A New Angle on L2 Texts: A Statistical Approach to Translation Universals

Younghee Cheri Lee and Yong-Hun Lee

September 12, 2019

A New Angle on L2 Writers’ Texts: A Statistical Approach to Translation Universals

Younghee Cheri Lee
Yonsei University
50 Yonsei-ro, Seodaemun-gu
Seoul 03722, S. Korea
cheriberry@yonsei.ac.kr

Yong-hun Lee
Chungnam National University
99 Daehak-ro, Yuseong-gu
Daejeon 34134, S. Korea
yleeuiuc@hanmail.net

Abstract

Studies on the second language (L2) writing from the angle of *translation universals* (TU) have substantial empirical research prospects, but as yet only limited literature explains what linguistic factors shape *non-nativeness* in L2 writing. This article claims to prove that key TU indices may predict non-nativeness, more particularly *translationese* in L2 writers’ texts. The ultimate aim is thus to classify texts using the indices of translationese, which will, in turn, signify shared, universal features detectable in L2 texts. To that end, we used a collection of multi-factorial analysis methods to compare native scholars' L1 corpora, respectively with two varieties of non-Anglophone scholars’ non-translated L2 corpora (L1 English vs. Quasi-L2 English vs. L2 English). The results provided evidence that the 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 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.

1. Introduction

Since the last decade facilitated significant advances in the accessibility of large digitized corpora along with robust algorithmic techniques, a corpus-based approach has been manifestly operative in mapping the methodological structure of empirical research queries in L2 text-based studies such as L2 writing and translation studies.

A vast array of research on L2 writers’ texts to date has been mostly dependent on the cognitive behavior models of the second language writing process to lay the theoretical groundwork (see Lee, 2017). In particular, crosslinguistic influences have been the focus of ongoing research endeavors in second language writing (Cumming, 1990). Key research strands include interruptions, transfer, code-switching, positive interplay, and translation strategy (e.g. Bagheri & Fazel, 2011; Connor, 1999; Cumming, 1990; Grabe, 2001; Jarvis & Pavlenko, 2008; Kellog, 1987; Reid & Findlay, 1986; Sasaki, 2000; Swales, 1990; Silva, 1993; Uzawa, 1996; Ventola & Mauranen, 1991; Wang & Wen, 2002; Woodall, 2002).

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

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

Despite the earlier research efforts, however, there remain unmet research needs for further exploration. Although L2 writing studies hinged on the notion of *translation universals* (TU) have substantial empirical research prospects, only limited studies incorporated an interdisciplinary approach to intermingling L2 writing research with descriptive translation studies. Such an approach is worthwhile in the sense that *non-nativeness* can be defined by unveiling universality shared in 'non-translated' L2 texts (e.g. Lee, 2017, 2018). [1]

Driven by the motivation to gain higher insights into such universality, therefore, we aim to discuss the potential linguistic attributes that shape 'non-nativeness' in L2 writers' texts by way of assessing the feasibility of the *translationese* indices (e.g., Lee, 2018). [2] To further develop a 'baseline' notion of non-nativeness, we will measure our selected TU indices in 'non-translated' L2 texts, not in L2 translations in an effort to define the universals of L2 texts through the prism of translationese.

In consonance with Baker's (1993) notion of translation universals to further augment of the previous findings, this study thus claims to prove the following two propositions:

(1) By using the TU indices, translationese will be detectable in non-native L2 writers' non-translated English texts, when compared to native L1 writers' original English texts, so that those indices will be valid to figure text types.

(2) By using the TU indices, intergroup homogeneity (i.e., similarities among different groups) will be measurable, so that the TU indices will classify text groups that share universal traits of L2 English to define linguistic non-nativeness.

Bearing the research aims in mind, we will use the self-constructed CCERA corpora which are based on the three varieties of English texts (L1 English vs. quasi-L2 English vs. L2 English) in two academic disciplines (linguistics and English literature). With reference to previous findings from the first author's two prior research (see Lee, 2017, 2018), we 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.

2. Related Work

2.1 Translation Universals

Beginning in the early 1990s, the advancement of multilingual corpora has invigorated empirical research interests into the *translational language*, in the discipline of translation studies. 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 own typical linguistic characteristics.

Such scholarly attention to the peculiar traits of a translational language has triggered a further development of a robust conceptual framework. The proposal regarding translation universals 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 typical 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 investigating the interrelationship between source texts and their target texts by using parallel corpora, as well as the linguistic relation

[1] Non-nativeness is a newly coined term by the first author, to collectively refer to any linguistic properties or behaviors apparent in L2 writers' 'non-translated' texts, which are perceptively different from those of L1 writers' original texts (Lee, 2018).

[2] 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 further extended its original definition to mean any 'linguistic fingerprints' or 'awkwardness' that L2 writers' target language leaves on their 'non-translated' L2 written production.

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 translation universals 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).[3] It was also widely accepted that translated versions may '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 Translation Universals

In contemporary descriptive translation studies, translation scholars have been constantly engaged in conducting empirical studies to discover what factors and indices can represent translational attributes. Most potential indicators involve the simplification, normalization, explicitation, and convergence hypotheses. Translation scholars are constantly engaged in continuous research to find language indices that represent transitive attributes.

Simplification is the tendency to consciously or unconsciously make target texts simpler lexically, syntactically and/or stylistically by using simpler translational language to increase readability of target texts. (Baker, 1996). Contrary ideas such as lexical diversity, lexical density, lexical richness, structural sophistication, and stylistic complexity are all associated with simplification. Some good parameters of simplification involve STTR 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; Malmkjæ r, 2012).

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

Explicitation is the most investigated feature among the 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.g. placing a clause in the position of a subject or an object of a sentence).

Comparatively less scholarly attention was paid to research into convergence (also called leveling-out) compared to the other features of universals (Laviosa, 2002). 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 (i.e. dispersion) of lexical variety, lexical density, type/token ratio, readability indices, and mean sentence length are the most studied indicators of convergence (c.f. Pym, 2008).

3. Methods

3.1 Corpus Construction

Comparable monolingual corpora were constructed with the specific aim of observing recurrent linguistic features that might render Korean scholars' L2 English compositions perceptively different from those of native scholars' L1 English compositions. The English abstracts data for the Comparable Corpora of English Research Abstracts of Scholarly Journal Articles (CCERA) were taken from acclaimed scholarly journal articles in the two English-related disciplines of linguistics and English literature. The CCERA was

[3] There have been diverse views raised to resolve the controversy over the notion of translation universals such as 'translationese' (Gellerstam, 1986), 'the third code' (Frawley, 1984), 'laws' (Toury, 1995), 'core patterns' (Laviosa, 1998a), and many more.

designed to be composed of three variants of texts and compiled to have balanced size, time span, genre representation, and search terms using simple random sampling so as to make equitable comparisons. [4] The three sub-corpora include native scholars' L1 English abstracts (NE), Korean scholars' L2 English abstracts of which articles were written in English (QE, meaning quasi-L2 English), and finally Korean scholars' L2 English abstracts of which articles were produced in L1 Korean (KE). 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 we use is L2 English 'compositions', not L2 translations. The reason is that the ultimate goal of this study is to see if translationese appears in expert non-native writers' English compositions. The scale of the CCERA is mapped out in Table 1.

Sub-Corpus	Domain	Abstract (#)	Token (#)	Type (#)
NE ***N**ative L1 **E**nglish* *(L1 English abstracts with L1 English articles)*	Linguistics	600	105,535	7,594
	Literature	530	106,851	9,743
	Sub Total	**1,130**	**212,386**	**17,337**
QE ***Q**uasi L2 **E**nglish* *(L2 English abstracts with L2 English articles)*	Linguistics	605	106,195	6,139
	Literature	440	107,869	8,538
	Sub Total	**1,045**	**214,064**	**14,677**
KE ***K**orean L2 **E**nglish* *(L2 English abstracts with Korean articles)*	Linguistics	603	106,545	5,898
	Literature	435	105,769	9,086
	Sub Total	**1,038**	**212,314**	**14,984**
	Total	3,213	638,764	46,998

Table 1: The Scale of the CCERA

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, 2018). 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 WordSmith Tools 7.0 and AntConc 3.4.4w, and all the statistical analyses were performed using R version 3.5.0. Table 2 and the information below briefly shows sets of hypotheses for each variable encoded.

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)

Table 2: Encoded Variables: Key TU Indices

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.

FUNCTION WORDS (TOTAL): LEXICAL SIMPLIFICATION

The QE and KE texts will have higher total values of function words than native scholars' NE texts.

HIGH-FREQUENCY WORDS (TOP 20): LEXICAL SIMPLIFICATION

The QE and KE corpora will have higher values of top 20 high-frequency words than the NE corpus.

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.

LEXICAL BUNDLES (TOTAL): LEXICAL NORMALIZATION

The total proportions of recurring lexical bundles will be higher in QE and KE than in NE.

LEXICAL BUNDLES (TOP 10): LEXICAL NORMALIZATION

The QE and KE corpora will hold a greater amount of top 10 lexical bundles than the NE corpus.

CONNECTIVES: SYNTACTIC EXPLICITATION

The ratio of connectives will be higher in the QE and KE corpora than in the NE corpus.

MEAN SENTENCE LENGTH SD: SYNTACTIC CONVERGENCE

The standard deviations of mean sentence length will be lower in both QE and KE than NE texts.

[4] The encoded corpus data for this study were drawn from the first author's two prior research projects. To carry out her doctoral dissertation project, she constructed the initial version of the CCERA and recently updated for her 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 (2018).

3.3 GLM Procedures

As a linear (regression) method, a Generalized Linear Model (GLM) can apply to the case either a dependent variable is not based on a ratio, or value does not fit a normal distribution. As the dependent variable TextType was categorical in this study, a GLM model was applied to our data so as to evaluate (linguistic) factors that play vital roles in identifying three variants of English texts: NE vs. QE vs. KE. For the implementation of a GLM model, an initial model was constructed first. The TextType was set as a dependent variable while the remaining factors became an 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. 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 lines in each plot graph.

Regarding the interpretation of the confidence intervals (CIs) given with the I-shaped lines, if the CI of one group does not overlap with that of the other group, the factor is statistically significant. It means that the factor behaves differently in the two groups. Conversely, if two CIs overlap, it indicates that the factor behaves similarly in the two groups.

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, we observed homogeneous group behaviors to predict whether Korean scholars' L2 writings share universal features of translationese.

4. Results

4.1 GLM Output

For multinomial regression analysis, the initial model was set up as in Table 3, followed by model selection procedures to select the most optimal model. The final model obtained was identical to the initial model below, and thus all the eight main factors survived in the final model.

TextType~STTR+Funct_Total_P+High_Top_20_P+Bottom_P+MSL_SD+Conn_Total_P+N_Gram_Total_P+N_Gram_Top_10_P

Table 3: Initial Model

Utilizing the final model, all the eight main factors were statistically analyzed. Table 4 outlines the 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 (NE vs. QE vs. KE) for the CCERA.

Variables	***χ*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	***

Table 4: GLM Output

4.2 Effect Plots: Text-Type Distinction[5]

Employing the method of effect plots, we further observed confidence intervals (CIs) of all the eight factors so that we can gain a better understanding of how each factor behaved differently in three different sub-corpora. It would be desirable to cover all measured variables, but due to space constraints, in the following section, we will only discuss five of the significant factors.

[5] 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.

STTR (Lexical Simplification)

As an indicator of lexical simplification, the factor of the STTR values was tested to evaluate the lexical diversity of three different types of English abstracts. Lexical simplification and lexical diversity and universals seem to be contradictory but related concepts. Shown in Figure 1, the I-shaped line in the effect plot graph above and below the dots indicates the level of 95% confidence intervals (CIs).

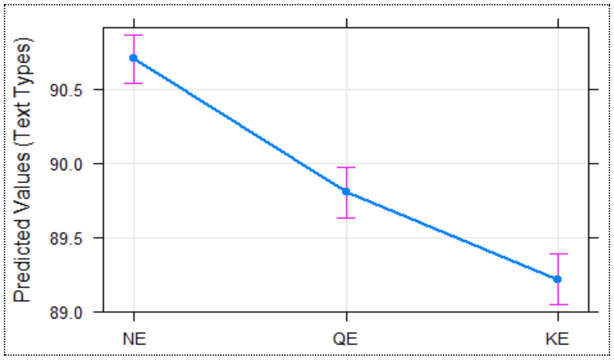

Figure 1: STTR

As shown in Figure 1, the native scholars' NE corpus had the highest value of STTR across the CCERA, and its value decreased as we went from NE, QE to KE. Both QE and KE corpora had lower STTR values than the NE corpus. In particular, the value of STTR in 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. In turn, it demonstrates that the factor STTR can be utilized as a valid TU indicator of lexical simplification (NE vs. QE vs. KE). Overall, the results indicate that both the QE and KE texts are far much 'simplified' than the NE texts, thus proving that Korean scholars' abstracts may share universal properties typical of translated texts.

High-Frequency Words (Lexical Simplification)

The factor of Top-20 High-Frequency Words was observed to evaluate the level of lexical richness. Illustrated in Figure 2, 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 this 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 be also utilized as a TU indicator to classify the types of texts. Consequently, the results demonstrate 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.

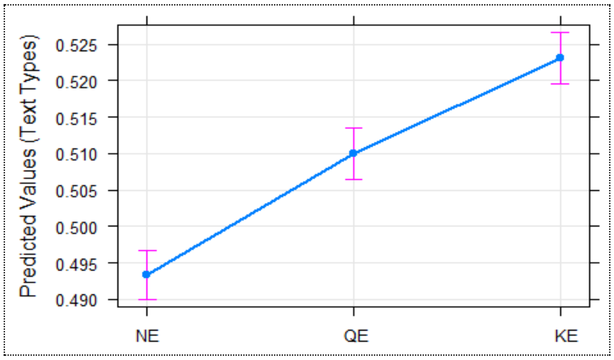

Figure 2: High_Top_20_P

Lexical Bundles (Lexical Normalization)

The factor of Lexical Bundles (*n*-grams) 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.

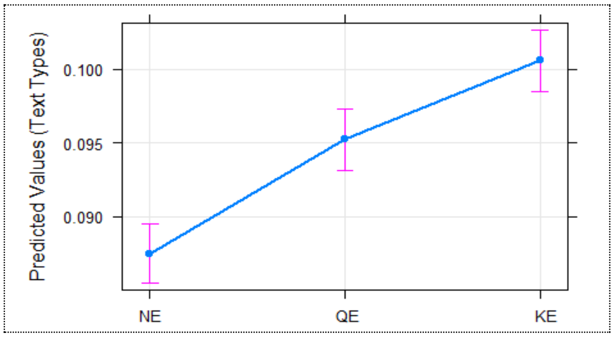

Figure 3: N_Gram_Top_10_P

The results imply that Korean scholars' texts seem to have been lexically simplified due to the behavior of repetitively using the lexical bundles of high-frequency 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' KE and QE sub-corpora may hold the similar linguistic qualities like those in translated texts, which may signify the instances of translationese.

CONNECTIVES (SYNTACTIC EXPLICITATION)

For syntactic explicitation, the factor of CONNECTIVES was tested across the CCERA. As illustrated in Figure 4, 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, which means 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 the three types of texts. 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’ sub-corpora may share the peculiar linguistic traits that translated texts may hold.

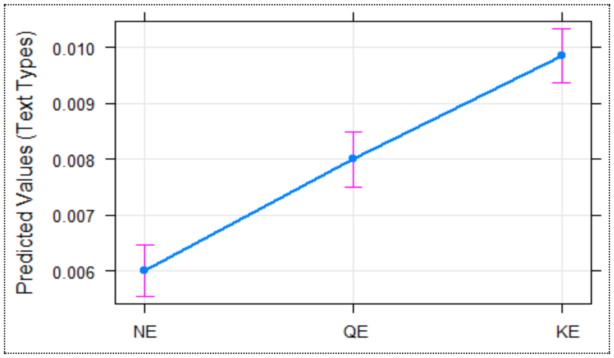

Figure 4: CONN_TOTAL_P

MEAN SENTENCE LENGTH SD (SYNTACTIC CONVERGENCE)

By using the effect plot graph, the factor of MEAN SENTENCE LENGTH SD was observed to evaluate the indicator of syntactic convergence. In Figure 5, the plot of MEAN SENTENCE LENGTH SD proved that the KE group had the lowest value compared to the other two sub-corpora QE and NE, but the difference between the KE and QE texts was not significant as the difference between the KE and NE sub-corpora. The value thus increased in the order of KE, QE, and then NE. Unlike the previous factors discussed, whereas the CI of the NE texts did not overlap with the CIs of the remaining factors, 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.

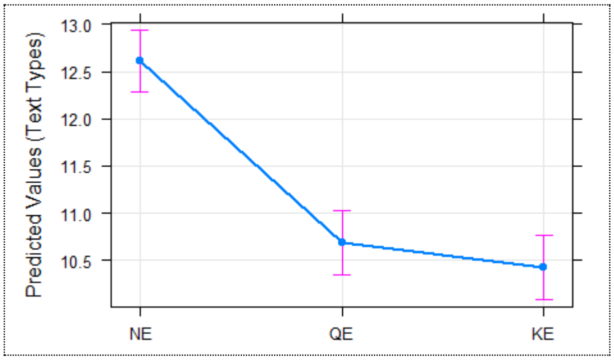

Figure 5: MSL_SD

4.3 BP Analysis: Intergroup Homogeneity

The analysis results in Section 4.2 demonstrate that all the eight TU indices listed in Table 2 can be utilized as robust and valid indicators to classify the three variants of texts (NE vs. QE vs. KE). As Table 2 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, we conducted a BP analysis to further investigate specific intergroup homogeneity. The dendrogram in Figure 6 was drawn based on the behaviors of all the eight linguistic factors in Table 2. 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.

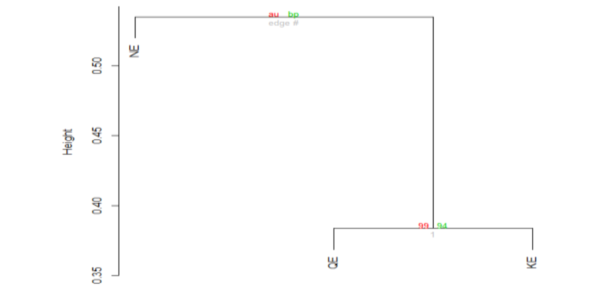

Figure 6: Intergroup Homogeneity

5. Discussion

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), we examined the validity of the eight TU indices to predict simplification, normalization, explicitation, and convergence 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 all the factors behaved distinctively across the three variants of English abstracts (NE vs. QE vs. KE). It can be deducible that the eight TU indices are feasible enough to make a text-type distinction, thereby proposing a high validity that the indices can be employed to discern translationese in non-translated L2 English compositions.

Additionally, the BP analysis drew entirely convincing results, thus consolidating our 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 6, 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 type of 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 traits, and concurrently these universal properties can be interpreted as universal 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 sub-corpora, the current findings seem to support the previous propositions reasonably. Even though expert L2 English writers may think they 'write' in English during the cognitive process of L2 writing, they may be engaged with the similar mental processing of 'translating' event during the task of L2 writing. 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, 2018). 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.

6. 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. Overall, this study has revealed that non-translated L2 English texts bear the properties of translationese, which renders those L2 English texts perceptively distinctive to L1 English texts. In turn, these instances of translationese seemed to shape 'non-nativeness' in L2 writers' texts.

References

Alister Cumming. 1990. Metalinguistic and Ideational Thinking in Second Language Composing. Written Communication 7(4): 482-511.

Marco Baroni and Silvia Bernardini. 2006. A New Approach to the Study of Translationese: Machine-learning the Difference between Original and Translated Text. Literary and Linguistic Computing 21(3): 259-274.

Mona Baker. 1993. 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.

Mona Baker. 1995. Corpora in Translation Studies: An Overview and Some Suggestions for Future Research. Target 7(2): 223-243.

Shoshana Blum-Kulka. 1986. 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.

Andrew Chesterman. 2004. Beyond the Particular. In Anna Mauranen and Pekka Kujamaki (eds), Translation Universals: Do They Exist?, 33-49. Amsterdam: Benjamins.

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

Ulla Connor. 1999. Linguistic/Rhetorical Measures for International Persuasive Student Writing. Research in the Teaching of English, 67-87.

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

Federico Gaspari and Silva Bernadini. 2008. 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.

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

Gwangyoon Goh, Younghee Cheri Lee, and Dongyoung Kim. 2016. A Corpus-based Study of Translation Universals in Thesis/Dissertation Abstracts. Korean Journal of English Language and Linguistics, 16(4): 819-849.

Martin Gellerstam. 1986. Translationese in Swedish Novels Translated from English. Translation Studies in Scandinavia, 88-95.

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

Stefan Gries and Naoki Otani. 2010. Behavioral Profiles: A Corpus-based Perspective on Synonymy and Antonymy. ICAME Journal 34:121-150.

Stefan Gries. 2010a. Behavioral Profiles: A Fine-grained and Quantitative Approach in Corpus-based Lexical Semantics. The Mental Lexicon 5(3):323-346.

Stefan Gries. 2010b. Behavioral Profiles 1.01: A Program for R 2.7.1 and Higher.

William Grabe. 2001. Notes toward a Theory of Second Language Writing. On Second Language Writing, 39-57.

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

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

Ronald Kellogg. 1987. Effects of Topic Knowledge on the Allocation of Processing Time and Cognitive Effort to Writing Processes. Memory and Cognition 15(3): 256-266.

Sara Laviosa. 1998a. The Corpus-based Approach: A New Paradigm in Translation Studies. Meta: Translators' Journal 43(4): 474-479.

Sara Laviosa. 1998b. Core Patterns of Lexical Use in a Comparable Corpus of English Narrative Prose. Meta: Translators' Journal, 43(4): 557-570.

Sara Laviosa. 2002. Corpus-based Translation Studies: Theory, Findings, Applications, Vol. 17. New York: Rodopi.

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

Younghee Cheri Lee. 2017. 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.

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

Baker, Mona. 2007. Patterns of Ｉdiomaticity in Translated vs. Non-translated Text. Belgian Journal of Linguistics 21(1): 11-21.

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

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

Mohammad Sadegh Bagheri and Ismaeil Fazel. 2011. EFL Learners' Beliefs about Translation and Its Use as a Strategy in Writing. Reading Matrix: An International Online Journal 11(3): 292-301.

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

Tony McEnery and and Zhonghua Xiao. 2007. 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.

Linn Øverås. 1998. In Search of the Third Code: An Investigation of Norms in Literary Translation, Meta: Translators' Journal 43(4): 571-588.

Maeve Olohan. 2004. Introducing Corpora in Translation Studies. London: Routledge.

Anthony Pym. 2008. On Toury's Laws of How Translators Translate. In Anthony Pym, Miriam Shlesinger, and Daniel Simeoni. (Eds.), Beyond Descriptive Translation Studies: Investigations in Homage to Gideon Toury, 311-328. Amsterdam: John Benjamins.

Joy Reid. 1992. A Computer Text Analysis of Four Cohesion Devices in English Discourse by Native and Nonnative Writers. Journal of Second Language Writing 1(2): 79-107.

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

Stephen Reid and Gilbert Findlay. 1986. Writer's Workbench Analysis of Holistically Scored Essays. Computers and Composition 3(2): 6-32.

Johan Swales. 1990. Genre Analysis: English in Academic and Research Settings. Cambridge: Cambridge University Press.

Miyuki Sasaki. 2000. Toward an Empirical Model of EFL Writing Processes: An Exploratory Study. Journal of Second Language Writing 9(3): 259-291.

Scott Andrew Crossley and Danielle McNamara. 2011. Shared Features of L2 Writing: Intergroup Homogeneity and Text Classification. Journal of Second Language Writing 20(4): 271–285.

Tony Silva. 1993. Toward an Understanding of the Distinct Nature of L2 Writing: The ESL Research and its Implications. TESOL Quarterly 27(4): 657-677.

Gideon Toury. 1995. Descriptive Translation Studies and Beyond. Amsterdam: John Benjamins.

Kozue Uzawa. 1996. Second Language Learners' Processes of L1 Writing, L2 Writing, and Translation from L1 into L2. Journal of Second Language Writing 5(3): 271-294.

Vivian Cook. 1992. Evidence for Multicompetence. Language Learning 42(4): 557-591.

Billy Woodall. 2002. Language-switching: Using the First Language while Writing in a Second Language. Journal of Second Language Writing 11(1): 7-28.

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

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