A Corpus-based Diachronic Study on the Linguistic Variation of Corporate Annual Reports from the Real Estate Industry in China

Chaowang REN[1,] Zixu CHEN[2,] Wanting HUANG[3], Cannan CHEN[4], Rui GUO[5]

[1]*School of International Studies, Guangdong University of Technology, E-mail: paulren@gdut.edu.cn*
[2]*School of International Studies, Guangdong University of Technology, E-mail: 1143475944@qq.com*
[3]*School of International Studies, Guangdong University of Technology, E-mail: 2013520170@qq.com*
[4]*School of International Studies, Guangdong University of Technology, E-mail: 2631863196@qq.com*
[5]*School of International Studies, Guangdong University of Technology, E-mail: 986079905@qq.com*

The research is financed by the Grant (No. GD21CWY05) from the Guangdong Provincial Fund for the Philosophy and Social Science Planning Project and the project of Teaching Innovation (*The Research on the Ideological and Political Education Teaching Model in English Major Courses Empowered by Educational Information Technology*) from Guangdong University of Technology.

Abstract

This project compiles a comprehensive corpus by collecting annual reports from Chinese listed companies over multiple years. Based on the characteristics of these reports, the study examines historical trends, industry-specific variations, and differences between Chinese and Western cultural contexts to identify key factors influencing the translation of annual reports in China. It conducts a multi-level analysis of these reports and summarizes the forms and distribution of related events. Tools like AntConc, WordSmith, and other corpus analysis software are used to examine the data. Finally, the study makes reasonable predictions regarding potential future trends based on the research findings.

Keywords: Annual reports, Industry Differences, Diachronic Changes

1. Introduction

This study explores the temporal and industrial variations in the annual reports of Chinese listed companies, set against the backdrop of the 12th and 13th Five-Year Plans. Fuoli (2017) believed the annual corporate report is a formal document that contains legally required and voluntary disclosure information, both digital and textual, about a company's financial position and future prospects. It aims to examine how these reports have evolved over time and across industries, considering the objectives and policies outlined in China′s national economic development plans. By investigating the changes and adaptations companies have made in response to these goals, as well as the broader economic and regulatory landscapes, this research seeks to provide a nuanced understanding of corporate reporting dynamics in China.

The primary objective of this research is to analyze the temporal and industrial dynamics within the annual reports of Chinese listed companies, within the context of the 12th and 13th Five-Year Plans. By examining reports from different time periods and industries, the study aims to identify trends, shifts, and sector-specific practices in reporting, aligned with the strategic objectives outlined in these national plans. Additionally, it seeks to answer the following questions:

How have the keywords in the annual reports of Chinese real estate companies changed over time, compared to BNC?

What are the specific reporting practices and trends in industries such as real estate, manufacturing, finance, and technology, across various time periods, in light of the strategic objectives set forth in the national economic development plans?

2. Literature Review

Corpora have become valuable tools for applications in various fields. The term "corpus" originated from Latin and was initially used in linguistic research. The earliest definition of a corpus can be traced back to 1982, when Professor Francis of Brown University (1982) described a corpus as a collection of texts used for language analysis, representative of a particular language, dialect, or specific aspect of a language. Later, Professor J. Sinclair from the University of Birmingham (1991) in the UK proposed his own definition of a corpus— a naturally occurring collection of language used to reflect the state and changes of a language. Professor Mona Baker from the University of Manchester (1993) was the first to apply corpora to translation studies. In her paper Corpus Linguistics and Translation Studies: Implications and Applications, she elaborated on the theoretical value, practical significance, and specific pathways for using corpora in translation studies. This paper is hailed as the seminal work in corpus-based translation studies, sparking a surge of research in the field.

The annual report plays a crucial role in communicating and shaping the reputation of companies to the public. According to Fulkerson′s (1996) research on the importance of annual reports to investors, two-thirds of portfolio managers and 54% of security analysts consider annual reports the most important document a public corporation can produce.

Gui (2006) believed that high-frequency words refer to words that are frequently used during a specific period. These words are characterized by high usage, wide distribution, strong stability, broad circulation, and a strong combinatory ability. According to statistics from the Brown Corpus, the top 3,000 words cover 84% of the text, with the top 1,000 words covering 72%.

While there is some existing literature on the English translation of corporate annual reports based on corpus analysis, much of the domestic research in corpus-based translation studies has focused on literary genres, with comparatively little attention given to non-literary genres, especially corporate texts. The translation of corporate profiles has attracted increasing scholarly attention in recent years, with most studies adopting qualitative research methods. These studies primarily focus on the differences between Chinese companies and foreign-listed companies in three areas: vocabulary, syntax, and discourse. However, there is a lack of sufficient literature on the diachronic analysis of annual reports from China.

Therefore, this study aims to fill this gap by constructing a corpus of annual reports from China's real estate industry and analyzing its linguistic changes. The study will focus on the use of high-frequency words and their changes over time to reveal industry trends and the evolution of language styles. Through this method, we can gain a deeper understanding of the linguistic characteristics of corporate annual reports and how these characteristics reflect the business conditions and market strategies of enterprises.

3. Methodology

This research builds a specialized corpus of annual reports from listed companies in China, covering the period from 2001 to 2021. For in-depth analysis, we divide the corpus into four periods, each defined by the national Five-Year Plans. We then use corpus analysis tools, such as AntConc (2023) and WordSmith (2009), to examine patterns within the corpus. Wei (2002) believed that the keyword table function of AntConc is useful for analyzing high-frequency words to explore the themes and content features of the corpus.

First, we compare the corpora of companies within the same industry across different time periods by creating keyword lists. For example, Zhang Xuhua (2021) published Patterns and Meanings of High-Frequency Nouns in English and Chinese. Through corpus-driven research, the book analyzes the typical usage of high-frequency nouns in both languages, exploring the pragmatic meanings they convey, the attitudes of language users, and the co-selective relationships between form and meaning. It also innovatively compares the use and meanings of semantically corresponding nouns in English and Chinese. Creating keyword lists enables us to observe changes and trends in the industry over time.

Secondly, by analyzing the collocations of words that show significant differences, we further investigate the reasons behind the changes in the industry. Firth (1957) said, "You shall know a word by the company it keeps." According to Firth, the law of companionship between lexical items, the mutual expectations and attractions between them, and the class associations of collocating elements are all formal properties of word collocation and key aspects of collocation research. The study of collocation focuses on lexical items. By observing, analyzing, and summarizing the typical behavior of words in a given context, we can identify their collocational partners, common grammatical forms, and associated meanings and functions. This approach helps us better understand the dynamics and mechanisms driving change within the industry.

Overall, these analyses provide a comprehensive understanding of the diachronic changes and differences in the real estate industry in China, offering valuable insights for future research and decision-making.

4. Results and Discussion

In order to reveal the process of development of the real estate industry in China, we compared the annual reports with the BNC to generate the keyword lists in different periods. Then, the top frequency words were summarized to show the features of the real estate industry during the past two decades.

4.1 Comparison between the Corpus of Annual Reports from Chinese Real Estate Industry and BNC

First, we selected the annual reports of five publicly listed real estate companies from the 10th Five-Year Plan period (2001-2005) and built the Subcorpus 1, as our target corpus, focusing specifically on the "Management's Discussion and Analysis" section. The BNC corpus was used as the reference corpus. We generated a list of keywords by comparing the two corpora.

Table 1. The Keywords of Real Estate Companies during the 10th Five-Year Plan Period

Type	Rank	Freq_Subcorpus 1	Freq_BNC	Range_Subcorpus 1	Range_BNC	Keyness (Likelihood)
company	3	1765	0	28	0	2731.736
RMB	6	1006	0	28	0	1552.893
group	8	642	1	24	1	976.066
profit	10	576	0	28	0	887.803
year	11	680	21	28	1	885.797
capital	12	548	0	27	0	844.564
development	13	509	0	27	0	784.352
report	15	468	0	20	0	721.07
ltd	16	444	0	27	0	684.035
co	17	443	0	27	0	682.492
HK	18	400	0	16	0	616.154
property	19	385	0	25	0	593.017
board	19	385	0	21	0	593.017
directors	21	372	0	23	0	572.967
business	22	378	1	27	1	569.59
period	23	460	23	26	1	552.331
Shenzhen	24	322	0	19	0	495.87
project	25	335	2	26	1	493.901
sales	26	303	0	23	0	466.579
meeting	27	315	2	24	1	463.312

The top 10 keywords, "company," "RMB," "group," "profit," "year," "capital," "development," "report," "Ltd," and "Co.", highlight the real estate industry's emphasis on financial transparency and capital management. This focus on investment planning and company structure suggests a trend toward standardization and strategic management, particularly evident during the 10th Five-Year Plan period, which indicates an upward trajectory in the sector.

Next, we collected the annual reports of five listed real estate companies from the 11th Five-Year Plan period (2006-2010) and focused on the "Management's Discussion and Analysis" sections to build Subcorpus 2 as our target corpus. The BNC corpus was used as the reference corpus. Using AntConc, we generated a keyword list by comparing the two corpora, highlighting significant linguistic differences and trends.

Table 2. The Keywords of Real Estate Companies during the 11th Five-Year Plan Period

Type	Rank	Freq_Sub corpus 2	Freq_BNC	Range_Sub corpus 2	Range _BNC	Keyness (Likelihood)
company	5	3119	0	53	0	2389.389
RMB	6	2954	0	56	0	2262.364
group	8	2438	1	56	1	1850.262
year	10	1987	21	56	1	1334.057
development	12	1197	0	56	0	914.067
December	15	1089	0	40	0	831.446
approximately	16	1101	1	41	1	826.909
sales	18	1048	0	55	0	800.089
project	19	979	2	51	1	723.134
property	20	871	0	52	0	664.764
total	22	920	5	56	1	651.51
projects	23	848	0	53	0	647.186
profit	24	834	0	52	0	636.486
increase	26	994	22	56	1	597.093
business	27	797	1	56	1	595.143
area	28	788	1	41	1	588.288
management	29	741	0	56	0	565.424
construction	30	782	3	53	1	564.242
HK	31	716	0	29	0	546.325
financial	32	708	1	52	1	527.382

The top 10 keywords — "company," "RMB," "group," "year," "development," "December," "approximately," "sales," "project," and "property" — emphasize financial performance, project specifics, and market dynamics. This suggests a trend toward greater transparency and a strategic focus on growth and sales, particularly evident during the 11th Five-Year Plan period.

Next, we selected the annual reports of five listed real estate companies from the 12th Five-Year Plan period (2011-2015) and focused on the "Management's Discussion and Analysis" sections to create Subcorpus 3. The BNC corpus served as the reference corpus. Using AntConc, we generated a keyword list by comparing the two corpora to identify significant linguistic trends and shifts in focus during this period.

Table 3. The Keywords of Real Estate Companies during the 12th Five-Year Plan Period

Type	Rank	Freq_Sub corpus 3	Freq_BNC	Range_Sub corpus 3	Range_ BNC	Keyness (Likelihood)
RMB	5	2612	0	59	0	2263.31
group	6	2539	1	55	1	2184.155
company	8	1964	0	58	0	1699.733
year	10	1762	21	59	1	1339.882
development	11	1240	0	59	0	1071.688
December	12	1029	0	45	0	888.975
project	13	1051	2	48	1	883.142
sales	14	1006	0	59	0	869.067
property	17	911	0	54	0	786.857
total	18	892	5	58	1	719.016
properties	20	818	0	50	0	706.407
approximately	23	755	1	38	1	638.763
area	24	742	1	39	1	627.557

management	25	690	0	59	0	595.726
projects	25	690	0	57	0	595.726
profit	28	638	0	55	0	550.777
land	29	608	0	55	0	524.848
investment	30	597	0	59	0	515.342
HK	31	593	0	27	0	511.885
business	32	588	1	59	1	494.904

The top 10 keywords - "RMB," "group," "company," "year," "development," "December," "project," "sales," "property," and "total" - primarily reflect financial aspects. Notably, the emphasis on terms related to development and growth indicates that the real estate industry experienced an upward trend during the 12th Five-Year Plan period.

Finally, we selected the annual reports of five listed real estate companies from the 13th Five-Year Plan period (2016-2020) and focused on the "Management Discussion and Analysis" sections as the Subcorpus 4. The BNC corpus served as the reference corpus, and we utilized AntConc to generate the keyword list comparing the two corpora. This analysis aimed to identify significant trends and shifts in focus within the real estate sector during this period.

Table 4. The Keywords of Real Estate Companies during the 13th Five-Year Plan Period

Type	Rank	Freq_Sub corpus 4	Freq_BNC	Range_Sub corpus 4	Range_BNC	Keyness (Likelihood)
RMB	4	3311	0	60	0	2812.749
group	5	3161	1	58	1	2668.57
year	8	2054	21	61	1	1550.671
development	10	1392	0	61	0	1178.339
company	11	1292	0	58	0	1093.486
December	12	1226	0	55	0	1037.501
property	13	1043	0	58	0	882.34
project	14	994	2	50	1	816.226
total	15	983	5	59	1	779.284
sales	16	876	0	61	0	740.836
business	17	871	1	61	1	723.187
billion	21	820	0	48	0	693.405
projects	22	807	0	58	0	682.396
management	23	781	0	60	0	660.379
increase	25	954	22	58	1	643.334
approximately	26	766	1	47	1	634.522
properties	27	701	0	54	0	592.647
land	29	674	0	55	0	569.792
investment	30	657	0	60	0	555.403
market	32	636	2	60	1	514.832

The top 10 keywords-"RMB", "group", "year", "development", "company", "December", "property", "project", "total", and "sales"-are primarily finance-related. Notably, the focus on terms indicating growth and development suggests that the real estate industry experienced an upward trend during the 13th Five-Year Plan period.

4.2 Wordlist Analysis of the Annual Reports in Different Periods

We selected the Management Analysis and Discussion section from the annual reports of five companies during the 10th Five-Year Plan period (2001-2005) from the corpus of listed companies and used AntConc to generate a word list for this section during that period.

Table 5. The Top Frequency Words of Management's Analysis and Discussion from Real Estate Companies during the 10th Five-Year Plan Period

Type	Rank	Freq	Range
company	6	1765	28
RMB	8	1006	28
year	16	680	28
group	17	642	24
profit	19	576	28
capital	21	548	27
development	22	509	27
report	24	468	20
period	25	460	26
ltd	26	444	27
co	27	443	27
HK	28	400	16
board	29	385	21
property	29	385	25
business	33	378	27
directors	34	372	23
project	36	335	26
Shenzhen	38	322	19
total	40	318	28
meeting	41	315	24

The ten most frequently used words in the chairman's speech during this period include "company", "RMB", "year", "group", "profit", "capital", "development", "report", "period" and "ltd". These terms indicate the company′s focus on financial performance and annual reports, showing the industry is experiencing stable growth and capital management, with an emphasis on profit and capital. It can also be inferred from "development" that the real estate industry was in an upward trend during this period.

Then, we selected the Management Analysis and Discussion section from the annual reports of five companies during the 11th Five-Year Plan period (2006-2010) from the corpus of listed companies and used AntConc to generate a word list for this section during that period.

Table 6. The Top Frequency Words of Management's Analysis and Discussion from Real Estate Companies during the 11th Five-Year Plan Period

Type	Rank	Freq	Range
company	7	3119	53
RMB	8	2954	56
group	11	2438	56
year	16	1987	56
development	21	1197	56
approximately	24	1101	41
December	25	1089	40
will	26	1071	54
sales	27	1048	55
increase	29	994	56
project	30	979	51
total	32	920	56
property	33	871	52

projects	34	848	53
profit	35	834	52
business	36	797	56
area	38	788	41
construction	39	782	53
management	40	741	56
HK	42	716	29

The ten most frequently used words in the chairman's speech during this period include "company," "RMB," "group," "year," "development," "approximately," "December," "will," "sales," and "increase." These terms reflect the company′s focus on annual performance, sales growth, and future outlook, indicating that the industry is expanding steadily and actively planning for future development.

Similarly, we selected the "Management Analysis and Discussion" section from the annual reports of five companies during the 12th Five-Year Plan period from the corpus of listed companies and used AntConc to generate a word list for this section during that period.

Table 7. The Top Frequency Words of Management's Analysis and Discussion from Real Estate Companies during the 12th Five-Year Plan Period

Type	Rank	Freq	Range
RMB	8	2612	59
group	9	2539	55
company	12	1964	58
year	13	1762	59
development	21	1240	59
project	22	1051	48
December	23	1029	45
sales	24	1006	59
property	27	911	54
total	28	892	58
will	29	854	59
properties	31	818	50
approximately	33	755	38
area	34	742	39
management	36	690	59
projects	36	690	57
increase	38	686	55
profit	39	638	55
land	40	608	55
investment	42	597	59

The ten most frequently used words in the chairman's speech during this period include "RMB," "group," "company," "year," "development," "project," "December," "sales," "property," and "total." These high-frequency words reflect the real estate industry's focus on financial performance, project progress, and annual summaries, indicating the industry's emphasis on capital operations, corporate structure, and market sales trends.

Finally, we selected the "Management Analysis and Discussion" section from the annual reports of five companies during the 13th Five-Year Plan period from the corpus of listed companies. Using AntConc, we generated a word list for this section during that period.

Table 8. The Top Frequency Words of Management's Analysis and Discussion from Real Estate Companies during the 13th Five-Year Plan Period

Type	Rank	Freq	Range
RMB	6	3311	60
group	7	3161	58
year	12	2054	61
development	19	1392	61
company	20	1292	58
December	21	1226	55
property	23	1043	58
project	24	994	50
total	26	983	59
increase	27	954	58
sales	30	876	61
business	31	871	61
billion	33	820	48
will	33	820	60
projects	35	807	58
management	36	781	60
approximately	37	766	47
properties	38	701	54
land	39	674	55
investment	40	657	60

The ten most frequently used words in the chairman's speech during this period include "RMB", "group", "year", "development", "company", "December", "property", "properties", "project" and "total". These high frequency words show that the real estate industry focuses on financial performance, project management and annual summary, reflecting the industry's continuous attention to capital operation, asset management and project development.

5. Conclusion

Based on an in-depth analysis of the corpus, we conclude that the real estate industry in China has shown an overall upward development trend. Despite the various challenges encountered during the industry’s development, steady growth momentum has been maintained. This growth can be attributed to positive shifts in both the domestic and international economic environments, as well as strong policy support for the real estate market.

Indeed, the results of this analysis are highly consistent with the observed facts, further demonstrating the value and effectiveness of corpus analysis in revealing industry trends and predicting future developments. By conducting a thorough analysis of the vast amounts of data in the corpus, we can more accurately grasp the industry′s development context and trends, thereby providing strong support for decision-making.

Therefore, corpus analysis has become an invaluable tool for various types of research. It not only helps us understand the current state and future directions of the industry but also offers valuable data-driven insights. It is believed that, in future research, we will continue to fully leverage the advantages of corpus analysis, contributing further insights and support to the industry′s development and progress.

References

Anthony, L. (2023). *AntConc* (Version 4.2.4) [Computer software]. Waseda University. https://www.laurenceanthony.net/software.

Baker, M. (1993). *Corpus Linguistics and Translation Studies: Implications and Applications*. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), Text and Technology: In honor of John Sinclair. Amsterdam (pp. 233-250). John Benjamins. https://doi.org/10.1075/z.64.15bak.

Firth, J.R. (1957) *Papers in linguistics 1934–51*. Oxford: Oxford University Press.

Fulkerson, J. . (1996). How investors use annual reports. *American Demographics, 18*(5), 16-19.

Francis N. (1982). Problems of assembling and computerizing large corpora. In *Computer Corpora in English Language Research*. Bergen: Norwegian Computing Centre for the Humanities.

Fuoli, M. (2017). Building a trustworthy corporate identity: a corpus-based analysis of stance in annual and corporate social responsibility reports. *Applied Linguistics, 39*(6), 846-885.

Gui, S. (2006). An overview of English vocabulary learning: Answers to common questions. *Foreign Language World,* 01, 57-65. [In Chinese: 桂诗春. (2006). 英语词汇学习面面观——答客问. 外语界(01), 57-65.]

Scott, M. (2009). *WordSmith Tools* (Version 5) [Computer software]. Lexical Analysis Software. http://www.lexically.net/wordsmith/

Sinclair J. (1991). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.

Wei Naixing. (2002). Corpus-based and corpus-driven word Collocation Research. *Contemporary Journal of Linguistics, 4*(2), 101-114,157. [In Chinese: 卫乃兴. (2002). 基于语料库和语料库驱动的词语搭配研究.当代语言学, 4(2), 101-114,157.]

Zhang Xuhua. (2021). *Patterns and meaning of High Frequency Nouns across English and Chinese*. Shanghai: Shanghai Foreign Language Education Press. [In Chinese: 张绪华. (2021). 型式与意义— —语料库驱动的英汉高频名词对比研究. 上海：上海外语教育出版社]

.