形容词修饰语语义计算理论及其在对外汉语学习词典编纂中的应用

李学宁 ◎著

世界图书出版公司

江南大学自主科研计划项目资助（编号：JUSRP20912）

形容词修饰语语义计算理论及其在对外汉语学习词典编纂中的应用

李学宁　著

世界图书出版公司
北京·广州·上海·西安

图书在版编目(CIP)数据

形容词修饰语语义计算理论及其在对外汉语学习词典编纂中的应用/李学宁著. —北京：世界图书出版公司北京公司,2012.5
ISBN 978-7-5100-3991-1

Ⅰ.①形… Ⅱ.①李… Ⅲ.①计算机应用—对外汉语教学—词典编纂法 Ⅳ.①H195

中国版本图书馆 CIP 数据核字（2011）第 201989 号

形容词修饰语语义计算理论及其在对外汉语学习词典编纂中的应用

著　　者：李学宁
责任编辑：陈晓辉
封面设计：春天书装图文设计工作室

出　　版：世界图书出版公司北京公司
出 版 人：张跃明
发　　行：世界图书出版公司北京公司
（地址：北京朝内大街 137 号　邮编：100010　电话：64077922）
销　　售：各地新华书店和外文书店
印　　刷：北京博图彩色印刷有限公司

开　　本：787 mm × 1092 mm　1/16
印　　张：10.5
字　　数：150 千
版　　次：2012 年 5 月第 1 版　2012 年 5 月第 1 次印刷

ISBN 978-7-5100-3991-1/H · 1255　　定　　价：26.00 元

序

形容词在汉语中虽然总的数量并不很多，但其使用却很广泛，使用频率也比较高，单个形容词的平均使用频率要远高于汉语中的另外两大类词——名词和动词[①]。因此形容词研究具有重要的意义。在形容词的研究中，语义研究还是一个薄弱环节。而形容词语义计算研究则更是少之又少。

从语义方面看，形容词可以分为两大类，一类是单义形容词，一类是多义形容词[②]。单义形容词往往与某个默认的特征相联系，在行文中一般倾向于省略与形容词相联系的特征名。但这类词毕竟是少数，更多的是多义形容词，它们往往与多个相关但不同的属性相联系。此时人往往能够通过上下文来理解多义形容词的含义，但计算机却不行，留学生也由于缺乏足够的背景知识，往往不能准确把握其含义。因此，如何准确把握多义形容词在上下文中的准确含义，成为计算语义学的一个重要课题。

然而语言的意义经常是只可意会而难以言传的，其主要原因就是语言的多义性。多义性，或者从另一个角度说歧义性，是自然语言和人工语言的最大差别，人工语言的意义是确定的、单一的，而自然语言则经常是模糊的、多义的。

计算语义学的一个重要课题就是如何消解自然语言的多义性，即所谓消歧。而要想消歧，则首先需要有合适的多义性表征理论。本书作者在第三章集中探讨了形容词多义性的表征问题，介绍了四种多义性表征理论，从最早的经典理论开始，到后来的原型理论、关系理论，以及最近的汉语内涵逻辑理论（值得一提的是，作者在

---

① 参见郭曙纶，《汉语语料库的建设及应用》，上海：上海外语教育出版社，2011 年，第 137 页。

② 参见郭曙纶等，“长”与“久”的语义分析及其他，《宁夏大学学报》，2011 年第 3 期。

概述这些表征理论的时候紧紧抓住了它们的一个共同点——特征来展开，因而为作者的理论构建提供了很好的基础，也为读者清晰地展现出了语义理论发展的脉络），从中概括出两种主要的语义分析方法——语义关系分析法和语义特征分析法；接着在第四章中作者则提出了更能刻画汉语形容词语义规律的汉语形容词概念语义模型——“值—特征—实体”的联结模型。这个模型的主要特点在于“值”和“特征”之间不是一一对应的，因而能够表征出形容词的多义性。而这个模型正是作者在后面展开形容词修饰语语义分析的理论基础。

在实践上，作者从《现代汉语规范词典》的形容词释义中概括出一种典型的形容词表征方法——同义词、反义词加特征。作者在此基础上展开了一个具体的实践——把《现代汉语规范词典》转换为学习词典的实验。在这个过程中，作者提出了一些非常实用的方法，成为本书的一个亮点，如在5.3.2.2节讨论“例词与例词组在模型中的区分”时，作者提出了一种非常有效的区分自由与非自由组配的方法。作者是以“大”和“小”为例来解释说明这种方法的。

从《现代汉语规范词典》中，可以获得许多与“大”和“小”组配的名词，如“巴”、“白菜”、“班”、“半”等，根据在与“大”和“小”组配的两个结构中特征的变化，这些形名组合可以分为三类：

1. 两者在名称和数目上完全一致。例如：

大巴：大型客车（巴：巴士）。

小巴：小型公共汽车（巴：巴士）。

大半：比一半大或比半数多的部分。

小半：比一半小或比半数少。

2. 两者在特征的名称和数目上不一致。但是特征之间具有密切的联系，或者在一个没有直接提供特征名的结构中可以接受对方的特征名。例如：

大姑：年龄最大的姑母。

小姑：排行最小的姑姑。

大班：幼儿园里的最高班级。

小班：幼儿园里年岁较小的孩子组成的班级。

3. 在特征的名称上完全不同，之间也没有密切的联系。例如：

大户：旧指有钱有势的人家。

小户：人口少的家庭。

大道：古代指儒家最高的政治理想，即“天下为公”的社会；也指最高的治世原则。

小道：狭窄的路。

第一种情况是典型的自由组配，第三种情况是非自由组配，而第二种情况可以认为是一个中间状态。这样的区分对于学习汉语的外国学生来说，显然是非常有用的。也就是说，对外汉语学习词典编撰时非常需要区分这三种情况。

我认为不同专业背景的人都可以从阅读本书中得到不同的收获。语言学专业的读者从中可以学到如何从普通内向词典释义中提取相关特征，进而更为精确地把握某个或某些形容词的语义，并且更好地理解形名组合的语义；计算机专业的读者从中可以学到如何具体地把普通内向词典自动或半自动地转换为学习词典，并且还可以学到形名组合的语义计算方法；而对于更为专业的读者，如研究形容词修饰语语义的读者，或者是中文信息处理研究者，又或者是词典学家，尤其是计算词典学家，本书则更是一本不可多得的好书，因为本书既有相关语义理论的探讨，也不乏具体实践的探索。

本书所研究的完全是一个全新的领域，这正是本书的价值所在。本书首次对作为修饰语的形容词语义做了跨学科的研究，其应用前景广阔。然而局限也在这里：许多问题，作者并没有展开论述，也并未真正解决，还有待于进一步的研究。如：怎样真正实现形名组合的语义计算，它的具体做法和算法如何？内向词典转换成对外汉语学习词典的具体步骤以及转换效果如何？等等。好在作者已经迈出了坚实的第一步，相信在不久的将来，我们能看到作者更多新的研究成果。

郭曙纶

2012 年 2 月 17 日于上海

前　言

在自然语言理解与计算语言学的研究中，一个难点和热点问题是形容词修饰语，即充当定语修饰名词的形容词。由于语义性质复杂，如何构建一个模型来表征其语义并计算形名语义的组合就吸引了众多语义学家、计算机专家和词典学家的共同关注。

本书从计算词典学的角度，对现代汉语形容词修饰语进行了跨学科的研究：

一、从语义学的角度，提出了一个“值—特征—实体”相联结的概念语义模型。

它具有两个重要的创新点。第一点是表征了形容词的多义性。在以往经典的 AVS 语义模型中，人们关注的是形容词的典型性、语境性和否定性等方面的语义性质。存在的一个主要问题是把形容词都处理为单义词，忽视了自然语言的一个普遍性质是多义性，即一词多义。因此，从理论上说，这些研究缺乏的一个基本前提条件是需要先表征出形容词的多义性。因为只有正确消歧后，才能进行后续的研究工作。

本书的解决方案是把一个值与多个特征相联结。通过这些特征，可以与其他的值（表征了该形容词的同义词、反义词）相联结。通过对《现代汉语规范词典》中 127 个高频形容词词条的释义方法的研究，证实了这种表征方法的有效性。

第二个创新点体现在特征的界定与设置上。如何界定特征？如何设置特征？这两个问题在一定程度上决定了后续的计算机自动抽取及其准确率、召回率的统计工作。在语义学中，经典理论、原型理论和关系理论提出了不同的特征。本书发现在这些特征之间存在着一个连续统的关系，因此具有相对性。这从根本上决定了 How-Net、CCD、《现代汉语语义词典》等各家设置的特征在数量和命名上不尽相同，在某些方面甚至存在较大的差异。

本书的解决方案是：基于《现代汉语规范词典》中形容词的释义方式——同义词、反义词加特征，总结、归纳出来一组全部由词典编纂者提供的特征。这些特征随着词典的长期广泛的使用，能够为普通大众所接受。

二、从计算机科学的角度，研究了如何采用 NLP 技术从机读词典中自动提取形容词词条的概念语义模型。

虽然形容词语义模型具有相同的基本结构，但是每个具体的形容词的值和特征都不尽相同。鉴于手工建构每个汉语形容词的语义模型费时费力，本书随机标注了小部分高频形容词的释义，以获取自动抽取的模板。此后使用其余的形容词作测验，获得了比较理想的准确率和召回率。从实验结果来看，从现有的文本词典中通过模板抽取的方法来自动生成形容词概念语义模型是可行的。

三、从词典学的角度，研究了如何把该语义模型运用于对外汉语学习词典的编纂。

对外汉语学习词典指的是在对外汉语教学中供外国学生使用的外向型词典。新语义模型可以作为新型形容词同义词、反义词学习词典的微观结构，在如下三个方面展现了一定的优越性：

1. 能够简便地区分形容词同义词的异同。两词相同，是由于联结了相同的特征和值；不同之处在于各自联结了其他的特征或值。

2. 能够简便地区分形容词词条下所配置的是例词还是例词组。如果某个实体（表征名词）可以与不同的值（表征形容词）相联结，那么它们是例词组，即具有组合性。否则很可能是例词，不具备组合性。

3. 在原来的文本词典中，由于受到版面的限制，某个形容词词条下收录的同义词、反义词和特征的数量是相当有限的。在本模型中，能够把分散在其他形容词词条下的同义词、反义词及其特征也能全部收集起来，从而便于查询和使用。

上述三个方面的研究存在紧密的联系，其意义和价值是研究如何把一本文本词典自动改编为形容词电子学习词典，从而服务于对外汉语电化教学的需要。

目　录

第一章　形容词修饰语研究现状 …… 1
1.1　形容词修饰语简介 …… 1
1.2　形式语义学研究综述 …… 1
1.2.1　谓词及其演算 …… 2
1.2.2　叠置原理 …… 3
1.2.3　形容词修饰语的挑战 …… 4
1.2.4　两种处理：函子与带参数的谓词 …… 5
1.2.5　使用条件问题及其研究动向 …… 8
1.3　认知语义学研究综述 …… 9
1.3.1　选择性修饰模型 …… 10
1.3.2　联结主义选择性修饰模型 …… 12
1.3.3　意义生成模型 …… 14
1.4　现代汉语形容词研究中的相关问题 …… 16
1.4.1　形容词词类的独立性和完整性 …… 16
1.4.2　形容词词类划分的理据 …… 17
1.5　小结 …… 19
第二章　对外汉语学习词典学视点下的形容词修饰语研究 …… 20
2.1　引言 …… 20
2.2　词典在对外汉语教学中的运用 …… 21
2.3　对外汉语学习词典学研究动态 …… 23
2.3.1　学习词典的定位和特点、原则 …… 23
2.3.2　双语词典编纂的改进 …… 25
2.3.3　计算机技术的采用 …… 27
2.3.4　遗留问题 …… 28
2.4　本书的进一步研究 …… 30
2.4.1　研究对象的转换 …… 30

2.4.2 研究方法的改进 …… 32
2.4.3 研究目的、意义的明确 …… 32
2.4.4 研究的主要内容及其章节安排 …… 34
第三章 形容词多义性的表征 …… 36
3.1 四种多义性表征理论 …… 36
3.1.1 经典理论 …… 36
3.1.2 原型理论 …… 37
3.1.3 关系理论 …… 38
3.1.4 “内涵逻辑”理论 …… 39
3.2 两种主要的语义分析方法 …… 40
3.2.1 语义关系分析法 …… 41
3.2.2 语义特征分析法 …… 42
3.3 汉语形容词的综合表征方法 …… 45
3.3.1 研究范围的确定 …… 45
3.3.2 《现代汉语规范词典》的释义体系 …… 46
3.3.3 汉语形容词的表征方式 …… 48
3.4 本章小结 …… 51
第四章 “值—特征—实体”的联结 …… 53
4.1 汉语形容词概念语义模型 …… 53
4.2 值与多个特征的联结 …… 54
4.2.1 多义性的表征 …… 54
4.2.2 多义形名组合的计算 …… 54
4.3 特征设置的相对性 …… 56
4.4 语义工程中特征的设置 …… 58
4.4.1 WordNet …… 59
4.4.2 CCD …… 62
4.4.3 《现代汉语语义词典》 …… 64
4.4.4 HowNet …… 66
4.5 基于《现代汉语规范词典》的特征库 …… 67
4.5.1 设置思路的转变——从专家范畴到大众范畴 …… 67
4.5.2 《现代汉语规范词典》中提供的特征 …… 68
4.6 特征名的补全 …… 69

4.6.1　特征名的显现 …… 70
4.6.2　特征名的预测 …… 70
4.7　本章小结 …… 73
第五章　“大”字模型在学习词典编纂中的运用 …… 75
5.1　“大”的同义词、反义词和特征 …… 75
5.2　“大”字语义模型简图 …… 79
5.3　在对外汉语学习词典编纂中的运用 …… 83
5.3.1　同义词的处理模式 …… 83
5.3.2　示例的区分 …… 85
5.4　本章小结 …… 88
第六章　形容词概念属性自动提取模型研究 …… 90
6.1　概念属性的提取模型 …… 90
6.2　词典预处理 …… 91
6.2.1　词典释义分词 …… 91
6.2.2　切分标注结果的表示 …… 91
6.2.3　切分标注结果的表示 SEGPOS 系统实现 …… 92
6.3　训练集 …… 94
6.4　模板生成算法 …… 96
6.4.1　词频分析法原理 …… 96
6.4.2　词频分析法性能分析 …… 99
6.4.3　最长公共子串 …… 99
6.4.4　最长公共子串性能分析 …… 101
6.4.5　模式匹配法原理 …… 103
6.5　结果筛选和分析 …… 104
6.5.1　对词频分析法的结果筛选 …… 104
6.5.2　对最大公共子串的结果筛选 …… 107
6.5.3　对模式匹配法的结果筛选 …… 109
6.5.4　结果分析 …… 112
6.6　概念属性的应用 …… 113
6.6.1　同（近）义词的抽取 …… 113
6.6.2　反义词的抽取 …… 114
6.7　本章小结 …… 116

第七章 结论 …… 118
7.1 总结 …… 118
7.2 展望一 …… 119
7.2.1 语义模型的心理学测验 …… 119
7.2.2 语义模型的计算机模拟 …… 120
7.3 展望二 …… 120
附录 …… 122
参考文献 …… 141
后记 …… 149

第一章　形容词修饰语研究现状

1.1　形容词修饰语简介

形容词的一个重要用法是修饰名词。当它修饰名词的时候，称之为形容词修饰语（adjectival modification）。

形容词修饰语展现出了极为复杂的语义性质。其中，一个重要的特点是具有“可变性”，即修饰不同名词的时候，语义会发生一定程度的变化。例如，“大个子”指的是“块头”大，而“大房子”一般指的是“面积”大。

这些性质为自然语言的理解带来了困难。目前，国外主要采用形式语义学、认知语义学的理论方法，研究了它的语义表征、形名语义组合的计算、语义模型的建构等三个核心问题。

表 1-1　形容词修饰语研究概括

理论方法 / 内容	形式语义学	认知语义学
形容词修饰语的表征	谓词	Attribute-Value Structure（AVS 结构）
形名语义组合的计算	叠置原理	
形容词修饰语语义模型的建构	Smith et al（1988）; Franks（1995）; Blutner et al（2004）	

1.2　形式语义学研究综述

从逻辑的角度研究语义，在语言学的领域中形成了一个理论流派——形式语义学，包括真值条件语义学、模型理论语义学、可能世界语义学、情景语义学等等。它的核心计算机制是叠置原理

(Compositionality Principle)，也称组合原理。

在形容词修饰语的研究中，原理的运用所遇到的困难是形容词的语义具有语境敏感性，形名之间逻辑关系复杂。在语义表征方面，两种基本的解决方案是将形容词处理为一个算子或者是一个带参数的谓词。在简化形名关系的同时，一个重要的动态是将逻辑规则和百科知识结合起来。

1.2.1 谓词及其演算

为了精确地研究语义，有必要采用某种形式语言作为表征体系。在形式语义学中，人们一般采用谓词演算体系。首先，它具有精确性和单义性。谓词与所表征的意义之间是一一对应的关系。而在自然语言中是一对多，绝大部分词汇具有多个义项。此外，谓词逻辑比命题逻辑更能刻画自然语言的内部结构。命题逻辑的基本单位是命题，对其内部成分和结构不作进一步的分析。而在谓词逻辑中，原子命题被进一步分解为个体词和谓词。这样一来，就能够深入研究词句的意义而不是只停留在句际逻辑关系上。

在经典的谓词理论中，形容词和所修饰的普通名词都被处理为谓词。以“红”和“苹果”为例，它们分别表示某种颜色和水果的特征、性质。只有具有这些性质的颜色和水果才能够称之为“红”和“苹果”。

谓词表征的意义既有性质的一面，也有指称的一面。传统的形式语义学从语言符号与客观世界的关系这个角度来把握意义，因此认为指称的对象是客观世界中的实体。

随着可能世界语义学的提出，实体不再局限于客观世界。Sebastian (2002) 明确区分了“世界中的实体”和“思维中的实体”(即概念)。这样一来，作为意义表征体系之一的谓词就可以指称概念了。在当代语义学文献中，用谓词指称概念，将概念表征为谓词就变得比较普遍了。

Jaap van der Does 和 Michiel van Lambalgen (1998) 提出了一个模型，以阐述谓词的两个不同指称之间的关系：

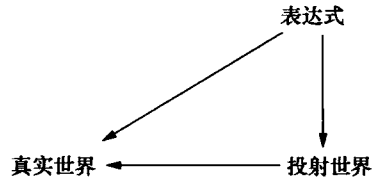

图 1-1　表达式与意义

图中的表达式即谓词，可以指称真实世界和投射世界（即主观世界、心理世界）中的实体。这两个世界之间存在如下映射关系：表达式在投射世界中得到直接解释，在客观世界中得到间接解释。

在建立了两种不同意义类型之间的投射关系之后，有必要指出谓词和 Attribute-value Structure（AVS 体系）是两种并行不悖的表征体系。从逻辑的角度来看，认知语义学中常采用的 AVS 体系所表达的概念、特征、值本身都是谓词。换言之，一个谓词可以分解为一组子谓词。另一方面，逻辑学中有个分支——特征逻辑（Rounds，1997），所研究的就是如何将谓词分解为特征。两种表征体系存在转换的关系。

这样一来，形式语义学中一些相对而言比较完善的理论模型就可以运用于概念，尤其是概念组合的研究。这有利于在一定程度上弥补目前一些认知理论在可计算性、可实现性等方面的不足。在形式语义学和相关的自然语言理解、人工智能科学中，一个重要的数理逻辑基石是集合论：研究的是集合以及集合之间的逻辑关系。从这个角度出发，谓词所指称的其实是一个实体对象集。例如，谓词“苹果”所指称的并不只是一个，还可以是多个苹果。将这个集合的共同性质分解为一组特征后，这些特征也构成了一个特征集合。两个谓词以及所指称的对象集、特征集之间的关系，就可以用“与”、“或”、“差”、“补”等基本的逻辑手段来揭示。这奠定了语义计算的基础。

1.2.2　叠置原理

叠置原理由德国哲学家、逻辑学家 Frege 提出。主要思想是：复合成分的语义取决于成分的语义和构成的方式（参见方立，2008）。

这条原理把握了语义计算过程中所涉及的两大关系。一是复合成分与组成成分之间的整体—部分关系：较大成分的意义由较小成分的意义组成。二是语义规则与句法规则同构对应，可以按照句法规则来计算复合成分的意义。

由于这条原理的逻辑学基础是谓词演算体系，一般采用谓词公式表示如下：

结构上：C = A + B

语义上：|C| = |A| · |B|

其中，|A|、|B|为构成成分的语义。通过“复合”，就可以获得复合语义|C|。

1.2.3 形容词修饰语的挑战

形容词修饰语指的是在名词前充当定语的形容词。它在语义上具有语境依赖性，并且与名词组合的关系复杂。在语义表征和语义操作两个方面给叠置原理的运用提出了挑战。

1. 语义表征

语境性指的是形容词的语义会随着修饰对象的改变而发生变化。这在程度形容词上体现得十分明显。以“大青蛙”、“小象”为例，“大”由于和“青蛙”组合后反而在实际尺寸上变小，“小”和“象”组合后变大。其实，绝对形容词也不绝对。Quine（1960）发现“红苹果”与“红头发”中的“红”并不完全一样，后者只要稍微带点红就可以了。这个观点为后来的真值条件语用学派所继承和发展（Blutner, 2004）。进一步的研究发现形容词的语义还受到其他因素的影响。

这给形容词的语义表征带来了困难。根据 Janssen（1997）的观点，叠置原理要求词在独立的情况下具有意义。只有这样才能够将这个意义处理为一个算符，表征为一个谓词，它的意义应该不受其变目（论元）的影响。

2. 语义操作

在一阶逻辑中，形名之间的语义组合一般处理为谓词的合取。从集合论的角度来看，这适用于形、名两个集合在外延上是交、内涵上是并的情况。然而，形名组合还存在两种其他类型：包含和

否定。

试看下面三组例句（Ivonne Peters and Wim Peters, 2000）：

1）The car is a red Volkswagen.（这辆车是红色大众。）
The car is red.（这辆车是红色的。）
The car is a Volkswagen.（这辆车是大众。）

2）It is a really big spider.（它是一个很大的蜘蛛。）
* It is really big.（*它是很大。）
It is a spider.（它是蜘蛛。）

3）Victor is a former Catholic.（维克多是个前天主教徒。）
* Victor is former.（*维克多是个前。）
* Victor is a Catholic.（*维克多（现在）是个天主教徒。）
Victor was a Catholic.（维克多（过去）是个天主教徒。）

第一组例句中，一辆红大众，它必定是红的，且是大众。“red Volkswagen”是“红”与“大众”两个集合的交集。

第二组中的大蜘蛛本身并不大，只是蜘蛛中的大者。因此，“big spider”所指称的对象集是名词所指称的对象集的子集。即“spider”与“big spider”之间是包含关系。

第三组中的“former Catholic”是否定关系。如果某人从前是个天主教徒，那么他现在就不是了。形名组合所指称的对象集，不属于（包含在）名词所指称的对象集内，而是包含在后者的补集内（谓词否定式所指称的），即非天主教徒。从前和现在的指称不同是由“former”这一表示时态的词引起的。

1.2.4　两种处理：函子与带参数的谓词

上述问题的产生，根源是将形容词处理为一个谓词。谓词所表达的意义是固定不变的，不能够反映形容词语义在语境中的细微动态变化。此外，谓词一般对应一个内涵和外延确定（和语境无关）的集合。而程度形容词和否定形容词在外延或内涵上是不确定的，处理为集合存在一定的周折和问题。这直接影响到了形名之间的语义关系和操作。