Linguistic Dating of Biblical Texts using Supervised Machine Learning

Thesis Submitted in Partial Fulfillment
of the Requirements
of the Jay and Jeanie Schottenstein Honors Program

Yeshiva College
Yeshiva University
May 2013

Toviah Y. Moldwin

Mentor: Moshe Koppel, Professor of Computer Science at Bar-Ilan University

Abstract: One of the major concerns in contemporary Bible scholarship is the problem of dating biblical texts. Our aim is to use an algorithmic approach known as supervised machine learning to utilize linguistic features of a biblical text to assign it either to the category of Early Biblical Hebrew or Late Biblical Hebrew, thereby informing us as to whether the text was written before or after the Babylonian exile. In addition to using this methodology to classify every book in the Bible, we perform a detailed case study on the book of Joel and demonstrate how our approach can be used to solve the long-debated question as to whether the book was written in the pre-exilic or post-exilic period. We also use the algorithm to individually classify each psalm in the book of Psalms , as most scholars believe that the book of Psalms was written over a long period of time by a variety of authors.

Part I: Introduction

A. Dating Biblical Texts: The Problem

When studying a book of the Bible, as when studying any piece of literature, one of the most important questions to ask even before opening the book is "what was the historical context in which this work was written?" This question is especially important when studying the Bible because so much of the Bible's meaning can only properly be understood in light of the background against which the biblical texts were authored. The Bible contains references to people, places, and events that cannot be fully appreciated without an accurate historical frame of reference.

Unfortunately for scholars of the Bible, determining the circumstances of authorship of any section of the Bible can be an incredibly difficult task. Even when a book of the Bible indicates the name of its author in its text, one is often left to question whether the book was written pseudepigraphically, by multiple authors, or if some other complex writing and editing process was necessary before the text arrived in the final form we see today. Not only is it challenging to pinpoint the exact author of a biblical text, but even the general time period in which a book of the Bible was written can elude even the most discerning of

readers. Nevertheless, scholars of the Bible have made painstaking efforts to assign a time frame for when each part of the Bible could have been written.

B. Methods of Dating: Content and Style

Barring external information, which we often don't have in the case of the Bible, there are two types of internal data that can be used to date a text: content and style. The content of a book of the Bible can tell us a good deal of information about the setting in which it was written. By identifying the characters, setting, and historical events contained within a book of the Bible, we can assign a lower bound to the date of its authorship. That is to say, if a book contains references to people, places, or events that we know from external sources happened at a particular time, we can assume that the book was written after that time[1]. Content usually can't tell us as much about the upper bound for when a biblical text was written, for a simple reason: later authors can—and often do—write about earlier time periods. Though we can make use of the fact that a book of the Bible makes no mention of events that occurred after a particular date, that knowledge is in no way a clear indication of an upper limit for the date of the book's authorship.

Style can be also very revealing about a text's authorship in subtle yet important ways. Authors have unique styles of writing which they often cannot change even if they want to. There are aspects of the way that a person writes which are endemic to the manner in which he expresses himself, as if they are ingrained in his psyche. Style has been convincingly used in a number of cases to demonstrate the authorship of anonymous texts (see, for example, Koppel 2006). Style is useful not only for finding the identity of a specific

[1] Assuming we rule out the possibility that those parts were added later.

author, but it can also be very helpful in determining a great deal of information about the document, including—as in our case—the time period in which the document was written.

Style, of course, is a difficult term to define. In the context of dating biblical texts, we are interested in stylistic features that give us information about the general time period in which the text was written. Primarily, this means looking carefully at linguistic features of the text to see if we can associate the words, grammar, and phraseology with a particular stage in the development of the Hebrew language. Linguistic dating, however, doesn't come without its own set of complications. We have very few external Hebrew texts that were written contemporaneously with the Bible, which means that most of the linguistic information that can be used to date biblical texts must come from within the Bible itself. The challenges that emerge when attempting to use linguistic features to date biblical texts have spurred a significant volume of discussion among Bible scholars.

C. Linguistic Dating of Biblical Texts: Early and Late Biblical Hebrew

Over the course of the deliberations about the use of the linguistic features to date biblical texts, one classification scheme has gained a great deal of traction. This scheme divides biblical Hebrew into two broad categories: Early Biblical Hebrew (also called Standard Biblical Hebrew) and Late Biblical Hebrew. Early Biblical Hebrew, according to this scheme, consists of the Hebrew that was used before the majority of the Israelites in the land of Israel were exiled to Babylon in the early part of the sixth century BCE. Late Biblical Hebrew, then, is the dialect of Hebrew that was spoken and written after this exile occurred. The historical presumption behind this scheme is that during the Babylonian exile, the Jewish people picked up words, phrases, and other linguistic affects from their foreign captors, thus creating an identifiable shift in the use of the Hebrew language. Working with this

assumption, scholars have tried to identify parts of the Bible as having been written before, during, or after the Babylonian exile.

One of the pioneers of the field of linguistically dating biblical texts is Avi Hurvitz, a professor emeritus at Hebrew University in Jerusalem. Hurvitz devised a set of guidelines to serve as a general guide for scholars in the proper use of linguistic data to classify a text as a post-exilic (that is, Late Biblical Hebrew) document (Hurvitz 2006).

The first guideline Hurvitz posits is that for a word to be classified as belonging to Late Biblical Hebrew, it must appear "exclusively or predominantly" in the books of the Bible that we already presume to have been written in the post-exilic period, such as Esther and Ezra—if a term appears regularly in earlier works, we can presume that it is simply part of Standard Biblical Hebrew.

Additionally, a term can only be definitively classified as belonging to Late Biblical Hebrew if it can be demonstrated that there is an equivalent term in the Hebrew of earlier biblical texts, but the text in question chose the unusual, "new" term instead of the standard one. If there is no semantic equivalent to the term in earlier parts of the Bible, it could easily be claimed that the word or phrase does indeed belong to "Standard Biblical Hebrew", but the Bible simply didn't have occasion to use the term in any text but the present one.

For a term to unambiguously belong to the class of Late Biblical Hebrew, Hurvitz also requires that it be independently attested to in Late Hebrew works, such as the Dead Sea scrolls, rabbinic literature, or Hebrew apocryphal books. This is necessary to demonstrate that the term is in fact a "late" term; not only do we have to show that the term *isn't* a Standard Biblical Hebrew Term, we also need know that it *is* a Late Hebrew term, which means it must appear in late sources.

Finally, Hurvitz only allows for a positive identification of a text within the biblical corpus as a post-exilic text if it contains a "heavy accumulation" of Late Hebrew terms; one or two neologisms don't constitute sufficient proof that a text was written at a late date.

One example Hurvitz frequently cites as an illustration of his method is the Hebrew name used for the city of Damascus (Hurvitz 2006). The Bible uses two different names for Damascus: דמשק and דרמשק. Though Damascus is mentioned 34 times in the Bible, the term דרמשק is used only on six occasions, all of which are found in the book of Chronicles. Thus, the term does not appear in any text presumed to have been written at an early date; in every early text, the semantically identical but phonologically different דמשק is used. Moreover, דרמשקappears in a variety of later Hebrew works—including the Dead Sea Scrolls[2], the Genesis Apocryphon[3], and the Mishnah[4]—confirming that the term was indeed used in later periods of the development of Hebrew. Hurvitz also notes that early texts in Egyptian, Akkadian, and Aramaic use only דמשק to refer to Damascus. According to Hurvitz, this information makes a compelling case that דרמשק is a Late Biblical Hebrew word, and the word can thus be used, in conjunction with other evidence, to verify the late authorship of Chronicles.

Hurvitz's methodology is not without its detractors. A number of scholars are more skeptical than Hurvitz about the extent to which linguistic features can be accurately used to date a text. Young, for example, argues that linguistic data, while useful, cannot be used to conclusively date biblical texts (Young 2006). Young disputes the premise that the words and forms ascribed to Late Biblical Hebrew only belong to the post-exilic era. Instead, Late

[2] In 1QIsa[a], the Qumran Isaiah scroll, the word דמשק, when it appears in the MT, is changed to דרמשק, "updating" the word to a later form.
[3] In its citation of Genesis 14:15.
[4] Yadayim 4:3.

Biblical Hebrew could have existed as a stratum of Hebrew before the exile; there is no reason to say that it was the Babylonian exile itself that was the sole source of the linguistic shift associated with Late Biblical Hebrew. Young also points out that Standard Biblical Hebrew seems to have been used in biblical texts that were unquestionably written in the post-exilic period, such as Haggai and Zechariah 1–8. Young suggests that this might have been done intentionally, because Standard Biblical Hebrew may have been seen as a more appropriate register in which to compose literary/religious works.

In addition to claiming that Standard Biblical Hebrew and Late Biblical Hebrew may have been used concurrently, Young argues that scribal emendations may be responsible for the differences in language between various books of the Bible. As evidence for this contention, Young points to the Dead Sea Scrolls, which have versions of a number of books of the Bible that differ from the Masoretic Text. One of these texts is 1QIsa^{a}, the Isaiah Scroll, which differs linguistically from the MT version of Isaiah more frequently than Kings (a presumed Standard Biblical Hebrew work) differs from Chronicles (a presumed Late Biblical Hebrew text) when Kings and Chronicles are discussing the same subject. Thus, if a scribal process occurred to the MT that was similar to what occurred with 1QIsa^{a}, we would have no way of knowing whether a Late Biblical Hebrew term was part of the original text or if it was added by a scribe.

A slightly different approach to linguistic dating is proposed by Joosten, which circumvents some of the issues raised by Young. Joosten notes that in addition to *lexical* differences between the language of pre-exilic and post exilic texts, there are also grammatical and syntactical differences which are independent of vocabulary. Joosten points out that even if post-exilic writers intentionally used Early Hebrew vocabulary to give their

works a more religiously authoritative tone, it is unlikely that these writers had a mastery over the subtle nuances of the syntactic usage and grammar of Early Biblical Hebrew to the extent that they would have attempted to reproduce them (Joosten 2005). A later editor would also be less likely to "update" an Early Biblical Hebrew work in order to make a minute change in syntax.

One of the syntactic features that Joosten points to as an example of a grammatical distinction between Early Biblical Hebrew and Late Biblical Hebrew is the use of a w^e before a second person-prefixed verb form. In Early Biblical Hebrew (which consists of Genesis–II Kings, according to Joosten) there are only two occasions in which a w^e is used prior to a second person verb. In contrast, Late Hebrew texts (consisting of Ecclesiastes, Esther, Daniel, Ezra, Nehemiah, and Chronicles) uses a second person prefixed verb preceded by a w^e no fewer than nine times, even though the corpus of Late Biblical Hebrew works is considerably smaller than that of Early Biblical Hebrew (Joosten 2005).

To summarize, Hurvitz and other proponents of linguistic dating of biblical texts feel that by identifying neologisms—that is, words that seem to only appear in late sources—we can assign parts of the Bible to the either the pre- or post-exilic period. There are others, however, who feel that this approach toward dating texts is flawed, or at least not always compelling, for two reasons. First, "Standard Biblical Hebrew" is not necessarily a distinct dialect from "Late Biblical Hebrew" (and even if it is, "Late Biblical Hebrew" may not necessarily be a post-exilic dialect). Second, the appearance of Late Biblical Hebrew in biblical texts could be due to scribal emendations and thus tells us nothing about the authorship date of the original text. Non vocabulary-related syntactical features, such as the grammatical features proposed by Joosten, might be valuable to help avoid some of the

pitfalls that come along with using lexical features to date biblical texts. Nevertheless, the criticisms of the traditional method of linguistically dating texts have made it difficult to use linguistic dating to assign an approximate date of authorship to any part of the Bible with certainty.

Part II: A Different Approach: Supervised Machine Learning

Given the problems associated with the standard approach to dating biblical texts, we propose the use of a different process, known as supervised machine learning. In brief, "supervised machine learning" refers to a category of computer algorithms which "learn" from data which is given to them in order to answer questions about new data. One of the major applications of machine learning algorithms is text categorization—in other words, classifying documents into different categories by examining various aspects of their content. With a few simple specifications, supervised machine learning algorithms for text categorization can be brought to bear on the problem of dating biblical texts.

A. Introduction to Supervised Machine Learning in the context of Authorship Attribution

To demonstrate how a supervised machine learning algorithm would operate for a text categorization problem, consider a simple classification program that classifies articles as either having been written by author A or by author B. This type of problem is called an "authorship attribution problem", where we are looking to identify the author of a particular text.

To "train" the algorithm, we first provide it a set of documents which contain both articles written by A and articles written by B, each of which are identified as such to the algorithm. The documents of each set are then turned into vectors ("vector" in this case

means a list that can be represented as a set of data points), of features and their respective values. A feature is any quantifiable aspect of a document, like a word or a phrase. The value of a feature is dependent on the frequency with which it appears in a document. For example, if document X is 100 words long and uses the word "and" five times, "the" three times, and "but" zero times, we would create a vector representing document X that looks like this: [and: 0.05, the: 0.03, but: 0.00] (each value has been divided by the document length to give us a normalized frequency). It is also possible to use binary values to simply determine whether a feature appears or does not appear in the document. In our case, if we were to use binary values, our vector would look like this: [and: 1, the: 1, but: 0].

Generally, when performing supervised learning experiments, we don't want to use every single feature in a document. Rather, we choose those features which are most likely to be accurate predictors of which category an unclassified document belongs to. In the case of an authorship attribution problem, we are interested in using features that will help us identify the unique style of an author. As such, we will generally want to focus on "function words"—words like "and", "the", and "but", which are content-independent. The frequency with which an author uses particular function words is often an endemic part of the author's style and can thus help us classify a document of unknown authorship. In addition to function words, other commonly used feature types include character n-grams (a set of *n* consecutive letters, like "wh*"* [a 2-gram] or "les*"* [a 3-gram]) and word n-grams (a set of *n* consecutive words, such as "just like" or "matter of fact").

Once we have created feature vectors from the training documents and selected the features we would like to use, we are now ready to use the data from the vectors to construct a model that will allow us to classify the document. There are a variety of algorithmic

techniques that can be used to create such a function. As a very basic example of a classification algorithm, we consider the Euclidean distance equation. Euclidean distance is a formula that calculates the length of a line drawn between two points in a Cartesian coordinate system. If we have two points, point a and point b, where (a_1, $a_{2,}\ldots a_n$) are the coordinates of point a and (b_1, $b_2\ldots b_n$) are the coordinates of point b, the distance (d) between the two points is defined as $d = \sqrt{\sum_{i=1}^{n}(b_i - a_i)^2}$. When a and b lie on a plane, this formula is equivalent to the Pythagorean Theorem.

To illustrate how we would use Euclidean distance formula in the example that we have been using thus far, let us consider the following scenario. We are interested in classifying document Y, whose word frequency vector is [and: 0.012, the: 0.023, but: 0.014]. From the training data, we have gathered that the word frequency vector for author A is [and: 0.02, the: 0.012, but: 0.05] and for author B is [and: 0.017, the: 0.02, but: 0.036]. Each of these vectors can easily be thought of as a point in 3-dimensional space: Author A is the point (0.02, 0.012, 0.05), Author B is the point (0.017, 0.02, 0.036), and Y, our query text, is the point (0.012, 0.023, 0.014).

To classify document Y, we want to compare the distance between the point that represents document Y to the point that represents author A with the distance between the point that represents document Y and the point that represents author B. To find the distance between document Y and author A, we use the Euclidean distance formula:

$$d = \sqrt{(0.02 - 0.012)^2 + (0.012 - 0.023)^2 + (0.05 - 0.014)^2} = 0.038484.$$

We then do the same thing for the distance between document Y and author B:

$$d = \sqrt{(0.017 - 0.012)^2 + (0.02 - 0.023)^2 + (0.036 - 0.014)^2} = 0.02276.$$

The distance between document Y and author B is smaller than the distance between document Y and author A, thus leading us to conclude that there is a higher likelihood that document Y was written by author B.

It should be noted that there are many other ways to create a classification function other than using pure Euclidean distance. Euclidean distance has been used here as a simple example of a classification function, but in practice, the most accurate types of classification functions are usually a bit more complex.

B. Testing the Efficacy of a Classification Model

Now that we have presented the basic structure of how supervised machine learning is used for authorship attribution queries, we demonstrate how one would verify the accuracy of a particular classification model. To this end, we consider an experiment performed by Koppel (Koppel et al 2006). This experiment involved classifying rabbinic responsa as having written by one of two medieval Spanish rabbis: Rabbi Shlomo Ben Aderet (also known as Rashba, lived 1235-1310) or by his student, Rabbi Yom Tov ibn Asvilli (also known as Ritba, lived 1250-1330).

To perform this experiment, 209 responsa were collected from each author and turned into feature vectors. 304 features were selected to be "function words" in the context of these documents, and these features were then used to construct to construct a model to classify a document as having been written by one of the two authors. A technique called Balanced Winnow (Littlestone 1988) was used to construct the model for this experiment.

To verify the accuracy of using Balanced Winnow with these function words as features, a method called k-fold cross-verification[5] was used. This approach entails dividing up the corpus of training documents into *k* parts, then using *k-1* of those parts to classify the remaining part. This process is iterated *k* times, each time using a different part of the corpus as the query set. The efficacy of the classification model in classifying each part of the training corpus in this fashion gives us a good idea as to how well the model will perform on unseen query texts.

In this experiment of classifying responsa written by Rashba and Ritba, Balanced Winnow performed with an accuracy of 85.8% when using 5-fold cross-verification. The accuracy was improved to 90.5% by removing features from the classification model that were not good discriminants between the two authors. The high level of accuracy that was achieved after performing the five-fold cross-verification indicates that the Balanced Winnow method, using these features, is an excellent algorithm for determining whether a document was written by Rashba or Ritba.

This technique can be applied to many classification problems to verify the accuracy of a classification model. K-fold cross-verification is also an easy method of finding the most useful feature types and classification algorithms for solving a particular categorization problem. If a particular feature type performs better than another feature type (e.g. if a function word model performs better than a letter n-gram model), we can make the claim that that feature type is a more accurate discriminant between the categories.

C. Applying Machine Learning Techniques to Linguistic Classification of Biblical Texts: Methodological Considerations

[5] Any number of folds can be used; 5 folds are seen as sufficient in many cases to attain a good predictor of a model's accuracy.

We now turn to the question of why and how supervised machine learning can be used to aid us in our task of classifying biblical texts as Early or Late Biblical Hebrew. Though we are attempting to classify the documents by date and not by author, the basic principle is the same: just like we use features of a document to tell us about the identity of a specific author, we can similarly use the same features to tell us whether a document was written in the era of Early Biblical Hebrew or Late Biblical Hebrew by analyzing the features of documents we know to have been written in each of those periods.

In a general sense, we can design a format of how we might apply supervised machine learning to solve a query about the dating of a part of the Bible. We would construct a training corpus consisting of one set of biblical texts we know to have been written in the pre-exilic era and one set of biblical texts we presume to have been written in the post-exilic era. We would then turn these sets into vectors of features, create a classification model, and use that model to classify the text that we want to date.

Using supervised machine learning to tackle this problem has several advantages over doing linguistic analysis by hand. The primary benefit lies in the fact that this technique allows us to do linguistic dating in a more holistic fashion. Instead of finding one or two words in a document that lead us to believe that the document contains Early or Late Biblical Hebrew, supervised machine learning allows us to look at the entire document and compare it to large sets of texts from the biblical corpus. Virtually every single feature of our query text can play a role in its classification, and each feature in the training sets can likewise be used to provide insight into the difference between Early and Late Biblical Hebrew.

Though the process seems simple enough, it is complicated by a number of factors unique to our problem. First of all, to even consider using supervised machine learning for this application, we have to work with initial premise that Standard Biblical Hebrew and Late Biblical Hebrew are, in fact, distinct strata in the development of the Hebrew language by which we can group together two sets of texts. If this is not the case, our classification is meaningless; the fact that a query text is classified closer to one set of documents to another tells us nothing about the dating of the text. We can presume, however, that even if the difference between language in pre-exilic and post-exilic Hebrew are minute, they nevertheless exist, and that should be sufficient to allow us to use a linguistic comparison of documents of unknown dating to documents of known dating for the purposes of classification. The k-fold cross-verification process can also assist us in determining the extent to which the two sets of training texts are distinguishable from each other.

There are also a number of complications that arise when choosing the texts to create the training sets. Usually when performing classification experiments, we try to use large sets of training texts to ensure that we have an accurate representation of each class. In our case, however, the Bible contains only a few documents that we know with a high degree of certainty belong to either the pre or post-exilic period. As such, we have to be careful that the results aren't being skewed because of the unique features of one text in the training set that aren't necessarily representative of the whole class. To avoid this problem, it is necessary to ensure that we include a large number of texts in each of the training sets so as to ensure that our training sets are indeed representative of Late and Early Biblical Hebrew, respectively.

It is also important to account for the possibility that even if we do find that there are features that can serve as discriminants between the two classes of documents, the difference

between the two classes is not due to a discrepancy in their date of composition but rather because of some other reason, like a difference in regional dialect. There is no algorithmic way to discount this possibility, but we must use our best judgment to determine whether a regional or a chronological factor is the source of a particular set of linguistic differences. Again, using a diverse set of training texts for each of the classes of Biblical Hebrew will help ensure that the common denominator between the texts in each category is their date of composition and not any other factor.

D. Creating a Model for Classifying Biblical Hebrew: Word Frequency

With the guidelines discussed in the previous section in mind, we now turn to creating a model for classifying Early and Late Biblical Hebrew. The first step in this process is to choose the texts for each training set. This task is not a simple one, as it is necessary to find parts of the Bible about which there is a consensus about whether it was written in the pre-exilic or post-exilic period.

Strictly based on content, there are some books of the Bible that we can easily date to the post-exilic period. Esther, Daniel, Ezra, and Nechemiah, and Chronicles all clearly contain descriptions of events occurred after the Babylonian exile. Though Chronicles in places seems to copy directly from earlier parts of the Bible, we work under the presumption that the language of Chronicles is primarily representative of Late Biblical Hebrew. We also include in this category the book of Ecclesiastes, despite the fact that Jewish tradition attributes the authorship of the book to King Solomon, as the consensus of scholarship about Ecclesiastes is that it was written in the post-exilic period, based on the presence of Persian loanwords and other factors (See Longman 1998).

Finding training texts for Early Biblical Hebrew is more difficult, as we cannot use the fact that a book only contains references to pre-exilic events as proof that it was written in the pre-exilic period. However, the consensus of Bible scholarship would indicate that we can presume that the following books are pre-exilic: Joshua, Judges, Samuel I and II, and Kings I and II. The book of Jonah is also considered by many scholars to be a pre-exilic work, in part because of its focus on the city of Nineveh, which had been destroyed by Nebuchadnezzar at around the time of the Jewish exile to Babylon. In theory, we could also include the five books of Moses in this class as well. However, because the date of authorship (as well as the unity of the text) of the Pentateuch is disputed, with some scholars claiming that the final redaction of the book occurred around the time of Ezra, we will leave it out for the time being. Later on, we show that our classification approach puts all of the five books of Moses in the category of Early Biblical Hebrew.

As of now, our training sets consist of the following:

Early Biblical Hebrew	Late Biblical Hebrew
Joshua	Daniel[6]
Judges	Esther
Samuel I and II	Ezra[7]
Kings I and II	Nehemiah
Jonah	Chronicles
	Ecclesiastes

[6] Though the books of Daniel and Ezra contain significant sections in Aramaic, we leave those sections in our training texts because the presence of Aramaic (especially the dialect of Aramaic that was spoken in the time of Ezra and Daniel) is itself indicative of a document belonging to the post-exilic period, even though it isn't part of Late Biblical "Hebrew" per se. In the vast majority of cases, the Aramaic features play no role in the classification of a document, because Aramaic isn't used very often in the Bible other than in Daniel and Ezra.

[7] See note 5

Our next task is to choose the types of features we will use to build our model. For now, we will focus on word frequency—the number of times that particular words appear in a text. To perform our word frequency experiment, we find "common" words in our corpus of training texts, which we define here as words that appear 10 or more times in the corpus. We are left with a vector of 1,563 common words. Using these words, we employ a Bayesian multinomial regression algorithm (Madigan 2005) to create a model to classify documents.

To see how well the model performs in classifying texts, we divide the corpus of biblical documents into chunks of 500 words each (for a total of 251 documents) and perform ten-fold cross-verification[8]. Ten-fold cross-verification yields a result of around 94% accuracy[9]. In other words, 235 of the 251 documents were classified correctly.

When judging the efficacy of a classification function, it is important to take note of the function's recall and precision. Recall measures the proportion of documents in each class that the algorithm classifies correctly, and it is defined as the number of correct assignments to a particular class divided by the total number of documents in the class. In our case, out of the 152 documents belonging to the Early Biblical Hebrew category, 145 of them were classified correctly, yielding a recall of 95%. Our algorithm did almost as well for Late Biblical Hebrew, correctly classifying 90 documents out of a total of 99 documents belonging to Late Biblical Hebrew, for a recall of 91%.

Precision measures the extent to which the algorithm returns a correct answer as opposed to an incorrect answer when it returns a particular result, and it is defined as the number of correct assignments to each category divided by the number of total assignments

[8] We use ten folds instead of five to get a more precise estimation of the model's accuracy.
[9] All percentages are rounded to the nearest whole number. These numbers can vary depending on how the corpus is divided into its ten folds.

made by the algorithm to that category. Out of a total of 154 documents that our algorithm classified as belonging to Early Biblical Hebrew, 145 of those were correctly assigned, yielding a precision of 94%. For Late Biblical Hebrew, out of 97 documents that were classified as Late Biblical Hebrew, 90 were correctly classified, giving us a precision of 92%.

What emerges from the results of the ten-fold cross-verification is that not only is it possible to distinguish between Early and Late Biblical Hebrew, we can do so a high level of accuracy simply by examining the frequency with which common words appear in a text, even when the query texts are relatively small (only 500 words in length).

E. Classification Models using other Feature Types: Word Bigrams and Morphology

Despite the high accuracy that was achieved in the ten-fold cross-verification process simply by employing word count, we would like to create models using other feature types to independently verify the results of the word count classification. If we achieve the same classification results using different types of features, we can be even more confident that our classification is correct.

Using the same training set, we create a model using word bigrams, that is, two-word Hebrew phrases. We use bigrams that appear over four times in the training corpus to ensure that the phrases are meaningful as a unit (as opposed to every permutation of two consecutive words). This gives us a set of 1,955 meaningful two-word phrases. We once again construct a classification model using Bayesian multinomial regression. We test the model using ten-fold cross-verification, which yields an accuracy of 89% (224 out of 251 documents were classified correctly). The precision and recall of this model are noted in the table at the end of this section.

We would also like to construct a model using morphological features – in other words, the frequency with which certain grammatical forms appear within the Bible. To create our morphology model, we utilize the Groves-Wheeler Westminster Morphological Database, which encodes each words of the Bible as string of letters representing its grammatical form. For example, the Westminster encoding for the second word in Genesis, בָּרָא (he created), would be @vqp3ms, which means "verb, qal, perfect, third person, masculine singular". We make our corpus of training sets with the Westminster Bible encodings just as we did for the plain text of the Bible, dividing the texts belonging to each training set into chunks of 500 words each. We then represent the text using vectors of unigrams, each of which represents a grammatical encoding of morphological form (the Westminster data also contains encodings for certain non-word items, like different types of paragraph breaks). The training texts contained 260 unique morphological encodings.

We create a model from the morphological features using Bayesian multinomial regression. Just as we did for word and bigram frequency, we perform ten-fold cross-verification to verify the accuracy of the morphology model and achieve an accuracy of around 84%. The accuracy, precision, and recall of each of our models are listed in the chart below.

	Accuracy	Precision Early Hebrew	Recall Early Hebrew	Precision Late Hebrew	Recall Late Hebrew
Word Unigram Frequency	94%	94%	95%	93%	91%
Word Bigram Frequency	89%	89%	89%	84%	83%

Morphology Unigram Frequency	84%	86%	89%	0.81%	75%

F. Most Valuable Features for Distinguishing Between Early and Late Biblical Hebrew

In text categorization experiments, some features end up being more revealing about a document's authorship than others. We present here a selection of the most significant features from each feature type (word frequency, bigram frequency, and morphology) as determined by T-test value[10]. These lists give us an insight into how the use of certain words changed from the pre-exilic to the post-exilic period.

1. Word Frequency

The following words, which had the highest T-test score in our word frequency model, are all significantly more common in Early Biblical Hebrew than in Late Biblical Hebrew.

Feature	Count	Freq_EarlyHebrew	Freq_LateHebrew
אֶל	1,464	0.015973	0.006096
וַיֹּאמֶר	1,157	0.013334	0.003756
אִישׁ	560	0.006308	0.002057
דָּוִד	451	0.006166	0.00002
הִנֵּה	161	0.001913	0.000444

[10] A T-test is a test of statistical significance to determine the probability that two samples come from different populations. In our case, it measures the likelihood that the presence of a feature is indicative of its belonging to a particular class.

We also present some of the words with a high T-test value that are more common in Late Biblical Hebrew than Early Biblical Hebrew:

Feature	Count	Freq_EarlyHebrew	Freq_LateHebrew
דָּוִיד	207	0.000028	0.004343
וְהַלְוִיִם	53	0.000027	0.001077
הַלְוִיִם	88	0.000137	0.001644
דָוִיד [11]	35	0	0.000744
לְבֵית	86	0.000302	0.001355

Just by examining these features, we can learn about the development of the Hebrew language in the biblical period. One obvious observation is the difference in spellings of the name of King David: Early Biblical Hebrew uses the defective spelling, without the letter *yod*, whereas Late Biblical Hebrew includes it. This is a manifestation of a phenomenon, noted by many scholars, that later Hebrew works prefer to use the plene spelling, which serves as an aid to reading and pronunciation (see Rooker 1994).

It would also seem that certain function words dropped out of use as the Hebrew language progressed. This is a relatively common phenomenon in the development of languages; words which aren't especially meaningful (such as הִנֵּה, which is generally translated as "behold") or words which can be replaced by contractions or ignored (such as אֶל , which can be replaced by a –ל prefix, or sometimes dropped entirely) tend to fall out of favor as a language is used over a period of hundreds of years.

[11] David's name with a plene spelling but without a *dagesh*.

There is also content-related information that can be derived from the word frequencies. For example, the frequency with which the Levites are mentioned in Late Biblical Hebrew works can be seen as indicative of the prominence of the priesthood during the post-exilic phase of the Biblical period.

2. Word Bigrams

The phrases below have the highest scoring T-test phrases for Biblical Hebrew two-word phrases, and they all appear more frequently in Early Biblical Hebrew than Late Biblical Hebrew.

Feature	Count	Freq_EarlyHebrew	Freq_LateHebrew
אֶל_דָּוִד	59	0.004059	0
וַיֹּאמֶר_דָּוִד	59	0.004052	0
הַיּוֹם_הַזֶּה	42	0.002595	0.000419
אֲשֶׁר_אִתּוֹ	31	0.002108	0
בְּנֵי_יִשְׂרָאֵל	97	0.005791	0.001269
חַי_יְקֹוָק	31	0.002039	0.000105

The top two phrases both contain the defective spelling of the name of King David, which itself is an indicator, as we saw earlier, of Early Biblical Hebrew. This highlights the fact that word bigram frequency and word frequency aren't entirely independent features; sometimes a word bigram will be an indicator of one class or another precisely because it contains a single word which is more indicative of that class.

It is also evident why the phrase בְּנֵי_יִשְׂרָאֵל is more indicative of Early Hebrew, as "children of Israel" is a term that would seem to refer to the unified nation comprised of all the tribes of Israel, which effectively ceased to exist subsequent to the split between the Northern and Southern kingdoms of Israel. Some of the other phrases, such as הַיּוֹם_הַזֶּה or אֲשֶׁר_אִתּוֹ don't have obvious reasons why they should be more indicative of Early Hebrew than Late Hebrew; instead, it would seem that these phrases are less common in Late Biblical Hebrew simply because they fell into disuse, for whatever reason.

We also examine the highest-scoring bigrams from Late Biblical Hebrew:

Feature	Count	Freq_EarlyHebrew	Freq_LateHebrew
בֵּית_הָאֱלֹהִים	28	0.000067	0.002845
עַל_כָּל	64	0.001435	0.004543
מִן_בְּנֵי	24	0	0.002532
הַכֹּהֲנִים_וְהַלְוִיִּם	20	0	0.002108
רָאשֵׁי_הָאָבוֹת	17	0	0.001813

We again note the prominence of the priestly class in Late Biblical Hebrew documents. The phrase בית_הָאֱלֹהִים as an appellation for the Temple also seems to be a phenomenon endemic to Late Biblical Hebrew.

3. Morphology

The following table contains the some of the highest T-test scorings for morphology unigrams that are more frequent in Early Hebrew than Late Hebrew.

Feature	Count	Freq_Early	Freq_Late

@vqw3ms (verb, qal, waw consecutive, third person, masculine, singular). [e.g. וַיֹּאמֶר, Gen 1:3]	4,320	0.03389	0.014175
@Pi (pronoun, independent). [e.g. , הִנֵּה Gen 1:29]	619	0.004939	0.001737
@Pg (Pronoun, interrogative) [e.g. אַיֶּכָּה, Gen 3:9]	311	0.002485	0.000931
@vqi1cs (verb, qal, imperfect, first person, gender neutral, singular) [e.g. אָשִׁית, Gen 3:15]	284	0.002319	0.000753
@vqw3mp (verb, qal, third person, masculine plural) [e.g. וַיִּהְיוּ , Gen 2:25]	1,506	0.011086	0.006237

We also have the following features with high T-test scores which are more frequent in Late Biblical Hebrew:

Feature	Count	Freq_Early	Freq_Late

@ncmpa (noun, common, masculine, plural, absolute) [e.g. אֱלֹקִים, Gen 3:9]	3,165	0.016043	0.024514
@ncmpc (noun, common, masculine, plural, construct) [e.g. חַיֵּיךְ, Gen 3:14]	4,544	0.023167	0.037029
@vhPmpa (verb, hiphil, participle, masculine, plural, absolute) [e.g. מַשְׁחִתִים, Gen 19:13]	96	0.000243	0.001158
@vpc (verb, piel, construct) [e.g. לְדַבֵּר , Gen 17:22]	284	0.001269	0.002494
@x (paragraph break) [e.g. after Gen 1:19]	1,393	0.006918	0.011974

It is not immediately apparent why some of these features should be more indicative of one class of Hebrew than another, though a few observations can be made. In some cases, the frequency with which a particular morphology appears is strongly tied to the frequency with which particular words appear (such as independent pronouns and the word הִנֵּה, both of which are highly scoring features for Early Biblical Hebrew).

The fact that @x -- the presence of a paragraph break (both *petuchot* and *setumot)*-- comes out as a highly significant feature for indicating Late Biblical Hebrew is a fascinating result, though it is difficult to explain why paragraph breaks would be more frequent in Late

Hebrew than Early Hebrew[12]. This may be due to the nature of the specific content of the books that were used for our training sets (though that is, of course, a possibility for every feature that is used) but it is also possible that this represents a real shift in stylistic convention in terms of how frequently paragraph breaks were utilized.

Part IV: Case Study: The Book Of Joel

As an illustration of how our models can be used to date a part of the Bible, we consider the Book of Joel. The book of Joel is chosen because of the wide range of positions held by scholars about the date of the book's authorship, as discussed below. Our aim is to demonstrate that supervised machine learning algorithms can be used to convincingly place Joel within the post-exilic era.

A. The Book of Joel and the Controversy Surrounding its Date of Authorship

The book of Joel is one of the twelve books known as "The Twelve Minor Prophets". Joel is a short book, consisting of only three chapters[13]. Scholars have divided the book's content into two thematic sections. The first section, which extends from the beginning of the book until verse 2:17, describes a plague of locusts and an exhortation by the prophet Joel encouraging the Israelites to repent in the face of this calamity. Though many scholars take the description of the locust plague literally, some consider the possibility that the locust plague may in fact be a metaphor for a military invasion (see Andiñach 1992). The tone of

[12] There is no consensus as to when paragraph breaks were introduced into the Bible, though they seem to date back at least to the time of the Dead Sea Scrolls, though the paragraph breaks in the Dead Sea Scrolls (as well as the Samaritan Pentateuch) are different than the ones found in the Masoretic text (See Weigold 2009).

[13] It should be noted here that the chapter divisions in most Hebrew Bibles do not match up to the chapter divisions in Christian Bibles. Hebrew Bibles generally have four chapters of Joel, while Christian English Bibles have three. In Hebrew Bibles, the third chapter of Joel consists of five verses which Christian Bibles have as the last five verses of the second chapter. Thus, the second chapter of the Christian Bible's book of Joel is 32 verses long, while the second Chapter of the Hebrew Bible only has 27 verses. This paper uses the Christian chapter divisions to conform to the rest of the scholarship on the book of Joel.

the book changes in verse 2:18, when God takes pity on His people and blesses them with abundant produce. This section also contains an eschatological prophecy[14] that describes various blessings that were to be bestowed upon the people of Israel and promises of retribution against Israel's enemies among the nations of the world.

There is little explicit information in the Book of Joel that gives a clear indication as to when the text was written. Scholars have nevertheless come to a wide variety of conclusions about the dating of the book based on various subtle clues within the text. The opinions for the date of the book of Joel range from the time of King Joash (836-797 BCE) as late as 350 BCE well within the Second Temple Period) (Jewish Encyclopedia 1906; Stephenson 1969). Jewish tradition and medieval Jewish Bible commentaries also posit a number of possibilities for when the prophet Joel might have lived.

The Midrash[15] identifies Pethuel, Joel's father, with the prophet Samuel. Other than exegetical wordplay, the Midrash gives no textual support for associating Joel with the time of Samuel. Kimhi[16] suggests that Joel lived in the time of King Jehoram of Judah (849-842 B.C.) as the Bible records a famine during that time, which accords with Joel's prophecy about a famine. Kimchi also cites Seder Olam as placing Joel during the time of the rule of Menasheh (7th century BCE). The Talmud[17] seems to favor a late date of authorship, stating that the books of the twelve Minor Prophets were written by the members of the Great Assembly[18], allowing for a relatively wide range of post-exilic dates. (It should be noted that the Talmud is only commenting on when the books of the twelve Minor Prophets were

[14] Treves (1957) disagrees with this characterization, arguing instead that this is simply a general description of God rewarding his people and punishing their enemies at a later date, which need not occur during the "end of days."
[15] Bamidbar Rabbah 10, s.v. "הדבר אל"..
[16] Radaq to Joel 1:1
[17] Baba Batra 15a
[18] This was a religiously authoritative body that existed from the end of the time of the prophets to the middle of the Second Temple Period.

written, not when the prophets actually lived and delivered their prophecies. Presumably, the Talmud means to claim that the prophecies of the Minor Prophets were transmitted orally until they were finally written down by the men of the Great Assembly.)

In the 19th century, many scholars supported the date of the monarchy of Joash for the authorship of the book of Joel, which would make Joel the first Jewish prophet to write a book of prophecies. Though no mention of Joash is made in the book, scholars have suggested that the events in the book occurred during the time that Joash was a minor, thus making the Joash himself an unimportant player in the prophecy. The events of Joel having taken place when Joash was a minor would explain why Joel directs his prophecies to the priests as opposed to the king (Allen 1976). One flaw with this dating, as some scholars have pointed out, is the fact that Judah is called "Israel" in a number of places in the book[19] which would indicate that the book was written after the Northern Kingdom of Israel ceased to exist, namely after 721 BCE.

Some have suggested the possibility that the book was written in the pre-exilic period, but slightly later, during the time of King Josiah. Among the reasons given for this dating is that the book of Jeremiah, which takes place during the time of King Josiah, mentions a famine that would concur with the description of the famine mentioned in Joel[20]. Moreover, the mention of Egyptians in the prophecy of Joel (3:19) might be a reference to Josiah's military campaign against the Egyptians. The fact that neither the Assyrians nor the Babylonians are mentioned in Joel somewhat undermines this dating, however, because all of the other pre-exilic prophets who prophesize about this time period talk about the impending conquests of Israel by the Assyrians and the Babylonians (Jewish Encyclopedia 1906).

---

[19] Joel 3:27, 3:2, 3:16.
[20] See Jeremiah 16:2-6

A few other suggestions have been proposed for late pre-exilic dates. A.S. Kapelrud, for example, dated the book to the time of Zedekiah (c. 600 BCE), because of the common references to Philistines in Joel 4:4 and in Zephaniah 1:24–18 and Jeremiah 47:4. In Jeremiah, the Philistines are mentioned together with Tzor and Tzidon, just like they are in Joel, suggesting a possible association between the time periods of the texts (Allen 1976).

Many other scholars favor a post-exilic date for the Book of Joel. Jacob Myers suggests that Joel was written around the time of Haggai and Zechariah, at the end of the sixth century BCE. Myers points to the mention of the Greeks in the book as an indication that the book was written at a time when the Greek international commerce was widespread, making the sixth century a likely date. Moreover, Joel has a number of thematic similarities with the book of Haggai, including a famine, an exhortation to repent, and a prophecy of judgment being visited upon the nations of the world (Allen 1976).

A number of other arguments have been advanced in favor of a post-exilic dating. The lack of any mention in the book of the Jewish monarchy indicates to some scholars that the book must have been written after the Jewish people were no longer a sovereign nation in Israel. Moreover, Joel makes mention of "the captivity of Judah and Jerusalem" (Joel 3:1) which would seem to indicate that the book is written after the exile of Jerusalem (Treves 1957). Joel 3:2 similarly says that the Israelites have been scattered around the world (Stephenson 1969).

The book also mentions the wall of the city (2:7, 2:9) which some scholars see as a reference to the walls rebuilt by Nehemiah around the end of the fifth century BCE, which would mean that the book was written after the time of Nehemiah. This argument is repudiated by Assis, who points out that even if we are to assume that the events in the Book

of Joel took place after Jerusalem's walls were breached by the Babylonians, there is no indication that the walls of Jerusalem were *completely* destroyed before they were rebuilt by Nehemiah (Assis 2011).

The latest suggestions for the date of Joel stand at around 350 BCE. Stephenson, one of the proponents of this position, points to Joel 2:31, which states, that "the sun shall be turned to darkness, and the moon to blood, before the coming of the day of the Lord, the great and the awesome." Stephenson understands this to refer to a complete solar eclipse, which he argues would only be visible in Israel during 402 BCE, 357 BCE, and 336 BCE (Stephenson 1969). Stephenson's argument is far from conclusive, however, because we can in no way be sure if the verse is actually referring to a solar eclipse as opposed to a creative metaphor for darkness and destruction. Even if the verse does refer to an eclipse as Stephenson claims, there is no reason to assume that the eclipse must have happened during Joel's lifetime.

In addition to the general controversy surrounding the date of Joel's authorship, there is also a debate about whether Joel can even be dated as a cohesive unit. Some scholars have argued that the book should not be taken as a unified text, but rather the two sections should be treated at separate documents, possibly each written by a separate author. J.W. Rothstein, for instance, posited that the first section of the book was written in the pre-exilic era whereas the second part of the book was written after the Babylonian exile. Others have taken this position even further, claiming that there are a number of portions of the book that that seem out of place, which could indicate that they were added to the original text at a later time. These theories have been rejected by others who prefer to see the book as a unified text. The unity of Joel--or the lack thereof--is still a disputed issue.

B. Using our Models to Classify the Book of Joel

Using the first model we created previously (word unigrams with Bayesian multinomial regression), the book of Joel is turned into a vector of words and categorized as belonging to the class of Late Biblical Hebrew. Of course, this does not mean that every word in the book is indicative of Late Biblical Hebrew; on the contrary, the book contains many words which are more indicative of Early Biblical Hebrew than they are of Late Biblical Hebrew. Our model, however, examines the book in a holistic fashion, classifying the document as being overall more linguistically similar to Late Biblical Hebrew than to Early Biblical Hebrew.

We now verify the conclusion we achieved using our word frequency model by classifying Joel with our other models: bigram frequency and morphology. Joel is similarly classified using the word bigram model and the morphology model with the same result: Late Biblical Hebrew.

We would also like to take into consideration the claim that the two parts of the book of Joel were written at different times, namely, that the first section was written in the pre-exilic era and the second section was written after the exile. This claim is easily evaluated using our algorithm: we simply break up the book into its two sections and run our classification models on each part separately. Even when breaking up the book in this way, we achieve the same result when classifying according to both word unigram and word bigram frequency: both sections of the book are classified as Late Biblical Hebrew.

However, when we use the morphology model to classify the two sections of Joel separately, we find an interesting result: the first section is classified as Late Biblical Hebrew, while the second section is classified as Early Biblical Hebrew, exactly the reverse of what Rothstein suggested. Nevertheless, because all three models classified the entire

book as Late Hebrew, and because both the word frequency model (which has a higher level of accuracy than the morphology model) and the bigram model classified both sections individually as belonging to Late Biblical Hebrew, it stands to reason that the classification of the second section of Joel probably does belong to Late Biblical Hebrew despite the fact that the morphology model indicates that it belongs to Early Biblical Hebrew.

At this point, we should note that even if we have demonstrated that the *book* of Joel was written during a late period, this does not necessarily imply anything about when the *person* Joel lived and prophesied (assuming Joel was a person who indeed existed). The content-based arguments of those who favor a pre-exilic dating could still be relevant to demonstrating that Joel lived during the time before the exile, which would then mean that the book of Joel is a book written in the post-exilic period about pre-exilic circumstances.

Part V: Classifying all the Books of the Bible

Now that we have demonstrated how supervised machine learning gives us a result for the book of Joel, we turn to classifying all of the books of the Bible not included in our training sets. We present here the conclusions achieved with each of the feature types we tested: word frequency, bigram frequency, and morphology. The books with an asterisk next to their names are books for which the different classification models produced different results.

Book	Word Unigram	Word Bigram	Morphology Unigram	Book	Word Unigram	Word Bigram	Morphology Unigram
Amos	Early	Early	Early	Leviticus	Early	Early	Early
Deuteronomy	Early	Early	Early	Malachi*	Early	Early	Late
Exodus	Early	Early	Early	Micah*	Late	Late	Early

Ezekiel*	Early	Late	Early	Nahum	Late	Late	Late
Genesis	Early	Early	Early	Numbers	Early	Early	Early
Habakkuk	Late	Late	Late	Obadiah*	Late	Late	Early
Haggai	Early	Early	Early	Proverbs	Late	Late	Late
Hosea*	Late	Early	Early	Psalms*	Early	Late	Early
Isaiah	Early	Early	Late	Ruth*	Early	Late	Early
Jeremiah	Early	Early	Early	Canticles	Late	Late	Late
Job*	Late	Late	Early	Zechariah	Late	Late	Late
Joel	Late	Late	Late	Zephaniah *	Late	Late	Early
Lamentations	Late	Late	Late				

Most of the classifications here turn out as we would expect them to. All of the books of the Pentateuch are unanimously assigned to Early Hebrew, Canticles is unanimously classified as Late Hebrew, and so on, just as the much of the scholarship on those books would indicate. There are some strange results, like the books of Haggai being dated as Early Biblical Hebrew even though the book of Haggai clearly discusses the post exilic-period[21]. The short length of Haggai (the book is only two chapters long) may have contributed to this inaccurate classification.

The fact that the book of Jeremiah was classified as Early and that Lamentations was classified as Late is a curious result. Both Lamentations and Jeremiah are traditionally attributed to the prophet Jeremiah, who lived when the Babylonians came into Israel and conquered Jerusalem. Given the results of our algorithm, however, it would appear rather

[21] This is in line with Young's observation that the linguistic features of Haggai seems to be more akin to Early Biblical Hebrew than to Late Biblical Hebrew. It is possible that Haggai wrote in a style more similar to Early Biblical Hebrew because he saw himself as an inheritor of a tradition of a particular style of prophetic writing and thus chose to write in a more archaic form. (Thanks to Professor Joshua Berman for this suggestion.)

unlikely that both books were written by the same author. Apparently, Lamentations was written by an author who lived long enough after the destruction of Jerusalem to have adopted the linguistic nuances endemic to Late Biblical Hebrew.

Some books, like Ezekiel, were probably written during a transition period between Early and Late Biblical Hebrew, which is why we get different results when using different feature types. This is an important consideration to keep in mind whenever classifying books according to the Early-Late classification scheme: even if Early and Late Biblical Hebrew are distinguishable classes of the Hebrew language, in some cases the most accurate approach is to treat biblical Hebrew as a spectrum, with some texts clearly belonging to Early Biblical Hebrew, some texts clearly belonging to Late Biblical Hebrew, and some which are somewhere in the middle. Using content-based arguments, as well as looking at different types of linguistic data, can be of assistance in determining whether a text belongs to such an in-between period.

Part VI: Classifying Psalms by Chapter

In addition to classifying all of the books in the Bible, we show how our approach can be used to classify individual chapters of the book of Psalms, as the psalms in the Psalter are generally presumed to have been written over a long period of time, with some psalms having been written before the exile and some having been written after it.

Psalms can difficult to date for a number of reasons. For one thing, many psalms lack any sort of content that would allow us to assign a likely date to their authorship. Moreover, the fact that the psalms use poetic language, which often differs from the Hebrew in the narrative and legal sections of the Bible, makes them difficult to classify linguistically. Psalms are also quite short, which could affect the likelihood that an analysis of word

unigram, word bigram, or morphology unigram frequency can really tell us very much about the dating of the individual psalms. Nevertheless, we present the results of our models' classifications for the purposes of further research.

Chapter	Word Unigram	Word Bigram	Morphology Unigram	Chapter	Word Unigram	Word Bigram	Morphology Unigram
Ch001	Early	Late	Early	Ch077	Late	Late	Early
Ch002	Early	Late	Late	Ch078	Late	Late	Late
Ch003	Early	Early	Early	Ch079	Late	Late	Late
Ch004	Early	Early	Early	Ch080	Early	Late	Late
Ch005	Early	Early	Early	Ch081	Early	Late	Early
Ch006	Early	Late	Early	Ch082	Late	Early	Late
Ch007	Late	Late	Early	Ch083	Late	Late	Late
Ch008	Late	Late	Late	Ch084	Early	Late	Early
Ch009	Early	Late	Early	Ch085	Early	Late	Late
Ch010	Early	Late	Early	Ch086	Late	Early	Late
Ch011	Late	Late	Early	Ch087	Early	Late	Late
Ch012	Early	Late	Early	Ch088	Late	Late	Late
Ch013	Early	Late	Early	Ch089	Late	Late	Late
Ch014	Late	Late	Early	Ch090	Late	Late	Late
Ch015	Late	Late	Early	Ch091	Late	Late	Late
Ch016	Early	Late	Early	Ch092	Early	Late	Late
Ch017	Early	Late	Early	Ch093	Early	Late	Late
Ch018	Early	Early	Early	Ch094	Early	Early	Early
Ch019	Late	Late	Late	Ch095	Early	Late	Early
Ch020	Early	Late	Late	Ch096	Late	Late	Late
Ch021	Late	Late	Late	Ch097	Late	Late	Late
Ch022	Late	Early	Early	Ch098	Late	Late	Late

Ch023	Late	Early	Early	Ch099	Early	Late	Late
Ch024	Early	Late	Late	Ch100	Late	Early	Late
Ch025	Early	Early	Early	Ch101	Early	Late	Early
Ch026	Early	Late	Late	Ch102	Late	Late	Late
Ch027	Early	Early	Early	Ch103	Early	Late	Late
Ch028	Early	Late	Late	Ch104	Late	Late	Late
Ch029	Early	Late	Late	Ch105	Early	Late	Early
Ch030	Early	Late	Late	Ch106	Late	Late	Late
Ch031	Early	Early	Early	Ch107	Late	Early	Late
Ch032	Early	Late	Early	Ch108	Late	Late	Late
Ch033	Early	Early	Late	Ch109	Late	Late	Early
Ch034	Early	Late	Early	Ch110	Late	Early	Late
Ch035	Early	Late	Early	Ch111	Late	Late	Late
Ch036	Late	Late	Late	Ch112	Early	Late	Late
Ch037	Late	Late	Late	Ch113	Early	Late	Late
Ch038	Late	Late	Late	Ch114	Late	Early	Late
Ch039	Early	Late	Early	Ch115	Late	Late	Early
Ch040	Early	Late	Early	Ch116	Early	Early	Early
Ch041	Early	Late	Early	Ch117	Early	Late	Late
Ch042	Late	Late	Early	Ch118	Early	Late	Early
Ch043	Late	Early	Late	Ch119	Late	Late	Early
Ch044	Late	Late	Late	Ch120	Early	Early	Early
Ch045	Late	Late	Late	Ch121	Early	Late	Early
Ch046	Early	Late	Late	Ch122	Late	Late	Early
Ch047	Early	Late	Late	Ch123	Early	Early	Early
Ch048	Late	Late	Late	Ch124	Early	Late	Early
Ch049	Late	Late	Early	Ch125	Early	Early	Late

Ch050	Late	Late	Late	Ch126	Early	Early	Late
Ch051	Late	Late	Late	Ch127	Early	Late	Early
Ch052	Late	Late	Late	Ch128	Early	Late	Early
Ch053	Late	Late	Early	Ch129	Early	Late	Early
Ch054	Early	Late	Early	Ch130	Early	Early	Late
Ch055	Late	Late	Early	Ch131	Early	Early	Early
Ch056	Early	Late	Early	Ch132	Early	Late	Early
Ch057	Late	Late	Early	Ch133	Late	Early	Early
Ch058	Early	Late	Early	Ch134	Early	Late	Early
Ch059	Late	Early	Early	Ch135	Early	Late	Late
Ch060	Late	Late	Early	Ch136	Late	Late	Late
Ch061	Late	Early	Early	Ch137	Early	Late	Early
Ch062	Late	Late	Early	Ch138	Late	Late	Late
Ch063	Late	Late	Early	Ch139	Early	Late	Early
Ch064	Early	Late	Early	Ch140	Early	Late	Early
Ch065	Late	Late	Late	Ch141	Early	Late	Early
Ch066	Early	Late	Early	Ch142	Early	Early	Early
Ch067	Late	Late	Late	Ch143	Early	Early	Early
Ch068	Late	Late	Late	Ch144	Early	Late	Early
Ch069	Late	Late	Late	Ch145	Early	Late	Late
Ch070	Early	Late	Early	Ch146	Early	Late	Early
Ch071	Late	Late	Late	Ch147	Early	Late	Late
Ch072	Early	Late	Late	Ch148	Late	Late	Late
Ch073	Late	Late	Late	Ch149	Late	Late	Late
Ch074	Late	Late	Early	Ch150	Early	Late	Late
Ch075	Late	Late	Late				
Ch076	Late	Late	Early				

In most cases, our models yield conflicting results as to the dating of particular chapters in Psalms. This is not especially surprising, as the poetic language and content of Psalms almost put the book in its own linguistic category, distinct from both Early and Late Biblical Hebrew. Nevertheless, the conclusions of our classification models should prove to be of at least some use, in conjunction with other evidence, in determining the date of authorship of certain chapters in the book of Psalms.

Part VIII: Conclusions

The results of our cross-verification analysis, as well as the conclusions obtained when classifying all the books of the Bible, demonstrate that supervised machine learning can be a useful tool to classify biblical texts as belonging to either Early or Late Biblical Hebrew. It is not, however, a foolproof method for accomplishing this task. We have shown that in many cases, our classification models have differed in their classification of a document, and in a few cases (such as the book of Haggai), all of the models clearly misidentified a document.

The accuracy of the classification models likely can be improved, but only to a certain point. In theory, one could look at all sorts of different feature types -- such as letter (as opposed to word) bigrams and trigrams, or individual aspects of the morphology of certain words, like the frequency of particular verb constructions—in an attempt to achieve a more accurate classification[22]. It is improbable, however, that a "silver bullet" feature type

[22] To improve the accuracy of the classifications, one could also be more selective in terms of the specific features chosen to create the models, such as focusing purely on function words as opposed to just words pass a certain frequency threshold. It could also be valuable be to expend more effort on developing the training sets by including more texts in each class and removing parts of books which are suspected as not having been

(or even combination of feature types) will be found that will discriminate between the two classes of Hebrew with complete accuracy. This is not due to the machine learning approach, per se, but rather to the nature of the problem itself. Early Biblical Hebrew and Late Biblical Hebrew, while discernible in many situations, aren't completely independent linguistic sets. Multiple dialects of Hebrew undoubtedly were used contemporaneously in ancient Israel during both the pre-exilic period and the post-exilic period, some of which might have been closer to Early Biblical Hebrew and some of which were probably more similar to Late Biblical Hebrew. Moreover, as Young suggested,a later author might conceivably feign an earlier stratum of Hebrew. It is unsurprising, therefore, that we should not be able to distinguish between them with perfect precision.

The utility of supervised machine learning algorithms for biblical studies is not limited to the problem of classifying Early and Late Biblical Hebrew. One could envision machine learning techniques being utilized to solve all sorts of problems, such as finding whether there is a linguistic dividing line between biblical poetry and biblical narrative, or if it is possible to more precisely dating books of the Bible relative to each other. Supervised machine learning can offer a whole world of possibilities to the Bible scholar who wishes to use linguistic features to help tell him something about the text. In conjunction with an understanding of the content and the history of the Bible, supervised machine learning can be a fresh and rigorous way to explore questions about the biblical text that have long been thought unsolvable.

---

written contemporaneously with the rest of the book. In line with Hurvitz's approach, one might also consider using post-biblical Hebrew, such as rabbinic or Qumran Hebrew, to include in the Late Biblical Hebrew training set in order to have a bit more data to work with.

References

Allen, L. C. (1976). *The books of Joel, Obadiah, Jonah, and Micah*. Grand Rapids, MI: Eerdmans.

Andiñach, P. R. (1992). The Locusts in the Message of Joel. *Vetus Testamentum*, *42*, 433-441.

Assis, E. (2011). The Date and Meaning of the Book of Joel. *Vetus Testamentum*, *61*, 163-183.

Hurvitz, A. (1997). The Historical Quest for "Ancient Israel" and the Linguistic Evidence of the Hebrew Bible: Some Methodological Observations. *Vetus Testamentum,*, *47*, 301-315.

Hurvitz, A. (2006). The Recent Debate on Late Biblical Hebrew: Solid Data, Experts' Opinions, and Inconclusive Arguments :. *Hebrew Studies*, *47*, 191-210.

Joel, Book Of. (1906). In *Jewish Encyclopedia*. Retrieved from http://www.jewishencyclopedia.com/articles/8703-joel-book-of

Joosten, J. (2005). The Distinction Between Classical And Late Biblical Hebrew as Reflected in Syntax. *Hebrew Studies*, *46*, 327-339.

Koppel, M., Mughaz, D., & Akiva, N. (2006). New Methods for Attribution of Rabbinic Literature. *Bar Ilan University*. Retrieved from http://u.cs.biu.ac.il/~koppel/papers/balshanut-26.5.04.pdf

Koppel, M., Schler, J., & Argamon, S. (2009). , Computational Methods in Authorship Attribution. *JASIST*, *60*(1), 9-26.

Lange, A., Weigold, M., Zsengellér, J., & Tov, E. (2009). *From Qumran to Aleppo: A discussion with Emanuel Tov about the textual history of Jewish scriptures in honor of his 65th birthday*. Göttingen: Vandenhoeck & Ruprecht.

Littlestone, N. (1988). Learning Quickly when Irrelevant Attributes Abound: A new linear-threshold algorithm. *Machine Learning*, *2*(4), 285-318. doi: 10.1007/BF00116827

Longman, T. (1998). *The book of Ecclesiastes*. Grand Rapids, MI: W.B. Eerdmans.

Madigan, D., Genkin, A., Lewis, D. D., & Fradkin, D. (2005). Bayesian Multinomial Logistic Regression for Author Identification. :. *Bayesian Inference and Maximum Entropy Methods in Science and Engineering*, *803*, 509-516.

Rooker, M. F. (1994). Diachronic Analysis and the Features of Late Biblical Hebrew.*Bulletin for Biblical Research*, *4*, 135-144. doi: http://www.ibr-bbr.org/files/bbr/BBR_1994_10_Rooker-LateHebrew.pdf

Stephenson, F. (1969). The Date of the Book of Joel. *Vetus Testamentum*, *19*(2), 224-229.

Treves, M. (1957). The Date of Joel. *Vetus Testamentum*, *7*(2), 149. doi: 10.2307/1515837

Young, I. (2005). Biblical Texts Cannot be Dated Linguistically. *Hebrew Studies*, *46*, 341-351.

Acknowledgements

This thesis could not have been written without the contribution of many people who provided me with invaluable assistance, ideas, and constructive criticism. First and foremost, I would like to thank Professor Moshe Koppel of Bar-Ilan University for introducing me to the field of machine learning approaches to text categorization and being my mentor for the research and writing of this thesis. I would also like to thank Professor Joshua Berman of Bar-Ilan University, who proposed the project of classifying biblical texts into Late and Early Biblical Hebrew and for allowing me to tap into his wealth of knowledge about biblical studies to help me better understand the historical and linguistic aspects of this project. I am also grateful to Professor Idan Dershowitz for providing comments to an earlier draft of this thesis.

I owe a debt of gratitude to Kfir Zigdon, who designed the text categorization software which made this thesis possible. I am also grateful to Ofir Tadmor, with whom I worked daily over the summer of 2012 on an updated version of the text categorization software and who patiently helped me understand the intricacies of implementing machine learning algorithms.

Many of my professors at Yeshiva University also provided me with valuable suggestions about various aspects of my thesis, including Professors Moshe Bernstein, Shalom Holtz, and Jeremy Wieder. I owe a special thanks to Professor Wieder for providing me with Bible texts that I could easily scan using the text categorization software.

I would like to acknowledge the Yeshiva University and the Jay and Jeanie Schottenstein Honors Program for providing me with the framework that enabled me to produce this thesis. In this vein, I would like to express my appreciation to the honors

program faculty members who have been helpful throughout the process of writing this thesis, including Professors Gabriel Cwilich, Samuel Gellens, and Gillian Steinberg, all of whom were incredibly supportive of the efforts of all the thesis writers in the honors program.

Finally, I would like to express my everlasting gratitude to my parents, who have supported me throughout my college career and continue to make enormous sacrifices on my behalf. I couldn't have written this thesis (or accomplished much of anything else) without the foundation that they have given me.