中文概念词典的结构

于江生　　俞士汶
北京大学计算语言学研究所，100871

摘要：中文概念词典[1]（Chinese Concept Dictionary，简称 CCD）是北京大学计算语言学研究所开发的与 WordNet 兼容的汉语语义词典。本文着重描述了 CCD 的结构：CCD 中的“概念”用同义词的集合定义，CCD 的主关系---概念之间的继承关系（即上下位关系）和一些附加关系使得 CCD 形成一个概念网络，其上的演绎规则是严格形式化了的，可应用于中文的语义分析。

关键词：概念，同义词集合，CCD，WordNet，计算词典学

The Structure of Chinese Concept Dictionary

YU Jiangsheng　　YU Shiwen
Institute of Computational Linguistics
Peking University, Beijing, 100871

Abstract: Chinese Concept Dictionary (CCD) is a WordNet-like semantic lexicon, developed by the Institute of Computational Linguistics, Peking University. This article focuses on the structure of CCD, which presents a concept defined by a set of synonyms (SynSet) and a network of concepts based on the hypernymy hierarchy, the basic relationship, with other supplementary relationships. The deductive rules on this semantic network are mathematically formalized, which could be well applied to Chinese semantic analysis.

Keywords: concept, SynSet, CCD, WordNet, Computational Lexicology

引言

二十世纪八十年代末至今，自然语言语义分析的理论和技术一直是计算语言学的研究热点，而语义分析所依赖的语言知识表示中最重要的初始环节就是语义词典。给定语言 L 的词语的有限集合 Σ，人们可以从不同的基本假设或应用目的出发构建 L 的语义词典。[2]泛泛地说，构建语义词典就是由 Σ 诱导出集合 Lex，并在 Lex 上定义各种关系使之结构化。语义词典的一个非常好的范例是 WordNet，它是在普林斯顿大学认知科学实验室 G. Miller 教授的指导下开发的，真正的、有实际意义的工作始于 1985 年。作为心理语言学家的试验品，WordNet 的最初设计并不是直接受计算语言学的影响或直接为自然语言处理服务的。八十年代末，语义计算需要一部结构良好的概念词典[3]，计算语言学家发现了 WordNet 并将之应用于自然语

[1]本项目得到了 973（G1998030507-4）、国家自然科学基金（69483003）和北京大学 985 项目的支持。
[2]语义词典（如 WordNet，FrameNet，MindNet，HowNet 和北大计算语言学研究所开发的面向汉英机器翻译的语义词典）在结构描述上差异也是计算词典学关注的主要问题之一。
[3]在通常情况下，人们根据应用所需要的语义信息和语义推理来确定语义词典的结构，CCD 的构建主要面向

言处理中涉及语义分析的诸多领域。WordNet 的基本思想是简单明确的，它的形式化做得很彻底（详见[2],[5],[6],[12],[13]等）。目前，WordNet 已经成为一个事实上的国际标准，从 EuroWordNet（多语的概念词典，详见[7]和[20]）发展的事实不难看出，WordNet 框架的合理性已被词汇语义学界和计算词典学界所公认。CCD 的描述语言和基本关系与 WordNet 是兼容的[4]，对 CCD 的大致描述是：

1. 在线的词汇语义的索引系统：词汇关系在词之间体现，语义关系在概念之间体现。
2. 用同义词集合（SynSet）表示一个概念，这个特征常用来区别于其他的语义词典。
3. 概念间的上下位关系是 WordNet 和 CCD 结构中的主关系，上下位关系所确定的概念标记森林附加上其他关系（如，对立关系、部分整体关系等）构成整个概念网络。

显然，我们无法脱离词语谈论概念。一个词语可以承载多个概念，一个概念也可能由多个词语来体现，即词语集合Σ与概念集合Γ之间的对应关系是多对多（不是映射）。不论对语言学研究还是计算语言学研究，都要搞清楚下面三个基本问题：

三个基本问题	CCD 的回答
什么是L中的概念？	L中的一个概念就是一个同义词集合
$\forall w\in\Sigma$，什么是w的词义？	一个词语w的词义$\Delta(w)$就是包含w的所有 SynSet 的集合，即w在所有语境中的所有可能的语义
概念与词义的关系？	虽然概念集合Γ和词义集合Δ之间不存在映射关系，但如果$C\in\Delta(w_1)\cap\Delta(w_2)$，则$C\in\Gamma$且$\{w_1,w_2\}\subseteq C$

Wittgenstein 在[29]中认为词的意义等同于词的使用[5]，越来越流行的观点是将概念作为对象的功能化（在这种意义上概念和对象可以视为同一个东西）。外延（指称）和内涵（意义）是概念的两个对立统一的表现，在 Frege 那里就已经搞清楚了，[6]二者对于概念体系的构造应该同等地重要。一旦区分了意义理论和指称理论，意义理论所要研究的首要问题就仅限于语言形式的同义性[7]和陈述的分析性，而意义本身则丝毫不重要了。以此观点为哲学基础，三个基本问题就可以有若干个不同的回答，这是由对语言形式的同义性的不同理解造成的（详见[26]）。

1 CCD 中的概念

CCD 中的名词实际上将包括语法上的体词、一部分区别词（金、银、男、女）、一部分后接

信息提取、信息检索和机器翻译等自然语言处理领域。

[4] CCD 不是 WordNet 的汉化，它的知识表示反映的是汉语的特点，主要面向的也是中文信息处理领域。

[5]这个观点可追溯到 Leibniz，在[9]中他说，“……，而我毫不怀疑，一个瞎子也能确切地来谈论颜色，并且作一篇成本大套的演说来颂赞他所不认识的光，因为他曾学会了解了它的效果和有关的种种情况”。事物和观念都是用词语来标志的，Leibniz 认为二者没有什么不同，“人们有时甚至就实质性地谈到词语，而在这种场合不能确切地以词语的意义、词语与观念或事物的关系来代替词语；这种情况不仅当人们作为语法学家来说话时是这样，而且当人们作为词典编撰者来说话，给名词以解释时也是这样。”而 Russell 在[18]中主张将常识中的事物等同于靠共现的关系联系起来的性质的集合。

[6]外延相同并不能推导出意义相同，譬如**晨星**和**暮星**，**有心脏的动物**和**有肾脏的动物**等等。

[7]对同义性的定义方式就是对意义的理解方式。

成分（性、器、仪、机）、一部分语素（民、柿）、一部分成语（八拜之交、铜墙铁壁）、一部分习用语（木头疙瘩、光杆司令）和一部分简称略语（政协）。CCD 名词还包括动词中的一个子类“名动词”和形容词的一个子类“名形词”。因此，CCD 中的名词应该是“广义的名词”。为了简便，仍叫名词（这也是为了同 WordNet 兼容）。同样，CCD 中的动词包括语法上的动词、谓词性代词。

1.1 可替换性原则与概念

我们假设只有那些经常使用的概念才会有承载它们的词语。从Σ构建Γ，一个直接的思路是：是否可以构造$Lex \subset 2^{\Sigma}$和单射$i: Lex \to \Gamma$使得$i(Lex)$在Γ中是高频的？由于Lex与$i(Lex)$之间是一一对应的，所以形式上可以规定Γ的子集$i(Lex)$用Lex来表示。形式刻画可以稍微再复杂一点：用二元组$< x, R_x >$来表示Γ中的某些概念，其中$x \in Lex$且R_x是集合x上的关系的集合，譬如 MindNet 和 HowNet。词典结构一方面受词典应用范围的制约，另一方面受构造代价的限制。WordNet 和 CCD 利用词语的同义性来构建Lex，即$\forall x \in Lex, R_x$仅仅包含同义关系。

定义 两个词语是同义的当且仅当它们在某个（或某些）语境中可以相互替换而不改变该语境的语义。[8]

在自然语言中不存在绝对的同义性。例如，**电脑**和**计算机**，在下面的语句中就不可替换：$S =$ **计算机**这个词由三个汉字组成。只要词语w_1和w_2的形式不同，我们总可以构造出例句使得它们不可替换。所以，CCD 中词语之间的同义关系中也包含近义关系，例如，{老师,教师,教员}和{春天,春季}。在 CCD 中，一个概念是用一个同义词集合（SynSet）来表示的（判定两个概念C和C'不同的依据是表示C和C'的同义词集合不同）。

定义 1.1.2 可替换性原则：对于 CCD 中任何合理的（well-defined）SynSet，必然存在某个非平凡的语境S使得这个 SynSet 中所有的词语在语境S中两两可替换而不改变S的语义。Quine 在[16]中论述了可替换性原则是认识同义性的充分条件。但是，判定两个语言形式的同义性完全是后验的，必须存在足够数量的可能语境使得可替换性原则得以验证。

1.2 上下位关系－CCD 中概念之间的主要关系

词语只是参与构造概念，所以从概念构造的角度看，CCD 刻画了词语之间的聚合关系。CCD 中的各种关系都是概念之间的，这并不妨碍组合性原则的使用---与传统的语义词典不同的是，CCD 虽然不是直接谈论词语之间的组合关系，但组合关系可以投射到词语之间。当确定给定语句中的某个词语w所承载的概念为$C = \{w_1, w_2, \cdots, w_k\}$时，保持语义不变的同义词替换从表象上看似乎机器“理解”了w的语义（即C在给定语境中的正确使用）。

定义 CCD 中名词（或动词）概念之间的主要关系是上下位关系：概念C'称为概念C的下位概念（hyponymy concept）或概念C是概念C'的上位概念（hypernymy concept）当且仅当命题C' is a kind of C为真。

[8]按照 Leibniz 的说法，同义性就是保全真值的相互替换性。从这个例子可以看出 Leibniz 的定义限制过强，并且只停留在逻辑层面，本文中的定义较之更广泛和实用一些。

定义 1.2.1 概念C是概念C'的祖先概念或概念C'是概念C的子孙概念当且仅当存在概念$C_1,C_2,\cdots,C_n$使得C'是C_1的下位概念，$\cdots,C_n$是C的下位概念。

例 概念{树,树图}的祖先概念为：

树，树图---（从一个根出发的所有枝杈构成的图形；“家谱树”）

$\Rightarrow$ 平面图，二维图形---（2-维图形）

$\Rightarrow$ 图形---（点、线、面所组合而成的可视的形）

$\Rightarrow$ 形状，形式---（事物的空间排列；“几何是关于形状的数学科学”）

$\Rightarrow$ 属性---（属于一个个体的抽象或特征）

$\Rightarrow$ 抽象---（从具体的事例中抽取公共特征从而形成的一般概念）

上下位关系所确定的概念层次结构将应用于信息提取与信息检索语等涉及语义推理的系统：当检索{灾难}时，系统也将搜索诸如{地震}、{暴风雪}、{洪水}等下位概念---这是理解词义的一个方面，正如 Russell 在[19]中所论述的，即听到或看到该词时能够“合适地行动”。面向自然语言理解的语义词典，为达到“合适地行动”，必然要在词语之间和概念之间定义许多关系。关键的问题是如何使用这些关系，即考虑元规则系统（如果元规则系统过于复杂，所付出的代价是语义词典变得难以应用）。CCD 规定在每个语义范畴中，上下位关系所确定的概念层次结构是一棵标记树，名词（或动词）概念按照上下位关系形成一个森林，其他关系都是次要地附着于标记森林上（见[21]）。由于 CCD 中概念之间的关系较少，并且每个关系的逻辑含义都是清晰的，所以 CCD 上的演绎体系也是可以严格刻画的。

性质 上位关系（下位关系、祖先关系和子孙关系）是传递和反对称的。

1.3 概念之间的对立关系

CCD 中概念之间的对立关系不是名词（或动词）概念的基本组织原则。

定义 概念C与C'是对立的（opposite）当且仅当C的内涵是$\neg C'$。

两个名词（或动词）概念C和C'如果是对立的，则存在共同的祖先概念。另外，名词（或动词）之间的反义关系是基于概念对立关系的，即两个词语w和w'有反义关系的必要条件是w所在的某个概念C与w'所在的某个概念C'之间有对立关系。

例 对立概念中词之间的反义关系的例子如下：

1. {冬天,冬}$=\neg$ {夏天,夏}$\Rightarrow <$冬天;夏天$> \oplus <$冬;夏$>$
2. {国王,男君主}$=\neg$ {王后,女君主}$\Rightarrow <$国王};王后$> \oplus <$男君主;女君主$>$
3. {丈夫,老公,夫君}$=\neg$ {妻子,老婆,内人}$\Rightarrow <$丈夫,老公;妻子,老婆$>$

其中，$<a_1,a_2,\cdots,a_m;b_1,b_2,\cdots,b_n>$表示词语$a_i$与$b_j$有反义关系，$i=1,2,\cdots,m$且$j=1,2,\cdots,n$。

性质 概念间的对立关系是对称的。

2 CCD 中的结构

计算词典学主要解决在某个词汇语义的框架内如何实现词典到语法的形式映射的问题，词典设计者应该至少完成下列目标，本节也将分别从这五个方面介绍 CCD 的结构：（1）语义范畴的分类；（2）词语与词义的形式表示之间的对应关系；（3）语义关系及其性质；（4）动词变元的闭合语义约束；（5）组合性原则。

2.1 CCD 的语义范畴

下面的两个表描述性地[9]给出了 CCD 名词和动词的初始概念：

动作行为	交流通信	位置处所	过程	形状
动物	事件	动机	植物	状况状态
人工物	感觉情感	自然物	所有物	时间
性质属性	食物	自然现象	数量度量	物质
知识认知	团体群体	人物人类	关系	身体躯体

25 个名词初始概念

身体动作	变化	通信
竞争	消费	接触
认知心理	创造	运动
情感心理	状态	感知
领属	社会交互	气象

15 个动词初始概念

CCD 动词初始概念就是动词的语义范畴。在名词初始概念中，有一些是有共性的，按照上位关系可以为它们找到一个共同的父结点，例如， {动物}，{人类,人}，{植物}都是{生物体,生物}的下位概念；{人工物,人造物}，{自然物}，{物质}都是{非生物体}的下位概念；而{生物体,生物}和{非生物体}又是{实体}的下位概念（见附录）。

例 同一个概念可以出现在两个不同的初始概念的子孙概念中，例如附录中图 1 所示，范畴 A 中的节点 011 对应的概念 C 与语义范畴 B 中的节点 012 对应的概念 C 相同。

初始概念是从心理语言学的角度得到的，尽可能地做到其子孙概念集合两两没有重叠的部分。初始概念集合定义得是否合理应该有一个衡量标准，CCD 要求名词（或动词）初始概念的子孙概念集合构成名词（或动词）概念的一个弱划分[10]。当然还可以引进范畴标记从形式上区分 SynSet 相同但初始概念不同的两个概念，譬如该例中的两个概念可以用分别用 C_A 和 C_B 来表示。总而言之，初始概念的子孙概念集合要求尽量做到是概念集合的划分，即便 SynSet 做不到，引入范畴标记也可以帮助我们从形式上做到一个划分。

[9]在 CCD 中，每个初始概念都由 SynSet 精确定义。

[10]集合 $P \subset 2^S$ 称为集合 S 的一个划分（partition）当且仅当 $\cup P = S$ 并且 $\forall a,b \in P$，如果 $a \neq b$，则 $a \cap b = \varnothing$。划分确定一个等价关系，反之亦然。集合 $P \subset 2^S$ 称为集合 S 的一个弱划分（weak partition）当且仅当 $\cup P = S$ 并且 $\forall a,b \in P$，如果 $a \neq b$，则 $|a \cap b| / \min\{|a|,|b|\} << 1$。显然，对于任意集合 S，S 的一个划分也是 S 的一个弱划分。

2.2 词语与词义的形式表示之间的对应关系

Frege 在 Grundlagen der Mathematik（1884 年）中指出，一个词的语义应该在某个具体的语境中被确定，而不应该孤立地分析。也就是说，离开具体的语境，我们无法谈论 w 的词义。由于词义消歧（word sense disambiguation，简称 WSD）是语义分析的关键步骤，因而首先必须搞清楚什么是词义。CCD 中一个词语 w 的词义 $\Delta(w)$ 就是包含 w 的所有 SynSet 的集合，即 w 所有可能的语义的集合，映射 $w \mapsto \Delta(w)$ 给出了词语与词义之间的对应关系。例如，词语树所对应的词义是集合{{树}, {树,树图}}。WSD 的过程就是在具体给定的语境中挑选出 $\Delta(w)$ 中某个元素的过程。一旦词语 w 的语义在给定的语境中被确定为概念 C，[11]机器就能够用 C 中 w 的同义词替换 w 而造出新的句子，表现出它“理解”了 w 的语义。

性质 如果 $C \in \Delta(w_1) \cap \Delta(w_2)$，则 $C \in \Gamma$ 且 $\{w_1, w_2\} \subseteq C$。

2.3 CCD 概念间的关系及其性质

在 CCD 中，名词（或动词）概念之间除了有上下位关系和对立关系，还有一些其他的关系[12]。名词概念之间有部分整体关系（meronymy-holonymy relation）。动词概念之间按照时段分为四种关系：有上下位关系的两个动词概念在时段上是重合的；一个动词概念的时段如果完全嵌入到另一个动词概念的时段中，它们之间的关系就可以类比为名词概念间的部分整体关系；一个动词概念发生的必要条件是另一个动词概念发生；一个动词概念发生致使另一个动词概念发生。最后，我们将谈论名词概念和动词概念之间的联想关系。需要说明的是名词概念对动词概念的闭合语义约束是通过约束句子框架中的角色变元实现的（详见 2.4 节），比本节描述的概念间的关系要复杂得多。

2.3.1 名词概念的部分整体关系

定义 名词概念 C' 是 C 的整体（holonym）或 C 是 C' 的部分（meronym）当且仅当命题 C is a part of C' 为真。CCD 中共有六种部分整体关系：

1. $C \#_p C'$ 表示个体 C 是个体 C' 的一部分，例如，{车轮} $\#_p$ {车}
2. $C \#_m C'$ 表示个体 C 集合 C' 的元素，例如，{树} $\#_m$ {森林}
3. $C \#_s C'$ 表示 C 是 C' 组成材料，例如，{钢} $\#_s$ {钢板}
4. $C \#_a C'$ 表示 C 是 C' 的部分区域，例如，{石家庄} $\#_a$ {河北}
5. $C \#_e C'$ 表示事件 C 是事件 C' 的一部分，例如，{开幕式} $\#_e$ {会议}
6. $C \#_t C'$ 表示时间 C 是时间 C' 的一部分，例如，{唐朝} $\#_t$ {古代}

WordNet 名词概念的部分整体关系只有前三种。部分整体关系与上下位关系是不同的，即如果两个名词概念有上下位关系，那么它们之间就不可能有部分整体关系，反之亦然。

---

[11] CCD 动词概念的句子框架经过了闭合语义约束，有利于确定给定语境中的 w 的词义。

[12] [21]描述了名词概念标记树中兄弟节点之间某些偏序关系，本文不作介绍。

性质 部分整体关系满足传递性、反对称性。

2.3.2 动词概念间的关系

在自然语言中，有些动词概念之间存在蕴涵关系，可以分为四种情形（又见附录）：

1. 两个动词概念发生的时段是一样的（具有上下位关系的两个动词概念显然是时段重合的）。{跛行}蕴涵了{行走}，可以认为{跛行} is a kind of {行走}。
2. 一个动词概念在另一个动词概念发生的时段内。{打鼾,打呼噜}蕴涵了{睡觉}。动词概念间的部分整体关系可以类比于名词概念间的部分整体关系。
3. 两个动词概念没有时段上的包含关系，但由动词概念 C 做反向推理可以得到动词概念 C'，或者说如果没有 C 的发生就没有 C' 的发生。{成功,胜利}蕴涵了{尝试}，即如果没有{尝试}就没有{成功,胜利}。
4. 两个动词概念没有时段上的包含关系，但由动词概念 C 做正向推理可以得到动词概念 C'。例如，{提高}蕴涵{上升}，即{提高}某物导致（或意图是使）这个物{上升}；{给,送,赠,捐}蕴涵{拥有}，即{给,送,赠,捐}某人某物导致（或意图是使）这个人{拥有}这个物。

WordNet 没有区分第二种和第三种情形，统一用 C entails C' 表示。由于考虑到时间在动词概念推理中的重要地位，CCD 动词概念间的关系如上所述。

2.3.3 名词概念和动词概念间的联想关系

定义 联想关系指的是描述同一个事件的名词概念和动词概念之间的关系。[13]
例 名词概念{战争}与动词概念{打仗,战斗}存在联想关系，简记为{战争}↔{打仗,战斗}。其他例子还有{车祸}↔{撞车}，{会议}↔{开会}等等。
性质 联想关系是对称的、传递的。

联想关系不同于共现关系。一般地，如果概念 C 和概念 C' 之间有联想关系，则与 C 有共现关系的概念也可以与 C' 共现。联想关系既可以直接应用于信息提取和信息检索，也可以与共现关系结合起来应用于信息提取与信息检索。

2.4 动词概念的闭合语义约束

在上下位关系所确定的一棵标记树（labeled tree）中，节点与概念之间是一一对应的。[14]一个节点 C 的邻域（即，C 的上位概念、下位概念的全体和 C 的兄弟节点的全体）从上下位的角度反映了 C 在整个概念体系中的地位---我们姑且将之称为上下位知识表示。可替换性原则决定了 SynSet 必然在句法和语义上反映了语言规律，譬如，给定一个动词概念 C，其中动词的配价和变元一定都分别相同。事实上，除了同一词性概念间的上下位知识表示，动词概念与充当其变元的名词概念之间相互约束的关系[15]对于自然语言语义分析也十分地重

[13]心理语言学所定义的联想要比 CCD 中的联想更宽泛。
[14]在语义范畴确定的情况下，有时我们用概念来表示节点或用节点来表示概念（参见[21]和[23]）。
[15] WordNet1.7 版本对名词概念和动词概念之间的关系还没有任何的描述，动词概念的句子框架也很简单。

要：

定义 用概念集合Γ的子集合Γ_N（名词概念集合）刻画或约束动词概念的句子框架中的角色变元称为*动词概念的闭合语义约束*（closed semantic constraint of verb concept）。

定义 CCD 中一个动词概念C的*句子框架*（sentence frame）就是包含C的最长语块序列（the longest chunk sequence）。其中，CCD 的语块共有 26 个角色变元（参见[10]）：

缩写	角色		缩写	角色		缩写	角色	
AGN	agent	施事	PAT	patient	受事	COS	course	经事
ESS	essive	当事	GOL	goal	向事	EXP	experience	感事
RSN	reason	缘由	INT	intention	意图	RST	result	结果
IMP	impellee	致事	TIM	time	时间	SPA	space	空间
CON	content	内容	MAN	manner	方式	BRL	belongings	属事
PAR	part	分事	MAT	material	材料	INS	instrument	工具
CAT	category	类事	SOR	source	源事	COM	comitative	涉事
QNT	quantity	数量	FRQ	frequency	频次	DUR	duration	历时
SCP	scope	范围	GEN	genitive	领事			

例 概念{吃,食用}的语块序列为 AGN+V+PAT，其中角色 AGN 的闭合语义约束为{动物}和{人,人类}，角色 PAT 的闭合语义约束为{食物,可吃物}---这意味着{动物}和{人,人类}的子孙概念都可以具体实现该语块序列中的角色 AGN，同理，{食物,可吃物}的子孙概念都可以具体实现该语块序列中的角色 PAT。CCD 将描述所有动词概念的语块序列及其角色变元的闭合语义约束，有两项后续的研究工作：一是语块序列的分布；二是闭合语义约束对名词概念结构的影响。

2.5 CCD 中的组合性原则

在自然语言的形式语义学中，Frege *组合性原则*[16]是一个非常基础的假设，主要考虑如何从有限多个表达式语义的形式描述演绎出潜在无穷多个复合表达式语义的形式描述。语义词典构建中所能涉及到的组合性原则主要是指从已有的概念诱导出新概念，这里我们把新概念等同于已有概念的某个复合表达式。

定义 给定概念$C_i, i=1,2,\cdots,n$，如果存在函数f使得$C=f(C_1,C_2,\cdots,C_n)$，则称C为$C_i, i=1,2,\cdots,n$按照方式f的*派生概念*（derived concept）。

例 {爷爷,祖父}$=CC$且{外祖父,姥爷}$=CC'$，其中$C=${父亲,爹,爸爸}，$C'=${母亲,娘,妈妈}。按照定义，{爷爷,祖父}和外祖父,姥爷}都是派生概念。

定义 如果名词概念C是一个二元关系，则

1. C是*对称的*（symmetric）当且仅当$xCy \rightarrow yCx$
2. C是*传递的*（transitive）当且仅当$xCy, yCz \rightarrow xCz$

例 {朋友,哥们儿}是对称的，但不是传递的；{长辈,前辈}是传递的，但不是对称的。

性质 如果一个名词概念C是传递的，则派生概念$CC\cdots C=C$。

[16]一个复合表达式的语义是其组成部分的语义的函项。

3 结论

本文着重论述了 CCD 的结构，即 CCD 中的概念和概念间的关系。CCD 与语料库语言学的关系（参见[21]）、CCD 的具体应用和 CCD 的词典演化不是本文的重点话题，并不意味着它们不重要，相反地，CCD 的设计思想直接受这几个问题的影响，作者后续的工作将对它们分别进行详细的论述。另外，CCD 与其他语义词典的比较研究也有待深入。

致谢

在“词典的可视化辅助构建”思想的指导下开发的应用软件 VACOL，大大改进了 CCD 的结构自动检查、语义知识半自动检查并提高了词典构建速度。2001 年 5 月至 10 月，北大计算语言学研究所完成了 17,000 个名词概念的构建。感谢北大计算语言学研究所的 CCD 小组成员刘扬、张化瑞、宋春燕和鲁川教授，感谢 CCD 的所有开发人员，CCD 的一切成果都离不开他们辛勤的工作。同时不能忘记的还有参加第二届汉语词汇语义学研讨会的朋友们，他们对 CCD 的关心和帮助一直激励我们不断深入探索 CCD 构建中的新课题并改进已有的工作。新加坡国立大学的杨宜谨副教授在北大计算语言学研究所访问期间也参与了这个项目，我们也对她出色的工作表示赞赏和感谢。

参考文献

[1] Aristotle. (1941). *Categoriae*, in The Basic Works of Aristotle, R. McKeon (ed). Random House, New York

[2] Beckwith, R. (1993). *Design and Implementation of the WordNet Lexical Database and Searching Software*, in the attached specification of WordNet 1.6.

[3] Carnap, R. (1966). *Der Logische Aufbau Der Welt*, Felix Meiner Verlag, Hamburg

[4] Carpenter, B. (1992). *The Logic of Typed Feature Structures*, Cambridge University Press

[5] Fellbaum, C. (1993). *English Verbs as a Semantic Net*, in the attached specification of WordNet 1.6.

[6] Fellbaum, C. (ed) (1999). *WordNet: An Electronic Lexical Database*, The MIT Press.

[7] Huang, C.R. et al (2001). *Linguistic Tests for Chinese Lexical Semantic Relations: Methodology and Implications*, report in the Second Workshop on Chinese Lexical Semantics.

[8] Lakoff, G. (1987). *Women, Fire, and Dangerous Things: What Categories Reveal about the Mind*, University of Chicago Press.

[9] Leibniz, G.W. (1981). *New Essays on Human Understanding*, P. Remnant and J. Bennett (Ed. and trans.), Cambridge University Press.

[10] 鲁川等, (2000). 《交易类四价动词及汉语谓词配价的分类系统》, 《汉语学习》第 6 期, P7-17

[11] Lucas, W.F. (ed) (1983) *Modules in Applied Mathematics*, Spring-Verlag New York, Inc.

[12] Miller, G.A. et al (1993). *Introduction to WordNet: An On-line Lexical Database*, in the

attached specification of WordNet 1.6.
[13] Miller, G.A. (1993). *Nouns in WordNet: A Lexical Inheritance System*, in the attached specification of WordNet 1.6.
[14] Partee, B.H. et al (1990). *Mathematical Methods in Linguistics*, Kluwer Academic Publishers.
[15] Priss, U. (1999). *The Formalization of WordNet by Methods of Relational Concept Analysis*, in *WordNet: An Electronic Lexical Database*, Fellbaum C. (ed), The MIT Press. 179-196
[16] Quine, W. (1980). *From a Logical Point of View*, Harvard University Press
[17] Rosch, E. (1975). *Cognitive Representations of Semantic Categories*, Journal of Experimental Psychology, 104, 192-233.
[18] Russell, B. (1948). *Human Knowledge --- Its Scope and Limits*, Simon and Schuster
[19] Russell, B. (1989). *Logic and Knowledge*, Unwin Hyman Ltd
[20] Vossen, P. (ed.), (1998). *EuroWordNet: A Multilinugual Database with Lexical Semantic Networks*, Dordrecht: Kluwer.
[21] Yu, J.S. (2001). *Structures in CCD*, report of Second Workshop on Chinese Lexical Semantics, held in Peking University
[22] Yu, J.S. (2001). *Algebraic Structures in Linguistics*, report of ICL-salon of Computational Linguistics, Peking University
[23] Yu, J.S. and Yu, S.W. et al (2001). *Introduction to Chinese Concept Dictionary*, in International Conference on Chinese Computing (ICCC2001), 361-367
[24] Yu, J.S. (2000). *Hidden Markov Model and its Applications in NLP*, report of ICL Seminar of Natural Language Processing, Peking University.
[25] Yu, J.S. (2000). *Machine Segmentation Ambiguities and Dynamic Lexicon*, in Associated Conference AI2000
[26] Yu, J.S. (2001). *Construction of Semantic Lexicon*, draft in the Yu Jiangsheng's homepage http://icl.pku.edu.cn/yujs
[27] Yu, J.S. (2001). *Specification of CCD*, draft of ICL, Peking University, 2000
[28] Yu, S.W. (2000). *The Comprehensive Chinese Language Knowledge Base and its Applications in the Teaching of Chinese Language*, in the 4th Global Chinese Conference on Computers in Education.
[29] Wittgenstein, L. (1953). *Philosophical Investigations*, Basil Blackwell Ltd.

附录

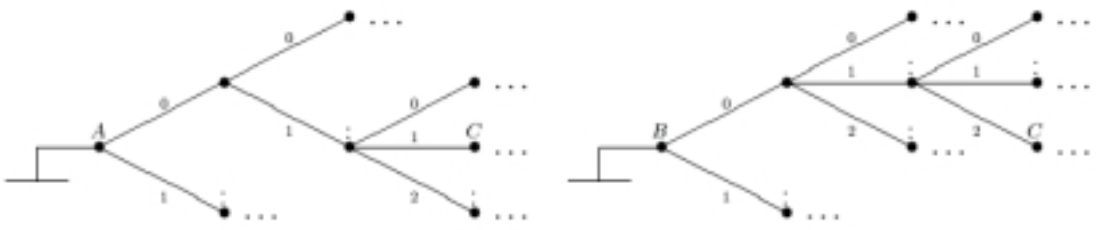

图 1：初始概念分别为 *A* 和 *B* 的概念标记树

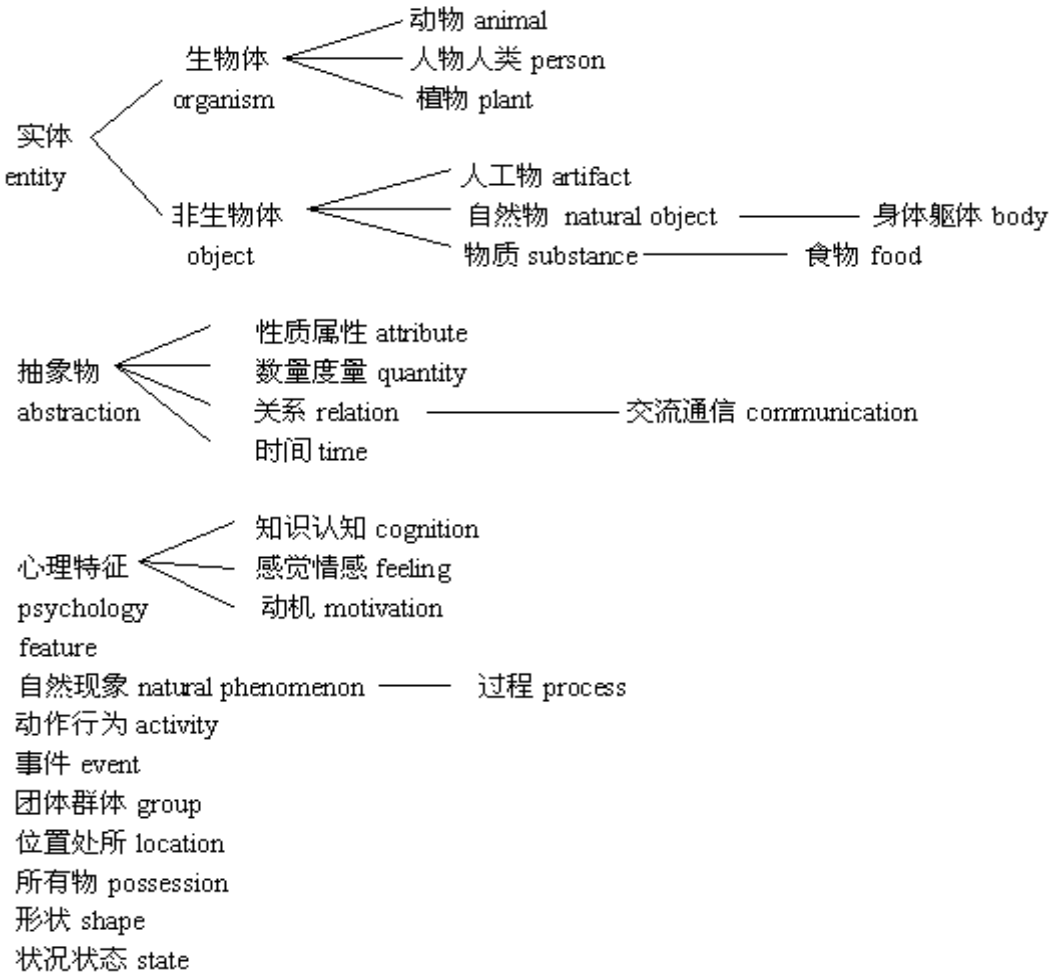

图 2：CCD 名词概念的初始结构

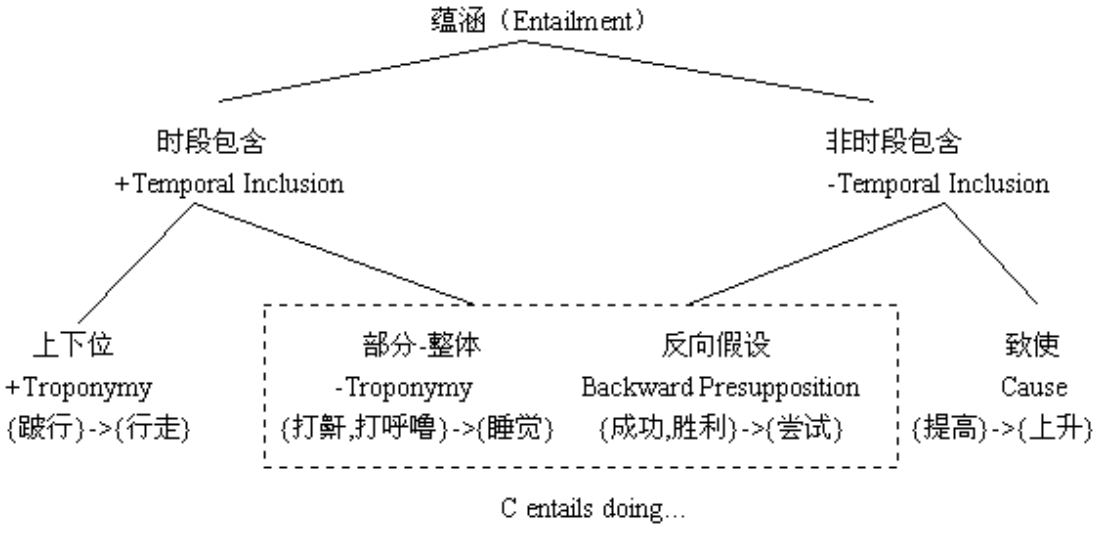

图 3：CCD 动词概念间的四种关系