OtoBERT:
Identifying Suffixed Verbal Forms in Modern Hebrew Literature

Avi Shmidman[1,2,†], **Shaltiel Shmidman**[1,‡]
[1]DICTA, Jerusalem, Israel
[2]Bar Ilan University, Ramat Gan, Israel
[†]avi.shmidman@biu.ac.il
[‡]shaltieltzion@gmail.com

Abstract

We provide a solution for a specific morphological obstacle which often makes Hebrew literature difficult to parse for the younger generation. The morphologically-rich nature of the Hebrew language allows pronominal direct objects to be realized as bound morphemes, suffixed to the verb. Although such suffixes are often utilized in Biblical Hebrew, their use has all but disappeared in modern Hebrew. Nevertheless, authors of modern Hebrew literature, in their search for literary flair, do make use of such forms. These unusual forms are notorious for alienating young readers from Hebrew literature, especially because these rare suffixed forms are often *orthographically identical* to common Hebrew words with different meanings. Upon encountering such words, readers naturally select the usual analysis of the word; yet, upon completing the sentence, they find themselves confounded. Young readers end up feeling "tricked", and this in turn contributes to their alienation from the text. In order to address this challenge, we pretrained a new BERT model specifically geared to identify such forms, so that they may be automatically simplified and/or flagged. We release this new BERT model to the public for unrestricted use.

1 Introduction

A primary obstacle for readers of Hebrew literature is the use of pronominal verbal suffixes. Hebrew allows the use of a bound suffix in place of a direct-object pronoun for virtually all object-taking verbs. For instance, the two-word Hebrew sentence **ראיתי אותו** (*raiti oto*, "I saw him") can be condensed into a single verb with pronominal suffix, with the equivalent meaning: **ראיתיהו** (*re'itihu*, "I saw him"). Although such suffixes are often utilized in Biblical Hebrew, their use is quite rare in modern Hebrew. Nevertheless, authors of modern Hebrew literature, in their search for literary flair, do select such forms at times. These unusual forms pose substantial difficulty for readers.

Prima facie, an effective solution to this obstacle in Hebrew literature would be to simplify the text (i.e., to convert instances of these rare suffixed verb forms into the equivalent pairs of non-suffixed verb followed by direct-object pronoun), or, alternatively, to add a gloss alerting the reader to the fact that a pronominal suffix is wrapped up inside the word.

Unfortunately, this is not a trivial procedure; because, it is not just that these forms are *rare*, but rather that they are often *ambiguous*; that is, they are often orthographically identical to common Hebrew words with very different morphological properties. To take a few examples:

- **הזמינו** ("they ordered" and "he ordered it")
- **לימדו** ("they taught" and "he taught him")
- **הגישה** ("she offered" and "he offered her")
- **ניהלה** ("she managed" and "he managed her")

Thus, we cannot automatically simplify or flag such words based on their letters alone; we can only do so if it can be inferred from the context that the suffixed analysis is intended. Nor would it make sense to flag every instance of such words as a cautionary measure; for, even in literary works, the overwhelming majority of these ambiguous forms are not in fact suffixed verbs. Flagging all of them would flood the reader with unnecessary alerts. Furthermore, as we will see below, existing NLP systems for Hebrew are not equipped to make this determination, because there are so few cases of suffixed verbs in their training data. To be sure, the question of how to optimally annotate Hebrew suffixed forms in training corpora for NLP systems has been explored (Tsarfaty and Goldberg, 2008). Nevertheless, at the end of the day, when faced with an ambiguous form that may or may not be a suffixed verb, existing morphological tagging systems for Hebrew too often blindly choose the usual non-suffixed form. Due to the fact the benchmarks used to evaluate these systems barely contain any cases of suffixed verbs, this myopic approach does

Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024), pages 12–19
November 15, 2024

not impact accuracy scores, and hence there is little motivation for the developers to address this shortcoming.

These ambiguous forms pose a formidable challenge for would-be readers of Hebrew literature. Upon encountering such words, readers naturally select the usual analysis of the word; indeed, in most cases, the analysis with the pronominal suffix is one that many readers will never before have encountered. Upon completing the sentence, they will find themselves confounded. Because there is no direct indication in the text that anything is special about the word, readers end up feeling "tricked" by the text, and this in turn contributes to the younger generation's alienation from Hebrew literature.

For example, in his novel *A Simple Story*, S. Y. Agnon writes (Agnon, 1953a, p. 76):

צייר גדול צייר את דמות תבניתה של בלומה וקבעה בלבו של הירשל...

"A great artist painted the image of Bluma *and he set it* in the heart of Hershel..."

The italicized phrase is the translation of a suffixed verb in the Hebrew. However, that same word is virtually always used as a non-suffixed verb, meaning "and she set". Upon first encountering the word, the reader will naturally adopt this latter analysis (with Bluma as the subject). Only at the sentence's completion will the reader be perplexed that the expected object of the verb "set" never materialized; this will hopefully trigger a reread, during which the reader will recognize the word as a suffixed verb.

In order to quantify just how big the gap is between standard Hebrew and literay Hebrew with regard to these suffixed verbs, our human annotators analyzed two corpora of daily Hebrew newspapers, as well a corpus of high-register Hebrew literature.[1] The results (table 1) reveal a sharp contrast: the literature corpus has over 35 times as many cases of suffixed verbal forms as do either of the newspaper corpora. Furthermore, within the literature corpus, a substantial number of the forms were ambiguous (64 cases), while each of the news corpora had but a single instance of an ambiguous suffixed verb. Thus, it is conceivable that a person could be completely proficient in reading Hebrew newspapers, yet never have had to cope with parsing an ambiguous suffixed verb.

[1]The literature corpus includes works of Hebrew novelists S. Y. Agnon, Haim Beer, Amoz Oz, and David Grossman. The news corpora are drawn from the daily Hebrew newspapers *Maariv* ("News1") and *Ynet* ("News2").

Corpus	**Corpus Size** (Words)	**Suffixed Verbs**	**Freq** (Per 10K Words)	**Ambig Cases**	**Freq of Ambig** (Per 10K)
News1	185K	7	0.38	1	0.05
News2	43K	2	0.47	1	0.23
Lit	135K	222	17.76	64	4.74

Table 1: Hebrew verbs with pronominal suffixes are exceedingly rare in regular newspaper text, but far more likely to appear in literary texts. We note the overall number of such verbs, and also the number of ambiguous cases which can be alternatively read as a more usual Hebrew word.

The present paper presents a new BERT model pretrained from scratch with the goal of providing a solution to this challenge. This new model provides a method to identify most of these troublesome forms with a high degree of confidence, in order to make these treasured literary works accessible to today's readers. We dub our model *OtoBERT* (after the Hebrew direct-object pronoun אותו (*oto*, "him").

2 Related Work

The challenge that we address in this paper - the task of identifying ambiguous Hebrew forms which unexpectedly serve as suffixed verbs - is a case of Complex Word Identification (CWI). CWI entails the identification of words which pose a challenge to readers of a given text, due to their unusual or complex usage within the context. CWI is a critical step within Text Simplification tasks, because it identifies the words that need to be simplified in order to make the text more accessible.

For a historical overview of machine-learning methods utilized for CWI, see North et al. (2023, pp. 14-23). Current methods for addressing CWI generally utilize BERT pretrained language models, leveraging the capabilities of the existing BERT models in a variety of different methods. For instance, Kelious et al. (2024) train a classifier for English CWI using embeddings produced by the pretrained model DeBERTa.

Other researchers have proposed emsemble methods which combine output from multiple BERT models. For instance, Bani Yaseen et al. (2021) utilize two different BERT models (standard BERT and RoBERTa) in order to produce four measurements for each word. Each word is evaluated in each model twice - once on its own, and once with its full context. All of these measure-

ments are then combined in a weighted scheme to produce an optimal measure. In a similar vein, Pan et al. (2021) fine-tune a wide series of BERT models (BERT, ALBERT, RoBERTa, and ERNIE) for the CWI task, and they combine the fine-tuned models via a stacking mechanism in order to produce a final prediction.

A novel method of leveraging BERT models is proposed by Kumbhar et al. (2023). They implement a procedure which follows the information flow of a given word through the hidden layers of the BERT model, measuring the complexity of the computation needed to process a given word with its context. This measured complexity is then used as a basis on which to perform CWI.

On the backdrop of these studies, our unique angle is that rather than building upon the foundation of existing pretrained BERT models, we pretrain a new dedicated BERT model from scratch, specifically designed to address our CWI challenge. We add a specialized step to the BERT tokenization training stage - prior to the pretraining of the model - in order to optimize its ability to identify the rare and elusive cases of Hebrew suffixed verbs.

3 Our Approach

In order to address the challenge of identifying verbs with a pronominal suffix, we seek to create a BERT model which is particularly sensitive to contexts which entail a <verb, direct-object pronoun> pair. We thus pretrain a new BERT from scratch, tokenizing it such all such cases of direct-object pronouns are combined together with the preceding word. For instance, the two-word sequence ראיתי אותו ("I saw him") would be stored as a single token in the BERT vocabulary, with an underscore in place of the space. The inclusion of these compound tokens in BERT's vocabulary allows the BERT model to directly learn representations for combined units of <verb, direct-object pronoun>, which correspond precisely to the meaning of the elusive suffixed verbs which we seek to identify. Furthermore, this allows BERT's masked language model (MLM) head to predict multiword tokens which consist of a verb followed by a direct-object suffix. We hypothesize that the prediction of such tokens at the position of an ambiguous Hebrew word will indicate that the word consists of a verb with pronominal suffix. This is because the two-word sequence of the verb followed by the direct-object pronoun is semantically equivalent to the verb with the bound pronominal suffix; they are two alternate ways of expressing the same thing in the Hebrew language.

Essentially, this extra tokenization step provides BERT with a new expressive power. The single-word vocabularies of existing Hebrew BERT models do not provide sufficient expressiveness to disambiguate Hebrew verbs with pronominal suffixes; even if the models properly analyze the word and predict a series of suffixed verbs for that word position, those words themselves will generally be ambiguous, and hence cannot serve to disambiguate the nature of the verb. In contrast, the vocabulary of our new model contains a large set of two-word phrases which each contain a verb together with a direct-object pronoun; thus, our model is equipped with the capacity to express token predictions that directly indicate that a given word position within the sentence can be occupied by a verb with pronominal suffix.

In the section below we provide details regarding the pretraining of this new BERT model; afterward we describe our experiments with it, our results, and our proposal for applying this model in practice to make Hebrew literature more accessible.

4 Model

4.1 Tokenizer

The first stage involves training a new word-piece tokenizer for BERT, optimally suited for encoding Hebrew texts in general, and for solving the issue of suffixed Hebrew verbs in particular. We use the Word-Piece tokenization method proposed by Song et al. (2021), with adjustments to handle the apostrophe and double-quote marks, which mark Hebrew abbreviations, and which otherwise would have been tokenized into separate word pieces.

Further, as discussed at length in the previous section, we add a preprocess procedure to the tokenizer which combines all cases of direct-object pronouns with the preceding word, treating these words pairs as single compound tokens.

Following previous work on Hebrew BERT models, (Gueta et al., 2023; Shmidman et al., 2023), the tokenizer was trained with a vocabulary size of 128,000 tokens.

4.2 Architecture

The model's architecture is based on BERT-base (Devlin et al., 2019a), trained using the same data and objectives as Shmidman et al. (2023), with the

adjustment of using the custom tokenizer described in the §4.1. For full details regarding the training details please see Appendix A.

5 Experimental Setup

Corpus: In order to properly evaluate OtoBERT, we assemble a corpus of naturally-occurring Hebrew sentences with ambiguous verbal forms, that is, homographs which can be analyzed either as a verb with pronominal suffix, or as a verb with no suffix at all. The homographs are manually annotated as to their correct analysis. The corpus contains a total of 2,589 instances of non-suffixed verbs, and 264 instances of suffixed verbs.

Classification based on BERT's MLM predictions: We run the MLM head of our new BERT model on each of the aforementioned homographs. For each case, we retrieve K predictions. If at least N of these predictions include compound tokens with direct-object pronouns, then we classify the word as a suffixed verb. We experiment with a range of values for both K and N.

For example, take the following sentence from S. Y. Agnon's *Only Yesterday* (Agnon, 1953b, p. 280): הקיפו וחזר והקיפו ונכנס ("*he encircled him*, and once again *he encircled him*, and entered"). Here, the (doubled) italics phrase "he encircled him" (a suffixed verb) corresponds to an ambiguous Hebrew word that can also mean "they encircled" (the usual analysis). If we take the initial occurrence of this ambiguous word in the sentence and run it through the MLM head of OtoBERT, the top 1000 predictions include numerous tokens with direct-object pronouns, including: ראו_אותו ("they saw him"), מצאו_אותו ("they found him"), and ראה_אותו ("he saw him"). OtoBERT's choice of these tokens among its predictions indicates that the context entails the use of a verb with pronominal suffix.

In contrast, take this sentence from the same novel (Agnon, 1953b, p. 294): אמרה שפרה עייף מר מן החום ("*Said* Shifra, master is tired from the heat"). The italicized word "said" corresponds to an ambiguous Hebrew verb which can also function as a suffixed verb, meaning "he said it". However, OtoBERT's top 1000 predictions for this word position don't include a single instance of a compound token with a direct-object pronoun, indicating that the context entails a regular non-suffixed verb.

Alternate Methods of Classification: In order to evaluate whether it was in fact necessary to train a completely new BERT model for this task, we also attempt to address this challenge using two standard methods of resolving Hebrew ambiguity. First, we use the SOTA morphological tagger available for Hebrew, DictaBERT (Shmidman et al., 2024) to tag the sentences in the corpus, and we measure whether it correctly assigned the "suffix" tag to the relevant verbs.

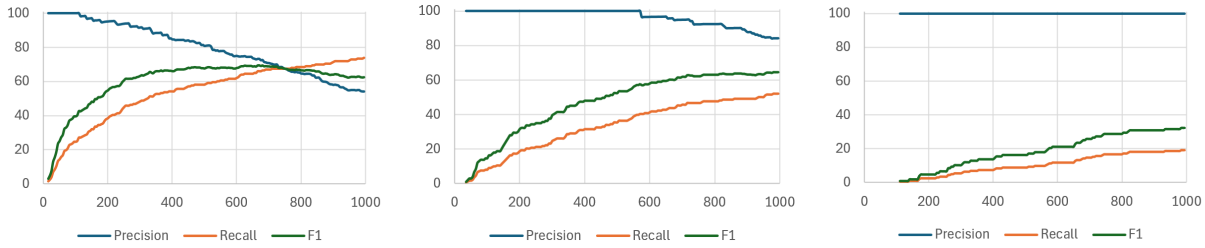
Second, we train a classifier to distinguish between cases of suffixed verbs and cases of non-suffixed verbs, based on the BERT embedding of the word. In order to avoid possible bias of one specific BERT model, we train classifiers separately for each of three BERT models with Hebrew support: mBERT, the original multilingual BERT, based upon Devlin et al. (2019b); AlephBERT, the impactful dedicated Hebrew BERT model produced by Seker et al. (2021), and finally with DictaBERT (Shmidman et al., 2023), the current SOTA of Hebrew BERT models. With each model, we train an MLP to recognize embeddings for the "suffixed verb" class by providing it with a corpus of sentences which have a verb followed by a direct-object pronoun, and we train it for the "non-suffixed verb" class via cases of unambiguous verbs which can only function as non-suffixed words. We then evaluate its ability to distinguish between suffixed verbs and non-suffixed verbs on the aforementioned test corpus.

6 Results

Results are displayed in Figure 1. We plot our method's performance across a range of values for the K and N thresholds. The performance is measured vis-a-vis the Suffixed Verb class; that is, the precision and recall lines depict the method's ability to pinpoint cases of Suffixed Verbs without falsely flagging non-suffixed verbs. In Table 2, we compare the results of our method (at K=1000, the highest-recall setting) with that of the two alternate methods mentioned above.

Model	Precision	Recall	F1
OtoBERT, K=1000, N=1	15.75	**95.57**	.270
OtoBERT, K=1000, N=5	54.15	73.89	.625
OtoBERT, K=1000, N=10	84.13	52.22	**.644**
OtoBERT, K=1000, N=25	**100**	19.21	.322
mBERT w/ Classifier	28.23	62.12	.388
AlephBERT w/ Classifier	38.92	62.50	.480
DictaBERT w/ Classifier	48.27	73.86	.584
DictaBERT Morph Tagger	88.73	23.86	.376

Table 2: Precision, recall, and F1 vis-a-vis the class of suffixed verbal forms for each of the evaluated methods.



(a) N=5 (b) N=10 (c) N=25

Figure 1: Precision and Recall for three different settings of N (the threshold for the number of compound tokens that must be predicted in order to support the classification of a word as a suffixed verb). The X-axis represents the K value (the number of predictions we retrieve form the MLM), and the Y-axis represents the Precision/Recall score.

7 Applying it to the Texts

Based upon the results in Table 2, we propose that OtoBERT, when run with K=1000, is sufficiently robust to automatically annotate Hebrew literary texts in a helpful and accessible way, as follows:

(1) With N set to 25, we achieve a precision of 100% on the test corpus; we conclude that given 25 or more predictions of compound tokens from the MLM head, we can be confident that a suffixed verb is intended. Given this confidence, we can directly simplify the text, breaking up the single suffixed verb into a non-suffixed verb with a subsequent direct-object pronoun. Alternatively, if we wish to avoid tampering directly with the literary text, we could add the simplified version as an interlinear gloss on top of the suffixed verb. This method provides recall of close to 20% of the cases.

(2) With N set to 10, we achieve a precision of over 84 percent. Although not sufficient to tamper directly with the text, this precision is sufficient to justify adding an interlinear gloss above the word qualified by the word "likely"; for instance, the gloss might read, "alert: likely a verb with bound suffix". Together with the previous step, this provides coverage of over half of the suffixed verbs in the test set.

(3) With N set to 5, we have a wide recall of over 72 percent, and we still have a precision of 54.15%. This might justify a gloss qualified by the word "perhaps".

Thus, OtoBERT can potentially provide a substantial readability boost to a Hebrew literary text. It can confidently identify a substantial set of suffixed verbs within a text, and for many other cases, it can mark the possibility of a suffixed verb while reliably indicating a lower confidence level. This paves the way for an automatic system which insert confident interlinear glosses where relevant, and qualified glosses at other places. In contrast, the other two methods don't provide a similarly versatile platform for simplifying the text. Our attempts to train classifiers on top of BERT models produce results of low precision (62.58%), which would not allow for any high-confidence glosses; and the DictaBERT Morphology Tagging system has very poor recall (22.11%), which would leave most suffixed forms unmarked.

8 Conclusion

In sum, OtoBERT provides a practical and effective method to automatically address a critical obstacle in the accessibility and readability of Hebrew literature. Through the creation of this new and specialized BERT model, we are able to identify suffixed verbs with a high degree of accuracy, enabling the simplification and/or glossing of most instances. Using this method, we can remove a primary stumbling block which alienates the younger generation of readers, ultimately paving the way for the new generation to enjoy the treasures that the Hebrew literary tradition has to offer.

We release OtoBERT to the public on huggingface, for both commercial and educational use, under an Apache license. Additionally, we release our dataset of naturally occurring Hebrew sentences containing ambiguous words which can be analyzed either as a verb without suffix or as a suffixed verb, with human annotations indicating the correct analysis for each.[2] We hope that this dataset will pave the way for additional studies further enhancing our ability to address this accessibility challenge of Hebrew texts.

[2]The model is available here: https://huggingface.co/dicta-il/otobert, and the dataset is available here: https://huggingface.co/datasets/dicta-il/hebrew_suffix_verbal_forms

9 Limitations

The method presented here relies on the ability to generate Masked Language Model predictions for the token which represents the ambiguous word. This entails the existence of the ambiguous word within the BERT vocabulary. If it is not present in the vocabulary, and the word is broken up into word pieces, then the MLM head cannot be relied upon to produce a reliable prediction for our purpose here, no matter which word piece we submit for the MLM predictions. Thus, the method presented here is limited to cases where the ambiguous word is contained in the BERT vocabulary as a single token.

Fortunately, in practice, this limitation affects only a small minority of cases. First of all, we pretrained the BERT model presented here with a substantially sized vocabulary, of 128K words, which means that from the get-go, most words in a modern Hebrew text need not be split into word pieces. Furthermore, the suffixed verbs that we focus upon here - the ones which are generally analyzed as a common non-suffixed Hebrew word, and which also contain the possibility of analysis as a verb with pronominal suffix - are, by their very nature, frequent words, which are most likely included in the vocabulary.

10 Lay Summary

In this paper, we address a specific obstacle which makes Hebrew literary texts difficult for students and youth: complex Hebrew words which are actually a series of multiple words combined together into one. For instance, instead of using multiple Hebrew words to say "and he threw it", they would all be combined into one complex Hebrew word. The problem is twofold. First of all, such complex words are exceedingly rare in modern Hebrew, outside of literary contexts. This already poses a difficulty for student readers who are not used to encountering such words. However, the real difficulty is that these complex words are often ambiguous: the very same Hebrew letters can be read as a different and non-complex Hebrew word, and that is the usual way that the word is used. Thus, it's not just that the students will be unfamiliar with the possibility of the complex word and not know how to understand it. Rather, it is that the students will recognize the word as a standard Hebrew word that they are used to seeing, and they will continue to read the sentence with that understanding. Yet, when they reach the end of the sentence, they will find themselves perplexed. When they are finally taught that the word in question actually doubles as a complex word, different in meaning from what they are used to, they feel tricked by the text, and this ends up alienating them from the literary treasures of the language.

To bridge this gap for student readers, we wish to design a system that automatically annotates these literary texts, adding little alerts or warnings in between the lines of the text in order to alert the reader to the fact that these words don't function here as they normally do. However, in order to do so, we need an automatic method to identify these complex words; and, because the words are ambiguous, this is not easy to do. As we demonstrate, the regular computational processes for clarifying text don't work well here, due to the extreme rarity of the complex words. We have therefore trained a new dedicated neural network language model, designed from the ground up specifically to identify this type of complex word. We release our new model here to the public.

Acknowledgements

This paper has been funded by the Israel Science Foundation (Grant No. 2617/22) and by the European Union (ERC, MiDRASH, Project No. 101071829), for which we are grateful. Views and opinions expressed are, however, those of the authors only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

S. Y. Agnon. 1953a. *At the Handles of the Lock*. Schocken Publishing House.

S. Y. Agnon. 1953b. *Only Yesterday*. Schocken Publishing House.

Tuqa Bani Yaseen, Qusai Ismail, Sarah Al-Omari, Eslam Al-Sobh, and Malak Abdullah. 2021. JUST-BLUE at SemEval-2021 task 1: Predicting lexical complexity using BERT and RoBERTa pre-trained language models. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 661–666, Online. Association for Computational Linguistics.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019a. Bert: Pre-training of

deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019b. Bert: Pre-training of deep bidirectional transformers for language understanding. *Preprint*, arXiv:1810.04805.

Eylon Gueta, Avi Shmidman, Shaltiel Shmidman, Cheyn Shmuel Shmidman, Joshua Guedalia, Moshe Koppel, Dan Bareket, Amit Seker, and Reut Tsarfaty. 2023. Large pre-trained models with extra-large vocabularies: A contrastive analysis of hebrew bert models and a new one to outperform them all. *Preprint*, arXiv:2211.15199.

Abdelhak Kelious, Mathieu Constant, and Christophe Coeur. 2024. Complex word identification: A comparative study between ChatGPT and a dedicated model for this task. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3645–3653, Torino, Italia. ELRA and ICCL.

Atharva Kumbhar, Sheetal Sonawane, Dipali Kadam, and Prathamesh Mulay. 2023. CASM - context and something more in lexical simplification. In *Proceedings of the 20th International Conference on Natural Language Processing (ICON)*, pages 506–515, Goa University, Goa, India. NLP Association of India (NLPAI).

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *Preprint*, arXiv:1907.11692.

Kai North, Marcos Zampieri, and Matthew Shardlow. 2023. Lexical complexity prediction: An overview. *ACM Comput. Surv.*, 55(9).

Chunguang Pan, Bingyan Song, Shengguang Wang, and Zhipeng Luo. 2021. DeepBlueAI at SemEval-2021 task 1: Lexical complexity prediction with a deep ensemble approach. In *Proceedings of the 15th International Workshop on Semantic Evaluation (SemEval-2021)*, pages 578–584, Online. Association for Computational Linguistics.

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. *Preprint*, arXiv:2104.04052.

Vitaly Shalumov and Harel Haskey. 2023. Hero: Roberta and longformer hebrew language models. *arXiv:2304.11077*.

Shaltiel Shmidman, Avi Shmidman, and Moshe Koppel. 2023. Dictabert: A state-of-the-art bert suite for modern hebrew. *Preprint*, arXiv:2308.16687.

Shaltiel Shmidman, Avi Shmidman, Moshe Koppel, and Reut Tsarfaty. 2024. MRL parsing without tears: The case of Hebrew. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 4537–4550, Bangkok, Thailand and virtual meeting. Association for Computational Linguistics.

Xinying Song, Alex Salcianu, Yang Song, Dave Dopson, and Denny Zhou. 2021. Fast wordpiece tokenization. *Preprint*, arXiv:2012.15524.

Reut Tsarfaty and Yoav Goldberg. 2008. Word-based or morpheme-based? annotation strategies for Modern Hebrew clitics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

A Appendix: Training Details

The model's architecture is based on the BERT-base architecture (Devlin et al., 2019a), trained on a DGX-A100 with 4xA100 40GB cards. The training was done with the fused lamb optimizer combined with AMP (Automatic Mixed Precision). A polynomial warmup learning rate scheduler was used to warm up for a portion of the training steps and then decay the learning rate over the total steps.

A.1 Training Data & Objectives

We train our model using the same training data & objectives as done for training `DictaBERT` (Shmidman et al., 2023). The training dataset is a mixture of several sources (such as the HeDC4 corpus (Shalumov and Haskey, 2023)), summing to a total of three billion words (3.8B tokens) of naturally occurring texts.

We trained the model using only the MLM (masked language modeling) training objective, as done by Liu et al. (2019). In addition, we adjusted the construction of the training examples for the MLM objective according to the guidelines specified by Shmidman et al. (2023). The main adjustments were:

1. We don't mask tokens that are broken up into multiple word-pieces since the non-masked word-pieces provide valuable information and make the task less challenging.

2. We never truncate part of a sentence, documents are always truncated with sentence units so that a training example is never cut off in the middle.

A.2 Training Details and Hyperparameters

We trained our model with the HuggingFace architecture wrapped with NVIDIA libraries[3] which are highly optimized for training compute-heavy machine learning models on NVIDIA hardware. We pre-trained the model on 4 A100 40GB GPUs for a total of 32,800 iterations, completing a total of 1.85 epochs. The training was done with sequences of up to 256 tokens, with 2 phases. First phase with a learning rate of 6e-3 for 1 epoch, followed by a second phase with a learning rate of 1e-4 for 0.85 epochs.

The total training time was 4.5 days. The training was done with a global batch size of 8,192 and a warmup proportion of 0.2843 for both phases.

[3] https://github.com/NVIDIA/DeepLearningExamples/tree/master/PyTorch/LanguageModeling/BERT