Korean Journal of English Language and Linguistics
Vol. 19, No. 3, Fall 2019, 452-474
10.15738/kjell.19.3.201909.452

A Comparative Analysis of Koreans' English Writings and Google Translations Using Coh-Metrix 3.0

Joonkoo Kim (Chung-Ang University)
Kyuyun Lim (University of British Columbia)*

Kim, Joonkoo and Lim, Kyuyun. 2019. A comparative analysis of Koreans' English writings and Google Translations using Coh-Metrix 3.0. *Korean Journal of English Language and Linguistics* 19-3, 452-474. The purpose of the present study is to identify the corpus-based differences between Koreans' English writings and their corresponding Google translations. For this purpose, the present study utilized Coh-Metrix 3.0 and conducted comparative analyses on two types of writings in terms of 12 benchmarks of text analysis. Coh-Metrix 3.0 provided numeric values for the following selected categories of text analysis: (a) basic counts (i.e., DESSC, DESWC, and DESSL), (b) lexical aspects (i.e., WRDFRQc and LDTTRc), (c) readability (i.e., RDFRE and RDFKGL) (d) syntactic complexity (i.e., SYNLE, SYNNP, and SYNSTRUTa), and (e) cohesion (i.e., CRAFAOa and LSASS1). Each output for 5 categories computed by Coh-Metrix 3.0 was then statistically processed in order to find statistically significant differences. The quantitative findings, given the small sample size associated with lower statistical power and non-normality of some data sets, were interpreted together with results from a robust technique of bootstrapped independent *t*-tests since the employment of bootstrapping has been empirically justified in the field of applied linguistics (Plonsky 2013, 2014). The overall findings indicated that Google translations tend to produce significantly more words before main verbs and longer sentences compared to human writings. Furthermore, it was also found that Google translations were significantly less readable, but more cohesive. However, there were no significant differences observed in lexical aspects.

Keywords: Coh-Metrix 3.0, Koreans' English writings Google translations, human writings, Google Translate

1. Introduction

The use of machine translation has been dramatically increased with the rapid advancement of automatic machine translation technologies. Among them, Google

* Joonkoo Kim is the first author, and Kyuyun Lim is an MA student in language and literacy education.

Translate has been predominantly used due to its accessibility, speed, and relatively accurate results (Hampshire and Porta 2010). Since Google Translate released the engine of Neural Machine Translation (NMT) in 2016, it has been able to provide a more sophisticated and precise translation service than its previous engine, Statistical−Based Translation (SMT). As the SMT system statistically matches the most similar words or phrases from a corpus of source texts, it often fails to translate polysemous words into the proper equivalence due to lack of its understanding of the contexts. On the contrary, the NMT system does not break up a sentence into words or phrases but recognizes the whole sentence and reflects the contextual meaning into translation (Ramati and Pinchevski 2017).

With this advancement, Google Translate has been an integral part of the multilingual world. It has supported translations between 103 languages, and more than 500 million people have used the service (Turovsky 2016). Google Translate has transformed a way of communication by breaking language barriers; it not only changes the daily lives of people but also supports linguistic minority communities such as immigrant newcomers or refugees via information access and advocacy (Kapadya 2018, Rodreguez−Castro, Salas and Benson 2018). Furthermore, its potential employment in second language acquisition and education has gained significant attention as it can be a useful resource for second language writings (Groves and Mundt 2015).

The advancement of Google Translate brings light to a question of whether Google Translate can replace human translators or writers. Regarding the question, there have been attempts to prove the quality of machine translated texts in comparison with human translations or writings. The notable incident of comparing human translations and machine translations took place in South Korea in 2017 (Kruolek 2017). The International Association of Interpretation and the Sejong University held 'Human vs. Artificial Intelligence Translation Showdown', and four professional translators and three popular machine translation programs including Google Translate, Systran, and Papago translated four pieces of texts, two from Korean to English and the other two from English to Korean. The quality of writings was gauged based on three characteristics: accuracy, language expressions, and logical organization. The result was an absolute win of human translators, who scored 49 points as opposed to machine translators, which scored 28 points. Although the event showed that machine translations were far from reaching the level of human writings, it drew public attention to the potentials and distinctive features of machine translations such as its promptness and relatively accurate translation of texts in the non−fiction genre.

The previous research has explored comparison focusing on linguistic accuracy of machine−translated texts (Correa 2014, Ducar and Schocket 2018, Grove and Mundt 2015). Understanding the varied aspects of Google translations beyond the linguistic features is significant to advance the field of translation studies as it uncovers the potentials and limitations of technology in terms of replacing or complementing human efforts in translation and interpretation of languages. In addition, the comparison with human writings will add insights to comprehending the way in which machines and humans work similarly or dissimilarly in translating languages. However, it still lacks empirical research conducive to apprehending various dimensions of textual features of machine translated texts. Therefore, this study aims to provide a corpus−based comparison between Google translated texts and human texts by analyzing lexical, syntactic, and discourse features.

2. Literature Review

2.1 Google Translate

This study utilizes Google Translate as the target machine translation service when comparing to human writings because Google Translate outperforms other freely available machine translators (Hampshire and Porta 2010). Google Translate is a web−based machine translator which translates between varieties of languages. Among several available Machine Translation tools, Google Translate gains its popularity due to its speed (Google 2012), cost (Sheppard 2011), adaptability (Duncar and Schocekt 2018), and most of all, its collection of large amounts of data sets from target texts and source texts (Bellos 2012).

Due to its low reliability and accuracy at its initial developmental stage, Google Translate has been perceived as a less reliable tool compared to translations by professional human translators. Previous studies pointed out the limitations of Google Translate in that it produced a relatively large number of syntactical errors in translating grammatically different languages (Kaltenbacher 2000). From this, it has been determined that Google Translate results in different translation quality depending on language (Aiken and Balan 2011, Costa−jussa, Farrus and Pons 2012). It has been emphasized that Google translations are far from being able to produce error−free texts (Grove and Mundt 2015) and it requires post−editing by human translators

(Kirchoff, Turner, Axelrod and Saavedra 2011).

Google Translate has recently achieved technological advances in its translating quality since it adopted artificial neural networks in 2016 (Lewis-Kraus 2016). Google's implementation of AI algorithms resulted in a 60% reduction in translation errors in contrast to the previous technology, a statistical machine translation model, according to Castelvecchi (2016). Ducar and Schocket (2018) highlighted that the machine translations have overcome the earlier limitations including translation of proper nouns, unnatural writing, a literal translation of polysemous words, untranslated misspelled words, and discursive inaccuracy. Although Google Translate achieved a certain level of accuracy, the challenges still lie in improving the translation quality of several linguistic, discourse and pragmatic aspects of translated texts such as translating rarely used idioms or phrasal verbs into proper equivalence, recognizing register, and misinterpreting metaphorical meanings. It implies that its capability to produce quality text is still far from human writers' abilities. Therefore, in order to explore the potential application of Google Translate to academic or professional fields such as supplementing second language learning or replacing human translators, it is required to closely investigate the capability of the machine translation in producing readable texts in comparison with human writings. In this study, Coh-Metrix (http://cohmetrix.com) is utilized to compare the differences between human writings and Google translations in terms of linguistic and lexical features, readability and cohesion.

2.2 Coh-Metrix as a Tool for Textual Analysis and Comparison

Coh-Metrix (http://www.cohmetrix.com/) is an automated language analysis tool that can analyze a wide range of measurements of discourse or texts (McNamara, Graesser, McCarthy and Cai 2014). It aims to acquire a deeper understanding of discourse and underlying psychological mechanisms using the multilevel framework, which include "the surface code, the explicit textbase, the situation mode (sometimes called the mental model), and the discourse genre and rhetorical structure, and the pragmatic communication level" (Grasser and McNamara 2011, p. 1).

As Coh-Metrix enables to gauge deeper levels of textual features and their relationships as well as surface levels of characteristics such as lexical diversity, syntactic complexity, and difficulty of written texts using readability formulas, it has been widely utilized to analyze and compare the written and oral texts in the field of linguistics and applied linguistics. For instance, Coh-Metrix has been employed in previous research which investigates the language varieties between native-speaker produced texts and non-native-speaker produced ones (Lee 2018, Ye 2013), distinctive characteristics of texts written in a specific discipline (Zhang 2015), the text difficulty and readability of ESL/EFL textbooks and assessments (Ahn and Ma 2015, Kim 2014), and comparison of the cohesion level of the original or simplified version of literature (Sohn 2018). Application of Coh-Metrix to examine the textual, linguistic, and discourse features of diverse sources have contributed to reading and writing pedagogy, language assessment, textbook evaluations, and discourse analysis. In this vein, the present study employs Coh-Metrix to compare the linguistic, syntactic and textual features of Google translated texts and human created texts.

2.3 Comparison of Machine Translations and Human Writings

The emerging body of literature has shown that Google Translate is achieving a relatively accurate quality of translations. Correa (2014) points out that online translations are skillful at verb conjugation, basic agreement, and translating common idioms. Particularly, with respect to spelling, it is able to correct misspelled words as well as produce correct spelling. However, Google Translate still “falls short of matching human production” (Ducar and Schocket 2018, p. 783). It formulates errors in meaning and punctuations, has difficulties with finding a pragmatic equivalence to idioms, and translates the formal register of discourse appropriately. Because of the long-lasting perceptions that Google Translate is less reliable than human writings by bilinguals or professional translators, there have been empirical studies comparing Google translations with human translations and writings.

Grove and Mundt (2015) focus on the error analysis of translated texts from Malay and Chinese to English to examine the linguistic accuracy using the taxonomy of errors introduced by Ferris, Liu, Sinha and Senna (2013). The findings reveal that Google Translate can produce clear and formal English texts composed of grammatically correct sentences written in an academic style. To be specific, the number of errors were less than the number of sentences, indicating that some translated sentences do not contain errors. Most of the errors were associated with sentence structure and word choice. The study marked the results of Google translations using the IELTS test scoring rubrics, and their grammatical accuracy was scored as 6.0, the equivalent to the level of acceptance to tertiary institutions. It indicates that although translated writings are not at the level of native speakers, it is close to the level of acceptable academic competence. Ahrenberg (2017) highlights that machine translation has similar length, information flow, and structure to the source; however, it exhibits restricted output that requires about three edits per sentence. The accuracy of machine translated outcomes is not considered sufficient for the publication purpose as compared to human translations. Overall, the previous studies conclude that machine translators have not reached the level of human translations in terms of linguistic accuracy.

As well as accuracy, Turner, Bergman, Brownstein, Cole and Kirchhoff (2014) compare human translations and machine translations of public health materials using a criteria of time, cost and quality. Although the quality of human translations is more preferred and guaranteed, considering the fact that machine translated materials require much less cost and time, there is a great need for machine translation. The research shows that machine translation will be of great benefit when it comes to the efficiency of the service.

While there have been attempts to evaluate the quality of Google translations regarding its efficiency and accuracy focusing on lexical and sentence−level structure, it has been little explored how machine translations and human writings are different in discourse and pragmatic level. Concerning the issues of formality, Google Translate has been criticized that it cannot properly translate languages with multiple levels of formality such as Arabic. The previous research has also pointed out the discursive inaccuracies of Google Translate (Correa 2014, Ducar and Schocket 2018). As investigating such issues, Li, Graesser and Cai (2014) show the contrasting results to the previous studies given that Google translations are highly correlated with human translations and original texts in terms of formality and cohesion. By using Coh−Metrix and LIWC, the study uncovers that Google Translate demonstrates good performance in producing formal and cohesive texts nearly at the level of human translations, which implies that Google translated texts are readable and decipherable in general. While the study investigates the readability of the Google translated texts, it is limited only to analysis of formality and cohesion. In addition, the data only includes texts written by an author, Mao; therefore, the broader range of documents written by different authors need to be further explored.

In this regard, the present study pays attention to the corpus−based comparison of Google translated texts with human writings. Although the preceding studies have investigated the quality of Google translations in terms of linguistic precision and productivity, there have been hitherto few studies that compare the wide range of features of Google translations based on the corpus. As there have been efforts to explore the potential employment of machine translations in the field of translation studies to gain benefits of cost and time free tools, it will progress the development of machine translators by uncovering similarities and differences in lexical diversity, syntactic complexity, readability and discourse between Google translations and human writings. Therefore, this study aims to analyze the multi-level comparison of Google translations and human writings by addressing the following research questions:

(1) In comparison with Koreans' English writings, what are the lexical and syntactic features of Google translated texts?
(2) In comparison with Koreans' English writings, how readable and cohesive are Google translated texts?

3. Methods

3.1 Instrument

This study utilized some of the Coh-Metrix indices selected in light of the previous studies (e.g., Jeon and Lim 2009, Kim 2018) that relied upon them to analyze and compare texts from the perspective of corpus-linguistics.

Coh-Metrix aims to measure the cohesion and coherence of written and spoken discourse on multiple levels by providing scores on linguistic characteristics. Coh-Metrix uses the multilevel framework of discourse comprehension including the surface code, the textbase, the situation model, and the genre (Grasser and McNamara 2011). Therefore, Coh-Metrix is able to gauge deeper levels of textual features and their relationships as well as surface levels of characteristics such as lexical diversity, syntactic complexity, and difficulty of written texts using readability formulas. As Coh-Metrix was developed to identify cohesion, coherence and language processing at all levels, it provides indices to assess such characteristics. In this context, cohesion is defined as "characteristics of the explicit text that play some role in helping the readers mentally connect ideas in the text" (Grasser and McNamara 2011, p. 379).

Coh-Metrix includes 108 indices categorized into 11 groups: (1) descriptive statistics (2) text easability (3) referential cohesion (4) latent semantic analysis

(LSA) (5) lexical diversity (6) connectives (7) situation model (8) syntactic complexity (9) syntactic pattern density (10) word Information and (11) readability (McNamara, Graesser, McCarthy and Cai 2014). Out of 108 variables, 12 indices were chosen for this study drawing on previous studies which used Coh−Metrix in order to measure linguistic features and writing quality pertaining to second language writing, textbook analysis or academic writings (Aryadoust 2016, Jeon and Lim 2009, Kim 2018, Lee 2015, Solnyshkina, Harkova and Kiselnikov 2014) as this study aims to provide textual similarities and differences between Google and human translations in terms of usability of Google Translate in the field of academic writing. These indices offer information on basic counts, word frequency, readability, syntactic complexity, and cohesion. Each of the categories will be elaborated in the following section (McNamara, Graesser, McCarthy and Cai 2014).

(1) Basic Counts

Basic counts include descriptive statistics of written texts such as the number of sentences (DESSC), the total number of words in the text (DESWC) and the mean number of words in each sentence (DESSL) in order to check the Coh−Metrix output and interpret data patterns. This information offers the surface−level comparison between Google translations and human writings and predicts the difficulty and complexity of the input texts.

(2) Lexical Aspects (Word frequency and lexical diversity)

Lexical aspects of the text are calculated by two indices in this study; the average word frequency for content words (WRDFRQc) and lexical diversity (LLTTRc). Lexical diversity is measured by the type−token ratio for content words (LLTTRc). Type−token ratio (TTR) is the number of unique words (types) divided by the number of entire content words (tokens).

(3) Readability

Readability is analyzed based on two formulas; Flesch Reading Ease (RDFRE) and Flesch Kincaid Grade Level (RDFKGL). The score of RDFRE ranges from 0 to 100, and a higher score indicates easier reading. Flesch Kincaid Grade Level is measured based on Reading Ease Score converted to a U.S. grade−school level with a higher score indicating harder reading. The scales of readability provide information about whether the text has the appropriate difficulty for the target readers, and it is a significant indicator of Google translated texts' abilities to produce readable texts for target audiences.

(4) Syntactic complexity

Syntactic complexity is measured by indices such as words before the main verb (SYNLE), the mean number of modifiers per noun−phrase (SYNNP), and the syntactic structure similarity between all adjacent sentences (SYNSTRUTa), reflecting the characteristics of the syntactic difference between simple sentences and complex sentences. As simple and short sentences require less mental processing of readers, more complex sentences will produce more difficult texts.

(5) Cohesion

Coh−Metrix assesses two kinds of cohesion: referential cohesion and semantic cohesion. Referential cohesion measures co−reference, overlapped content words between sentences. The present study focuses on argument overlap (CRFAOa), which occurs when a noun in one sentence is overlapped with the same noun in another sentence. It also includes the overlap between pronouns. Another measurement of cohesion is Latent Semantic Analysis (LSA), which gauges semantic overlap between sentences or between paragraphs. Among eight LSA indices which Coh−Metrix provides, LSA sentence adjacent (LAASS1) is chosen for this study as it measures how semantically related each sentence is to the next sentence. According to Lee (2018), second language writers tend to produce more cohesive texts than first language writers because of excessive repetition of content words; cohesion can be an indicator of fluency and proficiency. Thus, it is noteworthy to compare Google translations and human writings in terms of cohesion.

The 12 indices of Coh−Metrix used for the purpose of the present study are briefly presented in Table 1 below. As presented in Table 1, this study conducted comparative analyses of human writings and corresponding Google translations on the basis of 5 analytical categories consisting of 12 specific indices provided by Coh−Metrix 3.0. The results and discussion section addressed each category to shed a light on insights about corpus−based differences between the two types of writings.

Table 1. Coh-Metrix Indices Used for the Study

Category	Indices	Description
Basic counts	DESSC	Number of sentences
	DESWC	Number of words
	DESSL	Sentence length
Lexical aspects	WRDFRQc	CELEX word frequency for content words
	LDTTRc	Type-token ratio for content words
Readability	RDFRE	Flesch Reading Ease(FRE)
	RDFKGL	Flesch-Kincaid Grade Level(FKGL)
Syntactic complexity	SYNLE	Words before main verb
	SYNNP	Number of modifiers per noun phrase
	SYNSTRUTa	Sentence syntax similarity for adjacent sentences
Cohesion	CRFAOa	Argument overlap
	LSASS1	LSA overlap

3.2 Data Collection and Analyses

In order to collect 60 samples of human writing, the authors implemented an extensive search for relevant doctoral dissertations to applied linguistics through RISS. Considering that the authors may not be able to understand the dissertations on other fields of discipline, the scope of the sampling was narrowed down to the dissertations only on the topic of applied linguistics. The sampled dissertations were published between 2006 and 2018, and all of them contained both Korean and English abstracts since an inclusion of both versions of abstracts was another key yardstick for screening a number of qualified dissertations. The English abstracts for these sampled dissertations were used to build a corpus of human writing that consisted of 60 writing samples. Subsequently, the corresponding Korean abstracts were translated into English through Google Translate, thereby establishing another corpus of Google translations consisting of 60 translation samples. The translated versions of abstracts were rigorously checked by three experts with English-related majors including the authors to filter out excessively incomprehensible or irrelevant translations. None of them were excluded from the analyses as a result. Each corpus was then analyzed using Coh-Metrix 3.0 against the selected 12 indices mentioned earlier.

As for the quantitative data analyses, the first step taken was to check the normality of data sets because the normality should be satisfied for obtaining credible results from independent *t*-tests. Because Tsai (2019) did independent *t*-tests in his study that compared students' English writings and Google translated versions, this study implemented the same statistical procedure. To do so, the Shapiro−Wilk test was conducted on all the data sets collected. It turned out that data sets of seven indices (i,e., DESSL, WRDFRQc, FRE, SYNNP, SYNSTRUTa, CRFAOa, and LSASS1) were normally distributed and allowed for independent *t*−tests. The inferential statistics by independent *t*−tests on the normally distributed data were interpreted based upon Student's *t* or Welch's *t* in accordance with homoscedasticity of the data sets. In addition, given the relatively small sample size of the present study that might cause low statistical power (Wilcox 2001), the findings from independent *t*−tests on normally distributed data sets were compared with those of bootstrapped independent *t*−tests. The non−normally distributed data sets were first analyzed through a non−parametric test such as Mann−Whitney U, and its results were further compared with those of bootstrapped independent *t*−tests in order to enhance the reliability of quantitative findings and validity of interpretations. Krishnamoorthy, Lu, and Matthew (2007) argued that the bootstrapping technique with 100 repetitions can be an alternative statistical solution to a non−normality issue. In L2 research practices, Laflair, Egbert and Plonsky (2015) highlighted the need for adopting bootstrapping in quantitative analyses, stating that "small samples/low power and nonnormal data have been described as two common problems that bootstrapping may partially overcome...*bootstrapping, which is argued to mitigate the negative effects of all these conditions simultaneously* [emphasis added]" (p. 593). Furthermore, bootstrapping is known as being "less sensitive to irregularities such as outliers, thus *providing descriptive and test statistics that are robust to deviations from normality in the original sample* [emphasis added]" (Laflair et al. 2015, p. 593). The statistical analyses were conducted using Jamovi version 0.9.1.6 and SPSS for Mac 25.

4. Results and Discussion

4.1 Differences in terms of Basic Counts

The descriptive statistics for the basic counts that include DESSC, DESWC, and DESSL is presented in Table 2.

Table 2. Descriptive Statistics for Basic Counts

Index	Group	N	M	SD
DESSC	Human	60	25	12.40
	Google	60	24	8.82
DESWC	Human	60	548	247.90
	Google	60	643	271.99
DESSL	Human	60	21.90	4.82
	Google	60	26.60	3.33

The inferential statistics for DESSC, DESWC, and DESSL is tabulated in Table 3 as follows.

Table 3. Inferential Statistics for Basic Counts

Index	Group	Statistical procedure	statistic	*df*	*p*	Mean differences
DESSC	Human Google	Mann−Whitney U	1716	N/A	.66	1.00
DESWC	Human Google	Mann−Whitney U	1260	N/A	.01	−110.00
DESSL	Human Google	Welch's *t*	−6.23	105	.00	−4.71

As shown in Table 3, a non−significant mean difference in DESSC was observed through Mann−Whitney U, *U* = 1716, *p* = .66, Cohen's *d* = .16. This non−significant difference was consistent with the result of bootstrapped independent *t*−test, *p* = .43. Therefore, it can lead to a conclusion that there are no significant differences in the number of sentences between human English writing (M = 25, SD = 12.4) and Google translations (M = 24, SD = 8.82). On the other hand, the difference in DESWC was found to be significant in light of Mann−Whitney U, *U* = 1260, *p* = .01, Cohen's *d* = 0.43. In order to check the credibility of the findings, a bootstrapped independent *t*−test was subsequently conducted on DESWC data and the difference turned out to be significant, *p* = .04. All in all, these results showed that Google translations (M = 643, SD = 271.99) are likely to produce significantly more words than human English writings (M = 548, SD = 247.9) do. However, the effect size signified by Cohen's *d* suggested that these differences are small. Lastly, the mean differences in terms of DESSL were confirmed through Welch's *t* as the data suggests a violation of the assumption of equal variances. The result found a significant difference in DESSL, *p* = .00, Cohen's *d* = 1.14. In other words, the Google translations tend to yield significantly longer sentences (M = 26.9, SD = 3.33) than human English writings (M = 21.9, SD = 4.82) do. The effect size suggested that the observed difference is large.

4.2 Differences in terms of Word Frequency and Lexical Diversity

The descriptive statistics for word frequency measured by WRDFRQc and lexical diversity by LDTTRc is presented in Table 4.

Table 4. Descriptive Statistics for Word Frequency and Lexical Diversity

Index	Group	N	M	SD
WRDFRQc	Human	60	2.01	.11
	Google	60	1.99	.11
LDTTRc	Human	60	.53	.09
	Google	60	.51	.08

The inferential statistics for WRDFRQc and LDTTRc is summarized in Table 5 as follows.

Table 5. Descriptive Statistics for Word Frequency and Lexical Diversity

Index	Group	Statistical procedure	statistic	*df*	*p*	Mean differences
WRDFRQc	Human Google	Student's *t*	1.74	118	.09	.02
LDTTRc	Human Google	Student's *t*	1.07	118	.29	.02

As demonstrated in Table 5, the inferential statistics for WRDFRQc corroborated that the difference in WRDFRQc is not statistically significant, $t(118) = 1.07$, $p = .29$, Cohen's $d = .20$. The bootstrapped result supported the finding, $p = .29$. Therefore, Google translations (M = 1.99, SD = .11) are estimated to employ the same level of content words as human English writing (M = 2.01, SD = .11) given the fact that WRDFRQc refers to CELEX word frequency for content words. In addition, the inferential statistics for LDTTRc found that the difference is not statistically significant either, $t(118) = 1.72$, $p = .09$, Cohen's $d = .31$. The bootstrapped independent t-test result yielded a non-significant finding as well, $p = .08$. Therefore, it can be concluded that Google translations (M = .53, SD = .09) are not significantly different from human English writings (M = .51, SD = .08) in terms of type−token ratio for content words. The two types of writings do not seem to be distinguishable in the use of content words from the perspective of frequency and diversity.

4.3 Differences in terms of Readability

The descriptive statistics for readability that were calculated by both RDFRE and RDFKGL is presented in Table 6. It should be noted that the higher the value of RDFRE, the easier to comprehend the texts. In case of RDFKGL, the lower value means that it is easier to understand.

Table 6. Descriptive Statistics for Readability

Index	Group	N	M	SD
RDFRE	Human	60	29.50	9.89
	Google	60	26.50	8.80
RDFKGL	Human	60	14.70	2.32
	Google	60	16.30	1.79

The inferential statistics for FRE and FKGL is represented in Table 7 as follows.

Table 7. Inferential Statistics for Readability

Index	Group	Statistical procedure	statistic	*df*	*p*	Mean differences
RDFRE	Human Google	Student's t	1.74	118	.09	2.97
RDFKGL	Human Google	Mann−Whitney U	983	N/A	.00	−1.69

As shown in Table 7, the mean difference in FRE was found to be non−significant, *t*(118) = 1.74, *p* = .09, Cohen's *d* = .32. The non−significance was also found in the result of the bootstrapped independent t−test, *p* = .09. Therefore, it can be concluded that Google translations (M = 29.50, SD = 9.89) are not significantly more difficult to read than human English writings (M = 26.50, SD = 8.80) are. However, the findings for RDFKGL demonstrated that Google translations (M = 16.30, SD = 1.79) are significantly more difficult to read than human English writings (M = 14.70, SD = 2.32), *U* = 983, *p* = .00, Cohen's *d* = .73. The effect size calculated by Cohen's *d* suggested that the difference was moderate. The bootstrapped result was also found to be significant, p = .01, thereby supporting the validity of the result. Overall, the outputs of Google Translate tend to be more difficult to comprehend in terms of their readability computed by RDFKGL. These results seem to be relevant to the findings on DESSL which refers to mean sentence length in light of the fact that the formula underlying RDFRE and RDFKGL is known to be sensitive to sentence length (Jeon and Lim 2009). It has already been confirmed that Google translations tend to produce significantly longer sentences than human English writings do.

4.4 Differences in terms of Syntactic Complexity

The descriptive statistics for the syntactic complexity measured by SYNLE, SYNNP, and SYNSTRUTa is presented in Table 8.

Table 8. Descriptive Statistics for Syntactic Complexity

Index	Group	N	M	SD
SYNLE	Human	60	5.99	2.15
	Google	60	7.58	1.72
SYNNP	Human	60	1.20	.19
	Google	60	1.21	.13
SYNSTRUTa	Human	60	.10	.03
	Google	60	.10	,02

The inferential statistics for SYNLE, SYNNP, and SYNSTRUTa is represented in Table 9 as follows.

Table 9. Inferential Statistics for Syntactic Complexity

Index	Group	Statistical procedure	statistic	*df*	*p*	Mean differences
SYNLE	Human Google	Mann-Whitney U	938	N/A	.00	−1.72
SYNNP	Human Google	Welch's *t*	−.250	107	.80	.03
SYNST RUTa	Human Google	Welch's *t*	1.54	102	.13	.00

As presented in Table 9, the mean difference in SYNLE was statistically significant, U = 938, p = .00, Cohen's d = .82 This significant finding was also confirmed by the bootstrapped independent t-test, *p* = .01. The effect size suggested that this difference is large. As SYNLE refers to left embeddedness or mean number of words before main verbs, Google translations (M = 7.58, SD = 1.72) tend to employ significantly more words before main verbs than human English writings do (M = 5.99, SD = 2.15). However, the mean difference in SYNNP was not statistically significant, *t*(107) = -.25, *p* = .80, Cohen's *d* = .05. The bootstrapped independent *t*-test also found non-significant difference between them in SYNNP, *p* = .79. Therefore, it can be inferred that two versions of writings are not significantly different from the perspective of SYNNP that refers to the mean number of modifiers per noun phrases. Furthermore, it was found that the two types of writings did not show significant difference regarding SYNSTRUTa as well, *t*(111) = -3.99, *p* = .13, Cohen's *d* = .28. The difference was also found to be non-significant in the bootstrapped independent t-test, *p* = .15. This means that both of the writings are not significantly different in terms of "the average parse tree similarity between adjacent sentence pairs in a text" (McNamara, Graesser, McCarthy & Cai, 2014, p. 71). Overall, these quantitative findings suggest that sentences in Google translations are likely to be more syntactically complex than those of corresponding human English writings in that Google translations place more words before main verbs as supported by the significant difference in SYNLE. The difference in syntactic complexity seems to serve as another contributing factor to lower readability of Google translations.

4.5 Differences in terms of Co-referential and Semantic Cohesion

The descriptive statistics for the co-referential and semantic cohesion that were measured by CRFAOa and LSASS1 is presented in Table 10.

Table 10. Descriptive Statistics for Co-referential and Semantic Cohesion

Index	Group	N	M	SD
CRFAOa	Human	60	.58	.16
	Google	60	.65	.12
LSASS1	Human	60	.35	.01
	Google	60	.39	.01

The inferential statistics for CRFAOa and LSASS1 is summarized in Table 11 as follows.

Table 11. Inferential Statistics for Co-referential and Semantic Cohesion

Index	Group	Statistical procedure	statistic	*df*	*p*	Mean differences
CRFAOa	Human Google	Welch's *t*	−2.54	108	.01	.03
LSASS1	Human Google	Student's *t*	−2.10	118	.04	.02

As shown in Table 11 above, the mean difference in CRFAOa was statistically significant, *t*(108) = −2.54, *p* = .01, Cohen's *d* = .05. This significant finding was also confirmed by the bootstrapped independent t-test, *p* = .03. However, the observed effect was moderate. Given the fact that the index of CRFAOa measures co-referential cohesion based upon arguments (i.e., nouns, pronouns, and noun phrases) overlap, Google translations (M = .65, SD = .12) are significantly more cohesive than human English writing (M = .58, SD = .16) in terms of the degree to which arguments overlap, albeit a small observed effect. The two kinds of writings were also observed to be significantly different in their numerical values for semantic cohesion computed by LSASS1, *t*(118) = −2.10, *p* = .04, Cohen's *d* = .40. The result was consistent with the significant finding of the bootstrapped independent *t*-test, *p* = .02. Therefore, it can be inferred that Google translations (M = .39, SD = .01) tend to demonstrate significantly more semantic cohesion than human English writings (M = .35, SD = .01) do. All in all, these quantitative findings indicate that Google Translate can achieve a high level of co-referential cohesion of texts through appropriate uses of the arguments, and that propositions in Google translations are conceptually and semantically related to each other rather than fragmented.

5. Conclusion

The present study was carried out with a specific aim to identify corpus-based differences between human writings and their corresponding Google translations. For this purpose, 60 abstracts for doctoral dissertations on topics of applied-linguistics, which were searchable through RISS, were used to build corpora. The corpora were quantitatively analyzed in terms of basic counts (i.e., DESSC, DESWC, and DESSL), lexical aspects (i.e., WRDFRQc and LDTTRc), readability (i.e., RDFRE and RDFKGL), syntactic complexity (i.e., SYNLE, SYNNP, and SYNSTRUTa), and cohesion (i.e., CRFAOa and LSASS1). The statistical procedure was chosen in light of normality and heteroscedasticity of each data set. The findings of inferential statistics were confirmed by the results of bootstrapping to enhance the validity of the interpretations. There were several statistically significant differences between human writings and Google translations.

In basic counts, two types of writings were significantly different only in terms of DESSL that measures the average number of words in each sentence within the text. Google translations tend to produce significantly longer sentences, and the difference was large in light of the effect size. However, no significant differences were found between them in lexical aspects that were computed by CELEX word frequency and type-token ratios for content words in the texts. It implies that Google Translate is quite advanced in its choice of content words and stylistic sensitivity to diversify lexical items.

Google translations were found to be significantly less readable (i.e., or more difficult to read) than human writings. This finding can be partially explained by the significant difference between the two types of writings in sentence length measured by DESSL. Regarding syntactic complexity, they were significantly different in terms of SYNLE, which means that Google Translate produced more words before main verbs. It also seems to provide a partial account of the lower readability of the Google translations.

Regarding cohesion measured by CRFAOa and LSASS1, Google translations were significantly more cohesive than human writings. This result indicates that Google Translate is able to use arguments such as nouns and pronouns appropriately for achieving co-referential cohesion and to reflect conceptual or semantic relatedness of ideas and propositions of a given text in their translations.

This study has identified some differences between human writings and Google Translate using Coh-Metrix 3.0. Those findings are expected to lay out the foundation stones for the areas of L2 research that needs an in-depth understanding of distinctive features of Google Translate and human writings, and for L2 practitioners who are interested in pedagogical applications of Google Translate in teaching L2 writing. As Google Translate has been proven to have a great lexical density and a higher level of vocabulary (Tsai 2019), which aligns with the findings of the present study that Google Translate has achieved the advanced level of content word choice, it illustrates that Google Translate has a great potential to aid vocabulary expansion of second language writers. Furthermore, as previous research has mainly focused on its capability to offer initial advice on word choice or sentence structure, the less explored areas of incorporating Google Translate into L2 writing pedagogy such as cohesion, readability, or syntactic complexity will benefit from this study.

However, there are some perceived limitations in this study. Even though the authors tried to compensate for the expected statistical issues related to the small sample size by adopting bootstrapping, the 60 samples seem to be relatively small in ensuring external generalizability of the major findings of the present study. In addition, the quality of writing was not taken into consideration in pursuit of Coh−Metrix−based differences between them. It is hoped that future studies on similar topics may build a larger corpora and consider the quality of writings or semantic aspects as a variable for the design of their studies. Lastly, in this paper, the authors considered the abstracts of dissertations written by Koreans as human writing samples. This may constitute a methodological flaw in the research design because the authors had not been able to observe the actual writing processes and failed to take into consideration the question of how well the samples' characteristics reflect the true nature of English writing of L2 learners of Koreans. In this regard, the methodological limitation especially in selecting representative samples of human writings needs to be rigorously compensated for in the future studies. All in all, this study can be seen as a preliminary investigation that just brings light to hitherto less studied areas of doubts over Google Translate.

References

Ahn, B. and Y. Ma. 2015. A Coh−metrix analysis of elementary school English textbooks. *English 21* 28(3). 435−460.

Ahrenberg, L. 2017. Comparing machine translation and human translation: A case study. *Proceedings of The First Workshop on Human−Informed Translation and Interpreting Technology (HiT−IT),* 21-28.

Aiken, M. and S. Balan. 2011. An analysis of Google Translate accuracy. *Translation Journal* 16(2), 12−34.

Aryadoust, V. 2016. Understanding the growth of ESL paragraph writing skills and its relationships with linguistic features. *Educational Psychology* 36(10), 1742−1770.

Bellos, D. 2011. *Is That a Fish in Your Ear? Translation and the Meaning of*

Everything. New York: Faber and Faber.

Castelvecchi, D. 2016. Deep learning boosts Google translate tool. *Nature.* Retrieved from https://www.nature.com/news/deep−learning−boosts−google−translate−tool−1.20696

Colina, S. 2009. Further evidence for a functionalist approach to translation quality evaluation. *Target* 21(2), 235−264.

CORREA, M. 2014. Leaving the 'peer' out of peer−editing: Online translators as a pedagogical tool in the Spanish as a second language classroom. *Latin American Journal of Content & Language Integrated Learning* 7(1), 1−20.

Costa−jussá, M., M. Farres and J. Pons. 2012. *Machine translation in medicine: A quality analysis of statistical machine translation in the medical domain.* Paper presented at the Advanced Research in Scientific Areas.

Ducar, C. and D. H. Schocket. 2018. Machine translation and the L2 classroom: Pedagogical solutions for making peace with Google Translate. *Foreign Language Annals* 51(4), 779−795.

Enkin, E. and E. Mejías−Bikandi. 2016. Using online translators in the second language classroom: Ideas for advanced−level Spanish. *Latin American Journal of Content & Language Integrated Learning* 9(1), 138−158.

Ferris, D. R., H. Liu, A. Sinha and M. Senna. 2013. Written corrective feedback for individual L2 writers. *Journal of Second Language Writing* 22(3), 307−329.

Garcia, I. 2010. Is machine translation ready yet? *Target−International Journal of Translation Studies* 22(1), 7−21.

Garcia, I. and M. I. Pena. (2011). Machine translation−assisted language learning: Writing for beginners. *Computer Assisted Language Learning* 24(5), 471−487.

Google. 2012, April 26. *Breaking down the language barrier − six years in.* Retrieved from https://translate.googleblog.com/2012/04/breaking−down−language−barriersix−years.html

Graesser, A. C. and D. S. McNamara. 2011. Computational analyses of multilevel discourse comprehension. *Topics in Cognitive Science* 3(2), 371−398.

Graesser, A. C., D. S. McNamara, M. M. Louwerse and Z. Cai. 2004. Coh−Metrix: Analysis of text on cohesion and language. *Behavior Research Methods, Instruments, & Computers* 36(2), 193−202.

Groves, M. and K. Mundt. 2015. Friend or foe? Google translate in language for academic purposes. *English for Specific Purposes* 37, 112−121.

Hampshire, S. and C. Porta Salvia. (2010). Translation and the internet: Evaluating the quality of free online machine translators. *Quaderns: Revista De Traducci* 17, 197-209.

I. Solnyshkina, M. V., E. Harkova and S. A. Kiselnikov. 2014. Comparative Coh-metrix analysis of reading comprehension texts: Unified (Russian) state exam in English vs Cambridge first certificate in English. *English Language Teaching* 7(12), 65.

Jeon, M-G. and I-J. Lim. 2009. A Corpus-based Analysis of middle school English 1 textbooks with Coh-Metrix. *English Language Teaching* 21(4), 265-292.

Kaltenbacher, M. 2000. Aspects of universal grammar in human versus machine translation. In A. Chesterman, N. San Savador and Y. Gambier eds., *Translation in Context,* 221-230. Amsterdam: John Benjamins.

Kapadya, A. 2018, Sep 20. *Bringing hope to a refugee family, using Google Translate* [Blog post]. Retrieved from https://www.blog.google/products/translate/refugee-family-translate/

Kim, J. 2014. Continuity problems of elementary and secondary English education: Sequence analysis of English textbooks. *Language Research* 50(1), 161-184.

Kim, J-K. 2018. A Corpus-based investigation of reading passages in the national assessment of educational achievement English test using coh-metrix. *Secondary English Education* 11(2), 27-51.

Kirchhoff, K., A. M. Turner, A. Axelrod and F. Saavedra. 2011. Application of statistical machine translation to public health information: A feasibility study. *Journal of the American Medical Informatics Association* 18(4), 473-478.

Krishnamoorthy, K., F. Lu and T. Mathew. 2007. A parametric bootstrap approach for ANOVA with unequal variances: Fixed and random models. *Computational Statistics & Data Analysis* 51(12), 5731-5742.

Kroulek, A. 2017, Feb 28. *A Translation Showdown: Man vs Machine Translation*. Retrieved February 27, 2019, from https://www.k-international.com

LaFlair, G. T., J. Egbert and L. Plonsky. 2015. Bootstrapping. In L. Plonsky ed., *Advancing Quantitative Methods in Second Language Research,* 135-150. Routledge.

Lee, J. 2018. A Coh-metrix analysis of lexical, syntactic and discourse aspects in the newspaper articles of Korean and British university students. *The Modern English Education Society* 19(4), 17-26.

Lewis-Kraus, G. 2016, Dec 14. *The great A.I. awakening*. Retrieved from https://search.proquest.com/docview/1848431038

McCarthy, P. M., C. Hall, N. D. Duran, M. Doiuchi, Y. Fujiwara, B. Duncan and D. S.

McNamara. 2011. Analyzing journal abstracts written by Japanese, American, and British scientists using Coh-metrix, and the Gramulator. *The ESPecialist: Research in Language for Specific Purposes* 30(2). 141-173.

McNamara, D. S., A. C. Graesser, P. M. McCarthy and Z. Cai. 2014. *Coh-Metrix: Automated Evaluation of Text and Discourse*. Boston: Cambridge University Press.

Plonsky, L. 2013. Study quality in SLA: An assessment of designs, analyses, and reporting practices in quantitative L2 research. *Studies in Second Language Acquisition* 35(4), 655-687.

Plonsky, L. 2014. Study quality in quantitative L2 research (1990-2010): A methodological synthesis and call for reform, *Modern Language Journal* 98(1), 450-470.

Polio, C. and H. Yoon. 2018a. The reliability and validity of automated tools for examining variation in syntactic complexity across genres. *International Journal of Applied Linguistics* 28(1), 165-188.

Ramati, I. and A. Pinchevski. 2018. Uniform multilingualism: A media genealogy of Google translate. *New Media & Society* 20(7), 2550-2565.

Rodreguez-Castro, M., S. Salas and T. Benson. 2018. To Google translate™ or not? Newcomer Latino communities in the middle. *Middle School Journal* 49(2), 3-9.

Sheppard, F. 2011. Medical writing in English: The problem with Google translate. *La Presse Medicale* 40(6), 565-566.

Tsai, S-C. 2019. Using Google Translate in EFL drafts: A preliminary investigation, *Computer Assisted Language Learning* 32(5-6), 510-526.

Turner, A. M., M. Bergman, M. Brownstein, K. Cole and K. Kirchhoff. 2014. A comparison of human and machine translation of health promotion materials for public health practice: Time, costs, and quality. *Journal of Public Health Management and Practice* 20(5), 523-529.

Turovsky, B. 2016, April 28. *Ten years of Google translate* [Blog post]. Retrieved from https://www.blog.google/products/translate/ten-years-of-google-translate/

Wang, Z. and S-A. Lee. 2018. Machine translation versus human translation: The case of English-to-Chinese translation of relative clauses. *Language and Information* 22(1), 175.

Ye, D. 2013. A Coh-metrix analysis of language varieties between the journal articles of Chinese and American scientists. *International Journal of English Linguistics* 3(4), 63-70.

Zhang, R. 2015. A Coh−metrix study of writings by majors of mechanic engineering in the vocational college. *Theory and Practice in Language Studies* 5(9), 1929−1934.

Examples in: English
Applicable Languages: English
Applicable Level: Tertiary

Kim, Joonkoo
Ph.D Student, Graduate School of English Education
Chung−Ang University/Changdeok Girls' Middle School
22, Jeongdong−gil, Jung−gu, Seoul, Republic of Korea
Tel: 010) 9027−3362
E−mail: viali@sen.go.kr

Lim, Kyuyun
MA Student, Language and Literacy Education Department
University of British Columbia
6445 University Blvd, Vancouver, Canada
Tel: +1−604−785−8132
E−mail: kyuyun.lim@alumni.ac.kr

Received: July 25, 2019
Revised: September 10, 2019
Accepted: September 19, 2019