Good morning. Over the past several years, I have been exploring the use of statistical tools in ancient Hebrew linguistics. At the 2015 annual meeting of SBL, I presented and then published a paper highlighting the usefulness of hierarchical clustering and dendrograms in Hebrew studies. Then, last year I applied those techniques to the study of the diachronic change of Hebrew. That paper focused on ten linguistic features in biblical Hebrew. In this paper, I expand upon those studies to analyze 13 features of Biblical and Qumran Hebrew. I also move beyond Hierarchical clustering to other cluster methods as well as some further statistical tests that I will discuss in a moment. I will first lay out the design of my project, then move on to describing the linguistic profile of biblical and Qumran texts. I will end by presenting some conclusions and further questions that need exploration.

The research question that is driving this project is: In what ways are Qumran Hebrew and Biblical Hebrew related? In order to explore this question, I built linguistic profiles for all of the biblical texts and two texts from Qumran – The War Scroll and The Community Rule. These linguistic profiles consist of data for 13 linguistic features. Most of the features that I have selected have been considered by at least some scholars to reveal diachronic development in the Hebrew language. Selecting features that could be considered useful for linguistically dating Hebrew texts will allow me to explore diachronic explanations for differences between biblical texts and Qumran texts. This paper will consider the linguistic profiles for each text holistically first and then will turn to the specific features that differentiate texts.

Before turning to the data that this paper is based on, I would like to outline the design of my project. The first corpus I have chosen to work with includes all the biblical books found in the modern Hebrew Bible, which at times, I separate into two subcorpora: the so-called “Early” Biblical Hebrew texts consisting of Genesis through Kings and the so-called “Late” Biblical texts including Esther through Chronicles. Throughout this paper, I will refer to these two corpora as “EBH” and “LBH” respectively out of convenience while recognizing that there is significant debate surrounding the dating of these corpora. The second corpus utilized for this project consists of The War Scroll and The Community Rule. Note that this is a deviation from what my abstract proposed. This came about due to the database upon which I have based this work. Originally, I had planned to gather the data for this project using one of the prominent Bible software programs which would have given me access to more potential corpora for this study. I gave up that access in favor of using the open access syntactic database developed by the Eep Talstra Centre for Bible and Computer (known as the ETCBC database). This database includes linguistic information encoded hierarchically at the word, phrase, clause, and text levels for the Hebrew Bible as well as 1QS, 1QM, and some inscriptions. I came to the conclusion that utilizing this data, even with its limited access to Qumran texts, was necessary for several reasons. When scholars utilize digital tools to conduct research, it is imperative that their work be replicable by other researchers. In order for data rich research projects to be replicable, the tools upon which those projects are based need to be open source so that all scholars can gain access. Since the popular Bible software programs are not open access, I decided to utilize a different database to collect the data for this project. Another requirement for reproducibility is clear documentation of how the data was collected. At present there is no standard documentation style for linguistic searches in programs like Accordance or Logos. And even if there were, search results will change as new versions of these programs are developed. These changes from version to version make the replication of research difficult if not impossible at times. I have therefore become convinced that linguistic studies of Hebrew utilizing large datasets should start by doing three things: first, the software used should be open source, second, the means by which the data is collected should be published, and third all data upon which the project is based needs to be readily available for scholarly examination. The ETCBC database allows for all of this while also having other significant advantages.

Through Shebanq and LAF-Fabric, the ETCBC database can be analyzed through powerful programming tools such as Python, Jupyter, and R. Python is a programming language that can be used to manipulate the linguistic data encoded in ETCBC. Jupyter is a software program that is used for interactive computing across programming languages. One of the main advantages to using Python and Jupyter is the ability to publish the code through which one gathers data. This code can then be checked by other researchers and the data can be reproduced exactly as in the original study. Another advantage of using these programs is their ability to integrate with the R software environment for statistical computing and graphics. R provides the researcher with access to powerful statistical tools to process and analyze large amounts of data such as hierarchical clustering, dendrograms, PAM clustering methods, and other tools that allow for hypothesis testing. The final digital tool that I have utilized for this project is Github which is a development platform for open source project management. All of the code utilized and data collected for this paper will be published shortly to Github and will be freely accessible to the public.

Now I will turn to the analysis of the data. As has already been mentioned, I collected data for 13 linguistic features, including: the אני אנכי interchange, *plene* and defective spelling of the negation לא, the direct object marker, the directive-*he*, the infinitive absolute, word order of the subject and predicate in independent clauses, the interchange of the particles אל and ל and אל and על, the infinitive plus the preposition ל, the interchange between מלכות and ממלכה, the relatives אשר and ש, *vav* plus imperfect, and *vayyiqtol*, known sometimes as the preterite or vav consecutive.

I begin my analysis of these data by focusing on their use in the core “EBH” and “LBH” texts as well as in 1QS and 1QM. The statistical tool that I apply first to these data is hierarchical clustering. Hierarchical clustering is a method of cluster analysis that starts with a set of individuals (in our case Ancient Hebrew books). Then, in a series of steps, this method combines the individuals into clusters. At each step, the individual clusters that are most similar based on some criteria (in our case linguistic characteristics) are joined together until one large cluster is reached. This process is often summarized by a dendrogram, which is a tree-like structure that illustrates the process taken in hierarchical clustering. There are four main versions of hierarchical clustering: average, single, complete, and ward linkage. I begin with the ward linkage method of hierarchical clustering simply by default. When the data for the 13 features I’m working with is analyzed through this method for the core texts, this is the resulting dendrogram. This dendrogram contains three main clusters. In the red cluster we find all the “EBH” texts as well as Esther. In the purple cluster, we find most of the “LBH” texts. 1QM and 1QS cluster together and as we can see they appear to be closely related to the “LBH” cluster. Ecclesiastes is also closely related to the “LBH” texts, but appears as an outlier, being on its own. Using the ward method of hierarchical clustering, we get a result that is consistent with the general consensus about the linguistic typology of these texts. The one main exception to that is the place of Esther. After examining other methods of hierarchical clustering, I will return to the issue of Esther.

The second method of hierarchical clustering that I will apply to this data is the “average” method, which produces this dendrogram. We see the same basic clustering structure in this dendrogram as was seen on the previous slide. Each method of hierarchical clustering produced three main clusters, with Ecclesiastes as an outlier. However, in this dendrogram, those clusters are organized a little differently. On the left, we see Ecclesiastes entirely on its own. Then 1QM and 1QS again cluster together. The large middle cluster contains all the “EBH” texts and finally, on the right, all the “LBH” texts group together, including Esther this time. The main differences between these two dendrograms is the place of Esther, and the relative similarity of Ecclesiastes, 1QM and 1QS. In the previous version, we found these three texts grouped more closely with the “LBH” cluster while in this version they have been group apart from the other texts. The remaining two hierarchical clustering methods produce variations close to these two options.

Now we must return to Esther. Scholars have long debated the linguistic nature of this book. Some have argued that Esther is clearly written in so-called “late” biblical Hebrew, while others have claimed that the author of this text intentionally utilized what has been called “early” biblical Hebrew. The two dendrograms presented here align each one with these two theories about the linguistic nature of Esther. Further statistical analysis can help parse out this challenge.

We can calculate the agglomerative coefficient in order to gauge the cluster strength of each dendrogram. The coefficient value can range from 0 to 1, with values closer to one suggesting strong clustering structure. A dendrogram with strong cluster structure suggests that the results will not change with the addition of more data (such as if we added more linguistic features or more texts to our analysis). Therefore, we will favor the method of hierarchical clustering that produces the highest cluster strength. This is vital at this stage in the present study because in the next stage, we will be adding in more texts. Thus, if we begin with strong cluster structure we reduce the amount of variation that we will see when we move from focusing upon the core texts to analyzing all of the texts in the target corpora.

On this slide, I present a comparison of two dendrograms and at the bottom the agglomerative coefficient for all four possible dendrograms. We see that the “average” method has the lowest coefficient. This is the method that grouped Esther with the other “LBH” books. The ward method, which grouped Esther with the “EBH” books, has the highest coefficient. Thus, we will move forward utilizing the ward method for our analysis with the knowledge that Esther appears to be linguistically somewhere between “EBH” and “LBH”.

Hierarchical clustering is known as an unsupervised clustering method. It is called that because no matter what data you feed through the hierarchical clustering algorithm, clusters will result. Even random data will be forced into clusters. Because of this, we must test the validity of the clusters that are produced. One way to do that is to analyze the data at hand utilizing a supervised clustering method for comparison. Supervised clustering methods require the researcher to pre-assign the number of clusters, then the algorithm places the data into those clusters as appropriate. For this study, I will utilize the Partitioning Around Medoids algorithm, also known by its acronym PAM. In order to utilize this algorithm, I must set the number of clusters into which the data will be grouped. Any number can be selected, but since hierarchical clustering resulted in four main clusters (Ecclesiastes, 1QM and 1QS, “LBH”, and “EBH”) I select four for the PAM analysis. This cluster plot is the result. Note here how the four clusters align with the results of the hierarchical clustering analysis. Also note how Esther groups with the “EBH” books, but appears to stand somewhat apart from the main group. We can also highlight how the Qumran texts continue to cluster together apart from the biblical texts.

Further statistical analysis can confirm these results, while helping us to visualize more thoroughly how these texts are linguistically related. The next tool that I will utilize is a silhouette plot of PAM as seen here. This silhouette plot assigns a number to each text from 0 to 1. The closer the number is to 1, the stronger that text is related to the other texts in its cluster. 1QS and 1QM have the highest silhouette width, suggesting that they are very closely related to one another. The most helpful aspect of this analysis is the result for Esther. Note how Esther is grouped with the “EBH” books here, but it has the lowest silhouette width. This suggests that it is in between clusters, confirming the result of our hierarchical clustering analysis.

Other methods could be utilized to statistically confirm the validity of the results already presented, such as the application of heat mapping as seen here on the left or multiscale bootstrap resampling as visualized on the right. I won’t continue to dwell on these statistical methods in detail here. Suffice it to say, that these tools confirm in their own way the validity of the results presented in the dendrogram and PAM plots.

Before moving onto the next step of this paper, I will pause to draw some basic conclusions thus far. First, based on the 13 linguistic features being analyzed, the “EBH”, “LBH”, and Qumran texts all group into their own clusters, with Ecclesiastes being an outlier. Esther, is the one text that appears to be between clusters, revealing it to have a mixed linguistic profile. We can now take a look at the individual features being analyzed to better understand the linguistic character of these four groups. First, it needs to be noted that these data are for the clusters, which is an average of the data for each text in each cluster. Thus, the internal variation of each cluster is set aside here, while the dendrograms and other cluster plots presented up to this point help to highlight that internal variation. With that said, we can conclude that generally speaking, the “EBH” cluster uses features traditionally considered to be early, the “LBH” cluster is characterized by so-called “late” features, and 1QM and 1QS also use “late” features, sometimes, but not always at the highest rate. The graph at the bottom of this slide highlights three example features. The first common plural interchange of אני and אנכי, reveals the “EBH” cluster using אני the least, the “LBH” texts using אני at a high rate, the Qumran texts in between those two, with Ecclesiastes only using the late feature of אני. The data for the directive-*he*, here labeled “hl” shows a reduction in the use of this particle from “EBH” to Qumran. We see a similar reduction in the use of the infinitive absolute, labeled here “infa”. With these observations in place, I will now move onto the next stage of this analysis.

By first focusing on the core texts of the Hebrew Bible as well as our two Qumran texts, we have seen that the traditional approach to these texts is supported by statistical analysis, except for the clustering of the certainly “late” biblical book of Esther with the core so-called early corpus. The next step is to see how other biblical books align with these core texts. First, I will add the Song of Songs to the analysis. Nearly all scholars agree that the Song of Songs is a post exilic text. The traditional approach to the language of this book is to label it as late. This dendrogram shows that the linguistic profile of Song of Songs causes it to cluster with the core “LBH” texts. This is what most scholars would have assumed would be the case. Contrary to this example, another post exilic text clusters with the “EBH” books as seen when Zechariah is added to the core texts dendrogram. Similarly, Hosea which is commonly dated to before the exile is placed within the “LBH” cluster. Ruth presents another challenging example, although there is no firm consensus about this book’s date. When Ruth is analyzed along with the the core texts it is place in the “EBH” cluster.

Another possible approach that I would like to explore is to divide up texts that are presumed to have been written over a large period of time. The book of Psalms is a good example of this. In this dendrogram, I have divided the psalms into three groups: pre exilic (Ps\_A), exilic (Ps\_B), and post exilic (Ps\_C) and we see the pre exilic Psalm’s group clustering with the “LBH” books while the “late” psalms are grouped with Ecclesiastes. I will return to this type of analysis in a moment.

Before drawing some general conclusions, I will analyze all the texts in our corpora at once, which produces this dendrogram. Here we find mixed results. In the “EBH” cluster we find five post exilic texts and two exilic. In the “LBH” cluster we find three pre exilic texts mixed in with texts commonly assigned to after the exile. Note that Obadiah is joined with the Qumran texts. This appears to be caused by the length of Obadiah. The 21 verses in this book just do not provide enough data to confidently analyze its linguistic character statistically. One final dendrogram displays those biblical texts that have been divided into sections along pre exilic, exilic, and post exilic lines: including Kings, Job, Proverbs, Isaiah, and Psalms. From this dendrogram we can see how these different sections are linguistically similar to and different from the other texts in our corpora. A couple things stand out. First, Job 1 and 2, typically thought of as a post exilic expansion, aligns with the core “EBH” books while Job 3-42, usually thought of as a pre exilic text, aligns with the “LBH” books. This grouping seems to suggest that the genre of these sections of Job is the main driving force behind where they cluster. Second, the three sections of Isaiah all cluster with the “LBH” texts, forcing us to conclude that Isaiah as a whole is linguistically consistent (at least when the 13 features analyzed for this study are taken together).

Before concluding this paper, I would like to focus in on the distinctive linguistic character of 1QS and 1QM. At this point it is important to note that I am in no way assuming that these texts are representative of Qumran Hebrew as a whole. Many scholars have noted the linguistic diversity within the Qumran corpus. But with that said, these two Qumran texts share very similar linguistic profiles, especially compared with the variety of Hebrew found in the biblical texts.

On this slide, I have summarized the data for all 13 features analyzed for this project. In the table, I have provided the individual data for 1QM and 1QS and then place an average of those two texts in the context of the “EBH” and “LBH” books as well as Ecclesiastes. I will highlight a few points regarding these data before turning to some final conclusions. First, as I have already noted, these Qumran texts favor “LBH” features. As an example, both 1QM and 1QS have the highest rate of preference for אל over על. Similarly, these texts have lower rates of the preterite per word than “EBH” or “LBH”. However, when the 1st common singular pronoun is considered, these Qumran texts fall between the “EBH” and “LBH” texts. The same is true for word order. “EBH” favors verb subject word order for main clauses, while “LBH” has the subject first 56% of the time. 1QM falls just below “LBH” with the subject first 54% of the time, while 1QS is closer to “EBH” with the subject first in only 39% of main clauses. These various linguistic features are enough to place 1QM and 1QS into their own cluster apart from the other texts analyzed in this study.

Now for some general conclusions. The core “EBH” and “LBH” books align with the traditional view of the typological classification of these texts. The one exception is the core “LBH” Esther aligning with the core “EBH” books. The results for the core texts may support the traditional diachronic understanding of Hebrew, which considers the different typology of the two corpora to be due to their chronological relationship. However, when all the biblical books are analyzed the issue is more complicated. If one attempted to linguistically date biblical books by comparing their linguistic profile to that of the “EBH” and “LBH” texts, then a number of challenges would arise. Books that are commonly believed to be post exilic, such as Malachi and Zechariah, are dated linguistically as pre exilic. Opposite examples appear as well, such as the early books of Habakkuk and Hosea being linguistically dated as late. If we are willing to accept that these books were written in a different era than what is traditionally held, we may conclude that linguistic dating of texts is valid. Alternatively, we could postulate that there existed at least two main styles of literary Hebrew before and after the exile, which would allow for dating some texts as late even though they have a linguistic profile similar to “EBH” books. This theory would make it difficult to date texts of unknown period, however, since alignment with either the “EBH” or “LBH” corpus would not be a clear sign of periodization. A third option would be that the linguistic profiles of the books are not authorial, but editorial, and that therefore an “early” book like Hosea evidences linguistic revision toward a later variety of Hebrew.

Yet, more work needs to be done before more solid conclusions can be developed. First, I intend to refine the data for the features I am currently working with. As an example, instead of focusing only on מלכות, I intend to expand my analysis to all abstract nouns suffixed with ות. Second, I want to expand my list of features analyzed to a total of 30. Statisticians generally argue that working with at least 30 features is ideal. Statistical analysis is still possible with less features, but 30 tends to be the goal, especially in Corpus Linguistic studies. Thus, I would like to add to my analysis features such as the shift away from the Qal to the Piel, loan words, and word order in the object clause. I am also interested in hearing any suggestions on what features you think might help deepen my analysis. Finally, I would of course like to expand my analysis of Qumran Hebrew beyond just 1QM and 1QS. However, this is outside of my control at present as we wait for the ETCBC team to continue their important work on making grammatical information on these texts freely available to all researchers.

Thank you