BEReshiT: an Ancient Hebrew Model based on DictaBERT

Iglika Nikolova-Stoupak[1], **Maxime Amblard**[1], **Frédérique Rey**[2]
[1]LORIA, UMR 7503, Université de Lorraine, CNRS, Inria, 54000 Nancy, France
[2]Zentralinstitut für Katholische Theologie, Humboldt-Universität zu Berlin, 10099 Berlin, Germany
{iglika.nikolova-stoupak, maxime.amblard}@univ-lorraine.fr, frederique.rey@hu-berlin.de

Abstract

This project addresses the general absence of Natural Language Processing (NLP) tools when it comes to historical languages as a subset of low-resource languages that is relevant to an array of academic disciplines from linguistics to textual criticism. In particular, we train an Ancient Hebrew language model, BEReshiT, as well as BEReshiT-morph, a submodel for morphological annotation. BEReshiT is achieved through the fine-tuning of DictaBERT, a state-of-the-art model for Modern Hebrew that has also proved useful in Biblical Hebrew tasks. Layer freezing is applied in order to achieve maximal results while gaining insight about the adaptation process. In the context of an elaborate cloze test, BEReshiT demonstrates increased performance and notions of the Ancient Hebrew language compared to the source model as well as a selection of additional relevant models. The submodel BEReshiT-morph performs highly on tasks of morphological classification, reaching an F1 score of 0.97 for part-of-speech (POS) tagging. We release the main and morphological models as well as the datasets used at training and evaluation.

Keywords: BERT, layer freezing, Ancient Hebrew, cross-variety transfer, model interpretability

1. Introduction and Motivation

It is well-known that the NLP field is dominated by data and tools that pertain to high-resource languages (Joshi et al., 2020). Whilst low-resource languages have been receiving increased attention in recent years, work with them still constitutes a challenge (Hedderich et al., 2021; McGiff and Nikolov, 2025). In particular, historical languages are associated with a limited number of texts, which are moreover typically homogenous in register and complexity. Yet, these languages are of key importance within a number of research disciplines, such as linguistics, archaeology, theology and textual criticism.

We are focusing on Ancient Hebrew, the language of the Hebrew Bible and of additional texts similar in time frame and linguistic characteristics. Ancient Hebrew differs from Modern Hebrew in terms of morphology, syntax and, most notably, vocabulary. For more information about the Hebrew language and its stratification, please refer to: Khan et al. (2013), Hurwitz (2014), Hornkohl (2014) and Rezetko and Young (2014).

We are herein presenting our language model, BEReshiT, along with the submodel BEReshiT-morph, which predicts the morphological characteristics of Ancient Hebrew words in context. BEReshiT is achieved through the fine-tuning of a state-of-the-art BERT-based model specialising in Modern Hebrew, DictaBERT (Shmidman et al., 2023). In order to make optimal use of the source model as well as to gain insight about the relationship between BERT architecture and the Hebrew language (in its different varieties), we trained twelve versions of the model, each corresponding to a different scenario of layer freezing.

The trained models, as well as a representative selection of established Hebrew or multilingual models, were evaluated on specially crafted datasets, which underline the specificities of the Ancient Hebrew language as well as notable differences with its Modern counterpart. These datasets, as well as the strongest BEReshiT model and the BEReshiT-morph submodel, are publicly released in the following repository: https://zenodo.org/records/18750374. Our main contribution consists in the Ancient Hebrew model and associated submodel artefacts rather than in novelty in terms of aspects such as training and evaluation. The models may be used in future work with Ancient Hebrew texts that have been newly discovered or digitalised or that have not yet received high levels of attention (notably, morphological annotation).

2. Background

There are several Hebrew language models to date. In particular, the Dicta team[1] has developed high performing models that focus on different aspects and varieties of the language. DictaLM are recent Hebrew-based large language models (LLMs) (Shmidman et al., 2024b). In this study, we chose to focus on a BERT-based model (Devlin et al., 2019) for several reasons. BERT models systematically capture rich morphological and syntactic information across layers, which is particularly important for a morphologically complex language such as Hebrew. In addition, the architecture has been shown to allow for successful experiments that imply explainability and interlingual comparison. Also, task- and language-specific models such as DictaBERT

[1]https://dicta.org.il/

Proceedings of LT4HALA 2026 @ LREC 2026, pages 72–88.

have achieved strong performance on NLP tasks, often dominating the state of the art. Finally, the relatively small size of BERT models makes them more suitable than other high-performing models, such as LLMs, in view of computational and environmental constraints.

MsBERT is trained for the task of filling in lacunae in Hebrew manuscripts (Shmidman et al., 2024a). In turn, BEREL (Shmidman et al., 2022) specialises in Rabbinic Hebrew, a stratum that has significant links with Ancient Hebrew in terms of linguistics and register. However, we opted for the Modern Hebrew model DictaBERT (Shmidman et al., 2023) as a model to finetune for the following reasons: it is associated with a strong morphological submodel, whose performance has brought improvement even on tasks related to Ancient Hebrew (Nikolova-Stoupak et al., 2025); also, BEREL has already been trained on the Hebrew Bible as well as additional sources that reference it and is therefore not a good candidate for domain adaptation through fine-tuning on available Ancient Hebrew text. To our knowledge, no uniquely Ancient/Biblical Hebrew language model exists to date.

BERT- and RoBERTa-based models have been trained for a large variety of languages, including historical ones such as Latin (Bamman and Burns, 2020) and early modern French (Gabay et al., 2022). HmBERT is a multilingual model (a.k.a. one-model) of historical languages[2] that specialises in Named Entity Recognition (NER) (Schweter et al., 2022). The model improves on single-model benchmarks for the German language whilst falling short for the other languages. Fine-tuning of previously trained models has also been used in adaptation to low-resource historical languages. For instance, Lendvai et al. (2023) fine-tune several multilingual BERT models on historical varieties of Slavic languages for the purpose of identification of manuscripts and related tasks, achieving highest results when using models that already specialise in Cyrillic[3].

Various experiments have been carried out as a quest to gain insight about BERT's structure with regards to language knowledge. Jawahar et al. (2019) conclude that the model captures linguistic information about English in tree-like structures. Working with English, French and Swedish on the task of proficiency level determination, Muñoz Sánchez et al. (2024) explore layer freezing at BERT fine-tuning, discovering that optimal architectures differ for each language.

[2]historical German, English, French, Swedish and Finnish

[3]KoichiYasuoka/bert-base-slavic-cyrillic-upos and anonsubmissionmk/bert-base-macedonian-bulgarian-cased

3. Data

3.1. Model Training

We trained the BEReshiT models on the whole Hebrew[4] text of the Hebrew Bible (373 093 words/28 379 verses) and clean[5] Hebrew verses from the Dead Sea Scrolls (DSS), a series of both Biblical and extra-Biblical manuscripts discovered from 1947 onwards (67 806 words/5955 lines)[6]. The two sets of texts were used as available, respectively, through the Open Scriptures Hebrew Bible (OSHB)[7] and ETCBC's The Dead Sea Scrolls Text-Fabric dataset[8]. In accordance with DictaBERT, we used the established Hebrew character set consisting of consonantal letters only, excluding diacritical marks such as vowel signs (*nikkudot*) and cantillation marks. In the case of the Hebrew Bible, we duplicated one quarter of the data and included it a second time along with *nikkudot* in order to ensure that the model has knowledge of this later convention whilst not overburdening it with what it would treat as new vocabulary.

For the purpose of training the morphological submodel, we created a token-level dataset based on the same data along with morphological information as present in the two data sources and mapped in cases of difference in notation. To limit the number of morphological features per token, prefixes were separated from the word they modify.

3.2. Model Evaluation

We composed a 200-verse dataset that seeks to capture the general qualities of the Ancient Hebrew language. For half of the dataset, we selected Biblical verses as divided into discrete categories in relation to literary genre or time-period that have been standardly distinguished in Biblical Hebrew scholarship as exhibiting different linguistic features: "narrative"[9], "legal"[10], "prophetic"[11], "poetry"[12], "wisdom'[13] and "late"[14]. Note that the list is not unequivocally accepted and that the same text

[4]Aramaic parts were discarded.

[5]Lines with reconstructed or missing text were discarded.

[6]For more information about the Dead Sea Scrolls, please refer to djd (1955–2009).

[7]https://hb.openscriptures.org/

[8]https://github.com/ETCBC/dss

[9]30 verses; from "Gen", "Exod", "Num", "Josh", "Judg", "Ruth", "1 Sam", "2 Sam", "1 Kgs", "2 Kgs", "Esth", "Jonah"

[10]15 verses; from "Deut", "Lev"

[11]15 verses; from "Isa", "Jer", "Ezek", "Hos", "Joel", "Amos", "Obad", "Mic", "Nah", "Hab", "Zeph", "Hag", "Zech", "Mal"

[12]15 verses; from "Lam", "Ps", "Song"

[13]10 verses; from "Eccl', "Job', "Prov'

[14]15 verses; from "1 Chr", "2 Chr", "Dan", "Ezra", "Neh"

could be classified as belonging to more than one category ("Eccl" may be "late" as well as "wisdom"), in which case we have made selections in view of quantitative balance.

In order to specifically test our models' proficiency in Ancient as opposed to Modern Hebrew, we based the second half of the dataset around five phenomena as absent/rare in the latter but commonly encounterable in the former (20 verses per phenomenon): "wayyiqtol narrative chains", "combination of infinitive absolute and finite verb", "verb subject object", "suffixed possessive pronouns" and "particles". The last category includes verses that contain the particles ה[15], הנה־נה[16], נא[17] and the relative particle אשר (in place of the particle –ש, which came to gradually be more common). The included phenomena are not meant to be exhaustive of the differences between the concerned language varieties.

For the purpose of evaluating the performance of the BEREL and GPT models on data that thay have not been trained on, we also created an alternative dataset of 200 DSS lines (100 Biblical and 100 extra-Biblical). In the case of the Biblical lines, BEREL has already encountered the same text with alternative orthography, and the extra-Biblical lines are entirely unfamiliar to it. We do not have this level of knowledge concerning the GPT model's training, but a similar tendency is expected in view of the data's general online presence.

The morphological model was evaluated on 10% of the original dataset.

The data used for model evaluation was removed from the training datasets.

4. Methods

12 different versions of the BEReshiT model were trained, each corresponding to a different number of "frozen" layers within the original DictaBERT model. As with DictaBERT, standard BERT-based training was followed. Tokenisation was based on full words rather than subword tokens, as proven optimal in relation to DictaBERT (Shmidman et al., 2023). A simple grid search was used in the configuration of each model. The strongest performing model in terms of accuracy, which had 9 frozen layers (3 trainable layers), was trained for 3 epochs with a learning rate of 0.0001 and word-mask probability of 0.2. All models were trained on one GPU Nvidia L40S 45GB. Training time ranged between 16.72 min (for 12 trainable layers) and 79.42 min (for 9 trainable layers). The models were evaluated on a cloze test based on the evaluation dataset as presented in 3.2. A random word was masked within each verse/line.

For the purpose of inter-model comparison, we opted for the following selection of models: DictaBERT, BEREL[18], mBERT (Devlin et al., 2019) and GPT-4o (OpenAI, 2024). A comparison of our models with DictaBERT allows for clear evaluation of the improvement resulting from the fine-tuning process. It is also meaningful to juxtapose our models to BEREL, which specialises in Rabbinic Hebrew, a stratum that resembles Ancient Hebrew more than Modern Hebrew does. Whilst remaining within the framework of BERT models, mBERT[19], which was trained on 2k tokens of Modern Hebrew data, allows for a comparison between language-specific and multilingual transformer models in relation to the Hebrew language[20]. Finally, OpenAI's GPT-4o is included as a representative of LLMs. The model is well documented as exhibiting strong performance on non-English text (Zhu et al., 2024; Harigai et al., 2025; Shvartz et al., 2025). Within our prompt, the model is instructed to return solely the missing word in each example and to make a guess even if unsure. In the case of GPT and BEREL, evaluation is also performed on the separate DSS-based dataset as likely unseen by them during train time.

As a way to gain further insight about the compared models' characteristics, we also performed qualitative evaluation based on a sample of their predictions, focusing on the following: orthographic variations, (near-)synonyms, complex verb forms and infrequent words.

To develop the morphological model, task fine-tuning was applied using the BEReshiT encoder[21] and multi-task morphology heads for all morphological features. Loss was computed only for non-NA labels. Tokens were predicted with reference to their local context rather than in isolation. The values for features irrelevant for the predicted POS (e.g. "conjugation" for non-verbs) were forced to be "NA". Once again, a grid search was conducted. The strongest model (mean F1 0.76) was trained for 5 epochs at a learning rate of 0.0001 and with maximal input length of 256 tokens. Macro-F1 was utilised as primary metric as a means to provide a comprehensive view on the model's capacities, independent of the relevant complexity of the predictable information. A comparison was performed between our morphological model (named BEReshiT-morph) and DictaBERT's counterpart,

[15] question particle

[16] presentative particle/discourse marker; related forms are also included

[17] denotes entreaty/politeness

[18] as per its most recent version, 3.0

[19] as per BERT-BASE-MULTILINGUAL-CASED

[20] As the model uses WordPiece subwords, in the case of the gold label becoming multiple tokens, a pseudo-likelihood is computed through masking one token at a time.

[21] Our strongest model was used.

DictaBERT-morph. For the purpose, the relevant morphological categories were mapped to the best of our abilities (see Table 3).

5. Results

5.1. Quantitative

5.1.1. Main Models

Please refer to the results presented in Table 1. For more detailed results, including per category in addition to dataset, please refer to Appendix A. Overall, the reported accuracies are low, which is explainable through the challenging nature of the evaluation dataset (e.g. a large percentage of low-frequency words in Biblical text). Consistently lowest performance by a large margin is associated with mBERT. The best global performance is demonstrated by the GPT and BEREL models. BEReshiT scores better than BEREL solely for the "prophetic" category. However, it is crucial to note that BEREL is trained on the Sefaria repository[22], which includes the full text of the Hebrew Bible (Shmidman et al., 2022), thereby causing train-test contamination. When only the DSS examples from the same dataset are considered, the model's accuracy drops significantly to 0.18 (2 correct predictions out of 11); to go further, both of the model's correct answers come from Biblical scrolls[23], which it has already encountered during training, albeit with different orthography. In turn, GPT is trained on large amounts of undisclosed online data, which is likely to also include Biblical texts. Once again, when only DSS texts are considered, the accuracy drops to the low value of 0.09, the only correct answer being from the Biblical scroll 1QIsaa.

When GPT and BEREL results are disregarded due to train-test contamination, BEReshiT demonstrates the highest accuracy for our main evaluation dataset, surpassing DictaBERT by a margin ranging from 0.02 to 0.09 in its different configurations. Increase is especially high for the following phenomena: "particles", "vso", and "wayyiqtol"[24]. The globally strongest BEReshiT model is the one with 9 frozen layers (for subset "by phenomenon", the models with 3 and 4 frozen layers perform slightly better).

We went on to evaluate the two contaminated models as well as DictaBERT on the additional DSS-only dataset[25]. In this context, performance decreased significantly, results being by far worse in the case of extra-Biblical texts as likely completely unseen by the models. The fact that DictaBERT's performance was identical to BEREL's on unseen text reassures us that the choice of DictaBERT as the model to fine-tune has not constituted a compromise in terms of potential performance.

An accuracy of 0 is occasionally demonstrated for portions of the data: notably, by mBERT as well as by most models for the "wisdom" category. In these cases, the alternative metric "probability of gold" (i.e. the probability that the model attributes to the correct label, whether it is its first prediction or not) helps provide an interpretable non-zero (albeit still very low) value. "Probability of gold" is close in value to "accuracy" for all experiments, implying that there are no significant cases of correct labels being rated highly whilst not representing the top choice. Interestingly, the metric is typically lower than "accuracy" for all models except mBERT, for which the tendency is reversed. The model can therefore be judged to spread its probabilities more evenly among candidates. Note that "probability of gold" is not retrievable for the proprietary model GPT.

5.1.2. Morphological Submodel

Table 2 shows the performance of the model BEReshiT-morph in predicting each of 9 morphological features (see Appendix B.1 for more detailed results). The mean F1 score across the full evaluation dataset comes at 0.76. When the different coverage of each feature is accounted for (column "cov.-weighted"), the value increases to 0.82[26]. Notably, the result for POS tagging is 0.97. Albeit not uniformly, performance is slightly lower for the DSS portion of the dataset and slightly higher when only extra-Biblical DSS tokens are considered. Note that here, as train-test contamination is not an issue, these trends reflect solely linguistic differences in the subsets.

[22] https://www.sefaria.org/texts

[23] 1QIsaa and 4Q5, which denote, respectively, the books Isa and Gen.

[24] respectively, from 0.15 to 0.35, 0.15 to 0.35, and 0.3 to 0.5

[25] Naturally, BEReshiT models, which were trained on this data, were not evaluated.

[26] Not every category is relevant for every evaluated token. For instance, POS tags are available for 0.84 of the evaluation data, whilst "suffix type" labels - for 0.02.

	main dataset			DSS-only		
	overall	by category	by phenomenon	overall	Biblical	extra-Biblical
GPT-4o	**0.54 (-)**	**0.58 (-)**	0.50 (-)	**0.24 (-)**	**0.39 (-)**	**0.09 (-)**
mBERT	0.05 (0.06)	0.04 (0.05)	0.06 (0.07)			
DictaBERT	0.27 (0.21)	0.33 (0.25)	0.22 (0.16)	0.12 (0.07)	0.16 (0.09)	0.08 (0.05)
BEREL	**0.54 (0.49)**	0.53 (0.49)	**0.54 (0.49)**	0.20 (0.16)	0.32 (0.26)	0.08 (0.05)
BEReshiT (f.l.:1-11)	0.32 (0.22)	0.36 (0.26)	0.27 (0.17)			
BEReshiT (f.l.:1-10)	0.34 (0.25)	0.37 (0.29)	0.30 (0.22)			
BEReshiT (f.l.:1-9)	0.36 (0.25)	0.40 (0.29)	0.32 (0.21)			
BEReshiT (f.l.:1-8)	0.35 (0.25)	0.39 (0.28)	0.31 (0.21)			
BEReshiT (f.l.:1-7)	0.33 (0.24)	0.38 (0.28)	0.28 (0.20)			
BEReshiT (f.l.:1-6)	0.30 (0.22)	0.35 (0.26)	0.25 (0.19)			
BEReshiT (f.l.:1-5)	0.29 (0.21)	0.33 (0.24)	0.25 (0.18)			
BEReshiT (f.l.:1-4)	0.34 (0.25)	0.34 (0.26)	0.33 (0.23)			
BEReshiT (f.l.:1-3)	0.34 (0.25)	0.34 (0.27)	0.33 (0.23)			
BEReshiT (f.l.:1-2)	0.32 (0.21)	0.33 (0.24)	0.30 (0.19)			
BEReshiT (f.l.:1)	0.33 (0.23)	0.35 (0.26)	0.30 (0.20)			
BEReshiT (no f.l.)	0.32 (0.23)	0.36 (0.26)	0.27 (0.19)			

Table 1: Results of the evaluation of the discussed models on the main dataset and the DSS-only dataset (as well as their subsets). Primary metric: accuracy. Secondary metric (noted between parentheses): probability of the gold label, where retrievable. The best global results are indicated in **bold** and the best results when models with likely train-test contamination are disregarded are underlined. Numbers are rounded to the second digit after the decimal point. f.l.: frozen layer(s).

	by morphological feature (F1)									**overall (F1)**	
	pos	subpos	stem	conj.	person	gender	number	state	suffix type	**mean**	**cov.-weighted**
full	0.97	0.88	0.72	0.79	0.74	0.73	0.72	0.64	0.63	0.76	0.82
DSS	0.94	0.76	0.72	0.54	0.72	0.71	0.67	0.62	0.62	0.70	0.77
extra-Biblical DSS	0.94	0.78	0.73	0.60	0.96	0.92	0.96	0.62	0.49	0.78	0.85

Table 2: Morphological evaluation results for BEReshiT-morph: full dataset, DSS subset (14.98%) and extra-Biblical part of the DSS subset (5.66%). The reported metric is macro-F1. Numbers are rounded to the second digit after the decimal point.

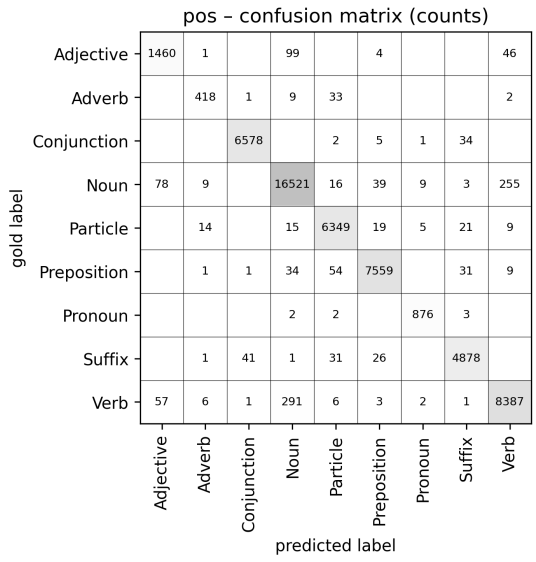

Figure 1: Confusion matrix of model BEReshiT-morph for the feature “pos”.

What follows is an overview of the model’s most common errors. Feature “pos” is mistakenly labelled “noun” when the gold label is “verb” 3.32% of times, and the reverse takes place 1.51% of times (see Fig. 1). “2nd person” is occasionally marked as “3rd” (1.31%) and vice-versa (0.65%). “Plural” number is marked as “singular” in 1.87% of cases, and “feminine” gender is marked as “masculine” in 4.46%, whilst the reverse occurrences are much less significant. Within feature “subpos”, “proper” is wrongly labelled as “common” in 2.33% of instances. Also, “common” is sometimes predicted as “UNK” (1%). Wrong “UNK” labels are also prominent when it comes to gold values “participle” (feature “conjugation”), “piel” (feature “stem”) and

"pronoun" (feature "suffix type")[27]. For the full confusion matrices per morphological feature, please refer to Appendix B.2.

As a next step, we took on to compare the performance of BEReshiT-morph to that of DictaBERT-morph, a morphological submodel based on DictaBERT. The first challenge came in the face of the two models' different tokenisation (subword-based for our model vs word-based for DictaBERT-morph). To guarantee fairness, we carried out the evaluation on DictaBERT-morph using a non-segmented version of the dataset, following which we segmented all prefixes and associated them with the labels that the model indicated as relevant to them. We proceeded to reduce the evaluation dataset to only the verses where tokenisation was now identical between the two models, which left us with a total of 920 verses. Finally, we mapped the morphological labels as existent in the two models (for the original label sets, please refer to Appendix C). This implied the occasional removal, combination and division of labels as well as the replacement of others with "NA". Please see Table 3 for the derived labels. As expected due to the morphological differences between Ancient and Modern Hebrew, DictaBERT-morph's performance in the task is significantly inferior (see Table 4). When only DSS tokens as well as only extra-Biblical DSS tokens are considered, the gap between the two models' performance increases further.

5.2. Qualitative

We conducted an additional microscopic, qualitative-based analysis, with the goal of gaining deeper knowledge about the performance of the investigated models, including different versions of BEReshiT (namely, the strongest performing version, where 9 layers are frozen during training, and the version that underwent full training).

See Table 5 for a representative example of the models' output. Common problems that are associated with all models, but most prominent with mBERT, are predictions consisting merely of punctuation, word fragments and simple particles, prepositions or pronouns. mBERT proposed punctuation in 73 out of the 200 examples of the main evaluation dataset, followed by DictaBERT (23), whilst for the BEReshiT models with 3 and 12 trained layers, there are respectively only 2 and 1 examples of the phenomenon.

Orthographic differences. The BEREL model is associated with the highest number of "wrong answers" that are in fact simply spelling variations of the gold label (15 over all 400 examples) i.e. one is the plene (including matres lectionis, typically more archaic) and the other is the defective spelling of the same word. Interestingly, there is no specific tendency as to the type of spelling preferred by the model as well as by DictaBERT and GPT, which are also associated with several examples of a predicted orthographic variant of the gold label[29]. There are only two examples of orthographic differences with the gold label for the BEReshiT

Feature	DictaBERT-morph label	Mapped label	BEReshiT-morph label
POS	VERB AUX	verb	verb
	NOUN	common noun	noun (common)
	PROPN	proper noun	noun (proper)
	ADJ NUM	adjective	adjective
	ADV	adverb	adverb particle (accusative-marker)
	ADP	preposition	preposition particle (object-marker)
	CCONJ	conjunction	conjunction
	DET	definite article	particle (article/definite)
	PRON	pronoun	pronoun
	SCONJ	relative particle	particle (interrogative)
		interrogative particle	particle (interrogative)
	INTJ	*NA*	
	SYM	*NA*	
	X	*NA*	
	suffix[28]	suffix	suffix
tense	Past	perfect	perfect
	Pres	participle	participle
	Fut	imperfect,wayyiqtol, jussive	imperfect wayyiqtol jussive
	Imp	imperative	imperative
		NA	cohortative
		NA	infinitive-absolute
		NA	infinitive-construct
person	1	1	1
	2	2	2
	3	3	3
	1, 2, 3	*NA*	
gender	Masc	masc	masc
	Fem	fem	fem
	Fem,Masc	comm	comm
number	Sing	sing	sing
	Plur	plur	plur
	Dual	dual	dual
suffix type	PRON ADP_PRON	pronoun	pronoun
		NA	directional *he*

Table 3: Mapping of the morphological features of models DictaBERT-morph and BEReshiT-morph. The values between parentheses denote BEReshiT-morph "subpos" labels.

[27] Respectively, in 4.1%, 6.86%, and 7.42% of cases.

[29] e.g. BEREL predicted תשכון (tiškon "dwell-FUT.2SG.M") in place of תשכן but also דוד (david "David") in place of דויד.

	by morphological feature (F1)						overall (F1)	
	pos	tense	person	gender	number	suffix type	**mean**	**cov.-weighted**
full dataset								
BEReshiT-morph	0.96	0.72	0.74	0.73	0.73	0.50	0.76	0.82
DictaBERT-morph	0.54	0.42	0.68	0.49	0.46	0.46	0.51	0.52
DSS								
BEReshiT-morph	0.94	0.37	1.00	0.95	0.93	0.49	0.78	0.91
DictaBERT-morph	0.55	0.34	0.61	0.46	0.43	0.49	0.47	0.50
extra-Biblical DSS								
BEReshiT-morph	0.95	0.48	1.0	0.94	0.95	1.0	0.89	0.93
DictaBERT-morph	0.30	0.26	0.40	0.30	0.24	0.21	0.28	0.29

Table 4: Comparison of the morphological submodels BEReshiT-morph and DictaBERT-morph using mapped features and labels for the full evaluation dataset vs the DSS subset (8.48%) vs the extra-Biblical DSS subset (4.09%). The reported metric is macro-F1. Numbers are rounded to the second digit after the decimal point.

Reference	**GPT-4o**	**mBERT**	**DictaBERT**	**BEREL**	**BEReshiT**
משק	משק	*punctuation*	*punctuation*	משק	סורר
mešeq	mešeq	–	–	mešeq	sorer
“steward, heir”	“steward, heir”	–	–	“steward, heir”	“rebellious”
ויאמר אברם אדני יהוה מה תתן לי ואנכי הולך ערירי ובן *משק* ביתי הוא דמשק אליעזר (Gen 15:2)					
But Abram said, “O Lord GOD, what will you give me, for I continue childless, and the *heir* of my house is Eliezer of Damascus?” (Gen 15:2)					

Table 5: Representative example of predictions provided by the investigated models. Typically for train-test contamination, BEREL and GPT guess the gold label. mBERT and DictaBERT propose punctuation. BEReshiT proposes a wrong vocabulary item.

model that underwent full training: לא/לוא (lo “no”) and אלהם/אליהם (alehem “to them”). In contrast, the model with three trained layers predicts the gold label for the later (the defective spelling), remaining with only a single example of the phenomenon.

Synonyms. Another type of prediction that deviates from the gold label but that may be considered correct is that of a synonym or close synonym. In the context of the verse ואל־בני ישראל תאמר איש איש מבני ישראל ומן־הגר הגר בישראל אשר יתן מזרעו למלך מות יומת עם הארץ ירגמהו באבן (Lev 20:2)[30][31], the gold label איש (iš “man, one”), which is predicted correctly by the contaminated BEREL model[32], is synonymous with the prediction put forward by both BEReshiT models as well as by their ancestor, DictaBERT - כל[33]. To go a little further, there are cases where an incorrect prediction is, while not strictly speaking synonymous, still relevant in terms of both morphology and semantics. For instance, the two BEReshiT models propose the word מים[34] in place of ארץ[35] in the following context: פנית לפניה ותשרש שרשיה ותמלא ארץ (Ps 80:10)[36]. Here, only the GPT model makes the correct prediction, whilst both BEREL and Dicta propose merely punctuation.

Complex verb forms. Morphologically complex words, such as conjugated verbs in less frequently used tenses, are challenging for all models. For instance, instead of the verb יחפרו[37], the

[30] “Any of the Israelites or of the aliens who reside in Israel who give any of their offspring to Molech shall be put to death; the people of the land shall stone them to death.”

[31] Biblical references follow the New Revised Standard Version (NRSV) versification. English translations are taken from the NRSV.

[32] but, interestingly, not by GPT, which predicts מזרעו (mizaro “from his seed”)

[33] kol “all, everyone”

[34] mayim “water”

[35] ereṣ “land”

[36] “You cleared the ground for it; it took deep root and filled the land.”

[37] yaḥperu “dig-FUT.3PL.M”; יחפרו בעמק וישיש בכח יצא לקראת נשק “It paws violently, exults mightily; it goes out to meet the weapons.” (Job 39:21)

models GPT and DictaBERT propose other POS (respectively, the noun גיבור[38] and the noun/adjective הקרב[39]), whilst BEREL, mBERT and BEREshiT select verbs with more common meanings and conjugations (respectively, בא[40], יצא[41] and וישב[42]). Occasionally, in cases of complex verb forms as gold labels, the two BEReshiT models make different predictions. For example, the word כפופים (kefufim "bend.down-PTCP.M.PL"[43]) is predicted as a fragmentary plural suffix by the BEReshiT model with three trainable layers and as אפים (apaim, "faces; noses") by the fully trainable one. In both cases, the morphological aspects of number and gender (plural masculine) are correct.

Infrequent words. Finally, we were interested in finding out how the different models behave in relation to rarely occurring words. Models that exhibit train-test contamination are likely to perform significantly better with such examples, as with the outputs of models BEREL and GPT in Table 5. In particular, we looked at hapax legomena (i.e. words that appear only a single time in a given text), a common phenomenon in the Hebrew Bible. Within the main evaluation dataset, we defined as hapax legomena masked words that appear only once in the full Hebrew Bible dataset[44]. For the Biblical DSS examples, we included words that do not reappear in the dataset and appear up to once in the full Hebrew Bible dataset[45]. In the case of extra-Biblical DSS examples, we counted only words that do not appear in the Hebrew Bible or elsewhere in the dataset. However, it is possible that words with the same root but different morphological features are present within the concerned data.

In the case of the main evaluation dataset, out of the 10 discovered hapax legomena, BEREL and GPT each made two correct predictions, whilst the two BEReshiT models and DictaBERT all guessed correctly solely the word מסיג (masig "remove/shift-PTCP.M.SG.SUBST")[46]. mBERT made no correct predictions. BEREL and GPT did significantly worse in relation to the DSS dataset, once again demonstrating the effects of train-test contamination. 13 Biblical and 8 extra-Biblical hapax legomena were found in the DSS dataset. No model made a correct prediction for an extra-Biblical hapax, and BEREL made a single correct prediction for a Biblical hapax: ויושביה (ve-yošveha "and-inhabit-PTCP.M.PL.CSTR+3FS")[47]. It is worth noting, however, that the word appears 7 times in the Hebrew Bible dataset without the prefixed conjunction and is therefore, strictly speaking, not a hapax.

6. Discussion

The BEReshiT model demonstrates high results on a specially crafted dataset that emphasises the characteristics of Ancient Hebrew and its differences with the language's Modern counterpart. GPT-4o and BEREL surpass the model in performance, but their behaviour on alternative data proves that the root cause for their high results is train-test contamination. The multilingual model mBERT performs very poorly, pointing at the higher suitability of a language-specific BERT-based model given the Hebrew language's resourcedness and characteristics.

Starting from the model with 11 frozen layers, performance increased until reaching its peak for the model with 9 frozen layers (best overall result as well as best result for the "by category" portion of our main evaluation dataset). Then, the results started deteriorating gradually, whilst seeing another increase for 4 and 3 frozen layers. Traditionally, BERT's upper layers have been associated with syntax-related information (Tenney et al., 2019; Hewitt and Manning, 2019). Our models' increased performance when these layers are fine-tuned speaks of efficient re-learning of syntax, notably captured by the subset "by category". Interestingly, the subset "by phenomenon" exhibited a differing trend, reaching a peak in results when earlier layers were also trained, possibly proving a higher relevance for the model's knowledge of vocabulary and semantics, language aspects that have been associated with these layers.

7. Conclusion and Future Work

We presented the Ancient Hebrew model BEReshiT, which was achieved through the fine-tuning of the state-of-the-art Modern Hebrew model, DictaBERT. When evaluated on elaborately

[38] gibbor "hero, warrior"

[39] ha-qerav/ha-qarov "the battle; the nearby"

[40] ba "come-PST.3SG.M"

[41] yaṣa "go.out-PST.3SG.M"

[42] vayešev "sit-PST.3SG.M"; both BEReshiT models make this prediction

[43] יהוה פקח עורים יהוה זקף כפופים יהוה אהב צדיקים "the Lord opens the eyes of the blind. The Lord lifts up those who are bowed down; the Lord loves the righteous." (Ps. 146:8)

[44] Words that appear twice due to the doubling of part of the dataset were not counted as hapax legomena, although they may be such within the Hebrew Bible, as the BEReshiT models have already encountered them during training.

[45] as the data is the same with possibly different orthography

[46] ארור מסיג גבול רעהו ואמר כל העם אמן " 'Cursed be anyone who moves a neighbor's boundary marker.' All the people shall say, 'Amen!' " (Deut 27:17

[47] ויושביה כמו כן ימותון וישועתי לעולם תהיה וצדקתי לוא תחת "and those who live on it will die like gnats,but my salvation will be forever, and my deliverance will never be ended." (1QISAA: 42 line 21, part of Isa 51:6)

constructed tasks, BEReshiT exhibits better performance than its ancestor model as well as a selection of other relevant Hebrew and multilingual models (when train-test contaminated models are disregarded). A closer qualitative look shows that even when wrong, the model's predictions tend to be sound in terms of morphology and semantics. The highest performing BEReshiT model is the one with 9 frozen layers. We are releasing this model and BEReshiT-morph, a strong submodel trained for morphological tagging of Ancient Hebrew text.

The proposed BEReshiT model would benefit from further evaluation on different tasks and possibly different data. The latter would be tricky to realise due to the scarcity of Ancient Hebrew data. Possible directions may include the use of letters and inscriptions from the same historical period, a selection of which is present through the Qumran-Digital repository [48] or of synthetic data (possibly derived from Modern Hebrew based on hand-crafted rules or style-transfer methods). Our future plans also include the application of BEReshiT within a pipeline of stemmatological analysis of Ancient Hebrew text that uses morphological tagging in the definition of textual variants.

8. Ethics Statement

The datasets used for model training and evaluation are based on publicly available data.

9. Limitations

The exhibited results are highly dependent on the utilised evaluation datasets. Different data as well as different task settings, such as masking at the subword level (e.g. of isolated suffixed pronouns), could have decreased the challenge. The definition of hapax legomena would have also differed given alternative tokenisation.

The comparison of our morphological submodel and DictaBERT-morph also comes with limitations: the evaluation dataset was reduced to identically tokenised verses only; also, the two models retrieve different morphological information. Although we have mapped the represented morphological features in an exhaustive way, the logic of label attribution is partly subjective and may differ between the models.

10. Bibliographical References

1955–2009. *Discoveries in the Judean Desert*. Clarendon Press, Oxford.

David Bamman and Patrick J. Burns. 2020. Latin bert: A contextual language model for classical philology.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT)*, pages 4171–4186.

Simon Gabay, Pedro Ortiz Suarez, Alexandre Bartz, Alix Chagué, Rachel Bawden, Philippe Gambette, and Benoît Sagot. 2022. From freem to d'alembert: a large corpus and a language model for early modern french.

Ayaka Harigai, Oshitaka Toyama, Mitsutoshi Nagano, Mirei Abe, Masahiro Kawabata, Li Li, Jin Yamamura, and Kei Takase. 2025. Response accuracy of gpt-4 across languages: insights from diagnostic radiology questions. *Japanese Journal of Radiology*.

Michael A. Hedderich, Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. A survey on recent approaches for natural language processing in low-resource scenarios. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2545–2568. Association for Computational Linguistics.

John Hewitt and Christopher D. Manning. 2019. A structural probe for finding syntax in word representations. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4129–4138, Minneapolis, Minnesota. Association for Computational Linguistics.

Aaron D. Hornkohl. 2014. *Ancient Hebrew Periodization and the Language of the Book of Jeremiah: The Case for a Sixth-Century Date of Composition*, volume 74 of *Studies in Semitic Languages and Linguistics*. Brill, Leiden.

Avi Hurwitz. 2014. *A Concise Lexicon of Late Biblical Hebrew: Linguistic Innovations in the Writings of the Second Temple Period*. Vetus Testamentum Supplements. Brill, Leiden.

Ganesh Jawahar, Benoît Sagot, and Djamé Seddah. 2019. What does BERT learn about the structure of language? In *Proceedings of the*

[48] https://www.qumran-digital.org/

57th Annual Meeting of the Association for Computational Linguistics, pages 3651–3657, Florence, Italy. Association for Computational Linguistics.

Pratik M. Joshi, Sebastin Santy, Ameya Godbole, Kalika Bali, and Chandra Sekhar Mukherjee. 2020. The state and fate of linguistic diversity and inclusion in the nlp world. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6282–6293. Association for Computational Linguistics.

Geoffrey Khan, Shmuel Bolozky, Steven E. Fassberg, Gary A. Rendsburg, Aaron D. Rubin, Ora R. Schwarzwald, and Tamar Zewi, editors. 2013. *Encyclopedia of Hebrew Language and Linguistics*, 1 edition, volume 1-4. Brill, Leiden.

Piroska Lendvai, Uwe Reichel, Anna Jouravel, Achim Rabus, and Elena Renje. 2023. Domain-adapting BERT for attributing manuscript, century and region in pre-Modern Slavic texts. In *Proceedings of the 4th Workshop on Computational Approaches to Historical Language Change*, pages 15–21, Singapore. Association for Computational Linguistics.

Josh McGiff and Nikola S. Nikolov. 2025. Overcoming data scarcity in generative language modelling for low-resource languages: A systematic review. *arXiv preprint*.

Ricardo Muñoz Sánchez, David Alfter, Simon Dobnik, Maria Irena Szawerna, and Elena Volodina. 2024. Jingle BERT, jingle BERT, frozen all the way: Freezing layers to identify CEFR levels of second language learners using BERT. In *Proceedings of the 13th Workshop on Natural Language Processing for Computer Assisted Language Learning*, pages 137–152, Rennes, France. LiU Electronic Press.

Iglika Nikolova-Stoupak, Maxime Amblard, Sophie Robert-Hayek, and Frédérique Rey. 2025. A classifier of word-level variants in witnesses of biblical Hebrew manuscripts. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 21313–21329, Vienna, Austria. Association for Computational Linguistics.

OpenAI. 2024. Gpt-4o system card. https://openai.com/index/gpt-4o-system-card/.

Robert Rezetko and Ian Young. 2014. *Historical Linguistics and Biblical Hebrew: Steps toward an Integrated Approach*, volume 9 of *Ancient Near East Monographs*. SBL Press, Atlanta.

Stefan Schweter, Luisa März, Katharina Schmid, and Erion Çano. 2022. hmbert: Historical multilingual language models for named entity recognition.

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

Avi Shmidman, Ometz Shmidman, Hillel Gershuni, and Moshe Koppel. 2024a. MsBERT: A new model for the reconstruction of lacunae in Hebrew manuscripts. In *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, pages 13–18, Hybrid in Bangkok, Thailand and online. Association for Computational Linguistics.

Shaltiel Shmidman, Avi Shmidman, Amir DN Cohen, and Moshe Koppel. 2024b. Adapting llms to hebrew: Unveiling dictalm 2.0 with enhanced vocabulary and instruction capabilities.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art BERT suite for modern Hebrew.

Elad Shvartz, Leah Attal, Omri Zur, Zaki Nujeidat, Gilad Plopsky, and Daniel Bahir. 2025. Bilingual comparison of the performance of gpt-4o and gpt-4 on multilingual medical questions. *European Journal of Radiology*.

Ian Tenney, Dipanjan Das, and Ellie Pavlick. 2019. Bert rediscovers the classical nlp pipeline. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4593–4601, Florence, Italy. Association for Computational Linguistics.

Wenhao Zhu, Hongyu Liu, Qingxiu Dong, Jingjing Xu, Shujian Huang, Lingpeng Kong, Jiajun Chen, and Lei Li. 2024. Multilingual machine translation with large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*.

A. Detailed Evaluation Results of the BEReshiT Model

A.1. Evaluation of dataset “Syntactic” by category

	overall					late					legal					narrative					poetry					prophetic					wisdom				
	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold
GPT-4o	**0.58**	**0.41**	**0.41**	**0.41**	–	0.47	0.30	0.30	0.30	–	0.47	0.33	0.33	0.33	–	**0.70**	**0.54**	**0.56**	**0.53**	–	**0.73**	**0.58**	**0.58**	**0.58**	–	**0.60**	**0.43**	**0.43**	**0.43**	–	**0.30**	**0.18**	**0.18**	**0.18**	–
mBERT	0.04	0.05	0.03	0.02	0.05	0.00	0.00	0.00	0.00	0.06	0.00	0.00	0.00	0.00	0.02	0.07	0.04	0.05	0.04	0.07	0.00	0.00	0.00	0.00	0.07	0.13	0.07	0.07	0.07	0.07	0.00	0.00	0.00	0.00	0.00
DictaBERT	0.33	0.19	0.20	0.20	0.25	0.26	0.15	0.15	0.17	0.14	0.40	0.26	0.26	0.26	0.30	0.53	0.33	0.34	0.34	0.40	0.27	0.17	0.17	0.17	0.25	0.20	0.12	0.12	0.12	0.19	0.00	0.00	0.00	0.00	0.00
BEREL	0.53	0.36	0.35	0.37	**0.49**	**0.53**	**0.37**	**0.36**	**0.38**	**0.42**	**0.60**	**0.47**	**0.47**	**0.47**	**0.62**	0.63	0.43	0.42	0.43	**0.59**	0.67	0.56	0.56	0.56	**0.66**	0.33	0.20	0.20	0.20	**0.28**	0.20	0.12	0.12	0.12	**0.20**
BEReshiT (f.l.:1-11)	0.36	0.21	0.21	0.22	0.26	0.20	0.11	0.10	0.12	0.15	0.47	0.32	0.32	0.32	0.33	0.53	0.33	0.34	0.33	0.41	0.33	0.20	0.20	0.20	0.23	0.33	0.20	0.20	0.20	0.18	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-10)	0.37	0.21	0.20	0.22	0.29	0.33	0.19	0.17	0.22	0.23	0.47	0.32	0.32	0.32	0.43	0.57	0.35	0.35	0.35	0.44	0.20	0.12	0.12	0.12	0.20	0.33	0.19	0.19	0.21	0.18	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-9)	0.40	0.22	0.22	0.24	0.29	0.33	0.19	0.17	0.22	0.23	0.53	0.38	0.38	0.38	0.44	0.6	0.38	0.38	0.38	0.43	0.27	0.15	0.15	0.15	0.20	0.33	0.20	0.20	0.23	0.17	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-8)	0.39	0.22	0.22	0.23	0.28	0.33	0.19	0.17	0.22	0.22	0.53	0.38	0.38	0.38	0.43	0.57	0.36	0.36	0.36	0.43	0.27	0.16	0.16	0.16	0.19	0.33	0.20	0.20	0.23	0.18	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-7)	0.38	0.22	0.22	0.24	0.28	0.40	0.25	0.24	0.26	0.27	0.40	0.26	0.26	0.26	0.37	0.50	0.30	0.31	0.30	0.40	0.27	0.15	0.15	0.15	0.19	0.40	0.26	0.25	0.27	0.23	0.10	0.06	0.06	0.06	0.04
BEReshiT (f.l.:1-6)	0.35	0.20	0.21	0.22	0.26	0.20	0.12	0.12	0.12	0.17	0.40	0.26	0.26	0.26	0.36	0.53	0.34	0.35	0.34	0.41	0.27	0.15	0.15	0.15	0.18	0.40	0.26	0.25	0.29	0.17	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-5)	0.33	0.18	0.18	0.19	0.24	0.20	0.11	0.10	0.12	0.19	0.47	0.32	0.32	0.32	0.32	0.53	0.33	0.34	0.33	0.39	0.20	0.12	0.12	0.12	0.17	0.27	0.15	0.14	0.17	0.17	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-4)	0.34	0.19	0.19	0.20	0.26	0.20	0.11	0.11	0.11	0.17	0.47	0.33	0.33	0.33	0.41	0.57	0.38	0.39	0.38	0.41	0.13	0.07	0.07	0.07	0.13	0.33	0.19	0.19	0.21	0.21	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-3)	0.34	0.19	0.20	0.20	0.27	0.27	0.15	0.14	0.16	0.21	0.40	0.26	0.26	0.26	0.35	0.53	0.35	0.36	0.36	0.43	0.20	0.12	0.12	0.12	0.15	0.33	0.20	0.20	0.22	0.25	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1-2)	0.33	0.18	0.18	0.19	0.24	0.27	0.16	0.16	0.16	0.18	0.47	0.32	0.32	0.32	0.38	0.50	0.31	0.32	0.30	0.38	0.13	0.08	0.08	0.08	0.11	0.33	0.20	0.19	0.22	0.16	0.00	0.00	0.00	0.00	0.00
BEReshiT (f.l.:1)	0.35	0.18	0.18	0.20	0.26	0.27	0.15	0.14	0.16	0.22	0.53	0.38	0.38	0.38	0.43	0.57	0.38	0.38	0.39	0.39	0.20	0.12	0.12	0.12	0.16	0.20	0.10	0.10	0.12	0.15	0.00	0.00	0.00	0.00	0.00
BEReshiT (no f.l.)	0.36	0.20	0.20	0.21	0.26	0.33	0.20	0.20	0.20	0.20	0.40	0.26	0.26	0.26	0.34	0.53	0.33	0.34	0.33	0.42	0.20	0.12	0.12	0.12	0.16	0.40	0.26	0.25	0.27	0.19	0.00	0.00	0.00	0.00	0.00

Table 6: Evaluation of dataset “Syntactic” by genre. Metrics: accuracy, F1 score (macro), precision (macro), recall (macro), average probability of gold label. The best global results per model and per subset are indicated in **bold** and the best results when models with likely train-test contamination are disregarded are underlined Numbers are rounded to the second digit after the decimal point. f.l.: frozen layer(s)

A.2. Evaluation of dataset “Syntactic” by phenomenon

	overall					inf.abs.fin.					part.					poss.					vso					wayyiqtol				
	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold	acc	f1	pr	r	prob.gold
GPT-4o	0.50	0.33	0.33	0.33	–	**0.55**	**0.39**	**0.39**	**0.39**	–	0.35	0.21	0.21	0.21	–	**0.55**	**0.38**	**0.38**	**0.38**	–	0.55	0.38	0.38	0.38	–	0.50	0.34	0.34	0.34	–
mBERT	0.06	0.03	0.03	0.04	0.07	0.10	0.05	0.06	0.05	0.06	0.10	0.05	0.06	0.05	0.01	0.00	0.00	0.00	0.00	0.05	0.00	0.00	0.00	0.00	0.01	0.10	0.05	0.05	0.06	0.13
DictaBERT	0.22	0.13	0.13	0.14	0.16	0.40	0.25	0.25	0.25	0.26	0.15	0.08	0.08	0.08	0.10	0.10	0.06	0.06	0.06	0.03	0.15	0.10	0.10	0.10	0.12	0.30	0.18	0.18	0.18	0.29
BEREL	**0.54**	**0.39**	**0.39**	**0.39**	**0.49**	**0.55**	**0.39**	**0.39**	**0.39**	**0.52**	**0.55**	**0.40**	**0.41**	**0.39**	**0.46**	0.40	0.25	0.25	0.25	**0.45**	**0.65**	**0.50**	**0.50**	**0.50**	**0.62**	**0.55**	**0.38**	**0.38**	**0.38**	**0.51**
BEReshiT (f.l.:1-11)	0.27	0.15	0.14	0.16	0.17	0.40	0.26	0.25	0.26	0.25	0.25	0.14	0.13	0.15	0.12	0.15	0.09	0.09	0.09	0.04	0.2	0.10	0.10	0.11	0.12	0.35	0.23	0.23	0.23	0.33
BEReshiT (f.l.:1-10)	0.30	0.16	0.16	0.17	0.22	0.35	0.21	0.20	0.22	0.30	0.30	0.17	0.16	0.18	0.16	0.15	0.08	0.08	0.08	0.07	0.30	0.18	0.18	0.18	0.19	0.40	0.25	0.25	0.25	0.36
BEReshiT (f.l.:1-9)	0.32	0.18	0.17	0.19	0.21	0.35	0.21	0.20	0.22	0.30	0.35	0.21	0.20	0.22	0.15	0.15	0.08	0.07	0.09	0.08	0.35	0.22	0.22	0.22	0.19	0.40	0.26	0.26	0.26	0.35
BEReshiT (f.l.:1-8)	0.31	0.17	0.17	0.19	0.21	0.40	0.25	0.24	0.26	0.30	0.25	0.13	0.13	0.15	0.14	0.15	0.08	0.07	0.09	0.09	0.35	0.22	0.22	0.22	0.19	0.40	0.25	0.25	0.25	0.34
BEReshiT (f.l.:1-7)	0.28	0.15	0.14	0.17	0.20	0.35	0.21	0.20	0.22	0.26	0.25	0.16	0.16	0.16	0.16	0.15	0.08	0.08	0.08	0.05	0.25	0.15	0.15	0.15	0.19	0.40	0.26	0.26	0.28	0.35
BEReshiT (f.l.:1-6)	0.25	0.13	0.12	0.14	0.19	0.35	0.21	0.21	0.22	0.26	0.20	0.10	0.10	0.11	0.14	0.05	0.03	0.03	0.03	0.04	0.25	0.14	0.14	0.15	0.15	0.40	0.26	0.26	0.26	0.35
BEReshiT (f.l.:1-5)	0.25	0.13	0.12	0.14	0.18	0.35	0.21	0.21	0.22	0.26	0.25	0.14	0.14	0.14	0.15	0.10	0.05	0.05	0.05	0.05	0.15	0.08	0.07	0.09	0.13	0.40	0.26	0.26	0.26	0.34
BEReshiT (f.l.:1-4)	0.33	0.19	0.18	0.20	0.23	0.45	0.31	0.30	0.33	0.32	0.30	0.17	0.16	0.18	0.17	0.25	0.15	0.15	0.15	0.09	0.30	0.18	0.18	0.18	0.20	0.35	0.22	0.21	0.23	0.35
BEReshiT (f.l.:1-3)	0.33	0.19	0.18	0.21	0.23	0.40	0.26	0.25	0.28	0.32	0.30	0.19	0.18	0.20	0.15	0.20	0.11	0.10	0.12	0.09	0.35	0.23	0.23	0.23	0.22	0.40	0.25	0.24	0.26	0.35
BEReshiT (f.l.:1-2)	0.30	0.17	0.16	0.18	0.19	0.35	0.21	0.20	0.22	0.26	0.30	0.17	0.17	0.17	0.14	0.15	0.08	0.08	0.08	0.06	0.20	0.11	0.11	0.12	0.14	0.50	0.34	0.34	0.34	0.36
BEReshiT (f.l.:1)	0.30	0.16	0.16	0.19	0.20	0.30	0.17	0.17	0.18	0.23	0.25	0.14	0.14	0.14	0.16	0.15	0.08	0.07	0.09	0.09	0.30	0.18	0.18	0.18	0.16	0.50	0.35	0.34	0.36	0.36
BEReshiT (no f.l.)	0.27	0.14	0.13	0.16	0.19	0.35	0.21	0.21	0.22	0.26	0.20	0.11	0.11	0.11	0.15	0.10	0.05	0.04	0.06	0.04	0.30	0.19	0.19	0.19	0.15	0.40	0.26	0.25	0.27	0.35

Table 7: Evaluation of dataset “Syntactic” by phenomenon. Metrics: accuracy, F1 score (macro), precision (macro), recall (macro), average probability of gold label. The best global results per model and per subset are indicated in **bold** and the best results when models with likely train-test contamination are disregarded are underlined Numbers are rounded to the second digit after the decimal point. f.l.: frozen layer(s)

B. Morphological Model Evaluation

B.1. BEReshiT-morph: Detailed Evaluation Results

	Morphological features																																							
	POS					sub-POS					stem					conj.					person					gender					number					state					suffix type				
	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.	acc	pr.	r.	f1	cov.
full	.98	.97	.97	.97	.84	.98	.89	.88	.88	.38	.93	.80	.68	.72	.12	.92	.82	.77	.79	.11	.98	.74	.74	.74	.16	.98	.73	.73	.73	.36	.99	.73	.72	.72	.36	.95	.65	.63	.64	.19	.98	.64	.61	.63	.07
OSHB	.98	.97	.97	.97	.86	.98	.89	.89	.89	.38	.93	.69	.59	.63	.12	.92	.82	.77	.80	.11	.99	.74	.74	.74	.17	.98	.73	.73	.73	.36	.99	.73	.73	.73	.36	.96	.65	.64	.65	.19	.98	.65	.61	.63	.08
DSS	.96	.94	.94	.94	.77	.96	.78	.76	.76	.35	.89	.74	.72	.72	.13	.89	.56	.52	.54	.13	.97	.72	.72	.72	.11	.97	.72	.71	.71	.35	.98	.69	.65	.67	.35	.92	.63	.61	.62	.18	.97	.62	.62	.62	.02

Table 8: Detailed evaluation results for model BEReshiT-morph (full dataset, OSHB subset, DSS subset). The following metrics are considered: accuracy, precision (macro), recall (macro), F1-score (macro), coverage (the fraction of evaluated token positions where the gold label is applicable). Numbers are rounded to the second digit after the decimal point.

B.2. BEReshiT-morph: Confusion Matrices

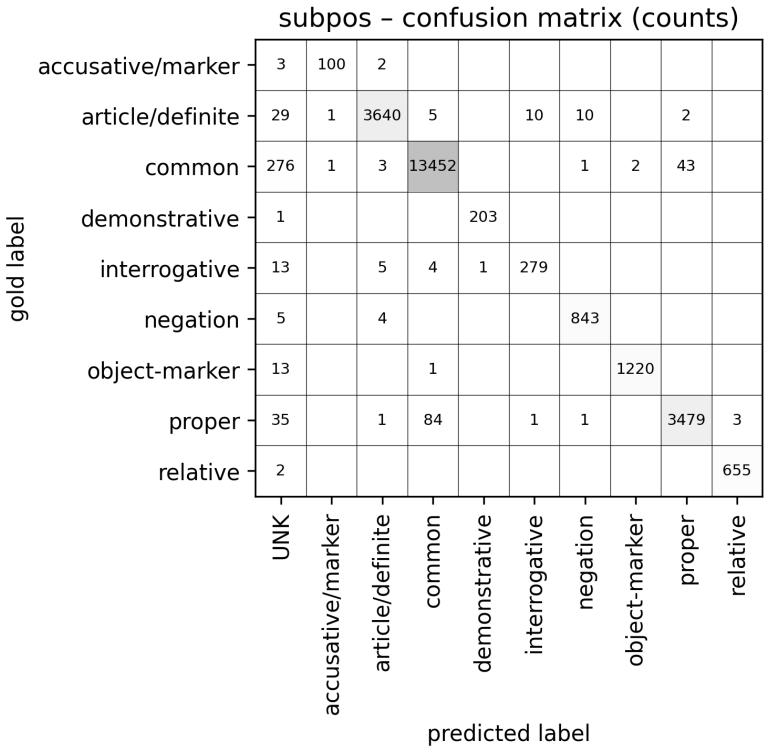

Figure 2: Confusion matrix for the feature “sub-POS”

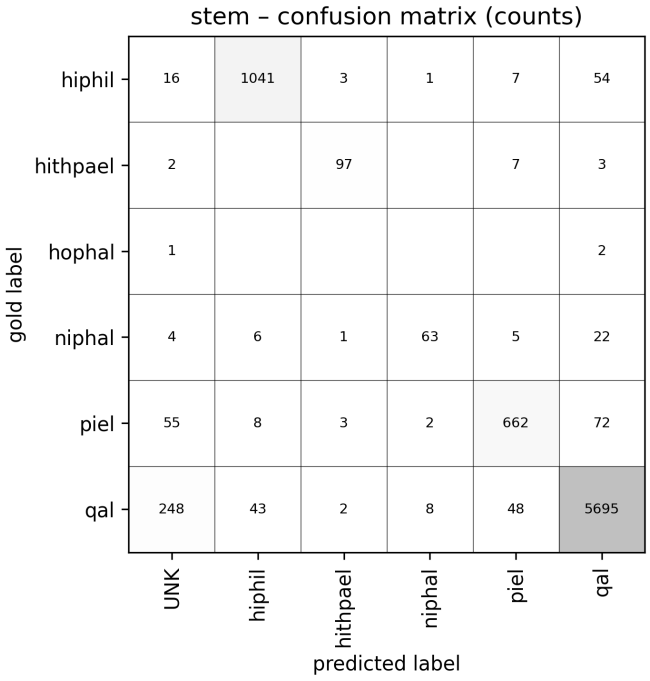

Figure 3: Confusion matrix for the feature “stem”

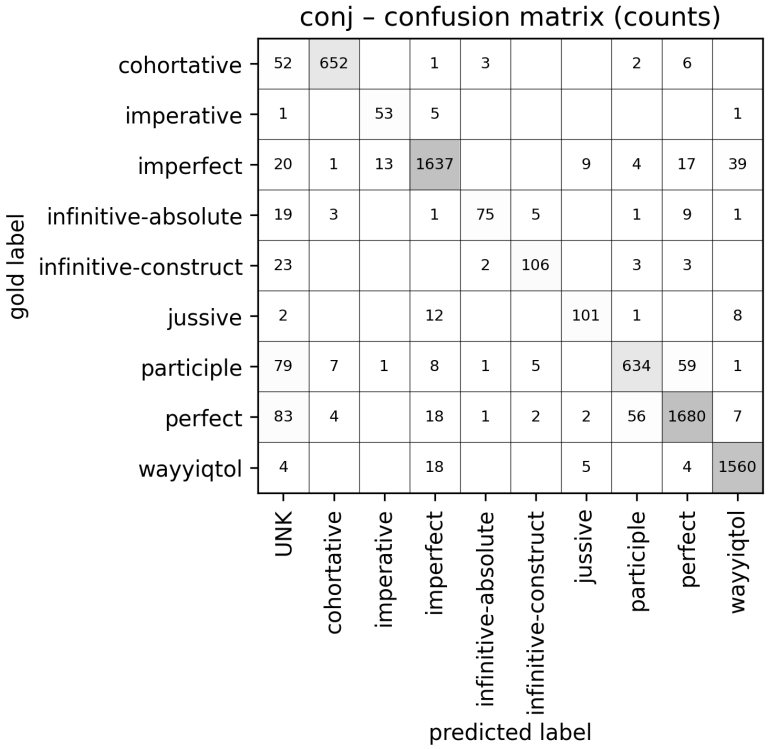

Figure 4: Confusion matrix for the feature “conjugation”

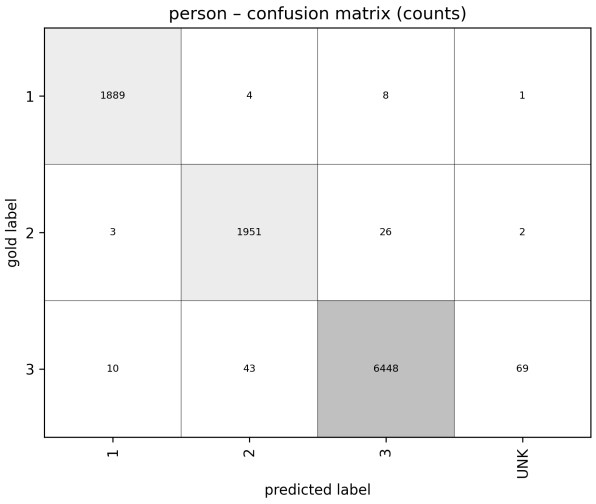

Figure 5: Confusion matrix for the feature “person”

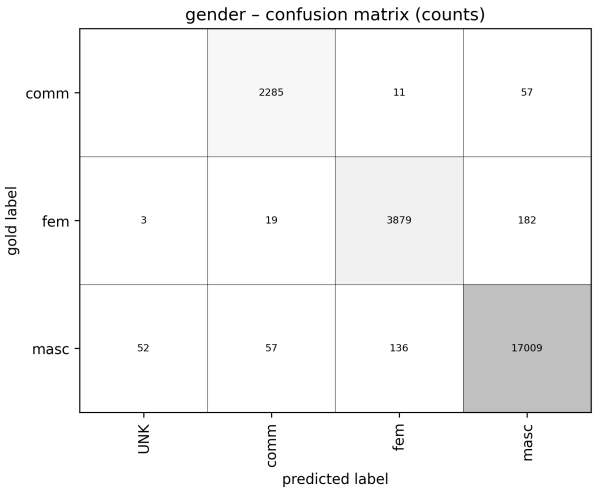

Figure 6: Confusion matrix for the feature “gender”

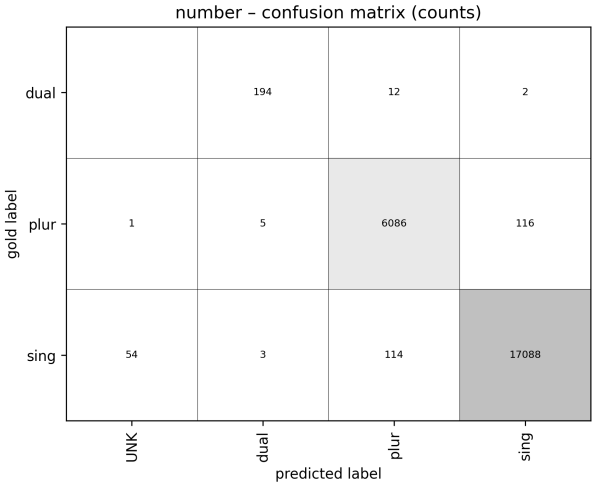

Figure 7: Confusion matrix for the feature “number”

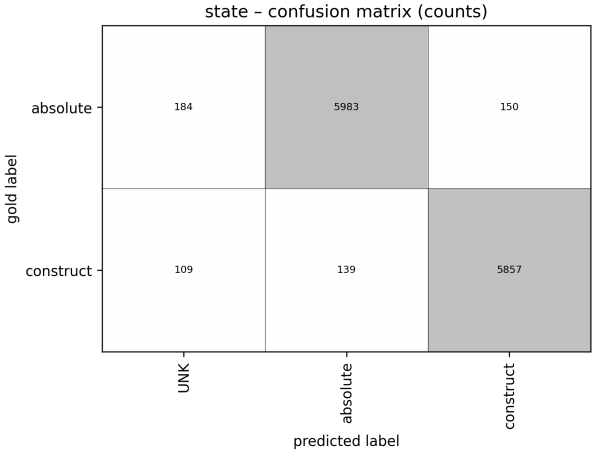

Figure 8: Confusion matrix for the feature “state”

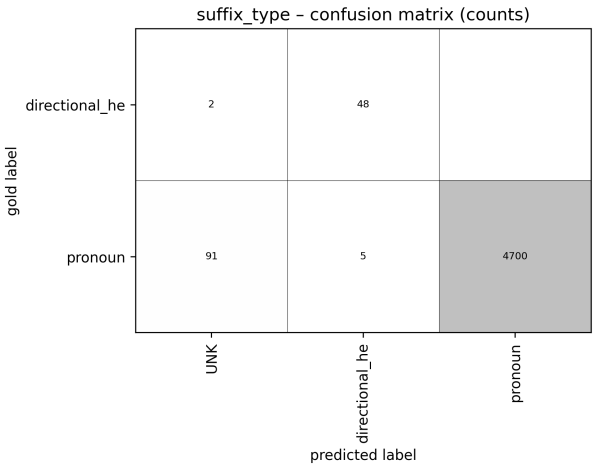

Figure 9: Confusion matrix for the feature “suffix type”

B.3. DictaBERT-morph: Confusion Matrix for POS

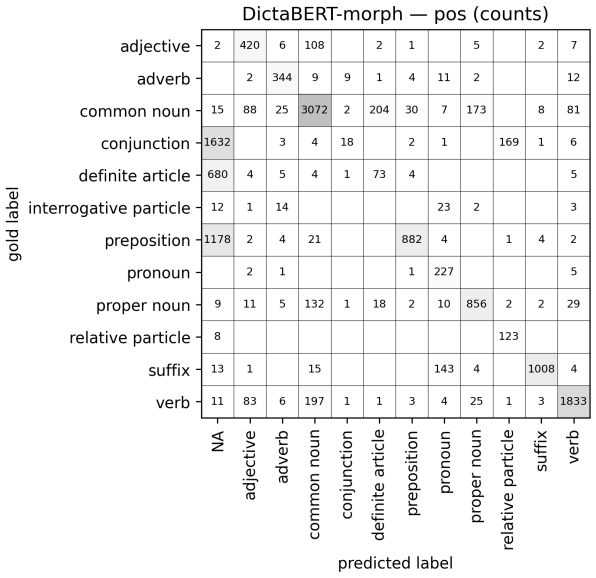

Figure 10: Confusion matrix for the feature “POS” for the model DictaBERT-morph.

C. Features and Labels as Present in the BEReshiT-morph and DictaBERT-morph Models

Feature	Label	Example	Ref.
POS	verb	יאמר	2 Sam 1:3
	noun	אזנים	Ps 40:7
	adjective	רקים	Prov 28:19
	adverb	שם	Gen 26:23
	pronoun	הוא	Num 9:13
	preposition	על	1 Kgs 6:35
	conjunction	ו	2 Chr 24:10
	particle	לא	Ps 119:157
	suffix	י	Ps 35:3
sub-POS	common	ימים	Judg 14:14
	proper	זבולן	Num 1:30
	article / definite	ה	Ps 24:10
	demonstrative	הנה	Esth 6:5
	interrogative	ה	2 Sam 3:33
	relative	אשר	Deut 23:11
	negation	לא	Lam 1:10
	accusative / marker	רק	2 Chr 28:10
	object-marker	את	Judg 19:29
conjugation	perfect	דבר	Jer 27:13
	imperfect	יהיו	2 Chr 12:8
	wayyiqtol	יערכו	2 Chr 14:9
	participle	באים	Jer 31:27
	imperative	נאכלה	Dan 1:12
	infinitive-construct	כתוב	Ps 149:9
	infinitive-absolute	רצח	Hos 4:2
	jussive	יחשב	2 Sam 19:20
	cohortive	המרות	Job 17:2
stem	qal	אמר	Exod 1:22
	niphal	יבא	Gen 2:22
	piel	ידברו	Jer 9:4
	hiphil	יזעק	Judg 4:10
	hophal	יכלכל	2 Chr 2:5
	hithpael	התרגז	Isa 37:29
person	1	אקח	Gen 14:23
	2	שמעת	Lam 3:61
	3	הוא	Jer 25:12
gender	masc	יבאו	2 Sam 20:14
	fem	ה-	Ezek 22:22
	common	אני	Job 33:6
number	sing	שקל	Exod 38:29
	pl	יחשבו	Ps 41:8
	dual	שני	Jer 46:12
suffix type	directional *he*	ה	Num 10:29
	pronoun	י	Eccl 2:3

Table 9: Morphological features determinable by the model BEReshiT-morph. Examples are retrieved from the golden labels (some tokens are at subword level).

Feature	Label	Example	Ref.
POS	VERB	ימול	Gen 17:13
	NOUN	בני	1 Sam 9:2
	ADJ	מלאה	Num 7:62
	ADV	מאד	Jer 48:29
	PRON	זה	Jer 52:28
	PROPN	חור	1 Chr 4:1
	ADP	ב	2 Kgs 18:17
	ADP_PRON	חללו	Ezek 7:22
	DET	ה	Josh 8:5
	NUM		
	AUX	היתה	1 Chr 4:10
	CCONJ	ו	1 Sam 14:37
	SCONJ	כי	Isa 52:5
	INTJ	הו	Isa 3:5
	PUNCT		
	SYM		
	X		
tense	Past	שגיתי	Job 6:24
	Pres	משמרים	Jonah 2:9
	Fut	יאכל	Eccl 5:16
	Imp	רדה	Gen 46:3
person	1	עלי	Job 27:23
	2	הביאת	Isa 43:23
	3	היתה	1 Chr 4:10
	1, 2, 3	חצובים	Neh 9:25
gender	Masc	אחי	Deut 2:8
	Fem	היתה	Gen 17:16
	Fem,Masc	בני	Exod 30:19
number	Sing	ראיתי	1 Sam 25:25
	Plur	ידם	Lev 10:3
	Dual	שנתים	1 Kgs 22:52

Table 10: Morphological features determinable by the model DictaBERT-morph. Examples are taken from the model's predictions, where available and correct (some tokens are at subword level).