Stylistic Comparison Between Fiction Written in English and Fiction Translated into English

Jae Kim

Introduction

The motivation for this study comes from a 2011 study by James M. Hughes and others called "Quantitative patterns of stylistic influence in the evolution of literature." The study examined whether there was any evidence to support the notion of a literary "style of time" and whether authors writing in the same time period were more similar in style than authors writing in different time periods. This study will explore a variation to the Hughes study and ask whether there is a significant difference in the styles of literature written in English (which I will sometimes refer to as non-translated literature) and literature translated into English (which I will sometimes refer to as translated literature).

For example, there are students in creative writing classes who complain, as soon as they learn that a short story has been translated from another language, that the story "sounds weird." But can a reader really tell whether they're reading a translated text? It doesn't seem impossible that this might indeed be the case. Perhaps translators are less stylistically concerned than original authors, and translations start to sound similar to one another. This study attempts to examine whether such a similarity really exists.

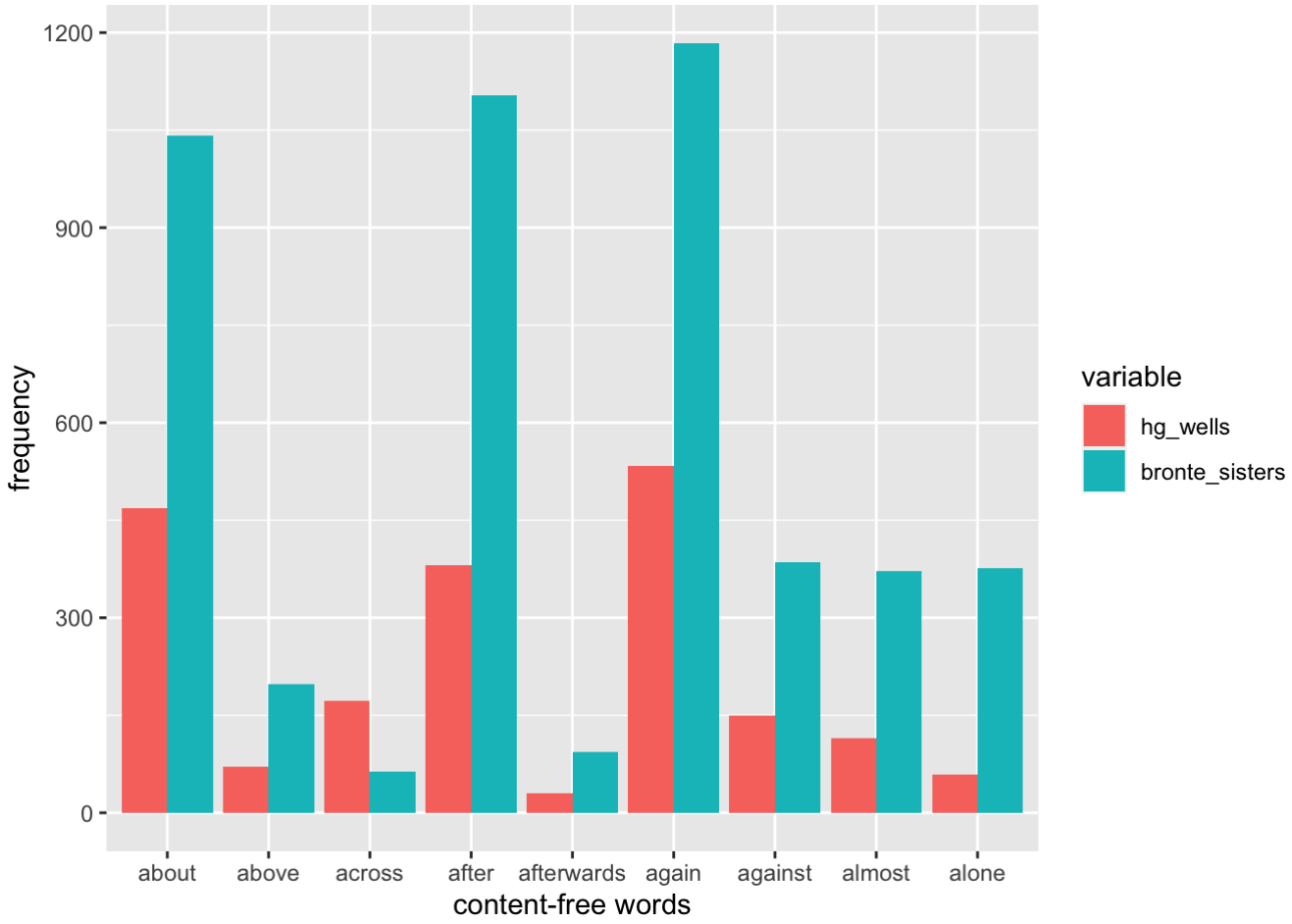
Data

The data comes from Project Gutenberg. It has its limitations, as I will discuss in the concluding section, but it is one of the only repositories of literature that are available for free. Since Hughes et al. have also used Project Gutenberg, the results of this study can easily be compared to theirs.

Since the Project Gutenberg r package doesn't identify translators as such, I've manually scraped their web page that contains the list of authors, where they also list the translators. I gathered the set of works that are labeled "(English) (as Translator)". This is the set of translated texts. I then subtracted this set from the full set of books whose language is labeled as English, which can be obtained via the r package. The remainder is the set of non-translated texts. I further filtered each set so that only those works considered to be fiction remain in the sets. I found there are **787** translated works of fiction and **12454** non-translated works of fiction.

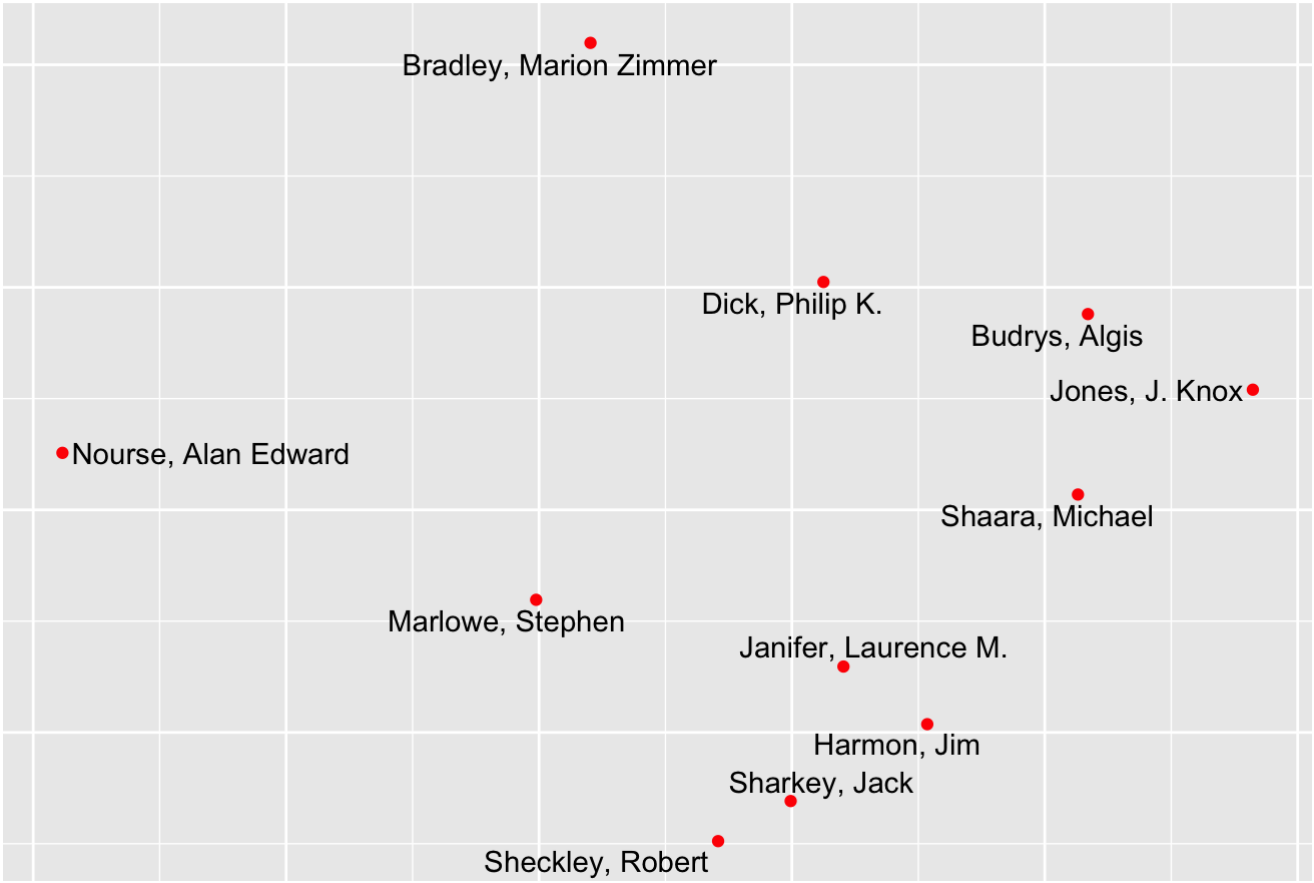
Methods

As Hughes et al. have done in their study, I used content-free words, and I borrow from their list of 307 such words that are considered devoid of content. They include, but are not limited to, prepositions, conjunctions, pronouns, and articles. As a point of illustration, below is a histogram containing the frequency of occurrence of ten content-free words in four novels by H.G. Wells (*The Time Machine*, *The War of the Worlds*, *The Invisible Man*, and *The Island of Doctor Moreau*) and in five novels by the Bronte sisters (*Jane Eyre*, *Wuthering Heights*, *The Tenant of Wildfell Hall*, *Villette*, and *Agnes Grey*).



Using the distribution of frequencies of the content-free words, one can visualize the stylistic difference among all the authors. The chart below, for example, arranges the authors in terms of their stylistic differences, using a method called Multi-Dimensional Scaling (MDS), treating each content-free word as a different factor that determines the authors' difference from one another. What multidimensional scaling allows one to do is to collapse the many dimensions in which these authors stand apart from one another onto, in this case, two dimensions, so that we can have a visual representation. I've also included the corresponding table with content-free words and counts. I'm only using the ten content-free words I've used above, and I have selected only those authors born after 1927, so that, for the sake of visualization, the number of represented authors is reasonably small.

MDS Representation of 11 authors from Projecgt Gutenberg



rownames	about	above	across	after	afterwards	again	against	almost	alone
Sheckley, Robert	314.00	42.00	60.00	185.00	1.00	209.00	79.00	75.00	29.00
Bradley, Marion Zimmer	248.00	45.00	89.00	169.00	0.00	398.00	120.00	104.00	71.00
Janifer, Laurence M.	233.00	7.00	15.00	130.00	0.00	190.00	56.00	48.00	43.00
Nourse, Alan Edward	564.00	54.00	241.00	284.00	0.00	578.00	198.00	234.00	51.00
Budrys, Algis	58.00	12.00	22.00	57.00	1.00	93.00	36.00	35.00	10.00
Sharkey, Jack	283.00	21.00	52.00	135.00	0.00	187.00	75.00	50.00	17.00
Marlowe, Stephen	362.00	25.00	90.00	201.00	9.00	338.00	107.00	140.00	56.00

Dick, Philip K.	165.00	60.00	91.00	157.00	1.00	211.00	93.00	78.00	21.00
Jones, J. Knox	0.00	9.00	1.00	2.00	0.00	0.00	0.00	2.00	0.00
Harmon, Jim	202.00	17.00	38.00	105.00	1.00	111.00	60.00	61.00	31.00
Shaara, Michael	93.00	11.00	17.00	68.00	0.00	73.00	13.00	40.00	37.00

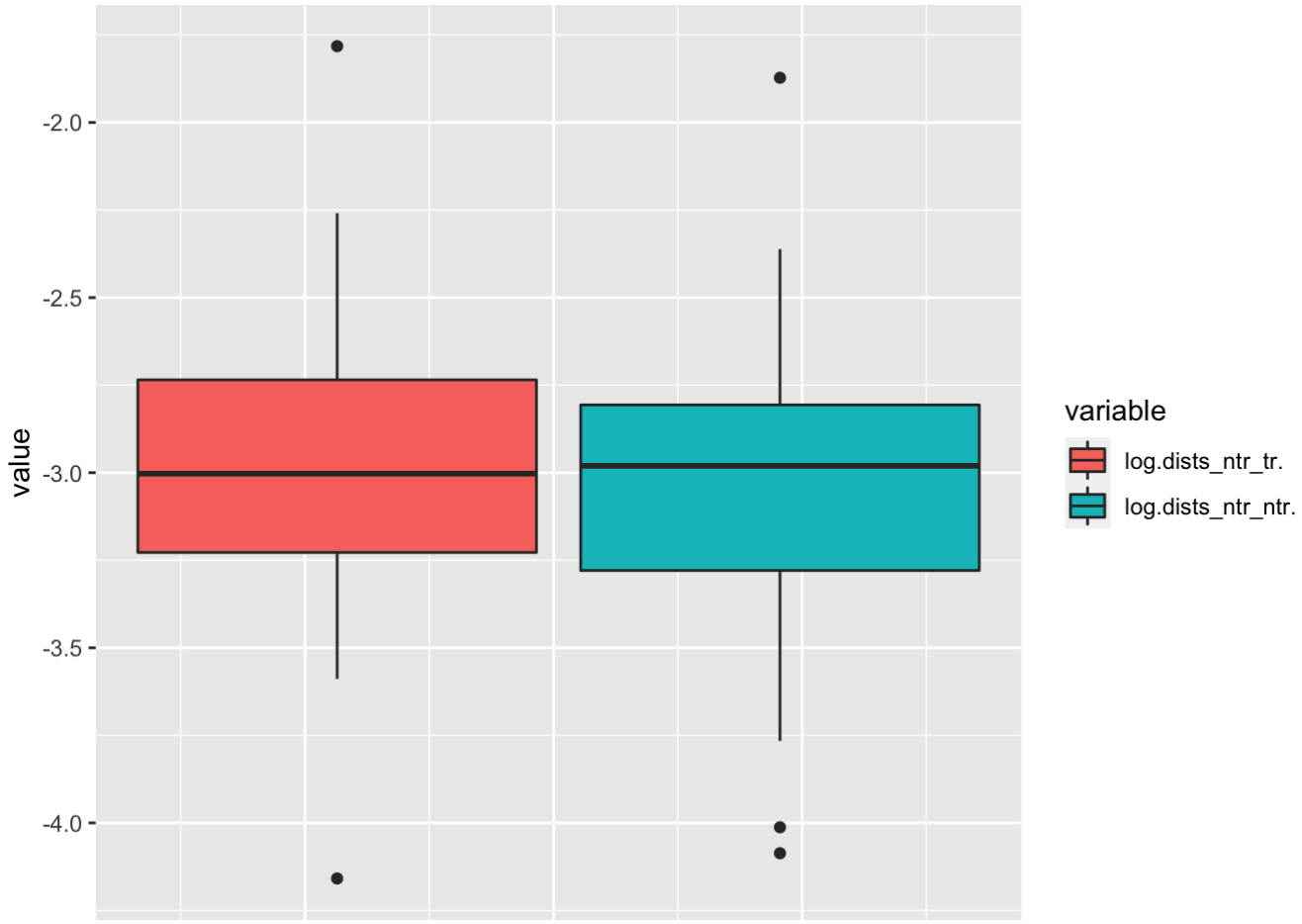
To further quantify the stylistic difference in a more rigorous manner, I have used the Kullback-Leibler divergence (DKL), as Hughes et al. have, which measures the distance between two distributions, much like a Euclidean distance between two points in many dimensions. The formula for computing DKL is as follows:

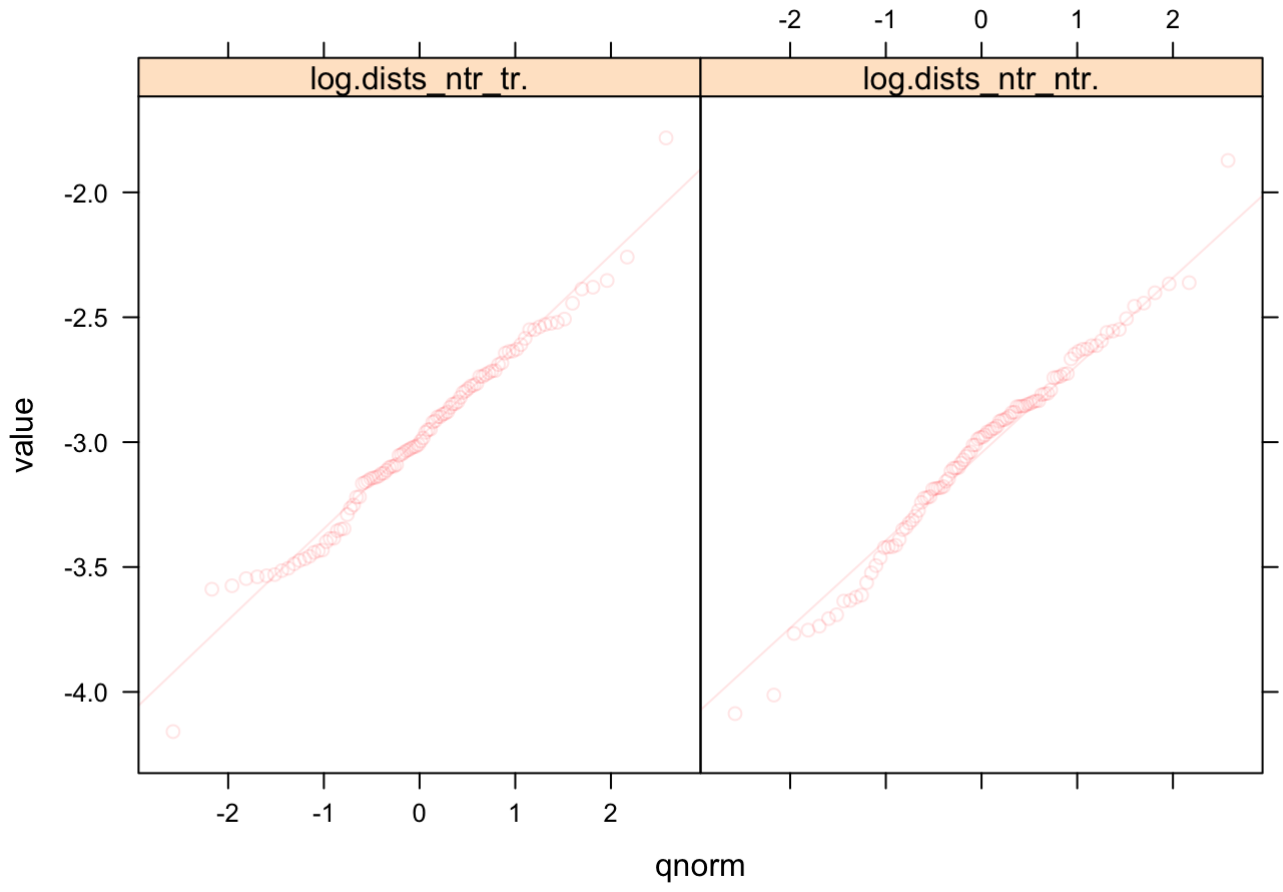
$$D_{KL}(P_i, P_j) = \frac{1}{2} \sum_{w \in \Omega} (P_i(w) log \frac{P_i(w)}{P_j(w)}) + (P_j(w) log \frac{P_j(w)}{P_i(w)})$$

The distance, in the case of H.G. Wells and the Bronte sisters (see above), is **0.0448**. We can use this method to compute the distances between many different pairs of authors. In our case, we'll use the set of translated fiction and the set of non-translated fiction, and compute the two sets of distances, in order to examine the effect of translation on style.

Analysis

In order to compare whether there is a significant stylistic difference between translated fiction and non-translated fiction, I've drawn two sets of 100 random pairs. For the first set, the pairing is between a translated work and a non-translated work; for the second set, the pairing is between two non-translated works. I've examined whether there is a significant difference between those sets of distances. For each pair, I've computed the Kullback-Leibler distance and performed a logarithmic transformation. dists_ntr_tr indicates the distribution of distances between a non-translated work and a translated work; dists_ntr_ntr indicates the distribution of distances between two non-translated works.


-2.0
-2.5
-3.0
-3.5
-4.0
value
variable
log.dists_ntr_tr.
log.dists_ntr_ntr.


log.dists_ntr_tr.
log.dists_ntr_ntr.
-2
-1
0
1
2
-2.0
-2.5
-3.0
-3.5
-4.0
value
qnorm

The QQPLOTs seem roughly linear, though not perfectly linear, and I'll assume that the log of the KL Distances approximately follows a normal distribution. I will not assume that the two sets of distances have the same variance. Judging from the box plots, the variances are similar enough that I should be able to use the two-sample t-test.

```
##
##  Welch Two Sample t-test
##
## data:  df1$log.dists_ntr_tr. and df1$log.dists_ntr_ntr.
## t = 0.86567, df = 197.46, p-value = 0.3877
## alternative hypothesis: true difference in means is not equal to 0
## 95 percent confidence interval:
##  -0.05955446  0.15274989
## sample estimates:
## mean of x mean of y
## -2.990398 -3.036996
```

```
##
## Call:
## lm(formula = binary ~ value, data = df2)
##
## Residuals:
##      Min       1Q   Median       3Q      Max
## -0.59235 -0.50267 -0.00641  0.49883  0.59265
##
## Coefficients:
##             Estimate Std. Error t value Pr(>|t|)
## (Intercept)  0.74386    0.28392   2.620  0.00948 **
## value        0.08092    0.09347   0.866  0.38772
## ---
## Signif. codes:  0 '***' 0.001 '**' 0.01 '*' 0.05 '.' 0.1 ' ' 1
##
## Residual standard error: 0.5016 on 198 degrees of freedom
## Multiple R-squared:  0.003771,   Adjusted R-squared:  -0.001261
## F-statistic: 0.7494 on 1 and 198 DF,  p-value: 0.3877
```

Shown above are two tests. First, a two-sample t-test whose null hypothesis is that the two sets of distances have the same mean. In other words the test assumes equivalence and looks for evidence to the contrary. What we find here is a high p-value of **0.3877**, well above the typical 0.05 threshold, meaning there is not enough evidence to suggest that the means are significantly different. In other words, the observed data does not indicate that the translated works are stylistically different from non-translated works. The test, by nature, also does not allow us to claim that they are significantly similar.

Second test shown above is a regression test that tests whether the type of pairing (translated-nontranslated or nontranslated-nontranslated) is correlated with the distance. We observe that the R-squared values are very small, suggesting there is hardly any correlation between the type of pairing and the stylistic distance.

In order to then test for equivalence, I used the TOST (Two one-sided tests) method, setting the interval of equivalence to +/- 0.25%, as shown below.

```
##
##  Welch Two Sample TOST
##
## data:  df1$log.dists_ntr_tr. and df1$log.dists_ntr_ntr.
## df = 197.46
## sample estimates:
## mean of x mean of y
## -2.990398 -3.036996
##
## Epsilon: 0.25
## 95 percent two one-sided confidence interval (TOST interval):
##  -0.04235935  0.13555478
## Null hypothesis of statistical difference is: rejected
## TOST p-value: 0.0001043313
```

The average difference between the two groups is 0.0466, and an epsilon of 0.25% examines a window as small as +/-0.000129. Since the p-value of **0.000104** is much smaller than the conventional 0.05 benchmark, we are able to conclude that the two groups are effectively equivalent, meaning the distance between a translated work and a nontranslated work is effectively equivalent to the distance between a nontranslated work and another nontranslated work. This implies that there is no significant stylistic difference between translated works and non-translated works.

Conclusion

The study finds that there is no significant stylistic difference between translated works and non-translated works, as observed by the distribution of content-free words. Any readerly claim to be able to recognize a translated text as such, or that a certain short story "sounds translated," appears to be unjustifiable.

One of the assumptions that could be further examined is whether the logarithms of Kullback-Leibler distances are indeed normally distributed. The tail ends of the QQPLOTs appear to be nonlinear. Another assumption that should be studied is the specific set of content-free words. It is a specific variation on Cyril Labbe's method for calculating "intertextual distance," which considers every word in all the texts, not just the specific set of content-free words that have been subjectively determined. Of course, there is also the question of whether we can meaningfully associate content-free words and the bag-of-words approach to an author's style.

One may expand upon this study in terms of stratifying the data according to temporal, geographical, and linguistic adjacency. Hughes et al. have found that there is greater similarity between the styles of similar time periods, for example, so it would be worth examining whether there is significant stylistic difference between a translated text and a nontranslated text in a similar time period. It would also be great to expand the study beyond the texts available in Project Gutenberg. The organizers of Gutenberg appear to have decided there is no need for more than one translation of the same original text. An interesting study could be conducted on multiple translations of a single author-text, or across all the translations by the same translator. Even to study this latter case, Project Gutenberg has too little data on translated works.

There are many possibilities for interpreting the result of this studies. One can conjucture, for example, that what makes a Russian literature sound Russian is not through the style as characterized by the translator's use of prepositions, pronouns, and so on, that translation does not communicate culture through such a style. Perhaps what tips off a reader that a certain text has been translated are the content words, rather than the content-free words. There is no doubt that the themes and motifs that recur in Russian literature remind us of its cultural backdrop, so one might be tempted to conclude that the Russian-ness has more to do with meaning-filled nouns and verbs. However, one cannot ignore the fact that different translations sound different from one another, and that there are schools that manifestly aim to render their translations sound "more natural" – more like a

non-translated work. What is it, then, that makes something sound more like a non-translated work? How can we characterize this aspect of style? How do we tease apart the original author's contribution and the translator's contribution to the style of the translated text? I hope these questions can be addressed in further studies.

Appendix

The list of the content-free words used:

```
a about above across after afterwards again against all almost alone along already also
although always am among amongst amoungst amount an and another any anyhow anyone anythi
ng anyway anywhere are around as at back be became because become becomes becoming been
before beforehand behind being below beside besides between beyond both bottom but by ca
ll can cannot cant con could couldnt cry describe detail do done down due during each ei
ght either eleven else elsewhere empty enough etc even ever every everyone everything ev
erywhere except few fifteen fify fill find fire first five for former formerly forty fou
nd four from front full further get give go had has hasnt have he hence her here hereaft
er hereby herein hereupon hers herself him himself his how however hundred ie if in inc
indeed into is it its itself keep last latter latterly least less ltd made many may me m
eanwhile might mine more moreover most mostly move much must my myself name namely neith
er never nevertheless next nine no nobody none noone nor not nothing now nowhere of off
often on once one only onto or other others otherwise our ours ourselves out over own pa
rt per perhaps please put rather re same see seem seemed seeming seems serious several s
he should show side since six sixty so some somehow someone something sometime sometimes
somewhere still such take ten than that the their them themselves then thence there ther
eafter thereby therefore therein thereupon these they thin third this those though three
through throughout thru thus to together too top toward towards twelve twenty two under
until up upon us very via was we well were what whatever when whence whenever where wher
eafter whereas whereby wherein whereupon wherever whether which while whither who whoeve
r whole whom whose why will with within without would yet you your yours yourself yourse
lves
```