LT3220 Corpus Linguistics

Individual Report

Instructor: Dr. FANG, Chengyu Alex

LI Wang Yau

wangyauli2

Table of Contents

Chapter 1

1. Empirical studies on "D. Religion" and "J. Learned" sections in the Brown Corpus

1.1. Introduction

This is an empirical study of the first computerize corpus, the Brown Corpus(Francis & Kucera, 1964). In this chapter, two sections in the informative prose of the Brown Corpus, namely "D. Religion" and "J. Learned", will be compared. Several measurements, including TTR, sTTR, coverage, vocabulary size, etc will be calculated and used to locate the differences and similarities of the two chosen sections. The research question of this section is ***"In D. Religion and J. Learned, which prose of the Brown Corpus is more difficult?"***

This report is based on the presentation slides of our group[1] during week 5 to week 7 in LT3220. However, the comparison will not be presented according to week orders, instead, the sections will be ordered according to the key concepts in the discussions.

Overall, after considering the limitations, most of the results supports our original hypothesis that "J. Learned" is more difficult than "D. Religion". While at the same time, this research pointed out some important factors what we have to take into account when using corpus approach in measurement.

[1] Other group members including Chang Chia Chi, Lam Wing Shan, Ma Yi Lun Ian, and Wong Po Ying

1.2. Corpus Design

The Brown University Standard Corpus of Present-Day American English (Brown Corpus) was the first computerised corpus which compiled in the 1960s by Henry Kučera and W. Nelson Francis at Brown University. It contains 1 million words with approximately 2000 words in each of 500 texts. It was divided into two main portions, including 374 samples form the informative prose and 126 samples form the imaginative prose. Similar categories are used in some sister corpus such as the Australian Corpus of English (ACE), and the Wellington Corpus of English (WCE).

According to the website of Lancaster university, the composition of the Brown corpus is as follow:

	Broad text category	Text category letter and description ("genre")		Brown
Informative	Press	A	Press: Reportage	44
		B	Press: Editorial	27
		C	Press: Reviews	17
	General Prose	D	Religion	17
		E	Skills, Trades and Hobbies	
		F	Popular Lore	
		G	Belles Lettres, Biographies, Essays	
	Learned writing	H	Miscellaneous: Government documents, industrial reports etc.	30
		J	Science	80
Imaginative	Fiction	K	General Fiction	29
		L	Mystery and Detective Fiction	24
		M	Science fiction	6
		N	Adventure and Western	29
		P	Romance and Love story	29
		R	Humour	9

Table 1. Brown Corpus composition

1.3. Chosen sections

In this study, the "D. Religion" and "J. Learned" sections have been chosen for comparison. Both of the sections were taken from the informative prose, while they differ in terms of their categories and text size. The composition of the two sections are listed below(Francis & Kucera, 1979).

D. Religion	
Books	7
Periodicals	6
Tracts	4
Total	**17**

Table 2. Composition of "D. Religion" sections

J. Learned	
Natural Sciences	12
Medicine	5
Mathematics	4
Social and Behavioral Sciences	14
Political Science, Law, Education	15
Humanities	18
Technology and Engineering	12
Total	**80**

Table 3. Composition of "J. Learned" sections

It is shown that the religion section contains mainly books and periodicals regarding religions, while the learned section mainly contains essays in scientific aspects and humanities. One important difference is that the number of texts contained in "J. Learned"(80) is nearly five times that of "D. Religion"(80). The difference in text size is one crucial factor in later comparisons.

1.4. Definitions

a. Type

- *Type* is the total number of unique words. Only unique words are counted as *type*, thus any repeated words would be counted once only. For example, there are only four types in the sentence "*The apple hit the boy"*. Given the total number of *types* is the estimation of vocabulary size of that corpus/text.

b. Token

- *Token* is the total number of words in a corpus. Unlike *type*, repeating words can be double counted in *token*. For example, there are five tokens in the sentence "*The apple hit the boy"*. The total number of *tokens* can be an estimation of corpus size.

c. Type token ratio(TTR)

- *Type token ratio(TTR)* serves as indicator of lexical diversity. However, direct comparison of *TTR* will only be meaningful when comparing similar sized corpora. It is calculated by the following formula:

$$TTR = \frac{total\ number\ of\ types}{total\ number\ of\ tokens} \times 100\%$$

d. Standardized type token ratio(sTTR)

- *Standardized type token ratio(sTTR)* is used when comparing corpora in different size. Since *TTR* varies hugely with corpus size, *sTTR* is needed for fair comparison. It is calculated by dividing the larger text into subsections which contains similar number of tokens as the smaller sized text. Then calculate the *TTR* for each subsections of the larger corpus.

e. Coverage

- *Coverage* is the percentage of text that could be understood by knowing a given number of vocabulary. The *individual percentage* is calculated by

$$\text{Individual percentage} = \frac{frequency\ of\ one\ type}{total\ number\ of\ tokens} \times 100\%$$

  The *cumulative percentage* is the sum of every *individual percentage.* It is used to calculate the *cumulative coverage* of *types* in a text.

- *Coverage* is usually used in two types of comparisons of texts. For example, given a **fixed** *coverage* and compare the required *types* in understanding the texts or given a **fixed** *vocabulary size* and compare their *coverage* in different texts. The interpretations will be shown in table 4 and 5.

-	Coverage	VS	Result
Text A	Higher	Fixed	Easier
Text B	Lower	Fixed	Harder

Table 4. Comparison of coverage under fixed vocabulary size

	Coverage	VS	Result
Text A	Fixed	Higher	Harder
Text B	Fixed	Lower	Easier

Table 5. Comparison of vocabulary size under fixed coverage

f. Vocabulary Size(VS)

- *Vocabulary size(VS)* is the number of *types* required to understand a particular amount/percentage of text. In order to compare the vocabulary size, an ordered wordlist according to word frequency is needed.
- Similar to the usage of *coverage*, *vocabulary size* is usually used in two types of comparisons of texts. For example, given a **fixed** *vocabulary size* and compare their *coverage* in different texts, or given a **fixed** *coverage* and compare the required *vocabulary size* in understanding that amount of text. The interpretations are the same as table 1 and 2.

g. Zipf's Law

- *Zipf's law* stated that the frequency of a type is related to the **inverse** of its rank (Zipf, 1936, 1949).

$$f(r) \propto \frac{1}{r^{\alpha}}$$ *(with $\alpha = 1$)*

This relation can be clearly shown by plotting the log graph as follow:

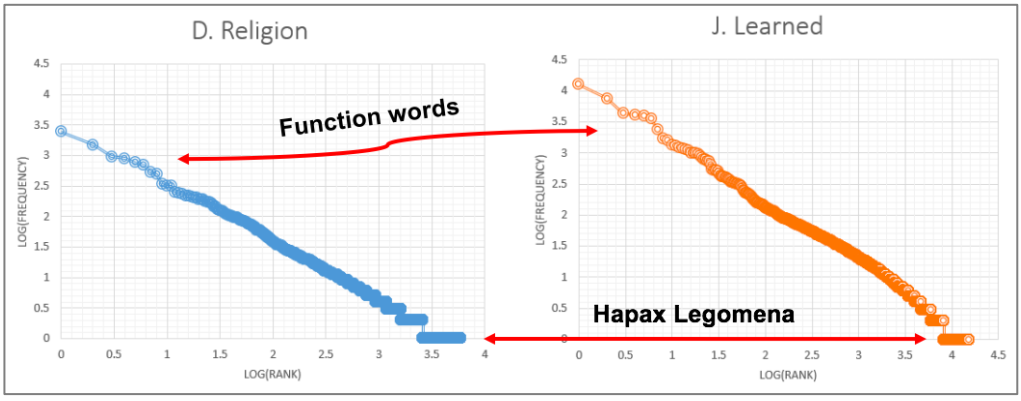


Figure 1. Illustration of Zipf's law in this study

In figure 4, it is shown that log(frequency) and log(rank) are in inverse relation.

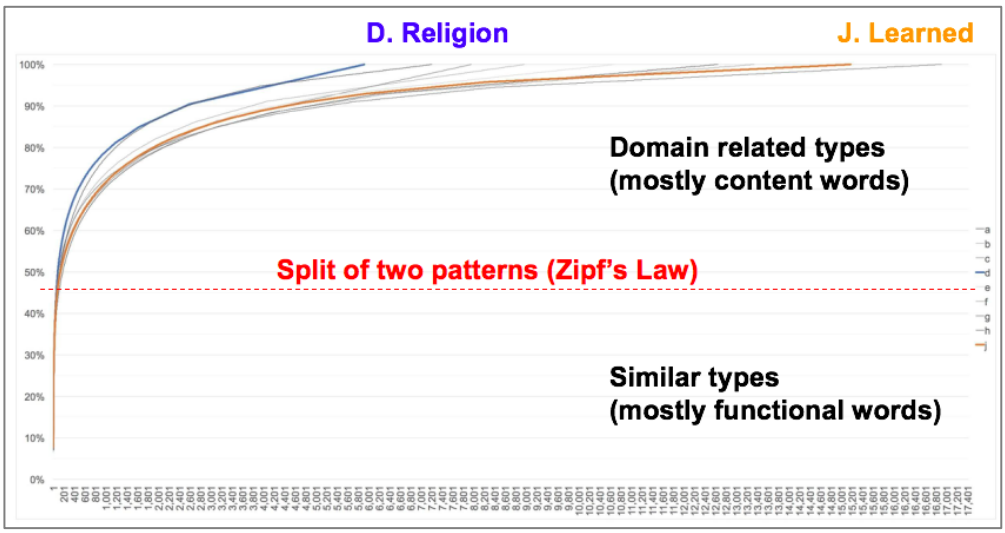


Figure 2. Illustration of Zipf's law in this study

Based on *Zipf's law*, we can also locate the split of function words and content words in figure 2.

h. Semantic Density(SD)

- *Semantic Density(SD)* is the degree of condensation of meaning. It is calculated by the following formula:

$$\text{Semantic density} = \frac{number\ of\ content\ words}{total\ number\ of\ tokens} \times 100\%$$

- In the above formula, *content words* only cover the words in the open class. Thus, functional words/grammatical words such as *he, she, the, a, has, of, in* would not be counted.[2]

i. Hapax Legomena Ratio(HLR)

- *Hapax Legomena Ratio(HLR)* is used to indicate vocabulary richness. The word *hapax* means a word that occurs only once in a text or a corpus (Lardilleux & Lepage, 2007: 458). This ratio can be calculated as follow:

$$\text{Hapax Legomena Ratio} = \frac{number\ of\ hapax\ legomena}{total\ number\ of\ tokens} \times 100\%$$

j. Repeat Rate (RR)

- *Repeat Rate(RR)* is used to express the diversity of relative frequencies of the elements of a closed system. It can be calculated as follow:

$$\text{Repeat rate} = \frac{number\ of\ tokens}{number\ of\ types} \times 100\%$$

[2] Function words can be filtered out by using a stop list stated in -grep -v -f command.

1.5. Hypothesis

A basic comparison table is shown in table 6.

Categories	D. Religion (17)	J. Learned (80)
Context	Contains more **stories** with **repeated names**	Contains **professional knowledge** in **different scopes**
Vocabulary	Easier and less complicated as stories contains less terms	More difficult due to a wider coverage of professional terminology
Purpose(s)	For general education/ entertainment	For academic /professional purposes

Table 6. Basic comparisons of two sections

We hypothesize that *"J. Learned"* is more difficult. Since it has a higher lexical diversity due to a wider coverage of professional knowledge/terminologies in different scopes, and thus result in a higher TTR/sTTR, lower coverage under fixed vocabulary size, higher semantic density, higher hapax legomena ratio, and lower repeat rate.

In contrast, *"D. Religion"* is easier. It contains more stories with possibly highly repeated names, and the stories should be easily enough to the general public, thus we hypothesize that *"D. Religion"* has a lower TTR/sTTR, higher coverage under fixed vocabulary size, lower semantic density, lower hapax legomena ratio, and higher repeat rate.

1.6. Discussions

1.6.1. Type Token Ratio (TTR)

Under direct comparison, the TTR of *"D. Religion"* and *"J. Learned"* showed the exact opposite to our hypothesis. Table 7 shows the overall distribution of size, tokens, types, and TTR of all sections in the Brown Corpus.

Text	Size	Tokens	Types	TTR
A. Press: Reportage	44	91063	11926	13.1%
B. Press: Editorial	27	55615	8627	15.5%
C. Press: Reviews	17	36103	7732	21.4%
D. Religion	**17**	**34984**	**5806**	**16.6%**
E. Skill and Hobbies	36	74618	9957	13.3%
F. Popular Lore	48	98965	12686	12.8%
G. Belles Lettres	75	154476	16137	10.4%
H. Miscellaneous	39	63510	6888	10.8%
J. Learned	**80**	**164891**	**14240**	**8.6%**
K. Fiction: General	29	59536	8354	14.0%
L. Fiction: Mystery and Detective Fiction	24	49712	6163	12.4%
M. Ficiton: Science	6	12332	2979	24.2%
N. Fiction: Adventure and Western	29	60081	7906	13.2%
P. Fiction: Romance and Love Story	29	60397	7528	12.5%
R. Humor	9	18731	4632	24.7%

Table 7. Size, tokens, types, and TTR of all sections in the Brown Corpus

According to table 7, it is clear that the TTR of *"D. Religion"*(16.6%) is almost double of *"J. Learned"(*8.6%). This result is totally different from our original hypothesis. However, this result may not truly reflect the difficulty of the two selected text. These mainly due to four reasons: hapaxes distribution, tokenization, sampling and size difference.

One factor is that the sampling method of taking only 2000 words for each text may not truly reflect performance of particular genre. We expected some text types to repeat in long term e.g. names in religion books, but each text only takes 2000 words. Thus this result in a TTR higher than expected in *"D. Religion"*.

However, the most important factor is the differences in text size. An illustration of the internal relation of token, type, and TTR is shown in figure 3.

(next page)

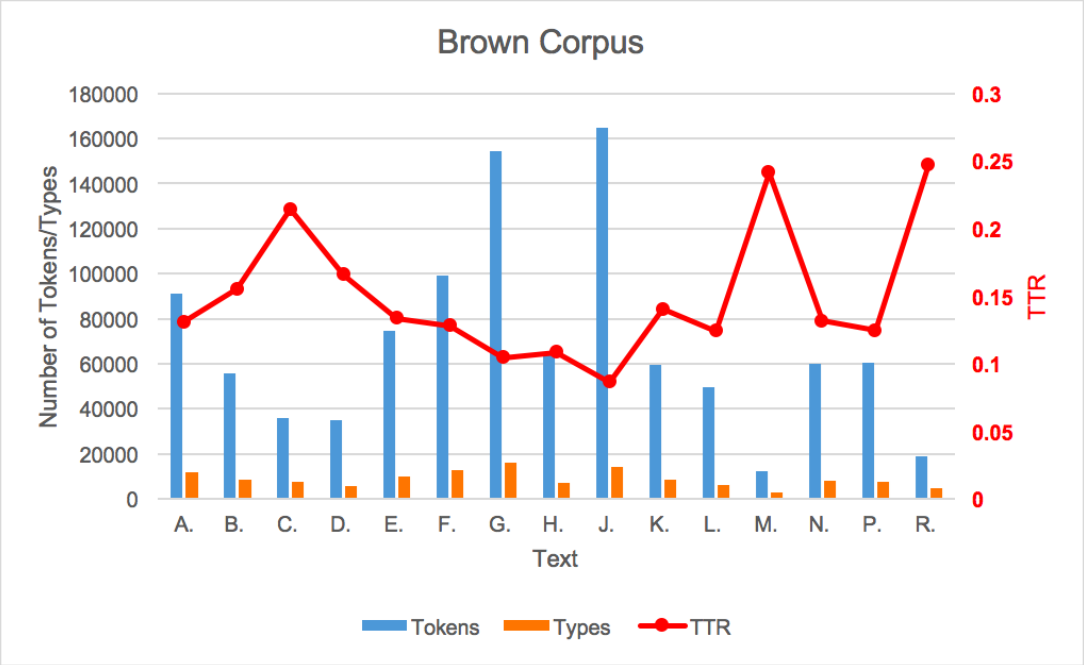


Figure 3. Token, type, and TTR in the Brown Corpus

According to figure 3, *"D. Religion"* has a higher TTR than *"J. Learned"*. More specific, *"D. Religion"* has a high TTR while *"J. Learned"* has the lowest TTR among all sections. While at the same time, *"D. Religion"* has a relatively small number of tokens(34,984), and *"J. Learned"* has the highest number of tokens(164,891) among all texts.

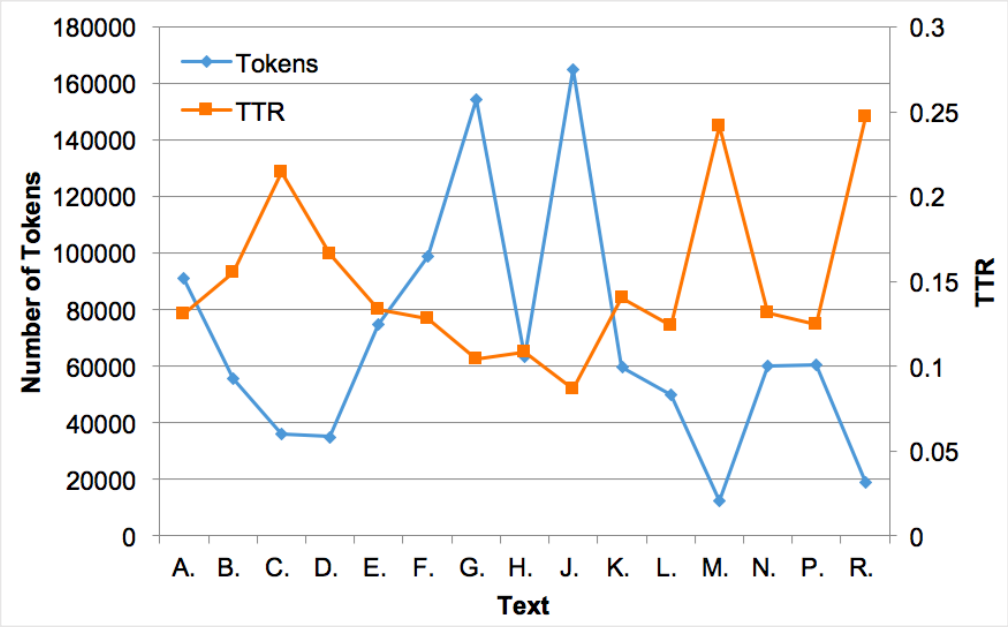


Figure 4. Tokens and TTR in inverse relations

Figures 4 shows that texts with high TTR usually has low token number, vice versa. Figure 5 and 6 show this relation by sorting the texts in ascending order of tokens.

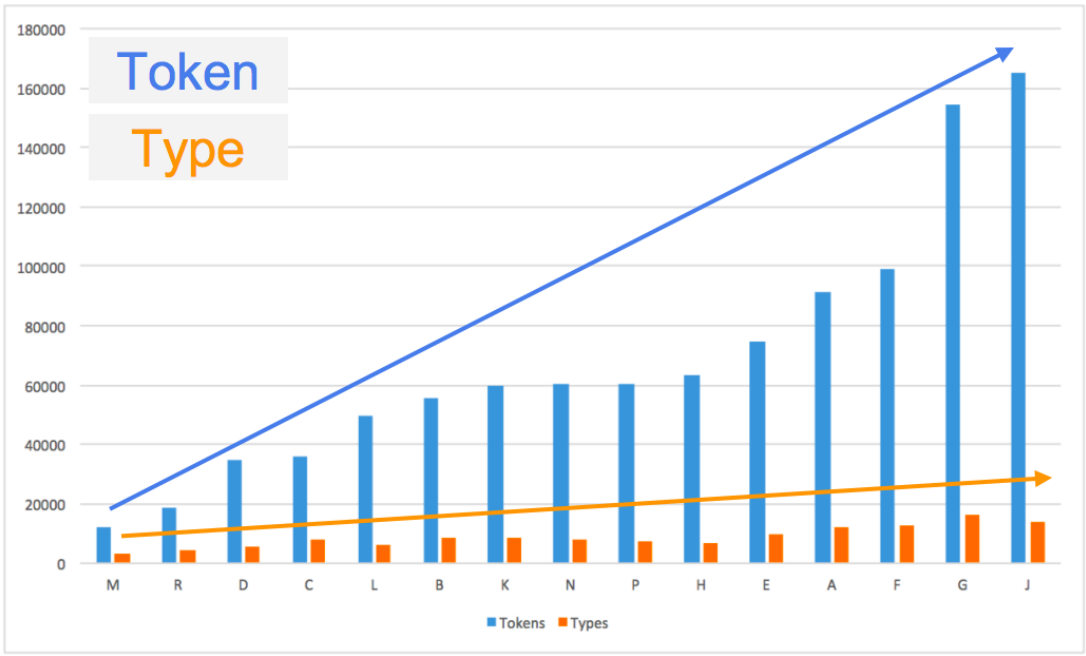


Figure 5. Text tokens and types in ascending order of tokens

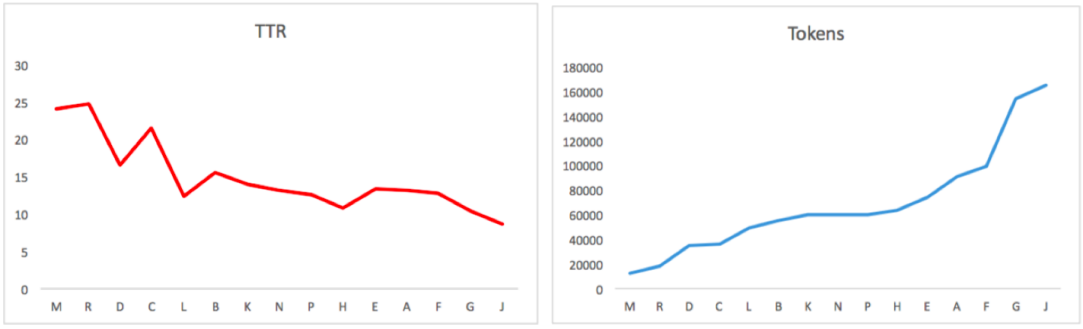


Figure 6 and 7. Relations of TTR and tokens

According to figure 5 - 7, even though types do vary, but they do **NOT** grow **linearly** with tokens. This leads to the problem that TTR varies enormously depending on corpus size. Therefore, TTR comparison is only meaningful between comparably sized corpora. One improvement is using Moving-Average-TTR(MATTR) as proposed by Covington & McFall(2010), an illustration is shown in table 8,

Text	Tokens	Types	TTR
D - all	34984	5806	16.5962
J - all	164891	14240	8.6360
J - beginning	35004	5166	14.7583
J - end	34999	5897	16.8491

Table 8. Demonstration of MATTR in similar sized text

By taking only the first and last 35000 tokens in *"J. Learned"*, the TTR(approximately 14.75-16.85) is much closer to that(approximately 16.60) of *"D. Religion"*.

1.6.2. Standardized Type Token Ratio (sTTR)

Standardized TTR is introduced after realizing the problem of directly comparing TTR of texts. Table 9 shows the comparisons of the two selected texts under standardizing their size.

Section	Type	Token	sTTR
D	5924	34839	**17.00%**
J_1	5286	34453	15.34%
J_2	4863	34709	14.01%
J_3	5226	34311	15.23%
J_4	6559	34735	18.88%
Average TTR of J			**15.87%**

Table 9. Comparison of sTTR

Under this comparison, it is shown that different part of the text has different TTR. Some portions of *"J. Learned"* has a higher TTR than *"D. Religion"*. While the J_4 has much higher TTR than *"D. Religion"*. The overall TTR of *"J. Learned"* (15.87%) is still lower than that of *"D. Religion"*(17.00%) but their differences is reasonably closer after eliminating the size factor.

(next page)

1.6.3. Coverage and Vocabulary Size

Our group have chosen 80% as coverage percaentage and 3000 words as the standard vocabulary size to compare the lexical complexity of *"D. Religion"* and *"J. Learned"*.

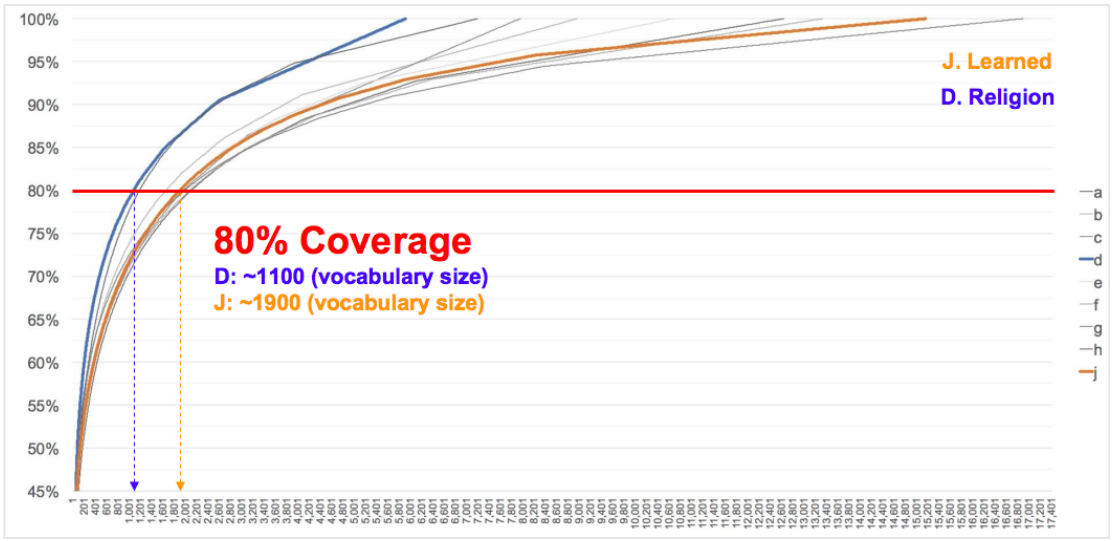


Figure 8. Comparing vocabulary size under fixed coverage (80%)

According to figure 8, *"J. Learned"* requires a larger vocabulary size(approximately 1900) than *"D. Religion"*(approximately 1100) under 80% coverage. Since a larger amount of word types is needed to understand the same portion of text in J, this shows that *"J. Learned"* is more difficult. Similarly, the same pattern is shown in another comparison,

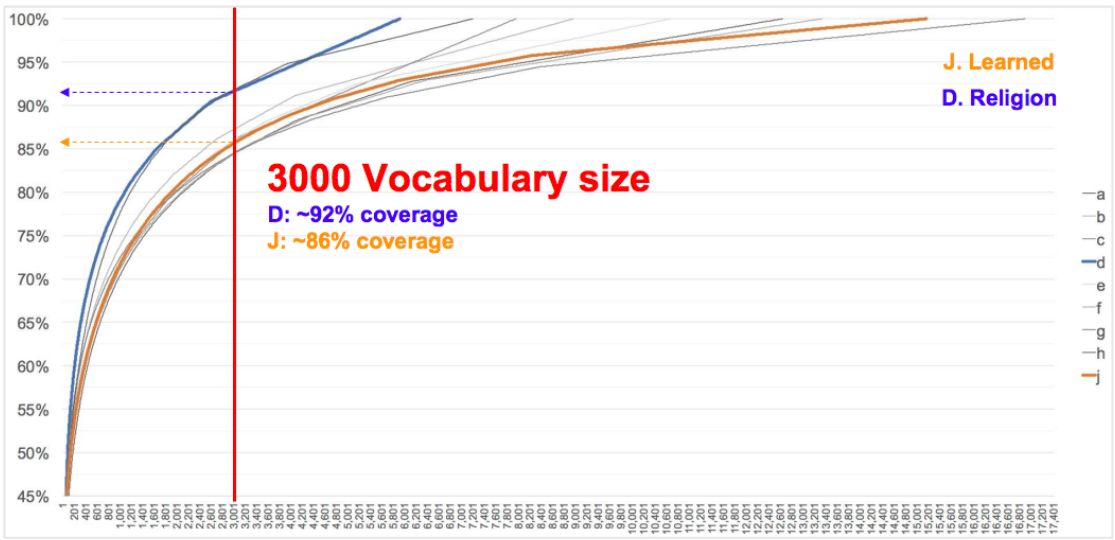


Figure 9. Comparing coverage under fixed vocabulary size(3000)

According to figure 9, *"J. Learned"* has a smaller coverage(approximately 86%) than *"D. Religion"*(approximately 92%) under 3000 vocabulary size. Since a smaller amount of text can be understood by the given fixed amount of vocabularies, *"J. Learned"* is more difficult. However, the same problem led by size difference occurs here.

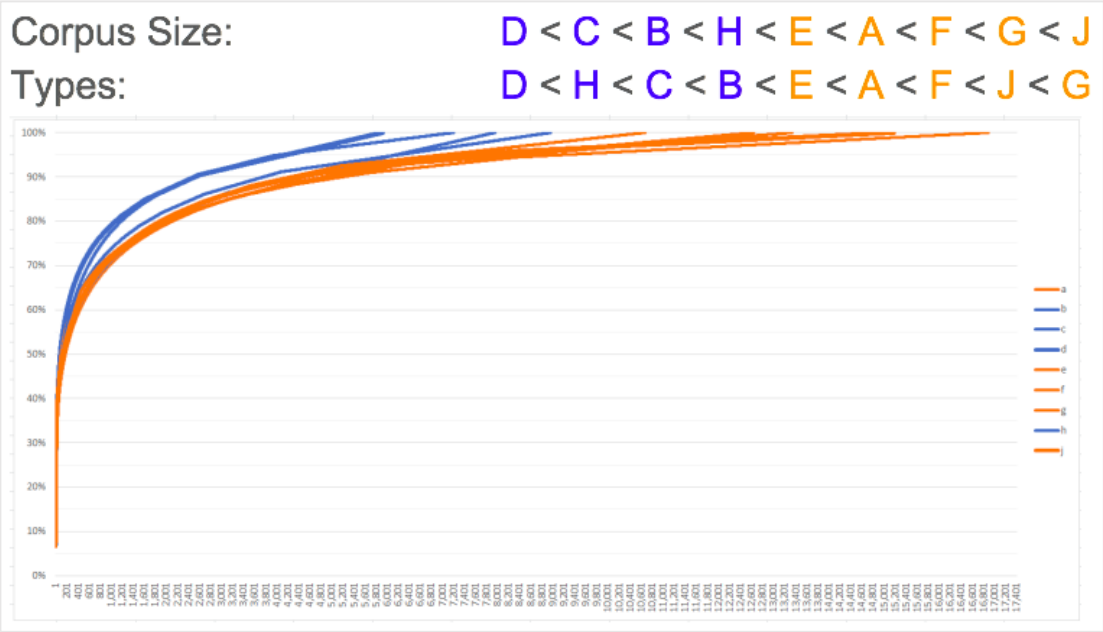


Figure 10. Cumulative coverage with colors

According to figure 10, it seems that texts with larger size would most likely have more types and thus its line would shift to the right of the graph. As a result, the comparison of coverage and vocabulary size may not be fair for texts with different size, as the larger sized corpus would most likely get a lower coverage and higher vocabulary size.

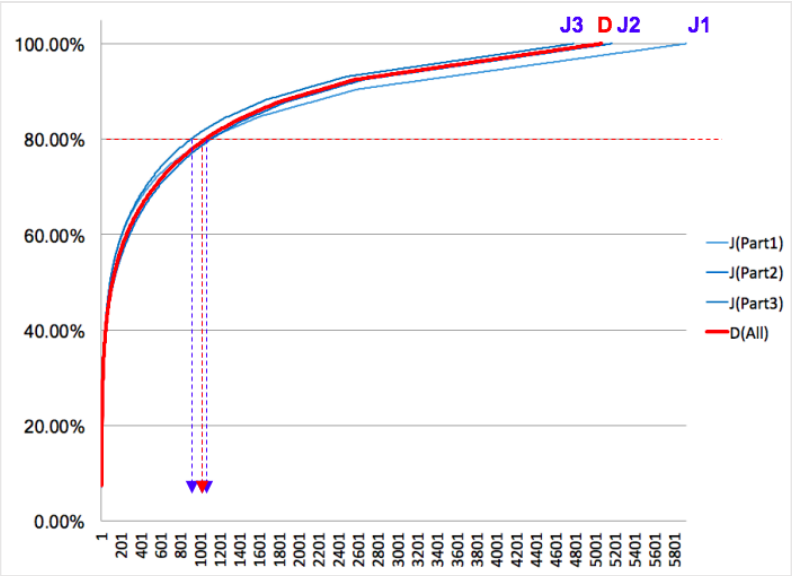


Figure 11. Compare under similar sized subsections

In figure 11, our group tried to compare the two texts in similar sized environment. By separating *"J. Learned"* into similar size of *"D. Religion*", which means approximately 35000 words per part, we obtained the same pattern shown in sTTR comparison. According to the coverage percentage, the two selected sections are of similar difficulty. *"J. Learned"* actually lies somewhere in the middle among different subsections of *"J. Learned"*.

Up to this point, some contradictions have been found in between TTR and word frequency studies. In order to compare the difficulties of the two sections, more measurements are needed for confirmation.

1.6.4. Zipf's Law

One important similarity of all texts can be pointed out by using Zipf's law. Zipf's law stated that the frequency of a type is related to the inverse of its rank (Zipf, 1936, 1949). This relation can be clearly shown by plotting the log graph as follow:

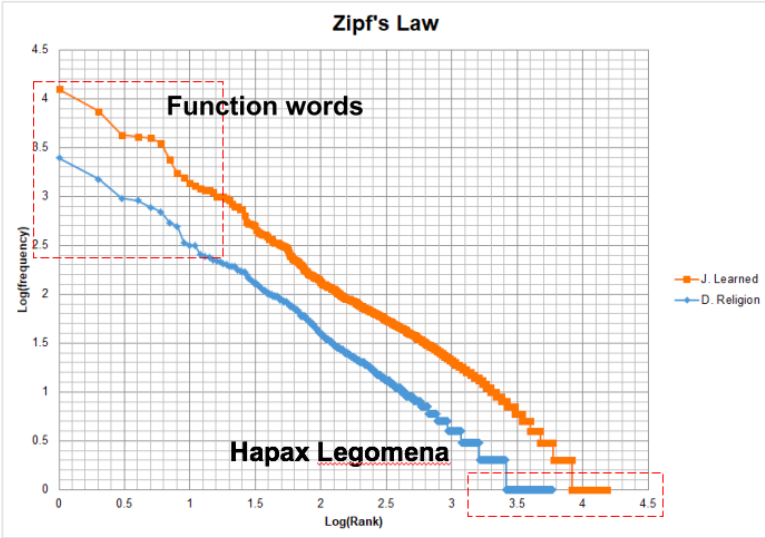


Figure 12. Illustration of Zipf's law in this study

In figure 12, it is shown that log(frequency) and log(rank) are inversely related. We can also see the differences between high-ranked words and low-ranked words. There are always some function words with high frequency while many hapax legomena with low frequency.

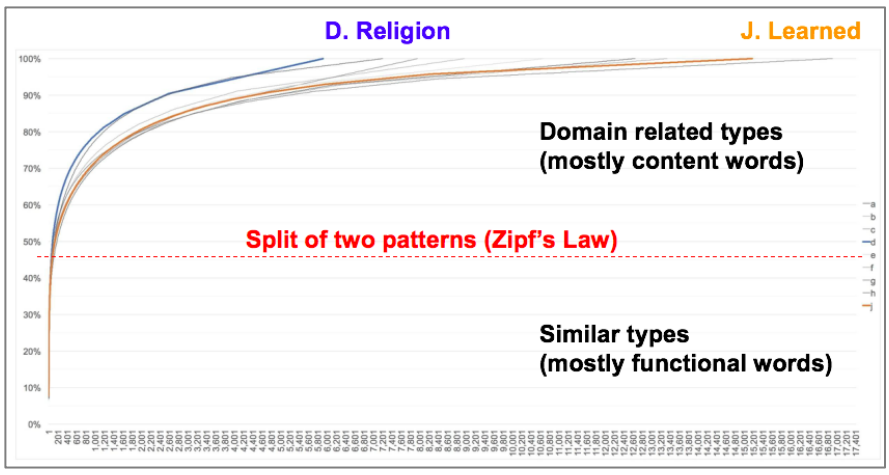


Figure 13. Illustration of Zipf's law in this study

Based on *Zipf's law*, we can also locate the split of function words and content words as shown in figure 13. Table 10 - 14 shows the wordlist in these two portions respectively.

D. Religion				J. Learned			
Freq.	Type	Ind %	Cum %	Freq.	Type	Ind %	Cum %
2479	the	7.12%	7.12%	12537	the	7.67%	7.67%
1505	of	4.32%	11.44%	7452	of	4.56%	12.23%
963	and	2.76%	14.20%	4283	and	2.62%	14.85%
904	to	2.60%	16.79%	4099	in	2.51%	17.36%
776	in	2.23%	19.02%	3948	to	2.42%	19.78%
699	a	2.01%	21.03%	3525	a	2.16%	21.94%
537	is	1.54%	22.57%	2408	is	1.47%	23.41%
495	that	1.42%	23.99%	1715	that	1.05%	24.46%
337	it	0.97%	24.96%	1568	for	0.96%	25.42%
315	for	0.90%	25.86%	1360	be	0.83%	26.25%

Table 10-11. Rank 1-10 types in "D." and "J."

According to table 10-11, the top 10 words across different texts would almost always be the same set of function words, such as *"the, of, and, to, in"*. Even more, each of their individual coverage is similar, for instance, the definite article "the" almost always account for around 6-7% in texts. Top 100 types usually cover around 50% of the coverage. These grammatical words serve as functional "skeleton" of the sentence structure. The occurrences of these words are highly governed by syntax.

Figure 14. Word cloud of function words in D. Religion (left) and J. Learned (right)

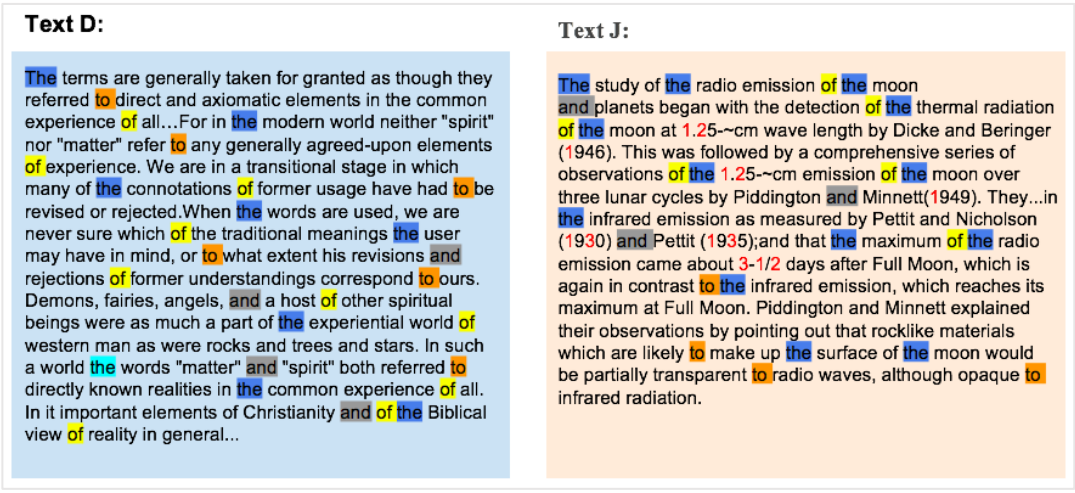


Figure 15. Actual usage of the function words in the two texts

Figure 14 show the word cloud illustration of the common function words. Figure 15 show the actual usages. Since both selected texts are from the informative prose of the Brown corpus, the function words are commonly used to serve the purpose of handling information. For example, "**the**" is used to show information that is mentioned in previous sections, "**of**" is used to show the possession, origin, and belonging, "**and**" is used to join related statements together.

D content word	
Freq.	Types
145	god
99	new
96	world
95	church
80	man
60	christ
59	spirit
57	also
55	life
50	members
49	power
49	christian
44	men
43	faith

J content word	
Freq.	Types
308	1
267	2
250	two
243	first
227	also
210	time
202	used
183	system
173	number
170	state
164	made
157	3
147	new
127	social

Table 12-13. Content words in "D." and "J."

While it is not the same case for lower ranked words, the word types after a certain percentage would mostly be content words which highly reflect the genre and authors' intentions. For example, the words in table 12 are very different form those in table 13. Table 12 contains mainly words related to religion, such as god, spirit, church, all these focused on religious and spiritual information in the western society. While table 13 contains words associated with numbers, such as 1, 2, time, number. These words are temporal related and used to recorded the years of new discoveries and references. Furthermore, words related to action are also frequently used in text J. These words are used to describe the materials/method used in discovery.

1.6.5. Semantic Density (SD)

According to the measurement of semantic density, *"D. Religion*" seems to be easier. Table 14 shows the SD of the two selected texts, a standardized version of *"J. Learned"* is calculated by averaging the SD of different standardized subsections of *"J. Learned"*.

	Semantic Density(SD)	
D	**45.97%**	**(smaller, easier)**
J	**49.75%**	**(higher, harder)**
J(standized)	**49.05%**	**(higher, harder)**

Table 14. Semantic density

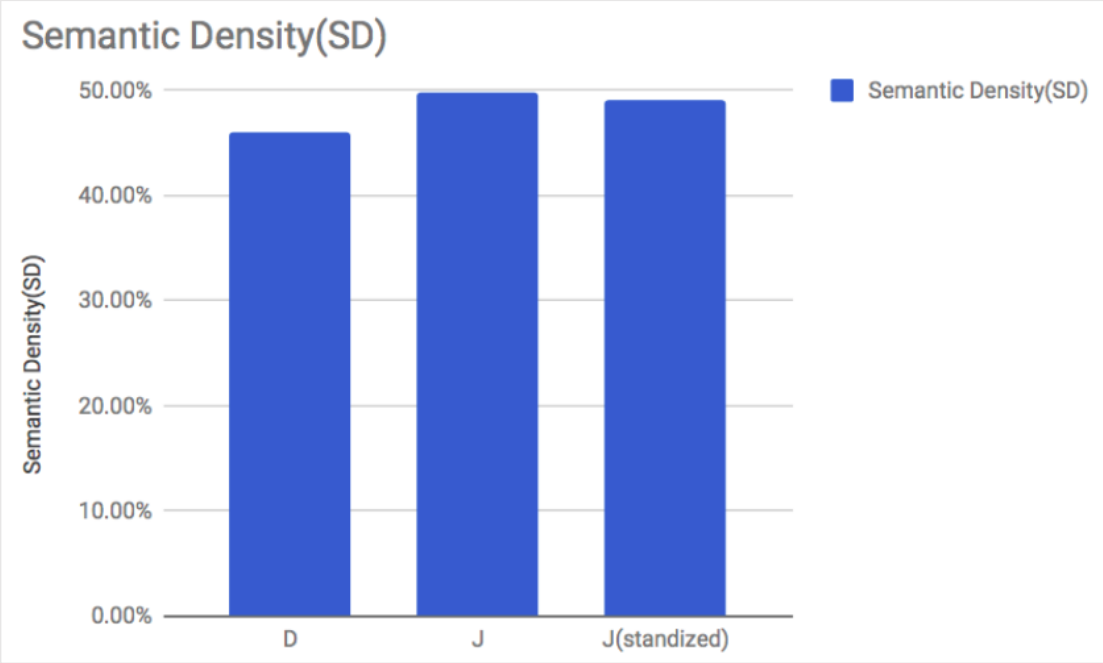


Figure 16. Semantic density

According to figure 14, "*J. Learned"* (49.75%) & "*J. Learned (standardised)*" (49.05%) have higher SD than "*D. Religion*" (45.97%), thus *"J. Learned*" is more difficult.

1.6.6. Hapax Legomena Ratio (HLR)

However, the results in hapax legomena ratio again shows the opposite pattern.

	Hapax Legomena Ratio(HLR)	
D	9.56%	(higher, harder)
J	4.25%	(smaller, easier)
J(standized)	8.08%	(slightly smaller, easier)

Table 15. Hapax legomena ratio

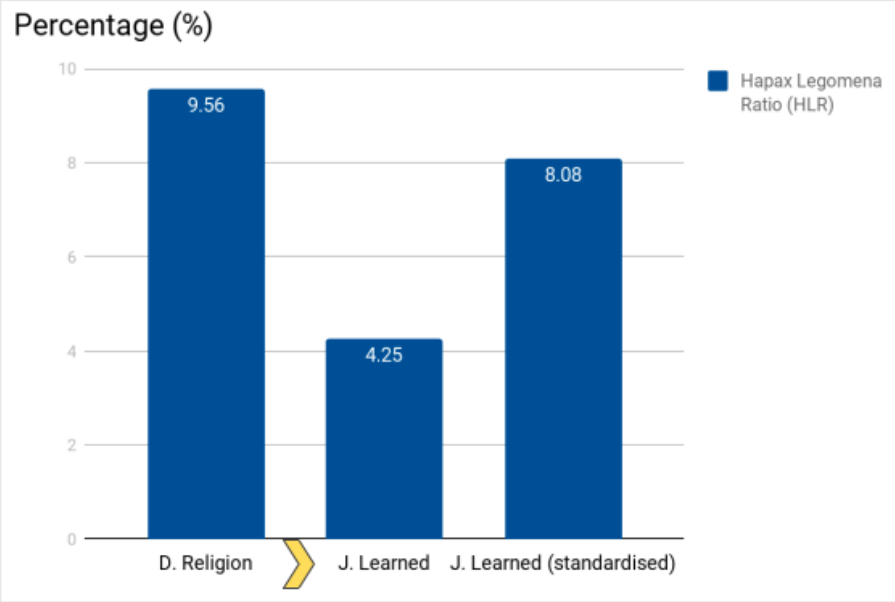


Figure 17. Hapax legomena ratio

Since size factor highly affects the ratio, standardized version is calculated for *"J. Learned"*. According to table 15 and figure 15, *"D. Religion"* (9.56%) has a higher hapax legomena ratio than "*J. Learned*"(4.25%) & "*J. Learned (standardised)*"(8.08%). Thus "*D. Religion*" is more difficult.

1.6.7. Repeat Rate (RR)

	D. Religion	J. Learned
Types	5924	15179
Tokens	34839	163403
Repeat rate	588.10%	1076.51%

Table 16. Repeat rate

Table 16 shows that "*J. Learned*"(1076%) has a much higher repeat rate, indicating *"D. Religion"* is more difficult.

1.7. Summary of all measurement methods

	D. Religion	J. Learned	Harder?
TTR	Higher	Lower	D
sTTR	Higher / Lower	Lower (J1-3) / Higher (J4)	D / J
Coverage (VS fixed)	Higher	Lower	J
Vocabulary size (C. fixed)	Lower	Higher	J
Semantic density	Lower	Higher	J
HLR	Higher	Lower	D
Repeat rate	Lower	Higher	D

Table 17. Summary of all measurements

Table 17 summarized the results and implications of all the seven measurements. It is found that different measurements contradict with each other. Nearly half of the measurements show that *"D. Religion"* is more difficult, and the other half show the opposite. Each of these measurements reflect different quality of the text, thus the actual level of difficulties depends on the area we are focusing on.

To sum up, "*D. Religion*" has higher vocabulary richness(HLR) and it reequires more specific knowledge to understand specific terms(TTR). While "*J. Learned*" has higher condensation of meaning(SD), and more types are required to get a fair amount of understanding of the text(VS and coverage).

1.7.1. Caveats and Limitations

Three main limitations that affecting all measurements were found: i) Corpus size, ii) Tokenization, iii) annotation, and iv) lemma form.

i) Corpus size

As shown in TTR, coverage, and repeat rate, most of the measurements are only comparable for similar sized corpora. There will be huge difference or bias when we are comparing texts with different size. The improvement is using standardize version or taking average of subsections.

ii) Tokenization

Since all the calculation are based on token and type counting, improper tokenization may affect the result of frequency count and thus affect the final result. It is found that some

of the numbers “61.2” , short form with apostrophes “ ‘ ” , and abbreviations “ U.S.A.” were improperly tokenized. For example, “Mary’s” will become “Mary” and “s”, “61.2” will become “61” and “2”. These errors would double count some of the tokens and lead to unusual frequency in some single letters, see table 18.

J. Learned					
Original		Type	Freq.	Ind %	Cum %
[f, **f]	→	f	999	0.61%	0.61%
[s, **s, ' s]	→	s	519	0.32%	0.93%
[1, 20.1, 1.1]	→	1	308	0.19%	1.12%
[2, 20.2, 1.2]	→	2	267	0.16%	1.28%
[c, **c]	→	c	159	0.10%	1.38%
[3, 20.3, 1.3]	→	3	157	0.10%	1.47%

Table 18. Top ranked character tokens

The improvement is to follow certain **rule sequences** or **flowcharts** in text processing (Fang, 2007, p.45-46) shown in figure 18.

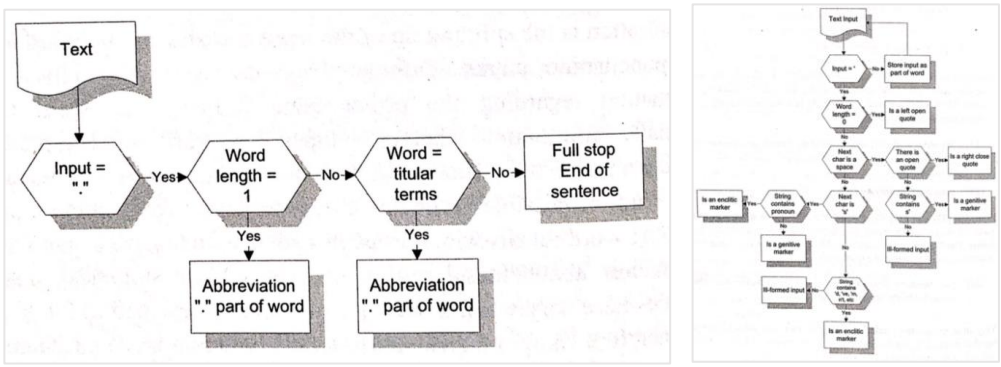


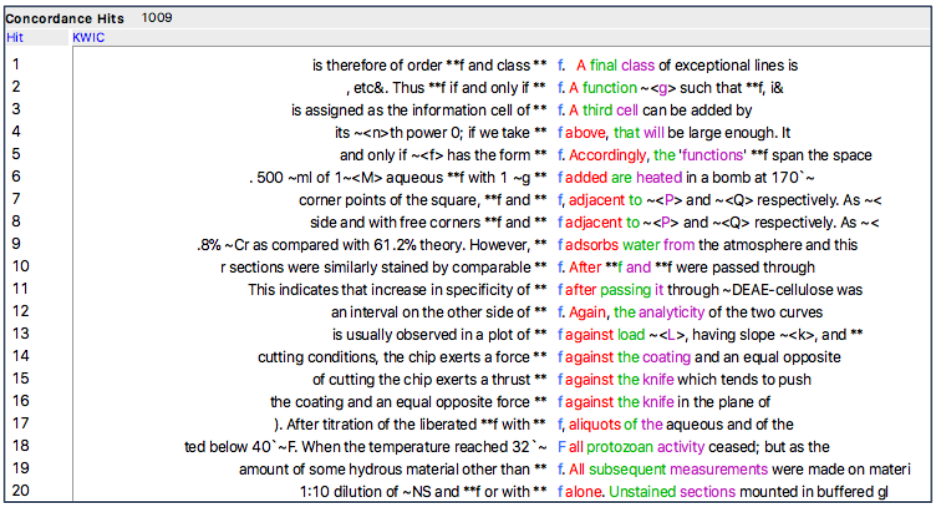
Figure 18. Flowcharts for the processing of “.” and “ ‘ ”

iii) Annotation

According to table 19, formulas and symbols in the Brown Corpus(Francis & Kucera, 1979) become **f and **s, high token count of “ f ” and “s” would be overrepresented in figure 19.

**A = ' (apostrophe, single quote)
**B = uncoded character(s)
**C = :
**D = (diaeresis or umlaut)
**F = formula
**I = ?
**J = begin symbol
**K = %
**N = begin major heading
NP = begin major division without heading
**P = end major heading
**Q = " (begin quotation)
**R = begin minor heading
RT = begin minor division without heading
**S = ;
**T = end minor heading

Table 19. Annotation in Brown corpus



Concordance Hits 1009

Hit	KWIC		
1	is therefore of order **f and class **	f.	A final class of exceptional lines is
2	, etc&. Thus **f if and only if **	f.	A function ~<g> such that **f, i&
3	is assigned as the information cell of **	f.	A third cell can be added by
4	its ~<n>th power 0; if we take **	f	above, that will be large enough. It
5	and only if ~<f> has the form **	f.	Accordingly, the 'functions' **f span the space
6	. 500 ~ml of 1~<M> aqueous **f with 1 ~g **	f	added are heated in a bomb at 170`~
7	corner points of the square, **f and **	f,	adjacent to ~<P> and ~<Q> respectively. As ~<
8	side and with free corners **f and **	f	adjacent to ~<P> and ~<Q> respectively. As ~<
9	.8% ~Cr as compared with 61.2% theory. However, **	f	adsorbs water from the atmosphere and this
10	r sections were similarly stained by comparable **	f.	After **f and **f were passed through
11	This indicates that increase in specificity of **	f	after passing it through ~DEAE-cellulose was
12	an interval on the other side of **	f.	Again, the analyticity of the two curves
13	is usually observed in a plot of **	f	against load ~<L>, having slope ~<k>, and **
14	cutting conditions, the chip exerts a force **	f	against the coating and an equal opposite
15	of cutting the chip exerts a thrust **	f	against the knife which tends to push
16	the coating and an equal opposite force **	f	against the knife in the plane of
17	). After titration of the liberated **f with **	f,	aliquots of the aqueous and of the
18	ted below 40`~F. When the temperature reached 32`~	F	all protozoan activity ceased; but as the
19	amount of some hydrous material other than **	f.	All subsequent measurements were made on materi
20	1:10 dilution of ~NS and **f or with **	f	alone. Unstained sections mounted in buffered gl

Figure 19. KWIC concordance or **f

The improvement is to check carefully what special annotation is used in the corpus and include these special usage of symbol in a stop list if they are not useful for the research.

iv) lemma form

Since our assumption of all comparison is that type frequency is correlated to the text difficulty, without further processing it ignored deviation complexity and forms variation. Even though some of the word types are grammatically related, such as *"Game, games, drive, driving, driven"*, they would still be considered as 5 different types. An improvement is constructing wordlists by lemma form or word family. For example, lemmatized wordlists would consider *"goes, went, gone, going and go"* as one form and list as go.

1.8. Applications

The studies in this section provide a clear overview of corpus approach in word level. The measurements can be applied to other linguistics researches as well as researches in other aspects, such as business and computational studies. It is noted that corpus approach is no only beneficial to linguistics studies, by manipulating the annotation method and source data, the same techniques can reveal customer behaviors, develop Artificial Intelligent, and even locate medical treatment solutions.

1.9. Conclusions

It is found that different measurements contradict with each other. Nearly half of the measurements show that *"D. Religion"* is more difficult, and the other half show the opposite. Each of these measurements reflect different quality of the text, thus the actual level of difficulties depends on the area we are focusing on. To sum up, "*D. Religion*" has higher vocabulary richness(HLR) and it reequires more specific knowledge to understand specific terms(TTR). While "*J. Learned*" has higher condensation of meaning(SD), and more types are required to get a fair amount of understanding of the text (VS and coverage).

After considering the limitations, most of the results supports our original hypothesis that "J. Learned" is more difficult than "D. Religion". While at the same time, this research pointed out some important factors what we have to take into account when using corpus approach in measurement.

1.10. References

Covington, M. A., & McFall, J. D. (2010). *Cutting the Gordian knot: The moving-average type–token ratio (MATTR)*. Journal of quantitative linguistics, 17(2), 94-100.

Covington, M. A. (2008). *Text Statistics*. [PDF file]. Retrieved from http://www.covingtoninnovations.com/mc/8570/TextStatistics.pdf

FANG, C. A. (2007). *English corpora and automated grammatical analysis*. Commercial Press.

Francis, W. N. & Kucera, H. (1979). *Brown Corpus Manual*. Department of Linguistics, Brown University, Providence, Rhode Island, US .

Gatt, A. (2011). *Corpora and Statistical Methods.* [Powerpoint slides]. Retrieved from http://staff.um.edu.mt/albert.gatt/teaching/dl/statLecture3a.pdf

Gatt, A. (2017). *LIN 3098 – Corpus Linguistics Lecture 5* [Powerpoint slides]. Retrieved from https://slidedocument.org/the-philosophy-of-money.html?utm_source=lin-3098-corpus-linguistics-lecture-5

Han, N. (2017). *Introduction to Corpora, Key Concepts.* [Powerpoint slides]. Retrieved from http://www.pitt.edu/~naraehan/ling1330/Lecture6.pdf

Lardilleux, A., & Lepage, Y., (2007). The contribution of the notion of hapax legomena to word alignment. *In Proceedings of the 4th Language and Technology Conference* (LTC'07), 458-462.

Zipf G. (1936). *The Psychobiology of Language*. London: Routledge.

Zipf G. (1949). *Human Behavior and the Principle of Least Effort.* New York: Addison-Wesley.

Chapter 2

2. Brief summary of Empirical studies of LOB Corpus and ICE Corpus

2.1. LOB Corpus

Our group also did a study on the Lancaster-Oslo-Bergen Corpus(LOB Corpus). It is an annotated corpus with POS information. It follows the same corpus design of Brown Corpus, thus it is also a British sister corpus of Brown Corpus.

For LOB corpus, our group compared *"J. Learned"* and *"P. Romance Fiction"*. We hypothesize that the usage of words would be quite different for these two corpora, since they are in the informative and imaginative prose respectively.

	J	P
V	Describing (objective)	Mental (subjective)
V	Less verbal	More Verbal
V	Action / Using tool	Motion
N	Inanimate	Animate
N	Abstract noun	Body part
PN	It, He, We, They	I

Table 20. Summary of the LOB corpus

In the result, we found that *"J. Learned"* has more INANIMATE and ABSTRACT nouns, while *"P. Romance Fiction"* has more ANIMATE and BODY PARTS nouns. This pattern matches the nature of the text type. INANIMATE and ABSTRACT nouns are used to describe the tools and phenomenon in science, while ANIMATE and BODY PARTS are frequently used in novel to describe the interaction between characters.

Another pattern is the usage of pronoun and verbs. *"J. Learned"* has more 3rd person pronoun, while *"P. Romance Fiction"* has more 1st person pronoun. Also, "*J. Learned"* has more describing, action(using tool) and instrumental verbs ,while *"P. Romance Fiction*" has more mental, motion, and communication verbs. It is because academic articles in *"J. Learned"* tends to be more objective by using 3rd person pronouns, while subjective mental thought or action is always involved in fictions.

2.2. ICE Corpus

The last corpus we used in the lecture is the International Corpus of English Corpus(ICE). It was a set of corpora containing different national or regional variety of English, including Hong Kong. It is mainly used for comparative studies of English in a worldwide scale. Each corpus consists of 1 million-word of spoken(60%) and written(40%) English.

In the spoken section, both dialogues and monologues are included. In the written section, both printed and non-printed text are included. It is a great tool for comparing formal and informal usage of languages.

By comparing the syntactic categories and functions, it is found that the spoken section used more gap fillers, pronouns as NPHD, and conjunctions. Overall the spoken section is more casual and conversational. In contract, for the written section, it is more formal in terms of higher usage of determiner and prepositions. The complex sentence structure and more informational text are the distinctive factures of formal usage.

Furthermore, we located that more adverbial clauses were used in written text than spoken sections. This pattern has been mentioned by Fang (2006).

By studying corpus data with respect to their categories and functions, the identified patterns can be benifical to other studies in rhetorical relations, lexical model of coherence and generation (Nikitina & Padó, 2013).

Fang, A. C. (2006). A corpus-based empirical account of adverbial clauses across speech and writing in contemporary british english. In Advances in Natural Language Processing (pp. 32-43). Springer, Berlin, Heidelberg.

Nikitina, O., & Padó, S. (2013). A corpus study of clause combination. In Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers (pp. 179-190).

Chapter 3

3. Extra Application of learnt concepts in Amazon book review dataset

3.1. Introduction

In this chapter, the measurements in chapter 1 will be applied to the Amazon book review dataset. The research question is that ***"What is the differences between high and low helpfulness reviews?"*** The aim of this study is to find out the distinctive features of high-helpfulness reviews by corpus approach. At the end, it is hoped that this analysis could help companies like Amazon to locate the potential high-hitting reviews which may help other customers, and thus enhance their marketing strategy.

3.2. Corpus design

3.2.1. Sampling

An open source dataset of Amazon Book Reviews, which was originally constructed by J. McAuley(2015), would be used in this project. It contains around 8.9 million book reviews (approximately 1.5 billion words) from Amazon spanning May 1996 - July 2014.

In the dataset, each review entity consists of the following items:

Annotation in the file	*Meaning*
(1) Reviewer ID	*- ID of the reviewer, e.g. A2SUAM1J3GNN3B*
(2) asin	*- ID of the product, e.g. 0000013714*
(3) reviewerName	*- name of the reviewer*
(4) helpful	***- helpfulness rating of the review, e.g. 2/3***
(5) reviewText	***- text of the review***
(6) overall	*- rating of the product*
(7) summary	*- summary of the review*
(8) unixReviewTime	*- time of the review (unix time)*
(9) reviewTime	*- time of the review (raw)*

Among these, *(4) and (5)* would be most useful to this study. The helpfulness ratings are given by other customers, a high helpfulness rating means the review is well written and helpful for other customer's judgements.

Two subsections have been selected from the original dataset and form two contrasting datasets according to the following criteria: 100000 reviews with the highest helpfulness rating and 100000 reviews with the lowest helpfulness ratings. At the same time, 100000 reviews from the remaining are selected randomly as control set.

3.2.2. Part of speech (POS) Tagging

A filter list[3] based on the lecture documents was used for further categorization. Part-of-Speech tagging was performed by TagAnt (Anthony, 2015).

[3] It consists of pronouns, function words except for negation markers "not" and conjunctions "but"

3.3. Hypothesis

The hypothesis is that the "high helpfulness reviews" are more complex in structure and lexical usage. Therefore, there would be differences between the two subsections in terms of TTR, SD, etc.

3.4. Discussions

	Top	Bottom	Random
Entities	100000	100000	100000
Token	21513810	13050988	19827789
Types	213506	168268	208709
Token per entry	215.1381	130.50988	198.27789
TTR	0.992%	1.289%	1.053%
SD	46.046%	45.666%	45.874%
Content(tokens)	9906226	5959854	9095754
Function(types)	397	403	407
HL	114723	91793	114145
HLR	0.533%	0.703%	0.576%

Table 20. Overall measurements

According to table 20, it is shown that there are a few differences between the top and bottom helpfulness reviews. Some figures will be discussed in the following sections.

3.4.1.1. Tokens

	Top	Bottom	Random
Entities	100000	100000	100000
Token	21513810	13050988	19827789
Types	213506	168268	208709
Token per entry	215.1381	130.50988	198.27789

Table 21. Types and tokens

The total number of entries in the two subsets are the same as 100000, however, the total number of tokens in the two subsets differs. The total tokens in low helpfulness reviews are much less than that of "top" and "random". This difference is much clearer in the measurement of token per entry.

3.4.1.2. Token per entry

According to table 21, the two subsets differ greatly in terms of the review length. The number of tokens per entry in high helpfulness reviews(215 tokens) are higher than that of the low helpfulness reviews(130 tokens) and random reviews(198 tokens).

The implication of this figure is that short reviews cannot provide enough information for other customers. Given the fact that "*short reviews are usually not helpful*", *"the longer the review, the more helpful it is"* seems to be true as well. However, there should be other criteria which determine the high helpfulness of reviews.

3.4.2. Semantic density

	Top	Bottom	Random
TTR	0.992%	1.289%	1.053%
SD	46.046%	45.666%	45.874%
Content(tokens)	9906226	5959854	9095754
Function(types)	397	403	407
HL	114723	91793	114145
HLR	0.533%	0.703%	0.576%

Table 22. Overall measurements

Since the subsections are of different size, it is not comparable by direct comparison of tokens. For semantic density, SD of *"Top"* is slightly higher than the two others.

3.4.3. Standardized Type Token Ratio (sTTR)

According to table 22, even though some of the measurement show that the three subsets are different from each other, these measurements could all be affect by the corpus size. Therefore, sTTR is needed for further comparison.

	Top	Bottom	Random
sTTR	1.234%	1.289%	1.226%
Type(stand.)	13050988	13050988	13050988
Token(stand.)	161092	168268	159953

Table 23. sTTR comparison

By using a standardized subsection of each part, the sTTR show that the low-helpfulness reviews have a slightly higher sTTR(1.289%). The high helpfulness reviews and random control are much closer(1.234% and 1.226%).

3.4.4. Part of speech (POS) distribution

Since overall distribution does not show many significant differences among these sections, POS distribution also take into account. The POS tagging is done by using TagAnt.

POS	Top/Bottom	Percentage
JJ	T	7.65%
JJ	B	7.65%
NN	**T**	**26.35%**
NN	**B**	**26.49%**
V	**T**	**12.76%**
V	**B**	**13.20%**
RB	T	5.79%
RB	B	5.76%

Table 24. POS comparison on the first 950 types

In table 24, the POS distribution show that among the most frequent types, some of the part of speech account for the same percentage of the text, while high helpfulness reviews use slightly less noun(26.35%) and verb(12.76%) than that of low helpfulness reviews(26.49% and 13.20%). However, further investigation with detailed tagging is needed to locate their differences in functions.

3.5. Conclusion and Impacts

This project demonstrated an approach of using part of the Amazon Book review dataset to produce key figures on word usage. It is found that high helpfulness reviews are usually longer in length (~215 tokens), slightly higher in semantic density, slightly lower in sTTR, and the commonly used types account for lower percentage in noun and verb with respect to other POS. Further research could be carried out by using a larger portion of the dataset or even other categories including music reviews, CDs, and clothes.

3.6. Online Applications

The result of this study together with my course work in another course regarding sentiment analysis in the same dataset could be used to develop review analysiser/predictor which help locating the potential high-hitting reviews. It is important to locate the potential high-hitting reviews for new products, especially when there are only a few new reviews. Companies can display the helping reviews to facilitate customer's choice.

For real applications, full wordlists would be released on github after final submission. It will be combined with other findings from another course and used in my sentiment analysis project[4].

[4] https://github.com/kennethli319/Sentiment-analysis-tool

3.7. References

Anthony, L. (2015). TagAnt (Version 1.1.0) [Computer Software]. Tokyo, Japan: Waseda University.

McAuley, J., Targett, C., Shi, Q., & Van Den Hengel, A. (2015, August). Image-based recommendations on styles and substitutes. *In Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information* Retrieval (pp. 43-52). ACM.