AI 技术支撑下非洲区域国别研究英语报刊语料库的构建与应用

陈蕾
西北大学外国语学院，陕西 西安 710127
DOI:10.61369/EIR.2026100035

摘　　要： 在全球化不断深化的背景下，非洲作为资源丰富、文明多元的重要大陆，对于认识全球政治、经济和文化格局具有重要意义。随着非洲国家更积极参与全球治理和“一带一路”合作，中非关系也受到广泛关注。深入研究非洲各国的政治、经济和文化状况，不仅有助于深化学界对非洲的认识，也能够为中非交流与国际合作提供理论支持。本文以非洲英语报刊为对象，在梳理相关研究的基础上，探讨 AI 技术支撑下区域国别研究语料库的构建方法、质量控制和应用价值。研究认为，人工智能可提高新闻语料采集和处理效率，并通过词性标注、命名实体识别、主题建模和情感分析，为非洲政治、经济和文化研究提供新资料与新视角；但新闻语料仍需与统计数据、政府文件和国际组织报告交叉验证，研究结论不能简单由机器输出替代。建设面向非洲区域国别研究的英语新闻语料库，有助于从丰富新闻文本中观察非洲国家发展变化。

关 键 词： 人工智能；自然语言处理；非洲区域国别研究；英语报刊；语料库

Construction and Application of English Newspaper Corpus for African Regional and Country Studies Supported by AI Technology

Chen Lei
School of Foreign Languages, Northwest University, Xi'an, Shaanxi 710127

Abstract: With the continuous deepening of globalization, Africa, a vital continent endowed with abundant resources and diverse civilizations, plays a significant role in understanding the global political, economic and cultural pattern. As African countries engage more proactively in global governance and Belt and Road cooperation, China-Africa relations have garnered extensive attention. In-depth exploration of the political, economic and cultural conditions of African countries can not only enrich academic research on Africa, but also provide theoretical support for China-Africa exchanges and international cooperation. Taking English newspapers and periodicals concerning Africa as the research subject, this paper explores the construction methods, quality control and application value of the corpus for regional and country studies based on a review of existing research, with the support of AI technology. The research indicates that artificial intelligence can enhance the efficiency of news corpus collection and processing. Through part-of-speech tagging, named entity recognition, topic modeling and sentiment analysis, it offers novel materials and perspectives for studies on African politics, economy and culture. However, news corpus data must be cross-validated with statistical figures, official documents and reports issued by international organizations, and research findings cannot be directly substituted by machine-generated results. The establishment of an English news corpus tailored to African regional and country studies facilitates the observation of developmental changes in African countries from massive news texts.

Keywords: artificial intelligence; natural language processing; African regional and country studies; English newspapers and periodicals; corpus

引言

人工智能技术的发展，使大规模文本的自动采集、清洗、分类和分析成为可能。在语言学及相关人文社会科学研究中，AI 已用于文本挖掘、情感分析和主题建模等任务，从而提高文本处理效率；在媒体传播领域，AI 也开始进入国际精准传播实践[4]。区域国别研究需

作者简介：
陈蕾（1980.10—），女，陕西岐山人，汉族，毕业于西北大学，英语语言文学硕士，就职于西北大学外国语学院，讲师，研究方向：大学英语教学理论研究和实践改革、区域国别研究、外文报刊阅读、英美文学。

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

要长期积累政治、经济、社会和文化资料，非洲英语报刊时效性强、信息丰富，是观察非洲国家现实议题的重要来源。已有研究以非洲国家英文媒体为对象分析“中非关系”报道，发现埃及更关注双边政治经济交往，埃塞俄比亚更关注劳工和就业，赞比亚更关心农业和农村发展[1]。因此，建设面向非洲区域国别研究的英语新闻语料库，既能为研究提供系统信息支持，也可借助 AI 从海量文本中发现潜在主题和话语特征。

非洲媒体也是观察中非相互认知的重要窗口。有关中国减贫理念的研究表明，国际理念进入非洲媒体语境后会产生本土化表达[3]；“人类命运共同体”话语研究也说明，英文媒体文本可用来考察国际传播中的意义建构[2]。因此，建设该语料库既是非洲区域国别研究数字化发展的基础工作，也能够为中非相互理解和国际传播研究提供新的文本资料。

相关研究也表明，英文媒体文本可以成为考察国际话语传播和本土化建构的重要材料。

一、AI 技术支撑下非洲区域国别研究英语报刊语料库构建

（一）语料库构建的目标与意义

建设该语料库，旨在为相关研究提供全面、连续、可检索的英文新闻信息。非洲国家数量众多，不同国家在政治制度、经济结构和文化传统等方面存在较大差异，系统整理分散于不同媒体的新闻，可为跨国家、跨地区和跨时期研究提供数据基础。新闻语料具有动态性，能反映特定时期政治议题、经济问题和社会热点；通过对新闻文本进行长期积累和比较，可以更加直观地观察相关议题的变化过程。非洲英语媒体对中非关系的报道，也为研究非洲国家对华认知提供了直接案例[1]。同时，媒体报道还涉及媒介感知、理解和话语翻译[3]，因此语料库可从非洲本土媒体话语中考察其对华认知和中非关系理解。

特别是在中非关系研究中，本土媒体话语能够补充官方文件和统计资料之外的社会认知维度。

（二）语料选取原则与范围

语料选取应遵循权威性、时效性和多样性原则。权威性上，优先选择影响力较高、出版较稳定的非洲英语媒体，如《非洲报告》（The Africa Report）、《埃及每日新闻》（Daily News Egypt）等，并按媒体性质分类。已有研究基于自建非洲国家英文媒体数据库，用文本挖掘分析9国报道，为国别覆盖、检索设计和分析流程提供方法参照[1]。时效性上，以近10年为主，围绕基础设施、贸易、投资、能源、农业和人文交流等议题收集报道。多样性上，兼顾埃及、埃塞俄比亚、赞比亚等重点国家和其他国家，覆盖综合日报、周刊及经济、商业、文化媒体；同时通过抽样和分类避免少数国家语料过多影响统计结果。同时，应根据研究目的对媒体性质进行分类，以保证不同类型媒体之间具有可比性。

表1 非洲英语报刊语料库建设的方法设计

维度	设计建议	说明
研究对象	非洲英语报刊公开新闻	以国家、媒体、时间为抽样单位
国别范围	埃及、埃塞俄比亚、赞比亚为重点并扩展	兼顾北非、东非和南部非洲
时间范围	近10年为主，可按议题回溯	观察议题变化和中非合作重点
主题范围	政治、经济、文化及中非关系专题	关键词检索与人工筛选结合
处理流程	采集—清洗—去重—标注—分类—复核	AI 处理与人工抽样复核结合
质量控制	记录来源、日期、作者、栏目、去重依据	便于复核、引用和更新

注：表1为方法设计示意，具体比例根据实际媒体资源确定。

（三）基于 AI 技术的语料采集与清洗

语料采集和清洗是语料库建设的基础。传统人工搜集需要逐篇查找整理，工作量大且难以持续更新；利用网络爬取、自动检索和自然语言处理技术，可以提高采集效率。采集时可依据目标网站页面结构和检索规则设置程序，获取标题、正文、日期、作者、栏目等信息并统一入库。专题研究可建立关键词表，以 cooperation、development、infrastructure、trade、investment 等词检索，并结合国家、人物和机构名称扩展范围。原始网页常含广告、菜单、HTML 标签和推荐信息，需用 HTML 解析、正则表达式清洗，再进行句子切分、词语规范化和相似度去重。AI 清洗可能误删正文或识别错误，因此宜采用“AI 自动处理 + 人工抽样复核”。本文聚焦这些技术在语料库建设中的流程位置和应用方式，不展开算法细节。

AI 自动清洗虽然能够提高效率，但也可能出现误删或正文识别错误，因此需要在关键环节保留人工复核。

入库字段应保持一致，便于后续检索、统计和引用。

（四）语料标注与分类

清洗后的语料需标注分类。词性标注（Part-of-Speech Tagging）可为句法和词汇研究提供基础；命名实体识别（NER）可提取人名、地名、国家和机构，用于分析政治人物、国际组织和国家关系。主题建模（Topic Modeling）可依据词汇共现和语义关系划分主题，如政治、经济、文化及外交关系、贸易往来、基础设施、农业发展、文化交流等子主题。“人类命运共同体”相关研究说明，新闻语料能呈现同一话语在不同媒介语境中的意义差异[2]。情感分析可辅助观察报道倾向，但 AI 结果必须结合语境人工判断，不能把文本情感等同于国家或社会整体态度[3]。

这种多层次分类有助于研究者快速定位感兴趣内容，也能揭示语料中的潜在模式和变化趋势。

Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License

二、AI 技术支撑下英语报刊语料库的应用

（一）在政治研究中的应用

语料库可整合不同国家英语报刊，系统观察政治议题、外交互动和全球治理参与。文本挖掘能发现特定国家或时期热点，关键词共现可揭示政治主体关联。孔蕾、秦洪武的研究表明，不同非洲国家媒体在“中非关系”报道中议题侧重不同[1]；据此比较埃及、埃塞俄比亚、赞比亚等国新闻，可进一步观察政治经济交往、劳工就业、农业农村发展等议题差异。“人类命运共同体”话语研究提示，还可分析相关理念在非洲媒体中的出现语境和表达方式[2]。此外，AI 可将关键词、主题和人物关系可视化，形成政治舆论图谱，帮助研究者掌握海量文本中的议题变化。因此，英语报刊语料库能够成为非洲政治研究的重要辅助工具。

通过统计政治关键词出现频率并结合主题建模，可以分析不同阶段的政治关注重点。

政治决策支持方面，历史新闻与现势新闻的横向比较，有助于观察政治环境变化并识别潜在风险或发展机遇。

（二）在经济研究中的应用

语料库可为经济研究提供统计数据之外的补充。围绕贸易、投资、基础设施、能源、农业和产业发展收集报道，可观察经济议题变化和社会关注程度。中非关系报道中，基础设施和经济合作是重要方面[1]；主题聚类、关键词分析和语义识别可提取与非洲经济发展相关的信息。特别是在中非经贸合作研究中，交通、电力、通信等基础设施项目受到较多关注，通过语料库可分析非洲媒体如何报道这些项目以及不同国家媒体之间的报道差异。对企业而言，目标国家新闻可反映产业政策、投资环境、市场需求和监管变化；结合赞比亚媒体更关注农业和农村发展的发现[1]，相关报道可为产业研究和投资决策提供线索。但新闻语料不能替代正式统计和政策文件，更适合用于发现问题、追踪线索和形成假设。

这些信息能够帮助研究者进一步认识中非经贸合作的重点领域，也能为观察非洲市场变化提供参考。

对政策和行业新闻持续监测，也有助于企业识别潜在风险。

因此，语料库不仅具有学术价值，也可为中国企业了解非洲市场提供辅助信息。

（三）在文化研究中的应用

语料库可为非洲文化研究和中非文化交流提供新资料。通过文化新闻，可了解非洲国家文化传统、社会价值和发展趋势；在中国相关报道中，文化、社会发展和理念传播都可成为观察中非文化认知的切入点。例如，春节、中医、电影、音乐等符号如何被选择和呈现，可以反映非洲媒体对中国文化的关注重点。黄敏、廖威从媒介感知和话语翻译角度分析中国减贫理念在非洲媒体中的传播，说明研究者不能只统计文化词频，还要分析具体语境中的意义[3]。情感分析可初步判断报道倾向，再结合文本细读解释积极或消极评价的原因。这对改进中非文化交流策略具有参考价值。

通过语料库分析，可以考察非洲媒体如何选择和呈现中国文化，以及不同国家对中国文化和相关理念的关注重点。

从跨文化传播角度看，文化传播并不是简单的信息传递，而是一个本土化理解和再阐释的过程。

（四）问题与优化方向

目前需解决语料分布不均、媒体立场差异和 AI 语义误判等问题。不同非洲国家在媒体发展水平、英语使用情况和数字资源方面存在差异，容易造成语料分布不均；媒体报道具有自身立场和受众特征，不能简单等同于国家或社会整体观点。未来应扩大国家和媒体覆盖，建立科学抽样分类机制；将媒体话语与政府文件、国际组织报告和统计数据交叉验证；对隐喻、反讽和复杂政治语境加强人工复核；同时重视新闻文本版权、规范使用和数据引用机制。

在此基础上，可以进一步利用大语言模型、知识图谱和可视化技术，提高语料库的智能检索和分析能力。

三、结束语

AI 技术支撑下的非洲英语报刊语料库，为认识非洲提供了新数据资源和方法。已有关于非洲英文媒体、中非关系报道和非洲媒体话语的研究，证明了新闻文本作为区域国别和国际传播资料的价值[1-3]；AI 也为媒体信息处理和国际传播提供技术条件[4]。语料库可用于政治议题观察、经济信息追踪和文化认知分析：在政治研究中，可通过文本挖掘、关键词分析和命名实体识别观察政治议题及国际关系；在经济研究中，可利用新闻文本了解贸易、投资、基础设施和产业发展等情况；在文化研究中，则可考察非洲媒体对中国文化以及中非文化交流的认知和评价。但它不能替代研究者判断。未来应扩大语料来源，改善国别平衡，提高语义分析准确性，并加强语料库与区域国别知识体系结合。

在中非关系不断深化以及全球治理格局持续变化的背景下，进一步建设开放、动态和智能化的非洲区域国别研究语料库，对于深化对非洲的认识、促进中非交流互鉴以及推动区域国别研究数字化发展具有积极意义。

因此，该类语料库更适合作为区域国别研究的基础资料平台，而不是替代田野调查、统计分析和政策文本的单一证据。其价值在于把分散新闻文本转化为可检索、可比较、可复核的研究资料，并在此基础上形成持续更新和共享使用的建设机制。

参考文献

[1] 孔蕾，秦洪武 . 非洲国家英文媒体 " 中非关系 " 报道分析 [J]. 北京第二外国语学院学报 ,2022,44(6):51-67.

[2] 程浩，张云笛，黄瑚 . 超越差异与区隔 :" 人类命运共同体 " 话语的他塑性建构研究 [J]. 当代传播 ,2023(3):14-20.

[3] 黄敏，廖威 . 媒介感知与话语翻译 : 非洲媒体有关中国减贫理念报道的研究 [J]. 现代传播 (中国传媒大学学报),2022,44(12):79-88.

[4] 王希 . 主流媒体基于 AI 技术的国际精准传播研究 [J]. 传媒 ,2023(18):51-52+54.

 Copyright © This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License