信息处理用现代汉语语义分类体系(之一)：属性分类

陈群秀　清华大学计算机科学与技术系
张　普　北京语言学院信息处理研究所

摘　要

本文首先论述了研究信息处理用现代汉语语义分类体系的意义和作用，然后对汉语语义分类体系研究中的几个问题作了说明。最后文章提出了汉语语义分类体系中最感困难的部分--属性分类的一种具体构想。

一. 前言

自然语言理解和处理在现代信息社会舞台上应该扮演和已开始扮演重要的角色，这是因为：①科学技术的高度发展和信息情报的激增迫使人们不得不用高速度大容量现代化工具--计算机来进行语言信息处理；②办公室自动化的兴起对自然语言理解和处理的迫切需求；③计算机的普及和推广应用对自然语言理解和处理的殷切期望（使用自然语言会话方式和计算机交换信息最方便最自然）。

为使计算机具有理解和处理自然语言的能力，必须使计算机拥有丰富的语言知识（例如词汇知识、句法知识、语义知识、语用知识、语境知识）和非语言知识（例如情景、常识等）。其中语义知识的获取和表达是当前一个核心的且又远未解决的问题。在语义知识中词汇意义占很重要的位置，是整个语义系统的基础。因此研制一部信息处理用现代汉语语义词典，不仅是具有奠基性的知识工程，而且也是相当重要的基础理论研究。

在语义分析的诸理论和方法中，语义场理论是有价值的理论之一，义素分析法是基本方法之一。义素分析法不仅揭示了词的语义结构，而且可以说明词的语义特征和各种语义关系，例如同义、多义、下义和不相容[1]。语义场体现义位的关系和区别，也体现语义的系统性。这个系统性是客观世界的系统性在语义中的反映[2]。研制一个以分类语义场为基础的多层次、多类型、多关系、多变化的静态语义网来分析词汇的聚合关系，研究一个基于谓词框架的动态语义网来分析词汇的组合关系，两者结合起来是语义知识获取和表达的一个路子。

研究汉语分类语义场必须先研究汉语语义分类体系。分类的意义和作用在于：分类是人类对事物认识的一种结果，也是人类对事物认识的一种手段。分类从根本上来说是人类认识客观世界的产物，因为茫茫宇宙无边无际，时间空间无始无终，大千世界的万事万物千姿百态千差万别，相互联系相互转化相互依存相互制约。这纷纷繁繁的万事万物及其形形色色的运动或性状的共性或个性，反映到语义系统中就形成分类。可以说分类反映了人们认识客观世界的成果。反之，人们又可利用分类作为对事物认识的一种手段，作为研究语义、表示某种语义知识的手段[3]。分类法刻划事物简洁、清晰、信息密度大，反映了词汇意义中最基本也是信息量最大的属性。当然并不是说分类可以解决全部的问题。"任何简单的分类、单一的关系描述、完全静态的分析都是无济于事的，必须把多层次、多类型、多关系、多变化这些性质综合加以考虑。[4]"不过分类体系描述既可节省存储空间（下位结点可以继承所有上位结点的语义特征），又可避免单个义项描写时遗漏语义特征或不同人描写出现的分歧。

国内对汉语语义分类体系的研究已有一些成果。例如，上海外语学院梅家驹等编写《同义词词林》（12大类、94中类、1428小类、3925词群，近七万个义项）及进行的语义形式化探讨[5]，中国人民大学的林杏光等编写的《简明汉语义类词典》（18大类、1730小类、6万词）和《汉语多用词典》（2大类、38中类、3500小类，　6万词），清华大学黄昌宁等1988年进行的汉语语义词典的研究[6]，北京师范大学汪培庄、陈军90年提出一种可能性语

义表示模型，确定了近四千个语义原语，对现代汉语六万常用词条进行了分类并建造了一个信息库[7]，等等。但目前尚未有一个汉语分类体系完全令人满意。因此，继承国内外已有的研究成果，在借鉴的基础上创造出有汉语自己特色的较为科学合理又具可操作性的语义分类体系是我们的愿望和努力的方向。国外的研究我们参考了日本的四个义类体系：一是日本国立国语研究所的《分类词汇表》（4大类、13中类、798小类、3万词）；二是日本角川书店出版的《类语新辞典》（10大类、100中类、1000小类、6万词）；三是日本CICC的概念词典；四是日本电子化辞书研究所的日本语的意味分类体系[8]。国内的研究我们参考了上海梅家驹等人的《同义词词林》、林杏光等人的《简明汉语义类词典》和台湾中研院的分类体系以及北京语言学院牵头的八五攻关项目的分类体系。在以上基础上初步确定了一个汉语语义分类体系。

二. 几点说明

1. 本论文不想就我们的语义分类体系的整体进行讨论，而只就其中的一部分（即属性的分类）进行讨论（其余部分留待以后再行讨论），以期得到专家和同行们的批评和指点。

2. 在语义学的论著或学术讨论中，常常看到或听见“概念”、“义位”、“义项”等术语。“概念”与“义位”、“义项”既有联系又有区别，在不同场合使用不同的术语。“义项”是词典学中的术语，指“字典词典中同一个条目内按意义列举的项目”。“义位”则是语义场理论中的术语。在静态语义网中，“语义场中具有区别性特征的一个基本概念称为一个义位[9]”，即把一个义项称为一个义位。而“概念”则是思维科学的一个术语，是“思维的基本形式之一，反映客观事物的一般的、本质的特征。”一个概念有时就是一个义项或者一个义位，但有时不是，例如“复合概念”。本文采用语义场理论中的“义位”为主要术语。

3. “义素”、“语义因子”、“语义特征”、“属性”也是一组常常混用的术语。“义素”、“语义因子”、“语义特征”是语义分析时进行义位描述的术语，任何一个义位总是由一个以上的义素（或语义因子）构成的。“属性则是哲学或思维科学使用的术语。属性是指事物所具有的性质、特点”。这些性质和特点在进行义位描述时，常常也就是很重要的义素。属性义素是义素中很重要的一类，但也不是全部，还有领属义素、功能义素、构件义素等。本文对属性采用狭义定义。属性本身也是义位，构成一个庞大的属性义场，属性也有自己的体系和分类。作为义位时，属性有其语义指向；作为义素时，属性有其描述对象。因此分析属性是至关重要的。

4. 建立汉语语义分类体系采用什么作为分类标准？是一个标准还是可以是多个？

如果能用一个标准对世界上的万事万物进行分类一贯到底，那当然是好。但世界上是否存在这种标准？客观世界包含着无限多样的物质形态，一切物质形态都依一定条件互相转化，每种物质形态都有其特殊的属性和结构而与别的物质形态相区别，又与别的物质有各种联系。人类总是通过不同的角度、从不同层面的聚合不断认识事物的各种属性，这些都无法体现在一种标准中。如果只用一个标准来进行分类，则很可能容纳不了所有的事物，也不符合人类认识客观世界的事实，这种分类就没有多大意义和作用。例如，如果只用“物态”这一个标准来分类，虽然“具体物”可全部容纳在其中，但“事”却容纳不了，“抽象物”、“运动类”、“性状类”也包括不进去。即使是“具体物”，也因此节点只具有“固体”、“液体”或“气体”这单一的共同特征而必须在下位各义位上详细描述许多其他的义素，因而失去分类体系的意义和作用。因此，只有从多个角度、多个方面才能认识丰富、复杂、纷繁的世界，因而分类标准就可能不是一个。此时有可能有的义项会同时出现在两个上位之下，即出现交叉现象（兼类）。有人认为此时可作为两个不同的义位，虽然在词典中是一个义项。分类体系中出现部分交叉现象是正常的，符合客观世界的存在规律，也符合人们的认识规律：语义场本身就具有层次性、交叉性和变化性的特点[1]。语义网本来就是三维的而不是二维的，语义分析与

语法分析最大不同点是我们必须从二维表达进入三维表达。例如“国家”这个义项，由于既具有人的意志力等特征因而归类于[人类]这个义位的下位[集体]的下位[国家地区]，而又因具有空间、场所的特征因而也归类于[时空]这个义位的下位[空间]的下下位[公共场所]。这是人类从不同角度认识同一事物的结果，也是事物形成复杂特征集的原因。一个义项就是一个义素集，其中的义素与其他义项的义素有着各种联系。因此分类体系中象“国家”这样一个义位同时出现在两个不同上位之下是正常的。当然我们应尽可能把兼类（分叉现象）控制在少量范围。

为使读者对我们的语义分类体系有个概貌，在详细介绍属性分类之前将我们语义分类体系作一个简略的总体介绍。

首先按辨证唯物主义观点把语义类分为运动类和事物类（含性状类）两大类（即顶层、0层）。运动类作为动态语义网另行研究。事物类静态语义体系根据“主要在时间上延展的事物”、“主要在空间上（包括思维空间）延展的事物”和“时空”本身一分为三：[事]、[物]、[时空]（第一层）。[物]根据“±形”、“±色”、“±质量”等基本义素分为[具体物]、[抽象物]两类（第二层）。[抽象物]则分为[精神活动工具]、[精神活动因和源]、[精神活动和产物]、[其他抽象物]、[属性]、[关系]六类（第三层）。[关系]则分为[社会关系]、[位置关系]、[时间关系]、[因果条件目的关系]、[逻辑关系]等七类（第四层）。[社会关系]又分为[血缘关系]、[地缘关系]、[业缘关系]、[爱缘关系]、[利缘关系]、[其他社会关系]等六类（第五层）。[属性]又分为[具体物属性]、[抽象物属性]两类（第四层）。另外还有两种属性（即“构件属性”和“功能属性”），因它们的特性不属于分类义场类型，将在“构件子义场”和“功能子义场”另行研究。本文只对事物类（含性状类）分类体系中的[属性]这一部分进行展开，而不涉及构件和功能属性。

5. 在实际进行分类时，有时碰上一些特殊问题进行某些特殊处理。例如“人类”按常识应放在[动物]的下位，但考虑到人与动物有许多本质区别，而且人是社会活动的主体，其自然属性和社会属性比其他动物多得多，是语义描述的轴心，因而由[动物]的下位提升到与[动物]并排成为同位。又如，有的动物类至今快绝迹或下属动物太少，人们又知之甚少或在信息处理中碰上的机率太低，因而不单独成类而归入一个混和类[其他动物]。还有“家禽家畜”和“宠物”，虽与[兽]、[鸟]等相关，但因其或有较大经济价值或与人类生活有很大关系因而也单独成类[家禽家畜]、[宠物]。再者，语义场中每个节点按理说都是义位，应有继承性，但由于时间原故和研究尚未完全成熟，因而有的中间结点现在或者用的是复合概念，或者用的是并列词组，有的甚至不大合适，这个问题有待今后继续修改完善。另外，为了分类体系的可操作性，有时作一点技术性处理。

三. 属性分类

属性分类是语义分类系统中最感困难的部分，也是需要认真对待的部分。下面表示的是试作的汉语语义分类体系中属性分类的部分。

2.2.5 属性
2.2.5.1 具体物属性
2.2.5.1.1 具体物一般属性
2.2.5.1.1.1 定量属性
2.2.5.1.1.1.1 数
2.2.5.1.1.1.2 量
2.2.5.1.1.1.3 数量名
2.2.5.1.1.1.4 频度名
2.2.5.1.1.1.5 顺序名
2.2.5.1.1.1.6 比例名
2.2.5.1.1.1.7 数型
2.2.5.1.1.2 定性属性

2.2.5.1.1.2.1 数量定性　　2.2.5.1.1.2.2 频度定性
2.2.5.1.1.2.3 顺序定性　　2.2.5.1.1.2.4 空间定性
2.2.5.1.1.2.5 时间定性　　2.2.5.1.1.2.6 物理定性
2.2.5.1.1.2.7 化学定性
2.2.5.1.1.3 空间属性
2.2.5.1.1.3.1 形状
2.2.5.1.1.3.1.1 外状　　2.2.5.1.1.3.1.2 内状
2.2.5.1.1.3.2 大小
2.2.5.1.1.3.2.1 长度　　2.2.5.1.1.3.2.2 面积
2.2.5.1.1.3.2.3 体积
2.2.5.1.1.3.3 位置
2.2.5.1.1.3.4 距离
2.2.5.1.1.3.5 范围
2.2.5.1.1.4 时间属性
2.2.5.1.1.4.1 时点　　2.2.5.1.1.4.2 时段
2.2.5.1.1.4.3 历时
2.2.5.1.1.5 物理属性
2.2.5.1.1.5.1 物态重量质量
2.2.5.1.1.5.1.1 物态　　2.2.5.1.1.5.1.2 重量
2.2.5.1.1.5.1.3 质量
2.2.5.1.1.5.2 颜色色调新旧度
2.2.5.1.1.5.2.1 颜色　　2.2.5.1.1.5.2.2 色调
2.2.5.1.1.5.2.3 新旧度
2.2.5.1.1.5.3 温度湿度浓度
2.2.5.1.1.5.3.1 温度　　2.2.5.1.1.5.3.2 湿度
2.2.5.1.1.5.3.3 浓度
2.2.5.1.1.5.4 硬度强度密度
2.2.5.1.1.5.4.1 硬度　　2.2.5.1.1.5.4.2 强度
2.2.5.1.1.5.4.3 密度
2.2.5.1.1.5.5 光洁透明亮度
2.2.5.1.1.5.5.1 光洁度　　2.2.5.1.1.5.5.2 透明度
2.2.5.1.1.5.5.3 亮度
2.2.5.1.1.5.6 弹性延展性松紧性
2.2.5.1.1.5.6.1 弹性　　2.2.5.1.1.5.6.2 延展性
2.2.5.1.1.5.6.3 松紧性
2.2.5.1.1.5.7 新旧完整纯度
2.2.5.1.1.5.7.1 新旧度　　2.2.5.1.1.5.7.2 完整度
2.2.5.1.1.5.7.3 纯度
2.2.5.1.1.5.8 导电导热耐磨性
2.2.5.1.1.5.8.1 导电性　　2.2.5.1.1.5.8.2 导热性
2.2.5.1.1.5.8.3 耐磨性
2.2.5.1.1.5.9 气味味道音质
2.2.5.1.1.5.9.1 气味　　2.2.5.1.1.5.9.2 味道
2.2.5.1.1.5.9.3 音质

2.2.5.1.1.5.10 型式构件
2.2.5.1.1.5.10.1 型式　　2.2.5.1.1.5.10.2 构件
2.2.5.1.1.6 化学属性
2.2.5.1.1.6.1 酸碱性　　2.2.5.1.1.6.2 可燃性
2.2.5.1.1.6.3 耐酸性　　2.2.5.1.1.6.4 耐碱性
2.2.5.1.1.6.5 化合性质
2.2.5.1.2 非生物社会属性
2.2.5.1.2.1 价值　　2.2.5.1.2.2 价格
2.2.5.1.2.3 功能
2.2.5.1.3 生物属性
2.2.5.1.3.1 生理属性
2.2.5.1.3.1.1 种族　　2.2.5.1.3.1.2 性别
2.2.5.1.3.1.3 年龄　　2.2.5.1.3.1.4 体形
2.2.5.1.3.1.5 体格　　2.2.5.1.3.1.6 仪容
2.2.5.1.3.1.7 健康　　2.2.5.1.3.1.8 肤色发色
2.2.5.1.3.1.9 血型　　2.2.5.1.3.1.10 生理成熟度
2.2.5.1.3.2 心理属性
2.2.5.1.3.2.1 智力　　2.2.5.1.3.2.2 习性
2.2.5.1.3.2.3 气质风度　　2.2.5.1.3.2.4 样态表情
2.2.5.1.3.2.5 兴趣爱好　　2.2.5.1.3.2.6 性格素质
2.2.5.1.3.2.7 感情心情　　2.2.5.1.3.2.8 心理成熟度
2.2.5.1.3.2.9 心理健康程度
2.2.5.1.3.3 社会属性
2.2.5.1.3.3.1 基本社会属性
2.2.5.1.3.3.1.1 姓名　　2.2.5.1.3.3.1.2 出生地日期
2.2.5.1.3.3.1.3 民族　　2.2.5.1.3.3.1.4 国籍
2.2.5.1.3.3.1.5 籍贯
2.2.5.1.3.3.2 家庭情况
2.2.5.1.3.3.2.1 家庭出身　　2.2.5.1.3.3.2.2 婚姻状况
2.2.5.1.3.3.2.3 家庭住址　　2.2.5.1.3.3.2.4 社会关系
2.2.5.1.3.3.2.5 经济状态
2.2.5.1.3.3.3 学历简历职称
2.2.5.1.3.3.3.1 学历文化程度　　2.2.5.1.3.3.3.2 学位
2.2.5.1.3.3.3.3 职称职务军衔　　2.2.5.1.3.3.3.4 职业专业
2.2.5.1.3.3.3.5 就业工资　　2.2.5.1.3.3.3.6 外语种类水平
2.2.5.1.3.3.4 品德才能态度
2.2.5.1.3.3.4.1 品德品质作风　　2.2.5.1.3.3.4.2 才能
2.2.5.1.3.3.4.3 态度(待人)　　2.2.5.1.3.3.4.4 态度(政治)
2.2.5.1.3.3.4.5 人间关系　　2.2.5.1.3.3.4.6 社会地位身份
2.2.5.1.3.3.4.7 政治面貌　　2.2.5.1.3.3.4.8 宗教信仰
2.2.5.1.3.3.4.9 涵养心胸
2.2.5.2 抽象物属性
2.2.5.2.1 抽象物状态
2.2.5.2.1.1 社会状态　　2.2.5.2.1.2 环境状态

2.2.5.2.1.3 活动状态　　2.2.5.2.1.4 意境状态
2.2.5.2.1.5 气氛状态　　2.2.5.2.1.6 思想状态
2.2.5.2.2 抽象物评价
2.2.5.2.2.1 科学系统正确性
2.2.5.2.2.1.1 科学性　　2.2.5.2.2.1.2 系统性
2.2.5.2.2.1.3 全面性　　2.2.5.2.2.1.4 一致性
2.2.5.2.2.1.5 规范性　　2.2.5.2.2.1.6 正确性
2.2.5.2.2.1.7 精确性　　2.2.5.2.2.1.8 真伪性
2.2.5.2.2.1.9 深刻性　　2.2.5.2.2.1.10 充分性
2.2.5.2.2.2 稳定重要必然性
2.2.5.2.2.2.1 稳定性　　2.2.5.2.2.2.2 安全性
2.2.5.2.2.2.3 重要性　　2.2.5.2.2.2.4 有效性
2.2.5.2.2.2.5 必然性　　2.2.5.2.2.2.6 必要性
2.2.5.2.2.2.7 典型性　　2.2.5.2.2.2.8 基础性
2.2.5.2.2.3 复杂详尽困难度
2.2.5.2.2.3.1 复杂度　　2.2.5.2.2.3.2 详尽度
2.2.5.2.2.3.3 困难度　　2.2.5.2.2.3.4 透明度
2.2.5.2.2.3.5 清晰度　　2.2.5.2.2.3.6 新旧度
2.2.5.2.2.4 周期规律逻辑性
2.2.5.2.2.4.1 周期性　　2.2.5.2.2.4.2 顺序性
2.2.5.2.2.4.3 规律性　　2.2.5.2.2.4.4 逻辑性
2.2.5.2.2.4.5 条理性　　2.2.5.2.2.4.6 周密性
2.2.5.2.2.4.7 随机性
2.2.5.2.2.5 质量规模价值
2.2.5.2.2.5.1 质量　　2.2.5.2.2.5.2 规模
2.2.5.2.2.5.3 价值　　2.2.5.2.2.5.4 水平
2.2.5.2.2.5.5 范围　　2.2.5.2.2.5.6 功能
2.2.5.2.2.6 可读可测可计算性
2.2.5.2.2.6.1 可读性　　2.2.5.2.2.6.2 可测性
2.2.5.2.2.6.3 可计算性　　2.2.5.2.2.6.4 可解性
2.2.5.2.2.6.5 可操作性　　2.2.5.2.2.6.6 可懂度
2.2.5.2.2.6.7 可视性
2.2.5.2.2.7 伟大光荣辉煌性
2.2.5.2.2.7.1 伟大性　　2.2.5.2.2.7.2 光荣性
2.2.5.2.2.7.3 远大性　　2.2.5.2.2.7.4 辉煌性
2.2.5.2.2.7.5 奇妙性
2.2.5.2.2.8 亲密和谐协调性
2.2.5.2.2.8.1 亲密度　　2.2.5.2.2.8.2 和谐性
2.2.5.2.2.8.3 协调性
2.2.5.2.2.9 内容结构情节
2.2.5.2.2.9.1 内容　　2.2.5.2.2.9.2 结构
2.2.5.2.2.9.3 情节　　2.2.5.2.2.9.4 手段
2.2.5.2.2.9.5 目的

四. 结语

由于时间、水平、占有的资料所限，上面提示的有关属性分类的构想还不够成熟和完善。我们撰写本文的目的仅仅在于期望得到专家和同仁们的批评、指导和建设性意见，以使汉语语义分类体系更加成熟、更加完善。今后努力方向有三：一是继续听取专家同行的意见，修改充实完善分类体系；二是针对汉语常用词语主要义项做具体归类工作，以验证、修改、完善、充实这一体系；三是把汉语语义分类体系应用到实际的自然语言理解和处理系统中去。例如，应用到中文智能输入系统中去解决同音字词的自动选择问题；应用到机器翻译系统中去解决多义词选择义项和介词译文选择问题；应用到汉语语素自动标注中去选择义项进行自动标注；应用到汉语述语动词机器词典的研究方面去帮助研究动词论旨角色的语义限制和选择问题；应用到句法分析器中去解决语义指向、多义排歧问题，等等。希骥通过大家的努力能得到一个较为科学合理的、有汉语特色的、可操作性强的汉语语义分类体系。

参考文献

〔1〕 王家国:“论义素分析”，上海大学国际商业学院，《汉语语义学论文集》，湖南人民出版社，1986年，长沙。

〔2〕 贾彦德:《语义学导论》，北京大学出版社，1986年7月，北京。

〔3〕 陈群秀:“有关语义分类体系研究的几个问题，《机器翻译研究进展》，电子工业出版社，92年8月。

〔4〕 张普:“信息处理用现代汉语语义分析的理论和方法，”《中文信息学报》1991年第三期。

〔5〕 梅家驹、高蕴琦:“语义形式化的研究”，上海外国语学院。

〔6〕 黄昌宁、陈祖舜:“关于语义辞典构造的一些初步设想”，清华大学计算机科学与技术系，《中文信息学报》1988年第3期。

〔7〕 陈军:“可能性语义表示与自然语言理解”，北京师范大学数学系，博士论文，导师：汪培庄，1990年2月。

〔8〕 荻野孝野：日本语的意味分类体系，计量国语学第十六卷第三号，1987年。

〔9〕 张普:“论语义场”，《机器翻译研究进展》，电子工业出版社，1992年8月。

A Thesaurus System of Contemporary Chinese in Information Processing (Part One): The thesaurus of Property

Chen Qunxiu

Dept. of Computer Science & Technology, Tsinghua University

Zhang Pu

Language Information Institute, Beijing Language Institute

ABSTRACT

At first, this paper introduces the significance and usage on research of thesaurus system of Contemporary Chinese in information processing. Then the paper explains some problems in the research of thesaurus system of Contemporary Chinese. At last, the authors give a suggest on dealing with the hardest part in thesaurus system of Contemporary Chinese : thesaurus of Property.