Modern Linguistics 现代语言学, 2026, 14(5), 805-812
Published Online May 2026 in Hans. https://www.hanspub.org/journal/ml
https://doi.org/10.12677/ml.2026.145462

Hans汉斯

基于Python的能源新闻英语语料库建设与应用

邓　淼，白嘉硕

中国矿业大学(北京)文法学院外语系，北京

收稿日期：2026年4月3日；录用日期：2026年5月13日；发布日期：2026年5月27日

摘　要

全球能源转型、低碳目标及能源安全相关的公共话语，在近年来的国际新闻媒体上热度不减。本文利用Python技术自主构建能源新闻英语语料库，定量结合定性地分析了能源新闻的词汇分布及其高频词搭配的规律性特点。研究结果初步展现了该语料库具有较好的词汇覆盖率，其高频率短语组合也反映出宏观上聚焦绿色发展与多边合作的话语模式是当前主流新闻报道的典型特征；同时本文探讨了将该领域语料库应用于行业趋势预判、政策解读以及商务英语翻译等领域的可行性与应用前景，以期为电力企业预测分析及人才培训提供案例支持与理论参考。

关键词

能源，新闻语料库，Python，话语分析，专门用途英语

Construction and Application of an English Corpus of Energy News Based on Python

Miao Deng, Jiashuo Bai

Department of Foreign Languages, School of Humanities and Law, China University of Mining and Technology (Beijing), Beijing

Received: April 3, 2026; accepted: May 13, 2026; published: May 27, 2026

Abstract

Public discourse on energy transition, low-carbon goals, and energy security has remained prominent in international news media in recent years. Using Python, this study independently built an English corpus of energy news and conducted both quantitative and qualitative analyses of the lexical distribution and high-frequency collocational patterns in the corpus. The results demonstrate the corpus's solid lexical coverage, indicating that a discourse pattern characterized by green development and multilateral cooperation serves as a typical feature in current mainstream news reporting. In addition,

文章引用: 邓淼, 白嘉硕. 基于 Python 的能源新闻英语语料库建设与应用[J]. 现代语言学, 2026, 14(5): 805-812.
DOI: 10.12677/ml.2026.145462

this paper explores the potential applications of the corpus in industry forecasting, policy interpretation, and Business English translation, providing theoretical references and case support for predictive analysis and talent training in power enterprises.

Keywords

Energy, News Corpus, Python, Discourse Analysis, ESP

1. 引言

本研究旨在构建并利用能源新闻英语语料库，以实现三方面核心目标：其一，从语料学视角揭示能源新闻的词汇、搭配与构式特征[1] [2]，为能源话语研究提供实证依据；其二，为能源英语教学提供基于真实语料的教材资源与教学设计方案[3] [4]；其三，探索语料库在产业趋势识别与政策话语研判中的可操作化应用路径。理论层面，本研究将验证搭配/构式理论与语域模型[5]在能源新闻话语领域的适用性；实践层面，研究成果可直接服务于课堂教学、行业预判与媒体传播分析，为能源行业预判与人才培养提供实证支撑。

近年来，随着全球气候变化议题的升温，能源领域的话语权建构逐渐成为学界关注的焦点。然而，目前针对专门用途英语(ESP)的能源类自建语料库相对较少，且大多依赖于现成的通用语料库，难以精准捕捉国际能源市场瞬息万变的专业词汇与话语策略。尽管目前国内学界已在学术英语[6]、地方新闻[7]、材料[8]、化工[9]及财经[10]等特定领域开展了语料库建设的积极尝试，这为本研究提供了宝贵的方法论借鉴，但针对国际能源动态新闻的专属语料库仍存在空白。因此，自主建设一个时效性强、专业覆盖面广的能源新闻英语语料库具有重要的现实意义。它不仅能够帮助研究者追踪国际能源合作机制的演变，还能为理解复杂的国际能源政策提供客观的数据支撑。

2. 语料库建库流程简述

(注：为突出研究重点与话语分析的深度，本节仅对建库流程进行宏观概述，关于系统架构、爬虫策略、反爬机制及数据预处理等详细技术细节，请参见本文附录部分。)

本文所建为中型能源类英语新闻语料库，语料来源涵盖《中国日报》、路透社能源频道、Energy Voice、OilPrice 等主流中英文新闻网站的英文报道；其选材广泛涉及新能源、碳中和、石油天然气、电力产业、国际能源交流等方面。项目严格遵循“采集－存储－加工－应用”四阶段建设模式[11]。

在采集阶段，系统基于 Python 的任务调度与分布式爬虫架构，实现对目标网站的自动化抓取；在存储阶段，采用 Excel 与 TXT 双格式并行策略，确保元数据(如标题、来源、时间)与正文数据的结构化管理；在加工阶段，利用 FragmentAnt 与 TreeTagger 等自然语言处理(NLP)工具进行深度文本清洗、分词、词元化及词性标注[6]，从而将原始非结构化文本转化为可供机器检索与分析的标准化语料库；最终在应用阶段，依托可视化工具与统计软件，深度挖掘语料背后的语言规律与行业态势。

3. 能源新闻英语语料库的结果与讨论(核心应用深度分析)

本研究的核心价值在于从海量数据中提炼语言规律与话语特征。下文将依托已建成的语料库，通过

词频统计、搭配网络及语境共现等路径，对能源新闻的深层特征进行详尽的量化与质性分析，并探讨其在实践中的具体应用。

3.1. 语料库基本统计特征与核心词汇分布

在完成语料库建设后，本研究首先通过计算机检索工具(如 AntConc)对整体语料进行了宏观的描述性统计。统计结果显示，本语料库总形符数(Token)为 2510，总类符数(Type)为 1115，根据类符/形符比计算公式 TTR = Type/Token，本语料库的 TTR 约为 0.4442，标准化类符形符比(STTR)维持在 46.5%左右。这一数据显著高于一般性日常英语语料库，客观数据直接证明了能源新闻文本具有极高的词汇密度与多元化的专业术语特征，阅读和理解此类文本需要较高的专业背景知识与词汇储备。

为了更直观地反映语料核心主题，本研究在原始频次基础上计算了每万词标准化频率(公式：单词原始频次/语料总形符数\times 10,000)。通过去除冠词、连词、代词以及介词等功能词后，本文提取了语料库中出现频率最高的实义词。排名分布如下表 1 所示。

Table 1. Top-frequency lexical words in the corpus
表 1. 语料库中出现频率最高的实义词

排名	单词	原始频次	每万词标准化频率
1	China	52	207.17
2	ener-gy	48	191.24
3	global	22	87.65
4	OPEC	18	71.71
5	economic	16	63.75
6	cooperation	15	59.76
7	renewable	14	55.78
8	transition	13	51.79
9	security	12	47.81
10	investment	11	43.82

根据齐夫定律(Zipf's Law) [12] [13]，高频关键词不仅揭示了本语境中密集的专业性词汇特征，而且大致描绘出了现阶段能源报道的叙事框架：

1) 核心议题：以 energy 和 transition 为主轴，表明“能源转型”是当前国际媒体最核心的关注点。

2) 空间与格局：global 和 China 的高频出现，反映了能源问题的全球化属性，以及中国在全球能源治理与绿色转型中扮演的日益重要的参与者角色。

3) 行动策略：cooperation 与 investment 强调了解决能源问题必须依赖跨国合作与大量资本注入。

4) 风险考量：security 与 OPEC 的共现，揭示了传统化石能源市场依然存在多边供需博弈与供应安全焦虑。

这一高频词分布完全符合英语新闻客观、严谨的语体特点，其呈现出的聚焦区域协同与绿色发展的话语模式，反映了当前主流新闻机构的报道重心，为后续的话语分析提供了坚实的数据支撑。

3.2. 关键节点词的搭配与语义韵分析

基于弗斯(Firth)的“词汇相伴”理论[14]，一个词的意义在很大程度上由其结伴而行的词汇决定。为

了深挖高频词背后的态度导向，本文选取了 transition (转型)和 security (安全)两个核心节点词，进行了深入的搭配分析(Collocation Analysis)。

(1) 关于“transition”的搭配网络分析

在语料库中提取 transition 左右各跨距为 4 的搭配词，利用互信息(Mutual Information, MI)得分大于 3 且频数大于 5 的标准进行筛选。数据表明，transition 最显著的左侧搭配形容词为 clean (MI = 7.8)、green (MI = 7.2)、smooth (MI = 6.5)和 global (MI = 5.9)；其最显著的左侧搭配动词为 accelerate (MI = 8.1)、facilitate (MI = 6.3)和 drive (MI = 6.1)。

此外，核心词 energy 与 transition 形成的[Adj/N1 + N2]名词性构式在语料中高度稳定。通过语境共现(Concordance)提取具体句子可以发现：

Example 1: The government urged continuous efforts to accelerate the green transition of the energy sector.

Example 2: A smooth transition requires massive investment in grid infrastructure.

以上数据和用例清晰地揭示了媒体在报道“能源转型”时呈现出积极的语义韵(Semantic Prosody)。“加速(accelerate)”与“平稳(smooth)”的组合，表明主流媒体不仅强调转型的紧迫性，同时也呼吁在转型过程中保持经济与民生的稳定，这种务实的话语特征通过语料库数据得到了精确地验证。

(2) 关于“security”的话语建构分析

与 transition 的积极、进取色彩不同，节点词 security 的搭配词汇更多展现出防御性与危机意识。统计显示，security 与 energy 构成的搭配凸显了“供应保障”的核心关切。其最紧密的名词除了 energy 外，还包括 supply (供应，MI = 7.5)、food (粮食，MI = 5.8)以及 national (国家，MI = 6.2)；其高频共现动词主要为 ensure (确保，MI = 8.5)、threaten (威胁，MI = 6.9)和 safeguard (捍卫，MI = 6.1)。

这一分析结果表明，国际媒体在探讨能源问题时，始终保持着对供应链脆弱性的高度警惕。特别是在涉及全球突发公共事件和地缘冲突的报道子库中，energy security 往往与 supply chain disruption (供应链中断)、price volatility (价格波动)等负面词簇(Lexical bundles)共现。这一分析结果表明，国际媒体在探讨能源问题时，始终保持着对供应链脆弱性的高度警惕，能源安全依然是各国制定宏观政策时的底层逻辑。

3.3. 语料库在产业动向识别与政策解读中的尝试与应用探讨

基于上述扎实的语料分析方法，我们可以将高频词、典型搭配与共现关系转化为可量化的研究证据[1] [10]，从而探讨将语料库应用于议题识别与行业趋势研判的潜在价值。

从宏观上看，研究者可以通过提取特定时间段(如特定季度)的突增词频(Keywords surge)，探究某一话题的发展趋势，从而发现媒体对于能源技术迭代的报道侧重点[15] [16]。例如，若语料库监控到 hydrogen (氢能)、solid-state battery (固态电池)或 carbon capture (碳捕集)等词汇的频次在短期内显著上升，并伴随着 breakthrough (突破)、commercialization (商业化)等搭配词，行业分析师便可尝试据此预测该领域资本市场的活跃度，为投资决策提供早期信号。

配合共现图谱和关键词簇的分析结果，语料库也可辅助政策研究人员研判各利益相关主体之间的发言联系。通过对比不同国家新闻机构在报道同一气候峰会时的词汇选择偏好，可以作为跨国政策传播、舆情应对策略的辅助参考依据。这与近年来学界利用文本挖掘与语料库技术，深入分析欧洲媒体对中欧电动汽车等具体能源争端报道[22]的研究思路不谋而合，进一步印证了该方法在政策解读中的前瞻性价值。

3.4. 语料库驱动的 ESP 教学设计与人才培养

就外语教学而言，该语料库能够为专门用途英语(ESP)教学提供海量的、真实的、可检索的专业语料

[17]。基于语料库调查语言结构与实际使用的经典范式[18]以及语料库在应用语言学中的深度整合[19]表明，这种方法不仅能够克服传统教材语料陈旧、脱离真实语境的弊端，还能大幅提升电力企业复合型国际化人才的培养质量。

根据 Nation [20]以及 Tognini-Bonelli [21]的“数据驱动学习(Data-Driven Learning, DDL)”理念，教师可尝试依托本语料库设计以下三个阶段的教学活动：

1) 观察与发现阶段(Observation)：教师通过检索系统提供包含目标术语(如 capacity)的 20 个真实索引行(Concordance lines)。学生通过观察这些真实语境，自主发现 capacity 在能源英语中不仅指“能力”，更常指“装机容量”(如 solar capacity, installed capacity, generating capacity)。

2) 归纳与总结阶段(Generalization)：学生利用语料库软件的搭配(Collocates)功能，自行归纳出能源英语中表达“增加装机容量”的高频动词群(expand, boost, add, double)以及表达“削减”的动词群(cut, reduce, phase out)，从而建立专业词块意识[22]。

3) 应用与产出阶段(Application)：在商务英语翻译训练中，学生面临中译英任务“我国将大力提升风电装机容量”。借助语料库中习得的地道构式，学生可以准确地将其翻译为 China will vigorously boost its installed wind power capacity，而非中式英语的逐字翻译。

本研究的实践表明，利用该语料库建立学科词汇表及教学例证，能够有效培养学生的专业阅读能力、术语敏感度以及自主探究式的学习能力[23]。

4. 结语

综上所述，本文自主搭建的中型能源新闻英语语料库，在语料收集机制和数据加工方面实现了自动化与结构化。通过详细的描述性统计与词汇搭配分析，本文初步验证了相关语料库模型可用于精准解释复杂的能源新闻话语[24]。实证数据证明：能源主题、跨国合作、转型与投资是构成当前该领域新闻叙事的核心要素。

尽管本研究在语料挖掘方面取得了一定进展，但目前仍存在局限性：第一，由于目前的语料库规模属于中型，语料的年代跨度及特定子领域的文本数量仍有扩充空间，这在一定程度上限制了历时性规律的统计信度；第二，自然语言处理过程中的自动化标注由于专用术语库尚在完善中，仍存在偶发的词性误判，需要进一步辅以更大规模的人工抽查校验；第三，在语义逻辑与篇章结构的深层分析上，还需结合更多质性的话语分析方法。

未来的建设工作将聚焦两方面：一是拓展语料规模，对接预训练大语言模型(如 BERT)，开发具有更高准确率的半自动化修正标注程序，提升语料纯净度；二是面向一线教师与企业分析师，开发高度封装的简易图形化查询系统，进一步拓展该语料库在能源预测、舆情监控及跨文化沟通等维度的应用场景。

基金项目

中国矿业大学(北京)大学生创新训练项目“基于 python 的能源新闻英语语料库建设与应用”阶段性成果(项目编号：202508019)。

参考文献

[1] Sinclair, J. (1991) Corpus, Concordance, Collocation. Oxford University Press.

[2] Goldberg, A.E. (1995) Constructions: A Construction Grammar Approach to Argument Structure. University of Chicago Press.

[3] Hou, H. (2014) Teaching Specialized Vocabulary by Integrating a Corpus-Based Approach: Implications for ESP Course Design at the University Level. *English Language Teaching*, **7**, 26-37. https://doi.org/10.5539/elt.v7n5p26

[4] Basturkmen, H. (2010) Developing Courses in English for Specific Purposes. Palgrave Macmillan.

[5] Halliday, M.A.K. and Hasan, R. (1989) Language, Context, and Text: Aspects of Language in a Social-Semiotic Perspective. Oxford University Press.

[6] 陈修琪, 吕佳忱, 陈曼. 能源领域学术英语语料库的建设及应用[C]//外语教育与翻译发展创新研究(第八卷). 北京: 中国矿业大学, 2019: 526-528.

[7] 方耀, 高琪娟. 基于 Python 的合肥英语新闻语料库的建设与应用[J]. 合肥师范学院学报, 2022, 40(6): 49-53+100.

[8] 李秀文. 材料英语语料库的建设及应用——评《复合材料与工程专业英语》[J]. 材料保护, 2021, 54(3): 174-175.

[9] 陈峰, 黄勇, 王和私. 化工英语语料库的构建与应用前景[J]. 材料保护, 2021, 54(3): 198-199.

[10] 冯正斌, 王峰. 财经英语新闻语料库的建设构想与教学应用[J]. 外语电化教学, 2016(2): 54-58+39.

[11] McEnery, T. and Hardie, A. (2012) Corpus Linguistics: Method, Theory and Practice. Cambridge University Press. https://doi.org/10.1017/cbo9780511981395

[12] Zipf, G.K. (1949) Human Behavior and the Principle of Least Effort. Addison-Wesley.

[13] Hu, Q., Yue, M. 用齐夫定律解读教材词表——评 Xiao *et al*. (2017)《基于语料库的小学英语认识率及教材选词策略研究》(英文) [J]. 信息与电子工程前沿, 2017, 18(7): 863-867.

[14] Firth, J.R. (1957) Papers in Linguistics 1934-1951. Oxford University Press.

[15] Baker, P. (2006) Using Corpora in Discourse Analysis. Continuum. https://doi.org/10.5040/9781350933996

[16] Xu, J. (2023) A Corpus-Driven Study of the Ecological Discourse Analysis of Energy Narrative in News: The New York Times as Example. *International Journal of Linguistics*, *Literature and Translation*, **6**, 54-60. https://doi.org/10.32996/ijllt.2023.6.10.8

[17] Basturkmen, H. (2025) Core Concepts in English for Specific Purposes. Cambridge University Press. https://doi.org/10.1017/9781009376723

[18] Biber, D., Conrad, S. and Reppen, R. (1998) Corpus Linguistics: Investigating Language Structure and Use. Cambridge University Press. https://doi.org/10.1017/cbo9780511804489

[19] Hunston, S. (2002) Corpora in Applied Linguistics. Cambridge University Press. https://doi.org/10.1017/cbo9781139524773

[20] Nation, I.S.P. (2001) Learning Vocabulary in Another Language. Cambridge University Press. https://doi.org/10.1017/cbo9781139524759

[21] Tognini-Bonelli, E. (2001) Corpus Linguistics at Work. John Benjamins Publishing Company. https://doi.org/10.1075/scl.6

[22] Hyland, K. (2006) English for Academic Purposes: An Advanced Resource Book. Routledge.

[23] Gries, S.T. (2024) Collostructional Methods. In: Chapelle, C.A., Ed., *The Encyclopedia of Applied Linguistics*, Wiley, 1-6.

[24] Fu, J. and Yang, M. (2025) Integrating Corpus Linguistics and Text Mining to Analyze European Media Coverage on China-EU Electric Vehicle Dispute. *Journalism and Media*, **6**, Article No. 196. https://doi.org/10.3390/journalmedia6040196

附　录

语料库系统爬虫与文本处理技术细节详述

(说明：为保证正文学术探讨的流畅性与篇幅聚焦，特将本研究中涉及的底层 Python 开发细节、爬虫调度机制以及自然语言处理预处理技术完整移至本附录，以供工程实现参考。)

附录 1. 基于分布式调度的爬虫系统架构与防反爬策略

本系统由爬虫采集模块、语料加工模块和可视化分析模块组成，在 Python 平台下完成自动采集、更新及存储，并利用 Python 的第三方库及其跨平台性实现了多数据源语料的扩充以及实时更新等功能。首先在顶层设计方案中，数据收集层通过对语料进行有针对性的解析和多重策略匹配从而实现高效爬取。

本系统基于任务调度与分布式爬虫架构搭建，系统通过统一接口搭配自建调度核心，实现任务调度与进度实时监控，以任务链形式发布关键词爬取任务，全程监测爬取进度与状态。在具体工具的选择中，网络通讯使用 requests，HTML 解析使用 BeautifulSoup 和 lxml，数据分析使用 pandas 和 openpyxl，路径操作和日志记录使用 os 和 json。

防重与反爬策略：

为了防止多次抓取及不必要的开销，引入 URL 去重模块及日志记录，并根据网页加载时间动态控制爬虫频率，以达到对抗反爬虫策略的目的。爬虫运行过程具有统一性：首先从新闻检索页面或者特定分类主页提取词条索引，识别标题、日期、网址等内容，之后进行内容采集；多种方式抓取信息，优先按照 id 的方式查找，如果找不到则使用 class 方式查找，再者用 article 层级查找，并对找到的信息进行清洗，去除不必要的广告以及 JavaScript 脚本程序，最终获得干净整洁的文章内容。同时，考虑到网络不稳定因素，在访问过程中出现异常或者无法正确解析网页的情况，将异常记录后继续爬取，提高软件可靠性。

设计动态关键词能够保证信息抓取的全面性准确性，不造成语料漏抓及重复。为提高系统的运行效率，在传输过程中使用 session 来重用连接来节省请求耗时；将数据集中保存以避免频繁访问数据库而消耗大量 I/O 时间；利用多线程技术加快爬虫的执行速度，并根据网关返回的数据自动调节爬虫的请求频率，保证爬虫在一定速率下进行抓取而不影响被爬网站的正常运行；支持断点续爬以及爬虫任务重启等功能，可以根据采集结果文件继续未完成的爬虫工作。

为实现长期稳定无人值守运行，系统配备日志监听与健康检测机制，爬取完成后自动生成日志。异常请求自动归入重试列表，爬取间隔可自定义配置。在反爬应对上，系统模拟浏览器真实 User-agent，滚动浏览器标识，并配置完整 HTTP 请求头(accept, accept-language, referer 等)，让请求特征贴合人类正常访问行为，避免被机器检测拦截。实际测试中每一条新闻在单线程下抓取速度约为：China Daily 25~30 条/分钟；Xinhua 20~25 条/分钟。开启 5 线程多线程并发后，采集效率提升 3~4 倍。

附录 2. 数据清洗去重与 NLP 加工技术方案

在数据预处理阶段，如果新闻没有发布日期，则从 URL 中提取、HTML 文件头中提取或者正文里面提取。在构建语料库过程中，首先对重复的数据根据 URL 进行第一次去重，然后根据 Levenshtein 距离算法判断两个字符串之间的相似程度来进行第二次去重(相似度阈值设为 0.85)，最后再根据文本特征进行第三次去重。基于字符串匹配技术检测相似度较高的重合内容，最大限度保证数据源的质量，以免影响最终的统计结果。

多层文本清洗规则：

针对采集的新闻原始语料存在 HTML 标签、脚本、广告语等噪音问题，系统设计三层文本清洗规则：结构清洗剔除 HTML 标签、JavaScript 代码；格式清洗替换换行空行、去除多余空格与不可见字符；内容

清洗清理版权信息、广告语，仅保留正文及段落。经三次清洗后，语料语义更紧凑。

分词与标注：

分词阶段采用 FragmentAnt 分词工具批量处理，针对英语新闻中大量专属词汇，分词前导入专用词典，对专有名词、缩略语做预处理，避免分词错误；分词后完成词形标准化，去除多余符号与空格，再借助 TreeTagger 实现词形、词性、词元三列词性标注，为后续数据分析提供支撑。

数据存储采用 Excel + TXT 双格式并行：Excel 存储编号、标题、新闻类型、出处、时间、关键词、文本长度等结构化信息，便于利用 pandas 进行统计分析；TXT 单独存储正文，文本头以 XML 格式嵌入元信息标签，适配计算机读取。程序使用 RESTful 风格的设计，提供关键词搜索、时间限制、分类限制等功能，支持 JSON 和 CSV 格式输出。系统预留了 NLP 接口，可对接 BERT、RoBERTa 等预训练模型进行深度分析。整体开发基于模块化、松耦合的方式，极大提高了系统的可拓展性和易维护性。