面向语言能力提升的成语知识库建构及扩展 *

俞士汶[1] 罗鳳珠[2] 朱学锋[1] 王雷[3] 常宝宝[4]

[1] 北京大学计算语言学教育部重点实验室

[2] 元智大学中國語文學系

[3] 北京大学外国语学院

[4] 北京大学计算语言学研究所

Email: {yusw, wangleics, chbb}@pku.edu.cn, gefjulo@mail2000.com.tw

摘 要：汉语成语知识库是北京大学计算语言学研究所研制的综合型语言知识库大家庭中的一名新成员，无论对于人还是机器的语言能力的提升，它都可以发挥作用。本文比较详细地介绍汉语成语知识库的内容、构建理念和发展历程。将汉语成语知识库与元智大学罗凤珠建置的诗词曲典故资料库相连接，两个知识库相得益彰。本文还提出了基于汉语成语知识库值得进一步研究的若干课题。

关键词：综合型语言知识库，语言能力，成语，成语知识库，典故资料库

Construction and Expansion of Idiom Knowledge Base for Improving language ability

Yu Shiwen[1] Lo Fengju[2] Zhu Xuefeng[1] Wang Lei[3] Chang Baobao[4]

[1]Key Laboratory of Computational Linguistics at Peking University, Ministry of Education

[2] Yuan Ze University

[3] School of Foreign Languages, Peking University

[4]Institute of Computational Linguistics, Peking University

Email: {yusw, wangleics, chbb}@pku.edu.cn, gefjulo@mail2000.com.tw

Abstract: The Chinese Idioms Knowledge Base (**CIKB**) is a new member of a large family known as the Comprehensive Language Knowledge Base (**CLKB**) developed by the Institute of Computational Linguistics at Peking University. It can play an important role for improving language ability of both humans and computers. In this paper we introduce in detail the content, construction ideas and development process of **CIKB**. The connection of **CIKB** with the Allusion Knowledge Base of Chinese Poetry built by Lo Fengju from Yuan Ze University makes the two knowledge bases complement each other. Based on CIKB, a few research proposals for further study are also provided.

Key Words: Comprehensive Language Knowledge Base, language ability, idiom, Chinese Idioms Knowledge Base, allusion knowledge base.

1 引言

语言知识库是自然语言处理系统的基础设施，其规模和质量在很大程度上决定了自然语言处理系统的成败。北京大学计算语言学研究所积20多年之努力研制的**综合型语言知识库**[1]，为支持中文信息处理的原创性科学研究和应用开发做出了贡献。**CLKB** 仍在发展中。**汉语成语知识库**是植根于**CLKB** 沃土的新株，是 **CLKB** 大家庭中的一名新成员。人为了准确理解和正确运用成语，免不了要查成语词典等工具书，现在还可以上网查询。不过，无论是词典还是网上资料，都是给人看的，不适合机器使用。**汉语成语知识库**与 **综合型语言知识库**一样，是面向机器语言能力提升的。自然语言处理系统中的语言知识库都具有规模大、格式化、周遍性、显性化等特点（周

* 基金项目：国家重点基础研究发展计划(2014CB340504)和国家自然科学基金项目（61272221，61170163）

遍性指给人看的词典收了“哀鸿遍野”，可以不收“哀鸿遍地”，但**CIKB**都收；又如，词典通常只收“排山倒海”，不收“倒海排山”，**CIKB**也在“变体”字段中列出。显性化指采用标注的办法将隐含在语料中的词法、句法、语义知识显现出来。)， **CIKB**亦是如此。这些特点使得它在汉语教学和汉语国际教育中也可以发挥作用[2]。

诗词曲典故资料库[3]是元智大学罗凤珠建置的，旨在借助电脑强大的存储、记忆、搜索、比对功能，提升人对文学作品中的典故的理解与运用能力，又基于资源共享的理念，在建置之初便建设了**诗词曲典故网站**（http://cls.hs.yzu.edu.tw/orig/），已产生广泛影响。建置**诗词曲典故资料库**的初衷虽然是为提升人的语言能力服务的，特别是为了提高古典文学作品的鉴赏能力。由于它也采用了数据库的结构化描述形式，可以将它扩充到面向自然语言处理的语言知识库中以提升机器对诗词曲等汉语古典文学作品的解读能力。

某些成语和典故之间有共同的渊源，字面或意义有某种程度的相似性，将成语知识库和典故资料库连接，两个知识库相得益彰，信息相互补足，可以发挥一加一大于二的作用。

2 汉语成语知识库的内容与成长历程

2.1 汉语成语知识库之样例

表1 给出成语知识库中3个成语的样例。

表1 成语知识库样例

词语	国色天香	开门揖盗	嫣然一笑
拼音	guo2se4tian1xiang1	kai2men2yi1dao4	yan1ran2yi1xiao4
变体	天香国色		一笑嫣然
结构	并列	连动	状中
褒贬	褒	贬	褒
子类	IA	IV	IV
主语			
宾语			
谓语	谓	谓	谓
定语	定		
状语			
补语	补		
近义	花容月貌、美若天仙、雍容华贵	引狼入室、开门延盗	莞尔一笑
反义	其丑无比、丑若无盐、其貌不扬	拒之门外	面目狰狞
直译英文	national beauty and heavenly fragrance	to open the door and invite robbers.	(of a woman) to give a charming smile
意译英文	beautiful and elegant female	to invite disaster by letting in evildoers.	
近似英文		Opportunity makes the thief.	
释义	原形容牡丹花的颜色和香气。喻指容貌极其美丽的女子。	揖：作揖；拱手行礼。意指打开门请盗贼进来。	嫣然：美好的样子。形容女子笑得很美。也用于形容花。
出处	唐·李濬《松窗杂录》：“臣尝闻公卿间多吟赏中书舍人李正封诗曰：天香夜染衣，国色朝酣酒。”	晋·陈寿《三国志·吴志·孙权传》：“况今奸宄竞逐，豺狼满道，乃欲哀亲戚，顾礼制，是犹开门而揖盗，未可以为仁也。”	战国·楚·宋玉《登徒子好色赋》：“腰如束素，齿如含贝，嫣然一笑，惑阳城，迷下蔡。”
例句	国色天香，乃牡丹之富贵。	把他招进公司，无异于开门揖盗。	她对他嫣然一笑，转身而去。
频次	81	53	45
级别	1	1	1

成语知识库采用关系数据库的二维表格式描述每个成语的各项属性。为方便阅读，表1 对数据库

文件中的行、列进行了转置。

《现代汉语词典》对“成语”的释义是：“人们长期以来习用的、简洁精辟的定型词组或短语。汉语的成语大多由四个字组成，一般都有出处。有些成语从字面上不难理解，如‘小题大做’、‘后来居上’等。有些成语必须知道来源或典故才能懂得意思，如‘朝三暮四’、‘杯弓蛇影’等。”成语的界定有狭义、广义之分，狭义的仅指那些难以从字面推断整体意义的成语，不过狭义、广义的界限并非泾渭分明。成语库收录广义的成语，实际上囊括了一部分习用语。

成语库的属性字段名称的含义大部分可以顾名思义，有些简要说明一下。词法部分中的**结构**指最外层的句法结构，像“国色天香”的“结构”为“并列”，意指“国色”与“天香”并列，没有进一步深究“国色”与“天香”的结构。四字格的成语多为“2+2” 的格式，但并非都是如此，像“感同身受”是“1+3”的主谓结构，而“一衣带水”是“3+1”的定中结构。句法部分的几个字段将在2.3中说明。附注部分中的**频次**的值是成语在1947年到2006年《人民日报》语料中出现的次数。**级别**字段用于区分成语的典型性。**级别**为1者指被广泛认可的最典型的成语，2，3者次之。

成语库仍在建造中。目前36000多个记录中，已有10000多个记录的各个字段的信息已经齐全。

2.2 构建汉语成语知识库的基本思路

成语在现代汉语的书面语言和口语中并不少用。成语的理解成为汉语理解的一道难关，原因有：1）成语中有些字不常用，像“人稠物穰”、“锱铢必较”、“罄竹难书”中的“穰”、“锱”、“铢”、“罄”都不是常用字；2）含费解的词，如“膏火自煎”中的“膏火”、“乌合之众”中的“乌合”；3）意义不能根据字义推断，如“倚马可待”、“铄金毁骨”、“胸有成竹”、“固若金汤”；4）望文生义可能导致似是而非或一知半解，如“一丝不苟”、“司空见惯”；5）同义或近义成语的使用语境大不相同，如“坚如磐石”与“固若金汤”；6）意义相同或相近的成语褒贬色彩可能相左，如“一丝不苟”与“斤斤计较”；7）近似成语的辨析相当困难，如“昨日黄花”是否为“明日黄花”之误？又如“每下愈况”同“每况愈下”的区别在哪里？8）意义、用法、褒贬义随时代迁移而发生变化，如“金屋藏娇”。另一方面，成语具有言简意赅、寓意深刻、形象生动、琅琅上口等特点，为求言谈、文笔优雅，富有文化内涵，中国人喜欢使用成语，也不乏化用成语的例子。荣获2013年台湾第七届myfone行动创作奖情书组首奖的“**你别来，我无恙。**”化用成语“别来无恙”，成为历届得奖作品中最精简的。俞士汶感慨于当今Email（《现代汉语词典》第5版以“**伊妹**儿”的词形收入）功能，诌打油诗一首：“伊妹传书超雁飞，鸟语花艳音乐美；待到网络传香来，虚拟世界人亦醉。” 其中也化用了两个成语：“鸿雁传书”和“鸟语花香”。无论对于人还是机器，成语的正确理解和运用对语言能力的提升具有重要意义。在自然语言处理系统中，配置汉语成语知识库是提升自动处理能力的一个现实可行的解决方案。

汉语成语蕴含着丰富的历史、社会和文化知识，是现代汉语语汇中最具民族特色的词语，其中很大一部分还有典故，可从文献中找到出处，是珍贵的民族文化遗产。建设成语知识库也是民族文化遗产传承的一种方式。

成语知识库是一项知识密集型的语言工程，采用人机互助的构建方法。

2.3 成语知识库构建的三个阶段

成语知识库的成长可划分为3个阶段。这里只给出各阶段的起始时间，因工作都还在继续。

第一阶段始于1986年。北京大学计算语言学研究所在其研制的《现代汉语语法信息词典》[4]（以下简称**GKB**）中就包含有成语库（6800多条）与习用语库(2400多条)。

GKB 是面向自然语言处理的机器词典，包含8万多现代汉语的词语。**GKB**的结构特点是在词语归类的基础上，按词类（名词、动词、形容词、区别词、副词、成语、习用语等）分别建数据库文件，详细描述每个词语的多种语法属性，另建一个总库，描述各类词语的共同属性。**GKB**中的成语库的格式和内容大致相当于**2.1**的表1中表述句法信息的字段名及其字段值。“主语”、“宾语”、“谓语”、“定语”、“状语”、“补语”分别表示该成语可否充当句法结构中的相应成分。由于

名词、动词、形容词、副词等基本词类是根据语法功能划分出来的，而成语、习用语不是。在这个意义上，成语、习用语与基本词类不在同一个层次上。为适应句法分析的需要（句法树通常以反映语法功能的词类的代码为终结符），有必要对成语、习用语进行更细致的区分。**GKB** 对成语、习用语划分了子类。划分成语、习用语子类的准则与基本词类是一致的，即依据它们的句法功能的优势分布[4]。例如，“良师益友”、“丰功伟绩”、“八拜之交”属于名词性子类；“别有洞天”、“此起彼伏”、“开门揖盗”属于动词性子类；“富丽堂皇”、“井井有条”、“语重心长”属于形容词性子类；“不由自主”、“自始至终”属于副词性子类，等等。

对超过9000条的成语和习用语详细地给出它们的句法属性信息，并把它们归入不同的子类，在**GKB**之前，这件事是没人做过的。

第二阶段从2004年算起。成语知识库单独建库，以**GKB**中的成语库和习用语库为基础，大幅度增加了成语的数量，达36000多条。数量多了，典型性难免参差不齐，也有鱼目混珠者。已对这些成语在60年的《人民日报》语料范围内统计了频次并划分了级别。这个阶段的工作重心是增添语义信息字段：“近义”、“反义”、“直译英文”、“意译英文”、“近似英文”等。

第三阶段始于2010年。在“历代语言知识库”的规划下，发展成语知识库的视野更开阔了。与“历代语言知识库”中的另一知识库“诗词曲典故资料库”进行连接[5]，实现信息互补，交相辉映，可以让汉语学习者同时领悟成语与相关典故的含义，不仅有助于理解现代汉语，还能增强赏析古典文学的能力。这个阶段也更加重视发挥成语-典故知识库在语文学习中的作用[6,7,8]。

3 与诗词曲典故资料库的连接

诗词曲典故资料库共收录典故近2万，表2给出了一个示例（资料库中的字体已经转换）。

表2　典故资料库示例

典故	类别	同义典故	相关典故	参见典故	朝代	人物	典籍	典籍内容节录
嫣然一笑	语典	一笑嫣然、嫣然凝笑	楚女窥墙、东家一笑、东墙窥宋、窥宋玉、惑阳城、迷下蔡、莫把一分增减。		战国（楚）	宋玉、楚王	宋玉〈登徒子好色赋〉	玉曰：“天下之佳人莫若楚国，楚国之丽者莫若臣里，臣里之美者莫若臣东家之子。东家之子，增之一分则太长，减之一分则太短，着粉则太白，施朱则太赤。眉如翠羽，肌如白雪，腰如束素，齿如含贝。嫣然一笑，惑阳城，迷下蔡。然此女登墙窥臣三年，至今未许也。”

苏轼咏海棠的名诗《寓居定惠院之东杂花满山有海棠一株土人不知贵也》:“嫣然一笑竹篱间，桃李漫山总麤俗。”便使用了这个典故。

典故资料库的“类别”字段指典故可分为3类：（1）来自于古书词句的“语典”；（2）来自于历史人物或事迹的“事典”；（3）兼具语、事两种来源的“语事混合典”。典故用于诗词曲韵文，其平仄声调、字数、句法都受到格律限制，因此没有单一的固定词形。典故有同义异形的情形，称为“同义典故”；也有一些典故来自于相同的语、事，但后人引用时，因含义的侧重面不同，而成为词形或相同或不同的典故，称之为“相关典故”；还有语、事出处不同，诗句中的词形不同，但所取的词义相同或相近，称之为“参见典故”。

成语与典故都是汉语中有来历、有出处的特殊语言形式，两者的差别是成语有固定词形，多数用于语体文或口语，典故没有固定词形，多数用于诗词曲韵文。因此，有的成语就可能和若干个典故有共同的渊源，其语义有内在的联系，有一定的相关度。基于这样的认识，将成语知识库同典故资料库相连接，建置融为一体的成语-典故知识库。借助这样的知识库，学习者在学习一个成语或典故的同时，便可了解数种不同词形的典故，再通过“出处典籍”查找文学原著，视野就更广了，还可以发现成语和典故产生、演变与运用的脉络[8]。

由于成语知识库和典故资料库的规模都相当大，所含记录皆数以万计，实现两者的有效连接并不是一件轻而易举的事。已经开发了一个自动映射软件可以由计算机辅助实现这种连接[5]。

在成语知识库和典故资料库之间建立映射关系的原理是计算成语和典故所在的记录之间的相似度。首先，成语和典故的字形本身是考察对象，像成语“嫣然一笑”与典故“嫣然一笑”字形完全相同。成语“邯郸学步”与典故“学步”部分相同。尽管成语“邯郸匍匐”与典故“学步”一个字也不同，但对照成语的“出处”字段和典故的“朝代”、“人物”、“典籍”、“典籍内容节录”字段的字面、意义、文献属性，发现都有相同之处，因此两者之间也有一定的相似度。自动映射软件给出了成语和典故所在记录之间相似度的综合评价策略与算法。对软件自动发现的相似度较高的 6000 个映射关系进行人工抽样检查，准确率达 87.5%。

已从成语库中精选 2005 条典型的常用成语，与典故库实现连接，建成“多语成语-典故知识库”，同时建立了分阶多语成语典故知识库网站（http://cls.hs.yzu.edu.tw/DLKB/idiom_indexlist.aspx）[7]。

4 有待进一步研究的课题

基于富含汉语语言知识的成语知识库，可以对很多相关课题进行更深入的研究。

⑴ 关于语言演化与社会环境变迁的交互影响

元智大学、北京大学、香港理工大学、早稻田大学和西江大学的学者进行“历代语言知识库”研究，其初衷就是期望探讨语言演化的脉络及其同社会环境变迁的关系。由于成语库浓缩了大量的中华历史文化知识，成语库可以为这个课题的研究提供线索。度量衡制度的改革、生活用具的变化、社会生态环境的变迁都影响了成语的产生和使用。像“海水不可斗量”、“得寸进尺”、“半斤八两”、“斤斤计较”这些成语都源自以往的市制度量衡，与日常生活密切相关，浅显好懂，易于流行。当将1斤等于16两改为1斤等于10两时，“半斤八两”就不符合其本意了，便出现了“半斤五两”的说法。现在改用公制，当从未使用过市制度量衡的世代成为社会主体时，这些成语还有人理解和使用吗？“斤斤计较”在60年《人民日报》语料中出现707次，而同义的“锱铢必较”只有68次，因为当代已经很少有人知道锱=1/4两，铢=1/24两。现在已有人造出“克克计较”，沿用久了，会不会也升格为成语？像“光阴似箭”、“同室操戈”、“化干戈为玉帛”这些成语显然同古代兵器密切相关。过去人们用“箭”这种飞行物形容速度快，现在人们更熟悉比“箭”快得多的其他飞行物，因此一位精通汉语的德国朋友在给我的邮件中写出了带有调侃意味的句子“光阴犹如子弹”。在礼教兴盛时期，“三从四德”曾是女子品德与行为的规范，现在大概很少有人知道，更没人遵从了。网上流传的“三从四得”（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”）也从一个侧面反映了当代社会生态的变化。

⑵ 成语与隐喻的关系

大量成语使用了隐喻（“光阴似箭”、“堆积如山”、“国色天香”、“刀山火海”、“举棋不定”、“寸土必争”等等）。考察这些成语，有助于深化对隐喻的认识，可以了解隐喻的类型（明喻、暗喻、转喻等）、哪些词语指称的事物经常作喻体（源域）、哪些经常做本体（目标域）以及成语中的隐喻对语体文中隐喻形成的影响，从而对隐喻计算提供支持，进而提升机器的语言能力。

像“知识的海洋”、“他是老狐狸”、“姑娘花一样”等人们现在常使用的隐喻词句的形成，固然符合隐喻形成的一般认知机制（即用身边的熟悉的具体的事物喻指不熟悉的抽象的事物），但也可能受到“学海文林”、“狐假虎威”、“花容月貌”这些成语的影响，反映了中华民族文化传承的脉络。

⑶ 成语的汉外翻译

无论是机器翻译，还是人翻译，成语的翻译都是拦路虎。成语知识库提供了一万多条成语的

英文翻译（直译、意译或近似的英文），分阶多语成语典故知识库网站提供了2005条成语的英、日、韩翻译[7]，这些资料是有参考价值的，不过当运用到文本的翻译时，还会有很多问题，并不能生搬硬套。以动词性子类的成语为例，成语库中通常提供的英译是英语动词的不定式，到了句子中，至少要根据上下文改变为适当的限定形式。

成语库中的翻译字段尚不完备，已有的译文也还需要订正、推敲。

⑷ 成语的生命期与发展趋势

成语同其他事物一样，也有生命期。对成语在语料中进行逐年的统计就能了解成语的生命轨迹。有些成语虽然在辞书可以查到，但可能从某个时代起就无人使用了。汉语成语知识库之所以未收入“丑若无盐”，原因就在于此。“百度知道”网站列了 15 个形容丑女的成语，也没有“丑若无盐”。久而久之， “丑若无盐”可能就销声匿迹了。《中国 2012 语言生活状况报告》给出了 2011 年前 50 个高频成语[9]，都通俗易懂，如“前所未有”、“见义勇为”等，而且这些成语在 2002 年出版的《新华成语词典》中都能查到。这些现象是否反映了成语通俗化的趋势？近年来网络上流行的一些用语，如“男默女泪”、“喜大普奔”、“细思恐极”等以及荣获 2013 年台湾第七届 myfone 行动创作奖生活笔记组首奖的“婚前朱丽叶，婚后玛利亚” 会不会演化为能被广泛认可的成语？

5 结语和谢辞

综合型语言知识库已经为推动中文信息处理事业的发展做出了自己的贡献[1]，于2011年获得中国国家科学技术进步奖二等奖。作为“综合型语言知识库”的衍生成果。期望汉语成语知识库同样能为提升人和机器的语言能力发挥作用。

本文相关研究除得到第一页标注的在研项目支持外，汉语成语知识库的研制还得到多个项目的支持，如2004年至2009年间执行的973项目“文本内容理解的数据基础”和2010年至2013年间执行的蒋经国国际学术交流基金“历代语言知识库建置”计划。

曾在北京大学计算语言学研究所工作或学习的业界同仁如李芸博士、王治敏博士等都曾为成语知识库的建设贡献过力量，在此致以诚挚的谢意。

参 考 文 献

[1] 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 25 卷第 6 期. 2011 年 11 月，12-20

[2] 俞士汶，朱学锋.综合型语言知识库及其在国际汉语教育中的应用初探.《国际汉语教育》，2013 年第一辑，174-180.

[3] 罗凤珠，蔡宛纯.以资源共享的观点建构数字文史工具书的方法：以诗词典故辞典网站为例.《汉学研究通讯》,2005 年，24（2）（总 94 期），17-29.

[4] 俞士汶，朱学锋等着.《现代汉语语法信息词典详解》（第 2 版），北京：清华大学出版社，2003 年 2 月

[5] 白易.《汉语成语与典故知识库的自动映像策略与实现》. 北京大学本科毕业论文，2011 年 5 月

[6] 王雷编着. 《汉英对照中国成语 1000》. 北京：北京大学出版社，2011 年 9 月第 1 版

[7] 罗凤珠，砂冈和子，姜柄圭，俞士汶，王雷，常宝宝. 分阶多语成语典故知识库教学设计. 《台湾华语教学研究 》，2013 年•第一期•总第六期，1-30

[8] 俞士汶，罗凤珠，朱学锋，王雷，常宝宝. 汉语成语及典故知识库在语文学习中的应用. 《台湾华语教学研究 》，2013 年•第二期•总第七期，13-36

[9]教育部语言文字信息管理司 组编.《中国2012语言生活状况报告》.北京：商务印书馆，2012年10月第1版，242-243