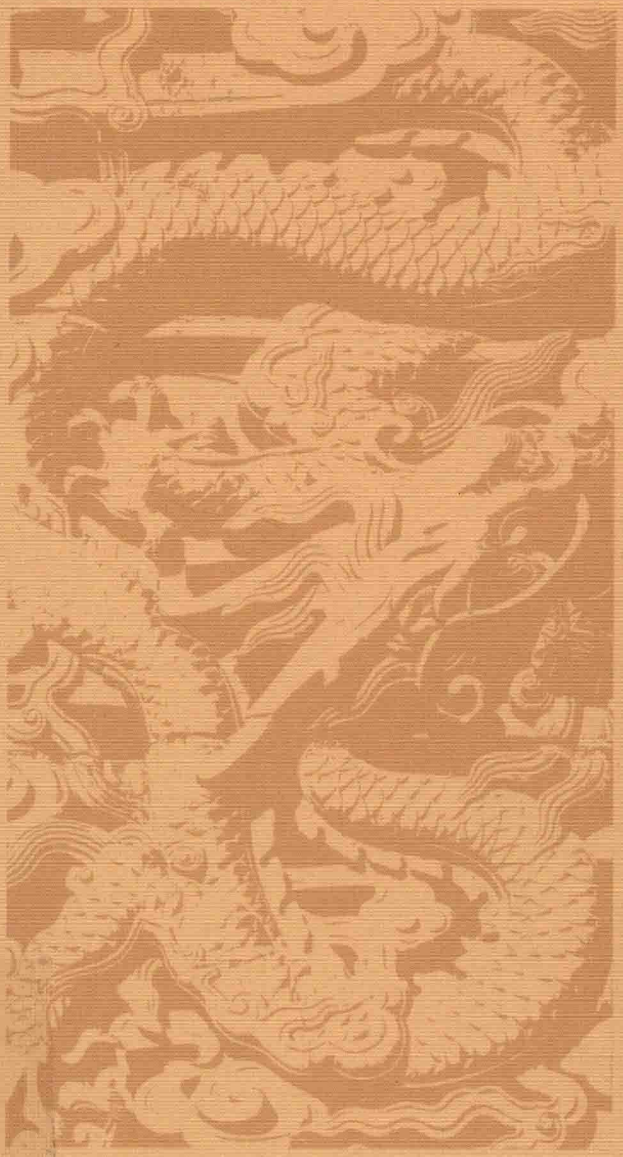


国别化汉语中介语动态语料库建设与研究

胡晓清等◎著

中国社会科学出版社

国别化汉语中介语动态语料库建设与研究

胡晓清等◎著

中国社会科学出版社

图书在版编目(CIP)数据

国别化汉语中介语动态语料库建设与研究 / 胡晓清等著 . —北京：中国社会科学出版社，2018. 12

ISBN 978-7-5203-3767-0

Ⅰ. ①国… Ⅱ. ①胡… Ⅲ. ①汉语-中介语-语料库-研究 Ⅳ. ①H1

中国版本图书馆 CIP 数据核字(2018)第 287828 号

出 版 人 赵剑英
责任编辑 任 明
责任校对 李 剑
责任印制 郝美娜

出 版 中国社会科学出版社
社 址 北京鼓楼西大街甲 158 号
邮 编 100720
网 址 http：//www. csspw. cn
发 行 部 010-84083685
门 市 部 010-84029450
经 销 新华书店及其他书店

印刷装订 北京君升印刷有限公司
版 次 2018 年 12 月第 1 版
印 次 2018 年 12 月第 1 次印刷

开 本 710×1000 1/16
印 张 26.25
插 页 2
字 数 429 千字
定 价 90. 00 元

凡购买中国社会科学出版社图书，如有质量问题请与本社营销中心联系调换
电话：010-84083683

本书作者

（按音序排列）

董婷婷　胡晓清

刘丽媛　王　艳

许小星

目　录

绪论（代序）

语料库（corpus）是以电子形式保存的语言数据库，是语言研究的一种普遍资源。目前国内外已经建成了许多大规模语料库，为语言研究提供了极大便利。国外语料库在四个阶段的发展后与应用紧密结合。自 1959 年英国 Quirk 开始建立“英语用法调查”（SEU）开始，在经历手工语料库、第一代电子语料库和第二代电子语料库阶段后，英语语料库语言学研究范围不断扩大，建库实践不断丰富，库容规模不断拓大。目前，国外语料库除著名的第一代电子语料库 BROWN 语料库、LOB 语料库和 LONDON-LUND 口语语料库，第二代电子语料库 COBUILD 语料库、朗文语料库网外，比较成熟、有特色的还有 AHI 语料库、OTA 牛津文本档案库、BNC 英语国家语料库、ACL/DCI 美国计算语言学学会数据采集计划、LDC 语言数据联合会、RWC 日语语料库、亚洲各语种对译作文语料库等。上述语料库有的进行了详细的韵律标注，有的为辞典编纂而建，有的使用了 TEI 编码和 SGML 通用标准置标语言的国际标准。在使用上，有的实施会员制，会员间共享语料库；有的付费使用；也有的部分开放共享。而美国计算语言学学会倡议的数据采集计划 ACL/DCI 则成为以动态性、流通性为主要特征的第三代语料库的代表。第四代语料库则注重语料的言思情貌整一（顾曰国，2013）[①]，以多模态语料库为建设目标。同时，众多学者基于语料库开展了大规模的研究，特别是对语料库进行了语法标注研究，将自动语法标注的正确率由 77%提高到 99.5%，超过了人工标注所能达到的最高正确率。（冯志伟，2002）[②] 在众多语料库中，朗文语料库

① 顾曰国：《论言思情貌整一原则与鲜活话语研究——多模态语料库语言学方法》，《当代修辞学》2013 年第 6 期。

② 冯志伟：《框架核心语法与自然语言的计算机处理》，《汉语学习》2002 年第 2 期。

网由三大语料库组成，包括朗文/兰开斯特英语语言语料库、朗文口语语料库、朗文学习者英语语料库，形成了一个覆盖极广，分则自成体系、合则可靠互助的库群，是众多语料库中特色较鲜明的一种语料库建构形式，值得我们学习和借鉴。

而国内语料库的建设与研究几十年来进展迅速。国内20世纪20年代开始建设的语料库主要为非机读语料库，自1979年始建可机读语料库。在英语研究领域，建立了JDEST英语专门用途语料库、中国学习者英语语料库（CLEC）、中国学习者英语口语语料库（COLSEC）、中国学生英语口笔语语料库（SWECCL）、香港科技大学学习者语料库（HKUST Learner Corpus）、中国英语专业语料库（CEME）等及一批双语平行语料库。在汉语本体研究领域，最初建立汉语语料库的目的仅是为字频、词频统计服务，代表性的为北京语言大学《汉语频率词典》项目专用语料库。之后建成的台湾中央研究院平衡语料库（Sinica Corpus）、中文五地区共时语料库（LIVAC语料库）、“现代汉语研究语料库”（北京语言文化大学）、北京大学现代汉语语料库、“汉语精加工语料库”以及“面向辞书编纂的大型通用语料库”（鲁东大学中文信息研究所）、《人民日报》语料库、《作家文摘》语料库、中科院现代自然口语语料库、北京语言大学的第三代动态流通语料库、中国传媒大学的广播电视文本语料库、传媒有声语言语料库等现代汉语通用或专用语料库，则已经在语料库语言学和汉语语言学本体研究中发挥了重大作用。而在研的国家社科重大招标项目“汉语国际教育背景下的汉语意合特征研究与大规模知识库和语料库建设”更是将汉语特征研究与语料库建设融为一体，使本体语料库建设更加凸显目标性。

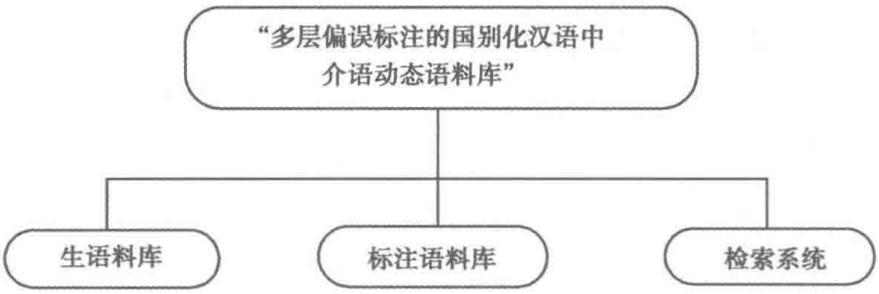
在汉语教学与研究领域，中介语语料库建设与研究卓有成效。1993年北京语言学院开始建设“汉语中介语语料库”，1995年完成，是为国内第一个中介语语料库。该语料库收集了不同国别、不同语言背景、不同学级、不同年龄性别的外国留学生原始语料350万字，其中经过标注加工的熟语料100万字。作为国内第一个大规模汉语中介语语料库，虽然它未完全对外开放，但在汉语语料库的研究、汉语中介语研究和对外汉语教学理论研究上已经发挥了重大的作用。基于该语料库进行的多项研究，为建立和发展汉语作为第二语言的语言学习理论奠定了基础。2008年，北京语言大学又建成了以424万字的HSK高等考试中作文考卷为原始语料的

“HSK 动态作文语料库”，目前已向用户免费开放。国内其他高校如南京师范大学、中山大学、暨南大学、鲁东大学、南京大学等也陆续建设了90 万字到 400 万字语料不等的各类中介语语料库。基于上述语料库开展的一系列研究丰富了中介语语料库建设理论，对汉语词汇语法习得研究、汉语认知研究起到了重要的推动作用。北京语言大学在建的“全球汉语中介语语料库”联合国内外众多高校、机构拟采集全球汉语生语料 5000 万字，加工熟语料 2000 万字，成为超大规模汉语中介语语料库的试水者。同时基于上述中介语语料库进行的中介语语料库建设研究、汉语中介语研究不断走向深入，并进入了汉语中介语语料库建设及中介语研究的宏观理论探索阶段。已建成或在建的汉语中介语语料库建设呈现多元化发展。从用途来看，除通用语料库（general corpus）外，也有为某一特定目的而设计的专门语料库（specific corpus），如美国莱斯大学建立的医患对话语料库和课堂教学语料库；有面向某一国别的国别化语料库，如鲁东大学胡晓清教授团队的“国别化汉语中介语动态语料库”；从语料分布时间来看，有共时性的语料库，也有历时性的语料库；从语料语体来看，有书面语语料库也有口语语料库。在语料库建设的基础上，语料库语言学作为语言研究的一种新的方法与策略，得到了广泛运用。特别是国内外关于语料库建构研究、语料库应用研究已经进入一个厚积薄发的阶段，取得了丰硕的成果。

在语料库建设和语料库语言学不断推进的大背景下，本团队 2011 年申请了国家哲学社会科学项目“国别化汉语中介语动态语料库建设与研究”（项目编号 11BYY050），优秀结项后，又于 2016 年申请了“多维参照的国别化汉语中介语动态语料库库群构建与研究”（项目编号 16BYY108）。在项目进行过程中，围绕国别化汉语中介语语料库的建设和研究，项目组进行了大量的工作，取得了若干成果。本书对“多层偏误标注的国别化汉语中介语动态语料库”的建设及基于该语料库进行的相关研究进行了介绍和总结，既是对前期工作的一个全面梳理，也是对后期工作的一个前瞻和规划。

本书分为上下两篇，上篇是多层偏误标注的国别化汉语中介语动态语料库的建设情况，主要介绍了语料库的整体框架和建设过程。

“多层偏误标注的国别化汉语中介语动态语料库”主要包括以下三个模块：



生语料库中保存有400万字原始语料，语料版本分为图片版和文字版，语料来源于学习者在作业和考试中的作文和造句。

标注语料库中保存有300万字熟语料，主要进行了基础标注层面的机器自动词性赋码、人工纠偏、基本句式标注；偏误标注层面的标点偏误标注、字偏误标注、词偏误标注、句法偏误标注、篇章偏误标注。

检索系统由语料加密软件以及用户检索平台两部分组成，其中，用户检索平台包括检索界面和底层的算法软件包两部分。

下篇是基于多层偏误标注的国别化汉语中介语动态语料库，以汉字层面、词汇层面、句法层面、对外汉语教学方面为对象所开展的一些初步的示例性的研究，还不够系统和完善。在后期的研究中，我们将进一步对总库进行描述，不断深入对不同层面的国别化汉语中介语研究。

上篇

多层偏误标注的国别化汉语中介语动态语料库建设

第一章

多层偏误标注的国别化汉语中介语动态语料库建设的必要性

语料库语言学的研究不但要基于本体语料库，也需要大规模中介语语料库的支撑。目前国内中介语语料库的数量尚满足不了语料库语言学发展的需求。而第二语言习得与汉语教学研究更迫切要求建立规模大、种类全、功能细的汉语中介语语料库。

截至目前，国内已开发的中介语语料库主要有“汉语中介语语料库”、“HSK 动态作文语料库”以及南京师范大学汉语中介语偏误信息库、中山大学中介语语料库、暨南大学留学生语料库等。其中“HSK 动态作文语料库”全开放使用，其他语料库部分开放或封闭使用。很多人包括部分学者曾对单国别汉语中介语语料库的建设有所质疑，认为多国别中介语语料库中已包含的国别其中介语情况可在多国别语料库中检索、提取，与其建设单国别中介语语料库，不如加大多国别语料库的规模。对此，我们有不同意见。

从语料数量来看，目前语料库中韩国留学生中介语语料不够充足。如“汉语中介语语料库”100 万字加工语料中朝鲜语占 15%（陈小荷，1996）[①]，即韩国学生汉语中介语加工语料为 15 万字。其他中介语语料库未见国别抽样具体数据，但只要是平衡语料，韩国留学生语料应不超过 100 万字。（基于“HSK 动态作文语料库”总规模 400 万字、南京师范大学语料库 100 万字、中山大学语料库 100 余万字、暨南大学 400 万字的初步数据信息）如语料再进行程度分级，分布到每个层级的韩国学生中介

① 陈小荷：《“汉语中介语语料库系统”介绍》，《第五届国际汉语教学讨论会论文选》，北京大学出版社 1996 年版，第 9 页。

语语料会更少，这样无法为单国别中介语偏误研究和国别化汉语教学提供足量的语料。因此有必要建设较大规模的针对韩国留学生的国别化汉语中介语语料库。

从语料层级来看，目前规模最大的“HSK 动态作文语料库”采自高级汉语水平考试作文语料，因此，语料均为高级学段作文。“汉语中介语语料库”中 15 万字韩国学生语料若分布到初、中、高三级，每一层级语料量会更少。其他类同。这就使基于中介语语料库进行汉语字、词、句、篇的难度序列研究受到分层级语料数量不足的制约。因此有必要对分层级中介语语料库予以关注。

从语料动态性来看，文中所涉中介语语料库均为动态语料库，但侧重点各有不同。“HSK 动态作文语料库”的动态性偏重于历时的可扩充性，即可随着 HSK 高级考试的逐年进行不断补充新的语料。然而，库中很难收录同一学习者的历年动态语料。“汉语中介语语料库”本意也要对同一学习者不同学习阶段语料进行跟踪收集，以便开展跟踪性调查研究。然而在取样时为了“使核心语料中各种属性的语料分布比较均匀”不得不“损有余而补不足”（陈小荷，1996）[①]，规定同一作者的语料一般最多抽取 4 篇。如此便无法开展学习者个案跟踪研究。要想使中介语语料既能满足面向全体学习者的偏误规律研究需要，同时可展开面向单一学习者的个案跟踪研究，语料库建设中的动态性就要既考虑一般意义的历时动态，也要注意针对部分学时较长，学级跨初、中、高三段的学生，对其进行语料的足量跟踪收集。我们在建的语料库即关注于此。

从语料加工情况来看，各中介语语料库基于不同研究目的和用途加工项目也不尽相同。“汉语中介语语料库”主要进行了文字预处理、断句、分词、词性标注等加工，未对学习者偏误进行标注。其他语料库有的主要进行了句法属性和偏误标注；有的侧重于偏误标注，未进行分词。如果考虑到全面研究的必要进行更多层面的加工标注，语料库会更高效、实用。我们的语料库则试图在语料加工上更加全面、细致。

另外，多国别中介语语料库在语料加工中制定的规范和规则应该是面向所有汉语学习者的普适性规律，为此有时要排除、忽略只影响某一国别

① 陈小荷：《“汉语中介语语料库系统”介绍》，《第五届国际汉语教学讨论会论文选》，北京大学出版社 1996 年版，第 9 页。

的特殊情况。而单国别语料库可根据单一国别语料的实际情况，制定最适合该国别偏误研究的标注规范，避免宝贵的个性化偏误现象湮没于宽泛的规则中。

因此，建设一个规模大、数量充足、层级鲜明、加工细致的单国别动态语料库是非常必要的，也是完全可行的。