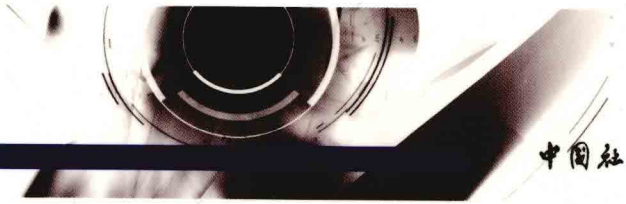
概念变体及其形式化描写

Alloconcepts and Their Formalized Description

胡惮 著



中国社会科学出版社

● “985工程”拓展项目“语言科学技术与当代社会建设跨学科创新平台”（项目编号：985YK006）成果
● 第46批中国博士后基金资助项目“词汇语义知识的形式化表达策略研究”（项目编号：20090460993）成果

概念变体及其形式化描写

Alloconcepts and Their Formalized Description

胡惮 著

中国社会科学出版社

图书在版编目（CIP）数据

概念变体及其形式化描写／胡惮著．—北京：中国社会科学出版社，2011.4

ISBN 978-7-5004-9663-2

Ⅰ.①概… Ⅱ.①胡… Ⅲ.①汉语-语法-研究②英语-语法-研究 Ⅳ.①H14②H314

中国版本图书馆 CIP 数据核字（2011）第 054505 号

出版策划 任 明
特约编辑 李晓丽
责任校对 李 莉
封面设计 弓禾碧
技术编辑 李 建

出版发行 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号 邮 编 100720
电 话 010-84029450（邮购）
网 址 http://www.csspw.cn
经 销 新华书店
印 刷 北京奥隆印刷厂 装 订 广增装订厂
版 次 2011 年 4 月第 1 版 印 次 2011 年 4 月第 1 次印刷
开 本 880×1230 1/32
印 张 8.375 插 页 2
字 数 215 千字
定 价 28.00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社发行部联系调换

序

这里我要说的第一句话是：本书《概念变体及其形式化描写》，是近年来语言信息处理和计算语言学词汇语义资源建设的重要理论成果之一。

语言信息处理，对汉语来讲，涉及字的处理、词的处理、句的处理和篇（语篇）的处理。字的处理一般只涉及形式，词、句、篇的处理则既涉及语言形式又涉及语言意义。

语言的意义是一个古老且复杂的问题，20 世纪以来，语言的意义成为哲学、语言学和逻辑学的中心问题之一。关于意义，哲学家、逻辑学家和语言学家，提出了各种各样的定义。观念论者认为，语言表达式的意义就是和语言表达式相联系的观念；指谓论者认为，语言表达式的意义就是语言表达式所指谓的对象；行为论者认为，语言表达式的意义就是语言表达式所产生的刺激和反应；形式逻辑认为，语句的意义就是语句的真值条件；实证主义逻辑认为，语言表达式的意义就是语言表达式的用法。中国的逻辑学家认为，语言表达式的意义就是语言的使用者应用这个语言形式所表达或传递的思想和感情。①

词、句、句联（句子联结体）、语篇，是不同层级的语言表达

---

① 参看周礼全主编《逻辑——正确思维和有效交际的理论》（国家“七五”哲学社会科学规划重点项目成果），人民出版社 1994 年版，第 15—16 页。

式，其中词和词的意义是语言处理和自然语言理解的第一站。对于词或词汇意义的理解同上面对所有语言意义的理解一样，可以从多种角度观察和展开，但是我们认为，语言是人类最重要的思维工具，思想和感情是思维的结果，词和其他语言表达式都是思维过程和结果让人可以觉察的形式和载体，并且它们分别对应着思想的不同形式和单位，词对应的是思想的最小单位和要素——概念。著名逻辑学家周礼全指出："概念是属于语言使用者的思想方面的，但是概念所描述的或词语所指谓的事物则是思想之外的客观存在"，并且"一个词的意义，就是这个词语所表达的概念"①。在这个意义上，语言信息处理的意义处理，最重要、最有效的入口和理论抉择，是从概念的角度研究词的意义。一些面向机器的词汇语义资源常常冠之"概念词典"、"概念语义网"，等等，就是这类意义定位的实据。

《概念变体及其形式化描写》一书，不只是其研究的理论切入，符合语言信息处理的理性选择，而其最为重要的价值当推"概念变体"概念的提出、论证和研究，对于词汇语义资源建设以及语言研究甚至逻辑研究的理论意义。

20 世纪最突出的词汇语义资源成果，应首推美国普利斯顿大学的词网（WordNet）。该词网在语言、认知和方法论方面具有标志性意义的贡献，应该是 Synset 概念的提出及其应用。所谓 Synset 用汉语讲就是具有同义关系的词语的集合②。这种集合的建立，首先，是使词典中具有相同意义（其实是义项）的词（严格地讲是词的形式——词形），第一次有了一个个团体"户口"，并使之由原始形态的隐形部落，而系统化为具有现代包装的若干显性群体；其次，是使在词典中分散于各个词条之下的义项得到归并。应该

① 周礼全：《逻辑——正确思维和有效交际的理论》，人民出版社 1994 年版，第 17 页。

② 考虑到汉语的表述习惯，我们一般称之"同义词群"。

说，Synset 在简化了词典义项个数的同时，能让我们发现同义词群中的词一个是另一个的概念变体。并且这种发现，在武汉大学催生了概念变体的工程研究、形式化描写和“词群—词位变体理论”的建构。①

概念变体涉及同义词，本书以传统语言学同义词研究成果为基础，结合语言信息处理的特点，研究了同义词的性质和方便操作的确定标准。

概念变体涉及概念之间的关系，本书从范畴化与范畴层次视角研究了概念间的关系。

概念变体涉及不同概念差异的构成及其描述，本书建构了差异系统，论述了词语不同语义关系形成的四种结构：二分对立结构、三分对称结构、多分量差结构和多向异征结构。

词语是语言概念和语法单位，概念毕竟是逻辑概念和思维单位。把概念变体作为词汇语义资源的研究对象，实事求是地讲，是词义研究对象的转变，是词义研究及词语和概念相互关系在认识论上的一种阶段性飞跃。关于研究对象，笔者在《汉语语法研究论——汉语语法研究之研究》（华中师范大学出版社 2001 年版）的前言中曾写道：“学科发展到一定阶段，个人研究进展到一定阶段，研究对象问题是不能不思考的。因为这种思考是学术坐标的自我定位。有位学者曾经这样说过：宇宙是一个时空无限的坐标，

① 关于“词群—词位变体理论”参看萧国政《动词“打”本义的结构描写及其同义词群建构》、《中文计算技术与语言问题研究》，电子工业出版社 2007 年版；Construcing Verb Synsets for Language Reasoning Based on Synst-Allolexeme Theory（一种人机共享的词汇语义理论：“词群—词位变体”论），*Recent Advance of Chinese Computing Technologies*，Coplis Publications，Singapore：2007）。在该理论中，一个概念在语义上是一个“词位”，同一集合中的不同词（音义统一体）就是该词位的语形语义变体。从思维角度讲，概念不易再分，但是在语言层面义位可再分解为若干语义基元的组合。［关于这方面的内容可参看萧国政等《从概念基元空间到语义基元空间的映射》，《华东师范大学学报》（哲社版）2011 年第 1 期。］

人类的一切努力都是在寻找和改变人在这个坐标中的位置。人生活和为之奋斗的社会，是一个特定时空的坐标，人的一切努力都是在重新确立自己在这个坐标中的坐标点。学科是以研究对象和成果性质划分的社会坐标象限，一个学科、一个学者任何有价值的质的努力，都是在移动学科或自己所在的学术坐标点。”

概念变体还引起我们对于思维单位和工具的进一步思考，这是逻辑学家应该也必须探讨的问题。语言是逻辑的载体，语言与逻辑的交合处，或从语言看逻辑，我们会惊喜地发现单从逻辑系统本身不易发现的东西。①

语言信息处理研究是交叉学科，涉及多个领域。胡惮本科毕业于英国语言文学专业，硕士读的是计算机辅助语言教学方向，博士是语言学及应用语言学专业语言信息处理方向，博士后的研究方向是计算机软件与理论。其文理工多学科交叉的知识结构使胡惮在这个领域能得心应手。

胡惮勤于思考，具有多学科的知识结构，工于语言表达，在教学、研究和产品开发方面都有不少成就。读完这本书，读者会跟笔者一样，深切地感受到，他的这些优势、特点和潜力，在本书不同方面均有不同程度的体现。也正因为这一点，该书可以作为有志于从事词汇的语言信息处理研究的文理科师生的入门参考书之一。

胡惮从当访问学者到博士、博士后，我们共同探讨切磋有七个年头，他参加了武汉大学语言与信息研究中心的多项建设，在武汉大学现在他除了博士后身份外，还是语言与信息研究中心主任助理。在武汉大学语言与信息研究中心的网页上，我们师生在

① 20 世纪 80 年代，我们从时间性定语的研究发现时种概念、时量判断和时段推理，为三段论推理的内容和概念划分的认识，作出了我们的贡献。[参看肖国政《试论时种概念》，《华中师范大学学报》（哲社版）1986 年第 4 期，中国人民大学复印资料《逻辑》1986 年第 8 期。]

首页共同写下了我们的宗旨：决战前沿，造福人类。其基本意思是：在我们的行为和信念中，争取我们的每一点努力都是社会应用、学科建设和研究亟需的内容，我们每做一件事，完成的每一项工程，开发的每一个软件，都是为减轻他人劳动，方便所有用众即人类的努力，绝不亮虚招。

在我们看来，自从盘古开天地，三皇五帝到如今，对人类生存和发展影响最重要的技术有三个：种养殖、蒸汽机和电子计算机。与之相应的人类社会也经历了三个重要阶段：农业时代、工业时代和信息时代。第一阶段的种养殖及其代表的农业的诞生，解决了人的生存问题，使人类结束了没有稳定食物的游牧阶段，能生存下来，并繁衍下去。但人类不会只满足于生存下来和繁衍下去的原始需求，还希望劳动轻松，生活美好。劳动，一般分为体力劳动和脑力劳动。第二阶段的蒸汽机不只是开创了工业时代，其最重要的价值是把人从繁重的体力劳动中解放了出来。第三阶段的电子计算机的诞生，其终极使命绝不仅仅是解决复杂的计算，而是要把人类从繁重的脑力劳动中解放出来，并同时帮助人类建立除自然空间和社会空间之外的第三空间——虚拟空间，让人类拥有更加广阔无垠的精神领空及更多空间的延伸。而要实现计算机的这个最终使命，首先让要计算机能像人一样思考和与人对话，其面向语言信息处理和人工智能的语言研究，是通往这条光辉前景不可替代的坦途，因此语言研究就是在营造进入人类天堂的通天塔。为了计算机的最终使命的早日实现，让我们大家和胡惮一起来修造这座通天之塔！

萧国政

2010 年冬于武汉大学珞珈山麓

摘　要

随着语言信息处理研究的不断深入，自然语言理解的关键技术已经逐渐聚焦到语义分析研究。作为语义计算立根之本的大规模语义知识库的研究与建设，是当今计算语言学的重要发展方向之一。迄今为止，国内外现有语义知识库的研究提出了处理概念间宏观语义关系的多种理论模式，但却普遍忽略了概念的语用变体以及与概念变体相关的各种微观语义关系。这种现状成了制约提高语义知识表示颗粒精细度与语义计算准确度的瓶颈。为了解决这个问题，本书以概念变体的形式化描写为基础，提出了一种新型的语义知识库——基于多维特征集的概念语义词网的建构理论及其工程实现的技术方案。

本书的研究共分为六章，每章主要内容分述如下：

第一章：绪论。阐述选题的背景与意义，简介本书的理论依据，对比几种主流的大规模语义资源的网络结构，分析其中存在的共同问题，交代本书的研究对象、目标追求、技术路线与语料来源。

第二章：范畴化视角中的词义聚类研究。范畴层次理论是语义知识库建设的重要理论依据之一，但现有认知语言学的范畴化理论并不足以为我们建构基于多维特征集的概念语义词网提供足够的理论支撑。本书对范畴层次理论进行了拓展，认为人类认知系统中实际上存在三种不同性质的概念次范畴：逻辑次范畴、语

用次范畴和元次范畴，并进一步论证这三种次范畴的性质、特征、交际价值以及它们在概念语义词网建设中的重要地位。

第三章：面向信息处理的汉语同义词群的确定与区别性特征描写。概念具有多个维度的属性特征，在不同的语用场合不同维度的特征会得到凸显，从而产生概念的语用变体。概念语义词网中的每一个概念节点对应着自然语言中的一个同义词集合，集合中的每个元素代表一个概念变体。对同义词的认定是建立概念网络节点的关键，而对同义词区别性特征的描写，是构建概念内部变体的微观语义关系的重要手段。在传统语言学同义词研究成果的基础上，结合语言信息处理特点，约定信息处理的汉语同义词确定标准。论证同义词的语义差异呈四种系统性结构分布：二分对立结构、三分对称结构、多分量差结构和多向异征结构。阐述根据这些结构特征设计概念变体特征属性的形式化描写方法。

接下来分两章具体探讨语用次范畴的形式化描写技术。

第四章：基于实例的语用次范畴描写（上）。以名词为例，探讨概念语用变体特征属性的描写变量。概念的特征属性维度分共享属性维度和次类属性维度，名词性概念的共享属性维度共分 11 种：义域广度属性、关联义位属性、语意强度属性、语义重心属性、表情倾向属性、语体色彩属性、搭配限制属性、地域变体属性、隐性表量属性、历时等义属性、绝对等义属性。详细阐述这 11 种属性各自的语义差异分布结构、属性变量的命名、变量的数据类型、变量的取值范围、属性变量的 XML 描写方法。

第五章：基于实例的语用次范畴描写（下）。以指人名词次类为例，描写概念语用变体在各个维度上的属性特征。除第四章定义的 11 种共享属性外，指人名词还有一种次类属性特征：称谓能力属性。从《汉语水平词汇与汉字等级大纲》和《汉语 8000 词词典》穷尽性地选取 137 个指人名词，构成 55 个同义词集，采用 XML 语言对这 55 个概念的 137 个变体进行 12 个特征属性的全方

位描写。

第六章：结语。概括本书的主要创新观点和结论，总结研究中尚存的不足之处，提出作者进一步的研究计划。

关键词：语义知识库；概念语义词网；概念语用变体；特征属性维度

目　录

第一章

绪　论

第一节　语言信息处理与概念的语义知识表达

在信息化高度发展的当代社会，知识的产出以井喷式的方式和速度增长。尤其是互联网上五花八门的电子信息呈几何级数日益递增，人们获取信息的途径和手段越来越依赖机器。自然语言作为承载与传递人类知识与文化的主要载体，其处理效率已经成为关涉整个人类社会文化与科学发展进程的瓶颈。借助自动化手段处理海量语言信息的社会需求日益凸显，推动语言信息处理技术迅速崛起，成为学界关注的焦点。

机器理解自然语言的根本前提，是我们应该首先告诉机器足够多的语言知识，因此，语言资源建设是自然语言处理技术取得进步的重要基础。

迄今为止，中文信息处理历经了字处理、词处理、句处理的阶段，经过几代学者的努力，取得了辉煌的成就，目前已经全面进入了语义处理阶段，新的一轮学术发展高潮正在逐步掀起。在这种背景下，作为汉语语义计算立根之本的大规模语义知识库的建设，受到了学界的广泛重视，成为整个汉语语言资源建设工程的主要环节。目前，信息处理所需的汉语语义知识库仍然比较缺乏，计算语言学界的一些知名学者们迫切地指出，应当把建立类

似于 WordNet[①] 的各种汉语语义资源列为中文信息处理近期最重要的发展战略之一（孙茂松，2004）。这正是本书立论的根本缘起。

被誉为“现代计算机之父”的冯·诺依曼[②]曾经精辟地指出：人类的语言不是数学的语言。他认为“神经系统是基于两种类型的通讯方式的。一种是不包含算术形式体系的逻辑指令的通讯，另一种是算术形式体系的数字的通讯。前者可以用语言叙述，后者则是数学的叙述”（Von Neumann，1958）。冯·诺依曼十分关注电脑和人脑在解决同一问题时共同的算法，在他设计和建造最早的电子数字计算机时就试图模拟人脑已知的运算过程。

如果说早期的计算机只是为了解决大量的计算工作而设计的，所以计算机对人脑的模拟主要是数字通讯的模拟，那么，我们今天的计算机所要面对的不仅仅是数字的计算，而是整个人类的自然语言，所以计算机还需要模拟人脑的语言通讯方式。对概念系统在人脑中的认知和储存方式的模拟，是语言信息处理的重要研究领域之一。

黄曾阳先生认为，“语言概念空间是存在于人类大脑之中的一个符号体系，这个符号体系既是人类进行语言思维的载体，也是人类进行语言交际的引擎。计算机要获得理解自然语言的能力就必须也拥有一台在功能上类似的引擎”。“语言概念空间是一个四层级——基层、第一阶层、第二阶层和上层——的结构体。这个四层级的概念空间和概念共同构成语言思维的载体，即概念层次网络。语言概念空间的基本特征就是它的层次性和网络性。”（黄曾阳，2004）黄先生的理论称为 HNC（概念层次网络）理论。

黄先生不是唯一持这种观点的学者。在语言信息处理领域，

① 指普林斯顿大学开发的词网，下文将专门论及。

② 我们今天所使用的计算机都建立在一种被称为“冯·诺依曼”范式的理论基础之上。