文章编号：1003-0077（2017）00-0000-00

现代汉语形容词资源库的构建*

饶 琪[1,2]，王厚峰[1,2]，汪梦翔[1,2]，李 慧[3]

（1. 北京大学 计算语言学教育部重点实验室，北京 100871；

2. 北京大学 计算语言学研究所，北京 100871；

3. 中国社会科学院 语言文字应用系 北京 100971）

摘要：形容词与名词、动词构成汉语实词的主体组成部分，在句法上表现出了对“名词”的极度依赖，其核心功能是在概念层面上，在认知注意机制的调适作用下对名词的特征进行“评价”。该文主要报告汉语形容词知识库构建相关的工作。首先是考察已有的形容词的收词情况，并结合语言演变中新产生的形容词，构建了一个较为全面的形容词词集；其次是详细阐述知识库的构建理念；再次是具体阐述知识库的特征描述体系；最后是对该知识库的应用场景进行展望。

关键词： 形容词；联想；知识库

中图分类号：TP391 **文献标识码：**A

Abstract: Adjectives, nouns, and verbs are basic subcategories of content words in Chinese, while syntactically t he nouns are more relied than other words, which is to evaluate the characteristics of the nouns under the effect of adjustment of cognitive attention mechanism with core function lying in concepts. This paper demonstrates the ach ievements relevant to the construction of the repository of Chinese adjectives. It first investigates the recording of previous adjectives as combining the new adjectives generated in the course of language change to build a relativel y comprehensive album of adjectives. Later it elaborates the philosophy of the construction of this repository and t hen elucidates the characteristic descriptive system of the repository. Finally, it prospects the application of the rep ository.

1. 引言

自然语言处理研究的诸多分支领域中，知识库的构建无疑具有基石意义，是整个自然语言处理研究系统架构中不可或缺的组成部分。在某种程度上来说，知识库的规模与质量很大程度上决定了自然语言处理系统的成败[1]，这已成为自然语言处理技术研究者和系统开发者的共识。但对于语言资源建设最为核心的问题：构建一个什么样的知识库以及怎样构建？不同的语言知识工程却存在显著分歧，这可以从普林斯顿大学研发的“WordNet”、加州伯克利分校的“FrameNet”、麻省理工学院的“ConceptNet”等英语世界里最具有代表性的三大知识工程的构建理念与细节方面清晰看出。

汉语资源建设情况具有相似性，近年来面向中文信息处理的大型知识库构建工作不断得到推进，先后形成了若干具有代表性的大型知识工程：如知网（HowNet）、同义词词林（扩展版）、北京大学综合型语言知识库（GLKB）等。但它们却有着各自的构建侧重点：1）HowNet 致力于通过汉语来全局刻画人类的概念体系；2）“同义词词林”主要是从“语义”

* **基金项目：**国家 863 项目（2015AA015402）；中国博士后科学基金 61 批面上项目（2017M610004）；国家自然科学基金（61602040）

的角度来实现汉语词汇“同义”汇集；3）GLKB 则是以“词类”为纲，主要是描述“词”的各种语法信息。这三大汉语知识库都是以“词”为基础构建对象的。词是语言的基本单位，也是“短语”、“小句”、“篇章”等更大语言单位衍生的基础，为这些语言单位意义的“在线构建”提供“有源之水”。认知神经研究也表明了“词”在心理词库的表征与长时记忆存储、提取中的基础地位（杨亦鸣等，2006）[2]，事件电位相关技术（ERPs）的实证调查也表现了对这一观点的支持（张珊珊，2010/2012）[3]。以“词”为基础表示单元，为汉语构建一个覆盖面大、加工精良的汉语知识库，无疑能够有效推动面向汉语的智能问答、文本理解、文本生成等多方面研究的深入。

虚词与实词是汉语词类的基础两分体系。在汉语虚词知识库方面，郑州大学构建了一个涵盖“副词、介词、连词、助词、语气词、方位词”等六类词，包括词典、虚词用法规则库、虚词语料库等组成部分的三位一体的汉语虚词知识库[4]。在实词知识库方面，名词、动词是汉语知识库构建的重点方面，对形容词的关注显得还不够。具体的说，同义词词林（扩展版）和北大综合型知识库（GLKB）虽然也对形容词有所涉及，但“同义词词林”主要是从“语义”的角度揭示形容词是如何组织的，GLKB 中的形容词部分更多追求是对“词例”各种语法功能信息的展示。还缺乏对形容词全景知识图景的展示，更为遗憾的是，截止目前为止还未见到专门的形容词知识库构建情况的报道。本文主要讨论与现代汉语形容词知识库（以下简称 PAKB）构建相关的问题。

2. 相关问题

2.1 形容词知识库构建的目标

在知识库构建的总体目标上，我们不仅追求知识库对中文信息处理的推动作用，也特别在意知识库在汉语本体研究中的基础平台意义。从目前已有的汉语大规模知识库来看，这些知识库构建最初的出发点与落脚点都是适用于中文信息处理，在语言学本体研究领域中表现出了适用面过窄的特征。但是需要看到的是汉语的形容词研究也积攒了丰富的学术成果，也有不少的下一步研究需要在一个“公共”的资源平台上进行比较，我们所构建的形容词知识库也希望能够成为汉语本体研究领域中一个可供比较的、基础资源平台。比如说，形容词的重叠现象是汉语本体研究中的一个重要问题，在现代汉语中，形容词的重叠类型有 AA、ABB、A 里 AB、AABB、BBAA 类型，以 ABB 型形容词为例，《现代汉语八百词》的附录部分《形容词生动形式表》、《现代汉语词典》、《重叠形容词用法例释》在收录 ABB 形容词方面就存在差异，具体情况可见下表：

表 1：三种不同辞书 ABB 型形容词的收词情况

词型	《现代汉语八百词》	《现汉》	《重叠形容词例释》
ABB	312	201	483

从表1中可以清晰看出,这三种辞书在收录ABB型形容词数量方面是存在较大的差异的。但无论以何种视角来介入汉语形容词重叠问题研究,对重叠在现象层面上的观察与把握都是首要的问题，如果每一位研究者都从调查。构建词表开始入手，无疑会费时费力，研究的结论也缺乏相互的可比性。在PAKB的构建目标上很重要的一条就是提供可以用来比较研究的基础资源集。

2.2 形容词集构建

在开始讨论如何构建汉语形容词知识库之前，有一个前提性问题需要首先予以回答：现代汉语中到底有哪些形容词？一个对现代汉语语料高覆盖率、完善的形容词词集是构建汉语形容词资源库的首要前提。为了回答这一问题，我们进行了分“两步走”的工作：

第一步是“求全”。上述问题的答案最显然来源于各类辞书，首先调查了目前出版的两部专门形容词词典：郑怀德、孟庆海编撰的《汉语形容词用法词典》收形容词1067条[5]；安汝磐、赵玉玲编撰的《新编汉语形容词词典》收词2268条[6]，整体收词规模较小。其次利用《现代汉语词典》（以下简称《现汉》）带词性标注的特征，以《现汉》(第7版)为蓝本，对词典中所收录的形容词进行了人工整理，共得到形容词5069条；同时也考察了《现代汉语语法信息词典详解》对形容词的收录情况，在该书中形容词被细分为形容词、状态词和区别词三个子类，分别收有形容词1473个、状态词203个；区别词194个。综合这四种工具书对“形容词”收词情况，取它们的合集作为构建汉语形容词知识库的词条基础。

第二步是“补全”。应该看到，任何一种汉语工具书囿于其自身的局限，在事实上难以穷尽枚举日常语言生活中所有的词。与此同时，行进中的语言演变也会造成“词汇总藏”中新成员的涌现,其中的一部分留存到语言中来,就是得到语言的过程(Steven Pinker,2013)。就形容词来说，其中的一个子类“区别词”（也称之为“非谓形容词”或《现汉》词性标注体系中的“属性词”），如“大型、中型、小型、大中型、中小型”等是汉语新词的一个重要“出生地”（吕淑湘、饶长溶，1981），其繁殖率仅次于名词[7]。新的形容词与已有的它类词扩张出形容词用法是现代汉语形容词词集版图扩大的两条最主要途径,下面各举一例略加说明：先说新的形容词。如：【结构化】

在《现汉》的“结”字头下共收录有53个词，未有收录“结构化”一词。“结构化”一词指的是在思考分析解决问题时，以一定的范式或者流程顺序进行，以假设为先导，对问题进行正确的界定,假设并罗列问题构成的要素,其次对要素进行合理分类,排除非关键分类,对重点分类进行分析，寻找对策，制订行动计划。如下面几例：

（1）广东教师招聘结构化面试模拟题(29)：如何遏制幼儿园暴力事件。（http://gd.offcn.com）

（2）目前各种类型的结构化金融产品的规模已经达到了十多万亿元，并且这种结构化产品的设计思路，在鼓励民间资金进入基础设施领域的PPP投融资模式中得到了进一步推广。（http://opinion.jrj.com.cn）

（3）人行长春支行举办“我与行长面对面”结构化研讨活动。（http://finance.jrj.com.cn/）

在上面三例中，“结构化”均是属性形容词，使用在“面试”、“产品”、“研讨”等名词前头，对这些名词进行次范畴的分类，用来凸显与强调了这三个名词的就有一定的范式、或者按照一定的流程进行的特征。

次说已有词形容词用法的涌现。如：【旗舰】

（4） 第一个屏幕下指纹识别？三星新**旗舰**机 Galaxy Note 8。
（新浪手机，2017-6-8）

（5） 吉利新款**旗舰**轿车最新谍照，年内将上市（新浪汽车，2017-6-8）

（6） 首家全系列、全品类穗宝**旗舰**店国庆节盛大开业（房天下，2016-12-1）

（7） CCL 是国内最大的自然语言处理专家学者的社团组织——中国中文信息学会（CIPS）的**旗舰**会议，全国计算语言学会议从 1991 年开始每两年举办一次，从 2013 年开始每年举办一次，经过 20 余年的发展历程，已形成了十分广泛的学术影响，成为国内自然语言处理领域权威性最高、口碑最好、规模最大的学术会议，今年注册参会人数超过 500 人。（http://www.scholat.com/vpost.）

在过去，“旗舰”是一个名词，指的是海军舰队司令、编队司令所在的军舰。舰队一般而言是由多所军舰构成的集合，“旗舰”的名词语义体现的是该军舰在整个舰艇集中的重要性。但近年来，“旗舰”可以与部分名词组配，如上几例中的“～机”、“～轿车”、“～店”、“～会议”。对于生产厂商来说，生产的“手机”、“轿车”也是一个集合，通常是多种多款，但它们的重要性并不一致，“旗舰”与“手机”、“轿车”、“店”、“会议”等名词组配，实际上是对这款“手机”、“轿车”在整个产品集中重要性的一种评价，其实这也是人类类比认知能力对“旗舰：军舰”关系对不同名词域的扩展，这个步骤如下：

1） 具体的：[旗舰：军舰] 刻画了旗舰在{ 军舰$_1$军舰$_2$, 军舰$_W$,…}中的重要性

2） 类比关系的转域：[旗舰：军舰] 映射到 CCL 在中文信息处理学会举办的会议集中地位、价值中来；

3） 域的扩张：专卖店、手机、轿车、会议；

4） 用法的习得：评价 X 在{X_1, X_2, X_W,…}中的重要性

从旗舰店到旗舰会议，“旗舰”的形容词用法在广泛使用的过程中得到不断的强化，从而沉淀于汉语之中。最近十几年来，汉语的载体形式发生了颠覆性的改变，网络媒体正在日益成为汉语的一种重要载体形式，中国互联网信息中心 2017 年发布的《中国互联网发展状况统计报告》（第 39 次）显示，截止 2016 年 12 月，中国网民规模已达 7.31 亿，互联网普及率为 53.2%。粘性极高的交互性互联网应用为互联网用户提供了高强化的汉语阅读机会与规约度降低的表达空间。活跃于互联网空间的“新词”与“旧词新用”给中文信息处理带来了新的挑战，这意味着我们需要在知识库构建上能够有效对这些崭新的语言事实予以追踪。我们对最近十年来的新词年度报道类工具书《汉语新词语》（2006～2015）进行了全面的考察[8-9]，手工遴选出了近年来产生的形容词 98 个，列表如下：

表 2:汉语形容词(2006～2015)

音节	词例（Token）
单音节	帅 萌 裸 潮 水 火 酷 囧 靓 面 牛 爽 铁 炫 妖 拽 雷
双音节	黄金 白金 黑金 八卦 憋屈 超值 扯淡 出彩 到位 低调 低迷 丰俏 丰挺 感冒 搞笑 骨感 杯具 花心 火爆 娇挺 矫情 紧俏 紧实 劲道 劲爆 惊艳 拉风 老套 老土 雷人 靓丽 灵动 另类 卖座 叫座 闷骚 内敛 拧巴 疲软 前卫 抢手 抢眼 轻薄 轻闲 缺失 热辣 热销 山寨 煽情 闪亮 上镜 生猛 时尚 舒爽 爽透 私密 酸楚 威猛 喜剧 小资 新锐 性感 休闲 阳光 养眼 拥堵 有序 晕菜 知性 走俏 绿色 有机 旗舰 较真儿 掉份儿 掉价儿
三音节	超白金 次顶级 结构化 标准化 程序化 旗舰级

结合这两个步骤的工作，得到了一个含有 5671 条词条的形容词词集。为了进一步验证这一词汇集的规模，我们使用了清华大学研发的中文词法分析工具包 THULAC，该分词包具有分词、词性标注一体化特征。在出版物和互联网两类载体形式语料上进行了覆盖率的考察，情况如下：

表 3：三类语料中形容词词表覆盖情况

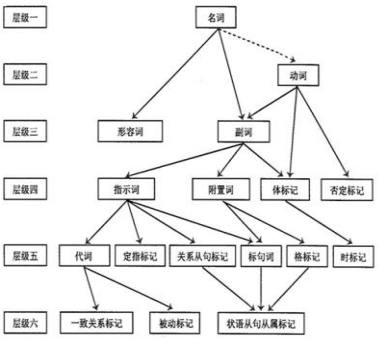
语料	覆盖率
现当代小说语料（131M）	98.1%
搜狗新闻语料（历史版本）（130M）	97.3%
新浪 微博语料（128M）	92.1%

尽管严格意义上来讲，我们构建的形容词词库也没有做到完全的覆盖，但已经在不同类型的语料下跨越了 90%的覆盖率阈值，表现出了针对不同语料的一定的适用性。进一步分析原因，主要是没有收录若干强势方言中，典型常用的形容词，如“苕、二、尖货”等。

2.3 现代汉语形容词的兼类

通常意义上，名词、动词、形容词是汉语的三大类实词，同时也是汉语从“词库”到“句法”得以实现的骨架力量。需要指出的是，部分形容词存在兼类现象，与形容词发生兼类现象的主要是名词、动词。比如，“超前”在现代汉语中就兼有形、动两类词的特征：动词性的如“～绝后”等，形容词性的如“～消费、～意识、～教育”；“绿色”有名、形两类词的标签，作为名词的“绿色”指的是“绿的颜色”；而作为形容词的“绿色”，通常指的是符合环保要求，无公害，无污染或简便、安全、快捷的途径或渠道，如“～食品、～经济、～通道”等。这在另一个侧面也说明了形容词与名词、动词之间的天然联系。这一点能够得到

人类语言中词类演化的证据支持，Heine and kuteva（2007）对跨语言中词类的演化进行了模拟[10]，如下图所示：



图一：词类范畴演化图

从上图中，可以清晰看出，作为词类的形容词处于名词的下级节点中。换句话说，从历史来源的角度上来讲，语言中形容词的涌现是早期名词分化的后果。与此同时，郭伏良（1983）就注意到，“端正、丰富、密切、孤立、健全、状大”等词在 20 世纪 40 年代的汉语中只有形容词义项，但在 20 世纪五六十年代开始起就常用作动词。我们也注意到，20 世纪八十年代以降，“潇洒、清洁、方便、规范、完善、突出”等词具有了动词用法。现代汉语中到底有哪些形容词存在兼类现象，是和名词发生兼类、还是和动词发生兼类，或者是与名词、动词均发生兼类？厘清这些语言事实，对汉语本体研究以及词性自动标注问题都具有十分重要意义。因此在构建汉语形容词知识库过程中，我们极其注重形容词的兼类现象信息的揭示。

2.4 形容词的子类与核心功能

词类子类的出现体现与反映了研究者对该词类认识的深入。在汉语研究文献中，存在有不少的术语来表征形容词的子类体系，如简单形容词、复杂形容词的两分（朱德熙，1956）[11]；一般形容词、非谓形容词的两分（吕叔湘、饶长溶，1981）；性质形容词与状态形容词的两分（朱德熙，1982）。尽管这些术语在表述上有参差，但都清晰的指出了形容词内部存在有差异，并且这种差异是可以得到真实文本里句法上的验证。比如非谓形容词（区别词）在句法层面上一般只能做定语，如“活期存款”中的“活期”，在句法上如果要进入谓语的位置，是需要存在于“是…的”构式之中的，并且在句法上有一个重要的约束条件，就是它前头出现的否定词只能是“非”，不能是“不”。这些充分说明在 PAK 知识库构建的过程中需要对具体个案词条的子类打上标签。

长久以来，在词类本质问题认识上，较多考虑的是句法上的分布式特征，而对同类词在语义层面的共性缺乏足够的讨论。事实上，词类一方面与句法分布表征有着密切关联，另一方面也与语义类存在的对应关系，这种对应背后体现并反映了人类对不同语义类语法编码的共性认知基础。Dixon（1977）所提出了鉴定形容词的三条标准：1）与动词和名词有语法上的区别；2）语义上包括部分或全部典型形容词的语义类型，如“维度、年龄、价值、颜色”等；3）具有充当不及物谓语和/或充任系词补语、名词短语的修饰语的功能。Croft（1991）

在此基础上将句法范畴、语义和语用功能结合起来，提出了句法范畴的原型关联。如下表所示：

表 4：句法范畴的原型关联（引自 Croft,1991）

句法范畴	名词（Noun）	形容词（Adjective）	动词(Verb)
语义类	事物 Object	属性 Property	动作 Action
语用功能	指称 Reference	修饰 Modification	陈述 Predication

对汉语形容词的本质认识问题，在部分承认 Croft（1991）观点基础上，对汉语形容词提出了如下的主张性认识：

1）句法上，汉语的形容词与名词具有天然的联系。在句法层面上，形容词几乎离不开名词，单独的形容词不具备成句的功能，除非出现问答对子中；

2）语义上，形容词是对名词多维属性中的一个侧面的刻画，对名词多维侧面中的某一维进行评价，如“大房子”中的“大”就是对具有多维属性的“房子”在空间维度上进行评价，这是形容词的核心本质；

4）语用上，形容词多具有情感性。正面、负面两分并不能很好的传导出形容词的情感，部分形容词的情感体现的是话语的态度，如“A 里 AB”类形容词多体现话语的言说者埋怨、责怪的语气；

5）非谓形容词是以团簇的方式存在，在语用上主要是实现对名词的再分类，起到次范畴化的效果，如“男、女”永远是相对存在。

3.形容词知识库的表示体系

词具有多种信息标签。以往研究主要关注词的形式和意义两端，这种观察无疑生发在静态层面，而动态的语用通常会赋予词几何维度上的信息，从而建构起词的整体知识图景。词的信息标签以外显和内隐的方式存在：外显是能够被直接感知的，如韵律、结构、高熟悉度的语义等信息；内隐是需要进一步挖掘才能获得的信息，如频率、情感、语体、极性等信息。这也是我们所需要知道和最大程度上试图表示出来的信息，同时也是计算机所需要配备用来学习的知识库。PAKB 试图从多个层面来展示汉语形容词的知识全景。

3.1 形式特征集

PAKB 对现代汉语形容词形式层面特征的刻画主要包括语音、音节数、重叠形式、语义、重叠情感、语体表现等六个方面。在汉语的形容词词汇集，不少的形容词存在“语体”使用偏置现象，胡明扬(1995)从语体风格方面区分了形容词在口语和书面语中功能上的差异。比如“哀戚”，就是一个典型的只使用于汉语书面语的形容词，但更多的形容词表现出书面语、口语两可的分布，如“哀伤”。有关形容词“语体”信息的标签是已有的汉语大规模知识库所未有刻画过的，同时也可以自动问答研究提供有效的口语形容词汇集。综合起来，汉语形容词知识库的形式特征集刻画示例如下图：

	A	B	C	D	E	F	G	H	I
1	词语	拼音	义项	音节数	重叠	重叠形式	重叠情感	语义	语体
2	哀愁	āichóu	1	双音节	不能		负面	悲哀、忧愁	书面语
3	哀戚	āiqī	1	双音节	不能		负面	悲痛伤感	书面语
4	哀伤	āishāng	1	双音节	不能		负面	悲痛忧伤	书面语
5	哀痛	āitòng	1	双音节	不能		负面	哀伤,悲痛	书面语
6	哀婉	āiwǎn	1	双音节	不能		负面	悲伤婉转	书面语
7	哀怨	āiyuàn	1	双音节	不能		负面	哀婉愁怨	书面语
8	嗳哟	āiāi	1	双音节	不能		中性	形容霜、雪洁白	书面语+口语
9	矮	ǎi	1	单音节	能	AA	负面	身材短，高度小，级别、地位低	书面语
10	矮矮墩墩	ǎiaiīdūndūn	1	四音节	不能		负面	身材矮胖的样子	书面语
11	矮墩墩	ǎidūndūn	1	三音节	不能		负面	过于粗胖的样子,矮胖得难看的样子	书面语
12	矮胖	ǎipàng	1	双音节	能	AABB/ABB	负面	又矮又胖	书面语+口语
13	矮小	ǎixiǎo	1	双音节	不能		负面	又矮又小	书面语+口语
14	碍事	àishi	1	双音节	不能		负面	指使人不方便，或妨碍人。	书面语+口语
15	碍眼	àiyǎn	1	双音节	不能		负面	1）看着不舒服；不顺眼。东西乱堆在那里很碍眼的。2）妨碍别人的事情,使人感到不方便模	
16	暧昧	àimèi	1	双音节	不能		负面	1）（态度、用意）含糊，不明白；	书面语+口语
17	暧昧	àimèi	2	双音节	不能		负面	不光明，不可告人的关系	书面语+口语
18	安安静静	ānānjìngjìng	1	四音节	不能		正面	性格安稳，平静，不受打扰	书面语+口语
19	安安生生	ānānshēngshēng	1	四音节	不能	AABB	正面	形容生活平稳，安定	书面语+口语
20	安安稳稳	ānānwěnwěn	1	四音节	不能	AABB	正面	形容生活安稳，不若是生非	书面语+口语

图 2：汉语形容词知识库形式特征示例（前 20）

3.2 基于名词的形容词组织与表征

在上文，我们提出了形容词的核心本质是对名词的某一侧面维度进行评价。进一步在跨语言的角度来观察，不同语言里形容词数量上存在多寡的差异。尼日利亚的伊博语是已报道出来的形容词数量最少的语言，只有八个，分别是“大”、“小”、“黑(暗)”、“白(明)”、“新”、“老”、“好”、“坏”。据 Bhat(2000）考察显示在形容词数量较少的语言中，如 Supyire 语 10 个，Bamha 语约 20 个，Luganda 语约 30 个，Acoli 语约 40 个，Kilivila 语约 50 个，Sange 语约有 60 个。尽管这些语言形容词数量较少，但仍然对名词评价、刻画了如下的属性：

1） 维度：大、小；高、低；宽、窄；深、浅；长、短；粗、细；厚、薄；

2） 价值：好、坏、纯洁、好吃

3） 年纪：新、老(旧) 、小(年轻)

4） 物理属性：硬、重、光滑

5） 颜色：红、白、黑

6） 速度：快、慢、迅速

这说明了这些抽象概念的表征具有跨语言的共性，这些有限的抽象概念可能是词汇组织与表征的重要指针。大脑究竟如何安置词汇和概念？ 最近的一项研究利用 985 个英语常用词汇来绘制大脑的“语义地图”(Huth. et al.,2016)，这项研究表明，并不存在一个单独的大脑区域来储存一个词汇或者概念与许多相关词汇存在联系，而是每一个单独词汇会点亮许多不同的大脑位置，形成了一张词汇会聚网络。研究结果一共识别出 12 个簇群(clusters)，其中每个簇群均保存着与特定概念相对应的词语，这些词语以相关的方式存在。比如，大脑左边，耳朵小面积区域代表着单词“受害人（victim)”，同时这块区域会对诸如 [杀害（killed)]、[宣告有罪（convicted)]、[谋杀(murdered)]、[认罪（confessed)] 有反应。

从早期汉语的“幽、黄、黑、白、赤、大、小、多、少、新、旧、高”等 12 个单音形容词（杨逢彬，2001）到今天汉语里面数量几千的形容词，书面语的高度发达催生了汉语中形容词数量几何级数的增长。在 PAKB 的形容词如何分类组织的问题上，我们是以名词为观察视点，将表征与共享了相同“概念空间”的形容词看成是“自组织”性的簇。举个例子，汉语里面存在有数量众多的形容女性外貌的词，如单音节的“美”；双音节“美丽，好看，漂亮”；四音节的“楚楚动人、闭月羞花、沉鱼落雁、冰清玉洁、粉妆玉琢、国色天香、国色天姿、惊鸿一瞥、明眸皓齿、明眸善睐”等（限于篇幅，不能够列举出所有的词)。这些形容女子外貌的词构成一个自组织的集，“美”是这个集合中是最常用的代表者。与形容女

子外貌的集相比较，汉语里面用来形容男子外貌的词在数量上就要少得多，如单音节的“帅”；双音节的有“英俊、潇洒”，以及通用性的“好看”；四音节的有“一表人才、眉清目秀、气宇轩昂、风流倜傥、高大威猛 、温文尔雅”等 ，在这些词语中，“帅”是该集合的代表。

我们以常用的现代汉语单音形容词，以及 PAKB 中形容词解释的元语言作为指针。同时也结合了认知中注意力机制（Attention Mechanism），需要指出的是，这里的注意力机制与通常意义上“深度学习”中的注意力机制不同，事实上，机器学习中的这一术语也借用自视觉图像认知领域。在这里引入“注意力机制”是想说明： “名词”通常具有不同的侧面维度，但汉语的使用者在观察、刻画名词的这些不同侧面的时候，总会将注意力聚焦在几个有限的维度之上。比如“房子”，人们注意的焦点一般都是“空间的大小、价格、地段价值、舒适程度”等几个维度。这几种维度的注意力将名词映射到形容词之中，就会构成“大房子、豪宅、交通方便、空气好”等“形+名”或“名+形”组配上。因此，我们在构建 PAKB 的过程中，以“名词”为观察视点，构建了一个形容词所表征的抽象概念体系：

1） 人：外貌、性格、气质、品德、情绪、态度、关系、年纪；

2） 物：价值、作用、评价、水平、垂直；

3） 事：性质、状态、结果；

4） 时间：长短、快慢、性质；

5） 空间：大小、长短、宽窄、高底、远近、深浅、厚薄；

6） 感官：视觉、味觉、嗅觉、听觉、触觉

7） 心理：哀、愁、烦、恨、羞、愧、惊、慌、骄

在上面这个分类体系下，我们对 PAKB 中所有的形容词进行了人工的分类与聚类。

3.3 搭配特征集

词汇通常会在更加抽象的语言能力层面上构建起一个涵盖范围极广的知识库，主要包括语音知识、词义知识、词类范畴知识、句法知识、形态知识以及与论元组配的可能与限制等方面。语言使用者通过基于概率的统计学习来学得这一知识库，因此这一知识库兼具有公共性和个体性：公共性指的是对于某一语言来说，这个词汇的知识库的构成是基于所有语言使用者的经验浮现，对单一的语言使用者来说具有不可逆转性；与此同时，个体对于知识库学得的情况又不尽相同，有程度的深浅和范围宽窄的区分。但个体的词汇知识库来源并服从于集体的词汇知识库。一项来自英语个人词汇知识库如何构建的研究表明：在随机游走学习过程中，词能和什么样的论元，以及与不同类型论元分布式搭配情形，在长期记忆中会以概率框架的抽象形式留下痕迹，并且与单词的频率水平呈现出正相关，在高频效应的催化下这种记忆痕迹会得到加强（D.Kemmerer.et.al，2012）。沿着这一思路，来理解汉语形容词划分会看到不一样的风景，形容词及其毗邻成分的共现刻画是汉语形容词资源库构建中重点关注的问题，具体包括两看：1)一看给定的形容词能够和什么名词组配；2）二看给定的名词能够与什么形容词组配，如下图三、四所示

	词例	搭配	类型1	类型2
1	哀愁	心情 情绪	情绪	人
2	哀戚	心情 情绪	情绪	人
3	哀伤	心情 情绪	心情	人
4	哀痛	心情 情绪	心情	人
5	哀婉	叹息	情绪	人
6	哀怨	人	气质	人
7	皑皑	白雪	状态	物
8	矮	个头、物品	空间	人、物
9	矮矮墩墩	胖子	空间	人
10	矮墩墩	胖子	空间	人
11	矮胖	男孩	空间	人
12	矮小	个子； 个头	空间	人
13	碍事	人 ；东西	结果	人、物
14	碍眼	人 ；东西	结果	人、物
15	暧昧	眼神	关系	人
16	暧昧	关系	关系	人
17	安安静静	性格	状态	事
18	安安生生	生活	状态	事
19	安安稳稳	生活	状态	事

图 3：汉语形容词知识库搭配特征示例（前 20）

	词儿	拼音	义项	释义	近义词集	相关联想词
1	阿姨	āyí	1	〈方〉母亲的姐妹。小姨		[姨夫、姨妈、家族、亲戚、长辈]n ;[投靠、抚养、拜见]v ;
2	阿姨	āyí	2	称呼跟母亲辈分相匠妇女，		[王、售票员]n ;[遇见、看见]v ;[热情、乐于助人]a
3	阿姨	āyí	3	对保育员或保姆的称保育员，保姆		[保洁、保姆]n ;[拜托、需要]v ;[认真、正直]a
4	癌症	áizhèng	1	生有恶性肿瘤的病：绝症，癌		[食管、乳腺、口腔]n ;[身患、预防、解密]v ;[可怕、夺人]a
5	艾滋病	àizībìngˋ	1	获得性免疫缺陷综合无		[病毒]n ;[感染、身患、对抗]v ;[恐怖、可怕]a
6	爱情	àiqíngˋ	1	男女相爱的感情。 恋情，恋爱，相爱，相恋，感情		[情感、伦理]n ;[渴望、寻找、期待]v ;[美好、幸福、甜蜜]a
7	爱心	àixīn	1	指关怀，爱护他人的善心，暖心，热心，关心，关怀		[名言、理论]n ;[奉献、捐献、传递]v ;[温暖、积极]a
8	安眠药	ānmiányàoˋ	1	催眠药的通称。 无		[安定]n ;[服用、依靠、吞下]v ;[困乏、疲童]a
9	岸	àn	1	江，河，湖，海等水岸边、河岸、湖岸、两岸、堤岸		[河、海、两]n ;[上、起]v ;[傲、高、异]a
10	岸	àn	2	（àn）姓。 岸边、河岸、湖岸、两岸、堤岸		[姓、氏]n ; ;
11	按键	ànjiàn	1	用手按的键；键?。 无		[虚拟、网络]n ;[按下、安装、敲击]v ;[简单、强大、兼容]a
12	案件	ànjiàn	1	有关诉讼和违法的事件，案子，罪案，案例		[刑事、诉讼、经济]n ;[审理、操控]v ;[重大]a
13	奥秘	àomì	1	深奥的尚未被认识的奥妙，奇奥，奇妙，秘密		[宇宙、外星人]n ;[探索、发掘]v ;[神秘、高深莫测]a
14	奥运会	àoyùnhuì	1	奥林匹克运动会的简亚运会，运动会		[世界、运动员、夏季、冬季]n ;[参加、竞争、拼搏]v ;[团结、友爱]a
15	八卦	bāguà	1	我国古代的一套有象无		[五行、阴阳、九宫、风水]n ;[论述、占卜]v ;[高深、神秘]a
16	把柄	bǎbǐngˋ	1	器物上便于用手拿的手把；弱点，痛处，短处等		门，刀具，器物
17	把柄	bǎbǐngˋ	2	比喻可以被人用来进过失，		[主意、凭据]n ;[操纵、霸占、抓到]v ;[看急、失落]a
18	白菜	báicài	1	一年生或二年生草本小白菜、大白菜、青菜、油菜		[蔬菜、种类]n ;[栽培、繁殖]v ;[爽脆、营养]a
19	白菜	báicài	2	特指大白菜。 大白菜，		[蔬菜、种类]n ;[栽培、繁殖]v ;[爽脆、营养]a

图 4：汉语名词联想示例（前 20）

4. PAKB 的应用

4.1 面向本体研究的基础资源平台

从业已构建现代汉语形容词知识库来看，并非所有的形容词都能够重叠。在整个形容词知识库中，观察到 1212 个形容词是可以重叠的：在 227 个单音节形容词中，可以重叠的 114 个，约占 50%；985 个双音节形容词中，309 个是重叠的，约占 31%。由于双音节形容词远远多于单音节形容词，所以总的来看，可以重叠的形容词约占形容词总数的 35%。形容词重叠问题是面向本体的汉语形容词研究中的一个重要问题，全面调查清楚汉语形容词中到底哪些是可以重叠的，哪些是不能够不能重叠，不仅是观察汉语形容词重叠式语法意义的一个基本点，也可以为后续的有关形容词研究提供可以用来比较的、基础资源平台。下面是结合北京大学的《人民日报》标注语料中 230 个形容词重叠的频率，取其中的前十位例示：

表 5：汉语形容词重叠使用情况表

	形容词	重叠频次
1	慢	290
2	好	255
3	轻	147

4	远	122
5	小	84
6	深	50
7	静	49
8	长	46
9	高	46
10	紧	45

4.2 面向自动问答的口语形容词集

近年来，自动问答已成为自然语言处理中的一个热门研究领域。在如何让计算机模拟人的进行对话问题上，已经有多种方法、手段介入。但还未有见到有报道针对性的使用了自然口语中的对话语料作为训练集，当然很可能是由于这类资源目前较为稀缺。在构建 PAKB 知识库的过程中，我们采样了 20M 的自然口语对话语料用来对知识库中形容词的语体性质进行辅助判断，在这个过程中，我们发现汉语的书面语形容词与口语形容存在着交集，也存在着差异。但与汉语书面语相比较，自然口语对话中，形容词主要存在于下面六类构式之中。

1） 说 A 不 A,说 B 不 B；

2） 那叫一个 A；

3） 要多 A 有多 A；

4） 还能再 A 点吗？

5） 要不这么 A?

6） 是有多 A?

4.3 面向基础教育领域的形容词集

语言是一个精密的逻辑自洽系统，蕴涵其间的“经济原则”提醒了这个系统不会有一个多余的词，因此严格意义上的“等价词”是不存在的。Schmitt(1998) 将词汇知识定义为六个方面：1）形式；2）词义（包括同义、反义、上下义）；3）语法图景；4）搭配信息；5）使用；6）语体风格与语域限制。对应的认为语言使用者在词汇能力分为感知能力与产出性能力，前者对应了语言理解，后者对应了语言表达。在词汇感知上包括词汇的深度和词汇量，词汇的产出则包括词汇的宽度和质量，体现了构成语篇的能力。PAKB 能够在最大程度上显示某一抽象概念空间下汉语形容词集，这对于基础教育领域中作文教学具有参考价值 。

5. 结语

过去这些年的自然语言处理研究每一次大的进展与飞跃，都再一次强调了通过人工构建的方式为计算机提供有效的语言知识库的重要性。但是，从中文信息处理的终极目标——计算机能够“理解汉语”与“表达汉语”来看，让计算机初步具有类人的语言使用能力现在来看仍是一件具有非常挑战性的事情。目前计算机处理自然语言的能力仅仅停留在“处理”层

面，还远不能达到“理解”的水平，未来的任务艰巨而充满挑战。这在一个侧面说明了，有必要对已有的汉语资源构建的理念、方式、规模与手段进行检讨，在这个意义上来说，本文的工作可以看成是一种初步的尝试，试图在局部层面上模拟人类是如何使用语言的，为计算机构建一个与人脑更接近的可以用来增强学习、预测学习的汉语形容词资源库。

参考文献

[1]俞士汶,段慧明,朱学峰等. 综合型语言知识库的建设与利用[J].中文信息处理学报,2004,18(5):1-10.

[2]杨亦鸣，张珊珊，刘涛等. 综合型语言知识库的建设与利用[J].语言科学，2006,5（3）：3-13.

[3]张珊珊,杨亦鸣. 从记忆编码加工看人脑中基本语言单位——一项基于单音节语言单位的ERPs研究[J].外语与外语教学，2012,11（2）：1-6.

[4]昝红英，张坤丽，柴玉梅等.现代汉语虚词知识库的研究[J].中文信息学报,2004，21(5):107-111.

[5]郑怀德，孟庆海.汉语形容词用法词典 [M].北京：商务印书馆,2003.

[6]郑怀德，孟庆海.新编汉语形容词词典[M].北京：经济科学出版社,2003.

[7]吕叔湘，饶长溶.试论非谓形容词[J].中国语文,1981,10（2）.81-85.

[8]周荐.2006汉语新词语[M].北京：商务印书馆,2007.

[9]候敏，周荐.2007汉语新词语[M].北京：商务印书馆,2008.

[10]B,Heine. Tania,K The Genesis Of Grammar: A Reconstruction (Studies In The Evolution Of Language) [M].oxford: Oxford University Press

[11]朱德熙.现代汉语形容词研究[J].语言研究,1956,1（1）.1-37.

[12] Alexander G. Huth, Wendy A. de Heer, Thomas L. Griffiths,et al, Semantic information in natural narrative speech is represented in complex maps that tile human cerebral cortex [J]. Nature. 532,453 - 458.