上海外国语大学规划基金项目资助

日语语料库研究的理论与实践

毛文伟 著

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS
www.sflep.com

上海外国语大学规划基金项目资助

日语语料库研究的理论与实践

毛文伟 著

上海外语教育出版社
外教社 SHANGHAI FOREIGN LANGUAGE EDUCATION PRESS

图书在版编目(CIP)数据

日语语料库研究的理论与实践 / 毛文伟著. --上海:上海外语教育出版社,2009

ISBN 978-7-5446-1221-0

Ⅰ. 日… Ⅱ. 毛… Ⅲ. 日语—语言学—研究 Ⅳ. H36

中国版本图书馆 CIP 数据核字（2009）第 009576 号

出版发行：上海外语教育出版社

（上海外国语大学内） 邮编：200083

电　　话：021-65425300（总机）

电子邮箱：bookinfo@sflep.com.cn

网　　址：http://www.sflep.com.cn http://www.sflep.com

责任编辑：应　允

印　　刷：上海外语教育出版社印刷厂

经　　销：新华书店上海发行所

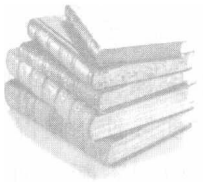
开　　本：850×1168 1/32 印张 6.375 字数 156千字

版　　次：2009 年 5月第 1版 2009 年 5月第 1次印刷

印　　数：2 100 册

书　　号：ISBN 978-7-5446-1221-0 / H · 0495

定　　价：18.00 元



前　言

依据实际语料对语言进行研究有着悠久的历史，最早可以追溯到中世纪。即便从现代意义上的第一个语料库布朗语料库建成至今，也已经经历了近半个世纪。20 世纪 90 年代以来，随着个人电脑的普及和功能的不断增强，利用语料库进行语言研究在技术条件上已经逐渐成熟。而随着研究的深入，人们也越来越清楚地认识到内省法研究的局限。一些研究者开始逐步尝试利用报刊、小说等素材，对语言现象进行实证性研究，涌现出一大批优秀的科研成果。

综观该领域，我们可以发现，语料库研究法的有效性不仅在理论上得到广泛认同，在实际的研究活动中也得到了越来越广泛的应用。但是，毋庸讳言，其中还存在着一些不尽如人意之处。尤其是相对于英语语料库在架构、内容和研究手段方面的日臻充实、完善，日语语料库的建设和应用无论是在研究理念，还是在具体实施上都无疑是较为落后的。

首先，缺乏经过周密规划、语料来源广泛、具有较好代表性的大型日语语料库。只有在强有力的理论支撑下，语料库才有可能具备较好的平衡性和代表性，全面、忠实地反映语言的本来面貌。为此必须进行科学规划，并在语料的选择、录入、校对以及后期维护方面投入极大的人力、物力。这也是目前日语研究界尚不具备

一般意义上的大型均衡语料库的原因之一。

其次，缺乏高效、实用的检索工具。从理论上说，语料库规模越大、取材越广泛，就越能客观、全面地反映语言的实际使用情况。但是，这也导致研究者花费在筛选、整理例句上的时间不断增加。因此，必须配备符合语言研究要求的检索和统计工具，实现对语料数据的部分自动处理，提高研究效率。此外，也有一些语言现象缺乏形态特征，如果单纯地依靠关键词进行抽取的话，会混入大量冗余信息，给进一步分析带来困难。这就需要我们借助形态素自动分析、词性自动辨识、语法信息自动添加等各种智能化技术，以实现各种丰富多彩的功能，满足更高层次的检索、统计要求。在这方面，欧美语言学研究者走在了前面。运用 Wordsmith Tools 等工具可以方便地进行制作词表、提取主题词等操作。而由于日语的特殊性，现有的一些软件均无法很好地满足语言研究日趋复杂多样的需要。

第三，存在着重应用、轻理论的倾向。研究者的注意力往往集中在从语言素材中就事论事地寻找结论，而疏于方法论方面的思考。这导致部分基于语料库的研究质量不高，甚至出现谬误。因此，建立一套语料库应用方面的理论体系是当务之急。

本书正是为了解决以上问题而作的一种尝试。在第一部分理论篇中，笔者探讨了语料库研究的理论意义和学科定位，分析了日语语料库建设的现状以及存在的问题。在此基础上，提出了设计和构建一个结构合理、内容完备、便于使用和维护的语料库所需注意的一些具体问题。接着，介绍了语料库信息自动抽取技术的原理及实现方式。最后，讨论了语料库素材的性质以及研究手法对结论的信度可能造成的影响及对策。

在第二部分实践篇中，笔者尝试运用语料库研究法解决一些

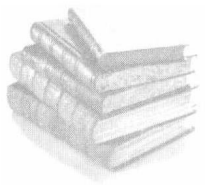
具体的研究课题。通过这些探索，不仅再次印证了语料库在各类语言研究中可能起到的巨大作用，也是对理论篇部分内容的实践、补充和扩展。

当然，本书绝不可能涵盖语料库研究的全部内容。受学识所限，在具体的论述过程中，难免会挂一漏万或有失偏颇。在此仅作抛砖引玉，恳请广大研究者不吝指正。

本书是上海外国语大学规划基金项目“日语语言研究与语料库开发应用”的研究成果之一。并在上海外国语大学学科建设规划项目“面向日语语言研究的语料库建设”的资助下，得到了进一步完善。上海外国语大学科研处、学科办、日本文化经济学院的各位领导以及学界的前辈专家学者为笔者创造了良好的条件，提供了宝贵的指导和支持。在此，谨表诚挚谢意。

毛文伟

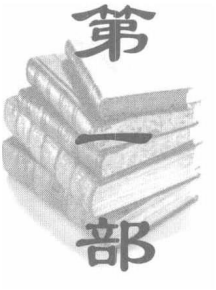
2009 年 1 月



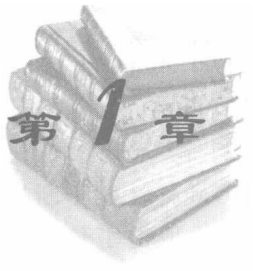
目　录

第一部　理论篇

第二部 实践篇



理论篇



语料库研究概述

1.1 何谓“语料库”

关于语言学研究应该以何为依据的问题，迄今为止人们进行过许多探讨。从素材来源以及使用途径来看，大致可以分为依赖本人内省、进行问卷调查以及对语料库素材进行大规模检索三种。这些方式可以说各有千秋，至今尚无万全之策。

尽管研究者的语感对于语言研究是不可或缺的，但正如人们屡屡指出的那样，基于个人内省的判断很容易受到地域、年龄、个人表达习惯等诸多因素的影响，有时难以得出正确结论。于是，研究者开始试图采用问卷调查的方式探究某种语言形式的被认可程度，并对调查结果进行量化和统计。但是，由于受到预算和时间的限制，研究者可能接触到的被试者人数往往会低于必需的数量。

以丰田丰子(1985)为例，该论文以日语母语使用者为对象实施了问卷调查，并以此为依据分析了接续助词「と」、「ば」、「たら」、「なら」在使用上的区别。但是值得注意的是，被调查者仅有31人。如果进一步按照年龄、出生地和性别等进行分类，每组的人数就更少。因此，其结果的可信程度令人怀疑。此外，如果不对问卷调查的内容和表达方式进行仔细推敲，也容易对接受调查者构成暗示，从而导致调查结果失真。

正是由于基于问卷调查的研究方法存在着以上不足，自 20 世纪 80 年代以来，研究者开始尝试运用存储于计算机内的大量实际语料对语言现象进行考察，涌现出大量基于语料库(corpus)的研究成果。这一方面是由于计算机技术日新月异，硬件性能不断提高，海量存储设备逐渐普及，软件在易用性方面也取得了长足进步，在很大程度上降低了研究者进入的门槛。另一方面也是由于人们对于实际语料的价值有了新的认识。

这些研究的共同之处在于，它们都是在大量现有的出版物或口语素材中检索包含某种语言现象的实例。通过进一步的观察、分析和统计，总结出其中蕴含的客观规律。因此，从本质上看，都属于实证性研究方法。近年来，基于语料库的实证性研究已经产生了许多成果，在语法、词汇以及句法等各个领域奠定了自己稳固的地位。

对于这种建立在观察分析自然语言现象基础上的实证性研究来说，语料库是不可或缺的研究基础。实证性研究方法在语言学研究的各个领域广受推崇也使得语料库的重要性不断提高，逐渐成为研究者不可或缺的工具之一。

语料库一词来源于拉丁语 corpus，意为“资料的总体”。英语中的 corpus 一词继承了拉丁语的原意。但我们现在所说的语料库显然不再是简单的资料的集合，它被赋予了新的意义。

Crystal(1991)认为，语料库是语言资料的集合，其资料来源既可以是书面语篇，也可以是话语的记录脚本。Sinclair(1991)将语料库定义为对自然发生的语篇的收集，目的在于描述一种状态或某种语言中的变化。Biber(1998)认为语料库是对自然语篇大宗的、有原则的收集。顾曰国(1998)则将语料库定义为放置语言材料的仓库，存放在计算机里的原始语料文本或经过加工后带有语言学信息标注的语料文本。

以上这些定义分别涉及了语料库的素材来源、使用目的、存在

形式和收集方法。将其综合起来，就可以得到较为完整的语料库的定义。笔者认为，语料库是以服务于语言研究为目的、按照某种标准收集并以数据形式存储在计算机内[1]的大量被实际使用过的书面语或口语素材。

1.2 关于语料库在语言研究中的价值的论争

依据实际语料对语言进行研究有着悠久的历史，最早可以追溯到中世纪。例如，克拉克夫人曾为莎士比亚作品制作索引。这可以视为语料库研究的雏形。初期的语料库主要被应用于五个领域：圣经与文学研究、辞典编撰、方言研究、语言教育研究和语法研究(Kennedy，1998)。除此以外，通过实际调查采集语料也是进行儿童语言习得、拼写、语言教学等方面研究常用的方法。

关于语言学研究中实际语料的价值问题，人们曾经走过两个极端。20 世纪 50 年代，受到实证主义和行为主义的影响，部分美国语言学家将语料的价值绝对化，忽视甚至否定直觉判断在语言研究中的作用。他们认为，对于语言学研究来说，收集足够多的自然语言素材不仅是必需的，而且是自足的。直觉是第二位的，甚至可以完全予以抛弃。

然而，自然语言的句子具有无限的可能性。语料库规模的有限性决定了语料难以被穷尽。如果只重视自然语料，忽视直觉的作用，便无法解决语料有限而语言无限的矛盾。同时，完全否定研究者通过内省对于语言规律的主动认知能力，又使得语言学研究成为对语言现象的简单罗列，无法加以总结，并找出规律。实际上，很多研究都是始于由直觉产生的语言学假设，或是在观察实际语料的过程中，通过研究者的直觉形成初步判断，继而在分析大量实例的基础上，对其进行检验或是修正，并最终得出结论。如果忽视甚至否定直觉判断在语言研究中的作用，必然会造成研究者面

对大量语料无从着手，找不到研究的出发点和方向。

以本书第8章为例，为了考察日语机能辞[2]「てならない」、「てしようがない」、「てたまらない」的异同，笔者将出现在「てたまらない」之前的表达形式[3]整理后得到下表[4]。

表1 「てたまらない」的先行词一览

先行表达	例句数[5]	先行表达	例句数
たい	53(15.1)	かわいそう	7(2.0)
うれしい	23(6.5)	残念	7(2.0)
いや	16(4.5)	こわい	6(1.7)
心配	13(3.7)	腹がすく	6(1.7)
気の毒	9(2.6)	～にくい	6(1.7)
さびしい	9(2.6)	不愉快	6(1.7)
癪に障る	9(2.6)	愉快	5(1.4)
かわいい	8(2.3)	気になる	5(1.4)
不安	8(2.3)	不快	5(1.4)
おかしい	8(2.3)	不思議	5(1.4)
すき	8(2.3)	いたい	4(1.1)
くるしい	8(2.3)	ほしい	4(1.1)
さむい	7(2.0)	面白い	4(1.1)
くやしい	7(2.0)	うるさい	3(0.9)

在表1所列举的先行表达中，有动词、形容词、形容动词，甚至于还有一些惯用词组，似乎杂乱无章。同时，正如下文中将着重指出的那样，不能仅凭搜索到的例句多寡就断言哪些表达形式更容易出现在「てたまらない」之前。这时，必须发挥内省的力量，从中