Linguistic Knowledge within Handwritten Text Recognition Models: A Real-World Case Study[⋆]

Samuel Londner[1⋆⋆], Yoav Phillips[2], Hadar Miller[3],
Nachum Dershowitz[4], Tsvi Kuflik[3], and Moshe Lavee[2]

[1] School of Engineering, Tel Aviv University, Israel
[2] Department of Jewish History and Thought, University of Haifa, Israel
[3] Department of Information Systems, University of Haifa, Israel
[4] School of Computer Science, Tel Aviv University, Israel

Abstract. State-of-the-art handwritten text recognition models make frequent use of deep neural networks, with recurrent and connectionist temporal classification layers, which perform recognition over sequences of characters. This architecture may lead to the model learning statistical linguistic features of the training corpus, over and above graphic features. This in turn could lead to degraded performance if the evaluation dataset language differs from the training corpus language.
We present a fundamental study aiming to understand the inner workings of OCR models and further our understanding of the use of RNNs as decoders. We examine a real-world example of two graphically similar medieval documents but in different languages: rabbinical Hebrew and Judeo-Arabic. We analyze, computationally and linguistically, the cross-language performance of the models over these documents, so as to gain some insight into the implicit language knowledge the models may have acquired. We find that the implicit language model impacts the final word error by around 10%. A combined qualitative and quantitative analysis allow us to isolate manifest linguistic hallucinations. However, we show that leveraging a pretrained (Hebrew, in our case) model allows one to boost the OCR accuracy for a resource-scarce language (such as Judeo-Arabic).
All our data, code, and models are openly available at `https://github.com/anutkk/ilmja`.

Keywords: Optical character recognition · Handwritten text recognition · Transfer learning · Language model · Hebrew manuscripts

1 Introduction

Modern optical character recognition (OCR) algorithms have come a long way in their ability to accurately recognize handwritten text. However, it remains an

[⋆] Supported in part by the Deutsch Foundation, the Israeli Ministry of Science and Technology (grant number 3-17516) and a grant from Tel Aviv University Center for AI and Data Science (TAD) in collaboration with Google, as part of the initiative “AI and DS for Social Good”.

[⋆⋆] Corresponding author: `samuell@mail.tau.ac.il` open question whether these algorithms are able to capture linguistic features of the text in addition to graphical features. These algorithms use neural networks, specifically ending with recurrent layers and connectionist temporal classification (CTC) layers [16,27]. This architecture may lead to the model learning statistical linguistic features of the training corpus, over and above graphic features. And this would lead to sensitivity of the model towards the document language and to degraded performance if the evaluation dataset language differs from the training corpus language.

This paper investigates this question by examining the performance of OCR algorithms on two manuscripts written by the same scribe, one in medieval Judeo-Arabic and the other in medieval rabbinic Hebrew, but both in the same Hebrew script. Manuscripts in Hebrew script demonstrate high variability due to the wide dispersion of Jewish communities across different geo-cultural milieus. The use of manuscripts written by the selfsame person allows us to control for graphical features and focus on the rôle of linguistic knowledge in OCR performance. Our hypothesis is that OCR algorithms that are able to capture linguistic features will show higher accuracy in recognizing the handwritten text. By analyzing this real-world experimental design, we aim to shed light on the extent to which linguistic knowledge is incorporated in modern OCR algorithms.

If our hypothesis is supported, it would have important practical implications for the development and deployment of OCR algorithms. Specifically, it would suggest that it may not be possible to use a single OCR model for multiple languages with comparable accuracies, but rather a separate model for each language would be required. Accordingly, building multilingual OCR systems and making them more cost-effective so as to support a wider range of languages requires additional research and engineering.

One potential application of this idea is to use a model trained on a relatively data-rich language as a starting point for recognizing other, poorer languages resource-wise. For example, a model trained on Hebrew could be fine-tuned on Judeo-Arabic, a related Semitic language with relatively little available data. This approach would allow us to leverage the larger amount of available training data for Hebrew to improve OCR performance for Judeo-Arabic. The Ktiv database of the National Library of Israel[5] lists 61,096 known, extant manuscripts and fragments in Judeo-Arabic.

Overall, the results of this study have the potential to inform the design and implementation of OCR systems for multiple languages, with implications for a range of applications including historical document preservation, digital humanities, and language learning.

2 Related Work

2.1 Handwritten Text Recognition

We use off-the-shelf methods for automatic page segmentation, layout analysis, and line segmentation. Machine-learning based systems have seen wide use re-

[5] `https://www.nli.org.il/en/discover/manuscripts/hebrew-manuscripts`

cently for these tasks [2,6,9,10,12,17,31,35,45], the majority using combinations of CNNs and LSTMs. Traditional computer-vision methods have advantages for some types of manuscripts [33,35]. State-of-the-art methods have been implemented in kraken [22] and eScriptorium [23] for mixed models in various scripts, including Hebrew, and for a wide range of manuscript types.

The current best transcription results for such manuscripts are achieved by combinations of CNNs and BLSTMs [11,19,22]. OCR efforts working with medieval Hebrew manuscripts include [25,23,24]. The Sofer Mahir project (`https://sofermahir.hypotheses.org`) applied kraken's OCR to 20 large manuscripts of early rabbinic compositions. In the Tikkoun Sofrim project [24,44], crowdsourcing and machine learning have been used to correct the errors of the automatic transcriptions of several large manuscripts of medieval exegetical literature. Character error rates (CER) of 2–3% were attained usually for manuscripts with homogeneous layout and script but only around 9% when there were complications. Modern end-to-end systems (segmentation plus OCR) include [5,20].

2.2 Implicit Linguistic Knowledge in OCR Models

Previous works have employed synthetic data to show OCR models' sensitivity to language, and thus that they implicitly learn linguistic features. The authors of [43] test the performance of an LSTM-based OCR trained on one language and tested on other languages. The difference in performance is indicative of the model's reliance on an implicit language model (LM). However, no explanation or linguistic analysis is provided. Moreover, no attention is paid to the fact that the languages being compared, English and French, share linguistic features and even complete lexemes to a significant degree. The authors of [32] established and characterized the strength of the implicit LM in LSTM-based OCR systems by synthesizing printed English text and shuffling the characters in each sentence. This approach, although proving the existence of the implicit internal language model, is not applicable to evaluate the cross-lingual generalization capability (or lack thereof) of pretrained language models. Furthermore, in the experiment described in [32], shuffling characters does not affect the distribution of characters, thus leaving some linguistic hints to the (hypothesized) LM.

In this work, we combine an in-depth linguistic qualitative analysis and a quantitative approach to examine the degree of reliance of OCR models on linguistic features. We account for similarities and differences between the languages and present specific examples of seemingly linguistic "hallucinations" in OCR models. To the best of our knowledge, this is the first time such a hybrid approach is applied in the field of digital humanities with the goal to isolate and quantify the influence of language on the OCR model's performance. Taking into account the fact that synthesizing data is less relevant for historical manuscripts, we leverage a real-world case of two manuscripts, in two different languages, which share the same graphical features, having been written by the same scribe.

2.3 Transfer Learning

Manuscript handwriting styles are highly dependent on time, place, training, and individual predilections. Improving over state-of-the-art models by leveraging transfer learning is an obvious choice. Models pretrained over a large corpus are fine-tuned on the first few annotated pages of a manuscript in order to help decipher the rest of the manuscript. In this way, the representation learned over a *source* dataset can be refined to solve the *target* task, namely transcribing documents of a smaller, disjoint dataset [14]. Recent research [1,18] shows that the optimal method to improve accuracy is to fine-tune the parameters of the whole recognition model, while the first layer can be frozen without any meaningful performance degradation. In [15], the authors successfully apply this concept for Latin-alphabet handwriting to historical handwritten Italian titles of plays. The technique also allows one to transfer the representation from Arabic printed text to genuine handwriting [29]. Transductive methods, using purely synthetic data with data rendering and augmentation, along with domain adaptation, cycle-consistent adversarial networks, and a combination of a domain-adversarial neural network approach with a convolutional recurrent neural network architecture, have been used to advantage in [20] for Tibetan Buddhist historical texts in a variety of scripts.

3 Linguistic Background

Judeo-Arabic is a general term describing an Arabic-based Jewish language or ethnolect, with a wide variety of regional dialects, which gradually developed in Jewish communities across Arabic-speaking Islamic regions, from the 8th century until the mid-20th century. Although these dialects were influenced by local variants of Arabic, they had their own distinct characteristics that distinguished them as a unique communal dialect. On the other hand, most Judeo-Arabic dialects shared common features forming a distinctive Jewish ethnolect. The most common distinctive feature is the Hebrew orthography that was common to all Judeo-Arabic dialects (apart of some Karaite writings that used Arabic characters). The implementation of Hebrew orthography was mostly phonetic; therefore, it may have differed from one Jewish community to another due to different local pronunciation tendencies. Another common feature was the grammatical and syntactical integration of Hebrew roots, words, and phrases into the Arabic. Most manuscripts written in the Middle Ages, roughly between the 10th and the 13th centuries, as is the case for the manuscripts with which we will be working, were written in a relatively high register defined as Classical Judeo-Arabic (CJA). Simply put, this means that the core Arabic elements of the text are similar to its literary Arabic counterpart, while the differences between the various dialects within CJA are relatively mild [21,36].

For the purpose of this investigation, words were classified into four different linguistic categories:

1,2. The two basic groups are Hebrew and Judeo-Arabic. Under Hebrew we included the odd Aramaic words that are frequent in Hebrew medieval works and hence are assumed to be part of the linguistic knowledge of a model trained on Hebrew manuscripts. Each word in the manuscript is classified either as Hebrew (including Aramaic) or Judeo-Arabic.

3. A third category comprises homographs (distinct words that are written in the same manner): Since our manuscripts, like most Hebrew and Judeo-Arabic texts, lack vowels (the Hebrew and Arabic alphabets are partial abjads), many of them can be read both as a Hebrew word and as an Arabic word with divergent meanings.
4. The fourth group classified consists of abbreviated words. Our manuscripts, like most Hebrew and Judeo-Arabic texts, have a tendency to abbreviate words by dropping one or more letters at the end and adding an apostrophe or dot on top of the last letter of the shortened word, as in גו' (*gō*') for גומר (*gōmer*). Shortening, which is not common in Arabic texts, is also applied to Arabic words in Judeo-Arabic texts, as for instance, ק' (*q*') for קאל (*qāla*). Thus, we have a Hebrew textual convention applied to both Hebrew and Arabic words. For completeness of the comparison of the model's performance between both languages, we group these potentially ambiguous strings in a separate category.

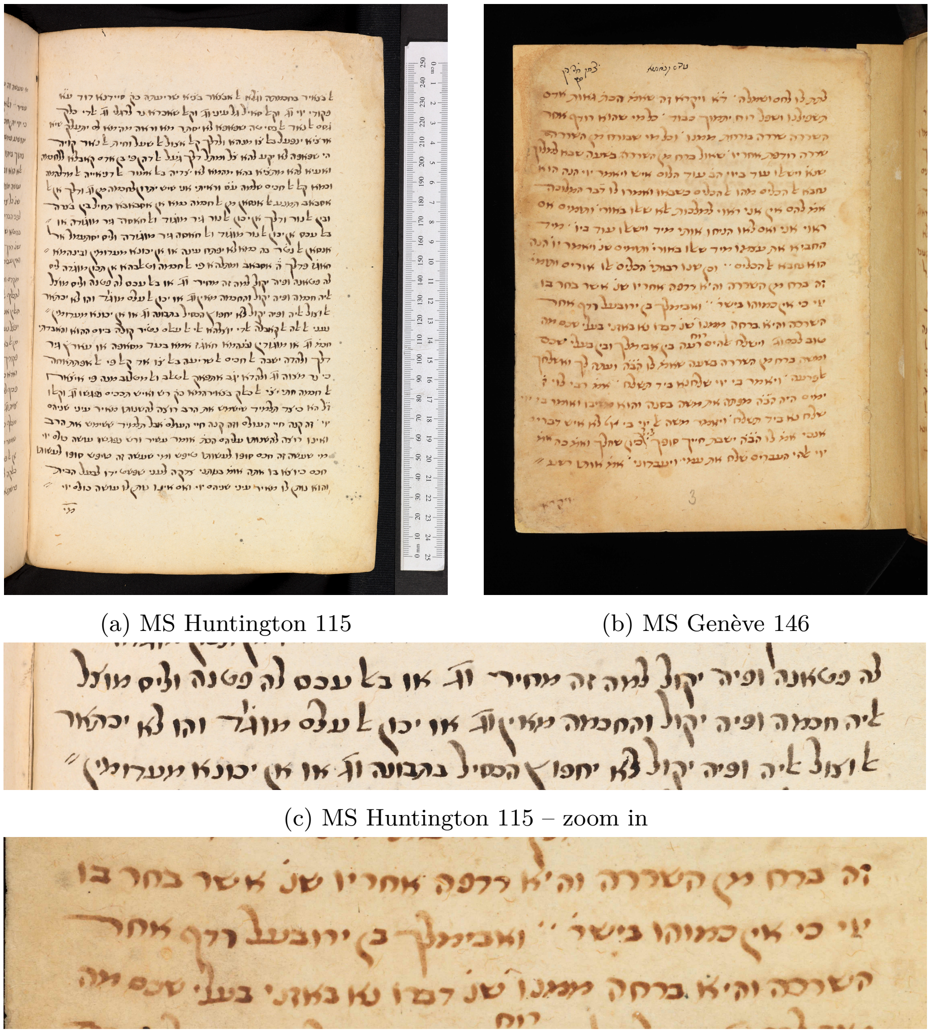
As in many Judeo-Arabic manuscripts, our scribe tended to separate the Arabic definitive article from the rest of the word. The abundant use of the definitive article in Arabic with the graphical effect of this Judeo-Arabic phenomenon was analysed separately. It should be noted that the definitive article [Arabic ال (*ʾal*), which in our manuscript may be signified by the ﭏ (*ʾal*) ligature], stripped of its context, was usually classified as a homograph since it can be read as Hebrew or Arabic, although the adjacent word to which it refers was not necessarily Arabic. In a case like ﭏ חכמים (*ʾal ḥaḫamīm*), the definitive form may be classified as a homograph and the noun חכמים (*ḥaḫamīm*) as Hebrew [7].

4 Data

For our experiments, we use the manuscripts, MS Genève Comites Latentes 146 [3] and Oxford Bodleian Library MS Huntington 115 [30]. See Fig. 1. MS Genève 146 contains a rabbinic homiletic work from late antiquity, *Midrash Tanḥuma*. MS Huntingtion 115 contains *Kitab al-Tuffāḥa*, an unpublished Judeo-Arabic homiletic work by Shamariah Hacohen (d. between 1124–1137) [26,28,13]. The majority of MS Huntington 115 (from p. 103r on) was copied by the same scribe who copied MS Genève 146, in an Oriental Hebrew Script of the 14th century.

The main evaluation set is composed of 5 pages from MS Huntington 115. It amounts to 1559 words, or 5818 characters. The manuscript was first transcribed using a base OCR system. The transcription was then manually corrected by two experts of the language and the relevant literature. The resulting ground-truth text is not corrected, that is, it includes "typos" that actually appear in the data. Labeling was performed using eScriptorium [23].

A character *k*-gram, also known as a "*k*-mer", is a sequence of *k* consecutive letters of the alphabet or other characters (spaces and punctuation). As de-



(a) MS Huntington 115

(b) MS Genève 146

(c) MS Huntington 115 – zoom in

(d) MS Genève 146 – zoom in

Fig. 1: Sample pages of the manuscripts used.

tailed below, for advanced analysis, we compare *k*-mer distributions of our texts with the distributions within two larger literary bases: one for rabbinic Tanḥumic Hebrew (which parallels the language of MS Genève 146) and another for Judeo-Arabic (parallel to MS Huntington 115). The Tanḥumic Hebrew corpus is a subset of Sefaria's dataset [34],[6] and the Judeo-Arabic corpus is from the Friedberg Judeo-Arabic Corpus [42].[7]

5 Methodology

5.1 Training

We fine-tune a pretrained model over Hebrew and test it over Judeo-Arabic and conversely. Transfer learning is an efficient approach to attain state-of-the-art OCR performance over a specific data distribution with a limited amount of data. The pretrained model, which is composed of four convolutional layers, three LSTM layers and a CTC layer, has been trained over a heterogeneous batch of generic medieval manuscripts [41]. We fine-tune it to get optimal performance over a specific manuscript, using the Adam optimizer (constant learning rate: 0.001, momentum: 0.9). Fine-tuning is performed using the kraken package.[8]

Fine-tuning the models' parameters [41] over the first few pages of the manuscript (whose ground-truth text is known) indeed improves performance dramatically. Preliminary results show that character accuracy can be boosted by around 18% by fine-tuning the recognition models over only three labeled pages (see Fig. 2). It appears that the maximum achievable accuracy with the current architecture and limited data scope is approximately 96–98%, as evidenced by state-of-the-art results for pretrained models in larger datasets [41]. When fine-tuning a model on a manuscript that exhibits a similar graphical and linguistic distribution to the pretraining dataset, only a minimal quantity of data is necessary to optimize the model's weights for the new manuscript, which accounts for the observed "saturation" phenomenon. As such, the particular choice of the source model does not seem to impact performance, nor does adding more labeled data. We note that the same technique can be applied to segmentation models.

We use a model pretrained on a corpus of biblical and rabbinical Hebrew [41]. The same base model is used for fine-tuning over Hebrew as well as Judeo-Arabic.

n.b. The original models were taken from [41] and are available from kraken's Zenodo archive [37,38,39,40].[9]

5.2 Inference

OCR is generally composed of two steps: segmentation of the image into lines and recognition of the identified segments as text. The model is applied to images and their corresponding ground-truth segmentation, generating output through

[6] See `https://www.sefaria.org/texts`. We selected all the available texts from books that belong to the Tanḥumic Hebrew corpus: *Tanḥuma*, *Pesikta Rabbati*, *Shemot Rabbah*, *Bemidbar Rabbah*, and *Devarim Rabbah*.

[7] See `https://ja.genizah.org/Home.aspx`.

[8] `https://kraken.re/`, `https://github.com/mittagessen/kraken`.

[9] `https://zenodo.org/communities/ocr_models`.

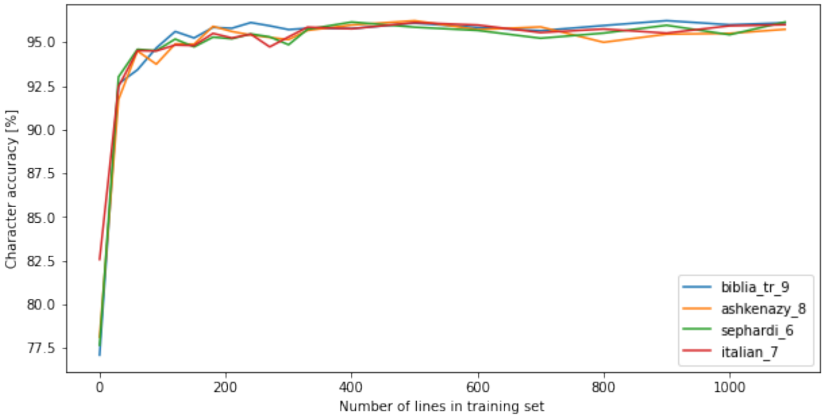


Fig. 2: Character accuracy achieved by transfer learning, as a function of additional labeled lines used for fine-tuning. Models courtesy [41].

CNN, RNN, and CTC layers. These outputs are exported to files and subsequently evaluated against ground truth, as elaborated next.

To neutralize the impact of incorrect segmentation as much as possible, we use manual ground-truth segmentation and focus only on the recognition network.

5.3 Evaluation

Some characters in Judeo-Arabic do not exist in Hebrew, mainly diacritics. We ignored these signs in the comparison, since a model trained on Hebrew material cannot generate Judeo-Arabic–specific symbols.

We compare character error rate (CER) in Table 1 and the word error rate (WER) in Table 2 over the complete evaluation sets. Although previous work [43,32] dealt only with CER, we include WER in our analysis, since we expect the hypothesised implicit language model to affect WER more significantly than CER.

We present results for four subsets of words in the Judeo-Arabic evaluation set: (*a*) all words; (*b*) Hebrew words; (*c*) homographs (Judeo-Arabic spelled like other Hebrew words); (*d*) words in Judeo-Arabic that do not exist in Hebrew. This classification was performed manually by experts. It allows us to infer the level – if any – of the linguistic features the model may have learned: character level, part-of-word level, or word level. For example, if the model learned features related to k-mer distributions, but not features related to word n-gram distributions, we would except the homograph group error rate to be similar to the Hebrew error rate. On the other hand, if the model learned language modeling features related to context, we may expect the homograph group to have a higher error rate, since the inter-word context in the evaluation set is very

Table 1: CER [percent].

Set	Hebrew model	Judeo-Arabic model
Hebrew MS	6.7	9.1
Judeo-Arabic MS – All	8.2	6.3
Judeo-Arabic MS – Hebrew	5.2	–
Judeo-Arabic MS – Hebrew homographs	5.8	–
Judeo-Arabic MS – Arabic	8.0	–

Table 2: WER [percent].

Set	Hebrew model	Judeo-Arabic model
Hebrew MS	13.9	24.6
Judeo-Arabic MS – All	17.1	14.0
Judeo-Arabic MS – Hebrew	12.7	–
Judeo-Arabic MS – Hebrew homographs	10.0	–
Judeo-Arabic MS – Arabic	21.2	–

dissimilar from the training-set context. To facilitate the manual comparison, we used Dicta's Synopsis Builder [4,8].

For reference, we include the resulting error rates of the reciprocate Judeo-Arabic model over the whole Hebrew and Judeo-Arabic datasets. Note that diacritics are ignored in the evaluation.

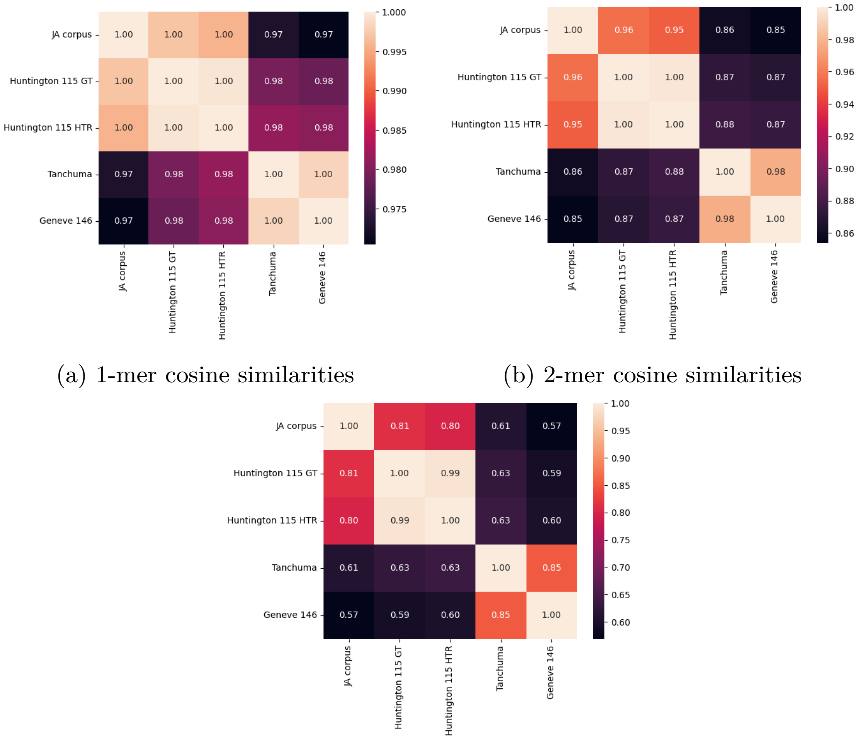
We also compare distributions of errors of the model trained over MS Genève 146 over the MS Genève 146 holdout test set and the MS Huntington 115 dataset. See Fig. 4 for confusion matrices. Moreover, to account for the different distribution of characters in the two languages, we normalize each column in the confusion matrix by the number of respective characters in the ground truth; see Fig. 5. We also report the actual error rate distribution by character in Fig. 6.

To see if the model reproduces statistical patterns from Tanḥumic Hebrew, we compare the distribution of 1,2,3-mers in the transcription in Hebrew and Judeo-Arabic. For this specific comparison, we ignore differences of ligature; specifically, ﭏ (*ʾal*) is considered identical to אל (*ʾal*). The numerical scores are cosine metrics between the (sorted) distributions. Results are detailed in Fig. 3.

6 Results and Analysis

6.1 Error Rates

The main result leading our analysis is the difference in the error rates between the Hebrew model's transcriptions over Hebrew and Arabic words (the first and last rows in Tables 1 and 2). We note that the CER difference, although existent (around 2% – consistent with previous results [43,32]), is modest compared to the WER gap of more than 8%. This gap is preserved in the overall error rates, without distinction between subsets of words (second row in Tables 1 and 2). This is a strong indication that an implicit language model exists and is sensitive



(a) 1-mer cosine similarities

(b) 2-mer cosine similarities

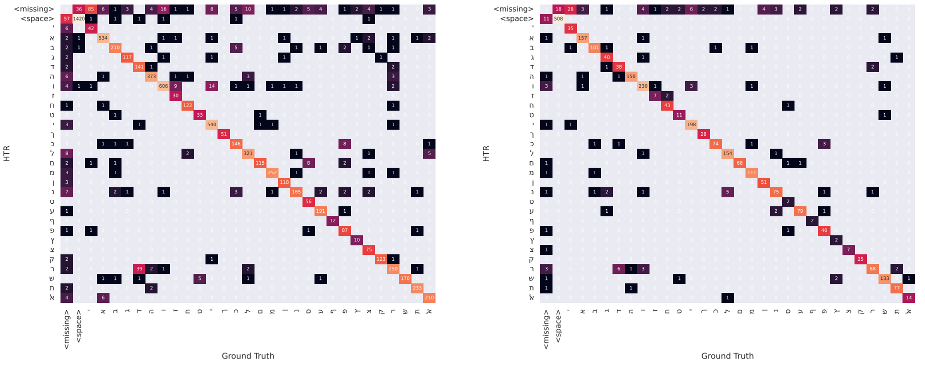
(c) 3-mer cosine similarities

Fig. 3: Comparison of character distributions. “Huntington 115 GT” denotes the distribution of ground-truth text in the Judeo-Arabic manuscript MS Huntington 115, whereas “Huntington 115 HTR” denotes the distribution of the transcription performed by the model trained on MS Genève 146.

to the specific language of the transcribed text. The fact that the error rates for the Hebrew and homograph words (third and fourth rows in Tables 1 and 2) are similar to the pure Hebrew error rate indicates that the learned linguistic features are intraword and not interword, that is, they are on the *k*-mer level.

An additional finding is the difference between the normalized error rates per character between the holdout Hebrew text and the whole Judeo-Arabic dataset (Fig. 6). The modest but significant gaps may be explained by the sensitivity of the model to language.

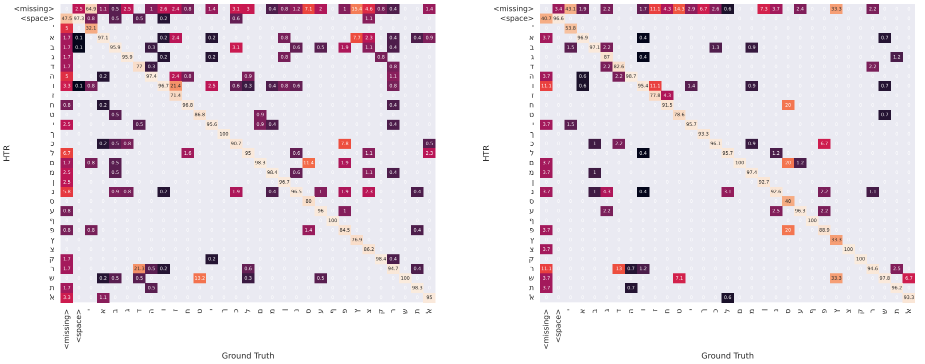
This conclusion is further reinforced by the converse finding that the Judeo-Arabic model performs much better on the Judeo-Arabic holdout test set that on the Hebrew text, by a margin of more than 10%. Incidentally, since the base pretrained model was trained on Hebrew data only, and fine-tuned on a limited amount of Judeo-Arabic data, this shows that the implicit language model can be relatively easily updated. This means that – provided the graphemes are



(a) Judeo-Arabic (Huntington 115)

(b) Hebrew (holdout MS Genève 146)

Fig. 4: Confusion matrices of the Hebrew model, evaluated over Judeo-Arabic and Hebrew.



(a) Judeo-Arabic (Huntington 115)

(b) Hebrew (holdout MS Genève 146)

Fig. 5: Normalized confusion matrices of the Hebrew model, evaluated over Judeo-Arabic and Hebrew. Units are in percent of corresponding characters in GT.

close enough – transferring the graphical knowledge and updating the language model by transfer learning may allow the leveraging of pretrained models for the benefit of data-scarce languages. Further research may analyze the influence of fine-tuning only the part of the model that is suspected to act as an implicit language model, namely the recurrent layers.

On the other hand, the *k*-mer distribution of the transcribed text (see Fig. 3) is significantly more similar to the corresponding distribution of the ground-truth Judeo-Arabic text than to Hebrew distributions (Tanḥuma or MS Genève 146).

We theorize that although the model's output mainly depends on purely graphical features, in case of ambiguous readings linguistic features "tip the

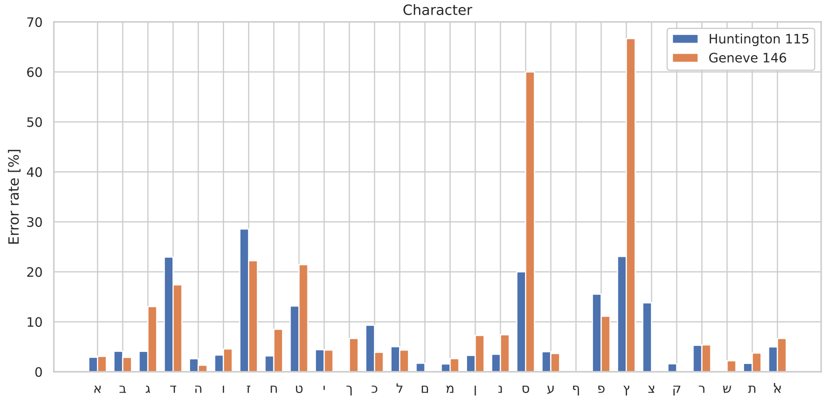
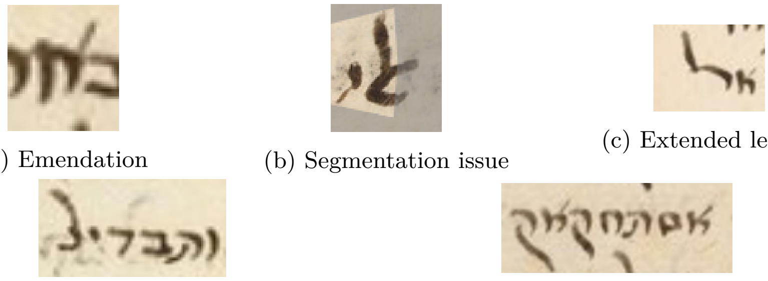


Fig. 6: Comparison of error rate per character. The outlier frequencies of ס and ץ are due to the low number of these characters in the holdout MS Genève 146 test set. Kolmogorov–Smirnov test after normalization: *statistic* $= 0.129$, $p = 0.963$.



(a) Emendation

(b) Segmentation issue

(c) Extended letters

(d) Ink bleeding from other side of the page – model read ותבדילו

(e) Final character hallucination: אסתחקאק; model read אם תחקאק

Fig. 7: Examples of erroneous readings.

scale". A potential explanation for the observed phenomenon could be the following: The RNN, positioned at the conclusion of the model, likely functions as a self-supervised conditional language model, primarily utilizing the target text during the training process. To validate this theory, we performed a semi-qualitative analysis of the identified errors.

6.2 Graphical Errors

In many cases, a graphic issue can explain mistakes: the ink bleeding from the reverse side of the page causes confusions (Fig. 7d), or the scribe made a slight emendation or wrote the letter in a manner that resembles another letter, and so forth. For example, in Fig. 7a it seems that the scribe wrote the letter ח (*ḥ*) by mistake, and then emended it to the correct letter א (ʾ) by adding the upper right

Table 3: Errors excluding graphical issues.

Language	Words	Error count	Error rate [%]
Hebrew	524	58	11.07
Arabic	499	90	18.04
Homographs	405	27	6.67
Other	54	11	20.37
Total	1482	186	–

stroke, which is hardly seen. Indeed the model read ח. Another type of mistake that is not related to acquisition of the language is due to segmentation and hence should not be counted as evidence for the question of semantic knowledge. For instance, in Fig. 7b, the segmentation missed the exact beginning of the line, and hence the ligature ﭏ (*ʾl*) was read as the last letter in it, ל (*l*).

The scribe tends to write wide letters to fill the space at the end of line, so it comes out adjusted (see Fig. 7c). The model did not "learn" this feature, frequently failing for such stretched letters. However, a human reader who knows the language has the ability to overcome graphic issues, and our assumption is that since we taught the model full lines, we should expect some knowledge about frequencies of letters, that would help the model to overcome graphic issues.

6.3 Evidence for Linguistically Triggered Errors

To validate our assumption, we excluded mistakes obviously caused by a graphical reason (i.e. segmentation, ink bleed). Indeed, the ratio between the mistakes in both language did not change. See Table 3.

Subject to the danger of the rule of small numbers, the following report suggests an analysis based on a qualitative review of the material and some quantitative related analyses. We examined all mistakes and noted the following phenomena.

The most frequent mistake is the replacement of ר (*resh*) in place of ד (*dalet*). The shapes of these two letters are very similar, and our scribe writes them in an extremely inconsistent manner. In many cases, without the semantic context, a human reader will be unable to distinguish between them. The directionality of the mistakes is very clear. Only one time ר (*r*) was read as ד (*d*), while ר (*r*) was read as ד (*d*) 35 times (19% of its appearances in the examined pages).

There are 20 mistakes in Judeo-Arabic words (out of 89 words containing the letter and 499 words in total) and 15 mistakes are in Hebrew words or homographs (out of 84 words containing the letter and 954 words in total). This clearly shows a typical cause for the larger proportions of mistakes in Judeo-Arabic. At least in 11 cases the mistake created a valid word in Hebrew, so if there is any accumulation of linguistic knowledge it could not support the model decision making in those cases. Of special importance are two cases in which the model also split the word wrongly, so that a valid Hebrew word is created: וגדנאהא (*wagadnāhā*) becomes וגר נאהא (*ve-ger nāhā*), ﭏדאר (*ʾd-dār*) becomes ﭏ ראה (*ʾel raʾah*). Note that וגר (*ve-ger*) and ראה (*raʾah*) are valid Hebrew words.

In both Rabbinic and Biblical Hebrew, the frequency of ר *resh* is double that of ד *dalet*. This explains why the model mistakes *dalet* for *resh* and not vice versa. In Judeo-Arabic the ratio changes significantly, possibly because the Hebrew letter ד (*dalet*) represents, in Arabic, both د (*dāl*) and ذ (*ḏāl* – usually written with a diacritic, ד̇).[10] In this case, which is the most glaring one, we can clearly see that the frequency of single letter is the cause of the different ratio of mistakes in the two languages.

Another frequent mistake is the reading of ס (*samekh*) as ם (final *mem*). These characters are graphically similar. This case is important because final *mem* always comes at the end of a word. Indeed, 11 out the 12 errors are ones in which the model read final *mem* as the last letter rather than the actual *samekh*. Another case of reading final *mem* mistakenly was also at the end of the word. Out of these mistakes, there are three striking cases in which the *samekh* was in a middle of the word, but the model both read it as final *mem* and split the word wrongly after the final *mem*, clearly demonstrating an inclination to represent the frequency of appearance of a space after final *mem*. For instance, in Fig. 7e, the OCR model erroneously mistook a ס for a ם, and hallucinated a space after the mistaken ם, turning אסתחקאק (*ᵓistiḥqāq*) into אם תחקאק (*ᵓim tiḥqāq*). A model trained on Judeo-Arabic would probably be familiar with the sequence אסת (*ᵓist*) which is part of the conjugation אסתפעל (استفعال = *ᵓistifᶜāl*). Additional errors of this genre include וק' סיידנא (*qaw' sayīdnā*) being mistaken for וק'ם יירנא (*waqa'm yīrnā*), and מע סעאדה (*maᶜa saᶜādah*) for מננים עאדר (*mananīm ᶜādir*).

Except once, all final *mem* mistakes are in Judeo-Arabic words. More telling is the following observation: Out of a total nine cases of *samekh* at the end of a Judeo-Arabic word, seven were wrongly read! *Samekh* appears in the middle of a word in Judeo-Arabic 36 times, and only one of them was read as final *mem* in the middle of word. In the three other cases, *samekh* in the middle of the word was read as final *mem* and followed by an imaginary space (presented above).

In Hebrew, the corpus has only 2 words ending with *samekh*, one read as final *mem*, and none of the total 24 cases of *samekh* in the middle of a Hebrew word was read as final *mem*. Indeed the frequency of final *mem* versus *samekh* at the end of the word in the Hebrew MS Genève 146 gold transcription is about 30:1. It seems obvious that the model "learned" that *samekh* hardly ever appears at the end of a word and that a final *mem* is final.

Another hint for a certain acquisition of knowledge concerns the frequency of letters and sequences is the reading of *zayin* (ז) as *vav* (ו), as shown in Table 4. Once again, these are two similar letters, though much more distinguishable to the human eye in the hand of this specific scribe. *Zayin* is very rare, whereas *vav* is very frequent, and hence the clear directionality of the mistakes (as in the case of *resh* and *dalet*). The model read *zayin* as *vav* 9 times, 5 of them

[10] As mentioned, since diacritics are not used in standard Hebrew, and do not appear in the Hebrew model's training data, we should ignore them in our analysis and error rate computation.

Table 4: *Zayin*/*vav* confusions.

	Hebrew	Judeo-Arabic
Total (mistakes/words) *zayin*	5 /25	4 /18
Beginning of a word	5 /16	1 /2
Middle of a word	0 /9	3 /16

in the beginning of a word. This is related to the frequent function of *vav* as a conjunction, which appears at the beginning of a word.

Zayin is much more frequent in Hebrew than in Judeo-Arabic, especially at the beginning of a word. As a result, this is a rare case where the OCR model's character error rate is significantly higher in Hebrew than in Judeo-Arabic.

7 Conclusion

This paper presents a fundamental study aiming to understand the inner workings of OCR models and further our understanding of the use of RNN as decoders. We find that a network concluded by a RNN, trained to recognize words in one language, suffers a bias for that language, and therefore performs less well on texts in another natural (not artificial) language with the same alphabet and distribution of letters. Specifically, our combined quantitative and qualitative analysis shows that although OCR models mainly base their output on graphical features, linguistic features play a significant rôle in the transcription process and affect the final word accuracy by around 10%. By combining a qualitative approach to the linguistic features of the transcription and a quantitative analysis of the error distributions, we were able to isolate specific cases of seemingly linguistic hallucinations. We surmise that the decoder functions as a self-supervised conditional language model, primarily utilizing the target text during the training process.

The results demonstrate the need to train specific models for languages other than Hebrew in Hebrew script. Our conclusions are probably relevant to other Jewish languages in Hebrew script, such as Yiddish and Ladino (Judeo-Español), to Aramaic, and perhaps to the different languages written in Arabic characters.

Moreover, the existence of a low-level internal language model in OCR models suggests that post-OCR correction using a character-level or k-mer language model may be less likely to be helpful than using a semantic language model.

It may be feasible to moderate the extent of learning, such as by training on multilingual datasets or randomized synthetic data, although this may result in reduced accuracy for the original target language due to the implicit language model's capacity to "pre-correct" errors. An alternative approach involving training on a data-rich language and subsequently fine-tuning all or part of the network on a closely related data-poor language may yield superior outcomes. In fact, the similarities between the languages leave the door open for fine-tuning pretrained models over less data-rich datasets, although special attention needs to be given to language-specific glyphs such as diacritics.

References

1. Aradillas, J.C., Murillo-Fuentes, J.J., Olmos, P.M.: Boosting offline handwritten text recognition in historical documents with few labeled lines. IEEE Access pp. 76674–76688 (2021), `https://ieeexplore.ieee.org/stamp/stamp.jsp?arnumber=9438636`
2. Barakat, B., Droby, A., Kassis, M., El-Sana, J.: Text line segmentation for challenging handwritten document images using fully convolutional network. In: Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 374–379. IEEE (2018)
3. Bibliothèque de Genève: Comites Latentes 146: Midrash Tanhuma (Leviticus-Numbers-Deuteronomy). `https://www.e-codices.unifr.ch/en/list/one/bge/cl0146` (2015)
4. Brill, O., Koppel, M., Shmidman, A.: FAST: Fast and accurate synoptic texts. Digital Scholarship in the Humanities **35**(2), 254–264 (2020)
5. Carbonell, M., Mas, J., Villegas, M., Fornés, A., Lladós, J.: End-to-end handwritten text detection and transcription in full pages. In: Proceedings of the International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 5, pp. 29–34. IEEE (2019)
6. Chen, K., Seuret, M., Liwicki, M., Hennebert, J., Ingold, R.: Page segmentation of historical document images with convolutional autoencoders. In: Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1011–1015. IEEE (2015)
7. Connolly, M.M.: Splitting definitives: The separation of the definite article in medieval and pre-modern written Judeo-Arabic. Journal of Jewish Languages **9**(1), 32–76 (2021)
8. Dicta: Synopsis Builder. `https://synoptic.dicta.org.il`
9. Diem, M., Kleber, F., Fiel, S., Grüning, T., Gatos, B.: cBAD: ICDAR2017 competition on baseline detection. In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1355–1360. IEEE (2017)
10. Droby, A., Kurar Barakat, B., Madi, B., Alaasam, R., El-Sana, J.: Unsupervised deep learning for handwritten page segmentation. In: Proceedings of the 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 240–245. IEEE (2020)
11. Dutta, K., Krishnan, P., Mathew, M., Jawahar, C.V.: Improving CNN-RNN hybrid networks for handwriting recognition. In: Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 80–85. IEEE (2018)
12. Fink, M., Layer, T., Mackenbrock, G., Sprinzl, M.: Baseline detection in historical documents using convolutional u-nets. In: Proceedings of the 13th IAPR International Workshop on Document Analysis Systems (DAS). pp. 37–42. IEEE (2018)
13. Gan-Zvi, M.: Parashat Pinchas in Kitáb-al-Tuffaẖa and the Early Judeo-Arabic Homiletics. Master's thesis, The University of Haifa (2018)
14. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), `http://www.deeplearningbook.org`
15. Granet, A., Morin, E., Mouchère, H., Quiniou, S., Viard-Gaudin, C.: Transfer learning for handwriting recognition on historical documents. In: Proceedings of the 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM) (2018)

16. Graves, A., Fernández, S., Gomez, F., Schmidhuber, J.: Connectionist temporal classification: Labelling unsegmented sequence data with recurrent neural networks. In: Proceedings of the 23rd International Conference on Machine Learning (ICML). pp. 369–376 (2006)
17. Grüning, T., Leifert, G., Strauß, T., Michael, J., Labahn, R.: A two-stage method for text line detection in historical documents. International Journal on Document Analysis and Recognition (IJDAR) **22**(3), 285–302 (2019)
18. Jaramillo, J.C.A., Murillo-Fuentes, J.J., Olmos, P.M.: Boosting handwriting text recognition in small databases with transfer learning. In: Proceedings of the 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 429–434. IEEE (2018)
19. Kahle, P., Colutto, S., Hackl, G., Mühlberger, G.: Transkribus–a service platform for transcription, recognition and retrieval of historical documents. In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 4, pp. 19–24. IEEE (2017)
20. Keret, S., Wolf, L., Dershowitz, N., Werner, E., Almogi, O., Wangchuk, D.: Transductive learning for reading handwritten Tibetan manuscripts. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR). pp. 214–221. IEEE (2019)
21. Khan, G.: Judeo-Arabic. In: Handbook of Jewish Languages, pp. 22–63. Brill (2016)
22. Kiessling, B.: Kraken–An universal text recognizer for the humanities. In: Digital Humanities (DH2019) (2019)
23. Kiessling, B., Tissot, R., Stokes, P., Stökl Ben Ezra, D.: eScriptorium: An open source platform for historical document analysis. In: Proceedings of the International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 19–19. IEEE (2019)
24. Kuflik, T., Lavee, M., Stökl Ben Ezra, D., Ohali, A., Raziel-Kretzmer, V., Schor, U., Wecker, A., Lolli, E., Signoret, P.: Tikkoun Sofrim combining HTR and crowdsourcing for automated transcription of Hebrew medieval manuscripts. In: Digital Humanities (DH2019) (2019)
25. Kurar Barakat, B., El-Sana, J., Rabaev, I.: The Pinkas dataset. In: Proceedings of the International Conference on Document Analysis and Recognition (ICDAR). pp. 732–737. IEEE (2019)
26. Lavee, M.: Literary canonization at work: The authority of aggadic midrash and the evolution of havdalah poetry in the Genizah. AJS Review **37**(2), 285–313 (2013)
27. Liwicki, M., Graves, A., Fernàndez, S., Bunke, H., Schmidhuber, J.: A novel approach to on-line handwriting recognition based on bidirectional long short-term memory networks. In: Proceedings of the 9th International Conference on Document Analysis and Recognition (ICDAR) (2007)
28. Nahra, R.: Kitab al-Tuffāḥa: A Collection of Judaeo-Arabic Homilies on the Torah, from the End of the 11th or the Beginning of the 12th Century. Introduction with an Edition of the Homilies on the Book of Bereshit. Ph.D. thesis, Hebrew University of Jerusalem (2016), [Hebrew]
29. Noubigh, Z., Mezghani, A., Kherallah, M.: Transfer learning to improve Arabic handwriting text recognition. In: Proceedings of the 21st International Arab Conference on Information Technology (ACIT). pp. 1–6. IEEE (2020)
30. Oxford University, Bodleian Library: MS. Huntington 115. `https://www.e-codices.unifr.ch/en/list/one/bge/cl0146` (2015)
31. Reul, C., Christ, D., Hartelt, A., Balbach, N., Wehner, M., Springmann, U., Wick, C., Grundig, C., Büttner, A., Puppe, F.: OCR4all–An open-source tool providing

a (semi-) automatic OCR workflow for historical printings. Applied Sciences **9**(22), 4853 (2019)

32. Sabir, E., Rawls, S., Natarajan, P.: Implicit language model in LSTM for OCR. In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 7, pp. 27–31. IEEE (2017)
33. Sadeh, G., Wolf, L., Hassner, T., Dershowitz, N., Stökl Ben Ezra, D.: Viral transcript alignment. In: Proceedings of the 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 711–715. IEEE (2015)
34. Sefaria, Inc.: A living library of Torah texts online (Dec 2021), `https://github.com/Sefaria/Sefaria-Export`
35. Seuret, M., Stökl Ben Ezra, D., Liwicki, M.: Robust heartbeat-based line segmentation methods for regular texts and paratextual elements. In: Proceedings of the 4th International Workshop on Historical Document Imaging and Processing. pp. 71–76 (2017)
36. Stillman, N.A.: The Judeo-Arabic heritage. In: Zion, Z. (ed.) Sephardic & Mizrahi Jewry: From the Golden Age of Spain to Modern Times, pp. 40–54 (2005)
37. Stökl Ben Ezra, D.: Medieval Hebrew manuscripts in Ashkenazi bookhand. `https://zenodo.org/record/5468478` (2021), [Online; accessed 31-Jan-22]
38. Stökl Ben Ezra, D.: Medieval Hebrew manuscripts in Italian bookhand, version 1.0. `https://zenodo.org/record/5468573` (2021), [Online; accessed 31-Jan-22]
39. Stökl Ben Ezra, D.: Medieval Hebrew manuscripts in Sephardi bookhand, version 1.0. `https://zenodo.org/record/5468665` (2021), [Online; accessed 31-Jan-22]
40. Stökl Ben Ezra, D.: Medieval Hebrew manuscripts, version 1.0. `https://zenodo.org/record/5468286` (2021), [Online; accessed 31-Jan-22]
41. Stökl Ben Ezra, D., Brown-DeVost, B., Jablonski, P., Lapin, H., Kiessling, B., Lolli, E.: BiblIA–a general model for medieval Hebrew manuscripts and an open annotated dataset. In: Proceedings of the 6th International Workshop on Historical Document Imaging and Processing (HIP). pp. 61–66 (2021)
42. The Friedberg Jewish Manuscript Society: The Friedberg Judeo-Arabic Project. `https://ja.genizah.org/` (2014), accessed: 2022-01-08
43. Ul-Hasan, A., Breuel, T.M.: Can we build language-independent OCR using LSTM networks? In: Proceedings of the 4th International Workshop on Multilingual OCR. pp. 1–5 (2013)
44. Wecker, A.J., Raziel-Kretzmer, V., Kiessling, B., Stökl Ben Ezra, D., Lavee, M., Kuflik, T., Elovits, D., Schorr, M., Schor, U., Jablonski, P.: Tikkoun Sofrim: Making ancient manuscripts digitally accessible: The case of Midrash Tanhuma. ACM Journal on Computing and Cultural Heritage (JOCCH) **15**(2), 1–20 (2022)
45. Xu, Y., He, W., Yin, F., Liu, C.L.: Page segmentation for historical handwritten documents using fully convolutional networks. In: Proceedings of the 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 541–546. IEEE (2017)