基于 AI 技术的 CVC 中文视听语料库设计与应用
(The Design and Application of a Chinese Audio-Visual Corpus Based on AI Technology)

Wang, Tao
(王涛)
Beijing International Studies University
(北京第二外国语学院)
toddy@bisu.edu.cn

摘要：AI 赋能下的语言教学生态环境带来了新型人机协作关系，学习者获取信息与知识的方式变得更加场景化、智能化，教师需要利用多种资源媒介和 AI 技术手段引导学生形成主动探究问题、整合知识的能力。CVC 中文视听语料库通过采集来自汉语母语者生活中使用的视频语言材料，将教材内容、本体知识和视频语料数据相关联，为用户提供场景化的教学资源应用服务。基于 AI 技术的视听语料库应用可提供以视频内容检索为核心的资源型、智慧化语言教学手段。在服务语言教学同时，视频语料标注结果将有助于自然语言处理（ NLP ）和计算机视觉（ CV ）交叉领域的语言模型训练，在言语行为识别、多模态分析、情感分析等方面满足人工智能对多模态大数据的需求，反哺人工智能领域的未来发展进程。

Abstract: Cultivating an AI-powered language teaching ecosystem has introduced a new model of human-computer collaboration. Many learners are acquiring information and knowledge in a manner that is more contextualized and intelligent. Many educators are adaptive to explore a variety of media resources and AI technologies to guide learners in developing capabilities in problems resolving and knowledge integration. This article describes the design and application of a Chinese Audio-Visual Corpus (CVC) that collects visual language materials from native Chinese speakers in their daily lives to associate textbook content, ontological knowledge, and video materials data in order to provide learners with context-aware teaching resource applications. This AI-based audio-visual corpus offers resourceful and intelligent language teaching methods with the priority of video content retrieval. In addition to serving language teaching, the annotated results from this audio-visual corpus will aid in the training of language models in the interdisciplinary field of Natural Language Processing (NLP) and Computer Vision (CV). It meets the demand for multimodal big data in artificial intelligence for applications such as multimodal discourse analysis, speech act recognition, and sentiment analysis, thereby contributing to the future development of the field of artificial intelligence.

关键词：视听语料库；视频检索；人工智能；多模态；国际中文教育

Keyword: Audio-Visual Corpus, Video Retrieval, Artificial Intelligence, Multimodal, International Chinese Language Education

1. 引言

当前，AI 赋能下的语言教学生态环境带来了新型人机协作关系，学习者获取信息与知识的方式变得更加场景化、智能化。教师的角色也随之转变，从传统的知识提供者转化为引导者，利用多种资源媒介和 AI 技术手段引导学生形成主动探究问题、整合知识的能力。传统语言教学多采用文本材料，教师讲解语言知识规则的同时缺乏真实语言实例的使用，教学中难免存在一定的局限性，不利于第二语言习得效果。视频语料在呈现情景语境和社会语境的同时，可以还原真实交际过程中的语言要素和非语言要素，包含语音、文字、情境、表情、肢体动作、交际身份、文化背景等不同符号系统。Ginsburgh（ 1935 ）Hendrix（ 1939 ）Palomo（ 1940 ）Kern（ 1959 ）和 Fallahkhair 等（ 2004 ）在相关研究中均提到真实视频在语言教学中的优势，指出有声电影、电视节目可以在有机语境中呈现目标语言，并推荐在外语教学中开展应用。冯惟钢（ 1995 ）沈履伟（ 1995 ）唐荔（ 1997 ）王飙（ 2009 ）张璐（ 2011 ）刘立新、和邓方（ 2018 ）等国内学者也先后进行了影视资源在对外汉语教学中的应用研究，并就视听说教材的选材依据、编制理念进行了深入探讨。

CVC 中文视听语料库（ www.chinafoucs.net.cn ）以下简称 CVC 语料库 ）采集汉语母语者生活中使用的真实语言材料，提供以视频节目内容检索为核心的资源型、智慧化语言教学工具。教师可针对教材词汇、语法等级大纲选取具有典型语境的视频语料展示语言功能实例，通过情景语境和文化语境促进学习者认知发展过程。通过结合情境设计练习活动，突出教学重点、解决教学难点、提高教学效率。将师生从单纯使用教材转向学材资源的开发、利用，变为教学活动的设计者和参与者，实现在用中学、以用促学、学用合一的目的。本文对 CVC 语料库的设计理念和使用方法进行详细说明。

2. 设计理念

近年来，随着现代外语教学理念和新文科建设的发展，语言学和语言教学也更为关注跨学科应用和话语研究转向，研究对象从平面媒体扩展到新媒体，从单一模态纸质出版物到多模态数字文化出版领域。研究方法从侧重静态语言形式结构描写到结合动态功能解释，从单一学科到多学科的交叉融合也是必然趋势。王涛（ 2012 ）提出视频语料库建设的必要性，并对视频语料采集、标注、检索实现过程进行了说明。2017 年，在北京第二外国语学院举办的“第一届汉语视听说教学理论与应用研讨会”上，王涛发表题为“多模态视角下的视听说教材立体化建设及教学创新”的主旨

报告，就视听说教材编写理念、教学模式创新进行了总体阐述。在“第二届汉语视听说教学理论与应用研讨会暨新媒体数字环境下的汉语教学创新研究学术会”上，王涛（ 2018 ）发表题为“视听说课程大纲设计与教学实践研究”报告，指出视听说课程和教学系统建立在视频语料库基础之上，教学内容全部来自真实语料，是汉语母语者生活中使用的语言。论文从系统功能语言学视角对视听语篇类型进行了阐述，进一步明确了课程开发与教学系统中视频语料库的作用（ 见图 1 ）。

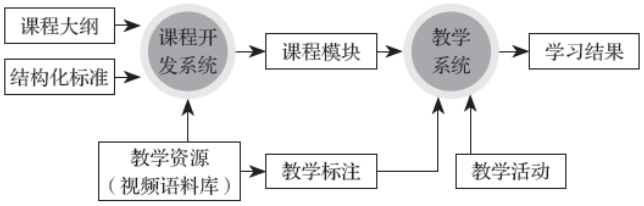

图 1 视听说课程开发与教学系统动力学模型

中文视听语料库研究项目正式被国际中文教学领域关注是在美国初中级中文教学兴趣组（ CLTA-SIG ）举办的线上讲座，王涛（ 2022 ）发表题为“中文视听语料库应用”的专题报告，详细介绍了视听语料库的核心理念及检索功能。讲座由耶鲁大学梁宁辉主持，相关资料整理发布在耶鲁大学初、中级中文教学交流网站[1]。在美国 TCLT《科技与中文教学》[2]期刊举办的科技教学系列讲座中，王涛（ 2024 ）发表题为“视听材料选取、教材编写及相关 AI 技术工具介绍”报告，介绍了 CVC 语料库的最新进展及教学应用成果。讲座由宾州印第安纳大学刘士娟主持，多位中外一线教师就语料库应用展开了深入探讨。

2.1 语料库设计

语料库语言学是国际中文教育领域的基础学科，通过对大规模口语或书面语真实语料统计分析，挖掘语言事实在意义和表达形式上的内在规律，为语言教学提供应用平台和实证性研究支持。根据建设目标、用途、语料来源、采集加工路线等不同方面，语料库存在多种类型。由于音视频介质语料采集、加工、存储成本较高，国内现有大型语料库多为文本形式，数据来源于古汉语、文学作品、新闻、报刊、社交媒体等书面语材料。詹卫东,郭锐等（ 2019 ）、荀恩东,饶高琦等（ 2016 ）研究结果显示，北京语言大学 BCC 语料库，口语语料主要来自新浪微博和影视字幕，占比为 6.3%。北京大学 CCL 语料库，现代文献中文学语料占比高达 92.15%，口语语料占比仅为 0.26%。普遍存在语料库采样不平衡、媒介形式单一、多模态语料库建设相对滞后等问题。另外，传统文本语料库查询系统采取的是一种基于字频、词频概率统计的科研量化手段，工具性地分析文本聚合关系并不能反映语言交际使用过程的全貌，无法满足智慧教育背景下语言学和语言教学研究对真实语料的需求。

CVC 语料库是一套面向国际中文教育领域的大规模音视频数据库，由北京第二

[1] 美国耶鲁大学初、中级中文教学交流网站 https://campuspress.yale.edu/exchange/

[2] 美国《科技与中文教学》期刊网址 http://www.tclt.us

外国语学院汉语学院王涛设计，北京视听说科技有限公司和自然语义（ 青岛 ）科技有限公司联合开发。该语料库依据《国际中文教育水平等级标准》（ GF0025-2021 ）和《汉语视听说课程大纲的研发与应用案例》研究成果，结合词汇等级、平均语速、百字生词率等参数，通过优化算法实现语料自动标注及检索功能。系统整体架构由采集层、数据层、系统中台、用户管理层和应用接口层五部分组成：

1) 采集层由语料采集和多数据源语料对齐两大模块组成，包括视频对象文件存储、文本存储、同步管理模块。
2) 数据层主要由语料加工数据库和结构化元数据库两大模块组成，包括文本对象存储、文本纠错、原数据模块、分词切分、词性标注以及视频类型数据、语言形式数据、语篇类型数据、语体类型数据等元数据项组成。
3) 系统中台主要由算法池和应用程序两大模块组成，包括分词算法、词性标注算法、文本纠错算法、语言量化算法、语料维度、词汇图谱、多语种翻译以及接口管理等模型。自然语言处理（ NLP ）算法采用 HanLP 框架，是全球 NLP 开源领域（ Github ）用户量最大最受欢迎的 NLP 框架，具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点。
4) 用户管理层包括用户基本信息管理、角色管理、个性化管理、应用管理、多维度检索功能、扩展管理功能等模块。
5) 应用接口层负责提供标准 API 接口，与其他系统进行集成交互。

2.2 语料采集

视频语料采集来自合作版权机构、公共媒体和网络视频节目，通过采集层字幕提取模块及语音识别技术转换为可针对语言内容检索的数据文本。用户输入关键词可以得到相应视频语料检索结果。视听语料库具有更强的交际互动性和功能阐释性，在语境、语用研究、多模态功能话语分析、互动语言学领域更符合语言学及语言教学需求。

2.3 语料分布

根据王涛（ 2018 ）视听说课程大纲研究，CVC 语料库分为电影、电视剧、情景剧、纪录片、综艺、访谈、辩论、朗诵、演讲、讲座、新闻、歌曲 12 类，将视频语料与语篇类型、语体程度以及语言表达形式关联，以便提供分类精准检索服务。截至 2024 年 6 月，语料分类及规模统计如下（ 见表 1 ）：

表 1 CVC 视听语料库规模分布统计

视频类型		字节数	百分比
电影		990651	6.96%
电视剧		9969426	70%
综艺		7080	0.05%
访谈		39189	0.28%
纪录片		2942924	20.66%

辩论		0	0.00%
朗诵		6339	0.04%
新闻		17511	0.12%
讲座		20727	0.15%
情景剧		55497	0.39%
演讲		166257	1.17%
歌曲		26626	0.19%
总计		**14242227**	**100%**

2.4 搜索引擎

CVC 语料库搜索引擎面向全球个人用户免费开放，支持通用检索、上下文全文检索、组合条件筛选检索三种模式，用户可通过以下两种方式登录：

1) 电脑 web 浏览器方式，输入网址 https://client.chinafocus.net.cn，使用微信或 WeChat 扫码登录（ 见图 2 ）。

图 2 CVC 语料库首页

2) 手机移动端使用微信搜索“中文视听”公众号，点击“语料检索”登录。

CVC 语料库首页采用通用检索模式，支持词汇和常见构式检索，可识别超过 35 万条核心词库词汇。如搜索结果显示“抱歉没有找到相关的视频语料资源”，需点击开启“上下文检索”模式，可输入短语、关键词组合进行字符串全文匹配方式检索（ 见图 3 ）。

图 3 CVC 语料库上下文全文检索模式

高级用户模式仅向《中国微镜头》教材用户和合作机构院校教师开放，如有教学、科研用途需求可通过邮件发送至 mail@chinafocus.net.cn 申请 VIP 权限。VIP 授权用户可使用云计算软件服务，实现语料分布结果、视频类型、语言形式、语篇类型、语体类型、适用等级等自定义组合条件检索，并提供拼音标注、多语种机器翻译、词性标注、等级分布、语义图谱、语用标签、教材信息标注、视频课件编辑等 AI 语言辅助教学功能（ 见图 4 ）。

图 4 CVC 语料库 VIP 模式

CVC语料库采用基于神经网络的机器翻译模型，目前VIP模式支持语种包括英语、日语、韩语、德语、法语、俄语、西班牙语、阿拉伯语字幕翻译，可提供葡语、泰语、越南语、缅甸语等小语种机器翻译定制化服务（ 见图5 ）。

图 5 CVC 语料库多语言机器翻译

2.5 检索式

CVC 语料库提供中文语料库通用规则检索式，查询符合条件的语料结果。支持关键词、通配符、词性符号、空格或“+”搜索及常见语法构式检索。分词算法、词性标注算法和文本纠错算法采用 HanLP 框架算法模型，线上模型训练数据来自 9970 万字的大型综合语料库，覆盖新闻、社交媒体、金融、法律等多个领域。检索式规则说明及词性符号对照表如图 6、图 7 所示。

检索式	用法解释
白/a	检索形容词性“白”
白/d	检索副词词性“白”
吃*饭	检索离合词“吃饭”的用法
洗. 澡	离合词“洗澡“中间有一个单字
参加n	检索动宾结构“参加”的搭配
我+. /c+你	“我”和“你”之间有一个单音节连词
跑. /v	跑作前缀的双音节动词
. . /v办法	检索双音节动词与“办法“的搭配
越*越	检索“越……越……”结构句型
爱v不v、一v就	检索相关构式
非[a v n]不可	检索“非”后加形容词或动词或名词，再接“不可”

图 6 CVC 语料库检索式规则说明

词性符号	词性类别	词性符号	词性类别	词性符号	词性类别	词性符号	词性类别
Ag	形语素	i	成语	o	拟声词	vn	名动词
a	形容词	j	简称略语	p	介词	w	标点符号
ad	副形词	k	后接成分	q	量词	x	非语素字
an	名形词	l	习用语	r	代词	y	语气词
b	区别词	m	数词	s	处所词	z	状态词
c	连词	Ng	名语素	Tg	时语素	un	未知词
Dg	副语素	n	名词	t	时间词	h	前接成分
d	副词	nr	人名	u	助词	g	语素
e	叹词	ns	地名	Vg	动语素	nz	其他专名
f	方位词	nt	机构团体	v	动词	vd	副动词

图 7 CVC 语料库词性符号对照表

3. 语料库应用

视频语料为学习者呈现了虚拟自然目的语的教学环境，有利于构建师生双方共享认知环境。从而在有效降低认知负荷前提下实现可理解性输入，为学习者创造有意义的输出机会。虞莉（ 2020 ）认为体演文化教学法通过身临其境的“体演”活动以及周密的课程设计，将语言教学与文化教学紧密结合，使语言学习者不仅能掌握语言技能，而且能获取跨文化交际能力，从而有效并得体地与母语者交流。CVC 语料库不仅可以应用于语音、词汇、语法、文化教学，还可以和体演法、任务型教学法相结合，丰富教学手段和教学设计，让课堂教学延伸到课前、课后环节，弥补常规课堂语言教学模式和平面教材的不足，进一步将“结构—功能—文化”为核心的教学理念场景化、话题化、实例化。教学应用示例如下：

3.1 语音教学

语音是中文教学的基础，传统听力技能教学内容普遍采用人工录制的教学语言，与母语者现实生活中使用的自然语言存在一定差异。CVC 语料库更为关注词汇、语法在真实口语交际活动中的表现形式，涉及音系-句法接口研究。即意义是如何通过音系层表达的，包括声调、音节组合、语调、节奏、轻重、停连等韵律结构特征。例如上声在自然语流中的调值多是半三或变调，学习者如果仅仅把注意力放在课文标注的声调符号上，容易忽视对语音本身的辨识。通过视频语料，学习者可以在真实情景中模仿正常语流中的发音，完成语音自然习得过程。用户在阅读教材注释时，可使用微信扫码观看视频语料，第一次扫码为登录语料库，以后扫码可直接观看（ 见图 8 ）

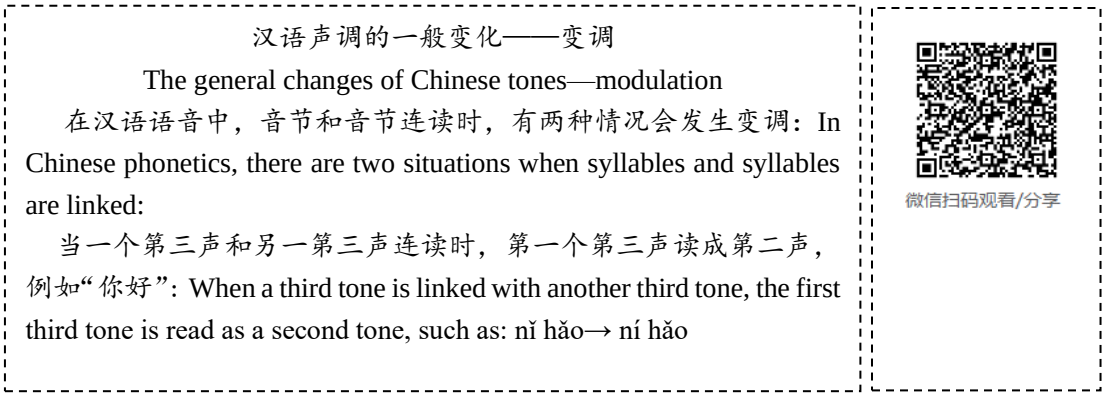

图 8. 语音教学示例

3.2 词汇教学

兼类词、虚词、副词是词汇教学的难点，CVC 语料库包含了大量真实语料数据，可以帮助学习者看到词汇在真实语境中的使用实例，增强学习者的自然表达能力和应用创造能力。值得注意的是，由于兼类词通常具有两种或两种以上不同语法特征，在不同语境中承担不同句法成分和词类属性，检索时需要标注词性符号以便提供精

准结果。例如根据《国际中文教育水平等级标准》,形容词词性“白”是 HSK1 级词汇，可通过“白/a”进行检索。副词“白”是 HSK3 级词汇，则需要输入“白/d”检索。另外，现有教学辞书注释中通常仅关注语义结构方面的描写，并没有给出主观情态功能的解释。例如语气副词“明明”，除了表示客观实事显然如此，还常用于表达自责或抱怨、指责别人的语气（ 见图 9 ）。

语气副词“明明”

《现代汉语词典》第 7 版中解释“表示显然如此或确实（下文意思往往转折）”

《现代汉语八百词》“表示显然这样。用‘明明’的小句前或后常有反问或表转折的小句。

微信扫码观看/分享

图 9 词汇教学示例

3.3 语法教学

演绎法是常见的语法教学方法，教师通常先展示语法形式结构，之后再进行讲解、操练。传统教学方法固然有助于形成规范的教学思路和教学模式, 但也往往容易被教材和教学模式所约束，容易使讲练停留在机械性操练层面。CVC 语料库搜索引擎支持常见语法构式检索，以“把字句”教学为例，教师在讲解语法规则的同时，往往还需要将句法结构与功能、语境结合进行句型操练。如果直接输入“把”字搜索，目前可得到 9492 条结果，其中还包含了量词、动词结果；输入“把/p”检索，可以得到 9104 条介词词性结果；输入“把 n+v”可以缩小到 2119 条结果；输入“把 n+v+在”仅为 81 条结果。同理，输入“所＋v+的”可以得到构成“所……的”字短语作主语、宾语的例句；输入“所 v*的+n”可以得到“所”用于动词短语前作定语的用法。

与演绎法不同的是，归纳法、情境法、任务型教学法重视语言在交际中的实际使用，教学操作过程中可借助视频语料提供一些具有真实情境的语言用例，丰富教学手段和教学设计。例如疑问代词“怎么”表达主观情态意义的非疑问用法（ 见图 10 ）

疑问代词“怎么”的非疑问用法

形式结构：

1. np/vp+怎么+v
2. S+怎么+vp/ap

释义：用于质问、责骂，带有生气的语气。例句：

1．你怎么开车的！

2．你怎么说话呢！

微信扫码观看/分享

图 10 语法教学示例

3.4 文化教学

CVC 视听语料库既包含饮食、茶、京剧、节日等中国传统元素，也涉及当代家庭、教育、婚姻、医疗、体育、贸易、一带一路、城市化等社会话题节目。视听材料可以通过融媒体视角聚焦中国社会发展进程，增强中华文化感召力和话语说服力，进一步完善中华优秀传统文化和中国式现代化进程的传播路径，推动中文国际话语体系多模态构建的故事化、形象化建设，发挥国际中文教育的“社会窗口”作用。例如《中国微镜头》视听说教材中级下《家庭篇》中介绍了现代年轻人的婚恋话题，课文的文化链接部分对“喝喜酒”和“交杯酒”进行了文化注释（ 见图 11 ）。

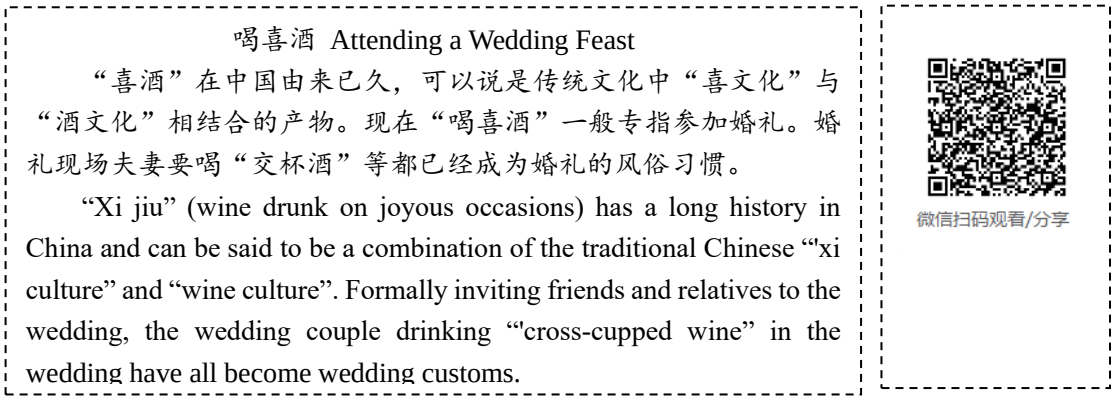

图 11 文化教学示例

4. 结语

智慧教育背景下的视听语料库研制，可以将教材内容、本体知识和视频语料相关联，为用户提供场景化的教学资源数据应用服务。探索人工智能技术与教育的融合创新，驱动教育理念、教学模式创新，培养学生自主学习能力，完成传统课堂教学向线上、线下混合模式的转变。从教学视角来看，视听语料库通过信息技术对教学资源进行重组，优化传统课堂教学流程，进一步拓展课前/课后学习环境，提高教学效率，提升教学质量，推动 ERP（ Education Resource Planning ）的实施；从学习者视角看，视听语料库通过数据驱动学习者主动构建认知过程，可以帮助学习者在真实情境中运用所学知识，发挥主动感知、思维、创新能力，促进深度学习方式的开展。

随着大数据和人工智能技术的不断发展，语料库建设及应用成果广泛运用于语言学、翻译、二语教学、融媒辞书编撰、教材开发等领域。CVC 语料库建设将有助于填补国际中文教育领域多模态平衡语料库研究空白。本文着重介绍了该语料库的设计理念和语料检索功能使用说明，后续将进一步结合具体案例进行教学应用经验总结分享。未来围绕视听语料库应用研究将从以下三个层面展开：宏观层面以语言学理论为基础，结合语用研究、互动语言学、多模态功能话语分析方向；中观层面可以结合任务型教学法、情境法、沉浸式教学法、体演文化教学法（ PCA ）产出导向教学法（ POA ）Backword Design、混合教学模式等；微观层面从视频语料在不

同课型中的使用入手，进行教学设计、立体化教材/学材开发、教学资源、教学评价等方面的应用研究。在服务语言教学同时，视频语料标注结果还将有助于自然语言处理和计算机视觉交叉领域的语言模型训练，在言语行为识别、多模态分析、情感分析等方面满足人工智能对多模态大数据的需求，反哺人工智能领域的发展进程。

致谢: 本项目获教育部中外语言交流合作中心 2023 年国际中文教学实践创新项目资助,《国际中文视听教材国际传播能力创新与实践研究》YHJXCX23-024。

参考文献

Feng, W. (1995). Audio-visual speaking teaching and the compilation of teaching materials. *Chinese Teaching in the World*, 4, 95-100. [冯惟钢. (1995). 视听说教学及其教材的编写. *世界汉语教学*, 4, 95-100.]

Liu, L., Deng, F. (2018). Compiling audio-visual-oral teaching materials with authentic data. *TCSOL Studies*, 3,31-37. [刘立新, 邓方. (2018). 基于“真实”材 料的视听说教材编制. *华文教学与研究*, 3, 31-37.]

Shen, L (1995). A brief discussion on the 'audio-visual speaking' teaching of Chinese as a foreign language. *Journal of Tianjin Normal University (Social Sciences)*, 1, 78-80. [沈履伟. (1995). 浅谈对外汉语的“视听说”教学. *天津师大学报(社会科学版)*, 1, 78-80.]

Tang, L. (1997). An initial exploration of Chinese 'audio-visual speaking' course teaching. *Journal of Open Learning*, 3, 20-23. [唐荔. (1997). 汉语“视听说”课程教学初探. *北京广播电视大学学报*, 3, 20-23.]

Wang, B. (2009). The audio-visual TCFL products published in Mainland China: A review and prospect. *Chinese Teaching in the World*, 2, 252-261. [王飙. (2009). 中国大陆对外汉语视听教材评述与展望. *世界汉语教学*, 2, 252-261.]

Wang, T. (2012). An initial exploration of the construction of a video corpus for teaching Chinese as a foreign language. *Series of International Research on Chinese Language(I)*, 1,175-181. [王涛. (2012). 对外汉语教学视频语料库建设初探.*国际汉语研究论丛（ 一 ）*, 1, 175-181.]

Wang, T. (2018). The development of a syllabus for the Chinese visual-audio-oral course and an application case. *Journal of International Chinese Teaching*, 4, 51-58. [王涛. (2018). 汉语视听说课程大纲的研发与应用案例. *国际汉语教学研究*, 4, 51-58.]

Xun, E., Rao, G., Xiao, X., & Zang, Ji. (2016). The construction of the BCC corpus in the age of big data. *Corpus Linguistics*, 1, 93-109. [荀恩东, 饶高琦, 肖晓悦, 臧娇娇. (2016). 大数据背景下 BCC 语料库的研制.语料库语言学, 1, 93-109.]

Yu, L. (2020). The performed culture approach: Intellectual history and core concepts. *Journal of International Chinese Teaching*, 2,42-49. [虞莉. (2020). 体演文化教学法: 渊源与核心. *国际汉语教学研究*, 2, 42-49.]

Zhan, W., Guo, R., Chang, B., Chen, Y., & Chen, L. (2019). The building of the CCL corpus: Its design and implementation. *Corpus Linguistics*, 1, 71-86. [詹卫东, 郭锐, 常宝宝, 谌贻荣, 陈龙. (2019). 北京大学 CCL 语料库的研制. *语料库语言学*, 1, 71-86.]

Zhang, L. (2011). A Study on the selection of content and topics for audio-visual speaking materials in teaching Chinese as a foreign language. *Modern Chinese*, 1, 142-145. [张璐. (2011). 对外汉语视听说教材内容取材和话题选择研究. *现代语文*, 1,142-145.]

