InContext 2:1 (2022), PP. 57-83. ISSN 2799-6190 | eISSN 2799-8952
https://doi.org/10.54754/incontext.v2i1.11

Looking under the Hood for Evidence of Normalization

Multivariate exploratory analysis of lexical bundles

Changsoo Lee
Hankuk University of Foreign Studies, Seoul, Korea
soolee@hanmail.net

Abstract: The study investigated the hypothesis of normalization and stylistic variation across translators as manifested in the use of lexical bundles between translated and non-translated English literary texts. Normalization is a hypothesis originally proposed as 'conservatism' by Baker (1996) which states that the translator tends to conform to linguistic patterns and conventions typical of the target language even to the point of exaggeration, and lexical bundles are sequences of three or four words recurring with high frequency in natural discourse. The study was carried out in two stages. The first stage replicated previous studies that relied on simple frequency tests to confirm the normalization hypothesis. Contrary to these earlier studies, the present study's frequency tests on lexical bundles failed to provide clear support for the normalization hypothesis. The second stage employed two types of multivariate exploratory analysis, principal component analysis (PCA) and hierarchical cluster analysis (HCA), to examine the underlying relationships among individual texts, lexical bundles, and translated and non-translated group categories. Following the failed frequency tests, it was hypothesized here that normalization might be still present in the translated corpus but restricted by types of lexical bundles. PCA confirmed this hypothesis by revealing that normalization occurred in the use of a particular functional type of lexical bundles, called discourse bundles, which are relatively free from the thematic content of the text in which they occur. This ascertains the traditional idea that statistical tests of translation hypotheses must deal with linguistic features unrelated to the thematic content of the corpus. Additionally, PCA revealed variation across the types of lexical bundles preferred by individual translators. HCA further identified the presence of a subgroup of translated texts that cluster with non-translated texts, rather than with their fellow translated texts. This was taken as indicating that the use of lexical bundles varied among the translators and that the division between translated and non-translated texts is not clear-cut.

Keywords: corpus-based translation studies, normalization, principal component analysis, hierarchical cluster analysis, Korean-English literary translation

논문초록: 본 연구는 번역과 비번역 영어 문학 텍스트 간에 어휘 번들(lexical bundle)의 사용 패턴에 기초하여 번역 표준화 가설과 번역가 별 문체 차이를 연구하는데 목적이 있다. 표준화란 베이커(1996)가 초기에 보수성이라는 명칭으로 제한한 가설로 번역가는 과장될 정도로 도착어의 언어 패턴이나 규범을 따르는 경향이 있다는 주장이다. 어휘 번들은 자연 담화에서 고빈도로 발생하는 어휘 패턴으로 몇 개의 단어가 연쇄적으로 연결된 단위를 일컫는다. 본 연구는 2단계로 진행되었는데, 1단계에서는 빈도 분석에 기초하여 표준화 가설을 입증한 이전 연구를 재현하여 연구결과를 검증하였다. 이전 연구와 달리 본 연구의 빈도 분석에서는 표준화를 입증하는 결과가 도출되지 않았다. 두번째 단계에서는 다변수 탐구적 통계분석법인 주성분분석(PCA)과 계층적 클러스터분석(HCA)를 사용하여 개별 텍스트, 어휘 번들, 번역–비번역 집단 범주 간의 기저 관계를 분석하였다. 표준화 입증에 실패한 빈도 분석에 이어 2단계에서는 번역코퍼스에 여전히 표준화가 존재하지만 특정 번들 형태에 제약을 받는다는 가설을 세워 검증했다. PCA분석에서는 텍스트의 주제 내용에서 비교적 자유로운 담화 번들이라는 특정한 종류의 어휘 번들에서만 표준화가 목격되어 동 가설을 뒷받침하는 결과가 도출되었다. 이는 번역보편소에 대한 통계 분석은 코퍼스의 주제 내용의 영향을 받지 않는 언어 자질을 사용해야 한다는 기존 주장을 뒷받침한다. PCA에서는 번역자 간에 선호되는 어휘 번들의 종류에서도 차이가 나타났다. HCA 분석에서는 번역 텍스트 중 일부가 다른 번역 텍스트와 거리를 두고 비번역 텍스트와 군집을 형성하는 것이 추가로 확인되었다. 이는 번역자 간에도 어휘 번들 사용 양상에서 차이가 존재하며 번역과 비번역 텍스트 간의 경계가 명확하지 않다는 것을 보여주는 결과로 해석된다.

핵심어: 코퍼스기반 번역연구, 표준화, 주성분분석, 계층적 클러스터분석, 한영문학 번역

1. Introduction

Lexical bundles, also known as word clusters or n-grams, are sequences of words that recur frequently in natural discourse (Biber et al., 1999, p. 990). Some examples of three-word lexical bundles frequent in English discourse are *are going to, at the end of* and *one of the*. As can be seen in these examples, lexical bundles are usually incomplete text fragments. Nevertheless, their high frequency and function of bridging phrases or clauses make them basic building blocks of discourse (Biber et al., 2004, p. 376).

Lexical bundles were first noticed as an object worth investigating in translation studies by Baker (2004). She used a computer program to extract recurring lexical sequences from translated and non-translated English corpora in order to test Venuti's (1995) hypothesis that an emphasis on fluency in translation would result in greater reliance on fixed expressions in translated English texts than in non-translated English texts. Despite Baker's pioneering work, lexical bundles have gone largely unnoticed by translation scholars. Recently, however, interest in lexical bundles has been rekindled as they were linked up with research on the hypothesis of normalization, which is defined as a 'tendency to exaggerate features of the target language and to conform to its typical patterns' (Baker, 1996, pp. 183-184). Xiao (2010) investigated bundles of two to six words in translated and non-translated Chinese corpora and found that the former was more frequent in the use of typical lexical bundles than the latter. Xiao interpreted this finding as supportive of the hypothesis of normalization. Xiao's work was followed by Lee (2013), whose case studies also found greater reliance on lexical bundles in translated Korean media texts than in comparable non-translated texts.

Methodologically, research on lexical bundles in translation has relied on a rather simple process of counting up high-frequency lexical bundles in translated and non-translated corpora and comparing their frequencies to determine whether the two corpora are different or not. Such confirmatory statistical analysis can provide a test of statistical significance but are seriously limited in allowing the researcher to probe deeper into his/her data and investigate its internal structure. Furthermore, frequency tests on such underlying relationships in the data can produce misleading results. It should be emphasized that the relationships among corpora, individual sample texts, and lexical bundles are not simple but highly complex and multidimensional, as is the case with all other linguistic phenomena (De Sutter et al., 2012, p. 327; Jenset & McGillivray, 2012, p. 303). Modern statistics provides us with a battery of multidimensional statistical techniques to explore with great success such complex relationships among linguistic features, texts and corpora.

Against this backdrop, the aim of the present study is to use two such techniques, namely, principal component analysis and hierarchical cluster analysis, to investigate the hypothesis of normalization as manifested in lexical bundles, using a corpus of English translations of Korean novels and a comparable corpus of authentic English fiction.

2. Lexical Bundles and Methods of Research in Translation

As defined earlier, lexical bundles are sequential combinations of words usually ranging from two to five words which recur in written or spoken discourse. They are a form of fixed expressions but differ from idioms on two important accounts. First, lexical bundles are highly frequent. In fact, high frequency is the sole basis for identifying lexical bundles. In contrast, idioms occur rarely in natural discourse (Biber et al., 1999, p. 183; Hyland, 2008, p. 6). Second, most lexical bundles are structurally incomplete as was shown in earlier examples. This makes them almost unrecognizable by speakers. Idioms, on the other hand, are independent structural units that are salient and easily discernible in discourse (Biber & Conrad, 1999, p. 184; Biber et al., 1999, p. 990). Incomplete as they are, lexical bundles are building blocks in discourse, providing pragmatic heads for phrases and clauses and serving as discourse frames for expressing new information (Biber & Barbieri, 2007, p. 277). Research has found that different genres, for example, academic prose and conversation, are distinguished from each other in the types and frequencies of lexical bundles used in them (cf. Biber et al., 2004; Biber et al., 1999; Scott & Tribble, 2006). This shows that lexical bundles typify language use.

Typicality is what links lexical bundles to the translation research on normalization. Baker (1996, pp. 183-184) stresses typicality as the most important requirement of a linguistic feature manifesting normalization. Over the years, supporting evidence for normalization has come from studies that have looked into such linguistic features as collocations (Kenny, 2001, 2000, 1999, 1998), lexical choices (Malmkjaer, 1998) and binominals of near synonyms (Toury, 1995, pp. 102-112; 1980, p. 131). The commonality among these features is that they represent typical language use either in source or target languages. As building blocks of natural discourse and 'extended collocations' (Biber & Conrad, 1999, p. 183), lexical bundles are an equally effective test bed for studying the phenomenon of normalization in translation.

The techniques of analysis used so far in the investigation of lexical bundles in translated texts are largely borrowed from corpus stylistics (cf. Biber et al., 2004; Scott & Tribble, 2006). Typically, the researcher uses a computer program to extract lexical bundles from translated and non-translated corpora and compares the frequencies to examine if there is a difference. When the translated corpus turns out to be more frequent, it is taken as evidence that translated texts are normalized by exaggerated use of lexical bundles typical of the target language. This is the essence of the process of analysis adopted by Xiao (2010) and Lee (2013) as well as by Baker (2004).

Statistical analyses used in such studies have been predominantly of the confirmatory kind, such as the log-likelihood (G2) test used by Xiao (2010). These tests can tell the researcher whether observed differences are statistically significant or not but have little to say about the data itself. In lexical bundle analysis, the researcher typically deals with a large amount of sample texts and hundreds of lexical bundles from different corpora, which give rise to multidimensional relationships. Confirmatory statistical analyses tend to reduce such complex relationships to a simple issue of statistically significant or non-significant differences. But rarely do corpora split clearly into two separate entities. They are close to each other in some areas while clearly disparate in others. This relates to the issue of variation across individual texts and translators, which Baker (2004, pp. 181-182) regards as important as overall patterns of similarity or dissimilarity. Another point to consider is the fact that not all high-frequency lexical bundles have an equal effect on the relationship between corpora. Some would have little effect as they are shared, while others will be clear differentiators as they are distinctly associated with one particular corpus. Understanding such internal structure of the data under investigation is crucial to uncovering the true nature of the relationship between translated and non-translated language. This objective can be better served by various methods of exploratory multivariate analysis (Husson et al., 2011). The usefulness of such techniques for translation research has been demonstrated by some studies such as De Sutter et al. (2012), Forsyth and Lam (2014), Grabowski (2013), Jenset and McGillivray (2012), Lee (2021), Rybicki (2006), and Rybicki and Heydel (2013).

3. Methods

The present study makes use of a translated English corpus and a comparable non-translated English corpus. The translated corpus (hereafter referred to as KTT) consists of the full texts of the English translations of 21 Korean novels. The translated sample texts were chosen primarily on the basis of availability, i.e. from the texts the author was able to have access to through libraries. The original Korean novels were published between the 1960s and the early 2000s, but most translations were done in the 2000s. The novels are general fiction in terms of literary subgenres as they deal with such generic themes as families, romance, society, action and the Korean War. The translated corpus totals 1,449,422 words. The comparable English corpora (hereafter referred to as CET), similarly made up of 21 original literary works, was constructed to match the translated corpora approximately in published periods and subject themes. Their total words amount to 1,968,670. The difference in corpus size results from the texts in CET tending to be lengthier than the KTT texts, but it will not be a problem for the present study as it will use normalized frequencies. This will be elaborated on in Section 4.1. Table 1 summarizes the general properties of the two corpora.

Table 1: General information about KTT and CET

	KTT	CET
Number of sample texts	21	21
Total words	1,449,422	1,968,670
STTR	95.64	95.95
Average words per sentence	12.30	10.62

* STTR = standardized type-token ratio (type-token ratio per 1,000 words in %)

Data analysis will proceed in two stages. In the first stage, we try to partially replicate the results reported by Xiao (2010) and Lee (2013), using conventional confirmatory statistical tests as the authors did, to set the stage for comparing these tests with the exploratory tests to be run in the second stage and highlighting the latter's value in terms of the new insights they add to our analysis. The first-stage analysis compares the two corpora in overall and high-frequency three-word lexical bundles. Three-word bundles are most popularly used in corpus stylistics research (cf. Scott & Tribble, 2006). Wordsmith Tools (Version 6) will be used to extract total raw frequencies from each corpus. The frequencies are normalized to one million to permit direct comparison between the two corpora which differ in size. Our main concern here is to test if there is any statistically significant difference in the frequencies of lexical bundles between the two corpora, using the R statistical software (http://www.r-project.org).

The second stage of analysis analyzes the 70 most frequent three-word lexical bundles, using two exploratory tests. The number 70 is an arbitrary choice. Burrows (2002, 1987) worked with 40 up to 150 most frequent English words in authorship attribution experiments, showing that roughly 100 most frequent words are enough to effectively identify authors. This, however, was in the case of single word lists. With lexical bundles, the list becomes sparser as we go from one- to two-, and to three-word bundles and so on, which requires the researcher to curtail the list against increased sparsity.

The data in the second stage is extracted from the combined pool of KTT and CET texts, using the *stylo* package in R. This will give us a 42x71 data frame, made up of 42 rows and 71 columns, including one additional column identifying the texts as translated or non-translated. The dataset is then subjected to PCA and HCA. Our aim in these analyses is to look into the internal structure of the data by analyzing how the corpora, lexical bundles and individual text samples are associated with one another. Each of these points will be elaborated on in the relevant sections of data analysis.

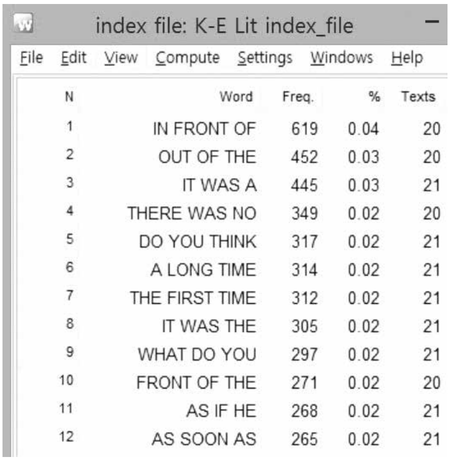
4. Data Analysis

4.1 Frequency Analysis

We will start with a comparison of overall frequencies of three-word bundles between KTT and CET. Extracting repetitive word sequences from a corpus would be an impossible task without a computer. Wordsmith Tools makes the task a simple process. The program first creates an index list of words for the corpus we want to investigate and then uses the index list as a basis for computing lexical bundles of a given length. The program can produce lexical bundles of different word lengths at the same time. But to make the situation simpler, we will focus on three-word lexical bundles, which are the standard length of bundles used in corpus stylistics. Figure 1 illustrates the top 12 on the KTT and CET three-word bundle lists respectively. The list shows the raw frequency of each bundle ('Freq'), the percentage the bundle represents in the corpus' total word count ('%'), and the number of the sample texts in which it occurs ('Texts').

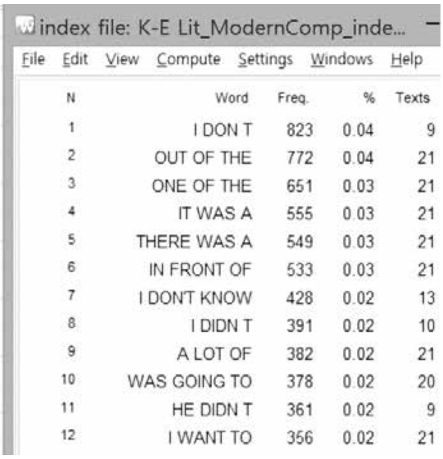
Figure 1: Three-word bundle lists for KTT and CET

KTT


index file: K-E Lit index_file
File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts
1	IN FRONT OF	619	0.04	20
2	OUT OF THE	452	0.03	20
3	IT WAS A	445	0.03	21
4	THERE WAS NO	349	0.02	20
5	DO YOU THINK	317	0.02	21
6	A LONG TIME	314	0.02	21
7	THE FIRST TIME	312	0.02	21
8	IT WAS THE	305	0.02	21
9	WHAT DO YOU	297	0.02	21
10	FRONT OF THE	271	0.02	20
11	AS IF HE	268	0.02	21
12	AS SOON AS	265	0.02	21

CET


index file: K-E Lit_ModernComp_inde...
File Edit View Compute Settings Windows Help

N	Word	Freq.	%	Texts
1	I DON T	823	0.04	9
2	OUT OF THE	772	0.04	21
3	ONE OF THE	651	0.03	21
4	IT WAS A	555	0.03	21
5	THERE WAS A	549	0.03	21
6	IN FRONT OF	533	0.03	21
7	I DON'T KNOW	428	0.02	13
8	I DIDN T	391	0.02	10
9	A LOT OF	382	0.02	21
10	WAS GOING TO	378	0.02	20
11	HE DIDN T	361	0.02	9
12	I WANT TO	356	0.02	21

Since KTT and CET differ in size, their raw frequencies cannot be compared directly. This problem is handled by normalizing the raw frequencies to one million words by using the simple formula below.

Normalized freq. = (raw frequency/corpus total word count) x 1,000,000
(Biber & Barbieri, 2007, p. 264, fn 4)

Also, there is the question of 'how many times should a word sequence occur in a corpus to be considered as a lexical bundle?' Setting such a frequency cut-off point is arbitrary, and researchers use different levels, ranging from ten per million (Biber et al., 1999, p. 994) to 40 per million (Biber et al., 2004, p. 376). Let us adopt the middle-of-the-road level of 30 per million. The normalized frequency of 30 per million equals 44 for KTT and 59 for CET in raw frequency. These cut-off points give us 623 lexical bundles for KTT and 601 for CET. The difference appears to be insignificant, but we need to take into account the fact that KTT is about 500,000 words fewer than CET. The difference is actually statistically highly significant in a log-likelihood test, as can be seen in Table 2.

Table 2: Comparison in number of bundles (with a cut-off of 30 per mil)

	KTT	CET	Log-likelihood test
No. of bundles with freq. of 30 per mil and above	623	601	p=2.273e-09 (df=1, G2=35.724) ** Cramer's V=0.003

The p-value, however, only tells us how likely it is to observe the differences in our data under certain assumptions and tells nothing about how strong the relationship is between the two categories of corpora and lexical bundles (Jenset, 2008, p. 11). One measure of effect size for log-likelihood statistics is Cramer's V. Cramer's V ranges between 0 and 1. Very much like correlation, a value close to 1 indicates strong significance and effect size, and a value close to 0 means no relationship (Jenset, 2008, p. 13). Cramer's V for our frequency data is 0.003, which is very close to 0. This means that the observed difference has little to do with the categories in our data.

Now, let us move on to comparison in high-frequency lexical bundles. Xiao (2010, p. 145) defines high frequency lexical bundles as 'those accounting for at least 0.01% of the respective corpus.' The lists in Figure 1 give us these percentages in the % column, making it easy to find out how many lexical bundles come within the 1% or even 2% high-frequency ranges in each corpus. The respective frequencies and the results of statistical tests are given in Table 3. We need to look at the normalized frequencies marked as 'n' in the table to compare the corpora directly. KTT are more frequent in both 1% and 2% high-frequency bundles, but the difference is statistically significant only in the 1% range. Even in that case, the effect size measured by Cramer's V is nearly zero, which makes the meaning of the p-value questionable.

Table 3: Comparison in numbers of high-frequency three-word lexical bundles

	KTT	CET	Log-likelihood test
1% bundles	n: 39 (r: 56)	n: 25 (r: 49)	p=0.024452 (df=1, G2=5.0623) * Cramer's V=0.001
2% bundles	n: 14 (r: 20)	n: 11 (r: 21)	p=0.25643 (df=1, G2=1.2879)

n = normalized count, r = raw count

The results of the frequency comparisons above, even if we go strictly by p-values, do not provide us with much confidence about whether KTT is truly more reliant on typical lexical bundles than CET, as would be predicted by the hypothesis of normalization. Additionally, it is uncertain whether the occurrence of a greater number of lexical bundles in KTT than in CET (above the cut-off of 30 per mil) can be used as supporting evidence for normalization. As was discussed in Section 2, typicality is the most important requirement of a linguistic feature to be a touchstone for testing normalization. In our case, typical lexical bundles would be those 1% or 2% ones that we tested in Table 3. The hypothesis of normalization would predict that KTT would rely more heavily on these bundles than CET, which is rejected by the results in Table 3. KTT's greater variety runs counter to the notion of normalization because a normalized translated text would be narrower in the range of choice, gravitating to the most typical options.

To sum up, our frequency analyses and significance tests have failed to clearly prove that our translated corpus is normalized in the use of lexical bundles vis-à-vis the non-translated corpus. This contrasts with Xiao (2010) and Lee (2013) who find greater use of typical lexical bundles in translated texts. The discrepancy may be attributable to the effect of language direction as Xiao (2010) and Lee (2013) looked at translation into Chinese and Korean respectively or to the effect of genres as Xiao analyzed a collection of text samples from diverse genres and Lee worked with journalistic texts. Or it could be that KTT and CET differ not so much in overall frequencies as in the types of lexical bundles they prefer and that there is wide variation across text samples which do not show up in general frequency tests. In fact, there is a serious problem in comparing frequencies in lexical bundles between translated and non-translated corpora without considering their functional types, as will be discussed in the next section. To address all these concerns, we now turn to multidimensional exploratory data analysis.

4.2 Exploratory Multivariate Analysis: PCA and HCA

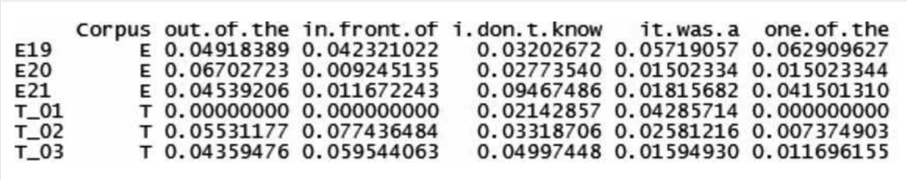
The term, exploratory multivariate analysis, basically says two things, first that the analysis addresses a dataset with multiple variables and, second, that it is designed to explore the structure of a dataset, instead of testing statistical significance. Several statistical methods fall under this category, such as principal component analysis (PCA), correspondence analysis (CA), exploratory factor analysis (EFA) and multidimensional scaling (MS). While differing in specific assumptions and mathematical formula, these methods all share the same goal of simplifying the structure of a multidimensional dataset by consolidating intercorrelated variables into a much smaller set of new variables that can account as effectively for the variation in the original data (Jenset & McGillivray, 2012, p. 304). In this section, we will use PCA, which will be complemented by another method of grouping variables, called 'hierarchical cluster analysis' (HCA). PCA is recommended as more suitable to corpus data than EFA (Jenset & McGillivray, 2012, pp. 306-308). CA is irrelevant to our numerical data as it takes categorical data as input.

Before progressing with our analysis, let us examine our dataset as illustrated in Figure 2. The list in Figure 2 is a part of our dataset, showing six rows in the middle and the first six columns. The first column with no label contains the row names representing the individual texts in KTT and CET. There are a total of 42 rows, the first 21 corresponding to the samples from CET, ranging from E01 to E21, and the next 21 matching the samples from KTT, from T_01 to T_21. The columns total 989. The first column, 'Corpus', is a categorical variable consisting of two values, E and T (short for CET and KTT respectively). The remaining 987 columns represent individual lexical bundles arranged in the descending order of total frequency. The numbers in the data frame in Figure 2 represent normalized frequencies of individual lexical bundles for individual texts.

The data was extracted from the corpora, using the *stylo* R package.

The package offers a suite of statistical analyses tailored to computational stylistics (or stylometry) and authorship attribution. One of the basic things it does in carrying out a statistical analysis is computing most frequent words or lexical bundles. Using this function, the 1,000 most frequent three-word lexical bundles were collected from the pool of KTT and CET texts. The *stylo* package can perform PCA, but it is rather limited for our purpose. We will just take the frequency list and use different R functions to carry out the analysis. Cross-checking of the principal component maps from the *stylo* analysis and the other R analysis revealed no significant difference. The 1,000 bundle list was manually scanned, and those obviously associated with a particular text and, thus, occurring exclusively in one or two texts, such as *yu jin said* (*yu jin* being the name of a character in one of the translated Korean novels), were eliminated. This reduced the list to 989. Even this number is impossibly too big to subject to PCA for 42 sample texts. Since we are concerned with the most typical bundles, we will take the top 70. This gives us a 42x71 data frame, including one categorical variable distinguishing KTT and CET.

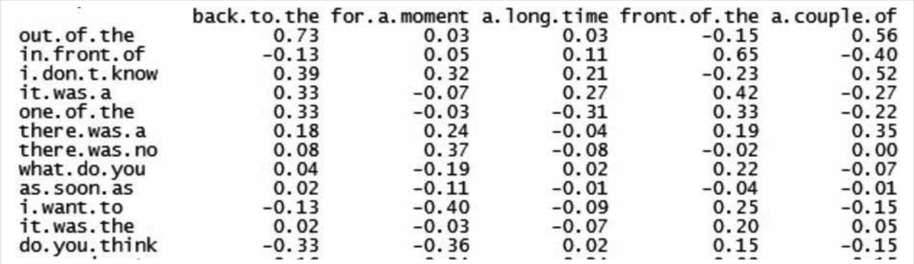
Figure 2: List of three-word lexical bundles from both KTT and CET



```
      Corpus out.of.the in.front.of i.don.t.know   it.was.a  one.of.the
E19        E 0.04918389 0.042321022   0.03202672 0.05719057 0.062909627
E20        E 0.06702723 0.009245135   0.02773540 0.01502334 0.015023344
E21        E 0.04539206 0.011672243   0.09467486 0.01815682 0.041501310
T_01       T 0.00000000 0.000000000   0.02142857 0.04285714 0.000000000
T_02       T 0.05531177 0.077436484   0.03318706 0.02581216 0.007374903
T_03       T 0.04359476 0.059544063   0.04997448 0.01594930 0.011696155
```

Each of the 70 bundles constitutes a variable, which can be thought of as a dimension. If we just take two variables and use them as the x and y axes, we can plot the 42 texts in the two-dimensional space and see how close or distant they are from each other individually or as groups (i.e. translated vs. non-translated). Plotting them in a 50-dimension space is not only impossible, but, even if we could, the graph would be beyond any meaningful interpretation. One characteristic of multidimensional datasets is that they are likely to have redundancy. This means that some variables are correlated with each other. Take the partial correlation matrix of lexical bundles for CET in Figure 3. Some pairs of lexical bundles show high correlations, e.g. 0.73 between *out of the* and *back to the*. What PCA does is look for such closely correlated variables in the data and packs them into a smaller set of more fundamental measures, which are called 'principal components' (PCs). The PCs are then used to produce informative graphs about the data. For more technical and detailed information about PCA, refer to Raykov and Marcoulides (2008: chpt. 7) for overall theory and principles, Everitt and Hothorn (2011: chpt. 3) and Husson et al. (2011: chpt. 1) for PCA in R, Baayen (2008: chpt. 5) for application of PCA to language study and Jenset and McGillivray (2012) for a case of PCA in a translation study.

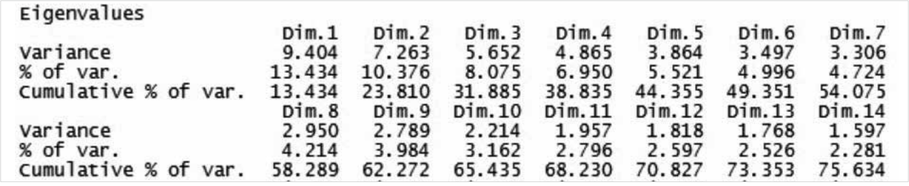
Figure 3: A partial correlation matrix for CET



```
               back.to.the for.a.moment a.long.time front.of.the a.couple.of
out.of.the            0.73         0.03        0.03        -0.15        0.56
in.front.of          -0.13         0.05        0.11         0.65       -0.40
i.don.t.know          0.39         0.32        0.21        -0.23        0.52
it.was.a              0.33        -0.07        0.27         0.42       -0.27
one.of.the            0.33        -0.03       -0.31         0.33       -0.22
there.was.a           0.18         0.24       -0.04         0.19        0.35
there.was.no          0.08         0.37       -0.08        -0.02        0.00
what.do.you           0.04        -0.19        0.02         0.22       -0.07
as.soon.as            0.02        -0.11       -0.01        -0.04       -0.01
i.want.to            -0.13        -0.40       -0.09         0.25       -0.15
it.was.the            0.02        -0.03       -0.07         0.20        0.05
do.you.think         -0.33        -0.36        0.02         0.15       -0.15
```

In our PCA analysis, we will use the *PCA* function from the FactoMineR package in R, which is well explained in Husson et al. (2011: chpt. 1). There are other R functions that can perform PCA, such as *prcomp* and *princomp* from the *stats* package and *principal* from the *psych* package. The biggest merit of the FactoMineR *PCA* is that it makes it easy to plot PCA results by group and that the PCA results can be directly fed into another FactoMineR function to perform cluster analysis.

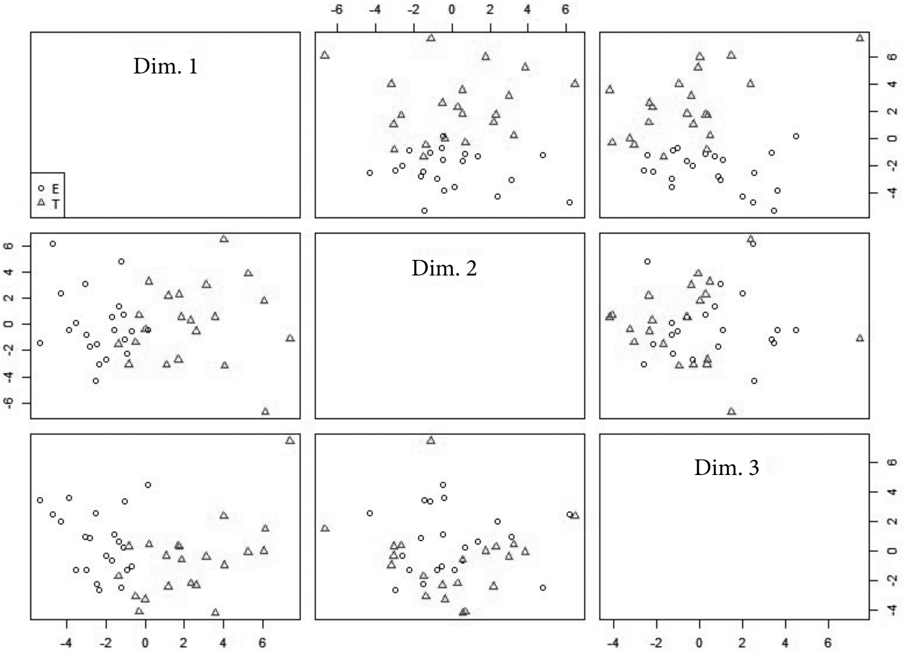
Figure 4: Information about PCs



```
Eigenvalues
                       Dim.1   Dim.2   Dim.3   Dim.4   Dim.5   Dim.6   Dim.7
Variance               9.404   7.263   5.652   4.865   3.864   3.497   3.306
% of var.             13.434  10.376   8.075   6.950   5.521   4.996   4.724
Cumulative % of var.  13.434  23.810  31.885  38.835  44.355  49.351  54.075
                       Dim.8   Dim.9  Dim.10  Dim.11  Dim.12  Dim.13  Dim.14
Variance               2.950   2.789   2.214   1.957   1.818   1.768   1.597
% of var.              4.214   3.984   3.162   2.796   2.597   2.526   2.281
Cumulative % of var.  58.289  62.272  65.435  68.230  70.827  73.353  75.634
```

One of the results of PCA is a list of PCs or Dimensions as shown in Figure 4. As explained earlier, PCs can be thought of as new variables that pack a series of original values with relative weights. Theoretically, there could be as many PCs as the original variables. In our case, *PCA* produced 41 PCs. The PCs are ordered in such a way that the first PC accounts for the largest portion of the variance in the original data, the second PC the second largest portion and so on. Figure 4 lists the first 14 of the PCs created by PCA from our variables. The numbers in the first row (*variance*) are eigenvalues, which are numeric estimations of how much variation or inertia in the original data each PC explains. The second row represents the same information in percentages, and the third are cumulative percentages. The first two PCs explain about 24 percent of variance, which is low but not terribly bad given the large number of variables. Besides, in a PCA based on correlations as is our case, these percentages are less meaningful (Jenset & McGillivray, 2012, p. 314).

Figure 5: Individual texts in PC spaces among Dimensions 1, 2 and 3



The next step in regular PCA is to determine how many PCs to choose for interpretation. This is necessary if we intend to use the PCs as variables for other subsequent statistical analyses such as regression. But since we are mostly concerned with exploring the underlying relationships in our data, let us choose the first three and scatterplot them in a matrix to see how individual texts from the two corpora are positioned in the three PC subspaces. The graph is provided in Figure 5. We will use the lower three plots above the diagonal line formed by the three Dimension label boxes. In the plots, circles represent CET texts, while triangles locate KTT texts. The plot between Dim. 1 and Dim. 2 successfully separates the texts between the two corpora on the Dim. 1 axis. The situation is similar with the plot between Dim. 1 and Dim. 3. In contrast, the texts are rather intermingled in the Dim. 3 space. Since the pair of Dim. 1 and 2 account for a greater portion of variance than Dim. 1 and Dim. 3, let us focus on the Dim. 1-2 space.

The results of PCA contain the coordinates of individual texts and variables for each PC, which allows them to be plotted in the PC subspaces we examined above.

Figure 6: Plot of individual texts on Dim. 1 and Dim. 2, with confidence ellipses

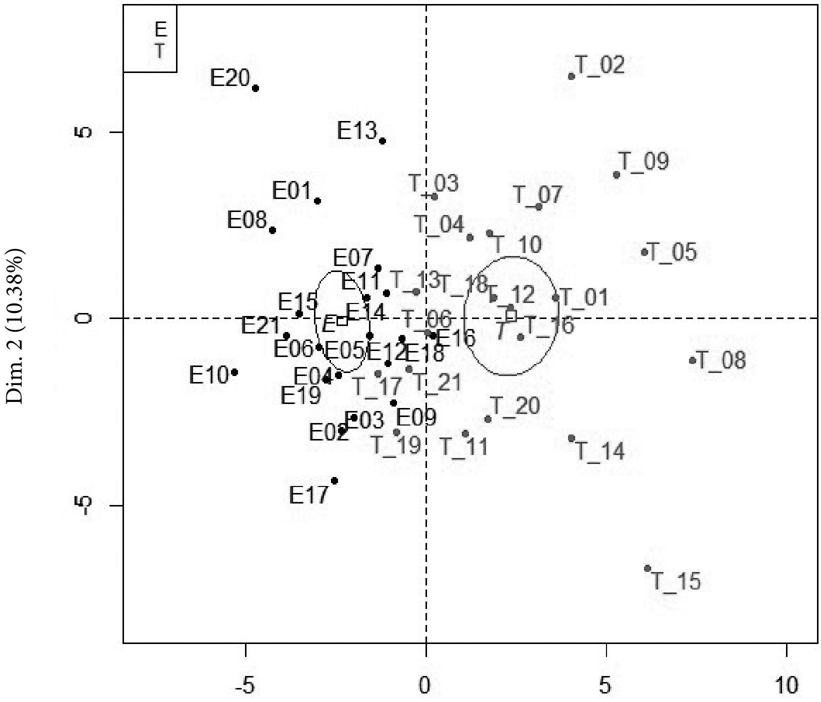


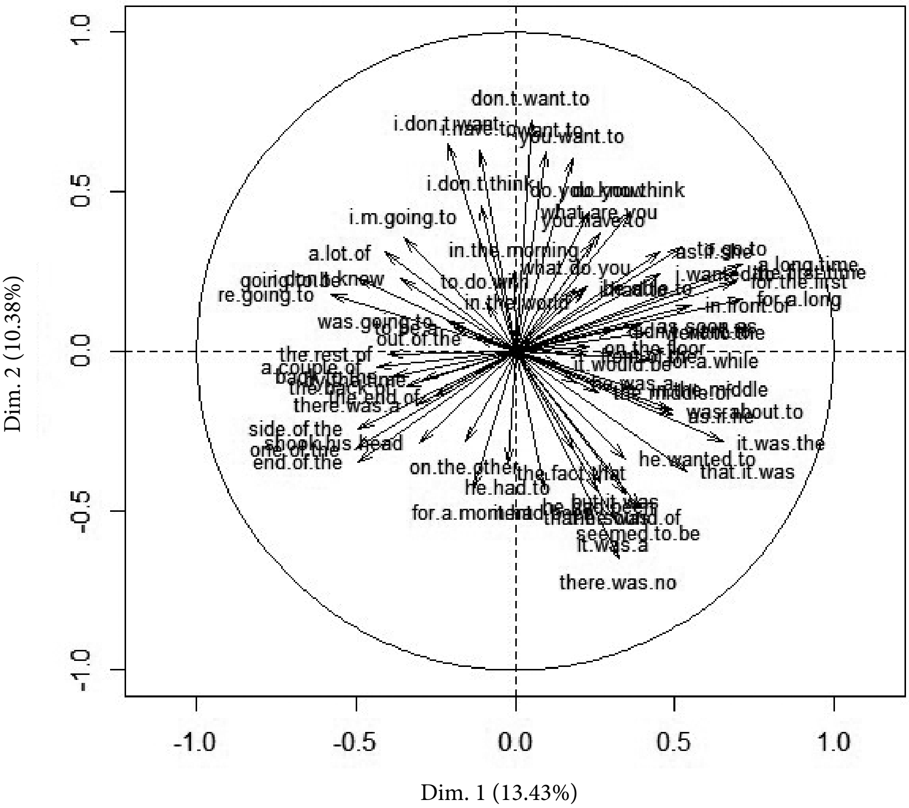
Figure 6 above is a plot of individual texts in the Dim. 1-2 space. It is a blowup of the Dim. 1-2 plot in Figure 5. Again, we can clearly see that the texts from the two corpora are separated on Dim. 1, with the KTT texts predominantly on the right side of the zero point, the CET texts on the left side. Dim. 2 on the other hand, is not discriminatory. We conclude from this that Dim. 1 mostly represents the distinction between our two corpora. The oval circles are called '95% confidence ellipses'. They provide visualization of whether the two groups differ significantly (Husson et al., 2011, p. 36). Much like 95% confidence intervals in regular statistics, statistical significance is achieved when the two ellipses do not overlap.

In Figure 6, the two ellipses are not only clearly separated from each other, but they are centered right on the Dim. 1 axis. This is another strong proof that Dim. 1 is mostly related to the distinction between the two corpora. With that fact established, we now need to find out which original variables (i.e. lexical bundles) are positively or negatively loaded on Dim. 1. This will enable us to identify lexical bundles that distinguish the two corpora.

For this purpose, we plot variables in the same Dim. 1-2 space. The plot in Figure 7 provides a visual representation of positive and negative links between variables and Dimensions. The arrow point of each line locates the position of a specific variable on the plane, and the length of the line represents the strength of its correlation with the Dimensions. Four clouds of variables are discernible. First, there is a group of variables positively linked with Dim. 1, on its right side. These are variables associated with KTT. The second group is negatively linked with Dim. 1, on its left side. These variables are associated with CET. The third and fourth smaller groups are positively and negative linked with Dim. 2, located on either side of the Dimension. These variables are on the borderline between KTT and CET, with no distinct association with either. Let us focus on the first and second groups of variables because they are the ones loaded on Dim. 1 that distinguishes KTT and CET.

Judging from the density of the clouds, it appears that more variables are positively linked with Dim. 1 than negatively linked. But the graph is rather too crowded to interpret easily. The *dimdesc* function from the FactoMineR package comes in handy in such a situation as it provides us with a list of the variables significantly correlated with each PC. The list in Figure 8 shows that 24 bundles are positively correlated with Dim. 1, and 13 negatively correlated. Again, the positively correlated ones are representative of KTT, while the negatively correlated ones characterize CET. The difference between 24 and

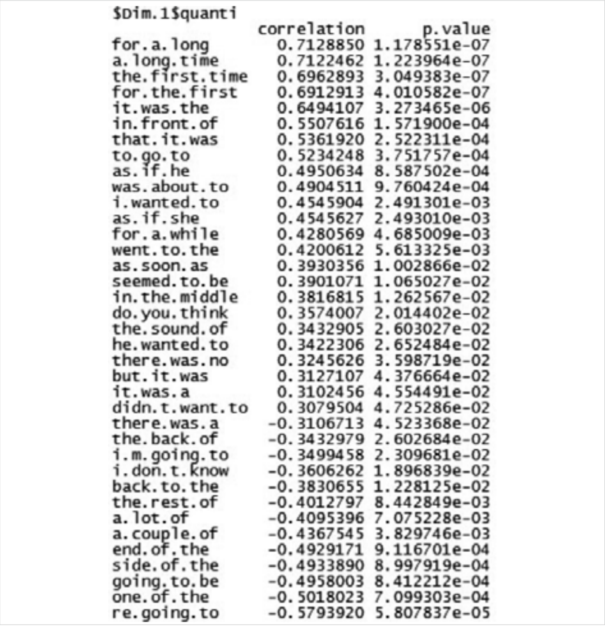
Figure 7: Plot of variables on Dim. 1 and Dim. 2



13 seems big (p=0.0059927, df=1, G2=.5525) but as discussed in Section 4.1, this statistical significance is spurious as Cramer's V is extremely small at 0.001. What is more interesting is the types of bundles associated with each corpus.

On the side of KTT, lexical bundles related with time reference stand out—*the first time, a long time, for a long, for the first, as soon as* and *for a while.* Among them, the first four are actually interrelated as it is apparent that they are parts of longer lexical bundles, namely, *for the first time* and *for a long time.* It is a property of lexical bundles that shorter ones are often incorporated in longer ones (Biber et al., 1999, p. 990). For CET, lexical bundles of location appear prominent—*side of the, end of the, back to the* and *the back of.* KTT also includes two place bundles, *in front of* and *in the middle.* The latter one may be a time bundle, too. These lexical bundles of time and place are what Biber and Barbieri (2007, pp. 271-272) call 'referential bundles'. These bundles make direct reference to concrete or abstract entities or actions. Many other bundles on the CET side are also of this type, namely, *the rest of, a lot of, a couple of* and *one of the.* There are two additional referential bundles on the KTT side, which are *to go to* and *the sound of.* These bundles are related to propositional elements in statements, which are primarily decided by the author, not by the translator. In this sense, they are 'noise', irrelevant to the study of translation universals, which should be excluded from analysis (Baker, 2004, p. 174). This is also another reason that a translated corpus cannot be compared with a non-translated one simply in overall frequency of lexical bundles, high-frequency or not, because we do not know how much of any observed difference can be attributed to the act of translating.

Figure 8: List of lexical bundles correlated with Dim. 1



```
$Dim.1$quanti
                correlation      p.value
for.a.long        0.7128850 1.178551e-07
a.long.time       0.7122462 1.223964e-07
the.first.time    0.6962893 3.049383e-07
for.the.first     0.6912913 4.010582e-07
it.was.the        0.6494107 3.273465e-06
in.front.of       0.5507616 1.571900e-04
that.it.was       0.5361920 2.522311e-04
to.go.to          0.5234248 3.751757e-04
as.if.he          0.4950634 8.587502e-04
was.about.to      0.4904511 9.760424e-04
i.wanted.to       0.4545904 2.491301e-03
as.if.she         0.4545627 2.493010e-03
for.a.while       0.4280569 4.685009e-03
went.to.the       0.4200612 5.613325e-03
as.soon.as        0.3930356 1.002866e-02
seemed.to.be      0.3901071 1.065027e-02
in.the.middle     0.3816815 1.262567e-02
do.you.think      0.3574007 2.014402e-02
the.sound.of      0.3432905 2.603027e-02
he.wanted.to      0.3422306 2.652484e-02
there.was.no      0.3245626 3.598719e-02
but.it.was        0.3127107 4.376664e-02
it.was.a          0.3102456 4.554491e-02
didn.t.want.to    0.3079504 4.725286e-02
there.was.a      -0.3106713 4.523368e-02
the.back.of      -0.3432979 2.602684e-02
i.m.going.to     -0.3499458 2.309681e-02
i.don.t.know     -0.3606262 1.896839e-02
back.to.the      -0.3830655 1.228125e-02
the.rest.of      -0.4012797 8.442849e-03
a.lot.of         -0.4095396 7.075228e-03
a.couple.of      -0.4367545 3.829746e-03
end.of.the       -0.4929171 9.116701e-04
side.of.the      -0.4933890 8.997919e-04
going.to.be      -0.4958003 8.412212e-04
one.of.the       -0.5018023 7.099303e-04
re.going.to      -0.5793920 5.807837e-05
```

This is not to say that the translators have nothing to do with referential bundles in KTT. Translators have some leeway to phrase a propositional element from the source text in different ways. For instance, the phrase *for a long time* could be re-expressed as other similar adverbial phrases such as *for an extensive period of time* or *for a long stretch of time.* Or, they could be phrased as an adjective as in *having a lengthy conversation* instead of *talking for a long time.* In this sense, the translators may be partially responsible for the prominence of the time-related lexical bundles in KTT. Yet, their occurrences are so strongly influenced by the propositional content of the source texts that they cannot be used as a credible basis for differentiating translated texts from non-translated texts.

Lexical bundles, relatively free from source text interference, are what are called 'stance bundles' and 'discourse bundles' (Lee, 2013, p. 383). Stance bundles express epistemic modality, evaluation or attitude (Biber et al., 2004, p. 393). They cover *i wanted to*, *seemed to be, do you think* and *didn't want to* on the KTT side and *re going to, going to be* and *i'm going to* and *i don't know* on the CET side. Discourse bundles point to overall discourse structure, serving to frame clauses and sentences. KTT contains many of this type—*it was the, that it was, as if he, was about to, as if she, there was no, but it was* and *it was a*. In sharp contrast, CET has only one—*there was a*.

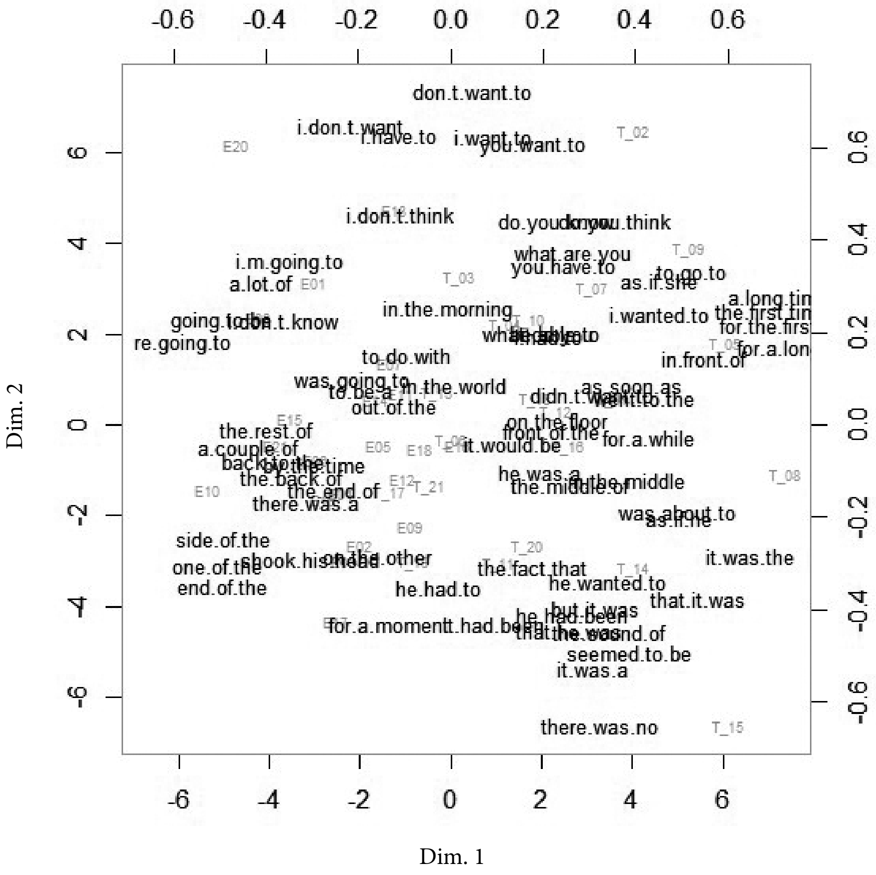
The typological analysis of lexical bundles above reveals that the principal source of distinction between KTT and CET is in the use of discourse bundles. KTT tends to overuse them when compared with CET. Take *as if he* and *as if she* for instance. The *as if* phrase is a typical means of expressing analogy in English fiction. But the fact that this phrase appears strongly correlated with KTT texts means that they use them far more significantly than average English novels. This observation can be ascertained by t-tests as in Table 4. The tests show that KTT is greater than CET in the mean frequency of both phrases and that the differences are statistically significant. The exaggerated use of discourse bundles by KTT is a clear sign of normalization. Our analysis reveals that normalization in our data occurs mostly at the structural level with an excessive reliance on typical structure-framing phrases. This important characteristic of the underlying relationship between the two corpora went undetected in our earlier frequency tests.

Table 4: T-tests on *as if he* and *as if she*

	Mean frequency		T-test
as if she	KTT: .02027329	CET: 0.01003841	p-value = 0.01947 (t = -2.4499, df = 34.786)
as if he	KTT:.014436220	CET: .006188142	p-value = 0.02043 (t = -2.43, df = 34.64)

Now, let us turn to the second issue Baker (2004, pp. 181-182) raises in connection with the study of translational language, that is stylistic variation across translated texts and translators. Here, we are more concerned with stylistic characteristics of individual texts and translators, rather than their overall patterns. To explore this issue, we can use the results of PCA to produce a biplot. The graph in Figure 9 plots the individual texts of both corpora and the lexical bundles in the same PC space. This allows us an easy visual inspection of how strongly individual texts and variables are correlated with each other on the basis of the distance between them. The closer a text and a variable are located to each other, the stronger they are correlated. By the same logic, the closer two texts are located to each other, the more they

Figure 9: Biplot of texts and lexical bundles on Dim. 1 and Dim. 2



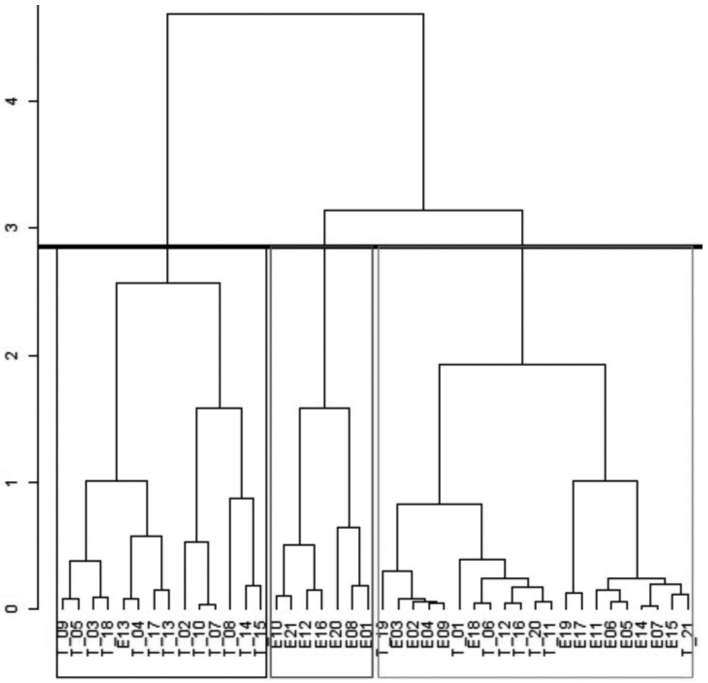
resemble each other. For example, the prominent set of time bundles in KTT form a cluster near the right-hand edge of the plot between 0.2 and 0.4. They are right on top of T_05, while distant from other translated texts. This means that these lexical bundles come predominantly from this single translated text, exerting an unbalanced influence on the entire data. In comparison, the discourse bundles, *as if he* and *as if sh*e are not associated with any particular text(s) but rather surrounded by clusters of translated texts. This means that these lexical bundles are broadly shared by the translated texts, making them a genuine stylistic marker of the translated corpus.

The biplot is particularly useful if we want to investigate stylistic difference among different translations (and translators) of the same source text. The two texts in KTT, T_20 and T_21, are different translations of the same Korean novel. T_20 was rendered by a Korean translator and published in 1980. T_21 came out in 2005, translated by a team of American and Korean translators. In relative terms, T_21 reads more naturally, if we define 'natural' as resembling authentic English novels. It departs frequently from the wording and structure of the ST, using shorter and, often, incomplete sentences, which creates a dynamic narrative flow. In comparison, T_20 is more formal and closer to the ST. These stylistic differences between the two texts are reflected in their relative positions in the PC space. T_21 is somewhat mingled with non-translated texts, locating itself on the borderline between the two corpora. This contrasts with T_21, which is rather removed from the borderline and nestled among other translated texts.

Interestingly, T_20 is strongly correlated with the bundle, *the fact that*. This particular phrase occurs nine times in T_20, while only one instance shows up in T_21. The heavier use of this rather formal structure, thus, appears to be a stylistic idiosyncrasy of the translator of T_20. The text is also located close to other bundles containing the *that* complementizer, namely *that it was* and *that he was*. They are on the opposite side of T_20 from where T_21 is located. This suggests that T_20 is distinguished from T_21 by more frequent use of *that*-embedded clauses. This relates to the fact that T_20 tends to use longer sentences than T_21, with its average number of words per sentence being 12.38 as opposed to 11.87 of T_21.

As the final step in our multidimensional analysis, let us carry out hierarchical cluster analysis (HCA). This analysis is performed by calling the *HCPC* function from the FactoMineR package. The analysis creates a tree-

Figure 10: A cluster tree of individual texts



shaped diagram, called 'dendrogram', as shown in Figure 10. The tree in Figure 10 groups closely correlated individual texts under the same branch, progressively as we go from bottom to top. The thicker horizontal line in the middle is where *HCPC* cuts the tree, indicating the level where the texts could be clustered into groups, with meaningful differences in variance among them. The analysis has decided on three clusters of texts, represented by three boxes enclosing the respective texts. The boxes on the left and in the center are composed exclusively of translated and non-translated texts respectively. But the last box, the one on the right, is a blend of both translated and non-translated texts.

While *HCPC* has chosen three clusters, we can look inside each cluster for further information about how individual texts are interrelated within that group. Particularly, the blending group warrants close inspection. We find three subgroups inside this cluster. The translated texts are mostly located in the second clump, indicating even within the blending cluster translated texts tend to hang together as a group. Interestingly, T_21, which we compared with T_20 as translations of the same source text above, does not belong in this group and instead aligns itself with other non-translated texts in the third subgroup. This is consistent with our earlier observation from the biplot in Figure 9 that T_21 is more akin to authentic English novels than T_20 in the use of lexical bundles.

The most important message we get from the cluster analysis is that the text samples in our two corpora do not split neatly into two groups as the plot in Figure 6 might lead us to believe. If we take out the group variable and let the texts cluster freely among themselves, more than a third of the translated texts blend with non-translated texts. It is the remaining 13 texts that are chiefly responsible for the differences that show up in general patterns. In fact, if we take another look at the plot in Figure 6 without considering the group values, the majority of texts form a cloud together in the center of the space. Among this cluster, the difference between translated and non-translated texts is very small. It is the minority of wayward texts outside the periphery of this cloud that pull the centers of the translated and non-translated groups away from each other. The influence of these wayward texts appears to be greater with KTT as they are more distant from the cloud. This indicates that KTT has greater variance among its texts than CET. This is proven by the fact that the KTT texts have greater standard deviations in their coordinates on both Dimensions of the PC plot than the CET, with 2.474163 vs. 1.432025 on Dim. 1 and 2.972591 vs. 2.530990 on Dim. 2. Incidentally, this runs counter to the hypothesis of leveling-out, which says translated texts are more like one another than non-translated texts (Baker, 1996, p. 184). In our data, the translated texts are more dispersed than the non-translated texts.

5. Conclusion

Our analysis in the preceding section has proven the usefulness of multivariate exploratory analysis for translation research. Translation corpora, like any other language corpora, are multidimensional in nature. They contain texts from different sources, classified at different levels (genres, translators, source languages, etc.) with different linguistic features. Generic frequency analysis can provide us with information about general patterns in the data. But these patterns often mask intricate underlying relationships, which may be more important and informative to us. Moreover, indiscriminate application of frequency analysis and statistical significance testing to selective data can skew the reality. Multivariate exploratory analysis can address many of these concerns successfully. As the name suggests, its primary aim is to explore the data, instead of proving some preconceived assumptions, and it rewards the researcher with discoveries and revelations. In our analysis, PCA has revealed that the difference between translated and non-translated texts lies in the use of a particular type of lexical bundles, rather than in their overall distribution. This has led us to the conclusion that the translated texts are structurally normalized by an exaggerated use of some typical structure-framing phrases. Additionally, PCA has provided us with valuable insights into how individual translated texts and translators are related to one another. This has led us to discover some distinct stylistic features that set particular translated texts and translators apart from others. Finally, cluster analysis has identified the presence of a subgroup of translated texts that behave like non-translated texts, suggesting that the division between translated and non-translated texts is not as clear-cut as significant p-values of frequency tests might lead us to believe.

Funding

This work was supported by the Hankuk University of Foreign Studies Research Fund of 2022.

References

Baayen, R. Harald. (2008). *Analyzing Linguistic Data: A Practical Introduction to Statistics Using R*. Cambridge University Press.

Baker, Mona. (2004). A corpus-based view of similarity and difference in translation. *International Journal of Corpus Linguistics, 9*(2), 167-193. https://doi.org/10.1075/ijcl.9.2.02bak

Baker, Mona. (1996). Corpus-based translation studies: The challenges that lie ahead. In Harold Somers (Ed.), *Terminology, LSP, and Translation: Studies in Language Engineering* *in Honour of Juan C. Sager* (pp. 175-186). John Benjamins.

Biber, Douglas and Federica Barbieri. (2007). Lexical bundles in university spoken and written registers. *English for Specific Purposes, 26*(3), 263-286. https://doi.org/10.1016/j.esp.2006.08.003

Biber, Douglas and Susan Conrad. (1999). Lexical bundles in conversation and academic prose. In Hilde Hasselgård & Signe Oksefjell (Eds.), *Out of Corpora: Studies in Honour of Stig Johansson* (pp. 181-189). Rodopi.

Biber, Douglas, Susan Conrad and Viviana Cortes. (2004). If you look at ...: Lexical bundles in university teaching and textbooks. *Applied Linguistics, 25*(3), 371-405. https://doi.org/10.1093/applin/25.3.371

Biber, Douglas, Stig Johansson, Geoffrey Leech, Susan Conrad and Edward Finegan. (1999). *Longman Grammar of Spoken and Written English.* Pearson Educated Ltd.

Burrows, John. (2002). 'Delta': A measure of stylistic difference and a guide to likely authorship. *Literary and Linguistic Computing, 17*(3), 267-287.

Burrows, John. (1987). Word-patterns and story-shapes: The statistical analysis of narrative style. *Literary and Linguistic Computing, 2*(2), 61-70.

De Sutter, Gert, Isabelle Delaere and Koen Plevoets. (2012). Lexical lectometry in corpus-based translation studies: Combining profile-based correspondence analysis and logistic regression modeling. In Michael P. Oakes & Meng Ji (Eds.), *Quantitative Methods in Corpus-based Translation Studies: A Practical Guide to Descriptive Translation Research* (pp. 325-346). John Benjamins. https://doi.org/10.1075/scl.51.13sut

Everitt, Brian and Torsten Hothorn. (2011). *An Introduction to Applied Multivariate Analysis with R.* Springer. https://doi.org/10.1007/978-1-4419-9650-3

Forsyth, Richard S. and Phoenix W. Y. Lam. (2014). Found in translation: To what extent is authorial discriminability preserved by translators? *Literary and Linguistic Computing, 29*(2), 199-217. https://doi.org/10.1093/llc/fqt018

Grabowski, Łukasz. (2013). Interfacing corpus linguistics and computational stylistics: Translation universals in translational literary Polish. *International Journal of Corpus Linguistics, 18*(2), 254-280. https://doi.org/10.1075/ijcl.18.2.04gra

Husson, François, Sébastien Lé and Jérôme Pagès. (2011). *Exploratory Multivariate Analysis by Example Using R.* CRC Press.

Hyland, Ken. (2008). As can be seen: Lexical bundles and disciplinary variation. *English for Specific Purposes, 27*(1), 4-21. https://doi.org/10.1016/j.esp.2007.06.001

Jenset, Gard B. (2008). Basic R for corpus linguistics [Seminar handout]. University of Bergen. Retrieved February 12, 2014 from http://folk.uib.no/gje037/tutorialR.pdf

Jenset, Gard B. and Barbara McGillivray. (2012). Multivariate analyses of affix productivity in translated English. In Michael P. Oakes & Meng Ji (Eds.), *Quantitative Methods in Corpus-based Translation Studies: A Practical Guide to Descriptive Translation Research* (pp. 301-324). John Benjamins.

Kenny, Dorothy. (2001). *Lexis and Creativity in Translation: A Corpus-based Study.* St. Jerome Publishing.

Kenny, Dorothy. (2000). Translators at play: Exploitations of collocational norms in German English. In Dodd Bill (Ed.), *Working with German Corpora: With a Foreword by John Sinclair* (pp. 143-160). University of Birmingham Press.

Kenny, Dorothy. (1999). The German-English parallel corpus of literary texts (GEPCOLT): A resource for translation scholars. *Teanga, 1*(18), 25-42.

Kenny, Dorothy. (1998). Creatures of habit? What translators usually do with words. *Meta, 43*(4), 515-523. https://doi.org/10.7202/003302ar

Lee, Changsoo. (2021). How do machine translators measure up to human literary translators in stylometric tests? *Digital Scholarship in the Humanities,* 1-17. https://doi.org/10.1093/llc/fqab091

Lee, Changsoo. (2013). Using lexical bundle analysis as discovery tool for corpus-based translation research. *Perspectives, 21*(3), 378-395. https://doi.org/10.1080/0907676x.2012.657655

Malmkjaer, Kirsten. (1998). Love thy neighbour: Will parallel corpora endear linguists to translators? *Meta, 43*(4), 534-541. https://doi.org/10.7202/003545ar

Raykov, Tenko and George A. Marcoulides. (2008). *An Introduction to Applied Multivariate Analysis.* Routledge.

Rybicki, Jan. (2006). Burrowing into translation: Character idiolects in Henryk Sienkiewicz's trilogy and its two English translations. *Literary and Linguistic Computing, 21*(1), 91-103. https://doi.org/10.1093/llc/fqh051

Rybicki, Jan and Magda Heydel. (2013). The stylistics and stylometry of collaborative translation: Woolf's Night and Day in Polish. *Literary and Linguistic Computing, 28*(4), 708-717. https://doi.org/10.1093/llc/fqt027

Scott, Mike and Christopher Tribble. (2006). *Textual Patterns: Key Words and Corpus Analysis in Language Education.* John Benjamins. https://doi.org/10.1075/scl.22

Toury, Gideon. (1995). *Descriptive Translation Studies and Beyond.* John Benjamins.

Toury, Gideon. (1980). *In Search of a Theory of Translation.* Porter Institute for Poetics and Semiotics, Tel Aviv University.

Venuti, Lawrence. (1995). *The Translator's Invisibility.* Routledge. https://doi.org/10.4324/9780203360064

Xiao, Richard. (2010). Idioms, word clusters, and reformulation markers in translational

Chinese: Can "translation universals" survive in Mandarin? In Xiao Richard (Ed.), *Proceedings of the International Symposium on Using Corpora in Contrastive and Translation Studies* (pp. 1-40). http://www.lancs.ac.uk/fass/projects/corpus/UCCTS2010Proceedings

Professional Profile

Changsoo Lee is Professor at the Graduate School of Interpretation & Translation, Hankuk University of Foreign Studies, Seoul, Republic of Korea. He received his PhD in Applied Linguistics from Boston University in 1996, with specialization in discourse and conversation analysis. His academic interest and research have focused on translation from the perspectives of systemic functional linguistics, discourse analysis, corpus linguistics and semiotics.