Systematic Textual Availability of Manuscripts

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

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

Abstract. The digital era has made millions of Hebrew manuscript images available to all. However, despite major advancements in handwritten text recognition (HTR) over the last decade, an efficient pipeline for large scale and accurate conversion of these manuscripts into useful machine-readable form is still sorely lacking.

We propose a cyclic pipeline to continuously improve the HTR models for automatic transcription of handwritten manuscripts. We use transfer learning to fine-tune pretrained models. And we leverage text reuse, a common phenomenon in medieval Hebrew manuscripts, for post-HTR correction. Our novel reuse framework, specifically designed for rabbinical Hebrew, utilizes fuzzy search over an inverted positional index followed by an approximate alignment algorithm.

Our framework successfully handles noisy HTR and consistently suggests alternate, better readings. Preliminary results show that word level accuracy increased by 10% for new readings proposed by the post-HTR text reuse corrections. Moreover, the HTR character level accuracy improved by 18% by fine-tuning a model on the first few pages of a manuscript.

Keywords: Handwritten text recognition · Transfer learning · Text reuse.

1 Introduction

The enormous collection of extant Hebrew manuscripts, kept in libraries and private collections worldwide, is a most valuable asset of historical, cultural and intellectual heritage. The digital era has brought a new renaissance to ancient and medieval handwritten manuscripts, heretofore available for study only to limited scholarly circles. The last decade's advancement in digitization has made images of these manuscripts accessible from every computer, notably through the Ktiv project of the National Library of Israel [43]. Yet, the colossal amount of manuscripts – more than 10 million images are expected with the completion of Ktiv project – along with the complexity of materials constitute a major hindrance on the way to full textual accessibility. Despite major advancements in optical character recognition (OCR) abilities in the recent decade, we are still

lacking an efficient framework for large-scale and accurate conversion of these manuscripts into a machine-readable form.

OCR paved the way to making printed texts available in machine readable representation. In recent years, with the rising prominence of artificial neural networks (ANN) and their application to OCR and handwritten text recognition (HTR), the accuracy of the automatic processes is continuously improving [8, 53]. The "Tikkun Sofrim" project [36] designed and tested an ANN based, automatic transcription pipeline for Hebrew manuscripts. The project leveraged open-source tools kraken [31] and eScriptorium [32], off-the-shelf methods for automatic page segmentation, layout analysis, and line segmentation followed by a crowd-sourcing platform developed to validate and correct the automatic transcriptions. However, kraken is designed to train a specific LSTM neural network model for each manuscript. This requires large efforts to prepare training labeled data for each manuscript. To dramatically reduce the quantity of manual annotation effort needed to create training sets for handwritten Hebrew text recognition we employ a bank of pretrained models as an ensemble of models in parallel, combining their results. Moreover, when minimal labeling of manuscripts is available, we use transfer learning to refine the accuracy of the pretrained models.

The crowd-sourcing efforts needed for transcription validation and correction are labor intensive. We aim to increase the pipeline efficiency by dramatically reducing transcription error rate using post-HTR correction algorithms. The most effective method at our disposal for automatically improving HTR transcriptions is the use of sequence alignment methods to line up the imperfectly deciphered texts with compositions in existing corpora or previously transcribed manuscripts of similar composition. This approach has been suggested in [69]. High-performance sequence alignment algorithms have long been used [24, 44, 47]. A recent adaptation targeted Hebrew [5]. Existing text alignment tools, however, generally assume accurate transcriptions, rather than error-riddled post-OCR texts. We propose a novel text reuse detection framework, designed for ancient Hebrew language, which utilizes fuzzy search on inverted positional index followed by a non-identical text's alignment algorithm to handle noisy OCR and propose new and better readings.

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 recently for these tasks [3, 9, 13, 14, 16, 21, 49, 57, 71], the majority using combinations of CNNs and LSTMs. Traditional computer vision methods have advantages for some types of manuscripts [52, 57]. State-of-the-art methods have been implemented in kraken [31] and eScriptorium [32] 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 [15, 29, 31]. HTR efforts working with medieval Hebrew manuscripts include [38, 32, 36]. 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 [36], 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 HTR) include [7, 30]. By appropriate data rendering and augmentation, deep learning models trained solely on synthetic manuscripts achieved good performance on original Tibetan Buddhist historical texts. Transductive methods, domain adaptation, cycle-consistent adversarial networks, and a combination of a domain-adversarial neural network approach with a convolutional recurrent neural network architecture were used to advantage [30].

Moreover, given an undeciphered manuscript, we can achieve the best possible reading by the use of the latest available bank of HTR models and algorithms. Aggregation and selection algorithms need to learn how to select the best automatic transcription model or combination of models for each specific manuscript [31, 49]. Letting OCR engines vote on readings has been done since at least the early 1990s [26]. Varying parameters of the input images (resolution, size, contrast) for each page can also have an impact, and image enhancement prior to OCR is commonplace. An attempt to apply this for Arabic was reported in [34]; automatically choosing the most successful among a variety of image enhancements was found to yield twice the improvement of lexical post-OCR correction.

2.2 Transfer Learning

Manuscript handwriting styles being highly dependent on time, place and individual scribes' 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 [19]. Recent research [2, 28] 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 [20], the authors successfully apply this concept on historical handwritten Italian titles of plays. The technique also allows one to transfer the representation from Arabic printed text to genuine handwriting [46]. Transductive transfer learning is used to good effect in [20].

2.3 NLP-Based Correction

Post-HTR error correction based on NLP techniques is a well-researched field. Pretrained language models of various kinds have been used to correct and refine OCR and HTR [37, 72], as well as optimized dedicated neural networks [18, 68]. This approach can be further improved by adding a classifier and a weighted confusion matrix [33]. In [41] an end-to-end jointly trained neural network for transcription and correction is proposed. State-of-the-art pretrained transformer-based contextual language models such as BERT [12] have been successfully used to detect and correct OCR errors [45] in English. Although a Hebrew version of BERT has recently been released [56], the Wikipedia-based dataset used to train it differs significantly in orthography and grammar from the old Hebrew used in manuscripts.

2.4 Text Reuse-Based Correction

Text reuse detection algorithms are used to locate the content of a manuscript within a library of reference texts [6], followed by alignment of the text against the most similar known text [1, 25], and to tackle potential failures of the automatic transcription [73].

Text Reuse Detection Manuscripts comprise human knowledge to be transmitted to others. The written transmission of information relied on a text reuse process whereby texts were either copied entirely or were borrowed partially to inspire new ideas. This leads to a phenomenon of many witnesses available for a single segment of text. The chance that several witnesses have been converted into a machine-readable form is high. For example, a manuscript segment could be matched with quoted fragments in later works, appear in dialog with previous authors [35, 47, 60] and also fit to spreading and amplifying ideas and opinions [60, 70]. Text reuse attracts the attention of researchers when considering ancient manuscripts for a large variety of languages (Greek [39], Chinese [67], Tibetan [35], Hebrew [48], Urdu [54]). Most of the studies so far have focused on exploring the potential of information technology to automate text-reuse detection in a specific domain. The text reuse detection framework usually expects the texts to be similar on a verbatim level or some level of paraphrase in between. However, we propose a novel framework that handles noisy HTR, thanks to which a gibberish-looking transcribed sentence could accurately be matched to reuses in other corpora.

Text Alignment The most effective method at our disposal for improving HTR is the use of sequence alignment methods to line up the imperfectly deciphered texts with compositions in existing corpora fetched by the text reuse detection framework. This approach has been suggested in [69]. Alternatively, one can align with previously transcribed manuscripts of the same composition. High-performance sequence alignment algorithms have long been in use [24, 44, 47].

An early work on aligning OCR text with ground-truth transcriptions is [50]. More recent work on aligning text with text and/or images includes [6, 10, 4]. A recent adaptation targeted Hebrew [5]. Existing text alignment tools, however, generally assume accurate transcriptions, rather than error-riddled post-OCR texts. Exceptions include [51, 66].

Language of our Corpus Practical text reuse detection and alignment challenges stem from the language of our interest. Hebrew is an orthographically and morphological complex language [27]. The number of valid inflected forms in Hebrew is 70 times larger than in English [23]. There is no orthographic standard in Hebrew. More specifically matres lectionis are mostly optional, a word may include it in one manuscript while it will be absent in another. We cannot know if a discrepancy is due to poor HTR or to an actual textual variant. Morphological analysis has been implemented in the text-reuse detection framework [58] to convert the tokens into its base form. Acronyms are ubiquitous in ancient written Hebrew. There are 17,000 different abbreviations in rabbinic literature, 35% of which are ambiguous [22] which challenge the alignment process. Furthermore, a Hebrew sentence can be written in multiple permutations while preserving its meaning, therefore reuses may take on different forms, which may be scored by the framework like [5, 59, 11].

3 Methodology

We designed a pipeline which extends [36]. Figure 1 illustrates our five step pipeline. (1) In the first step, the manuscripts are digitized. We rely on projects like Ktiv [43] digitizing the entire Hebrew manuscript corpus. Therefore, this step is external to our pipeline. (2) The second step deals with the transcription of the manuscript's pages. To minimize the transcriptions' character error rate, an interplay between the automatic transcription module and the text-reuse based correction module is employed. The former is based on [31] to segment and transcribe images, while the later rely on frequent text-reuses in Hebrew manuscripts to propose a new and better reading which in turn utilized to improve the transcriptions models. (3) In the third step, a group of experts correct the automatic transcriptions errors, if there are any, and approve the machine readable format. We use the user interface of [32] for this task. (4) In the fourth step, a text-reuse detection framework kicks in again to map all interconnections between the manuscript and other documents in the corpus. (5) Finally, the last step provides the outputs of the pipeline, openly, to digital humanities researchers.

4 Automatic Transcription

4.1 Handwritten Text Recognition

The automatic generation of transcribed text is achieved by the combination and integration of a variety of state-of-the-art algorithms. Core HTR is performed by

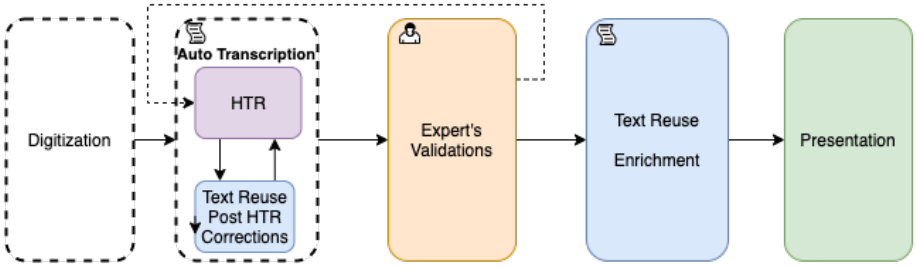

Fig. 1. Overview of the proposed pipeline.

the segmentation and recognition models trained on crowdsourced datasets [65] in the Sofer Mahir effort. The accuracy is boosted by automatically selecting the most appropriate model, either via a semi-automatic recommendation system or by unsupervised analysis of graphical features.

Furthermore, by manually labeling the first pages of the manuscript and fine-tuning the models' parameters [65] we improve the performance dramatically. Preliminary results show that character accuracy can be boosted by around 18% by fine-tuning the recognition models over three labeled pages (see Fig. 2). The particular choice of the source model does not seem to impact performance, nor adding more labeled data. We note that the same technique can be applied to segmentation models.

n.b. The original models were taken from [65] and are available from kraken's Zenodo archive [61–64]. Our data and source code will be made available as well.

4.2 Post-HTR: Text-Reuse Based Corrections

We leverage text reuse, a common phenomenon in Hebrew manuscripts, and run the HTR data through a text reuse detection framework which finds repetition pairs in the corpus and then align them based on a sequence alignment algorithm and propose a new and better reading for the HTR. Frameworks for short reuse detection first split large texts into small parts and try to detect reuses for each, commonly, n-gram over a sliding window [17]. However, kraken automatically segments the manuscript into rows. Therefore, we utilized these rows as our varying-size, non-overlapping sliding windows. In the rest of this section, we describe the pipeline of our text reuse detection framework as illustrated in Fig. 3. Our framework is tailored to Hebrew, on the one hand; on the other hand, it handles the expected noisy HTR inputs.

Prepossessing We use the Sefaria digital corpus [55] as our reference library. We expect our reference corpus to grow over time as we plan to insert new transcribed manuscripts into the reference library. We preprocess the reference library by removing special characters from the data as in [35]. Next we generate

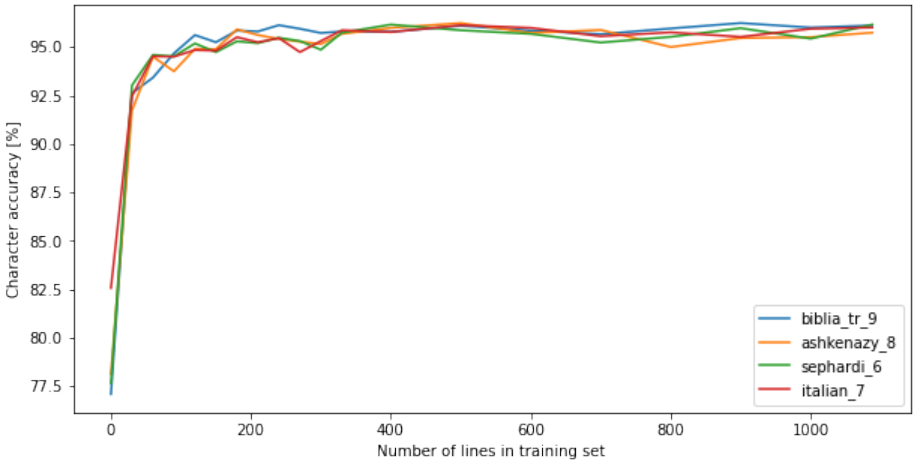

Fig. 2. Character accuracy achieved by transfer learning, as a function of additional labeled lines used for fine-tuning. Models are courtesy [65].

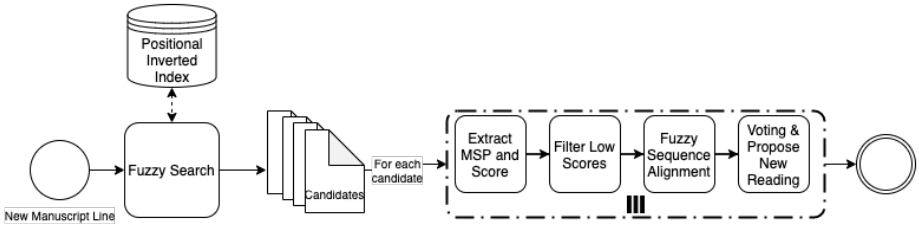

Fig. 3. Text reuse detection for post-HTR corrections.

a positional inverted index for fast candidate retrieval. In addition, we create a lexicon for each word in our corpus. A lexicon entry holds the inflected word as it appear in the corpus as well as its base form extracted by using a morphological analyser [42]. Each lexicon entry is enriched with the probability of its appearance in the corpus.

Candidate Retrieval For each manuscript line, we execute a fuzzy search against the positional inverted index. For each token in the input line, we seek orthographically close tokens to allow for transcription errors as well as Hebrew's orthographic variability. We end up with a list of all candidates suspected to have a text reuse relation with the tested row.

Scoring a Candidate The next step is to score the similarity between the tested line and each of the candidates. First we need to extract from each candidate a maximal segment pair (MSP), the most similar piece of text from the candidate with identical length to the tested line [1]. Then the score between the MSP and the input line is measured by edit distance [40]. At this stage, we also measure the similarity between the candidate and the previous and following rows of the manuscript. We boost the candidate's score relative to the similarity with the neighboring rows. The intuition here is that the longer a passage is shared between documents the higher the probability of a text reuse relation between them. We employ predefined similarity thresholds for the decision to move the candidate forward to the alignment stage, an approach used by most text reuse detection frameworks [17].

Fuzzy Sequence Alignment This stage aims to align all candidates against the tested row. As explained above, we align non-identical tokens. For example tokens with different orthographically, abbreviations or even synonyms are detected and aligned. Therefore, we assign a probability for each token's alignments measuring the confidence level of the framework about this alignment. For example, when two tokens with the same orthography are aligned, the probability will be set to 1.0. Aligned synonyms, acronyms or abbreviations will share the confidence level of their surrounding tokens. Tokens with different orthography will get their alignment probability as the edit distance ratio between them. However, if one of the tokens is missing from the Hebrew lexicon, the probability will be boosted.

The alignment stage starts with a "traditional" sequence alignment, which aligns tokens that share the same orthography [1]. The probability of these tokens' alignment is set to 1.0. Next we try to detect missing spaces. By its nature, manuscripts include varied sizes for word separators. That in turn cause the HTR, occasionally, to merge two words into a single one (miss the space in between) or wrongly detect space and split a single word into two. The framework will split/merge the tokens according to the missing spaces and reduce the alignment probability relatively. Lastly, we try to align non-identical tokens and assign the aliment probability accordingly.

Proposing a New Reading The final step is to choose the best reading for each token. Here we use majority vote between all available readings for each token. In this step only alignments that exceed a predefined threshold are included in the voting process. Preliminary results shows that our framework reduced the word error rate (WER) by 10%. The texts generated by the automatic transcription reached 81% of word level accuracy, while the new reading proposed by our text reuse framework boosted the accuracy to 91%.

4.3 Post-HTR: NLP-Based Correction

We apply post-HTR correction techniques, which draw on recent advancements in the field of natural language processing (NLP). The text reuse framework (detailed below) finds parallel citations and use majority vote to correct improbable readings. Language models trained over an appropriate corpus are used to automatically correct transcription errors. In both cases, if insufficiently certain of the correction, the improbable reading can be highlighted to a proofreading expert while possible alternate versions are explicitly suggested. Moreover, the advance of neural language models allows us to use the correction to fine-tune the whole transcription pipeline at the training stage.

4.4 Expert Proofreading

Following automatic machine transcription, a semi-automatic component allows experts to proof-read uncertain results. As detailed above, the suspect results and possible corrections are suggested by the automatic component. This integration of a machine-aided man-in-the-loop component allows to efficiently allocate the scarce resource of human expertise and attention.

5 Conclusion

The pipeline proposed in this paper aims to improve the accessibility of historical manuscripts in a machine readable form. Text reuse detection algorithms substantially improve the automatic transcription as a post processing component. The immediate gain are twofold. (1) First it minimizes the manual labor required by experts to validate and correct the transcription. Which in turn utilize to fine tune HTR models and improve the automatic transcription accuracy. (2) Second, the accuracy level reached automatically by our pipeline could be acceptable and deposit to open access repositories and public libraries, as is. This in turn will make the machine readable form available to all with an appropriate transparency declaration about the generation process. We expect that with additional fine tuning of the compared reference library against. which the manuscript is checked and text reuse thresholds selections the accuracy of the post processing could be increased even more. Given the flexibility of contemporary search engines we expect that even a non-perfect text will significantly improve the accessibility of knowledge both to scholars and the wider public.

Since the text reuse detection also includes mapping of the texts at the granularity of a line, it will also be possible to enable direct access to all images of each sentence.

The efficiency of the model we designed depend of the type of the text: Manuscripts of familiar works only demand identification of the work and alignment of the entire work with the manuscript text, expected to produced; manuscripts of anthological nature demand further scrutiny: identifying the most probable source of each paragraph. Manuscripts of paraphrase style will pose a complex challenge to our model. In the next stages of this research we intend to add an element of identifying the type of the text and automatically assigning fitting procedure in accordance (i.e automatic selection of the reference library for correction of the work or of each paragraph). We also intend to examine the efficiency of text reuse based feedback vs that of NLP based language models or the combination of both, and the adaptations needed to generalize the model and make it fit to other languages and corpora.

6 Acknowledgement

The work was partially supported by the Israeli Ministry of Science and Technology – grant number 3-17516.

References

1. Altschul, S.F., Gish, W., Miller, W., Myers, E.W., Lipman, D.J.: Basic local alignment search tool. Journal of Molecular Biology **215**(3), 403–410 (1990)
2. Aradillas, J.C., Murillo-Fuentes, J.J., Olmos, P.M.: Boosting offline handwritten text recognition in historical documents with few labeled lines. IEEE Access (2021)
3. Barakat, B., Droby, A., Kassis, M., El-Sana, J.: Text line segmentation for challenging handwritten document images using fully convolutional network. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 374–379. IEEE (2018)
4. Ben-Shalom, A., Silberpfennig, A., Dershowitz, N., Wolf, L., Choueka, Y.: Querying Hebrew texts via word spotting. In: World Congress of Jewish Studies. Jerusalem, Israel (Aug 2017)
5. Brill, O., Koppel, M., Shmidman, A.: FAST: Fast and accurate synoptic texts. Digital Scholarship in the Humanities **35**(2), 254–264 (2020)
6. Büchler, M., Burns, P.R., Müller, M., Franzini, E., Franzini, G.: Towards a historical text re-use detection. In: Text Mining, pp. 221–238. Springer (2014)
7. Carbonell, M., Mas, J., Villegas, M., Fornés, A., Lladós, J.: End-to-end handwritten text detection and transcription in full pages. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 5, pp. 29–34. IEEE (2019)
8. Charles, P.K., Harish, V., Swathi, M., Deepthi, C.: A review on the various techniques used for optical character recognition. International Journal of Engineering Research and Applications **2**(1), 659–662 (2012)

9. Chen, K., Seuret, M., Liwicki, M., Hennebert, J., Ingold, R.: Page segmentation of historical document images with convolutional autoencoders. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 1011–1015. IEEE (2015)
10. Cohen, R., Rabaev, I., El-Sana, J., Kedem, K., Dinstein, I.: Aligning transcript of historical documents using energy minimization. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 266–270. IEEE (2015)
11. Colavizza, G., Infelise, M., Kaplan, F.: Mapping the early modern news flow: An enquiry by robust text reuse detection. In: International Conference on Social Informatics. pp. 244–253. Springer (2014)
12. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
13. Diem, M., Kleber, F., Fiel, S., Grüning, T., Gatos, B.: cBAD: ICDAR2017 competition on baseline detection. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 1355–1360. IEEE (2017)
14. Droby, A., Kurar Barakat, B., Madi, B., Alaasam, R., El-Sana, J.: Unsupervised deep learning for handwritten page segmentation. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 240–245. IEEE (2020)
15. Dutta, K., Krishnan, P., Mathew, M., Jawahar, C.V.: Improving CNN-RNN hybrid networks for handwriting recognition. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 80–85. IEEE (2018)
16. Fink, M., Layer, T., Mackenbrock, G., Sprinzl, M.: Baseline detection in historical documents using convolutional u-nets. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS). pp. 37–42. IEEE (2018)
17. Foltỳnek, T., Meuschke, N., Gipp, B.: Academic plagiarism detection: A systematic literature review. ACM Computing Surveys (CSUR) **52**(6), 1–42 (2019)
18. Ghosh, S., Kristensson, P.O.: Neural networks for text correction and completion in keyboard decoding. arXiv preprint arXiv:1709.06429 (2017)
19. Goodfellow, I., Bengio, Y., Courville, A.: Deep Learning. MIT Press (2016), http://www.deeplearningbook.org
20. Granet, A., Morin, E., Mouchère, H., Quiniou, S., Viard-Gaudin, C.: Transfer learning for handwriting recognition on historical documents. In: 7th International Conference on Pattern Recognition Applications and Methods (ICPRAM) (2018)
21. 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)
22. HaCohen-Kerner, Y., Kass, A., Peretz, A.: Baseline methods for automatic disambiguation of abbreviations in Jewish law documents. In: International Conference on Natural Language Processing (in Spain). pp. 58–69. Springer (2004)
23. HaCohen-Kerner, Y., Schweitzer, N., Mughaz, D.: Automatically identifying citations in Hebrew-Aramaic documents. Cybernetics and Systems: An International Journal **42**(3), 180–197 (2011)
24. Haentjens Dekker, R., Middell, G.: Computer-supported collation with CollateX. In: Supporting Digital Humanities (2011)
25. Hakala, K., Vesanto, A., Miekka, N., Salakoski, T., Ginter, F.: Leveraging text repetitions and denoising autoencoders in OCR post-correction. arXiv preprint arXiv:1906.10907 (2019)

26. Handley, J.C., Hickey, T.B.: Merging optical character recognition outputs for improved accuracy. In: Intelligent Text and Image Handling, pp. 160–174 (1991)
27. Itai, A., Wintner, S.: Language resources for Hebrew. Language Resources and Evaluation **42**(1), 75–98 (2008)
28. Jaramillo, J.C.A., Murillo-Fuentes, J.J., Olmos, P.M.: Boosting handwriting text recognition in small databases with transfer learning. In: 2018 16th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 429–434. IEEE (2018)
29. Kahle, P., Colutto, S., Hackl, G., Mühlberger, G.: Transkribus–a service platform for transcription, recognition and retrieval of historical documents. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 4, pp. 19–24. IEEE (2017)
30. Keret, S., Wolf, L., Dershowitz, N., Werner, E., Almogi, O., Wangchuk, D.: Transductive learning for reading handwritten Tibetan manuscripts. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 214–221. IEEE (2019)
31. Kiessling, B.: Kraken–an universal text recognizer for the humanities. In: Digital Humanities (DH2019) (2019)
32. Kiessling, B., Tissot, R., Stokes, P., Stökl Ben Ezra, D.: eScriptorium: An open source platform for historical document analysis. In: 2019 International Conference on Document Analysis and Recognition Workshops (ICDARW). vol. 2, pp. 19–19. IEEE (2019)
33. Kissos, I., Dershowitz, N.: OCR error correction using character correction and feature-based word classification. In: 2016 12th IAPR Workshop on Document Analysis Systems (DAS). pp. 198–203. IEEE (2016)
34. Kissos, I., Dershowitz, N.: Image and text correction using language models. In: 2017 1st International Workshop on Arabic Script Analysis and Recognition (ASAR). pp. 158–162. IEEE (2017)
35. Klein, B.E., Dershowitz, N., Wolf, L., Almogi, O., Wangchuk, D.: Finding inexact quotations within a Tibetan Buddhist corpus. In: DH (2014)
36. 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)
37. Kukich, K.: Techniques for automatically correcting words in text. ACM Computing Surveys (CSUR) **24**(4), 377–439 (1992)
38. Kurar Barakat, B., El-Sana, J., Rabaev, I.: The Pinkas dataset. In: 2019 International Conference on Document Analysis and Recognition (ICDAR). pp. 732–737. IEEE (2019)
39. Lee, J.S.Y.: A computational model of text reuse in ancient literary texts. In: Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics. pp. 472–479 (2007)
40. Levenshtein, V.I., et al.: Binary codes capable of correcting deletions, insertions, and reversals. Soviet Physics Doklady **10**(8), 707–710 (1966)
41. Mahpod, S., Keller, Y.: Auto-ML deep learning for Rashi scripts OCR. CoRR **abs/1811.01290** (2018), http://arxiv.org/abs/1811.01290
42. More, A., Seker, A., Basmova, V., Tsarfaty, R.: Joint transition-based models for morpho-syntactic parsing: Parsing strategies for MRLs and a case study from modern Hebrew. Transactions of the Association for Computational Linguistics **7**, 33–48 (2019). https://doi.org/10.1162/tacl_a_00253, https://www.aclweb.org/anthology/Q19-1003

43. National Library of Israel: Digitized Hebrew manuscripts (Dec 2021), https://web.nli.org.il/sites/nlis/en/manuscript
44. Needleman, S.B., Wunsch, C.D.: A general method applicable to the search for similarities in the amino acid sequence of two proteins. Journal of Molecular Biology **48**(3), 443–453 (1970)
45. Nguyen, T.T.H., Jatowt, A., Nguyen, N.V., Coustaty, M., Doucet, A.: Neural machine translation with Bert for post-OCR error detection and correction. In: Proceedings of the ACM/IEEE Joint Conference on Digital Libraries in 2020. pp. 333–336 (2020)
46. Noubigh, Z., Mezghani, A., Kherallah, M.: Transfer learning to improve Arabic handwriting text recognition. In: 2020 21st International Arab Conference on Information Technology (ACIT). pp. 1–6. IEEE (2020)
47. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29**(1), 19–51 (2003)
48. Porat, E., Koppel, M., Shmidman, A.: Identification of parallel passages across a large Hebrew/Aramaic corpus. Journal of Data Mining & Digital Humanities (2018)
49. 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)
50. Rice, S.V., Kanai, J., Nartker, T.A.: An algorithm for matching OCR-generated text strings. In: Document Image Analysis, pp. 263–272. World Scientific (1994)
51. Romero-Gómez, V., Toselli, A.H., Bosch, V., Sánchez, J.A., Vidal, E.: Automatic alignment of handwritten images and transcripts for training handwritten text recognition systems. In: 2018 13th IAPR International Workshop on Document Analysis Systems (DAS). pp. 328–333. IEEE (2018)
52. Sadeh, G., Wolf, L., Hassner, T., Dershowitz, N., Stökl Ben Ezra, D.: Viral transcript alignment. In: 2015 13th International Conference on Document Analysis and Recognition (ICDAR). pp. 711–715. IEEE (2015)
53. Sahu, V.L., Kubde, B.: Offline handwritten character recognition techniques using neural network: A review. International Journal of Science and Research (IJSR) **2**(1), 87–94 (2013)
54. Sameen, S., Sharjeel, M., Nawab, R.M.A., Rayson, P., Muneer, I.: Measuring short text reuse for the Urdu language. IEEE Access **6**, 7412–7421 (2017)
55. Sefaria, I.: A living library of Torah texts online (Dec 2021), https://sefaria.org
56. Seker, A., Bandel, E., Bareket, D., Brusilovsky, I., Greenfeld, R.S., Tsarfaty, R.: AlephBERT: A Hebrew large pre-trained language model to start-off your Hebrew NLP application with. arXiv preprint arXiv:2104.04052 (2021)
57. 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)
58. Siegal, M.B.A., Shmidman, A.: Reconstruction of the Mekhilta Deuteronomy using philological and computational tools. Journal of Ancient Judaism **9**(1), 2–25 (2018)
59. Smith, D.A., Cordel, R., Dillon, E.M., Stramp, N., Wilkerson, J.: Detecting and modeling local text reuse. In: IEEE/ACM Joint Conference on Digital Libraries. pp. 183–192. IEEE (2014)
60. Smith, D.A., Cordell, R., Dillon, E.M.: Infectious texts: Modeling text reuse in nineteenth-century newspapers. In: 2013 IEEE International Conference on Big Data. pp. 86–94. IEEE (2013)

61. Stoekl Ben Ezra, D.: Medieval Hebrew manuscripts in Ashkenazi bookhand. https://zenodo.org/record/5468478 (2021), [Online; accessed 31-Jan-22]
62. Stoekl Ben Ezra, D.: Medieval Hebrew manuscripts in Italian bookhand version 1.0. https://zenodo.org/record/5468573 (2021), [Online; accessed 31-Jan-22]
63. Stoekl Ben Ezra, D.: Medieval Hebrew manuscripts in Sephardi bookhand version 1.0. https://zenodo.org/record/5468665 (2021), [Online; accessed 31-Jan-22]
64. Stoekl Ben Ezra, D.: Medieval Hebrew manuscripts version 1.0. https://zenodo.org/record/5468286 (2021), [Online; accessed 31-Jan-22]
65. Stoekl 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: The 6th International Workshop on Historical Document Imaging and Processing. pp. 61–66 (2021)
66. Stökl Ben Ezra, D., Brown-DeVost, B., Dershowitz, N., Pechorin, A., Kiessling, B.: Transcription alignment for highly fragmentary historical manuscripts: The Dead Sea scrolls. In: 2020 17th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 361–366. IEEE (2020)
67. Sturgeon, D.: Unsupervised identification of text reuse in early Chinese literature. Digital Scholarship in the Humanities **33**(3), 670–684 (2018)
68. Suissa, O., Elmalech, A., Zhitomirsky-Geffet, M.: Optimizing the neural network training for OCR error correction of historical Hebrew texts. In: iConference 2020 Proceedings. iSchools (2020)
69. Villegas, M., Toselli, A.H., Romero, V., Vidal, E.: Exploiting existing modern transcripts for historical handwritten text recognition. In: 2016 15th International Conference on Frontiers in Handwriting Recognition (ICFHR). pp. 66–71. IEEE (2016)
70. Wilkerson, J., Smith, D., Stramp, N.: Tracing the flow of policy ideas in legislatures: A text reuse approach. American Journal of Political Science **59**(4), 943–956 (2015)
71. Xu, Y., He, W., Yin, F., Liu, C.L.: Page segmentation for historical handwritten documents using fully convolutional networks. In: 2017 14th IAPR International Conference on Document Analysis and Recognition (ICDAR). vol. 1, pp. 541–546. IEEE (2017)
72. Zenkel, T., Sanabria, R., Metze, F., Niehues, J., Sperber, M., Stüker, S., Waibel, A.: Comparison of decoding strategies for CTC acoustic models. arXiv preprint arXiv:1708.04469 (2017)
73. Zhicharevich, A., Dershowitz, N.: Language classification and segmentation of noisy documents in Hebrew scripts. In: Proceedings of the 6th Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities. pp. 112–117 (2012)