汉语成语知识库与汉语教学*

王雷[1,2] 俞士汶[1] 朱学锋[1] 罗凤珠[3]
北京大学计算语言学教育部重点实验室[1] 北京大学英语系 北京 100871[2]
台湾元智大学中国语言文学系 32003[3]
Email: {wangleics, yusw}@pku.edu.cn; gefjulo@ saturn.yzu.edu.tw

摘 要：汉语成语一直被誉为“民族语言和文化的瑰宝”，在汉语的词汇体系中占有重要的地位。汉语成语具有数量庞大、结构固定、构词简洁、寓意丰富、表现力强等特点，不但蕴含着中国丰富的社会、历史、文化知识，还具有使语言表达生动、形象鲜明的独特作用。本文介绍了北京大学计算语言学研究所（ICL/PKU）建设的汉语成语知识库（CIKB），包括其收录的条目、相关的属性内容信息，以及基于该知识库利用语言统计方法所发现的知识如何应用在汉语文教学和对外汉语教学领域。

关键词：成语知识库；汉语文教学；对外汉语教学

Chinese Idiom Knowledge Base and Chinese Language Teaching

Wang Lei[1,2] Yu Shiwen[1] Zhu Xuefeng[1] Luo Fengzhu[3]
Key Laboratory of Computational Linguistics of Ministry of Education[1]
Department of English of Peking University Beijing 100871[2]
Department of Chinese Linguistics & Literature of Yuan Ze University Taiwan 32003[3]

Abstract: Chinese idioms are commented as the treasure of national language and culture and serve as a very important of Chinese vocabulary. With a pretty large number and fixed structure, Chinese idioms are semantically concise and pragmatically expressive, embody rich social, historical and culture knowledge, which enable them to present vivid image when in use. This paper introduces the entries and fields selected for the Chinese Idiom Knowledge Base (CIKB) built by the Institute of Computational Linguistics at Peking University (ICL/PKU) as well as its construction methods, research and applications so far.

Key words: Chinese idiom knowledge base; Chinese language teaching; Teaching Chinese as a foreign language

1 引言

成语在语言表达中有生动简洁、 形象鲜明、喻义深刻的特点，本身蕴含着丰富的历史、社会和文化知识，是一个民族语言最具有特色的组成部分。汉语历史悠久，大部分成语是从古代历史典故、寓言传说、经典文献中相承沿用下来的，通常有着几千年的历史，是珍贵的民族文化遗产；此外，汉语成语数量多，使用频率高，这也是汉语不同于其他语言的一个显著特点。在语言教学中，成语教学是不可忽视的重要组成部分，教好、学好成语可以使学生掌握有关成语的社会、历史和文化知识，开阔眼界，提高表达、阅读和写作能力。

在汉语中，成语占有非常重要的地位，研究类似成语、习语、谚语等多词表达（Multi-word Expression）并建设这种语言单位的知识库对于语言教学（Lo 1997）、词典编纂（Fellbaum 2007）、自然语言处理（Lin 1999）等领域的研究和发展会具有实质性的意义。近年来随着中文电化教学理论日益发展，相关实践与方法日益得到推广与普及，大规模、高质量的汉语语言知识库（包括各种形式的语料库）不断开发研制出来并应用于实际语言教学中，这些因素对于推动汉语文教学、对外汉语教学起了非常大的作用。

* 本文相关研究得到国家自然科学基金项目(项目号：61170163)与蒋经国国际学术交流基金会奖助（2009）的支持。

二十多年来，北京大学计算语言学研究所（以下简称“计算语言学所”）一直致力于各种语言知识库的建设，目前已经拥有一系列质量上乘、内容丰富的语言知识库，并经过优化统一整合成为综合型语言知识库（俞士汶等 2011）。对于汉语成语，俞士汶（2004）曾指出：“成语在现代汉语中频繁出现，对成语的理解（包括确切翻译）是文本内容理解的一个重要组成部分。成语庞大，毕竟有限；成语难懂，毕竟可查。只要建设好成语知识库，绝大部分成语的理解问题就会迎刃而解。”正是认识到了文本中成语理解的重要性，他提出了构建成语知识库的设想，并在国家重点基础研究课题（973）“文本内容理解的数据基础”（课题编号：2004CB318102）中实践了这一主张。

2 成语知识库的条目及属性信息

在综合型语言知识库中，《现代汉语语法信息词典》（俞士汶等 2003）（以下简称“《语法信息词典》”）和基本标注语料库最具有基础性和代表性，而基本标注语料库又是在《语法信息词典》的基础上开发的。《语法信息词典》中也包含一个成语库，这表明成语作为汉语语言中非常重要的一部分，被单独搜集起来并已经初步形成了自己的语言知识库。在此基础上，计算语言学所进一步收集更多的成语，增加多项语义属性特征，构建了现代汉语成语知识库（以下简称“成语知识库”）。目前该知识库收录的成语达到 36559 条。在大规模语料的基础上，对成语的使用频度进行了统计，并参考了其他有关汉语成语的分类与描述资料，对其中的 11705 条成语添加了详细的属性特征，作为成语知识库的完整库。该库中每个条目都基本描述了该成语的语法信息、语义信息与翻译信息等，其中的英译字段就分为直译、意译和英语近似成语三类。另外又挑选了其中的 3458 条成语形成核心库，作为常用成语供学习和研究参考使用。目前基本建设完成的汉语成语知识库实体的层次结构请参见图 1。

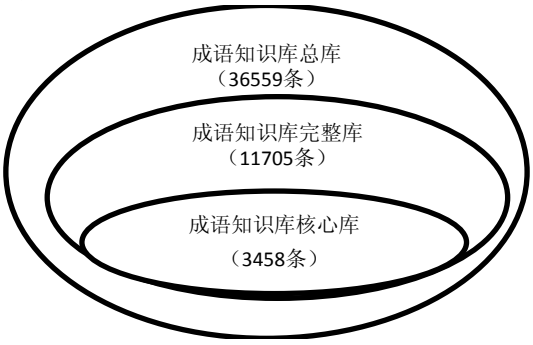

图 1 成语知识库层次结构图

有关成语条目的选择，周荐（1997）认为，判定成语的一个重要标准就是其“经典性”，提出“成语多出自权威性的著作，如十三经、官修和私撰的正史、子书和集书中的名家名作，因而具有经典性；成语之外的别类熟语，如谚语、歇后语、惯用语，则很少出自权威性的著作而多是随机的口头创作或出自俗白的作品，因而不具经典性。”虽然事实上古代的经典文献是成语的主要来源，但是从语言发展的历史角度看，近现代的一些经典文学作品虽然大多以白话写成，但这主要与汉语在近代所经历的由文言到白话的历史转变有关，并不是作者刻意舍弃文言而使用白话文，因此其中形成的成语我们认为也应该具有“经典性”。从这个意义上讲，我们兼顾古代经典文献和近现代经典文本进行综合考虑，也把这些成语收录进来。

目前成语知识库中，四字的成语占绝大多数，共有 34709 个，占总数的 95%，这一事实与语言学家们的认识是一致的。吕叔湘和朱德熙（1979）认为，成语多数是四个字的, 最普遍的格式是上下两截用对对子的办法连在一起。例如：“远走高飞”、“摩肩接踵”等。对于如何选择其他字数成语条目的问题，在查阅相关参考资料时我们发现，一些成语词典收录了两字的成语，如“陵迟”、“横失”等。但事实上这些两字组合是否符合成语的定义值得商榷。经过研究我们发现，它们共有的一个特点是在现代汉语中几乎都不再被使用，所以我们决定在成语知识库也不予以保留。比较难选择的是相当数量的三字组合，一些词典收录了而另一些却没有，其是否应该归入成语也缺乏明确、公认的标准，所以我们从成语应具有“经典性”这一原则出发，出

于实用（主要是参考使用频次）、易于操作的角度考虑，放弃了如“卷铺盖”、“耍嘴皮”等条目，仅把少数有明确来源出处的，如“借东风”等，收录在成语知识库中。我们统计了成语知识库中除四字成语外其他字数的成语，其分布见图 2。

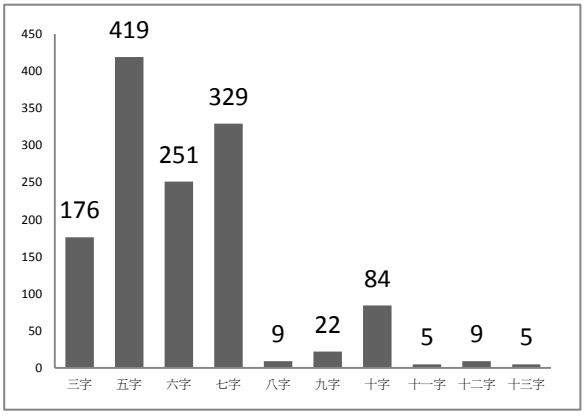

图 2 除四字外其他字数的成语数目

目前成语知识库包括新增字段共有属性字段 16 个，按照词汇、语义、语法和语用四项进行分类，具体分类情况见表 1：

表 1 目前成语知识库包含的属性字段

类别	属性字段
词汇	成语、 拼音、全拼音、 变体、 释义、 来源
语义	近义、 反义、 直译、 意译、 英语近似
语法	语法结构、 句法功能
语用	等级、 频次、 情感色彩

属性字段中的“变体”是指一些成语具有与其结构近似的其他形态，例如“山盟海誓”也写作“海誓山盟”。“等级”属性是针对前文提到的成语知识库的三个层次结构而设立的字段。“频次”字段是一个整数值，统计的是该成语在《人民日报》语料库（1947-2002 年）中出现的频次。成语的“情感色彩”字段共有三个属性：“褒义”、“贬义”和“中性”，对成语自身表现出的情感色彩做了描述。

3 成语知识库与汉语文教学

在汉语文教学中，首先可以从成语知识库中获得的是语言学的相关知识。利用字频统计的方法，从用字的角度看，成语知识库的成语条目共用汉字 150850 个，出现汉字个数为 5103。同时我们还可以利用 ACCESS 数据库模糊查询的方法设计相应的查询界面，查找成语知识库中的成语条目并统计出含有某种动物、植物、器物的成语，也可以查找出符合某种结构形态的成语，从而获得成语结构方面的知识。具体参见下表 2：

表 2 用数据库查询的方法获得有关成语的知识

动物		植物		器物		形态	
狗(共 140 条)	虎(共 287 条)	花(共 314 条)	树(共 93 条)	杯(共 42 条)	箭(共 40 条)	一 A 一 B(共 139 条)	A 不 BC(共 364 条)
白云苍狗	虎踞龙蟠	花落谁家	枯树逢春	杯弓蛇影	光阴似箭	一弛一张	力不能及
鸡飞狗跳	虎入羊群	花无百日红	玉树临风	杯觥交杂	箭在弦上	一言一行	衣不解带
狗嘴吐不出象牙	虎虎生风	花信年华	蚍蜉撼树	杯中之物	一箭双雕	一颦一笑	入不敷出
…	…	…	…	…	…	…	…

其次，利用成语知识库还可以获得与汉语相关的历史、文化知识。利用成语知识库的出处来源字段，我们考查了成语知识库中的各个成语条目的出处来源信息，将成语出现数量最多的

前十个经典文献列入表 3：

表 3 成语出现数量最多的前 10 个经典文献

文献序号	文献名称	出现成语数量	文献年代	文本长度	出现成语比例
1	《史记》	1567	西汉（公元前 93 年）	533505 字	0.28%
2	《汉书》	1181	东汉（公元 83 年）	742298 字	0.15%
3	《后汉书》	1067	南北朝（公元 488 年）	894020 字	0.12%
4	《左传》	941	春秋末年(公元前 770—前 476 年)	196845 字	0.48%
5	《晋书》	872	隋唐（579 年-648 年）	1158126 字	0.07%
6	《庄子》	847	西晋（265 年—316 年）	80400 字	1.05%
7	《论语》	833	战国初(公元前 475 年－公元前 221 年)	21683 字	3.84%
8	《孟子》	698	战国(公元前 290 年)	34685 字	2.01%
9	《诗经》	634	西周（公元前 1046 年-公元前 771 年）	41500 字	1.53%
10	《礼记》	625	西汉（前 202 年-9 年）	99010 字	0.63%

从表 3 中的文献年代一栏我们可以看出，汉代是成语出现最集中的时期，这也充分证明“汉语言”、“汉民族”这些名称是来自于汉朝这一历史时期。从成语生成的角度看，汉语文献的经典性位于第一位的是《论语》，随后依次为《孟子》、《诗经》、《庄子》等，儒家传统对中国文化的影响可见一斑。

在汉语言文学教学方面，在汉语词汇体系中，另一种与成语性质相似的汉语特殊语言形式是典故。一般辞典对典故的解释是：“诗文等作品中引用的古代故事和有来历出处的词语”。与典故相比，成语的词形与词义相对固定。典故的词形会随着作品的句型、句法改变，典故的词义须附着在作品的语境中才能显示，而且同一个出处来历的典故词语，词义也可能随着作品的语境而有所侧重。成语常用于语体文或口语，典故常用于诗词曲等韵文，几乎不用于口语。

以成语“倾城倾国”为例，它出自《汉书》卷九十七上〈外戚列传上・孝武李夫人〉，用以形容女子容貌绝美，而作为典故使用时在诗词曲等韵文里使用的词形有：一顾倾城、绝代佳人、国自倾、思倾国、倾人城、倾城色、倾城国、佳人绝世、倾国风流、倾国、倾城等。若将成语“倾城倾国”与不同词形的典故词汇对应，学会成语的语义，同时也就学会很多典故的语义。由蒋经国国际学术交流基金会资助，由元智大学、北京大学、香港理工大学、日本早稻田大学、韩国首尔市立大学共同合作的“历代语言知识库建置计划”，在北大成语知识库及元智大学的诗词曲典故数据库（http://cls.hs.yzu.edu.tw/orig/）基础上，进行了成语、典故的对应工作，让使用者从学习成语，扩展到理解典故，从能读懂语体文的成语，扩充到能理解诗词曲典故的词义，进而能赏析中国文学的宝藏——诗词曲韵文。

在教学方法上，左东琳（2002）认为，针对现代学生的心理特点，教学宜由简入繁、由浅入深、由易到难。而在这个过程中，培养学生学习兴趣，掌握适当的学习方法是很重要的。从成语入手，可以培养学生的学习兴趣。除了利用统计的方法了解成语知识库中的成语所涉及的语言学和文化相关信息外，成语知识库还可以在汉语文教学中引入趣味的元素。例如，作为一个存储容量、查询效率都远大于人的大脑的知识库，利用成语知识库可以完成我们熟知的“成语接龙”的游戏。从严格的意义上来讲，接龙中的成语其首字必须与前一个成语的尾字相同，而其尾字必须与后一成语的首字相同。我们利用图的深度优先理论设计了相关算法并实现了计算机程序，计算出了以每个条目为开头的成语接龙。最后共获得数目大于二的接龙 10253 条，其中最长的接龙包含成语 647 个，具体接龙情况如下：

胸有成竹→竹报平安→安富尊荣→荣华富贵→贵而贱目→目无余子→子虚乌有→
有目共睹→睹物思人→人中骐骥→骥子龙文→文质彬彬→彬彬有礼→礼贤下士→
士饱马腾→腾云驾雾→雾里看花→花言巧语→语重心长→长此以往→往返徒劳→
劳而无功→…

我们可以把这些接龙存储在数据库中，在教学中随机选择成语在课堂上让学生进行接龙游戏，既可以让学生口头做也可以让学生利用计算机查询数据库，边参与游戏边学习，然后利用上面自动获得的接龙结果进行提示或者核对答案。

4 成语知识库与对外汉语教学

成语一般都有很强的语义，其语用的主要目的是进行夸张和描写，而在实际生活中夸张有着严格的语境限制，因此无法像日常用语那样方便灵活的使用。尤其对于汉语作为外语的学习和使用者来说，把握不好这种成语使用的语境可能会给语言交流造成很多问题和困扰（谢新卫 2006）。王雷（2011）基于成语知识库中成语出现的频次等信息对 1000 条常用成语进行了整理、编辑，完成了用于对外汉语教学的参考书《中国成语 1000（汉英对照）》。在本书中，给每一条选择的成语都提供了例句，帮助读者判断语境和使用条件。为了帮助母语是英语的汉语学习者更好地理解成语的含义，对成语一般都采用了直译和意译两种方式，直译是对构成成语的各个语素字进行了单独的翻译，各个字用“\”分隔，让阅读者思考并联想语素字和成语语义的关联关系。

然而这种方法并不能适用于所有的成语，汉语的成语有着自身的一些结构特点。从结构上看，相当一部分成语具有一种类似于对称的结构，例如“欢天喜地”、“不折不扣”等，而其中一些处于对称位置的语素构成了后来汉语中很多的合成词，如前面两个例子中的“欢”和“喜”构成了“欢喜”，“折”和“扣”构成了“折扣”。因此，在对这类成语进行翻译时，如果这两个语素无法单独翻译，就采用了语素位置指示加完整的合成词翻译的形式。另外，汉语成语中的一些语素只用来表示或者夸张某种程度，并非机械使用这些词的本义，比如“千军万马”中的“千”和“万”，“一心一意”中的“一”，这类语素通常是按照这个字的本义进行翻译的。

而意译字段则对该成语的语义进行了整体上的翻译，帮助读者理解成语整体的含义。为了帮助读者全面地从多角度理解该成语，我们又利用现有的一部电子英语成语辞典，把其中与汉语成语语义近似的英语成语用自动的方法先找出来若干候选，再经过人工筛选、校对给一部分成语补充了“英语近似成语”信息。对一些有故事、典故的成语，为了帮助读者更好的理解成语的来源和形成背景，我们提供了一个简单的故事及其译文，并且把两者的对应关系以句子为单位用标号“①…②…”标示出来，形成了中英文对应的“句对齐”结构，方便读者阅读并同时进行学习。下面的图 3 就是《中国成语 1000（汉英对照）》上述体例的一个示例。

杯/弓/蛇/影 bēi gōng shé yǐng

<Literal>cup/bow/snake/shadow. The shadow of a bow is mistakened as a snake in one’s cup.

<故事>①晋朝河南人乐广十分好客。②一天他发现一个朋友好久不来他家里，感到十分奇怪，就去拜会他的朋友。③朋友说上次在你家喝酒发现杯中有蛇，喝后回来后就生病了。④乐广很困惑，回到家找原因：原来是挂在墙上的弓影子倒映在酒杯里。⑤朋友听说原因后病就好了。

<解释>疑心非常重而引起恐惧。

<例子>连续经过了三次失败，他变得很胆小，几乎到了杯弓蛇影的地步。

<Story>①In Jin Dynasty, a man named Yue Guang in Henan was very hospitable. ②One day he found a friend stopped visiting him for a long time. Surprised, he decided to pay his friend a visit.③His friend told him that he had found a snake in his cup when they drank in Yue’s house last time and turned sick ever since.④Confused by this, Yue returned home to find the reason: It was the shadow of a bow hanging on the wall.⑤After his friend learned this, he got well soon.

<Explanation>to be so suspicious as to arouse fear

<Example>After three failures in a row, he turned to be very timid and almost became ~.

<English Equivalent>to be afraid of one’s shadow

图 3 形成中英文“句对齐”结构的成语故事和译文

作者希望通过这本书，能够帮助汉语学习者更方便地使用成语，提高阅读写作能力和口语表达技巧，进而对成语所蕴含的中国悠久历史和灿烂文化产生更浓厚的兴趣。

5 结语

本文介绍了由北京大学计算语言学所建设的汉语成语知识库，包括成语知识库的内容及目前在汉语文教学和对外汉语教学中的应用。成语知识库作为北大计算语言学所综合型语言知识库的一个重要组成部分，2011 年 CLKB 获得了国家科学技术进步奖二等奖（证书号：2011-J-220-2-02）。总而言之，作为一门语言中最具表现力的一部分，成语浓缩了古人的经验智慧、人生哲理、道德操守、审美情趣等，是语言教学中最具有知识性、趣味性的部分。所以如何在汉语教学中充分重视成语教学，值得每一位教育工作者仔细研究、认真思考。虽然成语具有很大的稳定性，但从成语的整体看，还是在不断发展、变化的。这主要是指新的成语不断生成、部分旧的成语的逐渐消亡以及成语意涵、内容和形式不断发展变化。现在我们在着手考虑并制定计划在进一步完善现有成语知识库的基础上，动态调整成语知识库的内容，进一步对成语在语文教学中的应用深入研究，把成语知识库发展成促进汉语教学水平不断提高的重要的学习资源。

参考文献

[1] Fellbaum, Christiane. Idioms and Collocations: Corpus-based Linguistic and Lexicographic Studies (Research in Corpus and Discourse)[M]. London: Continuum International Publishing Group Ltd.2007：157-196.

[2] Lin, Dekang. Automatic Identification of Noncompositional Phrases[A]. In Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics on Computational Linguistics[C].1999. Maryland, USA: 317-324.

[3] Lo, Wing Huen. Best Chinese Idioms (Vol. 3)[M]. Hong Kong: Hai Feng Publishing Co,1997: 20-38.

[4] 吕叔湘, 朱德熙.《语法修辞讲话》[M].北京：中国青年出版社,1979：41.

[5] 王雷.《中国成语 1000（汉英对照）》[M].北京大学出版社, 2011：65-86.

[6] 俞士汶.国家重点基础研究发展计划 973 计划课题任务书[R].北京大学,2004：6.

[7] 俞士汶, 穗志方,朱学锋. 综合型语言知识库及其前景[J]. 中文信息学报,第二十五卷第六期. 2011 年 11 月:12-20.

[8] 俞士汶,朱学峰,王惠.《现代汉语语法信息词典详解（第二版）》[M].北京:清华大学出版社,2003：51.

[9] 谢新卫.第二语言教学中成语教学探析[J]. 语言与翻译(汉文) ,2006 年第 4 期：64-67.

[10]周荐.论成语的经典性[J]. 天津:南开大学学报,1997 年第 2 期：29-35.

[11]左东琳. 语文教学中的成语教学[D]. 硕士毕业论文. 辽宁师范大学图书馆. 2002 年：2.