Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-024-02986-7 OPEN

Does simplification hold true for machine translations? A corpus-based analysis of lexical diversity in text varieties across genres

Jiang Niu[1,2] & Yue Jiang[1]✉

Extensive studies have described the linguistic features of human translations and verified the existence of the simplification translation universal. However, little has been known about the linguistic features of machine translations, although machine translation, as a unique modality of translation, has become an integral part of translation practice. This study is intended to test whether the simplification translation universal observed in human translations also holds true for machine translations. If so, are simplification features in machine translations different significantly from those in human translations? And does genre significantly affect simplification features? To this end, we built a balanced comparable corpus containing three text varieties, i.e., machine translations, human translations and target-language originals across three genres namely contemporary novels, government documents and academic abstracts. Based on the corpus, we conducted a systematic comparison of lexical diversity, as a proxy for simplification, of different text varieties. The results show that simplification is corroborated overall in both machine and human translations when compared with target-language originals, and machine translations are more simplified than human translations. Additionally, genre is found to exert a significant influence on the lexical diversity of different text varieties. This study is expected to expand the scope of corpus-based translation studies on the one hand and to offer insights into the improvement of machine translation systems on the other hand.

[1] School of Foreign Studies, Xi'an Jiaotong University, Xi'an, China. [2] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, Xi'an, China. ✉email: yuejiang58@163.com

Introduction

Big data, strong computing power, and deep learning algorithms have brought about a significant improvement in the output quality of neural machine translation (NMT) in the past few years. This is particularly true since the pre-trained technique (Radford et al., 2018) and self-attention mechanism (Vaswani et al., 2017) have been used in the modeling of NMT systems. The quality assurance of NMT has made it popular in translation practice. On a daily basis, machine translation is used as a productive and cost-effective aid by millions of people (Way, 2018). Human-computer interaction has become the norm of translation practice and is reshaping the landscape of translation studies, bringing machine translation and its related studies from the periphery to the center of translation studies (Jiang and Niu, 2022).

Despite that NMT has become an integral part of translation practice, the linguistic features of machine translations have not yet received widespread academic attention. Human translation features, however, have been extensively investigated since corpus linguistics and descriptive translation studies were integrated to form corpus-based translation studies (CBTS). One typical research agenda in CBTS is the quest for the distinctive features of the translational language, i.e., translation universals (TUs). Baker (1993) referred to the universal features of translation as certain linguistic features that typically occur in translated texts, resulting from the translation process itself rather than from the confrontation of specific linguistic systems. Such features initially proposed by Baker (1993) encompass a marked rise in the level of explicitness, disambiguation, and simplification, preference for conventional 'grammaticality', repetitions avoidance, and exaggeration of the target-language features. As the relevant studies go further, more TUs have been introduced such as source language shining-through (Teich, 2003), the unique item hypothesis (Tirkkonen-Condit, 2004), and constrained language universal (Kruger and van Rooy, 2016; Kajzer-Wietrzny and Ivaska, 2020), to name a few.

Among the prolific translation universal features, simplification is the widely examined one. Baker (1996: p. 176) defined simplification as "the idea that translators subconsciously simplify the language or message or both". There is a growing body of evidence that human translation is simplified, although some studies are controversial. If simplification is inherent in the translation process (Baker, 1993), we may ask, analogically, whether simplification also occurs in machine translation, which, as a unique modality of translation, has a different operating mechanism from that of the human mind. Nevertheless, few studies have delved into this issue to date, leaving room for further investigation into the distinctive characteristics of machine-generated texts, which is crucial and necessary in the current AI era, where machine translation tools are extensively employed in diverse scenarios. Against this background, the current study is intended to test the simplification universal in machine translations with human translations and original texts in the target language as references. It aims to address the following research questions:

1. Does simplification exist in machine and human translations when compared with target-language originals?
2. If yes, to what extent does simplification in machine translations differ from that in human translations?
3. Does genre significantly influence the features in terms of simplification in translations?

This present study shifts the research focus from the traditional subject of human translation to machine translation, aiming to broaden the scope of corpus-based translation studies. Additionally, a systematic investigation of the features of machine translations could empower developers to improve the current machine translation systems.

Literature review

Simplification in human translations. As the more controversial translation universal feature compared to other TU candidates (Liu and Afzaal, 2021), simplification has attracted a large amount of research attention ever since it was proposed. The relevant studies mainly focus on the features at the lexical and syntactic levels.

By observing lexical features, some earlier studies have described simplification as "the process and/or result of making do with less word" (Blum-Kulka and Levenston, 1983: p. 119) and using informal, colloquial, and modern lexis to translate formal, literate, and archaic words in the source text (Vanderauwera, 1985). Using a corpus-based method, Laviosa (1998a, 1998b) examined simplifications in translated newspapers and narrative texts by comparing them with the comparable non-translated counterparts. The studies found that the translated texts exhibited a higher level of simplification, characterized by a lower lexical density (the ratio of content words to all running words) and type-token ratio, although the latter appeared to be only marginally lower. Following Laviosa's studies, Williams (2005) further compared type-token ratio and lexical density (the ratio of content words to all running words) in translated and non-translated government texts in both English and French. The empirical results supported the simplification hypothesis in English texts, as a lower type-token ratio and lexical density were found. In French texts, on the other hand, opposite results were obtained, and the simplification hypothesis was refuted. Similarly, Cvrček and Chlumská (2015) examined whether there was a phenomenon of simplification in translated Czech literary texts. They demonstrated that the translated texts had a slightly less diverse lexicon namely a smaller type-token ratio than the non-translated Czech texts, which supported the simplification hypothesis. Kajzer-Wietrzny (2015) continued to test the simplification hypothesis more comprehensively in English translations and simultaneous interpreting from German, Dutch, French, and Spanish. Compared with English original texts, the interpreted texts showed simplification in list head coverage but not in lexical density (the ratio of content words to function words) and the proportion of high-frequency words. However, simplification was confirmed in all three indicators when it comes to translated texts. Apart from the above-mentioned languages, simplification was also analyzed in translated Chinese. For example, Xiao (2010) and Xiao and Yue (2009) found that the lexical variability gauged by the type-token ratio in Chinese translated texts did not differ significantly from that in Chinese original texts. Thus simplification was not evidenced. However, Chinese translated texts exhibited a lower ratio of lexical over function words, a greater accumulated proportion of high-frequency words, a higher ratio between high- and low-frequency word tokens, and a higher repetition rate of high-frequency words, lending support to the existence of simplification.

Earlier studies on syntactic features of translated texts mainly focused on the mean sentence length. For example, by comparing translated and non-translated original texts, Laviosa (1998a) observed a shorter mean sentence length in translated newspapers. However, she obtained conflicting results when analyzing translated narrative texts (Laviosa, 1998b), as a longer mean sentence length was found in such texts. Likewise, Williams, (2005) found that the mean sentence length of French-translated text was shorter than that of French non-translated texts, whereas an opposite trend was observed when it comes to English texts. Xiao (2010) found a significantly greater mean sentence length in translated Chinese than in native Chinese. However, Xiao and Yue (2009) found the difference in mean sentence length between the two text types was not statistically significant. To sum up, there is still no conclusive evidence for a reduction in mean sentence length and, thus, simplification of translational language.

More recent studies also used some other syntactic features to test the simplification universal. For example, Liu and Afzaal (2021) examined syntactic complexity in translated English and English original texts. By analyzing 13 syntactic complexity measures in four genres, they found a lower syntactic complexity in translated texts, supporting the simplification hypothesis in translations. Additionally, a significant influence of genre on syntactic complexity was found. In later studies, Liu et al. (2022) and Liu et al. (2023) continued to test simplification in translated texts and interpreted speeches, respectively. The results showed that the translated texts tended to be simpler than their non-translated counterparts in unigram entropy but not in part-of-speech entropy. The interpreted speeches had significantly lower scores in most syntactic complexity measures than non-interpreted speeches. The existence of simplification was thus confirmed in the interpreted speeches but partially in the translated texts.

To recap, simplification in human translations has been investigated in different translated languages by using various lexical and syntactic measures. However, the influence of genre on simplification features has not been thoroughly examined, despite that some studies have used materials from multiple genres such as Xiao and Yue (2009), Liu and Afzaal (2021), and Liu et al. (2022). In addition, the results regarding simplification in human translations have been somewhat inconclusive at different linguistic levels.

Simplification in machine translations. In contrast with the prolific study of simplification in human translations, that in machine translations remains understudied. Lapshinova-Koltunski (2015) was perhaps the first to quest for translation universals in both human and machine translations. By building a multi-genre corpus, she compared English source texts, German human translations, German machine translations, and comparable German originals. The results showed that machine translations had a higher average lexical density (the ratio of content words to grammatical words) but a lower standardized type-token ratio (STTR) than human translations and English and German originals. Thus the simplification hypothesis in machine translations was corroborated for STTR but not for lexical density. Moreover, the average lexical density in human translations was greater than in German originals but smaller than in English source texts; on the other hand, the average STTR in human translations was lower than in both German originals and English source texts. Hence simplification in human translation was also partially supported. Han and Jiang (2016) compared the Chinese machine translation (Baidu Translate) and human translation of an English novel. The study found that the machine translation had a higher lexical diversity (STTR) and a lower lexical density than the human translation. What's more, both human and machine translations exhibit a higher lexical variety and lexical density than Chinese original texts. Therefore the simplification universal in translations was not verified in this study. In a more recent study, Luo and Li (2022) compared Chinese-English translations by WeChat Translate with English original texts. They found that the machine translation had a higher standardized type-token ratio, a higher lexical density gauged by the proportion of content words to all running words, more lemmas in the top 100 and 200 list heads, and a lower proportion of the most frequently used words. These findings indicated that the machine translations had a broader lexical range and a greater lexical variety, which could not support the tendency of simplification in the machine translations.

The above review of the previous literature demonstrates that the quest for simplification in machine translations is still in its infancy. Pioneering and inspiring as they are, these previous studies are limited in the following aspects. First, contradictory results regarding simplification in machine translations have been obtained, which warrants further testing. Second, similar to the studies on simplification in human translations, research on simplification in machine translations is predominantly conducted within a single genre and seldom takes into account the influence of genre on translation universal features, which, though, has been noted by some previous researchers (Kruger and van Rooy 2012; Delaere et al., 2012; Liu and Afzaal, 2021). Third, the reference texts of machine translations are typically either original texts in the target language or human translations. It is essential to include simultaneously the two types of reference texts to obtain a complete picture of the linguistic features of machine translation.

In this study, simplification is further examined in machine translations, which opens up a new research agenda for the description of the linguistic features of translated texts—a heated research topic in corpus-based translation studies and, in turn, expands the scope of translation studies. In addition, the study of linguistic features inherent in machine translations is beneficial to gain a deeper understanding of the nature of machine translation, which could empower translation and post-editing practitioners and instructors, as well as assist the machine translation system developers in improving the current ones.

Methods

Materials. To analyze machine translations with human translations and original texts as references and to delve into the effect of genre on the linguistic features of these text varieties, we built a balanced comparable corpus (Xiao, 2011) with multiple genres. As shown in Table 1, the corpus contains three text varieties, i.e., Chinese-English machine translation, Chinese-English human translation, and English original texts across three genres namely contemporary novels, government documents, and academic abstracts.

Human translations and English original texts were selected in accordance with the following principles: (1) written, natural and published texts should be included; (2) the publishing years of human translations and English original texts should fall within a similar time frame; and (3) at least 50 samples should be included for each genre to guarantee credible results. In total, the corpus contains over 600 texts across the three genres, comprising about 0.6 million words.

Machine translations were generated using the 'Document Translation' module of Google Translate and DeepL, which ensures that textual elements were produced in a continuous and coherent manner. The reasons for using the two neural machine translation systems are as follows. First, both are general-purpose NMT systems rather than domain-specific ones. Second, they are the widely used mainstream NMT systems known for their high-quality assurance. Google Translate is currently the commercial machine translation system with the highest volume of daily translation (Way, 2018). DeepL, trained on high-quality data collected by Linguee, boasts the highest output quality among commercial MT systems (Krüger, 2020). Third, the two systems have been widely used in linguistic research, especially in comparative studies of translations (e.g., Krüger, 2020; Frankenberg-Garcia, 2022; Loock, 2020; etc.). All the above considerations guarantee the representativeness of the two NMT systems. Notably, the machine-translated texts did not undergo any pre- or post-editing.

Table 1 Basic information on the corpus used.

Text variety		No. of text	Range of text length	Avg. text length	Total No. of words	Publication time
Contemporary novels						
Human		52	340-6816	1718	89,336	2000-2015
Machine	Google	52	335-7203	1587	82,537	transl.2023.2
	DeepL	52	309-6998	1546	80,400	
Original		53	311-4834	1637	86,765	2004-2010
Total		209	/	1622	339,038	/
Government documents						
Human		51	263-3930	1282	65,417	2010-2012, 2014-2015
Machine	Google	51	272-3992	1149	58,622	transl.2023.2
	DeepL	51	280-3871	1141	58,202	
Original		51	237-1817	818	41,695	2010-2015
Total		204	/	1098	223,936	/
Academic abstracts						
Human		52	76-242	135	7045	2000-2015
Machine	Google	52	63-247	150	7,817	transl.2023.2
	DeepL	52	57-251	158	8229	
Original		52	75-302	148	7693	2000-2015
Total		208	/	148	30,784	/

Hereafter detailed information about texts from each genre is introduced.

Contemporary novels. Chinese contemporary novels were selected from the masterpieces written by Chinese literary masters such as Yan Mo (a Nobel Prize winner), Pingwa Jia, Hua Yu, Jia Mai, etc. The human translators of these works include prestigious sinologists such as Howard Goldblatt, Michael Berry, Olivia Milburn, etc. English original texts were chosen from the works written by famous contemporary novelists in the English-speaking world, such as Julian Barnes, Richard Francis, Colum McCann, Joshua Ferris, etc. To exclude the factors pertaining to individual authors or translators, we chose a wide range of authors, translators, and subject matters.

Government documents. Government documents (GD) were selected from the *Report on the Work of the Government* (RWG) and the *State of the Union Address* (SOTU). RWG is an annual report given by the Chinese Premier at the session of the National People's Congress of the People's Republic of China. Exhibiting a high degree of formality, it is a vital document for the Chinese government to report achievements and allocate work (Xie and Yuan, 2013). Human translations of RWG were initially produced by professional Chinese translators. The translations were then reviewed by English native speakers and finally edited by high-level Chinese professional translators. Human-translated works were retrieved from the official website of 'Theory China: Resource for Understanding China' (http://en.theorychina.org/), affiliated with the Research Institute of the History and Literature of the Central Committee of the Communist Party of China.

Comparable to RWG, SOTU is an annual report given by the American President, covering topics such as the economy, education, politics, and so on. The relevant materials were collected on the website of 'The American Presidency Program' organized by the University of California, Santa Barbara (https://www.presidency.ucsb.edu).

Academic abstracts. Academic abstract (AA) is a condensed version of an entire article. It is usually a brief, natural, complete text and a meaningful way to promote academic viewpoints. Due to the interdisciplinary differences in the convention and style of abstract writing (Li, 2020), we collected academic abstracts exclusively within translation studies. The reason for choosing this discipline is that the translators of the abstracts, also as the researchers of translation studies, could consciously manipulate the translation process. Therefore the translated abstract could exhibit more noticeable features of translational language (Li, 2020).

The abstracts were selected in the light of the following principles. First, abstracts from leading journals in translation studies should be collected to ensure the quality of the materials. Second, journals that do not have word count limits and strict formatting requirements should be included to keep the materials as natural and authentic as possible. Third, general-purpose journals rather than theme-specific ones should be chosen to exclude the impact of a particular theme on the linguistic features of abstracts. Based on all these considerations, we chose the *Chinese Translators Journal* and *Target: International Journal of Translation Studies* as the target journals. Articles in the former are published in Chinese with an English translation of each abstract, while articles in the latter are published all in English. When selecting translated abstracts, we singled out the translations reformulated from Chinese source texts because there were no equivalences between translations and their source texts. When collecting English original abstracts, we determined whether the authors were English native speakers based on their affiliations and personal profiles. Moreover, too short texts were excluded since it is difficult to observe linguistic regularities in such texts.

It is worthy of note that since contemporary novels and government documents are typically very long texts consisting of relatively independent chapters or sections of a certain length, we took relatively shorter samples from the long texts without compromising the integrity and homogeneity of the contents.

Indicators for simplification. Since the goal of this study is to test whether simplification found in human translations also exists in machine translations, type-token ratio (TTR), an indicator commonly examined in corpus-based studies of human translations, is chosen to achieve the research goal. TTR is the ratio of unique words (types) that occur in a text to the total number of words (tokens) (McEnery and Hardie, 2011; McNamara et al., 2014). A higher TTR indicates that most words in a text are different, i.e., more new words are used. Conversely, a lower TTR occurs when the words in a text are repeated frequently.

As reviewed in section "Literature review", type-token ratio has been considerably investigated since Baker (1995) introduced it from corpus linguistics to corpus-based translation studies. According to Baker (1996), a lower TTR means a text uses a less varied or narrower range of vocabulary, indicating that the text is lexically easier and simpler to process. Previous studies have widely employed TTR as a proxy for lexical diversity or lexical variability to test the simplification hypothesis. Unfortunately, conflicting findings have been yielded in both human and machine translations, leaving room for further study.

One inherent weakness of the type-token ratio is that it is sensitive to text length (McEnery and Hardie, 2011). Specifically, with the text length becoming longer, the number of tokens must increase while types might be repetitive. As a result, some other indicators based on TTR but uncorrelated with text length have been created. The present study uses two such indicators, namely *vocd* and MTLD (Measure of Textual Lexical Diversity) (McCarthy and Jarvis, 2010), to measure lexical diversity and overcome the impact of text length. The two indicators, using different approaches and capturing unique aspects of lexical information, have been proven to be the more robust approach to lexical diversity assessment (McCarthy and Jarvis, 2010).

The value of *vocd* is calculated using a computational algorithm that fits TTR random samples with ideal TTR curves (McNamara et al., 2014). Specifically, it is measured by taking from a text 100 random samples of 35 tokens. The TTR for each of these samples is then calculated and a mean TTR is yielded. The same procedure is iterated for varying sample sizes, ranging from 36 tokens to 50 tokens. The mean TTR values for each sample size are used to create a random-sampling TTR curve for the text. Following that, a formula, *D* coefficient (see Malvern et al., 2004, p. 51), is employed to produce a theoretical curve that most closely fits the created TTR curve formed from the random samples. The best fit between the theoretical curve and the random-sampling TTR curve is referred to as the value of *vocd*. For a more comprehensive understanding of this indicator, additional information can be found in Malvern et al. (2004) and McCarthy and Jarvis (2007).

MTLD is calculated as the mean length of sequential word strings in a text that maintains a given TTR value (McNamara et al., 2014). To calculate the TTR value, each word of a text is evaluated sequentially for its TTR. When the default TTR factor size value is reached, the factor count increases by a value of 1 (a full factor), and the TTR evaluations are reset. The final MTLD value is obtained in a way that the total number of words in the text is divided by the total factor count. The calculation process is carried out in two ways, i.e., forward processing and reverse processing. The TTR value is calculated for each processing direction, and the mean of the two values represents the final MTLD value (McCarthy, 2005; McCarthy and Jarvis, 2010).

The data about MTLD and *vocd* were extracted via Coh-Metrix (Graesser et al., 2004). As a ready-made computational program, Coh-Metrix currently is among the most sophisticated textual assessment tools available on the web (McNamara et al., 2014). As it can be applied to almost any text, its metrics can be used to compare texts belonging to various genres (Graesser and McNamara, 2011).

Statistical analysis. This study involves two categorical independent variables, i.e., text variety and genre. Text variety is categorized into three groups namely machine translation (Google Translate and DeepL), human translation, and comparable original texts. Genre also falls into three categories namely contemporary novels, government documents, and academic abstracts. As for the dependent variable, i.e., lexical diversity, it is measured by two indicators of continuous type that are interrelated with each other. Therefore, a two-way analysis of variance (ANOVA) was carried out to test whether the differences in the lexical diversity of the three text varieties are statistically significant and whether the lexical diversity of different text varieties is significantly influenced by genre. Before carrying out the ANOVA test, a Shapiro–Wilk normality test was performed, and all indicators were found to meet the assumptions of normality. Also, the data passed the Levene's test of homogeneity. The statistical analyses were conducted with SPSS 18.

To further probe into the specific way of interactions between text variety and genre, i.e., how the lexical diversity of the text varieties differs across genres, we performed simple effect analysis using programming instructions in SPSS.

Table 2 Main effect and interaction effect of independent variables on lexical diversity.

Independent variables	Indicators	*df*	*F*	*p*	Partial η^2
Genre	MTLD	2	54.072	0.000	0.172
	vocd	2	91793.909	0.000	0.508
Text variety	MTLD	3	137.364	0.000	0.441
	vocd	3	32779.975	0.000	0.356
Genre × Text variety	MTLD	6	8.170	0.000	0.086
	vocd	6	5402.222	0.000	0.154

Results

According to the statistics shown in Table 2, genre exerts a significant main effect on each of the indicators of lexical diversity ($p < 0.001$), with its effect size reaching 0.172 in MTLD and 0.508 in *vocd*. This manifests that genre significantly affects lexical diversity, accounting for 17.2% of variances in MTLD and 50.8% of variances in *vocd*. Likewise, text variety exerts a significant main effect on the two indicators of lexical diversity ($p < 0.001$), with its effect size amounting to 0.441 and 0.356 in MTLD and *vocd*, respectively. That is, text variety also plays a significant role in lexical diversity. And 44.1% of variances in MTLD and 35.6% of variances in *vocd* could be explained as resulting from different text varieties. According to Li et al. (2014), the effect size of 0.14 is a large one. Thus the two independent variables both have a substantial impact on lexical diversity. In brief, the above results reveal that the lexical diversity of different text varieties and genres is significantly different.

Effect size in this study can also be regarded as a measure of the validity of the indicators in differentiating texts. According to Table 2, the effect size of genre is larger than that of text variety in *vocd*, while the effect size of text variety is larger than that of genre in MTLD. This finding suggests that *vocd* is more effective than MTLD in distinguishing texts of different genres, whereas MTLD is more reliable than *vocd* in distinguishing different text varieties. This discrepancy might be attributed to the two indicators' different calculation methods, though they are all derived from TTR.

As shown in Table 2, the two independent variables also significantly interact with each other in both MTLD and *vocd* ($p < 0.001$). This implies that the lexical diversity of text varieties differs significantly in different genres. To put it in another way, genre can significantly moderate the lexical diversity of different text varieties. This finding confirms the effect of genre on translation features, as assumed by some scholars such as Kruger and van Rooy (2012) and Delaere et al. (2012). In addition, text varieties and genre have a higher interaction effect in *vocd* than in MTLD, revealing that the lexical diversity of different text

Table 3 Descriptive statistics on lexical diversity of texts across genres.

Indicators	Genre	Text variety	Mean	SD	No. of text
MTLD	novel	Human	9.326	0.653	52
		Google	8.125	0.729	52
		DeepL	7.974	0.777	52
		Original	9.334	0.869	53
		Total	8.693	0.993	209
	GD	Human	8.260	0.669	51
		Google	7.563	0.791	51
		DeepL	7.572	0.691	51
		Original	10.053	0.761	51
		Total	8.362	1.250	204
	AA	Human	8.263	1.303	52
		Google	6.932	1.054	52
		DeepL	6.726	0.823	52
		Original	8.834	1.544	51
		Total	7.683	1.493	207
vocd	novel	Human	105.981	11.052	52
		Google	88.756	12.967	52
		DeepL	88.507	12.678	52
		Original	102.919	16.356	53
		Total	96.571	15.537	209
	GD	Human	79.384	12.785	51
		Google	69.765	12.804	51
		DeepL	69.521	12.207	51
		Original	120.697	13.613	51
		Total	84.842	24.689	204
	AA	Human	60.635	28.348	52
		Google	47.003	15.900	52
		DeepL	43.589	14.011	52
		Original	71.590	29.588	51
		Total	55.628	25.444	207

varieties measured by *vocd* is influenced more by genre than that measured by MTLD.

Comparison between translated and original texts. A simple effect analysis was conducted to investigate further the specific way of interaction between text variety and genre in lexical diversity, i.e., to analyze how the lexical diversity of different text varieties differs across genres.

As demonstrated in Tables 3, 4, and Fig. 1, for government documents and academic abstracts, both human translations and machine translations (DeepL and Google) are significantly lower than English original texts in MTLD and *vocd*. Specifically, the mean values of MTLD and *vocd* in original government documents are 10.053 and 120.697, respectively, while those in human-translated government documents are 8.260 and 79.384, with a mean difference of 1.794 and 41.313. The mean values of MTLD and *vocd* in Google-translated government documents are 7.563 and 69.765, respectively, with a mean difference from English original texts amounting to 2.491 and 50.931. The mean values of MTLD and *vocd* in DeepL-translated government documents are 7.572 and 69.521, respectively, with a mean difference from English original texts amounting to 2.481 and 51.175. In terms of the mean values of the lexical diversity of the text varieties in academic abstracts, almost the same trend with government documents can be found.

In contrast, when it comes to contemporary novels, there are significant differences in MTLD and *vocd* between machine translations and English original texts, while there are no significant differences between human translations and English original texts. To be specific, the mean values of MTLD and *vocd* in English original texts are 9.334 and 102.919, respectively, while those in human translations are 9.326 and 105.981, with a mean difference of 0.008 and 3.062. The mean values of MTLD and *vocd* in Google translations are 8.125 and 88.756, respectively, with a mean difference from English original texts amounting to 1.209 and 14.162. The mean values of MTLD and *vocd* in DeepL translations are 7.974 and 88.507, respectively, with a mean difference from English original texts reaching 1.360 and 14.411.

Comparison between translated texts. When comparing machine translations with human translations and Google translations with DeepL translations, a consistent trend is found across the three genres. Specifically, both Google and DeepL translations are significantly lower than human translations in MTLD and *vocd* across all three genres, as shown in Tables 3 and 4 and Fig. 1. In addition, the differences in lexical diversity between the two machine translations are statistically non-significant, with that of Google translations higher than that of DeepL translations on the whole. As shown by the detailed statistics, the mean values of MTLD in Google and DeepL translations of novels are 8.125 and 7.974, those of government documents are 7.563 and 7.572, and those of academic abstracts are 6.932 and 6.726. As far as *vocd* is concerned, the mean values in Google and DeepL translations of novels are 88.756 and 88.507, those in government documents are 69.765 and 69.521, and those in academic abstracts are 47.003 and 43.589. All these statistics indicate that there are negligible differences between Google and DeepL translations.

Comparison between machine translations and human-produced texts. Also, the statistics demonstrate a tendency of division between machine translations (Google and DeepL) and human-produced texts (human translations and English originals), as shown in Fig. 1. This is particularly true when it comes to novels. For example, in the two indicators of lexical diversity, both Google- and DeepL-translated novels are significantly different from human translations and English originals, with the former lower than the latter ones. However, the difference between human translations and English originals and that between the two modes of machine translations are statistically non-significant. These findings suggest that there seems to be a cut-off between machine-produced texts and human-produced texts. However, this is far from conclusive because the remarkable division is observed only in novels and machine translations are used only as the representative of machine-produced texts. It is necessary to include machine-produced original texts to ascertain whether machine productions deviate from human productions. Only in this way can the factors of translated VS. original and machine-produced VS. human-produced and their interactions be fundamentally explored.

Discussion

This study aims to test the simplification hypothesis in translations across three genres. As shown by the comparison between translated texts and original texts, the lexical diversity of both human- and machine-translated government documents and academic abstracts is all significantly lower than that of their comparable original counterparts. Thus the simplification hypothesis is corroborated in both human- and machine-translated government documents and academic abstracts. Herein the evidence of simplification measured by TTR-related indicators in human-translated texts is in line with Laviosa (1998a, 1998b), Williams (2005), Kajzer-Wietrzny (2015), Cvrček and Chlumská (2015) and Lapshinova-Koltunski (2015). By contrast, for contemporary novels, the lexical diversity of the human translations is not significantly different from that of

Table 4 The results of simple effect analysis.

Indicators	Genre	Text variety	Mean difference	SE	*p*
MTLD	novel	DeepL-Original	−1.360***	0.181	**0.000**
		Google-Original	−1.209***	0.181	**0.000**
		Human-Original	−0.008	0.181	0.965
		Human-Google	1.200***	0.182	**0.000**
		Human-DeepL	1.352***	0.182	**0.000**
		Google-DeepL	0.152	0.182	0.404
	GD	DeepL-Original	−2.481***	0.184	**0.000**
		Google-Original	−2.491***	0.184	**0.000**
		Human-Original	−1.794***	0.184	**0.000**
		Human-Google	0.697***	0.184	**0.000**
		Human-DeepL	0.688***	0.184	**0.000**
		Google-DeepL	−0.009	0.184	0.959
	AA	DeepL-Original	−2.108***	0.183	**0.000**
		Google-Original	−1.902***	0.183	**0.000**
		Human-Original	−0.571**	0.183	**0.002**
		Human-Google	1.331***	0.182	**0.000**
		Human-DeepL	1.537***	0.182	**0.000**
		Google-DeepL	0.206	0.182	0.257
vocd	novel	DeepL-Original	−14.411***	3.337	**0.000**
		Google-Original	−14.162***	3.337	**0.000**
		Human-Original	3.062	3.337	0.359
		Human-Google	17.225***	3.352	**0.000**
		Human-DeepL	17.473***	3.352	**0.000**
		Google-DeepL	0.249	3.352	0.941
	GD	DeepL-Original	−51.175***	3.385	**0.000**
		Google-Original	−50.931***	3.385	**0.000**
		Human-Original	−41.313***	3.385	**0.000**
		Human-Google	9.619*	3.385	**0.005**
		Human-DeepL	9.863*	3.385	**0.004**
		Google-DeepL	0.244	3.385	0.943
	AA	DeepL-Original	−28.001***	3.369	**0.000**
		Google-Original	−24.587***	3.369	**0.000**
		Human-Original	−10.955**	3.369	**0.001**
		Human-Google	13.632***	3.352	**0.000**
		Human-DeepL	17.046***	3.352	**0.000**
		Google-DeepL	3.414	3.352	0.309

Note. $p < 0.05$*; $p < 0.01$**; $p < 0.001$***.
Bold *p* values represent statistically significant differences between pairs of text varieties.

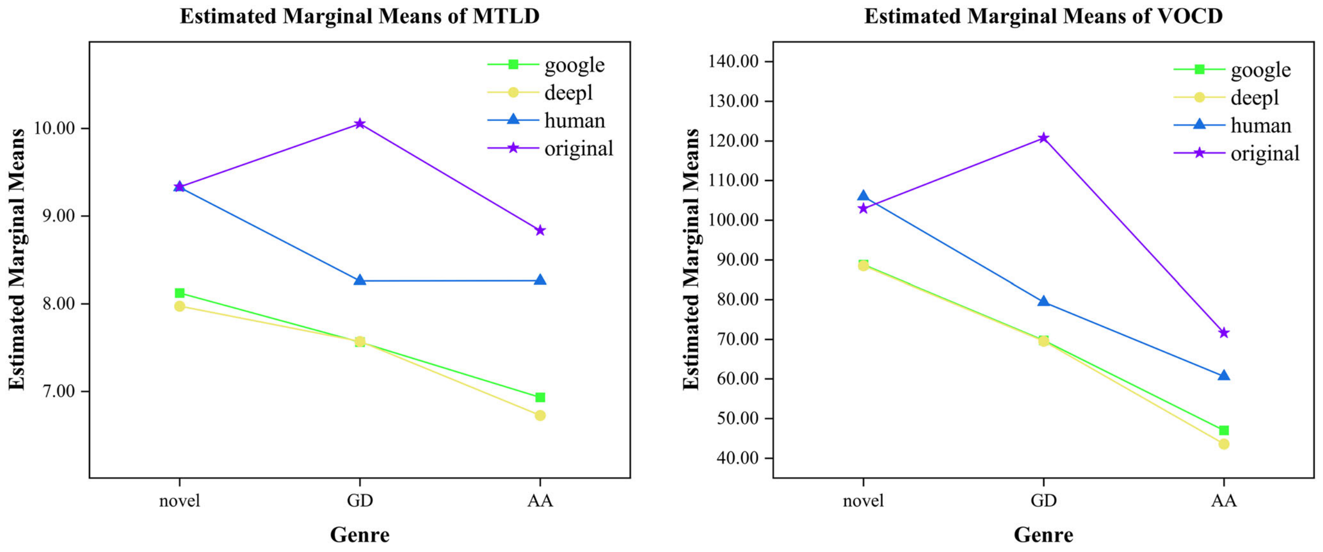

Fig. 1 The interaction between text variety and genre. The left shows MTLD values of different text varieties across genres. The right panel shows *vocd* values of different text varieties across genres.

English original texts. However, the lexical diversity of machine translations is significantly lower than that of English original texts. Hence, simplification is confirmed in machine-translated novels but not in human-translated novels. Furthermore, machine translations are more simplified than human translations across all three genres. These findings lend support to Lapshinova-Koltunski's finding (2015) that STTR (standardized type-token ratio) in machine translations is lower than in English and German originals and human translations. However, our findings conflict with the higher STTR in machine translations than in English original texts (Luo and Li, 2022) and human translations (Han and Jiang, 2016).

As shown by the statistics, the lexical diversity of human translations of contemporary novels is the highest compared with that of human translations of the other two genres, showing a non-significant difference between human translations and English original novels. This might be attributable to the fact that human translators can consciously manipulate their use of words when they translate novels. More specifically, novels are characteristic of language use with esthetic flavor, which requires creative wording and elaborate organization of information. Human translators, conscious of these defining features of novels, can utilize particular strategies such as flexibly switching among diverse words to avoid lexical repetition and redundancy. Moreover, they can take into consideration the target-language readers' acceptability of language uses. These factors might result in the higher lexical diversity of human translations, thereby challenging the simplification hypothesis.

The fact that machine translations of all three genres have lower lexical diversity than human translations and English originals might be ascribed to the following factors. In contrast with human translators, the operating mechanism of machine translation systems is the digital encoding and decoding of language symbols and probability computing by using algorithms, which is void of human-like intelligence or intentionality from the perspective of strong AI (Jiang and Niu, 2022). Therefore, during their output process, MT systems solely mechanically undertake the task of language transfer, but they are incapable of modifying translations according to the unique features of a particular genre and the acceptability of translations from the standpoint of target readers. In addition, as held by Vanmassenhove et al. (2021) and Vanmassenhove et al. (2019), the inherently probabilistic nature or algorithm of MT systems can make them overgeneralize by repeating more frequent words while disregarding less frequent ones in the training data when translations are produced, namely a tendency of algorithmic bias or statistical bias. Furthermore, repetitions are used more frequently in Chinese than in English (Lian, 2010), and machine translations exhibit a stronger effect of source language shining-through than human translations (Bizzoni et al., 2020; Ahrenberg, 2017). All these possible factors might cause the repetition of words in machine-translated English texts, thus resulting in the lower lexical diversity of these texts.

A few examples taken from each genre are given below to illustrate that human translators can deliberately avoid repetitive and redundant word usage, but machine translation systems are incompetent at this strategy. In Example 1, the name of the protagonist in the novel 韩文举 (*Han Wenju*) was repeatedly used in two adjacent sentences in the source text, which is typical of Chinese. The human translator used the relative pronoun *who* to replace the second *Han Wenju*. In contrast, both DeepL and Google merely adhered to the diction and sentence structure of the source text, thus resulting in word redundancy and lower lexical diversity in English translations. Example 2, taken from government documents, used the Chinese character 稳 (adj., 'stable') repeatedly in the source text to create a rhythmic and smooth flow of information. However, the two Chinese characters essentially express the same meaning in this context. DeepL and Google Translate repeated the second 稳, using other parts of speech such as *steady* and *stability*. Contrastively, the human translator, who can understand the context of the source text, creatively integrated the two clauses without repeating *steady*. In the third example taken from academic abstracts, the machine translation systems mechanically imitated the Chinese source text by repeatedly using *Chinese literature* ('中国文学'), while the human translator used the indefinite pronoun *ones* to avoid redundancy.

Example 1 (《浮躁》 ('*Turbulence*'), Jia Pingwa, 2009)

ST: 爹娘死得早, 小水就跟伯伯韩文举过活。韩文举能说会道, 但性情敏感而胆怯。

Literal translation: Xiaoshui's parents died when she was still very young, so she lived with her uncle Han Wenju. Although he has a silver tongue, he is sensitive and timid.

DeepL: *When his parents died early, Xiao Shui lived with his uncle, Han Wenju. Han Wenju could speak well, but he was sensitive and timid.*

Google: *Parents died early, and Xiao Shui lived with his uncle Han Wenju. Han Wenju is articulate, but his temperament is sensitive and timid.*

Human: *After her parents had died, when she was still very young, she'd gone to live with Han Wenju, who, although blessed with the gift of gab, was a moody, timid man.*

Example 2 (*Chinese Report on the Work of the Government*, 2015)

ST: 一年来, 我国经济社会发展总体平稳, 稳中有进。

Literal translation: Over the past year, the economic and social development of China generally remained stable and also made progress.

DeepL: *Over the past year, China's economic and social development has been generally stable, with steady progress.*

Google: *Over the past year, my country's economic and social development has been generally stable, with progress while maintaining stability.*

Human: *During the past year, China has, overall, achieved a stable performance while at the same time securing progress in its economic and social development.*

Example 3 (*Literary translation and China's "going-out" cultural strategy: current situation, existing problems and suggestions for improvement*, Chinese Translators Journal, issue 6, volume 31, 2010)

ST: 中国文学, 特别是中国当代文学在国际上的译介情况如何?

Literal translation: How is Chinese literature especially contemporary Chinese literature being translated internationally?

DeepL: *What is the situation of Chinese literature, especially contemporary Chinese literature, in international translation?*

Google: *How is the international translation and introduction of Chinese literature, especially Chinese contemporary literature?*

Human: *How are Chinese literary works in general and the contemporary ones in particular being translated?*

Lastly, it is notable that although Google Translate and DeepL are designed and trained differently, their translations exhibit no significant difference across all three genres, demonstrating a convergence of the two machine translation systems. In addition, as reported above, machine translations differ significantly from human translations across all three genres. These findings indicate that machine translation has unique features distinct from those of human translation, though they are all translated texts. To be specific, machine translations are less lexically diverse and thus more simplified when compared to human translations and original texts in the target language. In a sense, such simplified machine-translated language could be diagnosed as a symptom of an impoverished language (Vanmassenhove et al., 2021), the impact of which on human language, in the long run, should be paid keen attention to (Jiang and Niu, 2022).

To avoid over-simplifications in machine translations, the following solutions might be helpful in the improvement of MT models. The first solution is external in nature. The encoder-decoder paradigm has recently become a standard architecture of neural machine translation. The input language sequences from source texts are first encoded into real number vectors and then decoded into output language sequences. To address the problem of machine translations barely using pronouns or other forms of wording to replace repeated words, external linguistic knowledge, like knowledge about reference relations, can be integrated into either the encoder or decoder of the machine translation models. This integration can facilitate MT models better capturing the semantic links among different language units in a particular context. The second solution is an internal one for machine translation models. Collecting more training data as a huge language repository can allow MT models to internally learn relevant linguistic knowledge. Lastly, we propose the use of a recently introduced paradigm namely pre-training-fine-tuning in MT model design. By leveraging the transformer architecture (Vaswani et al., 2017) and a vast amount of training data, commercial MT developers can first pre-train a large language model (LLM) to better learn linguistic patterns and statistical laws at the lexical, syntactic, and even semantic levels. The LLM, as a general-purpose language model, can then be fine-tuned and optimized for specific translation tasks.

Conclusion

The present study attempts to examine whether the simplification universal observed in human translations also holds true for machine translations and whether the variable genre has a significant impact on the simplification features. Based on a self-built comparable corpus containing machine translations, human translations, and target-language originals across multiple genres, lexical diversity as a proxy of simplification was analyzed. The results show that simplification is found in both human- and machine-translated government documents and academic abstracts when compared with English original texts. However, in terms of contemporary novels, simplification is only verified in machine translations, while human translations are found to have no significant difference from English original texts. More importantly, machine translations (Google Translate and DeepL) across the three genres have no significant difference between each other and are significantly more simplified than human translations. In addition, the variable genre remarkably moderates the lexical diversity of different text varieties. In other words, the lexical diversity of machine translations, human translations, and English originals differs significantly in different genres.

With the fast development of artificial intelligence, machine translation, as an integral way of translation, is reshaping the landscape of translation studies and shaking the discipline ‘tree’ depicted by Holmes (Munday et al., 2022). Against this backdrop, the search for unique features of machine translations can offer new insights into corpus-based translation studies, especially the descriptive studies of translational features. On the other hand, a thorough understanding of the linguistic features of machine translations can shed light on the cultivation of MT literacy in translation education, the post-editing of MT outputs in translation practice, and the development of MT models in NLP as well.

This study is limited in the following ways. First, compared to the two commercial translation models used in this study, i.e., Google Translate and DeepL, large language models, a new AI paradigm, can generate translations with higher quality due to their capacity to connect longer contexts and the use of a larger amount of training data. Therefore, translations generated by large language models should be included in future studies to depict a more comprehensive picture of the distinct features of machine translations in this fast-evolving AI age. Second, this study delves into simplification in machine translations by merely looking at two lexical diversity indicators. Linguistic features at syntactic and textual levels should also be examined to further validate the simplification hypothesis.

Data availability

The datasets generated during and/or analyzed during the current study are available from the corresponding author on reasonable request.

Received: 24 December 2023; Accepted: 25 March 2024;
Published online: 03 April 2024

References

Ahrenberg L (2017) Comparing machine translation and human translation: a case study. In: Temnikova I, Orasan C, Corpas G, et al. (eds). Proceedings of the First Workshop on Human-Informed Translation and Interpreting Technology (HiT-IT). Association for Computational Linguistics, Bulgaria, pp. 21–28

Baker M (1993) Corpus linguistics and translation studies: Implications and applications. In: Baker M, Francis G, Tognini-Bonelli E (eds) Text and technology: in honour of John Sinclair. John Benjamins, Philadelphia, pp. 233–252

Baker M (1995) Corpora in translation studies: an overview and some suggestions for future research. Target 7(2):223–243

Baker M (1996) Corpus-based translation studies: the challenges that lie ahead. In: Somers H (ed.) Terminology, LSP and translation: studies in language engineering in honour of Juan C. Sager. John Benjamins, Amsterdam and Philadelphia, pp. 175–186

Bizzoni Y, Juzek TS, España-Bonet C et al (2020) How human is machine translationese? Comparing human and machine translations of text and speech. In: Federico M, Waibel A, Knight K et al. (eds) Proceedings of the 17th International Conference on Spoken Language Translation, online, Association for Computational Linguistics, pp. 280–290

Blum-Kulka S, Levenston EA (1983) Universals of lexical simplification. In: Faerch C, Kasper G (eds) Strategies in interlanguage communication. Longman, London and New York, p 119–139

Cvrček V, Chlumská L (2015) Simplification in translated Czech: a new approach to type-token ratio. Russian Linguist 39(3):309–325

Delaere I, De Sutter G, Plevoets K (2012) Is translated language more standardized than non-translated language?: Using profile-based correspondence analysis for measuring linguistic distances between language varieties. Target 24(2):203–224

Frankenberg-Garcia A (2022) Can a corpus-driven lexical analysis of human and machine translation unveil discourse features that set them apart? Target 34(2):278–308

Graesser AC, McNamara DS (2011) Computational analyses of multilevel discourse comprehension. Top Cogn Sci 3(2):371–398

Graesser AC, McNamara DS, Louwerse MM et al. (2004) Coh-metrix: Analysis of text on cohesion and language. Behav Res Method Instrum Comput 36(2):193–202

Han HJ, Jiang Y (2016) A corpus-based comparison of general language features of human translation and online machine translation. Foreign Lang Teach 37(5):102–106

Jiang Y, Niu J (2022) A corpus-based search for machine translationese in terms of discourse coherence. Acro Lang Cult 23(2):148–166

Kajzer-Wietrzny M (2015) Simplification in interpreting and translation. Acro Lang Cult 16(2):233–255

Kajzer-Wietrzny M, Ivaska I (2020) A multivariate approach to lexical diversity in constrained language. Acro Lang Cult 21(2):169–194

Kruger H, van Rooy B (2012) Register and the features of translated language. Acro Lang Cult 13(1):33–65

Kruger H, van Rooy B (2016) Constrained language: a multidimensional analysis of translated English and a non-native indigenised variety of English. Eng World-Wide 37(1):26–57

Krüger R (2020) Explicitation in neural machine translation. Acro Lang Cult 21(2):195–216

Lapshinova-Koltunski E (2015) Variation in translation: evidence from corpora. In: Fantinuoli C, Zanettin F (eds.) New directions in corpus-based translation studies. Language Science Press, pp. 81–99

Laviosa S (1998a) The English comparable corpus: a resource and a methodology. In: Bowker L, Cronin M, Kenny D, et al. (eds.) Unity in diversity: current trends in translation studies. Routledge, London, pp. 101–112

Laviosa S (1998b) Core patterns of lexical use in a comparable corpus of English narrative prose. Meta 43(4):557–570

Li WL, Zhang HC, Shu H (2014) Quantitative research methods and statistical analysis of education and psychology. Beijing Normal University Publishing Group, Beijing

Li XD (2020) Mediating cross-cultural differences in research article rhetorical moves in academic translation: a pilot corpus-based study of abstracts. Lingua 238:102795

Lian SN (2010) Contrastive study of English and Chinese. Higher Education Press, Beijing

Liu KL, Afzaal M (2021) Syntactic complexity in translated and non-translated texts: a corpus-based study of simplification. PLoS ONE 16(6):e0253454

Liu KL, Liu ZZ, Lei L (2022) Simplification in translated Chinese: an entropy-based approach. Lingua 275:103364

Liu Y, Cheung AKF, Liu KL (2023) Syntactic complexity of interpreted, L2 and L1 speech: a constrained language perspective. Lingua 286:103509

Loock R (2020) No more rage against the machine: How the corpus-based identification of machine-translationese can lead to student empowerment. J Spec Transl 34:150–170

Luo JR, Li DC (2022) Universals in machine translation? A corpus-based study of Chinese-English translations by WeChat Translate. Intern J Corp Ling 27(1):31–58

Malvern D, Richards B, Chipere N et al. (2004) Lexical diversity and language development. Palgrave Macmillan, New York

McCarthy PM (2005) An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (MTLD). Dissertation, The University of Memphis

McCarthy PM, Jarvis S (2007) *vocd*: a theoretical and empirical evaluation. Lang Test 24(4):459–488

McCarthy PM, Jarvis S (2010) MTLD, vocd-D, and HD-D: a validation study of sophisticated approaches to lexical diversity assessment. Behav Res Method 42(2):381–392

McEnery T, Hardie A (2011) Corpus linguistics: Method, theory and practice. Cambridge University Press

McNamara DS, Graesser AC, McCarthy PM et al. (2014) Automated evaluation of text and discourse with Coh-Metrix. Cambridge University Press

Munday J, Pinto SR, Blakesley J (2022) Introducing translation studies: Theories and applications (Fifth edition). Routledge, London and New York

Radford A, Narasimhan K, Salimans T, et al. (2018) Improving language understanding by generative pre-training. Available via Google Scholar. https://www.mikecaptain.com/resources/pdf/GPT-1.pdf

Teich E (2003) Cross-linguistic variation in system and text: a methodology for the investigation of translations and comparable texts. Mouton de Gruyter, Berlin

Tirkkonen-Condit S (2004) Unique items: over- or under-represented in translated language? In: Mauranen A, Kujamäki P (eds.) Translation universals: Do they exist? John Benjamins Publishing Company, Amsterdam/Philadelphia, pp. 177–184

Vanderauwera R (1985) Dutch novels translated into English: the transformation of a ’minority’ literature. Rodopi, Amsterdam

Vanmassenhove E, Shterionov D, Way A (2019) Lost in translation: loss and decay of linguistic richness in machine translation. https://doi.org/10.48550/arXiv.1906.12068

Vanmassenhove E, Shterionov D, Gwilliam M (2021) Machine translationese: Effects of algorithmic bias on linguistic complexity in machine translation. https://doi.org/10.48550/arXiv.2102.00287

Vaswani A, Shazeer N, Parmar N, et al. (2017) Attention is all you need. In: Proceedings of the 31st International Conference on Neural Information Processing System (NIPS 2017). Long Beach, USA, December 2017, pp. 6000–6010

Way A (2018) Quality expectations of machine translation. In: Moorkens J, Castilho S, Gaspari F, et al. (eds) Translation quality assessment. Springer, 159–178

Williams DA (2005) Recurrent features of translation in Canada: a corpus-based study. Dissertation, University of Ottawa

Xiao R (2010) How different is translated Chinese from native Chinese?: a corpus-based study of translation universals. Intern J Corp Ling15(1):5–35

Xiao R (2011) Word clusters and reformulation markers in Chinese and English: implications for translation universal hypotheses. Lang Contr 11(2):145–171

Xiao R, Yue M (2009) Using corpora in translation studies: The state of the art. In: Baker P (ed.) Contemporary corpus linguistics. Continuum, London, pp. 237–262

Xie Q, Yuan J (2013) A Genre analysis of report on the work of government. Theory Res10:155–156

Acknowledgements

This work is supported by the Social Science Foundation of Shaanxi Province (Grant No. 2023K001), the Fundamental Research Funds for the Central Universities (Grant No. SK2023031) and China Postdoctoral Science Foundation (Grant No. 2023M732817).

Author contributions

NJ: Conceptualization, data collection, methodology, data analysis and interpretation, writing-original draft. JY: Writing—revising and proofreading.

Competing interests

The authors declare no competing interests.

Ethical approval

Ethical approval was not required as the study did not involve human participants.

Informed consent

This article does not contain any studies with human participants performed by any of the authors.

Additional information

Correspondence and requests for materials should be addressed to Yue Jiang.

Reprints and permission information is available at http://www.nature.com/reprints

Publisher’s note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.