汉语成语知识库的建构理念与新进展

王雷[1,2] 俞士汶[1] 朱学锋[1] 罗鳳珠[3] 砂冈和子[4] 姜柄圭[5]
[1]北京大学计算语言学教育部重点实验室
[2]北京大学外国语学院
[3]台湾元智大学中国语言文学系
[3]早稻田大学政治与经济学院
[4]西江大学
Email: {wangleics, yusw}@pku.edu.cn[1,2];gefjulo@saturn.yzu.edu.tw[3] ; ksunaoka@gmail.com[4];
kbg4335@gmail.com[5]

摘要：在汉语中，成语是非常特殊的一个组成部分，其历史悠久、形态稳定、结构固定且多用比喻义。本文以描述汉语成语的特点为起点，详细辨析了成语、熟语、习语等多词表达的共同点和差别，给出了汉语成语面向中文信息处理的准确定义。重点介绍了北京大学计算语言学研究所建设的汉语成语知识库。作为一个重要的汉语语言资源，成语知识库除了能在机器翻译、机器辅助翻译、跨语言检索等自然语言处理任务中发挥重要作用外，它还对汉语语言学研究、对外汉语教学以及语言对比研究等工作也有重要的指导意义。

关键词：汉语成语知识库；建构理念；新进展

Principle and New Development of Constructing Chinese Idiom Knowledge Base

Lei Wang[1,2] Shiwen Yu[1] Xuefeng Zhu[1] Fengju Lo[3] Kazuko Sunaoka[4] Byeongkwu Kang[5]
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]
School of Political Science and Economics of Waseda University Tokyo 1698050[4]
Sogang University[5]
Email: {wangleics, yusw}@pku.edu.cn; ksunaoka@gmail.com; kbg4335@gmail.com

Abstract: Idioms are distinctive in Chinese for its long history, fixed constitution, continuity and metaphorical meaning in its context. This paper starts with a description of the characteristics of Chinese idioms and analyzes the similarities and differences of multi-word expressions such idioms, idiomatic expressions and phrases, which results in a clear definition of Chinese idioms for the purpose of Chinese information processing. We focus on a Chinese idiom knowledge base built by the Institute of Computational Linguistics at Peking University. As an important Chinese language resource, our idiom knowledge base will not only play a major role in NLP tasks such as machine translation, computer-aided translation, but also provides valuable assistance to cross-language research, linguistic research, teaching Chinese as a foreign language etc.

Key words: Chinese idiom knowledge base; principle of construction; new development

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[1]。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[2]、词典编纂[3]、自然语言处理[4]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。

此外，随着互联网的普及，搜索引擎已经逐渐成为人们学习、工作乃至日常生活的一部分。遇到问题，一些人首先想到用搜索引擎搜索一下，但是当前搜索引擎的能力和表现都还不能尽如人意。在进行搜索时，搜索对象限定为网页中用字符串表示的文字，而我们真正要搜索的是其表达的内容，并不是文字符号串本身。当搜索引擎尝试对类似于成语这样有固定组成结构的多词表达进行深入的语法、语义分析时，效果并不理想。因此从机器理解自然语言的角度，深入研究类似词组、习语、成语、谚语等这样的多词表达对于自然语言处理技术的提升会具有实质性的意义；而中文信息处理尤其需要大规模、高质量的、具有固定结构的词组、习语、成语等语言知识库的支持。

2 现代汉语中的成语及其特点

根据《现代汉语词典》[5]，汉语成语的定义为“人们长期以来习用的、简洁精辟的定型词组或短语。”其中“习用”一词表明成语须具有一定的历史渊源，亦有一个演变过程，通常时代感较强。从结构上看，成语书面语言用字较多，通常以四字格的形式出现，而其中“2+2“的联合结构又占大多数。从语义角度来看，正是由于成语具有“简洁精辟”的特点，导致其较普通词语难懂。有些可根据字面意义推断，有些必须知道来源或典故才能懂得其含义。描述性成语一般情感色彩比普通词语强，感情表达强烈。从修辞的角度看，很多成语具有隐喻，具有生动形象，寓意深刻的特点。从使用情况上看，中国国家语言文字工作委员会自 2006 年起，每年发布的《中国语言生活状况报告》，都包含各种媒体使用成语的情况。如在 2011 年的 10 亿汉字的语料中，成语出现近 200 万次，覆盖率为 0.32%。

但是如果我们仔细观察上述对成语的定义，我们发现其只是一个描述性的定义，存在模糊性。如何给汉语成语下一个精确的定义，无论是从语义、语法还是语用的角度，一直都是一个难题。在引入多词表达概念之前，汉语对于一些难以明确定义为成语的这类固定结构也称为“熟语”或者“习语”。《现代汉语词典》对于“熟语”的定义为：固定的词组，只能整个应用，不能随意变动其中成分，并且往往不能按照一般的构词法来分析，如“慢条斯理、无精打采、不尴不尬、一来二去、乱七八糟、八九不离十等。”[6]在一部有关“习语”的专著中，将其定义为：一种多词的语言单位，常为习惯用法，具有相对固定的句法-—语义结构。语言的使用者惯于将它作为一个整体来用，以增强语体效果。总体上，语言学家们对于熟语的一些特点达成了共识。文献[7]认为，熟语是语言中定型的词组和句子，使用时一般不能任意改变其组织，包括成语、谚语、格言、歇后语等。文献[8]认为，词汇当中，除了许多独立运用的词以外，还有一些固定词组为一般人所经常使用的，也作为语言的建筑材料和词汇的组成部分,这些总称熟语。熟语的范围相当广，包括惯用语、成语、歇后语、谚语、格言等。。

从以上描述中我们看到，“固定性”是这类语法结构的共同特点，而且熟语应该包含成语。不承认熟语（或按英语称为“多词表达”）的固定性，在自然语言处理任务中会出现很多问题。例如对汉语文本中的句子进行分词，一些成语或者熟语如果按照组成成分进行切分和标注，将会给理解造成很大的困难。例如汉语成语“鸡飞狗跳”，我们用 ICTCLAS1进行切分并标注会得到以下结果：

鸡/n　飞/v　狗/n　跳/v

而实际上这个成语只是利用两种动物“鸡”和“狗”来进行比喻，本身并非和这两种动物相关，把它切分开来会让人觉得这个成语的语义和这两种动物相关。再如谚语“只要功夫深，铁杵

[1] www.ictclas.org

磨成针。”同样切分后的结果为：

只要/c　功夫/n　深/a　，/w

铁杵/n　磨/v　成/v　针/n　。/w

也容易让人无法得到该谚语的真正含义。

问题在于熟语和成语的界限究竟在哪里？我们认为，汉语成语的定义应该符合国际通用的对成语的定义[9]：An idiom is a multi-word expression that has a figurative meaning that is comprehended in regard to a common use of that expression that is separate from the literal meaning or definition of the words of which it is made.如其所言，是否归入成语关键是该多词表达的语义不能从其组成成分——无论是字还是词——中推测出来，亦即无法从成语的字面知道其比喻义。这样汉语中“杯弓蛇影”为成语，而“兴高采烈”则不是。

3 现代汉语成语知识库的建设理念

人学习第二语言要掌握大量语法、语义知识，让计算机理解人类语言，也要给计算机配备语言知识库，使之成为计算机处理语言的知识基础和依据。而给计算机用的语言知识与给人学习的语言知识是要有区别的。针对利用计算机对自然语言进行处理，主要要解决三个问题：一、计算机需要什么样的语言知识？二、怎样描述这些语言知识，计算机才能接受？三、如何建设实用型语言知识库以便让计算机能够方便地处理这些知识？

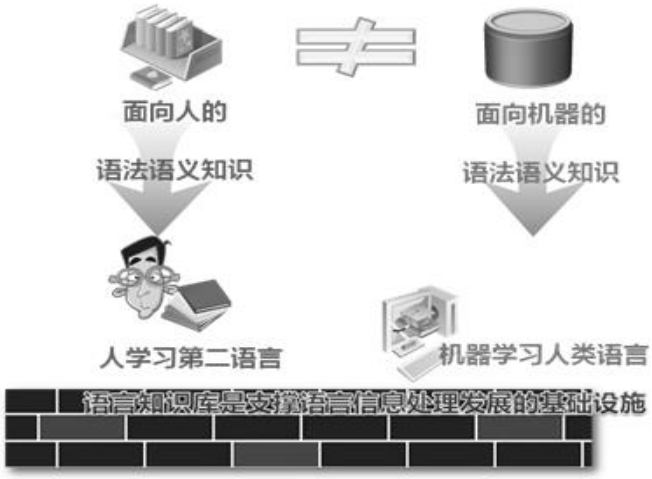

图 1　语言知识库是支撑语言信息处理发展的基础设施

在自然语言处理领域中，语言知识库就好比人类大脑中存储语言知识的记忆区域，是支撑语言信息处理发展的基础设施（如图 1 所示）。恰当的知识表示、快速有效的存储与读取机制、灵活高效的算法等都构成了计算机用语言知识库的必要要素。因此语言知识库是自然语言处理系统中不可或缺的组成部分，是这类系统成败的关键。在用语言知识库搭建的平台上可以上演威武雄壮生动活泼的应用系统的剧目（图 2）。

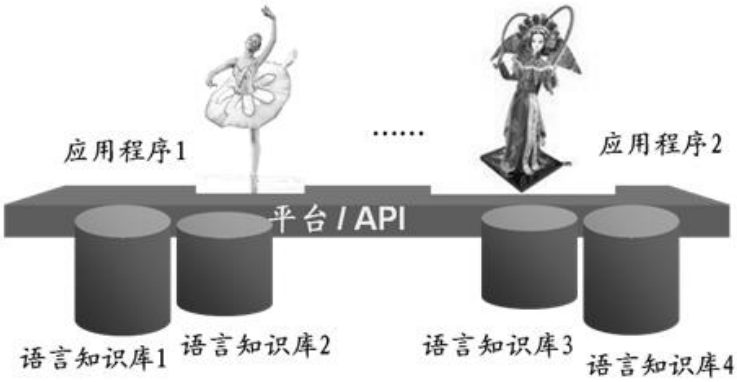

图 2 应用程序需要建立在语言知识库的平台上

从上世纪八十年代起二十六年来，北京大学计算语言学研究所（以下简称“计算语言学所”）立足北大文理结合的基础，发挥对母语知识和文化的认知优势，日积月累，终于建成综合型语言知识库（Comprehensive Language Knowledge Base，以下简称“CLKB”）。CLKB 的语言知识覆盖词、词组、句子、篇章各级语言单位和词法、句法、语义各个知识层面，从汉语向多语言辐射，从通用领域深入到专业领域[10]。一直以来，综合型语言知识库没有停止发展。已有的知识库的瑕疵不断被剔除，质量不断提高。新的知识库也在建造中。应用也在不断推进[2]。

对于针对汉语成语构建相应的语言知识库，俞士汶教授曾指出：“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，建成了一个汉语成语知识库，并基于该成语知识库开展了多词表达、比较语言学以及计算机辅助翻译方法的研究。在国家973课题的支持下，计算语言所构建了汉语成语知识库（Chinese Idiom Knowledge Base，以下简称“CIKB”）。

4 成长中的成语知识库

计算语言学所构建的汉语成语知识库，其发展历程共分三个阶段。第一阶段（1986 年—2003 年）所搜集标注的成语是作为《现代汉语语法信息词典》[3]（以下简称“语法信息词典”）的组成部分。当时《语法信息词典》收了 8 万余汉语词语，其中包含的成语和习语共有 9000 多条（见图 3）。清华大学出版社出版了介绍这部电子词典的专著[11]。

[2]自 1996 年起，正式签署对外转让协议，至今已持续 16 年。2001 年又签署了好几个协议，包括浙江大学、北京理工大学、解放军信息工程大学、韩国首尔大学，SONY 公司正在洽谈中。

[3] 《现代汉语语法信息词典》是一部面向语言信息处理的大型电子词典。它按照语法功能和意义相结合的准则收录了 7.3 万余词语。依照语法功能分布的原则，建立了词类体系，完成了这 7.3 万词语的归类。并在此基础上，分类描述每个词语的各种语法属性[12]。

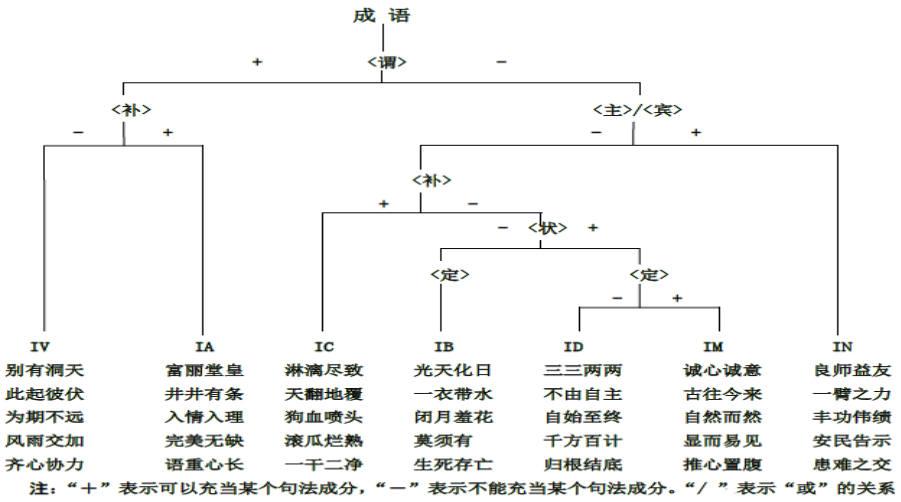

图 3 《语法信息词典》中所收录的成语语法属性标注

发展的第二阶段（2004 年—2009 年）是在 973 课题中所提出的“综合型语言知识库”的规划下，单独建库。收入成语及熟语 36000 多条。其中除《语法信息词典》原有的“主语”、“谓语”、“句首”等句法属性信息外，增设了 11 个新的属性字段：成语、级别、变体、近义、反义、直译、意译、英语近似、译者、释义、词频、来源。至此，成语知识库中共计有 23 个属性字段。特别是“直译”、“意译”、“英语近似” 字段（见图 4）既重要，又难填。现已完成 1 万多条成语完整的属性值填写，其中英语字段自译的就有 2600 多条。

成语：自暴自弃
直译： **to expose and throw oneself away**
意译： **to abandon oneself to despair**
英语近似： **to cut off one's nose to spite one's face**
近义：妄自菲薄、自惭形秽
反义：妄自尊大、自高自大、自命不凡

成语：开门揖盗
直译： **to open the door and invite robbers**
意译： **to invite disaster by letting in evildoers**
英语近似： **Opportunity makes the thief**
近义：自讨苦吃、引狼入室
反义：关门打狗

图 4 成语知识库中英译字段标注示例

发展的第三阶段（2010 年—至今）缘于计算语言学所与台湾元智大学合作的“历代语言知识库建置”计划。自 2010 年加入“历代语言知识库建置”计划起，成语知识库的发展有了更开阔的视野，更加重视成语知识库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。其进展如下：（1）成语知识库与历代语言知识库中的另一成果“诗词曲典故资料库”进行连接，相互参照。两个知识库中实现成员的优势互补，提高知识库整体品格。例如条目“倾城倾国”经过与典故资料库进行影射可以得到如“倾国风流、一顾倾城、名花倾国、倾人城、倾城色、倾城国”等古诗词和文学作品中的典故。成语与典故的对应使两个自立的知识库交相辉映，可以让学习者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）建设了成语典故分阶多语教学网站。网站的内容比较丰富，例如进行了成语形态对比与教学关系的探讨4（见图 5）。汉语的“走马看花”，韩语是“走马看山”，汉语的

[4]主要是在与汉语亲缘关系相近的日语和韩语进行教学实践活动。

“异口同声”日语是“异口同音”，词汇组成成分不同。汉语的“堂堂正正”，而日、韩语中的形态是“正正堂堂”，意义相同而词序不同；日语同时用[正々堂々]的写法,读音则为“せいせいどうどう(seiseidoudou)”；韩语的写法“正正堂堂”，读音则是“정정당당(jung jung dang dang)”。

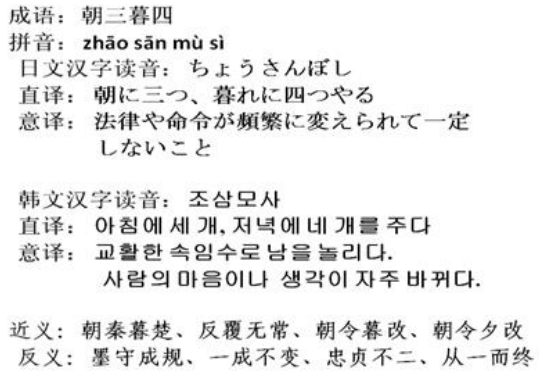

图 5 成语知识库中多语形态比较示例

（3）进行了基于成语知识库的汉语成语教学实践活动。其中包括王雷著《中国成语 1000（汉英对照）》[13]以及发表的相关汉语成语知识库与汉语教学的论文[1]。

5 结语与未来研究

目前，无论是从人的角度还是从机器的角度，成语的理解与运用还存在一定的困难。例如，成语中包含的非常用字：另辟蹊径、高屋建瓴、言简意赅、锱铢必较、罄竹难书……；含费解的词：膏火自煎、乌合之众、独具匠心、固若金汤、 司空见惯、格物致知……；隐喻的广泛使用：洛阳纸贵、罄竹难书、一丝不苟、金屋藏娇等。一些成语与历史典故关系密切，在应用时非常依赖语境，稍加不注意就可能造成应用不当甚至是错误。例如：胸有成竹、金屋藏娇、朝三暮四、杯弓蛇影、班门弄斧……等等。

基于成语知识库所开展的研究可以分为两个角度，从小视野来看主要是成语的理解与运用，尤其是面向中文信息处理的应用，从而做到既面向机器又面向人，以面向人的研究为基础，以机器自动理解为最终目标，两者相辅相成、相互促进。从大视野来看，则须紧扣历代语言知识库的构建，对历代汉语语言知识进行深层次的分析和研究，探索汉语言演化规律与社会环境变迁的交互影响。

为了支持成语知识库继续发展，计算语言学所也制定了一些新计划，其中包括：1）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）；2）中国国家自然科学基金项目“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）；3）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）。

成语知识库是一项已历时二十余年的大型语言工程，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。同时，成语知识库又是一项知识密集型的高级语言工程。单纯依赖自动技术建构的语言知识库的质量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。

致谢

本研究工作得到国家自然科学基金（项目编号 61170163，61272221, 蒋经国基金会(2009)以及北京大学计算语言学教育部重点实验室开放课题（项目编号 201302）。得到国家高科技研究与发展项目（863 项目）（项目编号 2012AA011101）部分支持。

参考文献

[1] 王雷，俞士汶，朱学锋，罗凤珠，汉语成语知识库与汉语教学[A]，第八届中文电化教学国际研讨会论文集，第83-89页，2012

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

[3] 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.

[4] 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.

[5] Fiedler, S.. English Phraseology: A Coursebook[M]. Turbingen: Gunter Narr Verlag(2007).

[6] 社科院词典编辑室，现代汉语词典（第三版）[M]，北京：商务印书馆，1998

[7] 辞海编辑委员会，辞海（1979年版）[M]，上海：上海辞书出版社，1979

[8] 胡裕树. 现代汉语[M] . 上海: 上海教育出版社,1998.

[9] McArthur, Tom. 1992. The Oxford Companion to the English Language[M]. Oxford University Press, Oxford, UK.

[10]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.

[11]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.

[12]中国工程院编，《20世纪我国重大工程技术成就》[M]，广州：暨南大学出版社，2002年，第一版31页

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