Digital Scholarship in the Humanities, 2024, **00**, 1–17
https://doi.org/10.1093/llc/fqae030
Full Paper

OXFORD

Entropy-based syntactic tree analysis for text classification: a novel approach to distinguishing between original and translated Chinese texts

Zhongliang Wang [1], Andrew K. F. Cheung [2], Kanglong Liu [2,*]

[1]School of Foreign Languages and Commerce, Guangzhou Railway Polytechnic, Guangzhou, China
[2]Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China
*Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong, China. E-mail: kl.liu@polyu.edu.hk

Abstract

This research focuses on classifying translated and non-translated Chinese texts by analyzing syntactic rule features, using an integrated approach of machine learning and entropy analysis. The methodology employs information entropy to gauge the complexity of syntactic rules in both text types. The methodology is based on the concept of information entropy, which serves as a quantitative measure for the complexity inherent in syntactic rules as manifested from tree-based annotations. The goal of the study is to explore whether translated Chinese texts demonstrate syntactic characteristics that are significantly different from those of non-translated texts, thereby permitting a reliable classification between the two. To do this, the research calculates information entropy values for syntactic rules in two comparable corpora, one of translated and the other of non-translated Chinese texts. Then, various machine learning models are applied to these entropy metrics to identify any significant differences between the two groups. The results show significant differences in the syntactic structures. Translated texts have a higher degree of entropy, indicating more complex syntactic constructs compared to non-translated texts. These findings contribute to our understanding of the effect of translation on language syntax, with implications for text classification and translation studies.

Keywords: syntactic rules; entropy; machine learning; text classification; translation studies.

1. Introduction

Translation plays a pivotal role in facilitating cross-cultural communication, breaking down language barriers, and fostering global understanding (House 2014). Beyond the mere transposition of words, translation involves the complex interplay of meaning, structure, and nuance, requiring a deep understanding of both source and target languages (Toury 1995). Furthermore, the influence of translation spans various fields, including the dissemination of scientific knowledge (Tabrizi and Pezeshki 2015), financial reporting (Wang, Liu, and Moratto 2023), and the preservation and transmission of literature across different cultures (Li, Zhang, and Liu 2011). Recognizing the importance of translation, it becomes increasingly important to unravel the intricacies of translated language (Laviosa 2002). Researchers have made considerable progress in identifying specific linguistic features that are consistently found in translated texts, regardless of the languages involved in the translation process (Baker 1993; Liu and Afzaal 2021). These unique features, commonly known as translation universals, are defined as “the features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems” (Baker 1993: 243). Given the growing demand for translation services in diverse fields such as business, academia, entertainment, and diplomacy (Pym et al. 2012), it is crucial to undertake thorough studies to analyze the linguistic features in translated texts. Such research not only offers valuable insights for translation theory and practice (Gambier 2016), but also raises important questions about language transfer during the translation process (Gile 2009).

Among the proposed translation universals, simplification is perhaps the most debated hypothesis (Liu and Afzaal 2021). This refers to the tendency of translators to subconsciously simplify the language, the message, or both (Baker 1996: 176). Previous research has primarily relied on specific and simple linguistic parameters as evidence to support the existence of translation universals, including simplification. One such parameter, average sentence length, has often been used as an indicator of simplification (Laviosa 2002; Kajzer-Wietrzny, Whyatt, and Stachowiak 2016). However, the inconsistency of this metric across different language pairs (Pym 2008; Xiao and Yue 2009) suggests that relying solely on individual linguistic features can lead to inconsistent results. Moreover, past studies on translation universals have frequently focused on literary texts, neglecting genre as a significant variable that can influence the features of translational language (Blum-Kulka and Levenston 1983; Chesterman 2004; Zasiekin 2016). Therefore, genre should be taken into account when investigating translation universals (Delaere, De Sutter, and Plevoets 2012).

Text classification, a crucial field in computational linguistics, focuses on automatically categorizing texts into predefined categories and has significant implications across various domains, such as digital humanities and legal forensics. One of the less explored, yet increasingly important challenges within this field is distinguishing between translated and non-translated texts (Wang, Liu, and Liu 2024). This distinction is essential for training more nuanced machine translation systems. Despite its importance, traditional text classification methods often fall short as they typically rely on surface-level features, such as basic n-grams (Baroni and Bernardini 2006). Deeper syntactic features, encapsulated in structures like syntactic trees, can provide valuable insights into the inherent differences between original and translated texts (Hu, Li, and Kübler 2018). To address these gaps, recent studies have employed information-theoretic measures like entropy or tree-based dependency measures to quantify complexity (Liu, Liu, and Lei 2022; Liu et al. 2022; Xu and Liu 2023). This study builds upon these methodologies, using information entropy to meticulously calculate syntactic rules. By integrating measures of entropy in these syntactic structures, our approach aims to provide a more robust framework for identifying these nuances, potentially leading to improvements in both academic research and practical applications in text analysis. Specifically, the application of entropy as a methodological tool presents a novel approach to distinguishing translational language from original language and further explores this area from a computational linguistic perspective by building on previous work (Liu, Liu, and Lei 2022; Liu et al. 2022). The application of syntactic rules in Context-Free Grammar in our study enables a structural analysis of sentences. By applying these rules to translated Chinese texts, we aim to gain valuable insights into the syntax and structure of translated Chinese, which may exhibit differences from original Chinese. Furthermore, the incorporation of machine learning algorithms could yield more accurate and sophisticated models for text classification. To ensure the broad applicability of our findings to translated Chinese texts in general, we utilize a genre-balanced corpus, thereby mitigating potential bias from genre-specific characteristics. Through this refined methodology, we aim to enhance the reliability and generalizability of our study's results.

2. Related work

2.1 Simplification in translation

Simplification, in the context of translation, is often characterized as a subconscious process wherein translators inadvertently streamline the language, the message, or both (Baker 1996: 176). In the pre-digital era, simplification is assessed using a parallel corpus methodology, which allows for a comparative analysis between original and translated texts. However, before the widespread availability of corpora and computational tools, the collection and analysis of texts were primarily manual tasks, and the chosen source and target texts were often limited in size. In this context, Blum-Kulka and Levenston (1983: 119) examined the translation of Hebrew and English pairs and found the occurrence of lexical simplification, a process that conveys the same meaning with fewer words. Focusing on syntactical simplification, Vanderauwera (1985) observed that translated texts often simplify complex syntactic patterns by substituting finite clauses with non-finite ones. However, due to the small sample size and the absence of statistical methods, the findings from this period cannot be generalized. Despite this, these early studies laid crucial groundwork for understanding simplification in translation.

Contrasting the parallel corpus approach, Baker (1993) innovatively proposed the comparable corpus approach. In this method, translated texts are compared with non-translated original texts to examine simplification. Following Baker's proposal, research on translation universals began to adopt quantitative methodologies supported by statistical techniques. Malmkjær (1997) found that translated texts tend to use stronger punctuation and simpler clauses instead of complex syntactic structures. Laviosa (1998) used four main indicators (lexical density, core vocabulary coverage, list head coverage, and average sentence length) to investigate the lexical features of English translations of narrative prose. Her findings revealed that translated English differed from native English in these four areas, providing evidence for simplification in translations. Similarly, Olohan (2004) used lexical diversity to compare translated and native English fiction and found that translated fiction used fewer color synonyms. Pastor et al. (2008) employed natural language processing tools, readability indices, and other measures to investigate simplification. Their results suggested that non-translated original texts exhibited higher lexical density and richness than their translated counterparts. The collective findings from these studies, made possible by the application of quantitative methods and statistical analysis, provide support for the simplification hypothesis.

Despite extensive research on simplification within the framework of translation universals, conflicting findings have called its validity into question. Some studies have reported longer average sentence lengths in translated texts (e.g. Xiao and Yue 2009), contradicting previous findings by Malmkjær (1997) and Laviosa (1998). Mauranen (2006) found that non-translated texts exhibited clearer and more stable multi-word patterns compared to their translated equivalents. Jantunen (2004) found no clear or consistent evidence supporting translation universals. Ferraresi et al. (2018) reported that translated texts were more complex and had a higher lexical density. These conflicting findings highlight that simplification is a dynamic phenomenon in translation, hinting at the possibility that variables such as the specific language pair and the genre of the text in question may significantly affect the degree and nature of simplification.

The lack of consensus in research on translation universals may stem from the selective use of language indicators to support specific translation universals (Liu, Liu, and Lei 2022). To overcome this limitation, the use of quantitative measures is recommended (Liu et al. 2022). Accordingly, we have opted to use entropy, a well-established quantitative measure of linguistic complexity, to investigate simplification in translated texts. Through the application of entropy analysis, the intricacies of simplification in translated texts become quantifiable, thereby facilitating a rigorous investigation into how it interacts with other linguistic phenomena. The findings from such research could provide a more definitive, evidence-based understanding of the simplification process, enhancing the ongoing dialogue regarding the impact of translation on linguistic complexity.

2.2 Entropy in language and translation studies

Shannon's concept of entropy (1948) plays a pivotal role in quantifying information and mapping the content within a specific dataset. Essentially, entropy is a measure of uncertainty or informational content within a particular dataset, providing a numerical representation of the average information or surprise derived from an event or observation. The formula for entropy, as introduced by Shannon, calculates the average information content of a message, where the potential outcomes or messages of a random event are represented as probabilities. Outcomes with higher probabilities have less impact on the overall entropy than those with lower probabilities.

In the field of language research, entropy has been extensively applied, beginning with Genzel and Charniak (2002) who proposed the "constancy rate principle." This principle indicated that local entropy increases with sentence number, providing insights into the linguistic principles associated with entropy. Tanaka-Ishii (2005) expanded on this exploration, illustrating how the uncertainty of tokens following a sequence is crucial in determining context boundaries. Subsequent studies incorporated entropy in their linguistic analyses, demonstrating its wide applicability (Juola 2008; Mehri and Darooneh 2011; Suo et al. 2012; Yang et al. 2013; van Ewijk and Avrutin 2016; Bentz and Alikaniotis 2016; Bentz et al. 2017; Lowder et al. 2018; Friedrich, Luzzatto, and Ash 2020; Friedrich 2021). In cultural studies, entropy is used as a tool to assess cultural complexity based on its degree of freedom, considering both the number of states in a system and their frequency distribution (Kockelman 2009). Further studies by Juola (2013), and Zhu and Lei (2018) also utilized entropy in their analysis of American and British cultures, respectively. In translation research, entropy has garnered significant attention due to its influence on cognitive and linguistic processes. Wei (2022) introduced surprisal (ITra) and entropy (HTra) as metrics approximating cognitive load, with ITra found to be a more accurate predictor of translation production time and HTra more effectively predicting source text reading time. Chen, Liu, and Altmann (2017) used entropy to demonstrate how text types exhibit unique linguistic profiles, and Yerkebulan et al. (2021) developed an entropy-based approach to detect patterns in multilingual texts. More recent studies by Liu, Liu, and Lei (2022) and Liu et al. (2022) showcased the effectiveness of an entropy-based approach in examining translation universals in Chinese texts. These studies provide a compelling argument for the potential benefits of employing more sophisticated entropy-based methodologies in future research.

2.3 Machine learning and classification between translated and non-translated texts

Machine learning models have proven effective in differentiating between translated and non-translated texts due to their capacity to identify complex relationships between features and labels (Bernardini and Baroni 2005; Volansky, Ordan, and Wintner 2015; De Clercq et al. 2021; Liu et al. 2022). These models utilize large datasets of labeled texts, each identified as either original or translated, to predict the classification of new texts. They use a variety of statistical and linguistic features extracted from the text, enhancing the classifiers' robustness and accuracy. Research has employed machine learning with various features and languages. For example, Baroni and Bernardini (2006) used word or part-of-speech n-grams to classify Italian geopolitical journal articles, achieving an accuracy rate of 86.7 per cent with Support Vector Machines (SVMs). Ilisei et al. (2010) used different classifiers to distinguish between translated and non-translated Spanish texts, finding that all classifiers performed better when simplification features were included. Volansky, Ordan, and Wintner (2015) operationalized the simplification hypothesis using various features, including TTR, mean word length, syllable ratio, lexical density, mean sentence length, proportion of most frequent words, and mean word rank. Mean word rank was calculated using a list of the 6,000 most frequent English words, with words not on the list either assigned the highest rank of 6,000 (mean word rank (1)) or ignored altogether (mean word rank (2)). The authors found that mean word rank (2), which ignores words not on the frequency list, was the most effective feature for distinguishing between translated and original texts, achieving an accuracy of 77 per cent.

Research in translation studies has increasingly applied machine learning to study translated Chinese texts. Earlier work by Nisioi and Dinu (2013) and Rabinovich and Wintner (2015) employed clustering techniques to pinpoint features typical of translation language. Similarly, Rubino, Lapshinova-Koltunski, and Van Genabith (2016) sought to distinguish between novice and professional translators, noting that a blend of various features enhanced classification accuracy and underscored the need for additional investigation. Likewise, Hu and Kübler (2021) utilized SVM classifiers to recognize unique features in translated Chinese texts, confirming their distinct status within the Chinese language. However, many studies, such as those by Baroni and Bernardin (2006), have limited themselves to a predefined set of features. There is a need for more research that systematically investigates and compares a wider range of potential features (Baker 1993). More recent research (Liu et al. 2022) has advanced the field by combining various machine learning models with entropy-based metrics to differentiate between original and translated Chinese texts. This innovative approach has shed light on the intrinsic syntactic properties that are characteristic of translated texts. This study aims to expand upon these groundbreaking findings by incorporating a wider array of machine learning algorithms and leveraging entropy measures to conduct a more thorough examination of tree-based syntactic features.

3. Research gaps and questions

The literature review underscores the significant potential of entropy-based metrics to provide meaningful insights in translation studies, particularly considering recent empirical evidence from research on typologically diverse language pairs such as English and Chinese (Xiao and Dai 2014). Incorporating entropy-based indicators in translation studies allows for the quantitative evaluation of text complexity and the degree of simplification in translations. By applying the fundamental concept of entropy from information theory, researchers gain a unique perspective for examining how linguistic information is compressed and transformed during the translation process. Although relatively underutilized in the field, entropy-based metrics present a promising avenue for innovative research that can substantially improve our understanding of translation universals and refine methods for assessing translation quality.

First, methods for measuring syntactic complexity, such as using Part of Speech (POS) entropy, have certain limitations. While these approaches provide a general assessment of complexity, they may not fully capture the intricate syntactic rules and structures specific to individual languages. This measure assesses the degree of uncertainty or randomness in the distribution of POS tags within a text. However, there is considerable room for improvement by integrating more refined parameters into the entropy calculation for syntactic structures. Our study departs from these conventional methods by concentrating on the entropy of syntactic rules themselves. We transcend simple POS tag distribution analysis to delve into the deeper structural intricacies of language. This entails a thorough investigation of the complexity and variation in syntactic structures within the analyzed texts, offering a more nuanced comprehension of linguistic complexity. This focus on syntactic structures enables us to capture the diverse ways in which sentences are constructed and phrases are combined, leading to a more comprehensive grasp of syntactic complexity. Additionally, while machine learning algorithms have proven to be efficient in pinpointing characteristics of "translationese" and can be synergized with entropy-based measures for more precise evaluations of translational simplification, the majority of existing studies are limited to the deployment of SVM. To bridge this gap, this study introduces a novel methodology that computes the entropy of principal syntactic constructions using a tree-based algorithm. By deconstructing sentences into specific constructions, this research endeavours to determine whether the syntax of Chinese translations, derived from English source texts, is indeed more simplified in contrast to native Chinese writing. The investigation employs a dual approach that integrates machine learning with entropy-based analysis. This study aims to address the following two research questions:

RQ1: Can machine learning algorithms differentiate between translated and non-translated Chinese

texts utilizing the complexity of syntactic rule features from an entropy-based perspective?
RQ2: If the answer to the first question is affirmative, which features are most crucial for this classification task?

4. Materials and methods

The study utilized entropy-based syntactic tree analysis and machine learning models to distinguish between original and translated Chinese texts. Figure 1 illustrates the workflow of the methods in the study.

4.1 Corpora

This study employs two corpora: the Lancaster Corpus of Mandarin Chinese (LCMC) and the Zhejiang University Corpus of Translational Chinese (ZCTC), representing native and translated Chinese texts, respectively. Both corpora were modeled after the Freiburg-LOB (FLOB) Corpus. The FLOB Corpus comprises approximately one million tokens of written British English across fifteen text categories, published in the early 1990s (Hundt, Sand, and Siemund 1998). The LCMC and ZCTC were developed as Chinese counterparts to FLOB, using identical sampling techniques and matching the corresponding sample duration (McEnery, Xiao, and Mo 2003; Xiao and Hu 2015). These balanced corpora, each containing one million words, are comparable in size and publicly available. As two major corpora in the Chinese language, numerous studies have utilized these resources, as evidenced by research references (Xiao and Hu 2015; Liu, Liu, and Lei 2022). Both corpora consist of 500 texts, each about 2,000 words long, across 15 text categories. These categories cover four macro genres: press, general prose, academic writing, and fiction. The text types contained in both corpora, presented in Table 1, while not exhaustive, are considered representative of both translation and non-translation texts (Liu, Liu, and Lei 2022). They are deemed sufficiently diverse to meet the needs of the current research.

4.2 Feature extraction

This study builds on the research conducted by Hu, Li, and Kübler (2018) and extracts syntactic rule features for each text of LCMC and ZCTC using Context-Free Grammar through the StanfordNLP package (Qi et al. 2018). In Context-Free Grammar, a context-independent syntax $G = (N, \Sigma, R, S)$ is defined, where $N \in \mathbb{R}^{G1}$is a set of non-terminal symbols, Σ is a set of terminal symbols, R is a set of rules of the form $(R : X \rightarrow \{Y^1, Y^2, \ldots, Y^n\}, for\, n \geq 1, X \subset N \in \mathbb{R}^G, Y^i \in \ \mathbb{R}^G)$, and $S \subset N$ is a distinguished start symbol.

For instance, Σ =('浙江大学有七个校区'), and $N = \{S, IP, NP, NR, NN, VP, VE, QP, CD, CLP, \ M\}$. The syntactic rules and corresponding Chinese parts of speech are displayed in Tables 2 and 3, respectively. The syntactic tree is illustrated in Fig. 2. The description of syntactic components and their abbreviations are shown in the Appendix. This sentence is a simple declarative clause which consists of a simple clause headed by INFL. It consists of a noun phrase (NP; 浙江大学) and a verb phrase (有七个校区). The NP (浙江大学) consists of a proper noun (浙江) and a noun (大学), the verb phrase (有七个校区) consists of a verb (有) and a NP (七个校区). The NP (七个校区) consists of a quantifier phrase (七), a classifier phrase (个), and a NP (校区). They are represented by a cardinal number, a measure word, and a noun respectively. In this example, the {NR, NN} can be regarded as a syntactic rule, and it is a syntactic rule of the NP category (NP is the node of the NR—NN structure).

4.3 Data analysis

We utilized StanfordNLP (Levy and Manning 2003), an open-source tool, to extract occurrences of syntactic rules in each text within the LCMC and ZCTC corpora. The analysis produced a comprehensive set of thirty-one general syntactic categories, categorized based on the syntactic component that acts as the node for each structure. These thirty-one categories encompass a total of 7,056 identified syntactic rules across the two corpora. To analyze the complexity and

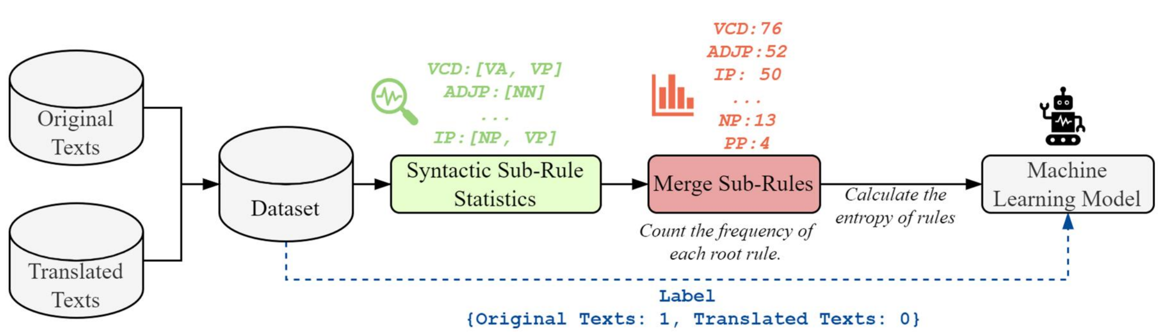

Figure 1. Overview of methodological approach employed in the study.

Table 1. Design and genre composition of LCMC and ZCTC.

Genres	Text types	Number of texts	Proportion (%)
Press	Press reportage	44	8.80
	Press editorial	27	5.40
	Press reviews	17	3.40
General Prose	Religious writing	17	3.40
	Instructional writing	38	7.60
	Popular lore	44	8.80
	Biographies and essays	77	15.40
	Reports and official documents	30	6
Academic writing	Academic prose	80	16
Fiction	General fiction	29	5.80
	Mystery and detective fiction	24	4.80
	Science fiction	6	1.20
	Adventure fiction	29	5.80
	Romantic fiction	29	5.80
	Humor	9	1.80
Total		500	100

Table 2. Illustration of the syntactic rules.

Syntactic category	Syntactic rule
NP	{NR, NN}
NP	{NN}
NP	{QP, CLP, NP}
VP	{VE, NP}
IP	{NP, VP}
QP	{CD}
CLP	{M}

Table 3. An example of parts of speech.

POS	Chinese word
NR	浙江
NN	大学
VE	有
CD	七
M	个
NN	校区

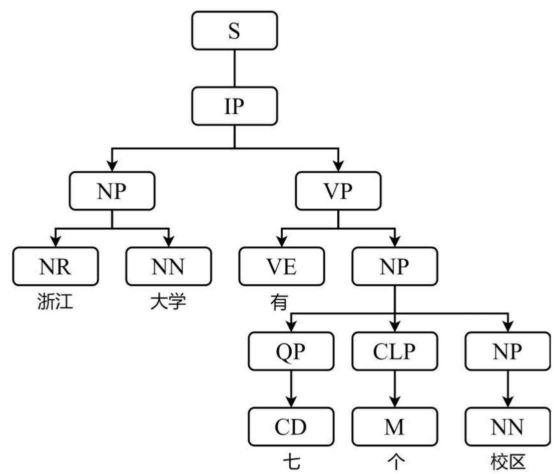

Figure 2. Syntax tree diagram for the provided example.

variability in the texts, we computed entropy values for both the individual syntactic rules and the overarching syntactic categories within each text across the corpora by taking into account their frequency of occurrence. The entropy of rule i is expressed as follows: (The formula is as follows:)

$$H(i) = -p_i log_2 p_i \tag{1}$$

where p_i is the probability of rule i appearing in a certain text, calculated as the ratio of the frequency of rule i to the total frequency of all rules observed in the text.

- Syntactic rules: the rules under the syntax tree in a sentence of text, such as `{'VP': ['VRD', 'NP']}` and `{'DP': ['NN', 'CLP']}`. In our analysis, we define the scope of these segments as one layer within the syntax tree.
- Syntactic categories: aggregation of syntactic rules (the node of the syntactic rule in the tree), such as VP and DP.

As shown in Fig. 3, our calculation is divided into three steps:

1) Use the StanfordNLP package to analyze syntactic trees and summarize syntactic rules for the two types of texts (i.e. A and B).

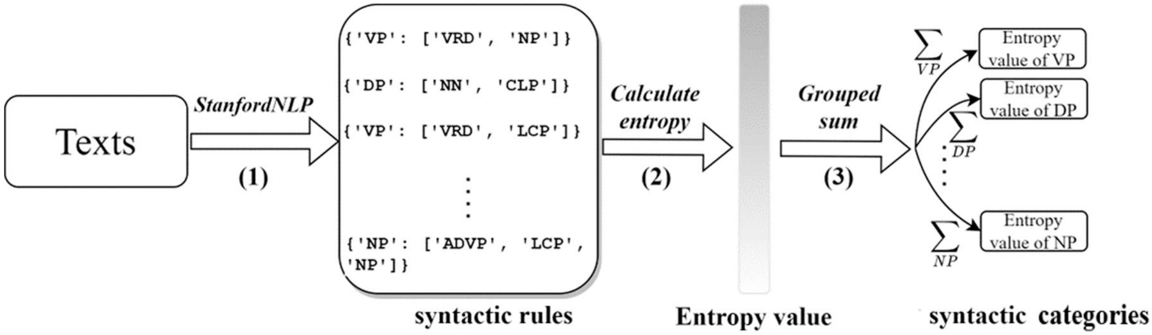

Figure 3. Process of calculating entropy values.

2) We calculate the entropy of each syntactic rule based on Equation X. After that, we obtain a collection of syntactic rules $\mathbb{I}$.
3) The node of a syntactic rule in the tree is a syntactic category (e.g. `{'VP': ['VRD', 'NP']}-> VP)`. The syntactic rules with the same type of syntactic structure as the node are classified into the same syntactic category. We group the entropy of each syntactic rule and then sum them to obtain the entropy of the syntactic category.
4) The entropy calculation of syntactic category $\mathcal{I}$ is as follows:

$$H(\mathcal{I}) = \sum_{i \in \mathcal{I}} -p_i log_2 p_i, \quad \text{for all } i \in \mathbb{I} \qquad (2)$$

In this section, we describe the methodology for calculating the entropy values associated with syntactic rules and categories. We provide the following example to illustrate how we calculate the entropy of NPs, assuming the text contains two sentences:

Text: '(1) 因特网的问世现在得以使任何地方的任何人几乎即刻可以获得技术和最新的商业方式。
(2) 因特网还具有潜力, 通过加快信息流通速度来完善全球竞赛规则。'

Figure 4 presents the syntax trees corresponding to the two sentences. For the purpose of this demonstration, we focus on the syntactic category "NP." Within the two sentences, eight syntactic rules belong to the syntactic category "NP." We analyze the frequency of each rule as follows:

{'NP': ['DNP', 'NP']}: 2
{'NP': ['NN']}: 7
{'NP': ['NT']}: 1
{'NP': ['DNP', 'DP', 'NP']}: 1
{'NP': ['DP', 'NN']}: 1
{'NP': ['NP', 'CC', 'NP']}: 1
{'NP': ['NN', 'NN']}: 1
{'NP': ['NN', 'NN', 'NN']}: 2

The total frequency count for all syntactic rules observed in the two sentences is 47. The entropy values for these eight syntactic rules are calculated using the formula for entropy in information theory:

{'NP': ['DNP', 'NP']}: $-\frac{2}{47} log_2 \frac{2}{47} = 0.1937$

{'NP': ['NN']}: $-\frac{7}{47} log_2 \frac{7}{47} = 0.4062$

{'NP': ['NT']}: $-\frac{1}{47} log_2 \frac{1}{47} = 0.1182$

{'NP': ['DNP', 'DP', 'NP']}: $-\frac{1}{47} log_2 \frac{1}{47} = 0.1182$

{'NP': ['DP', 'NN']}: $-\frac{1}{47} log_2 \frac{1}{47} = 0.1182$

{'NP': ['NP', 'CC', 'NP']}: $-\frac{1}{47} log_2 \frac{1}{47} = 0.1182$

{'NP': ['NN', 'NN']}: $-\frac{1}{47} log_2 \frac{1}{47} = 0.1182$

{'NP': ['NN', 'NN', 'NN']}: $-\frac{2}{47} log_2 \frac{2}{47} = 0.1937$

We aggregate the entropy values of the eight syntactic rules to determine the entropy value for the "NP" syntactic category, which is calculated to be 2.3391. This analytical method can be similarly applied to additional syntactic categories.

To further our analysis, we implemented four sophisticated machine learning algorithms: Adaptive Boosting (AB), Random Forests (RFs), Gradient Boosting (GB), and Extremely Randomized Trees (ET). These algorithms were employed to calculate the feature importance coefficients for the thirty-one identified syntactic categories. This methodology was instrumental in the initial phase of identifying the most significant syntactic rule features.

Upon analyzing the feature importance coefficients generated by the four machine learning models, we identified four top-ranking syntactic categories: IP (simple clause headed by INFL), VCD (coordinated verb compound), PP (prepositional phrase), and ADJP (adjective phrase), consisting of 2,023 syntactic rules. The four machine learning algorithms were then used

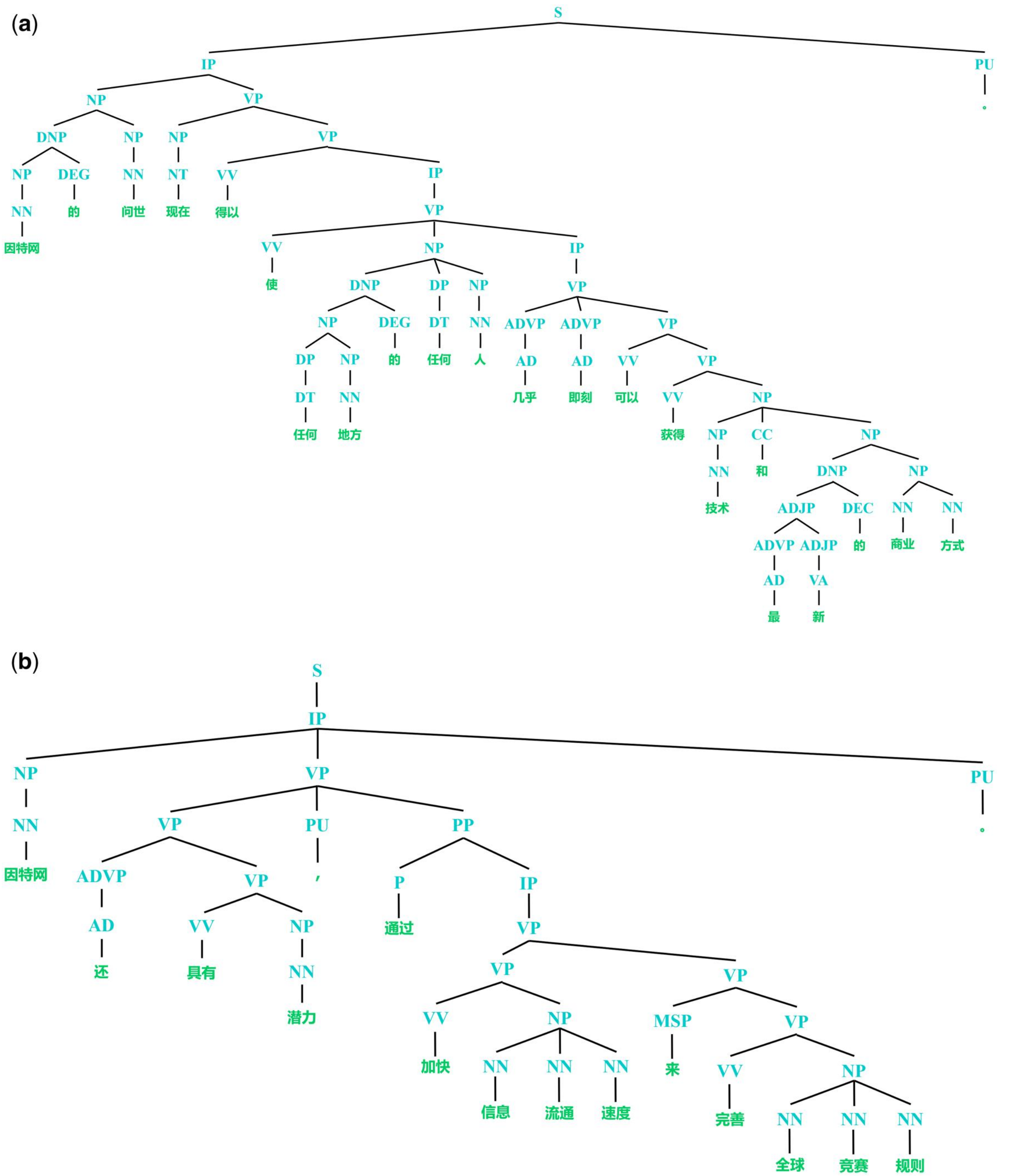

Figure 4. Syntax trees of the example: (a) Sentence (1); (b) Sentence (2).

to calculate the feature importance coefficients for the 2,023 syntactic features within these four syntactic categories. Subsequently, we selected the twenty syntactic rules with the highest feature importance coefficients. These rules were then used as the features to build models using the four machine learning algorithms: AdaBoost, ET, RF, and GB. To validate our models, we employed a 5-fold cross-validation approach and divided the dataset into training and testing subsets at an 8:2 ratio. We made binary predictions for both original and translated texts. The accuracy and area under the curve (AUC) values were then calculated and

Input: a model ϕ, $entropy\ of$ syntactic category $\mathcal{H}$ and label Y.
Output: Sub-syntactic rule importance within important syntactic categories

```
// Train model
```
$\phi_1 \longleftarrow \phi(\mathcal{H}, Y)$
```
// Obtain features importance in model ϕ
```
$\mathcal{S}_1$=feature_importance(ϕ_1)
```
// Obtain important syntactic categories
```
$\mathbb{C} \longleftarrow \underset{\mathcal{H}}{arg\,max}\, \mathcal{S}_1$
```
// Retrain the model for all sub-syntactic rules within ℂ.
```
$\phi_2 \longleftarrow \phi(\mathcal{H}(\mathbb{C}), Y)$
```
// Analyze the importance of sub-syntactic rules.
```
$\mathcal{S}_2$=feature_importance(ϕ_2)
Return $\mathcal{S}_2$

Figure 5. Pseudocode for calculating feature importance coefficients in machine learning models.

compared across the different models to identify the model with the best performance.

The pseudocode for implementing the calculation of feature importance coefficients in machine learning models is shown in Fig. 5.

4.4 Machine learning algorithms

In this study, we employed four machine learning algorithms for text classification: AB (AdaBoost), RFs, GB, and ET.

4.4.1 AB

AB (AdaBoost), introduced by Freund and Schapire in 1997, is a well-established ensemble machine learning algorithm known for its robust performance in various applications (Freund and Schapire 1997). AdaBoost combines a series of weak classifiers to form a more powerful and accurate classifier. Its iterative learning process adjusts the distribution of training data in each iteration, giving more weight to previously misclassified instances (Freund and Schapire 1997). AdaBoost assigns adaptive weights to each classifier, signifying their contribution to the final decision (Zhou 2012). The algorithm's flexibility allows its application to both classification and regression tasks, making it a versatile tool in machine learning (Hastie, Tibshirani, and Friedman 2009).

4.4.2 RFs

RFs, an ensemble learning method introduced by Breiman (2001), has become a key technique in machine learning due to its strong performance and versatility. It creates multiple decision trees and outputs the mode of the classes (for classification) or the mean prediction (for regression) of the individual trees, reducing model variance and improving accuracy (Breiman 2001). RFs' strength lies in bagging, which creates diverse training sets from the original dataset, addressing overfitting (Breiman 1996), and the random subspace method, which selects random feature subsets at each node, encouraging feature diversity and enhancing generalization (Ho 1998).

4.4.3 GB

GB, a powerful machine learning algorithm introduced by Friedman (2001), has earned widespread acclaim in the field of predictive modeling. It is praised for its ability to handle various data types, its resistance to overfitting, and its exceptional predictive accuracy. As part of the broader family of boosting algorithms, GB works on the principle of transforming weak learners into strong ones through a systematic, iterative process (Schapire 1990). The distinctiveness of GB lies in its strategy that uses the gradient of the loss function to guide the sequential construction of weak learners. Each subsequent learner is trained to correct the residual errors left by its predecessor, thereby gradually improving the model's performance in a stage-wise manner (Friedman 2001). This iterative process continues until an acceptable level of error is achieved or a predetermined number of learners are included in the model. The versatility of GB applies to various tasks, including both classification and regression.

4.4.4 ET

ET, proposed by Geurts, Ernst, and Wehenkel (2006), is an ensemble learning method that extends the RF algorithm by incorporating an additional level of randomness in decision tree construction. This approach offers benefits in computational efficiency and model robustness. ET distinguishes itself by using a randomized selection process for cut-points within each feature during node splitting, differing from traditional decision tree algorithms and RF that select optimal cut-points. This promotes diversity among individual trees, potentially enhancing model generalization

(Geurts, Ernst, and Wehenkel 2006). ET is applicable to various tasks, including classification, regression, and feature selection (Bosch, Zisserman, and Munoz 2007; Statnikov, Wang, and Aliferis 2008; Joulin et al. 2017), highlighting its versatility and significant role in machine learning.

5. Result

The 7,056 syntactic rule features extracted from the two corpora were categorized into thirty-one general syntactic rule feature categories. To perform preliminary feature mining within these categories, four machine learning models were employed. These models also calculated the feature importance coefficients for each of the thirty-one syntactic categories, aiding in the preliminary screening of syntactic rule features.

The study's findings are presented in Fig. 6, where the analysis using four machine learning models identifies the syntactic categories IP (simple clause headed by INFL), VCD, PP, and ADJP as the primary features differentiating translated and non-translated texts. Table 4 outlines the entropy values for these syntactic categories, and Fig. 7 visually represents these findings using boxplots. The entropy comparison shows that translated Chinese texts (ZCTC) have higher entropy values in these syntactic categories compared to non-translated Chinese texts (LCMC), indicating more complex and informationally dense syntactic structures in translated Chinese. This higher entropy, reflecting both frequency and distribution, suggests a wider and more frequent use of these syntactic constructions in translated Chinese, while their usage in native Chinese seems relatively limited.

After the initial analysis, IP, VCD, PP, and ADJP, which include 2,023 syntactic rule features (1,841 for IP, 79 for VCD, 43 for PP, and 60 for ADJP, respectively), were selected for further examination. The GB model, recognized for its precision, was used to identify the most critical syntactic rule features within these categories. This step was crucial for developing a binary prediction model with heightened accuracy. Table 5 presents the results of this comprehensive analysis, displaying the top twenty significant syntactic rule features identified by the four machine learning models for the IP, VCD, PP, and ADJP categories. These features are ranked based on their importance coefficients, providing a clear representation of the syntactic elements that play a crucial role in distinguishing between translated and non-translated Chinese texts. This approach not only enhances the model's accuracy but also offers deeper insights into the syntactic features characteristic of translation.

We then employed four machine learning algorithms, using the twenty significant syntactic rules as features for model training and testing. The ROC Curve of the four models is depicted in Fig. 8, with the model evaluation results presented in Table 6. In our comparative analysis, the RF model achieved the highest AUC value at 92.93 per cent, closely followed by the GB model at 91.6 per cent, and the Extremely Randomized Trees (ET) model at 89.63 per cent. In terms of accuracy, both the RF and GB models excelled, each with 88.5 per cent. This was notably higher than the ET model, which reached an accuracy of 87.3 per cent. The AdaBoost model reported an AUC of 85.85 per cent and an accuracy rate of 80 per cent, making it the least accurate among the models evaluated, yet still showing significant predictive power. These findings highlight the capability of the machine learning models to effectively use the identified syntactic rule features for binary classification tasks between translated and non-translated texts.

6. Discussion

In this study, we examined the complexity of translated texts by analyzing the entropy of four categories of syntactic rules: IP (simple clause headed by INFL), VCD, PP, and ADJP. These syntactic rules are essential for sentence construction, and our findings reveal a notable pattern: certain syntactic features of English appear to be transferred into and retained in translated Chinese texts, resulting in syntactic structures that are more complex and categorically distinct from those in native Chinese. These outcomes imply that translators may demonstrate a diverse range of syntactic choices when translating from English, especially in the IP, VCD, PP, and ADJP constructions. This increased variability and frequency in syntactic structures contribute to the overall higher entropy values observed in the translated Chinese texts compared to their native counterparts.

The integration of entropy and syntactic structures has improved classification performance, increasing the AUC to 92.93 per cent and the accuracy to 88.5 per cent. This represents an advancement over previous methods, such as those described by Liu et al. (2022), which achieved an AUC of 90.5 per cent and an accuracy of 84.3 per cent by employing entropy values from character, wordform, and POS n-grams. Likewise, this approach outperforms the one by Baroni and Bernardini (2006), which utilized word and POS n-grams and attained an accuracy of 86.7 per cent.

The study employs entropy, drawing on principles from the field of information theory, which primarily focuses on the measurement, storage, and transmission of information. Our research has demonstrated that entropy can serve as a tool for distinguishing between

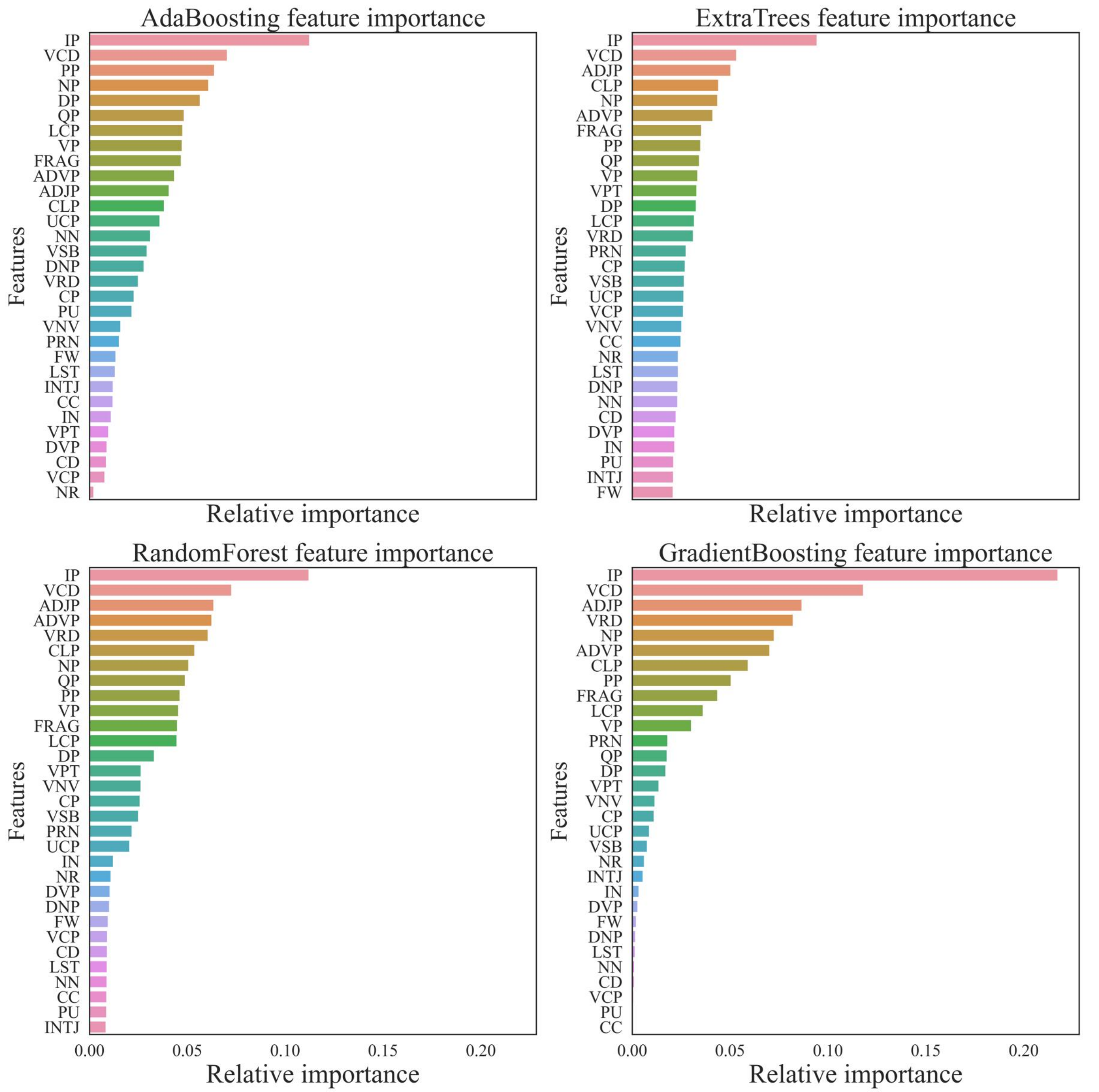

Figure 6. General rule features importance.

Table 4. Descriptive statistics of entropy values of IP, VCD, PP, ADJP of LCMC, and ZCTC,

	IP		VCP		PP		ADJP	
	LCMC	ZCTC	LCMC	ZCTC	LCMC	ZCTC	LCMC	ZCTC
Count	500	500	500	500	500	500	500	500
Mean	9.804	10.160	0.429	0.441	0.222	0.230	0.415	0.429
Std	0.673	1.028	0.031	0.047	0.017	0.025	0.038	0.040
Min	7.412	7.344	0.336	0.317	0.166	0.163	0.318	0.309
Max	12.351	14.501	0.534	0.626	0.295	0.332	0.561	0.565

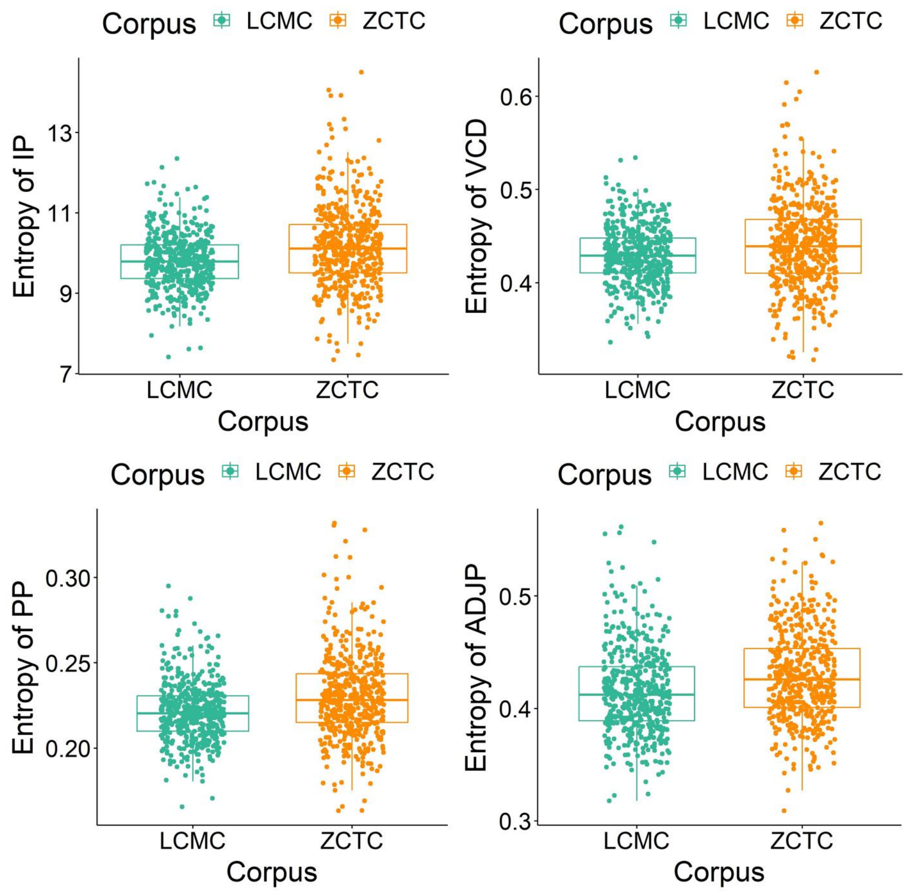

Figure 7. Boxplots of entropy of IP, VCD, PP, ADJP of LCMC, and ZCTC.

translated and native texts. The importance of empirical evidence in research is highlighted, as it provides concrete examples to support theoretical concepts. Unlike traditional methodologies that focus on individual linguistic elements, entropy offers a comprehensive view of text complexity (Shi and Lei 2022). Conventional linguistic analysis may be subject to the researcher's interpretation, while entropy provides a quantitative measure that can be systematically compared across a range of texts, thereby enhancing the objectivity and consistency of the analytical process (Liu, Liu, and Lei 2022).

In addition, we have demonstrated the significant contributions of machine learning models in distinguishing between translated and non-translated texts, advancing the fields of translation studies. As Dhar (2013) highlights, machine learning models efficiently process and analyze large datasets, a task that can be challenging for human analysts. This capability allows for the examination of large text corpora, leading to more robust and generalizable insights about the distinctive characteristics of translated texts. Furthermore, Bishop (2006) notes that these models excel at identifying complex patterns in data, a feature particularly useful in our context where subtle linguistic differences might elude traditional analysis. The models' ability to learn from a range of syntactic features enhances our understanding of textual nuances. Our findings have implications for both theory and practice in text classification and translation research. Moving beyond traditional models that rely on basic textual features, our research introduces syntactic rule features as a novel and effective tool for improving classification accuracy. This approach not only allows for nuanced differentiation between text types but also enhances the sophistication of text profiling.

The novel approach of employing entropy-based syntactic tree analysis for text classification not only enhances our understanding of linguistic structures in translation but also facilitates the development of more sophisticated tools for automated text analysis. In computational linguistics, the introduction of

Table 5. Top twenty important syntactic rule features,

Feature	AB	ET	RF	GB
{'IP': ['PU', 'VP']}	0.259	0.024	0.072	0.261
{'IP': ['PU']}	0.118	0.023	0.058	0.211
{'IP': ['PU', 'IP']}	0.075	0.019	0.050	0.211
{'VCD': ['VV', 'VV']}	0.067	0.016	0.023	0.074
{'IP': ['VP']}	0.031	0.011	0.016	0.062
{'ADJP': ['ADVP', 'ADJP']}	0.026	0.011	0.015	0.043
{'ADJP': ['JJ']}	0.025	0.010	0.015	0.036
{'IP': ['VP', 'IP']}	0.015	0.010	0.014	0.024
{'IP': ['ADVP', 'NP', 'VP']}	0.012	0.009	0.013	0.013
{'IP': ['IP']}	0.011	0.008	0.013	0.011
{'IP': ['ADVP', 'VP']}	0.011	0.008	0.012	0.007
{'IP': ['IP', 'IP']}	0.010	0.007	0.011	0.006
{'IP': ['VP', 'VP']}	0.010	0.007	0.010	0.005
{'IP': ['VV', 'VP']}	0.010	0.007	0.010	0.004
{'IP': ['IP', 'VP']}	0.009	0.007	0.010	0.004
{'IP': ['NP', 'VP', 'PU']}	0.008	0.007	0.009	0.004
{'IP': ['NP', 'VP']}	0.008	0.007	0.009	0.003
{'VCD': ['VA', 'VA']}	0.007	0.006	0.008	0.003
{'ADJP': ['NN']}	0.007	0.006	0.008	0.002
{'IP': ['QP', 'IP']}	0.006	0.006	0.007	0.002

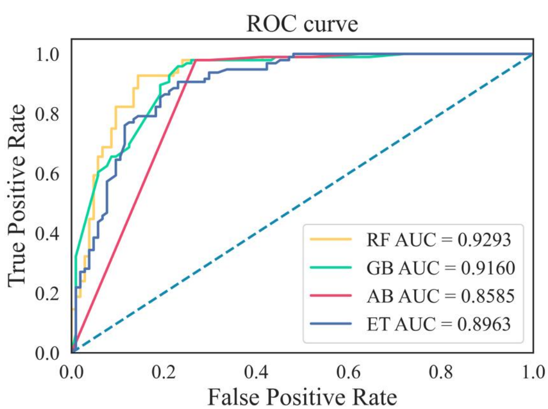

Figure 8. ROC curve of the four machine learning algorithms.

Table 6. Performance evaluation of the four classifiers.

Machine learning algorithm	Accuracy (%)	AUC (%)
RF	88.50	92.93
GB	88.50	91.60
ET	87.30	89.63
AdaBoost	80.00	85.85

entropy-based methods for syntactic analysis provides a nuanced metric for understanding complexity and variability within a language. This is particularly advantageous when analyzing translated texts, which often exhibit syntactic patterns that differ from native compositions due to translation norms and strategies (Baker 1993; Laviosa 1998). By integrating entropy measurements, our approach refines computational models, making them more sensitive to subtle linguistic shifts, which are crucial in tasks such as machine translation, text summarization, and authorship attribution. From the perspective of translation studies, this research illuminates the inherent syntactic differences between original and translated texts, which have been extensively discussed in the literature under the concept of translation universals (Toury 1995; Chesterman 2004). These include simplification, explicitation, and normalization, which can now be quantitatively analyzed through syntactic entropy. By applying computational techniques to problems in translation studies, this research fosters a more interdisciplinary approach, encouraging collaboration between the two fields. Such synergy is crucial as it harnesses the power of computational methods to address complex linguistic and cultural challenges, leading to advancements in automated translation software and a deeper understanding of cross-linguistic differences. The integration of entropy-based syntactic tree analysis for text classification not only enriches our understanding of language and translation processes but also paves the way for the development of innovative computational tools. The interdisciplinary nature of this research showcases the potential for computational linguistics and translation studies to work hand in hand, leveraging their respective strengths to push the boundaries of our understanding of language and translation.

7. Conclusion

In conclusion, this study's innovative approach, combining machine learning models and entropy-based syntactic rule features, provides valuable insights and methodological advancements in the study of syntactic simplification or complexification in translations. By employing entropy-based metrics, we have gained a more comprehensive understanding of translational language that goes beyond the narrow focus on individual linguistic features. The effectiveness of machine learning models in differentiating between translated and non-translated texts deepens our understanding of the complexities and nuances within translation studies. Furthermore, this research lays the groundwork for future studies in quantitative linguistics aiming to quantify simplification or complexification. However, it is crucial to acknowledge that our findings are specific to English-Chinese translations and may not fully capture the intricacies of other language pairs. Additionally, our focus on four general syntactic structures (IP, VCD, PP, and ADJP) may not encompass the entire complexity of translated texts. This highlights the need for future research to employ a diverse range of measures and approaches to provide a more comprehensive and nuanced understanding of translation phenomena.

Author contributions

Zhongliang Wang (Investigation, Visualization, Writing—original draft), Andrew Cheung (Investigation, Project administration, Validation, Writing—review & editing), and Kanglong LIU (Conceptualization, Data curation, Funding acquisition, Investigation, Methodology, Project administration, Supervision, Validation, Writing—review & editing)

Conflict of interest statement. None declared.

Data availability

Corpus data concerning the study are publicly available on Open Science Framework (https://osf.io/mu6vs/).

Funding

Funding support for this article was provided by the The Hong Kong Polytechnic University (P0046370, P0051009).

Note

1. $\mathbb{R}^G$ represents the part-of-speech space.

References

Baker, M. (1993) 'Corpus Linguistics and Translation Studies: Implications and Applications', in M. Baker, G. Francis and E. Tognini-Bonelli (eds) *Text and Technology. In Honor of John Sinclair*, pp. 233–50. Amsterdam, the Netherlands: John Benjamins.

Baker, M. (1996) 'Corpus-based Translation Studies: The Challenges that Lie Ahead', in J. C. Sager and H. L. Somers (eds) *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, pp. 44–54. Amsterdam, the Netherlands: John Benjamins.

Baroni, M., and 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**: 259–74.

Bentz, C., and Alikaniotis, D. (2016) 'The Word Entropy of Natural Languages', *arXiv preprint arXiv:1606.06996*.

Bentz, C. *et al.* (2017) 'The Entropy of Words—Learnability and Expressivity Across More than 1000 Languages', *Entropy*, **19**: 275.

Bernardini, S., and Baroni, M. (2005) 'Spotting Translationese. *A Corpus-Driven Approach Using Support Vector Machines*', In: *Proceedings of Corpus Linguistics Conference Series 2005*, Vol. 1, pp. 1-12. Birmingham: University of Birmingham.

Bishop, C. (2006) *Pattern Recognition and Machine Learning*, Vol **2**, pp. 531–7. Berlin, Heidelberg, Germany: Springer.

Blum-Kulka, S., and Levenston, E. (1983) 'Universals of Lexical Simplification', in C. Faerch and G. Kasper (eds) *Strategies in Interlanguage Communication*, 119–39. London: Longman.

Bosch, A., Zisserman, A., and Munoz, X. (2007) 'Image classification using random forests and ferns', *2007 IEEE 11th International Conference on Computer Vision*, pp. 1–8. Piscataway, NJ: IEEE.

Breiman, L. (1996) 'Bagging Predictors', *Machine Learning*, **24**: 123–40.

Breiman, L. (2001) 'Random Forests', *Machine Learning*, **45**: 5–32.

Chen, R., Liu, H., and Altmann, G. (2017) 'Entropy in Different Text Types', *Digital Scholarship in the Humanities*, **32**: 528–42.

Chesterman, A. (2004) 'Hypotheses about Translation Universals', *Benjamins Translation Library*, **50**: 1–14.

Delaere, I., De Sutter, G., and Plevoets, K. (2012) 'Is Translated Language More Standardized than Non-translated Language: Using Profile-based Correspondence Analysis for Measuring Linguistic Distances between Language Varieties', *Target. International Journal of Translation Studies*, **24**: 203–24.

De Clercq, O. *et al.* (2021) 'Uncovering Machine Translationese Using Corpus Analysis Techniques to Distinguish between Original and Machine-translated French', *Translation Quarterly*, 101: 21–45.

Dhar, V. (2013) 'Data Science and Prediction', *Communications of the ACM*, **56**: 64–73.

Ferraresi, A. *et al.* (2018) 'Simplified or not Simplified? The Different Guises of Mediated English at the European Parliament', *Meta*, **63**: 717–38.

Freund, Y., and Schapire, R. E. (1997) 'A Decision-theoretic Generalization of On-line Learning and an Application to Boosting', *Journal of Computer and System Sciences*, **55**: 119–39.

Friedman, J. H. (2001) 'Greedy Function Approximation: A Gradient Boosting Machine', *Annals of Statistics*, **29**: 1189–232.

Friedrich, R. (2021) 'Complexity and Entropy in Legal Language', *Frontiers in Physics*, **9**: 671882.

Friedrich, R., Luzzatto, M. and Ash, E. (2020) 'Entropy in legal language', *NLLP 2020 Natural Legal Language Processing Workshop 2020. Proceedings of the Natural Legal Language Processing Workshop 2020 Co-located with the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining (KDD 2020)*, Vol. 2645, pp. 25–30. CEUR-WS.

Gambier, Y. (2016) 'Translations| Rapid and Radical Changes in Translation and Translation Studies', *International Journal of Communication*, **10**: 887–906.

Genzel, D. and Charniak, E. (2002) 'Entropy Rate Constancy in Text', *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pp. 199–206. Pittsburgh, PA: Association for Computational Linguistics.

Geurts, P., Ernst, D., and Wehenkel, L. (2006) 'Extremely Randomized Trees', *Machine Learning*, **63**: 3–42.

Gile, D. (2009) *Basic Concepts and Models for Interpreter and Translator Training*. Amsterdam, the Netherlands: John Benjamins Publishing.

Hastie, T., Tibshirani, R., and Friedman, J. (2009) *The Elements of Statistical Learning: Data Mining, Inference, and Prediction*. Berlin, Heidelberg, Germany: Springer Science & Business Media.

Ho, T. K. (1998) 'The Random Subspace Method for Constructing Decision Forests', *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **20**: 832–44.

House, J. (2014) 'Translation Quality Assessment: Past and Present', in J. House (eds) *Translation: A Multidisciplinary Approach. Palgrave Advances in Language and Linguistics*, pp. 241–64. London: Palgrave Macmillan.

Hu, H., and Kübler, S. (2021) 'Investigating Translated Chinese and its Variants Using Machine Learning', *Natural Language Engineering*, **27**: 339–72.

Hu, H., Li, W., and Kübler, S. (2018) 'Detecting Syntactic Features of Translated Chinese'. *arXiv preprint arXiv*:1804.08756.

Hundt, M., Sand, A. and Siemund, R. (1998) *Manual of Information to Accompany the Freiburg-LOB Corpus of British English (FLOB)*. Freiburg: Albert-Ludwigs Universitat.

Ilisei, I. *et al.* (2010) 'Identification of Translationese: A Machine Learning Approach', *International Conference on Intelligent Text Processing and Computational Linguistics*, pp. 503–11. Berlin, Heidelberg, Germany: Springer.

Jantunen, J. (2004) 'Untypical Patterns in Translations. Issues on Corpus Methodology and Synonymity', in A. Mauranen and P. Kujamäki (eds) *Translation Universals: Do They Exist*, pp. 101–26. Amsterdam, the Netherlands: John Benjamins.

Joulin, A. *et al.* (2017) 'Bag of tricks for efficient text classification', *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics*, Vol. 2, *Short Papers*, pp. 427–31. Pittsburgh, PA: Association for Computational Linguistics

Juola, P. (2008) 'Assessing Linguistic Complexity', in M. Miestamo, K. Sinnemaki, and F. Karlsson (eds) *Language Complexity: Typology, Contact, Change, pp. 89-108*. Amsterdam, the Netherlands: John Benjamins Press.

Juola, P. (2013) 'Using the Google N-gram Corpus to Measure Cultural Complexity', *Literary Linguist Computing*, **28**: 668–75.

Kajzer-Wietrzny, M., Whyatt, B., and Stachowiak, K. (2016) 'Simplification in Inter-and Intralingual Translation–combining Corpus Linguistics, Key Logging and Eye-tracking', *Poznan Studies in Contemporary Linguistics*, **52**: 235–7.

Kockelman, P. (2009) 'The Complexity of Discourse', *Journal of Quantitative Linguistics*, **16**: 1–39.

Laviosa S. (1998) 'Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose', *Meta*, **43**: 557–70.

Laviosa, S. (2002) *Corpus-based Translation Studies: Theory, Findings, Applications*, Vol. **17**. Amsterdam, the Netherlands: Rodopi.

Levy, R., and Manning, C. D. (2003) 'Is it harder to parse Chinese, or the Chinese Treebank', *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pp. 439–46. Pittsburgh, PA: Association for Computational Linguistics.

Li, D., Zhang, C., and Liu, K. (2011) 'Translation Style and Ideology: A Corpus-assisted Analysis of Two English Translations of Hongloumeng', *Literary and Linguistic Computing*, **26**: 153–66.

Liu, K., and Afzaal, M. (2021) 'Syntactic Complexity in Translated and Non-translated Texts: A Corpus-based Study of Simplification', *PLoS One*, **16**: e0253454.

Liu, K., Liu, Z., and Lei, L. (2022) 'Simplification in Translated Chinese: An Entropy-based Approach', *Lingua*, **275**: 103364.

Liu, K. *et al.* (2022) 'Entropy-based Discrimination between Translated Chinese and Original Chinese Using Data Mining Techniques', *PLoS One*, **17**: e0265633.

Lowder, M. W. *et al.* (2018) 'Lexical Predictability during Natural Reading: Effects of Surprisal and Entropy Reduction', *Cognitive Science*, **42**: 1166–83.

Malmkjær, K. (1997) 'Punctuation in Hans Christian Andersen's Stories and their Translations into English', in F. Poyatos (ed.) *Nonverbal Communication and Translation: New Perspectives and Challenges in Literature, Interpretation and the Media*, pp. 151–62. Amsterdam, the Netherlands: John Benjamins.

Mauranen, A. (2006) 'Translation Universals', in K. Brown (ed.) *Encyclopedia of Language and Linguistics*, pp. 93–100. Amsterdam, the Netherlands: Elsevier.

McEnery, A., Xiao, Z., and Mo, L. (2003) 'Aspect Marking in English and Chinese: Using the Lancaster Corpus of Mandarin Chinese for Contrastive Language Study', *Literary and Linguistic Computing*, **18**: 361–78.

Mehri, A., and Darooneh, A. H. (2011) 'The Role of Entropy in Word Ranking', *Physica A: Statistical Mechanics and its Applications*, **390**: 3157–63.

Nisioi, S., and Dinu, L. P. (2013) 'A clustering approach for translationese identification', *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pp. 532–8. Shoumen, Bulgaria: INCOMA Ltd.

Olohan, M. (2004) *Introducing Corpora in Translation Studies*. London: Routledge.

Pastor, G. C. *et al.* (2008) 'Translation universals: do they exist? A corpus-based NLP study of convergence and simplification', *Proceedings of the 8th Conference of the Association for Machine Translation in the Americas: Research Papers*, pp. 75–81. Washington DC: Association for Machine Translation in the Americas

Pym, A. (2008) 'On Toury's Laws of How Translators Translate', in A. Pym, M. Shlesinger, and D. Simeoni (eds) *Beyond Descriptive Translation Studies: Investigations in Homage to Gideon Toury*, pp. 311–28. Amsterdam, the Netherlands: John Benjamins.

Pym, A. *et al.* (2012) *The Status of the Translation Profession in the European Union*. London: Anthem Press.

Qi, P. *et al.* (2018) Universal dependency parsing from scratch. *Proceedings of the CoNLL 2018 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pp.

160–70. Pittsburgh, PA: Association for Computational Linguistics.

Rabinovich, E., and Wintner, S. (2015) 'Unsupervised Identification of Translationese', *Transactions of the Association for Computational Linguistics*, **3**: 419–32.

Rubino, R., Lapshinova-Koltunski, E., and Van Genabith, J. (2016) 'Information density and quality estimation features as translationese indicators for human translation classification', *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 960–70. Pittsburgh, PA: Association for Computational Linguistics.

Schapire, R. E. (1990) 'The Strength of Weak Learnability', *Machine Learning*, **5**: 197–227.

Shannon, C. E. (1948) 'A Mathematical Theory of Communication', *The Bell System Technical Journal*, **27**: 379–423.

Shi, Y., and Lei, L. (2022) 'Lexical Richness and Text Length: An Entropy-based Perspective', *Journal of Quantitative Linguistics*, **29**: 62–79.

Statnikov, A., Wang, L., and Aliferis, C. F. (2008) 'A Comprehensive Comparison of Random Forests and Support Vector Machines for Microarray-based Cancer Classification', *BMC Bioinformatics*, **9**: 1–10.

Suo, J. J. *et al.* (2012) 'Study of Ambiguities of English-Chinese Machine Translation', *Applied Mechanics and Materials*, **157**: 472–5.

Tabrizi, H. H., and Pezeshki, M. (2015) 'Strategies Used in Translation of Scientific Texts to Cope with Lexical Gaps (Case of Biomass Gasification and Pyrolysis Book)', *Theory and Practice in Language Studies*, **5**: 1173.

Tanaka-Ishii, K. (2005) 'Entropy as an Indicator of Context Boundaries: An Experiment Using a Web Search Engine', *International Conference on Natural Language Processing*, pp. 93–105. Berlin, Heidelberg, Germany: Springer Berlin Heidelberg.

Toury, G. (1995) *Descriptive Translation Studies and Beyond*. Amsterdam, the Netherlands: John Benjamins.

van Ewijk, L., and Avrutin, S. (2016) 'Lexical Access in Nonfluent Aphasia: A Bit More on Reduced Processing', *Aphasiology*, **30**: 1264–82.

Vanderauwera, R. (1985) *Dutch Novels Translated into English: The Transformation of a 'Minority' Literature*. Amsterdam, the Netherlands: Rodopi.

Volansky, V., Ordan, N., and Wintner, S. (2015) 'On the Features of Translationese', *Digital Scholarship in the Humanities*, **30**: 98–118.

Wang, Z., Liu, K., and Moratto, R. (2023) 'A Corpus-based Study of Syntactic Complexity of Translated and Non-translated Chairman's Statements', *Translation & Interpreting*, **15**: 135–51.

Wang, Z., Liu, M., and Liu, K. (2024) 'Utilizing Machine Learning Techniques for Classifying Translated and Non-translated Corporate Annual Reports', *Applied Artificial Intelligence*, **38**: 234039.

Wei, Y. (2022) 'Entropy as a measurement of cognitive load in translation', *Proceedings of the 15th Biennial Conference of the Association for Machine Translation in the Americas (Workshop 1: Empirical Translation Process Research)*, pp. 75–86. Washington DC: Association for Machine Translation in the Americas.

Xiao R., and Dai G. (2014) 'Lexical and Grammatical Properties of Translational Chinese: Translation Universal Hypotheses Reevaluated from the Chinese Perspective', *Corpus Linguistics and Linguistic Theory*, **10**: 11–55.

Xiao, R., and Hu, X. (2015) *Corpus-based Studies of Translational Chinese in English-Chinese Translation*. Berlin Heidelberg, Germany: Springer.

Xiao, R., and Yue, M. (2009) 'Using Corpora in Translation Studies: The State of the Art', in P. Baker (ed.) *Contemporary Corpus Linguistics*, pp. 237–62. London: Continuum.

Xu, H., and Liu, K. (2023) 'Investigating Lexical Simplification', *Corpora in Interpreting Studies: East Asian Perspectives*, Vol. **197**. Abingdon: Taylor & Francis.

Xue, N. *et al.* (2005) 'The Penn Chinese Treebank: Phrase Structure Annotation of a Large Corpus', *Natural Language Engineering*, **11**: 207–38.

Yang, Z. *et al.* (2013) 'Keyword Extraction by Entropy Difference between the Intrinsic and Extrinsic Mode', *Physica A: Statistical Mechanics and its Applications*, **392**: 4523–31.

Zasiekin, S. (2016) 'Understanding Translation Universals', *Babel. Revue Internationale de la Traduction/International Journal of Translation*, **62**: 122–34.

Zhou, Z. H. (2012) *Ensemble Methods: Foundations and Algorithms*. Boca Raton, FL: CRC Press.

Zhu, H., and Lei, L. (2018) 'Is Modern English Becoming Less Inflectionally Diversified? Evidence from Entropy-based Algorithm', *Lingua*, **216**: 10–27.

Appendix

The syntactic components and tags (adapted from Xue et al. 2005).

Tag	Syntactic component
ADJP	Adjective phrase
ADVP	Adverbial phrase headed by AD (adverb)
CC	Coordinating conjunction
CD	Cardinal number
CLP	Classifier phrase
CP	Clause headed by C (complementizer)
DNP	Phrase formed by 'XP + DEG'
DP	Determiner phrase
DNP	Phrase formed by 'XP + DEG"
DVP	Phrase formed by 'XP + DEV'
FRAG	Fragment
FW	Foreign word
IN	Preposition or subordinating conjunction
INTJ	Interjection
IP	Simple clause headed by INFL
LCP	Phrase formed by 'XP + LC'
LST	List marker
M	Measure word
NN	Noun
NP	Noun phrase
NR	Proper noun
PP	Prepositional Phrase
PRN	Parenthetical
PU	Punctuation
QP	Quantifier phrase
S	Simple declarative clause
UCP	Unidentical coordination phrase
VCD	Coordinated verb compound
VCP	Verb compounds formed by VV + VC
VE	Verb
VNV	Verb compounds formed by A-not-A or A-one-A
VP	Verb phrase
VPT	Potential form V-de-R or V-bu-R
VRD	Verb resultative compound
VSB	Verb compounds formed by a modifier + a head

Digital Scholarship in the Humanities, 2024, 00, 1–17
https://doi.org/10.1093/llc/fqae030
Full Paper