The Development of Hebrew in Antiquity – A Computational Linguistic Study*

Hallel Baitner[†], **Dimid Duchovny**[‡]
Tel Aviv University

Lee-Ad Gottlieb[§] **Amir Yorav**[¶]
Ariel University Afeka College

Nachum Dershowitz[‖], **Eshbal Ratzon**[**]
Tel Aviv University

Abstract

The linguistic nature of Qumran Hebrew (QH) remains a central debate in the study of the Dead Sea Scrolls (DSS). Although some scholars view QH as an artificial imitation of Biblical Hebrew (BH), others argue that it represents a spoken dialect of ancient Judea.

The present study employs computational linguistic techniques, clustering, classification, and machine learning, to analyze the relationship of QH with Biblical and Mishnaic Hebrew. Preliminary findings confirm existing scholarly conclusions regarding the linguistic affinity of certain texts. This demonstrates that our methodology has a fundamental capacity to identify linguistic relationships. They also contribute new leads, on which we are now working to refine and enhance our analytical methods so as to provide founded insights into the historical development of Hebrew and the process of DSS textual composition.

1 Introduction

The study of Qumran Hebrew (QH) has long attracted scholars because of its linguistic complexity. Early analyses revealed QH's dual nature: It shares features with Biblical Hebrew (BH), while also displaying unique traits that align with later forms such as Mishnaic Hebrew (MH) and Samaritan Hebrew. This intricate blend has sparked an ongoing debate about QH's origins and its place in the historical development of the Hebrew language. This project aims to leverage computational language tools to deepen our understanding of QH, clarify its relationship to other Hebrew dialects, and refine the relative dating of specific scrolls within the corpus of the Dead Sea Scrolls (DSS).

2 The Nature of Qumran Hebrew

Initial scholarly evaluations of QH highlighted both its association with BH and its inclusion of linguistic traits found in later Hebrew forms. Scholars faced the challenge of explaining this duality in a comprehensive way. The predominant view, led by scholars such as Yalon (1967), Kutscher (1974), and Blau (2000), posits that QH represents a literary attempt to replicate BH. They argue that due to the cessation of BH as a living language before the composition of the DSS, this endeavor was only partially successful, allowing contemporary Hebrew features to penetrate. Some of these features are also known from MH. These scholars advocate for focusing on these contemporary linguistic features subtly embedded within QH to reconstruct the historical development of Hebrew during this period.

In contrast, scholars such as Ben-Ḥayyim (1958), Morag (1988), Rendsburg (2015), and notably Qimron (1992, 2018) propose a different model. They argue that QH authentically represents a spoken Hebrew dialect prevalent in ancient Judea. They position QH as a natural continuation of Late Biblical Hebrew (LBH), suggesting these are sequential points along the historical continuum of Hebrew language development. Qimron challenges the notion of shared morphological features between QH and MH, emphasizing their differences and proposing that MH originated from an unidentified Hebrew dialect in the Galilee, rather than from the DSS.

The scholarly debate thus centers on the inter-

*This research was funded in part by the Tel Aviv University Center for AI and Data Science and by the European Union (ERC, MiDRASH, Project No. 101071829). Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

[†]hallel.baitner@mail.huji.ac.il
[‡]dimidd@gmail.com
[§]leead@ariel.ac.il
[¶]amiryorav@gmail.com
[‖]nachumd@tauex.tau.ac.il
[**]eshbal@gmail.com

Proceedings of the Second Ancient Language Processing Workshop associated with NAACL 2025, pages 65–70
May 4, 2025.

pretation rather than the validity of the evidence. Scholars generally agree on the affinity between QH and LBH, as well as the shared lexical features between QH and MH. This situation underscores the need to expand and deepen comparative analyses of QH against both LBH and MH to provide new evidence regarding the relationships between these dialects. A global quantitative analysis, in addition to qualitative assessments and specific examples, will offer a more comprehensive understanding of these linguistic relationships. Utilizing digital analysis tools promises significant contributions to this discussion. In addition, while the majority of scholarship addresses the language of the scrolls as a whole, only limited research focuses on the distinctive language of specific scrolls, such as Kutcher's work on the Isaiah Scroll and Qimron's on 4QMMT (*Miqsat Ma'ase ha-Torah*). As the composition of the DSS is dated to a period of several centuries, we find this path of research to be promising.

3 Computational Linguistics for Hebrew

Before detailing our methodology, it is crucial to review past attempts to use computational linguistic tools for Hebrew text analysis. Early efforts focused on natural language processing (NLP) methods, requiring researchers to create morphological or syntactic descriptions for computers. Later, the field adopted machine-learning techniques, enabling computers to learn data descriptions from large training sets automatically.

Several tools have been adapted for Hebrew tasks, including automated transliteration, root identification, and opinion extraction. Notably, Santacruz (2017) used a bidirectional long-short term memory (LSTM) network to differentiate between Hebrew and Aramaic words. Similar techniques were used by HaCohen-Kerner et al. (2010) to classify Hebrew documents by historical period and ethnic origin, achieving high success rates. Liebeskind and Liebeskind (2020) further refined this approach, using more advanced techniques like recurrent neural networks and convolutional neural networks to differentiate between texts from different centuries. Koppel et al. (2011) and Yoffe et al. (2023) applied NLP methods to computerized source criticism of Biblical texts, focusing on identifying and distinguishing between different source materials within the Bible. Fono et al. (2024) used transformer-based models to reconstruct ancient Hebrew and Aramaic inscriptions, trained on the Hebrew Bible. Additionally, Dicta's Tiberias tool[1] applies modern machine learning to Bible datasets (though not to the DSS), providing stylistic comparisons and classifications based on detailed syntactic and morphological information.

Van Hecke (2018) and Van Hecke and de Joode (2021) explore the use of computational stylometric techniques to analyze BH texts and the DSS, highlighting the methodological challenges and the potential to identify distinct authors and textual variations.

4 Approach, Methods and Goals

The linguistic material we have used is based on the linguistic analysis provided by *Accordance*,[2] which includes annotated texts from ancient Hebrew works. We have developed a method to organize the linguistic data from these databases into standardized tables, facilitating computational analysis. Many compositions from the Dead Sea Scrolls have survived only in fragmentary form. We accept the scholarly decisions made in this dataset regarding doubtful letters, but our data is based solely on preserved ink, excluding reconstructions.

Our study involves two distinct types of clustering tasks: general clustering based on overall linguistic features and clustering based on specific morphological criteria. We began with a general clustering analysis of the three corpora based on word frequency. We converted each biblical book, scroll, and mishnaic tractate into a vocabulary vector, a mathematical representation of its lexical profile based on the frequency of word lemmas. To compare the compositions, we sequentially employ the following statistical approaches:

- Raw frequency analysis. Each document of the corpus is represented by a vector: With each word (or more precisely, lemma) of the entire corpus, we associate the same unique coordinate of the document vectors, and so the vector lengths are precisely the number of unique lemmas in the corpus. A document vector v contains in its i-th coordinate the number of occurrences in the document of the corresponding lemma.

- TF–IDF (term frequency–inverse document frequency). The raw vector is then normalized

[1] https://tiberias.dicta.org.il
[2] https://www.accordancebible.com

using this method, which reduces the weight of common words while emphasizing unique terms in each book. The TF term for document vector v and coordinate i is $v(i)$ divided by the total number of words in the document (i.e. $\sum_j v(j)$). The IDF term for coordinate i in all document vectors is the logarithm of the percentage of documents containing the corresponding lemma. We normalize the value $v(i)$ to be its TF value, multiplied by the IDF value of coordinate i (i.e. TF$*$IDF).

- Cosine similarity. Having computed the representative normalized vector for each document, we can then measure their similarity. For a pair of document vectors v, w, their similarity value is given as

$$\frac{\sum_i v(i) * w(i)}{\sqrt{\sum_i v(i)^2} * \sqrt{\sum_i w(i)^2}}$$

For clustering, we use hierarchical clustering with the Ward method, which groups texts based on lexical similarity while minimizing variance within clusters. The results are visualized as dendrograms, where proximity between texts indicates linguistic similarity.

In addition to general clustering, we focus on two specific morphological criteria: (1) the distribution of verb stems (*binyanim*), as previous research has shown shifts in stem usage across different periods of Hebrew (Fassberg, 2001), and (2) verbal valency patterns, which capture variations in the complements verbs can take. To analyze *binyanim*, our algorithm calculates the percentage distribution of each stem relative to the total number of verbs in each text. We then compute the Euclidean distance between these distributions across different texts, identifying those with the smallest inter-distribution distances as the most similar in stem usage. This methodology will be further refined as the research progresses.

To analyze valency, our algorithm systematically processes each verb, inspecting up to four subsequent words to determine whether it is followed by a prepositional particle, an object marker, or a pronominal suffix. Results are stored with detailed morphological attributes, enabling a structured comparison of valency patterns across texts and offering insights into syntactic shifts in Hebrew over time. This method is not yet perfect. In a sample review of the results, compared to a manual examination of the occurrences of the given verb, we observed that some complements were either not covered or incorrectly identified. However, the distribution of the various complements provides a sufficiently accurate representation of their actual occurrence. We will continue working to improve this algorithm.

Beyond clustering, our goal is to train machine learning models on the Hebrew Bible and the Mishnah to identify distinct linguistic features of Classical Biblical Hebrew (CBH), LBH, and MH. Special attention is given to distinguishing literary genres within these corpora to enhance the precision of linguistic classification.

For dialect classification, we aim to leverage recent deep learning models such as ELMo, BERT, XLNet, and RoBERTa, integrating expert knowledge of Hebrew morphology and syntax into statistical learning frameworks. These models, pretrained on large corpora and fine-tuned for specific tasks, will be validated against traditional classification algorithms using metrics such as accuracy, precision, recall, F_1-score, and clustering coherence measures like silhouette score and adjusted Rand index. Once the classifier is trained, it will be applied to the DSS to assess linguistic affinity with CBH, LBH, or MH. Special considerations include handling biblical quotations and multiple manuscript versions, ensuring that linguistic features are analyzed independently for each text. To account for textual transmission variations, we will compare rewritten or paraphrased biblical texts separately from non-biblical compositions, assessing linguistic deviations from the original biblical material and applying normalization techniques where necessary.

Since the data on which we relied to build our data set was taken from *Accordance*, it cannot be published without permission. However, the scripts we developed for data extraction will be released at the end of the project, enabling researchers to replicate our experiments

5 Preliminary Results

The clustering analysis of three major ancient Jewish textual corpora—the Hebrew Bible, the Mishnah, and the Dead Sea Scrolls—revealed nuanced insights into their linguistic and stylistic structures. The algorithm identified patterns that align with previously observed textual groups, such as the grouping of biblical books (e.g., 1 & 2 Samuel,

1 & 2 Kings, 1 & 2 Chronicles). The Mishnah's tractates generally stood out as a separate cluster. However, in an experiment conducted using the raw frequency model, the tractates *Tamid*, *Middot*, and *Yoma* distinctly differed from the rest of the Mishnah and showed a greater affinity with Qumranic compositions such as the Temple Scroll and Pseudo-Jubilees. This finding aligns well with Mishnah research, which has identified *Tamid*, *Middot*, and *Yoma* as among the earliest tractates (Epstein, 1957).

Additionally, the fragmentary copies of *Miqsat Ma'ase ha-Torah* exhibited, according to the tf–idf model, a closer linguistic proximity to Mishnaic tractates than to any other Qumranic composition (see Figure 1). This finding is consistent with prior research on this text, which has highlighted its distinctive language—deviating from the typical Qumranic linguistic style and resembling Rabbinic Hebrew more closely (Mizrahi, 2020).

Regarding the relationships among Qumranic compositions, further research is required. Preliminary results indicate, on the one hand, a clear affinity between texts such as the *Hodayot* and 4Q511 (The Song of the *Maskil*), as noted in previous studies (Angel, 2012). At the same time, unexpected connections emerged, such as the affinity between the Temple Scroll and a fragment from the Book of Jubilees (4Q219).

Future research should investigate the extent to which content and genre influence the clustering of these texts and strive to develop methodologies that minimize such biases as much as possible.

The analysis of verb stem distribution is still in its early stages. As expected, a close linguistic affinity was observed between related biblical books (e.g., 1 & 2 Samuel). However, other results indicate unexpected connections between compositions whose language appears to be significantly different. These findings require further investigation, and it may be necessary to integrate verb stem distribution data with additional types of linguistic analysis to refine the methods for identifying linguistic affinities between texts.

Valency patterns analysis is also still ongoing. Initial findings indicate distinct patterns in verb complement diversity. Some verbs display clear distributional tendencies, and certain books exhibit marked preferences for specific valency structures. For example, the algorithm successfully identified the various complements of the verb *byn* (*hiphil* stem, "to understand") and correctly detected the tendency of certain biblical books—such as Nehemiah, Daniel, and Chronicles—to use the preposition *b-* as a complement, in contrast to other biblical texts, such as Psalms and Proverbs, which regularly use a pronominal suffix, a direct object, or the preposition *l-*. Future analyses will compare the distribution of valency patterns across different works and corpora, further refining our understanding of verb usage in ancient Hebrew.

Figure 2 presents the normalized distribution of the complements of the verb *byn* in the *hiphil* stem across different books of the Hebrew Bible. The y-axis represents the relative distribution of each complement, while the x-axis lists the biblical books. The various colors indicate different complements attached to the verb, as shown in the legend.

We have similar graphs for 810 different verbs (where "verb" refers to a specific root in a particular stem), allowing us to quickly map the diversity of valency patterns for each verb.

6 Conclusion

This study employs an innovative combination of general clustering, morphological-based clustering, and machine learning techniques to investigate the linguistic landscape of the Dead Sea Scrolls. Our research aims to establish Qumran Hebrew's position within the broader development of ancient Hebrew, while providing new methodologies for the relative dating of scrolls based on linguistic features. By identifying previously unnoticed shared linguistic patterns among dialects and developing a chronological scaling of scrolls from the Hellenistic period to 70 CE, we seek to uncover potential literary connections between scrolls based on linguistic affinity. Our algorithmic approach reveals clusters of texts that share linguistic features with pre- and post-Qumranic corpora, suggesting possible social or chronological commonalities. This methodological framework not only deepens our understanding of Hebrew linguistic development but also contributes significantly to broader discussions on diachronic and dialectal variations in ancient Hebrew.

References

Joseph L. Angel. 2012. Maskil, community, and religious experience in the *Songs of the Sage* (4Q510–511). *Dead Sea Discoveries*, 19:1–27.

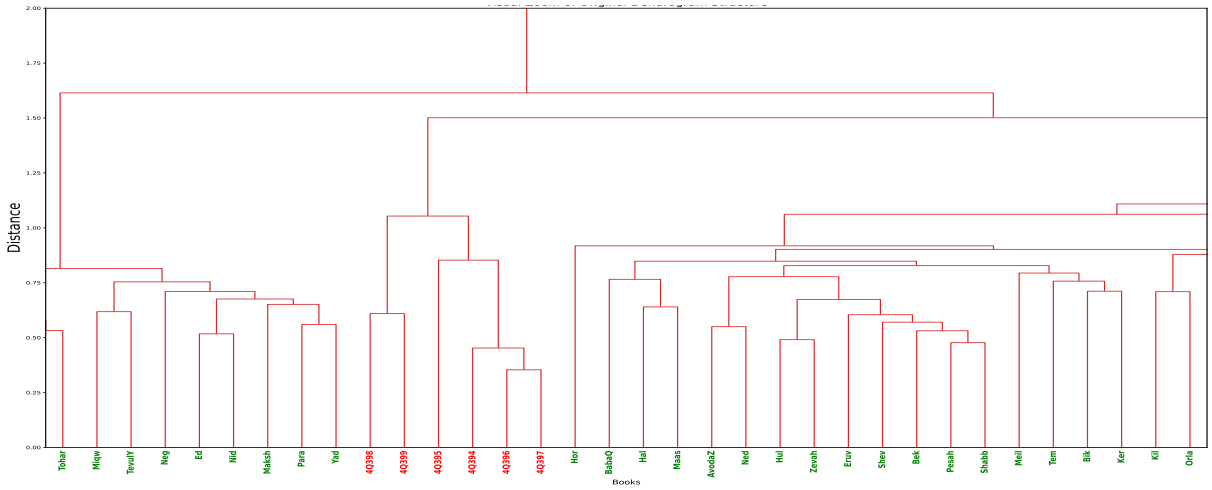

Figure 1: A small section of the dendrogram from the clustering analysis performed using tf–idf, illustrating the affinity between *Miqsat Ma'ase ha-Torah* (4Q394–399, labeled in red) and Mishnah tractates (green). The x-axis arranges the different texts; the level of the common ancestor on the y-axis indicates the degree of affinity.

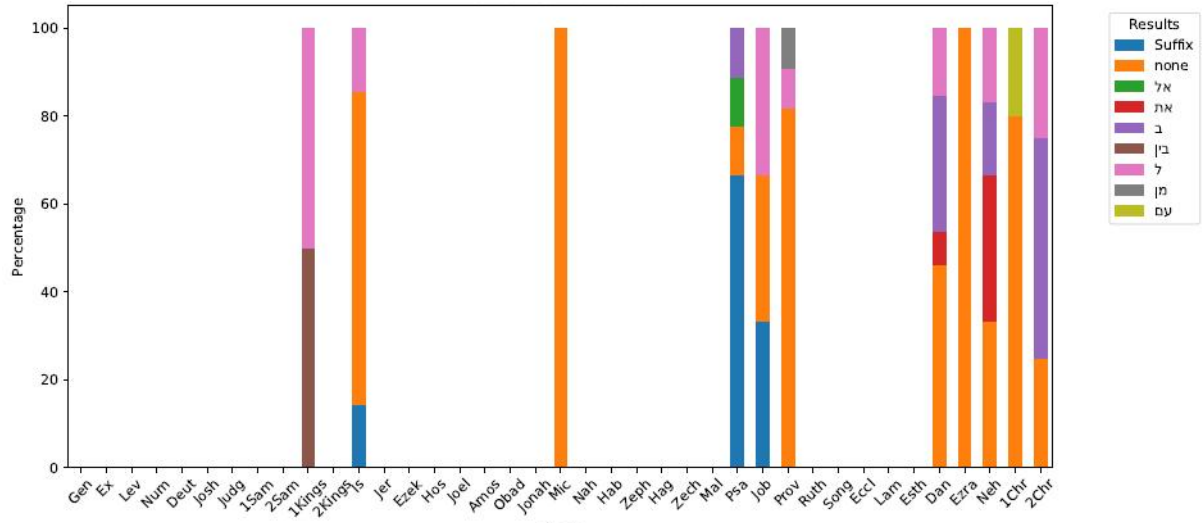

Figure 2: The normalized distribution of the complements of the verb *byn* (in *hiphil*).

Ze'ev Ben-Ḥayyim. 1958. Traditions in the Hebrew language, with special reference to the Dead Sea Scrolls. *Scripta Hierosolymitana*, 5:200–214.

Joshua Blau. 2000. A conservative view of the language of the Dead Sea Scrolls. In J. F. Elwolde and T. Muraoka, editors, *Diggers at the Well: International Symposium on the Hebrew of the Dead Sea Scrolls and Ben Sira*, pages 20–25. Brill, Leiden.

Yaakov N. Epstein. 1957. *Introductions to Tannaitic Literature: Mishnah, Tosefta, and Halakhic Midrashim*. Magnes, Jerusalem. [Hebrew].

Steven E. Fassberg. 2001. The movement from qal to pi"el in Hebrew and the disappearance of the qal internal passive. *Hebrew Studies*, 42:243–255.

Niv Fono, Harel Moshayof, Eldar Karol, Itai Assraf, and Mark Last. 2024. Embible: Reconstruction of ancient Hebrew and Aramaic texts using transformers. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 846–852, St. Julian's, Malta. Association for Computational Linguistics.

Yaakov HaCohen-Kerner, Hananya Beck, Elchai Yehudai, and Dror Mughaz. 2010. Stylistic feature sets as classifiers of documents according to their historical period and ethnic origin. *Applied Artificial Intelligence*, 24:847–862.

Moshe Koppel, Navot Akiva, Idan Dershowitz, and Nachum Dershowitz. 2011. Unsupervised decomposition of a document into authorial components. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 1356–1364, Portland, OR. Association for Computational Linguistics.

Eduard Y. Kutscher. 1974. *The Language and Linguis-*

tic Background of the Isaiah Scroll (1QIsaa). Brill, Leiden.

Chaya Liebeskind and Shmuel Liebeskind. 2020. Deep learning for period classification of historical Hebrew texts. *Journal of Data Mining & Digital Humanities*, 2020:2.

Noam Mizrahi. 2020. The language of 4QMMT. In R. G. Kratz, editor, *Interpreting and Living God's Law at Qumran: Miqsat Ma'aśe Ha-Torah, Some of the Works of the Torah (4QMMT)*, volume 37 of *Sapere*, pages 67–83. Mohr Siebeck, Darmstadt, Germany.

Shlomo Morag. 1988. Qumran Hebrew: Some typological observations. *Vetus Testamentum*, 38:148–164.

Elisha Qimron. 1992. Observations on the history of early Hebrew (1000 B.C.E.–200 C.E.) in the light of the Dead Sea documents. In A. Rappaport and D. Dimant, editors, *The Dead Sea Scrolls: Forty Years of Research*, pages 349–361. Brill, Leiden.

Elisha Qimron. 2018. *A Grammar of the Hebrew of the Dead Sea Scrolls*. Yad Yizhak Ben-Zvi, Jerusalem.

Gary A. Rendsburg. 2015. The nature of Qumran Hebrew as revealed through Pesher Habakkuk. In E. Tigchelaar and P. Van Hecke, editors, *Hebrew of the Late Second Temple Period: Proceedings of a Sixth International Symposium on the Hebrew of the Dead Sea Scrolls and Ben Sira*, pages 132–159. Brill, Leiden.

Noah Santacruz. 2017. PSHAT – part of speech handling for Aramaic in the Talmud. Master's thesis, The Cooper Union for the Advancement of Science and Art, New York, NY.

Pierre Van Hecke. 2018. Computational stylometric approach to the Dead Sea Scrolls: Towards a new research agenda. *Dead Sea Discoveries*, 25:57–82.

Pierre Van Hecke and Johan de Joode. 2021. Promises and challenges in designing stylometric analyses for Classical Hebrew. In S. Fassberg, editor, *Hebrew texts and Language of the Second Temple Period*, pages 349–374. Brill, Leiden.

Hanoch Yalon. 1967. *Studies in the Dead Sea Scrolls: Philological Essays (1949–1952)*. Shrine of the Book/America-Israel Cultural Foundation/Kiryath Sepher, Jerusalem.

Gideon Yoffe, Axel Bühler, Nachum Dershowitz, Thomas Romer, Eli Piasetzky, Israel Finkelstein, and Barak Sober. 2023. A statistical exploration of text partition into constituents: The case of the Priestly source in the books of Genesis and Exodus. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1918–1940, Toronto, Canada. Association for Computational Linguistics.