

文章编号: 1003-0077 (2011) 00-0000-00

对外汉语教学领域可比语料库的构建及应用研究 ——以“把”字句的句法语义标注及应用研究为例*

谭晓平

(上海师范大学对外汉语学院, 上海 200234)

摘要: 本文构建了一个 11 万字的包含自然语料、对外汉语教材语料、汉语中介语语料的“把”字句可比语料库, 对 1556 个“把”字句进行了句法语义标注, 探讨了其在对外汉语教学与研究中的应用。数据显示: 汉语教材中表示具体空间转移、描述动作的方式和次数、信息转移、致使类的“把”字句偏多, 但表示抽象空间转移、判断或认同类的“把”字句偏少。学习者对于结果类“把”字句掌握较好; 教材中 S+把+N+V+状态补语、S+把+N₁+V+N₂ 等 5 种结构形式的“把”字句偏多, 但 S+把+N₁+V 为+N₂、S+把+N₁+V 到+N₂ 等 6 种结构形式的“把”字句偏少。学习者须依次加强 S+把+N+V+趋向补语、S+把+N₁+V 到+N₂ 等 17 种结构形式的“把”字句的学习。

关键词: 可比语料库; “把”字句; 对外汉语教学

中图分类号: TP391

文献标识码: A

The Construction and Application of the Comparable Corpus for TCSL: A Case Study of Semantic and Syntactic Annotation and Its Application of Ba-sentence

TAN Xiaoping

(International College of Chinese Studies, Shanghai Normal University, Shanghai 200234, China)

Abstract: This paper builds a comparable corpus of 110000 characters. It contains three types of corpus data which are natural data, textbook corpus of TCSL, Inter-language Corpus. 1556 Ba-sentences have been annotated and its application has been discussed. The data shows that the Ba-sentence of expressing specific space transfer, describing the way and frequency of action, expressing information transfer and causative meaning in the Chinese-texts are more than we need. However, the Ba-sentence of expressing abstract space transfer and judgment in the Chinese-texts are less than we need. The learners are familiar with the ba-sentence of expressing the result meaning. S+ba+N+V+state complement, S+ba+N₁+V+N₂ and another 3 types of Ba-sentence in Chinese-texts are more than we need. However, S+ba+N₁+V+wei+N₂, S+ba+N₁+V+dao+N₂ and another 4 types of ba-sentence in Chinese-texts are less than we need. The learners should pay more attention on S+ba+N+V+directional complement, S+ba+N₁+V+dao+N₂ and another 15 types of Ba-sentence.

Key words: comparable Corpus; Ba-sentence; TCSL

* 收稿日期:

定稿日期:

基金项目: 上海师范大学文科科研项目“面向汉语作为第二语言教学的语法点知识库构建研究”
(A-0230-17-001003)

作者简介: 谭晓平 (1982—), 女, 讲师, 对外汉语教学、语料库语言学。

1 引言

可比语料库(Comparable Corpus)是指内容或形式上具有某些相似属性的语料的集合,可分为多语可比语料库和单语可比语料库。前者多应用于翻译研究、机器翻译、信息检索等,后者常应用于语言对比研究(Rapp et al. 2016)。本文所构建的汉语可比语料库属于后者。

在国外,单语可比语料库应用于:1)翻译研究。英国曼彻斯特大学构建的英语可比语料库(ECC)包括英文译文语料子库和原创英文语料子库。通过该资源可以更好地考察翻译英语的独特性(Laviosa 1998)。2)区域语言对比研究。英国伦敦大学学院构建的国际英语语料库(ICE)采集了美国、英国、印度、新加坡等23个地区的英语语料。基于该语料库可以更好地研究不同地区英语的使用特点(Greenbaum 1991)。3)外语教学与研究。美国密歇根大学英语学术论文语料库(MICUSP)收录了英语母语者撰写的已发表的英语学术论文以及不同母语背景英语学习者撰写的尚未发表的英语学术论文。基于该语料库,可以更好地分析不同母语背景学习者在学术英语写作上的特点,提高学习者学术英语的写作水平(Römer & Swales 2010)。此外,在此类语料库的基础上,学者们还进行了学习者句法复杂度、语言流利性及准确性的研究(Mancilla & Nihat 2017; Güz 2015)。

汉语可比语料库的建设及应用主要集中在:1)翻译研究。西南大学构建的当代汉语翻译小说语料库(CCTFC)收录了译自英、美、法、日等11个国家8种语言的汉语翻译小说(胡显耀 2007)。英国兰卡斯特大学构建的汉语可比语料库由汉语译文语料库(ZCTC)和汉语本族语语料库(LCMC)构成(Xiao 2010)。2)华语对比研究。例如香港理工大学的两岸三地汉语语料库、香港城市大学的五地共时语料库(LIVAC)以及暨南大学的全球华语语料库(刁晏斌 2016; Tsou 2015; 刘华等 2012)。

可见,与国外可比语料库的建设相比,我国尚缺少服务于汉语二语教学的可比语料库资源。现有对比研究主要依赖于:1)通用平衡语料库。如,国家语委现代汉语通用平衡语料库、CCL语料库、BCC语料库(靳光瑾等 2005; 荀恩东等 2016)。2)汉语中介语语料库。如,HSK动态作文语料库、汉字偏误标注的汉语连续性语料库、暨南大学中介语语料库。3)对外汉语教材语料库。如,北京师范大学、中山大学的对外汉语教材语料库(杨丽姣 2015; 周小兵 2017)。此外,部分学者还自建了可用于对比研究的语料库(马明艳 2017; 梁珊珊等 2016; 江新等 2016; 崔希亮 2005)。鉴于此本文以“把”字句为例,探讨对外汉语教学领域可比语料库的构建及应用问题。

2 语料的选取与规模

本文认为面向汉语二语教学的可比语料库应包含自然语料、教材语料、中介语语料,原因在于:教材语料与母语语料的对比,有助于发现教材语料与母语语料之间的差异;中介语语料与自然语料的对比,有助于了解学习者与母语者的差距;中介语与教材语料的对比,有助于探究语言输入与输出的关系。

本研究的自然语料来自于CCL语料库(网络在线版)。为保证语料的平衡性,我们先从CCL语料库中检索出50万条包含“把”字的语料,然后采用等距提取语料的方法,每隔1000条语料提取1条,共获得语料500条。删除其中不是“把”字句的语料,此外,若一条语料中包含多个“把”字句,则分别对其进行标注。最后获得标注语料534条。

教材语料来源于北京师范大学的对外汉语教材语料库。该库包含有141464条对外汉语教材语料及HSK样题文本语料。本文从语料库中提取了3套(17册)对外汉语教材,分别是《新实用汉语课本》(1-6册)、《博雅汉语》(初级汉语起步篇、准中级加速篇、中级冲刺

篇、高级飞翔篇)、《汉语纵横·精读课本》(0-6册)。之所以选择这三套教材,原因有二:一是它们都是经典的对外汉语教材,能反映出教材编写的实际情况;二是它们属于精读类教材,与口语或报刊类教材相比,与自然语料进行对比研究时,更具可比性。

中介语语料来源于 HSK 动态作文语料库。为了保证三类语料规模上的一致性,我们从语料库中随机提取了 500 条包含“把”字句的语料。三类语料的规模如表 1 所示。

表 1. 可比语料库的规模及构成

单位 \ 类型	自然语料	教材语料	中介语语料	总计
字符数(个)	53146	26409	32958	112513
语料数(条)	534	522	500	1556

需要指出的是,这三类语料的相似或相关性表现在:均为汉语语料,规模相当,都包含有“把”字句,并进行了统一的标注加工。

3 可比语料库的标注加工

3.1 可比语料库的标注框架

可比语料库的标注分为外部标注和内部标注两部分,标注框架如表 2 所示。

表 2. 可比语料库的标注框架

语料类型	标注框架	
	外部标注	内部标注
自然语料	语料的类型、来源、体裁、语体	分词与词性标注、语法项目的语义类别、结构形式
教材语料	语料的类型、来源、体裁、语体、课型、使用地区、汉语水平	分词与词性标注、语法项目语义类别、结构形式
中介语语料	语料的类型、来源、学习者的母语、汉语水平	分词与词性标注、语法项目的语义类别、结构形式、使用正确、使用错误

语料类型指标注的语料是自然语料、教材语料,还是中介语语料。来源信息需要指出所标注的语料来源于哪个语料库或网站。体裁包括:叙述文、说明文、议论文、应用文、演讲稿。语体包括:对话体、叙述体。教材语料中的课型包括:综合、口语、听力、阅读、写作、文化。教材语料中的适用地区包括:大陆、港澳台、东南亚、北美洲、南美洲、欧洲、非洲、大洋洲。教材语料和中介语语料中的汉语水平分为:初、中、高三个等级。中介语语料中的使用错误分为:冗余、漏用、错用。需要指出的是,中介语语料中只标注使用正确的语法项目的语义类别及结构形式信息。

语法项目是汉语二语教学中的重要内容,涉及到各类课型,也贯穿于初、中、高三个阶段。本研究重点进行语法项目的标注。语法项目不同,其句法语义的标注框架也不同,本文以“把”字句为例进行说明。

3.2 “把”字句的句法语义标注框架及标注示例

3.2.1 “把”字句的语义标注框架

吕文华(1994)、崔希亮(2005)、吕必松(2010)从汉语作为第二语言教学的角度分别将“把”字句分为 6 类、5 类和 3 类。本文在上述分类的基础上,结合标注实践,形成了“把”字句语义类别的标注框架。与以往分类不同之处在于本研究将“把”字句的语义类别分为三个级别。第一级别:处置类、判断或认同类、致使类。第二级别:将处置类分为强调动作的结果、强调转移、强调动作的方式、次数。第三级别:将强调转移类的“把”字句分

为：具体空间转移、抽象空间转移、时间上的转移、信息转移。释义及示例¹如下所示。

- ①强调动作的结果。例如：请把门关上
- ②具体空间的转移（处所、人物、单位或组织）。例如：把书放在外边|把书给他
- ③抽象空间的转移（范围、领域、目标）。例如：他们把 DNA 比对技术应用到了相亲上
- ④时间上的转移。例如：把参观者带回到那个年代
- ⑤信息的转移。例如，快把这个消息告诉他
- ⑥强调动作的方式、次数等。例如：把价格一次性提高|把这本书读三遍。
- ⑦表示判断或认同。例如：把学生当成孩子
- ⑧表示致使。例如：这事儿把他愁死了。

3.2.2 “把”字句结构形式的标注框架

吕文华（1994）、崔希亮（1995）、李英等（2005）、丁崇明等（2007）分别将“把”字句的结构形式分为 18 类、10 类、16 类、20 类。本文在以上分类的基础上，结合语料标注实践，将“把”字句的结构形式分为 4 大类，40 小类。分类及示例如下所示。

第一类：S+把+N+V+补语

- ①S+把+N+V+趋向补语 例如：把学生的精神力量发动起来
- ②S+把+N+V+结果补语 例如：如何把我们这个农村中的城市建设好？
- ③S+把+N+V+状态补语 例如：把一方土地组合得鲜艳而且丰满
- ④S+把+N+V+数量补语 例如：把他打了一顿
- ⑤S+把+N₁+V+数量补语+N₂ 例如：把他的情况作了一下介绍。
- ⑥S+把+N+V+时量补语 例如：他们把我的护照手续耽误了一个星期。

第二类：S+把+N₁+V+N₂+(来/去)

- ⑦S+把+N₁+V+N₂ 例如：有人把萧军在社会上的所作所为告诉了鲁迅。
- ⑧S+把+N₁+V+N₂+N₃ 例如：把我们公司的客户名单给他一份
- ⑨S+把+N₁+V 在+N₂ 例如：把声学系统应用在军事技术中
- ⑩S+把+N₁+V 到+N₂ 例如：把它推广到欧洲
- ⑪S+把+N₁+V 到+N₂+来/去 例如：把对外开放工作落实到效益上来
- ⑫S+把+N₁+V 为+N₂ 例如：把思想解放作为“总开关”
- ⑬S+把+N₁+V 成+N₂ 例如：把它变成不朽的经典
- ⑭S+把+N₁+V 给+N₂ 例如：把困难留给自己
- ⑮S+把+N₁+V 作/做+N₂ 例如：把它们称为形式主义的作品
- ⑯S+把+N₁+V 向+N₂ 例如：不幸把他推向了嬉皮和崩克的行列
- ⑰S+把+N₁+V 于+N₂ 例如：把受苦的原因归罪于外部困难的阻碍逼迫
- ⑱S+把+N₁+V 往+N₂ 例如：把喜报寄往部队
- ⑲S+把+N₁+V 至+N₂ 例如：把差错及各种污染的可能降至最低限度
- ⑳S+把+N₁+V 过+N₂ 例如：把窗幕拉过一边

第三类：S+把+N₁+V+补语+N₂+(来/去)

- ㉑S+把+N₁+V 进+N₂ 例如：把手伸进衣袋
- ㉒S+把+N₁+V 进+N₂+来/去 例如：把舞台搬进茶馆来
- ㉓S+把+N₁+V 上+N₂ 例如：父母把我抱上场院
- ㉔S+把+N₁+V 下+N₂ 例如：把我送下楼
- ㉕S+把+N₁+V 下+N₂+来/去 例如：把他们推下水去
- ㉖S+把+N₁+V 入+N₂ 例如：把许多成人和孩子引入误区

¹文中示例来自 CCL 语料库。

- ⑳S+把+N₁+V 回+N₂ 例如：把 4 名伤员送回医院
 ㉑S+把+N₁+V 回+N₂+来/去 例如：坚持把冠华拉回车旁去了
 ㉒S+把+N₁+V 出+N₂ 例如：宇航员把一颗卫星释放出航天飞机。
 ㉓S+把+N₁+V 出+N₂+来/去 例如：把我送出门去
 ㉔S+把+N₁+V 满+N₂ 例如：把桌底都摆满酒瓶

第四类：把+N+V+（其他）

- ㉕S+把+N+V+了 例如：把这三个人杀了
 ㉖S+把+N+V 例如：把头发软化
 ㉗S+把+N+AD+V 例如：把东西重叠地往上放
 ㉘S+把+N+V+着+（N） 例如：把椅子靠着墙
 ㉙S+把+N+一+V 例如：大贵把胸脯一拍
 ㉚S+把+N+V 一 V 例如：把这儿的情况在报纸上登一登
 ㉛S+把+N+V 了 V 例如：把请柬撕了撕
 ㉜S+把+N+VV 例如：把这些盆景修整修整
 ㉝S+把+N+V₁+V₂ 例如：我特地把外孙带来见识一下这个庄严的场面

3.2.3 “把”字句的标注示例

可比语料库的外部标注存储为 XML 格式，内部标注采用 EXCEL 表格的形式。图 1 是自然语料中“把”字句句法语义信息的标注示例。

编号	自然语料	语义类别	结构形式
1	管理对象的某些方面相似的模拟模型，然后对模拟模型进行试验，再【把】试验结果类推到实际管理对象上去的管理方法。	③处置-转移-抽象空间	⑩S+把+N ₁ +V到+N ₂
2	流思想和感情，优美的音乐使人们获得艺术的享受，工程技术人员还【把】声学系统应用在工业生产和军事技术中，成为颇为重要的信息之一。	③处置-转移-抽象空间	⑨S+把+N ₁ +V在+N ₂
3	女的……比如说现在咱们现在已经定位好了，这是个女作者。首先要【把】这个定位好，是个女作者。	①处置-结果	②S+把+N+V+结果补语
4	岁月将【把】它磨砺成一个完美的成熟的体系，留给人们的只剩下深深的崇敬和无限...	①处置-结果	⑬S+把+N ₁ +V成+N ₂

图 1. 自然语料中“把”字句的句法语义标注示例

4 可比语料库的数据分析与应用研究

4.1 数据统计与对比分析

基于标注语料的统计研究与对比分析，可比进一步了解教材编排的特点，发现学习者的习得规律。下文以“把”字句的数据分析为例，进行说明。

4.1.1 “把”字句语义类别的数据统计与对比分析

“把”字句各种语义类别在三类语料中的分布如表 3 所示。

表 3. 语义类别的统计

语料类型 \ 语义类别	自然语料		汉语教材语料		中介语语料	
	语料数	百分比	语料数	百分比	正确数	百分比
强调动作的结果	230	43.1%	209	40.0%	102	48.3%
具体空间的转移	150	28.1%	200	38.3%	38	18.0%
抽象空间的转移 (范围、领域、目标)	65	12.2%	19	3.6%	44	20.9%

时间上的转移	3	0.6%	3	0.6%	3	1.4%
信息的转移	5	0.9%	15	2.9%	1	0.5%
描述动作的方式、次数	16	3.0%	29	5.6%	2	0.9%
表示判断或认同	64	12.0%	40	7.7%	21	10.0%
表示致使	1	0.2%	7	1.3%	0	0.0%
语料总数	534	100.0%	522	100.0%	211	100%

需要指出的是在表3中,中介语语料的数据是基于正确语料的统计结果。500条语料中,正确使用了“把”字句的语料共有211条,正确率为42.2%。为了更清楚地显示“把”字句的语义类别在三类语料中的分布情况并进行对比研究,我们先以自然语料的百分比为基准,将统计结果按降序从高到低排列,再将排列后的数据转换为图2所示的折线图。

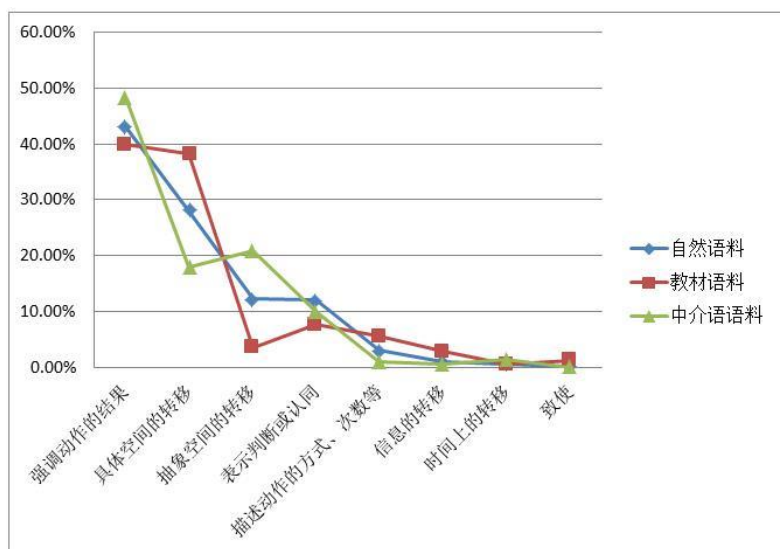


图2. “把”字句语义类别的分布

从上图可以看出: 1)“强调动作结果”的“把”字句在三类语料中的使用率在40%及40%以上,教材设置基本反映了自然语言的实际情况,学习者掌握得也较好。2)表示“具体空间转移”的“把”字句在自然语料、教材语料、中介语语料中的使用率分别为:28.1%、38.3%、18.0%,差别较大。与自然语料相比,教材中该类“把”字句的设置偏多,而学习者使用得较少。3)表示“抽象空间转移”的“把”字句在自然语料、教材语料、中介语语料中的使用率分别为:12.2%、3.6%、20.9%,差别较大。与自然语料相比,教材中该类“把”字句的设置偏少,学习者的使用率却较高。但从教学经验判断,表示“抽象空间转移”的“把”字句的习得难度应大于表示“具体空间转移”的“把”字句的习得难度,但为什么从数据来看“抽象空间转移”比“具体空间转移”的使用率要高呢?本文认为原因有二:一是学习者对“抽象空间转移”的表达需求高于对“具体空间转移”的表达需求;二是在语料中,出现了22条“……放在第一位”的语料,使用语境也基本相似。本文推测,大概是因为研究所用的语料取自HSK动态作文语料库,可能作文题目中出现了这一句型,而考生在写作过程中模仿着使用了这一句型,从而造成了统计结果与教学经验的相背离。4)“表示判断或认同”类的“把”字句在自然语料、教材语料、中介语语料中的使用率分别为:12%、7.7%、10.0%。与自然语料相比,教材中该类“把”字句的设置偏少,中介语语料中该类“把”字句的使用率基本能反映了自然语言的特点。5)“描述动作的方式、次数等”的“把”字句在自然语料、教材语料、中介语语料中的使用率分别为:3%、5.6%、0.9%,与自然语料相比,教材中该类“把”字句在的设置偏多,而学习者的使用率偏低。6)表示“信息转移”的“把”字句在

自然语料、教材语料、中介语语料中的使用率分别为：0.9%、2.9%、0.5%，在自然语料、中介语语料中，该类“把”字句的使用率都在1%以下，使用率低，而在教材语料中使用率偏高，说明教材中该类“把”字句的设置偏多。7)表示“时间上的转移”的“把”字句在自然语料、教材语料、中介语语料中的使用率分别为0.6%、0.6%、1.4%，教材中该类“把”字句的设置反映了自然语言的实际情况，但在中介语语料中的使用率略高，说明学习者有一定的表达需求。8)“致使类”的“把”字句在自然语料、教材语料、中介语语料中的使用率分别为0.2%、1.3%、0%。在自然语料、中介语语料中，该类“把”字句的使用率低，教材语料中使用率较高，说明教材过于强调“致使类”“把”字句的学习。从以上分析可以发现：1)教材设置偏多的“把”字句依次为：表示具体空间转移、描述动作的方式和次数等、信息转移、致使类的“把”字句。教材设置偏少的“把”字句依次为：表示抽象空间的转移、判断或认同类的“把”字句。2)学习者对于结果类的“把”字句掌握得较好，但仍须加强对其他类别“把”字句的学习。

4.1.2 “把”字句结构形式的数据统计与对比分析

“把”字句的各种结构形式在三类语料中的分布如表4所示。需要指出的是，中介语语料的数据是基于正确语料的统计结果。

表4. 结构形式的统计结果

语料类型 结构形式	自然语料		教材语料		中介语语料	
	语料数	百分比	语料数	百分比	语料数	百分比
S+把+N+V+趋向补语	84	15.70%	75	14.40%	21	4.2%
S+把+N+V+结果补语	68	12.70%	66	12.60%	46	9.2%
S+把+N+V+状态补语	16	3.00%	41	7.90%	3	0.6%
S+把+N+V+数量补语	8	1.50%	12	2.30%	2	0.4%
S+把+N ₁ +V+数量补语+N ₂	0	0.00%	1	0.20%	0	0.0%
S+把+N+V+时量补语	0	0.00%	0	0.00%	0	0.0%
S+把+N ₁ +V+N ₂	13	2.40%	25	4.80%	8	1.6%
S+把+N ₁ +V+N ₂ +N ₃	0	0.00%	1	0.20%	0	0.0%
S+把+N ₁ +V 在+N ₂	52	9.70%	62	11.90%	44	8.8%
S+把+N ₁ +V 到+N ₂	41	7.70%	28	5.40%	10	2.0%
S+把+N ₁ +V 到+N ₂ +来/去	6	1.10%	9	1.70%	0	0.0%
S+把+N ₁ +V 为+N ₂	38	7.10%	13	2.50%	7	1.4%
S+把+N ₁ +V 成+N ₂	37	6.90%	36	6.90%	32	6.4%
S+把+N ₁ +V 给+N ₂	28	5.20%	35	6.70%	17	3.4%
S+把+N ₁ +V 作/做+N ₂	19	3.60%	19	3.60%	9	1.8%
S+把+N ₁ +V 向+N ₂	11	2.10%	4	0.80%	2	0.4%
S+把+N ₁ +V 于+N ₂	10	1.90%	4	0.80%	0	0.0%
S+把+N ₁ +V 往+N ₂	2	0.40%	0	0.00%	0	0.0%
S+把+N ₁ +V 至+N ₂	1	0.20%	0	0.00%	0	0.0%
S+把+N ₁ +V 过+N ₂	0	0.00%	1	0.20%	0	0.0%
S+把+N ₁ +V 进+N ₂	12	2.20%	11	2.10%	0	0.0%
S+把+N ₁ +V 进+N ₂ +来/去	0	0.00%	1	0.20%	0	0.0%
S+把+N ₁ +V 上+N ₂	8	1.50%	3	0.60%	0	0.0%
S+把+N ₁ +V 下+N ₂	0	0.00%	2	0.40%	0	0.0%
S+把+N ₁ +V 下+N ₂ +来/去	0	0.00%	1	0.20%	0	0.0%

S+把+N ₁ +V 入+N ₂	7	1.30%	5	1.00%	1	0.2%
S+把+N ₁ +V 回+N ₂	4	0.70%	3	0.60%	0	0.0%
S+把+N ₁ +V 回+N ₂ +来/去	1	0.20%	0	0.00%	0	0.0%
S+把+N ₁ +V 出+N ₂	3	0.60%	2	0.40%	0	0.0%
S+把+N ₁ +V 出+N ₂ +来/去	0	0.00%	1	0.20%	0	0.0%
S+把+N ₁ +V 满+N ₂	0	0.00%	1	0.20%	0	0.0%
S+把+N+V+了	16	3.00%	19	3.60%	4	0.8%
S+把+N+V	12	2.20%	5	1.00%	0	0.0%
S+把+N+AD+V	6	1.10%	12	2.30%	1	0.2%
S+把+N+V+着+(N)	2	0.40%	2	0.40%	0	0.0%
S+把+N+一+V	2	0.40%	0	0.00%	0	0.0%
S+把+N+V 一 V	2	0.40%	2	0.40%	0	0.0%
S+把+N+V 了 V	1	0.20%	0	0.00%	0	0.0%
S+把+N+VV	0	0.00%	1	0.20%	0	0.0%
S+把+N+V ₁ +V ₂	24	4.50%	19	3.60%	4	0.8%
总数	534	100%	522	100%	211	42.2%

为更清楚地显示“把”字句各种结构形式在三类语料中的分布情况进行对比研究，我们先将结构形式分为两类：一是自然语料中使用率在1%以上的结构形式；二是使用率在1%以下的结构形式。然后以自然语料为基准，分别将两类统计结果按降序从高到低排列，最后将数据转换为如图3、图4所示的折线图。

1. 自然语料中使用率在1%以上的结构形式

图3是自然语料中使用率在“1%”以上的结构形式在三类语料中的分布情况。

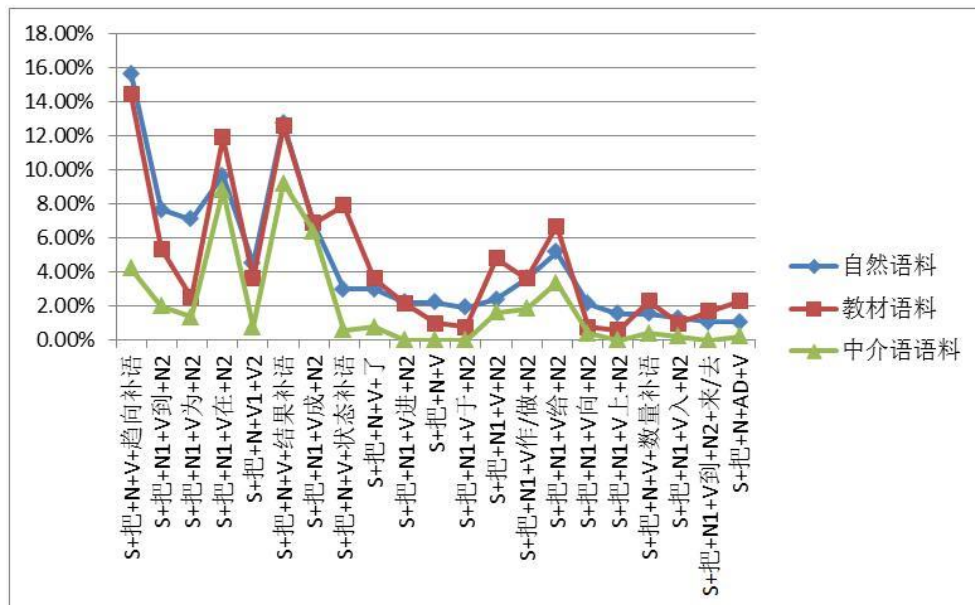


图3. 结构形式的分布 1

从上图可以看出：1)自然语料中使用率在1%以上的结构形式有21种，它们在三类语料中的使用率有较大差异。2)从教材对“把”字句结构形式的设置来看，本文认为如果某个结构形式在教材中的使用率比在自然语料中的使用率多1个百分点，则认为教材对于该结构形式的设置偏离了自然语言的实际情况，也就是说用教材语料的百分比减去自然语料的百分比，

当它们之间的“差”在“1%”以上时，说明教材中该结构形式的设置偏多，而且“差”越大，偏离得越严重。数据显示：在教材中设置偏多的结构形式依次为：S+把+N+V+状态补语（4.90%）²、S+把+N₁+V+N₂（2.40%）、S+把+N₁+V 在+N₂（2.20%）、S+把+N₁+V 给+N₂（1.50%）、S+把+N+AD+V（1.20%）。同样，用教材语料的百分比减去自然语料的百分比，当它们之间的“差”在“-1%”以下时，说明教材设置偏离了自然语言的实际情况，而且“差”越小，偏离得越严重。数据显示：在教材中设置偏少的结构形式依次为：S+把+N₁+V 为+N₂（-4.60%）、S+把+N₁+V 到+N₂（-2.30%）、S+把+N+V+趋向补语（-1.30%）、S+把+N₁+V 向+N₂（-1.30%）、S+把+N+V（-1.20%）、S+把+N₁+V 于+N₂（-1.10%）。3)从中介语语料来看，各结构形式在中介语语料中的使用率均低于自然语料中的使用率。如果某个结构形式在中介语语料中的使用率小于在自然语料中的使用率，则认为需要加强该结构的教与学，也就是说用中介语语料的百分比减去自然语料的百分比，当它们之间的“差”在“-1%”以下时，说明该结构形式是须加强教与学的结构形式，而且“差”越小，说明越重要。统计结果表明须要加强教与学的结构形式依次为：S+把+N+V+趋向补语（-11.50%）、S+把+N₁+V 到+N₂（-5.70%）、S+把+N₁+V 为+N₂（-5.70%）、S+把+N+V₁+V₂（-3.70%）、S+把+N+V+结果补语（-3.50%）、S+把+N+V+状态补语（-2.40%）、S+把+N+V+了（-2.20%）、S+把+N₁+V 进+N₂（-2.20%）、S+把+N+V（-2.20%）、S+把+N₁+V 于+N₂（-1.90%）、S+把+N₁+V 作/做+N₂（-1.80%）、S+把+N₁+V 给+N₂（-1.80%）、S+把+N₁+V 向+N₂（-1.70%）、S+把+N₁+V 上+N₂（-1.50%）、S+把+N+V+数量补语（-1.10%）、S+把+N₁+V 入+N₂（-1.10%）、S+把+N₁+V 到+N₂+来/去（-1.10%）。

2. 自然语料中使用率在 1%以下的结构形式

图 4 是自然语料中使用率在“1%”以下的结构形式在三类语料中的分布情况。

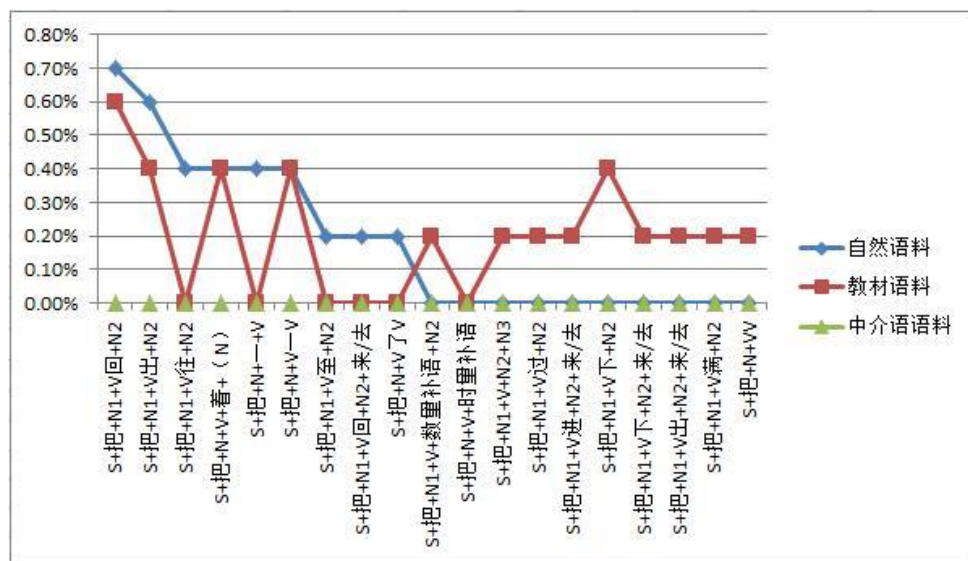


图 4. 结构形式的分布 2

从上图可以看出：1)自然语料中使用率在“1%”以下的结构形式有 19 种，它们属于低频用法，并且在三类语料中的分布差异较大。2)从教材对“把”字句结构形式的设置来看，S+把+N+V+着+(N)、S+把+N+V+一+V这两种结构形式在教材语料和自然语料中的使用率均为 0.4%，S+把+N₁+V+时量补语在教材语料和自然语料中的使用率均为 0%，说明教材中这 3 种结构形式的“把”字句的设置基本反映了自然语言的使用特点。另外，S+把+N₁+V+数量补语+N₂、S+把+N₁+V+N₂+N₃、S+把+N₁+V+过+N₂、S+把+N₁+V+进+N₂+来/去、S+把+N₁+V+下+N₂、S+

² 括号内的百分数是教材语料的百分比减去自然语料的百分比的差。

把+N₁+V 下+N₂+来/去、S+把+N₁+V 出+N₂+来/去、S+把+N₁+V 满+N₂、S+把+N+VV 在自然语料中的使用率为 0，而它们在教材语料中的使用率在 0.2%~0.4%之间，说明教材中这 9 种结构形式的“把”字句设置偏多。此外，S+把+N₁+V 往+N₂、S+把+N+一+V、S+把+N₁+V 至+N₂、S+把+N₁+V 回+N₂+来/去、S+把+N+V 了 V 在自然语料中的使用率在 0.2%~0.4%之间，而教材语料中的使用率均为 0%，说明教材中这 5 种结构形式的“把”字句设置略微偏少。3)从中介语语料来看，这 19 种结构形式的“把”字句的使用率均为 0%，可见学习者对于这 19 种结构形式较为陌生。但因它们在自然语料中的使用率也较低，因此可不作为重点学习的内容。

4.2 可比语料库的实践应用

本文在标注语料的基础上，开发了可比语料的检索工具，检索界面如图 5 所示。教师通过该检索工具可以获取教学所需的例句，比较教材语料及自然语料的差别，预测学习者可能出现的偏误。学习者也可以通过该检索工具选择适合自己难易程度的语料，进行自主学习。

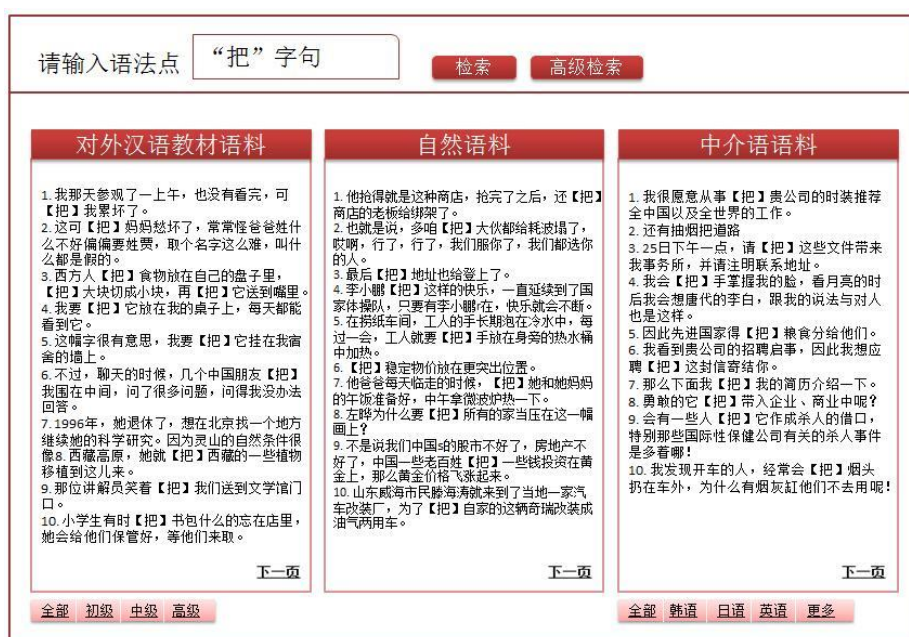


图 5. 可比语料库的检索界面

当用户在文本框中输入“把”字句并点击检索按钮后，可以获得包含该语法项目的教材语料、自然语料和中介语语料。对外汉语教材语料的文本框下设置有全部、初级、中级、高级四个按钮，用户可以根据学习者的汉语水平，选择合适的语料。中介语语料文本框下设置有学习者的母语信息，用户可以根据学习者的母语来查看汉语中介语语料。此外，在高级检索中，用户还可以根据“把”字句的语义信息、结构形式来查询相应语料。

5. 小结

本文以“把”字句为例，探讨了汉语可比语料库的构建及在汉语二语教学中的应用问题。本文构建了一个包含自然语料、对外汉语教材语料、汉语中介语语料的小型“把”字句可比语料库，规模约为 11 万字，制订了“把”字的句法语义标注框架，对语料库中的 1556 个“把”字句进行了句法语义标注，开发了检索界面，探讨了该资源在“把”字句教学与研究中的应用。研究发现：对外汉语教材中表示具体空间转移、描述动作的方式和次数、信息转移、致使类的“把”字句偏多，但表示抽象空间转移、判断或认同类的“把”字句偏少。学习者对于结果类“把”字句掌握较好；教材中 S+把+N+V+状态补语、S+把+N₁+V+N₂ 等 5 种结构形式的“把”字句偏多，但 S+把+N₁+V 为+N₂、S+把+N₁+V 到+N₂ 等 6 种结构形式的“把”字句

偏少。学习者需要依次加强 S+把+N+V+趋向补语、S+把+N₁+V 到+N₂ 等 17 种结构形式的“把”字句的学习。研究认为，汉语可比语料库可为教材研究、教材编排、二语习得研究、课堂教学提供数据支持与语料基础。此外，可比语料库的构建还有利于提高现有资源的利用率，推动资源共享的发展。

在今后的研究中，将进一步扩大语料的规模，从词汇、句子、篇章层面进行对比研究，并进一步开发语料库的检索功能，探讨其在教师备课、课堂教学、数据驱动学习中的应用。

参考文献

- [1] Rapp R, Zweigenbaum P, Sharoff S. Introduction to BUCC 2016[C]//Ninth Workshop on Building and Using Comparable Corpora. xi.
- [2] Laviosa S. The English Comparable Corpus: A resource and a methodology[J]. *Unity in diversity*, 1998: 101-112.
- [3] Greenbaum S. ICE: the International Corpus of English[J]. *English Today*, 1991, 7(4):3-7.
- [4] Römer U, Wulff S. Applying corpus methods to written academic texts: Explorations of MICUSP[J]. *Journal of Writing Research*, 2010, 2(2): 99-127.
- [5] Mancilla, Rae L., Nihat Polat, and Ahmet O. Akcay. "An investigation of native and nonnative English speakers' levels of written syntactic complexity in asynchronous online discussions." *Applied Linguistics* 38.1 (2017): 112-134
- [6] Götz S. Fluency in ENL, ESL and EFL: A corpus-based pilot study[J]. *Proceedings of Disfluency in Spontaneous Speech*, DISS, 2015.
- [7] 胡显耀. 基于语料库的汉语翻译小说词语特征研究[J]. *外语教学与研究*, 2007, (03):214-220+241.
- [8] Xiao, Richard. "How different is translated Chinese from native Chinese?: A corpus-based study of translation universals." *International Journal of Corpus Linguistics* 15.1 (2010): 5-35.
- [9] 刁晏斌. 关于进一步深化两岸四地语言对比研究的思考[J]. *北京师范大学学报(社会科学版)*, 2016, (02):68-75.
- [10] Tsou B K. Augmented Comparative Corpora and Monitoring Corpus in Chinese: LIVAC and Sketch Search Engine Compared[J]. *ACL-IJCNLP 2015*, 2015: 1.
- [11] 刘华, 郭熙. 海外华语语言生活状况调查及华语多媒体语言资源库建设[J]. *语言文字应用*, 2012, (04):125-133.
- [12] 靳光瑾, 肖航等. 现代汉语语料库建设及深加工[J]. *语言文字应用*, 2005, 02:111-120.
- [13] 荀恩东, 饶高琦等. 大数据背景下 BCC 语料库的研制[J]. *语料库语言学*, 2016, 01:93-109+118.
- [14] 杨丽姣, 肖航. 面向语义搜索的语料库语境信息标注研究[J]. *语言文字应用*, 2015, (01):107-116.
- [15] 周小兵, 薄巍等. 国际汉语教材语料库的建设与应用[J]. *语言文字应用*, 2017, (01):125-135.
- [16] 马明艳. 汉语学习者书面语作文“口语化”倾向的语体表征[J]. *汉语学习*, 2017, (01):81-90.
- [17] 梁珊珊, 杨峥琳. 韩国学生口语多重因果转折语篇使用情况分析[J]. *世界汉语教学*, 2016, (03):356-367.
- [18] 江新, 房艳霞等. 汉语母语者和第二语言学习者名名组合的理解[J]. *世界汉语教学*, 2016, (02):226-238.
- [19] 崔希亮. 欧美学生汉语介词习得的特点及偏误分析[J]. *世界汉语教学*, 2005, (03):83-95+115-116.
- [20] 吕文华. “把”字句的语义类型[J]. *汉语学习*, 1994, (04):26-28.
- [21] 崔希亮. “把”字句的若干句法语义问题[J]. *世界汉语教学*, 1995, (03):12-21.
- [22] 吕必松. “把”字短语、“把”字句和“把”字句教学[J]. *汉语学习*, 2010, (05):76-82.
- [23] 李英, 邓小宁. “把”字句语法项目的选取与排序研究[J]. *语言教学与研究*, 2005, (03):50-58.
- [24] 丁崇明, 荣晶. 汉语第二语言学习者应学的“把”字句及其变换[J]. *语言文字应用*, 2007, (S1):11-18.

作者联系方式：谭晓平，上海市徐汇区桂林路 100 号上海师范大学对外汉语学院西部校区 5 号楼 108-2 室，200234，18501037881，tanxiaoping@shnu.edu.cn