方法工具与日语教学研究丛书

丛书主编 张威

日语偏误研究的方法与实践

本书是“方法工具与日语教学研究丛书”之一，主要介绍了批改作文软件TNR_WritingCorrection2014、标注标签系统TNR_JapaneseErrorCorpusTagger2014、日语偏误语料库TNR_ErrorCorpusCon2014的功能和使用方法，并在此基础上总结了日语偏误的分布、规则与学习难点。

于康 著

浙江工商大学出版社

内容简介

本书是“方法工具与日语教学研究丛书”之一。主要介绍了批改作文软件TNR_WritingCorrection2014、标注标签系统TNR_JapaneseErrorCorpusTagger2014、日语偏误语料库TNR_ErrorCorpusConc2014的功能和使用方法，并在此基础上对偏误用法进行自动统计，总结偏误的分布、规则和学习难点。

方法工具与日语教学研究丛书

丛书主编　张威

日语偏误研究的方法与实践

王康 著

浙江工商大学出版社

图书在版编目(CIP)数据

日语偏误研究的方法与实践 / 于康著. — 杭州 : 浙江工商大学出版社，2018.7

(方法工具与日语教学研究丛书 / 张威主编)

ISBN 978-7-5178-2860-0

Ⅰ. ①日… Ⅱ. ①于… Ⅲ. ①日语－研究 Ⅳ. ①H36

中国版本图书馆 CIP 数据核字(2018)第 154691 号

日语偏误研究的方法与实践

于 康 著

责任编辑 姚 媛
封面设计 林朦朦
责任印制 包建辉
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 虎彩印艺股份有限公司
开 本 710mm×1000mm 1/16
印 张 58
字 数 1126 千
版 印 次 2018 年 7 月第 1 版 2018 年 7 月第 1 次印刷
书 号 ISBN 978-7-5178-2860-0
定 价 168.00 元(全 4 册)

卷 首 语

进入21世纪以来，国内外学者在语料库的研发与建设方面都取得了不少成就，这个领域正在发生着巨大的变化，而且已经逐渐成为语言学研究的一个必然发展趋势。

继《语料库的制作与日语研究》和《加注标签软件与日语研究》之后，“方法工具与日语教学研究丛书”的第三部著作《日语偏误研究的方法与实践》又要与读者见面了。这部著作是于康教授及其研究团队近年来在推动加注标签语料库研发与日语语言偏误研究相结合的方面所取得的又一个标志性成果。在这部著作里，作者详细地介绍了如何使用批改作文软件 TNR_WritingCorrection2014、标注标签系统 TNR_JapaneseErrorCorpusTagger2014 和日语偏误语料库 TNR_ErrorCorpusConc2014 这3个配套组合软件批改作文；如何将批改结果自动转换为正误标签，然后利用标签一览表标注研究用标签，继而自动识别标签；如何进行各类复杂性检索和对偏误用法的自动统计；如何显示偏误的分布、规则和学习难点。以上3个配套组合软件的研发和运用，无论在理念、方法还是应用规模上，都将给传统的日语偏误研究带来前所未有的变化和发展。而且随着大规模带标签学习者语料知识库的建设与推广，人们将会更加科学地从偏误的机制去观察语言，从而实现更好地从偏误研究的角度来阐明语言的生成机制，并以此探究语言之间的共有规则和非共有规则的目标。

我们相信，以上3个配套组合软件和书中所倡导的学术理念、研究方法和应用技术，将为推动我国日语偏误研究的不断深化发展发挥不可忽视的作用。

张 威

2014年10月

前　言

写论文有两个目的，一个是摆事实，一个是讲道理。摆事实指的是发现并阐明一个别人没有发现的事实；讲道理指的是从理论上解释这些事实。描写研究大多应该属于摆事实，理论研究大多应该属于讲道理。所以，论文只要能够摆清事实，就足以看作是独具特色的研究。

摆事实指的是从大量的现象中去抽取能产性高、经得起验证的规则。抽取规则有两个方法，一个是依靠目视或自省，一个是利用软件系统。如今，日语本体研究已经渐入佳境，而汉日对比研究，特别是日语偏误研究还停留在手工收集例句、目视抽取规则的阶段。因此，利用软件系统统计偏误的分布，以此阐明偏误的规则和学习的难点，可以说是一块尚待开垦的处女地。

本书提供并介绍的批改作文软件 TNR_WritingCorrection2014、标注标签系统 TNR_JapaneseErrorCorpusTagger2014、日语偏误语料库 TNR_ErrorCorpus-Conc2014 这三个软件系统为一条龙服务：批改作文，将批改结果自动转换为正误标签，利用标签一览表标注研究用标签，自动识别标签，进行各类复杂性检索，对偏误用法进行各类自动统计，显示偏误的分布、规则和学习难点。

研究需要切磋，切磋需要切磋者居于同一个平台，这样才能保证切磋的平等性。也就是说，大家共享相同的语料，共享相同的研究工具，这样才能做到平等条件下的切磋。我们公开这些工具就是为了提供一个平等的研究平台，希望这些工具能够有利于大家的日语教学和研究。

于　康

2014 年 10 月

重要声明

（此声明适用于本套丛书所有书籍）

软件下载地址：

http://www.zjgsupress.com/

目　录

第1章 日语偏误研究的步骤与工具

重要提示：

1. 语言研究大致可以分为语言本体研究、语言对比研究、语言习得研究3大角度。
2. 偏误研究是一个非常重要且不可轻视的研究领域。这个领域的研究不仅有助于语言本体的研究，还可以促进第二语言习得的研究。特别是可以用来阐明偏误反复出现及石化的过程和机制。
3. 批改作文软件 TNR_WritingCorrection2014 可以用来批改作文，将批改结果自动转换为正误标签，并自动保存为文本文件。
4. 标注标签系统 TNR_JapaneseErrorCorpusTagger2014 可以用来标注研究用的标签，自备标签一览表，并可以自动识别标签，确认标签的定义、主要用法和常用词语。此外，还可以自制标签等。
5. 日语偏误语料库 TNR_ErrorCorpusConc2014 可以进行各类复杂性检索，并能够按照研究的需要显示各种形式的检索结果，还能对标签进行各类自动统计。

1.1 语言研究的3个角度

语言研究大致可以分为3个角度：

①语言本体研究
②语言对比研究
③语言习得研究

第1个角度是语言本体的研究。比如从汉语的角度研究汉语、从日语的角度研究日语、从英语的角度研究英语等。第2个角度是语言的对比研究。比如汉日语言对比研究、汉英语言对比研究、汉韩语言对比研究等，通过与其他语言的对比来研究语言。第3个角度是语言的习得研究。从语言的习得过程和习得机制来研究语言的形成和语言的获得。语言习得研究中的语言偏误研究与语言教学和语言研究直接相关，是十分重要的基础研究。

语言的偏误研究还可以分为2个小类，即母语的偏误研究和非母语的偏误研究。母语为汉语者的汉语偏误研究为母语的偏误研究，而母语为汉语者的日语偏误研究为非母语的偏误研究。

语言本体研究和语言对比研究一直是语言研究的热点，且研究成果丰硕。特别是语言对比研究，尽管这个研究领域刚刚步入成长期，但是，有的研究成果已经显示出语言对比研究不仅可以发现语言本体研究无法发现的问题，还可以解决语言本体研究无法解决的问题。

语言本体研究是从语言自身的角度来观察语言，语言对比研究是从2种及以上的语言角度来观察语言：这2个角度对语言研究来讲都非常重要。但是，语言研究中还有一个不容忽视的角度，这就是偏误研究的角度。语言的偏误研究不仅可以从偏误的机制中去观察语言，即通过对最易出现的偏误和为何出现偏误的研究来阐明语言的生成机制，同时还可以在探索石化(fossilization)途径的过程中，揭示个体石化(individual fossilization)和群体石化(group fossilization)，以及暂时性石化(temporary fossilization)和永久性石化(permanent fossilization)的真实面貌，以此阐明语言之间的共有规则和非共有规则。

1.2 偏误研究的3个步骤

如果要进行偏误语言研究，从偏误语料的积累到偏误的研究，至少需要以下3个步骤：

①收集偏误语料
②给偏误语料标注供统计和分析用的标签
③统计和解析标签，以此阐明语言偏误的规律

没有偏误语料等于无米之炊；有了偏误语料，没有供统计和解析用的标签，结果还是得依赖目视来观察偏误语料，很难抽取出具有普遍意义的规则；有了标注标签的偏误语料，如果没有专门用来统计和解析语料的工具，等于有了材料却没锅炒菜。所以说，这3个步骤环环相扣，缺一不可。

偏误语料绝大多数来自批改后的学生作文。但是，迄今为止，批改作文依旧还是依靠笔头作业，即便是使用word的修订和批注功能来批改作业，其结果也如同手改一样，批改后的作文一旦还给学生，就如同泼出去的水一样一去不返。如果批改者不是有意识地收集批改后的句子，通常都会将非常宝贵的偏误语料白白浪费掉，令人扼腕痛惜。而且，个人收集的偏误例句，量少，有一定的偏向性，所

以，这些针对个别性偏误语料的分析，充其量只能算作个案分析，很难得出具有普遍意义的结论。

收集偏误语料虽不容易，但是只要花时间、下功夫，日积月累，还是可以积少成多的。不过，即便收集到了相当数量的偏误语料，如同使用日语语料库收集日语语料一样，如果不能有效地处理这些大量的语料，换句话说，如果没有一个能够处理大量偏误语料的有效方法，其结果只能是面对宝贵的素材而束手无策，望洋兴叹。

处理偏误语料的最大目的是从偏误语料中确定偏误的倾向和规律，以此阐明偏误的事实与学习难点的真正所在。要做到这一点，就需要对偏误语料进行统计和解析。要进行统计和解析，就需要有统计和解析的对象，而这个对象就是标签。以往标注标签有 2 个方法：一是手工标注，二是以英文字母为代号标注。手工标注太费时，代号标注难以辨认且缺乏直观性。因此，需要有一个既省时、直观性又强的标注标签的系统。

有了标注标签的偏误语料还尚未完整。统计和解析标签是不能仅凭目视和大脑来进行的，需要有一个能够处理大规模标签数据并能够对这些数据进行统计和解析的系统。也就是说，有了偏误语料和对偏误语料标注标签，然后又有一个对标签数据进行统计和解析的系统，这样才能算是一个良好的研究平台。

然而非常遗憾的是，日语偏误研究的现状可以说是极其不容乐观。批改作文依靠笔头作业，收集偏误语料依靠个人行为，标注标签依靠手工，统计和解析偏误语料依靠研究者的主观判断。也许正是由于这个原因，批改作文成为重负，日语偏误研究也一直滞后。

1.3 偏误研究所需的工具

要解决日语偏误研究的这种现状，就需要铺路搭桥。所谓铺路搭桥，指的是提供能够解决问题的有效工具和方法，这些工具和方法必须简便、容易操作才行。也就是说，需要一个批改作文的软件、一个标注标签的系统，以及一个带统计和解析功能的偏误语料库。

为了推动日语偏误研究的发展，近几年来，我们在研究和开发批改作文软件、标注标签系统和日语偏误语料库上花了一些工夫。

为了做到有的放矢，在进行研究开发之前，我们对偏误研究的需求进行了一些调查。结果如下：

①带保存功能的批改作文的软件

②批改作文的软件能够将批改的结果自动转换为正误标签
③标注研究用标签的软件
④标注的标签必须是文字标签而不是英文符号
⑤提供标签一览表，点击一览表中的标签便能够标注标签，而无须手动键入标签
⑥读者可以在现有标签一览表的任意位置补充标签
⑦读者可以自制标签一览表
⑧可以随时根据需要确认标签的定义和典型用法
⑨可以自动识别标签
⑩读者可以自制供自动识别用的标签一览表
⑪具备各种自由组合式检索功能
⑫具备各种统计功能
⑬具备自动转换成 XML 格式的功能
⑭全自动批改作文和标注研究用标签

在上述 14 个需求中，第 14 项需求是一个理想化的目标，就目前的技术来讲还很难实现，特别是全自动批改作文。正确说法往往未必只有一种，而且，有的时候还需要根据文章的上下文和语境来判断正误，就如同翻译软件一样，要达到全部机改，恐怕还需要几代人的努力。与此相比，其他 13 项需求基本可以实现。于是我们以前 13 项需求为基础，研究和开发了以下 3 个软件：

①批改作文软件 TNR_WritingCorrection2014
②标注标签系统 TNR_JapaneseErrorCorpusTagger2014
③日语偏误语料库 TNR_ErrorCorpusConc2014

1.4 3 个工具的主要功能

1.4.1 批改作文软件 TNR_WritingCorrection2014

批改作文软件 TNR_WritingCorrection2014 主要具备以下 3 个功能：

①批改作文的功能
②一次性将批改的作文结果自动转换为正误标签的功能

③保存批改结果并自动生成文本文件的功能

1.4.1.1 批改作文

启动批改作文软件 TNR_WritingCorrection2014 后，出现主画面。如图 1-1 所示，主画面有左、右 2 个区域，左侧区域用来显示需要批改的作文，右侧区域用来键入批改结果。

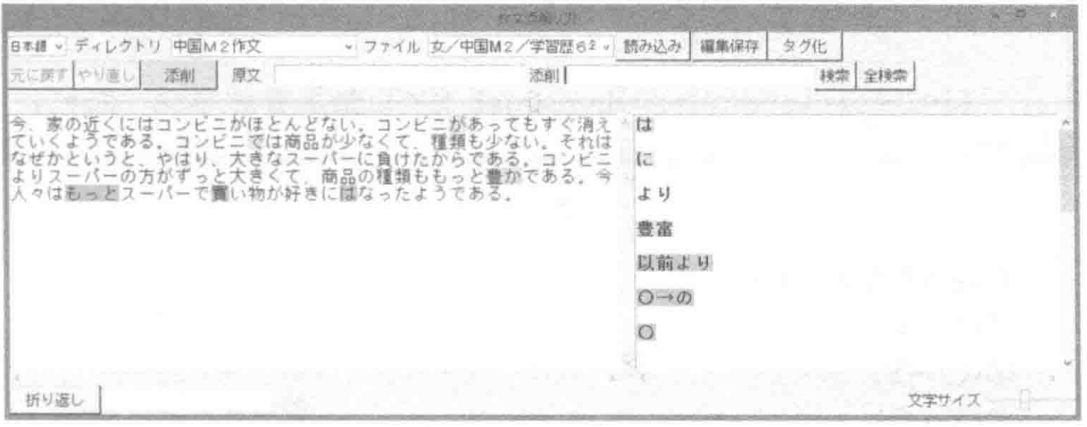

图 1-1

1.4.1.2 正误标签的自动转换

批改完作文后，只要点击图 1-2 画面中右上方的「タグ化」，所有的批改结果就会一次性自动地转换为〈X→Y〉形式的正误标签。X 表示错误的用法，Y 表示正确的用法。如图 1-3。

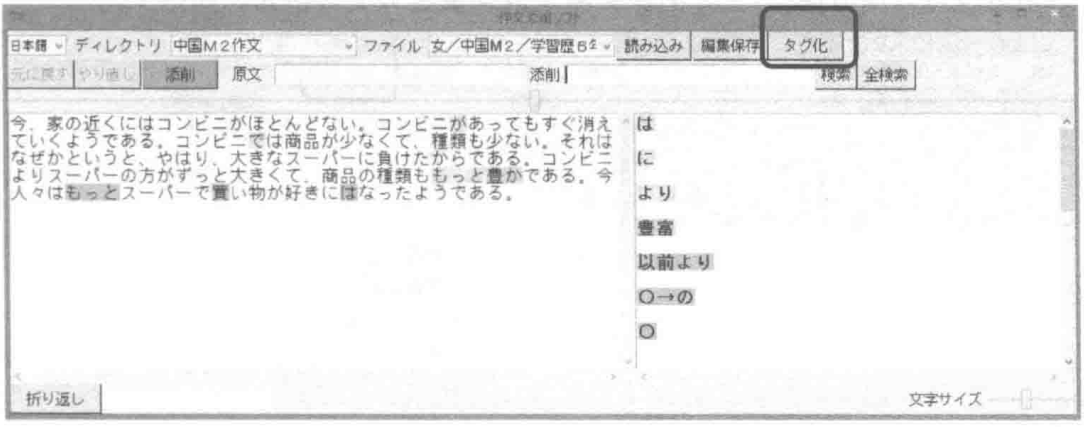

图 1-2

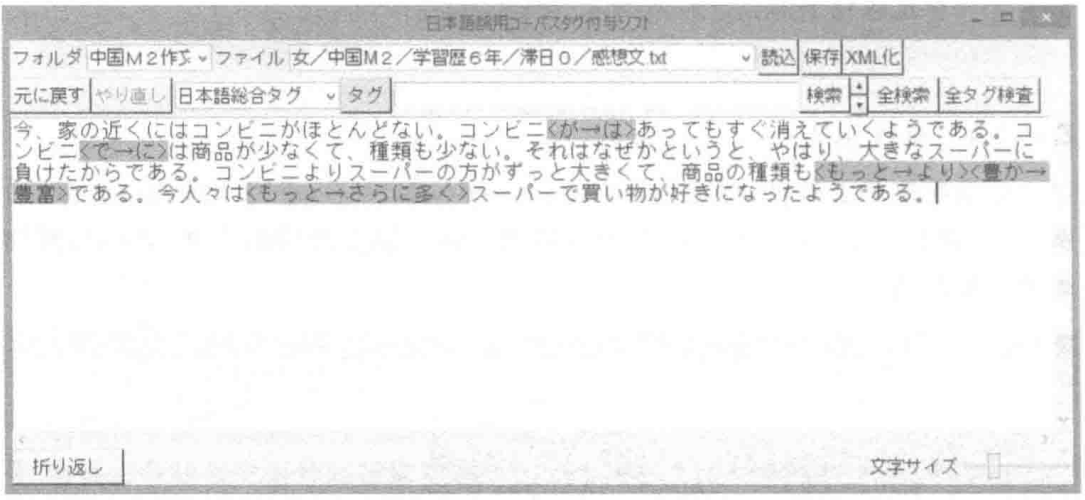

图 1-3

正误标签有 3 种形式：

①〈X→Y〉

②〈○→Y〉

③〈X→○〉

〈X→Y〉表示该用 Y 的地方却用了 X；〈○→Y〉表示该用 Y 的地方却什么都没有用；〈X→○〉表示什么都不该用的地方却用了 X。

1.4.1.3 自动保存为文本文件

批改完作文后，点击图 1-4 画面中右上方的「編集保存」，即可保存文件，批改的中途也可以随时保存，文件的保存格式自动设定为文本格式。如图 1-5。

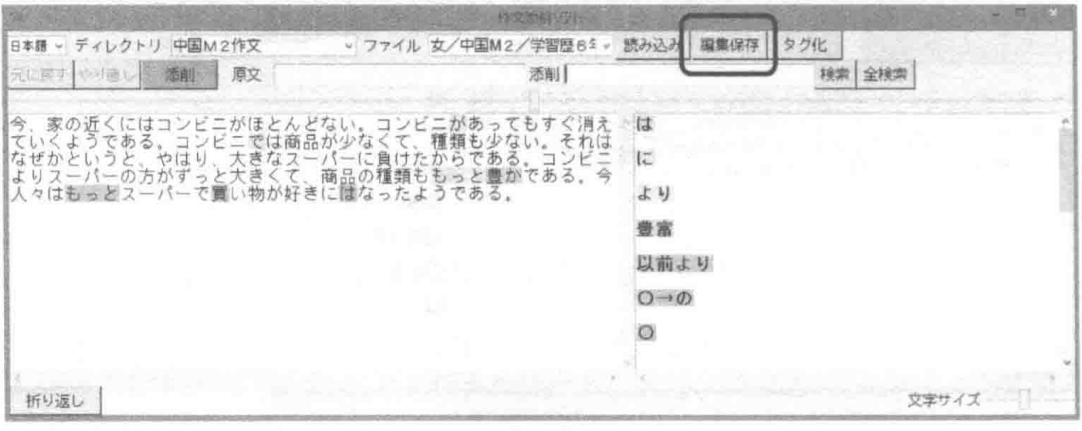

图 1-4