

Journal of Quantitative Linguistics

ISSN: 0929-6174 (Print) 1744-5035 (Online) Journal homepage: www.tandfonline.com/journals/njql20

Quantifying Syntagmatic Patterning in Translated and Native Chinese: An R-Motif Approach Based on POS Sequences

Yiyang Hu, Kanglong Liu & Lei Lei

To cite this article: Yiyang Hu, Kanglong Liu & Lei Lei (28 Feb 2026): Quantifying Syntagmatic Patterning in Translated and Native Chinese: An R-Motif Approach Based on POS Sequences, Journal of Quantitative Linguistics, DOI: 10.1080/09296174.2026.2634477

To link to this article: https://doi.org/10.1080/09296174.2026.2634477

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.

Published online: 28 Feb 2026.

Submit your article to this journal

View related articles

View Crossmark data

Full Terms & Conditions of access and use can be found at
https://www.tandfonline.com/action/journalInformation?journalCode=njql20

JOURNAL OF QUANTITATIVE LINGUISTICS
https://doi.org/10.1080/09296174.2026.2634477

Routledge
Taylor & Francis Group

OPEN ACCESS Check for updates

Quantifying Syntagmatic Patterning in Translated and Native Chinese: An R-Motif Approach Based on POS Sequences

Yiyang Hu [a], Kanglong Liu [a] and Lei Lei [b]

[a]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China; [b]Institute of Language Sciences, Shanghai International Studies University, Shanghai, China

ABSTRACT

Research on translation universals has traditionally focused on isolated linguistic features along paradigmatic dimensions due to ease of interpretation. However, syntagmatic approaches, which examine how linguistic elements combine sequentially, remain underexplored. This corpus-based study addresses this gap by analysing R-motifs, defined as recurring sequences of part-of-speech tags, across four genres in translated and native Chinese texts. We investigate both the rank-frequency distributions of R-motif types and motif lengths as potential indicators of translation universals. Our analysis shows that R-motif frequencies in both text types follow the right-truncated Zeta distribution, whereas motif length distributions conform to the Pólya model. Random Forests are used to establish the text classification model where texts are represented by the POS R-motif distribution parameters and attributes. The experiments show that the combination of features from distribution parameters and attributes can detect the translationese efficiently. Future research may extend this approach by exploring more granular features beyond part-of-speech sequences.

ARTICLE HISTORY Received 31 May 2025; Accepted 16 February 2026

1 Introduction

Research on translation universals has traditionally focused on isolated linguistic features within translated texts (Liu, Liu, et al., 2022). Much of this work has adopted a language-in-the-mass approach, which extracts discrete linguistic features from texts without adequately considering the interrelationships among linguistic elements within the broader language system (Herdan, 1966; Kretzschmar, 2021). This methodological orientation may limit the understanding of the systemic nature of translation phenomena. In contrast, the present

CONTACT Kanglong Liu kl.liu@polyu.edu.hk

© 2026 The Author(s). Published by Informa UK Limited, trading as Taylor & Francis Group.
This is an Open Access article distributed under the terms of the Creative Commons Attribution-NonCommercial-NoDerivatives License (http://creativecommons.org/licenses/by-nc-nd/4.0/), which permits non-commercial re-use, distribution, and reproduction in any medium, provided the original work is properly cited, and is not altered, transformed, or built upon in any way. The terms on which this article has been published allow the posting of the Accepted Manuscript in a repository by the author(s) or with their consent.

study aligns with a central principle of quantitative linguistics, namely that language operates as a self-adaptive and self-regulating system. To reflect this perspective, we adopt a language-in-the-line approach that emphasizes syntagmatic patterns captured through motifs (Köhler, 2015). By focusing on sequential structures rather than isolated features, we aim to examine whether motifs can serve as effective indicators for distinguishing translated Chinese from native Chinese.

1.1 Translation Universals

Translation universals have been a central topic in corpus-based translation studies since the 1990s. The term 'universals of translational behavior' was first introduced by Toury (1980), who aimed to develop a theory capable of systematically explaining translational phenomena. Later, translation universals were more explicitly defined 'rules or facts true of the translation process regardless of the languages involved' (Newman, 1987, p. 69). With the rise of corpus approaches, the study of translation universals has evolved from intuition-based hypotheses to more systematic and empirical investigations. One of the earliest hypotheses, explicitation, was proposed by Blum-Kulka (1986) to account for the tendency of translated texts to express information more explicitly than their source texts. As corpus-based methods gained prominence in the 1990s, researchers began to test such hypotheses using large collections of translated texts. This methodological shift led to the identification of recurring features in translated language, including simplification (Baker, 1993, 1995; Laviosa, 1998, 2002), explicitation (Baker, 1995, 1996), normalization and levelling out (Baker, 1996). To further refine the conceptual framework, Chesterman (2004) introduced a distinction between S-universals and T-universals by considering the relationship between source and target texts. S-universals describe differences between translations and their source texts, whereas T-universals refer to differences between translations and comparable non-translated texts in the target language. Collectively, these studies have established a robust foundation for understanding the distinctive linguistic patterns that characterize translated texts.

Early conceptualizations of translation universals also attracted criticism for their tendency towards overgeneralization and their neglect of cultural and psychological factors. Tymoczko (1998) argued that linguistic patterns observed in translation corpora could only reflect the interests of certain translation specialists or research groups within specific contexts. Chesterman (2004) emphasized the necessity of considering language-bound, translator-bound, culture-bound and cognitive psychology criteria in the exploration of translation universals. Similarly, House (2008) questioned the existence of translation universals by highlighting the

inadequacy of purely linguistic analyses and the failure to account for factors such as directionality, genre variation and diachronic change. Becher (2010) criticized the explicitation hypothesis for lacking theoretical motivation and explicit argumentation. Drawing on cooperative principles in pragmatics, Becher proposed the asymmetry hypothesis, which suggests that explicitation phenomena are associated with cultural distance and strategies of risk management. These critical perspectives prompted researchers to refine their research designs and to incorporate insights from other disciplines in order to address the limitations identified in earlier studies.

Recent studies on translation universals have made methodological advances to address earlier critiques concerning overgeneralization. A clear trend of controlling multiple variables can be observed in corpus compilation. Translated texts, which have been described as a type of constrained language (Lanstyák & Heltai, 2012), mediated language or as a third code (Frawley, 1984), have been systematically compared with both native and non-native texts (Kruger & van Rooy, 2016a, 2016b). Other variables and their interactions, including modes of production (Kajzer-Wietrzny, 2015) and language pairs (Ferraresi et al., 2018), have also been taken into account in order to reveal more subtle linguistic patterns. These studies have contributed to a more complex understanding of translation universals, where the primary interest of researchers has shifted from the existence of universals to an exploration of the factors that may play a greater role in shaping the linguistic features of translated languages.

Methodological advances in the study of translation universals extend beyond corpus compilation to the measurement of linguistic features. Drawing on corpus linguistics, quantitative linguistics and information theory, researchers have emphasized the construction of inventories comprising multiple local features for linguistic analysis. The co-occurrence patterns of these features can be captured through multidimensional analyses, as developed in register studies (Biber, 1988), revealing various aspects of the differences between translated and native languages (Kruger & van Rooy, 2016a; Wu et al., 2023; Xu & Li, 2024). Although these approaches represent methodological progress, many studies have continued to focus on isolated features or have constructed inventories without adequately considering interactions among features beyond the predefined sets. The success of multidimensional analyses has largely depended on the careful selection of variables within the feature inventory (Kretzschmar, 2021). Despite their systematic nature, these studies often included features that were not fully representative of the broader language system. If languages, including translated languages, are regarded as complex adaptive systems, then linguistic features at various levels may exert attractive or repellent forces upon one another.

To address these limitations, recent research on translated texts has introduced holistic indicators that capture global linguistic organization. Measures such as mean dependency distance (Fan & Jiang, 2019; Ma et al., 2024) and entropy (Liu, Liu, et al., 2022) have been applied to the investigation of syntactic properties in translated texts, offering new insights into the structural complexity of translated language.

1.2 Syntagmatic Dimensions of Languages

Most studies in corpus-based linguistics, particularly in translation studies, have applied the 'bag-of-words' model or the 'bag-of-phrases/syntactic features' model. According to Herdan (1966), these studies follow a language-in-the-mass approach, where the frequency of words, phrases or syntactic structures represents certain linguistic aspects of texts. By contrast, the language-in-the-line approach tests whether the organizational properties of linguistic units in sequences conform to specific laws within the language system. This approach explicitly considers the dimension of time in the exploration of syntagmatic properties. Previous research has suggested that the syntagmatic organization of linguistic features is subject to laws identified in quantitative linguistics. In syntagmatic dimensions, both speaking and writing are inherently constrained by the temporal linearity of linguistic organization. In Herdan's (1966) derivation of grammar, he argued that pure sequences of segments cannot form conceptual structures for listeners without grammatical forms. Temporal linearity in word sequences, according to Herdan (1966), must be 'counteracted by the grammar indications' (p. 8). If grammatical forms are governed by quantitative linguistic laws, then the relations among segments in linear sequences are similarly subject to general laws.

Although few studies in classical corpus linguistics have explored syntagmatic features, recent years have seen the application of motifs in investigating the sequential properties of languages. Motifs were first introduced in the analysis of the duration of musical pieces by Boroda (1982). Drawing on Boroda's work, Köhler (2006) adapted the concept of motifs for use in quantitative linguistics. It should be noted that sequential features had already been explored through other approaches prior to Köhler's introduction of motifs, such as word length balance (Anderson, 2005) and time-series analyses (Pawłowski, 2005). In recent studies, motifs have become the predominant units for the investigation of sequential linguistic features, mainly due to their flexibility in transforming measurable observations of linguistic properties into segments of a sequence (Köhler, 2008). According to Köhler and Naumann (2016), the general definition of linguistic motifs is:

> the longest continuous sequence of equal or increasing values representing a quantitative property of a linguistic unit. (p. 1)

The concept of motifs was later extended to transform categorical data of sequential features. For example, an R-motif is defined as 'an uninterrupted sequence of unrepeated elements' (Köhler & Altmann, 2014, p. 21). Sequential elements at various linguistic levels, including syllables (Köhler & Naumann, 2008, 2010), word frequency (Sanada, 2010), part-of-speech tags (Chen, 2017; Yan, 2017), valency (Čech et al., 2017; Zhang, 2023), syntactic features (Liu & Fang, 2016; Yan, 2017) and rhetorical discourse structures (Beliankou et al., 2012; Zhang & Liu, 2016), have been abstracted into motifs in an exhaustive manner. Based on different observable measurements such as length (L-motif), frequency (F-motif), polytextuality (T-motif) and relations (R-motif), motifs can also be operationalized iteratively, for instance, through the analysis of the length motifs of F-motifs (LF-motifs).

The flexible operationalization of motifs has facilitated their application in text classification, authorship identification, typology, stylometry and pedagogy. Čech et al. (2017) examined the rank-frequency distribution of L-motifs based on the valency frames of predicative verbs in Hungarian and Czech. A more detailed exploration of motifs based on valency frames is found in Zhang's (2023) analysis of Chinese and English news texts. These studies aimed to test hypotheses about language systems and to identify typological differences. One of the most practical applications of motifs has been in text classification, employing motifs formed by word length (Köhler & Naumann, 2008, 2010) and part-of-speech tags (Chen, 2017; Yan, 2017). As these motif types were shown to conform to established distributions (e.g. Zipf-Mandelbrot distribution and Zipf-Alekseev distribution), their status as linguistic units within the language system was further confirmed.

The diversification process is a hypothesis that has been validated across various linguistic entities in quantitative linguistics. Following confirmation, such a hypothesis is regarded as a law capable of explaining and predicting linguistic phenomena (Altmann, 2005). Certain motif types have also been incorporated into the diversification process, as they have been shown to conform to specific distributions. For example, in Zhang and Liu's (2017) study of valency motifs, the right-truncated modified Zipf-Alekseev model captured the distribution of valency motifs, while the Hyperpoisson model captured the distribution of their lengths.

1.3 Syntagmatic Features in Translation Studies

Syntagmatic features have yielded valuable insights in recent translation studies. Among various holistic indicators, entropy – which accounts for both the frequency and dispersion of elements in

a sequence (Pawłowski, 2005) – has been widely applied. Lin and Liang (2023) used entropy measures for words and word categories to compare the informativeness of consecutive and simultaneous interpreting, thereby informing the selection of interpreting modes in practical settings. Liu, Liu, et al. (2022) calculated entropy values for word unigrams and part-of-speech tags to compare lexical and grammatical features of translated and native Chinese. Their results showed that translated Chinese was simpler at the lexical level, based on unigram entropy, but more complex at the syntactic level, based on POS entropy. Moving beyond textual description, Liu, Ye, et al. (2022) applied machine learning techniques to distinguish translated from native Chinese using a range of entropy-based indicators. These studies demonstrate the practical value of syntagmatic analysis in translation research.

Recent studies have linked motifs to psychological and sociological factors that may influence the linguistic features of translations. Fang (2017) used the type-token ratio of L-motifs to measure vocabulary richness in two sections of *Hongloumeng* and its two translated versions. Based on statistical results, the author proposed possible explanations for the observed differences in lexical richness. Liang et al. (2019) distinguished consecutive from simultaneous interpreting by fitting the rank-frequency distribution of F-motifs derived from word frequencies. They also calculated L-motif lengths and analysed word frequencies in the dominant length categories. These studies introduced motifs into translation research as potential linguistic units within the language system, showing that F-motifs conform to the principle of least effort.

Motifs can be constructed using elements from various linguistic levels. This study selects POS tags as the basis for motif construction, as POS has been a focus in both classical corpus-based translation studies (Laviosa, 1998; Xiao & Hu, 2015) and recent sequential analyses (Lin & Liang, 2023; Liu, Liu, et al., 2022). Xiao and Hu (2015) found that lexical density in translated Chinese was lower than in native Chinese, while Liu, Liu, et al. (2022) observed greater syntactic complexity in translated Chinese based on POS unigram entropy. Given that the proportion of function words tends to be higher in most translated Chinese texts, we hypothesize that POS R-motifs and their derived attributes exhibit different rank-frequency distributions in native and translated Chinese due to the diversification process. It remains to be seen whether distributional parameters can be used to distinguish the two text types. To our knowledge, POS motifs have rarely been applied in translation studies. However, since POS motifs have proven effective in text classification (Chen, 2017; Yan, 2017), this study aims to evaluate their usefulness in classifying translated and native Chinese. The research questions are as follows:

RQ1: Is there any difference between native and translated Chinese in the rank-frequency distributions of POS R-motifs and their lengths?

RQ2: What statistical models best capture the rank-frequency distributions of POS R-motifs and their lengths in Chinese?

RQ3: Can the parameters and distributional attributes of these models be used to distinguish translated Chinese from native Chinese?

2 Methodology

2.1 Corpus Compilation

To examine the universality of quantitative linguistic laws, we selected four genres for both translated and native Chinese, resulting in eight text groups. The source language of the translated Chinese corpus was English. The four genres were academic prose, popular prose, news and fiction. Given the presence of multiple subgenres within each category, specific subgenres were selected to ensure comparability between translated and native Chinese. The fiction group included science fiction, romantic fiction and mystery and detective fiction. The news group was limited to press reportage. The popular prose group consisted of biographies and popular lore, while academic prose was drawn from research articles in CNKI and translated academic texts. All texts were published between 2000 and 2023.

The total number of characters and samples was adjusted through splitting and reorganization to improve comparability and maintain statistical operability. Balanced sample sizes were not maintained, in accordance with the scale-free nature of language as a complex system (Kretzschmar, 2015). After manual noise removal, the tokenization and POS tagging were performed by HanLP (He & Choi, 2021), and punctuations were removed following annotation. As for tokenization, we chose the fine-grained mode in HanLP which followed the format of Penn Chinese Treebank (Xia, 2000). Basic information about the eight groups after tokenization and POS annotation is presented in Table 1.

2.2 Motif Calculation

Since part-of-speech (POS) annotation schemes in Chinese are typically based on word units, the text samples listed in Table 1 were first segmented and annotated using HanLP. The Penn Chinese Treebank (CTB) scheme was adopted for POS annotation, as it provides more fine-grained tags that

Table 1. Corpus composition and descriptive statistics.

Group	Status	Genre	Num. of samples	Total word frequency	Average word frequency	Range of word frequency
n_academic	Native	Academic prose	50	63,201	1,264	939–2,784
n_fiction	Native	Fictions	50	63,153	1,263	647–2,552
n_news	Native	News	50	63,273	1,265	211–3,727
n_prose	Native	Popular prose	50	63,202	1,264	627–2,315
t_academic	Translated	Academic prose	50	63,138	1,263	762–2,475
t_fiction	Translated	Fictions	50	63,079	1,262	728–2,675
t_news	Translated	News	50	63,138	1,263	453–3,690
t_prose	Translated	Popular prose	50	63,229	1,265	862–1,858

facilitate the formation of more detailed R-motif types. POS tagging was performed using the pretrained Electra base model (Clark et al., 2020). R-motifs and their lengths were extracted from each sentence in each sample using Python scripts (see the Open Science Framework: https://osf.io/serqj/). Table 2 presents an example illustrating the extraction of POS R-motifs and motif lengths.

R-motifs are extracted from uninterrupted sequences of unrepeated POS tags and their lengths are calculated based on the number of POS tags in each motif. A new R-motif is formed whenever a POS tag reoccurs within the same sequence. As shown in Table 2, the second R-motif (AD + NN) is formed because its initial POS, AD, appears in the first R-motif. Similarly, the fourth R-motif is created when NN, which appears in the third R-motif, recurs. To identify the POS tags that trigger the segmentation of R-motifs, the first POS of each non-initial R-motif is also extracted for each sentence. As the formation of new R-motifs depends on existent R-motifs, we posit that the R-motif length x also depends on the existent length $x-1$, which can be mathematically formulated as $P_x = g(x)P_{x-1}$. It can be hypothesized that both R-motifs and their lengths result from the diversification process (Altmann, 2005, p. 649).

2.3 Statistical Analyses

After extracting two types of motifs, we first examined the overall rank-frequency distributions across groups. To observe the difference between rank-frequency distributions more explicitly, the Jensen-Shannon divergence (Lin, 1991) is calculated for each genre across the translated and native groups. The formula for the Jensen-Shannon divergence is defined as

$$JSD(P \parallel Q) = \frac{1}{2}D(P \parallel M) + \frac{1}{2}D(Q \parallel M)$$

Table 2. Extraction of POS R-motifs and motif lengths from a sample sentence.

Word	还是	真的	只有	地球	文明	是	宇宙	中	唯一	的	绿洲
Pinyin	(háishì)	(zhēnde)	(zhǐ yǒu)	(dìqiú)	(wénmíng)	(shì)	(yǔzhòu)	(zhōng)	(wéiyī)	(de)	(lǜzhōu)
English gloss	or	truly	only	Earth	civilization	is	cosmos	in	only	DE	oasis
Literal translation	(Is Earth truly the only oasis of civilization in the cosmos)										
Id	1	2	3	4	5	6	7	8	9	10	11
POS	CC	AD	AD	NN	NN	VC	NN	LC	JJ	DEG	NN
R-motif	CC + AD		AD + NN		NN + VC		NN + LC + JJ + DEG				NN
Length of R-motifs	2		2		2		4				1
1st POS of non-initial R-motifs			AD		NN		NN				NN

$$D(P \parallel M) = \sum_{i}^{n} P_i \times \log_2 \frac{P_i}{M_i}$$

where P and Q are two distributions, and $M = \frac{1}{2}(P + Q)$ is a mixture distribution of P and Q. An additive smoothing is applied by adding a value of $\epsilon = 1 \times 10^{-10}$ to each probability in P and Q.

Given the fractal nature of rank-frequency distributions in language systems, we inferred that the synergetic variation of these distributions would also be reflected in textual samples. Altmann-Fitter (1997) was used to fit the rank-frequency distributions of motif types in each sample to appropriate models. On the one hand, we sought to evaluate whether these sequential features could serve as linguistic units within the language system by examining goodness-of-fit measures. On the other hand, the parameters and attributes derived from the fitted models were used as features in the classification algorithm to assess their effectiveness in classifying translated texts from native texts. It should be noted that multiple models can fit motif distributions well, including the Zipf-Mandelbrot distribution (Mandelbrot, 1965) and the Zipf-Alekseev model (Hřebíček, 1996).

In this study, the rank-frequency distributions of R-motifs and motif lengths conform predominantly to the right-truncated Zeta and Pólya models, respectively. These two models are chosen for the following reasons. Among the 400 samples in the corpus, motif data of 391 samples fit the two models well, compared with the sample size that fits other models. Considering multiple variables involved in samples, the loss of a few samples, due to the poor coefficients of determination, is acceptable. Table 3 presents the mean values of R^2 fitted by the two distributions for each text groups.

The rank-frequency distribution of R-motifs is best described by the right-truncated Zeta distribution, with R^2 values ranging from 0.853 to 0.983. The distribution of R-motif lengths conforms closely to the Pólya model, with R^2 values ranging from 0.947 to 0.999. The right-truncated Zeta distribution is typical for modelling word frequency distributions (Popescu et al., 2009), while the Pólya model is a quadratic extension of the mechanism of the word

Table 3. Mean values of R^2 of 391 samples of POS R-motifs and their lengths fitted by two distributions.

Status	Genre	Sample size	Right-truncated Zeta	Pólya
Native	Academic prose	49	0.931	0.997
Native	Fictions	47	0.918	0.992
Native	News	49	0.939	0.997
Native	Popular prose	49	0.926	0.993
Translated	Academic prose	49	0.908	0.992
Translated	Fictions	50	0.922	0.993
Translated	News	50	0.926	0.994
Translated	Popular prose	48	0.900	0.990

Table 4. Features used in the random forest algorithm for text classification.

Index	Features	Explanation
A	a	Parameter a of the right-truncated Zeta distribution of R-motifs
B	entropy [R-motifs]	Entropy value of the frequency distribution of R-motifs
C	hapax [R-motifs]	Proportion of hapax legomena of R-motifs
D	s	Parameter s of the Pólya distribution of R-motif lengths
E	P	Parameter P of the Pólya distribution of R-motif lengths
F	entropy [R-motif length]	Entropy value of the frequency distribution of R-motif lengths
G	hapax [R-motif length]	Proportion of hapax legomena of R-motif lengths
H	genre	Genre information from the corpus metadata

length distribution (Wimmer & Altmann, 1996). As shown in Motif Calculation, the extraction of R-motifs and their lengths is based on POS tags assigned to word units. Consequently, their rank-frequency distributions exhibit patterns similar to those of word frequencies and word-length frequencies. The formula for the right-truncated Zeta distribution (Wimmer & Altmann, 1999) is defined as follows

$$P_x = \frac{x^{-\alpha}}{F(R)}, x = 1, 2, 3, \ldots, R$$

Here $\alpha \epsilon \mathbb{R}, \mathrm{R} \epsilon \mathbb{N}$, $F(R) = \sum_{i=1}^{R} i^{-\alpha}$.

The Pólya model is expressed as

$$P_x = \frac{\binom{-P/s}{x}\binom{-Q/s}{n-x}}{\binom{-1/s}{n}}, x = 0, 1, 2, \ldots n$$

Here $s > 0, n \in \mathbb{N}_0, 0 \ll P \ll 1, Q = 1 - P$.

The parameters and distributional attributes, together with the genre variable, are represented as vectors for text classification in Table 4, following Köhler and Naumann's (2010) design. A tree ensemble method based on random forests (Breiman, 2001) is adopted to assess the applicability of R-motifs in text classification. Compared with individual tree-based methods, random forests reduce the likelihood of overfitting by introducing two layers of randomness. One layer concerns the selection of data points used to fit each tree, and the other concerns the selection of predictors at each split. Since random forests can accommodate both numeric and categorical predictors, the genre variable, which may increase classification difficulty, can be included when training the classifier. To capture the data structure among predictors more parsimoniously, interactions among predictors are examined using the R package *randomForestSRC* (Ishwaran & Kogalur, 2024). The minimum node size is set to 3 to prevent overfitting. All other parameters follow the default settings in the *randomForestSRC* package.

The feature selection uses a backward elimination strategy that begins with all features included in the baseline model, because the manageable number of features allows potential interactions to be considered. After comparing the variable importance scores and the interactions among features in the baseline model, the model is retrained with the least important feature removed. The retraining process stops when the removal of a feature leads to a substantial drop in performance.

3 Results

This section is divided into two parts. Section 3.1 examines the distribution of POS R-motifs and R-motif lengths at the corpus level. Section 3.2 tests the application of sample fitting results in detecting translationese by building a classifier based on random forests.

3.1 POS R-Motifs and Lengths in Translated Chinese and Native Chinese

The overall frequencies of POS R-motifs on a logarithmic scale are presented in Figure 1. The rank-frequency distributions of the top 100 R-motifs in Figure 2 show that in academic prose, news and popular prose, the frequencies of the top three R-motifs in native Chinese are higher than those in translated Chinese. For fiction, indicated by green lines, no substantial differences are observed among the top ranks. The JSD for each genre in

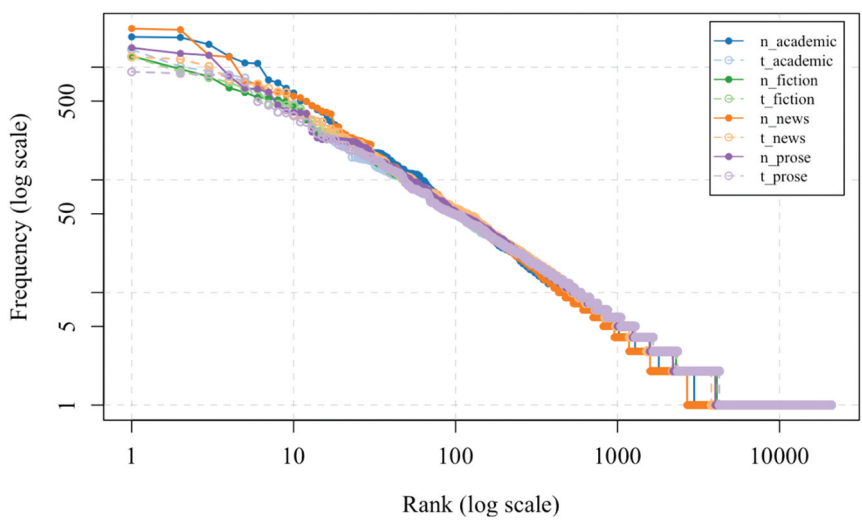

Figure 1. Rank-frequency distributions of POS R-motifs across genres in translated and native Chinese texts.

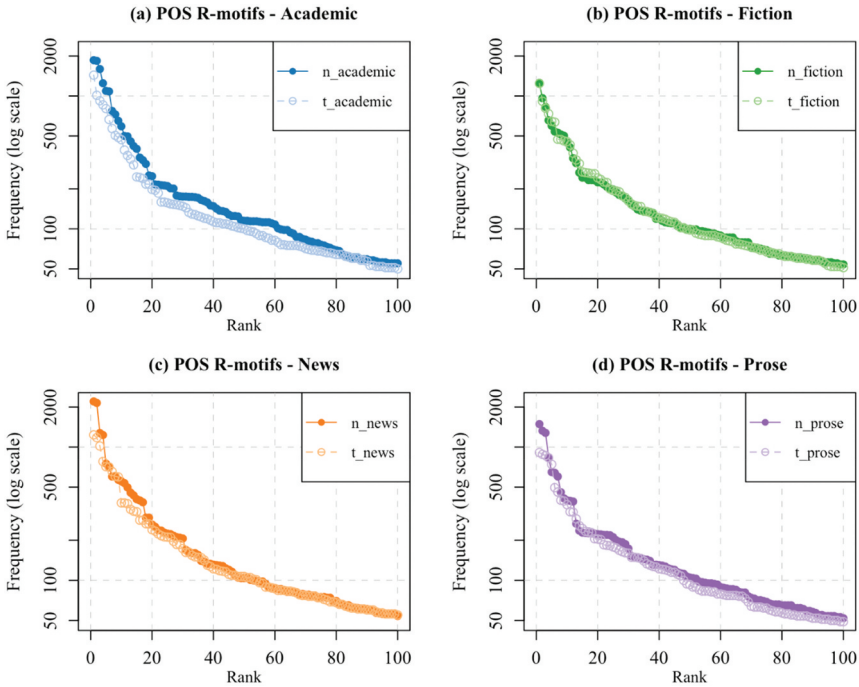

Figure 2. Rank-frequency distributions of the top 100 POS R-motifs across genres in translated and native Chinese texts.

Table 5. Jensen-Shannon divergence between overall rank-frequency distributions of R-motifs and their lengths in native Chinese and translated Chinese.

Genre	JSD for R-motif distributions	JSD for R-motif length distributions
Academic prose	0.0633	0.0186
Fictions	0.0049	0.0001
News	0.0468	0.0118
Popular prose	0.0252	0.0042

Table 5 provides evidence for the previous observation that the rank-frequency distributions of R-motifs in translated fictions and native fictions are similar to each other, while greater difference exists for other genres. To save space, a more detailed table listing the 10 most frequent R-motifs is provided in the Appendix. In general, common nouns (NN) and common verbs (VV) function as the core elements of high-frequency R-motifs. Compared with other genres, academic prose contains a notably larger number of R-motifs that involve the genitive or associative marker *de* (DEG).

According to the definition of POS R-motifs in this study, the first POS tag in each non-initial R-motif serves as the trigger for sequence truncation. To examine the overall role of different POS tags in motif segmentation, the

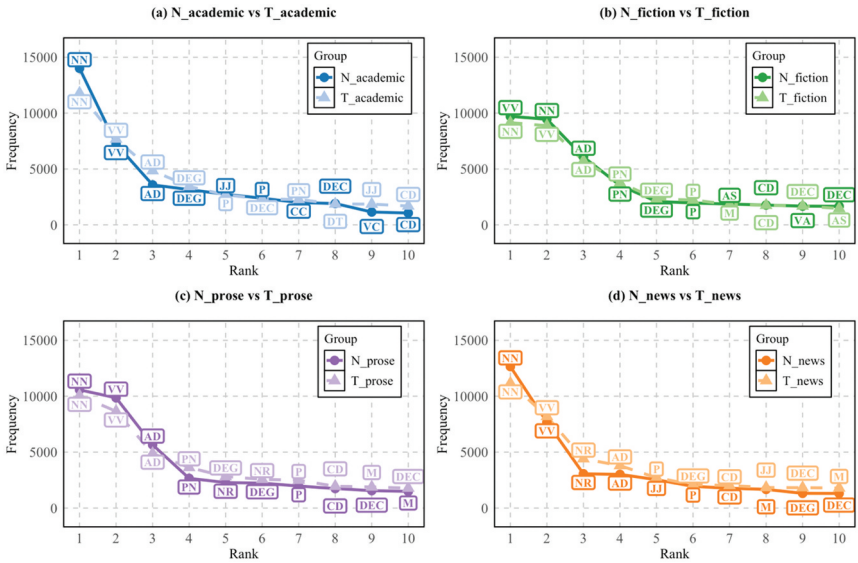

Figure 3. Rank-frequency distributions of the top 10 POS types that truncate motifs.

distribution of these elements is shown in Figure 3. Among the 10 most frequent POS types that truncate motifs, most are content words, including common nouns (NN), common verbs (VV) and adverbs (AD). The frequencies of VV and AD in translated Chinese vary across genres. They are higher than in native Chinese for academic prose and news, but lower for popular prose and fiction. This fluctuation across genres contributes to the changing positions of these tags in the rank frequency distributions. Function words such as the genitive or associative marker de (DEG), pronouns (PN) and prepositions (P) do not show substantial variation between translated and native Chinese across genres. Overall, content words are the primary contributors to R-motif formation and to the observed differences in rank frequency distributions between translated and native Chinese texts.

After extracting the lengths of POS R-motifs, we rank their frequencies across all genre groups, as shown in Figure 4. In all genres except fiction, the most frequent R-motif length (ranked first) occurs less frequently in translated texts than in native texts. Conversely, for R-motif lengths ranked between third and eighth, translated texts tend to exhibit higher frequencies than their native counterparts. The JSD values for R-motif length distributions in each genre also support the observation (see Table 5). The rank corresponds to motif length plus one. The high frequency of short R-motifs reflects the difficulty of sustaining sequences of unique POS tags within the language system, which aligns with the dominance of shorter R-motifs. Overall, R-motifs in translated Chinese tend to be longer than those in native

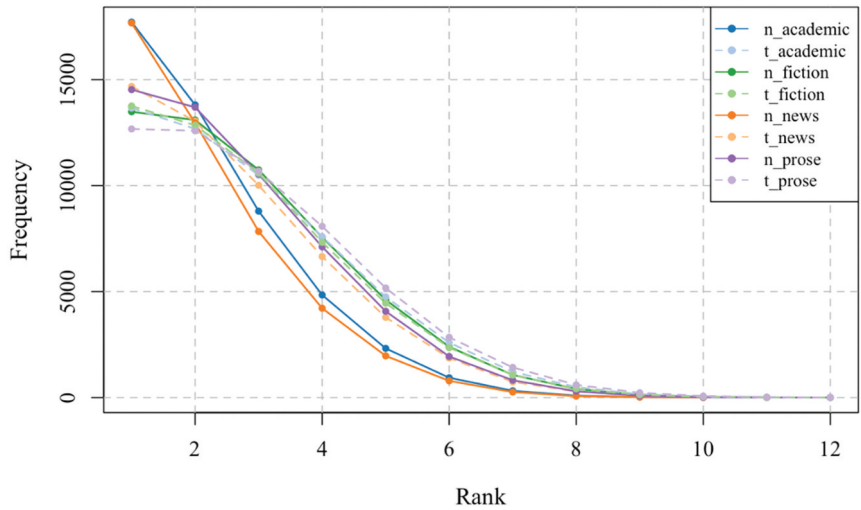

Figure 4. Rank frequency distributions of the length of POS R-motifs.

Chinese, with the exception of fiction, where no substantial difference is observed.

3.2 Application of Fitting Results in Detecting Translated Texts

Section 3.1 examined the overall distributional patterns of R-motifs and their lengths in translated and native Chinese. To minimize sample-level bias, the rank-frequency distributions of R-motifs and motif lengths are fitted to statistical models using Altmann-Fitter for each individual sample. As shown in Table 3, POS R-motifs and their lengths in 391 samples (above 97%) can fit the right-truncated Zeta distribution and the Pólya distribution well, respectively. These results support the treatment of POS R-motifs as linguistic entities within the language system. The following section assesses the applicability of R-motifs in detecting translationese by establishing the random forests classifier based on the features in Table 4.

The baseline model is established by including all features in Table 4 in the random forest algorithm, resulting in an error rate of 25.58%. For the baseline model, we calculate the estimated variable importance of each feature by using the subsample function in the package *randomForestSRC*. The subsample function uses the default settings, in which the number of subsamples is 100 and the size of each subsample is equal to the inverse square root of the sample size. Figure 5(a) summarizes the permutation importance of the predictors, where a higher value indicates a stronger contribution to the prediction. Among the eight features in Table 4, the proportion of hapax legomena of R-motifs (Feature C) provides the greatest contribution to the

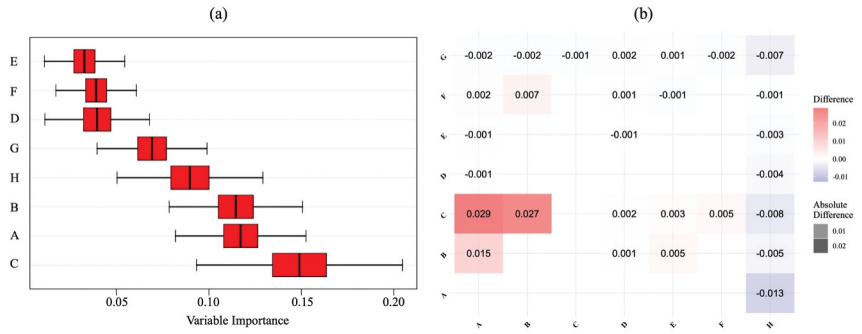

Figure 5. (a) Variable importance with subsampling inference for features in the baseline model; (b) Heatmap for the difference between paired and additive variable importance.

prediction. This is followed by the parameter α in the right truncated zeta model (Feature A) and the entropy value of the rank frequency distribution of R-motifs (Feature B). The genre feature (Feature H) shows a moderate level of predictive importance, while the features associated with R-motif lengths make only a modest contribution to the prediction.

Given that the variable importance score does not explicitly show how one predictor interacts with others (Gries, 2020), we examine the interactions among the eight predictors in the baseline model, as shown in Figure 5(b). Pairwise interactions are indicated by the large absolute values of the differences between the additive and the paired variable importance scores for two predictors (Ishwaran, 2007). Positive differences are observed for features related to the rank frequency distributions of POS R-motifs, including the parameter α in the zeta model (Feature A), the entropy of the rank frequency distribution of R-motifs (Feature B) and the proportion of hapax legomena of R-motifs (Feature C). These three features are therefore preferred for consideration in the next stage of feature selection. For large negative differences, the genre predictor reduces the additive predictive power when combined with other features. However, the negative differences are relatively small compared with the overall variable importance of the genre feature itself. Finally, features with lower variable importance scores (Features D, E, F) do not show strong interactions with other features, and they will therefore be considered in the backward elimination stage of feature selection.

The backward elimination of features in Table 6 shows that Model 3 achieves the lowest out-of-bag identification error rate, which is 23.27%, in identifying text status. Compared with the baseline model, Model 3 removes two features related to the rank frequency distributions of R-motif lengths. These are Feature E, the parameter *P* from the Pólya model and Feature F,

Table 6. Classification results using different feature sets.

Model	Feature sets	Identification error rate (%)
1	A + B + C + D + E + F + G + H (baseline model)	25.58%
2	A + B + C + D + F + G + H	23.79%
3	A + B + C + D + G + H	23.27%
4	A + B + C + G + H	23.79%
5	A + B + C + D + H	29.67%
6	A + B + C + D + G	26.60%

Table 7. Confusion matrix for the classification in model 3.

	Predicted as Native	Predicted as Translated	Recall (%)
Observed as Native	139	55	71.65%
Observed as Translated	36	161	81.73%
Precision (%)	79.43%	74.54%	

the entropy of R-motif length distributions. Interestingly, removing Feature G, the proportion of hapax legomena of R-motif lengths, increases the identification error rate from 23.27% in Model 3 to 29.67% in Model 5. Table 7 presents the confusion matrix for Model 3, which is constructed from the predicted probabilities using Bayes' rule. Model 3 demonstrates strong performance in detecting translationese, with a macro F1 score of 76.65%.

Figure 6 shows how each predictor in Model 3 influences the prediction of text status. Each panel reports the mean predicted probability of the levels of text status on the y axis. The three panels in the first row suggest that translated texts are much more likely if the proportion of POS R-motif

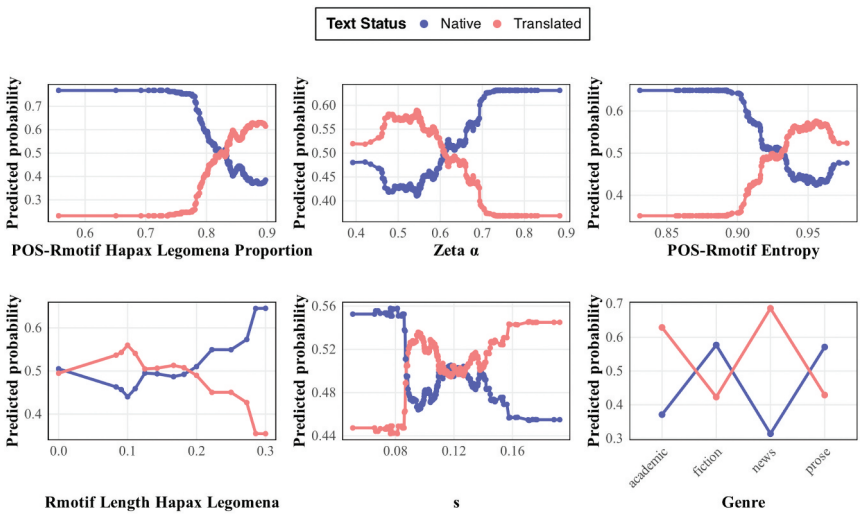

Figure 6. Partial dependence plots for the predictors in the refined random forests model.

hapax legomena is higher than 0.83, if the parameter Zeta α is lower than 0.6 or if the entropy of the POS R-motif distribution is higher than 0.93. For the R-motif length hapax legomena and the parameter *s* from the Pólya model, the prediction trends are more complex, with high predicted probabilities appearing only in certain intervals. Regarding genres, it is noteworthy that if a genre tends to exhibit idiosyncratic linguistic features, it is more likely to be predicted as native text. If it tends to follow composition conventions such as those found in academic prose and news, it is more likely to be predicted as translated text.

4 Discussion

4.1 POS R-Motifs in Translated Chinese

The JSD values in POS R-Motifs and Lengths in Translated Chinese and Native Chinese quantify the divergence in the overall rank frequency distributions of native and translated Chinese. This divergence at the corpus level is not obvious in fiction, but it is present in academic prose, news and popular prose. The overall divergence guides the further investigation of rank frequency distributions at the sample level. In Application of Fitting Results in Detecting Translated Texts, the rank frequency distributions of R-motifs in more than 97% of the samples follow the right truncated Zeta model well, which supports the interpretation of R-motifs as linguistic units in the language system. The probability of R-motif lengths decreases as the length increases across all groups, which resembles the distribution of word lengths.

Since there are mutual influences among these motifs and among their lengths, we regard the distributions of R-motifs and their lengths as outcomes of diversification processes (Altmann, 2005). The higher value of parameter α in the first model suggests a greater concentration of probability mass in the distribution, with short R-motifs formed mainly by common nouns and common verbs. The higher values of parameters *s* and *P* in the Pólya model indicate that a greater distance between repeated POS categories tends to occur along the sequence. The partial plots in Application of Fitting Results in Detecting Translated Texts highlight the contribution of lower α values and higher *s* values in predicting translated Chinese texts.

The longer POS R-motifs observed in translated Chinese may result from a greater variety of POS categories in the motif. Since most R-motifs are truncated by content words, longer motifs suggest an increased presence of function words within the sequences. Previous studies have confirmed the higher proportion of function words in translated Chinese from various perspectives. Xiao and Hu (2015) compared the lexical density of native and translated Chinese by calculating the ratio of content words to total

words, finding that all 15 genres exhibited higher lexical density in native Chinese. Similarly, Liu, Liu, et al. (2022) identified syntactic complexification in translated Chinese by calculating the entropy of POS unigrams.

In Figure 3, the longer R-motifs observed in translated Chinese suggest that function words weaken the truncating effect typically exerted by content words. This pattern reflects the bilingual processing constraints inherent in translation activities (Chen et al., 2024; Kruger & van Rooy, 2016a), which differ fundamentally from the production conditions of spontaneous native language use. Due to the high frequency and high dispersion of function words in the limited language exposure of translators, their activation remains elevated during translation, where short-term bilingual priming effects predominate. This sustained activation disrupts the usual truncation patterns mainly established by content words, thereby contributing to the formation of longer R-motifs in translated Chinese.

4.2 The Usage of POS in Detecting Translationese

Drawing on random forests models in Application of Fitting Results in Detecting Translated Texts, this study also evaluates the application of sequential features of POS in detecting translationese. Previous studies on translationese have proved the feasibility of POS features in classifying translated texts from native texts, mainly from the language-in-the-mass approach. For instance, Baroni and Bernardin (2006) compared the classifiers of Support Vector Machines (SVM) built upon different combinations of feature sets, where POS features were combined with wordforms, lemma in the formation of n-grams. They found that the distributions of function words and shallow syntactic patterns, without lexical features, could help classifiers detect the translated texts efficiently. Similar results of POS n-grams were also revealed in the usage of SVM in classifying native English from the English translated from 10 languages (Volansky et al., 2015). Hu et al. (2018) tested the SVM classifiers based on POS and its combination with other syntactic features for translated and native Chinese. These SVMs using syntactic features achieved roughly the same high accuracy with those using lexical n-gram features. Our study confirmed the potential usage of POS in text classification from the language-in-the-line approach. The dynamic changes in rank-frequency distribution of POS R-motifs and their length, reflected by the model parameters and attributes, can serve as holistic predictors with relatively excellent performance for detecting translationese.

The feature selection for the baseline model reveals strong pairwise interactions among the features of POS R-motif distributions in predicting text status. The predictive power of the combined POS R-motif features is greater than the sum of the individual features. Their collective value is also reflected

in the permutation importance scores, in which they rank among the top three. Compared with the features related to R-motif lengths, the features related to R-motifs contribute more to the classification task. Not all features related to R-motif lengths are eliminated when refining the model, such as the proportion of hapax legomena of R-motif lengths. The features included in the final model suggest that second-order motifs also have the potential to enhance prediction performance.

5 Conclusion

This study confirmed the status of POS R-motifs and their lengths as syntagmatic linguistic units within the language system. The rank-frequency distributions of R-motifs in the comparable Chinese corpus mainly followed the expected patterns of the diversification process. Specifically, the rank-frequency distributions of POS R-motifs and R-motif lengths fitted the right-truncated Zeta distribution and the Pólya distribution, respectively. On the one hand, the descriptive attributes of rank-frequency patterns and the associated model parameters provided a systematic explanation for the difference in POS R-motif formation between native texts and translated texts. On the other hand, the inventory of R-motif features demonstrated considerable potential for application in detecting translationese.

Nonetheless, this study has certain limitations. The corpus could be expanded to include a broader range of genres and languages to further test the universality of applying mathematical model parameters to translation status classification. Moreover, the effects of source texts and individual translators were not considered. Future research could explore motifs extracted from parallel and comparable corpora to investigate finer-grained variation within the diversification process during translation. Attributes based on rank-frequency distributions of motifs could also be integrated into other classification approaches to enhance the automatic identification of translated texts.

Disclosure Statement

No potential conflict of interest was reported by the author(s).

Funding

The work was supported by the Hong Kong Polytechnic University [P0051009]; Hong Kong Polytechnic University [P0046370].

ORCID

Yiyang Hu http://orcid.org/0009-0001-4101-8667
Kanglong Liu http://orcid.org/0000-0003-3962-2563
Lei Lei http://orcid.org/0000-0002-3366-1855

References

Altmann, G. (2005). Diversification processes. In R. Köhler, G. Altmann, & G. Piotrowski (Eds.), *Quantitative linguistik: Ein internationales handbuch/quantitative linguistics: An international handbook* (pp. 646–658). De Gruyter Mouton.

Altmann-Fitter. (1997). *Iterative fitting of probability distribution*. RAM.

Anderson, S. (2005). Word length balance in texts: Proportion constancy and word-chain-lengths in Proust's longest sentence. *Glottometrics*, *11*, 32–50.

Baker, M. (1993). Corpus linguistics and translation studies: Implications and applications. In M. Baker, G. Francis, & E. Tognini-Bonelli (Eds.), *Text and technology: In honour of John Sinclair* (pp. 233–250). John Benjamins Publishing Company.

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

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 honour of Juan C. Sager* (pp. 175–186). John Benjamins Publishing Company.

Baroni, M., & Bernardini, S. (2006). A new approach to the study of translationese: Machine-learning the difference between original and translated text. *Literary and Linguistic Computing*, *21*(3), 259–274. https://doi.org/10.1093/llc/fqi039

Becher, V. (2010). Abandoning the notion of "translation-inherent" explicitation: Against a dogma of translation studies. *Across Languages and Cultures*, *11*(1), 1–28. https://doi.org/10.1556/Acr.11.2010.1.1

Beliankou, A., Köhler, R., & Naumann, S. (2012). Quantitative properties of argumentation motifs. In I. Obradović, E. Kelih, & R. Köhler (Eds.), *Methods and Applications of Quantitative Linguistics, selected papers of the 8th International Conference on Quantitative Linguistics (QUALICO)* (pp. 35–43). Belgrade, Serbia. April 26–29, 2012

Biber, D. (1988). *Variation across speech and writing*. Cambridge University Press.

Blum-Kulka, S. (1986). Shift of cohesion and coherence in translation. In L. Venuti (Ed.), *The translation studies reader* (pp. 298–313). Routledge.

Boroda, M. G. (1982). Die melodische Elementareinheit [the melodic elementary unit]. In J. K. Orlov, M. G. Boroda, & I. I. Š. Nadarejšvili (Eds.), *Sprache, text, kunst: Quantitative analysen* [language, text, art: Quantitative analyses] (pp. 205–222). Brockmeyer.

Breiman, L. (2001). Random forests. *Machine Learning*, *45*(1), 5–32. https://doi.org/10.1023/A:1010933404324

Čech, R., Vincze, V., & Altmann, G. (2017). On motifs and verb valency. In H. Liu & J. Liang (Eds.), *Motifs in language and text* (pp. 13–36). De Gruyter Mouton.

Chen, J., Li, D., & Liu, K. (2024). Unraveling cognitive constraints in constrained languages: A comparative study of syntactic complexity in translated, EFL, and native varieties. *Language Sciences*, *102*, 1–14. https://doi.org/10.1016/j.langsci.2024.101612

Chen, R. (2017). Quantitative text classification based on POS-motifs. In H. Liu & J. Liang (Eds.), *Motifs in language and text* (pp. 65–86). De Gruyter Mouton.

Chesterman, A. (2004). Beyond the particular. In A. Mauranen & P. Kujamäki (Eds.), *Translation universals: Do they exist* (pp. 33–50). John Benjamins Publishing Company.

Clark, K., Luong, M. T., Le Q. V., & Manning, C. D. (2020). *Electra: Pre-training text encoders as discriminators rather than generators*. In ICLR. https://openreview.net/pdf?id=r1xMH1BtvB.

Fan, L., & Jiang, Y. (2019). Can dependency distance and direction be used to differentiate translational language from native language? *Lingua*, *224*, 51–59. https://doi.org/10.1016/j.lingua.2019.03.004

Fang, Y. (2017). L-motif TTR for authorship identification in Hongloumeng and its translation. In H. Liu & J. Liang (Eds.), *Motifs in language and text* (pp. 181–200). De Gruyter Mouton.

Ferraresi, A., Bernardini, S., Petrović, M. M., & Lefer, M. A. (2018). Simplified or not simplified? The different guises of mediated English at the European Parliament. *Meta*, *63*(3), 717–738. https://doi.org/10.7202/1060170ar

Frawley, W. (1984). Prolegomenon to a theory of translation. In W. Frawley (Ed.), *Translation: Literary, linguistic, and philosophical perspectives* (pp. 159–175). Associated University Presses.

Gries, S. T. (2020). On classification trees and random forests in corpus linguistics: Some words of caution and suggestions for improvement. *Corpus Linguistics and Linguistic Theory*, *16*(3), 617–647. https://doi.org/10.1515/cllt-2018-0078

He, H., & Choi, J. D. (2021). The stem cell hypothesis: Dilemma behind multi-task learning with transformer encoders. In Moens M. F. Huang X. Specia L. Yih S. W. (eds.), *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing* (pp. 5555–5577). Association for Computational Linguistics.

Herdan, G. (1966). *The advanced theory of language as choice and chance*. Springer.

House, J. (2008). Beyond intervention: Universals in translation. *Trans-kom*, *1*(1), 6–19.

Hřebíček, L. (1996). Word associations and text. In Schmidt P. (ed.), *Glottometrika 15: Issues in general linguistic theory and the theory of word length* (pp. 12–17). Wissenschaftlicher Verlag Trier.

Hu, H., Li, W., & Kübler, S. (2018). Detecting syntactic features of translated Chinese. In J. Brooke, L. Flekova, M. Koppel, & T. Solorio (Eds.), *Proceedings of the Second Workshop on Stylistic Variation* (pp. 20–28). Association for Computational Linguistics. https://doi.org/10.18653/v1/W18-1603

Ishwaran, H. (2007). Variable importance in binary regression trees and forests. *Electronic Journal of Statistics*, *1*, 519–537. https://doi.org/10.1214/07-EJS039

Ishwaran, H., & Kogalur, U. B. (2024). Fast unified random forests for survival, regression, and classification (RF-SRC), R package version 3.3.1.

Kajzer-Wietrzny, M. (2015). Simplification in interpreting and translation. *Across Languages and Cultures*, *16*(2), 233–255. https://doi.org/10.1556/084.2015.16.2.5

Köhler, R. (2006). The frequency distribution of the lengths of length sequence. In J. Genzor & M. Bucková (Eds.), *Favete linguis: Studies in honour of Victor Krupa* (pp. 145–152). Slovak Academic Press.

Köhler, R. (2008). Sequences of linguistic quantities: Report on a new unit of investigation. *Glottotheory*, *1*(1), 115–119. https://doi.org/10.1515/glot-2008-0018

Köhler, R. (2015). Linguistic motifs. In M. Georgios & J. Mačutek (Eds.), *Sequences in language and text* (pp. 89–108). De Gruyter Mouton.

Köhler, R., & Altmann, G. (2014). *Problems in quantitative linguistics 4*. RAM–Verlag.

Köhler, R., & Naumann, S. (2008). Quantitative text analysis using L-, F- and T-segments. In C. Preisach, H. Burkhardt, L. Schmidt-Thieme, & R. Decker (Eds.), *Data analysis, machine learning and applications* (pp. 637–646). Springer.

Köhler, R., & Naumann, S. (2010). A syntagmatic approach to automatic text classification: Statistical properties of F- and L-motifs as text characteristics. In P. Grzybek, E. Kelih, & J. Mačutek (Eds.), *Text and language: Structures, functions, interrelations, quantitative perspectives* (pp. 81–90). Praesens.

Köhler, R., & Naumann, S. (2016). Syntactic text characterization using linguistic S-motifs. *Glottometrics*, *34*, 1–8.

Kretzschmar, W. A. (2015). *Language and complex systems*. Cambridge University Press.

Kretzschmar, W. A. (2021). Complex systems for corpus linguistics. *ICAME Journal*, *45*(1), 155–177. https://doi.org/10.2478/icame-2021-0005

Kruger, H., & van Rooy, B. (2016a). Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide*, *37*(1), 26–57. https://doi.org/10.1075/eww.37.1.02kru

Kruger, H., & van Rooy, B. (2016b). Syntactic and pragmatic transfer effects in reported-speech constructions in three contact varieties of English influenced by Afrikaans. *Language Sciences*, *56*, 118–131. https://doi.org/10.1016/j.langsci.2016.04.003

Lanstyák, I., & Heltai, P. (2012). Universals in language contact and translation. *Across Languages and Cultures*, *13*(1), 99–121. https://doi.org/10.1556/Acr.13.2012.1.6

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

Laviosa, S. (2002). *Corpus-based translation studies: Theory, findings, applications*. Rodopi.

Liang, J., Lv, Q., & Liu, Y. (2019). Quantifying interpreting types: Language sequence mirrors cognitive load minimization in interpreting tasks. *Frontiers in Psychology*, *10*, 285. https://doi.org/10.3389/fpsyg.2019.00285

Lin, J. (1991). Divergence measures based on the Shannon entropy. *IEEE Transactions on Information Theory*, *37*(1), 145–151. https://doi.org/10.1109/18.61115

Lin, Y., & Liang, J. (2023). Informativeness across interpreting types: Implications for language shifts under cognitive loads. *Entropy*, *25*(2), Article 243. https://doi.org/10.3390/e25020243

Liu, H., & Fang, Y. (2016). Quantitative aspects of hierarchical motifs. In E. Kelih, R. Knight, J. Mačutek, & A. Wilson (Eds.), *Issues in quantitative linguistics 4* (pp. (pp. 9–26). RAM–Verlag.

Liu, K., Liu, Z., & Lei, L. (2022). Simplification in translated Chinese: An entropy-based approach. *Lingua*, *275*, Article 103364. https://doi.org/10.1016/j.lingua.2022.103364

Liu, K., Ye, R., Liu, Z., & Ye, R. (2022). Entropy-based discrimination between translated Chinese and original Chinese using data mining techniques. *PLOS ONE*, *17*(3), Article e0265633. https://doi.org/10.1371/journal.pone.0265633

Ma, R., Jiang, Y., & Qian, J. (2024). Syntactic and typological properties of constrained language: A study based on dependency treebanks of translated English and non-native English. *Across Languages and Cultures*, *25*(1), 91–111. https://doi.org/10.1556/084.2023.00363

Mandelbrot, B. (1965). Information theory and psycholinguistics. In B. B. Wolman Nagel E. (Eds.), *Scientific psychology* (pp. 550–562). Basic Books.
Newman, A. (1987). Translation universals - perspectives and explorations. In M. G. Rose (Ed.), *Translation perspectives III: Selected papers, 1985-86* (pp. 69–83). Binghampton.
Pawłowski, A. (2005). Modelling of sequential structures in text. In R. Köhler, G. Altmann, & R. G. Piotrowski (Eds.), *Quantiative linguistik: Ein internationales handbuch/quantitative linguistics: An international handbook* (pp. 738–750). De Gruyter Mouton.
Popescu, I., Altmann, G., Grzybek, P., Jayaram, B. D., Köhler, R., Krupa, V., Mačutek, J., Pustet, R., Uhlířová, L., & Vidya, M. N. (2009). *Word frequency studies*. De Gruyter Mouton.
Sanada, H. (2010). Distributions of motifs in Japanese texts. In P. Grzybek, E. Kelih, & J. Mačutek (Eds.), *Text and language: Structures, functions, interrelations, quantitative perspectives* (pp. 183–194). Praesens.
Toury, G. (1980). *In search of a theory of translation*. The Porter Institute for Poetics and Semiotics, Tel Aviv University.
Tymoczko, M. (1998). Computerized corpora and the future of translation studies. *Meta*, *43*(4), 652–660. https://doi.org/10.7202/004515ar
Volansky, V., Ordan, N., & Wintner, S. (2015). On the features of translationese. *Digital Scholarship in the Humanities*, *30*(1), 98–118. https://doi.org/10.1093/llc/fqt031
Wimmer, G., & Altmann, G. (1996). The theory of word length: Some results and generalizations. In P. Schmidt (Ed.), *Glottometrika, 15: Issues in general linguistic theory and the theory of word length* (pp. 112–133). Wissenschaftlicher Verlag Trier.
Wimmer, G., & Altmann, G. (1999). *Thesaurus of univariate discrete probability distributions*. Stamm.
Wu, J., Liu, K., Hu, R., & Zhou, W. (2023). 翻译汉语和原创汉语句法复杂度对比研究 [A comparative study of the syntactic complexity of translated Chinese and original Chinese]. *外语教学与研究* [Foreign Language Teaching and Research], *55*(2), 264–275.
Xia, F. (2000). *The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0)*. (IRCS Technical Reports Series, 38).
Xiao, R., & Hu, X. (2015). *Corpus-based studies of translational Chinese in English-Chinese translation*. Shanghai Jiaotong University Press & Springer.
Xu, C., & Li, D. (2024). More spoken or more translated: Exploring the known unknowns of simultaneous interpreting from a multidimensional analysis perspective. *Target*, *36*(3), 445–480. https://doi.org/10.1075/target.22028.xu
Yan, J. (2017). The rank-frequency distribution of part-of-speech motif and dependency motif in the deaf learners' compositions. In H. Liu & J. Liang (Eds.), *Motifs in language and text* (pp. 181–200). De Gruyter Mouton.
Zhang, H. (2023). *基于依存关系的汉英句法计量特征比较* [Quantitative syntactic features in Chinese and English: A comparative study based on dependency relations]. 浙江大学出版社 [Zhejiang University Press].
Zhang, H., & Liu, H. (2016). Motifs in reconstructed RST discourse trees. *Journal of Quantitative Linguistics*, *24*(2–3), 107–127. https://doi.org/10.1080/09296174.2016.1265794
Zhang, H., & Liu, H. (2017). Motifs in generalized valencies. In H. Liu & J. Liang (Eds.), *Motifs in language and text* (pp. 231–260). De Gruyter Mouton.

Appendix

Corpus data concerning the study are publicly available on Open Science Framework (https://osf.io/serqj/).