国际中国语言学学会第 21 届年会(IACL 21)的工作坊--——歷代语言知識库的建构与应用
蒋经国国际学术交流基金会《历代语言知识库建置计划》（2009）

特殊词语：成语知识库的建构

俞士汶[1] 王雷[1] 朱学锋[1] 罗鳳珠[2] 砂冈和子[3] 姜柄圭[4]

1.北京大学；2.元智大学；3.早稻田大学；4.西江大学

摘要

在汉语的词汇体系中，成语是一类颇具特色的词语。汉语成语知识库(**C**hinese **I**diom **K**nowledge **B**ase, **CIKB**)在融入历代语言知识库(**D**iachronic **L**anguage **K**nowledge **B**ase，**DLKB**)大家庭之前，已在北京大学经历了两个阶段的发展。

第一阶段（自1986年起）它以《现代汉语语法信息词典》[1]（**G**rammatical **K**nowledge **B**ase of Contemporary Chinese, **GKB**）中的成语库和习用语库的形式呈现。**GKB**的研制在词典类通用型汉语语言知识库的建设中实现了第一次转向，即过去词典都是面向人的，而**GKB**是面向自然语言处理系统的。当前**GKB**收了8万余汉语词语，其中成语和习用语共有9千多。

第二阶段(自2004年起)根据综合型语言知识库(**C**omprehensive **L**anguage **K**nowledge **B**ase，**CLKB**)[2]的规划，成语知识库单独建库，**CIKB**在继承**GKB**的设计思想及其成语库和习用语库的全部成果的基础上，新增了大量成语(包括习用语)，对成语的描述内容更加丰富。当前**CIKB**包含3.6万余记录(records)，描述每个成语的词法、句法、语义和语用的属性信息字段(fields)约有30项，已有1.1万余条记录完成了全部属性字段信息的赋值，其中“英语意译”、“英语直译”、“英语近似成语”3个字段不仅可以支持机器翻译、机器辅助翻译等跨语言信息处理研究，也让**CIKB**可以在国际汉语教学领域中发挥作用。

第三阶段即自2010年加入“历代语言知识库建置”计划起，**CIKB**的发展有了更开阔的视野，更加重视 **CIKB**在汉语教学特别是东亚地区的汉语教学领域所能发挥的潜能。到目前为止，**CIKB**有如下进展：

（1） **CIKB与历代语言知识库中的另一重要成员“诗词曲典故资料库”进行了连接，相互参照。**以成语“倾城倾国”为例，与之相关的在古典诗词曲等作品中使用的典故（词形）很多，如“一顾倾城”、“名花倾国”、“倾国风流”、“倾人城”、“倾城色”、“倾城国”等等。成语与典故的对应使两个自立的知识库交相辉映，可以让人同时领悟成语与典故的含义，不仅有助于理解使用成语的汉语文本的内容，还能增强赏析中国古典文学瑰宝的能力。

（2） **建构汉语成语教学网站。**从**CIKB**中选择了2千个高使用频率，而且具有典型意义的成语及其完整的属性信息，作为构建汉语成语教学网站的基础资料。网站由罗凤珠设计并负责实施，上网资料增加繁体字形、台湾所使用的注音符号、日语、韩语等很多新内容。网站上的成语知识库将以便于检索、浏览的方式呈现。在多方交流的过程中，对流传在多地域的汉语成语增加了新的认识，如成语在不同地域的使用频率是有差异的，又如某些成语在海峡两岸的读音也发生了变化（如“不可收拾”，大陆出版的《现代汉语词典》的汉语拼音是 bu2ke3shou1shi5, 其中的“拾”读轻声，而罗凤珠提供的台湾读音是阳平；又如“从容不迫”大陆读 cong2rong2bu4po4，其中的“从”读阳平，而台湾读音是阴平)。

（3） **进行了汉语成语教学实践活动。**对汉语成语知识库在汉语文教学和对外汉语教学领域的应用进行了探索[3]。王雷还出版了《中国成语1000（汉英对照)》(北京大学出版社, 2011年)，

该书内容丰富，编排新颖，每个成语不仅有对应的英译和逐字的直译，就连“故事”、“解释”和“例子”也有英译，而且是句子对齐的。日本和韩国的教学实践案例有待补充。

汉语成语知识库的建构是一项庞大的语言工程，建构方法本身也是一项重要的研究课题。自语言知识库开篇之作**GKB**开始研制，20多年来贯穿**CLKB**建构的全过程都采用人机互助的方法。自动建构本质上是机器辅助构建，尽可能采用适用的成熟的软件技术，如数据库技术，机器学习技术等等，可以保证工程的规模和进度。不过，单纯依赖自动技术建构的语言知识库的质量尚不能满足应用的需要，因此必须投入相当多的人力，必须投入高水平的专家的力量。专家的知识和奉献才是语言知识库质量的保证。对此，我们始终保持清醒的认识，坚持不懈。

汉语成语知识库还有广阔的发展空间。现在已纳入两个中国国家自然科学基金项目“隐喻识别与理解的理论与方法研究”（2012年-2015年，王治敏博士主持, 俞士汶参加）和“汉语全文词义标注关键技术研究”（2013年-2016年，曲维光教授主持，朱学锋参加）以及北京大学计算语言学中国教育部重点实验室开放课题“汉语和英语多词表达中的隐喻研究”（2013年起，王雷主持）。基于**CIKB**还可以衍生出很多有意义的、有趣的研究题目，如系统地研究成语对社会生活形态的反映（一个有趣的例子：“三从四德”是封建礼教对女子品行的规范，早已过时。现在有了新“三从四德”：老婆出门要跟“从”，老婆命令要服“从”，老婆讲错要盲“从”；老婆化妆要等“得”，老婆花钱要舍“得”，老婆生气要忍“得”，老婆生日要记“得”。这固然是调侃之词，却也折射了某种环境中男女地位的变化。又如度量衡制度的改变使得很多成语表达的意义与当代的生活形态不相适应，变得难以理解。）也许这样的研究正是当初构建历代语言知识库计划所追求的目标之一吧？

参考文献

[1] 俞士汶，朱学锋等. 现代汉语语法信息词典详解(第二版). 北京：清华大学出版社，2003 年 2 月

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

[3] 王雷, 俞士汶，朱学锋，罗凤珠. 汉语成语知识库与汉语教学. 第八届中文电化教学国际研讨会，中文教学现代化学会主办. 会议地点: 上海，2012 年 8 月 8-11 日.