HiLL PUBLISHING

Translation and Foreign Language Learning, 2025, 1(5), 906-910
https://www.hillpublisher.com/Journals/tfll/
ISSN Online: 3069-0315
ISSN Print: 3070-3077

Corpus-based LDA Topic Modeling for Interdisciplinary Language-data Research: A Case Study of News Article Analysis

Xudi Zhao, Feijing Han, Tiedong Yang*

School of Foreign Languages, Inner Mongolia University of Technology, Hohhot 010051, Inner Mongolia Autonomous Region, China.

How to cite this paper: Xudi Zhao, Feijing Han, Tiedong Yang. (2025) Corpus-based LDA Topic Modeling for Interdisciplinary Language-data Research: A Case Study of News Article Analysis. *Translation and Foreign Language Learning*, 1(5), 906-910. DOI: 10.26855/tfll.2025.12.027

Received: October 23, 2025
Accepted: November 16, 2025
Published: December 31, 2025

***Corresponding author:** Tiedong Yang, School of Foreign Languages, Inner Mongolia University of Technology, Hohhot 010051, Inner Mongolia Autonomous Region, China.

Abstract

This paper uses Latent Dirichlet Allocation (LDA) topic modeling on a corpus of 967 Word documents. It looks at how data-driven computational methods can fit into humanities research through a “Language + Data” framework. That framework means mixing linguistic analysis with computational data techniques so we can close the gap between language studies and data science. Using corpus linguistics and Natural Language Processing (NLP) techniques the study finds five semantically coherent topics inside a mixed corpus of 967 documents. Those documents include ISBN-related academic research papers on economics stock market news and a selection of general news articles. The LDA model got a coherence score of 0.72 and it identified five distinct topics. Most of the documents – 89.5% – clustered around financial content. Another 8.5% focused on academic research and the rest were split between general news topics.

Keywords

Corpus linguistics; LDA topic modeling; interdisciplinary research; language data; NLP

1. Introduction

The rapid advancement of computational tools and large-scale digital text corpora have transformed humanities and social science research. In China, the "New Liberal Arts" (新文科) initiative calls for integrating data science methodologies into traditional disciplines, creating demand for researchers who can bridge linguistic expertise and computational proficiency (Baker, 2000; Hu, 2012). The "Language + Data" interdisciplinary paradigm has emerged as a key framework for equipping language professionals with corpus construction skills, data extraction competencies, and the capacity to conduct empirical research (Hu & Yang, 2019).

This paper applies this framework to a systematic LDA topic modeling analysis of a 967-document corpus. Latent Dirichlet Allocation—the most widely adopted probabilistic topic modeling algorithm in the NLP literature (Blei, Ng, & Jordan, 2003)—is implemented via Python's scikit-learn library. The study aims: (1) to demonstrate the feasibility of corpus-based LDA topic modeling for language researchers; (2) to identify and interpret five latent topics within a mixed-domain corpus; and (3) to reflect on the broader implications of "Language + Data" integration for empirically grounded competency development in foreign language education and translation studies.

2. Literature Review

2.1 Corpus Linguistics and Language Research

Corpus linguistics provides the methodological foundation for data-driven language research. Hu (2012) argues that corpus translation studies is not merely a methodology but a new paradigm integrating linguistic and cultural approaches. Baker (2000) pioneered corpus methodologies for investigating literary translator style, showing that quantitative analysis can reveal stable individual linguistic preferences. Huang and Shi (2018) extended this approach through a corpus-based comparison of two Chinese translations of To the Lighthouse, demonstrating that macro-linguistic features illuminate stylistic differences at both textual and narrative levels. Hu and Li (2017) applied corpus methods to China's international image in translated texts, while Liu (2010) examined translator style and strategies through reporting verbs in translations of Dream of the Red Chamber. Hu and Sheng (2020) further argued that corpus-based literary translation criticism enhances the scientific rigor of translation evaluation.

2.2 Topic Modeling in Text Analytics

Blei, Ng, and Jordan (2003) proposed LDA as a generative probabilistic model representing documents as mixtures of latent topics and topics as word distributions. Griffiths and Steyvers (2004) validated LDA's efficacy in scientific text analysis, while Mimno et al. (2011) highlighted the importance of preprocessing—especially stopword removal and low-frequency term filtering—for coherence improvement. Chen, Zhou, and Yang (2018) demonstrated LDA's effectiveness in extracting market-relevant themes from financial news. For evaluation, Röder, Both, and Hinneburg (2015) showed that word co-occurrence-based coherence metrics best reflect human-judged thematic quality; Newman et al. (2010) established that coherence scores above 0.5 indicate acceptable interpretability and above 0.7 indicate strong semantic consistency.

2.3 The "Language + Data" Paradigm

The "Language + Data" paradigm foregrounds the complementarity between linguistic expertise and data science, arguing that rigorous research emerges when humanistic interpretation is grounded in large-scale empirical evidence (Hu & Yang, 2019; Kim & Zhu, 2023). Baker's (2011) collected works trace the interdisciplinary evolution of corpus-based translation studies across technologically and politically complex research contexts. Despite significant advances, a persistent gap remains between available computational tools and the methodological training offered in traditional humanities programs. The "Language + Data" orientation addresses this gap through structured engagement with corpus tools, data retrieval software (AntConc, Wordsmith, Paraconc), and quantitative presentation methods (Standardization Administration of China, 2021). The present study represents a direct application of this integrated approach.

3. Methodology

3.1 Corpus Construction

In corpus, we collected 967 Word documents in three categories: (1) ISBN-related academic papers on library science, publishing research, and book standards; (2) stock market news about financial performance, corporate earnings, and market indices; and (3) general news on social, cultural, and daily life topics. This three-part structure can allow analysis within and across categories, following the national corpus design standards (Standardization Admin- istration of China, 2021).

3.2 Preprocessing Pipeline

All documents were processed via a Python NLP pipeline. Text was extracted using python-docx, lowercased, and stripped of punctuation, numerals, and special characters. Tokenization used NLTK, followed by removal of English stopwords and domain-specific terms (e.g., "paper," "news," "isbn," "stock"). Tokens were lemmatized using WordNetLemmatizer (e.g., "trading" → "trade") to reduce lexical redundancy (Mimno et al., 2011). Words in fewer than two documents were excluded. The corpus was vectorized using CountVectorizer with 5,000 features and unigram/bigram tokenization.

3.3 LDA Model Configuration

The LDA model (scikit-learn's Latent Dirichlet Allocation) was configured with five topics (n_components = 5) based on domain knowledge and coherence optimization. Key hyperparameters: random_state = 42, max_iter = 50, batch learning method. Top 10 keywords per topic were extracted by ranking component matrix weights; document-topic assignments were determined by argmax of the topic probability distribution.

3.4 Evaluation

We checked model performance using perplexity for model fit, silhouette score (cluster separation), and coherence score (semantic interpretability). Coherence scores above 0.7 can show good semantic consistency (Newman et al., 2010; Röder et al., 2015).

4. Results and Analysis

4.1 Topic Keywords and Labeling

Table 1 presents the top 10 keywords for each of the five latent topics.

Table 1. Top 10 Keywords for Each Latent Topic

Topic	Label	Top 10 Keywords
0	Stock Market Basic Metrics & NYSE Trends	company, market, price, month, past, gas, largest, index, nyse, per
1	Corporate Financial Performance & Tech Spending	billion, revenue, expense, technology, company, area, spending, solution, impact, purpose
2	Academic Research & University Collaboration	university, research, technology, student, new, initiative, collaboration, project, development, including
3	Industrial Production & Corporate Growth	company, project, market, million, year, production, power, billion, new, growth
4	Corporate Press Releases & Financial Disclosures	year, share, price, company, source, website, release, section, press release, press

Note: Output from LDA topic modeling analysis (N = 967 documents).

We examine Topic 0 regarding stock market basic metrics and NYSE trends. Keywords include "market" and "price" together with "index". Also "nyse" appears. These can center this topic on core stock market indicators and NYSE trading activity.Topic 1 is about corporate financial performance and tech spending. Terms such as "billion" and "revenue" can be found. Also "expense" is included. These cover corporate earnings and technology sector investment. Topic 2 relates to academic research and university collaboration. Keywords include "university" and "research". Also we find "student" and "collaboration". This topic associates with ISBN academic papers subcorpus. It is worth noting that this is the only non-financial topic. Topic 3 includes industrial production and corporate growth. Terms such as "production" and "power" can define this area. Also "growth" is used. This is a cross-domain topic and it combines financial coverage with industrial elements. Topic 4 concerns corporate press releases and financial disclosures. Keywords include "press release" and "share". Then "price" also appears. These mark formal corporate financial communications.

4.2 Topic Distribution and Category Associations

Topic 0 accounts for 46.3% of documents (~448), Topic 1 for 33.0% (~319), Topic 2 for 8.5% (~82), Topic 3 for 6.4% (~62), and Topic 4 for 5.8% (~56). Topics 0, 1, and 4 consist almost exclusively of stock news; Topic 2 is dominated by ISBN academic papers; Topic 3 is a hybrid of financial and industrial general news content.

4.3 Model Evaluation

The model achieved a coherence score of 0.72, exceeding the 0.7 threshold for strong semantic consistency (Newman

et al., 2010). The perplexity score of 1,508.84 reflects the corpus's mixed-domain complexity. The silhouette score of 0.14 indicates moderate cluster separation, consistent with LDA's tendency to produce overlapping topic distributions in cross-domain corpora (Liu et al., 2017). The clear separation of academic Topic 2 from the four financial topics validates the model's discriminative capacity.

5. Discussion

5.1 Implications for "Language + Data" Research

Our results show that LDA topic modeling can be used by language researchers even if they have different computational backgrounds, and in principle they can produce meaningful results for analysis. The five topics are consistent with the document categories we defined before. This shows the pipeline can work well for the humanities researchers. Also, we find that combining quantitative corpus analysis with qualitative thematic interpretation has value, and this method is central to the corpus translation studies paradigm (Hu, 2012; Baker, 2011). Topic keywords give a objective, data-driven basis for thematic labeling. At the same time, domain knowledge is still needed to contextualize findings in the financial journalism and academic publishing discourses. So the complementarity between computational output and humanistic interpretation shows the core strength of the "Language + Data" research orientation. Also, if language scholars engage with the full research workflow from corpus construction, preprocessing, model training, evaluation and interpretation, they can develop empirically grounded mindset and the replicable methodological toolkit which is in fact demanded by the "New Liberal Arts" agenda (Hu & Yang, 2019).

5.2 Limitations and Future Directions

The corpus is heavily skewed toward stock news (~89.5%), limiting thematic diversity. Some keywords (e.g., "company," "new") appear across multiple topics, reducing distinctiveness. The five-topic configuration was determined by domain knowledge rather than automated optimization; future work should apply coherence score plotting across a range of topic counts. Corpus balance could be improved by adding academic and general news documents. Advanced methods such as BERTopic could capture more nuanced semantic relationships. The "Language + Data" framework could also be extended to parallel corpus analysis, enabling investigation of translation strategies and cross-linguistic representation patterns through integrated corpus-computational methods.

6. Conclusion

This study demonstrates the productive integration of corpus-based LDA topic modeling into language research through a 967-document mixed-domain case study. Five semantically coherent topics were extracted and interpreted, with a coherence score of 0.72 confirming strong thematic consistency. The study contributes to discussions on the role of "Language + Data" interdisciplinary methods in transforming humanities research paradigms, showing how the convergence of corpus linguistics and NLP can equip language researchers with the computational competencies required by the "New Liberal Arts" agenda. By bridging corpus linguistics, topic modeling, and translation studies, this work points toward a productive future for empirically grounded, interdisciplinary language-data research in Chinese foreign language education.

Funding

This study is supported by the Basic Research Fund for Universities Directly under the Inner Mongolia Autonomous Region Government: Research on the Cultivation Model of AI-Empowered College Students' Foreign Language Competence in Telling China's Stories from the Perspective of Mediationism (JY20250009); the 2025 General Education and Teaching Reform Project of Inner Mongolia University of Technology: Development and Practice of an Interdisciplinary "Foreign Language + Regional Culture" Curriculum Module (2025246); and the General Project of the Research Base for Forging a Strong Sense of Community for the Chinese Nation at Inner Mongolia University of Technology: Research on Northern Frontier Narratives in Western Travelogues and the Integration of Northern Frontier Culture into Foreign Language Teaching from the Perspective of Forging a Strong Sense of Community for the Chinese Nation (NGDZLJDY2504).

References

Baker, M. (2000). Towards a methodology for investigating the style of a literary translator. *Target*, *12*(2), 241-266.

Baker, M. (2011). Corpus linguistics and translation studies: Implications and applications. In M. Baker (Ed.), Critical concepts in linguistics: Translation studies (Vol. 2, pp. 73-98). Routledge.

Blei, D. M., Ng, A. Y., & Jordan, M. I. (2003). Latent Dirichlet allocation. *Journal of Machine Learning Research*, *3*, 993-1022.

Chen, D. H., Zhou, Z., & Yang, J. (2018). Financial news mining based on LDA and sentiment analysis. *Data Analysis and Knowledge Discovery*, *2*(10), 92-100.

Griffiths, T. L., & Steyvers, M. (2004). Finding scientific topics. *Proceedings of the National Academy of Sciences*, *101*(Suppl 1), 5228-5235.

Hu, K. (2012). Corpus translation studies: Connotations and implications. *Journal of Foreign Languages*, *35*(5), 59-70.

Hu, K., & Li, X. (2017). Corpus-based research on translation and China's image: Connotations and implications. *Foreign Languages Research*, (4), 70-75.

Hu, K., & Sheng, D. (2020). Corpus-based literary translation criticism: Connotations, implications and future. *Technology Enhanced Foreign Language Education*, (5), 19-24.

Hu, K., & Yang, F. (2019). Corpus-based literary studies: Connotations and implications. *Journal of Zhejiang University (Humanities and Social Sciences)*, *49*(5), 5-18.

Huang, L., & Shi, X. (2018). A corpus-based comparison of the translation styles of two Chinese versions of *To the Lighthouse*. *Journal of PLA University of Foreign Languages*, *41*(2), 11-19.

Kim, K. H., & Zhu, Y. (Eds.). (2023). Researching translation in the age of technology and global conflict: Selected works of Mona Baker. Routledge.

Liu, Z. (2010). Corpus-based study of translator's style and translation strategies: A case of reporting verbs in translations of *Dream of the Red Chamber*. *Journal of PLA University of Foreign Languages*, *33*(4), 87-92.

Liu, H., Huang, Y., Zhao, S., et al. (2017). Research on topic model optimization for Chinese short texts. *Computer Engineering & Science*, *39*(3), 550-557.

Mimno, D., Wallach, H. M., Talley, E., et al. (2011). Optimizing semantic coherence in topic models. In *Proceedings of EMNLP 2011* (pp. 262-272). Association for Computational Linguistics.

Newman, D., Lau, J. H., Grieser, K., & Baldwin, T. (2010). Automatic evaluation of topic coherence. In *Proceedings of EMNLP 2010* (pp. 100-109). Association for Computational Linguistics.

Röder, M., Both, A., & Hinneburg, A. (2015). Exploring the space of topic coherence measures. In *Proceedings of the Eighth ACM International Conference on Web Search and Data Mining* (pp. 399-408). Association for Computing Machinery.

Standardization Administration of China. (2021). General technical specifications for corpus. Standardization Administration of China.

Zhao, W. X., Jiang, J., Weng, J., et al. (2010). Comparing Twitter and traditional media using topic models. In *Proceedings of the 33rd International ACM SIGIR Conference on Research and Development in Information Retrieval* (pp. 338-345). Association for Computing Machinery.