我们需要什么样的汉语中介语语料库
(What Kind of Chinese Interlanguage Corpus Is Needed)

张宝林
(Zhang, Baolin)
新疆师范大学/北京语言大学
(Xinjiang Normal University / Beijing Language and Culture University)
zhangbl@blcu.edu.cn

摘要：汉语中介语语料库自问世以来极大地促进了汉语二语教学与习得研究的发展，其自身建设的设计水平和整体功能也随着研究的深入得到了很大提升，跨入了 2.0 时代。然而目前存在的单语种，横向语料，语料不平衡等问题，无法给“母语负迁移”之类的研究结论、汉语二语习得过程研究等提供充分的证据与支持。汉语习得研究正在向复杂动态系统理论指导下的二语发展研究转变，急需建设多语种、纵向语料、查询便捷、功能丰富的平衡语料库，把语料库建设由 2.0 时代推进到 3.0 时代，为汉语教学和习得研究提供适用而充足的语料资源支持。

Abstract： Since its advent, the Chinese interlanguage corpus has greatly promoted the development of research on Chinese as a second language (CSL) teaching and acquisition. Its own construction has also seen significant improvements in design level and overall functions, stepping into the 2.0 era. However, existing problems such as monolingualism, cross-sectional corpus, and unbalanced corpus cannot provide sufficient evidence and support for research conclusions like "negative transfer of mother tongue" and studies on the process of CSL acquisition. Research on Chinese acquisition is shifting towards the study of second language development under the guidance of Complex Dynamic Systems Theory (CDST), and there is an urgent need to construct a balanced corpus with multilingualism, longitudinal corpus, convenient query, and rich functions, advancing the corpus construction from the 2.0 era to the 3.0 era, so as to provide applicable and sufficient corpus resource support for CSL teaching and acquisition research.

关键词：汉语中介语语料库；多语种；纵向；平衡；3.0 时代

Keywords: Chinese interlanguage corpus; multilingual; longitudinal; balance; 3.0 Era

1. 语料库的作用与发展

1995 年 11 月，第一个汉语中介语语料库“汉语中介语语料库系统”在北京语言学院（北京语言大学前身）问世，立即引起了汉语学界的广泛关注，《世界汉语教学》（1995 年第 4 期）《中国语文》（1996 年第 2 期）均予报道。进入本世纪以来，汉语中介语语料库（以下简称“语料库”）以其庞大的语料规模和便捷的查询手段，为汉语二语教学与习得研究提供了量化研究的坚实基础，推动了汉语二语习得研究从主观思辨性研究范式向基于大规模真实语料的定量研究与定性研究相结合的实证性研究范式转变，也推动了基于语料库的汉语二语习得研究的发展，取得了大量的研究成果。例如 2006 年底建成开放的 HSK 动态作文语料库（简称“HSK 库”）[1]，截至 2025 年 11 月 20 日，注册用户为 122956 人，访问量达 1822704 人次；在中国知网（CNKI）查询，基于该库进行研究发表的各类论文达 10602 篇[2]（年度发文量详见图一）。全球汉语中介语语料库（简称“全球库”）[3]于 2019 年正式开放，注册用户为 31106 人，访问量达 219082 人次；基于该库进行研究发表的各类论文达 1226 篇[4]（年度发文量详见图二）。这些数据表明，汉语中介语语料库在汉语二语教学与习得研究中发挥了重大作用。

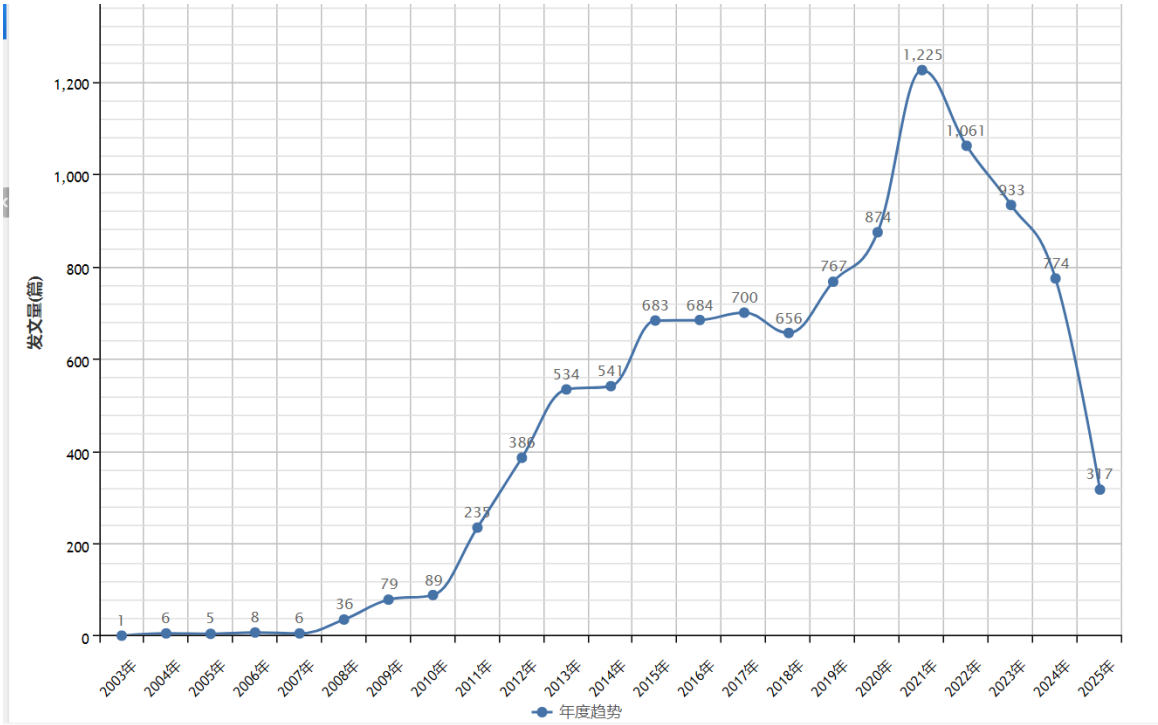


图 1　HSK 库年发文量分布图

[1] 网址：http://hsk.blcu.edu.cn。
[2] 检索方式：句子检索；检索式：HSK+语料库。
[3] 网址：https://qqk.blcu.edu.cn。
[4] 检索方式：句子检索；检索式：全球+汉语中介语语料库。

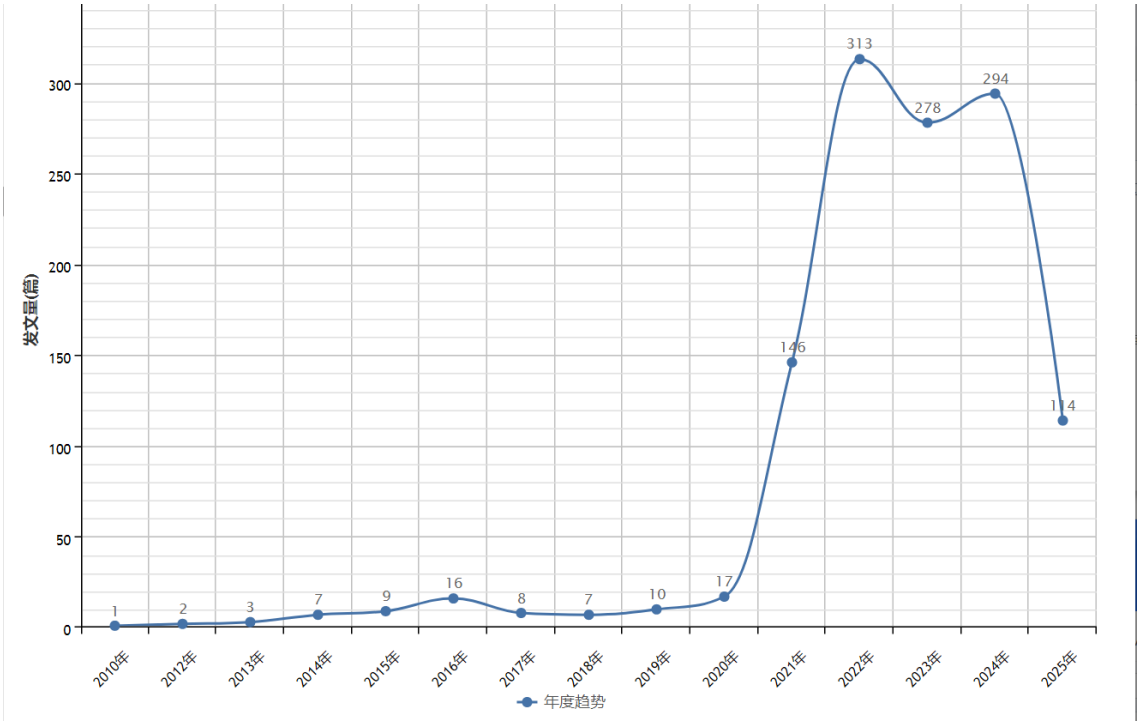


图 2 全球库年发文量分布图

同时，基于语料库的汉语二语习得研究的发展也促进了语料库建设的发展，“汉语中介语语料库建设渐成高潮，‘成为语料库研究中的热点’（谭晓平，2014），汉语中介语语料库建设正在跨人一个繁荣发展的重要时期”（张宝林、崔希亮，2015）。语料库的设计水平和整体功能得到很大提升，从“1.0 时代”跨入了“2.0 时代”（张宝林，2019）。二者的主要区别见表 1。

表 1 汉语中介语语料库 1.0 时代与 2.0 时代特征对照表

对照项	1.0 时代特征	2.0 时代特征
建设目的	自建自用为主	共建共享，服务学界为主
建设方式	离线，分包，固化	在线，众包，迭代
语料规模	百万字级	千万字级
语料类型	一种，笔语/口语	多种，笔语+口语+视频
标注内容	少数语言层面	追求全面标注
标注模式	偏误标注	偏误标注+基础标注
检索方式	简单检索，2 种	复杂检索，9 种
应用研究	偏误分析为主	“三性”分析，表现分析
总体概括	简单粗放	精细而丰富
起止时间	2000-2017	2018-现在
代表性语料库	HSK 库（1.1 版）	全球库

语料库建设推动了汉语二语习得研究发展，而汉语二语习得研究的发展又促进了语料库建设，可谓良性循环，相得益彰。

2. 存在的问题

2.1 语料库建设存在的问题

1）语种问题

目前的汉语中介语语料库种类繁多，包括笔语库与口语库，通用型库与专用型库，单体语料库与多维参照的汉语中介语语料库库群（胡晓清，2016），为汉语教学与研究服务的语料库与既为汉语教学和二语习得研究提供数据支持和检索服务，也可作为语法自动纠错算法的训练与评测数据，服务于智能辅助写作技术研究的语料库（王莹莹等，2023），1.0 时代所建之库与 2.0 时代所建之库。凡此种种，皆为单语库，即汉语中介语库。带来的问题是只看中介语语料和目的语语料，而没有学习者母语语料，根据什么做出“母语负迁移”的结论？显而易见，单语语料库是无法为语言迁移研究提供学习者母语语言事实的支持的。在这种情况下得出的“母语负迁移”之类的结论只能是既有理论的翻版复制，使偏误成因的研究变成了一种对号入座的固定套路；且论述简略，缺乏深度和参考借鉴价值。

2）语料连续性问题

从语料库建设整体情况看，语料缺乏连续性，多为共时语料库，而非历时语料库。可供中介语的静态研究之用，而不能为二语习得过程的动态考察提供支持，不能充分满足汉语二语教学与研究的多方面需求。从语料库建设本身来看，其设计水平和建设水平都是不高的。

3）平衡性问题

语料库中各类语料的数量及其平衡性十分重要，决定着不同类型的语料之间是否具有可比性，研究结论是否可靠。可见，语料的平衡性在一定程度上决定着语料库的功能和使用价值。从目前公开的语料库来看，这方面做得并不好。例如 HSK 库自 2006 年建成后即向学界免费开放，在中介语研究方面发挥了很大作用，但在语料产出者的国籍分布方面极不平衡，语料多者达数千篇，少者仅有几篇甚至一篇（任海波，2010）。语料太少不仅无法进行不同母语学习者习得汉语的对比分析，也不能反映学习者的习得规律，几乎没有使用价值。全球库的语料平衡性有较大改进，但问题依然存在，并未彻底解决（张宝林，2022）。有的语料库注意到了这一问题，但其并不对外开放，因而不能发挥其应有的作用。

4）标注问题

1.0 时代的语料库一般只做偏误标注，不做基础标注，即正确语言现象的标注（张宝林，2010）。且标注的内容很少，一般只有字、词、句等少数语言层面的标注；且不充分，有的只做几个句式的标注，其他句式即弃之不顾。作为 2.0 时代语料库的典型代表，全球库贯彻全面标注的原则，进行了字、词、短语、句、篇、语体、辞格、标点符号、语音、体态语等 10 个层面的标注，扩大、提高了语料库的功能与使用价值（张宝林、崔希亮，2022）。但能如此标注的语料库很少，尚属凤毛麟角。关于偏误分类，有研究以“遗漏、误加、误代、错序”四大偏误类型为参照，将偏误类型确定为“成分缺失、成分冗余、词汇误用、语序错误”四类。认为如此分

类“大大简化了偏误标注的难度，更有助于训练 GEC 模型”（王莹莹等，2023）。这一看法与做法不无道理，但从习得角度看，是会加剧目前基于语料库的偏误分析中套用“四大分类”（即遗漏、误加、误代、错序），不做具体、深入分析的不良倾向的，也就难以发现新的中介语现象，得出新的研究结论。即便对于 GEC（语法自动纠错）来说，只能处理这四种偏误现象，功能并不强大。况且，“遗漏、误加、误代、错序”的分类本身也还存在“太概括”的缺点，“学生的错误事实上比这个要复杂得多”（盛炎，1990，130）。在标注方法方面，大多数层面的语料标注为人工标注或人标机助，标注的一致性、准确性难以充分保证。全球库建成后曾专门组织人力进行审核修改，大大提高了标注正确率；但如果没有足够的人力和经费支持，这种审核修改工作是难以进行的。总体来看，标注质量问题尚未彻底解决。

5）检索问题

一般的语料库检索方式十分简略，只有字符串一般检索和对标注内容的检索，只能处理对一个查询对象的检索，因而对一些库存语料中存在的语言现象却无法检索。例如对离合词“离”的用法、“不……不……”等半固定格式、“是……的”句等有两个检索对象的语言现象即无法检索。全球库根据用户需求研发了 9 种检索方式，大大增强了检索能力。但只有分类标注检索可以检索到偏误语料和正确语料，其他检索方式则不能分别检索两种语料，使用上仍然不大方便。至于“A 的 A，B 的 B”结构（例如“跳舞的跳舞，唱歌的唱歌”）尚无法直接检索。目前需要进一步改进检索方式，以满足用户的使用需求。

6）一些基础性问题

（1）语料的分词与词性标注问题

在中文的自然语言处理中，分词与词性标注研究最为成熟，分词正确率可达 98%左右（刘开瑛，2000），甚至 99%左右（黄昌宁、李涓子，2002）。其中分词是词性标注的前提，词性标注又是实现“按词性检索”的基础，分词和词性标注的水平制约着按词性检索的实际效果。然而时至今日，汉语中介语语料库建设一直没有自己的分词规范和专用词表，而是借用母语语料库建设或中文信息处理用的规范和词表。由于中介语中存在的字词偏误，机器自动分词存在分词错误是必然的，在错误分词基础上所做的词性标注存在错误也是必然的。例如：由于别字形成的“有宜（友谊）、知说（知识）”，由于语素顺序颠倒形成的“忘淡（淡忘）、爱亲（亲爱）”，由于学习者臆测形成的“慈脸（慈祥的脸）、高量（大量）”在汉语词汇中并不存在，在各个分词系统的词表中也不可能有。因而在分词时会将这些组合切分开，并错误地标记不正确的词性代码。显而易见，研制汉语中介语语料库建设专用的分词规范与词表是提高语料库建设水平的当务之急。

（2）语料的自动分级问题

为了保证研究结果的客观性、稳定性和普遍意义，库存语料越多越好（杨惠中主编，2002），来源越广越好，类型越丰富越好。存在的问题是，不同来源的语料所标明的语言水平可能评价标准不一，因而缺乏可比性，进而影响到研究结论的可靠性。这就需要对学习者语料进行等级水平的自动分级，而目前这样的自动分级系

统并不多见，且不对公众开放，难以用到；系统的质量与效能尚有待提高与完善，例如有的系统语料分级的有效性只有 70%（胡韧奋、冯丽萍，2023），远未达到实用水平。因而急需开发优质高效的语料自动分级系统。

语料库中语料的背景信息大多来自学生的入学登记表、成绩登记表之类教学管理文件，有国籍信息而无母语信息，也没有参加 HSK 考试的分数和等级水平。从语料采集的角度看，是需要增加这些背景信息的。

2.2 语料库应用存在的问题

在 CNKI 中通过“主要主题”来看 HSK 库和全球库的使用情况（2025 年 11 月 20 日查询），“偏误分析”“对外汉语教学”“留学生”/“习得研究”位列三甲，表现出研究的基本趋向（详见图三、图四）：依据语料库的研究始终集中在汉语二语教学、偏误分析和习得研究等方面。

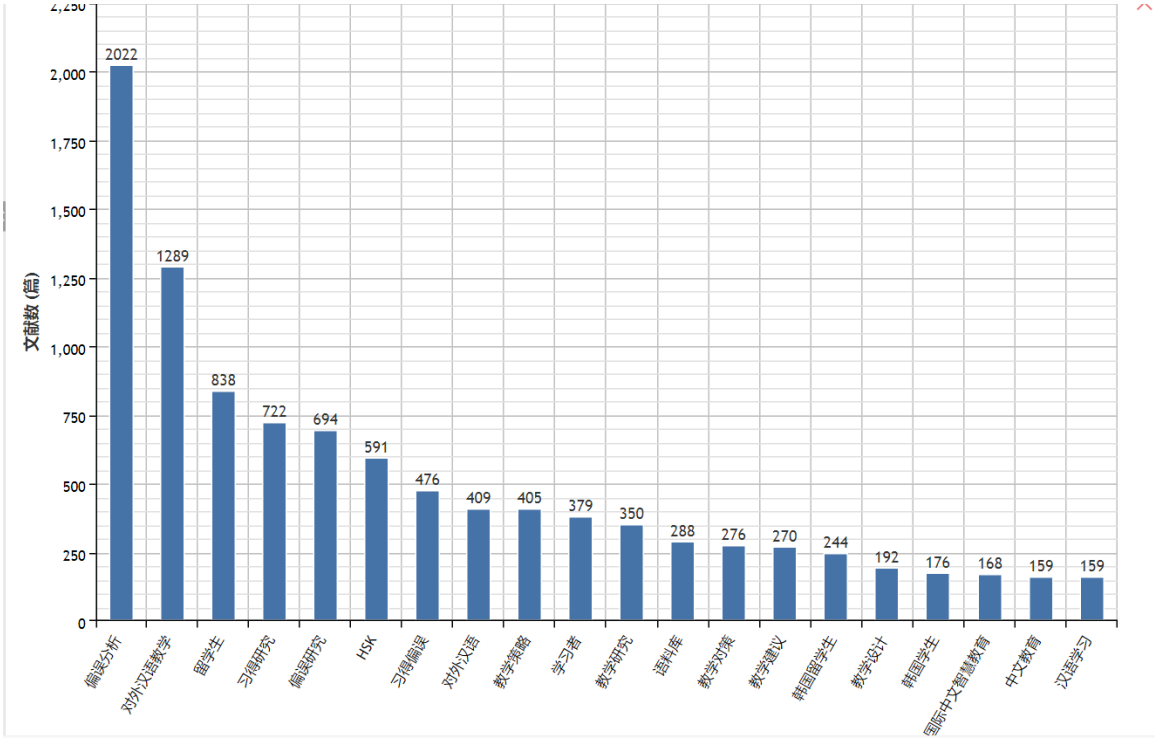


图 3 HSK 库主要主题分布图

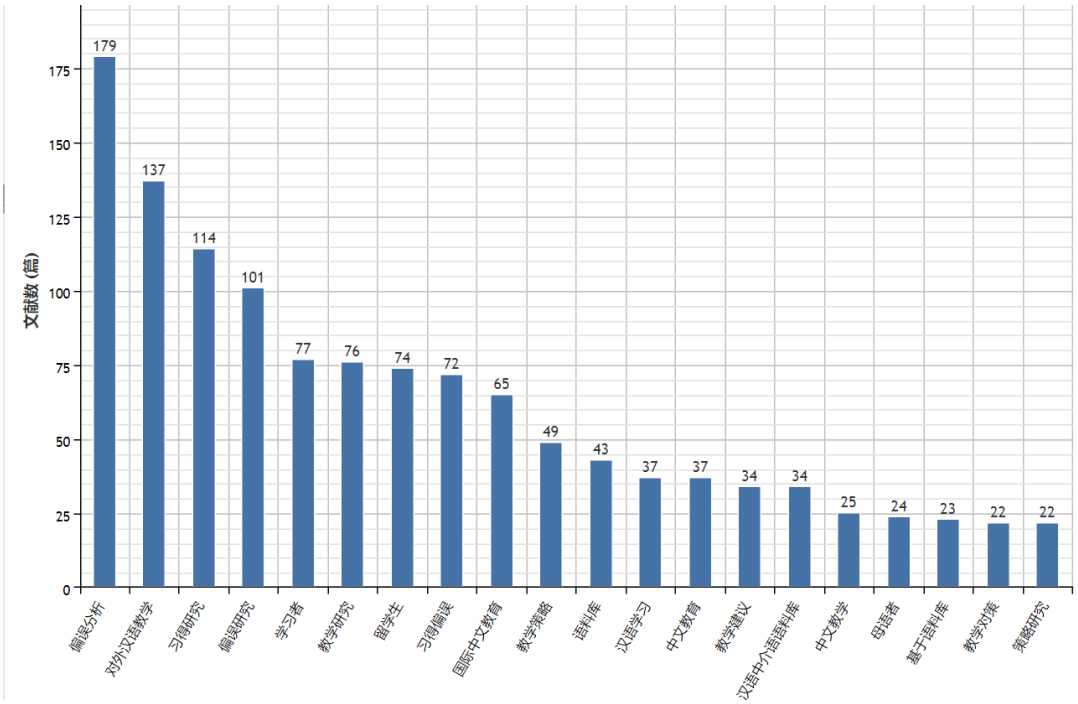


图 4 全球库主要主题分布图

其他相关研究也得出类似的结论，“偏误分析”“偏误”“偏误研究”“习得研究”等“排名均靠前”，“出现的频次最高”。（参见尤易、曹贤文，2022；王立，2022；李娟、谭晓平、杨丽姣，2016；蔡武、郑通涛，2017）

偏误分析无疑是有重要意义的，因为“偏误分析（Error Analysis）是对第二语言习得过程中所产生的偏误进行系统的分析，研究其来源，解释学习者的中介语体系，从而了解第二语言习得的过程与规律。”（刘珣，2000，p191）“偏误分析成为研究学习者习得过程的重要手段和方法，成为观察学习者习得过程的窗口。”（王建勤主编，2009，p37）“错误分析是研究学习过程的捷径，也是研究学习过程的第一步。”（盛炎，1990，p119）“偏误分析法形成了一套颇为有效的分析方法和程序，成为第二语言习得的重要研究方法，直到今天仍具有生命力。”（赵杨，2015，p47）

然而，基于语料库的偏误分析在对偏误进行分类和探讨偏误成因时，基本不会超出遗漏（少成分）、增添（多成分）、替代（所用不当，应用正确的替换下来）、错序（词序有误）等“四大类型”和母语干扰、过度泛化、母语文化干扰、学习策略、教学失误等“五大原因”的范围（参见鲁健骥，1999，p13-14）。这似乎已经成了固定“套路”，甚至不看研究过程都能预测到这样的结果。这就使研究走进了死胡同：研究变成了一种对号入座的过程。带来的问题是：既然偏误类型与产生偏误的原因如此整齐划一、千篇一律，还有什么必要进行这种研究与探讨？而对偏误原因的研究又常常是比较笼统的，非常缺乏具体深入的研究（张宝林，2011）。十几年过去了，这种情况并无根本性的改变，反而有变本加厉的趋势。其实质性的影响是：汉语二语习得研究始终停留在“研究学习过程的第一步”，而未能迈出第二步、第三步，未能深入了解汉语二语习得的全过程，限制了汉语习得研究的进一步发展。

3.破解之道

3.1 应用研究对语料库建设的新需求

语料库的建设目的是为汉语二语教学与研究服务，教学与研究的实际需求是语料库建设的驱动源泉与不竭动力，决定着语料库建设和发展的方向。曹贤文（2020）从二语习得研究“需求侧”角度，提出要加强汉语中介语多维语料库、汉语中介语动态发展语料库、中介语及其影响变量联动数据库、学习者多语发展语料库、汉语学习者网络交际语料库等的建设，以满足“三性/四性”（准确性、流利性和复杂性+多样性）分析、对学习者中介语系统的动态发展轨迹做出完整的描述和解释等研究的需要。这些观点基于语料库建设的现实情况，结合二语研究理论的发展，具有很强的针对性和敏锐的前瞻性，对语料库建设具有十分重要的指导意义。

郑通涛、曾小燕（2016）从大数据视角审视汉语中介语语料库存在的问题，主要包括语料库建设缺乏跨学科视角、缺乏高质量且真实的口语语料资源、语料数据来源存在局限性、缺少建设学习者的历史语料库、语料库数据尚不能充分共享等五个方面。指出在六对十二类语料库中包括单语语料库和多语语料库、不同变体语料库和集母语与二语为一体的语料库。这些认识站在时代发展的高度，反映出相关研究对语料库建设的需求。

李娟、谭晓平、杨丽姣（2016）关于“要注重收录语料层级的平衡性和国别的平衡性。除文本语料外，还需加强学习者语音语料的收集”，“要积极做好自动标注软件的研究开发工作”的见解，王立（2022）关于“共时研究较多，基于语料库的历时研究缺失”的认识，尤易、曹贤文（2022）关于“加强自动评量系统、智能写作评估等方面的建设及研究”的观点，都颇具建设性。

汉语二语习得研究需要理论突破，梁茂成（2018）认为，近年来偏误分析法和中介语对比分析法遇到了前所未有的挑战，而复杂理论（Complexity Theory）和多因素分析（Multi-factorial Analysis）方法将成为中介语语料库研究的新趋势。依据复杂动态系统理论，语言学习的本质是其非线性特点，学习频率是习得获取的主要原因，效果只能在多次重复后被发现（郑通涛，2014）。这为收集连续性语料建设历时的纵向语料库提供了充分的理论根据。

3.2 建设创新型汉语中介语语料库

1）新型语料库是以汉语（中介语+母语）为核心的多语语料库。放眼整个语料库语言学领域，多语语料库虽有，但多为双语，少见三语，罕见多语者；双语语料库或是平行/对应语料库（parallel corpora），或是对比/类比语料库（comparable corpora）。新型语料库将收集学习者产出的汉语中介语语料、学习者产出的和汉语中介语语料同题的学习者母语语料、学习者完成的汉语和其母语的翻译语料，将平行语料库和对比语料库融为一体。这样的多语语料库将为语际迁移研究提供直接证据，不仅在汉语中介语语料库的建设与发展史上尚无先例，在以往的各类语料库建

设中同样没有先例，具有鲜明、突出的创新性。

2）新型语料库是收集学习者连续性语料的纵向语料库。本文所谓连续性语料是指以固定时间长度为间隔周期收集的同一批学习者单位时间内产出的语料。例如以一个月或半个月为间隔周期收集的同一批学习者半年、1 年或数年内产出的汉语语料，最理想的情况是收集同一批学习者从初级阶段到高级阶段或从一年级到四年级的整个本科阶段的所有语料。这样收集到的语料是持续产出的连续性语料，用这样的语料建设的语料库是无可争议的真正的纵向语料库，而非用分层截面数据来取代纵向数据的“伪纵向数据”建设的“类历时语料库”。依据这样的语料库可以观察学习者的二语习得/发展过程与习得顺序，为此类研究提供充足而确凿的证据。

3）新型语料库是语料来源与相关属性均匀的平衡语料库。收集语料应严格遵循平衡性原则。例如学习者（即语料产出者）的国籍、母语、汉语水平等级应确保平衡，不能出现某些国家或水平等级的学习者的语料过多而另外某些国家或水平等级学习者的语料太少的情况（参见张宝林，2022）。学习者国籍是最基本的背景信息，必须具备；否则，收集到的语料再多也是无法使用的。有些国家的语言种类及其分布比较单纯，有些则比较繁杂，没有母语信息同样难以对学习者的二语习得情况进行具体深入的研究。学习者的汉语水平等级同样十分重要，关系到语料的可比性。如果没有清晰可靠的学习者水平等级，就无法对收集到的语料进行具体深入的分类与分析，得到的研究结论必然是含混不清的。

与此相关的问题是，由于语料来源广泛，有些语料可能没有收集到水平等级；有些语料虽然有此信息，但在不同学校、不同汉语教学单位、不同国家学习汉语的学习者，其所谓初级、中级、高级，或一、二、三、四年级的汉语水平等级标签的实际含义可能并不相同，其结果仍然无法进行可靠的对比分析。解决办法有二：其一，在收集语料的同时，从听说读写等方面对学习者进行语言能力测试，由此了解其实际的汉语水平。这个办法有效，但可行性较低。因为面对国内外诸多提供语料的汉语教学单位，逐一进行这种测试并组织专家队伍进行水平鉴定，是非常细致复杂的工作过程，需要大量的人力、财力和时间。其二，通过自动评分系统对收集到的语料进行水平等级评定。这种办法速度快，评定结果的一致性高，也无需投入很多的人力、财力。这种系统目前是有的，只是其准确性不高，尚未达到实用水平，需要进一步研究实验。当其评定的准确性达到 90%时，方可投入实用。

4）新型语料库是检索功能强大、便于用户查询使用的语料库。全球库共有字符串一般检索、按词性序列检索、特定形式检索、搭配检索、对比检索、离合词检索、重叠结构检索、按句末标点、按标注内容检索等九种检索方式，大大提升了对库存语料的查询能力与效率。而新型语料库由于收集了多语种语料和纵向语料，上述九种检索方法还需确保能从纵向（即对同一个/同一批学习者的多篇连续性语料的检索）和多语种（汉语中介语+学习者母语+汉语母语）角度进行检索。这样的检索系统功能强大，可以为用户提供查询语料的极大方便，是在全球库之后具备新的创新性的检索系统。

5）新型语料库是可以增加内容、扩充功能的成长型语料库。学界对语料库的需求是多方面的，有些需求可能是语料库建成之后产生与提出的。因此，新型语料库应具备可扩展性。语料库建成之后，如果需要增加新的语料，添加新的标注内容，拓展新的应用功能，应该都是可以的。这就要求语料库软件系统的研发预做考虑，在架构软件系统的基础框架时预留“扩展槽”，以便后期的扩展应用。

4. 新型语料库的作用与影响

4.1 拓展与深化语料库建设的理论研究

任何一种新型语料库的产生都是需求催生的产物。从研究的角度看，这种语料库的创意是在何种背景下产生与如何形成的？能够满足哪些需求？总体设计是怎样的？建设方式与技术路线是怎样的？多种语料如何放置、对齐、调用与呈现？为什么是这样的？为何如此设计？对上述一系列问题的研究与解答，将极大地推动语料库建设的理论研究。

4.2 为应用研究提供支持

以往的偏误分析、习得研究在讨论偏误成因时，常常是从既有理论出发，把“母语负迁移”作为首要原因。然而这样的结论只是套用现成理论，并无学习者产出的汉语中介语和其母语之间实际语料的具体对比分析，其是否正确难以证明。而新型语料库采集了学习者产出的汉语中介语语料、学习者用其母语所写的与汉语中介语语料同题的对比语料、汉语中介语语料与学习者母语语料的双向翻译语料，这就给母语迁移的证明或证伪提供了语料支持，使其结论更加客观、可信、可靠。

4.3 推动应用研究的转型与发展

以往的纵向研究所依据的多为“类历时语料库（quasilongitudinal corpus）”，其中的语料数据被称为“伪纵向数据”（pseudolongitudinal data），用分层截面数据来取代纵向数据，其有效性充满争议。（Gass & Selinker，2008）因为“类历时语料库有一个基本假设：二语是线性发展的，习得过程是线性渐增的。然而二语发展并非总是连续上升的过程，学习者的进步模式除了线性上升或下降以外，也包括 N 形、Ω 形、V 形、U 形等不同模式（文秋芳、胡健，2010），非线性过程是二语发展的常态”（曹贤文，2020）。可见依据“类历时语料库”进行二语发展研究是不可靠的。新型语料库将“花大力气采集中介语发展过程中的多波纵向数据”，“来支撑相关二语习得研究，尤其是深入考察中介语在时间轴上的变异和变化表现，对学习者中介语系统的动态发展轨迹做出比较完整的描述和解释。”（曹贤文，2020）这将极大地推动汉语二语教学与习得研究从中介语理论指导的偏误分析向复杂动态系统理论指导的汉语二语发展研究转型与发展。

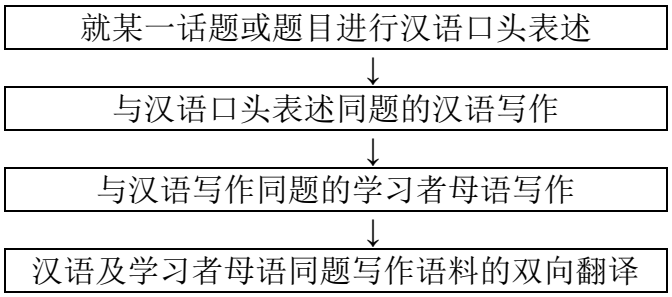
4.4 推动“以效果为导向”的课程改革

建设新型语料库须采集同一批学习者的汉语中介语口语和笔语历时语料、对该语料的学习者母语翻译语料，学习者的同题母语写作语料及其中文翻译语料，以便对学习者的汉语中介语进行包括语体、语言迁移等内容在内的全面而深入的考察。目前需要研究的问题是：这些语料分属汉语中介语、学习者母语、口语、笔语、翻译等不同类型，如何收集这些语料？通过现有课程类型能否收集到这些语料？

目前中国大陆的汉语二语语言技能课的课堂教学主要采用的是“主干课+分技能课”的模式，主干课即所谓精读课、综合课，分技能课包括听力、口语、阅读、写作、翻译等课程。显而易见，建库所需要的不同类型的语料是无法通过现有的任何一门课程来收集的，也许只有改变这种课程体系与类型才能收集到类型多样且密切相关的语料。而这种改变课程体系与类型的想法是否存在现实可行性？是否存在改变的理据？从目前的实际情况看，“主干课+分技能课”的课堂教学模式和课程设置流传已久，根深蒂固，为学界普遍接受，似乎难以改变。然而，这种教学模式和课程设置是否符合语言能力增长的实际过程？其实际效果究竟如何？似乎尚无定论，甚至无人关注。

郑通涛（2014）指出：“从复杂动态系统来解释大脑功能的分区，我们也会发现一个传统误区，即对大脑语言学习的功能分区的认识。以往人们一直简单地认为，左大脑或者右大脑，一部分是偏于视觉功能的处理，另一部分是偏于语言功能的处理。其实，这是一种非常肤浅的说法。因为任何一个功能都是各个部分通过共同协作来实现的，不可能单独运用某部分的功能。大脑的功能全部都是在协同工作中。”既然大脑语言学习的功能是其各个部分协同作用的结果，分技能课的设置似乎就值得探讨，因此并非不可改变。由于“学习各系统循环效果制约语言学习方向，效果是复杂动态系统能维持下去的推力。这要求我们以效果为导向进行教材编写、以效果为导向进行交际能力的重新定义，我们目前的交际能力并不是以效果为导向。此外，还应以效果为导向研究教材法、课堂组织法以及进行教学评估。这种以效果为导向的教学思想转变将挑战目前几乎所有的对外汉语教学领域。”（郑通涛，2014）这就为改革现有课程设置提供了理论依据。

如以翻译课作为课程体系改革的尝试，其基本教学过程是：



经过定期的多波积累，即可得到学习者产出的汉语中介语口笔语语料、翻译语

料、学习者母语语料，解决作为建库前提的语料收集问题。

4.5 推动语料库建设从 2.0 时代向 3.0 时代转变

相比于语料库建设的 1.0 时代，2.0 时代的语料库建设已经获得了长足的进步，然而仍然存在单语种、语料缺少连续性和平衡性、标注方法与质量欠佳等方面的诸多不足，因而需要继续改进。这种改进从语料库的总体设计思路到建库的技术路线，从语料库的基本性质到具体功能，从语料库建设到语料库应用研究，都将是一种质的飞跃，所建设的将是新一代语料库，即 3.0 时代的语料库。它与 2.0 时代语料库的区别主要体现在下列诸方面，详见表 2。

表 2 汉语中介语语料库 2.0 时代与 3.0 时代特征对照表

对照项	2.0 时代特征	3.0 时代特征
语料库性质	横向静态库为主	纵向动态库为主
语料采集	非连续性共时语料	连续性历时语料
语料语种	单语种（汉语）	多语种（依学习者母语而定）
语料类型	多种，非同题	多种，同题
语料平衡性	不严格	严格
语料加工	手工标注为主	AI 大模型自动标注为主
技术路线	重在语料标注	重在检索方式研发
应用研究	中介语理论为主	复杂动态系统理论为主
总体概括	单语种共时静态库	多语种历时平衡动态库

4.6 AI 大模型对语料库建设的影响

以 ChatGPT 和 DeepSeek 为代表的 AI 大模型，正日益广泛应用于各个领域，有望成为人们生活、学习与工作的高效助手。它们在语言生成与理解、逻辑推理等方面的卓越能力，为第二语言学习与研究带来了新的可能。例如，研究者可借助 AI 工具识别中介语语料中的各类偏误，或依据特定标注规则实现语料的自动化标注。

然而也应看到，目前 AI 大模型的语言知识体系尚不完善，对学习者偏误的判断与分类仍不够准确，需辅以人工核查；同时，AI 也难以主动、系统地收集大规模中介语语料，更无法独立总结语言习得规律。因此，AI 大模型尚无法取代汉语中介语语料库的作用，语料库在二语习得研究中仍具有不可替代的价值。

“工欲善其事，必先利其器”。AI 大模型为语料库建设提供了强有力的技术支撑，能显著提升语料处理与构建的效率。在新型语料库的开发中，应充分借助其能力，推动语料库研究向更智能、更高效的方向发展。

5. 结论

汉语习得研究正在由以中介语理论为主导转向以复杂动态系统理论为主导，由以横向的静态研究为主转向以纵向的动态研究为主，走向具体、细致、深入的二语发展研究。针对这一转变，急需建设多语种、纵向、平衡、成长型的动态语料库；建设规模适度、设计精密、标注准确、质量优异、功能丰富的通用型语料库。学界应顺应教学与研究的新需求，改进语料库设计，拓展语料库功能，把语料库建设由2.0 时代推进到 3.0 时代，建设新型语料库，为汉语习得/二语发展研究提供充足的、强有力的语料资源支持。在此过程中，AI 大模型将发挥重要作用，为新型语料库建设增添浓墨重彩的一笔。

Reference

Cai, W., & Zheng, T. T. (2023). The research status and hot topics of Chinese inter-language corpora: A Visualization Analysis Based on CiteSpace. *TCSOL Studies,* 03, 79-87. [蔡武, &郑通涛. (2017). 我国汉语中介语语料库研究现状与热点透视——基于 CiteSpace 的可视化分.*华文教学与研究*，03, 79-87.]

Cao, X. W. (2020). On the construction of Chinese language learner corpus from the perspective of “demand side” of second language acquisition research. *TCSOL Studies,* 01, 38-46. [曹贤文. (2020). 二语习得研究“需求侧”视角下的汉语学习者语料库建设. *华文教学与研究*, 01, 38-46.]

Gass, S. M., & Selinker, L. (2008). *Second language acquisition: An introductory course* (3rd ed.). Taylor & Francis.

Hu, R. F., & Feng, L. P. (2023). L2C-Rater: Research on automatic scoring system for L2 Chinese composition. Plenary report at the 7th International Symposium on the Construction and Application of Chinese Interlanguage Corpus, Shanghai.[胡韧奋，&冯丽萍. (2023). L2C-Rater：*汉语二语作文自动评分系统研究.* 第七届汉语中介语语料库建设与应用国际学术研讨会大会报告，上海。]

Hu, X. Q. (2016). The idea of the construction of multi-dimensional Chinese inter-language corpus network. In Li, X. Q., Jin, X. Z., & Xu, J. (Eds.), *Proceedings of the 10th international symposium on modernization of Chinese language teaching* (pp. 384-389). Tsinghua University Press. [胡晓清. (2016). 多维参照的汉语中介语语料库库群的建立构想. 载于李晓琪，金铉哲，徐娟（主编）.*第十届中文教学现代化国际研讨会论文集* (pp. 384-389). 清华大学出版社.]

Huang, C. N., & Li, J. Z. (2002). *Corpus linguistics*. The Commercial Press. [黄昌宁, & 李涓子. (2002). *语料库语言学.* 商务印书馆.]

Li, J., Tan, X. P., & Yang, L. J. (2016). Study on the application and development of Chinese interlanguage corpus. *Journal of Qujing Normal University*, *35*(2), 86-91. [李娟, 谭晓平, 杨丽姣. (2016). 汉语中介语语料库应用及发展对策研究. *曲靖师范学院学报*，35，02, 86-91.]

Liang, M. C. (2018). *Research on interlanguage corpus: History, challenges, and development trends.* Plenary Report at the 5th International Symposium on the

Construction and Application of Chinese Interlanguage Corpus, Nanjing. [梁茂成. (2018). *中介语语料库研究——历程、挑战与发展趋势*. 第五届汉语中介语语料库建设与应用国际学术研讨会大会报告，南京.]

Liu, K. Y. (2000). *Automatic word segmentation and annotation of Chinese text.* The Commercial Press. [刘开瑛. (2000). *中文文本自动分词和标注*. 商务印书馆.]

Liu, X. (2000). *Introduction to teaching Chinese as a foreign language*. Beijing Language and Culture University Press. [刘珣. (2000). *对外汉语教育学引论*. 北京语言大学出版社.]

Lu, J. J. (1999). *Collection of reflections on teaching Chinese as a foreign language.* Beijing Language and Culture University Press. [鲁健骥. (1999). *对外汉语教学思考集*. 北京语言大学出版社.]

Ren, H, B. (2010). Towards to the construction of the inter-language corpus of Chinese——Using the dynamic corpus of compositions from HSK as an example. *Language Teaching and Linguistic Studies*, 06, 8-15. [任海波. (2010). 关于中介语语料库建设的几点思考：以“HSK 动态作文语料库”为例. *语言教学与研究*，06, 8-15.]

Sheng, Y. (1990). *Principles of language teaching.* Chongqing Press. [盛炎. (1990). *语言教学原理*. 重庆出版社.]

Tan, X. P. (2014). Review of research on Chinese corpus construction in recent ten years. Proceedings of the 7th Beijing Regional Postgraduate Forum on *TCFL* (pp. 26-31). Peking University, Beijing. [谭晓平. (2014). 近十年汉语语料库建设研究综述. 第七届北京地区对外汉语教学研究生论坛论文集 (pp. 26-31). 北京大学，北京.]

Wang, J. Q. (Ed.). (2009). *Studies in second language acquisition.* The Commercial Press. [王建勤(主编). (2009). *第二语言习得研究*. 商务印书馆.]

Wang, L. (2022). A Corpus-based bibliometric analysis of international Chinese language education research papers. *Journal of International Chinese Teaching*, 02, 44-55. [王立. (2022). 基于语料库的国际中文教育研究论文文献计量分析. *国际汉语教学研究*, 02, 44-55.]

Wang, Y. Y., Kong, C. L., Yang, L. E., Hu, R. F., Yang, E. H., & Sun, M. S. (2023). The construction of Chinese multi-dimensional learner corpus: YACLC. *Applied Linguistics*, 01, 88-100. [王莹莹，孔存良，杨麟儿，胡韧奋，杨尔弘，&孙茂松. (2023). 汉语学习者文本多维标注语料库建设. *语言文字应用*, 01, 88-100.]

Wen, Q. F., & Hu, J. (2010). *The patterns and characteristics of oral English competence development among Chinese college students*. Foreign Language Teaching and Research Press. [文秋芳, &胡健. (2010). *中国大学生英语口语能力发展的规律与特点*. 外语教学与研究出版社.]

Yang, H. Z. (Ed.). (2002). *Corpus linguistics*. Shanghai Foreign Language Education Press. [杨惠中(主编). (2002). *语料库语言学*. 上海外语教育出版社.]

You,Y., & Cao, X . W. (2022). Analysis of domestic and international learner corpora construction and applied research in 20 Years. *International Chinese Language Education,* 02, 5-14. [尤易, &曹贤文. (2022). 20 年来国内外学习者语料库建设

及应用研究分析. *国际中文教育（中英文）*, 02, 5-14.]

Zhang, B. L. (2010). The content and method of basic annotation. In Zhang, P., Song, J. H., & Xu, J. (Eds.), *Digitized teaching of Chinese as a foreign language practice and reflection* (pp. 376-382). Tsinghua University Press. [张宝林. (2010). 基础标注的内容与方法. 张普、宋继华、徐娟（主编）.*数字化对外汉语教学实践与反思* (pp. 376-382). 清华大学出版社.]

Zhang, B. L. (2011). On methodology of the Chinese sentences acquisition by foreigners. *TCSOL Studies,* 02, 23-29+45. [张宝林. (2011). 外国人汉语句式习得研究的方法论思考.*华文教学与研究*，02, 23-29+45.]

Zhang, B. L. (2019). From 1.0 to 2.0: The construction and development of Chinese interlanguage corpus. *Journal of International Chinese Teaching,* 04, 84-95. [张宝林. (2019). 从 1.0 到 2.0：汉语中介语语料库的建设与发展. *国际汉语教学研究*, 04, 84-95.]

Zhang, B. L. (2022). Ways to expand the sources of Chinese interlanguage corpus. *International Chinese Language Education*, 7, 02, 30-37. [张宝林. (2022). 扩大汉语中介语语料库语料来源的途径. *国际中文教育（中英文）*,7, 02, 30-37.]

Zhang, B. L., & Cui, X. L. (2015). On the standards of building a Chinese interlanguage Corpus. *Applied Linguistics*, 02, 125-134. [张宝林，崔希亮. (2015). 谈汉语中介语语料库的建设标准. 语言文字应用, 02, 125-134.]

Zhang, B. L., & Cui, X. L. (2022). Features and functions of global Chinese interlanguage corpus. *Chinese Teaching in the World*, 36,01, 90-100. [张宝林，崔希亮. (2022). 全球汉语中介语语料库的特征与功能. *世界汉语教学*, 36, 01, 90-100.]

Zhao Y. (2015). *Second language acquisition*. Foreign Language Teaching and Research Press. [赵杨. (2015). *第二语言习得*. 外语教学与研究出版社.]

Zheng, T. T. (2014). The study of complex dynamic systems in teaching Chinese as a foreign language. *International Journal of Chinese Studies*, 5, 02, 1-16. [郑通涛. (2014). 复杂动态系统与对外汉语教学. *国际汉语学报*, 5, 02, 1-16.]

Zheng, T. T., & Zeng, X. Y. (2016). Construction of Chinese inter-language corpora based on big data. *Journal of Xiamen University (Arts & Social Sciences)*, 02, 53-63. [郑通涛，曾小燕. (2016). 大数据时代的汉语中介语语料库建设. *厦门大学学报（哲学社会科学版）*, 02, 53-63.]

