Linguistic Complexity Comparison Between British and American English

Wenjie Sun (ws854)
Xinyan Yang (xy975)
Code: https://github.com/ws854/textasdata

1. Introduction

English is the most popular and the dominant language in the world. It is the official language for the United States, U.K., Australia, Canada, Singapore, New Zealand, Barbados, etc. over time, the language written and spoken in these countries varies from each other, including accent, spelling (such as “program” and “programme”), grammar, and even vocabulary (“soccer” and “football”) and idioms.

There are quite a few different ways in modern days to measure the linguistic complexity for English language. Some common techniques include Dale-Chall’s language score and FRE. Dale-Chall’s measures how many words in a document are not an easy word from a pre-defined dictionary of about 3000 words and what is the word-to-sentence ratio. FRE also measures the syllables and sentence ratio. While they provide a good indicator as a measurement of language complexity, it misses a very important piece of how different English native speakers use this language.

This paper will be attempting to study the difference between these two languages in terms of the complexity underlined in these two languages. The purpose of this project is to research whether language complexity measurement can be used on documents from two different languages and whether there is a weak prior that determines one language is more complicated than the other. All of the code and results we used in this paper can be found at the github page of the authors: https://github.com/ws854/textasdata.

2. Data

This project needs a fairly large corpus to equally represent both language. Our initial thought was to web scripting headline news from mainstream media for the study. But, eventually, we opted to use two popular and publicly-available database for the study. For the british english, we used the British National Corpus (BNC), XML edition [1]. For the american english, we choose the open American National corpus (OANC) [2]. Both of the database provide one of the most diversified documents in each language.

The BNC contains a wide range sources of modern British English corpus with a 100 million word, including both spoken part (910 texts) and written part (3144 texts). The texts are from the later part of the 20th century. In this project, we only focus on the corpus in written part whose subject is "social_science", "world_affairs", "arts", "leisure" or "commerce_finance".

The BNC was encoded in XML files, which contain both the text and its structural properties. Unfortunately, it does not contain the plain text files so we have to convert these XML files into plain text files by removing the structural properties. We will talk about it in details in the next part.

The OANC contains of about 15 million words of American English corpus, also including both spoken part (2410 texts) and written part (8832 texts). We can find all the plain texts files in the corpus repository. In this project, we choose the Slate journal. Slate is an magazine whose topics including News, Politics, Arts, Business, etc. The ANC contains 4694 articles from the Slate archives published between 1996 and 2000[2]. The topics and published time of the sub-corpus we chose is similar with what we chose in British English part, so it is reasonable to choose the Slate corpus to represents the American English corpus for our project.

3. Data Preprocessing

We used the nltk package available in python to pre-process the BNC corpus from XML format to TXT format. nltk package provides a function to read the xml format of the BOC corpus and process to list string. There is a bug in the code that it adds an extra space in front of punctuations. This problem was handled by a rule base filtering to remove the extra space. We also used BNCmeta (Meta data of BNC) in the corpora package in R to map the meta information back to the corpus by document id.

For OANC, the data process was much more straightforward: we only kept the TXT format files in the Slate magazine's folder. The meta data was not accessible in this database, thus there is no further process for this database.

Before starting the experiments, both corpus were processed into dfm by quanteda package. We opted to not stem the word, but removed punctuation and lowercased words so that we can study each individual words contribution to a document.

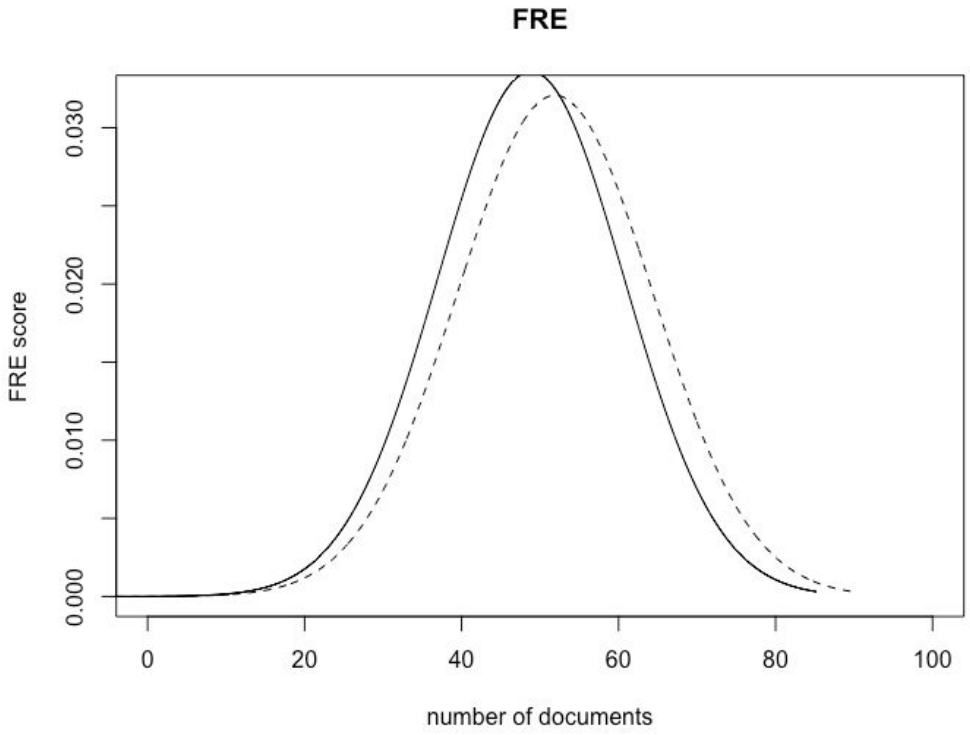
4. Linguistic Complexity Comparison

We will use several different models to compare the linguistic complexity of the two corpus we chose.

4.1 Flesch Reading Ease (FRE)

Flesch Reading Ease (FRE) is a measurement of linguistic complexity. FRE uses the number of total words, total sentences and total syllables to measure the readability of a text. Since the ratio of number of total words to total sentences and total syllables to total words is independent with the length of a text, we are able to use this measurement regardless of the length of our articles. This is important since the average length of texts in British English corpus is much greater than American English corpus.

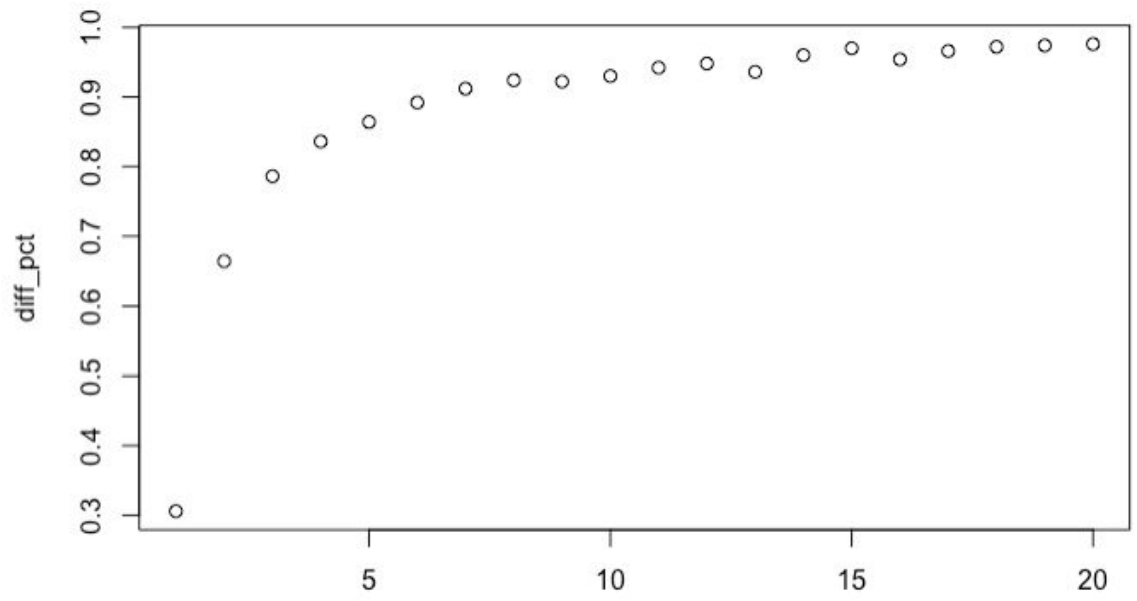
We used the quanteda package to calculate the FRE score of each text in both corpus for a basic understanding of these two corpus. By plotting a smooth histogram line for both corpus, the line (BNC) and the dashed line (OANC) very similar to each other in terms of the distribution. In fact, the mean of the FRE scores is 48.8381 and 51.91524 respectively, and the standard deviation are 12.43899 and 11.87719 respectively. However, we use t-test to find out these two means of the FRE scores are significantly different with a p value of about 10^(-16).



4.2. Bag-of-word

We took the intersection of words in both corpus, which now downsize to 62,116 features that are mutual in both corpus and calculate the weight of each word in each corpus respectively.

To better analyze the difference for each corpus, we opted in the first 10000 for this analysis. Below is a plot of the number of different words in each chunk of 500 words sorted by the ranking. In the first a few chunks, the line very quickly converged to almost 100% after the first 5 chunks (top 2,500 features).



We opted to the ranking technique learned in the first homework to subtract the rank of the same word in the OANC by the rank in BNC. Below is the a part of the table of the 10000 words from each corpus with highest difference in the ranking. The rank takes a descending order (which 10000 is the most frequent word while 1 is the least frequent word in this 10000 words).

features	rank_in_american	rank_in_british	difference
percent	9839	336	9503
u.s	9821	904	8917
dole	9267	418	8849
quiz	8712	171	8541

vs	8936	427	8509

One of the words caught our attention is “dole”. In british, it stands for “benefit paid by the government to the unemployed” on google translation, however, in Americans, it appears in more context with the Bob Dole.

4.3. Bigram

We can find out the most popular bigrams in both corpus using the same method we do on the bag-of-word model. To accelerate the speed of computation, we only focus on the diagrams that appears in the top 5000 bigrams in both of the corpuses. The most popular bigrams in both corpus are the combination of stop words like “of the”, “in the”, “by the”, etc. But as for bigrams with the most different rank in both corpus, the answer is different.

features	rank_in_american	rank_in_british	difference
los_angeles	3069	15	3054
times_and	2747	91	2656
supreme_court	2613	25	2588
the_us	163	2722	2559
in_washington	2558	9	2549

It is surprising that “the_us” have a higher rank in the British English corpus. Maybe it is because the authors of the articles written in British English are more likely live outside the US so they focus the US as a whole, while those who writes American English articles are more

likely live in the US so they focus more on some specific location in the US such as Los Angeles and Washington.

Looking at the bigram list, we can find some bigram that has a high difference in both language while its difference is not very high when we look at each of the single words. For example, “supreme_court”, refers to the highest federal court of the United States, are not widely used in British English. Beside the proper nouns, we can find that “claims_that”, “argues_that” is more popular in the American English and “the_nature”, “means_of” are more popular in the British English. We can find out more examples in the list (which has been uploaded to our github page).

4.4. Dale Chall Easy Word list

We also tried to learn more from the dale chall method. We are taking a different approach to study this method. Instead of running the dale-chall’s score, we analyzed the weight of words not in the list[3]. The rationale is that the list is supposed to include words that commonly used and easy to understand. If the percentage of difficult words is different in both languages, then the dale-chall score is very likely to be different without considering the words-to-sentence ratio.

After removing words that were included in the Dale Chall Easy Word list, there were 33.18% total occurrences of tokens remained in British English and 35.09% words remained in American English. So, without considering the words-to-sentence ratio, the american english has a slightly higher percentage difficult words than the british english.

One of the findings is that location or particular politics words pop up at the top when compared the weight difference between BNC and OANC, which proved our hypothesis that the different weights of vocabularies in both that heavily influenced the language complexity measure.

features	BNC_weight	OANC_weight

council	0.041%	0.005%
development	0.041%	0.009%
england	0.032%	0.006%
british	0.054%	0.029%
uk	0.022%	0.000%
minister	0.039%	0.017%
community	0.033%	0.012%

5. Conclusion & Limitation of Our Models

There is fundamentally difference between these two languages, at least in terms of the samples we selected. We were able to prove the average FRE scores in the two corpus are statistically significant different. The FRE scores of American English is a higher than British English. Based on this result, we conclude the FRE score of a document in American is likely to be 6.25% higher than British english. We reached to an almost similar conclusion that American english is likely to have a little bit (6%) more complicated words than British english when applying Dale-Chall's easy word list. We also believe the main reason by studying FRE, bi-gram, and Dale-Chall all pointed to a different distribution of words used in these two languages.

However, this study has its own limitation:

1) The datasets has a biased difference. OANC documents can not be filtered by domains, instead, it can only be filtered by the media source. Even though when we selected the documents in BNC, we tried to match relevant domains to Slate magazine, there exist an unavoidable difference between corpus. So, these documents may cover some different topics that influenced to the result.

2) Another major problem is the computational power. The corpus we worked on is large enough to represent a distribution of words in the real life. However, our computers were not able to run large computations such as bootstrap to reduce the variance of our result. So we limit our corpus to some specific topics.

7. Reference and Data Source

[1] British National Corpus, XML edition, http://ota.ox.ac.uk/desc/2554

[2] The Open American National Corpus, http://www.anc.org/data/oanc/download/

[3] DaleChallEasyWordList http://countwordsworth.com/download/DaleChallEasyWordList.txt

[4] Murphy, Tara, and James R. Curran. "Experiments in mutual exclusion bootstrapping." *Proceedings of the Australasian Language Technology Workshop (ALTW)*. 2007.