Universals in Machine Translation?

A corpus-based study of Chinese-English translations by *WeChat Translate*

Jinru Luo and Dechao Li
The Hong Kong Polytechnic University

By examining and comparing the linguistic patterns in a self-built corpus of Chinese-English translations produced by *WeChat Translate*, the latest online machine translation app from the most popular social media platform (WeChat) in China, this study explores such questions as whether or not and to what extent simplification and normalization (hypothesized Translation Universals) exhibit themselves in these translations. The results show that, whereas simplification cannot be substantiated, the tendency of normalization to occur in the WeChat translations can be confirmed. The research finds that these results are caused by the operating mechanism of Machine Translation (MT) systems. Certain salient words tend to prime WeChat's MT system to repetitively resort to typical language patterns, which leads to a significant overuse of lexical chunks. It is hoped that the present study can shed new light on the development of MT systems and encourage more corpus-based product-oriented research on MT.

Key words: Translation Universals, machine translation, *WeChat Translate*, simplification, normalization

1. Introduction

Along with the development of Machine Translation (MT) paradigms, phrase-based

statistical MT and neural network MT systems have made online MT systems available to different kinds of users and are now prevalent in the market (Schwartz, 2018). Some popular MT systems, such as the *EUROTRA Multilingual Translation System*, *Youdao*, *Google Translate*, and *Baidu Translate*, have been widely used by professional translators and ordinary users alike around the world (Cronin, 2013; Feng, 2018). In 2017, a similar translation system, *WeChat Translate*, was launched by WeChat, the most popular social media app in China with about one billion users. *WeChat Translate* is an advanced neural machine translation system developed by WeChat in cooperation with *Youdao Translate*, which can translate according to contexts (Liu, 2019). Given that a large number of users might use different languages to write and post contents, many social media platforms "now provide machine translations to allow posts to be read by an audience beyond native speakers" (Gupta, 2021: 96). *WeChat Translate* was born against this background. Considering the wide use of MT, machine translations have been widely acknowledged as new variants of translation (Cronin, 2013: 119; Lapshinova-Koltunski, 2015: 99).

The present study focuses on these new variants of translation, aiming to explore whether or not and to what extent MT systems, such as *WeChat Translate*, produce translations that exhibit common translational features (i.e., hypothesized Translation Universals) of human translations in general. Specifically, the present research attempts to answer the following questions: (i) Do simplification or normalization, two of the related hypothesized Translation Universals, exist in WeChat translations from Chinese to English? (ii) If one of these features can be found, then to what extent does it occur and what are the linguistic patterns of it? (iii) What are the underlying reasons for these features? It is hoped that this research will also serve as an investigation into the applicability of Translation Universals as parameters for the assessment of MT, so as to provide more insights into the future development of this increasingly important area of translation studies.

2. Translation Universals

Translation Universals, which refer to the common features of translated texts, were first discussed by Baker (1993), who pioneered corpus-based translation studies. Some of the universal features of translation that have been posited so far include explicitation, simplification, normalization, levelling out, the law of interference, and the unique item hypothesis (Baker, 1996; Laviosa, 2010; Laviosa, 2011).

Simplification, a term first put forward by Blum-Kulka and Levenston (1983), is a widely studied hypothesis of translational features in corpus-based translation studies. It refers to the fact that translators tend to "simplify the language or message or both" in translation (Baker, 1996: 176). It is usually examined by comparing translated texts with non-translated texts of a target language. Both Baker (1996) and Laviosa (1998; 2002) have summarized some statistical indicators that might suggest simplification, which usually include type/token ratio, lexical density, mean sentence length, list head coverage, core vocabulary coverage, and frequencies of hapax legomena (Laviosa, 1998; Bernardini et al., 2016; Hu, 2016). A number of recent studies based on comparable corpora (Grabowski, 2013; Kajzer-Wietrzny, 2015; Bernardini et al., 2016) have lent support to the existence of simplification in translated texts through the lens of different statistical indicators and language pairs.

Normalization refers to translators' tendency to "conform to patterns and practices which are typical of the target language, even to the point of exaggerating them" (Baker, 1996: 176). According to Baker, normalization can be reflected by the use of "typical grammatical structures, punctuation and collocational patterns or clichés" (1996: 183). However, the indicators used in recent studies of this feature are diverse. Most recent research can be classified into two types, according to indicators on phrasal levels and those on syntactic levels. The indicators on phrasal levels include creative expressions and peculiar collocations being normalized (Kenny, 2001), the use of coinages (William, 2005; Kruger & Van Rooy, 2012), the use of loan words and Anglicisms (Laviosa, 2006; Bernardini & Ferrasesi, 2011), frequencies of lexical

bundles (Kruger & Van Rooy, 2012), the proportionality of nominal phrases and verbal phrases (Lapshinova-Koltunski, 2015), frequencies of hapax legomena (Wang & Li, 2016), and frequencies of phrasal verbs (Cappel & Loock, 2017). Research based on indicators on syntactic levels includes studies of the *Ba* structure in Chinese (Hu & Zeng, 2011) and studies of prepositional phrase placement (Van Oost et al., 2016). The research methods used in these studies have developed from using monolingual comparable corpora to test hypotheses to a combined usage of monolingual comparable corpora and parallel corpora; the former is used for identifications of features and the latter for qualitative analyses and further explanations (Lapshinova-Koltunski, 2015; Van Oost et al. 2016; Cappel & Loock, 2017).

Although these recent studies prove that the claim of normalization does make sense and can be identified in many language pairs and different registers, some in the field still have doubts about the existence of normalization (Tikkonen-Condit, 2002). Questions have also been raised about the hypotheses of Translation Universals as a whole, especially the terminology used. For instance, both Toury (2004) and Chesterman (2004) considered that Translation Universals can at best present themselves as conditioned and probabilistic. Halverson (2003) pointed out that Translation Universals were second levels of generalization, whereas the cognitive basis of translation can offer a higher level of generalization and explanation. Even so, conditional claims about Translation Universals can still be valuable in the seeking of generalizations (Chesterman, 2004: 43). In this paper, the term Translation Universals is used with their probabilistic nature in mind and their current status as hypotheses acknowledged.

Up until now, almost all corpus-based studies of Translation Universals derive their findings from translations carried out by human translators; only a few are from MT (Lapshinova-Koltunski, 2015; Vanmassenhove et al., 2019; Zhang & Toral, 2019). It will therefore be interesting and worthwhile to discover whether online MT systems, which are not subject to similar cognitive influences as human translators, produce texts that share some features of Translation Universals as those exhibited in human

translations. Lapshinova-Koltunski (2015) is one of the first scholars to study the features of translation from English into German using corpora made up of both human translations and machine translations. In the study, she compared the lexical density and standardized type/token ratio (STTR) of German non-translated texts and German translated texts from both human and machine outputs. She concluded that simplification cannot be confirmed (Lapshinova-Koltunski, 2015). As for normalization, she used comparable corpora to study the proportion between nominal phrases and verbal phrases of both human translations and machine translations, as well as German non-translated texts before using the parallel corpora for evidence of further explanations. The results showed that normalization can be confirmed only as a feature of MT and not for human translation (Lapshinova-Koltunski, 2015). Though this study is valuable and pioneering, the operational indicators used only include lexical density, STTR, and the distribution of nominal and verbal phrases, which are not comprehensive enough to reveal the full linguistic profile of the data. Moreover, more diverse studies on different language pairs with more comprehensive indicators are needed to confirm the existence of Translation Universals in this type of translation.

3. Corpora and Methodology

This section outlines the data and methodology used in the present study. The compilation of the corpora is delineated in Section 3.1. The methodology employed to address the research questions set out in Section 1 is described in Section 3.2 .

3.1 Corpora Compilation

All the data making up the analysis corpus were contributed voluntarily by 120 college English majors from Sun Yat-sun University in China. The authors first invited these students to choose from their WeChat history certain texts that they were willing to share with the public for research purposes. The students checked and deleted private

The following publication Luo, J., & Li, D. (2022). Universals in machine translation? A corpus-based study of Chinese-English translations by WeChat Translate. International Journal of Corpus Linguistics, 27(1), 31-58 is available at https://dx.doi.org/10.1075/ijcl.19127.luo. International Journal of Corpus Linguistics is available at https://www.jbe-platform.com/content/journals/15699811

information (if any) before they submitted their data, and all the data were collected with their consent. They were then asked to use the *WeChat Translate* app to translate all of the chosen Chinese texts into English. Each student finally contributed English translated texts of about 2,000 words.

Two corpora were developed based on these raw data. The first is the WeChat Translation Corpus (hereafter referred to as the analysis corpus when the full name the WeChat Translation Corpus is not used), which consists of English texts translated by *WeChat Translate* (of 253,065 tokens in total). The second is a parallel corpus that includes the original Chinese texts and the English texts translated by *WeChat Translate*.

The authors also gave the students instructions to classify the text types of the translated texts contributed. The authors then manually checked the classification of all contributed texts. Table 1 shows the composition of text types in the WeChat Translation Corpus.

Table 1. Text types of the WeChat Translation Corpus

Formality	Text Types	Number of Pieces
Formal	Official instructions for academic activities (including lectures, homework and exams)	381
	Descriptions or explanations of disciplinary knowledge	27
	Essays	14
	Introduction to lecturers or colleges	13
	Job descriptions	11
	News	4
	Total	**450**
Informal	Casual talk (between people with close relationships about daily issues)	460
	Informal notices (from students to students, with simple and short sentences or typical spoken words)	148
	Homework discussions (between classmates)	27
	Informal speech (about casual online gatherings, with simple and short sentences or typical spoken words)	7
	Total	**642**

As shown in Table 1, the WeChat Translation Corpus is composed of two types of texts:

formal and informal. In the formal text types, there are six sub-types: official instructions about academic activities, descriptions or explanations of disciplinary knowledge, essays, introductions to lecturers or colleges, job descriptions, and news. The classification of these texts as formal is based on the categorizations of formal and informal text types by Joos (1967). All six sub-types share some basic features of formal texts, having long or complex informative sentences, participant detachment and standard language uses. Example (1) below is part of one text extracted from the WeChat Translation Corpus. As an explanation of the subject of "teaching Chinese as a foreign language", this text showcases some of the usual features of formal language mentioned above, including the informative function, personal detachment, and long and complex sentences ("He even expressed the hope that…and that…so that…"). The proportion of formal texts in the corpus data is high because WeChat functions not only as a social media platform but also as a platform for knowledge sharing.

(1) The enthusiasm for Mandarin is high among the British, and former Prime Minister David Cameron was active in promoting it to British students during his tenure. He even expressed the hope that the British would be able to keep in touch with the world's fastest-growing economies, and that attention should be shifted from traditional French and German to Chinese, so that Chinese could be learned in future business.

In the informal types, there are four sub-types: casual talk, informal notices, homework discussions and informal speech. The four sub-types of informal texts share some basic linguistic features, including short, simple, or even broken sentences, and words including typical spoken language features, such as *Haha* (an onomatopoeic sound of laughter in Chinese), *La* (a modal particle) and so on. In China, WeChat as an Instant Messaging (IM) tool has increasingly become a preferred mode of communicating, evolving into a textualized form of oral communication (Mao, 2014). Thus, part of the language used in IM tools bears certain features of spoken language (Tagliamonte & Denis, 2008; Sánchez-Moya & Moya, 2015). Example (2) below is part of a casual conversation extracted from the WeChat Translation Corpus. This example demonstrates some common features of spoken language, including short and simple

sentences, words from spoken language (*Hahaha* and *Ha-ha-ha*), no punctuation (the second line), and a broken sentence (*Playing that group game?*). Apart from the common linguistic features, all of the four sub-types of texts are communicated between people of close or equal relationships on casual occasions, resulting in the casual and informal nature of the language being used.

(2) -- Last night, you were on stage at the Magic Guild, right? Hahaha
-- Stop talking
-- It's embarrassing
-- Ha-ha-ha. It's okay
-- I'm embarrassed, too. Playing that group game?
-- The last one out!

In summary, the constitution of the different text types in Table 1 shows that the WeChat Translation Corpus is a hybridization of formal and informal registers. As Teich (2003) pointed out, researchers should choose a register-controlled corpus as the reference corpus to ensure closer comparability between the analysis corpus and the reference corpus, so as to produce more trustworthy results. However, in reality, the most common practice is to use a reference corpus of limited comparability (Bernardini & Ferraresi, 2011). In the present study, the authors also used a reference corpus of limited comparability. The reference corpus for this study is one half of BNC Baby (two million tokens), which is composed of both academic discourse and transcriptions of spoken language. In other words, this corpus is also a combination of formal and informal registers, thus ensuring comparability with the analysis corpus. The reference corpus is referred to as BNC Baby (Part) (hereafter referred to as the reference corpus when the full name BNC Baby (Part) is not used).

3.2 Methodology

The present study combined the indicators for simplification used in research by Laviosa (1998) and Kajzer-Wietrzny (2015). Laviosa (1998: 8) proposed that the core

patterns of simplification in translated texts included lower levels of lexical density, higher proportions of high-frequency words, and larger list head coverage with fewer lemmas. 'List head coverage' refers to the "proportion of the sum of the hundred most frequent words" in the analysis corpus to the total sum of words in the whole analysis corpus (Kajzer-Wietrzny, 2015: 243). The operational indicators of simplification for the present study include STTR, lexical density (content word proportion in the overall corpus), list head coverage (top 100 words coverage and top 200 words coverage), the number of lemmas in list heads, and the proportion of the 200 most frequently used English words. To calculate STTR, we used the corpus tool *WordSmith Tools 6.0* (Scott, 2012). We calculated the lexical density by using the corpus tool *Wmatrix 4.0* (Rayson, 2008; https://ucrel-wmatrix4.lancaster.ac.uk) to obtain the overall POS frequencies of the WeChat Translation Corpus and BNC Baby (Part) respectively. Then, the sums of content words (nouns, adjectives, adverbs, and non-auxiliary verbs) were counted and divided by the sums of the content words and function words to produce the content word proportions in the analysis corpus and the reference corpus respectively. The list heads were obtained from the word lists produced by *Wordsmith Tools 6.0*. We not only calculated the proportion of the sum of the hundred most frequent words (Top 100 words coverage) in the analysis corpus and the reference corpus respectively, but also calculated the respective Top 200 words coverages in the same way for a more detailed comparison. A lemmatization of the list heads was manually performed to obtain the lemmas of the list heads, since the number is manageable. Finally, the 200 most frequently used English words were taken from the list of the 200 most frequently used words in English according to Stubbs (1996: 36–37; see Appendix A). This list has been considered as the core vocabulary of English in later research (Kajzer-Wietrzny, 2015). Mean sentence length is not used as an index for the present research, as some data from the WeChat Translation Corpus do not have any ending punctuation, a typical feature of WeChat language (Mao, 2014).

The indicators used for the research on normalization include three different sets. The first is the proportion between nominal phrases (including prepositional phrases)

and verbal phrases. Lapshinova-Koltunski (2015) studied the proportions of nominal phrases (including prepositional phrases) and verbal phrases as an index for normalization or source language interference. The prepositional phrases were included as nominal phrases in her study according to the knowledge of contrastive linguistics regarding the German-English language pair (Lapshinova-Koltunski, 2015: 98; 106). Similarly, in translations from Chinese to English, some verbs used consecutively in one sentence are often nominalized as noun phrases or prepositional phrases, so as to conform to the rule of verb usage in English (Lian, 2010: 133). Both noun phrases and prepositional phrases are considered as nominal phrases in these two studies.

The present study adopted the index of proportion between nominal phrases (including prepositional phrases) and verbal phrases (hereafter referred to as the 'NV Ratio'). To calculate the NV Ratio, we used *Wmatrix* to obtain the POS frequencies of both the analysis corpus and the reference corpus, in order to further extract occurrences of nominal phrases, prepositional phrases, and verbal phrases. The sum of the nominal phrases and prepositional phrases was divided by the sum of the verbal phrases to obtain the NV Ratio. The second set is the correction of punctuation, specifically exclamation and question marks. Both Vanderauwera (1985) and Baker (1996) pointed out that the use of punctuation in translated texts tends to conform to traditional norms of the target language. However, almost all previous studies on normalization, such as those mentioned in Section 2, have focused on language features, rather than punctuation. Only Vanderauwera (1985: 94–95) studied the changes made to different punctuation in translated novels to conform to the norms of the target language. With the development of social media platforms, punctuation has become an important device for meaning making; it works side-by-side with language itself in many texts (Page, 2012; Wang et al., 2022). Although previous studies on normalization have bypassed punctuation, punctuation deserves more scholarly attention, as there have been increasingly diverse uses of punctuation as expressions of linguistic meanings among social media users (Wang et al, 2022: 1240). Since the raw data of the analysis corpus is extracted from the social media WeChat, it is likely that the diverse use of punctuation

may occur more often as part of linguistic data in the analysis corpus than in the reference corpus. For the second set of indicators, exclamation and question marks were chosen, as they are favored among young language users (Page, 2012). The third set is the overuse of typical grammatical structures and lexical bundles, an index following Baker's (1996) conceptualization of normalization.

4. Findings and Discussion

This section presents the findings and discussion. The results for simplification and normalization are introduced in Section 4.1 and Section 4.2, respectively. Section 4.3 presents a discussion of relevant results.

4.1 Results for simplification

As shown in Table 2, the WeChat Translation Corpus and the reference corpus are not comparable in terms of token sizes. Therefore, STTR is used instead of TTR for a more valid comparison. The STTR for the WeChat Translation Corpus is 40.60, which is higher than the BNC Baby (Part) value of 36.42. The lexical density (namely, the content word proportion in the overall corpus) is 58.31% for the analysis corpus, whereas the lexical density of the reference corpus is 54.81%. By contrast, the content word proportion in the WeChat Translation Corpus is a little higher than that in BNC Baby (Part). STTR and lexical density are indices for lexical richness and informativeness. The higher value of these two indices in the analysis corpus means that the analysis corpus is lexically denser, with more content words.

Table 2. Indicators for simplification

Items	WeChat Translation Corpus	BNC Baby (Part)
Tokens	253,065	1,917,751
STTR	40.60	36.42
Lexical density (overall content word	58.31%	54.81%

proportion)		
List head coverage (top 100)	51.12%	51.73%
List head coverage (top 200)	60.61%	60.10%
200 most frequent words coverage	55.63%	57.49%

In terms of the list head coverages, the top 100 words in the word list of the WeChat Translation Corpus cover 51.12% of the total words, with 85 lemmas (see Appendix B), while the top 200 words cover 60.61% of the total, with 167 lemmas (see Appendix B). For BNC Baby (Part), the top 100 words list head covers 51.73% of the total, with 78 lemmas (see Appendix C), while the top 200 words list head covers 60.10% of the total, with 154 lemmas (see Appendix C). There are more lemmas in both the top 100 and top 200 list heads of the analysis corpus. Figure 1 presents the visualized results of the similarity of the list head coverages between the two corpora. It can be seen in Figure 1 that the analysis corpus and the reference corpus share very similar percentage rates in terms of their list head coverages, from the top 100 list head to the top 200 list head. These similar results reflect the way in which the analysis corpus and the reference corpus share fairly similar rates of lexical repetitiveness. However, there are more lemmas in the list heads of the analysis corpus, resulting in greater lexical variety than in the reference corpus. Stubbs' (1996) list of the 200 most frequently used words in English is considered to cover the core vocabulary of English. The proportion of this core vocabulary reflects the lexical sophistication of a corpus (Kajzer-Wietrzny, 2015: 248). The greater the value of the proportion is, the narrower the vocabulary range of the corpus. The proportion of these 200 most frequently used words in the analysis corpus is 55.63%, whereas the proportion of these words in the reference corpus is 57.49%. The lower value of this core vocabulary proportion in the analysis corpus means that there is a broader range of vocabulary in the analysis corpus.

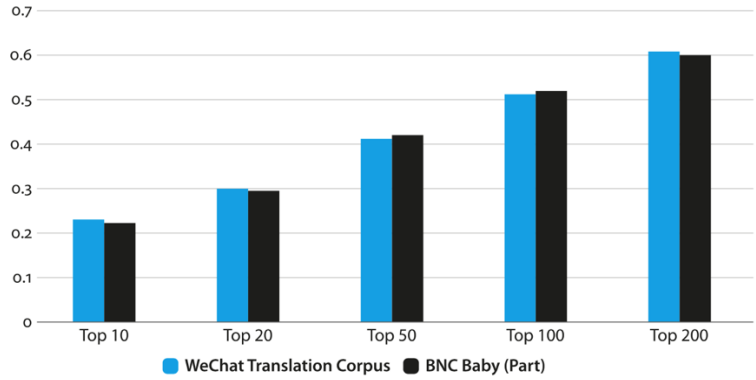

Figure 1. Coverage rate of list heads

In summary, the above statistical results cannot lend support to the tendency of simplification in the analysis corpus when compared to the reference corpus. The higher STTR and lexical density reveal a greater lexical richness. The similar list head coverages, with more lemmas in the analysis corpus, do not support the hypothesis that translated texts tend to have a greater percentage of repetition, with fewer lemmas. The lower core vocabulary proportion in the analysis corpus demonstrates broader lexical range and greater lexical variety. These results echo the findings of Lapshinova-Koltunski's study (2015), which could not confirm simplification tendencies.

4.2 Results for Normalization

This section introduces the results for normalization with regard to the NV ratio (4.2.1), the correction of punctuation (4.2.2), and the overuse of typical grammatical patterns (4.2.3).

4.2.1 *NV Ratio*

Table 3 shows the proportions of the nominal phrases (nominal and prepositional phrases) and the verbal phrases in the corpora. The results indicate that the nominal phrases occupy a greater proportion in the WeChat Translation Corpus, as its NV Ratio

(1.53) is greater than that of BNC Baby (Part) (1.01).

Table 3. Nominal phrase and verbal phrase proportion (NV Ratio)

Items	WeChat Translation Corpus	BNC Baby (Part)
Nominal phrase frequencies	66,392	411,213
Verbal phrase frequencies	43,395	405,411
NV Ratio	1.53	1.01

Chinese is a language with more frequent uses of verbs, while English is a language with more prominent uses of nominal phrases, which may further lead to frequent uses of prepositional phrases (Lian, 2010: 133). By contrast, nominal phrases take up a greater proportion of the analysis corpus than the reference corpus, which indicates that *WeChat Translate* renders more nominal and prepositional phrases in the translated texts instead of producing translations with more verbal phrases as influenced by the source language. This feature conforms to the grammatical rules of the English language. In other words, the normalization tendency is confirmed. This greater proportion of nominal phrases may also indicate that shifts from verbal phrases to nominal phrases have occurred in the translation process. Example (3) below is extracted from the parallel corpus.

(3) 我们学校没有“教务处”，只有“教务与科研部”，此外教务与科研部截止目前是不会以个人邮箱向学生个人直接发送电子邮件的，请各位知悉。
(There is no “Academic Affairs Office” in our school, only “Academic Affairs and Scientific Research Department”. Besides, the Academic Affairs and Scientific Research Department will not send any direct e-mail to students in their personal email address at present. Please be informed.”

In the Chinese sentence, the two words that are underlined – “只有” (*Zhi You*) and “截止” (*Jie Zhi*) – are classified as verbs. However, the verbal phrase with the verb “只有” (*Zhi You*) is translated as “only”, plus a noun phrase, while the verbal phrase with the verb “截止” (*Jie Zhi*) is translated as the prepositional phrase “at present” in the English translation. The two verbs have been changed to nominal phrases during the translation

process, a clear indication of the normalization tendency. Although the translations of these two phrases may not be obligatory choices, these choices under the mechanism of MT could become the most likely options, as explained in Section 4.3.

4.2.2 *Correction of punctuation*

As mentioned in Section 3.2, one typical feature of WeChat language is its sparing use of ending punctuation (Mao, 2014), a characteristic retained in *WeChat Translate* outputs. However, when the punctuation involves exclamation marks and question marks, *WeChat Translate* will automatically correct the abnormal use of punctuation. Figure 2 is a screenshot of *WeChat Translate* in operation. *WeChat Translate* operates in a dialogue model. Language users can type in the source text via the chat box. For example, all the lines with the green background are source texts that have been typed in. Once the *WeChat Translate* function is selected, the translation will be produced automatically under the relevant source text. As Figure 2 shows, the first line of the English translation (white background) retains the lack of ending punctuation from the original line (first line in green). This is also true for the third English translation (the third sentence in white). Both the first and third sentences in the source language are declarative sentences. However, when these two original sentences included question indicators, "吗" (*Ma*) and "吧" (*Ba*), as shown in the second and fourth sentences in the source text (in green), the two sentences were translated as questions. *WeChat Translate* renders the unpunctuated sentences with correct question marks, as is shown in the second and fourth English translations (second and fourth sentences in white).

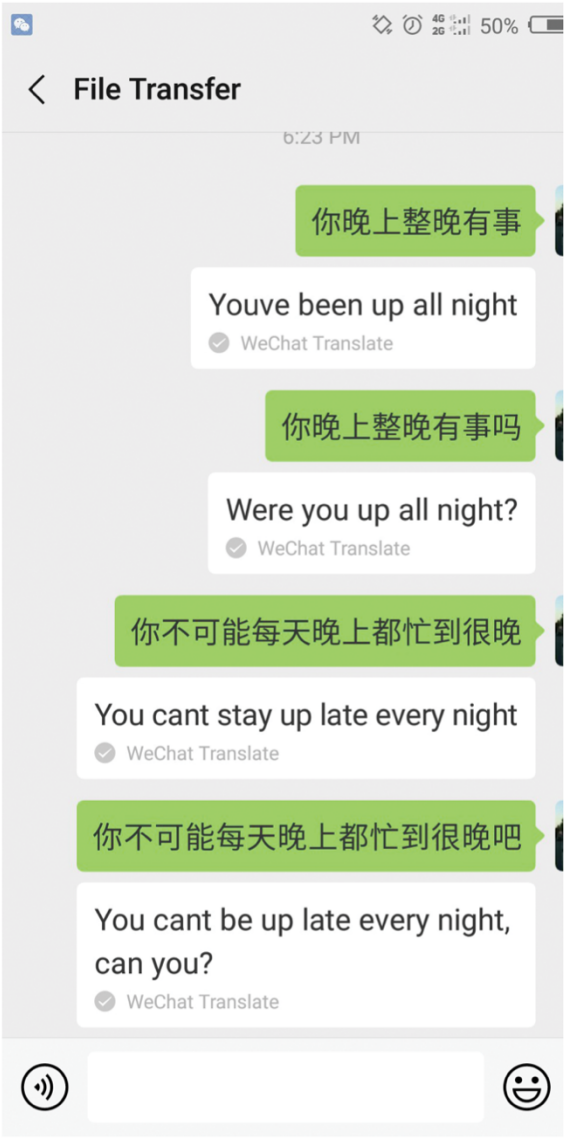

Figure 2. A screenshot of translations by *WeChat Translate*

Table 4 summarizes all the cases in the parallel corpus when the unpunctuated Chinese sentences with the co-occurrence of the words from (a) to (q) were identified as questions in the source text and *WeChat Translate* corrected the lack of ending punctuation by adding question marks to the translated text. The words from (a) to (q) in Table 4 are typical indicators of questions in Chinese. In standard Chinese question forms, these words should be accompanied by a question mark. Thus, question forms without question marks in these cases are abnormal and are usually considered as unfinished.

Table 4. Instances of correction of punctuation marks

Word Indicators of Questions	Instances of adding question marks to questions without punctuation
(a). 吗 (*Ma*)	383
(b). 吧 (*Ba*)	113
(c). 啊 (*A*)	67
(d). 呢 (*Ne*)	26
(e). 嘛 (*Ma*)	22
(f). 哈 (*Ha*)	14
(g). 呗 (*Bei*)	9
(h). 呀 (*Ya*)	5
(i). 谁 (*Shei*)	39
(j). 几 (*Ji*)	26
(k). 什么 (*Shen Me*)	110
(l). 还是 (*Hai Shi*)	38
(m). 多少 (*Duo Shao*)	22
(n). 有没有 (*You Mei You*)	33
(o). 是不是 (*Shi Bu Shi*)	27
(p). 会不会 (*Hui Bu Hui*)	7
(q). 够不够 (*Gou Bu Gou*)	3

Vanderauwera (1985) found that unorthodox punctuation was changed to conform to the traditional norms of the target language. Baker (1996) pointed out that both rounding off unfinished sentences and grammaticizing ungrammatical sentences were signs of normalization. The normalization tendency in this sense is actually what Chesterman (2004: 39–40) termed "S-Universals": features reflecting how translators process the source text in translation processes. Similarly, in this example, by adding question marks to the questions with no ending punctuation, *WeChat Translate* rounds off the unfinished questions in the process and normalizes the questions forms in the production.

Another abnormal use of punctuation that is normalized by *WeChat Translate* is the use of exclamation marks. Table 5 demonstrates how the excessive use of exclamation marks in the source texts was reduced by *WeChat Translate* in translation.

Table 5. Reduction of overuse of exclamation marks

Overuse of Exclamation Marks in Originals	Reduction in Translation	Instances
!!!!!!!!	!!!!!	3

!!!!!!!	!!!!	4
!!!!!	!!!!	2
!!!!	!!!	3

In Table 5, 12 instances of such reductions were summarized. This finding reveals that *WeChat Translate* tends to reduce the overuse of this punctuation mark when more than three exclamation marks are used together. Overuse of punctuation marks is also a feature of IM language (Mao, 2014), especially among young female language users, who are inclined to overuse punctuation marks as an expression of their emotion (Page, 2012). The overuse of exclamation marks may be a new linguistic phenomenon that is closely associated with the widespread use of social media platforms. In translation, *WeChat Translate* still conforms to the norms of punctuation use in standard English by reducing the overuse of it to a general level of acceptance.

4.2.3 *The Overuse of Typical Grammatical Patterns*

The third set of indicators of normalization found in the corpus is the overuse of certain typical grammatical structures. The five patterns in Table 6 were found when we studied the abnormal uses of question marks. They are words or lexical chunks commonly used in typical question forms, either general questions or *wh*-questions (Bo, 2002: 480–485).

Table 6. Overuse of grammatical patterns

Patterns	WCTC (Norm. Freq.)	BNC Baby (Part) (Norm. Freq.)	Loglikelihood	Significance
1. Why don't…?	570.06	43.60	308.99	0.0000
2. A(a)ren't you	187.18	82.64	16.50	0.0000
2(a). Aren't you…?	106.35	16.73	37.37	0.0000
2(b). …, aren't you?	80.83	65.91	0.27	0.6017
3. …, all right?	29.78	0.51	24.65	0.0000
4. Huh?	574.31	37.01	335.04	0.0000
5. …, right?/Right?	170.17	40.56	40.48	0.0000

As shown in Table 6, the WeChat Translation Corpus (WCTC) displays an exaggerated use of five grammatical patterns. The frequencies in Table 6 are normalized frequencies

per million. The results of a loglikelihood test reveal that the differences in the frequencies of all patterns (except Pattern 2(b)) are statistically significant. By comparison, WeChat Translation Corpus displays a greater frequency of the use of the first question pattern *Why don't…?* with 570.06 occurrences per million, which is much greater than that in BNC Baby (Part) (43.60). For the second pattern, the WeChat Translation Corpus shows a more frequent use of the question form *Aren't you…?* with 106.35 instances, while there are only 16.73 instances in BNC Baby (Part). For the third pattern, *all right?*, WeChat Translation Corpus also demonstrates a greater frequency (29.78) than BNC Baby (Part) (0.51). There are 574.31 instances in the analysis corpus and 37.01 instances in the reference corpus for the frequency of the fourth pattern, *Huh?*, a modal particle frequently used in question utterances. Finally, for the fifth pattern, *…, right?/Right?*, there are 170.17 instances in the WeChat Translation Corpus and 40.56 in BNC Baby (Part). In summary, the statistical results of these five grammatical patterns lend support to Baker's (1996) proposal of an indicator of normalization, namely, the exaggerated use of typical grammatical patterns in translated texts.

Moreover, the grammatical patterns under study are commonly used in question forms. The frequent use of them is related to another normalization tendency mentioned in Section 4.2.2: the addition of question marks to questions without ending punctuation. Below are some extracted sentences from the parallel corpus.

(4) 你还是去百度云吧
(*Why don't* you go to Baidu Cloud?)

(5) 嗯嗯，你们快到点军训了吧
(Well, *aren't you* almost at your military training?)

(6) 我把教案弄完再过去看你们哈
(I'll come by after I've finished the lesson plan, *all right?*)

(7) 这个三脚架好像不是蓝牙控制的哈
(I don't think this tripod is Bluetooth-controlled, *huh?*)

(8) 平常程度的危险吧

(Normal level of danger, *right?*)

Example (4) to Example (8), do not contain any ending punctuation in the source texts. Yet, there are some commonly used words as indicators of questions, such as "吧" (*Ba*) and "哈" (*Ha*), both of which are presented in Table 4. When presented with these indicators, *WeChat Translate* decodes the sentences concerned in the source language as questions, and thus renders them into question forms with typical question formats.

In summary, the NV Ratio (the proportion of nominal phrases and verbal phrases) proves that more nominal phrases than verbal phrases are produced in translations by *WeChat Translate* when compared to the reference corpus, indicating that the source language features have been adjusted in translation to conform to standard English written forms. The correction of punctuation and the overuse of grammatical patterns in the translation clearly reveal that abnormal question sentences in the source language have been normalized in translation, which again testifies to the tendency of normalization in the analysis corpus.

4.3 Discussion

As summarized in Section 4.1, the WeChat Translation Corpus displays greater lexical informativeness, richness, and variety. These results cannot confirm the existence of simplification, but can we draw the conclusion that simplification tendencies cannot be found in MT? To answer this question, we first need to look at the commonly proposed causes for simplification. Hypothesized simplification is probably caused by the principle of least effort, which refers to the fact that translators tend to, consciously or unconsciously, use "concise expressions, commonly used words and simple sentences", so as to exert less effort during the translation process (Hu, 2016: 103). However, this explanation is clearly not applicable to MT, the working mechanism of which operates differently from human beings. All of the currently dominant phrase-based statistical

machine translation systems (SMT) and the latest machine translation systems based on neural networks (NMT) can be broadly categorized as statistical corpus-based approaches to MT (Schwartz, 2018: 171–178), which generally follow two basic steps in the translation process: analysis (decoding) and generation (encoding). The SMT has one more transfer step in between the two (Schwartz, 2018: 171–173). Since the operating systems of MT enable them to work within a very short amount of time in both steps, they would not experience the same problems regarding effort as human translators do during the translation process. Therefore, it is reasonable to state that the translated texts produced by *WeChat Translate* cannot reflect the same tendency of simplification as those produced by human translators.

Moreover, given that the generation step of MT systems is based on pre-existing parallel corpora as references, it is natural that their translated texts are influenced by the composition of the reference parallel corpora. In the current study, the texts translated by *WeChat Translate* are influenced by the reference parallel corpora built into its system. This factor may also be the cause of the result that simplification cannot be confirmed in this study. Some scholars view the cognitive basis in translation processes as a deeper reason for the presence of hypothesized Translation Universals (Halverson, 2003; Szymor, 2018) and Lapshinova-Koltunski's study (2015), as well as our own, cannot lend support to the existence of simplification in the corpora of MT outputs. Therefore, we hypothesize that simplification does not exist in the corpora of machine translated texts, because MT is not subject to the same cognitive mechanism as human translators. More studies on different MT systems need to be conducted in order to confirm or refute our hypothesis.

All of the operational indicators examined above point to the fact that a normalization tendency exists in the translations carried out by *WeChat Translate*. Scholars have proposed that normalization in translation is mainly caused by the need to conform and the actual practice of conforming to the norms of target languages and cultures, especially when the target language enjoys a more superior status in society, as it can exert stronger influences on translators than the source language does (Baker,

1996; Toury, 2012). Although MT systems operate differently from human translators, there are similar reasons for the existence of the normalization of their translation products. To begin with, the analysis and generation steps of MT are based on its algorithm, the reference parallel corpora, and the machine learning system (Hu & Li, 2016: 10). To ensure the quality of MT, the reference parallel corpora must be carefully chosen to include as many correct translation examples as possible. The scale and quality of the data of the reference parallel corpora (built into the systems) are influential factors of the quality of the SMT and NMT systems (Feng, 2020). The reference parallel corpora of *WeChat Translate*, which makes full use of Youdao's developed online translation system, could be general corpora that are constituted by standard English with conventional language features. The standard English data forms the language model to be simulated and worked with alongside the reference parallel corpora in the algorithm of MT in translation (Feng, 2021). In SMT, the language model that the algorithm refers to may be manually designed and engineered, while, in NMT, the referred language model is the result of machine learning based on the reference parallel corpora (Feng, 2021). Either way, the language model bears the norms and conventional features of the target language. The algorithm in MT systems will simulate human language models and choose the most likely option in the generation step (Feng, 2021). As a result, it is highly probable that the texts translated from Chinese to English by the MT system of *WeChat Translate* will present typical features of standard English simulated by its language model, such as the proportion of nominal phrases and verbal phrases.

The correction of the abnormal use of punctuation marks found in the corpus is caused by the mechanism of *WeChat Translate* and the lack of such abnormal use in its language model. Both the use of no ending punctuation and the overuse of exclamation marks are two newly emerging language features from computer-mediated communication contexts during the most recent decade (Page, 2012; Mao, 2014). Compared to the built-in reference parallel corpora, which consist of data of correct translations and standard language models, the newly emerging use of syntactic patterns

may be considered as abnormal by WeChat's MT system. When the reference parallel corpora and the language model are processed by the algorithm in the translation process, standard language use will be produced as the most likely option and, thus, abnormalities will be corrected.

The mechanism of MT stipulates that some words may be given heavier weight toward the generation of certain optimal translation choices. For example, in Example (3) in Section 4.2.1, the verb phrase with the verb "只有" (*Zhi You*) is translated as "only" plus a noun phrase, and the verb phrase with the verb "截止" (*Jie Zhi*) is translated as "at present". Such translation choices could be translation equivalences that already exist in the reference parallel corpora of WeChat's MT system. When the MT system decodes a text with similar usage of these two verbs, the already existing translation equivalences will be processed and given heavier weight as the most likely option for the final output. Even though such translation choices are not obligatory, they could be the most optimal options. For another example, as shown in Table 4 in Section 4.2.2, the unpunctuated questions translated correctly with question marks are sentences with salient question indicators. These indicators place heavier weight on the generation of questions in SMT's transfer steps, or heavier weight in the "potentially language-independent vector" toward the generation of questions in NMT's encoding steps (Schwartz, 2018: 178). Such weight is given by the MT system according to the data in the reference parallel corpora and the rules of machine learning. Similar to the operation of lexical priming (Hoey, 2005), these salient word indicators with heavier weights prime the generation of question patterns as the optimal choice in MT. Therefore, *WeChat Translate* renders these unpunctuated sentences as questions according to the standard formats in its language model.

Furthermore, in Table 4, Instances (a) to (h) are Chinese modal particles that are typical in questions, especially in questions asking for information or confirmation, e.g. "吗" (*Ma*), or question forms functioning as offered suggestions, e.g. "吧" (*Ba*). Instance (i) is the Chinese equivalence of "who" in English; Instances (j) and (m) are equivalences of "how many"; Instance (k) is an equivalence of "what"; Instance (l) is

an equivalence of "or"; and the rest are equivalences of "whether or not". These words are given heavier weight to prime the system of *WeChat Translate* to generate certain specific types of questions in English. This may explain why there are salient overuses of certain grammatical patterns in the analysis corpus. In other words, the overuse of such patterns is related to the translation of certain question forms without ending punctuation. For example, the word "吧" (*Ba*) is an indicator of Chinese question forms functioning as offered suggestions. In the self-built parallel corpus, we have found that 12 such instances (without ending punctuation) are translated into the "Why don't…" pattern (as mentioned in Table 6 and Example (4)). The normalized frequency of the "Why don't…" pattern is much greater in the analysis corpus (570.06) than in the reference corpus (43.60), which suggests an exaggerated use of this pattern in WeChat's MT outputs. The translation of these 12 instances into the same English pattern contributes to the overuse phenomena discussed above. Since certain words may be given heavier weight toward the output for certain types of grammatical patterns by the MT mechanism, it is highly probable that normalization will exist and significantly present itself in MT outputs in the form of overuses of lexical chunks, some of which are related to syntactic patterns.

5. Conclusion

Our research shows that, whereas the simplification hypothesis cannot be confirmed in the analysis corpus, the tendency of normalization exists. It was argued that such findings result from the working mechanism of MT. On the one hand, since MT operates within seconds, there are no such influences as human cognitive factors, and there is no need to spend less effort as human translators, which are the most probable causes for the existence of simplification in translation. As a result of the research, we hypothesize that simplification cannot be confirmed in MT outputs and hope that more studies will be conducted in the future to confirm or refute this hypothesis. On the other

hand, the prioritized acceptability of translation acts as a gravitational pull toward conformation to the norms of target languages (Toury, 2012: 203–204). The mechanism of MT also follows this trend. To produce acceptable translations, the reference parallel corpora for the current MT systems usually include examples of standard language models and translation models, which offer standard formats in terms of diction, punctuation, and syntactic and textual patterns with which the most optimal option is produced. As a result, the features of the target language will be simulated. For example, in our study, the texts translated by *WeChat Translate* conform to the norms of standard English, as shown by the proportions of nominal phrases to verbal phrases, and syntactic patterns, including punctuation. In the process of translation, if certain salient words in the original texts prime MT systems to repetitively choose typical grammatical patterns in the target language, then normalization occurs more obviously in forms of the overuse of certain lexical chunks, as with the case of the overuse of question patterns in the present study. In this sense, the tendency of normalization to overuse lexical chunks as representatives can be hypothesized as a more prominently probabilistic feature of MT outputs, which is worthy of further exploration.

According to Holmes (1988), 'process oriented research' and 'product oriented research' are two sub-branches in descriptive translation studies: whereas the former attempts to uncover what is happening in the translation process, the latter tries to find out the patterns of existing translations. Corpora have been used in the process oriented research on MT (on the analysis and generation steps) for a long time (Feng, 2021). By investigating the hypothesized Translation Universals in MT, the present study showcases how corpora can also be applied in the product research of MT (of features of MT outputs). Corpus-based product research regarding MT can provide quantitative metrics, which might include the patterns found in corpus data and the statistics of language features, which may in turn be useful for developing MT systems in the future.

This study has investigated the linguistic features of MT outputs through the lens of two translation universals including normalization and simplification. It is argued that the research paradigm for corpus-based studies on translation universals can

be further extended by exploring the linguistic patterns not only of human renditions, but also of MT outputs. The present study is also expected to shed new light on the development of MT systems and encourage more corpus-based product-oriented research on MT outputs. Since the corpus used in this study is register-specific and *WeChat Translate* is a new MT system, further studies on the translations produced by other MT systems for different language pairs are needed to support or refute the hypothesized Translation Universals in MT.

Acknowledgements

We are grateful to the two anonymous reviewers for their valuable comments and suggestions which helped to shape this paper. The work described in this paper was partially supported by a grant from the Research Grants Council of the Hong Kong Special Administrative Region, China (PolyU/RGC 15602621).

References

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

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *In Terminology, LSP and Translation: Studies in Language Engineering, in Honour of Juan C. Sager* (pp. 175–186). John Benjamins.

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*, *28*(1), 61–86.

Blum-Kulka, S., & Levenston, E. (1983). Universals of lexical simplification. In C. Faerch & C. Gabriele (Eds.), *Strategies in Inter-language Communication* (pp. 119–139). Longman.

Bo, B. (2002). *English Grammar.* Kai Ming Press.

Cappelle, B., & Loock, R. (2017). Typological differences shining through: The case of phrasal verbs in translated English. In G. De Sutter, M. A. Lefer, & I. Delaere (Eds.), *Empirical Translation Studies: New Theoretical and Methodological Traditions* (pp. 235–264). Mouton de Gruyter.

Chesterman, A. (2004). Beyond the particular. In A. Mauranen & P. Kujamäki (Eds), *Translation Universals: Do They Exist?* (pp. 33–50). John Benjamins.

Cronin, M. (2013). *Translation in the Digital Age*. Routledge.

Feng, Z. (2018). Parallel development of machine translation and artificial intelligence. *Journal of Foreign Language*, *41*(6), 35–48.

Feng, Z. (2020). Rosetta Stone and machine translation. *Foreign Language Research*, *1,* 1–17.

Feng, Z. (2021). Two-wheel driven natural language understanding. *Frontiers in Corpus Studies*, *1,* 148–175.

Grabowski, L. (2013). Interfacing corpus linguistics and computational stylistics: Translation universals in translational literary Polish. *International Journal of Corpus Linguistics*, *18*(2), 254–280.

Gupta, A. (2021). User-controlled content translation in social media. In *26th International Conference on Intelligent User Interfaces—Companion* (pp. 96–98). Association for Computing Machinery. https://doi.org/10.1145/3397482.3450714

Halverson, S. (2003). The cognitive basis of Translation Universals. *Target*, *15*(2), 197–241.

Hoey, M. (2005). *Lexical Priming: A New Theory of Words and Language*. Routledge.

Holmes, J. (1988). *Translated!: Papers on Literary Translation and Translation Studies*. Rodopi.

Hu, K. (2016). *Introducing Corpus-Based Translation Studies*. Springer; Shanghai Jiao Tong University Press.

Hu, K., & Li, Y. (2016). Features of machine translation and its relations with human translation.

Chinese Translators Journal, *5,* 6–14.

Hu, X., & Zeng, J. (2011). 从“把”字句看翻译汉语的杂合特征 [Hybridization of translated Chinese as observed in the use of “ba” constructions]. *Foreign Language Research*, *130,* 69–75.

Joos, M. (1967). *The Five Clocks*. Harcourt, Brace & World.

Kajzer-Wietrzny, M. (2015). Simplification in interpreting and translation. *Across Languages and Cultures*, *16*(2), 233–255.

Kenny, D. (2001). *Lexis And Creativity In Translation: A Corpus-Based Study*. St. Jerome.

Kruger, H., & Rooy, B. (2012). Register and the features of translated language. *Across Languages and Cultures*, *13*(1), 33–65.

Lapshinova-Koltunski, E. (2015). Variation in translation: Evidence from corpora. In C. Fantinuoli & F. Zanetti (Eds.), *New Directions in Corpus-Based Translation Studies* (pp. 93–114). Language Science Press.

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.

Laviosa, S. (2006). Data-driven learning for translating anglicisms in business communication. *IEEE Transactions on Professional Communication*, 49(3), 267-274.

Laviosa, S. (2010). Corpus-based translation studies 15 years on: Theory, findings, applications. *Synaps*, *24*(2010), 3–12.

Laviosa, S. (2011). Corpus-based translation studies: Where does it come from? Where is it going? In A. Kruger, K. Wallmach, & J. Munday (Eds.), *Corpus-based Translation Studies: Research and Applications* (pp. 13–32). Continuum.

Lian, S. (2010). *英汉对比研究* [Contrastive Studies of English and Chinese]. Higher Education Press.

Liu, F. (2019, Oct. 24). *微信牵手网易有道* [WeChat and Youdao in Cooperation]. ZOL Soft. http://soft.zol.com.cn/489/4891071.html.

Mao, L. (2014). 微信与微语言生活[WeChat and WeChat language]. *Social Science Front*, *12*, 136–141.

Page, R. (2012). *Stories and Social Media: Identities and Interaction*. Routledge.

Rayson, P. (2008). From key words to key semantic domains. *International Journal of Corpus Linguistics*, *13*(4), 519–549.

Sánchez-Moya, A., & Cruz-Moya, O. (2015). "Hey there! I am using WhatsApp": A preliminary study of recurrent discursive realisations in a corpus of WhatsApp statuses. *Procedia: Social and Behavioral Sciences*, *212*(C), 52–60.

Schwartz, L. (2018). The history and promise of machine translation. In I. Lacruz & R. Jääskeläinen (Eds.), *Innovation and Expansion in Translation Process Research* (pp. 168–198). John Benjamins.

Scott, M. (2012). *WordSmith Tools* (Version 6.0) [Computer software]. Lexical Analysis Software. https://lexically.net/wordsmith/downloads/.

Stubbs, M. (1996). *Text and Corpus Analysis: Computer-assisted Studies of Language and Culture.* Blackwell.

Szymor, N. (2018). Translation: Universals or cognition? *Target: International Journal of Translation Studies*, *30*(1), 53–86.

Tagliamonte, S., & Denis, D. (2008). Linguistic ruin? LOL! Instant messaging and teen language. *American Speech*, *83*(1), 3–34.

Teich, E. (2003). *Cross-Linguistic Variation in System and Text, A Methodology for the Investigation of Translations and Comparable Texts*. Mouton de Gruyter.

Tirkkonen-Condit, S. (2002). Translationese: A myth or an empirical fact? A study into the linguistic identifiability of translated language. *Target*, *14*(2), 207–220.

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

Toury, G. (2012). *Descriptive Translation Studies and Beyond* (2nd ed.). John Benjamins.

Vanderauwera, R. (1985). *Dutch Novels Translated into English*. Rodopi.

Van Oost, A., Willems, A., & De Sutter, G. (2016). Asymmetric syntactic patterns in German-Dutch translation: A corpus-based study of the interaction between normalisation and shining through. *International Journal of Translation*, *10*, 1–18.

Vanmassenhove, E., Shterionov, D., & Way, A. (2019). Lost in Translation: Loss and Decay of Linguistic Richness in Machine Translation. In M. Forcada, A. Way, B. Haddow, & Sennrich (Eds.), *Proceedings of Machine Translation Summit XVII: Research Track* (pp. 222–232). European Association for Machine Translation. https://aclanthology.org/W19-6622

Wang, X., & Li, X. (2016). A corpus-based study of normalization in Chinese translations of Shakespeare's plays. *Journal of Foreign Languages*, *39*(3), 106–112.

Wang, B., Shan, D., Fan, A., Liu, L., & Guo, J. (2022). A sentiment classification method of web social media based on multidimensional and multilevel modelling. *IEEE Transactions on Industrial Informatics*, *18*(2), 1240–1249.

Williams, D. (2005). *Recurrent Features of Translation in Canada: A Corpus-based Study*. University of Ottawa.

Zhang, M., & Toral, A. (2019). The effect of Translationese in machine translation test sets. In O. Bojar, R. Chatterjee, C. Federmann, M. Fishel, Y. Graham, B. Haddow, M. Huck, A. J. Yepes, P. Koehn, A. Martins, C. Monz, M. Negri, A. Névéol, M. Neves, M. Post, M. Turchi, & K. Verspoor (Eds.), *Proceedings of the Fourth Conference on Machine Translation (Volume 1: Research Papers)* (pp. 73–81). Association for Computational Linguistics. https://aclanthology.org/W19-5208/

Appendices

Appendix A. Stubbs' (1996: 36-37) List of the 200 Most Frequent Words in English

a, about, after, again, against, all, also, always, an, and, another, any, are, around, as, at, away, back, be, because, been, before, being, between, both, but, by, came, can, children, come, could, course, day, did, didn't, do, does, don't, down, each, end, er, even, every, fact, far, few, find, first, for, from, get, go, going, good, got, great, had, has, have, he, her, here, him, his, home, house, how, I, I'm, if, in, into, is, it, its, it's, just, kind, know, last, left, life, like, little, long, look, looked, made, make, man, many, may, me, mean, men, might, more, most, Mr, much, must, my, never, new, no, not, nothing, now, of, off, oh, old, on, once, one, only, or, other, our, out, over, own, part, people, perhaps, place, put, quite, rather, really, right, said, same, say, says, see, she, should, so, some, something, sort, still, such, take, than, that, that's, the, their, them, then, there, these, they, thing, things, think, this, those, though, thought, three, through, time, to, too, two,

under, up, us, used, very, want, was, way, we, well, went, were, what, when, where, which, while, who, why, will, with, without, work, world, would, year, years, yes, you, your.

Appendix B. Lemmas in List Heads of WeChat Translation Corpus

1. the 2. to 3. you / your 4. I / me / my 5. a / an 6. and
7. of 8. is / be / were / are / am / was / been 9. in 10. it / its
11. not 12. for 13. have / has / had 14. can / can't 15. on 16. that
17. at 18. do / don't / did 19. so 20. time 21. this 22. will
23. go / going / gonna / went 24. if 25. we / our / us 26. with
27. all 28. what 29. it's 30. good / better 31. there 32. no
33. but 34. I'm 35. want 36. class 37. just 38. one
39. up 40. get / got 41. ha 42. know 43. like 44. about
45. please 46. then 47. when 48. or 49. take 50. out
51. more / much 52. as 53. school 54. from 55. back
56. come 57. first 58. tomorrow 59. how 60. right
61. students 62. who 63. after 64. now 65. think 66. well
67. people 68. see 69. by 70. two 71. oh 72. work
73. too 74. day / days 75. very 76. need 77. little 78. also
79. send 80. next 81. why 82. group 83. night 84. some
85. teacher ← (top 100)
86. ask 87. because 88. other 89. he 90. say / said
91. make 92. only 93. really 94. you're 95. thank 96. today
97. should 98. I'll 99. English 100. they / their / them 101. okay
102. eat 103. give 104. everyone 105. still 106. look
107. department 108. any 109. before 110. she / her
111. remember 112. yes 113. long 114. card 115. let
116. tonight 117. ill 118. three 119. number 120. over
121. examination 122. many 123. talk 124. where 125. way
126. than 127. tell 128. hard 129. information 130. own
131. morning 132. lot 133. every 134. pay 135. last
136. hope 137. again 138. here 139. must 140. each
141. sign 142. buy 143. meeting 144. even 145. college
146. that's 147. same 148. sister 149. wechat 150. home
151. money 152. early 153. find 154. there's 155. something
156. afternoon 157. new 158. training 159. feel 160. write
161. test 162. week 163. phone 164. life 165. friends
166. help 167. sure → (top 200)

Appendix C Lemmas in List Heads of BNC Baby (Part)

1. the 2. of 3. and 4. to 5. a / an 6. in 7. I / me / my
8. you / your 9. it / its 10. that 11. is / be / was / are / been / were / being
12. for 13. yeah / yes 14. on 15. have / had / has / haven't
16. as 17. this / these / those 18. no / not 19. with 20. oh / ooh
21. one 22. but 23. we / us / our 24. what 25. they / them / their
26. well 27. there 28. he 29. or 30. It's 31. by 32. so

33. do / don't / done / doing / did / didn't / does 34. know 35. at
36. can / could / can't 37. which 38. if 39. go / went / going / gonna
40. all 41. from 42. that's 43. then 44. mm 45. like
46. she / her 47. just 48. two 49. up 50. get / got 51. when
52. right 53. about 54. er / erm 55. think / thought 56. out
57. some 58. more 59. now 60. said / say 61. see 62. will / would 63. other 64. only 65. his / him 66. I'm 67. time
68. where 69. mean 70. may / might 71. now 72. than
73. very 74. here 75. want 76. cos / because 77. any
78. there → (top 100)
79. who 80. such 81. put 82. good 83. really 84. come
85. down 86. I've 87. look 88. he's 89. I'll 90. should
91. four 92. there's 93. five 94. much 95. way 96. you're
97. into 98. also 99. first 100. over 101. between 102. work
103. back 104. people 105. why 106. something 107. many
108. bit 109. take 110. off 111. alright 112. use / used
113. they're 114. most 115. need 116. even 117. you've
118. home 119. little 120. must 121. make / made 122. different
123. she's 124. same 125. before 126. new 127. quite
128. ah 129. both 130. long 131. number 132. give
133. nice 134. six 135. though 136. sort 137. after
138. another 139. what's 140. through 141. hundred 142. thing
143. example 144. too 145. however 146. again 147. point
148. still 149. case 150. each 151. social 152. last
153. never 154. rather → (top 200)

Address for correspondence

Jinru Luo
AG518 Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Hung Hom
Hong Kong
jin.ru.luo@connect.polyu.hk

Co-author information

Dechao Li
AG 518b Department of Chinese and Bilingual Studies
The Hong Kong Polytechnic University
Dechao.li@polyu.edu.hk