Reverse Engineering Critical Apparatuses for HTR Ground Truth Creation: The Case of Kennicott's Collation of the Hebrew Bible

Luigi Bambaci[1], Nachum Dershowitz[2], Daniel Stökl Ben Ezra[3]
[1] École Pratique des Hautes Études, Paris Sciences & Lettres, Paris, France, luigi.bambaci@ephe.psl.eu
[2] School of Computer Science and AI, Tel Aviv University, Ramat Aviv, Israel, nachum@tau.ac.il
[3] École Pratique des Hautes Études, Paris Sciences & Lettres, Paris, France, Daniel.Stoekl@ephe.psl.eu

ABSTRACT (ITALIANO)

Gli apparati critici sono strumenti fondamentali per lo studio della storia della tradizione e critica del testo. Grazie alla registrazione delle varianti testuali emerse durante il processo di trasmissione, essi offrono dati imprescindibili per l'analisi filologica, linguistica, paleografica, e non solo. Tuttavia, la loro complessità strutturale e l'elevata mole di dati rendono difficile l'analisi su larga scala con approcci manuali. In questo contributo, presentiamo un piano di lavoro per la digitalizzazione e la codifica automatica degli apparati critici, basato su tecnologie all'avanguardia di OCR e trattamento automatico del linguaggio. Oltre alla digitalizzazione e alla codifica, perseguiamo un obiettivo innovativo: la ricostruzione automatica di testi manoscritti a partire dalle varianti registrate in apparato, con l'intento di generare nuova ground truth per migliorare l'addestramento di modelli di Handwritten Text Recognition (HTR). Il nostro caso studio si concentra sull'edizione di Benjamin Kennicott (1776-1780), la più ampia raccolta di varianti testuali mai realizzata per la Bibbia Ebraica, e oggetto del progetto *Reverse Engineering Kennicott* (REK). Dimostriamo come sia possibile segmentare, trascrivere e codificare automaticamente grandi volumi di dati testuali, generando testi manoscritti completi con intervento manuale contenuto. Affrontiamo le principali sfide tecniche e metodologiche emerse durante tutto il processo, facendo infine qualche previsione per il lavoro futuro.

Parole chiave: apparato critico; riconoscimento del testo manoscritto; trattamento del linguaggio naturale; manoscritti della Bibbia Ebraica; collazione di Kennicott

ABSTRACT (ENGLISH)

Ingegneria Inversa degli Apparati Critici per la Creazione di Ground Truth per l'HTR: Il Caso della Collazione di Kennicott della Bibbia Ebraica

Critical apparatuses are essential tools for the study of textual tradition and textual criticism. By recording the variants that emerge during the transmission process, they provide indispensable data for philological, linguistic, paleographic, and other analysis. However, their structural complexity and the sheer volume of data they contain make large-scale analysis through manual methods particularly challenging. In this paper, we present a workflow for the digitization and automatic encoding of critical apparatuses, based on state-of-the-art technologies in OCR and natural language processing. Beyond digitization and encoding, we pursue an innovative objective: the automatic reconstruction of manuscript texts from the variants recorded in the apparatus, with the goal of generating new ground truth data to improve the training of Handwritten Text Recognition (HTR) models.

Our case study focuses on Benjamin Kennicott's edition (1776–1780), the most extensive collection of textual variants ever compiled for the Hebrew Bible, and the object of the *Reverse Engineering Kennicott* (REK) project.

We demonstrate how large volumes of textual data can be segmented, transcribed, and encoded automatically, generating complete manuscript texts with limited manual intervention. Finally, we discuss the main technical and methodological challenges encountered throughout the process and outline directions for future work.

Keywords: critical apparatus; Handwritten Text Recognition; Natural Language Processing; Hebrew Bible manuscripts; Kennicott's collation

1. INTRODUCTION

Critical apparatuses are an indispensable resource for textual and philological research. Initially conceived as interfaces to justify editorial choices and provide readers with a synthetic view of textual traditions, they have progressively evolved into (semi-)structured data repositories serving purposes far beyond traditional textual philology—including the study of scribal practices, paleography, historical linguistics, and the computational analysis of textual transmission.

Despite their value, critical apparatuses remain challenging to manage. Their structural complexity and the volume of data they contain often limit accessibility, requiring analytical efforts that are not always scalable for large or data-driven projects. In recent years, digital technologies have opened new possibilities for utilizing critical apparatuses. Advances in Optical Character Recognition (OCR) have enabled automated transcription of multiformat and multilingual documents with impressive results. Natural Language Processing (NLP) techniques facilitate the organization and interpretation of complex information, while encoding systems like the Text Encoding Initiative (TEI) provide robust standards for preserving data in interoperable formats, ensuring long-term usability and exchangeability.

Among the studies on the digital processing of critical apparatuses relevant to our purposes, we highlight (Boschetti and Zemanek, 2007), which addresses the challenge of aligning textual variants present in critical apparatuses with their reference texts. By employing edit distance algorithms, the study proposes methodologies to accurately identify the position of each variant within the reference text and link it to additional annotations, such as morphological or metrical data. (Boschetti, 2007) examines approaches to enrich classical digital corpora by integrating textual variants and conjectures from critical apparatuses, using XML markup and parsing algorithms.

On the OCR/HTR front, (Boschetti et al., 2009) propose a workflow that combines multiple alignment techniques with customized OCR engines to improve the transcription of critical Greek and Latin editions, with particular attention to complex layouts including critical apparatuses. (Romanello et al., 2021) compare the performance of different OCR pipelines on 19th-century classical commentaries, highlighting the importance of ground truth datasets and advanced post-processing for optimizing character recognition in multilingual texts.

While the digitization and computational processing of critical apparatuses has been—as these studies testify—the focus of considerable attention and progress, the use of textual variants to automatically reconstruct manuscript texts for HTR training remains a largely unexplored field.

In a previous study (Bambaci and Stoekl Ben Ezra 2023), we took our first steps in this direction, using the critical apparatus from Kennicott's collation of the Hebrew Bible (Kennicott 1776–1780). Through the transcription and analysis of the first of Kennicott's two volumes, we successfully reconstructed complete texts for over 100 witnesses of the book of Genesis, generating data corresponding approximately to 5,800 manuscript pages.

In this paper, we aim to expand and refine those results, providing a detailed report on the progress made so far and the prospects for future development.

The approach presented here has been developed as part of the "Reverse Engineering Kennicott" (REK) project, dedicated to digitizing and encoding Kennicott's work and reconstructing the texts of his witnesses.

2. KENNICOTT'S COLLATION OF THE HEBREW BIBLE

The Hebrew Bible is among the most influential texts in human history, forming the foundation of Jewish and Christian religious and cultural heritage. Originally written in Hebrew and Aramaic, and translated into numerous languages since antiquity, it comprises approximately 300,000 tokens across 24 books.

One of the most ambitious scholarly efforts ever undertaken on its textual tradition is Benjamin Kennicott's *Vetus Testamentum Hebraicum cum variis lectionibus*, published in Oxford between 1776 and 1780. Drawing on over 600 manuscripts and 70 printed editions, this two-volume work documents an estimated 1.5 million textual variants, meticulously organized in a critical apparatus spanning roughly 1,400 pages—making it the most comprehensive and systematic collation ever assembled for the Hebrew witnesses of the Bible.

Besides its scope, three features make Kennicott's edition stand out from both earlier and later works: the clear distinction between fully and partially collated witnesses; the high level of granularity of the collation; and the highly formalized language of the apparatus.

The first two features have clear philological relevance, as they legitimise interpreting the absence of a variant as agreement with the reference text—effectively compensating for the negative structure of the apparatus, which records only deviations. The third, instead, plays a key role from a computational perspective: the use of a controlled vocabulary and consistent syntactic structures, combined with minimal reliance on natural language, facilitates parsing and structured data extraction.

Taken together, these characteristics make Kennicott's collation particularly well-suited for digital processing. We demonstrate this in the following section, where we outline the key operational phases of

our workflow, namely: image acquisition, segmentation and transcription, parsing and encoding, and finally manuscript text reconstruction followed by alignment with the original manuscripts.

3. WORKFLOW

The image acquisition phase involved two complementary datasets. First, we collected high-resolution images of both volumes of Kennicott's *Vetus Testamentum*, primarily sourced from Archive.org. In parallel, we obtained approximately 15,000 images from 20 Hebrew Bible manuscripts fully collated by Kennicott, retrieved from Ktiv[1] and selected for their historical and textual significance, as well as for the regularity of their layout and script.

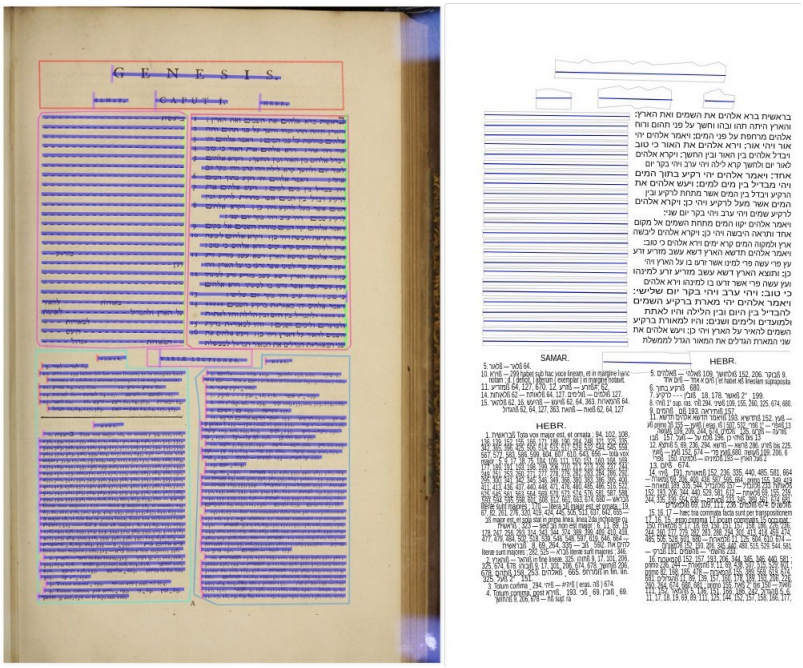

Figure 1: Segmentation and transcription of the *Vetus Testamentum*

The next step was segmentation and transcription. Kennicott's edition presents a particularly complex layout (Fig. 1), with two distinct textual zones: at the top, one or two columns of Hebrew containing the reference text used as the basis for his collation; at the bottom, two columns dedicated to the critical apparatus, where textual variants are recorded. The reference text is primarily in Hebrew with RTL (right-to-left) directionality, while the apparatus includes Hebrew and Latin text with LTR (left-to-right) directionality. Using eScriptorium cum Kraken (Stokes et al. 2021, Kiessling et al. 2019), we segmented the regions of interest (reference text and apparatus), excluding irrelevant paratextual elements such as headers, page numbers, and catchwords. We trained our segmentation models through iterative cycles of manual correction and retraining, using batches of 30–50 pages per round, until the desired level of segmentation accuracy was achieved.

Automatic transcription presented significant challenges, with notable differences between the treatment of the reference text and the critical apparatus. The transcription of the reference text was relatively straightforward, as it benefited from the biblical Hebrew HTR models developed within the BiblIA project (Stoekl Ben Ezra et al., 2021), which delivered highly accurate results. Transcribing the critical apparatus, by contrast, was significantly more complex, owing to its multilingual content, symbolic notation, and mixed text directionality (RTL Hebrew and LTR Latin). To address these challenges, we trained custom HTR models on targeted samples, following an iterative process similar to the one adopted for segmentation.

We then proceeded with parsing. For the reference text, we used simple Python scripts to align the automatic transcription with an already existing digital version of the biblical text, producing XML code structured into chapters, verses, and tokens. For the critical apparatus, a much more sophisticated approach was required. Using context-free grammars (CFGs) and the ANTLR4 software (Parr 2010, 2012),

[1] An international digital initiative that provides centralized access to thousands of Hebrew manuscripts from collections around the world, developed by the National Library of Israel. Available at: https://www.nli.org.il/en/discover/manuscripts/hebrew-manuscripts

and following the methodology outlined in detail in (Bambaci and Boschetti 2020, Bambaci 2021), we developed a rule-based parser capable of identifying apparatus components—such as lemmas, variants, and witness sigla (Fig. 2)—and automatically producing XML code (Fig. 3).

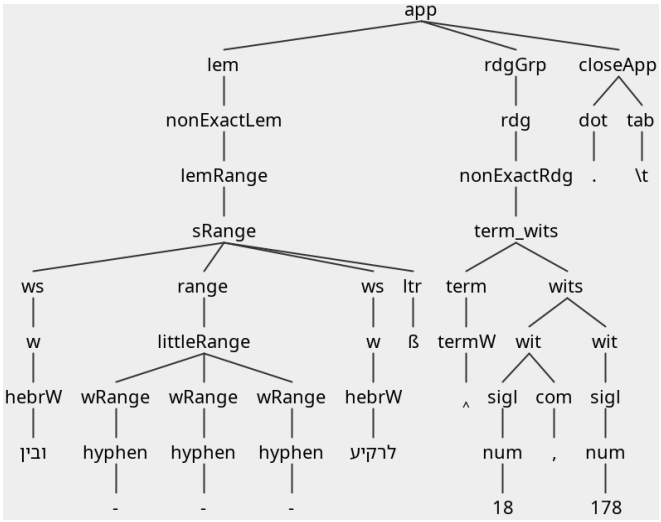

Figure 3: Parser's syntactic tree for Genesis 1:7

```
<app line="7" part="466569">
  <lem>
    <nonExactLem>
      <lemRange>
        <sRange>
          <ws>
            <w>
              <hebrW>וב׳ן</hebrW>
            </w>
          </ws>
          <range>
            <littleRange>
              <wRange>
                <hyphen>-</hyphen>
              </wRange>
              <wRange>
                <hyphen>-</hyphen>
              </wRange>
              <wRange>
                <hyphen>-</hyphen>
              </wRange>
            </littleRange>
          </range>
          <ws>
            <w>
              <hebrW>לרקי׳ע</hebrW>
            </w>
          </ws>
          <ltr>ß</ltr>
        </sRange>
      </lemRange>
    </nonExactLem>
  </lem>
  <rdgGrp>
    <closeApp>
      <dot>.</dot>
      <tab>\t</tab>
    </closeApp>
  <rdg>
    <nonExactRdg>
      <term_wits>
        <term>
          <termW>□</termW>
        </term>
        <wits>
          <wit>
            <sigl>
              <num>18</num>
            </sigl>
            <com>,</com>
          </wit>
          <wit>
            <sigl>
              <num>178</num>
            </sigl>
          </wit>
        </wits>
      </term_wits>
    </nonExactRdg>
  </rdg>
</rdgGrp>
</app>
```

Figure 2: XML output generated by the parser

The next phase focused on reconstructing complete manuscript texts from the variants recorded in the critical apparatus. This process involved three main steps, carried out using alignment techniques based primarily on Levenshtein distance: mapping each lemma from the apparatus onto its corresponding position in the reference text; reconstructing the Hebrew text of the variant readings when necessary; and inserting the variants into the reconstructed manuscript texts at the appropriate points, thus generating full versions of the witnesses with variants presented *in textu* (Figg. 4, 5).

```
<app line="7" part="466569">
  <lem>
    <exactLem>
      <recoLem>
        <s>
          <ws>
            <w>וב׳ן</w>
            <w n="2">המ׳ם</w>
            <w n="2">אשר</w>
            <w>מעל</w>
            <w n="2">לרקי׳ע</w>
          </ws>
          <ltr>ß</ltr>
        </s>
        <recoLemSep>] </recoLemSep>
      </recoLem>
    </exactLem>
  </lem>
  <rdgGrp>
    <rdg>
      <exactRdg>
        <rdgContent>
          <w line="7"/>
          <w line="7"/>
          <w line="7"/>
          <w line="7"/>
          <w line="7"/>
        </rdgContent>
        <wits>
          <wit>
            <sigl>
              <num>18</num>
            </sigl>
            <com>,</com>
          </wit>
          <wit>
            <sigl>
              <num>178</num>
            </sigl>
          </wit>
        </wits>
      </exactRdg>
    </rdg>
  </rdgGrp>
  <closeApp>
    <dot>.</dot>
    <tab> </tab>
  </closeApp>
</app>
```

Figure 5: XML apparatus with reconstructed lemma

```
<ab chap="1" n="7">
  <w>ויעש</w>
  <w>אלהים</w>
  <w>את</w>
  <w>הרקיע</w>
  <w>ויבדל</w>
  <w>בין</w>
  <w>המים</w>
  <w>אשר</w>
  <w>מתחת</w>
  <w>לרקיע</w>
  <app>
    <lem>
      <w line="7"/>
      <w line="7"/>
      <w line="7"/>
      <w line="7"/>
      <w line="7"/>
    </lem>
    <rdg>
      <w>וב׳ן</w>
      <w n="2">המ׳ם</w>
      <w n="2">אשר</w>
      <w>מעל</w>
      <w n="2">לרקי׳ע</w>
    </rdg>
  </app>
  <w>ויהי</w>
  <w>כן</w>
  <w>:</w>
</ab>
```

Figure 4: XML output of the manuscript text (ms. no. 18)

The final phase is textual alignment. We first automatically transcribe the manuscript images using the dedicated segmentation and transcription models mentioned above. We then align these automatic transcriptions with the reconstructed texts using Passim (Smith 2015, 2023), a text-reuse detection tool recently integrated into
eScriptorium (Fig. 6).

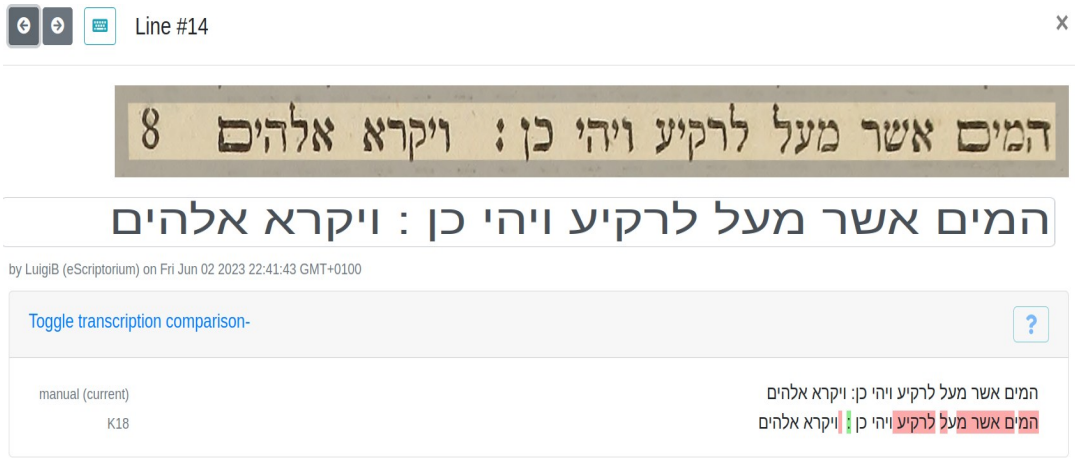

Figure 6: Passim alignment in eScriptorium: reference text vs. reconstructed text (ms. no. 18)

This alignment enables us to systematically compare the two outputs, identifying both errors—whether introduced during the automatic transcription phase or during the reconstruction of the manuscript texts from the apparatus—and actual variants. A portion of these aligned transcriptions will be manually corrected to produce a gold-standard dataset, which will be then used as ground truth to refine and train new Hebrew HTR models.

4. RESULTS

To date, we have completed the segmentation and transcription of both volumes of the *Vetus Testamentum*, amounting to approximately 90,000 lines for the reference text and 115,000 lines for the critical apparatus.

Ten of the 20 manuscripts fully collated by Kennicott have also been entirely segmented and automatically transcribed.

Parsing has been completed for the entire first volume, covering nine books of the Hebrew Bible, from Genesis to Kings (the Enneateuch). The automatic encoding of the critical apparatus produced approximately 50,000 lemmas (<lem>) and 100,000 variants (<rdg>).

In terms of reconstruction, we generated complete texts for 153 manuscripts, amounting to around 250,000 lemmas, an equal number of variants, and approximately 14 million tokens of text.

We are now aligning the reconstructed texts with the automatic transcriptions of the 10 manuscripts using Passim. We are still in the early stages of this process: while preliminary tests have confirmed the technical feasibility of the procedure, most of the alignment work remains to be completed. The coming months will be crucial for assessing the overall reliability of the reconstruction process and, consequently, for the creation of ground truth data.

Once this phase is completed, we will move on to Kennicott's second volume and to the remaining ten manuscripts, through which we expect to approximately double the volume of extracted and reconstructed data units. All relevant textual data, currently encoded in custom XML, will be converted into TEI-compliant format to ensure broader reusability.

5. CONCLUSION

The workflow we have implemented has proven effective across the various phases of the project. Segmentation and transcription models achieved an overall accuracy of around 98%, significantly reducing the need for manual correction. The *Vetus Testamentum* required considerable training effort during the initial phases, but the results—refined through multiple revision cycles—are now highly satisfactory. As for the ten manuscripts, the results also appear promising, thanks both to the regularity of their layout and the quality of their script (which guided their selection), as well as to the performance of the pretrained segmentation and transcription models used.

The parser achieved approximately 97% accuracy—a notable result given the volume and complexity of the data. The highly formalized language of Kennicott's apparatus allowed for the automatic processing and encoding of tens of thousands of entries, with minimal manual intervention.

The automatic reconstruction of manuscript texts from variant data proved to be a viable solution, allowing for the systematic generation of large textual outputs.
Finally, the alignment system based on Passim is proving effective for matching the reconstructed texts with the original manuscript transcriptions. Its seamless integration into eScriptorium's user-friendly interface, in particular, is greatly facilitating the ongoing correction.
While the results are promising, the workflow has also revealed a number of challenges and unresolved issues that deserve careful consideration.
First, although Kennicott's critical apparatus is highly formalized, some residual cases remain—due either to occasional use of natural language, to rare patterns, or to a combination of both—which require specific adjustments to the original transcriptions for proper parsing. However, such adjustments are negligible in quantitative terms and not critical in qualitative ones, as the philological information is preserved.
Numerically more significant are certain systematic ambiguities (approximately 3% of cases), where variants are mistakenly parsed as lemmas. Fortunately, we are able to identify and resolve them automatically by cross-checking the apparatus data against the reference text. Spot checks on particularly complex apparatus entries have not revealed additional semantic errors, though their presence elsewhere cannot be entirely ruled out.
Substantial manual intervention is instead proving necessary in the phase of textual reconstruction. Among these are cases in which multiple lemmas are returned with identical Levenshtein distance scores. Although numerous, they are generally easy to disambiguate manually, as they merely require selecting the correct match. Far more complex to handle are the approximately 4% of lemmas and variants that fall outside our mapping rules, due to irregular or complex structures. While modest in relation to the total, this percentage still amounts to several hundred cases that have required—and will likely continue to require—not only manual intervention, but also a certain degree of subjective judgment.
A final consideration concerns the overall reliability of Kennicott's collation. While extraordinary for its time, the accuracy of his work—as many scholars have long noted—is not always consistent, and the reconstructed texts will inevitably reflect this.
Since the alignment phase is still underway, we are not yet able to provide the necessary statistics on such flaws, nor on errors introduced during transcription, parsing, or automatic reconstruction. However, we anticipate that these will be unevenly distributed: minimal in the case of transcription and parsing, where both the models and the source data have proven reliable, and more frequent in the reconstruction process, where manual intervention has been greater and where Kennicott's collation may at times prove less reliable.
While the alignment phase is still ongoing, the results obtained so far confirm the viability of reconstructing manuscript texts from structured variant data, provided a sufficient degree of formalization in the apparatus. As this phase advances, we expect it to consolidate the workflow, clarify its limitations, and support the creation of reliable ground truth datasets for future applications in HTR of Hebrew biblical manuscripts.
All the textual data relevant to the project as well as the tools implemented and used will be shared on our GitHub repository,[2] and possibly on Zenodo.

6. ACKNOWLEDGEMENT

Our research received generous funding from the Agence Nationale de Recherche as part of the Programme d'investissements d'avenir within the France 2030 framework, under the reference ANR-21-ESRE-0005. Additionally, we benefited from funding by the European Union through the MiDRASH project (ERC, project number 101071829). Please note that the views and opinions expressed in this paper are those of the author(s) alone and do not necessarily represent the views of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for the expressed views.

7. REFERENCES

Bambaci, L. (2021). Critical Apparatus as Domain Specific Languages. A Rule-based Parser for Encoding an Eighteenth-Century Collation of Hebrew Manuscripts. *International Journal of Information Science and Technology*, *5*(1), 22–33.

Bambaci, L., & Boschetti, F. (2020). Encoding the Critical Apparatus by Domain Specific Languages – The Case of the Hebrew Book of Qohelet. In C. Marras, M. Passarotti, G. Franzini, & E. Litta (Eds.), *Atti del*

[2] https://github.com/LuigiBambaci/

IX Convegno Annuale AIUCD. La svolta inevitabile: Sfide e prospettive per l'Informatica Umanistica (pp. 7–13). Università Cattolica del Sacro Cuore. http://amsacta.unibo.it/6316/

Bambaci, L., & Stoekl Ben Ezra, D. (2023). Enhancing HTR of Historical Texts through Scholarly Editions: A Case Study from an Ancient Collation of the Hebrew Bible. In A. Šeļa, F. Jannidis, & I. Romanowska (Eds.), *Proceedings of the Computational Humanities Research Conference 2023* (Vol. 3558, pp. 554–576). CEUR. https://ceur-ws.org/Vol-3558/#paper6310

Boschetti, F. (2007). Methods to Extend Greek and Latin Corpora with Variants and Conjectures: Mapping Critical Apparatuses Onto Reference Text. *Proceedings of the Corpus Linguistics Conference CL, 14-17 Jul. 2007*. https://www.birmingham.ac.uk/documents/college-artslaw/corpus/conference-archives/2007/150Paper.pdf

Boschetti, F., Romanello, M., Babeu, A., Bamman, D., & Crane, G. (2009). Improving OCR Accuracy for Classical Critical Editions. In M. Agosti, J. Borbinha, S. Kapidakis, C. Papatheodorou, & G. Tsakonas (Eds.), *Research and Advanced Technology for Digital Libraries* (pp. 156–167). Springer Berlin Heidelberg.

Boschetti, F., & Zemanek, P. (2007). Alignment of variant readings for linkage of multiple annotations. *Proceedings of the ECAL 2007 Electronic Corpora of Ancient Languages, Prague 16-17 November 2007*, 11–24.

Kennicott, B. (1776-1780). *Vetus Testamentum Hebraicum cum variis lectionibus*, 2 Voll. Clarendon.

Kiessling, B., Tissot, R., Stokes, P., & Stoekl Ben Ezra, D. (2019). eScriptorium: An Open Source Platform for Historical Document Analysis. *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*, *2*, 19–24. https://doi.org/10.1109/ICDARW.2019.10032

Parr, T. (2010). *Language Implementation Patterns: Create Your Own Domain-specific and General Programming Languages*. Pragmatic Bookshelf.

Parr, T. (2012). *The Definitive ANTLR 4 Reference*. Pragmatic Bookshelf.

Romanello, M., Najem-Meyer, S., & Robertson, B. (2021). Optical character recognition of 19th century classical commentaries: The current state of affairs. *The 6th International Workshop on Historical Document Imaging and Processing*, 1–6.

Smith, D. A., Cordell, R., & Mullen, A. (2015). Computational Methods for Uncovering Reprinted Texts in Antebellum Newspapers. *American Literary History*, *27*(3), E1–E15. https://doi.org/10.1093/alh/ajv029

Smith, D. A., Murel, J., Allen, J. P., & Miller, M. T. (2023). Automatic Collation for Diversifying Corpora: Commonly Copied Texts as Distant Supervision for Handwritten Text Recognition. *Proceedings of the Computational Humanities Research Conference 2023, Paris, France, December 6-8, 2023.*, 206–221. https://ceur-ws.org/Vol-3558/paper1708.pdf

Stoekl Ben Ezra, D., Brown-DeVost, B., Jablonski, P., Lapin, H., Kiessling, B., & Lolli, E. (2021). BiblIA - A General Model for Medieval Hebrew Manuscripts and an Open Annotated Dataset. *The 6th International Workshop on Historical Document Imaging and Processing*, 61–66. https://doi.org/10.1145/3476887.3476896

Stokes, P., Kiessling, B., Stoekl Ben Ezra, D., Tissot, R., & Gargem, E. (2021). The eScriptorium VRE for Manuscript Cultures. *Ancient Manuscripts and Virtual Research Environments*, *Special issue of Classics@ 18*. https://classics-at.chs.harvard.edu/classics18-stokes-kiessling-stokl-ben-ezra-tissot-gargem/