华中语学论库

◎邢福义 主编

第二辑

汉语虚词知识库的建设

刘云 著

华中师范大学出版社

华中语学论库(第二辑)

邢福义　主编

国家社会科学基金项目成果
霍英东教育基金会青年教师基金项目成果
华中师范大学出版基金全额资助
华中师范大学语言与语言教育研究中心成果

Hànyǔ Xūcí Zhīshíkù de Jiànshè

汉语虚词知识库的建设

刘　云　著

华中师范大学出版社
2009·武汉

新出图证(鄂)字 10 号

图书在版编目(CIP)数据

汉语虚词知识库的建设/刘云著. —武汉:华中师范大学出版社,2009.5

(华中语学论库(第二辑)/邢福义主编)

ISBN 978-7-5622-3885-0

Ⅰ. 汉… Ⅱ. 刘… Ⅲ. 汉语—虚词—研究 Ⅳ. H146.2

中国版本图书馆 CIP 数据核字(2009)第 002601 号

Hànyǔ Xūcí Zhīshíkù de Jiànshè

汉语虚词知识库的建设

责任编辑:夏兴通 **责任校对**:刘 峥 **封面设计**:罗明波

编辑室:文字编辑室 **电话**:027—67863220

出版发行:华中师范大学出版社

社址:湖北省武汉市珞喻路 152 号

电话:027—67863040(发行部) 027—67861321(邮购)

传真:027—67863291

网址:http://www.ccnupress.com **电子信箱**:hscbs@public.wh.hb.cn

印刷:湖北恒泰印务有限公司 **督印**:章光琼

字数:260 千字

开本:850mm×1168mm 1/32 **印张**:10.5

版次:2009 年 5 月第 1 版 **印次**:2009 年 5 月第 1 次印刷

印数:1—2000 **定价**:22.00 元

欢迎上网查询、购书

序

邢福义

随着历史的发展，社会的进步，科技的发达，语言学在整个世界范围内越来越展示出强大的活力和能量。中国语言学是世界语言学的重要组成部分。为了对中国语言学事业有所推动，我们组织撰写“华中语学论库”。作为专用名称，这里的“语学”主要指汉语语言学，近期的15年时间里以现代汉语语法专题研究为重点。“语学论库”，这是汉语语言学研究的一个系统工程，如果将来主客观条件具备，在研究范围上可以不断扩大，在研究时间上可以无限延展，在研究队伍上可以辈辈交接，代代传承。“华中”一词，既跟研究队伍的华中群体相关，又跟华中师范大学出版社的名称相关。

汉语语言学源远流长。千百年来，特别是《马氏文通》出版以来，尤其是20世纪70年代之后，由于一代代学者的不懈努力，汉语语言学沿着“创业——拓新——发展”的轨道不断推进。目前，汉语语言学所统括的汉语语法学、汉语语音学、汉语方言学、汉语词汇学、汉语语用学等等学科，都已出现了初步繁荣的喜人局面。

但是，初步繁荣并不意味着已经成熟。对于语言学这样一门

社会科学来说，成熟与不成熟的突出标志，应该是学派或流派是否已经形成。在这一点上，科学跟艺术情况相同。比方说，我国的京剧表演艺术已经达到了成熟的高峰，最基本的表现就是形成的这“派”那“派”，只要一提到“梅派”和“程派”，稍有京剧表演艺术知识的人就会知道这是两个具有各自特点的著名流派。又比方说，我国的书法艺术早已达到了成熟的高峰，最基本的表现就是形成了这“体”那“体”，只要一提到“颜体”，稍有书法艺术知识的人就会知道它是不同于“柳体”“欧体”等的有独特风格的书写体，甚至还会知道颜真卿打破了“书贵瘦硬”的传统书风，开创了二王体系之外的新体。然而，汉语语言学的各门学科，即使是其中发展速度最快的现代汉语语法学，仍然缺乏显示成熟的任何标志，距离真正成熟实际上还十分遥远。

当今的汉语语言学，面临的主要问题是“二求”：一求创建理论和方法，二求把事实弄清楚。这是互补互促而又互成因果的两个问题。没有理论和方法的成熟，一门学科不可能是成熟的。而理论和方法的创建，是学者们长期深入研究的成果，是有效地进行群体性思考、独立性思考和开拓性思考的结晶。因此，必然带有鲜明的个性，带有学派的印记，反映一派学者的思想体系、研究特点和总体成就。另一方面，没有对事实的清楚了解，理论和方法的创建便成为空中楼阁。以现代汉语语法研究来说，之所以至今尚未成熟，自成体系的理论和方法之所以尚未创立起来，最根本的原因还是对事实的了解基本上仍然处于朦胧的状态。真正适合于我国语言文字的理论和方法，最终只能产生在我国语言文字的沃土之上。因此，应该强调“研究植根于汉语泥土，理论生发于汉语事实”。不然，我国的汉语语言学在世界语言学中就可能永远处于附庸的地位，就永远不会有跟国外理论对等交流的时候。

学术派别的产生，起码应该具备三个条件：第一，有特定的

学术领地，提示标帜性的理论和主张；第二，有鲜明的治学特点，形成一套自己的研究方法；第三，有良好的学风，形成一支富有活力的队伍。近年来的研究状况表明，我国的学者们已经或多或少地显示了各自的风格特点，但是，顶多只能说其中孕育着某些派别意识，或者顶多只能说预示了某种派别意识的萌芽。汉语语言学的真正成熟，需要经历很长很长的历史阶段，有赖于众多的学者群策群力，更有赖于一辈一辈的学者发扬愚公移山的接力精神。我们华中研究群体人数很少，力量单薄，起点不高，功力不足，对于汉语语言学的发展起不了多大的作用，但是，我们愿意跟在前辈学者的后头，跟在全国各地学者的后头，尽心竭力地做点力所能及的工作。如果把建设富于特色的汉语语言学比作建筑一座大厦，那么，我们组织撰写“华中语学论库”，便是想为这座大厦的建筑献上几根钢筋几块石头。通过参加大厦的建筑，使我们这支小小的队伍受到训练，这是我们的最大愿望。各部著作在内容上具有独立性，但我们希望，在出版了以上二十部之后可以看到研究风格上的某些特色和理论方法上的某种网络。

“华中语学论库”的撰写和出版，得到华中师范大学出版社的大力支持。年初，出版社社长朱峰先生和中文编辑室主任陈昌恒先生到我家，鼓励我牵头编写一套关于汉语语言学的丛书，要我拟订一个初步的计划。不久之后，新上任的总编辑王先霈先生了解了有关情况，立即审定计划，并且从内容到选题都提出了好些中肯的意见。他们为发展学术事业所作的决策，他们在出版事业上的决心、魄力和历史责任感，不管是对我个人还是对华中语言研究群体的所有成员，都是极为有力的鞭策。

千里之行，始于足下！

贵在努力，贵在坚持！

1996 年 5 月 4 日

前　言

20世纪90年代至今，随着席卷全球的“信息高速公路”和信息化浪潮，人类加速迈向信息化社会。信息技术（特别是网络技术）的进步推动了全球化的进程，全球化又反过来对信息技术提出更多新的要求，其中一个主要的要求是希望计算机能够尽可能理解人类的自然语言。我国的中文信息处理研究经过“字”处理期和“语”处理期，在理论探索、基础知识库和语料库的建设、词语的统计与分析以及相关的实验和应用系统的开发等方面都取得了较大的成绩，但相对快速膨胀的实际需求而言，中文信息处理技术依然滞后于实际需求，一个主要的原因是我们为计算机储备的语言知识不够。语言知识库（如机器词典、句法规则库等）是自然语言处理系统的重要组成部分，其规模与质量是自然语言处理系统成败的关键。对于汉语来说，由于缺乏形态变化，自动分析相对困难，尤其需要重视语言知识库的建设。目前中文信息处理领域的知识库，主要是实词的语法词典、实词的语义词典、句法规则库和各种语料库，遗憾的是，国内还没有建立系统的汉语虚词知识库。汉语虚词知识库的建设是目前中文信息处理领域里的薄弱环节，要想把中文信息处理推向一个新的高度，虚词知识库的建设是一个无法回避的基础性工程。

本书共分五章。

第一章“面向信息处理的语言研究”主要介绍了中文信息处理的发展与成就、汉语自动分析的难点、计算机背景下的汉语语

法研究以及“句管控”与汉语信息处理等内容。我国的中文信息处理经历了“字”处理和“语”处理两个时期，现在集中力量对“句”处理进行攻坚。信息时代对语言学研究提供了新的机遇，同时也提出了新的挑战，由于研究对象、研究目的、研究手段和研究视角的转换，注定了汉语语法研究要面向计算机，为信息处理服务。“句管控”理论对我们的启示是能否改进或改变以往的研究范式，把从小到大的策略与从大到小的策略结合起来。这一部分的内容反映了作者对当前中文信息处理现状的思考，意在既要“埋头拉车”，又要“抬头看路”。

第二章“虚词知识库的建设”主要介绍了虚词与虚词研究概貌、汉语虚词知识库的重要性、虚词知识库建设的难点以及虚词知识库建设的构想等内容。虚词知识在中文信息处理的词法分析、句法分析和具体应用等方面都有重要作用。但由于汉语虚词的个性很强，运用范围很广，使用频率较高，有的还一词多类兼多义，而且汉语虚词使用很灵活且缺省现象比较严重，因此汉语的虚词特别是信息处理用虚词词典的研究有很大难度。本书提出了一个三位一体的虚词知识库建设方案。首先，按照虚词的“用法”填写虚词机器词典；再在虚词词典的基础上标注语料库，同时利用语料库也可以检验虚词词典的填写；最后在虚词词典和标注语料库的基础上提炼出虚词规则库。

第三章“虚词词典的内容”主要介绍了副词机器词典、连词机器词典和介词机器词典等内容。对这几类虚词都给出了字段的设立情况、词表和机器词典的样例。

第四章“复句层次和关系的自动分析”主要介绍了复句自动分析的目标和意义、难点和对策、二重复句的自动分析、复句关系词语的离析度以及关系词语驱动的复句关系和层次自动判定。复句自动分析的目标是用树形图的方式把复句的关系和层次表现出来，复句自动分析对复句的理解与生成、单句句法分析、篇章

分析和对语言学成果的检验都有重要的意义。复句分析的难度在于关系词语自身的复杂性、单复句的纠结、复句与篇章的纠结、关系词语的省略、关系词语的复用、关系词语的嵌套、关系词语位置的灵活性、关系词语扩展的自由性和复句关系的复杂性。针对这些难点，提出了一个总的策略是：关系词语驱动；加强预处理；规则与统计结合。从省略能力、扩展能力、嵌套能力、停顿能力、连接能力、对应能力、位置因素和换位能力八个方面具体探讨了关系词语的离析度，并从单双音节、合用位置、关系类型、词性和个体差异的角度探讨了复句关系词语离析度不同的原因。最后，探讨怎样利用关系词语判定复句关系和层次，指出利用复句关系词语自动分析复句最大的困难在于关系词语的灵活多变，主要包括四种情况：复用、单用、虚用、不用，并具体考察了这四种情况。在此基础上，设计了一个汉语复句层次和关系自动分析的流程图，这个流程图的核心是位于中间的复句关系词语知识库。

第五章是全书的结语，主要是对已有研究工作的总结和进一步的研究计划。

刘　云

2008 年 5 月 18 日

目　录

第一章　面向信息处理的语言研究

第一节　中文信息处理的发展与成就

一、引　言

迄今为止人类社会经历了五次信息革命，第一次信息革命的标志是语言的产生，第二次信息革命的标志是文字的使用，第三次信息革命的标志是造纸术和印刷术的发明，第四次信息革命的标志是电信和无线电技术的发明，第五次信息革命的标志是计算机、互联网与现代通信技术的普及应用。每一次信息革命都扩大了信息和知识的传播，每一次信息革命都极大地推进了社会的文明和进步。20世纪90年代至今，随着席卷全球的“信息高速公路”和信息化浪潮，人类加速迈向信息化社会。① 这个时代的显

① 中国互联网络信息中心2008年1月发布的《中国互联网络发展状况统计报告》显示，截至2007年12月，中国的网民数已达到2.1亿人，IP地址数达到1.35亿个，域名总数是1 193万个，网站数量已有150万个，网页总数已经有84.7亿个。

著特点是，计算机在人类生活的各个方面起着越来越大的作用，人类的生产、生活、学习、工作和思维方式也随之发生了深刻变化。信息技术（特别是网络技术）的进步推动了全球化的进程，全球化又反过来对信息技术提出更多新的要求，其中一个主要的要求是希望计算机能够尽量理解人类的自然语言。如果计算机能够理解自然语言，那么人类社会就有可能突破语言壁垒，整个世界成为真正意义上的地球村；如果计算机能够理解自然语言，哪怕只是某种程度上的理解，那么就有可能实现人与机器之间更高水平的智能合作，就有可能使生产力达到前所未有的高度，从而导致人类社会的新飞跃。

物质、能源与信息是人类社会生存和发展的三大要素。信息是客观物质世界存在的形式、状态及各种关系，具有资源性、共享性、可传递性、增生性、可压缩性、符号性和工具性的特点。① 为了能适应信息社会发展的需要，信息社会的新型人才必须具有较强的信息获取、信息分析和信息加工的能力。自然语言是人类最重要的交际工具，是人类处理信息和交换信息的重要媒体，我们日常工作环境中的信息有 80%以上是以语言文字作为媒介来传播交换和记载的；就计算机的应用而言，据统计用于数学计算的仅占 10%，用于过程控制的不到 5%，其余 85%左右都是用于语言文字等的信息处理。② 面对庞大而且急剧膨胀的海量信息，如何高效地组织处理和管理这些信息，并快速、准确、全面地从中获得用户所需要的信息，是当前信息科学与技术领域面临的一大挑战。

---

① 关于信息及其属性，可参见盛玉麒编著《语言文字信息处理》，山东大学出版社，2006 年。

② 参见宗成庆、夏威《计算机能理解自然语言吗——关于人工智能问题的哲学思考》，《山东工业大学学报》（社会科学版）1997 年第 2 期。

随着计算机的推广应用，由数据处理、信息处理发展到知识处理，相应地，对语言文字处理要求的深度和广度越来越高，这一挑战大大地促进了自然语言信息处理研究工作的快速发展。自然语言信息处理研究的核心问题是语言的自动理解和自动生成，自动理解指从句子表层的词语符号串识别句子的句法结构，判断成分之间的语义关系，最终弄清句子表达的意思；自动生成指从要表达的意思出发选择词语，根据词语间的语义关系构造各个成分之间的语义结构和句法结构，最终造出符合语法和逻辑的句子。机器翻译是计算机最早应用的非数值领域之一。计算机问世不久，就开始了机器翻译实验。但无论同计算机技术本身的发展速度相比较，还是同计算机在其他领域应用技术的发展速度相比较，语言信息处理的发展是相当缓慢的，道路也比较曲折。20世纪50年代后期及60年代前期，美国出现过机器翻译研究的第一次热潮。1966年美国科学院语言自动处理咨询委员会发表了《语言与机器》报告（简称ALPAC报告），该报告宣称："在目前给机器翻译以大力支持还没有多少理由。"报告还指出，机器翻译研究遇到了难以克服的"语义障碍"（semantic barrier），认为全自动机译在较长时期内不会取得成功。ALPAC报告给机器翻译泼了一瓢冷水，语言信息处理又有过一段沉寂期。自20世纪70年代后期以来，由于计算机技术的飞速进步和语言学理论的发展，也由于一些机器翻译系统进入实用阶段①，更由于社会需求的推动，语言信息处理研究重新进入繁荣期，其显著标志是已有相当多的语言信息处理产品进入市场。近年来，互联网迅速发展，大量的信息如潮水般涌来，这些信息的主要载体仍然是自

① 例如1976年，加拿大蒙特利尔大学与联邦政府翻译局开发出英法翻译系统TAUM-METEO，该机译系统不经编辑处理就可以发布天气预报信息的译文。

然语言，人们渴望发展自然语言处理技术以实现文本自动分类、自动文摘、自动勘校、信息检索、信息提取、计算机辅助教学、自然语言理解、语音自动识别与合成、机器翻译等，加速信息、知识与文化的交流，促进社会、经济、科学的进步，显然这是每一个国家都面临的挑战。从这个意义上说，信息社会中语言文字信息处理的技术水平和每年所处理的信息总量是衡量一个国家现代化技术水平的重要标志之一。

我国的自然语言处理主要是针对中文进行的，亦即“中文信息处理”，其中“中文”指中国通用的所有语言种类，包括汉语及其他少数民族的语言，但一般指汉语。所谓“中文信息处理”，指的是用计算机对汉语（包括口语和书面语）进行转换、传输、存贮、分析等加工的科学，大致可以划分为两个层次：一个是文字层次，即汉字信息处理；另一个是语言层次，即汉语信息处理问题。它是一门与语言学、计算机科学、心理学、数学、认知科学等多种学科相联系的边缘交叉性学科，是自然语言信息处理的一个分支，需要以大量的语言知识、背景知识为依据，对中文信息的人脑处理过程进行模拟。同样是作为人类社会交际和思维认知工具的自然语言，汉语与其他语言有一定的共同性，因此，汉语信息处理也与其他自然语言的信息处理有一定的共性。不过，跟西方语言的信息处理相比，中文信息处理在许多方面有自己的特点，例如：（一）汉字的特殊性。西方语言的书写符号一般只有几十个字母，属于小字符集。汉字数量多，现代常用汉字有几千个，处理古籍要用到几万个，属于大字符集；而且，汉字字形复杂，因此，汉字输入、汉字识别等处理特别困难。① （二）书

① 正是由于汉字的特殊性带来处理汉语的计算机与处理西方语言的计算机一系列的差异，例如键盘输入、汉字打印与显示、内部代码、汉字识别、程序语言的数据类型、数据库的检索和排序等。