English Teaching, Vol. 78, No. 2, Summer 2023, pp. 55-81
DOI: https://doi.org/10.15858/engtea.78.2.202206.55
http://journal.kate.or.kr

Feature Importance Ranking of Translationese Markers in L2 Writing: A Corpus-Based Statistical Analysis Across Disciplines

Younghee Cheri Lee* and Soomin Jwa**

Lee, Younghee Cheri., & Jwa, Soomin. (2023). Feature importance ranking of translationese markers in L2 writing: A corpus-based statistical analysis across disciplines. *English Teaching, 78*(2), 55-81.

In recent years, an array of studies has focused on 'translationese' (i.e., unique features that manifest in translated texts, causing second language (L2) writings to be similar to translated texts but different from native language (L1) writings). This intriguing linguistic pattern has motivated scholars to investigate potential markers for predicting the divergence of L1 and L2 texts. This study builds on this work, evaluating the feature importance ranking of specific translationese markers, including standardized type-token ratio (STTR), mean sentence length, bottom-frequency words, connectives, and n-grams. A random forest model was used to compare these markers in L1 and L2 academic journal article abstracts, providing a robust quantitative analysis. We further examined the consistency of these markers across different academic disciplines. Our results indicate that bottom-frequency words are the most reliable markers across disciplines, whereas connectives show the least consistency. Interestingly, we identified three-word lexical bundles as discipline-specific markers. These findings present several implications and open new avenues for future research into translationese in L2 writing.

Key words: L2 Writing, Translationese, L1-L2 Translation, Translation Universals, Nativeness, Feature importance ranking

*First author: Younghee Cheri Lee, Teaching Professor, Dasan University College, Ajou University
**Corresponding author: Soomin Jwa, Assistant Professor, Department of English Education, Kongju National University, 56, Gongjudaehak-ro, Gongju-si, Chungcheongnam-do, Korea 32588; E-mail: smjwa@kongju.ac.kr
Received 31 March 2023; Reviewed 17 April 2023; Accepted 30 May 2023

1. INTRODUCTION

Over the last few decades, the use of L1 in L2 writing has been noted and considered in the study design of L2 writing research (see Göpeferich, 2017). A traditional view of L1 use in Second Language Acquisition (SLA) posits that reliance on L1 is a strategy that L2 writers use to alleviate their cognitive burden, thus, it is often favored by low language proficiency L2 writers. Despite the negative sentiments of some modern language practitioners, however, empirical investigations have revealed that the use of L1 (more specifically, translating from L1[1]) is natural and readily observable in the L2 composing process (Cohen & Brooks-Carson, 2001; Kobayashi & Rinnert, 1992; Liu, 2009; Roca de Larios, Murphy, & Manchón, 1999; Sasaki, 2002; Wang & Wen, 2002). Even when there is a discrepancy between L2 students' conceptual capacities and their language competencies, translating from L1 into L2 comprises a "knowledge-constituting" function (Galbraith, 1999, p. 141). L2 writers who use L1 as a supposed "crutch" (Kern, 1994), then, can also be creative without letting themselves be confined to the thoughts of their accessible linguistic repertoires.

It has been argued that the L2 writing translation process textually manifests in the form of translationese, a set of "fingerprints" that a source language leaves on a translated work (Gellerstam, 1986). Several scholars have suggested that particular linguistic features in L2 writers' texts bear a striking resemblance to those found in translated texts; they have also suggested that particular linguistic features in both L2 writers' texts and translated texts are markedly different from those in native L1 texts (see Ivaska & Bernardini, 2020; Rabinovich, Nisioi, Ordan, & Wintner, 2016). More specifically, they claimed that the difference between those L1 and L2 arises from translation activity. Based on the prevailing consensus that L2 writing is an interplay of L1 and L2 linguistic systems, SLA researchers have also shown interest in which linguistic features create a sense of non-nativeness in L2 writing, and concerted efforts have been made to generalize linguistic features identifiable in L2 writing (Crossley & McNamara, 2011; Hinkel, 2001). Even so, the measures selected for past studies were not designed to find the traces of translation from L1 to L2; rather, they were designed to diagnose a learner's progress by putting L2 writers of different proficiencies in juxtaposition.

Some applied linguistics researchers have recently attempted to ascertain signs of translationese in non-native L2 texts based on Baker's (1993, 1996) work on universal translation features. Several corpus-driven quantitative studies have found that there are, indeed, several linguistic elements that mark L2 texts as simplified, normalized, and

[1] The notion of translation is not limited to converting a textual representation in L1 into an equivalent textual representation in L2; it also includes the intervention of L1 in a series of cognitive operations involved in L2 writing such as idea generation, development, and reformulation, etc. (see also Chamot, 1987; Liu, 2009).

explicitated (see Rabinovich et al., 2016). These corpora, however, examined texts written by highly advanced L2 writers; any notable textual features could not necessarily be attributed to the participants' L2 proficiency. The few translationese markers that appeared consistently across the L2 corpora were related to lexical variety, sentence length, lexical bundles, and connectives. With this in mind, we designed the present study to build upon these prior findings, emphasizing the "feature importance ranking" of translationese markers in L2 writing as opposed to L1 writing. Here, "feature importance ranking" refers to the degree to which each marker contributes to the distinction between L1 and L2 writing, with some markers having a greater impact than others. This term clearly indicates that we not only aimed to measure the importance of each feature, but we also attempted to compare them to each other and rank them in order of importance. Utilizing a random forest analysis, we sought to determine the importance ranking of translationese markers, thereby identifying the markers contributing the most (and the least) to differences between L1 and L2 writing. For our research objectives, we compared native L1 and non-native L2 journal article abstracts in the disciplines of linguistics and literature. We also investigated whether the markers' feature importance ranking remained consistent within each discipline.

2. LITERATURE REVIEW

2.1. Translation as a Strategy in the L2 Writing Process

It is widely believed that with two languages at their disposal, the L2 writer's writing process becomes a "bilingual event" (Wang & Wen, 2002, p. 239). In such a process, the writer's L1 is consciously or subconsciously acknowledged and acted upon whenever needed; the role of the writer's L1 becomes apparent when the writer faces writing challenges, especially in light of insufficient linguistic resources. L2 writers' reliance on their L1 has been empirically explored through both self-reports (e.g., Cohen & Brooks-Carson, 2001; Kobayashi & Rinnert, 1992; Uzawa & Cumming, 1989) and think-aloud data (Cumming, 1989; Roca de Larios et al., 1999; Wang & Wen, 2002). Although the ways in which and the degree to which L2 writers use their L1 vary depending on a variety of factors (e.g., level of L2 proficiency, writing expertise in general, task types, etc.), it seems natural and—even inevitable—that L2 writers will make use of their L1 throughout their entire writing process (Kobayashi & Rinnert, 1992; Qi, 1998; Roca de Larios et al., 1999; Wang & Wen, 2002; Whalen & Ménard, 1995; Woodall, 2002).

In Wang and Wen's (2002) cross-sectional study, the researchers asked Chinese L2 writers, both lower-level and higher-level, to think aloud while composing. When calculating the ratio of L1 words uttered during each of the five identified sub-processes (i.e.,

task-examining, idea-generating, idea-organizing, text-generating, and process-controlling), researchers found that all participating groups used L1 least often when engaging in text-generating activities (13.5%; e.g., process-controlling [81.5%], idea-organizing [70%], idea-generating [61.5%], and task-examining [21%]). The analysis of L1 use during text production across the groups showed that the amount of L1 use dramatically declined for proficient L2 writers (to 3%): "less proficient writers construct sentences through L1-to-L2 translation, while proficient writers generate text directly in L2" (p. 240). On the other hand, not unlike novices who translated word-for-word from L1, advanced L2 writers in Sasaki's (2002) study enacted L1 to L2 translating strategies quite often; the only difference identified between the proficiency groups was the writers' pause frequency during translation. While novice L2 writers stopped writing to translate the ideas they developed in their L1, advanced L2 writers wrote more fluently with little pausing while translating (see also Sasaki & Hirose, 1996). Interestingly, Wang (2003) reported more language switching between L1 and L2 among high-proficiency L2 writers than among their lower-proficiency counterparts.

Liu (2009) extended such studies by detecting different functions of translation in the L2 composing process; she examined the think-aloud data of three high-proficiency and three low-proficiency Chinese L2 learners of English. Liu found that the number of translating performances was inversely proportional to L2 proficiency level, confirming similar results in prior studies (Beare & Bourdages, 2007; Sasaki, 2002; Sasaki & Hirose, 1996; Wang & Wen, 2002; Woodall, 2002). Her close analysis indicated that high-proficiency L2 writers' translation use was mainly concentrated at the semantic level—such as word/concept retrieval—whereas low-proficiency L2 writers translated significantly more at the syntactic level (Lu, 2011). Indeed, as shown in prior studies, advanced L2 writers often make reference to their L1 primarily for word retrieval, word choice, and coherence of their writing (Wang & Wen, 2002; Whalen & Ménard, 1995) or for idea generation (Beare & Bourdages, 2007; Uzawa & Cumming, 1989). For novice L2 writers, translation is often a compensatory strategy for encounters with lexical and morphosyntactic challenges (Cohen & Brooks-Carson, 2001; Pietila, 2015; Roca de Larios et al., 1999). Of course, such translating behaviors are not perpetually distinct for each proficiency benchmark. Liu (2009) suggested that advanced L2 writers may revert to novice L2 writers' translation strategies when confronted with writing challenges; additionally, novice L2 writers may use translation strategies at the semantic level for their composition needs. Depending on L2 proficiency, then, the degree of L1 involvement may vary; however, L2 writing processes—and, thus, L2 written products as well—are both inseparable from and intertwined with L1. As Kern (1994) argued, the process of "mental translation" is "inevitable" for L2 users (p. 442).

Overall, previous findings warrant the assumption that advanced L2 writers are likely to directly compose their intended meaning in their L2—meaning that has been formulated with access to both L1 and L2 repertories. Given this, our goal in the present study is to

evaluate linguistic predictors of translation in L2 texts produced by advanced L2 writers. We are particularly interested in highly advanced L2 writers' well-versed with academic written registers, such as L2 scholars; past studies, rather, have often investigated learner groups with L2 proficiency levels that are relatively higher than other learner groups. This article pursues determining if L2 texts produced by highly advanced L2 writers are products driven by thinking/composing in the L2 devoid of L1 influence.

2.2. Translationese as a Sign of L1-L2 Text Discrepancies

With the fundamental differences between L1 and L2 writing processes in mind, several SLA researchers have examined the links between L1 and L2 texts to identify L2-specific syntactic, lexical, and rhetorical features. Such research endeavors have focused either on textual attributes transferred from L1 (Jarvis & Pavlenko, 2008; Ringbom, 2007) or on general features characterizing texts written in any L2 (Crossley & McNamara, 2011, 2014; Hinkel, 2001). Researchers believe writers' L1 populates their L2 texts with "unnatural" linguistic attributes differing from those of native-like products. In the field of translation studies, such linguistic awkwardness is captured in the notion of *translationese*[2]. According to Gellerstam (1986), translationese is characterized by systematic influences of the source language (L1, in this study) exerted upon target language use. Gellerstam (1986) defined translationese as a set of "fingerprints" that a source language leaves on a translated text, especially during translation. The language aspects of translationese are theoretically arranged around the concept of *translation universals* (TU). Baker (1993), a foundational scholar for translation universals, suggested that some universal linguistic features arise as by-products of translation's mediating process between source and target languages—regardless of language pairs. These features, which are typical of variant translated texts, are far removed from what characterizes not only source texts but also comparable texts in the target language (Mauranen, 2007; Munday, 2016). Translation universals have, thus, been studied by comparing the corpora of translated texts with those of non-translated ones—when both corpora satisfy the same domain, genre, and time for corpus construction.

Notably, in the past two decades, corpus-based textual analyses have made great strides in identifying factors and indices that represent translational manifestations (Laviosa, 2002). The following three translational language universals have been used in empirical examinations: simplification, explicitation, and normalization. The simplification of

[2] In the Korean language, the word translationese is closely translated as 번역투 [beonyeog-tu] in which the expletive "tu" is attached to the predicate "beonyeog" meaning "translate." The expletive "tu" carries a negative connotation in most Korean contexts; thus, if one's L2 writing is said to sound like beonyeog-tu, this gives the impression that his or her writing is somewhat inferior to that of native writers.

translationese refers to the tendency to simplify language use lexically, syntactically, and stylistically by recycling high-frequency words, shortening sentences, and using less diverse vocabulary (Baker, 1993, 1996). Given these linguistic tendencies, translated texts tend to exhibit lower lexical variety, sophistication, richness, and density than non-translated texts (Al-Shabab, 1996; Baker, 1996; Laviosa, 1998). Some potential indicators that detect simplified L2 texts involve lower values of the Standardized Type/Token Ratio (STTR), a higher portion of top-frequency words, and a shorter mean sentence length (e.g., Baker, 1996; Laviosa, 1998, 2002). Among the other translational language universals, the explicitation of translationese is the most researched and least contentious. A large number of previous studies have outlined translators' behavioral patterns in making lexical, syntactic, and semantic relations more explicit and cohesive in translated texts—using extra discourse markers like connectives and conjugating words instead of leaving them ambiguous—all of which are meant to promote the clarity of conveyed information (Baker, 1996, 2007; Blum-Kulka, 1986; Hinkel, 2001; Øverås, 1998). Lastly, normalization is built on the assumption that formulaic language is more prominent in translated texts than in non-translated ones. According to Baker (1996), normalized texts tend to adhere to prevalent patterns and behaviors in the target language—even to the point of distortion. Normalized features of translationese can be detected by the overuse of prefabricated phrases, clichés, idioms, lexical bundles, collocations, common target language grammatical structures, and the frequent occurrence of typical generic features (Baker, 2007; Olohan, 2004; Øverås, 1998).

Over the last several decades, translation universals researchers have shifted their focus toward the proposition that the typical linguistic attributes initially assumed to be unique to translated texts may also arise in other forms of utterances, such as "constrained communication" (Chesterman, 2004) and "bilingual processing" (Granger, 2015; Halverson, 2003). As Ivaska and Bernardini (2020) claimed, SLA and TU researchers share similar views—especially on simplification and explicitation—in exploring language use in which multiple language systems operate concurrently. Though these two disciplines have operated independently in the past, renewed attention to similarities has contributed to the evolution of a new research area known as "constrained language use" (Kolehmainen, Meriläinen, & Riionheimo, 2014; Kruger & van Rooy, 2016; Lanstyák & Heltai, 2012). Researchers have argued that if translated texts are, indeed, associated with L2 utterances as a result of bilingual processing, identical linguistic attributes would be noticeable in L2 writers' texts as in translated ones when compared to native L1 texts (Ivaska & Bernardini, 2020; Kruger & van Rooy, 2016).

Additional empirical research efforts have been made to discern where there are linguistic discrepancies or shared linguistic attributes between native L1 and non-native L2 (constrained language) texts. Gaspari (2013), for example, posited "phraseological disparities" given the portion of four-word lexical bundles was much greater in non-native

news (the ANSA and Adnkronos online news reports) than in native news corpora (Reuters and United Press International), suggesting that non-native texts are much more formulaic than their counterparts. Further, Gaspari and Bernardini (2008) found potential shared properties between two variants of a constrained language, such as non-native and translated corpora vis-à-vis a native corpus; they observed that one of the English connectives, "therefore," which signifies coherent and argumentative connections in written academic discourse, was more prominent in constrained communication. Similarly, Koppel and Ordan (2011) suggested that non-native texts have far more cohesive markers than native texts but fewer markers than translated texts. Grabowski (2012) also investigated lexical variety among bottom-frequency words in contemporary translational and non-native literary Polish and confirmed that a constrained language has lower lexical diversity than its original, specifying one of the key linguistic features in translation universals. Finally, in research comparing two constrained languages (highly advanced non-native language and translated language) alongside native language, Rabinovich et al. (2016) observed shared traits of constrained communication: translated corpora are more equivalent to those of advanced non-native texts than to those of native texts. In this research, they made use of linguistic factors like lexical richness (e.g., TTR), collocations, cohesive markers, mean word rank, and personal pronouns.

Though prior research has not concurrently considered all of the conventional translationese features of simplification, explicitation, and normalization in a single research study, several studies attested to the co-existence of these three features by employing standard TU variables in a series of research studies contrasting native writers' L1 texts and non-native writers' L2 texts (see Goh & Lee, 2016; Lee, 2014, 2018, 2019, 2021). For example, in the studies by Lee (2014) and Goh and Lee (2016) with comparable corpora of English newspaper texts and indicators such as STTR, bottom-frequency words, top-frequency words, mean sentence length, lexical bundles, and N-grams, conventional TU results were affirmed in that non-native texts were more simplified, explicitated, and normalized than native texts. Similar findings were corroborated with different corpora—such as dissertation/thesis abstracts by Goh and Lee (2016) and scholarly journal articles by Lee (2018, 2019, 2021)—using similar indicators: STTR, bottom-frequency words, top-frequency words, connectives, standard deviations of mean sentence length, and N-grams.

As previously discussed, research into L2 texts has sought to assess whether tested variables predict L1 and L2 text disparities. In order to augment earlier findings and report expanded results, the present study aims to identify the feature importance ranking of translationese variables by comparing L1 and L2 academic journal abstracts corpora. Further, we intend to determine if the feature importance ranking of translationese variables remains consistent across disciplines. Our motivations, thus, lead to the following research questions:

1) To what extent does each translationese marker contribute to the distinction between L2 and L1 writing in terms of feature importance ranking?
2) Is the feature importance ranking of the markers consistent across disciplines?

3. METHODS

3.1. Corpus Data

In line with previous corpus-driven research adopting quantitative analysis methods, we compiled large-scale, natural language corpora to secure both data validity and reliability. We extracted the primary corpus resources from the purpose-end Comparable Corpora of English Research Abstracts of Scholarly Journal Articles (CCERA), which had been compiled with random sampling and updated for several previous research projects (Lee, 2018, 2019). As monolingual corpora, the CCERA comprises three variants of 2,243 English abstracts containing 638,764 tokens generated by native scholars' L1 English, highly expert non-native scholars' L2 English, and less expert non-native scholars' L2 English from two academic disciplines: linguistics and literature. The present study used CCERA's first two sub-corpora, which include 2,175 abstracts with 424,350 tokens: native scholars' L1 English (henceforth, L1) and non-native scholars' L2 English (henceforth, L2), respectively, in each discipline. Table 1 shows the textual statistics of the data adopted for the present study.[3]

TABLE 1
Corpus Scale and Textual Statistics

Domain	Text Type	Sub-corpora	Abstract (#)	Token (#)	Type (#)	TextLength Ave.
LINGUISTICS	Native Writers' L1 English	L1_LING	600	105,535	7,594	176
	Non-native Writers' L2 English	L2_LING	605	106,195	6,139	176
	Sub Total	•	1,205	211,730	13,733	•
LITERATURE	Native Writers' L1 English	L1_LIT	530	106,851	9,743	202
	Non-native Writers' L2 English	L2_LIT	435	107,869	8,538	245
	Sub Total	•	970	212,620	18,281	•
	Total		2,175	424,350	32,014	

Encoded variable selection was based on theoretical considerations and previous corpus-driven research findings on L2 writers' language use compared to L1 writers' language use.

[3] Corpus data sources and a more detailed description of the CCERA's construction process can be found in Lee's (2018) prior research projects.

We tested translationese simplification using three variables: STTR (e.g., Baker, 1996; Laviosa, 2002; Rabinovich et al., 2016), mean sentence length (e.g., Laviosa, 1998; Rabinovich et al., 2016), and bottom-frequency words (e.g., Grabowski, 2012). Measuring explicitation involved using connectives (e.g., Gaspari & Bernardini, 2008; Koppel & Ordan, 2011). In addition, examining normalization meant starting with the factor of lexical bundles in reference to the findings in Gaspari (2013), where four-word lexical bundles were computed; however, in the present study, we computed three-word lexical bundles according to Conrad and Biber (2004), who argued that three-word bundles occur more commonly than four-word bundles in academic discourses: three-word bundles appeared "60,000 times per million words and four-word bundles over 5,000 times per million words" (p. 61). We then encoded the five translationese variables with high significance to evaluate variable importance. Table 2 describes the five encoded variables and provides the baseline speculation suggested in translation studies.

TABLE 2

Encoded Variables

TU Feature	Variable	Description	TU Speculation
Simplification	STTR	Standardized Type/Token Ratio (%)	L1 > L2
	MSL	Mean Sentence Length (in words)	L1 > L2
	BOTTOM_P	Bottom-frequency Words (one-time occurrence) (%)	L1 > L2
Explicitation	CONN_ALL_P	Connectives: Cohesive Devices (%)	L1 < L2
Normalization	N_GRAM_TOP50_P	Top 50 Lexical Bundles: Trigrams (%)	L1 < L2

STTR: The Standardized Type/Token Ratio (STTR) of both the L2_LING and L2_LIT sub-corpora is lower than that of the L1_LING and L1_LIT corpora, so that non-native writers' L2 English shows lexical simplification.

MSL: The mean sentence length of both the L2_LING and L2_LIT sub-corpora is shorter than those of the L1_LING and L1_LIT sub-corpora, so non-native writers' L2 English typifies syntactic simplification.

BOTTOM_P: Both the L2_LING and L2_LIT sub-corpora exhibit fewer bottom-frequency words than the L1_LING and L1_LIT sub-corpora, so non-native writers' L2 English shows lexical simplification.

CONN_ALL_P: The ratio of connectives is higher in the L2_LING and L2_LIT sub-corpora than in the L1_LING and L1_LIT sub-corpora, which confirms non-native writers' syntactic explicitation.

N_GRAM_TOP50_P: The L2_LING and L2_LIT sub-corpora exhibit a greater portion of the top 50 trigram lexical bundles than the L1_LING and L1_LIT sub-corpora, which marks the lexico-grammatical normalization of non-native writers' L2 English.

3.2. Statistical Analyses

We conducted three-fold statistical analyses in the present study. We performed the first statistical analysis using descriptive statistics to observe interaction effects by computing each variable's overall mean values and confidence intervals (CIs). During the second stage, we used inferential statistics with a linear mixed-effects model approach. We applied a linear mixed-effects model to confirm whether the selected translationese variables played vital roles in identifying two domain variants (Linguistics versus Literature) and two text writer group variants (L1 writers versus L2 writers). We treated the different groups (L1 vs. L2) and different domains (LING vs. LIT) as fixed effects, research abstracts as random effects, and the five variables (STTR, mean sentence length, bottom-frequency words, connectives, and top-50 N-gram) as dependent variables.

Although the linear mixed-effects model demonstrated that L1 texts are different from L2 texts based on several linguistic features, this analysis did not examine the importance level of each linguistic feature in each dataset. Given that the model was limited to identifying valid translationese markers to classify different writer groups and domains, we performed an advanced statistical analysis in the present study's final stage to further determine the feature importance ranking of each variable in representing the difference between L1 and L2 writers' corpora. Therefore, we conducted a random forest analysis to identify the translationese markers' variable importance. A random forest model is used to enhance interpretability and understand which features drive predictions. Like a decision tree classifier, this model distinguishes the most important factors contributing to differences between L1 and L2 texts across different academic disciplines; thus, the present study signifies the feature importance ranking of the translationese markers discernable in L2 writing. The higher the factor is placed in the forest sets, the greater its importance is to given activities. A factor of greater importance in the present study means that textual manifestations related to the factor are apparent in L2 texts, clearly distinguishing L2 texts from L1 counterparts. We adopted a significance measure called Mean Decrease Accuracy (MDA) to rank variable importance. It stems from the idea that if a ranked variable is unimportant, then rearranging its values should not degrade prediction accuracy (Breiman, 2001). Logically, the MDA must determine how much accuracy decreases when training the model by eliminating each variable.

We facilitated text processing and baseline computational analyses using WordSmith Tools 7.0 (Scott, 2019) and AntConc 3.4.4w (Anthony, 2020). We conducted the statistical analyses, including the linear mixed-effects model and random forest analysis, using a lmerTest package (Kuznetsova, Christensen, & Brockhoff, 2013) and a RandomForest package (Liaw & Wiener, 2002) in R (R Core Team, 2020).

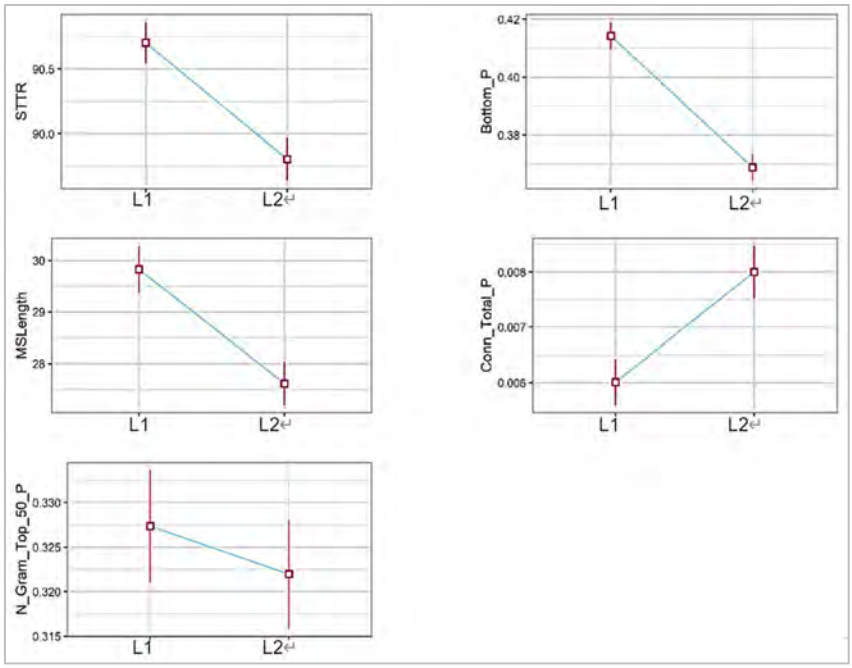
4. RESULTS

4.1. Feature Importance Rankings for Universal Translationese

We designed the present study to identify the feature importance rankings of universal translationese markers across academic disciplines. To this end, we first compared L1 writers with L2 writers across disciplines. Preliminary data analyses involved examining each variable's mean values and confidence intervals (CIs) to compare L1 writers and L2 writers in the combined academic domain. In Figure 1, the squared points in the plot graphs of the initial descriptive statistical analyses indicate means, and I-shaped error bars represent a 95% CI level. In each plot graph, when the CI of one group does not overlap with the CI of the other group, the variable is statistically significant between the means at a 0.05 significance level, which demonstrates that the variable behaves differently in the two independent groups being compared and that the variable can be utilized to distinguish the groups.

FIGURE 1

Means and 95% CIs: L1 and L2 Groups, Combined Domain (L1_All vs. L2_All)



In the second phase of analysis, we observed interaction effects between L1 and L2 writers

in the combined domain. As depicted in Figure 1, there were no interaction effects between the groups for four encoded variables (STTR, MSL, BOTTOM_P, and CONN_ALL_P). There were, however, interaction effects for one variable (N_GRAM_TOP50_P). The mean values of the four non-interacting variables graphed with the L1_All group in higher means than the L2_All group for three variables (STTR, MSL, and BOTTOM_P) but graphed with the L1_All group in lower means than the L2_All group for the CONN_ALL_P variable. As shown, the four encoded variables behaved differently, showing no interactions, which indicated a statistical significance for separating writer groups.

Meanwhile, the CIs of the N_GRAM_TOP50_P factor showed a partially overlapping pattern, indicating no statistical significance when compared between L1 and L2 writers. In order to evaluate interaction effects, we applied a linear mixed-effects model to interrogate inferential statistics. Table 3 outlines the summary of each encoded variable's statistical values for L1 and L2 writers in the combined domain.

TABLE 3

Variable Effects: Combined Domain (*df*=1)

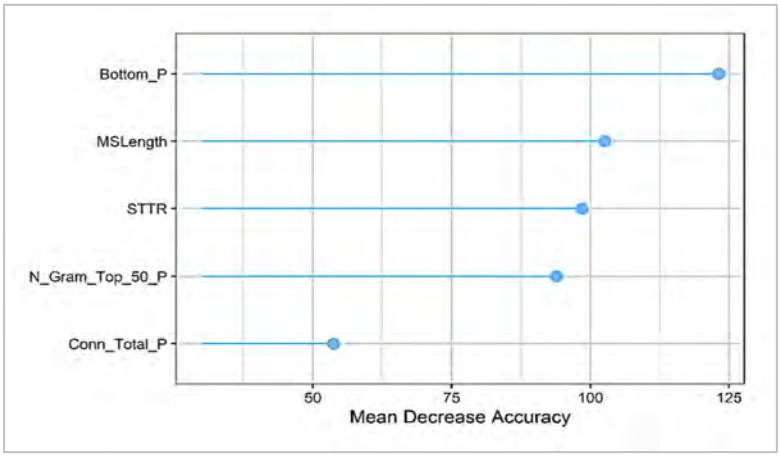
	L1_All vs. L2_All	
	X^2	*p*-value
STTR	59.992	<0.001
MSL	47.678	<0.001
BOTTOM_P	183.620	<0.001
CONN_ALL_P	39.659	<0.001
N_GRAM_TOP50_P	1.753	=0.185

For the third phase of statistical analyses, we utilized a random forest model to rank the variables' importance in distinguishing L2 writers' texts from L1 writers' texts. Results identified the most critical predictors contributing to differences in the combined domain setting. In Figure 2, a plot graph computed using a random forest model, the MDA represents how much the model's accuracy decreases if we drop a variable. The higher the value of the MDA, the higher the variable's importance in the model. The variable positioned at the top is typically considered the most important variable in the plot, whereas the variable positioned at the bottom is considered the least important.

Figure 2 illustrates the feature importance ranking of translationese markers between L1 and L2 writers. As shown, the five variables leveled from BOTTOM_P, MSL, STTR, N_GRAM_TOP50_P to CONN_ALL_P in a decreasing pattern. The BOTTOM_P variable was considered to be the most robust predictor of translationese amongst the others.

FIGURE 2

Feature Importance Ranking: Combined Domain (L1_All vs. L2_All)[4]



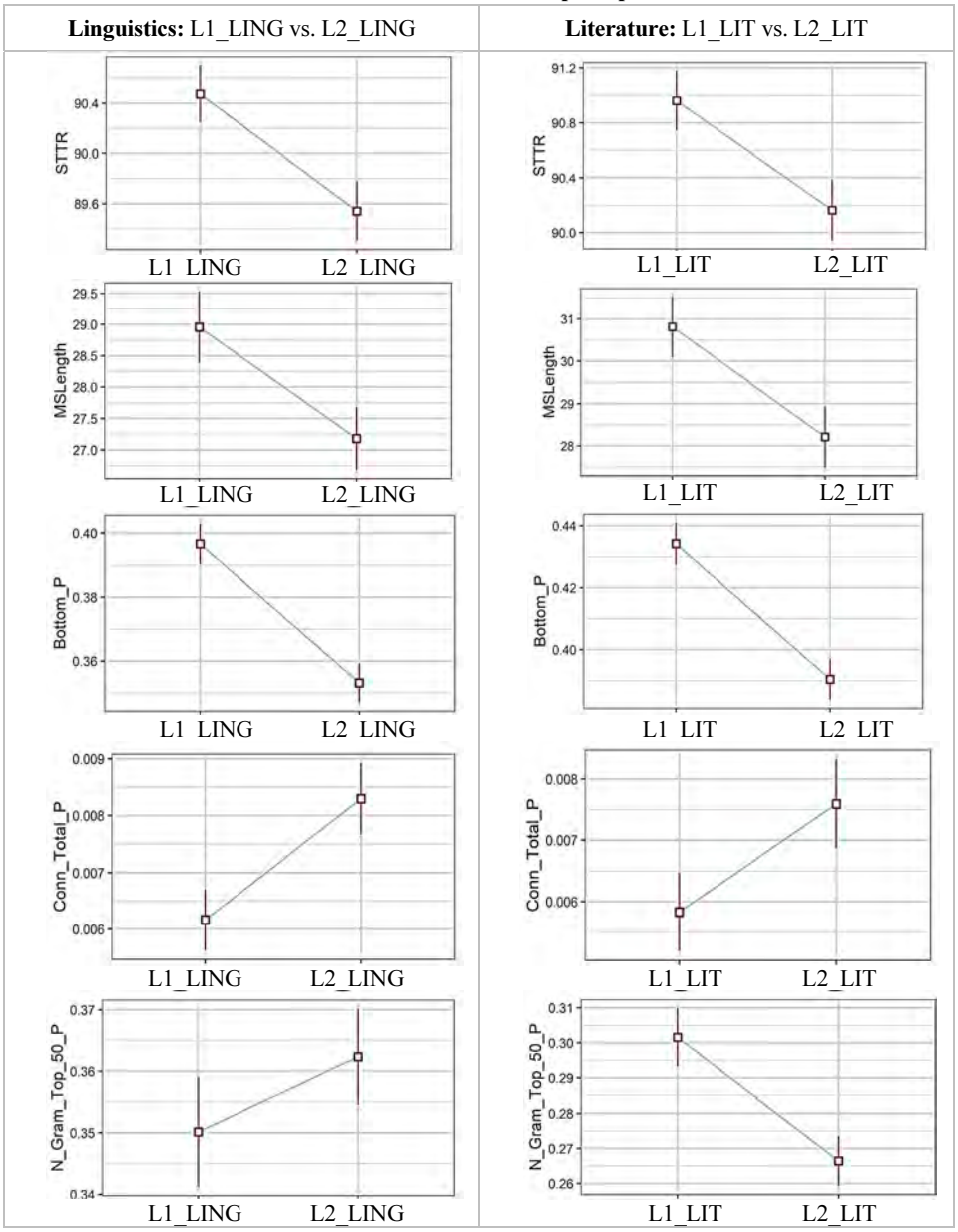
4.2. Feature Importance Rankings for Discipline-specific Translationese

In preliminary data analyses, we examined overall mean values and CIs in each academic discipline to evaluate variable importance rankings for discipline-specific translationese. Figure 3 presents a range of plausible values for the selected variables within each academic discipline, respectively. In the discipline of linguistics, as shown in the left column of Figure 3, the mean values of three encoded variables (STTR, MSL, and BOTTOM_P) graphed with the L1_LING group in higher means than those of the L2_LING group, whereas the opposite patterns prevailed for the remaining two variables (CONN_ALL_P and N_GRAM_TOP50_P). In literature, however, the mean values of four variables (STTR, MSL, BOTTOM_P, and N_GRAM_TOP50_P) graphed with the L1_LIT group in higher means than those of the L2_LIT group; only one variable (CONN_ALL_P) came out conversely.

[4] Confidence intervals (CIs) and the feature importance ranking in a random forest model serve different purposes and are not directly comparable. CIs provide an estimated range for population parameters, while the feature importance ranking measures each feature's contribution to model accuracy. In this study, a factor with overlapping CIs (e.g., N_GRAM_TOP50_P) could have higher feature importance ranking than another factor with statistically significant CIs (e.g., CONN_ALL_P) due to the random forest model considering a combination of factors when making predictions. It is crucial to interpret the results from both analyses within their respective contexts rather than drawing direct comparisons. Instead, view them as complementary information, providing a comprehensive understanding of the relationships between factors and the target variable.

FIGURE 3

Means and 95% CIs: L1 and L2 Groups, Separate Domains



Then, we further analyzed with a linear mixed-effects model to confirm whether the two groups could be observed separately. We found no interactions, indicating a statistical difference between the academic domains and clearance to examine them separately. As

outlined in Table 4, the linguistics discipline demonstrated a highly significant effect on four factors (STTR, MSL, BOTTOM_P, and CONN_ALL_P) and a significant effect on one variable (N_GRAM_TOP50_P); however, inferential statistical analyses using a linear mixed-effects model indicated that all five selected variables turned out to be highly significant in the literature domain. Overall, each variable behaved differently with statistical significance except for the N_GRAM_TOP50_P factor, which validated our separate observation of the two domains.

TABLE 4

The Effects of Each Variable in Separate Domains (*df*=1)

	Linguistics L1_LING vs. L2_LING		***Literature*** L1_LIT vs. L2_LIT	
	X^2	***p*-value**	X^2	***p*-value**
STTR	31.466	<0.001	25.945	<0.001
MSL	20.897	<0.001	24.817	<0.001
BOTTOM_P	91.665	<0.001	78.030	<0.001
CONN_ALL_P	26.075	<0.001	13.158	<0.001
N_GRAM_TOP50_P	4.303	<0.05	37.456	<0.001

The random forest analysis results identified the most important factors contributing to L1 and L2 textual differences in each discipline as well. As clearly shown in Figure 4, in the linguistics discipline, the five variables ranked in descending order from BOTTOM_P, N_GRAM_TOP50_P, MSL, STTR, to CONN_ALL_P. The BOTTOM_P factor was the most robust predictor of translationese, which means the L2 texts were more simplified than their counterparts in that they contained a lower portion of one-time occurring single frequency words. In other words, it can be deduced that lexical sophistication was a feature of the L1 texts.

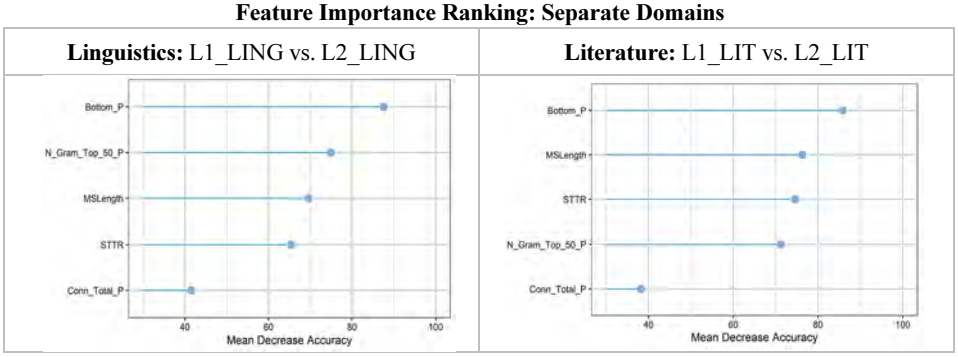
On the other hand, the CONN_ALL_P variable was the least essential predictor compared to other variables, meaning that the use of connectives may not have fully explained how different the L2 texts were from the L1 texts. In addition, the N_GRAM_TOP50_P variable functioned as the second-most important indicator of translationese, showing that greater use of lexical bundles was, presumably, an indicator of translational manifestations in the present study's L2 writings. The mean sentence length variable, abbreviated in MSL, was positioned in the middle, a yet valid but not quite prominent, robust indicator of translationese in the linguistics sub-corpus of the present study. Figure 4's left column illustrates the feature importance ranking of translationese markers between L1 and L2 writers in linguistics.

In the discipline of literature, the random forest analysis results demonstrated the five translationese variables ranked in descending order from BOTTOM_P, MSL, STTR, N_GRAM_TOP50_P, to CONN_ALL_P. Identical to the linguistics sub-corpus in the present study, BOTTOM_P predominated with the most potential in the literature field while

CONN_ALL_P showed minor potential among the five variables. MSL and STTR were yet effective translationese predictors but not as potent as BOTTOM_P. The right column in Figure 4 visualizes the feature importance ranking of translationese predictors between L1 and L2 writers in the literature domain.

FIGURE 4

Feature Importance Ranking: Separate Domains

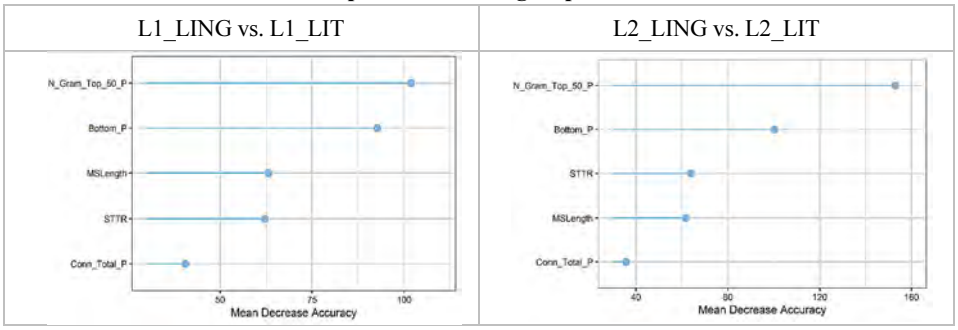


The normalization classifier yielded mixed results. Surprisingly, the N_GRAM_TOP50_P variable performed oppositely in the disciplines. The N_GRAM_TOP50_P variable, then, indicated ambivalent linguistic behavior in each discipline, implying the possibility that N_GRAM_TOP50_P may not have fully served a predictive role in universal translationese behaviors irrespective of academic disciplines in the present study.

As observed in section 4.1, of the five selected variables, N_GRAM_TOP50_P proved the only factor with no statistical significance in the combined domain. On the other hand, when we tested the variable in separate domains, as shown in section 4.2, it showed significance in linguistics and high significance in literature. To better understand the unstable behavior of N_GRAM_TOP50_P, we conducted two additional random forest analyses by performing comparisons between L1 writers and L2 writers, respectively: one test compared L1 writers in the linguistics discipline to those in the literature discipline ($t(1,128) = 58.592$, $p < .001$), and the other test established patterns between the same group of L2 writers in linguistics and literature ($t(1,038) = 260.46$, $p < .001$). Interestingly, as shown in Figure 5, the N_GRAM_TOP50_P factor ranked at the top, suggesting the highest predictive power of translationese in these two additional experimental settings.

FIGURE 5

Feature Importance Ranking: Separate Domains



5. DISCUSSION

We used a random forest model to determine the feature importance ranking of five specified translationese markers in L2 writing. The results demonstrated that all concerned factors, except N_GRAM_TOP50_P, accounted for differences between the L1 and L2 written texts. Further, among the five variables, the factor of BOTTOM_P was the most robust predictor, and CONN_ALL_P was the least. A close look at the data within each distinct discipline allowed us to confirm the findings across disciplines as a whole: the factor of BOTTOM_P held as the most robust predictor, and the factor of CONN_ALL_P held as the least. Overall, the present study's findings show that linguistic features concerned with simplification (BOTTOM_P, MSL, and STTR) are more vital predictors of translationese than those concerned with explicitation (CONN_ALL_P) and normalization (N_GRAM_TOP50_P); moreover, it is worth noting that in both the combined and separate disciplinary settings, the MSL factor prevailed as a more potent predictor of translationese than STTR. We also found that the N_GRAM_TOP50_P factor was discipline-specific, indicative of genre effects on written texts.

Given the nature of research papers (comprising discipline-specific jargon and registers), we did not expect that bottom-frequency words would hold as the most potent factor distinguishing L1 and L2 texts; however, the results of the present study countered our expectations and, instead, strongly supported the factor's robust predictability of translationese within and across disciplines. While holding the most potent predictive power in all settings, the BOTTOM_P factor pertains to the idea of a "hapax legomenon," often referred to as a "hapax," meaning a word occurring only one time in a body of texts. A hapax's relative frequency is typically concerned with lexical variety (see Baker, 1996, 2007). In the present study, L1 writers' texts contained a greater incidence of bottom-

frequency words than did L2 writers' texts in both combined and separate disciplinary settings, with the L1 texts exhibiting greater vocabulary richness than L2 texts. Based on the distribution of words by frequency, the occurrence of low-frequency words has generally been understood as an index that discriminates between L2 proficiency levels (Crossley & McNamara, 2011, 2012; Jarvis, 2002; Zareva, chwanenflugel, & Nikolova, 2005) and often determines the overall writing quality (Kyle & Crossley, 2014; Vögelin, Jansen, & Keller, 2019). From an L2 developmental perspective, several studies further suggested that L2 writers' lexical use, with some growth in their L2 competence, follows a pattern analogous to L1 writers' lexical use (Nasseri & Thompson, 2021; Pietilä, 2015); however, the pattern of lexical use identified in the L2 scholars' texts of the present study was not identical to—rather, far more simplified than—that of the L1 scholars' texts. In this case, we assume that decreased use of bottom-frequency words reflected the L1 literacy experiences of the L2 writers (and as a sign of translationese), not their low L2 proficiency or text quality. Yazici (2013) argued that simplification due to translation activity is out of line with what we usually perceive to be a "simple style." An L2 writer's recourse to plainness in language—as observed in the mean differences of BOTTOM_P, MSL, and STTR in the present study—is "to disambiguate the information load" in the transfer of knowledge from a foreign language rather than indicative of "stylistic impoverishment" (p. 1101).

As a second robust predictor of translationese, the MSL factor also appeared to characterize the L2 texts of the present study, whereby the mean length of sentences in the L2 texts was lower than that in the L1 texts. This result corroborates previous findings: the average sentence length in research papers generated by native speakers of English is generally higher than that of non-native speakers (Deveci, 2019; Mertens, 2008). Deveci's (2019) study added that native speakers' inclination to write longer sentences does not necessarily lead to a preponderance of complex and compound sentence types; in his study, non-native speakers used complicated sentence types more frequently to exhibit their sophisticated use of the English language. A growing body of research suggests the average sentence length in academic writing falls between 20 and 25 words—the readability range—and sometimes within 15 words, especially in scientific fields (Garner, 2000; Griffies, Perrie, & Hull, 2013). In the linguistics sub-corpus of the present study, however, the average length of sentences produced by the L1 writers was nearly 29 words, whereas the mean length of the sentences produced by the L2 writers exceeded no more than 27 words. The case was no different for the other discipline; in the literature sub-corpus, the mean sentence length of the L1 texts reached almost 31 words, and that of the L2 texts was around 28 words. Whether the texts came from L1 or L2 writers, the literature field's mean sentence length was relatively more extended than the linguistics field's mean sentence length, and there was no remarkable difference in sentence length between the L1 linguistics scholars and L2 literature scholars. It is conceivable that the style of writing preferred in linguistics—as a

social science subfield—aligns with the recent trend in scientific writing of lowering word counts (see Moore, 2011). The MSL factor, then, could be discipline-dependent—but not so clearly as the N_GRAM_TOP50_P factor.

As shown in Figure 4, N_GRAM_TOP50_P proved to be the second most robust predictor of translationese for the linguistics field in the present study but the fourth most robust predictor for the literature field. In addition, as shown in Figure 5, the N_GRAM_TOP50_P factor held as the strongest predictor of disciplines even when compared between L1 writers in linguistics and literature and among L2 writers. After confirming N_GRAM_TOP50_P as a discipline-specific predictor of translationese through a random forest analysis, we would like to draw particular attention to highly recurring three-word lexical bundles such as cliches, idioms, prefabricated structures, and untypical language (see Conrad & Biber, 2004). Lexical bundles have been described as ready-made expressions (Baker, 1996), prefabricated linguistic devices (Øverås, 1998), and typical collocations (Olohan, 2004). This linguistic feature was discernible in the present study and turned out to be the marker most influenced by disciplinary nature. Whether in the L1 or L2 texts, top-50 trigram lexical bundles occurred more often in linguistics than in literature, thereby constituting a property of translationese strongly associated with the linguistics discipline. The linguistics field utilizes a systematic approach for studying the structural aspects of language. In contrast, the literature field engages in the analysis and appraisal of many genres, ranging from poetry and dramas to novels across historical periods, thus devoid of structural rigidity. The salient disparity in the use of recurrent word combinations or lexical phrases, then, may bear on the distinctive characteristics of—and stylistic preferences developed in—each discipline, in terms of disciplinary thinking, theoretical approaches, and genres in use.

A related finding that warrants further examination is that in the linguistics sub-corpus, the L2 texts contained more trigram lexical bundles than did the L1 texts, but, in the literature sub-corpus, trigrams appeared more in the L1 texts than in the L2 texts. A close look at the linguistics sub-corpus indicates that the trigrams that incorporated verb phrase fragments characterized the L1 texts. On the other hand, the high-ranking trigrams in the L2 texts were mainly composed of noun phrases and prepositional phrase fragments (see Biber, Conrad, & Cortes, 2004) for the structural taxonomy of lexical bundles). The n-gram analysis showed that the verb phrasal expressions in the L1 texts were aimed at establishing an author's epistemic stance (*it is argued*, *is argued that*, *i argue that*, *we argue that*, *argue that the*, *argues that the*, *we show that*), none of which was noted in the L2 texts. Yet, most trigram expressions in the L2 texts were intended to direct readers' attention to the stages or sequences of textual elements (*the purpose of*, *purpose of this*, *results showed that*, *the results of*, *results show that*, *this study investigates*, *this paper is*). These types of expressions, which hardly appeared in the L1 texts of the present study, are often labeled "frame markers" (Hyland, 2005) or "discourse organizers" (Biber et al., 2004).

The results from our inspection of high-ranking instances across n-gram sizes in the literature sub-corpus were a mirror image of those from the linguistic sub-corpus. The general pattern identified in the L1 texts of the literature sub-corpus showed increased use of noun phrases with *of*-phrase fragments such as *part of the*, *one of the*, *the case of*, *the role of*, *the context of*, *the notion of*, *the use of*, and *the history of*, which corresponds to Hyland's (2008) finding that noun phrases with *of*-phrase fragments are the most defining feature of academic registers. That said, it is noteworthy that there were only three verb phrasal trigrams ranked in the top 50 of the present study's L1 texts. Noun phrases with *of*-phrase fragments occupied a high portion of the L2 texts as well; however, as compared to the L1 texts, there was a frequent occurrence of verb phrasal trigrams as shown in the examples of *this paper is*, *this paper aims*, *paper aims to*, *this paper examines*, and *paper is to*. In both the linguistics and literature sub-corpora, the L2 writers employed linguistic means or lexical items—in the form of either noun or verb phrase fragments—to signal text boundaries and announce discourse goals. This remarkable difference between the L1 and L2 texts within and across disciplines brings us back to the larger picture of translationese imprinted on L2 texts. In light of normalization (see Baker, 2007; Laviosa, 1998; Øverås, 1998), L2 writers' adherence to cliché-ridden, formulaic expressions can be interpreted as their conscious, or subconscious, attempts to attenuate a sense of foreignness by strictly complying with the norms governing the target genre.

Last, as observed in the statistical differences between the L1 and L2 texts, the CONN_ALL_P factor, indeed, functioned as a translationese predictor within and across disciplines—but not as markedly as other factors. A general understanding in SLA posits that L2 writers' attempts to unify their ideas manifest in a preponderance of particular types of cohesive devices, considered a linguistic pattern born of their limited syntactic and lexical repertoires (Hinkel, 2001; Paquot, 2008); however, given that the written texts in the present study's corpora were published journal articles, viewing L2 writers as short of available (linguistic) means for text coherence is not warranted. Alternatively, connectives are thought to indicate textual explicitness, a property often seen in texts produced through translation processes (Blum-Kulka, 1986; Øverås, 1998). Dimitrova (2005) explained, based on her psycholinguistic investigation of translation, that a writer engaged in translation attempts to make implicit meanings explicit, triggered by lexico-grammatical and pragmatic contrasts between two paired languages. Another motivation for explicitation, she added, may be strategically approaching the difficulty of processing from one language to another; when encountering processing difficulties, a writer can explicitate and reformulate text to make processing less cognitively demanding instead of reverting to the original ideas and finding equivalent linguistic chunks for translation.

6. CONCLUSION AND IMPLICATIONS

Previous studies have demonstrated that the L2 writing process is not entirely independent of the influence of the L1 and/or L1 writing process and that signs of translationese in an L2 text distinguish it from an L1 text. A critical aspect of the present study was identifying the covert signs of universal translationese in L2 texts, particularly in texts produced by highly advanced L2 writers. It is important to note that translational properties inherent in L2 texts should not be considered a 'stain' to be removed; rather, they should be viewed as mere textual features that make L2 writing a unique form of linguistic expression among various others. Nevertheless, in a classroom setting, teachers can help L2 writers expand their linguistic repertories, which are constructed and contained by L1 processing, to manipulate the language into different styles and registers more effectively and efficiently. Moreover, for this pedagogical aim, it is suggested that L2 writers become aware of available linguistic varieties beyond those related to L1 processing (i.e., translationese markers) and employ metacognitive strategies for optimal choice.

In this context, the present study offers practical guidelines for curriculum development concerning scope and sequence. Study findings revealed bottom-frequency words are the most potent predictor, while connectives are the least potent predictor of translationese across two different academic settings. Teachers may thus consider sequencing the curriculum in the order of feature importance ranking of translationese markers, as identified in the study. In light of the nature of scholarly journal articles, which encompass discipline-specific registers, we did not, in fact, expect bottom-frequency words to be the most prominent, universal force that distinguishes L1 texts from L2 texts; however, the present study's findings contradict these expectations and convincingly corroborate the hapax's robust predictive power of translationese within and across disciplines. Given the significance of vocabulary size and proficiency, it is suggested that we ensure the occurrences of mid- and low-frequency vocabulary in learning materials and guide L2 writers toward "noticing" and "engagement" with the target vocabulary in a principled manner.

Some caution is warranted in interpreting the findings. First, these findings do not provide a comprehensive explanation because representing the cognitive process of L2 writing in corpora is not a direct substitute for human cognition. Moreover, the results could be strengthened by investigating other translationese features and generating converging evidence. Future studies may explore the interaction of such features in other academic disciplines. As for the context of writing, we only included the fields of linguistics and literature in the present study. Future research could benefit from examining the timed writing of advanced L2 writers or collecting multiple pieces of writing from each writer. Such methodological changes would allow for additional evidence to consider when

examining the role of translation, either consciously or subconsciously, in the L2 writing process.

Regarding the corpus dataset selection, we acknowledge concerns about potential limitations in distinguishing between L1 scholars, highly proficient non-native scholars, and those who could be more proficient in the language. Nevertheless, our research primarily focuses on investigating the presence and feature importance ranking of translationese markers in L2 writing rather than individual authors' proficiency levels. While verifying a researcher's native speaker status based on his/her affiliation or residency may not be flawless, our study aims to uncover general patterns and trends across a large dataset, helping to mitigate the potential impact of individual inaccuracies. The challenge of distinguishing between highly proficient non-native speakers and less proficient speakers exists. However, this difficulty underscores the significance of our findings, as it suggests the possible presence of translationese markers even in the writing of highly proficient non-native speakers. In the context of our research, we maintain that our dataset selection and classification methodology are suitable for identifying the feature importance ranking of translationese markers in L2 writing.

Applicable level: Tertiary

REFERENCES

Al-Shabab, O. S. (1996). *Interpretation and the language of translation: Creativity and conventions in translation*. London: Janus Book Publishers.

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 Publishing.

Baker, M. (1996). Corpus-based translation studies: The challenges that lie ahead. *Benjamins Translation Library, 18*, 175-186.

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

Blum-Kulka, S. (1986). Shifts of cohesion and coherence in translation. In J. House & S. Blum-Kulka (Eds.), *Interlingual and intercultural communication: Discourse and cognition in translation and second language acquisition studies* (pp. 17-35). Tübingen, Germany: Narr.

Beare, S., & Bourdages, J. S. (2007). Skilled writers' generating strategies in L1 and L2: An exploratory study. In G. Rijlaarsdam (Series Ed.) & M. Torrance, L. Van Waes & D. Galbraith (Volume Eds.), *Writing and cognition: Research and applications* (pp. 151-161). Amsterdam: Elsevier.

Biber, D., Conrad, S., & Cortes, V. (2004). If you look at...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405.

Breiman, L. (2001). Random forests. *Machine learning*, *45*(1), 5-32.

Chamot, A. U. (1987). The learning strategies of ESL students. In A. L. Wenden & J. Rubin (Eds.), *Learner strategies in language learning* (pp. 71-83). New York: Prentice-Hall.

Chesterman, A. (2004). Hypotheses about translation universals. *Benjamins Translation Library*, *50*, 1-14.

Cohen, A. D., & Brooks-Carson, A. (2001). Research on direct versus translated writing: Students' strategies and their results. *The Modern Language Journal, 85*, 169-188.

Conrad, S., & Biber, D. (2004). The frequency and use of lexical bundles in conversation and academic prose. *Lexicographica, 20*, 56-71.

Crossley, S. A., & McNamara, D. S. (2011). Shared features of L2 writing: Intergroup homogeneity and text classification. *Journal of Second Language Writing, 20,* 271-285.

Crossley, S. A., & McNamara, D. S. (2012). Predicting second language writing proficiency: The roles of cohesion and linguistic sophistication. *Journal of Research in Reading, 35*(2), 115-135.

Crossley, S. A., & McNamara, D. S. (2014). Does writing development equal writing quality? A computational investigation of syntactic complexity in L2 learners. *Journal of Second Language Writing, 26*(4), 66-79.

Cumming, A. (1989). Writing expertise and second language proficiency. *Language Learning, 39*, 81–141.

Deveci, T. (2019). Sentence length in education research articles: A comparison between Anglophone and Turkish authors. *The Linguistics Journal, 13*(1), 73-100.

Dimitrova, B. E. (2005). *Expertise and explicitation in the translation process* (Vol. 64). Amsterdam/Philadelphia: John Benjamins Publishing.

Galbraith, D. (1999). Writing as a knowledge-constituting process. In M. Torrance & D. Galbraith (Eds.), *Knowing what to write: Conceptual processes in text production* (pp. 139-160). Amsterdam: Amsterdam University Press.

Garner, G. A. (2000). *The Oxford dictionary of American usage and style*. New York: Oxford University Press.

Gaspari, F. (2013). A phraseological comparison of international news agency reports published online: Lexical bundles in the English-language output of ANSA, Adnkronos, Reuters and UPI. *Varieng. Studies in Variation, Contacts and Change in*

English, *13*. Retrieved on March 31, 2023, from https://varieng.helsinki.fi/series/volumes/13/gaspari/

Gaspari, F., & Bernardini, S. (2008). Comparing non-native and translated language: Monolingual comparable corpora with a twist. *Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies* (pp. 215-234). Hangzhou, China.

Gellerstam, M. (1986). Translationese in Swedish novels translated from English. In L. Wollin & H. Lindquist (Eds.), *Translation studies in Scandinavia* (pp. 88-95). Lund, Sweden: CWK Gleerup.

Goh, G-Y., & Lee, Y. C. (2016). A corpus-based study of translation universals in English translations of Korean newspaper texts. *Cross-Cultural Studies*, 45, 109-143.

Goh, G-Y., Lee, Y. C., & Kim, D-Y. (2016). A corpus-based study of translation universals in English translations of Korean newspaper texts. *Cross-Cultural Studies*, 45, 109-143.

Göpeferich, S. (2017). Cognitive functions of translation in L2 writing. In J. W. Schweiter & A. Ferreira. (Eds.), *The handbook of translation and cognition* (pp. 402-422). New York: John Wiley & Sons.

Grabowski, Ł. (2012). On translation universals in selected contemporary Polish literary translations. *Studies in Polish linguistics*, *7*(1), 165-183.

Granger, S. (2015). Contrastive interlanguage analysis: A reappraisal. *International Journal of Learner Corpus Research*, *1*(1), 7-24.

Griffies, S. M., Perrie, W. A., & Hull, G. (2013). Elements of style for writing scientific journal articles. *Publishing Connect, Elsevier.* Retrieved from https://virayeh.com/SampleDoc/Virayeh20153204.pdf

Halverson, S. L. (2003). The cognitive basis of translation universals. *Target: International Journal of Translation Studies*, *15*(2), 197-241.

Hinkel, E. (2001). Matters of cohesion in L2 academic texts. *Applied Language Learning, 12,* 111-132.

Hyland, K. (2005). *Metadiscourse: Exploring interaction in writing*. London: Continuum.

Hyland, K. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(4), 4-21.

Ivaska, I., & Bernardini, S. (2020). Constrained language use in Finnish: A corpus-driven approach. *Nordic Journal of Linguistics*, *43*(1), 33-57.

Jarvis, S. (2002). Short texts, best-fitting curves, and new measures of lexical diversity. *Language Testing*, *19*(1), 57-84.

Jarvis, S., & Pavlenko, A. (2008). *Crosslinguistic influence in language and cognition*. New York: Routledge.

Kern, R. (1994). The role of mental translation in second language reading. *Studies in Second Language Acquisition, 16*, 441-461.

Kobayashi, H., & Rinnert, C. (1992). Effects of first language on second language writing: Translation versus direct composition. *Language Learning, 42*, 183-215.

Kolehmainen, L., Meriläinen, L., & Riionheimo, H. (2014). Interlingual reduction: Evidence from language contacts, translation and second language acquisition. In H. Paulasto, L. Meriläinen, H. Riionheimo & M. Kok (Eds). *Language contacts at the crossroads of disciplines* (pp. 3-32). Cambridge, England: Cambridge Scholars Publishing.

Koppel, M., & Ordan, N. (2011, June). *Translationese and its dialects*. Paper presented at the Forty-Ninth Annual Meeting of the Association for Computational Linguistics: Human Language Technologies, Oregon, USA.

Kruger, H., & van Rooy, B. (2016). Constrained language: A multidimensional analysis of translated English and a non-native indigenised variety of English. *English World-Wide*, *37*(1), 26-57.

Kuznetsova, A., Christensen, R. H. B., & Brockhoff, P. B. (2013). Different tests on lmer objects (of the lme4 package): Introducing the lmerTest package. *Journal of Statistical Software*, *55*(13), 1-27.

Kyle, K., & Crossley, S. A. (2014). Automatically assessing lexical sophistication: Indices, tools, findings, and application. *TESOL Quarterly, 49*(4), 757-786.

Lanstyák, I., & Heltai, P. (2012). Universals in language contact and translation. *Across Languages and Cultures*, *13*(1), 99-121.

Laviosa, S. (1998). Core patterns of lexical use in a comparable corpus of English narrative prose. *Meta: Translators' Journal*, *43*(4), 557-570.

Laviosa, S. (2002). *Corpus-based translation studies: Theory, findings, applications*. Amsterdam: Rodopi.

Lee, Y. C. (2014, November). A corpus-based study of translation universals in English translations of Korean newspaper texts. *Proceedings of the 2014 International Conference of the English Language and Literature Association of Korea: ELLAK* (pp. 22-23), Seoul, Korea.

Lee, Y. C. (2018). The hallmarks of L2 writing viewed through the prism of translation universals. *Linguistic Research*, *35*, 171-205.

Lee, Y. C. (2019). Spotting non-nativeness in L2 texts: A statistical approach to translationese. *Studies in English Language and Literature*, *45*(1), 367-388.

Lee, Y. C. (2021). Function words as markers of translationese: A corpus-based approach to mental translation in second language writing. *Korean Journal of English Language and Linguistics*, *21*, 261-281.

Liaw, A., & Wiener, M. (2002). Classification and regression by RandomForest. *R news*, *2*(3), 18-22.

Liu, Y. (2009). *Translation in second language writing: Exploration of cognitive process of translation*. Saarbrücken, Germany: VDM.

Lu, X. (2011). A corpus-based evaluation of syntactic complexity measures as indices of college-level ESL writers' language development. *TESOL Quarterly, 45*(1), 36-62.

Mauranen, A. (2007). Universal tendencies in translation. In G. Anderman & M. Rogers, (Eds.), *Incorporating corpora* (pp. 32-48). Bristol, England: Multilingual Matters.

Mertens, S. (2008). The check-up before publication. *Deutsches Ärzteblatt International, 5*(51-52), 897-899.

Moore, A. (2011). The long sentence: A disservice to science in the Internet age. *Bioessays, 33*(12), 193-193.

Munday, J. (2016). *Introducing translation studies: Theories and applications*. London: Routledge.

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

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

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

Paquot, M. (2008). Exemplification in learner writing: A cross-linguistic perspective. In F. Meunier & S. Granger (Eds.), *Phraseology in language learning and teaching* (pp. 101-119). Amsterdam: John Benjamins.

Pietila, P. (2015). *Lexical issues in L2 writing*. New Castle upon Tyne, England: Cambridge Scholars Publishing.

Qi, D. S. (1998). An inquiry into language-switching in second language composing processes. *The Canadian Modern Language Review, 54*, 413-435.

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna Austria: R Foundation for Statistical Computing. https://www.R-project.org/

Rabinovich, E., Nisioi, S., Ordan, N., & Wintner, S. (2016, August). *On the similarities between native, non-native, and translated texts*. Paper presented at the Fifty-Fourth Annual Meeting of the Association for Computational Linguistics, Berlin, Germany.

Reid, J. R. (1992). A computer text analysis of four cohesion device in English discourse by native and nonnative writers. *Journal of Second Language Writing, 1*, 79-107

Ringbom, H. (2007). *Cross-linguistic similarity in foreign language learning*. Clevedon, England: Multilingual Matters.

Roca de Larios, J., Murphy, L., & Manchón, R. M. (1999). The use of restructuring strategies in EFL writing: A study of Spanish learners of English as a foreign language. *Journal of Second Language Writing, 8*, 13–44.

Sasaki, M. (2002). Building an empirically based model of EFL learners' writing processes. In G. Rijlaarsdam (Series Ed.) & S. Ransdell & M. Barbier (Volume Eds.), *Studies in writing, volume 11: New directions for research in L2 writing* (pp. 49-80). Dordrecht, The Netherlands: Kluwer Academic Publishers.

Sasaki, M., & Hirose, K. (1996). Explanatory variables for EFL students' expository writing. *Language Learning, 46*, 137-174.

Uzawa, K., & Cumming, A. (1989). Writing strategies in Japanese as a foreign language: Lowering or keeping up the standards. *The Canadian Modern Language Review, 46*, 178–194.

Vögelin, C., Jansen, T., Keller, S. D., Machts, N., & Möller, J. (2019). The influence of lexical features on teacher judgments of ESL argumentative essays. *Assessing Writing*, *39*, 50-63.

Wang, L. (2003). Switching to first language among writers with differing second-language proficiency. *Journal of Second Language Writing, 12*, 347–375.

Wang, W., & Wen, Q. (2002). L1 use in the L2 composing process: An exploratory study of 16 Chinese EFL writers. *Journal of Second Language Writing, 11*, 225–246.

Whalen, K., & Ménard, N. (1995). L1 and L2 writers' strategic and linguistic knowledge: A model of multiple-level discourse processing. *Language Learning, 45*, 381–418.

Woodall, B. R. (2002). Language-switching: Using the first language while writing in a second. *Journal of Second Language Writing, 11*, 7–28.

Yazici, M. (2013, September). *Simplification as a translation universal.* Paper presented at the Thirteenth International Language, Literature, and Stylistics Symposium: Simple Style, Helsinki, Finland.

Zareva, A., Schwanenflugel, P., & Nikolova, Y. (2005). Relationship between lexical competence and language proficiency: Variable sensitivity. *Studies in Second Language Acquisition*, *27*(4), 567-595.