A Dataset for Metaphor Detection in Early Medieval Hebrew Poetry

Anonymous ACL submission

Abstract

The corpus of late antique and medieval Hebrew texts is vast and represents a crucial linguistic and cultural bridge between Biblical and modern Hebrew. Poetry is prominent in this corpus and one of its main characteristics is the frequent use of metaphors. Distinguishing figurative and literal language use is a major task for scholars of the Humanities, especially in the fields of literature, linguistics and hermeneutics. This paper presents a new, challenging dataset of late antique and medieval Hebrew poetry with expert annotations of metaphors, as well as some baseline results, which we hope will facilitate further research in this area.[1]

1 Introduction

The Hebrew language has a long and rich history, from Biblical Hebrew, through Rabbinic-Medieval Hebrew, to modern Hebrew. Since poetry was a prominent genre in late antique and medieval Hebrew literature, the corpus is rich in figures of speech like similes and metaphors. Active research in this area is conducted by scholars in the Humanities, especially Digital Humanities, who wish to understand not only the literal meaning of the text but also its figurative meaning (Münz-Manor, 2011). The current practice involves manual annotation of the texts, a process that is both time-consuming and requires significant human resources.

Scholars of Hebrew literature and Hebrew linguists would thus benefit greatly from a tool that automatically detects figurative language in these texts. Furthermore, these tools could be used by non-specialists who want to better understand these texts by highlighting figurative language. Since the literary and linguistic tradition of Piyyut runs throughout the Middle Ages, working on the early strata of this tradition would enable us to extend the impact of metaphor detection also to later periods and other genres. Yet, to the best of our knowledge, there are no previous studies that deal with this task, in either modern or pre-modern Hebrew.

To fill this gap, the main contribution of this work is a medieval Hebrew dataset with metaphor annotations of Hebrew liturgical poetry from the fifth to eighth centuries CE, also known as Piyyut.[2] The dataset consists of two corpora of ancient Piyyut, with 309 poems and 73,179 words, with expert annotations for metaphors at the word level.

We develop and evaluate several transformer-based models for detecting metaphors in the dataset, based on two pre-trained Hebrew language models: AlephBERT, which was pre-trained on modern Hebrew (Seker et al., 2021), and BEREL, pre-trained on ancient Jewish texts that are closer in style to the Piyyut texts (Shmidman et al., 2022). We substantially improve naïve baselines, with our best model achieving F1 scores of 48.7 and 49.4 on the two corpora. Considering the difficulty of the task, attested through an inter-annotator agreement study we conducted, we find the results encouraging while leaving ample room for improvements.

2 Background

2.1 Literary and Linguistic Background

Jewish liturgy took shape in the Near East in the first centuries of the Common Era and by the end of the 3rd century began to take on fixed forms. In the late 4th century, poets began to embellish liturgical prose, infusing religious meaning with poetic beauty. By the 7th century, Piyyut (Jewish liturgical poetry) became an integral medium of religious discourse and Payytanim (liturgical poets) evolved into prominent cultural figures (Lieber, 2010).

The study of Piyyut is relatively young and rather small in scale, since most of the Payytanic texts from this period were discovered towards the end of the 19th century in the Cairo Genizah. Throughout

[1] Code and data will be publicly released upon publication under the CC-BY Creative Commons license.

[2] From Greek *poietes*, to create, versify. Plural, Piyyutim.

most of the twentieth-century scholars of Piyyut focused on literary and linguistic investigations of the texts (Van Bekkum, 2008). In essence, the Payytanic language constitutes a separate stratum in the history of the Hebrew language although it is much closer to biblical Hebrew than to contemporaneous Rabbinic Hebrew. Importantly, there are significant differences between Piyyut and modern Hebrew, at syntactic and lexical levels.

In summary, metaphors play an important role in the literary fabric of Piyyut and at later stages, most notably in the Islamic East, they become increasingly central and innovative. The study of figurative language in Piyyut and more broadly in medieval Hebrew literature remains a major task and computational tools would greatly help advancing this area (Münz-Manor, 2011).

2.2 Hebrew NLP

Hebrew is a low-resourced morphologically-rich language with few labeled datasets, which are typically in modern Hebrew (Keren and Levy, 2021; Litvak et al., 2022). Notable unlabeled Hebrew corpora are the Ben-Yehuda project, a heterogeneous collection of medieval and mostly modern Hebrew literature;[3] and the Sefaria collection[4] and the Dicta Library,[5] which are composed of ancient Jewish texts.

Several Hebrew language models have been released, most of them trained on limited data compared to English language models (e.g., HeBERT; Chriqui and Yahav, 2021). A prominent model is AlephBERT (Seker et al., 2021), which was trained on a 1.9 billion words corpus of modern Hebrew. Fine-tuning it led to high performance on multiple sequence labeling tasks. A more recent model is BEREL (Shmidman et al., 2022). It was pretrained on Rabbinic Hebrew texts from Sefaria and the Dicta Library, which are more similar to Piyyut than modern Hebrew. BEREL's training set is an order of magnitude smaller than AlephBERT's (220 million compared to 1.9 billion words).

2.3 Metaphor Detection

Early work on metaphor detection has been based on feature engineering approaches, using for example frequency and co-occurrence of words with metaphorical words (Sardinha, 2006), or abstractness, supersenses, and unsupervised vector-space word representations (Peng et al., 2021).

Several attempts have been made to detect metaphors with pre-trained transformers (Vaswani et al., 2017). Su et al. (2020) augmented BERT (Devlin et al., 2018) with local representations of candidate words and linguistic features such as part of speech. Choi et al. (2021) used the gap between a word's representation in context and that without context, as well as the gap between the metaphor word and its neighboring words.

3 The Dataset

3.1 Construction and Annotation

The dataset consists of two separate corpora of Piyyut: (1) 172 poems by various poets (all anonymous except for one, Yosei ben Yosei) that were composed during the 5th century CE in the Galilee. This is the earliest corpus of Piyyut and it represents the formative phase of this poetic tradition, and referred to here as Pre-Classical Piyyut. (2) 137 poems by Pinchas Ha-Cohen (the Priest), who lived in the first half of the 8th century CE in Tiberias, and is regarded as the last major poet of the classic payytanic tradition (Elizur, 2004). Both corpora were recovered from medieval manuscripts that were unearthed towards the end of the 19th century in a medieval synagogue in Cairo.

The entire corpus was manually analyzed and annotated by an expert, who studied the literary aspects of the corpus with a special emphasis on figurative language and metaphors in particular. It was digitized using the CATMA annotation tool (Meister et al., 2017). The annotations appear at the word level, with each word tagged as metaphoric or literal. For metaphors that span several words, each word gets its own separate label. This allows for easier evaluation of the model in cases where only part of the words were detected as metaphors. Table 1 contains examples of texts and metaphor annotations from the dataset.

Since the identification of metaphors is to some extent interpretative, we asked another literary expert to annotate part of the corpora so we can calculate inter-annotator agreement and have a benchmark to evaluate the results of the models. The second expert annotated 27.7% of the first corpus and 18.5% of the second. The calculated Cohen's kappa score is 0.618 for the Pre-Classical Piyyut corpus and 0.628 for the Pinchas corpus. Although considered as a "substantial" agreement, the score reflects non-negligible variations between the two

[3] https://benyehuda.org

[4] https://www.sefaria.org.il/texts

[5] https://library.dicta.org.il

Examples (Hebrew)	Literal Translation	Meaning
טבענו בגזרות	We **drowned** in decrees	There are too many decrees
עצבון במשלח ידנו	Irritation is **in our hands**	We are sad at work
אחפס קרביים כליות אחקור	I'll explore **kidney guts**	Investigate the true intentions
הצית נשמה ויבער נר	**Ignite** a soul, **file a candle**	Activate a soul
לא עשו פרי	Did not **bear fruit**	Did no good deeds

Table 1: Examples from our dataset. Bold words in the middle column refer to annotated metaphors.

annotators. These variations should be further investigated, and it should be noted that while in some cases they are due to human errors, in more complex setups, variations are plausible and abundant and may lead to better models (Plank, 2022).

We fixed each label after receiving the data to begin at the first character of the word and end at the last character. We deleted some duplicated texts with minor changes and kept only one copy.

3.2 Statistics and Standard Splits

Our corpora size and metaphor percentage are summarized in Table 2. We note that 16.3% and 21.3% of the words are annotated as metaphors in the Pre-Classical Piyyut and Pinchas corpora, respectively. A few texts have an unusual high percentage of metaphors (Appendix A).

	Pre-Classical	Pinchas
# texts	172	137
# sentences	$6,836$	$6,881$
# words	$43,697$	$29,482$
% metaphor	16.3	21.3

Table 2: Overall statistics of the two corpora.

To facilitate reproducible research with the corpus, we define standard splits into training, validation, and test sets. The dataset is first split 80/20 into training and test, and then training is split again 80/20 to the final training and validation sets. (Table 4 in Appendix A.1 provides exact sizes.) We randomly split by text, so each text is only found in one split. We stratify by text length and metaphor ratio, to ensure that every split consists of texts with similar distributions. Of the words annotated as metaphors in the test sets of Pre-Classical Piyyut and Pichas, respectively, 55% and 52% do not appear as metaphors in the corresponding training sets.

3.3 Limitations

As aforementioned, metaphor detection involves human interpretation, making ambiguity common in both human and automatic metaphor detection.

The Pre-Classical Piyyut corpus was reconstructed from a somewhat arbitrary collection. The poems we have are the only ones that survived from the 5th century and in most cases we cannot identify the poets. Therefore, the corpus is not homogeneous and its literary and linguistic aspects can differ considerably and complicate the process of metaphor detection, whether manual or automatic. The Pinchas corpus, on the other hand, even if not complete because some poems may have been lost throughout the ages, represents the poetic works of one poet, hence it is much more homogeneous.

4 Experimental Evaluation

4.1 Problem Formulation and Metrics

We treat metaphor detection as a sequence labeling task, with each word labeled as metaphor or non-metaphor. When using models that split words into sub-word units, we assign all sub-words the word's label. At test time, we tag a word as metaphor if at least one sub-word was tagged as metaphor.

Due to the imbalanced nature of the dataset (Section 3.2), we primarily focus on the F1 score, but report also precision, recall, and accuracy.

4.2 Naive Baselines

Due to the novelty of this task, we report two naïve baselines. The majority baseline always assigns non-metaphor, obtaining around 80% accuracy, but its F1 score is zero. Another baseline is assigning the most frequent tag of the word in the training set for seen words, and a non-metaphor tag for unseen words. This baseline achieves a 24 F1 score. See Table 3 for F1 scores and other metrics in Appendix A.5. In general, the trends on both of our corpora are similar.

4.3 Transformer-based models

We experiment with two pre-trained Hebrew language models—AlephBERT and BEREL— which we fine-tune on the metaphor detection task. Both models are encoder-only with 12 layers. As explained in Section 2.2, the two models differ in the pre-training data, as well as their tokenizers and vocabularies (50K items in AlephBERT, 128K items in BEREL). The results in this section are the average of five runs with different seeds. To examine the effect of the latter, we first trained randomly-initialized versions of the two models on metaphor detection, obtaining poor F1 results of about 30–34. Details about the training and hyperparameters can in found in A.4

Next, we fine-tuned the pre-trained models, yielding substantial improvements: 40.8/42.2 F1 with AlephBERT on the two corpora, 43.7/46.5 with BEREL. We attribute the superior performance of BEREL both to its pre-training data being closer to the Piyyut language compared to AlephBERT's modern Hebrew pre-training data, and to its vocabulary size. This is especially noteworthy given that BEREL was pre-trained on 10x less data.

The fact that BEREL outperforms AlephBERT despite being pre-trained on less data suggests that adaptation to the target genre is crucial. Following Gururangan et al. (2020), we adapted AlephBERT to Piyyut by training it with the masked language modeling task on texts more similar to Piyyut. We first trained it on texts from Project Ben-Yehuda, a collection of modern and medieval Hebrew literature. We then continued training it on our Piyyut corpus (without metaphor labels). Finally, we fine-tuned the adapted model on metaphor detection. This process improved results by 1–2% ("adapted" rows, Table 3).

In view of the highly unbalanced data (metaphors are only 16% in Pre-Classical Piyyut and 21% in Pinchas), we used a weighted cross-entropy (WCE) loss. By increasing the loss of the wrong prediction of the less frequent class (metaphors), we encourage the model to identify more metaphors. This modification hurts precision and increases recall, resulting in an increase in F1 scores of 3–4 points (WCE rows in Table 3; Tables 6 and 7 in Appendix A.5). Fine-tuning BEREL with WCE provided the best results in our experiments in terms of F1.

Model	Pre-Classical	Pinchas
Global majority	0.0	0.0
Most frequent tag	24.2	24.7
BEREL rand	30.7 ± 2.1	34.4 ± 2.3
AlephBERT rand	31.6 ± 2.2	31.3 ± 3.4
BEREL	43.7 ± 0.6	46.5 ± 2.0
+ WCE	$\mathbf{48.7 \pm 1.4}$	$\mathbf{49.4 \pm 0.8}$
AlephBERT	40.8 ± 2.0	42.2 ± 1.2
+ WCE	45.9 ± 0.7	45.5 ± 2.0
+ adapted	42.8 ± 1.3	44.8 ± 0.7
+ adapted+WCE	47.2 ± 0.9	47.3 ± 1.0

Table 3: Metaphor detection average F1 scores. Each experiment was repeated five times with different seeds.

4.4 Performance Analysis

We examined how the best model (BEREL, trained with WCE) performs on words in the test set (of the Pre-Classical corpus) that do not exist in the training set ("unseen" words). Similarly, we examined its performance on "seen" words (i.e., words in the test set that exist in the training set). While the F1 score for seen words (54.6) is greater than unseen words (44.3), the latter score is still substantial, indicating that the model has learned to generalize to new words and metaphors.

Next, we examined the model's common mistakes. For example, the word מים (water) is the most common false positive. In the test set, it is correctly predicted as not being a metaphor 29 times (true negative), with only two false negative predictions. However, with regard to positive predictions, the model correctly detected the word as a metaphor 8 times, with 12 false positive predictions. This may be explained by WCE, which encourages metaphor detection.

5 Conclusion

We presented a corpus of medieval Hebrew poetry with metaphor annotations. The corpus can serve literary scholars who wish to study figurative language use in this genre. We also evaluated basic approaches for automatic metaphor detection based on this corpus, emphasizing the importance of adapting models to this particular genre. We hope to facilitate further research in this area, both in designing more sophisticated methods for metaphor detection in a challenging corpus and in improving the workflow of literary scholars interested in this body of texts.

References

Minjin Choi, Sunkyung Lee, Eunseong Choi, Heesoo Park, Junhyuk Lee, Dongwon Lee, and Jongwuk Lee. 2021. Melbert: metaphor detection via contextualized late interaction using metaphorical identification theories. *arXiv preprint arXiv:2104.13615*.

Avihay Chriqui and Inbal Yahav. 2021. Hebert & hebemo: a hebrew bert model and a tool for polarity analysis and emotion recognition. *arXiv preprint arXiv:2102.01909*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Shulamit Elizur. 2004. The liturgical poems of rabbi pinhas ha-kohen.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Omri Keren and Omer Levy. 2021. Parashoot: A hebrew question answering dataset. *arXiv preprint arXiv:2109.11314*.

Laura S Lieber. 2010. *Yannai on Genesis: An Invitation to Piyyut*, volume 36. ISD LLC.

Marina Litvak, Natalia Vanetik, Chaya Liebeskind, Omar Hmdia, and Rizek Abu Madeghem. 2022. Offensive language detection in hebrew: can other languages help? In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 3715–3723.

Jan Christoph Meister, Evelyn Gius, Jan Horstmann, Janina Jacke, and Marco Petris. 2017. Catma 5.0 tutorial. In *DH*. Alliance of Digital Humanities Organizations (ADHO).

Ophir Münz-Manor. 2011. Figurative language in early piyyut. In *Giving a Diamond*, pages 51–67. Brill.

Xutan Peng, Chenghua Lin, and Mark Stevenson. 2021. Cross-lingual word embedding refinement by l1 norm optimisation.

Barbara Plank. 2022. The 'problem' of human label variation: On ground truth in data, modeling and evaluation.

Tony Berber Sardinha. 2006. Collocation lists as instruments for metaphor detection in corpora. *DELTA: Documentação de Estudos em Lingüística Teórica e Aplicada*, 22:249–274.

Amit Seker, Elron Bandel, Dan Bareket, Idan Brusilovsky, Refael Shaked Greenfeld, and Reut Tsarfaty. 2021. Alephbert: A hebrew large pre-trained language model to start-off your hebrew nlp application with. *arXiv preprint arXiv:2104.04052*.

Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Eli Handel, and Moshe Koppel. 2022. Introducing berel: Bert embeddings for rabbinic-encoded language. *arXiv preprint arXiv:2208.01875*.

Chuandong Su, Fumiyo Fukumoto, Xiaoxi Huang, Jiyi Li, Rongbo Wang, and Zhiqun Chen. 2020. Deepmet: A reading comprehension paradigm for token-level metaphor detection. In *Proceedings of the second workshop on figurative language processing*, pages 30–39.

Wout Jac Van Bekkum. 2008. The hebrew liturgical poetry of byzantine palestine: Recent research and new perspectives. *Prooftexts: A Journal of Jewish Literary History*, 28(2):232–246.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

A Appendix

A.1 Additional Statistics

Figures 1 and 2 show histograms of texts in the two corpora, binned by the ratio of metaphors they contain. While a few texts contain a very high ratio of metaphors, most texts have a small such ratio. Table 4 presents the division of the dataset into training, validation, and test splits.

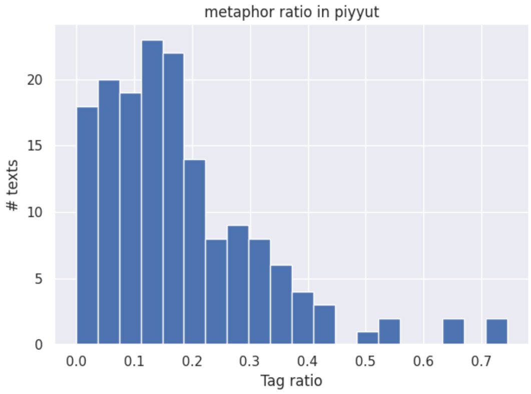

Figure 1: Distribution of the metaphor ratio in the Pre-Classical Piyyut corpus.

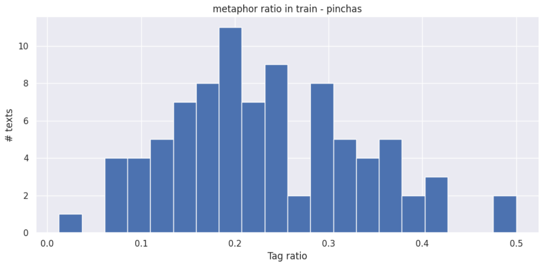

Figure 2: Distribution of the metaphor ratio in the Pinchas corpus.

	Training	Validation	Test
		Pre-Classical	
Metaphor	4707	1070	1070
Non-Metaphor	26485	26485	5103
Total	31192	6322	6183
		Pinchas	
Metaphor	4105	867	1225
Non-Metaphor	15552	2932	4801
Total	19657	3799	6026

Table 4: Number of tokens in each split for each corpus.

A.2 Intended Use

The work utilizes open-source models and resources that are in the public domain. We will release the dataset and associated models under the CC-BY Creative Commons license, in a GitHub repository that will include usage guidelines.

A.3 Potential Risks

We release a dataset from the 7th century. Many of the texts from that time period are biased, and some may find them offensive. The use of this dataset for metaphor detection does not appear to pose risks; however, it may result in biased or offensive models when it is used for other purposes.

A.4 Training Details

In this study, there were two kinds of training: fine-tuning and model adaptation. Using transformers hyperparameter search, we found the best hyperparameters for fine-tuning. Refer to Table 5 for the complete list of hyperparameters. We completed the hyperparameter search for each model and dataset pair. Since the hyperparameters were similar across experiments, we used the same hyperparameter throughout. We repeated the experiments five times with seeds 41-45. The final results can be found in tables 6, 7. The training was composed on Nvidia RTX 2080. A total of 16 experiments were conducted, five times each (different seeds), resulting in 13.5 hours of GPU time.

For model adaptation, we used a learning rate of 1e-4, batch size 128, 3 epochs, and 10000 warmup steps. The training was composed on Nvidia RTX 2080, with 10 hours of GPU time.

	Range	Best
learning rate	$1e-6:1e-3$	$5.4e-4$
epochs	$2:10$	8
batch size	$16, 32, 64, 128$	32
metaphor weight	$1:20$	9

Table 5: Hyperparamets searched (range) and chosen (best) for fine-tuning. The metaphor weight is the weight for weighted cross entropy.

A.5 Detailed Results

Tables 6 and 7 show detailed results on both corpora, including accuracy, precision, and recall, in addition to F1 scores, which were given in the main body.

Model	Accuracy	Precision	Recall	F1
Global majority	82.5	0.0	0.0	0.0
Most frequent tag	71.5	48.5	16.1	24.2
BEREL random init	78.5 ± 1.3	37.1 ± 2.3	26.6 ± 4.0	30.7 ± 2.1
AlephBERT random init	78.7 ± 0.7	37.3 ± 1.4	27.6 ± 3.3	31.6 ± 2.2
BEREL	$\mathbf{82.2 \pm 0.4}$	$\mathbf{51.1 \pm 1.4}$	38.2 ± 1.2	43.7 ± 0.6
BEREL WCE	77.2 ± 3.4	41.7 ± 3.9	$\mathbf{62.5 \pm 5.8}$	$\mathbf{48.7 \pm 1.4}$
AlephBERT	78.5 ± 2.0	48.1 ± 2.1	35.6 ± 3.7	40.8 ± 2.0
AlephBERT WCE	76.2 ± 0.1	38.5 ± 1.4	56.4 ± 2.6	45.9 ± 0.7
AlephBERT adapted	81.8 ± 0.5	49.4 ± 2.0	38.0 ± 2.9	42.8 ± 1.3
AlephBERT adapted WCE	76.2 ± 1.7	40.3 ± 2.6	59.5 ± 4.4	47.2 ± 0.9

Table 6: Results on Pre-Classical Piyyut corpus: Average Accuracy, Recall, Precision, F1, and standard deviations for all described methods. Each experiment was repeated five times with different seeds. WCE refers to weighted cross-entropy loss.

Model	Accuracy	Precision	Recall	F1
Global majority	79.7	0.0	0.0	0.0
Most frequent tag	79.6	49.9	16.4	24.7
BEREL random init	73.2 ± 2.6	36.7 ± 2.9	33.1 ± 6.3	34.4 ± 2.3
AlephBERT random init	74.8 ± 1.1	36.5 ± 1.7	25.8 ± 4.1	31.3 ± 3.4
BEREL	$\mathbf{79.7 \pm 1.1}$	$\mathbf{53.6 \pm 4.1}$	41.6 ± 5.2	46.5 ± 2.0
BEREL WCE	71.2 ± 3.5	40.0 ± 2.9	$\mathbf{65.7 \pm 7.6}$	$\mathbf{49.4 \pm 0.8}$
AlephBERT	79.1 ± 0.8	50.9 ± 2.8	36.1 ± 1.7	42.2 ± 1.2
AlephBERT WCE	75.6 ± 2.5	43.9 ± 3.9	48.7 ± 8.7	45.5 ± 2.0
AlephBERT adapted	79.7 ± 0.9	52.5 ± 2.9	39.3 ± 2.4	44.8 ± 0.7
AlephBERT adapted WCE	75.4 ± 2.5	43.9 ± 3.6	52.4 ± 6.5	47.3 ± 1.0

Table 7: Results on Pinchas corpus: Average Accuracy, Recall, Precision, F1, and standard deviations for all described methods. Each experiment was repeated five times with different seeds. WCE refers to weighted cross-entropy loss.