# Does simplification truly exist in translated Chinese? An exploration from a multilingual perspective

**Guo Zhen[*], Beijing Foreign Studies University**

## ABSTRACT

Since the introduction of the translation universal hypothesis, validation efforts have predominantly focused on English and other European languages. However, studies on translated Chinese, particularly those involving language pairs beyond English and Chinese, remain limited. This study investigates the simplification hypothesis by examining Chinese fiction translated from six different source languages. The findings suggest that the hypothesis is not universally applicable from a multilingual perspective, especially at the lexical and syntactic levels. This provides further evidence of a potential tendency toward complexification in translated Chinese. A clear disparity between Chinese translated from European and non-European languages has been observed, highlighting the significance of typology in this domain. A comparison with research on the news genre suggests that simplification and complexification are not mutually exclusive. Rather, translation complexity is shaped by multiple factors, including the typology of the source language, text genre, and the linguistic features selected for analysis. Future research should move beyond simply confirming or rejecting translation universals, and instead explore the interaction between linguistic features and contextual variables in order to uncover deeper patterns in translated language.

## KEYWORDS

Translation universals, simplification, complexification, typology, genre, multilingual corpus.

## 1. Introduction

In the field of translation studies, the concept of translated language, often referred to as a ‘third language code’ (Frawley,1984) or ‘translationese’ (Gellerstam, 1986), is characterised by linguistic features distinct from those of the original text. Building upon previous theoretical work (Toury, 1978; Blum-Kulka & Levenston, 1983; Blum-Kulka, 1986), Baker (1993) proposed the hypothesis of ‘translation universals’, which has since become a central focus of corpus-based translation studies over the past three decades. Various hypotheses, including explicitness, simplification, normalisation, among others, have been formulated and subjected to scrutiny across numerous language pairs. The concept of translation universals is characterised as ‘universal’ due to its consistent manifestation across language pairs, irrespective of specific linguistic systems. However, several unresolved issues remain within this domain.

First and foremost, there is a notable lack of horizontal comparisons across different language pairs. To comprehensively generalise the hypothesis of translation universals, research must go beyond English and its closely related languages (Xiao & Dai, 2010, p. 52). At present, international research on translation universals is largely confined to validation between English and various European languages. In contrast, studies conducted by Chinese scholars tend to focus primarily on the English–Chinese pair, with limited attention given to other language pairs. Second, the neglect of textual genre and the influence of translation direction (Hu et al., 2020, p. 273) has hindered a comprehensive understanding of how translation universals manifest across different genres and directionalities. In his investigation of explication

[*] ORCID 0009-0009-1520-2418, e-mail: guo-mark@bfsu.edu.cn

https://doi.org/10.26034/cm.jostrans.2025.8473

and implication in translation, Ke (2005) emphasised the crucial role of textual factors in shaping explicit and implicit patterns. Furthermore, ongoing challenges such as conceptual ambiguity, limited coverage of linguistic features, and inconsistent research findings underscore the need to investigate translation universals across a wider range of language pairs and genres in Chinese. Such efforts would substantially enhance the explanatory power of translation universals.

To address this research gap, the present study investigates Chinese fiction translated from six source languages, with a particular focus on the simplification hypothesis. The study aims to explore how this hypothesis manifests in multilingual contexts. Fiction is selected as the primary genre because, compared to legal and political texts, it tends to employ more colloquial language. This raises the question of whether such colloquial registers exhibit a consistent tendency toward simplification across different language pairs, which warrants further investigation. Although existing research on translated Chinese has extensively examined fiction, it has largely concentrated on English-to-Chinese translations. A broader cross-linguistic comparison therefore offers valuable insights. By engaging in comparative analysis with previous studies, this research seeks to shed light on the factors influencing the manifestation of translation universals and addresses the following research questions:

（1） Does the simplification hypothesis hold universally in Chinese translated from multilingual sources?

（2） How does linguistic complexity vary across different source languages, and which feature emerges as the most distinctive?

（3） Are there differences in linguistic complexity between translated fiction and other genres? How do these differences shed light on the interplay among source language, target language, and genre?

## 2. Literature review

The concept of simplification refers to the tendency to simplify the language used in translated text (Baker, 1996), a phenomenon observed at both lexical and syntactic levels. Laviosa (1998) compared translated English with original English and identified four indicators of simplification: lower lexical density, a higher proportion of high-frequency words to low-frequency words, increased repetition of more frequent words, and fewer lemmas in the list head of translated texts. Olohan & Baker (2000) further confirmed this tendency in translated English, demonstrating simplification across both lexical and syntactic dimensions. This study focuses on the simplification of translated Chinese fiction, a phenomenon supported by a series of empirical studies (Hu, 2007; Wang & Hu, 2008; Xiao, 2010; Hu & Kübler, 2021). Findings suggest that translated Chinese generally exhibits lower average sentence length, shorter clause length, reduced lexical density, and lower Standardised Type-Token Ratio (STTR), among other simplified features. Jiang et al. (2021) discovered that the average dependency distance in translated Chinese exceeds that of English source texts but is shorter than that of original Chinese, lending further support to the simplification hypothesis. However, the universality of simplification remains contested. Laviosa (1998) observed that translated texts can exceed original texts in mean sentence length, a finding

https://doi.org/10.26034/cm.jostrans.2025.8473

echoed by Xiao & Yue (2009), who reported that translated Chinese fiction tends to have longer mean sentence lengths than original Chinese fiction. Mauranen (2000) found that there are more atypical collocations in translated English, challenging the simplification hypothesis. Furthermore, Qin & Wang (2009) and Xiao & Dai (2014) also indicated that the simplification hypothesis may not be universally applicable to certain linguistic features. Wu et al. (2023) even proposed the co-existence of simplification and complexification in translated Chinese, based on an analysis of syntactic complexity. These findings indicate that there is no academic consensus on the simplification hypothesis. Its applicability appears to vary depending on language pair, genre, and the specific linguistic features under investigation, thereby warranting further empirical exploration.

Toury's 'law of interference' (Toury, 1995) posits that the source language influences the linguistic features of the translated language. This view aligns with Teich's (2003) concept of 'shining through', which similarly emphasises the structural imprint of the source language on the translation. In recent years, linguistic typology has emerged as a valuable perspective in corpus-based translation studies (Huang & Wang, 2023, p. 768). Due to language barriers and data limitations, comparative studies from a multilingual perspective remain relatively scarce, though some researchers have undertaken preliminary explorations. For example, Cappelle & Loock (2017) and Molés-Cases (2019) examined the influence of source language typology by comparing original texts with translated English and Spanish in the contexts of phrasal verbs and manner-of-motion expressions, respectively. Their findings suggest that language typology is a crucial variable restricting translation universal. Hu & Zeng (2017) compiled a multilingual corpus covering 20 source languages to examine source language interference in translated English. Their investigation revealed both similarities and differences in how various source languages influence translated English, shaped by factors such as linguistic typology, the relative status of languages and their literatures, and cognitive factors. In the context of translated Chinese, Hu & Kübler (2021) built a corpus of Chinese news translated from seven languages and compared it with original news texts from Xinhua News Agency. Their study supported the explicitation hypothesis at the lexical level but not at the syntactic level. Chen (2023) constructed a corpus of translated Chinese fiction and tested the 'levelling out' hypothesis across multiple language pairs. While the overall results aligned with the hypothesis, translations from Japanese deviated in aspects such as average sentence length and frequency of adverb usage. Although many studies have confirmed the impact of source language typology on translation, further research based on larger corpora is needed to explore the existence of translation universals in multilingual and cross-genre contexts.

While the simplification hypothesis has been widely examined in the context of translated Chinese, there remains a notable lack of comparative studies across different language pairs. This study revisits the hypothesis and extends its investigation to a multilingual setting. To enhance the rigor of our analysis, we draw insights from the study conducted by Hu & Kübler (2021), which investigated translated Chinese within the context of news genres.

## 3. Research design

https://doi.org/10.26034/cm.jostrans.2025.8473

To address the above research questions, this study employs a self-constructed corpus and extracts 15 linguistic features across three levels. This section outlines the research design in three parts: corpus composition, linguistic features under investigation, and research methodology.

### 3.1 Corpus composition

To address the research questions, this study compiled the Chinese Fiction Corpus (CNC), consisting of two main components: one comprising Chinese fiction translated from six source languages (five Indo-European and one non-Indo-European), and the other comprising a comparable corpus of original Chinese fiction. All texts were drawn from representative works of literary fiction in each language and were translated by native Chinese translators within the past two decades. Following data collection, all texts were converted into plain text (TXT) format and underwent cleaning procedures. For consistent comparison, the texts were segmented into units of approximately 5,000 characters, ensuring comparability in terms of production period, text length, and translator background. To maintain balance across sub-corpora with varying text volumes, 200 texts were randomly sampled from each sub-corpus, as detailed in Table 1. In total, the CNC comprises seven sub-corpora, amounting to nearly 8 million Chinese characters. The composition of the corpus is as follows:

| **Corpus component** | **Sub-Corpus** | **Number of texts** | **Total no. of characters** |
|---|---|---|---|
| Translated Chinese Fiction | Russian (RU) | 200 | 1,005,012 |
| | English (EN) | 200 | 1,195,070 |
| | German (DE) | 200 | 1,189,013 |
| | Japanese (JP) | 200 | 1,003,636 |
| | French (FR) | 200 | 1,196,052 |
| | Spanish (ES) | 200 | 1,195,022 |
| Original Chinese Fiction | Chinese (CN) | 200 | 1,193,903 |
| Total | 7 | 1,400 | 7,977,708 |

**Table 1: Composition of CNC**

### 3.2 Linguistic features under investigation

In previous examinations of simplification, there has been a notable emphasis on the lexical level, whereas the syntactic level has received comparatively less attention or has often been limited to features such as mean sentence length and mean clause length. To comprehensively evaluate the complexity of translated Chinese fiction, this study examines three dimensions: the lexical, syntactic, and collocational levels, incorporating a total of 15 indicators (see Table 2). At the lexical level, the features include average word length, Standardised Type-Token Ratio (STTR), lexical density,

https://doi.org/10.26034/cm.jostrans.2025.8473

and the percentage of four-character words, reflecting both vocabulary length and variability. These features are extracted using specific Python packages based on their operational definitions. For the syntactic and collocational levels, the analysis relies on the L2C-Rater tool developed by Beijing Normal University (Wang & Hu, 2021), which facilitates the extraction of approximately 90 syntactic complexity indicators. Given the inclusion of seven sub-corpora in this comparative study, six representative syntactic indicators and five collocational indicators are selected to ensure a focused and comprehensive analysis. The definitions of these indicators are drawn from Hu & Xiao (2019), Hu (2021), and Hu et al. (2022).

The first three features of the Collocational level represent the diversity of collocations. Among them, TOTAL_RTTR is a comprehensive evaluation of general and unique collocations in Chinese. General collocations include four types of common combinations: verb-object, subject-predicate, adjective-noun, and adverb-predicate. Unique collocations refer to those specific to the Chinese language, including classifier-noun, preposition-postposition, preposition-verb, predicate-complement and connective-connective. According to Hu (2021), RTTR (Root Type-Token Ratio) is a modified version of the traditional Type-Token Ratio, used to evaluate the variety of collocations in a given text. It is calculated by dividing the number of types by the square root of the number of tokens, which reduces the influence of text length. The final two features refer to the ratios of unique collocations and low-frequency collocations, respectively.

| Dimension | Code | Feature |
|---|---|---|
| Lexical level | AWL | Average word length |
| | STTR | Standardised type-token ratio |
| | LD | Lexical density |
| | FCW | Percentage of four-character words |
| Syntactic level | MLS | Mean length of sentence |
| | MLC | Mean length of clause |
| | NCPS | Number of clauses per sentence |
| | MLTU | Mean length of T-unit |
| | NTPS | Number of T-units per sentence |
| | MTD | Mean tree depth |
| Collocational level | TOTAL_RTTR | Diversity of total collocations |
| | GENERAL_RTTR | Diversity of general collocations |
| | UNIQUE_RTTR | Diversity of unique collocations |
| | UNIQUE_RATIO | Ratio of unique collocations |

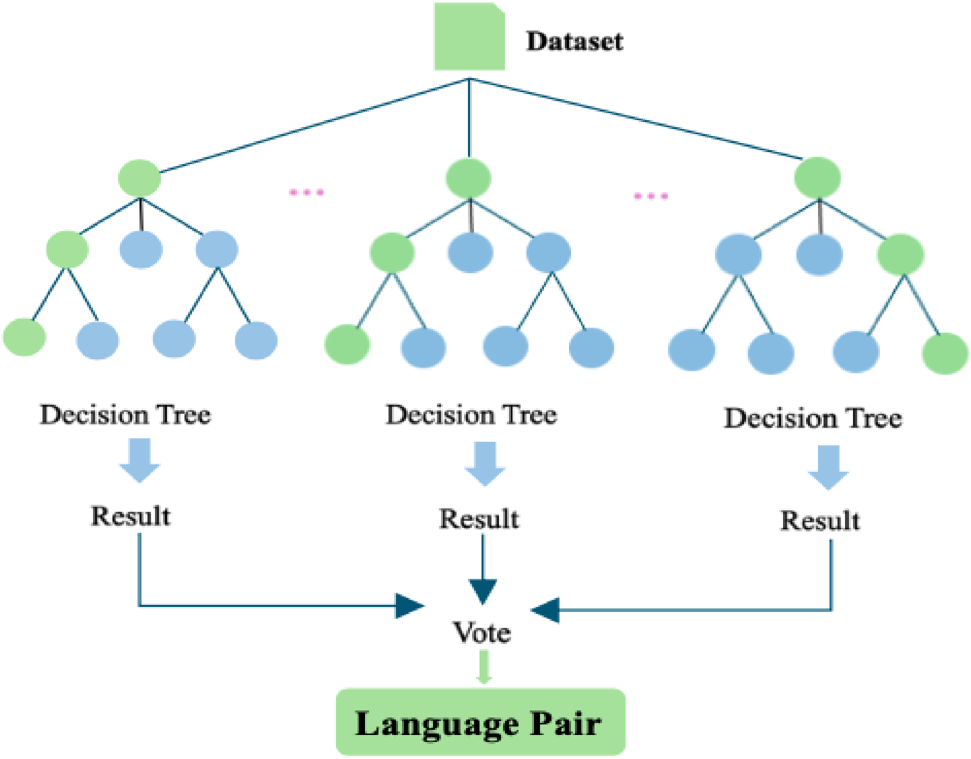
| | | |
|---|---|---|
| | LOWFREQ_RATIO | Ratio of low frequency collocations |

**Table 2. Features extracted**

### 3.3 Research method

Python was used to extract the values of these 15 features for each of the 1,400 texts across the sub-corpora. Before comparing translated fiction and original fiction, the normality of data distribution within each group was assessed using the Shapiro-Wilk test. As certain groups deviated from a normal distribution, Kruskal-Wallis non-parametric tests were employed for within-group comparisons. Post hoc comparisons were conducted using Mann-Whitney U tests, with the significance level set at 0.05.

In addition, Random Forest text classification, a machine learning method, was applied to test whether the extracted features could effectively distinguish between different language pairs. Random Forest is an ensemble learning algorithm used for classification and regression tasks. It uses bootstrap resampling to generate multiple subsets from the original dataset, constructs a decision tree for each subset, and aggregates the predictions of all trees through majority voting to produce the final output (see Figure 1). Compared to a single decision tree, Random Forest improves the model's generalisation ability by randomly sampling the dataset with replacement and aggregating predictions from multiple decision trees. This method is widely applied in fields such as medical diagnosis, financial risk assessment, and customer behaviour analysis. The model training uses the scikit-learn library in Python (Pedregosa et al., 2011), utilizing the Random Forest Classifier for classification.



https://doi.org/10.26034/cm.jostrans.2025.8473

**Figure 1. Random Forest text classification**

In this study, 80% of the texts were used for training, while the remaining 20% served as the test set. By comparing the model's predicted results with the actual outcomes, we evaluate its precision, recall, and F1-score. If the classification can accurately distinguish among different language pairs, it will not only confirm the impact of typology on translated language, but also validate the adopted feature framework. The model also provides the weights of linguistic features in the classification, which helps identify the most salient features across the six language pairs. This approach not only captures the influence of typology, but also reveals systematic differences in translated Chinese associated with different source languages. Ultimately, this contributes to advancing research on translation universals by identifying the most significant complexity features, which is one of the innovative aspects of this study.

## 4. Results

The results of the Kruskal-Wallis tests revealed significant differences across all features at the lexical, syntactic, and collocational levels ($p < 0.001$), underscoring the impact of source language differences on the linguistic patterns of translated Chinese across these dimensions. To visualise the data distribution and support post hoc comparisons, violin plots were employed (the point within each plot represents the median value of each feature, providing a complementary perspective on the data distribution). For clarity and conciseness, we adopted a convention in which the source language label is used to represent the corresponding language pair. For example, 'English' refers to 'Chinese translated from English'.

### 4.1 Lexical level

Table 3 presents the mean and standard deviation of different lexical features, while Figure 2 illustrates the overall distribution of the data across language pairs.

| Feature | AWL | | STTR | | LD | | FCW | |
|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| German | 1.53 | 0.05 | 51.57 | 5.34 | 0.56 | 0.02 | 0.03 | 0.01 |
| Russian | 1.50 | 0.05 | 51.75 | 4.45 | 0.58 | 0.02 | 0.02 | 0.01 |
| French | 1.51 | 0.05 | 54.44 | 3.87 | 0.59 | 0.02 | 0.02 | 0.01 |
| Japanese | 1.51 | 0.04 | 52.71 | 3.70 | 0.61 | 0.02 | 0.02 | 0.01 |
| Spanish | 1.57 | 0.10 | 52.20 | 3.84 | 0.58 | 0.04 | 0.03 | 0.02 |
| English | 1.45 | 0.06 | 49.85 | 4.44 | 0.57 | 0.02 | 0.01 | 0.01 |
| Chinese | 1.48 | 0.06 | 49.62 | 4.64 | 0.62 | 0.03 | 0.02 | 0.01 |
| Kruskal-Wallis Test | H = 840.79<br>p < 0.001 | | H = 385.00<br>p < 0.001 | | H = 533.33<br>p < 0.001 | | H = 750.48<br>p < 0.001 | |

**Table 3. Lexical complexity**

https://doi.org/10.26034/cm.jostrans.2025.8473

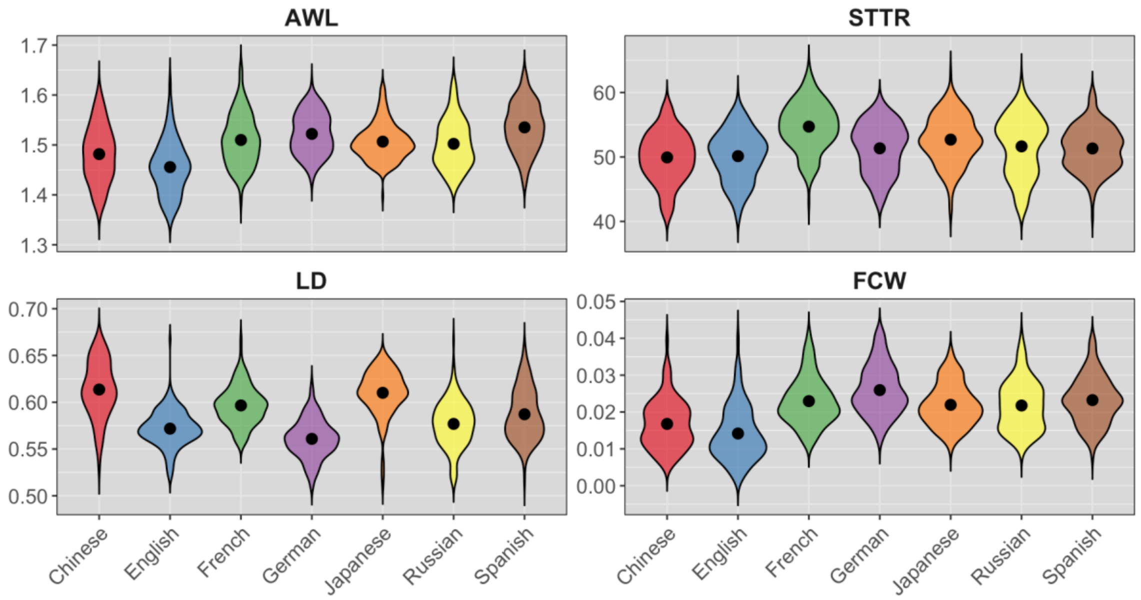


**Figure 2. Lexical complexity distribution**

According to table 3, among the six translation sub-corpora, only English exhibited a shorter AWL (average word length) compared to the original texts, aligning with the simplification hypothesis and consistent with findings from previous research on the English-Chinese pair. The mean values of the other sub-corpora exceeded those of the original texts, with statistically significant differences (p < 0.05). Specifically, Spanish had the longest AWL, followed by German. Thus, with regard to AWL, five out of the six language pairs did not support the simplification hypothesis. This result corresponds with the data on the FCW (percentage of four-character words), where significant differences were observed between all translated and original texts (p < 0.05). Notably, Chinese translated from English exhibited the lowest FCW usage, consistent with the simplification hypothesis ($z$ = −6.30, $p$ = 0.000 < 0.05), while the other five out of the six language pairs used more FCW, indicating a tendency towards ‘complexification’. Moreover, no significant differences were found among translations from Japanese, French, and Russian (p > 0.05), nor between Spanish and German ($z$ = −1.387, $p$ = 0.166 > 0.05). After further examination of the percentage of one-character words, we found that original Chinese texts had the highest proportion (0.46), indicating that original Chinese fiction had a shorter average word length due to a high proportion of one-character words. This is primarily due to the dominance of monosyllabic vocabulary in colloquial fiction, as opposed to the polysyllabic and disyllabic words prevalent in general translated Chinese.

The STTR (Standardised Type-Token Ratio), proposed by Scott (2004), was used to evaluate lexical richness. It was calculated by dividing the number of types (unique words) by the number of tokens, based on successive segments of 1,000 words. Across all six language pairs, the STTR values exceeded those of the original Chinese. English closely resembled the original Chinese in this respect, with post hoc analysis revealing no significant difference between them ($z$ = 0.225, $p$ = 0.822 > 0.05), which aligns with Xiao’s (2010) findings on STTR in English-Chinese translations. However, significant differences were found between the other language pairs and the original Chinese (p < 0.05). This contrasts with Hu’s (2007) earlier research, which suggested

https://doi.org/10.26034/cm.jostrans.2025.8473

lower lexical variation in translated Chinese fiction compared to original fiction, indicating that the simplification of lexical richness is not consistent across multilingual pairs. In five out of six language pairs, a tendency towards greater lexical diversity was observed, suggesting a shift towards 'complexification' in lexical usage.

Contrary to the performance on STTR, translated texts, although exhibiting higher lexical richness, consistently demonstrated lower lexical density (LD) compared to the original Chinese. Stubbs (1986, p. 33) defines lexical density as the ratio of content words to the total number of tokens in a corpus, and it is also regarded as an indicator of lexical richness. The data show that the original Chinese had the highest LD, followed by Japanese, with no statistically significant difference between them ($z = -1.604$, $p = 0.109 > 0.05$), thereby contradicting the simplification hypothesis. In contrast, Chinese translated from Indo-European languages exhibited significantly lower LD than the original Chinese ($p < 0.05$). This finding aligns with previous examinations of vocabulary in translated and original Chinese texts (Hu, 2007; Wang & Hu, 2008; Xiao & Yue, 2009; Xiao, 2010), and is consistent with Xiao & Dai's (2010) study across multiple genres, which found that the original Chinese consistently displayed higher lexical density than translated Chinese. These results reflect the more frequent use of content words—such as nouns, verbs, and adjectives—in original fiction, and a heavier reliance on function words—such as prepositions, conjunctions, and adverbs—in translated fiction. This pattern has been widely observed in both multilingual and cross-genre studies. This study further confirms the hypothesis that translated Chinese has lower lexical density at the multilingual level, consistent with Laviosa's (1998) findings. As Hu (2007, p. 219) notes, "the decreased lexical density in translated fiction, as compared to originals, is largely due to translators' efforts to mitigate the complexity of the translated text by reducing the informational content conveyed by content words, thereby enhancing its acceptability." However, this study also found that the average lexical density of Chinese translated from Japanese closely resembled that of the original Chinese, with no statistically significant difference ($z = –1.604$, $p = 0.109 > 0.05$). This result again challenges the simplification hypothesis and suggests a potential correlation between language typology and levels of linguistic complexity.

The aforementioned lexical features indicate that the simplification hypothesis is not fully supported in colloquial fiction genres. Among these features, lexical density is the only metric that consistently supports the simplification hypothesis. The original Chinese fiction tends to have higher lexical density, primarily utilising content words to convey information. In contrast, translated Chinese fiction exhibits greater vocabulary diversity and longer word lengths, indicating a trend towards 'complexification'. However, despite this lexical diversity, the proportion of content words in translated Chinese fiction is lower than that in original Chinese, leading to reduced information density. The disparities observed between Japanese and Indo-European languages underscore the impact of source language typology on translated language. It is also essential to acknowledge that variations in the selection of features may yield different results.

## 4.2 Syntactic level

Table 4 reports the mean and standard deviation for different syntactic features,

https://doi.org/10.26034/cm.jostrans.2025.8473

whereas Figure 3 provides an overview of their distribution across language pairs.

| Feature | MLS | | MLC | | NCPS | | MLTU | | NTPS | | MTD | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| German | 37.18 | 12.26 | 11.11 | 1.44 | 3.30 | 0.79 | 14.07 | 2.49 | 2.60 | 0.50 | 6.68 | 0.67 |
| Russian | 26.67 | 5.82 | 9.91 | 0.84 | 2.68 | 0.45 | 12.15 | 1.24 | 2.18 | 0.30 | 5.87 | 0.43 |
| French | 25.85 | 5.95 | 9.82 | 1.01 | 2.61 | 0.43 | 11.97 | 1.33 | 2.14 | 0.34 | 5.94 | 0.50 |
| Japanese | 18.57 | 3.74 | 9.03 | 1.00 | 2.04 | 0.22 | 10.68 | 1.34 | 1.73 | 0.16 | 5.42 | 0.36 |
| Spanish | 27.68 | 9.76 | 10.90 | 2.20 | 2.50 | 0.53 | 14.01 | 3.59 | 1.96 | 0.38 | 6.21 | 0.65 |
| English | 29.28 | 10.27 | 10.40 | 1.51 | 2.78 | 0.81 | 12.95 | 2.39 | 2.22 | 0.57 | 6.22 | 0.76 |
| Chinese | 27.27 | 6.72 | 9.28 | 0.96 | 2.94 | 0.63 | 10.75 | 1.22 | 2.53 | 0.51 | 6.18 | 0.46 |
| Kruskal-Wallis Test | H = 452.25<br>p < 0.001 | | H = 325.31<br>p < 0.001 | | H = 490.89<br>p < 0.001 | | H = 443.02<br>p < 0.001 | | H = 580.22<br>p < 0.001 | | H = 435.38<br>p < 0.001 | |

**Table 4. Syntactic complexity**

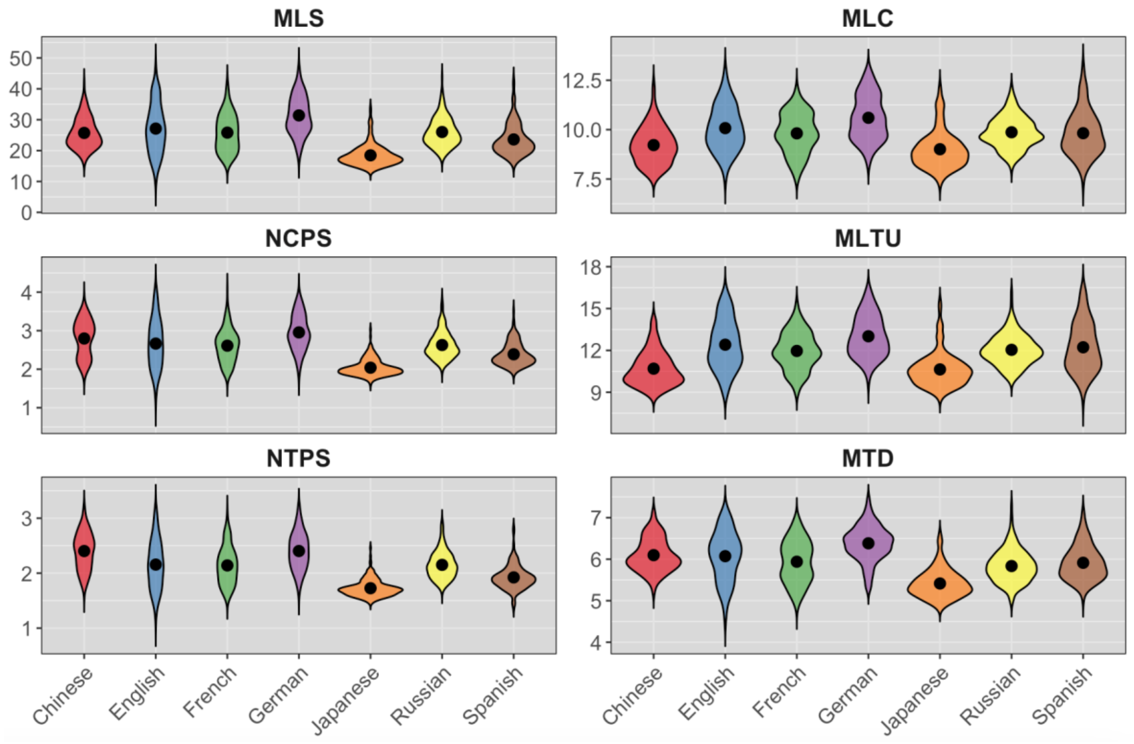


**Figure 3. Syntactic complexity distribution**

As depicted in Table 4, the mean length of sentences (MLS) in Chinese translated from German, English, and Spanish seemed to be longer than that in the original Chinese. However, the Mann-Whitney U test showed that there was no significant difference between the Chinese translated from Spanish and the original Chinese ($z = -1.14$, $p = 0.256 > 0.05$). In contrast, the mean lengths of Russian, French, and Japanese were shorter than those of the original Chinese. Among them, only Japanese had a significant difference ($p < 0.05$) from the original Chinese, while there were no significant differences between Russian–Chinese and French–Chinese translations compared to the original Chinese. Given the paratactic nature of Chinese, in which commas are used to link multiple clauses, a full sentence typically comprises several clauses. Thus, the MLC (mean length of clauses) or segments may better reflect the true structure of Chinese sentences. The results revealed that the mean length of clauses for all Indo-European languages exceeded that of the original Chinese, displaying significant differences ($p < 0.05$), contradicting the simplification hypothesis. This finding supports the conclusion by Xiao & Dai (2014), indicating that translated Chinese tends to have longer MLC in the English–Chinese pair, extending this observation to a broader range of Indo-European languages. The prolonged MLC in translated languages stems from the “overuse of structural extension in Chinese” (Wang & Qin, 2009, p. 105), wherein the structural capacity is expanded to accommodate complex modifier elements from the source language. However, unlike Indo-European languages, Japanese exhibited shorter MLS than the original Chinese, with significant differences ($z = -14.382$, $p = 0.000 < 0.05$), and also shorter MLC, with significant differences ($z = -3.188$, $p = 0.001 < 0.05$), contradicting previous findings based on research focusing on Indo-European and Chinese pairs.

The NCPS (number of clauses per sentence) indicates the number of clauses within a sentence. The data reveal that the Chinese translated from German exhibits a higher NCPS than the original Chinese ($z = 4.356$, $p = 0.000 < 0.05$), consistent with the previously observed longer MLS for the German–Chinese pair. Conversely, the NCPS for other language pairs was lower than that of the original Chinese, showing significant differences ($p < 0.05$). In this regard, five out of the six language pairs supported the simplification hypothesis, with the Japanese-Chinese pair displaying the lowest NCPS. Lu (1979, pp. 23–24) proposed that Chinese parataxis exhibits the characteristic of being ‘connectable’ and ‘separable’—a concept further summarised by Wang (2019) as ‘sentential chunkiness and discreteness’, contrasting with the English structural pattern of ‘connection and continuity’ in syntax. This inherent structural difference between English and Chinese is reflected in translated Chinese texts. When expanding beyond the English–Chinese pair, the syntactic differences observed across multilingual pairs underscore the influence of language type on translated language.

The MLTU (mean length of T-unit), an indicator widely used to evaluate Chinese syntactic complexity, has been validated and applied in numerous studies (An, 2015; Hu, 2021; Wu et al., 2022; Wu, 2023). Its performance consistently reflects the trends observed in MLC. The intricate syntactic structures of Indo-European languages are evident in translated Chinese, leading to higher MLTU values. Specifically, five out of six translations surpass the original Chinese, demonstrating significant differences ($p < 0.05$) and suggesting a tendency towards complexification. For Chinese translated from Japanese, the mean length of T-unit (10.68) is slightly lower than that of the original Chinese (10.75), although this difference is not statistically significant ($z = -0.861$, $p = 0.389 > 0.05$). Similarly, the distribution of NTPS (Number of T-units per Sentence) mirrors that of NCPS. German exhibits the highest values, followed by original Chinese, with no significant difference between them ($z = 0.689$, $p = 0.091 > 0.05$). Conversely, translations from other languages demonstrate lower values compared to the original Chinese, supporting the simplification hypothesis. Regarding the MTD (mean tree depth), translations from German, Spanish, and English exceed the original Chinese, contradicting the simplification hypothesis. However, the English–

https://doi.org/10.26034/cm.jostrans.2025.8473

Chinese ($z = -1.22$, $p = 0.223 > 0.05$) and Spanish–Chinese ($z = 0.708$, $p = 0.475 > 0.05$) pair show no significant differences. In contrast, translations from French, Russian, and Japanese fall short of the original Chinese ($p < 0.05$), further supporting the simplification hypothesis.

In summary, at the syntactic complexity level, Indo-European languages consistently demonstrate a trend towards increased complexity, as indicated by NTPS and NCPS. Particularly, German consistently exhibits complexification across all six indicators. Conversely, Japanese consistently demonstrates simplification across these indicators. The significant variation in complexity among different language pairs underscores the influence of the linguistic structures of the source language on translated language. It should also be noted that factors such as literary traditions and an author's style may influence sentence complexity. However, as a corpus-based study, the findings of this research are derived from large-scale statistical analysis. Therefore, when a general trend emerges across the selected texts, it suggests that the results are not merely isolated cases and, to some extent, reflect the characteristics of certain groups. The comparison between original and translated texts both confirms and contradicts the simplification hypothesis, highlighting the possibility of divergent outcomes depending on which features are selected.

### 4.3 Collocational level

Table 5 reports the mean and standard deviation for different collocational features, whereas Figure 4 provides an overview of their distribution across language pairs.

| **Feature** | **TOTAL_RTTR** | | **UNIQUE_RTTR** | | **GENERAL_RTTR** | | **UNIQUE_RATIO** | | **LOWFREQ_RATIO** | |
|---|---|---|---|---|---|---|---|---|---|---|
| | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* | *M* | *SD* |
| German | 28.99 | 1.66 | 25.94 | 1.49 | 12.25 | 1.12 | 7.89 | 1.05 | 4.36 | 0.51 |
| Russian | 25.97 | 1.00 | 23.38 | 0.90 | 10.57 | 0.88 | 6.57 | 0.91 | 3.82 | 0.45 |
| French | 29.54 | 1.02 | 26.48 | 0.89 | 12.36 | 1.02 | 7.57 | 1.11 | 4.14 | 0.43 |
| Japanese | 27.32 | 0.97 | 24.85 | 0.85 | 10.98 | 5.47 | 6.02 | 0.94 | 4.34 | 3.38 |
| Spanish | 28.23 | 3.02 | 25.35 | 2.63 | 11.70 | 1.83 | 7.43 | 1.52 | 4.48 | 0.65 |
| English | 29.23 | 1.60 | 26.34 | 1.49 | 11.91 | 1.27 | 7.22 | 1.26 | 3.74 | 0.56 |
| Chinese | 29.60 | 1.89 | 26.28 | 1.66 | 12.90 | 1.33 | 8.55 | 1.50 | 4.33 | 0.49 |
| Kruskal-Wallis | H = 610.49<br>p < 0.001 | | H = 570.41<br>p < 0.001 | | H = 494.46<br>p < 0.001 | | H = 437.97<br>p < 0.001 | | H = 303.06<br>p < 0.001 | |

**Table 5. Collocational complexity**

https://doi.org/10.26034/cm.jostrans.2025.8473

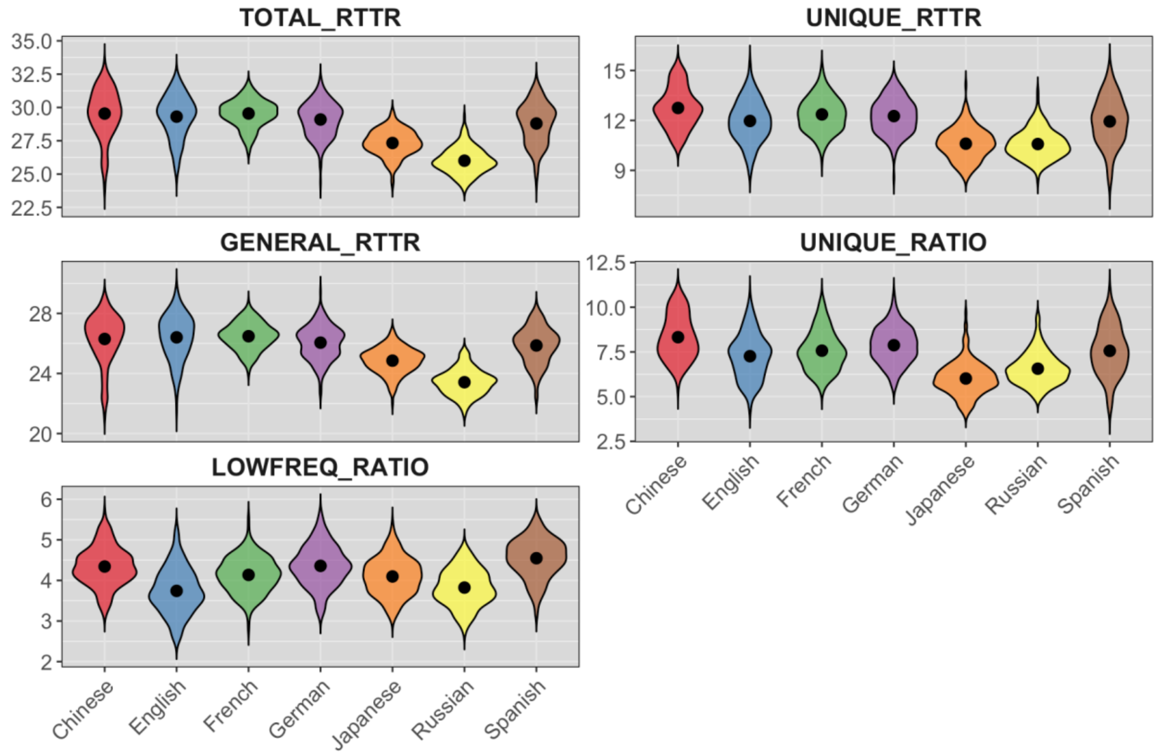


**Figure 4. Collocational complexity distribution**

As shown in Table 5, significant differences are evident among all groups at the collocational complexity level (p < 0.001). The original Chinese consistently ranks higher across all five indicators, with the highest values observed in TOTAL_RTTR, UNIQUE_RTTR, and UNIQUE_RATIO. Except for French, which shows no significant difference from the original Chinese in TOTAL_RTTR ($z = -1.445$, $p = 0.148 > 0.05$), all other languages exhibit significant differences, thereby supporting the simplification hypothesis in translated Chinese. In certain features, such as GENERAL_RTTR, the mean value of the original Chinese is lower than those of English ($z = 0.290$, $p = 0.772 > 0.05$) and French ($z = 0.655$, $p = 0.512 > 0.05$), but these differences are not statistically significant. Similarly, regarding LOWERFREQ_RATIO, although the mean value of the original Chinese is lower than those of Spanish and German, there is no significant difference between Chinese and German ($z = -0.603$, $p = 0.547 > 0.05$). Overall, only Spanish significantly exceeds the original Chinese in LOWERFREQ_RATIO ($z = 3.937$, $p = 0.001 < 0.05$). Notably, the original Chinese fiction excels at the collocational level, displaying greater diversity in collocations and a higher frequency of Chinese-specific expressions. The influence of the source language is thus reaffirmed: differences in language systems mean that the collocations of foreign languages cannot fully align with those of Chinese, resulting in a lower frequency of unique collocations in translated texts. Consequently, translated Chinese tends to simplify, often relying on repetitive or homogeneous collocations. In conclusion, the simplification hypothesis applies to translated Chinese at the collocational level, with noticeable variation across different language pairs, further underscoring the impact of typology.

## 4.4 Random forest text classification

https://doi.org/10.26034/cm.jostrans.2025.8473

Through the above analysis of lexical, syntactic, and collocational features, we observe fluctuating complexity in translated Chinese. It remains difficult to determine which linguistic feature diverges the most and which set of features can most reliably classify translated Chinese by source language so as to elucidate typological influence. Hence, this study applies a Random Forest model for text classification.

| **Language** | **Precision** | **Recall** | **F1-score** | **Text** |
|---|---|---|---|---|
| German | 0.57 | 0.65 | 0.60 | 40 |
| Russian | 0.92 | 0.85 | 0.89 | 40 |
| French | 0.76 | 0.80 | 0.78 | 40 |
| Japanese | 0.71 | 0.80 | 0.75 | 40 |
| Spanish | 0.66 | 0.60 | 0.63 | 40 |
| English | 0.81 | 0.75 | 0.78 | 40 |
| Chinese | 0.85 | 0.83 | 0.85 | 40 |
| **Mean** | 0.75 | 0.75 | 0.75 | **Total：** 280 |

**Table 6. Result of the Random Forest Classification Model**

Precision: The proportion of true positive predictions out of all positive predictions made by the model; Recall: The proportion of true positive predictions out of all actual positive instances in the dataset; F1-Score = [ Precision × Recall / (Precision + Recall)] × 2

As shown in Table 6, the model achieves an average precision, recall, and F1-score of 0.75, indicating a relatively accurate classification performance. The F1-score, regarded as a more comprehensive evaluation metric, shows that the model performs best when classifying Russian texts (0.89), followed by Chinese (0.85), French (0.78), and English (0.78), with the lowest performance when classifying German texts (0.60). Figure 5 illustrates the contribution of 15 linguistic features to the classification model, with higher values indicating greater importance in classification.

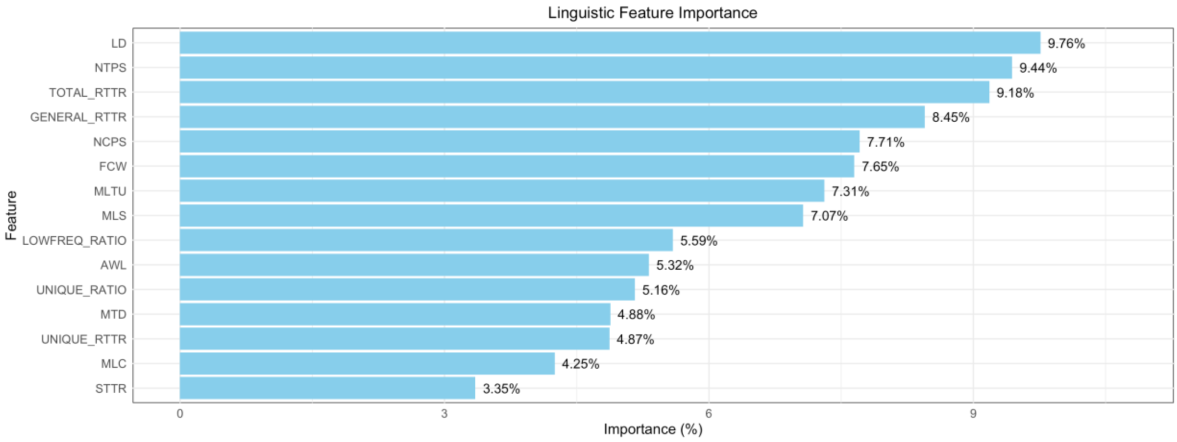


**Figure 5. Linguistic feature importance**

https://doi.org/10.26034/cm.jostrans.2025.8473

Among all indicators, LD, NTPS, TOTAL_RTTR, GENERAL_RTTR, and NCPS stand out as the most distinguishing features. Lexical density, the most significant factor in text classification, underscores differences in the use of content words in translated Chinese fiction rendered from various languages, potentially reflecting the distinct language systems of the source texts. In contrast, frequently used indicators for testing the simplification hypothesis, such as STTR and MLC, appear to have the lowest importance. These results suggest that collocational and syntactic indicators must be included when investigating the simplification or complexity of translated language. Examining only lexical features or relying on shallow syntactic measures such as MLC, which is based on the number of Chinese characters, does not yield accurate conclusions. The accurate classification achieved by the Random Forest model across different language pairs further substantiates the influence of typology on the complexity of translated language and demonstrates that the feature framework used in this study is robust, as it effectively captures differences among selected language pairs.

## 5. Discussion

The preceding analyses of lexical, syntactic, and collocational complexity reveal a nuanced interplay between simplification and complexification in translated Chinese fiction across different language pairs. This pattern challenges the notion of a straightforward simplification hypothesis, instead suggesting varying degrees of complexity across linguistic features and language pairs. Indo-European languages demonstrate a propensity for greater complexity at the lexical and syntactic levels, while Japanese, as a non-Indo-European language, tends towards simplification overall. However, at the collocational level, the original Chinese exhibits greater complexity, aligning with the simplification hypothesis in translation universals. The degree of simplification and complexification appears to be intricately linked to the characteristics of the source languages and is also influenced by the selection of indicators. Assessing language complexity is a multifaceted endeavour. Previous studies, such as those by Xiao & Dai (2014) and Fu & Wang (2021), have highlighted the multidimensional nature of simplification, indicating that translated languages may exhibit both simplified and complex features compared to the originals. In addition, “the performance and degree of explication are related to the nature of the source and target language” (Qin & Wang, 2009, p. 136). Consequently, the concept of translation universals necessitates a comprehensive examination, taking into account numerous factors and requiring exploration from diverse perspectives.

Hu & Kübler (2021) investigated the simplification of translated Chinese news from a multilingual perspective. Their findings supported the simplification hypothesis across features such as mean word length, type-token ratio, lexical density, and mean sentence length. They also highlighted that lexical density appears to be influenced by the source language, as evidenced by the differing performance of Japanese and Korean compared to other Indo-European languages. However, our study's results diverged on several key features, including TTR, mean word length, and mean sentence length. At the syntactic level, Hu & Kübler (2021, p. 355) noted that “the simplification hypothesis may be too simplified to capture the whole picture.” When comparing the language utilised in the original news texts and that in the original fictional texts, it becomes apparent that the news genre is intrinsically more complex.

https://doi.org/10.26034/cm.jostrans.2025.8473

In comparison to fictional texts, news texts demonstrate higher values in aspects like average word length, lexical density, and average sentence length. However, the distribution of simplification and complexification of translated Chinese in these features between the two genres is opposite. While the fiction texts tend to be simpler (without considering semantics), translated fiction texts exhibit more pronounced features of complexification. Based on the above comparisons, we can draw the following inferences:

（1） Language typology plays a significant role in influencing the degree of simplification or complexification in translated language. For instance, the features of Japanese often shine through and manifest differently in translated Chinese. Due to its similarity to Chinese, the Japanese–Chinese pair shows distinct patterns compared to translations from other language types. Such findings are consistent with Nida's (1964, pp. 160–161) classification of linguistic and cultural distance, where Japanese and Chinese fall into the category of 'culturally close and linguistically related', which may influence the translation process.

（2） The degree of simplification or complexification can fluctuate within the same language pair across different genres. The complexity of the source text affects the translator's ability to align the target language with the original. For example, translating children's literature poses unique challenges in reproducing the original text's simplicity and clarity, which may inadvertently lead to increased complexity in the translated text.

（3） The diverse findings across lexical, syntactic, and collocational complexity underscore the need for a more comprehensive selection of language features for investigating translation universals. For coarse-grained complexity indicators like AWL, STTR and MLC, which are calculated based on the number of Chinese characters, Chinese translated from Indo-European languages appears more complex. However finer-grained indicators like LD and other collocational complexity features reveal that original Chinese tends to be more complex. Hu et al.'s (2020) investigation on translated English across several genres and features found that only a fraction of selected language features conformed to translation universals across genres. This highlights the significant influence of features selection on research conclusions and emphasises the need for further exploration in defining simplification and complexification and selecting appropriate features.

## 6. Conclusion

This study examines the applicability of the simplification hypothesis in translated Chinese fiction from a multilingual perspective. Analyses at the lexical, syntactic, and collocational levels reveal nuanced patterns that challenge the universality of the simplification hypothesis. These findings provide valuable insights into translation universals, underscoring the significant role of language typology in shaping the linguistic complexity of translated languages. They also highlight the importance of text genres and feature selection in any analysis of translated languages. Although the study does not fully support the existence of simplification, it underscores the

https://doi.org/10.26034/cm.jostrans.2025.8473

continuing importance of researching the distinct features of translated languages. Rather than merely confirming or refuting hypotheses, future studies should focus on the interplay between linguistic features and other variables using multifactorial methods. A comparative approach across different translated languages is needed to uncover their unique characteristics, which is crucial for understanding the underlying processes that shape translated languages.

The formation of translated languages results from the interplay of various factors. In the fiction genres analysed in this study, alongside the influence of linguistic structures, the literary traditions of both the source and target languages exert a significant impact. The asymmetry between these traditions can also affect the eventual realisation of the translated language. However, this aspect was not explored in the present study due to the author’s limited expertise. Future research could include collaboration with scholars from diverse linguistic backgrounds to establish a multilingual parallel corpus. Such a corpus, combined with more in-depth studies of source languages, could further elucidate how literary traditions influence the translation process.

## Acknowledgments/Funding

This research was funded by the National Social Science Fund of China to The Project “Construction and Application of the Mega-sized Multi-Language Parallel Corpora with Chinese as Pivot Language” (21&ZD290) and the Fundamental Research Funds for Central Universities (2024JX088). I extend my sincere gratitude to Professor Wang Kefei for his invaluable guidance in both selecting the research topic and refining this paper. I would also like to thank Li Mengfei for his assistance in ensuring the smooth operation of the code used in this study.

## References

An, F. (2015). Analysis of fluency, grammatical complexity, and accuracy of CSL writing: A study based on T-unit analysis. *Language Teaching and Linguistic Studies*, *37*(3), 11–20.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and Technology: In Honor of John Sinclair* (pp. 233–250). John Benjamins. https://doi.org/10.1075/z.64.15bak

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and Translation Studies in Language Engineering: In Honor of Juan C. Sager* (pp. 175–186). John Benjamins. https://doi.org/10.1075/btl.18.17bak

Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In J. House & S. Blum-Kulka (Eds.), *Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition Studies* (pp. 17–35). Gunter Narr.

Blum-Kulka, S., & Levenston, E. (1983). Universals of lexical simplification. In C. Faerch & G. Kasper (Eds.), *Strategies in Interlanguage Communication* (pp. 119–139). Longman.

Cappelle, B., & Loock, R. (2017). Typological differences shining through: The case of phrasal verbs in translated English. In G. De Sutter, M.-A. Lefer, & I. Delaere (Eds.), *Empirical Translation Studies: New Methodological and Theoretical Traditions* (pp. 235–264). De Gruyter Mouton. https://doi.org/10.1515/9783110459586-009

Chen, Y. (2023). *Levelling out in translational literary Chinese: A corpus-based study across multiple*

https://doi.org/10.26034/cm.jostrans.2025.8473

*language pairs* (Master's thesis, Beijing Foreign Studies University). https://kns.cnki.net/KCMS/detail/detail.aspx?dbname=CMFD202401&filename=1023061159.nh

Frawley, W. (1984). Prolegomenon to a theory of translation. In W. Frawley (Ed.), *Translation: Literary, Linguistic and Philosophical Perspectives* (pp. 159–175). Associated University Press.

Fu, R., & Wang, K. (2021). Lexical patterns in interpreted and spontaneous English speeches: A comparable, intermodal, and corpus-based study [in Chinese]. *Foreign Language Teaching and Research*, *53*(6), 912–923. https://link.oversea.cnki.net/doi/10.19923/j.cnki.fltr.2021.06.010[1]

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin & H. Lindquist (Eds.), *Translation Studies in Scandinavia* (pp. 88–95). C.W.K. Gleerup.

Hu, H., & Kübler, S. (2021). Investigating translated Chinese and its variants using machine learning. *Natural Language Engineering*, *27*(3), 339–372. https://doi.org/10.1017/S1351324920000182

Hu, R. (2021). On the relationship between collocation-based syntactic complexity and Chinese second language writing [in Chinese]. *Applied Linguistics* (《语言文字应用》), *30*(1), 132–144. https://link.oversea.cnki.net/doi/10.16499/j.cnki.1003-5397.2021.01.017

Hu, R., & Xiao, H. (2019). The construction of Chinese collocation knowledge bases and their application in second language acquisition [in Chinese]. *Applied Linguistics* (《语言文字应用》), *1*(1), 135–144. https://link.oversea.cnki.net/doi/10.16499/j.cnki.1003-5397.2019.01.017

Hu, R., Wu, J., & Lu, X. (2022). Word-combination-based measures of phraseological diversity, sophistication, and complexity and their relationship to second language Chinese proficiency and writing quality. *Language Learning*, *72*(4), 1128–1169. https://doi.org/10.1111/lang.12511

Hu, X. (2007). A corpus-based study on the lexical features of Chinese translated fiction. *Foreign Language Teaching and Research*, *39*(3), 214–220.

Hu, X., & Zeng, J. (2017). A stylo-statistical analysis of source language interference in translational English. *Foreign Language Teaching and Research*, *49*(4), 595–607.

Hu, X., Xiao, R., & Hardie, A. (2020). A corpus-based multi-feature stylo-statistical analysis of translational English [in Chinese]. *Foreign Language Teaching and Research*, *52*(2), 273–282. https://link.oversea.cnki.net/doi/10.19923/j.cnki.fltr.2020.02.010

Huang, L., & Wang, K. (2023). Development stages and trends of corpus-based translation studies. *Foreign Language Teaching and Research*, *55*(5), 764–776.

Jiang, Y., Fan, L., & Wang, Y. (2021). Research on syntactic features of translated language based on a dependency treebank [in Chinese]. *Foreign Language Education*, *42*(3), 41–46. https://link.oversea.cnki.net/doi/10.16362/j.cnki.cn61-1023/h.2021.03.007

Ke, F. (2005). Implication and explication in translation. *Foreign Language Teaching and Research*, *37*(4), 303–307.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta*, *43*(4), 557–570. https://doi.org/10.7202/003425ar

Lu, S. (1979). *Problems of Chinese grammatical analysis*. Commercial Press.

Mauranen, A. (2000). Strange strings in translated language: A study on corpora. In M. Olohan (Ed.), *Intercultural Faultlines*: Research Models in Translation Studies 1: Textual and Cognitive Aspects (pp. 119–141). St. Jerome. https://doi.org/10.4324/9781315759951-9

https://doi.org/10.26034/cm.jostrans.2025.8473

Molés-Cases, T. (2019). Why typology matters: A corpus-based study of explicitation and implicitation of manner-of-motion in narrative texts. *Perspectives*, *27*(6), 890–907. https://doi.org/10.1080/0907676X.2019.1580754

Nida, E. A. (1964). *Toward a science of translating: With special reference to principles and procedures involved in Bible translating*. Brill.

Olohan, M., & Baker, M. (2000). Reporting that in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures*, *1*(2), 141–158. https://doi.org/10.1556/Acr.1.2000.2.1

Pedregosa, F., Varoquaux, G., Gramfort, A., Michel, V., Thirion, B., Grisel, O., Blondel, M., Prettenhofer, P., Weiss, R., Dubourg, V., Vanderplas, J., Passos, A., Cournapeau, D., Brucher, M., Perrot, M., & Duchesnay, É. (2011). Scikit-learn: Machine learning in Python. Journal of Machine Learning Research, 12, 2825–2830. http://www.jmlr.org/papers/volume12/pedregosa11a/pedregosa11a.pdf

Qin, H., & Wang, K. (2009). A parallel corpus-based study of Chinese as target language in E-C translation. *Foreign Language Teaching and Research*, *41*(2), 131–136.

Scott, M. (2004). *The WordSmith Tools* (Version 4.0). Oxford University Press.

Stubbs, M. (1986). Lexical density: A computational technique and some findings. In M. Coulthard (Ed.), *Talking about Text: Studies Presented to David Brazil* (pp. 27–42). English Language Research, University of Birmingham.

Teich, E. (2003). *Cross-linguistic variation in system and text: A methodology for the investigation of translations and comparable texts*. Mouton de Gruyter. https://doi.org/10.1515/9783110896541

Toury, G. (1978). The nature and role of norms in translation. In J. Holmes, J. Lambert, & R. van den Broeck (Eds.), *Literature and Translation: New Perspectives* (pp. 83–100). Acco. https://doi.org/10.4324/9780429280641-24

Toury, G. (1995). *Descriptive translation studies and beyond*. PA: John Benjamins. https://doi.org/10.1075/btl.100

Wang, K., & Hu, X. (2008). A parallel corpus-based study on lexical features of translated Chinese. *Chinese Translators Journal*, *29*(6), 16–21.

Wang, K., & Qin, H. (2009). A parallel corpus-based study of general features of translated Chinese. *Foreign Language Research*, *32*(1), 102–105.

Wang, W. (2019). *On the spatio-temporal differences between English and Chinese*. Foreign Language Teaching and Research Press.

Wang, Y., & Hu, R. (2021). A prompt-independent and interpretable automated essay scoring method for Chinese second language writing. In S. Li et al. (Eds.), *Chinese Computational Linguistics* (pp. 450–470). Springer. https://aclanthology.org/2021.ccl-1.107/

Wu, J. (2023). Measurement of linguistic features of academic Chinese writing by CSL learners [in Chinese]. *Applied Linguistics* (《语言文字应用》), *32*(3), 51–61. https://link.oversea.cnki.net/doi/10.16499/j.cnki.1003-5397.2023.03.005

Wu, J., Hu, R., & Lu, X. (2022). Effects of production modalities on syntactic complexity in intermediate level CSL learners. *Chinese Teaching in the World*, *36*(3), 399–415.

Wu, J., Liu, K., Hu, R., & Zhou, W. (2023). A comparative study of the syntactic complexity of translated Chinese and original Chinese [in Chinese]. *Foreign Language Teaching and Research*, *55*(2), 264–275.

https://doi.org/10.26034/cm.jostrans.2025.8473

https://link.oversea.cnki.net/doi/10.19923/j.cnki.fltr.2023.02.011

Xiao, R. (2010). How different is translated Chinese from native Chinese? A corpus-based study of translation universals. *International Journal of Corpus Linguistics*, *15*(1), 5–35. https://doi.org/10.1075/ijcl.15.1.01xia

Xiao, R., & Dai, G. (2010). In pursuit of the 'third code': A study of translation universals based on the ZCTC corpus of translational Chinese. *Foreign Language Teaching and Research*, *42*(1), 52–58.

Xiao, R., & Dai, G. (2014). Lexical and grammatical properties of translational Chinese: Translation universal hypotheses reevaluated from the Chinese perspective. *Corpus Linguistics and Linguistic Theory*, *10*(1), 11–55. https://doi.org/10.1515/cllt-2013-0016

Xiao, R., & Yue, M. (2009). Using corpora in translation studies: The state of the art. In P. Baker (Ed.), *Contemporary Corpus Linguistics* (pp. 237–262). Continuum.

## Data availability statement

All relevant data in this study are available at https://osf.io/tgcf4/files/osfstorage.

## Disclaimer

The authors are responsible for obtaining permission to use any copyrighted material contained in their article and/or verify whether they may claim fair use.

[1] CNKI (China National Knowledge Infrastructure) provides region-specific access domains. This article uses the link.oversea.cnki.net domain to ensure accessibility for international readers, as the standard link.cnki.net domain may not resolve outside mainland China. Readers in China may substitute link.oversea.cnki.net with link.cnki.net in the URL or access the content directly via https://www.cnki.net using the article's DOI.

https://doi.org/10.26034/cm.jostrans.2025.8473