https://doi.org/10.24303/lakdoi.2024.32.2.153

Sensitivity of Translation Universals to Genre/Register Variations: Focused on Corporate Reporting*

Jin Yim & Yong-hun Lee**
(Ewha Womans University & Chungnam National University)

Yim, Jin & Lee, Yong-hun. (2024). Sensitivity of translation universals to genre/register variations: Focused on corporate reporting. *The Linguistic Association of Korea Journal*, *32*(2), 153-173. This article aims to determine which specific linguistic feature more effectively distinguishes translated texts (TTs) from non-translated texts (NTTs) in the corporate reporting genre by adopting representative and comparable corpora with the help of conditional inference trees (CIT) and random forests (RF). Based on the widely explored translation theory of translation universals (TUs) and ample empirical evidence in this area, we selected linguistic features that were proven effective in showing the differences between TTs and NTTs. Among the numerous features known to represent the four categories of TUs—simplification, explicitation, normalization, and leveling out—we selected ten factors to find out how effective they are in distinguishing TTs from NTTs. After encoding ten factors in two monolingual, specialized corpora that consist of a total of 58 English management forewords of sustainability reports including TTs and NTTs, we analyzed the encoded data with conditional inference trees (CIT) and random forests (RF). The results show that, unlike our expectations, only simplification plays a significant role in identifying TTs from NTTs. This contrasts sharply with previous studies on the Korean-English language pair. The results not only confirm the considerable impact of genre variations on TU behaviors but also present the effectiveness of CIT and RF in corpus-based translation studies.

Key Words: Translation universal, genre variation, corpus linguistics, corporate reporting, conditional inference tree, random forest

* We wish to thank the three anonymous reviewers for their insightful and meaningful comments. Despite their contributions, any errors in this article remain our own.

** Jin Yim is the first author and Yong-hun Lee, the corresponding author.

1. Introduction

Translated texts have generally been regarded as different from texts that are not translated. Many researchers agree that translated texts have linguistic features that distinguish them from non-translated ones. In this sense, translated texts have been referred to by various labels. Frawley views them as a "third code" (Frawley, 1984, p. 168) between the source and target codes, while Toury (1979, p. 223) tried to delineate translation's unique aspects by borrowing the linguistic theory of "interlanguage" originally put forward by Selinker (1972). However, the most sought-after theory to denote the distinctive nature of translation is inarguably translation universals (TUs), proposed by Baker (1993).

According to various scholars who have tested the theory, translated texts exhibit distinctive linguistic features compared to their comparable non-translated texts. The hypothetical theory of TUs has been extensively explored on around four major tendencies —simplification (Blum-Kulka & Levenston, 1978), explicitation (Blum-Kulka, 1986; Klaudy, 1996), leveling out (Laviosa, 1998a), and normalization (Malmkjær, 1997). However, unlike the label implies, the four TU hypotheses have been empirically supported[1] or vigorously denounced by many. Critics of TUs—namely, House (2008) and Tymoczko (1998)—questioned the possibility of universalism that is supposed to exist all genres, language pairs, translators, different contexts of translations, etc.

Nevertheless, it is still worth carrying out TU-based empirical analyses that have revealed countless findings and intriguing research questions about translation, as rightly argued by Chesterman (2010, pp. 45-46). Although TU hypotheses are not always supported universally across all genres and all language pairs, research on TU and related linguistic features can provide a solid starting point for exploring translated, non-translated, or other types of texts. This is what this article aims to accomplish. Instead of viewing TUs as universal features across all texts, we see them as a useful analysis framework for our research objective of finding out linguistic features that are most effective in distinguishing TTs from NTTs in a particular genre and language pair.

For this purpose, we compiled two comparable corpora with the introductory part of 58 sustainability reports, which have been published in both the US and South Korea. After the compilation, each text is encoded with ten TU-related factors, and the encoded data are statistically analyzed using both random forests (RFs) and conditional inference

1) See Chapter 2 of this article to find out empirical evidence that supports TU hypotheses.

trees (CITs). Our corpus design of a less-studied genre and the language pair (non-European source language) is expected to contribute to the existing TU literature.

2. Literature Review

This chapter consists of three parts. First, the notion of TUs is briefly provided with specific linguistic features used to claim TUs. The second part discusses how the linguistic features linked to TU hypotheses were observed in translations from Korean into English. Last but not least, we seek to overview what has been reported about TUs in the business reporting genre investigated in this article.

2.1. Indicators of TUs

The development of corpus and corpus linguistics in recent decades has provided several opportunities to apply corpus-based techniques to translation research, especially translational language. The notion of TU was first proposed by Baker (1993). Because of the innate processes of translation itself, it is natural that distinctive and typical linguistic attributes can be observable in TTs in comparison with NTTs. Her claim is that those attributes exist in any translated text regardless of genres, source-target language pairs, etc. (Baker, 1993, 1995, 1996), and that those features could be labeled as TUs. Driven by the mediation between the target and source languages (Baker, 1993, 1995, 1996; Laviosa, 1998a, 1998b, 2002), those universal properties are "features which typically occur in translated text rather than original utterances and which are not the result of interference from specific linguistic systems" (Baker, 1993, pp. 243-244).

During the past two decades since Baker (1993), the claim of TUs has been shared and empirically supported by many scholars who believe translated texts differ not only from their source texts but also from non-translated target texts (Malmkjær, 2012; Mauranen, 2007; McEnery & Xiao, 2007; Xiao & Dai, 2014). Although varying across researchers, their analyses have mostly centered around four TU claims proposed first by Baker (1996), each of which is closely linked to specific linguistic features.

Among others, simplification is the tendency to make target texts much simpler lexically, syntactically and stylistically using simpler translational language (Baker, 1996). The simplification process can occur either consciously or unconsciously, and the goal of

simplification is to increase the readability of target texts. Some useful factors for measuring simplification are STTR for lexical diversity, the ratio of function words over content words for lexical density, high- or low-frequency words for lexical richness, sentence length for structural sophistication, and various punctuation marks (such as semicolons or full stops over commas) for stylistic complexity (Baker, 1996; Laviosa, 1998b; Malmkjær, 1997, 2012).

Leveling out refers to the linguistic property that translated texts tend to "gravitate around the centre of any continuum" (Baker, 1996, p. 176) and that they have greater closeness to one another lexically and syntactically. Some relevant linguistic factors include type/token ratio, standard deviations of lexical variety, lexical density, readability indices, and mean sentence length (Pym, 2008).

Explicitation refers to the tendency that translated texts have explicit and concrete translational language to increase the clarity of content (Klaudy, 1996; Olohan & Baker, 2000; Xiao & Dai, 2014). It can be accomplished by making lexical, syntactic, and/or semantic additions to the target texts or making grammatical relations more explicit and cohesive. The relevant features include various connective devices such as conjunctions or complementizers.

Normalization means the tendency that untypical language use is more salient in target texts than in their source texts. The untypical language use often causes awkwardness in the translated texts. Some useful factors of normalization include overuse of clichés, idioms, prefabricated language structures, lexical bundles, and collocations (Baker, 2007; Olohan, 2004; Øveras, 1998).

Although various researchers could successfully support TU claims with empirical findings, it is true that not all findings converge in one direction. The findings often diverge and contradict each other in different texts and language pairs. This gave rise to skepticism and criticism about the universality of TU claims. House (2008, p. 6), one of the harsh critics of TUs, argued that "the quest for specific translation universals is futile" because the linguistic features suggested by TUs are just general traits of language. Tymoczko (1998) raised questions about the boundary of translations to be included in the study corpus.

Despite the criticism, TUs have been used as an analysis framework for many researchers seeking to discover distinctive features of translation. In particular, research on TUs can provide a solid starting point for exploring translated, non-translated, or other types of texts. Lou and Li (2022) tested simplification and normalization hypotheses on machine translation of social media content, and found the tendency of normalization.

This article is in sync with those attempts, trying to capitalize on TU-related findings with the aim of unveiling still uncharted aspects inherent in translation.

2.2. TUs in Korean–English Translation

As stated in the previous section, the occurrence of TUs largely depends on source text (ST) genres and language pairs. This section explores the findings from existing studies in the Korean-English language pair in particular.

Overall, translation from Korean into English has been underrepresented compared to translation in the opposite direction, presumably because English is an L2 language in South Korea's research circles.

Table 1. TUs in Korean-English Translation

Genre	Simplification	Explicitation	Normalization	Leveling Out
Academic Prose	Goh et al. (2016)	Goh et al. (2016)	Goh et al. (2016)	Goh et al. (2016)*
Newspaper	Goh & Lee (2016)*		Goh & Lee (2016)	
Literature	Choi (2016)*	Lee (2014); Choi (2016)	Choi (2016)	
Business Reporting	Yim (2019)*; Lee & Yim (2019)*		Yim (2019)	

* Note that the asterisk (*) indicates that TTs show significantly different linguistic patterns from NTTs in the direction opposite to what TU hypotheses suggest.

Table 1 shows the findings regarding TUs in Korean-English translation, which are hard to generalize due to the limited number of studies and the limited coverage of ST genres. However, it is worth mentioning that all four TU hypotheses are supported in academic prose, whereas other genres support TU hypotheses partially at best.

Regardless of genre, normalization appears to be supported widely in this language pair. This means translated English from Korean tends to contain a higher proportion of high- frequency, typical lexical bundles and collocations than their NTT counterparts. Those features are typically measured by the proportion of those lexical bundles in a given text group.

Also notable in Table 1 are the studies with the asterisk (*) sign. In Goh et al. (2016), leveling out was investigated by the standard deviation of the mean sentence length in each text. The value of TT is supposed to be lower compared to NTTs to support the TU

hypothesis, but was found to be significantly lower. The same was observed in simplification, where TTs deviated significantly from NTTs, but in the direction exactly opposite to TU hypothesis, which will be discussed in detail in 2.3 and 5. Although those studies reject TU hypotheses, the linguistic features investigated give researchers great insights into the nature of translation as well as intriguing research questions. This article also focuses on the distinctive features in TTs for assessing their ability to classify TTs from NTTs.

2.3. Business Reporting Genre and TUs

Businesses in operation generally publish several materials to communicate with various stakeholders such as their customers, partners, media, regulators, investors, creditors, etc. They document mandatory materials such as regulatory filings and also voluntary materials such as annual reports and sustainability reports for not only regulatory authorities but also other stakeholders. Sustainability reports investigated in this article fall under voluntary corporate reporting. They used to cover environmental and social values and actions under different labels such as "'sustainability reports', 'CSR reports', or 'environmental reports', etc." (Zappettini & Unerman, 2016, p. 522). However, they have become a major tool that integrates other types of reports into a "hybrid text which brings together [⋯] financial, social, environmental reports" (Zappettini & Unerman, 2016, pp. 522-523). As Integrated Reporting (IR), which literally integrates the contents of an annual report into a sustainability report, has taken hold as a norm during the recent decade, more and more companies have been following the practice (Vaz et al. 2016).

Almost always in this genre, management forewords (or CEO letters) are attached at the beginning of a report, offering an introductory remark. These letters have attracted wide academic attention. Despite abundant literature about CEO letters mostly focusing on the relations to business performance, the linguistic features of CEO letters have only become the object of research since Hyland (1998). In translation studies, the letters' potential variations in expressing corporate cultures and philosophies across cultures have caught researchers' attention (Olohan, 2009), but only limitedly within the field of commercial translation (House, 1977, 2015; Wawra, 2007; Leibbrand, 2015; Poole, 2017; Sun et al., 2018, etc.).

Sustainability reports written in Korean and translated into English are significantly underrepresented except for a few studies on TUs in sustainability reports, given that IR

is the latest reporting practice. Yim (2019) took a corpus-based approach to TUs in about sixty sustainability reports. She manually encoded thirty-eight linguistic factors and evaluated them with monofactorial statistics (independent t-tests and Mann-Whitney tests). The results showed that only eight out of thirty-eight factors were statistically significant. More specifically, TTs were significantly different from NTTs in the proportion of conjunctions (explicitation), sentence length and the frequency of function words (simplification), and the proportion of the most frequently used 5/10/20/30/50 trigrams (normalization). The intriguing part was that normalization was the only TU hypothesis supported by the results: TTs were found to include a higher percentage of typical, standardized lexical bundles compared to NTTs. By contrast to TU hypotheses, TTs included significantly fewer conjunctions, longer sentences, and fewer function words than NTTs, exactly opposite to what TUs suggest. While English management forewords are reported to contain more personal pronouns, which are functional words to foster interactions and develop relationships with prospective readers (Hyland, 1998), personal pronouns are hardly used in formal, written Korean texts (Jiang & Seo, 2023). Also, it is possible that long sentences in Korean ST led translators to split sentences more with connectives (Yim, 2019, pp. 144-145). This suggests the possibility that genre-specific features in this language pair affect TU tendencies.

Although the aforementioned linguistic features could identify the differences between NTT and TT in this genre, such confirmatory data analysis was inadequate to show the interactions between the variables. Lee & Yim (2019), on the other hand, took a multifactorial approach and analyzed the data with a logistic regression. The results show that, among the eight TU factors which were statistically significant in Yim (2019), four of them were eliminated due to the multicollinearity effects. Lee & Yim (2019) also found that the probability of a text being TT increases with higher Type and STTR, fewer function words, shorter mean sentence lengths, and higher proportions of the 20 most frequently used words, all of which are related to simplification.

What those two studies fail to show us is how effective one factor is compared to the other in distinguishing TT from NTT. This could be effectively addressed by adopting predictive modeling such as the conditional inference trees (CITs) and random forests (RFs) that have been increasingly popular in corpus linguistics. CITs and RFs are known to be effective in finding out which linguistic factors determine various variants (Levshina, 2020, p. 611). However, those methods have been barely adopted to explore the difference between TTs and NTTs except for very few: For example, Lee (2021) could accurately

classify human translations from machine translations in literary texts with RF after he extracted 100 most frequently used words from each text. In a different study, Lee (2014, p. 215) pointed out the need for a shift away from a confirmatory approach towards more a multidimensional explanatory approach in studying linguistic features of translation. A more sophisticated statistical approach, Lee (2014, p. 227) argues, offers greater potential to understand and analyze the interactions of multiple variables and factors. Nevertheless, confirmatory data analysis appears to still overwhelmingly dominate translation studies, which emphasizes the necessity of this study.

3. Research Method

3.1. Corpus Representativeness

To examine the linguistic behaviors of TUs in sustainability reports, two types of corpora were compiled from introductory letters in sustainability reports published by South Korean and US companies. As it is impossible to collect every report available, we compiled a comparative sample that has representative power in this genre. The objective was achieved by selecting the companies listed on KTOP30 and Dow Jones Industrial Average, the two benchmark stock indexes in South Korea and the US. The Dow Jones index includes a total of 30 companies that represent the whole listed firms in the US using market capitalization (Wikipedia, n.d.). The KTOP30 index is a South Korean index that includes a total of 30 South Korean companies listed both on the KOSPI and KOSDAQ. Because the KTOP30 index was constructed by benchmarking the Dow Jones index, it is perfectly safe to conclude that the sample of the companies listed on the two indices are not only comparable, but also representative.

Table 2. Summary of the Corpora

Text	TT	NTT
Number of Texts	26	32
Number of Tokens	16,433	26,635
Number of Sentences	648	1,169

After looking for the corporate websites of the sample, authors found that most of the

firms are publishing sustainability reports on an annual or biannual basis. For comparability of the corpora, TTs were collected from KTOP 30 companies, while NTTs were from the Dow Jones-listed companies. The names of companies are enumerated in Table 3, while the summary of each corpus is shown in Table 2.

Table 3. Publishers of the Reports in Corpora

Type	Company
TT	KB Financial Group, LG Display, LG Electronics, LG Chemical, NAVER, POSCO, SK Innovation, SK Telecom, SK Hynix, Lotte Chemical, Mirae Asset Daewoo, Samsung SDI, Samsung C&T, Samsung Life Insurance, Samsung Electtro-Mechanics, Samsung Electronics, Samsung Fire & Marine Insurance, Shinhan Financial Group, Amore Pacific, Yuhan, Hankook Tire, Hyundai Engineering & Construction, Hyundai Glovis, Hyundai Mobis, Hyundai Heavy Industries, Hyundae Motors
NTT	3M, American Express, Apple, Boeing, Caterpillar, Chevron, Cisco, Coca-Cola, The Walt Disney Company, DowDuPont, ExxonMobil, General Electric, Goldman Sachs, The Home Depot, IBM, Intel, Johnson & Johnson, JPMorgan Chase, McDonalds, Microsoft, Nike, Pfizer, Procter & Gamble, Travelers Companies, Inc., United Technologies, Verizon, Visa, Wal-Mart

Some US firms such as JPMorgan and DowDupont included multiple management forewords in one report, generally one by the president and the other by the chief environmental officer. Because our research purpose is to explore differences between TTs and NTTs, we included all the letters in the NTT corpus.

3.2. Encoded Factors and TU Hypotheses

After two corpora were compiled using the reports published by Korean and US companies listed above, a total of ten factors shown in Table 4 were manually encoded for each of the 58 texts.[2)3)] Note that the factors in Table 4 cover all the four categories of TUs in Table 1. To determine which TU factor is included in our analysis and which one is excluded, we reviewed previous studies in this particular genre (Yim, 2019; Lee & Yim, 2019), including but not limited to Table 1 to pick out linguistic features of TTs that are generally known to be distinctive from NTTs. It is worth mentioning that the ten factors

2) The statistical analyses were conducted with R (R Core Team, 2019).

3) The data were encoded for the previous study by one of the authors (Yim, 2019). For details on the encoding process, please refer to Yim (2019, p. 140).

listed in Table 4 were those that survived in the multicollinearity test on 26 factors in Lee & Yim (2019, pp. 88-91).

Table 4. Encoded Factors

Category	Variables	Description
Simplification	TYPE	Total Number of Types
	STTR	Standardized Type/Token Ratio
	FUNCT_TOTAL_P	Function Words (%)
	HIGH_TOP_20_P	Top 20 High-Freq. Words (%)
	BOTTOM_P	Bottom-Freq. Words (%)
	MSLENGTH	Mean Sentence Length
Leveling Out	STTR_SD	Standard Deviation of STTR
	MSL_SD	Standard Deviation of Mean Sentence Length
Explicitation	CONN _TOTAL_P	Number of Connectives (%)
Normalization	N_GRAM_TOTAL_P	Lexical Bundles: Trigrams (%)

To perfectly support all four TU hypotheses of simplification, explicitation, leveling out and normalization, TTs are supposed to have lower values in TYPE, STTR, BOTTOM_P, MSLENGTH, STTR_SD, and MSL_SD, while higher values should be observed in FUNCT_TOTAL_P, HIGH_TOP_20_P, CONN_TOTAL_P, and N_GRAM_ TOTAL_P. Our estimation was that simplification and normalization would play a certain role, based on the results of existing studies in this genre.

3.3. Statistical Analysis

In this paper, two types of statistical analyses were applied to the encoded data: Random forests (RF) and conditional inference tree (CIT). The former was used for identifying how important the role of each linguistic factor is in identifying TT vis-à-vis NTT and the latter for how each linguistic factor actually works in the classification.

Although similar to classification and regression trees (CART), RF is a considerably improved approach. It is a partitioning approach that successively splits the data into two groups, based on various independent variables such that the split maximizes the classification accuracy. The process repeats itself until no further split would increase the classification accuracy sufficiently. RF in turn adds two layers of randomness to the analysis, which help (i) recognize the impact of each variable or the combinations of variables and (ii)

protect against overfitting: On the one hand, the algorithm constructs a different tree each time, each of which is fitted to a different bootstrapped sample of the full data. On the other hand, each split in each tree structure chooses from only a randomly-chosen subset of predictors. Our overall results came from the amalgamation of all 10,000 trees that have been generated.

An analysis using CIT is a recursive partitioning approach towards classification and/or regression which tries to classify or compute predicted values on the basis of multiple binary splits of the data (Bernaisch et al., 2014, p. 14). More specifically, a set of corpus data is recursively inspected to determine which independent variable serves the best to classify the whole data into two groups for predicting the known outcomes of the dependent variable. This process of splitting the data is repeated until no further split would be possible, and the final result is a flowchart-like decision tree (Bernaisch et al., 2014, p. 14).

4. Analysis Results

4.1. Random Forests

The RF analysis result is illustrated in Figure 1, which is based on the TTs and NTTs. After RF analysis, the importance of each TU factor was calculated. Then, the influence of each factor was computed based on the maximum value. The plot was drawn with relative importance, not with the absolute values.

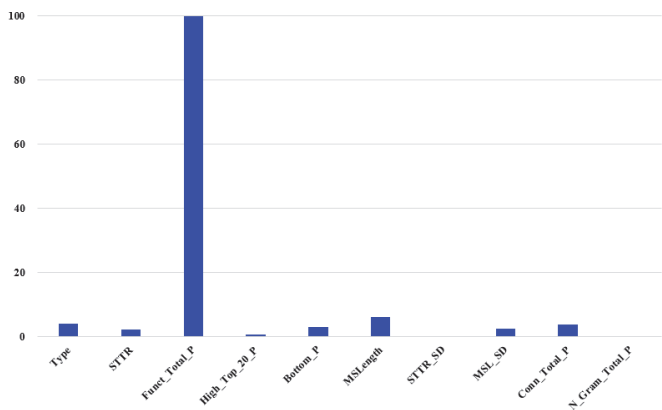

Figure 1. RF Analysis Results (Overall)

As this plot shows, the factor FUNCT_TOTAL_P has the maximum value, and the values of the other factors are less than 10.

This plot also demonstrates that not only the factors in simplification (STTR, FUNCT_TOTAL_P, HIGH_TOP_20_P, Bottom_P, and MSLENGTH) but also the factors in leveling out (MSL_SD) and explicitation (CONN_TOTAL_P) play a certain role in the identification of NTT from TT texts. The factors in normalization (N_GRAM_TOTAL_P), however, play no role in the given data set.

In order to examine how each factor plays a role in each type of text, the variable importance of each factor was closely examined in TT and NTT separately. The following plot shows the results. The values shown in Figure 2 indicate absolute importance here. As you can observe, the general patterns in TT are similar to those in NTT. It implies that each factor behaves similarly in both TT and NTT.

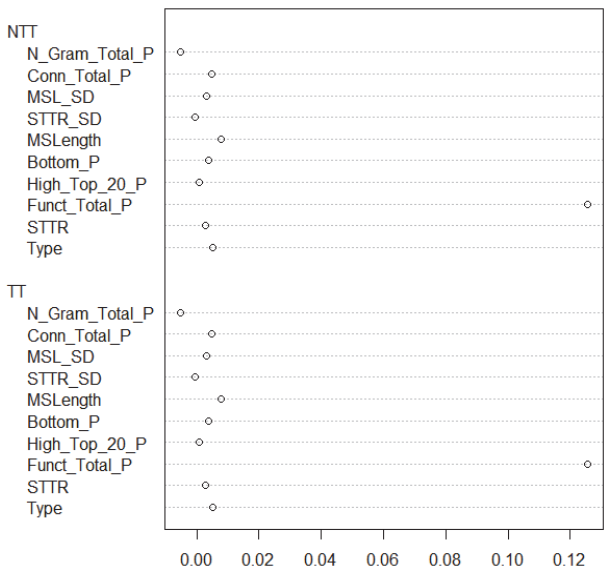

Figure 2. RF Analysis Results (TT vs. NTT)

4.2. Conditional Inference Tree

Although the analysis results in both Figure 1 and Figure 2 demonstrate how much importance each linguistic factor has, the plots do not provide information on how each linguistic factor actually behaves in the given data set. For the purpose of investigating

the (linguistic) behaviors of each factor, a CIT analysis was conducted. Figure 3 illustrates the results.

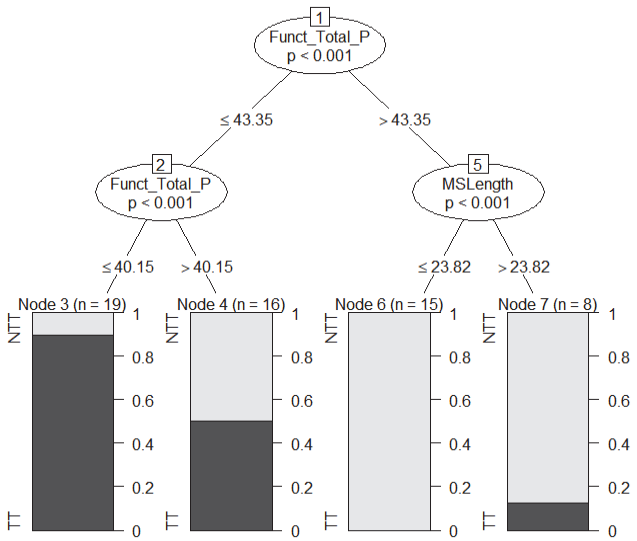

Figure 3. CIT Analysis Results

Among the ten factors in Figure 1, only two factors (FUNCT_TOTAL_P and MSLENGTH) appeared in Figure 3. It implies that the influences of these two factors overwhelm those of the other factors. It also means that these two factors are enough to identify NTTs from TTs.

Figure 3 can be interpreted as follows. At the root node, if the value of FUNCT_TOTAL_P is equal to or less than 43.35 (i.e., FUNCT_TOTAL_P£43.35%), we have to go to the left tree (node number 2). If not (i.e., 43.35%< FUNCT_TOTAL_P), on the other hand, we have to go to the right tree (node number 5). At node 2, if the value of FUNCT_TOTAL_P is equal to or less than 40.15 (i.e., FUNCT_TOTAL_P£43.35% and FUNCT_TOTAL_P£40.15%), we have to go to the left tree (node number 3). If not (i.e., FUNCT_TOTAL_P£43.35% and 40.15%<FUNCT_TOTAL_P), on the other hand, we have to go to the right tree (node number 4). In node 3, among the 19 texts, about 90% belong to NTT, and about 10% of texts are TT. In node 5, among the 16 texts, about 50% belong to NTT (8 texts), and about 50% of texts are TT (8 texts). The interpretation of node 5 is similar. If the value of MSLENGTH is equal or less than 23.82 (i.e., 43.34%<FUNCT_

TOTAL_P and MSLENGTH£23.82), we have to go to the left tree (node number 6). If it is not the case (i.e., 43.34%<FUNCT_TOTAL_P and 23.82<MSLENGTH), on the other hand, we have to go to the right tree (node number 7). In node 6, all the 15 texts belong to NTT. In the node 7, among the 8 texts, about 90% belong to NTT (7 texts), and about 10% of texts are TT (1 texts). The analysis results of RF and CIT reveal some interesting properties that cannot be observed in the monofactorial approach of Yim (2019) and the (multifactorial) logistic regression in Lee & Yim (2019).

5. Discussion

In this paper, the linguistic behaviors of TUs were analyzed with two statistical analyses, RF and CIT. The linguistic behaviors of TUs in the sustainability reports, however, seem to be different from those in other genres/registers of texts.

Usually, if the TUs are investigated in the TT vs. NTT distinctions, several TU factors play a considerable role even in the case where the multicollinearity effects are removed. In the data set for sustainability reports of this paper, however, only two of the simplification factors were found to play a crucial role in the identification of NTT from TT texts: FUNCT_TOTAL_P (the proportion of function words linked to lexical density), and MSLENGTH (the mean sentence length related to syntactical complexity). It is worth noting that lexical diversity features—STTR and HIGH_TOP_20_P—in the simplification hypothesis were als ovalid markers for translation in the previous multifactorial analysis (Lee & Yim, 2019).

The findings in this article show that lexical density (FUNCT_TOTAL_P) and syntactical complexity (MSLENGTH) play a more significant role in distinguishing TTs from NTTs comparad to lexical diversity (STTR and HIGH_TOP_20_P). Although the two features were found to be efficient markers for translation, their patterns were opposite to what TU hypotheses suggest: Function words in TTs were underrepresented compared to NTTs, while the mean sentence length in TTs was longer than that in NTTs.

These two features could possibly be attributed to the characteristics of the translated business reporting genre. First, the underrepresentation of function words in TTs could stem from the overrepresentation of relational markers as part of interpersonal metadiscourses in English management forewords (Hyland, 1998). The underrepresentation of those markers in Korean management forewords was also confirmed by Kim (2012).

The fact that relational markers tend to include quite a few functional words, namely, personal pronouns, partly explains the underrepresentation in our data. Second, lengthy sentences in the source texts of South Korean management forewords made the TT sentences lengthier than comparable NTTs.

From a comparative perspective across genres, function words supported the existing TU hypothesis only in academic prose (Goh et al., 2016). They were underrepresented in translations of business reporting in the current study, literature (Choi, 2016), and news (Goh & Lee, 2016). Regarding the mean sentence length, TTs contained shorter sentences in academic prose (Goh et al., 2016), news (Goh & Lee, 2016) as suggested by the TU hypothesis. In the genres of business reporting in this article and literature (Choi, 2016), TT sentences were longer than NTT sentences.

Then, where do the differences originate from? One possible answer is that the genre differences may result in different behaviors of TUs, in TTs and NTTs. In the literature of corpus linguistics, there have been many findings that the distributions and uses of certain linguistic factors are heavily influenced by specific registers and genres (Biber, 1988; Biber 1991; Biber & Finegan, 1994, Conrad, 1994, Reppen, 1994; Tribble, 1999). The fact that the distributions and uses of certain linguistic factors are heavily influenced by specific registers and genres was also observed in non-Western languages (Besnier, 1988; Biber & Hanrad, 1992, 1994; Kim & Biber, 1994). Especially, Biber (1988) demonstrated that the genre differences could be analyzed with Factor Analysis and that each genre could effectively be identified with the distributions and uses of certain linguistic factors.

The differences in genre/register can influence the distributions of TU factors. The analysis results in this paper clearly demonstrate that genre/register differences also have to be fully considered in the study of TUs in translation studies. In particular, it is important to note that individual linguistic features labeled under one TU hypothesis moved differently across genres. This calls for a more detailed classification of TU hypotheses.

6. Conclusion

In this paper, the behaviors of TUs were closely examined in TTs and NTTs. For this purpose, two small but representative and comparable corpora were compiled with management forewords of English sustainability reports translated from Korean and

originally written in English. The texts were chosen from the KTOP30 and Dow Jones indices, respectively. Ten linguistic factors were manually encoded and they were analyzed with RF and CIT.

The analysis results showed that only two TU factors (FUNCT_TOTAL_P and MSLENGTH) play a crucial role in the identification of NTTs from TTs, both of which are factors supporting the hypothesis of simplification. These results contrast starkly with those of previous TU studies, where most TU factors play a certain role in the identification of TTs from NTTs. However, our findings might be insufficient to make a strong case for supporting simplification against the other three hypotheses. The variables tested in this article are admittedly tilted towards simplification, which has been investigated most extensively in TU literature and been empirically supported by many studies in the Korean-English language pair including the authors' previous article (Lee & Yim, 2019). Rather, our findings could pinpoint which simplification features count more than others.

The results in this paper have two implications. First, the stark difference in the results from previous studies strongly reaffirms the importance of genre/register in the investigation of TU factors and, more broadly, the linguistic features of translation itself. Second, the fact that different statistical approaches lead to different findings necessitates a more thorough approach in our exploration of translation. It is noteworthy that our RF and CIT approach could pinpoint two simplification factors (the proportion of function words and the mean sentence length) out of four simplification factors that seemed valid in Lee & Yim (2019). This certainly suggests the approach's revealing potential in our quest of how translation differs from non-translation.

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). Amsterdam: John Benjamins.

Baker, M. (1995). Corpora in translation studies: An overview and some suggestions for future research. *Target*, *7*(2), 223-243.

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

Baker, M. (2007). Patterns of idiomaticity in translated vs. non-translated text. *Belgian Journal of Linguistics*, *21*(1), 11-21.

Bernaisch, T., Gries, S. T., & Mukherjee, J. (2014). The dative alternation in South Asian English(es): Modelling predictors and predicting prototypes. *English World-Wide*, *35*(1), 7-31.

Besnier, N. (1988). The linguistic relationships of spoken and written Nukulaelae registers. *Language*, *64*, 707-736.

Biber, D., & Finegan, E. (1994). *Sociolinguistic perspectives on register*. New York: Oxford University Press.

Biber, D., & Hanrad, M. (1992). Dimensions of register variation in Somali. *Language Variation and Change*, *4*, 41-75.

Biber, D., & Hanrad, M. (1994). Linguistic correlates ofthe transition to literary in Somali: Language adaptation in six press register. In D. Biber & E. Finegan (Eds.), *Sociolinguistic perspectives on register* (pp. 182-216). New York: Oxford University Press.

Biber, D. (1988). *Variation across speech and writing*. Cambridge, MA: Cambridge University Press.

Biber, D. (1991). Oral and literate characteristics of selected primary school reading materials. *Text*, *11*, 73-96.

Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In L. Venuti (Ed.), *The translation studies reader* (pp. 298-313). Routledge.

Blum-Kulka, S., & Levenston, E. A. (1978). Universals of lexical simplification. *Language Learning*, *28*(2), 399–415.

Chesterman, A. (2010). Why study translation universals? *Acta Translatologica Helsingiensia 1*, 38-48.

Choi, H.-K. (2016). Hanyeongmunhak beonyeok munche yeongu (A study of the stylistics of translated Korean literature: A corpus-based analysis). *Beonyeokagyeongu*, *17*(3), 193–216.

Conrad, S. (1994). Variation in academic writing: Textbook and research articles across disciplines. In *Proceedings of the Annual Conference of the American Association of Applied Linguistics*, Baltimore, Maryland.

Frawley, W. (1984). Prolegomenon to a theory of translation. In W. Frawley (Ed.), *Translation: Literary, linguistic and philosophical perspectives* (pp. 159–175). Associated University Press.

Goh, G.-Y., Kim, D., & Lee, Y. C. (2016). A corpus-based study of translation universals in thesis/dissertation abstracts. *Korean Journal of English Language and Linguistics*, *16*(4), 819–849.

Goh, G.-Y., & Lee, Y. (2016). Hanguk sinmunui yeongeo beonyeoge natanan beonyeok bopyeonsoui kopeoseu giban bunseok (A corpus-based study of translation universals in English translations of Korean newspaper texts). *Bigyomunhwayeongu*, *45*, 109–143.

House, J. (1977). A model for assessing translation quality. *Meta*, *22*(2), 103–109.

House, J. (2008). Beyond intervention: Universals in translation? *Trans-kom*, *1*(1), 6–19.

House, J. (2015). *Translation quality assessment: Past and present*. Routledge.

Hyland, K. (1998). Exploring corporate rhetoric: Metadiscourse in the CEO's letter. *The Journal of Business Communication*, *35*(2), 224–244.

Jiang, L., & Seo, S.-K. (2023). Hangugeoui tekseuteu yuhyeonggwa bunpo yangsang (A study on text types and distribution patterns in Korean). *Language Facts and Perspectives*, *58*, 159–181.

Kim, H. (2012) Beonyeok mit bibeonyeoke guchukdoen jeojawa dokjaui sanghojagyong: tekseuteujeok metadamhwa bunseogeul jungsimeuro (Reader-writer interaction in translated and non-translated Letters to Shareholders with an analysis of textual metadiscourse) *Tongbeonyeokagyeongu*, *16*(2), 115–137.

Kim, Y.-J., & Biber, D. (1994). A corpus-based analysis of register variation in Korean. Sociolinguistic perspectives on register, 1994, 157-81.

Klaudy, K. (1996). Back-translation as a tool for detecting explicitation strategies in translation. In K. Klaudy, J. Lambert & S. Anikó (Eds.), *Translation studies in Hungary* (pp. 99-114). Budapest: Scholastica.

Laviosa, S. (1998a). The corpus-based approach: A new paradigm in translationstudies. *Meta*, *43*(4), 474-479.

Laviosa, S. (1998b). 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.

Lee, C.-S. (2014). Cachawontonggyebunseokbeobeul hwaryonghan beonyeokbopyeonso saryeyeongu (Multidimensional explanatory analysis of translation universals). *Beonyeokagyeongu*, *15*(3), 211–232.

Lee., C.-S. (2021). Gigyehakseup algorijeumeul hwaryonghan munhakbeonyeogeseoui gigyebeonyeokgwa ingan beonyeok gyeolgwamul bullyu yeongu (Machine learning

classification of literary translation samples by human and machine translators). *Beonyeokagyeongu*, *22*(1), 199–217.

Lee, Y.-H., & Yim, J. (2019). A multifactorial analysis of translation universals in management forewords of sustainability reports. *English Language and Linguistics*, *25*(3), 79-105.

Leibbrand, M. P. (2015). The language of executive financial discourse. *Studies in Communication Sciences*, *15*(1), 45–52.

Levshina, N. (2020). Conditional Inference Trees and Random Forests. In M. Paquot & S. T. Gries (Eds.), *A practical handbook of corpus linguistics* (pp.611–643). Springer International Publishing.

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.

Malmkjær, K. (1997). Punctuation in Hans Christian Anderson's stories and in their translation into English. In F. Poyatos (Ed.), *Nonverbal communication and translation: New perspectives and challenges in literature* (pp. 151–162). Amsterdam & Philadelphia: John Benjamins.

Malmkjær, K. (2012). Language philosophy and translation. In Y. Gambier & L. Doorslaer (Eds.), *Handbook of translation studies* (Vol. 3)(pp. 89-94). Amsterdam: John Benjamins.

Mauranen, A. (2007). Universal tendencies in translation. In M. Rogers & G. Anderman (Eds.), *Incorporating corpora: The linguist and the translator* (pp.32-48). Clevedon: Multilingual Matters.

McEnery, T., & Xiao, R. (2007). Parallel and comparable corpora: What is happening?. In G. Anderman & M. Rogers (Eds.) *Incorporating corpora: The linguist and the translator (pp.* 18-31). Clevedon: Multilingual Matters.

Olohan, M. (2004). *Introducing corpora intranslation studies*. London: Routledge.

Olohan, M. (2009). Commercial translation. In M. Baker & G. Saldanha (Eds), *Routledge encyclopedia of translation studies* (pp. 40–43). London & New York: Routledge.

Olohan, M. & Baker, M. (2000). Reporting 'that' in translated English: Evidence of or dubliminal processes of explicitation. *Across Languages and Cultures 1*(2), 141-158.

Øverås, L. (1998). In search of the third code: An investigation of norms in literary translation. *Meta*, *43*(4), 557-570.

Poole, R. (2017). "New opportunities" and "Strong performance": Evaluative adjectives in letters to shareholders and potential for pedagogically-downsized specialized corpora. *English for Specific Purposes*, *47*, 40–51.

Pym, A. (2008). On Toury's laws of how translators translate. In A. Pym, M. Shlesinger, & D. Simeoni (Eds.), *Beyond descriptive translation studies: Investigations in homage to Gideon Toury* (pp. 311-328). Amsterdam: John Benjamins.

R Core Team. (2019). R: A language and environment for statistical computing. R Foundation for Statistical Computing, Vienna.

Reppen, R. (1994). V*ariation in elementary school writing*. Unpublished doctoral dissertation. Northern Arizona University.

Selinker, L. (1972). Interlanguage. *International Review of Applied Linguistics*, *10*(3), 209-231.

Sun, Y., Jin, G., Yang, Y., & Zhao, J. (2018). Metaphor Use in Chinese and American CSR Reports. *IEEE Transactions on Professional Communication*, *61*(3), 295–310.

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

Tribble, C. (1999). *Writing Difficult Texts*. Unpublished doctoral dissertation. Lancaster University.

Tymoczko, M. (1998). Computerized corpora andthe future of translation studies. *Meta*, *43*(4), 652-660.

Wawra, D. (2007). On course for the nextstage of success: The annual report of US and Japanese companies. In C. Ilie (Ed), *The use ofEnglish in institutional and business settings: An intercultural perspective* (pp. 127–146). Peter Lang Bern.

Wikipedia. (n.d.) DowJones Industrial Average. Retrieved April 30, 2024, from https://en.wikipedia.org/wiki/Dow_Jones_Industrial_Average.

Xiao, R., & Dai, G. (2014). Lexical and grammatical properties of translational Chinese: Translation universal hypotheses reevaluated from the Chinese perspective. C*orpus Linguistics and Linguistic Theory*, *10*, 11-55.

Yim, J. (2019). Jisokganeungseong bogoseo hanyeongbeonyeogui beonyeokbopyeonso yangsanggochal (Translation universals in translated CEO letters in sustainability reports). *Beonyeokagyeongu*, *20*(5), 131–162.

Zappettini, F., & Unerman, J. (2016). 'Mixing' and 'Bending': The recontextualisation of discourses of sustainability in integrated reporting. *Discourse & Communication*, *10*(5), 521–542.

Jin Yim
Adjunct Lecturer
Department of Interpretation and Translation
Graduate School of Translation and Interpretation, Ewha Womans University
52, Ewhayeodae-gil, Seodaemun-gu, Seoul 03760, Republic of Korea
Phone: +82-2-3277-3662
E-mail: jy2812@gmail.com

Yong-hun Lee
Research Professor
Department of Linguistics
Chungnam National University
99 Daehak-ro, Yuseong-gu, Daejeon 34134, Republic of Korea
Phone: +82-42-821-5318
E-mail: yleeuiuc@cnu.ac.kr

Received on June 1, 2024
Revised version received on June 18, 2024
Accepted on June 30, 2024