RESEARCH PROPOSAL

Title: **Integrating linguistic-stylistic and semantic features for computerized study of authorship questions in the dead sea scrolls**

Project coordinator: Jonathan Ben-Dov, Department of Bible, Tel Aviv University & Roded Sharan, School of Computer Science, Tel Aviv University.

1. Abstract

We suggest advancing the state of the art in computerized study of Qumran Hebrew using deep learning methods that have not hitherto been applied to it. Our novel approach integrates linguistic-stylistic and statistical features with semantic features derived from pre-trained language models for building deep models for authorial classification. The integrative approach which uses cutting-edge modeling techniques is expected to yield accurate and robust clustering and classification tools. If successful, these tools will produce unprecedented results in classifying the Dead Sea Scrolls, shedding light not only on traditional authorship questions such as the distinction between sectarian and non-sectarian, but also informing border cases along this spectrum (e.g. *Instruction*), as well as different types of Hodayot, different sources of Serekh Hayahad, redaction layers in sectarian literature (M, D), and the relations between various wisdom texts (Mysteries vs. Instruction).

2. Scientific background and state of the art

One of the main characteristics of the Hebrew scrolls from Qumran (Dead Sea Scrolls, henceforth *DSS*) is a baroque-like elaboration of the biblical poetic style, that intensifies many of the grammatical and stylistic features and leads to changes in the vocabulary, morphology and syntax. This branch of ancient Hebrew was studied recently, for example in the grammar by Qimron (2018) and in a series of columns on the Hebrew of the Second Temple Period (e.g. (Fassberg, 2021)). In turn, scholars pointed out groups of documents that employ divergent language and style. These distinctions led to the establishment of common categories, such as: sectarian / non-sectarian literature, or Hodayot of the teacher vs. Hodayot of the community (although the latter are contested, (Newsom, 2021), but new divisions may be pertinent, e.g. (Harkins, 2018).

Stylometric study of biblical Hebrew and the language of the DSS emerged alongside the traditional research. Statistical methods have been carried out since the 1970s (see summary in (Starr, 2016)), while machine learning methods have recently gained momentum. (Dershowitz *et al.*, 2015) used a learning algorithm in order to enact *author recognition* in various branches of biblical Hebrew. They achieved good results in clustering Jeremiah and Ezekiel apart, and in clustering Priestly / non-priestly material in the Pentateuch. Their method is essentially based on lexical usage. However, as current linguistic study on biblical Hebrew shows, the syntactic features are no less – and probably even more – effective than the lexicon in manually clustering linguistic layers of Hebrew. The linguistic classifier Tiberias (https://tiberias.dicta.org.il/#/) uses a large variety of linguistic features including syntactic and stylistic markers. Recent scholarship (Berman, 2021) used Tiberias to examine questions of authorship in biblical research.

Starr (2016; appendix) presented a series of parameters oriented towards quantifying the richness of the language in various books of the Bible, in both lexical and syntactic aspects. His list contains 22 quantified parameters from various domains of language, such as: ratio of construct to absolute state nouns; ratio of unbound to bound pronouns; proper noun percentage; persuasive verbs (imperative, jussive, cohortative) percentage, etc. A statistical aggregation of these 22 parameters led to the elucidation of a linguistic register for each book of the Bible, while pointing out the most effective parameters. He then classified biblical books into small groups, some of them quite intriguing, e.g. the clustering of Hosea, Isaiah and Job to the same group.

The stylometric distinctions are different from the more traditional distinctions of the ages of biblical literature (CBH. LBH), because two authors in the same stage of Hebrew may use different styles. Starr's list of features is oriented towards the stylistic distinction, while for the latter question we use the latest studies: (Hendel and Joosten, 2018), (Fassberg, 2019), as well as Qimron (2018) and (Muraoka, 2020) for Qumran Hebrew.

In a statistical study of Qumran Hebrew, (Van Hecke, 2018) introduced the use of trigram letter sequences. Although this method does not relate to any intelligible feature of the text, the author demonstrates its superiority in DSS analysis over grammatical stylometric features. Van Hecke produced good clustering results, generally validating the scholarly categories of Sectarian / non-sectarian literature. He further set the main open questions that remain: is there a joint linguistic profile to the main sectarian compositions (S, Sa, D, H)? where do borderline compositions like *Instruction* and The Songs of the Sabbath Sacrifice stand with regard to that core cluster (if established)? Further work by van Hecke and de Joode (2021) added vigor to their use of trigrams, and pointed out the unique methodological problems of working with the DSS.

Construct phrases (CPs) are an essential part of any language, but they become more frequent in late (post-classical) biblical Hebrew from the times of the Second Temple, and are particularly significant in the language of the DSS. In this sectarian corpus, many (if not most) of the sectarian stock phrases are coined as CP, for example מורה הצדק, בני אור, הוית עולה, מטעת אמת, בית קודש, etc. Moreover, the Qumranic authors use a large variety of such phrases, for example embedding them in chains of 2, 3 or more CPs or linking two or more nouns on either side of the phrase. The authors thus command a robust new morphological-syntactic means of expression, which forms part of their unusual language. Since the language of the DSS community forms a conspicuous part of its identity ((Schniedewind, 2013; Van Hecke, 2018); Ben-Dov 2009), the use of CP turns out to be a marker of identity as well.

Previous studies by linguists explored the use of CP from a variety of methods, exploring the cognitive procedures practiced by speakers when a CP is deciphered as well as the scale of chained vs. free constructs with regard to establishing their lexical value (Berman 2009; Bleiboim and Shatil 2014; (Aitchison and Jackendoff, 1998); (Jackendoff and Audring, 2020)). The morphological value of CPs for Qumran Hebrew was initially explored by Talshir (2013).

Deep learning methods have become the standard in many fields, most notably in natural language processing (NLP), outperforming other machine learning approaches in fundamental tasks such as text annotation and clustering. Many current approaches to these tasks are based on pretrained models that use the BERT language model in various forms to generate semantic features and fine-tune them to a specific task at hand (Devlin (Seker *et al.*, 2022), 2018). In particular, several modern Hebrew models have been developed based on the BERT architecture and their utility and superiority were demonstrated (Seker *et al.*, 2022). However, their application to ancient Hebrew texts was so far not investigated and could present considerable challenges (Bamman, 2017). In particular, the present study seeks to pinpoint differences between rather similar documents to a very fine resolution, with neither of the corpora (biblical Hebrew, Qumran Hebrew) amply studied before to a level that allows pre-training language models.

Text classification is one of the fundamental problems in NLP with state of the art methods employing transductive learning methods such as graph convolutional networks (GCNs) to make use of both labeled and unlabeled examples in the training process (Yao, Mao and Luo, 2019). Recently, several BERT-based methods for text classifications were suggested, including BertGCN that combines the advantages of a GCN model with the benefits of a pre-trained language model (Lin *et al.*, 2021). Nevertheless, their graph architecture is based on connecting documents to words that occur in them, which may be too noisy for long texts and too sparse for short texts, and was so far applied (to our knowledge) to modern English only.

3. Research objectives and specific aims

Our ultimate goal is to build a classification framework for Qumran texts that integrates, for the first time, three different methods of clustering: linguistic-stylistic, semantic and Ngrams, and uses state-of-the-art deep learning methods. Moreover, we suggest complementing this set of features with a novel layer of construct phrase based features that are central to the language of the DSS. This framework will generate more informed answers to traditional research questions about this corpus. Our specific aims are as follows:

1. Construct a map of construct phrases (CPs) and identify its main hubs and clusters to derive CP features.
2. Apply topic modeling methods to construct a graph of documents and their respective topics
3. Develop a graph convolutional network algorithm using the graph of Aim 2 that propagates linguistic-stylistic, CP and semantic features for clustering and classification of the DSS.

4. Detailed description of the proposed research

For the purpose of this project we will use a tagged corpus of the Qumran scrolls, provided courtesy of Prof. Martin Abegg from Canada. This is an updated version of the corpus represented in the Accordance program. For biblical Hebrew we will use the tagging in https://hb.openscriptures.org/.

Our proposed methodology integrates van Hecke's trigrams with AlephBERT derived semantic features and a linguistic profiling based on a list of indicative linguistic-stylistic features. The linguistic features will be chosen so as to provide effective clustering between various textual units that are apriori quite similar, such as different units of Serekh Hayahad (1QS). We have already experimented with a preliminary set of features that yielded good results, as we detail in the Preliminary Results. The features were selected from two sources. An initial list was compiled from Starr (2016: appendix) which is in itself a compilation of earlier Stylometric studies on Biblical Hebrew since 1980; we deducted from Starr's list those features that seem to us irrelevant, or those that may cause ambiguities when applied to

Qumran Hebrew. In addition we defined several features based on the traits of Late biblical Hebrew in Fassberg (2019:5). Some of these features are hard to extract from the data in their present form but in the course of the project we will perform advanced pre-processing that will allow more precise tagging.

We selected the following linguistic-stylistic features from Starr (2016): (1) Number of different lemmas per $\log_e$ word count, indicating vocabulary richness; (2) Ratio of construct to absolute state nouns (identical to our count of CPs); (3) Noun to verb ratio; (4) Unbound to bound pronoun ratio; (5) Independent pronoun percentage; (6) Definite article percentage; (7) Direct object marker percentage; (8) Pronouns bound to nouns or verbs percentage; (9) Percentage of modal forms (imperative, jussive, cohortative); (10) Preterite percentage; (11) Percentage of the particle (אשר; 12) Percentage of the particle *ʿm* ; (13) All conjunctions percentage; (14) Passive verb forms percentage; and (15) Non-finite to finite verb ratio. Additional features we propose to use are: (1) Percentage of periphrastic tenses (to be + participle); (2) Percentage of infinitive construct with *lamed*; (3) Percentage of negation word (לא) + infinitive construct with *lamed*; and (4) Percentage of elongated *yqtl* forms.

Aim 1, modeling construct phrases (Ben-Dov & Sharan)

We will add another linguistic feature layer of Qumran Hebrew to the list of aforementioned linguistic features: the intensive use of construct phrases (CPs). Many of the sectarian stock phrases are coined as CP. The sectarian authors practice much flexibility with such phrases, for example embedding them in chains of 2, 3 or more CPs or linking two or more nouns on either side of the chain, and using plural participles such as מזוקקי שבעתים. The scrolls thus attest to a robust new morphological-syntactic means of expression, which forms part of their unusual language and concomitantly also of their group identity. A PhD student of PI1 (Ben-Dov, in joint guidance with the expert linguist Prof. Tamar Sovran) is now writing her dissertation on construct phrases in Qumran Hebrew, and its insights will be employed in the project.

We will aim to create a directed network of nouns and their CP pairings. Visualizing and studying this network, e.g., by means of clustering and module detection (Sharan *et al.* 2005), will allow us to derive insights about network hubs and their typical positions in the CP phrase, as well as associate style and meaning with the computed clusters. Alongside the corpus of the DSS we will apply this analysis pipeline also to the Biblical text and perform a comparative analysis of the CP usage patterns derived from the two corpora. We expect this analysis to reveal novel uses of hubs and phrases in the DSS corpus and its different sub-classes as compared to the Bible. In addition, the analysis will yield notable CP clusters that are correlated with different DSS documents and could serve as features for the analysis in the subsequent aims.

Aim 2, topic modeling (Sharan)

An important component of our suggested classification scheme is a skeleton graph which is used to propagate information between similar DSS documents through a graph convolutional network. To this end, we propose to construct a graph of DSS documents and their topics. For this purpose we will manually define a list of "*documents"*, i.e. shorter textual sections within the large literary compositions; e.g. the covenant ceremony and the Treatise of the two spirits in Serekh Hayahad (Hempel, 2020), column 1 in the War Scroll 1QM, or distinct Hodayot as conventionally defined . Following van Hecke, we will represent each document using a vector of trigram counts, and apply topic modeling (Blei, 2012) to identify the main topics, which are distributions over trigrams, and their frequency in each of the documents. Specifically, we will use a multinomial mixture model (aka probabilistic latent semantic indexing) for the data and infer topics via a maximum likelihood approach using the Expectation-Maximization algorithm. We will also experiment with the biterm topic model for shorter texts (Yan *et al.*, 2013), inferred using a maximum likelihood approach.

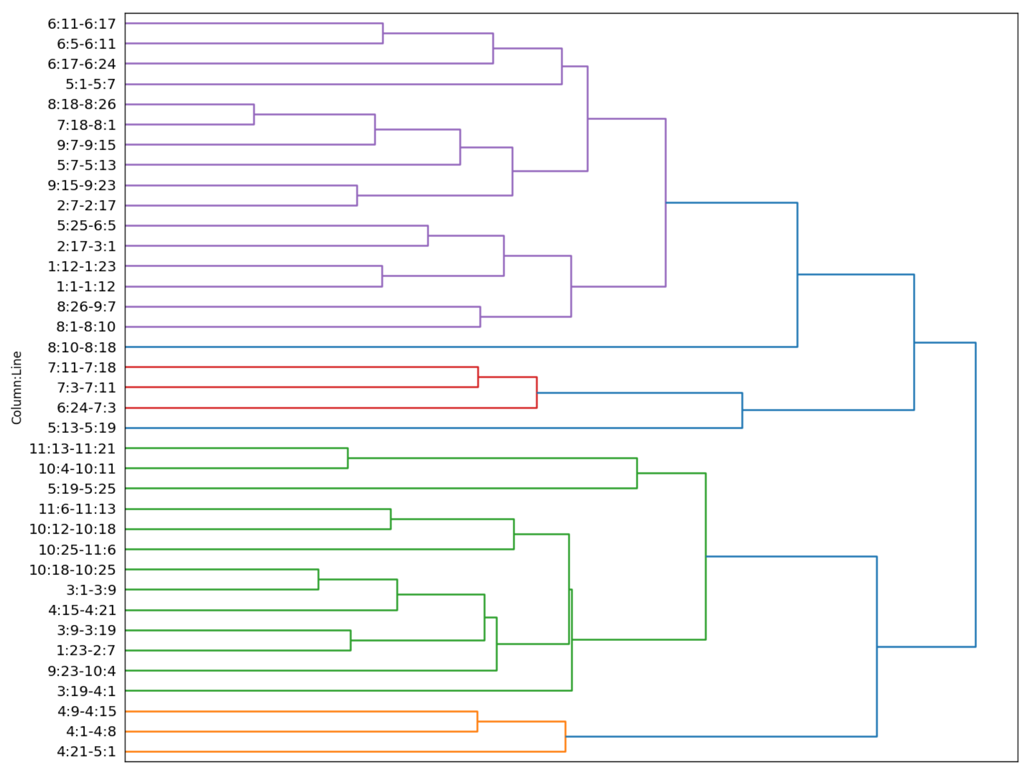
We will connect in the graph every document to the topics whose predicted frequency in that document is above some threshold, weighted by that frequency; we will also connect pairs of topics whose cosine similarity is above some threshold, weighted by that similarity. As an alternative skeleton we will also consider the document-word (occurrence) and word-word (co-occurrence) connections of (Yao, Mao and Luo, 2019) and compare between the two choices when evaluating model performance as detailed in Aim 3. We will also compare our full pipeline to a layman approach that clusters documents based on their topic distributions as inferred above.

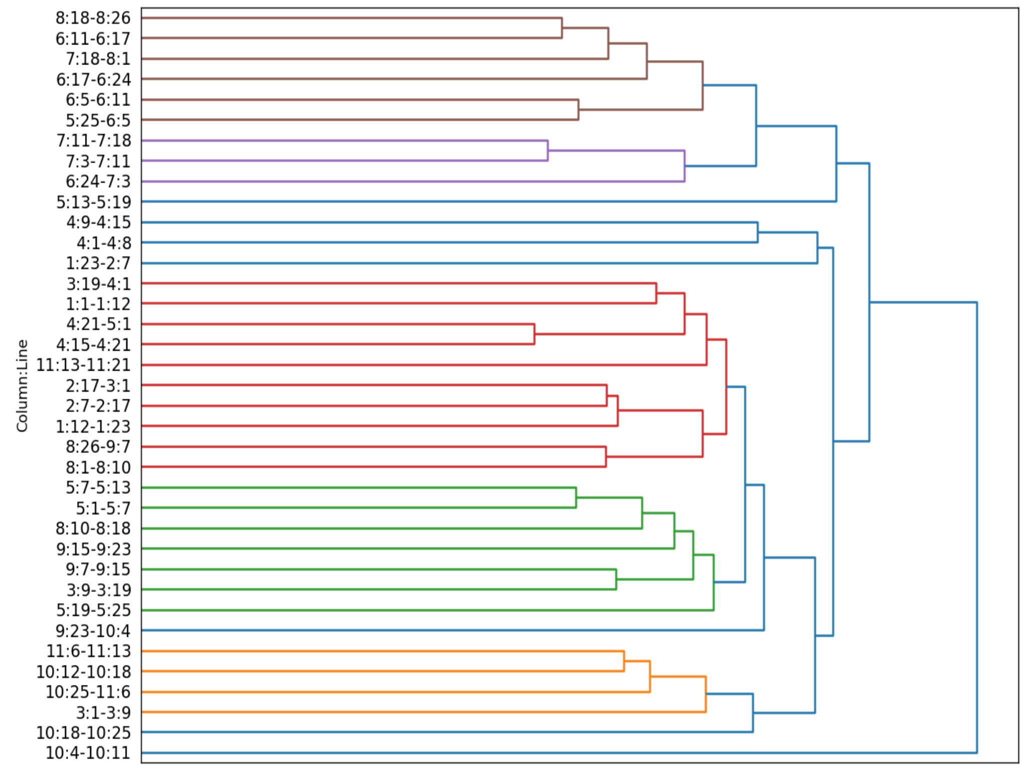
Aim 3, integration and classification (Ben-Dov & Sharan)

In the final stage of the work, we will integrate the CP-based features with Starr's linguistic features and AlephBERT's semantic features and use them for a detailed unsupervised and semi-supervised analysis of the DSS. To this end, we will use the graph constructed in the previous aim as a skeleton, where document features are complemented by topic features that are derived by averaging over the documents connected to it, weighted by the corresponding frequency. We will then apply graph convolutional network analysis which will propagate the features of every document to its neighbors and so on, thus producing a latent representation of the documents that can be used for multi-label classification. To guide the model architecture choice and hyperparameter tuning we will aim at first to cluster the Bible chapters into books or other delineated literary units. To assess the power of our integrated framework, we will compare its performance to direct clustering of each set of features (linguistic, trigram and semantic) using standard quality measures such as Dasgupta's objective for hierarchical clustering (Dasgupta, 2016) and the Jaccard measure for hard clustering. The comparison between the different feature sets is especially important with respect to the semantic features: while recent NLP work focused on semantics and the richness of language in order to absolve the need for manually curating linguistic features, its benefit in fields such as Qumran Hebrew, where the amount of data for pre-training BERT models is insufficient, is of question.

After establishing our framework, we will use core sectarian and non-sectarian scrolls as training data to assign labels to documents that are in dispute, or investigate other authorship debates in the DSS, for example the authorship of groups of Hodayot, columns 19-20 in CD, literary sections of S, or redactional layers in all of the above compositions.

Preliminary results





We made a preliminary exploration of a subset of the linguistic features from Starr described above to assess their predictive power with respect to Qumran hebrew. We focused on 1QS, divided it arbitrarily to blocks of 100 words and annotated each block with the 15 features selected from Starr. We then randomly permuted the blocks and applied hierarchical clustering based on their features using the average linkage algorithm. The results are summarized in the above figure and demonstrate the utility of the features in clustering the blocks, in particular placing to a large extent consecutive blocks together. Indeed when computing a normalized clustering score based on Dasgupta's cost function, where blocks are deemed similar if they are consecutive, the above clustering received a score of 0.55 while a random clustering gets an average score of 0.32 (with std 0.05).

As another proof-of-concept result, we clustered the 1QS blocks based on their AlephBERT features, using the embedding of the CLS token to represent each block. This clustering received a similar score as the feature-based one

(0.51), demonstrating the potential applicability of these features to ancient Hebrew and perhaps suggesting that one way to further improve the clustering is by integrating the different feature sets.

In order to assess the preliminary results we compare them with a large-scale thematic division of 1QS as follows: (a) 1QS 1:1 - 3:12 Introduction and covenant ceremony; (b) 3:13 - 4:26 treatise of two spirits; (c) 5:1 - 7:25 various rules and penal code; (d) 8:1 - 9:11 community charters; (e) 9:12 - 11:22 the maskil. Each of these sections can be divided further into smaller units, which are more relevant for assessing the proposed clustering. The division by Hultgren (2019) is as follows: 1:1-15 introduction; 1:16-3:12 covenant; 3:13-4:26 two spirits; 5:1-13 rules for joining the community; 5:14-6:8 rules for life in the community; 6:8-23 rules for the Many; 6:24-7:27 community discipline (penal code); 8:1-16 charter for the community; 8:16-9:2 discipline; 9:3-11 charter for the community; 9:12-25 regulations for the instructor (Maskil); 9:26-11:22 hymn of the Maskil.

Several valid clusters emerge. For example, in the BERT-based clustering, the first 10 entries all contain legal materials from columns 5-8; the next 4 entries cluster together sections from the Two Spirits Treatise with a similar section from the covenant ceremony; sections from the hymn in columns 10-11 are clustered together at the bottom of the graph together with a poetic section from the top of column 3.

The first clustering (based on Starr’s features) similarly groups legal sections as the first 9 entries, although not the same ones as in the BERT graph. It would be interesting to examine the reasons for this different classification, which could be achieved by isolating the most effective features in each method. Also similarly, the poetic sections from columns 10-11 group together on the graph, and so do the three sections of column 4 at the bottom of the graph.

In summary, both BERT and the stylometric features have successfully distinguished legal from poetic material. This is a promising start, achieved without fine-tuning any of them for the unique characteristics of Qumran Hebrew. The preliminary results thus offer much yet leave much to be explored for achieving finer results by integrating the different feature sets.

5. Significance, innovation and potential benefits of the proposed research

Our main innovation is in three aspects: (i) the introduction of a novel layer of construct phrase clusters for improved classification of the scrolls; (ii) the integration of stylistic-linguistic, statistical and semantic features; and (iii) the development of a deep learning framework for classification. The availability of a cutting-edge classification mechanism will provide scholars with an unprecedented research tool, advancing the study of DSS Hebrew to state of the art machine learning and ultimately providing a more informed approach to the linguistic corpus. Importantly, we will make our model and code publicly available to facilitate future work in this domain.

6. Applicability

We will apply the machine learning framework suggested here to study the DSS and the Bible to yield improved classifications of these texts. The general design of our framework allows its application for researching various other corpora, both in Hebrew and in other languages.

7. Work plan and Gantt

In the following we provide the work plan for the two groups (Ben-Dov: a, Sharan: b), numbered by months according to the three research aims.

#	Task	6	12	18	24	30	36
1	1a	x	x				
2	1b		x	x			
3	2b	x	x				
4	3a			x	x	x	x
5	3b			x	x	x	x

8. Bibliography

Aitchison, J. and Jackendoff, R. (1998) ‘The Architecture of the Language Faculty’, *Language*, p. 850. doi:10.2307/417010.

Bamman, D. (2017). “Natural Language Processing for the Long Tail”, *ADHO 2017.* https://dh2017.adho.org/abstracts/408/408.pdf

Ben-Dov, J. (2009). “Hebrew and Aramaic Writing in the Pseudepigrapha and the Qumran Scrolls: The Ancient Near Eastern Background and the Quest for a Written Authority”, *Tarbiz* 78: 27-60 (Hebrew).

Berman, J. (2021) ‘Measuring Style in Isaiah: Isaiah 34–35 and the Tiberias Stylistic Classifier for the Hebrew Bible’, *Vetus Testamentum*, pp. 303–316. doi:10.1163/15685330-12341070.

Berman, R. (2009). “Acquisition of Compound Construction,” in R. Lieber & P. Stekauer (eds.) *Handbook of Compounding*, Oxford, pp. 298-322.

Blei, D.M. (2012). “Probabilistic topic models”, *Communications of the ACM*, 55:77-84.

Bleiboim, R. and Shatil, N. (2014). “Between Juxtaposition and Construct State,” *Leshonenu* 76: 345-370 (Hebrew).

Dasgupta, S. (2016) ‘A cost function for similarity-based hierarchical clustering’, *Proceedings of the forty-eighth annual ACM symposium on Theory of Computing* [Preprint]. doi:10.1145/2897518.2897527.

Dershowitz, I., A. Navot, M. Koppel and N. Dershowitz (2015). “Computerized Source Criticism of Biblical Texts”. *JBL* 134, pp. 253-271.

Devlin, J., Chang, M.-W., Lee, K. and Toutanova, K. (2018). “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding”, arXiv:1810.04805.

Fassberg, S. (2021) ‘Hebrew Texts and Language of the Second Temple Period’. doi:10.1163/9789004447981.

Fassberg, S.E. (2019) *Mavo le-taḥbir leshon ha-Miḳra: An introduction to the syntax of biblical Hebrew.*

Harkins, A.K. (2018) ‘Another Look at the Cave 1 Hodayot: Was CH I Materially Part of the Scroll 1QHodayota?’, *Dead Sea Discoveries*, pp. 185–216. doi:10.1163/15685179-12341480.

Hempel, C. (2020) ‘The Community Rules from Qumran’. doi:10.1628/978-3-16-157027-8.

Hendel, R. and Joosten, J. (2018) *How Old Is the Hebrew Bible?*. New Haven: Yale UP. doi:10.2307/j.ctv7cjvjc.

Hultgren S. (2019). “Serekh Hayahd (S), in *T&T Clark Companion to the Dead Sea Scrolls*. Ed. C. Hempel and G. Brooke. London, 344-346.

Jackendoff, R. (1997). *The Architecture of the Language Faculty***,** Cambridge, Mass.

Jackendoff, R. and Audring, J. (2020) ‘Relational Morphology: A Cousin of Construction Grammar’, *Frontiers in psychology*, 11, p. 2241.

Keren, O., Avinari, T., Tsarfaty, R. and Levy, O. (2022). “Breaking Character: Are Subwords Good Enough for MRLs After All?”, arXiv:2204.04748.

Lin, Y. *et al.* (2021) ‘BertGCN: Transductive Text Classification by Combining GNN and BERT’, *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021* [Preprint]. doi:10.18653/v1/2021.findings-acl.126.

Muraoka, T. (2020) ‘A Syntax of Qumran Hebrew’. doi:10.2307/j.ctv1q26jm3.

Newsom, C.A. (2021) ‘A Farewell to the Hodayot of the Community’, *Dead Sea Discoveries*, pp. 1–19. doi:10.1163/15685179-bja10002.

Qimron, E. (2018). *A Grammar of the Hebrew of the Dead Sea Scrolls*, Jerusalem: Yad Ben Zvi.

Schniedewind, W. (2013). *A Social History of Hebrew*, New Haven: Yale UP. doi:10.12987/yale/9780300176681.001.0001.

(*) Seker, A. *et al.* (2022) ‘AlephBERT: Language Model Pre-training and Evaluation from Sub-Word to Sentence Level’, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)* [Preprint]. doi:10.18653/v1/2022.acl-long.4.

Sharan, R. et al. (2005). *"Conserved patterns of protein interaction in multiple species". PNAS* 102, 1974-79.

Starr, J. (2016) *Classifying the Aramaic Texts from Qumran: A Statistical Analysis of Linguistic Features*. Bloomsbury Publishing.

Talshir, D. (2013). “Syndetic Binomials in Second Temple Period Hebrew”, in *The Hebrew in the Second Temple Period: The Hebrew of the Dead Sea Scrolls & of Other Contemporary Sources,* eds. S. E. Fassberg, M. Bar-Asher & R. Clements, Leiden: Brill, 225- 239.

(*) Van Hecke, P. (2018) ‘Computational Stylometric Approach to the Dead Sea Scrolls’, *Dead Sea Discoveries* 25, 57-82. doi:10.1163/15685179-12341464.

Van Hecke, P. and J. de Joode (2021). “Promises and Challenges in Designing Stylometric Analyses for Classical Hebrew”, in *Hebrew Texts and Language of the Second Temple Period. Proceedings of an Eighth Symposium on the Hebrew of the DSS and Related Literature*. Ed. S. Fassberg. Leiden: Brill, 349-374.

Yan, X. *et al.* (2013) ‘A biterm topic model for short texts’, *Proceedings of the 22nd international conference on World Wide Web - WWW ’13* [Preprint]. doi:10.1145/2488388.2488514.

Yao, L., Mao, C. and Luo, Y. (2019) ‘Graph Convolutional Networks for Text Classification’, *Proceedings of the AAAI Conference on Artificial Intelligence*, pp. 7370–7377. doi:10.1609/aaai.v33i01.33017370.