《中国学术期刊网络出版总库》及CNKI系列数据库入选期刊

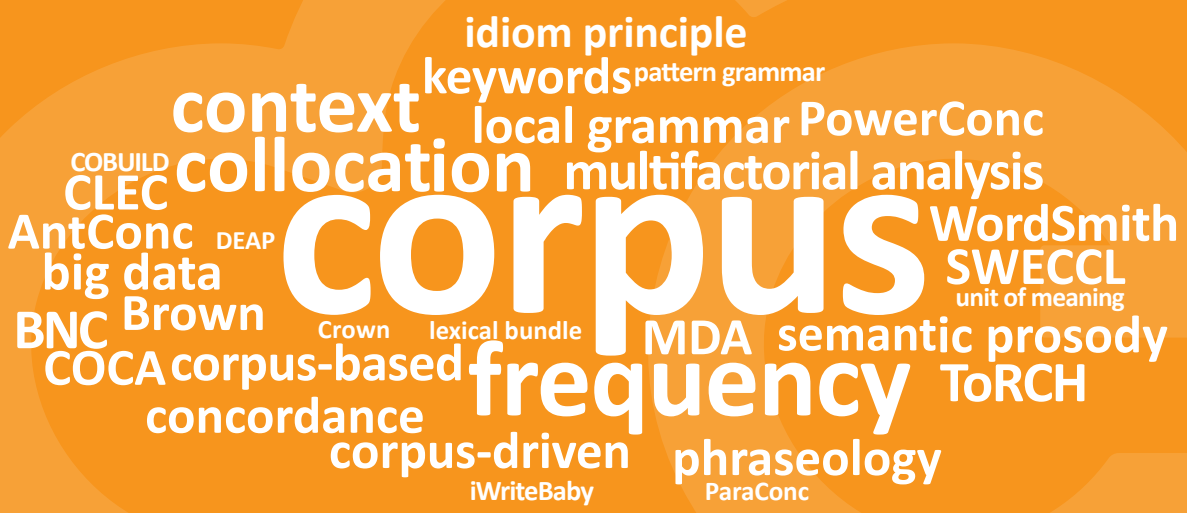
第17辑 二〇二二

语料库语言学

CORPUS LINGUISTICS

第17辑
2022

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



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

语料库语言学

外研社

语料库语言学
（半年刊）

Corpus Linguistics
(Biannual)

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

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

刊名题字：崔希亮
主　　编：许家金
责任校对：王　斌、郝美佳

Journal Name Calligraphy: Cui Xiliang
Editor: Xu Jiajin
Proofreaders: Wang Bin & Hao Meijia

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
　　中国外语与教育研究中心
　　《语料库语言学》编辑部（100089）

语料库语言学

CORPUS LINGUISTICS

2022 年　第 17 辑

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

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

《语料库语言学》

2022年　第9卷　第1期

目　录

语境共选

研究论文

研究综述

研究开发

CORPUS LINGUISTICS

Volume 9, Number 1, 2022

Table of Contents

“把”字句下位构式原型语义及构式化程度对比研究*

北京外国语大学 钱一华 熊文新

提要：本文在构式语法视角下，采用基于语料库的数据统计方法，对“把”字句下位构式的原型语义以及构式化程度展开对比，并在此基础上探讨“把”字句的语义结构。研究发现：下位构式均继承了上位“把”字句原型语义中的“致使-结果”义。其中，对“致使-结果”凸显度较高的构式，构式化程度也较高，并居于“把”字句中的核心位置；“把”字句具有表达“致使-位移”和“致使-变化”的双原型结构，并沿不同的下位构式形成两条相对独立的扩展路径。

关键词：语料库、“把”字句、构式语法、原型语义

1 引言

“把”字句作为现代汉语中的一种特殊句式，多年来一直是语法学界关注的焦点，关于其句式语义的探讨也从未间断。如王力（1943）最早提出的“处置说”，以及其他学者在此基础上提出的“广义处置说”（马真 1981；宋玉柱 1981）、“主观处置说”（沈家煊 2002）、“原型处置说”（王璐璐 2013）等。叶向阳（1997，2004）、郭锐（2003）则提出“致使说”，认为“把”字句并非强调主语对受事的处置行为，而是一种致使，即前一事件的发生导致了后一事件的发生。与“处置说”相比，“致使说”弱化了主语的意志性，强化了事件的结果性。施春宏（2010）进一步在致使类句式群中对比和观察“把”字句，将其概括为“凸显致使结果”的致使性句式。另一些学者则从原型范畴理论的视角出发，将“把”字句看作由典型空间位移扩展至其他语义类型的一个连续统（张伯江 2000；张旺熹 2001；吕文茜 2015）。

“把”字句句式语义的复杂性源于其内部结构的异质性。构式语法视角下，形式和意义/功能之间具有配对性，不同的形式必然对应于不同的意义，即构式无同义原则（Goldberg 1995）。“把”字句内部仍包含多类不同的VP结构，如“V（在

* 熊文新为本文通讯作者。
作者贡献：
钱一华：研究方法、数据收集、数据分析、讨论结论、初稿撰写、字数占比（80%）；
熊文新：选题构思、研究方法、字数占比（20%）、修改润色。

/到/向…）+时地补语”“V+趋向补语”“V+结果补语”等，不同VP结构的“把”字句可被视为不同的下位构式，应有其各自的原型语义。因此，本文拟从下位构式出发，对各下位构式的原型语义及构式化程度进行对比分析，并尝试在此基础上构建“把”字句的语义结构，或能加深我们对“把”字句复杂句义的认识。

已有研究也有不少聚焦于“把”字句中的特定小类，如“V+着/了/过”类“把”字句（史金生 1988；王惠 1993；史金生、胡晓萍 1998）、动词重叠式“把”字句（高平平 1999；曾祥喜 2020）、光杆动词式“把”字句（刘承峰 2003；高艳 2011）、“一V”式“把”字句（徐峰 2014）、保留宾语类“把”字句（施春宏 2015；玄玥 2017）、述结式“把”字句（王璐璐、袁毓林 2016）等。但以上研究一方面对各类“把”字句的考察较为分散，难以看到不同小类之间的关联；另一方面则较少采用大规模语料数据作为支撑，且未涉及对原型语义和构式化程度的探讨。本研究将从这些角度对现有研究予以补充。

2 研究设计

2.1 研究问题

本文拟回答以下三个问题：

（1）“把”字句各下位构式在原型语义上具有怎样的关系？

（2）“把”字句各下位构式在构式化程度上具有怎样的关系？

（3）从以上两方面的结果来看，“把”字句内部具有怎样的语义结构？

2.2 语料采集

本文所用语料来自BCC语料库（荀恩东等 2016）中的多领域模块，该模块收集了报刊、文学、微博、科技四个领域的现代汉语文本，总库容19.6亿字，能够较好地反映现代汉语使用的全貌。直接使用字面标记“把”进行检索，从检索结果中随机抽取无重复的6万条，并进一步删去其中的非“把”字句，最终得到研究样本52,678条。

根据构式的形义匹配原则，本文以不同的VP结构为依据，将“把”字句初步分为10类下位构式，为行文方便，将其分别命名为“位移”“趋向”“转移”“转化”“结果”“情态”“完成”“动量”“方式”“简单处置”。其他极低频结构均归入“其他”类。各类构式对应的VP结构及相应信息见表1。

表1　“把”字句下位构式信息[1]

类别	VP结构	例句	频次	百分比
位移	V（在/到/向）+时地补语	把做好的食物摆在桌子上	15,428	29.29%
转化	V（成/作/为）+结果宾语	把铁饭碗变成金饭碗	11,835	22.47%
趋向	V+趋向补语	把一盒饼拿出来	6,675	12.67%
结果	V+结果补语	把下午的事情处理完	5,649	10.72%
转移	V（给）+与事宾语	把复印的材料递给我	4,148	7.87%
情态	V+情态补语	把一个个官兵打得飞跌开去	2,402	4.56%
完成	V了	把一个区给拆了	1,796	3.41%
方式	状语+V	把一个科研工作者像修鞋匠一样对待	1,146	2.18%
动量	V+动量/时量补语、一V、V（一/了）V	把下巴在领子角上蹭了两下、把不停眨着的眼睛一瞪、把上边的土扫了扫	952	1.81%
简单处置	V	把一部分政府职能分解	825	1.57%
其他	V+保留宾语、习语、谓宾动词+V、V+名量补语、V着……	把煮嫩的鸡蛋剥去壳、把一切事情都附耳相告、把上海同这些城市进行比较、把一份申请书复印8到10份、把三个杯子摆着……	1,822	3.46%

由表1可见，不同下位构式的使用频率差别较大。使用频率最高的是位移类和转化类，这两类占所有“把”字句使用的一半以上。频率最低的是方式类、动量类和简单处置类，三类均仅占2%左右，不到高频构式的十分之一。可见不同下位构式在“把”字句中的地位悬殊。

2.3　原型语义考察方法

本文采用Stefanowitsch & Gries（2003，2009）构式搭配分析法中的共现词分析法（collexeme analysis）考察“把”字句各下位构式的原型语义。该方法基于构式的语义一致原则（Goldberg 2006：39），认为构式内部词语与构式存在语义层面的一致性，统计方法层面的相互预测性（房印杰 2018）。本文通过比较词项在构式中的观察频次和期望频次，计算出词项与构式的搭配强度，搭配强度最高的一系列词项可用以预测构式的原型语义。搭配强度在原始频次的基础上滤去了常用

词的影响效应，使得对构式具有特别偏好的词语凸显出来。Gries *et al.*（2005）通过实验研究证明，使用搭配强度比使用原始词频预测出的原型语义更符合我们的认知。鉴于谓语动词在构式中的核心地位，本文取各构式谓语动词槽位的词语进行共现词分析。计算搭配强度有多种可选算法，本文使用对数似然率进行计算[2]。

2.4 构式化程度考察方法

Casenhiser & Goldberg（2005）提出的“偏态频率假设”认为，共现词的偏态分布（即齐普夫分布）是构式的区别性特征之一。偏态分布表明构式具有明确的语义倾向，体现为出现少数高频共现词；同时又具有较强的能产性和延展性，体现为出现大量低频共现词。Gries（2012，2015）同样将共现词的偏态分布看作构式形成的方式之一，并提出可采用信息熵（entropy）来衡量偏态分布的程度。信息熵是一种离散趋势的度量，取值范围为0—1，词汇分布越不均衡（即词频差异越大），信息熵越低。信息熵的计算公式为：

$$H=-\frac{\sum_i^n(p_i\cdot lnp_i)}{ln\mathrm{n}}$$

（Gries 2013：120）

其中，n为共现词的类符数，p_i为第i个共现词的使用频率。公式中的分母起到归一化作用，使得信息熵不受类符数量的影响。

3 结果与讨论

3.1 原型语义分析

为可视化呈现下位构式原型语义之间的相似度，本文选取各类原型动词（搭配强度最高的15个谓语动词）合成特征词表，以搭配强度作为特征值构建特征向量，并在此基础上进行对应分析。对应分析既能横向对比各列（构式）之间的相似度，又能纵向对比各行（动词）之间的相似度，并将两类相似度距离对应到同一个二维空间中（Desagulier 2014；Levshina 2015：370）。该方法在语义学研究中具有重要地位（Glynn 2010a，2010b；Desagulier 2014）。具体操作采用R中的ca程序包实现，结果见图1。

由图1可见两个比较明显的聚类，分别为左下角由“转移”“位移”“趋向”“方式”四类构式形成的聚类（简称“聚类1”），以及其右侧由“结果”“情态”“完成”和“动量”四类构式形成的聚类（简称“聚类2”）。转化类和简单处置类相对独立。[3]

聚类1以包含大量具有位移义的动词为主要特征。其中位移类和趋向类构式相似度最高，二者均大量吸引典型的空间位移动词[4]，如“拿”“搬”“挂”“带”

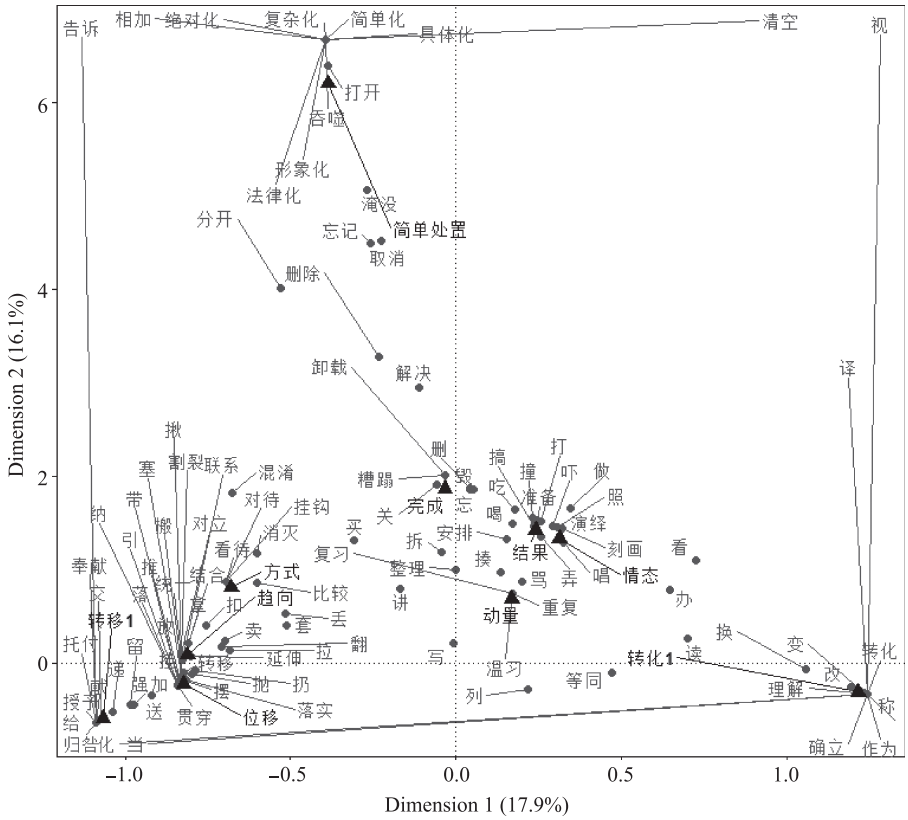


图1　构式与动词搭配关系的对应分析[5]

“推”“扔”等。其中位移类还吸引部分抽象位移动词如“纳（入）”“贯穿”等，可看作典型空间位移的隐喻扩展；趋向类同时吸引一类带有明显方向性的动词，包括“结合”“对立”“联系”“割裂”“统一”等，此类动词表达抽象事物的分合关系，同样可看作物理空间位移的一种隐喻扩展（张旺熹 2001）。结合VP结构，位移类构式的原型语义可概括为：主体对客体施加移动类操作，致使客体位移至某终点，如例（1）和例（2）。

（1）把一台笔记本电脑放到会议桌上。

（2）把一切金融活动纳入规范化、法制化轨道。

趋向类构式的原型语义可概括为：主体对客体施加移动或分合操作，致使客体朝特定方向发生位移，如例（3）和例（4）。

（3）把一块二尺左右的红布拿出来。

（4）把“依法治国”和“以德治国”结合起来。

方式类构式则同时吸引以上几类动词，包括表位移的“塞”“套”“扔”“延伸”“转移”等，以及表分合关系的“结合”“分开”“挂钩”“混淆”等，此外还吸引了一些其他类别的双音节动词，如“对待”“看待”“比较”“消灭”“删除”等。与这些动词搭配的状语大多带有一定的方向性，如“与/同……结合/分开/挂钩/混淆/比较”“向/朝/往……塞/套/扔/延伸”“当……对待/看待”等。据此，可将方式类构式的原型语义概括为：主体以特定的方向处置或对待客体，如例（5）和例（6）。

（5）把一颗糖往它口里塞。

（6）把hele当真朋友看待。

转移类构式除了吸引典型的位移动词“递”“扔”“塞”等外，以吸引表达事物转移或传递的双及物动词为主要特征，包括“交”“送”“给”“告诉”“托付”“授予”等。位移和转移分别表达物体的空间位置变化和领属者变化，其中，空间位置变化是领属者变化的必要条件和方式，包含在转移义之中，如“把一只乌龟小玩偶扔给始源”，位移动词“扔”为转移实现的方式，整体仍表达转移义。据此，转移类构式的原型语义可概括为：主体对客体施加转移或传递类操作，致使客体为新的领属者所有，如例（7）和例（8）。

（7）把一顶帽子送给对方。

（8）把一切真实情况告诉晓彤。

以上四类构式均带有位移性和方向性，其中位移类和转移类分别以地点和对象来指明位移的终点，趋向类仅指明位移的方向，方式类不限于表达位移事件，但常通过状语来表达行为的方向。

聚类2以包含各类动作行为动词为主要特征。其中，结果类和情态类构式的相似度最高，二者以吸引单音节动作行为动词为主，较具体的如“吃”“喝”“打”“骂”“唱”等，较抽象的如“弄”“搞”“做”“办”等。结合VP结构，可将这两类构式的原型语义概括为：主体对客体施加动作行为，致使客体产生某种变化，区别在于情态类构式能够更加具体地描绘出变化结果的情貌，如例（9）和例（10）。

（9）把一个小男孩的充气艇打翻（结果类）。

（10）把一件新衣服弄得千疮百孔（情态类）。

动量类构式除了吸引具体动作动词“看”“读”“骂”“揍”等外，还以吸引位移动词和含有明显重复义的行为动词为特征，前者包括“放”“扔”“塞”“套”等，后者包括“复习”“重复”“温习”等。对位移类动词的吸引将动量类在图中拉扯至相对靠近聚类1的位置。结合VP结构，其原型语义可概括为：主体对客体施加一定量的动作行为，如例（11）至例（13）。

（11）把《月光宝盒》和《大圣娶亲》都认真地看了两遍（特定量）。

（12）把三十几条意见整理整理（少量）。

（13）把一大袋东西往旁边的椅子上一放（瞬时量）。

与方式类相似，动量类原型语义的涵盖面也较广，不限于特定类型的行为。

完成类构式吸引的动词以包含“去除”义为主要特征，其搭配强度最高的15个动词中除了“看”和“买”之外，剩下的13个动词均含有“去除”义，即“删”“忘”“丢”“毁”“拆”“删除”“卸载”“卖”“糟蹋”“忘记”“关”“扔”“解决”。张伯江（2000，2001）指出“把”字句具有完全影响义，这主要体现在完成类构式中。试比较“他卖了苹果”和“他把苹果卖了”，后者意指卖掉了全部的苹果，而前句无此意。正因如此，完成类“把”字句中的原型动词经常与“一切”“所有”“都”共现，如例（14）至例（16）。

（14）把一切都毁了。

（15）把ipad里所有东西都删了。

（16）把一大盒都扔了。

据此，可将完成类构式的原型语义概括为：主体致使客体发生完全的损耗。与结果类和情态类构式类似，完成类的原型语义同样为“致使–变化”义，只是进一步限定为完全损耗类的变化。

下面再看图中相对孤立的转化类和简单处置类。

转化类以吸引改变类动词和含有主观认识义的动词为主，前者如“变”“改”“换”“转化”“译”等，后者如“作为”“称”“理解”“等同”“当”等。以上两类动词用于转化类构式，分别表达主体将客体在客观层面转化为另一物/形态，如例

（17）和例（18），以及主体将客体在主观层面转化为另一物/形态，如例（19）和例（20）。

（17）把铁饭碗变成金饭碗。

（18）把旧房子改成生产车间。

（19）把一场灾难当成游戏。

（20）把问题看成钉子。

这两类动词与其他构式的原型动词极少交叉，使得转化类在图中自成一类。除此之外，转化类还吸引部分动作行为动词，如“读”“办”“写”“列”等，这类动词也含有对客体形态上的影响义，用于转化类构式中同样表达将客体转化为或看作另一物/形态，如例（21）和例（22）。

（21）把“English”读成“应给利息”。

（22）把杀菌涂料列为分支产品。

由于这类动词与聚类2有较多交集，使得转化类在图中与聚类2相对靠近。类似于结果类、情态类和完成类，此类构式同样表达一种“致使–变化”义，但仅限于客体的完全质变。

简单处置类构式则以吸引多音节动词为主要特点，其中三音节的均为“形容词/名词+化”结构，如“具体化”“简单化”“绝对化”“法律化”等，此类动词与其他构式无交集，使得简单处置类在图中自成一类。双音节的则有“打开”“删除”“淹没”“取消”“吞噬”等，此类动词与聚类1中的方式类和聚类2中的完成类有较多交集，使简单处置类基本位于方式类和完成类的中线上。用于简单处置类“把”字句中的动词均表达主体致使客体产生某种变化，如例（23）和例（24）。

（23）把上级精神具体化。

（24）把新买回来的藤篮子打开。

尽管结果类和情态类构式的原型语义也为“致使–变化”，但原型动词多为单音节具体行为动词，而简单处置类由于音节的限制作用，原型动词多为多音节抽象动词，其原型语义也相对抽象。

总体来看，各下位构式存在一个统一的共性，即表达“致使–结果”义。其

中致使义体现为各类构式的原型动词均以强影响性的及物动词为主。结果义则在VP结构和动词的互动关系中有所体现。其中，VP结构为“动词+补充成分”的构式可由补充成分表达结果，因而原型动词大多不含结果义，此类构式包括位移类、趋向类、转移类、结果类、情态类和转化类；而VP结构中动词后不含补充成分的构式，则大量吸引含结果义的动词，此类构式包括完成类、方式类和简单处置类。其中，受动词后“了”的影响，完成类以吸引含“去除”类结果义的动词为主，此类动词加“了”即可实现结果义，如“删了”“扔了”“毁了”（王惠 1993；杨素英1998a，1998b；叶向阳 2004）。简单处置类由于动词前后不带任何成分，因而多吸引动补结构的双音节动词以及“形容词/名词+化”结构的三音节动词，其中前者由动词中的补充成分表达结果义，如“打开”“删除”“淹没”等；后者由形容词或名词指向结果的状态，如“具体化”“简单化”“绝对化”等。方式类的动词后无补充成分，但动词前有状语，而状语也能在一定程度上隐含结果义，如“把一颗糖往它口里塞”，由方向指示出客体位移的终点——“它口里”，再如“把hele当真朋友看待”，由方式预示出客体的最终状态——“真朋友”，因而方式类除了像简单处置类一样吸引动补结构的双音节动词外，还吸引了少量不含结果义的动词，如“塞”“套”“延伸”等。与以上这些小类相比，动量类似乎是唯一不具有结果义倾向的构式小类。但叶向阳（2004）指出，具有强影响性的动词作用于受事，即已蕴含使受事产生一定的结果，如“把S揍了一顿”暗含受事受伤的结果，“把C盘整理一下”暗含C盘变整洁的结果。由于动量类构式同样以吸引强影响性的动词为主，我们可以推测此类构式也倾向于表达一定的结果义。当然，与以上其他构式小类相比，动量构式表达结果义的程度相对较弱。

Lakoff（1987）指出，相对具体的构式会继承其上位构式的某些形式和功能，同时又具有上位构式所没有的一些特征。据此可推测，各下位构式的“致使–结果”义正是从上位“把”字句继承而来，由“把”字句的基本形式“（NP1）把+NP2+VP”赋予，并传导至共有该形式的下位构式中，这进一步证实了叶向阳（1997，2004）、郭锐（2003）的“致使说”以及施春宏（2010）的“致果说”。下位构式之间的差异则源于不同的及物性关系，由不同的VP结构赋予，使得具有相似及物性关系的小类倾向于拥有相近的原型语义，如结果类和情态类构式，位移类和转移类构式等。

3.2　构式化程度分析

采用散点图呈现各类构式中谓语动词的词频分布情况，并将在此基础上计算出的信息熵标示于图中，绘图采用R中的“plot（ ）函数”进行，结果见图2。

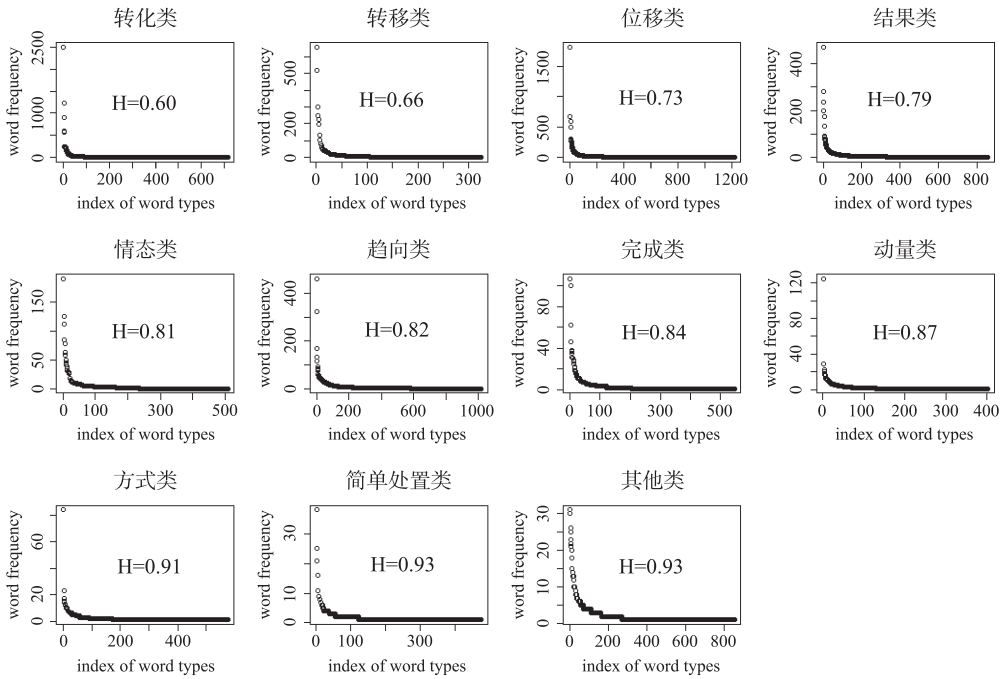


图2　谓语动词的词频分布[6]

由图2可见，不同下位构式的谓语动词均呈现典型的齐普夫分布，但在偏态的程度上具有差异。进一步分析发现，偏态分布的程度与构式的认知凸显度[7]之间存在关联：认知凸显度较高的构式，偏态分布也更显著，信息熵较低；认知凸显度较低的构式，偏态分布也较弱，信息熵较高。位移类、转移类和趋向类均为表达客体位移的构式，其中位移类和转移类分别通过时地补语和与事宾语显性表达客体位移的终点，如“把电脑放桌上（终点）”“把衣服拿给她（终点）”，位移路径的认知凸显度较高；趋向类仅说明位移的方向而不指明终点，如“把衣服拿出来（方向）”，位移路径的认知凸显度较低。相应的信息熵为位移类（0.73）、转移类（0.66）<趋向类（0.82）。转化类、结果类、情态类和完成类均为表达客体状态变化的构式，其中转化类通过“把”后宾语和结果宾语显性表达客体从A态到B态的质变过程，如“把铁饭碗（A态）变成金饭碗（B态）”，变化路径完整，认知凸显度最高；结果类和情态类构式仅通过补语表达终结态B，未表达起始态A，如“把充气艇打翻（B态）”“把水喝得一滴不剩（B态）”，变化路径的凸显度稍低；完成类构式同样仅表达终结态B，且此终结态由“谓语动词+了”进行隐含表达，如“把水喝了”，隐含“水没了”的结果，认知凸显度更低。相应的信息熵为转化类（0.60）<结果类（0.79）、情态类（0.81）<完成类（0.84）。动量类和方式类“把”字句在结构上仅凸显行为的量或方式，如“把桌子整理一下（量）”“把糖往口里塞（方式）”，在直观上不如客体的位移或状态变化凸显，因而动量类和

方式类的信息熵又比以上各类更高，分别为0.87、0.91。最后是简单处置类“把”字句，此类构式在结构上仅凸显动作行为本身，认知凸显度最低，其对应的信息熵也最高（0.93）。

根据以上分析，可将“把”字句各下位构式按照构式化程度归纳为以下的层级模式，见图3。

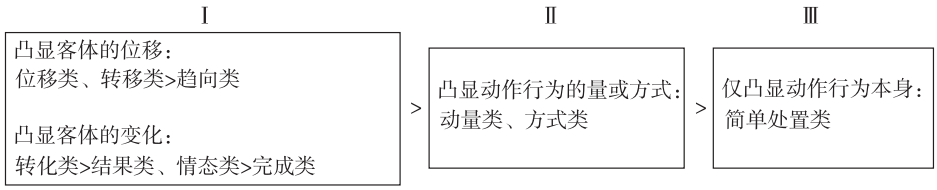


图3　构式化程度层级模式

对此可能的解释是，凸显客体的位移或变化均为凸显“致使–结果”义的方式，因而相应构式位于“把”字句中的核心位置。一方面，其在“把”字句中的使用频次较高；另一方面则具有较清晰的认知图式，表现为鲜明的语义倾向以及较强的能产性，进而形成共现词较强的偏态分布。以上两方面可能是相互影响的，即高频使用有利于促成清晰的认知图式，而清晰的认知图式又能进一步促进其更多的使用。结果凸显度较低的构式，则位于“把”字句中相对边缘的位置，一方面使用频次较低，另一方面认知图式也较为模糊。

3.3 “把”字句语义结构分析

以上原型语义和构式化程度的分析结果均表明，“把”字句的核心语义为“致使–结果”，但此核心语义内部又可分为“致使–位移”和“致使–变化”两个原型图式。根据Lakoff（1987）的看法，原型范畴内部可能具有多个原型，如张懂（2018）研究得出汉语双及物构式具有表达“给予”和“获取”的双原型语义结构。汉语“把”字句也可能为双原型语义结构，其中“致使-位移”义主要沿着“位移类、转移类→趋向类→动量类、方式类”的路径扩展；“致使–变化”义主要沿着“转化类→结果类、情态类→完成类→动量类、方式类→简单处置类”的路径扩展。两类原型在向外围扩展，同时也在对结果凸显度逐渐降低的过程中，边界逐渐模糊并发生交融。

张伯江（2000）、张旺熹（2001）和吕文茜（2015）均证明“把”字句的原型语义为空间位移，这与本文双原型结构的结论并不完全冲突。“位移”为原型，但不排斥可能与另一种原型语义共存。张伯江（2000）统计得出的表位移的动趋类“把”字句占绝大多数，而本文统计得出的表“致使–位移”的位移类和“致使–变化”的转化类为两类最高频构式，此差异极有可能与语料类型有关。张伯

江（2000）采用的语料来自小说，本文语料则来自报刊、文学、微博、科技四类语体。张旺熹（2001）总结的“把”字句五类图式中，位移图式和系联图式均以表达位移为主，等值图式、变化图式和结果图式均以表达状态变化为主，这两类在该文语料中的占比分别为62.8%和35.4%，尽管前者占比更大，后者占比也不小。吕文茜（2015）通过构式–搭配分析法发现，“把”字句的显著搭配动词中有49.4%为位移动词，形符占比67.71%，但也不排斥与另一种原型语义共存。此外，该文中举例的位移动词有部分我们认为并没有明显的位移义，如“到达”类的“撞”“踩”，“敲”类的“拍”“敲”“戳”等。这些动词用于“致使–位移”类构式中可表达位移，如“把花瓶撞到地上”；用于“致使–变化”类构式中则可表达内在状态的变化，如“把花瓶撞碎了”。甚至典型的位移动词用于“致使–变化”类构式中，也可由构式的压制作用表达变化义，如“把门推坏了”“把冰棍放成了一摊水”，等等。以上三项研究均以频次信息为主要依据，本文则结合了下位构式的原型语义和构式化程度来进行分析。尽管视角不同，但两种分析路径的结果可以相互佐证。

至于“致使–位移”和“致使–变化”两类图式之间是否存在隐喻关系，我们暂时持保守观点。尽管Goldberg（1995：83）通过英语实例指出结果构式由位移构式隐喻而来，如“Bob fell asleep”“Bob went crazy”中通过表达位移的“fell”和“went”来表达状态变化路径，汉语中也同样存在由位移义隐喻表达结果义的现象，如“找出来”“热起来”“安静下去”，由趋向动词的引申义来表达事件的结果。但汉语中的此类隐喻现象仅限于趋向类构式，对于转化类、结果类、情态类等构式，则找不到从位移义隐喻而来的证据。从认知上来看，也很难说物体的位置变化就一定比其质和量的变化更具有直观性和凸显性。因此，本文暂且将“致使–位移”和“致使–变化”看作两个相对独立的原型图式。当然，无论二者之间是否存在隐喻关系，不可否认的是，两类图式均处于“把”字句中的核心位置。

4 结语

本文在构式语法视角下，采用基于语料库的数据统计方法，对比分析了“把”字句各下位构式的原型语义以及构式化程度，并在此基础上探讨了“把”字句的语义结构。主要有以下几点发现：（1）“把”字句各下位构式的原型语义均含有一定的“致使–结果”义，表现为原型动词的强影响性以及原型语义蕴含结果义的倾向，这验证了“致使–结果”为上位“把”字句固有的原型语义，并由各下位构式所继承；（2）对“致使–结果”义凸显度较高的构式，构式化程度也较高，并居于“把”字句中的核心位置；（3）“把”字句具有表达“致使–位移”义和“致使–变化”义的双原型结构，并沿不同下位构式形成两条相对独立的扩展路

径。本研究从构式语法、语料库驱动等角度进一步检验并丰富了对“把”字句复杂句义的认识。

本文仅对“把”字句中频率较高的10类构式进行了研究，未涉及其他低频构式，包括VP结构为“V着”“V过”“V+保留宾语”等的“把”字句，这些构式作为“把”字句中的特殊类、边缘类，同样具有重要的参考价值。此外，本文通过语料数据得出的假设性结论有待通过其他研究路径进一步证实或证伪。

注释

1　按频次由高至低排序。

2　尽管Stefanowitsch & Gries（2003）更推荐对数据分布要求较低的费舍精确检验（Fisher exact test）算法，但该算法难以解决当P值过小时搭配强度显示为无穷大/小（Inf）的问题。本文在Gries的建议下采用原理相似的对数似然率算法。

3　下文分析以图1为依据，同时也参考了搭配强度表，以确认某词是否在某类构式的原型动词中。

4　关于位移动词的研究参考栗爽（2008）。

5　图1中灰色圆点代表动词所在的位置，黑色三角代表构式所在的位置。由于部分动词在构式中的表现近似，导致位置重叠，我们使用R中的ggrepel程序包来自动分开重叠的标签。

6　图2中各类构式按照信息熵由低至高进行排序，横坐标为动词类符的序号，纵坐标为其在构式中的共现频次。

7　这里指的是从结构上来看的认知凸显性，不局限于原型语义。

参考文献

CASENHISER D, GOLDBERG A. Fast Mapping of a phrasal form and meaning [J]. Developmental Science, 2005, 8 (6): 500-508.

DESAGULIER G. Visualizing distances in a set of near-synonyms: rather, quite, fairly, and pretty [C]//GLYNN D, ROBINSON J. Corpus methods for semantics: quantitative studies in polysemy and synonymy. Amsterdam: John Benjamins, 2014: 145-178.

GLYNN D. Corpus-driven cognitive semantics: introduction to the field [C]// GLYNN D, FISCHER K. Corpus-driven cognitive semantics: quantitative approaches. Berlin: Mouton de Gruyter, 2010a: 1-42.

GLYNN D. Synonymy, lexical fields, and grammatical constructions: a study in usage-based cognitive semantics [C]//SCHMID H, HANDL S. Cognitive foundations of

linguistic usage-patterns: empirical studies. Berlin: Mouton de Gruyter, 2010b: 89-118.

GOLDBERG A. Constructions: a construction grammar approach to argument structure [M]. Chicago: University of Chicago Press, 1995.

GOLDBERG A. Constructions at work: the nature of generalization in language [M]. Oxford: Oxford University Press, 2006.

GRIES S. Frequencies, probabilities, and association measures in usage-/exemplar-based linguistics [J]. Studies in Language, 2012, 36 (3): 477-510.

GRIES S. Statistics for linguistics with R: a practical introduction [M]. Berlin: Walter de Gruyter, 2013.

GRIES S. More (old and new) misunderstandings of collostructional analysis: on Schmid and Küchenhoff (2013) [J]. Cognitive Linguistics, 2015, 26 (3): 505-536.

GRIES S, HAMPE B, SCHÖNEFELD D. Converging evidence: bringing together experimental and corpus data on the association of verbs and constructions [J]. Cognitive Linguistics, 2005, 16 (4): 635-676.

LAKOFF G. Women, fire, and dangerous things: what categories reveal about the mind [M]. Chicago: University of Chicago Press, 1987.

LEVSHINA N. How to do linguistics with R: data exploration and statistical analysis [M]. Amsterdam: John Benjamins, 2015.

STEFANOWITSCH A, GRIES S. Collostructions: investigating the interaction between words and constructions [J]. International Journal of Corpus Linguistics, 2003, 8 (2): 209-243.

STEFANOWITSCH A, GRIES S. Corpora and grammar [C]//LÜDELING A, KYTÖ M. Corpus linguistics: an international handbook (Volume 2). Berlin: Mouton de Gruyter, 2009: 933-951.

房印杰. 搭配构式分析——应用与发展[J]. 现代外语，2018（3）：425-435.

高平平. 谈“把”字句中的动词叠用[J]. 汉语学习，1999（5）：22-23.

高艳. 光杆动词进入把字句的条件论略[J]. 沈阳师范大学学报（社会科学版），2011（5）：86-89.

郭锐. 把字句的语义构造和论元结构[J]. 语言学论丛，2003，28：152-181.

栗爽. 现代汉语位移动词研究[D]. 上海：上海师范大学，2008.

刘承峰. 能进入“被/把”字句的光杆动词[J]. 中国语文，2003（5）：422.

吕文茜. 基于组配—构式分析法的“把”字句典型构式义研究[J]. 外语研究，2015（5）：22-25.

马真. 简明实用汉语语法[M]. 北京：北京大学出版社，1981.

沈家煊. 如何处置“处置式”？——论把字句的主观性[J]. 中国语文，2002（5）：

387-399.

施春宏. 从句式群看“把”字句及相关句式的语法意义[J]. 世界汉语教学，2010（3）：291-309.

施春宏. 边缘“把”字句的语义理解和句法构造[J]. 语言教学与研究，2015（6）：53-66.

史金生. 谈“把”字句中的“过”[J]. 汉语学习，1988（3）：27-30.

史金生，胡晓萍. 动词带“着”的“把”字结构[J]. 语言教学与研究，1998（4）：39-50.

宋玉柱. 关于“把”字句的两个问题[J]. 语文研究，1981（2）：39-43.

王惠. “把”字句中的“了/着/过”[J]. 汉语学习，1993（1）：6-12.

王力. 中国现代语法[M]. 北京：商务印书馆，1943.

王璐璐. 基于变换的“把”字句自动释义研究[D]. 北京：北京大学，2013.

王璐璐，袁毓林. 述结式与“把”字句的构式意义互动研究[J]. 语言教学与研究，2016（3）：54-63.

徐峰. “把NP一V”的句法、语义和语用功能[J]. 汉语学习，2014（4）：46-54.

玄玥. 保留宾语类把字句与完结短语理论[J]. 语言教学与研究，2017（3）：28-39.

荀恩东，饶高琦，肖晓悦，等. 大数据背景下BCC语料库的研制[J]. 语料库语言学，2016（1）：93-109.

杨素英. 从情状类型来看“把”字句：上[J]. 汉语学习，1998a（2）：10-13.

杨素英. 从情状类型来看“把”字句：下[J]. 汉语学习，1998b（3）：10-12.

叶向阳. “把”字句的致使性解释[D]. 北京：北京大学，1997.

叶向阳. “把”字句的致使性解释[J]. 世界汉语教学，2004（2）：25-39.

曾祥喜. “把+N+vv（双音节）”构式及构式压制[J]. 吉林大学社会科学学报，2020（5）：223-234.

张伯江. 论“把”字句的句式语义[J]. 语言研究，2000（1）：28-40.

张伯江. 被字句和把字句的对称与不对称[J]. 中国语文，2001（6）：519-524.

张懂. 基于语料库的汉语双及物构式原型语义模式实证研究[J]. 外语与外语教学，2018（5）：79-88.

张旺熹. “把”字句的位移图式[J]. 语言教学与研究，2001（3）：1-10.

通信地址：100089　北京市　北京外国语大学中国语言文学学院

基于动态图的英语近义词搭配分析*

大连海事大学 李雯静 孟庆楠

提要：本研究基于GloWbE语料库，利用动态图这一可视化方法，探究与quick和fast共现的搭配名词在20种英语变体中的使用特征和分布情况。研究发现：这一对近义词的右一搭配词存在较为明确的语义分工。quick主要倾向与look、fix、access、response等抽象名词搭配使用，涵盖范围更广且多用于描述经济、科技和社会发展状况，而fast则常与food、bowler、bowling、track等较为具体化的名词共现，多描述人们的日常生活方式。除此之外，由于受社会文化和历史地理等因素影响，这两个词语在不同国家和地区各自呈现出了较为独特的使用特征。

关键词：近义词、搭配分析、语义倾向、动态图、GloWbE语料库

1 引言

几乎每一种语言中都存在含义重叠的单词，这些词通常被视为近义词（synonyms）。英语作为国际通用语言之一，在不同国家和地区产生了多种方言变体，这使得英语近义词的种类和数量数不胜数，对近义词的分析也成为语言学研究中的热门话题之一。然而，仅仅依靠传统意义上的权威词典进行词义辨析显然已经无法满足时代的需求。近年来，随着诸多大规模语料库的建立，基于语料库的实证研究范式开始逐渐盛行，这为近义词辨析提供了丰富的语料素材。通过在语料库中查找与近义词共现的搭配词及相应的句法结构，可以了解不同词语之间语义和语用方面的细微差别。因此，借助大型语料库的近义词搭配分析业已成为主流的研究范式。

Pichler（2016）基于美国英语历时语料库（The Corpus of Historical American

* 本研究系辽宁省社会科学规划基金青年项目“基于原美国杨百翰大学系列语料库的英语构式交替现象研究”（L21CYY004）、辽宁省教育厅高等学校基本科研业务费项目“概率语法视角下英语情态构式交替现象多元定量研究”（LJKQR2021002）及中央高校基本科研业务费项目“基于R软件的语言学历时研究及可视化呈现”（3132022331）的阶段性研究成果。孟庆楠为本文通讯作者。
作者贡献：
李雯静：数据收集、数据分析、讨论结论、初稿撰写、字数占比（80%）；
孟庆楠：选题构思、研究方法、字数占比（20%）、修改润色。

English，简称COHA）对六组英语近义词在1810—2009年搭配词的变化情况进行了探究。Pichler在研究形容词类近义词quick、swift、rapid和speedy时，通过观察每个词的频数分布与变化情况得出：quick与rapid较为相似，quick与rapid相比，适用范围更广，使用频数也更高。然而这一研究并没有将另一较常使用且具有“快”含义的形容词fast涵盖其中，因此笔者选用fast与quick组合，构成全新的研究对象。通过查询在线朗文词典[1]发现，fast与quick所重合的释义更多，含义更为接近，主要义项有moving or done with speed、taking or lasting a short time、happening soon or without any delay等。很多时候对一些习惯性用语的理解并不能仅根据词典中的定义来确定，因此有必要进一步探究quick和fast的搭配和用法，根据语境来窥视这组词在意义上的细微差别。以往对近义词的搭配研究多涉及历时层面，利用历时语料库探究这一语言现象的演变规律。为拓展近义词横向层面的研究，本文采用在线版的世界网络英语语料库（The Corpus of Global Web-based English，简称GloWbE）[2]对quick和fast开展共时研究，将其高频搭配词在20种英语变体中的分布及使用情况进行可视化呈现，从而揭示两词的语义差别及使用规律。

2　理论框架

在语言学研究中，并不存在绝对意义上的同义词。因为绝对同义词在任何语境中都可相互替代且丝毫不改变交际效果，然而这一现象十分罕见，学界也并未达成共识。大多数学者认为，语言研究仅针对近义词，即在外延义上看似相同的两个或多个词，实则在文体风格（style）、信息焦点（emphasis）、适用语域（register）等方面存在一定程度的区别，在某些情况下不能替换使用。DiMarco *et al.*（1993）对近义词的外延义和内涵义都进行了区分，并将这些差异概括总结为12个维度，其中内涵意义主要由语义韵（semantic prosody）进行区分。语义韵这一术语最初由Louw（1993）提出，探讨的是抽象的语用层面的词语共选问题[3]。在他看来，词项因受常用搭配词语义的影响，逐渐获得了与搭配词相同的语义色彩。通过观察与某一词语共现的高频搭配词的语义特征，可以概括出这些搭配词的语义倾向（semantic preference），从而进一步勾勒出节点词所在的意义单位的语义韵特征，层层递进，逐渐完成对词语意义的分析。

在语义韵和语义倾向等概念出现之前，就有关于词语共选方面的研究。英国伦敦学派的创始人、语境论的提出者J. R. Firth（1957）首次将搭配问题上升至语言学理论高度，提出了“由词之结伴而知其义”的经典论断，并提出“通过搭配研究意义”（meaning by collocation）的论断。Sinclair（1991）深化了Firth的观点，不仅提出“扩展意义单位模型”，也对搭配给出了更为可操作化的定义。以上观点表明，任何意义的解读都不能脱离语境。

在如今计算机技术飞速发展的时代背景下，语言学研究也逐渐向实证研究

和定量研究方向转变。语料库语言学逐渐兴起并向纵深发展，各种类型的大规模在线语料库开始不断出现，这为在语境中研究词语及搭配的语义倾向和语义韵提供了更加现实的解决方案。因此，基于上述理论框架，本文将利用19亿库容的GloWbE语料库，对quick和fast后搭配词的语义倾向及分布情况进行探究，以揭示这对近义词的语义差别。

3 语料来源与提取

本研究所选取的语料均来自GloWbE在线语料库。GloWbE语料库由原杨百翰大学（Brigham Young University，简称BYU）的语言学教授Mark Davies提供，发布于2013年4月，自发布之日起每月超17万人次使用，目前已经成为使用最广的在线语料库之一。该语料库中包含了2012—2013年来自全球20个国家和地区的英文文本，总数约19亿单词，涵盖超过180万个网页的内容。所涉及的语言变体范围既包括以英语为母语的国家，如英国、美国、加拿大等，也包括将英语作为官方第二语言的国家，如印度、巴基斯塔、牙买加等。所包含的文本数量约是同类型的国际英语语料库（The International Corpus of English，简称ICE）的100倍，这就为学者进一步研究某些低频出现的语言现象提供了有利条件。

为探究与quick和fast共现的名词搭配在世界英语中的分布情况及规律，本研究采用了以下做法：首先，在GloWbE语料库在线页面中选择“compare”功能，在下面的“Word1”和“Word2”搜索框中依次输入quick和fast，并点击其后的“POS词性标注”，在下拉列表中选择“adj. ALL”，将待检索的两个词全部定性为形容词。其次，在“collocates”条框中选择右一（R1）搭配词，并在POS标签的下拉式菜单中选择“noun. SG”，代表所有单数名词。为保证数据充足及可解读性，本研究将阈值设置为3（即每个搭配词在各变体子库中至少出现3次），将检索结果数设为100。然后依次选取“sections”列表中20种英文变体的名称，并将每一次搜索的两组搭配词的原始频数分别进行记录，存储至Excel表格中，按照搭配词与检索词的相关性从高到低进行排序。最后，为使不同变体的数据具有可比性，本研究以百万词为单位，对上述数据进行了标准化处理。

表1为处理后的部分数据。其首列为每种英语变体的国家名称缩写，第二列为搜索得出的主要右一搭配词。第三列和第四列分别代表与quick和fast共现的搭配词的标准化频数，最后一列为这些搭配词与检索词共现的总标准化频数。

表1　绘制动态图所需的数据片段

Variety	Noun collocate	Collocate frequency with *quick*（X-axis values）	Collocate frequency with *fast*（Y-axis values）	Combined frequency
US	LOOK	1.326234729	0.01809677	1.344331499
US	FIX	0.97722559	0.01809677	0.99532236
US	SEAECH	0.736797072	0.028437782	0.765234853
US	REPLY	0.199064472	0.020682023	0.219746495
US	WALK	0.111165874	0.012926264	0.124092138
US	WORK	0.168041437	0.020682023	0.18872346
US	STUDY	0.09823961	0.012926264	0.111165874

4　研究方法与过程

为了将与近义词quick和fast频繁共现的搭配词的语义特征及共时分布规律以一种更加直观的方式展现，本研究将对上一部分生成的Excel表格和相应文本文件进行进一步分析和处理。最终的分析结果以“语言学动态图”（linguistic motion chart）的可视化方式呈现。这种可视化手段的原型是一系列动态气泡图（animated bubble charts），由Hans Rosling在2006年的一场TED演讲中所使用，以展现过去近50年国际经济、社会和人口的变化发展情况。受此启发，这种动态图发展成为一种展示大量二元或多元数据在一段时间内变化情况的交互式图表，于2011年由Gesmann & de Castillo研发。自此，在媒体、商业、经济学、语言学等领域，动态图被越来越多的人所熟知并接受。Martin Hilpert是最早将动态图应用于语言学领域研究的学者之一，他早期主要利用动态图对兼类词（ambicategorical words）和动词补语（complement-taking predicates）的演变和发展进行了实证研究。Hilpert（2011）指出，作为分析大型语料库数据的新技术，动态图能够更加直观清晰地揭示复杂语言现象背后的发展规律。然而，由于期刊论文在排版方面的限制，本研究的数据分析结果主要使用R统计软件中的“plot（函数）”，最终以一系列按20个国家及地区英语变体名称首字母排列的静态气泡分布图的形式有序并排呈现在二维平面上。

语言学的实证研究不可避免地涉及大量数据的分析与处理，与传统图表相比，动态图所囊括和传递的信息量更大，更加具有优势。因此近些年来，这一可视化技术手段逐渐在语言学领域应用拓展开来。如上所述，这些运用动态图的研究大多集中在对二元数据的观察与分析之上，例如：孟庆楠、李基安（2019）基于浮

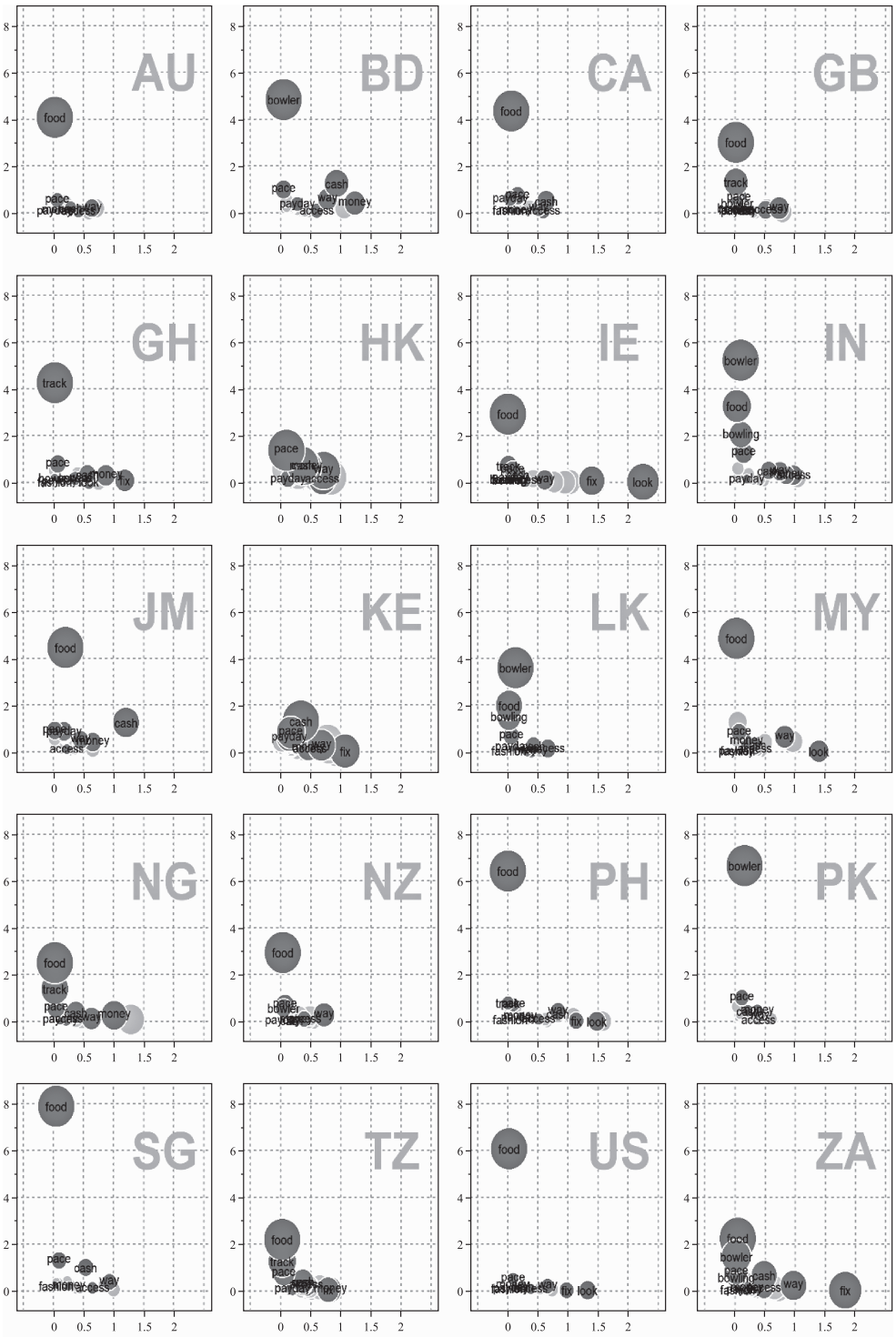
现语法视角对英语临界情态动词的历时变化研究；邵斌、黄丹青（2016）借助美国英语历时语料库对英语不规则动词过去式在近200年来的频数变化研究等。与此同时，学界利用该方法进行近义词辨析的研究也在逐渐增多，例如：Primahadi-Wijaya-Rajeg & Rajeg（2014）基于COHA语料库对近义词hot和warm右一名词性搭配词分布规律的历时研究；Pichler（2016）对形容词aware和conscious、动词attempt和try、名词couch和sofa等三类近义词的搭配词语义倾向历时变化的研究；孟庆楠等（2020）借助GloWbE语料库对海事英语近义词maritime和marine在世界英语中的用法研究等。

不过，上述绝大部分的研究都是从历时角度探索语言学现象的规律与变化，而在共时层面运用可视化手段的学术探究则少之又少。为进一步拓宽动态图的适用研究范围，本研究从共时视角出发，立足于右一搭配名词，探索quick和fast在世界英语变体中的分布规律及语义差别。在语料的筛选方面，GloWbE语料库中包含的所有20种国家和地区的英语变体皆为本文的数据来源。由于形容词fast包含与“快的”意思不相关的其他含义，比如“牢固的”“固定的”“可靠的”等，为保证quick和fast两个词数据的可比性，与这些含义共现的搭配词及频数被手动剔除。

5 结果与讨论

图1为本研究所需数据通过可视化技术呈现的最终结果，以一系列气泡散点图的形式，清晰直观地展现了quick和fast在20种英语变体中的搭配使用情况。其中，各个方框的横、纵坐标刻度范围根据该变体中搭配词标准化频数的最大值和最小值确定，各个方框右上角的大写字母代表20种英语变体所对应的国家及地区代码[4]，图形按照缩写的首字母横向进行排序分布。在每张图中，各气泡的大小代表该名词与quick和fast搭配出现的标准化频数之和，各气泡圆心所在的横、纵坐标分别对应该搭配词与quick、fast共现的标准化频数。为了避免相邻气泡标签堆叠影响图形的可解读性，笔者经过反复尝试，仅选取了部分与quick和fast相关性较高的搭配词赋予标签，并将这些标签的气泡颜色设置为深灰色，其余未被选中的词语气泡则呈浅灰色。

首先，通过观察图1可知，近义词quick和fast的名词性搭配词在GloWbE语料库20种英文变体范围内的分布情况既蕴含共性，又存在差异之处。总体而言，在所选取的绝大部分国家（地区）英语变体中，频繁与fast搭配使用的名词包括food、bowler、bowling、track、pace等，而经常与quick搭配共现的词语主要有look、fix、access、money、way等。鲜有气泡分布于每个图框的对角线附近，由此可知近义词quick和fast在右一搭配名词方面存在较为明显的语义分工。通过观察图形中主要气泡标签的含义可知，quick主要与含有抽象意义的名词搭配使用，



（注：横轴代表quick，纵轴代表fast）

图1　quick和fast的右一搭配词在GloWbE语料库20种英语变体中的频数分布情况

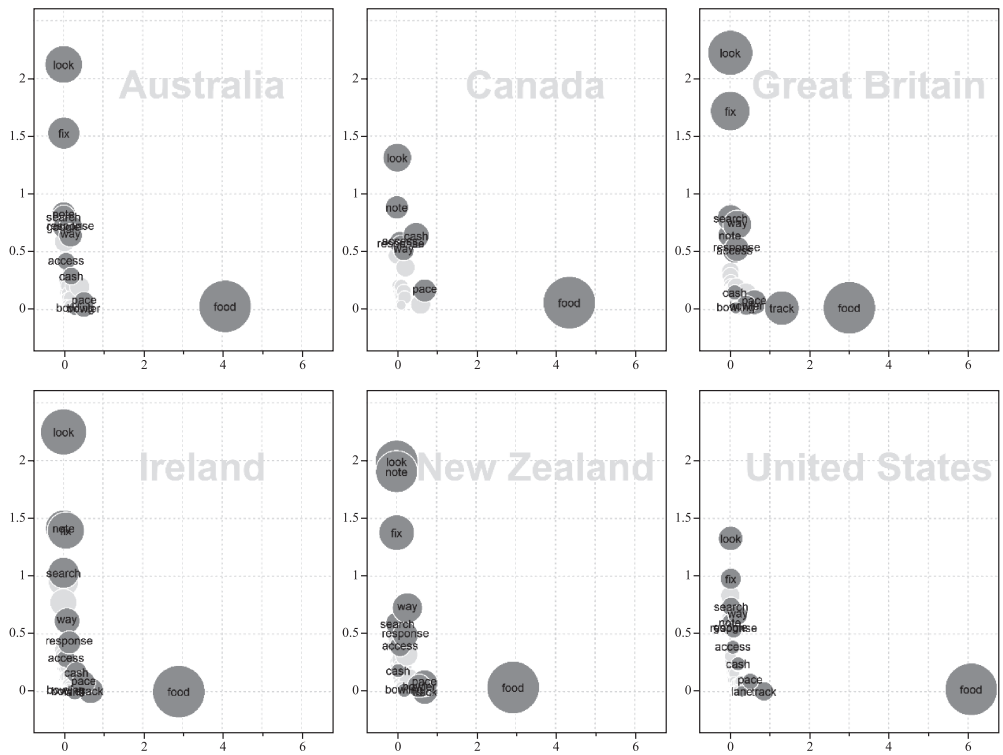
所涵盖的范围更宽泛，多用来描述经济社会发展及人们的生活方式，例如quick money（快钱；外快）、quick access（快速访问；快速存取）等，而形容词fast则多与较为具体化的词语共现，且含义多涉及体育运动、休闲娱乐等方面，比如fast bowling（板球快速投球）、fast bowler（快速投球手）、fast food（快餐；速食）等。观察横轴和纵轴的取值范围以及搭配词气泡大小后可知，在多数英语变体中，fast相较于quick使用的频数更高一些。

其次，这两个近义词在20国（地区）英语变体中的使用情况存在着略微不同。在图1第一行的四种英语变体中，唯一一个以英语为官方第二语言的国家——孟加拉国——其动态图呈现出了较为不同的搭配词分布情况。在其方框内，与fast搭配最频繁的名词是bowler，其次为pace，而最常与quick共现的词语是money，其次为way和access。单词cash更加倾向于与fast搭配使用。然而，在其他三种变体中，fast food是最常见的组合。在英式英语中，fast track排在第二位，可译为“快车道”“快速晋升之道”“捷径”等。在加拿大英语中，cash的气泡由对角线平分，说明在该地词组fast cash和quick cash出现的频数大致相等。在第二行中，只有爱尔兰以英语为母语，因此展现出了略微不同的气泡分布，名词look成为与quick搭配频数最高的名词，其次是fix，而纵轴fast上最为明显的气泡则为food。与孟加拉英语相似，在印度英语中，fast bowler是出现最多的名词搭配，其相应的动词性表达fast bowling也逐渐增多。随后通过观察剩余变体可知，斯里兰卡英语及巴基斯坦英语中皆以fast bowler或fast bowling居多，均呈现出与孟加拉英语和印度英语较为相似的趋势，这可能说明这些国家的板球体育运动有着较高的受关注度和话题度。笔者在查阅相关资料后发现，这四个国家无一例外都位于南亚地区，在19世纪都曾沦为英属殖民地，其中印度、巴基斯坦和孟加拉国原先同属于一个国家，因此在独立之后他们的联系相对较为密切。板球运动（cricket）起源于英国，被称为“绅士的游戏”，后盛行于印度、巴基斯坦、尼泊尔、南非等地。在这些国家独立之后，市场开放，资本主义逐渐发展，新兴的城市中产阶级开始追求闲暇时的娱乐，因此板球运动顺理成章地流行起来。通过查询原始语料素材后发现，关于fast bowling和fast bowler的记录多来自当地著名的体育网站，如Cricket Country、ESPNcricinfo、Sportskeeda等，这也印证了这些国家对板球这一体育项目或休闲方式关注度较高的观点。综上来看，fast bowler或fast bowling在这四国英语变体中的使用频数明显高于其余16种英语变体，其背后的原因也是有迹可循。

最后，在英式英语、加纳英语、尼日利亚英语、坦桑尼亚英语中，fast与track的组合也占部分比例，并且在原始语料中多以快速审判法庭（Fast Track Courts，简称FTCs）的形式出现。据了解，在这些国家中，英国的法制与经济较为发达，FTCs早已普遍使用，以提高司法机构的效率。而加纳、尼日利亚和

坦桑尼亚分属西非和东非，经济较为落后，其中加纳是非洲较早独立并进行法律变革的国家，FTCs及调解制度的引入为稳定社会秩序和保障女性权利做出了一定贡献。

在GloWbE语料库所收录的六种以英语为母语的变体内部，quick和fast似乎也存在着一定程度的分布规律。由于图1少数变体在方框左下角的部分标签较难辨认，因此笔者调换了原先横轴和纵轴所代表的近义词（即此时横轴代表fast，纵轴代表quick），将以英语为母语的六种变体搭配词分布图进行了放大处理，并添加了几处气泡标签，如图2所示。



（注：横轴代表fast，纵轴代表quick）

图2　quick和fast在6个以英语为母语的国家中的频数分布及使用情况

不难看出，在这6个核心国家中，使用频数最高的依然是fast food和quick look。而fast 和food的组合在美式英语中最为常见，其主要原因在于西式快餐业最初起源于20世纪50年代初的美国。当时美国经济复苏，工业文明和全球扩张步伐加快，人们热衷于创造并积累财富，生活节奏越来越快，对速度和效率的崇尚，推动了餐饮行业的发展。20世纪80年代末到90年代初，快餐发展在美国国内达到顶峰，而美国也在此时成为世界上快餐业最发达的国家。20世纪末，国内市

场饱和，现代快餐转向海外市场拓展，并逐渐风靡全球。原始语料中，与fast food存在于同一语境的名词多包括sandwich、chocolate、hamburger、pizza等流行美式快餐，McDonald's、KFC、Taco Bell等美国餐饮品牌，以及healthy、unhealthy、junk food等表评论性的词语。随着网络的进一步发展，人们对只求速度不求内涵的快餐文化也产生了不同程度的争议，因此fast food成为频数较高的搜索词也不无道理。

除此之外，英式英语、新西兰英语、爱尔兰英语和澳大利亚英语的气泡分布情况类似，与quick搭配出现的词语类型较fast更多，排在前几位的有look、fix、note、search、response、access。在线剑桥词典[5]对于quick fix的解释为：something that seems to be a fast and easy solution to a problem but is in fact not very good or will not last long，指不完善的应急解决办法，另外还指VSCode源代码编辑器的快速修复快捷键。

在加拿大英语中，与quick及fast共现的名词种类相对较少，且频数也相对较低。在上述6种英语变体中，quick access与quick response的气泡分布较为稳定，使用频数相近，原始语料中前者多与toolbar连用，表示软件中的快捷工具栏。而后者多与code连用，指一种叫作快速响应矩阵图码的二维码，广泛应用于物品识别、文档管理等方面的读码操作。这皆与现代信息技术息息相关，因此quick的这两种搭配在日常生活中也较为常见。综上，美式英语中fast food使用最多，加拿大英语中搭配词类别最少，与quick共现的词语多涉及现代生活与信息技术，这表明这些母语国家的语言使用与时代背景下快节奏的社会生活休戚相关。

6 结语

本研究借助GloWbE语料库，对quick和fast右一搭配词在20种英语变体中的语义倾向及分布情况进行了探究，并将数据分析结果进行了可视化呈现。研究结果表明：在绝大多数英语变体中，其右一搭配名词存在较为明显的语义分工。quick主要与含有抽象意义的名词搭配使用，如look、fix、access等，所涵盖的范围更宽泛，多用于描述经济社会发展及人们的生活方式；而fast则多与较为具体的词语共现，如bowling、bowler、food等，且含义多涉及体育运动、休闲娱乐等方面。在以英语为官方语言的四个南亚国家中，fast bowler和fast bowling的使用频数明显高于其他英语变体，其背后原因可以追溯至英国殖民统治时期引入的板球运动。在英式英语和加纳、坦桑尼亚、尼日利亚英语变体中，fast与track的搭配相对较高，且形成fast track courts这一较为固定的搭配。在6种以英语为母语的变体中，使用频数最高的是quick look和fast food，并且与quick搭配出现的名词的类符频数要多于fast。美式英语中fast food的使用频数最高。quick fix多出现于英

国、爱尔兰、新西兰和澳大利亚的英语变体中。加拿大英语中与quick和fast共现的名词种类相对较少，频数也较低。这些用法均与相关国家或地区的经济发展状况及社会生活环境息息相关。

本研究所采用的动态图这一可视化手段，在语言学研究领域具有广阔的应用前景。除近义词研究之外，亦可将其运用至对近义词组、近义词缀等的探究。本研究也存在一些不足之处。首先，GloWbE语料库仅包含20种以英语为母语或官方第二语言的国家（地区）的英语变体，上述使用情况是否在其他讲英语的国家（地区）中存在，仍需进一步研究；其次，除地域及社会经济文化因素之外，语言使用者的性别、年龄、受教育程度以及语域、语体风格等因素也有可能影响近义词的搭配分布及语义倾向。后续研究还需运用其他专用或通用语料库对本文分析结果进行交叉验证。

注释

1　详见https://www.ldoceonline.com。

2　详见https://www.english-corpora.org。

3　四个不同层面的共选关系依次为："搭配"（词语层面的共选）、"类联结"（语法层面的共选）、"语义倾向"（语义层面的共选）和"语义韵"（语用层面的共选）。

4　各英语变体所对应的国家及地区代码及全称如下：US（United States）、CA（Canada）、GB（Great Britain）、IE（Ireland）、AU（Australia）、NZ（New Zealand）、IN（India）、LK（Sri Lanka）、PK（Pakistan）、BD（Bangladesh）、SG（Singapore）、MY（Malaysia）、PH（Philippines）、HK（Chinese Hong Kong）、ZA（South Africa）、NG（Nigeria）、GH（Ghana）、KE（Kenya）、TZ（Tanzania）、JM（Jamaica）。

5　详见https://dictionary.cambridge.org。

参考文献

DIMARCO C, HIRST G, STEDE M. The semantic and stylistic differentiation of synonyms and near-synonyms [C]//AAAI Spring Symposium on Building Lexicons for Machine Translation. 1993, 1: 114-121.

FIRTH J R. A synopsis of linguistic theory, 1930-55 [C]//FIRTH J R. Studies in linguistic analysis. Oxford: Blackwell, 1957: 1-32.

GESMANN B M, DE CASTILLO D. Using the Google visualisation API with R [J]. The R Journal, 2011, 3(2): 40-44.

HILPERT M. Dynamic visualizations of language change: motion charts on the basis of bivariate and multivariate data from diachronic corpora [J]. International Journal of

Corpus Linguistics, 2011, 16(4): 435-461.

LOUW B. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies [C]//BAKER M, FRANCIS G, TOGNINI-BONELLI E. Text and Technology. Amsterdam: John Benjamins, 1993: 157-176.

PICHLER K. A diachronic perspective on synonymy [D]. Vienna: University of Vienna, 2016.

PRIMAHADI-WIJAYA-RAJEG G, RAJEG I M. Visualising diachronic change in the collocational profiles of lexical near-synonyms [C]//SUDIPA I N, PRIMAHADI-WIJAYA-RAJEG G. Cahaya Bahasa: in honour of Prof. I Gusti Made Sutjaja. Denpasar: Swasta Nulus, 2014: 247-258.

SINCLAIR J. Corpus, concordance, collocation [M]. Oxford: Oxford University Press, 1991.

SINCLAIR J. The lexical item [C]//WEIGAND E. Contrastive lexical semantics. Amsterdam: John Benjamins, 1998: 1-24.

孟庆楠，李基安. 英语临界情态动词构式变化研究——基于历时语料库数据的动态图分析[J]. 外语电化教学，2019（6）：103-112.

孟庆楠，钱景，周晨煜. 海事英语中近义词的搭配分析及可视化呈现——以marine和maritime在世界英语中的用法为例[J]. 大连海事大学学报（社会科学版），2020（6）：112-118.

邵斌，黄丹青. 基于语料库的英语不规则动词演变的可视化研究[J]. 山东外语教学，2016（4）：12-20.

通信地址：116026　辽宁省大连市　大连海事大学外国语学院

数学科普文本中的英语介入标记语研究

中国科学院大学 于 华

提要：科普文本一般包括科学知识和背景知识两个主要内容，其中普及科学知识是科普文本的主要目的。本研究以《素数之恋：黎曼和数学中最大的未解之谜》为研究素材，在语料库分析的基础上，检索、统计并分析介入标记语在该书奇偶章节中的分布情况、语用功能和使用特征。研究发现：该书中使用频率最高的介入标记语为读者人称，说明科普文本较多使用显性的读者介入策略。奇数章节比偶数章节显著多用介入标记语，尤其是读者人称和提问这两种类型，说明与背景知识相比，科普文本的科学知识更期待读者的介入、参与和理解。本研究还发现不同介入标记语的共现可提高介入力度。

关键词：介入标记语、科普文本、语料库分析

1 引言

科技发展是国际综合国力竞争的焦点。为了提高公众科学素质，吸引人才投身科学事业，科学普及已被视为科技创新发展的重要途径。科学普及是指“采用公众易于理解、接受和参与的方式，普及自然科学和社会科学知识，传播科学思想，弘扬科学精神，倡导科学方法，推广科学技术应用的活动”（中华人民共和国人大法工委 2002）。科普作品是一种以科学普及为主要目的的文本，注重提升读者的阅读体验，从而更有效地普及科学知识。优秀的科普作品常常具有科学性和艺术性（焦国力 2009；陈浩 2014）。科学性指的是，从内容来看，让读者学习到科学现象、知识或原理，这部分信息相对更重要；科普作品的艺术性体现为通过生动有趣的、与科学知识相关的背景知识或故事表达作者的情感，令读者获得阅读愉悦感和审美感受（Rakedzon & Baram-Tsabari 2017；张继红、李云海 2010；陈浩 2014）。目前有关这一双重属性是否会影响科普作家在讲解科学内容时和在介绍背景知识时选择不同的语言机制的研究甚少（Hyland 2010）。

作为一种学术文本，科普作品和其他大多数文本一样具有对话性（Bakhtin 1981），需要作者与读者之间进行互动（Hyland 2004；Swales 2004）。科普作者在明确写作目的和目标读者的情况下，有效运用人际功能的语言机制使得读者介入到文本中（Hyland 2010）。近年来，英语已成为国际学术圈交流和沟通的语言（Gordin

2015：2）。因此，研究英文科普文本中介入标记语的使用情况有助于更好地了解英文科普作者如何通过语言的使用达到向读者普及科学知识的目的。

介入标记语是一系列显性修辞手段（Hyland 2001，2005a）。主要包括五大成分：读者人称（direct reader reference）、指令语（directives）、共享知识（reference to shared knowledge）、提问（questions）和插入语（asides）（Hyland 2005a；Hyland & Jiang 2016）。

自Hyland（2001）提出介入标记语的概念以来，诸多学者深入探讨了介入标记语的语用功能（Hyland 2005a，2010；Hyland & Jiang 2016；Masroor & Ahmad 2017；Jiang & Ma 2018；Zou & Hyland 2020；徐昉2013；娄宝翠、王亚丽 2019）。其中，Hyland（2001）认为读者人称和插入语能够将读者视为文本的参与者，满足读者参与学科知识构建的期望，指令语、提问和共享知识能够引导读者解读文本信息。另外，一些学者研究认为介入标记语的使用与文本体裁有密切关系，介入标记语在广告、新闻报道、公共标识用语、学术论文、网络博文、科普作品等中被广泛使用（Hyland 2001；Masroor & Ahmad 2017；Jiang & Ma 2018）。近年来，关于介入标记语的研究主要集中在学术文本上（Hyland 2005a，2005b；Jiang & Ma 2018；Zou & Hyland 2020；周雅 2012；娄宝翠、王亚丽 2019；徐昉 2013）。如Hyland（2010）对比分析了科普文本和研究性论文中介入标记语的使用情况。目前以科普作品作为主要研究对象的研究很少，从科普文本的科学性和艺术性为出发点对比分析其介入标记语使用情况的更是寥寥无几（Plikington 2018）。

《素数之恋：黎曼和数学中最大的未解之谜》（简称《素数之恋》）是由约翰·德比希尔创作的关于“黎曼猜想”的科普书籍（Derbyshire 2003）。正如题目主标题所提示，该书分为两部分，一是“素数”，一是“痴迷”。该书在结构上巧妙地呼应了主标题，采用了双线索结构，一条线索围绕“素数”，另一条则是“痴迷”。全书共22章，其中奇数章节讲解与素数及“黎曼猜想”紧密相关的数学知识；偶数章节中则依次介绍了与“黎曼猜想”有关的数学史以及黎曼、高斯、欧拉、希尔伯特、哈代、兰道等数学大师们的轶事，生动有趣。这两条线索平行推进，一条代表了数学严谨的演算、逻辑和理性思维，体现了该书的科学性；另一条则透过赤诚的文字描写了数学家们的情感、直觉和他们对数学的痴迷和迷恋，同时也呈现了作者对数学大师们的敬意和崇拜，感性之光温暖读者，体现了该书的艺术性。

本研究在语料库检索、统计和分析的基础上，探讨《素数之恋》一书的奇数章节（科学知识）和偶数章节（背景知识）内容和结构上的不同是否也体现在语言使用上，尤其是介入标记语的使用在奇偶章节中是否存在差异。另外，本文还重点分析了《素数之恋》奇偶章节中介入标记语的语用功能和使用特征。本研究期待从互动性元话语的角度深入解读科普文本科学性和艺术性的语言使用机制。

2 研究设计和步骤

2.1 研究语料

本研究基于2003年出版的《素数之恋》英文版一书建库。为便于对比分析，共建两个语料库，均由11个章节组成，一是奇数章节语料库（Odd-number Chapters Corpus，简称OCC），库容为41,814词；二是偶数章节语料库（Even-number Chapters Corpus，简称ECC），库容为49,108词。所有语料均进行文本正规化处理，例如将一般公式替换为\FORMULA，将图表分别替换为\FIGURE和\TABLE。

2.2 研究步骤

首先，使用Natural Language Toolkit（简称NLTK）（Loper & Bird 2002）对正规化的文本进行初步处理，并以句子为单位切分文本。初步统计各章节的平均句长、平均词长和类符/形符比。

接着，确定检索词/句型。基于Hyland（2002，2005a）和Hyland & Jiang（2016）的分类框架以及Jiang & Ma（2018）中的附录，本研究考察的五类介入标记语包括：（1）读者人称9个；（2）指令语，包括祈使动词62个、情态动词7个和“It is +形容词/名词+to do”构式14个；（3）共享知识21个；（4）提问；（5）插入语。

然后，检索并提取索引行。利用Python脚本程序匹配检索词/句型，同时提取所有含有检索词/句型的索引行。接着，我们采用NLKT（Marcus *et al.*1993）和Wmatrix（Rayson 2009）工具包中给出的词性标注结果进行检索词词性的匹配和识别，排除不符合词性要求的索引行（如例1）。不同工具包匹配词性时会出现词性不一致的情况。另外，一些检索词在上下文中并非用作介入标记语（如例2）。人工逐条排查上述情况，以保证索引行的准确率。

（1）Like many other performances, this one begins with a deck of cards.（该索引行中one指“this performance”，并非用作读者人称。）

（2）Having shown you the Golden Key, I now have to begin preparations for turning it.（该索引行中的have to主语为I，因此并不具有读者介入的功能。）

最后，统计并分析数据。统计介入标记语的原始频数和每万词的标准化频数，使用对数似然比计算工具（梁茂成等 2010）和卡方检验计算并分析介入标记语在OCC和ECC两个语料库中的Log值和P值。

3 研究结果与讨论

3.1《素数之恋》奇偶章节的基本语言特征和差异

平均句长、平均词长和类符/形符比能够衡量文本语言的正式程度，它们的数值越大，说明语言越正式。表1显示，偶数章节在这三个维度上均显著高于奇数章节（P<0.001）。

表1 奇偶章节的基本语言特征

	OCC	ECC	P值
平均句长	15.31	19.24	0.000***
平均词长	4.31	4.72	0.000***
类符/形符比	17.95%	25.14%	0.000***

（注：*P<0.05, **P<0.01, ***P<0.001）

《素数之恋》奇数章节的平均句长较短，说明该部分的语言表达倾向口语化，例如较多使用短语式的不完整句子；平均词长较短，较多使用简单易懂的高频词汇；类符/形符比较低表示词汇重复率较高或词汇丰富度较低。奇数章节这三个维度的数值说明作者介绍数学知识时使用的语言比较口语化，符合教师课堂授课的语言特点，营造了课堂话语的氛围，使得读者阅读时身临其境。

相比之下，偶数章节的平均句长较长，使用了较多复杂句式，因此句子长度显著加长；平均词长较大，较多使用正式的书面词汇；类符/形符比较高表示词汇重复率较低或词汇丰富度较高。在介绍数学史和数学大师们的生平与成就时，语言较正式，体现了作者对数学家的景仰之情。

通过对比基本语言特征，可以看出《素数之恋》的奇偶章节在语言上存在显著差异。正如《素数之恋》中文版简介中所说的“极其明晰的数学阐释文字与行文优雅的传记和历史篇章交替出现”，“迷人而流畅”地叙述了黎曼猜想（德比希尔2014），奇数章节语言随意简洁，偶数章节语言正式考究。

奇偶章节在基本语言特征方面的明显差异引起了我们的关注，接下来我们将在介入标记语方面进行更细致的探究。

3.2 《素数之恋》中介入标记语的整体使用频率

本研究检索了五类介入标记语并统计了它们的原始频率和标准化频率。《素数之恋》中介入标记语总计每万词263个，频率从高到低依次为读者人称、插入

语、提问、共享知识和指令语（见表2）。作为使读者介入文本的最为显性的修辞手段（Hyland 2010），读者人称在《素数之恋》中的使用频率最高，标准化频率为114.24个，占介入标记语总数的近44%，这与第二人称在科普作品中的较多使用关系较大。相反，《素数之恋》中的指令语使用最少，标准化频率为18.80个，仅占7.15%，其主要原因是指令语包含命令、指使等强硬的语气，被认为是一种冒险的人际策略（Hyland 2002；娄宝翠、王亚丽 2019），会直接威胁读者面子（Brown & Levinson 1987）。考虑到指令语的使用频率在《素数之恋》中过低，本文的3.4节将重点关注除指令语之外其他四类介入标记语的使用情况。

表2 《素数之恋》中介入标记语的使用频率

	读者人称	指令语	共享知识	提问	插入语	总计
原始频率	1,038	171	248	285	649	2,391
标准化频率	114.24	18.80	27.27	31.34	71.36	263
百分比	43.44%	7.15%	10.37%	11.91%	27.13%	100%

前人研究曾发现数学研究性论文中共享知识是一种较多使用的介入标记语（McGrath & Kuteeva 2012）。而我们发现，《素数之恋》中共享知识作为一种较为隐性的策略，使用频率仅排在第四位。这说明科普文本中很可能倾向于使用更为显性的策略，比如读者人称、插入语或提问。

在检索的过程中，我们发现同一个索引条可能重复出现。比如，例（3）中的索引条同时出现在指令语、读者人称和插入语这三个不同的分类中。作者通过综合使用不同类别的介入标记语，显著增强了读者介入文本的力度。

（3）Imagine you are a very small — infinitesimal, if you can manage it — homunculus, climbing up the graph of the log function from left to right.

3.3 介入标记语在奇偶章节中的使用频率

考虑到《素数之恋》奇偶章节在内容、结构和基本语言特征等方面均呈现出明显的差异，我们进一步观察了奇偶章节中介入标记语的使用情况。表3显示，奇数章节中介入标记语的标准化频率为354.19个，而偶数章节为185.51个；奇数章节比偶数章节显著多用介入标记语（LL=84.22，P<0.001）。这一结果与《素数之恋》结构和内容的组织安排相吻合，体现了该书的写作目的。科普作品的主要目的是普及科学知识，奇数章节的内容正是以数学知识为主。因此，在解说数学

知识时，是否能够让读者介入到文本中是传递和普及科学知识的一个重要环节，需要充分运用诸如介入标记语这样的语言机制来实现该书的写作目的。相比之下，偶数章节在叙述与数学知识相关的人物、历史、地理、文化等背景知识时，读者是否能够介入语篇显得不如奇数章节那么必要，因此这部分显著少用介入标记语。

表3　奇偶章节中介入标记语的使用频率

介入标记语	OCC		ECC		Log值	P值
	原始频率	标准化频率	原始频率	标准化频率		
读者人称	699	167.17	339	69.24	191.07	0.000***
指令语	111	26.55	60	12.22	24.75	0.000***
共享知识	138	33.00	110	22.40	9.27	0.002**
提问	198	47.35	87	17.72	64.08	0.000***
插入语	335	80.12	314	63.94	8.25	0.004**
总计	1481	354.19	911	185.51	84.22	0.000***

具体来看，奇数章节中介入标记语使用频率从高到低依次为读者人称、插入语、提问、共享知识和指令语；偶数章节中介入标记语使用频率从高到低依次为读者人称、插入语、共享知识、提问和指令语。相较于偶数章节，奇数章节尤其显著多用读者人称（LL=191.07，P<0.001）和提问（LL=63.94，P<0.001）。如前文所述，在奇数章节中，作者化身为一位循循善诱、耐心细致的数学教师，在讲台上向读者讲解数学知识。为营造课堂讲解的氛围，奇数章节中较多使用口语化的表达，例如称呼读者为you，或使用提问的方式以引起读者注意。读者人称和提问以显性的方式使得读者非常直接地介入语篇，激发读者主动参与数学概念的思考或数学公式的推导。

3.4　奇偶章节中介入标记语的功能特征和使用特征

3.4.1　读者人称

首先，不论是奇数章节还是偶数章节，读者人称中you/your和we/us/our的标准化频率远远高于one/one's和reader/readers（见表4），这说明科普文本中作者倾向于以第一人称或第二人称直接称呼读者，而不是以第三人称称呼读者。由此可见，第一人称和第二人称的读者介入力度远大于第三人称，更容易拉近作者与读者的距离。同时，我们发现，第二人称在《素数之恋》中的标准化频率为每万词75.17个，是第一人称（每万词36.38个）的两倍多，与Hyland（2010）的研究发现一致。

表4 读者人称频率分布情况

读者人称	OCC		ECC		Log值	P值
	原始频率	标准化频率	原始频率	标准化频率		
you/your	523	125.08	161	32.78	264.34	0.000***
we/us/our	170	40.66	161	32.99	3.62	0.057
one/one’s	1	0.24	7	1.43	4.15	0.042*
reader/readers	5	1.20	10	2.04	0.99	0.319

（注：*P<0.05, **P<0.01, ***P<0.001）

第一人称we/us/our在奇偶章节中的使用无显著差异（LL=3.62，P=0.057），而第二人称you/your的使用呈现明显的差异（LL=264.34，P<0.001）。这说明《素数之恋》奇偶章节中读者人称使用频率的差异主要与第二人称的使用多少有关。

由于作为读者人称的you在句子中既可以用作主格，又可以用作宾格，本研究进一步检索并统计了you的句法使用特征。在《素数之恋》中，用作主格的you为每万词67.53个（占比为89.77%），而用作宾格的you仅为每万词7.70个（占比10.23%）。这一占比情况在奇偶章节中未呈现差异。

同样，本研究发现第一人称we的使用频率（占比70.69%）明显高于宾格us或物格our，在奇数章节中we占比达80%。这在一定程度上说明在英文科普文本中，读者人称（包括第一人称和第二人称）用作主格的情况明显多于宾格的情况。一般情况下，英语句子的主语位于句首或句子开始的位置，而宾格往往位于句尾或句子的后半部分。句首的位置显然比句尾更能引起读者关注。因此，使用主格的读者人称比宾格的人称更为显性，介入力度也更大。

另外，我们还发现，在《素数之恋》中，you经常出现在“if …”或者“as …”从句中，尤其以“if …”居多，几乎与指令语的使用频率相当。“if you …”从句具有与指令语相似的引导或建议读者采取某种行动的功能（如例4）。与指令语不同的是，“if you …”从句更委婉得当，不会伤及读者面子。另外，“if you …”从句也可以帮助读者调用自己的已知信息来建构现有的新信息，树立读者获取新知识的信心（如例5）。

（4）If you differentiate any fixed number, you get zero.

（5）If you ever did a calculus course, this all sounds pretty familiar.

3.4.2 共享知识

根据Hyland & Jiang（2016），共享知识类的介入标记语主要有三种：常规条

件、熟知传统和逻辑推理。通过检索统计OCC和ECC两个语料库中这三种共享知识介入标记语，我们发现，与ECC相比，OCC中显著多用逻辑推理共享知识（见表5）。而常规条件和熟知传统在两个语料库中不存在明显差异。这说明，《素数之恋》奇数章节中为讲解数学知识，作者使用了逻辑推理的途径（如of course），使得读者认识到目前的这一信息推理明了，容易把握，具有可获取性（如例6）。

（6）As far as I am concerned, the iron rule is, one argument, at most one value (no value at all, of course, when the argument isn't in the function's domain.)

表5　共享知识频率分布情况

共享知识	OCC		ECC		Log值	P值
	原始频率	标准化频率	原始频率	标准化频率		
常规条件	21	5.02	36	7.33	1.95	0.163
熟知传统	11	2.63	9	1.83	0.65	0.420
逻辑推理	106	25.35	65	13.24	17.62	0.000***

（注：*P<0.05, **P<0.01, ***P<0.001）

3.4.3　提问

表3显示OCC比ECC显著多用提问（LL=64.08，P<0.001）。奇数章节中，提出或展示与数学知识相关的问题可以使得读者进入特定的数学领域，体验数学家的思维过程（Hyland 2010）；偶数章节以叙述背景知识为主，这些背景知识大多为既成事实，叙述平实客观。

另外，我们发现，提问的读者介入策略经常与读者人称策略共现。经进一步的检索统计发现，第一人称与第二人称相比，问句中较常见的人称为第一人称we，其中奇数章节中，we为主语的问句占11.5%，you为主语的占8.0%；偶数章节中we为主语的问句占9.6%，you为主语的占6.6%。这说明，虽然第一人称单独使用时其介入力度不及第二人称，但当提问时使用第一人称作主语，以鼓励地语气邀请读者进一步思考，提升了表达的介入强度（如例7）。

（7）Can we solve it with closed-form solutions like the vis viva equation?

3.4.4　插入语

本研究分别检索了插入语的三种主要情况：双连线、双括号和by the way短

语。首先，表6显示它们的使用频率从高到低依次为双括号（71.65%）、双连线（26.5%）和by the way短语（1.85%），其中双括号在奇数章节中的使用情况显著多于偶数章节（LL=13.90，P<0.001）。同时，作为介入插入语，by the way短语的使用相对独立，而双括号和双连线的使用可以与其他介入标记语连用，例如读者人称（如例8中的you）或共享知识（如例9中的of course）。

（8）Here, the remainder (once you get past \FORMULA) can be either 1 or 2, and the bias is to 2.

（9）（But not, of course, Gauss.）

表6 插入语频率分布情况

插入语	OCC		ECC		Log值	P值
	原始频率	标准化频率	原始频率	标准化频率		
双连线	77	18.41	95	19.35	0.10	0.748
双括号	254	60.75	211	42.97	13.90	0.000***
by the way	4	0.96	8	1.63	0.79	0.373

（注：*P<0.05, **P<0.01, ***P<0.001）

4 结论和启示

本研究从介入标记语的角度对比分析了《素数之恋》这一科普文本有关科学知识和背景知识内容的语言使用特征。科普作品以普及科学知识为主要目的。一般来说，科学知识比背景知识更加期待读者介入文本，从而有效建构对科学信息的理解。《素数之恋》奇数章节和偶数章节分别叙述了科学知识和背景知识。在语料库分析的基础上，我们首先发现《素数之恋》中奇偶章节的基本语言特征存在显著差异。接着，我们检索、统计并分析了介入标记语在奇偶章节的分布、语用功能和使用特征。研究发现：读者人称的使用频率最高，其中第二人称和第一人称显著多于第三人称，这说明科普文本较多使用显性读者介入策略；奇数章节比偶数章节显著多用介入标记语，尤其是读者人称和提问这两种类型，说明介绍科学知识时作者更期待读者的介入、参与和理解，因此使用了更多的介入标记语。另外，我们还发现了介入标记语的词汇共现现象，不同介入策略的共现提高了介入力度。

未来研究可在扩充语料库的基础上对科普文本的元话语作更深入的探讨。本研究可帮助科普作家更有效地选择语言机制。同时，细致观察科普文本中介入标记语的使用特征和语用功能，对学术英语教学具有一定启发。

参考文献

BAKHTIN M M. Discourse in the novel [C]//HOLQUIST M. The dialogic imagination: four essays. Austin: University of Texas Press, 1981: 259-422.

BROWN P, LEVINSON S C. Politeness: some universals in language usage [M]. Cambridge: Cambridge University Press, 1987.

DERBYSHIRE J. Prime obsession: Bernhard Riemann and the greatest unsolved problem in mathematics [M]. Washington: Joseph Henry Press, 2003.

GORDIN M D. Scientific Babel: science was done before and after global English [M]. Chicago: University of Chicago Press, 2015.

HYLAND K. Bringing in the reader: addressee features in academic articles [J]. Written Communication, 2001, 18(4): 549-574.

HYLAND K. Directives: argument and engagement in academic writing [J]. Applied Linguistics, 2002, 23(2): 215-239.

HYLAND K. Disciplinary discourses: social interactions in academic writing [M]. Ann Arbor: University of Michigan Press, 2004.

HYLAND K. Stance and engagement: a model of interaction in academic discourse [J]. Discourse Studies, 2005a, 7(2): 173-191.

HYLAND K. Representing readers in writing: student and expert practices [J]. Linguistics & Education, 2005b, 16(4): 363-377.

HYLAND K. Constructing proximity: relating to readers in popular and professional science [J]. Journal of English for Academic Purposes, 2010, 9(2): 116-127.

HYLAND K, JIANG F. “We must conclude that…”: a diachronic study of academic engagement [J]. Journal of English for Academic Purposes, 2016, 24: 29-42.

JIANG F, MA X. “As we can see”: reader engagement in PhD candidature confirmation reports [J]. Journal of English for Academic Purposes, 2018, 35: 1-15.

LOPER E, BIRD S. Nltk: the natural language toolkit（arXiv:cs/0205028）[CP]. 2002.

MARCUS M, SANTORINI B, MARCINKIEWICZ M A. Building a large annotated corpus of English: the Penn Treebank [J]. Computational Linguistics, 1993, 19(2): 313–330.

MASROOR F, AHMAD U K. Directives in English language newspaper editorials across cultures [J]. Discourse, Context & Media, 2017, 20: 83-93.

MCGRATH L, KUTEEVA M. Stance and engagement in pure mathematics research articles: linking discourse features to disciplinary practices [J]. Journal of English for Specific Purposes, 2012, 31: 161-173.

PLIKINGTON O A. Presented discourse in popular science: professional voices in book for lay audiences [M]. Boston: Koninklijke Brill NV, 2018.

RAKEDZON T, BARAM-TSABARI A. To make a long story short: a rubric for

assessing graduate students' academic and popular science writing skills [J]. Assessing Writing, 2017, 32: 28-42.

RAYSON P. Wmatrix: a web-based corpus processing environment [CP]. Lancaster University, 2009. http://ucrel.lancs.ac.uk/wmatrix/.

SWALES J. Research genres: exploration and applications [M]. Cambridge: Cambridge University Press, 2004.

ZOU H, HYLAND K. "Think about how fascinating this is": engagement in academic blogs across disciplines [J/OL]. Journal of English for Academic Purposes, 2020, 43.

陈浩. 科学的艺术作品：科普作品的文学性与科学性[J]. 科普研究，2014（3）：79-84.

德比希尔. 素数之恋[M]. 陈为蓬，译. 上海：上海科技教育出版社，2014.

焦国力. 引进文学手法，创立科普美学[J]. 科普创作通讯，2009（1）：7-10.

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

娄宝翠，王亚丽. 学习者英语学术写作介入标记语使用特征[J]. 当代外语研究，2019（4）：58-69.

徐昉. 二语学术写作介入标记语的使用与发展特征：语料库视角[J]. 外语与外语教学，2013（2）：5-10.

张继红，李云海. 浅谈现代科普创作的基本理念[C]// 中国科普研究所. 中国科普理论与实践探索——2010科普理论国际论坛暨第十七届全国科普理论研讨会论文集. 北京：科学普及出版社，2010：680-685.

中华人民共和国人大法工委. 中华人民共和国科学技术普及法[Z]. 北京：法律出版社，2002.

周雅. 关于二语学术写作中介入标记语的比较性研究[D]. 南京：南京大学，2012.

通信地址：100049　北京市　中国科学院大学外语系

小学教师课堂用语情态及人际意义研究*

重庆师范大学 王家锋 西南大学 肖开容

提要：本研究以系统功能语言学情态理论为基础，结合汉语情态研究成果，从情态类别、情态值和情态隐喻三个维度对小学教师课堂用语的情态资源使用特征及人际意义等进行分析。研究发现：教师使用义务情态最多，情态较多服务于提问和命令的表达；教师通过低值的义务情态、纯能力情态的意态化疑问形式和显性主观情态隐喻实现委婉提议；他们倾向于通过使用高值概率情态来维持命题的可协商性与导向性之间的平衡;不同类别的教师间还存在一些情态使用上的显著差异。

关键词：小学教师、课堂用语、情态、人际意义

1 引言

Halliday创立的系统功能语言学认为，情态指“是”和“否”之间各种程度的集合，它构成了肯定和否定之间所有不确定性的区域，体现了讲话者对命题成功性和有效性的判断，或在命令中要求对方承担的义务，或在提议中要表达的个人意愿（Halliday & Matthiessen 2014）。情态是实现人际功能的重要手段之一，另外两个手段是语气和语调（Halliday 1994）。人际功能是指语言所具有的“表达讲话者身份、地位、态度、动机和他对事物的推断、判断和评价等功能”（胡壮麟等2017：110），是语言的参与功能，体现了讲话者作为参与者的“意义潜势”以及交际双方所扮演的角色关系。系统功能语言学研究注重情态表达形式和人际功能相结合这一思路值得国内学者借鉴（杨曙、常晨光2012；何鸣、张绍杰2019）。

国内不少学者运用了Halliday的情态理论来分析语篇的情态及人际功能，如吴平（1995），王振华（2004），王和私等（2011），江玲（2013），任凯，王振华（2017）等。在理论维度方面，这些研究主要关注情态类别、情态值及情态取向等

* 本文系2019年重庆市教委人文社科项目“小学教师话语人际功能探究”（19SKGH041）阶段性成果、2018年重庆市儿童发展与教师教育研究中心重点项目“小学教师话语人际意义研究”（JSJY1801）成果。王家锋为本文通讯作者。
作者贡献：
王家锋：选题构思、数据收集、数据分析、讨论结论、初稿撰写、字数占比（80%）；
肖开容：研究方法、字数占比（20%）、修改润色。

维度，但大都不考虑或比较忽略与人际意义紧密联系的言语功能这一维度。在研究的语言类型方面，现有研究多集中在英语语篇（吴平 1995；王振华 2004）或中英语篇对比方面（王和私等 2011；任凯、王振华 2017），但全面分析汉语语篇的研究较少，其中江玲（2013）对汉语法官语言的情态分析稍有代表性。在分析的语篇内容方面，关注政治、新闻语篇的相对较多，很少有研究关注以汉语作为课堂用语的情态使用情况。教育部2012年2月颁发的《小学教师专业标准（试行）》指出，教师应能“较好使用口头语言”，“使用符合小学生特点的语言进行教育教学工作”，将教师课堂语言能力纳入了专业素质要求。关注和分析汉语作为教学用语的情态特征能为提升教师语言技能提供有益参考。有鉴于此，本研究将运用系统功能语言学情态理论来分析小学教师课堂用语语料中表达人际意义的情态资源使用情况，同时还将分析情态资源在教师职业阶段、性别和授课科目等维度上可能存在的差异。

2　系统功能语言学的情态理论

2.1　情态分类

表1显示了系统功能语言学的情态分类及情态表达成分。Halliday将情态分为情态化（modalization）和意态化（modulation）两大类，其中情态化包含概率和频率，意态化包括义务和意愿。英语的情态表达成分主要有情态操作语、情态形容词、情态附加语等。彭宣维（2000）对英汉情态表达做了对比，他指出汉语情态意义主要通过情态动词（对应情态操作语），又称“能愿动词”和情态副词（对应情态附加语）来表达，同时汉语中语气助词、形容词、名词和实义动词也可能表达情态意义。本研究在情态意义的词汇表达层面主要关注情态动词和情态副词两类形式。

表1　情态类型及表达手段（胡壮麟等2017：148）

基本类别	情态类型	表达手段	例子
情态化	概率	限定性情态动词	They must have known.
		表概率的情态副词	They certainly know.
		以上两者相结合	They certainly must have known.
	频率	限定性情态动词	It must happen.
		表频率的情态副词	It always happens.
		以上两者相结合	It must always happen.

（待续）

（续表）

基本类别	情态类型	表达手段	例子
意态化	义务	限定性情态动词	You must be patient.
		被动谓语动词	You are required to be patient.
	意愿	限定性情态动词	I must win.
		谓语形容词	I’m determined to win.

较多汉语学者（如彭利贞 2007；徐晶凝 2008；范伟 2017；朱斌 2017）都比较接受Lyons（1977）和Palmer（1986）等人基于模态逻辑（model logic）建立的传统语义学情态分类框架。Halliday认为其情态分类与Lyons 和Palmer等人的分类有一定的对应关系。Palmer在继承Lyons的基础上，将情态分为三类：（1）认识情态（epistemic modality），指说话人根据自己所知对命题的真假进行判断；（2）道义情态（deontic modality），表达说话人对行为或事件实现或发生的可能性的判断，判断标准为说话人的要求、指令或社会道义等；（3）动力情态（dynamic modality），涉及行为主体的内部特征，包括该行为主体的能力、意愿以及惯常习性。Halliday认为其情态分类中的概率相当于认识情态，义务相当于道义情态，意愿相当于动力情态。而他所说的频率实际上只是动力情态中惯常情态的一个子类别而已。惯常情态不仅仅表示频率，还可以表示规律性或固化的生活习性，如例（1）和例（2）（范伟 2017：124）。

（1）我骑自行车上班。　　（2）室友睡觉打呼噜。

值得注意的是，Halliday & Matthiessen（2014：696）以及Thompson（2008：67）等倾向于把can/be able to等表示能力/潜力的形式划为低情态值的意愿，但Palmer则将能力和意愿划为动力情态下两个不同的次级情态类别。实际上，具备能力并不一定意味着愿意施展能力，如例（3）和例（4）。

（3）a—这里没有桥你打算怎么过河呢？　　b—我可以游泳。

（4）a—你会什么运动？　　b—我能/会游泳。

例（3b）中的“可以”同时表达了能力和意愿，而例（4b）中的“能/会”仅表达了能力，与意愿无关。由此可见，能力/潜力作为个人的智力、技能或体力仅是个体完成某项任务的前提，并不能表达个人是否具有完成此事的意愿（杨曙、

常晨光 2012：15）。鉴于这种将“能力/潜能”类情态全部划为意愿类情态的主张值得商榷，本研究在认定语料中的“能力/潜力”语义的情态类别时，将包含意愿语义的划分为意愿情态，而将不包含意愿语义的姑且命名为“纯能力”情态。

2.2　情态值

Halliday用高、中、低三个等级的值来标识情态的强弱，表2大致归纳了其量级判定标准。

表2　情态值判断标准（Halliday & Matthiessen 2014：691）

情态值	情态类		意态类	
	概率	频率	义务	意愿
高	certainly	always	required	determined
中	probably	usually	supposed	keen
低	possibly	sometimes	allowed	willing

在Halliday看来，所有情态都是有情态值的。汉语学者将情态值称为“情态强度”（林刘巍 2019）或“情态梯度”（徐晶凝 2008），但汉语学界普遍认为汉语中表达能力语义的情态（即“纯能力”情态）是没有强度可言的（范伟 2017；林刘巍 2019）。例如，我们无法从例（3b）和例（4b）中的“可以”“能/会”来判断技能的熟练程度。

2.3　情态隐喻

系统功能语言学为情态设定了两个维度来判定情态取向（orientation），一是主观/客观维度，二是显现/隐性维度，这两个维度交叉构成了四个情态取向（见表3）。

Halliday将显性主观和显性客观两类取向的表达形式称为情态隐喻，这是由于显性取向的情态使用如“I think”“It’s likely”等投射小句代替了情态动词和情态副词等词汇层面的情态成分，因而被视为情态隐喻。英语情态的显性客观取向往往通过“it”作形式主语的结构实现，如表4中的“It’s expected...”结构。但这类情态表达结构较难在汉语中找到大致对等的结构。因为“It’s expected that John goes.”对应的汉语很可能是“迈克应该去”，汉语中的情态语义表达已经没有形式上的“显性”可言了。因此，本研究在分析教师用语情态隐喻时暂不讨论显性客观形式。另外，如表3所示，Halliday等人认为系统中的显性主观频率和显性主观意愿两个类别是空白的[1]。

表3　情态类型与情态取向的结合（Halliday & Matthiessen 2014：693）

	主观		客观	
	显性	隐性	隐性	显性
概率	**I think**→Mary knows; [**In my opinion** Mary knows]	**Mary'll** know	Mary **probably** knows. [in all probability] [Mary **is likely** to]	**It's likely** that [[Mary knows]]
频率		Fred'**ll** sit quite quiet.	Fred **usually** sits quite quiet	**It is usual** [[for Fred to sit quite quiet]]
义务	**I want**→Mike to go.	John **should** go	Jon's **supposed** to go	**It's expected** [[that John goes]]
意愿		Jonh'**ll** help	Jane's **keen** to help	

2.4　情态与言语功能

Halliday认为语言交流中的交流角色和交换物两个变项的交叉构成了四种言语功能：提供（offer）、命令（command）、陈述（statement）和提问（question）（见表4）。

表4　交流角色和交换物构成的言语功能（Halliday & Matthiessen 2014：136）

交流角色	交换物	
	物品与服务	信息
给予	提供 Would you like this teapot?	陈述 He's giving her the teapot.
求取	命令 Give me that teapot.	提问 What is he giving her?

较之于Lyons和Palmer等人建立的传统语义学情态理论视角，系统功能语言学情态理论的突出特征在于它关注情态的人际功能。以上四类言语功能为分析情态的人际意义提供了基本语义框架，因此本研究在探讨情态的人际意义时，会紧密结合该情态所在小句（命题或提议）的言语功能来分析。

基于以上理论分析，本研究拟对教师课堂用语的情态类别、情态值和情态隐喻三个维度的使用频率进行统计分析，并结合情态小句言语功能分析其人际意义。

在情态类别方面，本研究会在Halliday提出的四类情态基础上增加一类“纯能力”情态，且“纯能力”情态的分析不涉及情态值。在情态隐喻方面，本研究只讨论显性主观概率和显性主观义务。基于该分析框架，本研究拟探究以下问题：（1）教师用语中情态类别、情态值、情态隐喻的分布情况如何？（2）在言语功能基础上，情态资源表达了怎样的课堂人际意义？（3）情态类别、情态小句言语功能及情态隐喻使用频数各自在教师职业阶段、授课科目和性别等分类变量上有何差异？

3　研究方法

3.1　语料采集

分析语料来自32位小学教师的课程，包括优秀在职教师的示范课和小学实习教师汇报课各16节（40分钟/节）。优质课来源于教视网（http://www.sp910.com/）视频。实习教师汇报课视频为重庆师范大学小学教育专业四年级师范生精心准备的展示课现场录像。所有授课均为新授课，全部来自小学高段（五、六年级），教师详细信息见表5。

表5　授课教师基本信息

职业阶段	授课科目		教师性别		授课年级	
	语文	数学	男	女	五	六
在职	8	8	2	14	7	9
实习	8	8	3	13	1	15
合计	16	16	5	27	8	24

3.2　语料处理及差异检验方法

将语料中师生话语转写为电子文本文档，并对每句话进行编号，转写文本总计156,856字，其中教师用语100,639字，共7,546句，句子的单位可能是独立小句或小句复合体。依据教师职业阶段、授课科目和性别对每位教师的语料文档进行编号，然后再按照自制的赋码表对含有情态的句子的情态类别、情态值、情态隐喻和情态小句的言语功能进行手动标注。为降低分析的复杂性，当句子为小句复合体时，仅标注主句，不标注次句。对转写的学生用语未进行标注，仅作为分析情态及人际意义的语境参考。语料统计和分析主要采用Antconc和SPSS等软件。教师间差异检验依据样本特征选择性地使用独立样本t检验或 Mann-Whitney U检

验。为了避免因语速不同造成的教师间总字数差异对频率差异检验的影响，在检验前对每位教师有关检测变量频数以1,000字为单位进行了标准化处理（频数的取值是每1,000字中出现的频数均值）。

4 结果与讨论

4.1 情态基本分布

在教师的7,546句话中，包含情态的有1,005句，占比13.3%。表6显示了各类情态的频数和百分比，其中义务情态的频次占所有情态的60.1%，是使用最多的情态，且义务情态句中以学生主语（如“大家”“同学们”“你”等）的占70.7%，说明情态较多用于教师对学生的义务要求，反映了课堂中教师权力的主导地位。概率、意愿和纯能力三类情态的比例都相对较小，都仅略高于10%。频率情态比例非常小，仅为0.8%。

表6 情态频次及百分比

情态类别	概率	频率	义务	意愿	纯能力	合计
频次	118	8	605	136	138	1,005
百分比	11.7%	0.8%	60.1%	13.5%	13.7%	100%

4.2 情态小句的言语功能分布

表7为情态小句的言语功能频次及百分比情况。可以看出，在教师用语中，情态更多地服务于提问（42.6%）和命令（35%）功能，教师无论是通过提问作为起始话步以获取信息，还是通过命令引导课堂行动，都体现了教师作为课堂主导者的角色。例如，教师使用例（5）中的“想不想”表达提问功能，以获取学生关于“探讨圆的面积公式”的意见，并引导其关注此话题，使用例（6）中的“不要”也体现了教师对学生行动的决定权。陈述功能占21.2%，在这当中教师主要使用概率来实现命题中信息的表达。提供功能仅占1.19%，说明教师用语情态很少服务于提供物品和劳务等功能。

表7　情态小句的言语功能频次及百分比

言语功能	命令	陈述	提问	提供	合计
频次	352	213	428	12	1005
百分比	35%	21.2%	42.6%	1.2%	100%

（5）那你们**想不想**和老师一起探讨圆的面积公式呢?（编号：6-100）

（6）该做笔记的**不要**只顾着听啊！（编号：1-122）

4.3　情态与言语功能交叉统计分析

为了分析情态的具体人际意义，下面将对各类情态及其情态值与情态小句的言语功能进行交叉统计分析。由于频率类情态的使用频次太少（0.8%），以下不作具体讨论。此外，因纯能力情态的情态值缺失，讨论时不会涉及情态值。

4.3.1　义务

教师总计使用义务情态605次。表8显示，其言语功能较多集中于命令（237次）和提问（254次）。

表8　义务的情态值和言语功能交叉频次及百分比

情态值	陈述功能		命令功能		提问功能		提供功能	
	频次	百分比	频次	百分比	频次	百分比	频次	百分比
低	50	45.0%	143	60.3%	142	55.9%	3	100%
中	8	7.2%	17	7.2%	39	15.4%		
高	53	47.7%	77	32.5%	73	28.7%		
合计	111	99.9%	237	100%	254	100%	3	100%

义务情态小句主要通过描述学生义务强度来实现命令功能，如例（7）中的“要”和例（8）中的“可以”，都规定了对学生义务强度的要求。在情态值方面，有60.3%的命令都采用了低值的情态成分，如“可以”。教师使用低值义务情态可使得命令具有建议性和委婉性，体现了教师有意放下权威，拉近与学生的距离，是教师亲和力的表现。

（7）你们**要**认真听讲。（编号：13-188）

（8）大家**可以**自己积累一下，记一记这些词语。（编号：3-15）

义务情态表达提问功能时，体现了客观条件（如事理和事实）对学生所承担义务强度的规定。如教师通过例（9）中的“应该”所表达的义务基于“算术法”的事理或条件；例（10）中的“可以”表达的义务则是基于“这篇课文的主要内容”。由此可见，与表达命令功能的义务来源是教师主观要求不同，表达提问功能的义务来源为事理或条件等客观事实。使用义务表达提问功能给予了学生对义务进行判断和反馈的机会，比仅由教师表达命令更有利于营造良好的课堂互动氛围。在情态值方面，表达提问功能的义务情态有一半以上为低值（55.9%），这增强了提问中商榷的语气。如例（10）中的低值情态动词“可以”就比使用中值的“应该”或高值的“必须”给予了更大的协商空间。

（9）第二题如果用算术法，我们**应该**怎样列式？（编号：4-89）

（10）根据这篇课文的主要内容，同学们**可以**（将其）划为几个部分？（编号：2-30）

4.3.2　意愿

意愿情态总计使用136次。表9显示，大部分意愿服务于提问（53次）和命令（45次）功能。教师使用意愿情态表达提问功能的目的是获取学生关于意愿的信息，如例（11）、例（12）。这一方面体现了教师所扮演的“意愿关照者”角色，另一方面也有助于教师在学生反馈基础上作出适时的教学策略调整。在情态值方面，有62.3%的表达提问功能的意愿情态为中值，例如中值的情态助动词“想”就在总计33次意愿提问中使用了25次，这说明教师在询问学生意愿时，主要充当了一个“中立”的信息索取者角色。

表9　意愿的情态值和言语功能交叉频次及百分比

情态情态值	陈述功能		提问功能		命令功能		提供功能	
	频次	百分比	频次	百分比	频次	百分比	频次	百分比
低	7	24.1%	8	15.1%	19	42.2%	1	11.1%
中	10	34.5%	33	62.3%	21	46.7%	2	22.2%
高	12	41.4%	12	22.6%	5	11.1%	6	66.7%
合计	29	100.0%	53	100%	45	100%	9	100%

（11）同学们，在这个园子里你还**想**扮演什么呢？（编号：32-93）

（12）那你们**愿意**和老师一起探讨一下圆的面积公式吗？（编号：6-100）

而在使用意愿情态表达命令时，教师并未使用与命令功能一致的语气，即祈使语气，而是用如例（13）、例（14）的疑问语气来表达委婉的命令，Halliday将其称为意态化的疑问语气，并将这种使用疑问语气代替祈使语气表达命令的形式看作一种语气隐喻。尽管在（14）中使用了特指疑问词“谁”，但小句语义等价于“你们当中谁如果愿意的话，就请补充”。因而言语功能仍是命令。从情态值来看，表达命令的意愿情态采用中值（46.7%）和低值（42.2%）的频次较多，这些中低值情态成分（如中值“愿意”和低值的“可以”等情态动词）的命令则显得更加委婉。

（13）还有同学**愿意**来试一下吗？（编号：15-240）

（14）谁还**想**补充？（编号：19-176）

4.3.3　概率

概率情态总计使用118次，教师使用概率情态表达了教师对命题真假判断的把握性。表10显示，概率情态主要在表达陈述功能（54次）和提问功能（60次）的小句中出现。

表10　概率的情态值和言语功能交叉频次及百分比

情态值	陈述功能		提问功能		命令功能	
	频次	百分比	频次	百分比	频次	百分比
低	12	22.2%	9	15%	1	25%
中	19	35.2%	8	13.3%	2	50%
高	23	42.6%	43	71.7%	1	25%
合计	54	100%	60	100%	4	100%

表达陈述功能和提问功能的高值概率情态分别占42.6%和71.7%，是两种功能各自使用比例最高的情态值。在表达陈述功能时，教师多用高值概率说明他们对自己给予的命题真假判断把握较大，如在例（15）中，教师通过情态成分“会”

暗示教师对命题判断有较大信心。而在表达提问功能时，多用高值概率说明教师更多地向学生索取对高值概率的判断，如在例（16）中，教师索取的信息是关于“要”变大的事物，而非“或许”变大的事物。由此可见，在教师给予和索取信息的过程中，师生在信息交流中的关系大都为引导者和被引导者间的关系。尽管概率较之于极性命题判断（绝对的肯定或否定）显得更具有开放性，但高值概率情态频次高于中值和低值概率频次，说明这种开放性受教师引导意图的限制。

（15）对，说到非洲啊，很多人都**会**把它与贫穷落后联系到一起。（编号：28-5）

（16）还有什么**要**变大？（编号：157-23）

4.3.4 纯能力

表11显示，纯能力情态主要用于表达提问（42.8%）和命令（47.1%）功能。另据统计，当纯能力服务于这两种言语功能时，大都采用疑问语气（96.8%）。在表达提问功能时，纯能力情态用于询问学生做某事的能力和潜力，如例（17）。其目的往往在于确认学生的能力情况，为下一步教学提供参考。当纯能力服务于命令功能时，往往通过意态化疑问语气表达委婉的命令，如例（18）的目的在于要求学生读得更好一点，但其命令强度明显比祈使语气更低。

表11 纯能力及其言语功能频次及百分比

言语功能	陈述	提问	命令	合计
频次	14	59	65	138
百分比	10.1%	42.8%	47.1%	100%

（17）你**能**理解他这句话的意思吗？（编号：25-41）

（18）**能不能**读得更好一点？（编号：20-79）

4.4 情态隐喻

如前所述，本研究只讨论显性主观概率和显性主观义务两类隐喻。表12显示，表达概率和义务的显性主观隐喻分别占54.2%和45.8%，这类隐喻主要通过投射结构来表达。例（19）、例（20）投射小句使用“觉得”和“相信”等心理动词提示教师对命题的判断具有主观性，从而使教师成为命题的协商者，而不是强势的知识灌输者形象。例（21）、例（22）的投射小句使用“想”“希望”等将学生

的义务（命令功能）投射为教师的想法，从而使命令变得委婉。在Halliday看来，这些通过心理动词来表达情态语义的方式体现了从及物过程到心理过程的投射。表12还统计了语料中此类结构的心理动词的使用频次，有时名物化结构也会代替心理动词来实现投射，如例（23）、例（24）。

表12　显性主观情态隐喻的频次及构成

隐喻类型	频次	百分比	投射结构心理动词频次	名物化结构频次
显性主观概率	26	54.2%	觉得（19）相信（4）想（1）以为（1）	我的理解是.....（1）
显性主观义务	22	45.8%	希望（10）想（8）要（2）要求（1）	老师的要求是.....（1）

（19）你用“压抑”这个词，我**觉得**倒还不至于。（编号：18-101）

（20）我**相信**他和盲姑娘一样，内心十分的激动。（编号：18-173）

（21）老师**希望**有更多的同学愿意举手回答问题。（编号：11-29）

（22）我**想**请这个小组来回答一下。（编号：31-78）

（23）我的**理解**是：既是出于礼貌，也是出于对盲姑娘的尊重。（编号：18-132）

（24）现在老师的**要求**是：用简洁的语言来说一下这个故事的主要内容。（编号：8-6）

4.5　差异分析

分别以授课科目、职业阶段和性别为分类变量，对各类情态类别、情态隐喻和情态小句言语功能等维度中各使用频数变量进行差异检验，发现部分检测变量存在0.1或0.5水平的显著性差异。

4.5.1　职业阶段差异

独立样本t检验结果显示，实习教师和在职优秀教师间在情态类别和情态小句言语功能维度的一些检测变量上存在显著性差异（见表13）。

在情态类别方面，概率和纯能力情态都存在差异。在职优秀教师比实习教师使用更多的概率情态（P=0.000<0.01），这在一定程度上说明，在职优秀教师在与学生进行命题信息交换时更具可协商性，而实习教师则可能更愿意直接使用包含肯定或否定的两极命题，这样的命题显然压缩了协商空间。例如，若将例（25）中表概率的情态副词“好像”去掉，就会在一定程度上降低可协商性。

表13　职业阶段差异独立样本t检验结果

维度	检测变量	在职（N=16）		实习（N=16）		均差	t值	P值
		均值	标准差	均值	标准差			
情态	概率	1.6350	0.7642	0.4394	0.4315	1.1956	5.449	0.000
	纯能力	1.9050	1.4526	0.7731	0.6751	1.1319	2.826	0.008
言语功能	陈述	2.3781	1.0475	1.5169	1.1233	0.8613	2.243	0.032

（25）他发现了有个公交小艇，**好像**有上下两层，对吗？（编号：29-124）

另外，在职优秀教师的纯能力情态使用频率也高于实习教师（P=0.008 < 0.01），结合4.3.4的分析结果来看，说明在职优秀教师更多地通过疑问语气掌握学生的能力和技能情况，以及更多地通过意态化提问表达委婉的命令。

在情态小句言语功能方面，在职教师比实习教师更多地使用表达陈述功能的情态小句（P=0.032<0.05）。由于陈述功能意味着通过命题给予信息，涉及的情态主要为概率，因此两类教师在陈述功能上的差异与在概率使用上的差异是相呼应的，说明在职优秀教师课堂用语更具协商性和开放性。

4.5.2　授课科目差异

语文和数学教师在情态类别和情态小句言语功能中的部分检测变量也存在显著性差异（见表14）。在情态类别方面，两类教师间在义务和意愿两类情态的使用上都存在显著性差异：数学老师比语文老师使用更多义务情态（P=0.000<0.01），而语文老师比数学老师使用更多的意愿情态（P=0.038<0.01）。这说明数学老师比较强调课堂义务，而语文老师更关注师生意愿。此外，在情态言语功能方面，数学老师比语文老师更多地使用情态来实现提问功能（P=0.000<0.01）。

表14　授课科目差异独立样本t检验结果

维度	检测变量	语文（N=16）		数学（N=16）		均差	t值	P值
		均值	标准差	均值	标准差			
情态	义务	4.1113	1.9800	7.6038	2.6798	–3.4925	–4.193	0.000
	意愿	1.6919	1.0148	1.0031	0.7583	0.6888	2.175	0.038
言语功能	提问	2.6838	1.5344	5.4769	1.3734	–2.7931	–5.425	0.000

4.5.3　性别差异

由于样本中仅有5位男教师，受样本特征限制，性别差异检验采用的是Mann-Whitney U 检验（见表15）。

表15　性别差异Mann–Whitney U检验结果

维度	检测变量	女（N=27）		男（N=5）		U值	Z值	P值
		均秩次	秩次和	均秩次	秩次和			
言语功能	提供	15.07	407	24.20	121	29	–2.359	0.018

检验结果显示，在情态小句表达的言语功能中，仅提供功能使用频率存在性别差异（P=0.018< 0.05），结合均秩次看，男教师使用情态服务于提供功能的频次高于女教师。通过观察语料发现，提供功能中使用的情态大都是意愿，如例（26）、例（27），说明男教师在课堂上比女教师更愿意扮演物品和服务的提供者角色。

（26）下课前，老师还**要**送大家一副对联。（编号：25-179）

（27）第一个图形我**可以**用画来解决。（编号：24-330）

4.6　结果与讨论

教师使用的情态资源主要出现在提问功能和命令功能的小句中，充分体现了情态在课堂用语互动功能中的重要作用。教师的情态使用有两个较突出的特点：（1）教师通过使用低值的义务情态，意态化的纯能力情态和显性主观情态隐喻等多种手段来实现委婉提议。（2）教师倾向于使用高值概率情态，说明教师一方面希望使用概率情态保持命题判断的可协商性和开放性，另一方面又通过预设高值概率情态来对命题判断实现导向，这或许说明教师试图追求命题的开放性和导向性之间的平衡。另外，教师间的情态使用差异也是值得关注的，尤其实习教师和在职优秀教师间的差异能对实习教师课堂用语能力的提升有所启示。例如，实习教师在使用情态资源实现课堂用语的可协商性方面与在职优秀教师有明显差距，说明实习教师在该方面的意识和语言技能都需要提升。

5　结语

本研究运用系统功能语言学理论分析小学教师课堂用语情态使用特征，并结合情态小句的言语功能，分析了情态表达的人际意义。一方面，本研究在运用系

统功能语言学理论分析汉语情态方面进行了有益的尝试；另一方面，本研究对小学教师课堂用语多个维度的统计和描写以及所发现的教师间内部差异对职前和在职教师课堂用语能力的培养都能提供有益的参考。由于人际意义的表达还要涉及语气、语调等因素，未来有关教师用语的研究应关注这些维度及其和情态资源交互影响的研究。

注释

1　常晨光（2001）探讨了这两类隐喻存在的可能性。

参考文献

HALLIDAY M A K. An introduction to functional grammar (2nd edition) [M]. London: Edward Arnold, 1994.

HALLIDAY M A K, MATTHIESSEN C. Halliday's introduction to functional grammar [M]. London: Routledge, 2014.

LYONS J. Semantics (Volume 2) [M]. London: Cambridge University Press, 1977.

PALMER F R. Mood and modality (1st edition) [M]. London: Cambridge University Press, 1986.

THOMPSON G. Introducing functional grammar [M]. Beijing: Foreign Language Teaching and Research Press, 2008.

常晨光. 英语中的人际语法隐喻 [J]. 外语与外语教学，2001（7）：6-8.

范伟. 现代汉语情态系统与表达研究 [M]. 北京：中国社会科学出版社，2017.

何鸣，张绍杰. 国外情态研究对汉语语气研究的借鉴与启示 [J]. 外语教学，2019（5）：13-17.

胡壮麟，朱永生，张德禄，等. 系统功能语言学概论（第三版）[M]. 北京：北京大学出版社，2017.

江玲. 情态与身份：功能语言学视角下的法官语言分析 [J]. 语言文字应用，2013（2）：63-71.

林刘巍. 汉语情态强度研究 [M]. 北京：社会科学文献出版社，2019.

彭利贞. 现代汉语情态研究 [M]. 北京：中国社会科学出版社，2007.

彭宣维. 英汉语篇综合对比 [M]. 上海：上海外语教育出版社，2000.

任凯，王振华. 系统功能语言学视角下的英汉情态对比研究——以政治新闻语篇为例 [J]. 当代外语研究，2017（2）：20-26.

王和私，尹丕安，王芙蓉. 中英文政治演说的情态对比研究 [J]. 西安外国语大学学报，2011（2）：38-42.

王振华. 法庭交叉质询中的人际关系——系统功能语言学“情态”视角[J]，外语学刊，2004（3）：51-59.

吴平. 试论情态表达在商业广告英语标题中的运用[J]. 外国语，1995（4）：67-72.

徐晶凝. 现代汉语话语情态研究[M]. 北京：昆仑出版社，2008.

杨曙，常晨光. 情态的评价功能[J]. 外语教学，2012（4）：13-17.

朱斌. 现代汉语情态语气成分的关联机制研究[M]. 北京：中国社会科学出版社，2017.

通信地址：400700　重庆市　重庆师范大学初等教育学院（王家锋）

400715　重庆市　西南大学外国语学院（肖开容）

《中国日报》扶贫报道中的国家形象自塑研究*

西南石油大学 王淑雯 颜镇源

提要：本研究以我国英语主流媒体*China Daily*在“十三五”期间有关“扶贫”的新闻报道为研究对象，将批评隐喻分析与国家形象建构相结合，分析“扶贫”新闻报道中的概念隐喻类型、分布特征和自塑的国家形象。研究发现：我国“扶贫”新闻报道中的概念隐喻共有七种类型，包括冲突隐喻、旅程隐喻、人体隐喻、建筑隐喻、植物隐喻、书籍隐喻和圆形隐喻。概念隐喻类型分布不均衡，源域共鸣值及使用比例最大的是冲突隐喻，其次为旅程隐喻，而人体隐喻、建筑隐喻、植物隐喻、书籍隐喻和圆形隐喻的总占比不足10%。*China Daily*中有关“扶贫”的新闻报道自塑了勇于担当、稳步发展、持续成长及和谐共建的国家形象。本研究对于我国新闻媒体对外自塑国家形象、传播中国声音、提升软实力建设具有一定的借鉴意义。

关键词：扶贫、概念隐喻、批评隐喻分析、新闻报道、自塑国家形象

1 引言

国家形象是一个国家的自我认知与国际体系中他者对其认知的结合（Boulding 1959：123），是体现国家软实力的重要组成部分。在信息社会，新闻媒体是民众获得信息的重要途径，是建构国家形象的重要媒介。国家媒体形象是媒体文本中呈现的具有象征性、隐喻性和代表性的形象群（刘丹凌 2014）。其中，隐喻既是一种语言修辞手段，也是人类思维和认知世界的方式（Lakoff & Johnson 1980：3-6）。新闻报道借助概念隐喻将其所代表的观点间接传递、渗透给读者，影响读者对新闻事件的态度、判断和立场（van Dijk 1988；Moon 1998），具有深层次

* 本文系教育部人文社会科学一般项目“中美硕博学位论文摘要的语类特征对比研究”（19XJA740008）、中国学位与研究生教育学会项目“实验型英语学术论文语料库的建设及其在EAP写作教学中的应用研究”（2020MSA51）、西南石油大学人文社科专项“杰出人才”项目“实验类英语学术论文语料库的建设及其在 EAP 写作教学中的应用研究”（2020RW038）和西南石油大学国际油气资源区语言文化研究中心研究生创新研究项目“我国主流英语媒体关于‘扶贫’报道的概念隐喻研究——以*China Daily*为例”（YQCX2020004）的阶段性成果。王淑雯为本文通讯作者。
作者贡献：
王淑雯：选题构思、研究方法、数据分析、讨论结论、初稿撰写、字数占比（60%）、修改润色；
颜镇源：数据收集、数据分析、初稿撰写、字数占比（40%）。

的“劝说”功能（Charteris-Black 2004：41），有助于实现新闻的传播效果（Stern 2000）、塑造报道对象的形象（Koller 2005：201）。

贫困问题一直是全世界关注的社会问题之一，中国政府在致力于消除本国贫困的同时，也为推进国际减贫进程做出了突出贡献。本研究以批评隐喻分析为指导，以我国英语主流媒体*China Daily*在“十三五”期间的“扶贫”新闻报道为研究对象，分析概念隐喻类型、分布特征，进而探究这些隐喻所自塑的中国国家形象。

2　文献综述

隐喻是一种思维和行为方式，而语言是隐喻的外在表现形式（Lakoff & Johnson 1980：153）。因此，隐喻普遍存在于人们的认知中，对人们认识世界有着深刻的潜在影响。受社会文化的影响，人们通常会选择性地使用一种熟悉的、具体的事物去理解陌生的、抽象的事物，以便更好地认识和掌握外界事物。故而，隐喻被视为意识形态的载体，隐含了隐喻使用者的观点态度（Lakoff & Johnson 1980；Fairclough 1989；Goatly 1997，2007；Moon 1998），具有深层的劝说功能（Charteris-Black 2004：41）。

Charteris-Black（2004）将Fairclough（1995）提出的批评性话语分析模式与Lakoff & Johnson（1980）建构的概念隐喻理论相结合，提出了批评隐喻分析概念，旨在通过对文本中的隐喻进行“识别（identification）—阐释（interpretation）—解释（explanation）”，揭示语言使用者的潜在意图（Charteris-Black 2004：34）。这就意味着，批评隐喻分析是一种揭示潜在意识形态、态度和信仰的方式，也是一种理解语言、思维和社会背景之间复杂关系的重要手段（Charteris-Black 2004：42）。该方法论一经提出便引起国内外学者的关注。研究发现：概念隐喻兼具认知功能、社会功能和文化属性，能够反映特定社团内部共享的价值、信念、态度和意识形态，在解释和改变世界的过程中扮演了重要角色（Koller 2005；Goatly 2007；Meadows 2007；Musolff 2007；Carver & Pikalo 2008；Hart 2008；Koller & Davidson 2008）。

新闻报道借助语言传递人们对世界的认知，通过隐喻手段表达对新闻事件的评价、判断和立场，进而影响读者群体，成为建构国家形象的重要传播手段。国家形象是自我认知与他者认知的结合，是一系列信息输入和输出的结果，是一种信息资本（Boulding 1959）。国家形象被视为国际关系中的软权力，是一个国家“软实力”的重要组成部分。因此，国家形象的塑造与传播研究也受到各国政府的重视。然而，在以西方国家为主导的新闻传播领域，我国国家形象通常被他塑为“好战者”“威胁者”“挑战者”和“扩张者”等负面形象（Su 1991；潘志高 2003；杨雪燕、张娟 2003；胡爱清 2013；赵秀凤、冯德正 2017；汪徽、辛斌 2019；江

进林、贾盼盼 2020），误导了世界民众对中国的认识。但这些研究都是国外媒体对中国国家形象的他塑研究，探讨国内媒体自塑中国形象的研究较少。总体上看，从批评隐喻视角对中国形象的自塑研究刚刚兴起，且很少有专门探讨中国媒体对某一特定事件长期报道的研究。

贫困问题一直是全世界最关注的社会问题之一，我国政府也一直致力于消除贫困。始于2016年的“十三五”规划明确了脱贫是全面建成小康社会的底线任务和标志性指标，是我国全面建成小康社会的收官计划。不过，从概念隐喻视阈展开“扶贫”新闻话语研究的成果仅有王亚聪（2021），她对国内15幅“扶贫”系列平面广告中的多模态隐喻进行了认知研究，发现其反映了人们对同一事件不同的识解方式，具有普遍性以及文化和语境的特殊性。但她并未探讨广告所建构的国家形象。

本研究以我国英语主流媒体*China Daily*在“十三五”期间有关“扶贫”的新闻报道为研究对象，将批评隐喻分析与自塑国家形象相结合，统计分析概念隐喻类型以及分布特征，探究*China Daily*在“扶贫”报道中通过概念隐喻自塑的国家形象。

3 研究设计

3.1 语料库建设

*China Daily*是国内英语主流新闻媒体之一，对于重大事件的报道具有较高的信度和效度，是国外媒体转载率最高的中国报纸，也成为自塑国家形象的重要媒介。本研究将poverty、poverty alleviation、poverty elimination、anti-poverty program、off-poverty等作为关键词对*China Daily*进行检索，收集了2016年3月16日（“十三五”规划在该日正式通过）至2020年12月31日的相关新闻报道，建立语料库。本研究是对新闻话语的单一模态研究，故删除了报道中的标题、图表、图表介绍词等，按日期对新闻语料进行编号命名，如20160316表示2016年3月16日。若当天有多篇相关报道，则编号为2016031601、2016031602等。通过人工筛选，可用语料共计1,250篇，形符数为518,517。

3.2 研究问题

本研究以*China Daily*关于“扶贫”新闻报道为研究对象，运用批评隐喻分析框架，旨在回答以下三个问题：

（1）*China Daily*关于“扶贫”新闻报道中使用了哪些类型的概念隐喻？

（2）*China Daily*关于“扶贫”新闻报道中的概念隐喻呈现的分布特征是什么？

（3）*China Daily*关于“扶贫”的新闻报道自塑了怎样的国家形象？

3.3　研究方法

本研究基于自建语料库，采用定量研究与定性研究相结合的研究方法，采纳Charteris-Black（2004）的批评隐喻分析方法和步骤，对*China Daily*关于“扶贫”新闻报道中的概念隐喻进行识别和阐释。首先，在细读语料的基础上，将人工识别与MIP软件相结合，先识别出候选隐喻（candidate metaphor）。然后，结合语境，确定具有隐喻意义的关键词，并检索统计隐喻词频。本研究采用两人交叉与多次评定相结合的方法，利用信度测试公式“R=n+K/[1+（n-1）+K]”进行计算，得到的评判信度是R=0.933>0.90，说明研究的统计数据具有可信度。

4　结果与分析

4.1　“扶贫”隐喻类型及分布特征

*China Daily*的“扶贫”报道中共出现了七类主导概念隐喻（见表1）。

表1　*China Daily*“扶贫”新闻报道中的隐喻统计

隐喻类型	隐喻关键词数量	出现频次	频次百分比	源域共鸣值[1]	源域共鸣值百分比
冲突	43	1,923	45.7%	82,689	53.07%
旅程	54	1,128	26.8%	60,912	39.10%
人体	25	206	4.90%	5,150	3.30%
建筑	8	615	14.6%	4,920	3.16%
植物	7	249	5.92%	1,743	1.12%
书籍	6	41	0.97%	246	0.16%
圆形	3	46	1.09%	138	0.09%
合计	146	4,208	99.98%	155,798	100.00%

表1显示，*China Daily*“扶贫”新闻报道中运用了大量的隐喻话语策略。首先，我们在语料库中共识别出146个隐喻关键词，使用频次为4,208次，即本语料库中，每一万个词中就有81.2个词具有隐喻意义。其次，隐喻类型丰富，既有常规隐喻，包括冲突隐喻、旅程隐喻、人体隐喻、建筑隐喻和植物隐喻，还出现了两种富有中国文化特色的新奇隐喻（novel metaphor）——书籍隐喻和圆形隐喻。最后，源域共鸣值显示，七种隐喻的分布极不均衡，冲突隐喻和旅程隐喻的源域共鸣值的百分比就达到了92.17%，而另外五种隐喻总体占比仅有7.83%。

4.1.1　冲突隐喻

长期以来，物质抗争始终是人类社会行为的特征之一（Charteris-Black 2004：208）。“冲突”有两个目标，一是为了实现积极的社会目标，如权利、自由、信念等；二是为了反对消极的社会现象，如贫困、不公、疾病等。人类维持自身生存的最基本要求是解决衣、食、住、行等方面的生理需求，“贫困”威胁到了人类的生存，被概念化为“敌人”。“扶贫”是与“贫困”的直接交锋，最终目的是“打败”贫困，消除贫困，保护社会群体的利益。因此，“扶贫是解决冲突问题”这一隐喻在“扶贫”新闻报道中出现的次数、百分比、源域共鸣值及其百分比都是最高的（见表2）。

表2　冲突隐喻统计数据

隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数
fight	361	struggle	31	blow	7	rivalry	2
campaign	265	warn	31	battlefield	7	shield	2
battle	263	deploy	19	resist	6	triumph	2
challenge	187	frontline	15	alliance	5	enemy	1
win	179	guard	15	front-runner	4	foe	1
strategy	135	forge	12	threat	4	hostage	1
victory	86	safeguard	12	battleground	3	rampart	1
war	66	beat	11	shot	3	rival	1
force	64	barrier	9	weapon	3	trigger	1
protect	45	defeat	8	defend	2	victim	1
combat	42	survive	8	defense	2		

表2显示，“扶贫”新闻报道中最高频使用的五个冲突隐喻关键词是fight、campaign、battle、challenge和win，这表明我国政府明确向“贫困”这一敌人宣战，反映了贫困的严重性、扶贫工作的严峻性、我国政府消除贫困的必胜信念，以及已取得的成就。我们根据Charteris-Black（2004：69）基于隐喻关键词的语义取向，将冲突隐喻分为三个次类别：防御隐喻（metaphors of defense）、进攻隐喻（metaphors of attack）、抗争隐喻（metaphors of struggle）。统计显示，防御隐喻占比为5.31%，如protect、defend、fight for等，主要用于反映中国政府为保护社会价值观和贫困人口基本权利所做出的努力。进攻隐喻占比为47.60%，如attack、destroy、defeat、fight against等，该次类将“贫困”视为敌人和进攻目标，

显示了我国政府对于贫困的零容忍态度，主动出击，消除贫困。抗争隐喻占比为47.09%，反映了消除贫困的困难性和艰巨性，以及我国政府不屈不挠的态度。

（1）Shanxi University has recently been playing an active part in Shanxi province's targeted poverty alleviation **campaign**.（来源：2016120801）

（2）China's **fight against** poverty remains tough and the country has entered a crucial stage in its efforts.（来源：2017033101）

（3）The **battle** against poverty is one of the "three tough **battles**" that the country must **win** to build a moderately prosperous society by 2020.（来源：2018053101）

例（1）至例（3）中，campaign、fight against、battle和win都是“扶贫”报道中最为广泛使用的冲突隐喻关键词，其始源域是各种冲突，如战斗、战役、争斗等，目标域是与贫困的抗争和抗争的结果。用“战争”概念来描绘贫困等问题，可以体现出贫困对人类安全和社会稳定可能造成的极大破坏性（贾玉娟 2015）。这些冲突隐喻衬托出扶贫工作刻不容缓，各级政府背负着共同的使命——消除贫困，实现共同富裕。国人不畏艰难、勇于奋斗、敢于抗争、迎难而上的价值观和公民责任感也随之体现。

4.1.2　旅程隐喻

Lakoff（1993：63）将旅程隐喻表述为“有目的的活动是沿着一条通往目的地的旅行”。这不仅强调了旅程的运动特征，还突出了目标导向。在旅程隐喻中，将起点、旅者、目的地、道路、障碍等作为源域，贫困、国家和人民、全面脱贫、扶贫进程、扶贫困难等为目标域。“扶贫”就像是一场旅行，政府、贫困人员和普通民众作为“旅者”携手前行，在国家政策的引导下，克服万难，朝着全面建成小康社会的“目的地”砥砺前行（见表3）。

表3　旅程隐喻统计

隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数
step	156	track	28	landmark	6	turning point	2
drive	90	pace	27	lead	6	finishing line	1
forward	89	guide	26	catch up	4	footprint	1
move	82	course	25	obstacle	4	fuel	1

（待续）

（续表）

隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数
advance	64	milestone	16	passenger	4	harness	1
back	61	smooth	14	crossroads	3	hurdle	1
path	61	direction	13	march	3	kick-start	1
road	46	driving force	12	movement	3	navigate	1
come	43	destination	11	pave	3	pathway	1
start	36	map	11	ride	3	sailing	1
follow	33	channel	10	shift	3	steer	1
explore	31	starting point	9	pass	2	stride	1
journey	30	engine	8	rush	2		
burden	28	route	7	halt	2		

表3显示，“扶贫”新闻报道中最高频使用的五个旅程隐喻关键词分别是step、drive、forward、move和advance，都具有积极评价意义。Charteris-Black（2004）的研究也发现，step是旅程隐喻中使用频率最高的隐喻关键词。step表明向着既定目标稳步前行，并将获得圆满成功；drive则隐含“预设路线、宽阔大道、技术支持、规则意识、快速前行”之意，表达了我国政府全力以赴消除贫困的决心和信心。报道中还用了一些消极评价意义词汇，如burden、obstacle、hurdle等，这表明我国政府充分认识到了“扶贫之旅”任重道远。

（4）Moreover, as China moves **forward** with a nationwide urbanization plan, there are new challenges, such as the lack of jobs and problems with building quality in newly-created and newly-developed urban areas.（来源：2016031105）

（5）Xi stressed the need to strengthen top-level design, adopt more vigorous measures and pool more strength to **advance** rural vitalization, a task no less challenging than poverty alleviation.（来源：2020123001）

（6）The ultimate purpose of poverty alleviation through education is to solve this fundamental **obstacle** to a better life for impoverished children.（来源：2018030602）

例（4）中forward一词有towards a good result、towards the future（朝好的结果发展；面向将来）之意，隐含着国家与人民对全面脱贫的肯定与向往，更是对建成富强民主文明的社会主义国家的美好憧憬。例（5）中的advance传递了只要付出努力，就可发展进步的积极语义，体现了我国“有付出，才有收获”的价值观。而且，forward和advance都隐含“向前”的空间和时间延展趋势，更深层次地展现了我国政府迎难而上的革命乐观主义精神。而例（6）中的obstacle表明了教育扶贫的重重困难，强调了只有教育才能帮助贫困家庭的孩子过上美好生活。

4.1.3　人体隐喻

人类的身体有各种器官，彼此协调，共同维持人体健康。如果某个器官出现病症，就意味着产生了问题，破坏了平衡。身体部位隐喻常常被用来解释器官在功能、重要程度上相似的概念。例如，源域hand的目标域是“携手共进”“帮助”“不可分割”，与所有实践活动相关；heart喻指“核心”“关键”，强调重要性。另外，人类有情感表达、维护社会关系的需求，如face、spirit、friendship等。人体隐喻将国家看作一个有思想、有感情的有机体，即“国家是人”（见表4）。

表4　人体隐喻统计

隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数
hand	27	embrace	9	powerful	4	blood	1
body	27	celebrate	8	dream	3	pressure	1
willing	25	seize	8	hurt	2	disappointed	1
face	21	friendship	8	fear	2	wound	1
heart	15	foot	7	head	2		
force	10	spirit	6	hand out	2		
eye	9	weakness	5	backbone	2		

表4显示，“扶贫”新闻报道中最高频使用的五个人体隐喻关键词是hand、body、willing、face和heart，既有人体部位，又有情感表达，体现了我国政府以人为本的“扶贫”理念。

（7）Poverty distribution in China will undergo major changes in characteristics that the relatively poor replacing the absolute poverty to be the main **body** of poverty population.（来源：2019101507）

（8）China Eastern extends a helping **hand** to Va ethnic group in poverty fight.（来源：2019110701）

例（7）中body是人体隐喻出现频率较高的隐喻关键词，其目标域是贫困人口构成的主体将发生改变。例（8）借助hand表达了中国政府对少数民族佤族脱贫致富的重视，强调了各民族情同手足，传递了共同富裕的国家发展理念。

4.1.4　建筑隐喻

“住宅”是维持人类生活的必需品之一，可泛指人类的物质需求。“建筑物”具有地基扎实、稳固耐用、规划设计、建设时间长等特征。从空间上看，建筑物不断向上方延展，通常被赋予积极内涵，传递了期盼社会经济改善的美好愿景。Charteris-Black（2004：73）指出，“有价值的活动是一栋建筑”。“扶贫”是“有价值的活动”，规划蓝图、打牢基础、规范设计、协同合作等是“建楼”的必要条件，政府是“高楼”的设计者、规划者和践行者，与人民携手成为建设者，埋头苦干，循序渐进，最后建成一栋栋建筑——脱贫致富。建筑隐喻可以激起人们对美好未来的向往，调动人民投身社会主义建设的积极性（黄秋林、吴本虎 2009）（见表5）。

表5　建筑隐喻统计

隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数
build	405	structure	43	threshold	18	bridge	7
foundation	105	framework	19	pillar	11	ground	7

表5中检索到的建筑隐喻关键词除了threshold（门槛）具有消极意义——扶贫遇到的困难和问题外，其余都具有积极内涵。

（9）In 2019, all the 3,613 villages in the county were connected by cement roads, laying a solid **foundation** for further poverty alleviation efforts.（来源：2020011001）

（10）This spectacular achievement came as a result of the Chinese poverty reduction **framework**...（来源：2019111502）

任何建筑如果要坚不可摧、稳固长久，都需有牢固的地基，例（9）中的laying a solid foundation（打下牢固的基础）揭示了乡村基础建设的重要性。例

（10）中的framework（框架）具有支撑性和约束性，表明了我国扶贫所取得的辉煌成就得益于国家政策的正确引导。此外，“基础”和“框架”表明国家注重实现目标之前的准备活动，体现了中国文化“凡事预则立，不预则废”的思想观念。

4.1.5　植物隐喻

植物是自然界中常见的生物物种，生命力旺盛。为了让植物健康成长，人们需谨慎干预，适时除草、施肥、灌溉、修剪等。“贫困”是妨碍植物健康生长的有害“杂草”，必须予以根除，但“除草”任务艰巨，需防止出现“春风吹又生”（返贫）现象（见表6）。

表6　植物隐喻统计

隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数
eradicate	152	root	24	root out	4	take root	1
grow	48	remove	16	nurture	4		

表6中的隐喻关键词依赖语境，都具有积极意义，其中，eradicate、root、remove、root out凸显了我国政府除贫务尽的决心和行动，而grow和nurture反映了国家为帮扶改善贫困户生活生存条件和扶助贫困地区发展生产所付出的努力及取得的成果。

（11）The local government will attempt to **eradicate** poverty in 14 counties and offer villagers more job opportunities.（来源：2017020601）

（12）With the gathering of labor, capital and natural resources in villages, counties and small-sized cities, those places are likely to **nurture** industries based on their natural resources.（来源：2020072901）

例（11）表明地方政府为彻底根除（eradicate）“贫困”这一杂草所做出的承诺和采取的具体行动，隐性传递了巩固脱贫攻坚成果、杜绝“返贫”现象的决心。例（12）中的nurture隐含了长远性、保护性、鼓励性、资助性等积极语义取向，表达了全国人民齐心协力，精心扶持企业发展，帮助贫困群众合理利用得天独厚的自然资源脱贫致富的意愿。

4.1.6　书籍隐喻

本研究还发现了一个新奇隐喻——书籍隐喻。书籍是人类文明的载体，人类生活的精神必需品，可记录描述、传承延续、回顾反思等。要完成一本书则并非易事，需全面规划、选择主题、表明目的、撰写大纲、制定规划、执行细节、斟酌权衡等

（见表7）。

表7　书籍隐喻统计表

隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数
picture	17	chapter	5	content	2
context	10	page	5	profile	2

表7显示，书籍隐喻聚焦书籍的宏观结构（如picture、context、content）和微观构成（如chapter、page、profile）。

（13）If all of them were on the same **page** as China, every obstacle in the path toward development could be overcome.（来源：2020072202）

（14）..., each providing a piece of the overall **picture** of China's accomplishment in lifting people out of poverty and China's contributions to global poverty reduction.（来源：2020093001）

例（14）中的page是书籍的组成部分，体现了中国的发展进程，从历时视角说明中国此时的发展状况，以便掌握国家的发展进程，并更好地做出下一阶段的计划和部署等。例（15）中的picture则是以全面视角展现国家扶贫的既有成就，体现了国家对于全局的把控。

4.1.7　圆形隐喻

圆形作为一种几何图形普遍存在于人类的日常生活中。一个圆形包括圆心、圆周以及从圆心到圆周的部分。圆形隐喻是一种富有中国文化特色的新奇隐喻。在中国传统文化中，“圆”被赋予团圆、美满、幸福、快乐、富足、稳定、团结等积极意义，成为重要的精神原型。在“扶贫”圆形隐喻中，国家就是一个圆形，党和政府就是这个圆形的圆心，人民填满了圆心到圆周的部分。为了全面建设小康社会，人民与政府团结一心，群策群力，期盼从根本上摆脱贫困，过上圆满、富足、稳定的小康生活（见表8）。

表8　圆形隐喻统计表

隐喻关键词	出现次数	隐喻关键词	出现次数	隐喻关键词	出现次数
core	31	round	10	center	5

表8中的两个隐喻关键词core和center在语境中都喻指以习近平同志为核心的党中央，表明只有坚持中国共产党的领导，坚决贯彻执行党的战略安排与部署，才能脱贫致富，最终实现伟大的中国梦。

（15）The whole Party and nation will closely unite around the CPC Central Committee with Xi as the **core**.（来源：2017030202）

例（15）中将习近平总书记喻为圆心（core），周围围绕着中共中央、各级政府和普通民众，共同形成了一个完整的圆形。由于任何缺失都不能构成一个完整的圆，所以圆形隐喻也凸显了中国各民族的高度凝聚力和团结意识。

4.2 “扶贫”隐喻自塑的国家形象

统计分析发现，在*China Daily*“扶贫”新闻报道中高频使用了七种概念隐喻。依据语境分析，大部分隐喻关键词的语义韵都呈现出了积极意义，折射出我国党和政府以人为本、共同富裕的宗旨，“除贫务尽”的信心，以及求真务实、迎难而上的躬行实践。少数具有消极意义的隐喻关键词表明了贫困的严重性和消除贫困的艰巨性，传递了中国英语主流媒体对“扶贫”事件实事求是的态度（见表9）。

表9 “扶贫”新闻报道隐喻及其自塑的国家形象

隐喻类型	语义韵	“自塑”国家形象
冲突隐喻	积极	实事求是、不畏艰难、坚韧不屈、积极主动、担当使命
旅程隐喻	积极	求真务实、脚踏实地、砥砺前行、稳步推进
人体隐喻	积极	以人为本、协同发展、共同富裕
建筑隐喻	积极	打牢基础、质量为重、精准设计、稳扎稳打、踏实肯干
植物隐喻	积极	遵循规律、躬行实践、除“贫”务尽
书籍隐喻	积极	高屋建瓴、深思熟虑、精心设计
圆形隐喻	积极	勇于担当、群策群力、团结一致、共同富裕

冲突隐喻体现了我国政府和人民“反贫困”的心理，凸显了必定完成脱贫任务的内生动力，为推进国际减贫进程、推动人类命运共同体的建设努力奋斗，自塑了实事求是、勇闯难关的大国形象。旅程隐喻突出了国家扶贫进程与未来发展目标，体现了人民在追求国家繁荣富强道路上砥砺前行，构建了一个求真务实、踏实奋进

的“发展”国家形象。书籍隐喻、建筑隐喻和植物隐喻凸显了以人为本的理念和除贫务尽的目标，塑造了一个躬行实践、踏实肯干、务实进取的“成长”国家形象。人体隐喻和圆形隐喻则塑造了一个胸怀大志、团结一致、同心同德的“和谐”国家形象。然而，由于冲突隐喻常与战争、牺牲、痛苦、灾难等联系起来，如果不依赖语境，过度使用冲突隐喻可能会将一个国家塑造成好战形象。相关研究发现：外媒对我国的新闻报道中多采用战争隐喻，他构的中国形象多是消极负面的（汪徽、辛斌 2019；江进林、贾盼盼 2020）。

概言之，中国主流英语媒体*China Daily*的“扶贫”新闻报道，通过七种概念隐喻自塑了勇于担当、稳步发展、持续成长及和谐共建的正面国家形象，生动形象地向读者传递了中国为国内脱贫攻坚、推进国际减贫进程、推动构建人类命运共同体所做出的重大贡献。

5 结语

媒体在国家形象的建构与传播中发挥着重要作用，由媒体塑造的国家形象会长期影响人们的思维方式和观点态度。本研究以我国英语主流媒体*China Daily*在“十三五”期间有关“扶贫”的新闻报道为研究对象，将批评隐喻分析与自塑国家形象相结合，统计分析概念隐喻类型以及分布特征，探究*China Daily*在“扶贫”报道中通过概念隐喻自塑的国家形象。研究发现：（1）隐喻类型丰富。相关报道中既有冲突隐喻、旅程隐喻、人体隐喻、建筑隐喻和植物隐喻等五种常规隐喻，还出现了书籍隐喻和圆形隐喻两种富有中国文化特色的新奇隐喻。（2）七种隐喻的分布极不均衡。使用比例最大的是冲突隐喻（53.07%），其次为旅程隐喻（39.10%），而人体隐喻、建筑隐喻、植物隐喻、书籍隐喻和圆形隐喻的总占比仅有7.83%。（3）概念隐喻自塑了勇于担当、稳步发展、持续成长及和谐共建的积极国家形象。

然而，本文虽然通过语境分析发现，冲突隐喻传递了我国不畏艰难、勇于奋斗、敢于抗争、迎难而上的价值观和公民责任感，但过度使用有可能会被误解，甚至曲解为“好战”“威胁”。在今后的相关报道中，可适当提高具有积极语义特征的人体隐喻和建筑隐喻的使用比例，引导舆论导向，更多使用具有中国文化特色的书籍隐喻和圆形隐喻，践行中国文化走出去的战略方针，提升国家“软实力”建设。

注释

1 源域共鸣值=隐喻关键词数 × 出现频次，用于计算隐喻的使用情况并确定源域的普遍性。

参考文献

BOULDING K E. National images and international system [J]. Journal of Conflict Resolution, 1959, 3(2): 120-131.

CARVER T, PIKALO J. Political language and metaphor [M]. London: Routledge, 2008.

CHARTERIS-BLACK J. Corpus approaches to critical metaphor analysis [M]. New York: Palgrave Macmillan, 2004.

FAIRCLOUGH N. Language and power [M]. London: Longman, 1989.

FAIRCLOUGH N. Critical discourse analysis [M]. Boston: Addison Wesley, 1995.

GOATLY A. The language of metaphors [M]. London/New York: Routledge, 1997.

GOATLY A. Washing the brain: metaphor and hidden ideology [M]. Amsterdam: John Benjamins, 2007.

HART C. Critical discourse analysis and metaphor: toward a theoretical framework [J]. Critical Discourse Studies, 2008, 5(2): 91-106.

KOLLER V. Critical discourse analysis and social cognition: evidence from business media discourse [J]. Discourse & Society, 2005, 16(2): 199-224.

KOLLER V, DAVIDSON P. Social exclusion as conceptual and grammatical metaphor: a cross-genre study of British policymaking [J]. Discourse & Society, 2008, 19(3): 307-331.

LAKOFF G. The contemporary theory of metaphor [C]//ORTONY A. Metaphor and Thought. Cambridge: Cambridge University Press, 1993: 59-107.

LAKOFF G, JOHNSON M. Metaphors we live by [M]. Chicago: University of Chicago Press, 1980.

MEADOWS B. Distancing and showing solidarity via metaphor and metonymy in political discourse: a critical study of American statements on Iraq during the years 2004-2005 [J]. Critical Approaches to Discourse Analysis across Disciplines, 2007, 1(2): 1-17.

MOON R. Fixed expressions and idioms in English: a corpus-based approach [M]. New York: Oxford University Press, 1998.

MUSOLFF A. Popular science concepts and their use in creative metaphors in media discourse [J]. Metaphorik.de, 2007, 13: 67-85.

STERN J. Metaphor in context [M]. Cambridge: MIT Press, 2000.

SU S. Changing American images of China as reflected in the New York Times, the Washington Post and the Christian Science Monitor, 1972-1985 [D]. Manoa: University of Hawaii, 1991.

VAN DIJK A. News analysis [M]. Hillsdale: Lawrence Erlbaum Associates, 1988.

胡爱清. 美国新闻报道中他者形象建构的隐喻作用[J]. 广东技术师范学院学报，2013（8）：89-92.

黄秋林，吴本虎. 政治隐喻的历时分析——基于《人民日报》（1978—2007）两会社论的研究[J]. 语言教学与研究，2009（5）：91-96.

贾玉娟. 战争隐喻广泛性之理据分析[J]. 学术界，2015（12）：148-153.

江进林，贾盼盼. 西方媒体中的北京环境——基于语料库的批评隐喻分析[J]. 语料库语言学，2020（1）：32-43.

刘丹凌. 论国家形象的三重内涵——基于三种偏向的分析[J]. 南京社会科学，2014（5）：106-114.

潘志高.《纽约时报》对华报道分析：1993—1998[J]. 贵州师范大学学报（社会科学版），2003（3）：52-55.

汪徽，辛斌. 美国媒体对中国形象的隐喻建构研究——以“美国退出TPP”相关报道为例[J]. 外语教学，2019（3）：32-38.

王亚聪.“扶贫”系列平面公益广告中多模态隐喻的认知研究[J]. 中州大学学报，2021（1）：54-59.

杨雪燕，张娟. 90年代美国大报上的中国形象[J]. 外交学院学报，2003（1）：41-48.

赵秀凤，冯德正. 多模态隐转喻对中国形象的建构——以《经济学人》涉华政治漫画语篇为例[J]. 西安外国语大学学报，2017（2）：31-36.

通信地址：610500　四川省成都市　西南石油大学外国语学院

语域与语料规模在语义韵研究中的影响*

中南财经政法大学 李中正

提要：在语义韵相关研究中，语域的区分与语料规模的控制长期以来为部分研究者所忽视。本文以最高程度副词entirely为节点词，分别在COCA语料库的总库，以及小说、学术、报刊、口语、小样本、中样本、大样本七个子库中进行检索，进而探究其语义韵在不同语域和不同语料规模中的差异。研究结果显示，就语域而言，节点词entirely的语义韵在小说、学术、报刊和口语四大语域中均存在较为显著的差异；就语料规模而言，节点词entirely的语义韵仅在大样本当中部分还原了COCA总库中的原貌，而在小、中样本中与在COCA总库中相去甚远。因此，研究者应将语义韵研究置于特定的语域当中，并尽量避免因样本数量过小而产生的负面影响。

关键词：语义韵、扩展意义单元、语域、语料规模

1 引言

语义韵（semantic prosody）的概念滥觞于Sinclair（1987）对短语set in的研究，而后由Louw（1993）用为术语，并在接下来的20余年中受到国内外众多学者的广泛关注，成为语料库语言学的重要发现之一（卫乃兴 2011）。相关研究认为，节点词与搭配词的语义相互渗透，相互影响，在语境当中形成特定的语义氛围，以此表达说话者的态度与交际目的（卫乃兴 2002）。这种语义氛围被称为语义韵，是节点词与周围相关搭配词的语义特征交互影响所产生的意义（卫乃兴 2011）。

随着语义韵理论探索的深入与研究方法的拓展，国内外众多学者对语义韵的相关研究取得了大量成果，语义韵研究整体呈现初步繁荣与快速发展的态势。然而，在大量的语义韵相关研究中，仅有少数学者（Stubbs 2001；Partington 2004；Hunston 2007；Bednarek 2008；杨晓琳、程乐 2016）在对比语义韵的过程中区分了语域，大部分学者使用的对比语料为通用语料库而非专业文本语料库。此外，许多学者在语义韵研究当中频繁使用随机抽取或隔行抽取的方法获取语料，致使研究中体现的语义韵为特定规模样本中的语义韵，而非整个语料库中的语义韵。那么，在语义韵研究中，不同的语域和不同的语料规模是否会对研究结果形成干

* 本文系中南财经政法大学中央高校基本科研业务费专项资金资助（2722021EK011）项目成果。

扰呢？本研究基于COCA语料库及其所生成的不同语域、不同语料规模子库，以最高程度副词entirely为例，探究语域与语料规模在语义韵研究中的影响，并为日后的语义韵研究提出相关建议。

2 语义韵与扩展意义单元相关研究

在Firth（1957）对搭配（collocation）和类联接（colligation）论述的基础上，Sinclair（1987）发现了词语间语义的相互影响，而Louw（1993）则将这一现象提炼为语义韵这一概念。随后，Sinclair（1996）提出语义趋向（semantic preference），对搭配词语义特征进行概括描述，并将以上诸多概念整合在一起，构建了扩展意义单元（extended unit of meaning）研究的基本框架。Sinclair（1996）将词项的扩展意义单元定义为该词项的搭配、类联接、语义韵和语义趋向。语义韵传递观点、立场、态度，具有典型的评价功能，考察说话人的情感态度和情感倾向。语义趋向则反映出对共现词项语义特征的概括。类联接描述的是节点词与共现词项的语法结构。搭配体现出节点词与共现词项在词汇层面的组合关系。可以看出，扩展意义单元研究囊括了词汇、语法、语义、语用诸多层面的研究，是典型的形式、意义与功能的复合体（卫乃兴 2008）。

作为扩展意义单元模型中的核心，语义韵制约词汇、语法选择，界定意义单位的边界（Sinclair 2004）。在扩展意义单元模型中，语义韵是由整个扩展意义单元共同营造的，每个要素都与语义韵的形成有着千丝万缕的联系，尤以语义趋向为甚。语义趋向是语义韵形成的基础，而语义韵反过来又影响和制约着语义趋向（王均松、田建国 2016）。因此，国内外对语义韵的研究已经从早期单纯研究语义韵本身，扩展到了对整个扩展意义单元的研究。

在研究对象方面，国内外许多学者已经将目光从对单一语言中语义韵的描述扩展到了跨语言的语义韵对比研究（Xiao & McEnery 2006；卫乃兴 2011；李晓红、卫乃兴 2012；赵朝永 2014；余渭深、李中正 2017；高歌、卫乃兴 2019）。Partington（2004）认为，语义韵研究应控制在相同的语域中进行，但鲜有学者在语义韵对比研究中进行语域的区分，大部分研究中的对比语料库均为通用语料库而非专用语料库，致使研究语料与对比语料产生了语域差别。例如，伍晓飞（2019）在进行ghost一词的语义韵对比研究时，使用的研究语料为小说《歌剧魅影》，而对比语料则为COCA语料库总库。研究语料为小说，对比语料是否应该为相应的英语原创小说库呢？对比语料中存在的语域差别是否会对研究结果造成影响？产生影响的原因有哪些？这些问题尚未得到解释，值得学界关注。

在研究方法方面，学界关于语义韵的研究总体上遵循以下步骤（陆军、卫乃兴 2014）。首先，利用统计抽样手段（随机抽样或隔行抽样）从语料库中提取足

够数量的索引行。其次，观察索引行并确定类联接；再次，参照类联接检查和概括搭配词语义特征。最后，归纳语义韵。值得思考的是，抽样提取多少索引行才能称得上是“足够数量”的索引行呢？语料规模要达到何种程度才能有效地代表整个语料库呢？笔者对2010—2019年在中国知网（CNKI）上发表的有关语义韵研究期刊论文进行了梳理，从中挑选出采用抽样方法提取语料的期刊论文，标记出每篇论文所抽取的语料规模（见图1）。可以看出，大多数研究者从语料库中抽取的索引行数目在200条以下，小部分研究者抽取400条左右的索引行，而抽取1,000条以上的研究屈指可数。

诚然，语义韵研究发展态势一片大好，发表论文数量逐年增多，研究内容、研究视角、研究方法逐渐呈现多元化趋势（戴建春 2018），然而在成熟的研究范式之下，隐藏着的是众多研究者对相应语料库的语域、语料规模等要素的忽视。因此，语域和语料规模对语义韵研究的影响还亟待学界进一步探索。

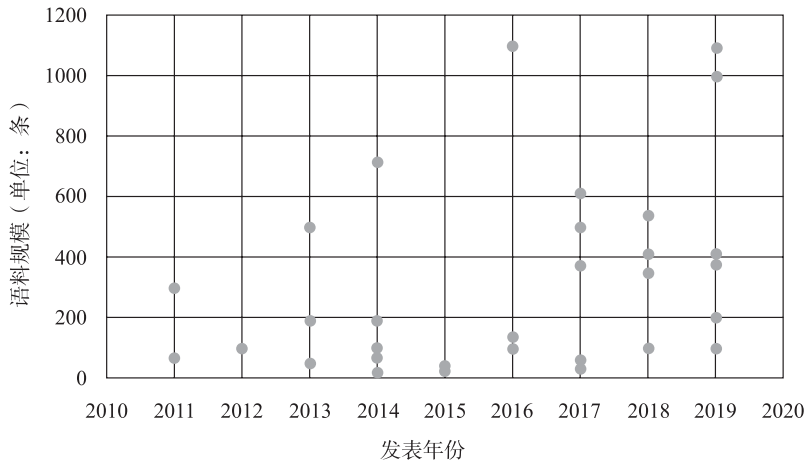


图1　2010—2019年CNKI检索中语义韵相关期刊论文语料规模分布

3　研究设计

3.1　研究问题

为探索语域与语料规模在语义韵研究中的影响，本研究提出以下三个问题：

（1）语域对节点词语义韵有何影响？

（2）节点词语义韵在不同语域之间存在何种差异？

（3）节点词语义韵因语料规模的不同会产生何种变化？

3.2 工作定义

本研究以扩展意义单元模型为框架，在确定类联接、概括语义趋向的基础上归纳节点词的语义韵。在语义韵的分类方面，本文采用Stubbs（1996）和卫乃兴（2002）对语义韵的三种分类：积极语义韵、消极语义韵和中性语义韵，即若节点词吸引具有积极/消极/中性语义色彩的搭配词，则该节点词所在的扩展意义单元具有相应的语义韵。语义韵特征判定标准采用李晓红、卫乃兴（2012）提出的语义韵力度（prosodic strength），即节点词表达态度意义的频数与总频数的比值。语义趋向判断标准采用Stubbs（2001）提出的语义归纳法，对词项高频搭配词的语义进行分析和归纳，进而揭示构成节点词的语义趋向。

3.3 研究对象与语料选取

本文选取最高程度副词entirely为研究对象，为该词项在各主流语料库中高频出现，语义韵特征较为明显（余渭深、李中正 2017），便于进行比较和深入观察。语料选取于COCA语料库，该语料库由杨伯翰大学Mark Davis教授研制，包含小说、学术、报刊、口语几大子库，词容量达4.5亿，便于提取构建出不同语域、不同语料规模的子库。为了探究不同语域和不同语料规模对节点词entirely语义韵构建的影响，本研究将COCA中的子语料库分为两组，一组为区分语域的子语料库，选取COCA中的小说、学术、报刊、口语四大子库作为不同语域语料库；一组为区分语料规模的子语料库，从COCA中以随机抽取的方式，自建小样本语料库、中样本语料库与大样本语料库作为不同语料规模语料库。根据笔者对国内语义韵研究抽样规模的调查结果（见图1），本研究拟将小样本COCA语料库规模确定为100条索引行，中样本COCA语料库规模确定为500条索引行，大样本COCA语料库规模确定为1,000条索引行。本研究中的参照语料库选用COCA总库，而语料库检索操作利用COCA自带检索功能即可实现。

3.4 研究步骤

首先，进行索引行提取、观察与统计：将最高程度副词entirely作为节点词，分别在COCA总库、COCA小说子库、COCA学术子库、COCA报刊子库、COCA口语子库、COCA小样本语料库、COCA中样本语料库以及COCA大样本语料库中进行检索，左右跨距选取为4，取MI值大于等于3的搭配词为显著搭配词（汪腊萍 2006；李文中 2017）。观察各个语料库中节点词entirely的类联接共选形式，概括各自搭配词语义趋向，进而归纳语义韵。其次，进行对比分析：分别将四类不同语域子库和三类不同语料规模子库中节点词entirely的类联接、语义韵力度和语义趋向与COCA总库中的检索数据进行对比，探索节点词语义韵因语域和语料规模不同而可能产生的差异。最后，讨论节点词语义韵因语域和语料规模而变化

的影响，并为日后的语义韵相关研究提出建议。

4 数据统计

4.1 搭配词特征

将节点词entirely分别在COCA总库和七个子库（COCA小说子库、COCA学术子库、COCA报刊子库、COCA口语子库、小样本COCA语料库、中样本COCA语料库、大样本COCA语料库）中进行检索，得到的显著搭配词个数（类符）与例数（形符）如表1所示。为考察节点词entirely在各个子库与总库中显著搭配词的相似程度，本研究提取了各个子库中排名前50位的高频显著搭配词，与COCA总库中的前50位高频显著搭配词相比较，梳理出相同的显著搭配词个数作为考察子库与总库中搭配词相似性的依据。

表1 节点词entirely在各个语料库中显著搭配词个数与例数

语料库	显著搭配词类符	显著搭配词形符	COCA总库高频搭配词个数
COCA总库	288	7,981	50
COCA小说子库	143	1,585	26
COCA学术子库	157	2,004	36
COCA报刊子库	205	3,205	39
COCA口语子库	147	1,420	28
小样本子库	83	98	6
中样本子库	55	133	10
大样本子库	126	421	19

显而易见，显著搭配词的个数和例数与语料规模息息相关，语料库规模越大，显著搭配词的形符与类符越多。在子库与总库中搭配词相似性方面，四类不同语域语料库与三类不同语料规模语料库的情况截然不同：前者语料规模均较大，因此与COCA总库高频搭配词相同的个数也较多，后者语料规模较为有限，与COCA总库高频搭配词的相同个数相去甚远。小样本语料库中仅有6个高频搭配词与COCA总库相同，即便是抽取索引行数目达到1,000条的大样本子库，与COCA总库相同的高频搭配词也只有19个。而在四种不同语域子库当中，COCA学术子库和COCA报刊子库两种语域中的高频搭配词与COCA总库中的相似性较

高，COCA小说子库和COCA口语子库中的搭配词相似程度较低。

4.2 类联接特征

节点词entirely词性为副词（Adverb），根据语料库检索结果并结合英语语法规则，发现该节点词的搭配词均为动词（Verb）或形容词（Adjective）。因此，节点词entirely的常见共选形式为entirely+V和entirely+Adj两类。两组语料库中节点词entirely的类联接特征分别如图2、图3所示。

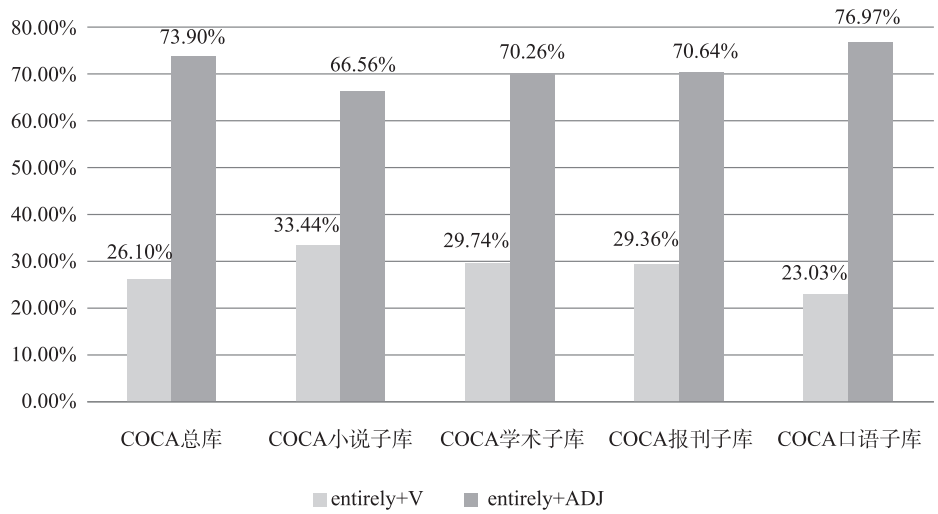


图2 不同语域中节点词entirely的类联接特征

图2显示了节点词entirely在COCA总库与四个不同语域的COCA子库中的类联接特征。在COCA总库中，节点词entirely的显著搭配词类符出现了288个，显著搭配词形符则为7,981例，分布在entirely+V和entirely+Adj两种类联接当中。其中，entirely+V的类联接形式出现了2,084例，高频显著搭配词如separate、eliminate、disappear等，占总例数的26.10%；entirely+Adj的类联接形式出现了5,897例，高频显著搭配词如different、possible、clear等，占总例数的73.9%。可见，在COCA总库中，节点词entirely与形容词共现的倾向性极强，与动词共现的倾向性则稍弱。

与COCA总库中的检索数据相比，节点词entirely在不同语域子库中呈现的类联接特征存在普遍差异，但差异并不大，大体能与COCA总库中的数值保持正负10%的浮动，维持住entirely+V和entirely+Adj两种类联接形式三七开的局面。在COCA小说子库中，entirely+V的类联接形式出现了530例，高频显著搭配词如forget、compose等，占总例数的33.44%，这一数值也是所有语域

子库中最高的；entirely+Adj的类联接形式出现了1,105例，高频显著搭配词如different、true、possible等，占总例数的66.56%。COCA学术子库与COCA报刊子库的类联接特征几乎相同，前者entirely+V和entirely+Adj的类联接形式占比分别为29.74%和70.26%，后者entirely+V和entirely+Adj的类联接形式占比分别为29.36%和70.64%。各自的高频显著搭配词也无显著差异，高频共现动词均为depend、eliminate等，高频共现形容词均包括different、clear等。COCA口语子库中entirely+Adj的类联接形式出现例数为1,093，高频共现形容词有different、possible、clear等，占总例数的76.97%，这一比例为所有子库中最高值；entirely+V的类联接形式出现了327例，高频共现动词包含agree、depend、focus等，占总例数的23.03%。

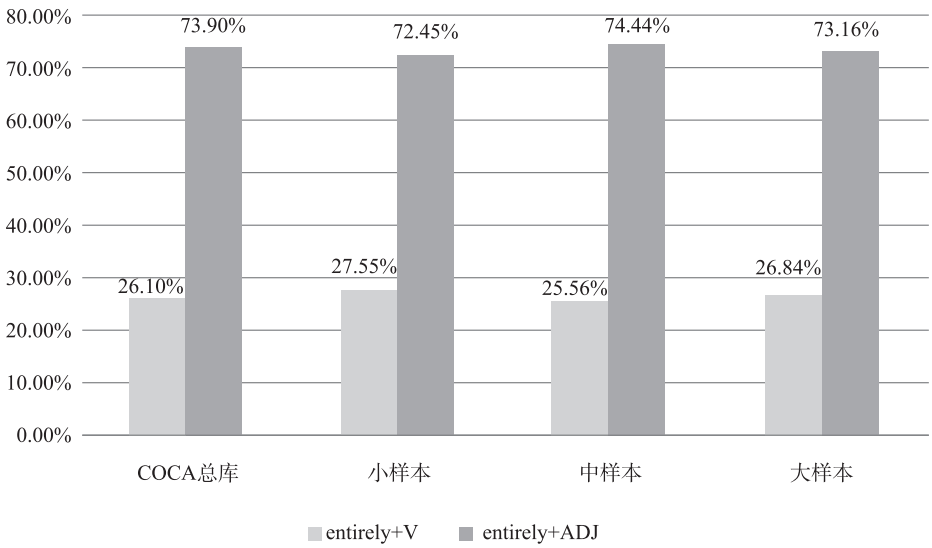


图3　不同语料规模中节点词entirely的类联接特征

图3显示了节点词entirely在小、中、大三种规模语料库中的类联接特征。在小样本COCA语料库当中，entirely+V型类联接出现27次，高频搭配词主要为occur、matter、give等，占总例数的27.55%；entirely+Adj型类联接则出现了71次，高频搭配词例如different、clear、domestic等，占总例数的72.55%。两种类联接形式在中、大样本COCA语料库中的分布大致与小样本语料库中的数据相当，分别为entirely+V型占比25.26%、entirely+Adj型占比74.44%和entirely+V型占比26.84%、entirely+Adj型占比73.16%。高频搭配形容词也均为different、clear、possible等。然而，高频搭配动词差异显著，在中样本COCA语料库中节点词entirely的高频搭配动词为forget、surprise和separate等，而在大样本COCA语料库中则为free、depend、clean等。可见，三种规模语料库中节点词entirely类联接类

型与总库并无较大差异，主要的差异仍体现于搭配词方面。

4.3 语义韵特征

语义韵体现出搭配词积极、消极或是中性的语义色彩，具有某一语义色彩的搭配词频数与总频数之比为该节点词的语义韵力度（李晓红、卫乃兴 2012）。如图4所示，节点词entirely在COCA总库中吸引具有消极语义色彩搭配词的倾向性最高，例数为4,231例，占总例数的53.05%，其高频搭配词为different、separate、eliminate、absent等；具有中性语义色彩的搭配词次之，例数为2,721例，占比34.12%，如compose、depend、focus等；具有积极语义色彩的搭配词较少，例数为1,023例，占比12.83%，如appropriate、comfortable、devote等。

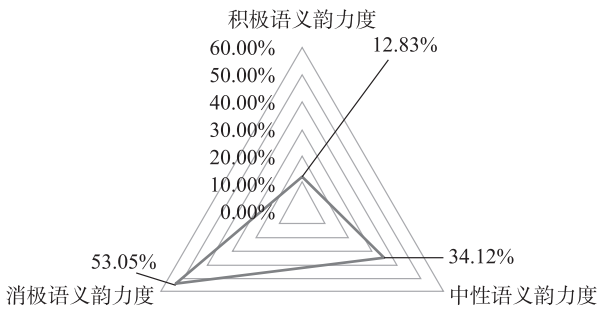


图4 COCA总库中节点词entirely的语义韵特征

四种语域子库中节点词entirely的语义韵特征不尽相同（见图5）。与COCA总库的检索数据相比，在COCA小说子库中，消极语义韵力度同样呈现主导趋势，例数861例，占比54.32%，然而积极语义韵力度跃升至24.42%，例数387例，甚至超过了中性语义韵力度，例数337例，占比21.26%。值得一提的是，COCA小说子库中节点词entirely的积极语义韵力度是众多子库中最高的，而且是唯一超过中性语义韵力度的。此外，节点词积极语义韵力度较高的还有COCA口语子库（22.18%），其消极语义韵力度和中性语义韵力度则均低于COCA总库，分别为45.56%与32.25%。在COCA学术子库中，节点词entirely的积极语义韵力度（10.88%）与COCA总库中的数据（12.83%）较为接近，然而其消极语义韵力度与中性语义韵力度差别不大，分别为48.95%和40.17%。三种语义韵力度最接近COCA总库的为COCA报刊子库，其积极、消极、中性语义韵力度分别为11.76%、51.11%和37.13%。

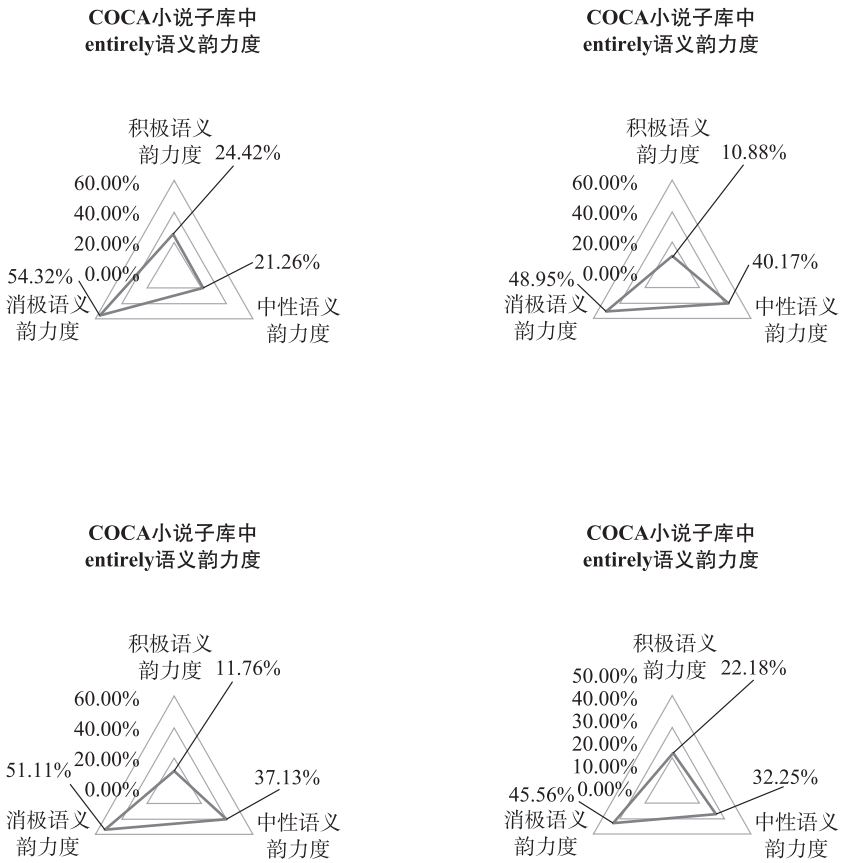


图5　四种语域子库中节点词entirely的语义韵特征

节点词entirely的语义韵力度在小样本、中样本、大样本三个语料库中亦存在较大差异（见图6）。在小样本COCA语料库中，中性语义韵力度（74.49%）占据了绝对的统治性地位，而消极语义韵力度仅以19.39%位居次席，积极语义韵力度则为6.12%。中样本COCA语料库中的主导语义韵仍为中性语义韵，但其语义韵力度为47.38%，不复在小样本之中的绝对优势；与COCA总库的数据相比，积极语义韵力度恢复到11.28%的正常水平，而消极语义韵力度（41.14%）仍然处于较低的水平。在大样本COCA语料库中，消极语义韵力度（47.12%）恢复了COCA总库中的主导地位，中性语义韵（38.24%）与积极语义韵（14.49）亦较为接近COCA总库中的对应数据。

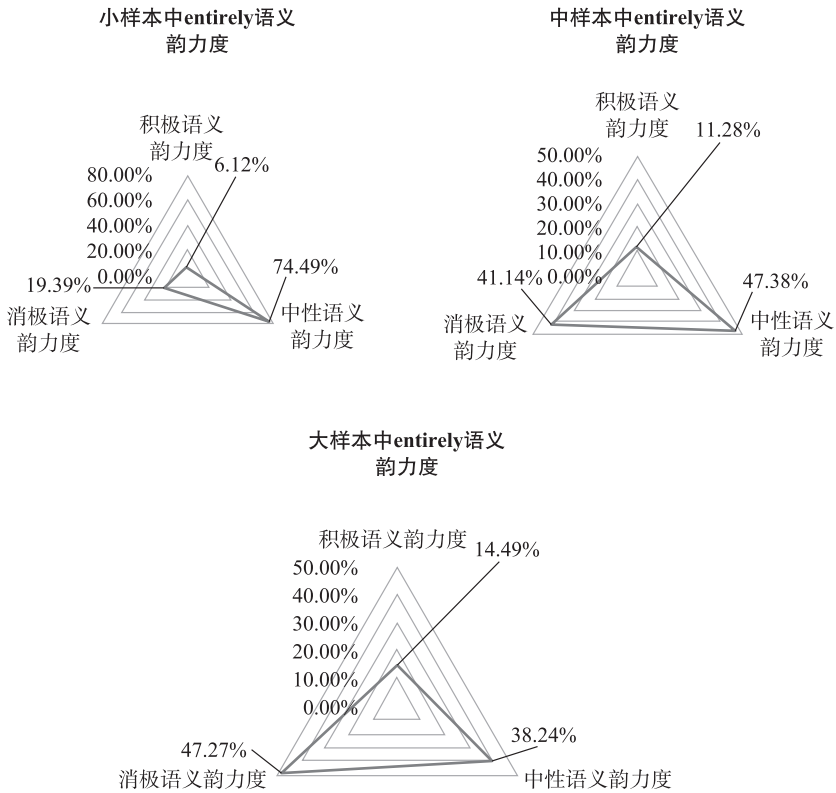


图6　三种不同语料规模语料库中节点词entirely的语义韵特征

4.4　语义趋向特征

作为对共现搭配词语义特征的概括，语义趋向分析需要对节点词的高频显著搭配词进行语义归纳，进而总结其语义特征。本研究分别统计了节点词entirely在各个语料库中的高频显著搭配词，并将其语义特征整理如下（见表2、表3、表4）。

表2　节点词entirely在COCA总库中的语义特征

语料库	语义特征	高频搭配词
COCA总库	异同（2,366例）	different，consistent，compatible，consonant
	消除（860例）	disappear，forget，eliminate，vanish，absent，dismiss
	准确性（633例）	clear，accurate，convinced，plausible
	聚散（583例）	compose，separate，consist，construct，exclude，comprise

（待续）

（续表）

语料库	语义特征	高频搭配词
COCA总库	依靠（534例）	depend，rely，dependent，reliant
	可能性（505例）	possible，conceivable

表3　节点词entirely在不同语域中的语义特征

语料库	语义特征	高频搭配词
COCA小说子库	异同（437例）	different，unlike，consistent
	消除（208例）	forget，vanish，disappear，erase，devoid
	聚散（100例）	compose，separate，consist，comprise
	可能性（73例）	possible
	正确性（76例）	true
COCA学术子库	异同（521例）	different，consistent，compatible，incompatible
	消除（300例）	eliminate，ignore，disappear，devoid，vanish
	依靠（187例）	dependent，depend，rely，reliant
	准确性（160例）	clear，accurate，convinced
	聚散（147例）	compose，consist，comprise
COCA报刊子库	异同（875例）	different，compatible
	消除（232例）	eliminate，disappear，devoid，vanish，banish
	准确性（192例）	clear，accurate，convinced
	聚散（216例）	consist，compose，separate，comprise，exclude
	依靠（190例）	depend，dependent，reliant，rely
	可能性（196例）	possible，conceivable
COCA口语子库	异同（537例）	different，agree，disagree，consistent
	准确性（157例）	clear，accurate，circumstantial，convinced
	可能性（133例）	possible，conceivable

表4　节点词entirely在不同语料规模中的语义特征

语料库	语义特征	高频搭配词
COCA小样本库	异同（7例）	different
	震惊（5例）	surprise，shock
	改变（5例）	change，switch
COCA中样本库	异同（39例）	different
	可能性（10例）	possible
	聚散（7例）	focus
COCA大样本库	异同（81例）	different，compatible，agree，disagree
	消除（26例）	eliminate，disappear，ignore，vanish，destroy
	准确性（32例）	clear，accurate，convinced，plausible
	聚散（23例）	consist，compose，separate
	依靠（21例）	depend，dependent，reliant，rely
	可能性（19例）	possible

显而易见，所有语料库中与节点词entirely共现频率最高的搭配词当属different，而且例数极多，因此可以将节点词entirely的主导语义趋向概括为“异同”，这也与余渭深、李中正（2017）的发现相似。同时，与节点词entirely共现的其他高频搭配词也体现出不同的语义特征。在COCA总库当中，与different（不同的）一同表达“异同”语义的还有consistent（一致的）、consonant（一致的）、compatible（兼容的）等词。除此之外，节点词entirely的高频搭配词还构建出五类语义特征：高频搭配词consist（组成）、compose（组成）、separate（分离）、construct（组成，构建）等表现出整体中特定部分“组成”或“分散”的概念，其语义特征可勾勒为“聚散”；高频搭配词disappear（消失）、vanish（消失）、eliminate（消灭）等体现出主体的消失或毁灭，而forget（忘记）指记忆的消除，dismiss（解雇）指职务的消除，因此可将这一类高频搭配词的语义特征概括为“消除”；高频搭配词clear（清楚的）、accurate（准确的）等明显体现出“准确性”的语义特征；possible（可能的）和conceivable（可能的）表示对“可能性”的判断；高频共现动词depend（依靠）、rely（依赖）与其各自的形容词dependent（依靠的）、reliant（可靠的）则体现出“依靠”。在六种语义特征当中，语义特征“异同”例数为2,366例，而其余五种语义特征例数则在500例左右。可见，在COCA总库当中，节点词entirely的语义趋向呈现以“异同”（2,366

例）为主导，“消除”（860例）、“准确性”（633例）、“聚散”（583例）、“依靠”（534例）、“可能性”（505例）五种语义特征并行存在的特点。

与在COCA总库中相比较，节点词entirely的语义趋向在不同语域中体现出了不同的特点。在COCA报刊子库中，节点词entirely的语义趋向几乎与COCA总库中的完全相同，均体现为以“异同”为主，其他五种语义特征为辅的特点。就每种语义特征的例数而言，主导语义特征“异同”的例数为875例，而其余五种语义特征例数均为200例上下，是主导语义特征例数的四分之一左右，这一数据也与COCA总库相吻合。然而，在其他三种语域中，除主导语义特征仍为“异同”，其余语义特征均有较大差异。例如，与在COCA总库中相比，节点词entirely在COCA小说子库中的语义趋向并未包含“依靠”与“准确性”，取而代之的是“正确性”，其对应高频搭配词为true（正确的）。值得注意的是，true这一高频搭配词仅仅出现在了COCA小说子库当中，而在其他语域中均未见与“正确性判断”相关的搭配词，故而“正确性”为节点词entirely在COCA小说子库中独有的语义特征。在COCA学术子库中，节点词entirely并未体现出“可能性”语义特征，其他五种语义特征则均有出现。就例数而言，表达“消除”语义特征的高频搭配词例数（300例）明显高于“依靠”（187例）、“准确性”（160例）与“聚散”（147例）三类高频搭配词例数。节点词entirely在COCA口语子库中的语义趋向体现出最为显著的差异，其语义特征仅仅体现在“异同”（537例）、“准确性”（157例）和“可能性”（133例）三种类型中，与COCA总库索引行相比，缺少了“消除”“聚散”和“依靠”三种语义特征。

节点词entirely的语义趋向在COCA大样本库中与在COCA总库中十分相近，均为以“异同”为主导语义特征，其他五种语义特征并行存在，仅在搭配词类符上存在细微差异。然而，随着索引行的减少，在COCA中样本语料库中，节点词entirely的语义趋向仅仅体现为“异同”（39例）、“可能性”（10例）与“聚散”（7例）。在COCA小样本语料库中，“异同”（7例）这一语义特征甚至被剥夺了主导地位，和“震惊”（5例）、“改变”（5例）一同构成节点词entirely的语义趋向。由此可见，与在COCA总库中相比较，节点词entirely的语义趋向在不同语料规模中的差异主要体现在小样本与中样本之中，而在大样本语料库中并未体现出明显差异。

5 讨论

在传统的语义韵研究当中，语域的区分以及语料规模的控制往往被研究者们所忽视。然而，研究发现：节点词的语义韵会因语域以及语料规模的不同而变化，其所在的整个扩展意义单元在搭配词、类联接、语义韵、语义趋向上均存在显著差异。

在搭配词方面，节点词entirely在不同语域语料库和不同规模语料库中均有形式各异的高频搭配词，与COCA总库中搭配词的相似性也较为有限。就类联接而言，节点词entirely搭配词的类联接模式在不同语域中体现出了较为显著的差异，然而，在不同语料规模中却并未体现出差异。在不同语域中的语义韵特征方面，节点词entirely在COCA学术子库、COCA报刊子库中的语义韵特征与COCA总库中较为相似，而在COCA小说子库中积极语义韵力度过高，在COCA口语子库中中性语义韵力度过高。在小、中、大三类样本中，虽然体现出随着样本容量增加，语义韵特征逐渐向着COCA总库中数据靠拢的趋势，但即便在COCA大样本语料库中，与COCA总库数据相比，其积极语义韵力度与中性语义韵力度仍明显偏高，整体语义韵特征存在较大差距。在语义趋向方面，COCA总库中节点词entirely的语义趋向由“异同”主导，“消除”“准确性”“聚散”“依靠”“可能性”五种语义特征并行存在。然而，在不同语域当中，部分语义特征却出现了缺失，例如，节点词entirely在COCA小说子库中缺失了“依靠”与“准确性”语义特征，在COCA学术子库中缺少了“可能性”语义特征，在COCA口语子库中缺少了“消除”“聚散”和“依靠”三种语义特征。同时，就语料规模而言，节点词entirely在大样本中的语义趋向与COCA总库中完全相同，而在中、小样本中均出现了严重的语义特征缺失。

从数据的对比中可以发现，节点词entirely的语义韵确实会受到语域和语料规模两个因素的影响。若研究者在语义韵研究，尤其是语义韵对比研究中，未能进行明确的语域区分，那么通过语料库检索得到的数据很可能会掺杂“水分”。试想，为探索某文学作品中特定节点词的语义韵，某研究者将文学作品中该节点词的语义韵与COCA总库中的数据相对比，进而得出结论：该文学作品中某节点词出现了严重的语义韵冲突。这种结论具有足够的说服力吗？显然答案是否定的。文学作品中的语义韵研究应在相应的语域中开展，其对比语料库应为专用的文学语料库而非通用语料库，这样才能避免语域对研究结论的干扰。此外，本研究还发现，节点词entirely的语义韵在小样本（100条索引行）和中样本（500条索引行）中与在COCA总库中的语义韵差异极大，即便是在大样本（1,000条索引行）中，也未能完全描绘出总库中的语义韵全貌。可见，采用随机抽取索引行或隔行抽取索引行的方法进行语义韵研究，很可能会导致研究仅勾勒出了节点词语义韵的某个侧面，进而影响后续结论与探讨。

6 结语

本文在探讨语域与语料规模对语义韵研究影响的基础上，以节点词entirely为例，进行了基于语料库的语义韵研究，从搭配词、类联接、语义韵特征、语义趋向等方面对比了不同语域和不同样本规模中节点词entirely的语义韵差异。结果表

明，节点词entirely的语义韵在不同的语域中差异显著，并且在样本规模有限的前提下无法完全呈现语义韵的全貌。可见，语域和语料规模会严重影响语义韵的形成，是语义韵研究中不可忽视的重要因素。因此，未来的语义韵研究应在区分具体语域的前提下开展，并尽量在大样本中进行检索分析。

参考文献

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

FIRTH J R. Papers in linguistics 1934-1951 [M]. London: Oxford University Press, 1957.

HUNSTON S. Semantic prosody revisited [J]. International Journal of Corpus Linguistics, 2007, 12(2): 249-268.

LOUW B. Irony in the text or insincerity in the writer? The diagnostic potential of semantic prosodies [C]//BAKER M, FRANCIS G, TOGNINI-BONELLI E. Text and technology: in honour of Sinclair. Amsterdam: John Benjamins, 1993: 157-176.

PARTINGTON A. “Utterly content in each other’s company”: semantic prosody and semantic preference [J]. International Journal of Corpus Linguistics, 2004, 9(1): 131-156.

SINCLAIR J. Looking up [M]. London/Glasgow: Collins, 1987.

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

SINCLAIR J. Trust the text: language, corpus and discourse [M]. London: Routledge, 2004.

STUBBS M. Text and corpus analysis: computer-assisted studies of language and institutions [M]. Oxford: Blackwell, 1996.

STUBBS M. Words and phrases: corpus studies of lexical semantics [M]. New York: Blackwell, 2001.

XIAO R, MCENERY T. Collocation, semantic prosody and near synonymy: a cross-linguistic perspective [J]. Applied Linguistics, 2006(1): 103-129.

戴建春. 国内语义韵研究回顾：趋势、问题与建议[J]. 外语电化教学，2018（3）：20-25.

高歌，卫乃兴. 汉英翻译界面下的语义韵探究——来自《红楼梦》英译本的证据[J]. 解放军外国语学院学报，2019（1）：48-56.

李文中. 搭配的界定、测量与中国学习者语料库搭配分析[J]. 外语教学，2017（2）：70-74.

李晓红，卫乃兴. 汉英对应词语单位的语义趋向及语义韵对比研究[J]. 外语教学与

研究，2012（1）：20-33.

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

汪腊萍. 词项搭配的定量分析方法[J]. 上海师范大学学报（哲学社会科学版），2006（6）：117-122.

王均松，田建国. 基于扩展意义单位模型的量词语义韵研究[J]. 外语教学，2016（4）：39-43.

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

卫乃兴. 语料库语言学的弗斯学说基础[J]. 外国语，2008（2）：23-32.

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

伍晓飞. 基于语料库的“Ghost”语义韵对比研究——以小说《歌剧魅影》为例[J]. 周口师范学院学报，2019（4）：75-78.

杨晓琳，程乐. 英汉翻译不同语域下被动标记形式及语义韵变化中的“Translationese”[J]. 中国翻译，2016（6）：5-12.

余渭深，李中正. 基于语料库的中英文最高程度副词的扩展意义研究[J]. 外语教学，2017（5）：32-37.

赵朝永. 基于汉英平行语料库的翻译语义韵研究——以《红楼梦》“忙XX”结构的英译为例[J]. 外语教学理论与实践，2014（4）：75-82.

通信地址：430205　湖北省武汉市　中南财经政法大学外国语学院

数据驱动学习对于中国学生外语学习成效影响的元分析*

浙江大学 杨玲玲

提要：本研究对2009—2019年28项国内有关数据驱动学习对外语学习成效影响的实证研究进行了元分析。研究共提取了39个样本，涉及2,365名学生。研究显示：（1）数据驱动学习对于中国学生外语学习成效的总效应值为中等效应值，具有中等程度的正面促进作用，低于国际上的相关研究；（2）学段、语言技能、实验人数和实验时间都对中国学生使用数据驱动学习的有效性具有影响；（3）直接或间接使用数据驱动学习对其有效性没有影响。本研究对于进一步优化数据驱动学习在我国外语教学实践中的应用具有一定借鉴意义。

关键词：数据驱动学习、语料库、外语学习、学习成效、元分析

1 引言

数据驱动学习（Data Drive Learning，简称DDL）的概念是在20世纪末由Tim Johns（1991）提出的，主要是通过语料库索引（Concordance）来引导学生对所学语言进行观察和探索的教学方法。DDL有助于提高学习者语言的准确性（Granger 1998），契合以学生为主的建构主义教学理念（陈坚林、史光孝 2009），并且符合外语教学信息化潮流的学习方式。DDL概念引入国内后，国内学者做了许多实证性研究，涵盖了词汇、语法、写作、翻译等多个领域（俞燕明 2009；张德凤 2013；王雯秋 2013；梁兵 2014；罗凌等 2014）。这些研究肯定了DDL学习方式对于提高学生外语学习成效的正面促进作用，但其成效不一。有些研究表明，与传统教学方法相比，DDL对于外语学习成效显著（俞燕明 2009），有些研究则显示只有轻微的提高（梁兵 2014）。因为这些实证研究间的学习者特征以及研究设计差异较大，无法直接对其结果进行比较和整合。因此，本文通过元分析对已有的实证研究结果进行定量分析，探究DDL对于中国学生外语学习成效的总效应值大小，了解对于DDL的有效性具有促进作用的影响因素，以期进一步优化DDL在我国外语教学实践中的应用。

* 本研究为中央高校基本科研业务费专项资金资助项目“数据驱动学习对中国学生外语学习成效影响研究”（S20220077）成果。

2 文献综述

2.1 DDL对于外语学习成效的影响

21世纪以来，国内外出现了许多有关DDL对于外语学习者学习成效影响的实证研究，这些研究设计多样，涉及的DDL使用方式、外语技能、实验人数和学段均不相同（Boulton 2017）。通过传统的文献综述方式无法对其进行定量比较，也无法从整体上了解DDL对于外语学习成效的影响，更难以进一步分析和确定造成这些实证研究差异的因素。因此，一些学者（Mizumoto & Chujo 2015；Boulton & Cobb 2017）尝试使用元分析对已有研究结果进行整合，统计DDL对于外语学习成效影响的总效应值，分析造成这些实证研究结果差异的可能因素。

Mizumoto & Chujo（2015）和Boulton & Cobb（2017）认为，从整体上看，DDL对于外语学习成效具有显著的正面影响。但是Mizumoto & Chujo（2015）的研究只讨论了DDL对于日本学习者外语学习成效的影响，并且只研究了学习者在使用DDL前后学习成效的差异，没有研究DDL与传统教学的差异。Boulton & Cobb（2017）的研究较为全面，搜集了全世界范围内的相关文献，研究了与传统教学方法相比，DDL教学方式的促进效果，但其纳入的中国样本多是港台地区，缺失了对中国其他地区的样本研究。因港台地区的教育体制和大陆（内地）差异较大，DDL对于中国学习者外语学习成效的总效应值依然不明，需要通过元分析对现有的研究结果进行整合，并与国际上的相关研究进行对比，以进一步总结和了解DDL对于中国学生外语学习成效的影响，并为未来的研究和实践提供借鉴和启发。

2.2 潜在的影响变量

DDL的有效性受到多种变量影响，如学习者特征、实验设计、DDL的使用方法、语言技能、学段、实验人数和实验时间等。根据使用方法，DDL可以分为直接DDL和间接DDL。直接DDL指的是学生使用计算机进入语料库，并利用相关的语料分析工具来进行语料搜索、观察和分析。在这个过程中，教师不对语料进行任何筛选。而间接DDL指的是教师事先对语料进行筛选，并将选定的语料在课堂上供学生使用。Johhs本人推崇直接DDL，认为应该让学习者直接面对语料库里的数据，使学习者成为语言的研究者（Johns 2002）。但也有不少研究者认为，在进行直接DDL学习时，语料库操作繁琐（王均松 2010），语料搜索时产生的结果过于庞杂或超出学生语言能力，可能使学生产生抵触情绪（陈怡 2011）。使用间接DDL可以为学生提供适合其水平的语言材料，降低操作难度，同时也保留了DDL发现式学习的特点（张北镇、周江林 2012）。这两种使用方法对于外语学习成效的影响是否有差异，一直缺乏相关的实证研究。

早期DDL教学方式常见于词汇学习。随着语料库技术的发展，这种方法也开始在其他语言技能教学中使用，如写作、翻译、听说等（俞燕明 2009；张德凤 2013；王雯秋 2013；梁兵 2014；罗凌等 2014）。目前国内已有许多关于这些语言技能的DDL实证研究，但是缺乏对这些实证研究结果进行对比分析，因此也无法精确了解语言技能DDL对于不同语言技能的学习促进效应是否存在差异，以及造成这些差异的可能原因。

受年龄因素的制约，不同学段（如小学、中学、大学）学习者的认知能力和语言水平不同，会对外语学习成效造成影响。国内外有关DDL的实证研究大多集中在大学，有关中小学DDL实践的实证研究较少（Boulton 2019；张晓燕 2015）。这些实证研究都缺乏对学段影响的比较分析。

此外，Boulton & Cobb（2017）的研究显示，实验人数和实验时间会对DDL的有效性造成影响。因此，本研究将实验人数和实验时间纳入潜在影响变量，并对其进行分析，以了解这两个因素是否会对中国学习者的DDL学习有效性造成影响。

本文尝试使用元分析来对国内2009—2019年28项实证研究结果进行量化整合，共39个样本，涉及2,365名学生，主要回答以下两个问题：

（1）和传统教学方法相比，DDL对于中国学习者外语学习成效的总效应值有何影响?

（2）影响DDL有效性的调节变量有哪些？哪些变量能够提高DDL的使用效果?

3 研究方法

3.1 文献筛选

首先，本研究使用中国知网中国学术期刊数据库、维普数据库和万方数据知识服务平台作为文献来源，通过精确搜索关键词“数据驱动学习”得到568篇文献。检索设置的时间跨度为2009—2019年，检索时间为2020年2月15日。对文献进行去重后得到147篇，初筛之后，去除理论研究和文献综述类文献，共获得104篇文献。

其次，结合本文的研究目的，在仔细阅读分析后对文献进行进一步筛选。本研究的文献纳入标准如下：第一，研究考察的是对中国学生外语学习成效的影响；第二，研究方法为定量研究，研究需要同时有实验组和对照组；第三，DDL是主要研究变量；第四，研究报告了实验组和对照组的学习成效信息；第五，研究提供了足够的数据信息。基于以上筛选标准，最终纳入分析的文献数量为28篇，因为部分研究测量了两个以上的语言技能效应值，最终提取了39个样本，涉及2,365人。

3.2 文献编码

参考Boulton & Cobb（2017）的编码系统，根据本文的研究问题，笔者首先制定了本研究的编码表（见表1）。编码信息包括：作者、发表时间、学段、使用方式、语言技能、实验人数和实验时间。编码工作由笔者和一名经过训练的编码员分别进行。其次，笔者对编码结果进行Cohen Kappa一致性系数计算，得出系数为0.90，表明编码结果可信。最后，两人对编码不一致的地方进行协商，选取最佳的编码结果，形成最终的编码结果。

表1 编码表

编码特征	描述
作者	作者姓名,如梁兵
发表时间	发表年份,如2014
学段	中学/大学
使用方式	直接DDL/间接DDL
语言技能	词汇/写作/翻译/听说/语法
实验人数	50人以下/50—100人/100—150人
实验时间	少于1个月/1—3个月/3—6个月

3.3 数据分析

本研究采用元分析软件Comprehensive Meta Analysis 3.0进行数据分析，并使用Cohen's d值作为效应值指标。在计算效应值时，使用的数据主要为：实验组和对照组的样本量、学习成效的平均值和标准差。为保证结论准确，在进行效应值计算前，先对纳入样本作了发表偏倚和异质性检验。

本研究采用失安全系数（Nfs Classic fail-safe N）对纳入样本进行发表偏倚检验。失安全系数由Rosenthal（1979）提出，衡量标准是5n+10，其中n指纳入分析的文献数量。如果得出的值远大于5n+10，表明未发表研究结果的效应值对已发表的总体效应值影响不大，即不存在发表偏倚。本研究的失安全系数为3,859，远大于5n+10，说明本研究所纳入的文献不存在明显的发表偏倚，元分析的结果具有可靠性。

此外，本文采用I-squared作为指标进行异质性检验。Higgins *et al.*（2003）将I-squared值25%，50%，75%作为异质性低、中、高的标准。本研究的异质性检验结果为Q=175.812（$p<0.05$），I-squared=78.486（I-squared>75%），表明文献样本

间存在着明显异质性。当异质性显著时，采用随机效应模型计算总效应值；当异质性不显著时，使用固定效应模型（Rücker *et al.* 2008）。文献样本间的异质性还表明，DDL和外语学习成效间存在着潜在的调节变量，因此需要对调节变量进行效应值分析。

4 研究结果

4.1 DDL对于外语学习成效的总体影响

本研究纳入样本中效应值最低为–0.183，最高为2.175。由于文献样本间存在着明显异质性，本研究使用随机效应模型统计总效应值（见表2）。

表2 DDL对外语学习成效影响的总效应值

模型	效应值和95%的置信区间						异质性			
	样本量	效应值	标准误	方差	下限	上限	Q-value	df（Q）	P	I-squared
固定	39	0.723	0.037	0.001	0.651	0.795	175.812	38	0.000	78.386
随机	39	0.747	0.080	0.006	0.591	0.904				

表2显示，DDL对于外语学习成效的总效应值为0.747，且P<0.05。根据Cohen（1988）的效应值大小评定标准，小于0.2为小效应值，0.2和0.8之间为中效应值，大于0.8为大效应值，本研究所得的效应值0.766为中效应值。这说明与传统教学方法相比，DDL整体上对于中国学习者外语学习成效有中等程度的正向促进作用，其效应小于Boulton & Cobb（2017）的研究结果（总效应值为0.95）。

4.2 潜在调节变量对DDL外语学习成效的影响

为了确认影响DDL有效性的因素，本研究对潜在调节变量的效应值进行了统计分析（见表3）。

表3　调节变量效应值和异质性检验结果

调节变量		样本量	效应值	标准误	95%置信区间		组间异质性Q值	P
					下限	上限		
学段	中学	9	0.562	0.068	0.428	0.696	7.850	0.005
	大学	30	0.789	0.044	0.704	0.874		
使用方式	直接DDL	15	0.665	0.059	0.549	0.782	1.545	0.214
	间接DDL	24	0.759	0.047	0.668	0.851		
语言技能	词汇	21	0.749	0.047	0.656	0.842	17.837	0.003
	写作	10	0.639	0.076	0.490	0.788		
	翻译	3	1.319	0.183	0.961	1.677		
	听说	3	0.514	0.150	0.220	0.808		
	语法	2	0.472	0.163	0.152	0.793		
实验人数	小于50	5	0.463	0.133	0.203	0.723	7.868	0.020
	50—100	26	0.794	0.046	0.704	0.884		
	大于100	8	0.635	0.069	0.501	0.770		
实验时间	1月内	10	0.624	0.070	0.487	0.761	12.204	0.002
	1—3个月	7	0.513	0.097	0.323	0.703		
	3—6个月	19	0.844	0.052	0.743	0.946		

学段的影响：中学的效应值为0.562，大学的效应值为0.789，且两者的异质性显著（Q=7.850，P<0.05），说明不同学段对DDL的有效性有显著影响。

使用方式的影响：直接DDL和间接DDL的效应值分别为0.665和0.759，两者的异质性不显著（Q=1.545，P=0.214>0.05）。这意味着使用方式对于DDL有效性没有显著影响。

语言技能的影响：词汇、写作、翻译、听说和语法的效应值分别为：0.749、0.639、1.319、0.514和0.472。这五种语言技能的效应值异质性显著（Q=17.837，p=0.003）。这说明语言技能对于DDL有效性具有显著影响。其中，利用DDL进行翻译学习的促进效果最大，而语法学习的促进效果最小。

实验人数的影响：实验人数小于50人、介于50—100人以及大于100人的效应值分别为0.463、0.794和0.635。三者的异质性显著（Q=7.868，P=0.020），这表明实验人数对于DDL有效性的影响显著。且实验人数和效应值之间并非线性关

系。其中，人数介于50—100时，DDL的效应值最高。

实验时间的影响：实验时间为1个月、1—3个月以及3—6个月的效应值分别为0.624，0.513和0.844，三者间的异质性显著（Q=12.204，p=0.002），这表明实验时间对于DDL有效性的影响显著。实验时间为1个月和1—3个月均为中等效应值；实验时间为3—6个月时，效应值最高且为大效应值，这说明较长的实验时间能够取得更好的学习成效。

5 讨论

本研究运用元分析方法探究了DDL对于中国学生外语学习成效的总效应值及其影响因素。结果显示，与传统教学方法相比，DDL对于中国学生外语学习成效总效应值为0.766，即有中等程度的促进作用。这说明DDL方法从整体上来说，对于中国学生的外语学习有较好的促进作用，只要使用得当，可以有效提高学生的外语学习成绩。但是，该效应值小于Boulton & Cobb（2017）研究得出的大效应值。这说明，DDL对于中国学生外语学习成效的促进作用低于国际上的整体水平。未来有必要做进一步研究以了解造成这种差异的原因。

通过对调节变量效应值的分析，本研究发现：使用方式对于其有效性没有显著影响。无论是使用纸质材料还是使用电脑进行DDL的外语学习，都能够对外语学习成效起到中等程度的正向促进作用。因此，教师可以根据具体的教学需求和环境来选择使用方式。即使是在技术条件不足的偏远地区，也可以使用间接DDL来帮助学生提高外语学习成效。

学段、外语技能、实验人数和实验时间都会对DDL的有效性产生影响。学段对DDL产生明显影响的主要原因是，DDL是一种以学生探索为主的学习方法，强调“自我管理”“自我监督”和“自我评估”（王雯秋 2013），同时要求学生具有一定的计算机和语料库使用素养，因此更适合自主学习能力强且计算机使用熟练的大学生。Johns（1986）本人也指出，DDL适合学习动机强和学习能力高的成年人。在国内，中学课程任务重，DDL的语料分析方式较为耗时，对于中学英语教育的挑战较大。

外语技能对DDL的有效性影响显著，表明DDL对不同外语技能学习的促进效果不同。其中，翻译学习的效应值最高且为大效应值，这和Bouton & Cobb（2017）的研究结果相近，说明DDL对于翻译学习具有显著的促进效果。DDL的使用对于学生提高翻译水平有明显优势，一方面学生可以利用语料库了解翻译涉及的专业内容，了解专业术语；另一方面学生可以观察专业译者所采用的翻译策略，提高自己的翻译技巧（李德超、王克非 2011；巩雪先 2016）。但是目前国内外的相关实证研究仍然较少，需要进一步研究。词汇的效应值位居第二，为中等

效应值，与Boulton & Cobb（2017）的研究结果较为接近。DDL学习方法能够提供给学生词频、搭配、语义韵和语体特征等信息，这是国内传统教学方法所缺乏的，因此对于外语词汇学习有突出优势（吴江、江兰 2012；倪修璟 2016）。写作的效应值位居第三，高于Boulton & Cobb（2017）研究显示的小效应值。这主要是因为真实的语料能够更加系统地展示语言衔接和篇章技巧，有助于中国学生建立系统的语篇和句子结构意识，提高写作技巧（梁兵 2014）。听说和语法也具有中等效应值，但是和其他语言技能相比，促进效果偏低。这主要是因为国内缺乏适合听说的多媒体语料库资源，且多媒体语料库对教室硬件设备要求较高。而对于语法学习来说，语法规则较为隐晦，不易归纳，且传统的KWIC居中显示模式不利于凸显语法结构（罗凌等 2014）。

实验人数对于DDL的有效性也有显著的调节效应。但与Boulton & Cobb（2017）的研究结果不同的是，本研究显示实验人数和效应值之间并非线性对应关系。当实验人数介于50—100时，效应值最高。实验人数和班级规模有紧密的关系。与国外的外语班级规模相比，国内的外语班级人数较多，这会带来课堂教学以教师讲解为主、教学活动单一、学生的学习主动性差等问题（文健 2003；李洁莉 2007；王健 2015）。在课堂中引入DDL教学方式可以在一定程度上改善这些问题，增加学习活动的多样性，使学生获得“自主学习”的乐趣并促进师生交流（李德超、王克非 2011），与传统教学方法相比，可以获得更好的学习效果。

实验时间对DDL的有效性影响显著。实验时间越长，DDL的有效性越显著。这主要是因为语料库的操作以及语料分析的学习方式对于多数学生来说较为复杂（刘萍等 2016）。此外，DDL适合语言水平较高且有一定研究能力的学习者使用，随着学习时间的增加，学生的语言水平逐步提高，能够更好地使用这种方式来提高学习效果。

6 结论

本文利用元分析对2009—2019年国内28项DDL实证研究做了定量的整合分析，研究结果如下。（1）DDL对中国学生外语学习成效具有中等程度的正向促进作用。与传统的教学方法相比，DDL可以更好地促进中国学生的外语学习。但是该效应值低于国际相关研究中显示的大效应值，未来需要进一步研究来了解造成这种差异的原因。（2）使用方式对中国学生的外语学习成效没有显著影响。（3）学段、语言技能、实验人数和实验时间均对DDL的有效性具有显著影响。其中，翻译学习的促进效果最好，但是相关实证研究偏少。（4）DDL对于听说和语法学习的促进效果较低，主要原因是因为缺乏合适的语料库，未来需要增加多媒体语料库以及适合语法学习的语料库的建设。

本文存在以下不足：（1）未对语料库类型以及地区分布等因素进行讨论；（2）没有深入探究造成国内外DDL促进效应差异的原因。希望未来的研究能够对这些方面进行更深入的探讨和研究。

参考文献

BOULTON A. Data-driven learning and language pedagogy [C]//THORNE S, MAY S. Language, education and technology (3rd edition). New York: Springer, 2017: 181-192.

BOULTON A. Data-driven learning for younger learners: obstacles and optimism [C]// CROSTHWAITE P. Data-driven learning for the next generation: corpora and DDL for pre-tertiary learners. London: Roultledge, 2019: 14-20.

BOULTON A, COBB T. Corpus use in language learning: a meta-analysis [J]. Language Learning, 2017, 67(2): 348-393.

COHEN J. Statistical power analysis for the behavioral sciences [M]. Hillsdale: Lawrence Erlbaum Associates, 1988.

GRANGER S. The computer learner corpus: a versatile new source of data for SLA research [C]//GRANGER S. Learner English on computer. New York: Addison Wesley Longman, 1998: 1-16.

HIGGINS J P T, THOMPSON S G, DEEKS J J, et al. Measuring inconsistency in meta-analyses [J]. British Medical Journal, 2003, 327 (7414): 557-560.

JOHNS T. Micro-concord: a language learner's research tool [J]. System, 1986, 14 (2): 151-162.

JOHNS T. Should you be persuaded: two examples of data driven learning [C]//JOHNS T, KING P. Classroom concordancing. Birmingham: University of Birmingham, 1991: 1-16.

JOHNS T. Data-driven learning: the perpetual challenge [C]//KETTEMANN B, MARKO G. Teaching and learning by doing corpus analysis. Amsterdam: Rodopi, 2002: 107-117.

MIZUMOTO A, CHUJO K. A meta-analysis of data-driven learning approach in the Japanese EFL classroom [J]. English Corpus Studies, 2015, 22: 1-18.

ROSENTHAL R. The file drawer problem and tolerance for null results [J]. Psychological Bulletin, 1979, 86 (3): 638-641.

RÜCKER G, SCHWARZER G, CARPENTER J R, et al. Undue reliance on I^2 in assessing heterogeneity may mislead [J/OL]. BMC Medical Research Methodology, 2008. https://doi.org/10.1186/1471-2288-8-79.

陈坚林，史光孝 . 对信息技术环境下外语教学模式的再思考——以DDL为例 [J]. 外语教学 , 2009（6）：54-57.

陈怡. 融合多视角的影视剧本语料库与英语口语教学——一种教师引导的DDL模式构想[J]. 西安外国语大学学报, 2011（4）：62-66.

巩雪先. 基于在线英汉平行语料库的DDL翻译教学——以英语被动句翻译教学为例[J]. 西华大学学报（哲学社会科学版）, 2016（5）：107-112.

李德超，王克非. 基于双语旅游语料库的DDL翻译教学[J]. 外语电化教学，2011（1）：20-26.

李洁莉. 大学英语大班化教学中合作学习可行性的实证研究[J]. 西安外国语大学学报，2007（3）：91-94.

梁兵. 基于DDL的功能语篇写作教学研究[J]. 长沙大学学报，2014（6）：146-148.

刘萍，吴良平，刘丽亚. CQPweb在ESP写作教学中的应用研究[J]. 外语界，2016（5）：11-19.

罗凌，温善毅，朱永红. 英语语法引导式数据驱动学习平台的构建与成效[J]. 外语电化教学，2014（5）：16-21.

倪修璟. 数据驱动学习方法对大学生英语写作水平的影响[J]. 上海理工大学学报（社会科学版），2016（3）：267-271.

王健. 班级规模对课堂教学和学习影响及其作用机制研究——以大学英语课堂为例[J]. 湖北函授大学学报，2015（19）：175-177.

王均松. 数据驱动学习模式与词汇自主学习能力培养——基于COCA语料库的一项教学实验[J]. 中国外语教育，2010（1）：24-32.

王雯秋. 基于数据驱动学习的大学英语翻译教学模式实验研究[J]. 重庆理工大学学报（社会科学），2013（8）：103-107.

文健. 论英语大班教学的利弊及交际教学法的运用[J]. 北京第二外国语学院学报，2003（6）：75-78.

吴江，江兰. 试论语料库数据驱动学习在外语词汇教学中的应用[J]. 长江师范学院学报，2012（8）：93-99.

俞燕明. 数据驱动词汇教学——基于计算机和语料库的研究性教学探索[J]. 外语电化教学，2009（2）：58-62.

张北镇，周江林. 数据驱动学习的课堂实现模式研究[J]. 外语与外语教学，2012（3）：41-45.

张德凤. 数据驱动模式在英语写作教学中的运用[J]. 大学英语（学术版），2013（2）：70-73.

张晓燕. 纸质材料数据驱动法在英语教学中的应用[J]. 内蒙古师范大学学报（教育科学版），2015（4）：114-116.

通信地址：310000　浙江省杭州市　浙江大学外国语言文化与国际交流学院

国内不同导向媒体新冠肺炎疫情报道批评话语分析*

北京外国语大学 常芳玲

提要：本文基于Fairclough批评话语分析三维框架，自建《人民日报》和《人民日报》（海外版）新型冠状病毒肺炎疫情报道语料库，对国内受众导向的《人民日报》和国外受众导向的《人民日报》（海外版）疫情相关报道就“描写”“阐释”和“解释”三个层面展开分析，旨在了解国内不同受众导向媒体疫情报道的特色所在。研究发现：两个媒体在疫情的相关报道中呈现共性与个性并存的特点，其中，共性特征主要受新闻报道的一般规律、媒体性质与报道事件影响，个性特征则多取决于两者受众的不同。本文提供了一种详细了解国内疫情报道的途径，同时也在一定程度上揭示了新闻报道背后隐藏的意识形态问题。

关键词：新冠肺炎、媒体话语、批评话语分析、语料库

1 引言

新型冠状病毒肺炎是近百年来人类所遭遇的一次影响范围最广的全球性大流行病，严重威胁了人类的生命安全与健康。面对突如其来的疫情，中国果断打响了疫情防控阻击战。在此期间，国内媒体紧跟疫情动态，为民众提供即时疫情信息，一定程度上缓解了疫情状态下人民的恐慌情绪。

批评话语分析（Critical Discourse Analysis，简称CDA）是Fairclough（1989）在批评语言学（Fowler *et al.*1979）基础上提出的一种话语分析方法。与传统话语分析、篇章语言学的研究方法不同，批评话语分析不仅涉及话语或篇章本身，而且重视话语实践过程及其社会语境分析，注重从社会制度和社会构成方面来解释话语（郭松 2011）。该方法旨在分析语言、权力和意识形态三者之间的关系，揭示语篇与社会结构、权力关系的互动方式（辛斌、高小丽 2013）。批评话语分析中的“批评”与文学批评概念中的“批评”相同，两者都具有解释性，是对语篇进行解读的过程（吕万英 2005）。借助语料库开展批评话语分析是媒体话语分析的常见方法之一。唐丽萍（2011）指出，批评话语分析与语料库语言学在以下三个方面存在对话基础：承认语言的社会属性、重视意义的累积效应、强调词汇语

* 本文为2018年度教育部人文社会科学研究青年基金项目“基于三元组可比语料库的对象类介词研究”（18YJC740077）的阶段性成果。

法的共选关系。语料库语言学量化研究方法与批评话语分析质性研究方法的结合能够从不同方面反映数据的原貌，更深层次地揭示语言背后的特点与意识形态问题。纵观以往借助语料库开展的媒体报道批评话语分析，中外媒体的对比分析及外媒对中国相关事件报道的研究占比较大，而有关国内媒体的本土化研究相对较少。基于此，本文选取国内代表性媒体对新冠肺炎的相关报道为研究对象，以《人民日报》和《人民日报》（海外版）为报道源，借助语料库与批评话语分析相关知识对国内外不同导向媒体话语的传播特点予以分析，旨在考察我国媒体话语的本土化特点。该研究分析国内不同导向媒体对新冠肺炎疫情的动态报道，有利于增进对媒体报道话语特点的理解与掌握。

2 研究设计

2.1 理论框架

Fairclough（1989）在系统功能语言学的基础上，将语篇、话语实践和社会实践相联系，提出了一个三维分析框架，尝试为批评话语分析建立一个可供参考的理论范式。具体而言，该理论范式可分为三个层次：（1）描写文本的语言形式和结构特征；（2）阐释语篇与话语实践过程的关系，如语篇的生产、传播与接受；（3）解释语篇的社会语境，分析语篇的建构过程以及背后所隐含的意识形态。其中，“描写”层属于微观层面，仅包括对文本表层形式和结构特征的分析；“阐释”层介于微观层面和宏观层面中间，既超出了单纯的文本分析范围，又不限于宏观层面的语篇探讨，该层面是“描写”层和“解释”层相互联系的纽带；“解释”层聚焦语篇与宏观层面的互动，需要结合社会结构来说明权力和意识形态在文本生成中的作用。本文选用Fairclough批评话语分析三维框架为理论指导，分析国内媒体关于新冠肺炎疫情报道的话语特征。

2.2 语料来源及研究问题

本文借助人民日报图文数据库和人民日报海外版平台，以“新冠肺炎”为关键词，检索时间截至2020年6月23日，自建《人民日报》与《人民日报》（海外版）新冠肺炎疫情报道语料库。经查重，返回《人民日报》有效报道3,623篇（共计3,021,881词），《人民日报》（海外版）1,889篇（共计1,547,994词），将两个语料库所在文件夹分别命名为RMRB_NCP和RMRBHWB_NCP。基于此，本文拟回答以下三个问题：

（1）《人民日报》与《人民日报》（海外版）的新冠肺炎疫情报道在“描写”层有何特点？是否存在差异？

（2）《人民日报》与《人民日报》（海外版）的新冠肺炎疫情报道在“阐释”层有何特点？是否存在差异？

（3）《人民日报》与《人民日报》（海外版）的新冠肺炎疫情报道在“解释”层有何特点？是否存在差异？

3 研究结果与分析

本节依据Fairclough（1989）提出的批评话语分析三维框架，分别从“描写”层、“阐释”层和“解释”层入手，分析《人民日报》和《人民日报》（海外版）中有关新型冠状病毒肺炎疫情的报道特点。

3.1 “描写”层文本分析

本节借助语料库语言学工具BFSU PowerConc（许家金等 2012），从高频词、主题词和搭配分析三个维度出发，对《人民日报》和《人民日报》（海外版）的疫情相关报道开展“描写”层面的文本分析。

3.1.1 高频词分析

高频词指某一语料库中出现次数较多的词语集合，这一指标能够反映研究文本中复现率较多的词块。词频能够帮助人们辨别最基本的语言特征，这些特征往往包含话语的意义（McEnery *et al.* 2006）。笔者将经过segtag分词处理的《人民日报》与《人民日报》（海外版）语料分别导入BFSU PowerConc，运用N-gram List选项生成词频列表。限于篇幅，本文仅考察两个语料库中出现频次位于前50位的一元词组[1]，结果如表1、表2所示。

表1 《人民日报》疫情报道高频词（前50位）[2]

序号	一元词组	词频	标准化频率[3]	序号	一元词组	词频	标准化频率	序号	一元词组	词频	标准化频率
1	的	116,045	384.02	18	要	11,501	38.06	35	到	7,726	25.57
2	和	36,415	120.50	19	有	11,050	36.57	36	个	7,691	25.45
3	疫情	36,202	119.80	20	中	10,835	35.86	37	人	7,384	24.44
4	在	33,161	109.74	21	经济	9,954	32.94	38	大	7,150	23.66
5	了	25,273	83.63	22	冠	9,843	32.57	39	医院	7,038	23.29
6	是	24,080	79.69	23	我们	9,560	31.64	40	就	6,939	22.96

（待续）

（续表）

序号	一元词组	词频	标准化频率[3]	序号	一元词组	词频	标准化频率	序号	一元词组	词频	标准化频率
7	中国	21,722	71.88	24	人民	9,364	30.99	41	将	6,876	22.75
8	控	18,406	60.91	25	上	9,336	30.89	42	人员	6,674	22.09
9	防	17,809	58.93	26	多	9,147	30.27	43	说	6,621	21.91
10	一	17,253	57.09	27	肺炎	8,870	29.35	44	更	6,404	21.19
11	新	17,029	56.35	28	社会	8,791	29.09	45	好	6,390	21.15
12	为	16,010	52.98	29	企业	8,560	28.33	46	国际	6,369	21.08
13	对	15,009	49.67	30	与	8,557	28.32	47	能	6,124	20.27
14	等	13,750	45.50	31	国家	8,390	27.76	48	疫	5,977	19.78
15	工作	13,001	43.02	32	这	8,154	26.98	49	都	5,792	19.17
16	不	12,837	42.48	33	也	8,031	26.58	50	合作	5,681	18.80
17	发展	11,501	38.06	34	名	7,744	25.63				

表2　《人民日报》（海外版）疫情报道高频词（前50位）

序号	一元词组	词频	标准化频率	序号	一元词组	词频	标准化频率	序号	一元词组	词频	标准化频率
1	的	65,258	421.56	18	中	5,773	37.29	35	将	4,047	26.14
2	在	20,476	132.27	19	冠	5,608	36.23	36	发展	4,040	26.10
3	疫情	16,580	107.11	20	也	5,460	35.27	37	疫	3,967	25.63
4	和	16,451	106.27	21	这	5,265	34.01	38	大	3,820	24.68
5	了	14,601	94.32	22	多	5,166	33.37	39	就	3,742	24.17
6	是	13,252	85.61	23	工作	4,996	32.27	40	经济	3,692	23.85
7	中国	13,070	84.43	24	肺炎	4,876	31.50	41	都	3,433	22.18
8	一	9,192	59.38	25	要	4,803	31.03	42	年	3,341	21.58
9	新	8,516	55.01	26	与	4,618	29.83	43	更	3,300	21.32
10	为	7,771	50.20	27	国家	4,593	29.67	44	人民	3,281	21.20
11	等	7,151	46.20	28	企业	4,526	29.24	45	国际	3,143	20.30

（待续）

（续表）

序号	一元词组	词频	标准化频率	序号	一元词组	词频	标准化频率	序号	一元词组	词频	标准化频率
12	对	7,057	45.59	29	我们	4,474	28.90	46	人员	3,104	20.05
13	控	6,582	42.52	30	人	4,346	28.08	47	向	3,015	19.48
14	有	6,541	42.25	31	个	4,233	27.35	48	抗	2,981	19.26
15	不	6,477	41.84	32	到	4,225	27.29	49	武汉	2,949	19.05
16	防	6,309	40.76	33	我	4,190	27.07	50	还	2,931	18.93
17	上	5,848	37.78	34	说	4,052	26.18				

表1为《人民日报》有关新冠肺炎疫情报道的前50个高频词，表2是《人民日报》（海外版）有关疫情报道的前50个高频词。总体来看，表1、表2同质性较高，出现的高频词均可分为两大类：一是虚词，如“的”“在”等，这类词在任何语料库中均占比较大，这是由语言自身的特点决定的；二是与新冠肺炎报道相关的词汇，如“疫情”“中国”等。两家报纸前50个高频词中重合词有44个，重合率达88%，且多数重合词在两个语料库高频词表中排列次序相差不大，其标准化频率也趋于一致。

除上述44个重合高频词外，《人民日报》独有词包括“社会”“名”“医院”“好”“能”和“合作”，《人民日报》（海外版）独有高频词为“我”“年”“抗”“武汉”和“还”。《人民日报》中，“社会”一词出现频率偏高。一方面是由于《人民日报》用于刊登新冠肺炎报道的版面名称之一为“社会”，另一方面则是由于国内社会与国际社会在此次新冠疫情报道中的凸显；“名”常跟在数字后用来修饰疫情相关人员，如患者、医务人员、志愿者和党员等；“医院”一词在《人民日报》报道中的凸显表明了该媒介对疫情防控主战场——医院的重视；“好”在文本中常用于表达国家和政府采取一系列积极措施促使疫情向良性方向发展；“能”最为显著的搭配词是“才”，“才能”一词常出现于若干假设条件之后，辅助表达抗疫胜利所需条件；“合作”存在于各方力量之间，只有相互合作，才能更好、更快地战胜疫情。《人民日报》（海外版）将“我”作为前50高频词中的独有词之一，主要出现在个体疫情故事中，这也表明，《人民日报》（海外版）刊登的叙事语篇较多；“年”用于与疫情有关的历史时间线梳理，如“五千多年来……”；“抗”最常出现在“抗疫”一词中；《人民日报》（海外版）在报道中呈现了“武汉”作为疫区而开展的相关活动；“还”的高频使用在一定程度上说明海外版的报道内容存在信息叠加的现象。综合两个语料库中独有的高频词来看，《人民日报》多从宏观层面报道疫情，而《人民日报》（海外版）在关注宏观层面的同时也聚焦

了个体疫情故事等微观层面的报道。

3.1.2　主题词分析

主题词分析可用于描述某一语体并在语言中找出话语轨迹（Baker 2004）。与高频词不同，主题词指某一语料库与参照语料库相比，统计出的具有特殊词频的词（钱毓芳 2010）。本文将国家语委现代汉语平衡语料库汉语词频表作为参照语料库词表，分别生成《人民日报》疫情报道主题词表和《人民日报》（海外版）疫情报道主题词表。

据统计，《人民日报》疫情主题词表中具有显著意义的主题词共计2,266个。限于篇幅，本文仅列举显著性较高的前50个主题词（见表3）。

表3　《人民日报》疫情报道主题词（前50位）

序号	主题词	序号	主题词	序号	主题词	序号	主题词	序号	主题词
1	控	11	标题	21	工作	31	物资	41	记者
2	防	12	卫生	22	保障	32	服务	42	防护
3	中国	13	医疗	23	人民日报	33	落实	43	推动
4	冠	14	抗击	24	日期	34	公共	44	就业
5	肺炎	15	名	25	安全	35	全面	45	提供
6	新	16	人员	26	总书记	36	经济	46	加强
7	医院	17	合作	27	推进	37	复	47	作者
8	患者	18	全球	28	支持	38	湖北	48	病例
9	武汉	19	版	29	企业	39	做好	49	一线
10	抗	20	国际	30	人民	40	产	50	治理

由表3可知，“控”和“防”是《人民日报》报道主题性最强的两个词，结合语境发现，这两个词多源于“防控”一词。该词的广泛使用体现了该媒介对疫情防控工作的宣传力度。结合新闻体裁六要素以及主题词的所在语境，《人民日报》疫情报道主题词可分为以下五类。

（1）WHO类词汇：患者、人员、总书记、人民；

（2）HOW类词汇：控、防、抗、抗击、合作、保障、推进、支持、落实、全面、复、做好、产、防护、推动、提供、加强、治理；

（3）WHAT类词汇：冠、肺炎、新、卫生、医疗、工作、安全、企业、物资、

服务、公共、经济、就业、病例；

（4）WHERE类词汇：中国、医院、武汉、全球、国际、湖北、一线；

（5）新闻元话语与新闻特色词汇：标题、名、版、人民日报、日期、记者、作者。

本研究按照以上标准对《人民日报》非前50主题词进行校验，发现上述类别基本能够覆盖该媒介主题词表的全部词汇。在WHO类主题词中，“患者”是《人民日报》最为关注的主体，体现了报道的人本主义关怀。除“患者”一词外，凸显度较高的主题词还包括“人员”（主要指医务人员）和“总书记”，前者与患者并肩作战抗击疫情，后者则负责整个国家层面疫情大局的统筹安排。HOW类主题词多与疫情防控方式有关，以动作类词汇为主，此类主题词较为丰富，在一定程度上彰显了国内疫情防控工作的多手段性。不过，在疫情好转且趋向平稳阶段，HOW类部分主题词也涉及对复苏国内经济的提倡，建议有条件的地区可以复学、复工、复商和复产等。WHAT类词汇关注疫情和疫情状态下社会各方面的基本情况，是HOW类词汇作用的具体对象。与HOW类词汇相似，WHAT类词汇在国内疫情发展中后期也出现了一些经济类词汇，如“经济”“就业”等。WHERE类词汇指出了《人民日报》疫情相关报道中主题性较为突出的地点或范围，这些地点不仅出现了“武汉”“湖北”等直接指示国内疫区的词，同时也涉及“全球”和“国际”等国际化词汇，说明此次疫情的影响并不仅局限于中国，而是全人类共同面对的一场战役。新闻元话语与新闻特色词汇作为主题词出现，与新闻报道的语篇形式密不可分，是区别于其他体裁的标志。

《人民日报》（海外版）疫情报道主题词表中具有显著意义的主题词共计2,235个，表4为主题性较为突出的前50位。

表4 《人民日报》（海外版）疫情报道主题词（前50位）

序号	主题词	序号	主题词	序号	主题词	序号	主题词	序号	主题词
1	中国	11	物资	21	患者	31	湖北	41	网络
2	控	12	医院	22	防护	32	海外	42	服务
3	防	13	版	23	平台	33	检测	43	数据
4	冠	14	人民日报	24	记者	34	提供	44	工作
5	肺炎	15	人员	25	线	35	复	45	做好
6	新	16	国际	26	支持	36	产	46	香港
7	武汉	17	卫生	27	病例	37	战	47	华人

（待续）

（续表）

序号	主题词	序号	主题词	序号	主题词	序号	主题词	序号	主题词
8	抗	18	抗击	28	安全	38	隔离	48	措施
9	全球	19	合作	29	保障	39	相关	49	公共
10	医疗	20	企业	30	新华社	40	推进	50	表示

从类别上来看，表4中的主题词同样可借助表3的框架分类，具体如下所示：

（1）WHO类词汇：人员、患者、华人；

（2）HOW类词汇：控、防、抗击、合作、防护、支持、保障、检测、提供、复、产、战、隔离、推进、服务、做好、表示；

（3）WHAT类词汇：冠、肺炎、新、医疗、物资、卫生、企业、平台、线、病例、安全、相关、网络、服务、数据、工作、措施、公共；

（4）WHERE类词汇：中国、武汉、全球、医院、国际、湖北、海外、香港；

（5）新闻元话语与新闻特色词汇：版、人民日报、记者、新华社。

《人民日报》（海外版）关注度较高的群体包括医务人员、患者以及海外华人华侨，没有刻意凸显总书记的领导角色。海外版报道中HOW类主题词与《人民日报》同质性较高，但在疫情防控方式及举措的相关表达中，除与《人民日报》相同的一些较为笼统的方式外，海外版着重强调了“检测”和“隔离”两种具体措施。此外，海外版报道中涉及的中外国家首脑间交流较多，“表示”一词多在此类语境中出现，用于引出某一方观点。海外版WHAT类主题词在关注疫情和社会的基础上，与网络相关的主题词时有出现，如“平台”“线”和“数据”，多数语篇涉及互联网相关平台、线上教学和问诊等云服务在疫情期间的作用。在WHERE类词汇中，海外版报道除《人民日报》提及的地点和范围外，还特意提及香港地区。最后一类主题词中“新华社”的出现表明海外版的部分报道转载于新华社。同时，海外版报道中出现“人民日报”一词主要是“人民日报海外版”分词的原因。

从《人民日报》和《人民日报》（海外版）的主题词分析来看，两个媒介关于疫情的报道有同有异，且同大于异。具体来说，两家有关疫情报道的总体方向没有变化，主题词多有重复且分类基本一致。不过，它们的侧重点也存在一些不同。首先，《人民日报》中主题性最高的词为“控”和“防”，而海外版报道中“中国”一词的主题性超过“控”和“防”排名首位，这也表明人民日报社在面对国内外不同读者时，选登内容有所差异。面向国内的报道更强调疫情防控工作的现状与进展，而在面向海外读者时则更强调中国这一主体在疫情中的相关活动。其次，《人民日报》主题词虽涉及国内外两个方面，但以国内报道为主，而海外版对国外

的关注要多于《人民日报》，如对华人华侨的关心以及中外领导人就疫情的相互交流等。再者，海外版报道中与互联网相关的词汇主题性较高，这在《人民日报》前50主题词中未有体现。海外版主题词的这一特点展示了中国在疫情中充分利用了网络这一信息时代的产物，使部分工作在无接触的前提下有条不紊地开展。

3.1.3 搭配分析

将“新冠肺炎”作为检索词，分别在RMRB_NCP和RMRBHWB_NCP两个语料库中分析其搭配情况。为了更清晰地呈现检索词左右的搭配情况，本文将左搭配和右搭配分开处理。在RMRB_NCP语料库中，“新冠肺炎”一词左搭配（跨距为5）显著性最高的10个词分别是“应”“在”“的”“中国”“统筹”“中央”“这次”“受”“推进”和“收治”；右搭配（跨距为5）显著性最高的10个词为“疫情”“的”“以来”“中”“中国”“带来”“在”“是”“给”和“诊疗”。依据上述显著性较高的搭配词，同时结合上下文语境分析，发现“新冠肺炎”一词在《人民日报》报道中的几个典型搭配形式如下：

（1）（实体）在（动词性成分）新冠肺炎疫情中，……

在统筹推进新冠肺炎疫情防控和经济社会发展的斗争中，各级党组织和广大党员、干部坚决贯彻落实习近平总书记重要指示精神和党中央决策部署，自觉践行初心使命，勇于担当、攻坚克难、无私奉献，充分展现出新时代共产党人的政治本色。

（2）……应对新冠肺炎疫情……

新华社报道，李克强主持召开中央应对新冠肺炎疫情工作领导小组会议，要求抓好巩固防控成效各项工作，突出做好无症状感染者防控。

（3）新冠肺炎疫情发生以来，……

新冠肺炎疫情发生以来，中国政府本着对中国人民和世界人民生命安全和身体健康高度负责的态度，采取最全面、最严格、最彻底的防控举措，同疫情开展坚决斗争。

在第一种典型搭配型式中，“实体”可以是国家、机构、身份凸显度较高的人或群体，也可以是药物等促进疫情好转的一些因素，该成分也可以省略不写。“动词性成分”在第一个搭配型式中也不是必须的，可选择性添加或删除，高频出现

的动词性成分包括“统筹推进”“抗击”等。在“（实体）在（动词性成分）新冠肺炎疫情中”这一结构后，通常是对疫情防护工作的具体描述。总的来说，这一搭配型式显示了不同实体在此次新冠肺炎疫情中扮演的角色与作用。第二种搭配型式在“应对新冠肺炎疫情”左边出现的显著性成分为“中央”，常见右搭配为会议、峰会以及领导小组或其代表性成员的活动。该型式表明中共中央在应对疫情时常借助会议讨论相关事宜。最后一种型式以“新冠肺炎疫情发生以来”开始，其后多是对中国在国家、社会和群体不同层面积极性应对措施的解释与说明，同时也包括中国与他国来往的一些活动。

在RMRBHWB_NCP的报道语篇中，“新冠肺炎”显著性较高的前10个左搭配词（跨距为5）依次是“应”“在”“受”“的”“中央”“中国”“这次”“关于”“指出”和“第”；右搭配（跨距为5）显著性最高的十个词为“疫情”“的”“以来”“中”“在”“中国”“带来”“流行”“形势”和“治疗”。梳理发现，“新冠肺炎”一词在《人民日报》（海外版）报道中典型搭配型式与《人民日报》总体相同，但在微观表达上存在差异。《人民日报》（海外版）除了有关国内情况的一些描写，还增加了从国际视角出发的一些报道，这类报道尤其在第三个搭配型式中最为常见，如：

（4）习近平强调，新冠肺炎疫情发生以来，中法保持了高水平战略协调。

（5）此次新冠肺炎疫情发生以来，民进党当局和“独”派分子更是迫不及待地与大陆切断往来和联系，限制口罩向大陆出口……

（6）新冠肺炎疫情发生以来，国际社会向中国提供了真诚、友善的帮助。

除上述搭配型式的微观差异外，《人民日报》（海外版）在与“新冠肺炎”的显著搭配用词上也与《人民日报》存在差异，如“指出”多出现在习近平主席与他国首脑的电话交谈中；“流行”用于表达疫情全球大流行的特征；关注疫情“形势”的变化；报道新冠肺炎的“治疗”方式，其中着重强调了中药和中医的作用。上述提及的搭配词多出现在海外版报道中，一方面展示了疫情的全球性影响及我国与其他国家的交流沟通，另一方面在报道相关事实的同时也向世界展现了以“中药”“中医”等为代表的中国特色产物在疫情抗击中的重要作用，有利于建构积极的中国国家形象。

3.2 “阐释”层话语实践分析

“阐释”层关注语篇生产、传播和消费过程的话语实践，本文对话语的“阐释”主要着眼于互文性这一语篇特征。互文性也被称作引语，是一个语篇中包含

他人话语或其他语篇的片段，即一个文本中除了作者的声音，还存在他人的声音（窦卫霖、陈丹红 2009）。这些引语来自不同的信息渠道，代表不同群体的利益和意识形态，因此必然在一定程度上通过话语反映在语篇中，并形成某些关系（辛斌 2008）。引语从形式上可分为直接引语和间接引语，直接引语常与双引号共现，能够较好实现定位与查找，但间接引语隐蔽性较高，直接提取的难度较大。为兼顾语篇中直接引语和间接引语识别的准确性，本文采取质性研究方法对“阐释”层的话语实践进行分析。具体来说，笔者分别从RMRB_NCP和RMRBHWB_NCP两个语料库中随机抽取30个文本作为分析对象，通过报道中引语的转述来源和转述方式来考察其互文性。经统计，在抽取的30篇语料样本中，《人民日报》报道互文性数目共计191条，海外版共计216条，这些互文性条目的转述来源及转述方式分布如表5所示。

表5　RMRB_NCP和RMRBHWB_NCP引语转述来源及方式分布

	转述来源			转述方式		
	具体	略具体	不具体	直接	间接	混合
RMRB_NCP	158（82.7%）	11（5.8%）	22（11.5%）	60（31.4%）	98（51.3%）	33（17.3%）
RMRBHWB_NCP	152（70.4%）	42（19.4%）	22（10.2%）	87（40.3%）	107（49.5%）	22（10.2%）

依据明确程度，新闻话语的转述来源可分为具体来源、略具体来源和不具体来源。具体来源指引语清晰指向具体的某个人，略具体来源不提及姓名，而仅涉及职务等一些模糊信息，不具体来源则常使用“相关人士”等一些意义较为宽泛的用词。信息来源的具体程度影响其内容可信度，信息来源越具体，内容可信度越高。由表5可知，《人民日报》与其海外版在引文中均倾向于指明引语的具体来源以增强新闻的可信度和说服力，这一特点与两份报纸机关报的性质也是紧密相关的。不过，具体来源在《人民日报》抽取样本引语中的占比要大于《人民日报》（海外版）。就比重均较小的略具体来源和不具体来源来说，《人民日报》样本引文中不具体来源比重大于略具体来源，与海外版情况相反。通过进一步观察发现，具体来源和略具体来源话语的发出方多是此次疫情中角色较为重要或有特色贡献的个体，而不具体来源的发出方均是一些对当前文本不太重要的对象，如民众等。在《人民日报》和其海外版的报道中，具体来源和略具体来源话语发出者身份较为趋同，但在缺乏具体来源的引语中，《人民日报》中以民众居多，而海外版无具体来源的话语则主要源于外媒报道。虽然缺乏具体来源的话语可信度不高，但从另一个角度来说同一类主体话语在某媒介中的频繁引用也说明了该类主体的影响

力之大。

从转述方式上来看，两份报纸报道引文涉及直接转述、间接转述以及直接转述和间接转述混合使用三种方式。《人民日报》中间接转述比重最大，直接转述次之，混合转述情况比重最小。间接引语是对事物的间接描述或转载说明，相对于直接引语而言客观性更高，既能让读者了解事实，又没有主观感情的表露，符合新闻真实、客观的立场，因此在实际报道中，间接引语的使用频率最多。但有时为了突出报道人物的过往经历，增强读者的直观感受，同时也为了缓解间接引语频繁使用带来的枯燥感，直接引语在新闻报道中也时有出现，直接引语所扮演的这一角色在直接引语、间接引语混合使用的转述方式中体现得尤为明显。

综合来看，《人民日报》和《人民日报》（海外版）的互文性特点可以从两个方面来考察：一方面是新闻报道或机关报的共有特点，如引语中标记具体来源信息增强可信度、使用间接转述提升客观性，这也是上述两份报纸报道的共性所在；另一方面是不同报纸因其侧重点不同而表现出的一些个性特点，如《人民日报》和《人民日报》（海外版）报道中缺乏具体来源转述话语发出者的类别差异。

3.3 “解释”层社会实践分析

“解释”层是Fairclough批评话语分析三维分析框架中最为宏观的一层，同时也是与具体语言表述关联最少的一层。为了解《人民日报》和《人民日报》（海外版）疫情报道背后隐藏的意识形态问题，本节借助两份报纸疫情报道引语对其关注主体进行了比较。

从《人民日报》报道中引语所属主体来看，出现最多的首先是以习近平为核心的党中央领导集体及各级政府，这类主体在此次疫情抗击战中起着统领大局的作用，国内疫情工作能够有条不紊地开展离不开党和政府的正确领导，《人民日报》报道中此类主体的频繁出现体现了该媒介对政府声音的关注。除政府声音外，《人民日报》疫情报道中关注度最高的主体是医生群体，他们的抗疫经历与建议有利于增强民众抗击疫情的信心与决心。此外，组织机构、企业、代表委员以及相关文件、工作会议等主体凸显度也较高。上述所列实体均是国内不同群体或对象关于疫情的发声，但《人民日报》报道中不仅关注了国内的声音，同时也引用了部分国外的观点，如外国政府、企业和世卫组织等多种不同群体。总的来说，《人民日报》所引用的这些国外声音多是对中国的正面评价，高度赞赏了中国在此次疫情中发挥的作用。

《人民日报》（海外版）疫情报道中较为关注的主体与《人民日报》差异较大，党和政府的声音要明显少于《人民日报》，同时对医生群体的声音关注度也较少，出现较多的主体是组织机构和企业。在所出现的组织机构中，外交部、商务部和税务部凸显程度最高，外交部发文内容多是为因疫情滞留在国外的留学生送去关

怀，商务部和税务部则重点关注疫情恢复期中国的经济情况。在与企业相关的引文中，一部分内容关注疫情期间防控物资的供应问题，一部分涉及企业运营在疫情后的恢复问题。除组织机构和企业外，海外版报道涉及较多的国内主体还包括政府、研究人员和网友。在海外版涉及的国外主体中，“外媒”的引用次数最多，所引用的外媒报道主要涉及对中国疫情的关注。世界卫生组织由于在此次突发公共安全卫生事件中的特殊作用在海外版报道中也频繁出现。

从以上分析可知，虽然《人民日报》与《人民日报》（海外版）在关注主体的类别上差异不大，均涉及国内外政府、企业和组织机构等，但不同类别间所占比重差异较大。《人民日报》有关疫情报道的引文多源于政府和医生，而海外版则更关注组织机构、企业和外媒。《人民日报》和《人民日报》（海外版）报道关注主体的不同，在很大程度上是由于媒体性质的不同。《人民日报》是党和政府的喉舌，宣传党和政府的政策主张、凸显党和政府的核心领导地位始终是其主要任务，因此政府在疫情报道中频繁出现不足为奇。另一方面，《人民日报》面向的国内读者相比海外人士更关心中国疫情的动态发展情况，而在此次疫情防控中以钟南山院士为代表的医生群体享有很高的话语权，在报道中引用此类群体的话语能够使受众了解疫情现时状况，安抚国人情绪。《人民日报》（海外版）虽然与《人民日报》同属中国共产党中央委员会机关报，但其更是对外开放的综合性报纸，面向的国外受众对中国疫情防控的政策与过程关注度相对不高，因此海外版未凸显政府和医生群体的声音，而是结合受众特点重点宣传了组织机构、企业和部分外媒的观点。

4 结语

为了解国内不同导向新闻媒体对新冠肺炎疫情的报道特点，本文基于Fairclough批评话语分析三维分析框架，以《人民日报》和《人民日报》（海外版）报道为例，分别从“描写”层、“阐释”层和“解释”层三个层面进行了分析与比较。研究发现：无论从哪一层面上来说，《人民日报》与《人民日报》（海外版）两个媒介的疫情报道均呈共性个性并存的特点，其中，共性特征主要受新闻报道的一般规律、媒体性质与报道事件影响，而个性特征则多由两个媒体面向受众的差异所致。本研究通过批评话语分析和语料库相结合的方式，同时涉及定量和定性两种分析方法，较为全面地讨论了《人民日报》和《人民日报》（海外版）疫情报道的相关特征。该文提供了一种详细了解国内疫情报道的途径，同时也有利于揭示新闻报道中隐含的意识形态差异。

注释

1　词组指分词后左右两边均被空格隔开的单位。

2　分词程序无法准确识别一些新词或人名等特殊词类，如“新冠肺炎”一词被分为“新”“冠”和“肺炎”三个词。因此，在分析时需要对一些分词结果进行人工处理。

3　表1与表2中的标准化频率按照每万词计算，计算公式为词频/语料库总词数*10,000。

参考文献

BAKER P. Querying keywords: questions of difference, frequency and sense in keywords analysis [J]. Journal of English Linguistics, 2004, 32(4): 346-359.

FAIRCLOUGH N. Language and power [M]. London: Longman, 1989.

FOWLER R, HODGE B, KRESS G, et al. Language and control [M]. London: Routledge & Kegan Paul, 1979.

MCENERY A, XIAO R, TONO Y. Corpus-based language studies: an advanced resource book [M]. London: Routledge, 2006.

窦卫霖，陈丹红. 对中美国家领导人演讲中的互文性现象的批评性话语分析[J]. 外语与外语教学，2009（11）：12-15.

郭松. 基于语料库的批评话语分析[J]. 天津外国语大学学报，2011（5）：12-17.

吕万英. 英文新闻标题批评性分析[J]. 广东外语外贸大学学报，2005（3）：49-52.

钱毓芳. 语料库与批判话语分析[J]. 外语教学与研究，2010（3）：198-202.

唐丽萍. 语料库语言学在批评话语分析中的作为空间[J]. 外国语，2011（4）：43-49.

辛斌. 语篇研究中的互文性分析[J]. 外语与外语教学，2008（1）：6-10.

辛斌，高小丽. 批评话语分析：目标、方法与动态[J]. 外语与外语教学，2013（4）：1-5.

许家金，梁茂成，贾云龙. BFSU PowerConc 1.0 [Z]. 北京外国语大学：中国外语教育研究中心，2012.

通信地址：100089　北京市　北京外国语大学中国语言文学学院

德语话语分析的语料库转向*

北京外国语大学 徐泽茗 葛囡囡

提要：近二十年，德语话语分析发生了语料库转向，各学派有了新的发展。本文梳理了当代德语话语分析的缘起，对目前与语料库结合最为紧密的海德堡学派、杜塞尔多夫学派、维也纳学派和话语语言学多层面分析法依据研究传统、分析方法和研究主题进行了归纳梳理，并凝练出德语话语分析发生语料库转向后的新特点：“话语”概念与语料库联系紧密；研究范式与研究方法多样；重理论建构，轻实证研究；影响限于德语区与德语语言学界。最为重要的是，语料库的加入不仅影响到德语话语分析一个研究分支，还引发了德语语言学理论和方法的大变革。

关键词：话语分析、语料库转向、学派、德语语言学

1 引言

德语话语分析在Michel Foucault思想的影响下，形成了不同于英语地区的特色：以跨篇章结构为研究对象，关注语言与社会知识建构的关系。近二十年，在语料库语言学迅猛发展的背景下，德语话语分析发生了语料库转向，形成了新的发展趋势。本文首先厘清德语话语及话语分析概念，继而通过德语话语分析的不同学派发生语料库转向后的变化和发展脉络，探究德语话语分析的新特点，以期为其他相关研究提供借鉴。

2 当代德语话语分析的缘起

当代语言学界所称的话语分析（discourse analysis，简称DA）源于Harris（1952）提出的话语概念。以描写主义语言学的研究方法为基础，Harris将语言学的研究对象从句子扩展到超句结构，从而探索语篇的基本单位及其在语篇结构中发挥的作用，进而探索语篇组织中的规律[1]。20世纪60年代起，伴随语用学与社会语言学的发展，话语分析也进入蓬勃发展的阶段。这一时期，一度有对话语分

* 本文系北京市社会科学基金项目“基于语料库的‘一带一路’话语对比研究”（18YYC015）的阶段性成果。葛囡囡为本文通讯作者。
作者贡献：
徐泽茗：研究方法、数据收集、讨论结论、初稿撰写、字数占比（70%）；
葛囡囡：选题构思、研究方法、讨论结论、字数占比（30%）、修改润色。

析与会话分析（conversation analysis，简称CA）的区分。其中，话语分析的代表性研究主要有van Dijk（1972）的语篇语法（text grammar）和以de Beaugrande & Dressler（1981）为代表的篇章语言学（text linguistics），主要研究篇章结构、篇章衔接与连贯等问题，是将句子层面的语法研究拓展到篇章层面的尝试。而会话分析的主要代表为Sacks *et al.*（1974）、Schegloff *et al.*（1977），该路径从口头语言数据出发，探究口语会话的组织规律，主要研究范畴有话轮转换、会话模式等。在Levinson（1983：286-294）看来，这一时期话语分析与会话分析的研究对象都是语篇的衔接与序列，但在方法论上，话语分析偏向于对语言学中已较为成熟的理论原则与概念进行扩展，而会话分析则是从语言数据出发进行归纳性研究。但无论是话语分析还是会话分析，英语语言学界都更为关注口头篇章，因此二者发展至今，已经没有显著分界。20世纪80年代，受英语学界影响，德语学界也曾将“话语”解释为社会交际语境下的口语表达，从而将话语分析基本等同于会话分析（李彬 2014：17）。这是德语话语分析的一段小插曲[2]，这一研究方向也未成为德语话语分析的主流，因而不属于本文探讨的德语话语分析范畴。

20世纪90年代，在法兰克福学派批评理论、Foucault的后结构主义理论等思想的影响下，Fairclough、van Dijk、Wodak等欧洲话语研究者开始关注语言与社会权力结构的关系，形成了批评话语分析（critical discourse analysis，简称CDA）这一研究范式（Wodak & Meyer 2009），他们对discourse的定义也逐渐超越了单一的语篇，逐渐将抽象的社会话语也称为discourse。与英语语言学界不明确区分具体的语篇和抽象的社会话语不同，德语学界在Foucault影响下，对话语（Diskurs）和篇章（Text）两个概念进行了区分，认为话语是超越篇章层面的结构，在跨篇章层面进行的研究称为话语分析（Diskursanalyse），对篇章结构进行的研究则是篇章语言学（Textlinguistik）（Warnke 2008）。

德语学界主流的话语分析研究路径为批评话语分析（kritische Diskursanalyse，简称KDA）和语言学话语分析（linguistische Diskursanalyse，简称LDA），前者的特点是批判导向，关注社会问题，而后者更加注重跨篇章结构与社会知识建构（Niehr 2014：50-56）。不过，无论是批评性还是描述性的语言学研究，都基于一定的规范或视角，即使是从语言学视角出发的话语分析也必然受到价值判断的影响（Reisigl & Warnke 2013），可谓“殊途同归”。二者均关注语言与知识建构、社会结构的关系，区别仅在于其侧重点是批评还是描述，实质上都是英语学界所称的CDA。鉴于此，本文不再进一步区分LDA和KDA。

早期德语学界的话语分析以批评话语分析为主，主要是以Wodak为代表的维也纳学派和Jäger为代表的杜伊斯堡学派。在理论出发点上，两个学派与欧洲其他地区的批评话语分析理论有着相似的理论基础，都认为话语是社会建构的产物，致力于揭示乃至应对话语背后的社会问题，侧重批评视角。在选题上，德语国家

的种族主义思潮和战后外籍劳工的流入使两个学派都较为关注种族主义话语，如Wodak *et al.*（1990）对奥地利战后反犹主义话语的研究和Jäger（1993）对种族主义话语的研究。在方法上，两个学派均以质性研究方法为主，如维也纳学派发展出了话语历史分析法（discourse-historical approach，简称DHA），注重分析话语现象的历史变化，而杜伊斯堡学派确立并完善了定因分析法（dispositive analysis），通过对话语片段的质性分析探究整体社会话语，进而探寻以话语为载体的社会知识的特征（Jäger 2004）。

传统的话语分析研究在理论上不断完善，在其关注的话题领域取得了一系列研究成果。然而，单纯依靠传统的质性研究方法，选取的篇章数量有限，分析结果常被质疑主观解读性过强，普适性、客观性与可信度较低（李莎莎 2019）。语料库语言学的研究方法能够处理大规模语言数据，注重话语运用的具体语境，因而是适宜的话语分析研究路径（许家金 2019：6-9）。近二十年来，德语话语分析学者将语料库语言学方法引入话语分析领域，为阐释定性的话语分析提供定量的实证数据支持，从而提高了分析结果的客观性和可信度。由此，德语学界的话语分析发生了语料库转向，各学派也在已有研究基础上发展出了新的研究路径。

3　德语话语分析诸学派的语料库转向

3.1　海德堡学派

海德堡学派的LDA研究起源于海德堡–曼海姆研究团队（Die Heidelberg-Mannheimer Gruppe），特别是Busse（1987）等提出的历史语义学（historische Semantik）对整个学派影响深远。Busse主张跳出结构主义语义学对语义稳定性的假设，从历史语境的角度对语义进行观察，揭示体现在语义变化中的社会意识变迁。Busse & Teubert（1994）发展出了基于语义的话语概念，并开始探索与语料库的结合，将语义相关篇章组成的虚拟语料库称为话语，从而将语义学的研究对象扩展至跨篇章层面，这一定义构成了海德堡学派的共同理论基础，也为其与语料库的融合创造了先决条件。Hermanns（1995）厘清了话语与语料库的关系，认为与某一主题相关的所有书面与口语篇章组成了假想语料库（imaginäres Korpus），但其中有一部分因从未被记录下来而无法研究；某一历史时期得以保留的篇章则组成了虚拟语料库（virtuelles Korpus），即Busse和Teubert所称的话语，但这些篇章数量通常较大，无法全部研究，用于研究的那一部分是实际语料库（konkretes Korpus）。

Konerding（1993，2005）基于Busse和Teubert的话语定义，认为框架在人类知识结构的组织中发挥了重要作用，提出了基于宏观框架（Makro-Frame）确定话语主题间的语义相关性。Felder（2006，2015，2018）同样认同Busse和Teubert

的话语概念，认为社会中的权力结构会以语言为载体，体现在语义的冲突中，在具体话语中体现为不同的冲突中心（agonale Zentren）。但Felder（2012）认为，Konerding提出的语义类型对于具体分析来说较为抽象，因此提出可以通过明确话语主题、生成次级主题、确定冲突中心三个步骤来进行更为具体的语义分析。而在确定冲突中心时，Felder提出应当围绕事实建构、事实关联、事实评价三种言语行为类型在词、句法单位、句、篇章乃至图片等不同的符号层级上提取冲突中心，因此将这一分析方法命名为语用符号学篇章分析法（pragma-semiotische Textarbeit）。该方法的突出特点是结合语料库开展研究，主要用于分析媒体话语与法律话语，研究主题有辅助死亡（Felder *et al.* 2016）、能源政策（Jacob 2017）等德国社会热点问题。另外，Felder *et al.*（2010）创建了海德堡语料库（HeideKo），并基于这一语料库对媒体、法律、科技伦理话语中的语义冲突开展了一系列研究，旨在发现并揭示社会结构与语言的互动关系。

Vogel（2010）继承并发展了Konerding对认知框架的理解，赞同框架在知识建构中起到了认知原型的作用，尤其关注媒体话语所建构的公共形象（public image），将其视为整体话语现象的一部分，体现在大众媒体广泛传播的、重复出现的典型语言模式中。为此，他总结出了聚焦媒体话语的语言学形象分析法（linguistische Imageanalyse，简称LIma）。LIma在理论上强调重复性与典型性，因而在实证研究中着重发挥了语料库的作用，并有专门的研究工具（Vogel 2012）。LIma已应用于国家形象（Vogel 2010）与组织形象（Vogel 2014）等不同领域的形象分析中。

3.2 杜塞尔多夫学派

杜塞尔多夫学派同样基于Busse的历史语义学研究传统，注重研究概念和语义的历史流变，将语言的历史视为话语主题的历史（Stötzel & Wengeler 1995：14）。不过，Jung（1996）在Busse和Teubert的话语定义基础上，提出了新的“话语”定义，认为话语不是由篇章组成的语料库，而是由各个篇章中关于同一主题的具体陈述组成的语料库。这一定义将篇章视为单个陈述与整体话语之间的中间层，认为篇章中与主题无关的部分不属于话语的一部分。但实际操作中，借助于语料库语言学的手段更容易从大量篇章中提取与某一主题相关的陈述（Jung 2006），因此该学派积极地将语料库方法融入其话语分析。

该学派的研究范式较为复杂，其中影响最大的是论式分析法、隐喻研究和框架语义分析。Wengeler（2003，2015）、Niehr（2004）认为，论式（Topos）即论证模式，是某一特定时期在社会中较为广泛传播的思维模式，能够折射出社会思潮的变化，因而采用论式分析法对公共领域内话语进行分析。在认知语言学隐喻概念影响下，Böke（1996）将对隐喻的分析融入其话语分析研究范式中，观察

隐喻反映出的认知模式。在同属认知语言学的框架语义学影响下，Ziem（2008，2014）认为，框架作为一种知识结构，能够在话语中体现。他们完善了将框架分析和语料库结合的研究路径，发展了包括语料标注、述谓分析、确定上义词类型、述谓分类四个步骤的分析方法。

该学派注重考察话语的历史变化，早期研究关注政治或公共政策讨论，如核能（Jung 1994）、联邦德国政治话语（Böke *et al.*1996）、德国外来移民（Niehr & Böke 2000）等。近年来，该学派较关注经济领域话语，如Wengeler（2013）对不同时期德国经济危机相关报道中的论式进行了对比分析，Ziem（2014）研究了德国纸媒报道中金融投资者的隐喻框架。

3.3　维也纳学派

本文所称的维也纳学派，是形成于维也纳大学、采用DHA的话语分析流派，代表人物是Wodak。该学派一方面受Foucault思想的影响，区分了篇章和话语两个概念，但Wodak（2008）认为Foucault所用的话语概念不能理解为依据主题相关性确定的对象，而是话语事件之间的一系列关系；另一方面，受到法兰克福学派批评理论的影响，该学派有着更强的社会问题导向意识并积极开展跨学科研究。该学派前期致力于通过传统的质性方式从事话语分析。2004年起，Wodak同时任教于维也纳大学和兰卡斯特大学，她以兰卡斯特大学为中心，与团队开展了一系列基于语料库的话语分析，形成了基于语料库的话语分析中的重要一支（Baker & McEnery 2015）。

在研究方法上，该学派尤为注重历史语境对话语的影响，主要从内容和主题、话语策略、语言手段与具体语言形式三个维度入手，对话语进行质性为主的分析（Reisigl & Wodak 2016）。在DHA研究思路的基础上，Baker *et al.*（2008）提出了语料库辅助批评话语分析的研究策略，基于1.4亿词规模的英国新闻语料库，通过词频统计、关键词表、搭配与索引行等方法，分析了关于难民或移民的媒体报道中话语呈现的特点，形成了DHA与语料库方法结合的标志性成果。

该学派的研究主题多围绕英国、奥地利乃至欧洲地区的话语历史，重视对具体社会问题的研究与批评，特别关注欧洲范围内的身份认同、种族主义与歧视思潮，如欧盟话语中的欧洲概念建构及历史变化（Forchtner & Kølvraa 2012）、英国极右翼政党的自我形象与移民话语建构（Engström & Paradis 2015）以及奥地利民族身份的话语建构（De Cillia *et al.* 2020）。

3.4　话语语言学多层面分析法

不同于前三个学派坚持各自的话语概念并延续了相近的分析范式，Warnke & Spitzmüller（2008）提出的话语语言学多层面分析法（diskurslinguistische Mehr-

Ebenen-Analyse，简称DIMEAN）并非基于特定理论的某一学派，而是对德语地区多种话语分析理论的汇总，旨在为研究者提供一套可因需取用的分析模型。语料库语言学[3]因其具有重视上下文、侧重定量分析的特点，能够对传统的质性话语分析起到补充作用，是目前通过DIMEAN对跨篇章结构进行分析所采用的主要方法。

DIMEAN主要包括篇章内层面、话语参与者、跨篇章层面三个研究层面。篇章内层面指的是对篇章及其内部元素的分析，可进一步分为词汇、命题、篇章；话语参与者层面的分析是从话语中的各类行为者出发，从参与者在话语行动中的交互身份、持有的话语立场与话语行为的媒介三方面入手，探究话语参与者怎样沟通话语与篇章两个层面；跨篇章层面则围绕话语这一超越篇章的结构，包括了互文性、认知框架、论式、社会符号、社会意识等分析范畴。DIMEAN作为一种整合模型，并不是一套具体方法的合集，而是一种方法论，旨在为话语分析和邻近学科的研究提供指导（Spitzmüller & Warnke 2011：197-200）。因其具有丰富的层次和分析维度，完整地按照该模型在实证研究中进行详尽分析难度较大，研究者多选择与其研究范围、研究对象最相适应的研究层面。与语料库方法结合后，DIMEAN为各类话语分析提供了更直接的指导，例如Krüger（2016）从篇章内层面出发，分析了德国媒体报道中的老龄化问题，Glausch（2017）从语用、篇章内与跨篇章层面开展了对德国、意大利企业沟通中环保可持续性话语的研究。

4　德语话语分析语料库转向后的特点

德语话语分析的不同学派发生语料库转向后，形成了以下特点。

第一，“话语”概念与语料库联系紧密。虽然不同学派对“话语”的定义不同，但本质接近，均与语料库有着天然的联系。德语语言学界所称的话语概念受Foucault影响较深，故而强调跨篇章性，所以单一的质性分析方法不能充分满足跨篇章研究的需要，不能较好地揭示篇章中重复出现的语言模式。Busse和Teubert的话语概念在德语学界产生了较大影响，而这一概念自身已经融合了语料库语言学的思想，使得基于这一概念的诸多话语分析路径与语料库语言学更易结合。

第二，研究范式与研究方法多样。德语话语分析虽基于相似的话语概念，但不同学派产生了多样的研究范式与研究方法。德语话语分析尤为关注社会与语言的互动关系、知识建构，因此也将目光投向了关注社会认知的认知语言学。认知语言学中的框架概念对海德堡学派和杜塞尔多夫学派的Ziem产生了较大影响，而隐喻概念则成为杜塞尔多夫学派Böke研究路径的重点。在语料库手段的辅助下，德语学界对论式、框架、隐喻等概念的研究也更为全面。

第三，重理论建构，轻实证研究。整体来看，德语话语分析并不强调对语言

本体的研究或对语言数据的分析，而是注重诠释其所依据的理论模型及哲学渊源，试图建构较为完善的理论体系。多数研究呈现的实证部分篇幅较少，对语料库步骤以及对实证数据的客观介绍较为简略，其重点在于揭示话语体现的社会因素。但另一方面，宏大的理论框架与不同理论间的差异也使得部分研究成果的相互可理解性和可验证性受到影响。这集中体现在DIMEAN的理论模型中，该方法提供了层次丰富的分析模型，但正因为其分析层次过于复杂，直接开展实证研究的难度很大。

第四，影响限于德语区与德语语言学界。德语话语分析的国际影响力小，大多在德语区以及德语语言学界传播。究其原因，一方面在于德语话语分析诸学派均有着较为复杂的理论体系，客观上对理论的相互可理解性造成了负面影响。另一方面是因为多数德语话语分析的研究成果仅以德文发表，极大限制了其国际影响力。Busse和Teubert的“话语”定义几乎未对德语语言学界以外造成影响。然而，当德语话语分析的经典理论与英美学界接轨，就很可能产生有影响力的研究成果。海德堡学派的Teubert后期到英国伯明翰大学工作，研究重心转移到语料库语言学领域，他因此成为英语语言学界语料库语言学的代表人物之一，话语也始终是其研究的重要主题之一。同样地，Wodak将研究重心转移至兰卡斯特大学后，推动了基于语料库的话语分析研究，因而DHA方法是当前发源于德语国家的诸话语分析方法中国际影响最大的学派之一。

5　结语

德语话语分析的理论基础深受Foucault后结构主义思想影响，强调话语的跨篇章性，关注话语与社会知识建构的关系。近二十年来，德语话语分析发生了语料库转向，丰富了话语分析的理论，发展出了新的分析方法，产生了一系列研究，是世界话语分析重要的组成部分。更为重要的是，德语语言学界内篇章语言学与话语分析间的界限也因语料库的加入而模糊。一方面，篇章语言学将目光投向了更上一级的跨篇章结构，例如Warnke（2002）提出话语性（Diskursivität）也是篇章的基本属性之一。另一方面，大量篇章语言学的研究范畴被引入话语分析，用于单一篇章的分析方法得以拓展至对语料库内所有篇章的考察，如Rheindorf & Wodak（2019）基于DHA对奥地利德语篇章类型演变的研究。可以说，语料库不仅影响到话语分析一个研究分支，而且引起了德语语言学理论和方法的大变革。

虽然国内对于德语话语分析的理论引介和研究总体不多，但在其与语料库融合后，一些影响较大的学派在中国得到了比以往更加积极的关注，还产生了若干与中国主题相关的实证研究。中德合作的“中德形象报告”（Chinesisch-Deutscher Imagereport，简称CDI）项目建设了德语区媒体报道组成的CDI语料库，集中分析了报道中关于中国的刻板印象，提供了进一步促进中欧人民之间跨文化交流的

建议（Vogel & Jia 2017）。李媛、章吟（2018，2019）向国内学界介绍了Wengeler的论式分析法，并分析了德国媒体对中国核能的报道，发现了其中论式的变迁与集体认知的演变。此外，DHA还得到中国英语学界的关注，如杨敏、王敏（2019）系统梳理和介绍了DHA的理论源流和哲学、社会学基础。杨敏、符小丽（2018）基于语料库对美国纸质媒体关于希拉里邮件门的报道开展了实证分析。

语料库转向后的德语话语分析发展活跃，是一个非常有潜力和活力的研究领域，目前仍有许多需要深入探讨、研究和扩展的主题。期待未来国际和国内学界有更多关于德语话语分析的理论引介和实证研究。

注释

1 在中国亦有研究者将discourse analysis称为语篇分析，参见姜望琪（2011），冉志晗、冉永平（2015）。

2 Konerding（2009）认为英语的discourse analysis不是德语的Diskursanalyse，而应当翻译为Gesprächsanalyse（会话分析）。我们认为，尽管与欧陆国家学者相比，英美话语分析学者对口头篇章更为重视，但这不代表可以将话语分析与会话分析两个概念等同起来。

3 Spitzmüller & Warnke（2011）认同话语语言学应当成为一种语言学分支学科的观点，并对话语语言学与语料库语言学进行了区分，将话语语言学视为对语言含义的研究，是一种扩展的语义学，而语料库语言学是在跨篇章层面对语言本体进行的研究，但这并不代表语料库语言学与话语分析领域的主流观点。Teubert（2007）认为，Foucault提出的社会话语是语料库语言学的核心概念，对意义的研究不能从中分离。许家金（2014）也认为，对意义的研究是语料库语言学的本体关切之一。

参考文献

BAKER P, GABRIELATOS C, KHOSRAVINIK M, et al. A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press [J]. Discourse & Society, 2008, 19(3): 273-306.

BAKER P, MCENERY T. Introduction [C]//BAKER P, MCENERY T. Corpora and discourse studies: integrating discourse and corpora. Basingstoke: Palgrave Macmillan, 2015: 1-19.

BÖKE K. Überlegungen zu einer Metaphernanalyse im Dienste einer “parzellierten” Sprachgeschichtsschreibung [C]//BÖKE K, JUNG M, WENGELER M. Öffentlicher Sprachgebrauch. Praktische, theoretische und historische Perspektiven. Opladen: Westdeutscher Verlag, 1996: 431-452.

BÖKE K, LIEDTKE F, WENGELER M. Politische Leitvokabeln in der Adenauer-Ära [M]. Berlin/New York: de Gruyter, 1996.

BUSSE D. Historische Semantik [M]. Stuttgart: Klett-Cotta, 1987.

BUSSE D, TEUBERT W. Ist Diskurs ein sprachwissenschaftliches Objekt? Zur Methodenfrage der historischen Semantik [C]//BUSSE D, HERMANNS F, TEUBERT W. Begriffsgeschichte und Diskursgeschichte. Methodenfragen und Forschungsergebnisse der historischen Semantik. Opladen: Westdeutscher Verlag, 1994: 10-28.

DE BEAUGRANDE R, DRESSLER W. Einführung in die Textlinguistik [M]. Tübingen: Niemeyer, 1981.

DE CILLIA R, WODAK R, RHEINDORF M, et al. Österreichische Identitäten im Wandel: empirische Untersuchungen zu ihrer diskursiven Konstruktion 1995–2015 [M]. Wiesbaden: Springer VS, 2020.

ENGSTRÖM R, PARADIS C. The in-group and out-groups of the British National Party and the UK Independence Party: a corpus-based discourse-historical analysis [J]. Journal of Language and Politics, 2015, 14(4): 501-527.

FELDER E. Semantische Kämpfe in Wissensdomänen [C]//FELDER E. Semantische Kämpfe: Macht und Sprache in den Wissenschaften. Berlin/New York: de Gruyter, 2006: 13-46.

FELDER E. Pragma-semiotische Textarbeit und der hermeneutische Nutzen von Korpusanalysen für die linguistische Mediendiskursanalyse [C]//FELDER E, MÜLLER M, VOGEL F. Korpuspragmatik: thematische Korpora als basis diskurslinguistischer Analysen. Berlin/Boston: de Gruyter, 2012: 115-174.

FELDER E. Lexik und Grammatik der Agonalität in der linguistischen Diskursanalyse [C]//KÄMPER H, WARNKE I. Diskurs – interdisziplinär. Zugänge, Gegenstände, Perspektiven. Berlin: de Gruyter, 2015: 87-121.

FELDER E. Linguistische Diskursanalyse im Paradigma der pragma-semiotischen Textarbeit. Agonale Zentren als Deutungskategorien [C]//HAGEMANN J, STAFFELDT S. Pragmatiktheorien II. Diskursanalysen im Vergleich. Tübingen: Stauffenburg, 2018: 19-42.

FELDER E, LUTH J, VOGEL F. “Patientenautonomie” und “Lebensschutz”: eine empirische Studie zu agonalen Zentren im Rechtsdiskurs über Sterbehilfe [J]. Zeitschrift für Germanistische Linguistik, 2016, 44(1): 1-36.

FELDER E, MÜLLER M, VOGEL F. Das Heidelberger Korpus. Gesellschaftliche Konflikte im Spiegel der Sprache [J]. Zeitschrift für Germanistische Linguistik, 2010, 38(2): 314-319.

FORCHTNER B, KØLVRAA C. Narrating a “new Europe”: from “bitter past” to self-

righteousness? [J]. Discourse and Society, 2012, 23(4): 377-400.

GLAUSCH D. Nachhaltigkeitskommunikation im Sprachvergleich. Wie deutsche und italienische Unternehmen zum Thema Nachhaltigkeit Kommunizieren [M]. Wiesbaden: Springer VS, 2017.

HARRIS Z S. Discourse analysis [J]. Language, 1952, 28(1): 1-30.

HERMANNS F. Sprachgeschichte als Mentalitätsgeschichte. Überlegungen zu Sinn und Form und Gegenstand historischer Semantik [C]//GARDT A, MATTHEIER K, REICHMANN O. Sprachgeschichte des Neuhochdeutschen. Gegenstände, Methoden, Theorien. Tübingen: Niemeyer, 1995: 69-101.

JACOB K. Diskursive Kehrtwenden in der Energiepolitik [C]//ROSENBERGER N, KLEINBERGER U. Energiediskurs. Perspektiven auf Sprache und Kommunikation im Kontext der Energiewende. Bern: Peter Lang, 2017: 199-224.

JÄGER S. Kritische Diskursanalyse. Eine Einführung [M]. Duisburg: DISS, 1993.

JÄGER S. Kritische Diskursanalyse. Eine Einführung（4th edition）[M]. Münster: Unrast, 2004.

JUNG M. Öffentlichkeit und Sprachwandel: zur Geschichte des Diskurses über die Atomenergie [M]. Opladen: Westdeutscher Verlag, 1994.

JUNG M. Linguistische Diskursgeschichte [C]//BÖKE K, JUNG M, WENGELER M. Öffentlicher Sprachgebrauch. Praktische, theoretische und historische Perspektiven. Opladen: Westdeutscher Verlag, 1996: 453-472.

JUNG M. Diskurshistorische Analyse – eine linguistische Perspektive [C]//KELLER R, HIRSELAND A, SCHNEIDER W, et al. Handbuch sozialwissenschaftliche Diskursanalyse. Wiesbaden: VS Verlag für Sozialwissenschaften, 2006: 31-53.

KONERDING K-P. Frames und lexikalisches Bedeutungswissen [M]. Tübingen: Niemeyer, 1993.

KONERDING K-P. Diskurse, Themen und soziale Topik [C]//FRAAS C, KLEMM M. Mediendiskurse. Frankfurt am Main: Peter Lang, 2005: 9-38.

KONERDING K-P. Diskurslinguistik – eine neue linguistische Teildisziplin [C]// FELDER E. Sprache. Berlin: Springer, 2009: 155-177.

KRÜGER C. Diskurse des Alter(n)s: öffentliches Sprechen über Alter in der Bundesrepublik Deutschland [M]. Berlin/Boston: de Gruyter, 2016.

LEVINSON S C. Pragmatics [M]. Cambridge: Cambridge University Press, 1983.

NIEHR T. Der Streit um Migration in der Bundesrepublik Deutschland, der Schweiz und Österreich [M]. Heidelberg: Winter, 2004.

NIEHR T. Einführung in die linguistische Diskursanalyse [M]. Darmstadt: WBG, 2014.

NIEHR T, BÖKE K. Einwanderungsdiskurse [M]. Wiesbaden: Westdeutscher Verlag,

2000.

REISIGL M, WARNKE I H. Diskurslinguistik im Spannungsfeld von Deskription, Präskription und Kritik: eine Einleitung [C]//MEINHOF U H, REISIGL M, WARNKE I H. Diskurslinguistik im Spannungsfeld von Deskription und Kritik. Berlin: Akademie Verlag, 2013: 7-35.

REISIGL M, WODAK R. The discourse-historical approach (DHA) [C]//WODAK R, MEYER M. Methods of critical discourse studies (3rd edition). London: Sage, 2016: 23-61.

RHEINDORF M, WODAK R. Genre-related language change: discourse- and corpus-linguistic perspectives on Austrian German 1970-2010 [J]. Folia Linguistica, 2019, 53(1): 125-167.

SACKS H, SCHEGLOFF E A, JEFFERSON G. A simplest systematics for the organization of turn-taking for conversation [J]. Language, 1974, 50(4): 696-735.

SCHEGLOFF E A, JEFFERSON G, SACKS H. The preference for self-correction in the organization of repair in conversation [J]. Language, 1977, 53(2): 361-382.

SPITZMÜLLER J, WARNKE I H. Diskurslinguistik. Eine Einführung in Theorien und Methoden der transtextuellen Sprachanalyse [M]. Berlin, 2011: de Gruyter.

STÖTZEL G, WENGELER M. Kontroverse Begriffe [M]. Berlin/New York: de Gruyter, 1995.

TEUBERT W. Parole-linguistics and the diachronic dimension of the discourse [C]// HOEY M, MAHLBERG M, STUBBS M, et al. Text, discourse and corpora: theory and analysis. London/New York: Continuum, 2007: 57-87.

VAN DIJK T. Some aspects of text grammars [M]. The Hague: Mouton, 1972.

VOGEL F. Linguistische Imageanalyse (LIma). Grundlegende Überlegungen und exemplifizierende Studie zum öffentlichen Image von Türken und Türkei in deutschsprachigen Medien [J]. Deutsche Sprache (DS). Zeitschrift für Theorie, Praxis, Dokumentation, 2010 (4): 345-377.

VOGEL F. Das LDA-Toolkit. Korpuslinguistisches Analyseinstrument für kontrastive Diskurs- und Imageanalysen in Forschung und Lehre [J]. Zeitschrift für Angewandte Linguistik, 2012 (3): 129-165.

VOGEL F. Die Zukunft im Visier. Zur medialen Selbstinszenierung der Bundeswehr gegenüber Jugendlichen [J]. Medien & Kommunikationswissenschaft, 2014, 62(2): 190-215.

VOGEL F, JIA W. Chinesisch-Deutscher Imagereport: das Bild Chinas im Deutschsprachigen Raum aus Kultur-, Medien- und Sprachwissenschaftlicher Perspektive (2000–2013) [M]. Berlin/Boston: de Gruyter, 2017.

WARNKE I H. Adieu Text – bienvenue Diskurs? Über Sinn und Zweck einer

poststrukturalistischen Entgrenzung des Textbegriffs [C]//FIX U, ADAMZIK K, ANTOS G, et al. Brauchen wir einen neuen Textbegriff? Antworten auf eine Preisfrage. Frankfurt am Main: Peter Lang, 2002: 125-141.

WARNKE I H. Text und Diskurslinguistik [C]//JANICH N. Textlinguistik: 15 Einführungen. Tübingen: Gunter Narr, 2008: 35-52.

WARNKE I H, SPITZMÜLLER J. Methoden und Methodologie der Diskurslinguistik – Grundlagen und Verfahren einer Sprachwissenschaft jenseits textueller Grenzen [C]//WARNKE I H, SPITZMÜLLER J. Methoden der Diskurslinguistik. Sprachwissenschaftliche Zugünge zur transtextuellen Ebene. Berlin/New York: de Gruyter, 2008: 3-54.

WENGELER M. Topos und Diskurs [M]. Tübingen: Max Niemeyer, 2003.

WENGELER M. Historische Diskurssemantik: das Beispiel Wirtschaftskrisen [C]// ROTH K S, SPIEGEL C. Angewandte Diskurslinguistik. Felder, Probleme, Perspektiven. Berlin: Akademie Verlag, 2013: 43-60.

WENGELER M. Patterns of argumentation and the heterogeneity of social knowledge [J]. Journal of Language and Politics, 2015, 14(5): 689-711.

WODAK R. Introduction: discourse studies – important concepts and terms [C]// WODAK R, KRZYŻANOWSKI M. Qualitative discourse analysis in the social sciences. Basingstoke: Palgrave, 2008: 1-29.

WODAK R, MEYER M. Critical Discourse analysis: history, agenda, theory, and methodology [C]//WODAK R, MEYER M. Methods for critical discourse analysis (2nd edition). London: SAGE, 2009: 1-33.

WODAK R, NOWAK P, PELIKAN J, et al. “Wir Sind alle Unschuldige Täter”. Diskurshistorische Studien zum Nachkriegsantisemitismus [M]. Frankfurt am Main: Suhrkamp, 1990.

ZIEM A. Frames und sprachliches Wissen. Kognitive Aspekte der semantischen Kompetenz [M]. Berlin: de Gruyter, 2008.

ZIEM A. Frames of understanding in text and discourse: theoretical foundations and descriptive applications [M]. Amsterdam/Philadelphia: John Benjamins, 2014.

姜望琪. Harris的语篇分析[J]. 外语教学，2011（4）：13-17.

李彬. 福柯话语理论关照下的德语话语语言学的源起与发展[J]. 德语人文研究，2014（2）：16-22.

李莎莎. 德国主流媒体对中国“一带一路”倡议认知——一项语料库批评话语分析[J]. 德国研究，2019（2）：99-114.

李媛，章吟. 论式话语分析：理论与方法[J]. 中国外语，2018（1）：42-50.

李媛，章吟. 论式话语分析视域下的德国主流媒体中国核能话语嬗变研究[J]. 德国

研究，2019（3）：85-101.

冉志晗，冉永平. 语篇分析视域下的元话语研究：问题与突破[J]. 外语与外语教学，2015（2）：38-44.

许家金. 许家金谈语料库语言学本体与方法[J]. 语料库语言学，2014（2）：35-44.

许家金. 语料库与话语研究[M]. 北京：外语教学与研究出版社，2019.

杨敏，符小丽. 基于语料库的“历史语篇分析”（DHA）的过程与价值——以美国主流媒体对希拉里邮件门的话语建构为例[J]. 外国语，2018（2）：77-85.

杨敏，王敏. Ruth Wodak话语-历史分析法中的哲学社会学思想探索[J]. 外语教学与研究，2019（3）：39-47.

通信地址：100089　北京市　北京外国语大学德语学院

多媒体、多模态语料库协作管理平台的设计与实现*

中国社会科学院语言研究所 张永伟
北京外国语大学 刘沛鑫
北京大学 程 璐
北京外国语大学 顾曰国

提要：多媒体、多模态语料库协作管理平台服务于多媒体、多模态语料库建设，专供语料库建设者使用，支持多用户在线协作。平台拟帮助用户在建设多媒体、多模态语料库时降低门槛、节约成本、加速进程、提高入库语料质量。文章详细介绍了系统的研发背景、目标、架构、功能设计与实现，重点突出了对多媒体、多模态语料库的多维度支持和对语料库协作管理的支持，可为同类系统的研制提供参考借鉴。

关键词：多模态、多媒体、语料库管理系统、设计与实现

1 引言

随着语料库建库需求的增多、规模的扩大，语料库建设效率的问题日益凸显，对功能丰富的语料库协作管理系统的需求日益迫切。目前的语料库管理系统要么只关注文本语料，忽略了对多媒体、多模态语料的支持（Gleim *et al.* 2009；陈华辉 2001b），要么只关注语料的切分和标注，忽略了对建库流程的协作管理支持。随着基于多媒体、多模态语料库的研究日益增多（Garofolo *et al.* 2004；McCowan *et al.* 2005；Oertel *et al.* 2013），现有语料库管理系统无法满足大规模多媒体、多模态语料库的建设需求。

语料库管理系统的概念有狭义和广义之分，狭义的语料库管理系统只包含语料库和语料的增、删、改、查和基础统计等功能（何婷婷 2003）；广义的语料

* 本文系国家社科基金重大项目“中国老年人语言能力的常模、评估及干预体系研究”（21 & ZD294）、国家社科基金重大项目“近40年来两代大规模北京口语调查的多模态语料库建设及应用研究”（20 & ZD300）和国家语委“十三五”科研规划2020年度项目“辅助语文辞书编纂的人工智能关键技术研究”（WT135-69）阶段性研究结果。程璐为本文通讯作者。
作者贡献：
张永伟：研究方法、初稿撰写、字数占比（30%）、修改润色；刘沛鑫：数据收集、数据分析、字数占比（10%）；程璐：初稿撰写、字数占比（50%）、修改润色；顾曰国：选题构思、讨论结论、字数占比（10%）。

库管理系统还包含语料的采集、加工、检索、统计分析、输出等功能（胡凤国 2007）。本文按使用对象的不同，将广义的语料库管理系统划分为语料库使用者使用的服务平台（包含检索、统计分析、输出等功能）和语料库建设者使用的管理平台（包含检索、统计分析、输出以外的其他功能）。本文设计与实现的语料库管理系统聚焦于后者，供语料库建设者使用。

能将文字语料，音频语料和动、静态图像语料集成，研究者可以通过多模态方式加工、检索和统计的语料库被称为多模态语料库（顾曰国 2013；黄立鹤 2015），而多模态话语需要通过多媒体形式呈现。传统的语料库管理系统限于文本语料，对多媒体、多模态语料库并不适用，主要因为：（1）多媒体语料文件通常比文本语料更大，传输和存储耗费资源更多、难度更大；（2）多媒体语料库包含的媒体类型多、格式多、编码种类多，与文本语料的切分和处理均不相同；（3）多媒体语料的多模态标注具有物理时空属性，在标注格式、内容和标注方法上与文本语料不同。刘剑、胡开宝（2015）阐述多模态语料库创建时存在两大问题：（1）缺乏专门的建库工具；（2）在处理语料时耗费大量人力成本。

为了在建设多媒体、多模态语料库时降低门槛、节约成本、加速进程、提高质量，北京外国语大学人工智能与人类语言重点实验室研制了“多媒体、多模态语料库协作管理平台”（以下简称“平台”）。该平台是“多语言、多模态、多媒体、多环境”云端实验室的重要组成之一，拟协助语料库建设者异地线上协作，便捷共建高质量多媒体、多模态语料库。

2　前人工作述评

近二十年，国内外学者对语料库管理系统的设计已经做出了一些尝试，如表1所示。

表1　前人工作概要

研制者	应用概念	媒体类型	架构	元数据管理	语料标注	是否公开
Rüdiger Gleim	广义	文本	C/S	×	×	×
陈华辉	狭义	文本	C/S	×	×	×
胡凤国	广义	音频	C/S（主体）	√	√	×
于娜娜	广义	音频	B/S	×	√	×

Rüdiger Gleim研制了辅助人文学科计算需求的语料库管理和分析系统。该系统仅支持基本的文本语料管理功能，不支持多媒体、多模态语料。此外，对于文

本语料，该系统也不支持语料标注和元数据管理。陈华辉（2001b）研制了英语教学语料库管理系统，功能包含语料库的增、删，语料的增、删、改、查、格式转换、分类、加工、检索等功能。陈华辉（2001a）的设计主要强调检索功能，管理功能有所欠缺，且仅支持英语文本生语料的管理。胡凤国（2006）研制了传媒语言语料库管理系统，包含采集、标注、导入、检索、输出子系统。该系统虽然整体架构较为完整，但用途单一，是专门为“传媒语言”定制开发的系统。该系统语种只支持汉语，熟语料库管理模块的许多功能尚未实现，并且语料采集和元数据标注依赖人工操作，缺乏计算机自动化或辅助支持。于娜娜（2017）研制了基于B/S（浏览器/服务器）架构的语音语料库管理系统，包含语音语料的收纳、格式检查、标注、特征提取、训练集管理等模块。为了应对用户通过Web访问语料库管理系统时的协作问题，还创建了登录注册和用户管理的模块。但于娜娜（2017）的设计仅适用于语音语料，缺乏对语料元数据的管理，在协作方面存在用户角色过少、权限和角色强绑定等问题。

上述系统主要有两方面不足。（1）难以用于多媒体、多模态语料库的建设。上述系统难以满足多媒体、多模态语料的传输、存储、展示、加工、元信息管理、语料标注对平台的要求。（2）难以用于团队协作建库。现代语料库建设项目往往由团队成员共同完成，但上述系统对此缺乏考虑。上述系统大多采用C/S（客户端/服务器）架构，客户端与服务器点对点的连接将消耗大量服务器资源，不适用于大规模在线协作，且后期系统使用、维护、升级较为困难。此外，这些系统建成后往往不公开，难以获取使用，用户只能另求他法，如此限制了向其他语料库建设者提供建库服务。

本文设计和实现的系统拟解决上述两大问题。

3　平台整体设计

3.1　架构设计

顾曰国（2002）在进行多模态语料库建设时严格区分了资源库和语料库。资源库对内容不进行限定，有经过一定标准筛选后才可以进入语料库。这种区分在系统实现中同样适用。语料库建设项目中，团队不同成员扮演的角色不同，采集资源时，采集人员不需要掌握系统的语言学理论知识，只需要按照资源采集标准执行即可。资源库中的内容在导入语料库时则需要由具有语言学背景的专业人员筛选。资源库和语料库的区分降低了平台的使用门槛。此外，语料的采集往往费时费力，同一资源库中的资源可以用于建设不同研究目标的语料库，合理的资源复用可以将语言学背景的专业人员从采集资源的繁琐劳动中解放出来。根据以上观点，平台的整体架构设计如图1所示。

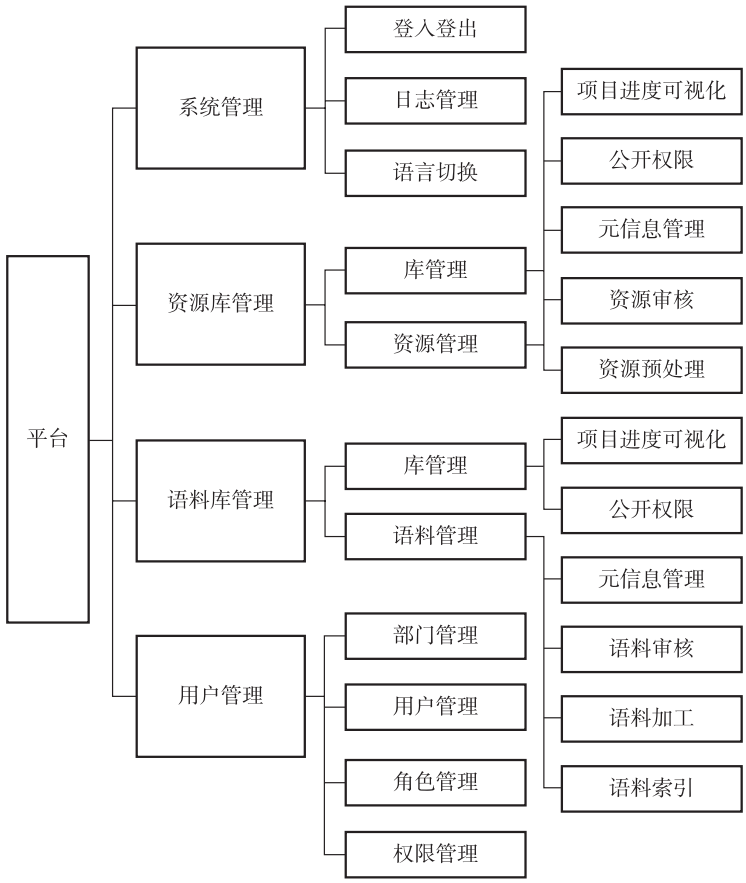


图1　平台整体架构

平台主要由系统管理、资源库管理、语料库管理、用户管理四个部分组成。

（1）系统管理包含三个子模块：系统登录、日志管理和语言切换。其中，日志管理使得平台操作有迹可循（见图2）；语言切换用于变更平台界面语言，支持不同语种的用户使用平台。

（2）资源库管理和语料库管理均分为库管理和库内容管理两个子模块。库管理支持在建库时限定库的公开权限，支持库的增、删、改、查等基本功能以及建库项目进度可视化和内部“成员管理”等功能。资源库内容的管理包含资源的增、删、改、查等基本功能以及资源的预览、审核、预处理、转码等功能。

（3）语料库内容的管理包含语料的增、删、改、查等基本功能和语料的预览、审核、加工、标注和索引等功能。

（4）用户管理包含部门管理、用户管理、角色管理和权限管理四个子模块。平台使用基于角色的权限访问控制，即RBAC模式（Sandhu 1998）对平台用户进行管理。

部分重点功能模块将在后续部分阐述。

图2　日志管理模块

3.2　技术选型

平台基于B/S架构，前端图形界面采取响应式设计，支持用户使用浏览器访问并线上协作；采用SpringCloud微服务框架开发，使用Java和JavaScript作为主要开发语言，前端使用Vue进行视图层展示；基于SpringCloud的微服务架构，采用Docker容器化部署，方便运维，在用户访问量提升后可以方便进行横向扩展来提升系统性能。根据保存数据的不同特点，平台采用了三种数据存储中间件。

（1）MinIO对象存储：主要用于存储外部数据，包括用户上传或平台自动采集的资源文件、语料文件及标注数据。多媒体语料，尤其是音频语料和视频语料往往文件较大，使用MinIO可以随着存储文件规模的扩大而动态扩容，具有良好的扩展性。

（2）MySQL关系型数据库：主要用于存储结构化的元数据、用户信息、部门信息和系统操作日志等。

（3）Redis缓存数据库：用于缓存数据字典、系统参数、用户登录Token等内部数据。

平台数据存储中间件的实现过程如图3所示。

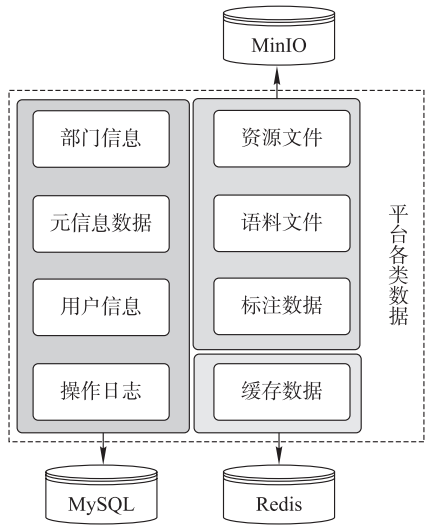


图3　数据存储实现过程

3.3　整体流程设计

平台的整体流程如图4所示，左边一列在资源库中进行，右边一列在语料库中进行。

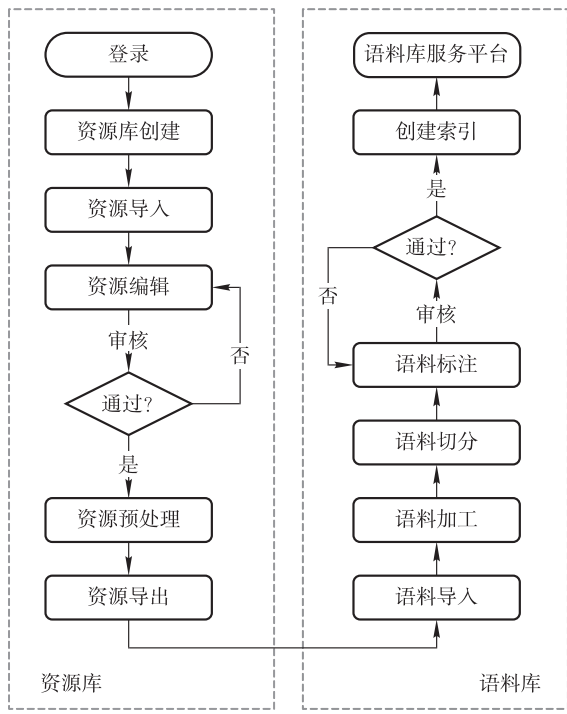


图4　平台整体流程

用户登录平台后可以创建资源库，导入资源。平台支持多媒体语料的在线预览。对于需要审核的资源库，拥有审核权限的用户可以对资源进行提取审核，判断是否通过。审核未通过的资源需要重新编辑后再审，审核通过的资源可以进行预处理。符合语料库入库标准的资源可以导入语料库，成为生语料。

语料库中的生语料经加工、切分、标注等处理后，成为熟语料。经过审核的熟语料将用于建立索引，为语料服务平台提供支持。

4　平台对多媒体、多模态语料的支持

4.1　多媒体语料加工

多媒体语料的加工主要指多媒体语料的切分、编码转换、格式转换等。平台为不同媒体类型的语料提供了不同的切分处理，比如：为文本语料库提供了分段、分句、分词等，为音频语料提供了音段切分，为静态图像提供空间区域切分，为视频语料同时提供时间片段以及空间区域的切分。编码转换主要针对图像、音频和视频文件进行编码转换、压缩、视频关键帧提取等处理，转码后的图像、音频、视频文件将被大部分主流浏览器支持预览。格式转换主要针对音频文件和视频文件的转写文本、标注文本，指将它们转为预览、建立索引时支持的格式。

4.2　语料多模态标注

对于多模态语料的标注而言，平台面临的最大难题是标注哪些内容、采用什么标签集方面缺乏系统、统一的规范标准。不同研究者的研究目的不同，多模态标注的需求不同。此外，不同媒体语料标注的内容也存在不同的倾向性，比如文本语料可以进行词法信息、句法信息、语义信息的标注，图像语料可以进行图像中的人物、背景的标注，音频语料可以进行讲话人识别、内容转写、音段语音特征的标注，视频语料可以进行肢体动作、面部表情、场景变化的标注等。

平台同时提供了自动标注和手工标注功能。自动标注依据媒体类型、语种调用自动标注工具进行标注，比如：对于日语、爱尔兰语等文本语料，可以进行词性标注、短语结构句法标注和依存句法标注等处理。但对于部分小语种，如萨摩亚语等文本语料目前限于缺少自动标注工具，尚无法进行自动标注。在手工标注方面，平台通过集成开源工具Label Studio实现基于XML自定义标注对象，对不同媒体类型提供不同的标注模板，满足用户多模态标注的需求。

4.3　多媒体、多模态语料预览

语料预览对语料使用和语料审核起支撑作用。针对语料的多模态特征，平台

支持语料本身和标注界面同步展示，便于用户手动标注语料、审核语料。语料预览允许用户在同一个界面预览具有时间、空间属性的原始语料及其标注，以及和时间、空间属性无关的原始语料及其标注。

5　平台对团队协作的支持

语料库建设需要团队配合，为实现高效的在线多人协作，平台在实现时从平台整体和资源/语料库两个层级重点考虑了两个问题：（1）谁可以访问？（2）谁可以操作？保证了平台在对外公开使用时，团队内权责清晰，团队间资源隔离。

5.1　平台层成员和权限管理

平台整体采用RBAC模式进行权限管理。平台管理员可以创建不同角色并赋予数据权限和操作权限。每个用户在创建时属于某个部门，将用户与角色关联即可限定用户平台权限。此处的权限可以控制用户平台可见的菜单项目，用户所在部门影响到下文组内公开可见的库内容。

5.2　资源/语料库层成员和权限管理

资源库和语料库在创建时需要限定库的公开情况，平台也支持以库为单位的成员和权限管理的功能，相较平台层来说，用户可以通过库层级的功能实现对资源/语料操作更细致的限定。

（1）资源/语料库公开情况。资源库和语料库在创建时需要限定库的公开情况。公开情况分为三种，在默认情况下包括：

a. 个人私有：仅对建库者本人开放；

b. 组内公开：对组内所有用户开放；

c. 完全公开：对注册平台的所有用户开放。

资源/语料库公开权限的限定可以对库内容起到一定的保护作用，将没有权限的用户隔离开来。对于用户来说，系统自动隐藏用户无权访问的库，减轻了用户查找的成本，降低了视觉负担。

（2）库权限管理。库的权限分为只读、编辑、审核、管理。创建者默认具有管理权限，即可以对库进行任何操作。用户对于面向自己公开的资源/语料库默认具有只读权限（例如组织内公开或完全公开的库），仅可以看见库中内容，如果需要进行其他操作，则需要管理者赋权。

（3）库成员管理。具有“管理”权限的用户可以通过库成员管理功能浏览所有平台用户，选中并添加至本库成员列表，或移除本库中的成员。“个人私有”情

况的资源/语料库也可以通过添加库成员分享给其他用户。

5.3 项目进度可视化展示

收集语言资源是一个费时费力的工作，为了让用户能够清晰了解目前语料库建设进展，平台采用特色的项目进度可视化展示界面。以资源/语料库为单位展示当前进度、库容量、语料类型占比等内容。用户可以通过创建项目文件夹，设定语言资源收集目标，收集过程中，进度即同步在项目进度可视化页面。

6 其他重点功能设计

6.1 审核模块

入库的资源和语料的质量关乎学术研究，因此质量把控至关重要。平台在资源库和语料库中分别提供两次审核功能，以保证资源和语料的质量。对于只有只读和编辑权限的用户，平台将自动隐藏提取和审核入口。具有审核或管理权限的用户可以通过提取相应文件，在界面中预览该资源，填写审核意见，选择通过或驳回，在审核窗口可以对资源进行放大、缩小、播放、暂停等操作。审核通过的资源可以进行下一步操作，例如在资源库中的资源预处理操作和在语料库中的创建索引操作；审核不通过的资源需要修改后再进行审核，否则不再支持进一步操作。

6.2 元信息管理模块

平台将元信息限定为对资源文件或语料文件的整体描述。元信息管理有利于对资源和语料更好地描述和检索。平台元信息分为两种类型。

（1）物理元信息。物理元信息指资源文件的物理属性信息，通用的如类型、文件名、大小等，个性化的如静态图像的拍摄日期、分辨率等信息，音频的编码、时长、比特率等信息。平台自动提取资源文件的物理元信息，允许部分元信息空缺，例如有的静态图像经过某些软件处理后，诸如拍摄日期等元信息会丢失。

（2）基本元信息。基本元信息指资源库和语料库在创建时，指定用户必填或选填的元信息。基本元信息可以分组管理，支持用户根据研究目的的不同进行自定义。比如创建某中介语作文语料库时，每个语料需指定作者信息和作文信息两个分组，作者信息分组包括作者国籍、性别等信息，作文信息分组包括写作时间、题目、分数等信息。基本元信息既可以内置在语料文件中，也可以在平台中填写并与语料文件关联。在使用时，不同位置保存的元信息将进行合并处理，内容不一致时，以平台中填写的元信息为准（见图5）。

图5　语料元信息管理

6.3　语料库配置模块

语料库配置分为索引配置和检索配置两种类型。

（1）索引配置。索引配置指配置创建语料库索引时需要使用的参数，如语料文件类型，文字书写方向（从左向右还是从右向左）、索引文件保存路径、索引名称、元数据及其解析方式、索引的标注及其解析方式。其中标注最重要，包括标注字段名、标注字段描述、是否在索引中保存、是否建立前向索引、大小写敏感性等信息。

（2）检索配置。检索配置指配置查询语料库索引时需要使用的参数，比如索引文件保存路径，分析功能默认设置，性能、缓存、日志等信息等。其中分析功能默认设置和具体应用有关，比如检索功能的默认页面大小、默认大小写是否敏感、检索行居中显示时左右侧词语数量等，共现搭配功能的默认关键性计算公式、跨距（span）值等。

7　结语

随着基于多媒体、多模态语料库的研究日益增多，多媒体、多模态语料库建设难度大、复杂度高的矛盾日益突出，使得多媒体、多模态语料库管理平台的研制显得尤为重要。

本文首先针对多媒体、多模态语料库建设的特点，系统阐述了多媒体、多模态语料库管理平台的整体架构、技术选型和数据库实现，为相关平台的研制提供参考。其次，从“多媒体、多模态”和“协作”两个角度出发介绍了平台对于这

三方面的重点考虑及功能设计。最后，介绍了平台部分其他重点功能的实现，保障平台设计的完整性。

本平台真正实现了用户在线资源创建、预览、审核、预处理、加工、标注、建立索引全流程的无缝衔接操作，大大提高了研发人员的建库效率和内容管理效率，降低了语料库工程项目的人力成本，保障了语料库内容质量，为后续基于多媒体、多模态语料库的学术研究打下基础。

参考文献

GAROFOLO J, LAPRUN C, MICHEL M, et al. The NIST meeting room pilot corpus [DB/OL]. International Language Resources and Evaluation Conference (LREC), 2004. https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=150471.

GLEIM R, WALTINGER U, ERNST A, et al. eHumanities desktop - An online system for corpus management and analysis in support of computing in the humanities [C]// Proceedings of the Demonstrations Session at EACL 2009. Athens: Association for Computational Linguistics, 2009: 21-24.

MCCOWAN I, CARLETTA J, KRAAIJ W, et al. The AMI meeting corpus [C]// NOLDUS L, GRIECO F, LOIJENS L, ZIMMERMAN P. Proceedings of the 5th international conference on methods and techniques in behavioral research. Wageningen: Noldus Information Technology, 2005, 88-100.

OERTEL C, CUMMINS F, EDLUND J, et al. D64: a corpus of richly recorded conversational interaction [J]. Journal on Multimodal User Interfaces, 2013, 7(1): 19-28.

SANDHU R S. Role-based access control [J]. Advances in Computers, 1998, 46: 237-286.

陈华辉. 一个中英文全文搜索引擎的设计与实现[J]. 计算机应用研究，2001a，（3）：131-133.

陈华辉. 英语教学语料库管理系统CMS的设计与实现[J]. 计算机工程，2001b，（5）：174-176.

顾曰国. 北京地区现场即席话语语料库的取样与代表性问题[C]//中国社会科学院世界经济研究中心. 全球化与21世纪. 北京：社会科学文献出版社，2002：484-500.

顾曰国. 论言思情貌整一原则与鲜活话语研究——多模态语料库语言学方法[J]. 当代修辞学，2013（6）：1-19.

何婷婷. 语料库研究[D]. 武汉：华中师范大学，2003.

胡凤国. 多层级语料库管理系统研究[D]. 北京：中国传媒大学，2006.

胡凤国. 多层级一体化语料库管理系统的开发[C]//萧国政，何炎祥，孙茂松. 中国

计算技术与语言问题研究——第七届中文信息处理国际会议论文集. 北京：电子工业出版社，2007：274-277.

黄立鹤. 语料库4.0：多模态语料库建设及其应用[J]. 解放军外国语学院学报，2015（3）：1-7.

刘剑，胡开宝. 多模态口译语料库的建设与应用研究[J]. 中国外语，2015（5）：77-85.

于娜娜. 基于B/S架构的语料库管理系统[D]. 哈尔滨：哈尔滨理工大学，2017.

通信地址：100029　北京市　中国社科院语言研究所/语料库暨计算语言学研究中心（张永伟）

100089　北京市　北京外国语大学人工智能与人类语言重点实验室（刘沛鑫）

102600　北京市　北京大学软件与微电子学院（程璐）

100089　北京市　北京外国语大学人工智能与人类语言重点实验室（顾曰国）

多轮对话的篇章级抽象语义表示标注体系研究*

南京师范大学 黄 彤 陈 瑾 谢媛媛 李 斌 曲维光

提要：对话分析是智能客服、聊天机器人等自然语言对话应用的基础课题，而对话存在大量情感短语、省略、语序颠倒等现象，对句法和语义分析器的影响较大，对话自动分析的准确率相对书面语料一直不高。其主要原因在于缺乏严整的多轮对话形式化描写方式，不利于后续的分析计算。因此，本文在梳理国内外针对对话的标注体系和语料库研究的基础上，提出了基于抽象语义表示的篇章级多轮对话标注体系，探讨了篇章级的语义结构标注方法，提出词语和概念关系的对齐方案，为称谓语和情感短语增加了相应的语义关系和概念，调整了表示主观情感词语的论元结构，并规定了对话中的一些特殊现象，设计了人工标注平台，以期为大规模的多轮对话语料库标注与计算研究奠定基础。

关键词：抽象语义表示、多轮对话、标注体系

1 引言

近年来，伴随人工智能的浪潮，问答系统、智能助手、聊天机器人等成为研究热门，人们希望机器能够像人一样思考，与人类对话，这就要求机器要能够理解、处理人的对话内容。对话分析是自然语言对话应用的基础，口语对话的分析正逐渐受到重视（宗成庆等 1999）。

就目前而言，多轮对话的篇章分析存在以下问题：首先，目前对话语义分析往往以处理普通文本的方式分析对话，导致自动分析效果较差。语义分析大多仍处在相对规范的书面文本层面，口语对话不同于书面语，对话中存在更多省略、语法不规范等现象。分析口语对话不能局限于对单句的分析，需要考虑上下文的信息。这些都给机器自动分析对话增加了难度。Adams（2017）尝试使用不同模型对对话语料进行依存解析，得到的F值仅有85.7%和80.3%，而常规语料均能达

* 本文系国家社科基金项目“中文抽象语义库的构建及自动分析研究”（18BYY127）的阶段成果。李斌为本文通讯作者。
作者贡献：
黄彤：选题构思、研究方法、讨论结论、初稿撰写、字数占比（40%）；陈瑾：数据收集、数据分析、讨论结论、字数占比（20%）、修改润色；谢媛媛：数据收集、讨论结论、字数占比（10%）；李斌：选题构思、讨论结论、字数占比（20%）；曲维光：字数占比（10%）、修改润色。

到90%以上，存在一定差距，对话解析的效果不甚如意。其次，多轮对话缺乏整体的篇章表示体系和语料建设。目前的语料库资源大多以书面语料为主，专门的对话语料较少，而面向对话的语料库和语料标注规范研究主要集中在对话行为、篇章关系等特定领域，一般只标注说话人、话轮信息、词性或句法分析结构，而忽视话轮间应答关系、话轮内部小句的关系，以及省略恢复、指代消解等难题。对话在篇章层面上的语义结构、应答逻辑没有得到有效的研究和表述，因此需要高质量的口语对话资源，以推动语义理解模型的发展（郑桂东 2018）。

本文提出了一种针对对话的语义表示方法——对话抽象语义表示（Dialogue Abstract Meaning Representation，简称DAMR）——来解决篇章级多轮对话的语义表示问题。这个方法基于中文抽象语义表示（Chinese Abstract Meaning Representation，简称CAMR）改进而来。抽象语义表示（Abstract Meaning Representation，简称AMR）作为一种新兴的句子语义表示方法，采用单根有向无环图来表示句子的语义结构（Banarescu *et al.* 2013），能够有效解决句子中的论元共享、省略、冗余、语序错乱等难题，并进行了多语言的理论和计算实践（Oepen *et al.* 2019），标注了上万句英文语料和汉语语料。不过，AMR虽然已经能较好地表示句子语义，但由于对话语料和常规书面语存在较大差异，例如省略、独立的称呼语、情感短语、冗余等，且目前CAMR仅针对单句进行了标注，而对话标注必然是篇章级别的，因此不能直接套用CAMR的规范来表示中文对话的语义，需要根据对话特点对CAMR的框架和规范进行调整、改进和扩充，使之能够表示多轮的对话语料。

因此，我们提出的DAMR继承了CAMR的框架和理论。DAMR是一种针对中文对话的篇章级句子语义表示方法，DAMR从四个方面进行了改进：（1）改进概念关系对齐的语法，将篇章信息融合到语料标注中；（2）针对对话特点，增加概念标签和关系标签；（3）调整了部分词语的论元结构；（4）对一些对话中的称呼语、情感短语特殊现象进行了规定。

2 相关工作

专门针对对话的标注体系和语料库较少，由于主要面向智能机器人，因此多为限定领域的标注，且标注重点为对话行为、篇章关系等，而完整的对话标注体系还需要包括对话的基本信息、指代信息、句法语义信息等。表1给出了下文提到的对话语料库的标注内容。

表1　语料库标注信息

语料库	对话基本信息	指代信息	篇章结构	句法语义信息	对话行为
ISO24617-2	√				√
LUNA（2007）		√		√	√
MATE（2009）		√		√	√
Martinez（2002）	√			√	√
Zhou（2010）	√				√
周小强（2018）	√		√		√

2.1　对话行为信息标注及语料

多层对话行为标注（Dialogue Act Markup in Several Layers，简称DAMSL）是应用最为广泛的一种面向任务的通用领域标注体系，DAMSL在四个维度上对对话行为进行了标注，包括：交际状态记录话语是否完整，信息层面标注话语的特征，向前功能记录当前话语与之后话语的联系，向后功能记录当前话语与之前话语的联系（Allen & Core 1997）。在DAMSL提出之后，部分学者使用DAMS 标注体系对语料库进行标注，其中最为著名的是Switchboard（简称SWBD）电话语料库，其目的是进一步提高自动语音识别的语言模型（Jurafsky *et al.*1997）。MRDA（Meeting Recorder Project）对话标注体系则是在SWBD-DAMSL的基础上修改的标注体系，用于标注ICSI（International Computer Science Institute）项目的英语会议多人对话内容，形成了ISCI-MRDA语料库（方称宇等 2013）。

Bunt *et al.*（2010）认为DAMSL的维度存在模糊性，提出了一种新的体系DIT++。DIT++细分十个维度，包括联系管理、任务/活动行为、自我反馈、启他反馈、话轮管理、时间管理、社会义务管理、自我交际管理、语篇构建和伙伴交际管理，并规定了每个维度下的交际功能，设计了两类标签：通用目的功能和特定维度功能标记集，两个标记集下又分多层多个标签。DIT++体系已应用于多个语料库中，如DIAMOND人人对话库、面向任务的AMI人人对话库等（方称宇等 2013）。随着对话行为标注体系的不断发展，Bunt *et al.*（2010）根据DAMSL和DIT++等多个对话行为标注体系的特点，提出了多维度的对话行为标注国际标准：ISO24617-2，借鉴DIT++设定了九大维度，包括任务、自我反馈、启他反馈、话轮管理、时间管理、社会义务管理、自我交际管理、语篇构建和伙伴交际管理，各个维度下也设计了相应的对话行为标签。除了通用领域的对话行为标注，还有部分针对特殊领域的标注语料库，如美国基于查询铁路交通的人机对话语料TRIANS（Allen & Heeman 1995）、查找路线的人人对话语料、英国HCRC语料（Carletta *et al.*

1996）。这些标注体系都根据各自的语料特点设定了限于该领域的标签。

汉语对话行为标注的相关研究有限。王珊、刘锐（2016）建立了一个电视台访谈节目语料库，基于国外对话行为的研究，通过对语料库中问答句子的分析，设计了汉语单层级对话行为的类别。周强（2017）基于国外DAMSL、SWBD-DAMSL等标注体系，借鉴了ISO标准中的维度设计，设计了五大标记集，各个标记集下面再分不同标记。

AMR可根据40多种语义关系标签来表示相应的意图或语用功能，例如语气关系标签“mod”可以表示说话人“祈使”“询问”等意图，因此DAMR暂时不引入对话行为的标签，更注重使用原有体系表达说话者的实际语义。

2.2 篇章关系信息标注及语料

对话中篇章关系标注主要沿用宾州篇章树库（Penn Discouse TreeBank，简称PDTB）、修辞结构篇章树库（Rhetorical Structure Theory Discourse Treebank，简称RST）两大体系。

PDTB仅考虑相毗邻句子之间的关系，借鉴了谓词论元结构，以连接词为核心分别定义了两个论元arg1和arg2，连接关系包括显性关系（Explicit）、隐性关系（Implicit）、替代关系（AltLex）、实体关系（EntRel）和无关系（NoRel）。如果没有显性的连接词，标注人员要根据自己的判断表示出其连接关系，同时设定了多层多类语义关系标签。Tonelli *et al.*（2010）将PDTB体系用于LUNA口语对话语料库中，针对对话语料的特征对意义标签进行了调整。Xue *et al.*（2016）也同样将PDTB体系用于标注SMS短信息对话，根据信息对话的特点对标签进行增删。

RST将篇章关系称为修辞关系，设定了单核心和多核心两种修辞关系。修辞关系所连接的篇章单位如果存在主次区别，那么就是单核心关系，反之是多核心关系。RST与PDTB的最大区别在于，RST篇章结构树有层次，每个修辞关系都可以连接两个或多个篇章单位，这些篇章单位又可以组成更大的篇章单位和其他篇章单位形成修辞关系，最终形成一个有层次的篇章结构树（Carlson *et al.* 2001）。Stent（2000）首次将RST用于标注面向任务的对话语料中，针对对话或是标注领域特点，新增了一些修辞关系（如问答关系），并为某些范围过于广泛的标签设置了更具体的下级标签。

中文AMR中规定了10种篇章关系，我们将沿用这些关系来标注对话中的篇章关系。因为对话话题较为分散，因此存在篇章关系的两个或多个句子不仅局限于相毗邻的两个句子中，同时也会根据对话的实际特点增加相应的篇章关系标签。

2.3 综合信息标注语料库

综合标注的对话语料库指标注了多种信息的语料库，包括上文提到的对话行

为、篇章结构，还有语义信息、共指等信息。

LUNA语料库是一个跨语言（意大利语、波兰语、法语）、跨领域的人人、人机对话语料库。LUNA语料库采用了层标注，第一层为语义标注，第二层为领域属性标注，以及非必须的其他层，包括谓词结构框架、对话行为、指代信息等（Raymond *et al.* 2007）。如图1所示，领域属性标注层标注句子中语义块所属的领域及其属性，以“属性-值”对构成，语义块来自第一层的语义标注，谓词结构框架借鉴了FrameNet框架标注语义结构，为预先设定好的领域设定框架，再填入相应的元素。对话行为沿用DAMSL体系标签，同时，LUNA语料库标注了共指信息，将可标记共指的元素标记为given或new，如果标为given，则找出最近发生的对象并增加指针指向它。

buongiorno lei [pu`o iscriversi]$_{concept1}$ [agliesami]$_{concept2}$ [oppure]$_{concept3}$ [otteneredelle informazioni]$_{concept4}$ come la possoaiutare（早上好，你可以报名参加考试，也可以获取一些我能帮上忙的信息）	buongiorno [lei]$_{fe1}$ [pu`o iscriversi]$_{fe2}$ [agli esami]$_{fe3}$ [oppure otteneredelle informazioni]$_{fe4}$ come la possoaiutare
	set={id1，id2，id3}
<concept1 action: inscription>	…
<concept2 objectDB: examen>	set={id4}
<concept3 conjunctor: alternative>	frame=info-request
<concept4 action: obtain_info>	frame-element: {student，addressee，topic}
	<fe4 frame=“info-request”>

图1 LUNA标注方法示例

其他较为知名的语料库还有Martinez-Hinarejos *et al.*（2002）在铁路信息系统的对话语料上标注了三层标签，分别为对话行为、框架和实例，对话行为基于TRAINS体系的标签进行了调整，框架借用FrameNet的思想，为具体任务设置相应框架，实例则用来填充框架的槽；MATE语料库标注了语义信息、对话行为、共指信息（Poesio *et al.*1999）。Zhou（2010）建立了一个汉语的旅游领域语料库，共标注了十三层信息，包括话轮、主题、说话者信息、分词词性信息、拼音、语音转录、语音边界、句子重音、音量、非语言信息、基于ICSI-MARA体系的对话行为、形式错误信息和情绪。

LUNA、MATE、Martinez建立的语料库都是面向任务的语料库和标注方法，因此其意义标注仍是从对话行为出发，更加注重抽取出说话者所要实现的功能意图，再根据意图设定论元结构，无法完整地表示句子的语义。在指代标注上，LUNA等其他标注共指的语料库都只涉及名词，将上下文中共指的元素用同样的

ID连接依赖，忽略了指向一个完整事件的代词，因此不利于判断指示词和先行语之间的关系和指代消解的实现。另外，这些语料库的重点仍然是单句的语义标注，没有将有相应问答或其他对应关系的句子表示出来。

周小强等（2017）设计了一个交互式问答语料的关系结构标注体系。除标注了对话行为类别外，还标注了问答中的语义匹配关系和语义补充关系。其对应关系仅限于句子和句子之间的关系，但在实际语料中情况更为复杂，有补充和匹配关系的不一定为整个句子，可能只是句子中的一部分，因此这种方法存在问答点对应不明确的问题。

3 数据来源及AMR体系介绍

3.1 数据来源

我们在改进的中文抽象语义表示标注平台上试标部分中文短信息SMS对话语料，以分析对话标注中存在的问题。该语料共有15,000篇对话，我们从中选取了10篇对话、994个句子进行试标注，语料信息包括话语编号、说话人编号、时间信息。我们对其进行预处理，增加了话轮编号信息。选其作为试标语料是因为短信对话保留了日常对话的基本要素和特征，同时避免了肢体语言或现实环境语境对录音转写语料内容的影响。

3.2 AMR体系介绍

AMR可将句子中的实词抽象为概念节点，实词之间的关系则抽象为带有非核心语义关系标签的有向弧，忽略虚词和一些较虚的语义（冠词、时态、单复数），允许增加、删除或修改概念（Bonial *et al.* 2019）。

在这个基础上，O'Gorman *et al.*（2018）提出了标注多句AMR（Multi-sentence AMR，简称MSAMR），即将AMR拓展到篇章层面，但只关注了篇章中的共指现象，标注了名词、动词、代词、隐形角色的共指关系，MSAMR设定了三种共指关系：一致关系、部分–整体关系、成员–集合关系。Bonial *et al.*（2020）针对人机对话语料对AMR进行了改进，构建了对话AMR体系，主要有以下几点扩充：（1）在AMR的最上层设定了36个对话行为标签；（2）增加了时、体标签；（3）针对该人机对话语料的用途设定了空间参数。

李斌等（2017）在AMR体系的基础上提出融合概念对齐的一体化标注方案，针对汉语特有现象进行了改进，形成了CAMR。CAMR的改进内容如下：（1）为量词、时、体新增了语义关系标签；（2）还原重叠式，如“试试”还原为“试”；（3）组合离合式，如把“睡一会觉”合成概念“睡觉”；（4）为复句关系增加了关

系概念标签。

使用CAMR能够更完整、合理地标注对话。

第一，AMR关注的并非句子中的具体词语，而是句中抽象的概念和关系，允许增加、删除或修改概念，利用这个特点，可以在一定程度上解决对话中的倒序、冗余等情况，也可以对对话中省略的概念进行恢复，将话语中的语义合理地表示出来。

第二，CAMR进行了对齐改进，采用句中的序列进行编号，实现了概念与句中单词的对齐，有利于合理地表示指代、省略等情况，也有助于照应语和先行语之间关系的标注。

第三，CAMR为复句关系添加了并列、因果、让步、条件、转折、解释说明、选择、目的、递进、时序10个标签，如图2中的并列复句关系and。DAMR可用这10个复句关系标签表示对话篇章结构关系。

第四，CAMR新增的dcopy和refer用来标注两个概念之间的关系，有助于省略和指代照应的标注。

但CAMR标注对话存在一些问题，比如目前CAMR仅标注单句，复句关系仅限于同一句中的标注，一些在句中没有充当明确成分的词语无法标注等。因此，我们在CAMR的基础上进行了改进，具体标注方法在第4节中说明。

```
运河[1]的[2]整治[3]改善[4]了[5]该[6]县[7]的[8]投资[9]环境[10]，[11]吸引[12]了[13]外商[14]投资[15]。
x18/and
    ::op1（ ）x4/改善-01
        :arg0（ ）x3/整治-01
            :arg1（x2/的）x1/运河
        :aspect（ ）x5/了
            :arg1（ ）x10/环境
        :mod（ ）x9/投资
        :poss（x8/的）x7/县
        :mod（ ）x6/该
    :op2（ ）x12/吸引-01
        :arg0（ ）x3/整治-01
            :arg1（ ）x15/投资-01
        :arg0（ ）x14/外商
        :aspect（ ）x13/了
```

图2　CAMR示例

4 对话AMR标注体系

我们在改进版的CAMR标注平台上试标注了994句中文短信息SMS对话语料，尽可能在现有CAMR体系的基础上标注，同时根据标注对话遇到的问题对其进行调整，以期扩充CAMR的兼容性。对话中会有一些特有的成分，如称呼、表示情绪的成分，存在指代照应的距离较远的现象，话轮间问答不直接对应，出现大量的省略，诸如说话人/听话人人称代词的省略。因此针对对话的特点从以下几点对CAMR做出了改进：实现双层标签概念对齐、增加若干个标签、修改部分词语的论元结构、规定了一些对话中特殊现象的标注。

4.1 概念对齐

DAMR的每个句子都包含以下字段：语篇编号、话轮编号、句子编号、说话人编号、问答位置信息（见表2）。

表2 对话基本信息

语篇编号	话轮编号	句子编号	说话人编号	句子	问答位置信息
3	48	83	151460	过[1]一会[2]和[3]你[4]说[5]	-
3	49	84	131525	好[1]	-

前5个字段根据语料顺序自动分配，问答位置信息由人工标注，标注答句所对应问句的位置信息，如果非问答对应则不标注（本文其他例句如不涉及问答则省略该字段）。CAMR的概念标签采用xn的形式，n是根据输入的已分词的原始句子序列分配的有序编号。人工补充的概念则由标注系统分配随机编号。目前的CAMR编号仅适用于单句，无法跨句子进行标注，因此为了实现篇章级别的标注，DAMR 采用了双层编号，即用sn_xn来对齐句中的概念，sn根据输入的句子序列分配，xn 则仍旧根据词语在句中的序列分配，样例见图3。

s83_x5/说-01 :arg3(x3) x4/你 :time() x1/过-01 :arg1() x2/一会 **s84_x1/confirm** :arg1() s83_x5/说-01	x5/说-01 :arg3(x3) x4/你 :time() x1/过-01 :arg1() x2/一会 **x1/好-01** :arg0() x3/说-01

图3 DAMR/CAMR概念对齐（左图为对齐后，右图为对齐前）

为减轻标注人员的操作量，当句子只出现当前句子的概念，则仅使用xn标签，当出现其他句子的概念时，才完整表示sn_xn。

4.2 新增标签

DAMR沿用了CAMR的5个核心语义关系标签、44个非核心语义关系标签和109个专名概念。argx（x∈[0,4]）表示核心语义角色关系，每个谓词的每个义项都有自己的核心语义角色框架。非核心语义关系是指核心语义关系之外的语义角色关系，CAMR在AMR的基础上规定了目的、处所、时间等44种对所有谓词通用的非核心语义关系。根据对话语料特点，DAMR新增了4个概念标签和2个非核心语义关系标签以兼容对话中会出现的语义关系。

4.2.1 说话：speak

DAMR为对话新增了speak、speaker和hearer概念，对话中每一个句子的根节点都为概念speak。概念speak规定了三个论元，分别为：arg0: speaker（说话人）；arg1: thing speak（说话内容）；arg2: hearer（听话人）。说话人speaker和听话人hearer 为新增概念，标注时需根据实际语义标注出话语的说话人和听话人，如图4所示。本文其他例句会省略speak、speaker、hearer概念，以使例子更清晰。

语篇编号	话轮编号	句子编号	说话人编号	句子
16	400	709	131525	太[1]土[2]了[3]

```
s709_x5/speak
   :arg0() x4/speaker
   :arg2() x6/hearer
   :arg1() x2/土-01
       :degree() x1/太
       :aspect() x3/了
```

图4 speak概念标注示例

4.2.2 肯定：confirm

对话是交互的，听话人会对说话人的话语表达态度，最常见的是对上一句的肯定（如“是的”“嗯嗯”等），针对这种情况，DAMR新增了一个confirm概念。具体示例见图5，句146的根节点是肯定概念confirm，“嗯”是对“复习好痛苦”的肯定，标注时将“嗯”抽象为概念“confirm”，不再单独表示出来。

语篇编号	话轮编号	句子编号	说话人编号	句子
2	85	144	135882	555[1]，[2]复习[3]好[4]痛苦[5]
2	85	145	135882	……
2	86	146	138459	嗯[1]

```
s144_x5/痛苦-01
      :feeling() x1/555
      :arg0() x6/speaker
      :arg1() x3/复习
      :degree() x4/好

s146_x1/confirm
      :arg0() s144_x5/痛苦-01
      :arg1() s144_x5/痛苦-01
```

图5 confirm概念标注示例

4.2.3 情感：feeling

对话中说话人会用多种形式表达自己的心情，如“哈哈”“呵呵”“呜呜”等，以及在线上对话文本中会出现的表情包，甚至是单纯的标点符号，如“。。。”“…”，DAMR新增了非核心语义关系标签“feeling”来表示这种语义。

如图5所示，“555”表示说话人的心情，将其置于根节点“痛苦”的下层。由于心情的表示形式太过复杂，例如“呵呵”可表示偏正向的愉悦情绪，而现在网络上的新兴用法也将“呵呵”表负面情绪，因此为了避免标注不统一，目前DAMR只使用“feeling”标签，不区分具体的情绪类别。另外，一部分表情并非表示心情，而是表示“再见”“你好”等概念，对于这部分表情，DAMR要求对其进行语义转写，将其真正的语义表示出来。

4.2.4 称呼：naming

存在称呼语是对话中突出的特点，称呼本身不存在于谓词概念的论元结构中，为了更好地表示对话中的称呼现象，DAMR引入非核心语义关系标签“naming”。注意与原标签“name”区分，称呼并不等同于实际名字“name”，称呼是动态的，“name”则是静态的。

如图6所示，根节点speak支配的论元分别为arg0说话人、arg1说话内容、arg2听话人，因为说话人称呼听话人为“王老师”，因此naming在arg2节点的下层。如果称呼就是听话人的名字，则“naming”和“name”同时出现，都在arg2节点的下层。

语篇编号	话轮编号	句子编号	说话人编号	句子
20	549	967	151461	王老师[1]你[2]是否[3]可以[4]给[5]出[6]更[7]多[8]的proposition[10]

s967_x12/speak
:arg0() x12/speaker
:arg2() x4/hearer
:naming() x1/王老师
:arg1() x4/可以-01
:arg0() x5/给-01
:arg0() x2/你
:arg1() x10/proposition
:quant(x9/的) x8/多
:degree() x7/更
:direction() x6/出
:mode() x3/interrogative

图6　naming关系标签标注

4.3　论元结构

对话中的话语常常带有说话人或者听话人的主观态度，在CAMR中没有将这种态度表示出来，因此，DAMR 对一部分谓词的论元结构进行了修改，增加了一个论元：argh（态度持有人）。目前修改的谓词是含有主观态度的形容词或短语，如“好看”“很有主意”，这部分词语原来只有一个论元：arg0（entity describe）。如图7所示，“好看”除了arg0（电影）外，还增加了态度持有人argh，根据语义，“好看”的态度持有者是听话人hearer。

语篇编号	话轮编号	句子编号	说话人编号	句子
20	567	997	138158	今天[2]的[3]电影[4]好看[5]么[6]？[7]

x5/好看-01
:arg0() x1/电影
:time(x3/的) x2/今天
:argh() x4/hearer
:mode() x6_x7/interrogative

图7 “好看”的论元结构

4.4 对话中的特殊现象

4.4.1 问答句的对应

问答是对话中常见的形式，但是在对话中，说话人可能同时进行两个或多个话题，问句和答句不一定为相邻句，问答的语义对应较为分散，并且答句不一定正面回答问句。为体现问句和答句之间的语义联系，DAMR在每个句子上增加了一个字段，可以将答句与问句联系起来。

问句位置信息有两个维度。第一个维度为答句所对应的问句编号，第二个维度为所对应问句的根节点。如图8，问句257的根节点为x12，因此答句259 的问句位置信息为s257_x12（见表3）。

```
s257_x12/or
    :op1() x4/难-01
    :      arg0() x7/整-01
                :arg1() x3/socio
                      :mod() x1/那
                      :cunit() x2/个
    :op2() x6/难-01
          :polarity() x5/-
          :arg0() x7/整-01
    :mode() x7_x8_x9/interrogative
```

图8 疑问句的标注示例

表3 DAMR语料字段

语篇编号	话轮编号	句子编号	说话人编号	句子	问句位置信息
4	149	257	138375	那[1]个[2]socio[3]难[4]不[5]难[6]整[7]？[8]？[9]？[10]	
4	149	258	138375	嗯[1]是[2]的[3]小[4]野兽[5]最近[6]也[7]满[8]虚弱[9]的[10]	
4	150	259	138194	哎[1]我[2]觉得[3]我[4]重新[5]上[6]了[7]一[8]次[9]socio[10]似的[11]	s257_x12

4.4.2 问句的省略

由于对话双方处于同一个语境中，完成对话理解所需的背景知识是两者共享的，因此对话中的一方提问常常会省略很多成分，在标注时，需要根据句子实际语义将省略的问句成分表示出来。如图9所示，句2仅用一个问号表示说话者的疑问，其完整语义为：为什么说atac疯了。在标注时需将实际语义标注出来。

语篇编号	话轮编号	句子编号	说话人编号	句子
1	1	1	151430	atac[1] 疯[2] 了[3]
1	2	2	131525	？[1]

s1_x3/疯-01
:arg0() x1/atca
:aspect() x3/了

s2_x2/amr-unknown
:cause-of() s1_x3 疯
:mode() x1/interrogative

图9　省略问句的标注示例

4.4.3　人称省略

在对话中说话人常常会省略自己或听话人的人称，标注时要将省略的说话人或听话人补出来。如图10，“弄完”的施事为听话人，在标注时我们将省略的hearer补充出来。

语篇编号	话轮编号	句子编号	说话人编号	句子
7	143	232	131525	嗯[1] 不过[2] 先[3] 把[4] 小说[5] 弄完[6] 吧[7]

s232_x18/contrast
:arg1() x20/confirm
:arg0() x21/speaker
:arg1() s231_x16/causation
:arg2(x6/不过) x10/弄完-01
:arg0() x24/hearer
:arg1(x4/把) x5/小说
:time() x3/先
:mode() x7/imperative

图10　人称省略的标注示例

4.5　小结

我们改进了原CAMR标注平台，加入了篇章对话信息（如语篇编号、话轮编

号、说话人编号），通过对994句对话语料的标注，针对对话特点新增了标签，处理了称呼语、情感短语等对话特有现象，规定了省略、话轮间应答关系的标注，使CAMR体系从单句拓展到篇章级别。

5 结论

近年来，对话系统的发展越来越受到重视，对话语义的形式化表示的作用愈发凸显，但国内目前还没有较完整的表示对话语义的标注体系。本文梳理了国内外对话标注体系和语料库的发展，在CAMR体系的基础上进行改进扩充：实现跨单句层面的概念对齐，新增适用于对话语料的概念标签和非核心语义关系标签，修改词语的论元结构，规定了问答句对应、省略等对话中特有现象的标注，形成了对话标注体系DAMR。这些改进有利于解决对话中的省略和跨句子指代等问题，使问答点的对应更明确，能更完整地表达说话者语义，对对话的自动理解与分析有较大价值。

在今后的工作中，第一，我们将加强对对话语义特点的研究，尝试标注语音对话转写语料，针对实际对话特点和新出现的问题完善DAMR标注体系，使之能够适用于各个领域的对话语料，以验证DAMR的效果；第二，使用DAMR标注体系标注语料，构建一个大规模的对话AMR语料库，并进行统计分析；第三，我们希望通过对对话标注语料库的学习，提高对话自动分析的效果。

参考文献

ADAMS A. Dependency parsing and dialogue systems: an investigation of dependency parsing for commercial application [D]. Uppsala: Uppsala Universitet, 2017.

ALLEN J, CORE M. Draft of DAMSL: dialog act markup in several layers [Z]. Rochester: University of Rochester, 1997.

ALLEN J, HEEMAN P. TRAINS Spoken Dialog Corpus [DB/OL]. Linguistic Data Consortium, 1995. https://doi.org/10.35111/vdzw-3271.

BANARESCU L, BONIAL C, CAI S, et al. Abstract meaning representation for Sembanking [C]//Proceedings of the 7th linguistic annotation workshop and interoperability with discourse. 2013: 178-186.

BONIAL C N, DONATELLI L, ERVIN J, et al. Abstract meaning representation for human-robot dialogue [J]//Proceedings of the Society for Computation in Linguistics, 2019, 2(1): 236-246.

BONIAL C, DONATELLI L, ABRAMS M, et al. Dialogue-AMR: abstract meaning representation for dialogue [C]//Proceedings of the 12th Language Resources and

Evaluation Conference. 2020: 684-695.

BUNT H, ALEXANDERSSON J, CARLETTA J, et al. Towards an ISO standard for dialogue act annotation [C]//Proceedings of the Seventh International Conference on Language Resources and Evaluation. 2010: 2548-2555.

CARLETTA J, ISARD A, KOWTKO J, et al. HCRC dialogue structure coding manual [Z]. Human Communication Research Centre, 1996.

CARLSON L, MARCU D, OKUROWSKI M E. Building a discourse-tagged corpus in the framework of rhetorical structure theory [C]// Proceedings of the Second SIGdial Workshop on Discourse and Dialogue. 2001: 1-10.

JURAFSKY D, SHRIBERG L, BIASCA D. Switchboard SWBD-DAMSL shallow-discourse-function annotation coders manual [C]//Institute of Cognitive Science Technical Report. 1997: 97-102.

MARTINEZ-HINAREJOS C D, SANCHIS E, GARCIA-GRANADA F, et al. A labelling proposal to annotate dialogues [C]//Proceedings of the 3rd LREC. 2002: 1566-1582.

O'GORMAN T, REGAN M, GRIFFITT K, et al. AMR beyond the sentence: the multi-sentence AMR corpus [C]//Proceedings of the 27th international conference on computational linguistics. 2018: 3693-3702.

OEPEN S, ABEND O, HAJIC J, et al. MRP 2019: cross-framework meaning representation parsing [C]//Proceedings of the Shared Task on Cross-Framework Meaning Representation Parsing at the 2019 Conference on Natural Language Learning. 2019: 1-27.

POESIO M, BRUNESEAUX F, ROMARY L. The MATE meta-scheme for coreference in dialogues in multiple languages [C]//ACL'99 Workshop Towards Standards and Tools for Discourse Tagging. 1999: 65-74.

RAYMOND C, RICCARDI G, RODRIGEZ K J, et al. The LUNA corpus: an annotation scheme for a multi-domain multi-lingual dialogue corpus [C]//Proceedings of the 11th Workshop on the Semantics and Pragmatics of Dialogue. 2007: 185-186.

STENT A. Rhetorical structure in dialog [C]//INLG'2000 Proceedings of the First International Conference on Natural Language Generation. 2000: 247-252.

TONELLI S. RICCARDI G. PRASAD R, et al. Annotation of discourse relations for conversational spoken dialogs [C]// Proceedings of International Conference on Language Resources and Evaluation (LRFC2010). 2010: 2084-2090.

XUE N, SU Q, JEONG S. Annotating the discourse and dialogue structure of SMS message conversations [C]//Proceedings of the 10th Linguistic Annotation Workshop held in conjunction with ACL 2016 (LAW-X 2016). 2016: 180-187.

ZHOU K, LI A, YIN Z, et al. CASIA-CASSIL: a Chinese telephone conversation corpus in real scenarios with multi-leveled annotation [C]//Proceedings of the Seventh

International Conference on Language Resources and Evaluation. 2010: 2407-2413.

方称宇，曹竟，刘晓月. 基于语料库的最新ISO会话行为标注体系的研究：从SWBDDAMSL到SWBD-ISO[J]. 当代语言学，2013（4）：439-458.

李斌，闻媛，宋丽等. 融合概念对齐信息的中文AMR语料库的构建[J]. 中文信息学报，2017（6）：93-102.

王珊，刘锐. 谈话节目语料库的构建与会话结构分析[J]. 中文信息学报，2016（6）：140-146.

郑桂东. 多轮对话语料构建中的离群对话分析[D]. 哈尔滨：哈尔滨工业大学，2018.

周强. 汉语日常会话的对话行为分析标注研究[J]. 中文信息学报，2017（6）：75-82.

周小强，王晓龙，陈清财. 交互式问答的关系结构体系及标注[J]. 中文信息学报，2017（5）：1-10.

宗成庆，吴华，黄泰翼，等. 限定领域汉语口语对话语料分析[C]//黄昌宁，董振东，计算语言学文集. 北京：清华大学出版社，1999：115-122.

通信地址：210000　江苏省南京市　南京师范大学文学院

基于语料库的机械工程学术词汇表创建研究*

沈阳理工大学 常 乐 沈阳建筑大学 吴明海 陈 颖

提要：随着语料库语言学和计算机检索技术的发展，利用语料库制定学术词汇表已成为外语教学研究中的主流。本文报告了在创建由120篇学术论文构成的机械工程学术语料库（MEAC）的基础上，制定包含398个单词、289个词族的机械工程学术词汇表（MEAWL），并将其与Coxhead的学术词汇表（AWL）进行对比。结果显示，MEAWL在MEAC中的覆盖率为12.82%，略高于AWL11.95%的覆盖率，但由于MEAWL包含的词族数仅为AWL的570个词族的一半左右，能够在很大程度上减轻学生负担，因而更具适用性；两个词汇表的第一级子表（最高频词汇）的重合率仅为31.7%，反映出MEAWL独具的专业性，同时也能为机械工程学术英语教学和教材编写提供参考依据。

关键词：学术词汇表、语料库、机械工程、覆盖率

1 引言

词汇是语言的构成基础和重要组成部分。按照Nation（2013）对词汇的分类，学习者在习得词汇时遇到的主要困难既不是频繁出现的高频词，也不是特殊领域技术词汇，而是介乎两者之间的学术词汇，即用以体现诸如提出问题、回顾文献、阐述观点、描述方法、讨论结果以及阐述结论等语篇功能的词汇（亦被称为“准专业词汇”或“半技术词汇”）。学术词汇的掌握程度是衡量学习者学术能力的重要标准，也是影响学术写作的一个主要因素。然而，已有研究证明，学术词汇很难通过阅读小说或新闻习得，因为学术词汇在非学术文本中只占1.4%（小说）至4.5%（新闻）（Coxhead & Byrd 2007）。可见，突出学术词汇的重要性并对其进行专门化训练在外语词汇教学中具有十分重要的现实意义。语料库是在语言实际使

* 本文系教育部产学合作协同育人项目“基于语料库的英汉对照机械工程学术词汇与专业术语表创建研究”（202102546001）、2020年辽宁省教育厅科学研究经费项目“基于语料库的机械英语词汇表创建”（LG202015）、沈阳理工大学研究生教育教学改革研究项目“基于学术词汇表的《机械工程研究生学术英语读写》教材建设”（2021Y040549）的阶段成果。常乐为本文通讯作者。
作者贡献：
常乐：选题构思、数据收集、讨论结论、初稿撰写、字数占比（55%）、修改润色；吴明海：选题构思、研究方法、数据分析、字数占比（25%）；陈颖：选题构思、研究方法、字数占比（20%）。

用中真实出现过的语言材料的集合。桂诗春（2013）指出，随着语料库语言学和计算机检索技术的不断发展，利用语料库技术辅以教师经验来制定反映词语频数和覆盖面的词汇表已成为外语教学的主流。

近年来，随着机械工业在经济体系中的蓬勃发展，机械专业人才阅读专业英语文献和规范要求、在国际上发表最新研究或应用成果的学术需求与日俱增。然而，由于国内机械专业学者对学术英语和语篇特点了解不足，导致国际学术交流受到限制。纵览相关文献，国内外迄今尚无“机械英语语料库”问世，基于语料库的“机械英语学术词汇表”则更是未见先例。显然，创建机械学术语料库并编制学术词汇表，进而助力中国机械学者和技术人员与国际接轨已是当下的迫切所需。

2 文献回顾

学术词汇表分为通用学术词汇表和专门学术词汇表。通用学术词汇表关注不同学科所共同使用的学术词汇，而专门学术词汇表所关注的是某个学科或专业的学术词汇。

2.1 通用学术词汇表

West（1953）创建了具有里程碑意义的通用学术词汇表（General Service List，简称GSL），其中包含的2,000个常用词族直接影响了对英语中核心词汇的定义。另一个被广泛应用于语言教学和词汇研究的词汇表当属Coxhead（2000）创建的学术词汇表（Academic Word List，简称AWL），是一个建立在涵盖四大学科、350万词的学术英语语料库之上，包含570个词族的通用学术词汇表。众多学者对AWL的适用性展开了大量研究，发现其覆盖率稳定保持在10%左右（Hyland & Tse 2007）。按照Nation（2013）的比例划分，如果学习者能够掌握在整个语篇中占比10%的AWL，占比76.1%的GSL，至少5%的技术词汇，再加上专有名词、缩略词等，就能够读懂本专业学术语篇中大约95%的内容，基本达到了有效阅读的词汇要求，可以实现无障碍阅读学术文本的目标（Schmitt *et al.* 2017）。

然而，有研究者发现，AWL对于工程类学科的学术语篇覆盖率偏低。吴瑾、王同顺（2007）用AWL对比了上海交大科技英语语料库，发现AWL的覆盖率仅有9.3%。Martinez *et al.*（2009）发现不同专业的语料库中学术词汇出现频率的分布情况与AWL中依照词频而生成的10个子表吻合度存在较大差异。为此，刘佳、韩丽娜（2014：68）提出：“AWL不应该作为一个通用学术词表完全不经过修改地应用于任何学科”。Hyland & Tse（2007），刘迪麟、雷蕾（2020）建议语言教师应基于学科特点开发更具针对性的专业词汇表，以满足不同学科的需求。

2.2 工程学术英语词汇表

工程类学术词汇表分为通用工程学术词汇表和专业工程学术词汇表（韩丽娜、刘佳 2014）。自2006年以来，国内外研究者先后创建了多个基于语料库的通用工程学术英语词汇表（见表1）。

表1 通用工程学术英语词汇表

研究者	语料库及其构成	词汇表名称	词表属性
Mudraya（2006）	学生工程语料库（Student Engineering Corpus，简称SEC）；13本教材；2,000,000个词标，18,000个词类。	学生工程词汇表（Student Engineering Word List，简称SEWL）	1,260个高频词族；8,850个词类。
Ward（2009）	工程语料库（Engineering Corpus，简称EC）；化工、土木、电气、工业和机械等5个专业共25本教材；271,000个词标，10290个词类。	基础工程词汇表（Basic Engineering List，简称BEL）	299个单词
Hsu（2014）	工程教材语料库（Engineering Textbook Corpus）；航空、生物技术、电气、信息、材料、机械等20个专业的100本教材；4,570,000个词标，229,000个词类。	工程英语词汇表（Engineering English Word List，简称EEWL）	729个词族
Veenstra & Sato（2018）	理工教材语料库（Science Textbook Corpus）；四个学科（生物、化学、物理、工程）共12本教材；704,237个词标。	理工教材词汇表（Science Textbook Word List，简称STWL）	309个词族

Mudraya（2006）首创了学生工程词汇表（SEWL），并提倡开展数据驱动下的词汇教学。但SEWL并未区分基础词汇（高中阶段的词汇）与非基础词汇，包含了高达1,260个词族（8,850个词类），对于学生负担过重。Ward（2009）制成的基础工程词汇表（BEL）以词为单位，适用于高中阶段英语水平欠佳但在大学阶段必须读懂工程类教科书的学生。但他强调必须提醒教师和学生不仅要学习单词，也要注意其词法和语法等出现的语境。Hsu（2014）编制的工程英语词汇表（EEWL）能够覆盖除核心词汇外14.3%的学术篇章，进而断定如果学生在高中阶段掌握了2,000词族的高频词汇（GSL），并在此基础上通过半年到一年的时间学完EEWL，就能达到有效阅读理解工程类教材的水平。Veenstra & Sato（2018）编制的理工教材词汇表（STWL）对于“理工教材语料库”的覆盖率高达13.4%。

由于大部分通用学术词汇表含词量较多，占据了学生大量的时间，为学习者创建基于具体专业语料库的学术词汇表势在必行。为此，近年国内外先后涌现出

基于石油、化学、航海等工程专业语料库的专业工程学术词汇表（见表2）。

表2 专业工程学术英语词汇表

研究者	语料库及其构成	词汇表名称	词表属性
江淑娟（2010）	石油英语语料库；5本专业英语教材；18万词	石油英语学术词表（PAWL）	498个词；覆盖率11.3%，高于AWL3.2%。
Valipouri & Nassaji（2013）	化学研究论文语料库（CRAC）；涵盖四个专业方向的1185篇论文；400万词。	化学学术词汇表（CAWL）	1400词族；包含683个GSL词汇、327个AWL词汇以及390个其他词汇。
赵志刚（2015）	航海英语语料库（MEC）；教材、论文、杂志等；300万词；11个专题。	航海英语学术词表（MEAWL）	641个词/438个词族；覆盖率11.78%，高于AWL3.04%。

江淑娟（2010）制成了比AWL少72个词，覆盖率比AWL高出3.2%的石油英语学术词表（PAWL）。Valipouri & Nassaji（2013）选取化学专业论文制成化学研究论文语料库，依照Coxhead（2000）的选词标准最终创建了化学学术词汇表（CAWL）。赵志刚（2015）收集了航海类的英语教材、论文、杂志、公文和公约等文本，建成了航海英语语料库（MEC），并制成航海英语学术词表（MEAWL）。MEAWL在MEC中的覆盖率高达11.78%，比AWL在MEC中的覆盖率高出3.04%，对航海专业的学生更具适用性。

2.3 不足与争议

尽管上述众多学术词表都参考了AWL的选词标准，但频率与覆盖率的标准往往由研究者自行制定，缺乏统一的客观标准。此外，词表的制定还存在两个争议，即是否该排除West（1953）的通用英语词表（GSL）以及呈现形式该采用词族还是词类（刘迪麟、雷蕾 2020）。

AWL包括的单词都是GSL之外的词汇，体现的理念为：学习者在掌握了通用英语词表（GSL）中2,000词族的高频词后再学习学术词汇。然而，英语高频词中也含有学术词汇，而且反之亦然；另外，由于有些高频词在通用英语和学术英语中的词义和用法可能完全不同（即同形异义），使得高频词汇和学术词汇之间很难划出明晰的界线。Masrai & Milton（2018）通过研究AWL的词汇分布，发现学术词汇大多存在于3,000级别的词汇中，因此认为AWL的重要性与其说是在于对学术词汇的体现，还不如说是第三层次的1,000个高频词汇。事实上，正是基于此点考虑，近几年新开发的一些词汇表（如：Gardner & Davies 2014）并未将GSL的

高频词排除在外。笔者认为，既不能“走极端”，也不可“一刀切”，应该灵活掌握——鉴于基础阶段的本科生词汇量较小，针对他们建词表时不应排除GSL；而对于面向高年级本科生或研究生的学术词表，则可以考虑排除GSL。

受GSL词汇呈现方式以及构词方式的影响，AWL采用了词族的呈现形式。其依据为：学生在掌握了词汇的基本构词规则后，很可能较易习得某一词族的所有单词（Nation 2016）。然而，不可否认同一词族的不同词汇可能意义差别较大，甚至截然不同，并非学生通过掌握词干与屈折变化规则就能够完全习得。同时，用词族呈现词汇的方式会加重学生的学习负担——由570个词族构成的AWL实际上包含了至少3,100个单词。此外，刘迪麟、雷蕾（2020）认为，既然学生掌握了构词规则后就能习得词族的所有单词，就没有必要在词族中列出屈折和派生变化的方式；同时，这样也会影响到词表的表面效度（face validity）。笔者发现，用单词体现词表的方式使词汇的教与学更具针对性，更能有效减轻学生的学习负担，而且此呈现方式已逐渐被学术圈所接受（如：Ward 2009；Gardner & Davies 2014）。

3 研究内容与方法

基于频数驱动理念的语料库及其衍生出的词汇表可以满足对语言习得的基本要求。本研究基于该理念，建立机械工程学术英语语料库并检验AWL在该语料库中的覆盖率，同时探索创建机械工程学术英语词汇表的必要性与可行性。

3.1 语料库创建

本研究自建的语料库属于单学科跨专业单语语料库，系专门为机械工程研究生学术英语教学设计。本语料库及预建词表着眼于研究生未来的学术阅读和写作需求，因此只收集一类语料，即学术论文。参照专用英语语料库的建设原则（梁茂成等 2010），为确保语料的代表性和平衡性，由机械工程学科专家帮助确定机械工程涵盖的三个专业方向，并分别推荐两种权威国际学术期刊，最终在保证每个专业方向有至少一本期刊的前提下确定了四本学术期刊，即*International Journal of Machine Tools and Manufacture*（《机械工具和制造国际学刊》）、*Journal of Mechanical Design*（《机械设计学刊》）、*Journal of Microelectromechanical Systems*（《微机电系统学刊》）以及*Mechatronics*（《机械电子学》）。从以上期刊中选取近5年由英语国家研究者撰写的学术论文共计120篇，并以全文形式保存。

在语料处理方面，首先，所有论文被转换为txt纯文本格式，并按照专业子方向分类；其次，通过语料整理与清洁除去文本中的图表、公式、符号、参考书目、附录等计算机分析软件所无法处理的内容；然后，将余下的所有正文部分按照专业方向组建成子库；最后，将三个子库合成为包含120篇文章、755,566个形符、20,426个类符的“机械工程学术英语语料库”（Mechanical Engineering Academic

Corpus，简称MEAC）。

3.2 词汇表创建

首先，词汇表创建的前期采用AWL（Coxhead 2000）的方法确定词汇专业性（specialized occurrence）、词汇覆盖面（range）以及频次（frequency）。具体而言，本词汇表为面向研究生的学术词汇表，排除了West（1953）的GSL中的单词。在词汇覆盖面方面，参考前人研究标准，选择了覆盖三个方向中至少两个方向的词汇。为了确定选词的频率标准，参考Francis和Kucera对Brown语料库的研究发现，即找出词频为100次的词族需要350万词的语料库（Coxhead 2000：218），经计算，对于库容75万词的MEAC而言，词频最终被确定在22次。

接下来，利用FileJoin将三个专业方向的文本分别合成为一个txt文本，并将其分别载入由Nation & Heatley（2002）研发的词汇检索软件Range_GSL_AWL，选择BASEWRD1和2（即去除第一层和第二层的各1,000个高频词汇）。然后，将运行结果中Type not found in any list中的词表拷入Excel，并对Range和Frequency作两次排序（即取Range≥2，Freq≥22的结果），将结果中的缩略词、数字、公式符号、专有名词等提取出来后，请机械学科专家对选出的词汇进行最后敲定，最终生成了包含398个单词（经统计为289个词族）的机械工程学术英语词汇表（Mechanical Engineering Academic Word List，简称MEAWL）。

4 结果与讨论

4.1 覆盖率

为了对比MEAWL和AWL对于MEAC的覆盖率，它们被分别置于Range软件中，结果显示：AWL在MEAC中的覆盖率为11.95%，完全符合“其在学术语篇中10%左右的覆盖率”的论断（见表4）。11.95%的比例高于许多前人的研究发现，证实了AWL在机械工程学术英语语料库中具有一定的适用性，说明AWL在机械专业中也能发挥一定作用。

相比之下，MEAWL的词族数几乎只有AWL的一半，其单词数量远低于AWL，在MEAC中的覆盖率却已达12.82%，略高于AWL。MEAWL的制定预设排除了2,000个GSL高频词汇，更适合研究生阶段的学习者；MEAWL只包含289个词族（或398个单词），极大地减轻了教师和学生的负担。此外，表4还显现出MEAWL和AWL共有151个重合的词族，即在AWL中只有151个词族（占AWL的26.5%），符合MEAWL的选词标准。换言之，在AWL中有将近3/4的词并不在MEAWL之中，这证明了在AWL之外存在大量的机械专业领域更常使用或更具专业性的学术词汇，同时也降低了普通学术词表在机械领域的适用性和重要性。

表4　MEAWL和AWL的对比

词汇表	词族数	在MEAC中的覆盖率	重合词族数	重合率
AWL	570	11.95%	151	26.5%
MEAWL	289	12.82%	151	52.2%

将MEAWL作为第三层级的词表加入Range中，结果显示：在GSL的2,000高频词族基础上，其整体覆盖率为82.05%，若加上至少占比5%的技术词汇，以及专有名词和缩写，则完全可以达到对有效阅读学术文本的词汇要求（见表5）。可见，在掌握更少词汇的情况下，MEAWL能够更加高效地帮助机械工程研究生读懂本专业学术论文，因而对他们具有更高的适用性。GSL的两个层级（即前2,000词）在MEAC中的覆盖率仅为69.23%，远低于其在AWL中76.1%的覆盖率（Coxhead 2000：223）。该结果从一个侧面证实了MEAC更加突出学术性，从另外一个侧面也与Masrai & Milton（2018）的观点相吻合。

表5　GSL和AWL/MEAWL对于AWC/MEAC的覆盖率

词汇表	对AWC的覆盖率	对MEAC的覆盖率
GSL 1-1000	71.4%	62.81%
GSL 1001-2000	4.7%	6.42%
合计	76.1%	69.23%
AWL/MEAWL	10.0%（AWL）	12.82%（MEAWL）
合计	86.1%	82.05%

4.2　高频词汇

Valipouri & Nassaji（2013）把CAWL中最常见的60个词与AWL第一级子表中的60个词相比较，只有18个重合词。同样，MEAWL中最常见的60个词（即出现频率400次以上的高频词汇）也被列为第一级子表（见表6），并与AWL第一级子表进行比较后发现，只有如approach、data、function、similar等19个重合词，重合率仅为31.7%，表明AWL在机械学术英语中适用的局限性。鉴于此，笔者有理由质疑Coxhead & Nation（2001）提出的“学生在掌握了GSL后可以学习AWL”的词汇学习顺序，建议学生优先学习与本专业更相关的词汇。

MEAWL第一级子表中的algorithm、axis、chip、fabrication、geometry、linear、machining、radius、sensors、vector、velocity等词在AWL中未出现，反映出MEAWL

所独具的专业性，即机械工程学术词汇在重要性上有其独特的顺序，因而有必要通过出现频率编排、制定机械工程学术英语词汇表为机械工程英语的教材编写和教学提供更具指导意义的依据，使教学更具针对性。与此相反，在MEAWL中无法找到AWL第一级子表中的contract、economy、export、formula、income、interpret、labour、legal、legislate、policy等词，显现出AWL中较为偏重的商科和法学等学科成分。此结果说明，由于每个专业都有自身独特的研究内容和研究方法，在词汇的使用习惯上也存在差异，创建机械工程学术词汇表具有很高的必要性与可行性。

表6 MEAWL的第一级子表

accuracy	design/designs	identified	required
algorithm	device	layer	research
analysis	dynamic	linear	section
approach	elements	machining	sensors
area	energy	maximum	significant
assembly	equations	mechanism	similar
axis	error/errors	method/methods	simulation
chip	evaluated	mode	specific
complexity	fabrication	obtained	stability
components	factor	optimization	structure
constant	features	parameters	technique
contact	final	prediction	thermal
create	function	process	variation
data	generated	radius	vector
defined	geometry	range	velocity

此外，MEAWL第一级子表中的大部分动词都以过去分词的形式出现，如defined、evaluated、generated、identified、obtained、required等，体现出理工科文本中多使用被动句表达客观事实的语篇特点（Hyland 2008）。

5 结论

语料库在提供大量真实数据方面具有优越性，其研究核心是发现在大量文本中频繁出现的潜在语言模式。编制词表的宗旨就是以最少的词汇量达到最大的覆盖率，以帮助读者在词汇及其所存在的语言模式的习得中提高效率。本研究的结果证明，尽管AWL在机械工程学术英语语料库中显示出了一定的覆盖率和适用性，但机械工程学术词汇表则体现出更高的覆盖率。

机械工程学术词汇表具有十分重要的教学价值。鉴于同一词汇在普通文章与专业文本中的意义有可能截然不同，而词汇学习与具体语境密不可分，词汇教学也应该置于语境当中。基于语料库生成的学术英语词汇表以大量真实的英语语料为基础，编制方法科学合理，能指出应该将哪些词汇包含在学术英语（EAP）教学中。同时，教师还可以充分利用语料库作为教学文本和编写练习的重要资源，将词汇还原到语料库中进行教学加工（梁红梅、何安平 2012），即从语料库中取材与教学文本相配套的语例，设置教学练习或评价活动，为学习者创造更多机会去强化、巩固已学知识，理解、使用自然语言，使词汇教学落实到深层知识的理解与应用之上，达到语言教学的最终目的。

创建基于语料库的机械学术词汇表对于整个机械行业颇具专业意义。机械英语语料库的创建在填补国内外空白、丰富和完善语料库类型的同时，也能够以专业知识内容为依托，探索语言特点和规律，进而推动语料库语言学在机械专业学术英语词汇、短语及语篇研究方面的发展。通过有针对性地训练机械专业学习者和研究人员逐步掌握专业领域文献的程序和特点，即话语或语篇的图式结构，可以从根本上助力科研和从业人员融入专业领域的话语共同体，积极参与国际学术交流，最终辅助我国机械行业摆脱目前面临的人才和技术上的英语语言瓶颈问题，更好地向“工业智造”和“工业4.0”迈进。

机械工程学术论文中涉及大量的公式以及数学和物理符号，目前的语料库软件还无法读取这些符号。因此，当后续研究者从语料库中选取语例进行教学加工时，也往往会刻意回避由于公式或符号被删除而变得不完整的语例。这无疑成为本语料库创建过程中的一件憾事，但同时也给语料库技术的未来发展指明了突破方向。

参考文献

COXHEAD A. A new academic word list [J]. TESOL Quarterly, 2000, 34(2): 213-238.

COXHEAD A, BYRD P. Preparing writing teachers to teach the vocabulary and grammar of academic prose [J]. Journal of Second Language Writing, 2007, 16 (3): 129-147.

COXHEAD A, NATION I S P. The specialized vocabulary of English for academic purposes [C]// FLOWERDEW J, PEACOCK M. Research Perspectives on English for Academic Purposes. Cambridge: Cambridge University Press, 2001: 252-267.

GARDNER D, DAVIES M. A new academic vocabulary list [J]. Applied Linguistics, 2014, 35(3): 305-327.

HSU W. Measuring the vocabulary load of engineering textbooks for EFL undergraduates [J]. English for Specific Purposes, 2014, 33(1): 54-65.

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

HYLAND K, TSE P. Is there an "academic vocabulary"? [J]. TESOL Quarterly, 2007, 41(2): 235-254.

MARTINEZ I A, BECK S C, PANZA C B. Academic vocabulary in agriculture research articles [J]. English for Specific Purposes, 2009, 28(3): 183-198.

MASRAI A, MILTON J. Measuring the contribution of academic and general vocabulary knowledge to learners' academic achievement [J]. Journal of English for Academic Purposes, 2018, 31: 44-57.

MUDRAYA O. Engineering English: a lexical frequency instructional model [J]. English for Specific Purposes, 2006, 25(2): 235-256.

NATION I S P. Learning vocabulary in another language (2nd edition) [M]. Cambridge: Cambridge University Press, 2013.

NATION I S P. Making and using word lists for language learning and testing [M]. Amsterdam: John Benjamins, 2016.

NATION I S P, HEATLEY A. Range and frequency programs [CP/OL]. 2002. http://www.vuw.ac.nz/lals/staff/ paul-nation.

SCHMITT N, COBB T, HORST M, et al. How much vocabulary is needed to use English? Replication of van Zeeland & Schmitt (2012), Nation (2006) and Cobb (2007) [J]. Language Teaching, 2017, 50(2): 212-226.

VALIPOURI L, NASSAJI H. A corpus-based study of academic vocabulary in chemistry research articles [J]. Journal of English for Academic Purposes, 2013, 12(4): 248-263.

VEENSTRA J, SATO Y. Creating an institution-specific science and engineering academic word list for university students [J]. The Journal of Asia TEFL, 2018, 15(1): 148-166.

WARD J. A basic engineering English word list for less proficient foundation engineering undergraduates [J]. English for specific purposes, 2009, 28(3): 170-182.

WEST M. A general service list of English words [M]. London: Longman, 1953.

桂诗春. 多视角下的英语教学 [M]. 上海：上海外语教育出版社，2013.

韩丽娜，刘佳. 国外学术英语词汇研究视角分析 [J]. 中共郑州市委党校学报，2014（5）：106-108.

江淑娟. Coxhead学术词汇表在石油专业英语教学中的适用性研究 [J]. 西南石油大学学报，2009（1）：53.

江淑娟. 石油英语学术词汇表创建研究 [J]. 中国石油大学学报（社会科学版），2010（6）：27-30.

梁红梅，何安平. 语料库的"教学加工"与教材编写 [J]. 当代外语研究，2012（10）：

36-40.

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

刘迪麟，雷蕾. 学术词表研究综述[J]. 外语教学，2020（2）：34-38.

刘佳，韩丽娜. 基于语料库的通用学术词表在专业英语学习中的适用性研究——以“学术词表”和环境科学专业为例[J]. 北京化工大学学报（社会科学版），2014（1）：64-68.

吴瑾，王同顺. Coxhead“学术词汇表”的适用性研究[J]. 国外外语教学，2007（2）：28-33.

赵志刚. 专门用途英语学术词表创建研究——以航海英语为例[J]. 重庆交通大学学报（社会科学版），2015（6）：140-144.

通信地址：110159　辽宁省沈阳市　沈阳理工大学外国语学院（常乐）

110168　辽宁省沈阳市　沈阳建筑大学外国语学院（吴明海、陈颖）

English abstracts

A contrastive analysis of the prototypical meaning and construction degree of subordinate *ba*-constructions

This corpus-based study compares the prototypical meanings and construction degrees of subordinate *ba*-constructions and discusses the semantic structure of Chinese *ba*-sentences from the perspective of Construction Grammar. The results show that lower constructions inherit the prototypical caused-result sense of the upper *ba*-sentence. Constructions indicating a higher salience of caused-result also have a higher construction degree, appearing in the center of a *ba*-sentence. Chinese *ba*-sentences have two prototypes, caused-motion and caused-change, which give rise to two relatively independent expansion paths among different subordinate constructions.

A collocational analysis of English near-synonyms based on the motion chart

Based on the GloWbE corpus and the linguistic motion chart, this study explores the usage and distribution of nominal collocates that co-occur with “quick” and “fast” in 20 varieties of English. The results show a clear tendency for the right-side collocates of “quick” to be used with abstract nouns, such as “look,” “fix,” “access,” and “response,” that cover a wider range and are mostly used to describe economic, technological, and social development, while “fast” is often used with more concrete nouns, such as “food,” “bowler,” “bowling,” and “track,” that mostly pertain to people’s daily lives. In addition, the usage of these two words shows distinct socio-cultural and historical-geographical influences in different countries and regions.

Engagement markers in English popular science texts

Popular science texts include two types of information, namely: scientific knowledge and relevant background information, with the exposition of scientific knowledge as the primary writing purpose. Based on corpus analysis, this study investigated the frequency, pragmatic functions, and use characteristics of engagement markers in the odd (scientific knowledge) and even chapters (background information) of a popular science book, *Prime Obsession: Bernhard Riemann and the Greatest Unsolved Problem in Mathematics*. The engagement marker most frequently

used in this book is direct reader reference, indicating that explicitness is preferred in popular science discourse. The odd chapters use significantly more engagement markers (particularly direct reader reference and questions), indicating that compared with background information, scientific knowledge expositions in popular science texts expect the reader's greater engagement, participation, and comprehension. We also found that the co-text of several engagement markers can enhance their strength.

Modality and interpersonal meanings of primary school teachers' classroom discourse

Based on modality theories in Systemic Functional Linguistics in combination with related Chinese modality research, this paper investigates the modality characteristics of primary school teachers' classroom discourse along the dimensions of modality type, modality value, modality metaphor, and their interpersonal meanings based on speech functions. The result shows that (1) obligation is the most frequently used modality and that modality clauses mainly serve the speech roles of question and command. (2) Teachers realize euphemistic proposals via low-value obligation, modularized interrogative of pure ability modality, and metaphors of explicit subjective orientation. (3) Teachers tend to use high-value probability in order to balance negotiability and guidance on proposition judgments. (4) Significant differences were found in modality use across teachers.

A study of the self-constructed national image about poverty alleviation in *China Daily*

Based on a corpus of news reports on "poverty alleviation" in *China Daily* during the 13th Five-Year-Plan period, this study combines critical metaphor analysis with national image construction to analyze the types, distribution, and self-constructed national image of the conceptual metaphors in "poverty alleviation" news reports, finding the following: First, there are seven types of conceptual metaphors, including conflict, journey, human body, building, plant, book, and circle metaphors. Second, the distribution of conceptual metaphors is imbalanced. Conflict metaphors with the highest resonance of source domain account for the largest proportion of total resonance, followed by journey metaphors, while human body, building, plant, book, and circle metaphors account for less than 10%. Third, these metaphors have created a positive national image of responsibility, steady development, sustainable growth, and harmonious cooperation. This study provides valuable reference data for Chinese media to establish our national image, spread China's voice, and enhance the construction of soft power.

The factors of register and corpus size on semantic prosody research

LI Zhongzheng (69)

In semantic prosody studies, some researchers have long failed to distinguish registers and neglected corpus size limits. Taking "entirely" as the node word, this study explores the differences in semantic prosody in seven sub-corpora of COCA of diverse registers and sizes, namely fiction, academic text, newspaper and magazine, spoken text, small sample, medium sample, and large sample. The results show that the semantic prosodies of the node word are significantly different in the four registers of fiction, academic text, newspaper and magazine, and spoken text. In terms of corpus size, the semantic prosody of the node word in the large sample sub-corpus only partially shares the original semantic prosody in COCA, which is also far different from that in small and medium samples. Therefore, researchers should examine semantic prosody in a specific register and avoid the negative effects of limited sample size.

A meta-analysis of the effects of data-driven learning on the learning performance of Chinese foreign language learners

YANG Lingling (85)

This paper reports a meta-analysis of 28 experimental and quasi-experimental investigations of the impact of Data-Driven Learning (DDL) on the foreign language learning performance of 2,365 Chinese students from 2009 to 2019 in 39 samples, finding that DDL has an overall moderate effect on the learning performance of such learners smaller than that reported in international studies. In addition, the moderation analysis shows that the learning stages, language abilities, number of students, and experiment time all have an impact on the effect of Chinese students' use of DDL in foreign language learning, but the mode of DDL implementation, that is, direct versus indirect DDL, has no effect on its effectiveness. This research is significant for optimizing the application of DDL in foreign language teaching practice.

语料库语言学

CORPUS LINGUISTICS

第17辑 2021

要 目

责任编辑：赵 雪
责任校对：孙凤兰
封面设计：锋尚设计

外研社·期刊中心
电话：010-88819267
E-mail：qkzx@fltrp.com
网址：www.bfsujournals.com

记载人类文明
沟通世界文化
www.fltrp.com

北外学术期刊

iResearch 微信公众号

ISBN 978-7-5213-3622-1

定价：35.00元

语料库语言学

外研社