Computational Detection of Intertextual Parallels in Biblical Hebrew

A Benchmark Study Using Transformer-Based Language Models

David M. Smiley

University of Notre Dame

dsmiley@nd.edu

Abstract

Identifying parallel passages in biblical Hebrew is foundational in biblical scholarship for uncovering intertextual relationships. Traditional methods rely on manual comparison, which is labor-intensive and prone to human error, particularly with the linguistic complexity of a morphologically rich language like biblical Hebrew. This study evaluates the potential of pre-trained transformer-based language models, including E5, AlephBERT, MPNet, and LaBSE, for detecting textual parallels in the Hebrew Bible. Focusing on known parallels between the books of Samuel/Kings and Chronicles, I assessed each model's capability to generate word embeddings that delineate parallel from non-parallel passages. Utilizing cosine similarity and Wasserstein Distance measures, I found that E5 and AlephBERT show significant promise, with E5 excelling in parallel detection and AlephBERT demonstrating stronger non-parallel differentiation. These findings indicate that pre-trained models can enhance the efficiency and accuracy of detecting intertextual parallels in ancient texts, suggesting broader applications for ancient language studies.

1 Introduction

[1] The identification of parallel passages in biblical Hebrew (BH) texts has long been a cornerstone of biblical scholarship, serving as a critical tool for understanding intertextual relationships and theological development (Fewell, 1992). Scholars have traditionally relied on manual comparisons to trace these connections (Harvey, 2004; Miller, 2011; Schnittjer, 2021). These parallels are not merely textual repetitions but often involve reinterpretations and recontextualizations that provide profound insights into the evolution of biblical narratives (Fishbane, 1985; Kalimi, 2004).

---

[1]All code and data can be found in the following Github repository: `https://github.com/dmsmiley/detect-bh`

Traditional methods for detecting these parallels, however, are labor-intensive and susceptible to human error, particularly when the connections are implicit, nuanced, or even subtle. The linguistic complexity of BH, with its rich morphology and intricate verbal structures, exacerbates these challenges (Singh et al., 2012). These difficulties are further magnified when scholars manually attempt to analyze texts spanning extensive chronological, thematic, and theological corpora, which push the limits of one's attention and lead to an inconsistency in intertextual findings.

Recent advances in natural language processing (NLP) present new possibilities for addressing these limitations. Large language models (LLMs), particularly transformer-based architectures, excel at capturing semantic relationships across vast datasets (Devlin et al., 2019). These models process large amounts of text with high accuracy, generating rich word embeddings that reflect nuances of meaning and context in a way that traditional methods cannot.

In this study, I evaluate the potential of pre-trained transformer models, such as E5, AlephBERT, MPNet, and LaBSE to aid detection of parallel passages in the Hebrew Bible (HB). By focusing on known parallels that the book of Chronicles (Chr) reuses from the earlier narratives of Samuel (Sam) and Kings (Kgs), I assess each models' ability to generate word embeddings that accurately delineate known parallel and non-parallel verses in the HB. This research not only seeks to provide benchmarks of model accuracy for tasks in textual similarity, but it also lays the groundwork for future NLP applications in biblical studies and can be applied to other ancient literature more broadly.

2 Previous Computational Attempts at Biblical Hebrew Parallels

Various computational methods have been applied to the challenge of identifying parallels in BH, though they have often been constrained by the tools and techniques available at the time. Notably, van Peursen and Talstra (2007) sought to systematically compare parallel texts from 2 Kgs, Isaiah, and Chr using a computer-assisted synopsis. Their approach highlights the difficulty in establishing a rule-based method for identifying parallel texts. Criteria of what counts for a parallel are often subjective and proper parameters are opaque.

In recent works, Shmidman, Koppel, and Porat (2018) have significantly refined rule-based approaches for detecting parallels in large Hebrew and Aramaic corpora. Their algorithms are designed to detect approximate matches, which account for rephrasing, orthographic differences, and interpolations. This method has proven effective in identifying textual reuse in complex corpora like the Babylonian Talmud, where near-identical passages often exhibit minor discrepancies.

Shmidman (2022) further refined his work by creating a hashing algorithm that breaks each lexeme into the smallest unit for word representation. However, as Shmidman acknowledges, this method is susceptible to numerous false positive matches. The inherent limitation of rule-based methods lies in their inability to capture deeper semantic relationships, often missing the broader contextual nuances in rewritten texts like those in Chr. This underscores the challenges faced by rule-based approaches in understanding the complex semantic

landscape of ancient texts.

These frequency- and rule-based approaches rely too heavily on surface-level lexical matching rather than more sophisticated linguistic representations (Mars, 2022). Furthermore, parallel passages are often not exact copies of previous texts, meaning that passages with significant omissions, additions, or rewrites, such as those in Chr, are not accurately captured by these methods. This highlights the need for a method(s) that can effectively capture semantics rather than solely being dependent on superficial matching.

Type of Textual Difference	Sam/Kgs Reference	Sam/Kgs Text	Chr Reference	Chr Text
Addition and Name Changes	2 Sam 5:6	The king and his men went to Jerusalem against the Jebusites, the inhabitants of the land.	1 Chr 11:4	David and all of Israel went to Jerusalem, which was Jebus. There the Jebusites were the inhabitants of the land.
Omission	2 Sam 6:15	David and all of the house of Israel brought up the ark of the covenant. . .	1 Chr 15:28	[ ] All of Israel brought up the ark of the covenant. . .
Addition and Omission	2 Kgs 11:4	Jehoiada. . . took the officers of one hundred from the Carites and the guard, and brought them to him in the house of YHWH, and he cut a covenant with them.	2 Chr 23:1	Jehoiada. . . took the officers of one hundred, Azariah son of Jeroham, Ishmael son of Jehohanan, Azariahu son of Obed, Maasieah son of Adaiahu, and Elishaphat son of Zichri into a covenant agreement with him.
Omission	2 Sam 7:14	I will be like a father to him, and he will be like a son to me. Whenever he goes astray, I will punish him with the staff of men and with the beatings of the songs of man.	1 Chr 17:13	I will be like a father to him, and he will be like a son to me. [ ]

Table 1: Examples from Kalimi's work (2004) on the rewriting patterns employed by the author of Chr to reshape earlier parallel passages.

Recent advances, particularly in transformer-based models, offer a solution to problem parallels, like those shown above. Unlike frequency-based methods, transformers can generate embeddings at the verse level based on the relationship of a word in a given context,

thus capturing the semantic variation (Vaswani et al., 2017). This ability to contextualize words semantically makes them particularly suited to uncovering parallel passages in BH.

Despite their potential, no systematic evaluation has been conducted to benchmark these modern language models on BH texts. While models like AlephBERT (Seker et al., 2021) have been trained on vast amounts of modern Hebrew (MH), their effectiveness on ancient Hebrew remains untested. The linguistic differences between MH and BH, combined with the absence of BH-specific language models, create a significant gap in our understanding of how well these pre-trained models can handle ancient language tasks.

This study addresses this gap by providing the first comprehensive benchmark of transformer-based models for detecting biblical parallels. By leveraging known parallels between Sam/Kgs and Chr, I systematically evaluate how different pre-trained LLMs perform when creating embeddings for BH verses. This benchmark study not only advances NLP techniques for ancient texts but also provides crucial empirical data for scholars seeking to apply these powerful tools to biblical scholarship without needing to develop models from scratch.

3   Objectives

The primary objective of this study is to assess the effectiveness of transformer-based language models in identifying parallel passages within BH texts. Specifically, I will evaluate how well pre-trained models such as E5, AlephBERT, MPNet, and LaBSE can capture the relationships between parallel texts, particularly those found between the books of Sam/Kgs and Chr.

By comparing how well these models create word embeddings that can distinguish true parallel passages from non-parallels, I can estimate their suitability for finding unknown connections which have traditionally relied on manual methods in biblical scholarship. In doing so, I can demonstrate how transformer-based models can enhance the accuracy and efficiency of detecting intertextual parallels in ancient texts. In addition to assessing the strengths and limitations of each model, this study contributes to the broader effort to integrate advanced NLP techniques into the field of digital humanities. Our findings offer an empirical method for detecting previously unknown parallel and intertextual connections on historical texts.

4   Data and Models

4.1   Dataset

The Hebrew text for this study is drawn from the Biblia Hebraica Stuttgartensia Amstelodamensis (BHSA) corpus, as compiled by the Eep Talstra Centre for Bible and Computer at Vrije Universiteit Amsterdam (Peursen et al., 2015). Additionally, there are 559 verses from Chr with recognized, known parallels in Sam/Kgs (Endres et al., 1998), ensuring a consistent and empirical set of passages for evaluating transformer-based models' performance.

4.2 Pre-Trained Models

Five pre-trained transformer models were used, each selected for their strengths in text classification and documentary similarity. Some are optimized for their scale and multilingual capabilities like Multilingual E5 (Wang et al., 2024). Another has its strength in representing larger sentences and paragraphs such as MPNet (Song et al., 2020). LaBSE is a language agnostic approach to embeddings (Feng et al., 2022). In contrast the final model, Aleph-BERT, is specifically pre-trained on MH (Seker et al., 2021). By selecting robust, diverse models I am able to observe which approaches generalize embeddings best for the task of text similarity in BH.

5 Model Evaluation

5.1 Cosine Similarity

Cosine similarity is the preferred measure for evaluating vector proximity in text similarity tasks by calculating the cosine of the angle between two vectors in high-dimensional space. This score is particularly useful for analyzing embeddings generated by LLMs, as it captures both syntactic and semantic relationships between words (Gomez et al., 2022).

In the context of BH, parallel passages often display variations in word choice, spelling, and theological interpretation, as shown above (Kalimi, 2004). So a later account is not always a simple "copy and paste" version of its earlier parallel text, which is why frequency-based methods are insufficient to account for finding parallels. By representing each passage as a vector, cosine similarity quantifies the semantics of a text, regardless of superficial additions, omissions, or spelling differences.

For this study, cosine similarity was computed in two ways. First, each passage in Chr was compared to its known parallel in the books of Sam/Kgs, generating a cosine similarity score:

$$\text{Cosine}(v_i^{\text{Chr}}, v_j^{\text{Sam/Kgs}})$$

Second, each passage from Chr was compared to every other verse in Sam/Kgs that is not a known parallel to compute the mean non-parallel cosine similarity:

$$\mu_{\text{NonParallel}}(v_i^{\text{Chr}}) = \frac{1}{N} \sum_{j=1}^{N} \text{Cosine}(v_i^{\text{Chr}}, v_j^{\text{Sam/Kgs}}), \quad \text{where } v_j \notin \text{Parallels}(v_i)$$

This approach evaluates how well the models detect true parallels and their ability to avoid false positives by assigning high similarity to unrelated passages.

5.2 Parallel and Non-Parallel Verse Cosine Similarities

The table below summarizes the performance of the pre-trained models, including the mean cosine similarity for parallel and non-parallel passages, and the statistical significance (p-value) between the differences of the mean cosine for parallel and non-parallel texts. The Appendix contains the histogram of the parallel cosine similarity scores for each model.

Model	Mean Cosine (Parallel)	Mean Cosine (Non-Parallel)	P-value	% Cosine $\geq$ 95%	% Cosine $\geq$ 98%
E5	0.966	0.882	1.06×10^{-284}	75.0%	40.11%
AlephBERT	0.914	0.638	9.98×10^{-314}	44.78%	17.27%
MPNet	0.903	0.649	1.61×10^{-231}	46.76%	25.72%
LaBSE	0.828	0.375	2.48×10^{-270}	26.07%	11.33%

Table 2: Cosine similarity, including p-value from t-test of the difference between the mean cosine similarity scores.

E5 consistently outperforms the other models, achieving a high mean cosine similarity of 0.966 for parallel passages, with 75.0% of passages exceeding a 95% similarity. However, its high similarity score for non-parallel passages (0.882) suggests an inability to properly delineate between the parallels and non-parallels. AlephBERT, while designed for Hebrew, underperforms relative to E5, with a mean cosine similarity of 0.914 for parallel passages. Only 17.27% of parallels exceeded 98% similarity. However, AlephBERT demonstrates stronger separation between parallel and non-parallel passages (mean cosine for non-parallels: 0.638), suggesting that it might be properly fitting better than E5.

MPNet and LaBSE exhibit the weakest performance with lower mean cosine similarities for parallel passages (0.903 and 0.828, respectively) and less distinct separation between parallel and non-parallel passages. This likely stems from their lack of optimization for Hebrew, which affects their ability to capture the linguistic nuances of the text. A t-test confirms that the observed differences between mean parallel and non-parallel cosine similarities are significant across all models, with p-values below 1e-100. For instance, E5's p-value of 1.06e-284 indicates a highly significant distinction between the mean cosine score for parallel and non-parallel texts. So the models are at least able to consistently achieve similarity scores that distinguish parallel and non-parallel passages, despite not being optimized on BH.

5.3   Distribution-Based Testing for Parallel and Non-Parallel Verses

To further assess model performance a distribution-based metric called the Wasserstein Distance was employed, which measures the separation between the cosine similarity distributions for parallel and non-parallel passages. A higher Wasserstein Distance indicates greater separation and a stronger proclivity to distinguish between parallel and non-parallel verses (Leo et al. 2023).

The results reveal important insights about model characteristics. E5 exhibits the lowest Wasserstein Distance (0.0812), indicating significant overlap between the cosine similarity distributions for parallel and non-parallel passages. This confirms E5's tendency to assign high similarity scores for true parallels broadly; however, it does so at the expense of assigning high cosine similarity scores to false positives quite consistently, as reflected in this limited distribution separation.

In contrast, AlephBERT shows a larger Wasserstein Distance (0.2764), indicating better

Model	Wasserstein Distance
E5	0.0812
AlephBERT	0.2764
MPNet	0.2540
LaBSE	0.4532

Table 3: Wasserstein Distance scores for each model.

separation between parallel and non-parallel passages. While AlephBERT's overall cosine similarity for parallel passages is lower than E5's, its greater separation suggests that it is less prone to false positives.

MPNet and LaBSE also exhibit larger Wasserstein Distances (0.25 and 0.45, respectively), reflecting better distinction between parallel and non-parallel passages. However, their lower overall performance in mean cosine similarity limits their effectiveness for detecting true parallels. MPNet has a slightly worse, but similar performance to AlephBERT on both measures. While LaBSE has the largest Wasserstein Distance, as the next section will confirm, it may be too conservative in identifying parallels, potentially missing subtle intertextual connections.

5.4 Evaluating False Positives with Classification Metrics

Cosine similarity was also used to find the closest match for each passage, helping to identify which specific text was most similar to our parallel text in Chr. This nearest-neighbor approach evaluates whether the models are finding true parallels or capturing incorrect matches. To further assess model performance, classification reports were created, focusing on precision, recall, and F1-score. The micro average metrics (Rainio et al., 2024) gave a clear picture of how well each model distinguished between true parallels and unrelated passages based on finding the closest cosine similarity match for each verse in Chr.

Model	Recall	Precision	F1-Score
E5	0.85	0.92	0.88
AlephBERT	0.82	0.92	0.87
MPNet	0.68	0.90	0.78
LaBSE	0.72	0.86	0.78

Table 4: Classification report of the closest parallel passage according to cosine similarity.

E5 achieved a micro average precision of 0.92 and a recall of 0.85, indicating that while the majority of detected parallels were correct, around 15% of true parallels were incorrectly categorized as the closest parallel text based on cosine similarity. Additionally, E5 only has a false positive rate of 8%. This balance resulted in a strong F1-score of 0.88, demonstrating effective performance overall.

AlephBERT demonstrated similar performance, with a micro average precision of 0.92 and a recall of 0.82, resulting in an F1-score of 0.87. The metrics for AlephBERT are close

to those of E5, suggesting that both models perform similarly in identifying true parallels and avoiding false positives.

MPNet and LaBSE showed lower effectiveness overall. They performed well with F1-scores of 0.78, but still fell short of the performance levels seen with E5 and AlephBERT.

These classification metrics complement cosine similarity and Wasserstein Distance analyses by providing insights into the balance between precision and recall for each model. E5 and AlephBERT both effectively detect parallels, with similar precision and recall metrics, indicating comparable performance. MPNet and LaBSE demonstrate limited reliability, suggesting a combined approach using E5 and AlephBERT may yield the best results for identifying intertextual parallels in BH texts.

5.5   Model Conclusions

Based on the evaluation of pre-trained models, E5 and AlephBERT stand out as the most effective models for detecting true parallel passages in BH texts. E5 excels at detecting explicit parallels, achieving high mean cosine similarity for known passages. However, its performance also reveals a tendency to catch false positives, with high similarity scores for non-parallel texts, as reflected in its relatively low Wasserstein Distance. This makes E5 highly capable of identifying parallels but a bit less effective in distinguishing true from false positives.

AlephBERT, while slightly behind E5 in mean cosine similarity for parallel passages, shows a slight edge in avoiding false positives. Its lower similarity for non-parallel texts and higher Wasserstein Distance indicate better separation between related and unrelated passages. This makes AlephBERT particularly suitable when minimizing incorrect matches is a priority, providing a complementary strength to E5.

MPNet, and LaBSE showed weaker performance in identifying true parallels, largely due to their lower optimization for BH. These models exhibited lower cosine similarities and greater difficulty in distinguishing parallel from non-parallel texts. Despite their strengths, they should be avoided in their current forms as it relates to creating word embeddings for BH.

These findings suggest that using both the E5 and AlephBERT models as a check and balance of one another offers a promising solution for enhancing parallel detection accuracy. Future work should focus on fine-tuning these pre-trained models for BH to maximize their complementary strengths, potentially revolutionizing the identification of intertextual parallels, not only in BH, but also in other ancient texts. This approach enables scholars to utilize advanced NLP tools without the need to develop new models from scratch, instead focusing on optimizing and adapting existing models for specific historical and linguistic contexts.

6   Implications for Future Studies

This study demonstrates the capability of pre-trained transformer-based models to identify parallel verses in BH using document similarity measures such as cosine similarity and Wasserstein Distance. E5 and AlephBERT show significant promise in detecting these passages despite differences in lexical choice or sentence structure. Although not perfect, by

adapting these models and fine-tuning them later I can bypass the issues of attempting to build our own BH language model from scratch.

Expanding this research to other ancient languages, such as Syriac, Greek, or Latin, could provide further insights into intertextuality in other areas of ancient language study. If these models generalize well with BH, then in theory they could also work well in other languages on which they are not explicitly trained. Developing embeddings from pre-trained models for ancient texts could transform the study of literary connections in the ancient world, offering new tools for scholars in the digital humanities.

7   Limitations

While this study demonstrates the promise of pre-trained transformer-based models for detecting biblical parallels, several limitations should be acknowledged. First, this evaluation focuses exclusively on narrative texts, specifically the known parallels between Sam/Kgs and Chr. The HB encompasses numerous, variegated genres, like poetry, wisdom literature, prophetic oracles, and legal codes, each with their own ways of displaying intertextual connections. The performance of these models on something like poetic parallelism in the Psalter, the covenant code in Exodus, or prophetic sayings of Jeremiah remains untested, thus limiting the generalizability of our findings across the full spectrum of biblical literature.

Second, the embeddings generated by these pre-trained models, while effective for similarity detection, remain largely opaque and uninterpretable. This "black box" nature of all language models built on the transformer architecture prevents scholars from understanding why certain passages are deemed similar, potentially obscuring important linguistic or theological nuances that traditional philological methods would capture. This limitation is particularly significant for biblical studies, where understanding the specific mechanisms of textual reuse is often as important as identifying the parallels themselves.

Finally, this study evaluates existing pre-trained models rather than developing a BH-specific language model. While our approach demonstrates that adaptation of existing models is viable and practical, a purpose-built model trained on BH corpora might better capture the unique morphological, syntactic, and semantic features of ancient Hebrew texts, potentially improving both accuracy and interpretability. Unfortunately, the HB is a small corpus and building a language model from the ground up is probably not feasible with the data needs of transformers, although fine tuning a model like AlephBERT or E5 for textual similarity on BH should not be ruled out and could actually point toward productive avenues for future research.
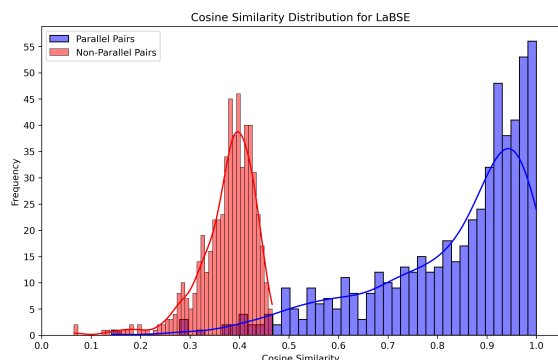
Appendix

Each graph contains the distribution of cosine similarity scores for all known parallels (in blue) alongside the mean non-parallel cosine similarity scores (in red) according to the embeddings created by the different models.
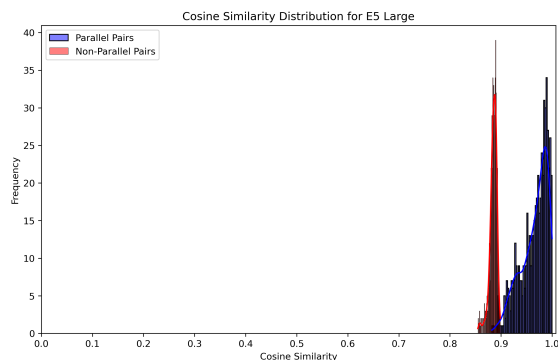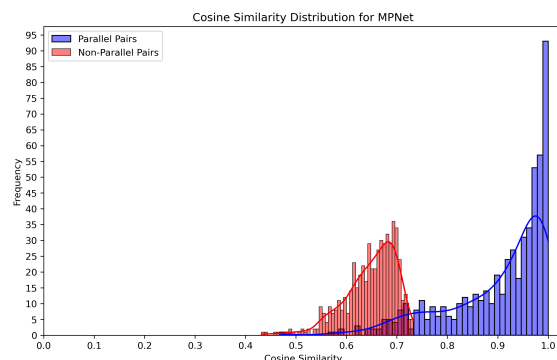

(a) E5
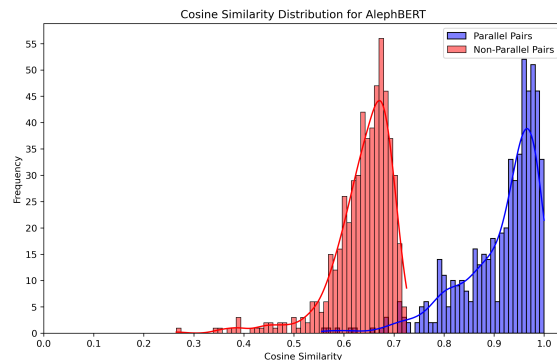
(b) AlephBERT

(c) LaBSE

(d) MPNet

References

- Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019) "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding", *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 4171–4186. `https://doi.org/10.18653/v1/N19-1423`

- Enders, J., Millar, W., & Burns, J. (1998). *Chronicles and Its Synoptic Parallels in Samuel, Kings, and Related Biblical Texts.* The Liturgical Press.

- Feng, F., Yang, Y., Cer, D., et al. (2022). Language-agnostic BERT Sentence Embedding. `https://doi.org/10.48550/arXiv.2007.01852`

- Fewell, D. (1992). *Reading between Texts: Intertextuality and the Hebrew Bible.* Louisville: Westminster John Knox Press.

- Fishbane, M. (1985). *Biblical Interpretation in Ancient Israel.* New York: Oxford Press.

- Gomez, J., and Vázquez, P. (2022). An Empirical Evaluation of Document Embeddings and Similarity Metrics for Scientific Articles. *Applied Sciences, 12*(11), 5664. `https://doi.org/10.3390/app12115664`

- Harvey, J. (2004). *Retelling the Torah: The Deuteronomistic Historian's Use of Tetrateuchal Narratives.* New York: Bloomsbury.

- Kalimi, I. (2004). *The Reshaping of Ancient Israelite History in Chronicles.* Eisenbrauns.

- Leo, J., Ge, E., and Li, S., (2023) Wasserstein Distance in Deep Learning. `http://dx.doi.org/10.2139/ssrn.4368733`

- Mars, M. (2022). From Word Embeddings to Pre-Trained Language Models: A State-of-the-Art Walkthrough. *Applied Sciences, 12*(17), 8805. `https://doi.org/10.3390/app12178805`

- Miller, G. (2011). Intertextuality in Old Testament Research. *Currents in Biblical Research, 9*, pp. 283–309.

- Rainio, O., Tueho, J., and Klén, R. (2024). Evaluation Metrics and Statistical Tests for Machine Learning. *Scientific Reports, 14*, 6086. `https://doi.org/10.1038/s41598-024-56706-x`

- Schnittjer, G. (2021). *Old Testament Use of the Old Testament.* Grand Rapids: Zondervan.

- Seker, M., Bandel, E., Barekat, D., et al. (2021). AlephBERT: A Hebrew Large Pre-Trained Language Model to Start-off your Hebrew NLP Application With. In *Proceedings of the 2022 Conference on Computational Natural Language Processing*, pp. 321–330. `https://doi.org/10.48550/arXiv.2104.04052`

- Shmidman, A., Koppel, M., and Porat, E. (2018). Identification of Parallel Passages Across a Large Hebrew/Aramaic Corpus. *Journal of Data Mining and Digital Humanities.* `https://doi.org/10.46298/jdmdh.1388`

- Shmidman, A. (2022). Automatic Identification of Biblical Citations and Allusions in Hebrew Texts. In *Jewish Studies in the Digital Age*, pp. 335-348. `https://doi.org/10.1515/9783110744828-015`

- Singh, N., and Habash, N. (2012). Hebrew Morphological Preprocessing for Statistical Machine Translation. In *Proceedings of the 16th EAMT Conference*, pp. 43-50.

- Song, K., Tan, X., Qin, T., et al. (2020). MPNet: Masked and Permuted Pre-training for Language Understanding. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, pp. 16857–16867. `https://doi.org/10.48550/arXiv.2004.09297`

- van Peursen, W., and Talstra, E. (2007). Computer-Assisted Analysis of Parallel Texts in the Bible. *Vetus Testamentum, 57*(1), pg. 45–72.

- van Peursen, W., Sikkel, C., and Roorda, D. (2015). Hebrew Text Database ETCBC4b. `https://doi.org/10.17026/dans-z6y-skyh`

- Vaswani, A., Shazeer, N., Parmar, N., et al. (2017). Attention is All You Need. In *Advances in Neural Information Processing Systems 30* (NIPS 2017), pp. 5998–6008. `https://doi.org/10.48550/arXiv.1706.03762`

- Wang, L., Yang, N., Huang, X., et al. (2024). Multilingual E5 Text Embeddings: A Technical Report. `https://doi.org/10.48550/arXiv.2402.05672`