12 Investigating lexical simplification

A corpus-based comparative analysis of interpreted, L2, and native speech

Han Xu and Kanglong Liu

12.1 Introduction

The tendency for translated texts to exhibit less complex language use compared to texts produced in the original language has garnered significant scholarly attention in recent decades. This phenomenon, known as simplification, has been extensively studied in corpus-based translation research since its emergence in the early 1990s (Baker 1996; Kruger and Van Rooy 2016; Laviosa 1998; Liu and Afzaal 2021; Liu, Cheung, and Liu 2023). The observed pattern of linguistic simplification in translation serves as the foundation for arguments supporting the potential existence of translation universals that distinguish translational language from other language varieties (Baker 1996).

More recently, researchers have expanded the focus of investigation from translation to interpreting in order to explore the presence and extent of simplification as a potential universal property of interpreting. Many studies in this line of research have employed lexical complexity indicators to operationalize and measure simplification (Bernardini, Ferraresi, and Miličević 2016; Dayter 2018; Ferraresi et al. 2018; Sandrelli and Bendazzoli 2005). This lexical approach was initially introduced by Laviosa (1998) to identify core lexical use pattern in translation by comparing lexical density, list head, and the proportion of high-frequency words in translated and non-translated English texts. Laviosa (1998) discovered that, in contrast to non-translated texts in the original language, translated texts tend to be less informative, more repetitive, and less lexically sophisticated. However, in the limited body of research on interpreting, findings have been inconsistent and even contradictory (Dayter 2018; Kajzer-Wietrzny 2015; Russo, Bendazzoli, and Sandrelli 2006; Xu and Li 2022). Empirical evidence further indicates that the manifestation of simplification in interpreting may be influenced by genre (Hu, Xiao, and Hardie 2019; Xu and Li 2022), mode (Kajzer-Wietrzny 2015; Lv and Liang 2019), the influence of orality (Shlesinger and Ordan 2012), and the specific measures used. Notably, most existing studies on simplification have employed the umbrella term "interpreting" to encompass various types of interpreting without making clear distinctions regarding their specific characteristics. This poses

DOI: 10.4324/9781003377931-13

a methodological challenge for researchers aiming to compare the results of different studies. Furthermore, even within the same type of interpreting, the final product may be influenced by factors such as the qualifications of the interpreter (Xu, Hale, and Stern 2020; Xu 2021), interpreting directionality (Dayter 2018; Sandrelli and Bendazzoli 2005), and language pair (Bernardini, Ferraresi, and Miličević 2016; Russo, Bendazzoli, and Sandrelli et al. 2006). Therefore, investigating simplification as a narrowly defined linguistic universal may not fully capture its dynamic nature. Instead of solely focusing on substantiating simplification as a hypothesis, it is more meaningful to profile its diverse manifestations using valid measures and to extrapolate the reasons behind observed variations or commonalities in order to gain a deeper understanding of the interpreting process. Similarly, Chesterman (2010, 6), who focused on translation universals, argued that "perhaps it would be more fruitful to search for less-than universal patterns in translation profiles, under different sets of conditions, and thus make more modest claims."

On a different note, most existing studies on simplification in translation and interpreting have primarily employed a monolingual comparative approach, drawing comparisons between translational language and non-translated original language following the practice of Laviosa (1998). However, some researchers argue that the tendency to simplify may not be exclusive to translational language and could also be a characteristic of other language varieties influenced by bilingualism, such as L2 speech or writing (Chesterman 2010; Kotze 2020; Lanstyák and Heltai 2012). These bilingualism-influenced language varieties have been extensively discussed under the notion of language contact, which refers to a situation where language production is constantly shaped by cross-linguistic influence due to the activation of two language systems by the producer (Thomason and Kaufman 1992). Viewed from this perspective, translation and interpreting can be considered as language-contact events (Kranich 2014). Kotze (2020, 121) posited that "in contexts of contact-influenced communication, cooperation and normativity may combine in unique ways to impose certain socio-cognitive constraints on language production, leading to the avoidance of complexity, increased explicitness and hyperstandardization or conservatism, in order to ensure effective communication." However, studies on simplification in translational language have rarely considered the inclusion of contact-induced language varieties as reference texts for comparison. Little is known about whether the observed simplification in translational language is similar to or different from its potential existence in other contact-induced language varieties.

With the aim of narrowing this research gap, this study sets out to explore simplification in interpreting by examining the lexical variation patterns among three language varieties: speech interpreted from Chinese into English (interpreted speech), original English speech produced by native speakers (L1 speech), and original English speech produced by advanced non-native speakers (L2 speech). By measuring the extent to which simplification manifests itself across these language varieties, the study aims to explore

potential similarities in the lexical use patterns between interpreted speech and L2 speech, considering their shared characteristics as bilingualism-influenced language activities.

12.2 Lexical simplification in interpreting

Existing studies on lexical simplification in interpreting reveal that its operationalization features great complexity and is subject to the influence of a wide range of intra- and extra-textual factors (Bernardini, Ferraresi, and Miličević 2016; Dayter 2018; Ferraresi et al. 2018; Lv and Liang 2019; Russo, Bendazzoli, and Sandrelli 2006; Sandrelli and Bendazzoli 2005; Xu and Li 2022). These factors are intertwined, giving rise to various lexical use patterns in interpreted speech. Sandrelli and Bendazzoli (2005) were among the first to adopt Laviosa's approach in examining lexical simplification in interpreting. They collected data from an open trilingual corpus (Italian, English, and Spanish) of European Parliament speeches (EPIC) and their simultaneous interpreting. Sandrelli and Bendazzoli found that there is little variation in lexical density when comparing interpreted speech to original English speech. Moreover, they observed a higher proportion of high-frequency words in the list head of interpreted English, indicating a more frequent use of high-frequency words and less lexical variation. However, this pattern did not apply to interpreted Italian speech, suggesting that the level of repetitiveness may be influenced by the specific language combination. They also discovered a higher percentage of lexical words in interpreted English, which may be attributed to interpreters adopting different coordination strategies, such as adding explanations, reformulations, and paraphrases. On the whole, Sandrelli and Bendazzoli's study reveals that the lexical simplification pattern does not consistently apply to simultaneously interpreted speech. A similar finding was reported by Kajzer-Wietrzny (2015) in another study. Comparing the lexical use of English speech interpreted from German, French, Dutch, and Spanish to that of the original English speech, Kajzer-Wietrzny found that regardless of the language combination, interpreted speech is not more simplified than original English. Instead, interpreted speech exhibited a higher level of lexical density and less use of high-frequency words. While English speech interpreted from Spanish showed a higher level of repetition than the original, supporting the simplification hypothesis, the opposite was observed for speech interpreted from French. This finding suggests that the tendency for repetition may depend on the specific language combination. In a similar vein, Dayter (2018), based on a parallel bidirectional corpus of original and simultaneously interpreted speech in Russian and English, demonstrated that interpreting directionality may also influence the lexical variation pattern in interpreting. Dayter's study revealed that compared to the original Russian speech, interpreted Russian had lower lexical density and higher list head coverage, confirming a tendency for simplification. However, the speech interpreted from Russian into English exhibited the opposite trend.

Shlesinger and Ordan (2012, 43) once posited that "modality may exert a stronger effect than ontology—i.e. that being oral (vs written) is a more powerful influence than being translated (vs original)." In light of the emergence of varying counterevidence, one may wonder whether the lexical approach, which was originally used to study translation, is suitable for interpreting. With this research agenda in mind, some researchers have become interested in exploring how lexical use may present different patterns between interpreting and translation or across interpreting types (Bernardini, Ferraresi, and Miličević 2016; Ferraresi et al. 2018; Lv and Liang 2019; Shlesinger and Ordan 2012; Xu and Li 2022). One major study in this group is from Bernardini, Ferraresi and Miličević (2016), who compiled an intermodal bidirectional corpus of interpreted and translated EU Parliament speeches (EPTIC). In contrast to the findings of previous research, they found that interpreted speech demonstrates a more simplified lexical use pattern than translated text and original speech, showing a strong mediation effect on the reduction of lexical complexity. Based on this result, Bernardini, Ferraresi and Miličević (2016, 81) argued that "simplification thus appears to be both of feature of orality and a feature of mediation, such that interpreted texts, being both spoken and mediated, occupy one extreme of the simplicity cline." This finding largely supports Laviosa's simplification hypothesis.

In addition, Bernardini and colleagues found that interpreted Italian speeches have lower lexical density and mean sentence length than their translated versions, while interpreted English texts make greater use of high-frequency words. This suggests that the lexical complexity measures may apply differently to different languages. In a more recent study that adopts an intermodal comparative approach, Xu and Li (2022), based on a corpus of interpreted and translated English speeches from the Hong Kong Legislative Council and original English speeches from the UK parliament, reported similar findings. According to Xu and Li (2022), except for the measure of lexical density, interpreted speech is consistently more simplified than translated original speech in terms of the standard type-token ratio, list head coverage, and average sentence length, indicating that interpreted speech is more repetitive and less diverse in lexical use. From a different perspective, Lv and Liang (2019) investigated the lexical use patterns across different interpreting modes. Their study examined a set of lexical complexity measures in consecutively interpreted English speech from Chinese, simultaneously interpreted English speech from Chinese, and original English speech. The results indicated that consecutively interpreted speech exhibited greater simplification than the original speech in all aspects, while simultaneously interpreted speech demonstrated an opposite trend that is more informative and lexically sophisticated than the original. These findings suggest that the lexical use pattern in interpreting may also be influenced by the interpreting mode.

Based on the available studies, it appears that the lexical use pattern in interpreting is highly complex, and its operationalization is heavily influenced by various factors inherent to the interpreting activity under investigation,

as well as the nature of the reference text used for comparison. Due to the challenges in accessing data for corpus-based interpreting studies, unraveling the intricate interplay of the various factors and testing their individual or collective influence on simplification may not always be feasible. Despite the diverse findings, most researchers agree that the mediation process, whether through interpreting or translation, tends to reduce lexical complexity, with simplification manifesting in different linguistic ways (Bernardini, Ferraresi, and Miličević et al. 2016; Ferraresi et al. 2018; Shlesinger and Ordan 2012). Furthermore, the majority of existing studies on simplification in interpreting focus solely on comparing interpreted texts to two speech varieties: original speech by native speakers and translated texts. Very few studies utilize other language varieties as reference texts. In light of this, the present study introduces advanced bilingual language use, specifically L2 speech, as a new reference text for comparison. Since L2 speech is produced under significant cross-linguistic influence, it shares similarities with interpreted speech, as both can be categorized as language-contact events (Kotze 2020; Kranich 2014). Specifically, this study aims to investigate the variation in lexical use patterns across interpreted speech, L1 speech, and L2 speech by addressing the following research questions:

RQ 1: How does the lexical use pattern vary across interpreted speech, L1 speech, and L2 speech?
RQ 2: Are there any similarities or differences between interpreted speech and L2 speech regarding the variation in lexical use patterns?

12.3 Methods and procedures

12.3.1 Corpus compilation

The data used in this study is sourced from the Political Debate English Comparable Corpus (PEDCC). PEDCC consists of three subcorpora representing different language varieties: Simultaneously Interpreted English Speech (L2I), Native English Speech (L1O), and L2 English Speech (L2O).[1] L2I comprises debates from the Legislative Council of Hong Kong, where professional interpreters interpret impromptu speeches from Cantonese (Chinese) into English. These interpreters are native Cantonese speakers who work into their second language during interpretation. Given that L2I was compiled from interpreted debates, the interpreting activity can be classified as "free simultaneous interpreting" (Dayter 2018). In this type of interpreting, the interpreters are required to interpret impromptu speeches delivered by the speakers. It is important to note that this differs from situations where source speech speakers read from a prepared script (Kajzer-Wietrzny 2015). The impromptu nature of the interpreted speeches adds an additional layer of complexity and spontaneity to the interpreting process. L1O consists of debates from the British Parliament's House of Commons, featuring native English speakers.

Table 12.1 Profile of the three subcorpora in PEDCC

Subcorpus	*Texts count*	*Token no.*	*Source*	*Format*	*Producer*	*Language*	*Availability*
L2I	50	103,097	HK LegCo debates	Q&A	L2 speakers	English	Publicly available
L1O	50	104,077	UK Parliament debates	Q&A	L1 speakers	English	Publicly available
L2O	50	105,186	HK TV interviews	Q&A	L2 speakers	English	Publicly available

L2O is derived from interviews conducted in English by native Cantonese speakers in two popular Hong Kong TV programs: "All About Money and Talk and Walk."

A total of 50 speeches, covering the period between 2016 and 2020, were collected from these sources. All speeches share a question-and-answer format and revolve around political, financial, and social topics. The three subcorpora (L2I, L1O, and L2O) are comparable and homogeneous in terms of size, genre, format, time span, and language type, thereby minimizing the potential influence of confounding factors. The detailed information of the three sub-corpora is presented in Table 12.1 below. To ensure accuracy, the collected speeches were automatically transcribed using iFLYTEK, an automatic transcription software. Manual verification was conducted to ensure transcription accuracy.

12.3.2 Data analysis

Following Laviosa's approach, the present study will adopt four mainstream lexical complexity indicators: type-token ratio, lexical density, list head, and core vocabulary coverage. The aim is to maximize comparability of results with previous research (Dayter 2018). The type-token ratio describes the proportion of word types to the total number of running words. It is frequently used in interpreting studies to indicate the level of lexical variety (Shlesinger and Ordan 2012). The type-token ratio can be automatically obtained using the Lexical Complexity Analyzer (Ai and Lu 2010; Lu 2012). Lexical density measures the informativeness of a text and is calculated as the proportion of content words to functional words (Stubbs 1986). Content words, also known as lexical words, include nouns, verbs, adjectives, and adverbs (Sandrelli and Bendazzoli 2005). The number of functional words can be obtained by subtracting all the content words from the total running words. Similar to the type-token ratio, lexical density can also be automatically calculated using the Lexical Complexity Analyzer. For the list head, it serves as an indicator of the repetitiveness level within a text. As a corpus-internal measure, it represents the percentage of the corpus that is accounted for by high-frequency

words. The list of high-frequency words for each of the three subcorpora is provided in Appendix 1. High-frequency words are typically defined as the 100 most frequently used words (Ferraresi et al. 2018). Core vocabulary coverage measures the proportion of high-frequency words to low-frequency words. The high-frequency words used in this study are the top 200 most frequently used words, obtained from an external reference corpus (Bernardini, Ferraresi and Miličević 2016; Ferraresi et al. 2018). To create this list, the spoken component of the BNC (Leech, Rayson, and Andrew Wilson 2001) was adopted as the reference corpus. The list of the most frequently used words from the BNC is provided in Appendix 2.

According to the definitions of the four lexical indicators, a higher value for type-token ratio and lexical density represents greater lexical diversity, while a higher value for list head and core vocabulary suggests a larger portion of the text is occupied by high-frequency words, indicating increased repetitiveness. Consistent with previous research conventions (Ferraresi et al. 2018; Lv and Liang 2019), the type-token ratio, lexical density, and core vocabulary were calculated on a textual basis to indicate lexical use patterns, while list head was calculated at the corpus level. However, the calculation approach for list head may not fully capture within-corpus individual textual variation (Biber and Jones 2009) or how it relates to the exhibition of repetitiveness at the corpus level. Therefore, in addition to calculating list head at the corpus level, this study also performed calculations on a textual basis. The results from the two calculation methods are compared to assess their compatibility.

12.4 Results

The calculation results of the four lexical indicators on a textual basis, including their mean values across the 50 texts in each subcorpus and standard deviation, are presented in Table 12.2. The calculation of list head at the subcorpus level is presented in Table 12.3. Both calculation methods yield the same results, except for a marginal distinction between the values in L2I. This confirms the validity of both calculation methods. The lexical variation pattern in the three subcorpora is depicted in Figure 12.1.

On the whole, the simplification phenomenon appears to be most pronounced in L2 speech, which exhibits the lowest type-token ratio and the highest list

Table 12.2 Lexical variation patterns across the three subcorpora (by-text calculation)

	M (SD) in L2I	*M (SD) in L1O*	*M (SD) in L2O*
Type-token ratio	0.257 (0.022)	0.285 (0.019)	0.238 (0.020)
Lexical density	0.499 (0.024)	0.490 (0.017)	0.502 (0.021)
List head coverage	0.560 (0.037)	0.562 (0.022)	0.605 (0.027)
Core vocabulary coverage	0.582 (0.049)	0.578 (0.023)	0.652 (0.034)

Table 12.3 List head coverage in the three subcorpora (by-corpus calculation)

Subcorpus	*Token no.*	*List head (freq.)*	*List head ratio*
L2I	102,312	57,331	0.5603
L1O	103,845	62,858	0.6053
L2O	101,907	57,312	0.5623

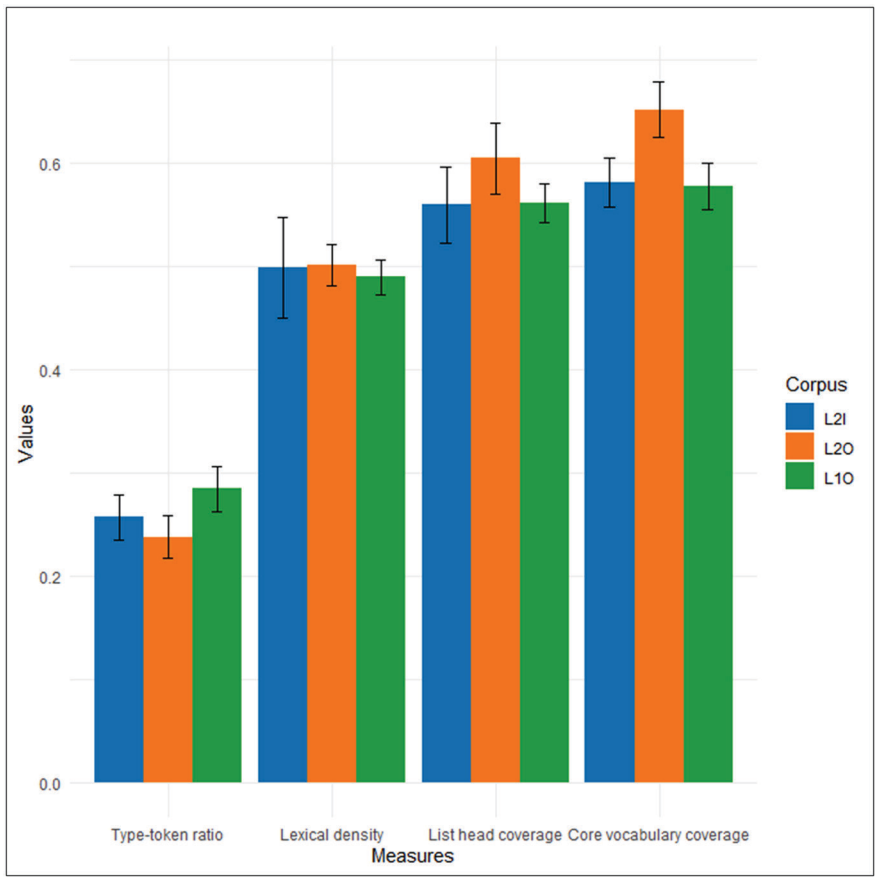

Figure 12.1 Comparison of lexical measures across interpreted, L2 and L1 speech

head and core vocabulary coverage. This suggests that L2 speech has the least lexical variety and is the most repetitive among the three language varieties, indicating a strong tendency for speakers to simplify their lexical use when speaking in their second language. In contrast, L1 speech is characterized by the highest type-token ratio and the lowest core vocabulary coverage, indicating

that native speakers have more linguistic resources available and are more capable of producing lexically sophisticated speech. Regarding interpreted speech, a tendency for interpreters to simplify their lexical use can be observed when compared to non-interpreted native speech. Interpreted speech shows a lower type-token ratio and a higher core vocabulary coverage. The portion covered by list head is almost the same between L2I and L1O. This lexical use pattern, as reflected by the type-token ratio and core vocabulary coverage, confirms the presence of a simplification phenomenon in interpreting. However, it is intriguing to note that the extent to which speakers simplify their speech is greater in L2 speech than in interpreted speech. Furthermore, while the simplification phenomenon can be largely confirmed through the type-token ratio and core vocabulary coverage, the lexical variation patterns revealed by lexical density show a completely opposite trend. Lexical density, which measures information load and lexical variety, is highest in L2 speech, followed by interpreted speech and L1 speech. This result suggests that L2 speech is the most informative among the three varieties in terms of lexical density.

In order to determine whether the mean values of the four lexical complexity measures are statistically different, a one-way ANOVA test was conducted for each lexical measure across the three subcorpora. The results of the ANOVA tests, presented in Table 12.4, indicate statistical significance ($p<0.01$) for all four lexical measures, confirming the distinctions in lexical use among the three language varieties. However, the ANOVA test alone does not provide information on the statistical significance of pairwise differences. Therefore, a series of Tukey post hoc tests were performed. The results, as shown in Table 12.5, reveal that all the differences between L1 and L2 speeches are statistically significant, indicating that these two types of language varieties are inherently different in terms of their lexical use. In the case of interpreted speech, not all pairwise comparisons with L1 and L2 speeches exhibit statistical significance. For instance, although the proportion of core vocabulary coverage in interpreted speech is higher than that in L1 speech, the difference between the two is not statistically significant. This suggests that interpreted speech occupies an intermediate position in terms of lexical complexity among the three language varieties, with certain aspects of its lexical use resembling both L1 and L2 speech. To enhance the comprehension of the ANOVA test results,

Table 12.4 ANOVA tests results of the lexical measures

Lexical measures	*F value*	*P value*
Type-token ratio	67.85	<0.01*
Lexical density	4.14	<0.01*
List head	37.52	<0.01*
Core vocabulary coverage	64.22	<0.01*

Note: * indicates a statistically significant difference in the measure.

Table 12.5 Tukey post-hoc tests results

Pair-wide comparison		*Difference*	*P value*
Type-token ratio	L2I-L1O	-0.027	<0.01*
	L2O-L1O	-0.047	<0.01*
	L2O-L2I	-0.019	<0.01*
Lexical density	L2I-L1O	0.008	0.12
	L2O-L1O	0.012	0.02*
	L2O-L2I	0.003	0.70
List head	L2I-L1O	-0.002	0.9
	L2O-L1O	0.043	<0.01*
	L2O-L2I	0.045	<0.01*
Core vocabulary coverage	L2I-L1O	0.005	0.79
	L2O-L1O	0.075	<0.01*
	L2O-L2I	0.070	<0.01*

Note: * indicates a statistically significant difference in the measure.

the means of the three language varieties and their respective distributions are visually depicted in four boxplots, as presented in Figure 12.2.

12.5 Discussion

By analyzing data from a corpus containing English speech simultaneously interpreted from Chinese, English speech produced by native speakers, and English speech produced by advanced non-native speakers, this study aims to investigate the phenomenon of simplification in interpreting. Consistent with previous research findings (Bernardini, Ferraresi and Miličević 2016; Dayter 2018; Kajzer-Wietrzny 2015; Russo, Bendazzoli, and Sandrelli 2006), the results of this study demonstrate that not all parameters uniformly indicate the presence of simplification in interpreting, thus highlighting the complexity associated with its manifestation in this context. When compared to the lexical use observed in non-interpreted L1 speech, interpreted speech exhibits lower lexical variety and increased repetitiveness, as indicated by a higher type-token ratio and lower core vocabulary coverage. These findings provide supporting evidence for previous research that confirms the tendency toward simplification in interpreting (Bernardini, Ferraresi and Miličević 2016; Lv and Liang 2019; Xu and Li 2022).

However, it is worth noting that the lexical density observed in interpreted speech exceeds that found in non-interpreted original speech, suggesting that interpreters employ content words to a greater extent than native speakers. Although this finding may initially seem counterintuitive, similar results have been reported in several previous studies (Dayter 2018; Kajzer-Wietrzny 2015; Russo, Bendazzoli, and Sandrelli 2006; Xu and Li 2022). In light of this counterevidence, researchers have proposed various explanations to

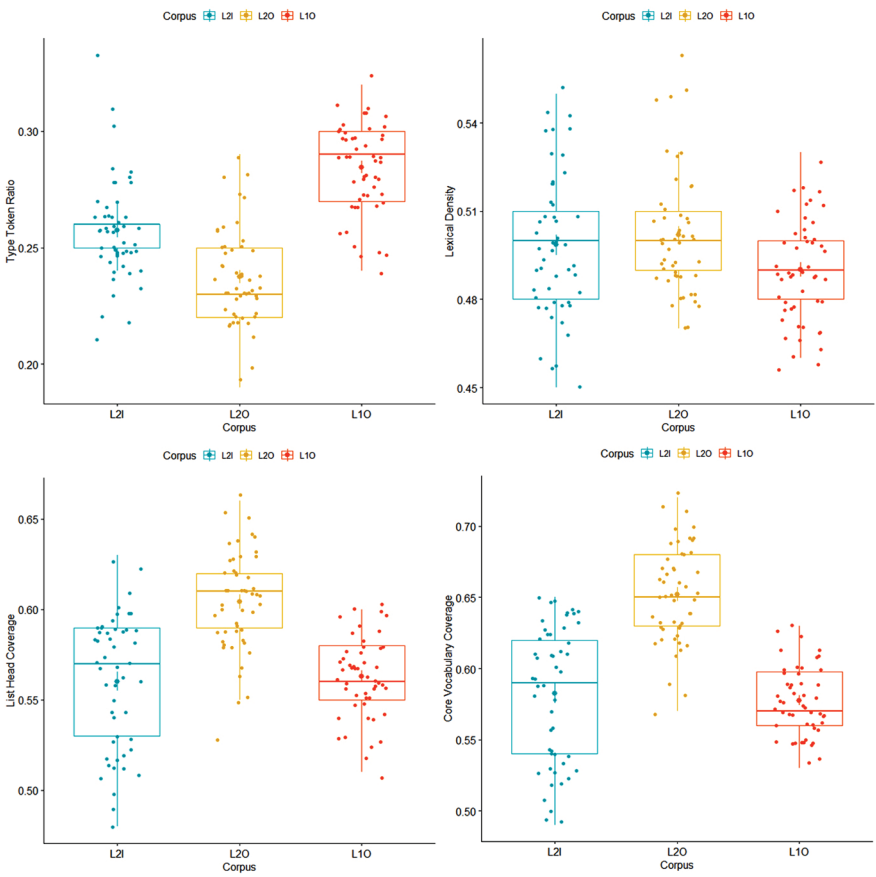

Figure 12.2 Comparison of mean and distribution of four lexical complexity indicators across three subcorpora

account for the increased lexical density, taking into consideration factors such as interpreting type, directionality, and specific strategies employed by interpreters (Dayter 2018; Kajzer-Wietrzny 2015; Shlesinger and Ordan 2012; Xu and Li 2022). For instance, Dayter (2018) discovered higher lexical density in situations where interpreters were required to engage in simultaneous interpreting "with text." This implies that the source speech was delivered from a meticulously planned script, featuring dense lexis and a more diverse structure. Consequently, the interpreter's output may exhibit higher lexical density. Additionally, as interpreters must work into their non-native language (B language), they tend to exercise caution and may resort to over-correction to ensure the quality of their rendition, thereby relying more heavily on intricate lexical options (Dayter 2018). However, it should be noted that in the present

study, the source speech consisted of impromptu speech in a question-and-answer format, which is different from Dayter's (2018) scenario.

Regarding the influence of interpreting directionality on lexical use, Sandrelli and Bendazzoli's study (2005) observed lower lexical density when the speech was interpreted from Italian into English, and higher lexical density in the opposite direction. However, it was unknown whether the interpreters were working into their B language in the instances of higher lexical density, making it challenging to compare the results with those of Dayter (2018). Some researchers expressed skepticism toward Dayter's explanation. Xu and Li (2022) pointed out that working into the interpreters' B language imposes greater cognitive demands, which may reduce lexical complexity and increase the use of high-frequency words in their output. Aligning the increased lexical density with interpreting strategy, Kajzer-Wietrzny (2015, 247) suggested that it "may be interpreters' avoidance of redundancy and the need to produce a very compact and dense text caused by severe time constraints in simultaneous interpreting." These assumptions about higher lexical density in interpreted speech appear valid to some extent. However, they also raise a fundamental question regarding the reliability of lexical density as an indicator of simplification in interpreting. Further empirical investigations are needed to examine how the proposed factors influence lexical density in interpreting.

Furthermore, it is intriguing to note that simplification is not exclusive to interpreted speech, as an even more pronounced manifestation is observed in L2 speech. Among all the investigated language varieties, L2 speech exhibits the least diverse lexical use and the highest degree of repetitiveness, as indicated by all the lexical indicators except for lexical density. This finding aligns with the patterns observed in previous studies on L2 use (Lanstyák and Heltai 2012; Liu et al. 2023). The observed simplification in L2 speech can be attributed to the nature of bilingualism-influenced communication and the associated high cognitive demands. According to processing efficiency theories (Hawkins, 2004), the tendency to simplify lexical options "is regarded as more efficient processing mechanism in contexts of high cognitive demand, since it reduces the processing effort of parsing syntactic and discourse relationships in linguistic production" (Kotze 2020, 120). When examining the three language varieties in terms of their cognitive demands, the production of L1 speech is highly automated due to the speaker's years of exposure to their first language (Liu, Cheung, and Liu 2023; Trebits 2014). In contrast, both interpreters and L2 speakers need to engage in significant cross-linguistic language processing and navigate various contextual challenges that do not exist in L1 speech production. This inevitably results in increased cognitive demands for the producers. Specifically, simultaneous interpreters must allocate attention to multiple sub-tasks simultaneously, such as active listening and analysis, memorization, speech production, and coordination (Gile 2009). Combined with time constraints, the need to keep pace with speakers, and the ethical requirement to constantly monitor the quality of their output, simultaneous interpreting poses significant cognitive demands on interpreters (Seeber 2011). This may

lead to behavioral pattern changes, whether consciously or unconsciously initiated by interpreters, in order to alleviate cognitive overload. A similar situation applies to L2 speech production. In order to generate proficient speech in their second language, L2 speakers must constantly address the inherent structural differences between the two languages and adapt to the contextual requirements of immediate oral communication (Vercellotti 2019). Due to the high cognitive demands, both interpreters and L2 speakers may not possess the same cognitive resources as L1 speakers do to enhance the lexical diversity of their output. Moreover, the presence of simplification in both interpreted speech and L2 speech may indicate that many observed features of translational language can also be present in other language varieties produced within similar socio-cognitive contexts (Chesterman 2010; Kruger and Van Rooy 2016; Lanstyák and Heltai 2012). Therefore, it is crucial to carefully examine the distinctions in their various manifestations to comprehend the characteristics of different language varieties.

12.6 Conclusion

The question of whether simplification can be considered a universal feature of translational language has been a central focus in corpus-based translation and interpreting studies. Laviosa (1998, 9) called upon researchers to explore simplification by examining lexical variation patterns across different text genres and mediation types, aiming to "establish whether and to what extent these regularities are subject field and/or modality specific and/or language specific, or whether they can indeed be considered universal features of translational English." In response to this call, the present study investigates the manifestation of simplification in simultaneous interpreting from Chinese into English using four lexical complexity indicators. Going beyond previous research, this study introduces L2 speech as a new reference language variety to explore whether simplification may also occur in other bilingualism-influenced language activities. While not all lexical indicators consistently support the presence of simplification in interpreting, there is a tendency observed in interpreted English speech to rely on less varied lexical options and more high-frequency words compared to L1 English speech. This finding partially supports the notion of simplification. Another intriguing result is the detection of a greater degree of simplification in L2 English speech, which can be attributed to the high cognitive demands involved in producing speech in a second language.

The findings of existing studies on simplification in interpreting reveal a wide range of perspectives, making it difficult to reach a definitive conclusion regarding its presence. Simplification is a phenomenon that heavily depends on the specific situation, posing methodological challenges in isolating and investigating the complex factors that contribute to its occurrence. To gain a more nuanced understanding of simplification in interpreting, further research efforts are needed to carefully consider the various contextual factors involved. Moreover, future studies could enhance their insights by incorporating a

broader set of simplification indicators, which would provide a more comprehensive picture of its operationalization and enable examination at different linguistic levels. Additionally, exploring the universality of simplification would benefit from the inclusion of additional reference language varieties. These efforts will contribute to a more coherent and comprehensive understanding of simplification in interpreting.

Notes

1 The L2I and L1O are existing corpora that were initially compiled in Xu and Li (2022). The L2O is a new sub-corpus added to PEDCC.
2 Unclassified words (with the tag Uncl, i.e., er, erm) and exclamations (with the tag Int, i.e., mm, mhm, ah, yeah, oh) were removed from the list of core vocabularies.
3 This represents infinitive marker "to" when it is combined with the preceding verb "gonna" and "wanna."
4 This is the initial part of "can't" (i.e., ca~ + n't).
5 This is the initial part of "gonna" (i.e., gon~ + ~na).
6 This is the initial part of "won't" (i.e., wo~ + n't).

References

Ai, Haiyang and Xiaofei Lu. 2010. "A Web-Based System for Automatic Measurement of Lexical Complexity." Paper presented at the 27th Annual Symposium of the Computer-Assisted Language Consortium (CALICO-10). Amherst, MA. June 8–12.

Baker, Mona. 1996. "Corpus-Based Translation Studies: The Challenges that Lie Ahead." In *Terminology, LSP, and Translation: Studies in Language Engineering in Honour of Juan C. Sager,* edited by Harold Somers, 175–86. Amsterdam: John Benjamins.

Bernardini, Silvia, Adriano Ferraresi, and Maja Miličević. 2016. "From EPIC to EPTIC—Exploring Simplification in Interpreting and Translation from an Intermodal Perspective." *Target* 28(1): 61–86.

Biber, Douglas, and James K. Jones. 2009. "Quantitative Methods in Corpus Linguistics." In *Corpus Linguistics: An International Handbook Volume 2*, edited by Anke Lüdeling and Merja Kytö: 1286–1304. Berlin: De Gruyter Mouton.

Chesterman, Andrew. 2010. "Why Study Translation Universals." *Acta Translatologica Helsingiensia* 1: 38–48.

Dayter, Daria. 2018. "Describing Lexical Patterns in Simultaneously Interpreted Discourse in a Parallel Aligned Corpus of Russian-English Interpreting (SIREN)." *Forum* 16(2): 241–64.

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): 717–38.

Gile, Daniel, 2009. *Basic Concepts and Models for Interpreter and Translator Training (Vol. 8)*. Amsterdam: John Benjamins.

Hawkins, John. 2004. *Efficiency and Complexity in Grammars*. Oxford: Oxford University Press.

Hu, Xianyao, Richard Xiao, and Andrew Hardie. 2019. "How do English Translations Differ from Non-Translated English Writings? A Multi-Feature Statistical Model

for Linguistic Variation Analysis." *Corpus Linguistics and Linguistic Theory* 15(2): 347–82.

Kajzer-Wietrzny, Marta. 2015. "Simplification in Interpreting and Translation." *Across Languages and Cultures* 16(2): 233–55.

Kotze, Haidee. 2020. "Translation, Contact Linguistics and Cognition." In *The Routledge Handbook of Translation and Cognition*, edited by Fabio Alves and Arnt Lykke Jakobsen, 113–32. Abingdon: Routledge. https://doi.org/10.4324/9781315178127-9.

Kranich, Svenja. 2014. "Translations as a Locus of Language Contact." In *Translation: A Multidisciplinary Approach,* edited by Juliane House, 96–115. Basingstoke: Palgrave Macmillan.

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 Helta. 2012. "Universals in Language Contact and Translation." *Across Languages and Cultures* 13(1): 99–121.

Laviosa, Sara. 1998. "Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose." *Meta* 43(4): 557–70.

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

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.

Liu, Yi, Andrew K. F. Cheung, Kanglong Liu. 2023. "Syntactic Complexity of Interpreted, L2 and L1 Speech: A Constrained Language Perspective." *Lingua* 286: 103509.

Lu, Xiaofei. 2012. "The Relationship of Lexical Richness to the Quality of ESL Learners' Oral Narratives." *The Modern Language Journal* 96(2): 190–208.

Lv, Qianxi and Junying Liang. 2019. "Is Consecutive Interpreting Easier than Simultaneous Interpreting?–A Corpus-based Study of Lexical Simplification in Interpretation." *Perspectives* 27(1): 91–106.

Russo, Mariachiara, Claudio Bendazzoli, and Annalisa Sandrelli. 2006. "Looking for Lexical Patterns in a Trilingual Corpus of Source and Interpreted Speeches: Extended Analysis of EPIC (European Parliament Interpreting Corpus)." *FORUM. Revue internationale d'interprétation et de traduction/International Journal of Interpretation and Translation* 4(1): 221–54.

Sandrelli, Annalisa, and Claudio Bendazzoli. 2005. "Lexical Patterns in Simultaneous Interpreting: A Preliminary Investigation of EPIC (European Parliament Interpreting Corpus)." In *Proceedings from the Corpus Linguistics Conference Series (Vol. 1)*. Birmingham: University of Birmingham.

Seeber, Kilian. 2011. "Cognitive Load in Simultaneous Interpreting: Existing Theories—New Models." *Interpreting* 13(2): 176–204.

Shlesinger, Miriam and Noam Ordan. 2012. "More Spoken or More Translated? Exploring a Known Unknown of Simultaneous Interpreting." *Target. International Journal of Translation Studies* 24(1): 43–60.

Stubbs, Michael. 1986. "Lexical Density: A Technique and Some Findings." *Talking About Text*: Studies presented to David Brazil on his retirement (Discourse Analysis Monograph 13). Birmingham: English Language Research, 27–48.

Thomason, Sarah and Terrence Kaufman. 1992. *Language Contact, Creolization, and Genetic Linguistics*. Berkeley: University of California Press.

Trebits, Aana. 2014. "Sources of Individual Differences in L2 Narrative Production: The Contribution of Input, Processing, and Output Anxiety." *Applied Linguistics* 37(2): 155–74.

Vercellotti, Mary Lou 2019. "Finding Variation: Assessing the Development of Syntactic Complexity in ESL Speech." *International Journal of Applied Linguistics* 29(2): 233–47.

Xu, Cui and Dechao Li. 2022. "Exploring Genre Variation and Simplification in Interpreted Language from Comparable and Intermodal Perspectives." *Babel* 68(5): 742–70.

Xu, Han. 2021. "Interprofessional Relations in Interpreted Lawyer-Client Interviews: An Australian Case Study." *Perspectives* 29(4): 608–24.

Xu, Han, Sandra Hale, and Ludmila Stern. 2020. "Telephone Interpreting in Lawyer-Client Interviews: An Observational Study." *Translation and Interpreting* 12(1): 18–36.

Appendix 1
List Head words in the three subcorpora

This list contains 100 most frequently used words extracted from L2I.

the	with	was	like
to	do	said	tax
and	can	up	whether
of	if	our	question
in	but	more	other
a	Hong	their	animal
that	Kong	when	how
we	has	some	out
is	or	because	any
for	about	these	very
i	by	all	which
have	government	what	under
you	would	them	want
be	n't	time	animals
it	also	members	services
are	should	may	many
will	people	one	thank
this	well	just	your
they	at	such	need
not	been	say	development
so	now	public	take
on	mr	my	president
as	then	years	new
there	from	think	work
's	an	no	secretary

This list contains 100 most frequently used words extracted from L2O.

the	right	from	an
to	but	more	been
and	with	also	say
of	about	if	out
a	very	was	see
you	Hong	actually	want
that	Kong	just	terms
we	on	our	different
in	there	when	by
is	or	their	my
so	people	lot	work
it	really	then	thing
i	because	one	get
’s	not	well	’ve
for	like	would	business
are	how	these	market
have	can	them	look
they	at	some	government
this	now	which	even
do	will	has	other
know	your	kind	going
think	all	time	go
as	’re	up	years
what	us	need	where
be	n’t	who	tell

This list contains 100 most frequently used words extracted from L1O.

the	on	they	friend
to	this	’s	their
and	be	what	an
that	with	there	all
of	not	from	would
in	as	minister	his
is	he	people	more
a	government	by	who
i	hon	which	very
we	has	right	work
for	my	so	am
will	but	been	those
are	can	do	was
have	at	house	or
it	our	about	if

uk	one	NHS	know
secretary	children	year	many
health	state	think	us
out	now	support	should
no	country	does	being
when	these	than	across
Ireland	northern	care	eu
need	also	time	him
make	them	important	she
ensure	up	new	Because

Appendix 2

This list contains the 200 most frequently used words extracted from the spoken component of the British National Corpus.[2]

the	so	now	an
I	got	your	time
you	've	had	who
and	not	were	want
it	are	about	like
a	if	two	come
's	with	said	really
to	no	'm	three
of	're	see	by
that	she	me	here
n't	at	very	put
in	there	out	has
we	think	my	good
is	yes	when	as
do	just	mean	does
they	all	right	cos
was	can	which	any
have	then	from	down
what	get	going	where
he	did	say	~na[3]
but	or	been	him
for	would	people	his
be	them	because	other
on	'll	some	five
this	one	could	something
know	up	will	their
well	go	how	these

our	okay	never	must
way	more	little	probably
ca~[4]	doing	than	saying
actually	done	tell	pounds
back	twenty	many	nine
four	went	same	Mr
'd	six	another	seven
should	give	money	also
gon~[5]	thought	point	wo~[6]
take	again	bit	big
thing	off	being	over
look	might	anything	of course
those	year	week	getting
why	her	too	made
things	last	ten	after
only	much	through	number
into	need	eight	old
us	day	new	find
quite	used	nice	coming
hundred	says	work	thirty
lot	still	always	fifty
make	sort	thank	as well
first	years	next	thousand