『영어영문학연구』 제45권 제1호
(2019) 봄 367－388

Studies in English Language & Literature
http://dx.doi.org/10.21559/aellk.2019.45.1.017

Spotting non-nativeness in L2 texts: A statistical approach to translationese* **

Younghee Cheri Lee
(Yonsei University)

Lee, Younghee Cheri. "Spotting non-nativeness in L2 texts: A statistical approach to translationese." ***Studies in English Language & Literature*** **45.1 (2019): 367-388.** Second language (L2) writing from the angle of translation universals (TU) offers substantial prospects of empirical research, but currently, only limited literature explains what linguistic factors shape non-nativeness in L2 writers' texts. This article claims to demonstrate that robust TU indices may predict non-nativeness, more particularly translationese from non-translated English texts produced by non-native scholars of English. The ultimate goal is, therefore, to classify text types using the indices of translationese, which will, in turn, signify linguistic factors of non-nativeness detectable in non-translated L2 texts. To this end, this study employed a collection of multi-factorial analysis methods to compare native scholars' L1 English corpora, respectively with two different variations of non-Anglophone scholars' non-translated L2 English corpora (L1 English vs. Quasi-L2 English vs. L2 English). The results provided evidence that most TU indices were valid to spot translationese as a signal of non-nativeness in expert non-native writers' journal abstracts. Additionally, the behavioral profiles of the selected TU indices demonstrated that the two variant L2 texts were clustered in higher mutual proximity due to intergroup homogeneity when compared to their native counterparts (Yonsei University).

Key Words: translationese, non-nativeness, translation universals, second language (L2) writing, L2 texts

---

* A preliminary version of this article was presented at the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32) held at the Hong Kong Polytechnic University, Hong Kong, SAR on December 2, 2018.

** I would like to express my heartfelt gratitude to Yong-hun Lee for his statistical support. I would also like to thank the three anonymous reviewers for their thoughtful review of the manuscript. All errors remain mine.

I. Introduction

Since robust algorithmic techniques made significant progress in the accessibility of large digitized corpora over the last decade, corpus-based and corpus-driven investigations have been clearly operational in mapping the methodological structure of empirical research queries in the second language (L2) text-based studies such as L2 writing and translation studies. A vast array of research on L2 texts to date has been primarily associated with the cognitive behavior models of the L2 writing process, thereby laying its theoretical groundwork. In particular, crosslinguistic influences have been the center of attractions and ongoing research concerns in second language writing (e.g. Cumming, 1990). Paramount research strands include interruptions, transfer, code-switching, positive interplay, and translation strategies (e.g. Bagheri & Fazel, 2011; Connor, 1999; Cumming, 1990; Grabe, 2001; Kellog, 1987; Jarvis & Pavlenko, 2008; Reid & Findlay, 1986; Sasaki, 2000; Silva, 1993; Swales, 1990; Uzawa, 1996; Ventola & Mauranen, 1991; Wang & Wen, 2002; Woodall, 2002).

Regarding the studies on translated L2 texts, nearly all investigations have yet centered on *descriptive translation studies*, in which research aims are predonminantly to delineate similarities and dissimilarities between natives' L1 originals and non-natives' L2 translations, thus automatically identifying group interactions (e.g. Baroni & Bernardini, 2006; Gaspari and Bernardini, 2008).

Meanwhile, a growing body of recent research has set in to draw particular attention to uncovering shared similarities among L2 texts to prove their peculiar and universal characteristics (c.f. Baker, 1993, 1995, 1996; Crossley & McNamara, 2011; Goh & Lee, 2016a; 2016b; Hinkel, 2002; Laviosa, 1998a, 1998b, 2002; Lee, 2017, 2018a).

Despite previous research efforts, however, unmet research needs remain further exploration. Although L2 writing studies hinged on the notion of *translation universals* (TU) have significant potential for empirical study, only limited studies

have included an interdisciplinary endeavor to intermingle L2 writing and descriptive translation studies. Such an approach is worthwhile in that it may render *non-nativeness* definable through the disclosure of universality in untranslated L2 texts (e.g. Lee, 2017, 2018a).

Driven by the motivation to gain higher insights into such universality, therefore, this article aims to discuss the potential linguistic attributes that form *non-nativeness* in L2 writers' texts by way of assessing the feasibility of the *translationese* indices (e.g., Lee, 2018a). The term *translationese* was initially raised by Gellerstam (1986), defining as the set of *fingerprints* that a source language leaves on a target language or vice versa, especially during the process of translation. This study extended its original definition to mean any 'linguistic fingerprints' or 'awkwardness' that L2 writers' target language leaves on their 'non-translated' L2 written production, aiming to collectively refer to any linguistic properties apparent in non-translated L2 texts, which are perceptively different from original L1 texts. To develop a 'baseline' notion of non-nativeness, this study will measure selected TU indices in *non-translated* L2 texts, not in L2 translations to define the non-nativeness of L2 texts through the prism of translationese. To further augment of the earlier findings, this study thus claims to prove the following two propositions in consonance with Baker's (1993) notion of translation universals:

(1) By using the TU indices, translationese will be detected in non-translated L2 texts of non-native English writers in comparison to original English language texts of native L1 writers, so that these indices will be proved to be valid for text-type identification.

(2) The intergroup homogeneity (e.g., similarities between different groups) of the TU indices will be measurable so that the TU indicators will be valid to classify text groups that share the universal characteristics of L2 English to define linguistic non-nativity.

To this end, this study will use the self-built CCERA corpora in two academic disciplines (linguistics and English literature), which is based on three different variations of English texts (L1 versus quasi-L2 versus L2 versus English). With reference to previous findings from my two prior research (see Lee, 2017, 2018a), this study will select eight key TU indices along with their encoded data and then statistically analyze using multi-factorial methods: Generalized Linear Model (GLM) to classify three variant text types and the Behavioral Profile (BP) analysis for clustering to observe intergroup homogeneity.

II. Related Work

2.1 Transationese as Universals of Translation

Since the early 1990s, in the field of translation studies, the development of multilingual corpora has strengthened empirical research interests into the *translational language*. As an apparatus language for a communicative event, the translational language is neither a target language (i.e., a language for translated texts) nor a source language (i.e., a language for original texts), having its typical linguistic characteristics. Such scholarly attention to the peculiar traits of a translational language has triggered a further advancement of a robust conceptual framework.

There have been diverse views raised over the terminology of translation universals such as ‘translationese’ (Gellerstam, 1986), ‘the third code’ (Frawley, 1984), ‘laws’ (Toury, 1995), ‘core patterns’ (Laviosa, 1998a), and many more. The proposal regarding the universals of translation was first put forward by its forerunner, Baker (1993). Reflecting that a translational language is pertinently associated with cognitive phenomena, she claimed that translation universals are any common linguistic attributes that are observable in translations rather than originals,

regardless of any language pairs (i.e., target and source languages) involved in the translating process (Baker, 1993. 1995, 1996). She meant those universal linguistic features as "by-products" driven by the mediating process between the target and source languages, rather than the effect of 'interference' caused by either target or source language (Baker, 1993, 1995, 1996; Laviosa, 1998a, 1998b, 2002, 2007).

Referring to Chesterman's (2004) assertion, the common proposals of translation universals are pertinent to unveiling the interrelationship between source texts and their target texts by using parallel corpora, as well as the linguistic relation between translated and non-translated texts both produced in the target language by utilizing monolingual, comparable corpora.

The last decades following the birth of translation universals have begun to share the commonly held notion that such universals of translation are linguistic characteristics that are typical of variant translated texts that differ not only from their source texts but also from comparable texts in the target language (Malmkjar, 2012; Mauranen, 2007; McEnery & Xiao, 2007; Munday, 2008; Xiao & Dai, 2014). It was also widely accepted that translated versions might 'under-represent' linguistic features of their counterparts which lack "obvious equivalents" in original texts (Mauranen, 2007). Consequently, such a viewpoint enables L2 writing scholars to infer that the effect of the source language on translations may be plausible enough to render translated texts perceptibly distinctive from original source texts.

2.2 Indicators of Translationese

In contemporary descriptive translation studies, translation scholars have continuously been engaged in conducting empirical studies to discover what factors and indices can represent translational attributes. Most potential indicators involve simplification, normalization, explicitation, and convergence.

Simplification is the tendency to consciously or unconsciously make target texts simpler lexically, syntactically and/or stylistically by using more straightforward

translational language to increase readability of target texts. (Baker, 1996). Opposing ideas such as lexical diversity, lexical density, lexical richness, structural sophistication, and stylistic complexity are all associated with simplification. Some useful parameters of simplification involve the STTR values for lexical diversity, function words over content words for lexical density, high to low-frequency words for lexical richness, and sentence splitting for structural sophistication, and semi-colons or full stops over commas for stylistic complexity (e.g. Baker, 1996; Laviosa, 1998b; Malmkjær, 2012).

Normalization centers on the idea that untypical language is more salient in target texts than their counterparts, thus causing awkwardness. The normalization indices include clichés, idioms, pre-fabricated structures of the target language, lexical bundles and collocations (Baker, 2007; Olohan, 2004; Øverås, 1998).

Explicitation is the most investigated feature among others. It is closely linked to translating strategies to increase the clarity of content in target texts by making lexical, syntactic, or semantic additions using more explicit and concrete translational language rather than leaving them implicit (Baker, 2006; Xiao & Dai, 2014), thereby making grammatical relations more explicit and cohesive. Most feasible indices predictable of the explicitation features involve connective devices such as conjunctions and complementizer (i.e. placing a clause in the position of a subject or an object of a sentence).

Meanwhile, comparatively less scholarly attention was paid to research into convergence (Laviosa, 2002). Often called leveling-out, convergence is pertinent to the idea that translated texts tend to group together towards the center of a continuum as they show greater closeness to one another lexically and syntactically. Some most feasible predictors of the convergence hypothesis include lower standard deviations of lexical variety, lexical density, type/token ratio, readability indices and mean sentence length.

III. Methods

3.1 Corpus Construction

Comparable monolingual corpora were constructed with the specific aim of observing recurrent, typical linguistic traits that might render Korean scholars' L2 English compositions perceptively different from those of native scholars' L1 English texts. The English texts were taken from acclaimed scholarly journal articles in two English-related disciplines to construct the Comparable Corpora of English Research Abstracts of Scholarly Journal Articles (CCERA).[1] Using simple random sampling, the CCERA was designed to be composed of three variants of L2 English texts and compiled so as to have balanced genre representation, size, and period to make equitable comparisons. The three sub-corpora include Korean scholars' L2 English abstracts whose research articles were written in L1 Korean (KE), Korean scholars' L2 English abstracts of which articles were produced in English (QE, meaning quasi-L2 English), and finally native scholars' L1 English abstracts (NE).

In particular, by the speculation that Korean scholars' Korean articles may have served as source texts, Korean scholars' L2 English abstracts have been separately categorized into two different groups to prevent such source-text effects, if any. The critical premise to note here is that the corpus data used in this study is non-translated L2 *compositions*, not L2 translations. The CCERA is mapped out in Table 1.

[1] The encoded corpus data for this study was drawn from the author's two prior research projects. The initial version of the CCERA was built for a doctoral dissertation, and it has been recently updated for the second project. The construction process including the list of databases assessed can be found in Lee (2017) and revised values of the dataset in Lee (2018a).

Table 1. The Scale of the CCERA

Sub-Corpus	Domain	Text (#)	Token (#)
KE *Korean L2 English*	English Linguistics	603	106,545
	English Literature	435	105,769
	Sub Total	1,038	212,314
QE *Quasi L2 English*	English Linguistics	605	106,195
	English Literature	440	107,869
	Sub Total	1,045	214,064
NE *Native L1 English*	English Linguistics	600	105,535
	English Literature	530	106,851
	Sub Total	1,130	212,386

3.2 Encoded Variables

A two-tier analysis was performed to select key TU indices indicative of translationese. As a preliminary analysis, probable variables that might explain universal features of translation were initially selected under theoretical considerations and previous empirical findings (see Lee 2017, 2018a). During the second tier, the eight TU indices that had shown high significance were encoded to identify the non-nativeness of L2 writers' texts. Baseline analyses were operated using the WordSmith Tools 7.0 and AntConc 3.4.4w programs, and all the statistical analyses were performed using R version 3.5.0 (2018). The information below and Table 2 briefly show sets of hypotheses for each variable encoded.

(1) STTR (Lexical Simplification): The Standardized Type/Token Ratio (STTR) of both QE and KE sub-corpora will be lower than that of the NE sub-corpus.

(2) Function Words (Lexical Simplification): The QE and KE texts will have different degrees of function words compared to native scholars' NE texts.

(3) High-Frequency Words (Lexical Simplification): The QE and KE corpora will have higher values of top 20 high-frequency words than the NE corpus.

(4) Bottom-Frequency Words (Lexical Simplification): Differently from the case of high-frequency words, QE and KE will hold fewer bottom-frequency words with one-time occurrence than their counterpart.

(5) Total Lexical Bundles (Lexical Normalization): The total proportions of recurring lexical bundles will be higher in QE and KE than in NE.

(6) Top 10 Lexical Bundles (Lexical Normalization): The QE and KE corpora will hold a greater amount of top 10 lexical bundles than the NE corpus.

(7) Connectives (Syntactic Explicitation): The ratio of connectives will be higher in the QE and KE corpora than in the NE corpus.

(8) Mean Sentence Length SD (Syntactic Convergence): The standard deviations of mean sentence length will be lower in both QE and KE than NE texts.

Table 2. Encoded Variables: Key TU Indices

TU Indices	Variables	Description
Simplification	STTR	Standardized Type/Token Ratio
	Funct_Total_P	Function Words (%)
	High_Top_20_P	Top 20 High-Freq. Words (%)
	Bottom_P	Bottom-Freq. Words (%)
Normalization	N_Gram_Total_P	Lexical Bundles: Trigrams (%)
	N_Gram_Top_10_P	Top 10 Trigrams (%)
Explicitation	Conn _P	Connectives (%)
Convergence	MSL_SD	Mean Sentence Length_*SD* (sd)

3.3 GLM Procedures and Output

As the dependent variable TextType was categorical in this study, a Generalized Linear Model (GLM) model as a linear regression method was applied to this study to evaluate linguistic factors that play vital roles in identifying three variants of

English texts: KE versus QE versus NE. For the implementation of a GLM model, an initial model was constructed first. The TextType was set as a dependent variable as the remaining factors became independent. Then, step-wise model selection processes were applied to the initial model constructed, and then insignificant factors were eliminated to produce the best model. Along with the final model, each variable was observed to judge statistical significance using a summary table and effect plots.

For the behaviors of each factor, effect plots were additionally employed to observe the confidence intervals (CIs) by the I-shaped error bars in each plot graph. If the CI of one group does not overlap with that of the other group, it means the factor is statistically significant, which demonstrates that the factor behaves differently in the two groups observed. Conversely, if two CIs overlap, it indicates that the factor behaves similarly in the two groups.

For the multinomial regression analysis, the initial model was set up followed by model selection procedures to select the most optimal model. The final model obtained was identical to the initial model, and thus all the eight main factors survived in the final model. By utilizing the final model, all the eight main factors were statistically analyzed. Table 3 outlines the final output of a GLM analysis. As shown, the *p*-value of each variable was less than 0.05, showing statistical significance. The results indicate that each factor can serve as a valid indicator to classify the TextType (KE vs. QE vs. NE) for the CCERA.

Table 3. The Output of GLM

Variables	*c*2	*df*	*p*
STTR	26.692	2	<0.001
Funct_Total_P	24.012	2	<0.001
High_Top_20_P	15.269	2	<0.001
Bottom_P	15.743	2	<0.001
N_Gram_Total_P	37.406	2	<0.001
N_Gram_Top_10_P	12.677	2	0.002
Conn_P	93.538	2	<0.001
MSL_SD	44.003	2	<0.001

3.4 BP Analysis

As another multi-factorial approach, a Behavioral Profile (BP) analysis was adopted. Developed by Gries and Otami (2010) and Gries (2010a), the BP analysis examines the behavioral properties of each linguistic factor by representing the similarity or dissimilarity of components in the form of a dendrogram. The BP method can be viewed as a hierarchical clustering algorithm where the behavioral profiles of each linguistic factor are adequately reflected (Gries, 2010a). The values in the dendrogram are not the *p*-values but the probabilities by which intergroup homogeneity is determined. In the dendrogram, if A converges with B rather than C, it indicates that the behaviors of (linguistic) factors in A are closer to those in B, rather than those in C. Therefore, similar group behaviors were observed to predict whether Korean scholars' L2 writings share universal features of translationese.

IV. Results and Discussion

4.1 Effect Plots: Text-Type Distinction

Employing the method of effect plots, confidence intervals (CIs) of all the eight factors were further observed in an effort to gain a better understanding of how each factor behaved differently in three different sub-corpora. The I-shaped error bars in the effect plot graph above and below the dots indicate the level of 95% confidence intervals (CIs).[2]

The TU Indicator (1) STTR

The factor of the 'STTR' values was tested to spot lexical simplification as a sign

[2] Gravetter and Wallnau (2013) suggest two distinct methods of data normalization. One is to adopt z-scores while the other is to convert (semi-)raw scores into z-scores. This study employed the second method with zero-one scaled by total-sum normalization so as to maintain the characteristics of each linguistic factor.

of translationese among the three different types of English abstracts. Figure 1 shows that the native scholars' NE corpus had the highest value of STTR, and its value decreased as the values went from NE, QE to KE. Both QE and KE corpora had lower STTR values than the NE corpus. In particular, the value of the KE corpus was far lower than that of QE. The CIs of the three groups did not overlap, implying that the three language variants of English texts can be separable using the factor of STTR. The results indicate that the factor STTR can spot translationese in that both the QE and KE texts are far much 'simplified' and 'lexically less diverse' than the NE texts. Overall, it can be deducible that Korean scholars' non-translated L2 texts may hold the properties of translated texts, thus shaping non-nativeness.

The TU Indicator (2) Function Words

The factor of 'Total Function Words' was observed to evaluate the level of lexical density. Figure 2 shows that the CIs partially overlap, suggesting that the overlap could be caused by a change or a higher variability in the datasets. It can be inferred that the QE and KE sub-corpora may hardly bear the properties of translationese. The factor of function words should be further investigated by observing specific types of function words, rather than the total numbers of them across the three different sub-corpora.

Figure 1. STTR

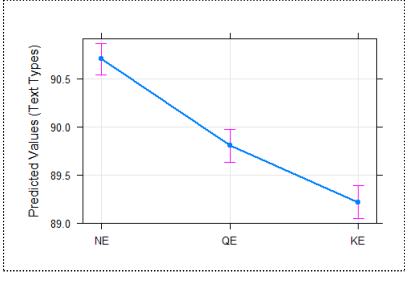
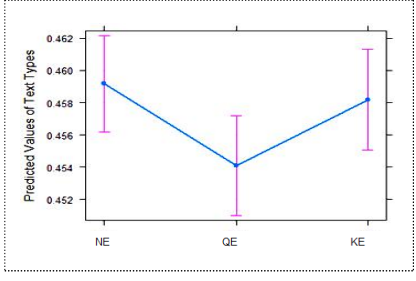


Figure 2. Funct_Total_P



The TU Indicator (3) High-Frequency Words (Top 20)

The factor of 'Top-20 High-Frequency Words' was observed to evaluate the level of lexical richness. Illustrated in Figure 3, the shape of the effect plot came out as starkly opposed to the case with STTR. The factor of Top-20 High-Frequency Words showed the greatest value in the KE texts, and the value decreased as it went from KE to QE, and then to NE in order. Being compatible with the universals of lexical simplification, the effect plot of Top-20 High-Frequency Words supported that Korean scholars' texts might have recycled highly recurring vocabulary repetitively throughout both QE and KE sub-corpora. As the highly recurring vocabulary, especially ranked at top 20, increased across the Korean scholars' texts, the level of lexical richness might have become lower, causing the QE and KE texts to become simplified. Seeing that the CIs of three different sub-corpora did not overlap, the factor of Top-20 High-Frequency Words can also be utilized as a TU indicator to classify text types. Consequently, the results imply that the QE and KE sub-corpora bear the properties of lexical simplification with a lower lexical richness which is not the typicality of native scholars' original texts but the behavior of translated texts.

The TU Indicator (4) Bottom-Frequency Words

The factor of 'Bottom-Frequency Words' was observed to evaluate the indices of lexical simplification. One-time occurring words were paid particular attention. Depicted in Figure 4, the behavior of the factor Bottom-Frequency Words were similar to the case with standard deviations of mean sentence length shown in Figure 8. The factor value of the KE group was lowest, and the NE was the highest among the three sub-corpora. As the CIs of three different sub-corpora did not overlap, it can be concluded that the factor of Bottom-Frequency Words can be also utilized as a TU indicator to classify the types of texts.

Overall, the results indicate that the QE and KE sub-corpora obviously bear the properties of lexical simplification, which is not the typicality of native scholars'

original texts but the behavior of translated texts identifying non-nativeness.

Figure 3. High_Top_20_P

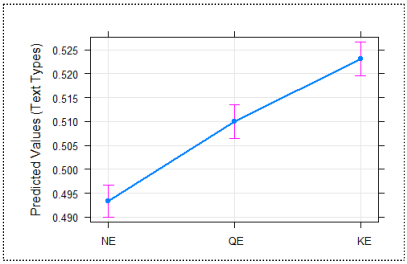
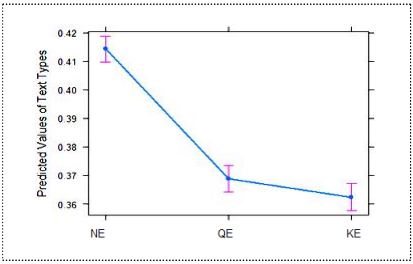


Figure 4. Bottom_P



The TU Indicator (5) Lexical Bundles (Total)

The factor of 'Total Lexical Bundles' was observed to evaluate the indices of lexical normalization. The total number of N-grams were paid particular attention. As Figure 5 illustrates, the behavior of the factor 3-Gram Lexical Bundles seemed to be starkly opposed to the case with Function Words. The figure shows that the CI between the NE and QE sub-corpora do not overlap while the CI between the NE and KE the CIs partially overlap. It can be interpreted that the factor of Lexical Bundles can be possibly considered a TU indicator to make distinctions of the variants of sub-corpora. The factor of Lexical Bundles needs to be further investigated by observing the different combinations of lexical bundles including 4-Grams or 5-Grams, though.

The TU Indicator (6) Lexical Bundles (Top 10)

The factor of 'Top-10 Lexical Bundles' was observed to evaluate the indices of lexical normalization. Highly recurring trigrams ranked up to top 10 were paid particular attention. Depicted in Figure 3, the behavior of the factor 3-Gram Lexical Bundles seemed to be identical to the case with Top-20 High-Frequency Words. The factor value of the KE group was higher than that of the QE corpus, and again the QE was higher than that of the NE corpus.

The results imply that Korean scholars' texts seem to have been lexically simplified due to the behavior of repetitively using high-frequency lexical bundles that have already been pre-fabricated. Likewise, the CIs of the three groups did not overlap, so that the variable of 3-Gram Lexical Bundles could be considered a possible TU indicator to make distinctions of the three variants of sub-corpora. Overall, it can be deducible that Korean scholars' non-translated texts may hold similar linguistic qualities like those in translated texts, which may shape the hallmarks of non-nativeness.

Figure 5. N_Gram_Total_P

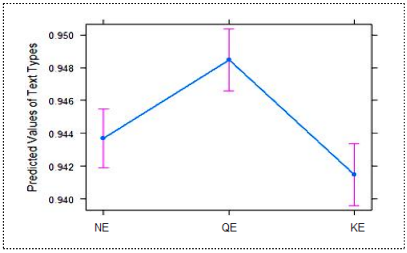
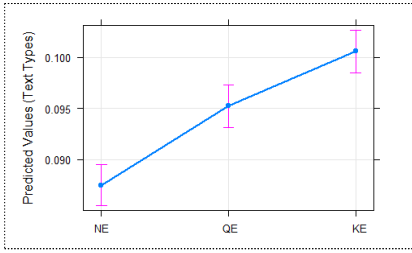


Figure 6. N_Gram_Top_10_P



The TU Indicator (7) Connectives

For syntactic explicitation, the factor of 'Connectives' was tested. As illustrated in Figure 7, the NE group had the lowest value compared to the other two sub-corpora, and the values increased from NE to QE, and then to KE in order. The CIs of the three sub-corpora groups did not overlap as well, meaning that the three groups can be separable according to the different behaviors of each sub-corpus. The results indicate that the variable of *Connectives* could be used as a valid TU indicator to classify text types. Now that cohesive devices such as connectives are frequently used to make sentences more 'explicit' in translated texts, accordingly, it can be deducible that the Korean scholars' writing may share the peculiar linguistic traits that translated texts may hold.

The TU Indicator (8) Mean Sentence Length SD

The factor of 'Mean Sentence Length SD' was observed to spot translationese in syntactic convergence. The effect plot in Figure 8 shows that the KE group had the lowest value compared to the other sub-corpora, but the difference between the KE and QE texts was not as significant as their counterpart. Unlike the previous factors discussed, the CI of the NE texts did not overlap with the remaining factors while the CIs of QE and KE overlapped.

The results indicate that the factor Mean Sentence Length SD can be applied as a valid TU indicator to separate the native group (NE) from the non-native groups (QE and KE), but not to classify the two non-native groups in that the factors in QE and KE might have behaved similarly. It can be thus interpreted that the texts in QE and KE may share the universal attributes of typical translations, which are quite distinctive to the behavior of native writers' original texts.

Figure 7. Conn_Total_P

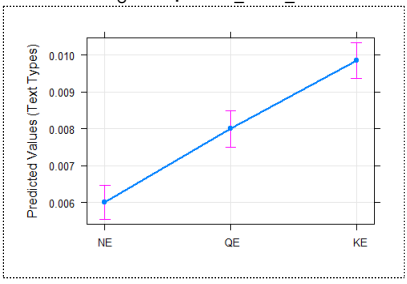
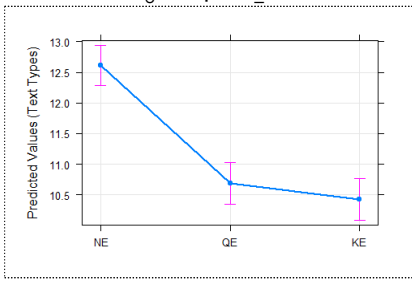


Figure 8. MSL_SD

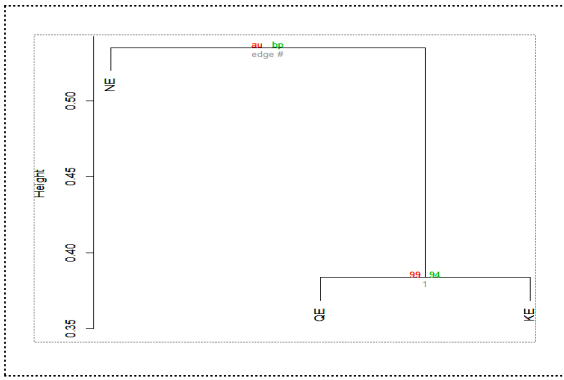


4.2 BP Analysis: Intergroup Homogeneity

The analysis results in Section 4.1 demonstrate that most of the TU indices listed in Table 3 can be utilized as robust and valid indicators to classify the three variants of texts (NE vs. QE vs. KE). As Table 3 indicates the behaviors of each factor but not the overall tendency of each sub-corpora, thus, it can be assumed that there might be a possibility that the TU variables may behave similarly among pairs of groups.

Therefore, this study conducted a BP analysis to investigate specific intergroup homogeneity further. The dendrogram in Figure 9 was drawn based on the behaviors of all the eight linguistic factors listed in Table 3. As observed, the QE and KE corpora were grouped first and represented as {QE, KE}. Then, the NE sub-corpus was merged with them, forming {NE, {QE, KE}}. The results imply that the QE and KE texts can be clustered in higher proximity due to intergroup homogeneity when compared to their native counterparts, representing non-native writers' L2 English texts are significantly different from native writers' L1 English texts.

Figure 9. Intergroup Homogeneity



Attempting to identify the factors that shape 'non-nativeness' in L2 writers' texts, the present study explored how the indicators of translationese behave differently in three different variants of English journal abstracts. In consonance with translation universals postulated by Baker (1993), this study examined the validity of the eight TU indices to spot translationese in non-translated L2 English texts by using the two multi-factorial methods. The GLM analysis proved that the eight TU indices selected were valid, demonstrating that the seven factors behaved distinctively across the three variants of English abstracts (NE vs. QE vs. KE). It can be thus deducible that the TU indices are feasible enough to make a text-type distinction, thereby being

employed as indicators of translationese to discern non-nativeness in non-translated L2 English compositions.

Additionally, the BP analysis drew entirely convincing results, thus consolidating the initial proposition regarding the manifestations of non-nativeness, which is premised to be starkly opposed to nativeness in written production. In the dendrogram in Figure 9, the QE and KE sub-corpora were bound first, and then the NE sub-corpus has joined them, forming {NE, {QE, KE}}. The results further imply that irrespective of the language involved to search resources during the L2 writing process, both L2 English abstracts from Korean articles and L2 English abstracts from English articles might have gone through universal linguistic behaviors, and concurrently these universal properties can be interpreted as shared features of L2 English compositions that might shape non-nativeness. Baker (1993, 1995) claims that translation universals are cognitive phenomena in that they are caused in and by the process of translation. Likewise, Chesterman (2004, 2010) argues that writers' language awareness (either in an L1 or an L2) of the conscious or unconscious cognitive process is pertinent to the direct or indirect translational activity.

Given that the first grouping occurred between the QE and KE, the current findings seem to support the previous propositions reasonably. Even though expert L2 English writers think they 'write' in English during the cognitive process of L2 writing, they may be engaged with the similar mental processing of the 'translating' event in the L2 writing process. Though Korean L2 scholars' abstracts in both groups were placed in two different source-text settings, it can be interpreted that those text writers might have been sharing quite an identical mode of mental translation consciously or unconsciously, which has indeed caused L2 writers' English compositions salient of translationese (e.g., Cook, 1992; Lee, 2017, 2018a). If it had not been for the case, the TU properties of the QE group should have been much closer to those of the NE group.

VI. Conclusion

Driven by the motivation to define what linguistic factors and behaviors shape the identity of non-nativeness, this study questions whether the TU indices are indicative of translationese even in non-translated L2 English compositions produced by highly competent L2 scholars in the English-related disciplines. On a substantial level, the premises on the nature of linguistic behaviors shared between non-translated L2 texts and translated L2 texts were proved to be valid. This study has thus provided evidence that text-type distinction and intergroup homogeneity are universal attributes that exist in non-translated L2 English texts when compared to native writers' L1 English texts. This study has revealed that non-translated L2 English texts may bear the properties of translationese, thereby rendering those L2 texts perceptively distinctive to L1 originals. In turn, these instances of translationese seemed to shape 'non-nativeness' in L2 writers' texts.

Works Cited

Bagheri, Mohammad Sadegh and Ismaeil Fazel. "EFL Learners' Beliefs about Translation and Its Use as a Strategy in Writing." *Reading Matrix* 11.3 (2011): 292-301. Print.

Baker, Mona. *Corpus Linguistics and Translation Studies. Implications and Applications*. In Mona Baker, Gill Francis, and Elena Tognini-Bonelli (eds.), Text and Technology: In Honour of John Sinclair, 233-250. Amsterdam: John Benjamins, 1993. Print.

_____. "Corpora in Translation Studies: An Overview and Some Suggestions for Future Research." *Target* 7.2 (1995): 223-243. Print.

_____. *Corpus-based Translation Studies: The Challenges That Lie Ahead*. In Harold Somers (ed.), Terminology, LSP and Translation, 175-186. Amsterdam: John Benjamins, 1996. Print.

_____. "Patterns of Idiomaticity in Translated vs. Non-translated Text." *Belgian Journal of Linguistics* 21.1 (2007): 11-21. Print.

Baroni, Marco and Silvia Bernardini. "A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text." *Literary and*

Linguistic Computing 21.3 (2006): 259-274. Print.

Blum-Kulka, Shoshana. *Shifts of Cohesion and Coherence in Translation*. In Juliane House and Shoshana Blum-Kulka (eds.), Interlingual and Intercultural Communication: Discourse and Cognition in Translation and Second Language Acquisition studies, 17-35. Tübingen: Gunter Narr, 1986. Print.

Crossley, Scott Andrew and Danielle McNamara. "Shared Features of L2 Writing: Intergroup Homogeneity and Text Classification." *Journal of Second Language Writing* 20.4 (2011): 271-285. Print.

Connor, Ulla. "Linguistic/Rhetorical Measures for International Persuasive Student Writing." *Research in the Teaching of English* (1999): 67-87. Print.

Cook, Vivian. "Evidence for Multicompetence." *Language Learning* 42.4 (1992): 557-591. Print.

Cumming, Alister. "Metalinguistic and Ideational Thinking in Second Language Composing." *Written Communication* 7.4 (1990): 482-511. Print.

Chesterman., Andrew. Beyond the Particular. In Anna Mauranen and Pekka Kujamaki (eds), *Translation Universals: Do They Exist?* Amsterdam: Benjamins, 2004. Print.

Chesterman, Andrew. *Why Study Translation Universals?* In Ritva Hartama-Heinonen and Pirjo Kukkonen (eds.), Kiasm. Acta Translatologica Helsingiensia Vol 1, 38-48. Helsinki: Helsingfors Universitet: Nordica, 2010. Print.

Frawley, William. "Prolegomenon to a Theory of Translation." *Translation: Literary, Linguistic and Philosophical Perspectives* (1984): 159-175. Print.

Gaspari, "Federico and Silva Bernadini. Comparing Non-native and Translated Language. Monolingual Comparable Corpora with a Twist." In *Proceedings of the International Symposium on Using Corpora in Contrastive Translation Studies,* 2008. Print.

Gellerstam, Martin. "Translationese in Swedish Novels Translated from English." *Translation Studies in Scandinavia* (1986): 88-95. Print.

Goh, Gwangyoon and Younghee Cheri Lee. "A Corpus-based Study of Translation Universals in English Translations of Korean Newspaper Texts." *Cross-Cultural Studies* 45 (2016): 109-143. Print.

Goh, Gwangyoon, Younghee Cheri Lee, and Dongyoung Kim. "A Corpus-based Study of Translation Universals in Thesis/Dissertation Abstracts." *Korean Journal of English Language and Linguistics*, 16.4 (2016): 819-849. Print.

Grabe, William. "Notes toward a Theory of Second Language Writing." *On Second Language Writing* (2001): 39-57. Print.

Gravetter, Frederick and Larry Wallnau. *Statistics for Behavioral Sciences*. Belmont, CA: Wadsworth, 2013. Print.

Gries, Stefan and Naoki Otani. "Behavioral Profiles: A Corpus-based Perspective on

Synonymy and Antonymy." *ICAME Journal* 34 (2010): 121-150. Print.

Gries, Stefan. "Behavioral Profiles: A Fine-grained and Quantitative Approach in Corpus-based Lexical Semantics." *The Mental Lexicon* 5.3 (2010a): 323-346. Print.

Hinkel, Eli. *Second Language Writers' Text: Linguistic and Rhetorical Features*. London: Routledge, 2002. Print.

Jarvis, Scott and Aneta Pavlenko. *Crosslinguistic Influence in Language and Cognition.* London: Routledge, 2008. Print.

Kellogg, Ronald. "Effects of Topic Knowledge on the Allocation of Processing Time and Cognitive Effort to Writing Processes." *Memory and Cognition* 15.3 (1987): 256-266. Print.

Laviosa, Sara. "The Corpus-based Approach: A New Paradigm in Translation Studies." *Meta: Translators' Journal* 43.4 (1998a): 474-479. Print.

_____. "Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose." *Meta: Translators' Journal,* 43.4 (1998b): 557-570. Print.

_____. *Corpus-based Translation Studies: Theory, Findings, Applications*, Vol. 17. New York: Rodopi, 2002. Print.

Lee, Younghee Cheri. "The Hallmarks of Expert L2 Writers' Texts Viewed through the Prism of Translation Universals: A Corpus-based Approach to English Research Abstracts of Scholarly Journal Articles." PhD Dissertation. Yonsei University, 2017. Print.

Lee, Younghee Cheri. "The Hallmarks of L2 Writing Viewed through the Prism of Translation Universals." *Linguistic Research 35(Special Edition)* (2018a): 171-205. Print.

_____. "A New Angle on L2 Writers' Texts: A Statistical Approach to Translation Universals." In *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation* (2018b). Print.

Malmkjaer, Kristen. *Language Philosophy and Translation*. In Yves Gambier and Luc van Doorslaer (Eds.), Handbook of Translation Studies (Vol. 3). Amsterdam: John Benjamins, 2012. Print.

Mauranen, Anna. *Universal Tendencies in Translation*. In Margaret Rogers and Gunilla Anderman (Eds.), Incorporating Corpora, The Linguist and The Translator, 32-48. Clevedon: Multilingual Matters, 2007. Print.

McEnery, Tony and and Zhonghua Xiao. *Parallel and Comparable Corpora: What is Happening?*. In Margaret Rogers and Gunilla Anderman (Eds.), Incorporating Corpora. The Linguist and the Translator, 18-31. Clevedon: Multilingual Matters, 2007. Print.

Munday, Jeremy. *Introducing Translation Studies: Theories and Applications.* 2nd ed.. New York, NY: Taylor & Francis, 2008. Print.

Olohan, Maeve. 2004. *Introducing corpora in translation studies*. London: Routledge, 2004. Print.

Øveras, Linn. "In Search of the Third Code: An Investigation of Norms in Literary Translation." *Meta: Translators' Journal* 43.4 (1998): 571-588. Print.

R Core Team. R: *A Language and Environment for Statistical Computing*. R Foundation for Statistical Computing, Vienna. 2018.

Reid, Joy. "A Computer Text Analysis of Four Cohesion Devices in English Discourse by Native and Nonnative Writers." *Journal of Second Language Writing* 1.2 (1992): 79-107. Print.

Reid, Stephen and Gilbert Findlay. "Writer's Workbench Analysis of Holistically Scored Essays." *Computers and Composition* 3.2 (1986): 6-32. Print.

Sasaki, Miyuki. "Toward an Empirical Model of EFL Writing Processes: An Exploratory Study." *Journal of Second Language Writing* 9.3 (2000): 259-291. Print.

Silva, Tony. "Toward an Understanding of the Distinct Nature of L2 Writing: The ESL Research and its Implications." *TESOL Quarterly* 27.4 (1993): 657-677. Print.

Swales, John. *Genre Analysis: English in Academic and Research Settings*. Cambridge: Cambridge University Press, 1990. Print.

Toury, Gideon. *Descriptive Translation Studies and Beyond. Amsterdam.* John Benjamins, 1995. Print.

Uzawa, Kozue. "Second Language Learners' Processes of L1 Writing, L2 Writing, and Translation from L1 into L2." *Journal of Second Language Writing* 5.3 (1996): 271-294. Print.

Wang, Wenyu and Qiufang Wen. "L1 Use in the L2 Composing Process: An Exploratory Study of 16 Chinese EFL Writers." *Journal of Second Language Writing* 11 (2002): 225-246. Print.

Woodall, Billy. "Language-switching: Using the First Language while Writing in a Second Language." *Journal of Second Language Writing* 11.1 (2002): 7-28.

Xiao, Richard and Guangrong Dai. "Lexical and Grammatical Properties of Translational Chinese: Translation Universal Hypotheses Reevaluated from the Chinese Perspective." *Corpus Linguistics and Linguistic Theory* 10 (2014): 11-55. Print.

Lee, Younghee Cheri (Yonsei University/Adjunct Lecturer)

Address: English Education, Graduate Program in Cognitive Science, Yonsei University, 50 Yonsei-ro, Seodaemun-gu, Seoul 03722, Korea

E-mail: cheriberry@yonsei.ac.kr

Received: December 31, 2018 / Revised: January 31, 2019 / Accepted: February 07, 2019