Humanities & Social Sciences Communications

ARTICLE

https://doi.org/10.1057/s41599-025-05562-9 OPEN

Assessing lexical and syntactic simplification in translated English with entropy analysis

Zhongliang Wang[1], Andrew K. F. Cheung[2], Han Xu[2] & Kanglong Liu[2]✉

This study investigates the lexical and syntactic complexity of translated and non-translated native English texts, with a particular focus on evaluating the simplification hypothesis in translation. Drawing on entropy as a quantitative measure rooted in information theory, we conduct a comparative analysis of two balanced corpora, each consisting of 500 texts, with a focus on how translation status and text type interact to shape informational complexity. The findings challenge the widely held assumption of universal simplification in translated texts. Contrary to expectations, translated English texts exhibit greater lexical complexity, as evidenced by higher wordform entropy, compared to native English texts. However, no significant differences emerge between the two groups in terms of syntactic complexity, as measured by part-of-speech entropy. These results contribute to the ongoing debate on translation universals, highlighting the nuanced nature of simplification as a construct. The study underscores the need to account for factors such as source language influence, translator's native language, text genre, and translation direction in future research. Furthermore, the use of entropy provides a useful and consistent means of assessing text complexity, contributing to ongoing methodological developments in translation studies and related areas.

[1] Guangzhou Railway Polytechnic, Guangzhou, China. [2] Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Hong Kong SAR, China. ✉email: kl.liu@polyu.edu.hk

Introduction

Translation plays a pivotal role in bridging linguistic and cultural divides, facilitating the global exchange of ideas, knowledge, and values (House, 2015). As globalization accelerates, this role becomes increasingly central: the demand for skilled translation has expanded exponentially, with profound implications for geopolitics, economic systems, educational practices, and cultural production. In this context, translation is far more than a conduit for communication—it operates as a dynamic sociocultural practice that (re)constructs cultural identities, negotiates power dynamics in public discourse, and mediates policy formulation. Its multifaceted impact is evident across domains: it underpins the dissemination of scientific innovation (Olohan, 2007), enables multinational corporate operations (Wang et al., 2023), and ensures the cross-cultural transmission and preservation of literary heritage (Wu and Li, 2022a).

Despite its critical role in global communication, translation is frequently misrepresented as a superficial, mechanical word-for-word transfer between languages. In reality, it constitutes a dynamic and multifaceted process that entails negotiation across divergent linguistic systems, cultural norms, and sociocultural contextual constraints (Bassnett, 2013). Given its centrality to cross-cultural mediation, existing research on the nature of translated language (as distinct from non-translated language) is imperative for optimizing the efficacy of translation practices. Research in translation studies has thus focused on the systematic analysis of linguistic and discursive patterns in translated texts, particularly their implications for translation quality, communicative equivalence, and cross-cultural communication strategies (Laviosa, 2002; Wu and Li, 2022b; Huang and Li, 2023). Such investigations not only enhance our understanding of translation as a unique mode of communication but also provide valuable insights for refining translation theory and practice in an increasingly interconnected world. Scholars in the field have identified specific linguistic characteristics that consistently appear in translated texts across language pairs (Baker, 1993; Chou et al., 2023; Su and Liu, 2022; Su et al., 2023; Wang and Liu, 2024), as well as in the interpreting process (Li et al., 2022; Xu and Liu, 2023, 2024). These distinctive features, 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, p. 243). The concept of "translation universals" bears a close resemblance to that of "translationese" (Gellerstam, 1986). Originally employed as a somewhat pejorative term to denote unnatural or deficient language use, translationese has since been redefined as a neutral, descriptive concept within translation studies, to describe the linguistic features or "fingerprints" that are characteristic of translated texts (Toury, 1995). Translation universals have been the subject of extensive scholarly inquiry in the field. Baker (1993) systematically identified several such universals, including simplification, explicitation, and normalization, while Toury (1995) further advanced the theoretical framework by introducing the laws of standardization and interference. Teich (2003) conducted a corpus-based investigation into translation universals, notably exploring the phenomenon of "shining-through", where specific features of the source language remain visible in the target text, preserving the source language patterns. Unlike interference, "shining-through" is considered a systematic pattern rather than a translation error.

Translation universals are understood as inherent patterns and tendencies intrinsic to the translation process itself, independent of the linguistic systems of either the source or target languages (Fan and Jiang, 2019; Liu and Afzaal, 2021). Recent advancements in the application of digital humanities approaches to translation studies have provided new perspectives on these patterns, yielding insights that have significantly enriched both translation theory and practice (Gu, 2023; Sun and Li, 2020). This progress has prompted critical examinations of the intricacies of language transfer during translation, underscoring the complexity of the process. Consequently, there is a pressing need for further research into the linguistic complexities embedded within translated texts. Such investigations are essential to advancing our understanding of the translation process and, ultimately, to fostering more effective cross-cultural communication.

Extensive research on the distinctive characteristics of translated texts has primarily focused on four key translation universals: explicitation, simplification, normalization, and levelling-out. Simplification, defined as "the idea that translators subconsciously simplify the language, or message, or both" (Baker, 1996, p. 176), stands as a particularly debated hypothesis within the study of translation universals (Liu and Afzaal, 2021). Explicitation, on the other hand, refers to "an overall tendency to spell things out rather than leave them implicit" (Baker, 1996, p. 180). Normalization, also known as "conventionalization" (Mauranen, 2007), involves exaggerating "features of the target language to conform to its typical patterns" (Baker, 1996, p. 183). Consequently, translated texts exhibit a semblance of normalcy compared to non-translated native texts in the target language (Xiao, 2010). This normative influence is notably evident in the overuse of clichés, typical grammatical structures of the target language, genre-specific features, punctuation adjustments, and treatment of various dialects in dialogs within the source texts (Xiao, 2010, pp. 10–11). Levelling-out signifies translation's tendency to "steer a middle course between any two extremes, converging toward the centre" (Baker 1996: 184). Numerous studies have sought to investigate these universals (Liu and Afzaal, 2021; Wang et al., 2023; Wang et al., 2024c; Xiao and Dai, 2014); however, the findings remain inconclusive, and no definitive consensus has been reached regarding their existence.

As Liu and Afzaal (2021) observe, the majority of earlier studies have relied on isolated and simplistic linguistic features to substantiate the existence of translation universals in translated texts. For instance, average sentence length has often been employed as an indicator of simplification, as noted by Laviosa (2002). However, this measure has proven inconsistent across different language pairs, such as Spanish-English (Pym, 2008) and Chinese-English (Xiao and Yue, 2009), highlighting the limitations of relying on individual linguistic features for conclusive evidence. Additionally, early research on translation universals predominantly focused on literary texts (e.g., Blum-Kulka and Levenston, 1983; Laviosa, 1998). Yet, the genre of texts has since been identified as a critical variable influencing translational language features, necessitating its consideration in studies on translation universals (Delaere et al., 2012; Liu and Afzaal, 2021). To address these methodological challenges, researchers have increasingly turned to information-theoretic measures, such as entropy (Liu et al., 2022a; Lin and Liang, 2023; Wang et al., 2024a), and tree-based dependency measures (Xu and Liu, 2023, 2024) to quantify the complexity of translated texts. These approaches have yielded new insights, providing a more nuanced understanding of the translation process. The present study builds on this body of research by incorporating the concept of information entropy from information theory, utilizing balanced corpora across multiple genres to investigate whether simplification occurs in translated English. Entropy, which quantifies the complexity or uncertainty inherent in a linguistic system, serves as a robust analytical tool for examining the linguistic features of translated texts. By applying entropy as a measure, this study aims to offer a more precise understanding of the degree of simplification or complexification in translated English,

contributing a novel perspective on the linguistic characteristics of translational language. Furthermore, this study extends prior research (Liu et al., 2022a; Liu et al., 2022b) by exploring the application of computational methods to deepen our understanding of translation's impact on language.

Related work

Simplification in translation. Simplification is one of the most extensively studied translation universals in the field of translation studies. This process typically involves the use of simpler lexical, grammatical, and syntactic structures while maintaining the intended meaning. Prior to the advent of corpora and computational tools, text collection and analysis were primarily conducted manually, with research often limited to small samples of source and target texts. In this context, Blum-Kulka and Levenston (1983) investigated translations between Hebrew and English, with Hebrew as the source language and English as the target language. Their study identified lexical simplification in translations, evidenced by the use of fewer words to convey equivalent meanings. This early research highlighted the phenomenon of simplification in translation, establishing a foundation for more comprehensive studies enabled by modern computational methods and larger corpora. Vanderauwera (1985) expanded the investigation of simplification to include syntactic aspects in English translations of Dutch texts, observing that translated texts frequently replaced finite clauses with non-finite ones, thereby reducing syntactic complexity. Additionally, Vanderauwera (1985) examined simplification from a stylistic perspective, noting how translators modified stylistic elements to enhance the accessibility of the text. However, due to the restricted sample sizes and the absence of advanced statistical techniques, the findings of these early studies lacked generalizability.

With advancements in corpora and computational linguistics, quantitative approaches have increasingly been adopted to examine the manifestation of translation universals from a statistical perspective. The assessment of simplification has been conducted using a comparable corpus approach, which allows for comparisons between translated and original texts (Baker, 1993). Malmkjær (1997) found that translated English texts from Danish tend to employ stronger punctuation, such as replacing commas with semicolons or periods, and substituting complex syntactic structures in source texts with shorter, less complicated clauses in translated texts. Laviosa (1998) examined the linguistic features of English-translated narrative prose by analyzing criteria such as lexical density, average sentence length, and the frequency of commonly used words. By analyzing a section of the English Comparable Corpus (ECC), a monolingual, multi-source-language English corpus, Laviosa (1998) identified four key aspects in which translated English differs from native English. Firstly, translated texts exhibited lower lexical density and a higher proportion of grammatical terms compared to original texts. Secondly, they made greater use of high-frequency words. Thirdly, these high-frequency words were repeated more often in translated texts. Lastly, the most commonly used terms in translated texts included fewer lemmas than those found in original texts. Olohan (2004) employed lexical diversity as a metric to compare translated English fiction from German with native English fiction, revealing that translated works tended to use fewer synonyms for colors than their native counterparts. Similarly, Pastor et al. (2008) explored simplification using natural language processing tools, readability formulas, and other indices, finding that non-translated Spanish texts exhibited higher lexical density and richness compared to Spanish texts translated from English. Collectively, these studies provide empirical support for the simplification hypothesis. The adoption of quantitative methods and statistical analysis in such research has enabled a more systematic and rigorous examination of simplification in translation. By measuring diverse linguistic features and comparing translated texts with their native counterparts, researchers have identified consistent patterns of simplification, thereby advancing our understanding of translation universals.

While simplification has been extensively studied as a potential translation universal, it remains a contentious topic within translation studies. Some research findings challenge the traditional view of simplification, suggesting that the phenomenon is more nuanced than previously thought. Lexical complexity, often associated with vocabulary richness and diversity, has been extensively studied (Baker, 1996; Blum-Kulka and Levenston, 1983; Kruger, 2019; Laviosa, 2002). Early studies, such as Blum-Kulka and Levenston (1983), suggested that translated texts tend to exhibit simplified lexical choices in their comparison of Hebrew and English. However, more recent investigations have challenged this view. For instance, Kruger (2019) found that translated English texts from Afrikaans display higher lexical diversity than comparable non-translated English texts.

Syntactic complexity in translation has been examined through various approaches. Vanderauwera (1985) found reduced complex syntax in Dutch-English translations. This finding was supported by Pastor et al. (2008), who showed that sentences in translated Spanish texts were statistically shorter than those in non-translated Spanish texts. However, Wang et al. (2023) found that chairman's statements translated from Chinese into English exhibited higher syntactic complexity than non-translated English chairman's statements, as evidenced by longer production units, higher frequency of coordinate phrases, and more complex nominal structures.

The relationship between text type and complexity has also received considerable attention. Biber, Conrad (2019) demonstrated how different genres exhibit distinct linguistic patterns. In the field of translation studies, Liu and Afzaal (2021) showed that text type significantly influences syntactic complexity in English texts translated from Chinese. Translation status and its impact on text complexity have been investigated through various methodological approaches. Studies employing machine learning techniques have successfully distinguished between translated and non-translated texts based on their linguistic features (Baroni and Bernardini, 2006; Volansky et al., 2015; Wang et al., 2024a; Wang et al., 2024b). Additionally, multidimensional analyses have shown that translated texts often display distinct linguistic patterns compared to non-translated texts (Wang and Liu, 2024).

The exploration of translation universals, however, has largely been conducted from a Eurocentric perspective, as noted by Xiao and Dai (2014). This Eurocentric focus is reflected in the significant number of studies that have concentrated on European languages (De Camargo, 2016; Laviosa, 2002), particularly in their examination of lexical and syntactic features. The variations in these features within European languages may not be as pronounced as those observed in more linguistically diverse language pairs, such as English and Chinese, where the differences in vocabulary, grammar, and sentence structure are considerably more substantial (Xiao and Dai, 2014). The substantial linguistic disparities between English and Chinese pose unique challenges and opportunities for translators, underscoring the necessity for a nuanced analysis of translation universals that carefully considers these distinctions. The characteristics of translated texts often vary significantly across different language pairs, necessitating a broader and more inclusive perspective in translation studies. Notably, prior research comparing Chinese-English translations with original

English texts has identified compelling evidence of syntactic variations between translated and non-translated texts (Liu and Afzaal, 2021). These variations are shaped not only by linguistic factors but also by cultural, social, and contextual elements that influence the translation process. The current lack of a coherent picture in research on translation universals can, in part, be attributed to the selective use of linguistic indicators chosen to support specific hypotheses (Liu et al., 2022a).

To address this gap, we draw on the methodologies proposed by Liu et al. (2022a) and Lin and Liang (2023), employing entropy as a robust quantitative measure of linguistic complexity. Entropy, introduced by Shannon (1948), was developed to quantify information and assess the information content within a given source. It serves as a metric to measure the degree of uncertainty or randomness in a specific dataset. Essentially, entropy provides a numerical representation of the average information or "surprise" associated with an event or observation. The calculation involves taking the logarithms of the probabilities assigned to each potential outcome or message in a random event, with these probabilities serving as weights. Outcomes with higher probabilities contribute less to the overall entropy compared to those with lower probabilities. Shannon's formula for calculating entropy is as follows:

$$\mathrm{H} = -\sum_{i=1}^{n} P_i \log_2 P_i$$

In the provided formula, the total entropy denoted by H represents the aggregate measure of entropy for all elements present in a message. P_i within the formula represents the probability of a specific element occurring, which can be computed using its relative frequency within the dataset. The value of n corresponds to the total count of elements within the message. The formula allows for the calculation of entropy based on the probabilities associated with different elements, providing a quantitative assessment of the overall uncertainty or information content in the given dataset. Entropy, widely used in fields such as ecology, computational linguistics, and information science (Ben-Naim, 2019; Bentz et al., 2017; Cushman, 2021), also offers a systematic and objective method for analyzing linguistic phenomena. Among the diverse quantitative approaches available for analyzing translation simplification, entropy-based measures are particularly advantageous due to their capacity to simultaneously capture both lexical and syntactic complexity through the quantification of linguistic elements' predictability and distribution patterns (Liu et al., 2022a). While traditional analytical methods often focus on isolated features such as lexical density, sentence length, or type-token ratio, entropy measures offer a comprehensive assessment of text complexity by integrating information content, structural organization, and the probabilistic nature of language patterns (Wang et al., 2024a). This methodological advancement toward sophisticated quantitative analysis, specifically through the application of entropy measures, enables researchers to precisely quantify the degree of simplification in translated texts. The shift from potentially subjective qualitative evaluations to mathematically grounded metrics not only enhances the reliability and replicability of findings but also facilitates cross-linguistic comparisons in translation research.

Entropy in language and translation research. The concept of entropy has been extensively applied in language research, highlighting its versatility and significant contributions to linguistic analysis. Genzel and Charniak (2002) introduced a foundational principle of language generation known as the entropy-rate constancy principle. Their research demonstrated that, when sentences are analyzed out of context, sentence entropy increases with sentence number, a finding that aligns with this principle. This discovery provided new insights into linguistic patterns associated with entropy, particularly in written English. Specifically, Genzel and Charniak focused on the entropy of text, conceptualizing each word as a random variable whose distribution depends on all preceding words. They observed that the average entropy of these variables remains constant throughout a text. Building on these findings, Tanaka-Ishii (2005) expanded the understanding of entropy in language by showing that the uncertainty of tokens following a sequence is crucial for determining context boundaries. This study assessed the uncertainty of successive tokens using the concept of branching entropy, thereby underscoring the importance of entropy in shaping linguistic contexts. Juola (2008) further applied entropy to quantify linguistic complexity at the levels of lexicon, morphology, and syntax, demonstrating its utility across diverse linguistic dimensions. Mehri and Darooneh (2011) explored entropy's role in word ranking, emphasizing its systematic application in text mining and its relevance to understanding word patterns. More recently, Yang et al. (2013) advanced this line of research by introducing a new entropy-based metric for evaluating word relevance. They found that significant words closely reflect the author's intent, distinguishing them from irrelevant, randomly distributed words. Thus, analyzing word distribution is pivotal for understanding text complexity and for discerning how meaningful words are shaped by an author's objectives.

In the field of speech therapy, entropy has been applied to evaluate word complexity, with a particular focus on individual verbs and verb paradigms (van Ewijk and Avrutin, 2016). These researchers employed inflectional entropy as a measure, framing the informational complexity of a word as a combination of the information inherent in the target word itself and the information embedded within its morphological paradigms. Bentz et al. (2017) made significant contributions by exploring convergence points in word entropy, enabling quantitative language comparisons and enhancing translation systems, thereby underscoring entropy's fundamental role in comparative linguistic studies. Lowder et al. (2018) applied entropy-reduction techniques to enhance lexical prediction, extending its applications to predictive text analysis. They used entropy reduction as a complexity metric to quantify the amount of information gained with each word. The concept posits that if entropy decreases from one word to the next, communicative uncertainty is reduced, indicating active information processing by the reader. Friedrich et al. (2020) introduced a neural language model to measure word ambiguity in legal texts, finding lower entropy in the German Federal Court of Justice corpus due to the frequent use of technical language, thereby emphasizing entropy's relevance in legal language analysis. Friedrich (2021) continued this exploration, investigating various linguistic features of legal language and highlighting its high complexity and low entropy, which present challenges for non-experts in understanding specialized language domains. The consistent application of entropy across this wide spectrum of research underscores its critical importance and broad implications in language analysis and computational linguistics.

Entropy has gained considerable attention within the field of translation research due to its demonstrated relevance to cognitive and linguistic processes. Carl and Schaeffer (2017) proposed a noisy channel model for the translation process, demonstrating how entropy measurements can reveal cognitive processing patterns during translation. Their work demonstrated that information-theoretic measures like entropy are proportional to cognitive processing effort, particularly when combined with other metrics. Further developing this line of research, Carl (2021) investigated the first universal translational response, finding that the cognitive effort required for initial translation responses correlates with information content and cross-linguistic

similarity. His study revealed that typing pauses, which indicate cognitive processing time, are influenced by both information content of the source text and the degree of literal translation possible between language pairs. Wei (2022) introduced two important metrics, surprisal (ITra) and entropy (HTra), which approximate cognitive load. ITra was identified as a more accurate predictor of translation production time, while HTra was found to be more effective in predicting the reading time of the source text. More recently, Deilen et al. (2023) examined cognitive aspects of compound translation, using entropy measurements alongside other metrics to quantify cognitive effort. Their research provides empirical evidence that Information-theoretic Translation (ITra) measures, when combined with other indicators, can effectively predict cognitive processing load during translation tasks. These findings highlight the utility of entropy-based indicators in accounting for cognitive processes underlying translation activities.

Chen et al. (2017) explored the linguistic profiles of different text types using entropy. By examining word and part-of-speech entropy, as well as the entropy of aspect markers in Chinese and English corpora, they revealed distinct distribution patterns in both languages. Notably, Chinese exhibited higher entropy in aspect markers, underscoring grammatical differences compared to English and aiding in distinguishing between narrative and expository texts. Expanding the application of entropy in translation studies, Yerkebulan et al. (2021) developed an entropy-based methodology to detect patterns in multilingual texts and identify translations. Their innovative approach, which involved calculating text proximity using centers of parametric means, proved effective in distinguishing genuine translations from "pseudo" ones based on distance measures. This underscores the potential of entropy in translation verification and analysis. Recent studies by Liu et al. (2022a) and Liu et al. (2022b) further demonstrate the effectiveness of entropy-based approaches in examining translation universals, particularly in Chinese texts translated from English. These researchers suggest that further exploration of entropy-based methods could yield valuable insights when applied to other translated languages beyond Chinese.

Entropy has been employed as an inverse indicator of simplification: higher entropy values indicate lower levels of simplification and increased linguistic complexity, while lower values denote higher simplification and reduced complexity (Liu et al., 2022a; Liu et al., 2022b; Wang et al., 2024a). When translated texts exhibit lower entropy values compared to non-translated texts, this supports the simplification hypothesis, indicating less diverse lexical choices and more predictable language structures. Conversely, higher entropy values suggest greater lexical diversity and less predictable patterns, pointing to a lower degree of simplification.

Research questions

The use of entropy-based metrics to analyze translational simplification has gained traction, as evidenced by Liu et al. (2022a), who demonstrated their effectiveness in revealing translation phenomena in Chinese texts. Building on this work, our study extends the investigation to translated English, aiming to provide comparative data that enhances our understanding of translational simplification across different languages. We employ entropy-based metrics to quantitatively assess the complexity and simplification of texts, aiming to identify patterns of simplification across various linguistic structures. This research incorporates data from typologically diverse language pairs, such as English and Chinese, which Xiao and Dai (2014) suggest could lead to robust and insightful results. Our comparative analysis seeks to determine whether translated English, derived from Chinese sources, exhibits simplification compared to native English in four distinct genres. This approach advances our understanding of cross-linguistic complexity patterns and reveals how genre-specific features are preserved or simplified during the Chinese-to-English translation process. The following research questions will guide our investigation:

RQ1: Are there significant differences in the lexical and syntactic complexity of texts based on translation status?

RQ2: Are there significant differences in the lexical and syntactic complexity of texts based on text type?

RQ3: Is there an interaction effect between translation status and text type on the lexical and syntactic complexity of texts?

Materials and methods

Corpora. The study utilizes two primary corpora: the Freiburg-LOB (LOB stands for Lancaster-Oslo/Bergen corpus) Corpus of British English (FLOB) (Hundt et al., 1998) and the English component of the Corpus of Chinese into English (COCE). FLOB, established in the early 1990s, aimed to create a parallel corpus to the original LOB and Brown corpora with one million words reflecting the diversity of written British English from that period. It includes a variety of genres, such as press reportage, editorials, religion, hobbies, popular lore, biographies, and academic prose, all from 1991 to provide a dataset comparable to the original LOB Corpus from 1961 (Hundt et al., 1998). COCE, designed to closely match FLOB in terms of size and genre components, serves as a parallel corpus containing Chinese source texts and their English translations, aligned at the sentence level for enhanced representativeness (Liu and Afzaal, 2021). The focus of our research is on the English translations within COCE, which comprises 500 texts averaging 2,000 words each. These texts cover four main genres and 15 subgenres, offering a broad spectrum for analysis. Further details about the English component of COCE, including genre distribution, text types, and token counts, are summarized in Table 1. This corpus setup allows for a rigorous comparative analysis of translational simplification between native and translated English across distinct genres. It is worth noting that all translators in this study were native Chinese speakers with advanced English proficiency, rather than native English speakers. While this approach is common practice in China, it differs from many translation studies in which translators are typically native speakers of the target language. This characteristic of our translator pool represents a methodological limitation that should be considered when interpreting our results, as non-native English translators may process and produce translations differently from native speakers. Their linguistic backgrounds might influence the translation outcomes in ways that native English translators' work would not.

Following Baker's (1993) comparable corpus approach, the present study compares the English translations in COCE with the original English texts in FLOB. This comparative analysis is instrumental in comprehensively examining the degree of simplification utilizing the entropy-based approach.

Calculation of wordform entropy and POS entropy. Previous research by Shi and Lei (2020) has shown that text length significantly affects text entropy, underscoring the importance of maintaining consistent text lengths for precise analysis. Although initial estimates suggested that each text in our corpora should average 2000 words, a detailed examination revealed discrepancies in text lengths between FLOB and COCE. To address this, we standardized text lengths across both corpora, aiming for uniformity in our analysis. In line with the methodology proposed by Liu et al. (2022a), we set a maximum word count of 1500 words per text, excluding punctuation. This adjustment was crucial to preserve the original structure of the corpora, which comprises 500 texts.

Table 1 Details of the english part of COCE.

Genres	Text types	Number of texts	Tokens	Average tokens
News	News (reportage, editorial, review)	88	183,101	2081
General prose	Religious writing	17	34,968	2057
	Skills, trades, and hobbies	36	78,702	2186
	Popular lore	48	91,134	1899
	Essays and biography	75	161,455	2153
	Miscellaneous	30	61,616	2054
Academic writing	Academic prose	80	164,704	2059
Fiction	General fiction	29	59,815	2063
	Mystery and detective stories	24	49,566	2065
	Science fiction	6	12,380	2063
	Adventure fiction	29	58,869	2030
	Romantic fiction	29	59,834	2063
	Humor	9	18,825	2092
Total		500	1,034,969	26,864

Where texts exceeded this limit, they were truncated to meet the new word count criterion. This approach ensures that our entropy measurements are not skewed by variations in text length, facilitating a more reliable comparison of translational simplification between the corpora. For POS tagging, we employed the Penn Treebank POS tagger implemented through Natural Language Toolkit (NLTK). Subsequently, we computed wordform entropy and POS entropy values for both FLOB and COCE using a Python program. It should be noted that during this computation, all punctuation marks were omitted to eliminate potential confounding effects caused by variations in punctuation usage between the corpora. This standardized approach to text length and preprocessing ensures a rigorous and comparative analysis of word and POS entropy across the two corpora.

The calculation of wordform entropy for a text is outlined as follows:

1. Count the occurrences of each word: This is achieved by tabulating the frequency of all words in the corpus.
2. Calculate the probability of each word: The probability of a word, $P(w_i)$ is determined by dividing the frequency of the word by the total number of words in the corpus.

$$p(w_i) = \frac{\text{Frequency of} w_i}{\text{Total number of words in the text}}$$

3. Calculate entropy using the Shannon entropy formula:

$$\mathrm{H(W)} = -\sum_{i=1}^{n} p(w_i)\log_2 p(w_i)$$

where H(W) is the entropy of the word distribution, $P(w_i)$ is the probability of word i, and n is the number of unique words.

As for the POS entropy of a text, its calculation is presented below:

1. Tag each word with its part of speech: Use a POS tagger to label each word in the corpus with its corresponding part of speech. Punctuation is removed prior to this step.
2. Count the occurrences of each POS tag.
3. Calculate the probability of each POS tag: The probability of a POS tag, $p(t_j)$, is computed as the frequency of the tag divided by the total number of POS tags in the text.

$$p\left(t_j\right) = \frac{\text{Frequency of} t_j}{\text{Total number of POS tags in the text}}$$

4. Calculate POS entropy:

$$\mathrm{H(T)} = -\sum_{j=1}^{n} p\left(t_j\right)\log_2 p\left(t_j\right)$$

where H(T) is the entropy of the POS distribution, $P(t_j)$ is the probability of POS tag j, and n is the number of unique POS tags.

The use of wordform entropy and Part of Speech (POS) entropy in evaluating textual complexity has been well established in prior research (Liu et al., 2022b; Shi and Lei, 2020). Wordform entropy, calculated from the distribution of each word's occurrences within a text, quantifies the unpredictability or randomness of word usage. High wordform entropy signifies a text that utilizes a broad vocabulary in a relatively even distribution, suggesting linguistic richness and a higher level of lexical sophistication (Liu et al., 2022a). Such texts may be more challenging to comprehend due to the wide range of vocabulary required. Conversely, low wordform entropy, indicating repetitive use of certain words, denotes simpler texts that are often more suitable for beginner readers or contexts where clarity and repetition are valued. Therefore, wordform entropy can effectively capture both the depth and utilization of a text's vocabulary, reflecting its lexical complexity, which influences reader comprehension and engagement (Liu et al., 2022b; Shi and Lei, 2020). Similarly, POS entropy assesses the diversity of grammatical structures by analyzing the distribution of parts of speech within a text (Liu et al., 2022a). Utilizing a probability-based entropy formula, it measures the uncertainty in the choice of grammatical categories. High POS entropy reflects the use of a variety of grammatical constructs, indicating complex sentence formations that contribute to syntactic richness. This complexity can enhance the cognitive load for readers, as intricate grammatical structures often involve variations in sentence length, clause embedding, or innovative uses of language, characteristics typical of advanced texts. Conversely, lower POS entropy indicates a predominance of specific parts of speech, leading to uniform and potentially simpler syntactic structures. Such simplicity may limit the expressive breadth and syntactic diversity of the language. Taken together, these measures provide a nuanced view of textual complexity by integrating both lexical and grammatical dimensions, which is essential for analyzing language use across different types of texts.

Data analysis. To address the research questions, two two-way ANOVA tests were conducted. The first analysis assessed the influence of translation status and genre on lexical complexity, which served as the dependent variable. Specifically, translation status was categorized into translated and non-translated texts, and genre was divided into four categories: press, general prose, academic prose, and fiction. This ANOVA aimed to identify any main effects of translation status and genre, as well as any interaction between these factors on lexical complexity, quantified through wordform entropies. Following the identification of an

interaction effect, a Tukey post hoc test was performed to pinpoint specific differences between the groups.

Similarly, the second two-way ANOVA explored the effects of translation status and genre on syntactic complexity, which served as the dependent variable in this analysis. Translation status was again categorized into translated and non-translated texts, while genre included press, general prose, academic writing, and fiction. This test aimed to determine whether translation status and genre independently or interactively influenced syntactic complexity, measured by POS entropies. Like the first test, upon detecting significant interaction effects, a Tukey post hoc analysis was conducted to elucidate specific differences between the groups. This methodical approach ensures a thorough examination of how translation status and genre may shape syntactic structures in various text types.

Results

The descriptive statistics for wordform entropy and POS entropy across the two corpora can be observed in Tables 2 and 3, and Fig. 1. We calculated the wordform entropy and POS for the four genres, encompassing 15 text types, within both FLOB and COCE. The findings reveal that native English texts (FLOB) consistently exhibit lower average wordform entropy values compared to translated texts (COCE) across all four genres. Conversely, translated texts display lower mean POS entropy values than non-translated texts in general prose, academic

Table 2 Descriptive statistics for wordform entropy of FLOB and COCE.

Corpus	Genre	Mean	Std. deviation	Min	Max	N
FLOB	News	8.227	0.124	7.948	8.586	88
	Prose	7.948	0.243	7.076	8.552	206
	Academic	7.809	0.213	7.237	8.201	80
	Fiction	7.952	0.165	7.512	8.417	126
	Total	7.976	0.239	7.076	8.586	500
COCE	News	8.399	0.113	8.067	8.6	88
	Prose	7.949	0.291	6.897	8.453	206
	Academic	7.873	0.238	7.205	8.39	80
	Fiction	8.149	0.182	7.6	8.659	126
	Total	8.067	0.295	6.897	8.659	500
Total	News	8.313	0.147	7.948	8.6	176
	Prose	7.949	0.268	6.897	8.552	412
	Academic	7.841	0.227	7.205	8.39	160
	Fiction	8.05	0.199	7.512	8.659	252
	Total	8.021	0.272	6.897	8.659	1000

Table 3 Descriptive statistics for POS entropy of FLOB and COCE.

Corpus	Genre	Mean	Std. deviation	Min	Max	N
FLOB	News	3.974	0.087	3.804	4.18	88
	Prose	3.93	0.108	3.581	4.192	206
	Academic	3.859	0.121	3.533	4.175	80
	Fiction	4.016	0.072	3.779	4.155	126
	Total	3.948	0.111	3.533	4.192	500
COCE	News	3.98	0.079	3.777	4.105	88
	Prose	3.92	0.152	3.383	4.215	206
	Academic	3.845	0.121	3.349	4.099	80
	Fiction	3.986	0.071	3.789	4.164	126
	Total	3.935	0.128	3.349	4.215	500
Total	News	3.977	0.083	3.777	4.18	176
	Prose	3.925	0.131	3.383	4.215	412
	Academic	3.852	0.121	3.349	4.175	160
	Fiction	4.001	0.073	3.779	4.164	252
	Total	3.942	0.120	3.349	4.215	1000

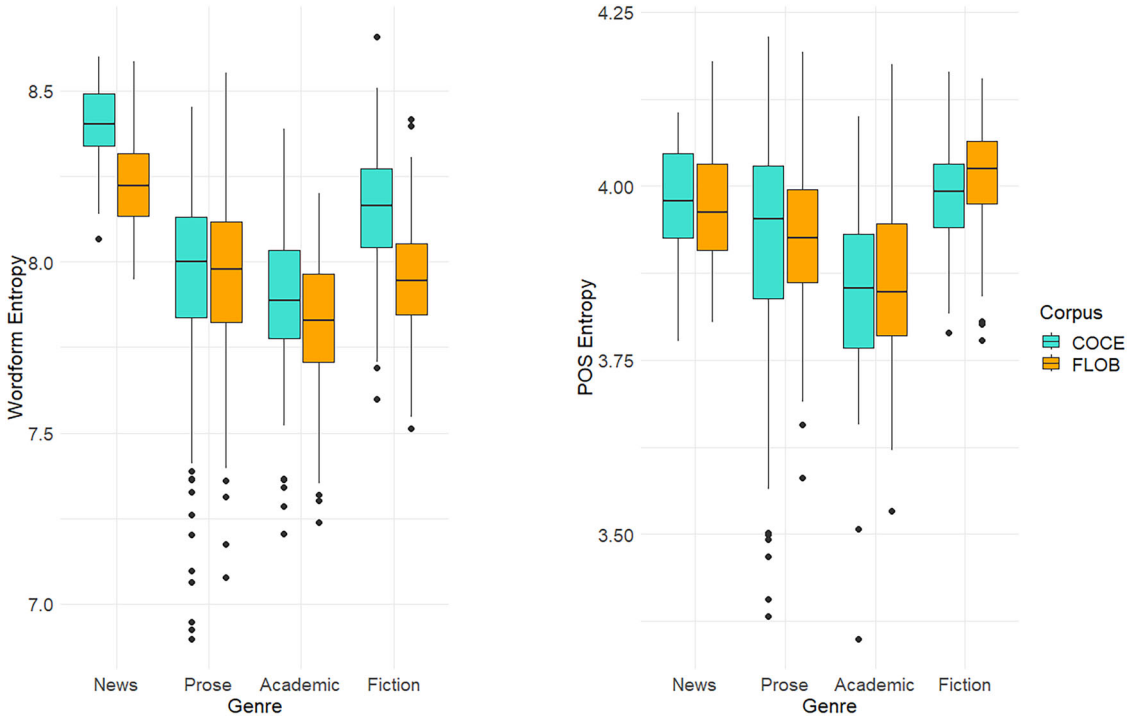

Fig. 1 Boxplots of wordform entropy and POS entropy of FLOB and COCE. Boxplots comparing wordform entropy (left) and POS entropy (right) across four genres (News, Prose, Academic, Fiction) in the FLOB and COCE corpora. COCE is represented in turquoise, and FLOB in orange. Medians are indicated by lines within the boxes, with outliers shown as dots.

writing, and fiction, while registering higher mean POS entropy values in news.

To analyze the statistical differences in wordform entropy between the two corpora, a two-way ANOVA was conducted, with corpus and genre as the independent variables. The results revealed a significant main effect of corpus, indicating a statistically significant difference in wordform entropy between the two corpora ($F(1, 992) = 42.91.8$; $p < 0.001$). Specifically, the FLOB corpus exhibited lower wordform entropy (Mean = 7.976) compared to the COCE corpus (Mean = 8.067). Furthermore, a main effect of genre was observed, indicating significant differences in wordform entropy across distinct genres ($F(3, 992) = 156.90$; $p < 0.001$). The interaction between corpus and genre was also statistically significant ($F(3, 992) = 12.84$; $p < 0.001$), indicating notable variations in wordform entropy values across diverse genres within the two corpora (see Fig. 1). Additionally, as shown in Table 2, the genre of news had the highest mean wordform entropy (8.313), followed by fiction (8.05), and then general prose (7.949) and academic writing (7.841).

Table 4 Tukey comparisons of wordform entropy.

Pair	Mean difference	Lower bound	Upper bound	Sig.
Prose-news	−0.364	−0.415	−0.314	<0.001**
Academic-news	−0.472	−0.534	−0.410	<0.001**
Fiction-news	−0.263	−0.318	−0.208	<0.001**
Academic-prose	−0.108	−0.160	−0.055	<0.001**
Fiction-prose	0.102	0.056	0.147	<0.001**
Fiction-academic	0.209	0.152	0.266	<0.001**
COCE:News-FLOB:News	0.172	0.072	0.273	<0.001**
COCE:Prose-FLOB:Prose	0.001	−0.064	0.067	1
COCE:Academic-FLOB:Academic	0.064	−0.041	0.169	0.584
COCE:Fiction-FLOB:Fiction	0.196	0.113	0.280	<0.001**

*$p < 0.05$, **$p < 0.001$.

As the ANOVA test indicated a significant interaction, a Tukey's test was conducted to further explore the influence of corpus (FLOB, COCE) and genre (news, academic prose, fiction) on wordform entropy. The objective was to analyze significant variations within the same genre across the two corpora (see Table 4). The results demonstrated statistically significant differences across all four genres. When comparing FLOB and COCE, significant differences were observed in the genres of news and fiction ($p < 0.001$), while no significant difference was found in the genres of academic writing and general prose between the two corpora ($p > 0.05$). Figure 2 illustrates that COCE exhibits significantly higher wordform entropy in the genres of news and fiction. However, this difference is not statistically significant in the genres of general prose and academic writing.

A similar two-way ANOVA was performed to explore the statistical differences in POS entropy between the two corpora. A main effect of genre was evident, denoting significant disparities in POS entropy across various genres ($F(3, 992) = 69.314$; $p < 0.001$). As indicated in Table 3, the genre of fiction exhibited the highest mean POS entropy (4.001), followed by news (3.977), and then general prose (3.925) and academic writing (3.852). However, there was no significant main effect of corpus ($F(1, 992) = 3.493$; $p = 0.0619$) or a significant interaction between corpus and genre ($F(3, 992) = 0.967$; $p = 0.4074$) (see Fig. 2). The Tukey's test further revealed significant differences between news and general prose, academic writing and news, academic writing and general prose, fiction and prose, fiction and academic writing

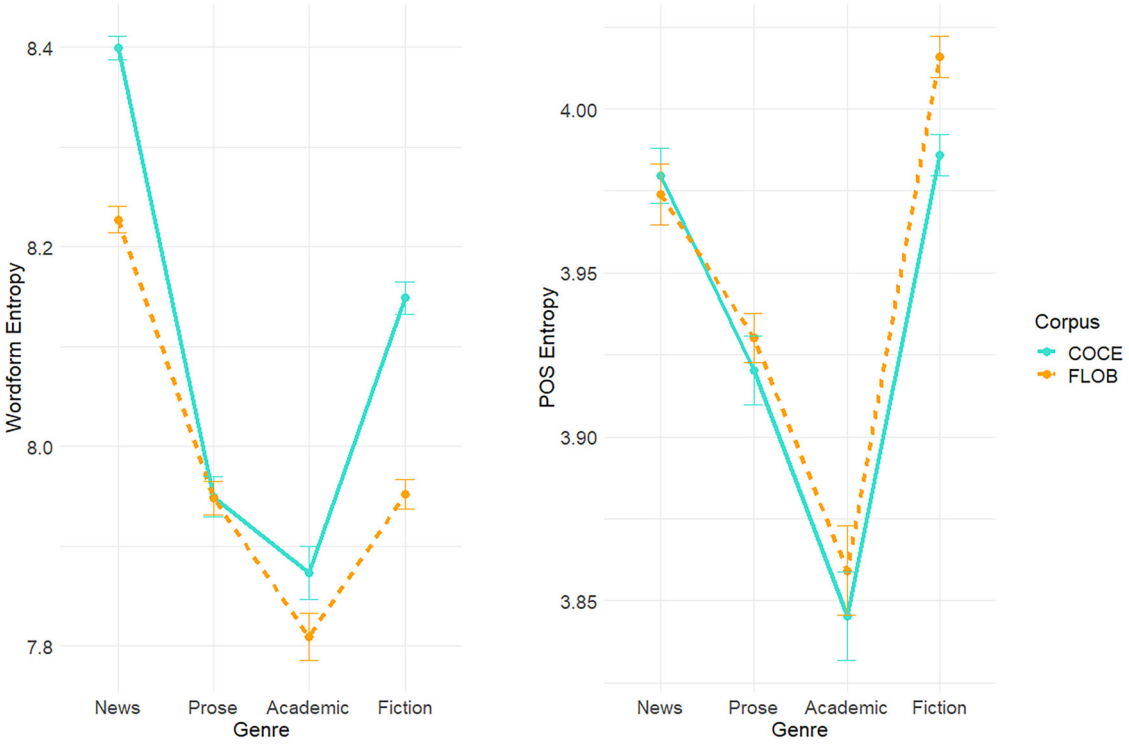

Fig. 2 Interaction between corpus and genre for wordform entropy and POS entropy. Interaction plots showing the relationship between corpus (COCE vs. FLOB) and genre (News, Prose, Academic, Fiction) for wordform entropy (left) and POS entropy (right). Solid lines represent COCE, and dashed lines represent FLOB. Error bars indicate standard error of the mean.

($p < 0.005$), as illustrated in Table 5. However, no significant difference was observed between fiction and news ($p = 0.113$).

Discussion

In this study, we employed an analysis utilizing information entropy, following the comparable corpus approach outlined by Baker (1993), to compare translated English with native English. Our research reveals significant distinctions in lexical complexity, gauged through wordform entropy, between translated and native texts. Notably, translated texts exhibited higher lexical complexity than their native counterparts, contrary to initial expectations. Interestingly, this heightened complexity was not reflected in the syntactic structures, where we found no notable differences. These findings stand in contrast to Liu et al. (2022a) investigation of translated Chinese versus native Chinese, which employed entropy measures and revealed a tendency for translated Chinese to be simpler than its native equivalent. Furthermore, our outcomes diverge from earlier research by Laviosa (1998) and Olohan (2004), both of which identified lexical simplification as a characteristic feature of translated texts.

In light of these unexpected findings, we conducted a case study aimed at explaining why the wordform entropy of the COCE corpus surpasses that of the FLOB corpus. To ensure a fair comparison, we randomly selected one text from each corpus, both from the news genre. Our analysis revealed an overall wordform entropy of 8.01 for the FLOB sample and 8.51 for the COCE sample. We proceeded to calculate the entropy contribution of each word in these two texts, subsequently creating a descending list of word entropy contribution rankings. Figure 3 visually represents this data, illustrating the word entropy contribution of each sample from the FLOB and COCE corpora via points and lines. Interestingly, the COCE sample contains more unique words than its FLOB counterpart, with counts of 765 and 673 words, respectively. Notably, while the highest-ranking word in the FLOB sample contributes more to entropy than the top word in the COCE sample, the COCE sample compensates by incorporating an additional 92 unique words. Each new word increases the system's overall entropy, suggesting that these extra words have elevated the COCE sample's entropy beyond that of the FLOB sample. Our findings appear to challenge the simplification hypothesis in translation, which posits that translated texts should exhibit lower complexity than non-translated ones. Rather than demonstrating a reduced vocabulary range, our translated texts (COCE) show higher lexical complexity, indicated by higher wordform entropy. According to Baker (1996), translation often involves explicitation, where translators use a more extensive vocabulary or additional phrasing to convey meaning more precisely. This tendency may be particularly pronounced in Chinese-to-English translation, overriding simplification and resulting in higher entropy values that reflect greater complexity.

Source texts can significantly influence subsequent translations, leading to potential divergences between translated and non-translated texts (Toury, 1995). Teich (2003) posits that the impact of source texts is a key determinant in shaping target texts, setting them apart from content originally crafted in the target language. This phenomenon, termed "source text shining through," occurs when the translated text mirrors the structural and stylistic properties of the source language more closely than the target language's norms (ibid.). For instance, this effect is particularly noticeable in the corpus of COCE, where translated texts seem to be more "source language-oriented" (Teich, 2003). Chinese, known for its complex syntactical structures and extensive vocabulary, tends to have higher word entropy than English (Chen et al., 2017). Consequently, when translations are made from Chinese to English, they often exhibit higher word entropy values than texts originally written in English. This is a manifestation of the "source text shining through" effect as the

Table 5 Tukey comparisons of POS entropy.

Pair	Mean difference	Lower bound	Upper bound	Sig.
Prose-news	−0.051	−0.077	−0.026	<0.001**
Academic-news	−0.124	−0.155	−0.094	<0.001**
Fiction-news	0.024	−0.004	0.052	0.113
Academic-prose	−0.073	−0.099	−0.047	<0.001**
Fiction-prose	0.076	0.053	0.098	<0.001**
Fiction-academic	0.149	0.120	0.177	<0.001**

*$p < 0.05$, **$p < 0.001$.

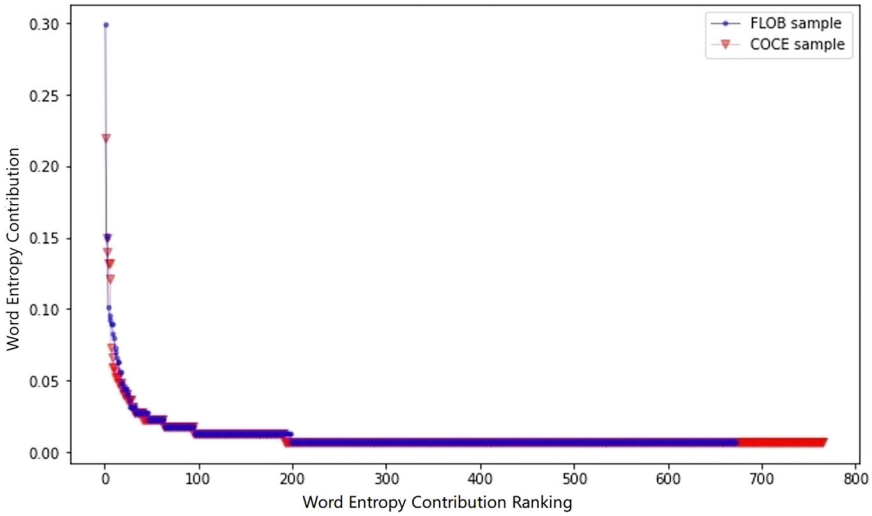

Fig. 3 Word entropy contribution of FLOB and COCE sample. Word entropy contribution for a selected FLOB and COCE sample from the news genre. The plot shows the ranking of word entropy contributions, with the FLOB sample (in blue) and the COCE sample (in red) displayed via points and lines. The COCE sample includes more unique words (765) compared to the FLOB sample (673), which results in a higher overall wordform entropy.

complexities of the source language (in this case, Chinese) leave their mark on the target language (English) (Xu and Liu, 2024).

Previous studies have demonstrated that textual complexity correlates with cognitive effort required for both the creation and comprehension of text (Fan and Jiang, 2019; Liang et al., 2017). This framework helps explain the higher word entropy found in the COCE corpus, which consists of English translations by native Chinese speakers (Liu and Afzaal, 2021). Such translations show increased entropy due to the complex cognitive processes involved in translating from Chinese (L1) into English (L2), a task that is inherently more challenging than translating into a native language.

POS entropy, as detailed by Liu et al. (2022a), provides insights into the syntactic variability and complexity of languages. Their study found no significant difference in POS entropy between translated and native English texts, indicating that the syntactic structures of translated texts are closely aligned with those of native English compositions. This similarity is particularly notable in the COCE corpus, which comprises translations by advanced Chinese speakers. These translators, who are proficient in English, adhere to translation norms and standards that ensure syntactic consistency with the target language, despite their deep understanding of Chinese linguistic structures. The COCE corpus is unique because it contains published translations by Chinese speakers, highlighting the advanced language skills of its translators. These individuals are capable of producing translations that meet high linguistic standards, facilitated by their professional proficiency and the rigorous editing processes typical of published works. While POS entropy is a valuable metric for understanding variability and predictability in part-of-speech usage, it does not encompass all aspects of syntactic complexity, such as the hierarchical relationships among parts of speech. To further explore syntactic complexity in translated and non-translated texts, future research should incorporate more nuanced syntactic analyses. Techniques such as entropy-based syntactic tree analysis (Wang et al., 2024a) and studies on dependency distance (Liu et al., 2017; Xu and Liu, 2023, 2024) could provide a more comprehensive examination of syntactic structures, enhancing our understanding of translational language.

Translation extends beyond the internal cognitive processes of the translator's mind; it is also a sociocultural endeavor that serves as a conduit between cultures (House, 2014). This activity involves navigating both linguistic and cultural nuances to create texts that are faithful to the source language while resonating with the target language context (Jakobson, 1959). A key theoretical framework that offers insights into this intricate process is the Hypothesis of Gravitational Pull (Halverson, 2003), which combines elements of bilingual theory and cognitive linguistics. This hypothesis posits that translation is influenced by three interconnected dynamics. The first dynamic, the magnetism effect, pulls translators toward adopting the dominant linguistic features of the target language. The second dynamic, the gravitational pull of the source language, counteracts this by urging translators to retain aspects of the source language in their work. The third dynamic, the connection effect, emerges from the frequent co-occurrence of translation equivalents across the source and target languages, facilitating a natural translational flow. These dynamics interact to determine the composition of the final translated text, striking a distinctive balance between adherence to the source and fluency in the target language (Halverson, 2017). The findings of this study suggest that the gravitational pull effect is particularly pronounced at the lexical level in translations from Chinese to English. This effect results in a richer and more varied vocabulary in the translated texts, likely because the translators are working from their native language (L1) into their second language (L2). This orientation enhances the influence of the source language, encouraging the preservation of its linguistic features in the translations—a phenomenon known as source-language shining through—thereby enriching the target language text with complex lexical items.

The research utilizes observable data to illustrate that entropy is an effective tool for analyzing variations between translated and untranslated texts. Empirical evidence is crucial in research as it provides tangible examples to support theoretical frameworks. Unlike traditional methods that may focus on specific linguistic elements, entropy offers a broader perspective on textual complexity. Traditional linguistic analysis often relies on subjective interpretation, whereas entropy assigns a numeric value, facilitating more systematic comparisons across texts. This contributes to a more objective and consistent analytical process. By incorporating principles from information theory, which addresses the quantification, storage, and communication of information, the research is grounded in a solid theoretical base. This ensures that the findings are both objective and repeatable. Overall, this study highlights the benefits of using a mathematically robust and comprehensive tool like entropy to explore differences between translated and untranslated texts, potentially leading to more reliable and insightful outcomes in translation studies.

Our entropy-based analysis of lexical and syntactic complexity in translated English texts offers several insights for translation practice. The finding that translated texts exhibit higher lexical complexity challenges the traditional assumption that translators tend to simplify their vocabulary choices (Kruger and Rooy, 2012). We also observed that the syntactic complexity of translated texts was comparable to that found in native texts, though further research is needed to confirm this pattern across different genres. Moreover, variations in text types may necessitate different approaches to managing complexity levels. Finally, entropy measurements could serve as a helpful tool for establishing complexity benchmarks and promoting consistency in large-scale translation projects.

Conclusion

The present study employs an entropy-based approach to analyze lexical and syntactic simplification in translated texts, offering practical insights into translation universals. Using entropy-based indicators broadens our understanding of translational language compared to traditional analyses that focus on individual linguistic features. However, some limitations should be noted. The findings primarily apply to Chinese-English translations and may not necessarily extend to other language pairs. Furthermore, the study uses translations produced by Chinese translators, which could subtly influence the results. Future research could replicate this study with native English translators to determine whether the patterns observed here are related to the translators' linguistic background. Such comparative work would help clarify the role of this variable in shaping translation universals and linguistic patterns in translated texts. Finally, while the corpora used are robust, they do not guarantee perfect temporal alignment, which may introduce some variability in the findings. These limitations suggest areas for further research to refine and expand upon the current study's conclusions.

Data availability

All relevant data of this study are available at https://osf.io/z6nh5/.

Received: 22 November 2023; Accepted: 15 July 2025;
Published online: 31 July 2025

References

Baker M (1993) "Corpus linguistics and translation studies: Implications and applications." In: Baker M, Francis G, Tognini-Bonelli E (eds) Text and technology. In honor of John Sinclair, Amsterdam/Philadelphia, John Benjamins, pp 233–250

Baker M (1996) "Corpus-based translation studies: the challenges that lie ahead." In: Sager JC, Somers HL (eds) Terminology, LSP and translation: studies in language engineering in honour of Juan C. Sager, Amsterdam, John Benjamins, pp 44–54

Baroni M, Bernardini S (2006) A new approach to the study of translationese: machine-learning the difference between original and translated text. Lit Linguis Comput 21(3):259–274

Bassnett S (2013) Translation studies. 4th edn. Routledge, Taylor & Francis group, pp 208

Ben-Naim A (2019) Entropy and information theory: uses and misuses. Entropy 21(12):1170

Bentz C, Alikaniotis D, Cysouw M, Ferrer-i-Cancho R (2017) The entropy of words—Learnability and expressivity across more than 1000 languages. Entropy 19(6):275. https://doi.org/10.3390/e19060275

Biber D, Conrad S (2019) Register, genre, and style. Cambridge University Press

Blum-Kulka S, Levenston E (1983) "Universals of lexical simplification." In: Faerch C, Kasper G (eds) Strategies in interlanguage communication, London, Longman, pp 119–139

Carl M (2021) "Micro Units and the First Translational Response Universal." In: Carl M (ed) Explorations in empirical translation process research. Cham, Springer, pp 233–257

Carl M, Schaeffer M (2017) Sketch of a noisy channel model for the translation process. Empir Model Trans. Interpret. 7:71

Chen R, Liu H, Altmann G (2017) Entropy in different text types. Digit. Sch. Humanit. 32(3):528–542. https://doi.org/10.1093/llc/fqw008

Chou I, Li W, Liu K (2023) Representation of interactional metadiscourse in translated and native English: a corpus-assisted study. Plos one 18(7):e0284849

Cushman SA (2021) Entropy in landscape ecology: a quantitative textual multivariate review. Entropy 23(11):1425

De Camargo DC (2016) Language of translation and interculturality for a corpus-based translation pedagogy. Signata Ann Semiot 7:155–173. https://doi.org/10.4000/signata.1191

Deilen S, Lapshinova-Koltunski E, Carl M (2023) Cognitive aspects of compound translation: Insights into the relation between implicitation and cognitive effort from a translation process perspective. Ampersand 11:100156. https://doi.org/10.1016/j.amper.2023.100156

Delaere I, De Sutter G, Plevoets K (2012) Is translated language more standardized than non-translated language?: Using profile-based correspondence analysis for measuring linguistic distances between language varieties. Target Int J Transl Stud 24(2):203–224. https://doi.org/10.1075/target.24.2.01del

Fan L, Jiang Y (2019) Can dependency distance and direction be used to differentiate translational language from native language? Lingua 224:51–59

Friedrich R (2021) Complexity and entropy in legal language. Front Phys 9:671882. https://doi.org/10.3389/fphy.2021.671882

Friedrich R, Luzzatto M, Ash E (2020) Entropy in legal language. In: 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 and Data Mining (KDD 2020), vol 2645. CEUR-WS Organization, San Diego, p 25–30

Gellerstam M (1986) Translationese in Swedish novels translated from English. Transl Stud Scand 1:88–95

Genzel D, Charniak E (2002) Entropy rate constancy in text. In: 40th annual meeting of the Association for Computational Linguistics. Association for Computational Linguistics, Philadelphia, Pennsylvania, pp 199–206

Gu C (2023) One-third of a century on: the state of the art, pitfalls, and the way ahead relating to digital humanities approaches to translation and interpreting studies. Digit Schol Humanit 39:154–161

Halverson SL (2003) The cognitive basis of translation universals. Target Int J Transl Stud 15(2):197–241. https://doi.org/10.1075/target.15.2.02hal

Halverson SL (2017) "Gravitational pull in translation: Testing a revised model." In: De Sutter G, Lefer M, Delaere I (eds) Empirical translation studies: New methodological and theoretical traditions. Berlin, Mouton de Gruyter, p 9–46

House J (2014) "Translation quality assessment: past and present." In: House J. (ed) Translation: a multidisciplinary approach. palgrave advances in language and linguistics. London, Palgrave Macmillan, p 241–264

House J (2015) Translation as communication across languages and cultures. Routledge

Huang Y, Li D (2023) Translatorial voice through modal stance: a corpus-based study of modality shifts in Chinese-to-English translation of research article abstracts. Lingua 295:103610

Hundt M, Sand A, Siemund R (1998) Manual of information to accompany the Freiburg-LOB Corpus of British English (FLOB). Freiburg, Albert-Ludwigs Universitat

Jakobson R (1959). On linguistic aspects of translation. In: Reuben AB (ed) On Translation. Cambridge, Harvard University Press, p 232–239

Juola P (2008) "Assessing linguistic complexity." In: Miestamo M, Sinnemaki K, Karlsson F (eds) Language complexity: typology, contact, change. Amsterdam, John Benjamins Press, pp 89–108

Kruger H (2019) That again: a multivariate analysis of the factors conditioning syntactic explicitness in translated English. Across Lang Cult 20(1):1–33

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

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

Laviosa S (2002) Corpus-based translation studies: theory, findings, applications (Vol. 17). Rodopi

Li R, Cheung AK, Liu K (2022) A corpus-based investigation of extra-textual, connective, and emphasizing additions in English-Chinese conference interpreting. Front Psychol 13:847735

Liang J, Fang Y, Lv Q, Liu H (2017) Dependency distance differences across interpreting types: implications for cognitive demand. Front Psychol 8:2132

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

Liu H, Xu C, Liang J (2017) Dependency distance: a new perspective on syntactic patterns in natural languages. Phys Life Rev 21:171–193

Liu K, Afzaal M (2021) Syntactic complexity in translated and non-translated texts: a corpus-based study of simplification. PLoS ONE 16(6):e0253454. https://doi.org/10.1371/journal.pone.0253454

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

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

Lowder MW, Choi W, Ferreira F, Henderson JM (2018) Lexical predictability during natural reading: Effects of surprisal and entropy reduction. Cogn Sci 42:1166–1183. https://doi.org/10.1111/cogs.12597

Malmkjær K (1997) "Punctuation in Hans Christian Andersen's stories and their translations into English." In: Poyatos F (ed) Nonverbal Communication and Translation: New Perspectives and Challenges in Literature, Interpretation and the Media Amsterdam/Philadelphia, John Benjamins, p 151–162

Mauranen A (2007) "Investigating English as a lingua franca with a spoken corpus." In: eds. Campoy MC, Luzón MJ (eds) Spoken corpora in applied linguistics. Berlin, Peter Lang, pp 33–56

Mehri A, Darooneh AH (2011) The role of entropy in word ranking. Phys A Stat Mech Appl 390:3157–3163. https://doi.org/10.1016/j.physa.2011.04.013

Olohan M (2004) Introducing corpora in translation studies. London and New York, Routledge

Olohan M (2007) The status of scientific translation. J Transl Stud 10(1):131–144

Pastor GC, Mitkov R, Afzal N, Pekar V (2008) Translation universals: do they exist? A corpus-based NLP study of convergence and simplification. In: 8th conference of the Association for Machine Translation in the Americas: research papers, pp 75–81

Pym A (2008) "On Toury's laws of how translators translate." In: Pym A, Shlesinger M, Simeoni D (eds) Beyond descriptive translation studies: investigations in homage to Gideon Toury. Amsterdam, John Benjamins, pp 311–328

Shannon CE (1948) A mathematical theory of communication. Bell Syst Tech J 27(3):379–423. https://doi.org/10.1145/584091.584093

Shi Y, Lei L (2020) Lexical richness and text length: an entropy-based perspective. J Quant Linguist 29(1):62–79. https://doi.org/10.1080/09296174.2020.1766346

Su Y, Liu K (2022) Orality in translated and non-translated fictional dialogues. In: Advances in corpus applications in literary and translation studies, Routledge, pp 119–137

Su Y, Liu K, Cheung AK (2023) Epistemic modality in translated and non-translated English court judgments of Hong Kong: a corpus-based study. J Spec Transl 40:56–80

Sun Y, Li D (2020) Digital humanities approaches to literary translation. Comp Lit Stud 57(4):640–654

Tanaka-Ishii K (2005) Entropy as an indicator of context boundaries: an experiment using a web search engine. In: International Conference on Natural Language Processing. Berlin, Heidelberg, Springer Berlin Heidelberg, pp 93–105

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

Toury G (1995) Descriptive translation studies and beyond. John Benjamins, Amsterdam

van Ewijk L, Avrutin S (2016) Lexical access in nonfluent aphasia: a bit more on reduced processing. Aphasiology 30(11):1264–1282. https://doi.org/10.1080/02687038.2015.1135867
Vanderauwera R (1985) Dutch novels translated into English: the transformation of a "minority" literature. Amsterdam, Rodopi, pp 170
Volansky V, Ordan N, Wintner S (2015) On the features of translationese. Digit Sch Humanit 30(1):98–118
Wang Z, Liu K (2024) Linguistic variations between translated and non-translated English chairman's statements in corporate annual reports: a multi-dimensional analysis. SAGE Open 14(2):21582440241249349
Wang Z, Liu K, Moratto R (2023) A corpus-based study of syntactic complexity of translated and non-translated chairman's statements. Int J Transl Interpret Res 15(1):135–151
Wang Z, Cheung AK, Liu K (2024a) Entropy-based syntactic tree analysis for text classification: a novel approach to distinguishing between original and translated Chinese texts. Digit Sch Humanit 39(3):984–1000
Wang Z, Liu M, Liu K (2024b) Utilizing machine learning techniques for classifying translated and non-translated corporate annual reports. Appl Artif Intell 38(1):2340393
Wang Z, Xu H, Liu K (2024c) Lexical complexity in corporate communication: a corpus-based study of translated and non-translated chairman's statements. Corpus-Based Stud Across Humanit 2(2):265–284
Wei Y (2022) Entropy as a measurement of cognitive load in translation. In Proceedings of the 15th biennial conference of the Association for Machine Translation in the Americas (Workshop 1: Empirical Translation Process Research), Association for Machine Translation in the Americas, pp 75–86
Wu K, Li D (2022a) Are translated Chinese Wuxia fiction and Western heroic literature similar? A stylometric analysis based on stylistic panoramas. Digit Sch Humanit 37(4):1376–1393
Wu K, and Li D (2022b) "The reception of the English translations of Hongloumeng". In: Moratto R, Liu K, Chao DK, eds. Dream of the red chamber: literary and translation perspectives. London, Routledge, Taylor and Francis Group, pp 225–242
Xiao R, Dai G (2014) Lexical and grammatical properties of translational Chinese: translation universal hypotheses reevaluated from the Chinese perspective. Corpus Linguist Theory 10(1):11–55. https://doi.org/10.1515/cllt-2013-0016
Xiao R (2010) How different is translated Chinese from native Chinese?: a corpus-based study of translation universals. Int J Corpus Linguist 15(1):5–35. https://doi.org/10.1075/ijcl.15.1.01xia
Xiao R Yue M (2009) "Using corpora in translation studies: The state of the art." In: Baker P (ed) Contemporary Corpus Linguistics. London, Continuum, pp 237–262
Xu H, Liu K (2023) Syntactic simplification in interpreted English: dependency distance and direction measures. Lingua 294:103607
Xu H, Liu K (2024) The impact of directionality on interpreters' syntactic processing: insights from syntactic dependency relation measures. Lingua 308:103778
Yang Z, Lei J, Fan K, Lai Y (2013) Keyword extraction by entropy difference between the intrinsic and extrinsic mode. Phys A Stat Mech Appl 392(19):4523–4531. https://doi.org/10.1016/j.physa.2013.05.052
Yerkebulan G, Kulikova V, Kulikov V, Kulsharipova Z (2021) Devising an entropy-based approach for identifying patterns in multilingual texts. East Eur J Enterp Technol 2(2):110. https://doi.org/10.15587/1729-4061.2021.228695

Acknowledgements

This research was funded by the Guangdong Provincial Philosophy and Social Sciences Planning (GD25YWY19), Tertiary Education Scientific Research Project of Guangzhou Municipal Education Bureau (2024312550), and Guangzhou Railway Polytechnic Newly Introduced Talent Research Project (GTXYR2426).

Author contributions

Conceptualization: Andrew K.F. Cheung, Kanglong Liu, Supervision: Han Xu, Kanglong Liu, Data collection: Zhongliang Wang, Formal analysis: Zhongliang Wang, Kanglong Liu, Visualization: Zhongliang Wang, Validation: Han Xu, Andrew K.F. Cheung, Kanglong Liu, Writing—Original draft: Zhongliang Wang, Writing—Revising and editing: Han Xu, Andrew K.F. Cheung, Kanglong Liu, Project administration: Kanglong Liu.

Competing interests

The authors declare no competing interests.

Ethical approval

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

Informed consent

This article does not contain any studies with human participants performed by any of the authors. Informed consent is thus not applicable in the context of our specific study.

Additional information

Correspondence and requests for materials should be addressed to Kanglong Liu.

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

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