现代汉语语义词典（SKCC）的新进展*

王惠　　俞士汶　　詹卫东

北京大学计算语言学研究所　北京大学中文系

whui@pku.edu.cn;　yusw@pku.edu.cn;　zwd@pku.edu.cn

摘 要："现代汉语语义词典"(SKCC)是一部面向中文信息处理的语义知识库，1998年底完成一期工程，收词48,835条。从2001开始，北大计算语言学研究所与中文系合作进行SKCC的二期开发。针对机器翻译的需要，对原有的语义分类体系作了较大改造，语义属性描述也得到全面修订，并新增了1.8万常用词语，以及大量的词义解释与真实用例。目前，顺利完成了6.6万多词语的语义归类及组合搭配信息的填写与校对。整个词典的规模和质量有了显著提高，可为计算机语义分析提供更强有力的支持。

关键词：语义词典　语义属性　语义分类　计算词典学　中文信息处理

New Progress of the Semantic Knowledge-base of Contemporary Chinese（SKCC）

Wang Hui,　Yu Shiwen　　Zhan Weidong

Institute of Computational Linguistics
Peking University, Beijing 100871

Department of Chinese Language & Literature
Peking University, Beijing 100871

whui@pku.edu.cn;　yusw@pku.edu.cn　　zwd@pku.edu.cn

Abstract: *The Semantic Knowledge-base of Contemporary Chinese* (SKCC) is a large machine-readable dictionary developed by the Institute of Computational Linguistics and Chinese Department of Peking University. Through the continuous development in the last two years, the scale and quality of the knowledge-base have been improved remarkably. Currently, it can provide a large amount of semantic information such as semantic hierarchy and collocation features of 66,539 Chinese words. Its POS and semantic classification system represents the latest progress in Chinese linguistics and language engineering. The descriptions of semantic attributes are fairly thorough, comprehensive and authoritative. The paper introduces the outline of new SKCC, and indicates that it can not only provide valuable semantic knowledge for Chinese language processing, but also has great theoretical significance in Chinese lexical semantics and computational lexicography research.

Key words: Semantic knowledge-base, lexical semantic, semantic hierarchy computational lexicography, Chinese language processing

* 本研究得到国家973重点基础研究项目（G1998030507-4）和（G1998030507-1）的支持。

1 前 言

在自然语言处理中，语义分析占有很重要的位置。北京大学计算语言学研究所与中科院计算所自 1994 年联合开发“汉英机器翻译模型系统”开始，就着手研制面向汉英机器翻译的“现代汉语语义词典”（SKCC），目的是在语法分析的基础上，为计算机自动分析汉语句子和生成英语句子提供更深入的语义信息。1996 年至 1998 年，双方共同承担了国家 863 高科技项目“通用机器翻译开发平台和汉英机器翻译系统”课题。作为该课题的一个重要组成部分，现代汉语语义词典进入到大规模开发阶段，并取得阶段性成果，完成了 4.9 万汉语常用实词语的语义分类和搭配信息描述[1]。IBM、Intel、Fujitsu, Toshiba, NTT, Canon, Sail-labs 等 20 多家公司与大学先后从北大购买了该词典的许可使用权。

4 年多来，北京大学计算语言学研究所在积极应用、推广该词典的同时，仍不断地投入力量进行词典本身的发展。从 2001 年 11 月开始，“现代汉语语义词典”的二期开发工作受到了国家 973 重点基础研究项目“面向新闻领域的汉英机器翻译系统”和“面向中文信息处理的现代汉语动词论旨结构系统和汉语词语语义分类层级系统研究”的支持，由计算语言学研究所和中文系联合承担，对词典规模进行较大幅度的扩充，并对全部词语的语义分类及属性描述进行全面修订。在双方的积极努力下，项目进展得非常顺利。目前，词典规模已达到 6.6 万余条，同时语义属性描写质量有了显著提高。在一个汉英机器翻译系统中的实际应用表明，新版 SKCC 可以为句义分析、词汇歧义消解提供更全面的语义知识，有效地提高了翻译精度。

2 词典规模的扩充

2.1 增加词语

语义词典（SKCC）原有词条 48,835 个，它们全部来自于北大计算语言学研究所开发的《现代汉语语法信息词典》。但 1999 年，后者的规模已由 5 万词扩充到了 7 万词[2]，此后的继续改进又使得属性信息的质量有了很大提高[3]。相比之下，语义词典却仍然停留在原来的水平上，无论数量还是质量上，二者都已不太协调，不能满足与语法词典配套使用的实际需要。

SKCC 的二期工程及时吸收了语法信息词典的最新成果，对原有的“词语”、“词类”、“同形”、“拼音”、“兼类”等字段进行了统一检查、修订，而且增加了 14,663 个名词、动词、形容词，以及 1993 个区别词、时间词、处所词、方位词、副词、数词。现在的语义词典比原来增加了 1.6 万词语。

2.2 词条细分

开发语义词典的直接目的是帮助机器翻译系统消解句法和词汇分析中的歧义问题，因而，语法词典中的同一个词条，若对应不同的英语译文，或者属于不同的语义类，语义词典都将其进一步细分为不同的条目。

2.3 语义词典 SKCC 的现有规模

SKCC 现已收词 6.6 万余条,分为 12 个数据库，其中包含全部词语的总库 1 个，每类词语各建一库，计 11 个。每个库文件都详细刻画了词语及其语义属性的二维关系。总库中包括词语、拼音、同形、义项、语义类、词类、子类、兼类等 8 个字段。每类词的特有属性填在各类词库中，如名词库设 15 个属性字段，动词库设 16 个属性字段，如此等等（见表 1）。总库与各类词库可以通过"词语、词类、同形、义项"这 4 个关键字段进行链接。

库 名	词条	属性字段
名　词	375	15
时间词	567	15
处所词	185	15
方位词	204	15
代　词	236	15
动　词	2114	16
形容词	382	15
区别词	753	15
状态词	997	15
副　词	997	11
数　词	109	11
总　库	**665**	**8**

表 1 语义词典 SKCC 的规模

3 语义分类体系的改进

与基于常识的各种语义分类相比,SKCC 的一个突出特点就是其语义分类的深度与广度取决于语法分析的需要[1]。经过 4 年来的应用检验与研究，我们发现，对于汉语信息处理来说，这种分类法是很有前途和实用价值的。但毋庸讳言，语义词典原有的分类中还存在一些地方没有完全贯彻这个原则，因而需要按照语言分析的实际要求进行调整。

比如，动词"吃"的客体是"可食物"，原语义词典中"苹果、面包、青霉素、强心剂"都是"可食物"，但我们只能说"吃苹果、吃面包"，而不能说"*吃青霉素、*吃强心剂"。因此，需要把"药物"类独立出来,作为单独的一类。

再如，"构件"类原是"生物"下面的一个子类，与"人类、动物、植物、微生物"等并列，但显然"构件"类名词不仅可以是生物的，如"鼻子、脸、腿"等，也可以是非生物的，如"袖子、封面"等。因此，现在把它的位置提升了上来，并进一步细分为"生物构件"和"非生物构件"。

此外，考虑到以后便于与 Wordnet [4]和"中文概念辞书（CCD）[5]"兼容，同时与"知网（hownet）[6]"、"同义词词林"等已有的多种语义词典实现资源共享，我们在参照现有各家语义分类的基础上，针对汉英机器翻译的需要，对语义词典的原分类体系作了较大的调整：

（1） 名词上下位关系更加系统化：首先，将具体事物、抽象事物与过程、时间、空间并

列为 5 大类；然后再逐层细分：具体事物分为生物、非生物 2 类，生物里再把人与动物、植物、微生物相并列，非生物中则进一步区分开人工物、自然物、排泄物和外形。然后，根据 Wordnet 与“知网”中的内容，补充了一些较低层的名词小类*。

（2） 把 Wordnet 中的动词分类借鉴过来，但根据汉语的实际作了相应改造；

（3） 形容词的分类更加细化，由原来的 7 类发展成为现在的 5 大类 29 小类，与名词的分类互相照应，从而可以更细致地刻画形名搭配关系。

总的来说，调整后的新语义分类更趋合理，名词的分类相对较细，动词、形容词、数词、副词的分类较粗，只要能揭示出与名词性成分、动词性组合成分的不同组合类型即可。目前我们已实际完成了 6.6 万词语的语义类划分与属性描述。

4 语义属性描写的完善

4.1 语义属性的增补与修订

原语义词典的重点是对名词、动词、形容词这 3 类词进行语义类划分，并在配价理论的基础上，描述其语义搭配限制。但是由于多种原因，其中有些属性项目还没有填写内容，或只是填写其中一部分的值。比如，名词、形容词库的“参照体”、“对象”字段都没填；名词和形容词都有 1 价和 2 价之分，但“配价数”字段中只填“1”，没有考虑 2 价；3 价动词（如“补偿、补助、发给、奉送、赠送、转告”等）的“邻体”字段也基本未填。所有这些，在本次修订之中，都要求严格按照新说明书予以填平补齐，同时进行至少两遍校对。表 2 是名词库样例：

词语	词类	同形	义项	语义类	配价数	参照体	对象	英语译文
老虎	n			动物	0			tiger
腿	n	1	1	生物构件	1	人/动物		leg
腿	n	2	2	非生物构件	1	用具		leg
意见	n	1	1	认知	2	人	实体|抽象物	view
意见	n	2	2	认知	2	人	人|事件	objection

表 2 语义词典 SKCC 的名词库部分属性

* Wordnet 在每类词的基本分类下，只有大大小小的、彼此具有上下位关系的同义词集合（synset），而不再设立低层的语义类名称。因此，我们对 Wordnet 语义类的借鉴主要限于名词、动词的基本语义类上，然后，根据汉语句子分析的需要适当地补充一些 synset 作为小类，如“Artifact（人工物）”下面的“创作物、药物、设施、工具”等，而不可能也没有必要把该语义类的直接下位概念（以下 18 个 synset）全部都照搬过来：Antiquity（古代遗产）、block（块状物）、covering（遮盖物）、creation（创作物）、decoration（装饰物）、drug（药物）、enclosure（圈）、excavation（挖掘的地洞）、excavation（纺织品）、facility（设施）、fixture（固定设备）、float（飘浮物）、instrumentality（工具）、toy（玩具）、way（道路）、keepsake（纪念品）、notion（小饰物）、prize（奖品）。

此外，新版 SKCC 中还增加了对区别词、状态词、时间词、处所词、方位词、数词、副词等 7 类词语的语义描述，并详细刻画每个词的配价数以及其在上下文中的语义搭配限制。

4.2 新 SKCC 的语义属性描述

目前， SKCC的语义属性描写非常完整，全部的属性项目都已经按照要求填满信息，并基本上至少经过了两遍校对。所描述的属性大致可归纳为以下4类：

（1）词语本身的一些基本特征，如该词的词形、拼音、词性、兼类、有无同形词、例句等。它们均是从北京大学计算语言学研究所开发的《现代汉语语法信息词典》（2002 版）[3] 中直接继承而来。这不仅保证了语义词典收词的规范性、注音与词性标注的准确性，而且也使得它可通过“词语、词类、同形”3 个关键字段与语法信息词典进行链接，相互配合使用，从而使系统获得更加完备的语法、语义信息。

（2）词语意义的基本刻画，如所属的语义类、词义解释、是否属于多义词的一个义项等。如表 4 对“找”两个义项的描写，可以为汉语词义消歧和词义研究提供了丰富的知识。

（3）描述一个词语跟其他实词发生语义联系的能力，主要包括配价数（能支配多少名词性成分）、配项成分的语义角色（主体、客体、与事）和语义约束几个方面。这是语义词典的重点开发内容，可直接服务于计算机语义自动分析。

（4）每个词条的英语译词及其词类。若对应的是英语短语，还要指出其中心语(用!表示)。如表 3：

词语	同形	释义	语义类	主体	客体	与事	英语译文	英语词类
找	A	寻找	身体活动	人	具体事物		look for	!V+P
找	B	退还	领属转移	人	“*钱”	人	give change	!V+N

表 3 多义词“找”的两个义项

5 词典质量的保证措施

质量是词典的生命。由于 SKCC 中的属性种类繁多，信息量庞大，开发周期长，编纂人员多，词典质量的保证更是非常关键的问题。为此，课题组除了制定一套严格的人工检查流程外，还设计了一个计算机辅助编辑和校对系统，用于词典信息浏览、填写、校对、检测及版本比较。利用它，词典开发人员可以很方便地填写各项语义属性，而且还可从各个角度对词典的属性值进行检查，有效地提高了语义属性描述的正确性与一致性。

(1) 语义词典 SKCC 属性值的内部检测

“词典检测辅助工具”会自动对所填入的属性值进行有效性检查，如果发现某个属性项目漏填，或者超出规定的取值范围，计算机将会立即弹出错误警告，提醒编辑人员及时检查、修改错误。如果需要，计算机还可提供一份完整的错误记录报表

(2)参照已有语义知识资源进行横向检测

利用“同义词词林”、“知网”等词典内容对 SKCC 检测。比如，“罢工、罢课、罢市”在“知网”中属于同一个语义小类，但 SKCC 中却分别归入了不同的语义大类。这说明其中某个词的分类很有可能发生了错误。计算机就会把这一组词提交校对人员，及时审查、修正。

(3)不同版本的自动比较

在开发过程中，由于词典的属性信息一直要不断地维护和更新。通过版本比较，可以清楚地看到所有更改过的词条和字段，并以报表形式输出结果。这样，校对人员就可以全面掌握以前的各种修订情况，及时发现并纠正各种错误，防止错校、漏校。这对词典的质量保证显然也是非常重要的。

6 结语

作为北大计算语言学研究所的综合语言知识库的一个组成部分，“现代汉语语义词典”（SKCC）不仅可以应用于机器翻译，而且还可以在多种 NLP 系统（如自然语言接口、文献检索、信息自动提取、语音识别与合成、文本校对、语料库加工等）的语义分析中发挥重要作用。其中的语义信息在汉语分析的各个层面，包括多义词义项判断、短语结构层次和结构关系判定、以及成分之间语义关系的确定等等，都能起到重要的作用。同时，对于促进汉语词汇与语义学研究，开展汉语词义定量分析等也有很大的价值。

目前，SKCC 的二期工程已取得重要的成果，词典规模扩大到了 6.6 万多条词语，语义属性描写质量也有了显著提高，并已在一个汉英机器翻译系统中得到实际应用。但语义词典的开发毕竟是一项长期的语言工程，我们还应根据实际应用的反馈意见，不断地发现问题，总结经验，逐渐完善现有的语义分类体系及属性项目。同时，从大规模语料中自动抽取更多的语义搭配知识，检验并丰富我们现有的语义约束描述，在计算词义学方面进行更深入的探索。

参 考 文 献

[1] 王惠，詹卫东，刘群.“现代汉语语义词典的设计与概要”.《1998 中文信息处理国际会议论文集》. 清华大学出版社. 1998. pp 361-367.

[2] 俞士汶，朱学锋，王惠.“《现代汉语语法信息词典》的新进展”.《中文信息学报》2001 年第 1 期.

[3] 俞士汶，朱学锋，王惠，张化瑞等.《现代汉语语法信息词典详解（第 2 版）》.清华大学出版社. 2002

[4] Christiane Fellbaum. ed.. *WordNet: an electronic lexical database*. Mass: MIT Press. 1998

[5] 于江生，俞士汶.“CCD 的结构与设计思想”.《中文信息学报》. 2002, 16 (4). pp 12-20.

[6] 董振东，董强.“知网”（Hownet）. http:// www.keenage.com.

[7] 陆俭明.《现代汉语配价语法研究 • 序》，北京大学出版社 1995. pp1-7.

[8] 袁毓林.“一价名词的认知研究”，《中国语文》1994 第 4 期

[9] 袁毓林.“现代汉语二价名词研究”.《现代汉语配价语法研究》，北京大学出版社 1995. pp29-58.