# Counting the Jeremiahs: Machine Learning and the Jeremiah Narratives

NICHOLAS J. CAMPBELL (THE SOUTHERN BAPTIST THEOLOGICAL SEMINARY)

*Abstract*

*Scholars have long debated the redactional history of the prose sections of Jeremiah (chapters 26–45) but no consensus has been reached on the number of redactional layers in the text, the verses that comprise these layers or their sources. This study used a machine learning method to organise the chapters into sections based upon authorial word choices. The method used pairs of synonyms in a hierarchical clustering algorithm in the statistical program R. The goal of the study was two-fold. First, the division of the text by computerised model was used to analyse the divisions made by three other more traditional critical methods. Second, the validity of the method used in this study and previous synonym-based studies was analysed and critiqued. The conclusion is that this type of analysis can validate findings from other methods but some of the inherent biases and linguistic ambiguities make it dubious as a primary method of investigation for the Hebrew Bible.*

KEYWORDS: Computerised learning, Jeremiah, Synonyms, Source criticism, Redaction criticism

## A     INTRODUCTION

A number of scholars have attempted to identify sources and redactional layers in Jer 26–45. At the same time, mathematical linguistic studies have become increasingly popular as powerful statistical software has become widely available. This study will combine traditional redactional questions with a modern algorithmic clustering method. The goal is to analyse the utility of statistical software in general and, specifically, synonym sets, for the study of biblical texts. The results of a computer model of synonym clustering for Jer 26–45 will be compared with three prominent redaction theories to show the weaknesses and strengths of the different models. At the end, an assessment of this computer modelling technique will be provided to define the potential of this type of work within traditional biblical scholarship.

## B        SCHOLARLY DIVISIONS OF JEREMIAH

Though there are numerous proposed redactional divisions of the text of Jer 26–45, three of the most prominent ones will be discussed here and they, along with a number of other prominent views, will be presented in Table 1.[1] The three views treated in this study were selected for two reasons. First, each scholar uses a method that is widely accepted as valid even if the results are not always agreed upon (Deuteronomist language and Classical versus Late Hebrew). Second, each of these authors treat features in all the chapters in Jer 26–45 whereas many other scholars only discuss some of the chapters.

Thomas Römer describes a three-stage redaction of the book of Jeremiah. The first Deuteronomistic (dtr) redaction is chapters 7–35. This redaction, according to Römer, was done around the time of Amos and Hosea. Its vocabulary, composition and theology are similar to the Deuteronomistic Historian (DH).[2] The early edition of the text then travelled to Babylon and was combined with chapters 2–6 and 37–44 to create a second edition with a frame provided by chapters 1 and 45.[3] This second section strongly criticises the Egyptian diaspora, unlike chapters 7–35, and is dated to the end of the exilic period or beginning of the Persian period.[4] The third piece of the redaction history is chapter 36 which is the legitimising linkage of the new portion (37–45) with the earlier edition (7–35).[5] According to Römer, the linguistic evidence of this redactional structure is that 55% of Stulman's dtr diction not attested in the DH is found in chapters 1–6 and 36–45.[6]

Moving from Römer's redaction history, the second view is the dtr system developed by Louis Stulman. For Stulman, the objective was not to develop a general theory on the structure of Jeremiah but rather to catalogue and classify

---

[1]    There is a debate on the precise definitions of redactor and author. Redactor is used broadly here to refer to anyone making changes to the text whether it is additions, arrangement or some other reworking of a pre-existing text. For a survey of the debate, see John Van Seters, "Author or Redactor?," *JHS* 7 (2007): 1–23.

[2]    Thomas Römer, "How Did Jeremiah Become a Convert to Deuteronomistic Ideology?" in *Those Elusive Deuteronomists: The Phenomenon of Pan-Deuteronomism* (ed. Linda S. Schearing and Steven L. McKenzie; JSOTSup 268; Sheffield: Sheffield Academic Press, 1999), 197.

[3]    Chapters 2–6 and 37–44 were supposedly already circulating within the Babylonian community before they were combined with chapters 7–35. Chapters 1 and 45, however, were created out of whole cloth by the redactors to frame the two pieces that had been combined. Römer, "How Did Jeremiah Become a Convert?," 198.

[4]    Ibid.

[5]    This chapter, like chapters 1 and 45, was also created specifically for the book by the redactors. Ibid.

[6]    Ibid.

dtr linguistic data in the prose sermons.[7] The classification system is four-fold. Dtr I comprises words and phrases attested more than once in DH, dtr II contains words and phrases which occur only once in DH, dtr III has diction found in Deuteronomy proper but not in DH and dtr IV has words and phrases not attested in Deuteronomy or DH but are said to resemble their diction.[8]

The dtr data presented by Stulman is detailed in the first two columns under "Stulman" in Table 1.[9] The analysis is not as cohesive as Römer's for two primary reasons. First, Stulman focuses strictly on the prose speeches. Second, the focus is on dtr word preference, not on tracing the broad redactional or geographic history. This system will be dealt with in more detail in comparison with the computerised model. However, it is interesting to note that, contra Römer, the dtr IV diction is found in significant quantities in chapters 27, 29 and 35. This makes the redactional distinction between 26–35 and 37–44 difficult to identify, at least in terms of dtr linguistic traits.

The last study that will be analysed is Aaron Hornkohl's division of Classical Biblical Hebrew and Late Biblical Hebrew. Hornkohl uses words, phrases and spellings that are commonly believed to be indicators of either early or late Biblical Hebrew to date the book of Jeremiah.[10] The argument presented by Hornkohl is that much of Jeremiah is written in Classical Biblical Hebrew. This indicates that the text is probably from the sixth century when the style was transitioning. If the book were later than this, more Late Biblical Hebrew would be expected (i.e. Ezra, Nehemiah or Daniel).[11] Like Stulman, Hornkohl's division does not neatly classify large sections of Jeremiah or indicate a redactional history. Two of the most prominent lexical features presented by Hornkohl have been added to Table 1 and will be discussed in the results section.

---

[7]     Louis Stulman, *The Prose Sermons of the Book of Jeremiah: A Redescription of the Correspondences with Deuteronomistic Literature in the Light of Recent Text-Critical Research* (SBLDS 83; Atlanta: Scholars Press, 1986), 3.

[8]     Stulman also divides the text of Jeremiah between phrases found in both the Septuagint and the Masoretic Text and phrases found in only one of them. However, the concern here is only with the Masoretic Text as it currently stands. Therefore, the conclusions drawn about the relationship between the two versions will not be discussed. Stulman, *The Prose Sermons*, 32.

[9]     Two columns are associated with Stulman and as they present data from two different tables, each column should be read vertically. For example, in the left column, chapters 27, 29, 35, 39 and 45 are all similar because category IV diction comprises more than half the words. In the second column, chapters 27, 29, 33, 39 and 45 are related because less than 5% of their diction comes from categories I and II.

[10]     Aaron D. Hornkohl, *Ancient Hebrew Periodization and the Language of the Book of Jeremiah: The Case for a Sixth-Century Date of Composition* (Studies in Semitic Languages and Linguistics 74; Boston: Brill Nijhoff, 2014), 40.

[11]     Hornkohl, *Ancient Hebrew*, 52–60.

## C    BIBLICAL TEXTS AND MACHINE LEARNING

An often-cited work from computerised text criticism in biblical studies is D.L. Mealand's *Correspondence Analysis of Luke.* Mealand divided the Gospel into 500-word segments and analysed the text by function words, parts of speech, and letter variables. The parts of speech were analysed by frequency and synonymity. With the same method as this study, words were separated according to the part of speech (e.g. verb versus noun) and then paired with a synonymous term.[12] The frequencies of these terms were used to separate out different authors in the text. The results of each of the analyses show a distinct style in the infancy narrative but less pronounced distinctions between the Q material, Mark material and L material.[13]

Another recent study using sets of synonyms (synsets) was performed on the first four books of the Bible by Dershowitz et al. in 2015.[14] The goal was to separate out the Priestly source from the rest of the text.[15] The authors paired Hebrew synonyms based on Strong's Concordance which resulted in a list of 517 synsets.[16] The definition of "synonym" here is quite broad, as Dershowitz et al. explain: "the word for 'plant' (נטע) and the word for 'transplant' (שתל) are considered synonymous because they both share the concept of 'plant' even though they are not identical in nuance."[17]

With this list of synsets, the authors ran a clustering algorithm on the chapters of Genesis through Numbers to divide the Priestly from the non-Priestly chapters. After this, they ran the algorithm a second time using a verse classifier that assigned individual verses to the Priestly or non-Priestly source based on the synonym usage. To accomplish the verse-by-verse division, a nearest neighbour smoothing technique was used when no synset features were found in a particular verse.[18] The results were then compared against a benchmark of scholarly

---

[12] D.L. Mealand, "Correspondence Analysis of Luke," *Literary and Linguistic Computing* 10/3 (1995): 172.

[13] The results of this are surprising for New Testament source critical scholars. However, the discussion of the results and their acceptance or rejection is outside of the scope of this study, as the primary concern here is the precedent and influence of this study. Mealand, "Correspondence Analysis," 183. See also, D.L. Mealand, "The Seams and Summaries of Luke and of Acts," *JSNT* 38/4 (2016): 482–502.

[14] Idan Dershowitz et al., "Computerized Source Criticism of Biblical Texts," *JBL* 134/2 (2015): 253–271.

[15] Deuteronomy was not included because of the lack of texts assigned to the Priestly source.

[16] Dershowitz et al., "Computerized Source," 257.

[17] Ibid.

[18] The use of "nearest neighbor" in Dershowitz et al. is not the clustering function commonly known by this name. The nearest neighbour is not the most closely related

opinions and matched for 91.4% of the verses.[19]

The algorithmic model of Dershowitz et al. is the basis for the methodology used here with two exceptions. First, the text will not be broken down into verses and the smoothing technique will not be used. The paucity of synset data in individual verses makes analysis at this level of detail difficult and the results questionable.[20] Second, the number of clusters were limited to two by the authors before the algorithm was run. The process used here does not set a specific number of groups from the beginning and so the ideal number of clusters will be determined by the computer algorithm rather than a pre-set limitation.

## D    INITIAL METHODOLOGICAL CONSIDERATIONS

The methods cited above and the one used in this study rely on synonyms. To do this type of analysis, the texts must be related. The importance of this relationship is that they must use words that can be put into synsets. An extreme example of unrelated texts is an auto repair manual compared with a newspaper. One could make synsets for the words in those texts but the two topics of the texts are so disparate that there would not be enough overlapping synonym information to compare them.[21]

Another challenge to synonymity is genre. For example, Hebrew poetry frequently uses unique terms that are not found in narratives. Therefore, even

---

cluster (as it is in the algorithmic function) but literally the nearest verse. For example, if verse 1 has features that assign it to the Priestly source and verse 2 has no definitive features, it will be assigned to the Priestly source based on its nearest neighbour (verse 1). Ibid., 264.

[19]    The benchmark used were the text assignments in Theodor von Nöldeke, *Untersuchungen zur Kritik des Alten Testaments* (Kiel: Schwers, 1869); S.R. Driver, *An Introduction to the Literature of the Old Testament* (9th ed.; Edinburgh: T&T Clark, 1913); Richard Elliott Friedman, ed., *The Bible with Sources Revealed: A New View into the Five Books of Moses* (1st ed.; San Francisco: Harper San Francisco, 2003). Dershowitz et al., "Computerized Source," 267.

[20]    The smoothing technique uses the synset representation of surrounding verses to categorise a verse without a synset representation. However, this leads to the question of whether that verse truly belongs in that cluster. If the data is unavailable in the verse, automatically aligning it with the surrounding verses seems at odds with the assumption of multi-source redacted texts that can be divided between individual verses or even individual clauses. The authors of the article do not address this potential impact on the results.

[21]    For example, "motor" and "engine" would be a synset and "president" and "CEO" could be another synset. However, the first pair would rarely occur in a newspaper while the second synset would rarely, if at all, occur in an automotive repair manual. In this instance, the clustering algorithm would be unable to measure a degree of difference because the texts do not share enough synonymous words to make synsets.

though the topic of the texts might be similar enough to make synsets, biblical poetry and narrative typically share similar theological and historical themes and they will be separated based on unique features of the genres rather than author word preference. The chapters chosen from Jeremiah deal with the same topic, the prophetic ministry of Jeremiah and are primarily from the same genre, narrative. Therefore, both potential problems are minimised.

The division of the text used in this study follows the chapter organisation for two reasons. The primary reason is the use of chapter divisions in modern scholarly opinion. Many scholars divide the text by chapter, therefore, the results are comparable to current opinions. For the analyses that do not follow the chapter divisions explicitly, they are usually close enough to the divisions to still be compared with the total chapter analysis. An example of this is the division by Stulman cited above. Stulman uses Jer 27:1–22 in his analysis, which is the whole chapter. However, in Jer 44, Stulman explicitly discusses only verses 1–14 even though there are 30 verses in the chapter. The rest of the chapter is not discussed and the next sample for Stulman is 45:1–5, which again is the whole chapter.[22] Thus, chapter 44 does not require further subdivision in the clustering model because the remaining verses are not treated separately by Stulman. This additional data might create background noise that hinders the algorithmic output from perfect alignment with Stulman but a correlation will still appear. The remedy to this background noise is to draw out significant features in the creation of each cluster.

The second reason that the chapters are used in their entirety is sample size. The text in Jeremiah is a total of 4887 words, not counting proper names. If the chapters were identical in length, there would be about 244 words per chapter (they are not identical but they are similar in length except for chapter 45). Mealand's division of Luke was into 500-word segments which is more than the word count of the chapters in Jeremiah. However, the study of the Pentateuch by Dershowitz et al. used chapter divisions to initially divide their text regardless of the length of the chapters. A further division of the text may obscure the results by creating sample sizes that are too small to contain the representative features that are being studied, as noted in Mealand's study and maintaining chapter divisions follows the earlier precedent by Dershowitz et al.

The synset methodology utilised here has some interpretative limitations. The terminology preferences cannot provide direct insight into sources, dating or literary organisation. Many scholars divide text by type of literature. However, the synset data cannot assess literary changes unless there are significant differences in vocabulary (e.g. unique poetic terminology). To avoid this potential confusion between genre changes and word preferences, the chapters chosen are primarily narrative.

---

[22]   Stulman, *The Prose Sermons*, 123.

A second method of dividing texts that scholars often employ is redactional organisation. This might be inferred from the choice of synonyms in a text. However, the choice of synonym in itself does not identify an author or a date. Several scholars have attempted to identify dtr and DH words or phrases and some of these will be compared with the data gathered in the results section.[23] The differences in word choices can be noted with this method but the scribal school or historical significance of certain terms cannot be deduced solely from the clustering algorithm.

The advantage of the computerised learning method is that a large number of words can be analysed to test the textual divisions that are normally done with only a handful of terms. In clustering with synonyms, the use of a specific term over and against alternative options is accentuated.[24] For example, in the distribution of בטח and אמן ("believe"), בטח occurs once in Jer 39, אמן occurs once in chapter 40 and neither occurs in chapter 41. Hypothetically using just this synset, chapters 39 and 40 would be placed into two different clusters while the null value of chapter 41 would leave it unidentified.[25]

## E    METHODOLOGY

The synsets were created from suggested synonyms in Clines's *Dictionary of Classical Hebrew*.[26] Some of the synonyms have been merged when they are variant spellings of the same word (e.g., masculine and feminine nouns) or simply variant vowel pointings. A dataset of 306 synsets was created based on the words found in Jeremiah and Ezekiel. Words with no synonym in Jeremiah

---

[23] Stulman, *The Prose Sermons*; Moshe Weinfeld, *Deuteronomy and the Deuteronomic School* (Oxford: Clarendon Press, 1972); Hornkohl, *Ancient Hebrew.*

[24] The use of synsets relies on the instability of features cited by Efstathios Stamatatos. The point is that the fewer synonyms available, the more stable the feature. For example, in English words "and" and "the" are extremely stable because they have no clear synonyms. However, "large" has many rough synonyms, therefore, it is very unstable. The precise formula for this feature is given in Efstathios Stamatatos, "A Survey of Modern Authorship Attribution Methods," *JASIST* 60/3 (2009): 545.

[25] In reality, the algorithm cannot run without the appearance of at least one synset feature in each text block. If this example were actually run through the algorithm, an error would be returned because there is a chapter with no value and therefore no classification. This is part of the reasoning behind the smoothing technique of Dershowitz et al. when they classified verses. If there is a verse that contains no terms from any of the synsets (for example one comprised only of proper names), the clustering algorithm will not be able to classify the dataset and will return an error message.

[26] David J.A. Clines, *The Dictionary of Classical Hebrew* (9 vols.; Sheffield: Sheffield Academic Press, 1993). Note that synonymity is meant in a very broad sense. The point is overlapping meaning not precise/exact equality in meaning between two or more words.

or Ezekiel were not used in this study as they would not have a direct comparable word in another chapter.[27]

After the dataset was compiled, a training exercise was performed with Jer 26–45 and Ezek 1–12. The goal was to train the program to separate the chapters of Jeremiah and Ezekiel. This ensured that the dataset was working properly and, more importantly, that the algorithm was separating out the clusters correctly with the features that were input.

The computer modelling method used was hierarchical clustering (hclust) in R.[28] This was programmed to use an agglomerative approach[29] with a complete linkage[30] using the Euclidean distance.[31] Once the clustering model was created, a silhouette analysis was done to obtain the optimal number of clusters.[32]

---

[27]   The qualification of only words found within the texts of Jeremiah and Ezekiel is the reason behind the smaller synset from Dershowitz et al. Including words without a synonym in the text would not affect results. If part of a synset returns a null value in all the chapters, it would be excluded because of lack of comparable data.

[28]   See   https://www.rdocumentation.org/packages/stats/versions/3.5.2/topics/hclust. for the documentation on the hierarchical clustering package within the R program.

[29]   This builds clusters from the bottom up. Therefore, the program begins with 20 separate chapters and links them together based on similarity until they are all in one cluster.

[30]   Complete linkage is a comparison of the furthest pair of individuals within the clusters (single linkage is the opposite: clustering with the nearest pair). Since the distance of connection is larger, it is less sensitive to observational errors than single linkage. Brian S. Everitt et al., *Cluster Analysis* (5th ed.; Wiley Series in Probability and Statistics; London: Wiley, 2010). §4.2.2 and Lawrence Hubert, "Approximate Evaluation Techniques for the Single-Link and Complete-Link Hierarchical Clustering Procedures," *JASA* 69/347 (1974): 698–704.

[31]   Euclidean distance is the most common form of clustering. Manhattan and Squared Euclidean are less common but provide the same results on different scales. However, Euclidean has been shown to provide the clearest interpretation of the raw data compared with the other primary distance measures. Michael R. Anderberg, *Cluster Analysis for Applications* (Probability and Mathematical Statistics 19; New York: Academic Press, 1973), 55.

[32]   The algorithm process for silhouette is: 1. In each observation *I,* the average dissimilarity ai between *i* and all other points of the cluster to which *i* belongs. 2. All other clusters C, to which *i* does not belong, the average dissimilarity $d(i,C)$ of *i* to all observations of C is calculated. The smallest of these $d(i,C)$ is defined as $b_i = min_c d(i,C)$. The value of $b_i$ is the dissimilarity between *i* and its nearest neighbour cluster. 3. The silhouette width of the observation *i* is defined by the formula: $S_i = (b_i - a_i)/max(a_i, b_i)$.

## F    RESULTS

The clustering results have been organised graphically in a dendogram which represents the relationship of the chapters from the most closely related attached at the bottom to the most distantly related chapters which branch apart at the top. The final training set of Jeremiah and Ezekiel is shown in the appendix Figure 1 and the test set with Jeremiah alone in Figure 2. The dendogram in Figure 2 was cut at six clusters based on the analysis of the optimal clustering in the silhouette analysis (Figure 3). The six-part computerised division was also added to the list of scholarly opinions in Table 1 so that comparisons between them can be easily identified (under "Cluster Data").[33]

Comparing the cluster data and Römer's redactional divisions provided in Table 1, the algorithmic division is not as simplistic as Römer's hypothesis. In agreement with his theory, the legitimising linkage (36) between the first (27–35) and second (37–44) redactions is in a cluster alone which may indicate a mixture of two different styles. Additionally, half of the chapters in his second redaction are located in clusters outside of cluster one, while most of the first redaction chapters are located within cluster one (all but 31 and 32).

The computer model disagrees with Römer in the second redaction group because chapters 36–45 are not homogeneous within themselves. The chapters in clusters five and six (38, 40, 41) are the highest split on the dendogram in Figure 2 which means they are the most distinct from the rest of the chapters. They are unrelated even to chapters 37, 39 and 42–44 which are supposed to be in the same redactional unit.[34] Unfortunately, Römer does not list the words or phrases that create this split outside of the blanket claim that there are DH vocabulary, compositional techniques and theological distinctives.[35]

On the other hand, Stulman's narrative analysis has a stronger correlation with the algorithmic model. Though he does not deal with all the chapters, the ones that are listed match quite well. The chapters where more than 50% of the terms are from dtr IV are all within cluster one (27, 29, 35, 39, 45). All the chapters comprised of less than 5% dtr I and II material are in cluster one as well (27, 29, 33, 39, 45). Chapter 32 has dtr I and II terms comprising more than 50%

---

[33]    Cluster numbering will follow the numbers given in table 1, which means that chapter 31 is cluster two, chapters 32 and 44 are cluster three, et cetera.

[34]    The response to this might be that they are in fact different sources. Römer claims that chapters 37–44 were compiled by the redactors and not that they were written by them. Thus, it is possible that some of the chapters had different authors and the redactors then attached texts from a variety of authors onto the first kernel of homogeneous text (26–34). However, Römer does not specify which of the chapters are from different authors and why some of them seem to be closely related with the first redactional element. Römer, "How Did Jeremiah Become a Convert?," 198.

[35]    Römer, "How Did Jeremiah Become a Convert?," 197.

of the text and chapter 44 have a significant amount of dtr I and II language at more than 10%. Both are higher than any other chapter in this section and they are the only two chapters in cluster three.

A few examples from Stulman to illustrate this division are dtr I terms תועבה and שקץ against dtr IV terms עבד and נביא.[36] The dtr I terms were paired in a synset together as they both mean "abomination." However, they both only occur in cluster three (תועבה .45% of the total words, שקץ.15%). Conversely, the two dtr IV terms are much more common and spread throughout the clusters (עבד .87% of cluster one, .15% of cluster three and נביא 1.98% of cluster one, .45% of cluster three).[37] That the first two terms are found only cluster three creates a node of similarity where there is a null value in cluster one. The other two terms, dtr IV, are shared between both clusters one and three. However, the frequency of each term is significantly higher in cluster one. The difference between dtr I and IV occurrences in Stulman is illustrated by the dendogram from the computerised model. Cluster three contains terms distinct from cluster one (specifically dtr I) and they both share terms but at different rates (dtr IV is more frequent in cluster one). The split between the clusters is not the highest on the dendogram but it is still quite significant.

For his part, Hornkohl contrasts an Early Biblical Hebrew term שׂר with the Late Biblical Hebrew term חרים, both meaning "nobles."[38] The latter term is considered an instance of Late Biblical Hebrew because of the proliferation in Nehemiah more than any other book of the Old Testament while the former term is far more frequent overall and especially in earlier texts.[39] The late term only occurs in cluster one and even at that, it is only .09% of the total words but שׂר is far more interesting in its distribution. It is .87% of cluster one, not found in cluster two, .45% of cluster three, 1.74% of cluster four, 2.39% of cluster five and 1.63% of cluster six. Comparing these numbers with the dendogram, there

---

[36]   Stulman, *The Prose Sermons*, 35–43.

[37]   The total occurrence percentages for the dtr IV terms are עבד .87% of cluster one, .15% of cluster three, .65% of cluster six; נביא 1.98% of cluster one, .45% of cluster three, .69% of cluster four, 1.20% of cluster five. The only occurrences being addressed here are in clusters one and three since the passages that Stulman cites are only found in those two.

[38]   These two terms are focused on specifically for two reasons. First, Hornkohl provides both an early and late term, which he does not do with every Late Biblical Hebrew term that he lists. Second, both the early and late terms occur within chapters 26–45. Some of the pairs that Hornkohl lists have only one term appearing in Jeremiah or only one of the two terms occurs within this passage. These two, however, both occur in chapters 26–45 and with enough frequency to make the analysis valuable.

[39]   Hornkohl, *Ancient Hebrew*, 301.

is a definite correlation. Cluster five has a significantly higher frequency of שׂר than any other cluster and on the dendogram it is also the most distinct cluster. Clusters one, two and three are all under 1% and they are all close to each other on the dendogram split. Clusters four and six break from the expectation here. Both of them are about halfway between cluster five and clusters one and three. Clusters four and six are not near each other though. Cluster four is near cluster three while cluster six is near cluster five. The differences of other synsets are strong enough to separate these two clusters even though they are quite close on the frequency of this term.

Viewing this synset overall is interesting because the late term only occurs in cluster one while the early term occurs in all the clusters except two which has a null value because neither of the terms occurs in it. Combining this with Römer's redaction theory, there appears to be a conflict. The first redaction, chapters 26–35, is primarily cluster one which has the only occurrences of the Late Biblical Hebrew term (though only one occurrence of it is within 26–35). Conversely, the second redaction, chapters 37–44, has the Classical Biblical Hebrew term in almost every chapter (and the other occurrence of the Late Biblical Hebrew term). The division by Römer seems to be challenged not only by the computer model but also by the other two scholarly views. On the other hand, though the computer model lines up fairly well with Stulman and Hornkohl, it is not an exact match. To elucidate one cause of the minor dissonance, a synset of far more frequent terms in the text will be addressed.

The more frequent synset is אִישׁ, אדם, אנושׁ, נבר, נפשׁ meaning "man/human." Cluster three contains every term roughly the same number of times (.75%, .45%, .30%, .15% and .75% of the total words, respectively). However, cluster six is heavily skewed toward אנושׁ over the other synonyms (.98%, 0, 4.58%, .33% and .65%, respectively). As much as cluster six uses אנושׁ, there are no occurrence of it in clusters two and four; אנושׁ makes up .51% of the words in cluster one but אִישׁ is double that at 1.06%.

Turning to the dendogram, clusters one and three are both under 1% for אנושׁ while two and four have no occurrences. All of these are close together while cluster six at 4.58% is separated from them at the top of the dendogram. Cluster five has אנושׁ at 2.39% of the total words which is the second highest in occurrences after cluster six with which it is paired in the dendogram. Similarly, אדם is also interesting in that the only clusters that do not contain the term are five and six (cluster one .23%, cluster two .58%, cluster three .45%, cluster four .35%). The synset for "man/human" occurs at a fairly high frequency and the lexical choices match the expectation based on the dendogram even more closely than the keywords presented by the scholars. The two clusters that are

most different, five and six, are also the two with the largest difference in synonym frequency in the scholarly terms and the "man/human" synset but the smaller differences between clusters are seen more clearly in the "man/human" synset than the significant scholarly terms.

## G    DISCUSSION

The field of computerised learning algorithms in linguistics is still expanding as the software and processing power of computers are growing. The question is where does this type of analysis fit within traditional textual criticism?

First, the inherent bias in this method should be addressed. Computer modelling, especially the synset method, is not an unbiased standard to measure the other (biased?) scholarly methods. Dershowitz et al. makes the claim that, "Our method is also less liable to accusations of bias, thanks to its reliance on context-independent criteria."[40] However, bias is evident from the beginning when a text is assumed to be from multiple sources (cf. Dershowitz's and Mealand's articles) and especially when the number of sources is defined upfront, like the two sources defined by Dershowitz et al. The process of synonym pairing is also made with human bias. This goes from the source of the synonyms, my use of *The Dictionary of Classical Hebrew* versus Dershowitz's use of *Strong's Concordance*, to how much semantic overlap is needed for terms to be considered synonymous. Admitting the human element and inherent bias in this process is not meant to negate the truth of the results. In fact, it may allow the results to be assessed more accurately as the results of study by human scholars using modern tools instead of being accepted as objective, mathematical truth.

As shown in this study, and in the earlier works by Mealand and by Dershowitz et al., the computer model is always compared with current scholarship. The computer analysis can draw upon a much larger amount of data to bring out details that might have been missed by scholarship looking at only select words or phrases. It can thereby strengthen or question arguments being made in current scholarship.

The other side of this is that human scholarship is necessary to understand the algorithmic results. When thousands of words are involved, who designates which words are significant? The software can list the terms but it is the scholars who must interpret that data. Pulling out a handful of high frequency terms ignores the vast amount of lower frequency terms that also affect the algorithm. At the same time, searching through thousands of lower frequency terms to find a significant synset could be unfruitful; not only in the time spent searching but also in defining what lexemes are "significant" if they have not been deemed so by modern scholars. The difference between algorithmically significant and

---

[40]    Dershowitz et al., "Computerized Source," 253.

stylistically significant terms or phrases is unclear. It often appears as though scholars do not distinguish between these two categories but these are not necessarily related. In this study, "man/human" synonyms appear algorithmically significant but other, less frequent, terms have been identified as stylistically significant by scholars. How does one decide which terms are significant and deserving of more detailed study?

The final two critiques of this method concern Hebrew and synset modelling, specifically. The method of using ancient Hebrew roots to make synonyms can be problematic for two reasons. The first is the uncertainty of meaning in the ancient language. Certain words, especially *hapax legomena*, are unclear and their meanings are often derived from context or cognates in other Semitic languages. Though many of the unclear terms are rare, using words with questionable glosses for synonym pairing can create background noise at the very least.

The second problem is that Hebrew roots can have more than one meaning. For example, the dtr verb סור, identified by Weinfeld, could be paired with מוש, both roughly sharing the meaning "turn aside/back."[41] This synset appears valid with the use of סור in Jer 32:40 and מוש in 31:36. However, it is used in 32:31 in the phrase וְעַד הַיּוֹם הַזֶּה לַהֲסִירָהּ מֵעַל פָּנָי ("and until this day to remove it from before my face"). In this case, סור does not easily translate as "to turn aside (it from before my face?)." This is perhaps a related movement away from the subject but most likely not an exact synonym.[42]

An even more obvious example of this issue is the possible synset of נפל and קרא.[43] Whereas קרא has a meaning of "to befall/happen" that overlaps with נפל which can also mean to "fall/fall upon,"[44] this translation is valid for קרא in both usages of Jer 26–45[45] but this is not the case for נפל. This term is used for desertion to the enemy (37:13-14; 38:19; 39:9), presenting a request (36:7; 37:20; 38:26; 42:2, 9) and being killed (39:18; 44:12). These two terms are not interchangeable in every location. Though it could be argued that presenting a request could be done with קרא (even with the meaning "proclaim"), it is very

---

[41]    Weinfeld, *Deuteronomy and the Deuteronomic School*, 339.

[42]    For this secondary meaning of סור, see Clines, *The Dictionary of Classical Hebrew* 6: 140–141.

[43]    For the claim of נפל as a dtr term, see Weinfeld, *Deuteronomy and the Deuteronomic School*, 330.

[44]    See Clines, *The Dictionary of Classical Hebrew*, 5:716, 7:303.

[45]    Jeremiah 32:23 says, "you caused this calamity to befall them" and 44:23, "calamity has befallen you."

difficult to argue that it could replace נפל in the standard formula for being killed נפל + חרב. Synset pairs do not account for syntax nor multiple meanings of individual roots. For this reason, the computer model can guide researchers into profitable areas of inquiry but it cannot be accepted without corroboration by other methods and detailed analysis of the input data.

The results of the synset modelling in Jeremiah line up well with two of the three scholars—Hornkohl and Stulman. Though one is looking at Hebrew language periodisation and the other at Deuteronomic redaction, the computer model divides the text without reference to any historical background of the text. In this respect, both can be compared with the algorithm results even if they are not comparable to each other. Surprisingly, both have significant overlap with the computer model even though they utilise different terms and different methodologies. In this case, the computerised learning technique largely strengthens these scholars' arguments but the divergences of the clustering algorithm can be used to further examine the discrepancies of the differences noted by these scholars.

# APPENDIX

Table 1

| Chapter | Cluster Data | Römer[46] | Stuhlman[47] | Hornkohl[48] | Bowman[49] | Brueggeman[50] | Callaway[51] | Carroll[52] | Davidson[53] | Domeris[54] |
|---|---|---|---|---|---|---|---|---|---|---|
| 26 | Cluster 1 | First dtr Redaction | Prose Cycle | | | | | | | |
| 27 | Cluster 1 | | dtr I + II < 5% | | | | | | | |
| 28 | Cluster 1 | | dtr IV >50% | החם | | | | | | |
| 29 | Cluster 1 | | dtr IV >50% / dtr I + II < 5% | אשר | | Threat against Egypt | | King Narrative 1 | | |
| 30 | | | | | | | | | | |
| 31 | Cluster 2 | | | | | | | | | |
| 32 | Cluster 3 | | dtr I + II >50% | אשר | 32-33 are as Utopian as 30-31 (contra Sharp) | | | | | Land purchase verbs ga'al, qanah, chalaq |
| 33 | Cluster 1 | | Book of Consolation | | | | | | | |
| 34 | Cluster 1 | | dtr IV < 5% / dtr I + II < 5% | אשר | | | | King Narrative 2 | | |
| 35 | | | dtr IV >50% | אשר | | | | | | |
| 36 | Cluster 4 | Legitimizing Linkage | | אשר | | | | | | |
| 37 | Cluster 1 | Second dtr Redaction | | אשר | | | Highly Unique Vocabulary | | Similar | |
| 38 | Cluster 5 | | | אשר | | | | King Narrative 3 | Similar | |
| 39 | Cluster 1 | | dtr IV >50% / dtr I + II < 5% | חיים, אשר | | | | | Parallel | |
| 40 | Cluster 6 | | dtr IV >50% / dtr I + II < 5% | אשר | | | | | Parallel | |
| 41 | Cluster 1 | | | אשר | | | | | | |
| 42 | Cluster 1 | | Literary Unit | אשר | | Redaction 1 | | | | |
| 43 | | | | | | | | | | |
| 44 | Cluster 3 | Third dtr Redaction | dtr I + II >10% / dtr IV > 10% | אשר | | Threat against Egypt / Redaction 2 | | King Narrative 4 | | |
| 45 | Cluster 1 | | dtr IV >50% / dtr I + II < 5% | | | | | | | |

| Green[55] | Hill[56] | Holt[57] | Nicholson[58] | Perdue[59] | Seitz[60] | Sharp[61] | Wanke[62] | Yates[63] |
|---|---|---|---|---|---|---|---|---|
| Prophetic Scene | | | Rejection of the Word | Pre-597 | Jehoiakim — speech | Babylonian | Section 1 | A - prophetic word |
| Prophetic Scene | | | Fake Prophecy | Pre-597 | Zedekiah's Reign — activity | Foreign Dialect | Section 1 | B - prophetic conflict / Reign of Jehoiakim block 1 |
| Prophetic Scene | | Dialogue Engendered by Proclamations | Book of Consolation | post-597 redaction | Zedekiah's Reign — speech | Conspicuous latecomers | | C - aftermath of exile |
| Prophetic Scene | | | Disobedience | | | Judahite | | D - covenant infidelity |
| Prophetic Scene | 4th Year of Jehoiakim | | Rejection of the Word | Pre-597 | Jehoiakim — activity | | Section 1 | A - prophetic word |
| | | | Consequences | post-597 redaction | Zedekiah's Reign — activity | | Section 2 | B - prophetic conflict |
| | | History of the word during siege/fall of Jerusalem | Aftermath | Passion Narrative / post-597 redaction | Zedekiah's Reign — speech | Babylonian | | C - aftermath of exile / Reign of Jehoiakim block 2 |
| | 4th Year of Jehoiakim | | Egypt; Appendix | post-597 redaction | | | Section 3 | D - covenant infidelity |

[46]    Römer, "How Did Jeremiah Become a Convert?" 197–198.

[47]    Stulman, *The Prose Sermons*, 230; Louis Stulman, "The Prose Sermons as Hermeneutical Guide to Jeremiah 1–25: The Deconstruction of Judah's Symbolic World," in *Troubling Jeremiah* (ed. A.R. Diamond, Kathleen M. O'Connor and Louis Stulman; JSOTSup 260; Sheffield: Sheffield Academic Press, 1999), 37.

[48]    Hornkohl, *Ancient Hebrew*, 301.

[49]    Barrie Bowman, "Future Imagination: Utopianism in the Book of Jeremiah," in *Jeremiah (Dis)Placed: New Directions in Writing/Reading Jeremiah* (ed. A.R. Diamond and Louis Stulman; Library of Hebrew Bible/Old Testament Studies 529; New York: T&T Clark, 2011), 247.

[50]    Walter Brueggemann, "The 'Baruch Connection': Reflections on Jeremiah 43.1-7," in *Troubling Jeremiah* (ed. A.R. Diamond, Kathleen M O'Connor and Louis Stulman; JSOTSup 260; Sheffield: Sheffield Academic Press, 1999), 372–373, 380–381.

[51]    Mary Chilton Callaway, "Black Fire on White Fire: Historical Context and Literary Subtext in Jeremiah 37–38," in *Troubling Jeremiah* (ed. A.R. Diamond, Kathleen M. O'Connor and Louis Stulman; JSOTSup 260; Sheffield: Sheffield Academic Press, 1999), 175–177.

[52]    Robert P. Carroll, "The Book of J: Intertextuality and Ideological Criticism," in *Troubling Jeremiah* (ed. A.R. Diamond, Kathleen M. O'Connor and Louis Stulman; JSOTSup 260; Sheffield: Sheffield Academic Press, 1999), 228–229.

[53]    Steed Vernyl Davidson, "Ambivalence and Temple Destruction: Reading the Book of Jeremiah with Homi Bhabha," in *Jeremiah (Dis)Placed: New Directions in Writing/Reading Jeremiah* (ed. A.R. Diamond and Louis Stulman; Library of Hebrew Bible/Old Testament Studies 529; New York: T&T Clark, 2011), 150, 162.

[54]    William R. Domeris, "The Land Claim of Jeremiah: Was Max Weber Right?" in *Jeremiah (Dis)Placed: New Directions in Writing/Reading Jeremiah* (ed. A.R. Diamond and Louis Stulman; Library of Hebrew Bible/Old Testament Studies 529; New York: T&T Clark, 2011), 145.

[55]    Barbara Green, "Sunk in the Mud: Literary Correlation and Collaboration between King and Prophet in the Book of Jeremiah," in *Jeremiah Invented: Constructions and Deconstructions of Jeremiah* (ed. Else K. Holt and Carolyn J. Sharp; Library of Hebrew Bible/Old Testament Studies 595; New York: T&T Clark, 2015), 35.

[56]    John Hill, "The Dynamics of Written Discourse and of the Book of Jeremiah MT," in *Jeremiah (Dis)Placed: New Directions in Writing/Reading Jeremiah* (ed. A.R. Diamond and Louis Stulman; Library of Hebrew Bible/Old Testament Studies 529; New York: T&T Clark, 2011), 108.

[57]    Else K. Holt, "The Potent Word of God: Remarks on the Composition of Jeremiah 37–44," in *Troubling Jeremiah* (ed. A.R. Diamond, Kathleen M. O'Connor and Louis Stulman; JSOTSup 260; Sheffield: Sheffield Academic Press, 1999), 162.

[58]    Ernest W. Nicholson, *Preaching to the Exiles: A Study of the Prose Tradition in the Book of Jeremiah* (Oxford: Blackwell, 1970), 111.

[59]    Leo G. Perdue, *The Collapse of History: Reconstructing Old Testament Theology* (Overtures to Biblical Theology; Minneapolis: Fortress Press, 1994), 247–259.

[60]    Christopher R. Seitz, *Theology in Conflict: Reactions to the Exile in the Book of*

Figure 1
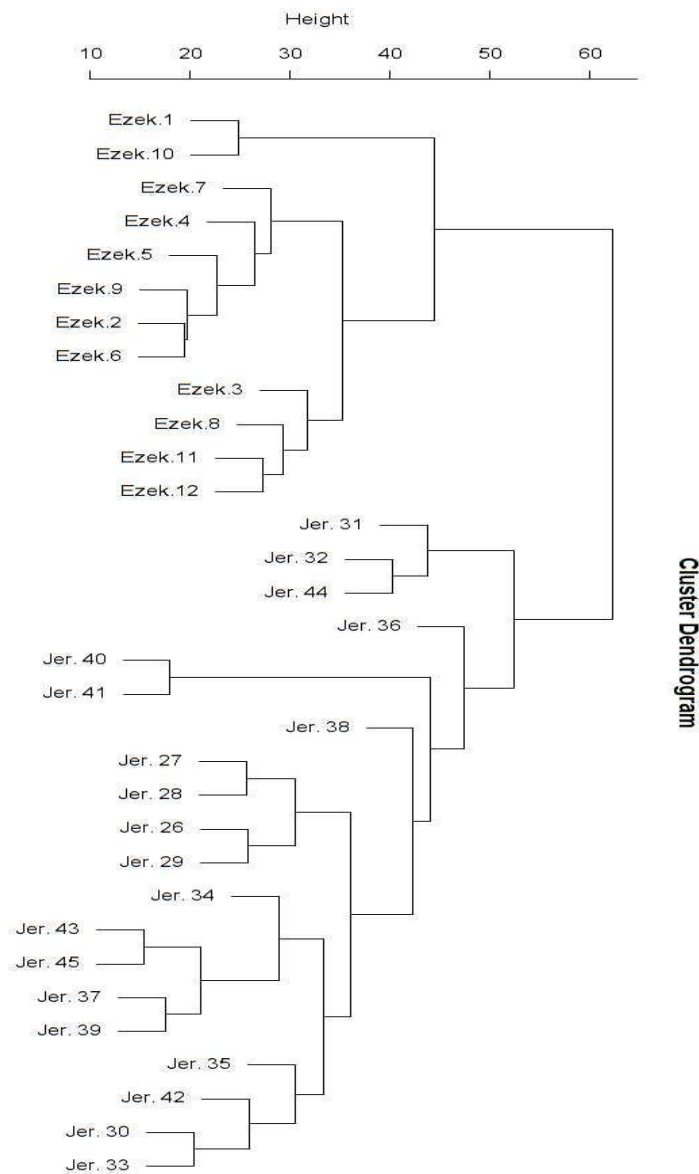


*Jeremiah* (BZAW 176; New York: De Gruyter, 1989), 223–234.

[61]    Carolyn J. Sharp, *Prophecy, and Ideology in Jeremiah: Struggles for Authority in Deutero-Jeremianic Prose* (OTS; New York: T&T Clark, 2003), 42–43.

[62]    Gunther Wanke, *Untersuchungen zur Sogenannten Baruchschrift* (Berlin: De Gruyter, 1971), 144.

[63]    Gary E. Yates, "Narrative Parallelism and the 'Jehoiakim Frame': A Reading Strategy for Jeremiah 26–45," *JETS* 48/2 (2005): 264; Gary E. Yates, "'The People Have not Obeyed': A Literary and Rhetorical Study of Jeremiah 26–45" (PhD. Diss., Liberty University Faculty Dissertations, Dallas Theological Seminary, 1998), 6.
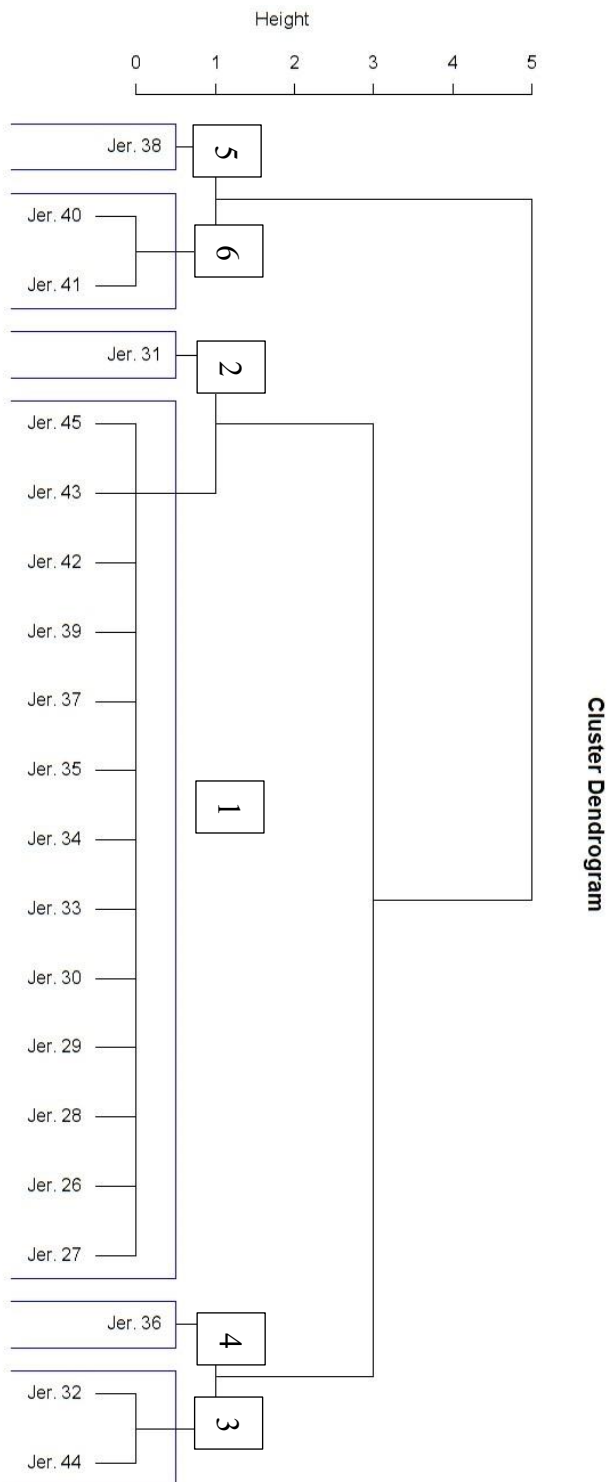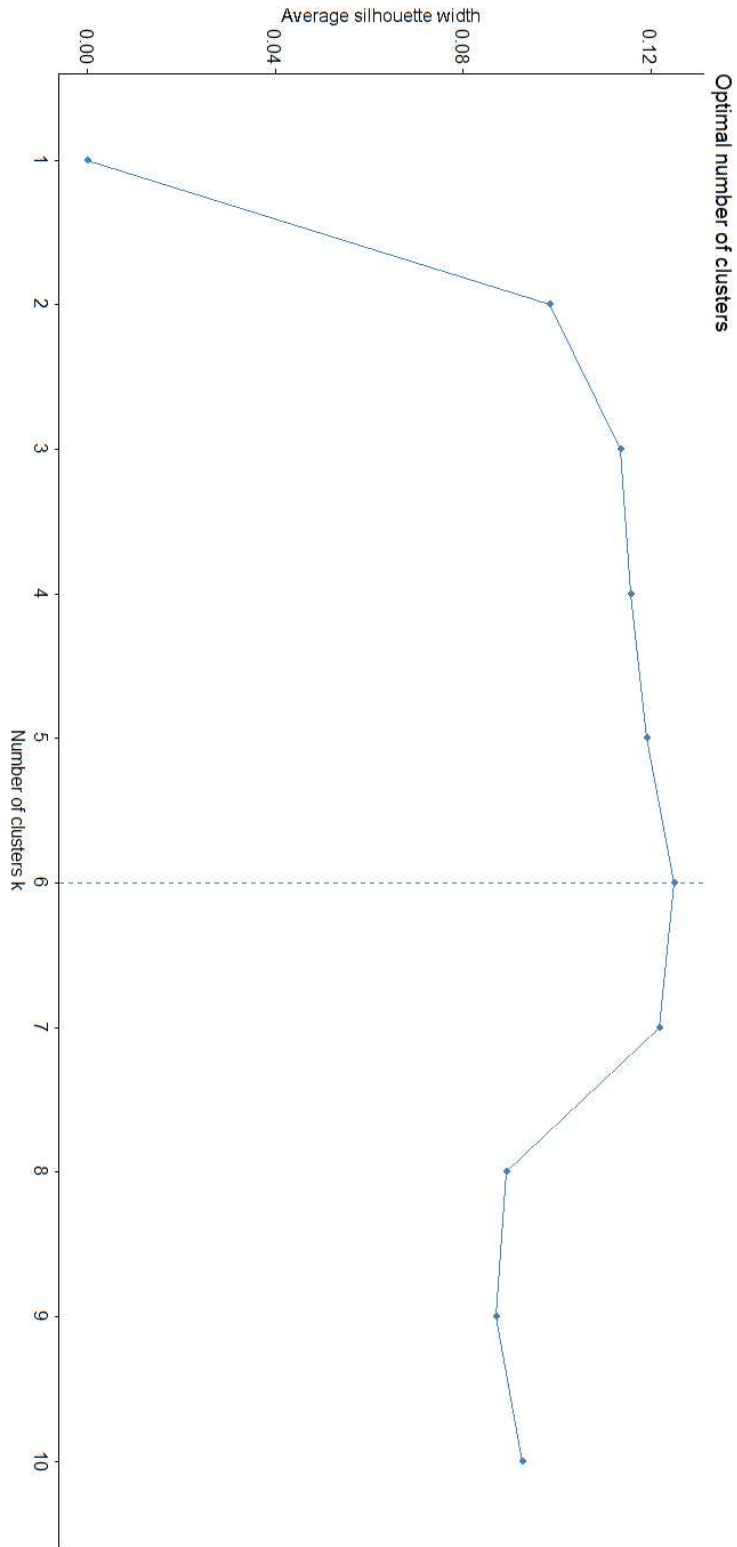
Figure 2

Figure 3

## I    BIBLIOGRAPHY

Anderberg, Michael R. *Cluster Analysis for Applications*. Probability and Mathematical Statistics 19. New York: Academic Press, 1973.

Bowman, Barrie. "Future Imagination: Utopianism in the Book of Jeremiah." Pages 243–249 in *Jeremiah (Dis)Placed: New Directions in Writing/Reading Jeremiah*. Edited by A.R. Diamond and Louis Stulman. Library of Hebrew Bible/Old Testament Studies 529. New York: T&T Clark, 2011.

Brueggemann, Walter. "The 'Baruch Connection': Reflections on Jeremiah 43.1-7." Pages 367–386 in *Troubling Jeremiah*. Edited by A.R. Diamond, Kathleen M. O'Connor and Louis Stulman. Journal for the Study of the Old Testament Supplement Series 260. Sheffield: Sheffield Academic Press, 1999.

Callaway, Mary Chilton. "Black Fire on White Fire: Historical Context and Literary Subtext in Jeremiah 37–38." Pages 171–178 in *Troubling Jeremiah*. Edited by A.R. Diamond, Kathleen M. O'Connor and Louis Stulman. Journal for the Study of the Old Testament Supplement Series 260. Sheffield: Sheffield Academic Press, 1999.

Carroll, Robert P. "The Book of J: Intertextuality and Ideological Criticism." Pages 220–243 in *Troubling Jeremiah*. Edited by A.R. Diamond, Kathleen M. O'Connor and Louis Stulman. Journal for the Study of the Old Testament Supplement Series 260. Sheffield: Sheffield Academic Press, 1999.

Clines, David J.A. *The Dictionary of Classical Hebrew*. 9 vols. Sheffield: Sheffield Academic Press, 1993.

Davidson, Steed Vernyl. "Ambivalence and Temple Destruction: Reading the Book of Jeremiah with Homi Bhabha." Pages 162–171 in *Jeremiah (Dis)Placed: New Directions in Writing/Reading Jeremiah*. Edited by A.R. Diamond and Louis Stulman. Library of Hebrew Bible/Old Testament Studies 529. New York: T&T Clark, 2011.

Dershowitz, Idan, Moshe Koppel, Navot Akiva and Nachum Dershowitz. "Computerized Source Criticism of Biblical Texts." *Journal of Biblical Literature* 134/2 (2015): 253–271.

Domeris, William R. "The Land Claim of Jeremiah: Was Max Weber Right?" Pages 136–149 in *Jeremiah (Dis)Placed: New Directions in Writing/Reading Jeremiah*. Edited by A.R. Diamond and Louis Stulman. Library of Hebrew Bible/Old Testament Studies 529. New York: T&T Clark, 2011.

Driver, S. R. *An Introduction to the Literature of the Old Testament*. 9th edition. Edinburgh: T&T Clark, 1913.

Everitt, Brian S., Sabine Landau, Morven Leese and Daniel Stahl. *Cluster Analysis*. 5th edition. Wiley Series in Probability and Statistics. London: Wiley, 2010.

Friedman, Richard Elliott, ed. *The Bible with Sources Revealed: A New View into the Five Books of Moses*. 1st edition. San Francisco: Harper San Francisco, 2003.

Green, Barbara. "Sunk in the Mud: Literary Correlation and Collaboration between King and Prophet in the Book of Jeremiah." Pages 34–48 in *Jeremiah Invented: Constructions and Deconstructions of Jeremiah*. Edited by Else K. Holt and Carolyn J. Sharp. Library of Hebrew Bible/Old Testament Studies 595. New York: T&T Clark, 2015.

Hill, John. "The Dynamics of Written Discourse and of the Book of Jeremiah MT." Pages 104–111 in *Jeremiah (Dis)Placed: New Directions in Writing/Reading*

*Jeremiah*. Edited by A. R. Diamond and Louis Stulman. Library of Hebrew Bible/Old Testament Studies 529. New York: T&T Clark, 2011.

Holt, Else K. "The Potent Word of God: Remarks on the Composition of Jeremiah 37–44." Pages 161–170 in *Troubling Jeremiah*. Edited by A.R. Diamond, Kathleen M. O'Connor and Louis Stulman. Journal for the Study of the Old Testament Supplement Series 260. Sheffield: Sheffield Academic Press, 1999.

Hornkohl, Aaron D. *Ancient Hebrew Periodization and the Language of the Book of Jeremiah: The Case for a Sixth-Century Date of Composition*. Vol. 74. Studies in Semitic Languages and Linguistics. Boston: Brill Nijhoff, 2014.

Hubert, Lawrence. "Approximate Evaluation Techniques for the Single-link and Complete-link Hierarchical Clustering Procedures." *Journal of the American Statistical Association* 69/347 (1974): 698–704.

Mealand, D.L. "Correspondence Analysis of Luke." *Literary and Linguistic Computing* 10/3 (1995): 171–182.

_____. "The Seams and Summaries of Luke and of Acts." *Journal for the Study of the New Testament* 38/4 (2016): 482–502.

Nicholson, Ernest W. *Preaching to the Exiles: A Study of the Prose Tradition in the Book of Jeremiah*. Oxford: Blackwell, 1970.

Nöldeke, Theodor von. *Untersuchungen zur Kritik des Alten Testaments*. Kiel: Schwers, 1869.

Perdue, Leo G. *The Collapse of History: Reconstructing Old Testament Theology*. Overtures to Biblical Theology. Minneapolis: Fortress Press, 1994.

Römer, Thomas. "How Did Jeremiah Become a Convert to Deuteronomistic Ideology?" Pages 189–199 in *Those Elusive Deuteronomists: The Phenomenon of Pan-Deuteronomism*. Edited by Linda S. Schearing and Steven L. McKenzie. Journal for the Study of the Old Testament Supplement Series 268. Sheffield: Sheffield Academic Press, 1999.

Seitz, Christopher R. *Theology in Conflict: Reactions to the Exile in the Book of Jeremiah*. Beiheft zur Zeitschrift für die Alttestamentliche Wissenschaft 176. New York: De Gruyter, 1989.

Sharp, Carolyn J. *Prophecy and Ideology in Jeremiah: Struggles for Authority in Deutero-Jeremianic Prose*. Old Testament Studies. New York: T&T Clark, 2003.

Stamatatos, Efstathios. "A Survey of Modern Authorship Attribution Methods." *Journal of the American Society for Information Science and Technology* 60/3 (2009): 538–556.

Stulman, Louis. *The Prose Sermons of the Book of Jeremiah: A Redescription of the Correspondences with Deuteronomistic Literature in the Light of Recent Text-critical Research*. Dissertation Series/Society of Biblical Literature 83. Atlanta: Scholars Press, 1986.

_____. "The Prose Sermons as Hermeneutical Guide to Jeremiah 1–25: The Deconstruction of Judah's Symbolic World." Pages 34–63 in *Troubling Jeremiah*. Edited by A.R. Diamond, Kathleen M. O'Connor and Louis Stulman. Journal for the Study of the Old Testament Supplement Series 260. Sheffield: Sheffield Academic Press, 1999.

Van Seters, John. "Author or Redactor?" *The Journal of Hebrew Scriptures* 7 (2007): 1–23.

Wanke, Gunther. *Untersuchungen zur Sogenannten Baruchschrift*. Berlin: De Gruyter, 1971.

Weinfeld, Moshe. *Deuteronomy and the Deuteronomic School*. Oxford: Clarendon Press, 1972.

Yates, Gary E. "Narrative Parallelism and the 'Jehoiakim Frame': A Reading Strategy for Jeremiah 26–45." *Journal of the Evangelical Theological Society* 48/2 (2005): 263–281.

_____. "'The People Have Not Obeyed': A Literary and Rhetorical Study of Jeremiah 26–45." PhD Dissertation, Dallas Theological Seminary, 1998.

Nicholas J. Campbell, The Southern Baptist Theological Seminary; Email: ncampbell570@students.sbts.edu.  ORCID:  https://orcid.org/0000-0002-8713-4647.