A Contrastive Corpus Study on Lexical Features of the English Translation of the Report of the 20th and 19th CPC National Congress

Yiyu Zhang[1], Lin Wang[1], Wenming Hu[1], and Xiajing Yao[1,*]

[1]School of Foreign Languages, China University of Geosciences, Wuhan, China, 430074

Abstract. As the channel spreads the Chinese voice and promotes Chinese culture, the status of political document translation is self-evident with the deepening of global communication. Taking the English translation of the reports of the 20th and the 19th CPC National Congress as examples, this paper makes a contrastive study of their lexical features in a corpus-based way. With the assistance of corpus analysis software such as AntConc, TagAnt and WordSmith4.0, the author investigates TTR (type-token ratio), lexical density, average sentence length, high-frequency words, and keywords between two reports. It is found that the report of the 20th CPC Nation Congress has more diversified and native expressions with fewer words used, which provides a guiding significance for the English translation of political documents.

1 Introduction

Delivered by President Xi Jinping, adopted at the 20th CPC National Congress and translated by the Central Institute of Party History and Literature, the document titled "*Hold High the Great Banner of Socialism with Chinese Characteristics and Strive in Unity to Build a Modern Socialist Country in All Respects*" (hereinafter referred to as "the report of the 20th CPC National Congress") further charted the way forward for the cause of the Party and the country, focusing, in particular, on the Party's strategic tasks in the coming five years.

As an authoritative document that serves as a political manifesto, the report of the 20th CPC National Congress played a vital guiding role in uniting and leading the Chinese people of all ethnic groups to uphold and develop socialism with Chinese characteristics on the new journey in the new era. To ensure the key points and content are accurate and understandable to target readers with different cultural backgrounds and language habits, China has invited foreign experts to help edit translated versions of the work [1].

It is worth noting that considerable attention has been paid to corpus-based political document translation research, to name a few, Jia Hui and Sun Minwei discussed translation innovations in the report of the 19th CPC National Congress [2], Wu Yana, Xie Yating, and Xia Mengjun studied explicitation techniques in the English translation of the report of the 19th CPC National Congress [3], and Shang Wenbo focused on the image of the CPC through corpus-based critical translation studies [4]. However, the research object needs to be updated over time and studies on lexical features are still insufficient.

Given that the importance of its English translation is universally recognised, the paper takes the English translation version of the reports of the 20th and the 19th CPC National Congress as research objects, exploring differences between them based on a corpus approach with the help of computer software, then discussing their lexical features by contrasting the research data. Through qualitative and quantitative analysis, it is hoped that some references can be provided for the English translation of political documents.

2 Research method

In this research, the report of the 20th and 19th CPC National Congress from *China Daily* (titled with *Full text of the report to the 20th National Congress of the Communist Party of China* and *Full text of Xi Jinping's report at 19th CPC National Congress* respectively) are selected as research objects [5-6].

The first step in the building process of corpora is text pre-processing, where noise reduction and text normalization take place. Rich text downloaded from *China Daily* websites is cleaned manually and converted into plain text. Then TagAnt is used for POS (Part-of-Speech) tagging so that AntConc, by which lexical density is calculated, can be further used. Apart from the functions mentioned above, WordSmith4.0 is also used to analyse the lexical features of both English and Chinese text in the current study. It is an integrated toolkit of modules comprising three functions: WordList, Concord, and KeyWord. The first module, WordList, as the most advocated program for data extraction, provides characteristics observed in the untagged text, including tokens, types, Standardized type-token ratio (STTR),

* Corresponding author: yaoxiajing@cug.edu.cn

mean sentence length, and a list of all words in frequency order. Noting that all text imported into AntConc and TagAnt utilises the UTF-8 character encoding method while all text imported into WordSmith utilises the UTF-16LE character encoding method for the best compatibility.

3 Statistics and analysis

3.1 Type-token ratio

Type-token ratio (TTR) is the ratio obtained by dividing the types (the total number of different words) occurring in a text or utterance by its tokens (the total number of words). A high TTR indicates a high degree of lexical variation while a low TTR indicates the opposite [7]. TTR scores were calculated as follows:

$$\text{TTR} = \frac{\text{Number of different words (types)}}{\text{Total number of words (tokens)}}$$

Standardised type-token ratio (STTR) calculates the mean TTR for every N-words (N is generally set to 1,000 by default but can be changed according to different needs). Those two kinds of ratios reflect the degree of word occurrence in the corpus. However, the basic problem of TTR is that it is affected by the length of the text sample. For larger text, TTR is rather meaningless in most cases. Therefore, the author uses the standardised type-token ratio to analyse two reports. The higher the standardised type-token ratio is, the more easily the word varies, and vice versa. Noting that "NCCPC" in the table is the abbreviation of "the National Congress of the Communist Party of China".

Table 1. Tokens and types

	The report of the 20th NCCPC (target text)	The report of the 19th NCCPC (target text)
Types	3,095	3,029
Tokens	25,346	25,059
Type-Token Ratio	12.21%	12.09%
Standardised Type-Token Ratio	39.77%	39.22%

From table 1, the TTR of the English translation of the report of the 20th and the 19th CPC National Congress is 12.21% and 12.09% respectively. However, the two reports are quite different in length, making TTR have less reference value. Therefore, the paper concentrates on the STTR of the two materials. The STTR data of the report of the 20th CPC National Congress is 39.77%, and the STTR data of the report of the 20th CPC National Congress is 39.22%. The standardised type-token ratio of the English version of report of the 20th CPC National Congress is higher than that of the 19th CPC National Congress report. This shows that the use of English vocabulary in the 20th CPC National Congress report has more variability than that of the 19th CPC National Congress translation, and the scope of vocabulary change is wider than that of the 19th CPC National Congress translation report. This also indicates that the translated version of the 20th CPC National Congress report is more expressive, which is conducive to readers' understanding.

For example, there are four different translation versions of Chinese word "*guan1jian4*" in the report of the 20th CPC National Congress:

(1) It takes place at a **critical** time as the entire Party and the Chinese people of all ethnic groups embark on a new journey to build China into a modern socialist country in all respects and advance toward the Second Centenary Goal.

(2) We have grown stronger in basic research and original innovation, made breakthroughs in some core technologies in **key** fields

(3) The next five years will be **crucial** for getting our efforts to build a modern socialist country in all respects off to a good start.

(4) Our Party **has a pivotal role** in building China into a modern socialist country in all respects and in advancing the rejuvenation of the Chinese nation on all fronts.

However, for the same Chinese word "*guan1jian4*", there are only two translation versions in the report of the 19th CPC National Congress:

(1) The 19th National Congress of the Communist Party of China is a meeting of great importance taking place during the decisive stage in building a moderately prosperous society in all respects and at a **critical** moment as socialism with Chinese characteristics has entered a new era.

(2) We have grown stronger in basic research and original innovation, made breakthroughs in some core technologies in **key** fields.

As is demonstrated by the comparison, vocabulary changes of the 20th CPC National Congress report translation are more diverse.

3.2 Lexical Density

Lexical density is defined as content words (lexical words) divided by the total number of words, which estimates the amount of information in the text [8]. Content words in English include nouns, verbs, adjectives, and adverbs; Function words include pronouns, conjunctions, numerals, mood words, and interjections. Content words indicate the difficulty and amount of information in the corpus. Lexical density is calculated as follows:

$$L_d = \frac{\text{Content words (lexical words)}}{\text{Total number of words}}$$

The more varied a vocabulary a text possesses, the higher lexical diversity. For a text to be highly lexically diverse, the writer has to use many different words, with little repetition of the words already used [9]. In this part, the author studies the lexical density of the translation text to explore lexical features.

Table 2. Lexical density

	The report of the 20th NCCPC (target text)	The report of the 19th NCCPC (target text)
Nouns	7,581	7,472
Adverbs	674	600
Verbs	3,958	4,003
Adjectives	3,164	3020
Content words	15,377	15,095
Function words	25,346	25,059
Lexical density	60.67%	60.24%

According to table 2, the lexical density of the translation material of the 20th CPC National Congress report is 60.67%, and the 19th CPC National Congress report is 60.24%. The lexical density of the English translation of the 20th CPC National Congress report is larger than the 19th CPC National Congress one. It means there are more content words in the 20th CPC National Congress report translation text. In other words, the 20th CPC National Congress report translation text is more concise and more readable.

3.3 Average Sentence Length

Average sentence length is calculated by dividing the total number of words in the text by the total number of sentences. The average sentence length is positively correlated with the syntactic structure of the translated text. Standard deviation reflects the degree of concentration and discreteness of the data, and the larger the standard deviation, the greater the variation in sentence length of the translated text. The data in table 3 is based on WordSmith calculations.

Table 3. Average sentence length

	The report of the 20th NCCPC (target text)	The report of the 19th NCCPC (target text)
Sentences	930	975
Mean (in words)	26.81	25.87
Std.dev.	14.21	17.37

In accordance with the data in table 3, the standard deviation of average sentence length in the 20th CPC National Congress report translation is larger than the translated text of the 19th CPC National Congress report. It indicates the length of sentences in translation material of the 20th CPC National Congress report has a richer variety of forms. The result is also corroborated by the previous statistics of type-token ratio: the vocabulary in the translation of the report of the 20th CPC National Congress is more varied and abundant.

For the same Chinese sentence "*qiang2diao4 jian1ding4 dao4lu4 zi4xin4 li3lun4 zi4xin4 zhi4du4 zi4xin4 wen2hua4 zi4xin4*" there are two different translations in two reports:

(1) ...strengthen our confidence in the path, theory, system, and culture of socialism with Chinese characteristics. (Abridged from the 20th CPC National Congress report translated text)

(2) It highlights the importance of fostering stronger confidence in the path, theory, system, and culture of socialism with Chinese characteristics. (Abridged from the 19th CPC National Congress report translated text)

After comparing the different translations of the same sentence in the 19th and 20th CPC National Congress report, it can be concluded that in the same sentence, the translation of the 20th CPC National Congress report is shorter than that of the 19th CPC National Congress report, which can also be concluded that the translation of the 20th CPC National Congress is more concise.

3.4 High-frequency words

Table 4 below shows the top twenty high-frequency words in the English translation of the report of the 20th CPC National Congress and that of the 19th CPC National Congress, which indicates that high-frequency words in two reports are "and", "the", "of", "to" and "in". In addition, two reports, both of which focus on Party, people, China, development and so on, maintain the consistency of politics. All high-frequency words are closely related to contents, highlighting the topic of these reports.

Table 4. High-frequency words

	The report of the 20th NCCPC (target text)		The report of the 19th NCCPC (target text)	
Rank	Freq	Word	Freq	Word
1	1832	and	1700	and
2	1541	the	1600	the
3	866	of	951	of
4	702	we	693	to
5	684	to	573	we
6	517	in	483	in
7	493	will	424	a
8	388	a	344	will
9	280	for	327	party
10	260	party	272	people
11	248	s	262	our
12	238	people	258	for
13	220	our	239	with
14	209	with	218	s
15	182	china	207	that
16	181	development	195	chinese
17	179	that	184	is
18	172	have	173	china
19	165	new	162	development
20	156	all	158	must

It worth noting that the frequency of the word "development" in the report of the 20th and 19th CPC National Congress are 181 and 162 respectively, showing

that the report of the 20th CPC National Congress is more concerned with the development issue. Especially considering the ongoing COVID-19 global outbreak, policymakers in China are addressing the pandemic's lasting effects and making efforts to spur green, resilient and inclusive growth while safeguarding macroeconomic sustainable development. There are 149 extra modal verbs "will" in the report of the 20th CPC National Congress, demonstrating that the CPC is positively seeking opportunities to provide guidance for the public and serve the country's overall interests.

3.4.1 Modal auxiliary verbs

Modal auxiliary verbs reflect the distance, social relationship, and status between the two sides of communication, revealing different power relations behind the utterance. Halliday and Hasan (1989) divided modal auxiliary verbs in terms of their pragmatic values [10]. High-value modal auxiliaries include "must", "ought to", "need", and "have to"; median-value modal auxiliaries include "will", "would", "shall", "should"; and low-value modal auxiliaries include "may", "might", "can", and "could". Different groups of modalities are related to different politeness degrees of speech. Modal auxiliary verbs of high value indicate impolite speech, which is liable to cause the reader's disfavour, whereas low-value modals suggest a most polite use of language.

In this research, the tagged text is imported into the concord tool of WordSmith, by which frequencies of modal auxiliaries are calculated. To simplify the analysis of results, all modal auxiliary verbs in two reports are classified into three value categories according to Halliday's theory. The following table illustrates the frequency of modal auxiliary verbs in two reports. Noting that frequencies are calculated per 1,000 words.

Table 5. Modal auxiliary verbs

Values	Modal auxiliary verbs	The report of the 20th NCCPC (target text)		The report of the 19th NCCPC (target text)	
		Hits	Freq (per 1,000)	Hits	Freq (per 1,000)
Low	can	16	0.64	21	0.85
	may	5	0.20	1	0.04
	could	1	0.04	0	0.00
	might	0	0.00	0	0.00
	dare	2	0.08	1	0.04
	have to	0	0.00	1	0.04
	Total	24	0.97	24	0.97
Median	would (wouldn't)	0	0.08	1	0.04
	should (shouldn't)	35	1.41	93	3.77
	is/was to (isn't/wasn't to)	5	0.20	9	0.36
	Total	40	1.69	103	4.18
High	must (mustn't)	108	4.35	158	6.41
	ought to (oughtn't to)	0	0.00	0	0.00
	need	5	0.20	12	0.20
	can't	0	0.00	0	0.49
	couldn't	0	0.00	0	0.00
	mayn't	0	0.00	0	0.00
	Total	113	4.55	170	7.09

The analysis of the results indicates a lower frequency of median-value and high-value modal auxiliaries in the 20th CPC National Congress report translation, with a total of 40 median-value modal auxiliaries and 113 high-value modal auxiliaries compared to the 103 median-value and 170 high-value found in the 19th CPC National Congress report translation. Besides, low-value modal auxiliaries, the quantity of which is 24, are consistent in the two reports.

High-value and median-value modal auxiliary verbs are less used in the 20th CPC National Congress report translation in number and frequency, mainly because the Chinese government is committed to building its own internationally friendly and positive image. On the one hand, less modal auxiliary verbs can be gentler and more comfortable for the audience, and on the other hand, using modal auxiliary verbs is a way to call for or express strong feelings. Given that most policies in the 20th CPC National Congress are sticking to prior trends and calls for change appear less than before, as a result, fewer modal auxiliary verbs are used.

3.4.2 Delexical verb "make"

The macro indicators discussed above only scratch the surface of lexical features. Little attention has been paid to specific linguistic items in the translated texts. Thus, this paper further discusses lexical features from the micro level of the delexical verbs to remedy this situation.

Many common words have lost their semantic meaning as a result of the complex mechanism known as delexicalization, and now they must rely on their co-occurring partners to provide meaning [11]. Together, the delexicalised words and their co-occurring partners communicate a common meaning by influencing each other's semantic content.

The author discovers that "have" and "make" have startlingly high frequencies by hesitantly exploring the occurrences of the most prevalent lexical verbs, such as have, get, give, take, make, and do, in WordSmith WordList and Concord modules. Additionally, a more thorough examination (i.e., searching "have/made + noun" structure) reveals that "have" acts as a delexical verb far less frequently than "make". In other words, the word "have" usually has a concrete meaning rather than a phraseological one. Therefore, only the more typical delexical is focused on.

To further instigate, the original Chinese texts of instances where delexicalised "make" is used in interpreted texts are analysed. Instances where the Chinese delexicalised verbs "*zuo4(1)*", "*zuo4(2)*", "*gao3*", and "*jin4xing2*", are interpreted into English delexicalised verb "make" are collected, with the result shown in table 6-1, 6-2 and 6-3:

Table 6-1. Typical Chinese delexicalised verbs in source text

Typical Chinese delexicalised verbs in source text	Number of them in the report of the 20th NCCPC (source text)	Number of them in the report of the 19th NCCPC (source text)
zuo4(1)	18	23
zuo4(2)	130	145
gao4	3	4
jin4xing4	3	12
Total	154	184

Table 6-2. Total number of delexicalised "make" in target text

Total number of delexicalised "make" in target text	
The report of the 20th NCCPC (target text)	The report of the 19th NCCPC (target text)
64	69

Table 6-3. Frequency of delexicalised "make" interpreted from typical Chinese delexicalised verbs in target text

Frequency of delexicalised "make" interpreted from typical Chinese delexicalised verbs in target text	
The report of the 20th NCCPC (target text)	The report of the 19th NCCPC (target text)
41.56%	37.50%

The three tables above show clearly that the delexical verb "make" is more frequently used in the 20th CPC National Congress report translation. It is found that the frequent use of Chinese delexicalised verbs in source text inevitably influences the diction of target text, resulting in the high frequency of delexicalised verbs in both two English translations. The author also believes that the heavy use of delexicalised verbs with collocates of abstract noun phrases is a significant feature of English diction, translators use delexicalised verbs frequently to collocate with different noun phrases to meet the readers' expectation of fully understanding the message. Consequently, target text of the 20th CPC National Congress report tends to be more native than the previous one.

3.5 Keywords

The Brown corpus, the full name of which is Brown University Standard Corpus of Present-Day American English, was the first text corpus of American English. The original Brown corpus was published in 1963-1964 by W. Nelson Francis and Henry Kučera (Department of Linguistics, Brown University Providence, Rhode Island, USA). Taking the Brown corpus as a reference, the words in the two English translations are retrieved.

Table 7. Keywords

	The report of the 20th NCCPC (target text)		The report of the 19th NCCPC (target text)	
Rank	Type	Freq	Type	Freq
1	and	1832	and	1700
2	we	702	we	573
3	will	493	will	344
4	party	260	party	327
5	s	248	people	272
6	people	238	our	262
7	our	220	with	239
8	development	182	s	218
9	china	181	chinese	195
10	have	172	china	173
11	new	165	development	162
12	chinese	156	must	158
13	all	156	have	149
14	system	130	all	136
15	improve	117	new	125
16	must	108	system	118
17	country	93	has	105
18	national	85	work	100
19	work	79	country	94
20	security	78	improve	93

In table 7, the top 20 keywords are selected for comparison, which illustrates that there are 16 same keywords including "and" "we" "will" "party" "people", "our", "Chinese", "China", "development" "must", "have", "all", "new", "system", "work" and "country". The similarity in the two translations is up to 80%, the result having a lot to do with the continuity of the reports and the relevance of their themes.

Compared with the report of the 19th CPC National Congress, the keywords "we" and "will" in the report of the 20th CPC National Congress occur more frequently, which illustrates the CPC is inclined to rally people of all ethnic groups to make efforts for the great rejuvenation of the Chinese nation.

In the top 20 keywords, the keyword "improve" occurs in the translation of the report of the 20th CPC National Congress but not in that of the report of the 19th CPC National Congress. It shows that due to the hit of the COVID-19 pandemic on the economy, China increasingly focuses on economic recovery.

4 Conclusion

Based on a corpus research method, the paper makes a contrastive study of the lexical features between the English translations of the report of the 20th CPC National Congress and that of the 19th CPC National Congress from both macro and micro levels. Lexical features are explored from five perspectives: TTR, lexical

density, average sentence length, high-frequency words, and keywords.

The research reveals that two reports have relevance in themes and are a continuum in their major tasks. Specifically, the author found that high-value and median-value modal auxiliary verbs tend to be less used in the 20th CPC National Congress report, the reason for this is that the Chinese government is committed to building its own internationally friendly and positive image. Besides, the frequency of using delexicalised verbs is much higher in the 20th CPC National Congress report, resulting in more native expressions. It is also found that words in the report of the 20th CPC National Congress are more diversified and readable as a consequence of shorter sentence length. The paper hoped to provide some thoughts and ideas on the English translation of political documents.

However, only a limited range of perspectives are examined as an example in this paper due to time constraints. The author remains confident that more lexical features will be found in the future with the growth of corpora and more in-depth research. Additionally, it is proposed that more emphasis be placed on diction, modal auxiliary verbs and delexicalised verbs in translator training because the discussion demonstrates how valuable they are in the translating process due to their high potential of creation.

References

1. CGTN. (2022, October 17). *Foreign Experts Invited to Help Copy Edit Translated Versions of the 20th CPC National Congress Work Report*. https://news.cgtn.com/news/2022-10-17/Foreign-experts-help-translate-the-20th-CPC-National-Congress-report-1ed3T5oXm5G/index.html
2. Jia Hui & Sun Minwei. (2018). Innovation in English Translation of Political Documents -- A Corpus-based Comparison of English Translation of the Reports of the 19th and 18th National Congresses. *Shanghai Translation*, (05), 35-40.
3. Wu Yana, Xie Yating & Xia Mengjun. (2020). A Corpus-based Study on the Manifestation of the 19th National Congress Report. *Overseas English*, (02): 24-25+64.
4. Shang Wenbo. (2020). Research on the Image of Our Party from the Perspective of Corpus Criticism Translation Studies -- Taking the English Translation of the Report of the 19th National Congress as an Example. *East Journal of Translation*, (03): 38-45.
5. The Xinhua News Agency. (2022, October 25). Full Text of the Report to the 20th National Congress of the Communist Party of China. *China Daily*. https://global.chinadaily.com.cn/a/202210/25/WS6357e484a310fd2b29e7e7de.html
6. The Xinhua News Agency. (2017, November 04). Full Text of Xi Jinping's Report at 19th CPC National Congress. *China Daily*. https://www.chinadaily.com.cn/china/19thcpcnationalcongress/2017-11/04/content_34115212.htm
7. Kettunen, K. (2014). Can Type-Token Ratio be Used to Show Morphological Complexity of Languages?. *Journal of Quantitative Linguistics*, 21(3), 223-245.
8. Ure, J. (1971). Lexical density and register differentiation. *Applications of linguistics*, 23(7), 443-452.
9. Johansson, V. (2008). Lexical diversity and lexical density in speech and writing: A developmental perspective. *Working papers/Lund University, Department of Linguistics and Phonetics*, 53, 61-79.
10. Halliday, M. A. K., & Hasan, R. (1989). *Language, Context and Text: Aspects of Language in a Social-semiotic Perspective* (2nd ed.). Oxford: Oxford University Press.
11. Bonelli, E. T. (2000). Corpus Classroom Currency. *Darbai ir Dienos* 24, 205-243.