首届全国学习者语料库专题研讨会综述

陈功

(北京外国语大学 中国外语教育研究中心, 北京 100089)

提要: 2009年12月26-27日“首届全国学习者语料库专题研讨会”在北京外国语大学召开。会议围绕四个议题展开研讨, 即学习者语料库建设、学习者语料库研究方法、学习者语料库分析技术、基于学习者语料库的相关研究。在回溯反思以往学习者语料库研究的基础之上, 与会者积极参与研讨, 会议表现出敢于创新求变、凸显团队合作、强调学术严谨、激发理论争鸣的四个特点。

关键词: 学习者语料库

中图分类号: H319 **文献标识码:** D

为了进一步推动我国学习者语料库的建设、研究及其在语言教学中的实际应用, 北京外国语大学中国外语教育研究中心主办、外语教学与研究出版社承办的“首届全国学习者语料库专题研讨会”于2009年12月26—27日在北京外国语大学召开。本次会议吸引了来自国内及台湾近百所高校及相关研究机构的100余名代表与会, 还有来自英国、新加坡等从事学习者语料库相关研究的学者参与了研讨。

在本次研讨会上, 卫乃兴教授(上海交通大学)、何安平教授(华南师范大学)、李文中教授(河南师范大学)、许家金博士(北京外国语大学)、濮建忠教授(解放军外国语学院)、洪华清博士(新加坡南洋理工大学)、梁茂成教授(北京外国语大学)等七位长期活跃在学习者语料库研究领域的专家分别作了主旨发言。另有60余名与会代表在分会场宣读了论文。

本次会议围绕四个议题(即学习者语料库建设、学习者语料库研究方法、学习者语料库分析技术、基于学习者语料库的相关研究)展开研讨。在回溯反思以往学习者语料库研究的基础之上, 与会代表表现出强烈的创新求变的意识。本文为会议的述评。

1. 研讨内容概要

1.1 学习者语料库建设的创新

李文中教授题为“Constructing a Learner English Portfolio Corpus on the Open Corpus Platform: A Research Proposal”的主旨发言中认为, 学习者在语言能力和学习问题上呈现多元化和个性化的特征, 而当前学习者语料库侧重于群体性特征的描述, 忽视了学习者的个体特征; 学习者的语言变化是一个动态发展的过程, 具有历时性, 而当前学习者语料库语料的收集往往只能抽取代表学生某一阶段水平的共时性语料, 在很大程度上无法对学生的学习过程和语言发展过程进行描述; 此外, 学习者的语言学习具有高度的语境化特征, 而当前学习者语料库由于受技术的限制在很大程度上是去语境化的产物, 既无教师或同伴的反馈, 又无声音图像等对意义的构建起着重要作用的非文本模态。基于这些想法, 李文中教授提出了构建基于网络学习系统的学生语料动态采集机制。这种前端为网络语言学习, 后端为形成性学生语料汇集的思路, 被李文中教授称为“电子档案袋语料库”。单个学习者的语言学习记录形成个体库, 所有个体库的集合为群体库。这种动态实时的学习者语料库具有历时性和追踪性, 可以对学习者的语言产出进行动态的记录和追踪。因此, 可作为对学生形成性评价的重要资源。此外, 李文中教授还演示了他的团队开发的结合文本和音频视频收集检索的“电子

档案袋语料库”的多模态语料收集检索原型平台。李文中教授团队所设计开发的学习者语料库已经不再是传统意义上的完全依靠文本的语料库,而是综合了学习者个体特征和群体特征、历时性和共时性、学习者产出和教师同伴反馈的,包含文本、声音、图像等多种模态的新型学习者语料库。

洪华清博士题为“**Multimodal Learner Corpus Construction: Challenges and Directions**”的主题发言指出,现在大多数学习者语料库关注的是学习者的产出,而很少关注学习者的语言是如何通过他们所处的语言学习环境发展起来的,目前很少有考虑到影响学习者学习因素的语料库。鉴于此,洪华清博士提出学习者语料库的建设要坚持“整体观”(Holistic Approach)。如果学习者语料库能够把教材、讲义、作业、课堂互动、教师话语等因素结合起来,就有可能发现影响学习者语言学习的积极因素和消极因素,以及各因素对学习者的语言水平发展的解释程度。本着这一思想,洪博士介绍了他的团队开发的新加坡教育语料库(SCoRE)。该语料库收集的语料包括课堂录像、教材讲义和学生作业,形成多模态语料库。为了充分挖掘语料库的价值,洪华清博士团队利用XML语言对课堂话语进行了多层次标注并开发出了相应的检索平台。该语料库充分整合了转写文本、课堂录像和语言学和教育学研究者的分析数据,使得教育研究者可以方便地检索课堂教学中的学习者或者教师的互动行为,从而为研究学习者语言、诊断教学和课程改革提供事实依据。

此外,在本次会议的分组发言中,另有六篇论文探讨了学习者语料库的建设问题。其中四篇主要围绕专门用途学习者语料库的建设展开,包括学习者英语学术论文语料库建设、英语专业学士学位论文语料库建设、理工科大学生英语口语多模态语料库建设、英语专业写作教学语料库建设。还有两篇探讨的是对外汉语教学语料库以及中国大学生德语中介语语料库的建设。

从本次会议的主旨发言和分组发言可以看出,学习者语料库建设已经从原有的文本语料库建设逐渐发展为多模态语料库建设,从共时语料库建设转向历时语料库建设,从学习者作文语料库建设转向多种专门语料库建设。同时,除了英语学习者语料库建设之外,其他语种的学习者语料库和对外汉语语料库建设也得到了重视,学习者语料库建设已从英语辐射到其他语种。

1.2 对学习者的语料库研究方法论的探讨

在学习者语料库研究不断深入的今天,如何拓展学习者语料库研究的深度和广度是研究者需要认真思考的一个问题。卫乃兴教授在题为“**Phraseology in Contrast: A Multi-dimensional Contrastive Analysis of Learner Phraseologies across Corpora**”的主旨发言中,系统总结了学习者语料库研究的主要特点,并在继承Sinclair思想的基础上提出建设典型的学习者语料库需要遵循的原则。他特别指出,学习者语料库研究和二语习得研究之间既存在差别又互为补充的关系,以及学习者语料库研究反哺教学的核心价值。之后,卫乃兴教授提出了学习者语料库研究的多维度模型,即从多个层面对学习者语言中的短语做对比研究,进而观察短语的形式、语义和语用等特征。他以**bent on**为实例,通过在SWECCCL和BNC中进行检索,分析了学习者和本族者所使用的**bent on**在词语搭配、类联接、语义倾向和语义韵等各个语言层面的特点。同时卫教授引入多个对比层面(中介语语料、母语语料、双语平行语料),进一步分析了**bent on**在原生英语和译自汉语的英语中不同语言层面上的特征。在此研究的基础上,他指出高级英语学习者语言使用的主要问题不是语法,而是语言搭配、语义倾向和语义韵的选择。由此可见,有关学习者短语研究的多维分析思路,不仅增强了学习者语料库研究分析的层次和深度,还可以对学习者的语言使用问题做适当解释。

何安平教授在题为“**过多使用, 过少使用, 错误使用, so what?**”的主旨发言中,直指学习者语料库研究中的一个突出问题,即重描写轻阐释。与本族者语料库相比,学习者语料库

研究首先发现的是某些语言形式的过多使用、过少使用或错误使用,至此如何将研究深入下去取得更有价值的结果是一个亟待解决的问题。何教授通过文献考察和个案分析展示了如何应用语料库语言学的短语理念将语言表层形式的频数统计延伸至语法结构、语义和语用等语言的较深层面。她以中国学生过多使用的 I think 为例,指出除了关注频数高低外,还可以观察 I think 的语义倾向和语义韵等周边语言环境,从而分析出这些词项的语用特点,发现中国学生表意不足情态有余的问题。此外,何教授还引用了 Nick Ellis 利用语料库建立学术话语词块表的研究,探讨了如何结合其他相关领域的研究方法对中介语特征的成因进行推广和论证,使学习者语料库的研究发现真正运用到教学中去,指导实际教学。

卫乃兴教授和何安平教授的主旨发言所涉及的共同问题是,学习者语料库研究不能拘泥于频数多少的比较。研究者应该进一步拓展思路,着眼于搭配、类联接、语义、语用等特征,分析学习者语言的深层次特点。

1.3 学习者语料库研究的拓展

1.3.1 学习者语料库驱动研究的拓展

许家金博士另辟蹊径,利用学习者语料库开展了有关中国英语专业大学生的话语认同的社会语言学研究。他的主旨发言题为“**Storied Self in Another Language: A Collocational Approach to Interlanguage Identity**”。该研究以社会建构理论为基础,试图通过话语身份的三个分析维度,即主体特征、社会网络、互动语力,构建一个抽象的中国英语专业大学生“施惟可(SWECCL)”的自我形象。在研究方法上,许博士指出,传统的话语认同和自我形象的研究通常为个案研究,通过访谈、人种志等质性研究的方法获得数据,是“显微镜”式的研究,而语料库方法可以称之为“望远镜”式的研究,能够更方便地从整体入手,而且语料库研究中的大量客观数据可以有效降低研究者的主观色彩,因此可以作为传统话语身份构建研究的有力补充。他采用框合结构(concgram)与关键主题词(key keywords)相结合的方法,提取出 SWECCL 口语子库中的典型词汇和短语模式,归纳这些模式的语义共性,有效地构建了“施惟可”的形象。框合结构是近年来语料库相关研究中的一个新视角,而关键主题词这一概念虽早就有人提出,但在实际研究中很少被研究者所使用。许博士独具匠心地将两种方法结合起来,从社会语言学视角出发,研究学习者的话语认同,无论从研究方法上看还是从研究视角上看,都是对学习者的语料库研究的极大拓展,对今后的研究具有很大启发性。

1.3.2 基于学习者语料库研究的拓展

濮建忠教授的主旨发言以“从词块的使用看学习者语言能力的发展”为题,为与会者展示了基于学习者语料库研究向纵深拓展的范例。濮教授首先指出短语式意义单位研究(即词块)在语言研究、语言教学与学习中的重要性,并详细解读了 Sinclair 关于扩展词汇单位的五大成分(核心词、词语搭配、类联接、语义倾向和语义韵)在语言研究中的地位,逐步解剖了 Sinclair 所提倡的观察方法。之后,濮教授以词块的使用为切入点,利用自己收集的历时学生语料,从多文本和单文本两个视角出发,从词语搭配、类联接、语义倾向和语义韵几个层面,对以 find、life、hard 几个词为核心的词块做了细致的分析,以观察我国英语专业大学生写作能力的发展情况。他的研究通过词丛比较和关键词丛分析,解析了中国英语学习者的词块特征。分析表明,中国英语学习者群体已经具备较好的词块意识,但经过两年左右时间的学习,学习者在词块的各个层面进步并不明显,应当引起教育工作者的重视。濮教授最后指出,词块模型具有非常大的潜在作用,可以构建篇章衔接理论、搭配理论、社会行为理论、关于人如何认识世界的理论以及构建数据驱动的意义理论等等。他的研究抓住词块这一热点问题,从搭配、类联接、语义倾向和语义韵等多方面对学习者的语言进行跟踪研究,不仅在研究方法上具有新意,也从一个侧面表明了历时语料库的价值所在。

1.4 学习者语料库分析技术

语料库分析技术是制约学习者语料库研究发展的瓶颈之一, 语料库分析技术的实现及其合理运用是科学呈现语料库研究数据的前提。梁茂成教授带来了他的研究成果 **Keywords+**, 并做了题为“**Keywords Plus: 主题词研究方法扩展及其实现**”的主题演讲。主题词方法是语料库对比研究中的最常见方法。在学习者语料库对比研究中常常以卡方检验或对数似然率为测量工具, 考察某些词在两个语料库中的使用频率是否存在显著差异。然而, 常见的语料库分析工具仅以单个的词为分析对象, 统计语料库容量时所考虑的也是形符数。随着当今语料库研究对短语和搭配越来越重视, “词”这一语言单位逐渐淡化, 因而以上分析方法所能解决的问题极为有限, 远不能通过对比的方法有效地发现文本间的差异。针对目前主题词分析工具存在的上述计算缺陷, 梁教授设计并实现了一种新型研究工具 **Keywords+**。该工具突破“词”的界限, 将语料库对比中的分析单位扩展到词簇乃至词性赋码序列, 根据词簇或结构的频数和 n 元组的总频数来发现观察语料库中的主题词、主题词序列 (**keyword sequences**) 乃至主题词性赋码序列 (**key POS tag sequences**)。此外, 该工具还可以将短语或赋码序列的对比、显著性计算和索引行析出结合起来, 一键输入, 完成全部分析, 对于有效分析文本特征具有重要意义。会上, 梁教授对该工具的上述功能进行了演示, 为广大参会代表展示了语料库分析技术的强大魅力。

会议上, 另有三位代表介绍了语料库分析工具在研究中的应用, 包括多模态信息检索工具 (**MCA**) 在外语教学中的应用, **Stanford NLP** 和 **Uplug** 在平行语料库词块提取和对齐中的应用, 以及语料库处理工具 **MonoConc Pro** 在提取学习者搭配错误中的应用。对语料库分析技术的探讨, 体现了研究者们对语料库研究方法的思考和突破研究瓶颈的需求, 也极大地激发了与会代表学习和探索语料库相关技术的热情, 使大家更有信心使用语料库技术和工具为自己的研究服务。

2. 会议评述

作为全国首届以学习者语料库研究为专题召开的会议, 本次大会集结了国内学习者语料库研究领域的一流专家学者, 吸引了一大批热衷于学习者语料库研究的专家、教师和研究生前来参会。可以说本次会议规模大、层次高, 对于学习者语料库研究在国内的发展将会起到极大的推动作用。综合起来看, 本次会议表现出了如下特点:

第一, 敢于创新求变。本次大会的参会代表们表现出了强烈的创新求变意识。在学习者语料库建设方面, 多位学者不满足于单纯基于文本的传统语料库, 而利用多媒体处理技术, 将音频、视频和文字整合起来, 构建多模态学习者语料库。如李文中教授团队设计的具有历时性、动态性和个性化的“学习者电子档案袋语料库”以及洪华清博士及其团队开发的、综合考虑影响学习者语言学习因素的新加坡教育语料库。在学习者语料库研究的方法论探讨中, 几位主旨发言专家不约而同地指出, 学习者语料库的分析层面需要从词项为基础的局部语言单位的研究扩展至词汇语法的共现、语义倾向和语义韵的多维度综合考察, 将学习者语料库研究向纵深推进。在学习者语料库研究方面, 许家金博士创造性地将学习者语料库研究和社会语言学研究相结合研究学习者的话语形象, 令参会代表感到耳目一新。在语料库分析技术方面, 梁茂成教授研发的工具打破了传统的主题词概念, 为语料库研究突破词的界限向深层次发展提供了技术支持。本次会议展现出来的诸多新的研究思路、视角和技术实现方法为今后中国学习者语料库的研究开辟了新的天地。

第二, 凸显团队合作。卫乃兴教授在本次会议中欣喜地谈到, 我国学习者语料库研究的团队已经基本形成。这些团队包括: 上海交通大学的卫乃兴教授团队、河南师范大学的李文中教授团队、华南师范大学的何安平教授团队、北京外国语大学的梁茂成教授团队等。语料库研究不同于语言研究的其他分支及流派, 它通常包括语料库的建设、方法论的探索、具体

应用性研究的开展, 还包括语料库分析技术和工具的开发及实现。团队的建立能够有力地保证语料库研究各个环节中人力资源的配置, 有利于整个团队的不断创新。此外, 团队间的交流和合作也是语料库建设与研究中的重要环节, 可以有效地促进思考、开拓视野、避免重复劳动。本次会议上各位专家不仅带来了各自团队的研究成果, 还进行了深入的探讨, 实现了团队之间的交流。团队精神和团队间频繁的深层次交流将使得学习者语料库建设与研究步入良性循环, 为中国学习者语料库研究源源不断地注入活力。

第三, 强调学术规范。语料库语言学研究以大量客观而科学的数据著称, 数据的可信度和可验证性是语料库研究的一大特点。梁茂成教授在他的发言中也提到了当前学习者语料库研究中的数据处理问题。比如, 计算某一语言现象频数的问题、利用 SPSS 计算某一语言特征差异性时的数据处理问题, 以及卡方检验本身的算法缺陷, 等等。作为语料库研究的一个基本前提, 数据及其算法的准确性具有至高无上的重要性。为此, 梁茂成教授以其开发的工具 **Keywords+** 为例, 重申应最大限度地避免研究中的数据问题和计算问题。此外, 梁茂成教授还强调, 研究中客观存在的问题是可以通过技术的进步来解决的, 但是对于某些人为操纵数据、扭曲数据的行为, 应坚决抵制。

第四, 激发理论争鸣。在大会主旨发言和小组发言中, 各位专家和与会代表就各自的研究展开了深入的探讨, 也激发了大家对语料库语言学研究一些理论问题的争鸣。如, 在谈到语料库的标注时, 与会代表探讨了语料库是否应该标注以及标注到什么程度的问题; 学习者语料是否是真实自然的语料; 学习者语料库对比分析是否应该坚持以本族者语言作为参照标准; 学习者语料库研究应该采用自上而下还是自下而上的方法等等。这些理论上的探讨进一步加深了我们对学习者语料库研究的了解, 虽然有些问题尚无定论, 但却在一定程度上提升了我们的理论高度, 开阔了与会者的研究视野, 为今后研究的深化打下了更好的理论基础。

从本次学习者语料库专题研讨会的研究热点看, 近期我国学习者语料库相关研究呈现出以下主要特点和发展趋势:

第一, 学习者语料库多模态化、专门化的趋势。传统的文本语料库提倡“清洁的文本”, 而文本之外的许多因素(如字体、图像、语音语调、人际互动等)也是意义构建的重要成分, 理应收集到语料库之中。与会专家和研究者极力提倡建设多模态学习者语料库和动态语料库, 全方位跟踪学习者的语言学习过程。此外, 随着研究的不断深入和不同研究需求的出现, 研究者对学术英语语料库等专门语料库的建设也越发重视。

第二, 研究方法系统化, 突破原来的中介语对比分析方法。与会专家和研究者普遍重视研究方法的系统化, 主张从搭配、类联接、语义倾向、语义韵等多个不同维度对学习者的语言进行研究。

第三, 研究中空前重视短语问题。从大会的主旨发言和分组发言看, “词”的概念越来越淡化, 取而代之的是搭配、类联接、词块、意义单位、主题词序列、词性赋码序列等与短语相关的概念。

第四, 研究视角多元化。第二语言习得视角是传统的学习者语料库相关研究中最重要、最核心的研究视角。在本次研讨会上, 语用视角、社会语言学视角、语言教学视角等为学习者语料库研究增加了新的取向。

第五, 技术开发紧密结合研究方法。在语料库语言学发展的早期, 语料库建设、语料库工具开发常常是技术人员需要关注的问题, 而语料库应用和基于语料库的研究则常常是语言研究者更为关注的问题。从本次会议上看, 研究者以团队为单位, 自主设计和开发语料库、研制相关工具成为一种新趋势。这种趋势不仅使得语料库建设呈现出百花齐放的繁荣景象, 而且使得语料库和研究工具能够更好地满足研究者的需要, 使得技术开发更加有的放矢。

Overview of the First National Symposium on Learner Corpora

CHEN Gong

(National Research Centre for Foreign Language Education, Beijing Foreign Studies University, Beijing 100089, China)

Abstract: On “the First National Symposium on Learner Corpora”(December 26-27, 2009, Beijing) four topics (learner corpus construction, learner corpus research methodology, learner corpus analysis techniques and learner corpus studies) were discussed. The participants actively participated in seminars and put forward a number of creative ideas on the basis of reflections on learner corpus research in the past.

Key words: learner corpus

收稿日期: 2009-12-28

作者简介: 陈功, 女, 北京外国语大学中国外语教育研究中心博士生。研究方向: 语料库语言学, 第二语言习得。