NLP in the DH pipeline: Transfer-learning to a Chronolect

Aynat Rubinstein
The Hebrew University
of Jerusalem
aynat.rubinstein@mail.huji.ac.il

Avi Shmidman
Bar-Ilan University
Dicta: The Israel Center for Text Analysis
avi.shmidman@biu.ac.il

Abstract

A big unknown in Digital Humanities (DH) projects that seek to analyze previously untouched corpora is the question of how to adapt existing Natural Language Processing (NLP) resources to the specific nature of the target corpus. In this paper, we study the case of Emergent Modern Hebrew (EMH), an under-resourced chronolect of the Hebrew language. The resource we seek to adapt, a diacritizer, exists for both earlier and later chronolects of the language. Given a small annotated corpus of our target chronolect, we demonstrate that applying transfer-learning from either of the chronolects is preferable to training a new model from scratch. Furthermore, we consider just how much annotated data is necessary. For our task, we find that even a minimal corpus of 50K tokens provides a noticeable gain in accuracy. At the same time, we also evaluate accuracy at three additional increments, in order to quantify the gains that can be expected by investing in a larger annotated corpus.

1 Introduction

Digital Humanities (DH) projects that deal with understudied languages, including many historical languages, often face a "throughput bottleneck" due to the lack of Natural Language Processing (NLP) resources trained for the specific language variety they document. In many cases, there is a closely related language, typically a modern standard variety, for which there exist large digital corpora, goldsets manually-annotated for various features, and NLP tools (e.g., morphological analyzers, syntactic parsers). It is an open question to what extent it is useful to adapt tools tailored for one *chronolect* (a language used in a particular point in time), in order to build resources for a closely related yet different chronolect (Baron and Rayson 2008; Schneider 2020 on spelling normalization for English chronolects; Pettersson et al. 2013 for Early Modern Swedish spelling, tagging, and parsing; Pettersson and Nivre 2015 for verb phrase extraction in Early Modern Swedish; Hana et al. 2011 for morphological tagging of Old Czech).

The researcher is faced in such a case with questions such as: how much data is needed for training? Is it better to fine-tune existing NLP models, which were trained to fit a different chronolect, or instead to train new models only on the chronolect of interest? These are the questions that arose in the framework of a DH project "The Jerusalem Corpus of Emergent Modern Hebrew" (JEMH; Rubinstein 2019), which aims to digitize and enable linguistic analysis of multi-genre archival material in Emergent Modern Hebrew (EMH; Hebrew of the early 20th century). Diacritization of this corpus is a critical step both in order to increase its accessibility (in particular because the diacritics are a necessary prerequisite for the application of text-to-speech algorithms), and also for downstream NLP tasks such as intelligent search, relation extraction, and linguistic analysis of the historical materials.

Although EMH is a severely undersourced Hebrew chronolect, diacritization models do exist for other Hebrew chronolects. We thus aim to answer the following questions: how much diacritized EMH data must be assembled, and how can we best leverage existing models from other Hebrew chronolects? We manually add annotations of diacritics to a representative sample of the JEMH Corpus and use this new goldset to train a neural-net diacritizer for EMH. The resulting diacritizer is the first NLP tool developed for this historical langauge variety. It can be incorporated into the pipeline of any DH project dealing with EMH, providing valuable disambiguations of the text.

We compare two approaches to the creation of the diacritizer for EMH. The *indirect* approach uses the EMH goldset to fine-tune existing dia-

Proceedings of the Workshop on Natural Language Processing for Digital Humanities (NLP4DH), pages 106–110
co-located with ICON-2021, December 16-19, 2021, Silchar, India.
https://doi.org/10.26615/978-952-94-5833-2_012

critizers on top of two well-resourced chronolects of Hebrew which are closely-related to our target chronolect (one variety predates EMH and one is the present-day standard variety). In the *direct* route, we train a new model for EMH using just our (small) annotated goldset. We provide, for each approach, an estimate of the lower bound on the amount of training data needed to reach acceptable performance, discussing the implications of our findings for planning DH projects.

2 Linguistic challenges

The use of Hebrew as a spoken language only began around the turn of the 20th century, and the norms of Modern Hebrew - lexical, morphological, syntactic, and orthographical - only crystallized during the decades afterward (Rosén, 1956; Blanc, 1968; Ben-Hayyim, 1992; Rabin, 1999; Reshef, 2013, 2015, 2016; Doron et al., 2019b). However, knowledge about previous stages of the language was accessible to learners through canonized texts which formed the basis of Jewish education throughout history (Doron et al. 2019a). Therefore, EMH constitutes a chronolect that is considerably distant from Modern Hebrew. First of all, on the lexical plane, neologisms for many everyday objects such as "kitchen" or "newspapers" were only just being invented for the first time, and were often referred to by multiword circumlocutions instead (e.g., for kitchen: בֵּית הַבְּשׁוּל *bet habišul*, lit. 'house of cooking', rather than the modern word מִטְבָּח *mitbaħ*, or, for newspaper, מִכְתָּב עִתִּי *mixtav ʕiti*, lit. 'periodical letter', rather than עִתּוֹן *ʕiton*). On the morphological plane, we find much use of nominal patterns now completely obsolete (e.g., מְפֹאָרָה *mefoʔarah* 'magificent', rather than מְפֹאֶרֶת *mefoʔeret*, or מַתְחֶלֶת *matħelet* 'start (participle, fem.)' rather than מַתְחִילָה *matħilah*). Finally, and most significantly, plene orthography - now normative in Modern Hebrew - had not yet been embraced, causing ambiguities to abound. As has been noted, Modern Hebrew by itself is already highly ambiguous morphologically (Wintner, 1998; Tsarfaty et al., 2019); however, without the norms of plene spelling, the ambiguity is amplified considerably. For example, in EMH חדשים *ħdšim* may be analyzed as one of three words: חֲדָשִׁים *ħadašim* 'new (pl.)', חֳדָשִׁים *ħodašim* 'months', or חָדְשַׁיִם *ħodšayim* 'two months', whereas in Modern Hebrew, each is represented by a distinct unambiguous spelling, via the addition of matres lectionis: חדשים *ħdšim*, חודשים *ħodšim*, or חודשיים *ħodšiim*.

Resolving these challenges is crucial for natural lanuage processing of the formative and highly influential EMH corpus of Hebrew. Adding diacritics can dramatically reduce ambiguity and is the tool we sought to develop.

3 Diacritizaton

Modern Hebrew (similar to other Semitic languages such as Arabic and Syriac) is generally written and published with many of the vowels omitted. In EMH, even fewer vocalizations are marked. *Diacritization* refers to the specification of all vowels as part of the written word. The set of diacritics generally used in Hebrew is termed "Tiberian diacritization", and consists of a set of a dozen essential marks placed below, within, or above the characters (Golinets 2013). Letters can be optionally geminated (marked with a dot in the center of the letter), leading to a total of 24 possible diacritic permutations for each letter.

A given non-diacritized Hebrew word generally admits to multiple possible diacritizations, each representing a different semantic and morphological analysis of the word. Thus, diacritization cannot be automated via a simple lookup table; rather, it is necessary to use contextual information to choose from among the multiple analyses. Several machine-learning systems have been developed to perform this task (Choueka and Neeman, 1995; Gal, 2002; Gershuni and Pinter, 2021). The current state-of-the-art for Modern Hebrew is the LSTM-based diacritizer developed by Dicta (Shmidman et al., 2020), as per external evaluations performed by Gershuni and Pinter (2021).

However, as noted, the EMH chronolect presents a new set of challenges, and, indeed, automated diacritization systems for Modern Hebrew falter on EMH. We therefore set out to determine how large a corpus would be necessary to train the same kind of LSTM specifically for EMH, and to determine whether it would be beneficial, alternatively, to attempt a transfer-learning architecture based upon the pre-trained Modern Hebrew LSTM model.

4 Experiments

Data Our data consists of a selection of texts from the JEMH Corpus (Rubinstein, 2019) to which diacritics have been added manually by experts. For manual annotation, we used the *Nakdan*

	Words	Years
Literature	129,453	1858-1932
Ephemera	14,993	1862-1941
Total	144,446	

Table 1: Goldset of EMH with manually-annotated diacritics

Pro interface by Dicta.[1] Most of the corpus represents a literary genre,[2] complemented by ephemera from the JEMH Street Ads supcorpus. Table 1 provides information about the size of the annotated corpus and the period it spans. We release this vocalized corpus to the public domain.[3]

Implementation and Results We divided our corpus of diacritized EMH literature into 120K words for a training set and 11K words for the test set. In order to track the effect of the size of the corpus, we train four separate LSTM models, using four subsets of the training corpus, of sizes 50K, 75K, 100K, and 120K. For each subset, we train two models: (1) We train the LSTM from scratch, using only the vocalized data in the training subset. We train for 100 epochs. (2) We adopt a transfer-learning approach: we start with Dicta's Modern Hebrew LSTM model, which was trained on over 2 million words of Modern Hebrew text. We then fine-tune this model for 100 additional epochs, using only the data in the EMH training subset. Regardless of the subset used for training, we always use the same test set, to ensure consistency.

In evaluating the accuracy of the resulting model, we separately consider two approaches: (1) We test the ability of the LSTM model to predict the correct vocalization without any constraints. For each word, we run a beam search across the top eight vocalization predictions for each letter, and we take the top scoring beam. (2) We test the ability of the LSTM model to choose the correct vocalization option from a set of options provided by a wordlist.[4] Thus, as opposed to the previous approach, here we constrain the LSTM to known valid vocalization options. For each word, we calculate the LSTM score for each of the vocalization options in the wordlist, and we take the top scoring option.[5] Results are shown in Table 2.[6]

Training Size	New Train		Fine-tune	
	LSTM	LSTM +Wordlist	LSTM	LSTM +Wordlist
0	—	—	78.30%	**84.57%**
50,000	70.47%	81.92%	78.60%	**85.86%**
75,000	73.39%	83.38%	80.20%	**86.80%**
100,000	76.14%	84.44%	80.92%	**87.06%**
120,000	77.29%	85.15%	81.98%	**87.38%**

Table 2: We test how much training data is necessary to train an LSTM model to diacritize the EMH chronolect. First (col 2-3) we show the effects of training a new model from scratch based on the specified number of tokens, with and without the use of a wordlist restricting choices to known valid forms. Next (col 4-5) we show the superior effects of transfer-learning from an existing robust model for Modern Hebrew. The initial row shows the performance of the existing model on the EMH test corpus. We then show the improvement gained by fine-tuning this model with increasing sizes of EMH texts.

Training Size	New Train		Fine-tune	
	LSTM	LSTM +Wordlist	LSTM	LSTM +Wordlist
0	—	—	80.66%	90.19%
120,000	64.98%	83.39%	74.01%	88.09%

Table 3: We evaluate an out-of-domain text (ephemera) within the target chronolect. Retraining shows no benefit at all, even with the entire 120K word corpus. Whether we retrain from scratch or fine-tune on top of the existing Modern Hebrew model, we find that the training from the EMH literary corpus only reduces the accuracy. With ephemera, it is preferable to stick with the existing Modern Hebrew model.

Next, we test the effect of the training corpus on an out-of-domain corpus within the chronolect. Whereas the previous experiment involved training and test corpora both drawn from literary EMH, here we keep the same literary training corpus, but we use a different genre of EMH (ephemera) as the test corpus. Results are shown in Table 3.

Finally, we examine the effects of doing the transfer-learning from an *earlier* chronolect, rather than from a later chronolect. In the previous experiments, we took a pre-trained model from Modern Hebrew, and we fine-tuning it for the EMH

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

[2] Obtained from the *Ben-Yehuda Project* (https://benyehuda.org/) snapshot in 2014.

[3] https://github.com/JEMHcorpus/corpora/tree/master/diacritized.

[4] We use a high-coverage Hebrew lexicon curated in-house at Dicta. For details regarding the lexicon, see (Shmidman et al., 2020), 199. We further augmented the wordlist for this project by adding support for morphological expansions typical of EMH, such as those described in section 2.

[5] If the word is not in our wordlist at all, then we default to the top LSTM beam-search prediction, as in the first approach; within our EMH corpus, this situation occurs regarding a small minority of cases, approximately 1.5% of the corpus.

[6] Percentages displayed here (and in all other tables as well) reflect word-level accuracy. For a given word in the text, we consider the prediction correct if and only if all the diacritic marks on the word are correct, including proper gemination and selection of the 'shin' dot, and including removal of all matres lectionis. Note that punctuation and non-Hebrew words within the text are not included in this calculation (because their inclusion would artificially inflate the score).

Training Size	Fine-tune over Rabbinic Hebrew		Fine-tune over Modern Hebrew	
	LSTM	LSTM +Wordlist	LSTM	LSTM +Wordlist
0	70.80%	80.03%	78.30%	**84.57%**
50,000	76.26%	83.89%	78.60%	**85.86%**
75,000	77.23%	84.04%	80.20%	**86.80%**
100,000	78.69%	84.91%	80.92%	**87.06%**
120,000	79.34%	85.39%	81.98%	**87.38%**

Table 4: We introduce the question of how transfer-learning from an earlier choronolect would differ from the case of a later choronolect. In columns 2 and 3, we evaluate the effect of fine-tuning on top of Rabbinic Hebrew, a Hebrew chronolect of the middle ages and the subsequent few centuries. In columns 4 and 5, we show the comparison to the results from Table 2 of fine-tuning on top of Modern Hebrew. As we can see here, from the outset, the pre-trained model of Modern Hebrew performs substantially better on the EMH corpus than the pre-trained model of Rabbinic Hebrew. Nevertheless, fine-tuning on top of the earlier Rabbinic Hebrew corpus does provide substantial results. Enlarging the training corpus progressively closes the gap between the Rabbinic and Modern models.

chronolect; essentially, we were testing what it would take to retroject the Modern Hebrew model to the Hebrew of some 100 years prior. However, the Dicta Nakdan also contains a robust model for Rabbinic Hebrew, a dialect of Hebrew found in abundance in Jewish legal texts from the middle ages, and the first few centuries afterward. Fine-tuning this model would be testing a transfer of the opposite direction: we evaluate what it would take to adapt the medieval and post-medieval Rabbinic model several centuries forward. We fine-tune the Rabbinic model with the same four subsets with which we fine-tuned the modern model, and we compare the results in Table 4.

5 Discussion

The results of these experiments lead us to the following observations:

First of all, in all cases, transfer-learning from a different Hebrew chronolect is better than training from scratch. This is true whether the source chronolect is later or earlier. At the same time, fine-tuning on top of a later chronolect was clearly superior to fine-tuning on top of an earlier chronolect, likely because a later chronolect will still have remnants of the earlier chronolect, while the same cannot be said of an earlier choronolect.

Additionally, in all cases, the wordlist filter improved scores considerably. Without the wordlist filter, the LSTM is free to choose any diacritization sequence at all, and in many cases ends up predicting a sequence that is not a valid word. Although many diacritization patterns are limited to certain letter configurations, and we expect the LSTM to learn these patterns, in other cases the choice of one pattern or another is fairly arbitrary, and just indicates natural development stages of the language. Thus, the wordlist provides the necessary knowledge to ensure that the predicted diacritics are indeed a pattern known to apply to the given word.

Regarding the question of how much training data is necessary: Of course, as expected, the larger the training corpus, the more accurate the result. However, the tables indicate just how much of an improvement we can expect with every additional 20,000-25,000 tokens. As we see, there is generally an increase of somewhere between half a percent and one and a half percent from stage to stage. Presumably, if we were to continue to enlarge our training corpus, the accuracy would continue to rise, until it reached a point of diminishing returns.

Importantly, although the full 120K training corpus yields best results, even a 50K corpus succeeds in providing a palpable improvement when fine-tuning on top of other chronolects. This suggests that investing in a small annotated corpus is a viable route for DH projects that seek to adapt NLP tools from neighboring chronolects.

Finally, the successful effects of the transfer-learning are limited to texts whose genre is represented within the training corpus. In contrast, when we tested our fine-tuned models on the out-of-domain ephemera corpus, we found that the transfer-learning actually lowered the accuracy of the model. This result may be related to the particular languages we tested, however, showing that the language of the EMH ephemera is closer to the present-day language than other EMH materials.

6 Conclusion

This paper examines a central concern in DH: the question of how much data is needed to train viable NLP tools for under-resourced chronolects. We addressed these questions for the specific case of a historical dialect of Hebrew, showing an advantage for fine-tuning a diacritization model using a small annotated corpus (i.e., transfer-learning) over direct training of a new model for the chronolect of interest. Our findings have implications for any project that aims to adapt an NLP algorithm to an alternate chronolect. Given access to existing NLP tools, it is likely that producing even a small annotated cor-

pus of historical materials from the relevant time period will result in substantial gains.

Acknowledgments

We acknowledge the substantial help of our programmer Cheyn Shmuel Shmidman, and of our linguistic experts, Binyamin Ehrlich and Tsion Eliash, who were responsible for the creation of the diacritized goldset of EMH texts.

This research was supported by the Israel Science Foundation (grant no. 2299/19) and by Dicta: The Israel Center for Text Analysis, headed by Prof. Moshe Koppel.

References

Alistair Baron and Paul Rayson. 2008. VARD2: a tool for dealing with spelling variation in historical corpora. Postgraduate Conference in Corpus Linguistics ; Conference date: 22-05-2008.

Ze'ev Ben-Hayyim. 1992. *The struggle for a language*. The Academy of the Hebrew Language, Jerusalem. In Hebrew.

Haim Blanc. 1968. The Israeli *koine* as an emergent national standard. In Joshua A. Fishman, editor, *Language problems of developing nations*, pages 237–251. Wiley, New York.

Yaacov Choueka and Yoni Neeman. 1995. Nakdan-text,(an in-context text-vocalizer for modern hebrew). In *BISFAI-95, The Fifth Bar Ilan Symposium for Artificial Intelligence*.

Edit Doron, Malka Rappaport Hovav, Yael Reshef, and Moshe Taube. 2019a. Introduction. In Edit Doron, Malka Rappaport Hovav, Yael Reshef, and Moshe Taube, editors, *Language Contact, Continuity and Change in the Genesis of Modern Hebrew*, pages 1–31. John Benjamins, Amsterdam.

Edit Doron, Malka Rappaport Hovav, Yael Reshef, and Moshe Taube, editors. 2019b. *Language Contact, Continuity and Change in the Genesis of Modern Hebrew*. John Benjamins, Amsterdam.

Ya'akov Gal. 2002. An hmm approach to vowel restoration in arabic and hebrew. In *Proceedings of the ACL-02 workshop on Computational approaches to semitic languages*, pages 1–7. Association for Computational Linguistics.

Elazar Gershuni and Yuval Pinter. 2021. Restoring hebrew diacritics without a dictionary.

Viktor Golinets. 2013. Masora, Tiberian. In Geoffrey Khan, editor, *Encyclopedia of Hebrew Language and Linguistics*. Brill, Leiden.

Jirka Hana, Anna Feldman, and Katsiaryna Aharodnik. 2011. A low-budget tagger for old Czech. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 10–18, Portland, OR, USA. Association for Computational Linguistics.

Eva Pettersson, Beáta B. Megyesi, and Jörg Tiedemann. 2013. An smt approach to automatic annotation of historical text. In *Proceedings of the Workshop on Computational Historical Linguistics at NODALIDA*, volume 87 of *NEALT Proceedings Series 18*, pages 54–69, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Eva Pettersson and Joakim Nivre. 2015. Improving verb phrase extraction from historical text by use of verb valency frames. In *Proceedings of the 20th Nordic Conference of Computational Linguistics (NODALIDA 2015)*, pages 153–161, Vilnius, Lithuania. Linköping University Electronic Press, Sweden.

Chaim Rabin. 1999. What was the revival of the Hebrew language? *Linguistic Studies*, pages 359–376. In Hebrew.

Yael Reshef. 2013. Revival of Hebrew: Grammatical structure and lexicon. In Geoffrey Khan, editor, *Encyclopedia of Hebrew Language and Linguistics*, volume 3, pages 397–405. Brill, Leiden.

Yael Reshef. 2015. *Hebrew in the Mandate period*. The Academy of the Hebrew Language, Jerusalem. In Hebrew.

Yael Reshef. 2016. Written Hebrew of the revival generation as a distinct phase in the evolution of Modern Hebrew. *Journal of Semitic Studies*, 61(1):187–213.

Haiim Rosén. 1956. *Ha-'ivrit šelanu*. Am Oved, Tel Aviv. In Hebrew.

Aynat Rubinstein. 2019. Historical corpora meet the digital humanities: the Jerusalem Corpus of Emergent Modern Hebrew. *Language Resources and Evaluation*, 53(4):807–835.

Gerold Schneider. 2020. Spelling normalisation of Late Modern English. In Merja Kytö and Erik Smitterberg, editors, *Late Modern English: Novel encounters*, pages 244—-268. John Benjamins, Amsterdam.

Avi Shmidman, Shaltiel Shmidman, Moshe Koppel, and Yoav Goldberg. 2020. Nakdan: Professional Hebrew diacritizer. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 197–203, Online. Association for Computational Linguistics.

Reut Tsarfaty, Amit Seker, Shoval Sadde, and Stav Klein. 2019. What's wrong with Hebrew NLP? and how to make it right. *CoRR*, abs/1908.05453.

Shuly Wintner. 1998. Towards a linguistically motivated computational grammar for Hebrew. In *Computational Approaches to Semitic Languages*.