文章编号：1003-0077（2011）00-0000-00

汉语二语教学领域词义标注语料库的研究及构建*

王敬，杨丽姣，蒋宏飞，苏靖杰，付静玲

（.北京师范大学中文信息处理研究所，北京市 100875）

摘要：汉语二语教学领域，词汇教学在其中占有极为重要的地位，其中多义词又是词汇教学的重点和难点。本研究通过分析三部经典领域词表，选取了 1181 个重点多义词，以《现代汉语词典（第六版）》为标注体系，制定了适合实际标注的多义词标注规范和形式，在 197 册经典汉语二语教材上进行了多义词词义标注，构建了一个规模约 350 万字的面向汉语二语教学领域的词义标注语料库，并在此基础上对 1811 个多义词、4323 个多义词义项进行了计量统计，分析了多义词不同词义的出现情况及其分布规律。为了更好地服务于汉语二语教学，开发了语料库检索系统，设计并实现了了多义词义项的查询功能。

关键词：汉语二语教学；语料库；多义词标注

中图分类号：TP391　　**文献标识码：**A

The Research and Construction of a Word Sense Annotation Corpus for Teaching Chinese as Second Language

Wang Jing[1], Yang Lijiao[1], Su Jingjie[1], Fu Jingling[1]

(1.Beijing Normal University, Beijing, 100875, China)

Abstract: In field of teaching Chinese as a second language, word teaching is very important, in which polysemous word teaching is harder and more important. By researching 3 classical word tables in this field, this paper selected 1181 polysemous words, and employed Modern Chinese Dictionary as the semantic system, and set out a tagging specification and form for practical annotation. The author tagged 1181 words on 197 classical Chinese textbooks and built a word sense annotation corpus including over 3.5 million characters. Based on the corpus, 1811 polysemous words and 4323 word senses was counted for a quantitative analysis of word sense distribution. In order to offer comprehensive reference for this field, a retrieval system of corpus was developed, which can search for polysemous word senses.

Key words: teaching Chinese as a second language；corpus；polysemous words annotation

1 引言

对外汉语教学领域主要包括语音、词汇、语法等方面的教学，词汇教学在其中占有极为重要的地位。李如龙，吴茗（2005）认为学习语言，词汇是基础，词汇体现了语音的结构和变化，组成语句又体现了种种语法关系，词汇教学的效果直接影响着留学生汉语的整体水平。其中多义词的教学又是词汇教学的重点和难点。在自然语言处理领域，如何识别文本中多义词的词义也一直是一个重要课题，词义消歧任务最早与 1950 年作为机器翻译的一个任务被提出。随着语料库语言学的兴起，语料库开始在无论是汉语二语教学领域和自然语言处理领域起到了至关重要的作用，因此需要建立一个高质量的多义词词义标注语料库。

词义标注语料库是指，根据某个词典对多义词各个义项的定义，在真实语料上标注多义词的正确义项。Leech（1993）指出词义标注是最实用的语义标注。词义标注语料库是机器翻译、信息检索等自然语言处理系统的基础性资源，在语言研究、词典编纂等方面也有重要应用。例如，Sinclair 等（1991）提出在 COBUILD 词典编纂中利用词义标注语料库统计得到词义频率信息编排义项。

词义标注语料库已经经过了十几年的建设，无论是英语还是汉语都有了自己的词义标注语料

* **收稿日期：**　　**定稿日期：**

基金项目：国家语委“十二五”科研规划项目“语言资源建设规划研究”（项目号 YB125-124）；国家高技术研究发展计划(863)(NO.2012AA011104)；中国博士后科学基金第 53 批面上资助（一等）（编号：2013M530026）

库。目前已经建设的词义标注语料库主要以采用词义知识库 WordNet 为主，著名的有 SemCor 语料库、SenseVal 语料库和 DSO 语料库等。采用传统语言词典进行词义标注的语料库数量很少，不成规模。

汉语的词义标注语料库建设起步较晚，主要有北京大学汉语词义标注语料库（Chinese Word Sense Tagging Corpus, STC ） 。该语料库由北京大学计算语言学研究所建设，所选语料是 2000 年 1～3 月和 1998 年 1 月的人民日报，共计 642 万字，所用词典是该所开发的《现代汉语语义词典》。该语料库标注了 966 个多义名词和动词的义项。其中名词 794 个、动词 168 个（金澎，2008）；肖航（2009）将新加坡国立大学“华文教材语料库”中的中小学语文教材作为语料库，选择传统语言词典——《现代汉语词典（第五版）》作为词义体系，对该语料库添加词义标记，该语料库总字数约为 200 万字。

目前，国内的面向汉语二语教学的语料库主要集中在中介语语料上，例如，北京语言大学开发的“HSK 动态作文语料库”，中山大学开发的“汉字偏误标注的汉语连续性中介语语料库”、南京大学开发的“外国留学生汉语口语纵向语料库”和“美国学生汉语作文纵向语料库”，中介语语料库主要是对语料进行字、词、句的偏误标注等。

但是国内目前还没有专门的面向汉语二语教学的词义标注语料库，因此本文选取《现代汉语词典》（第六版）为标注词典，《现代汉语规范词典》作为补充，在汉语二语教材语料库上进行词义标注，在标注实践的基础上，制定了一个比较完善的词义标注体系，规范了词义标注标准，并对标注结果进行了数据分析和统计，在此基础上设计了一个多义词词义检索系统。本研究弥补了汉语二语教学领域语料库类型单一的缺陷，并填补了汉语二语教学领域的基于语料库进行词义研究的空白。

2 标注语料及多义词选择

2．1 标注语料

本研究使用北京师范大学中文信息处理研究所开发的汉语国际教育动态语料库[1]中的外汉语教学领域经典教材部分，包括经典教材 58 套，共 189 册，约 350 万字（含字母、数字和汉字），12 万句。图 1 是教材信息库的部分截图：

教材名称	出版信息				学习者信息		教材信息		
	主编	作者	出版时间	出版社	适用年龄	汉语水平	适用课型	教材类型	教材性质
长城汉语·生存交际课本·一级	马箭飞		北京语言大学出版社	2005年第1版		初级	综合	语言技能	通用型教材
长城汉语·生存交际课本·二级	马箭飞		北京语言大学出版社	2005年第1版		初级	综合	语言技能	通用型教材
长城汉语·生存交际课本·三级	马箭飞		北京语言大学出版社	2005年第1版		初级	综合	语言技能	通用型教材
长城汉语·生存交际课本·四级	马箭飞		北京语言大学出版社	2005年第1版		初级	综合	语言技能	通用型教材
长城汉语·生存交际课本·五级	马箭飞		北京语言大学出版社	2005年第1版		初级	综合	语言技能	通用型教材
长城汉语·生存交际课本·六级	马箭飞		北京语言大学出版社	2005年第1版		初级	综合	语言技能	通用型教材
标准中文第一级第一册	课程教材研究所		人民教育出版社	1998年第1版	中小学	初级	综合	语言技能	通用型教材
标准中文第一级第二册	课程教材研究所		人民教育出版社	1998年第1版	中小学	初级	综合	语言技能	通用型教材
标准中文第一级第三册	课程教材研究所		人民教育出版社	1998年第1版	中小学	初级	综合	语言技能	通用型教材
标准中文第二级第一册	课程教材研究所		人民教育出版社	1998年第1版	中小学	中级	综合	语言技能	通用型教材
标准中文第二级第二册	课程教材研究所		人民教育出版社	1998年第1版	中小学	中级	综合	语言技能	通用型教材
标准中文第二级第三册	课程教材研究所		人民教育出版社	1998年第1版	中小学	中级	综合	语言技能	通用型教材
标准中文第三级第一册	课程教材研究所		人民教育出版社	1999年第1版	中小学	高级	综合	语言技能	通用型教材
标准中文第三级第二册	课程教材研究所		人民教育出版社	1999年第1版	中小学	高级	综合	语言技能	通用型教材
标准中文第三级第三册	课程教材研究所		人民教育出版社	1999年第1版	中小学	高级	综合	语言技能	通用型教材
博雅汉语·初级·起步篇Ⅰ	李晓琪	任雪梅、徐晶凝	北京大学出版社	2004年第1版	成人	初级	综合	语言技能	通用型教材
博雅汉语·初级·起步篇Ⅱ	李晓琪	徐晶凝、任雪梅	北京大学出版社	2005年第1版	成人	初级	综合	语言技能	通用型教材
博雅汉语·准中级·加速篇Ⅰ	李晓琪	黄立、钱旭菁	北京大学出版社	2004年第1版	成人	中级	综合	语言技能	通用型教材
博雅汉语·准中级·加速篇Ⅱ	李晓琪	黄立、钱旭菁	北京大学出版社	2005年第1版	成人	中级	综合	语言技能	通用型教材
博雅汉语·中级·冲刺篇Ⅰ	李晓琪	赵延风	北京大学出版社	2005年第1版	成人	中级	综合	语言技能	通用型教材
博雅汉语·中级·冲刺篇Ⅱ	李晓琪	张明莹	北京大学出版社	2006年第1版	成人	中级	综合	语言技能	通用型教材
博雅汉语·高级·飞翔篇Ⅰ	李晓琪	金舒年、陈莉	北京大学出版社	2004年第1版	成人	高级	综合	语言技能	通用型教材

图 1 汉语二语教材语料库教材信息库

动态语料库在采集教材时充分考虑了教材类型、适用水平、出版年代、影响因子等属性特征（杨丽姣，2015）。所选教材分别从出版年份、学习者适用年龄、学习者汉语水平以及教材性质、教材类型、适用课型等方面做了考虑。所有教材均是 1989 年到 2012 年共 21 年间的典型教材，学习者的使用年龄覆盖到儿童、小学、中学和成人，学习者的汉语水平也从

[1]汉语国际教育动态语料库由北京师范大学中文信息处理研究所和汉语文化学院共同建设，主要收录对外汉语教学领域经典教材和新 HSK 样卷文本语料，并提供多层次的语言信息标注，目前规模约 14 万句，240 余万词次。

零基础、初级、中级到高级水平，适用课型包括口语、听力、阅读、写作和综合课型，这些教材大部分是用来教授语言技能的，一小部分是医学汉语和商务汉语。

2．1 多义词的选取

本研究所说的多义词是广义概念下的多义词，即同一词形具有多种词义可能的均视为多义词。从词义标注和词义消歧角度来说，机器无法只从词形上判断两个词形相同的词是否为两个不同的词，这对留学生来说是一样的，因此采用广义多义词概念更适用于词义标注和汉语二语教学。

语料库词义标注根据标注词的不同，可分为部分词词义标注和全词（all-words）词义标注两种类型。本文的研究是部分词词义标注，选取了 1181 个多义词进行标注。这 1181 个多义词来自《新汉语水平考试大纲 HSK 词汇》（2009）、《汉语国际教育用音节汉子词汇等级划分》（2010）、《1700 对近义词语用法对比》，我们将这 1181 个多义词成为多义词词义标注词表（以下简称词表）。

词表中双音节词占多数，共 812 个，单音节词占少数，共 517 个，多音节词最少，只有 8 个。词表中多义词音节数和多义词的 HSK 等级分布如表 1 所示：

表 1 词表多义词音节数和 HSK 词汇等级分布情况

HSK 等级	双音节词	单音节词	多音节词	总计
1 级	14	41	0	55
2 级	18	47	0	65
3 级	66	49	2	117
4 级	135	61	1	197
5 级	222	91	2	315
6 级	296	35	2	333
超纲	50	48	1	99
总计	**801**	**372**	**8**	**1181**

3 标注体系的选择

语料库词义标注在语义体系和词典资源的选择上有多种不同做法，主要包括传统语言词典（例如《辞海》《现代汉语词典》）、语义词典（例如《同义词词林》）、用于信息处理用的词义知识库（例如 WordNet、HowNet）等（肖航，2010）。本文选择在释义方面具有代表性的、使用最为广泛的《现代汉语词典（第六版）》作为词义体系。《现代汉语词典》（以下简称《现汉》）是汉语语言研究、研究教学等使用最为广泛的词典。

《现代汉语词典（第六版）》对词义和语素义进行了区分。图 2 是《现汉》对多义词“白”的释义。图 2 中“白”的义项⑴、⑶、⑹、⑺、⑽和⑾是词义，而⑵、⑷、⑸、⑻和⑼是语素义。符淮青（2004）指出为了分析多义词不同义项的意义，首先要区分词义和语素义，词义能作为词独立运用，语素义只能存在于它所构成的词和固定结构中。词典对词义和语素义的区分，便利了语料库的词义标注和计算机的词义消歧。因为计算机词义消歧的输入（通常情形下）是经过了词语切分，因此真正成为消歧对象的是词义，而语素义则可以在词语或固定组合中自动得到消解（吴云芳、俞士汶，2009）。因此我们在进行多义词词义标注时，只标注词义，不标注语素义。


白 1 bái ❶ 形 像霜或雪的颜色(跟"黑"相对)。❷ 某些白色或近似白色的东西：葱～|茭～|卵～。❸ 形 光亮；明亮：东方发～|大天～日。❹ 清楚；明白；弄明白：真相大～|不～之冤。❺ 没有加上什么东西的；空白：～饭|～开水|～卷|一穷二～。❻ 副 没有效果；徒然：～跑一趟|～费力气。❼ 副 无代价；无报偿：～吃|～给|～看戏。❽ 象征反动：～军|～区。❾ 指丧事：～事。❿ 动 用白眼珠看人，表示轻视或不满：～了他一眼。⓫ (Bái)名 姓。

图 2 《现代汉语词典（第六版）》对多义词"白"的释义

4 多义词义项标注实践

4．1 标注形式

对于一个包含 n 个义项 S 的词 WORD，它在一定的上下文中被标注上词义 WS，其标注形式为：

@DUOYI_WORD/POS#WS（WS = Si、MH、UN）

其中，WORD 代表一个词，@DUOYI 表示这个词需要进行多义词义项标注，POS 是这个词的词性，WS 是这个词在其所在句子中的词义（WS 的取值将在下文说明）词性和词义用#隔开。

肖航（2010）从为语料库标注多义词词义的实践来看，词典普遍存在词义可区分性不足的情况，根据对《现代汉语词典》的分析，他认为词典中多义词的义项之间存在重叠、相离、包含等关系。本文根据该文献，以及在多义词词义试标注过程中的实践，对多义词在词典中的义项和语料中的义项进行了形式化的表示，如下：

假设一个多义词在词典中一共有 n 个义项，则其第 i 个义项就表示为 Si（i=1,2,3...n）。其中，词义 WS 和义项的关系有几种形式：

1、义项 i 可以准确表示词义，则 WS = Si；

2、义项 i 无法准确表示词义，则 WS = MH，其中 MH 的取值有三种：

A 义项 i 和义项 j 的交集表示词义，则 MH = Si+Sj；

B 义项 i 和义项 j 共同表示词义，但是义项 i 包含义项 j，则 MH = Si >> Sj；

C 义项 i 或义项 j 表示词义，则 MH = Si|Sj；

3、无法为语料中的多义词找到合适的义项，则 WS = QS，QS 的取值为：

A 义项 i 释义过窄，则 QS = Si-

B 合适的没有义项，则 QS = addS(n+i) （其中，QS 参照《现代汉语规范词典》进行义项补充）

4、义项 i 是语素义，无需进行标注，则 WS = UN（无需标注）

4. 2 标注方式

多义词标注采用机器和人工结合的方式进行。首先进行机器标注，再进行人工标注。机器标注是对多义词中可以通过词性决定词义的义项进行标注，例如："安定"这个词，凡是词性是形容词的都选择义项①，凡是词性是动词的都选择义项②。

【安定】	①_【形】（生活、形势等）平静正常；稳定：生活～|情绪很～|～的社会秩序。 ②_【动】使安定：～人心。

为了保证多义词词义的标注质量，我们组织了 20 名语言学及相关专业本科生、研究生参与人工标注，标注人员经过培训和试标注检验合格后，进入正式标注。标注结果需要经过 2 次人工校对，存疑处由专家讨论确认。如上文所述的词典中存在的义项重叠、相离、包含等关系，也在标注过程中不断地进行词表修订，以使结果更符合实际情况。具体工作可以分为以下几个步骤：

步骤 1. 对语料进行分词和词性标注；

步骤 2. 对分词和词性标注记性人工校对；

步骤 3. 对可以通过词性决定词义的义项进行机器标注；

步骤 4. 对剩余的多义词进行人工标注；

步骤 5. 对标注了 MH 和 QS 的义项进行讨论并修订词表；

步骤 6. 根据更新的词表修订标注结果；

步骤 7. 初次校对多义词词义标注结果；

步骤 8. 终校标注结果，收集讨论结果，确认标注词表；

为了方便进行人工标注，我们开发了辅助词义标注的工具，该词义标注工具的界面如图 3 所示：

图 3 多义词词义标注工具

该软件将多义词词义标注词表加载于其中，通过点选的方式对每个词进行词义标注。

4．3 标注过程中特殊语言现象的处理

在多义词词义标注过程中，会涉及到很多特殊语言现象，需要有一定的规范对这些现象进行统一处理，才能对多义词进行标注。

1. 多义离合词的处理：多义的离合词在没有分离时和普通的多义词一样进行标注，当离合词分离之后，只对分离出来可以单独成词的那一部分进行标注，不成词的部分不进行标注，另外，可离合的趋向动词在分离后都不进行标注。

表 2 多义离合词的标注规范及示例

离合情况	标注规则	实例
不离合时	正常标注。	1. 父亲/n 是/vl 一/m 个/q 胖子/n ，/w @DUOYI_走/v#① **@DUOYI_过去/vd#[3]**-① 自然/d @DUOYI_要/vu#[2]-

		⑧ 费事/v 些/q 。/w
分离时，两个部分都可单独成词	两个部分单独进行标注。	2. 他/r 上班/v 时/nt 完成/v @DUOYI_工作/n#③ ，/w **@DUOYI_下/v#[1]-⑲了/u @DUOYI_班/n#②** 就/d 闲着/v 。/w
分离时，一个可单独成词，一个不可单独成词	可单独成词的部分，按照该词进行标注，不可单独成词的部分，词性改为语素，并标注 UN。	3. 可/vu @DUOYI_一起/d# ② **@DUOYI_洗/v#[1]-①** @DUOYI_过/u#[3]-① **澡/g** 以后/nd
分离时，两个部分都不单独成词	词性修改为语素 g，并标注 UN。	4. 一些/m 人/n @DUOYI_确实/d#② 替/v @DUOYI_你/r#① **帮/g** 了/u 个/q 大/a **@DUOYI_忙/g#UN**
可离合的趋向动词分离时	两个词都不进行标注。	5. 便/d @DUOYI_回/v#[1]-③ **@DUOYI_过/g#UN** @DUOYI_脸/n#① **@DUOYI_去/g#UN** ，/w 不/d @DUOYI_再/d#① 理会/v 。/w

例 1 中，多义离合词“过去”没有分离，因此按照词典，将此句中的“过去”标注上词义。例 5 中，多义离合词“过去”分离为“过”和“去”，那么此时就将它们的词性标注为“语素 g”，并标注上“UN”表示无需标注。

2. 多义重叠词的处理：一个多义重叠词是否要进行标注，要看这个词的重叠形式是否能拆分成独立的词，能拆分出独立的词就进行标注，不能拆分或拆分后不能独立成词就不进行标注，是否拆分以《现汉》的收词标准进行判断，拆分规则如表 3 所示：

表 3 多义重叠词的标注规则及示例

重叠形式	动词重叠	形容词重叠	量词重叠	数词重叠
AA 式	走/v 走/v	慢慢/d 慢慢儿/d 早早/d 早早儿/d	个/q 个/q	一/m 一/m
AAB 式	跑/v 跑步/v 听/v 听/v 音乐/n	红/a 红/a 的/u 好好/a 的/u 雪白/a 雪白/a 的/u	/	/
ABB 式		香喷喷/a	一/m 个/q 个/q	/
ABAB 式	打扫/v 打扫/v	轻松/a 轻松/a	一/m 个/q 一/m 个/q	/
AABB 式	/	干干净净/a	/	/

5 多义词义项分布情况的统计和分布

按照上述步骤，我们将在 340 万字语料上进行 1181 个多义词的词义标注，并对这 1181 个多义词的义项标注数据进行了统计和分析，希望从中能够挖掘出其中的规律，并解释其中的原因，希望对汉语教学尤其是教材编写提供一定的参考。

1. 多义词义项复现情况统计

词表中共 1181 个多义词，根据《现代汉语词典（第六版）》，共包括 4213 个义项，平均每个义项出现 3.57 次。在所有语料中，待标注多义词共 538159 词次，其中需要标注的多义词共 537493 词次，无需标注的多义词共 666 词次。本文将每个义项出现的次数按照 100 的距离进行了统计，其结果如图 4a)所示。

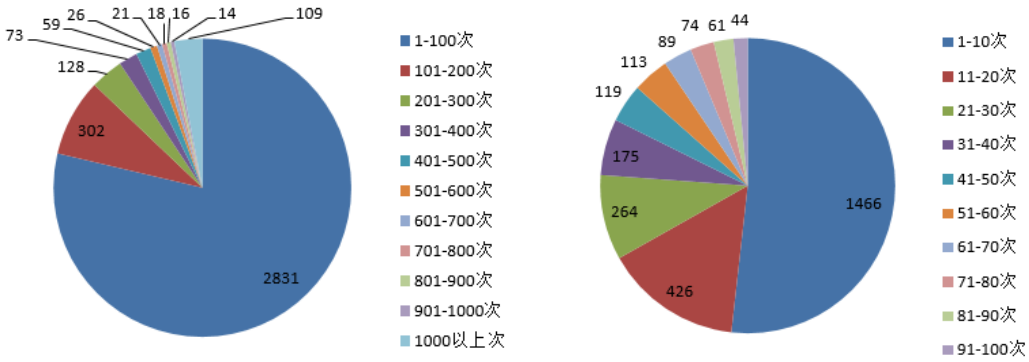

图 4a)　　图 4b)

图 4　多义词义项在语料中的分布情况

从图中可以看出，出现次数在 100 以内的义项是最多的，为 2831 个，共占总义项的 78.70%。其中出现次数排在前 10 位的词的义项分为别：在 1000 次以上的词分别为“在/p#⑦”（20173 次）、“你/r#①”（18191 次）、“有/v#①”（8796 次）、“和/c#[2]-③”（7790 次）、“看/v#[2]-①”（7288 次）、“有/v#②”（6664 次）、“能/vu#④”（6180 次）、“会/vu#[2]-⑤”（5403 次）、“把/p#[1]-⑾”（5124 次）、“什么/r#①”（5079 次），其中，“在/p#⑦”出现次数最多，为 20173 次。

再此基础上，本文又统计了出现次数在 1-100 次之间的义项，并以 10 为距离再次进行了统计，其结果见图 4b)。其结果表示，出现次数在 1-10 的义项最多，一共出现了 1466 次，占总义项的 51.78%，并且随着出现次数的增加，义项逐渐减少。

随后，本文继续统计了出现次数在 1-10 次之间的义项，其结果如图 5 所示。

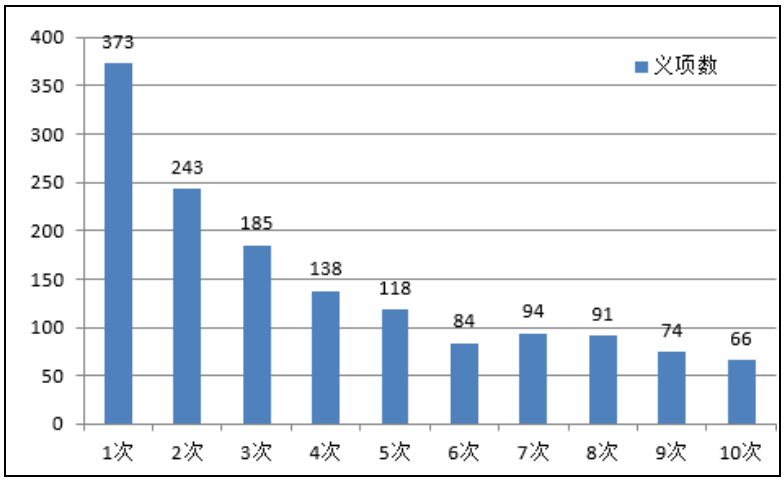

图 5　出现次数小于 10 的义项在语料中的的分布情况

这表明，所有义项中，出现一次的义项最多，为 370 个，占总义项的 25.44%。

2. 高频义项及低频义项的分布统计及分析

根据对多义词在真实语料下的词义情况分析，绝大部份多义词的义项频率分布是不均衡的。具体表现为只有个别义项高频，其他义项低频，对义项进行义频的统计分析，可以反映出教材中选择词义的状况，结合新汉语水平考试（HSK）词汇等级大纲中的词汇等级，可以为以后的编写者提供客观的数据基础。以下的数据，本文只统计了 773 个复现次数在 100 次以上的义项。高频义项及低频义项的统计如表 4 所示：

表 4 语料中高频及低频义项分布情况

频率阈值	总义项数	义项总占比	1 级义项数	2 级义项数	3 级义项数	4 级义项数	5 级义项数	6 级义项数	超纲义项数
=100.00%	35	4.53%	3	1	7	12	10	0	2
>=90%	189	24.45%	16	21	41	59	42	4	6
>=80%	274	35.45%	25	30	64	77	63	5	10
>=70%	348	**45.02%**	33	38	75	**105**	78	5	14
<=30%	220	28.46%	83	52	41	19	5	0	20
<=20%	150	19.40%	70	36	24	7	0	0	13
<=10%	90	11.64%	48	22	14	1	0	0	5
=0	1367	/	87	113	118	227	364	312	146

从上表可以看出，当高频阈值设置为 70%时，有 45.02%的多义词义项是高频义项，并且当词汇的等级是 4 级时，高频义项出现的最多。通过低频义项的分布可以看出，无论低频阈值设置为 30%、20%还是 10%，词汇等级是 1 级的词汇所包含的低频义项最多。

多义词义项频率分布的两个极端情况是：①一个多义词有多个词典义项但除一个高频义项外其他义项不出现；②一个多义词中个别义项不会出现在语料中。

对于第一种情况，通过调查发现，共有 35 个多义词只在语料中出现了一个义项，其他义项没有出现，并且这 35 个多义词只包括 2-3 个义项，除了出现的那个义项，剩下的 1-2 个义项基本都是不常见义项。

在所有的义项中，一共有 1367 个义项没用出现在语料中，涉及到 699 个多义词，这 1367 个义项所属词汇的 HSK 等级分布如表 5 所示：

表 5 词表中多义词的 HSK 等级分布

HSK 等级	双音节词	单音节词	多音节词	总计
1 级	14	41		55
2 级	18	47		65
3 级	66	49	2	117
4 级	135	61	1	197
5 级	222	91	2	315
6 级	296	35	2	333
超纲	50	48	1	99
总计	**801**	**372**	**8**	**1181**

通过分析语料，发现造成这种现象主要有以下几点原因：

1）该义项为其所在词的不常用义项，比如“扒”的第[2]-④个义项：烹调方法，现将原料煮到半熟，再用油炸，最后用文火煮烂：～羊肉|～白菜。这些义项主要是方言义、文言文义等，通过统计，这 1381 个义项中，共有方言义 63 个，文言义 24 个，口语义 11 个，旧时义 12 个，少数民族用语 1 个，用作姓氏 180 个，统计 291 个；

2）多义词义项释义太窄，导致该义项的适用范围太小，比如：“编辑”的第二个和第三个义项：“②_【名】做编辑工作的人”；“③_【名】新闻出版机构中编辑人员的中级专业职称。”第三个义项就是将义项限定的太窄，导致语料中出现“编辑/n”90%都都会选择义项②，只有在上下文很明确的条件下，才会选择义项③；

3）语料的限制，因为语料是来自汉语二语教学的教材，有一部分语料是面向初中级的留学生，一小部分是面向高级留学生，所以语料中出现的多义词的义项总是会集中在一些基本义或词的最常用义。例如， 义项“把/p#[1]-⑴”就一共出现了 5000 多次，而其他词义基本上很少出现

或不出现。这在一定程度上也显示了教材选词和词义的局限性，以及程度不均的词义复现率。

6 多义词义项查询功能的实现

通过系统的、大规模的语料标注实践，我们构建了一个面向汉语二语教学的词义标注语料库。资源的开发需要面向实际的需求，为了更好地服务于汉语二语教学及相关的研究工作，我们对该领域的用户需求进行了分析，开发了语料库检索系统，并依此设计并实现了了多义词义项的查询功能。图 6 不完全显示了多义词词表，通过词表可以查询某个多义词的义项。如图 7 所示，在检索框中输入“阿姨”，会显示出“阿姨”这个词所包含的义项。点击“阿姨”的义项 1，将会显示出义项 1 所包含的所有语料，如图 8 所示。

序号	多义词	序号	多义词	序号	多义词	序号	多义词
1	阿姨	301	干扰	601	麻	901	万万
2	挨	302	干涉	602	麻痹	902	往
3	矮	303	刚	603	麻烦	903	妄想
4	安定	304	高	604	麻木	904	忘记
5	安静	305	高潮	605	骂	905	危机
6	安宁	306	高低	606	卖	906	危险
7	安稳	307	高级	607	满足	907	为难
8	暗	308	高明	608	慢	908	维持
9	熬	309	高尚	609	忙	909	尾巴
10	扒	310	高兴	610	茫然	910	未来
11	把	311	搞	611	毛	911	未免
12	把握	312	告别	612	毛病	912	位置
13	把戏	313	搁	613	茂盛	913	味道
14	白	314	格外	614	没	914	喂
15	白白	315	隔离	615	没有	915	温暖
16	摆	316	个人	616	门	916	文化
17	班	317	个体	617	闷	917	文件

图 6 汉语国际教育动态语料库检索系统——多义词检索功能

图 7 多义词“阿姨”的检索结果

图 8 标注了“阿姨”义项 1 的语料详情

如图 8 所示，“阿姨”的义项 1 共包含 13 条语料，在所有包含“阿姨”的中占比 7.34%。当点击第一条语料时，会显示出这条语料的详情，包括该语料的来源、词性标注信息、多义词标注信息，以及该语料所包含的语法点信息。

7 结论

多义词是汉语二语教学领域词汇教学中的重点和难点，本文根据三张经典领域词表，筛选出 1181 个重点多义词，在 197 册经典汉语二语教材上，以《现代汉语词典（第六版）》为标注体系进行了标注，制定了一套多义词标注规范和形式，构建了一个包含约 350 万字的面向汉语二语教学的词义标注语料库。在该语料库上，本文对 1181 个多义词及其 4323 个多义词义项进行了数量统计，分析了多义词义项复现情况、高频及低频义项分布情况及其规律。并在此基础上研发了一个原料库检索系统，实现了多义词词义查询功能。

基于上述研究工作，我们希望从一下几个方面做出尝试，以改进和提升现有的资源，并探索新的应用空间：第一：扩大语料库规模，目前语料库仅包含汉语二语教学领域的教材语料，并未包含真正的母语语料，希望以后的工作中能够加入一些新闻语料、网络语料等，使语料覆盖更广；第二：加入更多的多义词，实现全词标注。因为人力物力有限，目前仅在语料上实现了部分词标注，有很多多义词并没有被标注上；第三：在第二部的基础上，开展多义词词义消歧研究，以节省人力，丰富现有的资源库建设维度和应用空间。

参考文献

[1] Ide, N., Wilks, Y. Making sense about sense. Word Sense Disambiguation. Dordrecht:：Springer，2007，33：47-73.
[2] Leech, G. Corpus annotation schemes. Literary and Linguistic Computing，1993，8(4)：275-281.
[3] Sinclair, J. Corpus, concordance, collocation . Oxford: Oxford University Press，1991.
[4] 符淮青. 现代汉语词汇 (增订本第二版) [J]. 北京: 北京大学，2004： 63.
[5] 国家汉办/孔子学院总部. 新汉语水平考试大纲[M]. 北京：商务印书馆，2009.
[6] 金澎, 吴云芳, 俞士汶. 词义标注语料库建设综述[J]. 中文信息学报，2008，22(03)： 16-23.
[7] 李如龙, 吴茗. 略论对外汉语词汇教学的两个原则[J]. 语言教学与研究，2005，2(41)： 21.

[8] 刘英林, 马箭飞. 汉语国际教育用音节汉字词汇等级划分[M]. 北京：北京语言大学出版社， 2010.
[9] 吴云芳, 俞士汶. 信息处理用词语义项区分的原则和方法[J]. 语言文字应用，2006， 2：126-133.
[10] 肖航, 杨丽姣. 基于词典的语料库词义标注研究[J]. 语言文字应用， 2010， 2：135-141.
[11] 肖航. 基于词典的语料库词义标注[D]. 新加坡：新加坡国立大学， 2009.
[12] 杨寄洲, 贾永芬. 1700 对近义词语用法对比[M]. 北京：北京语言大学出版社，2005.
[13] 中国社会科学院语言研究所词典编辑室.《现代汉语词典》（第 6 版）[M]. 北京：商务印书馆， 2012.

作者简介

1. 王敬（1988——），女，博士，主要研究领域为：中文信息处理，Email：wangjing1204@foxmail.com；

2. 杨丽姣（1973——），女，副教授，主要研究领域为：汉语国际教育、词汇语义学、语料库语言学，对外汉语教学，Email：yanglijiao@bnu.edu.cn；

3. 蒋宏飞（1982——），男，博士后，主要研究领域为：自然语言处理、智能问答、文本挖掘、机器翻译，Email：jianghongfei@dinfo.cn。