RiCL Research in Corpus Linguistics

Lexical simplification in learner translation: A corpus-based approach

Ho Ling Kwok[a] – Sara Laviosa[b] – Kanglong Liu[a]
The Hong Kong Polytechnic University[a] / China
University of Bari Aldo Moro[b] / Italy

Abstract –The advance of corpus-based methodology in translation studies has greatly enhanced our understanding of the nature of translational language. While most research efforts have focused on identifying the unique features of translations carried out by professionals, comparatively fewer studies have investigated the linguistic features of student translations. In this corpus-based study, we examine if learner translations carried out by Hong Kong students exhibit lexical simplification features *vis-à-vis* comparable written texts. The study is based on two comparable corpora: the *International Corpus of English in Hong Kong* (ICE-HK) and the *Parallel Learner Translation Corpus* (PLTC) compiled at The Hong Kong Polytechnic University. Following Laviosa (1998), we compare four main lexical features (lexical density, type-token ratio, core vocabulary coverage, and list head coverage) to investigate if student translations show a simplification trend. The results demonstrate that Chinese-to-English translation is not lexically simpler than English as a Second Language (ESL) writing. Furthermore, it is lexically denser than ESL writing. Our study aims to provide new insights into learner translation as a form of constrained communication.

Keywords – lexical simplification; learner translation; corpus-based approach; students' translations

1. INTRODUCTION[1]

Translational language is often regarded as a 'third language' (Duff 1981) or 'third code' (Frawley 1984) since it involves the bilateral consideration and accommodation of at least two different codes. In this regard, Baker (1993: 243) proposed the hypothesis of translation universals, referring to them as "universal features of translation […] which typically occur in translated text." Baker (1996) put forward four translation universals: 1) simplification (tendency to simplify language subconsciously), 2) explicitation (tendency to make information clearer), 3) normalization or conservatism (tendency to

[1] This research is sponsored by a General Research Fund (GRF) grant from the *Research Grants Council of Hong Kong* (Ref: 15605520).

Research in Corpus Linguistics 11/2: 103–124 (2023). ISSN 2243-4712. <https://ricl.aelinco.es>
Asociación Española de Lingüística de Corpus (AELINCO)
DOI 10.32714/ricl.11.02.06

conform to typical patterns of the target language), and 4) levelling out (tendency to be more homogeneous than the original texts). Many translation scholars have argued that the term ‘universals’ is not scientifically sound (e.g., Tymoczko 1998; Pym 2008; Saldanha 2011). House (2015: 62) even suggested that “the quest for translation universals is in essence futile.” In her opinion, the absence of careful comparative analyses is an inadequacy in most existing studies, and the terms used to denote them — ‘simplification’ and ‘normalization’— are overly general and lack a clear operational definition. Besides, House (2015: 62–63) argued that universality in translation is questionable as some translation universals are subject to the variables of translation directions and genres. This empirical evidence challenges the claim of translation universals.

Despite these controversies, Baker’s initial proposal suggested several research directions that have yielded new insights into the features of translational language (e.g., Olohan and Baker 2000; Xia 2014; Liu and Afzaal 2021). Baker (1993, 1995, 1996) also pioneered the application of corpus methods to identify features of translated texts, especially the use of comparable corpora to compare translated texts with non-translated ones. Over the years, the quest for translation features has been spurred by advances in corpus-based translation methods and the availability of large-scale computerized corpora. These developments have made it possible to study translation phenomena systematically instead of relying on the researchers’ own experience and subjective evaluation, thus enhancing our understanding of the nature and role of translational language.

In the past three decades, although the features of professional translations have been examined extensively, comparatively little effort has been made to investigate the linguistic features of student translations. As corpus research into learner translation can reveal potential learner problems, a systematic investigation of some central issues in this area has pedagogical implications. This line of research was pioneered by Bowker and Bennison (2003), who described the construction of a student translation archive. Since then, more scholars have devoted themselves to this research area. A recent effort is the *Multilingual Student Translation Project* (MUST; Granger and Lefer 2020), which aims at compiling a student translation corpus covering different language pairs. Overall, we have witnessed an increase in scholarly interest in the field of learner translation in recent years.

The current study is based in Hong Kong, which has been active in learner corpus research over the past two decades due to its bilingual environment. However, these learner corpus studies are mainly related to Second Language Acquisition (SLA) rather than to translation studies (Liu *et al.* 2022). As L1-L2 translation and L2 writing share common challenges and constraints in terms of L2 language production, we used various lexical simplification indicators to identify the extent to which L2 learner translation differs from L2 writing and uncover their possible relationship from a constrained language perspective.

1.1. Constrained communication

Lanstyák and Heltai (2012: 100) suggested the term 'constrained communication' to indicate "communication taking place under conditions where one or several of the potential limiting factors play a greater than average role." This framework implies that all communicative events are influenced by different types and degrees of constraints. However, some have exceptionally prominent constraints, such as language contact situations, including translation and bilingual communication (Lanstyák and Heltai 2012: 100). Based on this framework, Kruger and van Rooy (2016: 27) proposed the term 'constrained language' to denote the language produced under apparent constraints. They also pointed out that both translation and L2 language varieties share the same constraints in the form of bilingual activation and language contact.

These constraints are exhibited in two ways, namely psycholinguistic and social. From a psycholinguistic perspective, constraints are associated with language processing. Bilinguals activate languages along the continuum from a monolingual mode to a bilingual mode in different contexts (Grosjean 2013: 15). In this regard, translation is always operated in the continuous bilingual activation mode (Kruger and van Rooy 2016: 29). In addition, translation is restricted by the pre-existing source text, which can interfere with target language production (Toury 2012: 310–311). The constraints of L2 production are associated with the cognitively demanding environment experienced by non-native speakers in contact situations (Kruger and van Rooy 2016: 31).

From a social perspective, the constraints are also related to language and translation norms. Translators must ensure that the translation is faithful to the source language culture as well as acceptable in the target language culture. Similarly, L2 writers also need to conform to the perceived norms of L2 (Kruger and van Rooy 2016: 27). Against this background, L2 translation, especially the one done by student translators, can be a unique constrained language output that combines features of both translation and L2 production.

In translation, the rapid bi-directional switching involved in the translation process increases the demand for working memory. The lack of a common communication context due to linguistic differences can also lead to malcommunication or non-communication. Therefore, strategies for cognitive load reduction (Carl and Dragsted 2012) and risk minimization (Pym 2015), such as literal translation, explication, and simplification, are often applied by translators. Similarly, L2 writing also involves cognitive and communicative difficulties on the part of non-native speakers who need to use and process an additional language. Some scholars argue that simplification is one of the strategies to deal with these challenges (e.g., Kortmann and Szmrecsanyi 2009; McWhorter 2011). Under these claims, simplification is believed to be one of the features related to translational language and L2 varieties.

1.2. Lexical simplification

Simplification is defined as "the tendency to simplify the language used in translation" (Baker 1996: 181). Over the years, simplification has been studied at different levels, such as lexical (Laviosa 1998; Ferraresi *et al.* 2018; Nasseri and Thompson 2021) and syntactic (McWhorter 2011; Liu and Afzaal 2021). In the current research, we aim to investigate whether L2 translation and L2 writing are (dis)similar in terms of lexical simplification. Lexical simplification can be described as "making do with less words" (Blum-Kulka and Levenston 1983: 119). Lexical simplification is usually operationalized through indicators such as lexical density, the use of frequent words, type-token ratio, and mean sentence length, amongst others (Hu 2016: 101).

The lexical simplification hypothesis has been widely discussed in translation studies. Chesterman (2004) regarded simplification as a potential T-universal, which concerns the translation features in relation to non-translation in the target language. The simplification hypothesis thus assumes that translated texts are simpler than comparable non-translated native texts in the target language. Laviosa (1998) reported evidence that supports the lexical simplification hypothesis with certain parameters. She found that translated narrative prose has lower lexical density, a higher proportion of high-frequency words, and more repetition of list head words than original narrative works in English. Hu (2016: 12, 22) reviewed a few studies focusing on translated Chinese that confirm Laviosa's (1998) findings. For instance, translated fiction is found to have lower lexical variety, lower lexical density, and a higher percentage of high-frequency words than non-translated fiction (Hu 2007). Similarly, Wen (2009) showed that translated detective fiction has lower type-token ratio, lower lexical density, and lower mean sentence length than non-translated detective fiction.

A number of studies, however, have not confirmed the simplification hypothesis. These studies reported contradictory findings with some parameters, such as higher mean sentence length (Laviosa 1998), untypical lexical patterning (Mauranen 2000), and overuse of degree modifiers (Jantunen 2004) in translated compared with non-translated texts. The study by Ferraresi *et al.* (2018) even rejected Laviosa's (1998) findings. For example, they found that French-English translated texts are not simpler than non-translated texts and that Italian-English translated texts are more complex than the non-translated ones in that they contain fewer common words and are also lexically denser. Lexical simplification is thus a controversial translation hypothesis.

Lexical simplification (or complexity) is also a research topic in SLA. It is considered an indicator of lexical proficiency and language production quality of L2 users (Bulté and Housen 2012; Lu 2012). Generally speaking, texts that are lexically more complex are associated with higher L2 proficiency (Laufer and Nation 1995; Jarvis 2002; Crossley and McNamara 2012). Controversial results were obtained when comparing the lexical complexity of texts produced by L2 writers with those by native speakers. Gonzalez (2013) found that native texts show significantly greater lexical diversity and a higher proportion of low-frequency words than non-native texts. Jarvis (2002) also suggested that native texts generally have higher lexical diversity than non-native texts. By contrast, Nasseri and Thompson (2021) compared academic writing produced by English native, ESL, and English as Foreign Language (EFL) students, and showed that the texts produced by EFL students have the lowest lexical density and diversity despite the fact that the English native and ESL groups produced texts with similar lexical density and diversity. These findings seem to suggest that other factors, such as L1 background, L2 instruction, and L2 proficiency, probably have an influence on the lexical complexity of L2 writing (Jarvis 2002: 57).

1.3. Research questions

Translation and L2 writing are forms of constrained communication. Simplification, a strategy dealing with constraints, is conceivably a feature that characterizes both of them. However, translation and L2 communication are different in that the former is "dependent text production" while the latter is "independent text production" (Lanstyák and Heltai 2012: 101). As student translators usually have sufficient L2 proficiency but are not fully competent in translation, they might experience different degrees of constraints which affect their use of translation strategies during text production.

The review of the literature above makes it clear that there is a gap in corpus research into L2 learner translation and L2 writing. To bridge this gap, the present study examines how L2 learner translation and L2 writing might converge or diverge in lexical simplification. In this study, we address two research questions:

RQ 1: How is L2 learner translation (dis)similar to L2 writing in the four lexical simplification parameters?

RQ 2: What are the possible factors that account for the (dis)similarities?

The findings are expected to provide a better understanding on how L2 learner translation is possibly influenced by translation and L2 (interlanguage) factors. As prospective professional translators, student translators are "major stakeholders in translator training" (Li 2002: 513). The current study is thus important for the learners' development of L2 translation competence.

2. METHODOLOGY

2.1. Corpora

This study adopts a corpus-based methodology to investigate the lexical simplification of L2 learner translation and L2 writing. The investigation is based on two comparable corpora: the *International Corpus of English in Hong Kong* (ICE-HK; Bolt and Bolton 1996; Nelson 2006) and the *Parallel Learner Translation Corpus* (PLTC) compiled at The Hong Kong Polytechnic University.

ICE-HK is an existing corpus initiated by Bolt and Bolton (1996) in the early 1990s. It is part of the *International Corpus of English* project (ICE) initiated by Greenbaum (1988). The project aimed to collect comparative English data representing different regional varieties of English. ICE-HK follows the general structure of ICE[2] worldwide to collect English data from the Chinese population in Hong Kong, whose first language is Cantonese and whose primary and secondary education is in Hong Kong. ICE-HK contains a wide range of text categories, including different communication modes (i.e., spoken and written) and registers (e.g., direct conversations, broadcast news, business letters, academic writing, and novels, among others). ICE-HK thus represents English as a second language (ESL) in Hong Kong (Nelson 2006).

PLTC is a learner translation corpus being compiled at The Hong Kong Polytechnic University.[3] It is constructed to match the composition in text categories and proportion in size as the written component (printed subcategory) of ICE-HK. The aim of PLTC is to document learner translated English in different written registers (e.g., academic writing, novels, etc.) in Hong Kong. The compilation of PLTC consists of a two-stage procedure: preparation of Chinese textual materials and collection of learner translations done from Chinese to English. In the first stage, qualified translators first translate ICE-HK texts from English into Chinese. Copyeditors then check and edit the texts for quality control. The edited Chinese texts are then further checked and approved by the translator and copyeditors together, and the approved versions are used as source texts for translation in the next stage. In the second stage, second-year to fourth-year undergraduate students majoring in translation at The Hong Kong Polytechnic University are invited to participate in the current study via mass email. Interested students sign the consent form to indicate their intention to participate in the study. The researchers then select the participants by taking into consideration their language and educational background. The eligible participants must speak Cantonese proficiently and receive secondary education in Hong Kong. So far, a total of 28 eligible students have been recruited as participants, as shown in Table 1.

[2] https://www.ice-corpora.uzh.ch/en.html

[3] PLTC is still under compilation. More participants will be recruited to produce additional translated texts for the corpus. It is anticipated that PLTC will have a more balanced distribution of participants and registers in its completed version. Further information about the corpus project may be found at https://cerg1.ugc.edu.hk.

	Students (*n* = 28)
Age (years)	20.3 (*SD* = 1.7)
Gender	
Female	27 (96.4%)
Male	1 (3.6%)
Education level	
Year 2	15 (53.6%)
Year 3	8 (28.6%)
Year 4	5 (17.9%)

Table 1: Demographic data of the participants in PLTC

1. Please, translate the essay from Chinese into English. The target audience are native English speakers who are interested in learning more about this essay topic.

2. There is no time or word limit.

3. You can use different tools, including books, dictionaries, and internet resources, to help you complete the translation, but you cannot consult others for any translation solutions.

4. Please, record the approximate time and all translation tools you use to complete the translation.

5. Please, use proper wording and grammar. Make sure that the translation is complete and appropriate.

6. The register of the essay is [*register and sub-register are provided*].

Registers and sub-register	
Academic writing. (Humanities)	Non-academic writing. (Humanities)
Academic writing. (Social sciences)	Non-academic writing. (Social sciences)
Academic writing. (Natural sciences)	Non-academic writing. (Natural sciences)
Academic writing. (Technology)	Non-academic writing. (Technology)
Reportage. (Press news reports)	Instructional writing. (Administrative writing)
	Instructional writing. (Skills and hobbies)
Persuasive writing. (Press editorials)	Creative writing. (Novels and stories)

Table 2: Translation brief (translated from Chinese)

Participants are provided with a written translation brief in Chinese. The brief, which is shown in Table 2, above, states that their task is to translate a Chinese text into English for native English speakers who are interested in learning more about the essay topic. It also stated what the register of the source text belongs to. Participants are also instructed to use any resources and tools they think they are useful to assist them with their translation without time and word limits. However, they cannot consult other people about translation solutions. Each participant translates one to four texts, depending on the willingness to continue with the study. In order to ensure the representativeness of the translated texts for each register, no participant is allowed to translate more than one text for each register. Besides, no participants are allowed to translate more than four texts to ensure participant/subject representativeness.

At the time of writing, 53 texts have been collected for PLTC. This study is based on these 53 text samples to represent the L2 learner translation corpus (henceforth L2T) and 53 corresponding text samples extracted from ICE-HK to represent the L2 writing corpus (henceforth L2W). L2T and L2W cover six major registers: academic writing, popular writing, reportage, instructional writing, persuasive writing, and creative writing. Each text contains around 2,000 words. L2T has a total of 125,178 tokens (total number of items) and 11,880 types (number of unique items), while L2W has 127,835 tokens and 12,704 types, as shown in Table 3.

Corpora	**Label**	**Nature**	**Files**	**Tokens**	**Types**	**Type-token ratio (TTR)**	**Standardized type-token ratio (STTR)**
PLTC	L2T	L2 learner translation	53	125,178	11,880	9.49	41.05
ICE-HK	L2W	L2 writing	53	127,835	12,704	9.94	41.47

Table 3: Composition of the corpora

2.2. Parameters and analysis

Following Laviosa (1998) and Ferraresi *et al.* (2018), we examined four parameters of lexical simplification: lexical density, standardized type-token ratio, core vocabulary coverage, and list head coverage. The operational definition of each parameter is stated in Table 4, below. In this study, several tools were employed to obtain the quantitative data. Both corpora were annotated using *Stanford CoreNLP* (Manning *et al.* 2014) to retrieve a part-of-speech tag for each word. It helped distinguish lexical words from running words. *WordSmith 8.0* (Scott 2021) automatically calculated the standardized type-token ratio (STTR) of each text, generated a list head word list for each corpus, and counted the number of core vocabulary and list head words of each text.

Parameters	Operational definitions (see Ferraresi *et al.* 2018)
Lexical density	It is used to evaluate the information load of a text. It is calculated by dividing the number of lexical words by the number of running words.[4] $= \frac{no.of\ lexical\ words}{no.of\ running\ words}$
Standardized type-token ratio	It is used to measure the lexical diversity of a text. It is obtained by calculating the ratio of the number of unique words to the number of running words on the basis of 1,000 words. $= \frac{no.of\ unique\ words\ (types)}{no.of\ running\ words\ (tokens)}$
Core vocabulary coverage	It is used to measure lexical diversity by exploring patterns of frequent word use of a text in comparison to an external reference. It is obtained in two steps: 1) by establishing a list of 200 most frequent words (core vocabularies) from a reference corpus —the written component of the *British National Corpus* (BNC; Leech *et al.* 2001) was selected (see Appendix 1), and 2) by calculating the proportion of core vocabularies to the number of running words. $= \frac{no.of\ core\ vocabularies}{no.of\ running\ words}$
List head coverage	It is used to measure lexical diversity by exploring patterns of frequent word use from an angle of internal corpus measure. Unlike Ferraresi *et al.*'s (2018) study, which performed the analysis at a sub-corpus level, this study measures frequent word use at the text level. This is achieved in two steps: 1) by creating a list with the 100 most frequent words (list head words) from each corpus examined in the study, namely L2T (See Appendix 2) and L2W (Appendix 3), respectively, and 2) by dividing the number head words of a text by the number of running words. $= \frac{no.of\ list\ head\ words}{no.of\ running\ words}$

Table 4: Operational definition of each parameter

Preliminary checks on the normality showed that the data of the core vocabulary coverage was normally distributed, so a paired t-test was run to examine if significant differences exist between the two corpora. However, the data of the remaining three parameters were not normally distributed, so Wilcoxon tests were used. Paired t-test and Wilcoxon tests were useful because L2T and L2W are related to each other. The source texts of L2T are the Chinese translation of L2W, that is, the translated texts of L2T are back translations of L2W. To illustrate the quantitative findings, some examples were extracted to supplement the quantitative results.

3. RESULTS

The descriptive data and differences between the two corpora in terms of the four parameters are summarized in Table 5. The data show that the lexical density of L2T (*M*

[4] Running words represent the total numbers of items. They are based on the unit of part-of-speech tagging, that is, each tagged word is regarded as one running word (excluding symbols, digits, and punctuations). Lexical words are nouns, verbs, adjectives, and open-class adverbs (those that end in *-ly,* except *only*).

= 57.77, *SD* = 4.51) is significantly higher than that of L2W (*M* = 57.27, *SD* = 4.47), *V* = 1014, *p* = .008. However, L2T is not significantly different from L2W in terms of standardized type-token ratio, core vocabulary coverage, and list head coverage. Standardized type-token ratio of L2T (*M* = 40.99, *SD* = 4.46) is slightly lower than that of L2W (*M* = 41.33, *SD* = 4.94). The core vocabulary coverage of L2T (*M* = 51.48, *SD* = 5.66) is slightly lower than L2W (*M* = 51.58, *SD* = 5.22). L2T (*M* = 47.21, *SD* = 3.86) also shows a greater but not significant list head coverage than L2W (*M* = 46.98, *SD* = 3.75).

Parameters	**L2T** (*n* = 53)		**L2W** (*n* = 53)		**Wilcoxon test**		**Paired t-test**	
	***M* (%)**	***SD***	***M* (%)**	***SD***	***V***	***p***	***t***	***p***
Lexical density	57.77	4.51	57.27	4.47	1014	.008*	/	/
Standardized type-token ratio	40.99	4.46	41.33	4.94	526.5	>.05	/	/
Core vocabulary coverage	51.48	5.66	51.58	5.22	/	/	-0.46	>.05
List head coverage	47.21	3.86	46.98	3.75	841	>.05	/	/

Table 5: A comparative analysis of L2T and L2W

The similarities and differences between L2T and L2W are illustrated in (1)–(3) below. Lexical density denotes information loaded on a text. Unlike function words, which mainly perform grammatical functions, lexical/content words carry semantic information. A high proportion of lexical words (emphasis added in the example) indicates that the text is packed with dense information. In example (1), L2T is lexically denser than L2W. In the example, L2T contains participial phrases, while L2W uses apposition in the second half of the sentence, resulting in lexical density differences between the two text varieties. A noun phrase in the apposition often needs a determiner (function word) to mark the noun, like ‘*an* insistence’ and ‘*an* emphasis’ (L2W), but a participial phrase does not. Also, when a noun phrase is modified by another noun phrase, a preposition (function word) is needed to express the modification, such as ‘an emphasis *on* moral sensitivity’ (L2W). For the participial phrase, a preposition is added after the participle only when the participle is an intransitive verb. In (L2T), ‘emphasizing’ is a present participle with a transitive nature which can be followed by a direct object, ‘moral sensitivity,’ without a preposition. Due to the above two reasons, the sentence in L2T is lexically more packed than in L2W.

(1) **L2T**

In both ethics, there seems to be a lack of universal rules or general principles, insisting that rules are not absolute and overriding everything, and emphasizing moral sensitivity above principles and moral reasoning.

L2W

In both ethics, there seems to be an absence of universal rules or general principles, an insistence that rules are not absolute and overriding, and an emphasis on moral sensitivity over principles and moral reasoning.

Standardized type-token ratio, core vocabulary coverage, and list head coverage are the measures of lexical diversity or variation. A high proportion of unique words and low repetition of common words indicate that the text is composed of a wide variety of vocabulary. Examples (2) and (3) show similar lexical diversity in all three parameters. This indicates that discrepancies between two sets of comparable sentences are not great enough to result in a significant difference in lexical diversity.

(2) **L2T**

In fact, she **was** just **trying** to **have a joke** on **the animal**, the most **innocent** kangaroo she had ever seen. She knew how fragile her life was, and she understood the rules of the forest. **With any** luck, if she **was not** eaten **today**, she **might** be eaten the next **day** by **some** careless animal or big bird. Anyway, she **thought**, **looking** up **at** the sky.

L2W

In fact, she just **wanted** to **play a trick** on **this creature**, the most **naive** kangaroo she had ever seen. She knew **exactly** how fragile her life was and she understood the rules of the forest. **By** luck if she **avoided being** eaten **one day**, she **could** be eaten the next, by **another** careless animal, or a big bird. Anyway, **as** she **was thinking**, **she looked** up **to** the sky.

(3) **L2T**

The **plight of** many of Hong Kong’s elderly is a **worrying reflection of** a society that has traditionally **given** great care to the **elderly**. A study commissioned by the government on the **condition** of **older** citizens **produced disheartening results**. In Hong Kong, 30 percent of suicides involve the **elderly**, even though they make up only 14 percent of the population. On average, one elderly **person** is reported to **have committed** suicide every 1.5 days.

L2W

The **unhappy conditions in which** many of Hong Kong’s elderly **live** is **cause for concern, and reflects poorly on** a society that has traditionally **taken** great care of the **aged**. A study commissioned by the Government on the **state** of **our senior** citizens **makes depressing reading**. Thirty percent of the suicides in Hong Kong involve the **aged**, even though they make up only 14 percent of the population. On **an** average, one **case of** elderly suicide is reported every 1.5 days.

By examining the examples closely, we find that major differences between them (emphasis added in the examples) can be categorized into grammatical factors (e.g., verb tense and the use of function words) and non-grammatical factors (e.g., lexical word choice). Non-grammatical issues are likely to be the determining factors of lexical diversity. From a grammatical perspective, the expressions of verb tense and function words (e.g., prepositions and determiners) often follow certain rules. They may not lead to considerable differences in the use of the vocabulary in both texts. However, if we focus on the non-grammatical factors, it can be noticed that many lexical words have synonyms, hypernyms, and hyponyms. They allow for more variations in word choice. If a text is characterized by more synonyms and hyponyms, its lexical diversity will naturally increase. In example (2), 'creature' (L2W) and 'animal' (L2T and L2W) are synonyms, and 'kangaroo' (L2T and L2W) and 'bird' (L2T and L2W) are hyponyms of 'animal'. These words carry related meanings and allow for some variations in word choice. In (2), sentences in both L2T and L2W seldom repeat the same word, leading to a high diversity. Example (3) provides another instance in which synonyms are used instead of the repetition of the same word. In (3), various adjectives which describe negative emotions, i.e., 'worrying' (L2T), 'disheartening' (L2T), 'unhappy' (L2W), and 'depressing' (L2W), are synonyms. In example (3), both L2T and L2W are characterized by the use of synonyms instead of a single word to express ideas of similar meanings. Besides, (3) mainly reports the situations of older adults. Both L2T and L2W use different expressions to indicate this meaning in the example: 'elderly' (L2T and L2W), 'aged' (L2W), 'older citizens' (L2T), and 'senior citizens' (L2W). Both (2) and (3) show that L2T and L2W have a similar variety of lexical words, resulting in equally high lexical diversity.

4. DISCUSSION

This study compared lexical simplification patterns between L2 learner translation (L2T corpus) and L2 writing (L2W corpus) using a corpus-based approach. The four lexical simplification parameters examined can be roughly classified into two broad categories: informativeness —i.e., lexical density— and lexical diversity —i.e., standardized type-token ratio, core vocabulary coverage, and list head coverage— (Ferraresi *et al.* 2018: 727; Xu and Li 2022: 10–11). The results demonstrate that learners' Chinese-to-English L2 translation is not lexically simpler than L2 writing. While there was no significant difference between the two corpora in all three lexical diversity parameters, L2 learner translation was found to be even lexically denser, i.e., more informative, than L2 writing.

Our study shows that simplification is not confirmed in learner translation when compared to L2 writing. The examples further show that lexical density may be related to the syntactic structure of the texts, while lexical diversity is likely associated with a variety of lexical/content words. In what follows, we will address the possible motivations for these findings from the perspectives of constrained communication, the language background of writers and translators, source language influence, and comparable corpus construction.

The degree of constraints may influence the lexical simplification patterns in communication. As mentioned in Section 1.2, the simplification hypothesis is regarded as one of the potential translation universals in comparison with non-translated native texts (Baker 1996). A possible explanation for this hypothesis is that translation is a form of constrained communication while non-translated native text production is not. Therefore, translated texts are simpler due to a higher cognitive load (Carl and Dragsted 2012) and risk minimization (Pym 2015) on behalf of the translator. Our study mainly focuses on the comparison of L2 writing and comparable translation done by student translators. Such a corpus design maximizes the degree of constraints in the way that language production in L2 becomes a major restriction shared in the two corpora. Our results show that L2 translation and L2 writing share more similarities than differences, highlighting the similar constraints faced by translators and writers who come from a similar background.

From a language background perspective, the texts of L2W and L2T are collected from Hong Kong writers and student translators. Hong Kong has a unique language environment due to its colonial history. Adding to Cantonese/Chinese (L1), English is also an official language in Hong Kong. English is a compulsory course for primary and secondary school students and remains the predominant language in professional settings, such as tertiary education, business, and law (Liu *et al.* 2022: 80). Therefore, Hong Kong English is often regarded as ESL rather than as EFL (Nasseri and Thompson 2021). Examples (1)–(3) also show how L2 writers and L2 learner translators in Hong Kong are able to use various synonyms to express ideas with similar meanings, which results in a variety of lexical words in the texts. Lexical diversity, an indicator of lexical simplification, positively correlates with language proficiency (Jarvis 2002; Crossley and McNamara 2012). We postulate that similar language proficiency and vocabulary knowledge of the L2 writers and L2 learner translators narrows the differences in terms of lexical patterns.

From a source language perspective, the features of translational language can be subject to source language variation. Ferraresi *et al.* (2018) revealed that texts translated from Italian are lexically denser than original written texts. In contrast, similar results are not observed in the texts translated from French. Ferraresi *et al.*'s (2018) findings suggest that the source language can influence translation activity and alter the lexical density of a text. In our study, lexical density is the only parameter that distinguishes L2 learner translation from L2 writing, and the major difference between the two is the variable of source texts. Therefore, lexical density is likely to be subject to source language influence. Learner translators may be influenced by the source language (Chinese) to produce lexically denser texts than L2 writers. In addition, according to Laufer and Nation (1995: 309), lexical density "depends on the syntactic and cohesive properties of the composition. Fewer function words in a composition may reflect more subordinate clauses, participial phrases and ellipsis." Example (1), above, shows that participial phrases in L2T make a sentence more lexically packed. In short, the source language seems to play a critical role in the syntactic patterns of translations. Since the comparison at the syntactic level is not the focus of this study, further investigation is needed to uncover the relationship between lexical density and syntactic properties of L2 translation and L2 writing.

From the perspective of corpus construction, the degree of comparability may affect the comparison between the two corpora. As House and Kádár (2021: 4–5) argue, "[w]henever we use corpora compiled by others, we need to consider whether the generic, temporal and other features of the corpora are actually comparable." Also, "[i]n any rigorous […] research the size and other features of the corpora investigated need to be as comparable as possible." In traditional comparable corpus-based translation research, researchers mainly compile the corpora by considering the comparability of the genre and size. This study further enhances the comparability of the two corpora by ensuring their semantic sameness, as the texts of L2T are back-translated from that of L2W, that is, both originate from the same source. This may explain why examples (1)–(3) sometimes show similar text structure or vocabulary use. The corpus design may contribute to enhanced similarities between L2 learner translation and L2 writing in the findings. On the other hand, since the corpora examined in this study are highly comparable, we can be more confident that their differences are likely due to translation factors.

5. Conclusion

This study provides a preliminary picture of the relationship between L2 learner translation and L2 writing through a lexical simplification prism. Our analysis can be summarized in three main points: 1) L2 learner translation is not lexically simpler when compared with L2 writing in the Hong Kong context, 2) lexical density of L2 learner translation is higher than that of L2 writing, and 3) L2 learner translation and L2 writing have similar lexical diversity. We have also discussed that factors such as the degree of constraints in communication, language background of writers and translators, source language, and comparable corpus design may play a part in the results. Through the comparison of L2 learner translation and L2 writing, the findings hint at how L2 learner translation might be influenced by the translation factor (reflected in the differences between the two corpora) and the L2 factor (reflected in the similarities between the two corpora). This can be important for enhancing translation learners' L2 translation competence.

Despite the findings, there are some limitations to the study. First, since the sample size is relatively small, the limited number of texts does not allow to analyze how register might be a possible factor in affecting the various simplification parameters. For future research on the topic, we plan to collect more texts in different registers and consider how register as a variable may affect lexical and syntactic simplification. Second, this study compared L2 learner translation with L2 writing only. Professional L2 translation needs also to be taken into account in order to address the simplification hypothesis properly. The comparison of learner and professional translations will show the extent to which the two differ in the lexical simplification parameters. Third, as lexical density is not only associated with the proportion of lexical words, but also with the syntactic structure of the texts, it is also worthwhile to examine syntactic structures to gain a better understanding of the simplification phenomenon underlying learner translation. All this represents an avenue for future research.

References

Baker, Mona. 1993. Corpus linguistics and translation studies: Implications and applications. In Mona Baker, Francis Gill and Elena Tognini-Bonelli eds. *Text and Technology: In Honour of John Sinclair*. Philadelphia: John Benjamins, 233–250.

Baker, Mona. 1995. Corpora in translation studies: An overview and some suggestions for future research. *Target* 7/2: 223–243.

Baker, Mona. 1996. Corpus-based translation studies: The challenges that lie ahead. In Harold Somers ed. *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*. Philadelphia: John Benjamins, 175–186.

Blum-Kulka, Shoshana and Eddie A. Levenston. 1983. Universals of lexical simplification. In Claus Færch and Gabriele Kasper eds. *Strategies in Interlanguage Communication*. London: Longman, 119–139.

Bolt, Philip and Kingsley Bolton. 1996. The International Corpus of English in Hong Kong. In Sidney Greenbaum ed. *Comparing English Worldwide: The International Corpus of English*. Oxford: Clarendon Press, 197–214.

Bowker, Lynne and Peter Bennison. 2003. Student translation archive: Design, development and application. In Federico Zanettin, Silvia Bernardini and Dominic Stewart eds. *Corpora in Translator Education*. Manchester: St. Jerome Publishing, 103–117.

Bulté, Bram and Alex Housen. 2012. Defining and operationalising L2 complexity. In Alex Housen, Folkert Kuiken and Ineke Vedder eds. *Dimensions of L2 Performance and Proficiency: Complexity, Accuracy and Fluency in SLA*. Amsterdam: John Benjamins, 21–46.

Carl, Michael and Barbara Dragsted. 2012. Inside the monitor model: Processes of default and challenged translation production. *Translation: Corpora, Computation, Cognition* 2/1: 127–145.

Chesterman, Andrew. 2004. Hypotheses about translation universals. In Gyde Hansen, Kirsten Malmkjær and Daniel Gile eds. *Claims, Changes and Challenges in Translation Studies*. Amsterdam: John Benjamins, 1–13.

Crossley, Scott A. and Danielle S. McNamara. 2012. Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading* 35/2: 115–135.

Duff, Alan. 1981. *The Third Language: Recurrent Problems of Translation into English*. Oxford: Pergamon Press.

Ferraresi, Adriano, Silvia Bernardini, Maja Petrović and Marie-Aude Lefer. 2018. Simplified or not simplified? The different guises of mediated English at the European parliament. *Meta* 63/3: 717–738.

Frawley, William. 1984. *Translation: Literary, Linguistic, and Philosophical Perspectives*. Newark: University of Delaware Press.

Gonzalez, Melanie. 2013. *The Intricate Relationship between Measures of Vocabulary Size and Lexical Diversity as Evidenced in Non-native and Native Speaker Academic Compositions*. Florida: University of Central Florida dissertation.

Granger, Sylviane and Marie-Aude Lefer. 2020. The Multilingual Student Translation Corpus: A resource for translation, teaching and research. *Language Resources and Evaluation* 54/4: 1183–1199.

Greenbaum, Sidney. 1988. A proposal for an international computerized corpus of English. *World Englishes* 7/3: 315. https://doi.org/10.1111/j.1467-971X.1988.tb00241.x.

Grosjean, François. 2013. Bilingualism: A short introduction. In François Grosjean and Ping Li eds. *The Psycholinguistics of Bilingualism*. Oxford: Wiley-Blackwell, 5–25.

House, Juliane. 2015. *Translation as Communication across Languages and Cultures*. London: Routledge.

House, Juliane and Dániel Z. Kádár. 2021. Introduction. In Dániel Z. Kádár and Juliane House eds. *Cross-Cultural Pragmatics*. Cambridge: Cambridge University Press, 1–12.

Hu, Kaibao. 2016. *Introducing Corpus-based Translation Studies*. Heidelberg: Springer.

Hu, Shirong. 2007. *A Corpus-based Study of the Translation Strategies Used in the Chinese Translations of Hamlet and Othello*. Shanghai: Shanghai Jiao Tong University dissertation.

Jantunen, Jarmo Harri. 2004. Untypical patterns in translations: Issues on corpus methodology and synonymity. In Anna Mauranen and Pekka Kujamäki eds. *Translation Universals: Do They Exist*. Amsterdam: John Benjamins, 101–126.

Jarvis, Scott. 2002. Short texts, best-fitting curves and new measures of lexical diversity. *Language Testing* 19/1: 57–84.

Kortmann, Bernd and Benedikt Szmrecsanyi. 2009. World Englishes between simplification and complexification. In Thomas Hoffmann and Lucia Siebers eds. *World Englishes – Problems, Properties and Prospects*. Amsterdam: John Benjamins, 263–286.

Kruger, Haidee and Bertus van Rooy. 2016. Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide* 37/1: 26–57.

Lanstyák, István and Pál Heltai. 2012. Universals in language contact and translation. *Across Languages and Cultures* 13/1: 99–121.

Laufer, Batia and Paul Nation. 1995. Vocabulary size and use: Lexical richness in L2 written production. *Applied Linguistics* 16/3: 307–322.

Laviosa, Sara. 1998. Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta* 43/4: 557–570.

Leech, Geoffrey, Paul Rayson and Andrew Wilson. 2001. *Word Frequencies in Written and Spoken English: Based on the British National Corpus*. Harlow: Longman.

Li, Defeng. 2002. Translator training: What translation students have to say. *Meta* 47/4: 513–531.

Liu, Kanglong and Muhammad Afzaal. 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, Kanglong, Joyce Oiwun Cheung and Nan Zhao. 2022. Learner corpus research in Hong Kong: Past, present and future. *Corpora* 17/Supplement: 79–97.

Lu, Xiaofei. 2012. The relationship of lexical richness to the quality of ESL learners oral narratives. *The Modern Language Journal* 96/2: 190–208.

Manning, Christopher D., Mihai Surdeanu, John Bauer, Jenny Finkel, Steven J. Bethard and David McClosky. 2014. The Stanford CoreNLP Natural Language Processing toolkit. In Kalina Bontcheva and Jingbo Zhud eds. *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*. Baltimore: Association for Computational Linguistics: 55–60.

Mauranen, Anna. 2000. Strange strings in translated language: A study on corpora. In Maeve Olohan ed. *Intercultural Faultlines. Research Models in Translation Studies 1: Textual and Cognitive Aspects*. Manchester: St. Jerome Publishing, 119–141.

McWhorter, John H. 2011. *Linguistic Simplicity and Complexity: Why Do Languages Undress?* Berlin: Mouton De Gruyter.

Nasseri, Maryam and Paul Thompson. 2021. Lexical density and diversity in dissertation abstracts: Revisiting English L1 vs. L2 text differences. *Assessing Writing* 47: 100511. https://doi.org/10.1016/j.asw.2020.100511.

Nelson, Gerald. 2006. *The ICE Hong Kong Corpus: User Manual*. London: University College London.

Olohan, Maeve and Mona Baker. 2000. Reporting *that* in translated English: Evidence for subconscious processes of explicitation? *Across Languages and Cultures* 1/2: 141–158.

Pym, Anthony. 2008. On Toury's laws of how translators translate. In Anthony Pym, Miriam Shlesinger and Daniel Simeoni eds. *Beyond Descriptive Translation Studies: Investigations in Homage to Gideon Toury*. Amsterdam: John Benjamins, 311–328.

Pym, Anthony. 2015. Translating as risk management. *Journal of Pragmatics* 85: 67–80.

Saldanha, Gabriela. 2011. Emphatic italics in English translations: Stylistic failure or motivated stylistic resources? *Meta* 56/2: 424–442.

Scott, Mike. 2021. *WordSmith Tools Version 8.0*. Stroud: Lexical Analysis Software.

Toury, Gideon. 2012. *Descriptive Translation Studies – and Beyond*. Amsterdam: John Benjamins.

Tymoczko, Maria. 1998. Computerized corpora and the future of translation studies. *Meta* 43/4: 652–660.

Wen, Tinghui. 2009. *Simplification as a Recurrent Translation Feature: A Corpus-based Study of Modern Chinese Translated Mystery Fiction in Taiwan*. Manchester: University of Manchester dissertation.

Xia, Yun. 2014. *Normalization in Translation: Corpus-based Diachronic Research into Twentieth-century English-Chinese Fictional Translation*. Newcastle upon Tyne: Cambridge Scholars Publishing.

Xu, Cui and Dechao Li. 2022. Exploring genre variation and simplification in interpreted language from comparable and intermodal perspectives. *Babel* 68/5: 742–770.

Corresponding author
Kanglong Liu
The Hong Kong Polytechnic University
Department of Chinese and Bilingual Studies
11 Yuk Choi Road
Hung Hom, Kowloon
Hong Kong SAR
China
Email: klliu@polyu.edu.hk

received: November 2022
accepted: February 2023

APPENDICES

Appendix 1: The 200 most frequent words extracted from the reference corpus. The written component of the BNC.

The
Of
And
A
In
To
Is
Was
It
For
That
With
He
Be
On
I
By
’s
At
You
Are
Had
His
Not
This
Have
But
From
Which
She
They
Or
An
Were
As
We
Their
Been
Has
Will
Would
Her
There
n’t
All
Can
If
Who
Said
Do
What
One
Its
Into
Him
Some
Up
Could
When
Them
So
Time
Out
My
Two
About
Then
No
More
Other
Also
Only
These
Me
First
Your
May
Now
Did
New
Any
People
Should
Than
See
Very
Made
Like
Just
After
Between
Many
Years
Way
How
Our
Being
Those
Such
Down
Make
Through
Over
Even
Back
Must
Know
Year
Own
Still
Because
Too
Get
Good
Three
Last
Take
However
Government
Work
Go
Man
Well
World
Same
Most
Life
Against
Day
Might
Under
Here
Does
Another
Come
Us
Think
Old
While
Never
Where
Each
Again
Found
Mr.
Part
Say
House
Much
Used
Out of
Number
Without
Going
Different
Children
System
Put
During
Within
Came
Although
Few
Local
Small
Before
Got
Social
’ll
Place
Case
Great
Off
Always
’ve
’m
’re
Why
Something
Group
Went
Want
Thought
Company
End
Party
Per cent
Women
Next
Both
Men
Find
Information
Important
Five
Took
National
Often
Every
State

Appendix 2: The most frequent words extracted from L2T

The
Of
To
And
A
In
Is
For
That
I
It
Be
’s
On
Are
Hong
Kong
Was
With
As
By
This
From
He
Not
At
Have
Will
Said
They
You
Or
Can
Has
But
My
An
Their
We
She
n’t
Also
More
If
When
One
People
Do
There
His
Were
Me
Government
Her
Two
Only
All
Which
Other
Would
Who
Been
Had
These
About
Time
Should
No
Mr.
After
Them
Because
Some
Years
So
Its
What
New
Than
Into
First
Could
Year
Out
Like
Business
China
Services
Up
System
Most
May
Many
Any
Public
Did
Still
Such
However
Between

Appendix 3: The most frequent words extracted from L2W

The
Of
To
And
A
In
Is
For
I
It
That
Be
On
As
With
’s
Was
By
Are
Hong
Kong
From
Have
At
Not
He
You
Or
This
An
They
Will
Has
Said
But
Their
Can
One
My
Were
We
n’t
She
More
His
Had
Which
Also
Mr.
When
If
Her
Me
There
Do
All
Would
People
Been
Who
Two
Government
About
So
Only
Its
Up
Time
Should
No
Other
Into
What
Than
After
Out
Years
Them
New
Could
Because
Some
These
First
Such
Business
May
Many
Services
Did
Year
System
Most
Says
Public
Any
Now
Last
Chinese
China