计算语言学研究系列

陈小荷 主编

汉英平行语料库的平行处理

冯敏萱 著

出版公司

汉英平行语料库的平行处理

冯敏萱 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

汉英平行语料库的平行处理/冯敏萱著.—北京：世界图书出版公司北京公司,2011.11
ISBN 978-7-5100-4110-5

Ⅰ.①汉… Ⅱ.①冯… Ⅲ.①汉语—语法—对比研究—英语
Ⅳ.①H146 ②H314

中国版本图书馆 CIP 数据核字（2011）第 243681 号

汉英平行语料库的平行处理

著　　者：冯敏萱
责任编辑：梁沁宁

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：三河市国英印务有限公司

开　　本：711mm × 1245mm　1/24
印　　张：10.25
字　　数：230 千
版　　次：2011 年 11 月第 1 版　2011 年 11 月第 1 次印刷

ISBN 978-7-5100-4110-5/H · 1261　　定　　价：25.00 元

《语言科技文库》总序

李葆嘉

当代语言学已经进入了一个科学与技术的互补时代，信息处理水平成为衡量国家现代化水平的重要标志之一。知识世界的载体是语符系统，信息处理的根本对象是语言信息处理。与计算机的出现使得语言符号有可能成为数据处理对象相似，神经科学实验仪器设备的应用，使得在大脑神经层面探讨语言机制成为可能。这些无疑都引导语言研究走向科技化，“语言科技新思维”（李葆嘉 2001）应运而生。所谓“语言科学”包括理论语言学、描写语言学、历史语言学、应用语言学等分支学科，所谓“语言技术”指语言研究的现代技术手段，包括语言信息处理、语音实验分析，以及语言的神经、心理和行为实验分析的技术手段等。就语言信息处理而言，又可以分为语料库研制技术、知识库研制技术、知识挖掘和抽取技术、句法信息处理技术、词汇信息处理技术、语音信息处理技术、语义信息处理技术、语用信息处理技术等。

2001 年 5 月，南京师范大学文学院创办了史无前例的“语言科学及技术系”，率先迈出了从传统文科教育范型向现代科技教育范型转变的步伐。“十五”期间，南京师大“211 工程”重点学科建设项目“语言信息处理与分领域语言研究的现代化”（陈小荷教授主持），以基础平台建设、资源建设和理论探索等为主，迈出了语言科技研究的一大步。

“十一五”期间，南京师大文学院、外国语学院和国际文化教育学院联袂申报“211 工程”三期重点学科建设项目。该项目以“语言科技”为引导，以“多学科交叉、跨院系整合、开放型营运”为理念，建设具有前瞻性、原创性、成长性的语言科技高级工作平台。以典型课题的工作原理为核心，进行资源开发和系统研制，拓展语

音科技、二语习得的神经机制研究、言语能力受损儿童的语言能力研究等新方向。同时造就新一代学术领军人物和培养一批高层次复合型人才，以期形成一支高水平的交叉学科团队。该项目设计，体现了工作平台建设、理论创新、应用研究、人才培养、团队建设的学科发展一体化思路。其旨趣在于，加速语言研究从传统文科范型向现代科技范型的转变，以引领21世纪语言科技的新潮流。

作为新兴交叉学科项目，通过教育部组织的专家匿名评审，“语言科技创新及工作平台建设”（2008～2011）获批，总投入1 000万元。总体而言，这一“语言科技创新”团队，分支学科齐全，专业知识互补。涵盖了理论语言学、计算语言学、语义科技、语音科技、实验方言学、历史语言学、神经语言学、二语习得研究、话语行为语言学等领域。这一期间，项目组成员获批的国家级基金项目达20多项。该项目理念之前瞻、实力之雄厚、工程之浩大、经费之保障，为学界瞩目。

2008年秋，本项目以南京师范大学语言科技研究所为实施单位正式启动。主要有三大任务：建设一个领先性的语言信息科技实验室、建立一个独创性的语言科技工作平台、撰著一套有特色的语言科技文库。

从实验室方案设计到设备招标采购，再到实验室用房改造，经过8个月的努力，2009年12月，语言信息科技实验室建成，为语言研究从传统范型向科技范型的转变提供了基本保障。该实验室划分为实验工作区、科研工作区和管理服务区。实验工作区建有语音实验与计算室、神经认知实验与计算室、课堂话语实录室三个专门实验室。科研工作区建有语义科技工作室、语音科技工作室、方言实验工作室、知识工程工作室Ⅰ（先秦词汇）、知识工程工作室Ⅱ（中古词汇）、知识工程工作室Ⅲ（敦煌俗语言文字）、语言习得神经机制工作室、语言习得中介机制工作室，以及参研工作室。管理区服务包括办公室、管理室、编辑室和交流室。出席“语言科技高层论坛暨语言信息科技实验室落成仪式”（2009年12月14日）的专家认为，该实验室体现了语言学跨学科研究的当代性和先进性，具有整体性、科技型、开放型三个特点，处于全国领先地位，是“语言科技新思维”的又一体现。同时认为，该实验室的科研工作涵

盖了四个二级学科、四个博士学位点，有稳定明确的研究方向，有合理的设计规划和很好的科研基础；整体设计合理，功能齐备。以教育部重点实验室建设标准衡量，很多方面超过了指标。

语言科技工作平台是基于工作原理（课题定位—理论方法—技术路线—关键技术—评估方式）而建设的高级平台。一方面，从语言信息、语言知识和语言机制三个层面，围绕典型课题进行设备配置、资源建设和软件开发；一方面，将典型课题研究与工作平台建设融为一体，依据典型课题建设的子平台应具有解决同类课题的功能。

建设语言科技工作平台的目标是要实现语言研究手段的技术化和模型化，总体设计包括三个二级平台和八个子系统。

一、语言信息工作平台　1. 语义科技工作系统（李葆嘉教授主持）：基于词汇语义－句法语义的一体化研究思路，开发“人－机交互语义标注工具”，研制“深度语义标注信息库”；研制“幼儿（2～6）日常话语跟踪语料库”，完成幼儿语义系统和话语行为分析研究。2. 语音科技工作系统（顾文涛教授主持）：研制“多语言、多语境、多语用的语音语料库”，基于声学信号分析、感知实验和数学建模，完善语音韵律理论与相关技术应用。3. 方言实验工作系统（刘俐李教授主持）：完成“网络版汉语方言有声语料库”，拟定系统的可操作性语音、词汇、语法实验模型和研究方法，进一步完善新兴交叉学科“实验方言学”。

二、语言知识工作平台　1. 先秦词汇统计与知识检索系统（陈小荷教授主持）：研制“先秦文献语料库”、“专名知识库”、“汉语词汇档案库”等，开发先秦文献自动分词算法、古籍版本异文自动发现算法、同指专名检索软件工具等，完成“先秦汉语词汇统计与知识检索”。2. 中古词汇统计与知识检索系统（董志翘教授主持）：研制“中古文献语料库”、“专名知识库”、“中古汉语词汇档案库”等，开发中古文献自动分词和标注工具等，完成“中古汉语词汇统计与知识检索”。3. 敦煌俗语言文字统计与检索系统（黄征教授主持）：研制“敦煌文献资料库”、“敦煌文献俗词语档案库”，开发相应工具，完成“敦煌文献资料与知识检索”。

三、语言机制工作平台　1. 二语习得的神经机制研究系统（倪

传斌教授主持)：研制“英语受蚀词汇库”等，基于行为学、脑成像和脑电三维度模型，进行中国人英语习得与磨蚀的神经机制研究，完成“基于神经机制的英语个性化学习分析系统”。2. 二语习得的中介机制研究系统（肖奚强教授主持)：研制“留学生汉语口语中介语语料库”，基于中介语理论、对比分析理论、偏误分析理论以及二语习得影响因素等，完成“留学生汉语习得的中介机制研究”。

这一工作平台，既是科技研究平台，也是人才培养平台，即一个现代化的科学研究和人才培养工作体系。

作为本项目的文本成果，《语言科技文库》包括计算语言学研究、语义语法学研究、汉语方言学研究、古代汉语学研究、语言教学与研究、语言新专题研究六个系列。其总体特征为：领域的开拓性、理论的原创性、选题的新颖性、方法的交叉性、考据的精审性、成果的应用性。在研究过程中，除了数据采集分析、资源建设和软件开发，更重要的还是要有新思路、新理论和新材料。陈小荷提出的先秦文献信息处理新方法，从先秦典籍注疏文献中挖掘出用于自动分词和词义消歧的知识，再注入已开发的古汉语分词和词性标注工具中去，所取得的先秦古籍版本异文自动发现、先秦词汇知识自动挖掘等成果均具开拓性。李葆嘉提出的语义语法学理论和话语行为理论，基于研制专用语料库或语义信息库和技术手段，开拓了语义网络建构、深度语义分析和话语行为研究等新的领域。刘俐李建构的实验方言学理论和方法，为方言学向现代科技方法的转型研究提供了新路，并取得了一系列新成果。黄征多年来从事敦煌文献及其俗词语文字研究，古代汉语学研究系列中的敦煌文献校录整理，以及敦煌写本字词考释、以古佚和疑伪经为中心的敦煌佛典词语和俗字研究、两汉声母系统研究等新见迭出。肖奚强基于汉语中介语语料库的二语习得研究，在对外汉语教学研究界已经产生了影响。钱玉莲的汉语介词与相应英语形式比较研究等专著各有亮色。倪传斌依据语言测试和认知实验等数据，从行为学、生理学和语言学三个层面分析影响中国英语学习者外语磨蚀的相关因素。刘宇红基于隐喻的理论探讨，对各类隐喻形式的结构、特性和解读规律进行了多视角的深入探讨。

《语言科技文库》所收论著，由作者在 2008 年 12 月申报选题，

2011 年始逐步完稿。系列主编审读了书稿，主要就其学术价值、章节安排、内容关联、行文表述、图表绘制等方面，提出审阅意见。此后，作者们对书稿又进行了修改和润色。《语言科技文库》的作者，大多数是具有博士学位的年轻教师。对于我们这些 20 世纪 80 年代走进语言学研究领域的而言，出版论著可能已不足为道。然而，对于年轻学者而言，其论著的出版既是几年来研究的结晶，也是对其继续探索的促进。换而言之，“211 工程”重点学科建设的目的之一，就是为年轻教师搭建一个可持续发展的科研和教学平台。学科带头人的主要任务之一就是提携后进。

尽管从根本上来说，科学或学术研究是一种个人的探索行为，然而复杂问题的研究，无疑需要群体协作。“学科建设”或团队合作模式，是 20 世纪 90 年代后期出现的一个新概念。这种模式涉及总体规划、多方协调，是需要付出精力和心血的。2008 年，通过投票方式推举我担任该项目总负责时，就意识到自己成了一个“劳动班委”。2009 年，前往安徽大学拜访黄德宽教授时，曾谈到“学科负责人的任务就是规划设计，争取项目经费和提供科研设备设施”，得到黄教授的赞许。2010 年，申报江苏省高校哲学社会科学重点研究基地时，评审专家柳士镇教授提问的“作为一个交叉学科项目，各学科之间的协调是怎么考虑的，有什么做法”，可谓一语中的。作为后学，深知交叉研究之艰、学科整合之难。相关学科之间的整合协调需要借助行政机制，但凭借行政方式并非就能完成。当时的回答是，目前做到的是建成了一个可以合作研究的场所，至于学科之间的进一步沟通合作应有较长过程。有一点很明确，只有通过交叉项目，相应学科才能渗透，合作者才能逐步磨合。我们只是在一步步探索。

十一五期间的“211 工程”建设项目即将完成，但是学科建设的任务并没有结束。2010 年，“语言信息科技研究中心”被评审为江苏省高等学校哲学社会科学重点研究基地，为“语言科技”这一交叉领域注入了新的建设活力。重点研究基地建设，除了“跨院系整合、多学科交叉、开放型运行”理念，需要凸显“合作性攻关”。围绕交叉性项目，实施计算语言学、语音科技、神经语言学、语义科技等力量的联合攻关计划。只有通过全面开放以及和与国内外同

行的合作交流，才有望建成具有影响的语言科技研究、人才培养和学术交流基地。

十年前，我（2001）曾写道："语言科技"的内涵是以理论研究为指导，以描写研究为基础，以应用研究为枢纽，促使语言研究向计算机应用、认知科学和现代教育技术领域等延伸，沟通文理工相关学科以实现语言研究过程及其成果的技术化。"语言科技"的外延为语言工程科技、语言教育科技和语言研究科技。其中，"语言研究科技"是将语言研究活动与资源建设、软件开发相结合，其目标是实现语言学自身的科技化。还应包含语言实验、数据处理这些实验语音学、神经语言学研究的科技手段。

虽然语言学家不可能也不必要都转向语言计算或实验研究，尽管描写、考据和内省始终是最基本的方法，但是具有一定的语言科技意识却非常必要。语言学家只有了解有哪些可供利用的资源、软件或仪器，才能提高其研究深度、精度和效率。语言学家也只有了解到信息处理的语言研究需求，才有可能为之提供可资应用或参考的基础成果。"语言科技"是21世纪语言学研究的潮流。

此为出版缘起。是为总序。

2011年8月谨识于南都

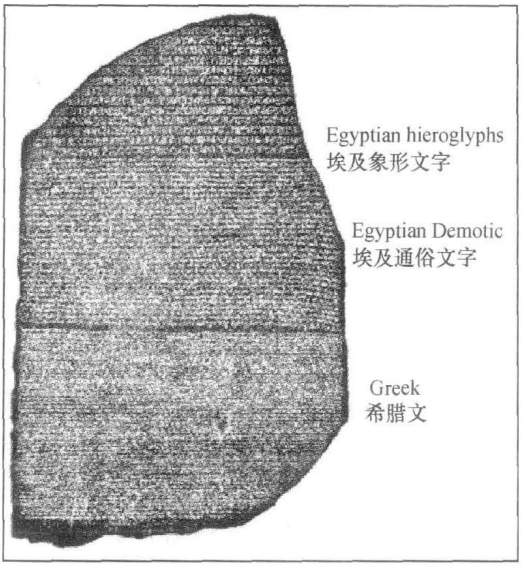
冯　序

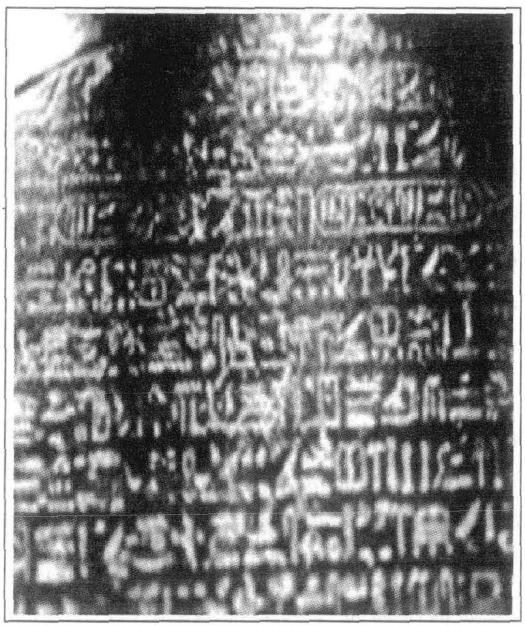
冯敏萱博士的《汉英平行语料库的平行处理》一书，以汉英双语平行语料库（parallel corpus）为研究对象，重点探讨了利用平行处理技术来加工汉英平行语料，排除双语各自语言中的歧义的理论与方法。这种平行处理思想的依据是达甘（Ido Dagan）等1991年提出的“两种语言的信息要比一种语言丰富”（two languages are more informative than one）[1]的主张，其目标是尽可能地利用另一语言来消除源语言中较难解决的歧义问题。作者采用规则与统计相结合的技术来实现平行处理，在对双语语料统计的基础上归纳规则、提取词例知识，获得了平行处理的宝贵的语言资源，使用了句珠层、语篇层、语料库层及外部知识层四个层次的语言资源来进行平行处理，在词汇分析、词性标注、词义标注、句法分析等方面进行排歧，都取得了良好的效果。本书的实践说明，基于平行语料库的平行处理方法是一种行之有效的自然语言处理方法。

这种基于平行语料库的研究方法源远流长。学者们在解读古文字的研究中，就使用过这种基于平行语料库的研究方法。解读密码（decipherment）是古典文献研究的一个重要内容，历代学者曾经依靠自己的聪明才智出色地解读了不少古代的铭文，或者通过铭文中已知的部分来解读铭文中未知的文字。罗塞塔石碑（Rosetta Stone）上古代埃及文字的解读，就是使用平行语料库方法来解读密码的一次成功范例。

罗塞塔石碑共刻有同一段诏书的三种语言版本，是用埃及象形文字（Egyptian Hieroglyphs，又称圣书体，代表献给神明的文字）、埃及通俗文字（Egyptian Demotic，又称草书体，是古代埃及平民使用的文字）和古希腊文（Greek，代表统治者的语言，这是因为当时的埃及臣服于希腊的亚历山大帝国，来自希腊的统治者要求统治领地内所有此类文书都添加希腊文的译版）三种不同的文字写成的，石碑刻于公元前196年，现藏于不列颠博物馆。我在2001年访问不

列颠博物馆时，曾经拍下了罗塞塔石碑的如下照片，读者从上面的照片可以看到这三种语言的版本，上层是埃及象形文字，中间是埃及通俗文字，下层古希腊文；下面的照片则展示了埃及象形文字的细部。





不列颠博物馆所见罗塞塔石碑

在公元4世纪结束后不久，尼罗河文明式微，不再使用埃及象形文字和埃及通俗文字，这两种文字的读法与写法都彻底失传了。虽然后来有许多考古专家与历史学专家极尽所能来研究，却一直解读不了这些神秘文字的结构与用法。直到1799年，法国远征军在埃及的Rosetta（罗塞塔）发现了罗塞塔石碑，才使埃及古代文字的解读工作获得了突破性的进展。罗塞塔石碑独特的三语对照写法，意外地成为解码的关键，因为这三种语言中的古希腊文是近代人类可以阅读的，利用这个关键来比对和分析碑上其他两种语言文字的内容，就可以了解这些失传的古代语言的文字与语法结构。在19世纪初期的英国物理学家汤马斯·杨（Thomas Young）和法国学者让-佛罕索瓦·商博良（Jean-François Champollion）的努力下，学者们依靠已知的古希腊文来解读未知的埃及象形文字和埃及通俗文字这两种埃及的古代文字，在1822年终于揭开了埃及古代文字的神秘面纱，成功地解读了埃及古代文字。

我们认为，罗塞塔石碑上面的三种文字就像三个彼此对应的平行语料库（parallel corpus），罗塞塔石碑也许就是世界上最早的三种语言的平行语料库。罗塞塔石碑的解读，是使用平行语料库解读密码的成功范例。

可惜，这样的成功范例当时在语言学研究中并没有得到推广，绝大多数语言学家仍然使用基于语感和个人语言经验的内省方式来研究语言。在20世纪90年代初，这种情况才发生了明显的改变。目前，双语平行语料库的建设已经引起了国内外语言学界的普遍重视，2011年国家社会科学基金重大招标课题中，就设有大规模英汉双语平行语料库的课题。这个课题现在已经正式立项了。

计算语言学兴起于20世纪50年代。20世纪90年代以前，从事计算语言学研究的绝大多数学者都把自己的目标局限于某个十分狭窄的专业领域之中，他们采用的主流技术是基于规则的句法-语义分析。尽管这些应用系统在某些受限的“子语言”（sub-language）中也曾经获得一定程度的成功，但是，要想进一步扩大这些系统的覆盖面，用它们来处理大规模的真实文本，仍然有很大的困难[2]。因为从自然语言处理系统所需要装备的语言知识来看，其数量之浩大、颗粒度之精细，都是以往的任何系统所远远不及的。而且，随着系

统拥有的知识在数量上和程度上发生的巨大变化，系统在如何获取、表征和管理知识等基本问题上，不得不另辟蹊径。这样，就提出了大规模真实文本的自动处理问题。

1990 年 8 月，在芬兰赫尔辛基举行的第 13 届国际计算语言学会议为会前讲座确定的主题是“处理大规模真实文本的理论、方法和工具”。这说明，实现大规模真实文本的处理将是计算语言学在今后一个相当长的时期内的战略目标，计算语言学正面临“战略转移”（strategic transit）的关键时刻。为了实现这样的战略转移，需要在理论、方法和工具等方面实行重大的革新。1992 年 6 月，在加拿大蒙特利尔举行的第四届机器翻译的理论与方法国际会议（TMI-92）将会议主题定为“机器翻译中的经验主义和理性主义的方法”。所谓“理性主义（rationalism）”，是指以生成语言学为基础的方法；所谓“经验主义（empiricism）”，是指以大规模语料库的分析为基础的方法。从中可以看出当前计算语言学关注的焦点[3]。

当前语料库的建设和语料库语言学的崛起，正是计算语言学战略目标转移的一个重要标志。随着人们对大规模真实文本处理的日益关注，越来越多的学者认识到，基于语料库的分析方法（即经验主义的方法）至少是对基于规则的分析方法（即理性主义的方法）的一个重要补充。因为从“大规模”和“真实”这两个因素来考察，语料库才是最理想的语言知识资源。但是，要想使语料库名副其实地成为自然语言的知识库，就有必要首先对语料库中的语料进行自动标注，使之由“生语料”变成“熟语料”，以便于人们从中提取丰富的语言知识。

可以看出，计算语言学当前面临着的这一场战略转移的关键是知识的获取方式和方法：从依靠“内省”方式转向依靠“语料”的方式，从“基于规则（rule-based）”的方法转向“基于语料库（corpus-based）”的方法，也就是“基于统计（statistics-based）”的方法[4]。

随着战略转移的深入，统计方法已经逐渐成为计算语言学的主流方法。

面对计算语言学的战略转移，我觉得，语言学研究中获取知识的方式方法也应当进行一场战略转移。

与战略转移以前的计算语言学相似，传统语言学家获取语言知识的方法基本上也是通过“内省（introspection）”进行，由于自然语言现象充满了例外，治学严谨的学者们提出了“例不十，不立法”（黎锦熙）和“例外不十，法不破”（王力）[5]的原则，这样的原则貌似严格，实际上却是片面的。在成千上万的语言数据中，只是靠十个例子或十个例外就来决定规则的取舍，难道真的能够保证万无一失吗？显然是不能保证的。因此，“例不十，不立法”、“例外不十，法不破”的原则只是一个貌似严格的原则，实际上是一个很不严格的原则。

在语料库出现之后，传统语言学的这个原则受到了严重的挑战！

语料库是客观的、可靠的语言资源，语言学研究应当依靠这样的宝贵资源。语料库中包含着极为宝贵的语言知识，我们应当使用新的方法和工具来获取这些知识。当然，前辈语言学家数千年积累的语言知识（包括词典中的语言知识、语法书中的语言知识等）是宝贵的，但由于这些知识是通过这些语言学家们的“内省”或者“洞察力”发现的，难免带有主观性和片面性，需要我们使用语料库来一一加以审查。英国著名语料库专家辛克莱（John Sinclair）一针见血地指出：“生造的例子看上去不管是多么的可行，都不能作为使用语言的实例。”[6]他大声疾呼：“我们总不能靠造几朵人造花来研究植物学吧！”[7]记得几年前在匈牙利的巴拉顿湖畔的美丽城市蒂哈尼（Tihany）开会的时候，在一次闲谈中，我对辛克莱说：“我们也同样不能根据赛璐珞造的玩具狗 Rex 来研究动物学。”当时他对于我的意见表示赞同。

如果搞语言研究不使用语料库或概率，很可能就只能使用自己根据“内省（introspection）”得到的数据，这是“第一人称数据（first person data）”；在使用第一人称数据时，语言研究者既是语言的数据的分析者，又是语言数据的提供者，有人把它称为“拍脑袋”得出的数据。或者使用根据“问卷调查”或“查词典”之类的“诱导（elicitation）”得到的数据，这是“第二人称数据（second person data）”；在使用第二人称数据时，语言研究者不充当数据的提供者，数据需要通过“作为第二人称的旁人”的诱导才能得到。如果使用语料库的数据作为语言研究的数据来源，那么，语言研究者就不再

充当数据的提供者或诱导者，而是充当数据的分析者了，这种“观察（observation）”得到的数据是“第三人称数据（third person data）”。

这是多年前威多逊（H. Widdowson）在 The Limitation of Linguistics applied[8]一文中提出的看法，我觉得这种看法很有价值，值得我们中国人思考。

当然，如果使用第三人称的观察数据，语言学研究者同时也可以充当数据的“内省者”或“诱导者”，所以，第一人称和第二人称与第三人称是难以分开的。这也就是我不反对“拍脑袋”这种第一人称数据的原因。不过，从总体上说来，第三人称数据显然是比较科学的。冯敏萱博士在本书中所用的就是“第三人称数据”。

乔姆斯基（N. Chomsky）的生成语法采用的是第一人称数据，他自己亲自来充当“理想的说话者”，由于他具有非凡的智慧，也可以取得卓越的成就；心理语言学、实验语音学采用的是第二人称数据，也取得了不少的成果，而我们现在则提倡第三人称数据。当然，与此同时，我们仍然要充分地尊重第一人称数据研究者和第二人称数据研究者的智慧和洞察力，我们并不反对第一人称的内省法和第二人称的诱导法。第一人称的“拍脑袋”方法固然会产生主观性，但是，脑袋拍得好也并不容易，语言研究中研究者的主观性往往显示了研究者的智慧和洞察力，不可忽视，所以，前辈语言学家的卓越的智慧和洞察力仍然是值得我们称道的。

不过，我们认为，语言学的一切知识，不论是过去通过“内省”或“诱导”得到的知识，最终都有必要放到语料库中来进行“观察”和“检验（verification）”，决定其是正确的，还是片面的或者错误的，甚至是荒谬的，从而决定其存在的必要性，决定其是继续存在，还是放弃其存在。

我们可以预见，语言学研究战略转移的时代必将到来！一种新的“基于语料库（corpus-based approach）”的研究方式或者“语料库驱动（corpus-driven approach）”的研究方式将逐渐代替传统的依靠“内省”和“诱导”的研究方式，传统的研究方式今后很可能只是基于语料库研究方式或语料库驱动研究方式的补充，而不是语言学研究的主流。当然，这种基于语料库的研究方式或者语料库驱动

的研究方式离不开语言学家的对于语言现象的“洞察力（insight）”，我们决不能忽视理性思维的重要作用。

传统语言学正处于面临战略转移的重要时刻，我们应当从高度的历史责任感出发，敏锐地认识到这个战略转移的重要时刻或迟或早总会来临，为此而调整我们的研究方法和研究计划，从而为世界的语言学宝库做出我们中国学者应有的贡献。

冯敏萱博士的这本书，是语言学战略转移中的一个引人注目的研究成果。冯敏萱博士要我写一个序言，我欣然同意了，写出了上面一些不成熟的看法，敬请方家批评指正。

冯志伟

2011 年 4 月 10 日

于德国 Heidelberg

注释：

[1] 冯志伟，1983，汉语句子的多叉多标记树形图分析法［J］，人工智能学报（2）：29-46

[2] 冯志伟，1996，自然语言的计算机处理［M］，上海：上海外语教育出版社

[3] 冯志伟，2010，自然语言处理的形式模型［M］，合肥：中国科学技术大学出版社

[4] 王力在《汉语史稿》（上册）（1980）中指出，“所谓区别一般与特殊，那是辩证法的原理之一。在这里我们指的是黎锦熙先生所谓‘例不十，不立法’。我们还要补充一句，就是‘例外不十，法不破’。”

[5] Dagan, I., Itai, A. & Schwall, U. 1991. Two languages are more informative than one. In *Proceedings of the 29th Annual Meeting of the ACL*, pp. 130-137.

[6] Sinclair, J. M. 1991. *Corpus Concordance Collocation*. Oxford University Press.

[7] Sinclair, J. M. 2007. Intuition and annotation: The discussion continues. In W. Teubert & R. Krishnamurthy, *Corpus Linguistics: Critical Concepts in Linguistics*. London and New York: Routledge, pp. 415-435.

[8] Widdowson, H. 2000. The limitation of linguistics applied. *Applied Linguistics*, 1, pp. 3-25.

陈　序

语料库是语言知识获取的源泉。对于平行语料库，以往的研究主要是从中获取翻译知识，并且为机器翻译服务，这当然是它的一个最重要的用处。冯敏萱的这部著作给我们打开了另一扇窗户，让我们看到也可以用它来解决单语处理的一些困难问题。这个思路以前也有人提出过，但是还没有人像作者这样系统地研究两种语言的平行处理。

所谓系统性，在论著中主要表现为两个方面：

第一，从未登录词识别、词性标注、词义标注到句法结构分析，几乎是自然语言处理的各个层面的歧义消解，作者都进行了汉英两种语言的平行处理研究，并取得了许多有价值的成果。

第二，不仅利用英语来帮助汉语消歧，而且还利用汉语来帮助英语消歧。一般人的印象是，英语形态比较丰富，英语的信息处理容易一些。作者让我们看到，英语中其实也有一些难解的问题，可以从汉语译文中取得帮助。当然，作者更注重的是汉语研究，在大体平衡的情况下，较多地侧重利用英语译文来解决汉语处理的问题，也是可以理解的。

如果在这一框架下两种语言都处理得不错，就可以更好地实现从平行语料库中获取翻译知识的目的。所以，平行语料库研究的两个目标并不矛盾，而是可以相辅相成的。

目前，基于统计的方法在自然语言处理中占据主导地位，自从在词性标注上被“战胜”之后，基于规则的方法还在用，但很少有人提及。其实，对于后者是怎样被“战胜”的，还有许多可议之处。在科学史上，没有哪种方法可以包打天下，或者被简单地否定。作者在平行处理中也使用统计方法，例如汉英人名的平行识别等，但主要使用的是基于规则的方法，并充分展示了这一方法的生命力。例如词性消歧和词义消歧，先用单语规则，然后使用译文中提取的