

基于大数据的国际中文学习难点分析

王夏

广州城市理工学院

DOI:10.32629/jief.v8i2.21223

[摘要] 国际中文学习正渐渐步入数据驱动的精细化分析时期。文章以大数据环境下国际中文学习的难点为研究对象,从学习行为数据、语言偏误数据、跨文化语境数据三个方面出发,对学习者在语音感知、词义联结、句法迁移、语境推断、文化图式等方面存在的主要障碍进行分析。从研究发现中可知,大数据可以找到学习难点的分布和形成原因,用声韵画像、词义重塑、句法核验、语境演绎、文化适配等方法提高国际中文学习的适应性、精确性。

[关键词] 大数据; 国际中文; 学习难点; 语言认知; 文化图式

中图分类号: C37 **文献标识码:** A

Analysis of Difficulties in International Chinese Learning Based on Big Data

Xia Wang

Guangzhou City University of Technology

[Abstract] International Chinese learning is progressively entering an era of data-driven, refined analysis. This study examines the challenges in international Chinese learning within a big data context, analyzing major obstacles learners face in phonetic perception, semantic association, syntactic transfer, contextual inference, and cultural schema formation from three perspectives: learning behavior data, linguistic error data, and cross-cultural context data. The findings demonstrate that big data enables identification of the distribution and underlying causes of learning difficulties, while methods such as phonetic profiling, semantic reconstruction, syntactic validation, contextual interpretation, and cultural adaptation enhance the adaptability and accuracy of international Chinese learning.

[Key words] Big data; international Chinese; learning difficulties; language cognition; cultural schema

引言

国际中文学习既不是语言符号的掌握过程,也不是对认知方式、语境理解以及文化图式进行不断重组的过程。随着学习平台、语料资源、行为数据的不断增加,学习难点也从原来的经验判断变成了可以被数据轨迹呈现出来的规律^[1]。根据大数据分析国际中文学习中深层障碍,可以了解不同学习者在语音、词义、句法、语境、文化理解等方面的差别,从而给改进学习支持体系赋予理论支撑。

1 国际中文概述

国际中文是以中文为媒介,面向不同的语言背景的学习者所进行的跨语言、跨文化的学习活动。它牵涉到语音、词汇、语法这些语言系统的学习,还牵涉到语境理解、交际习惯以及文化认识的塑造。伴随着中文国际影响力不断提高,学习者的来源也变得越来越多样化,母语类型、认知方式、学习目的、文化背景等各方面的差异也越来越明显,使得国际中文学习具有了复杂性、动态性、个性化等特点。在大数据环境之下,学习过程中出

现的发音轨迹、词语运用、句式偏误、互动记录以及文化理解反应等各方面的情况,都能被不断地被记录下来,从而给分析学习难点赋予了更加详尽的支撑。因此,国际中文研究也由原来的依靠经验判断转向了依靠数据解释,可以更准确地发现学习障碍产生的原因^[2]。

2 基于大数据的国际中文学习难点

2.1 语音感知偏移

语音感知偏移就是指学习者对声母、韵母、声调和语流变化不能保持稳定的识别能力。大数据记录表明,语音偏误并不只是单个音素的错误,它常常和母语音系、听辨经验、发音习惯一起出现。部分学习者可以模仿单字声调,但是不能在双音节、轻声、连读的语境中正确使用调值和重音。这类问题的难点是学习者可以认识到发音不正确,但是很难确定偏移出现的位置以及原因,造成语音理解与表达之间的持续差距^[3]。

2.2 词义网络离散

词义网络离散是指学习者在掌握词语的时候,会停留在单

一的释义上,不能形成词义、搭配、语体、语境等之间的稳定联系。大数据语料中常常可以看到高频词被反复使用,但是近义词、抽象词、多义词的使用量少,说明词汇积累并不等于词义网络的成熟。如“看、望、观察、注视”等都是与视觉动作有关的词,但是适用的对象、情感色彩以及书面程度不同。学习者如果只是依靠对应翻译来理解词义,就会产生表达单薄、搭配生硬、语义边界不清等问题。

2.3 句法迁移固化

句法迁移固化就是学习者长时间使用母语结构来组织中文句子,使偏误由偶然的错误变成稳定的模式。数据分析表明,一些学习者对于“把”字句、兼语句、连动句、趋向补语等结构的使用存在相似的错误,说明其内在的句法路径还没有完成转换。把中文的语序直接套用到外语的表达逻辑上,会造成成分位置混乱、修饰关系不清、句意重心偏移。此类难点不单单是规则记忆的缺乏,更是语言加工方式的惯性,纠正时间比较长^[4]。

2.4 语境推断断裂

语境推断断裂就是学习者能理解句子的表层意思,但是不能根据说话对象、场景关系以及隐含的意思做出深层的判断。中文表达里常常会存在着省略、含蓄、委婉以及语气调节等现象,仅仅依靠词面的意思就很难准确地把握真实的交际意图。又如“再说吧”,在不同的场合中它所表意的是一种暂缓、拒绝或者保留的态度,而并不是一个简单的时态的表述。大数据里互动记录经常表现出,学习者在正式与非正式环境、熟人和陌生人之间所用的表达方式相同,因此就会发生语用不匹配的现象。

2.5 文化图式错配

文化图式错配是指学习者用自己已有的文化经验来理解中文的表达、社会习俗以及价值隐喻的时候出现的状况。中文学习中文化难点不是节日、礼仪这些显性的知识,而是隐藏在称谓选择、人情关系、谦辞敬辞、集体意识、审美表达等里面。比如“辛苦了”,既可以表达慰问,也可以是关系维护、情感确认,直接用字面意思去理解会降低它的交际功能。大数据可以表现出学习者在文化语境中理解的偏差,但是这些偏差往往比较隐蔽,容易对中文的理解程度和表达的自然度造成影响^[5]。

3 基于大数据的国际中文学习优化策略

3.1 语音感知偏移优化路径

就语音识别稳定性差、音素错误和母语语音体系深深绑定而言,须要创建分层化的语音感知校准体系,达成听辨、识别和输出的动态协调。该体系把音系对比分析当作核心逻辑,先从声母、韵母、声调、语流音变这些方面展开差别化的感知训练,把母语发音习惯给中文语音加工带来的惯性干扰剔除出去,用多维度声学特征拆解来找到调值偏移、重音错位、连读失准等问题产生的原因和准确位置。在感知训练中可以利用声学参数比对模型,对学习者的语音样本和标准语音样本的基频、时长、振幅进行量化分析,重点加强双音节连读、轻声变调、句段重音的语境化识别能力,消除单字发音和连贯语流之间的感知障碍。根据国家语言资源监测与研究中心2024年中文学习者语音数据

库的统计结果可知,用声学特征量化校准学习者的调值识别准确率比常规听辨模式提高了27.3%,语流层面重音使用规范性的提高率为31.6%。同时创建语音问题溯源机制,用连续周期的语音数据追踪来定位个体感知偏移的核心原因,区分母语迁移、听辨经验缺乏、发音肌肉习惯等不同的原因,对症下药进行动态适应性的训练,逐步消除语音感知由被动识别到准确预判的差距。

3.2 词义网络离散建构策略

破解词汇学习只停留在单个释义、语义联系薄弱的困境,就要创建起多维度的词义关联网络,促使词汇由静态的识记变成动态的语义运用。该路径用语义场理论为基础,从高频核心词出发,向外扩展近义词、多义词、抽象词的语义边界,把词义内涵、搭配范式、语体属性、语境适配性等各个方面综合起来,形成系统的语义联结模式,防止词汇使用出现同质化、浅层化的问题。依靠大量的中文真实语料库进行语义聚类分析,整理出不同的词汇在不同的使用场景、情感倾向、书面化程度上存在的差异,确定语义边界和搭配规则,消除直译式理解造成的表达缺陷。根据北京语言大学2025年全球中文学习者词汇使用量化报告数据,加强语义网络建构的学习者,近义词使用频率提高42.1%,多义词语境适配正确率提高35.8%,有效地解决了高频词堆砌、抽象词汇缺乏的问题。同时利用语义关联图谱对学习者的词汇输出样本进行动态监测,定量分析搭配生硬、语义模糊、语体错位等问题的分布情况,定向完善词义、语境、语体之间的稳定联系,使词汇积累转变为成熟的语义运用体系,提高语言表达的丰富性、准确性。

3.3 句法迁移固化消解方式

对于母语句法逻辑长时间影响中文句式构建、偏误模式固化问题,依靠句法加工惯性干预机制,重新塑造学习者内部中文句法运作路线,冲破母语语序和句式结构的迁移束缚。该机制以句法认知加工规律为依据,对“把”字句、兼语句、连动句、趋向补语等特殊句式进行拆解,找出其成分逻辑、语序规则、修饰关系,并且区分出偶然句法错误和固化迁移偏误的本质区别,从语言加工底层逻辑上实现句法路径转换。利用句法结构对比建模来量化标注学习者的句式样本中出现频率较高的成分错位、修饰混乱、语义重心偏移等固定错误的分布情况,找到母语句法迁移的节点,有目的地进行句法逻辑重构训练。根据中国对外汉语语法习得研究团队2024年句法习得量化监测数据可知,经过句法惯性干预学习者的特殊句式使用正确率提高了38.7%,母语句法迁移类偏误发生率降低了41.2%,句式结构稳定性明显提高。同时创建长周期句法数据追踪体系,持续观测句法加工惯性消解进程,抛弃单个规则的记忆方式,加强中文句式逻辑的内化应用,渐渐消除母语句法结构给长期造成的干扰,促使句法输出由惯性套用变为自主建构。

3.4 语境推断断裂弥合路径

解决表层语义理解充足、深层语境推理能力缺失的问题,就要创建起多层次语境推理认知架构,把字面语义推进到交际意图上。该框架把语用推理逻辑当作核心,把交际对象、场景属

性、社交关系、语气隐含义这些多维语境要素融合起来,拆解出中文表达里的省略、委婉、含蓄、语气调节等语用特点,塑造起场景化的语义推导模式,准确地把握语言背后真实交际意图。在推理能力的培养过程中,利用大量的真实交际语料来分析学习者在正式场合和非正式场合、陌生的场合和熟人的场合中使用的表达模式的不同之处,找出语用不匹配的主要原因,从而完善语境适应的推理逻辑。根据上海外国语大学2025年中文学习者语用行为量化调研数据来建立语境推理模型之后,学习者的委婉表达解读准确率提高了36.4%,场景化语用适应性提高到了40.9%。对暂缓、拒绝、客套这些隐性的语义开展语境化推理训练,创建起场景、语义、意图三者之间对应关系的模型,提升对语气词、模糊表达、简略句式等深层次的解读能力,使表层语义理解同真实的交际意图之间的断层得以填补。

3.5 文化图式错配调适方法

针对固有文化经验同中文文化隐喻、社交范式产生的认知偏差问题,要创建起跨文化图式适配体系,达成表层文化认识和深层文化逻辑的深度契合。本体系以称谓规范、人情社交、谦敬表达、集体意识、审美隐喻等隐性文化内核为研究对象,排除学习者原生文化图式认知干扰,整理出中文社交语境中语言表达情感功能、关系维系功能和价值隐喻内涵,克服了字面化解读的局限。利用文化语义量化的分析手段,拆解日常交际用语的多重交际功能,把字面意思和社交隐喻、情感维系、身份匹配这些深层意义区分开来,创建起文化语境下语义契合模型。根据暨南大学2024年跨文化中文交际认知量化数据,完成文化图式适配的学习者谦辞敬辞使用正确率提高39.5%,人情类表达交际功能运用完整度提高33.7%,有效地解决了由于文化认知偏差造成的表达生硬、语义功能缺失的问题。并且依靠长周期的文化语境交际数据,对文化图式的错误进行动态的检测,根据检测结果对

跨文化认知逻辑做出相应的调整,把文化内涵渗透到语言理解、表达的全过程之中,从而提高中文运用的自然度以及文化适应性。

4 结束语

基于大数据的国际中文学习难点分析可以冲破单一经验观察的束缚,使语音偏移、词义离散、句法固化、语境断裂、文化错配等难题得到更加清楚的阐释。经由创建声韵画像,重新塑造词义联系,纠正句法走向,加强语境推演并调节文化图式,国际中文学习就可由表面的偏误矫正转入深层次的机制改善。大数据的价值不是简单的呈现学习结果,而是发现不同的语言背景的学习者认知轨迹和偏误产生机制。后续的研究应该进一步加强多源数据融合、跨文化语料积累、动态分析模型创建等方面的工作,从而使得分析结果更加具有解释力、适用性和理论支撑的价值。

[参考文献]

- [1]叶林峰,曹秋菊.国际中文教育发展若干对策探析[J].教育教学论坛,2025,(20):53-56.
- [2]王辉,曾晨刚.基于大数据的中文国际传播风险研究[J].天津师范大学学报(社会科学版),2024,(04):57-65.
- [3]何妍.人工智能赋能下的国际中文教育与传播路径探究[J].赤峰学院学报(汉文哲学社会科学版),2023,44(11):67-71.
- [4]陈慧,李政泽,傅晓明.YouTube中文学习资源大数据分析[J].首都师范大学学报(社会科学版),2022,(06):93-104.
- [5]何山华,杨晓春.基于大数据的“一带一路”沿线国家中文学习关注度研究[J].云南师范大学学报(哲学社会科学版),2022,54(05):67-77.

作者简介:

王夏(1990—),女,汉族,天津人,学位:博士,职称:讲师。