方法工具与日语教学研究丛书

丛书主编 张威

语料库的制作与日语研究

本书是方法工具与日语教学丛书之一。主要介绍如何利用免费软件来制作日语语料库和如何使用语料库。让每位读者拥有自己的语料库，以此进行各自的研究。内容主要包括语料的收集、PDF转换为文本文件的处理技巧、语料库的具体制作技术、语料库的使用、利用语料库研究日语的范例研究等。

于康 著

浙江工商大

方法工具与日语教学研究丛书

丛书主编 张威

语料库的制作与日语研究

本书是方法工具与日语教学丛书之一。主要介绍如何利用免费软件来制作日语语料库和如何使用语料库。让每位读者拥有自己的语料库，以此进行各自的研究。内容主要包括语料的收集、PDF转换为文本文件的处理技巧、语料库的具体制作技术、语料库的使用、利用语料库研究日语的范例研究等。

于康 著

浙江工商大学出版社

图书在版编目(CIP)数据

语料库的制作与日语研究 / 于康著. — 杭州：浙江工商大学出版社，2013.3

（方法工具与日语教学研究丛书 / 张威主编）

ISBN 978-7-81140-706-8

Ⅰ. ①语… Ⅱ. ①于… Ⅲ. ①日语－研究 Ⅳ. ①H36

中国版本图书馆 CIP 数据核字(2013)第 028134 号

语料库的制作与日语研究

于 康 著

责任编辑 罗丁瑞 姚 媛

责任校对 周敏燕

封面设计 王妤驰

责任印制 汪 俊

出版发行 浙江工商大学出版社

（杭州市教工路 198 号 邮政编码 310012）

（E-mail:zjgsupress@163.com）

（网址:http://www.zjgsupress.com）

电话:0571-88904980,88831806(传真)

排 版 杭州朝曦图文设计有限公司

印 刷 浙江云广印业有限公司

开 本 710mm×1000mm 1/16

印 张 12

字 数 235 千

版 印 次 2013 年 3 月第 1 版 2013 年 3 月第 1 次印刷

书 号 ISBN 978-7-81140-706-8

定 价 32.00 元

版权所有 翻印必究 印装差错 负责调换

浙江工商大学出版社营销部邮购电话 0571-88804227

卷 首 语

自上世纪末以来，我国开始注重对高层次外语研究人才的培养，许多大学相继增设了日语专业的硕士点，而且发展速度很快，如今已超过100余所。尤其2011年，教育部和国务院学位委员会又在全国批准设立了20多个外国语言学一级学科博士点，使国内具有日语语言研究方向博士学位授予权的大学由此前的7所一下子增至30余所。这样的发展态势充分表明，中国的日语研究人才的培养工作已经步入了一个新的发展阶段。

研究生阶段的人才培养具备以下三个主要特征：1）强调传授必要的专业理论知识；2）注重指导学生掌握理性思维的方法和学术研究技巧；3）重点培养学生自己动手从事科学研究的能力。但是，就目前我国的现状而言，有许多研究生和年轻教师都很容易在如何选择合适的研究课题、如何收集和处理所需要的语料、如何观察和分析所研究的语言现象、如何发现规则和导出结论等问题上感到非常迷茫和不知所措。这是一个具有普遍性的现象，严重地制约着研究生培养水平的提高。究其原因，我们发现：长期以来，人们往往更多地关注对日语语言本身的各种探讨，却时常忽略了对日语研究的方法和工具也同样应该给予必要的关注和研究。俗话说“磨刀不误砍柴工”。其实，关于方法与工具的研究就好比“磨刀”，它是为更好地完成日语语言研究服务的。如果掌握了一套比较好的研究方法，而且应用得当的话，就可以大大提高日语研究的效率，做到事半功倍。

为了更好地适应目前我国日语教学的发展态势和研究生培养方面的实际需求，我们策划编写了一套《方法工具与日语教学研究丛书》。其中包括《语料库的制作与日语研究》《汉日日汉翻译语料库的制作、应用与翻译教学研究》《加注标记软件的使用与日语、日语偏误和翻译教学研究》《计量统计解析与日语研究》《中国日语教材的计量解析与日语教材编写》《偏误语料库的制作、应用与中国学生的日语偏误研究》和《日语复合动词语义解析与教学方略研究》，并计划于2016年底以前完成各册的撰写工作。本丛书的主要特点是，积极倡导“方法工具”与“实际应用”相结合的研究理念，内容位于国内外的学术前沿，其中涉及两项国家社科基金项目的相关成果（①课题名称：《翻译教学理论、教学体系和教学模式的研究与翻译语料库的建设》；批准号：11BYY013；课题负责人：邱鸣。②课题名称：《日语

复合动词教学方略研究》;批准号:09BYY079;课题负责人:张威),力求通过推出这套具有实际应用价值和普及意义的学术系列著作,为推动我国的日语教学与研究事业做出实质性的贡献。

现在,由日本关西学院大学的于康教授执笔的《语料库的制作与日语研究》一书马上就要与读者见面了。实在可喜可贺。近几年来,于康教授一直致力于日语语料库的制作与应用研究,并为在国内普及相关的知识与方法做出了不懈地努力。此书对于帮助广大的日语研究者、青年教师、日语语言研究方向的硕士和博士研究生以及日语专业的高年级学生了解和掌握制作日语语料库的具体方法,学会如何通过给语料库加注标签的方法对分析日语中的具体现象进行应用性研究,都具有非常重要的指导作用和利用价值。现在,此类研究成果在国内外尚不多见。不仅在国内外的日语研究和日语教学研究中处于前沿位置,而且可以填补相关领域的空白。

我们衷心地希望这套丛书能够有助于我国日语研究的不断深化,为提高我国日语研究的整体水平做出贡献。

张　威

2013 年 3 月

此为试读,需要完整PDF请访问: www.ertongbook.com

前　言

以归纳法为主研究日语，例句是生命线。手工收集的例句有的时候难免出现疏漏，具有一定的局限性。网上的语料库虽然很方便，但同样具有一定的局限性，因为要受到能否上网的限制。因此，自己动手制作语料库就显得十分重要了。

自己动手制作语料库，有两个非常关键的问题不容忽视：一是技术上不能太难，因为并非所有的人都精通电脑；二是不能够只用来查阅例句，因为很多人在面对大量的例句时会不知所措，不知道该如何处理例句。

过去讲授语料库制作时，大都集中在如何制作上，很少涉及该如何活用语料库，该如何使用语料库的功能来解析收集到的例句。这好比只教如何砍柴，而不教如何使用砍柴的工具。

此次中国人民大学的张威教授主编一套研究日语的工具丛书，旨在介绍如何使用砍柴的工具，要我执笔语料库制作一册。虽然我与语料库打了十多年的交道，但仍旧还是一个初学者。只是在学习的过程中积累了一些心得，不知深浅地将这些心得编成讲稿在自己所在的学校讲授多年，又应邀在国内的一些大学做过连续讲座，故斗胆应允。希望投石问路，得到各位专家学者的指教。

本书有两个重点：一是教授没有太多电脑知识的读者也能够学会制作语料库，二是给例句加注标签并使用语料库来统计这些标签以此归纳出规则来。本书的目的不仅在于学会制作自己的语料库，更重要的是在于试图解决面对大量例句不知如何下手的问题。也就是说，如何使用语料库，如何最大限度地发挥语料库的作用是本书写作的主要焦点之一。当然，这里介绍的方法只是最基本的方法之一。如何活用语料库最终还是要靠读者的创造性的智慧，希望读者能够举一反三，活学活用。

于　康

2013 年 3 月

目　录

第1章 语言资源的利用与语料库建设

1.1 制作语料库的目的

语料库指的是储存各种类型文章的大型文字仓库。我们可以根据各种需要对语料库里储存的语言信息进行各种各样的检索，抽取所需要的信息。

过去，收集和查找例句的方法主要有两种：一种方法是制作卡片，然后将卡片按照所需的分类标准进行排列，以供查询。另一种方法是根据某种需要在报刊书籍等中寻找例句，然后将所需的例句抄写在本子上。前者受存放场所和时间的限制，能够收集的信息非常有限，制作和查找方法十分烦琐。后者受研究对象的制约，研究对象发生变化，例句就必须重新寻找，费时费力。二者都很不经济。特别是因为受到容量的限制，所收集的信息涵盖面窄，很难进行量化统计。

为了克服上述两种方法的短处，几十年来，研究者们一直在研制和开发语料库上下功夫。有了语料库就可以不受时间、场所和容量的限制，随时根据各种需要检索到所需的例句，以供学习和研究使用。

语料库大致可以分为两大类：一类是无标签语料库；另一类是有标签语料库。无标签语料库指的是没有对构句成分加注任何标记的语料库，有标签语料库指的是对构句成分加注各类所需信息标记的语料库。

现在，除了一小部分偏误语料库外，大多数语料库都是无标签语料库。这类语料库中，有专业语料库，也有业余语料库。专业语料库指的是有编程专家参与制作并具备各类复杂性检索功能的语料库。业余语料库指的是非编程专家制作，只能进行有限的复杂性检索的语料库。

专业语料库不仅可以抽取某个词汇实际使用的例句，而且还可以抽取各种句式与各类构句成分之间的搭配用法的例句，同时对出处、作者性别、作品的体裁、作品的发表时间等加以限定。业余语料库除了不能自由地抽取句式与各类构句成分之间的搭配用法的例句外，其他功能基本上与专业语料库相同。

学会制作语料库，可以不受时间和上网条件等限制，根据各种需要随时进行

检索，快速获取大量的例句。学会给例句加注标签，制作带标签的语料库，可以从根本上解决面对大量例句而束手无策的问题，提高发现规则的速度和精度。

1.2 现在可供使用的主要语料库

目前日本已正式公开的具有检索功能的主要语料库和近似语料库如下：

①『現代日本語書き言葉均衡コーパス』

日本国立国语研究所研制。可检索11种体裁的信息，共计1亿480万字。可在网上检索，但有容量限制。申请光盘时需要付费。网站的地址如下：
http://www.kotonoha.gr.jp/shonagon

②『太陽コーパス』

日本国立国语研究所研制。该语料库收录了1895年～1925年博文馆出版发行的月刊杂志『太陽』的大部分内容，共计1450万字，作者约一千多人，由博文馆新社发行，价格为9500日元(不含税)。网站的地址如下：
http://www.hakubunkan.co.jp/gengo/taiyoC.html

③『日本語話し言葉コーパス』

日本国立国语研究所、情报通信研究机构、东京工业大学联合研制。共17盘DVD—ROM。包括3302个演讲的声音资料和文字资料，以及词性信息、话语结构信息和说话者信息等。使用时需要购买光盘。网站的地址如下：
http://www.ninjal.ac.jp/products-k/katsudo/seika/corpus/releaseinfo

④『近代女性雑誌コーパス』

田中牧郎、小椋秀树、山口昌也、小木曾智信、笹原宏之、汤浅茂雄研制。收录了1894年～1925年发行的部分女性杂志。共计210万字左右。免费使用，但需要申请光盘。网站的地址如下：
http://www2.ninjal.ac.jp/lrc/index.php?%B6%E1%C2%E5%BD%F7%C0%AD%BB%A8%BB%EF%A5%B3%A1%BC%A5%D1%A5%B9

此为试读，需要完整PDF请访问：www.ertongbook.com

⑤『Webデータに基づく複合動詞用例データベース(開発版)』

国立国语研究所的山口昌也开发研制，现为开发版。专门用来检索和考察日语的复合动词。已收录日语复合动词 3037 个。网上检索，免费使用。网站的地址如下：

http://csd.ninjal.ac.jp/comp/index.php

⑥『青空文庫』

网上电子图书馆。收录作品共计 1 万 1144 个(2012 年 3 月为止)。网上检索，免费使用。网站的地址如下：

http://www.aozora.gr.jp

⑦『新潮文庫の100 冊』

新潮社出版发行，载体为 CD。共收录 100 部名作。现已绝版，但可通过旧书店购买。

⑧『明治の文豪』

新潮社出版发行，载体为 CD。共收录 40 部名作。现已绝版，但可通过旧书店购买。

⑨『大正の文豪』

新潮社出版发行，载体为 CD。共收录 40 部名作。现已绝版，但可通过旧书店购买。

⑩『新潮文庫の絶版 100 冊』

新潮社出版发行，载体为 CD。共收录 100 部名作。现已绝版，但可通过旧书店购买。

⑪『日本語動詞の結合価』

荻野孝野、小林正博、井佐原均研制。三省堂出版发行。约 15 万个例句。是研究日语动词配价的一个经典语料库。5 万 400 日元。网站的地址如下：

http://www.sanseido-publ.co.jp/publ/nihongo_dosi_ketugoka.html

⑫『朝日 DNA～聞蔵～』

朝日新闻社出版发行，收录了从 1879 年创刊号至今，包括日本全国和地方版的早报和晚报，以及杂志『週刊朝日』『AERA』『現代用語事典知恵蔵』。网上检索，收费。

⑬各大报纸网上语料库

除了上述朝日新闻社的『朝日 DNA～聞蔵～』之外，日本各大报纸都建立了自己的语料库。网上检索，收费。

⑭『国会会議録検索システム』

日本国立国会图书馆研制。可检索众议院和参议院各类会议的记录，并具备按照专题和指定发言者进行检索的功能。网上检索，免费使用，网站的地址如下：

http://kokkai.ndl.go.jp

⑮『日本語学習者による日本語作文と，その母語訳との対訳データベース(作文対訳 DB)』

日本国立国语研究所研制，收录了 21 个国家的 1575 篇日语作文，网上检索，免费使用。使用时需要提前登记。网站的地址如下：

http://jpforlife.jp/taiyakudb.html#p1

⑯『寺村誤用集データベース』

日本国立国语研究所研制，可检索非日语母语者学习日语时出现的偏误用法，网上检索，免费使用。网站的地址如下：

http://teramuradb.ninjal.ac.jp/db

⑰『KYコーパス』

以镰田修和山内博之两人的罗马字首字母命名，收录了 90 名汉语、英语、韩语母语者的录音文字记录，免费使用，但需要申请。网站的地址如下：

http://opi.jp/shiryo/ky_corp.html

⑱《中日对译语料库》

北京日本学研究中心研制。共计两千多万字，可在北京日本学研究中心购买《中日对译语料库》的 CD-ROM。

⑲『NINJAL-LWP for BCCWJ』

日本国立国语研究所和 Lago 语言研究所共同开发的网上检索系统。以日本国立国语研究所研制开发的大型语料库『現代日本語書き言葉均衡コーパス』为母体，专门用来检索动词、形容词、名词和其他词汇以及语法标记的搭配关系。网上检索，免费使用。网站的地址如下：

http://nlb.ninjal.ac.jp

除了上述语料库之外，很多研究者也在研发各类语料库。有的已经公开，有的属于有条件的公开或半公开，有的尚未公开。凡是正式公开的，在网上应该是很容易找到的。

1.3 制作自己的语料库

对在国内学习和工作的人来说，要想得到上述语料库可能不是一件易事。几经周折或得到光盘后，又会出现电脑不兼容的现象。由于无法请教相关人士得到解决问题的办法，只好弃而舍之。特别是网上使用的语料库，要受到各种条件的限制。如果自己拥有一个不受时间和地点等限制的语料库，就可以随时检索，使学习和研究达到事半功倍的效果。因此，制作自己的语料库就至关重要了。

有的人可能会认为，每个人都制作自己的语料库岂不是浪费时间、精力和财力吗？如果有人发扬雷锋精神制作一个可供大家使用的语料库，既节约又省事，还可以腾出更多的时间用于学习和研究。

这个想法是可以理解的。但是，实现起来却并非易事。首先会遇到的是版权问题。其次是个性化语料库的问题。

从理论上讲，版权的有效期通常定为 50 年。但即使过了 50 年期限，作者和版权继承者仍不同意自由使用，那么，在制作语料库时就会遇到障碍。换一个角度来讲，能够公开使用的语料库的语言资料，除了一部分已在网上公开的作品外，大多数也仅限于 50 年前的作品。如果想要制作以现代日语语言资料为主的语料库，仅仅使用这样的语料是不够的。

要解决这个问题只有一个办法，这就是制作个人专用的语料库。如果只是用于自己的学习和研究，不转让、不传播、不发行，就不会违反版权法。

上面说到，语料库可以分为“专业语料库”和“业余语料库”两大类，这是根据检索功能来分类的。如果根据语料库的内容来分类，语料库又可以分为“综合性语料库”和“个性化语料库”两类。

综合性语料库指的是包括各类体裁的语料库。个性化语料库指的是根据专

业对口标准制作的某一类或几类体裁(或作家、作品)的语料库。

例如,研究夏目漱石的人,可以把夏目漱石的所有作品或主要的作品制作成一个专供研究夏目漱石用的个性化语料库。也可以根据研究的需要将夏目漱石的作品和同时代的其他流派或作家的作品制作成一个个性化的语料库。

如果想要研究「ている」与动词「V」的共现情况和表义功能,可以先使用综合性语料库对「Vている」的使用情况进行检索。然后,再将得到的例句制作成一个专供研究「Vている」使用的个性化语料库,以便对「Vている」的使用情况进行集中观察。如果给 V 和其他共现的词汇加注标签,同时对「Vている」的表义功能也加注标签,就可以制成一个带标签语料库。这样不仅可以统计和分析不同类型的动词与「ている」的表义功能之间的关系,还可以统计和分析「Vている」表义功能与其他共现词之间的制约关系。这样就不会面对大量的例句望"洋"兴叹,无从下手了。

不过个性化语料库别人是无法越俎代庖的,只能自己动手制作。

1.4 语料库的多元化用途

讲到语料库,会有不少人认为是用来学习和研究语言用的,似乎与其他领域的研究无关,其实并非如此。

比如,我们将日本历任首相安倍晋三、福田康夫、麻生太郎、鸠山由纪夫、菅直人在国会所做的就职演说制作成一个个性化语料库,并就其中与政治、外交、社会福利、对内和对外政策等相关的用词进行检索和比较(如表 1-1),就可以发现许多属于语言研究之外的课题。表内的数字表示使用次数。

表 1-1 日本历任首相的就职演说用词比较

	安倍	福田	麻生	鸠山	菅
経済	11	9	14	4	38
金融	1	1	2	0	1
危機	0	0	0	0	2
消費税	1	1	0	0	0
改革	16	18	7	2	14
エネルギー	2	1	1	0	1
原子力	0	0	0	0	1
国民	17	23	25	37	29

续表

	安倍	福田	麻生	鸠山	菅
日本	35	5	24	9	8
我が国	12	11	7	0	17
外国	1	0	2	0	0
海外	3	1	0	0	1
アジア	9	8	1	0	11
日米	5	4	4	0	4
中国	1	1	1	0	2
日中	0	0	0	0	0
東南アジア	0	0	0	0	1
アセアン	0	1	0	0	0
高齢化	1	1	0	0	4
少子化	3	1	0	0	1
福祉	0	1	0	1	2
医療	4	4	4	8	4
教育	8	3	2	4	3
観光	0	1	2	0	4
自衛隊	3	1	1	0	0
防衛	0	0	0	0	4
沖縄	1	1	2	0	7
基地	0	0	0	0	5
強い＋NP	0	0	1	1	15
美しい＋NP	9	1	0	0	0
厳しい＋NP	4	3	1	0	1

首相的就职演说指的是就任首相后第一次在国会上公开发表的自己的施政方针。通过对表 1-1 中各类用词的使用频率进行比较，我们可以解读出各位首相的施政重点和各项政策的比重，从而把握各届政府在施政方针和政策上的差异。通过这个小小例子可以知道，语料库不仅可以用在语言的研究上，也可以用在日本政治、经济、社会和文化等方面的研究上。

总而言之，语料库仅仅是一个工具而已，如何灵活运用，将有赖于使用者的智慧和创造性思维。只要做到融会贯通，就可以举一反三，收到“一箭多雕”的效果。

第2章　电脑配置、所需软件及软件的安装

2.1　电脑的配置

语料库通常指的是文字语料库，由于无需处理图像，所以对电脑的配置要求并不是很高。一般情况下，只要满足以下条件，就可以制作语料库了。

① Windows® XP 以上（本文以 Windows®7 为例）

② CPU 为 Atom，或 Celeron，或 Core 皆可

③ 内存为 1GB 以上

当然，配置越高，检索的速度会越快。选择何种配置，读者可以根据自己的具体条件来决定，无需一味求全。

由于制作语料库所需要的软件都是日文版的，所以，使用中文版的 Windows® XP 或 Windows®7 等来制作语料库时，有时候会出现乱码。此时，只要改变语言的设定，基本上就可以解决问题。如果使用的是非正版的 Windows，或是试用版的 Windows，制作语料库的软件大部分不能正常工作。另外，有的时候国内的免费防火墙软件也会影响软件的正常工作。

2.2　所需软件

制作语料库需要 4 个基本软件，它们分别是：

① Java

② 秀丸（ひでまる）

③ えだまめ

④ ひまわり

此为试读，需要完整PDF请访问：www.ertongbook.com