CHINESE MODERN RESEARCH WORKS

Culture

| 中国当代研学丛书 |

文化

基于语料库的现代汉语指人名词研究

韩蕾 | 著

中央编译出版社
Central Compilation & Translation Press

| 中国当代研学丛书 |

文化

基于语料库的现代汉语指人名词研究

韩蕾 | 著

中央编译出版社
Central Compilation & Translation Press

图书在版编目（CIP）数据

基于语料库的现代汉语指人名词研究 /韩蕾著. —
北京：中央编译出版社，2020.3
ISBN 978-7-5117-3789-2

Ⅰ. ①基…
Ⅱ. ①韩…
Ⅲ. ①现代汉语—名词—研究
Ⅳ. ① H146.2

中国版本图书馆 CIP 数据核字（2019）第 285749 号

基于语料库的现代汉语指人名词研究

出 版 人：葛海彦
责任编辑：杜永明
执行编辑：纪宛伯
责任印制：刘　慧
出版发行：中央编译出版社
地　　址：北京西城区车公庄大街乙 5 号鸿儒大厦 B 座（100044）
电　　话：（010）52612345（总编室）　（010）52612339（编辑室）
（010）52612316（发行部）　（010）52612346（馆配部）
传　　真：（010）66515838
经　　销：全国新华书店
印　　刷：三河市华东印刷有限公司
开　　本：710 毫米 ×1000 毫米　1/16
字　　数：261 千字
印　　张：16
版　　次：2020 年 3 月第 1 版
印　　次：2020 年 3 月第 1 次印刷
定　　价：95.00 元

网　　址：www.cctphome.com　**邮　　箱**：cctp@cctphome.com
新浪微博：@ 中央编译出版社　**微　　信**：中央编译出版社（ID：cctphome）
淘宝店铺：中央编译出版社直销店（http://shop108367160.taobao.com）（010）55626985

本社常年法律顾问：北京市吴栾赵阎律师事务所律师　闫军　梁勤

作者简介

韩蕾，女，1973年生，籍贯江苏。汉语言文学博士，现为华东师范大学中文系教授、硕士研究生导师。2009年至2011年任韩国高丽大学客座教授。

研究兴趣为现代汉语中的汉字、词汇、语法、语义、语用与修辞。曾在国内外专业期刊、语文刊物上发表学术论文、札记40余篇，已出版专著一部、编著一部。承担并完成国家哲学社会科学基金项目、国家汉办、上海市教委、上海市语委等课题多项。

前　言

本书运用结构主义语言学、语料库语言学、认知语言学、功能语言学等多种理论，对现代汉语指人名词进行研究。

第一章介绍中文信息处理的基本背景，阐明现代汉语语料库建设与汉语语法研究相互依存的关系。

第二章指出面向中文信息处理的“名·名”词语串研究还比较薄弱，本书的主旨是帮助计算机正确完成跟指人名词构成的双名词语串有关的句处理任务。此外，还阐明了研究依据、性质定位和基本思路。

第三章强调语料库加工过程中应对名词进行多层级分类标注，并对提前分类标注的可行性、必要性等问题进行论证。

第四章在以往研究的基础上，阐明指人名词的判定标准及分类的必要性，提出框架测试法，论证该方法的理据并进行分类实践，归纳出现代汉语指人名词分类系统。

第五章探讨双项指人名词组合的内部制约。先是确立同位短语独立的语法地位，接着描写指人同位组构模式、模式间的连续统及组构规律，然后探讨同位短语的认知语义基础及作为独立类的理据。

第六章从大规模语料中提取对短语定界有用的外部规则，主要考察词语串邻近词语的属性、一定的句式对双名组构的外部限制。并论证边界确认规则与否认规则、外部规则与内部规则之间的关系。

第七章总结本书研究特点及其理论价值和实用价值，并探讨了由计算机处理汉语引发的语言工程的系统性、两可现象与语法规则、语法/非语法形式与静态/动态等问题。

韩蕾

序　言

（一）

应该公正地说，是计算机科学家的探索，开辟出了自然语言信息处理这一具有重大突破意义的领域，开创了计算语言学这一重要的分支学科。也正是因为计算语言学的探索进入了一个关键性的阶段，急切地寻求语言学家的协助和合作，这一合乎科学发展的动向得到了语言学界的迅速反响，计算语言学成为计算机科学与语言学研究相结合的一门交叉科学，成为应用语言学研究中的一门显学。

从20世纪八九十年代起，中国国家语委就开始了汉语语料库的建设，开始了一系列计算语言学的理论建设和人才培养工作。这部著作的作者就是在这样的大背景下成长起来的一位应用语言学的青年才俊。她参加了国家语委、语言文字应用研究所许嘉璐教授、傅永和教授领导的国家社科基金“九五”重大项目《信息处理用现代汉语词汇研究》的研究工作，参加了北京大学计算语言学研究所俞士汶教授领导的《人民日报标注语料库》的研制工作，参加了上海师范大学《当代汉语语料库（一期）》的研制工作，完成了她的博士论文《信息处理人名词研究》。

正式工作后，她又在博士论文的基础上，继续深入研究，对有关现代汉语指人名词的方方面面的属性描写和分析规则进行了一系列的研究。这部著作就是这一系列研究工作的集大成性质的成果。

（二）

语言信息处理的词切分、词性标注、短语结构标注和句子分析都会遇到语言单位的分类和句法语义属性的描写问题，需要探讨语言单位的分类机制和原则，语言结构的组配规律和应用限制问题，这些都涉及句法、语义、语用三个平面的属性描写与规则探讨。目前的研究主要是在句法语义属性的描写上，语用层面的研究比较少见。

汉语语言学在句法层面上的研究受到了汉语信息处理需求的严重挑战。传统的词分类研究的分析深度和句法语义属性描写的细度远远达不到汉语信息处理的精度要求。这就要求开展面向信息处理的理论、方法、机制、规则的研究，开展细致的句法语义语用属性分析描写的研究。这部著作就是这类面向汉语信息处理的现代汉语基础研究，旨在对现代汉语指人名词的再分类规则作深度探讨，对现代汉语指人名词的句法语义属性作细致的描写。更具有创新意义的是在进行指人名词的分类属性描写时，纳入了“指称性”的标准，探讨了句法组构中的指称性的原理和识别方法，深入到了句法——语用属性描写的层面。

（三）

跟传统的汉语本体研究相比，面向信息处理的汉语基础研究具有以下几方面的特征：

可验证性是面向信息处理的汉语基础研究的第一个特点。现代语言学认为操本族语者的语言直觉（intuition）是判断一种语言形式是否合乎语法，是否可以接受的根本标准。但是这种语言直觉是整个言语社团的社会语言心理，它是一种客观存在，但无法验证，所谓“只可意会，不可言传”。传统的语言研究实际上是依赖于语言学家的直觉（linguistic intuition），即“内省法”。内省法提高了语言分析的精确性和精致性，但仍没有解决可验证性这个根本问题，因此对某一个语言形式的合法性、可接受性，不同的语言学家往往有不同的判断和解

释，这是一种常见的现象。面向信息处理的汉语基础研究，强调从实际语料出发，运用归纳法抽绎出规则规律来，这就有了词切分、词性标注（辨歧）、短语结构（结构层次、结构功能、结构关系）分析等一系列的工作程序，运用的是“概率统计法”。概率统计法大大加强了可验证性。但是，即使是大规模的语料统计，也免不了遇到言语中的种种变异：临时性的出错、言语应用领域（语域）差异造成的统计偏差，等等。从理论上讲，全域性的或者超大规模的语料统计应该可以消解这些偏差，但实际上无法做到真正的“全域”和“超大规模”。这就涉及可操作性了。

可操作性是面向信息处理的汉语基础研究的第二个重要特点。这部著作运用概率统计法和内省法相结合的研究方法，用各自研究的结果互校，作对比分析，这是一种切实可行且有实效的好方法。实际上最强调“从言语实际出发”的语言学家在语言研究中也不排斥“内省式”语言学分析的“干预”。话语语言学家是最强调第一手口语语料的真实性和完整性的了，通过隐蔽性的现场录音、电话窃听、非暗示性的调查对话等方法，收集到大量语料，然后进行分析。然而在统计分析以前，还必须对录音记录稿进行整理，把“嗯、啊、这个这个”之类的无意义的“停顿性垫话”、反复啰唆的赘余成分，说了半截子话再重新起头说产生的“非结构性成分”等都删除。这种“整理”，实际上就是语言学家的“内省式”干预。此外，本书中采用的“多层级分类标注法”“分（小）类的属性描写法”等，也都是加强面向信息处理的汉语基础研究的可操作性的方法。

严密的系统性是面向信息处理的汉语基础研究的第三个重要特点。人们在听说读写的言语实践中得到自身语言知识库的全力支持，得到在长年的语言交际中积累起来的言语交际知识和语境知识的大力支持，这种知识库的建设和应用是随机的、开放式的。而计算机的运行则是一个相对封闭的系统，面向信息处理的汉语基础研究必须更强调严密的系统性。这部著作运用组合分析的原理进行分类的聚合分析，提出分类的句法语义属性描写的“框架测试法”，对关系类指人名词分析时运用了名词配价分析的原理；在对指人名词同位结构进行“内部限制”和“外部限制”的分析时，注意到了句法语义属性的结合处理，注意到了上下文因素、说话语境因素，这些都体现了汉语信息处理工程的系统性，是一种很有启发的探索。

（四）

诸多方家都说面向信息处理的汉语基础研究“很重要”，但是也“很难”。

既然“很重要”，就要有人来做；因为“很难”，更需要不畏难的志士仁人来做。

现在这方面的探索已经不再是“筚路蓝缕”的阶段了，“深入”“细化”的难度更大。

作者的辛勤劳动，只是面向信息处理的汉语基础研究的一个小课题，只是汉语信息处理工程的一块砖石。但是，宏伟大厦就是在这样的一块块砖石的基础上建筑起来的。

成果有价值，人才正年轻，前途灿烂辉煌。

范开泰

出 版 人　葛海彦
责任编辑　杜永明
执行编辑　纪宛伯
封面设计　中联华文

Contents

目 录

第一章

引言：语料库与语言学研究

第一节　自然语言处理与语料库

自然语言处理（Natural Language Processing，简称 NLP）就是以电子计算机为工具对自然语言信息进行各种类型处理和加工的技术。

1946 年第一台电子计算机诞生后不久，人们就想用计算机来研究和处理自然语言。从 20 世纪 50 年代初期到 60 年代中期，机器翻译一直是研究的中心。当时采用的主要是“词对词”的翻译方式，译文效果很差。机器翻译的困境使人们意识到：要让计算机真正具备类似于人那样的处理自然语言的能力，就必须对语言自身的规律进行深入挖掘。因此，自 20 世纪 60 年代中期以后，人们便开始重视研究自然语言的语法、语义和语用等基本问题，并尝试实现计算机的自然语言理解，即人机对话，也就是人用自然语言向计算机提出问题，相应的，计算机也能够理解并用自然语言做出回答。当前，除了机器翻译和自然语言理解之外，自然语言处理的内容还涉及情报自动检索、语音自动识别与合成、文字自动识别、词典自动编纂、自动文摘、计算机辅助教学等众多领域。（冯志伟，1996）

随着自然语言处理深度和广度的增加，语料库（corpus）的作用日益明显。语料库，顾名思义，就是存放语言材料的仓库，但严格意义上的语料库主要指熟语料库，即“由大量搜集的书面语或口语构成，经过计算机储存和处理，用于语言学研究的文本库”（Renouf，1987）。

语言学史上第一个大型电脑语料库是“英语用法调查”（Survey of English Usage，简称 SEU），该库由伦敦大学语言学教授伦道夫·夸克（R. Quirk）于 1959 年建立，共收集 200 个语篇，内容涉及各种不同的语体。几乎与此同时，美国英语语料库也在美国布朗大学诞生，1961 年，以弗朗西斯（N. Francis）和库塞拉（H. Kucera）为首的一批语言学家和计算机专家联合攻关，建成世界上最早的机读语料库——BROWN 语库，语篇取自 20 世纪 60 年代有代表性的美国英语出版物，选材严格按照随机原则，语域也非常全面均衡，迄今仍被视为标准语料库。一般认为，这两个库可视为现代语料库语言学开端的标志。（王伯浩，1998）但就语料库语言学自身短暂的几十年发展历程而言，20 世纪中期正是它的低谷期。

20 世纪 80 年代以来，英语语料库语言学（corpus linguistics）复兴，相继出现 COBUILD、英国国家语料库（British National Corpus，简称 BNC）等容量达上亿词的大型语料库，到了 90 年代末，世界上主要语种基本上都开发了各自的语料库。语料库发展迎来一个前所未有的高潮。“计算机语料库研究者们突然发现处在一个不断扩大的世界”，“这种发展应使那些语料库的先驱者们感到欣慰。他们就像是从一辆驴车突然坐到了游行队伍中的一辆花车上”（Leech，1991）。

这中间的一个重要原因就是，计算机科学的飞速发展与计算机技术的迅速普及和应用。（丁信善，1998）语料库与自然语言处理的关系十分密切：从大规模、高质量语料库中提取出的细粒度语言规则，是制作出高精度自然语言处理软件的基础；高精度的语言处理软件反过来又可以提高语料处理水平，保证语料库的质量。

第二节　语料库发展与语言学理论

自 1957 年乔姆斯基（Chomsky）发表《句法结构》一书后，以转换生成语言学为代表的形式主义就逐渐占据语言学界的主导地位。跟早期结构主义者不同，乔姆斯基反对布龙菲尔德（Bloomfield）学派信奉的经验主义哲学和行为主义心理学基础，其观点十分符合 17 世纪笛卡尔的理性主义哲学。他认为，人脑不是一张白纸，不是经验主义学派所说的被动接收器，在那儿等着外部印象和

数据印到上边。人脑天生具有一种非常丰富而且颇为细密的程序，用于接受、理解、贮存和使用来自感官的随意信息。人类之所以能学会语言，就在于大脑中先天赋予的语言习得机制（language acquisition device），具体地说就是普遍语法。因此，乔姆斯基区分语言能力（competence）和语言运用（performance），并认为语言学的中心任务就是前者。在材料的来源上，他主张从“内部”观察、获取本族语使用者的感觉和反应。这从乔姆斯基论著中的例句也能看出，他从不注明例句的出处，只要凭内省合乎规则的就是合格的句子，即使现在没有人说，将来也还可能有人说，这就使得有些例句显得十分古怪。这种对待例句的态度，充分反映了他面向理论（theory - oriented）而不是面向材料（data - oriented）的语言学立场。

正是由于转换生成理论把语料视作经验主义产物进行了全盘否定，并不遗余力地鼓吹研究者个人直觉在语言研究中的重要作用，因此，20 世纪中期第一代语料库工作者的努力被当时的主流看成不合时宜的徒劳，整个语言学界在随后 20 多年的时间里差不多唯直觉是从、唯思辨独尊。语料库建设虽未绝迹，却只能小规模、不成气候地进行，基于语料库的研究方法也大受打击、名誉扫地。

经过对转换生成语法的跟从、应用和反思，人们逐渐发现形式主义唯理方法的最大不足在于其不可验证性（丁信善，1998），拿自造的例句、想当然的推论以及未经验证的假设进行语言学研究无异于“拿一束塑料花去研究植物学”（Sinclair，1991）。

20 世纪末兴起的以韩礼德（Halliday）为代表的功能主义语法，则把语言看成是一种社会行为，而不是独立的客观存在。认为只有从语言功能，即语言使用的角度，才能对语言做出最终的解释。他继承了弗斯（Firth）的实证主义传统，其基石是对可观察的对象进行研究。显然，作为人们外部行为的语言运用是可观察的、可靠的依据；人们内在的语言能力则不可直接观察，只能通过语用实例进行推断。因此，功能主义特别看重语料库中语言的真实使用情况。

近年来，功能主义在跟形式主义的对峙中逐渐占据上风，语料库语言学也毫无疑问地由当初的边缘地位上升为语言研究的主流。因此，有学者认为，语料库正是语言学中形式主义与功能主义两大理论阵营“对垒天平上的一个举足轻重的砝码”（顾曰国，1998）。

第三节　汉语语料库与中文信息处理

一、汉语语料库加工思路

我国历来有注重语料的传统，20 世纪 20 年代就有学者手工建设语料库。比较有代表性的是著名教育家陈鹤琴，为了编选千字课本，他与助手用了两年多时间，建立了包含语文课本、通俗报刊、儿童用书、妇女杂志、小学生课外作品、古今小说等六种合计 55 万余字的语料库，并在此基础上进行字频统计，最终选定 4261 个单字，编成《语体文应用字汇》，于 1928 年 6 月由商务印书馆出版，这是第一本现代汉字字频统计的著作，为汉字的计量研究做出了宝贵的贡献。

在西方语料库语言学的影响下，20 世纪 80 年代以来，零星的机器可读汉语语料库的建设也开始起步。1991 年，国家语言文字工作委员会开始建立国家级的大型汉语语料库，以推进汉语的词法、句法、语义和语用的研究，同时也为中文信息处理的研究提供语言资源，计划规模为 7000 万汉字，当时宣称，这将成为世界上最大的汉语语料库。近年来，各高校、科研院所也纷纷开始了汉语语料库的建设工作（冯志伟，2002）。

大规模、高质量的汉语语料库建设是信息工程的重要基础工程，可以给中文信息处理研究提供更为有效的支持，在语料库开发的几个环节（即规划、设计、选材、建库和标注）中（刘连元，1996），最后一个阶段（即标注）对语料库能起多大作用至关重要。所谓标注，是指计算机系统自动对未加工的语料库进行分析，使其具有语言学结构语义信息和其他信息特征标记。正因为语料库的功能跟语料标注的深度有如此密切的关系，因此，目前国内外许多研究机构的主要精力都花在用大量的人力、物力来制作大规模汉语标注语料库。据我们所知，北京大学计算语言学研究所跟日本富士通研究开发中心共同制作的一年《人民日报》（约 2600 万汉字）标注语料库是迄今为止世界上规模最大的汉语语料库之一。同时，上海师范大学也正在积极筹建“当代汉语语料库”。

总起来看，当前国内语料库加工的主要思路有两种（许嘉璐，2000）：

主流作法是以传统计算语言学为基本理论，循序研究语素—词—短语—句子—语段—篇章。北京大学开发的语料库基本采用此法，有这样几个环节：生语料—自动分词—语法标注—句法分析—语义语用分析—语言知识库。其中，词语加工的两个环节（即自动分词、词性的语法标注）是结合在一起同时进行的，所以，从未加工的生语料到形成语言知识库（静态词典、语法规则库和动态的上下文相关信息），中间主要经过三个环节，即词语加工、句法加工和语义语用加工（周强、段慧明，1993）。另外，许嘉璐教授主持的国家社会科学“九五”重大项目“信息处理用现代汉语词汇研究”包含了九个子课题①，重在解决现代汉语词的构造、分词、词类、兼类、词的语法属性等一系列中文信息处理技术所需要解决的基础性问题，是这一处理思想比较突出和集中的体现。

跟传统的基于句法知识的语言表述及处理模式不同的有黄曾阳先生的概念层次网络理论（HNC）。该理论认为：人对语言的理解本质上是一种认知行为，计算机对自然语言的处理就应建立在模拟人脑的这种语言感知过程的基础上。而人脑的认知机制“绝不是语法或句法，而是概念联想网络”，对联想网络的表述是语言深层（即语言的语义层面）的根本问题。联想网络分为局部和全局两类，前者对应着词汇层面、后者对应着语句层面。语料库加工的基本步骤为：语义块感知和句类假设—句类分析—语义块构成分析（黄曾阳，1998）。此外，具有探索性的研究还有陆汝占先生的基于内涵模型论的语义分析理论，目标是把汉语的语句表达式转换成逻辑公式，并进行模型解释，也是要深入到语义层面来处理汉语。但总的来说，从意义方面处理汉语的思想还没有用于大规模语料库的加工。

二、中文信息处理的难题

当“自然语言 = 中文或汉语”时，自然语言处理就是中文信息处理。因此，中文信息处理，就是利用计算机处理汉语信息（包括书面的和口头的）。跟其他

① 这九个子课题是：信息处理用现代汉语分词词表；歧义切分与部分专有名词识别；信息处理用现代汉语词类及标记集规范；汉语词类兼类问题；现代汉语词的语法属性研究（之一）；现代汉语词的语法属性研究（之二）——现代汉语动词电子词典的扩充和名词槽关系；现代汉语知识词典的建立和词汇内部语义网络描述；现代汉语真实文本短语结构的人工标注；现代汉语词的构造研究。可参见许嘉璐：《现状和设想》，载《中国语文》，2000 年第 6 期。