Article

Comparative Analysis of CNPC and Saudi Aramco Annual Report Using Digital Tools

Hongkai Xu [1,*]

[1] Xi'an Shiyou University, Xi'an, China
* Correspondence: Hongkai Xu, Xi'an Shiyou University, Xi'an, China

Abstract: Annual report of a company covers several perspectives. It contains accomplishments, plights and predictions in the following year. Owing to the two companies' positions in global petroleum and gas industry, China National Petroleum Corporation and Saudi Aramco are selected. This paper explores the latest published 2024 annual report of the two companies as the selected material. The comparative research covers three aspects, including lexical, syntactic, and discourse levels. For discussion stage, retrieval results from AntConc, WordSmith Tool, and L2SCA are used to reveal the self-narrative discourse pattern. Ultimately, this paper may hope to make scholars notice the differences between domestic and international annual reports. In addition, it may help scholars to make adjustments when translating. By doing so, it may enhance the international communication.

Keywords: Company Annual report; Digital tools; Text analysis; Self-narrative discourse pattern

1. Introduction

For global petroleum landscape, national oil companies (NOCs) like China National Petroleum Corporation and Saudi Arabian Oil Company play pivotal but distinct roles. Their international roles are shaped by their unique national contexts and strategic priorities. The distinctive differences are visibly reflected in their corporate vision statements. In 2024 annual report of CNPC, it aspires "to become a world-class integrated international energy company built to last". On the contrary, 2024 annual report of Saudi Aramco, it aims "to be the world's preeminent integrated energy and chemicals company, operating in a safe, sustainable and reliable manner."

The two enterprises, however, target the global market, their visions may reveal different emphases. For exploring the nuanced differences, this paper focuses on overall comparison and partial comparison as well. For overall comparative analysis, this paper uses AntConc 4.3.1 to retrieve the quantity of modal verbs and top 20 nouns of the two annual reports. In addition to this, the paper also uses WordSmith Tool 9.0 for lexical diversity. With respect to partial comparison, this papers focuses on the messages from the Chairman and President sections. This paper uses L2SCA developed by Lu to retrieve macro and micro metrics [1]. For discourse level, this papers combines the top 20 nouns from AntConc to reveal the self-narrative discourse style of the two companies.

2. The Latest Research of NLP

Currently, literature related to Natural Language Processing (NLP) witnesses a high decrease. Therefore, this paper selects literature from the year 2025 on Web of Science database to do a visualization. Then the retrieved literature is visualized by Citespace and is shown in Figure 1. As demonstrated in Figure 1, it shows that NLP is applied in diverse fields, including environmental science, medical science, and sentiment analysis. In also covers communication studies, and linguistics.

Received: 25 March 2026
Revised: 05 May 2026
Accepted: 18 May 2026
Published: 22 May 2026

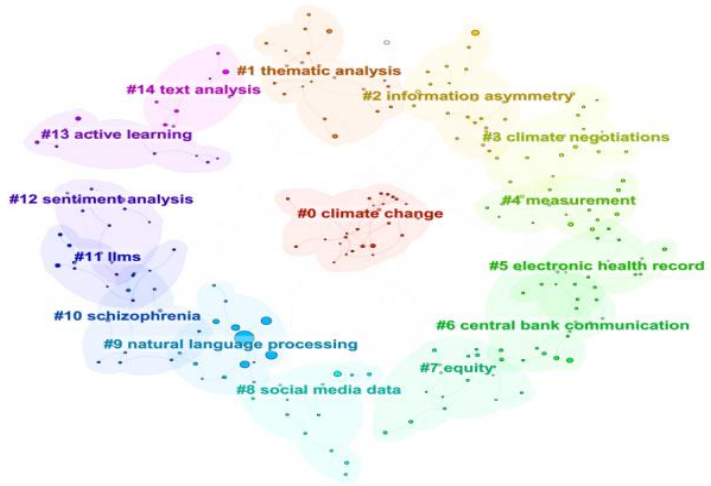

Figure 1. Visualization of Literature on Web of Science in the Year of 2025.

In sentiment analysis regard, Yang, Deng, Wei and et al conclude that for topic extraction, ChatGPT outperforms LDA and BERT models [2]. In linguistics, Yi affirms the potential of the digital humanities and argues that digital tools may enhance academic development [3]. In conclusion, these diverse applications underscore the versatility of NLP. Its applicability provides a foundation for comparative studies like the one conducted in this paper on annual reports in energy fields.

3. Technical Roadmap of This Paper

This paper employs digital tools to conduct a comparative analysis of the two annual reports from China National Petroleum Corporation and Saudi Aramco. The detail of the design is shown in Figure 2.

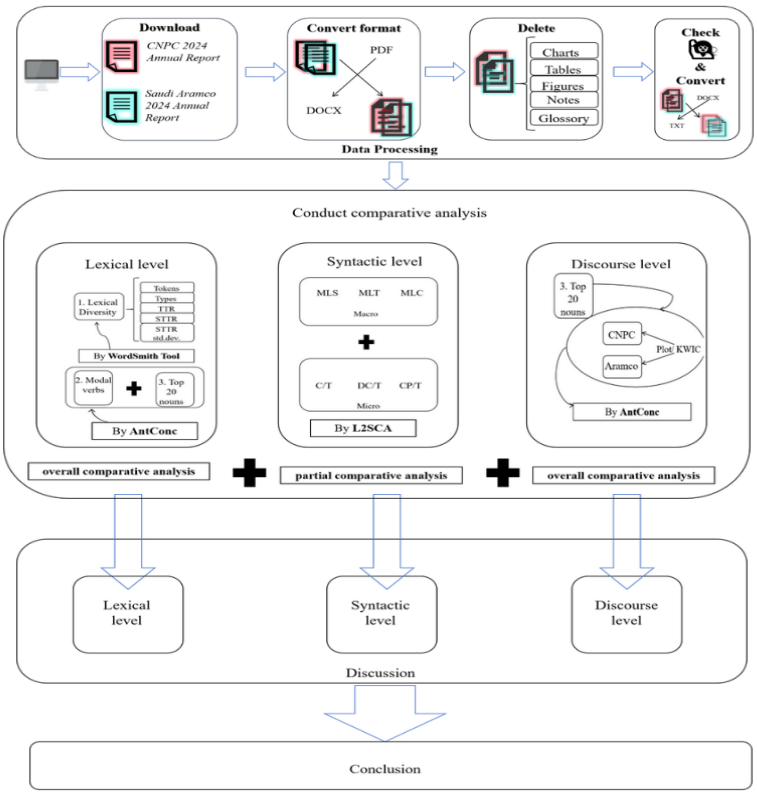

Figure 2. The Roadmap of This Paper.

As shown in Figure 2, the whole process is divided into four stages. To begin with, the conversion of the format of the two annual reports is required in that the digital tools need TXT format. Therefore, the selected material needs conversion twice, from PDF format to DOCX format, and from DOCX format to TXT format. Subsequently, charts, tables, figures and notes are deleted respectively. Through manual check, the DOCX format is finished respectively. In comparative analysis stage, it covers overall comparison and partial comparison. For lexical level, this paper applies WordSmith Tool 9.0 for macro analysis, including Tokens, Types, TTR, and STTR. It also uses AntConc 4.3.1 to conduct micro analysis, such as the quantity and frequency of the modal words and top 20 nouns. For syntactic level, this paper selects one section of the two annual reports to target macro and micro aspects. With respect to discourse level, this paper combines the top 20 nouns and the KWIC of the AntConc 4.3.1 to reveal the Self-narrative discourse pattern.

4. Comparative Analysis of the Two Annual Reports

In this section, this paper explores the differences between the 2024 CNPC annual reports and 2024 Saudi Aramco 2024 Annual Report. The comparison covers 3 aspects, including lexical, syntactic, and discourse levels. With respect to lexical level, this paper compares lexical diversity, the frequency and types of modal verbs of the selected material. In addition to this, this paper also lists the top 20 nouns to make preparation for the analysis of discourse level respectively. The comparative analysis is conducted by using WordSmith Tool 9.0 and AntConc 4.3.1. In terms of syntactic level, this paper applies L2SCA to compare the macro and micro aspects. Additionally, this paper lists mentioned countries of the two annual reports manually, and combines the top 20 keywords to reveal the two companies' narrative pattern.

4.1. Lexical Level

For lexical diversity, Token, Type, TTR and STTR in the two annual reports are compared by utilizing WordSmith Tool 9.0. The results are shown in Table 1. Mariko states that type-token ratio (TTR) is a simple measure of lexical diversity [4]. As shown in Table 1, it reveals that the number of Tokens and Types of 2024 Saudi Aramco Annual Report surpasses 2024 CNPC Annual Report. It shows that the length of the 2024 Saudi Aramco Annual Report is significantly higher than that of 2024 CNPC Annual Report, with the length being approximately 1.65 times that of the latter. In addition, the TTR and STTR scores reveal that 2024 CNPC Annual Report achieves greater lexical diversity. In addition to this, the slightly higher STTR standard deviation in 2024 Saudi Aramco Annual Report indicates somewhat greater fluctuations in vocabulary diversity in different sections within the report.

Table 1. Comparative Study on Lexical Level Using WordSmith 9.0

Comparative Analysis on Lexical Level					
Metrics	Tokens	Types	TTR (Type/Token Ratio)	STTR (Standardised TTR)	STTR Standard deviation (std.dev.)
2024 CNPC Annual Report	18, 255	2, 866	15.70 %	40.49 %	54.21
2024 Saudi Aramco Annual Report	30, 197	3, 506	11.61 %	38.37 %	57.30

For the type and frequency of modal verbs, the two cleaning DOCX texts are first tagged by using TreeTagger. The respective tagged texts is subsequently uploaded to AntConc 4.3.1. PatternBuilder is also used to identify the modal verb patterns. The regular

expression of modal verbs is "\S+_MD\s". In AntConc 4.3.1, it cannot retrieve modal verbs by using it. In comparison to the earlier version, AntConc 4.3.1 features a small adjustment for retrieving with regex. The regex can be achieved by replace "\s" with a space and the retrieved results are shown in Table 2 and Table 3.

Table 2. The Usage and Frequency of Modal Words in 2024 CNPC Annual Report

	Type	Rank	Freq	Range
1	will	1	12	1
2	can	2	1	1

Table 3. The Usage and Frequency of Modal Words in 2024 Saudi Aramco Annual Report

	Type	Rank	Freq	Range
1	may	1	85	1
2	could	2	72	2
3	will	3	49	3
4	can	4	18	4
5	should	5	7	5
6	would	6	7	6
7	must	7	3	7
8	shall	8	2	8
9	might	9	1	9

As shown in Table 2, 2024 Saudi Aramco Annual Report employs a wider range of modal verb types than 2024 CNPC Annual Report. In addition, the two reports both use low-value and medium-value modal verbs, as proposed by Halliday, in their narratives [5]. The findings, therefore, suggest that the two annual reports both show objectivity and credibility by using modal verbs. In addition, Shan argues that modal verbs have complexity in both syntactical form and semantic meaning [6]. Therefore, the frequency of the modal verbs in 2024 Saudi Aramco Annual Report may give a presentation of how modal verbs used in energy annual report. The scholars may notice the difference, and may guide them to use modal verbs accurately and appropriately when translating.

In English, words can be divided into content words and functional words [7]. Furthermore, this study employs AntConc to identify the top 20 nouns in the two annual reports respectively, as presented in Table 4 and Table 5.

Table 4. The Top 20 Nouns in 2024 CNPC Annual Report

NO.	Type	Freq
1	gas	181
2	oil	163
3	company	120
4	energy	106
5	production	96
6	development	95
7	cnpc	83
8	management	83
9	project	68
10	projects	66

11	operations	65
12	ton	58
13	carbon	57
14	exploration	48
15	capacity	46
16	year	46
17	products	45
18	industry	44
19	business	42
20	oilfield	37

Table 5. The Top 20 Nouns in 2024 Saudi Aramco Annual Report

NO.	Type	Freq
1	aramco	338
2	oil	205
3	gas	150
4	business	145
5	production	113
6	operations	112
7	company	111
8	energy	108
9	board	87
10	products	82
11	risk	74
12	cash	73
13	demand	72
14	committee	69
15	value	66
16	government	63
17	year	62
18	position	61
19	management	56
20	projects	56

As shown in Table 4 and Table 5, it holds similarities and differences. For instance, "oil", "gas", "operations", "energy", "company", "production", "year", "management", "products", "projects" and "business" are used in the two reports. From these common nouns, industry-specific terminology, including "oil," "gas," and "energy", confirms that both organizations operate within the energy sector and employ specialized discourse characteristic of this domain. In addition, nouns with suffixes such as "-ment" and "-tion" are frequently used, highlighting a formal and abstract style of professional annual reports. Third, the prominence of corporate identifiers, CNPC and Aramco, indicate a strong emphasis on branding. In addition, the term "company" often serves as a generic substitute across the two annual reports.

For differences, "oilfield", "ton", "exploration"and "capacity" may show that the lexical choice of 2024 CNPC Annual Report focus on technology, while "board", "committee", "cash" and "risk" may show that the lexical preferences of 2024 Saudi Aramco

Annual Report target finance. Second, the corporate identifiers of the two annual reports show that Saudi Aramco may have a stronger emphasis on brand promotion. These similarities and differences may help scholars better understand the domain and thematic focus of a text.

4.2. Syntactic Level

This paper selects the same section of the two annual reports, the messages from the Chairman and President, as the entire text of 2024 Saudi Aramco Annual Report could not be analyzed. This paper uses L2SCA to analyze the macro and micro aspects of sentences, and the results are shown in Figure 3 and Figure 4.

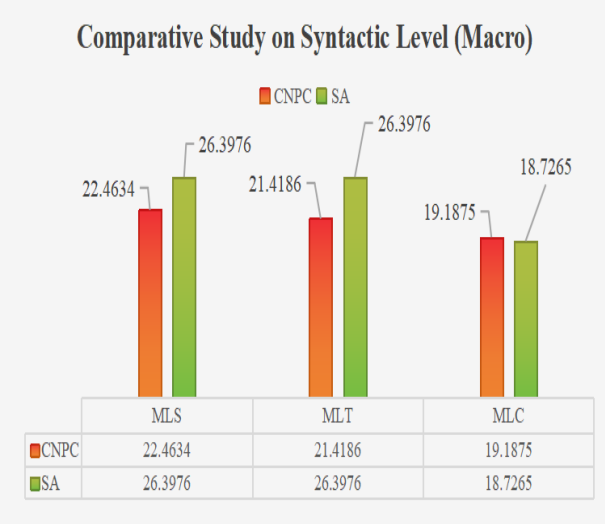

Figure 3. Comparative Study on Macro Aspects Using L2SCA

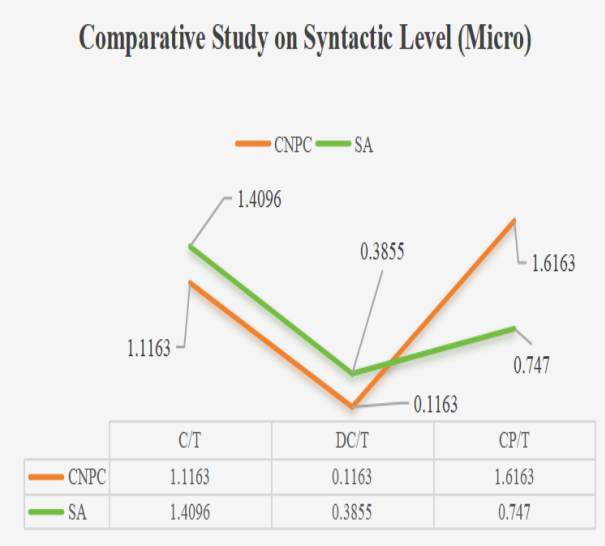

Figure 4. Comparative Study on Micro Aspects Using L2SCA

As shown in Figure 3, it shows that the Mean Length of Sentence (MLS) and Mean Length of T-unit (MLT) in 2024 Saudi Aramco Annual Report are higher than those of 2024 CNPC Annual Report. It may be concluded that the sentences in 2024 Saudi Aramco Annual Report are more complex.

Furthermore, as shown in Figure 4, metrics of Clauses per T-unit (C/T) and Dependent Clauses per T-unit (DC/T) attest to the syntactic complexity of the sentences from a micro-level perspective. CP/T ratio in 2024 CNPC Annual Report may suggest a greater reliance on employing coordinated phrases and structures to connect sentences.

Overall, the Mean Length of Clause (MLC) shows that although the sentences in 2024 Saudi Aramco Annual Report exhibit a higher number of clauses, they share a similar average clause length with those in 2024 CNPC Annual Report.

In conclusion, higher syntactic complexity is observed in 2024 Saudi Aramco Annual Report, marked by longer sentences and a higher frequency of subordination. Conversely, in 2024 CNPC Annual Report, it displays a simpler syntactic structure with a greater dependence on coordinated structures.

4.3. Discourse Level

For mentioned countries, this paper finds that 2024 CNPC Annual Report mentions 15 countries, while the 2024 Saudi Aramco Annual Report mentions 12 countries through manual labelling. The difference may reflect the two companies' partnerships. In addition to this, this paper uses KWIC of AntConc 4.3.1, with the combination of the top 20 nouns, to reveal the self-narrative style. To begin with, this paper selects the first noun in top 20 respectively. For 2024 CNPC Annual Report, "gas" ranks the first, and the KWIC retrieval results are shown in Figure 5. For 2024 Saudi Aramco Annual Report, "aramco" ranks the first in top 20 nouns, and the KWIC retrieval results are demonstrated in Figure 6.

KWIC | Plot | File View | Cluster | N-Gram | Collocate | Word | Keyword | Wordcloud | ChatAI

Total Hits: 185 **Page Size** 500 hits | 1 to 185 of 185 hits

	File	Left Context	Hit	Right Context
1	CNPC-清...	than 33 million tons of oil equivalent with steady growth. Natural	gas	production at Southwest Oil and Gas Field surpassed 40 billion
2	CNPC-清...	production from the Changbei project in Changqing Oilfield remained steady.	Gas	production at Southwest Oil and Gas Field's Chuandongbei (
3	CNPC-清...	in new energies to achieve sustained growth in oil and	gas	production and rapid advances in new energies. Domestic Exploration
4	CNPC-清...	in crude production capacity and 27.77 billion cubic meters in natural	gas	production capacity. Unconventional Hydrocarbons In line with the strategic
5	CNPC-清...	breakthroughs and important discoveries in exploration, with domestic oil and	gas	production exceeding the objectives of the Seven-Year E&
6	CNPC-清...	s South Sulige Project stood above 4 billion cubic meters. The	gas	production from the Changbei project in Changqing Oilfield remained
7	CNPC-清...	year-on-year decrease in purchases of foreign oil; natural	gas	production increased for the eighth consecutive year, exceeding 10 bcm.
8	CNPC-清...	the E&P activities in key oilfields. In 2024, oil and	gas	production of Changqing Oilfield was on the rise, standing
9	CNPC-清...	million tons, marking six consecutive years of production growth. Natural	gas	production reached 246.4 bcm, an increase of above 10 bcm for
10	CNPC-清...	billion cubic meters of tight gas in 2024. Mid-deep shale	gas	production remained stable while the development of deep shale
11	CNPC-清...	increase in Asia. " E&P spending fell as oil and	gas	production slightly increased. Due to escalating geopolitical conflicts, energy
12	CNPC-清...	stabilize oil production and boost gas output, bringing oil and	gas	production to a new high. 2024 Global Oil and Gas
13	CNPC-清...	came from Latin America and North America. Estimated global natural	gas	production was 4.39 tcm, an increase of 120 bcm or 2.8%. Refining
14	CNPC-清...	expand the capacity of gas storage facilities. In 2024, four new	gas	storage facilities were put into operation. By the end
15	CNPC-清...	storage facilities were put into operation. By the end of 2024, 23	gas	storage facilities were operational. In 2024, the total gas injection
16	CNPC-清...	rescue responses during production well leaks and fires at offshore	gas	storage facilities. In this way, employees and rescue teams
17	CNPC-清...	Storage Facilities The Company continued to expand the capacity of	gas	storage facilities. In 2024, four new gas storage facilities were
18	CNPC-清...	the successful commissioning of key projects such as the Tongluoxia	gas	storage facility and the first offshore gas storage facility
19	CNPC-清...	as the Tongluoxia gas storage facility and the first offshore	gas	storage facility Nanbao No.1. The product portfolio for green
20	CNPC-清...	of CBM in 2024 (including 2.3 billion cubic meters of deep CBM).	Gas	Storage Facilities The Company continued to expand the capacity
21	CNPC-清...	gas processing plant, the Dukouhe gas purification plant, the Wen-23	gas	storage project, and the Jidong Nanbao No.1 gas storage
22	CNPC-清...	the Wen-23 gas storage project, and the Jidong Nanbao No.1	gas	storage surface engineering project were successfully commissioned as planned.

Figure 5. KWIC of "Gas" in 2024 CNPC Annual Report.

KWIC | Plot | File View | Cluster | N-Gram | Collocate | Word | Keyword | Wordcloud | ChatAI

Total Hits: 547 **Page Size** 500 hits | 1 to 500 of 547 hits

	File	Left Context	Hit	Right Context
1	SA-清洗.txt	immaterial, which may in the future become material or affect	Aramco'	s business, financial position and results of operations, or
2	SA-清洗.txt	refineries are not sufficiently competitive in the geographies they serve,	Aramco'	s business, financial position and results of operations could
3	SA-清洗.txt	on favorable terms could have a material adverse effect on	Aramco'	s business, financial position, and results of operations. In
4	SA-清洗.txt	significant disruption, it could have a material adverse effect on	Aramco'	s business, financial position, and results of operations (see
5	SA-清洗.txt	any of which could have a material adverse effect on	Aramco'	s business, financial position, and results of operations. Aramco
6	SA-清洗.txt	Any such losses could have a material adverse effect on	Aramco'	s business, financial position, and results of operations. Aramco'
7	SA-清洗.txt	Aramco's operations and have a material adverse effect on	Aramco'	s business, financial position, and results of operations, could
8	SA-清洗.txt	be burdensome and could have a material adverse effect on	Aramco'	s business, financial position, and results of operations. Aramco
9	SA-清洗.txt	and expertise, it could have a material adverse effect on	Aramco'	s business, financial position, and results of operations. Aramco
10	SA-清洗.txt	any such measures may have a material adverse effect on	Aramco'	s business, financial position, and results of operations. In
11	SA-清洗.txt	as duties, which could have a material adverse effect on	Aramco'	s business, financial position, and results of operations. The
12	SA-清洗.txt	and gas industry could have a material adverse effect on	Aramco'	s business, financial position, and results of operations. In
13	SA-清洗.txt	the Concession, which would have a significant adverse effect on	Aramco'	s business, financial position, and results of operations. Additionally,
14	SA-清洗.txt	penalties or sanctions could have a material adverse effect on	Aramco'	s business, financial position, and results of operations. Aramco'
15	SA-清洗.txt	may impose new obligations on Aramco or otherwise adversely affect	Aramco'	s business, financial position, and results of operations. Aramco
16	SA-清洗.txt	costs or liabilities could have a material adverse effect on	Aramco'	s business, financial position, and results of operations. In
17	SA-清洗.txt	litigation. These consequences could also have an adverse effect on	Aramco'	s business, financial position, results of operations, and reputation.
18	SA-清洗.txt	of these actions could have a material adverse effect on	Aramco'	s business, financial position, and results of operations. The
19	SA-清洗.txt	Any such change could have a material adverse effect on	Aramco'	s business, financial position, and results of operations. Aramco
20	SA-清洗.txt	social unrest will not have a material adverse effect on	Aramco'	s business, financial position, and results of operations. In
21	SA-清洗.txt	international shipping routes could have a material adverse effect on	Aramco'	s business, financial position, and results of operations. Moreover,
22	SA-清洗.txt	which could in turn have a material adverse effect on	Aramco'	s business, financial position, and results of operations or

Figure 6. KWIC of "Aramco" in 2024 Saudi Aramco Annual Report.

As shown in Figure 5 and Figure 6, it may be concluded that narrative style in 2024 CNPC Annual Report pay more attention to achievements-oriented, while narrative style in 2024 Saudi Aramco Annual Report target risks-oriented. For example, the dominant structure related "gas" is Subject (project/field) + verb (steady/surpassed) + object (output) in 2024 CNPC Annual Report. In addition to this, the screenshot of the first 22 contexts show that each sentence introduces new project names, numbers, or locations. In contrast, the dominant structure related "Aramco" is conditional/risk clause + modal verb+ fixed consequence phrase in 2024 Saudi Aramco Annual Report. The screenshot of the first 22 sentences show a stable sentence pattern. Secondly, 2024 CNPC Annual Report prefer indicative mood, while 2024 CNPC Annual Report has a preference for modal verbs such as "could", "may" and "would".

5. Conclusion

This paper utilizes several digital tools, including AntConc 4.3.1, WordSmith Tool 9.0, L2SCA and Citespace to conduct a comparative study. For lexical level, this paper explores macro metrics such as TTR, STTR and STTR Standard deviation. It also lists modal verbs and top 20 nouns to reveal the lexical preferences of domestic and international annual report. For syntactic level, this paper uses L2SCA, concluding that the international annual report has greater reliance on subordination structures. With respect to discourse level, this paper selects the first noun in top 20 to reveal the narrative orientation of the two annual reports.

Furthermore, the study may hold value for the field of translation and international communication. It raises translators' awareness of the deep-seated structural conventions. These conventions may differ between domestic and international annual reports. Translators may leverage these insights to produce translation that are not only accurate but also rhetorically and culturally congruent with the target audience. This may enhance the effectiveness of international communication.

References

1. X. Lu, "Automatic analysis of syntactic complexity in second language writing," *Int. J. Corpus Linguistics*, vol. 15, no. 4, pp. 474–496, 2010.
2. K. Yang, R. Deng, Y. Wei, and S. Wang, "The power of ChatGPT in processing text: Evidence from analysis and prediction in the exchange rate markets," *Financial Innovation*, vol. 11, no. 1, p. 118, 2025.
3. R. Yi, "Tech-empowered equity: advancing linguistic justice through digital scholarship," *Digital Scholarship in the Humanities*, vol. 40, no. 1, pp. 381–399, 2025.
4. M. Mariko, "Comparative analysis of lexical density, lexical diversity, and multiword expressions in Russian, English, and French legal texts: implications for readability and understandability," *Russian Linguistic Bulletin*, vol. 7, no. 67, pp. 1–4, 2025.
5. M. A. Halliday and C. M. I. M. Matthiessen, *An Introduction to Functional Grammar*, 3rd ed., London: Arnold, 2004.
6. Y. M. Shan, "Analysis of a grammatical category in English—modal verbs," *Open J. Soc. Sci.*, vol. 9, pp. 271–278, 2021.
7. N. X. Nasridinova and D. M. Murodovna, "Notional and functional words: understanding their role in language," *PEDAGOGS International Research Journal*, vol. 69, no. 1, pp. 143–147, 2024.