现代汉语语义词典规格说明书*

王惠[1] 詹卫东[2] 俞士汶[1]

[1]北京大学计算语言学研究所
[2]北京大学中文系

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

Submitted on 8 March , 2003, Revised and Accepted on 16 May, 2003

摘 要

“现代汉语语义词典”(SKCC)是一部面向自然语言信息处理的语义知识库，它以数据库文件形式收录了 6.6 万余条汉语实词，不仅给出了每个词语所属的词类、语义类，而且以义项为单位详细描述了它们的配价信息和多种语义组合限制，可以为包括机器翻译在内的多种中文信息处理系统中的语义自动分析提供强有力的支持，同时，对于汉语词汇语义学和计算词典学研究也具有重要的意义。本文概要介绍这部语义词典的结构、内容，以及语义属性项目的填写规范。

关键词

语义知识库，语义类，配价信息，计算词典学，中文信息处理

1. 引言

随着语言处理技术的迅速发展，词义分析的重要性与迫切性也越来越突出。为了获取足够的词义知识，克服目前普遍存在的“词义瓶颈”难题，从 80 年代中期开始，世界上许多国家都大力投资开发机用语义词典，如：美国的 Wordnet (Fellbaum, 1998)、Mindnet (Richardson, 1998)、Framenet (Fillmore, 1998)、日本的 EDR 概念词典、新加

* 本文有关研究得到了 973 项目（G1998030507-4，G1998030507-1）和 863 项目（2002AA117010-08）的支持。

坡的 SenseWeb 等。中国也陆续开展了汉语语义词典的研究与开发，如“905”项目“信息处理用汉语语义词典”(陈力为, 袁琦, 1995)、“现代汉语述语动词机器词典”、“知网（Hownet）”（董振东, 1999）、“中文概念辞书（CCD）”（于江生, 俞士汶, 2002）等。此外，不少计算语言学家还尝试着从机器词典中自动抽取词义知识（Chodorow 1985，Ide 1993，黄居仁 1998 等）。但迄今为止，现有的规模较大的词义工程，基本上都是采用词义分类的办法，有些再加上为数不多的属性描述。而国内外研究工作者建立义类体系的方法，也基本上都是对词义进行静态的聚合分类，并没有把词义放到一定的组合框架中去观察，所以，在自然语言处理系统中起的作用是有限的。

为了给计算机自动分析提供更全面、深入的语义信息，我们应充分吸收现有的研究成果，在语法知识库的基础上构建语义知识库。不仅要进行系统的语义分类，而且对词义组合信息加以全面描述，进一步加强动态的语义组合知识的研究和总结，建立一个与语言工程应用紧密配合的、合理的语义知识描述框架。

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）。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 万汉语名词、动词、形容词及成语、习用语的语义分类和搭配信息描述(王惠等,1998)。IBM、Intel、Fujitsu, Toshiba, NTT, Canon, Sail-labs 等 20 多家公司与大学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，由北大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 万余词条，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，新版的 SKCC 可以为句义分析、词汇歧义消解提供更全面的语义知识，有效地提高了机器翻译的精度。

2. 现代汉语语义词典的内容概要

2.1 规模与结构

语义词典 SKCC 二期工程及时吸收了语法信息词典的最新成果（俞士汶等，2003），对原有的“词语”、“词类”、“同形”、“拼音”、“兼类”、“备注”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。现在语义词典的规模比原来增加了 1.7 万词语，达到了 6.5 万余条。

词典采用 Microsoft Forxpro 6.0 数据库实现，其中包含全部词语的总库 1 个，每类词语（实词）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。比如，总库中包括词语、拼音、同形、义项、释义、语义类、词类、子类、兼类等 8 个属性字段。名词库设 15 个属性字段，动词库设 16 个属性字段（见表 1）。所有的库都可以通过“词语、词类、同形、义项”这 4 个关键字段进行链接。

数据库	词条	属性字段
名　词	37485	15
时间词	567	15
处所词	185	15
方位词	203	15
代　词	235	15
动　词	20798	16
形容词	3557	15
区别词	315	15
状态词	993	15
副　词	996	11
数　词	108	11
总　库	**65442**	**8**

表 1 语义词典 SKCC 的规模

2.2 SKCC 的语义分类体系

国内外对汉语语义分类体系的研究已有了不少成果，如：梅家驹（1983）、林杏光（1987，1998）、陈群秀（1996）、陈晓荷（1998）、董大年(1998)、董振东（1998，1999）等。但由于各家分类体系的目的及应用范围不同，对同一事物可能有不同的定义与归类。如“动物”在一个语义体系中分为：“兽类、鸟类、鱼类、虫类、爬行类”，而在另一个体系中分为“脊椎动物、腔肠动物、软体动物”。但这些分类体系都是基于自然科学或常识而独立于语法的。在实际语言分析中，如何将这些语义知识与语法知识有机地结合起来是一件很困难的事情。

与这些基于常识的各种语义分类相比，SKCC 语义分类的突出特点就是分类的深度与广度取决于语法分析的需要。应用语义知识应着重于解决那些仅靠语法规则难以解决的问题。因而语义分类是在词的语法分类基础上进行的，并且只对名词、动词、形容

词等实词进行语义分类描述，而那些带有明显标志的、通常用句法形式就可以表示的语义关系，如各类虚词，则不作为语义分类研究的对象。

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，“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 和“中文概念辞书（CCD）”兼容，同时与“知网”、“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:

A. 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，补充了一些低层小类*。

B. 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；

C. 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。

---

* 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）。又如，“Substance（物质）”这个语义类下面也是直接分为 18 个 synsets：material（原材料）、allergen（过敏源）、mixture（混合物）、atom（原子）、molecule（分子）、chemical_element（化学元素）、activator（催化剂）、inhibitor（抑制剂）、compound（化合物）、fuel（燃料）、medium（酶）、eaven（酵母）、fluid（流体）、sludge（软泥）、refrigerant（制冷剂）、residue（残渣）、poison（毒物）、chemical_irritant（化学刺激物）、solid（固体）、emanation（放射物），但我们只吸收了其中的“原材料、食物”两个作为小类，其余的则归入“其他”类，暂不区分。

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。具体分类体系如下：

2.2.1 名词（Noun）

1 具体事物（entity）
- 1.1 生物（organism）
  - 1.1.1 人（person）
    - 1.1.1.1 个人（individual）
      - 1.1.1.1.1 职业（profession）：教师 秘书 会计 医生
      - 1.1.1.1.2 身份（identity）：华侨 外行 健将 模范
      - 1.1.1.1.3 关系（relation）：父亲 阿姨 长辈 朋友
      - 1.1.1.1.4 姓名（name）：爱因斯坦 毛泽东 鲁迅
    - 1.1.1.2 团体（group）
      - 1.1.1.1.1 机构（organization）：工厂 医院 商店 剧团
      - 1.1.1.1.2 人群（society）：人民 委员会 少先队 团伙
  - 1.1.2 动物（animal）
    - 1.1.2.1 兽（beast）：狗 猪 牛 羊 老虎 豹子 狐狸
    - 1.1.2.2 鸟（bird）：鸡 鸭 麻雀 杜鹃
    - 1.1.2.3 鱼（fish）：鲤鱼 河豚 鲸 泥鳅
    - 1.1.2.4 昆虫（insect）：蚯蚓 知了 蟑螂
    - 1.1.2.5 爬行动物（reptile）：青蛙 乌龟 甲鱼 蛇
  - 1.1.3 植物（plant）：树 花 草 牡丹 芍药
    - 1.1.3.1 树（tree）：白杨 水杉 芭蕉
    - 1.1.3.2 草（grass）：狗尾巴草 含羞草 蒲公英
    - 1.1.3.3 花（flower）：牡丹 芍药 杜鹃 映山红
    - 1.1.3.4 庄稼（crop）：蔬菜 小麦 高粱 棉花
  - 1.1.4 微生物（microbe）：细菌 病毒 霉菌
- 1.2 非生物（object）
  - 1.2.1 人工物（artifact）
    - 1.2.1.1 建筑物（building）：别墅 礼堂 会议室 水库 庙
    - 1.2.1.2 衣物（clothes）：服装 外套 衬衫 裙子 帽子
    - 1.2.1.3 食物（food）：面包 牛奶 菜 米饭 饮料
    - 1.2.1.4 药物（drug）：药片 阿斯匹林 酒精 镇定剂

1.2.1.5 创作物（works）：论文 书 杂志 文章 油画 电影
1.2.1.6 计算机软件（software）：操作系统 数据库 程序 软件
1.2.1.7 钱财（asset）：财产 钱 资金 报酬 罚款 美元 利息
1.2.1.9 票据（bill）：发票 单据 汇票 支票 包裹单
1.2.1.10 证书（certificate）：结婚证 执照 毕业证 驾驶证
1.2.1.11 符号(symbol)：签名 路标 箭头 句号
1.2.1.12 材料（material）：木材 钢铁 煤炭 玻璃 水泥
1.2.1.13 器具（instrument）
1.2.1.13.1 用具（tool）：剪子 刀子 钉子 拖把 改锥
1.2.1.13.2 交通工具（vehicle）：车 船 飞机 自行车
1.2.1.13.3 武器（weapon）：大炮 机关枪 鱼雷
1.2.1.13.4 家具（furniture）：桌子 椅子 沙发
1.2.1.13.5 乐器（musical-instrument）：钢琴 吉他 鼓
1.2.1.13.6 电器（electricity）：电视 空调 电冰箱
1.2.1.13.7 文具（stationery）：钢笔 橡皮 尺子
1.2.1.13.8 运动器械（sports- instrument）：足球 单杠
1.2.2 自然物（natural object）
1.2.2.1 天体（celestial body）：太阳 月亮 流星 星星
1.2.2.2 气象（weather）：云 彩虹 晚霞
1.2.2.3 地理（geography）
1.2.2.3.1 地表物（land）：原野 沙漠 山 山洞 陆地
1.2.2.3.2 水域物（water）：江 河 湖 海 河流
1.1.2.2.4 矿物（mineral）：煤矿 原油 铁矿
1.1.2.2.5 元素（element）：金 银 铜 铁
1.1.2.2.6 基本物质（substance）：水 土 灰
1.2.3 排泄物（excrement）：汗 尿 粪便 奶水 眼泪
1.2.4 外形（shape）：粉末 长方形 圆 窟窿 孔 洞 泡
1.3 构件（part）
1.3.1 身体构件（body-part）：头 脸 鼻子 嘴 耳朵 头发 血液 骨头
1.3.2 非生物构件（object-part）：梁 屋檐 车闸 车筐
2 抽象事物（abstraction）
2.1 属性（attribute）
2.1.1 量化属性（measurable）：体积 面积 重量 质量 价格
2.1.2 模糊属性
2.1.2.1 人性（property_of_human）：胆量 勇气 脾气 作风
2.1.2.2 事性（description_of_event）：境况 形势 状态 环节

2.1.2.3 物性（property_of_object）：性能 效用 品种 式样
2.1.3 颜色（color）：黑色 白色 浅色 素色
2.2 信息（information）：话 言语 信件 口信 密码 声明 借口
2.3 领域 (field)：社会 经济 法律 科学 艺术
2.4 法规（rule）：法律 条约 协议 制度 规章 合同 协议 条文
2.5 生理（physiological_state）：瘟疫 疾病 炎症 艾滋病
2.6 心理特征（psychol feature）
2.6.1 情感（feelings）：态度 感情 爱情
2.6.2 意识（cognition）：意图 幻想 兴趣 主意 见解
2.7 动机（motivation）：目的 原因 理由
3 过程（process）
3.1 事件（event）：学潮 球赛 晚会 课 早餐 战争 火灾
3.2 自然现象（natural phenomenon）
3.2.1 可视现象（visible phenomenon）：火 电 光 风 雨
3.2.2 可听现象（audible phenomenon）：声音 雷鸣 风暴
4 时间（time）
4.1 绝对时间（specific time）：宋朝 三国 清代
4.2 相对时间（relative time）：昨天 当代 古代 今天
5 空间（space）
5.1 处所（location）：浙江 西湖 黄山 中国 亚洲
5.2 方位（direction）：东南 前面 之间 途中 高空

2.2.2 形容词（Adjective）

1 事性值（description of event）：紧急 突然 困难 容易 错误 费时
2 物性值 (property of object)
2.1 量化属性值（measurable value）：
2.1.1 浓度（concentration）：浓 稀薄
2.1.2 温度（temperature）：热 冷 凉爽
2.1.3 速度（speed）：快 慢
2.1.4 长度 (length)：长 短
2.1.5 高度 (height)：高 矮 低
2.1.6 宽度 (width)：宽 窄
2.1.7 深度 (depth)：深 浅
2.1.8 厚度 (thickness)：厚 薄
2.1.9 硬度 (rigidity)：硬 软

2.1.10 湿度 (humidity)：潮湿 湿润 干燥
2.1.11 粗细 (degree of finish)：粗 细
2.1.12 松紧 (degree of tightness)：松 紧
2.1.13 大小 (size)：大 中 小
2.1.14 价值 (value)：贵 便宜
2.2 模糊属性值（unmeasurable value）
2.2.1 视感 (vision)：亮 醒目 清晰 混浊
2.2.2 触感 (tactility)：紧 松 粗糙 滑 柔
2.2.3 音质（tone）：响亮 低沉 刺耳
2.2.4 味道 (taste)：酸 甜 苦 辣 可口
2.2.5 性质（quality）：新 旧 真 假 好 坏 强 弱
2.2.6 内容 (content)：空洞 晦涩 清楚 浅显
2.2.7 外形 (shape)：方 圆 尖
2.3 颜色 (color)：红 黄 蓝 绿 鲜艳
3 人性值 (property of human)
3.1 年龄 (age)：年轻 幼小 老
3.2 品格 (character)：善良 博学 幼稚 优雅
3.3 关系 (relation)：亲密 疏远 热情 冷淡
3.4 境况 (condition)：繁忙 贫穷 危险 疲劳
4 空间值 (property of space)
4.1 一维值 (one dimension)：远 近
4.2 二维值 (two dimension)：平 斜 弯
4.2 三维值 (three dimension)：拥挤 杂乱 整齐 满 壮阔
5 时间值 (property of time)：古老 久远 短暂 早 晚

2.2.3 动词 (Verb)

1 静态关系（state）：是 有 等于 包括
2 心理活动（emotion/ cognition）：喜欢 尊敬 反对 同意 怀疑 思考 判断
3 动态行为（event）
3.1 变化（change）：死 病 下降 长高 缩小 变暗
3.2 气象（weather）：下雨 刮风 打雷 起雾
3.3 身体活动（bodily care and functions）：蹬 跳 推 笑 咳嗽 游泳
3.4 五官感觉（perception）：看见 听到 闻着 品尝
3.5 消耗（consumption）：吃 喝 饮
3.6 位移（motion）：跑 走 散步 飞 过来 回去 拉来

3.7 创造（creation）：制作　画　炒　写　创建　修筑
3.8 接触（contact）：触摸　撞击　打中　系　挖掘
3.9 领属转移（possession）：买　卖　赠送　给　转让　借
3.10 信息交流（communication）：告诉　询问　请求　转达　叮嘱　说
3.11 比赛（competition）：竞赛　赛跑　打仗　摔跤　辩论
3.12 社会活动（social behavior）：改革　调价　开会　联欢
3.13 其他行为（other event）

2.2.4 副词（adverb）

1 程度（degree）：很 挺 太 顶 更 最 极 十分 非常 稍 稍微 略微
2 范围（range）：都 也 总 共 一共 总共 统统 只 就 光 仅 仅仅
3 时间（time）：正 刚刚 就 先 曾经 已经 终于 立刻 马上 永远
4 处所（location）：到处　处处　暗中　当场　当面
5 频度（frequency）：常常 常 时常 又 再 还 重新 重
6 方式（manner）：渐渐 逐渐 挨次 挨个 逆时针　慢慢
7 否定（negation）：不 没有 没 未 莫 休 勿 别

2.2.5 数词（Numeral）

1 基数（cardinal number）
　1.1 系数：一 二 两 三 五 六 七 八 九 几
　1.2 位数：十、百、千、万、亿、万万
2 序数（ordinal number）：第一 第二 第十
3 概数（amount）：多半 多少 若干 很多 许多 好多 好几 好些 无数
4 助数 (auxiliary)：又　来　左右

2.3 词语的属性描写

2.3.1 各类词库的共同字段

以下说明中，左边的一列楷体汉字代表字段名，中间的一列数字表示各个字段所占的字节数。右边的词语则是对字段值的说明。

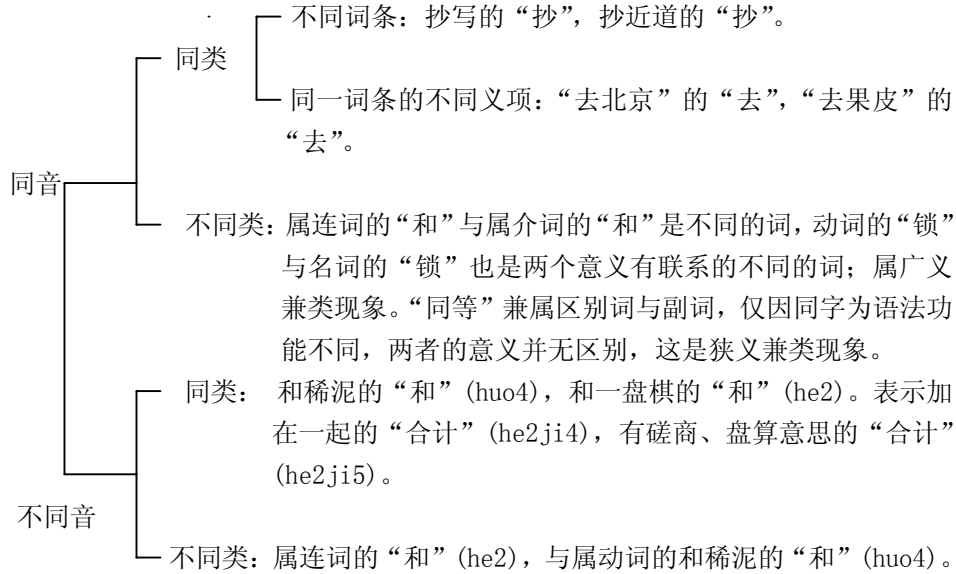
词语　　8　收录 1~4 个字的词语

拼音 24 填每个词语的汉语拼音，声调用 “1，2，3，4，5”表示，其中“5”表示轻声。如：“常识”的全拼音是“chang2shi2”，“尺子”的全拼音是“chi3zi5”。

词类 2 填词语所属词类的代码。如：名词填“n”，动词填“v”，形容词填“a”。

子类 2 填词语所属词类的子类代码。如：名词性成语填“IN”，动词性习用语填“LV”。

同形 2 词典中同形词(即汉字相同的词)的情况是很复杂：



在词典中，除了“同字同音同类”的情况外，上图中同形词的其他情况均作为不同记录收入词典。为了进一步区分同字同音同类的情况，专设了一个“同形”字段。对于同字、同音、同类但是应算不同词项的情况，在“同形”字段中填上字母 A，B，C 等。对于同字、同音、同类、同一个词的不同义项的情况，在“同形”字段中填上数字 1，2，3 等。为了提高同形词的处理效率，在“同形”字段中也用 A，B，C 等标识同字同类不同音的情况。总之，“同形”中的 A，B，C 等表示不同的词，数字 1，2，3 等表示同一个词的不同义项。当需要字母与数字并存时，则将字母置于数字之前，如 A1，A2，B1，B2 等。

兼类 4 填该词语兼属的词类代码，如：名词“锁”的兼类填“v”，动词“锁”的兼类填“n”。

义项 2 对“同形”字段相同的词条进一步加以区分，填上不同的义项编码，如“菜做得很清淡”中的“清淡”在本字段填“1”，“生意清淡”中的“清淡”则填“2”。

释义	10	填写该词语的简明释义，如：词典中收录了两个“天才”，就分别在本字段填上“人”和“智慧”。即前一个“天才”指人，比如可以说“他是一位数学天才”，后一个“天才”指“智慧”，可以说“他在数学方面很有天才”。
语义类	20	填写该词语的语义类别名称。意义明确的尽量填低层的小类；意义难以把握的可酌情填上层语义类。可以不止填一个类别名称，不同的名称之间用“/”隔开。如“校长”填“身份”，“刀”填“用具”，“青菜”填“植物/食物”；“是”填“静态关系”，“喜欢”填“心理活动”，“打雷”填“气象”。
WORD	40	填该词语对应的英语译词或短语，如:“安静”在本字段填“quiet”，“脏乱”填“dirty and messy”。
Ecat	40	填该词语的英语译词的词性代码，或短语组成结构，如:“安静” 在本字段填“A”，“脏乱”则填“!A+C+!A”(!表示中心词)。
备注	20	填写词语用法的简明示例或说明，用“~”代替该词，各示例之间用斜道“/”隔开。

注：除了“语义类、释义、WORD、 Ecat”4 个字段外，上述其他字段均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）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，从而使计算机获得更完备的语法、语义信息。

2.3.2 动词库（包括动词性的成语、习用语、简称略语）

配价数	2	一价动词填“1”， 二价动词填“2”， 三价动词填“3”。 本词典引进“配价”概念来说明一个动词能支配多少名词性成分(陆俭明,1995；袁毓林，1998)。从理论上说，在一个句子中，直接受谓语动词支配的名词性成分不得超过 3 个：主语、宾语[1]、宾语[2]。动词能支配几个名词性成分，它就是几价动词。如果能支配一个名词性成分，则为一价动词，如“奔跑、出差、劳动、前进、病、失败”等；如果能支配两个名词性成分，则为二价动词，如“搬、穿、发明、制定、听见、遗失、是”等；如果能支配 3 个名词性成分，则为三价动词，如“给、给予、问、回答、借、送给”等。配价是从静态的角度依据动词的词汇意义确定的，因此，就某个具体的动词而言，其配价是相对稳定的。 动词同配价成分之间的组合应该是有意义的、可理解的。比如，“他

		跑了一身汗”这个句子中，“跑汗”是无意义的。因而就可以判断出“汗”不是“跑”的配价成分，“跑”（义为“快速前进”）的配价成分只是主语“他”，是 1 价动词。因此，“跑”在本字段填“1”。需要说明的是，汉语中有些动词可以看作没有配价成分，像“例如、可见、天亮”。它们的配价数定为 0，本字段不填。
主体	20	填写动作主体所属的语义类名称。如“逃跑”在本字段填“人类/动物”，“刮倒”填“气象”，“死”填“生物”。 主体（agent）是动词的配价成分承担的一种语义角色，指动作行为（或状态）或自然现象的发出者。如：“敌人逃跑了 / 风刮倒了大树”中的“敌人、风”。 在句子中，主体一般占据主语位置，但有时也可处于宾语位置，如“死了一只兔子”值的“兔子”。
客体	20	填写二价和三价动词的客体语义类名称。如“擦”在本字段填“人为事物/构件”，“画”填“作品”，“丧失”填“抽象事物”。 客体（object）也是动词的配价成分承担的一种语义角色，指动作行为或变化所涉及的直接对象。如“擦玻璃 / 画了一幅画儿”中的“玻璃、画”。在句子中，客体一般占据宾语位置，但在受事主语句或被动句中则处于主语位置，如“玻璃被擦过了 / 画儿画好了 ”。
与事	20	填写三价动词的与事所属的语义类名称。如“给”在本字段填“人类”，“送”也填“人类”。 与事，指事件中有利害关系的间接客体，如受益者或受损者。如“给他一本书 / 送我 30 元钱”中的“他、我”。在句子中，邻体一般占据间接宾语位置，但在主谓谓语句、受事主语句或被动句中则处于一般宾语位置，如“那本书小李给他了/ 那本书给他了/ 那本书叫小李给他了”。

注：为了尽可能详细地描述主体、客体或与事的语义限制，本词典还引入了以下几种符号：

“/”表示“或”，如“叙述”的主体填“人类/作品”；

“~”表示“非”，如“越冬”的主体填“生物~人类”；

“ ”（双引号）表示具体的字、词，如“晒”的主体填“太阳”；

“*”表示任意汉字串，“吹拂”的主体是“*风”（微风、大风、春风、晚风……）。

这 4 种符号的定义同样适用于本词典名词库和形容词库中的“参照体”、“对象”、“主体”等字段。

2.3.3 名词库（包括名词性的成语、习用语、简称略语。时间词、处所词、方位词、代词、数词库的库结构与此相同）

配价数　2　一价名词填“1”， 二价名词填“2”，零价名词不填。

名词的配价表现为支配性名词要求语义上受其支配的从属名词与之共现（袁毓林 1994，1995）。要求一个从属名词与之共现，配价数为 1。如“老李的女儿回来了/ 小坡的爸爸病了”着两句话中的“女儿、爸爸”都是一价名词。因为，从意义上看，它们在表示某事物的同时，还隐含了该事物跟另一个事物之间的某种依存关系。当它在语句中出现时，它要求支配其配价成分。这也就是说，一价名词“女儿、爸爸”不仅是句法上的中心词，而且是语义上的支点，因而在句子中不能省略。如：“老李的女儿——*老李的”、“小坡的爸爸——*小坡的”。而一般名词（零价名词）则可以省略，如：

老李的拐杖—— 老李的

小坡的书包—— 小坡的

要求两个名词性成分与之共现，配价数为 2。如：“这件事老李有意见/ 他对刘刚一直没有好感”，这里的“意见、好感”都是二价名词。从语义上看，“意见、好感”一般是某人针对某人或某物的，涉及到两个个体。因而，在句子中要求两个配项与之共现，如果其中一个配项不出现，那么句子的语义就不完整，如：“老李有意见 / 他一直没有好感”。

汉语中绝大多数名词并不一定要求有任何配项与之共现，如“天下雨了/ 桌子坏了”，“天”和“桌子”就都是零价名词，它们在本字段均不填。

参照体　20　填写一价和二价名词参照体（配项成分）的语义类名称。零价名词在本字段不填。

如“女儿”、“看法”的本字段填“人类”，“桌子”是零价名词则不填。

对象　20　填写二价名词的对象的语义类名称。如“意见”在本字段填上“人类/事件”。一价和零价名词本字段不填。

在句子中，名词的对象一般可以用“对、对于”等介词标记出来，如“群众对他的意见很大”中的“他”。

直接上位　20　填写该名词的直接上位概念。如“雨鞋”、“皮鞋”在该字段均填“鞋”；“轿车”在本字段填“车”。

2.3.4 形容词库 （包括形容词性的成语、习用语、简称略语。状态词、区别词库的库结构与此相同）

配价数	2	一价形容词填“1”，二价形容词填“2”。 在句子中，只要求一个名词性成分与之共现，配价数为 1，如“大雨 / 花很红”中的“大、红”；要求两个名词性成分与之共现，配价数为 2，如“小李对人很热情 / 他对象棋的兴趣淡薄”中的“热情、淡薄”。 汉语形容词绝大多数都是一价的，二价形容词不多，在词义上主要是表示态度、效用及熟悉程度的，如“淡薄、淡漠、冷淡、恭敬、亲热、忠诚、耳熟、陌生、熟、有用、无益、面熟、友好、亲密、忠实”等。
主体	20	填写形容词的主体的语义类名称。如“红（一种颜色）”在本字段填“具体事物”，“友好（亲近和睦）”填“人类/动物”。 主体指性状的承当者，如:“花儿红了 / 她对兔子很友好”中的“花、她、兔子”。 在句子中，主体一般占据主语位置。一价形容词的主体也可以处于偏正结构的中心语位置，如“红花儿 / 大雨”中的“花儿、雨”。
对象	20	填写二价形容词的关涉对象的语义类名称。如“眼熟”在本字段填“具体事物”，“有利”填“人类/抽象事物”。 在句子中，形容词的对象一般可以用“对、对于”等介词标记出来，如“这份合同对甲方有利 / （对）这个人我有点眼熟”。

3 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价值。

目前，SKCC 的二期工程已取得了重要的阶段性成果，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。

致 谢

“现代汉语语义词典”的二期工程有幸得到了北京大学中文系陆俭明教授的大力支持，他不仅给予了我们很多深入细致的语言理论指导，而且从人力、物力上也给予充分扶持，使得课题组始终拥有一个稳定的高素质的开发队伍，为高质量地完成任务奠定了坚实的基础。在此，谨向陆俭明教授及语义词典全体开发人员致以深深的谢意。

北大“973”机器翻译课题组的技术负责人常宝宝博士、王厚峰博士、中国科学院计算技术研究所刘群副研究员、清华大学计算机系周强博士也都对语义词典开发工作给予了热情的关心与有力的支持，并对语义分类体系、属性字段设置等问题提出了很多宝贵的建设性意见。朱学锋副教授、段慧明高级工程师更是为语义词典课题组做出了很多无私的奉献。计算机系李康年同学实现的词典辅助编辑软件，大大加快了工程进度。郭涛小姐对词典数据库作了多角度的细心检查，有效地提高了词典内容的一致性。此外，中文系博士生应晨锦、英语系研究生方宏的认真负责、精益求精也给大家留下了深刻的印象。这里我们一并表示由衷的感谢！

参考文献

[1] Bake, C.F, C.J. Fillmore, and John B.Lowe, 1998, The Berkeley FrameNet Project, In *Proceedings of COLING'98*. pp. 86-90

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

[3] Richardson, Stephen D., 1998, MindNet: acquiring and structuring semantic information from text, In *Coling'98*. pp. 1098-1102

[4] 陈力为，袁琦主编. . 1995. 《中文信息处理应用平台工程》. 北京:电子工业出版社

[5] 陈群秀. 1996. . 信息处理用现代汉语语义分类体系的设计思想. 见：罗振声、袁毓林主编《计算机时代的汉语和汉字研究》，北京：清华大学出版社

[6] 陈晓荷，1998，一个面向工程的语义分类体系. *语言文字应用*，第 2 期

[7] 董大年主编. 1998. 《现代汉语分类词典》. 北京：汉语大词典出版社

[8] 董振东. 1998. . 语义关系的表达和知识系统的建造. . *语言文字应用*，第 3 期

[9] 董振东，董强. 1999. “知网”（Hownet）文献介绍. http:// www.keenage.com

[10]林杏光主编. 1987.《简明汉语义类词典》.. 北京: 商务印书馆
[11]林杏光. 1998. 中文信息界的语义研究谭要.. *语言文字应用*, 第 3 期
[12]陆俭明..1995..《现代汉语配价语法研究 · 序》. 北京: 北京大学出版社. pp..1-7
[13]梅家驹主编. 1983.《同义词词林》. 上海：上海辞书出版社
[14]王惠，詹卫东，刘群.. 1998. 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》. 北京：清华大学出版社. pp 361-367
[15]于江生，俞士汶.. 2002.. “CCD 的结构与设计思想”. *中文信息学报*, No.4. pp 12-20
[16]俞士汶，朱学锋，王惠，张化瑞等. 2003..《现代汉语语法信息词典详解》（第 2 版）. 北京：清华大学出版社
[17]袁毓林. 1994. 一价名词的认知研究.. *中国语文*, 第 4 期
[18]袁毓林. 1995.. 现代汉语二价名词研究.《现代汉语配价语法研究》. 北京：北京大学出版社. pp29-58
[19]袁毓林. 1998..《汉语动词的配价研究》. 南昌：江西教育出版社

附录：语义词典填写样例

词语	词类	同形	语义类	配价数	主体	客体	与事	WORD
发芽	v		变化	1	植物			sprout
修建	v	2	创造	2	人	建筑物		build
赠送	v		领属转移	3	人	实体	人	present
告诉	v		信息交流	3	人	信息	人	tell

表 2　语义词典 SKCC 的动词库部分属性

词语	词类	同形	义项	语义类	配价数	参照体	对象	WORD
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
看法	n			认知	2	人	实体/抽象物	opinion

表 3　语义词典 SKCC 的名词库部分属性

词语	词类	同形	义项	语义类	配价数	主体	对象	WORD
大	a		1	外形	1	具体物		big
大	a		2	性质	1	“雨”/“雪”		heavy
拥挤	a			境况	1	空间/建筑物		crowded
热情	a			关系	2	人	人	warm

表 4　语义词典 SKCC 的形容词库部分属性

The Specification of The Semantic Knowledge-base of Contemporary Chinese

Hui Wang[1], Weidong Zhan[2], Shiwen Yu[1]

[1]（Institute of Computational Linguistics, Peking University, Beijing 100871）

[2]（Dept. 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. It provides a large amount of semantic information such as semantic hierarchy and collocation features for 66,539 Chinese words and their English counterparts. Its 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 and specification of SKCC, and indicates that, as a large scale fundamental semantic resource of Chinese, SKCC will not only provide valuable semantic knowledge for Chinese language processing, but also play an important role in Chinese lexical semantics and computational lexicography research.*

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