International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011

INTERNATIONAL JOURNAL OF MODERN EDUCATION (IJMOE)
www.ijmoe.com

GLOBAL ACADEMIC EXCELLENCE

CORPUS-BASED TRANSLATION STUDY ON THE SIMPLIFICATION OF CHINESE SUBTITLES IN ENGLISH-LANGUAGE FILMS

Su Tingting[1*], Mohamed Abdou Moindjie[2], Manjet Kaur Mehar Singh[3]

[1] School of Languages, Literacies & Translation, Universiti Sains Malaysia, Malaysia
Email: sttserena@163.com
[2] School of Languages, Literacies & Translation, Universiti Sains Malaysia, Malaysia
Email: mohdmoindjie@usm.my
[3] School of Languages, Literacies & Translation, Universiti Sains Malaysia, Malaysia
Email: manjeet@usm.my
[*] Corresponding Author

Article Info:

Article history:
Received date: 30.09.2024
Revised date: 14.10.2024
Accepted date: 19.11.2024
Published date: 22.12.2024

To cite this document:

Su, T., Moindjie, M. A., & Singh, M. K. M. (2024). Corpus-Based Translation Study On The Simplification Of Chinese Subtitles In English-Language Films. *International Journal of Modern Education, 6* (23), 142-154.

DOI: 10.35631/IJMOE.623011

This work is licensed under CC BY 4.0

Abstract:

While previous research has explored simplification in literary translation studies, this study's focus on universals in audiovisual translation studies(AVT), especially in subtitling. In the era of advancing technology, where machine translation and automated translation analysis tools play a crucial role, the study of simplification as a universal phenomenon provides valuable insights. Understanding which translation features are universal across different language pairs can guide improvements in modern translation analysis tools and methods, and enhance translation accuracy. This study delves into the realm of translation studies to investigate simplification in Chinese subtitles of English-language films compared to Chinese subtitles of Chinese-language films. The research explores the presence of simplification in translation by examining three key aspects: lexical variety, lexical density, and high-frequency words. The findings reveal that Chinese subtitles of English-language film , as translated text, exhibit a reduced level of lexical variety when compared to Chinese subtitles of Chinese-language films, as the non-translations. Specifically, Chinese subtitles of English-language films employ more function words and fewer content words, resulting in lower lexical density than Chinese subtitles in Chinese-language films. The statistical analysis of high-frequency words further supports the notion that high-frequency words are used more frequently in translated Chinese subtitles than in the original Chinese subtitles. These results align with the simplification hypothesis proposed by Laviosa (1998) and Hu (2007). This research aims to enhance the comprehension of the intricacies of Chinese translation, not only across linguistic boundaries but also in the realm of technology, contributing to the reference of translation studies and practice.

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011

Keywords:

Audiovisual Translation, Corpus-based Translation Studies, Simplification, Subtitles, Universals

Introduction

Subtitling is a thriving research area within the broader discipline of translation studies, encompassing various fields such as technology, linguistics, cross-cultural communication, and translation. With the rapid evolution of technology, subtitling has seen new opportunities and challenges, drawing widespread academic attention. Many studies in subtitling have concentrated on areas such as translating humor (De Rosa et al., 2014), swear words (Mattsson, 2006), gender (De Marco, 2012), or multimodality (Chen, 2019). And most of these previous studies were based on individual films and case studies. But translation studies should not be limited solely to individual films or case studies, although they can offer some in-depth insights. Following the guidance of Descriptive Translation Studies (DTS), there is necessity to move beyond individual case studies and isolated descriptions in order to discover patterns (such as universals, laws, norms, strategies, and procedures) presents additional challenges associated with the creation, access, and analysis of corpora (Gambier & Pinto, 2018).

Thanks to the development of computer technology, there have been endeavors to create larger corpora to identify patterns across a broader range of audiovisual products in various genres, decades, and countries in the last decade, such as the Veiga corpus, a multimedia corpus of subtitled films used to analyse both English intralingual subtitling and English-Galician interlingual subtitling (Sotelo Dios, 2015), the CORSUBIL corpus by Rica Peromingo (2014) built to obtain a list of lexical units of English or Spanish subtitles in American films, and the OpenSubtitle2018 corpus (Lison & Tiedemann, 2016), which is a new collection of translated movie subtitles from http://www.opensubtitles.org. While there have been corpus-based translation studies of subtitles in many languages like English and Spanish, there is a notable lack of research on Chinese subtitles.

Baker (1993) served as a pivotal figure, marking the initiation of corpus-based studies in translation universals. Following Blum-Kulka's (1983) explicitation hypothesis, various new universals hypotheses have emerged. For instance, Vanderauwera (1985) explored disambiguation, simplification, and generalization in translated texts, while Laviosa-Braithwaite (1996) investigated simplification features within translation universals. Kenny(2014) compared the presence of generalization features in German literary texts and their English translations. From Toury's (1979) "inter-language" to Frawley's (1984) "third code" and further to Olohan's (2004)'s "features of translation", the research focus has shifted from socio-cultural and historical factors to the linguistic characteristics of translation itself (Hu & Zeng, 2011). The concept of translation universals has faced continuous scrutiny and challenges since its inception. Chesterman (2004) made a significant distinction in universals hypotheses between potential source-language universals (S-universals) and potential target-language universals (T-universals).

Chesterman (Chesterman, 2004) explored the essence of universals and linguistic relationships by introducing three types of textual comparisons: translations, source texts and non-translations, and two linguistic relations: equivalence and textual fit. Based on the two

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011
relationships among the three types of texts, Chesterman further distinguishes two types of universals: S-universals and T-universals. Potential S-universals refers to source text related universals, including: lengthening, the law of interference, the law of standardization, dialect normalization, reduction of complex narrative voices, the explicitation hypothesis , sanitization, the retranslation hypothesis, reduction of repetition, while potential T-universals are comparison between translations and non-translations, including: Simplification, Conventionalization, Untypical lexical patterning, Under-representation of TL-specifific items.

Literature Review

Simplification is often studied in two aspects. Firstly, as a translation technique involving the reduction of certain aspects of the original text. Simplification as a technique aims to enhance the readability by eliminating redundancies or inappropriate content present in the source text, such as repetitions, sensitive or offensive language. This simplification is between the source language and the target language. Secondly, simplification is also investigated as the universal in translation, referring to the overall tendency of translated texts to have a simplified language compared to non-translations as described by Baker (1996). Key simplification features in corpus-based translation studies include generalization, a decrease in lexical density (more functional words and fewer content words), simplification of sentence structures (Hu, 2020).

Simplification is defined by Blum-Kulka and Levenston (1983, p. 119) as "the process and/or result of making do with less words". Lavisa (1998) compared translated and non-translated texts within the English Comparable Corpus (ECC), focusing on specific linguistic features(lexical density, mean sentence length, list heads and high-frequency words). The research results indicates that the translated texts has lower lexical density and repetition of the most frequent words, which is a strong evidence of the simplification of the universals hypothesis. Zanettin (2000), based on an examination of a parallel corpus of 1.5 million words in English and Italian, found that translated texts exhibit lower lexical variation (type-token ratio) compared to original texts. Hu (2007)conducted research on the vocabulary characteristics of Chinese translated novels and identified a tendency for reduced lexical variation and increased word frequency for common words in translated Chinese. Ippolito (2014) built a comparable corpus containing classic fiction for children, comprising both translated and non-translated works. The findings indicate that within the translational subset, processes such as simplification, explicitation, and normalization do not dominate over the non-translational subset. Seracini (2021) conducted a research using a parallel corpus of EU legislative texts translated from English into Italian and a reference corpus of national non-translated legislation. The findings revealed evidence of simplification, manifesting in the omission of unnecessary repetitions, avoidance of complex sentence structures, and a preference for the active voice. Pastor et al (2008) initiated a project applying Natural Language Processing (NLP) techniques to comparable corpora of Spanish translated and non-translated texts. The research findings indicated that translated texts exhibited significantly lower lexical density and richness compared to non-translated texts, and appeared to be more readable. Surprisingly, translated texts had a notably lower proportion of simple sentences and shorter sentences. The simplification features were more pronounced in technical translation corpora and even more so in medical translation corpora.

The research on simplification mentioned above spans various languages and different domains, yielding different results. Some scholars have conducted their studies based on general corpora, while others have employed specialized or self-compiled corpora specific to

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011

particular fields. Regardless of the approach, all these studies serve as valuable supplements in the exploration of universals. They provide additional support and evidence for understanding common patterns in translation and offer insights into how simplification varies in different linguistic and domain contexts.

Although existing studies have provided many insights into the phenomenon of simplification in translation, they have also limited in scope, depth, and methodology. First, many studies focus on specific types of texts, such as literature or technical documents (Chen & Chang, 2023), while neglecting the increasingly important field of film subtitles. In addition, some studies use corpora of limited size, which may not fully reflect the diversity of simplification phenomena (Xu et. al, 2026). Therefore, this study aims to expand the scope of research on translated text genres, with a special focus on subtitle translation. It is an area that has been less explored in previous studies. Subtitle translation involves not only the conversion of languages, but also the bridging of cultural differences and the optimization of audience understanding. By shifting the research perspective from a single film case analysis to a comparative analysis of multiple films (corpus), this study seeks to reveal the prevalent patterns and trends in subtitle translation. In addition, this study has made methodological innovations, drawing on corpus tools widely used in grammar translation and linguistics research, in order to provide richer and more systematic data support for subtitle translation research.

Research Methods

Data Collection

The corpus constructed for this study is a comparable corpus, primarily consisting of two monolingual corpora with similar sizes: the Chinese Subtitles of English-language Films and the Chinese Subtitles of Chinese Films. There are many sampling methods, purposive sampling is commonly used in corpus-based studies (Saldanha & O'Brien, 2014). The samples choice mainly considers the following aspects: 1) Diversity: These films encompass various film genres, including drama, science fiction, action, horror, comedy, etc. This diversity helps in studying translation differences in different types of films; 2) Time Periods: The selected films span different decades, from classic films from the 1960s to contemporary modern films; 3) Directors and Actors: Some films are directed by renowned directors or feature well-known actors. These films may have unique language usage and translation characteristics and are of higher quality and research value. 4) Cultural Backgrounds: These films cover different cultural backgrounds, including the United States, the United Kingdom, Australia, etc; 5) Language Styles : Some films may contain informal, slang, or specialized language styles, which may require special handling in translation; 6) Audience : These films are usually widely popular, so their subtitle translations may undergo careful scrutiny to ensure they are suitable for a broad international audience; 7) Awards and Influence: Some films are Oscar-winning films or have had a profound impact, and they may have higher translation requirements to maintain the quality of the original work.

Taking the mentioned factors into consideration, this study has chosen the Chinese subtitles of 50 English-language films. Through a word count analysis of the text in these 50 film's subtitles, a total of 500,391 words were identified (after tokenization). It's important to note that comparability is a key characteristic of a comparable corpus, with the aim of minimizing potential interference from other variables that could affect the results (Saldanha & O'Brien, 2014). In order to enhance comparability, meaning similarity in type and size, a selection of 70

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011

Chinese language films was made, which yielded a word count of 511,117 words. The two corpora are comparable in terms of both type and quantity. The primary sources of data for this research are derived from OpenSubtitles(http://www.opensubtitles.org) for Chinese subtitles of English-language films and Shooter(https://assrt.net) for Chinese subtitles of Chinese-language films.

Data Collation

The procedures for data collation including text transition, data clean, data tokenization, and data annotation. Firstly, to facilitate the analysis process, the subtitles undergo format conversion using Aegisub, transitioning from .srt to .txt format. Secondly, in order to get clean text, various mistakes such as "\N" are replaced with spaces, as they serve as defaults generated by software. Thirdly, the collected data, after being tokenized and annotated using Sketch Engine (https://www.sketchengine.eu), becomes available for statistical analysis and can be used to generate word lists that include part-of-speech tags and annotations.

Data Analysis

This study focuses on comparing English-language film subtitles to Chinese-language film subtitles to explore the presence of simplification. The research primarily employs statistical analysis in the following areas: 1) lexical variety, 2) lexical density, and 3) high-frequency words. With the help of Voyant Tools (https://voyant-tools.org/), metrics such as TypesCountMean, TypesCountStdDev, Vocabulary density, RelativeFrequency can be employed to determine which corpus exhibits higher lexical diversity. Lexical density is primarily concerned with examining the distribution of content words and function words in the two corpora. High-frequency words, on the other hand, are compared using a word list generated through Sketch Engine to assess their usage across the two corpora.

Discussion and Findings

Lexical Variety

Currently, there are numerous readily available statistical tools that, upon input of text, can generate relevant data. From the data generated by Voyant Tools, metrics are directly exported for the three corpora, including TokensCount, TypesCount, TypesCountMean, and TypesCountStdDev, as shown in the table 1 below. Among these, TypesCount signifies how many different vocabulary words are present, while TokensCount represents the total number of occurrences of these words in the text. TypesCountMean, calculated as TokensCount divided by TypesCount, serves as a measure of lexical variety. When the value of TokensCount/TypesCount is higher, it indicates lower lexical variety in the text, meaning that relatively fewer different vocabulary types are frequently repeated. Conversely, if TokensCount/TypesCount has a smaller value, it signifies higher lexical variety in the text, with different vocabulary types being more evenly distributed throughout the text. TypesCountStdDev refers to the standard deviation of different vocabulary types (or words in the vocabulary list). Standard deviation is a statistical measure used to gauge the degree of dispersion or variation in a set of data. If TypesCountStdDev is higher, it suggests a significant variation in the distribution of different vocabulary types in the text, meaning that some vocabulary types may appear very frequently while others appear less often. This may indicate an uneven distribution of vocabulary in the text, with certain vocabulary types dominating. On the contrary, if TypesCountStdDev is small, it suggests a more uniform distribution of different vocabulary types in the text, with minimal variations in frequency among vocabulary types.

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011

This may indicate a relatively even distribution of vocabulary in the text, without any distinct dominant vocabulary types (Hu, 2020). CSEF here represents the Chinese Subtitles of English-language Films corpus, while CSCF represents the Chinese Subtitles of Chinese-language Films corpus.

Table 1: Basic Information of Two Corpora

Title	TokensCount	TypesCount	TypesCountMean	TypesCountStdDev
CSEF	500391	18890	26.48973	343.02615
CSCF	511117	19651	26.00972	326.06952

Through comparison, it can be observed that, between two corpora of similar sizes, the TypesCountMean in CSEF is higher than that in CSCF. Thus, from an overall perspective, the lexical variety in Chinese subtitles of English-language films (translated text) appears to be lower than in Chinese subtitles of Chinese films (non-translation). CSEF also exhibits a higher TypesCountStdDev compared to CSCF, indicating significant variations in the distribution of different types of vocabulary in CSEF texts. Some vocabulary types may occur very frequently, while others are less common. This suggests that the vocabulary distribution in CSEF is relatively uneven, with certain vocabulary types dominating. Both sets of data support the hypothesis of simplification in Chinese translations of English-language movie subtitles.

Lexical Density

Furthermore, the parts of speech tagging for these three corpora were coded for statistical analysis. Part of speech tagging, using Sketch Engine, was categorized into three main classes for analysis: Content words, which include verbs (v), nouns (n), adverbs (a), adjectives (j), and numerals (m); Function words, which consist of pronouns (d), prepositions (p), conjunctions (c); and Others, encompassing habitual language (i) and miscellaneous categories (x). The Chinese Penn Treebank part-of-speech tagset, which is integrated into Sketch Engine, is employed for the part of speech tagging and analysis.

Two methods were employed to calculate lexical density. The first method, TTR (Type/Token Ratio), is akin to TypesCountMean mentioned earlier and is not reiterated here. The second method involved calculating the proportion of content words and function words as a percentage of the total tokens in the two corpora. The results are summarized in the following table 2.

Table 2: Percentage of Content Words and Function Words

Corpus	Content Words	Function Words
CSEF	67.18%	19.84
CSCF	69.61	16.89

Through a comparison of the proportion of content words and function words, it was found that CSEF contain more function words and fewer content words compared to CSCF. This suggests that the number of content words used in the Chinese subtitles of English-language film is lower than in the original Chinese. On the other hand, the use of function words in English-language film subtitles is higher than in Chinese-language film subtitles. This indicates that the use of function words in translated Chinese is higher in the subtitles than in the original

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011

Chinese. An increase in the use of function words and a decrease in content words suggest lower lexical density. This, in turn, makes the subtitles more easily comprehensible to the movie's audience in a limited amount of time. The lower lexical density also supports the simplification hypothesis for translated texts. That is, Chinese translated texts tend to reduce complexity by lowering the information provided by content words, making the text more acceptable overall.

The specific performance of content words and function words in terms of frequency per million can be observed in Figure 1.

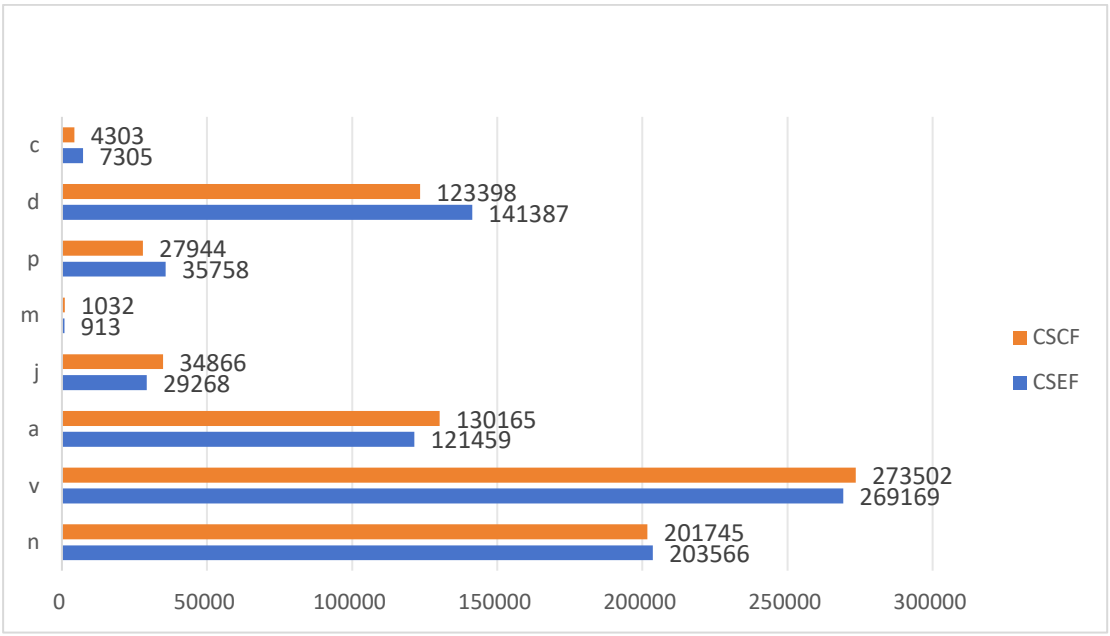

Figure 1: POS Frequency Per Million

Because the two corpora have some variation in the total number of tokens, and because as the size of the corpus increases, the number of tokens also increases while the number of types tends to stabilize, using frequency per million within a relative comparable range is a more informative metric.

According to the data provided by Sketch Engine, in the case of Content words (v, n, a, j, m), except for nouns, CSCF has slightly lower frequency per million values than CSEF. This might be due to the higher number of nouns, potentially stemming from the presence of loanwords, technical terms, creative vocabulary, and culturally loaded words in English-language films. For example, terms like "上帝 (God)," "麦克 (Michael)," "宝贝 (baby/sweetheart)," "人类 (human)," "国王 (king)," "魔戒 (the Ring)," are foreign terms based on English culture and may have minimal presence in Chinese culture. English-language film subtitles often contain many loanwords and technical terms that may require explanation or annotation to help the audience better understand the film's content. Some English-language films may include creative vocabulary, such as fictional place names and brand names. These terms might require special translation or annotation to convey their meanings to the audience. The goal of Chinese subtitles is to ensure that the audience comprehensively understands the content of the film, which may necessitate providing more vocabulary and information to bridge language and cultural differences.

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011

From figure 1, it is evident that function words (d, p, c), particularly in the case of CSEF, exhibit higher frequency per million values compared to CSCF. In CSEF, pronouns with frequency per million exceeding 1000 include "我 (I), 你 (you), 他 (he), 我们 (we), 他们 (they), 这 (this), 她 (she), 什么 (what), 你们 (you all), 那 (that), 它 (it), 谁 (who), 这里 (here), 自己 (oneself), 您 (you)." In CSCF, pronouns with frequency per million exceeding 1000 include "你 (you), 他 (he), 我们 (we), 她 (she), 你们 (you all), 什么 (what), 这 (this), 谁 (who), 他们 (they), 您 (you), 那 (that), 自己 (oneself), 这里 (here)." When examined individually, the frequency per million of "我 (I)" in CSEF is 45,826, while in CSCF, it is 39,980, indicating a significantly higher usage of pronouns in CSEF.

In CSEF, prepositions with a frequency per million exceeding 1000 include "在 (at/in), 对 (to/for), 跟 (with), 上 (on/above), 给 (to/give), 里 (inside), 从 (from)." In CSCF, the prepositions with a frequency per million exceeding 1000 include "在 (at/in), 跟 (with), 给 (to/give), 对 (to/for), 上 (on/above), 里 (inside)." The trend in prepositions suggests that, for the sake of maintaining sentence structure consistency and fluency, more function words may need to be added in Chinese sentences during translation to ensure that the sentences are coherent and easy to understand.

Conjunctions in CSEF with a frequency per million exceeding 100 include "和 (and), 如果 (if), 只要 (as long as), 要是 (if), 若 (if), 或 (or), 跟 (with), 还是 (or), 与 (and), 还有 (and), 或者 (or), 及 (and), 不管 (no matter), 或是 (or), 除非 (unless), 虽然 (although), 以及 (and)." In CSCF, conjunctions with a frequency per million exceeding 100 include "和 (and), 如果 (if), 要是 (if), 跟 (with), 只要 (as long as), 还是 (or), 既然 (since), 若 (if), 虽然 (although), 与 (and), 还有 (and)." This frequency trend indicates that, to maintain sentence structure consistency and fluency, additional function words may be necessary in Chinese sentences during translation to ensure that the sentences are coherent and easily understood.

High Frequency Words

Sketch Engine has the capability to automatically generate word lists sorted by token frequency. The statistics for the top 10 token frequencies in the two corpora are as showed in table 3.

International Journal of Modern Education
EISSN: 2637-0905
IJMOE
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011

Table 3: Top 10 Frequent Words in Two Corpora

CSEF					CSCF				
word	tag	pos	Freq. per million	% of concordance	word	tag	pos	Freq. per million	% of concordance
我	PN	d	45826	4.9	我	PN	d	39980	4.3
你	PN	d	37322	4.0	你	PN	d	39819	4.3
是	VC	v	25582	2.8	不	AD	a	25976	2.8
的	DEG	x	23445	2.5	是	VC	v	22928	2.5
不	AD	a	21056	2.3	了	AS	x	19538	2.1
了	AS	x	18466	2.0	的	DEG	x	14419	1.6
的	DEC	x	16404	1.8	的	DEC	x	12558	1.4
他	PN	d	12799	1.4	就	AD	a	10288	1.1
我们	PN	d	9646	1.0	一	CD	j	9153	1.0

These two corpora share a common high-frequency pronoun, "我(I/me)," followed by "你 (you)." In CSEF, there are additional pronouns, "他 (he/him)" and "我们 (we/us)," which are not present in CSCF. And "我 (I/me)" in CSEF exhibits higher frequency per million and a higher percentage of concordance compared to CSCF. The highest-frequency verb in both corpora is "是 (to be)," with CSEF showing higher frequency per million and a greater percentage of concordance for this verb compared to CSCF. Regarding the usage of "的," it has two forms in Chinese. One is "的" in a relative clause (DEC), for example: "我认识的人很多 (I know many people)". The other is "的" in an associative sense (DEG), for example: "一瓶水的价格很便宜 (The price of a bottle of water is very cheap)". In both cases, whether DEG or DEC, "的" in CSEF exhibits higher frequency per million and a higher percentage of concordance compared to CSCF.

Surprisingly, "不 (not)" in CSEF has a lower frequency per million and a lower percentage of concordance compared to CSCF. This led to an exploration of the contexts in which "不" is used, revealing instances where "不" is not employed for negation but rather to express questioning or a skeptical tone. For example, phrases like "饿不饿 (Are you hungry?)" and certain dialectal expressions like "中不？(Alright?)" use "不" as an interrogative modal particle. In phrases like "过不过分 (Is it excessive?)," the "不" is used to indicate a questioning or skeptical tone. In these cases, "不" is not used for negation but rather to emphasize the interrogative mood.

However, analyzing only the top 10 high-frequency characters/words is insufficient to fully understand the usage of high-frequency words. Therefore, high-frequency words in the top 10, top 30, top 50, and top 100 were analyzed in stages, measuring their frequency per million. The results are as follows:

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011

Table 4: Frequency Count of High Frequency Words

Ranking	CSEF	CSCF
10	219439	203571
30	343436	324523
50	403717	386977
100	481127	472644

As indicated by the data in the table 4, the statistics at each stage consistently demonstrate that the frequency of high-frequency words in CSEF is higher than in CSCF. Therefore, in summary, the overall trend suggests that the frequency of high-frequency words in the Chinese subtitles of English-language film is higher than in the Chinese subtitles of Chinese-language film. In other words, the usage frequency of high-frequency words in translated Chinese is higher than in original Chinese, supporting the hypothesis that translated texts employ more common words, thus aligning with the simplification hypothesis.

Cultural factors have a profound impact on the simplification of Chinese subtitles for English films, which is mainly reflected in differences in language structure (such as the omission of subjects), adaptability of cultural references (such as the preference for the use of four-character idioms), audience expectations and acceptance(such as the preference of short subtitles), handling of sensitive content (omission of profanity) and other translation strategies, as well as the balance between globalization and localization. Translators must take into account the cultural background of the target language and the reading habits of the audience while maintaining the original meaning. All these factors often lead to the use of more concise and direct expressions in subtitles to ensure the rapid transmission and easy understanding of information while avoiding cultural sensitivity issues and offensive content. This comprehensive consideration makes subtitle translation a complex cross-cultural communication activity that aims to achieve a harmonious unity between the effective communication of the source text information and the smooth acceptance of the target cultural audience.

Conclusion

This study has provided evidence for the universals of simplification in Chinese subtitles of English-language film through three main aspects: lexical variety, lexical density, and high-frequency words. The research demonstrates that the translated text in Chinese subtitles of English-language film exhibits lower lexical variety compared to the Chinese subtitles of Chinese-language film. Chinese subtitles of English-language film use more function words and fewer content words, resulting in lower lexical density than Chinese subtitles of Chinese-language film. The statistical analysis of high-frequency words also indicates that the translated text in English-language film employs high-frequency words relatively more frequently than non-translations of Chinese-language film. These research findings align with the hypotheses of simplification as validated by Laviosa (Laviosa, 1998) and Hu (Hu, 2007).

Despite its important implications for translation studies, this study has some limitations. First, the sample selection was limited to a certain number of films and may not be able to fully represent all types of film subtitle translation practices. In addition, the films in the sample set mainly originated from a specific time period and region, which may limit the generalizability of the research results. In terms of analytical methods, although quantitative analysis provides relatively persuasive insights into the linguistic characteristics of subtitles, it may not capture

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011

all contextual and emotional nuances in subtitle translation. In addition, the statistical tools and algorithms used may have inherent limitations, which may affect the interpretation of the data. In terms of cultural factors, this study may not fully consider the diversity of the target audience's cultural background and the complexity of dealing with cultural adaptation in subtitle translation. Future research can overcome these limitations by expanding the sample range, adopting a mixed method research design, and exploring in depth the impact of cultural factors on subtitle translation.

While the field of exploring universals in translation studies has seen a considerable volume of research on simplification, the differences between languages, specialized domains, text types, and more continue to make the exploration of universals a meaningful endeavor. Translation is a complex interplay of language and culture. Studying simplification in translation allows for a deeper understanding of the phenomenon, shedding light on the differences between languages, translation strategies, and the impact of cultural factors on translation processes. Simultaneously, these related studies can enrich translation studies and offer valuable insights into understanding and addressing common features in translation research. Furthermore, in the context of modern technological advancements, machine translation and automated translation tools have become increasingly significant. Investigating simplification as a universal can provide guidance for the enhancement of these tools. Understanding which translation features are universal across different language pairs contributes to improving the accuracy of automated translation.

Acknowledegment
The authors would like to acknowledge opensubtitles.com for the free subtitle resources.

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* (233-250). John Benjamins.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. In H. Somers (Ed.), *Terminology, LSP and Translation* (18, pp. 175-186). John Benjamins Publishing Company.

Blum-Kulka, S., & Levenston, E. (1983). Universals of lexical simplification. In C. Faerch & G. Kasper (Eds.), *Strategies in Interlanguage Communication*. Longman.

Chen, J., & Chang, H. (2023). Testing simplification in translated and creative writing texts: The case of Robert van Gulik's Judge Dee detective stories. Across Languages and Cultures. https://doi.org/10.1556/084.2023.00375.

Chen, Y. (2019). *Translating Film Subtitles into Chinese：A Multimodal Study*. Springer.

Chesterman, A. (2004). Hypotheses about translation universals. In G. Hansen, D. Gile, & K. Malmkjær (Eds.), *Claims, Changes and Challenges in Translation Studies* (1-14). Benjamins Translation Library. http://digital.casalini.it/9789027295552.

De Marco, M. (2012). *Audiovisual translation through a gender lens*. Rodopi.

De Rosa, G. L., Bianchi, F., De Laurentiis, A., & Perego, E. (2014). *Translating humour in audiovisual texts*. Peter Lang.

Frawley, W. (1984). Prolegomenon to a Theory of Translation. In W. Frawley (Ed.), *Translation: Literary, Linguistic, and Philosophical Perspectives* (159-175). Associated University Presses.

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011
Gambier, Y., & Pinto, S. R. (2018). *Audiovisual translation: Theoretical and methodological challenges*. John Benjamins Publishing. https://doi.org/10.1075/btl.108

Hu, X. (2007). A Corpus-based study on the lexical features of Chinese translated fiction. *Foreign Language Teaching and Research*(3), 214-220.

Hu, X. (2020). *Corpus Stylostatistics: Methods and Applications*. Foreign Language Teaching and Research Press.

Hu, X., & Zeng, J. (2011). New Trends in the Corpus-based Research on Translation Universals. *Journal of PLUniversity of Foreign Languages*(1), 56-62.

Ippolito, M. (2014). *Simplification, Explicitation and Normalization: Corpus-Based Research into English to Italian Translations of Children's Classics*. Cambridge Scholars Publishing.

Kenny, D. (2014). *Lexis and Creativity in Translation: A Corpus Based Approach*. routledge.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *META*, 43(4), 557-570. http://doi.org/https://doi.org/10.7202/003425arCopiedAn error ha

Laviosa-Braithwaite, S. (1996). *The English Comparable Corpus (ECC): A Resource and a Methodology for the Empirical Study of Translation* [PhD, University of Manchester].

Lison, P., & Tiedemann, J. (2016) *OpenSubtitles2016: Extracting Large Parallel Corpora from Movie and TV Subtitles*. Paper presented at the 10th International Conference on Language Resources and Evaluation (LREC 2016).

Mattsson, J. (2006) *Linguistic variation in subtitling: The subtitling of swearwords and discourse markers on public television, commercial television and DVD*. Paper presented at the MuTra 2006–Audiovisual Translation Scenarios

Olohan, M. (2004). *Introducing Corpora in Translation Studies* . Routledge.

Pastor, G. C., Mitkov, R., Afzal, N., & Pekar, V. (2008*Translation universals: do they exist? A corpus-based NLP study of convergence and simplification*. Paper presented at the the 8th Conference of the Association for Machine Translation in the Americas.

Rica Peromingo, J. P., Albarrán Martín, R., & García Riaza, B. (2014). New Approaches to Audiovisual Translation: The Usefulness of Corpus Based Studies for the Teaching of Dubbing and Subtitling. In E. Bárcena, T. Read, & J. A. Hita (Eds.), *Languages for Specific Purposes in the Digital Area (Series: Educational Linguistics)* (19, pp. 303-322). Springer-Verlag.

Saldanha, G., & O'Brien, S. (2014). *Research methodologies in translation studies*. Routledge.

Scarpa, F. (2006). Corpus-based QualityAssessment of Specialist Translation: A Study Using Parallel and Comparable Corpora in English and Italian. In M. Gotti & S. Sarcevic (Eds.), *Insights into Specialized Translation* (155-172). Peter Lang.

Seracini, F. L. (2021). Translation Universals in Legal Translation: A Corpus-based Study of Explicitation and Simplification. *Translation Quarterly*(101), 67-91.

Sotelo Dios, P. (2015). Using a Multimedia Corpus of Subtitles in Translation

Training: Design and Application of the Viega Corpus. In A. Lenko-Szymanska & A. Boulton (Eds.), *Multiple Affordances of Language Corpora for Data-Driven Learning* (238-264). Benjamins.

Toury, G. (1979). Interlanguage and its manifestations in translation. *META*, 24(2), 223-231.

Vanderauwera, R. (1985). *Dutch Novels Translated into English: The Transformation of a Minority Literature* . Rodopi.

Xu, W., Napoles, C., Pavlick, E., Chen, Q., & Callison-Burch, C. (2016). Optimizing Statistical Machine Translation for Text Simplification. Transactions of the Association for Computational Linguistics, 4, 401-415. https://doi.org/10.1162/tacl_a_00107.

International Journal of Modern Education
IJMOE
EISSN: 2637-0905
Volume 6 Issue 23 (December 2024) PP. 142-154
DOI: 10.35631/IJMOE.623011

Zanettin, F. (2000). Parallel corpora in translation studies: lssues in corpus design anoanalysis. In M. Olohan (Ed.), *Intercultural Faultlines: Research Models in TranslationStudies 1: Textual and Cognitive Aspects* (105-118). St. Jerome.