基于互联网的汉语认知属性获取及分析*

李 斌[1] 陈家骏[2] 陈小荷[1]

(1. 南京师范大学文学院 江苏 南京 210097;
2. 南京大学计算机软件新技术国家重点实验室 江苏 南京 210093)

[摘要]认知语义学强调词语的日常感知意义的描写,但传统的人工描写方式存在搜集难度大、主观性过强两大困难。本文基于互联网数据,利用知网收录的51020 个名词、27901 个动词和 12252 个形容词自动采集词语的认知属性,得到 120 多万条原始记录。对这些结果按词类进行详细的频率统计和数据分析,构建了 6000 多词条的汉语常用词语的认知属性库,尝试了夸张和反讽的自动生成。该研究可应用于计算机理解和生成修辞表达、语言教学、词典编纂和机器翻译等领域。

[关键词]认知属性;隐喻;显著度;认知计算
[中图分类号]H08 [文献标识码]A [文章编号]1003-5397(2012)03-0134-10

Collection and Analysis on Chinese Cognitive Properties Based on Web Data

LI Bin, CHEN Jiajun, CHEN Xiaohe

Abstract: Cognitive linguistics centers the description of the common cognitive meanings of words. However, to manually collect and describe the cognitive meanings are boring and subjective. In order to overcome the shortcomings of traditional researches, this paper employs the web to collect the cognitive properties of words in HowNet which contains 51020 nouns, 27901 verbs and 12252 adjectives. Over 1.2 million records are

[收稿日期] 2011-12-05

[作者简介] 李斌,南京师范大学文学院讲师,南京大学计算机系博士后,主要研究方向为计算语言学;陈家骏,南京大学计算机系教授,博导,主要研究方向为自然语言处理;陈小荷,南京师范大学文学院教授,博导,主要研究方向为计算语言学。

* 本文得到国家社科基金项目(10CYY021)、中国博士后基金(2012M510178)、江苏省博士后基金(1101065C)、南京大学计算机软件新技术国家重点实验室开放课题(KFKT2011B03)、江苏高校优势学科建设工程的资助。

gained and statistically analyzed. Thus, a cognitive property bank of over 6000 Chinese common words is constructed and tested in automatic generation of exaggerations and ironies. The research has further applications in natural language understanding and generation, language teaching, lexicography, machine translation and other fields.

Keywords: cognitive property; metaphor; salience; cognitive computation

一　引　言

词语的认知属性是指,在特定的语言中,语言使用者对词语代表的概念或实体的认知体验凝结到词义中的各种属性。例如汉语里“猪”这个词,除了语文词典所描写的“眼睛小,耳朵大,身体肥”之外,其认知属性还有“懒”“笨”“丑”等等。表面上看,后面三个属性不过是“猪”的常见特性而已,但这样的语言知识却很难在现有的语文词典和电子资源中找到。传统的词汇语义学,一般将这种意义作为附加在概念义或基本义之上的附加义或陪义(张志毅,2001:44;张慧晶,2003),并没有重视系统性的描写。认知语言学则不区分概念义和附加义,在隐喻研究的框架下,将“肥”和“笨”都作为“猪”的“显著特征”(Giora,97;Veale,2007)。显著特征可以用作隐喻的喻底,比如“他真是头猪”的意思是他像猪一样笨或像猪一样懒,等等。所谓“显著”是认知上显著,“特征”则表示事物的特殊之处,表现为属性和属性值的特异性(陈小荷,2005)。从理论上看,“显著特征”并没有得到良好的界定。为了凸出认知体验性和属性(值),并涵盖更多的词语属性,在本文中,我们将“显著特征”一般化为“认知属性”,以便进行阐述和分析。

词语的认知属性有助于外国人学习汉语词语的文化认知意义。通过查询词语的认知属性库,外国人可以快速地把握词语文化意义。同时,词语认知属性的研究也有助于计算机处理和理解自然语言。在互联网高度发达的今天,人们习惯于通过论坛、博客、微博发表对于事物的评价,有了“猪—笨、懒”的知识,计算机便可以理解“这个人真是头猪”表达的含义,甚至可以自动生成这样的句子。在人机对话系统中,认知属性的加入,也会让计算机了解人的隐喻性表达,同时生成有趣的话语。在机器翻译领域,借助跨语言的认知属性,有助于计算机翻译意译的、带有修辞手法的句子。

因此,本文力图通过对汉语常用词语认知属性的大规模搜集和统计分析,建立结构化的语言知识库,服务语言教学和词典编纂的需要,提高计算机对修辞表达的理解和处理能力。

二　研究现状

从上世纪 80 年代开始,词语的显著特征特别是名词的显著特征,不仅已成为国内外隐喻分析、自动理解和生成的重要依据(Weiner,1984;杨芸,2008;贾玉祥,2009),还应用于反讽等修辞手法的识别理解(Veale,2007)。对于汉语来说,显著特征还可以解释“太猪了”这样的“副 + 名”特异搭配(施春宏,2001)。因此,名词认知属性的获取成为认知隐喻计算的一个研究热点。

获取和分析词语的认知属性过程中,判定的主观性比较强。纯手工建立电子百科知识库的方式已为学界所抛弃,转而采用自动或半自动方法来建立新型语言知识库。Kintsch

(2000) 在语料库上利用潜在语义分析技术(LSA),寻找和名词密切相关的形容词,作为名词的显著特征。Veale(2007) 利用英语的明喻格式“A is as B as C”,从搜索引擎谷歌上抓取了大量的“目标域(名词)—属性值(形容词)”对。杨芸(2008) 利用词语相关度计算,从本地语料库获取了和名词相关的属性形容词,贾玉祥(2009) 则用明喻格式(如“A 像 C 一样 B”) 搜索百度,获取了汉语名词的显著特征形容词,用于明喻句的理解。这些研究极大地推进了隐喻中明喻研究的进展,但美中不足的是,在数据采集上忽视了频率信息和理论分析。

在语言学研究方面,显著度也逐步被认为是语言理解的重要机制。在 Giora(1997) 提出的“梯度显著度假说”(Graded Salience Hypothesis) 中,显著度高的义项往往被首先处理。Giora (1997) 认为,比喻性语言和非比喻性语言的理解都遵循梯度显著度假说:语言理解过程中首先处理显著度高的意义。意义的显著度受习俗、频率、熟悉程度以及上下文语境的影响。词语认知属性也存在显著度的问题,“猪”的各属性的显著程度如何,和语境的关系如何,直接影响着言语理解的过程。

综上所述,随着隐喻理论的发展,计算语言学界催生了对词语隐喻属性的获取研究,但是对“词语—属性”的分析尚不够细致,自动获取中也存在不少问题。在语言学界,也逐步使用显著度理论解释语言中的各种现象。我们试图将两种方法结合起来,形成基于概率分析的汉语常用词语认知属性库。

三 基于互联网的采集方法

认知属性的采集,我们采用已有的基于互联网搜索引擎的方法,以避免纯手工构建的弊端。为了与现有的语言知识库对接,便于进行语义分析并扩展至其他语言,我们采用了中英双语语义知识库知网的 2007 版(下简称“知网”) 作为词典。知网共收录了汉语的 51020 个名词、27901 个动词和 12252 个形容词。基于三种最简单有效的明喻句式“像 + 名词 + 一样”“像 + 动词 + 一样”“像 + 一样 + 形容词”,使用百度共提交查询 91173 次,每次查询最多返回 100 个结果,共得到 5637500 条记录[①]。

对于返回的记录,使用张华平的分词标注软件 ICTCLAS[②]进行全文自动分词和词性标注。然后提取正文中含“A 像 B 一样 C”的句子,将喻体 B 和属性相似点 C 导入数据库,得到{B, C} 对 3197624 例(tokens),1256430 型(types),得到的喻体型为 461865 个,属性型为 386009 个,参见表 1。

表 1 百度查询明喻格式返回结果

词类	词条数	返回原始记录数	含“喻体—属性”的记录数
名词 N	51020	2877777	800293
动词 V	27901	1658112	522208
形容词 ADJ	12252	1101611	348187
合计	91173	5637500	1670688/1258430[③]

这些喻体—属性数量庞大,由于使用了形容词搜索,还得到了知网未收录词语的明喻用法。不过,其中也含有大量的比较句、含代词的短语和一些错误条目。因此,我们在词性自动标注的基础上,用知网的名词和形容词进行过滤,剩余 22888 个“名词—形容词”型,119375 例,覆盖了 6022 个名词和 3539 个形容词。名词和形容词的双重过滤下,这些明喻的条目基本

正确,但是相比知网中的名词和形容词总量来说,数量大为减少。过滤后存在的另一个问题是,知网的收词范围有限,会遗漏不少正确的条目。如果只用形容词来过滤,则剩余 47869 个“喻体—形容词”型,其喻体会变得更为多样,但其中又包含不少错误的条目。因此,除了 22888 个基本条目外,我们会从 47869 个词型中补充名词以外的喻体词语的认知属性。

关于百度搜索的三点说明:(1) 频率问题:使用百度时,往往会搜索到其他词语的明喻用法,如搜索“像猪一样”,在检索结果中会出现“像猪一样蠢”,也可能额外出现“像母鸡一样蠢”等句子,所以很多条目的用例都多于 100 个。加之仅采集所有词的前 100 个结果,因此本文得到的各种词语认知属性的频率并不完全等于在整个互联网上的频率,虽然不特别准确,但基本可以看出这些词语的明喻频度。(2) 采集数量:知网的 51020 个名词和 12252 个形容词中,只采集到了 6022 个名词和 3539 个形容词构成的“名词—形容词”对。我们发现,没有采集到认知属性的名词数量很大,如“百衲本”“边际”“保存期”“家资”等大量不常用或特征不凸显的名词。形容词也是如此,如“极深”“即刻”“潸然”等。知网将区别词也作为形容词予以收录,而这些区别词,如“非常规、全日制、恶性、国立”等基本上没有明喻用法。当然,如果我们用 5 万多个名词和 1 万多个形容词逐一按照“像名词一样形容词”的格式去搜索百度,会得到更为全面的认知属性数据。但是,受限于搜索引擎的采集间隔时间一般为 20 秒/次,即使同时用 100 个独立 ip 同时采集,也无法在短期内完成 5 亿多次的搜索采集工作。(3) 数据发布:未经词性标注的认知属性采集结果,目前已封装为网络数据库,供学界研究使用,可分别检索喻体词语和属性词语。网址为 http://nlp.nju.edu.cn/lib/cog/ccb_nju.php。

四　汉语词语认知属性分析

词语的认知属性,主要体现在形容词上。我们将“名词—形容词”“动词—形容词”“其他词语—形容词”作为观察和分析对象。

(一) 名词的认知属性

名词的认知属性是最为典型的。表 2 分别给出了频率最高的前 10 个“名词 n—形容词 adj”搭配、名词喻体和形容词属性,均为比较常用的明喻,这些明喻体现名词多样的认知属性。从“喻体—属性”上看,“美玉—美丽”出现次数最多,“纸—薄”“雪—白”也都是人们所熟知的认知属性。从名词喻体上看,拥有最多形容词的是“水”,有 270 个不同的形容词,如“流畅、稀、清淡、纯净”等。本文开头谈到的“猪”,在采集到的认知属性里,“笨、懒、肥”的频次分别为 178、142 和 119,此外还有“幸福”42 次、“贪得无厌”22 次等上百个形容词。不过,“快乐”和“幸福”也排进了前 5 名,是一个很有趣的现象,说明当前人们对“猪”的评价的多样性和时代性。

从形容词属性上看,拥有名词最多的形容词是“大”。由于“大”的义项比较多,而且包含一般性的比较的含义,如“和蚂蚁一样大”。“多、快”也是如此,这些单音节形容词的条目还需要后续的人工校对。相比之下,义项少而基本不含比较含义的双音节词更适合作为常用认识属性代表词,如“简单、美丽、可爱”等。

表 2 名词的认知属性

前 10 个“名—形”			前 10 个名词			前 10 个形容词		
名词	形容词	频次	名词	形容词数	前 5 个形容词	形容词	名词数	前 5 个名词喻体(n)
美玉	美丽	840	水	270	流畅、稀、清淡、纯净、温柔	大	363	扫帚、太阳、蚂蚁、宰相肚子、铃铛
纸	薄	660	孩子	164	好奇、天真、快乐、任性、无助	简单	316	火焰、童贞、孩子、按键、游戏
雨点	密集	557	花儿	126	美丽、香、红、灿烂、美	美丽	271	美玉、花儿、天神、天使、花
雪	白	521	猫	116	贴心、顽皮、慵懒、敏捷、灵活	漂亮	230	花儿、花、妈妈、鲜花、电影
花儿	美丽	497	大海	115	深、宽广、深邃、深广、蓝	可爱	185	天使、新娘、小猫、小女孩、洋娃娃
妖精	温柔	466	山	107	高、沉重、深重、多、重	多	169	星星、海水、山、雨、雨水
细瓷	完美	450	猪	103	快乐、笨、懒、肥、幸福	快	167	兔子、胡子、箭、闪电、风
大海	深	402	花	92	美丽、美、灿烂、漂亮、多	美	153	花儿、花、诗、城市、紫荆花
阳光	灿烂	386	小孩子	91	兴奋、好奇、任性、淘气、天真	好	152	天气、妈妈、广告、家人、姐姐
天神	美丽	341	狼	85	凶狠、狠、果敢、凶残、机敏	白	141	雪、银、纸、卫生纸、牛奶

（二）动词的认知属性

和名词的认知属性相比，动词的认知属性较为特殊。例如，动词“呼吸”本身很难说具有什么认知属性，但是在采集到的结果中有“自然、自由、重要”等属性，体现了人们在进行“呼吸”的动作中体验到的情感。表 3 详细地给出了频率最高的前 10 个“动词喻体—形容词属性”、动词喻体及形容词属性。“过节、过年”的“高兴”和“热闹”，“抽筋”的疼痛，甚而是口语常用的“放屁”的“轻松”，都非常形象地表现出人们在这些活动中的认知体验。“过节”和“死”都超过了 30 个形容词，其体验的多样性是显而易见的。形容词大都是体验性的、高频的。

表 3　动词的认知属性

前 10 个“动—形”			前 10 个动词			前 10 个形容词		
动词(v)	形容词(a)	频次	动词(v)	形容词数	前 5 个形容词	形容词(a)	动词数	前 5 个动词
呼吸	自然	758	过节	34	热闹、高兴、开心、兴奋、快乐	简单	40	吃饭、泡茶、写字、呼吸、梳头
过节	热闹	294	死	32	安静、坚强、寂静、圣洁、冷酷	难受	37	晕车、便秘、冒烟、虚脱、脱臼
过年	热闹	165	呼吸	28	自然、自由、简单、重要、容易	疼	27	抽筋、岔气、撕裂、裂开、电击
过节	高兴	152	过年	27	热闹、高薪、开心、喜庆、兴奋	痛	23	抽筋、扭伤、撕裂、触电、裂开
抽筋	痛	151	吃饭	19	简单、平常、容易、正常、频繁	轻松	20	放屁、喝茶、刷牙、梳头、散布
过年	高兴	145	抽筋	17	痛、疼、疼痛、难受、麻木	快	19	飞、拉稀、加油、跑、变脸
放屁	轻松	123	做梦	17	不真实、恍惚、美妙、美丽、动听	疼痛	18	抽筋、触电、生育、落枕、散架
呼吸	自由	119	爱	17	寂静、深情、重要、细腻、醇香	容易	16	做爱、呼吸、吃饭、说话、放屁
打仗	紧张	108	唱歌	16	好听、有趣、动听、动人、快活	方便	16	拔牙、取款、加油、串门、喝茶
抽筋	疼	106	打仗	16	紧张、匆忙、麻利、忙碌、可怕	自然	12	呼吸、眨眼、睡觉、吃饭、打喷嚏

(三)其他词类和短语

借助词性自动标注的结果,我们可以进一步分析包含普通名词和动词之外的、没有被知网收录的词语的认知属性。表 4 给出了专有名词中的人名、地名及其他类型的 10 个最高频的“词语—形容词”对,表 5 给出了时间词、语素、字母词的 10 个最高频的“词语—形容词”对,表 6 给出了较长短语的相关信息。从表 4 中可以看出人物形象“可卿、凤姐、赫本”的认知属性分别为“漂亮、精明、优雅”,国家名、地名“美国、西湖、泰山”的认知属性分别为“强大、美丽、稳固”。其他专名“北斗星、蒙牛、春兰”等也都具有各自的认知属性。有些认知属性初看起来有些费解,如“两面针—蠢”,通过百度查看相关新闻和帖子之后才比较清楚其内涵。

表 4 专名词语的认知属性

前 10 个“人名—形”			前 10 个“地名—形”			前 10 个“专名—形”		
人名(nr)	形容词(a)	频次	地名(ns)	形容词(a)	频次	专名(nz)	形容词(a)	频次
可卿	漂亮	943	美国	强大	58	北斗星	弯曲	90
风儿	自由	201	西湖	美丽	44	蒙牛	成功	55
小山	高	152	台湾	精致	38	春兰	淡雅	16
凤姐	精明	128	泰山	稳固	35	春兰	幽远	15
蓝晶	好	117	长城	坚固	32	捷达	好	14
风标	漂亮	106	终南山	高	25	汉语	易学	14
赫本	优雅	102	西伯利亚	寒冷	24	两面针	蠢	12
曾参	孝顺	97	沁水	宽	23	麦当劳	强大	8
老徐	大	79	长城	长	23	虎豹	威武	7
刘邦	大方	62	泰山	稳	21	娃哈哈	稳定	6

时间词、语素名词、字母词的认知属性也很有趣。在人们的感知中，“春夏秋冬”分别代表着“温暖、炎热、爽朗、寒冷”。语素字的词性标记为 g，字母词的词性标记为 x。语素字“瓷—白”“箫—哀怨”“猴—机敏”都是非常典型的认知属性。字母词由于数量较少，我们给出了包含字母词的词语的认知属性，如“挂 QQ、打 CS、做了 SPA”等。这些名词和动词短语，显示了人们现实生活的诸多主体认知体验。

表 5 特殊词语的认知属性

前 10 个“时间—形”			前 10 个“语素—形”			前 10 个“字母词语—形”		
时间词(t)	形容词(a)	频次	语素(g)	形容词(a)	频次	字母词语	形容词(a)	频次
春天	温暖	256	瓷	白	53	SB/x	开心	203
白天	明亮	119	箫	哀怨	22	NB/x	开心	100
夏天	炎热	68	猴	机敏	19	本田/nz CBR600/x	醇厚	61
春天	美丽	56	乳	润泽	15	挂/v QQ/x	简单	55
冬天	冷	52	犬	忠诚	14	做/v 了/u SPA/x	光泽红润	31
秋天	爽朗	39	猿	敏捷	13	J8/x	长	23
春天	妩媚	31	貉	懒惰	12	打/v CS/x	舒服	21
春天	滋润	28	鼠	轻灵	11	小/a S/x	幸福	20
春天	美好	26	玑	爽朗	10	gt/x	高	18
冬天	寒冷	25	狐	妖娆	9	ThinkPad/x	严谨	17

自动采集的结果中也包含了较长短语的认知属性，这是以往的研究所忽略的部分。由于短语较长，往往是对特定对象、事件的小范围感知。表 6 给出了普通短语、书名号管辖的专名、引号管辖的专名的认知属性。普通短语里既有传统的“吃了蜜—甜”，也有近年来才出现的“打了鸡血—兴奋”。其中也包含一些小领域内的感知，如“同桌小丽—漂亮”。还有一些非明

喻表达,如“对夏鸥的母亲—亲热”。数据库中含有书名号或引号的专名多为书籍影视作品名称或人物,可以从上下文的形容词中观察到公众对这些专名的态度。如“《魔戒》—伟大”“谢大脚—漂亮”“犀利哥—出名”。

表 6　较长短语的认知属性

前 10 个短语—形			前 10 个书名—形			前 10 个引号—形		
短语	形容词(a)	频次	书名号专名	形容词(a)	频次	引号专名	形容词(a)	频次
在/p 电脑/n 上/f 登录/v	方便	331	一/m 公升/q 的/u 眼泪/n	感人	31	黑客/n	黑	50
吃/v 了/u 蜜/n	甜	233	窗/g 边/k 的/u 小豆/n 豆/n	美丽	29	赵/nr 丹阳/ns	清醒	36
清晨/t 的/u 雾/n	纯洁	216	天方夜谭/l	有趣	13	布达拉宫/ns	美丽	25
灌/v 了/u 铅/n	沉重	172	谍/x 海/n 风云/n	无聊	12	看家狗/n	忠实	24
同桌/v 小/a 丽/g	漂亮	171	泰山/ns 刻/v 石/g	工整严谨	9	水/n	软	23
对待/v 父母/n 生前/t	恭敬	154	命案/n 高悬/v	有趣精致	8	偷/v 菜/n	简单	21
对/p 夏鸥/nr 的/u 母亲/n	亲热	131	公主/n 小妹/n	美丽	8	康吉/nr	快乐	19
一/m 头/q 贪吃/an 的/u 熊/n	臃肿	130	魔/g 戒/g、星球/n 大战/n	伟大	7	谢/nr 大脚/n	漂亮	18
打/v 了/u 鸡/n 血/n	兴奋	110	士兵/n	火爆	7	上甘岭/ns	残酷	17
踩/v 了/u 棉花/n	软	108	飞升/v 之后/f	宏大	7	犀利/a 哥/n	出名	16

(四)总结分析

根据上文的统计分析,可以明显地看出,我们采集的词语已经大大超出了知网收录的词语范围,体现出互联网数据的巨大优势。从上述三类词语的认知属性,可以看出以下几个特点:

1. 词语认知属性的个性差异大,相同义类的词语的认知属性不同。如表中所列举的时间词“春天”和“秋天”、处所词“西伯利亚”和“长城”的认知属性差别很大。

2. 不同词语的认知属性可能相近,如“花儿”“天使”都有“美丽”的属性。

3. 名词和动词的认知属性差异大。名词的认知属性主要是名词所指称的概念在日常生活中的感知体验,在形容词上主要使用“美丽”“可爱”等。而动词的认知属性则是人们在这些动作、活动中体验到的各种感觉,在形容词上的使用差异也是非常明显的,如“疼”“痛”“紧

张”等。

4. 名词和动词的认知属性中有部分可互通，体现了认知对象和动作行为的一致性。“兔子”和“飞”“跑”的属性都有“快”，这体现出运动的典型主体和典型动作在认知属性上的一致性。“蜜”和“吃了蜜”都是“甜”，则体现出感知对象和感知过程在认知属性上的一致性。

5. 仅使用形容词作为认知属性的载体，还难以做到细致入微。例如，名词“孩子”和动词“呼吸”的属性都有形容词“简单”，但其“简单”的含义并不完全一样。作为抽象名词，“简单”的这两个含义，在《现代汉语词典》和知网中也没有细致的区分。这就要求在认知语义学的框架下，对形容词做更为细致的研究。

五 认知属性的理论意义和应用价值

中文词语认知属性库的建立，对于词汇语义学、英汉认知差异对比、夸张反讽的自动生成都有直接的理论和应用价值。

1. 丰富词汇语义学的研究。将传统的词语文化义、隐喻义在认知属性的理论框架下较为系统地描写出来，揭示心理词库的组织方式。认知属性甚至可以作为词语分类的新的依据。以属性“温暖”为依托，我们看到“阳光、太阳、家、家庭、春天、春风、火”等不同语义类下的词语可以拥有相同的认知属性，这可能是人们心理词库的重要组织方式之一。在这个意义上说，传统的词语相似度的计算，也可以借助认知属性得到更好的结果，把语义类上差别较大的词语，计算出较高的相似性。

2. 英汉认知属性的差异对比。我们把汉语认知属性库和 Veale(2007) 建立的英文数据库 sardonicus[④] 进行了初步的比较，利用知网收录的中英双语名词和形容词为中介，发现两个数据库仅有 1000 多条“名词—形容词”可以匹配上，可见英汉之间的认知差别是较大的。由于英文数据库没有频率信息，暂时难以做系统的比较，即使有相同的“名—形”对也不排除是偶然的巧合或是翻译表达。

3. 夸张的自动生成。使用典型的夸张句式“比 N 还 A”和认知属性库中的数据，可以很容易地生成“比美国还强大”“比凤姐还精明”“比打仗还紧张”“比西伯利亚还寒冷”等句子。这些句子显得非常自然，可以直接用于人机对话、机器翻译的译词选择任务。

4. 反讽的自动生成。利用知网提供的反义、对义关系，我们将认知属性应用于反讽的自动生成。在这个过程中，我们发现大多数反讽句都是难以接受的，如“像美国一样弱小(强大)”“像凤姐一样愚蠢(精明)”。这促使我们进一步思考，反讽的生成机制是较为复杂的。一般只能将不好的属性反讽为好的属性，如“跟猪一样聪明(笨)”“像豆腐一样硬(软)”，反之则不行。因此反讽的自动生成还需进一步研究。

六 结 语

词语的认知属性是在一种文化下人们对事物认知的语言表达，涵盖了传统词汇语义学的多种陪义(附加义)。由于这些陪义往往是词典所无或疏于描写的对象，描写难度也较大，本文则采用基于互联网的技术，通过明喻格式，从搜索引擎上获取了大量词语的认知属性，比过去仅靠人工总结、分析词语的文化意义和附加意义更为快捷和全面。特别是根据频率信息的统计，能看到一个词语不同属性的显著度，也可以看到一个认知属性支配不同名词的概况。这对于外国人学习汉语词语的文化认知意义、编纂认知型教学和语文词典、辅助机器翻译方面都

具有重要的应用价值。

不过,目前的研究仍有不足,还需要继续研究下列问题:(1)继续研究知网收录却没有采集到实例的那些名词和形容词,观察其语义类的分布特点和认知特点。(2)认知属性的结构化和形式化,考虑将作为认知属性的形容词进一步离析,从认知背景和认知角度等方面进行分析。(3)加强动态性、地域性的研究,不同时期、不同地域的不同主体对同一事物的认知属性可能是不同的,需要在地域、篇章的角度来建模,描写这种差异性。(4)使用谷歌等搜索引擎,采集英文的认知属性,形成英汉双语带频度的认知属性库,以进行英汉词语认知属性的深入比较。(5)对于不断涌现的新词语,如“林书豪”、形容词“囧”等,能够做到增量式采集,以观察认知属性的动态变化。

[附　注]

① 百度对中文搜索的结果较好。我们也尝试利用 Google Book 公开的中文历代图书中包含的 5Gram 数据(5 个词构成的所有词串)统计明喻句式,但仅得到数百条有效记录,数据量过低,无法使用。而谷歌的查询结果,往往会混入其他语言的翻译结果,且严格限制爬虫的速度。

② 下载地址:www. nlp. org. cn。

③ 1670688 条记录中,由于去除了上下文,存在部分重复。经去重后剩余 1258430 条记录。

④ http://afflatus. ucd. ie/sardonicus/tree. jsp。

[参考文献]

[1] Kintsch, W. Metaphor comprehension: A computational theory. *Psychonomic Bulletin & Review*, 2000, (7): 257 ~ 266.

[2] Kruse, D. Alan. *Lexical semantics* [M]. Cambridge: Cambridge University Press, 1986.

[3] Tony Veale, Yanfen Hao. Learning to Understand Figurative Language: From Similes to Metaphors to Irony [A]. *Proceedings of CogSci* 2007 [C], Nashville, USA, 2007.

[4] Weiner, E. J. A Knowledge Representation Approach to Understanding Metaphors [J]. *Computational Linguistics*, 1984, 10(1): 1 ~ 14.

[5] 陈小荷. 属性分析说略 [A]. 语言计算与基于内容的文本处理 [C]. 北京:清华大学出版社,2005.

[6] 贾玉祥. 基于实例的隐喻理解与生成 [J]. 计算机科学,2009,(3).

[7] 施春宏. 名词的描述性语义特征与副名组合的可能性 [J]. 中国语文,2001,(3).

[8] 杨　芸. 汉语隐喻识别与解释计算模型研究 [D]. 厦门大学博士学位论文,2008.

[9] 张志毅,张庆云. 词汇语义学 [M]. 北京:商务印书馆,2001.