

JIYU YULIAOKU DE
RIYU YANJIU

基于语料库的 日语研究

戴宝玉 著

学林出版社



本书得到沪江网校外语研究基金资助

JIYU YULIAOKU DE RIYU YANJIU

基于语料库的 日语研究

戴宝玉 著

学林出版社

图书在版编目(CIP)数据

基于语料库的日语研究 / 戴宝玉著. —上海: 学林出版社, 2012. 1

ISBN 978 - 7 - 5486 - 0273 - 6

I. ①基… II. ①戴… III. ①日语—研究—文集
IV. ①H36 - 53

中国版本图书馆 CIP 数据核字(2011)第 255675 号



基于语料库的日语研究

作 者—— 戴宝玉

责任编辑—— 胡小波

封面设计—— 周剑峰

出 版—— 上海世纪出版股份有限公司

学林出版社(上海钦州南路 81 号 3 楼)

(电话: 64515005 传真: 64515005)

发 行—— 上海世纪出版股份有限公司发行中心

(上海福建中路 193 号 www. ewen. cc)

印 刷—— 上海港东印刷厂

开 本—— 890×1240 1/32

印 张—— 7. 875

字 数—— 195 千字

版 次—— 2012 年 1 月第 1 版

2012 年 1 月第 1 次印刷

书 号—— ISBN 978 - 7 - 5486 - 0273 - 6/H · 22

定 价—— 23. 00 元

(如发生印刷、装订质量问题,读者可向工厂调换。)

日语语料库与日语研究

(前言)

1. 序言

在日语教学和研究中,对某个词或某种表达是否成立进行判断是常有之事,对一些简单的语言现象可以查阅辞典或请教他人,但是要解释、判断生僻的或者是复杂的语言现象,仅凭借一般辞典和普通人的知识是难以解决的。如果要对这类语言现象作进一步深入的研究,就不得不借助于用计算机来处理的大量的语言数据,简而言之,这就是语料库。

自 20 世纪 70 年代美国制作了 Brown Corpus 以来,世界各国都开始研制语料库,种类越来越多,规模也越来越大。语料库就种类而言,有文本语料库和语音语料库,口语语料库和书面语语料库,单语种语料库和多语种对译语料库^①,带标注语料库和不带标注的纯文本语料库等,最近比较受关注的还有学习者语料库。随着电脑技术的飞速发展,语料库制作不再限于国家研究项目或是某个大学的研究项目,个人只要投入足够的资金和时间,也可以制作出相当规模的语料库。而对语料库的追求是永无止境的,从规模和内容上来说,永远不会不偏不倚,尽善尽美,短期内难以实现一个理想的语

^① 笔者有幸参与北京日本学研究中心主持的中日对译语料库研制的部分工作,该语料库规模为中日文各 1 000 万字,详见徐一平(2002)。

料库。

除了对语料库建设本身的研究,更重要的是利用语料库进行的语言研究,在此且称为语料库的应用研究。语料库只是手段,研究才是终极的目的,利用现有的语料库,发挥它的长处,只要有新的切入点,就可以解决许多实际问题。本文就是对语料库应用研究的一个尝试,探讨如何利用各种语料库,如何去发现新的语言事实,从而充实和改善甚至是纠正现有的语言研究成果,以下从入门的角度探讨教学和研究两方面的语料库的应用。

2. 语料库应用举例

语料库的最直接的应用就是检验某个词或者某个表达方法是否实际存在。正如田村野(1995)所指出的,人通过内省可以联想起的事物十分有限,比如要求某人当场无一遗漏地举出所有友人的姓名几乎是不可能的,在语言教学中也不例外。理想的语料库涵盖各种文体和不同时代,包罗了所有的语言现象,而决不是简单地局限于语法或词汇。以下以从网上下载的日本国立国语研究所公开的语料库“himawari”为例,说明如何利用语料库来解决教学和研究中的实际问题。

日前,某教师编写了教材日语,笔者对其中引号的用法感到很不自然,但这只是基于长期阅读和积累形成的一种直观判断,本人并非研究标点符号的专家,没有确切的依据。据该教师称,他在教学中一直强调日语中的句号只能在引号之内,不能用在引号之外。如果就句号和引号关系而言,可以有以下三种组合,即:

(1a) ~た。」

(1b) ~た」。

(1c) ~た」

显然,这位教师赞成上述的(1)的写法,反对(2)的写法,但是笔者直观觉得真正最为常用的是(3)。如果用语料库来验证一下,结果就一目了然了。为此笔者分别用「た。」、「た」。和「た」进行检索,其结果是,(1)为126例,(2)为79例,(3)为4877例。可见,这位教师的见解没有错,但是他忽略了日语中句号和引号的另一种更为普遍的书写法。一个简单的调查足以改变长期以来形成对语言事实的误解。

其实,使用语料库来研究语言,实际上就是一场低成本高效率的大规模问卷调查,只不过不是面对面的调查而已,而且还可以根据自己的需要,不断地改变问卷内容,即改变关键词,从调查者中得到自己最需要的结果,直到满足研究的需要。比如在上述问卷调查中共抽出5082句例句,排除3例无法确认作者的例句外,这些例句共涉及1205位作者,由于himawari还提供了作者的出身年月和性别,等于同一时间内向1200位不同年龄不同性别、具有高度语言素养的人实施问卷调查,其结果是非常可信的。这种调查在现实社会中几乎无法实现,而语料库却可以轻松帮助我们达到这个目的。除了调查以上(1a)(1b)(1c)以外,还可以从年龄角度进行检索,检索结果显示,以上(1)(2)(3)的使用频度最高的前5位分别如下,数字表示其年代。

(2a) 1920,1950,1940,1960,1930

(2b) 1930,1940,1950,1920,1960

(2c) 1930,1920,1950,1950,1890

由此可见,以上三种标点符号的搭配虽然使用频度有所不同,但基本不受年龄的影响。那么这三种用法是否受到性别的影响呢?该语料库还提供了作者的性别信息,因此很容易从这个角度对上述标点符号再次进行更深入的统计,排除非个人作者和多个作者合作

的作品,结果是:

(3a) 男 84 例: 女 16 例

(3b) 男 55 例: 女 6 例

(3c) 男 3 230 例: 女 766 例

从这个结果可以看出,性别对上述三种书写法也没有特别的影响,男女作家都使用这 3 种搭配方式。通过以上调查可以认为,就使用频度而言, $3>1>2$ 是一种普遍存在的现象,而且不受年龄和性别的影响。这个调查结果对编写教材和指导学生作文具有很大的指导意义,防止了对语言理解的片面性和盲目性。

以上大致分析了标点符号的使用方法,如果进一步深入分析就会发现,大部分作家标点符号使用比较专一,如车谷长吉只用(1),赤川次郎只用(3),(2)由于使用者较少,所以没有专用(2)的作者。还有部分作者是兼用两种甚至更多的,比如安部公房使用(1)和(3),野坂昭如使用(2)和(3),而山口修则三者都使用,不过这样的作家人数极少。

如果采用传统的纸质问卷调查,虽然问题本身并不复杂,然而要完成这个调查,却是耗时巨大,而且要承受相当的经济负担,而用语料库来作同样的研究,由于上述语料库是免费下载的,所以没有经济负担,对数据的检索和统计分析仅在 20 分钟以内就可以完成,可谓省时高效。如果有一个好的切入点,一个崭新的创意,可以写成一篇很好的研究论文。当然使用小型语料库作研究,只限于常见的语言现象,如果是一个非常冷僻的语言现象,检索所得例句有限,就没有足够的参考价值。

作为比较简单的语料库应用的研究,以下再举一例。

如果有报刊语料库,那么,研究就不限于语言研究,还可以用于社会研究,因为报刊及时反映了社会的各个方面的情况,因此只要

选取的关键词得当,可以很直观地反映出各种社会问题。

比如笔者在使用「男性」「女性」为关键词观察它们与前接名词和后续名词构成的复合名词的调查中,如「单身女性」和「女性社員」,使用了日本数年的报刊,共抽取例句 34 983 句,以下主要涉及「男性社員」和「女性社員」这样的与后续名词构成复合名词的用法。

使用上述关键词,目的是为了从数据上来分析这些复合名词的使用状况,从而达到对日本社会男女平等现状的更为深入客观的了解,而不是停留在主观的臆测上,经过检索和数据处理后,得到的部分复合名词的实际使用状况如表 1。

表 1 「男性」「女性」与后续名词

	会社員	従業員	教諭	教師	医師	議員	記者	監督	研究者	弁護士	差別
男性	824	180	530	113	110	38	50	3	6	13	8
女性	535	296	327	163	113	377	278	115	102	67	116
	一般職業 男女比 1 757 : 1 434					知的職業 男女比 110 : 939					

在表 1 种,笔者将后续名词分为两类,将「男性会社員」「女性教諭」称为一般职业,将「男性議員」「女性記者」称为高级职业,从表中可以看出,在一般职业中,男女的比例是男性略多于女性,考虑到部分女性婚后放弃原有的职业,这个数字直接反映了目前日本男性和女性在一般职业中实际就业的情况。而在高级职业中,数据显示女性明显高于男性,而且职业越高,男女比例之差就越大。但是这并不意味在日本的高层次职业中女性占了明显优势,现实却是恰恰相反。本来,某些职业都是男性的天下,因此说到从事这个职业的人员,没有必要提及性别,不必一一说明「男性議員」,而当女性开始进入这些职业时,就开始有了区分性别的必要性。「女性議員」是为了区别大多数的男性议员而产生、存在的复合词,它的频繁出现正好

说明在日本,一方面女性议员有所增加,另一方面女性的议员在政界中还是占少数,正因为如此所以才一一注明。

表1中一般职业中的男女比例大致反映了实际的男女职业分布状况,而高级职业中的男女比例反映的却是数字背后的实际的不平等。因此从表1可以得出结论,在日本,在一般职业中基本上是男女平等的,而在高级职业中女性还远不及男性,这从最后一栏的「男性差别」和「女性差别」这两个复合名词的实际使用状况也可以得到证实,相对于男性歧视,在日本存在更多的还是对女性的歧视。

用前接名词进行研究同样可以获得有价值的线索,比如,在同一语料库中,「独身男性」和「独身女性」分别为55:110,而「未婚男性」和「未婚女性」的比例分别为9:82,这个数据从一个侧面说明了日本现在结婚适龄男女的婚姻状况,如果数据库足够大,对研究日本的婚姻状况一定会有很大帮助。

如果条件允许,比如具备1945年至今的报刊资料,那么就可以分别从不同年代的报刊中抽取上述关键词,然后按年代分别列出加以比较,可以非常客观地描述战后日本社会各个方面的变化,从中可以发现变化的规律。

综上所述,利用日本的报刊资料,只要有独到的视角和良好的切入点,可以获得很客观的研究成果,因为报纸可以说是涵容万象的,所以提供了我们利用语料库从各方面研究日本的可能性。现在日本主要报纸都发售光盘,记录该年度报刊的所有内容,为中国各大学和研究机构研究日本社会提供了非常方便和有效的工具。

3. 语料库与规模

以上涉及的都是一些很常见的语言现象,因此小规模语料库也足以应对,然而对一些很冷僻的语言现象,要想获得满意的结果,

就只有扩大语料库的规模。比如日语中有「ばこそ」,它分别接在假定形和已然形之后表示不同的意义,比较常见的是已然形,而假定形在 himawari 中只能检索到 5 例,而且不是所有检索所得例句都具有研究价值,所以研究「ばこそ」,himawari 就显得捉襟见肘了。此时,就必须使用规模更大的语料库。为此,笔者使用了自己开发的规模更大的语料库,有效的未然性例句达 25 句。假定形和已然形的「ばこそ」的例句分别例示如下。

- (4) 君というパートナーがいればこそ、できることだと思う。
(「女たちのジハード」)
- (5) 方針も指針もあらばこそ、無我夢中でここまで来た、というのが家内にとっても実感だろう。(「親父の全力投球の姿」)
- (6) 戸棚の口から弾丸の如く飛び出した者が、避くる間もあらばこそ、風を切って吾輩の左の耳へ喰いつく。(「吾輩は猫である」)

收入バコソ的词典很少,如《明镜国语词典》解释如下:

- (a) 理由を強調して示す。……からこそ
- (b) そんなことはまったくないと全面的に否定する。

《大辞泉》解释如下:

- (a) あとの説明を成立させるのに十分な理由を、前に提示する意を表す。
- (b) 中世以降終助詞的に用いて強い否定の意を表す。…などするものか。…絶対ない。

上述例句中,对于(4),学生利用现有的知识就可以理解。(5)在目前笔者接触到的高年级教材中没有出现,如果出现,可能要费一番周折。首先,(5)的解释根据前后关系,可以做不同解释,没有固定的译法,比如在此处,可以理解为「～ないまま」,也可以根据前后文理解为「どころか」等。对(5)这样的例句只有掌握更多的例句,才能对该用法有个比较全面的理解,因此就必须使用更大规模的语料库。

在未然性用法中,(6)是(5)的一种派生用法,在(6)中「間もあらばこそ」,直译的意思为「時間がない」。但是此处已经赋予一种全新的意义,表示两个动作之间的瞬间接续,而不是简单的否定,因此可以理解为现代日语的「～たとたんに」「～するやいなや」,由于始终以固定的形式出现,而且已经与特定的意义结合,所以作为一个新的复合助词将它和(5)区别对待更有利于理解。本次共检索到8句(6)这样的用例,虽然数量有限,但是足以发现バコソ的另一种新的作为复合助词的用法。

バコソ可以说是个极端冷僻的语言现象,相比而言,ニシテモ比较常用,但是如果使用规模较大的语料库还是会有新的发现。一般认为ニシテモ有三种用法,分别为:

- (7) 韓国は、日本とはいろいろなちがひがあります。道路交通にしても、[中略]人は左、車は右側通行です。(「海をわたった盲導犬ロディ」)
- (8) 彼女とは、親友というほどではないにしても、わりと親しかったわ。「三浦綾子全集」
- (9) アメリカにしてもイギリスにしても、名門学校はすべて私立である。(「ユダヤに学ぶ世界最強の勉強法」)

虽然简单地将它们纳入现有的语法体系有一定的困难,至少可

以分为副助词,接续助词和并列助词^①。此前笔者对以下(10)的词性判断感到困惑,因为ニシテモ在接在用言之后这点上属接续助词,然而,这个用言同时又接受「同じ」的修饰,意味着它不是一个简单的用言,而具有体言的性质,如「同じ人」。于是根据ニシテモ前出现的单词的词性来判断ニシテモ的词性,在此就失去了应有的作用,因为没有有一个有效的手段来区别这个「捨てる」究竟属于何种词性。

- (10) 同じ捨てるにしても、どこかべつの場所で捨てたほうがいい。(「海辺のカフカ」)

然而,笔者在整理语料库中发现了不少以下例句,如(11)。

- (11) この私ににしても、それはまったく同じだった。(「eの悲劇」)

这个例句说明ニシテモ作为复合助词虽然在形态上已经趋于稳定,但是有时候允许有所变化,成为ニシテミテモ,不过这种变化只限于(11)这样的接在名词之后的副助词,而不存在(12)这样的接在用言之后的用法。

- (12) *彼女とは、親友というほどではないにしても、わりと親しかったわ。

如果(10)也不能作(12)这样的改动,那么由此可以断定(10)中的「捨てる」虽然在形态上为动词,实质上是名词,因此它后续的ニ

① 参见戴宝玉(1986)。

シテモ是副助词,排除了接续助词的可能。

作为研究,仅限于通过语料库揭示(10)中的ニシテモ的词性还是不够的,对这种现象的观察结果不应限于该现象本身,而应举一反三,这样就可以对其他的复合助词研究有所裨益。比如借助(11)的启示,可以发现(13)中也存在同样的现象。

- (13) 彼女にとってみれば、それは腹の立つことだったろう。
(「人の砂漠」)

由此可见,部分复合格助词的形态还并非完全固定,是否还有其他的这类复合格助词,他们的存在对复合格助词整体研究的意义如何,将是一个新的研究课题。毋庸置疑,这种新的研究课题的发现,得益于一个大规模的语料库。

4. 语料库与年代

如果按年代搜集语料,那么就可以在时间坐标上对某个词进行观察,如果这个语料库具有一定规模且构成合理,在研究词汇时就具有积极的参考意义。

比如,「朝シャン」是「朝シャンプする」之略,这个词作为流行语极少被词典收入,使人无从了解这个词的确切含义,但是语料库却可以弥补这种不足。这个词在本人的语料库中最早出现在以下(14)中^①。

- (14) 朝シャンして会社へ出かけ、昼食抜きで美容院へ駆けつけて昼シャンをすませ、夜、帰宅するやいなや夜シャン

① 例句出处的数字表示该书的出版年月。

にいそしむ。(「はみだし人間の系譜」1996)

(15) 聞くところによるとかれらの清潔願望は大変なもので、朝シャンとやらで毎朝洗髪するのだそうだが…(「自分の顔を持つ人になる」2001)

(16) 朝シャンといって毎朝これで頭を洗うことが専ら男子高校生にはやった。(「一寸さきはヤミがいい」2003)

由此可见,「朝シャン」意为早晨洗发,不但有「朝シャン」,还有「昼シャン」「夜シャン」。从作品发表年代和(16)的刻意解释来看,这个新词大致出现在20世纪80年代末期。事实上,这个词在1987年曾经得到《流行语大奖》中的表现奖^①,不过作为研究,还需要更为深入的调查。

此外,「超」也曾是颇受关注的另一语言现象。关于「超」,米川(1996)介绍如下:

この「チョー(超)」が東京周辺では強意の副詞として若者たちによく使われている。井上史雄氏によれば、「チョー」は八〇年代半ばに東京の女子大生が使った例が記録されており、まもなく高校生や中学生も使うようになったという。

「超」作为前缀早已有之,《新明解国语词典》收入含「超」的单词20个,对「超」的解释是:

- (a) ある限度を超える。超越、超過、出超、入超
- (b) 普通の人とかけ離れている。超然、超俗、超人的
- (c) 比較を絶している。超特急、超満員、超短波、超国家主義、

① <http://zokugo-dict.com/01a/asashanpoo.htm>

超高層ビル

笔者检索结果得例句共 660 例、180 种,按使用频度排序前 10 名分别为:超能力(98)、超常現象(33)、超伝導(25)、超人(24)、超一流(24)、超高層(23)、超音波(22)、超満員(20)、超高級(13)、超自然(13)。从统计结果来看,它们的意义并未超出词典收词的范围,检索结果前 10 名都属词典列举的第 3 种用法。

「超能力」以异常高的频度位居榜首,是因为语料库收入了三篇相关内容的推理小说和科普文章。作品中频繁出现某个词,在检索结果中就会反映出来,使这个词的使用频度陡升,所以对语料库检索的结果必须结合语料的构成加以全面的斟酌考虑,这就是卷首提及语料库不可能尽善尽美的含义之所在。相对于语法研究,词汇研究更容易受语料库构成的影响,必须慎重对待。

通过对检索结果的分析可知,从 1960 年代起,「超」的接续范围从纯名词开始向表示性质状态的名词形容动词扩展,出现了「超大型の」「超巨大な」「超豪華ホテル」「超モダンな」「超ハードな」「超まじめ人間」等用法。20 世纪 80 年代前后,这种趋势越发明显,最终扩展到了后续形容词和动词,从名词到形容词动词的扩展经历了数十年的时间,以下为这方面的用例。

- (17) この時代の浅草の魅力は、浅草寺や仲見世といった古い物が、デパート、カフェー、シネマ、レビューといった超新しい物と隣合せになっていたことから発していたと思われる。(「私説東京繁昌記」1984)
- (18) 元祖東京版は「超やさっしー」のに対して、広島は「ぶちやさしんじゃー」。(「現代若者ことば考」1996)
- (19) 「超ムカツク」というのが、この節の流行り言葉だとは知っていたが、(「ものは言いよう」2000)

但是,即使「超」后续词的范围不断扩大,它仍只是一个前缀,而不是副词,由「超」构成的复合词结构紧密,中间插入「は」或与修饰的用言的倒置都不被容许,与副词和被修饰语之间只是一种松散的修饰关系大不相同。

此外,「むかつく」原意有二,即生理上的不快感和心理上的不快感,加上「超」后,意义就被限定于后者,不能说「胃が超むかつく」,而且后者在很大程度上丧失了动词的许多语法范畴,在意义和功能上更加接近形容词。在这个意义上,对后续词在意义和语法范畴上的限制与一般副词也不同。凡此种说明「超」仍停留在前缀上,它非常接近关西方言中的「ど根性」「ど阿呆」「どぎつい」「どえらい」,只是「超」与「ど」相比,它不受地域限制,接续词的范围更为广泛而已^①。不过,最近这个「超」似乎正在不断扩大使用范围,向着动作动词进军,如(20)。

(20) 外国人と話すのって超緊張しない? (「留学ってドキドキ!」2003)

以上例子表明,只要有一个大规模的按照年代排序的语料库对研究词的变化是有效的,如果这个语料库时间跨度达 100 年以上,会给我们带来许多有益的帮助。

5. 语料库与文体

口语和书面语在问题是不同的,由于如何界定口语和书面语可以有不同的见解,而且获得纯粹的口语的语料库目前尚有困难,

^① 网上可找到以下例句。「大学時代、西部の人は『超むかつく』っていうのを『どむかつく』って言ってたっけ。最近聞かねえなあ、この言葉」「どむかつくめっちゃいいます!」<http://dot.thebbs.jp/1057397440.html>

因此本文退而求其次,对以小说剧本为主的语料库(简称小说)与纯粹报刊语料库(《朝日新闻》,简称报刊)为例,加以比较,考察它们之间的异同。

为了方便比较分别从小说抽取了 2 000 万字,同时从报刊抽取了 2 000 万字,两个语料库的时间跨度和年代基本相同,并以动词「をつける」「を付ける」「を附ける」作为关键词,对使用过于个性化的汉字的「を尾ける」和复合动词「駆けつける」中的「つける」等本次不作为观察对象。

笔者从两个语料库中抽出了与动词「つける」共现的名词,并对他们进行了处理,得出的结果如表 2。

表 2 与「つける」共现名词检索结果

	延べ語数	異なり語数
小説	4 102	1 077
新聞	2 451	582

从表 2 可以看出,两个语料库规模和年代相同,而得出的结果却相去甚远,小说语料库无论在动词总数还是在「延べ語数」(名词总量),还是在「異なり語数」(名词种类),小说几乎都是报刊的两倍。在这个意义上,作为语言研究的对象,小说要远胜于报刊。这是因为小说涉及生活的各个方面,可谓包罗万象,而报刊则相对集中在时事、财经、体育、娱乐等方面,因此小说语料库比报刊语料库更具参考价值。

由于语料侧重的范围不同,自然与「つける」共现的高频度名词也不同,表 3 就反映了两者间的差异。

从表 3 可以看出,小说和报刊一致的名词只有 4 个,而且分布很不均匀,小说使用频度最高的是与日常生活紧密相关的名词,而报刊使用频度最高的是政治经济和体育娱乐。而且有些重要信息这