国家社会科学基金课题成果

中日对译语料库的研制与应用研究

论文集

徐一平 曹大峰 主编

外语教学与研究出版社

国家社会科学基金课题成果

中日对译语料库的研制与应用研究

论　文　集

外语教学与研究出版社

(京)新登字 155 号

图书在版编目(CIP)数据

中日对译语料库的研制与应用研究论文集/徐一平,曹大峰主编. －北京：外语教学与研究出版社,2002

ISBN 7－5600－2875－6

Ⅰ. 中… Ⅱ. ①徐… ②曹… Ⅲ. 日语－词语－翻译－研究－文集－日、汉 Ⅳ. H365.9

中国版本图书馆 CIP 数据核字(2002)第 041424 号

中日对译语料库的研制与应用研究论文集

徐一平 曹大峰 **主编**

* * *

责任编辑：张 溥

出版发行：外语教学与研究出版社

社 址：北京市西三环北路 19 号 (100089)

网 址：http://www.fltrp.com.cn

印 刷：北京外国语大学印刷厂

开 本：850×1168 1/32

印 张：15.75

版 次：2002 年 9 月第 1 版 2002 年 9 月第 1 次印刷

书 号：ISBN 7－5600－2875－6/H·1505

定 价：21.90 元

* * *

前　言

2002 年对于北京日本学研究中心来说，是非常重要的一年。首先，从中日两国关系而言，2002 年是中日邦交正常化 30 周年这样一个值得纪念的年份。作为中日两国政府合作的教学、科研机构，这个纪念同样具有十分重要的意义。北京日本学研究中心自 1985 年成立以来，经过中日双方的精诚合作，我们摸索出了一条共同培养人才的道路。由于合作的成功，北京日本学研究中心多次受到双方政府的表彰。也正是因为双方合作的成功，1999 年，当时的日本首相小渊惠三访问中国时，决定以日本政府无偿援助的形式，为支持北京日本学研究中心扩建，援建一座新的教学科研大楼，而这座大楼也将于今年 9 月底封顶。届时，北京日本学研究中心将举行题为“日本学研究的深化与拓展”的大型国际学术研讨会。我们相信，新楼的建成肯定会为北京日本学研究中心的发展创造更加优越的条件。我们也一定决不辜负全国日本学研究界对我们中心寄予的厚望，一定努力将北京日本学研究中心建设成为我国研究日本、培养日本学研究高级人才和中日两国学者进行学术交流的重要基地。

另外，2002 年对于北京日本学研究中心语言研究室来说，还有一层特殊的意义。那就是，我们从 1998 年起开展的一个大型研究项目《中日对译语料库的研制与应用研究》已经于去年年底基本完成，并在今年将其研究成果公诸于世。这一项目的形成要追朔到 1996 年。当年为了促进计算机在中国日语教学和日本语研究当中的应用，北京日本学研究中心举办了一个题为“计算机与日

本语研究”的国际学术研讨会。现在回忆起来，那次研讨会具有非常大的启蒙意义。会议结束以后，后来成为本课题组的主要成员聚集在一起，大家讨论着研讨会上没有尽兴的主题。大家一致认为，要想在今后的日本语研究中导入计算机的应用，必须形成一个具有一定规模的研究课题，而且这个研究课题必须从基础研究着手。渐渐地大家就提出了这样一个研制中日对译语料库的课题。从那以后，我们经过了一年多的准备，终于于1998年《中日对译语料库的研制与应用研究》项目正式立项。这一项目不仅立即得到了日本国际交流基金的大力支持，同年我们还申报了国家社会科学基金项目，并且得到了批准。2000年北京外国语大学成立中国外语教育研究中心时，也对此项目非常重视，项目负责人徐一平被聘为该中心的兼职研究员。

三年来，在全体课题组成员的共同努力之下，我们终于基本实现了当初的设计目标，研制完成了一个约2000万字的大型中日对译语料库。大家知道，在计算机事业飞速发展的今天，单语语料库已经随处可见，但是双语平行语料库由于它的特殊性，尚未得到大力的开发。尤其是中日对译语料库更是如此。在研制过程中，课题组成员们克服了中日双语种同窗显示、语料对齐、双语检索、语法标注等种种技术上的难关，终于研制出了一个比较令人满意的中日对译语料库。我们课题组成员中，除了北京日本学研究中心的研究人员以外，有来自洛阳外国语学院、上海外国语大学、北京第二外国语学院、北京大学、广东外语外贸大学等全国各地的大专院校的教学科研人员。在开发研制的过程中，更有福建师范大学、对外经济贸易大学、国际关系学院、南京师范大学、译林出版社、日本国立国语研究所、京都橘女子大学、东京外国语大学、和光女子大学、横浜国立大学、关西学院大学、立教大学等的教学科研人员分别以研究合作者或客座研究员的身份

参与和承担了我们的研制工作。北京日本学研究中心的硕士、博士研究生课程中的学生也积极地投入到了开发研制的过程中。同时，在检索软件、语料标注等技术研究方面，我们还得到了日本日立中央研究所、奈良先端科学技术大学院大学、日本大学、北京大学计算语言研究所等的大力支持。所以说，我们的这个语料库应该是中日两国科研人员大家努力协作的结晶。在此我们代表课题组全体成员，向所有参与和支持过我们项目的人，表示衷心的感谢。

本研究项目的预期成果是，一个约有 2000 万字的可检索的中日对译语料库和一部反映研制开发过程的论文集《中日对译语料库的研制与应用研究》。本论文集当中收入了各种研究论文共 32 篇。其中有直接涉及研制开发的论文，也有在语料库开发研制过程中，使用中期成果进行应用性语言研究的论文。特别是在 2001 年 9 月，于北京日本学研究中心召开题为“中日对译语料库的研制与应用研究”的国际学术研讨会时，来自中国、日本的学者围绕语料库的问题发表了许多最新的研究成果，其论文的大部分也都收入了本论文集。另外还有一些运用其他语料库进行研究的成果。我们收入这些论文，其目的就是为了能为今后语料库研究与语言研究的结合提供一个参考的基础。我们相信，今后基于语料库的语言学研究一定会有更加广阔的前景。但是，出于对知识产权问题的慎重考虑，我们的中日对译语料库的全部数据目前暂不能完全公开，此次只能将演示盘作为附录提供给大家，如果有在研究上需要和对此感兴趣的人，请与我们课题组联系。我们也希望有更多的人参与到我们今后的研究课题中来。在这里，我们也高兴地告诉大家，我们的语料库自从研制一开始，就受到了广泛的关注，现在韩国的国家语料库课题组也对我们的语料库产生了浓厚的兴趣，并提出一起开发中、日、韩多语种平行语料库。我

们也有计划开发研制中、日、英多语种平行语料库。

总之，我们的研究工作可以说刚刚迈出了第一步，我们热切地希望学界同仁继续给与我们大力的支持，我们也愿为我国的日本语研究事业和计算机语料库的研制继续贡献我们的微薄之力。

徐一平（中国外语教育研究中心兼职研究员）
曹大峰

2002 年 4 月于北京

中日对译语料库简介

《中日对译语料库》是北京日本学研究中心中日合作科研项目“中日对译语料库的研制与应用研究”的基础主干成果。本语料库于 1997 年开始规划设计，1999 年列入国家社会科学基金项目（99BYY007）后，经过 3 年多的研制工作，现已基本完成预定的目标。

1. 规模与性能

经过中日研究人员合作攻关，本语料库达到了下列规模和性能：

a. 收入多种体裁的现代日语和汉语的真实文本及其对译文本达 2000 万字规模；

b. 实现了原内码语料组库、跨平台日汉双语检索和平行显示的功能；

c. 语料全部实现了句段对齐和词性标注，其中 10%进行了句法标注尝试。

上述规模和性能的实现，从先进性和通用性的角度均填补了国内外同类研究的空白。本语料库不仅可供一般中日语言学习、语言及翻译研究等多种目的的应用，而且为日汉双语信息处理研究和二次性开发提供了重要的基础资料，因此，可以说在应用方面也具有重要意义。

2. 语料内容

考虑到语料的多样性、时间性和完整性，本语料库经过译本

调查和专家筛选共收入下列文章全文80篇,体裁包括近现代各时期的小说、诗歌、散文、传记、政论、法律条约等。

中国文章

家(上)(巴金)　上海的早晨(1)(周而复)　呐喊·彷徨(鲁迅)　骆驼祥子(老舍)　霜叶红似二月花(茅盾)　青春之歌(杨沫)　小鲍庄(王安忆)　金光大道(浩然)　天云山传记(鲁彦周)　活动变人形(王蒙)　关于女人(冰心)　盖棺(陈建功)　钟鼓楼(刘心武)　插队的故事(史铁生)　轮椅上的梦(张海迪)　人到中年(谌容)　人啊人(戴厚英)　红高粱(莫言)　棋王(阿城)　倾城之恋(张爱玲)

政府工作报告 96-99　我的父亲邓小平(毛毛)1-2　邓小平文选 1-3　网上飞鸿(人民日报·朝日新闻)　毛泽东选集 1-4　毛泽东传(金冲及)　中日建交联合声明　中日和平友好条约

日本文章

蒲団(田山花袋)　坊ちゃん(夏目漱石)　羅生門·鼻(芥川龍之介)　砂の女(安部公房)　青春の蹉跌(石川達三)　黒い雨(井伏鱒二)　高野聖(泉鏡花)　あした来る人(井上靖)　沈黙(遠藤周作)　野火(大岡昇平)　死者の奢り·飼育(大江健三郎)　雪国(川端康成)　破戒(島崎藤村)　痴人の愛(谷崎潤一郎)　斜陽(太宰治)　こころ(夏目漱石)　金閣寺(三島由紀夫)　雁の寺·越前竹人形(水上勉)　友情(武者小路実篤)　ノルウェーの森(村上春樹)

日本戦後百家詩集(罗兴典)　百言百話(谷沢永一)　ひとりっ子の育て方(中澤次郎　鈴木芳正)　激動な百年史(吉田茂)　日本経済の飛躍的な発展(大谷健)　ビジネスマンのための「心の危機管理術」(岡本常男)　近代作家入門(松澤信祐)　マッテオリッチ伝(平川祐弘)　日本列島改造論(田中角栄)　日本国憲法　サラダ日記(俵万智)　タテ社会の

人間関係(中根千枝)　適応の条件(中根千枝)　五体不満足(乙武洋匡)

3. 操作环境

为了便于中国、日本及其他国家的使用者使用，本语料库实现了在中日英三种文字版本的 WINDOWS95/98/ME/2000/XP 操作系统上使用的跨平台功能，使用者只需利用微软公司的 OFFICE2000（含 Access）办公套件及其免费提供的 Global IME 中文套件或日文套件即可实现中日双语检索和显示。具体操作环境如下：

CPU：推荐 Pentium3 以上机种

操作系统：WINDOWS95/98/ME/2000/XP（中日英三种文字版本）

办公套件：Microsoft Office2000（含 Access）

双语环境：Microsoft Global IME 中文套件或日文套件

4. 版本及使用方法

本语料库分为研制版和应用版两个版本，研制版包括本语料库的所有语料（原语料、对齐加工语料、词性标注语料、句法树加工语料等）和有关软件包（检索软件、抽样软件、专用词典、词性标注软件、文本整理软件等）等本语料库的全部研制成果，仅供项目开发研制和项目验收用。应用版是在本语料库全部语料的基础上经加工对齐后建立的具有双语检索和显示功能的光盘数据库，它以一般语言研究者为使用对象，在知识产权问题未全部解决以前暂且以小范围试用和征求意见的目的内部发行。关于应用版光盘的使用方法详见光盘内的说明文件（readmej. doc）。

5. 版权限制

根据中日两国有关知识产权的法规规定，本语料库只能在有限的范围内用于研究者个人的科学研究，本语料库及其收录的语料只能通过检索部分地利用，不得整体复制。使用者须与研制开

发者签署使用协议，承诺遵守知识产权法规并承担违规时的一切责任，不得任意复制转让他人使用，更不能用于商业目的。另外，希望使用者将使用时发现的问题及时地反馈给我们，以不断改进语料库的质量。

6．研制人员

项目负责人

徐一平　北京日本学研究中心

项目顾问

冯志伟　国家语言文字工作委员会

严安生　北京日本学研究中心

项目主要成员

曹大峰　北京日本学研究中心

施建军　洛阳外国语大学

戴宝玉　上海外国语大学

李　强　北京大学外国语学院

潘寿君　北京第二外国语大学

杨诎人　广东外语外贸大学

林　璋　福建师范大学

姚莉萍　对外经济贸易大学

王　信　国际关系学院

张　兴　北京日本学研究中心

周　浩　南京师范大学

项目合作人

中野洋　国立国语研究所

宫岛达夫　京都橘女子大学教授

平井和之　东京外国语大学

加藤三由纪　和光大学

村田忠禧　横浜国立大学

松本裕治　奈良先端科学技术大学院大学

河野胜也　日本大学

小田切文洋　日本大学

于　康　关西学院大学

王学群　立教大学

资金援助

(中国)国家社会科学基金

(日本)国際交流基金

技术合作

(日本)日立製作所中央研究所

北京日本学研究中心

2002.6.26

目　录

一、中日对译语料库的设计与研制

二、中日对译语料库的应用研究

三、相关研究

关于《中日对译语料库》的研制和应用研究

北京日本学研究中心　徐一平

计算机技术的飞跃发展为科学研究带来了一场革命。今天，计算机语料库的建设已成为语言研究现代化的重要内容之一，语料库语言学被作为语言研究的主流受到重视。在欧美，继著名的BROWN语料库和COBUILD大型语料库之后，上万亿字符的特大型语料库和监控语料库正在研制。在我国，由国家语委主持研制的现代汉语语料库已达7000万字，《人民日报》光盘数据库收集了该报46年的全部文字和图像内容并公开发行。在日本，由“新情报处理开发机构（RWCP）”研制并公布的RWC语料库收录了《每日新闻》报纸4年的全文信息，其语素标注量达1亿条。随着信息技术的进步和语言学的发展，语料库的研制正继续向着大规模、细标注和多功能的方向发展，而国际化的潮流和因特网的普及又要求我们对双语或多语语料库加大发展力度并加强其应用研究。

目前，世界上的双（多）语语料库均为英语与其他语言的平行语料库，汉语与日语的平行语料库尚未见开发。随着信息社会的到来和中日交流的不断扩大，越来越多的语言研究者、教育者以及机器翻译研究者急需一个大型的中日平行语料库。为此，北京日本学研究中心组织中日两国专家学者从1996年10月开始酝酿并经过反复调查论证，于1998年9月决定着手开发研制《中日

对译语料库》以填补这一空白。第一期工程拟录入2 000万字的对译文本，建立一个带有基本检索功能的文本数据库；第二期工程，对已录入语料进行平行对齐和语法标注等深加工；第三期工程，在第二期工程的基础上开发一个适合语言研究的多种功能检索系统。

本课题以研制一个具有千万字级规模的多用途中日对译语料库为基本目的，并将语料加工和应用研究贯穿于研制的全过程。拟研究的主要问题及其重点和难点为：

一、语料的选材、录入和校对

首先设计一个中日双语平行的大型语料库，收录有译文的汉语和日语平行语料2 000万字，为兼顾多种研究目的，收录的内容以有研究价值的中日文学名著为主，兼收剧本、散文、政论文等其他文体的文章，原文和译文全文收录。为满足文学和翻译学研究学者的研究需要，部分名著收录多个译本，语料错误率确保在千分之五以内。

录入阶段，中日文分别在中文版和日文版的windows平台上进行处理，各自使用GB或SHIFTJIS两个不同体系的内码系统建立基础文本数据库，构建对译语料库时再将中日文文本纳入同一个文字代码系统。这个问题的解决是本语料库研制过程中的主要难点之一，将在“四、中日文的兼容处理和检索软件的开发”一节中详细论述。

二、对译语料之间的文本对齐

所谓文本对齐，主要是指以句为基本单位，将原文和译文配对放在一起。文本对齐问题，是关系到双语语料库使用价值的重要问题。由于中日语差别很大，日语表达繁杂，汉语言简意赅，