在获取语言规则时，声称采用了型式语法（Pattern Grammar）的理念，并称所有的规则为“型式规则”（pattern rules）。

但是，仔细阅读之后，笔者发现，Huang, *et al.*（2011）研究中所表现出来的只是对型式语法狭义的理解，EdIt所基于的规则本质上只是通过统计手段提取出的共现程度较高的单词和/或词性赋码串，例如，play ~ role in Noun、he plays DET、look forward to V-ing等，与型式语法本身的理论主张（即词汇和语法不可分；型式和意义相联系）不尽相同。确切来说，应该是“基于实例”的语法检查。此外，该研究的实现方法主要是基于正确的模式匹配，即利用上述具体规则与输入文本进行匹配。这一方法在语法检查中虽有一定优势，但是却有着模式匹配无法克服的问题，有相当一部分的语法错误无法顾及。如果该系统在实现方法上不做补充，很难称得上是一个“普查型”的语法检查系统。

Lawley（2003）的语法检查系统也是基于模式匹配的。不过，与Huang, *et al.*（2011）的方法不同的是，Lawley（2003）采用的是错误实例，即从学习者作文中提取出错误词串（incorrect sequences），构成错误模式数据库。尽管这也是一个很好的尝试，但是从测试结果来看，系统表现却不尽如人意，模式匹配查错准确率高的优势并没有体现出来。笔者认为主要是三大原因造成的：（1）单纯的模式匹配一例一错，无法穷尽所有错误，若输入文本中的错误未包含在规则中，则无法查出；（2）错误词串匹配很容易导致误判。例如，该研究将on autumn认定为错误之一，而实际输入文本中若出现on autumn nights则也会被标记为错误；（3）该方法只能检查语言的线性错误，而且对部分线性错误的检查效率不高，如，主谓一致问题。从这三个方面来看，如果按照该研究的方法操作，必然会有相当一部分语法错误无法检查出来。

Brehony（1993）为法国的英语学习者设计的语法检查系统FSGC是在链语法分析器的基础上改造而成的，主要解决的是母语迁移导致的11个英语语法错误。为了对这些错误进行准确查找，研究者采用了两种办法：一是添加错误的链接规则；二是对原正确规则进行约束松弛，从而使链语法分析器能够处理含有目标错误的句子，并将错误识别出来。该研究对链语法分析器规则的改编是一个非常有益的尝试。但是，该研究对链语法规则的改动，并不是对原有正确语法规则的修订、扩充或细化，因此，并不能提高链语法分析器本身的分析能力，也无法处理更多的语法错误。

在通用型语法检查系统当中，基于统计方法的系统较少。Chodorow & Leacock（2000），Leacock & Chodorow（2001，2003）开发的语法错误自动检查系统ALEK（Assessment of Lexical Knowledge）采用的就是基于统计的方法。ALEK的训练语料来自北美报纸，语料规模约为3千万词；系统通过计算相邻两词或词性赋码之间的互信息值，构建起二元组（bigram）语言模型，作为评判输入文本是否包含语法错误的依据。不过，值得注意的问题是，该研究的训练语料为新闻文体，与学习者作文文体差异较大，必然会对检查结果产生影响。另外，该研究只考虑了

二元组，对于远距离关系上的错误则显得无能为力。此外，有的研究开始利用网络资源对学习者错误进行检查，More（2006）就试图利用网络搜索引擎来帮助判断输入文本是否合乎语法。该研究一方面利用了网络文本资源，另一方面则发挥了搜索引擎的统计功能（web-counts），是对传统语法检查系统研究的很好补充。不过，该研究所利用的网络文本属于充满大量噪音的文本，可能包含各种各样的错误，因此，将会极大地影响系统查错的准确率和召回率。

Gamon, *et al.*（2009）是文献中少有的将基于规则和基于统计的方法相结合的研究。该语法检查系统为模块化设计最终系统中将有四个模块采用基于统计的机器学习，包括介词、冠词、助动词以及动名词/不定式混淆；其余19个模块则采用基于规则的方法。不过，Gamon, *et al.*（同上）在文中只描述了冠词和介词模块的构建，包括特征提取、特征削减、分类器训练以及最后的测试。这两个模块的测试采用了两大类语料：英语本族语者语料和非本族语者语料。非本族语者语料又包括三小类：（1）CLEC（中国学习者英语语料库）中随机抽取的一万个句子；（2）母语为汉语、韩语或日语的非本族语者在网络上发布的英语语料，随机抽取一千句；（3）微软Outlook非本族语用户的电子邮件文本1,755句。测试结果表明，本族语者语料测试准确率要高于非本族语者语料。究其原因，可能是由于该系统的训练语料为本族语者语料，因此，分类器在对学习者文本进行判定时，容易受到各种错误的干扰，从而导致准确率的降低。虽然Gamon, *et al.*（同上）的研究只讨论了介词、冠词语法检查模块的情况，但是该研究却给了我们极大的启示，即要想开发一个真正意义上的通用型语法检查系统，可以考虑以下两点：（1）不同的语法错误检查采用不同的方法，让每种方法的优势最大化；（2）系统的模块化设计。大的系统由小的模块构成，每个模块既各自独立，又可以协同工作，共同完成任务。笔者文献调查发现，现有的通用型语法检查系统还没有一个能够完全覆盖所有语法错误，Gamon, *et al.*（同上）的建议非常值得尝试。

值得一提的是，有一些通用型语法检查系统是专门为中国学习者设计的。

和所有基于错误语法的研究一样，Liou（1991）、Liou, *et al.*（1991）在制定错误规则前，首先对125篇中国学生作文进行了错误分析，归纳出了14类错误，共93小类。不过，他们的研究只报告了7种类型错误的检查。该研究并未提供测试结果，只是在文中提及了制定错误规则时遇到的种种问题，例如，动词次范畴化错误（verb sub-categorization errors）的错误规则难以制定的问题（Liou, *et al.* 1991）。可见，并不是所有的错误都能很方便地用错误语法表示出来，而构建一部“完备的”错误语法则难上加难。廖信海（2003）开发的Wordhelp采用了基于规则的浅层句法分析、基于实例和负规则的模式匹配两种方法，准确率较高。但是笔者认为该研究对错误分析的描述不够具体，测试情况也有待进一步说明。

Chen & Xu（1990）为中国学生设计的Grammar Debugger则采用了另外一种

思路，即在已有的Parsifal句法分析器的基础上进行了两方面的操作：（1）对原本已经较为完备的正确语法模型进一步扩充、完善；（2）如果遇到部分不合语法的句子无法通过句法分析的情况，则采取约束松弛的办法。通过这两种方法，该研究便可以实现中国学生语法错误检查的目的。不过，遗憾的是，研究者并未提供系统测试结果，只是展示了一些目标错误句子的分析结果。该方法和Brehony（1993）的方法相似，即一方面利用了已有句法分析器及其较为完备的语法，另一方面又可以结合目标错误对原来的规则进行调整，达到为己所用的目的。

事实上，设计语法检查系统时，利用或改造已有的句法分析器是非常重要的。Vandeventer-Faltin（2003）在阐述她的语法检查系统所基于的The Fips Parser时，就谈到了使用已有工具的好处：（1）节省时间和资源。句法分析器的设计和开发是一个耗时耗力的工作，“要想开发一个能够提供准确的合语法性判断的NLP工具可能需要花费若干年时间，因此，使用已有工具一定是一个更为有效的办法”（Schulze & Hamel 2000：86）。（2）使用成熟的句法分析器可以让系统的鲁棒性和可靠性更好，语法的覆盖面更大。而这几方面对于语法检查系统来说都是至关重要的。另外，笔者认为，改造合适的句法分析器可以很好地验证自己的研究假设，并可以很快应用到实际中来。

2.2.2 专用型语法检查系统

近年来，越来越多的研究者开始专门研究某一类语法错误的自动检查，例如冠词错误、介词错误、限定词错误、动词短语错误，等等。笔者将这些语法检查系统定义为“专用型语法检查系统”。这种语法检查系统的不断出现在某种程度上表明，本领域的研究者们在不断反思以往的研究方法，同时在追求语法检查系统“大而全”和“小而精”的道路上不断地调整着研究思路。按照Gamon, *et al.*（2009）的说法，“大而全”和“小而精”这两个概念并不矛盾，前者是最终目标，后者则是在实现这个最终目标的道路上获得的阶段性成果。只有将一个个小的语法错误解决好，才有可能构建一个“大而全”的语法检查系统。

根据已有文献，专用型语法检查系统的研究主要集中于介词错误的自动检查，而实现方法也主要表现为基于统计的方法，因为判断介词错误通常需要邻词或局部上下文信息的帮助。由于数据稀疏问题，大多数研究的训练语料采用的是英语本族者语料，例如，Chodorow, *et al.*（2007）、Tetreault & Chodorow（2008）、De Felice & Pulman（2009）等等。不过，这一做法会给系统的实际应用带来问题，学习者语法错误的特殊性往往会对已经训练好的分类器或语言模型造成干扰（Gamon, *et al.* 2009），从而降低系统错误检查的准确率和召回率。在这些研究中，使用学习者文本作为测试语料所得结果基本上都低于本族语者文本的测试结果。

目前，只有为数不多的研究采用了具有错误标注的学习者语料库作为训练语料，如，Han, *et al.*（2010）。由测试结果来看，尽管该研究的准确率较高，达到了

93.3%，但是召回率却很低，只有14.8%。也就是说，如果目标错误共有100个，该系统只查到了14.8个，而在所查到的错误中，13.8个为准确判断（14.8*93.3%），其中一个为误判。换句话说，尽管该研究的训练语料采用了学习者语料库，但是其表现仍然无法令人满意[5]。为了突破训练数据的限制，有的研究者更是开始尝试利用庞大的网络资源对系统进行训练，如Yi, *et al.*（2008）对限定词错误检查的研究，就是通过网络搜索引擎来获得训练数据的，虽然测试结果不尽如人意，但是该研究发现，基于网络的方法还需要结合一些局部语言信息（local linguistic resources）才能够更好地对目标错误进行检查。

除了对训练语料进行改进，提高系统查错准确率和召回率的另一个方法就是，为基于统计的分类器或语言模型提供更有价值的语言特征。根据Leacock, *et al.*（2010: 47）的总结，现有语法检查研究所用到的语言信息大致可以分为以下三类：（1）形符上下文信息（token context information），即目标词左右的若干邻词；（2）句法上下文信息（syntactic context information），即词性信息和句法信息或组块信息（chunk information）；（3）语义信息。其中，形符上下文和词性信息是使用最为广泛的特征，Chodorow, *et al.*（2007）的介词错误自动检查研究采用的就是这两种特征。为了进一步完善系统设计，Tetreault & Chodorow（2008）和Tetreault, *et al.*（2010）在Chodorow, *et al.*（2007）已有研究的基础上进行了尝试性的改善。其中，Tetreault, *et al.*（2010）主要探讨的是在特征集中加入句法特征是否能够提高错误自动检查准确率和召回率的问题。结果表明，句法特征的加入能够显著提高本族语者文本中介词选择的准确率，并且可以提高学习者文本中介词错误的检查性能（尽管不显著，但是有提高）。由此也可以看出，句法分析器对于基于统计的语法检查系统来说具有非常重要的意义。因为目标词的句法信息是一项非常重要的特征，任何词类错误的自动检查都不能置之不顾。

值得一提的是，有的研究为了优化系统性能，甚至采用了语义信息，如De Felice & Pulman（2009）的介词错误自动检查研究，就是通过WordNet提取了介词周围的实义词语义信息，并将其运用到了系统训练当中。

近年来，专用型自动语法检查所涉及的词类错误，只是涉及了部分功能词类的错误检查，除了Lee & Seneff（2008）的动词形式错误检查研究，以及一部分实义词类搭配错误检查研究（如Chang, *et al.* 2008；Park, *et al.* 2008；Futagi, *et al.* 2008等）之外，较少有研究探讨实义词错误的自动检查。正如Leacock, *et al.*（2010：3）指出的，目前学习者语言错误的自动检查还是一片待开发的领域，需要我们的不断努力和探索。

3. 存在的问题

通过对已有研究的回顾，本研究发现，相关研究存在以下几个方面的问题：

第一，对学习者语法错误的认识不够全面。对基于错误模式匹配或错误规则的研究来说，由于研究方法的需要，研究者首先需要对学习者作文中的语法错误进行分析，进而确定系统所要解决的目标错误。但是，有的研究（如廖信海2003）在错误分析方面的描述较少，影响读者对研究的深入理解；有的研究在错误分类方面还有待改进，例如，Huang, *et al.*（2011）[6]将目标错误分成了三大类：（1）词形错误，包括“限定词—名词”不一致和动词形式错误；（2）介词错误；（3）及物性错误。而笔者认为，第一类词形错误中的动词形式错误和第三类的及物性错误有重合之处，如此划分不太恰当。

另外，在不少介词错误检查研究中，研究者并未考虑介词在担任不同角色时，会受到不同成分的制约（即做附语时受介词宾语的制约，做谓语论元标记时受谓语动词制约），也并未对相关错误进行区别对待。所有包含介词错误的用法都被认为是介词本身的错误，而没有探讨造成错误的根源是什么。这样的检查系统即便准确地查到了介词错误，也无法给学习者有价值的反馈。

第二，较少考虑短语学层面的错误。在已有研究中，有的研究主要解决的是句法成分（constituents）方面的错误，如Schneider & McCoy（1998）。还有的研究试图检查语法细节方面的错误，如Liou（1991）、Liou, *et al.*（1991）、Brehony（1993）、Lawley（2003），以及Gamon, *et al.*（2009）等等。还有一些研究主要解决的是词汇层面的错误，如Lee & Seneff（2008）、De Felice & Pulman（2009）、Huang, *et al.*（2011）等等。就笔者所阅文献，少有研究者从短语学层面入手进行错误自动检查研究。

应该说，学习者语法错误检查研究已经从“重语法”的阶段，进入了“既重语法，又重词汇”的阶段，只是研究者们仍旧将语法和词汇分割为两个不同的范畴。事实上，在今天的语言研究中，语法和词汇之间的界限已经逐渐模糊，“每个单词都有自己的语法”，语法和词汇是不可分的（Lewis 1993：142）。我们应该“将短语看作是一种正常的语言组织原则”（Hunston & Francis 2000：21），并尝试将这种语言组织原则应用到语法错误的自动检查当中。

第三，对语言本体的关注不够。已有研究大多是从技术操作的视角进行的尝试，这对学习者语法检查研究有一定的推动作用。但是，正如Milton & Cheng（2010）所言，对于应用语言学家来说，我们面临着一个很严肃的问题：对技术的热情往往会使研究走上纯粹的操作性路线，对算法的过度关注也会让研究者忽略二语学习者长期的教学需求。另外，笔者发现，对技术的探讨往往会让研究者忽略语言本身。而对于一个自然语言处理系统来说，如果没有对语言的深刻描写，仅仅依靠技术力量是难以实现根本性突破的。

基于统计的语法检查系统更是由于其对语言的忽视，而引发了一些研究者的批评。邢永康、马少平（2003）认为，截至目前，即使最成功的统计语言模型技

术也很少考虑什么才是真正的语言。赵正文、康耀红（2006）则认为，最常用的N元模型把语言作为无意的符号序列处理，只是通过对单词之间统计关系的挖掘来表示语言模型。尽管这些模型在某些应用领域表现不错，但这种先天不足最终会表现出来。

第四，对句法分析器语法模型改编不够。任何一个形式化模型或句法分析器都不是完美的，需要根据研究目的进行修订或改编，但是有的基于句法分析的语法检查研究对原分析器语法模型的改编不够。例如，Brehony（1993）对链语法分析器的改造只能检查研究者设定好的目标错误，而无法从根本上提高链语法分析器的句法分析能力。

4. 启示

根据上述问题，我们可以得到以下四个启示：

首先，自动语法检查研究者都应该深入分析学习者错误，对学习者错误的特殊性和需求了然于心。只有对错误有了深刻的认识，我们才能够在研究中充分考虑到系统对于学习者错误的判断力，并预见可能出现的问题。

其次，我们应该尊重语言事实，要认识到词汇和语法在很大程度上并不是非此即彼的关系，两者是不可分割的（Sinclair 1987；Francis 1993；Hunston & Francis 2000）。换句话说，“句法结构和词项之间具有共选关系，不能将其分开考察”（Francis 1993：142-143）。

再次，我们要认识到语言学理论的重要地位，用语言学的理论指导自动语法检查研究。我们的系统必须能够对语言有一个充分的描述，忠实于它所处理的所有语言行为（Bolt & Yazdani 1998）。

最后，要敢于对已有的语法模型进行改编。所有的句法分析都需要有一个形式化的语法模型，语法模型按最理想的目标来说，必须完全同真实语言相吻合，既不宽泛也不缩减（葛诗利、陈潇潇 2009）。这就要求研究者在对句法分析器进行改编时，要考虑其语法模型是否有问题，是否需要做进一步的完善或优化，而不是简单的规则增减。此外，为了更好地反映语言的真实用法，我们还需要不断地为句法分析器添加语法信息，甚至语义信息和语篇信息等等，从而不断完善其句法分析的能力。

注释

1. 根据笔者所阅文献，绝大多数语法检查系统还处于原型阶段。

2. 本研究回顾的学习者语法检查工具既包括独立的语法检查工具，也包括模

块式的语法检查工具（如内嵌于作文评分或批改系统等的语法检查模块）。

3. 目前还没有声称能够检查出所有语法错误的系统。

4. 不同的语法检查系统有不同的功能，即“检查”（detection）、“诊断”（diagnosis），甚至“纠错”（correction）。考虑到本研究所关注的是“检查”功能，因此将暂不探讨以往研究中与“诊断”、“纠错”的有关问题。

5. 当然，不同的研究所采用的方法、训练数据和测试数据都不尽相同，因此，当我们说某一个系统性能较好或较差时，并不是在和其他研究进行比较（也无法进行比较），而是就数据本身而言的。

6. 该研究是基于正确模式的，笔者认为，其错误分类的提出主要是为了限定模式提取的范围。

参考文献

Bolt, P. & M. Yazdani. 1998. The evolution of a grammar-checking program: LINGER to ISCA [J]. *Computer Assisted Language Learning* 11(1): 55-112.

Brehony, T. 1993. Francophone stylistic grammar checking using link grammars [D]. MSc. Thesis. Limerick: University of Limerick.

Chan, A. 2010. Toward a taxonomy of written errors: Investigation into the written errors of Hong Kong Cantonese ESL learners [J]. *TESOL Quarterly* 44(2): 295-319.

Chang, Y., J. Chang, H. Chen & H. Liou. 2008. An automatic collocation writing assistant for Taiwanese EFL learners: A case of corpus-based NLP technology [J]. *Computer Assisted Language Learning* 21(3): 283-299.

Chen, S. & L. Xu. 1990. Grammar-Debugger: A parser for Chinese EFL learners [J]. *CALICO Journal* 8(2): 63-75.

Chodorow, M. & C. Leacock. 2000. An unsupervised method for detecting grammatical errors [A]. In *Proceedings of the Conference of the North American Chapter of the Association of Computational Linguistics (NAACL)* [C]. 140-147.

Chodorow, M., J. Tetreault & N. Han. 2007. Detection of grammatical errors involving prepositions [A]. In *Proceedings of the Fourth ACL-SIGSEM Workshop on Prepositions* [C]. 25-30.

Connors, R. & A. Lunsford. 1988. Frequency of formal errors in current college writing, or Ma and Pa Kettle do research [J]. *College Composition and Communication* 39(4): 395-409.

De Felice, R. & S. Pulman. 2009. Automatic detection of preposition errors in learner writing [J]. *CALICO Journal* 26(3): 512-528.

Francis, G. 1993. A corpus-driven approach to grammar: Principles, methods and examples [A]. In M. Baker, G. Francis & E. Tognini-Bonelli (eds.). *Text and Technology: In Honour of John Sinclair* [C]. Amsterdam: John Benjamins. 137-156.

Futagi, Y., P. Deane, M. Chodorow & J. Tetreault. 2008. A computational approach to detecting

collocation errors in the writing of non-native speakers of English [J]. *Computer Assisted Language Learning* 21(4): 353-367.

Gamon, M., C. Leacock, C. Brockett, W. Dolan, J. Gao, D. Belenko & A. Klementiev. 2009. Using statistical techniques and web search to correct ESL errors [J]. *CALICO Journal* 26(3): 491-511.

Han, N., J. Tetreault, S. Lee & J. Ha. 2010. Using error-annotated ESL data to develop an ESL error correction system [A]. In *Proceedings of the Seventh International Conference on Language Resources and Evaluation (LREC)* [C]. 763-770.

Huang, C., M. Chen, S. Huang & J. Chang. 2011. EdIt: A broad-coverage grammar check using Pattern Grammar [A]. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)* [C]. 26-31.

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

Kann, V. 2002. CrossCheck – A grammar checker for second language writers of Swedish [OL]. http://www.csc.kth.se/tcs/projects/xcheck/ansokan02-1.pdf. (accessed 27/06/2010).

Lawley, J. 2003. The development of a grammar checker for Spanish secondary students of English as a foreign language [J]. *RESLA* 16: 127-138.

Leacock, C. & M. Chodorow. 2001. *Automatic Assessment of Vocabulary Usage without Negative Evidence (TOFEL Research Report RR-67)* [M]. Princeton, N.J.: Educational Testing Service.

Leacock, C. & M. Chodorow. 2003. Automated grammatical error detection [A]. In M. Shermis & J. Burstein (eds.). *Automated Essay Scoring: A Cross-Disciplinary Perspective* [C]. Mahwah, N.J.: Lawrence Erlbaum Associates. 195-207.

Leacock, C., M. Chodorow, M. Gamon & J. Tetreault. 2010. *Automated Grammatical Error Detection for Language Learners* [M]. San Rafael, CA.: Morgan & Claypool Publishers.

Lee, J. & S. Seneff. 2008. Correcting misuse of verb forms [A]. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics: Human Language Technology (ACL/HLT)* [C]. 174-182.

Lewis, M. 1993. *The Lexical Approach: The State of ELT and a Way Forward* [M]. Hove: LTP.

Liou, H. 1991. Development of an English grammar checker: A progress report [J]. *CALICO Journal* 9(1): 57-70.

Liou, H., H. Hsu, Y. Huang & V. Soo. 1991. Development of an automatic English grammar debugger for Chinese students [A]. In *Proceedings of Rocling IV Computational Linguistics Conference IV* [C]. 277-302.

Milton, J. & V. Cheng. 2010. A toolkit to assist L2 learners become independent writers [A]. In *Proceedings of the NAACL HLT 2010 Workshop on Computational Linguistics and Writing* [C]. 33-41.

More, J. 2006. A grammar checker based on web searching [OL]. http://www.uoc.edu/digithum/8/dt/eng/more.pdf. (accessed 27/06/2010).

Park, T., E. Lank, P. Poupart & M. Terry. 2008. "Is the Sky Pure Today?" AwkChecker: An assistive tool for detecting and correcting collocation errors [A]. In *UIST'08,* October (19-22) [C]. 121-130.

Schneider, D. & K. McCoy. 1998. Recognizing syntactic errors in the writing of second language learners [A]. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics (ACL) and the 17th International Conference on Computational Linguistics (COLING)* [C]. 1198-1204.

Schulze, M. & M. Hamel. 2000. Towards authentic tasks and experiences: The example of parser-based CALL [J]. *The Canadian Journal of Applied Linguistics* 3(1-2): 79-90.

Sinclair, J. 1987. Grammar in the dictionary [A]. In J. Sinclair (ed.). *Looking Up: An Account of the COBUILD Project in Lexical Computing* [C]. London: Collins ELT. 104-115.

Tetreault, J. & M. Chodorow. 2008. The ups and downs of preposition error detection in ESL writing [A]. In *Proceedings of the 22nd International Conference on Computational Linguistics (COLING)* [C]. 865-872.

Tetreault, J., J. Foster & M. Chodorow. 2010. Using parse features for preposition selection and error detection [A]. In *Proceedings of the ACL 2010 Conference Short Papers* [C]. 353-358.

Tschichold, C. 2003. Lexically driven error detection and correction [J]. *CALICO Journal* 20(3): 549-559.

Vandeventer-Faltin, A. 2003. Syntactic error diagnosis in the context of computer-assisted language learning [D]. Ph.D. Dissertation. Geneva: University of Geneva.

Yi, X., J. Gao & W. Dolan. 2008. A web-based English proofing system for English as a second language users [A]. In *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP)* [C]. 619-624.

葛诗利、陈潇潇，2009，大学英语作文自动评分研究中的问题及对策 [J],《山东外语教学》（3）：21-26。

廖信海，2003，基于实例和规则相结合的语法检查研究及系统实现 [D]。硕士学位论文。广州：中山大学。

邢永康、马少平，2003，统计语言模型综述 [J],《计算机科学》(9)：22-26。

赵正文、康耀红，2006，统计语言模型在信息检索中的应用 [J],《计算机工程与应用》（36）：158-161。

通讯地址：100029　北京市对外经济贸易大学英语学院

语言学研究中的多因素分析

北京邮电大学/北京外国语大学 房印杰

提要：多因素分析是一种新的语言学研究范式，本文尝试对语言学研究中的多因素分析进行界定，梳理多因素分析的两个发展阶段：基于人工的特征分析和基于统计模型的多因素分析，并对现有的多因素分析案例进行分类。一系列现有研究表明：虽然该范式还处于萌芽阶段，各类统计模型的应用具有探索性，但是多因素分析范式在语言学研究中已经体现出强大的生命力。

关键词：多因素分析、统计模型、特征分析、相关性

1. 引言

多因素分析（multifactorial analysis）是针对某一研究对象开展的全方位、整体性研究。其核心理念在于对影响该研究对象的所有潜在因素开展共时分析，包括：单个潜在影响因素对研究对象的影响权重；多种因素间的交互效应，这种交互效应可能是正向的，也可能是负向的（Dell & O'Seaghdha 1994）。对词汇、句法问题进行多方位的使用特征分析，这一研究范式源自不同的语言学领域：Dirven（1982）和 Rudzka-Ostyn（1989）从认知语言学的视角出发研究了一系列的近义词汇，Atkins（1987）和Hanks（1996）则从语料库语言学出发对词汇进行探讨。Gries（2003）第一次将多因素统计分析的研究方法与行为使用特征的分析范式相结合，以认知语言学理论和统计量化的方法厘清两种动词-介词构式：动词+宾语+介词构式和动词+介词+宾语构式。Gries（2003）有力地论证了上述两种构式有着不同的原型，其各自的使用可以通过复杂统计模型加以预测。Gries（2003）所倡导的认知语言学理论和复杂统计方法相结合的研究方式在近十年中逐步为更多学者所接纳、使用（Gries & Stefanowitsch 2006；Arppe 2008；Gries & Divjak 2009；Glynn 2009，2010；Divjak 2010；Glynn & Fischer 2010；Glynn & Robinson 2014等）。

2. 多因素分析的界定

Gries（2013：247）指出：我们生活在一个多种因素相互并存的世界里，似乎没有哪个现象只受到某一单独因素的制约。在该前提下，针对任何现象的科学研究，都必然需要综合考察所有对该现象有潜在影响的因素，面对几十个乃至上百

个潜在影响因素，学者自身的内省式分析会变得极为不科学。对海量因素变量的分析催生了现代统计学的发展。统计逐步脱离单纯对现象的描述，开始向推断性统计、预测统计发展。多因素分析正是伴随现代推断统计学的发展、成熟，而开始出现在各个学科的研究中。本文将语言学研究中的多因素分析界定为：基于描述、推断统计模型，对某一语言现象（如近义词语构式）的所有潜在因素开展的共时分析，该分析包括单个因素对该研究对象的影响权重和因素间交互作用。多因素分析的重要价值在于其从海量影响因素中客观、准确的剥离出对该研究对象具有显著影响的因素及因素间交互效应，为进一步的理论阐释奠定了坚实的基础。多因素分析的研究方法已经在诸多学科得到了广泛应用，如医学、社会学、经济学等。语言学研究作为一门以统计为基础的实证学科（Gries 2013：7），与多因素分析有着天然的接合点。Kuznetsova（2015: 2）将多因素分析范式在语言学研究中的崛起原因归纳为三点：第一，大规模语料库的日臻成熟；第二，基于使用的研究范式（usage-based）的崛起；第三，对语言使用的原型性的认识。

3. 语言学研究中多因素分析的发展阶段

语言学自成为一门独立学科之初，便强调在具体研究中使用实证性的研究方法（Bloomfield 1933；Harries 1951；Firth 1957），20世纪90年代实证性研究方法再次在语言学研究中受到重视后，众多语言学家逐步认识到：任何语言学现象均受到多种因素的共同制约。Vanhatalo（2003）提出：语言使用中的近义词选取受到语境因素的影响，该语境应当超越词语搭配（collocation）的范畴，而语境因素必然是多方面的，包括词语层面、词语-句法层面、语义层面、语用-篇章层面等。学界普遍认可语言选择过程中存在一系列的影响因素，但对这些因素之间的相互关系，对选择过程的决定性权重则关注不足，研究有待深入（Benor & Levy 2006：233）。总体来看语言学研究中的多因素分析经历了人工标注、人工分析和人工标注、统计模型预测两个阶段。

3.1 基于人工的特征分析

语言学家对多因素分析的关注由来已久，Bloomfield（1933）提出语言学研究中应当关注那些可以观察到的客观特征，如词汇、句法特征，该论断被发展为结构主义语言学中的分布分析（Harris 1951）。虽然分布分析排斥人类主观内省，但是其对外在语言特征的关注被其他学者进一步发展成为特征分析。Glynn（2010）将特征分析概括为：在语言学研究中，通过挖掘前人内省式研究积累的大量文献，提取出多种语言学特征（词汇、词汇-句法、语义、语用等），并对各类特征开展客观的实证分析，从而挖掘出语言现象背后的潜在规律。

早期的语言学研究中的多因素分析被称为特征分析（Glynn 2010），Dirven

（1982）和Rudzka-Ostyn（1989）从认知语言学角度作了特征分析；Atkins（1987）和Hanks（1996）开展了语料库语言学研究中的特征分析。以下分别选取Dirven（1982）和Hanks（1996）对认知语言学和语料库语言学研究中的早期特征分析简要说明。

3.1.1　语料库语言学层面

Hanks（1996）提出：在研究动词义项的过程中，应当充分考虑该动词所在的语境，并进一步认为，每个词的意义均由其整体语境决定（the semantics of each word is determined by the totality of its complementation patterns）。Hanks的论断被概括为语言学中的行为全貌研究范式（behavioral profile）（Gries 2003）。Hanks从词典编纂视角出发，分别选取一系列动词，通过抓取其相应的词汇、句法特征进而确定其义项分类。在对动词urge的义项描写中，Hanks提出：通过不断抓取urge所处语境的特征，urge的各个义项会逐步被确定下来。换言之，通过多种特征的相互交叉，urge的相应义项得以确定。语料库提供了观察词语型式（pattern）的手段，但是并不能直接观察到词语的意义。为了将意义和词语形式（form）相结合，就必须抽取出该词语所在的各类搭配、句法型式（Hanks 1996：116-117）。将形式与意义合二为一充分说明了Hanks以各类语言特征来分析、提取词语义项的行为属于全貌研究范式。

Hanks（1996）将其特征分析概括为若干步骤：使用统计方法提取目标词的显性搭配词，将抽取的显性搭配词按照一定义项加以分类，对各个义项分类命名，抽取更多索引行并对既有义项分类体系进行修正，观察搭配词的句法成分，记录搭配词出现频次、义项频次等。

在针对一系列动词的特征分析中，Hanks（1996：126）坦言：确定各类语言学特征并不容易，因为每一个动词的句法、语义交互都是不同的。Gries（2003）也认为：多因素分析的特征随研究对象的不同而变化。Hanks（1996）针对一系列动词的特征分析以真实语料为研究基础，通过人工对各类特征进行分类进而确定动词义项。搭配特征占据了Hanks研究的重点，同时Hanks指出：更多的句法特征也应被纳入分析范畴。

Hanks（1996）凸显了特征分析方法的重要价值，强调了词语研究中语义-句法交互的重要性。但人工开展特征分析使得型式抽取变得痛苦而缓慢（Hanks 1996：141）。

3.1.2　认知语言学层面

与Hanks（1996）相类似，Dirven（1982）将动词talk作为研究对象，通过特征分析来剥离其不同义项。通过对比talk与say、tell、speak的差异，Dirven提出：talk主要用于强调话题（discourse topic），在这一背景下，talk一经出现便暗示句子主语应当为人，对话题的强调使得talk表达具体言说内容的作用退化，并

最终形成其原型性句式：主语（人）+talk。该句式中并不包含宾语成分。而当talk需要表达言说动作的对象和动作接受者时，介词about与to便成为其高频搭配。Dirven（1982）发现，talk除去表达强调话题这一主要义项之外，还包含许多从属义项，其各自存在不同的句法表现形式。例如，当talk被用来表示言说行为时，更多地以状语形式出现。Dirven（1982：65）将其句法型式概括为六类：

a. Talk like that/this

b. Talk like NP (does)

c. Talk + way

d. Talk + adverb in –ly or ø

e. Talk as if

f. Talk in NP

Talk的主要功能在于提示话题，其提示话题的功能如此之强，以至其衍生出诸多表现形式：talk about、talk on、talk of。Talk与诸多介词形成的词组帮助其表达言说行为的具体内容。Dirven（1982：80）认为talk的主要义项与从属义项构成了一个统一的语义体系。

Dirven（1982）更多的从认知角度分析talk及其衍生词组所发挥的认知、交际功能：talk自身主要用于提示话题，并不涉及言说行为的具体信息；当需要表达具体言说信息时，talk与介词（如about、to）形成搭配来实现这一义项。Dirven（1982）基于真实语料，尝试将认知、语义和句法相融合。在整个研究过程中，Dirven充分利用了talk的不同句法特征来对其不同义项加以说明。但同时，Dirven的分析过程依赖作者人工完成，仍带有较为浓重的内省式色彩，语料库的频次优势也没有得到充分体现。

上述研究首次通过抓取、提炼与研究对象相关的语言学特征，对其开展人工分析，进而解释了相关语言现象背后的规律。但是传统特征分析依靠研究者自身来人工完成多因素的特征分析，这使得语言现象背后的潜在规律未能得到充分挖掘。Gries & Deshors（2015）提出：没有哪个分析者可以直接用大脑量化出各个因素对某语言学现象的影响权重，更不可能单凭大脑测算出因素间的交互效应。人工进行海量因素的特征分析必然造成分析过程中的过度简化、部分潜在规律被忽视。

3.2　基于统计模型的多因素分析

人工操作的特征分析之所以耗时耗力，而效果欠佳，根本原因在于其对统计方法的应用不足。伴随现代统计学的蓬勃发展，一些复杂的推断性统计模型（如逻辑斯蒂回归、线性分类器、决策树等）已经具备在一定程度上对研究对象的

描写、解释、预测的能力。在结合了传统特征分析与现代统计学模型的基础上，Gries（2003）首次将基于统计模型的多因素分析范式引入语言学研究。在针对动词+小品词构式的研究中，Gries充分发挥了推断性统计的优势，对“动词 + 小品词”构式的原型特征做了准确阐述。

Gries（2003）针对VPC（动词+小品词）构式的研究包括两个核心部分：针对VPC的因素特征标注体系；复杂统计模型的应用。整个标注体系的特征选取基于百余年中前人针对VPC进行的内省式、单因素分析。Gries将前人的研究发现梳理分类成语音、词汇-句法、语义、篇章等层面，并将各种语言学特征转化为可操作的变量，然后分别对从BNC中提取到的403句包含VPC构式的句子展开人工标注。Gries（2003）设计的标注体系较之Hanks（1996）和Dirven（1982）更为清晰、庞大，操作性也更强。在完成语料人工标注后，Gries采用复杂统计模型中的线性分类器（linear discriminant analysis）对动词+小品词构式的两种构式（动词+宾语+介词vs.动词+介词+宾语）做了准确剥离，有力地证明两种构式分属不同的原型，而非传统语言学研究所认为的二者同属一种原型。同时，线性分类器还对导入统计模型的各个语言学因素变量赋予了不同的影响权重，从而客观、准确地将具备统计显著性的因素筛选出来。

统计模型与传统特征分析的结合既实现了对前人内省式研究发现的检验，同时也为进一步挖掘语言现象背后的深层规律创造了可能。基于统计模型的多因素分析可以对传统语言学研究中的诸多争议提供较为客观、完满的解答（Gries & Stefanowitsch 2006：58）。

4. 现有多因素分析研究分类

自Gries（2003）开创基于统计模型的多因素分析范式之后，越来越多的学者将该范式纳入具体的语言学研究中（Cappelle 2006；Arppe 2008；Gries & Divjak 2009；Divjak 2010；Glynn 2009，2010；Deshors 2010）。现有的多因素分析研究案例可以大致按照研究对象和对象语言加以划分。

4.1 按照研究对象分类

现有多因素分析选取的研究对象多种多样，大体可以划分为两大层面：词汇层面和句法层面（Glynn 2010）。

4.1.1 词汇层面

针对词汇的研究是认知语言学和语料库语言学的焦点（如；Gries 2006），针对词汇层面的多因素分析也占据了该类研究中的主要位置。从表1中可以发现：表达近义关系的词语一直是研究的焦点议题。虽然各个研究所用多因素分析方法

各不相同，但是其研究目的有很大相似性。各个研究均从特征分布焦点出发，通过多因素统计模型，提取近义词语结构的细微差异，及其在认知、心理语言学层面的意义。

表1. 词汇层面的已有多因素分析研究

研究对象	多因素分析方法	出处
意向类动词	HCA	Divjak（2006）
尝试类动词	HCA	Divjak & Gries（2006）
认知类动词	逻辑斯蒂回归	Arppe（2008）
致使类动词	MER	Levshina（2011）
May与can	逻辑斯蒂回归	Deshors（2010）

4.1.2 句法层面

句法层面的多因素分析在数量上大大少于针对词汇层面开展的相关研究。造成这一现状的原因可能是多方面的。首先，各种句法结构的自动提取并不理想（Roland 2007），这使得针对句法结构的大规模研究难以开展。同时，针对句法结构的特征分析涉及因素更多，其中不乏尚存争议的问题。目前真正实现了句法层面标注的语料库仅有ICE-GB等少数几个，这使得针对不同题材、不同国籍背景学习者语言的研究难以开展。在研究对象上，现有研究多针对传统语法和生成语法研究中的热点或难点问题，如Gries（2003）的VPCs（动词+小品词）构式分析，直接针对前人研究中的争议开展多因素分析，并最终证明VPCs的两个形式：动词+介词+宾语vs.动词+宾语+介词分别存在不同的原型。

表2. 句法层面的已有多因素研究

研究对象	多因素分析方法	出处
动词+小品词构式	线性分类器	Gries（2003）
形容词顺序	主成份分析法	Wulff（2003）
时间状语从句	逻辑斯蒂回归	Diessel（2008）
并列结构	逻辑斯蒂回归	Lohmann（2014）

4.2 按照本族语/中介语分类

现有的多因素研究分析除了可以从词汇-句法层面切分，还可以按照研究

对象的所属语言进行划分。本研究将研究对象所属语言区分为两类：母语和中介语。

4.2.1　针对本族语者的多因素研究

绝大部分的多因素研究属于对母语的分析，如Gries（2003）从BNC语料库中抽取VPCs构式，这就决定了其所使用的语料为英语母语者的口、笔语。Divjak（2010）和Arppe（2008）则分别探讨俄语和芬兰语中的近义词类，Levshina（2011）和Lohmann（2014）分别考察荷兰语中表达let概念的使役动词和英语母语者的并列结构。纵观自Gries（2003）以来的多因素研究，不难发现针对不同本族语者的多因素研究长期占据主导地位。产生这一现状的原因可能在于针对母语的研究相对便于操作，母语者间的异质性便于控制。

4.2.2　针对中介语的多因素研究

Deshors（2010）自称第一次将多因素分析应用于中介语研究，在针对情态动词may和can的差异辨析中，Deshors初步证明了多因素分析方法在中介语研究中的价值。

面对前人针对may和can的大量文献，Deshors将基于统计模型的多因素分析引入该研究中。为了探讨may和can的潜在差异及学习者母语对中介语的潜在影响，Deshors选取LOCNESS（英美大学生议论文语料库）为参照语料库，ICLE（国际学习者英语语料库）中的法语子库为学习者语料库，CODIF（法国学生作文语料库）为法语母语语料库开展对比研究。仿效Gries（2003）、Deshors（2010：95）依托前人文献建立语言特征标注框架。该标注框架涵盖语义、句法、词汇三个方面的语言学特征共计21个变量。在数据分析部分，Deshors采用二元逻辑斯蒂回归来提取中介语语料库和英语本族语者语料库的显性差异变量。逻辑斯蒂回归结果显示：从句类型、动词语义、主语单复数、主语有灵、句式肯定/否定等五个语言学因素在两个语料库中有显著性差异，可以作为区分英语本族语者和法国英语学习者的语言学因素。更为重要的是，Deshors从认知加工角度对逻辑斯蒂回归提取到的显性因素做了解释：各个显性因素均表明，在认知加工负载增大时，法国英语学习者往往会做出与英语本族语者不同的may和can的选择。

Deshors（2010）是基于统计模型的多因素分析方法在中介语研究中的首次应用。该研究初步证明了多因素分析方法与中介语研究的契合性。

5. 多因素分析中的常用统计方法

多因素分析虽然在语言学研究中起步较早（Dirven 1982；Hanks 1996等），但基于统计模型的多因素研究范式在语言学中应用时间并不长，仍处于一个摸索阶段。Gries（2013）将多因素分析称为一种推断性分析方法，本质上属于一种探索

性研究。因此，不同学者使用的具体统计模型并不相同，本文主要介绍在语言学研究中使用较多的两种统计模型：线性分类器和逻辑斯蒂回归。

5.1 线性分类器

线性分类器（linear discriminant analysis）属于统计学中诸多分类方法的一种（薛毅、陈立萍 2006）。该方法尝试将两种句法结构按照各自的语言学特征权重进行分类。分类中参照的线性分类值源自每个语言学变量对两种句法结构分类中的贡献权重，语料集中的每一个具体句法结构都将被赋予一个分类值，然后所有句法结构的分类值将形成一个连续统。在该连续统的两端为两个句法结构的理想分类位置，句法结构的分类值位置越趋近于中央位置，则该结构区分度越小。连续统两端的极值代表两个构式的原型（Gries 2003：108）。

线性分类器显示出特定语言学因素对确定某一语法结构的权重。由于其能够同时处理所有人工标注的语言学因素，其在一定程度上模拟了语言使用者的在线加工过程。本族语者在语言使用中，总是会在潜意识中追踪、积累各个相关因素（Gries 2003：6）。线性分类器也存在不足：它对于多因素分析中数据的正态性分布有要求，Gries（2003）在使用线性分类器时也对这一点做了相应说明。

5.2 逻辑斯蒂回归

逻辑斯蒂回归是对线性回归的拓展，属于广义线性模型（generalized linear models）的一种。其操作原理在于：通过一系列自变量（predictors）的权重预测，来确定各个自变量及其变量交互效应（variable interaction）对因变量（dependent variable）的影响。逻辑斯蒂回归的优点在于其不要求数据的正态性分布，这使得该方法的使用范围更广。在语言学的研究中，二元逻辑斯蒂回归使用更加广泛（Deshors 2010；Hoffmann 2006）。逻辑斯蒂回归的应用使得诸多显性语言学因素能够被准确、客观地提取出来，并进一步反映出内在的语言学规律。但是逻辑斯蒂回归在语言学研究中的应用也并非毫无问题。一些学者在开展逻辑斯蒂回归分析时，并未将变量交互效应纳入考察范畴，这极大削弱了统计模型对语言学研究的解释力，甚至存在重要语言学因素被忽视的可能。

6. 小结

本文围绕语言学研究中的多因素分析这一主题展开，介绍了多因素分析的界定，语言学研究中多因素分析的发展阶段，现有多因素分析研究的分类和语言学研究中常用的两种多因素分析方法。

多因素分析被引入语言学研究是对后者向科学性迈进的一大提升。正是因为

有了基于统计模型的多因素分析方法，诸多语言学研究得以实现更加准确、客观的量化分析，得以超越对原始频次的笼统、单一依赖。现有研究已经证实：多因素分析适用于各类语言学研究对象，包括词汇层面、句法层面；也适用于不同语言研究，包括母语和中介语。但是，采用复杂统计模型分析语言现象的研究范式还处于萌芽阶段，针对语言学现象的数据分析、模型预测并没有一套固定不变的方法，数据分析本身就是一个探索过程（Gries 2013）。现有的多因素统计模型还只能挖掘各类变量因素间的相关性，不能反映因果性（梁茂成 2015：24）。综上所述，基于统计模型的多因素分析方法已经初步证明了其在语言学研究中的适用性和生命力。作为一种新的研究范式，更多的语言学研究亟需引入多因素分析，使得其在理论层面、方法论层面得以提高。

参考文献

Arppe, A. 2008. Univariate, Bivariate and Multivariate Methods in Corpus-based Lexicography: A Study of Synonymy [D]. Ph.D. Thesis. Helsinki: University of Helsinki.

Atkins, B. 1987. Semantic ID tags: Corpus evidence for dictionary senses [A]. In *The Uses of Large Text Databases: Proceedings of the Third Annual Conference of the UW Centre for the New Oxford English Dictionary* [C]. Waterloo: University of Waterloo. 17-36.

Benor, S. & R. Levy. 2006. The chicken or the egg: A probabilistic analysis of English binomials [J]. *Language* 82(2): 233-278.

Bloomfield, L. 1933. *Language* [M]. New York: Henry Holt and Company.

Cappelle, B. 2006. Particle placement and the case for "allostructions" [J]. *Constructions* SV1-7: 1-28.

Dell, G. & P. O'Seaghdha. 1994. Inhibition in interactive activation models of linguistic selection and sequencing [A]. In D. Dagenbach (ed.). *Inhibitory Processes in Attention, Memory, and Language* [C]. San Diego: Academic Press. 409-453.

Deshors, S. 2010. A Multifactorial Study of the Uses of May and Can in French-English Interlanguage [D]. Ph.D. Thesis. Brighton: University of Sussex.

Diessel, H. 2008. Iconicity of sequence: A corpus-based analysis of the positioning of temporal adverbial clauses in English [J]. *Cognitive Linguistics* 19: 457-482.

Dirven, R. 1982. "Talk": Linguistic action perspectivized as discourse [A]. In R. Dirven, L. Goossens, Y. Putseys & E. Vorlat (eds.). *The Scene of Linguistic Action and Its Perspectivization by Speak, Talk, Say, and Tell* [C]. Amsterdam: John Benjamins. 37-83.

Divjak, D. 2006. Ways of intending: Delineating and structuring near-synonyms. In S. Gries & A. Stefanowitsch (eds.). *Corpora in Cognitive Linguistics*: *Corpus-based Approaches to Syntax and Lexis* [C]. Berlin: Mouton de Gruyter. 19-56.

Divjak, D. 2010. *Structuring the Lexicon: A Clustered Model for Near-synonymy* [M]. Berlin: Mouton de Gruyter.

Firth, J. 1957. Modes of meaning [A]. In J. Firth (ed.). *Papers in Linguistics 1934-1951* [C]. Oxford: OUP. 190-215.

Glynn, D. 2009. Polysemy, syntax, and variation: A usage-based method for cognitive semantics [A]. In V. Evans & S. Pourcel (eds.). *New Directions in Cognitive Linguistics* [C]. Amsterdam: John Benjamins. 77-106.

Glynn, D. 2010. Synonymy, lexical fields, and grammatical constructions: A study in usage-based cognitive semantics [A]. In H. Schmid & S. Handl (eds.). *Cognitive Foundations of Linguistic Usage-patterns* [C]. Berlin: Mouton de Gruyter. 89-118.

Glynn, D. & K. Fischer (eds.). 2010. *Quantitative Cognitive Semantics: Corpus-Driven Approaches* [C]. Berlin: Mouton de Gruyter.

Glynn, D. & J. Robinson (eds.). 2014. *Corpus Methods for Semantics: Quantitative Studies in Polysemy and Synonymy* [C]. Amsterdam: John Benjamins.

Gries, S. 2003. *Multifactorial Analysis in Corpus Linguistics: A Study of Particle Placement* [M]. London: Continuum Press.

Gries, S. 2013. *Statistics for Linguistics with R* [M]. (2nd revised and extended edition). Berlin: Mouton de Gruyter.

Gries, S. & S. Deshors. 2015. EFL and/vs ESL? A multi-level regression modeling perspective on bridging the paradigm gap [J]. *International Journal of Learner Corpus Research* (1): 130-159.

Gries, S. & D. Divjak. 2009. Behavioral profiles: A corpus-based approach towards cognitive semantic analysis [A]. In V. Evans & S. Pourcel (eds.). *New Directions in Cognitive Linguistics* [C]. Amsterdam: John Benjamins. 57-75.

Gries, S. & A. Stefanowitsch (eds.). 2006. *Corpora in Cognitive Linguistics: Corpus-based Approaches to Syntax and Lexis* [C]. Berlin: Mouton de Gruyter.

Hanks, P. 1996. Contextual dependency and lexical sets [J]. *International Journal of Corpus Linguistics* 1(1): 75-98.

Harris, Z. 1951. *Methods in Structural Linguistics* [M]. Chicago: University of Chicago Press.

Hoffmann, T. 2006. Corpora and introspection as corroborating evidence: The case of preposition placement in English relative clauses [J]. *Corpus Linguistics and Linguistic Theory* 2(2): 165-195.

Kuznetsova, J. 2015. *Linguistic Profiles: Going from Form to Meaning via Statistics* [M]. Berlin: Mouton de Gruyter.

Levshina, N. 2011. Doe wat je niet laten kan: A Usage-based Analysis of Dutch Causative Constructions [D]. Ph.D. Thesis. Leuven: University of Leuven.

Lohmann, A. 2014. *English Coordinate Constructions: A Processing Perspective on Constituent Order* [M]. Cambridge: CUP.

Roland, D. 2007. Frequency of basic English grammatical structures: A corpus analysis [J]. *Journal of Memory and Language* 57(3): 348-379.

Rudzka-Ostyn, B. 1989. Prototypes, schemas, and cross-category correspondences: The case of ask [A]. In D. Geeraerts (ed.). *Prospects and Problems of Prototype Theory* [C]. Berlin: Mouton de Gruyter. 613-661.

Vanhatalo, U. 2003. Evaluating the semantic content of near synonyms: Population tests versus corpus linguistics [J]. *Virittäjä* 107(3): 351-369.

梁茂成，2015，梁茂成谈语料库言学与计算机技术[J],《语料库语言学》(2)：15-25。
薛　毅、陈立萍，2006,《统计建模与R软件》[M]。北京：清华大学出版社。

通讯地址：100876　北京市北京邮电大学人文学院/100089 北京市北京外国语大学中国外语教育研究中心

大数据背景下BCC语料库的研制

北京语言大学 荀恩东 饶高琦 肖晓悦 臧娇娇

提要：“北京语言大学语料库中心（BLCU Corpus Center，简称BCC）”是以汉语为主、兼有其他语种的在线语料库。BCC总规模达数百亿字，是服务语言本体研究和语言应用研究的在线大数据系统。BCC检索式由字、词和语法标记等单元组成，并且支持通配符和离合查询。本文将概述BCC的总体情况，包括语料库建设情况和检索引擎开发等，重点介绍BCC形式化检索语言和在线系统的使用方法。

关键词：BCC语料库、大数据、语言检索、检索式

一、引言

在大数据背景下，语言本体研究、语言教学和语言应用研究都离不开语料库的支持。在语言本体研究中，利用大规模语料，对语言现象进行穷尽式考察，可以归纳、完善、验证语言理论或观点，又可以通过实证方法，为语言理论的研究提供数据支撑和量化分析；在语言教学中，语料库可以提供真实的语言素材，用于教学内容制定和讲解，使语言教学内容选取和教学实施过程更加科学，并可以支撑辞书和教材的编纂；同时，语料库作为模型训练知识库，在语言信息处理各种应用中起着不可或缺的作用。

采用语料库进行实证研究历史悠久，国内外一系列语料库系统推动了语言研究的进步和发展。中文语料库方面，有“国家语委语料库”、“北京大学现代（古代）汉语语料库”、“中国台湾中央研究院语料库”、“兰卡斯特汉语语料库”等；在英语语料库方面，有“英国国家语料库（BNC）”、“美国当代英语语料库（COCA）”等。语料库发展到今天，出现了新的特点和需求：

1）语料库规模越来越大，逐渐进入大数据时代。随着信息社会的发展，个人微机的迅猛发展和存储数据的硬盘造价持续下降，使得能够记录语言生活的终端设备越来越普及，数据存储能力越来越强，网络传输速度越来越快，每天产生的语料数量大大超过以往。这些发展都为大规模语料库的采集提供了技术支持。

2）语料库成为语言技术进步的知识库。在语言大数据基础上，语言应用技术快速发展，人工智能在多个应用领域取得突破性进展。这些新技术进步，正在改

变社会语言生活，为语言研究不断提供新课题并提出新的挑战。

3）语料库形式多样。语料的领域越来越细化，语料加工越来越深入，网络社交语料异军突起。

4）语料库使用越来越便捷。在线语料库查询和统计功能更加人性化，除了面向个人在线使用外，语料库还利用云服务接口，通过云调用大大拓展了语料库的应用范围。

“北京语言大学语料库中心（BCC）”（http://bcc.blcu.edu.cn）是以汉语为主、兼有其他语种的语言大数据，目标是为语言本体研究提供一个使用简便的在线检索系统和构建大数据的语言应用基础平台。BCC支持云服务，通过API调用方式为开展知识抽取、模型构建等研究和应用工作提供便利。

本文首先概述BCC研制的总体情况，重点介绍BCC检索式，并在附录中给出了BCC检索式实例和中英文词性体系。

二、BCC语料库研制

一个语料库系统的建设，主要包括三方面工作：语料库资源建设、检索引擎开发和提供语料库检索服务。如图1所示，语料库的资源建设是构建语料库数据内容的基础。BCC主要包括三种类型语料：多语种单语语料库、双语对齐语料库和深加工的树库。语料库检索内核是实现语料库系统的技术基础，采用基于后缀串的全文检索算法，并且支持通配符和离合模式匹配。检索服务是指使用语料库系统的方式和方法。BCC提供两种服务方式：在线检索和云调用。

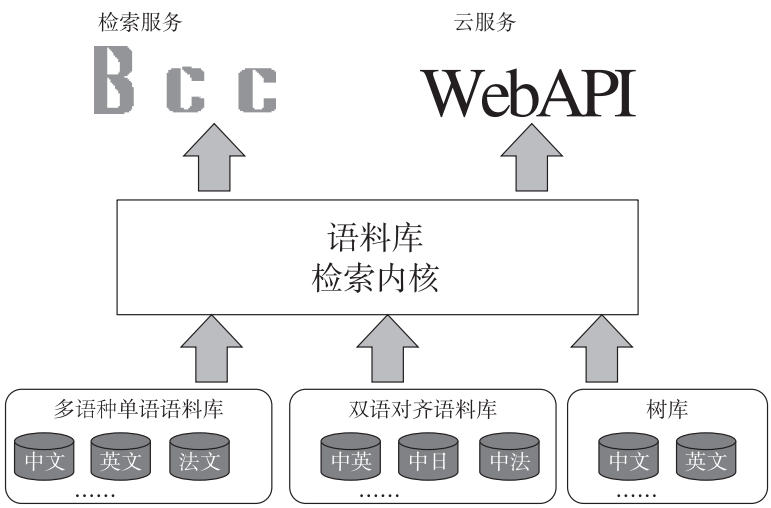


图1.　BCC语料库系统示意图

2.1　语料库资源建设

语料库建设是指在确定语料库内容、规模和形式后，对语料进行采集、加工和标注等，通过对自然语言文本的采集、存储、加工，可以凭借大规模语料库提供的客观语言事实为语言学研究提供支撑（黄昌宁、李涓子 2002）。BCC语料库具有以下特点：

语料库函盖多个语种

以汉语为主，兼顾其他语种的语料。目前BCC包含9种语言，如英语、西班牙语、法语、德语、土耳其语等。其中的英文语料主要采自《华尔街日报》，规模约为12亿单词。BCC语料以单语语料为主，也包括双语平行语料，如英汉、英德等双语对齐语料库。目前有9种语言互译，各类双语语料总规模约千万句。检索时，汉语最小的单位是汉字，其他语种最小的单位是单词，但单词不支持词形变化，保持原始语料中的形态，例如：英语The和the在语料库中是两个单词。

多层次语料加工

包括生语料、分词语料、词性标注语料和句法树。目前已对现代汉语、英语、法语的语料进行词性标注，除此以外的其他语料都是未加工的生语料；句法树包括中、英文树库，分别引自美国宾州大学的中文和英语树库。语料加工层次不同，支持检索的功能也不同，例如：生语料不支持带有词性信息的检索，树库支持短语类型标记的检索。

现代汉语语料和古代汉语语料兼具

对现代汉语语料进行了分词和词性标注，支持带有词性信息的检索；而古代汉语没有进行分词和词性处理，只能以字为单位进行检索。

汉语多语体

现代汉语语料涵盖新闻、口语（微博）、科技、文学、综合等多个语体。其中新闻、文学和综合语料标注时间、作者等组成信息，可以用BCC的“自定义”功能进行受限检索，即选择某一个子语料，限定在该语料中进行检索。

新闻语料：采自《厦门日报》、《厦门商报》、《厦门晚报》等；

口语（微博）语料：采自2013年新浪微博；

科技语料：采自国内学术期刊；

文学语料：采自国内外文学作品，对每个作品都标注了作品名称、作者、发表时间等信息。

综合语料：包括报刊、文学、微博、科技四个领域，语料内容独立，与其他语料不交叉，目标是建立一个“平衡”语料库。

共时语料和历时语料兼备

BCC对报刊语料和文学作品标注了时间信息，其中文学作品的时间信息体现在BCC的“自定义”功能应用上，用户可以选定某时间的文学作品进行限定检索；BCC“历时检索”主要是报刊语料，语料来自1945年至2015年的《人民日报》。历时检索是以图形可视化方式呈现的。

BCC语料库使用了语料采集、加工和语言分析处理等多种工具，例如对现代汉语进行分词和词性标注。为了完成语料采集、加工、标注等工作，开发了BCC语料库采集和加工平台，主要包括：

网上语料采集工具

BCC语料库中的语料主要源自互联网的页面文本，利用采集工具自动下载网页，把网页数据保存到本地。

语料加工整理工具

将网络作为语料库，是将以自然语言形式存在的整个网络电子文本当作一个庞大的语料库，可以通过征调主流搜索引擎的应用程序调用接口，获取搜索引擎的返回结果，再对其进行相应的语料库统计分析（熊文新 2015）。BCC语料加工整理的方式主要为：从网页中提取原数据信息，包括名称、出处等；网页数据清洗，从网页数据中剔除非内容数据，提取有效文本内容；对数据进行自动断句处理，为后续语言分析做准备；异常重复句子甄别和处理，剔除网页数据清洗阶段不能甄别的重复句子。

语言自动分析工具

原始语料完成断句后，在语言分析阶段对句子进行分词和词性标注处理。中文词性标注采用北京大学计算语言研究所提出的词性标注体系（俞士汶等 2000，2002），英文词性体系采用美国宾州大学词性体系。目前，BCC可以对现代汉语、英语、法语的语料进行自动分词和词性标注处理。

语料库标注平台

该平台的目标是通过人工标注来构建专门语料库。

2.2 BCC检索引擎

语料库建设是围绕内容进行的，用户通过检索使用语料库数据，而使用的检索功能是通过检索引擎实现的，因此检索引擎的性能直接影响语料库系统的使用体验。使用体验体现在多个方面，包括对数据规模的支持程度、语料类型的支持程度、响应检索的时空开销、检索式的支持功能、对服务器软硬件的适应性等。BCC检索引擎具有以下特点：

1）支持语言大数据。目前BCC检索内核支持建立超大规模语料库检索系统，单机可以索引的语料库规模最大可以支持64G（约320亿汉字），实际规模与机器内存相关。

2）支持多语种检索。BCC语料库检索内核技术支持中文、英文、日文等不同语种的语料库。

3）支持多种语料形式。BCC语料库包含原始语料、分词语料、词性标注语料，同时可以支持短语结构树的语料库检索。

4）支持功能强大的检索。BCC定义一种用户友好且功能强大的语料库检索语句，不仅具有模式查询和统计功能，支持带有词性的通配符和离合模式查询，还可以支持二次查询、自定义语料查询等，同时BCC还实现了在线统计以及在线反馈统计结果的功能。

2.3　语料库服务

BCC语料库服务包括两种形式：一种是在线检索，即在浏览器内使用BCC，输入检索式，以页面形式返回结果；另外一种是云服务，通过编程使用BCC的Web API接口形式来调用BCC。云服务一般用于BCC的二次开发，或者用于利用BCC进行语言的应用开发。

在BCC首页中可以选择不同语种的语料库，在输入框的上方，列出该语种的不同语料频道（如图2）如果想在某个频道中做更细化的查询，可以选择“自定义”搜索（如图3）通过点击语料库的组成窗口选择子语料库或者通过搜索定位子语料库。当用户选择一个子语料时，页面会给出该子语料库的语料规模，后续检索也会限定在该子语料库中进行。

图2.　BCC首页

图3. 通过“自定义”选择语料及查看语料组成和规模

图4. “帅气的n”检索结果页

输入检索式，点击“搜索”后得到检索页面，其中包括检索总条数信息、分页显示的检索实例等，如图4是“帅气的n”的检索结果。在搜索结果页面，BCC还提供在线统计、二次检索、下载结果、显示结果和查看原文等多个功能：

1）统计：BCC检索式中可以包含词性或者短语类型，也可以带有通配符。在结果页面，词性和通配符体现在具体的检索实例中。BCC通过统计实例，在线统计检索式在语料库中的分布情况，统计结果页如图5所示。

2）筛选：筛选提供在线二次检索的功能，即在现有的返回结果中保留或者剔除符合检索式的语料实例，得到二次检索结果。筛选的检索式同一般的BCC检索式一致。

3）下载：下载检索或统计结果，登录用户可以下载更多的结果。

4）高级：可以设置返回结果的显示形式；可以随机生成实例，如设定上下文显示字数，设定是否以句形式显示结果等。

5）全文：点击该按钮可以查看检索实例更多的上下文。

图5.　统计结果页

BCC语料已经过自动分词和词类标注，并加工成为全文检索索引，制作成为“现代汉语词汇历时检索系统”。历时检索第一代系统（荀恩东等 2015）于2012年上线，使用的语料来源于1946年到2012年的《人民日报》语料，检索结果以年为单位，显示检索式的次数，并以可视化的方式呈现给用户，如下页图6所示。点击图中的每个柱形可以具体显示该年的实例结果。语料仅有分词标注，无词类标注。

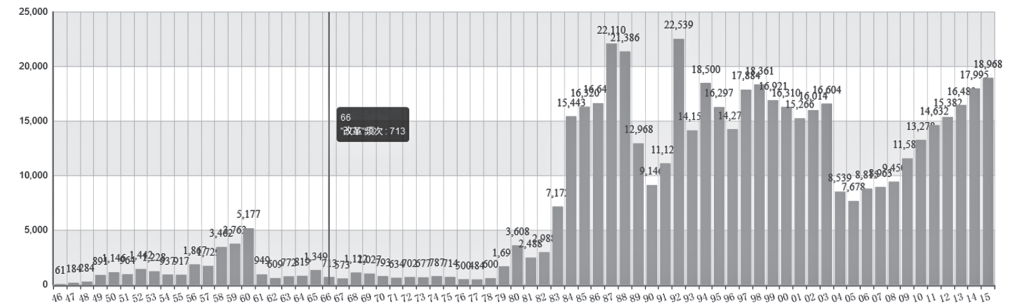


图6. 检索式“改革”频次历时结果

2015年底，历时检索第二代系统上线。历时语料库在分词的基础上增加了词类标注，在保留原有用户体验的同时开始提供多模态检索功能。在该功能的支持下，用户可以在对任意词串（不限于词）进行检索之外对词类串和字符词类混合串进行检索。如图7所示。第二代历时语料库在国内外引起强烈反响，为语言学和许多社会科学领域的相关研究提供了很大的便利（Rao & Xun 2015）。

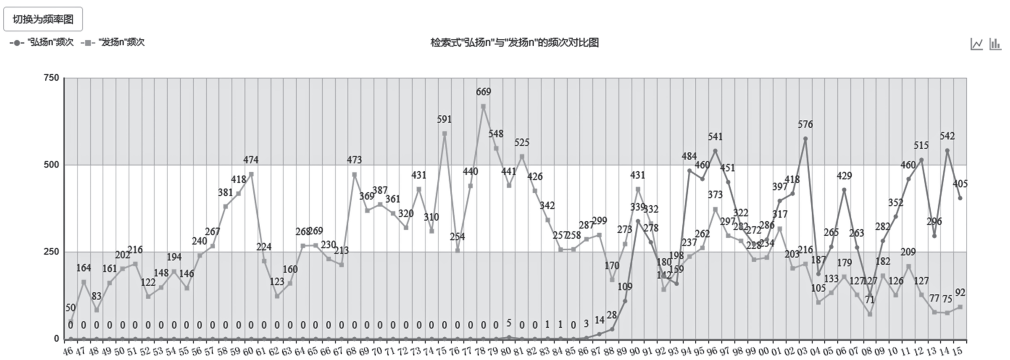


图7. 检索式“弘扬n”与“发扬n”频次历时对比

三、BCC检索式

一些语料库采用交互式生成检索式的设计，即以输入查询和界面控件设置相结合的方式进行查询。这种方式有利有弊：如果设置项少，往往限制检索功能的发挥；如果功能复杂，设置项过多，便会影响用户的使用体验。

BCC设计简洁，在界面中没有各种复杂的控件，选定语料后，输入符合语法的检索式可以直接搜索查询。检索式的设计也需要平衡考虑：一般来讲，检索式的语法直接影响语料库功能和用户友好性。复杂的检索式设计可以支持强大的检索功能，但是会对用户学习和使用造成负担。例如，检索系统采用正则表达式的方式，虽然语句标准、功能强，但是不易理解，需要付出更多的学习代价。

为平衡用户友好性和检索功能，BCC设计实现了一种简单的查询语言，即BCC检索式。选取语料，输入检索式，即可查询符合检索式的语言片段。在BCC中，这些语言片段通常是一个（检索式中无离合符号）或者两个（有离合符号）连续出现的字串或者词串。如果检索式中不包含标点或表示标点的词性，检索结果将会限定在一个标点句中。

BCC检索式主要由汉字串（或者词串）、属性符号、通配符、集合符号、离合符号、属性约束符号、空格或“+”组成。

汉字串（或者词串）：

汉字串不需要给出分词信息，如果在相邻汉字串之间加入空格或者“+”，BCC也会自动过滤掉这两个符号，检索时作为一个整体字符串匹配。这与通常搜索引擎的使用体验不同。例如：

检索式：“与其说是”，检索包含“与其说是”的实例；

检索式：make a promise，检索包含make a promise的实例；

检索式：“提高 水平”或“提高+水平”，等同于“提高水平”；

检索英语语料库时，按照词的原始形态进行匹配处理，BCC不对单词的形态做处理，也不对单词的大小写进行处理。

属性符号：

属性符号是指在标注语料库中，字、词或短语所具有的类型标记。它可以是词性符号、短语类型符号、语义符号等，具体与标注语料库的内容和标注体系相关。目前，BCC中的属性符号主要是指词性符号和短语类型符号，而短语类型符号仅用于具有短语标注的树库中。

BCC中汉语语料库采用北京大学的词性体系，英语语料库采用美国宾州大学的词性体系，具体见附录2。限于语料性质或加工工具，并不是所有BCC语料都有词性信息，即除现代汉语、英语、法语外的其他语种的语料都没有进行词性标注。

为了便于使用，在汉语（除古汉语）和英语（除标准词性标注集）语料库中，同一属性可以用多个等价符号表示，例如：汉语动词可以用v、V、verb、Verb、VERB等不同符号替代。例如：

检索式：“不得不说v”，检索“不得不说”后接属性符号v（动词）的实例。以下形式都是等价的：“不得不说 v”（加空格）、“不得不说 Verb”、“不得不说V”。

检索式：“w 吃 n w”，检索属性符号w（汉语标点符号）后接“吃”再接属性符号n（名词）和属性符号w（汉语标点符号）。这里n和w之间要用空格分隔，表示是两个独立的属性符号。

检索式：v dt problem，在英文语料库中检索属性符号v（动词）后接属性符号dt（定冠词）后面再接单词problem的实例。

通配符：

与属性符号相比，通配符代表更宽泛的语言单元，说明某个位置是任意一个汉字或单词。BCC支持三种通配符“.”、“~”和“@”。

1）“.”在汉语语料库中，表示任意一个汉字，检索其他语种的语料库时，表示一个单词。例如：

检索式：“洗...澡”表示检索“洗澡”中间插入三个字的实例。

检索式：make..promise表示检索make promise中间插入两个单词的实例。

2）“~”是汉语语料库专用符号，表示任意一个词。其他语种的语料库则是使用通配符“.”表示任意一个词，而“~”只作为普通符号。BCC限制该符号只能在检索式中出现一次，不支持多个连用的情况。例如：

检索式：“洗~澡”表示检索“洗”后接任意一个词再接“澡”的实例。该通配符通常用来统计上下文词语的搭配情况。该检索式的统计结果见图8示意。

检索式：“w 吃 ~ w”，表示属性符号w（标点），后接“吃”再接任何一个词，再接属性符号w（标点），即检索“吃”后接一个词并单独成小句的实例。该检索式的统计结果见下页图9示意。

图8. 检索式“洗~澡”的统计结果

：吃柚子，	348	：吃梨。	343
，吃东西，	289	、吃后，	264
，吃枣仁。	253	，吃核桃。	251
。吃苹果，	245	。吃芒果，	233
。吃山药，	233	。吃百合，	232
。吃胡椒，	231	，吃土豆。	224
，吃荔枝。	220	。吃葡萄，	220
“吃人”	164	→吃醋。	160

图9.　检索式“w 吃 ~ w”的统计结果

3）“@”在各种语料库中，都表示任意词，该符号往往用于统计的功能，即统计该位置对应不同词性出现的频次。BCC限制该符号只能在检索式中出现一次，不支持多个连用的情况。例如：

检索式：“w 吃 @ W”，在检索实例时，结果同“w 吃 ~ W”，不同的是检索式的统计结果，如图10所示。

w+吃+n+W	4546	w+吃+v+W	699
w+吃+y+W	370	w+吃+a+W	326
w+吃+u+W	276	w+吃+r+W	222
w+吃+m+W	96	w+吃+f+W	48
w+吃+Ag+W	48	w+吃+d+W	46
w+吃+nz+W	46	w+吃+Ng+W	41
w+吃+q+W	40	w+吃+nr+W	38
w+吃+ns+W	31	w+吃+z+W	28

图10.　检索式“w 吃 @ W”的统计结果

集合符号“[]”：

在符号“[]”内，可以写多个汉字字符串、单词或者词性，之间用空格分隔，表示可以对应括号内任意一项。例如：

检索式：“[美丽 靓丽]”表示检索包含“美丽”或者“靓丽”的实例。

检索式：“v[上来 下去]”表示检索动词后面接着“上来”或者“下去”的实例。

检索式：“打击[n vn]”表示检索动词“打击”后面接着名词n或者动名词vn的实例。

离合符号“*”：

通常，BCC检索式对应连续的字符串。引入该符号的目的是描述语言中的各种离合现象。使用该符号的一般形式为：“检索式1*检索式2”，表示在句子内（对于汉语是小句内），检索符合“检索式1”后接其他成分再接“检索式2”的实例。要注意离合表达的顺序和检索所表达的范围。BCC中限制该符号最多只能出现一次，即不支持多个语言片段连续出现的检索功能。例如：

检索式：“洗*澡”是检索“洗澡”离合出现的情况。

检索式：“见*面”是检索“见面”离合出现的情况。

属性约束符号“/”：

该符号作用于一个检索式，后面给出属性符号，约束检索式对应的实例所具有的特定属性，比如“检索式/属性符”的格式。例如：

“./Vg人”表示单音节动词后面接“人”的实例。

“打./v”表示以“打”字开头的双音节动词。

空格或者“+”：

除了“/”外，一般情况下，不同表达内容之间需要用“+”或者空格分隔，如果在没有歧义的情况下，也可以连接在一起。例如：

检索式：“我想吃n”，检索“我想吃”后面紧接着一个名词的语言实例。与通常搜索引擎含义不同，在BCC检索式中，有歧义表达时，需要加空格，起到分隔的作用。如“我想吃n”在汉字串后接一个半角的属性符号“n”没有歧义，所以在检索时可以省略空格。

例如：“我们 大家”等同“我们大家”，“打击 n”等同“打击n”。在检索式中，如果连续出现两个或多个词性标记，或在外文语料库检索时，单词之间要用空格分隔。例如：“一q n”，表示检索“一”后面连着一个量词，量词后面是一个名词的实例。多个词性相连时，用‘ ’（空格）分隔。另外，空格在集合符号“[]”中使用，用来分隔多项内容；配合“/”使用，可以用来表示词边界。

四、结语

BCC语料库为语言本体研究提供数据和技术支持，在大数据背景下，可以证实、证伪或者发现语言现象；BCC作为语言应用开发的基础平台，为信息抽取、构建知识图谱、语言自动分析等提供便利；同时，也为语言教学研究提供统计数据和实例支撑等。

BCC是动态发展的，本文没有提供现有BCC在线服务语料的细节信息，最新的语料和规模可以通过BCC的“自定义”功能或在线说明文档获得。今后BCC将会纳入更多的语种、更大规模的数据、更多形式的语料，从文本语料向多模态语料拓展，从语法属性为主的检索向语义信息检索方面发展。BCC的建设目标是打造一个大型知识库。从一个语料库发展成为一个知识库，这不仅能支持语言本体研究，也能为语言相关应用的研发提供知识支撑。

参考文献

Rao, G. & E. Xun. 2015. Words and characters in official newspapers since the founding of the PRC: *Guizhou Daily and People's Daily* as examples [J]. *International Journal of Knowledge and Language Processing* (2): 23-33.

黄昌宁、李涓子，2002，《语料库语言学》[M]。北京：商务印书馆。

熊文新，2015，《语言资源视角下的语料库建设与应用研究》[M]。北京：外语教学与研究出版社。

荀恩东、饶高琦、谢佳莉、黄志娥，2015，现代汉语词汇历时检索系统的建设与应用[J]，《中文信息学报》(3)：169-176。

俞士汶、段慧明、朱学锋、孙斌，2002，北京大学现代汉语语料库基本加工规范[J]，《中文信息学报》(6)：49-64。

俞士汶、朱学锋、段慧明，2000，大规模现代汉语标注语料库的加工规范[J]，《中文信息学报》(6)：58-64。

附录1. 检索式示例

构词

../v	双音节动词
打./v	以“打”为首的双音节动词
.性/n	以“性”为结尾的双音节名词
./v 货/n	单音节动词，后接名词“货”
./v ./n	单音节动词，后接单音节名词

搭配

讨论n	“讨论”后邻“名词”
~讨论	任意词后邻“讨论”
提高*n	“提高”后面离合接名词
@的提高	任意后接“提高”
提高 ../[vn n] w	提高句尾后接双音节名词或者动名词

离合

洗*澡	“洗”后接“澡”
洗.澡	“洗澡”中间有一个字
洗..澡	“洗澡”中间有两个字
澡*洗	“澡”后接“洗”

句型

是*的[。？ !]	“是”后接“的”，“的”后面是“。”或“？”或“！”
是*的w	“是”后接“的”，“的”是句尾
把*v[上下起].	“把”后接动词，动词后邻“上”或“下”或“起”，后面再接一个字
被*v[上下起]来	“被”后接动词，动词后邻“上来”或“下来”或“起来”
被n v一下	“被”后邻名词、动词和“一下”
被n v一下w	“被”后邻名词、动词和“一下”，“一下”是句尾

定界

w吃	“吃”做句首
w吃.W	“吃”做句首的二字短句
[，。]吃W	“吃”是单字短句，句首前标点“，”或“。”，句尾符号不限
吃W	“吃”做句尾
[，。]吃[，。]	“吃”是单字短句，句首前标点“，”或“。”，句尾符号是标点“，”或“。”

构式

a不到哪里去	形容词后邻“不到哪里去”
还n尼	“还”后邻名词，再接“尼”
v就v	动词后邻“就”，再接动词
v不着	动词后邻“不着”
v不到	动词后邻“不到”
n连n都	名词后邻“连”、名词、“都”
n连n也	名词后邻“连”、名词、“也”
有一种n叫n	“有一种”后邻名词、“叫”、名词
非[a v n]不可	“非”后邻形容词或动词或名词，再接“不可”
活活[a v n]死人	“活活”后邻形容词或动词或名词，再接“死人”
放着n不v	“放着”后邻名词、“不”、动词
v不过n	动词后邻“不过”、名词
n说起来v	名词后邻“说起来”、动词

附录2. 词性标注集

汉语词性列表

词性编码	词性名称	词性编码	词性名称	词性编码	词性名称	词性编码	词性名称
Ag	形语素	I	成语	o	拟声词	vn	名动词
a	形容词	J	简称略语	p	介词	w	标点符号
ad	副形词	K	后接成分	q	量词	x	非语素字
an	名形词	l	习用语	r	代词	y	语气词
b	区别词	m	数词	s	处所词	z	状态词
c	连词	Ng	名语素	Tg	时语素	un	未知词
Dg	副语素	n	名词	t	时间词	h	前接成分
d	副词	nr	人名	U	助词	g	语素
e	叹词	ns	地名	Vg	动语素	nz	其他专名

（待续）

（续表）

词性编码	词性名称	词性编码	词性名称	词性编码	词性名称	词性编码	词性名称
f	方位词	nt	机构团体	V	动词	vd	副动词

英语词性列表（缩减版）

词性编码	词性名称	词性编码	词性名称	词性编码	词性名称
WRB	Wh-副词	PRP	人称代词	DT	冠词
WP	Wh-代词	POS	所有格	CD	数词
WDT	Wh-限定词	PDT	后接成分	CC	连词
W	标点	NN	名词		
VB	动词	MD	情态词		
UH	语气词	LS	名语素		
TO	to	JJ	形容词		
SYM	符号（# $）	IN	介词		
RP	叹词	FW	外来语		
RB	副词	EX	there		

英语词性列表（完整版）

词性编码	词性名称	词性编码	词性名称	词性编码	词性名称	词性编码	词性名称
CC	并列连接词	MD	情态动词	RBR	副词，比较级	VBP	动词，非第三人称单数现在式
CD	基数	NN	名词，可数或不可数	RBS	副词，最高级	VBZ	动词，第三人称单数现在式
DT	限定词	NNS	名词，复数	RP	小品词	WDT	wh-限定词
EX	存在型there	NNP	专有名词，单数	SYM	符号（数学或科学）	WP	wh-代词
FW	外文单词	NNPS	专有名词，复数	TO	to	WP$	所有格wh-代词

（待续）

（续表）

词性编码	词性名称	词性编码	词性名称	词性编码	词性名称	词性编码	词性名称
IN	介词	PDT	前位限定词	UH	感叹词	WRB	wh-副词
JJ	形容词	POS	所有格结束词	VB	动词，基本形态	#	# 符号
JJR	形容词，比较级	PRP	人称代名词	VBD	动词，过去式	$	美元符号
JJS	形容词，最高级	PP$	物主代词，所有格代名词	VBG	动词，动名词/现在分词	.	句点
LS	列表项标记	RB	副词	VBN	动词，过去分词	,	逗号
:	冒号，分号	)	右括号	‘	左单引号	’	右单引号
(	左括号	“	双引号	“	左双引号	”	右双引号

通讯地址：100083 北京市北京语言大学大数据与语言技术研究所

《中国语境下的语料库语言学》述评

北京外国语大学 徐秀玲

Bin Zou, Simon Smith & Michael Hoey (eds.). 2015. *Corpus Linguistics in Chinese Contexts*. Basingstoke: Palgrave Macmillan. xxiii+203pp.

《中国语境下的语料库语言学》一书收录了多项基于语料库的中介语研究、英汉对比、语言教学等方面的最新成果。以往的语料库语言学论著大多关注欧美学者的研究成果，专辑讨论中国语境下语料库研究的英文论著还不多见。下文将介绍该书主要内容，并作简评。

1. 内容概述

全书由导言和9篇论文组成。前3篇与汉语或英汉双语语料库研究有关，后6篇与英语语料库研究有关。

导言由Wenzhong Li（李文中）和该书主编之一Simon Smith共同撰写。该节首先介绍了20世纪20年代以来的汉语语料库研究，并回顾了中国学者基于英语语料库所作的研究，尤其是英语学习者中介语研究。语料库研究在中国发展势头迅猛，该文集也将进一步推动语料库在中国语言教学与研究中的应用。

首篇论文“词汇触发：一个适用于英汉语言的语料库语言学假设背后的心理语言学理论”由Michael Hoey和Juan Shao（邵娟）执笔。作者指出，词汇触发（Lexical Priming）是一个基于语料库语言学和心理语言学研究而产生的理论，它揭示了词汇在使用模式中的共现行为。具体来看，任何词都受到以下要素的触发：与其他词共现（搭配），与特定的语义类别共现（语义关联），与特定的语法类别共现（类联接），与特定的语用功能共现（语用关联），等等。作者认为词汇触发背后有其心理学机制，虽然该理论是基于英语分析提出的，但应该同样适用于汉语。作者通过分析汉语语料发现，汉语词汇使用也存在搭配、语义关联、类联接和语用关联等现象。由此作者得出结论，尽管在语言类型上英汉语差别较大，但从词汇和心理语言学的角度来看，两者具有诸多相似之处。这一发现对语言描写和教学均有启示。假如两种不相关语言的词汇特征可以用同一组概念和步骤来描

写，那么我们离普遍语言描写（universal description of language）又近了一步。作者还指出，目前语言教材过于强调英汉差异，倘若能多关注两种语言的共性，将有助于增强词汇学习效果，激发学生们的语言学习动机。

Richard Xiao（肖忠华）的论文“对比语料库语言学：英汉语言对比研究”基于4个英汉语书面语和口语语料库，考察了英汉语中的被动句和量词。作者对比了英语中的be、get被动句和汉语中的“被”、“给”、“为……所”、“叫”、“让”几种被动句，发现英汉被动句在使用频率、语用含义、句法功能、语体分布等方面均表现出明显差异。在接下来的量词研究中，作者认为与汉语一样，英语中也存在量词。对比分析表明英汉语中的量词在使用频率、使用范围、语义类别、句法结构、语体分布等方面存在诸多不同。基于以上两个英汉对比个案分析，作者提出了一个“对比语料库语言学”的整合模型，并指出语料库语言学的对比本质（如主题词分析、搭配计算都涉及频数比较）与对比语言学十分契合。对比语料库语言学能够把对比分析与语料分析有机地结合起来，为语料库语言学、对比语言学、翻译研究和二语习得研究提供了一个共同的研究框架。

Adam Kilgarriff、Nichole Keng和Simon Smith的论文“借助Sketch Engine学习汉语”探讨了如何使用语料库技术进行汉语教学。过去几年里汉语教学得以迅速发展，如何将英语教学中使用的语料库技术迁移到汉语教学中来，以便汉语教学也能从中获益，值得学界探讨。文章详细介绍了如何将在线语料库检索系统Sketch Engine应用于汉语教学。据作者介绍，Sketch Engine预装了zhTenTen（20亿汉字）、Chinese Gigaword（2亿汉字）等多个汉语语料库，供用户查询使用。该系统具有以下几大功能：基本检索、检索包含某个汉字的词、通过词汇素描功能（word sketches）对检索词的语法特征及搭配特征进行汇总、使用近义词功能（thesaurus）列举与检索词意义相近的词、利用素描差异功能（sketch differences）进行近义词辨析，等等。鉴于Sketch Engine的强大功能，作者认为它能有效地帮助学生学习汉语词汇的搭配及句法结构。

Maocheng Liang（梁茂成）的论文“语篇内部的短语学特征分布模式探索：以学术论文为例”考察了学术文本各部分的短语学特征。囿于语料库分析软件的局限，以往语料库语言学领域的短语学研究大多不关注短语在语篇不同部分的使用差异，而体裁分析领域的研究虽然关注文本的不同部分，却很少考察文本内部的短语学特征。作者认为短语学特征在同一文本的不同部分会有所不同，相应研究工具的开发十分必要。为解决这一问题，作者所在的北京外国语大学中国外语教育研究中心语料库团队开发了专用语料库分析工具TextSmith Tools，可以将文本自动切分成多个部分，并将文本的某一部分作为观察语料库，其余部分作为参照语料库，实现文本不同部分之间相比较的目的。作者借助该工具对*Applied Linguistics*期刊的学术论文进行了分语步分析，发现论文各主要部分具有明显

不同的短语学特征。这些特征与论文各部分的语篇功能和内容密切相关。当然，TextSmith Tools也有自身的局限性，它对同一体裁、具有相同结构的同质文本分析效果最佳，可能不太适合分析不同体裁、结构差异较大的异质文本。

Anping He（何安平）的论文“短语学的语料库教学加工在外语教学中的应用：一个实施案例”介绍了一个将语料库资源和技术应用于课堂教学的实证研究。作者指出，直接将语料库应用于教学存在诸多弊端，最好能对语料库进行教学加工，即将语料、调查结果乃至语料库技术转化为日常语言教学的资源和教学手段。另一方面，Sinclair的短语学理念强调词汇教学应注重短语而不是孤立的单词，然而如何选择合适的短语以及构建相应的教学平台是一个亟待解决的问题。为此，作者带领团队试图利用语料库的方法提取与教学相关的短语及其典型特征，并将之转化为多媒体教学形式。语料库资源的教学加工可分为3个阶段：1）提取目标短语。基于英语教材语料库和常用词表，选取与教材相关且最常用的动词词组。2）分析短语分布及使用模式。按照不同的语言学习阶段，总结有关短语的典型用法。3）将语料库成果转化为多媒体教学资源。将阶段2的成果设计成多模态教学平台，既包括相关文本语料，又有与文字匹配的图像、视频和音频，并提供配套的语料库辅助练习。

Wangheng Peng（彭望蘅）的论文“基于语料库的西交利物浦大学学生名词使（误）用研究”是一项中外合作办学环境下的学习者中介语研究。作者首先简要介绍了西交利物浦大学学习者语料库（the XJTLU Written English Corpus，XWEC）的创建情况，接着基于两个英语本族语者语料库（BNC、BAWE）和两个中国英语学习者语料库（SWECCL、XWEC）考察了中国大学生英语写作中的名词可数性问题。鉴于英语学术写作在西交利物浦大学的重要性，作者选取了学术论文中15个常用的不可数名词，如advice、evidence、research、software。研究表明，无论是西交利物浦大学学生还是中国其他高校学生，把这些词当作可数名词使用的比例远远高于本族语者，说明中国大学生英语写作中存在一定程度的名词误用。有意思的是，本族语者写作中也使用一些名词的复数形式，如researches，但不一定是误用，而是有特定的含义。因此，作者提醒学生使用词典时需要谨慎，语言使用有特例，并且语言也在不断变化。语料库中大量真实的语言用例可以帮助学生学习更鲜活的语言。

Bin Zou（邹斌）和Wang Pengheng（彭望蘅）的论文“中英合作大学学术英语教学环境下的连词使用——一项基于语料库的研究”同样也是一项学习者中介语研究。作者利用西交利物浦大学学习者语料库（XWEC）、中国学生英语口笔语语料库（SWECCL）、英国学术书面语英语语料库（BAWE），分析了中国大学生英语写作中6种连词（包括因果、附加、对比/让步、时间、总结、介词及短语连词）的使用情况。研究发现，西交利物浦大学学生从入学到大学一年级结束的一

年时间里，连词使用水平取得了显著提高，接近于本族语者水平。而中国其他高校学生从大一到大四的几年时间里，连词使用提高幅度相对较小，与本族语者还有一定差距。具体表现为，西交利物浦大学学生已经学会在英语写作中使用较多的正式连词，如therefore、furthermore、however等；而中国其他高校学生在大四时仍然使用大量的非正式连词，如so、what’s more、but等。作者推测，这种差异可能与两种办学方式的英语教学环境有关。

Haiping Wang（王海萍）、Yuanyuan Zheng（郑媛媛）和Yiyan Cai（蔡懿焱）的论文“语料库方法在高级英语阅读和语篇分析技巧教学中的应用”是一项将语料库技术运用于阅读教学的实证研究。作者认为，基于语料库的语篇分析有助于提高学生自上而下的阅读能力及语篇分析技巧，并通过实验来验证这一假设。研究设立对照组和实验组，对照组使用传统的阅读教学方法，实验组则采用语料库方法。具体步骤如下：在教师指导下，学生收集与课本单元主题相关的文章，创建一个阅读文本语料库；然后逐篇标注文章的体裁、主题句、组织结构、衔接手段等信息；最后总结标注信息并在课堂上作演示。教学实验持续一个学期，实验前后分别对实验组和对照组的学生进行阅读能力测试。研究发现，使用语料库方法的学生阅读成绩明显提高，而使用传统方法的学生阅读成绩几乎没有变化。作者认为，学生参与创建阅读文本语料库，并对文本作详细的语篇信息标注，可以帮助他们建构内容图式，提高语篇分析技巧和阅读能力。调查问卷和访谈结果也表明，多数学生对语料库方法持肯定态度，阅读能力和阅读兴趣均有所提高。

Zhaoyang Mei（梅朝阳）、Ren Zhang（张韧）和Baixiang Yu（于柏祥）的论文“《纽约时报》对中国军事报道的评价分析”利用Martin的评价理论，从批判话语分析的角度分析了美国主流媒体对中国军事报道中的评价资源。作者收集了20篇《纽约时报》关于中国军事发展、中美军事关系、亚丁湾护航任务的新闻报道，并根据评价系统的三个子系统——态度、介入和级差对新闻报道中的评价资源逐一进行手工标注。作者结合定量与定性方法，分析了评价资源的分布特点。研究发现，该类新闻报道看似客观，却暗含着丰富的评价资源。在态度系统中，判断资源和鉴赏资源使用较多，情感资源相对较少；判断资源中又以积极的能力资源最多，可能与中国军事实力日益增强有关。在介入系统中，归属资源在数量上占绝对优势，主要是通过引用外界权威声音，使报道显得客观公正。在级差系统中，语势资源数量明显多于聚焦资源，多用于加强或锐化观点。作者认为，《纽约时报》有关中国军事的新闻报道并没有如实地反映事实，而是受到了新闻机构的立场和意识形态的影响。

2. 简评

纵览全书，该书主要有以下几方面特点：

第一，理论与应用兼顾。该书9篇论文既涉及语料库理论的探讨，例如将“词汇触发理论”扩展到汉语；又包括语料库在教学中的应用，或将语料库直接用于课堂教学，或对语料库进行教学加工转化成多媒体资源，或通过创建、标注语料库提高阅读能力。目前，中国的语料库辅助语言教学还远未到达预期的效果（何安平 2010），该书关于语料库应用于教学的深入探讨具有现实指导意义。

第二，研究内容丰富。该书收录的论文篇数虽不算多，但覆盖领域广泛，包括英汉对比、中介语分析、语言教学（词汇、短语、阅读教学）、体裁分析、媒体话语分析等。分析对象也比较全面，从单词、词类、短语、句法到篇章，均有涉及。

第三，研究工具新颖。该书的研究者十分重视语料库新技术。需要特别指出的是在线语料库检索系统Sketch Engine和Maocheng Liang（梁茂成）开发的专用语料库分析工具TextSmith Tools。新技术新工具不仅打破了以往的语料库技术壁垒，带来数据检索和分析的变革，也改变了我们的研究视角（李文中 2014）。

第四，研究方法多样。语料库语言学领域的研究以分析语料为主。该书的研究除了语料库分析方法，有的还运用了实验法和问卷调查法进行三角验证，使得研究结果更为可靠。这也说明，语料库方法与其他方法的结合具有可行性，对今后相关研究具有借鉴意义。

除了内容上值得一读外，该书语言清晰易读。当然，书中也难免会有一些笔误：第36页表2.1中的L类别应为Mystery and detective fiction而不是Adventure fiction；第57页第19行有位作者的姓氏Salkie误拼为Salki；第79页第14行的“Liu (2011)”应改为“Liu (2012)”；第129页第3行the other hand前缺少On。但瑕不掩瑜，这些细微的瑕疵无法掩盖该书的学术意义和教学价值，该书值得一读。

参考文献

何安平，2010，语料库的“教学加工”发展综述 [J],《中国外语》(4)：47-52，108。
李文中，2014，在变化中成长的语料库语言学 [J],《解放军外国语学院学报》(1)：3-4。

通讯地址：100089　北京市北京外国语大学中国外语教育研究中心

English Abstracts

The impact of translation upon modern Chinese: The case of inferential markers

QIN Hongwu, LIU Dandan & DU Xiaoying (1)

Based on a diachronic comparable corpus, this paper observes the role played by translation in the development of the discourse markers in modern Chinese, the focus being on inferential markers. Findings indicate that modern Chinese resorts more to discourse markers than ancient *baihua* does in framing the discourse, and Chinese discourse organization has undergone great changes, some of which have much to do with translation. It is also found that the language contact between translationese and Chinese yields highly isomorphic patterns, and enhances the frequency of their equivalents in Chinese. The study claims that Chinese only selectively accepts the influence of translation, which does not induce the substantial change. Therefore, the changes in the use of discourse markers come from the interaction between translation and modern Chinese.

Critical issues on corpus-driven machine dictionary creation

CAO Rong, PU Jianzhong & HUANG Jinzhu (13)

Corpus-driven linguistics tries to get rid of existing theories and to find out new realities which reflect the nature of language use. According to this idea, we try to re-examine several critical issues concerning machine dictionary creation: what is the core of language description, what is the fundamental unit, what is the appropriate model of explanation, etc. This paper proposes a new idea of machine dictionary creation which takes meaning as the core of description, language use (or text) as the essential source of meaning, and the extended unit of meaning as the fundamental unit of description, and makes use of enumerating and regular expressions to explain meaning.

Revisiting standards for the construction of Chinese interlanguage corpora

ZHANG Baolin (21)

The construction of Chinese interlanguage corpora has been gaining momentum in recent years, yet it is hampered by substantial problems, such as the lack of corpus construction standards. However, such standards have not received enough scholarly attention and remain largely unexplored. The present study, therefore, proposes some suggested corpus construction standards, with special reference to the content and methods. It is hoped that scholars can realize the

significance of this study and carry out extensive research to promote the standardization, and systematization of corpus construction, which will contribute to the teaching and research of Chinese.

Corpus linguistics and bibliometrics: Intersection and complementarity

Corpus linguistics and bibliometrics both bear on quantitative methods. The former is characterized by the stress on actual language use and the synthesis of relevant linguistic theories into research. Bibliometrics on the other hand focuses on the retrieval of information from academic discourse in science and technology, journal articles in particular. Despite their differences, the two fields have much in common in such aspects as the basic unit of analysis, similar requirements for research data and the genre of journal articles as a common research object. From an interdisciplinary perspective, the two fields are likely to learn from and enrich each other.

Mapping the intellectual structure of corpus linguistics: A co-word analysis (1971-2015)

This study aims to use co-word analysis to map the major research areas and their developmental trends in the field of corpus linguistics from 1971 to 2015. Hierarchical clustering analysis, multidimensional-scaling analysis and strategic diagram analysis were conducted on the word co-occurrence matrix constructed from the 23,078 literature records collected from LLBA (Linguistics and Language Behavior Abstracts) database. Ten major research areas were identified, including the study of language features, natural language processing, discourse analysis, historical linguistics, sociolinguistics, lexicography, child language acquisition, second language acquisition, text analysis, translation and contrastive studies. The results also indicate that corpus-based studies on language features, historical linguistics, sociolinguistics, lexicography, natural language processing and discourse analysis are relatively mature. In contrast, the application of corpora in the fields of child language acquisition, translation and contrastive studies, second language acquisition and text analysis needs to be improved.

A data-based study of L2 pragmatic knowledge: The case of state transition copulas in Chinese EFL learner English

Within the framework of the co-selection theory, this study seeks to examine the pragmatic knowledge of state transition copulas BECOME, GO and GET used by Chinese EFL learners in comparison to those used by natives. The results show that: 1) state transition copulas both in native English and Chinese EFL learner English have their specific pragmatic features; 2) their colligational patterns are not very different, but their semantic prosody presents marked deviation . The analysis illustrates that: 1) semantic prosody, which is the most abstract pragmatic knowledge, restricts the selecting of collocations and syntactic structures 2) semantic prosody is difficult to perceive for Chinese EFL learners, and their collaborative priming of the native tongue's pragmatic knowledge and the lexical and syntactic knowledge of second language causes the deviation.

An overview of the research on grammatical error automatic detection for English learners

Automatic detection of grammatical errors in learners' English is an important issue in the field of computational linguistics. This paper describes the specific features of the learners' grammatical errors. Then by reviewing relevant literatures, it sums up what the future study can learn from and what they should avoid. Finally, four improvements are discussed, hoping that this paper could be helpful to future research.

Multifactorial analysis in linguistic studies

This paper addresses multifactorial analysis, an emerging approach in linguistic studies. Multifactorial analysis falls into two phases: feature analysis based on manual calculation of linguistic features and modern multifactorial analysis relying on computerized statistical models. Previous applications of multifactorial analysis in linguistic studies concern lexical and syntactic dimensions, in both native language and interlanguage. Though in its primitive stage, multifactorial analysis has displayed a great feasibility in linguistic studies.

The construction of the BCC Corpus in the age of Big Data

.....*XUN Endong, RAO Gaoqi, XIAO Xiaoyue & ZANG Jiaojiao* (93)

Beijing Language and Culture University Corpus Center (BLCU Corpus Center, BCC) Corpus is a large full-text retrieval corpus with multiple languages, including Chinese and other languages as well. BCC is an online data system with a size of about ten billion words, ideal as a data source for studies in linguistics as well as applied linguistics. BCC search queries support wildcards, splittable words, as well as character-based, word-based and POS-tag based expressions. This paper introduces the BCC Corpus in detail, including the construction of the corpus and the design of the search engine, with a particular focus on the query language and tips for corpus search.

语料库语言学

CORPUS LINGUISTICS

要　目

责任编辑：毕　争
执行编辑：解碧琰
封面设计：外研社设计部

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

unipus

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 978-7-5135-7762-5

定价：12.00元