Do Pretrained Contextual Language Models Distinguish between Hebrew Homograph Analyses?

Anonymous NAACL-HLT 2021 submission

Abstract

Semitic morphologically-rich languages (MRLs) are plagued by word ambiguity; in a standard text, many (and often most) of the words will be homographs with multiple possible analyses. Previous research on MRLs claimed that standardly trained contextualized embeddings based on word-pieces may not sufficiently capture the internal structure of words with hugely ambiguous homographs. Taking Hebrew as a case study, we investigate the extent to which Hebrew homographs can be disambiguated using contextualized embeddings. We evaluate all existing models for contextualized Hebrew embeddings on 75 Hebrew homograph challenge sets. Our empirical results demonstrate that contemporary Hebrew contextualized embeddings outperform non-contextualized embeddings; and that they are most effective for disambiguating segmentation and morphological features, less so regarding pure sense disambiguation. We show that these embeddings are more effective when the number of word-piece splits is limited, and they are more effective for 2-way and 3-way ambiguities than for 4-way ambiguity. We show that the embeddings are equally effective for homographs of both balanced and skewed distributions. Finally, we show that these embeddings are as effective for homograph disambiguation with extensive supervised training as with a few-shot setup.

1 Introduction

Semitic morphologically-rich languages (MRLs) such as Arabic, Hebrew, and Aramaic are plagued by ambiguity at the word level (Wintner, 2014; Tsarfaty et al., 2020). In a standard text, many (and often most) of the words will be homographs with multiple possible analyses. The high ambiguity derives from several factors. First, prepositions, conjunctions, accusative pronouns, and possessive pronouns are often seamlessly affixed to words. Next, vowels are generally omitted in written texts. Also, proper nouns are not differentiated from common nouns (no capital letters).

Type	Form	Word (translation)	Morphology
Segmentation	הקפה	הַ+קָפֶה (the+coffee)	DET + Noun [M,S,abs]
		הַקָפָה (credit)	Noun [F,S,abs]
	שאף	שֶׁ+אַף (for+even)	Sconj + Cconj
		שָׁאַף (he aspired)	Verb [M,S,3,PAST]
Morph	אלימות	אַלִימוֹת (violent)	Adj [F,P,abs]
		אַלִימוּת (violence)	Noun [F,S,abs/cons]
	הרים	הֵרִים (he lifted)	Verb [M,S,3,PAST]
		הָרִים (mountains)	Noun [M,P,abs]
Semantic	הזמר	הַ+זֶמֶר (the+song)	DET + Noun [M,S,abs]
		הַ+זַמָר (the+singer)	DET + Noun [M,S,abs]
	הסופר	הַ+סוֹפֵר (the+author)	DET + Noun [M,S,abs]
		הַ+סוּפָר (the+market)	DET + Noun [M,S,abs]

Table 1: Examples of ambiguity types

Hebrew word ambiguities can be divided into three primary categories (Table 1): 1. Segmentation ambiguities, in which a raw token may be analyzed as a single standalone word, or segmented into multiple word units each bearing its own role (POS tag) in the sentence. 2. Morphological ambiguities, in which the segmentation of the token is not ambiguous, but the multiple analyses of the word reflect different morphological signatures (POS and morphological properties). 3. Semantic ambiguities, in which case the analyses have the exact same morphological signature, but differ on the semantic plane, in their *sense*; for a discussion of sense ambiguities, see Navigli (2009).

Previous research has claimed that pretrained contemporary language models (PLMs) that are based on word-pieces tokenization would not sufficiently capture the structure of MRLs in order to distinguish between internally-complex homograph analyses (Klein and Tsarfaty, 2020; Tsarfaty et al., 2020). In this work, we take Modern Hebrew, a Semitic language with rich and highly ambiguous morphology, as a case study, and investigate the extent to which Hebrew homographs can be disambiguated by contextualized embeddings.

Hebrew is a particularly challenging language on which to perform a homograph disambiguation due to the limited available corpora. First of all, the only currently existing Hebrew treebank contains less than 100K words, such that most of the words in the language are not amply represented. Further-

more, even regarding common Hebrew words, this corpus is problematic, because the nature of Hebrew homographs is that many of them are skewed in their distribution; thus, even if the primary analysis is sufficiently represented within a tagged corpus, the secondary analysis will often be hopelessly underrepresented. For instance, the ratio of the two analyses of the form מהם (*mhm*) in naturally-occurring Hebrew text is 1:187, and thus the instances of the secondary analysis within existing tagged corpora are not sufficient to allow for proper evaluation. To be sure, even English NLP suffers from this issue when it comes to less frequent homographs in the long-tail distribution (Chen et al., 2021); yet in Hebrew, this challenge is present with many of the most frequent words in the language. For analogous cases in other languages, researchers have proposed creation of dedicated challenge sets, containing hard-to-classify sentences not easily found in naturally-occurring text (Gardner et al., 2020; Elkahky et al., 2018). Here too, in order to evaluate the performance of disambiguation approaches for Hebrew homographs, it is critical to produce dedicated challenge sets with ample representation of all possible analyses.

A recent study (Shmidman et al., 2020) produced challenge sets for 22 Hebrew homographs, and demonstrated that a Bi-LSTM of non-contextualized embeddings can obtain high accuracy on this task, establishing the current SOTA. For the use of word2vec for disambiguation, see Iacobacci et al. (2016). In this paper, we extend the investigation by considering whether contextualized emeddings from pretrained language models (PLMs) can provide a more optimal solution. Previous studied conjectured that contemporary PLMs may not suffice for disambiguating homographs in MRLs (Klein and Tsarfaty, 2020). Here we consider all existing contextualized PLMs for Hebrew: the multilingual BERT (henceforth, "mBERT") (Devlin et al., 2019); HeBERT (Chriqui and Yahav, 2021); and AlephBERT (Seker et al., 2021) (Table 2), and assess their suitability for the task.

Our experiments demonstrate that contextualized PLMs trained on sufficiently large data and vocabulary size are excellent at disambiguating the word-internal structures of homographs, yet face some challenge with pure sense disambiguation. We show the efficacy of these models in cases of homographs with skewed distribution, and in few-shot learning. All in all, we provide new state-of-the-art results on this challenging task and confirm the adequacy of PLMs for morphological tasks.

Model	Vocab (Heb. tokens)	Corpus Size (Heb. sentences)
mBERT	2.5K	6.3M
HeBERT	30K	27.2M
AlephBERT	52K	98.7M

Table 2: Comparison of available Hebrew BERT models

2 The Data

The existing challenge sets produced by Shmidman et al. (2020) are limited in number (only 22 sets) and very unbalanced in terms of the types of ambiguities that they covered (only one of the sets involved a prefix-segmentation ambiguity). They are limited to binary cases, where only two possible analyses exist. Finally, they do not all represent frequent Hebrew words; the authors included a number of relatively infrequent words because the data happened to be easily accessible.

In contrast, for this study we employed field experts to choose the most critical homographs in the language. The experts chose 75 homographs from a list of the 3600 most frequent words in the language, balancing frequency of word occurrence with practical need for its disambiguation. Our challenge sets include homographs with 2-5 possible analyses. Our sets contain a wide representation of segmentation ambiguities (15 in number), as well as 5 cases of purely semantic ambiguities. For each of the 75 homographs, we collected 1000 naturally-occurring sentences attesting to the primary analysis, at least 500 sentences attesting to each secondary analysis, and at least 300 for each additional analysis. The sentences were culled from newspapers, Wikipedia, literature, and social media. All in all, our 75 challenge sets contain 161K tagged sentences. The full list of homographs and analyses is provided in Appendix A.

3 Experimental Setup

We set out to evaluate the ability of embeddings based on pre-trained language models to disambiguate the in-context analyses of morphologically rich and highly ambiguous tokens in Hebrew. In order to do so, we create dedicated "word expert" classifiers for each homograph (Zhao et al., 2020).

We use two types of PLMs, contextualized and non-contextualized. For the non-contextualized case, we replicate the method used by Shmidman et al. (2020). For each training example, we use a BiLSTM to encode the word2vec embeddings of the full sentence. An MLP is trained to predict the correct homograph analysis based on the BiLSTM encoding.[1] For the contextualized case, we run

[1]We use 100-dimension word2vec embeddings trained

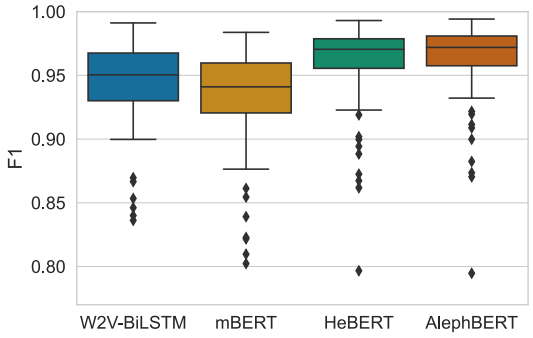

Figure 1: Comparison of previous SOTA (w2v-based Bi-LSTM method) versus BERT-based approaches.

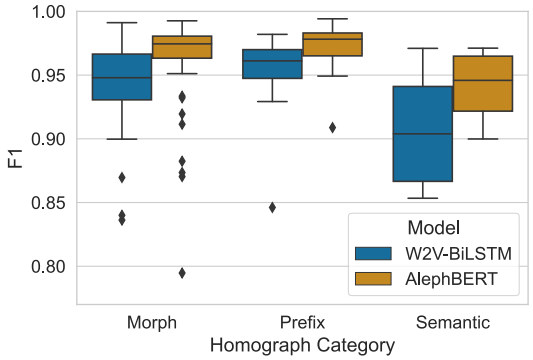

Figure 2: Categories of homograph ambiguity.

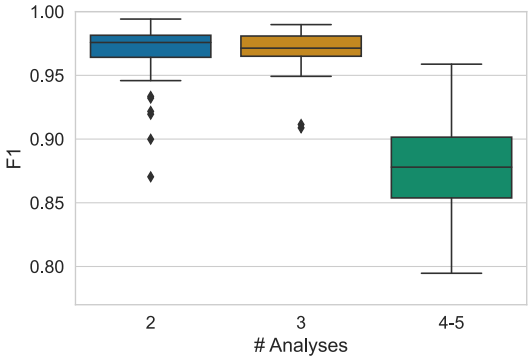

Figure 3: Homographs with differing option counts.

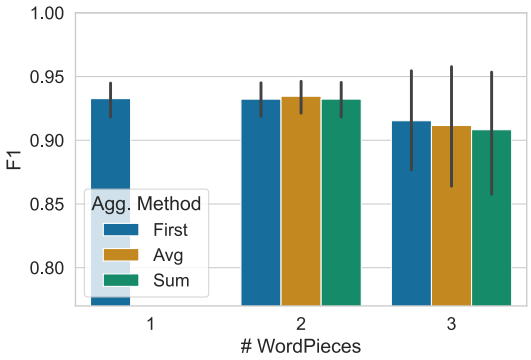

Figure 4: Word-piece splits using mBERT.

the sentence through a pretrained contextualized language model and retrieve the 768-dimension embedding representing the homograph in question. An MLP is trained to predict the correct analysis based on the homographs embeddings alone.

All BiLSTMs and MLPs are trained using dynet (http://dynet.io/). We use 2-layer MLPs with a hidden layer of size 100. We train with the Adam optimizer at a learning rate of .001, for 3 epochs.

We evaluate the performance of each given method on each given challenge set using 10-fold cross-validation. We calculate an F1 score for each homograph analysis, based upon the precision and recall scores micro-averaged across all folds. We then calculate the macro-average of the F1 scores for all possible analyses for a given homograph, and this is the score reported in the charts herein.

4 Results and Analysis

Figure 1 presents the cumulative F1 score obtained by the models for all challenge sets. Our results show that HeBERT and AlephBERT far outperform mBERT, with AlephBERT achieving the higher score. The poor performance of mBERT is likely due to its smaller pre-training data size and exceedingly lean Hebrew vocabulary (cf. Table 2). Furthermore, the HeBERT and AlephBERT models both substantially outperform the previous word2vec-based SOTA. It is thus apparent that *contextualized* language models do effectively capture Hebrew homograph distinctions, even those based on word-pieces, even for an MRLs, and they do so more effectively than non-contextualized models.

Figure 2 demonstrates AlephBERT's performance on different ambiguity types. AlephBERT performs equally well on cases of segmentation ambiguity and morphological ambiguity. In contrast, when it comes to ambiguities that are purely semantic, the scores are noticeably lower. This is in line with the findings of Ettinger (2020), who shows that BERT is stronger with syntax than semantics; Goldberg (2019) also notes BERT's strong syntactic abilities. Interestingly, the same gap exists with the W2V-based method. Thus, both contextualized and non-contextualized embeddings struggle to differentiate between senses which are morphologically equivalent. Although such cases are only of minimal importance when it comes to sentence parsing, they are critical for downstream tasks such as coreference resolution and relation extraction. It thus remains a desideratum to improve disambigua-

on a 500M-word Hebrew corpus using Yoav Goldberg's word2vecf, adding position info to context words (https://github.com/BIU-NLP/word2vecf). We also tried fastText embeddings, but results were inferior.

tion of purely semantic Hebrew homographs.

The results in Figure 3 demonstrate that AlephBERT performs equally well on cases of binary homographs as on cases of three-way homograph classification. However, when faced with cases of four-way classification, accuracy declines.

The Effect of Word-Pieces Previous studies have hypothesized that word-pieces are not adequate for capturing complex morphological structures due to arbitrary (non-linguistic) word-splits. To probe into this we investigate the question, do such splits affect performance. Our 75 homographs are all treated as single tokens in HeBERT and AlephBERT. However, many of the homographs are broken up into word pieces in mBERT, due to its meager Hebrew vocabulary. We thus compare mBERT's results on words treated as single tokens versus those that are broken up into two or three pieces, which are aggregated using first, sum, or average of the vectors. As shown in figure 4, the splitting of a homograph into three word-pieces appears to have a negative impact on the ability of the resulting embedding to differentiate between homograph analyses, for all aggregation methods.

Skewed Homographs As noted, many homographs are skewed, such that one analysis will appear dozens of times more often than the other analysis in naturally-occurring text. We consider whether the pretrained embeddings might be disproportionately influenced by the skewed distribution. Our tests show that AlephBERT's scores do not degrade even as the ratio of the homographs become more and more skewed (full data in Appendix B).

Few-Shot Scenarios In our experiments thus far, the 10-fold cross-validation allows the MLP to leverage 90% of the data in each fold (hundreds of sentences for each analysis) in order to learn the difference between the analyses. We now consider whether the AlephBERT embeddings can suffice on a few-shot basis, where the training stage has access to only 100, 50, 25, 10 or even 5 examples of each analysis. In these cases, we train an MLP based only on these few samples, and we use the rest of the sentences for evaluation. Astoundingly, as demonstrated in Figure 5, the AlephBERT embeddings provide a highly accurate solution even on this few-shot basis. Even when training with only 5 examples of each homograph analysis, AlephBERT reaches an accuracy that is not far below the accuracy achieved when performing full 10-fold CV across hundreds of sentences of each analysis.

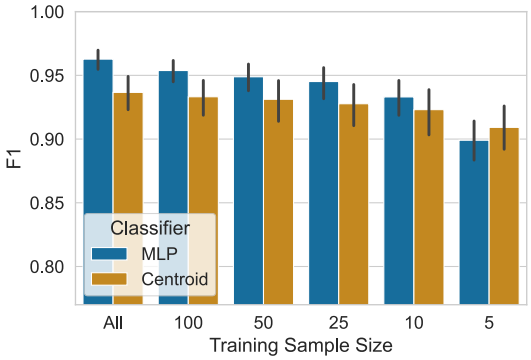

Figure 5: Use of AlephBERT embeddings to differentiate between homographs on a few-shot basis, contrasted with scores from the full 10-fold CV ("All").

Probing Scenarios Finally, we probe the pretrained AlephBERT embeddings (Yaghoobzadeh et al., 2019; Tenney et al., 2019; Klafka and Ettinger, 2020; Belinkov, 2021) to see whether in and of themselves they reflect clusters which correspond to different homograph analyses. We skip the MLP, and instead use the raw embeddings directly, classifying sentences based on their proximity to the centroid of the training samples for each homograph analysis. We use cosine distance to measure the proximity. As shown in the orange bars in Figure 5, this method generally does not perform as well as the MLP-based method; however, the degradation is limited to only a few percentage points, indicating that the raw embeddings are in fact clustered in groups which reflect the distinctions between the homograph analyses.

5 Conclusion and Future Work

In this study we have utilized a wide-ranging collection of Hebrew homograph challenge sets in order to evaluate the extent to which raw BERT embeddings can be leveraged to disambiguate Hebrew homographs. We found that contextualized embeddings can effectively disambiguate morphological analyses of homographs, much more so than non-contextualized ones. Yet, an increasing number of splits, or an increasing number of different possible analyses of a token, have a negative effect on this efficacy. We further discover that BERT embeddings can function effectively for this purpose on a few shot basis, with as little as 5 examples of each analysis. This indicates that with relatively modest effort, highly ambiguous homographs may be effectively treated. In the future we aim to consider zero-shot approaches as well, using clustering to differentiate between groups of embeddings, and using generic classifiers to determine the morphological properties of each of the clusters.

6 Ethical Statement

Creation of the Dataset As noted, our dataset contains over 161K sentences in all. Every sentence was reviewed and tagged by our team of human annotators, who chose the relevant homograph analysis for each instance of each of our 75 homographs. Our annotator team included members of diverse genders and sexual orientations. They were paid hourly wages with legal pay stubs. Their hourly wage was well above the minimum wage required by law.

We pre-filtered the corpus and removed sentences with offensive language, in order to ensure that our human annotators would not have to read offensive material. The pre-filter was based on a wide-ranging set of potentially problematic keywords. Nevertheless, we recognize that a keyword-based method cannot always succeed in filtering out every offensive sentence. We therefore also provided all the taggers with a "flag sentence" button in their graphic tagging interface. We encouraged them to press the button immediately and without hesitation upon encountering a sentence that seems at first glance to be offensive, so that they should not be forced to fully contemplate the sentence. Once flagged, the sentence is removed from our corpus and never again presented to our human taggers. Similarly, our taggers are encouraged to flag sentences which contain personal information about named individuals.

The sentences in the dataset are taken, in part, from Wikipedia (CC-BY-SA), and in part from copyrighted data scraped from public internet sites. The copyrighted data is used only for the purpose of this research evaluation, and will not be distributed. However, all tagged sentences originating from Wikipedia will be released with the acceptance of this article, together with the tagging information, under the CC-BY-SA license. To be sure, the original intended use of the Wikipedia texts was not for corpus-based research, but rather for the dissemination of knowledge to end-users. Nevertheless, the use of Wikipedia texts for corpus-based research is consonant with its access conditions.

Given that this paper is an empirical investigation, and that its primary purpose is to confirm specific hypotheses, we believe that this data split strikes the right balance between protecting the copyrighted rights of the content creators, and yet still providing the NLP community with a large set of Wikipedia-based sentences for evaluation and training of Hebrew homograph disambiguation systems.

Limitations of the Dataset We have made every effort to be as inclusive as possible in the creation of the dataset, making sure to include data from a widely diverse set of genres. A perennial challenge in corpus-based studies is that the lion's share of the available data tends to be authored by male writers. In order to offset this bias, we bolstered our corpus with a large corpus of texts specifically taken from blog sites devoted entirely to female bloggers. Nevertheless, it is likely that texts authored by women and by other minorities are underrepresented in our dataset.

A further limitation derives from the aforementioned filter regarding offensive language. Because we filtered out offensive-language sentences from the outset, our resulting tests necessarily do not reflect the performance of the systems when applied to sentences with offensive language, and resulting algorithms built upon our datasets would likely fail to properly parse sentences with offensive language.

Risks of the Research Ultimately, this data will enable end-users to automatically vocalize and parse large corpora of Hebrew text. For the most part, this will provide a beneficial contribution to the world: for the visually impaired, this technology will enable the development of more precise text-to-speech products; teachers will be able to provide children and second-language learners with accessible vocalized texts; and humanities and linguistics researchers can bolster their research with big-data analysis. However, there also is a risk of nefarious use, if an end user were to leverage these capabilities in order to produce anonymous texts or recordings containing threats to human life, liberty, or happiness.

References

Yonatan Belinkov. 2021. Probing Classifiers: Promises, Shortcomings, and Advances. *Computational Linguistics*, pages 1–13.

Howard Chen, Mengzhou Xia, and Danqi Chen. 2021. Non-parametric few-shot learning for word sense disambiguation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1774–1781, Online. Association for Computational Linguistics.

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

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

bidirectional transformers for language understanding.

Ali Elkahky, Kellie Webster, Daniel Andor, and Emily Pitler. 2018. A challenge set and methods for noun-verb ambiguity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2562–2572, Brussels, Belgium. Association for Computational Linguistics.

Allyson Ettinger. 2020. What BERT is not: Lessons from a new suite of psycholinguistic diagnostics for language models. *Transactions of the Association for Computational Linguistics*, 8:34–48.

Matt Gardner, Yoav Artzi, Victoria Basmova, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, Nitish Gupta, Hanna Hajishirzi, Gabriel Ilharco, Daniel Khashabi, Kevin Lin, Jiangming Liu, Nelson F. Liu, Phoebe Mulcaire, Qiang Ning, Sameer Singh, Noah A. Smith, Sanjay Subramanian, Reut Tsarfaty, Eric Wallace, Ally Zhang, and Ben Zhou. 2020. Evaluating models' local decision boundaries via contrast sets.

Yoav Goldberg. 2019. Assessing bert's syntactic abilities. *CoRR*, abs/1901.05287.

Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 897–907, Berlin, Germany. Association for Computational Linguistics.

Josef Klafka and Allyson Ettinger. 2020. Spying on your neighbors: Fine-grained probing of contextual embeddings for information about surrounding words. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4801–4811, Online. Association for Computational Linguistics.

Stav Klein and Reut Tsarfaty. 2020. Getting the ##life out of living: How adequate are word-pieces for modelling complex morphology? In *Proceedings of the 17th SIGMORPHON Workshop on Computational Research in Phonetics, Phonology, and Morphology*, pages 204–209, Online. Association for Computational Linguistics.

Roberto Navigli. 2009. Word sense disambiguation: A survey. *ACM Comput. Surv.*, 41.

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.

Avi Shmidman, Joshua Guedalia, Shaltiel Shmidman, Moshe Koppel, and Reut Tsarfaty. 2020. A novel challenge set for Hebrew morphological disambiguation and diacritics restoration. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3316–3326, Online. Association for Computational Linguistics.

Ian Tenney, Patrick Xia, Berlin Chen, Alex Wang, Adam Poliak, R Thomas McCoy, Najoung Kim, Benjamin Van Durme, Samuel R. Bowman, Dipanjan Das, and Ellie Pavlick. 2019. What do you learn from context? probing for sentence structure in contextualized word representations.

Reut Tsarfaty, Dan Bareket, Stav Klein, and Amit Seker. 2020. From SPMRL to NMRL: What did we learn (and unlearn) in a decade of parsing morphologically-rich languages (MRLs)? In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7396–7408, Online. Association for Computational Linguistics.

Shuly Wintner. 2014. Morphological processing of semitic languages. In *NLP of Semitic Languages*.

Yadollah Yaghoobzadeh, Katharina Kann, T. J. Hazen, Eneko Agirre, and Hinrich Schütze. 2019. Probing for semantic classes: Diagnosing the meaning content of word embeddings. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5740–5753, Florence, Italy. Association for Computational Linguistics.

Mengjie Zhao, Philipp Dufter, Yadollah Yaghoobzadeh, and Hinrich Schütze. 2020. Quantifying the contextualization of word representations with semantic class probing. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1219–1234, Online. Association for Computational Linguistics.

7 Appendix A: Table of Homographs

7.1 Homographs with Segmentation Ambiguity

Form	Word	Morphology	Translation	Sentences
האם	הַאִם	Interrogative	does	1,000
	הָ+אֵם	Det + Noun [F,S,abs]	the + mother	1,000
המראה	הַ+מַּרְאֶה	SConj + Participle [M,S]	that + indicates	1,000
	הַ+מַּרְאָה	SConj + Participle [F,S]	that + indicates	935
	הַמְרָאָה	Noun [F,S,abs]	takeoff	739
הקפה	הַ+קָּפֶה	Det + Noun [M,S,abs]	the + coffee	1,000
	הַקָּפָה	Noun [F,S,abs]	credit	750
הקשר	הַ+קֶּשֶׁר	Det + Noun [M,S,abs]	the + connection	1,000
	הַ+קַּשָּׁר	Det + Noun [M,S,abs]	the + signaler	1,000
	הֶקְשֵׁר	Noun [M,S,abs/cons]	context	633
השלמה	הַשְׁלָמָה	Noun [F,S,abs]	completion	1,000
	הַ+שְּׁלֵמָה	Det + Adj [F,S]	the + complete	915
ועד	וְ+עַד	Conj + Prep	and + until	1,000
	וַעַד	Noun [M,S,abs/cons]	committee	529
לשם	לְשֵׁם	Prep	for the purpose of	1,000
	לְ+שָׁם	Prep + Adverb	to + there	1,000
מבחינה	מִ+בְּחִינָה	Prep + Noun [F,S,abs]	from + point of view	1,000
	מַבְחִינָה	Participle [F,S,abs]	she notices	842
מסיבות	מְסִיבוֹת	Noun [F,P,abs]	parties	1,000
	מִ+סִּיבּוֹת	Prep + Noun [F,P,abs/cons]	due to + reasons	1,000
מפתח	מַפְתֵּחַ	Noun [M,S,abs]	key	1,000
	מְפַתֵּחַ	Participle [M,S,abs]	develops / developer	663
	מִ+פֶּתַח	Prep + Noun [M,S,abs/cons]	from + opening	405
שאלה	שְׁאֵלָה	Noun [F,S,abs]	question	1,000
	שָׁאֲלָה	Verb [F,S,3,Past]	she asked	1,000
	שֶׁ+אֵלֶּה	SConj + Pronoun [MF,P,3]	that + these	595
שאף	שֶׁ+אַף	SConj + CConj	for + even	1,000
	שָׁאַף	Verb [M,S,3,Past]	he aspired	672
שבה	שֶׁ+בָּהּ	Sconj + Prep [suf=F,S,3]	that + in it	1,000
	שָׁבָה	Verb [F,S,Present/Past]	she returns / she returned	1,000
שמן	שֶׁמֶן	Noun [M,S,abs/cons]	oil	1,000
	שָׁמֵן	Adj [M,S,abs]	wide	767
	שֶׁ+מִּן	SConj + Prep	that + from	464
	שְׁמָן	Noun [M,S,abs,suf=F,P,3]	their name	458
שמר	שֶׁמֶר	PropN	Shemer	1,000
	שָׁמַר	Verb [M,S,3,Past]	he guarded	1,000
	שֶׁ+מַּר	SConj + Titular [M,S]	that + Mr.	342

7.2 Homographs with Morphological Ambiguity

Form	Word	Morphology	Translation	Sentences
אהבה	אַהֲבָה	Noun [F,S,abs]	love	1,000
	אָהֲבָה	Verb [F,S,3,Past]	she loved	1,000
אוכל	אֹוכֶל	Noun [M,S,abs/cons]	food	1,000
	אוֹכֵל	Participle [M,S,abs/cons]	eats	1,000
	אוּכַל	Modal [MF,S,1,Future]	I can	1,000
אחדות	אֲחָדוֹת	Det [F,P,abs]	several	1,000
	אַחְדוּת	Noun [F,S,abs]	unity	1,000
אחיו	אֶחָיו	Noun [MF,P,abs,suf=M,S,3]	his brothers	1,000
	אָחִיו	Noun [MF,S,abs,suf=M,S,3]	his brother	1,000
אלימות	אַלִּימוֹת	Adj [F,P]	violent	1,000
	אַלִּימוּת	Noun [F,S,abs/cons]	violence	1,000
אם	אִם	Conj	if	1,000
	אֵם	Noun [F,S,abs/cons]	mother	1,000
אמצעי	אֶמְצָעֵי	Noun [M,P,cons]	centers of / methods of	1,000
	אֶמְצָעִי	Noun [M,S,abs] / Adj [M,S]	method / central	997
אמרה	אָמְרָה	Verb [F,S,3,Past]	she said	1,000
	אִמְרָה	Noun [F,S,abs]	a saying	520
אפשר	אֶפְשָׁר	Modal / Adv	possible	1,000
	אִפְשַׁר	Verb [M,S,3,Past]	he allowed	720
את	אֶת	ACC	accusative	1,000
	אַתְּ	Pronoun [F,S,2]	you	1,000
בהמשך	בְּ+הֶמְשֵׁךְ	Prep + Noun [M,S,cons]	in + continuation of	1,000
	בַּ+הֶמְשֵׁךְ	Prep [with Det] + Noun [M,S,abs]	in the + future	1,000
בחיי	בְּ+חַיֵּי	Prep + Noun [M,P,cons]	in + lives of	1,000
	בְּ+חַיַּי	Prep + Noun [M,P,abs,suf=MF,S,1]	in + my life	1,000
בעולם	בְּ+עוֹלַם	Prep + Noun [M,S,cons]	in + a world of	1,000
	בָּ+עוֹלָם	Prep [with Det] + Noun [M,S,abs]	in the + world	1,000
	בְּ+עוֹלָם	Prep + Noun [M,S,abs]	in + a world	948
בקרב	בְּ+קֶרֶב	Prep + Noun [M,S,cons]	in + midst of	1,000
	בַּ+קְרָב	Prep [with Det] + Noun [M,S,abs]	in the + battle	1,000
	בְּ+קָרַב	Prep + Verb [Bare Infinitive]	in + approaching of	847
	בְּ+קְרָב	Prep + Noun [M,S,abs]	in + a battle	314
גילו	גִּילוּ	Verb [MF,P,3,Past]	they discovered	1,000
	גִּילוֹ	Noun [M,S,abs,suf=M,S,3]	his age	865
די	דִּי	Prefix	di-	1,000
	דֵּי	Det [cons]	enough of	1,000
	דַּי	Adverb	sufficiently	1,000
הזקן	הַ+זָּקָן	Det + Noun [M,S,abs]	the + beard	1,000
	הַ+זָּקֵן	Det + Adj [M,S] / Det + Noun [M,S,abs]	the + old	911
החל	הֵחֵל	Verb [M,S,3,Past]	he began	1,000
	הָחֵל	Verb [Bare Infinitive]	starting (from)	1,000
המשנה	הַ+מִּשְׁנֶה	Det + Noun [M,S,abs]	the + deputy	1,000
	הַ+מִּשְׁנָה	Det + PropN [F,S,abs]	the + Mishna	1,000

הנחה	הַנָּחָה	Noun [F,S,abs]	placing	1,000
	הִנְחָה	Verb [M,S,3,Past]	he directed	735
	הֶנָּחָה	Noun [F,S,abs]	discount	522
הרים	הֵרִים	Verb [M,S,3,Past]	he lifted	1,000
	הָרִים	Noun [M,P,abs]	mountains	1,000
ואת	וְ+אֶת	Conj + ACC	and + accusative	1,000
	וְ+אַתְּ	Conj + Pronoun [F,S,2]	and + you	1,000
זר	זֵר	Noun [M,S,abs/cons]	bouquet	1,000
	זָר	Adj [M,S] / Noun [M,S,abs]	foreign / stranger	1,000
חברות	חֲבָרוֹת	Noun [F,P,abs]	companies	1,000
	חֶבְרוֹת	Noun [F,P,cons]	companies of	1,000
	חֲבֵרוֹת	Noun [F,P,abs/cons]	friends	676
	חֲבֵרוּת	Noun [F,S,abs/cons]	friendship	430
	חַבְרוֹת	Noun [F,P,cons]	friends of	340
חדר	חֲדַר	Noun [M,S,cons]	room of	1,000
	חֶדֶר	Noun [M,S,abs]	room	1,000
	חָדַר	Verb [M,S,3,Past]	penetrated	1,000
טוב	טוֹב	Adj [M,S]	good	1,000
	טוּב	Noun [M,S,abs/cons]	goodness	511
יהודי	יְהוּדִי	Noun [M,S,abs] / Adj [M,S]	a Jew / Jewish	1,000
	יְהוּדֵי	Noun [M,P,cons]	Jews	1,000
כיוון	כִּיּוּן	Noun [M,S,abs] / Noun [M,S,cons]	direction	994
	כִּיֵּוון	Verb [M,S,3,Past]	directed	963
	כְּיָוון	Conj	because	763
לו	לוֹ	Prep [suf=M,S,3]	to him	1,000
	לוּ	Conj	if only	1,000
לחם	לֶחֶם	Noun [M,S,abs]	bread	1,000
	לָחַם	Verb [M,S,3,Past]	he fought	1,000
לפנות	לִפְנוֹת	Prep / Verb [Infinitive]	facing / to turn	1,000
	לְפַנּוֹת	Verb [Infinitive]	to clear out	580
מדי	מִדֵּי	Det [cons]	every	1,000
	מְדַי	Adv	too much	1,000
	מַדֵּי	Noun [M,P,cons]	uniforms of	681
מהם	מֵהֶם	Pronoun [M,P,3]	from them	1,000
	מָהֵם	Interrogative	what are	623
מי	מִי	Interrogative / Pronoun [S,3]	who	1,000
	מֵי	Noun [M,P,cons]	waters of	1,000
מלך	מֶלֶךְ	Noun [M,S,abs/cons]	king	1,000
	מָלַךְ	Verb [M,S,3,Past]	he ruled	619
מעבר	מֵעֵבֶר	Prep	beyond	1,000
	מַעְבַּר	Noun [M,S,cons]	passage of	1,000
	מַעְבָּר	Noun [M,S,abs]	passage	1,000
מראה	מַרְאֶה	Participle [M,S]	he shows	1,000
	מַרְאָה	Participle [F,S]	she shows	1,000
מרכז	מֶרְכַּז	Noun [M,S,cons]	center of	1,000
	מֶרְכָּז	Noun [M,S,abs]	center	1,000
	מְרַכֵּז	Participle [M,S,abs/cons]	organizes / organizer	790

משחק	מִשְׂחָק	Noun [M,S,abs]	game	1,000
	מְשַׂחֵק	Participle [M,S,abs]	plays / player	947
נעשה	נַעֲשֶׂה	Verb [MF,P,1,Future]	we will do	1,000
	נַעֲשָׂה	Verb [M,S,3,Past]	was done	1,000
נשים	נָשִׁים	Noun [F,P,abs]	women	1,000
	נָשִׂים	Verb [MF,P,1,Future]	we will put	813
נתן	נָתַן	Verb [M,S,3,Past]	gave	1,000
	נָתָן	Propn	Nathan	701
עבר	עָבַר	Verb [M,S,3,Past]	he passed	1,000
	עָבָר	Noun [M,S,abs]	past	1,000
	עֵבֶר	Noun [M,S,abs/cons]	side	629
עד	עֵד	Noun [M,S,abs/cons]	witness	1,000
	עַד	Prep	until	1,000
עובדות	עֻבְדוֹת	Noun [F,P,abs/cons]	facts	1,000
	עוֹבְדוֹת	Participle [F,P]	they work / workers	1,000
עם	עִם	Prep	with	1,000
	עַם	Noun [M,S,abs/cons]	nation	1,000
פני	פְּנֵי	Noun [M,P,cons]	faces of	1,000
	פָּנַי	Noun [MF,P,abs,suf=MF,S,1]	my face	700
פרס	פְּרָס	Noun [M,S,abs]	award	1,000
	פֶּרֶס	Propn	Peres	1,000
	פָּרַס	Verb [M,S,3,Past]	he spread	845
	פְּרַס	Noun [M,S,cons]	award of	308
ציון	צִיּוֹן	Propn	Zion	1,000
	צִיּוּן	Noun [M,S,abs/cons]	mark	1,000
קודם	קֹודֶם	Adv	before	1,000
	קוֹדֵם	Adj [M,S]	previous	1,000
	קֻדַּם	Verb [M,S,3,Past]	was promoted	391
ראשי	רָאשֵׁי	Noun [M,P,cons]	heads	1,000
	רֹאשִׁי	Noun [M,S,abs,suf=MF,S,1]	my head	910
	רָאשִׁי	Adj [M,S,abs]	head	414
שירת	שֵׁירֵת	Verb [M,S,3,Past]	he served	1,000
	שִׁירַת	Noun [F,S,cons]	poetry of	911
שכר	שָׂכָר	Noun [M,S,abs]	salary	1,000
	שָׂכַר	Verb [M,S,3,Past]	rented	901
	שִׂכַר	Noun [M,S,cons]	salary of	798
שם	שֵׁם	Noun [M,S,abs/cons]	name	1,000
	שָׁם	Adv	there	1,000
	שָׂם	Verb [M,S,Present/Past]	he placed	1,000
תנאי	תְּנָאֵי	Noun [M,P,cons]	conditions of	1,000
	תְּנַאי	Noun [M,S,abs/cons]	condition	850

7.3 Homographs with Semantic Ambiguity

Form	Word	Morphology	Translation	Sentences
הזמר	הַ+זֶּמֶר	Det + Noun [M,S,abs]	the + music	1,000
	הַ+זַּמָּר	Det + Noun [M,S,abs]	the + musician	1,000
הסופר	הַ+סּוֹפֶר	Det + Noun [M,S,abs]	the + market	1,000
	הַ+סּוֹפֵר	Det + Noun [M,S,abs]	the + author	783
זמר	זַמָּר	Noun [M,S,abs]	musician	1,000
	זֶמֶר	Noun [M,S,abs]	song	609
חברה	חֲבֵרָה	Noun [F,S,abs]	friend	1,000
	חֶבְרָה	Noun [F,S,abs]	company	1,000
רשות	רְשׁוּת	Noun [F,S,abs/cons]	permission	1,000
	רָשׁוּת	Noun [F,S,abs/cons]	authority	1,000

8 Appendix B: Balanced vs. Unbalanced Homographs

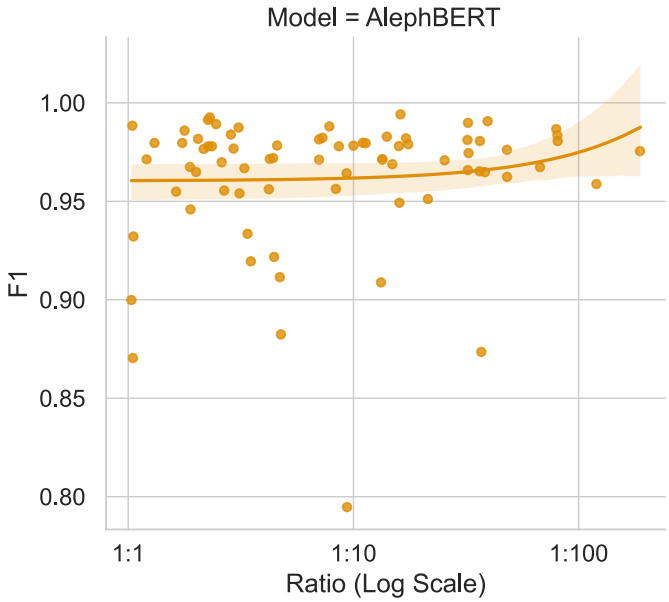

Figure 6: Comparison of AlephBERT performance on balanced vs. unbalanced homographs. An unbalanced homograph is one whose natural distribution is highly skewed, such that one analysis appears much more frequently than the other. In such a case, there is a concern that corpus-based tagging systems will be disproportionately influenced by the natural distribution and thus will be unequipped to handle the less frequent analysis. Indeed, Shmidman et al. (2020, section 3, table 2) found that the leading Hebrew morph-syntactic parser, YAP, faltered substantially on unbalanced homographs, compared with high performance on balanced homographs. We thus plot AlephBERT's performance on our 75 homograph sets against the natural distribution ratio of the homographs (Data regarding the distribution of homograph analyses is based upon an in-house annotated 2.4M word corpus maintained by DICTA.) This graph indicates that although BERT is based on a naturally-occurring Hebrew corpus, it nevertheless handles skewed homographs just as well as balanced homographs.

9 Appendix C: Computational Equipment

We performed all computations on a desktop workstation with an i9-10980XE processor and 256GB of memory. This system enabled us to run 36 experiments in parallel (the processor contains 18 hyperthreaded cores), and thus we were able to complete all of the relevant experiments and computations over the course of several weeks of calendar time.