https://dx.doi.org/10.1016/j.lingua.2022.103364

Simplification in translated Chinese: An entropy-based approach

Kanglong Liu[a], Zhongzhu Liu[b], Lei Lei[c]

[a] *Department of Chinese and Bilingual Studies, The Hong Kong Polytechnic University, Kowloon, Hong Kong, China*
[b] *School of Mathematics and Statistics, Huizhou University, Yanda Road, Huizhou, Guangdong 516007, China*
[c] *School of Foreign Languages, Shanghai Jiao Tong University, Shanghai 200240, China*

Abstract

For a long time, translation researchers, particularly those working in corpus-based translation studies, have held the presumption that translated texts tend to be simpler in lexical and syntactical features than non-translated native texts. Such claims have led to the formulation of the simplification universal hypothesis in translation studies. However, this line of research which focuses predominantly on the investigation of individual linguistic features has failed to provide sufficient evidence to confirm the existence of the simplification universal. To a large extent, the lack of global quantitative indicators for evaluating the complexity level of the translated and non-translated texts has hindered progress in this field. The current study, using entropy as an indicator, analysed the linguistic complexity between translated and native Chinese from the information-theoretical perspective. Our research found that translational Chinese tends to be simpler than its non-translated counterpart at the lexical level based on unigram entropy, but not the syntactic level based on part-of-speech entropy. Our study has confirmed the use of entropy as a reliable measure for lexical and syntactic complexity in the field of translation studies.

Keywords: Translation; Entropy; Word forms; POS forms; Computational linguistics

1. INTRODUCTION

1.1. Background

In the contemporary context of globalization, the role of translation in facilitating cultural communication can never be underestimated. Translation has played a crucial role in human and intercultural communication throughout history. Without translation, communication as we know it would not exist. On the other hand, the notion that translation is categorically different from original texts has been rooted in both the academic as well as professional world. Researchers have noticed that certain linguistic features typically occur in translated rather than original texts and are independent of the influence of the specific language pairs involved in the process of translation (Baker, 1993). The basic assumption of these researchers is that translation is fundamentally a mediating process and thus has its own unique features which might come under the interference through the recodification process (Toury, 1995). For example, Frawley (1984) believes that translation is essentially a "third code" which emerges from the bilateral consideration of the source and target codes (i.e., source and target languages). Over the years, various terms have been proposed to denote such linguistic features, such as "translation universals" (Baker, 1993), "laws of translation" (Toury, 1995) and "mediation universals" (Ulrych and Murphy, 2008). Despite the subtle nuances between the terms and different aspects of the same phenomenon each term emphasizes, the terminological variety clearly demonstrates that the investigation of the translational features has attracted a lot of attention from translation researchers in the field.

It is generally acknowledged that Baker (1993) is a pioneer in translation research who envisaged and proposed a new agenda for studying the linguistic features of translational languages, i.e., translation universals (TUs). Specifically, Baker (1993) contended that translated texts can be studied vis-à-vis the native non-translated texts using a comparable corpus to provide a better understanding of the peculiarities of translational language. Such a proposal has been innovative as research is no longer confined to a comparison of translated texts with their correspondent source texts. This method of comparing translation and comparable non-translation of the target language, is what Chesterman (2004) calls the T-universals which mainly aims at characterizing how translators use the target language.

Over the past two decades, the study of TUs has gained momentum in corpus-based Descriptive Translation Studies despite the controversies and debates surrounding the term and its possible existence. A number of researchers (House, 2015; Tymoczko, 1998) argue that translation-inherent universals is a pseudo-concept and the quest for TUs is in essence futile. Toury (2004) proposed that instead of making universal claims of translation, the search for general and observable regularities, i.e., translation laws, should be the fundamental task of descriptive translation studies. To a large extent, the central controversy related to the TUs research is on the use of the term "universal" (Chesterman, 2014) to

describe general norms and tendencies in translation. Nonetheless, corpus-based TUs research has proved its value on merit of its new research methods and agenda. As is argued by Chesterman, "the quest for universals is no more than the usual search for patterns and generalizations that guides empirical research in general" (Chesterman, 2014:87). Over the years, researchers have made great efforts in studying these unique features of translational language, including simplification which refers to the tendency of translation to simplify language use compared to native writing (Laviosa, 2002), explicitation which refers to the tendency of spelling out the information more explicitly in translation than native writing (Chen, 2004; Olohan and Baker, 2000), normalization (translation tends to overuse linguistic features typical of the target language) (Bernardini and Ferraresi, 2011) and levelling out (the tendency that translation is more homogeneous than native writing) (Cappelle, 2012). Despite the dearth of research on TUs, little consensus has been reached as to the existence of TUs. In recent years, researchers have become more vigilant of the limitations in this line of research. For example, Kruger and van Rooy (2012) specially mentioned language pairs and text type as two major limitations in TUs research. Firstly, most research is confined to the European context and based on European languages. The majority of TUs research has been done on European languages, where the linguistic patterning may not be as distinct as in typologically divergent languages like English and Chinese (Xiao and Dai, 2014). Second, earlier TUs research following Baker's (1993) proposal was mainly based on the literary text type, e.g., the Translational English Corpus held at the Centre for Translation Studies at UMIST (University of Manchester Institute of Science and Technology). In corpus-based TUs research, genre is an important variable affecting the profiling of translational language (Baker, 1999; Delaere et al. 2012; Kruger and van Rooy, 2012). More recent research has attached importance to such a variable. For example, Delaere et al. (2012) in their study on how text type (i.e., fiction, non-fiction, journalistic texts, instructive texts, administrative texts and external communication) and the translation status (i.e., translated and non-translated Belgian Dutch) serve as independent variables to affect the standard and nonstandard use of translational language, found that both translation status and text type have played a part in affecting the makeup of the translational language. Although more studies on European languages have paid attention to the variable of text type, it is still largely ignored in the Chinese context. Another limitation of TUs research that is closely relevant to this study is the use of linguistic features. A large amount of research tended to resort to the use of particular language features and prove their existence in translated texts to confirm certain TUs candidate. For example, the use of shorter sentence length has been treated as a simplification feature in Laviosa (2002), but has been found to be an unreliable indicator in many language pairs including Spanish-English (Pym 2008) and Chinese-English (Xiao and Yue 2009). This tells us that the use of individual language features might result in conflicting and ambivalent findings. Our research aims to complement such studies by using information entropy borrowed from information theory to identify if translation indeed demonstrates a simplification trend. So far, few researchers have advanced this line of enquiry using a computational linguistic perspective.

1.2. Simplification studies

Simplification is one of the most frequently tested TUs candidates in the quest for translational features. According to Baker (1996), simplification occurs when translators subconsciously simplify the language or message or both. In response to Baker's proposal, researchers have devoted much effort to the investigation of this translational feature. Chesterman (2004:39) distinguished translation universals into S-Universals, which "claim to capture universal differences between translations and their source texts, i.e. characteristics of the way in which translators process the source text" and T-Universals, which "make claims about universal differences between translations and comparable nontranslated texts, i.e. characteristics of the way translators use the target language". Of the four TUs (i.e., explicitation, simplification, normalisation and levelling-out) proposed by Baker (1996), explicitation is listed as a potential S-Universal while simplification serves as a typical T-Universal. As is pointed out by Delaere et al. (2012:237), simplification as a TUniversal "has most frequently, but not exclusively, been investigated with reference to non-translated texts". In other words, simplification is often investigated using a comparable corpus consisting of translated texts and non-translated native texts.

Interestingly, most of the existing literature has confirmed the existence of translational simplification which is found primarily at the lexical level. As mentioned in the foregoing review, there is no consensus as to what constitutes simplification and the linguistic parameters in connection with simplification tend to be decided by individual researchers. Over the years, lexical simplification has been defined as "making do with less words" (Blum-Kulka and Levenston, 1983); using informal, colloquial and modern lexis to translate formal, literate and archaic words (Vanderauwera, 1985) and lower type-token ratio in the translated texts (Cvrček and Lucie, 2015; Feng et al., 2018); In her pioneering research, Laviosa (2002) using the TEC (Translation English Corpus) and a comparable corpus of the same genre, found that lexical simplification existed in translation, evidenced by a limited range of vocabulary and lower lexical density used in the translated than non-translated texts. In her research, range of vocabulary was examined by studying high frequency words to low-frequency words ratio, proportion of the most frequent words (frequency list head) and the lemmas of the list head. Lexical density, on the other hand, is calculated by measuring the proportion of content words to grammatical words. Although there has been considerable research on pinpointing language features as potential indicators of simplification, there is no coherent evidence to support the existence of simplification. A plethora of existing studies have

identified linguistic features that contravene the simplification hypothesis including greater mean sentence length (Laviosa, 1998), untypical collocations (Mauranen, 2006) and more frequent use of modifiers (Jantunen, 2004). In the same vein, Xiao and Yue (2009), based on a corpus of translated Chinese fiction and native Chinese fiction, also found that translated texts have significantly higher average sentence length than non-translated ones. Their research shows that mean sentence length is not a reliable indicator for predicting the simplification in the translated texts. This is in line with the research by Xiao (2010) who found that translated Chinese has lower lexical density than native Chinese but shows no significant difference from native Chinese in mean sentence length, and the latter is believed to be genre sensitive. In a more recent research, Fan and Jiang (2019) used mean dependency distances (MDD) and dependency direction borrowed from quantitative linguistics as parameters to examine translated English texts against native texts in the same language. They found that the MDD of translated texts is much longer than non-translated English texts. Their research represents a shift from the previous ones which use a limited number of linguistic features and shows that it is possible to study translational language using mathematical and computational linguistic methods. One of the reasons why TUs research fails to yield a coherent picture is the use of linguistic metrics that are often randomly selected to prove the existence of certain TUs candidate. To avoid such an issue, more quantitative indicators that can minimize subjectivity should be used in this regard. For this reason, we adopted entropy, which has long been used as a quantitative indicator for measuring linguistic complexity, to investigate simplification in translational language. We believe that such a measure, which has been successfully used in various fields including ecology, communication, computational linguistics, can provide an alternative way for us to investigate the simplification phenomena in translation studies.

2. SHANNON'S ENTROPY AND COMPLEXITY RESEARCH

2.1. Defining Shannon's entropy

The concept of entropy was first introduced by Shannon (1948) to solve the problem of information quantification and compute the amount of information in the information source. In essence, Shannon's entropy measure is a weighted sum of the logs of the probabilities of each possible type in a random event or message. The weights used in the sum are derived from the probabilities of all the types, i.e., the types with high probabilities contribute less to the overall entropy of a message than types with low probabilities.

Shannon's entropy of a text H is calculated using the Formula (1) as follows:

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

where H is the total entropy of all the elements in the message, P_i is the probability of occurrence of a certain element (approached by its relative frequency) and n is the total number of the elements.

Shannon's entropy has been successfully applied to information theory (Shannon, 1948, 1951) and is now widely used in linguistics to study a wide array of topics including diversity of culture (Juola, 2008, 2013; Kockelman, 2009; Liu, 2016), authorship identification (Khmelev, 2000), loss of inflections (Zhu and Lei, 2018), the effects of text language on the metrics of word-length distributions and correlations (Kalimeri et al., 2015) and text complexity of different languages (Takahira et al., 2016). Shannon's entropy is based on the assumption that a more complex text contains more information and therefore requires more effort and time for readers to process. Traditionally, linguists often use type token ratio (TTR) to measure text richness. However, TTR measure only takes into consideration the number of distinct words while ignoring the frequency and distribution of a word in relation to other words. In this regard, the entropy measure will address both aspects as the distribution of words can make a difference in the final complexity values.

We provide a concrete example to show the differences between TTR and entropy measures in Table 1. As can be seen, the first row has a TTR of 0.2 while its information value is zero. As all data points have the same value, so they don't carry any meaningful information from an information-theoretical perspective. In other words, information value is determined by the degree of predictability.

It can be noted that all five data examples in Table 1 contain five letters (N = 5), meaning they have the same textual length. The difference between TTR and entropy is best reflected in the second and third data examples. The TTR for both yields a value of 0.4, as the number of types and tokens are the same. However, the entropy helps to distinguish the different informational value presented by A and B. In the following, we will explain how we calculated the entropy value of data 2 and 3 in Table 1. Based on Formula (1), H is the total entropy value of all the letters in the text, n is the types of letters (in the cases of data 2 and 3, n = 2), P_i is the probability of occurrence of the i-th letter within the text, $\log_2 P_i$ is the self-information of the i-th letter, and P_i ($-\log_2 P_i$) calculated the individual entropy of the i-th letter.

Data 2 AAAAB contains two letter types A and B, wherein A can be calculated using the formula (4/5)X(-log2(4/5)) = 0.25752, B can be calculated using the formula (1/5)X(-log2(1/5)) = 0.4644, thus the total entropy value of AAAAB is 0.25752 + 0.4644 = 0.72192. As for Data 3 AAABB which also contains two letter types A and B, wherein A can be calculated using the formula (3/5)X(-log2(3/5)) = 0.4422, B calculated using the formula (2/5)X(log2(2/5)) = 0.5288, thus the total entropy value of AAABB is 0.4422 + 0.5288 = 0.971.

As Shannon's information entropy measures the probability of different states, meaning that occurrence of events or states having high probability gives less information than occurrence of events or states having low probability. In other words, the events or states that have low probability will result in a higher entropy and thus be more complex. In this example, both AAAAB and AAABB contain A and B, which can be said to be equally diversified; however, the information entropy not only takes into consideration the types but also their frequencies and distribution. Since entropy is a measure of "disorder" or "randomness" in a system, a higher entropy in AAABB actually means that 3 As and 2 Bs will have more combinations than 4 As and 1B. In other words, the former is more complex than the latter. This has given entropy measure an advantage than the TTR measure which only takes the number of types and tokens into consideration.

2.2. Entropy as an index of complexity

The measurement of the complexity of a given system is in fact the measurement of the degree of freedom offered by the system. The complexity of a system is subject to the number of states of the system as well as how these states are distributed in terms of frequency. More states and a more even distribution of the states are positively correlated with the complexity level of a system (Kockelman, 2009). In the field of linguistics, quantifying morphological complexity has attracted the attention of the research community. As a reliable metric that can measure both frequency and distribution information, entropy has been effectively used to quantify different linguistic phenomena, in particular, the complexity of morphological systems (Ackerman and Malouf, 2013; Baerman, 2012). Based on Shannon's information theory, Juola (1998, 2008) applied an entropy measure to quantify language complexity at various levels, including lexical, morphological and syntactic ones. In the field of speech therapy, entropy has also been employed to explore the complexity of individual verbs and verb paradigms and its effect on lexical access in unimpaired people and people with aphasia (PWA) (van Ewijk and Avrutin, 2016). In the same vein, entropy has also been applied to evaluate child speech during phonological development by measuring various categories of different word complexity, including words containing consonant singletons and words containing consonant clusters (Babatsouli et al., 2016). Besides, this entropy-based measure has also been used to calculate the complexity of discourse patterns by studying the number of discourse patterns together with the frequency of each pattern (Kockelman, 2009). From a linguistic perspective, Juola (2013) used entropy as an index to calculate the bigrams in the American component of the Google Books N-gram Corpus and found that American culture has become more complex over the years. Similarly, based on a large data-set of speeches and debates from the British parliament, Zhu and Lei (2018) also applied this entropy measure and demonstrated that the British cultural complexity has increased from the perspective of spoken texts. In summary, entropy has been widely used as an index of complexity at morphological, lexical, syntactic, discourse, and cultural levels.

Table 1
A comparison between TTR and entropy.

No.	Data	Types	Tokens	TTR	Entropy
1	AAAAA	1	5	0.2	0
2	AAAAB	2	5	0.4	0.722
3	AAABB	2	5	0.4	0.971
4	AABBC	3	5	0.6	1.522
5	ABCDE	5	5	1	2.322

2.3. Shannons entropy in language and translation research

As an index for measuring information richness and complexity, the concept of entropy was first developed outside the field of linguistics. In 1948, Shannon used the concept of entropy in information theory to describe how much randomness or information content a signal or random event contains. Technically speaking, entropy is a measure of uncertainty, disorder, or of a large configuration of equiprobable choices (Kent, 1986).

Shannon's entropy has long been used in language and translation research. Entropy-based measures have been used to quantify writing styles of literary texts (Hoover, 2003; Thoiron, 1986). At first, the measure was used to quantify word-level information (Thoiron, 1986; Zhang, 2016). As research progresses, entropy has also been used as an indicator to measure the linguistic diversity of Internet information (Paolillo et al., 2005), morphological diversity and complexity from a psycholinguistic approach (Ackerman and Malouf, 2013; Juola, 2008; Xanthos and Gillis, 2010), complexity of Shakespeare's different genres including poems, comedies and tragedies (Rosso et al., 2009). Bentz et al. (2017) applied word entropy to 1259 languages and found that word entropies display relatively narrow and unimodal distributions. In computational linguistics, the concept of entropy has widely been used to tackle problems in relation to machine translation, including improving segmentation to boost machine translation quality (Xiong et al., 2011) and evaluation of machine translation quality (Carl and Schaeffer, 2014). Bangalore et al. (2016) proposed using the concept of translation entropy to measure all observed word translation probabilities of a given ST word into TT words. A source word would have higher word translation entropy if it has more different equally probable translations. However, such a concept is slightly different from the one that is used in linguistic research to measure language diversity and complexity.

In the field of genre research, entropy has also been used to quantify structure differences in Literature Nobel laureates and other famous authors (Febres and Jaffé, 2017). It was found that text genre influences the resulting entropy and diversity of the text apart from a correlation between entropy and word diversity with quality of writing. Chen et al. (2017) also investigated how the unique linguistic profile of different text types can be reflected in their respective entropy characteristics and identified a strikingly similar distribution pattern in Chinese and English concerning the relative entropy of word forms and POS forms. However, to the best of our knowledge, entropy has not been used to distinguish translated from non-translated texts from an information-theoretical perspective. Our research drew insights from the studies in the foregoing review and attempted to examine whether translated texts in different genres display different entropy characteristics.

2.4. Types of entropy in corpus studies

The current research adopts the indicator, i.e., Shannon's entropy, to measure text complexity. We calculated two types of entropy, i.e., unigram entropy which is based on word forms and POS (Part of Speech) entropy which is based on the POS forms. "As words are the morphological realizations of grammar, the means of the relative entropy of wordforms can tell us the average vocabulary richness [.. .] in different text types" (Chen et al., 2017:535). However, although word frequency and distribution can be measured by entropy to demonstrate lexical richness, it cannot be directly equalized with their degree of syntactic variations. Thus, researchers have proposed using POS forms to measure syntactic complexity. As argued by Chen et al. (ibid), POS forms serve as "a more reliable indicator of syntactical differences, as POS has already attained a certain degree of abstraction for words". For this reason, the current study uses word forms and POS forms to measure lexical and syntactic complexity of translational and non-translational texts. "From the linguistic point of view, lower relative entropy of POS-forms means more syntactic regularity, more stereotypical of the text type; while higher entropy indicates more syntactic freedom or variation, more peculiar of the text type" (Chen et al., 2017:535-6). For this reason, it is deemed appropriate for the current study to use POS forms as a reliable measure of syntactic complexity.

3. RESEARCH QUESTIONS

Based on the foregoing review, some research gaps can be summarized regarding the issue of simplification research. First, previous studies have limited their investigation of translational simplification to a comparison between translated and non-translated texts while ignoring the factor of genres. Second, a vast proportion of research has been undertaken based on European language pairs, thus the generalizability of the research is limited. Arguably, evidence from typologically distinct language pairs such as English and Chinese can yield more convincing evidence (Xiao and Dai, 2014). Finally, no research, to our knowledge, has used an indicator to measure translational simplification from a macro and computational linguistic perspective.

The purpose of the study reported in this paper aims to explore whether translational Chinese is simpler than native Chinese across four genres using Shannon's entropy as an indicator. The following three research questions will be addressed:

(RQ1) Do differences exist in the lexical and syntactic complexity of texts based on translation status?

(RQ2) Do differences exist in the lexical and syntactic complexity of texts based on text type?
(RQ3) Does interaction occur between translation status and text type in effecting the lexical and syntactic complexity of texts?

4. DATA AND METHOD

4.1. Corpora

The study is conducted on two corpora, i.e. The Lancaster Corpus of Mandarin Chinese (LCMC) comprising the native Chinese texts, and The Zhe Jiang University Corpus of Translational Chinese (ZCTC) comprising the translated Chinese texts. Both corpora were modelled after The Freiburg-LOB (FLOB) Corpus which consists of around one million tokens of written British English sampled from fifteen text categories published in the early 1990s (Hundt et al., 1998). The LCMC and The ZCTC were created as the native and translated Chinese matches for FLOB by using the same sampling techniques and matching the corresponding sample period (McEnery et al., 2003; Xiao and Hu, 2015). Being one-million word balanced corpora that are comparable in size, both corpora contain 500 texts of around 2,000 words each in 15 text categories, falling into four macro genres: press, general prose, academic prose, and fiction (Chen et al., 2017). Both word segmentation and POS annotation were conducted with LCMC and ZCTC. The text types of both corpora are presented in Table 2. The text types, although not exhaustive, are believed to be representative of translation and non-translation and adequate for the purpose of the current research needs.

Since the two corpora contain a wide variety of genres and text types, they can allow an effective comparison of translational against the native non-translational language. Also, the two corpora have been segmented and annotated using standardized methods, thus the comparability was greatly enhanced. These two corpora have been extensively studied in corpus-based research to investigate the similarities and differences between translated Chinese texts and comparable non-translated native Chinese texts, such as lexical density, information load, high frequency words, mean sentence length, and word clusters, word frequency and word length, keywords, word class distribution, the use of pronouns and prepositions, idioms and major types of punctuation (Xiao, 2010; Xiao and Dai, 2014; Xiao and Hu, 2015). These studies have yielded some insights into the unique features of native and translated Chinese. However, such types of research are not without their problems. The use of individual linguistic indicators has some limitations. One major problem with these types of research is that the results based on individual indicators will yield conflicting results and fail to provide a full picture as to the global features of translational language. There is so far no research aiming at studying the features of the translated texts against native texts using entropy as a global measure of text complexity. It is deemed that an entropy-based research comparing these two corpora would offer quantitative evidence of translated Chinese as opposed to native Chinese.

Table 2
Genres and Text types in LCMC and ZCTC.

Genres	Text types	Samples	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	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%

4.2. Segmentation and data processing

Text segmentation has been an essential task in natural language processing as well as corpus linguistics. Notable progress has been achieved in segmentation in the past two decades in both English and Chinese languages (Wong et al., 2009; Zhang et al., 2003). Unlike segmentation in English which is already considered to be a solved problem, word segmentation in Chinese is relatively more complicated as Chinese is highly ambiguous depending on the different

contextual aspects (McEnery and Xiao, 2004). One major difficulty is that sentences in Chinese consist of an uninterrupted string of characters without clear delimitations. Most of the Chinese word tokens are made of a combination of characters. It is therefore important to segment text strings into word tokens before POS tagging. In this study, we used the Stanford CoreNLP Natural Language Processing Toolkit, which consists of The Stanford Parser (Levy and Manning, 2003) and The Stanford Chinese POS Tagger (Tseng et al., 2005), to work with segmentation and POS tagging. The Tagger has been widely applied in various corpus-based Chinese studies and claims a high accuracy rate of 93.65%. In the following, we used one sentence from the corpus to show how the segmentation and POS tagging work in our study.

Chinese Sentence: 相对地, 中国的佛教, 也不全同于印度或其他国家的佛教;

(English Translation: relatively, Chinese Buddhism is not completely different from Buddhism in India or other countries;)

Tokenize: ['相对', '地', '，', '中国', '的', '佛教', '，', '也', '不', '全同于', '印度', '或', '其他', '国家', '的', '佛教', ';']

Part of Speech: [('相对', 'AD'), ('地', 'DEV'), ('，', 'PU'), ('中国', 'NR'), ('的', 'DEG'), ('佛教', 'NN'), ('，', 'PU'), ('也', 'AD'), ('不', 'AD'), ('全同于', 'VV'), ('印度', 'NR'), ('或', 'CC'), ('其他', 'DT'), ('国家', 'NN'), ('的', 'DEG'), ('佛教', 'NN'), (';', 'PU')]

As can be seen from the above example, the Chinese sentence is separated into three segments demarcated by two commas and there is no clear boundary between the words. In such a form, the computer cannot tell which ones are words as there is no space between them. The Stanford Parser helped tokenize the Chinese sentence into 14 individual words to be further POS tagged by The Stanford Chinese POS Tagger. Using such a method, we are able to segment and annotate the two corpora according to standardized criterion.

Note that previous studies have found that text length can be a major variable affecting the overall entropy values (Shi and Lei, 2020). Thus, it is important to safeguard an equal length in each file for analysis. It was found that the texts in the two corpora differ greatly in length despite the compilers' claim that each text contains around 2000 Chinese words. Thus, we tested different text length in order to ensure that each text contains the same number of words while retaining the same number of files as the original corpus design. Finally, we decided on 1500 Chinese words per text, which is the maximum number of words we could obtain in each text file in order to retain the same number of files (i.e. 500 files) as per Table 2. Files which have more than 1500 Chinese words were trimmed to meet the criteria. After tagging is finished, all the punctuation was removed in order to eliminate the confounding variable of possible disparity of punctuation use between the two corpora. In the current study, we mainly conducted two types of entropy calculations. The first one is based on the words in which we calculated the unigram entropy values of all 500 texts in each of the two corpora. The second one is based on POS forms in which we calculated the POS entropy values of the two corpora. All the calculations were performed in Python program using Formula (1).

4.3. Methodology

In order to answer the three research questions set out in Section 2.4, we conducted two two-way ANOVA tests. The first one is mainly aimed at testing the effect of two independent variables of translation status and text type on the dependent variable of lexical complexity. Specifically, this ANOVA test examined whether translation status (translation vs. non-translation) and/or text type (press vs. general prose vs. academic prose vs. fiction) have an effect on the lexical complexity (i.e., unigram entropy values). This test aimed to find out if there is a main effect of translation (RQ1) or a main effect of text type (RQ2) or an interaction of both factors (RQ3). If interaction is identified, we would then conduct a Tukey post hoc test with adjusted family-wise error rates to further examine which group means were significantly different (RQ3).

Likewise, the second ANOVA test is mainly aimed at testing the effect of the two independent variables of translation status and text type on the dependent variable of syntactic complexity. Similar to the first test, this ANOVA test examined if translation status (translation vs. non-translation) and/or text type (press vs. general prose vs. academic prose vs. fiction) have an effect on the lexical complexity (i.e., POS entropy values). We aimed to identify if there is a main effect of translation (RQ1) or a main effect of text type (RQ2) or an interaction of both factors (RQ3). Again, if there is an interaction of both factors, a Tukey post hoc test would be conducted to examine which group means were significantly different (RQ3).

5. RESULTS

The mean word and POS entropy values of the two corpora are summarized and presented in Table 3 and Fig. 1. We calculated the entropy values of unigram and POS of the four text dimensions (comprising 15 text types) in the two corpora. The results show that native texts (LCMC) are consistently higher in unigram entropy than translated texts (ZCTC) in all four genres. On the contrary, the translated texts are higher in POS entropy than non-translated texts in all four genres.

First, in order to further study if the differences are significantly different between the two corpora in unigram entropy, we

conducted a two-way ANOVA on unigram entropy using corpus and genre as independent variables. Results show there is a main effect of corpus, indicating a significant difference in unigram entropy between the two corpora ($F(1, 992) = 140.84$; $p < .001$), with a higher unigram entropy in LCMC than ZCTC (8.517 vs. 8.328). There is also a main effect of genre, showing that different genres also differed significantly in their unigram entropy ($F(3, 992) = 116.75$; $p < .001$). The interaction between corpus and genre was significant ($F(3, 992) = 12.88$; $p < .001$), suggesting that different genres have different unigram entropies across the two corpora (See Fig. 2). Examination of the four genres (see Table 3 and also Fig. 1) showed that the genre of press has the highest unigram entropy (8.602), followed by General Prose (8.489), and then Fiction (8.431) and Academic prose (8.161).

Table 3
Mean Unigram entropy and POS entropy of LCMC and ZCTC.

Corpus	Genre	Unigram Mean	Std. Deviation	POS Mean	Std. Deviation	N
LCMC	Press	8.674	0.293	3.466	0.105	88
	General Prose	8.633	0.258	3.508	0.167	176
	Academic	8.175	0.302	3.259	0.23	110
	Fiction	8.542	0.242	3.563	0.113	126
	Total	8.517	0.329	3.46	0.197	500
ZCTC	Press	8.53	0.182	3.629	0.097	88
	General Prose	8.345	0.205	3.675	0.123	176
	Academic	8.147	0.312	3.572	0.116	110
	Fiction	8.32	0.215	3.741	0.077	126
	Total	8.328	0.26	3.661	0.123	500
Total	Press	8.602	0.254	3.547	0.129	176
	General Prose	8.489	0.274	3.592	0.169	352
	Academic	8.161	0.307	3.416	0.24	220
	Fiction	8.431	0.254	3.652	0.131	252
	Total	8.422	0.311	3.56	0.192	1000

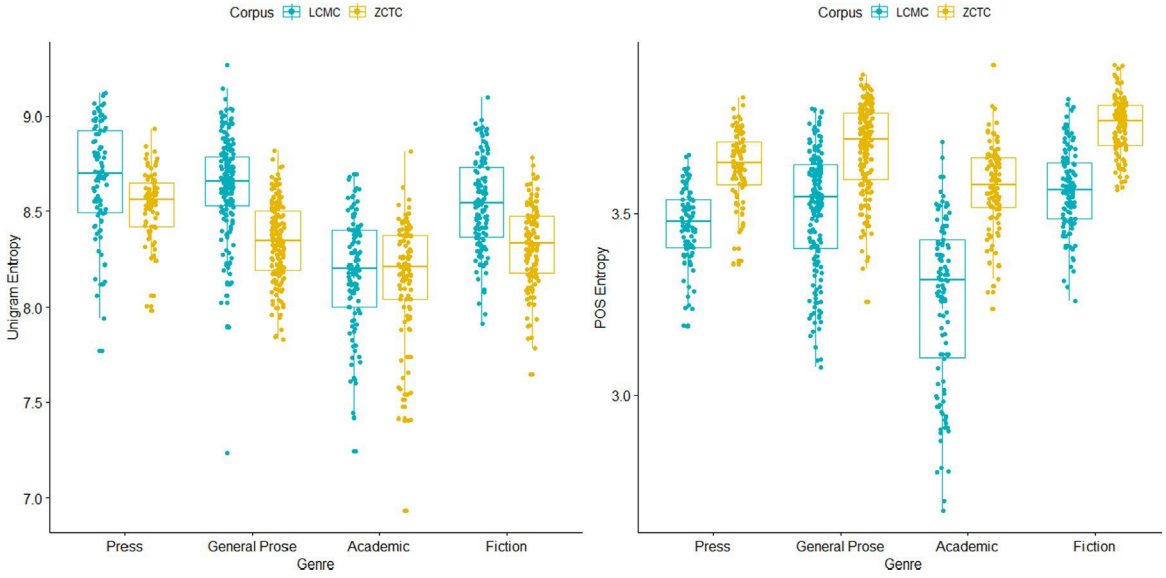

Fig. 1. Boxplots of mean unigram entropy and POS entropy of ZCTC and LCMC.

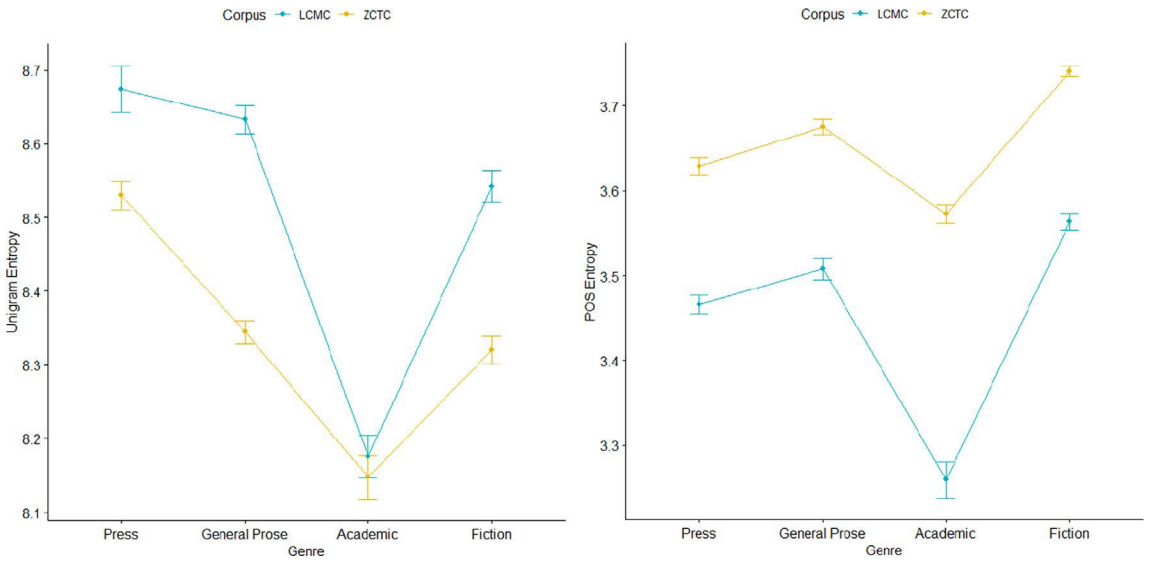

Fig. 2. Interaction between corpus and genre for unigram and POS entropy.

As the ANOVA revealed significant interaction, we further conducted a Turkey's post hoc test for multiple comparisons to further analyze the impact of two independent variables including corpus (translation, non-translation), genre (press, general prose, academic prose, fiction) on the unigram entropy values to examine potential differences between the same genre across the two corpora (see Table 4). Specifically, we found that there were significant differences between all the four genres. While comparing LCMC and ZCTC, significant differences are found in the three genres of press, general prose and fiction (ps < .005) while there was no significant difference in the genre of academic prose between the two corpora (p >.05). Fig. 2 shows that LCMC is consistently higher in unigram entropy in three genres (press, general prose, fiction), but such a difference is hardly noticeable in the genre of academic prose.

Table 4
Tukey multiple comparisons of unigram entropy means.

Pair	Mean Difference	Lower Bound	Upper Bound	Sig.
B-A	0.113	0.173	0.053	<0.001**
C-A	0.441	0.506	0.375	<0.001**
D-A	0.171	0.235	0.107	<0.001**
C-B	0.327	0.383	0.272	<0.001**
D-B	0.058	0.111	0.004	0.028*
D-C	0.270	0.210	0.330	<0.001**
ZCTC:A-LCMC:A	0.144	0.260	0.029	0.004*
ZCTC:B-LCMC:B	0.289	0.370	0.207	<0.001**
ZCTC:C-LCMC:C	0.028	0.131	0.075	0.992
ZCTC:D-LCMC:D	0.222	0.319	0.126	<0.001**

Note: A = Press; B = General prose; C = Academic prose; D = Fiction. *p ::; 0.05, **p ::; 0.001.

Table 5
Tukey multiple comparisons of POS entropy means.

Pair	Mean Difference	Lower Bound	Upper Bound	Sig.
B-A	0.044	0.011	0.077	0.003*
C-A	0.132	0.168	0.096	<0.001**
D-A	0.105	0.070	0.140	<0.001**
C-B	0.176	0.206	0.145	<0.001**
D-B	0.061	0.031	0.090	<0.001**
D-C	0.237	0.204	0.269	<0.001**
ZCTC:A-LCMC:A	0.163	0.100	0.226	<0.001**
ZCTC:B-LCMC:B	0.168	0.123	0.212	<0.001**
ZCTC:C-LCMC:C	0.313	0.257	0.370	<0.001**
ZCTC:D-LCMC:D	0.178	0.125	0.231	<0.001**

Note: A = Press; B = General prose; C = Academic prose; D = Fiction. *p ::; 0.05, **p ::; 0.001.

Next, we further conducted a similar two-way ANOVA on POS entropy using corpus and genre as independent variables. Results show there is a main effect of corpus, indicating translation and non-translation differ significantly in POS entropy ($F(1; 992) = 534.15$; $p < .001$), with a higher POS entropy in ZCTC than in LCMC (3.66 vs. 3.46). There is also a main effect of genre, showing different genres differed significantly in their POS entropy ($F(3; 992) = 124.69$; $p < .001$). Examination of the four genres (see Table 3 and also Fig. 1) showed that Fiction led to the highest entropy (3.652), followed by Academic prose (3.592), and then Press (3.54), with Academic prose having the lowest entropy (3.416). The interaction between corpus and genre was significant ($F(3; 992) = 15.54$; $p < .001$), suggesting that genres have different entropies across the two corpora (See Fig. 2).

A Turkey's post hoc test for multiple comparisons was conducted to further analyze the impact of two independent variables including corpus (translation, non-translation), genre (press, general prose, academic prose, fiction) on the POS entropy values to examine potential differences between the same genre across the two corpora (see Table 5). It reveals that all four genres differed significantly from each other (ps < .005). Besides, comparison of the specific genres of LCMC and ZCTC showed that all the four genres differed significantly across the two corpora (ps < .001), suggesting that translation differ from non-translation in all four genres in POS entropy categorically. Examination of Fig. 2 showed that, while genres are generally lower in POS entropy in the LCMC corpus than in the ZCTC corpus, academic prose was much lower in the former than in the latter corpus.

6. DISCUSSION AND CONCLUSION

The present study demonstrated that translated texts are different from non-translated native texts from an information-theoretic perspective using entropy as an indicator. Specifically, translation is simpler at the lexical but not at the syntactic level than native texts. This means that, compared to native texts, translation uses a limited range of words but has a higher syntactic complexity (i.e., more varied syntactic structures) as measured in terms of POS forms. Our research findings echo those of Xiao and Yue (2009) who found that translated texts have significantly higher average sentence length than non-translated ones and Xiao (2010) who found that translated Chinese has lower lexical density than native Chinese but shows no significant difference from native Chinese in mean sentence length. As a matter of fact, a large number of previous research has failed to distinguish between lexical and syntactic complexity in their quest of the simplification universal. The use of simplification as an umbrella term has failed to paint a genuine picture of a multi-faceted linguistic phenomenon. So far as Chinese is concerned, we have identified in the current study that translated Chinese texts are characterized by both lexical simplification and syntactic complexification. Our research has, to a certain extent, rejected the simplification hypothesis which is understood as a universal feature.

For a long time, researchers working on TUs have tended to affirm simplification as a universal feature (Bernardini et al., 2016; Laviosa, 1998). The current research has yielded some interesting results. Lexical simplification in translation is confirmed in our research while syntactic simplification is rejected. This shows that translation, as a mediation activity, is operating at both the lexical and syntactic levels. For a long time, researchers have come up with different models to explain lexical simplification in translation. For example, the Hypothesis of Gravitational Pull (Halverson, 2003, 2017) has been seen as one of the robust models on merit of its wide-ranging explanatory power for translational features. The model contains of three interrelated forces that are in interplay to shape the translational language. The first force known as the magnetism effect would tempt the translators to reproduce the prototypical salient language traits of the target language in their translations. Conversely, the translator is also faced with an opposing force known as the gravitational pull effect initiated by the source language which would resist the magnetism effect. The third important variable in this model that contributes to the profiling of translational language is the connectivity effect which occurs due to the impact of high frequency co-occurrence of translation equivalents in the source and target languages. The translational language emerges from an interaction and interplay of these three forces. This model has been found exceptionally powerful in explaining translation universals involving lexical features, such as the unique item hypothesis that postulates an underrepresentation of "unique items" in translations than native texts (Tirkkonen-Condit, 2005). The findings of this study also suggests that the model is also applicable to syntactic features. To a large extent, the findings show that the magnetism effect is at a significantly greater force at the lexical level, resulting in an underrepresentation of word forms in translated Chinese. Yet, the gravitational pull effect is operating more effectively at the syntactic level, resulting in more varied and complex structures in translated Chinese.

Although the Hypothesis of Gravitational Pull model mentioned above attempts to provide a systematic account of simplification phenomenon in translation, the description of the different effects is relatively vague as it fails to describe under what circumstances one effect is more forceful than the other two. In the field of psycholinguistics, there are two competing psychological frameworks to characterize the translation process. One is the vertical model (Fodor et al., 1974) which assumes that translation is a sequential process in which a message is deverbalized then reverbalized, and the final translation production is free from the influence of the source language. In this model, meaning is on the driver seat, shaping lexical and grammatical forms in the translated text. On the other hand, the Horizontal model of translation (Maier et al., 2017; Ruíz and Macizo, 2019) holds that features of the linguistic resources of the translator are linked "via shared memory representations and that cognitive processes during translation are specific to the combination of both languages involved" (Schaeffer and Carl, 2013). In horizontal translation, lexical and syntactic properties of the source and target languages are linked via shared representations, which means that both decoding

of source texts and encoding of target texts work in parallel. A number of studies have corroborated the horizontal model (Ruiz et al., 2008; Balling et al., 2014; Maier et al., 2017). This shows that translation is a result of interaction between the source and target language features. The final translation output is a result of competing forces occurring in source and target languages. Lexical simplification as found in this study shows that some lexical features of the target language which are not linked with source language via shared representation are not activated, leading to a limited variety of lexicon in translation. On the other hand, the higher syntactic complexity in translation as measured by POS entropy means that translation has more varied structures than non-translation. This corroborates the psycholinguistic demonstration that, in translating a sentence, translators tend to use a structure in the target language that is similar to the structure of the source sentence (Maier et al., 2017). This might be related to the uniqueness of Chinese which is often labelled as a paratactic language. Wang (1943/1984) used two concepts to compare Chinese and Western languages by stating that Chinese emphasizes "*yihe* (i.e., parataxis)" while Western languages including English stress "*xinghe* (i.e., hypotaxis)". According to Wang, Chinese is characterized by a lack of such function words as connectives. However, in the Europeanization process of the Chinese language, the use of connectives has increased considerably, particularly in translation.

We have shown in the current research that entropy can be adopted as an effective measure to study translational and non-translational language. By applying an entropy-based approach to the study of lexical and syntactic simplification in translation, the current study has practical and methodological implications for corpus-based investigations of TUs. First, the use of a global indicator (i.e. entropy) can paint a more holistic picture of translational language than using individual language features, which might lead to conflicting results depending on the features used. Second, we have shown that translated Chinese is characterized by both lexical simplification and syntactic complexification in comparison to non-translated Chinese, which is different from previous studies which investigated simplification using individual lexical features. Third, the current research has opened new avenues for future research in quantifying simplification from the perspective computational and quantitative linguistics.

Nonetheless, it should be noted that findings of the current research are limited to English-Chinese translations and the causes might be due to the differences between these specific languages involved. In future studies, more studies involving the other translation direction (i.e., Chinese-English translation) or language pairs should be conducted to test the simplification universal.

References

Ackerman, F., Malouf, R., 2013. Morphological organization: The low conditional entropy conjecture. Language 89 (3), 429–464.

Babatsouli, E., Ingram, D., Sotiropoulos, D.A., 2016. Entropy as a measure of mixedupness of realizations in child speech. Poznan Stud. Contemporary Linguist. 52 (4), 605–627.

Baerman, M., 2012. Paradigmatic chaos in Nuer. Language 88 (3), 467–494.

Baker, M., 1999. The role of corpora in investigating the linguistic behaviour of professional translators. Int. J. Corpus Linguist. 4 (2), 281–298.

Baker, M., 1993. Corpus linguistics and translation studies: Implications and applications. In: Baker, M., Francis, G., Tognini-Bonelli, E., Sinclair, J. (Eds.), Text and Technology: In Honour of John Sinclair. John Benjamins, Philadelphia.

Baker, M., 1996. Corpus-based translation studies: The challenges that lie ahead. In: Sager, J.C., Somers, H.L. (Eds.), Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager. John Benjamins, Amsterdam.

Balling, L.W., Hvelplund, K.T., Sjørup, A.C., 2014. Evidence of parallel processing during translation. Meta 59, 234–259.

Bangalore, S., Behrens, B., Carl, M., Ghankot, M., Heilmann, A., Nitzke, J., Schaeffer, M., Sturm, A., 2016. Syntactic variance and priming effects in translation. In: Carl, M., Bangalore, S., Schaeffer, M. (Eds.), New Directions in Empirical Translation Process Research: Exploring the CRITT TPR-DB. Springer, Cham, pp. 211–238.

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.

Bernardini, S., Ferraresi, A., 2011. Practice, description and theory come together-normalization or interference in Italian technical translation? Meta 56 (2), 226–246.

Bernardini, S., Ferraresi, A., Miličević, M., 2016. From EPIC to EPTIC–Exploring simplification in interpreting and translation from an intermodal perspective. Target. Int. J. Transl. Stud. 28 (1), 61–86.

Blum-Kulka, S., Levenston, E.A., 1983. Universals of lexical simplification. In: Faerch, K., Kasper, G. (Eds.), Strategies in Interlanguage Communication. Applied Linguistics and Language Study. Longman, London and New York, pp. 119-139.

Cappelle, B., 2012. English is less rich in manner-of-motion verbs when translated from French. Across Languages Cultures 13 (2), 173–195.

Carl, M., Schaeffer, M., 2014. Word transition entropy as an indicator for expected machine translation quality. In: Miller, K.J., Specia, L., Harris, K., Bailey, S. (Eds.), Proceedings of the Workshop on Automatic and Manual Metrics for Operational Translation Evaluation. MTE 2014, Pari, pp. 45–50.

Chen, W., 2004. Investigating explicitation of conjunctions in translated Chinese: A corpus-based study. Language Matters 35 (1), 295–312.

Chen, R., Liu, H., Altmann, G., 2017. Entropy in different text types. Digital Scholarship Humanit. 32 (3), 528–542.

Chesterman, A., 2014. Translation studies forum: Universalism in translation studies. Translation Studies 7 (1), 82–90.

Chesterman, A., 2004. Beyond the Particular. In: Mauranen, A., Kujamäki, P. (Eds.), Translation Universals: Do They Exist? John Benjamins, Amsterdam and Philadelphia, pp. 33–49.

Cvrček, V., Lucie, C., 2015. Simplification in translated Czech: A new approach to type-token ratio. Russian Linguistics 39 (3), 309–

325.
Delaere, I., Sutter, G.D., Plevoets, K., 2012. Is translated language more standardized than non-translated language?: Using profile-based correspondence analysis for measuring linguistic distances between language varieties. Target 24 (2), 203–224.

Fan, L., Jiang, Y., 2019. Can dependency distance and direction be used to differentiate translational language from native language? Lingua 224, 51–59.
Febres, G., Jaffé, K., 2017. Quantifying structure differences in literature using symbolic diversity and entropy criteria. J. Quantit. Linguist. 24 (1), 16–53.
Feng, H., Crezee, I., Grant, L., 2018. Form and meaning in collocations: A corpus-driven study on translation universals in Chineseto-English business translation. Perspectives 26 (5), 677–690.
Fodor, J.A., Bever, T.G., Garrett, M.F., 1974. The Psychology of Language. McGraw-Hill, New York.
Frawley, W., 1984. Translation: Literary, Linguistic, and Philosophical Perspectives. University of Delaware Press; Associated University Presses, Newark and London.
Halverson, S.L., 2003. The cognitive basis of translation universals. Target. Int. J. Transl. Stud. 15 (2), 197–241.
Halverson, S., 2017. Gravitational pull in translation: Testing a revised model. In: De Sutter, G., Lefer, M.A., Delaere, I. (Eds.), Empirical Translation Studies: New Methodological and Theoretical Traditions. Mouton De Gruyter, Boston, MA, pp. 9–46.
Hoover, D.L., 2003. Another perspective on vocabulary richness. Comput. Humanit. 37 (2), 151–178.
House, J., 2015. Translation quality assessment: Past and present. Translation: A Multidisciplinary Approach. Palgrave Macmillan, London, pp. 241–264.
Hundt, M., Sand, A., Siemund, R., 1998. Manual of information to accompany the Freiburg-LOB Corpus of British English (FLOB). Albert-Ludwigs Universitat Freiburg.
Jantunen, J., 2004. Untypical patterns in translations. Issues on corpus methodology and synonymity. In: Mauranen, A., Kujamäki, P. (Eds.), Translation Universals: Do They Exist? John Benjamins, Amsterdam and Philadelphia.
Juola, P., 1998. Measuring linguistic complexity: The morphological tier. J. Quantit. Linguist. 5 (3), 206–213.
Juola, P., 2008. Assessing linguistic complexity. In: Miestamo, M., Sinnemaki, K., Karlsson, F. (Eds.), Language Complexity: Typology, Contact, Change. John Benjamins Press, Amsterdam.
Juola, P., 2013. Using the Google N-gram corpus to measure cultural complexity. Literary Linguist Comput 28 (4), 668–675.
Kalimeri, M., Constantoudis, V., Papadimitriou, C., Karamanos, K., Diakonos, F.K., Papageorgiou, H., 2015. Word-length entropies and correlations of natural language written texts. J. Quanti. Linguist. 22 (2), 101–118.
Kent, T., 1986. Interpretation and Genre: The Role of Generic Perception in the Study of Narrative Texts. Bucknell University Press.
Khmelev, D.V., 2000. Disputed authorship resolution through using relative empirical entropy for markov chains of letters in human language texts. J. Quantit. Linguist. 7 (3), 201–207.
Kockelman, P., 2009. The complexity of discourse. J. Quant. Linguist. 16 (1), 1–39.
Kruger, H., van Rooy, B., 2012. Register and the features of translated language. Across Languages Cultures 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. Rodopi, Amsterdam and New York, NY.
Levy, R., 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. 439–446.
Liu, Z., 2016. A diachronic study on British and Chinese cultural complexity with Google Books N-grams. J. Quant. Linguist. 23 (4), 361–373.
Maier, R.M., Pickering, M.J., Hartsuiker, R.J., 2017. Does translation involve structural priming? Quart. J. Exp. Psychol. 70 (8), 1575–1589.
Mauranen, A., 2006. Translation universals. In: Brown, K. (Ed.), Encyclopedia of Language and Linguistics. Elsevier, pp. 93–100.
McEnery, T., Xiao, R., 2004. The Lancaster Corpus of Mandarin Chinese: A corpus for monolingual and contrastive language study. Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC) 2004, Lisbon, pp. 1175–1178.
McEnery, A., Xiao, Z., Mo, L., 2003. Aspect marking in English and Chinese: Using the Lancaster Corpus of Mandarin Chinese for contrastive language study. Literary Linguistic Comput. 18, 361–378.
Olohan, M., Baker, M., 2000. Reporting that in translated English. Evidence for subconscious processes of explicitation. Across Languages Cultures 1 (2), 141–158.
Paolillo, J.C., Pimienta, D., Prado, D., et al., 2005. Measuring Linguistic Diversity on the Internet. UNESCO, Paris. http://uis.unesco.org/sites/default/files/documents/measuring-linguistic-diversity-on-the-internet-ict-2005-en_0.pdf (Accessed April 30, 2021)/.
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. John Benjamins, Amstedam, pp. 311–328.
Rosso, O.A., Craig, H., Moscato, P., 2009. Shakespeare and other English Renaissance authors as characterized by Information Theory complexity quantifiers. Physica A 388 (6), 916–926.
Ruíz, J.O., Macizo, P., 2019. Lexical and syntactic target language interactions in translation. Acta Psychol. 199, 102924. https://doi.org/10.1016/j.actpsy.2019.102924.
Ruiz, C., Paredes, N., Macizo, P., Bajo, M.T., 2008. Activation of Lexical and Syntactic Target Language Properties in Translation. Acta Psychol. 128 (3), 490–500. https://doi.org/10.1016/j.actpsy.2007.08.004.
Schaeffer, M., Carl, M., 2013. Shared representations and the translation process: A recursive model.Translation and Interpreting Studies. J. Am. Transl. Interpret. Stud. Assoc. 8 (2), 169–190.
Shannon, C.E., 1948. A mathematical theory of communication. Bell Syst. Techn. J. 27 (4), 379–423.
Shannon, C.E., 1951. Prediction and entropy of printed English. Bell Syst. Techn. J. 30 (1), 50–64.
Shi, Y., Lei, L., 2020. Lexical richness and text length: An entropy-based perspective. J. Quantit. Linguist. 29 (1), 62–79.

Takahira, R., Tanaka-Ishii, K., Dębowski, ., 2016. Entropy rate estimates for natural language – A new extrapolation of compressed large-scale corpora. Entropy 18 (10), 364.

Thoiron, P., 1986. Diversity index and entropy as measures of lexical richness. Comput. Humanit. 20 (3), 197–202.

Tirkkonen-Condit, S., 2005. Do unique items make themselves scarce in translated Finnish? In: Károly, K., Fóris, Á. (Eds.), New Trends in Translation Studies: In Honour of Kinga Klaudy. Akadémiai Kiadó, Budapest, pp. 177–189.

Toury, G., 1995. Descriptive Translation Studies and Beyond. John Benjamins, Amsterdam.

Toury, G., 2004. Probabilistic explanations in translation studies: Welcome as they are, would they qualify as universals? In: Mauranen, A., Kujamäki, P. (Eds.), Translation Universals: Do They Exist? John Benjamins, Amsterdam and Philadelphia, pp. 15–32.

Tseng, H., Jurafsky, D., Manning, C.D., 2005. Morphological features help POS tagging of unknown words across language varieties. In: Proceedings of the Fourth SIGHAN Workshop on Chinese Language Processing, pp. 32–39.

Tymoczko, M., 1998. Computerized corpora and the future of translation studies. Meta: J. des traducteurs/Meta: Translators' J. 43 (4), 652–660.

Ulrych, M., Murphy, A.C., 2008. Descriptive translation studies and the use of corpora: Investigating mediation universals. In: Taylor Torsello, C., Ackerley, K., Castello, E. (Eds.), Corpora for University Language Teachers. Peter Lang, Bern, pp. 141–166. van Ewijk, L., Avrutin, S., 2016. Lexical access in nonfluent aphasia: A bit more on reduced processing. Aphasiology 30 (11), 1264–1282.

Vanderauwera, R., 1985. Dutch novels translated into English. The transformation of a 'minority' literature, Rodopi, Amsterdam.

Wang, L., 1943/1984. Zhongguo xiandai yufa (A grammar of modern Chinese). Wang Li Wenji (The Collected Works of Wang Li), vol. 2. Shandong Education Press, Jinan.

Wong, K.-F., Li, W., Xu, R., Zhang, Z.-S., 2009. Introduction to Chinese natural language processing. Synthesis Lectures Human Language Technol. 2 (1), 1–148.

Xanthos, A., Gillis, S., 2010. Quantifying the development of inflectional diversity. First Lang 30 (2), 175–198.

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.

Xiao, R., Dai, G., 2014. Lexical and grammatical properties of translational Chinese: Translation universal hypotheses reevaluated from the Chinese perspective. Corpus Linguist. Linguist. Theory 10 (1), 11–55.

Xiao, R., Hu, X., 2015. Corpus-based Studies of Translational Chinese in English-Chinese Translation. Springer, Berlin Heidelberg.

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

Xiong, D., Zhang, M., Li, H., 2011. A maximumentropy segmentation model for statistical machine translation. IEEE Trans. Audio Speech Lang. Process. 19 (8), 2494–2505.

Zhang, Y., 2016. Entropic evolution of lexical richness of homogeneous texts over time: A dynamic complexity perspective. J. Language Modell. 3, 569–599.

Zhang, H.P., Xiong, D., Liu, Q., 2003. HHMM-based Chinese lexical analyzer ICTCLAS. In: Proceedings of the Second SIGHAN Workshop on Chinese Language Processing, pp. 184–187.

Zhu, H., Lei, L., 2018. Is modern English becoming less inflectionally diversified? Evidence from entropy-based algorithm. Lingua 216, 10–27.