古汉语词义标注语料库的构建及应用研究

舒蕾[1,♠] 郭懿鸾[2,♣] 王慧萍[1,♡] 张学涛[1,2,◇] 胡韧奋[1,♠,†*]
[1] 北京师范大学中文信息处理研究所
[2] 北京师范大学人文宗教高等研究院
♠{202021090021, irishu}@mail.bnu.edu.cn
♣gylguoyiluan@163.com ♡18013752306@163.com
◇11112011118@bnu.edu.cn

摘要

古汉语以单音节词为主，其一词多义现象十分突出，这为现代人理解古文含义带来了一定的挑战。为了更好地实现古汉语词义的分析和判别，本研究基于传统辞书和语料库反映的语言事实，设计了针对古汉语多义词的词义划分原则，并对常用古汉语单音节词进行词义级别的知识整理，据此对包含多义词的语料开展词义标注。现有的语料库包含3.87万条标注数据，规模超过117.6万字，丰富了古代汉语领域的语言资源。实验显示，基于该语料库和BERT语言模型，词义判别算法准确率达到80%左右。进一步地，本文以词义历时演变分析和义族归纳为案例，初步探索了语料库与词义消歧技术在语言本体研究和词典编撰等领域的应用。

关键词： 古代汉语；语料库；词义标注；词义消歧

The Construction and Application of Ancient Chinese Corpus with Word Sense Annotation

Lei Shu[1] Yiluan Guo[2] Huiping Wang[1] Xuetao Zhang[1,2] Renfen Hu[1,†]
[1] Institute of Chinese Information Processing, Beijing Normal University
[2] Institute for Advanced Study of the Humanities and Religion, Beijing Normal University

Abstract

In ancient Chinese, monosyllabic words are dominant, and polysemy is very common, which brings a certain challenge for modern people to understand the meaning of ancient Chinese. Based on the linguistic facts reflected in traditional dictionaries and corpora, this paper designs the principles of semantic division of polysemous words in ancient Chinese, and arranges the knowledge of commonly used monosyllabic words in ancient Chinese, so as to annotate the meaning of polysemous words. So far, the corpus contains 38,700 labeled data with a scale of more than 1,176,000 characters, which enriches the language resources in the field of ancient Chinese. Experiments show that the accuracy of automatic word sense disambiguation based on the corpus with the BERT language model achieves about 80%. Furthermore, this paper explores the application of the corpus built and the word sense disambiguation technology in the fields of language ontology research and dictionary compilation by taking the diachronic evolution analysis of word meaning and the induction of sense families as examples.

Keywords: ancient Chinese , corpus , word sense annotation , word sense disambiguation

† Corresponding author.

第二十届中国计算语言学大会论文集，第549页-第563页，呼和浩特，中国，2021年8月13日至15日。

1 引言

词义标注语料库通常需要根据某个词典对多义词各个义项的定义，在真实的语料上标注多义词的准确义项(金澎等, 2008)。英语词义标注语料库的研究起步较早，由英国Sussex大学主办的SENSEVAL英语词义消歧评测推动了该领域的研究。英语词义标注语料库有基于词典义项的SENSEVAL-1语料库(Kilgarriff, 1999)和以WordNet为词义系统的Semcor语料库(Miller et al., 1993)、DSO语料库(Ng and Lee, 1996)、SENSEVAL-2语料库(Kilgarriff, 2001; Palmer et al., 2001)，以及结合WordNet和Wordsmyth知识库的SENSEVAL-3语料库(Mihalcea et al., 2004; Snyder and Palmer, 2004)。在SENSEVAL评测中，研究者进一步加入外部知识库，完善了竞赛提供的词义标注集，相关研究如Wu et al. (2004)和Palmer et al. (2007)。作为基础语言资源，词义标注语料库可以服务于有监督的词义消歧，进而为语言理解、机器翻译和词汇学研究提供支持。例如，Chan et al. (2007)利用词义标注语料库建立消歧模型，并应用于机器翻译系统，有效改善了翻译效果。Hu et al. (2019)利用牛津英语词典的例句建立词义标注语料库，并借助BERT语言模型实现了细粒度的历时词义演变分析，从而揭示了义项竞争和合作的规律。

现有的汉语词义标注语料库以现代汉语为主，如北京大学汉语词义标注语料库(STC) (Wu et al., 2006)、台湾"中研院"中文词义标注语料库SSMS (Huang et al., 2005)、新加坡国立大学华文教材词义标注语料库(肖航和杨丽姣, 2010)、汉语二语教学词义标注语料库(王敬等, 2017)等。北京大学的STC语料库基于《现代汉语语义词典》的词义体系，对1998年1月2000年1-3 月的《人民日报》（总计约642万字）进行多义词义项标注，共标注了966个多义名词和动词的义项。截至2005年底，台湾"中研院"词义标注语料库SSMS共包含约2000个现代汉语中频词，共涉及约5900个义项。新加坡国立大学的中小学华文教材词义标注语料库依据《现代汉语词典（第五版）》的词义体系，对新加坡国立大学的中小学华文教材语料库（约200万字）进行词义标记。汉语二语教学词义标注语料库以《现代汉语词典（第六版）》为词义区分体系，对197册汉语二语教材文本中的1181个多义词进行词义标注，构建了约350万字的词义标注语料库。

现代汉语词义标注语料库以词典为基础，对新闻、教材语料开展加工，有了较为充分的积累。与之相比，古汉语语言资源的建设仍然较为薄弱。古汉语以单音节词为主，其一词多义现象十分突出，且在不同历史时期的词义分布状况有较大差异。建设古汉语词义标注语料库不仅有助于研究古代词汇的使用状况，也可作为基础资源服务于词义消歧算法的研究，为古汉语信息处理技术、词汇学本体研究、词典编撰等提供参考。因此，本文选取了古汉语常用词汇，综合经典辞书和语料库实际使用状况对多义词进行义项区分和属性整理，并据此开展词义标注，建成了超过117万字规模的古汉语词义标注语料库[0]。以该库为基础，本文基于BERT语言模型研发了小样本情境下的词义消歧技术，准确率达到80%左右。进一步地，本文以词义历时演变分析和义族归纳为案例，初步探索了语料库与词义消歧技术在语言本体研究和词典编撰领域的应用，以期为自然语言处理技术在古汉语领域的应用，如文白机器翻译、文言文信息抽取、古汉语词汇语法现象研究等提供参考和借鉴。

2 基础词义知识库构建

2.1 选词的原则

本研究的目标词为古汉语常用单音节多义词。综合考虑词频和学术研究需要，筛选出了200个古汉语单音节实词，在后续研究中还将根据研究需要和用户反馈持续补充，进行版本迭代。根据国家语委古代汉语语料库字频表[1]，第一阶段选词有较高的使用频度，如表1所示。在频率排序上，51.5%所选词在古汉语字频表中排名前500，80.5%所选词在古汉语字频表中排名前1000。具体的选词以及频率信息可参见附录A。

2.2 义项的设立

词义知识库构建的关键任务是多义词义项的设立与区分。吴云芳，俞士汶 (2006)讨论了"面向人的"辞书义项和"面向汉语信息处理"的词语义项的区别，认为后者需要充分比较面向人和

[0]本文所构建的古汉语词义标注语料库参见：https://github.com/iris2hu/ancient_chinese_sense_annotation

[1]古汉语字频表：http://corpus.zhonghuayuwen.org/resources.aspx

字频排序	选词数量	累积百分比
前100	22	11.00%
101-200	27	24.50%
201-300	18	33.50%
301-400	18	42.50%
401-500	18	51.50%
501-600	16	59.50%
601-700	12	65.50%
701-800	11	71.00%
801-900	12	77.00%
901-1000	7	80.50%
1000以上	39	100.00%

Table 1: 选词的字频分布

面向机器的词语义项，抽取、概括而成一系列义项区分的原则。肖航和杨丽姣 (2010)提出，词义标注语料库建设主要有两个难点：一是词典词义区分不清晰，可能导致标注时出现两可的情况；二是词典义项不全面，无法包括真实语料中目标词所有的可能的含义。从前人研究可以看出，词义标注语料库中的义项设立，既需要尊重辞书描写，也需要考虑语言事实和后续信息处理加工的需要。同时，值得注意的是，古汉语词汇在数千年的使用中，产生了极为丰富的引申、活用、借用等现象。与现代汉语的词义归纳侧重共时用法有所不同，古汉语的词义描写具有时间跨度大、复杂性高等特点，这也就导致了不同的辞书对同一多义词的义项设立存在较大差异。

以“兴(xīng)”为例，《王力古汉语字典》(王力, 2000)、《汉语大字典》(汉语大字典编辑委员会, 2010)、《辞源》(辞源修订组, 1988)、商务印书馆《古代汉语词典(第2版)》(商务印书馆辞书研究, 2014)对其的义项划分差异较大。其中，《王力古汉语字典》分列4个义项，《辞源》6个，《古代汉语词典》8个，而《汉语大字典》则有14个义项。各词典的义项区分如表2所示。

辞书	“兴(xīng)”之义项
《王力古汉语字典》	1.起，起来。2.创立，创办。3.行动起来；引申指流行。4.兴旺，昌盛。
《辞源》	1.起来。2.兴起，发动。3.举。4.征集。5.兴盛。6.姓。
《古代汉语词典》	1.起来，起身。引申：升起，出现。2.兴起，产生。又：创办，建立。又：倡导。3.发动。4.作。5推举，选拔。6.征集，征敛。7.兴盛，昌盛。8.姓。
《汉语大字典》	1.兴起。2.起身。3. 升起。4. 动；发动。5.办理；创办。6.建立。7.推举；选拔。8.昌盛；繁盛。9.奋发。10.成功。11. 征发。12.流行；时行。13.副词。14.姓。

Table 2: 各辞书对“兴(xīng)”的义项区分

词典对标注质量有着极为重大的影响。词典的选择必须具有专业性、受认可度高、对词语义项描述清晰等特点。《王力古汉语字典》兼具“概括性”和“时代性”，可以直观地解释义项的类聚与引申。《王力古汉语字典.序》中提出字典具有“扩大词义的概括性”和“注意词义的时代性”的特点。就“概括性”而言，王力认为：“一般字典辞书总嫌义项太多，使读者不知所从，其实许多义项都可以合并为一个义项，一个是本义，其余是引申义。本书以近引申义合并，远引申义另列，假借义也另列。这样，义项就大大减少，反而容易懂了”。就“时代性”而言，《王力古汉语字典》在《凡例》指出：本字典的义项按照“本义在前，引申义在后；通用义在前，非通用义在后；实词义在前，虚词义在后；古义在前，后起义在后”的原则排列，体现出较强的时代性和系联性，体现出了义项之间的关系。

而《汉语大字典》具有“粒度细”、“涵盖广”的特点，恰好与《王力古汉语字典》在义项设立的宽严方面形成互补。《汉语大字典·第二版修订说明》称该字典力求“古今兼收、源流并重”，“不仅注重收列常用字的常用义，而且注意考释常用字的生僻义和生僻字的义项……是新中国成立以来形音义收录最完备、规模最大的一部汉语字典。”

结合《王力古汉语字典》和《汉语大字典》构建基础词义知识库，兼顾了“概括性”“时代性”和“涵盖性”，能有效应对古汉语的词义描写时间跨度大、复杂性高等特点，满足词义标注语料库的需要。因此，本文拟以《王力古汉语字典》为基础、《汉语大字典》为补充，对多义词的义项设立进行初步划分。除了基于辞书信息进行义项的设置之外，词义标注语料库还需要从语言事实和信息处理的需求出发，根据语料标注情况对词典义项进行一定程度的增补、删减与合并。

确立上述原则后，本研究首先设计了词义知识库的框架，如表3所示。除了词语和义项的基础属性外，还引入了义族、义项属性等信息，以呈现古汉语词义的类聚、引申和假借等特殊现象。同时，根据标注语料库中的义项出现情况设置了“义项频次”字段，为进一步的义项修订提供参考。

属性字段	字段描述	示例
词语id	以“w”+数字表示	w111
词形	以简体字记录的目标词	望
义项id	以“s”+数字表示	s6
读音	目标词义的读音	wang4
词性	目标词义的词性	〈名〉
义项	对义项的文字说明	名望
义族	义项在王力字典中的义族	5.0（远引申义）
义项属性	本义、引申或假借等	远引申义
例句	王力字典中对应例句	陆逊年幼望轻，恐诸公不服。
频次	语料库中该义项的频次	6
修订者	修订该义项的人员	（姓名）

Table 3: 词义知识库各属性字段

在义项整理的过程中，按照如下步骤进行词义知识库属性填充。第一步，根据《王力古汉语词典》确立基础义项，将词语和义项的属性填入表中。然后，根据词典中的义族信息确立义族编号和义项属性：义族以a.b的形式编号，a对应王力划分的义项，b对应同一义项内的小类。义项属性包括“本义”、“近引申义”、“远引申义”、“假借义”、“后起义”、“晚起义”、“偏僻义”，具体定义如下：

（1）本义：《王力古汉语字典》中的第一个义项
（2）近引申义：与本义合并在同一义项内的为近引申义
（3）远引申义：由本义引申，但列为另一个义项的引申义
（4）假借义：《王力古汉语字典》另列的假借义
（5）后起义：魏晋至唐宋产生的词义
（6）晚起义：元明以后产生的词义
（7）偏僻义：《王力古汉语字典》收录在备考栏中的少见的词义

以“假（jiǎ）”为例，《王力古汉语字典》中义项为：①借。引申为凭借。②暂摄职务为假。引申为非真的，伪的（后起义）。【备考】大。词义知识库与《王力古汉语字典》的对应如表4所示。

最后，词义知识库还需要根据语料的实际标注情况填充义项频次，并据此进行增、删、合并等操作，该步骤的操作方式可参见本文第3节。

2.3 义项整理中特殊语言现象的处理

2.3.1 同形词问题

区分义项时该如何处理同形词？吴云芳，俞士汶 (2006)认为，在面向中文信息处理的现代汉语词义区分体系中，可将同一个词的不同义项与同形异义词放在同一个平面上，而无需严格

《王力古汉语字典》	词义知识库
①借。	1.0（本义）
引申为凭借	1.1（近引申义）
②暂摄职务为假	2.0（远引申义）
引申为非真的，伪的（后起义）	2.1（后起义）
【备考】大	3.0（偏僻义）

Table 4: 词义知识库与王力义项的对应

区分同形和多义。在中文信息处理实践中，区分同形词与区分多义词的实际义项遵循相同的过程，即根据语境选择该词形下的某个含义。然而，在古代汉语中，同形词事实上由不同的古代词形表示，只是受到汉字简化的影响而变成了今天在简化字书写范畴下的古汉语同形词，如“后”（皇后）和“後”（先后），这些同形词不仅在传统辞典中有分立的词条，而且在各词内部也有相对独立的词义引申链条。因此，本文认为，吴云芳，俞士汶 (2006)应用驱动的观点是切实合理的，而本研究针对古汉语语言现象进行处理，也应兼顾同形词不同词形的独立性，在标注形式上有所体现。具体来说，本文通过如下方式进行同形词的义项梳理。以“后”为例，根据辞书记载，“后”这个字形共对应了2个同形词，在字形栏分别用“后1”、“后2”标注，“词语id”栏则用词语序号+字形序号标注。每个不同的“后”各自有本义、引申义，被看作是两个起点不同的引申链，互相之间没有联系，义项编号也各自从s1开始编写。特别地，在同形词各自的义项编号前，由一位数字来区分同形词。这样的标识方法在基于大规模语料库的信息处理实践中也具有一定的灵活度。下表5显示了同形词“后”的标注方法。

词语id	词形	义项id	读音	词性	义项	王力义族
w45-1	后1	1s1	hou4	〈名〉	君主。天子和诸侯都称后	1.0（本义）
w45-1	后1	1s2	hou4	〈名〉	君主的正妻	2.0（远引申义）
w45-2	后2	2s1	hou4	〈动〉	走在后面，迟到	1.0（本义）
w45-2	后2	2s2	hou4	〈名〉	位置在后，与“前”相对	2.0（远引申义）
w45-2	后2	2s3	hou4	〈副〉	时间在后的，与“先”相对	3.0（远引申义）
w45-2	后2	2s4	hou4	〈名〉	后代	4.0（远引申义）

Table 5: 同形词“后”的义项区分

2.3.2 临时用法或通假

张永言 (1982)在《词汇学简论》中认为，词的临时用法是词在个别的特殊的应用场合临时带上的含义，比如“行将就木”中的“木”临时具有了“棺材”意义。词的意义和词的用法存在一定差别，意义是稳定和普遍的，而用法是不稳定的、特殊的。所以我们在面对词义活用、通假和其他临时用法时，应根据它的出现频次判断是否需要设置义项，以确保词义的代表性和典型性。若词的某种临时用法较为常见，则需要为它设立新的义项，来保证词义知识库能涵盖尽可能多的用例。

比如，词语“殆”在《王力古汉语字典》中的义项“通‘怠’，懒惰，疲惫”属于假借义，例句如“学而不思则罔，思而不学则殆”。首先根据《王力古汉语字典》设立该义项，在随后的语料库标注过程中，有12句语料中的目标词“殆”属于该义项，因而确定设立该义项。又如“奇”的活用义“以……为奇，惊异”在《王力古汉语字典》中收录，且在语料库中有可观的频次，如例句“大将军邓骘奇其才，累召不应”，因此设立为义项。另外，我们亦设立了一些辞书未收录的临时用法义项，其考量标准是在语料中的频次。如“城”的活用意义“守城”并未在辞书中列出，但在语料库中的例句“（李）应庚发两路兵城南城”、“丞相尝使籍福请魏其城南田”等均应属于“守城”意义，共约10句语料，因此也设立该义项。

一些特殊的、不常见的临时用法则不收入知识库，例如“及其为天子三公，而立为诸侯贤相，乃始信于异众也”，高诱注“信，知也”，可知“信”在语境中是“知晓”的含义，属随文释义，意义具有临时性，因而不设立义项。又例如：“尚得推贤不失序”中的“得”应为“德”的借字，属名词用法，含义为“德，道德，有德之人”。考虑到“得”、“德”的借用在语料库中较为罕见，所

以不设为新义项。同理，“右”的“通‘侑’，劝食”义，“方”的“通‘谤’，指责别人的过失”义出现在极少量语料中，皆属此类，均不为临时用法新增义项。

2.3.3 专有名词

在实际语料标注中，发现不少词例为专有名词，例如“诵”在句子“冬十一月，遣使册高丽国王诵”中应当被解释为人名；“视”在句子“以真时南北差加减之，为食甚视纬”中属于天文术语；“孰”在句子“上诏王僧辩镇姑孰以御之”中属于地名“姑孰”。绝大部分做专名的用法并未被传统辞书收录，而使用频次却相当可观。为了服务于后续的语言学及信息处理研究，本研究对专有名词单独设立义项编号：s0，并按照表6所示规则标注具体的专有名词类别。

专有名词类型	义项编号	示例
人名	s0-1	[谢]安见其草，辄改之，由是历旬不就。
地名	s0-2	日入[信]陵宾馆静，赠君闲步月明中。
官名	s0-3	湖南观察[使]蔡袭为安南经略招讨[使]。
年号	s0-4	晋安帝元[兴]三年六月丙申，白雀见豫章新淦，获以献。
机构名	s0-5	故不加械，即若系尚[方]，于事为苦。
其他专名	s0-6	[左]传宣二年郑破宋师大棘，杜预曰在襄邑县南。

Table 6: 专名标注示例

在实际的语料标注过程中，共有约1800个例句的目标词被标注为专有名词，接近语料库规模的4.7%。

3 词义标注

完成了基础词义知识库的构建后，本研究依据词义知识，在语料库中标注目标词的义项，并根据标注结果对词义知识库中的义项进行增补、删除、合并等操作。

3.1 语料采样及预处理

从古汉语词义标注语料库的建设需求出发，本研究认为语料选取应符合如下原则：(1) 句子完整、句长适中，以提供较为明确的语境信息；(2) 语料均衡，覆盖了不同时代和文献类型，尽可能体现词义使用和分布状况；(3) 无文本内容之外的特殊符号和标记。根据上述原则，本文将语料采样的范围设定于“语料库在线”古代汉语语料库（国家语委语料库）和CCL古代汉语语料库，二者均为研究者广泛使用的古代汉语语料库，采用简体字加工，具有体量大、收录全、覆盖不同朝代等特点。从上述语料库中抽取含有目标词的句子，每个目标词随机抽取200条语料，并保证其朝代分布的均衡性。随后，去除语料中的特殊标记。

3.2 词义标注实践

根据基础词义知识库，由汉语言文字学、古典文献学专业研究生开展语料标注工作，具体遵循如下步骤。

(1) 标注义项。根据目标词在语境中的含义，从义项表中选择义项编号。对于无法找到对应义项的情况做如下标记：若目标词属于专有名词，则按上文所述专名编号标记；若目标词义属于知识库未收录的义项，则标为“其他”；若根据上下文难以判定义项归属则标为“待定”；若存在句子不完整情形或目标词在该语境中有歧义，则标记为“语料不宜”。

(2) 搜集标注反馈，统计义项频次信息，并结合词典描写调整知识库中的义项列表，对词义知识库中的义项做出新增、删除、合并等操作建议。具体来说，包括如下几种情形：(a) 若语料库中该义项出现至少2次，则在词义知识库中保留该义项。(b) 若义项在语料库中未出现或仅出现1次，参考《汉语大字典》的义项设立和例句情况，如果其为《汉语大字典》独立收录且有例句佐证用法，则保留，否则建议归并或删除：如该罕见义项与其他义项存在较高相似性，则建议归并，否则建议删除取消该义项的设立。(c) 针对标注中发现的“其他”义项，如果为《汉语大字典》收录且具有可观频次，则建议为其新增义项；如果两部辞书均未收录，且仅在在少量语料中出现该意义，则不设立新义项，例如：包含目标词“绝”的一条语料：“乡中少年闻其美，神魂倾动，媪悉绝之。（《聊斋志异》）”，根据文义应当属“拒绝”义，但王力、《汉语大

字典》“绝”字均未有“拒绝”义。考虑到此义项出现情况较少，且不宜和其余义项合并，因而不新立义项。

(3) 针对上述操作中给出的新增、删除、归并等建议，由汉语言文字学、中文信息处理专业教师再次审订后，确认词义知识库的修订。

(4) 根据修订后的词义知识库对语料标注结果进行修订，以确保修订后的词义知识库和语料标注中义项的一致性。同时，将词典中的例句也作为补充加入语料库。

(5) 开展知识库和语料库校对工作，首先由高年级汉语言文字学研究生对语料库中的“待定”、“其他”等条目进行校对，给出合理的标注建议；然后由项目组师生对词义知识库和语料标注结果做再次校对。

4 语料库整体规模和义项分布概览

4.1 整体规模

第一阶段的古汉语词义标注语料库共收录200个单音节多义词，词义知识库中收录的词语义项数量2007个，加上专名义项编号6种，共有2013个义项，平均每词义项数量10个。其中，有5个义项未出现在语料库标注中，这些义项被《王力古汉语字典》或《汉语大字典》认为属于本义，但未列出例句用法，如“尽”的本义“器物中空”。考虑到这些属于本义的义项在引申链的构建中具有较大的意义，因此保留这些低频义项备考。

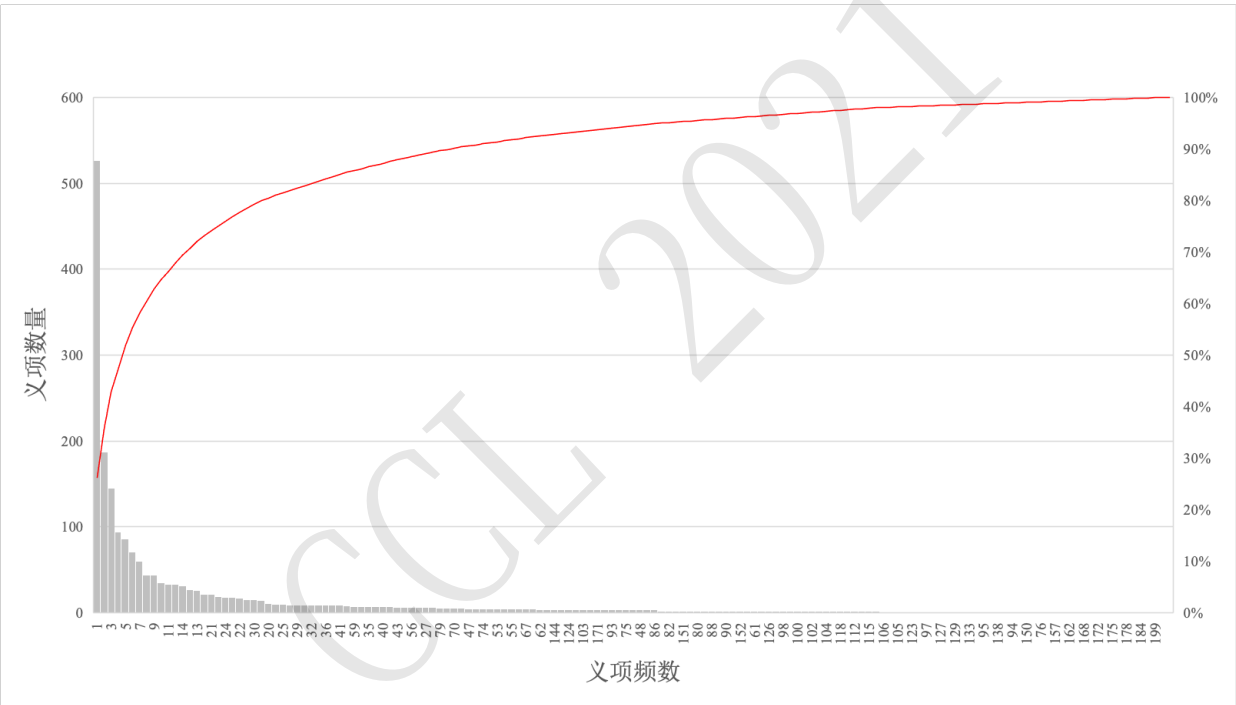

Figure 1: 语料库中的义项频次分布情况

目前，词义标注语料库收录38720条标注数据，总计117.6万字。除专名外，标注语料库中的总义项数量为2002个，每条语料仅对唯一的目标词进行标注。

4.2 义项分布概览

语料库中义项频度信息如图1所示，其中，大量的义项仅出现1次，出现次数在5次及以下的义项占比51.85%，主要原因推测有两方面：(1) 古汉语历时跨度长，不少义项仅在个别或少数朝代使用，整体的频次较低；(2) 在同一个词形下，存在使用优势的义项占据主导地位，使得其他义项比例较低。为了解词义分布的真实情况，仅依靠统计词义语料库中对应目标词的标注结果（约200条/词）是不够的，而我们可通过有监督的词义消歧技术，对大规模语料进行义项标注，从而获得义项真实的分布情况。

5 词义标注语料库的应用

5.1 古汉语词义消歧

依托词义标注语料资源，可以实现有监督的多义词消歧。Hu et al. (2019)以牛津英语词典的例句作为训练语料，将每个义项不多于10条例句作为训练集，通过BERT语言模型获得各个义项的语境向量表示。针对新语料中的目标词，将该词的语境向量与该词形各个义项的向量计算相似度，将相似度最高的义项确定为该句中目标词所属义项。相似地，本研究尝试将义项标注语料库资源划分为训练集和测试集，开展词义消歧实验。

本研究采用的语言模型来自胡韧奋等 (2019)构建的古汉语BERT模型，该模型由总计33亿字的殆知阁古代文献藏书2.0版语料库训练而成。由于训练语料库中繁简体混杂，考虑到繁体转简体的准确率更高，模型研发者将训练语料统一转换为简体。本研究选择该模型进行词义消歧，是因为其训练语料和本研究所使用的语料较为接近，均来自存世古代汉语典籍，且都有朝代跨度广、涵盖文体多的特点。

对目标多义词w_i的消歧过程如下。第一步，获得目标词w_i的所有义项的向量表示：对于特定义项s_j而言，将目标词w_i标注义项为s_j的句子$\{ Sent_1^{w_is_j}, Sent_2^{w_is_j}, \ldots, Sent_n^{w_is_j} \}$输入BERT模型，在最终层获得每个句子中目标词$w_i$的语境向量$e_i$，从而获得目标词$w_i$的义项$s_j$的向量表示集合$\{ e_1^{w_is_j}, e_2^{w_is_j}, \ldots, e_n^{w_is_j} \}$。对该集合中的所有向量取平均值，得到目标词语$w_i$的义项$s_j$的向量表示$e^{w_is_j}$。对目标词的每一个义项重复此步骤，获得该目标词的所有义项的向量表示集合$\{ e^{w_is_1}, e^{w_is_2}, \ldots, e^{w_is_m} \}$。在通过训练集获得目标词的义项表示后，可以对测试集中的多义词进行义项预测。具体来说，将测试集中的句子$Sent_k$输入BERT，获得待消歧的目标词w_i的语境向量$e_k^{w_i}$，将该语境向量与该目标词所有的义项向量逐一计算余弦相似度，取余弦相似度最大的义项s_j为目标词在该句中的义项。

考虑到古汉语词义标注语料库中，每个义项下的例句样本较小，实验设定了2-10共9种阈值，在不同阈值下进行词义消歧实验。阈值表示对于某一个义项，若例句数量超过该阈值，则将其纳入消歧实验。设立不同阈值可以较好地检验和对比小样本情境下消歧方法的效果。当某个义项的例句数量为2、3、4时，实验划分出1条例句作为测试，其余例句归入训练集。当阈值大于等于5时，按照8：2的比例划分训练、测试集。考虑到语料库中约52%的义项只有1-5条例句，这样的划分方法能够较为真实地反映词义消歧模型的效果。

最少训练测试句数	筛选后义项数	训练集、测试集总句数	词义消歧准确率
2	1569	37945	76.36%
3	1367	37541	77.62%
4	1213	37079	78.40%
5	1107	36655	79.77%
6	1018	36210	80.34%
7	942	35754	81.17%
8	876	35292	82.86%
9	834	34956	82.03%
10	786	34524	**84.84**%

Table 7: 词义消歧实验数据

不同阈值下词义消歧的数据划分结果及准确率如表7所示。句子数量阈值为2时，模型达到了高于75%的准确率，而随着阈值的增高，消歧准确率也进一步提高，当训练样本数量达到5（即阈值取6）时，词义消歧准确率达到80%以上。实验结果显示，本研究构建的古汉语词义标注语料库可以作为词义消歧技术的基础语言资源，基于BERT语言模型的小样本词义标注方法达到了一定的准确率。如能进一步有针对性地人工增补例句，确保每条义项的例句数量达到一定阈值以上，该方法将可能取得更好的效果。

接下来，我们对低阈值和高阈值下模型判断错误的数据进行人工分析，归纳总结出两种典型的情况。

典型情况一：阈值的提升纠正了原本判断错误的义项。例句“束书辞东山，改服临北风。”中的目标词“书”正确的义项应为“s4-书籍，装订成册的著作”。在阈值为2时，目标词被模

型自动标注为“s2-文字”，属于标注错误的案例。而当阈值为10时，义项被正确标注了。对此本文认为可能的原因是：阈值较高时，低频义项不参与训练，这减少了目标词在义项消歧时的候选义项数量，增加了消歧准确率。另外，相较于高频义项，低频义项由于参考例句较少，其义项向量难以得到充分的表示。

典型情况二：高阈值时仍然判断错误的义项。目标词“慕”在例句“汤、禹久远兮，邈而不可慕。”中的正确义项为“s2-羡慕”，而模型标注为“s1-思念，依恋”。原因可能是这两个义项本身较为接近，且上下文未提供足够信息。类似的误判有：“九者彼来加我，志在不报。”的“报”本应标为“s1-报答，报酬”，却被模型标为“s7-报复”；例句“子思，字众念，性刚暴，以忠烈自许。元天穆当朝权，以亲从荐为御史中尉。”中的目标词“朝”本应标为“s3-朝廷”，而被模型标为“s8-政事”。

5.2 古汉语历时词义演变

历时词义演变研究依托大规模的历时语料库，旨在还原多义词义项在一段历史时期内频率的变化，发现词语义项产生、消亡和义项之间的竞争等关系(Tahmasebi et al., 2018)。在本研究词义消歧模型获得一定准确率的基础上，可以使该模型自动标注大量历时语料中的目标词词义，从而获得义项的历时分布。

本文以多义词“使”为例，从国家语委古汉语语料库中随机抽取20000条带有时代信息、且包含目标词“使”的语料，以词义标注语料库中所有目标词为“使”的例句作为训练集，建立目标词“使”的词义标注模型。用该模型对20000条带有时代标签的语料进行义项自动标注，梳理各个主要义项的历时分布情况，对曲线进行四次多项式拟合，其结果如图2所示。

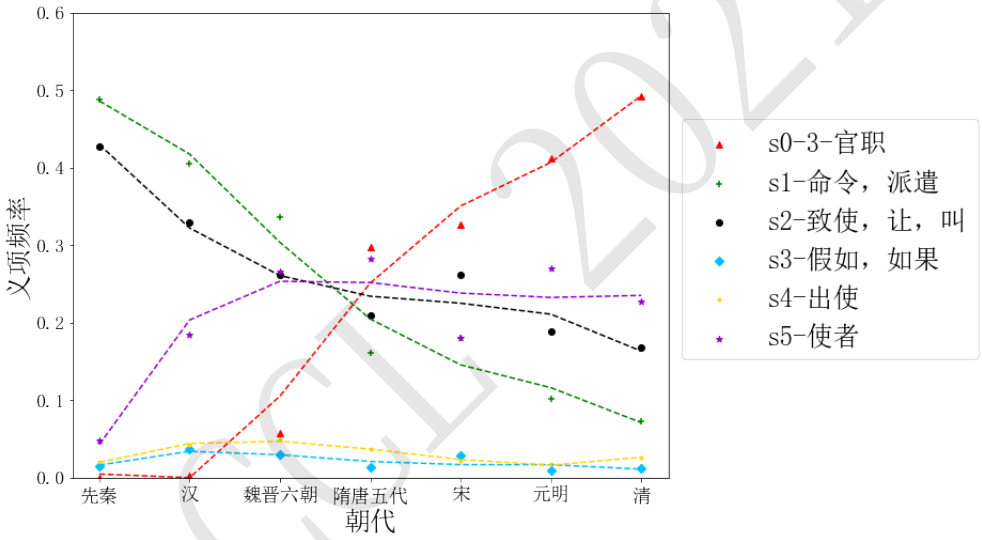

Figure 2: “使”各义项的历时变化趋势

从图中各义项的历时变化趋势可见，“使”作为（君主）使者的含义在先秦即有，而作为官职名称则可能在汉代以后出现，随后激增。到了清朝，“使”作为官职名称成为文献中最常见的义项。相反的，“命令、派遣”和“致使”意义在先秦频率较高，但二者的占比在后期总体呈现下降趋势。

5.3 辅助词典编撰

除了历时词义分析之外，各义项的向量表示也可以作为词典划分义族的参考。本文根据词义标注语料库，使用古汉语BERT语言模型获得了多义词“望”的各义项的向量表示。通过计算各义项向量之间的余弦相似度或采用层次聚类方法，可以获得各义项之间的亲疏关系。层次聚类图中，目标合并的先后顺序标志着所属类别的远近。另外，对词义向量做PCA降维，可以直观地在语义空间图中查看义项之间的位置远近。

以词语“望”为例。在《王力古汉语字典》中，“望”未单列“希图，企图”和“向，对着”义项，这两个义项被《汉语大字典》单列，且在实际标注过程中分别有22和13条例句被标为该义

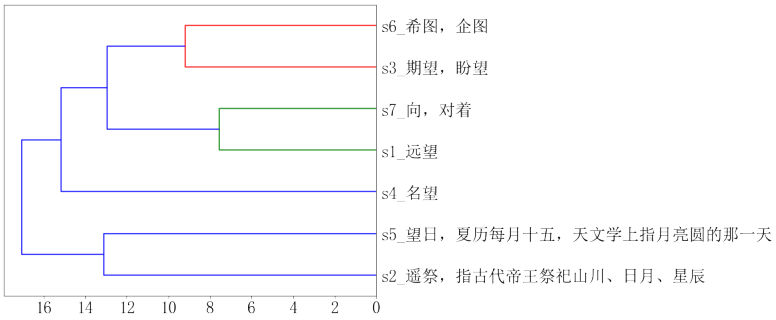

Figure 3: “望”各主要义项的层次聚类情况

项，因而我们的词义知识库的构建过程中新增了这两个义项。为了描述这两个义项与其他义项之间的关系，本文采用层次聚类的方法，以常用的欧式距离作为距离计算公式。如层次聚类图(图3)所示，首先目标义项“向，对着”和“远望”合并、另一个目标义项“希图、企图”和“希望、盼望”合并，接着这两个小类合并后，与“名望”合并，最后，两个边缘义项“望日”和“遥祭”合并后再并入其中。

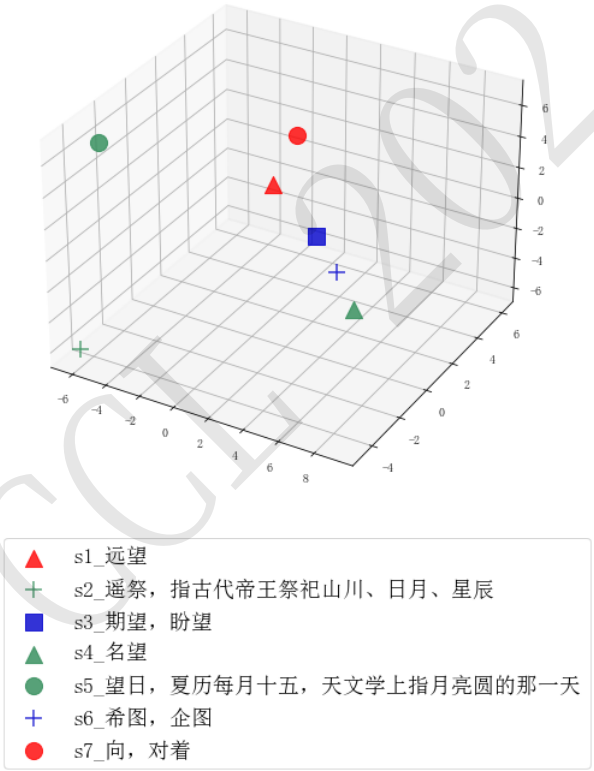

Figure 4: “望”各主要义项向量在降维后的语义空间中的相对位置

进一步地，如图4所示，降维后的语义空间反映了义项向量在三维空间中的相对位置关系，虽然降维过程丢失了高维空间中的一些细节，但是还是可以直观地看到义项“遥祭”和“望日”属于边缘义项，而“远望”和“向、对着”、“期望、盼望”和“希图、企图”之间两两具有紧密联系。因此本文认为“望”的义项“希图、企图”和“向，对着”有可能属于引申义，“希图、企图”与“期望，盼望”义项关系密切、“向，对着”和“远望”义项之间关系密切。考虑到义项“远望”在《王力古汉语字典》被认为是本义，而《汉语大字典》中义项“向，对着”的最早用例来自马王堆帛书的“日月相望”，则推测义项“向，对着”是由本义经过语法化的过程而产生的近引申义。

6 结论

本文以古汉语词义标注语料库为研究对象，基于传统辞书和语料库中的义项频率，设计了古汉语多义词的词义划分原则，以200个常用古汉语单音节多义词为例，构建了词义级别的知识库，并据此对包含多义词的语料开展词义标注。现有的语料库包含3.87万条标注数据，规模超过117万字，丰富了古代汉语领域的语言资源。实验显示，基于该语料库和BERT语言模型，词义消歧算法准确率可达到80%左右。在此基础上，本文介绍了该语言资源在古汉语词义历时演变研究、辅助词典编撰中的应用案例。未来，该资源和相关算法还为文白机器翻译、文言文信息抽取、古汉语词汇语法现象研究等提供参考和借鉴。

值得一提的是，本研究提出的古汉语词义标注语料库依然存在规模较小的问题，为确保提升该资源的应用价值，我们将在未来的研究中对其做进一步的扩充和更新。

致谢

本研究得到国家自然科学基金青年项目“面向古籍整理智能化的知识表示与加工研究”(62006021) 资助。在语料库建设中，北京师范大学曹媛南、段毓赜、何琪怡、黄芷晴、蒋瑞、李涔、李隽琪、罗涵柯、孙雨、杨济清、姚昊辰、张文强、张霄等同学（姓名按音序排列）为义项修订和语料标注工作作出了重要贡献，匿名审稿人对论文提出了宝贵的修改意见，在此一并表示感谢。

参考文献

Adam Kilgarriff. 1999. Senseval: An exercise in evaluating word sense disambiguation programs. 02.

Adam Kilgarriff. 2001. English lexical sample task description. *Proceedings of the senseval-2 workshop*, 10.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain, July. Association for Computational Linguistics.

Chu-Ren Huang, Chun-Ling Chen, Cui-Xia Weng, Hsiang-Ping Lee, Yong-Xiang Chen, and KehjiannChen. 2005. The sinica sense management system: Design and implementation. *International Journal of Computational Linguistics and Chinese Language Processing*, 24(11):503–512.

Dekai Wu, Weifeng Su, and Marine Carpuat. 2004. A kernel pca method for superior word sense disambiguation. In *Proceedings of the 42nd Annual Meeting on Association for Computational Linguistics*, ACL '04, page 637–es, USA. Association for Computational Linguistics.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, HLT'93, page 303–308, USA. Association for Computational Linguistics.

Hwee Tou Ng and Hian Beng Lee. 1996. Integrating multiple knowledge sources to disambiguate word sense: An exemplar-based approach. In *Proceedings of the 34th Annual Meeting on Association for-Computational Linguistics*, ACL'96, page 40–47, USA. Association for Computational Linguistics.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Dang. 2001. English tasks:All-words and verb lexical sample. 01.

Martha Palmer, Hoa Dang, and Christiane Fellbaum. 2007. Making fine-grained and coarse-grained sense distinctions, both manually and automatically. *Natural Language Engineering*, 13:137–163,06.

Nina Tahmasebi, Lars Borin, and Adam Jatowt. 2018. Survey of computational approaches to lexicalsemantic change. *arXiv preprint arXiv:1811.06278.*,

Rada Mihalcea, Timothy Chklovski, and Adam Kilgarriff. 2004. The senseval-3 English lexical sample task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 25–28, Barcelona, Spain, July. Association for Computational Linguistics.

Renfen Hu, Shen Li, and Shichen Liang. 2019. Diachronic sense modeling with deep contextualized word embeddings: An ecological view. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3899–3908,Florence,Italy,July. Association for Computational Linguistics.

Yee Seng Chan, Hwee Tou Ng, and David Chiang. 2007. Word sense disambiguation improves statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 33–40, Prague, Czech Republic, June. Association for Computational Linguistics.

Yunfang Wu, Peng Jin, Yangsen Zhang, and Shiwen Yu. 2006. A Chinese corpus with word sense annotation. In *Computer Processing of Oriental Languages. Beyond the Orient: The Research Challenges Ahead*, pages 414–421, Berlin, Heidelberg. Springer Berlin Heidelberg.

辞源修订组. 1988. 辞源. 北京:商务印书馆.

汉语大字典编辑委员会. 2010. 汉语大字典（第二版）. 武汉:崇文书局, 成都:四川辞书出版社.

胡韧奋, 李绅, 诸雨辰. 2019. 基于深层语言模型的古汉语知识表示及自动断句研究. 第十八届中国计算语言学大会（CCL2019）.

金澎, 吴云芳, 俞士汶. 2008. 词义标注语料库建设综述. 中文信息学报, 22(3):16.

商务印书馆辞书研究. 2014. 古代汉语词典（第2版）. 北京:商务印书馆.

王敬, 杨丽姣, 蒋宏飞, 苏靖杰, 付静玲. 汉语二语教学领域词义标注语料库的研究及构建. 中文信息学报, 31(1):221.

王力. 2000. 王力古汉语字典. 北京:中华书局.

吴云芳, 俞士汶. 2006. 信息处理用词语义项区分的原则和方法. 语言文字应用, (2):126–133, 8.

肖航, 杨丽姣. 2010. 基于词典的语料库词义标注研究. 语言文字应用, (2):135–141, 9.

张永言. 1982. 词汇学简论. 武昌:华中工学院出版社.

附录A.语料库选词的字频分布

词频排序	汉字	频率（%）	词频排序	汉字	频率（%）
5	是	0.9553	488	亡	0.0435
14	得	0.5671	490	乘	0.0434
16	如	0.5643	499	陵	0.0427
17	道	0.5221	510	诚	0.042
25	见	0.414	511	念	0.042
26	说	0.4139	518	穷	0.0415
31	何	0.397	519	卒	0.0414
36	知	0.361	522	固	0.0414
38	王	0.3387	523	存	0.041
41	去	0.3248	526	徒	0.0405
49	行	0.2911	532	任	0.0401
52	便	0.2762	543	休	0.0392
54	将	0.2717	570	加	0.0373
57	相	0.2687	577	工	0.0363
75	若	0.2294	579	备	0.0362
78	能	0.2248	583	贼	0.0357
80	使	0.2218	584	达	0.0356
88	当	0.1933	591	退	0.0353
89	多	0.1868	597	谢	0.0349
91	书	0.1793	602	胡	0.0346
92	从	0.1789	626	期	0.0333
94	长	0.1776	630	属	0.0333
101	方	0.169	648	除	0.032
108	理	0.1649	649	私	0.0319
109	非	0.1645	650	顺	0.0318
115	看	0.161	663	类	0.0312
121	乃	0.1529	678	克	0.0307
122	过	0.1514	689	顾	0.0302
123	诸	0.1507	692	视	0.0301
129	安	0.1455	693	读	0.0299
132	闻	0.143	700	疾	0.0296
138	要	0.1375	701	迁	0.0296
141	后	0.1351	711	池	0.029
145	归	0.1332	719	济	0.0286
148	兵	0.132	732	宾	0.0276
150	尽	0.1296	734	怨	0.0276
157	名	0.1272	738	竟	0.0275
165	复	0.1243	747	遗	0.0271
171	及	0.122	768	具	0.0263
172	阳	0.1216	769	素	0.0263
173	本	0.1216	785	恨	0.0255
175	令	0.1187	795	置	0.0249
179	却	0.1176	809	图	0.0242
182	会	0.1158	822	末	0.0239
183	城	0.1156	823	负	0.0239
186	发	0.1132	835	造	0.0233
193	朝	0.1099	849	怜	0.0229
194	即	0.1095	854	率	0.0226
200	就	0.1065	857	堪	0.0225

203	间	0.1047	858	质	0.0225
207	善	0.1028	873	徐	0.0215
224	数	0.0951	876	奇	0.0213
225	少	0.0951	879	适	0.0212
232	进	0.0911	881	奔	0.0211
233	莫	0.0909	912	卧	0.0203
245	内	0.0881	926	假	0.02
248	坐	0.0861	928	劝	0.0199
256	节	0.0838	932	患	0.0198
261	信	0.0825	950	博	0.0192
270	或	0.0804	960	察	0.0189
271	请	0.0797	962	敌	0.0188
273	传	0.0794	1003	短	0.0177
274	通	0.0788	1025	悉	0.0172
276	既	0.0785	1063	汤	0.0164
292	求	0.0752	1080	夺	0.016
293	居	0.0752	1081	睡	0.016
297	遂	0.0743	1163	族	0.0141
301	望	0.0721	1203	弥	0.0136
305	度	0.0709	1244	拔	0.0127
311	曾	0.0699	1270	暴	0.0123
318	恶	0.0681	1279	潜	0.0122
319	解	0.0681	1332	规	0.0115
326	举	0.0671	1335	旁	0.0115
332	易	0.0645	1344	眠	0.0114
340	阴	0.0631	1377	稍	0.0109
344	许	0.0624	1417	籍	0.0104
349	走	0.0614	1424	戚	0.0103
351	论	0.0608	1469	按	0.0097
361	辞	0.0588	1493	涉	0.0095
365	第	0.0579	1507	孰	0.0093
366	病	0.0578	1538	寝	0.009
372	兴	0.057	1582	判	0.0086
378	爱	0.056	1637	慕	0.008
394	治	0.0535	1654	倍	0.0078
399	胜	0.052	1721	勒	0.0072
404	左	0.0515	1767	笃	0.0069
407	修	0.0511	1805	诵	0.0066
419	临	0.0504	1816	涕	0.0065
421	识	0.0503	1819	鄙	0.0065
440	被	0.0484	1993	熙	0.0052
445	报	0.0474	2079	逾	0.0047
448	致	0.0473	2267	敝	0.0039
449	次	0.0473	2273	殆	0.0039
451	引	0.047	2319	怠	0.0037
454	静	0.0467	2355	寐	0.0036
455	微	0.0465	2629	俾	0.0027
456	右	0.0465	2651	讽	0.0026
465	宜	0.0457	2896	贻	0.0021
476	绝	0.0443	3097	饯	0.0017
481	比	0.0439	3158	弛	0.0016

附录B.词义知识库结构示例

词语id	w1	w1	w1	w1	w1	w1
词形	爱	爱	爱	爱	爱	爱
义项id	s1	s2	s3	s4	s5	s6
读音	ai4	ai4	ai4	ai4	ai4	ai4
词性	〈动〉	〈名、形〉	〈动〉	〈形〉	〈动〉	〈形〉
义项	喜爱，爱好	情爱，仁爱，恩惠；德行美好	爱护，加惠；钦慕；爱戴	吝惜，舍不得	通“[艹爱](ài)”，隐蔽，躲藏	通“暧”，昏暗
王力义族	1.0（本义）	1.1（近引申义）	1.2（近引申义）	2.0（远引申义）	3.0（假借义）	
义项属性	本义	引申义	引申义	引申义	通假义	通假义
示例	淳之少有高尚，爱好坟籍，太原王恭所称。	及子产卒，仲尼闻之，出涕曰：“古之遗爱也。”	父母之爱子，则为之计深远。	百姓皆以王为爱也。	爱而不见，搔首踟躇。	上台爱育通幽细。
频次	88	24	51	19	2	3

注：因版面限制采用转置排版