A Comparison of New Testament Bible Translations in English

Executive Summary

This project compares 48 translations of the New Testament using analytical methods for natural language. The questions include: Is there a qualitative change over time in the translations, Are there material changes in the quantity or quality of words used in translation, How different do translations look from one to the other?

Analysis will start with obtaining basic statistics about the texts in the data set. This will provide insight into word quantity and suggest if word selections have changed in terms of word size. It will also help identify volumes that may be problematic, for example, including a children's bible would not be representative of typical translations.

The next stage will be computing cosines of similarity. All texts should be very similar, but not exactly. Texts under different names but essentially the same can be removed from further analysis. This analysis will also suggest what texts are most different.

Finally, each text will be scored for polarity (net positive or negative), subjectivity (closer to 1) or objectivity (closer to 0), and net sentiment (negatively rated words versus positively rated ones). The publication date will be compared with the sentiment score to identify trends.

Data Preparation

Selections were made from seven websites offering free downloads of bibles published over the past 500 years. Most of the texts were not downloadable as simple text, but as some format like PDF. No effort was made to machine-read the PDF, rather, a manual copy and paste of the new testament into a .txt file was performed to create the data set which may be downloaded from https://github.com/bertoli517/DATA532 All texts were formatted into text files containing all new testament books from Matthew to Revelations. Automated processing consisted of the following steps:

1. Basic Statistics: All 48 bible text files were placed into one directory. 'Basic_Stats.py' was run from another directory. 'Basic_Stats.py' will need to have its path modified on line 17 to point to the data directory, and line 93 will need to point to a different output directory for a successful run. The successful run produced the information in Table 1 saved as a .csv file in the output directory. File names simplified to save space in the table. The python program used in this analysis used regular expressions to remove punctuation and generally clean the text. Four bar graphs (Figures 1 & 2) are plotted to illustrate word count, median word size, variance and standard deviation by text.

2. Cosine of Similarity: The first half of 'Data532Main.py' was dedicated to computing and presenting the cosine of similarity. Line 24 will have to be changed to point to the data directory where the texts are located. Regular expressions were used again to clean the text before processing. A list was compiled of all the words in all the texts in the data directory. Duplicate words were removed from the word list with the 'set' function. The term-frequency matrix was made using the de-duplicated word list and counting occurrences in each text. Then a normalized tf-IDF matrix was computed. The 'cosine_similarity' function of scikit-learn was used to compute the cosine matrix from

the tf-IDF matrix. The distribution of cosines of similarity is illustrated in Figure 5 and a colormap of the matrix is illustrated in Figure 6. Using Spyder's variable explorer one can also see the cosine matrix shaded with actual values (sample illustrated in Figure 6).

3. Polarity, subjectivity, and sentiment analysis: The other half of 'Data532Main.py' computed and plotted polarity (how positive or negative), subjectivity, and sentiment analysis of the words in the texts. These results are illustrated in Figures 7, 8 and 9.

Results

Basic statistics showed that most of the texts had similar characteristics for number of words and word size. Outliers had specific differences: Darby_1889 had more and smaller words because of formatting issues that broke words up with spaces. Douay-Rheims_1752 contains more words because there is commentary interspersed within the text. This is also true of EasyToReadVersion_2006, RevisedEnglishVersion_1989, TheAmplifiedBible_1984, and TranslationforTranslators_2008. The PeoplesNewTestament_1889 is the largest with extensive footnotes on the pages. TheEverydayBible_1924 appears to be an abridged bible, coming in as the smallest.

Figure 1 – Words Per Text

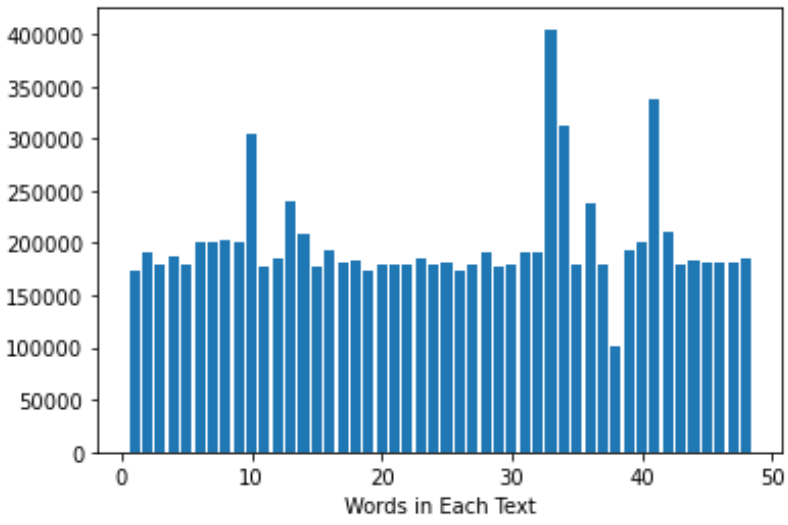


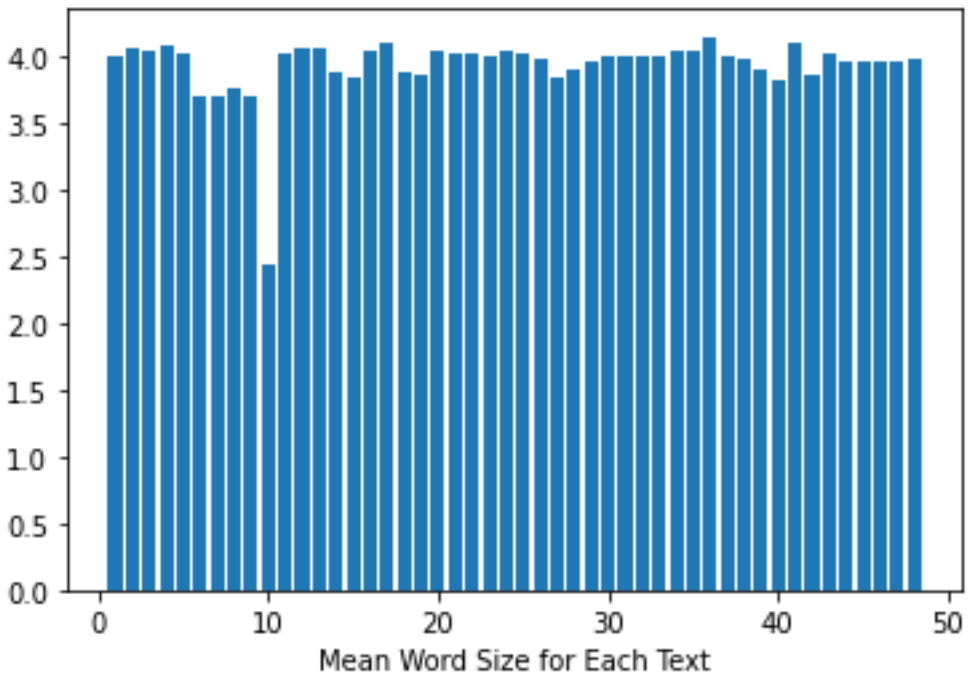
Table 1 – Bible Text Statistics

Text	Word Count	Mean Word Length	Variance	STDev	Z-Score
Darby_1961	185,890	4.0651	4.3193	2.0783	-0.0313
TT_2008	337,721	4.1003	4.2059	2.0508	-0.0489
ESV_2001	176,708	3.8420	3.7232	1.9296	0.0819
TNT_1534	210,941	3.8652	4.0074	2.0018	0.0673
NWB_1833	178,759	3.9905	4.0570	2.0142	0.0047
NLT_1996	191,711	3.8944	3.9189	1.9796	0.0533
WEBB_2014	181,893	3.9587	4.0556	2.0139	0.0205
LEB_2010	185,489	4.0087	4.3004	2.0738	-0.0042
NET_2016	181,357	4.0189	4.2458	2.0605	-0.0092
AS_1901	186,430	4.0841	4.2414	2.0595	-0.0408
ETRV_2006	207,817	3.8873	3.3175	1.8214	0.0619
NIV_XXXX	173,348	3.9699	4.0050	2.0012	0.0150
TMB_2002	201,321	3.8245	4.0290	2.0072	0.0874
BBE_1965	201,209	3.6989	3.0328	1.7415	0.1729
KJCB_2005	179,316	4.0393	4.0471	2.0117	-0.0195
GWN_1995	183,819	3.8872	3.8867	1.9715	0.0572
ABUB_1866	173,832	4.0016	4.1046	2.0260	-0.0008
OEB_US_2020	191,090	3.9990	4.1929	2.0477	0.0005
ASB_1881	192,053	4.0598	4.2814	2.0692	-0.0289
AV_1769	180,128	4.0232	4.0921	2.0229	-0.0114
TAB_1984	238,352	4.1453	5.0928	2.2567	-0.0644
ULB_2017	182,834	3.9686	3.9977	1.9994	0.0157
FBV_XXXX	193,086	4.0320	4.4898	2.1189	-0.0151
TEB_1924	100,553	3.9727	3.7372	1.9332	0.0141
ASV_1901	180,296	4.0419	4.0862	2.0214	-0.0207
WEBC_2014	181,893	3.9587	4.0556	2.0139	0.0205
NKJV_1982	178,847	3.8434	3.9098	1.9773	0.0792
DR_1752	240,646	4.0654	4.3493	2.0855	-0.0314
TGNB_1992	193,737	3.8982	3.6030	1.8982	0.0536
Tyndale_1530	179,808	4.0262	3.9452	1.9863	-0.0132
LSV_2020	179,491	4.0353	4.5570	2.1347	-0.0165
WEB_2014	181,893	3.9587	4.0556	2.0139	0.0205
OEB_2020	191,090	3.9990	4.1929	2.0477	0.0005
KJV_2006	179,469	4.0232	4.0649	2.0162	-0.0115
CEVUK_1995	201,209	3.6989	3.0328	1.7415	0.1729
HCSV_2003	173,632	3.8642	4.0513	2.0128	0.0674
YLT_2013	185,716	3.9825	4.1436	2.0356	0.0086
RVA_1895	178,838	4.0406	4.0789	2.0196	-0.0201
Darby_1890	177,816	4.0275	4.1843	2.0456	-0.0135
NTM_1913	177,340	3.9577	4.0410	2.0102	0.0210
GB_1599	180,480	4.0990	4.0121	2.0030	-0.0494

Text	Word Count	Mean Word Length	Variance	STDev	Z-Score
TEMTV_2014	179,145	4.0020	4.3455	2.0846	-0.0010
Darby_1889	304,478	2.4430	3.4500	1.8574	0.8382
BSB_2016	201,209	3.6989	3.0328	1.7415	0.1729
KJVBible_1989	178,497	4.0274	4.0415	2.0104	-0.0136
REV_1989	311,766	4.0312	4.1541	2.0382	-0.0153
CPDB_2010	202,567	3.7662	4.5494	2.1329	0.1096
PNT_1889	404,683	4.0059	4.2250	2.0555	-0.0029

Darby_1889's small word size (mean = 2.443) is due to formatting issues which inserted spaces into words causing word size to be calculated to be too small. All other texts have comparable mean word sizes.

Figure 2 – Mean Word Size



Examining the Cosine of Similarity Matrix on the first pass over the original 48 texts, several of the texts had a value of “1” off the main diagonal. These were removed since they were identical to other other texts under different titles. This was not expected and it affected five texts which were removed from further analysis.

Table 2 – Removed Texts

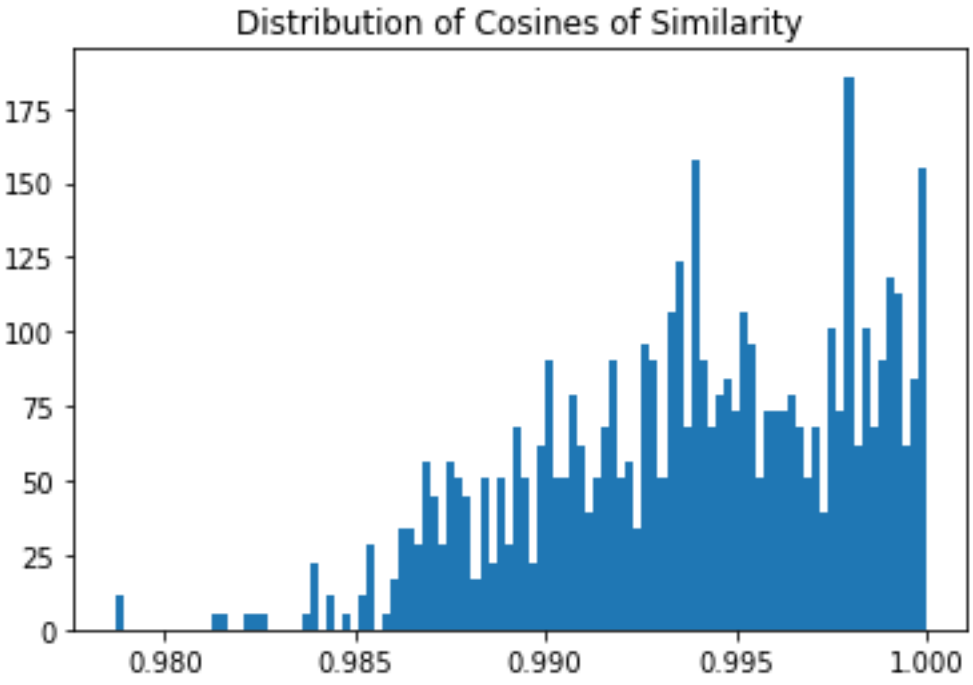
BereanStandardBible_2016
ContemporaryEnglishVersionUK_1995
OpenEnglishBible_US_2020
WorldEnglishBible_2014
WorldEnglishBibleCatholic_2014

Figure 3 – Cosine of Similarity Matrix, Partial

	3	4	5	6	7	8	9
0	0.997607	0.998256	0.995863	0.995863	0.988386	0.995863	0.990317
1	0.998963	0.999354	0.99369	0.99369	0.986201	0.99369	0.988416
2	0.999257	0.999741	0.993689	0.993689	0.985958	0.993689	0.987635
3	1	0.999078	0.99315	0.99315	0.985056	0.99315	0.987137
4	0.999078	1	0.993381	0.993381	0.985782	0.993381	0.987489
5	0.99315	0.993381	1	1	0.986876	1	0.987848
6	0.99315	0.993381	1	1	0.986876	1	0.987848
7	0.985056	0.985782	0.986876	0.986876	1	0.986876	0.981121
8	0.99315	0.993381	1	1	0.986876	1	0.987848
9	0.987137	0.987489	0.987848	0.987848	0.981121	0.987848	1
10	0.997389	0.997916	0.99578	0.99578	0.988257	0.99578	0.991198

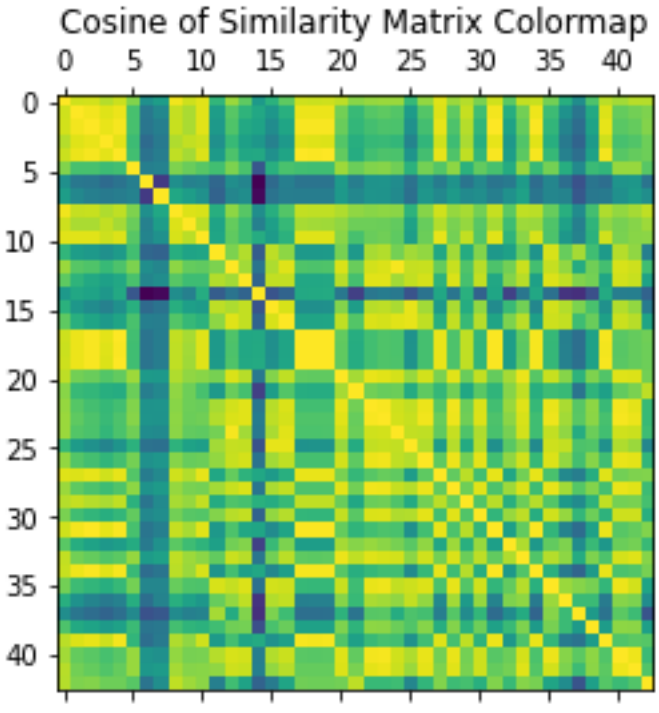
Re-running the cosine of similarity matrix after these five texts were removed, and converting the matrix into a list gave the bar chart in Figure 4. Considering the cosine values that fell 1.5 times below the inter-quartile range (< 0.988), four texts stand out as different. Darby_1889, which was already identified as having formatting issues, EasyToReadVersion_2006 and TranslationforTranslators_2008, which both have in-line footnotes on the verses, and the CatholicPublicDomainVersion_2010 which may be due to some character that was not stripped during clean-up, such as ‘{}’ used in citing verses.

Figure 4



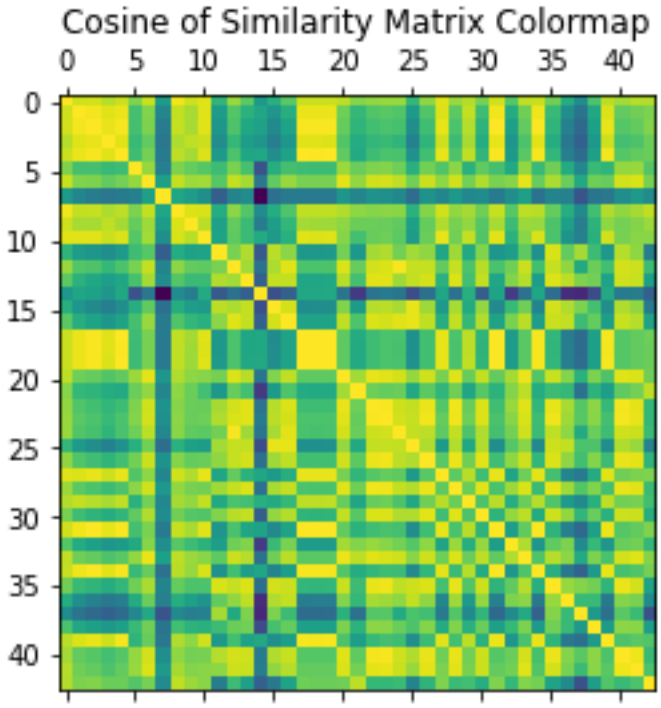
While the variable explorer of Spyder makes it easy to explore the values of the cosine matrix, placing the output into a readable table is not as practical as plotting a color map of the matrix, Figure 5. Here we can see that items 6, 7, and 37 are most different from all other texts. These three lines correspond to the three texts cited above that are most distant from the mean of cosines (< 0.988).

Figure 5



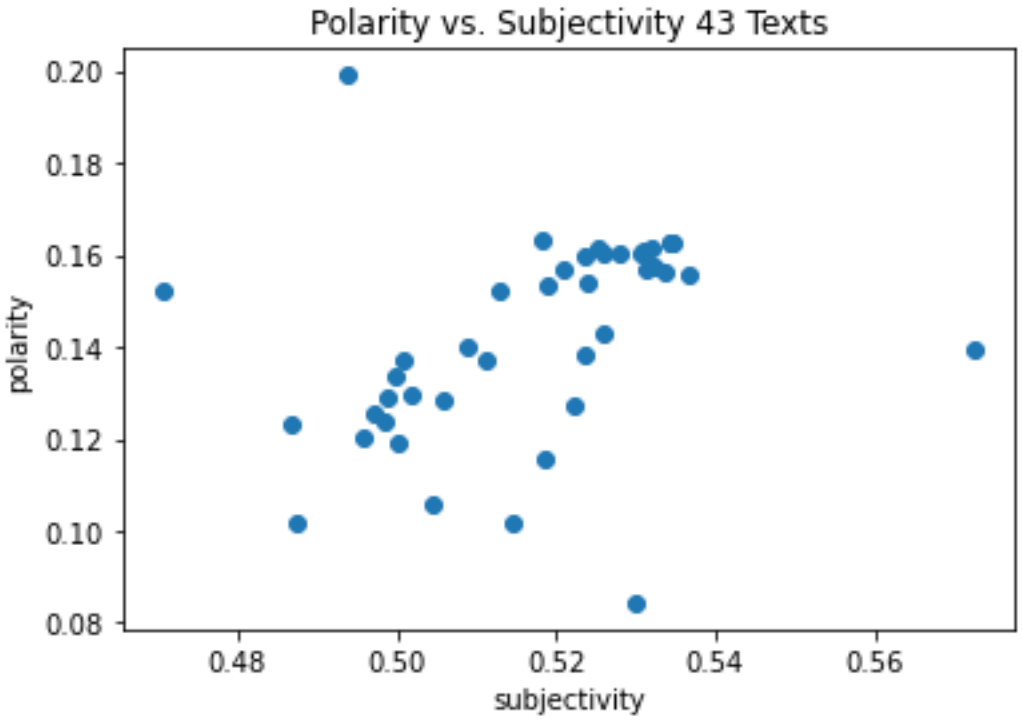
After re-running the program including '{}' in the cleaning regex, the CatholicPublicDomainVersion's cosines compared better with the other texts in this analysis (> 0.988). The color map in Figure 6 shows this in the change on line 6.

Figure 6



Polarity was exclusively positive, ranging from 0.084, BibleInBasicEnglish_1965, to 0.199 for the GenevaBible_1599. Subjectivity was also constrained into a tight range between 0.471, AmericanStandardBible_1881, and 0.573 with Darby_1889. If we remove Darby_1889 with its formatting errors, the highest value is 0.537 of the AmericanBibleUnionBible_1866. Plotting polarity against subjectivity (Figure 7) shows roughly two groupings but it may not be significant. There appear to be a mix of older and newer texts in both groups with no obvious explanation.

Figure 7



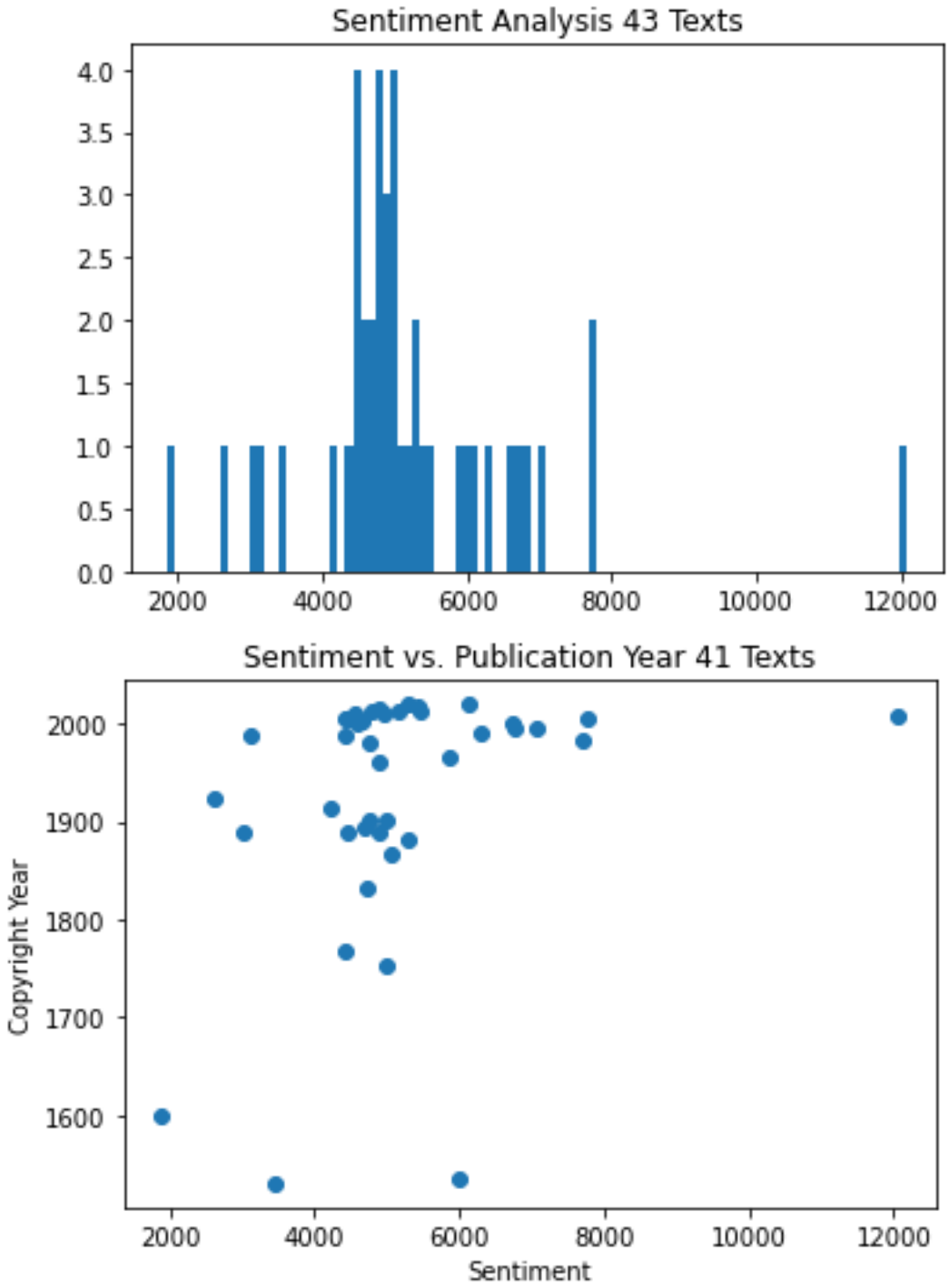
Sentiment analysis had a range of 1,872 to 12,063 where lowest scoring and highest scoring are shown in Table 3. These outliers do suggest some influence of publishing date on sentiment. Poorly formatted Darby_1889 turns up near the bottom of sentiment as does TheEverydayBible_1924, with its significant abridgement of the new testament. Highest scoring was TranslationforTranslators_2008, probably not a surprise with its in-line footnoting about the text that frequently adds to what is being said, amplifying the context. All of the texts with the highest scores are more recent (last 40 years).

Table 3 – Lowest and Highest Sentiment Scores

Score	Text	Score	Text
1872	GenevaBible_1599	6714	TheMessageBible_2002
2612	TheEverydayBible_1924	6778	GodsWordtotheNation_1995
3013	Darby_1889	7051	NewLivingTranslation_1996
3106	RevisedEnglishVersion_1989	7701	TheAmplifiedBible_1984
3446	Tyndale_1530	7781	EasyToReadVersion_2006
		12063	TranslationforTranslators_2008

Figure 8 shows the sentiment scores of 43 texts where we can see that most of the texts scored between 4,000 and 6,000. Plotting sentiment score against publication date, we can see that there are three groups. One is between 4,000 and 6,000 published about 100 years ago. The second is also between 4,000 and 6,000 but published more recently. The third is published recently and between 6,000 and 8,000.

Figures 8 & 9



Conclusions

This project was very enlightening for me. It is easy to pick up any bible today and read with understanding. This is thanks to people who translated, then transcribed and edited editions of the bible for ease and acceptance. An evangelist might say, “it is better to have a fair translation that many will read than a perfect translation that no one would read.”

Today’s bibles are built on centuries of translation, transcription, and editing by countless individuals each with perspectives, varying levels of competence, interests, and fallibilities. The overarching question for me is, how well does a document codifying behavior and relationships retain its original meaning over long time spans? This project did show some changes in sentiment for some translations in the most recent forty years. Does this change the message? Is this more an artifact of the tools used to measure sentiment? This work did not get answers to these questions, which will take additional study.

Things I would Do If Time Permitted

Koine Greek to English machine translation. I know nothing of Greek that isn’t part of math of physics. In researching this project, I learned that I am not the only one interested in this and some resources exist online, such as https://www.freelang.net/o_nline/koine_greek.php, which has a downloadable dictionary of 5,408 words. An extremely important reference for this project is https://www.codexsinaiticus.org/en/ which is the oldest copy of the bible in Greek, hand written 1,600 years ago. The more I see of this, the more I think it would be worth doing. When I started this project, the first Greek to English text I saw was Matthew’s genealogy of Jesus. Every “begat” after Abraham was negated in the Greek even though it was translated without the negation. I speculated that this was the fault of a typesetter who didn’t know better. Examining the Codex Sinaiticus, I found

that the text was rendered as written, 1,600 years earlier! And the English translation is even more embellished to include the names of wives.

A real comparison of translations should be measured against the original Greek. I would not expect this to be pretty or entertaining in a literary sense. I would expect translations to reasonably follow the key elements and relate the meanings of the verses. The availability of such an unpolished translation would enable a scoring of translations for readability versus deviation.

A “genealogic” tree of bible translations. Looking at the data here, I am wishing I could see the source material for each edition to see what had been changed then track and evaluate those changes. The identified changes could become markers of social conditions prevalent at the time. An example is the existence of the King James Bible which was translated under the direction of King James versus the Geneva Bible. It seems that there is also a path dependency at work here. The KJV became one of the most widely used bibles in the world. Is it really a superior version, or was it just the British empire’s success that made it popular?

Choice ratings: Not directly translation, but running through each bible version to compare word and phrase choices by assigning a comparative meaning and charting the differences in choices made by translators and editors to produce each new volume. These choices could be rated in several ways.

A corpus of unique bibles could be polled by the software to catalog words used at a particular verse. A value representing its frequency of use in those bibles at that verse would be assigned. Does the sum of these values put the particular bible closer or further from other bibles? The word choices could also be assigned to translators. Translators will have favorite words, one might expect them to be used preferentially where they have a choice.

Authorship: Is each book genuinely written by a different person? How is authorship analysis affected by translation? Can the original author still be “seen” in the translation? Recent work suggests that machine translation may still permit differentiating authorship.(Ouamour 2023) How many times can a text be transcribed, translated and edited before the evidence of the original author is erased?

Tool design for the task of analyzing the older texts: The sentiment tools used here may be too seated in modern culture to appreciate sentiment in older texts. It would be interesting to try to build tools that would account for the age of the text being analyzed. This would require a corpus of documents contemporary to the text under study and an understanding of the documents’ intent in context; to entertain, threaten, encourage, etc. The goal would be to make an algorithm that could recognize the intent in terms of attitudes contemporary to the text. Something uplifting from 2000 years ago might be revolting today. It would not be fair to score it as a negative.

Idioms and figures of speech detection: There are passages that don’t seem to make sense, or, describe odd behavior that are actually idioms that might not be very widely known. I would like to use text processing to catalog and identify possible new ones. (Tedeschi 2022, Zeng 2021) Many idioms have been cataloged (Lamsa 1985) , but I am not interested in things that have become idioms today. I want to look for things there were idioms long ago that have no meaning today. This might involve searching other documents to understand the usage.

Prepared .txt files are available at:

https://github.com/bertoli517/DATA532

Texts were sourced from these websites:

Text	Source
AmericanBibleUnionBible_1866	https://archive.org/details/newtestamentofou1867amer/page/n1/mode/2up
AmericanStandard_1901	https://archive.org/details/jstor-3137096/mode/2up
AmericanStandardBible_1881	https://biblesfree.org/Bibles/american-standard-bible.pdf
AmericanStandardVersion_1901	https://ebible.org/pdf/eng-asv/eng-asv_all.pdf
AuthorizedVersion_1769	https://archive.org/details/kjv-1769-oxford-edition-full-bible
BereanStandardBible_2016	https://ebible.org/pdf/engbsb/engbsb_all.pdf
BibleInBasicEnglish_1965	https://ebible.org/pdf/engBBE/engBBE_all.pdf
CatholicPublicDomainVersion_2010	https://openbible.com/pdfs/cpdv.pdf
ContemporaryEnglishVersionUK_1995	http://download.sabda.org/mobile/pdf/CEVUK.pdf
Darby_1889	https://archive.org/details/the-holy-bible-darby-translation
Darby_1890	https://archive.org/details/ENGDBY_DBS_HS
Darby_1961	https://bibles.dbs.org/ENGDBY/pdf/ENGDBY.pdf
Douay-Rheims_1752	https://www.gutenberg.org/cache/epub/1582/pg1582.txt
EasyToReadVersion_2006	https://ebible.org/pdf/engerv/engerv_all.pdf
EnglishStandardVersion_2001	https://biblesfree.org/Bibles/english-standard-version-Bible.pdf
FreeBibleVersion_XXXX	https://ebible.org/pdf/engfbv/engfbv_all.pdf
GenevaBible_1599	https://ebible.org/pdf/enggnv/enggnv_all.pdf
GodsWordtotheNation_1995	http://download.sabda.org/mobile/pdf/GWV.pdf
HolmanChristianStandardBible_2003	http://download.sabda.org/mobile/pdf/HCSB.pdf
KingJamesCambridgeParagraph_2005	https://ebible.org/pdf/engkjvcpb/engkjvcpb_nt.pdf
KingJamesVersion_2006	https://ebible.org/pdf/eng-kjv2006/eng-kjv2006_all.pdf
KJVBible_1989	https://www.gutenberg.org/cache/epub/10/pg10.txt
LexhamEnglishBible_2010	http://download.sabda.org/mobile/pdf/LEB.pdf
LiteralStandardVersion_2020	https://ebible.org/pdf/englsv/englsv_all.pdf
NewEnglishTranslation_2016	https://ebible.org/pdf/engnet/engnet_nt.pdf
NewInternationalVersion_XXXX	https://biblesfree.org/Bibles/NIV-Bible.pdf
NewKingJamesVersion_1982	http://download.sabda.org/mobile/pdf/NKJV.pdf
NewLivingTranslation_1996	https://biblesfree.org/Bibles/New-Living-Translation-NLT.pdf
NewTestamentMoffat_1913	https://archive.org/details/newtestamentnewt0000unse_d4p3
NoahWebsterBible_1833	https://ebible.org/pdf/engwebster/engwebster_all.pdf
OpenEnglishBible_2020	https://ebible.org/pdf/engoebcw/engoebcw_all.pdf
OpenEnglishBible_US_2020	https://ebible.org/pdf/engoebus/engoebus_all.pdf
PeoplesNewTestament_1889	https://archive.org/details/peoplesnewtestam0000unse
RevisedEnglishVersion_1989	http://download.sabda.org/mobile/pdf/REB.pdf
RevisedVersionwithApocrypha_1895	https://ebible.org/pdf/eng-rv/eng-rv_all.pdf
TheAmplifiedBible_1984	https://biblesfree.org/Bibles/Amplified-Bible.pdf
TheEnglishMajorityTextVersion_2014	https://ebible.org/pdf/engemtv/engemtv_all.pdf

Text	Source
TheEverydayBible_1924	https://archive.org/details/everydaybible0000unse
TheGoodNewsBible_1992	https://biblesfree.org/Bibles/good-news-bible.pdf
TheMessageBible_2002	http://download.sabda.org/mobile/pdf/MSG.pdf
TranslationforTranslators_2008	https://ebible.org/pdf/eng-t4t/eng-t4t_all.pdf
Tyndale_1530	https://www.gutenberg.org/ebooks/10553
TyndaleNewTestament_1534	https://ebible.org/pdf/engtnt/engtnt_all.pdf
UnlockedLiteralBible_2017	https://ebible.org/pdf/engULB/engULB_all.pdf
WorldEnglishBible_2014	https://ebible.org/pdf/engwebp/engwebp_all.pdf
WorldEnglishBibleBritish_2014	https://ebible.org/pdf/engwebpb/engwebpb_all.pdf
WorldEnglishBibleCatholic_2014	https://ebible.org/pdf/eng-web-c/eng-web-c_all.pdf
YoungsLiteralTranslation_2013	https://ebible.org/pdf/engylt/engylt_all.pdf

References

1. Bilisoly, Roger, and Wiley InterScience (Online service). *Practical Text Mining with Perl*. Wiley, 2008.

2. Ouamour, S., Sayoud, H. (2023). Effect of Machine Translation on Authorship Attribution. In: Bhateja, V., Yang, XS., Ferreira, M.C., Sengar, S.S., Travieso-Gonzalez, C.M. (eds) Evolution in Computational Intelligence. FICTA 2023. Smart Innovation, Systems and Technologies, vol 370. Springer, Singapore. https://doi.org/10.1007/978-981-99-6702-5_5

3. Tedeschi, S., Martelli, F., and Navigli, R. ID10M: Idiom Identification in 10 Languages. In Findings of the Association for Computational Linguistics: NAACL 2022, pages 2715–2726, Seattle, United States. Association for Computational Linguistics. https://aclanthology.org/2022.findings-naacl.208.pdf

4. Zeng, Z., Bhat, S. Idiomatic Expression Identification using Semantic Compatibility. Transactions of the Association for Computational Linguistics 2021; 9 1546–1562.https://doi.org/10.1162/tacl_a_00442

5. Lamsa, George M., and George M. Lamsa. Idioms in the Bible Explained ; and, A Key to the Original Gospel. Harper & Row, 1985.