Enhancing HTR of Historical Texts through Scholarly Editions: A Case Study from an Ancient Collation of the Hebrew Bible

Luigi Bambaci[1], Daniel Stökl Ben Ezra[2]

[1] *Archéologie & Philologie d'Orient et d'Occident UMR 8546, École Pratique des Hautes Études, Université Paris Sciences & Lettres (EPHE, PSL), Les Patios Saint-Jacques, 4-14 Rue Ferrus, 75014 Paris, France*

[2] *Archéologie & Philologie d'Orient et d'Occident UMR 8546, EPHE, PSL*

Abstract

Printed critical editions of literary texts are a largely neglected source of knowledge in computational humanities. However, under certain conditions, they hold significant potential for multifaceted exploration: First, through Optical Character Recognition (OCR) of the text and its apparatus, coupled with intelligent parsing of the variant readings, it becomes possible to reconstruct comprehensive manuscript collations, which can prove invaluable for a variety of investigations, including phylogenetic analyses, redaction history studies, linguistic inquiries, and more. Second, by aligning the printed edition with manuscript images, a substantial amount of Handwritten Text Recognition (HTR) ground truth can be generated. This serves as valuable material for paleography, layout analysis, as well as for assessing the quality of the collation criteria adopted by the editor. The present paper focuses on the challenges mastered in the processes of the OCR, the apparatus parsing, the text reconstruction, and the alignment with the manuscript images, taking as a case study the edition of the Hebrew Bible published by Kennicott in the late eighteenth century.

Keywords

layout analysis, automatic transcription, text encoding, Hebrew Bible manuscripts, textual criticism

1. Introduction

For centuries, critical editions have served as the backbone of the humanities far beyond philology, offering important insights into the textual evolution of numerous historical works and providing scholars with reliable texts for their academic inquiries. The advent of Optical Character Recognition (OCR) and Handwritten Text Recognition (HTR) technologies as well as Natural Language Processing (NLP) has opened a new era both in the preservation and in the exploration of these indispensable works.

Numerous OCR and HTR software solutions are available today, and multiple studies and projects have contributed to the advancement of digitizing and analyzing the cultural book heritage. Among the most well-known software, we can mention Transkribus [13], Monk [21],

CHR 2023: Fourth Conference on Computational Humanities Research, December 6–8, 2023, École pour l'Informatique et les Techniques Avancées, Paris, France

luigi.bambaci@ephe.psl.eu (L. Bambaci); daniel.stoekl@ephe.psl.eu (D. Stökl Ben Ezra)
0009-0009-2152-5221 (L. Bambaci); 0000-0001-5668-493X (D. Stökl Ben Ezra)

Aletheia [7], and Tesseract 4.0.[1] Notable research efforts relevant to ours include the work by Toselli *et al.* [26], who exploited huge datasets of existing OCRed printed books for self-supervised layout analysis, as well as projects like HORAE [5], which examined large amounts of Biblical texts or quotations in Latin to create HTR ground truth and conduct manuscript analysis.

The focus of these advancements has predominantly revolved around traditions within classical languages or modern languages in the Latin alphabet. However, there has been a recent shift towards including Hebrew texts in such endeavors: one eminent example is the BiblIA project [25], which examined a substantial corpus of Medieval manuscripts in Hebrew script, providing the first public dataset of transcriptions as well as efficient models for automatic segmentation and text recognition.

In this paper, we aim to contribute to the ongoing progress in digital Hebrew research, focusing in particular on the corpus of scholarly editions and biblical manuscripts. We will delve into the challenges faced in digitizing and encoding an ancient edition of the Hebrew Bible, namely the eighteenth-century collation by Benjamin Kennicott, and we will elucidate how we extracted from it a large amount of complete manuscript texts that we will be able to align with their manuscript images.

The significance of Kennicott's collation for biblical studies remains unparalleled. The wealth of data it offers is exceptional and its potential applications are manifold, as we will elaborate shortly (§ 2). Yet, the sheer volume and complexity of the data constitute a significant obstacle to analysis, compelling scholars to work on limited samples and to perform laborious manual processing.

Through the digitization of this edition, our aim is to provide the scholarly community with a digital resource for swift, efficient, and large-scale examinations of Hebrew Bible manuscripts. Additionally, we will leverage Kennicott's collation for an unprecedented purpose: enhancing the performance of HTR systems using automatically reconstructed texts derived from the critical apparatus data. The automatic generation of these texts will afford us a massive amount (approximately 75,000 pages) of ground truth for HTR of Hebrew manuscripts, while the alignment with the images will enable us not only to measure the degree of discrepancy between the original collation and the actual manuscripts, but also to correct errors, fill gaps, and produce more faithful and updated collation data.

In the next sections, we will provide a detailed account of our work. We will elaborate on how we conducted layout analysis for the purpose of segmentation and transcription (§§ 3.2, 3.3), and how we automatically encoded the data present in the critical apparatus using a rule-based parser (§ 3.4). Lastly, we will demonstrate how we successfully generated complete texts of fully collated witnesses and how we are going to use them as training data to improve and speed up the automatic transcription of a number of manuscripts of the Hebrew Bible (§ 4).

The method we are about to present here is part of an ongoing project entitled Reverse Engineering Kennicott (REK), funded by Biblissima+[2] and directed by the École Pratique des Hautes Études, Paris Sciences et Lettres University. The project is carried out in close synergy

[1] https://github.com/tesseract-ocr/tesseract.

[2] https://biblissima.fr/.

with Ktiv,[3] the most important online catalog of Hebrew manuscripts, and with the National Library of Israel,[4] and is centered on the web application eScriptorium.[5]

At the time of writing this article, the first of the two volumes of Kennicott's work has been encoded, and the text of the witnesses of the book of Genesis have been automatically generated.

Before illustrating the pipeline of our project, let us briefly outline Kennicott's work, in order to explain why it is so important for biblical research and how it can be used to fully recover the text of medieval manuscripts.

2. Kennicott's collation of the Hebrew Bible

The Hebrew Bible is a compilation of texts from the first millennium before the common era, written primarily in Hebrew with some sections in Aramaic, and totaling around 470,000 tokens. Sacred text in Judaism and later also Christianity, with numerous translations into many ancient languages such as Greek, Latin, Aramaic, Armenian, Coptic, Georgian, Arabic and, since the reformation era, into virtually all contemporary languages, it is one of the most important texts extant worldwide.

Kennicott was the first scholar to systematically gather and collate the Hebrew textual witnesses of the Bible.

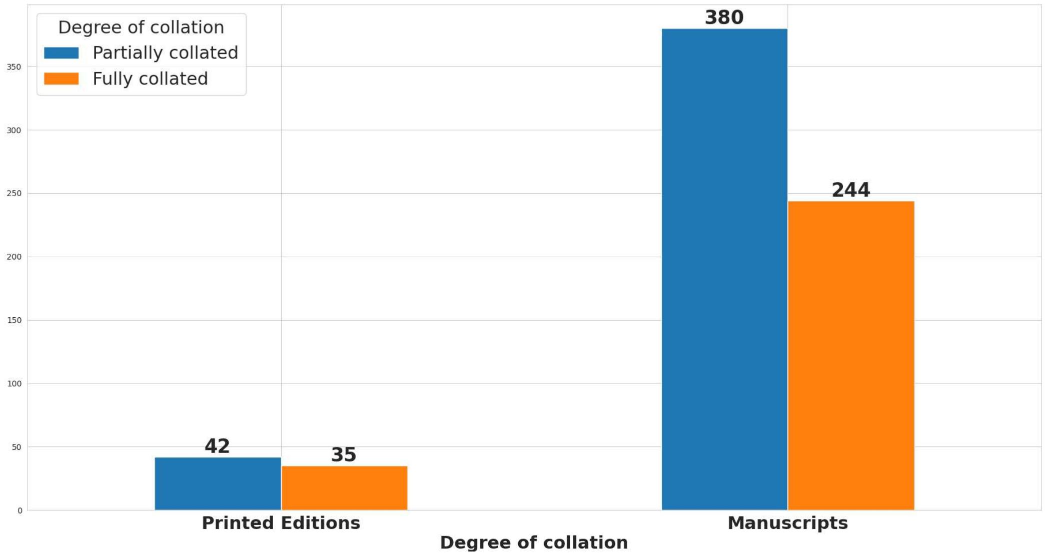


Figure 1: Count of the collated witnesses in the two volumes of the *Vetus Testamentum*. Witnesses are categorized by typology (printed editions versus manuscripts) and degree of collation (partially collated versus fully collated).

His two-volumes collation, published at Oxford between 1776 and 1780 and titled *Vetus Testa-*

[3] https://www.nli.org.il/en/discover/manuscripts/hebrew-manuscripts.
[4] https://www.nli.org.il/en.
[5] https://msia.escriptorium.fr/.

mentum Hebraicum cum variis lectionibus [14, 15], remains the largest of its kind to this day: its extensive critical apparatus, built upon the examination of no fewer than 600 manuscripts and 70 printed editions (Fig. 1), is estimated to contain something like 1,500,000 pieces of textual information.[6]

Kennicott's work has never been replaced: De Rossi's collations [10, 9], which were published shortly afterwards, are highly eclectic and present only a restricted selection of variants, while later editions either depend on these classical collations,[7] or drastically reduce the number of collated manuscripts,[8] or even dispense with the testimony of medieval manuscripts altogether.[9]

The use of Kennicott's data is not confined solely to consultation or the compilation of critical editions. Scholars have repeatedly demonstrated how it is possible to extract relevant research information out of Kennicott's apparatus: from textual history, enabling the reconstruction of the transmission process of the Hebrew Bible in the Middle Ages through stemmatological methods, such as clustering [6] and phylogenetics [3, 2]; to philology, for the study of common copying errors and scribal habits; from codicology and paleography, aiding in dating and localizing new manuscripts [19, 12]; to linguistics, allowing the analysis of variant spelling and orthography [8].

Kennicott's collation is a valuable resource for research across all these domains. Indeed, it stands as the inaugural and, up to present, solitary endeavor to provide a scholarly edition of the medieval Hebrew Bible text. Let us take a closer look at some of its key features.

The work is organized into sections, each dedicated to a biblical book (e.g. Genesis, Exodus etc.) or to a collection of biblical books (e.g. the Five Megilloth: Song of Songs, Ruth etc.). Each section comprises two main parts: a reference text[10] printed at the top page and a critical apparatus of variants printed at the bottom (Fig. 2).

In the apparatus, the witnesses are cited using unique alphanumeric *sigla*. Keys to these *sigla* are provided in the catalog reproduced in the introduction to the first volume, containing the most relevant bibliographical information, such as date and provenance.[11]

In addition to this catalog, Kennicott provides recapitulative lists of witnesses at the end of each book or collection of books. The purpose of these lists is to categorize the witnesses into manuscripts and printed editions, as well as to signal which of them have been collated in full (*per totum collati*) and which only partially (*in loci selectis collati*). This distinction by degree of collation, which is absent, for example, in De Rossi, is of fundamental importance and directly impacts our work: since only fully collated witnesses can provide the basis for a systematic gathering of variants, it permits us to identify which are the witnesses we can reasonably expect to obtain complete and reliable automatic transcriptions for.

[6] [4] 28ff.

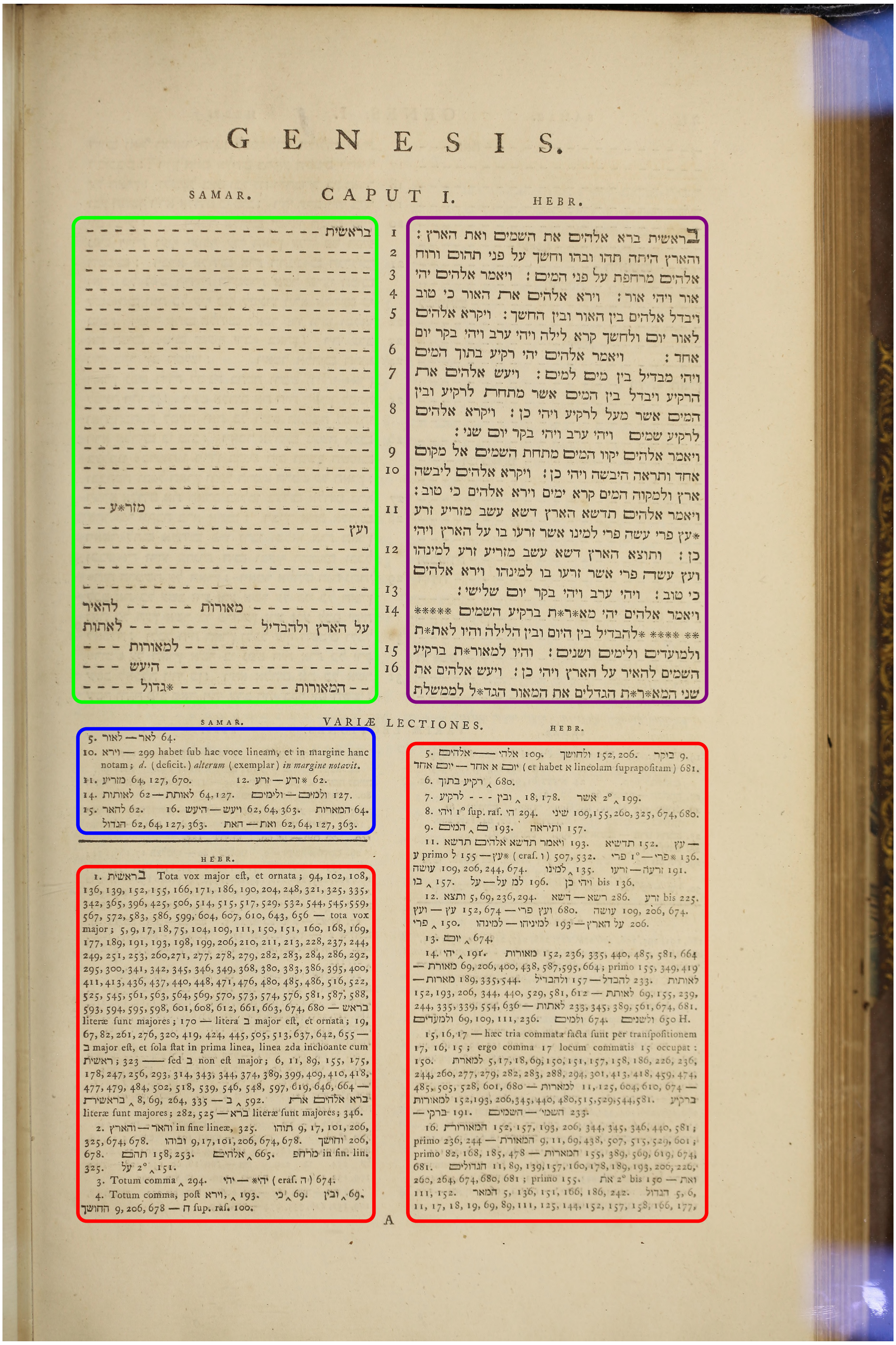
[7] So for example *the Biblia Hebraica Stuttgartensia* [11].

[8] Like the Hebrew University Bible [22].

[9] As the *Biblia Hebraica Quinta* [20].

[10] Taken from the most widely used edition at the time, that of E. van der Hooght (Amsterdam 1705), which Kennicott adopted as basis for his collation.

[11] Most of Kennicott's manuscripts have been identified: in 2020, Idan Dershowitz published a comprehensive list of these manuscripts containing URLs to Ktiv, where updated bibliographic information and, when available, images can be found. This list is accessible on the author's academia.edu page: https://www.academia.edu/37862623.



GENESIS.

SAMAR. CAPUT I. HEBR.

- - - - - - - - - - - - - - - - - בראשית	1	בראשית ברא אלהים את השמים ואת הארץ :
- - - - - - - - - - - - - - - - - - - -	2	והארץ היתה תהו ובהו וחשך על פני תהום ורוח
- - - - - - - - - - - - - - - - - - - -	3	אלהים מרחפת על פני המים : ויאמר אלהים יהי
- - - - - - - - - - - - - - - - - - - -	4	אור ויהי אור : וירא אלהים את האור כי טוב
- - - - - - - - - - - - - - - - - - - -	5	ויבדל אלהים בין האור ובין החשך : ויקרא אלהים
- - - - - - - - - - - - - - - - - - - -		לאור יום ולחשך קרא לילה ויהי ערב ויהי בקר יום
- - - - - - - - - - - - - - - - - - - -	6	אחד : ויאמר אלהים יהי רקיע בתוך המים
- - - - - - - - - - - - - - - - - - - -	7	ויהי מבדיל בין מים למים : ויעש אלהים את
- - - - - - - - - - - - - - - - - - - -		הרקיע ויבדל בין המים אשר מתחת לרקיע ובין
- - - - - - - - - - - - - - - - - - - -	8	המים אשר מעל לרקיע ויהי כן : ויקרא אלהים
- - - - - - - - - - - - - - - - - - - -		לרקיע שמים ויהי ערב ויהי בקר יום שני :
- - - - - - - - - - - - - - - - - - - -	9	ויאמר אלהים יקוו המים מתחת השמים אל מקום
- - - - - - - - - - - - - - - - - - - -	10	אחד ותראה היבשה ויהי כן : ויקרא אלהים ליבשה
- - - - - - - - - - - - - - - - - - - -		ארץ ולמקוה המים קרא ימים וירא אלהים כי טוב :
- - מזר*ע - - - - - - - - - - - - - - -	11	ויאמר אלהים תדשא הארץ דשא עשב מזריע זרע
- - - - - - - - - - - - - - - - - - ועץ		*עץ פרי עשה פרי למינו אשר זרעו בו על הארץ ויהי
- - - - - - - - - - - - - - - - - - - -	12	כן : ותוצא הארץ דשא עשב מזריע זרע למינהו
- - - - - - - - - - - - - - - - - - - -		ועץ עשה פרי אשר זרעו בו למינהו וירא אלהים
- - - - - - - - - - - - - - - - - - - -	13	כי טוב : ויהי ערב ויהי בקר יום שלישי :
להאיר - - - - - - מאורות - - - - - - - - -	14	ויאמר אלהים יהי מא*ר*ת ברקיע השמים *****
לאתות - - - - - - - - - - - - ולהבדיל על הארץ		** **** *להבדיל בין היום ובין הלילה והיו לאת*ת
- - - - למאורות - - - - - - - - - - - - - -	15	ולמועדים ולימים ושנים : והיו למאור*ת ברקיע
- - - - היעש - - - - - - - - - - - - - - -	16	השמים להאיר על הארץ ויהי כן : ויעש אלהים את
- - - - - גדול* - - - - - - - - - - - המאורות - -		שני המא*ר*ת הגדלים את המאור הגד*ל לממשלת

SAMAR. VARIÆ LECTIONES. HEBR.

5. לאור — לאר 64.
10. וירא — 299 habet ſub hac voce lineam, et in margine hanc notam; *d.* (deficit.) *alterum* (exemplar) *in margine notavit.*
11. מזריע 64, 127, 670. 12. זרע — זרע* 62.
14. לאותות 62 — לאותת 64, 127. ולמים — ולימים 127.
15. להאר 62. 16. היעש — ויעש 62, 64, 363. המארות 64.
הגדול 62, 64, 127, 363. האת — ואת 62, 64, 127, 363.

HEBR.

1. בראשית Tota vox major eſt, et ornata; 94, 102, 108, 136, 139, 152, 155, 166, 171, 186, 190, 204, 248, 321, 325, 335, 342, 365, 396, 425, 506, 514, 515, 517, 529, 532, 544, 545, 559, 567, 572, 583, 586, 599, 604, 607, 610, 643, 656 — tota vox major; 5, 9, 17, 18, 75, 104, 109, 111, 150, 151, 160, 168, 169, 177, 189, 191, 193, 198, 199, 206, 210, 211, 213, 228, 237, 244, 249, 251, 253, 260, 271, 277, 278, 279, 282, 283, 284, 286, 292, 295, 300, 341, 342, 345, 346, 349, 368, 380, 383, 386, 395, 400, 411, 413, 436, 437, 440, 448, 471, 476, 480, 485, 486, 516, 522, 525, 545, 561, 563, 564, 569, 570, 573, 574, 576, 581, 587, 588, 593, 594, 595, 598, 601, 608, 612, 661, 663, 674, 680 — בראש literæ ſunt majores; 170 — litera ב major eſt, et ornata; 19, 67, 82, 261, 276, 320, 419, 424, 445, 505, 513, 637, 642, 655 — ב major eſt, et ſola ſtat in prima linea, linea 2da inchoante cum ראשית; 323 —— ſed ב non eſt major; 6, 11, 89, 155, 175, 178, 247, 256, 293, 314, 343, 344, 374, 389, 399, 409, 410, 418, 477, 479, 484, 502, 518, 539, 546, 548, 597, 619, 646, 664 — בראשית ^ 8, 69, 264, 335 — ב ^ 592. ברא אלהים את literæ ſunt majores; 282, 525 — ברא literæ ſunt majores; 346.

2. והארץ — והאר in fine lineæ, 325. תוהו 9, 17, 101, 206, 325, 674, 678. ובוהו 9, 17, 101, 206, 674, 678. וחושך 206, 678. תהם 158, 253. אלהים ^ 665. מרחפ in fin. lin. 325. על 2° ^ 151.

3. Totum comma ^ 294. יהי — יהי* (eraſ. ה) 674.

4. Totum comma, poſt וירא, ^ 193. כי ^ 69. ובין ^ 69. החושך 9, 206, 678 — ח ſup. raſ. 100.

5. אלהים —— אלהי 109. ולחושך 152, 206. ביקר 9. יום אחד — יום א אחד (et habet א lineolam ſuprapoſitam) 681.

6. רקיע בתוך ^ 680.

7. לרקיע - - - ובין ^ 18, 178. אשר 2° ^ 199.

8. ויהי 1° ſup. raſ. הי 294. שיני 109, 155, 260, 325, 674, 680.

9. המים ^ ם 193. ותיראה 157.

11. ויאמר תדשא אלהים תדשא 193. תדשיא 152. עץ — ע primo ל 155 — עץ* (eraſ. ו) 507, 532. פרי 1° — פרי* 136. עושה 109, 206, 244, 674. למינו ^ 135. זרעו — זרעה 191. בו ^ 157. על — לם על 196. ויהי כן bis 136.

12. ותצא 5, 69, 236, 294. דשא — רשא 286. זרע bis 225. ועץ — עץ 152, 674 — ועץ פרי 680. עושה 109, 206, 674. פרי ^ 150. למינהו — למיניהו 193 — על הארץ 206.

13. יום ^ 674.

14. יהי ^ 191. מאורות 152, 236, 335, 440, 485, 581, 664 — מאורת 69, 206, 400, 438, 587, 595, 664; primo 155, 349, 419 — מארות 189, 335, 544. ולהבדיל — להבדל 157 — להבדל 233. לאותות 152, 193, 206, 344, 440, 529, 581, 612 — לאותת 69, 155, 239, 244, 335, 339, 554, 636 — לאתות 233, 345, 389, 561, 674, 681. ולמעדים 69, 109, 111, 236. ולמים 674. ולשנים 650 H.

15, 16, 17 — hæc tria commata facta ſunt per tranſpoſitionem 17, 16, 15; ergo comma 17 locum commatis 15 occupat: 150. למארת 5, 17, 18, 69, 150, 151, 157, 158, 186, 226, 236, 244, 260, 277, 279, 282, 283, 288, 294, 301, 413, 418, 459, 474, 485, 505, 528, 601, 680 — למארות 11, 125, 604, 610, 674 — למאורות 152, 193, 206, 345, 440, 480, 515, 529, 544, 581. ברקיע — ברקי 191. השמים — השמי׳ 233.

16. המאורות 152, 157, 193, 206, 344, 345, 346, 440, 581; primo 236, 244 — המאורת 9, 11, 69, 438, 507, 515, 529, 601; primo 82, 168, 185, 478 — המארות 155, 389, 569, 619, 674, 681. הגדולים 11, 89, 139, 157, 160, 178, 189, 193, 206, 226, 260, 264, 674, 680, 681; primo 155. את 2° bis 150 — ואת 111, 152. המאר 5, 136, 151, 166, 186, 242. הגדול 5, 6, 11, 17, 18, 19, 69, 89, 111, 125, 144, 152, 157, 158, 166, 177,

A

Figure 2: Sample page from the *Vetus Testamentum*, first volume, displaying the collation of the book of Genesis. At the top is the reference text of the Massoretic version (the standardized Hebrew text maintained by Jewish scribes known as the Masoretes, enclosed by the purple box in the image); at the bottom, the two columns apparatus of variants (in the red boxes). For the first five books of the Bible (Pentateuch), Kennicott provides a reference text also for the Samaritan version (the ancient version of the Hebrew Bible used by the Samaritans, in the green box) along with its variants (the blue box in the apparatus). In our project, we are only considering the Massoretic version.

Such systematic approach towards collation is the hallmark of Kennicott's method. In contrast to what De Rossi would later do, Kennicott goes beyond the most conspicuous phenomena of variation, encompassing all potential discrepancies between the reference text and each individual witness, such as spelling, the layout of paratextual elements, and various details of the *mise en page*. This choice, however philologically questionable, actually benefits us by assuring, at least theoretically and net of inconsistencies, errors, and omissions, that we have complete lists of variants at our disposal.

Finally, but most importantly, Kennicott organizes the variants in the apparatus in an extremely precise manner, minimizing the use of natural language and adopting a formalism that anticipates that of most recent editions. On this aspect, which is crucial for automatically extracting information from the critical apparatus, we will dwell at length later on (§ 3.4.1).

The features we have just listed effectively make Kennicott's work not only a rich source of data on the textual tradition of the Hebrew Bible, but also an ideal candidate for our computational treatment.

3. Pipeline

REK's main objectives are threefold:

1. to obtain a TEI-compliant encoding of both the reference text and the critical apparatus of Kennicott;
2. to reconstruct the text of 244 manuscripts fully automatically by way of encoding, for a total of approx. 75,000 pages;
3. to provide an accurate and complete transcription of the text of 10 Kennicott's manuscripts (approx. 7,500 pages) through alignment with these automatically reconstructed texts

To achieve these objectives, we devised the following 4-step pipeline:

1. acquisition of images of Kennicott's *Vetus Testamentum* and of the 10 chosen manuscripts;
2. automatic segmentation and transcription;
3. parsing and encoding of Kennicott's apparatus;
4. reconstruction of the witness texts

We will now discuss each of these points in detail, presenting the work done as well as outlining what is yet to be accomplished. Let us begin with the first step, image acquisition.

3.1. Image acquisition

Digital copies of Kennicott's *Vetus Testamentum* are freely available on the web on platforms such as Archive.org and Google Books, both in .pdf format and in various image formats. We chose the .jp2 images from Archive,[12] which are in an acceptable resolution, and converted

[12] First volume: https://archive.org/details/vetustestamentum01kenn; second volume: https://archive.org/details/vetustestamentum02kenn.

them to .jpeg, which is most widely supported and produces smaller file sizes which still suffice for OCR.

Among the manuscripts collated by Kennicott, we have identified about 20 that are important for their variants. Among these, we have selected 10, based on criteria of convenience such as simple layout, the absence of inline translations into targumic Aramaic, and, of course, the availability of the images (Tab. 1).

Table 1
List of the 10 selected manuscripts, arranged according to Kennicott's catalog number (first column). The 'Script' column refers to the type of writing and may indicate provenance: Sephardi corresponds to Iberian Peninsula and North Africa; Ashkenazi to Central and Eastern Europe; Italian to the Italian Peninsula. The century is represented by Arabic numerals when the exact date is known, otherwise by Roman numerals if dating is indicative. The last column provides the exact number of pages for each manuscript (total: 7,688).

Catalog no.	Script	Century	Library	Shelfmark	Pages
2	Sephardi	1304	Bodleian Library	Ms. Arch. Seld. A. 47	820
3	Sephardi	XV	Bodleian Library	Ms. Poc. 347-348	974
4	Ashkenazi	XIV	Bodleian Library	Ms. Hunt. 12	434
82	Sephardi	1306	Bodleian Library	Ms. Kennicott 2	872
99	Sephardi	1384	British Library	Kings MS 1	869
196	Ashkenazi	XIII-XIV	Ambrosian Library	Ms. E 52 Inf	214
209	Sephardi	1272	National Library of France	Ms. hebr. 26	909
225	Italian	XI-XII	Vatican Library	Ms. Urb ebr. 2	827
227	Italian	1287	Vatican Library	Ms. ebr. 9	836
602	Ashkenazi	XI-XII	State Library of Berlin	Ms. Or. fol. 1213	933

Among the different software mentioned in the Introduction, we have chosen to work with eScriptorium [24, 17, 23] and its OCR/HTR engine, Kraken [16],[13] which is optimized for historical and non-Latin script material.

To upload the images of these manuscripts into eScriptorium, we made use of the IIIF standard: for each chosen manuscript, we retrieved the IIIF manifest and then we used Python scripts to download the images and populate our database.

In the next section, we will discuss the segmentation (§ 3.2) and transcription process (§ 3.3). For the sake of clarity, we will devote separate subsections to segmentation and transcription of the *Vetus Testamentum* (§§ 3.2.1, 3.3.1) and of Kennicott's manuscripts (§§ 3.2.2, 3.3.2), respectively.

3.2. Segmentation

Once we uploaded the images of the *Vetus Testamentum* and the manuscripts onto eScriptorium, we proceeded with segmentation, which is indispensable for identifying those regions on the page where the text to be transcribed is located.

For both segmentation and transcription, we used models in .mlmodel model format trained with Kraken software. These models can be trained with Kraken and then imported into eS-

[13] https://kraken.re/4.0/.

criptorium. Alternatively, as in our case, they can be trained directly within the eScriptorium application.

3.2.1. *Vetus Testamentum*

The layout of the *Vetus Testamentum* is complex, but the segmentation was relatively straightforward. We started by defining a segmentation ontology, distinguishing running headers, titles, left and right main columns, and left and right apparatus for both region types and line types. Following this, we manually segmented approx. 30 pages and trained a model on this sample. With this model, we were able to automatically segment the entire first volume, keeping manual corrections to a bare minimum. Fig. 3 shows an example of segmentation of regions (3a) and lines (3b) taken from the first volume.

As can be seen, eScriptorium provides an intuitive graphical interface that allows the creation of an ontology to distinguish between different types of regions and lines, which are represented by different colors. This feature is extremely useful, as it enabled us to mark only the portions of text for which we wanted to obtain a transcription, namely the reference text (the two regions at the top) and the critical apparatus (the two regions at the bottom), while excluding titles, headers, page numbers, and catchwords.

Similarly, by marking the types of lines, we can express the order of columns and the textual flow. This permitted us to differentiate between lines we need to transcribe and those we do not (e.g., the Samaritan text with its variants, see Fig. 2).

In addition to its user-friendly graphical interface, eScriptorium offers a rich API, which makes it possible to automate numerous segmentation- and transcription-related operations. Using the API functions, we opted to replace the polygonic lineboundaries with parallelogrammatic ones, as they were found to enhance transcription accuracy (Fig. 4).

3.2.2. Kennicott's manuscripts

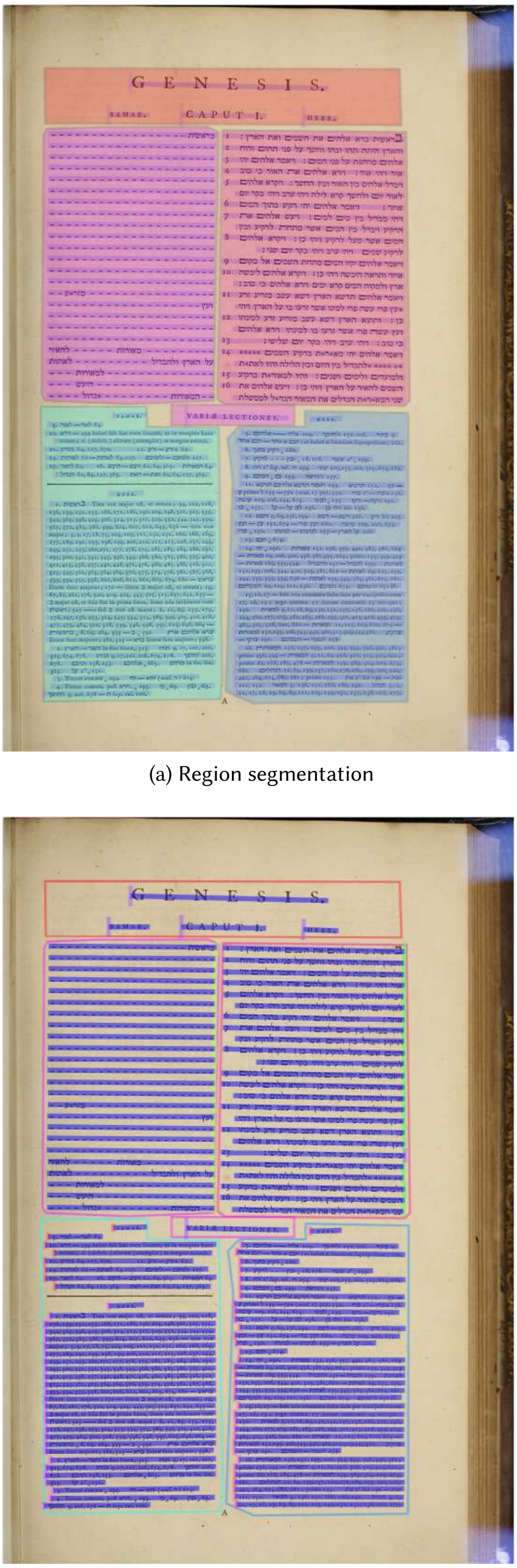
We have applied the same segmentation procedures to the medieval manuscripts. Unlike the *Vetus Testamentum*, which required us to create our own models from scratch, there already exist excellent segmenters as well as recognizers for Hebrew manuscripts, and ongoing research in this area continually improves their accuracy.[14] Only occasionally, for manuscripts with less regular layout, we had to train new models on top of these standard models.

An instance of automatic segmentation for one of the 10 manuscripts in our possession can be seen in Fig. 5.

3.3. Transcription

We proceeded with transcription next. Currently, we have completed the transcription of both the reference text and the critical apparatus of Kennicott's first volume, and we are now in the process of transcribing the manuscript texts.

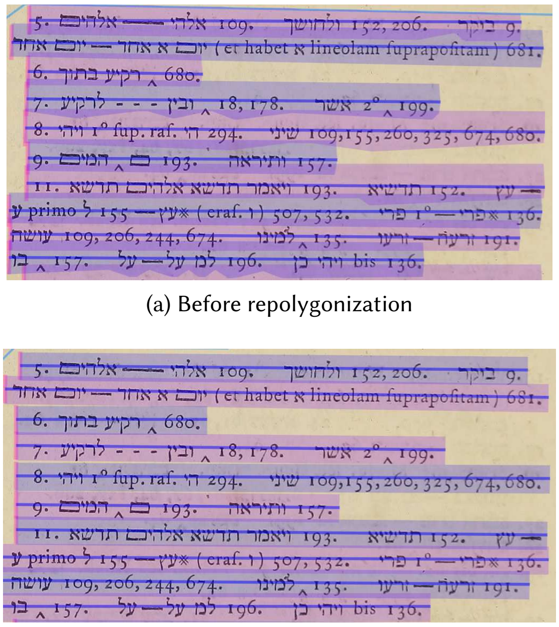
[14] The segmentation models we used are accessible here: https://github.com/dstoekl/sofer_mahir and the recognition models here: https://zenodo.org/record/5167263#.YhzNEtIo-po.



(a) Region segmentation

(b) Line segmentation

Figure 3: Layout analysis from the collation of the book of Genesis.



5. אלהים —— אלהי 109. ולחושך 152, 206. בוקר 9.
יום אחד — יום א אחד (et habet א lineolam ſuprapoſitam) 681.
6. רקיע בתוך ^ 680.
7. לרקיע - - - ובין ^ 18, 178. אשר 2° ^ 199.
8. ויהי 1° ſup. raſ. הי 294. שיני 109,155,260,325,674,680.
9. המים ^ ם 193. ותיראה 157.
11. ויאמר תדשא אלהים תדשא 193. תדשיא 152. עץ —
ע primo ל 155 — עץ* (eraſ. ו) 507, 532. פרי 1° — פרי* 136.
עושה 109, 206, 244, 674. למינו ^ 135. זרעו — זרעה 191.
בו ^ 157. על — על למ 196. ויהי כן bis 136.

(a) Before repolygonization

5. אלהים —— אלהי 109. ולחושך 152, 206. בוקר 9.
יום אחד — יום א אחד (et habet א lineolam ſuprapoſitam) 681.
6. רקיע בתוך ^ 680.
7. לרקיע - - - ובין ^ 18, 178. אשר 2° ^ 199.
8. ויהי 1° ſup. raſ. הי 294. שיני 109,155,260,325,674,680.
9. המים ^ ם 193. ותיראה 157.
11. ויאמר תדשא אלהים תדשא 193. תדשיא 152. עץ —
ע primo ל 155 — עץ* (eraſ. ו) 507, 532. פרי 1° — פרי* 136.
עושה 109, 206, 244, 674. למינו ^ 135. זרעו — זרעה 191.
בו ^ 157. על — על למ 196. ויהי כן bis 136.

(b) After repolygonization

Figure 4: Repolygonization of the critical apparatus, example from Genesis, chapter 1, verses 5-11.

3.3.1. *Vetus Testamentum*

Transcribing the text of the *Vetus Testamentum* posed numerous challenges. The reference text and the critical apparatus follow two distinct textual flows, each with its own peculiarities and complexities, and require different treatments. We opted, therefore, to transcribe them separately.

The main complexity of the critical apparatus lies in the presence of two different alphabets (Hebrew and Latin) with distinct directionality (right-to-left and left-to-right), as well as of punctuation, numbers, and special symbols that require exact reproduction for proper parsing (§ 3.4.1). Dealing with directionality proved particularly demanding, since RTL and LTR markers are invisible and therefore difficult to manage during correction. We successfully overcame this obstacle by employing a visible LTR marker to establish proper word order. After transcribing manually a sample of about 30 pages and training a recognition model on these sample pages, we finally managed to achieve a satisfactory accuracy of approx. 98%.[15] Thanks to the introduction of the LTR marker, the resulting transcriptions became much more manageable to correct.

Transcribing the reference text, on the other hand, proved notably smoother, since it is in a single alphabet, Hebrew,[16] and since it reproduces a standard text, that of the Hebrew Bible, for which excellent transcription models, as we mentioned (§ 3.2.2), already exist. The combination of these features resulted in an accuracy of 98%.

[15] Starting hereon, the accuracy percentages we provide for transcription are based on Character Error Rate (CER) metric, which is the one used by Kraken.

[16] Excluding verse and chapter numbers, which were added in post-processing, see § 3.4.

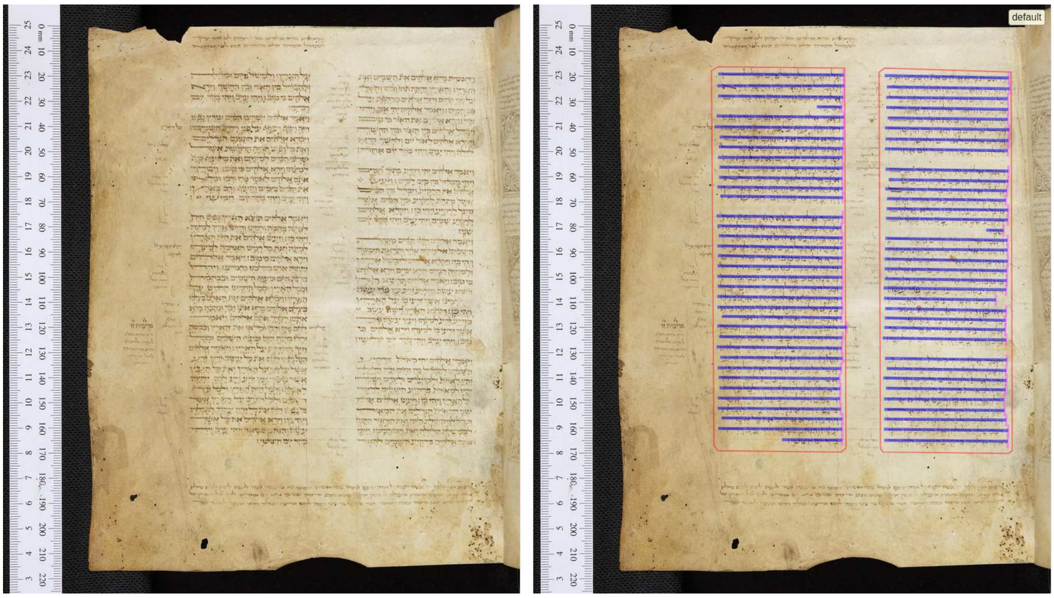
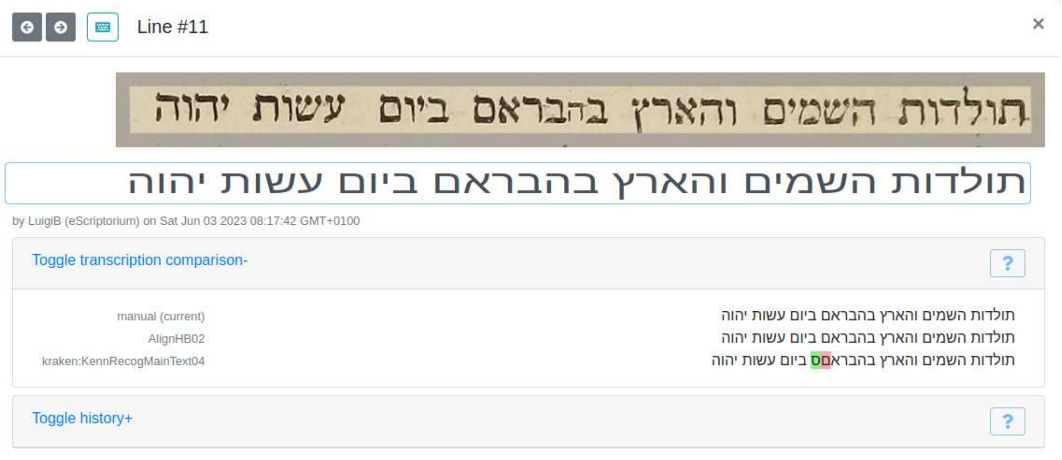


Figure 5: Layout analysis of a page from Kennicott manuscript no. 2. On the left, the original page; on the right, the output of automatic segmentation.



Line #11

תולדות השמים והארץ בהבראם ביום עשות יהוה

by LuigiB (eScriptorium) on Sat Jun 03 2023 08:17:42 GMT+0100

Toggle transcription comparison-

manual (current)	תולדות השמים והארץ בהבראם ביום עשות יהוה
AlignHB02	תולדות השמים והארץ בהבראם ביום עשות יהוה
kraken:KennRecogMainText04	תולדות השמים והארץ בהבראסם ביום עשות יהוה

Toggle history+

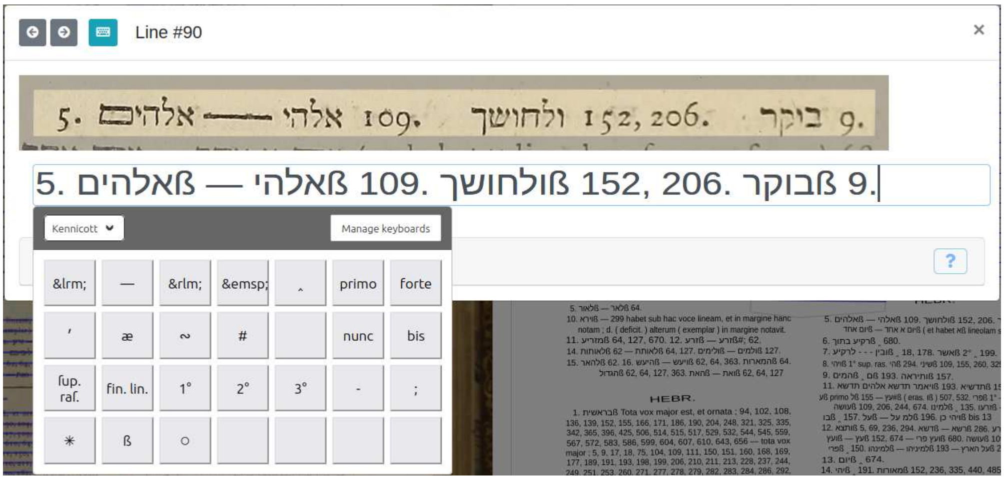
Figure 6: Alignment box. In order of appearance from the top: image of the text to be transcribed (from Genesis 1:5); current transcription (level "manual"); alignment of the current transcription with a standard version of the Bible ("AlignHB02") and with the automatic transcription from the Kraken model ("KennRecogMainText04"). The alignment with the standard version highlights an error made by the model, namely the substitution of ם with the similar letter ס in the word **בהבראם**.

As for the correction of the reference text, we took advantage of the recent integration of passim's[17] text-to-text alignment into eScriptorium, which allows loading an external version of the same text and aligning it with the output of the automatic transcription. This alignment

[17] https://github.com/dasmiq/passim.

significantly expedited the correction process: As depicted in Fig. 6, differences between the aligned versions are highlighted (deletions in red and additions in green), enabling easy identification of errors as well as variants. The exceptional benefits of this tool are evident, and we are confident that it will prove immensely helpful also for the correction of the reconstructed textual witnesses (§ 4).

Before going on to describe the treatment of medieval manuscripts, it is only right to spend a few words about the manual correction process, which is by far the most time-consuming for the human user.



Line #90

5. ßאלהים — ßאלהי 109. ßולחושך 152, 206. ßבוקר 9.

Figure 7: Correction box. At the top, the image from the apparatus of Genesis 1:5 with its corresponding transcription; at the bottom left, the keyboard containing some of the most frequent characters useful for correcting the automatic recognition output.

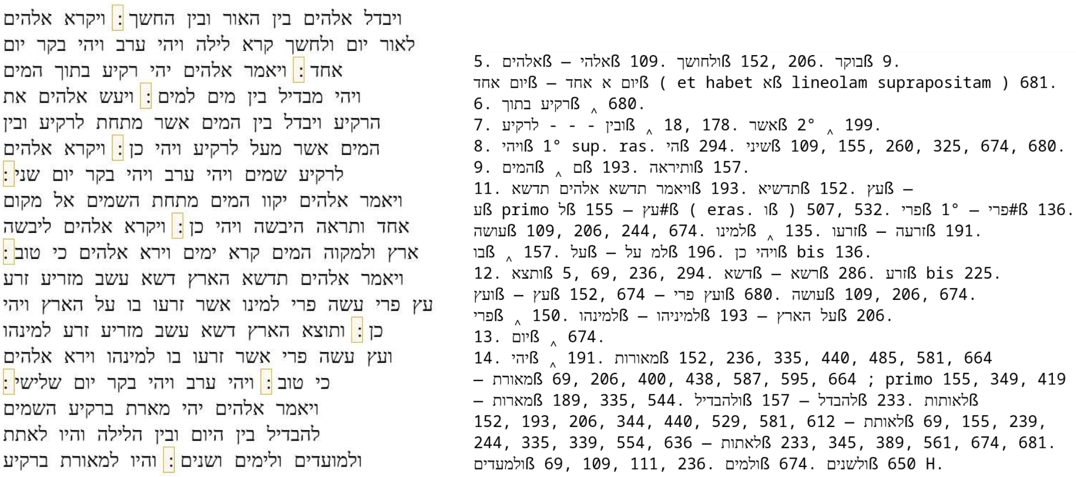
The graphical interface of eScriptorium is designed to make the manual correction process easier: As shown in Fig. 7, eScriptorium enables the user to scroll through the text line by line, with the original image alongside the result of the automatic transcription. Additionally, eScriptorium allows for the creation of customizable keyboards, which can be used to insert characters that are not easily reproducible otherwise. This utility proved exceptionally convenient for correcting the critical apparatus, which, as mentioned, contains many of these special characters.

Once we obtained correct transcriptions for both the reference text and the critical apparatus, we exported them using eScriptorium's API, so as to have pairs of .txt files (text + apparatus) for each treated biblical book (Figs. 8a and 8b).

Finally, we post-processed these files (removing hyphenations, regularizing newlines etc.) to obtain copies suitable for automatic encoding (§ 3.4). Examples of these post-processed texts are visible in Figs. 9a and 9b.

3.3.2. Kennicott's manuscripts

We are presently working in transcribing the 10 Kennicott's manuscripts (Fig. 10), using the models mentioned in Section 3.2.2. When we have their text, we plan to utilize the same align-



ויבדל אלהים בין האור ובין החשך : ויקרא אלהים
לאור יום ולחשך קרא לילה ויהי ערב ויהי בקר יום
אחד : ויאמר אלהים יהי רקיע בתוך המים
ויהי מבדיל בין מים למים : ויעש אלהים את
הרקיע ויבדל בין המים אשר מתחת לרקיע ובין
המים אשר מעל לרקיע ויהי כן : ויקרא אלהים
לרקיע שמים ויהי ערב ויהי בקר יום שני:
ויאמר אלהים יקוו המים מתחת השמים אל מקום
אחד ותראה היבשה ויהי כן : ויקרא אלהים ליבשה
ארץ ולמקוה המים קרא ימים וירא אלהים כי טוב:
ויאמר אלהים תדשא הארץ דשא עשב מזריע זרע
עץ פרי עשה פרי למינו אשר זרעו בו על הארץ ויהי
כן : ותוצא הארץ דשא עשב מזריע זרע למינהו
ועץ עשה פרי אשר זרעו בו למינהו וירא אלהים
כי טוב : ויהי ערב ויהי בקר יום שלישי:
ויאמר אלהים יהי מארת ברקיע השמים
להבדיל בין היום ובין הלילה והיו לאתת
ולמועדים ולימים ושנים : והיו למאורת ברקיע

(a) Reference text

```
5. אלהיםß — אלהיß 109. ולחושךß 152, 206. בוקרß 9.
אחד ßיום — אחד א ßיום ( et habet ßא lineolam suprapositam ) 681.
6. רקיע בתוךß ^ 680.
7. לרקיע - - - ובין ß ^ 18, 178. אשרß 2° ^ 199.
8. ויהיß 1° sup. ras. יהיß 294. שיניß 109, 155, 260, 325, 674, 680.
9. המיםß ^ םß 193. ותיראהß 157.
11. ויאמר תדשא אלהים תדשאß 193. תדשיאß 152. עץß —
עß primo לß 155 — עץ#ß ( eras. וß ) 507, 532. פריß 1° — פרי#ß 136.
עושהß 109, 206, 244, 674. למינוß ^ 135. זרעוß — זרעהß 191.
בוß ^ 157. עלß — למ עלß 196. ויהי כןß bis 136.
12. ותצאß 5, 69, 236, 294. דשאß — רשאß 286. זורעß bis 225.
ועץß — עץß 152, 674 — ועץ פריß 680. עושהß 109, 206, 674.
פריß ^ 150. למינהוß — למיניהוß 193 — על הארץß 206.
13. יוםß ^ 674.
14. יהיß ^ 191. מאורותß 152, 236, 335, 440, 485, 581, 664
— מאורתß 69, 206, 400, 438, 587, 595, 664 ; primo 155, 349, 419
— מארותß 189, 335, 544. ולהבדילß 157 — להבדלß 233. לאותותß
152, 193, 206, 344, 440, 529, 581, 612 — לאותתß 69, 155, 239,
244, 335, 339, 554, 636 — לאתותß 233, 345, 389, 561, 674, 681.
ולמעדיםß 69, 109, 111, 236. ולמיםß 674. ולשניםß 650 H.
```

(b) Critical apparatus

Figure 8: Automatic transcriptions of the reference text and apparatus from Genesis 1:5-14.

ment feature discussed in Section 3.3.1, which was used for transcribing Kennicott's reference text. This will help us locate transcription errors more effectively and speed up correction.

Upon completing the transcription process, we intend to align these texts with the texts automatically reconstructed from Kennicott's data. Subsequent sections will explain the details of this reconstruction.

3.4. XML encoding

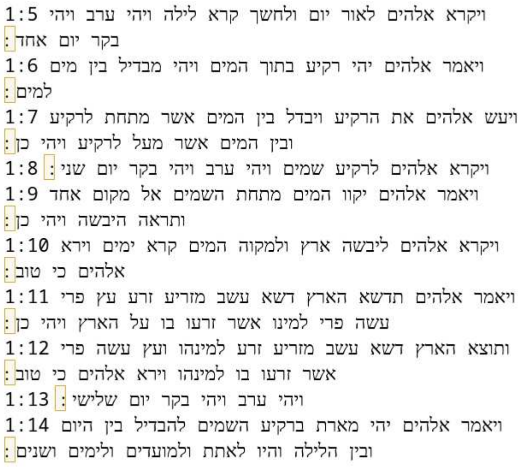
In order to process the data from a scholarly edition, mere machine-readability in the provided .txt files is insufficient. It is crucial for the data to be machine-actionable to enable automated processing and analysis. To achieve this essential feature, we opted for XML encoding, the most widely adopted practice in Digital Scholarly Editing.

As for the encoding of Kennicott's reference text, we used simple Python scripts to first divide the text of each biblical book into its hierarchical units, i.e. chapters, verses, and words. Then, we compared these segmented texts with a standard digital version of the Bible in order to determine the exact number of chapters and verses. An extract of encoded reference text is shown in Fig. 11.

The encoding of the critical apparatus was much more complex. We decided to follow and extend the methodology outlined in Bambaci [2, 1], which involves the development of a rule-based parser for automatic encoding. A detailed account of this methodology can be found there. In the following subsection, we will highlight the key points necessary for readers to understand how we manage to obtain XML files out of Kennicott's critical apparatus.

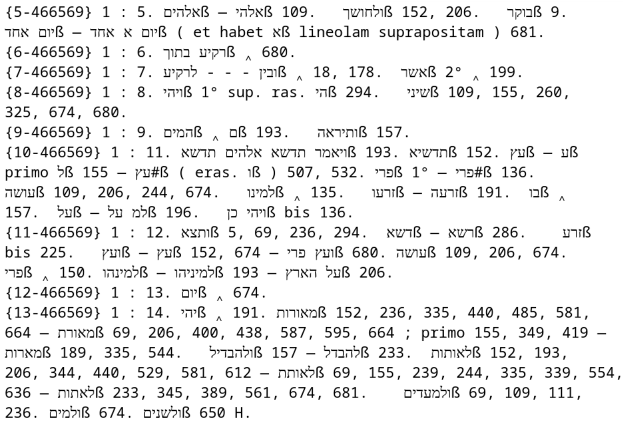
3.4.1. Parsing the critical apparatus

As anticipated in Section 2, Kennicott's critical apparatus proves to be highly suitable for automatic parsing due to its rigorous language and structured presentation of variants. Instead



```
1:5 ויקרא אלהים לאור יום ולחשך קרא לילה ויהי ערב ויהי
בקר יום אחד :
1:6 ויאמר אלהים יהי רקיע בתוך המים ויהי מבדיל בין מים
למים :
1:7 ויעש אלהים את הרקיע ויבדל בין המים אשר מתחת לרקיע
ובין המים אשר מעל לרקיע ויהי כן :
1:8 ויקרא אלהים לרקיע שמים ויהי ערב ויהי בקר יום שני :
1:9 ויאמר אלהים יקוו המים מתחת השמים אל מקום אחד
ותראה היבשה ויהי כן :
1:10 ויקרא אלהים ליבשה ארץ ולמקוה המים קרא ימים וירא
אלהים כי טוב :
1:11 ויאמר אלהים תדשא הארץ דשא עשב מזריע זרע עץ פרי
עשה פרי למינו אשר זרעו בו על הארץ ויהי כן :
1:12 ותוצא הארץ דשא עשב מזריע זרע למינהו ועץ עשה פרי
אשר זרעו בו למינהו וירא אלהים כי טוב :
1:13 ויהי ערב ויהי בקר יום שלישי :
1:14 ויאמר אלהים יהי מארת ברקיע השמים להבדיל בין היום
ובין הלילה והיו לאתת ולמועדים ולימים ושנים :
```

(a) Reference text



```
{5-466569} 1 : 5. אלהיםß – אלהיß 109.   ולחושךß 152, 206.   בוקרß 9.
יום אחדß – יום א אחדß ( et habet אß lineolam suprapositam ) 681.
{6-466569} 1 : 6. רקיע בתוךß ‸ 680.
{7-466569} 1 : 7. ובין - - - לרקיעß ‸ 18, 178.  אשרß 2° ‸ 199.
{8-466569} 1 : 8. ויהיß 1° sup. ras. היß 294.   שיניß 109, 155, 260,
325, 674, 680.
{9-466569} 1 : 9. המיםß ‸ םß 193.   ותיראהß 157.
{10-466569} 1 : 11. ויאמר תדשא אלהים תדשאß 193. תדשיאß 152. עץß – עß
primo לß 155 – עץ#ß ( eras. וß ) 507, 532. פריß 1° – פרי#ß 136.
עושהß 109, 206, 244, 674.   למינוß ‸ 135.   זורעוß – זורעהß 191.  בוß ‸
157.  עלß – למ עלß 196.   ויהי כןß bis 136.
{11-466569} 1 : 12. ותצאß 5, 69, 236, 294.  דשאß – רשאß 286.   זרעß
bis 225.   ועץß – עץß 152, 674 – ועץ פריß 680. עושהß 109, 206, 674.
פריß ‸ 150. למינהוß – למיניהוß 193 – על הארץß 206.
{12-466569} 1 : 13. יוםß ‸ 674.
{13-466569} 1 : 14. יהיß ‸ 191. מאורותß 152, 236, 335, 440, 485, 581,
664 – מאורתß 69, 206, 400, 438, 587, 595, 664 ; primo 155, 349, 419 –
מארותß 189, 335, 544.   ולהבדילß 157 – להבדלß 233.  לאותותß 152, 193,
206, 344, 440, 529, 581, 612 – לאותתß 69, 155, 239, 244, 335, 339, 554,
636 – לאתותß 233, 345, 389, 561, 674, 681.   ולמעדיםß 69, 109, 111,
236. ולמיםß 674. ולשניםß 650 H.
```

(b) Critical apparatus

Figure 9: Normalized text from Genesis 1:5-14, with addition of chapter numbers and relevant metadata. Each line corresponds to a single verse.

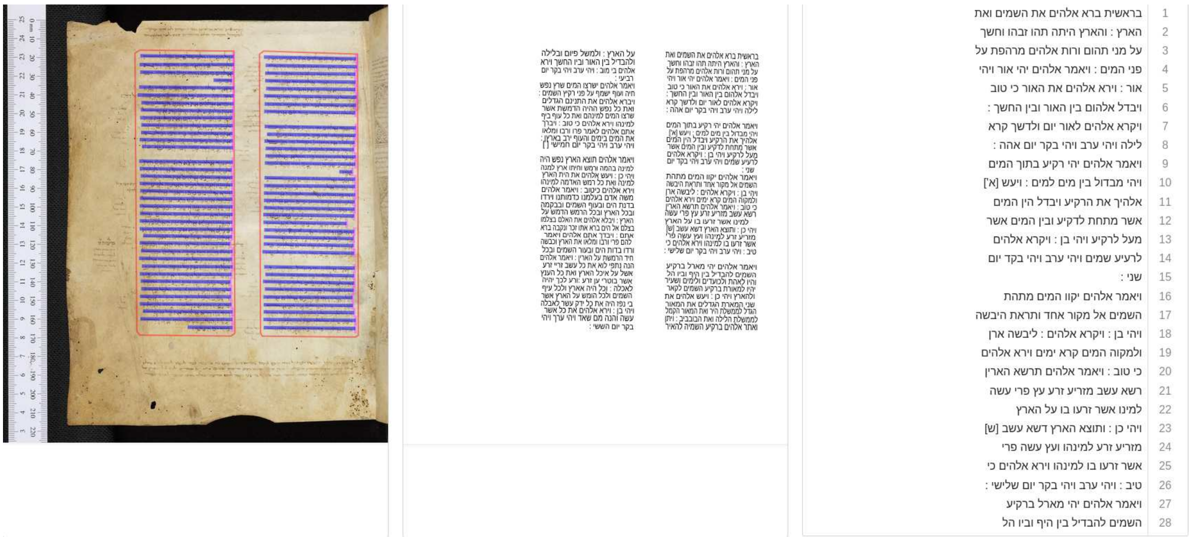
of using Latin commentary-like notes like De Rossi, Kennicott employs a highly formalized language, in which each element performs a precise function according to the position it occupies in the overall structure and according to the class of strings (letters, numbers, symbols) to which it belongs. Both the position and the class of strings can be "captured" by the rules of a Context-Free Grammar (CFG), and these rules "fed" to the parser in order to recognize the function of the individual apparatus components.

Let us consider a fragment of the apparatus as shown in Fig. 12.

For simplicity, let us focus on the first apparatus entry only:

5. **אלהים – אלהי** 109.

which informs us of the substitution of '**אלהים**' with '**אלהי**' in manuscript no. 109. The philologist will immediately recognise the following elements: the place of variation, expressed

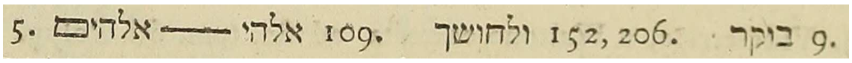


בראשית ברא אלהים את השמים ואת
הארץ : והארץ היתה תהו זבהו וחשך
על מני תהום ורות אלהים מרהפת על
פני המים : ויאמר אלהים יהי אור ויהי
אור : וירא אלהים את האור כי טוב
ויבדל אלהום בין האור ובין החשך :
ויקרא אלהים לאור יום ולדשך קרא
לילה ויהי ערב ויהי בקר יום אהה :
ויאמר אלהים יהי רקיע בתוך המים
ויהי מבדול בין מים למים : ויעש [א']
אלהיך את הרקיע ויבדל הין המים
אשר מתחת לדקיע ובין המים אשר
מעל לרקיע ויהי בן : ויקרא אלהים
לרעיע שמים ויהי ערב ויהי בקד יום
שני :
ויאמר אלהים יקוו המים מתהת
השמים אל מקור אחד ותראת היבשה
ויהי בן : ויקרא אלהים : ליבשה ארן
ולמקוה המים קרא ימים וירא אלהים
כי טוב : ויאמר אלהים תרשא הארין
רשא עשב מזריע זרע עץ פרי עשה
למינו אשר זרעו בו על הארץ
ויהי כן : ותוצא הארץ דשא עשב [ש]
מזריע זרע למינהו ועץ עשה פרי
אשר זרעו בו למינהו וירא אלהים כי
טיב : ויהי ערב ויהי בקר יום שלישי :
ויאמר אלהים יהי מארל ברקיע
השמים להבדיל בין היף וביו הל

Figure 10: Automatic transcription of a page from Kennicott manuscript no. 2. On the left panel, the segmented page; in the center panel, the transcription; on the right panel, the transcribed text.

```
<text>
  <head>Genesis</head>
  <div n="1" type="chap">
    ...
    <ab n="5">
      <w>ויקרא</w>
      <w>אלהים</w>
      <w>לאור</w>
      <w n="1">יום</w>
      <w>ולחשך</w>
      <w>קרא</w>
      <w>לילה</w>
      <w n="1">ויהי</w>
      <w>ערב</w>
      <w n="2">ויהי</w>
      <w>בקר</w>
      <w n="2">יום</w>
      <w>אחד</w>
      <w>:</w>
    </ab>
    <ab n="6">
      <w>ויאמר</w>
      <w>אלהים</w>
      <w>יהי</w>
      <w>רקיע</w>
      <w>בתוך</w>
      <w>המים</w>
      <w>ויהי</w>
      <w>מבדיל</w>
      <w>בין</w>
      <w>מים</w>
      <w>למים</w>
      <w>:</w>
    </ab>
    <ab n="7">
      <w>ויעש</w>
      <w>אלהים</w>
      ...
</text>
```

Figure 11: XML encoding of the reference text (Genesis 1:5-7) with division in chapters (<div>) and verses (<ab>), and tokenization (<w>). Identical words have been assigned their number of occurrence (@n) to enable text reconstruction.



5. אלהים —— אלהי 109. ולחושך 152, 206. ביקר 9.

Figure 12: List of apparatus entries from Genesis 1:5.

by the verse number ('5'), separated by a dot ('.'); the lemma of the reference text, expressed in Hebrew letters ('אלהים') and separated by a horizontal line ('–'); the variant ('אלהי'); the numerical *siglum* of the manuscript ('109'); and finally a dot followed by a long white space (encoded as a tabulation, see below), which closes the apparatus entry.

```
grammar kennicottCFG;
all: app;
app: loc lem var appSep;
loc: verse locSep;
lem: w lemSep;
var: w wit;
verse: NUM;
locSep: DOT;
w: HEBW;
lemSep: DASH;
wit: NUM;
appSep: DOT TAB;
```

```
NUM: [0-9]+;
HEBW: [\u0590-\u05ff]+;
DASH: '-';
DOT: '.';
TAB: '\t';
WHITESPACE : ' ' -> skip;
```

Figure 13: Example of Context-Free Grammar (CFG) for parsing the provided apparatus entry. The left column lists the parsing rules, while the right column contains the tokenization rules.

A CFG as the one shown in Fig. 13 can be formulated[18] in order to describe this apparatus entry and instruct the parser on how to recognize its individual elements correctly.

With the first rule (all) we describe the structure of the entire document, which in our example consists of a single apparatus entry, which we call app. This in turn consists of a variant location (loc), a lemma (lem), a separator for the lemma (lemSep), a variant (var), a witness number (wit), and finally a separator for the apparatus (appSep). A variant location consists in turn of a sequence of numbers (NUM); lemma and variant contain Hebrew words (HEB); separators consist of horizontal bars (DASH), dots (DOT), and tabulations (TAB).

With the first sequence of rules we established, the so-called parsing rules, we are able to define the order of succession of the elements (that is, their syntax), as well as to express their function using "speaking" names that make their meaning explicit for the philologist. With the second sequence of rules, called tokenization rules, we instead indicate the class of strings to which the individual elements belong, such as numerals ([0-9]), alphabetical letters ([\u0590-\u05ff]), punctuation etc.

By employing a CFG akin to the illustrated fragment and using ANTLR4 software [18],[19] we were able to automatically encode the entire critical apparatus of the first volume into XML, with minimum cost and very high accuracy (around 98%).[20]

[18] This CFG is designed just for explanation purposes. The CFG we used to parse Kennicott's apparatus is much more complex and will be published, along with all the relevant material, upon completion of the project (§ 4).

[19] https://www.antlr.org/.

[20] This percentage is indicative and is calculated for the book of Genesis by simply dividing the total number of XML elements correctly assigned by the parser by the total number of XML elements found in this book. To identify errors and derive the correct elements through subtraction, we use the element <lem> (lemma) as the unit of measurement. Here is the calculation: In Genesis, the total number of lemmata amounts to 6,866; out of these, 146 are cases of lemmata erroneously interpreted as readings (<rdg>) due to syntactic ambiguity; the parser correctly identified 6,720 lemmata; the accuracy is therefore equal to $\frac{6866-146}{6866} \times 100 = 97.87\%$. This value

An extract of XML code of the apparatus is shown in Fig. 14.

```
<all>
  ...
  <listApp>
    <loc>
      <chap>
        <num>1</num>
          <chapSep>:</chapSep>
      </chap>
      <verse>
        <singleVerse>
          <num>5</num>
        </singleVerse>
      </verse>
      <closeLoc>
        <dot>.</dot>
      </closeLoc>
    </loc>
      <app>
        <lem>
          <exactLem>
            <originalLem>
              <s>
                <ws>
                  <w>
                    <hebrW>אלהים</hebrW>
                  </w>
                </ws>
                <ltr>ß</ltr>
              </s>
            </originalLem>
          </exactLem>
          ...
    </all>
```

Figure 14: XML encoding from Genesis 1:5. The element names correspond to the parser rules established in our CFG.

Having been encoded, the reference text and the apparatus are now ready for the reconstruction of the witness text, which is our ultimate goal.

4. Text reconstruction

To reconstruct the witnesses, all variants in the apparatus must first be mapped onto the reference text using the lemmata, as it were, as foreign keys. Once the mapping has been performed, our textual reconstruction proceeds simply by replacing, for each manuscript, the lemma in the reference text with the variant in the apparatus.

Table 2
Textual reconstruction for manuscript 109 in the apparatus entry from Genesis 1:5. Lemmata are highlighted in red, variants are highlighted in green.

Main text:	1:5 ויקרא אלהים לאור יום ולחשך קרא לילה...
Apparatus:	5. אלהים – אלהי 109. ולחושך 152, 206. בוקר 9.
Reconstructed text of ms. no. 109:	1:5 ויקרא אלהי לאור יום ולחשך קרא לילה

An example of such a procedure is shown in Tab. 2 (see also Fig. 12), where the lemma

may naturally vary from book to book, but there is no significant reason to expect a substantial change. For instance, the same calculation performed on the book of Exodus yields 99.32%, and on Leviticus it is 98.38%. A comprehensive estimate of the parser's accuracy will only be feasible upon the project's completion.

'אלהים' corresponds to the variant 'אלהי' in manuscript no. 109. Textual reconstruction is straightforward here: using Python, we simply replace the lemma with the variant, as shown. Cases like this, where each apparatus entry corresponds to one and only one lemma, are the easiest to deal with and constitute the majority in Kennicott, accounting for about 70% of the entries.

In the remaining 30% of cases, on the other hand, we do not have any lemma provided, and we need to deduce it from the reference text before we can map the variants. Automatic deduction was possible in all but 3% of the cases.

Let us give one example of the most common case (Tab. 3).

Table 3
Identification of unspecified lemmata using Levenshtein distance. Example from the second apparatus entry from Genesis 1:5.

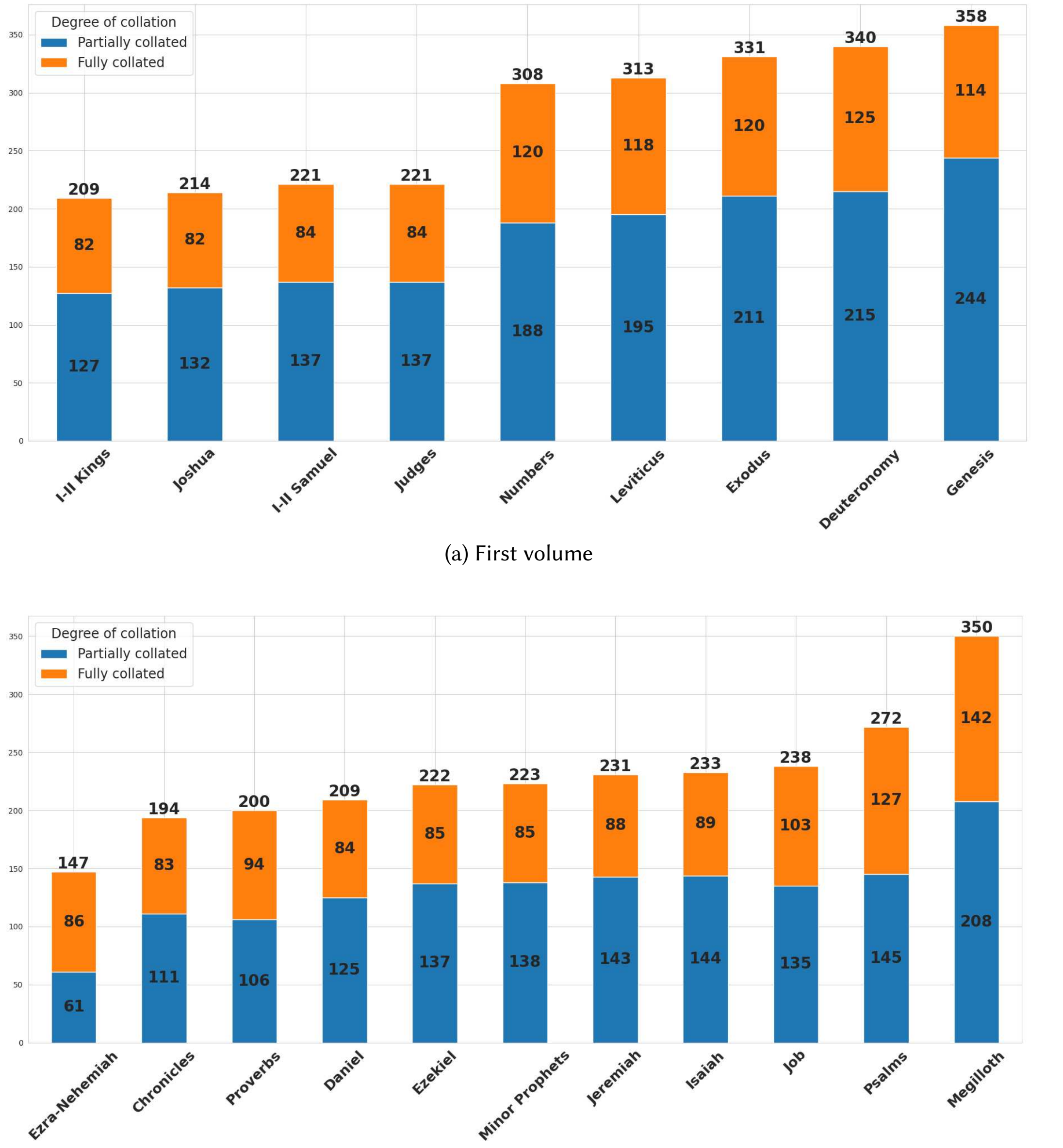
Main text:	1:5 ויקרא אלהים לאור יום ולחשך קרא לילה...
Apparatus:	5. אלהים – אלהי 109. ולחושך 152, 206. בוקר 9.
Reconstructed lemma (*ED* = 1):	ולחשך] ולחושך 152, 206.
Reconstructed text of mss. nos. 152, 206:	1:5 ויקרא אלהים לאור יום ולחושך קרא לילה

In the apparatus, as shown, the variant 'ולחושך' is cited for manuscripts no. 152 and 206, but the lemma of the reference text 'ולחשך' is missing. Due to the very close proximity of the two words (only one character difference), the reference is immediate for the human reader, and for this reason, it is omitted. To make this information available to the machine, we use Levenshtein distance (or edit distance, *ED*), which returns the correct solution in our example ('ולחושך', with *ED* = 1). Such an approach proved to be quite effective for our case study: 60% of all the variants in Kennicott are in fact graphical variants that involve only a few letters.

There are cases, however, where this approach returns multiple outputs with equal distance value, as well as more complex cases where the lemma spans across two or more verses, or where the lemma is not given explicitly in Hebrew, but is rather described by Latin phrases. At the time of writing this article, such residual cases account for about 3% of the total, but we are further improving their automatic treatment.

Using the procedures just described, we have been able to reconstruct the full text of 114 witnesses of the book of Genesis, including 97 manuscripts and 17 printed editions. We are now working on generating transcriptions for the entire Enneateuch (from Genesis to Kings, corresponding to Kennicott's first volume), which means an average of 100 witnesses per biblical book and approx. 35,000 manuscript pages obtained in a fully automatic manner (Fig. 15a and Tab. 4).

Next, we will align these automatically generated text with the automatic transcriptions of Kennicott's manuscripts, and then we will correct them using eScriptorium alignment feature discussed in Section 3.3.1. After correction, we will have approx. 7,500 pages of text, which



(a) First volume

(b) Second volume

Figure 15: Count of witnesses by biblical book and degree of collation in the two volumes of the *Vetus Testamentum*.

will allow us to train new and accurate models for automatic text recognition of Hebrew manuscripts.

Finally, once we have the XML files of all relevant texts (Kennicott's reference text and apparatus, reconstructed witness texts etc.), we will take care to convert our custom XML language to TEI standards using XSLT, in order to ensure data interchangeability and reusability.

We plan to make all the data generated throughout the project, from the HTR models to

XML and text files, publicly available. We envisage publishing all pertinent segmentation and recognition models on Kraken's Zenodo repository.[21] For the HTR and OCR results, we could either publish their different milestone stages in a separate repository on Zenodo with pointers from Biblissima+ (and e.g. HTRunited) or directly on Biblissima+. Moreover, we will post all the relevant material for the project at our GitHub address.[22]

5. Conclusion

We discussed how traditional scholarly editions could offer a viable pathway to improve the performance of current HTR models. Taking the concrete example of the REK project, we illustrated how, through encoding the critical apparatus, we were able to generate complete automatic transcriptions of witness texts, and how we plan to obtain a large amount of training data useful for HTR of biblical Hebrew manuscripts out of these transcriptions.

The accuracy values achieved so far are highly encouraging. All the Kraken models for segmentation and transcription that we used have proven to be exceptionally performant, even in handling highly complex texts such as the critical apparatus: their overall accuracy is never lower than 97%.

The decision to implement a rule-based parser for mining the apparatus has also been fruitful: thanks to it, we have been able to automatically encode a huge amount of data (more than 65,000 apparatus entries) that would have been unthinkable to encode manually, and this with an accuracy of 98%.

The automatic reconstruction of the texts of the witnesses has been equally efficient, and is fully automatable for 97% of the cases. The remaining portion that necessitates manual intervention is still substantial, considering the number and complexity of interventions needed for each biblical book, but we are confident that we can increase the automation of variant mapping (including, for example, the case of lemmata spanning multiple verses), thereby further reducing the need for manual correction.

The data and statistics presented here refer to the first volume of the *Vetus Testamentum*, the processing of which is nearing completion. Our intention is to extend the methodology discussed to the second volume (Fig. 15b), which will allow us to increase the number of reconstructible witnesses up to 244, for a total of approx. 75,000 manuscript pages (Tab. 4).

Moreover, we are intending to incorporate the remaining 10 out of the 20 identified manuscripts mentioned in Section 3.1. This will enable us to double the quantity of pages with highly accurate transcriptions, providing us with an augmented ground truth from which to develop further enhanced HTR models specifically tailored for Hebrew Bible manuscripts.

6. Acknowledgments

We would like to express our sincere thanks to Idan Dershowitz from the University of Potsdam for his invaluable collaboration, as well as to Uriel Aiskovich for his assistance with the

[21] https://zenodo.org/communities/ocr_models?page=1&size=20.

[22] https://github.com/LuigiBambaci/REK.

Table 4
Total number of fully collated manuscripts for each biblical book in the two volumes of the *Vetus Testamentum*, along with an estimate of the number of manuscript pages that it is possible to reconstruct by parsing the critical apparatus.

Book	**No. mss**	**No. pages**
First volume		
Genesis	97	5,820
Exodus	103	5,150
Leviticus	101	3,535
Numbers	103	5,150
Deuteronomy	108	4,644
Joshua	65	1,950
Judges	67	1,943
I-II Samuel	67	4,623
I-II Kings	65	4,745
Second volume		
Isaiah	72	3,528
Jeremiah	71	4,473
Ezekiel	69	3,795
Minor Prophets	69	3,243
Psalms	102	6,426
Job	87	2,262
Proverbs	76	1,748
Megilloth	126	4,032
Daniel	68	1,224
Ezra-Nehemiah	71	2,130
Chronicles	68	5,168
|Total|	**244**	**75,589**

manuscript segmentation task. We also wish to extend our appreciation for the support kindly provided by the National Library of Israel's staff.

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.

References

[1] L. Bambaci. “Critical Apparatus as Domain Specific Languages. A Rule-based Parser for Encoding an Eighteenth-Century Collation of Hebrew Manuscripts”. In: *International Journal of Information Science and Technology* 5.1 (2021), pp. 22–33.

[2] L. Bambaci. “Digitizing Kennicott’s Collation of the Hebrew Bible: Experiences of Encoding and of Computer-Assisted Stemmatic Analysis”. In: *Jewish Studies in the Digital Age*. Ed. by G. Zaagsma, D. Stökl Ben Ezra, R. Miriam, M. Michelle, and L. Amalia S. Studies in Digital History and Hermeneutics 5. De Gruyter, 2022, pp. 299–334. doi: 10.1515/9783110744828-014.

[3] L. Bambaci. “Is a Stemma Possible for the Hebrew Bible? Towards a Genealogy of Medieval Manuscripts Through Phylogenetic Analysis”. In: *Materia Giudaica – Rivista dell’Associazione Italiana per lo Studio del Giudaismo* Xxvi.2 (2021), pp. 3–30.

[4] D. Barthélemy. “Les manuscrits médiévaux et le texte tibérien classique”. In: *Critique textuelle de l’Ancien Testament, 3. Ézéchiel, Daniel et les 12 Prophètes*. Vol. 3. Orbis Biblicus et Orientalis 50. Fribourg/Göttingen: Éditions Universitaires/Vandenhoeck & Ruprecht, 1992, pp. xix–xcvi.

[5] M. Boillet, M.-L. Bonhomme, D. Stutzmann, and C. Kermorvant. “HORAE: An annotated dataset of books of hours”. In: *Proceedings of the 5th International Workshop on Historical Document Imaging and Processing*. 2019, pp. 7–12. doi: 10.1145/3352631.3352633.

[6] P. G. Borbone. “Appendice – La tradizione medievale”. In: *Il libro del profeta Osea – Edizione critica del testo ebraico*. Torino: Zamorani, 1990, pp. 183–227.

[7] C. Clausner, S. Pletschacher, and A. Antonacopoulos. “Aletheia – An Advanced Document Layout and Text Ground-Truthing System for Production Environments”. In: *Proceedings of the 2011 International Conference on Document Analysis and Recognition*. 2011, pp. 48–52. doi: 10.1109/icdar.2011.19.

[8] M. Cohen. “The ‘Masoretic Text’ and the Extent of Its Influence on the Transmission of the Biblical Text in the Middle Ages”. In: *Studies in Bible and Exegesis*. Ed. by U. Simon. Vol. 2. Ramat Gan: Bar Ilan University Press, 1986, pp. 229–256.

[9] G. B. De Rossi. *Scholia critica in V.T. libros, seu supplementa ad varias sacri textus lectiones*. Parma: Ex regio typographeo, 1798.

[10] G. B. De Rossi. *Variae lectiones Veteris Testamenti*. Parmae: Ex regio typographeo, 1784-1788.

[11] K. Elliger and W. Rudorf. *Biblia Hebraica Stuttgartensia*. 5th. Stuttgart: Deutsche Bibelgesellschaft, 1997.

[12] J. S. Penkower. “A Sheet of Parchment from a 10th or 11th Century Torah Scroll: Determining its Type among Four Traditions (Oriental, Sefardi, Ashkenazi, Yemenite)”. In: *Textus* 21.1 (2002), pp. 235–264. doi: 10.1163/2589255x-02101012.

[13] P. Kahle, S. Colutto, G. Hackl, and G. Mühlberger. "Transkribus – A Service Platform for Transcription, Recognition and Retrieval of Historical Documents". In: *1st International Workshop on Open Services and Tools for Document Analysis, 14th IAPR International Conference on Document Analysis and Recognition, OSTICDAR 2017, Kyoto, Japan, November 9-15, 2017*. Vol. 04. 2017, pp. 19–24. DOI: 10.1109/icdar.2017.307.

[14] B. Kennicott. *Vetus Testamentum Hebraicum cum variis lectionibus*. Vol. 1. Oxford: Clarendon, 1776.

[15] B. Kennicott. *Vetus Testamentum Hebraicum cum variis lectionibus*. Vol. 2. Oxford: Clarendon, 1780.

[16] B. Kiessling. "Kraken – An Universal Text Recognizer for the Humanities". In: 2019.

[17] B. Kiessling, R. Tissot, P. Stokes, and D. Stökl Ben Ezra. "eScriptorium: An Open Source Platform for Historical Document Analysis". In: *2019 International Conference on Document Analysis and Recognition Workshops (ICDARW)*. Vol. 2. 2019, pp. 19–24. DOI: 10.1109/icdarw.2019.10032.

[18] T. Parr. *The Definitive ANTLR 4 Reference*. Dallas/Raleigh: Pragmatic Bookshelf, 2012.

[19] J. S. Penkower. "A Tenth-century Pentateuchal MS from Jerusalem (MS C3), Corrected by Mishael ben Uzziel". In: *Tarbiz* 58.1 (1988), pp. 49–74.

[20] A. Schenker, Y. A. P. Goldman, G. J. Norton, A. Kooji Van Der, S. Pisano, J. De Waard, and R. D. Weis, eds. *Biblia Hebraica Quinta. General Introduction and Megilloth*. Stuttgart: Deutsche Bibelgesellschaft, 2004.

[21] L. Schomaker. "Design considerations for a large-scale image-based text search engine in historical manuscript collections". In: *it – Information Technology* 58.2 (2016), pp. 80–88. DOI: 10.1515/itit-2015-0049.

[22] M. Segal. "Methodological Considerations in the Preparation of an Edition of the Hebrew Bible". In: *The Text of the Hebrew Bible and Its Editions*. Leiden, The Netherlands: Interactive Factory, 2017, pp. 34–55.

[23] P. Stokes, B. Kiessling, D. Stökl Ben Ezra, R. Tissot, and E. Gargem. "The eScriptorium VRE for Manuscript Cultures". In: *Ancient Manuscripts and Virtual Research Environments,* Special issue of Classics 18 (2021). Ed. by C. Clivaz and G. V. Allen.

[24] P. A. Stokes, D. Stökl Ben Ezra, B. Kiessling, and R. Tissot. *EScripta: A New Digital Platform for the Study of Historical Texts and Writing*. 2019. DOI: 10.34894/bixswx.

[25] D. Stökl Ben Ezra, B. Brown-DeVost, P. Jablonski, H. Lapin, B. Kiessling, and E. Lolli. "BiblIA – A General Model for Medieval Hebrew Manuscripts and an Open Annotated Dataset". In: *The 6th International Workshop on Historical Document Imaging and Processing*. Hip '21. New York, NY, USA: Association for Computing Machinery, 2021, pp. 61–66. DOI: 10.1145/3476887.3476896.

[26] A. H. Toselli, S. Wu, and D. A. Smith. "Digital Editions as Distant Supervision for Layout Analysis of Printed Books". In: *Document Analysis and Recognition – ICDAR 2021*. 2021, pp. 462–476. DOI: 10.1007/978-3-030-86331-9_30.