科技与中文教学

# Journal of Technology and Chinese Language Teaching

Volume 16 Number 2, December 2025
二〇二五年十二月　第十六卷第二期



ISSN: 1949-260X
http://www.tclt.us/journal

科技与中文教学

**Journal of Technology and Chinese Language Teaching**

*A peer-reviewed online publication with in-print supplement*
*ISSN: 1949-260X http://www.tclt.us/journal*

## Volume 16 Number 2, December 2025

Managing editor for this issue: Jun Da

## Articles



## Columns

科技与中文教学

**Journal of Technology and Chinese Language Teaching**

*A peer-reviewed online publication with in-print supplement*
*ISSN: 1949-260X  http://www.tclt.us/journal*

## Sponsor

Department of World Languages, Literatures, and Cultures, Middle Tennessee State University

## Editorial board



## Editorial staff

## Contacts

URL: http://www.tclt.us/journal
Email: editor@tclt.us

Journal of Technology and Chinese Language Teaching Volume 16 Number 2, December 2025
http://www.tclt.us/journal/2025v16n2/xusunaoka.pdf pp. 1-22

# 数据驱动的初级汉语课堂优化研究：语言使用比例可视化与生成式 AI 的结合
# (A Data-Driven Approach to Optimizing Beginner Chinese Classrooms: Integrating Language-Use Proportion Visualization with Generative AI)

徐勤
(Xu, Qin)
京都大学
(Kyoto University)
xu.qin.4f@kyoto-u.ac.jp

砂冈和子
(Sunaoka, Kazuko)
早稻田大学
(Waseda University)
ksunaoka@waseda.jp

**摘要**：本研究以日本的大学初级汉语课堂的两段教学录音为样本，利用我们自主开发的语音可视化应用 Voice-to-Text App 进行分析。结果显示，该应用在处理中日语码混合的课堂录音时，具有较高的转写准确率与操作便利性，能迅速识别课堂中的 L2（汉语）使用比例。与传统的课堂语言行为评估框架相比，该系统在效率、可操作性与教师自主分析能力方面均表现出显著优势。在此基础上，研究将 APP 生成的数据结合两种类型的提示词（prompt），输入 ChatGPT，由生成式 AI 提供课堂改进建议，以探讨 AI 在汉语课堂设计中的潜在优势与应用局限。结果发现，AI 能够在语言形式重组与表层推理层面提出较为合理的课堂优化方案，但由于缺乏深层认知能力与创新语法概念的生成能力，难以提出具有启发性的教学设计。

**Abstract**: This study investigates the optimization of beginner-level Chinese language classrooms in Japan through a data-driven approach that integrates speech visualization and generative AI. Classroom recordings were analyzed using a self-developed voice-to-text app, which automatically transcribes classroom recordings of mixed Chinese-Japanese classroom discourse and then visualizes the proportion of L2 (Chinese) use. The app demonstrates high transcription accuracy and operational convenience, offering significant advantages in efficiency, usability, and teacher autonomy compared with traditional classroom language analysis frameworks. Based on the app-generated data and two types of instructional prompts, ChatGPT was used to generate feedback and suggestions for classroom improvement, in order to explore how AI could be applied in Chinese language class design and identifying its limitations. The findings reveal that while AI can effectively propose revisions at the linguistic and surface-level reasoning stages, it lacks deeper cognitive and creative capacities necessary for generating pedagogically insightful designs.

**关键词:** 数据驱动的汉语教学，课堂优化，语音转写 App，语用比例可视化，生成式人工智能

**Keywords:** Data-driven Chinese language teaching, Classroom optimization, Voice-to-Text App, Visualization of L1/L2 usage ratio, Generative AI

## 1. 引言

### 1.1 研究目的与背景

经济合作与发展组织（OECD）在 Learning Compass 2030 中提出以"学习者能动性（student agency）"为核心的教育理念，主张学习是学习者主动建构意义的过程，而非被动接受知识。知识与技能被重新界定为"认识论知识（epistemic knowledge）"，即学习者能理解、反思并应用的能力（OECD，2019 a；OECD，2019 b）。该框架促使教育从"知识传递"转向"能力建构"，强调学习过程的动态性与反思性。

然而，日本的大学第二外语教学仍主要采用以语法大纲为中心的"知识传递型教学"模式。课堂通常由教师主导，学生缺乏使用目标语的机会。受笔试评价体系的影响，教学往往重形式而轻意义，第一语言使用比例偏高，导致学习动机与成效均不理想。近年来，日本文部科学省的多次问卷调查结果[1]显示，超过半数的大学生对外语学习成效持否定态度，由此引发了对学习意义与教学方法的持续质疑（文部科学省，2020；2022；2023）。

本研究开发了一款语音可视化应用 ——Voice-to-Text App（以下简称 APP），可自动转写课堂语音并可视化呈现 L1/L2 的使用比例，从而帮助教师进行课堂语言使用的自我诊断与反思。此外，将 APP 生成的结果输入生成式 AI，由 AI 对课堂互动进行分析并提出改进建议，使教师借助数据驱动的外部反馈，实现持续的自我省察与教学重构。

本文第二节介绍 APP 的功能，第三节展示课堂应用效果，第四节探讨与 AI 结合的可行性，第五节总结研究结果。

[1] 该调查的有效回答率不足一成，样本代表性相对有限，但调查对象覆盖了日本全国范围内的国公立与私立大学及短期大学的学生。由于该调查由国家教育机构组织实施，其结果具有较高的权威性，因而可能会对今后的外语教育政策产生一定影响。

### 1.2 传统课堂语言行为评估框架及其局限

为改进外语课堂教学而广泛采用的分析框架主要包括：FLINT（Foreign Language Interaction：Moskowitz, 1971）、FOCUS（Foci for Observing Communications Used in Settings：Fanselow, 1977）、和 COLT（Communicative Orientation for Language Teaching：Spada & Fröhlich, 1995）等。这些框架通过系统分析课堂语言行为，帮助外语教师进行自我诊断、改进教学策略，并为教师培训提供参考。

然而，欧美开发的课堂分析系统多以交际导向型课堂为前提（Patsy & Spada, 2021），在以读写活动为主的日本外语课堂中，其适用性仍受限制（飯野厚，2009；飯窪真也等，2020）。此外，传统的课堂分析框架通常要求对整节课堂进行录像与转写，并依据预设类别进行人工编码。此过程不仅耗时费力，还需要一定的专业知识，因而难以实现实时分析。例如，COLT 原版包含 13 个主要类别及 40 余个子类别（Spada & Fröhlich, 1995）；FLINT 系统则以 15 个功能类别评估教师语言与反馈的教育功能（Moskowitz, 1971）。尽管这些框架结构严谨，但因分析成本过高，仍难在教师与研究者群体中被广泛推广（Pellerin et al., 2024）。作为数字化应对方案，Mobile COLT 等工具相继出现（石塚博規等，2021）。然而，此类工具仍依赖专业分析与大量时间投入，难以满足日常教学改进需求。因此，亟需开发能有效降低分析负担、使教师便捷且实时获取课堂分析结果的新型工具。

## 2. APP 的开发目的与历程

近年来，自动语音识别（Automatic Speech Recognition，ASR）技术的进步显著降低了语音转写的成本与时间，为缓解上述问题提供了新的契机。

Ferraro et al.（2023）采用单词错误率（WER：word error rate[2]）评估了多种开源 ASR 工具（如 Mozilla DeepSpeech[3]、Conformer[4]、HuBERT[5]、SpeechBrain[6]、WhisperX[7]、SpeechStew[8]）与商业 ASR 服务（如 Amazon Transcribe[9]、Microsoft

[2] WER 衡量的是模型转录错误的单词占参考文本总单词数的百分比，详见 Ferraro et al. (2023): “[T]he metric measures the percentage of words that are incorrectly transcribed by the model relative to the total number of words in the reference transcript”。

[3] 工具信息详见：https://github.com/mozilla/DeepSpeech。

[4] 工具信息详见：https://github.com/sooftware/conformer。

[5] 工具信息详见：https://github.com/facebookresearch/fairseq/tree/main/examples/hubert。

[6] 工具信息详见：https://speechbrain.github.io/。

[7] 工具信息详见：https://github.com/m-bain/whisperX。

[8] 论文参考：https://doi.org/10.48550/arXiv.2104.02133。该工具目前尚未公开官方代码。

[9] https://aws.amazon.com/transcribe/。

Azure Speech to Text[10]、Google Cloud Speech API[11]、IBM Watson Speech to Text[12]）在七个常用数据集上的语音转文本的性能，其结果表明，在大多数评测数据集上（包括 LibriSpeech，CommonVoice，WSJ 和 CHiME），开源 ASR 工具的表现优于商业 ASR 服务[13]。

商用 ASR 服务通常采用按处理分钟数计费或按需付费的定价模式，长期或大规模使用会产生较高的经济成本，且个性化定制能力有限。而开源 ASR 工具通常可供用户免费使用、成本低，并支持个性化定制和本地部署。其中，以 Whisper 为代表的开源 ASR 模型不仅支持多语种识别与翻译，还能自动检测语种并生成带时间戳（timestamp）的转写文本，在准确率和跨语言性能上已超越了大部分商用 ASR 服务。该模型以 30 秒音频片段为单位进行训练，其长音频转写策略是通过对连续 30 秒的音频片段进行逐段转录，并结合模型预测的时间戳信息进行窗口平移，从而实现对长音频的高效转写（Radford et al., 2023）。鉴于日本汉语课堂录音语料中频繁出现中日语码切换，以及课堂语言分析的实际需求，本研究最终选择 Whisper 作为核心语音转写模型，并将其集成至自主开发的 APP 中。

### 2.1 针对多语码转换的自动语音识别模型

语码转换（Code-switching，CS）是指在同一段话语中交替使用两种或两种以上语言的现象（Mustafa et al., 2022）。近年来，在外语教育中，学者们将学习者母语（L1）的语码转换（CS）视为教师互动的资源，并对其有效性进行评价（Macaro, 2009；Myers, 2002）。但日本的汉语课堂过度依赖母语（L1），从而导致提升（L2）语言运用能力的教学目标难以实现（砂岡和子等，2023a），即便是在英语课堂中，这一问题依然存在（田崎敦子，2006）。另外，多语码转换仍然对识别的准确率构成挑战，并且后续的编码环节仍存在较高的技术门槛（砂岡和子 & 徐勤，2023b）。

Whisper 是 OpenAI 于 2022 年 9 月发布的多语言自动语音识别（ASR）模型，提供 tiny、base、small、medium、large 五种规模。模型规模越大，识别精度越高，计算资源与处理时间的需求也相应增加。此后，OpenAI 于 2022 年 12 月推出 large-v2，并在 2023 年 11 月发布性能进一步提升的 large-v3。徐勤 & 砂岡和子（2024）将 large-v3 与“Pyannote.audio”[14] 的话者分离功能结合，成功应用于包含中日语码转换的汉语课堂录音，实现了文本转写与说话人区分，为多语码课堂的语音数据分析提供了有价值的范式。

[10] https://azure.microsoft.com/en-us/products/cognitive-services/speech-to-text/。
[11] https://cloud.google.com/speech-to-text。
[12] https://www.ibm.com/cloud/watson-speech-to-text。
[13] 详见 Ferraro et al. (2023): “[O]ur analysis also highlights that open-source solutions outperform paid services for most datasets, including LibriSpeech, CommonVoice, WSJ, and CHiME.”
[14]可参考项目主页: https://pyannote.github.io/pyannote-audio/

2024 年 10 月，OpenAI 又发布了 large-v3 的优化版本——Whisper large-v3-turbo（以下简称 “turbo”）。与 large-v3 相比，turbo 在识别精度略有下降的同时显著提升了转录速度[15]。在此基础上，砂岡和子 & 徐勤（2025）开发了基于 Whisper 多版本（tiny、base、small、medium、large-v2、large-v3、turbo）的语音转写 Web 应用——APP[16]（见**图 1**）。该应用基于 Python 与 Flask 构建，支持音频上传、多模型选择与时间戳转写，并集成 Matplotlib 进行可视化处理，可自动统计并展示课堂中汉语（L2）与日语（L1）以及教师授课时偶尔出现的英语的使用比例（见**图 3**，**图 4**），结果可供用户下载并自动保存为 Word 文档。

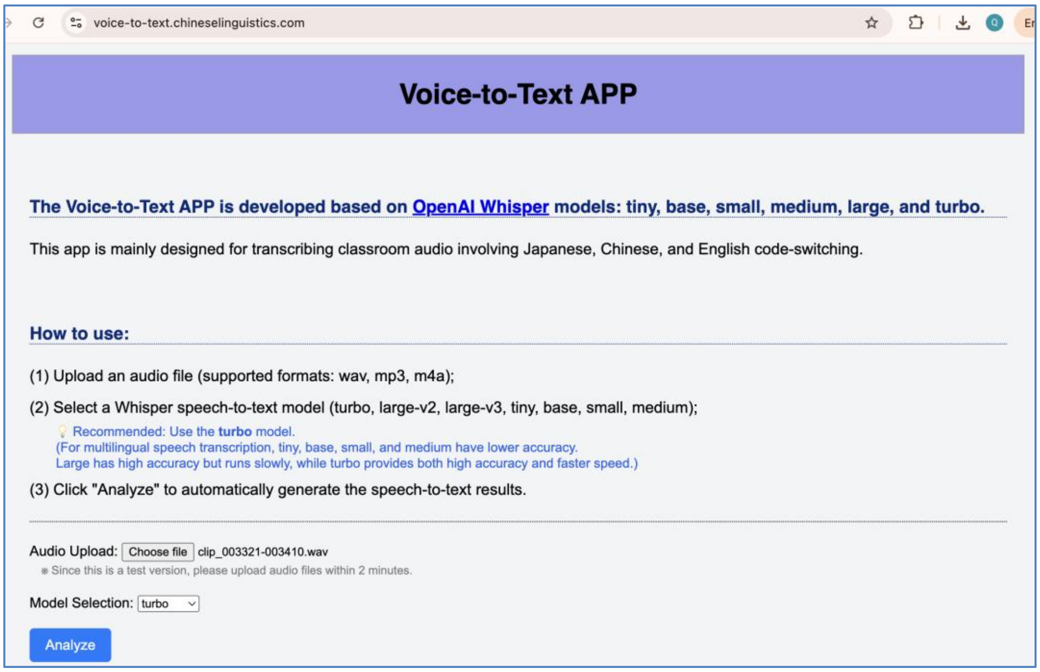


**图 1 Voice-to-Text APP 的页面**

## 3. 基于可视化分析的课堂流程优化研究

### 3.1. 通过 APP 实现课堂发话的可视化

与传统外语课堂观察工具相比，该系统在自动化与效率上具有显著优势。仅需数秒即可将发话音频转写为文本，识别精度可达 90% 以上（徐勤＆砂岡和子，2024；砂岡和子＆徐勤，2025）。其操作无需编程能力与专业知识，也无需对外公开教学内容，即可实现近实时的课堂自我诊断，在保障隐私与降低技术门槛的前提

[15] 可参考 large-v3-turbo 的测评结果：https://medium.com/@bnjmn_marie/whisper-large-v3-turbo-as-good-as-large-v2-but-6x-faster-97f0803fa933。

[16] 该 APP 目前仍处于测试阶段，目前公开测试版可通过以下网址访问：https://voice-to-text.chineselinguistics.com/。后续功能及界面尚有进一步调整与完善的可能。

下，减轻教师与学生的认知负荷，并支持教师开展自主、持续且常态化的课堂分析与教学改进。

若依托 FLINT 系统把时长 60 分钟的课程进行分析，则需要一名具备中日双语能力的研究助理耗时约 4 个月、累计工作时间超过 300 小时（曲明＆砂岡和子，2024）。**图 2** 展示了基于 FLINT 分类框架，对 2022 年 1 月 11 日于日本某大学开设的一年级初修汉语课程（以下简称 [2022 课堂]）的片段进行人工分析的结果（原始音频 0:34:10–0:37:07，约 3 分钟）。尽管人工转写在精确度上或许优于自动转写，但其所需时间至少为自动转写的二十倍，分析成本极高。相比之下，本研究开发的 APP 无需外部协助，教师即可在课堂结束后即时生成并查看分析结果，在效率、可操作性及教师自主性方面均具有显著优势。此外，APP 具备自动区分母语（L1）与目标语（L2）的功能，其性能明显优于 FLINT 和 COLT 系统。

| 開始 | 結束 | 経過時間 | 時長 | 行為分類 | 發言人 | 参與方式 | 轉寫文本 |
|---|---|---|---|---|---|---|---|
| 0:34:10 | 0:36:36 | 0:02:26 | 146 | 321 | 教師 | 線上＋線下 | はい、これは理解しやすいところだと思いますけれども、「誰々がどこどこにいる」「何何がどこにある」という文ですので、主語は人間、もしくは物になりますよね。それから、後ろは在です。で、これから後ろは場所がきます、場所は地名のときもあれば、名詞が場所として使われることもあります。@@@とか@@とかは地名ですよね、(中略) 名詞は場所として使うとき、先も簡単に説明しましたんですけれども、桌子は本来では、机で、物の名前ですけれども、これは場所として使われる時は、上もしくは裡どれか一つね、(中略) 中国語は、書くときは、名詞が場所として使われる時、この名詞の後ろに上もしくは裡をつける必要があります。ここは覚えてくださいね。で、これを、復習する、復習というか、これさらに確認するために、一番下に、「名詞の場所化」と言うところがあるので、こういうように使えますよ。フレーズ、名詞と裡、名詞と上、一緒に使う例があげられています。 |
| 0:36:36 | 0:36:42 | 0:00:06 | 6 | 311 | 教師 | 線上＋線下 | では、じゃあ、一人一個づつね。まず、「冷蔵庫の中」を読んでください。 |
| 0:36:42 | 0:36:45 | 0:00:03 | 3 | 310 | 教師 | 線上＋線下 | 山本 大郎！ |
| 0:36:45 | 0:36:48 | 0:00:03 | 3 | 314 | 教師 | 線上＋線下 | 沉默 (老師等待學生回覆的時間) |
| 0:36:48 | 0:36:52 | 0:00:04 | 4 | 310 | 教師 | 線上＋線下 | 山本 さんいますか。いないようで |
| 0:36:52 | 0:36:55 | 0:00:03 | 3 | 310 | 教師 | 線上 | いないようで、鈴木和夫。 |
| 0:36:55 | 0:36:58 | 0:00:03 | 3 | 330 | 學生* | 線上 | はい。 |
| 0:36:57 | 0:37:00 | 0:00:03 | 3 | 310 | 教師 | 線上 | 山本さん、山本さんね |
| 0:37:00 | 0:37:03 | 0:00:03 | 3 | 311 | 教師 | 線上 | はい、「冷蔵庫の中」を読んで欲しい。 |
| 0:37:03 | 0:37:06 | 0:00:03 | 3 | 330 | 學生** | 線上 | 冰箱裡。 |
| 0:37:06 | 0:37:07 | 0:00:01 | 1 | 313 | 教師 | 線上 | 非常 好ね，冰箱裡ね。 |

**图 2 FLINT 分析例[17]（部分）**

下文选取[2022 课堂]中的一个发话片段（实例 1，简称[2022 课堂]），以及另一所日本的大学汉语课堂发话片段（实例 2，简称[2025 课堂]），将二者导入 APP 进行分析[18]。

## 3.2. 实例 1：[2022 课堂]发话内容的语音转写

实例 1 为[2022 课堂]中，教师总结当日语法重点的片段。APP 将原始音频（00:52:01-00:52:35，共 34 秒）按时间戳（如**图 3** Transcription with timestamps）

[17] “行为分类”系指事先设定的发话类别。[321] 等编号为分析者对各类别所赋予的编号。“参与方式”：由于之一堂课为混合授课形式，故区分为“线上（远程参与）”与“线下（课堂参与）”。

[18] 截至 2025 年 11 月 5 日，目前的 APP 测试版每次可处理的音频文件上限约为 2 分钟，因此在分析前需对原始录音进行剪辑。

自动转写为文字，并展示课堂中的语种使用比例（Language Ratios）。ASR 技术将课堂中的发话细节真实地转写为文字。虽然极短的语气词有时可能被忽略，但整体上几乎所有语音内容都得到了准确转写。

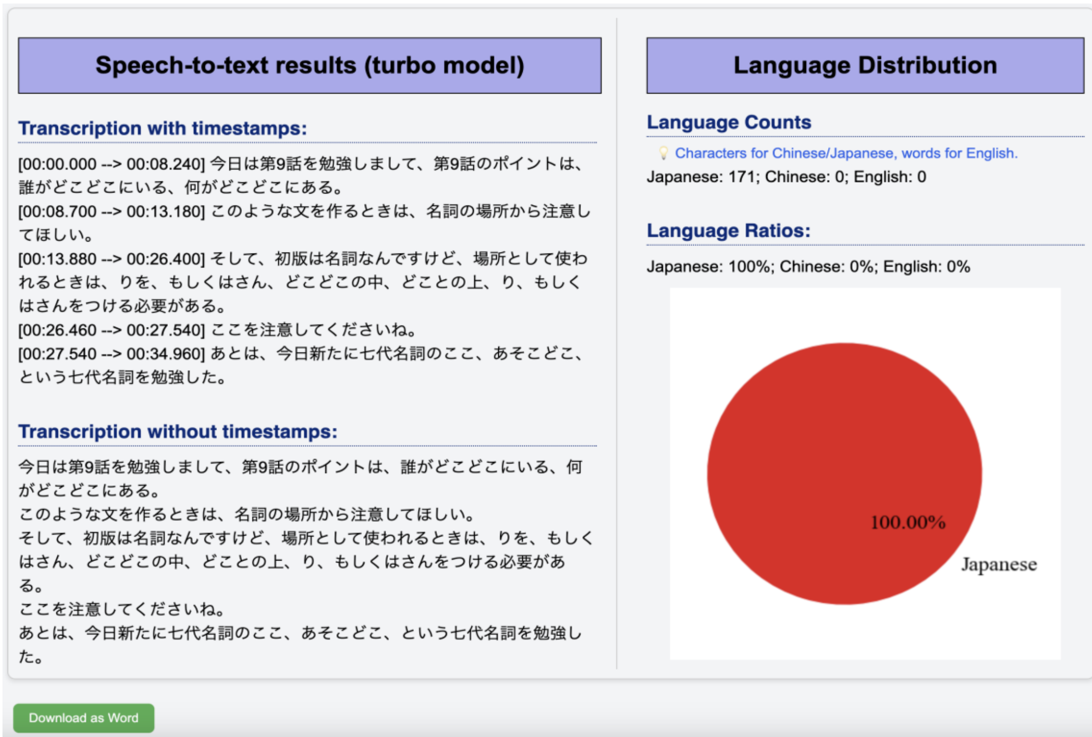


**图 3 实例 1[2022 课堂]转写部分（带时间戳）**

**表 1　实例 1 的人工校对（不带时间戳）**

| |
|---|
| 今日は第 9 **課**を勉強しまして、ええと、第 9 **課**のポイントは、ええ、誰々がどこどこにいる、何々がどこどこにある。このような文を作るときは、名詞の場所化を注意してほしい。そして、**书包**は名詞なんですけど、場所として使われるときは、**里**を、もしくは**上**、どこどこの中、どこどこの上、**里**、もしくは**上**をつける必要がある。ここを注意してくださいね。あとは、ええ、今日新たに指示代名詞のここ、あそこどこという指示代名詞を勉強した。 |

注：带下划线的文字为 APP 转写错误或遗漏部分。灰色底纹的黑体字为中文（下同）。

**表 2　实例 1 文本的中译**

| |
|---|
| 今天我们学习了第 9 课。嗯，这一课的重点是，嗯，表示“谁在某处”“什么东西在某处”的句型。造这种句子时，要注意名词的“场所化”。比如，“书包”是名词，当它表示场所时，需要加上“里”或“上”，“在某处里” “在某处上” 。 要加上“里”或“上” 。请大家注意这一点。另外，嗯，今天还学习了新的指示代词——“这里”“那里”“哪里”等用法。 |

结果显示，该段授课语言几乎全部为日语。APP 检测的语种比例为：日语（L1）100%，汉语（L2）及英语均为 0%（见**图 3**）。实际上，该片段中包含 5 个

汉语词语（见**表 1** 与**表 2** 中带底纹的黑体字），真实的汉语（L2）使用比例约为 3%。这是因为 Whisper 在处理句内语码转换（code-switching）时的识别效果不理想，常将句中嵌入的外语误判为整句的主要语种，从而导致 L2 使用比例被低估。例如，在**表 1** 中，日语语境中的汉语词语“そして、书包は名詞なんですけど”被误识为“そして、初版は名詞なんですけど”（文中下划线部分为笔者标示的识别错误，下同）；汉语“里”“上”分别被转写为日语读音相近的平假名“り”“さん”。此外，即便是日语部分，涉及专业术语时也容易出错，如“場所化”“指示代名詞”分别被误识为“場所から”“七代名詞”。

语码转换不仅可能导致机器识别错误，也可能对人类的听辨与理解造成干扰（Xiao & Park, 2021；Amrate & Tsai, 2025）。因此，外语教师在课堂中进行语码转换时应有策略，同时应尽量避免使用过多高难度术语，可将其转化为更易理解的表达，以确保学生能够真正听懂并理解课堂的内容。虽然 APP 的语音识别存在一定误差，但能即时将语种比例可视化，这为教师检查语言使用比例、调整语码转换策略提供了依据。APP 自动生成的时间戳不仅呈现语料的时间分布，还可用于分析教师发话的节奏与密度。例如，在实例 1 中，单句话语的最长持续时间约 13 秒，最短为 1.08 秒（**图 3**），这表明教师发话单位较短、语速较快，整体节奏紧凑。在此之前，教师已就“名词的处所化”进行了约 146 秒的讲解（见**图 2** 第一行），随后点名 9 名学生依次朗读课文例句并检查发音（**图 3** 展示了其中两名学生的互动）。其中 3 名学生未作答，实际仅 6 名学生各朗读一句，平均每句约 20 秒，其余学生仅处于聆听状态，缺乏发言机会，也失去了与教师进行意义协商的契机。[2022 课堂]的其他片段亦呈现类似情况。

总体来看，该课堂以教师讲解为主，虽有师生问答环节，但多停留在知识再现层面，缺乏交际性互动（**图 2**）。FLINT 分析结果进一步验证了这一倾向（曲明&砂岡和子，2024）：

（A）教学主导性强——教师发话次数占总量的 58%，时长占 67%，均高于学生的 42% 与 33%。

（B）缺乏支持学生主体性的发话——直接性发话（如指示、说明、订正）显著多于间接性发话（如提问、称赞、鼓励、重述改正），在次数和时长上分别高出约 28% 与 50%。

值得注意的是，无论是 FLINT 还是 APP，均只能在定量层面揭示课堂结构性问题，尚无法提供如何重构互动或激发学生能动性的具体策略。即便如此，教师仍可依据这些分析结果回顾自身教学行为，检查课堂互动状况，并据此调整策略，以促进教学改进与质量提升。

### 3.3. 实例 2：[2025 课堂]发话内容的语音转写

实例 2 取自日本的一位大学汉语教师（母语为汉语）于 2025 年 9 月 27 日为初级学习者进行的约 15 分钟模拟课堂片段，主题为讲解“是……的”句式。APP 自动转写了其中约 1 分 32 秒的片段，结果显示课堂语种比例为：L2（中文）42.18%，L1（日语）57.45%， 英语 0.36%（见**图 4**）。整体来看，[2025 课堂] 的教师在讲解中注重意义协商，并设计了使学生能够在真实语境中操练的语言活动（见**表 3**），因此目标语（汉语）的使用比例相对较高。尽管模拟课堂中没有学生互动环节，但讲解流畅、结构清晰。然而，教学内容仍主要停留在“是……的”句式的知识再现阶段，缺乏引导学生将语法形式与交际情境相结合的教学设计。

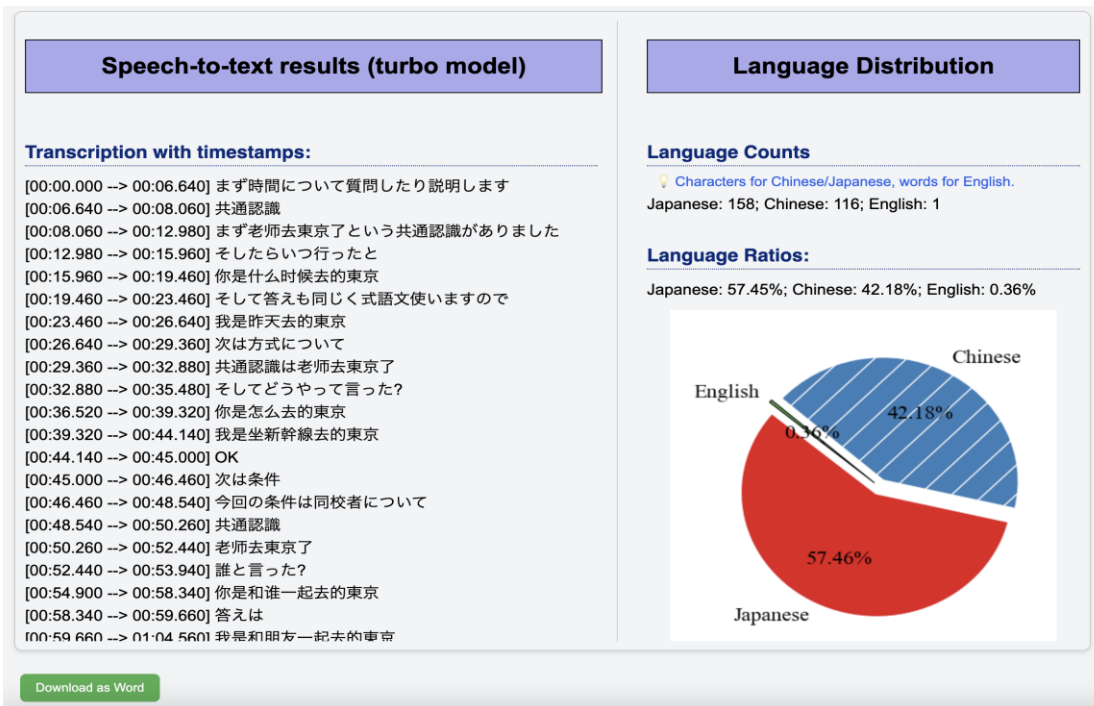


**图 4 实例 2 [2025 课堂] （带时间戳）**

**表 3 实例 2 的人工校对（不带时间戳）**

| |
|---|
| まず時間について質問したり説明します。共通認識。まず老师去東京了という共通認識がありました。そしたらいつ行ったと。——你是什么时候去的東京。でそして答えも同じく是的構文使いますので。——我是昨天去的東京。次は方式について。共通認識は老师去東京了。でそしてどうやって言った?——你是怎么去的東京。我是坐新幹線去的東京。OK。次は条件。今回の条件は同校者について。共通認識。——老师去東京了。誰と言った?——你是和谁一起去的東京。答えは我是和朋友一起去的東京。で次は場所について。共通認識は。——老师買了東京香蕉。先ほど皆さんの教室で。お土産ありますよって。——老师買了東京香蕉。じゃあどこで買った?。——東京香蕉是在哪買的? 東京駅ですね。——東京香蕉是在東京站买的。 |

**表 4 实例 2 文本的中译**

| |
|---|
| 首先，我来提问或说明“时间”这一部分的内容。这里有一个共通认知：老师去了东京。于是我们可以问——“什么时候去的？”——你是什么时候去的东京？那么，回答时同样使用“是……的”句式，比如：我是昨天去的东京。接下来是“方式”。共通认知仍然是“老师去了东京”。那么，可以问“怎么去的？”——你是怎么去的东京？回答：我是坐新干线去的东京。然后是“条件”。这次的条件是“同行者”。共通认知仍是“老师去了东京”。于是可以问“和谁去的？”——你是和谁一起去的东京？回答：我是和朋友一起去的东京。那么，接下来是关于“地点”的部分。共同认识是：老师买了东京香蕉。刚才在大家的教室里，我说“有礼物啊”，——老师买了东京香蕉。那，在哪里买的呢？东京香蕉是在东京哪里买的？——在东京站。——东京香蕉是在东京站买的。 |

对于以日语为母语的学习者而言，“是……的”句式以及实例 1 中出现的“处所名词”等结构，均属于新的语法概念（王亚新，2021；Pan & Liu, 2023）。若仅依赖形式性讲解与“共通认知”“处所化”等抽象术语，难以实现“可理解性输入（comprehensible input）”。部分以中文为母语的教师未能充分利用日语这一学习者的母语资源进行对比，导致学习者难以从语境中理解并认同“汉语母语者为何常用这些句式”，从而缺乏意义建构的动机。

这种依循教材进度、以知识传授为中心的授课模式，在其他课堂中亦屡见不鲜。如何避免陷入此类“惯性化教学”的盲点，仍有赖于教师的省察与反思能力（白水始等，2021）。目前，APP 虽能在定量层面揭示教学问题，但尚无法提供具体的课堂改进策略。为弥补这一不足，本研究进一步尝试借助生成式人工智能，自动生成课程优化建议，以探讨其在汉语课堂设计中的潜在优势与应用局限。

## 4. 生成式 AI 辅助课堂改进的效能评估

生成式人工智能通过分析海量文本的规律，基于上下文预测最合适的下一个词语或表达，从而生成回答。因此，许多面向第二语言习得的语言学习应用程序及提示词设计，均依托 AI 算法的优势，涵盖了翻译与摘要、误用检测、文本补全、习得难度预测等多种功能（Shan et al., 2024；连维琛等，2024）。在本研究中，我们对 ChatGPT 的期待并不限于这些表层的语言输出功能，而是希望其能够提供关于教学方法与课堂设计的启发性建议。

已有研究指出，明确且精确的提示（prompts）能够显著提升 AI 输出的针对性与质量（Poole & Coss, 2024）。然而，要实现课堂改进指令的明确化，教师需对自身的教学过程具有深入的理解、评估与调适能力，并能够将这种元认知觉察进行语言化与概念化（Mizumoto，2023）。本研究利用 AI 的目的，主要在于协助非语言教育专业背景的教师反思并诊断其教学实践。考虑到部分教师难以自行提出教学改进的关键词，我们设计了两种类型的提示词：（A）开放式提示与（B）结构化提示。（A）类提示词根据前述 FLINT [2022 课堂] 分析中揭示的两大问题——即“如

何促进学生更具主体性、积极参与与互动”——进行探索性设计。（B）类提示词则要求根据具体条件（据情况可增减条件项）提出相应的教学改进方案。两类提示词均要求提升课堂中的汉语使用比例，具体提示内容见**表 5**。带波浪下划线的部分为作者有意强调的重点，“//”符号之后所附的课堂文本为未经修改、未校正的 APP 自动转写原文，便于授课教师直接使用。正如后文所示，AI 已对其中的转写错误部分进行了基本修正。

**表 5 提示类型与具体提示示例**[19]

| 提示类型 | 具体提示例 | 生成结果号码 |
|---|---|---|
| （A）<br>开放式提示 | 以下是一位汉语教师在课堂上讲解（如处所词组／“是……的”句式；根据本课内容填写）语法点时的一段课堂发话文本。将其改写为能够促使学生发挥主体性并积极参与互动的课堂形式。同时将课堂中汉语的使用比例控制在 30%至 40%之间。//（此处附上该课堂的 APP 转写文本） | [实例 1**（A）-1]**<br>[2022 课堂]<br>[实例 2**（A）-2]**<br>[2025 课堂] |
| （B）<br>结构化提示 | 以下是一位汉语教师在课堂上讲解（如处所词组／“是…的”句；根据本课内容填写）语法点时的一段课堂发话文本。依据下列四项条件提出教学改进方案：1）引导学生通过母语（日语）的比较来深化汉语语法理解；2）避免使用抽象术语，确保输入内容易于理解；3）鉴于学生处于初级水平，课堂练习宜以简短回答或选项反应为主；4）将课堂中汉语的使用比例控制在 30%至 40%之间。//（此处附上该课堂的 APP 转写文本） | [实例 1**（B）-1]**<br>[2022 课堂]<br>[实例 2**（B）-2]**<br>[2025 课堂] |

（A）类提示要求 AI 分别将 [2022 课堂] 与 [2025 课堂] 改写为师生互动更频繁、汉语使用率更高的课堂脚本，整体难度较低。（B）类提示较（A）类多出三个条件，内容更具体、更严密。以下将展示在（A）与（B）两种不同提示词条件下所得回应的主要特点，并对比两种提示条件下的输出结果，以检验 AI 生成建议的适切性与启发性。鉴于生成式 AI 在多次运行中可能产生不同的输出结果，本文选取了 ChatGPT 5.0 于 2025 年 9 月 15 日至 10 月 26 日期间生成的数十个版本中的若干示例进行呈现与分析。本次提示词与回答均以汉语进行；若提示词改为日语或英语等其他语言，ChatGPT 亦可生成结构基本一致的结果。

[19] 尽管指令中要求中文使用比例保持在整体约三成。因 GPT 无法进行精确的数值计算，有时会超过三成。

### 4.1 AI 生成能力评估

总体而言，无论是（A）类提示还是（B）类提示，AI 在改进课堂设计方面均展现出一定优势，不仅能提供多维度的互动模式，还能够提出较为适切的语言使用比例方案。相较之下，AI 在语法讲解的生成方面仍显不足，往往仅对输入的提示内容进行表层重组，或提供来自网络检索的简短句例。这些句子既缺乏真实交际语境，也难以体现语法功能的深层关系（陶红印，2025）。AI 在教学改进建议中未能展现出具有启发性的教学构想，也难以生成新的潜在概念，甚至偶有错误出现。下面将分别介绍并分析 AI 方案的检验结果。以波形框标示的 ChatGPT 方案， 除对前后说明部分及个别换行位置进行了适度调整外，其余内容均与原文保持一致。

#### 4.1.1 AI 辅助课堂设计的效能

经过多次检验后，从下列[实例 1（A）-1]【A】【B】与[实例 2（A）-2]【C】的结果可见，ChatGPT 除语法讲解部分外，均能持续生成符合提示要求的方案。具体而言，【A】通过看图说话、全班齐读，【B】以小游戏活动，【C】则以伙伴互问等方式，均被改写为增强课堂互动的教学脚本。与原课堂几乎完全由教师单向讲授的情况相比，这些改写有效提升了学生的课堂参与度与教学效果。值得注意的是，教学改进示意均为 ChatGPT 自动生成的描述，可见 AI 具备提出课堂改进方案的能力。

【A】[实例 1（A）-1]通过看图说话、问答、全班齐读等方式，增强了师生间的互动。

教师：大家好！今天我们学习第九课。今天的重点是——“谁在哪儿”“什么在哪儿”。（黑板书写：谁在哪儿？什么在哪儿？）请看图片。（展示教室图片）这是谁？——对，是老师。老师在哪儿？（指向学生）小李，你说说看。
学生 A：老师在教室里。
教师：很好！“老师在教室里。”——对不对？（全班回应）那，“书”呢？“书在哪儿？”
学生 B：书在桌子上。
教师：非常好！大家一起来说：“书在桌子上。”
（全班齐读）

【B】[实例 1（A）-1]以小游戏提高学生的反应速度，增强课堂互动。

【练习活动】
教师：现在我们玩一个小游戏——“找一找”。
我说一个东西，你们说它在哪儿。比如——“铅笔”在哪儿？
（学生环视图片或实物，回答）
学生 C：铅笔在书包里！
教师：很好！那，“手机”呢？

学生 D：手机在老师的手上！😂
教师：哈哈，对！非常好！

【C】[实例 2（A）-2]用图片提示、学生间互问等形式，提升学生的课堂参与度。

教师:现在换一个问题。还是“老师去东京了”。那，怎么去的？（提示用图片或实物）大家问我吧！
学生:老师，您是怎么去的东京？
教师:很好！我是坐新干线去的东京。那你呢？
（学生间互问）“你是怎么去的东京？”——“我是坐飞机去的。”
（教师小结）👉“是……的”也可以强调方式。

### 4.1.2 AI 能把握语言使用比例

无论是（A）类还是（B）类提示，ChatGPT 从生成的教学示例均将汉语使用比例控制在 35～40% 之间，且兼顾了可理解性与实践性。可见，AI 在设计外语教学方案时，能够较好地掌握语言使用比例的适切性，并能主动提供教师课堂常用语。尽管部分语句（如“谁在桌子上？”参见【D】）并非常用句，但在维持汉语使用比例、不增加课堂难度方面，仍具有实用价值。

【D】[实例 1（B）-1] 改进要点：教师讲解或规则说明使用日语（60∽70%）；教师的提问、提示和练习指令使用汉语（30∽40%）；学生的回答以汉语短句为主。语言分配示例如下：

| 教学环节 | 教师主要语言 | 汉语比率 |
|---|---|---|
| 导入・比较 | 日语＋汉语例句 | 約 30% |
| 示范讲解 | 汉语 | 約 40% |
| 练习活动 | 教师问答・学生短答（汉语） | 約 40% |
| 总结・回顾 | 日语 | 約 30% |

教师用的固定句（保持汉语比率而不增加难度）：
“请看——”“谁在桌子上？”“对，不错！”“再说一遍。”“很好，大家一起说——”

## 4.2 语法认知生成的缺位

如上所见，在数据驱动的条件下，生成式 AI 在面向初级水平学习者的课堂设计与语言使用比例控制方面展现出一定优势。然而，其在语法教学生成方面仍存在局限，表现出认知生成的缺位。AI 所生成的汉语句式往往刻板且简短，常缺乏真实交际语境，难以体现语法功能的深层关系。有时还生成脱离教学语境的内容，导

致句子显得突兀、不自然，甚至偶尔出现判断错误。下面【E】方案中的“谁在哪儿？”、“学生在操场”、“猫在椅子上”等句子缺少谓语成分，名词也多未搭配量词，显得不够自然。应用“（量词）＋谁／什么＋在哪儿＋做什么”的句式，如“（两个）学生在操场上（跑步）”、“（一只）猫在椅子上（睡觉）”等表达更为自然。此外，【F】方案中，教师以“这句话对吗？”来让初级学生判断句式正误，语气略显突兀，不够自然。在“是……的”结构中，除强调主语外，“是”在口语或非正式语境中常被省略，因此 AI 无需刻意纠正“这句话还少一点儿‘是’”之类的表达。下面所举的方案中，笔者以***斜体粗字体***加＊号标注的部分，表示包含应答方式在内，均为不自然或错误的汉语表达（下同）。

【E】[实例 1（A）-2] 处所词组的教学改进方案

【练习一：***谁在哪儿？***】
教师：请大家看第二张图。（图片：＊***学生在操场，猫在椅子上***）我问——谁在操场？
（指名）　学生 A：＊***学生在操场***。
教师：很好！那猫呢？
（学生回答）学生 B：＊***猫在椅子上***。
教师：好！大家一起说一遍。
（全班齐读）

【F】[实例 2（A）-2]「是……的」句的教学改进方案

① 導入（情境設定）
教师（日语＋少量中文）:今天我们来学习一个非常有意思的句型——“是……的”。比如：“老师去东京了”。＊***这句话对吗？***
（学生：***对！***）
可是，如果我想问“什么时候去的？”，该怎么说？
学生:（思考后）“你什么时候去的东京？”
教师:很好！＊***但是***——（微笑）***这句话还少一点儿“是”***。听我说：“你是什么时候去的东京？”大家一起来说一遍。
学生（齐读）:你是什么时候去的东京？

另外，尤其是在（B）类提示的四项条件中，AI 难以有效实现“引导学生通过日汉比较加深语法理解”与“避免使用抽象术语以确保可理解输入”这两项教学要求。其生成脚本多停留在语言形式与表层推理层面，缺乏教学构想的启发性，未能形成新的教学概念（参见【G】-【I】）。理想的回答不仅应引导学生注意母语与目标语的差异，还应超越单纯的比喻式说明，能够以语言的形式阐明“日汉表达概念的差异”。例如，汉语一般名词不同于日语，不具“场所性”，因此需加方位词才能表示地点；又如，在讲解“是……的”句式时，AI 应进一步说明其与汉语时态概念的关联，从而帮助学生理解“是……的”是汉语中常用且必须掌握的句式。

ChatGPT 在语法认知生成方面的不足，主要源于其仅依赖输入文本（如 APP 转写语料）进行表层重组与检索式生成，缺乏深层语法认知与教学构想能力，因而难以提出具有启发性的教学方案。另外，AI 的优势主要体现在编写适合不同水平班级的例句、提供带拼音的词汇表、制作音频与图像教材以及设计测验等方面，从而为课堂改进提供有力支持（胡文菊，2021；Zhang, 2021）。鉴于篇幅所限，其他功能不再赘述。

【G】[实例 1（B）-1]以中日对照，促进学生理解处所词组的差异。

改进要点：首先呈现中日对照，让学生“自己发现”句式结构的差异。
例：汉语：老师在教室里。
日语：先生は教室にいます。
→ 教师提问：“日语中只用‘に’，但汉语中说‘在……里’，为什么呢？”
让学生自己发现汉语中需要“地点＋方位词”的结构。
意图：不是让学生死记语法形式，而是通过对母语与目标语差异的比较，引导学生从意义层面加深理解。

【H】 [实例 2（B）-2] 教学策略：教师先呈现中日对照句，帮助学生发现结构差异。

例如：汉语：“你是什么时候去的东京？”
日语：“あなたはいつ東京に行きましたか？”
教师引导学生观察：
👉 日语中只用“行きました”就能表示过去，
👉 而汉语则使用“是……的”来表达“过去＋强调”的意义。
活动设计：在黑板上并列书写两种语言，教师提问学生：“哪里不同？”“感觉有什么差别？”引导学生以意义为焦点进行比较，而非仅关注形式。
目的：让学生以母语结构为线索，理解汉语表达的独特性（即焦点化功能）。

【I】 [实例 1（B）-2] 不要频繁使用“名词＋助词”或“方位词”等抽象术语，而应借助教室中常见的图片、实物和动作来帮助学生理解。

教师可以在黑板上书写并配合图示：
• 老师（人）在教室里
• 猫（东西）在桌子上
• 书（东西）在包里
→ 一边指着图示，一边提问学生：“谁？在哪儿？”以此引导学生主动发话。
比喻性说明示例：“‘在’是告诉位置的魔法词，‘里・上・下’就像指示地方的小手指”。通过这种形象化的说明来避免使用抽象术语。

## 5. 总结

### 5.1 研究结论

本研究围绕自动转写与生成式 AI 在汉语课堂中的应用，构建并验证了一套集音频转写、L1 / L2 使用比例可视化与课堂诊断功能于一体的 APP。实证结果表明，该 APP 在处理中日语码混合的课堂录音时，具有较高的转写准确率与操作便利性，显著降低了人工分析成本，为教师即时掌握课堂语言分布、优化教学策略提供了量化依据。

在案例分析中，教师可借助 APP 快速了解课堂中 L2 使用比例及师生的互动情况。当 APP 结果页面显示 L2 使用偏低或师生互动不足时，教师可结合生成式 AI，将量化结果与课堂情境相结合，生成针对性更强的教学改进方案，从而推动数据驱动的课堂优化实践。该机制不仅为教师能动性的培养提供了新路径，也为语言教育研究带来了数据化与智能化的新方向。即便缺乏专业分析背景的教师，也能依据系统提供的数据做出基本的教学优化判断。

然而，如第四章所述，ChatGPT 仍受限于创新语法概念的生成。其“智能”缺乏元认知能力的原因，一方面与大型语言模型（LLMs）的架构设计有关，另一方面则在于 LLMs 尚未能有效吸收语法研究的最新成果，尤其是汉语本体研究与第二语言习得（SLA）研究之间的衔接不足，导致 AI 缺乏形成创新语法概念的优质知识资源。鉴于人工“智能”尚未成熟，教师应在教学情境中保持主体性，通过自主决策、反思与创新教学设计，引导课堂改进。与此同时，积极关注 SLA 研究的最新进展，在努力提升 L2 使用比例的同时，充分利用学生的母语资源，在当前的教学实践中尤为关键。

### 5.2 研究局限与未来展望

首先，在方法层面，本文以实例 1 与实例 2 的 APP 输出结果为基础，结合预设提示词输入至 ChatGPT 生成课堂改进草案，此过程主要基于假设设定而展开。未来研究有必要将该工具实际引入课堂场景，系统收集并分析教师与学习者的反思性数据，以开展更具实证意义的验证。

其次，在技术层面，如第 3.2 节案例分析所揭示，APP 在界面交互、功能扩展以及多语种混合语料的识别精度方面仍有不足。目前尚不支持视频直接上传与转写，也缺乏对不同发言人的区分功能。在语言识别机制上，基于字符范围的归属判定方法易高估汉语比例，尤其在日语文本中大量使用汉字时，可能导致偏差。

最后，在应用层面，尽管该工具已公开提供使用，但实际用户规模仍然有限，且尚未实现与生成式 AI 的自动联动功能，这在一定程度上限制了其推广与应用。

为评估该 APP 的功能实用性并明确后续改进方向，我们邀请 5 名教师[20]在试用该 APP 后填写问卷。因篇幅所限，以下仅呈现用户反馈的关于功能便利性、不便之处以及改进建议的反馈。

在功能便利性方面，用户普遍认可该 APP 在转录中日语码混合音频转录中高便利性（如可直接上传音频进行分析、可视化 L1/L2 使用比例）与较高的准确性，认为其在节省人工处理时间、提高课堂教学分析效率方面具有一定的应用价值。

用户反馈的功能不便主要集中在五个方面：（1）交互界面的设计相对简单；（2）对 APP 内可选语音识别模型的说明不足；（3）无法直接上传视频文件进行语音转写；（4）转写结果未区分不同的发言人；（5）英日语码混合场景下的语音识别精度偏低。

用户提出的改进建议主要包括三个方面：（1）扩展相关功能：如支持视频文件直接上传与转写、转写结果标注不同发言人，以及支持简体/繁体中文切换等；（2）增强界面交互体验：包括统一界面语言风格、在模型选择界面增加说明及模型推荐提示、丰富网页布局设计及转写结果的呈现方式；（3）优化多语种混合（如英日混合）以及句间语码转换场景下的语音识别精度。

结合用户的反馈和建议，以及本研究开发该 APP 的主要目的与当前存在的主要问题，未来研究拟从以下四个方面进行优化，以提升 APP 的适用性与推广价值。

（1）提升响应速度，在保证识别精度的前提下缩短模型加载与转写时间。目前，模型加载耗时较长，语音转写的响应速度较慢。Whisper 模型本身的参数量较大，尽管系统通过预加载机制 whisper. load_ model 避免每次请求时重复加载模型的问题，但整体转写过程仍存在延迟，尤其是使用 large-v2 或 large-v3 这两种大模型时，加载和转写时间明显增加。为兼顾识别精度与速度，建议用户选 turbo 模型进行语音转写。

（2）集成说话人分离功能，结合 Whisper 与 “pyannote.audio”，实现对课堂互动的细粒度分析。徐勤 & 砂岡和子（2024）通过 Whisper large-v3 与 “pyannote.audio” 的组合实现语音转写和说话人分离：即由Whisper生成带有时间戳标注的初始文本，再由“pyannote.audio”对音频进行话人分离。未来计划集成该功能，以拓展“谁在说什么”的细粒度口语语料分析需求。今后将在APP种集成该功能，在语音转写后的结果种区分不同的说话人。

（3）优化汉日混合语料的语言归属判定算法，减少比例偏差（见**表 3** 与**表 4**），提升中日汉字识别与语种归属识别的精度。当前版本的 APP 主要通过字符范围进

[20] 参与本研究的 5 位教师均为女性，均从事语言教育工作。其中，4 位分别在日本的不同大学教授汉语或英语，1 位在中国的高中教授日语。她们的教龄分布为：5 年以下 3 位，6－10 年 1 位，20 年以上 1 位。除 1 位教师提供了口语考试录音作为研究材料外，其余均提供了本人实际课堂的教学录音用于分析。

行语言归属划分，其中汉字被自动归类为汉语，平假名、片假名被归类为日语。这会导致语言判断偏差，尤其在处理日语中存在大量汉字的文本时，汉语的比例易被高估。未来计划进一步优化汉日混合语料文本的语言归属判定，以提升语言识别的准确率。此外，进一步提升系统在英日混合及句间语码转换场景下的识别准确性，也将有助于其在更多语言环境中的推广应用。

（4）扩展支持用户自定义分析时间段的功能，便于片段式语料研究。当前 APP 默认转写整个音频。未来计划在 Web 界面添加自定义时间区间的选项，使用户在上传音频后可自行指定音频需要分析的起止时间段，以拓展语言研究中常见的片段式语料分析需求。

如前所述（详见 1.1 章节），文科省的调查结果显示，日本的大学外语学习者普遍对自身的语言掌握程度感到不满意，这在一定程度上表明教师可能未能准确把握学生的实际理解水平。传统上，教师多依赖“客观测试”来检验学生的理解程度，但若测试仅限于词汇和语法的辨别性能力，生成式 AI 的表现往往优于学生（Mizumoto, 2023）。在技术快速发展的背景下，语言教学应更加注重认知能力的培养（He & Lin，2021）。因此，外语教师更应主动反思日常教学实践，改进教学设计，以确保提升学生的真实理解水平与学习成效。本研究开发的 APP，旨在为教师提供发现课堂问题的反思契机，从而支持持续的教学改进与优化。

**致谢**：本研究得到日本学术振兴会科学研究基金（JSPS KAKENHI：课题编号 24K04091、24K16129）资助。在此谨致谢忱。本文系根据2024年6月22日于TCLT12（The International Conference and Workshops on Technology and Chinese Language Teaching）上由砂冈和子与徐勤共同发表之报告〈多语码汉语教学课堂中的话者分离与文本转录——Whisper与“Pyannote.audio”的应用研究〉修订而成。

## 参考文献

Amrate, M., & Tsai, P. (2025). Computer-assisted pronunciation training: A systematic review. *ReCALL*, 37 (1), 22-42. https://doi.org/10.1017/S0958344024000181

Fanselow, J. F. (1977). Beyond Rashomon: Conceptualizing and describing the teaching act. *TESOL Quarterly*, 11 (1), 17-39. https://doi.org/10.2307/3585589

Ferraro, A., Galli, A., La Gatta, V., & Postiglione, M. (2023). Benchmarking open source and paid services for speech to text: an analysis of quality and input variety. *Frontiers in Big Data*, 6, 1210559. https://doi.org/10.3389/fdata.2023.1210559

He, F., & Lin, C. (2021). Supporting online Chinese narrative writing pedagogy through metacognitive writing process and approach: A design-based research. *Journal of Technology and Chinese Language Teaching*, *12*(1), 117-137.

Hu, W.-C. (2021). The theoretical foundation of virtual reality assisted language learning and its application in TCSL. *Journal of Technology and Chinese Language Teaching*, *12*(2), 66-85. [胡文菊. (2021). 虛擬實境科技運用於語言學習的理論背景與華語教學範例. *科技与中文教学*, *12*(2), 66-85.]

 

Iino, A. (2009). Review of classroom observation systems -focusing on FLINT, COLT, and FOCUS，*Bulletin of Seisenjogakuin Junior College*, 27, 13-29. http://purl.org/coar/resource_type/c_6501 [飯野厚. (2009). 語学授業観察法の概観—FLINT, COLT, FOCUS に焦点をあてて. *清泉女学院短期大学紀要*，27, 13-29. http://purl.org/coar/resource_type/c_6501]

Ishizuka, H., Koshie, M., Sakuai, Y., Kamada, R. & Kubo, M. (2021). An attempt to develop foreign language teaching by the use of classroom analysis tool which can provide immediate feedback ― By using mobile COLT. *Journal of Hokkaido University of Education (Humanities and Social Sciences)*, *72*(1), 69-78. [石塚博規，越江麻衣，櫻井靖子，鎌田亮祐， & 久保稔.(2021). 即時フィードバック可能な授業分析ツールによる外国語授業改善の試み：Mobile COLT を用いて.*北海道教育大学紀要（人文科学・社会科学編）*, *72*(1)，69-78.]

Lian, W., Zheng, M., & Xu, J. (2024). A comparative study on development models of AI Chinese language partners based on the ERNIE large language model. *Journal of Technology and Chinese Language Teaching*, *15*(2), 35-53. [连维琛，郑明鉴，& 徐娟 (2024). 基于文心大模型的 AI 中文语伴开发模式对比研究. *科技与中文教学*, *15*(2), 35-53. ]

Macaro, E. (2009) . Teacher use of codeswitching in the second language classroom: Exploring "optimal" use. In M. Turnbull & J. Dailey-O'Cain (Eds.), *First language use in second and foreign language learning* (pp. 35–49). Multilingual Matters. https://doi.org/10.21832/9781847691972-005

Ministry of Education, Culture, Sports, Science and Technology. (2020). *Results of the FY2019 national student survey (pilot implementation): Data report.* [文部科学省. (2020). *令和元年度「全国学生調査（試行実施）」結果【資料編】*] https://www.mext.go.jp/content/20200618-mxt_koutou01-000001987_03.xlsx

Ministry of Education, Culture, Sports, Science and Technology. (2022). *Results of the FY2021 national student survey (second pilot implementation): University edition — Data report.* [文部科学省. (2022). *令和 3 年度「全国学生調査（第 2 回試行実施）」結果（大学）【資料編】*.] https://www.mext.go.jp/content/20221110-koutou01-000001987_1.xlsx

Ministry of Education, Culture, Sports, Science and Technology. (2023). *Results of the FY2022 national student survey (third pilot implementation): University edition — Data report*. [文部科学省. (2023). *令和 4 年度「全国学生調査（第 3 回試行実施）」結果（大学）【資料編】*] https://www.mext.go.jp/content/20230712-koutou02-000001987_2.xlsx

Mizumoto, A. (2023). Data-driven learning meets generative AI: Introducing the framework of metacognitive resource use. *Applied Corpus Linguistics*, *3*(3), 100074. https://doi.org/10.1016/j.acorp.2023.100074

Moskowitz, G. (1971). Interaction Analysis-A New Modern Language for Supervisors. *Foreign language annals*, *5*(2), 211-221. https://doi.org/10.1111/j.1944-9720.1971.tb00682.x

Mustafa, M. B., Yusoof, M. A., Khalaf, H. K., Rahman Mahmoud Abushariah, A. A., Kiah, M. L. M., Ting, H. N., & Muthaiyah, S. (2022). Code-switching in automatic speech recognition: The issues and future directions. *Applied Sciences*, *12*(19), 9541. https://doi.org/10.3390/app12199541

Myers, S. (2002). *Contact linguistics; bilingual encounters and grammatical outcomes*. Oxford University Press.

OECD. (2019a). *OECD future of education and skills 2030/2040.* https://www.oecd.org/en/about/projects/future-of-education-and-skills-2030.html

OECD. (2019b). *The OECD learning compass 2030*. https://www.oecd.org/en/data/tools/oecd-learning-compass-2030.html

Pan, V. J., & Liu, C. (2023). Focus constructions involving *shì* in Mandarin Chinese. *Languages, 8*(2), 103. https://doi.org/10.3390/languages8020103

Patsy, M. L., & Spada, N. (2021). *How languages are learned* (5th ed.). Oxford University Press.

Pellerin, M., Ishizuka, H., & Cervantes, V. F. (2024). Language teaching supervision for a new era: AICOLT system's journey to AI-driven innovation. *Proceedings of the International CALL Research Conference*, 2024, 213–218. https://doi.org/10.29140/9780648184485-32

Poole, F. J., & Coss, M. D. (2024). Can ChatGPT reliably and accurately apply a rubric to L2 writing assessments? The devil is in the prompt(s). *Journal of Technology and Chinese Language Teaching*, *15*(1), 1-24. https://doi.org/10.35542/osf.io/3r2zb

Qu, M., & Sunaoka, K. (2024). Pedagogical analysis of Chinese hyflex classrooms: A focus on teacher' and students' talk. *The Journal of Modernization of Chinese Language Education*, *13*(2), 14-24. [曲明, &砂岡和子. (2024). 汉语 Hyflex 课堂教学分析—教师与学生发言行为的分析. *中文教学现代化学报*, *13*(2), 14-24.] https://waseda.repo.nii.ac.jp/records/2005755

Radford, A., Kim, J. W., Xu, T., Brockman, G., McLeavey, C., & Sutskever, I. (2023). Robust speech recognition via large-scale weak supervision. *In International conference on machine learning,* 28492-28518. https://doi.org/10.48550/arXiv.2212.04356

Shan, L., Pan, Z., & Weidman, R. (2024). Integrating task-based language teaching and generative AI: Design, implementation, and evaluation of the CFLingo Platform for Chinese learning. *Journal of Technology and Chinese Language Teaching*, *15*(2), 1-34.

Shinya, I., Moegi S., Hajime S., & Kimihiko H. (2020). Reality of interaction among teachers and researchers in lesson study. *Cognitive Studies: Bulletin of the Japanese Cognitive Science Society*, *27*(4), 461-486. [飯窪真也, 齊藤萌木, 白水始, 堀公彦. (2020). 授業研究における教師と研究者の相互作用のリアリティ. *認知科学*, *27*(4), 461-486.] https://www.jstage.jst.go.jp/article/jcss/27/4/27_2020.043/_pdf/-char/ja

Shirouzu, H., Iikubo S., & Saito M. (2021). Birth, growth, and future issue in learning sciences: In search of science of learning that supports practices. *The Annual Report of Educational Psychology in Japan*. 60, 137-154. [白水始，飯窪真也，& 齊藤萌木（2021）. 学習科学の成立，展開と次の課題 : ―実践を支える学びの科学を模索して. *教育心理学年報*, 60, 137-154. ]

Spada, N., & Fröhlich, M. (1995). *COLT Communicative Orientation of Language Teaching observation scheme: coding conventions and applications*. National Centre for English Language Teaching and Research, Macquarie University.

Sunaoka, K., Wang, S., Sugie, S., & Xu, Q. (2023a). Code-switching in Chinese language classes: Inclusive membership and L2 acquisition optimization. *Proceedings of the 72nd National Conference of the Chinese Language Society of Japan* (pp. 253-257). [砂岡和子，王松，杉江聡子，& 徐勤. (2023a). 中国語授業の Code-Switching—包摂的メンバーシップと L2 習得最適化. *日本中国語学会第 72 回全国大会予稿集* (pp.253-257).]

Sunaoka, K., & Xu Qin. (2023b). Speaker diarization and text transcription in Chinese classrooms containing multilingual code-switching: Applied study of Whisper and Pyannote.audio. *Proceedings of the 12th International Conference and Workshops on Technology and Chinese Language Teaching (TCLT12)*, 61-71. (砂岡和子, & 徐勤. (2023b). [多语码汉语教学课堂中的话者分离与文本转录 —Whisper 和 Pyannote.Audio 的应用研究. *第十二届国际汉语电脑教学研讨会论文集*， 61-71.]

Sunaoka, K., Xu Qin. (2025). Self-analysis of foreign language classes using a multilingual Voice-to-Text app and AI: Development of the multilingual Voice-to-Text app. *Proceedings of the 31st Annual Meeting of the Association for Natural Language Processing (NLP2025)*, 1799-1804. [砂岡和子, & 徐勤. (2025). 多言語音声転写アプリと AI による外国語授業の自己分析ー Multilingual Voice-to-Text App の開発. *言語処理学会第 31 回年次大会 (NLP2025) 予稿集*, 1799-1804. ]

Tao, H. (2025). Interactive competence: The integration of interactional linguistics and language teaching and its practice in teaching Chinese as a second language. *Journal of International Chinese Teaching*, 03, 2-16. [陶红印. (2025). 互动能力：互动理论研究与语言教学的结合及其在汉语二语教学中的实践. *国际汉语教学研究*, 03, 2-16.]

Tasaki, A. (2006). An overview of studies on code-switching: For ana1ysis of communication in multilingual societies. *Language, Culture, and Japanese Language Education: Special Supplementary Issue — Frontiers in Second Language Acquisition and Education Research,* 54-84. [田崎敦子. (2006). コードスイッチング研究の概観: 多言語社会のコミュニケーション分析に向けて. *言語文化と日本語教育(増刊特集号, 第二言語習得・教育の研究最前線)*, 54-84.]

Wang, Y. (2021). Pragmatic conditions and functions of the Chinese “shi…de” construction. *Bulletin of the Institute of Human Sciences*, *Toyo University*, 23, 17-37. [王亚新. (2021). 汉语“是…的”句的语用条件和功能. *東洋大学人間科学総合研究所紀要,* 23, 17-37.]

Xiao, W., & Park, M. (2021). Using automatic speech recognition to facilitate English pronunciation assessment and learning in an EFL context: pronunciation error diagnosis and pedagogical implications. *International Journal of Computer-Assisted Language Learning and Teaching (IJCALLT)*, *11*(3), 74-91.

 

Xu Q., & Sunaoka, K. (2024). Speech transcription and speaker separation with multiple language codes: Advanced automatic speech recognition with Whisper + Pyannote.audio. *Proceedings of the 30th Annual Conference of the Association for Natural Language Processing*, 3149-3154. [徐勤, &砂岡和子. (2024). 複数言語コードを含む発話転写と話者分離：Whisper＋Pyannote.audio による自動音声認識の高度化. *言語処理学会第 30 回年次大会論文集*, 3149-3154.]

Zhang, S. (2021). Integrating augmented reality into a task-based thematic language teaching unit. *Journal of Technology and Chinese Language Teaching, 12*(2), 29-48.

# 基于 BERT-LDA 的中文学习 APP 评价指标体系构建研究 (Construction of an Evaluation Indicator System for Chinese Learning Apps Based on BERT-LDA)

张邝弋
(Zhang, Kuangyi)
北京语言大学
(Beijing Language and Culture University)
zky.staybirds@foxmail.com

侯尚余
(Hou, Shangyu)
云南大学
(Yunnan University)
houshangyu@stu.ynu.edu.cn

宋靖雯
(Song, Jingwen)
云南大学
(Yunnan University)
songjingwen@stu.ynu.edu.cn

肖锐
(Xiao, Rui)
云南大学
(Yunnan University)
ruixiao@ynu.edu.cn

**摘要：**本研究围绕中文学习 APP 在信息化及智能化趋势下的评价标准和质量挑战，提出了一种基于 BERT-LDA 模型的主题聚类算法，并结合 LLMs 的专家模型主题提取方法，从评价内容（内容质量）、评价过程（用户体验）、评价效果（学习成效）等核心维度构建了中文学习 APP 的多维度、动态化评价指标体系，并在情感分析任务验证其有效性，最后从智能化、动态化以及安全性等方面指明了未来国际中文教育数字资源评价指标体系构建的未来方向及风险挑战。

**Abstract:** This study explores the evaluation standards and quality challenges associated with Chinese learning apps in the context of increasing informatization and intelligence. It introduces a topic clustering algorithm derived from the BERT-LDA model and integrates an expert model for topic extraction utilizing Large Language Models (LLMs). A multidimensional and dynamic evaluation indicator system for Chinese learning apps is developed, focusing on core dimensions such as evaluation content (content quality), evaluation process (user experience), and evaluation outcomes (learning effectiveness). The validity of this system is confirmed through sentiment analysis tasks. Lastly, the study identifies future directions and potential risk challenges for creating evaluation indicator systems in international Chinese education digital resources, emphasizing intelligent, dynamic, and secure approaches.

**关键词：**中文学习 APP, BERT-LDA, 大语言模型

**Keywords:** Chinese learning APP, BERT-LDA, Large language model

 

## 1. 引言

以 ChatGPT 为代表的大语言模型(Large Language Model, LLM)在智能教育助手、课程定制、学习评价和语言交互等多个领域的应用，进一步突显了人工智能技术在全球中文教育普及与深化进程中的核心驱动作用，并揭示了 LLM 作为通用人工智能发展的重要里程碑，对中文教学产生了深远的影响（Wu et al., 2023）。

聚焦在移动学习的特定领域，中文学习 APP 凭借其方便高效的学习模式和出色的内容个性化功能，正逐渐成为众多中文学习者掌握知识和提高语言能力的关键工具（郭晶等，2021）。然而，对于如何精确有效地评价这些中文学习 APP 的质量和效率，以及它们在推动中文教学向数字化、智能化转型中所做的实际贡献，仍缺少一套广泛适用的评价指标体系。

传统的评价方法如层次分析法(Analytic Hierarchy Process, AHP) （Kharat et al., 2016）和德尔菲专家咨询法(Delphi Method)（Alon et al., 2025）虽在一定程度上解决了评价复杂系统的问题，但在应对快速迭代更新的学习环境，尤其是融合了先进人工智能技术的中文学习 APP 时，这些方法的局限性日益显现（王春枝等, 2011; 邓雪等, 2012）。鉴于此，本研究旨在借鉴现有评价理论及方法的基础上，提出一种基 BERT(Bidirectional Encoder Representations from Transformer)-LDA(Latent Dirichlet Allocation)型[1]的主题聚类算法，同时基于 LLMs(Large Language Models)[2]的专家模型对聚类主题进行提取，从而构建动态适应性增强的中文学习 APP 评价指标体系，并在实际案例中对指标体系进行验证，以实现对中文学习 APP 的多维动态评价，最终为中文学习 APP 的持续优化改进与健康发展提供有力支持和科学依据。

## 2.文献综述

评价指标作为量化评价与决策支撑的重要依据，在数据分析和业务优化过程中扮演着核心角色（虞晓芬等，2004）。数据挖掘作为一种强大的工具和技术手段，为评价指标的精准量化设定与深层次洞察力发现提供了强有力的技术支持和实质性的改进空间。在现代教育信息化背景下，中文学习 APP 作为普及语言学习及促进文化交流的数字媒介，能够通过对海量数据的挖掘提炼出有价值的信息。因此，构建一套完善的指标体系至关重要， 这不仅能有效实现对教学效果的实时监测与精确度量，也能深入剖析用户行为特征。这一举措将有力驱动中文智能教学效率的提升、数字化教育资源管理水平的进步，使得相关领域的研究和实践逐步摆脱传统上过度依赖人工操作、孤立分散的数据分析方式和相对有限的个性化服务，进而迈向自动化、规模化以及高度集成化的智慧教育新时代。

[1] https://www.kaggle.com/code/dskswu/topic-modeling-bert-lda
[2] LLMs（Large Language Models）为多个大语言模型，指基于提示引导的群体智能；LLM（Large Language Model）为单个大语言模型。

## 2.1 传统指标体系构建的研究现状

经验驱动的传统指标体系构建方法主要依赖专家经验和定性分析手段，具有较强的主观性和过程复杂性。例如，梁宇等（2023）综合运用德尔菲法和层次分析法，从专家经验和逻辑推理出发构建了国际中文教材评价指标体系；杨甜等（2023）基于广泛的问卷调查和用户反馈定性数据，构建了国际中文教师智能素养指标体系；方紫帆等（2023）参照《国际中文教师专业能力标准》[3]，结合理论研究与实践需求，构建了国际中文教师数字素养指标体系；程涛等（2024）利用德尔菲专家咨询法，尝试性地建构了具有中国特色的跨文化职业胜任力评价指标体系；宫雪等（2023）运用词频统计、多词序列提取、搭配分析等量化手段改进了国际中文教材评价指标基础框架的构建方式，减轻了其原有的“重定性、轻定量”问题。由此可知，以层次分析法、德尔菲方法等为代表的经验主义与半定量研究策略，在语言教学评价、教育政策制定及课程质量评估等多个领域发挥了重要作用（袁海红等, 2014；杨绪辉，2019）。然而，此类方法同样存在显著局限性：首先，它们对大规模客观数据的利用不足，过度依赖专家的专业见解和判断，可能导致评价结果的主观性强、稳定性差；其次，建立指标体系的过程往往涉及多次循环的匿名咨询、意见整合、反馈调整等环节，周期长且成本高；最后，由于专家观点的主观偏倚以及数据采集阶段可能出现的操作不一致，所得到的评价指标权重分配和预测结果，在客观性和精确性方面可能与基于大数据挖掘方法所得出的结论存在一定差距。

## 2.2 基于数据驱动的指标体系构建研究现状

数据驱动(Data-Driven)是指利用大规模客观数据，结合统计学和机器学习技术，以数据内在规律为基础，自下而上地构建评价指标体系的过程（杨现民等，2017）。这种方法强调通过算法模型揭示数据间的深层关联和模式，克服传统经验主义方法的主观性和不确定性，从而提高评价体系构建的客观性、准确性和普适性。随着深度学习和自然语言处理技术的发展，数据挖掘和机器学习算法已在不同领域指标体系构建中广泛应用，并已历经多个发展阶段：（1）传统模型的独立应用。早期的数据驱动指标体系构建多依赖于 LDA 等单一的模型，这些模型在处理文本数据时，能够初步揭示数据中的隐含主题或模式。（2）模型融合与技术创新。随着对更深层次数据关联需求的增长，研究者开始探索模型的融合使用，旨在通过结合不同模型的优势来提升分析的全面性和准确性。这一时期 Convolutional Neural Networks（CNN）等深度学习模型因其强大的语境理解能力而被引入，与 LDA 等传统主题模型结合使用成为趋势。例如，贾海楠等（2023）的工作展示了 LDA 与扎根分析法的融合，Lai（2023）使用 CNN 和 Bi-LSTM 模型对已有指标体系进行验证，都是这一阶段创新的体现。此外，潘小宇等（2023）提出的 HBL-LDA 方法，则是模型集成思想的实践，它通过结合多种模型特性，提高了书法价值评估指标构建的效率与准确性。（3）面向特定领域的最优模型选择与定制化融合，研究更加注重模型优化，以适应特定领域的独特需求。李天义等（2024）等从文本特征融合的视角出发，创造性地结合了 BERT-LDA 与 K-means 聚类算法，针对绘画作品的价值要素

[3] https://shihan-org.chinese.cn/index/build/detail.html?id=239

进行深度挖掘，这种融合模型不仅继承了 BERT 对复杂语境的强大理解力，还利用 LDA 捕获主题结构，同时通过 K-means 进一步细化类别，实现了对绘画领域高度定制化的价值评估指标体系构建。这标志着数据驱动方法在特定领域应用趋向成熟，不仅追求技术的先进性，更强调模型与实际应用场景的紧密结合。由此可知，基于数据驱动与主题挖掘的研究方法与指标构建研究已结合得十分紧密。

## 2.3 中文学习 APP 研究现状

诸如 *Duolingo*、*HelloChinese* 等中文学习 APP 因其丰富的用户交互数据、多样的学习行为记录以及实时更新的内容反馈等数字化资源特征，为教育研究和个性化学习提供了前所未有的可能性和挑战。相关研究主要呈现出以下特点：第一，中文学习 APP 评价数量较少，覆盖面不足，难以全面反映各类产品的优劣（高传智等, 2025; 李姝姝等, 2025）；第二，中文学习 APP 评价维度较为单一，往往集中在功能设计或用户体验上，无法做到对教学内容、学习效果、技术性能等方面的综合评价（刘永俊, 2021）；第三，缺乏系统的评价理论作为支撑，容易导致评价标准不一、主观性强的问题（杨倩, 2018）。由此可知，借助大数据与人工智能技术高效、科学地构建更具针对性、动态适应性的中文学习 APP 指标评估体系显得尤为迫切且必要。

## 2.4 国际中文教育数字化资源多维评价

人工智能、大数据、云计算、虚拟现实等技术的不断进步与广泛应用正深刻重构国际中文教育生态，其不仅促进目标受众角色从传统语言习得者向具备多元文化表征的网络用户转型，更通过技术赋能的增效机制，显著提升了该群体对数字化学习工具的探索动能、应用黏性及其对技术的接受度和融合能力。在这一演变过程中，赵学铭等（2017）基于模糊层次分析法对学习 APP 的易用性进行评价；张熠等（2019）基于 D－S 证据理论，从用户体验视角构建了针对中国大陆学习 APP 的指标，验证了用户体验与 APP 使用、内容资源之间的紧密关系；蔡燕等（2022）基于技术接受模型(Technology Acceptance Model, TAM)，构建了解释和预测中文学习者在线直播课程学习意愿的理论模型；梁宇等（2023）则更进一步以技术接受扩展模型为理论框架，构建了中文数字学习资源使用意愿模型，并且特别强调了感知易用性、感知有用性、使用态度具有关键的中介作用。由此可知，中文学习 APP 作为数字教育资源的一种创新形式，显著增强了学习的便捷性和互动性，促进了个性化学习路径的发展。因此，从用户体验视角出发，系统性地评价与分析用户对该类新兴数字资源的应用效果及内容反馈对于优化产品设计、提升教学效果至关重要。

## 2.5 基于 LLMs 的专家模型主题提取与效果评价

LLM 在多项基准测试中展现出媲美人类专家的水平和表现（Achiam et al., 2023）。提示工程作为一种有效引导 LLMs 的方法，通过专门设计的提示词或短语，能够在零样本（Kojima et al., 2022）或少量样本（Brown et al., 2020）条件下显著提升模型在特定 NLP 任务上的表现。进一步而言，群体智能决策机制能进一步强化

LLMs 的性能，甚至在某些任务上超越人类（Wu et al., 2023; Jang et al., 2023）。例如，何多魁等（2025）提出了一种微调大语言模型驱动的短文本动态主题建模方法，通过结合指令微调、检索增强生成(Retrieval-Augmented Generation, RAG)和聚类技术，有效提升了主题识别的准确度，并揭示了主题的演化规律，为主题建模提供了新的思路和方法。翟洁等（2025）则针对计算机实验报告评阅过程中评语模板化、缺乏个性化内容等问题，提出了基于 LLM 的个性化实验报告评语自动生成框架，通过主题-评价决策-集成提示策略，实现了从实验要求和代码质量需求中抽取评价体系，自动生成具有可解释性的实验评语，提高了评阅效率和质量 。Reuter（2024）介绍了 GPTopic 软件包，利用 LLM 创建动态、互动的主题表征，通过聊天界面让用户能够探索、分析和优化主题，使主题建模更加易于访问、更加系统全面。这些研究均体现了大语言模型在不同领域的应用潜力，以及在提升数据分析和决策支持方面的显著优势。基于此，本研究着重关注在提示工程与群体智能决策双重赋能下的 LLMs 在文本聚类主题提取任务的应用潜力，旨在探索这一策略如何实现高效自动化处理并显著提升文本聚类和主题提取的准确性与鲁棒性，同时减少对大量标注数据的依赖。

### 2.6 已有研究的启示与本研究的具体问题

构建中文学习 APP 的评估指标体系是一项极具挑战性且意义深远的工作。它涵盖众多多维度与多层次，需要全方位、多角度的综合考量。传统构建方式往往依赖专家的见解与经验，然而在当下技术革新日新月异、用户需求瞬息万变的大环境下，这种模式逐渐显露出局限性。与之相对地，数据驱动的评价模型凭借深度学习算法的强大能力，能够更加精准且灵活地适配当前中文教学的发展态势以及用户不断变化的需求，为构建全新的评价指标体系提供了崭新的视野。近年来，以 ChatGPT 为代表的 LLM 与教育领域的深度融合发展，更是对中文学习 APP 评价体系的构建以及用户体验感的提升产生了深远且重要的影响。

基于上述情况，本研究旨在构建一套智能、客观、动态、综合的中文学习 APP 评价指标体系，以实现教学资源评价的科学化、精准化和自动化。为此，提出以下研究问题：

（1） 如何设计并实现一个覆盖用户多样性与教学场景多变性的中文学习 APP 效能评价框架，以精准监控中文学习者的学习过程并有效评价 APP 的应用效能？

（2） 如何在构建与优化中文学习 APP 评价指标体系中，整合 LLMs 促进群体智能决策，确保评价体系的高效性、准确性和对技术动态的敏捷响应能力？

（3） 如何通过情感分析技术，结合中文学习 APP 的特点，深入挖掘用户对中文学习 APP 的情感倾向和具体反馈，从而为评价指标体系的验证和优化提供更具针对性和实用性的依据，进一步提升中文学习 APP 的用户体验和教育效果？

## 3. 研究思路

本研究借鉴文本特征向量融合的理念，融合了 BERT 模型的语义特征向量和 LDA 主题特征向量，进而设计了一种适用于中文学习 APP 短评文本的主题识别、评估指标构建及验证的整体框架，该框架如图 1 所示，具体实施步骤如下：

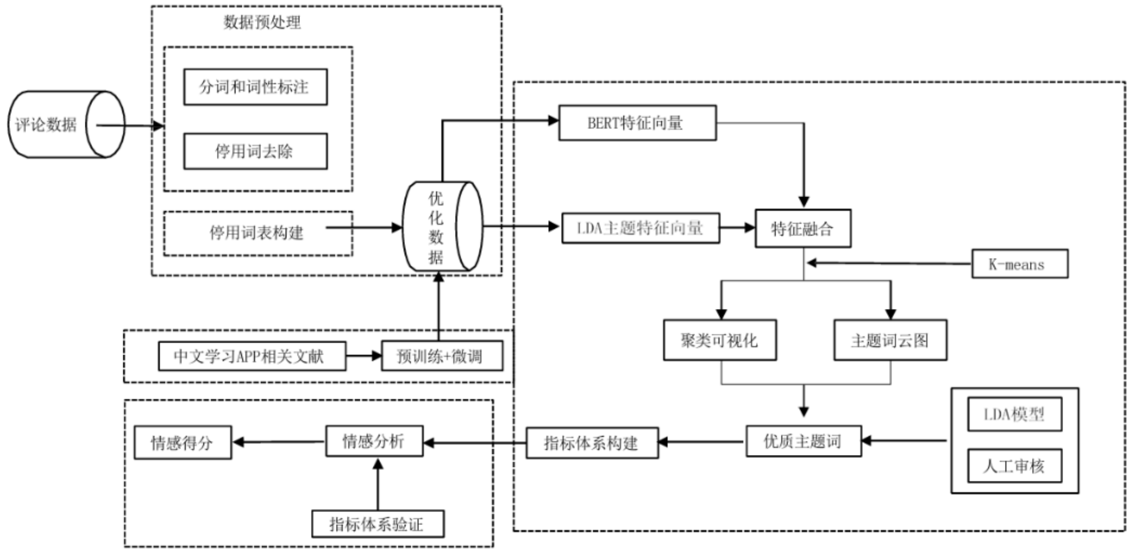


**图 1 中文学习 APP 评估指标体系构建和验证的整体框架**

### 3.1 数据采集与预处理阶段

第一，以爬虫软件“后羿”为数据采集工具[4]，从“七麦数据平台”[5]上抓取大量中文学习 APP 的用户短评文本，以此作为下游任务的训练数据集；第二，通过“中国知网”平台[6]整合中文学习相关的学术文献标题与摘要信息，从而构建预训练数据集；第三，整合百度、四川大学以及哈工大的通用停用词表[7]，并据此对原始数据进行深度筛选与结构化处理；第四，为增强模型主题提取方面的性能，本研究将 BERT 模型通过[CLS]标记符产生的综合文本向量与 LDA 模型生成的主题特征向量相结合，借助加权求和、拼接等方式进行特征融合，以构建融合深层语义信息及主题结构的复合特征向量。

### 3.2 K-means 聚类

首先，将语义和主题相近的关键词分成若干群组，通过此方法探究它们之间的深层联系。然后，通过计算困惑度来挑选 K-means 算法[8]的合适 K 值，以此来决定

---

[4] https://www.houyicaiji.com/

[5] https://www.qimai.cn/rank/featured

[6] https://www.cnki.net/

[7] https://www.csdn.net/

[8] https://baike.baidu.com/item/K%E5%9D%87%E5%80%BC%E8%81%9A%E7%B1%BB%E7%

恰当的主题数量；接着，为使聚类结果更加直观且易于理解，运用统一流形逼近与投影(Uniform Manifold Approximation and Projection，UMAP)算法[9]，对多维聚类结果进行了降维并进行了可视化处理；最后，在此基础上，构建对应主题的词云图，用以呈现各个主题的核心词汇组成及其相互之间的关系。

### 3.3 基于 LLMs 的专家模型主题提取方法

首先，通过文本聚类技术提炼出一系列主题，每个主题内都包含相关关键词。其次，引入群体智能体参与分析流程，以文本聚类主题下的关键词为处理对象，鉴定其作为构建中文学习 APP 评价指标体系的适用性。再次，引导智能体进一步深化执行关键词的语境分析任务，力图实现关键词内涵与既定评价理论体系的无缝对接，确保分析的深度与精度。最后，将所有智能体的分析结果集成为统一知识库，并经自一致性(Self-Consistency)投票机制进行过滤与强化，从而高信度地确立核心评价主题群集，为后续评价体系构建奠定坚实基础。

### 3.4 构建并验证中文学习 APP 评估指标体系

首先基于 LDA 主题模型挖掘用户评论中的核心主题特征，结合 BERT 语义特征与 LDA 主题特征进行多维度融合，构建涵盖功能体验、内容质量、用户情感等维度的评价指标框架。通过 K-means 聚类分析提炼高频主题词，筛选出与学习效果强相关的优质主题词作为核心评价维度。上述过程采用基于 SnowNLP[10]的情感分析技术对中文学习 APP 的短评文本数据进行情感得分量化处理。再将得出的情感得分与用户的实际评分进行对照，以此来检验评价指标体系的准确性与实效性。

## 4. 研究工具和方法

### 4.1 BERT 模型

（1）模型介绍：BERT 模型由 Google 公司在 2018 年 10 月推出，与传统的基于静态词嵌入的 Word2Vec 模型不同，BERT 在基于 Transformer 双向编码器架构的基础上将词在不同语境的文本特征纳入考虑（Devlin et al., 2019）。为了使模型能够进行跨任务应用，并能深入语境中捕捉文本语义联系，BERT 在其输入层融合了词向量(Token Embedding)、段落标识向量(Segment Embedding)以及位置向量(Position Embedding)这三种向量嵌入技术，同时融入独特的标记符[CLS]和[SEP]。一方面，在[CLS]的帮助下，模型可为整个序列创建统一的句子向量表征，有助于其执行分类任务。另一方面，[SEP]的作用在于划分和标识文本序列中的各个句子或片段，有助于模型在处理涉及多个句子或段落的情况下，依然能够维持文本顺序和结构信息的稳定性（如图 2 所示）。借助上述结构设计，BERT 模型有效实现了对文本内容

AE%97%E6%B3%95/15779627

[9] https://blog.csdn.net/CRUSH8496052/article/details/132926453

[10] https://developer.baidu.com/article/details/3330267

 

深入且全面的双向语境理解，增强了其在自然语言处理任务上的表现和精确度。

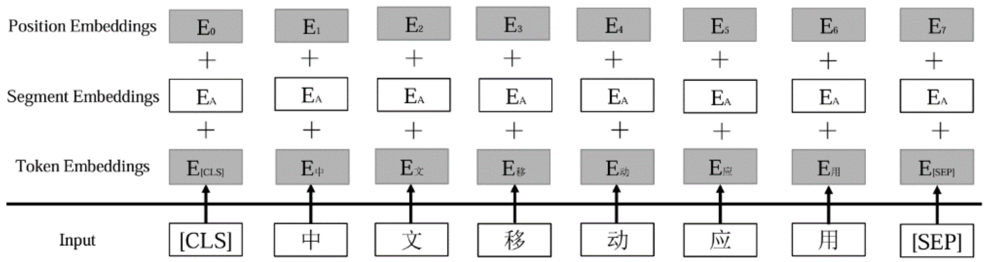


**图 2 BERT 的句子级表示**

（2）预训练: BERT 模型的训练起初采用大规模的无监督学习方法，这一过程涉及两个主要任务：一是遮蔽语言模型(Masked Language Model, MLM)；二是预测下一句(Next Sentence Prediction, NSP)。MLM 的目标是预测随机遮蔽情况下的词汇，模型必须依赖上下文信息来填充这些缺失，这样它就能学习到更加丰富的语言表达。而 NSP 任务则是评价两个连续的句子片段是否在逻辑上构成前后关系，其目的是辅助模型掌握文本的连贯性和整体结构。

（3）微调: BERT 模型在完成预训练之后，能够针对特定的自然语言处理任务进行微调。在微调过程中，经过有监督学习，模型会在某个特定的数据集上进行训练，这通常意味着在预先训练过的模型之上，仅需增加一个输出层便可满足任务需求，并在此基础上针对这一层进行细致的调整，可实现模型参数的精确优化。BERT 通过迁移学习的方法，在自然语言处理领域，例如文本分类、命名实体识别、情感分析等多个任务中都展现了广泛的应用价值。

### 4.2 LDA 模型

LDA 是一种无监督学习的概率生成模型，包含文档、主题和词语三层结构，其主要思想是：文档是由若干主题组成的，主题是由文档中一组特定词汇组成的，文档中的每个词都是以一定概率分布的，由此可将一篇文档的主题以出现频率最高的一组词汇表示。LDA 主题模型可以在文档、主题、词语三个层面进行概率建模，计算主题与文档、主题与词语之间的语义关联度，已在文本挖掘、信息检索和情感分析等领域得到了广泛应用（Blei et al., 2003），具体计算过程如图 3 所示。

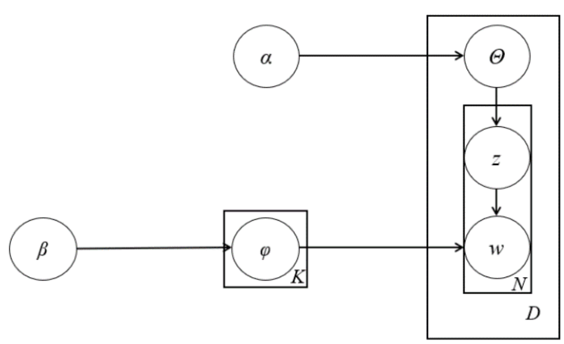


**图 3 LDA 主题模型**

图 3 中每个符号的含义见表 1，变量间的箭头表示条件依赖关系(Conditional Dependencies)，即文档(Documents)、主题(Topics)以及词语(Words)之间的生成概率关系。

**表 1 主题模型中各参数意义**

| 参数 | 描述 |
|---|---|
| $\alpha$ | 狄利克雷分布，$\theta$ 的超参数 |
| $\beta$ | 狄利克雷分布，$\varphi$ 的超参数 |
| $\theta$ | 评论-主题分布 |
| $\varphi$ | 主题-词分布 |
| $z$ | 评论中词语对应的主题 |
| $w$ | 评论中的词语 |
| $K$ | 主题数 |
| $M$ | 文档数目 |
| $N$ | 一篇文档的词数 |

### 4.3 BERT-LDA 模型

（1）构建 LDA 主题特征向量: 首先对原始文本数据集进行处理，运用 LDA 主题模型对其进行训练。通过无监督学习的方式，挖掘出文本潜在的主题分布。在 LDA 模型中，每个文本都被表示为一系列主题的概率分布，从而可以提取出每个文本对应的主题特征向量。这些向量记录了文本在各个主题上的权重信息。

（2）构建 BERT 语义特征向量: 使用 BERT 模型对预处理数据执行词嵌入操作，以此构建 BERT 语义特征向量。Transformer 编码器单元中，输入向量首先通过多头自注意力机制进行上下文依赖建模，随后经由残差连接与层归一化操作实现梯度稳定，继而通过前馈神经网络进行非线性空间变换并叠加二次残差连接，最终输出具有多层抽象特征的 BERT 语义向量表征（王秀红等, 2021）。

（3）BERT-LDA 特征向量融合: 借助加权求和、拼接以及深度神经网络融合等方法，融合文本特征表示，这种表示兼顾主题结构和深层语义信息，可优化自然语言处理任务的表现，如图 4 所示。

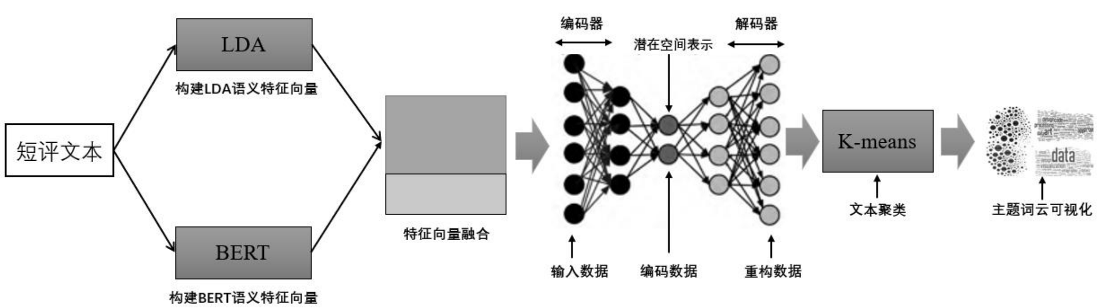


**图 4 BERT-LDA 模型示意图**

### 4.4 K-means 聚类及可视化

引入 BERT 语义特征向量对 LDA 主题特征向量进行补偿，虽然提升了文本高层次语义的保持和底层主题模式的捕捉能力，丰富了表达的多样性和深度，但向量拼接操作在信息稀少的高维空间中容易引发维度灾难[11]和过拟合[12]的问题。为此，本研究采用 K-means 算法，通过聚类实现降维，并提取关键词，以降低模型的复杂性且提升分类的效率。K-means 算法是一种无监督学习方法，擅长处理大量数据。该算法以欧氏距离为基准，将相似的数据点划分为同一类别，从而实现数据的聚类（Sinaga et al., 2020）。在进行聚类分析时，可通过评价潜在语义主题模型的困惑度找出最合适的主题数量，该数量将用作 K-means 算法中的 K 值。随着 K 值的逐步上升，模型的困惑度前期呈现降低趋势。然而，在达到一个局部最小值之后，如果继续提高K值，模型的表现会开始退化，出现过拟合现象，从而影响其泛化能力。本研究采用 UMAP 算法对特征空间进行非线性可视化处理，目的是为维护数据的全局与局部结构信息，直观展示高概率主题词及其对应概率，进而以此为基础，构建中文学习 APP 的评价指标体系。

### 4.5 基于情感计算的指标体系验证

基于情感词典的短评文本计算方法是一种利用预先建立的情感词汇库来量化分析文本情感倾向的技术，主要有以下步骤：首先对情感词权重赋值，其次对短评文本中情感词的位置进行定位，最后进行情感强度的计算。基于此，本研究调用 SnowNLP 库中的自然语言处理工具对中文学习 APP 短评文本进行情感打分，计算出每条短评文本的情感得分。接着，对大量短评文本的情感得分进行统计和分析，以获取整体的情感倾向分布。然后，根据分析结果对基于情感计算的指标体系进行验证和调整，确保其准确性和有效性。

## 5. 实验设计

### 5.1 数据集构建

本研究选择“中国知网”和“七麦数据”两个平台的中文学习 APP 相关文献以及短评文本作为数据采集对象。首先，在“中国知网”平台以“教育 APP”、“在线汉语/中文”、“汉语/中文技术”、“汉语/中文词典”以及“汉语/中文学习”为主题字段，检索得

---

[11] 维度灾难（Curse of Dimensionality），又称为维数灾难、维度诅咒，最早由美国数学家理查德・贝尔曼（Richard Bellman）在 20 世纪 50 年代末研究动态规划时提出。随着问题维度的增加，解决问题的难度呈指数级增长，计算量和存储需求等也急剧增加，使得问题变得难以处理。

[12] 过拟合（Overfitting）是指在机器学习和统计建模中，模型在训练数据上表现得过于完美，过度学习了训练数据中的噪声和细节，导致在新的、未见过的数据上表现不佳，泛化能力差的现象。

到 3459 篇，经人工筛选后删除了 356 无效文献，最终得到 3103 篇有效文本摘要数据，并将其作为语料用于增强 BERT 模型的语言理解能力，例如提升模型在识别特定 APP 名称、关键技术与教学模式等方面的准确性，更好地理解用户对“交互性”、“课程涉及”等维度的评价。其次，从“七麦数据”平台搜集了 17 款中文学习 APP 的用户短评，共 10866 条短评数据，将其作为特定领域语料用于下游任务的微调。这些 APP 涵盖了多种学习场景与用户群体，包括综合学习、词典查询、汉字书写、考试备考等类型，具备一定的市场代表性进而功能多样性。本次所选取的 17 款在用户基数、活跃度与功能类型上具有一定代表性，能够反映主流中文二语学习工具的使用体验和反馈特征。所选 APP 多数同时提供中文及英文名称，其开发者背景多样，既包括中国本土企业（如 HELLOCHINESE TECHNOLOGY CO., LTD.），也有 CHINEASY LTD.等国际团队(如表 2 所示)。这些应用主要面向非母语者，提供从零基础到高级水平的中文学习支持，包括词汇、语法、听力、阅读、写作等多个维度。所有 APP 均可通过 IOS APP Store 在全球多个地区下载，覆盖中国、美国、日韩、欧洲以及东南亚等广泛区域，具有较高的可获取性和使用普及度。

**表 2 中文学习 APP 评论数据信息**

| 序号 | 中文学习 APP 名称 | 开发者 | 评论数量 |
| --- | --- | --- | --- |
| 1 | ChineseSkill | YIYANTECHNOLOGY CO., LTD. | 3776 |
| 2 | HelloChinese | HELLOCHINESE TECHNOLOGY CO., LTD. | 2902 |
| 3 | PlecoChinese Dictionary | PLECO INC. | 1425 |
| 4 | LearnChineseEasily | CHINEASY | 1038 |
| 5 | Scripts:LearnChinesewriting | TOUCHSCREEN LEARNING LTD | 642 |
| 6 | Chineasy:LearnChineseeasily | CHINEASY LTD | 345 |
| 7 | DuChinese–ReadMandarin | SINAMON AB | 195 |
| 8 | ChineseParents | LITTORAL GAMES | 111 |
| 9 | DailyChinese\|Words&Idioms | MOJAY, LLC | 90 |
| 10 | MandarinChinesebyNemo | NEMO APPS LLC | 83 |
| 11 | LearnChinese-Mandarin | BRAINSCAPE | 79 |
| 12 | HSKStudyandExam-SuperTest | SHANGHAI YUXUAN INFORMATION TECHNOLOGY CO., LTD | 76 |
| 13 | DominoChinese | ZIMAD | 40 |
| 14 | HanYou-ChineseDictionary | Nomad AI OU | 28 |
| 15 | DotLanguages-LearnChinese | / | 17 |
| 16 | LearnChineseHSK1Chinesimple | KHANJI SCHOOL DIGITAL FACTORY SOCIEDAD LIMITADA | 13 |
| 17 | LearnChineseforBeginners | HECTOR GONZALEZ LINAN | 6 |
| | 合计 | / | 10866 |

### 5.2 数据预处理

鉴于从“七麦数据”平台所采集的短评文本数据涵盖了英语、俄语、法语等多种语言，为确保后续对这些文本进行一致性处理，首先将所有非中文的短评翻译转写为中文版本。然而，直译过程中往往难以避免情感色彩和初始语义信息的部分损失。为此，本研究利用 LLM 并结合提示技术，针对性地设计了适用于机器翻译任务的提示策略，旨在最大程度上缓解统一翻译过程中可能产生的语义流失问题。其次，针对收集到的中文学习 APP 短评文本，进行系统化的数据预处理步骤：第一，整合了百度、四川大学及哈尔滨工业大学发布的停用词表，通过去除文本摘要中的常见停用词，有效地减少无关噪音信息的影响。第二，通过统计分析文本中的高频词汇，并基于其对主题内容的实质性贡献度，过滤了诸如空格、标点符号等出现频率较高但贡献微弱的词汇单元。

### 5.3 文本聚类与主题分析

（1）困惑度计算:困惑度是确定主题模型最优主题数目的重要判断指标，困惑度值越小，模型泛化能力越强，当前主题数目就越优（关鹏等, 2016），然而随着主题数增大，提取的主题噪声也会随之增多。因此，本研究将模型最终主题数选定为 6 个。如图 5 所示：

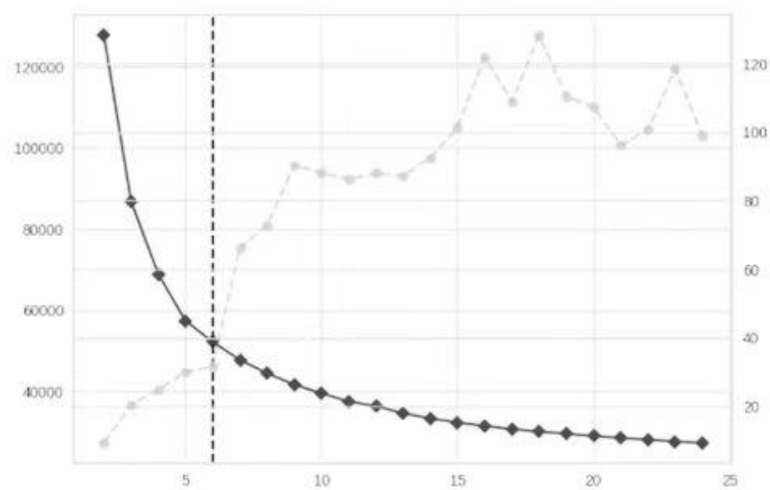


**图 5 BERT-LDA 模型不同主题数的困惑度变化**

（2）Umap 聚类可视化：通过 UMAP 算法将高维文本向量降维至二维空间并进行可视化。结果显示，不同主题对应的文本在低维空间中形成分布清晰的聚类簇，且各簇间边界较为明确，表明模型能够有效区分语义差异。尽管部分主题簇存在局部重叠，反映了主题间的潜在关联性，但整体聚类结构与预设的 6 个主题数较为吻合，如图 6 所示：

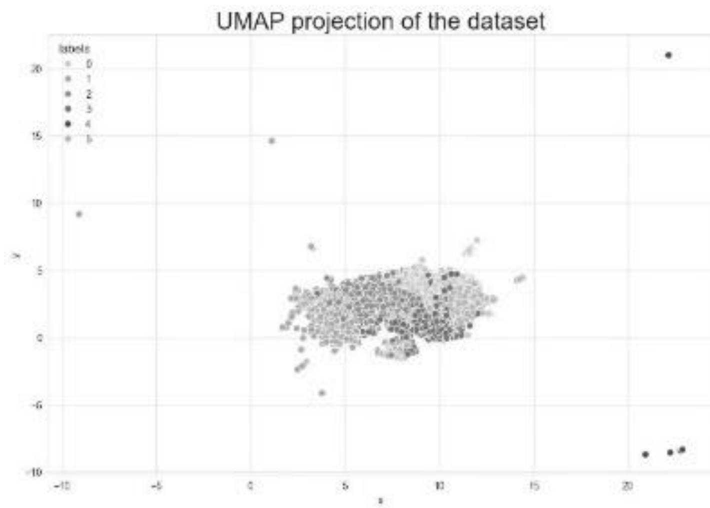


**图 6 基于 UMAP 的二维聚类可视化**

（3）词云图：基于 BERT-LDA 模型所识别出的 6 个相关主题，选择每个主题下的前 40 个核心词汇进行深入的可视化探索，这一过程旨在通过构建词云图，直观展示这 6 个主题的词汇分布特征(图 7)。


(a)主题 1　(b)主题 2　(c)主题 3


(d)主题 4　(e)主题 5　(f)主题 6

**图 7 中文学习 APP 短评词云图**

（4）主题关键词：由表 3(下页)可知，BERT-LDA 主题模型提取的主题为 6 个，6 个主题类型分别为“多媒体学习与交互设计”“学习效率与标准测试”“工具创新与社区互动”“语言技能与文化传播”“用户参与与个性化服务”“文化沉浸与深度学习”。

## 5.4 基于 LLMs 的专家模型主题提取与效果评价

基于 LLMs 的专家模型主题提取与效果评价的核心原理是基于提示去引导群体智能体，并利用自一致性投票机制协同自适应学习策略执行主题提取与识别任务，如图 8(下页)所示。这主要包括主题识别提取算法的定义、主题识别矩阵构建、效果评价等步骤，旨在通过智能化协同提升主题识别的精度与效率，并通过定量与定性分析确保评价的全面性与客观性。

**表 3 中文学习 APP 文本聚类主题关键词**

| 序号 | 主题 | 关键词 |
|---|---|---|
| 1 | 多媒体学习与交互设计 | 沉浸、听力、视频、翻译、闪卡、设计、界面、互动性、流利、教学 |
| 2 | 学习效率与标准测试 | 记忆 、最佳、笔画、关注、制作、激励、值得、直观、语音、HSK |
| 3 | 工具创新与社区互动 | 工具、程序、结构、出色、困难、分享、对话、声调、测验、轻松 |
| 4 | 语言技能与文化传播 | 语言、故事、教授、开发、声调、精彩、西班牙语、说话、答案 |
| 5 | 用户参与与个性化服务 | 购买、美好、关注、反馈、电子邮件、发送、故事、抖音、口音、老师 |
| 6 | 文化沉浸与深度学习 | 深刻、繁体字、教授、改进、笔画、视频、易用性、访问、文化、经典著作 |

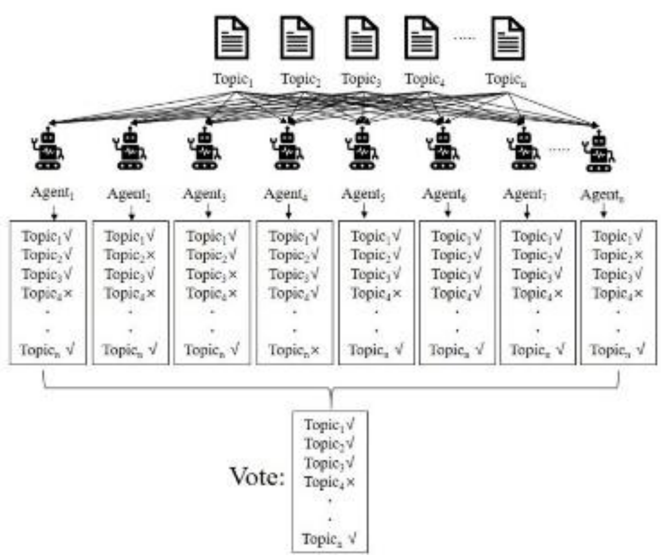


**图 8 LLM 自一致性投票机制**

（1）基于 LLMs 的主题识别提取算法：本研究定义了适用于 LLMs 处理的主题识别提取算法流程，如表 4 所示。具体来说，包括符号定义、智能处理、筛选准则、综合评价与排序以及集体决策与输出等 7 个流程。算法核心包括智能体执行函数，该部分在于引导智能体根据自身逻辑识别出与主题相关的关键词子集。随后，通过筛选准则去除与评价指标语义不相关的关键词，确保关键词的高关联度。综合评价与排序步骤中，每个智能体对其识别的主题根据相关关键词数量进行排序并优选，排除关键词数量不足的主题。最后，集体决策与输出阶段采用投票机制，比较各主题在不同智能体排序中的流行程度，仅保留那些出现频率超过预设阈值的主题，作为构建中文学习 APP 评价指标的最终主题集合。

**表 4 基于 LLM 的主题识别提取提示构建**

| | |
|---|---|
| （1）定义符号 | ·$T=\{T_1, T_2 \ldots T_i\}$，其中每个 $T_j=\{K_{j1}, K_{j2}, \ldots, K_{jn}\}$，表示序列 T 包含 $i$ 个主题，每个主题有 n 个关键词。$R=\{A_1, A_2, \ldots, A_a\}$，表示序列 R，由 a 个智能体组成。<br>·$C=\{I_1, I_2 \ldots I_d\}$，表示一级评价指标的集合。 |
| （2）智能体执行函数 | ·对于每个智能体 $A_r \in R$ 和每个主题 $T_j \in T$，定义识别函数 $f_r(T_j)=K'_{rj}$, 其中 $K'_{rj} \subseteq T_j$ 为智能体 $A_r$ 认定的相关关键词集合。 |
| （3）筛选准则 | ·定义筛选函数 $g(K'_{rj})= K''_{rj}$,使得 $K''_{rj} \in K'_{rj}$，且 $k \in K''_{rj}$，存在 $c \in C$ 使得 $k$ 与 $C$ 在语义上相关。 |
| （4）综合评价与排序 | ·对于每个智能体 $A_r$，定义排序和选择函数 $h(A_r)=S_r$,其中 $S_r$ 是按与一级评价指标相关的关键词数量降序排列的主题集合，且$\lvert S_r \rvert < I$,表示除去了一些关键词数量较少的主题。 |
| （5）集体决策与输出 | ·定义投票函数 $v(R,S)=F$，其中 $F \subseteq T$，基于所有智能体 $R$ 的排序结果 $S=\{S_1, S_2, \ldots, S_a\}$,通过比较各主题在不同智能体排序中的出现频率，选择频率高于预设阈值的主题作为最终结果。<br>·输出结果为$F$，代表经集体决策确定的，用于构建中文学习 APP 评价指标的主题集合。 |
| （6）输入主题 | $[T_1, T_2 \ldots T_i]$ |
| （7）输出结果 | $[T'_1, T'_2 \ldots T'_i]$ |

（2）基于 LLMs 的专家模型主题识别矩阵：表 5 展示了不同 LLM 模型对前述定义的各个主题（Topic1 至 Topic6）的支持情况。符号“✓”表示对应主题能够为 LLM 有效识别，并能够作为中文学习 APP 评价指标体系构建的基础，而“×”则表示支持度较低或不具备直接关联的主题。通过汇总各模型对聚类主题的支持情况，并在最后一行统计出每个主题的投票支持率，可直观反映各聚类主题识别提取情况。

**表 5 基于 LLM 的专家模型主题识别矩阵**

| | Topic1 | Topic2 | Topic3 | Topic4 | Topic5 | Topic6 |
|---|---|---|---|---|---|---|
| Agent1 | ✓ | ✓ | ✓ | ✓ | × | × |
| Agent2 | ✓ | ✓ | ✓ | ✓ | × | ✓ |
| Agent3 | ✓ | × | ✓ | × | × | ✓ |
| Agent4 | ✓ | × | ✓ | ✓ | ✓ | × |
| Agent5 | × | × | ✓ | × | ✓ | ✓ |
| Agen6 | ✓ | ✓ | × | ✓ | × | ✓ |
| Agent7 | ✓ | × | ✓ | × | × | ✓ |
| Agent8 | ✓ | ✓ | ✓ | ✓ | × | ✓ |
| Agent9 | × | ✓ | ✓ | × | ✓ | ✓ |
| Agent10 | ✓ | × | ✓ | × | × | ✓ |
| 投票支持率 | 80% | 50% | 80% | 50% | 30% | 80% |

（3）评估指标：为全面评价基于 BERT-LDA 模型的主题聚类效果，本研究采用了以下 3 个评价指标：查准率(*Precision*)、查全率(*Recall*)与 $F$ 值(*F-measure*)，以下分别用 $P$，$R$ 和 $F$ 表示。这些指标帮助本研究从不同维度理解模型识别主题的准确性和全面性。其中 $T_{correct}$ 是指 LLM 专家模型识别的正确主题数量，$T_{extract}$ 是指基于 LLM 专家模型提取或识别出的主题数量，$T_{standard}$ 是指专家总结出的主题数量。

（1）$P = \frac{T_{correct}}{T_{extract}}$

（2）$R = \frac{T_{correct}}{T_{standard}}$

（3）$F = \frac{2PR}{P+R}$

（4）评价结果：本研究评价对比了 BERT-LDA 模型与传统 LDA 模型在主题抽取任务上的性能。从结果可以看出前者在查准率、查全率以及 $F$ 值上均优于后者，分别高了 16.07%、33.3%以及 27.96%。BERT-LDA 模型识别出的主题类别更加清晰、准确，其中主题 1、主题 3 和主题 6 更有利于构建中文学习 APP 评价指标体系。

**表 6 不同模型主题抽取效果对比结果**

| | 主题数 | $T_{extract}$ | $T_{correct}$ | $T_{standard}$ | 查准率 | 查全率 | F 值 |
|---|---|---|---|---|---|---|---|
| BERT-LDA | 6 | 6 | 3 | 6 | 50% | 50% | 50% |
| LDA | 3 | 3 | 1 | 6 | 33.3% | 16.7% | 22.04% |

## 6. 中文学习 APP 评估指标体系构建及验证

### 6.1 中文学习 APP 评估指标体系构建

基于 BERT-LDA 模型和 K-means 聚类，同时运用基于 LLMs 的主题识别提取方法对“多媒体学习与交互设计”“学习效率与标准测试”“工具创新与社区互动”、“语言技能与文化传播”“用户参与与个性化服务”“文化沉浸与深度学习”等 6 个主题及其主题关键词进行了识别，并基于主题 1、主题 3 和主题 6，最终归纳总结得到中文学习 APP 评价指标体系。需要特别指出的是，本研究中归纳总结出的主题及最终构建的评价指标体现（如表 7 所示），均由机器基于数据挖掘和算法模型自动生成，整个过程未引入人类专家进行审核或验证。这种纯粹数据驱动的方法虽然保障了规模与效率，但也可能引入算法固有的偏见，例如不能捕捉到凭借专家经验才能洞察的深层教学逻辑与核心质量维度。

**表 7 中文学习 APP 评价指标体系**

| 一级维度 | 二级维度 | 三级维度 | 指标描述 |
|---|---|---|---|
| 内容评价（内容质量） | 内容组织 | 语言表达 | 清晰准确，符合语法规范 |
| | | 词汇覆盖 | 广泛多样，适应不同水平 |
| | 视听融合 | 音质协调 | 声音清晰，无杂音干扰 |
| | | 视频指导 | 情景模拟，直观展示应用 |
| 过程评价（用户体验） | 信息架构 | 页面布局 | 简洁高效，便于信息查找 |
| | | 导航设计 | 逻辑清晰，快速定位功能 |
| | 交互体验 | 流畅体验 | 响应迅速，操作无卡顿 |
| | | 平台兼容 | 多系统适配，运行稳定 |
| | 更新迭代 | 问题修复 | 及时反馈，解决用户难题 |
| | | 版本优化 | 持续改进，提升应用性能 |
| 效果评价（学习成效） | 实用工具 | 笔顺演示 | 正确书写顺序，动画展示 |
| | | 字典查询 | 例句丰富，快速查词解义 |
| | 技能应用 | 听力强化 | 多场景练习，增强交流能力 |
| | | 文化适应 | 跨文化融入，提升理解能力 |

## 6.2 中文学习 APP 短评情感分析

（1）中文学习 APP 短评文本评分分布：调用 SnowNLP 库对中文学习 APP 的短评进行情感计算，旨在对 17 款中文学习 APP 的情感倾向进行深入分析，即积极、消极或中性等评论数量的占比。基于此，本研究对 10866 条评论进行情感分析，并量化了每款 APP 的情感得分，进而得出中文学习 APP 总体短评情感分布情况，见表 8。

**表 8 中文学习 APP 短评情感分布情况**

| 情感类型 | 评论量 | 占比 |
|---|---|---|
| 积极情感 | 10132 条 | 93.25% |
| 中性情感 | 311 条 | 2.87% |
| 消极情感 | 422 条 | 3.88% |

（2）基于情感词典的中文学习 APP 短评文本计算：本研究选择了适合中文语境的情感词典，结合了否定词识别、程度副词权重调整等方法，以提高情感分析的准确性。并以 *HanYou-Chinese Dictionary* APP 的部分短评情感得分为示例结果，如表 9 所示。

 

**表 9 "HanYou - Chinese Dictionary"APP 部分短评情感得分**

| 序号 | 评论内容 | 情感得分 |
|---|---|---|
| 1 | 我特别喜欢绘图词典。 | 92.11% |
| 2 | 汉友确实帮助我增加了中文词汇量。闪存卡也很有用。 | 82.48% |
| 3 | 超级有用且易于使用。强烈推荐给汉语学习者和来中国的游客！ | 95.71% |
| 4 | 我建议该APP可以包含一些同义词及有关定义的更多详细信息。 | 24.69% |
| 5 | 即使它是免费的，它仍然是垃圾。 | 10.48% |

基于评论量的梯度分布性与情感均值的层级覆盖性双重筛选原则，选取用户评分排名前三的 3 款中文学习 APP 作为研究对象进行深入分析，如表 10 所示。通过该分析，可以深入了解用户反馈的情感倾向，为 APP 开发者提供改进建议，并帮助潜在用户做出更明智的选择。

**表 10 部分中文学习 APP 短评情感分布情况**

| 中文学习 APP 名称 | APP 分数 | 评论量 | 情感均值 |
|---|---|---|---|
| Du Chinese – Read Mandarin | 4.7 | 195 条 | 87.09% |
| Domino Chinese | 4.6 | 40 条 | 83.63% |
| Learn Chinese HSK1 Chinesimple | 4.7 | 13 条 | 92.22% |

*Learn Chinese HSK1 Chinesimple* 和 *Du Chinese–Read Mandarin* 两款 APP 不仅获得了较高的平均评分，而且用户情感均值也相对较高，表明这两款 APP 在用户满意度方面表现突出；相比之下，*Domino Chinese* APP 在内容深度及用户体验优化方面需进一步改进。（3）基于中文学习 APP 的多维评价：*Chinese Parents* APP 在所选的中文学习 APP 中评分最低，其在内容、用户体验以及学习成效等方面可能存在较多有待改进的地方，因此选择 *Chinese Parents* APP 作为验证中文学习 APP 评价框架的主要对象，期望通过对其进行详细评价，为提升此类 APP 的整体质量和用户体验提供有价值的案例借鉴。具体分析结果如表 11(下页)所示，展示了该 APP 在各个评价维度上的情感均值及用户反馈的详细情况。

## 7. 讨论与分析

### 7.1 研究价值

（1）理论贡献

第一，推动中文教育数字化与智能化转型。通过整合 BERT-LDA 模型与 K-means 聚类算法，并结合 LLMs 的主题识别与提取方法，本研究在"内容质量—用户体验—学习成效"三维框架下构建了系统的评价指标体系（见表 7）。这一过程不仅

验证了人工智能驱动评价体系构建的可行性，也为教育理论与技术融合提供了新路径。

**表 11 “Chinese parents”APP 短评文本多维评价结果**

| 一级维度 | 二级维度 | 情感均值 | 三级维度 | 情感均值 |
|---|---|---|---|---|
| 内容评价（内容质量） | 内容组织 | 83.21% | 语言表达 | 86.25% |
| | | | 词汇覆盖 | 80.17% |
| | 视听融合 | 47.71% | 音质协调 | 32.74% |
| | | | 视频指导 | 62.67% |
| 过程评价（用户体验） | 信息架构 | 50.03% | 页面布局 | 52.25% |
| | | | 导航设计 | 44.81% |
| | 交互体验 | 49.31% | 流畅体验 | 42.18% |
| | | | 平台兼容 | 56.44% |
| | 更新迭代 | 70.35% | 问题修复 | 65.47% |
| | | | 版本优化 | 75.22% |
| 效果评价（学习成效） | 实用工具 | 86.65% | 笔顺演示 | 84.74% |
| | | | 字典查询 | 88.51% |
| | 技能应用 | 77.84% | 听力强化 | 90.27% |
| | | | 文化适应 | 65.41% |

第二，扩展评价方法论的适用性。在对 10866 条用户短评的情感计算中，本研究验证了基于 SnowNLP 与情感词典结合的方法能够兼效率与精准性。例如，*HanYou-Chinese Dictionary* APP 在部分评论中的情感得分差异显著，显示了细颗粒度分析在揭示用户真实情感上的独特价值。这为教育技术评价提供了新的方法论支撑。

第三，推动个性化与动态化评价的形成。在 *Chinese Parents* APP 的多维评价结果中。用户对“听力强化（90.27%）”表现高度认可，但对“音质协调（32.74%）”则明显不满。这种维度差异凸显了动态指标不仅要面向整体水平，更要识别局部薄弱环节，为后续模型的个性化适配提供了理论依据。

（2）实践意义

第一，服务教师教学与资源甄选。对于外语教师而言，本研究的成果可直接应用于课堂资源筛选与教学辅助。例如，教师可参考 APP 在“词汇覆盖”或“文化适应”维度的情感均值，判断该应用是否适合于初级、中级或跨文化教学场景，从而提升教学资源配置的科学性。

第二，提升教育决策的数据驱动水平。基于情感分布结果（积极评估占比 93.25%），教师与教育管理者能够快速了解不同 APP 的整体口碑，并通过多维度细分（如“信息架构”“交互体验”）洞察学生学习中的真实痛点。这使得教学管理更具针对性和实效性。

第三，促进教育市场竞争与应用生态优化。评价框架不仅帮助开发者精准把握用户需求，还能为学生与教师提供直观的参考。例如，*Du Chinese–Read Mandarin* 和 *Learn Chinese HSK1 Chinesimple* 两款 APP 的情感均值分别达到 87.09%和 92.22%，对教师而言，这类数据有助于有限推荐更受认可的资源，提升课堂学习效果。

## 7.2 发展建议

优化和完善基于 BERT-LDA 的中文学习 APP 评价指标体系的构建，不仅能提升其实用性和科学性，还能为中文教学资源的优化发展提供有力的数据支持和决策依据。由此，本文从智能化、动态化、安全性三个层面对其评估指标体系构建提出建议：

（1）智能化导向。结合 LLM 与大数据分析，教师可依赖评价系统快速识别适配不同学习水平的资源。例如，在 APP 教学应用中，系统可自动生成"听力练习难度梯度"或"词汇拓展路径"，为教师布置差异化作业提供支持。同时，未来模型还应在跨语种支持与文化内容融合等方面加大投入，提升全球推广的适配度。

（2）动态化适应。面对互联网教育资源更新迅速的现实，指标体系必须保持灵活可扩展。例如，*Chinese Parents* APP 在"页面布局"和"导航设计"上评分偏低，若能及时反馈并优化，则可快速提升用户体验。教师在选择 APP 时，也能依靠这种动态评价，避免教学过程中因工具落后而导致的学习障碍。

（3）安全性保障。在教育应用的推广中，教师和学生的数据安全问题尤为关键。未来的评价框架需要纳入"隐私保护"维度，并遵循国际数据保护法规。这不仅关系到用户信任度，也直接影响教育资源的可持续发展和跨国推广。

（4）人机协同与专家介入。本研究所构建的指标体系完全基于机器算法，虽展现了自动化处理的潜力，但缺乏教育领域专家的深度干预。未来研究应积极探索"人机协同"(Human-AI Collaboration)的混合模式，将本研究的数据驱动方法与德尔菲法相结合。例如。可以在机器初步生成主题和指标以后，引入国际中文教育领域的专家和一线教师进行多轮审议、修正与验证，对机器可能存在的偏差进行校准，对指标的权重和表述进行优化。这种融合了算法广度与专家深度的模式，有望构建出既客观全面又符合教育理论与实践的评价体系。

# 8. 结语

本研究通过整合 BERT-LDA 模型与 LLMs，不仅科学构建了一套科学的中文学习 APP 评价指标体系，更重要的是，通过对真实用户数据的深入挖掘，验证了该体系的有效性和实用性。研究表明将 BERT-LDA 模型与基于 LLMs 引导的群体智能决策相结合的评价方法，能有效构建客观、动态的中文学习 APP 评价指标体系，

并通过情感分析验证了该体系在精准反映用户体验与学习成效方面的实用性与科学性。未来研究应进一步探索评价体系的自动化与自适应机制，融合更多维度的用户数据，并加强对数据安全与伦理问题的关注，以推动更加智能、全面的国际中文教育数字资源评价生态。

**致谢：**本文受教育部人文社会科学重点研究基地重大项目“国际中文教育数字资源综合评价理论与方法研究”（22JJD740016）的资助。

## 参考文献

Achiam, J., Adler, S., Agarwal, S., et al. (2023). Gpt-4 technical report. *arXiv Preprint arXiv*:2303.08774. https://doi.org/10.48550/arXiv.2303.08774

Alon, I., Haidar, H., Haidar, A., & Guimon, J. (2025). The future of artificial intelligence: Insights from recent Delphi studies. *Futures*, 165, 103514. https://doi.org/10.1016/j.futures.2024.103514

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent dirichlet allocation. *Journal of Machine Learning Research, 3*(Jan), 993-1022. https://dl.acm.org/doi/abs/10.5555/944919.944937

Brown, T., Mann, B., Ryder, N., et al. (2020). Language models are few-shot learners. *Advances in Neural Information Processing Systems*, 33, 1877-1901. https://doi.org/10.48550/arXiv.2005.14165

Cai, Y., & Wang, Z. (2022). Research on the Learning Intention of Chinese Learners in Live Courses Based on the Technology Acceptance Model. *Language Teaching and Research*, (5), 35-46. [蔡燕, 汪泽. (2022). 基于技术接受模型的中文学习者直播课程学习意愿研究. *语言教学与研究*, 5, 35-46.]

Cheng, T., & Wang, Z.Q. (2024). Research on the construction of a cross-cultural professional competence model for "Chinese + Vocational Skills". *China Vocational and Technical Education*, (1), 59-70. [程涛, 王正青. (2024). “中文+职业技能”人才跨文化职业胜任力模型构建研究. *中国职业技术教育*, 1, 59-70.]

Deng, X., Li, J.M., Zeng, H.J., et al. (2012). Analysis and application research on weight calculation methods in the analytic hierarchy process. *Mathematical Practice and Knowledge,* (7), 93-100. [邓雪, 李家铭, 曾浩健, 等. (2012). *层次分析法权重计算方法分析及其应用研究. 数学的实践与认识,* 7, 93-100.]

Devlin, J., Chang, M. W., Lee, K., et al. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv Preprint* arXiv:1810.04805. https://aclanthology.org/N19-1423/

Fang, Z.F., & Xu, J. (2023). Construction of an indicator system for digital literacy of international Chinese teachers. *Journal of Tianjin Normal University (Social Sciences Edition)*, (6), 25-33. [方紫帆, 徐娟. (2023). 国际中文教师数字素养指标体系建构研究. *天津师范大学学报(社会科学版)*, 6, 25-33.]

Gao, C.Z., & Ge, Z.Y. (2025). A study of the supply and demand of the interactive digital products of international Chinese-language education and its supply strategies. *Journal of Yunnan Normal University (Humanities and Social Sciences*

*Edition)*, (1), 53-60. [高传智, 戈兆一. (2025). 数字交互式国际中文教育产品供需状况与供给策略研究. *云南师范大学学报(哲学社会科学版)*, (1),53-60.]

Gong, X., & Liang, Y. (2023). Construction of a basic framework for evaluation indicators of international Chinese textbooks based on descriptive corpus. *Journal of Research on Education for Ethnic Minorities*, (3), 161-167. [宫雪, 梁宇. (2023). 基于描述语库的国际中文教材评价指标基础框架构建. *民族教育研究*, 3, 161-167.]

Guan, P., & Wang, Y.F. (2016). Study on the determination of optimal number of topics in LDA topic model in scientific and technological intelligence analysis. *Modern Library and Information Technology*, (9), 42-50. [关鹏, 王曰芬. (2016). 科技情报分析中 LDA 主题模型最优主题数确定方法研究. *现代图书情报技术*, 9, 42-50.]

Guo, J., Wu, Y.H., Gu, L., et al. (2021). Current status and prospects of digital resource construction in international Chinese education. *International Chinese Language Teaching Research*, (4), 86-96. [郭晶, 吴应辉, 谷陵, 等. (2021). 国际中文教育数字资源建设现状与展望. *国际汉语教学研究*, 4, 86-96.]

Jang, M. E., & Lukasiewicz, T. (2023). Consistency analysis of ChatGPT. *arXiv Preprint* arXiv:2303.06273. https://doi.org/10.48550/arXiv.2303.06273

Jia, H. N., & Chen, L. H. (2023). Topic mining and indicator construction of clothing brand information in online social networks. *Wool Textile Science & Technology, 51*(1), 121-129. [贾海楠, 陈李红. (2023). 在线社交网络中服装品牌信息主题挖掘及其指标构建. *毛纺科技, 51*(1), 121-129.]

Kojima, T., Gu, S. S., Reid, M., et al. (2022). Large language models are zero-shot reasoners. *Advances in Neural Information Processing Systems*, 35, 22199-22213. https://doi.org/10.48550/arXiv.2205.11916

Kharat, M. G., Kamble, S. J., Raut, R. D., & Kamble, S. S. (2016). Identification and evaluation of landfill site selection criteria using a hybrid Fuzzy Delphi, Fuzzy AHP and DEMATEL based approach. *Modeling Earth Systems and Environment*, *2*(2), 98. https://link.springer.com/article/10.1007/s40808-016-0171-1

Lai, W. (2023). Deep learning network-based evaluation method of online teaching quality of international Chinese education. *3c Tecnología: glosas de innovación aplicadas a la pyme, 12*(1), 87-106. https://dialnet.unirioja.es/servlet/articulo?codigo=8881472

Li, S. S., Liu, F., Cao, H. Y. (2025). An analysis of Chinese app teaching design from the perspective of mobile microlearning theory: A case study of *HelloChinese*. *International Chinese Language Education*, (1), 112-127+142. [李姝姝, 刘芳, 曹洪豫. (2025). 移动微学习理论视角下的中文 App 教学设计分析——以 HelloChinese 为例. *国际中文教育(中英文)*, 1, 112-127+142.]

Li, T. Y., & Liu, Q. M. (2024). Construction of an evaluation indicator system for painting works based on BERT-LDA and K-Means clustering. *Software Engineering,* (1), 68-73. [李天义, 刘勤明. (2024). 基于 BERT-LDA 和 K-means 聚类的绘画作品价值评估指标体系构建. *软件工程,* 1, 68-73.]

Liang, Y., & Li, N. E. (2023). Construction of an evaluation indicator system for international Chinese textbooks—Based on the Delphi Method and the Analytic Hierarchy Process. *Journal of Guizhou Normal University (Social Sciences Edition)*, (6), 30-40. [梁宇, 李诺恩. (2023). 国际中文教材评价指标体系构建——基于德尔菲法和层次分析法. *贵州师范大学学报(社会科学版)*, 6, 30-40.]

Liang, Y., & Li, N.E. (2023). Research on the intention to use Chinese digital learning resources and its influencing factors: Based on the extended TAM Model. *Language and Text Application*, (2), 23-35. [梁宇, 李诺恩. (2023). 中文数字学习资源使用意愿及其影响因素研究——基于 TAM 扩展模型. *语言文字应用*, 2, 23-35.]

Liu, Y. J. (2021). Research on the optimization path of lexicographical integrated publishing: A review of the Modern Chinese Dictionary (7th Edition) APP. *Journal of Beijing Union University (Humanities and Social Sciences Edition),* (2), 109-115. [刘永俊. (2021). 辞书融合出版的优化路径研究——兼评《现代汉语词典》(第 7 版)APP. *北京联合大学学报(人文社会科学版),* 2, 109-115.]

Pan, X. Y., Ni, Y., Jin, C. H., Zhang, J. (2023). Extraction of value elements and construction of an indicator system for calligraphy works based on Hyperplane-BERT-Louvain optimized LDA model. *Data Analysis and Knowledge Discovery,* (10), 109-118. [潘小宇, 倪渊, 金春华, 张健. (2023). 基于超平面-BERT-Louvain 优化 LDA 模型的书法作品价值要素提取及指标体系构建. *数据分析与知识发现,* 10, 109-118.]

Sinaga, K. P., & Yang, M. S. (2020). Unsupervised K-means clustering algorithm. *IEEE Access*, 8, 80716-80727. https://ieeexplore.ieee.org/abstract/document/9072123

Wang, C. Z., & Si, Q. (2011). Data statistical processing methods and their application research in the Delphi Method. *Journal of Inner Mongolia University of Finance and Economics (Comprehensive Edition),* (4), 92-96. [王春枝, 斯琴. (2011). 德尔菲法中的数据统计处理方法及其应用研究. *内蒙古财经学院学报(综合版),* 4, 92-96.]

Wang, X. H., & Gao, M. (2021). Key technology identification method based on BERT-LDA and empirical research: A case study of agricultural robots. *Library and Information Service,* (22), 114-125. [王秀红, 高敏. (2021). 基于 BERT-LDA 的关键技术识别方法及其实证研究——以农业机器人为例. *图书情报工作,* 22, 114-125.]

Wu, Q, Bansal, G, Zhang, J, Wu, Y, Li, B, Zhu, E, Jiang, L, Zhang, X, Zhang, S, Liu, J, Awadallah, A, White, R, Burger, D, Wang, C. (2023). Autogen: Enabling next-gen LLM applications via multi-agent conversation framework. *arXiv Preprint* arXiv:2308.08155. https://openreview.net/forum?id=BAakY1hNKS

Wu, T; He, S; Liu, J; Sun, S; Liu, K; Han, Q. (2023). A brief overview of ChatGPT: The history, status quo and potential future development. *IEEE/CAA Journal of Automatica Sinica, 10*(5), 1122-1136. DOI: 10.1109/JAS.2023.123618

Yang, Q. (2018). Implications for the construction of Chinese language network resources from Chinese learning APPs: A case study of iChinese APP. *Publishing Horizon*, (14), 77-79. [杨倩. (2018). 中文学习 APP 对汉语网络资

源建设的启示——以 iChinese APP 为例. *出版广角*, 14, 77-79.]

Yang, T., Xu, T., & Li, Q. (2023). Construction, empirical study, and optimization of an indicator system for intelligent competence of international Chinese teachers. *Journal of Yunnan Normal University (Teaching & Studying Chinese as a Foreign Language Edition),* (3), 41-52. [杨甜, 许桐, 李琴. (2023). 国际中文教师智能素养指标体系的构建、实证及优化. *云南师范大学学报(对外汉语教学与研究版),* 3, 41-52.]

Yang, X. H (2019). Design thinking training: A way out of the dilemma of thinking teaching. *China Educational Technology,* (7)*,* 54-59[杨绪辉. (2019) 设计思维培养: 基础教育思维教学困境的出路. *中国电化教育*, 7, 54-59.]

Yang, X. M., Luo, J. J., Liu, Y. X., Chen, S. C. (2017). Data-driven teaching: New directions of teaching paradigms in the era of big data. *Research in Electro-Education,* (12), 13-20+26. [杨现民, 骆娇娇, 刘雅馨, 陈世超. (2017). 数据驱动教学:大数据时代教学范式的新走向. *电化教育研究,* 12, 13-20+26.]

Yu, X. F., & Fu, D. (2004). Overview of multi-indicator comprehensive evaluation methods. *Statistics and Decision Making*, (11), 119-121. [虞晓芬, 傅玳. (2004). 多指标综合评价方法综述. *统计与决策*, 11, 119-121.]

Yuan, H. H., & Gao, X. L. (2014). Assessing the economic vulnerability to disasters of cities: A case study of Haidian District in Beijing. *Journal of Natural Resources*, (7)*,* 1159-1172. [袁海红, 高晓路. (2014), 城市经济脆弱性评价研究——以北京海淀区为例 [J]. *自然资源学报*, 7, 1159-1172.]

Zhang, Y., Zhu, Q., & Li, M. (2019). Construction of evaluation index system for domestic learning APPs from the perspective of user experience: Based on D-S Evidence Theory. *Journal of Intelligence,* (2), 187-194. [张熠, 朱琪, 李孟. (2019). 用户体验视角下国内学习学习 APP 评价指标体系构建——基于 D-S 证据理论. *情报杂志,* 2, 187-194.]

Zhao, X. M., Shu, J., & Zhang, Z. X. (2017). Research on the evaluation of learning APPs based on user experience. *Heilongjiang Science and Technology Information*, (2), 186-189. [赵学铭, 舒珺, 张振兴. (2017). 基于用户体验的学习学习 APP 评价研究. *黑龙江科技信息*, 2, 186-189.]

## 附录 中文 APP 情况简介

| APP | 发布日期 | 应用特点 |
|---|---|---|
| ChineseSkill | 2014-02-08 | 内置中文语音评估、汉字手写、动画技术。 |
| HelloChinese | 2015-06-18 | 致力于为初级中文学习者提供优质语言服务。 |
| PlecoChineseDictionary | 2009-12-17 | 集成词典/文档阅读器/单词卡系统，支持全屏手写输入及实时 OCR 功能。 |
| LearnChineseEasily | 2018-02-14 | 以“积木式”方法组织汉字学习的形式。 |

| | | |
|---|---|---|
| Scripts:LearnChinesewriting | 2018-10-10 | 配置极简的语言插图及快节奏的语言游戏。 |
| Chineasy:LearnChineseeasily | 2018-02-12 | 内置中文词汇学习游戏，且具备28 个贴合实际生活场景的汉语学习主题及 1834 个词汇。 |
| DuChinese–ReadMandarin | 2015-12-05 | Du Chinese 是一款分级阅读应用程序，为各级汉语学习者提供广泛的阅读练习。 |
| ChineseParents | 2022-04-22 | 以真实生活为背景，沉浸式体验中文学习 |
| DailyChineseWords&Idioms | 2019-06-20 | 遵循艾宾浩斯记忆曲线规律，间隔复习并逐步引入新单词，确保已学单词的有效记忆。 |
| MandarinChinesebyNemo | 2011-04-12 | 个性化追踪进度，重点学习高频词汇，逐步构建长期记忆，轻松应对日常对话。 |
| LearnChinese-Mandarin | 2011-05-19 | 应用内含超过 200 条常用词汇及短语，并由母语者录音，支持离线使用。 |
| HSKStudyandExam-SuperTest | 2018-02-14 | 应用结合 AI 技术提供精准水平测试与定制化课程，拥有丰富题库及模拟考试资源。 |
| DominoChinese | 2022-02-16 | 通过视频教程和真实情境，清晰解释并演示日常普通话使用形式 |
| HanYou-ChineseDictionary | 2014-09-19 | 具备强大的离线 OCR 功能，能识别万余个汉字，辅助阅读各种文本。 |
| DotLanguages-LearnChinese | 2021-04-12 | 通过丰富多样的 HSK 级别文章提升普通话水平。每日新增六篇以上文章确保学习材料充足。 |
| LearnChineseHSK1Chinesimple | 2020-03-11 | 告别枯燥课本与昂贵课程，透过宾果系统分析进步状况，集中练习要点，快速高效备考 HSK。 |
| Learn Chinese for Beginners | 2022-01-11 | 无需注册账号，所有内容完全免费且可离线使用。课程覆盖拼音、汉字及日常生活中的各种实用词汇与表达。 |

# Rethinking Technology Integration in Chinese Language Teaching: Insights from the Four-Level Feedback Theory (重新审视科技在中文教学中的应用：四级反馈理论的启示)

Huang, Shuwen
(黄淑雯)
District of Columbia Public Schools
(哥伦比亚特区公立学校)
shuwennn1123@gmail.com

Tian, Ye
(田野)
University of Pennsylvania
(宾夕法尼亚大学)
tianye1@sas.upenn.edu

**Abstract:** This study explores how technological tools support feedback mechanisms in Technology-Assisted Chinese Language Teaching (TACLT) by using Hattie and Timperley's (2007) Four-Level Feedback Theory (FLFT) as the evaluative framework. After reviewing 15 studies published in the *Journal of Technology and Chinese Language Teaching* (JTCLT) from 2022 to 2024, the research assesses the extent to which task-level, process-level, self-regulation, and self-level feedback are integrated into instructional designs. The findings reveal that while task-level feedback is widely implemented through correction-based technologies, process-level and self-regulation feedback are only moderately integrated, and self-level feedback remains largely underdeveloped. The paper argues that technology alone cannot fully address all feedback dimensions and advocates for teachers to actively design instructional activities that complement technological tools, especially in fostering metacognitive engagement and learner motivation.

**摘要：**本研究运用 Hattie 和 Timperley（2007）提出的"四级反馈理论作为评估框架，探讨了技术工具在科技辅助中文教学中对反馈机制的支持作用。通过分析期刊《科技与中文教学》在 2022 年至 2024 年间发表的 15 篇相关研究，本文评估了各个教学设计中科技运用在任务层面、过程层面、自我调控层面以及自我层面的反馈的整合程度。研究结果显示：虽然纠错类技术广泛用于任务层面的反馈，但过程层面与自我调控层面的反馈仅被中等程度地融入教学中，而自我层面的反馈仍明显不足。文章指出，仅靠科技技术本身无法有效全面覆盖所有反馈层面，因此教师应在教学中积极设计有效的教学活动，以弥补科技反馈的不足，尤其是在促进学生元认知投入与学习动机方面发挥关键作用。

**Keywords:** Technology-Assisted Chinese Language Teaching, Four-Level Feedback Theory, Instructional design, Pedagogical Effectiveness

**关键词：**科技汉语教学，四层级反馈理论，教学设计，教学效果

 

## 1. Introduction

Technology-Assisted Chinese Language Teaching (TACLT) has developed unprecedentedly after the global pandemic in 2022 and the sheer momentum of generative AI in 2023. More than ever, language educators regard technology as an instructional resource to supplement traditional classroom teaching across diverse age groups (Ma et al., 2023; Sun et al., 2023; Tan et al., 2022; Wang, 2024; Wu, 2022; Zhang, 2022). Such a direction varied from digital games creating new affordances for instructional design (Frederick et al., 2022) to the integration of text-to-speech technologies (Wang et al., 2022) and speech-to-text technologies (Feng & Tian, 2025) into Chinese language classrooms. Other research concerns ways to improve participants' experiences of synchronous online teaching (Bao & Chen, 2022; Gong et al., 2023) and asynchronous learning through information and communication technology (Luo, 2023) and social learning platforms (Ji & Lin, 2024). After the introduction of ChatGPT in November 2022, possibilities regarding integrating AI into language education opened up. Recently, several studies showcased its capabilities in promoting various aspects of language learning: writing development (Pool & Coss, 2024), oral proficiency (Li et al., 2024), vocabulary acquisition (Zhao et al., 2024), writing feedback (Yang & Tian, 2024), and many other aspects of language learning.

While there is much enthusiasm for technology integration, a significant gap exists between perceived potential and actual practice. For example, while many language educators, researchers, and instructors claim successful technology integration in their teaching practice, a few have experienced unexpected challenges that impede using such tools effectively to improve student learning outcomes (Tian, 2020). In this case, technology integration seems to focus on the sophistication of the technology itself rather than on pedagogical effectiveness and the actual learning outcomes, raising a serious question regarding its actual role in developing language proficiency. Such a gap underlines the pressing need for a systematic framework for reviewing and guiding technology integration into language teaching.

Previous research on technology integration in language education has mainly focused on elaborating the implementation strategies and measuring learning outcomes. Much less concern was given to developing the theoretical frameworks that would guide the educator's choices about technology integration. This study attempts to fill this void by arguing for applying Hattie and Timperley's (2007) four-level feedback model as a theoretical framework through which the implementation of educational technologies may be evaluated within language classrooms. Educators are also provided with a theoretical lens through which they can analyze the nature and quality of feedback different technological tools afford and develop more rigorous criteria for judgment and justification in implementing technologies in their teaching practices.

 

## 2. Literature Review

### 2.1 Feedback Theory in Language Learning

Feedback plays a significant role in language learning, bridging instructional input and learner output. As Brandl (2008) highlights, the primary role of feedback is to be informative, enabling learners to recognize discrepancies in their current target language (TL) use and guiding them toward repairing errors. It supports learners in testing and refining their understanding and hypothesis of TL rules, which is important in facilitating language acquisition. After reviewing second language acquisition theories in educational contexts, the current study found that Hattie and Timperley's (2007) four-level feedback model (FLFM) may provide a solid theoretical framework for evaluating the effectiveness of technology integration in language learning. The framework outlines four levels of feedback: task, process, self-regulation, and self. Each level contributes uniquely to the language learner's development and achievement of learning goals (See Table 1 on next page).

More specifically, Hattie and Timperley (2007) stated that task-level feedback (FT) relates directly to the performance of a task, such as how one distinguishes between right and wrong answers. This type of feedback is most common in language learning and might involve immediate corrections of language use, such as pronunciation and grammatical errors. This level is fundamental as it provides the basic information for language acquisition. Task-level feedback becomes most effective when it helps students identify and reject incorrect assumptions that have formed during their language learning, accompanied by particular information needed to acquire the correct forms. An emphasis on such task-level feedback runs the risk of creating a learner who becomes overly fixated on the immediacy of correctness and underdevelopment in broader strategic areas that support independent learning, hence creating a gap in the development of transferable skills associated with long-term language proficiency.

Hattie and Timperley (2007) also suggested that the process-level feedback (FP) addresses the main processes needed to understand or perform tasks. This includes feedback about strategies for language acquisition, techniques for oral communication, or approaches to reading comprehension. Process-level feedback is particularly important in language learning as it helps students develop effective learning strategies and understand the underlying mechanisms of language acquisition. Process-level feedback emphasizes deeper learning by focusing on the relationships between ideas, cognitive processes, and the transfer of knowledge to untried tasks (Marton et al, 1993). According to Earley et al. (1990), feedback at this level tends to be more powerful than task-level feedback in terms of promoting deeper learning and mastery of knowledge.

**Table 1 The Focus, Key Features, Benefits, and Drawbacks of FLFM**

| Feedback Level | Focus | Key Features | Potential Benefits | Potential Drawbacks |
|---|---|---|---|---|
| **Task-Level Feedback (FT)** | Correctness of task performance | Immediate correction of language use, e.g., pronunciation or grammar. Provides basic information for language acquisition. | Helps identify and reject false assumptions. Provides specific guidance for accurate information acquisition. | Overemphasis may lead to prioritizing correctness over broader learning strategies. |
| **Process-Level Feedback (FP)** | Main processes for understanding or performing tasks | Feedback on strategies, techniques, or approaches, e.g., strategies for oral communication or reading comprehension. | Promotes deeper learning, understanding mechanisms of language acquisition, and knowledge transfer. | None explicitly mentioned but requires learners to apply feedback effectively to maximize benefits. |
| **Self-Regulation-Level Feedback (FR)** | Self-monitoring and self-evaluation capabilities | Develops learner autonomy and metacognitive strategies. Creates internal feedback loops. | Enhances self-efficacy and learning outcomes by encouraging task focus and effort investment. | The impact depends on student confidence and willingness to process feedback. |
| **Self-Level Feedback (FS)** | Personal feedback directed at the "self." | Focuses on praise rather than task-related information. Rarely translates into improved learning unless tied to effort or strategies. | Improves motivation when linked to effort or strategies. | Minimal impact on performance unless explicitly connected to actionable insights. |

At the self-regulation level (FR), Hattie and Timperley (2007) explained that feedback directs attention to students' self-monitoring and self-evaluation capabilities. This level helps develop learner autonomy in language acquisition, improving students' metacognitive strategies for assessing their learning progress. Self-regulatory feedback involves creating internal feedback loops where learners evaluate their performance and the processes they use (Winne & Butler, 1995). However, the impact of self-regulation-level feedback depends on students' confidence in their responses and their willingness to invest effort in processing feedback (Kulhavy & Stock, 1989). Research indicates that feedback to enhance self-efficacy and self-regulation can significantly improve learning outcomes by encouraging students to redirect attention to tasks and invest greater effort (Kluger & DeNisi, 1996).

 

The fourth level, self-level feedback (FS), involves personal feedback directed at "self." This kind of feedback is generally regarded as the least effective because it contains little task-related information, which seldom translates into improvements in language learning outcomes. According to Hattie & Timperley (2007), while students tend to like praise, its impact on performance is minimal unless it specifically relates to the effort, processes, or strategies used to accomplish a task. In this regard, praise should be directed at effort and strategy to be effective, providing students with insight that can be applied to future tasks (Burnett, 2002). In this respect, self-level feedback is commonplace in traditional classroom settings, although it has uniquely problematic features within technology-enhanced learning environments, which require specific, actionable feedback.

### 2.2 Four-Level Feedback Theory: Current Applications and Research Gap

Feedback is an area that researchers have thoroughly investigated in education, and many reports have been made on how various strategies may affect students' achievement. However, the application of Hattie and Timperley's (2007) Four-Level Feedback Theory has been relatively limited in the literature, especially when it comes to fully incorporating all four levels: task, process, self-regulation, and self. This section provides an overview of the general landscape of feedback research, then narrows down to studies specifically engaging with the four-level framework, before finally identifying critical gaps in the current applications of this theoretical model.

Recently published literature views feedback in educational contexts through an array of lenses. Many studies have examined feedback strategies in education without directly referencing the four-level feedback model. Within higher education contexts, Sato et al. (2018) researched the role of instructor feedback in large-enrollment biology classes. In professional development settings, Johnson, Sondergeld, and Walton (2019) focused on the implementation of formative assessment across three urban districts. For vocational education, Peters et al. (2018) studied the role of formative assessment scripts in scaffolding peer feedback. These studies, along with others like Panadero et al. (2019) and Ritzhaupt et al. (2018), assume feedback as a general means toward improving learning but lack elaboration at a more detailed level.

A smaller subset of studies explicitly mentions the four-level feedback framework, but these studies often treat it as a reference point rather than fully utilizing it as an analytical tool. For example, in the study by Baadte (2019), the influence of short-term video-based interventions on the development of teacher feedback skills in support of students' self-regulated learning was investigated, taking a four-level framework into consideration but not fully applying it.

In contrast to these broader approaches, only a few studies have fully employed the Four-Level Feedback Theory as a primary analytical framework. Among them, Muthukrishnan et al. (2024) and Roby (2022) conducted research specifically using the four-level feedback framework within the context of English as a Second Language (ESL) instruction. These studies applied the four-level feedback concepts systematically to investigate the impact of feedback on language learning outcomes. For example, Muthukrishnan et al. (2024) examined the relationship between feedback types and growth mindset among secondary school ESL learners, emphasizing the role of process and self-regulation feedback in fostering student motivation and performance.

Of particular relevance to Chinese language education, Ding and Chew (2019) investigated online feedback practices in Chinese language learning, exploring how online feedback benefited learners through metaphorical perceptions. While their study shares similar interests in technology-enhanced feedback and Chinese language instruction, they only touched upon elements of the four-level model without fully applying it as an analytical framework. Despite their valuable insights into online feedback in Chinese language learning, their research further highlights the need for a more systematic theoretical approach to understanding feedback in technology-enhanced language instruction.

While feedback theory has been widely explored in education, a preliminary search in Google Scholar suggests a notable gap in the literature: a query for “four-level feedback” returned only a handful of relevant results, with just five directly engaging with Hattie and Timperley’s framework. The search was conducted using the keywords “four-level feedback,” “Hattie and Timperley,” and “language learning,” which together yielded fewer than twenty results published between 2007 and 2024. Although not exhaustive, this finding aligns with recent meta-analyses on feedback in language education (e.g., Panadero et al., 2019), which similarly note that the four-level model remains underrepresented in applied language studies. This gap is particularly evident in Chinese language education and technology-enhanced instruction, where no studies have comprehensively applied this theoretical model. Addressing this void, the present study systematically employs the four-level feedback framework to examine how feedback at the task, process, self-regulation, and self levels impacts learning outcomes in technology-mediated Chinese language instruction. By bridging this gap, the research offers both theoretical insights into the model’s applicability in language education and practical strategies for optimizing feedback in technology-enhanced Chinese teaching.

## 3. Research question

This study examines how feedback mechanisms are expressed within Technology-Assisted Chinese Language Teaching (TACLT) through the lens of Hattie and Timperley’s (2007) Four-Level Feedback Theory (FLFT). Specifically, it investigates how technological tools facilitate different levels of feedback in instructional design and to what extent current TACLT practices align with the FLFT framework. To provide a more focused analysis, the study addresses the following sub-questions:

1. How do recent TACLT studies incorporate the four feedback levels—task, process, self-regulation, and self—proposed by Hattie and Timperley (2007)?
2. To what extent do these studies demonstrate alignment or divergence between their instructional designs and the principles of FLFT?

3. What common trends and challenges emerge in the implementation of feedback mechanisms across different technological tools and learning contexts?

Together, these questions aim to clarify how effectively current technology-enhanced instructional designs in Chinese language education operationalize the multiple dimensions of feedback envisioned in the FLFT framework.

## 4. Methodology

### 4.1 Data Sources

To achieve this goal, this study primarily selects research published in the *Journal of Technology and Chinese Language Teaching* (JTCLT), a leading source of studies on the intersection of technology and Chinese language instruction in the U.S. [1] JTCLT provides a comprehensive perspective on the latest advancements in digital learning environments, AI-assisted language acquisition, and online language pedagogy, making it a highly relevant source for this investigation.

This investigation is based on JTCLT research from 2022 to 2024, when this research began. A total of 28 studies published during that period have been reviewed, supplemented by an additional study, Tian (2020), which was included due to its relevance as a typical counterexample in evaluating the Four-Level Feedback Theory. The selections cover a wide range of focuses, from AI applications and digital learning environments to online teaching, both synchronous and asynchronous, and even tool development for automated assessment, all towards the enhancement of the teaching of the Chinese language.

The current study reviews research on learners across multiple proficiency levels, from beginners to advanced, and in diverse instructional contexts, including distance learning, hybrid formats, and classroom-based technology integration. To manage this breadth, a systematic filtering process was applied to ensure alignment with the research objectives.

### 4.2 Data Collection Criteria

This study refines the dataset by considering the relevance of the articles on TACLT and the presence of instructional design elements in which technology is well integrated. The selected studies have been analyzed using Hattie and Timperley's (2007) Four-Level Feedback Theory to explore the feedback mechanisms concerning the specific mechanisms for feedback within the studies. The filtering process involved the following steps: The initial dataset consisted of twenty-eight JTCLT articles published between June 2022 and December 2024, plus Tian 2020, hence a total of twenty-nine. Studies would be included

[1] See *Journal of Technology and Chinese Language Teaching* at http://www.tclt.us/journal for details about its scope and recent issues.

in the current analysis based on whether they focused on TACLT, clearly showed instructional design elements in technology, discussed how technologies facilitate teaching, and were empirical about the discussion rather than the tool's theoretical argument or evaluation.

Studies were excluded if they primarily focused on technological tool evaluation or theoretical discourse without direct instructional design applications. This criterion excluded Ma et al. (2023), Poole & Coss (2024), Wang (2024), Li (2024), Qian (2022), and Juan (2023), as these writings either discussed the evaluation of AI models, digital tools, or applications of computational linguistics, or engaged in theoretical discussions without applying instructional design in Technology-Assisted Chinese Language Teaching. Additionally, studies emphasizing teaching methods within technological environments rather than technology-enhanced instruction were removed. Such studies include Bao & Chen (2022), Sun et al. (2023), Jiang & Xie (2022), Hu et al. (2023), and Lyu et al. (2023). These studies cover several pedagogical approaches, such as TPR and project-based learning, and their application to the online or digital context, but did not focus on how technology itself facilitated instructional feedback. Book reviews that do not examine instructional practice were also excluded (e.g., Kalyanov, 2024; Song, 2024).

After applying these criteria, 15 articles remained eligible for detailed analysis. Tian's (2020) research, which was also published on JTCLT, was added to the reference list here despite it being beyond the time scope because it serves as an illustrative case of an instructional design completely unaligned with FLFT. A further discussion of Tian's work will be undertaken in Section 5.4, as it represents a counterexample.

**4.3 Data Analysis**

The analysis was guided by Hattie and Timperley's (2007) Four-Level Feedback Theory (FLFT), which informed the criteria for evaluating how feedback was represented across the selected studies. Each paper was reviewed for the existence and effectiveness of feedback in the four dimensions and assigned a score from 1 (absent or minimally addressed) to 3 (explicitly described and well integrated), with 2 indicating partial or emerging integration. Rather than applying a rigid rubric, the grouping focused on the relative depth and clarity with which each study incorporated feedback within instructional design. For instance, higher scores reflected studies that explicitly demonstrated how technology supported feedback loops or learner reflection, whereas lower scores represented designs where such mechanisms were only briefly mentioned or implied. In this context, "well-integrated" refers to feedback that was systematically embedded in instructional activities and clearly connected to learning objectives rather than added as a peripheral feature. Illustrative examples of high- and low-integration cases are provided in Section 5 to demonstrate how these distinctions appeared across studies. This approach aimed to capture overall patterns and trends in feedback integration rather than to make fine-grained evaluative judgments about individual studies.

The scoring was conducted by the first author as a single-reviewer analysis, following consistent criteria across all studies to ensure interpretive coherence. Because this was a single-reviewer analysis, no formal inter-rater reliability test was conducted; however, the scoring process emphasized consistency and transparency in applying the framework to all cases. The aggregated scores were then analyzed to identify recurring feedback patterns, as presented in Section 5.

## 5. Findings and Discussion

### 5.1 Statistical Analysis and Key Findings

Table 2 summarizes the technology tools used, instructional design goals, and feedback scores across all four levels for the 15 selected studies, ranked in descending order by their total scores. The statistical results indicate varying degrees of feedback implementation across the four levels. Task-level feedback (FT) attained the highest mean score of 2.73 (SD = 0.59), suggesting that most of the studies incorporated technology-facilitated corrective feedback for language tasks. Process-level and self-regulation feedback both had a mean score of 1.87 (SD = 0.74), indicating moderate integration of feedback on learning strategies and self-monitoring. Self-level feedback had the lowest mean score, 1.13 (SD = 0.52), confirming that very few studies provided personalized, motivational feedback. Given the relatively small sample size (n = 15), this study reports descriptive statistics (means and standard deviations) rather than inferential analyses. The goal is not to establish statistical generalizability but to identify observable patterns and relative tendencies in feedback integration across the selected studies.

Total scores showed that the average study received 7.60 points out of 12, with a median score of 8.00 and a standard deviation of 1.64. The best-scoring study was Ji & Lin (2024), with a total score of 10, indicating good compliance with all four feedback levels. The lowest total score was 4, as in the case of Tian (2020), which represented an instructional design with minimal feedback incorporation.

Considering the overall picture, none of the teaching designs seem to fully incorporate all four feedback levels to the extent envisioned in the framework. Indeed, much better integration was found at the task and process levels, while gaps persist at both the self-regulation and self-levels, indicating that learners are often not provided with structured opportunities to monitor their own progress autonomously or receive motivational feedback that optimally engages them. The self-level feedback is also poorly addressed in the studies, and this poses critical implications for how technology can better facilitate the learning and engagement of students in the Chinese language. The following sections will further analyze a highly aligned study (Ji & Lin, 2024), a study with mid-level match (Chang & Tseng, 2023) that represents a typical image in current research, and a low-aligned study (Tian, 2020) to illustrate these findings in greater detail.

**Table 2 Summary of Technology Use, Design Goals, and Feedback Scores in 15 Studies**

| | Study | Tools | Design Purpose | FT | FP | FR | FS | Total Score |
|---|---|---|---|---|---|---|---|---|
| 1 | Ji & Lin, 2024 | *Yellowdig* | Examine the implementation of asynchronous online discussion (AOD) using the *Yellowdig* platform in a Chinese heritage language course, highlighting its role in community building, resource sharing, and enhancing student engagement in online language learning. | 3 | 2 | 2 | 3 | 10 |
| 2 | Shan et al, 2024 | *CFLingo* （Open AI API） | Explore how can task-based language teaching principles be effectively integrated with generative AI to create an adaptive language learning platform that enhances Chinese language acquisition through progressive task complexity and personalized feedback. | 3 | 3 | 3 | 1 | 10 |
| 3 | Qiu & Zhang, 2023 | “北语中文智慧系统” (*BLCU* AI System for International Chinese Education) | Examine the effectiveness of an AI-supported reading-aloud practice system in enhancing advanced CSL learners’ oral proficiency | 3 | 3 | 2 | 1 | 9 |
| 4 | Ni & Rovira, 2024 | digital dictionary | An analysis of digital Chinese dictionaries’ typologies, features, and applications in teaching Chinese as a foreign language. | 3 | 2 | 3 | 1 | 9 |
| 5 | Chang & Tseng, 2023 | Data-Driven Learning (DDL) (*Sketch Engine*, *Concordance*, *Word Sketch Difference*, *Thesaurus*) | Examine the effectiveness of integrating Data-Driven Learning approach into teaching Chinese confusable words through a combination of indirect and direct corpus consultation methods. | 3 | 2 | 2 | 1 | 8 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 6 | Gong et al, 2023 | *Zoom* | A case study exploring how Chinese as a Foreign Language (CFL) teachers utilize multilingual scaffolding, real-time interaction, and technology-enhanced feedback to promote behavioral, emotional, and cognitive engagement in online classrooms. | 3 | 3 | 1 | 1 | 8 |
| 7 | Luo, 2023 | *Skype*, *Wechat* | How can virtual exchange platforms (*Skype* and *WeChat*) be effectively integrated into Chinese language teaching to promote both linguistic and cultural learning outcomes while addressing practical challenges in implementation? | 2 | 2 | 3 | 1 | 8 |
| 8 | Frederick et al., 2022 | Digital RPG Game (*Legend of dragon*) | Explore how integrating a digital RPG game into Chinese dual language immersion classrooms affects both students' vocabulary/reading comprehension and creates pedagogical affordances for meaningful language interaction. | 3 | 2 | 2 | 1 | 8 |
| 9 | Tan et al., 2022 | Open Educational Resource (*STARTALK eTower*) | Introduce *STARTALK eTower* as useful cultural resources and digital tools to enhance Chinese language proficiency, learner autonomy, and cultural competence in K-16 education. | 3 | 2 | 1 | 1 | 7 |
| 10 | Zhang, 2022 | Online Accessible Resources (OAR) | Examine the effectiveness of intermediate CFL learners' use of online accessible resources to improve their language skills and cultural knowledge, while fostering autonomy and critical evaluation in their learning process. | 3 | 1 | 2 | 1 | 7 |
| 11 | Wu, 2022 | *Open Learning Initiative* by CMU | Introduce an online Chinese language learning platform and discuss how to effectively incorporate it into a pedagogically effective and efficient Chinese online curriculum | 3 | 2 | 1 | 1 | 7 |

| | | | | | | | | |
|---|---|---|---|---|---|---|---|---|
| 12 | Li et al, 2024 | *ChatGPT-3.5* | Explore the acceptance of *ChatGPT*-assisted oral language practices among CFL learners, emphasizing the role of learning motivation and willingness to communicate in enhancing the adoption of AI-driven language tools. | 3 | 1 | 2 | 1 | 7 |
| 13 | Zhao et al, 2024 | Large Langauge Modals (LLMs)，including *ERNIE4.0*、*Baichuan2-13B, and GPT3.5 Turbo* | Explore how can prompt engineering be optimized to enhance LLMs' effectiveness in identifying Chinese language learners' Zone of Proximal Development for near-synonym learning. | 3 | 1 | 2 | 1 | 7 |
| 14 | Wang et al, 2022 | Text-to-speech & Speech-to-text technology | Evaluate the intelligibility of Chinese synthesized speech and Chinese as a second language learners' attitudes toward its use in language learning and instruction to assess its potential as a pedagogical tool. | 2 | 1 | 1 | 1 | 5 |
| 15 | Tian (2020) | Machine Translation, including *Sogou Translate* | Use Machine translation as a self-editing tool to improve students' writing proficiency. | 1 | 1 | 1 | 1 | 4 |

## 5.2 Analysis of a High-Level Alignment Case

Among the selected studies, Ji & Lin (2024) stands out as a highly aligned case due to its comprehensive incorporation of feedback across all four levels of Hattie and Timperley's (2007) Four-Level Feedback Theory. Their study, which examines the use of asynchronous online discussions (AOD) in an online Chinese heritage language course, demonstrates a well-balanced instructional design that effectively incorporates technology to enhance both linguistic and metacognitive learning processes. The key strength of this study lies in its ability to integrate various forms of feedback through the *Yellowdig* social learning platform, making it one of the most successful examples of Technology-Assisted Chinese Language Teaching in terms of feedback design.

Ji & Lin (2024) effectively implement task-level feedback by providing corrective feedback on students' language use through asynchronous discussion activities. The *Yellowdig* platform enables students to receive peer and instructor feedback in an interactive format, reinforcing their language accuracy in a communicative setting. Furthermore, instructors review students' posts after each discussion cycle, identifying common linguistic errors and addressing them in subsequent synchronous sessions. This structured approach ensures that task-related feedback is explicitly provided and integrated into the instructional process, aligning with the highest level of task-feedback effectiveness in the FLFT framework.

The study demonstrates strong support for process-level feedback, as the AOD platform facilitates collaborative learning strategies and encourages metacognitive engagement. Students are required to share external resources (e.g., articles, videos, and songs) related to class topics, explain their relevance, and reflect on their meaning. This reflective component prompts learners to engage in deeper processing rather than merely completing tasks for participation. Additionally, the instructor uses student-generated content to shape future synchronous discussions and supplementary reading materials, effectively bridging online discussions with structured classroom learning. By allowing students to drive the learning process and connect new knowledge with prior understanding, the study successfully incorporates process-oriented feedback mechanisms.

A defining characteristic of Ji & Lin's (2024) teaching design is its emphasis on learner autonomy and self-regulation feedback. The asynchronous nature of *Yellowdig* allows students to participate in discussions at their own pace, providing opportunities for self-monitoring and independent reflection. The grading mechanism also supports this: it tracks student participation and rewards rather than penalizes failures, thus relieving learners of responsibility for contributions. Minimal instructor intervention in the discussions further promotes self-regulated learning, given that students dispose of all means of interaction—indirect support is facilitated through post-discussion reviews. Thus, it corresponds well with the principle of self-regulation feedback and makes an excellent model within this category.

Unlike most of the studies analyzed in this review, Ji and Lin (2024) effectively integrate motivational and affective support into their instructional design. A "like" function on the *Yellowdig* platform enables students to appreciate others' contributions. This mechanism of social validation helps build community and engenders students' motivation to recall the associated benefits of participation. The gamified grading system provides positive reinforcement by granting points for participation and interaction, and not punishing errors. This feature mimics informal learning behavior on social media. During this process, feedback is natural and thus facilitating rather than evaluative. This contrasts with Tan et al.'s (2022) teaching design with *eTower*, which basically employs a unidirectional information delivery model and lacks interactively engaging features to support students' mutual engagement. By integrating peer feedback and social validation, *Yellowdig* effectively compensates for the limitations of *eTower*, providing a more interactive and emotionally supportive learning environment. As a result, Ji & Lin's study is one of the few that meaningfully addresses the affective dimension of feedback, demonstrating a well-rounded implementation of the FLFT model.

The comprehensive integration of feedback in Ji & Lin (2024) highlights the potential of asynchronous learning environments in TACLT. Unlike many studies that primarily emphasize task-based correction, this study balances all four feedback levels, ensuring that students not only receive linguistic corrections but also develop higher-order learning strategies, self-regulatory skills, and intrinsic motivation. The interactive and student-centered design of the AOD component sets a strong example of how technology can be leveraged to optimize feedback mechanisms in online Chinese language instruction.

In contrast to lower-scoring studies, which often fail to integrate feedback beyond the task level, Ji & Lin (2024) successfully demonstrate how technology can create a dynamic and supportive learning environment. The following section (5.3) will analyze a case with mid-level match (Chang & Tseng, 2023) to illustrate both the potential and limitations of organized, technology-enabled feedback, before turning to a study demonstrating a low-level match (Tian, 2020) in Section 5.4.

### 5.3 Analysis of a Mid-Level Alignment Case

Four of the selected studies demonstrated a total feedback score of 8, placing them in the mid-range category within Hattie and Timperley's (2007) Four-Level Feedback Theory (FLFT). These are Chang & Tseng (2023), Luo (2023), Gong, Pang & Li (2023), and Frederick et al, (2022). While the total scores are the same, distribution across the four levels varies, making the selection of a representative mid-level case a deliberate process.

To identify the most suitable mid-level case, the current research considered studies that demonstrated a structured yet incomplete implementation of FLFT, where task-level (FT) feedback was strong, process-level (FP) and self-regulation (FR) feedback was present but not fully developed, and self-level (FS) feedback were weaker. This distribution reflects the most typical pattern among all analyzed studies, where task-level feedback tends to be the most systematically implemented, followed by process and self-regulation feedback, while self-level feedback remains the least developed. Among the three studies with a total score of 8, Chang and Tseng (2023) best exemplify this pattern (FT:3, FP:2, FR:2, FS: 1), making it the most representative mid-level case for analysis.

Chang and Tseng (2023) designed a five-week experimental course to investigate the role of Data Driven Learning (DDL) in helping learners distinguish between commonly confused Chinese word pairs. The first five sessions employed an indirect DDL in which the instructor pre-selected and organized corpus examples into paper-based materials for students to analyze collocations and grammatical patterns. The last five sessions employed an explicit DDL approach by having students directly work with *Sketch Engine* to discover linguistic patterns using the tools provided, such as the *Concordancer* and *Word Sketch Difference*. In this case, the teaching design merged the use of technological tools with task-based activities, guiding learners to notice usage differences in authentic contexts and inducing them to infer the lexical rules behind the usages through guided corpus exploration.

This teaching design reflects a very systematic form of task-level feedback, particularly in leading students to achieve more accurate lexical choices. Corpus tools such as *Sketch Engine*, *Concordancer*, and *Word Sketch Difference* engage participants in analyzing collocations, word frequencies, and semantic differences. These elements provide effective and obvious corrective feedback as learners compare their output directly with authentic linguistic materials and know how to discriminate between confusable words. The feature of these technological tools aligns closely with task-level feedback (FT), involving immediate and accurate correction and assuring that the students get explicit input regarding their lexical errors. This systematic correction provided high ratings for task-level feedback in the present study.

Beyond these immediate corrections, the study also sought to deepen students' understanding of word relationships. This deeper engagement aligns with process-level feedback, which involves helping learners reflect on how they learn, not just what they learn. In this study, corpus-guided tasks were provided that made learners pay attention to patterns in word usage. For instance, learners were instructed to check how target words occur in different contexts, compare collocations, and make hypotheses about their meanings and grammatical functions. Much of this process, however, remained teacher-controlled: rather than engaging in free exploration, learners were set on a structured path involving word lists and research tasks. Students did some analytical thinking, but the chances for the independent development of strategies were limited. The lack of open-ended inquiry constrained deeper cognitive involvement, which positioned process-level feedback at a moderate level.

Another notable challenge in this study was the limited mention of self-regulation. Although corpus tools were available, and students were encouraged to consult linguistic data independently, the highly structured course did not allow them to develop autonomous learning habits. Unlike the more organic process of monitoring and adjusting one's lexical choices in free or less guided practice, corpus-based exercises were embedded in fixed instructional tasks, which all but skimmed the surface of individual reflection processes in learning. This implies that self-regulation feedback (FR) was available but rather limited here. While students possessed the means for self-assessment, self-tracking of language development over time was only sporadic. In this respect, self-regulation feedback in this teaching design was present but considerably less salient.

The most apparent gap in this case is self-level feedback (FS). The preoccupation with linguistic accuracy meant that correctness would always take priority over motivation and engagement. Unlike studies incorporating peer interaction, gamified elements, or explicit praise, this approach offered no mechanisms for emotional or motivational support, making self-level feedback the weakest component. In this model, students were to initiate their engagement solely through a linguistic curiosity and an interest in task completion, with no recognition of effort, no encouragement, and no validation of progress. While this spell is effective in developing lexical accuracy, this model lacks affective scaffolding, which is often a strong determining factor in maintaining long-term language engagement.

As a mid-level case, Chang and Tseng (2023) illustrate both the potential and limits of organized, technology-enabled feedback. On the one hand, its task-level feedback is well-developed, ensuring that students receive precise linguistic corrections and guided analytical training. On the other hand, its process-level, self-regulation, and self-level feedback remain underdeveloped, making it difficult for students to take ownership of their learning progress or feel intrinsically motivated. Compared with highly aligned studies such as Ji & Lin (2024), in which explicit peer collaboration and interactive engagement are designed, this was a relatively more instructor-driven design. Compared with weakly aligned studies such as Tian (2020), in which feedback loops are unsuccessful or even largely absent, this provides a structure through which feedback can be given, assuring measurable learning outcomes.

### 5.4 Analysis of a Low-Level Alignment Study

Technological tools are undoubtedly valuable for enhancing Chinese language teaching. However, applying these tools in the classroom without considering the feedback that they provide may also impede learning. For example, Tian (2020) explored a failed teaching experiment to train intermediate-level Chinese language learners to use Machine Translation as a self-editing tool to improve their writing proficiency. The goal of this approach was to help students develop self-assessment skills using *Sogou Translate* for their homework. In this method, students wrote an essay in Chinese and then used *Sogou Translate* to convert their Chinese writings into English. By examining the English translations, students were expected to identify apparent mistakes in their Chinese essays. The underlying assumption was that, since *Sogou Translate* is highly accurate for intermediate-level texts, any incorrect English translation would indicate errors in the original Chinese sentences. Students would then revise their Chinese essays until they produced an acceptable English translation. However, Tian (2020) discovered that *Sogou Translate*'s advanced error tolerance often generated correct English translations despite errors in the original Chinese sentences. Consequently, students could not rely on machine translation to identify and correct mistakes in their Chinese writing, limiting the effectiveness of this approach in fostering writing proficiency.

FLFT provides a theoretical framework for understanding the failure of this teaching design. The primary issue was an over-reliance on task-level feedback from a technological tool that failed to accurately reflect students' language errors. The design required students to identify mistakes in their Chinese essays by comparing them with *Sogou Translate*'s English output. However, due to *Sogou Translate*'s error tolerance, it often generates accurate English translations despite errors in the Chinese input, rendering the task-level feedback ineffective. Students were not reliably informed about their mistakes, undermining the intended learning outcomes.

Additionally, the design lacked emphasis on process-level feedback. It did not equip students with strategies to understand the reasons behind their errors or guide them in revising their essays effectively. The reliance on Sogou Translate bypassed cognitive engagement with the editing process, a critical element for fostering deeper learning strategies. The design also aimed to promote self-regulation by encouraging students to self-assess their work using Machine Translation. However, the tool's error tolerance provided false-positive confirmations of correctness, preventing students from effectively self-monitoring and evaluating their progress. This hindered the development of autonomy and self-regulation skills. Finally, the absence of self-level feedback, such as praise or encouragement tied to effort or strategies, exacerbated the design's shortcomings. While self-level feedback is generally less impactful, its omission left students without motivational reinforcement to counterbalance the frustrations caused by the design.

## 6. Limitations

This study has several limitations that must be carefully observed when interpreting its findings. First, the study is limited in scope as it primarily examines research published in the *Journal of Technology and Chinese Language Teaching* (JTCLT). This journal is an important source of scholarship in this field, though it cannot represent the entirety of Technology-Assisted Chinese Language Teaching (TACLT) research. Some relevant studies published elsewhere may present different findings or alternative interpretations regarding technology integration. For instance, journals such as *CALICO Journal* or *Language Learning & Technology*, which often feature studies on English or multilingual contexts, may reveal stronger emphases on learner analytics, adaptive feedback systems, or cross-linguistic transfer—areas that are less frequently highlighted in JTCLT. Future comparative reviews could examine whether similar patterns of feedback integration emerge across these broader venues.

Second, while based on Hattie and Timperley’s (2007) landmark theoretical framework, the scoring process involves subjective interpretation. Assigning numerical scores to feedback levels depends on how well researchers document these feedback mechanisms in their studies. Although a structured rubric was used, in all likelihood, different evaluators might have slightly divergent impressions about the way feedback had been implemented and have rated it, which could bring variability to the results. Future studies should strive toward establishing firmer inter-rater reliability measures and elaborated rubrics on feedback implementation assessment.

Lastly, it uses secondary data rather than direct classroom observation. As a result, the analysis is constrained by the extent to which published studies explicitly describe their instructional designs and feedback mechanisms. Some articles may not explain so well how feedback was integrated, potentially affecting the accuracy of the study’s evaluation.

## 7. Pedagogical Implications, Reflections, and a Conceptual Model

This study’s findings offer several pedagogical implications for Chinese language educators seeking to integrate technology effectively while maintaining a coherent feedback mechanism.

First, the results highlight that pedagogical effectiveness should take precedence over technological novelty. Prior research suggests that many instructional designs emphasize technological innovation more than pedagogical impact (e.g., Tian, 2020; Bao & Chen, 2022; Sun et al., 2023; Wu, 2022). Although technology provides valuable affordances for language instruction, it rarely encompasses all dimensions of effective feedback. Therefore, instructors are encouraged to design classroom activities that deliberately complement the limitations of technological tools, particularly in supporting higher-level feedback. Specifically, at the self-regulation level (FR), AI-based tools can generate metacognitive prompts that guide learners to monitor progress and reflect on learning strategies. For instance, intelligent assistants may ask students to explain their reasoning or identify recurring errors, thereby fostering greater learner autonomy. At the self-level (FS), gamified systems—such as badges, point tracking, and peer recognition—can enhance motivation and engagement, addressing the reflective and affective dimensions that are often overlooked in current TACL designs.

Second, when selecting technological tools, educators should consider both their functional capabilities and their capacity to support multiple feedback forms. Tools ought to facilitate not only immediate correction but also longer-term strategic learning. Proper tool selection strengthens instructional design by aligning technological affordances with pedagogical objectives.

Third, this study underscores the importance of motivation and engagement in technology-mediated learning. Existing TACL designs often neglect self-level feedback, which, although less directly tied to language acquisition, plays a crucial role in sustaining learner motivation. Incorporating gamified elements, peer-interaction platforms, and incentive-based recognition can increase engagement and foster a more dynamic learning environment.

Fourth, while the emphasis on feedback levels varies across courses, a balanced approach encompassing all four levels is essential. Task-level feedback is generally well implemented, yet process-level, self-regulation, and self-level feedback should not be overlooked. Instructors should move beyond merely providing correct answers to designing activities that promote metacognitive awareness, independent learning strategies, and affective engagement. Combining automated correction with guided reflection, scaffolded feedback, and interactive discussion can deepen students' learning and autonomy.

To synthesize these pedagogical insights, Table 3 (next page) presents an adapted conceptual framework linking the four feedback levels with corresponding technological functions, instructional roles, and intended learning outcomes. Ultimately, aligning emerging technologies with Hattie and Timperley's (2007) Four-Level Feedback Theory ensures that innovations in Chinese language teaching not only enhance task performance but also foster deeper metacognitive reflection, learner autonomy, and sustained motivation across all levels of feedback.

**Table 3 Adapted Four-Level Feedback Framework for Technology-Assisted Chinese Language Teaching (TACLT)**

| Feedback Level | Focus in TACLT Context | Typical Technological Support | Instructor's Role | Intended Learning Outcome |
|---|---|---|---|---|
| **Task Level (FT)** | Accuracy of linguistic performance | Automated correction, AI-assisted speech or writing evaluation tools | Select appropriate tools and ensure correction aligns with learning objectives | Improved linguistic accuracy and immediate corrective awareness |
| **Process Level (FP)** | Learning strategies and comprehension processes | Interactive platforms, corpus tools, or adaptive tutorials guiding problem-solving | Scaffold strategy use and interpret feedback results for learners | Development of effective learning strategies and transfer of knowledge |
| **Self-Regulation Level (FR)** | Learner autonomy and metacognitive reflection | AI-driven reflective prompts, progress dashboards, self-assessment checklists | Design reflection activities and guide learners in interpreting analytics | Enhanced self-monitoring, planning, and evaluation skills |
| **Self Level (FS)** | Motivation and affective engagement | Gamified feedback systems (badges, peer recognition, point tracking) | Reinforce effort, persistence, and collaboration through recognition | Sustained motivation and positive learner identity formation |

## 8. Conclusion

This paper has systematically reviewed feedback mechanisms within technology-enhanced Chinese teaching designs, revealing both potential and limitations. Although feedback at the task level is generally well effectuated, there is still considerable potential at the process level and self-regulation level, and particularly poorly integrated are those pertaining to self-level feedback. Generally, self-level feedback is often neglected, thereby limiting technology's potential to boost learner motivation and engagement.

To fill these gaps, educators will need to take a more structured approach to designing feedback mechanisms in their classrooms, paying extra attention to ensure that technological tools are used not just for automation but as mechanisms to facilitate effective and deeper learning interactions. Future research could consider how FLFT might be more systematically included in TACLT, particularly through empirical classroom studies that assess the long-term impact of different feedback strategies.

Ultimately, effective technology integration in CLT should strike a balance between leveraging digital advancements and maintaining pedagogical integrity. By applying a structured feedback framework like FLFT, educators will be able to optimize the role of technology in Chinese language teaching, ensuring that it serves as a meaningful tool for linguistic and cognitive development rather than a superficial addition to instructional design.

## References

Baadte, C. (2019). Effects of short-term video-based interventions and instructions on teachers' feedback skills to support students' self-regulated learning. *European Journal of Psychology of Education*, *34*(3), 559–578. https://www.jstor.org/stable/48698237

Bao, Y., & Chen, Y. F. (2022). Optimizing remote synchronous learning in a Chinese language class: Theory and practice. *Journal of Technology and Chinese Language Teaching*, *13*(2), 39-63.

Brandl, K. (2008). Communicative language teaching in action: Putting Principles to work. Pearson Education.

Burnett, P. C. (2002). Teacher praise and feedback and students' perceptions of the classroom environment. *Educational Psychology, 22*(1), 1-16.

Chang, L., & Tseng, Y. (2023). A corpus-driven Chinese experimental course and effectiveness analysis. *Journal of Technology & Chinese Language Teaching*, 14(1), 1-25. [张莉萍, &曾钰婷. (2023). 语料驱动学习的华语实验课程与成效分析. *科技中文与教学*, 14(1). 1-25]

Ding, S. L., & Chew, E. (2019). Thy word is a lamp unto my feet: A study via metaphoric perceptions on how online feedback benefited Chinese learners. *Educational Technology Research and Development*, *67*(4), 1025–1042. http://www.jstor.org/stable/45217348

Earley, P. C., Northcraft, G. B., Lee, C., & Lituchy, T. R. (1990). Impact of process and outcome feedback on the relation of goal setting to task performance. *Academy of Management Journal*, 33(1), 87–105.

Feng, Y., & Tian, Y. (2025). Assessing the accuracy of Chinese speech-to-text tools for Chinese as foreign language learners. *Chinese as a Second Language, 60(2), 79-108.* https://doi.org/10.1075/csl.24013.fen

Frederick, P., Jody, C. M., & Siyu, J. (2022). Exploring the affordances and effectiveness of a digital game in the Chinese dual language immersion classroom. *Journal of Technology and Chinese Language Teaching, 13*(1), 46-73.

Gong, Y., Pang, Q., & Li, W. (2023). Engaging students in the online classroom: A case study on teachers of Chinese as a foreign language. *Journal of Technology and Chinese Language Teaching*, *14*(2), 25-43.

Hattie, J., & Timperley, H. (2007). The power of feedback. *Review of Educational Research*, *77*(1), 81-112. https://doi.org/10.3102/003465430298487

Hu, B., Xiao, S., & Chen, H. (2023). Online instructional strategies for teaching Chinese idioms under the framework of digital Bloom's taxonomy. *Journal of Technology and Chinese Language Teaching*, 14(2), 44-61. [胡波，肖诗俊，&陈红. (2023). 数字布鲁姆理论下国际中文教师的成语线上教学策略研究. *科技中文与教学*, 14(2), 44-61]

Ji, J., & Lin, C. (2024). Use of asynchronous online discussion in an online Chinese heritage language course. *Journal of Technology and Chinese Language Teaching, 15*(1), 82-102.

Jiang, Z., & Xie, K. (2022). Motivating online language learners: From theory to design strategies. *Journal of Technology and Chinese Language Teaching*, *13*(1), 1-25.

Johnson, D., Kakar, R., & Walton, P. (2019). Structured public health scenario analyses promote critical thinking in undergraduate students. *Pedagogy in Health Promotion*, *5*(4), 261–267.

Juan, L. T. (2023). A corpus study of internal modifications in Chinese request and its application in CSL instructional design. *Journal of Technology and Chinese Language Teaching*, *14*(1), 56-77.

Kalyanov, A. (2024). A hybrid approach to teaching Chinese through digital humanities, CALL, and project-based learning. *Journal of Technology and Chinese Language Teaching*, *15*(2), 101-106.

Kluger, A. N., & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin*, 119(2), 254–284.

Kulhavy, R. W., & Stock, W. A. (1989). Feedback in written instruction: The place of response certitude. *Educational Psychology Review*, 1(4), 279–308.

Li, J. (2024). The application of four major social media platforms in the informal Chinese learning of Thai university students: A case study of Khon Kaen University and Suratthani Rajabhat University. *Journal of Technology and Chinese Language Teaching*, 15(2), 54–74. [李娇. (2024). 四大社交媒体在泰国大学生中文非正式学习中的应用研究-以孔敬大学和室利佛逝皇家理工大学为例. *科技中文与教学*, 15(2), 54-74.]

Li, N., Zhang, L., Lau, K. L., & Liang, Y. (2024). Predicting Chinese language learners' ChatGPT acceptance in oral language practices: The role of learning motivation and willingness to communicate. *Journal of Technology and Chinese Language Teaching, 15*(1), 25-48.

Luo, H. (2023). Teaching Chinese language and culture through Chinese-American virtual exchange: A pedagogical reflection. *Journal of Technology and Chinese Language Teaching, 14*(2), 76-92.

Lyu, B., Chang, D., & Ma, X. (2023). A study on project-based learning in international Chinese culture instruction. *Journal of Technology and Chinese Language Teaching*, 14(2), 62-75. [吕伯宁，常大群，&马璇. (2023). 基于项目学习法的中文文化教学研究. *科技中文与教学*, 14(2), 62-75.]

Ma, R., Zheng, M., & Xu, J. (2023). Visualized evaluation and intelligent recommendation of international Chinese MOOCs based on learning data mining. *Journal of Technology and Chinese Language Teaching*, *14*(2), 1–24.

Marton, F., Dall'Alba, G., & Beaty, E. (1993). Conceptions of learning. *International Journal of Educational Research*, *19*(3), 277-300.

Muthukrishnan, P., Fung Lan, L., Anandhan, H., & Swamy D, P. (2024). The role of growth mindset on the relationships between students' perceptions of English language teachers' feedback and their ESL learning performance. *Education Sciences*, *14*(10), 1073.

Ni, Y., & Rovira-Esteva, S. (2024). Digital dictionaries: types, features, and applications in teaching Chinese as a foreign language. *Journal of Technology and Chinese Language Teaching*, *15*(2), 75–100. [倪蕴玲，& 罗斐岚. (2024). 数字词典：在对外汉语教学中的类型、特征和应用. *科技中文与教学*, *15*(2), 75-100.]

Panadero, E., Broadbent, J., Boud, D., & Lodge, J. M. (2019). Using formative assessment to influence self-and co-regulated learning: The role of evaluative judgement. *European Journal of Psychology of Education*, *34*(3), 535–557.

https://www.jstor.org/stable/48698236
Peters, O., Körndle, H., & Narciss, S. (2018). Effects of a formative assessment script on how vocational students generate formative feedback to a peer's or their own performance. *European Journal of Psychology of Education*, *33*(1), 117–143. http://www.jstor.org/stable/44951943
Poole, F. J., & Coss, M. D. (2024). Can ChatGPT reliably and accurately apply a rubric to L2 writing assessments? The devil is in the prompt(s). *Journal of Technology and Chinese Language Teaching, 15*(1), 1-24. https://doi.org/10.35542/osf.io/3r2zb
Qian, Z. (2022). The design of a web-based placement test for college-level Chinese language programs. *Journal of Technology and Chinese Language Teaching*, *13*(2), 17-38.
Qiu, J., & Zhang, J. (2023). A practical study on advanced-level Chinese reading-aloud training supported by instructional platforms. *Journal of Technology and Chinese Language Teaching*, *14*(1), 78-94. [邱经纬，& 张俊萍. (2023). 教学平台支持下的高年级朗读训练实践研究. *科技中文与教学*, *14*(1), 78-94]
Ritzhaupt, A. D., Pastore, R., Wang, J., & Davis, R. O. (2018). Effects of organizational pictures and modality as a feedback strategy on learner comprehension and satisfaction. *Educational Technology Research and Development*, *66*(5), 1069–1086. http://www.jstor.org/stable/26746305
Roby, Y. (2022). Teachers' written feedback in English: How does this relate to the pathways leading towards self-regulated learning in 11-13 year olds [Master's thesis, University of Oxford].
Sato, B. K., Dinh-Dang, D., Cruz-Hinojoza, E., Denaro, K., Hill, C. F. C., & Williams, A. (2018). The impact of instructor exam feedback on student understanding in a large-enrollment biology course. *BioScience, 68*(8), 601–611. https://www.jstor.org/stable/90023901
Shan, L., Pan, Z., & Weidman, R. (2024). Integrating task-based language teaching and generative AI: Design, implementation, and evaluation of the CFLingo platform for Chinese learning. *Journal of Technology and Chinese Language Teaching*, *15*(2), 1-34.
Song, F. (2024). Using data to drive Chinese language teaching research—Review of research on Chinese classroom teaching structure and process modeling. *Journal of Technology and Chinese Language Teaching*, *15*(2), 107–110.
Sun, L., Zhou, K., & Lin, C. H. (2023). Applying the total physical response (TPR) method to online one-to-one Chinese vocabulary instruction to children. *Journal of Technology and Chinese Language Teaching*, *14*(1), 26–55.
Tan, D., Wu, L., Huang, R., & Wang, Z. (2022). STARTALK eTower: An open educational resource for improving learner competence and autonomy. *Journal of Technology and Chinese Language Teaching*, *13*(1), 74-100.
Tian, Y. (2020). Error tolerance of machine translation: Findings from failed teaching design. *Journal of Technology and Chinese Language Teaching, 11*(1), 19-35.
Wang, T. (2024). The design and application of a Chinese audio-visual corpus based on AI technology. *Journal of Technology and Chinese Language Teaching*, *15*(1), 70–81.
Wang, Y., Da, J., & Yin, C. (2022). Intelligibility of Chinese synthesized speech and learners' attitudes towards its use in CSL learning and instruction: A preliminary

study. *Journal of Technology and Chinese Language Teaching*, *13*(2), 1-16.
Winne, P. H., & Butler, D. L. (1994). Student cognition in learning from teaching. *International Encyclopedia of Education*, *2*, 5738-5775.
Wu, S. M. (2022). Chinese online teaching and learning: The CMU OLI Chinese online program. *Journal of Technology & Chinese Language Teaching*, *13*(2), 64-77.
Yang, Y., & Tian, Y. (2024). Exploring the Chinese AWCF platform's value in improving CSL learners' writing performance in a ChatGPT context. *Chinese as a Second Language, 59*(1), 46-68. https://doi.org/10.1075/csl.24008.yan
Zhang, S. (2022). Intermediate-level language learners' use of online accessible resources to supplement learning: An exploratory study. *Journal of Technology and Chinese Language Teaching, 13*(1), 26-45.
Zhao, Q., Hsu, Y. Y., & Huang, C. R. (2024). Large language model and Chinese near synonyms: Designing prompts for online CFL learners. *Journal of Technology and Chinese Language Teaching, 15*(1), 49-69.

 

# 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。

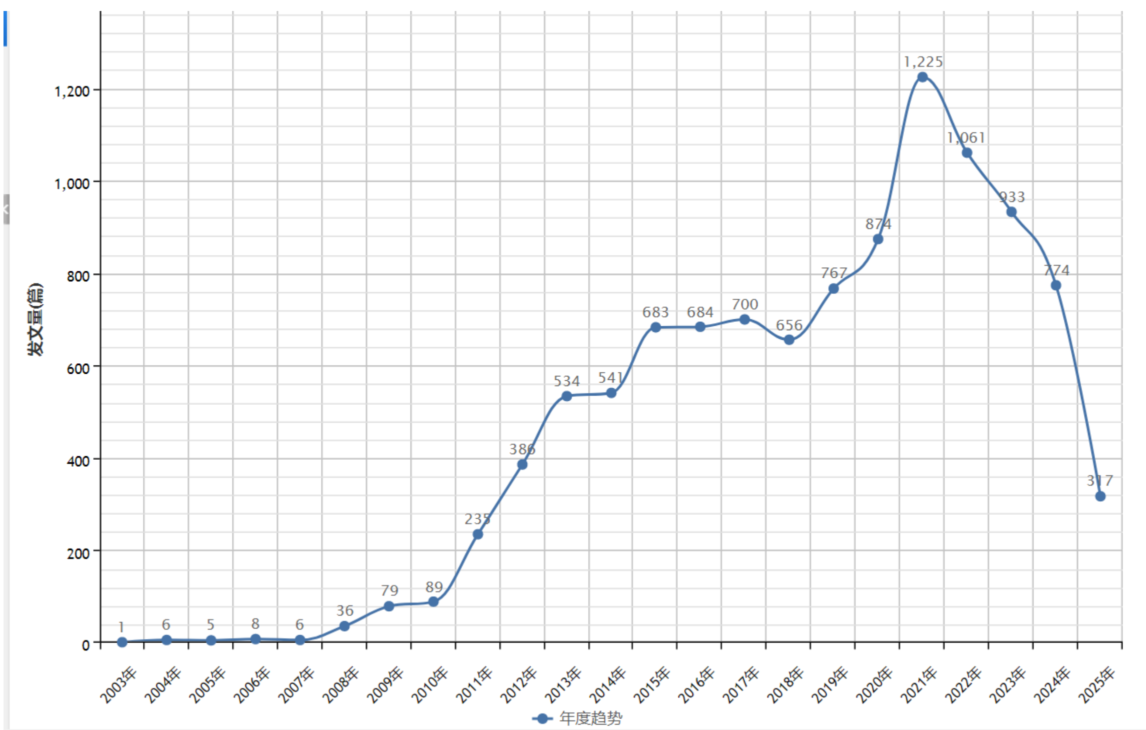


**图 1　HSK 库年发文量分布图**

[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。

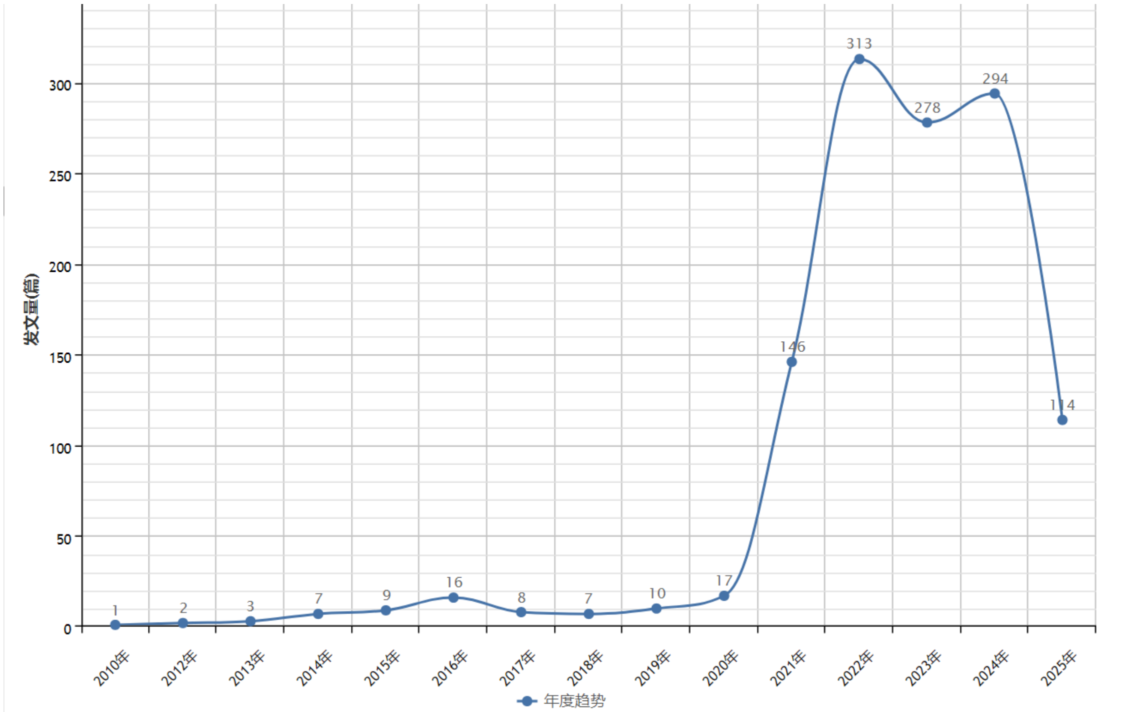


**图 2 全球库年发文量分布图**

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。

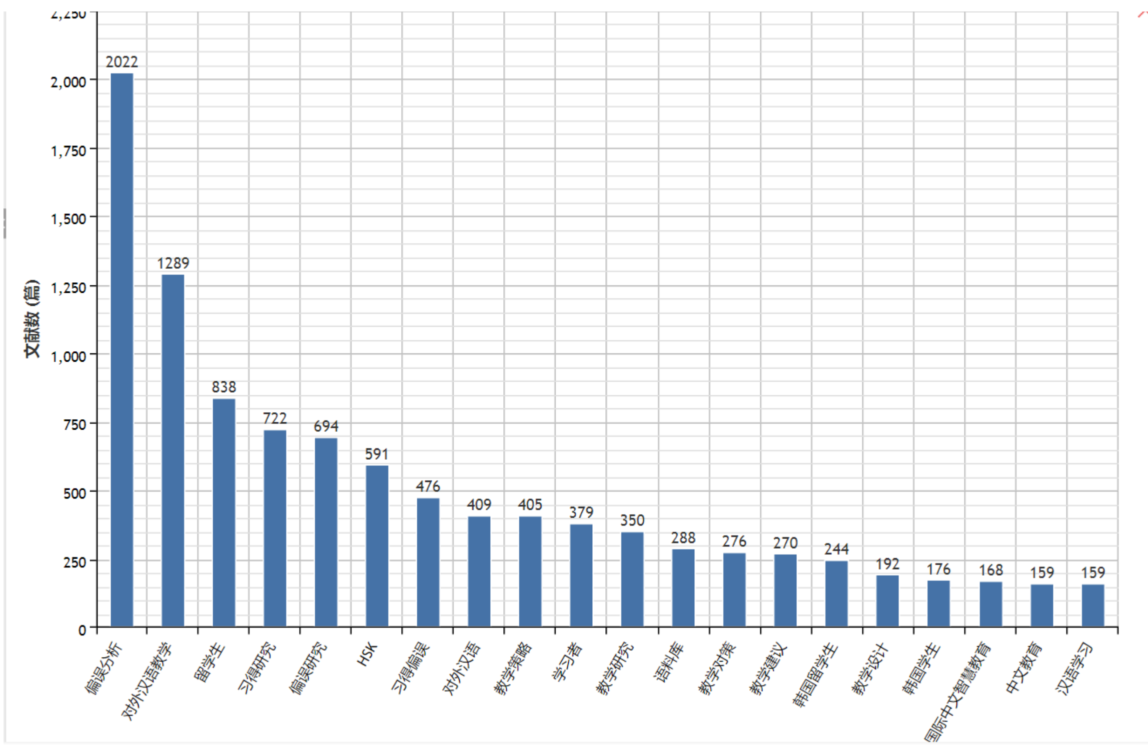


**图 3 HSK 库主要主题分布图**

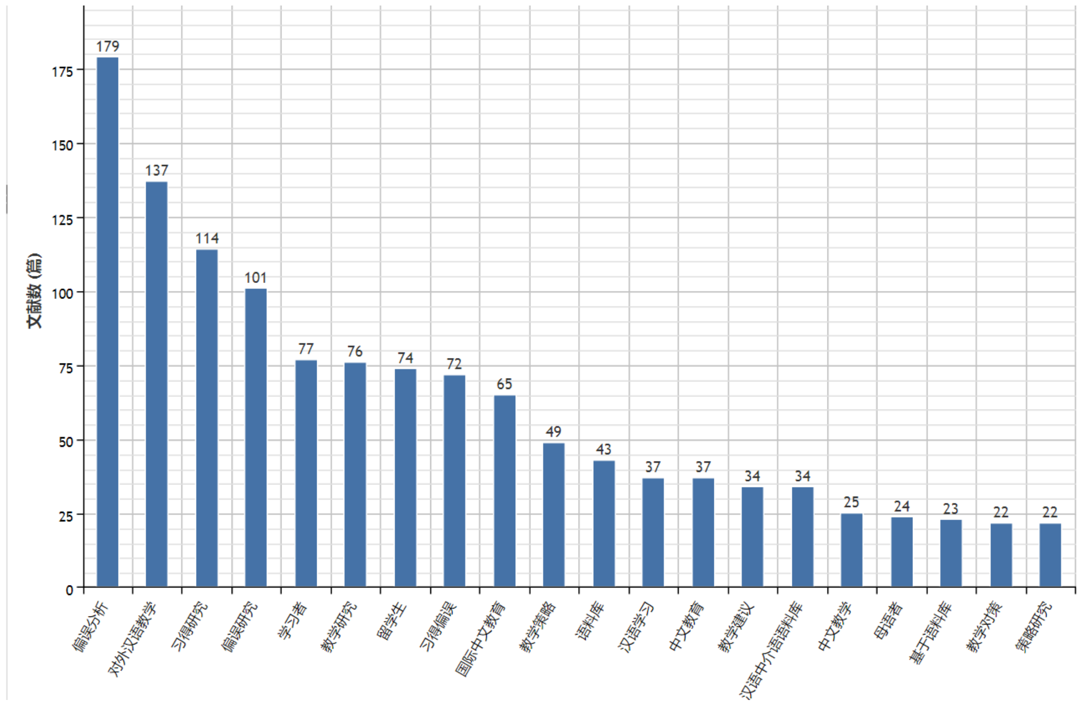


**图 4 全球库主要主题分布图**

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：

就某一话题或题目进行汉语口头表述
↓
与汉语口头表述同题的汉语写作
↓
与汉语写作同题的学习者母语写作
↓
汉语及学习者母语同题写作语料的双向翻译

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 interlanguage 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 interlanguage 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.]

# Digital Game-Based Chinese Language Learning for Adults: A Critical Review of Apps in the Apple App Store (面向成人的数字游戏化中文学习：对苹果应用商店中相关应用的批判性评估)

Gu, Sijia
(顾思佳)
YK Pao School
(包玉刚实验学校)
joicygu@outlook.com

Tian, Ye
(田野)
University of Pennsylvania
(宾夕法尼亚大学)
tianye1@sas.upenn.edu

**Abstract:** This study investigates the availability and suitability of Digital Game-Based Chinese Language Learning applications for adult learners, focusing on apps available in the Apple App Store. While English language learning apps were also examined, they serve primarily as a baseline for comparison rather than a central focus. Using a structured, multi-variable search under eight distinct conditions involving system language, search language, and App Store region, the research identified 32 unique applications. Findings reveal that most game-based language learning apps are designed for children, especially those labeled 4+, and prioritize basic vocabulary and grammar through simplified, repetitive mechanics. Very few applications cater to adult or university-level learners seeking context-rich, communicative, and cognitively engaging experiences. Among the limited options, "懒人英语 (Lazy English)" stands out for its dubbing feature, which aligns with communicative language teaching principles and illustrates the potential of game-based strategies for older learners. The paper concludes by discussing the pedagogical implications of these findings and calling for the development of more sophisticated game-based tools that address the linguistic needs and motivational profiles of adult learners of Chinese.

**摘要：**本研究调查了基于数字游戏的语言学习应用在成人中文学习领域中的可用性与适用性。虽然本文也考察了部分英文学习应用，但其作用仅限于提供基准对照，而非研究的核心。通过对苹果应用商店的多变量结构化检索，本研究在涉及系统语言、搜索语言和应用商店区域等八种不同搜索条件下，共筛选出 32 款独特应用。研究发现：绝大多数游戏应用专为儿童设计。这些应用，尤其是标注"4+"(四岁以上)的应用，通过简单重复的机制侧重基础词汇和语法训练；而能为成人或大学阶段学习者提供情境化、交际性及认知参与体验的应用屈指可数。在有限的选择中，"懒人英语"因其配音功能而脱颖而出，该功能契合交际语言教学的理念，展示了游戏化策略在成人学习者中

的潜在价值。本文最后讨论了这些发现的教学启示，并呼吁开发更多契合成人中文学习者语言需求和动机特征的高水平游戏化学习工具。

**Keywords:** Game-Based Language Learning, Chinese Language Acquisition, Adult Learners, Mobile Applications, Apple App Store

**关键词**：游戏化语言学习，中文学习，成人学习者，移动应用程序，苹果应用商店

## 1. Introduction

Digital Game-Based Learning (DGBL) has long been recognized as an effective and engaging pedagogical approach, particularly in second language acquisition (SLA), where it combines the motivational power of play with structured learning objectives (Prensky, 2007; Gee, 2004; Peterson, 2009; Shaffer, 2006). Widely adopted in English language instruction, applications such as *Duolingo, Kahoot*, and *LingQ* leverage gamification features—like level progression, reward systems, and adaptive feedback—to support vocabulary retention, grammar practice, and learner autonomy. These tools are grounded in robust theories of cognitive and social development. For instance, Piaget's (1962) stages of cognitive development and Vygotsky's (1978) concept of the Zone of Proximal Development (ZPD) emphasize the critical role of play and interaction in learning, providing theoretical support for the educational value of games.

Despite DGBL's growing use and its solid theoretical foundation, its application in Chinese as a Foreign Language (CFL) instruction—particularly for adult learners—remains significantly underexplored (Poole et al., 2022; Yang & Li, 2023). Chinese presents unique challenges for language learners, including character recognition, tonal pronunciation, and syntactic differences (Lan, 2015; Xu et al., 2022; Zhang et al., 2024), which may not be adequately addressed by game-based strategies originally designed for alphabetic languages. While some studies highlight the potential of DGBL to support CFL learners in areas like vocabulary retention and pronunciation (Yu & Tsuei, 2023), few have systematically evaluated whether existing applications are pedagogically appropriate for adult learners or tailored to the specific complexities of Chinese. Moreover, most current GBL applications available on digital platforms such as the Apple App Store were not originally designed with CFL instruction in mind. A preliminary review suggests that many focus on rote memorization or vocabulary drills without fully engaging learners in meaningful, contextualized language use—an essential component for adult learners aiming for real-world proficiency.

This study aims to bridge this gap by systematically investigating and categorizing Chinese language learning applications available on the Apple App Store. Through empirical analysis of their game mechanics, instructional design, and target user base, the research evaluates whether these tools effectively support adult CFL learners' linguistic and cognitive needs. By integrating developmental theory and practical application, this study contributes to the broader discourse on technology-enhanced language learning. It offers insights into how DGBL can be more effectively implemented in Chinese language education.

## 2. Literature review

### 2.1 Game-Based Learning (GBL)

Game-Based Learning (GBL) has gained significant attention in education due to its ability to enhance learner engagement and motivation. The theoretical foundations of GBL are rooted in constructivist theories, particularly those of Piaget (1962), who emphasized the importance of learning through active exploration, and Vygotsky (1978), who highlighted the social interaction aspect in cognitive development. Gee (2007) suggests that games provide immersive environments where learners can develop critical thinking and problem-solving skills through structured challenges and feedback mechanisms. These environments create opportunities for learners to actively engage with content actively, making learning more interactive and meaningful. Shaffer (2006) expands on this by arguing that GBL enables learners to participate in role-playing and simulated environments, fostering deeper understanding and encouraging the practical application of knowledge in realistic contexts.

Furthermore, well-designed educational games, as Egenfeldt-Nielsen (2007) noted, enhance cognitive development by aligning gameplay mechanics with learning objectives. This alignment ensures that learners face appropriate challenges and remain motivated to progress through game-based tasks that reinforce learning goals. Prensky (2007) highlights how the interactive nature of games captures learners' interest and fosters a sense of ownership in their learning journey. Additionally, Squire (2011) discusses the scaffolding potential of games, emphasizing their ability to provide incremental challenges and timely feedback, which promote active participation and sustained engagement. By integrating these elements, GBL creates a holistic learning environment that supports both cognitive development and learner motivation, making it a powerful tool in modern educational settings.

### 2.2 Game-Based Language Learning (GBLL)

The application of GBL in language education, known as Game-Based Language Learning (GBLL), has shown promising results in second language acquisition (SLA). Peterson (2009) explores how digital games create opportunities for immersive interaction in target language environments, allowing learners to practice conversational skills in authentic contexts. According to Reinders & Wattana (2015), game-based language learning promotes confidence and reduces communication anxiety by providing a low-stress setting for learners to experiment with language use. Yudintseva (2015) highlights the impact of game-based approaches on vocabulary acquisition, demonstrating that learners retain new words more effectively through contextual gameplay. Choo (2015) argues that games facilitate incidental vocabulary learning by exposing learners to repeated input within meaningful contexts. DeHaan (2005) examines the effects of video games on reading comprehension, showing that narrative-driven games can enhance learners' reading skills by engaging them with interactive storytelling. Rankin et al. (2006) provide evidence that game-based environments support reading comprehension by offering contextually rich text and visual cues. They further discuss how game-based activities can improve oral communication skills through interactive dialogues and decision-making scenarios. Despite these advantages, we argue that effective implementation of GBLL requires careful consideration of game selection to ensure alignment with pedagogical objectives and to maintain an appropriate balance between entertainment and educational value.

### 2.3 Game-based learning in Chinese language instruction

Chinese language learning presents unique challenges due to the complexity of its writing system, aural reception, and reading abilities, as highlighted by Gabbianelli and Formica (2017), who found that first-level Mandarin Chinese learners perceive the learning process as long and complex while maintaining high achievement expectations. GBL offers an effective solution by providing interactive and engaging methods tailored to these specific challenges. Lan (2015) found that repeated exposure and interactive exercises in game-based approaches significantly enhance conversation performance, reinforcing memory retention and aiding in the mastery of complex sentence structures. Similarly, Peterson (2009) highlights how digital games support listening skills development by combining contextualized audio input with visual cues, helping learners distinguish tones and improve pronunciation accuracy, which is particularly important given the tonal difficulties inherent in Chinese (Gabbianelli & Formica, 2017).

In speaking practice, Xu et al. (2022) underscore the value of role-playing scenarios and speech recognition features in games, which encourage consistent practice in a low-pressure environment, crucial for mastering tonal variations and reducing the anxiety often experienced by learners of tonal languages (Ng et al., 2022). Supporting the motivational and emotional aspects relevant to Chinese language learners, Zhang and Chen (2021) found that gamification helps visualize learning goals and creates a relaxing learning environment that reduces foreign language anxiety, a common barrier to oral proficiency and participation, especially in complex languages such as Chinese (Zhang & Chen, 2021).

Additionally, Bytheway (2014) emphasizes the importance of designing culturally adaptive games that align with learners' cognitive abilities and real-life applications, ensuring meaningful engagement. Building on this, Chen & Lin (2015) demonstrated that digital game-based situated learning, which simulates authentic historical and cultural contexts—such as those found in Tang Dynasty poetry—can deepen learners' understanding by immersing them in meaningful language use situations, addressing the cultural and contextual difficulties that often arise in Chinese language learning. Together, these findings demonstrate how game-based and gamified learning approaches can holistically address linguistic, cultural, and affective challenges inherent in Chinese language acquisition.

## 2.4 Digital game-based learning for CFL students

The rapid advancement of technology has significantly transformed the field of education, introducing innovative tools such as artificial intelligence (AI), virtual reality (VR), and mobile applications (Luckin et al., 2016; Shadiev et al., 2023). These technologies have revolutionized content delivery and learner engagement by enabling personalized and interactive learning experiences. AI-driven platforms analyze learners' progress and provide adaptive content, while VR immerses students in realistic environments that enhance cultural and linguistic understanding (Lan & Lin, 2015; Qiao & Zhao, 2023). Additionally, mobile applications provide on-the-go access to gamified learning materials, making language acquisition more accessible and flexible (Chen & Hsu, 2020).

The integration of these technological advancements has significantly enhanced the effectiveness of game-based learning in Chinese language instruction. VR, as highlighted by Zhang et al. (2024), fosters cultural immersion and provides contextualized language practice, reducing learners' anxiety and increasing motivation. AI-powered systems, according to Poole and Clarke-Midura (2020), offer personalized feedback and track progress, helping students adjust their learning strategies effectively.

# 3. Research gap

Existing studies on game-based learning and game-based language learning have primarily focused on theoretical frameworks and the general benefits of integrating games into language education. While prior research (Gee, 2007; Prensky, 2007) has established the cognitive and motivational advantages of DGBL, and studies specific to Chinese as a foreign language (Poole et al., 2022) have explored how games support areas such as character recognition, pronunciation, and cultural understanding, these works largely remain theoretical or exploratory in nature.

Moreover, much of the current literature tends to focus on games designed for children and beginners, with minimal attention to the adaptation and evaluation of game-based tools for advanced or professional language use. Preliminary observations indicate that many technology-enhanced language-learning games appear to target young learners (ages 4–12). This tendency will be examined and clarified in the findings section, but it points to a potential mismatch between the needs of adult CFL learners and the current resources available to them.

Therefore, this research aims to address this gap by identifying and analyzing technology-enhanced language learning games that are suitable for adult CFL learners. Through an examination of existing applications and their pedagogical features, this study seeks to provide practical recommendations and contribute a comparative evaluation of DGBL designed or adaptable for adult Chinese language education.

## 4. Methodology

This study employs a structured methodology for data collection and analysis to examine the availability and categorization of game-based Chinese and English language learning applications in the Apple App Store because Apple devices are widely used among U.S. university students, the target demographic for adult CFL learners (Denoyelles et al., 2023). Limiting the scope to a single platform also ensured methodological consistency, as rankings and app availability differ across systems. Future studies may extend the analysis to Android platforms for broader comparison.

### 4.1 Data collection

All data were collected in January 2024 using a single Apple iPhone 12 Pro running iOS version 18.3.1, in order to control for device- and system-related variations. A two-stage approach was used to ensure both breadth and depth in the data collection process. In the **first stage**, language learning applications were retrieved using the following predefined search terms:

- “English Learning App” / “Chinese Learning App”
- “English Learning Game” / “Chinese Learning Game”

In the **second stage**, search conditions were systematically varied to assess how search results might be influenced by key factors, including:

- **System Language**: English vs. Chinese
- **App Store Region**: United States vs. China
- **Search Language**: English vs. Chinese

Each search query was conducted under carefully controlled conditions to ensure the comparability of results. To minimize algorithmic fluctuations and regional personalization biases, all searches were performed on the same device within a short time frame. Combinations deemed improbable or irrelevant—such as a Chinese system language paired with the U.S. App Store, or an English system language paired with the China App Store—were excluded, as these configurations are unlikely to reflect the experience of typical users, particularly adult Chinese language learners based in the U.S.

For each valid search condition, the top four applications as ranked and displayed in the App Store search results were selected for analysis. In the Apple App Store, ranking refers to the order in which applications are presented to users in response to a search query. This display order is determined by Apple’s proprietary algorithm, which is not publicly disclosed but is generally understood to reflect a combination of factors such as total downloads, user ratings and reviews, update frequency, and, in some cases, paid promotions. As a result, higher-ranked applications are more visible to users conducting casual searches and are therefore more likely to be encountered and downloaded. Focusing on the top four results under each condition thus provided a representative sample of the applications that adult learners would realistically encounter when navigating the App Store without prior familiarity.

### 4.2 Data analysis

After collection, the selected applications were categorized and analyzed according to predefined criteria to evaluate their alignment with the principles of game-based language learning. The analysis focused on three key dimensions:

- **Target Age Group**: Based on Apple's user age recommendations (e.g., 4+, 12+, 17+), to identify whether the app was intended for children, adolescents, or adults.
- **Game Elements**: Documentation of the types of game mechanics used (e.g., flashcards, sentence construction, role-playing, or interactive mini-games).
- **Learning Approach**: Categorization of the pedagogical focus, such as pronunciation, grammar, vocabulary acquisition, or conversational practice.

A comparative analysis was then conducted to identify broader trends across the dataset and evaluate the suitability of these applications for adult learners. The analysis also considered cognitive and pedagogical factors relevant to university-level learners, including motivation, engagement, and communicative competence.

The findings are presented in Section 5, beginning with a summary table of the applications retrieved under different search conditions. The discussion then highlights significant gaps in the availability of game-based learning tools for adult users, evaluates the potential of existing apps, and identifies one application—*懒人英语 (Lazy English)*—as a particularly promising example for future instructional design.

## 5. Findings

### 5.1 Overview of identified applications

This study identified 32 game-based language learning applications through systematically varied searches on the Apple App Store. These applications were retrieved using targeted keyword combinations under different search conditions, such as system language (English vs. Chinese), App Store region (U.S. vs. China), and search language. These conditions were designed to reflect realistic usage scenarios for U.S.-based university learners of Chinese, as well as learners of English in the China App Store. The selected applications represent the top four results for each query, assuming they will most likely be encountered and downloaded by users unfamiliar with the app landscape. Table 1 presents a comprehensive overview of the top four search results across eight distinct search conditions, yielding 32 entries in total. Since some applications—such as *Studycat*—appear under both Chinese and English search results due to offering bilingual learning content, we consider them unique applications.

A close examination of the table reveals several key trends. First, most applications—regardless of target language—are geared toward young learners, with most marked as suitable for users aged 4+. This age classification aligns with the types of game mechanics employed, which include matching games, flashcards, phonics drills, and animated interactions—features commonly designed to appeal to children's developmental stages and learning preferences. While some apps, such as *懒人英语 (Lazy English)* and *Johnny Grammar Word Challenge,* target older users (12+ or 17+), they are exceptions rather than the norm.

In terms of functionality, the listed applications strongly focus on basic language acquisition, emphasizing vocabulary drills, pronunciation practice, and sentence construction. Game types vary from casual puzzles and swipe-based input to speech recognition and interactive lessons but remain generally limited in scope. Few apps incorporate immersive or context-rich scenarios appropriate for advanced learners or adult users seeking deeper linguistic engagement. This distribution underscores a significant gap in the marketplace: despite the prevalence of game-based learning in language education, there is a notable lack of well-designed, pedagogically sound applications tailored specifically for adult or university-level learners, particularly for those studying Chinese.

**Table 1 Top Four Language Learning Applications Identified Under Eight Distinct Apple App Store Search Conditions**

| Search condition | Ranking Search-result position[1] | APP name | User age | Game type | Target learning language |
|---|---|---|---|---|---|
| Chinese system, China Mainland App Store, Search in Chinese | 1 | 英语天天练 | 4+ | Reading & collection, text-based exercises, reward system | English |
| | 2 | Study cat | 4+ | Interactive animation, vocabulary drills, mini-games | English |
| | 3 | 懒人英语 | 12+ | Creative dubbing, pronunciation practice, speech mimicry | English |
| | 4 | 狐狸快跑 | 4+ | Adventure-based, character selection, progression levels | English |
| | 1 | Study cat | 4+ | Interactive animation, vocabulary drills, mini-games | Chinese |
| | 2 | 成语接龙 | 12+ | Word puzzle, idiom connection, level-based challenge | Chinese |
| | 3 | 成了个语 | 4+ | Word-matching, casual puzzle, visual memory game | Chinese |

[1] Search-result position reflects the order in which apps were displayed in the Apple App Store at the time of data collection and does not imply pedagogical quality or effectiveness.

| | | | | | |
|---|---|---|---|---|---|
| | 4 | 益智早教汉语拼音字母-教育游戏 | 4+ | Phonics-based, early literacy, pronunciation drills | Chinese |
| Chinese system, China Mainland App Store, Search in English | 1 | EWA | 12+ | Flashcards, interactive lessons, pronunciation practice | English |
| | 2 | Drops 点滴学语言 | 4+ | Visual vocabulary learning, swipe-based input, daily word practice | English |
| | 3 | Study cat | 4+ | Interactive animation, vocabulary drills, mini-games | English |
| | 4 | 博树 Busuu | 12+ | AI-driven feedback, structured lessons, pronunciation evaluation | English |
| | 1 | 博树 Busuu | 12+ | AI-driven feedback, structured lessons, pronunciation evaluation | Chinese |
| | 2 | HelloChinese | 4+ | Sentence-building, interactive grammar exercises, speech recognition | Chinese |
| | 3 | Learn Chinese-Chinese Skills | 4+ | Picture-word association, grammar challenges, topic-based exercises | Chinese |
| | 4 | 恐龙识字 | 4+ | Early literacy, character recognition, animation-based learning | Chinese |
| English system, US. App Store, Search in Chinese | 1 | Learning English With Momo | 4+ | Matching game, puzzle-style learning, interactive exercises | English |
| | 2 | Study Cat | 4+ | Interactive animation, vocabulary drills, mini-games | English |
| | 3 | Johnny Grammar Word Challenge | 17+ | Timed grammar challenges, vocabulary tests, real-time competition | English |
| | 4 | 懒人英语 | 12+ | Creative dubbing, pronunciation practice, speech mimicry | English |
| | 1 | Learn Chinese-Study cat | 4+ | Gamified exercises, character recognition, pronunciation drills | Chinese |

| | | | | | |
|---|---|---|---|---|---|
| | 2 | iHuman Chinese | 4+ | Character writing, phonetic guidance, interactive stroke order | Chinese |
| | 3 | Powder Game | 4+ | Physics-based puzzle, interactive learning through gameplay mechanics | Chinese |
| | 4 | Words of Wonders | | Crossword-style vocabulary building, word puzzle game | Chinese |
| English system, US. App Store, Search in English | 1 | Learn English US for beginners | 4+ | Multiple-choice quizzes, progressive difficulty, grammar exercises | English |
| | 2 | English Sentence Builder Game | 4+ | Sentence construction, word-order learning, level-based play | English |
| | 3 | Johnny Grammar Word Challenge | 17+ | Timed grammar challenges, vocabulary tests, real-time competition | English |
| | 4 | Duolingual | 4+ | Adaptive learning, sentence matching, gamified progression | English |
| | 1 | Hello Chinese | 4+ | Sentence-building, interactive grammar exercises, speech recognition | Chinese |
| | 2 | Learn Chinese-Chinese Skills | 4+ | Picture-word association, grammar challenges, topic-based exercises | Chinese |
| | 3 | WordMatch | 4+ | Word association, memory-based matching, spelling reinforcement | Chinese |
| | 4 | Study Cat | 4+ | Interactive animation, vocabulary drills, mini-games | Chinese |

## 5.2 Lack of DGBL apps for adult learners

Building on the findings presented in Section 5.1, this study further reveals a significant gap in the availability of game-based language learning applications tailored specifically for adult learners, particularly university students. Although the Apple App Store requires developers to assign an age specification (e.g., 4+, 12+, 17+) when uploading an app, this designation alone does not necessarily reflect the pedagogical design or intended audience. To address this, the learning content and game mechanics of each application were examined. The analysis confirmed that the vast majority of apps marked

“4+” employ features clearly oriented toward young children. Although English language learning games demonstrated a slightly broader age range, they similarly focused on beginner-level skills such as basic vocabulary acquisition, simple grammar drills, and repetitive task-based exercises—features not well-suited for more advanced or adult learners, who require more context-rich and cognitively demanding tasks. This dual consideration—both the default age ratings and the actual instructional content—strengthens the conclusion that most commercially available Chinese language learning games are not designed with adult learners in mind.

From a theoretical perspective, this lack of suitable adult-focused applications can be explained by the nature of existing game-based learning models. Research on GBL (Gee, 2007; Prensky, 2007; Squire, 2011) emphasizes the effectiveness of games in enhancing engagement and motivation, particularly through scaffolding, interaction, and reward-based learning. However, most commercially available language learning games simplify their mechanics to fit child-friendly, casual learning environments, where repetitive matching exercises, flashcards, and animated rewards serve as the primary learning mechanisms. While these features may be effective for young learners, they fail to address the cognitive and linguistic needs of adult learners, who often require more complex, context-driven, and goal-oriented language acquisition strategies (Ellis, 2003).

Additionally, motivation factors for adult learners differ significantly from those of children. While young learners often respond well to extrinsic motivation, such as point systems, badges, and colorful animations, adult learners tend to be more driven by intrinsic motivation, including practical language application, career-related skills, and real-world fluency (Reinders & Wattana, 2015). Most available games fail to integrate realistic, immersive scenarios that would make learning relevant for adult users, further reducing their effectiveness in a university setting. This gap highlights a critical need for the development of more sophisticated, adult-focused game-based applications that incorporate complex linguistic tasks, real-life communicative contexts, and higher-order thinking skills. Without such tools, university-level learners may be underserved by the current app ecosystem despite the growing body of research supporting the benefits of game-based learning in second language acquisition.

### 5.3 Case study: *懒人英语* (*Lazy English*)

Among the applications analyzed, *懒人英语 (Lazy English)* stands out as a rare exception to the dominant trend of child-oriented language-learning games. Unlike most applications categorized as "learning games," which primarily target young learners (ages 4+), 懒人英语 (Lazy English) is designed for users aged 12 and above, making it one of the few gamified language-learning tools explicitly catering to older learners. This distinction makes it a strong model for future DBGL applications in Chinese language acquisition.

From a GBL perspective, *懒人英语 (Lazy English)* effectively incorporates key pedagogical elements such as interactive engagement, scaffolding, and motivation, which are widely emphasized in GBL research (Gee, 2007; Prensky, 2007; Squire, 2011). For instance, the application features listening exercises, sentence-building challenges, and vocabulary drills, but its most distinctive feature is its creative dubbing function. This feature allows users to perform voice-over activities for short video clips, a process that aligns with communicative language teaching (CLT) theory (Hymes, 1971) by encouraging learners to produce spoken language in meaningful and authentic contexts actively. By engaging in dubbing, learners refine their fluency and pronunciation while receiving immediate feedback—an approach that aligns with Peterson's (2009) principles of game-based second-language acquisition.

Unlike standard pronunciation drills or passive listening exercises, dubbing requires learners to actively engage with spoken language, adjusting their pronunciation and tone to match the original dialogue. This process reinforces Swain's (1985) output hypothesis, which posits that language learners benefit most when they are required to produce language rather than merely receive input. The dubbing activities also facilitate noticing gaps in learners' pronunciation and intonation, promoting self-correction and refinement (Schmidt, 1990). Additionally, these exercises develop intonation awareness and fluency, aspects often overlooked in traditional classroom settings. Since adult learners typically seek practical, real-world language applications, this feature makes *懒人英语 (Lazy English)* a more engaging and effective alternative to memorization-based learning tools.

By creating an immersive learning environment, *懒人英语 (Lazy English)* supports Reinders and Wattana' (2015) findings that games can lower communication anxiety, allowing learners to practice speech in a low-risk, engaging setting. The app offers clips of varying difficulty levels, enabling a self-paced and adaptable learning experience. Additionally, its gamified elements, such as score tracking and progression incentives, align with Prensky's (2007) theory that interactive, goal-oriented tasks enhance learner motivation.

The dubbing feature also introduces a challenge-reward system, where learners engage in authentic language production while implicitly comparing their speech to native pronunciation. This approach aligns with Squire's (2011) concept of scaffolding in digital game-based learning, wherein learners progressively build their skills through structured gameplay mechanics. The interactive nature of dubbing ensures that learners are not passive recipients of language input but active participants, enhancing cognitive engagement and retention (Shaffer, 2006). Moreover, Ishak and Aziz (2022) argues that role-playing and decision-making tasks in games improve oral communication skills—an aspect directly reinforced by the dubbing exercises in *懒人英语 (Lazy English)*.

Another key strength of *懒人英语 (Lazy English)* is its potential for incidental vocabulary acquisition, as highlighted by Yudintseva (2015) and Choo (2015). Through dubbing tasks, learners are exposed to contextually rich, meaningful input, facilitating implicit learning of collocations, expressions, and pronunciation patterns. This supports the argument that game-based approaches enhance long-term vocabulary retention by providing repeated exposure in interactive, meaningful contexts.

 

In summary, *懒人英语 (Lazy English)* exemplifies how DGBL can be leveraged to create engaging, interactive, and effective language-learning experiences for older learners. Its dubbing feature, in particular, fosters active language production, fluency development, and pronunciation improvement, making it a valuable model for future Chinese language-learning applications.

## 6. Discussion

A key issue in the integration of Digital Game-Based Learning (DGBL) in Chinese language instruction is the widespread assumption that educational games are primarily designed for children. For instance, Apps like *HelloChinese* and *Study Cat* are primarily designed for beginners and casual learners, integrating interactive exercises with gamified elements. Despite their effectiveness, most of these applications are designed for younger learners or self-paced casual learners rather than formal academic settings. Notably, very few Chinese language learning apps explicitly target adult learners, particularly university students who require more in-depth, context-based instruction.

A search across different regional versions of the App Store reveals that when searching for English learning applications, well-known platforms like *Duolingo, Kahoot*, and *LingQ* dominate the results, catering to users aged 12 and above. However, searching for language learning games, whether for English or Chinese, overwhelmingly returns apps aimed at young children, typically rated 4+. This discrepancy suggests that game-based language learning is widely perceived as a tool for early-stage learners rather than advanced students or adults seeking proficiency.

The case of *懒人英语 (Lazy English)* illustrates how specific game-based features—such as dubbing—can be designed to align with communicative language teaching principles. While this study did not evaluate learner outcomes, the app serves as an illustrative example of how game-based strategies might be adapted to address challenges in adult Chinese language learning, particularly tonal accuracy. Game-based approaches in Chinese learning applications would need to incorporate specific phonetic scaffolding, AI-driven tone correction, and structured feedback mechanisms to support learners' mastery of tones. Additionally, as supported by Peterson (2009) and Reinders & Wattana (2015), the effectiveness of interactive pronunciation tasks in language acquisition suggests that similar methods could be integrated into Chinese learning apps to provide real-time, immersive pronunciation training. While challenges remain, particularly in adapting voice-based tasks to the unique phonetic demands of Chinese, the fundamental principles of game-based immersion, repetition, and situated learning suggest that such an approach has strong potential for adult CFL learners.

## 7. Pedagogical implications

Only a few applications, such as *懒人英语 (Lazy Englis*h), offer gamified learning experiences suitable for adults. The creative dubbing game in *懒人英语 (Laze English)* could also be effectively utilized in Chinese language learning, particularly for pronunciation practice and contextual speaking exercises. However, such applications remain rare, highlighting a major gap in the availability of game-based learning tools designed specifically for adult learners.

This raises a fundamental question: Is game-based learning truly effective in an adult Chinese language classroom? Given that most language learning games are tailored toward younger learners, their instructional design may not align with the cognitive and linguistic needs of adult learners, particularly those pursuing professional or academic language proficiency. While some English-learning games could be adapted for Chinese instruction, the lack of comprehensive, age-appropriate game-based programs tailored to adult CFL learners suggests an urgent need for further development in this area.

To make game-based learning a viable approach in adult-level CFL education, future advancements should prioritize the development of more sophisticated games incorporating real-world simulations, immersive role-playing, and AI-driven feedback tailored to the needs of adult learners. By addressing this gap, educators and developers can create more effective game-based language learning tools that cater not only to children and beginners but also to advanced learners seeking meaningful and engaging ways to master Chinese.

## 8. Limitations

This study has several limitations, primarily due to the numerous variables affecting search results in the app store. One key limitation is that the researcher conducting the searches is a woman in her twenties, which may have influenced the individualized search results due to AI and big data algorithms that personalize app recommendations. As search algorithms become increasingly tailored to individual users, different people may see different rankings and app suggestions, making it difficult to generalize findings across all users. Additionally, we found inconsistencies between search results on Mac computers and iPhones. The Mac App Store includes filtering options, allowing users to refine their search results, whereas the iPhone App Store does not. This discrepancy suggests that search conditions may significantly impact the visibility and ranking of language-learning applications. Another major limitation is the influence of advertisements on mobile app rankings. Many search results were directly affected by paid promotions, meaning that certain apps appeared at the top due to advertising rather than user engagement or pedagogical effectiveness. This highlights a challenge in evaluating app popularity and effectiveness solely based on search rankings. Given these limitations, future researchers should further explore how individualized search algorithms, device-based variations, and advertisement-driven rankings influence app store search results. A broader, multi-device, and multi-user study could provide more generalizable insights into the availability and effectiveness of GBL applications for CFL learners.

 

## 9. Conclusion

This study explored the role of Digital Game-Based Learning (DGBL) in Chinese as a Foreign Language (CFL) instruction, particularly for adult learners, and identified a gap in the availability of suitable applications for this demographic. Through a comparative analysis of language-learning applications, it became evident that most existing DGBL tools cater primarily to children and beginner learners, offering limited support for the advanced linguistic and cognitive needs of adult learners. The findings highlight that current game-based language learning applications emphasize basic language skills, such as vocabulary recognition and grammar drills, while lacking elements crucial for adult learners, including task-based interactions, pronunciation refinement, and real-world application scenarios. Moreover, factors such as app store ranking mechanisms, algorithmic personalization, and the influence of advertisements further limit the accessibility of effective CFL learning applications.

Despite these challenges, DGBL remains a promising approach to enhancing engagement and retention in language learning. This study underscores the need for more adaptive, interactive, and context-rich applications tailored to adult CFL learners. Future research should focus on the development of AI-driven feedback, immersive role-playing, and adaptive learning models to bridge this gap. By addressing these limitations, educators, developers, and researchers can work together to create more effective DGBL tools that align with the learning objectives of university students. The evolution of technological advancements in educational gaming presents an opportunity to refine and expand the potential of game-based CFL learning for adult learners, ensuring that DGBL is not just engaging but also pedagogically effective.

## References

Bytheway, J. (2014). In-game culture affects learners' use of vocabulary learning strategies in massively multiplayer online role-playing games. *International Journal of Computer-Assisted Language Learning and Teaching, 4*(4), 1–13. https://doi.org/10.4018/ijcallt.2014100101

Chen, H. & Lin, Y. (2015). An examination of digital game-based situated learning applied to Chinese language poetry education, *Technology, Pedagogy and Education.* https://doi.org/10.1080/1475939X.2015.1007077

Chen, Y. L., & Hsu, C. C. (2020). Self-regulated mobile game-based English learning in a virtual reality environment. *Computers & Education, 154*, 103910. https://doi.org/10.1016/j.compedu.2020.103910

Choo, K. F. (2015). The effects of game-based practice on young learners' vocabulary acquisition in learning Chinese language. *Journal of Research, Policy & Practice of Teachers & Teacher Education, 5*(1), 46–67.

DeHaan, J. W. (2005). Acquisition of Japanese as a foreign language through a baseball video game. *Foreign Language Annals, 38*(2), 278–282. https://doi.org/10.1111/j.1944-9720.2005.tb02492.x

Denoyelles, A., Brown, T., Seilhamer, R., & Chen, B. (2023). The evolving landscape of students' mobile learning practices in higher education. *EDUCAUSE Review*. https://er.educause.edu/articles/2023/1/the-evolving-landscape-of-students-mobile-learning-practices-in-higher-education

Egenfeldt-Nielsen, S. (2007). *Educational potential of computer games*. Continuum.

Ellis, R. (2003). *Task-based language learning and teaching*. Oxford University Press.

Gabbianelli, G., & Formica, A. (2017). Difficulties and expectations of first level Chinese second language learners. In I. Kecskes (Ed.), *Explorations into Chinese as a second language* (Educational Linguistics, Vol. 31, pp. 183–206). Springer. https://doi.org/10.1007/978-3-319-54027-6_8

Gee, J. P. (2004). *What video games have to teach us about learning and literacy*. Palgrave Macmillan.

Gee, J. P. (2007). *Good Video Games and Good Learning*. https://doi.org/10.3726/978-1-4539-1162-4

Hymes, D. H. (1971). *On communicative competence.* University of Pennsylvania Press.

Ishak, S. A., & Aziz, A. A. (2022). Role play to improve ESL learners' communication skills: A Systematic Review. *International Journal of Academic Research in Business and Social Sciences, 12*(10), 884–892.

Lan, Y. J. (2015). Contextual EFL learning in a 3D virtual environment. *Language Learning & Technology, 19*(2), 16–31. https://doi.org/10.64152/10125/44412

Lin, T. J., & Lan, Y. J. (2015). Language learning in virtual reality environments: Past, present, and future. *Educational Technology & Society, 18*(4), 486–497.

Luckin, R., Holmes, W., Griffiths, M., & Corcier, L. B. (2016). *Intelligence unleashed: An argument for AI in education*. Pearson.

Ng, M. P., Alias, N., & Dewitt, D. (2022). Effectiveness of a gamification application in learning Mandarin as a second language. *Malaysian Journal of Learning & Instruction, 19*(2), 183–211. https://doi.org/10.32890/mjli2022.19.2.7

Peterson, M. (2009). Computerized games and simulations in computer-assisted language learning: A meta-analysis of Research. *Simulation & Gaming, 41*(1), 72–93. https://doi.org/10.1177/1046878109355684

Piaget, J. (1962). *Play, dreams and imitation in childhood*. Norton.

Poole, F. J., & Clarke-Midura, J. (2020). A systematic review of digital games in second language learning studies. *International Journal of Game-Based Learning, 10*(3), 1–15. https://doi.org/10.4018/ijgbl.2020070101

Poole, F., Clarke-Midura, J., & Ji, S. (2022). Exploring the affordances and effectiveness of a digital game in the Chinese dual language immersion classroom. *Journal of Technology and Chinese Language Teaching, 13*(1), 46–73. http://www.tclt.us/journal/2022v13n1/pooleclarkeji.pdf

Prensky, M. (2007). *Digital game-based learning*. Paragon House.

Qiao, H., & Zhao, A. (2023). Artificial intelligence-based language learning: illuminating the impact on speaking skills and self-regulation in Chinese EFL context. *Frontiers in Psychology, 14, 1255594.* https://doi.org/10.3389/fpsyg.2023.1255594

Rankin, Y. A., Gold, R., & Gooch, B. (2006). Playing for keeps. *In ACM SIGGRAPH 2006 Educators Program on-siggraph'06* (p. 44). https://doi.org/10.1145/1179295.1179340


 

Reinders, H., & Wattana, S. (2015). The effects of digital game play on Second language interaction. *International Journal of Computer-Assisted Language Learning and Teaching, 5*(1), 1–21. https://doi.org/10.4018/ijcallt.2015010101

Rho, Y. A. (2024a). Teaching idiomatic expressions through storytelling based on narrative thinking. *STEM Journal, 25*(3), 17–29. https://doi.org/10.16875/stem.2024.25.3.17

Schmidt, R. (1990). The role of consciousness in second language learning. *Applied Linguistics*, 11, 17–46. http://dx.doi.org/10.1093/applin/11.2.129

Shaffer, D. W. (2006). *How computer games help children learn.* Palgrave Macmillan.

Shadiev, R., Wen, Y., Uosaki, N., & Song, Y. (2023). Future language learning with emerging technologies. *Journal of Computers in Education, 10*(3), 463–467. https://doi.org/10.1007/s40692-023-00285-9

Squire, K., & Jenkins, H. (2011). *Video games and learning: Teaching and participatory culture in the Digital age*. Teachers College Press.

Swain, M. (1985). Communicative competence: Some roles of comprehensible input and comprehensible output in its development. In S. Gass & C. Madden (Eds.), *Input in second language acquisition* (pp. 235–253). Newbury House.

Vygotsky, L. S. (1978). *Mind in society: The development of higher psychological processes*. Harvard University Press.

Xu, W., Zhang, H., Sukjairungwattana, P., & Wang, T. (2022). The roles of motivation, anxiety and learning strategies in online Chinese learning among Thai learners of Chinese as a foreign language. *Frontiers in Psychology, 13*, 962492. https://doi.org/10.3389/fpsyg.2022.962492

Yang, L., & Li, R. (2023). Contextualized game-based language learning: Retrospect and prospect. *Journal of Educational Computing Research, 62*(1), 357–375. https://doi.org/10.1177/07356331231189292

Yu, Y. T., & Tsuei, M. (2023). The effects of digital game-based learning on children's Chinese language learning, attention, and self-efficacy. *Interactive Learning Environments, 31*(10), 6113–6132. https://doi.org/10.1080/10494820.2022.2028855

Yudintseva, A. (2015). Game-enhanced second language vocabulary acquisition strategies: A systematic review. *Open Journal of Social Sciences, 03*(10), 101–109. https://doi.org/10.4236/jss.2015.310015

Zhang, L., & Chen, Y. (2021). Examining the effects of gamification on Chinese college students' foreign language anxiety: A preliminary exploration. *In Proceedings of the 2021 4th International Conference on Big Data and Education (ICBDE '21)*. Association for Computing Machinery, New York, NY, USA, 1–5. https://doi.org/10.1145/3451400.3451401

Zhang, L., Fang, L., & Shang, J. (2024). The application of gamified virtual scenarios in international Chinese education: A perspective on the utility of discourse cognition and emotional motivation (游戏化虚拟情境在国际中文教育中的应用：语篇认知与情绪动机的效用视角). *e-Education Research (电化教育研究)*, *45*(7), 121–128. https://doi.org/10.13811/j.cnki.eer.2024.07.016

URL (网址): http://www.tclt.us/journal
Email (电子邮件): editor@tclt.us