

A Historical Comparative Study on the Annual Reports of Chinese Public Companies

Chaowang REN¹; Haolin ZHENG²

¹ *Guangdong University of Technology, Guangzhou, China paulren@gdut.edu.cn*

² *Guangdong University of Technology, Guangzhou, China 2112316004@mail2.gdut.edu.cn*

Abstract

The significant changes in human social history will undoubtedly lead to alterations in language use. Concurrently, the diachronic study of language change can reflect historical changes in society. This paper employs the principles of corpus construction and utilizes AntConc and other corpus retrieval tools to conduct a diachronic comparative analysis of discourse variation in management discussion and analysis. This analysis is based on a self-built corpus and integrates findings from diachronic linguistics and sociolinguistics. We examine discourse variation in management discussions through three key levels: word frequency, keywords, and collocates. By delving into the text, we perform a diachronic comparative analysis of management discussions, emphasizing the dialectical relationship between discourse and society. This approach highlights language use across different time periods, reveals patterns of language change, and supports the development and transformation of society. The authors assert that corpus-based diachronic language change elucidates the nature of language evolution as a reflection of social change. Furthermore, it provides valuable insights for a deeper understanding of the complex causes of language variation and offers material for the standardization of language expression.

Keywords: a corpus-based approach, historical comparative study, language variation, annual reports

1. Introduction

Diachronic language research, a key aspect of usage-based linguistics, emphasizes the importance of utilizing real corpora. The rapid development of diachronic studies can largely be attributed to the ongoing construction of electronic diachronic corpora. In linguistic literature, this area of research is commonly referred to as "historical linguistics" or "diachronic linguistics," focusing on the mechanisms of language evolution (Campbell, 1999: xiv). Relevant studies encompass evidence of lexical evolution (etymology), the construction of language family trees (contrastive linguistics), and the exploration of mechanisms that facilitate the transformation of content words into functional words (grammaticalization).

Furthermore, there is a notable overlap between diachronic language studies and the field of sociolinguistics. Diachronic variation within sociolinguistics, archaic usage in dialectology, and contemporary language interpretation through the lens of language contact all exhibit diachronic characteristics. Language is inherently dynamic, and change is fundamental to its nature. Thus, conducting diachronic language research is essential within the field of linguistics.

The term "Management's Discussion & Analysis" (MD&A) denotes the management's evaluation and interpretation of a company's past operational performance, as well as forward-looking assessments concerning future trends, as presented in periodic reports by publicly listed companies. Often regarded as the "core and soul" of financial reporting, MD&A constitutes an essential element of the information disclosed by these entities. It provides insights that extend beyond the GAAP-compliant data found in tables and notes, thereby fulfilling the growing demand among investors for pertinent and prospective information.

This disclosure mechanism allows investors to better understand management's perceptions of significant historical events, assess key risks, and grasp the rationale behind projections regarding future developments. Due to its essential function in improving the quality of corporate information transparency, MD&A has been widely adopted in nations such as the United States, the United Kingdom, and Canada. Notably, it was recognized as one of the top ten financial reporting challenges by the American Association of Financial Managers in 2005. Additionally, in response to the Enron accounting scandal, the U.S. Securities and Exchange Commission instituted MD&A as a crucial corrective action. In China, the China Securities Regulatory Commission launched the MD&A framework in 2002 and subsequently reinforced regulations in 2005.

A comprehensive analysis of the discourse within annual reports of Chinese public companies can significantly enhance investors' understanding of these entities, thereby facilitating more effective resource allocation in the capital market. Moreover, a comparative study of MD&A sections in the annual reports of Chinese public companies, grounded in corpus analysis, has the potential to enrich diachronic research in China. This approach will expand the corpus field and allow for a multidimensional and multi-level examination of the linguistic features present in these reports. Consequently, this research

aims to systematically and objectively explore the characteristics of language change within the annual reports of public companies in China, as well as elucidate the dialectical relationship between language change and societal factors.

2. Literature review

Corporate annual reports represent a critical component of the information disclosed by publicly listed companies, playing an essential role in enhancing the quality of corporate information disclosure. In linguistic literature, diachronic language research is commonly referred to as "historical linguistics" or "diachronic linguistics," focusing on the mechanisms underlying language variation.

To refine and mature our Management's Discussion & Analysis (MD&A) system, there is an increasing necessity to improve the quality of information disclosure among listed companies. The academic community must urgently engage in theoretical research on MD&A to provide foundational support and guidance for future practices in this area.

2.1 Previous studies on the discourse of Management's Discussion and Analysis

Management's Discussion and Analysis (MD&A) serves as an essential element of a public company's annual report or quarterly filing, focusing on the organization's performance. In this section, the management team and executives—commonly referred to as the C-suite—offer an analysis of the company's performance employing both qualitative and quantitative metrics. Furthermore, the MD&A may include discussions related to compliance, risk factors, and future strategies, encompassing organizational objectives and upcoming initiatives.

MD&A constitutes the central component of the "Business Situation Discussion and Analysis" section in the annual reports of publicly listed companies. It incorporates essential financial and non-financial information, along with both historical and prospective data, thereby significantly enhancing the disclosure of corporate financial information.

Numerous studies have advanced the understanding of the effects and implications of MD&A. Caserio (2019) explored whether financial companies in the United States exhibited a tendency to manipulate the tone of their MD&A to engage in impression management practices. Lee (2019) investigated how the financial expertise of audit committees affected the quality of the textual information presented in the MD&A sections of corporate annual reports. Durnev (2020) analyzed the robustness of disclosures within the MD&A, concluding that firms in industries characterized by lower entry costs, larger scales, and less substitutable products were more likely to provide comprehensive disclosures. Jaspersen (2020) utilized supervised machine learning models to predict earnings management behaviors among property and casualty insurers based on their MD&A submissions. Lastly, Pajuste (2020) contended that increased scrutiny from market participants in more liquid companies motivated managers to obscure unfavorable financial outcomes through heightened complexity in reporting.

2.2 Previous corpus-based diachronic discourse analysis

Conrad (2002) offered an extensive examination of methodologies within corpus linguistics that pertain to discourse-level phenomena. The primary aim was to amalgamate two analytical perspectives: performing a thorough discourse analysis of individual texts while also generalizing insights across the entire corpus. Biber (1998) distinguished two principal approaches to this endeavor: 'top-down' and 'bottom-up'. The study employed relatively large corpora characterized by parallel structures and content from various contemporary periods, specifically utilizing the SiBol corpora to monitor changes in modern language use in relation to associated social, cultural, and political transformations as evidenced in linguistic expression.

Partington (2010) provided a comprehensive overview of corpus-assisted discourse studies (CADS) and multidimensional CADS (MD-CADS), with the objective of clarifying the methodologies typically employed in this area and offering insights into newspaper corpora. Brown et al. (2012) analyzed a corpus of Medieval Spanish texts to explore the factors affecting the outcome of the initial /f/ in Latin FV- words as they evolved into Modern Standard Spanish. Li (2013) introduced a discourse-level joint model for argument identification and role assignment, thereby advancing the understanding of inter-sentence arguments within discourse.

The challenges of achieving stable and reliable results from corpus analysis, particularly in studies of lexical variation, were highlighted by Miller (2015), who illustrated these difficulties through experiments analyzing a corpus related to university-level introductory psychology textbooks. Odebrecht (2017) introduced a multi-layer corpus architecture featuring multiple tokenizations, utilizing the open-source historical, diachronic corpus of German known as Register in Diachronic German Science. Furthermore, speeches delivered from 1950 onwards have often been made available online through video recordings, enabling new research avenues in multimodal linguistics. This includes studies on the integration of speech and gesture in meaning construction, as well as analyses of recurrent gestures and multimodal constructions (Debras, 2021).

2.3 Research questions

(1) What linguistic variations can be identified in the Management's Discussion and Analysis section of annual reports from Chinese public companies across different time periods?

(2) What are the key factors contributing to the linguistic variations observed in the Management's Discussion and Analysis sections of annual reports from Chinese public companies?

3. Data and methodology

By collecting and cleaning the target texts, we obtained a refined version to construct the Chinese corpus. Analytical tools were utilized to extract the linguistic features from the corpus.

3.1 Data

3.1.1 Data collecting

In this study, we excluded ST companies and those with incomplete data from the 145 Chinese firms listed in the 2022 Fortune Global 500. Ultimately, we selected the management discussion and analysis (MD&A) sections of annual reports from 30 A-share companies for the years 2007 to 2021 as our research data. All data were sourced from www.cninfo.com.cn. According to annual report disclosure guidelines, companies must include specific MD&A information in the "Report of the Board of Directors," such as: an overview of operational performance during the reporting period; a list of main business activities by product and region; an analysis of significant changes in operating results and profit composition compared to the previous year; statements on the impact of major changes in the production environment, macro-policies, and regulations on financial position and operating results; the board's business plan for the upcoming year; and the profit forecast for the New Year. Although some companies label their MD&A sections as "business situation discussion and analysis," the content is fundamentally similar. Therefore, these sections were also included in the study's data set.

3.1.2 Data cleaning

Data cleaning is to keep useful data in the corpus and delete noise data. First, the annual reports of 30 target companies were obtained in pdf from www.cninfo.com.cn. Then the pdf file was converted into Word version, and the management discussion and analysis section of the annual report was manually selected to delete the table content, English characters and other information that affected the search results.

3.2 Methodology

This chapter introduces the research methodology and analytic tools employed in this research. First a brief introduction to corpus and corpus-based approach is presented. Then the issues of corpus construction are discussed, including the principles, data selection and annotation. Finally, it is followed by the description of the analytic tools.

3.2.1 An introduction to corpus and corpus-based approach

Corpus is from the Latin word, meaning "body", hence a corpus refers to any body of text. But the term "corpus" has the specific connotation in modern linguistics.

According to Sinclair (1999), corpus is defined as "a collection of naturally occurring language text, chosen to characterize a state or variety of a language." Biber et al. (1998) emphasize the authenticity and purpose of a corpus by noting "It is a large and principled collection of natural texts." It is used to describe a collection of naturally occurring examples of language, consisting of written texts or tape recordings; recently it has been reserved for collections of texts that are stored and accessed electronically (Hunston, 2002).

There are two main approaches to corpus linguistics, a corpus-based approach and a corpus-driven approach. In the corpus-based approach, corpora are used mainly to "expound, test or exemplify theories and descriptions that were formulated before large corpora became available to inform language study". In contrast, corpus-driven approach claims that 'the theoretical statements are fully consistent with, and reflect directly, the evidence provided by the corpus (Tognini-Bonelli, 2001:65-84). The most notable difference is that the corpus-driven approach claims to be a new paradigm within which a whole language can be described, while the corpus-based approach has made no such claim. The corpus-based approach is widely applied in nearly all branches of linguistics. Therefore, the corpus-based approach has made it possible to conduct new kinds of investigations into language use.

The advantages of corpus-based approach are obvious. The corpus-based approach takes advantage of computers' capability for fast, accurate, and complex analysis; the extensive information about language use found in large collections of natural texts from multiple registers; and the rich descriptions that result from integrating quantitative findings and functional interpretations (Biber et al., 1998:233). Even though the corpus does not contain new information about language, the corpus-based approach offers a new perspective view with the help of software.

3.2.2 Constructing the MD&A corpus

At the "2003 International Conference on Corpus Linguistics" held in Shanghai Jiaotong University, John Sinclair mentioned that corpus construction became a popular practice among language teachers and researches. Although there are some

established-corpora available, yet those do not always suit the purpose of every research. Researchers can establish a mini-corpus with the availability of computer technology, software and on-line data to satisfy the aim of research.

As the purpose of this paper is to discuss the diachronic linguistic variation of the Management's Discussion and Analysis in annual reports of Chinese public companies during the 15 years, the Management's Discussion and Analysis in annual reports is selected as data for the construction of MD&A Corpus.

Allan Bell (1991) suggests that when establishing a news corpus, "Decisions are needed on three questions: the time period the sample will cover, the days to be sampled within that period, and the specific issues or bulletins to be sampled within those days."

For selecting data from MD&A in annual, random sampling is employed. The selected annual reports are from 2007 to 2021. As most of the websites can be restored as a text file, the easiest way of obtaining texts is downloading the electronic format from the internet.

3.3 Analytical procedures

The AntConc was used to retrieve and sort high-frequency words in the three subcorpora. Functional words refer to the words that only play a grammatical role without inflection and have no practical meaning in the literature, such as prepositions, articles, conjunctions, pronouns, auxiliary verbs, etc. They cannot reflect or represent the content of the literature, and their resolution is very low or even zero. Therefore, the analysis of this paper will exclude the function words and only analyze the content words. In this paper, content words with the top ten word frequency in the three subcorpora will be selected for analysis. If necessary, the collocate search function of AntConc will be combined for joint analysis and the search results will be sorted out in Excel, so as to obtain the similarities and differences of words in the three subcorpora. Two analytic tools were used.

(1) AntConc

AntConc, developed by Professor Laurence Anthony at Waseda University, is a free corpus retrieval tool widely used in corpus linguistics, translation studies, and foreign language teaching (Anthony, 2024). It is recognized as an authoritative software for linguistic research and supports various languages, including Chinese, Japanese, and Korean. AntConc can perform corpus statistics on UTF-8 and UTF-16 encoded texts and utilizes regular expressions, particularly beneficial for processing Chinese corpora, where it can batch-add spaces.

(2) Microsoft Excel

Microsoft Excel, part of the Microsoft Office suite, is a spreadsheet software utilized for data processing, statistical analysis, and decision-making. It is extensively used in management, finance, and statistical fields. In this paper, Excel will be employed for data statistics and organization.

4. The Diachronic Linguistic Variations of the Corporate Annual Reports in China

Corpus-based approaches to language study have introduced new dimensions to linguistic description and to various applications by permitting some degree of automatic analysis of texts. In this study, the self-established MD&A Corpus is used to describe discourse linguistic variation from wordlist level, keywords level, and collocation and word cluster level. In the following sections, MD&A Corpus is researched with the aid of the AntConc tool from both quantitative and qualitative perspectives.

4.1 The General Description of the MD&A Corpus

As mentioned in Chapter Three, the corpus for the current study is compiled based on the principles to guarantee the balance thus the validity of analysis result, it is still necessary to present some of the general information to verify it. MD&A Corpus is composed of three sub-corpora, with 26312315 words and 39755 types in total. The six sub-corpora line up chronologically. The general statistics of each sub-corpus are listed in Table 4-1.

Table 4-1 General statistics of three sub-corpora

Text file	File size	Tokens (running words)	Types (distinct words)	Type/ token ratio (TTR)	Standardized TTR	Mean word length
2007-2011	6724990	806660	16972	2.26	37.45	1.9
2012-2016	8794900	1044206	20661	2.13	38.55	1.92
2017-2021	10792425	1268889	26830	2.26	40.41	1.92

Type Token Ratio (TTR): is a measure of the ratio of different words to the total number of words in a text, sometimes used as a measure of the difficulty of a 32 passage or text. It is also referred to as lexical density which is normally expressed as a percentage and is calculated by the formula:

$$\text{Lexical Density} = \frac{\text{The number of different words}}{\text{The total number of words in the text}} \times 100$$

Standardized Type Token Ratio (STTR): is an average type/token ratio based on consecutive 1,000-word chunks of text, which can alleviate the influence of the number of tokens in the sub-corpus.

From Table 4-1, we can get the general view of the three sub-corpora. Even though the construction of corpus is based on the rules, yet there are some differences between each sub-corpus because of the different formats of each year. The tokens of the 3 sub-corpora are 806660, 1044206, and 1268889 respectively. It can also be found that the TTR varies from 2.13 to 2.26. Different numbers of tokens of each year influence the figure of TTR, thus, so STTR is more meaningful to show that the lexical diversity is comparatively stable during the past 15 years. The STTR of the 3 sub-corpora are 37.45, 38.55, and 40.41 respectively. The language difficulty of MD&A maintains the same during the 15 years.

4.2 The Diachronic Description of the Linguistic Features of the Subcorpora

The MD&A Corpus consists of three sub-corpora. When these sub-corpora are combined and analyzed, they reveal a new perspective. The total size of the MD&A Corpus is 26,312,315 words, with a total of 39,755 unique types. This chapter will explore the linguistic variation within the MD&A Corpus over the past 15 years by analyzing the three sub-corpora across different dimensions.

To conduct a word frequency analysis, we utilized the Wordlist function in AntConc. This tool counts all words in the corpus and presents them in an ordered list, allowing us to identify the most frequent terms. The frequency of these words reflects linguistic variations over the 15-year period. Once the Wordlist is generated, we can assess the frequencies. Additionally, the MD&A Corpus exhibits patterns in parts of speech that align with findings from other corpora. This section will provide statistical analyses and comparative insights into the top ten high-frequency nouns, verbs, and adjectives derived from the three sub-corpora.

4.2.1 The high-frequency nouns in the subcorpora

A noun is a word that represents a person, a place, or a thing. A thing in this definition can be a physical entity or it can be an abstract idea. A noun may be used as a subject of a sentence, as a direct object, as an indirect object, or it may be used as the object of a preposition (Quirk & Greenbaum, 1973:49). Thus, the nouns need to be discussed firstly. The top 10 nouns in each sub-corpus are listed in Table 4-2.

Table 4-2 The top 10 nouns in three sub-corpora

Rank	High-frequency nouns in 2007-2011	Frequency	High-frequency nouns in 2012-2016	Frequency	High-frequency nouns in 2017-2021	Frequency
1	公司 (Company)	13056	公司 (Company)	18293	公司 (Company)	16826
2	业务 (Business)	9842	业务 (Business)	13684	风险 (Risk)	14304
3	管理 (Management)	7850	管理 (Management)	11650	管理 (Management)	13322
4	风险 (Risk)	6925	风险 (Risk)	11078	业务 (Business)	13246
5	市场 (Market)	5380	资产 (Assets)	7881	金融 (Finance)	8615
6	人民币 (Renminbi)	5229	金融 (Finance)	6440	客户 (Client)	8227
7	资产 (Assets)	4428	市场 (Market)	6279	服务 (Service)	8113
8	银行 (Bank)	4180	银行 (Bank)	6205	资产 (Assets)	7647
9	集团 (Group)	4005	客户 (Client)	6127	集团 (Group)	7274
10	产品 (Product)	3921	集团 (Group)	5921	产品 (Product)	6985

From the words in Table 4-2, we can see that the focus of MD&A is about company and the annual reports serves as a window for the whole world to see the company. Through comparison, it is found that the top 10 nouns in the word frequency ranking of the three sub-corpora do not change very much. The word frequency ranking of most words still remains in the top 10 of the sub-corpora under the influence of time, but there are also some notable phenomena. For example, the word "公司 (company)" always ranks in the top one words frequencies of the three sub-corpora, which indicates that the part of

management's discussion and analysis focuses on the company and always focuses on discussing and analyzing all affairs related to the company.

Secondly, in the comparison, it is found that among the top ten words in the corpus in terms of word frequency in each time period, the word "管理(management)" is also ranking the third space in each sub-corpora. From this phenomenon, it can be concluded that the management of the company pays more attention to the management of the company and promotes the good operation of the company by improving the company's management system, so as to better attract investors to invest in the company.

In addition, we can see that the word frequency of the word "风险(risk)" ranked 4th in the two sub-corpora from 2007 to 2011 and 2012 to 2016, but jumped to the second place in the 2017-2021 sub-corpus. We can speculate that more and more enterprises are paying more attention to the standardization of enterprises and the perfection and improvement of investment system. In particular, since 2019, the impact of the novel coronavirus pneumonia epidemic has increased the downward risk of the global economy, which will also increase the business risk.

Moreover, it is worth noting that the word "服务(service)" suddenly appear in the third sub-corpus. In particular, the noun frequency of the word "服务(service)" ranks seventh in the third sub-corpus. After comparison, it is found that the word frequency ranking of "服务(service)" increases year by year in the three sub-corpora, which indicates that enterprises pay more attention to the improvement of service and pay attention to the provision and improvement of service. The frequency of "服务(service)" in each sub-corpus is listed in Table 4-3.(Note: This data is the ranking and frequency of "服务(service)" in the three sub-corpora, regardless of speech.)

Table 4-3 The frequency of "服务(service)" in each sub-corpus

Period	Frequency	Rank
2007-2011	2,919	24
2012-2016	5,520	18
2017-2021	9,444	8

4.2.2 The high-frequency verbs in the subcorpora

Table 4-4 The top 10 verbs in the three sub-corpora

Rank	High-frequency verbs in 2007-2011	Frequency	High-frequency verbs in 2012-2016	Frequency	High-frequency verbs in 2017-2021	Frequency
1	增长 (Grow)	4514	增长 (Grow)	5733	增长 (Grow)	5863
2	增加 (Increase)	3849	增加 (Increase)	4140	提升 (Enhance)	5659
3	实现 (Achieve)	2518	实现 (Achieve)	3757	实现 (Realize)	4762
4	提高 (Enhance)	2017	提升 (Enhance)	3714	增加 (Increase)	3599
5	提升 (Elevate)	1902	推进 (Advance)	2614	推进 (Promote)	3538
6	加强 (Strengthen)	1854	加强 (Strengthen)	2541	加强 (Strengthen)	2764
7	推进 (Advance)	1597	发展 (Develop)	2043	提供 (Provide)	2545
8	发展 (Develop)	1582	优化 (Optimize)	1926	优化 (Optimize)	2497
9	继续 (Continue)	1439	下降 (Decline)	1889	发展 (Develop)	2174
10	下降 (Decline)	1386	提高 (Improve)	1875	推动 (Drive)	2103

As can be seen from Table 4-4, the top ten verbs with word frequency in the three sub-corpora have not changed much except their ranking. However, we can also find that the top ten verbs with the highest frequency in the three sub-corpora have a common feature, that is, most of them express positive attitudes. Pava and Epstein (1993) thought most companies, in the disclosure of forward-looking information, tended to disclose good news not only for the most part, and the number of accuracy would more than double that of bad news. Many companies clearly have poor operating conditions, but if you only look at the MD&A, you will think that the situation of the company is promising.

4.2.3 The high-frequency adjectives in the subcorpora

Table 4-5 The top 10 adjectives in three sub-corpora

Rank	High-frequency adjectives in 2007-2011	Frequency	High-frequency adjectives in 2012-2016	Frequency	High-frequency adjectives in 2017-2021	Frequency
1	主要 (Major)	1772	新 (New)	2646	新 (New)	3950
2	新 (New)	1432	有限 (Limited)	2192	综合 (Comprehensive)	1887
3	有限 (Limited)	825	主要 (Major)	2013	大 (Large)	1817
4	相关 (Relevant)	718	综合 (Comprehensive)	1440	主要 (Major)	1544
5	综合 (Comprehensive)	657	大 (Large)	1171	高 (High)	1379
6	宏观 (Macro)	630	相关 (Relevant)	1062	相关 (Relevant)	1309
7	总 (Total)	607	净 (Net)	839	有限 (Limited)	1278
8	大 (Large)	536	重大 (Significant)	748	绿色 (Green)	1107
9	不良 (Adverse)	495	总 (Total)	711	重大 (Significant)	964
10	净 (Net)	494	不良 (Adverse)	686	总 (Total)	899

As can be seen from Table 4-5, the top 10 adjectives in the three sub-corpora have not changed much except their ranking. However, it is worth noting that the word "green" popped up in the third sub-corpus and ranked seventh.

Since the beginning of the 21st century, the report to the 17th National Congress of the Communist Party of China has listed ecological civilization as a new requirement for the goal of building a well-off society in an all-round way. The report to the 18th CPC National Congress made important breakthroughs in the overall layout of the cause of socialism with Chinese characteristics, including ecological civilization in the five-in-one overall layout of the cause of socialism with Chinese characteristics. At the Fifth Plenary Session of the 18th CPC Central Committee in 2015, the concept of innovative, coordinated, green, open and shared development must be firmly established. Taking the improvement of ecological environment as the starting point, this paper systematically expounds the policies, measures and measures to promote our green development. The report of the 20th National Congress of the Communist Party of China put forward: "To promote green development and promote the harmonious coexistence between man and nature", "we must firmly comb and practice the concept that lucid waters and lush mountains are gold and silver mountains, and plan development from the height of the harmonious coexistence between man and nature. As the country gradually improves and attaches more importance to the concept of green development, listed companies also conform to the trend of The Times and the national development strategy to promote green development and achieve high-quality development of the economy and society.

4.3 Discussions on the Reasons of Diachronic Linguistics Variations of the Corporate Annual Reports

The variations in linguistic characteristics of the Management's Discussion and Analysis (MD&A) sections over the past 25 years have been described and summarized with the aid of AntConc. Given the complexity of language change, this paper provides a preliminary explanation for these variations.

In examining high-frequency words, it is important to note that the Chinese language has a limited number of synonyms for certain noun expressions. This inherent characteristic restricts the precision and specificity of terms used in annual reports, which serve as formal documents reflecting a company's financial performance and related information. Consequently, companies must present their content with clarity and accuracy. In contrast, Chinese verbs often possess multiple expressions conveying similar meanings. To avoid redundancy, managers frequently employ varied verbs to articulate the same ideas. Furthermore, the influence of contemporary social conditions necessitates that corporate high-frequency terms evolve in tandem with societal changes and national policies.

Overall, the transformation of discourse in the MD&A sections of corporate annual reports can be attributed primarily to textual features, the companies' disclosure tendencies, and broader social changes

5. Conclusion

This study investigates the discourse linguistic variation in the Management's Discussion and Analysis (MD&A) sections of corporate annual reports from 2007 to 2021, focusing on five-year intervals. Through word frequency analysis, significant linguistic changes over the 15-year period have been identified.

The relationship between language and society is bidirectional; language both reflects and influences societal developments (Wardhaugh, 1986: 23). Dittmar (1976: 238) emphasizes that speech behavior and social structure are in constant interaction, with language serving as a mirror to societal evolution.

To comprehensively explore these linguistic changes, the research utilizes the MD&A Corpus, which consists of the MD&A sections from 30 corporate annual reports compiled at five-year intervals. The analysis employs AntConc software to establish a framework for diachronic examination of language shifts over the past 15 years. The study initiates with a literature review on diachronic studies and corpus-based discourse analysis, both domestically and internationally. Chapter 3 outlines the research methodology and the corpus development, while Chapter 4 provides a detailed description of the established corpora and summarizes key findings across three analytical levels. From the word frequency analysis, notable terms such as 公司 (enterprise), 风险 (risk), 管理 (management), and 业务 (business) indicate fluctuations in frequency that correspond to evolving corporate priorities over time. This shift reflects the changing emphasis of companies in response to broader societal dynamics. The rapid advancement of the internet in China has enabled stakeholders to access more direct and immediate corporate information, further contextualizing these findings.

This study is pioneering in its examination of discourse variation in Chinese corporate annual reports, offering fresh perspectives for future research. By collecting and comparing linguistic data over a 15-year span, the study visualizes historical trends through tables, creating a comprehensive overview akin to geological interpretations of earth's evolution through rock strata. Given that much of the existing research on historical corpora focuses on English and adheres to frameworks established by Western scholars, this work contributes significantly to the field of modern and contemporary Chinese by providing a diachronic comparative analysis of corporate discourse.

Ultimately, this research highlights the intricate nature of language change and its reflection of societal transformations. While some aspects of these changes remain difficult to explain, the study lays foundational insights for further exploration into the complex causes of language evolution and standardizes expressions within corporate communication. Future inquiries should continue to examine the multifaceted phenomena of language change, enhancing our understanding of its role in reflecting societal shifts.

Declarations and Acknowledgment:

The author declares there is no conflict of interest.

This work was supported by the Guangdong Provincial Fund for the Philosophy and Social Science (under grant number GD21CWY05).

References

- Allan, B. (1991). *The Language of News Media*. Oxford: Basil Blackwell Ltd.
- Anthony, L. (2024). *AntConc (Version 4.3.1)* [Computer software]. Waseda University. Available from <https://www.laurenceanthony.net/software>
- Biber, D., Conrad, S., & Reppen, R. (1998). *Corpus Linguistics: Investigating Language Structure and Use*. Cambridge: Cambridge University Press.
- Brown, E. L., & Raymond, W. D. (2012). How discourse context shapes the lexicon: Explaining the distribution of Spanish F-/H- words. *Diachronica*, 29(3), 297-317.
- Campbell, L. (1999). *Historical Linguistics: An Introduction*. Cambridge, MA: MIT Press.
- Caserio, C., Panaro, D., & Trucco, S. (2019). Management discussion and analysis: A tone analysis on US financial listed companies. *Management Decision*, 57(6), 1287-1304.
- Conrad, S. (2002). Corpus linguistic approaches for discourse analysis. *Annual Review of Applied Linguistics*, 22, 143-158.
- Debras, C. (2021). How to prepare the video component of the diachronic corpus of political speeches for multimodal analysis. *Research in Corpus Linguistics*, 9(2), 129-145.
- Dittmar, N. (1976). *Sociolinguistics: A Critical Survey of Theory and Application*. London: Edward Arnold.
- Durnev, A., & Mangen, C. (2020). The spillover effects of MD&A disclosures for real investment: The role of industry competition. *Journal of Accounting and Economics*, 69(1), 101344.
- Hunston, S. (2002). *Corpora in Applied Linguistics*. Cambridge: Cambridge University Press.
- Jaspersen, J. G., Richter, A., & Zoller, S. (2020). Predicting earnings management from qualitative disclosures. *S&P Global Market Intelligence Research Paper Series*. Retrieved from <https://www.spglobal.com>

- Lee, J., & Park, J. (2019). The impact of audit committee financial expertise on management discussion and analysis (MD&A) tone. *European Accounting Review*, 28(3), 475-498.
- Leech, G. (1991). The state of the art in corpus linguistics. In K. Aijmer & B. Altenberg (Eds.), *English Corpus Linguistics: Studies in Honour of Jan Svartvik* (pp. 8-29). London: Longman.
- Li, P., Zhu, Q., & Zhou, G. (2013). Joint modeling of argument identification and role determination in Chinese event extraction with discourse-level information. *IJCAI*. Retrieved from <https://ijcai.org>
- Miller, D. G., & Biber, D. (2015). Evaluating reliability in quantitative vocabulary studies: The influence of corpus design and composition. *International Journal of Corpus Linguistics*, 20(3), 275-298.
- Odebrecht, C., Belz, M., Zeldes, A., Lüdeling, A., & Krause, T. (2017). RIDGES herbology: Designing a diachronic multi-layer corpus. *Language Resources and Evaluation*, 51(3), 755-779.
- Pajuste, A., Poriete, E., & Novickis, R. (2020). Management reporting complexity and earnings management: Evidence from the Baltic markets. *Baltic Journal of Management*, 15(1), 85-104.
- Pava, M. L., & Epstein, M. J. (1993). How good is MD&A as an investment tool? *Journal of Accountancy*, 175(3), 51-53.
- Quirk, R., & Greenbaum, S. (1973). *A University Grammar of English*. London: Longman.
- Sinclair, J. (1999). *Corpus, Concordance, Collocation*. Oxford: Oxford University Press.
- Tognini-Bonelli, E. (2001). *Corpus Linguistics at Work*. Amsterdam: John Benjamins Publishing Co.
- Wardhaugh, R. (1986). *An Introduction to Sociolinguistics*. New York: Basil Blackwell.
- Author Bio:** Chaowang REN is currently an associate professor of Guangdong University of Technology, with a Ph.D. in Applied Linguistics. His research interest includes corpus linguistics, discourse analysis, and metaphor, with a publication of over 20 articles on corpus-based discourse studies in peer-reviewed journals. Haolin ZHENG is currently a postgraduate of Guangzhou University of Technology, majoring in translation and interpreting. His research interest lies in corpus-based translation and machine-aided translation studies.