Exploratory Text Analysis of the Biblical Book of Proverbs

Daniel Trevino
Applied Data Science

Abstract

This study identifies subtle similarities and differences between two classic Bible translations, the KJV and the NIV. These similarities and differences are assessed to identify if the overall meaning and context of the entire book of Proverbs is changed. Analysis is performed on the entire book and the study is identified as a baseline application for future exploratory text analysis on other religious books.

1 Introduction

The Protestant Christian Bible contains 66 books. The 66 books are broken into two parts, the Old Testament and the New Testament. The Old Testament ("Tanakh" in Hebrew) contains 39 of the books. The 39 books are broken into 4 further sections, or classifications. They are the Pentateuch ("Torah"), History ("Neviʾim"), Wisdom, and Prophets. Amongst the Wisdom section, there are 5 books, Job, Psalms, Proverbs, Ecclesiates, and Song of Soloman. The Book of Proverbs, contains proverbial statements written by King Solomon which are stemmed in morality, such as, Proverbs 22:6 "Start children off on the way they should go, and even when they are old they will not turn from it."

Proverbs is broken into 6 Topics by many mainstream pastors (Tautges, 2019), however there is a rabbinical topic model as pointed out by Michael Fox (Fox, 1968) originating from the Hebrew text. Fox states the critical term for Proverbs is the word חָכְמָה (chokmah), meaning wisdom, and it is broken into 4 topics, or senses, with the 1st topic having 3 subparts, or a total of 7 topics (Fox, 1968). Others such as Buchanan, propose Proverbs is broken into 44 topics (Buchanan, 1988). Given the variation in number of topics, part of this study aims to identify how many topics are valid and if topics vary between two different translations of the bible.

1.2 Purpose

For the purpose of analyzing the value of the text, the Biblical book of Proverbs is used for this study. However, the book of Proverbs is translated from Hebrew to English and in some cases from Hebrew, to Latin, to English. The translation from Hebrew to English will not be analyzed for this study, but a comparative between two of the largest Bible translations, the King James Version (KJV) and the New International Version (NIV). More specifically, the Book of Proverbs will be analyzed for topic modeling and tone to answer several research questions. 1) What are the main topics (themes) in each translation? 2) What are the major tones in each translation? 3) How much similarity are there between the two translations? 4) Are there differences in the two translations? 5) Is there an effect to the overall meaning of the text and proverbial statements by difference in translation? The intent behind the textual exploratory analysis of Proverbs is to establish a baseline for future text mining of the 5 wisdom books of the bible (Job, Psalms, Proverbs, Ecclesiastes, Song of Songs) and develop a foundation for exploration in non-Christian Religious literature.

2 Method

2.1 Toolkit

The primary source for tools to evaluate the text is python 3.7 utilizing SKLearn. Amongst the packages, the Latent Dirichlet Allocation (LDA) package ideally produced the most meaningful results after comparing the LDA model to the Non-negative Matrix factorization (NMF) and the Latent Semantic Indexing/Analysis (LSI). The LDA model will be the primary tool for topic modeling throughout this study. Additionally, to analyze the tone of each verse (915 verses), the IBM Tone analyzer is used.

2.2 Data Set

The text is processed from two GitHub repositories with the Bibles in XML format. To

structure the text by chapters and verse, the XML is converted to JSON and then normalized into a flattened Pandas Dataframes. The table below shows the original data structure as depicted in the dataframe. The same processes is repeated for the New International Version (NIV) of the Bible.

Table 1 King James Version Dataframe

	Chapter	Verse	KJV
0	1	1	The proverbs of Solomon the son of David, king...
1	1	2	To know wisdom and instruction; to perceive th...
2	1	3	To receive the instruction of wisdom, justice,...
3	1	4	To give subtilty to the simple, to the young m...

Analyzing basic statistics between the two Bibles show in Table 2 the KJV uses slightly longer words and is attributed to the type of English the KJV uses. The KJV was compiled around 1611 whilst the NIV was published in 1978 as a completed book. To illustrate further comprehension of the text and distribution between the two translations, more detailed counts and procedures will be assessed.

Table 2 Average Word Size

	Chapter	NIV_Avg_Word	KJV_Avg_Word
count	915.000000	915.000000	915.000000
mean	16.101639	4.313353	4.356395
std	8.931269	0.563491	0.539063
min	1.000000	2.615385	2.972222
25%	8.000000	3.930952	4.000000
50%	16.000000	4.250000	4.294118
75%	24.000000	4.642857	4.647059
max	31.000000	6.833333	7.100000

2.3 Experimental Procedures

2.3.1 Data Preprocessing and Vectorization

After importing the text from the book of Proverbs into a dataframe, the text mining process began. The first procedure identified the word count for both the KJV and the NIV. Table 3 shows the top 10 words for each version.

Table 3 Top Words in Proverbs

KJV		NIV	
shall	259	man	182
man	157	lord	87
thy	110	wicked	82
thou	102	heart	70
wicked	89	like	67
lord	87	righteous	64
heart	81	wise	61
wise	66	wisdom	51
unto	62	life	48
thee	61	son	45

Another technique to visualize the top words within the Proverbs is to show as a word cloud. The following two figures show the KJV first followed by the NIV word clouds.

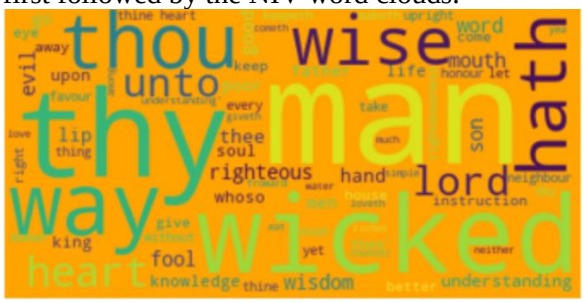

Figure 1 KJV Proverbs Word Cloud

Figure 2 NIV Proverbs Word Cloud

The interesting dynamic with the two word clouds and the top word count is the terms returned. Being familiar with the actual text, the word "wicked" stands out as being something bad, which in the context of the verse, it is. However, many verses often use "not" in front, therefore negating the word wicked. Thus, selected words such as "thy" and "hath", which are synonymous with "you" and "have", both stop words, are added and the word "not" and "will" is removed from the stop-words list to equate the top synonym terms. Recomputing the KJV and NIV word count results in the word "not" being the second most used word whilst will and shall are now the top words with 259 and 224 respectively.

The next step in preprocessing is the vectorization of the words. Using bag of words, the entire book of Proverbs is used for the vectorization. This process involved using the TFIDF vectorizer from sklearn.feature_extraction.text. Two iterations are performed with the vectorizer. The first iteration involved using Unigrams only. For the purpose of developing more coherence of the meaning and context, Unigram plus Bigram is used for analysis as the book of Proverbs contains many axiomatic sentences. Additionally, the following parameters (max_df=0.98, min_df=1, ngram_range=(1,2)) are adjusted to provide the most identifiable topics, discussed in the next section.

2.3.2 Topic Labeling and Interpreting

Establishing a baseline for the vectorization, the LDA model is used as it produced more identifiable topics than NMF and LSI. The next step involves choosing the number of topics. Given modern English interpretations do not have a set standard of how many topics, two computational attempts are made to identify the number of topics. The first involved analyzing the perplexity. The initial test involved choosing between 1 and 44 topics, based on the resources where topics are defined. The resulting figure illustrates, there are and infinite number of topics for the Proverbs.

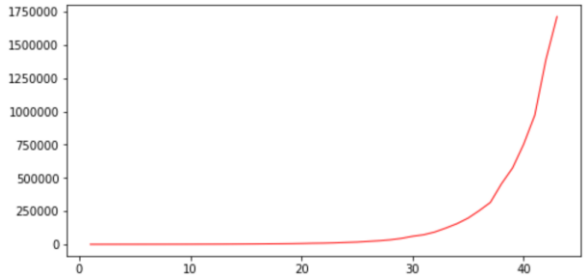

Figure 3 Perplexity Topic 1 – 44

The second attempt after perplexity is calculated involved narrowing the topics down to 7, based on the Hebrew lexicon for the word "wisdom". Figure 4 shows flattening of the line around 4 – 5. Therefore, based solely on perplexity, 4 topics are chosen.

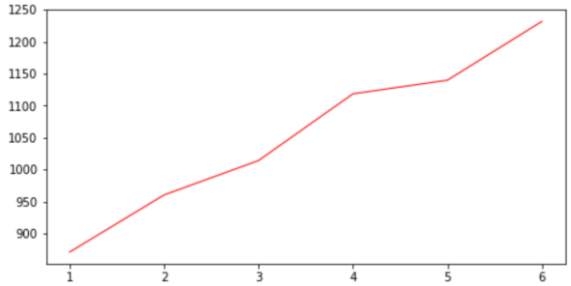

Figure 4 Perplexity Topic 1 - 7

The second method for analyzing the number of topics, is based on a graphical approach. Using pyLDAvis, the topics are plotted on a quadrant chart and where the bubbles overlap identifies a closer relationship in topics. Plotting pyLDAvis with 7 topics produces the following chart. Dimensionally, no topics overlap and there is one strong topic. This is also prevalent when assessing the KJV LDA [0.03338731 0.03339488 0.03339782 0.03339435 0.03338643 0.79964387 0.03339533] and the NIV LDA [0.79969964 0.03338266 0.03338154 0.03338826 0.03338265 0.0333827 0.03338255] which shows one strong topic at 0.799 and all other topics at 0.033. Therefore, as the Hebrew lexicon determined, 4 topics is chosen for further analysis and comparison between the KJV and the NIV.

Figure 5 Topic pyLDAvis 7 topics

2.3.3 IBM Tone Analyzer

The last experimental procedure for comparison involves the IBM tone analyzer. The IBM Tone Analyzer uses linguistic analysis to detect joy, fear, sadness, anger, analytical, confident and tentative tones found in text. This resulted in passing each verse (915 total) though the IBM Watson cloud tone analyzer and then labeling each verse for the tone. The following table is a representation of the tone analyzer for all text, minus the text the tone analyzer did not process.

Table 4 Tone Analysis for Proverbs

KJV		NIV	
Analytical	266	Analytical	268
Joy	176	Joy	177
Sadness	53	Tentative	68
Anger	50	Sadness	45
Confident	49	Confident	43
Tentative	36	Anger	40
Fear	24	Fear	29

3 Results

3.1 Topic Models KJV and NIV

LDA Topic modeling for the KJV and NIV book of Proverbs, with the number of topics set to 4, produced the below topics. The topic names were added based on familiarization with the text. Following are the topic by word distribution and the distribution of topics by verse. The first table and bar chart depict the KJV and the second table and bar chart depict the NIV version.

3.2 KJV Topic Models

The Topics in table 5 below highlight the top 10 words per the 4 topics. Given the words and the underlining theme of wisdom for Proverbs, all 4 topics will be oriented toward topics dealing with wisdom. Topic modeling in order is:

Topic 0 -> Perseverance and Hard work
Topic 1 -> Teach Children Virtues
Topic 2 -> Respect Family
Topic 3 -> Speak Wisely

Many words in the bible have dual meaning. The topics chosen above incorporate these dual meanings and addresses how each deal with wisdom.

Table 5 NIV Topic Key Words

Topic	Topic 1	Topic 2	Topic 3
hands	Man	not	shall
poor	excellest	household	not
shall	daughters	thine	man
fruit	virtuously	shall	lord
not	excel child	wicked	wicked
fruitful	daughters	hands	way
man	known	women	heart
night	Like	children	righteous

Plotting the distribution for the overall topics by verse shows Topic 3 being the most dominant topic in the Proverbs for the KJV with topic 0, then 2, then 1. Thus, by topic distribution, speaking wisely is the most predominant topic in the KJV while teaching children virtues is less prevalent.

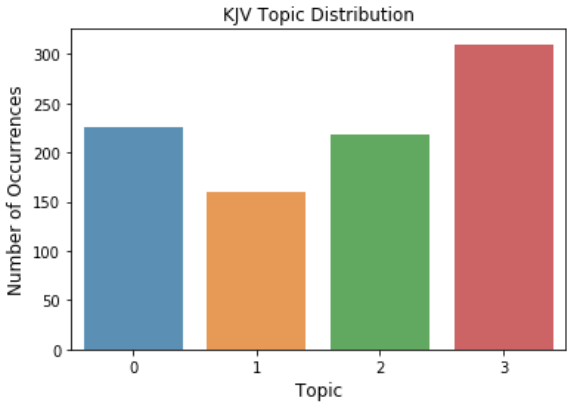

Figure 6 KJV Topic Distribution

3.3 NIV Topic Models

The following are the topics assigned to the NIV LDA model. The same contextual understanding of the dual meaning is used in assigning topics.

Topic 0 -> faith builds perseverance
Topic 1 -> Self Control
Topic 2 -> Seek Guidance
Topic 3 -> Speak Wisely

The topics chosen above illustrate the top models from the NIV version.

Table 5 NIV Topic Key Words

Topic	Topic 1	Topic 2	Topic 3
poor	will	man	not
makes	wicked	not	man
will	man	will	lips
linen	husband	like	will
noble	lord	son	wisdom
things	not	lord	tongue
holds	evil	food	does not
wife	eyes	wicked	like

As seen in the NIV topic distribution below, the predominate topic is referencing speaking wisely to others, while the second most important topic is about self-control. These topics illustrate the major themes of the Proverbs in the NIV.

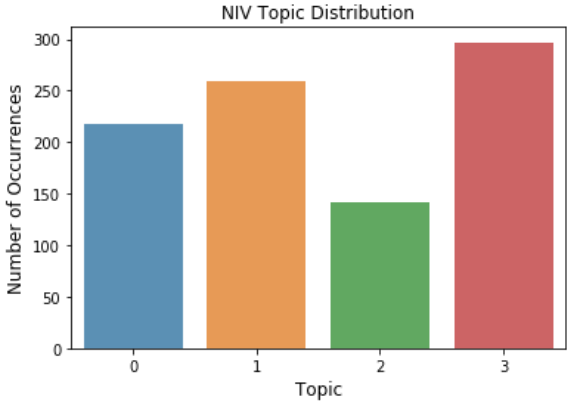

Figure 7 NIV Topic Distribution

3.4 KJV and NIV Topic Comparison

The KJV and NIV are similar in many aspects. Specifically analyzing the book of Proverbs in this study points out similarities between words, the term frequency, but also identifies differences. The largest differences are a result of the topic modeling, where using the same parameters with TFIDF vectorization, Topic modeling, and Tone Analysis produced different results. The next section will further elaborate on the similarities and differences.

4 Discussion

4.1 Similarities

The KJV and NIV closely match in term frequency, major topic, and in tone. After adjusting the stop words, the top 5 words all contained a similar frequency, which is expected. Both translations idealistically should be similar, as the original source of translation is the same. Additionally, when analyzing the topic, the LDA modeling produced one primary topic with roughly 80% being to speak wisely. This is prevalent in both the KJV and NIV version.

One insight of note with topic modeling, is when analyzing the perplexity, after about 30 topics, the perplexity grows exponentially. However, when analyzing for fewer topics, the perplexity graph flattens between 4 and 5 topics. Looking back at the original Hebrew, Fox explains the 4 main definitions of wisdom in Hebrew are (1) practical sagacity, (2) ethical-religious wisdom, (3) speculative wisdom, and (4) intellectual wisdom (Fox, 1968). Interestingly, these 4 wisdom types are more closely aligned with the NIV 4 topics than the KJV topics.

There are also similarities when assessing similarities with tone as Proverbs contains more Analytical and Joyful tone overall for both translations than the other 5 tones. Other than these similarities, there are more differences.

4.2 Differences

The topic modeling produced three separate topics between the two translations. Of the three topics, only topic 0 is close in relation to perseverance. However, the top words output by the LDA model does not produce enough similarities between the two translations. Whilst topics 1 and 2 are make up a larger portion of the overall topic distribution and are not similar. However, given the perplexity, the differences may be attenuated by the lower number of overall topics selected. Thus, given a low topic is selection, it is feasible for the two topics output by the KJV model to be included within the NIV model and the two within the NIV model are incorporated with the KJV.

Assessing the tones is the most surprising as the order of tone is not symmetrical. The NIV is more tentative when translating from each verse than the KJV. The tentative tone appears to be the largest difference between the two translations and may be indicative of different prose and period of publication. The NIV, based on tone and verse analysis also appears to explain words from the Hebrew translation in a more sensitive nature, while the KJV tone appears to be more direct and use negation of words more often.

5 Conclusion and Limitation

The book of Proverbs from the KJV and NIV is an instrumental religious and historical text. Whether 4 topics or 44, the overall book has similar topics, as it should. Additionally, there are many similarities between the two translations that speak of the unique nature of time which they were written. There are differences as well between the two translations, albeit the differences are not large in nature to conclude the two translations are semantically different. Further exploratory analysis is still needed to pinpoint the exact differences. Comparing several of the text where both tone and topic differentiated does address a by verse bases, there are semantic differences, but the focus of the study was to address the overall context of the entire book of Proverbs.

Several limitations, in regards to data processing hindered further analysis of the data. Additionally, the Watson Tone Analyzer did not take several verse from both the NIV and KJV, likely due to processing of the sentences without stopwords. Concluding, the text analysis using the Bag of Words (BOW) technique is insightful and progressive as it allows quicker comprehension of many text documents as well as enables identification of trends or elements of trends in text data. This study illustrated to me the need to compare the translated versions to their original and points to further research and personal study on religious text.

References

1) Buchanan, Hugh. “The Proverbs.” Http://Fridaysunset.net/Articles/ClassifiedProverbs.pdf, Http://Fridaysunset.net/Articles/ClassifiedProverbs.pdf, 1988.

2) Fox, Michael V. “Aspects Of The Religion Of The Book Of Proverbs.” *Hebrew Union College Annual*, Vol. 39, 1968, Pp. 55–69. *JSTOR*, www.jstor.org/stable/23503072

3) Hu, Wei. "Unsupervised Learning Of Two Bible Books: Proverbs And Psalms." Sociology Mind 2.03 (2012): 325. [4] McCallum, A. K. (2002).

4) Tautges, Paul. “6 Categories In Proverbs.” Counseling One Another, 20 Apr. 2019, Counselingoneanother.Com/2012/10/03/6-Categories-In-Proverbs/.

5) “The Proverbs.” Edited By The Editors Of Encyclopaedia Britannica, Encyclopædia Britannica, Encyclopædia Britannica, Inc., 16 June 2008, Www.Britannica.Com/Topic/The-Proverbs.

6) “Tone Analyzer: Using The General-Purpose Endpoint.” Edited By Ibm Watson Contributors, Ibm Cloud Docs, 7 Mar. 2019, Cloud.Ibm.Com/Docs/Services/Tone-Analyzer?Topic=Tone-Analyzer-Utgpe&Locale=En-Us.