College and Job 职业教育发展, 2026, 15(6), 162-167
Published Online June 2026 in Hans. https://www.hanspub.org/journal/ve
https://doi.org/10.12677/ve.2026.156257

Hans汉斯

“中文 + 人工智能”职业核心汉语词表构建研究

张新宇[1*]，于嘉林[2]

[1]浙江传媒学院国际教育学院，浙江 杭州
[2]浙江传媒学院新闻与传播学院，浙江 杭州

收稿日期：2026年4月16日；录用日期：2026年6月2日；发布日期：2026年6月10日

摘 要

本文立足“中文 + 职业技能”教育需求，聚焦人工智能领域，构建“中文 + 人工智能”职业核心汉语词表。研究选取人工智能本科五门核心课程，从课程视频、教材、期刊论文三类语料中提取术语。经语料转换、清洗分词后，以词频(≥40次)、专业性(排除HSK一到六级通用词)、跨子语料库分布(覆盖4个以上)为标准筛选，结合人工审核剔除非专业词汇，最终形成含365个词语的核心词表。研究弥补了现有“中文 + 职业技能”人工智能行业教学资源不足的问题，为“中文 + 人工智能”教学提供专业词汇参考。

关键词

中文 + 职业技能，人工智能，术语库构建，汉语词表

Research on the Construction of a Core Chinese Vocabulary List for “Chinese + Artificial Intelligence” Careers

Xinyu Zhang[1*], Jialin Yu[2]

[1]School of International Education, Communication University of Zhejiang, Hangzhou Zhejiang
[2]School of Journalism and Communication, Communication University of Zhejiang, Hangzhou Zhejiang

Received: April 16, 2026; accepted: June 2, 2026; published: June 10, 2026

Abstract

Based on the educational demand for “Chinese + Professional Skills”, this study focuses on the field

*通讯作者。

文章引用: 张新宇, 于嘉林. “中文 + 人工智能”职业核心汉语词表构建研究[J]. 职业教育发展, 2026, 15(6): 162-167.
DOI: 10.12677/ve.2026.156257

of artificial intelligence and constructs a core Chinese vocabulary list for "Chinese + Artificial Intelligence" professions. Five core undergraduate courses in artificial intelligence are selected, and relevant terms are extracted from three types of corpora: course videos, textbooks and academic papers. After corpus conversion, cleaning and word segmentation, the terms are screened according to three criteria: word frequency (≥40 occurrences), professionalism (excluding general words from HSK Level 1 to Level 6), and cross-corpus distribution (covering more than 4 sub-corpora). Non-professional words are further eliminated through manual review, and a core vocabulary list containing 365 words is finally formed. This research addresses the shortage of teaching resources for "Chinese + Professional Skills" in the artificial intelligence industry, and provides a reference for professional vocabulary in "Chinese + Artificial Intelligence" teaching.

Keywords

Chinese + Professional Skills, Artificial Intelligence, Terminology Base Construction, Chinese Vocabulary List

1. 引言

2019 年，国际中文教育大会正式提出“中文 + 职业技能”教育项目[1]。为解决“中文 + 职业技能”教学资源相对匮乏的问题，中外语言交流合作中心于 2021 年底发布《“中文+职业技能”教学资源建设行动计划（2021-2025 年）》[1]，以需求导向、服务发展、标准引领、多元合作为基本原则；以研制高质量教材、开发高水平课程、建设数字化资源、培训专业化开发队伍、完善推广机制为重点任务；目标在 3-5 年内出版完成 300 本“中文 + 职业技能”系列教材，建立多语种数字化教学资源库[2]。截至 2025 年 7 月，已有相当一部分“中文 + 职业技能”教材陆续问世，包括《职通中文》系列教材、《新丝路“中文 + 职业技能”》系列教材等。然而，现有教材及教学资源在匹配教师教学需求、满足学习者学习诉求、实现行业人才供给方面仍存在明显差距。

近年来，中国人工智能产业高度发展，截至 2025 年 7 月，中国已发布 1509 个大模型，在全球已发布的 3755 个大模型中数量位居首位；中国人工智能企业全球占比约为 15%，产业规模持续壮大[3]。同时，随着“中文 + 职业技能”教育国际化推进，面向人工智能领域的专业汉语教学需求显著增长，但现有教学资源难以满足人工智能技术传播与人才培养需求。本文以人工智能专业为例，以行业需求为导向，立足人工智能本科课程体系，提取该专业高频核心术语，进而构建“中文 + 人工智能”核心术语库，生成专门用于人工智能领域的中文词表，旨在弥补现有“中文 + 职业技能”教育中人工智能领域教学资源短板，为国际中文教师开展人工智能领域相关词汇教学提供参考，为跨境技术交流、职业技能培训提供标准化语言支撑，具有重要现实意义。

2. 语料来源

本研究参考全国十所高校开设的人工智能本科专业课程，筛选出人工智能导论、计算机视觉、机器学习、深度学习、自然语言处理等五门核心课程作为研究范围，从课程视频、教材、专业论文三类语料中进行专业术语提取。

[1]https://www.chinese.cn/uploads/file/20220125-1643091122329263.pdf

(一) 课程视频

Bilibili 平台在线公共课程资源：复旦 DISC 魏忠钰出品的《人工智能导论》课程；计算机视觉那点事出品的《计算机视觉》课程；尚硅谷出品的《机器学习》《深度学习》《自然语言处理》课程。

(二) 教材

每门课程选取两本教材，共十本，见表 1：

Table 1. Details of selected textbooks
表 1. 所选教材详情

教材	作者/主编	出版社
《人工智能导论（第五版）》	王万良	高等教育出版社
《人工智能导论》	李德毅	中国科学技术出版社
《Computer Vision: Algorithms and Applications（计算机视觉：算法与应用）》	Richard Szeliski	清华大学出版社
《2D 计算机视觉：原理、算法及应用》	章毓晋	电子工业出版社
《机器学习》	周志华	清华大学出版社
《统计学习方法（第二版）》	李航	清华大学出版社
《Deep Learning （深度学习）》	Ian Goodfellow/Yoshua Bengio/Aaron Courville	人民邮电出版社
《神经网络与深度学习》	邱锡鹏	机械工业出版社
《Speech and Language Processing （自然语言处理综论 第二版）》	Daniel Jurafsky/James H. Martin	电子工业出版社
《自然语言处理 基于预训练模型的方法》	车万翔/郭江/崔一鸣	电子工业出版社

(三) 专业论文

选取 CNKI 中国知网在线公开的(截至 2026 年 3 月)，以“人工智能”、“人工智能时代”、“人工智能技术”为主要主题；以“人工智能”、“机器学习”、“深度学习”、“神经网络”为次要主题，来源类别为北大核心和 CSSCI 的期刊论文 50 篇。

3. 提取过程

(一) 语料采集

在待处理的三类语料资源中，课程视频为在线视频；教材为图片内容的 PDF 扫描件；论文为 PDF 格式文件，三类语料均无法对其直接进行批量化编辑，首先需将全部语料转为 txt 文本格式。

1. 在线视频转 txt 文本

使用哔哩哔哩视频批量下载器 bilidown 下载在线视频后，选取开源 ASR 程序批量进行视频转文字。开源 ASR 程序是一类低成本、高可控，具有可定制性和安全性的自动语音识别工具，可将视频语音内容自动转换为可编辑、可分析的结构化文本，并允许其灵活修改程序做进一步处理，如语义理解、去除口语化冗余等，以提高视频语音转文本的准确性。

2. 图片形式 PDF 教材转 txt 文本

选取开源软件 Umi-OCR_Paddle 将图像中的文字提取出来并转化为可编辑的文本。Umi-OCR_Paddle 支持批量识图、文件 OCR 扫描等多种场景，具有识别精准度高、批量处理高效、支持多语种识别等优势，还能通过导入专业术语词典，进一步优化特定领域词汇的识别精度与准确度，且操作便捷，可离线使用

保障数据隐私。

3. PDF 格式论文转 txt 文本

使用我的 ABC 软件工具箱中的集成功能进行文本格式转换，与其他在线转换工具相比，其转换速度更快，且不受文件大小限制。转换后保留原文本的逻辑结构，可确保后续的文本处理能准确反应原始文档中术语出现的场景。

以上三类语料经过文本转换后，我们获得了 3,078,484 字符的原始语料，并分别归入 6 个子语料库，如表 2。

Table 2. Character count of raw corpus
表 2. 原始语料字符数

人工智能导论	计算机视觉	机器学习	深度学习	自然语言处理	专业论文
264,090	360,021	385,977	660,851	280,144	1,127,401

(二) 语料清洗与分词

语料清洗是保证术语提取准确的前提。原始语料中包含大量冗余信息，如特殊符号、重复内容、广告文本、乱码等，如不清洗干净将会直接干扰统计结果。本研究采用 PyCharm 对语料进行清洗与分词，并加入了哈尔滨工业大学停用词表、百度停用词表、中文停用词表(中国人民大学)，四川大学机器智能实验室停用词库等四大停用词表，共 2242 个停用词以及自定义词典。PyCharm 作为一款专业的 Python 集成开发环境(IDE)，可以在语料清洗和分词等自然语言处理任务中为开发者提供便利，还可以解决通用工具对特定领域术语错误拆分的问题，同时进行自动化处理，减少主观因素，实现专业术语的精准分词。显著提升从原始语料到高质量语言处理结果的转化效率。

经清洗和分词后各个子语料库的词语数量见表 3。

Table 3. Word count of sub-corpus
表 3. 子语料库词数

人工智能导论	计算机视觉	机器学习	深度学习	自然语言处理	专业论文
141,601	156,384	252,270	467,871	206,976	843,557

(三) 术语提取标准

要在数量庞大的语料中提取符合要求的术语进入术语库，则需制定严格的提取标准。本研究以词频、专业性、跨子语料库分布三个维度作为术语提取标准，后进行人工审核，通过严格筛选后的词语才可作为人工智能行业核心术语纳入职业核心汉语词表。

1. 词频

词语的出现频次是考量一个词是否具有教学价值的最直观的体现。Coxhead (2000)所设定的词频标准是出现频次高于 100，大约在每百万词中出现 28.6 次[4]；梁有蕾、梁焱(2025)经多次反复统计后发现，当最小词频是 54 时，才可以保证有更多的专业词汇被纳入[5]。本研究使用 AntConc 统计清洗、分词后的语料中人工智能术语的出现频率，筛选核心术语。清洗、分词后的总语料库容量约为 206 万词，参考已有研究中的词频和语料库容量的比例，并经过反复统计后，我们将 40 设定为最低频次，在总语料库中出现频次 40 次及以上的词语作为高频词保留，做进一步分析。经过筛选后，共有 2075 个词语符合词频标准。

2. 专业性

本研究旨在构建一个专业度高、行业应用性强的人工智能核心术语库，可用作“中文 + 人工智能”

教育教学专用词表，那么词表的构建必须考虑到教学对象的实际情况，接受“中文 + 职业技能”教育的学习者中文水平大多起步于 HSK 初级及以上，而达到 HSK 高级的学习者同汉语零基础学习者一样，在“中文+”教育模式的受众群体中都属于极少数情况。鉴于以上情况，本研究选定教育部、国家语言文字工作委员会于 2021 年 3 月联合发布、7 月 1 日正式实施的《国际中文教育中文水平等级标准》（GF0025-2021）[6]中一到六级的 5456 个词汇作为通用中文教学词汇依据。将 5456 个 HSK 词汇与经过词频筛选的 2075 个高频词汇一同导入 Excel，去掉重复值后，得到 1080 个人工智能术语。由于高频词汇没有与 HSK 高级词汇进行对比，所得的 1080 个词语难免会包含非人工智能领域词语，但本研究从“中文+”词汇教学角度考量，认为一些高级词语在此教学模式中具有一定的教学意义，因此保留此类词语在专业术语词表中。

3. 跨子语料库分布

跨子语料库的分布是指同一词语在多个子语料库中同时出现，词语跨子语料库的数量越多，该词的分布就越广，更能说明该词的行业通用性。一般来说，词语跨子语料库数量至少要在总语料库数量的一半以上，这样提取出来的词语会更具代表性。在本研究中，五门核心课程和专业论文的文本构成 6 个子语料库，我们将同时分布于 4 个以上子语料库的词语作为“中文 + 人工智能”重点教学词汇纳入到职业核心汉语词表中，共有 624 个。

4. 人工审核

使用计算机按照词频、专业性和跨子语料库分布三个提取标准自动筛选出术语后，为了让术语库更准确、更具有“中文 + 人工智能”教学实用性，我们专门进行了人工审核。这一步主要是解决在语料处理环节由于技术原因产生的分词拆分有误，专业术语对比不全面等问题。同时，要对词表中的非人工智能专业汉语词汇进行处理。

经计算机提取出的词表中存在一些短语被保留的现象。比如“一种”、“三个”、“各类”等数量短语；“更大”、“更深”、“广泛应用”、“因果关系”等偏正短语；“所说”、“所示”等“所”字短语等等，都需要剔除。而对一些人工智能行业常见的组合形式，如“训练样本”、“随机变量”、“概率分布”等，则作为一个整体保留。一些专业相关概念或公式的提出者，如“马尔可夫”、“高斯”、“贝叶斯”等予以保留。此外，还有一些“词根 + 词缀”形式的具有人工智能特色的附加词，如“生成式”、“结构化”、“归一化”、“解释性”等全部保留。

在对高频词汇进行专业度分析时，我们选用《国际中文教育中文水平等级标准》中一到六级词汇作为日常通用词依据，来与满足词频条件的词语进行对比，而《国际中文教育中文水平等级标准》官方文件为带有水印的 pdf 文件，在 pdf 转 txt 文本的过程中会因水印干扰而降低 txt 文本内容的纯净度，导致与之对比后的高频术语表中还掺有个别 HSK 一到六级通用词汇，影响专业术语筛选的准确性，此类词汇需经人工一一核对后删除。

词表中还存在两类内容：一是人工智能行业字母词，此类词汇虽具有全球通用性，但本研究重点聚焦汉语词表的提取与构建，因此不予保留。二是一些方位词、时间名词和具体的地名、国名，本研究认为无需保留。

经过人工逐一审核后，我们最终保留了 365 个词语收录进“中文 + 人工智能”专用核心汉语词表。

5. 研究局限

本研究构建的人工智能核心汉语词表虽然经过人工审核优化，但仍存在四个关键局限需要正视。

第一个局限是语料来源的实践代表性不够。术语提取完全依赖课程视频、教材和学术文献，缺少真

实行业应用场景下的实践性语料，如技术文档、项目报告、开发日志等，未能充分覆盖人工智能行业实际应用中的真实用语。这就无法确保最终的汉语词表是否可以满足行业实践要求。

其次，术语提取标准的多维性有待加强。当前依靠词频统计、词表对比和跨子语料库分布统计三个维度作为术语提取标准是不全面的：词频考察的是词语是否被高频使用；与 HSK 词汇表进行对比考察的是词语是否具有专业性；跨子语料库分布考察的是词语是否具有行业通用性。若要严格构建一个符合行业发展的专业词表，或许还应添加词语离散度和动态时效性两个标准。词语离散度考察词语在语料库中的分布是否均匀，是否能反映其行业普适性；通过对动态时效性的研究可以考察词语的活跃度，是否是过时术语等。

第三，人工审核由单人完成，存在认知局限。研究者虽然通过制定一定的筛选标准，对词语进行逐一审核，但由于研究者是非人工智能行业相关人员，在面对模糊性术语时，会受到个人经验的局限，无法准确把握一些词语是否应该收录。且由单人审核完成的词表未经过他人校对，难以保证词表中是否会存在其他错误或遗漏，需要通过交叉审核来平衡解决，避免这类问题。

尤为重要的是，本词表缺少行业专家验证。目前术语仅从学术文本角度进行筛选，没有邀请人工智能从业者参与核验。很多词语可能存在非从业人员不了解的实践内涵，那么此类词语在筛选的过程中就会因未被察觉而被舍弃，这种学术定义与行业实操的差距，可能会降低术语库的实践指导价值。

6. 结语

本研究以“中文 + 职业技能”教育的实际需求为导向，以人工智能专业的本科核心课程为基础，收集了课程视频、教材和学术论文三类语料。经过语料转换、清洗分词等预处理后，通过词频、专业性、跨子语料库分布三个维度筛选术语，再经人工审核进行优化，最终形成了包含 365 个词语的“中文 + 人工智能”职业核心汉语词表，为人工智能领域的中文教学提供专业词汇参考，且在一定程度上弥补了现有“中文 + 职业技能”教育中人工智能领域教学资源短板，初步实现了人工智能专业词汇与通用汉语教学词汇的区分。后续研究需针对行业实践性语料不足、提取标准不够多元、单人审核局限、缺少专家验证等问题进一步完善，不断提升词表的实践价值，为“中文 + 人工智能”教育资源建设提供更有力的支持。

参考文献

[1] https://kns.cnki.net/kcms2/article/abstract?v=R4S5WDEiKcsrNJC1-fHFOAnjXLCpYFHCa8cCAUS3YYCIy10EDkEyw9iwRmztkKm-f_HpB-7FEBMQ7Xm-Ke9spi4E5sg8024fbjrap0eHURk9IhibSI4Ss5X6NteoWrr7-_L9kF_z8NVy8pAZzgLPh9ri-vlw6_gCjDbCJA5QYeVS_K1mFKO8m9g==&uniplatform=NZKPT&language=CHS.

[2] 中外语言交流合作中心. “中文 + 职业技能”教学资源建设行动计划(2021-2025 年) [EB/OL]. https://www.chinese.cn/uploads/file/20220125-1643091122329263.pdf, 2026-04-30.

[3] 龚雯, 宋晨. 我国大模型数量超 1500 个[EB/OL]. http://www.news.cn/20250727/cbe78383b65a48499666d6009e73bdbf/c.html, 2025-07-27.

[4] Coxhead, A. (2000) A New Academic Word List. *TESOL Quarterly*, **34**, 213-238. https://doi.org/10.2307/3587951

[5] 梁有蕾, 梁焱. “中文 + 职业”背景下的跨境电商职业中文核心词表构建[C]//河南省民办教育协会. 2025 年高等教育发展论坛论文集(下册). 乌鲁木齐: 新疆大学国际文化交流学院, 2025: 12-16.

[6] 中华人民共和国教育部, 国家语言文字工作委员会. 国际中文教育中文水平等级标准[EB/OL]. https://hudong.moe.gov.cn/jyb_sjzl/ziliao/A19/202111/t20211118_580755.html, 2021-03-04.