Original Paper

A Corpus-based Study on Lexical Features in the English Version of Report on the Work of the Government (2003-2023)

Binyi Ye[1]

[1] Ningbo University, Ningbo City, Zhejiang Province, China

Abstract

In the context of globalization, China's voice has been recognized by more and more countries. As the most important political document of China, The Report on the Work of the Government is the main content of translation research of political papers. Its translation provides a significant window for people worldwide to understand China comprehensively. Based on the corpus method, this paper analyzes the lexical features of the English version of Report on the Work of the Government (2003-2023) in terms of the average word length, the type-token ratio, keyword and high-frequency words to explore the image of China constructed, and offer experience for the translations of other publicity materials.

Keywords

Report on the Work of the Government, Corpus, Lexical features, English translation

The forty-five years since the reform and opening up and socialist modernization has been truly momentous with China's economic and political status reaching remarkable heights. Since the 20th National Congress, Xi Jinping has attached great importance to accelerating the construction of China's international discourse system to convey China's stories globally in the new era. As a political paper, the *Report on the Work of the Government* (hereinafter referred to as the *Report*) is one of the most authoritative sources for the outside world to understand China's situation and political ideas, while its English version is an important part of China's publicity materials to help the world understand China

more comprehensively and objectively.

1. Literature Review

In recent years, there has been a boom of research on the translation of the *Report* carried by domestic scholars. Among them, Wu Guangjun compares the specific differences between the *Report* and US's *State of the Union Address* (hereafter referred to as *Address*) at the lexical and syntactic levels, and concludes the stylistic features of the *Report* (2011); Di Yanhua, Yang Zhong, Li Xiaoqian, Hu Kaibao and Shi Honghong focus on the keywords of the *Report*, and examine their collocation structure. Hu Kaibao and Tian Xujun further expand the scope of their study to analyse the linguistic features and textual effects of the *Report* (2000-2016) from the perspectives of words' frequency and word class. Hu Yuanjiang identifies the collocation characteristics of the *develop* and its family and explores the interaction and influence of language and politics in the annual reports of the Chinese and American governments. Si Bingyue, Li Yong and Wu Meixuan, analyze nominalization and the application of model verbs. Wang Liyang, Huang Weiqi and other scholars focus on the lexical and syntactic differences of the Report (2016-2020).

The *Report* is of time-sensitive and reflects the direction of China's development. However, the existing researches are based more on materials in small quantity with a limited time span. There is little of insight in the English translation of the *Report*. Therefore, this paper applies the corpus method, with the assistance of the analysis software AntConc4.2.0, to conduct a comparative analysis of the lexical features of the *Report* (2003-2023) and the *Address* (2003-2023) in the past two decades. The aim is to better understand the different standpoints and attitudes expressed in the *Report* and the *Address*, and to provide reference for the translation of Chinese external publicity texts in the future.

2. Research Design

2.1 Corpus Data

The Report on the Work of the Government is an official document issued by the State Council to summarize and plan the work of the government on an annual basis, covering important issues such as economic development, domestic and foreign affairs, and the people's livelihood. The *State of the Union Address* is the annual report of the President of the United States to Congress. Although there are some differences in the number of tokens, they are still comparable.

This paper is based on two self-built special corpora. One consists of 21 reports on the work of the government (about 316,100 words) in English version from 2003 to 2023 and another one contains 21 articles of the *State of the Union Address* of the United States (about 129,700 words) in the same period as the comparable corpus. Then here takes the CLOB, a member of the Brown family corpora, as the reference corpus to get the keyword list of the above 2 special corpora separately. The CLOB includes more than 1 million words. Although the total number of words in the two comparative corpora is different, they are somewhat comparable due to their types (both political documents) while they are all full text without tags.

64

The texts are obtained respectively from the website of the Central People's Government of the PRC and The White House. The general description of the two self-built corpora is shown in Table 1:

Table 1. General Description of the Data Selected

	Report on the Work of the Government	*State of the Union Address*
Time Span	2003—2023	2003—2023
Text_Count	21	21
Token_Count	316,107	129,700
In Total	445,807	

2.2 Methodology

This paper combines quantitative and qualitative research methods. Firstly, the corpus analysis tool AntConc4.2.0 developed by Lawrence Anthony is used to extract the basic information of the two self-built corpora each and generate two word lists to compare the general features, such as the average word length, Type-Token Ratio (TTR), and high-frequency words.

Then, take the CLOB as a reference corpus to extract the top 20 keywords of two self-built corpora each (Table 5). Different keywords represent different themes of the governments' work and focus. By comparing the two political documents, we can explore the differences in language collocation and political emphasis between the two political documents.

Later, according to the current social hot spots in the construction of ecological civilization, this paper selects the relevant words in the *Report*, such as *green* and *ecology*, and then extracts and analyzes their collocation, so as to verify the accuracy of the translation, and provide experience for the translation of relevant topics in the future.

3. Results and Analysis

According to the general characteristics of the words in the texts, this paper makes a comparison and specific analysis between the English version of the *Report* and the *Address* from five aspects: average word length, type-token ratio, keyword and frequency.

3.1 Average Word Length

The average word length is the ratio of the total number of letters to the total number of words in a text, which can reflect the stylistic characteristics of the text. The standard deviation of word length is the degree to which the word length fluctuates around the average word length. Formal texts often use long words to highlight the stylistic features of their texts in order to distinguish them from informal or colloquial texts. For example, the five-year average word length for the three different text categories in the Brown corpus is 4.73 for newspaper texts (Category A, B, C), with a standard deviation of 2.59; 4.93 for academic texts (Category J), with a standard deviation of 2.87; and 4.31 for fiction (Category

K, L, M), with a standard deviation of 2.18. The academic category is the most formal of the three, followed by the newspaper category, and then by the fiction category (Chen & Cui, 2010). This shows that there is a close relationship between the average word length and the degree of formality of the text. The following Table 2 shows the average word length and standard deviation of the two corpora of the *Report* and the *Address*.

Table 2. Average Word Length

	Report	*Address*
Average Word Length	5.47	4.61
Standard Deviation	2.98	2.45

The statistics in Table 2 illustrate that the average word length of the *Report* are higher than the *Address*, indicating that the Chinese report is more formal than the latter.

3.2 Type-Token Ratio (TTR)

Type-token Ratio (TTR for short) refers to the ratio of type to token in a specific corpus. The number of type is the number of different word forms and the number of token is the amount of words in the text. The mere type or token cannot reflect the essential feature of the text, but the ratio of the two can reflect the variability of words (Yang, 2002). Table 3 shows the information of the type, token, TTR and standardized TTR analyzed by AntConc4.2.0.

Table 1. Type-Token Ratio

	Report	*Address*
Token_Count	316,107	129,700
Type_Count	8,682	9,226
TTR	2.75	7.11
Standardized TTR（%）	29.31	32.07

From Table 3, we can see the 2 corpora's TTR are prominent indifferent. Due to the amounts of texts are relatively large in number, the mere TTR of the whole corpus can not scientifically reflect the difference in word richness so the Standardized TTR is taken as the basis. The standardized TTR of the *Report* (29.31) is smaller than that of the one in native language (32.07). On the one hand, this outcome reflects that native English speakers have larger vocabularies and use more various vocabularies; On the other hand, it shows that there are more replacement of frequently occurring words in actual use. For instance, in the *Address* (2018), the word *storm* is then replaced by *hurricane*. However, the translation of political materials in our country requires the use of strict terms and fixed collocations, which will lead to the repetition of synonyms and collocations.

For examples: (here O refers to Original; T refers to Translation)

O: "三农"

T: *Agriculture, rural areas and farmers*

O: "四个全面"

T: "*The Four-pronged Comprehensive Strategy*" (China's strategic plan for building socialism with Chinese characteristics, that is, to make comprehensive moves to complete a moderately prosperous society in all respects, to further reform, to advance the rule of law, and to strengthen Party self-governance).

In short, the diversity and variation of the language applied in the English version of the *Report* are generally lower than in those the *Address* written by native speakers. However, because of the repeated use of difficult terms, the English version of the *Report* is more difficult to understand than the *Address*.

3.3 Keyword

Keyword refer to the words with greater discrepancy in frequency in the target corpus compared with the reference corpus, and the more significant the discrepancy, the stronger the keyness. By examining the linguistic features of keyword, we can effectively grasp the theme of the texts as a whole. This paper applies the keyword function of AntConc4.2.0 and takes the CLOB corpus as the reference corpus to extract the keyword of the *Report* and the *Address* from 2003 to 2023 respectively. Table 4 and Table 5 are obtained:

Table 4. Keyword List of the *Report*

Table 5. Keyword Lists of the 2 Corpora

Rank	Report	Frequency	Percentage(%)	Range	Address	Frequency	Percentage(%)	Range
1	we	8714	2.76	100	the	7619	5.87	100
2	and	21751	6.88	100	and	4872	3.76	100
3	will	6046	1.91	100	to	5481	4.23	100
4	development	2674	0.85	100	we	3582	2.76	100
5	rural	1287	0.41	100	of	3509	2.71	100
6	improve	1373	0.43	100	in	2798	2.16	100
7	reform	1246	0.39	100	a	2751	2.12	100
8	economic	1265	0.40	100	is	1092	0.84	100
9	areas	1091	0.35	100	that	1869	1.44	100
10	china	1005	0.32	100	our	1218	0.94	100
11	system	1341	0.42	100	for	1827	1.41	100
12	promote	779	0.25	100	I	1617	1.25	100
13	government	1537	0.49	100	have	1765	1.36	100
14	strengthen	712	0.23	100	it	1281	0.99	100
15	enterprises	652	0.21	100	's	2184	1.68	100
16	yuan	583	0.18	100	you	843	0.65	100
17	increase	811	0.26	100	this	597	0.46	100
18	urban	680	0.22	100	on	1003	0.77	100
19	ensure	718	0.23	100	American	798	0.62	100
20	work	1362	0.43	100	will	1894	1.46	100

Table 5 illustrates that there are 19 content words (95%) in the list of the *Report*, and only 1 function word (5%), while there are only 11 content words (55%) and 9 function words (45%) in the *Address*.

Figure 1. Lexical Comparison

According to the statistics in the table 5 above, the keyword of the *Report* includes *development, rural, improve, reform, economic, china, system, government, strengthen, enterprises*, etc. while the one of the *Address* is *the, and, to, we, of, in, a, our, I, have, American, will*, etc. The difference between the two lists partly reflects that the focus of the *Report* is more concentrated and clear than that of the *Address*. By focusing on the content words appearing in the top 100 list in the *Report*, we can get a rough idea of the work priorities of the Chinese government from 2003 to 2023. Some of them are summarized in the following Table 6:

Table 2. List of Top 100 Content Words of the *Report*

we	implement	growth	accordance	promoting
will	develop	services	governments	stability
development	policies	policy	economy	china's
rural	central	regions	grain	new
improve	efforts	law	residents	ethnic
reform	our	progress	cultural	industrial
economic	chinese	should	improved	macao
areas	basic	encourage	public	income
china	investment	oversight	medical	fully
system	industries	deepen	administrative	supply
promote	major	employment	million	consumption
government	social	accelerate	year	technological
strengthen	people	billion	energy	security
enterprises	education	insurance	opening	tax
yuan	agricultural	production	strengthened	improving
increase	market	must	trade	interests
urban	cooperation	agriculture	percent	socialist
ensure	support	measures	quality	poverty
work	innovation	stable	overseas	programs
continue	projects	expand	national	capacity

Figure 2. A Wordcloud of the Keyword of the *Report*

Table 6 shows that the Chinese government is concerned about the people's livelihood, economy, development and other issues, especially rural issues. The content words in the top 100 high-frequency keywords of the *Address* can also reflect the focus of the US government (not listed here), but the

69

frequency and percentage of these content words in the top 100 high-frequency words indicate that the issues concerned by the *Address* are more, and more scattered. In addition, the pronouns *I* and *you* appear in the high-frequency words of the *Address*. Although both of them are obviously less than the *Report*, which belongs to the political discourse style, the *Report* is more formal the *Address*, which shows a certain tendency of colloquial language.

Meanwhile, we can see that in the *Report* (2003-2023), *we* ranked first, which could shorten the distance between the government and the people, indicating that China has always adhered to the purpose and principle of "people-oriented", and made due contributions to creating a harmonious society. The word *will* is mostly used in the future tense to indicate the work plan contained in the report and to express the nation's strong expectations of future development and progress. The words *we* and *will* also appear in the *Address* (2003-2023), with *we* ranking second, but relatively less thematic than the *Report*. In addition, *development*, *reform* and *China* are highly thematic in the R*eport*, which shows that China, as the largest developing country, attaches great importance to reform and development in economy, education and other aspects.

3.4 Selected Words

In 2022, the report of the 20th National Congress clearly points that nature is the basic condition for the survival and development of human being. Respecting, adapting to and protecting nature are inherent requirements for establishing a modern socialist country in an all-round way. Based on the demands of ecological civilization construction mentioned in the Five-Sphere Integrated Plan, China is currently focusing on ecological construction with the assets of "lucid waters and lush mountains are invaluable" and the goals of "reaching peak carbon emissions and carbon neutrality".

Therefore, this paper chooses words related to ecological construction, such as *green* and *ecological*, to explore the attention paid to ecology in the *Report*, and apply the corpus tool AntConc4.2.0 to query word frequency and retrieve its relevant collocation.

Table 3. Figures of Selected Words concerning Ecology

Word	Rank	Freq	Range	NormFreq	NormRange
environment	1	309	21	585227.273	1
environmental	2	188	21	356060.606	1
ecological	3	139	21	958620.69	1
sustainable	4	57	21	919354.839	1
green	5	56	14	1000000	0.667
carbon	6	34	9	1000000	0.429
circumstances	7	27	15	450000	0.714
environments	8	18	12	34090.909	0.571
environmentally	9	13	7	24621.212	0.333

ecology	10	3	3	20689.655	0.143
gynecological	11	2	2	13793.103	0.095
ecologically	12	1	1	6896.552	0.048

In this paper, several words with high correlation with ecological construction are selected, here comes the Table 7, which illustrates that China attaches great importance to ecology. Compared with traditional words such as *environment* and *ecological*, many a relatively new words such as *sustainable*, *green* and *carbon* have emerged successively.

In this section, we continue to use the corpus analysis tool AntConc4.2.0. In this software, In this paper, I use the Likelihood measurement method to take statistical tests, setting the minimum collocation frequency of node words as 1, with a span of 5-5. Then I choose two words—"ecological" and "green"— for index and select the top 20 collocation words as the main objects.

The word cloud map and related word lists are as follows:

Figure 3. Word Cloud of the Collocation

Table 4. Lists of Collocation of *Ecological* and *Green*

Word	*ecological*		*green*	
Rank	Collocate	Likelihood	Collocate	Likelihood
1	conservation	417.411	development	65.371
2	protection	143.392	carbon	60.51
3	environmental	121.807	smart	36.576
4	environment	92.111	low	35.242
5	projects	69.752	circular	29.338
6	improvement	66.916	finance	27.333
7	damage	63.376	conservation	24.635

8	compensation	48.395	develop	23.951
9	compensating	36.512	appliances	22.551
10	political	32.119	environmentally	22.551
11	restoration	29.025	advances	22.496
12	sound	24.434	emissions	22.246
13	forestry	23.041	packaging	21.03
14	strengthen	22.877	specializing	21.03
15	progress	22.617	green	20.784
16	grassland	21.883	big	19.991
17	emissions	19.726	blue	19.97
18	for	19.151	element	19.342
19	and	19.097	home	19.274
20	cultural	19.058	gas	19.161

Table 5. Index of KWIC of two Selected Words

By comparing the matching words with the text, we can find that China has added "strengthening

72

policy support" to the conventional environmental protection, and added "reducing the emission of carbon", "green development" and "environment-friendly" to the content of measures such as "ecological engineering construction". In addition to measures such as "construction of ecological projects", new contents such as "carbon reduction", "green development" and "environmental friendliness" have been added. We are required to plan our development on the basis of the original assertion that "lucid waters and lush mountains are invaluable" and in harmony with nature.

On the new journey of building a modern socialist country, China insists on promoting the construction of a beautiful China, ecological protection and systematic governance, coordinating industrial restructuring, pollution control and combating climate change, and promoting ecological priority, conservation and intensive, green and low-carbon development. This shows the goal and attitude of China towards environmental protection and high-quality development in the future. Compared to the translations of other aspects, the ecology-related translations tend to be simplified, with more concise wording and a tendency to translate directly from Chinese, which on the one hand indicates that China is increasingly having an ecological discourse in the international arena and its international influence is expanding. On the other hand the translations carry obvious Chinese characteristics, indicating China's adherence to the path of cultural self-confidence and opening up new ideas for he translation of other publicity materials.

4. Conclusion

Both our *Report* and the *Address* are among the most important political documents for promoting international communication. By analyzing and comparing the lexical features of the two political documents (2003-2023), and focusing on the ecological usage of words and collocations in the *Report* itself, this paper aims to understand the political, economic, cultural and social differences between the two countries more directly through a corpus perspective in order to better grasp the current development direction of China. The study found that: 1) the textual style of the *Report* is more formal than the *Address*, but the richness of vocabulary is also limited by the use of fixed expressions in translations; 2) there is a greater difference in the TTR between the *Report* and the *Address*; 3) the political focus of China and the US is different, with China focusing on development; 4); 4) the vocabulary used in the translation of China's ecological civilization is different from that used in traditional publicity texts with Chinese characteristics.

All in all, there is still a long way to go for the translation and promotion of publicity materials in China. This paper draws on the rational analysis of the corpora figures and summarizes some of the characteristics of existing translations. However, the practice of English translation of political literature still lacks systematic theories and practical exploration, which is an important factor of China's image shaping and needs to be further developed and improved.

Reference

Chen, J. S., & Cui Y. N. (2010). A Study on the Vocabulary Characteristics of the English Translation of the Chinese Government Work Report Based on Corpus. *Contemporary Foreign Language Research*, *2010*(06), 39-43+62.

Chen, Y., & Ruan, H. (2023). A Study of Corpus-Based Translation for China's Global Communication from the Perspective of Critical Discourse—A Case Study of the 20112022 Report on the Work of the Government. *International Journal of Languages, Literature and Linguistics*, *9*(3). https://doi.org/10.18178/IJLLL.2023.9.3.411

Di, Y. H., & Yang, Z. (2010). Corpus based research on core theme words in Chinese government work reports. *Journal of Foreign Languages*, *2010*(06), 69-72.

Hu, K. B., & Tian, X. J. (2018). A Study on the Language Characteristics and Text Effects of the English Translation of the Government Work Report: A Corpus Based Study. *Foreign Languages*, *34*(05), 1-11.

Laviosa, S. (1998). Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. *Meta: journal des traducteurs/Meta: Translators' Journal*, *43*(4). https://doi.org/10.7202/003424ar

Liang, M. C., Li, W. Z., & Xu, J. J. (2010). In *Corpus Application Tutorial*. Beijing: Foreign Language Teaching and Research Press.

Liu, M. (2023). Corpus assisted construction of China's ecological image. *Modern Foreign Languages*, *46*(01), 83-96.

Luo, X. Q. (2017). In *A Study on the Language Feature Differences between English and Western Translations of Chinese Political Texts Based on Corpus*. Beijing Foreign Studies University.

Mona, B. (1993). *Corpus: Linguistics and Translation Studies: Implications and Applications*. Bonelli Amsterdam: Pholadelphis Hohn Benjam Ins.

Mona, B. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target*, *7*(2), 223-243. https://doi.org/10.1075/target.7.2.03bak

Paul, B. (2008). A useful methodological synergy? Combining critical discourse analysis and corpus linguistics to examine discourses of refugees and asylum seekers in the UK press. *Discourse & Society*, *19*(3). https://doi.org/10.1177/0957926508088962

Shi, X. P., & Ma, Q. Y. (2022). Discourse Strategies in the Construction of National Image from the Perspective of Attitude Resources: A Case Study of Report on the Work of the Government in 2022. *Journal of Social Science and Humanities*, *4*(9). https://doi.org/10.53469/jssh.2022.4(09).23

Wang, Y., Zhang, R. E., Han, M. L. (2022). Research on China's National Image in the English Translation of the Government Work Report: A Corpus based Critical Discourse Analysis. *Journal of Beijing University of Chemical Technology (Social Sciences Edition)*, *2022*(04), 71-80+111.

Wu, G. J. (2014). Analysis of stylistic features of the English translation of the Chinese Government Work Report based on corpus. *Journal of Xi'an Foreign Studies University*, *22*(04), 118-121.