Clustering Biblical Texts Using Recurrent Neural Networks

Yanniek van der Schans
yanniekvdschans@gmail.com
Vrije Universiteit Amsterdam

David Ruhe
david.ruhe@gmail.com
Vrije Universiteit Amsterdam

Wido van Peursen
w.t.van.peursen@vu.nl
Vrije Universiteit Amsterdam

Sandjai Bhulai
s.bhulai@vu.nl
Vrije Universiteit Amsterdam

ABSTRACT

This study examines linguistic variation within Biblical Hebrew by using Recurrent Neural Networks (RNNs) to detect differences and cluster the Old Testament books accordingly. Various linguistic features are analysed that are traditionally considered to be of importance in analysing linguistic variation. The traditional division of books as either Early Biblical Hebrew or Late Biblical Hebrew is hereby put to the test. Results show that RNNs are a fitting method for analysing the (morpho)syntax of a language. The model works well on both separate features, as well as all the features combined. On the basis of the results the RNNs provide, we propose that the diachronic approach to Biblical Hebrew is indeed plausible. The clusters generally hint to the scholarly division made in the diachronic approach to linguistic variation.

KEYWORDS

Recurrent Neural Networks, Biblical Hebrew, Diachronic Linguistics, Computational Semantics

1 INTRODUCTION

Within the field of Old Testament scholarship, there is no consensus on the possibility of dating the Hebrew Bible according to the linguistic variation that appears in the text. A point of debate is the question whether the linguistic differences between the books can be a sufficient ground to date them, or if they merely show a stylistic difference. In this research, we explore the potential of machine learning to capture the linguistic variations by clustering the books according to those variations. Specifically, by using Recurrent Neural Networks (RNNs), we explore a possibility for providing new insights into the ongoing discussion within the field of Hebrew linguistics.

A Recurrent Neural Network can effectively model sequential data, such as language. It can output the next state by taking into account not only the current state, but also the preceding states. The model can, therefore, offer a rather precise way of studying the linguistic variation of a language. This research is a continuation of previous work that explored the possibility of linguistic dating of Old Testament books [15]. The authors used Markov Chains to model Old Testament texts. Our research furthers on this in the following ways.

(1) We use a more sophisticated approach that has proven to be very effective in modelling sequential data like language;
(2) Our approach can directly model the word usage. Since the number of words used in the Hebrew Bible is large, this is computationally very expensive and left out of previous research.

Our research focuses on the question: how can Recurrent Neural Networks help to give insight into the linguistic variation of the Hebrew Bible? To answer this question we will provide the following:

(1) An investigation of the effectiveness of the application of RNNs as language models to the Hebrew Bible;
(2) A clustering approach to a selection of the Old Testament books using RNNs;
(3) A comparison of the results of this approach with the established views of Old Testament scholars.

In the following sections, we first provide background information about the scholarly debate on linguistic variation in Biblical Hebrew. Second, we discuss the data that was used, and how we constructed our data set. We then discuss the methodology we have chosen. The results of our research are then presented and discussed, after which we will conclude with answering our research question.

2 BACKGROUND

Dating biblical texts has been a challenging endeavour in the study of the Old Testament. Trying to recover a chronological time frame has been part of biblical criticism for many years, with different methods and arguments along the way. One of the major points of discussion is the question whether the linguistic variation of the texts can provide warranty in the dating of the texts. Generally speaking, there are two positions in the discussion: those who believe that linguistic variation shows a division between Early Biblical Hebrew (EBH) and Late Biblical Hebrew (LBH), and those who believe that this variation merely points to a stylistic preference [4]. In this division, EBH embodies the texts that were written in the pre-exilic period, whereas LBH texts were written after the exile.

On the one hand, there are those who argue that linguistic variation in biblical texts shows a division between EBH and LBH, suggesting a diachronic explanation. Wilhelm Gesenius [2] pioneered the diachronic study of Biblical Hebrew, paving the way for diachronic analyses. He argued that there are two distinct layers in Biblical Hebrew that point to two successive stages which would later be labbeled as pre-exile (EBH) and post-exile (LBH) ([1] pp. 21). Avi Hurvitz continued Gesenius' line of thought[6]. Both Hurvitz and Robert Polzin found features that point to an LBH writing style, using the Chronicler's language as important representative of LBH. Hurvitz argued, as Kim shows in his book on the linguistic variation of biblical Hebrew: "Linguistic change in

Biblical Hebrew (BH) during the exile was so decisive as to render the post-exilic biblical writers unable to write Early Biblical Hebrew (EBH) of the pre-exilic period; and second, that since LBH of the post-exilic period was a linguistic body distinct from EBH both in form and in chronology, one can date biblical texts solely on the basis of linguistic data"([6] pp. 1.). In other words, by analysing the language of a text, one can deduce features that point to either an early dating (EBH) or a later dating (LBH). According to Hurvitz, BH is linguistically heterogeneous, because the language changed over time. Hurvitz argued that during the exile, the contact between Aramaic and Hebrew brought change to the latter, making pre-exilic (EBH) and post-exilic Hebrew (LBH) two distinct languages. In his view the chronological development of BH was supported by and extra-biblical evidence [6]. From these presumptions, Hurvitz constructed a method of defining LBH features that could even be used for dating texts of unknown origin on the basis of these features. His method was almost a mathematical formula; the input of linguistic data would produce the output of either EBH or LBH [6].

Up until the 2000s, the existing diachronic approach was a broadly accepted, even though there were also dissident voices. Whereas in the 1990s advocates of alternative views (sometimes labelled as "minimalists") focused in historical and archaeological data, in the first decade of the current century, the linguistic framework of the diachronic approach was seriously challenged by series of publications by Ian Young, Robert Rezetko and others (see [20] pp. 341-351, [22], [21]). These scholars argued that linguistically dating biblical texts is impossible [6, 20]. They agree with the diachronic approach on the linguistic differences between EBH and LBH, but they prefer to ascribe these differences to *style*, rather than to diachronic development. Rather than assuming that the two corpora were composed and edited in different periods, they argued that “scribes modified individual linguistic elements occasionally and unsystematically” ([21] pp. 597.) and that EBH and LBH are to be thought of as “co-existing styles of literary Hebrew throughout the biblical period" [21]. To avoid confusion, they started to use the labels "Standard Biblical Hebrew" (SBH) and "Peripheral Biblical Hebrew" (PBH), rather than EBH and LBH[11]. They state that various LBH features appear in the EBH corpus as well [6], and that only a few LBH features appear in all LBH texts. They, therefore, conclude that “it is often difficult to determine which feature in which book would represent truly late language” (see [22] pp. 86-87.), arguing that the choice between EBH and LBH was a matter of style, not time [21]. This implies that only the EBH books appear to be a strong group together, but that the rest of the books are not necessarily to be clustered. The distinction made is that of SBH and the peripheral books, the latter not forming a cluster on its own.

The debate between the traditional, chronological approach of Biblical Hebrew, and the challengers’ alternative, the stylistic approach is difficult to reconcile. By using RNNs, we hope to provide some insight that can help further the debate, and give a new perspective on the discussion.

3 DATA

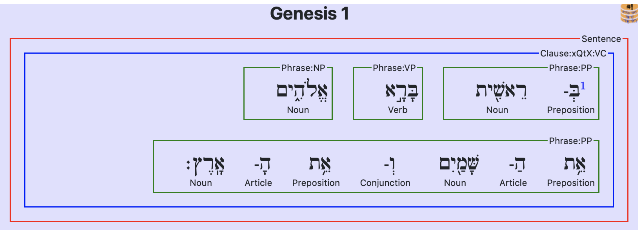


Figure 1: Example of annotation in Genesis.

In this research, we use data provided by the Eep Talstra Centre for Bible and Computer (ETCBC) [17], that provides a data package containing different annotations called Biblia Hebraica Stuttgartensia Amstelodamensis (BHSA). The BHSA is an open-source Hebrew Bible data set that can be accessed through Text-Fabric [13]. Text-Fabric works with the Python programming language, which offers different packages and tools that one can use in combination with Text-Fabric. The BHSA data set carries different annotations that are stored on different linguistic levels: word, phrase, clause, and text layers. These annotations are stored as features and values. Figure 1 shows the different layers of the annotations in the text. The red line presents the sentence structure, the blue one marks the clauses, where the green lines give away the structure of the different phrases. The extra information shows the specific clause and phrase types.

We will use the traditional division introduced in Section 2 as a basis for our research:[1]

- Early Biblical Hebrew: Genesis, Exodus, Deuteronomy, Leviticus, Judges, Joshua, Kings, Samuel.
- Late Biblical Hebrew: Chronicles, Daniel, Ecclesiastes, Esther, Ezra-Nehemiah, Song of Songs.

With compiling this list, we chose to leave out some books (e.g., Psalms, Isaiah) since they do not have a homogeneous nature, or the books which are debated when it comes to their categorisation. The books of Ezra and Nehemiah are considered as one book in their original construction, so we followed that line of thought in our research [14]. The books 1 and 2 Kings, 1 and 2 Samuel, and 1 and 2 Chronicles are also most commonly understood as to be single compositions [10, 15, 21]. Thus, we combined them in our data set as well. Since this study considers the Hebrew parts of the Old Testament, we filtered out the Aramaic sections and use the Hebrew sections only.[2]

3.1 Features

To be able to use the Recurrent Neural Networks to contribute to the ongoing debate on linguistic variation in the Bible, we selected features that have been considered typical of the differences between the EBH corpus and the other books. Extensive research has been done on the linguistic differences between EBH and LBH, resulting in lists and tables that show the variations (see [21] e.g., pp. 166-214.). The distinctions occur on multiple levels of the linguistic spectrum, e.g., from word forms to syntactic constructions.

[1] For sake of convenience we use the common designations of these corpora as EBH and LBH rather than SBH and PBH but that does not imply an a priori choice for the diachronic approach.

[2] The parts that include Aramaic are Genesis 31:47, Jeremiah 10:11, Daniel 2:4b-7.28, Ezra 4:8-6:18 and 7:12-26.

One such feature is the use of the verbal stem formations. Rooker [12] argued that there is an increase in the use of the Pi'el and a decrease in the use of Hif'il stem[3] and Kutscher [7] points out that there is an increase of Nif'al stem within the LBH books, replacing the passive Qal of EBH. Variation in the use of verbal stems in the alleged EBH and LBH corpora is acknowledged by both advocates and opponents of the diachronic approach [10, 21]. We would therefore expect that clear clusters appear when considering verbal stem predictions of the RNNs. Clustering on this feature could be a very neat one, assuming the hypothesis is correct.

A second feature concerns the *waw consecutive* sentences,[4] which in the ETCBC database are identified in the clause type labels. It has been argued that *wayyiqtol* sentences are less frequent in LBH. If that is correct, we expect clusters to emerge when training on this feature.

In addition to these two features that play a role in the scholarly debate, we have included some other syntactic features, which are less visible then verbal stem or verbal tenses, and therefore also less easy to be manipulated by an author who wants to write in a certain standard or archaic language. Because of the unconscious nature of these syntactic features, they are apt to reveal stylistic or chronological differences between the analysed books (see [8], pp. 244). These features include, part of speech, that executes on a word level, and phrase function, that looks at the syntactic function of the phrase.

The features that we will use in our research are then as follows:

- phrase: function – object, subject, relative, etc.;
- clause: type – nominal, participle, wayyiqtol, etc.;
- word: part of speech – noun, article, preposition, etc.;
- word: verbal stem – Hif'il, Pi'el, Qal, etc.

Since a RNN can model language directly on word usage, it is able to capture most of the linguistic variation directly from the language. Therefore, we also run the model directly on the Hebrew Bible. We hypothesise that the resulting modelled variation reflects the other features.

4 METHODOLOGY

In this section, we firstly discuss the chosen approach to model the corpus. Secondly, we discuss how the clusters were obtained.

4.1 Recurrent Neural Networks

Recurrent Neural Networks are designed to model sequential data. Language, in particular, exhibits a sequential structure. A straightforward and effective way of modelling language is by predicting the next word given a previous set of words. That is, the joint probability of a sequence of words $P(w_1, \dots, w_N)$ can be decomposed as follows:

$$P(w_1, \dots, w_N) = P(w_1) \prod_{i=2}^{N} P(w_i | w_1, \dots, w_{i-1}). \tag{1}$$

A previous iteration of this research modelled this probability distribution with Markov Chains [15]. A major drawback of this approach is that the model is reduced to

$$P(w_1, \dots, w_N) = P(w_1) \prod_{i=2}^{N} P(w_i | w_{i-1}). \tag{2}$$

In other words, the dependency of word w_i is modelled to be solely dependent on w_{i-1}; a highly unrealistic assumption. Contrarily to Markov Models, Recurrent Neural Networks are able to take previous states of the input into account by keeping track of a hidden state vector. Compared to previous research, the method used here is a more fitting one to give insight into linguistics.

Let $\boldsymbol{x}_1$ be an input vector at time-step 0 and $\boldsymbol{h}_0$ be an initialised hidden state vector. The RNN cell takes $\boldsymbol{x}_1$ and $\boldsymbol{h}_0$ and computes the first hidden state $\boldsymbol{h}_1$ as follows;

$$\boldsymbol{h}_t = \sigma_h(\boldsymbol{W}\boldsymbol{x}_t + \boldsymbol{U}\boldsymbol{h}_{t-1}), \tag{3}$$

where $\boldsymbol{W}$ and $\boldsymbol{U}$ are learned weight matrices. The hidden state vector captures the modified data (e.g., features) on the way to the final time-step. It can be seen as a representation of the previously observed words, capturing the required (syntactical) information to predict the next word. Parameters are learned by backpropagation through time [19]. Backpropagation through time can suffer from the infamous 'vanishing gradient problem' and fail to capture long-term dependencies effectively in its hidden state $\boldsymbol{h}_t$ [5]. Therefore, we use a Long Short-Term Memory model [3], which has proven to be very effective in retaining information for long sequences. By using a memory vector in addition to input, output and forget gates, we calculate hidden state $\boldsymbol{h}_t$ as follows.

$$\begin{aligned}
\boldsymbol{f}_t &= \sigma_g(\boldsymbol{W}_f \boldsymbol{x}_t + \boldsymbol{U}_f \boldsymbol{h}_{t-1} + \boldsymbol{b}_f) && (4)\\
\boldsymbol{i}_t &= \tau_g(\boldsymbol{W}_i \boldsymbol{x}_t + \boldsymbol{U}_i \boldsymbol{h}_{t-1} + \boldsymbol{b}_i) && (5)\\
\boldsymbol{o}_t &= \tau_g(\boldsymbol{W}_o \boldsymbol{x}_t + \boldsymbol{U}_o \boldsymbol{h}_{t-1} + \boldsymbol{b}_o) && (6)\\
\boldsymbol{c}_t &= \boldsymbol{f}_t \cdot \boldsymbol{c}_{t-1} + \boldsymbol{i}_t \cdot \sigma_c(\boldsymbol{W}_c \boldsymbol{x}_t + \boldsymbol{U}_c \boldsymbol{h}_{t-1} + \boldsymbol{b}_c) && (7)\\
\boldsymbol{h}_t &= \boldsymbol{o}_t \cdot \sigma_h(\boldsymbol{c}_t), && (8)
\end{aligned}$$

in which $\boldsymbol{f}_t$ denotes the activation to a forget gate, $\boldsymbol{i}_t$ the activation to an input gate, $\boldsymbol{o}_t$ the activation of an output gate, $\boldsymbol{c}_t$ the update to a memory cell, σ the logistic sigmoid function, and τ the hyperbolic tangent function. By using this structure, the LSTM unit is able to retain information for extended time periods.

The LSTM is trained on a corpus. In our work, this corpus consists of the books that compose the combination of EBH and LBH that we proposed in Section 3. By forwarding (parts of) the books, we can prime the model's hidden state solely on that book. If we then cluster the resulting hidden states, we get an overview of the modelled states for each book.

4.2 Hierarchical Clustering

To obtain clustering results, we use an agglomerative hierarchical clustering procedure: the Ward variance minimisation algorithm. First, we sample T sequences and save the resulting hidden states $\{\boldsymbol{h}_1, \dots, \boldsymbol{h}_T\}$ for each book in $\{b_1, \dots, b_M\}$ after forwarding the sequences through the RNN. Next, we take the mean hidden state of the book: $\boldsymbol{h}_i = \frac{1}{T}\sum_{t=1}^{T} \boldsymbol{h}_t$. Ward's algorithm starts with M mean hidden states that all represent a cluster e. In each clustering iteration, we compute the distance to the new cluster centroid for all cluster unions $\{e_i, e_j\}$. The pair that has the least distance to the

[3] He also points this out in the book *Biblical Hebrew in Transition*, in which the late grammatical features of the book Ezekiel are studied.

[4] See Polzin, Kropat, van Peursen, Rooker. Taken from Ian Young, and Robert Rezetko. *Linguistically Dating of Biblical Texts* vol. 2, pp. 166-214.

centroid of their union is united. We repeat this procedure until convergence. A mathematical overview is given in Algorithm 1.

```
Algorithm 1: Hidden state clustering procedure
  Data: $\{b_1, \ldots, b_M\} \leftarrow$ Set of books in corpus
  Function: $f(\boldsymbol{x}, \boldsymbol{\theta}) : \mathcal{X} \rightarrow (\boldsymbol{h}, y) \in \{\mathcal{Y}, \mathcal{H}\}$ where $\boldsymbol{X}$ is a word
            sequence, $y$ the predicted next word and $\boldsymbol{h}$ the
            hidden representation of the sequence.
  Result: Book clusters $\{e_1, \ldots, e_z\}$
  for $i \leftarrow 1$ to $M$ do
      sample $T$ sequences $\boldsymbol{x}_t = \{w_1, \ldots, w_N\}$ from $b_i$
      obtain $T$ hidden states using $f(\boldsymbol{x}_t, \boldsymbol{\theta})$
      $\boldsymbol{h}_i \leftarrow \frac{1}{T} \sum_{t=1}^{T} \boldsymbol{h}_t$
  end
  $E = \{e_1, \ldots, e_C\} \leftarrow$ Initialize a cluster for each book: $M = C$.;
  while not converged do
      $e^* \leftarrow \arg\min_{\{e_i, e_j\}} \sum_{\boldsymbol{h}_k \in e_i \cup e_j} \|\boldsymbol{h}_{\boldsymbol{k}} - \frac{1}{|e_i \cup e_j|} \sum_{k=1}^{|e_i \cup e_j|} \boldsymbol{h}_k\|_2, \quad \{i, j\} \in [1, \ldots, C] \times [i, \ldots, C]$
      $E \leftarrow E \cup e^*$
      $E \leftarrow E \setminus \{e_i\}, \quad \forall e_i \in e^*$
  end
```

5 RESULTS

The model proved to be very effective in modelling the Hebrew language. Some fully generated sentences are depicted in Figure 2. Their grammar is quite accurate. The average accuracy of the word language model was 90.02%. Running language models on the selected features yielded promising results. The accuracies were 87.45%, 81.48%, 62.14%, and 55.56% for verbal stem, clause type, phrase function, and word part-of-speech, respectively.

Looking at the separate features, as well as all of them combined, the method shows that there is a strong preference of a part of the classified EBH books, and a strong preference of some classified LBH books. There also appears a middle group, that moves between these two clusters depending on which features the RNNs analyse. This gives some support to the diachronic approach, because Hurvitz and others consider EBH and LBH as two clusters, whereas those who distinguish between standard and peripheral BH tend to stress the variety within the alleged LBH corpus. We will analyse the separate features below, after which we will conclude with what this could imply for the linguistic variation debate.

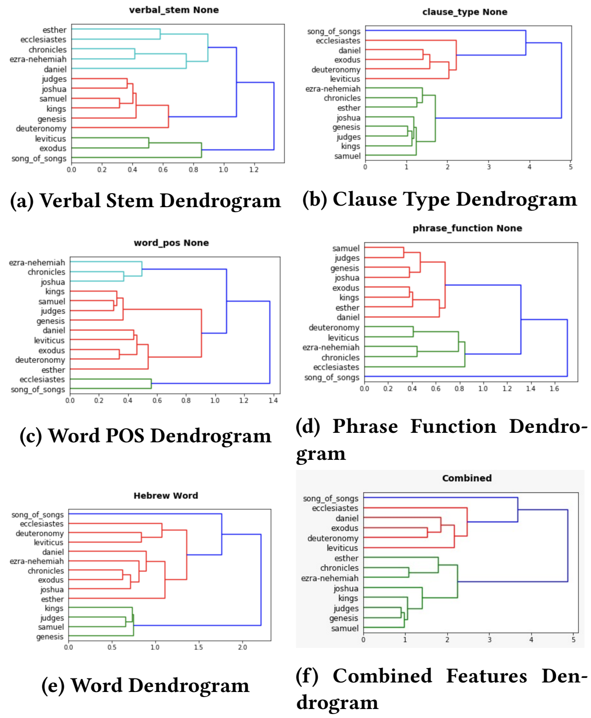
5.1 Verbal Stem

The dendrogram in Figure 3a shows a neat clustering of the books, according to the traditional division of EBH and LBH. All the classified LBH books form a cluster, with the Song of Songs being the only exception. This could be related to the fact that the book is a unique one in the selection of books we use. The book is made up entirely out of 'quotative' text (direct speech as opposed to narrated text), which entails that the verbal stem use is different as well[5]. Oddly enough, the book shows up in a cluster with Exodus and Leviticus, rather than forming a cluster on its own, as one would expect. An explanation for this could be that Exodus and Leviticus too, have a high percentage of clauses with quotative text. Maybe the verbal stem use in direct speech is more simple than in the narrative parts of the books, and thus explains the clustering of the Song of Songs with EBH books like Leviticus and Exodus. This could be an interesting subject for further investigations.

וַ יַּעַשׂ מֹשֶׁה כְּ כֹל אֲשֶׁר צִוָּה יְהוָה אֹתֹו כֵּן עָשָׂה

וּ בָאוּ עָלֶיךָ כָּל־ הַ בְּרָכֹות הָ אֵלֶּה וְ הִשִּׂיגֻךָ כִּי תִשְׁמַע בְּ קֹול יְהוָה אֱלֹהֶיךָ

Figure 2: Two generated Hebrew sentences.



(a) Verbal Stem Dendrogram

(b) Clause Type Dendrogram

(c) Word POS Dendrogram

(d) Phrase Function Dendrogram

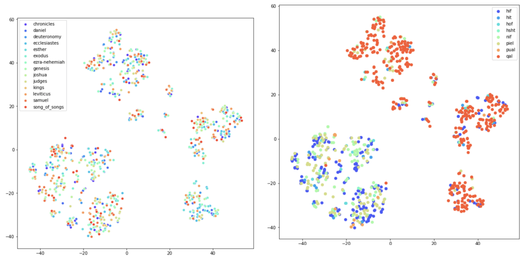
(e) Word Dendrogram

(f) Combined Features Dendrogram

Figure 3: Clustering results for all analysed features.

As stated above, there are scholars who argue that the use of the verbal stem is a feature that changed within the Hebrew language. With compiling a t-SNE visual [16] of the different verbal stems in the books, we can find out if these specific changes are reflected in our test case as well. Combining both of the t-SNE models gives insight into which book represents which verbal stem, providing an overview of changes that happen over time. When analysing both the t-SNE graphics, see Figure 4, the hypothesis of a preference for Pi'el in late, and Hif'il in early Hebrew can be confirmed by our model. Combining both (a) and (b) together, it follows that the EBH books show a higher percentage of Hif'il than the LBH books do. The opposite goes for the Pi'el stem, just as the literature on this feature suggested. A.J.C. Verheij hypothesises for example that Qal has gradually been replaced by Pi'el ([18] pp. 132.). He shows that the two co-occur relatively infrequently, and states that this could point to the fact that they substitute each other ([18] pp. 70-76; 79;

[5] For the various text types in the ETCBC database, see [9].



(a) Division with books **(b) Division with verbal stems**

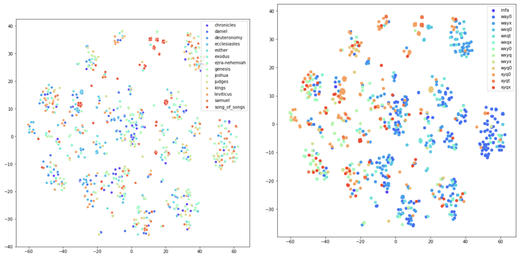
Figure 4: t-SNE clustering according to Verbal Stem.

131-132). This is a phenomenon that could be found in our results as well, making connections with the diachronic approach. The hypothesis that the use of the Nif'al has an increased use in the proposed LBH books, is one we can also detect in our results.

5.2 Clause Type

Another feature that plays an important role in the debate on linguistic variation in Biblical Hebrew is the clause type. It has been argued that the *waw consecutive* is used less frequently in LBH, being replaced by various other constructions. The *waw consecutive* is part of the clause type feature within the ETCBC database, and therefore we used the feature to cluster on. Within the feature, there are many different values available, with a total of 45 different possibilities.[6] For the t-SNE visualizations, we picked the most debated features when looking at linguistic variation, which came to a total of 13 most relevant values. This in order to keep the results and clustering from being clouded with features that appear in a similar way over all the books. Figure 3b shows the clustering according to the clause types, where an interesting set of clusters appear. Song of Songs again stands on its own compared with the other LBH books; the latter form two separate clusters, where Ecclesiastes and Daniel fall into the same cluster as Exodus, Deuteronomy, and Leviticus. The other LBH cluster consists of the books Ezra-Nehemiah, Chronicles and Esther, that form a separate cluster with the EBH cluster of Joshua, Genesis, Judges, Kings and Samuel. All of these books have a high percentage of narrative, where the other cluster consists of books that carry direct speech, poetry or laws in them. Accordingly, although some EBH and LBH clusters can be identified, the text type also has considerable effect on the clustering of the books.

Looking once again at the t-SNE (Figure 5b for the specific feature), the hypothesis is present in the results of the RNNs. The *waw consecutive* constructions are indeed less present in the classified LBH books, where more complex constructions appear. From the t-SNE, the division seems in line with the hypothesis, with the EBH group showing a strong favour for the *waw consecutive*. In a similar manner, we see that the *yiqtol* constructions are more prominent in the LBH, a sight that confirms the literature on the phenomenon. This, of course, does not warrant the diachronic approach, it merely confirms the fact that this indeed is a feature that is shown to be of importance in the division.



(a) Division with books **(b) Division with clause types**

Figure 5: t-SNE clustering of books according to Clause Type.

5.3 Word Part of Speech

The dendrogram in Figure 3c shows a clear distinction of Ecclesiastes and the Song of Songs as a separate cluster, both of these books, however, would be classified as poetry, where the remaining books of our data set are classified as prose. This might indicate that part of speech is a more genre sensitive feature. Likewise, this might explain the appearance of Daniel and Esther within the cluster that has mainly narrative EBH books. Joshua also appears in an unexpected spot with Ezra-Nehemiah and Chronicles. This could be related to the fact that these books all have a major section of a list of names in them, which could be why the books clustered together on the basis of the high number of words with the part of speech Proper Noun.

5.4 Phrase Function

The phrase function leads to a strong clustering. (see Figure 3d). The Song of Songs stands alone as a separate cluster. Again, this can most likely be explained from its poetic character and its quotative text type. There is a cluster of mainly LBH books, but including Leviticus and Deuteronomy. And there is a cluster of predominantly EBH books, but including Esther and Daniel. Again, this could be due to the dominant types of text in these books: the name lists in Chronicles, the enumerations of laws in Leviticus and Deuteronomy, the report style of Ezra-Nehemiah (including autobiographic sections), and the philosophical and reflexive style of Ecclesiastes. However we explain the unexpected clustering results, it is clear that there is no neat division along the borders of EBH and LBH.

5.5 Lexemes

When looking at word use, there are three main clusters that appear in the dendrogram, see Figure 3e. The clustering appears in roughly four spots, again with the Song of Songs as the odd one out. Perhaps this relates to the theme of the book, with a more passionate language, or to some unique lexemes that may be borrowings from Aramaic or other languages. The cluster of Kings, Judges, Samuel and Genesis again is strong, leaving a massive middle cluster, where LBH and EBH books mix with one another. Dividing this up in two sections shows a semi cluster of (a) LBH books, with the odd appearance of Joshua and Exodus, and (b) EBH, with Ecclesiastes breaking the cluster. The word clustering is largely in line with the clustering of the combined features, see Figure 3f. This supports our hypothesis that RNNs can directly model linguistic variation

[6] See the ETCBC database; https://etcbc.github.io/bhsa/features/hebrew/c/typ.html.

from language. This is an interesting finding for future analyses on data sets with less available annotation.

6 CONCLUSION

Within the field of Old Testament scholarship, there is no consensus on the possibility of dating the biblical books according to its linguistic variation. To shine new light on the matter, we explored the effectiveness of RNNs on modelling the Hebrew Bible. Additionally, we analysed the extent to which the results aligned with the established scholarly views. To this end, we trained multiple models on a set of features present in the BHSA corpus. We selected some features on the basis of the current scholarly debate, and added some features that are significant because of their syntactic, and hence less conscious nature: phrase function, clause type, part of speech and verbal stem in the analysis. We also went beyond existing research by modelling the text directly. Additionally, we proposed a clustering algorithm and analysed the interpretability of the results.

There is considerable agreement between the clustering of books based on words (without further grammatical annotation) and the clustering based in the combined features. This suggests that RNNs are also useful for data sets with less annotation available than the richly annotated ETCBC database.

Results show that RNNs can indeed be applied effectively on the Hebrew Bible to analyse its linguistic variation. Compared to the previous research pilot using Markov chains, [15], where clustering appeared only after combining all different features together, our model showed to be more fitting. The RNNs showed clustering on different levels, without combining the features together in order to get neat clustering. This makes it possible to get insight in which features turn out to be the most dominant ones when it comes to the linguistic changes. Especially on word level, the RNNs are more competent in showing clusters agreeing with current scholarly insights than the Markov Models were.

Furthermore, we obtained interesting insights that are relevant to the ongoing debate on linguistic variation in Biblical Hebrew. The obtained clusters appear to be generally in line with the established consensus regarding the alleged EBH and LBH corpora. A specific set of classified EBH books appear in a strong cluster throughout all the tested features; this group includes the books Genesis, Judges, Kings and Samuel. Joshua shows a preference to this cluster group as well, yet is not quite consistent enough to be adopted in the group. Another cluster occurs within the alleged LBH books, containing Ezra-Nehemiah, Chronicles, Ecclesiastes and Song of Songs. These four appear in close proximity to each other, with the duo Ezra-Nehemiah and Chronicles appearing together and the duo Ecclesiastes and Song of Songs sticking together over the various features. The results show clusters that partly reflect the division into EBH (or SBH) and LBH (or PBH). A strong cluster of EBH books consists of Genesis, Judges, Samuel and Kings.

That there exists such a cluster of EBH books is acknowledged by both the advocates and the opponents of a diachronic explanation of linguistic diversity in the Bible. Hence this outcome of our analysis does not favour one approach over the other. In most cases our analysis focuses on the distribution of linguistic features, rather than an intrinsic linguistic explanation that may refer to language development and hence support the diachronic approach (e.g. processes of analogy formation of weak forms), though in some cases a diachronic explanation is suggested by the distribution (e.g. the decreasing use of the passive qal). What remains is one outcome of our results that may be supportive of the diachronic approach, namely that we discovered not only a strong EBH cluster, but also one or more strong LBH clusters. This may support the view that these books reflect a certain language phase (and hence support the diachronic approach), rather than deviations from the standard language.

The traditional division between EBH and LBH is a particularly strong in the clustering based on *verbal stem* and *clause type*. In addition, there is a middle group that moves seemingly freely between the proposed EBH cluster and LBH cluster, depending on the feature analysed. This group consists of the following books; Exodus, Deuteronomy, Leviticus, Daniel, and Esther. This middle group requires further investigation. Within the diachronic framework, one may wonder whether the label "early" is appropriate for books such as Exodus, Leviticus, Deuteronomy. In some cases the alignment of books in certain clusters may be due to text type, and hence the legislative parts of these three books may have affected their linguistic profile. The narrative style of Daniel and Esther may have caused their position in this middle group. The following division is one that we found in our results and put forth as a hypothesis that could be a stepping stone for further research.

- Early Biblical Hebrew: Genesis, Judges, Joshua, Kings, Samuel.
- Middle group: Exodus, Deuteronomy, Leviticus, Daniel, Esther.
- Late Biblical Hebrew: Chronicles, Ecclesiastes, Ezra-Nehemiah, Song of Songs.

7 FURTHER RESEARCH

In our research, we used a selection of the 39 Old Testament books, mainly those of which there is a consensus of the main corpus to which it belongs and language use. Speaking of consensus is only possible here because both advocates and opponents of the diachronic approach acknowledge the main EBH (or SBH) and LBH (or PBH) corpora. It would be interesting, however, to train the model on more ambiguous books to see what kind of clustering appears. An challenge to such a project might be the small size of some books, especially of the twelve Minor Prophets, which probably will not provide enough information for RNNs to train well. However, there are some books that we left out that do provide enough data for a model to train on. Thus, expanding the data set with more books could be an interesting follow up on this research.

Since the RNNs proved to be an effective method for clustering historical texts according to their linguistic similarities, analysing different historical corpora could be of importance. In the field of Classical Hebrew the Dead Sea Scrolls (as a corpus somewhat larger than the complete Hebrew Bible) and the Rabbinic literature come to the fore. Finally, there is no reason to assume that the method we have tested in our pilot project on Hebrew texts would work less effective on other ancient Semitic languages.

REFERENCES

[1] Wilhelm Gesenius. 1815. *Geschichte der hebräischen sprache und schrift: Eine philologisch-historisch einleitung in die sprachlehren und wörterbücher der hebräischen sprache.* FCW Vogel.

[2] Wilhelm Gesenius. 1902. *Wilhelm Gesenius' Hebräische Grammatik*. Vol. 1. FCW vogel.

[3] S. Hochreiter and J. Schmidhuber. 1997. Long short-term memory. *Neural computation* 9, 8 (1997), 1735–1780.

[4] A. Hornkohl. 2017. All Is Not Lost: Linguistic Periodization in the Face of Textual and Literary Pluriformity. *Advances in Biblical Hebrew Linguistics* (2017), 53.

[5] A. Hurvitz. 2000. Can Biblical Texts Be Dated Linguistically? Chronological Perspectives in the Historical Study of Biblical Hebrew. *Lemaire, A. and M. Sæbo (eds.) Congress Volume, Oslo 1998* (2000).

[6] D. Kim. 2013. *Early Biblical Hebrew, Late Biblical Hebrew, and linguistic variability: a sociolinguistic evaluation of the linguistic dating of biblical texts.* Koninklijke Brill Nv.

[7] E.Y. Kutscher. 1982. *A History of the Hebrew language.* Magnes Press, The Hebrew University.

[8] Wido van Peursen. 2019. A Computational Approach to Syntactic Diversity in the Hebrew Bible. *Journal of Biblical Text Research* 44 (2019), 237–253.

[9] Wido van Peursen. 2019. Tracing Text Types in Biblical Hebrew. (2019).

[10] R. Polzin. 1976. *Late Biblical Hebrew: toward a historical typology of Biblical Hebrew prose.* BRILL.

[11] Robert Rezetko and Ian Young. 2014. *Historical linguistics and Biblical Hebrew: steps toward an integrated approach.* Vol. 9. Society of Biblical Lit.

[12] M.F. Rooker. 1994. Diachronic Analysis and Features of Late Biblical Hebrew. *Bulletin for Biblical research* 4 (1994), 135–144.

[13] D. Roorda. 2016. Text-Fabric.

[14] D.J. Shepherd and C.J.H. Wright. 2018. *Ezra and Nehemiah.* Wm. B. Eerdmans Publishing.

[15] E.P. van de Bijl, C. Kingham, W.T. van Peursen, and S. Bhulai. 2018. A Probabilistic Approach to Syntactic Variation in Biblical Hebrew.

[16] L. van der Maaten and G. Hinton. 2008. Visualizing data using t-SNE. *Journal of machine learning research* 9 (2008), 2579–2605.

[17] W.T. van Peursen and C. Kingham. 2018. The ETCBC Database of the Hebrew Bible.

[18] A.J.C. Verheij. 2000. *Bits, Bytes, and Binyanim. A Quantitative study of Verbal Lexeme Formations in the Hebrew Bible.* Uitgeverij Peeters en Department Oosterse Studies Leuven.

[19] P.J. Werbos. 1990. Backpropagation through time: what it does and how to do it. *Proc. IEEE* 78, 10 (1990), 1550–1560.

[20] I. Young. 2005. Biblical Texts Cannot Be Dated Linguistically. *Hebrew Studies* 46 (2005).

[21] Ian Young. 2017. *Linguistic Dating of Biblical Texts: Volume 2.* Routledge.

[22] I. Young and R. Rezekto. 2008. *Linguistic Dating of Biblical Texts: Volume 1.* Routledge.