昆明学院学报　2018，40（1）：127～132　ISSN 1674－5639　CN 53－1211/G4
Journal of Kunming University　【教育·课程·教学】

中国学习者韩语中介语语料库建设方案

徐中云
（徐州工程学院 外国语学院，江苏 徐州 221000）

摘要：目前我国韩语中介语语料库建设还处于起步阶段，教师在教学中由于缺乏科学可靠的中介语语料库支持，很难对学习者的二语认知特点及语言能力发展规律做出客观准确的判断。在建设科学而有效的韩语中介语语料库时，首先应全面采集共时语料，实时补充历时语料；其次，录入韩语中介语语料并编辑语料背景信息应真实客观；第三，预先设定偏误代码及规范，进行偏误标注；第四，实现语料库的持续动态发展和资源开放共享。

关键词：二语教学；韩语中介语；语料库；动态数据；资源共享；偏误标注
中图分类号：H55　**文献标识码：**A　**文章编号：**1674－5639（2018）01－0127－06
DOI：10.14091/j.cnki.kmxyxb.2018.01.021

The Korean Inter-language Corpus Construction plan of Chinese Learner

XU Zhongyun
（College of Foreign Languages，Xuzhou Institute of Technology，Xuzhou，Jiangsu，China　221000）

Abstract：The support from the scientific and reliable Korean inter-language corpus is not sufficient for language teaching since the corpus construction of the Korean inter-language is still at the initial stage in China and it is difficult to judge objectively and accurately the features of the second-language cognition and developing rules of language ability from the language learners. However，the construction of the scientific and efficient Korean inter-language corpus needs the collection of diachronic corpus and the supplement of the synchronic corpus，the input of the Korean inter-language corpus and the true and objective compilation of the background information，the presetting of the error codes and specifications and the tagging of the errors and then the realization of sustainable diachronic development of the corpus and resource sharing.

Key words：second language teaching；Korean inter-language；corpus；dynamic data；resource sharing；error tagging

一、引言

搜集二语学习者的语言材料并汇集成中介语语料库，基于语料库采用定性或定量的方法研究二语习得过程，是近年来二语教学研究领域的新趋势。随着计算机技术的不断发展，应用于语言学研究的各种语料库逐渐形成，中介语语料库在二语教学和研究中所起的作用日益受到学界的重视。全球范围内对中介语语料库的研究始于1980年代，至今建设成果以英语学习者语料库居多（Granger 1998年、2004年、2007年，Tono 2003年，McEnery&Hardie 2012年）[1]。我国通过建立学习者中介语语料库开展外语教学研究始于20世纪90年代初，现今已建成许多有规模的中国学生英语口笔语语料库（杜诗春、杨惠中2003年，李文中等2004年，王立非、孙晓坤2005年，王立非、文秋芳2007年）。同时，近年来汉语中介语语料库建设发展迅猛，内容涉及书面语语料库和口语语料库、横向共时语料库及纵向历时语料库建设（杨翼等2006年，张宝林2010年，曹贤文2013年，郑通涛、曾小燕2016年）。非通用语种韩语教育教学自20世纪40年代发展至今，全国已有200余所高等院校开设韩语专业，国内韩

收稿日期：2017－06－21
作者简介：徐中云（1979—），女，江苏丰县人，讲师，硕士，主要从事韩国语教育研究。

语教育发展迅速，然而相关理论研究滞后，许多研究仍然是基于经验型的，韩语学习者中介语语料库建设的研究更是少之又少，缺乏科学可靠的中介语语料库支持，教师很难对学习者的二语认知特点及语言能力发展规律做出客观准确的判断。因此建设韩语中介语语料库，基于中介语语料库开展教学研究是亟待进行的研究趋势。

二、韩语中介语语料库建设的基本原则

（一）真实客观性

真实客观的语料反映出二语学习者真实的语言习得状况和语言使用面貌。韩语中介语语料库的建设首先要坚持真实客观的原则，这是后期基于语料库的研究能够得出准确结论的基础和前提。

首先，真实客观性体现在语料的采集方面。语料的真实客观性是指语料必须是由学习韩语的中国学习者在自然状态下自主产出的韩语表达材料。自然状态所指的学习者，应该在轻松自然的氛围下自由的产出客观反映自身语言水平的语料。自主产出是指，不论写出来的文字还是说出来的口语，都是学习者主观思考的产物，而不是抄写或记录别人的话。[2]如主题讨论、即兴问答、课堂写作练习、考试作文等不提供任何参考资料和工具书的中介语语料。

其次，真实客观性体现在语料的形式方面。真实客观的中介语语料应该反映出学习者的韩语综合习得系统。中介语是合法的语言系统，它也是一个由内部要素构成的系统，就是说它有语音的、词汇的、语法的、语用的规则系统，而且自成体系。[3]因此韩语中介语语料库不仅包括书面语料，还应该包含口语语料的搜集建库，全面真实地反映学习者中介语的特性，从而为基于语料库的分析和研究得出有意义、有价值的结论奠定基础。

第三，真实客观性还要体现在语料的录入和标注上。在语料录入阶段应该最大程度地忠实语料原貌，尽量“原样照搬”，包括韩语中介语语料库的单词拼写、短语使用、助词及连接词尾的应用，句子及篇章的组织等，以求真实反映韩语学习者中介语的语言原始面貌。从保持语料真实性的角度来看，语料标注阶段也要尽可能保持中介语语料原貌。即只是指出语料中的偏误现象与偏误类型，而不做任何更改。[4]327

（二）动态发展性

中介语语言系统是学习者随着交际需要和自身努力不断向目标语靠近的动态发展的语言系统，反映出学习者的学习心理发展过程。通过研究韩语学习者的外在言语表达，即中介语语料，有助于研究体现学习者韩语认知规律的“内在大纲”，揭示中国学习者掌握韩语的普遍规律。韩语中介语语料库的建设应该有助于揭示中国学习者韩语习得过程与发展规律为宗旨，为进一步改革韩语教学模式、更新教学方法、提高韩语教育教学质量提供大量可参考的动态数据依据。

大多数二语习得过程研究都采用共时跨层语料，即通过初、中、高不同水平等级的共时语料来研究二语发展过程。[5]客观来讲，搜集共时跨层语料并开展相关二语习得研究的方法相对易于操作，且积极推动了二语习得研究的发展。然而正如 Doughty & Long（2003 年）指出：“由于纵向研究较少，大量的二语习得研究是横截面式的，使得在一些重要问题上所得出的结论存在严重的限制”[5]。

因此中介语语料库的建设不仅包含横向的语料搜集，也还是一个纵向的动态跟踪调查过程。只有将横向截面语料采集与纵向追踪调查结合起来，实时搜集纵向历时语料，实现中介语语料库的不断更新与完善，才能全面了解学习者由简单机械的模仿套话到自主生成韩语的动态渐进过程。然而，搜集跨时语料时间跨度大，必然耗费大量的人力和财力，建设全面的韩语中介语追踪语料库将会是一个长期而艰难的过程。

（三）协同创新性

协同创新性要求各独立的创新主体为实现共同目标打破壁垒，汇聚信息与资源，实现交流与互动，开展深入合作和资源整合，从而产生系统叠加的非线性效用。韩语作为非通用语种，与英语、汉语等通用语不同，国内各高校韩语专业初、中、高级学生人数及教研人员数量较少，如果单独由某一所高校采集语料和管理维护语料库，语料的数量和代表性方面都存在制约；另外，语料采集、入库、标注以及语料库管理与应用系统的维护也需要专门人力

和财力作保障。因此，构建反映中国学习者韩语中介语语言特征的中介语语料库单靠某一所高校或科研机构的力量很难实现。要走出这一困境，笔者以为可以由某高校牵头申报中介语语料库建设基金项目，获得教育管理部门的经济扶持，再通过各高校、研究机构间的协同合作，实现互动交流与资源共享，使各个环节之间发挥各自优势能力，协作开展韩语中介语语料库建设工作，从而服务于教学研究，提高教学质量，为社会培养更优秀的外语人才。

（四）开放共享性

开展韩语中介语语料库建设研究的目的就是汇集中国韩语学习者的学习语料，利用计算机技术对各类语言使用现象进行标注，依据检索手段发现学习者中介语中的共性问题，将经验教学和语料库研究的量化分析结合起来，从而更好地开展韩语教学、研究，促进韩语教育理论和实践的发展，而建设开放型语料库，实现语料库资源的共享则是实现这种目的的前提。

在计算机、互联网技术高度发展的今天，“互联网 +”、云空间、大数据等已成为这个时代的显著特征，信息时代的核心观念就是开放共享。目前中国国家图书馆和 HSK 动态作文语料库等数字化成果已面向社会开放。中介语语料库的建设也应当由“独建自营”向“合作共享”拓展，从国内语料库成果共享，到与世界各国的语料库建立共享关系。[6]“人文计算”是一个新兴的将现代信息技术深入应用于传统人文研究的跨学科领域。[2]在大数据、“互联网 +”背景下，韩语中介语语料库的建设属于人文计算的范畴，应该坚持语料库数据对外开放、资源共享的原则。

三、韩语中介语语料库的构建流程

由于国内韩语中介语语料库建设尚处于摸索阶段，可以借鉴英语、汉语中介语语料库的建设经验，尽量少走弯路。各语种语料的来源、学习者基础、错误类型等各有差异，因此在语料背景信息描述、语料标注等方面应体现各自特点。中国韩语中介语语料库的建设包括语料的采集与录入、语料的标注、语料库的管理与应用界面的维护等。

中介语语料库文本包括三种类型：首先是单模态中介语语料库，其次是多模态中介语语料库，最后是多维度中介语语料库。[1]韩语中介语语料库的建设应该是个循序渐进的过程，本文暂且只讨论单模态文字类韩语书面语中介语语料库的构建。我们以徐州工程学院韩语专业为试点，尝试建设“中国学习者韩语中介语语料库”。

（一）语料的采集与录入

语料是建库的基本前提，建设一个语料库首先要解决语料来源问题。[2]学习者中介语有书面语和口语形式，从学习者水平的角度又可以分为初、中、高级语料，因此中介语具有多样性和层次性的特点。另外，由于中介语动态发展的特征，中介语的搜集不仅包括横截面式的获取同时期不同级别学习者的语料，也应该是对学习者中介语的动态跟踪与监控过程。韩语中介语语料库的建设也并非把学生的语料网罗堆积就能完成，需要根据不同研究需要对语料进行归类。

1. 横向语料的采集

横向语料应遵循真实客观性的原则，广泛收集自然环境下韩语学习者各种类型的书面中介语语料。语料库建设初期，我们首先定期采集 2014、2015 届在校生的横截面书面语料，包括精读、写作、翻译等课堂及课后写作练习，内容涉及自然人物、景物描写，个人爱好介绍，平日见闻描述等初、中级阶段语料，以及对社会热点及抽象话题表达个人看法等高级阶段语料，根据学习者年级及语言水平进行归类，分别存放。

2. 纵向语料的采集

纵向语料的采集要根据研究时间、研究需要来确定采集的时间。国内韩语专业基本学制为四年，学习者在四年间经历从韩语入门到初级、中级、高级阶段的发展过程。建设“中国学习者韩语中介语语料库”正是为了研究中国学生习得韩语的发展过程，因此高校四年制是纵向语料采集最便捷的方式和时段。我们以 2017 届共约 50 名学生为受试对象，搜集他们四年间从初、中、高级阶段的造句、作文原语料，为了保证纵向语料库数据取样的均衡性，我们决定收集这 50 名被试的大一、大二、大三及大四阶段的作文各 10 篇，并按照时间先后

顺序进行编号。

3. 编辑语料背景信息及语料录入

成系统的语料能反映学习者的整个学习过程和完整的语言面貌，便于从各种角度对语料进行观察分析，对基于语料库的相关研究具有重要意义。[2] 语料的系统性不仅要求收集学习者在不同的语言学习阶段的语料，而且要求体现语料产出者的背景信息要齐全，且能完全对应。“中国学习者韩语中介语语料库”中对所收集的语料应注明语料产出者的性别、年龄、省份、民族、入学时间、语料产出时间、语料产出课程、韩语等级等背景信息。

语料录入应遵循真实客观性原则，如实反映学习者语言使用情况。“底层的不一致性在上层应用中会被放大几倍到几十倍”①，所以语料录入时应最大限度保持语料原貌（有些语料中可能会存在无法辨认的文字），同时要对已录入的语料进行严格核对，保证语料库的真实可用性。同时，为了给研究者呈现更真实的语料原貌，将学习者的原始语料以图片形式嵌入生语料库，更直观地展现语料的原始状态。

（二）语料的标注

语料库标注完成之日，就是问题研究清楚之时。①语料的标注为了使基于语料库的研究能够准确判断某一项或某一类语言表现形式，对生语料库的中的原始语料进行赋码注明，指出其语言特征。

中介语的特点决定了中介语语料库不同于母语语料库，建立韩语中介语语料库就是要为研究者们把握学习者的语言使用面貌及发展动态提供数据支持，从而更好地服务于二语习得研究及教学实践。要客观地把握学习者对于特定词汇或形态的使用频率、语言要素间的结合或连接、学习者语料错误类别及导致错误的因素分析，就要对生语料库进行必要的加工，开展语料的偏误标注工作。对于中介语语料库的标注问题，张宝林指出应坚持“全面性”原则，应在字、词、短语、句、篇、语体、语义、语用、标点符号等各个层面上对相关的语言现象进行标注，这样才能保证语料库功能的全面，主张“偏误标注 + 基础标注”的标注模式。[7] 肖奚强、周文华则从标注的广度、深度、角度和准确度四个维度探讨中介语语料库标注的全面性问题，主张采取“正确信息 + 错误信息”的标注方式。[8] 韩语中介语语料库从理论上可以根据研究的需要，从音韵、词素、单词、语节②、句法结构、语用信息等多个角度对语料进行分别赋码标注，研究目的不同，标注的内容与方法也不同。由于国内韩语中介语语料库建设仅处于摸索起步阶段，本文在坚持真实客观性的原则基础上，只讨论韩语中介语的词汇、语法层面的偏误标注，至于正确信息标注、语体、语用等其他层面的标注，只待韩语中介语语料库相关研究取得一定成果后再进行。

开展韩语中介语语料库偏误标注工作，首当其冲的问题应该是对偏误的判定与分类，即首先对学习者语料与目标语进行比对，对于被判定为偏误的部分区分是属于词汇偏误还是语法偏误，而词汇与语法各自的下位概念我们从形态、意义及使用范围三个方面去考量。词汇的形态主要指单词的构成与拼写（口语语料库中还包括单词发音），意义指单词的词典意义，使用范围指交流场景、前后语境下单词使用的恰当性；语法形态包括语法的拼写及结构，意义指如“时间名词 + 에”“地点名词 + 에”中“에”所表达的不同意义，使用范围指语法项目的使用限制条件或是否使用敬语法等。具体偏误标记如下表 1 所示：

表 1　偏误分类标注③

类型	项目
词汇偏误 VO	形态 FO
	意义 ME
	使用范围 US
语法偏误 GR	形态 FO
	意义 ME
	使用范围 US

①参见宋柔 2010 年讲座课件《文本语料库建设同语言教学和语言研究》。

②韩国语中有隔写法，每间隔开的一部分内容称为一个语节，即“어절”。

③此表格参考了김정숙, 김유정.한국어 학습자 말뭉치 구축을 위한 기초 연구—개인 정보 표지 체계와 오류 정보 표지체계를 중심으로的部分内容。

例如：

우리 언니는 내년에 결혼할（VOFO） 겁니다.
오늘 오전에 손님이 많았어요. 돈이 많이 팔았어요.（VOME）
제가 댁까지 데려다 줄게요.（VOUS）
어디를（GRFO） 재미있어요?
학교 근처에 있는 시장에（GRME）야채를 샀어요.
동생이 열심히 공부했더니（GRUS） 일등을 했어요.

对韩语中介语语料的偏误标注还应包括与目标语比对后具体差异的标注，包括遗漏（OMM）、添加（ADD）、语序错误（MISO）、替代错误（REP）等。

例如：

이 세상에 우정이란 _(게 OMM) 없으면 우리의 생활이 재미 없을 거예요.
내일에（ADD） 비가 올까?
우리는 많이 술을（MISO）마셨어요.
우리 중국 사람은 찬 반찬을 먹을 때 향유（REP）을 많이 사용해요.①

此外，韩语中介语语料库的偏误标注还应包括偏误范畴（词素、词、短语、句子、篇章等）、偏误原因分类（语际偏误、语内偏误）、语体（口语、书面语）、语用（尊敬语、非尊敬语）等相关内容的标注。因此，制定国内韩语中介语语料库标注规范、统一标注代码，以实现中国学习者韩语中介语语料标注的通用化与标准化，方便今后的语料库建设及语料库间的资源共享。

对韩语中介语语料库中的偏误做出精准的判断并进行科学的修正是语料库标注工作的关键。担任语料偏误标注工作的研究人员必须经过韩语语言（包括语音、词汇、语法、语用等）及韩国文化专业学习，并积累深厚的语言文化功底和丰富的韩语一线教学经验，同时要具备一定的二语习得理论知识，才能从外在的语言现象准确分析出现偏误的原因。此外，在标注之前应对标注人员进行严格的培训并通过相关考核，最大限度地提高语料标注的正确率。单词拼写、隔写法、语法规范、外来语标记规范等语料标注应以韩国国立国语院相关的语言文字规范为标准，如单词拼写、外来语标记、隔写法以《표준국어대사전》《한글 맞춤법》《외래어 표기법》等为标准，语法规范以《외국인을 위한 한국어 문법 1，2》《표준어 규정》等为主要依据。目前国内部分高校韩语语料库的偏误标注工作主要依赖朝鲜族教师，原因是朝鲜族教师母语为朝鲜语，在对偏误现象的识别及判断方面具有相对优势。非朝鲜族韩语教师由于亲身经历了从零基础到熟练掌握韩语的学习过程，因此能更准确把握大学生的韩语认知习得规律及偏误原因，在经过专业的语言文化学习和有效的培训之后，也可胜任语料库的偏误标注工作。

语料库数据信息庞大，仅靠人力对各种语言现象进行判断并根据预先设定的代码进行加工处理的话，对于工作人员的语言能力、精神状态、体力等都是巨大的考验，耗时耗财耗力也很难保证标注质量。因此，在语料的标注过程中，应由软件技术员根据预先设定好的标注规范及代码，开发自动机读标注软件，实现计算机自动标注，再由语言工作者进行检验和校对，对误标或漏标进行修改和补充。随着现代科技的不断发展，计算机自动标注和人工辅助校对将会成为较为理想的语料库标注方法。

（三）语料库管理与用户检索界面

语料库的建设不仅包括语言工作者的语料收集及标注工作，还需要计算机技术人员开发研制语料库管理软件及用户检索系统。因此“中国学习者韩语中介语语料库”应包括语料素材登录管理界面和用户检索界面两个功能模块，语料素材登录管理界面主要包括韩语中介语语料的录入、赋码标注、查对等功能，用户检索界面主要为使用者开展中介语研究提供语料和数据统计功能。

在语料素材登录管理界面的建设过程中，计算机技术人员完成软件的编程和开发之后，需要语言工作者与技术人员通力合作，将中介语语料与语料背景信息及统计数据入库，根据预先设定的标注代码开展语料自动标注及修订和补充工作，实现语料库资源的整合和功能的完善，整个过程需要语言工作者和技术人员的密切配合，最终确保用户顺利使用。

普通的语言教师和研究者在使用中介语语料库时普遍希望用户检索界面简洁易懂、操作方便，供检索的内容应分类清楚，可以随用户的意愿分别或同时查询与显现。[4]328 如实现韩语中介语语料库对

①例句来源于徐州工程学院 15 级朝鲜语专业 1 班 2016－2017 学年第二学期作业。

单词、语节、短语、句子、短文等不同级别语料的浏览和检索，从不同层面满足研究者的需求。此外，用户检索界面还应为用户提供多种不同检索条件，用户可以根据研究需要选择检索条件，自主设置显示内容并自动保存检索统计结果。

我们设想建设的“中国学习者韩语中介语语料库”将坚持开放共享的原则，并根据研究需要源源不断地增加语料资源，实现语料库建设的持续动态发展和可监控性，为国内韩语教学改革、教材编写、能力评估测试等提供有力的数据支持。

四、结语

大数据时代下二语教学观念应跟随社会发展步伐，目前国内、外语言学研究领域在语料库的开发与建设取得了显著的成绩，而国内韩语教学研究领域有关韩语中介语语料库的研究却一直很少有人关注，起步阶段的韩语中介语语料库建设研究应在借鉴其他语种语料库建设成功经验的基础上，在语料采集、语料存取与处理、语料库资源共享与继续建设方面，融合最新计算机技术（如大数据处理技术）实现韩语中介语语料库的智能优化与更新。然而，建设韩语中介语语料库只处于起步阶段，它将会是一项艰巨复杂的系统工程，任何个人和部门单枪匹马根本无法完成，需要广泛集聚学界的智慧和力量，共同探讨语料库的建设思路和标准，循序渐进地建成并不断完善韩语中介语语料库。

［参考文献］

［1］周文华. 汉语中介语语料库建设的多样性和层次性［J］. 汉语学习，2015（5）：97－105.

［2］张宝林，崔希亮. 谈汉语中介语语料库的建设标准［J］. 语言文字应用，2015（2）：125－133.

［3］杨连瑞，张德禄，等. 二语习得与中国外语教学［M］. 上海：上海外语教育出版社，2007.

［4］张宝林. 关于汉语中介语语料库建设的若干重要问题［C］//第八届中文电化教学国际研讨会论文集. 上海：中文电化教学国际研讨会，2012.

［5］曹贤文. 留学生汉语中介语纵向语料库建设的若干问题［J］. 语言文字应用，2013（2）：127－134.

［6］郑通涛，曾小燕. 大数据时代的韩语中介语语料库建设［J］. 厦门大学学报（哲学社会科学版），2016（2）：53－63.

［7］张宝林. 关于通用型汉语中介语语料库标志模式的再认识［J］. 世界汉语教学，2013（1）：128－140.

［8］肖奚强，周文华. 汉语中介语语料库标注的全面性及类别问题［J］. 世界汉语教学，2014（3）：368－377.