

Chapter 5

Similarity data: Bible translations



Abstract

Main topics. Many Bible translations from different times are based on the same *Masoretic text*. How alike are these translations, and how many differences have emerged over time? **Data.** The data for this study were collected by Zachary Bleemer from British, American, and German translations. First, Bible verses were matched, and subsequently parallel words were counted. **Research questions.** How can the similarities between texts based on the word counts be analysed and graphed, so that conclusions can be drawn with respect to influence between the various translations? **Statistical techniques.** Multidimensional scaling, cluster analysis.

Keywords Bible translations · Hebrew old testament · *Masoretic text* · *Tanakh* · King James Bible · Similarity coefficients · Multidimensional scaling · Cluster analysis · Centroid-linkage clustering

5.1 Background

An intriguing question¹ regarding language development is whether words have the same meaning and connotations at different points in time. Did the imagery used by Shakespeare in his plays and sonnets evoke the same sentiments and associations in his contemporaries as they do in readers nowadays? If we assume that cultural expressions such as translations are locked in the context of their own time; is there a way in which such an issue can be researched?

Such questions may be even more relevant when people interpret Bible texts and discuss the relevance of the texts in their own times. This point is reinforced when we consider that the Bible is known by virtually everyone in their own vernacular, so that its interpretation is often time and language dependent. The same thing can

¹This chapter was co-authored by Zachary Bleemer, Department of Economics, University of California, Berkeley, <http://zacharybleemer.com/>.

be seen in many Western paintings of religious events which took place in Jesus' time; these are portrayed as if situated in the painter's own time and surroundings, which is especially evident in the clothing of the depicted persons and the depicted landscapes.

Translations of the Hebrew Bible, the *Tanakh*,² are an interesting opportunity to get a view on what according to the translators was being described and expressed by the text. As over time the same basic Hebrew (*Masoretic*) text has been translated several times into several languages, shifts in interpretations can be compared and evaluated through these translations. An advantage of the Old Testament is its length and narrative variation: most of the interesting concepts appear in a wide variety of contexts throughout the text.

The study presented here is part of a project by Zachary Bleemer in which aesthetic concepts such as beauty and handsomeness, employed by translators from different countries or centuries, are assessed. Bleemer (2016, p. 3) discusses the idea that over time attributes of objects may be perceived differently, and thus translated into different adjectives. He then postulates that such differences are a reflection of the aesthetic vocabulary available, and perhaps of the 'latent aesthetic categories' that beautiful objects can belong to, according to those translators. This is generally not an idiosyncrasy of a single translator because translations were generally done by teams of scholars. Moreover, Bibles in the vernacular were intended for the general population and thus had to be translated in the then current common variant of the language. Therefore, it seems safe to assume that the translations represent if not the common language of the period, then at least the language which would be commonly understood at the time.

5.2 Research questions: Similarity of translations

In this chapter we will concentrate on one aspect of Bleemer's study: the extent of the variability of the translations, and the question whether period effects are reflected in the available translations. These queries refer to the more general questions posed in the original Bleemer (2016) study, which focusses on one particular Hebrew word (*yaphah*, to be beautiful), the way it is handled in different translations, and what this has to tell us about social and cultural differences between people over time.

²The *Tanakh* [...] is the canonical collection of Jewish texts, which is also the textual source for the Christian Old Testament. These texts are composed mainly in Biblical Hebrew, with some passages in Biblical Aramaic[...]. The traditional Hebrew text is known as the *Masoretic* text (www.DBpedia.org/page/Tanakh).

5.3 Data: English and German Bible translations

Acquiring the data for proper comparisons between the translations proved an arduous task, and impossible without proper digital support. Bleemer first collected 14 available Bible translations and divided them into verses, on the basis of the versification in the *King James Bible*—*KJV*. This is an intricate process, the full details of which are given in Bleemer (2016, Section 6.3.1).

The translations can be grouped into four types; see Bleemer (2016) for full details.

- *KJV, GNV*. Two Early Modern Protestant translations into English from the 16th and 17th centuries. These Bibles were translated by large teams of British Protestant scholars from the original Hebrew text rather than from prior Latin and Greek translations, as were some earlier renderings. The *King James Version (KJV)* was partially a revision of the *Bishop's Bible*, which itself was a partial revision of the *Geneva Bible (GNV)*, so that vocabulary decisions are not wholly independent.
- *ASV, NKJV, JPS*. Three 20th-century revision translations derived from the *KJV* (with contemporary emendations arising from comparison with the Hebrew text). The *American Standard Version (ASV)* and the *New King James Version (NKJV)* were produced by Protestant organisations, whereas the *Jewish Publication Society's version (JPS)* was published by a Jewish organisation.
- *NWT, NAB, NIV, NET, GNT*. Five 20th- and 21st-century American English *sui generis* translations, independently produced by unconnected organisations. Specifically, these were the *New World Translation of the Holy Scriptures (NWT)* by the Jehovah's Witnesses, the Catholic *New American Bible Revised Edition (NAB, also abbreviated as NABRE)*, the Protestant *New International Version (NIV)*, the Protestant *New English Translation (NET)*, and the interdenominational *Good News Bible (GNT)*.
- *Luth, EU, HFA, GNB*. Four German translations: the 16th-century Protestant *Luther translation (Luth, also Luther)* and three 20th-century translations: the Catholic *Einheitsübersetzung (EU)*, the Protestant *Hoffnung für Alle (HFA)*, and the *Gute Nachricht Bibel (GNB)*, jointly produced by Evangelical Protestants and Catholics.

Each Bible verse in a translation was compared with the same verse in the other translations via Dice's *coincidence coefficient*, which measures the *similarity* between two sets as the fraction of elements that are present in both sets. Details can be found in the original publication (Dice, 1945). What is important here is that the measure of similarity was defined in such a way that the higher the number the more similar the verses, with as its highest value 1 (the translations are identical) and lowest value 0 (the two translations do not share any words). It should be remarked that certain types of words are excluded from the comparisons, such as function words or articles and prepositions. This in contrast with authorship studies, where function words are the crucial information in the study; see Chapter 6 on the authorship of the Pauline Epistles, and Chapter 13 on the authorship of *The Royal Book of Oz*.

The coincidence coefficients for the verses of two translations were combined to create the average fraction of identical words. They are reproduced in Table 5.1,

where the translations are listed in alphabetical order, and in Table 5.2, in which that table has been reformatted for better visual interpretation. These tables are similarity matrices with only the lower triangle and diagonal shown, because the upper triangle is a mirror of the lower one and contains no new information. A simple perusal of Table 5.1 is not very enlightening, and the alphabetical order of presentation does not give much insight into the relationships between the translations. The main reason is, of course, that the alphabet is unrelated to the similarity between translations. It is a general phenomenon that tables are only insightful if the orders of rows and columns correspond to a relevant aspect of the objects presented; see Section 3.15.1, p. 93 for some rules to improve the presentation of a table, where Table 5.2 is also presented in a slimmed-down version (Table 3.6, p. 97).

Table 5.1 Similarities of Bible translations based on the *Average fraction of identical words*: Unsorted

	Translations													
	ASV	EU	GNT	GNB	GNV	HFA	JPS	KJV	Luth	NAB	NET	NKJV	NIV	NWT
ASV	1.00													
EU	0.05	1.00												
GNT	0.35	0.04	1.00											
GNB	0.04	0.39	0.04	1.00										
GNV	0.77	0.05	0.36	0.04	1.00									
HFA	0.04	0.36	0.04	0.43	0.04	1.00								
JPS	0.89	0.05	0.36	0.04	0.77	0.04	1.00							
KJV	0.87	0.05	0.36	0.04	0.84	0.04	0.86	1.00						
Luth	0.03	0.22	0.03	0.15	0.04	0.15	0.03	0.04	1.00					
NAB	0.53	0.04	0.43	0.04	0.53	0.04	0.55	0.53	0.03	1.00				
NET	0.46	0.04	0.45	0.04	0.46	0.04	0.48	0.46	0.03	0.56	1.00			
NKJV	0.73	0.04	0.42	0.04	0.70	0.04	0.74	0.76	0.03	0.62	0.56	1.00		
NIV	0.54	0.04	0.49	0.04	0.52	0.04	0.54	0.53	0.04	0.59	0.58	0.62	1.00	
NWT	0.53	0.04	0.37	0.04	0.47	0.04	0.51	0.49	0.03	0.51	0.48	0.53	0.52	1.00

To get a better insight into the patterns, Table 5.1 has been reformatted on the basis of the size of the similarity coefficients, see Table 5.2. Coefficients greater than .50 are in bold.

Several aspects now stand out much more clearly, such as the fact that there is no real similarity between the English and German translations, which is understandable as they are in different languages. In fact, we may wonder why the coefficients are not zero. The occurrence of names and places is the most likely source of such similarities.

Some other conclusions that may be drawn from the table are that the UK and US King James versions are rather more similar to each other than the US translations are to each other. As they intended, the translators of the *Good News Bible (GNT)* have succeeded in making their translation different from any other English translation.

Table 5.2 Similarities of Bible translations based on the *Average fraction of identical words*: Sorted and formatted

Trans	Year	UK		USA - KJV			USA - <i>sui generis</i>					Germany			
		KJV	GNV	ASV	JPS	NKJV	NAB	NIV	NWT	NET	GNT	EU	HFA	GNB	Luth
KJV	1560	1.00													
GNV	1611	0.84 1.00													
ASV	1900	0.87 0.77		1.00											
JPS	1917	0.86 0.77		0.89 1.00											
NKJV	1982	0.76 0.70		0.73 0.74 1.00											
NAB	2011	0.53 0.53		0.53 0.55 0.62			1.00								
NIV	1996	0.53 0.52		0.54 0.54 0.62			0.59 1.00								
NWT	1976	0.49	0.47	0.53 0.51 0.53			0.51 0.52 1.00								
NET	2005	0.46	0.46	0.46	0.48	0.56	0.58 0.48 1.00								
GNT	1978-	0.36	0.36	0.35	0.36	0.42	0.43	0.49	0.37	0.45	1.00				
EU	1980	0.05	0.05	0.05	0.05	0.04	0.04	0.04	0.04	0.04	1.00				
HFA	1996	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.36 1.00				
GNB	1982	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.04	0.39	0.43	1.00		
Luth	1534	0.04	0.04	0.03	0.03	0.03	0.03	0.04	0.03	0.03	0.03	0.22	0.15	1.00	

Denominations: Anglican: KJV; Protestant: GNV, ASV, NIV, GNT, HFA, Luth; Jewish: JPS; Jehova: NWT; Catholic: NAB, EU; Evangelical Protestant: NKJV, NET, GNB.

Furthermore, the German translations are not very similar among themselves. It may be that the greater word-level variation in German is the cause of this, as the language contains compound nouns and declensions that make overlap less likely, even after attempts to trace to their stems to make them more comparable.

These comparative observations could not have been derived from the alphabetically sorted Table 5.1 without hard graft.

5.4 Analysis methods

In this case study we wanted to compare word occurrences in various Bible translations without having a particular response variable in mind. Thus we are dealing here with an internal structure design (see Section 3.7).

The analysis of similarity matrices has a long tradition, and the main working horses are various types of multidimensional scaling (MDS) and cluster analysis. Here we will use only one method for each type of analysis, but there are many variants of both techniques, often specially geared towards specific analytic goals (see Section 3.20, p. 81).

5.4.1 *Characteristics of multidimensional scaling and cluster analysis*

Similarity and dissimilarity matrices are eminently suitable for both multidimensional scaling and cluster analysis. These techniques supplement each other, as the former produces a spatial representation and the latter a group representation. Thus, we can first make an appropriate multidimensional scaling graph and then draw contours around the points that the cluster analysis has indicated as belonging to the same group. The two techniques do not necessarily use exactly same information, because the dimensions which account for the maximum variance in the multidimensional scaling solution need not be those that provide the best cluster separation. Fortunately they often do. This makes combined graphing an insightful activity (see also Section 3.9.3).

5.4.2 *Multidimensional scaling*

Multidimensional scaling aims at a graphical representation of similarities, which have been first transformed into dissimilarities. They are subsequently portrayed as Euclidean (straight-line) distances in a preferably two-dimensional graph. Higher dimensional graphs are only used if a proper representation in two dimensions is not adequate. The specific method MDS employed here is implemented in SPSS (Heiser, 1988; Busing, Commandeur, & Heiser, 1997).

Note that very high similarity means a very low dissimilarity and a very short distance. Given the high similarities between the *King James Version* and the first four translations in Table 5.2, we expect that in the MDS graph they will be very close to each other. Furthermore, the German translations will be clearly separated from the English ones as similarities among the two languages are uniformly low. Moreover, it is unlikely that the German translations form a tight group, because the similarities among themselves are not very high, certainly not as high as among the English-language ones.

5.4.3 *Cluster analysis*

Cluster analysis attempts to group the items (here, translations) based on their similarities. Because cluster analysis uses the similarity information differently from multidimensional scaling, the produced grouping may not be clearly visible in the low-dimensional space derived by the multidimensional scaling; cluster analysis uses information contained in all dimensions. Nevertheless, the grouping found can often enhance the scaling results, as is the case in the present study. Here we used *Cen-*

triod linkage cluster analysis, which is fairly generally applicable. (see Section 3.9.3, p. 81), and Everitt et al. (2011), and Šulc and Řezanková (2019)).

5.5 Bible translations: Statistical analysis

5.5.1 Multidimensional scaling

Figure 5.1 shows the outcome of the multidimensional scaling and clustering procedures. The two-dimensional graph shows the patterns which we discussed after rearranging the table of similarities (Table 5.2). This is not surprising, because we actually rearranged the table using the multidimensional scaling solution. The more surprising effects are the stretched-out positions of the American and German translations and those of the German ones. The ellipses indicate the subgroups produced by the cluster procedure (see below). Note by the way, that the coordinate axes (dimensions) themselves need not be interpreted, but only the positions of the translations with respect to each other.

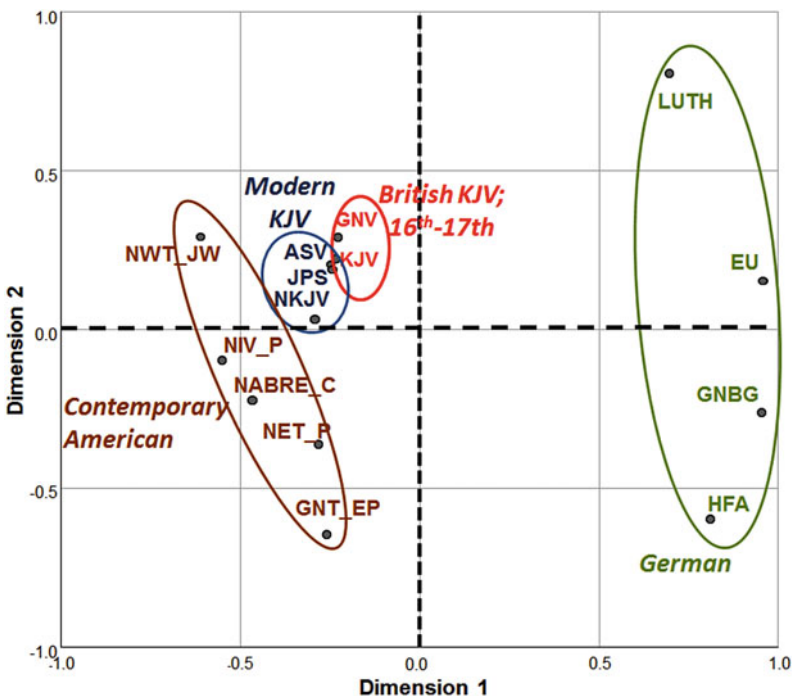


Fig. 5.1 Multidimensional scaling and cluster analysis of the Bible translations. The space based on the similarities in Table 5.2; clusters based on the visual clustering in the table.

5.5.2 Cluster analysis

Centroid-linkage cluster analyses were carried out and the six-cluster solution was chosen to be presented here. The result for the six-cluster solution is drawn in the space constructed by the multidimensional scaling of the similarity matrix. Figure 5.2 shows that the German translations are rather different, both from the English ones and among themselves, as could be expected from their relatively low similarities. The *Good News Bible (GNT)* is clearly distinct from the other English translation, in accordance with the intentions of the translators. The translations derived from the King James Version (*KJV*) group together, as do the four American translations (*NIV*, *NABRE*, *NET*, and the Jehovah's Witness *NWT*).

Continuing the analysis with less clusters via merging the already found clusters, the *EU* translation join the German group, and almost at the same time the British and America translations merge into one group. Finally, *GWT* joins the English-language translations and Luther the German ones. Then it takes a very long time before all translations amalgamate into one single cluster. The moment of translations or translations groups merging is regulated by similarities, so that the progression of 'time' is actually an expression of decreasing similarity.

5.6 Other approaches to analysing similarities

Similarity and dissimilarity matrices appear in many disciplines, including the humanities and the social and behavioural sciences, and many of these have been analysed with similar methods as have been used in this chapter.

There is a wealth of applied, technical, and statistical books and papers, and computer programs on multidimensional scaling and cluster analysis. A reasonable starting point to study multidimensional scaling is Borg et al. (2013), and for cluster analysis Everitt et al. (2011) may be fruitfully consulted. Several general introductory multivariate statistics books have chapters on multidimensional scaling and cluster analysis (e.g. Field, 2017; Tabachnick & Fidell, 2013).

5.7 Content summary

What probably makes the character of the Bible translation dataset unique is that the translations were all made from the same original text, the *Masoretic* text of the Hebrew Bible (or *Tanakh*) dating back in its final form to roughly the 9th century CE.³ This common starting point, plus the digitisation of the text and computer-assisted text analyses made it possible to compute the similarities used in the analysis.

³Wikipedia: https://en.wikipedia.org/wiki/Masoretic_Text.

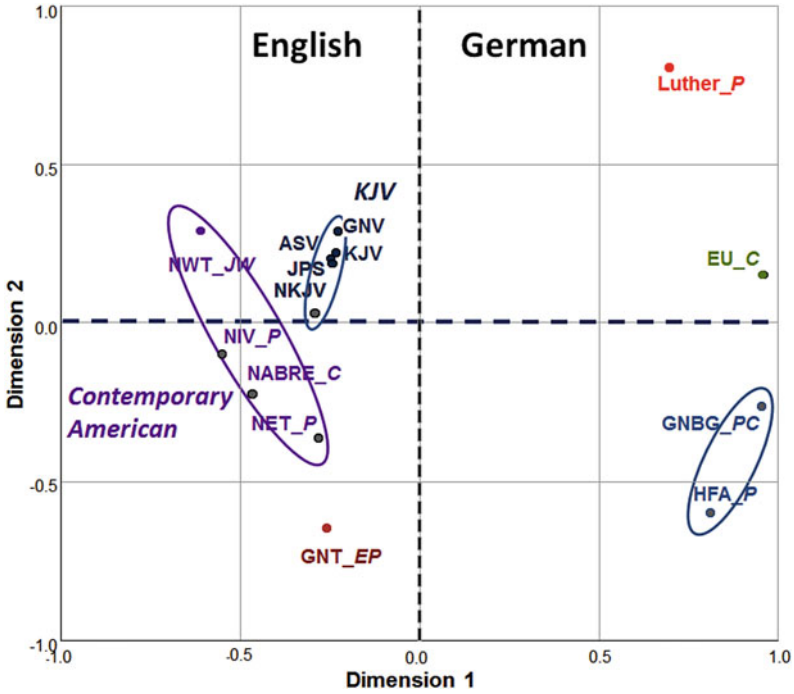


Fig. 5.2 Multidimensional scaling and centroid clustering of Bible translations. Space enhanced by a six-cluster solution; three clusters consist of a single translation. P = Protestant, C = Catholic, JW = Jehova Witness.

The differences between the translations come to the fore by applying standard multivariate analysis methods to the similarities. Some of the results are obvious, such as the lack of similarities between the English and German translations. Interesting is, however, that the German translations among themselves show less similarity than the English ones. Moreover, some other results, such as the clear differences between modern translations made by different Christian denominations, or with different groups of readers in mind (for example, the Jehovah Witness *New World Translation of the Holy Scriptures* versus the *Protestant Good News Bible*), were less obvious beforehand.

The analyses presented here are only the start of a project which examines changes in aesthetic categories over time and between denominational groups, as can be derived from different translations of the same text (see Background for this study, p. 133). Readers should turn to the work by Bleemer (2016) for such evaluations and interpretations.