Steps Towards Mining Manuscript Images for Untranscribed Texts: A Case Study From the Syriac Collection at the Vatican Library*

Luigi Bambaci[1,2,*,†], George Kiraz[3,4,†], Christine Roughan[5,6,†], Matthieu Freyder[2,7] and Daniel Stökl Ben Ezra[1,2,†]

[1]*Archéologie & Philologie d'Orient et d'Occident UMR 8546, Paris, France*
[2]*École Pratique des Hautes Études, Université Paris Sciences & Lettres, Paris, France*
[3]*Institute of Advanced Studies, Princeton, New Jersey*
[4]*Beth Mardutho: The Syriac Institute, New Jersey*
[5]*Manuscript, Rare Book and Archive Studies, Princeton University, New Jersey*
[6]*Center for Digital Humanities, Princeton University, New Jersey*
[7]*Institut Catholique d'Arts et Métiers, Strasbourg, France*

Abstract

Digital libraries and databases of texts are invaluable resources for researchers, yet their reliance on printed editions can lead to significant gaps and potentially exclude works without printed reproductions. The Simtho database of Syriac serves as a pertinent example: it is derived primarily from OCR of scholarly editions, but how representative are these of the language's extensive literary tradition, transmitted and preserved in manuscript form for centuries? Taking the Simtho database and a selection of the Vatican Library's Syriac manuscript collection as a case study, we propose a pipeline that aligns a corpus of e-texts with a set of digitised manuscript images, in order to ascertain the presence or absence of texts between the e-text and manuscript corpora and thus contribute to their enrichment. We delve into the complexities of this task, evaluating both effective tools for alignment and approaches to detect factors that can contribute to alignment failures. This case study is intended as a first step towards foundational methodologies applicable to larger-scale manuscript processing efforts.

Keywords

automatic text recognition (ATR), layout segmentation, text alignment, Syriac manuscripts

1. Introduction

Digital libraries and databases have become indispensable for researchers working with historical texts across various disciplines. Many such repositories predominantly rely on printed works, which are usually more feasibly made machine-readable through automated text recognition (ATR) technology. Many manuscript materials have become more accessible through

CHR 2024: Computational Humanities Research Conference, December 4–6, 2024, Aarhus, Denmark
*Corresponding author.
†These authors contributed equally.
luigi.bambaci@ephe.psl.eu (L. Bambaci); gkiraz@gorgiaspress.com (G. Kiraz); croughan@princeton.edu (C. Roughan); matthieu.freyder@gmail.com (M. Freyder); daniel.stoekl@ephe.psl.eu (D. Stökl Ben Ezra)
0009-0009-2152-5221 (L. Bambaci); 0000-0003-2338-5081 (G. Kiraz); 0009-0004-5999-8749 (C. Roughan); 0000-0001-5668-493X (D. Stökl Ben Ezra)

digitization efforts as well, albeit often not in the form of a machine-readable text (hereafter: e-text) such as a plain text file – the outputs of library manuscript digitization initiatives are usually digital images.

We propose that aligning these categories of digital materials can support greater enrichment of both e-text and manuscript collections. Identifying which manuscript images overlap with works in a database of e-texts would allow for supplementing the database with information about manuscript witnesses of particular works; conversely, it would facilitate the cataloguing of manuscript contents where such data is not yet available or complete. Meanwhile, finding spans of manuscript folios where their contents do *not* overlap with any part of an e-text corpus can identify texts that are not yet present in that corpus and which might merit future inclusion or scholarly attention – some of these might be unedited works, existing only in manuscript form.

This task is a form of manuscript alignment: mapping a machine-readable text to manuscript images. Much of the previous work on this topic has explored aligning the text to image directly: see a summary overview in [8]. Alternatively, the manuscript images and digital texts might be aligned via an intermediary dataset: (noisy) transcriptions of the image produced by handwritten text recognition (HTR) technology. This is the approach taken by [1, 11, 4] and is the approach explored in the present paper.

These past studies were aimed towards the alignment of a transcription with its manuscript images or towards alignment for the purpose of creating ground truth that could be used for training HTR tools. The present study's goals of detecting matching or missing texts between an e-text corpus and a manuscript image corpus requires such alignments to be leveraged on a larger scale.

A pipeline recently published by Smith, Murel, Allen, and Miller [12], Automatic Collation for Diversifying Corpora (ACDC), similarly addresses text-to-manuscript alignments at scale and demonstrates the feasibility of such an approach. Through the text-reuse software Passim,[1] this pipeline performs extensive alignments between manuscripts and existing digital editions of popular texts, using those results for distantly-supervised training of HTR models.

Building on this groundwork, we propose a pipeline that, like ACDC, exploits the capabilities of Passim to align automated transcriptions and digital editions, but that, unlike ACDC, offers greater flexibility: since our goals do not require the exact line matches essential for HTR ground truth creation, we allow for alignments that are more tolerant of the scribal variants and transcription errors found between manuscripts and print editions.

Given (1) a corpus of digital, machine-readable texts, (2) a set of digitized manuscript images, and (3) segmentation and transcription models for the language in question, our full workflow would proceed as follows:

1. Perform layout analysis with a segmentation model to locate the text on the manuscript image.
2. Run a transcription model to generate automated transcriptions of the manuscript text. These transcriptions do not have to be high quality.[2]

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

[2] Smith, Murel, Allen, and Miller [12] have already demonstrated success in aligning digital editions to automated transcriptions produced by an ATR model that achieved only 60.5% accuracy on a test set.

3. Use alignment software to find likely alignments between the automated transcriptions and each e-text in the given digital corpus.
4. For each page image, select the e-text with the best alignment results. This is necessary because the alignment software might return successful results for multiple works, as in the case of one work using direct quotations from another work.

The results are a set of files, one for each manuscript image, where each of the lines detected in the segmentation phase either contains a successfully-aligned line of text or contains nothing (indicating no alignment).

These alignment results can be further categorized as follows:

- **True positive:** the process returns an alignment where the aligned text matches what is in the manuscript.
- **False positive:** the process returns an alignment where the aligned text does not match what is found in the manuscript.
- **True negative:** the process does not return an alignment, and this is accurate because the text in the manuscript is not represented in the digital corpus.
- **False negative:** the process does not return an alignment, but the manuscript text is indeed represented in the digital corpus.

Minimizing false positives and negatives in such a pipeline is crucial. In our paper, we will specifically focus on the factors leading to such errors, and explore methods for reducing them. To achieve this, we will use a test dataset consisting of a random sampling of manuscript images from the Vatican Library's collection of Syriac manuscripts aligned with a digital corpus, the Simtho database of Syriac texts. This experiment will serve as feasibility test: by evaluating the performance of our pipeline on this subset, we aim to establish robust methodologies that can be scaled in future work to process entire collections of digitized Syriac manuscripts.

The use of true/false positives/negatives here will serve a qualitative evaluation of the challenges posed to such a process: for reasons we will explain in § 5, a full quantitative evaluation of such a binary classification is outside the scope of the present paper.

We will begin our work by providing context on the Syriac materials used for our case study (§ 2). Next, we will detail the data preparation process, including image selection (§ 3.1), automatic transcription (§ 3.2), and the preprocessing of Syriac e-texts from Simtho (§ 3.3). Finally, we will thoroughly describe our pipeline and the alignment experiment (§ 4), evaluate the results (§ 5), and offer concluding remarks on our findings and future work (§ 6).

The materials used in our study – segmentation data, automatic transcriptions, links to images from the Vatican Library, and the alignment outputs – are available in our Zenodo repository.[3]

[3] https://doi.org/10.5281/zenodo.13941501.

2. Syriac Manuscripts, Syriac Data

2.1. The Syriac language and script

Syriac emerged in the area of the city of Edessa (today Şanlıurfa in Turkey), the capital of the kingdom of Osrhoene (in today's South-East Turkey and Northern Syria) as a local daughter script of Imperial Aramaic after the demise of the Persian Empire. Its earliest attestations are inscriptions from the turn of the era (6 CE), while the oldest manuscripts in Syriac are documents from the third century.

Classical Syriac became the common cultural and liturgical language of Eastern Christianity and was used for literary texts in a vast area from Edessa to China and India.

Some of the local Aramaic dialects remain in use today, but with the rise of Islam and the increase of the usage of Arabic, Syriac's use as a spoken language decreased and gradually became restricted mainly to liturgy and literature. In some contexts, Arabic-language texts were written using the Syriac script – these are known as Garshuni texts. The Syriac script was also used for other languages such as Armenian, Malayalam, and Ottoman Turkish [10].

Luckily, huge amounts of Syriac and Garshuni manuscripts survived the vicissitudes of history, wars and persecutions, so that scholars need a repertoire of libraries and manuscript catalogues to navigate the collections [3].

The Syriac writing system is an abjad, written in a connected cursive from right to left. Vowels and other features can optionally be written with diacritical marks. Syriac makes use of a number of scripts which differ in how letterforms are written, the most important of which are the Estrangela, Serto and Eastern scripts.

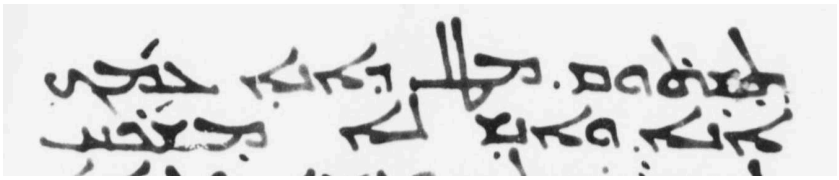
2.2. The Simtho database of Syriac

Over the course of several years, a team of students, postdocs, and heritage speakers at Beth Mardutho (The Syriac Institute) have built a 16-million token corpus of Syriac texts, now hosted at simtho.bethmardutho.org. The source for this data is printed editions digitized via OCR. Beth Mardutho teams have been involved in developing high-accuracy OCR models for printed Syriac; where these models do not achieve perfect results, the MelthoLab team (young women and men from Syriac heritage communities, mostly in the Middle East) have manually corrected the output.

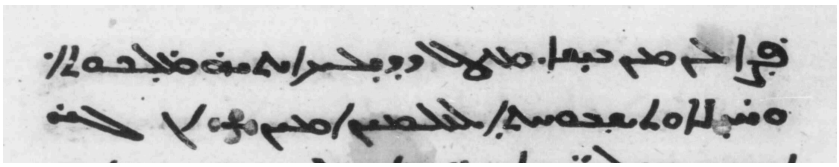
The texts currently included in Simtho encompass 1,214 works, almost all printed scholarly editions, coming to a total of 19,978,900 tokens. The team is additionally OCRing heritage-community-produced publications (both liturgical and literary texts). When complete, the Simtho database will include the vast majority of printed texts in Syriac.

2.3. The Syriac manuscript collection at the Vatican library

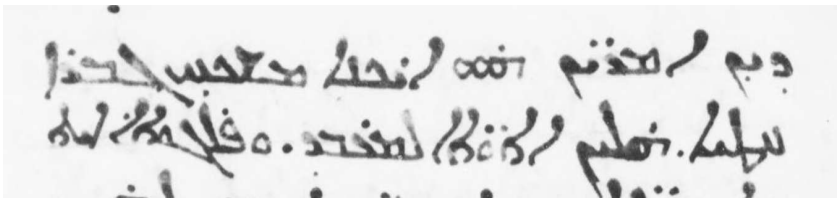
With ca. 850 items, the Vatican library (Biblioteca Apostolica Vaticana) contains one of the largest collections of Syriac manuscripts in the world. Much of the Vatican collection was acquired by Elia Assemani and Joseph Simon Assemani in the 18th century. On one of these trips, the boat of Elia Assemani capsized on the Nile and the manuscripts fell into the water



(a) Estrangela script (MS Sinai Syr. 15)



(b) Serto script (MS Paris syr. 56)



(c) Eastern script (MS Sinai Syr. 1)

Figure 1: Sample selections demonstrating the forms of the Estrangela, Serto, and Eastern scripts. [Public domain: Library of Congress Collection of Manuscripts in St. Catherine's Monastery, Mt. Sinai and gallica.bnf.fr / Bibliothèque nationale de France.]

[6]. Some manuscripts could be salvaged but were badly damaged resulting in darkened pages or ink washed away.

Digitization campaigns started in 2000 with a project lead by Brigham Young University [13, 7]. Most of the manuscripts are in the main collection of the Vatican Library, but some are in the Borgiana collection. Of the former, 371 manuscripts have been digitized so far; of the latter, 45 manuscripts. 144 of the main Vatican collection and 20 of the Borgiana collection are in high quality color images, the rest are older and lower quality greyscale scans. All are accessible via a powerful IIIF-server.

3. Preparation of the Data

For our experiment, we needed to prepare three types of data. The first of these was a small-scale slice of manuscript images representing the Vatican Syriac manuscript collection (§ 3.1). From these, we then generated automated transcriptions using ATR tools (§ 3.2). Third, it was necessary to preprocess the texts in the Simtho database so that they could serve as an effective source for alignment with the automated transcriptions (§ 3.3).

3.1. Image selection

The present experiments have been run on selections from the full-color digitizations in the Digital Vatican Library (so digitized microfilms have not yet been evaluated). Because our segmentation models perform best on folio layouts with one or two columns of text, we further narrowed our selection to only manuscripts with such layouts, which we identified manually. This resulted in a corpus of 121 manuscripts (69 single-column, 52 double-column).

For each manuscript in this collection, we randomly sampled ten images to be used in the following experiment. We then discarded covers, empty pages, and calibration images (color checkers, rulers), leaving us with our final test dataset of 1,115 images. The manual classification into single, double, empty and untreatable images was quickly done on a Windows 11 system by sliding images into subfolders. For future projects, we plan to train a layout classifier based on this data.[4]

3.2. Generating automated transcriptions

For the segmentation and recognition of the sample images we used the open-source ATR engine Kraken (version 5) [9].[5] We chose Kraken for its strengths in right-to-left scripts and the ease of integrating its outputs with Passim, discussed below (§ 4).

The segmentation and transcription models previously produced by a transcribathon workshop organized by Christine Roughan, Daniel Stökl Ben Ezra and George Kiraz from 25-28 March 2024 at Princeton University. These models and further details on their production will be published in a separate publication. Here it shall suffice to detail that the transcribathon produced training data for manuscript images from three collections: the Biblioteca Apostolica Vaticana (Vatican City, Vatican State),[6] the Bibliothèque Nationale de France (Paris, France), and the Sinai collection whose digitized microfilms are housed at the Library of Congress (Washington DC, USA). Participants corrected automatic text-to-text alignments of transcriptions that had been generated by earlier base-models. Among the models trained on this data are dedicated segmentation models for single- and double-column layouts as well as one generalized transcription model.

The transcription model had been trained on 108 sample folios from 38 manuscripts ranging in date from the 6th to the 20th centuries. The distribution of dates was not precisely even: the final training data included no manuscripts from the 18th or 19th century, and manuscripts from the 13th century were most frequent. Approximately 60% of the transcription ground truth were images of Serto scripts, 30% were Estrangela scripts, and 10% were Eastern scripts. This HTR model, therefore, is a generalized one that does not specialize in one particular manuscript but may perform better on Serto and Estrangela scripts. It demonstrated 97.4% accuracy on the test data during training.

[4] We envision a classifier such as the one developed by Gogawale, Bambaci, Kurar-Barakat, Vasyutinsky-Shapira, Stökl Ben Ezra, and Dershowitz [5] for the layout classification of Hebrew prints, see: https://github.com/TAU-CH/midrash_layout_classification_using_multilabel_vgg.

[5] kraken: https://github.com/mittagessen/kraken/.

[6] There is an overlap of 17 images between the training data that produced the transcription model used here and the manuscript images used to test alignment in the current experiment.

Using kraken to apply these segmentation and recognition models to the experiment's images produced transcriptions in ALTO XML totalling 215,859 tokens.

3.3. Preprocessing Simtho

Vocalization and diacritics may vary widely from manuscript to manuscript and are usually absent in the printed editions such as the ones we used for alignment. To aid the alignment process, we therefore stripped from the Simtho texts most such characters, i.e. diacritics, except for the combining diaeresis (U+0308) and the combining dot above (U+307); all punctuation, except the end paragraph sign (U+700), the period, and the semicolon; special Garshuni characters and any letters that do not fall within the main Syriac Unicode block (U+0710 to U+072C).

4. The Alignment Experiment

Once we obtained the ALTO files and the preprocessed texts to compare from Simtho, we proceeded to the alignment of the two datasets.

As anticipated in the introduction, we chose Passim as a key tool in our alignment pipeline for its ability to process large corpora of text efficiently.[7] While Passim excels in detecting instances of text reuse across multiple sources, it is not specifically designed for the task of text alignment, which requires matching specifically-sized sections of text to each other. In our case, it is necessary to align subsections of the longer text-reuse results on the level of the line in the manuscript. In addition, it is essential for us to obtain specific and detailed reports on the alignment results, both at the line level and the page level, for use in evaluation, statistical analysis, and more.

Recognizing these specific needs, we turned to a pipeline designed to complement Passim by providing robust and accurate text alignment functionality. This pipeline, developed by Matthieu Freyder,[8] performs the following essential tasks:

1. **Text Preparation:** It uses the set of texts from the input corpus (here, Simtho) as the source texts for alignment. It processes those e-texts as well as the automatic transcriptions from the ALTOs generated by Kraken to prepare them for ingestion into Passim.
2. **Text-reuse Data Production:** It runs Passim to detect instances of text reuse between the prepared e-texts and the automated transcriptions.
3. **Alignment Data Extraction:** It extracts text-reuse data generated by Passim and compares it with each line of the transcriptions. This comparison involves filtering based on a user-defined threshold of Levenshtein percentage to identify the best alignment candidates. These are reinserted back into the ALTOs, while unfit candidates are removed.
4. **Metrics Generation:** Finally, the pipeline generates dedicated metrics in the form of TSV files to enable evaluation and perform targeted analyses.

[7] See [2] and kitab-project.org for some use cases with extensive corpora. Further bibliography on Passim can be found there.

[8] TABA: https://github.com/Freymat/from_eScriptorium_to_Passim_and_back.

The first preparatory phase involves choosing the units of texts to compare. Passim offers flexibility in structuring and specifying the length of input data, from entire books down to finer units, making it suitable for our corpora. For our experiment, we used regions – i.e. the text zones on the page identified by our segmentation models – as our minimal text units for the automatic transcriptions. For the source e-texts from Simtho, we used plain text documents, with one TXT file per edition. Phase two then executes the main processing task by running Passim on the textual corpora (see Listing 1 in the appendix, § 7, for the arguments passed to Passim in our experiment).

The core component of our pipeline revolves around the third phase, where we aim to transform text-reuse findings into precise text alignments. Here, we use the Levenshtein percentage metric, meaning that we accept as potential alignments any Passim suggestions that overlap with the automatic transcription by a given Levenshtein percentage threshold. After carefully evaluating three of them — 70%, 80%, and 90% — we ultimately selected the lowest (70%), so as to manage variations and possible noise within the e-texts as well as the automatic transcriptions, while still ensuring pertinent alignment results.

The TSVs generated by the last phase of our pipeline provide insights into the quantity and quality of identified alignments. We will give a demonstration of that in Section 5.

5. Evaluation

Let us turn our attention now to evaluating the alignment results. We will start with a general overview (§ 5.1), and then move on to examine successful and unsuccessful alignments (§§ 5.2, 5.3).

5.1. Overview

These results come from the TSV reports generated at the end of our pipeline. A portion of one such TSV is shown in Table 1.

As can be seen, all alignments found for any of the e-texts are shown – thus, folio 1v of MS Vat. sir. 1 appears in nine rows, with nine e-texts as potential alignment candidates for that image. This does not mean that that one folio contains excerpts from nine different texts; rather, the entries with lower alignment ratios are likely to be false positives.

This report facilitates scaling up from evaluations of alignments on the manuscript line level to that of alignments on the folio level. Each folio image will contain a certain number of lines, and we can measure for how many lines per image our pipeline has found a particular alignment candidate. Total alignment success on the image level means an alignment candidate was found covering 100% of the lines in an image; total alignment failure means none were found covering any lines.

In the results for our test dataset, the pipeline found 100% alignment for 21 images; conversely, it found 0% alignment for 268 images. Table 2 gives an overview of how many images saw between 0-25%, 25-50%, 50-75%, or 75-100% successfully aligned lines. Although total folio alignments were a small proportion of the results, for approximately half of the test dataset we found an e-text from the Simtho database that aligned to at least 50% of the image's lines.

Table 1
Excerpt of the TSV report on text alignment across multiple manuscript pages. Each row corresponds to an alignment result for a specific manuscript page identified by its ALTO filename (column "Filename"), along with key metrics such as the number of OCR-generated lines in that page ("ocr_lines"), the total number of lines successfully aligned with a given e-text ("alg_lines"), and the corresponding alignment ratio expressed as a percentage ("alg_ratio"). The "max_cluster" column indicates the largest cluster of aligned consecutive text lines found within each text region, reflecting areas of significant agreement between OCR outputs and the aligned e-text. Finally, The "e-text_id" column indicates the e-text detected by the alignment process.

Filename	ocr_lines	alg_lines	alg_ratio	max_cluster	e-text_id
Vat.sir.1_0006_fa_0001v.xml	34	17	50	11	P_Gen [AB]
Vat.sir.1_0006_fa_0001v.xml	34	13	38.2	10	SyrHexapl
Vat.sir.1_0006_fa_0001v.xml	34	7	20.6	6	ChronZuq
Vat.sir.1_0006_fa_0001v.xml	34	3	8.8	3	Basil_SyrHex
Vat.sir.1_0006_fa_0001v.xml	34	1	2.9	1	GabrShahrz_MartMtBer'ain
Vat.sir.1_0006_fa_0001v.xml	34	1	2.9	1	JnPenk_BkMainPts1
Vat.sir.1_0006_fa_0001v.xml	34	1	2.9	1	ExampFath
Vat.sir.1_0006_fa_0001v.xml	34	1	2.9	1	JacSer_MemNativ2
Vat.sir.1_0006_fa_0001v.xml	34	2	5.9	1	ComGenExod
Vat.sir.1_0007_fa_0002r.xml	32	29	90.6	21	P_Gen [AB]
Vat.sir.1_0007_fa_0002r.xml	32	9	28.1	8	SyrHexapl
Vat.sir.1_0010_fa_0003v.xml	32	32	100	32	P_Gen [AB]
Vat.sir.1_0010_fa_0003v.xml	32	31	96.9	27	BarBahl_SyrLex
Vat.sir.1_0010_fa_0003v.xml	32	28	87.5	16	BarEbr_LampSanc
Vat.sir.1_0010_fa_0003v.xml	32	13	40.6	7	Eph_ComGen&Exd

Table 2
An overview of alignment results across all folio images. The alignment success group indicates the percentage of lines on that image for which an alignment was found (filtering for only the top alignment text ("alg_ratio" column in Table 1) in cases where alignments were found to multiple e-texts).

alignment success group	image count	percent of dataset
75-100%	477	42.8%
50-75%	84	7.5%
25-50%	114	10.2%
0-25%	440	39.5%

Of the 21 images for which the pipeline returned total alignment success, many of the aligned texts were Biblical ones. So for example fol. 25v of MS Vat. sir. 1 fully aligned with the book of Genesis from the Syriac version of the Old Testament (the Peshiṭta, "P_Gen[AB]"). Visual inspection confirmed the accuracy of this alignment. 100% alignment success presents an obviously strong case for an accurate text identification, but partial alignment successes offer text identification candidates as well.

The reasons for total alignment failure were meanwhile more varied. Alignment failure in general will be discussed in further detail in § 5.3 – here, manual inspection of the 268 images with zero alignment results found reasons ranging from physical damage to segmentation failures to Garshuni text contents. This inspection did additionally identify three Syriac texts in

the manuscript images of the test dataset that are not yet incorporated into the Simtho database: the Ecclesiastical History of Socrates Scholasticus, in MS Vat. sir. 145; the Ecclesiastical History of Theodoret of Cyrrhus, also in MS Vat. sir. 145; and the Syriac Grammar of Bar Hebraeus, in MS Vat. sir. 193.

We can also visualize the data for alignment success ratios per image (Figure 2). Across the approximately 10 random images selected for each manuscript in the test dataset, we can see that for some manuscripts alignment was highly successful: MS Vat. sir. 471 tops the list with nine of its folios achieving 100% alignment success and the tenth achieving 96.9%. This is a manuscript of the Peshiṭta New Testament, and all of the alignment results are relevant Biblical books.

For other manuscripts the pipeline returned negative results for attempted alignments. For instance, MS Vat. sir. 424, parts 1 and 2, achieved 0% alignment across all folio images. Inspection confirmed that the text contents were Garshuni, not Syriac.

The following subsections will discuss the successful and failed alignments in more detail, categorizing different types and discussing methods that address these categories. For various reasons, a full quantitative evaluation of the specific identifications of the texts is beyond the scope of this present article.[9]

5.2. Successful alignments

Alignments can be found by the pipeline for several reasons. The aligned text can either be:

1. A version of the automatically transcribed text, e.g. one of the manuscripts used for the edition with more or less variant readings, resolved abbreviations, etc.
2. A different (but close enough) recension, e.g., there are different translations of the Biblical books from Greek, Hebrew and Aramaic into Syriac.
3. A quotation or parallel, especially in the case of long biblical quotations, florilegia, dictionaries, biographies, histories, or in anthologies such as exegetical Catenae that are compilations of quotations from previously existing commentaries or homilies.
4. Not a true match, but perhaps a line that happens to share enough formulae between the e-text and automatic transcription to pass the Levenshtein percentage threshold.

Case [4] is a false positive and one we want to filter out. Similarly, case [3] does not serve the goals of either automatic cataloguing or automatic detection of out-of-corpus texts, and so should be filtered out as well. (Such alignments could, of course, serve HTR ground truth creation or text reuse analyses, but these are outside the goals of the present experiment.)

Fortunately, the majority of such cases can be filtered out simply by selecting only the alignment result with the highest alignment ratio for that folio. So, for Table 1's example of MS

[9]For some texts Simtho does not have any ground truth to be aligned with, for others the Simtho e-text covers only a part of the composition. Many liturgical, legal and exegetical manuscripts present anthological material that contain quotations or parallels of varying length; here a binary evaluation is not easily decided and probably should be avoided altogether. Some e-texts in Simtho are themselves anthologies and so the issue exists both ways. Text reuse in Syriac literature in general is quite extensive. A different challenge is posed by the limitations of catalogue data: for at least 40 pages from 8 manuscripts (e.g. Vat. sir. 107, 122, 318, 352, 529, 560 pt. 1, 567, 623) we do not have scholarly reports of their contents. For at least half of them, the pipeline seems to make good or excellent suggestions, but this requires more research. We will take up these challenges elsewhere.

Vat. sir. 1 folio 1v, we end up with the alignment for the book of Genesis from the Peshiṭta ("P_Gen[AB]") and discard the remainder.

This works when we do have a likely true positive alignment result for that image, but for images with less likely results we require other methods. This is where we leverage the "max_cluster" column, which expresses the total number of *consecutive* aligned lines between the automatic transcriptions and the digital edition: case [4] usually results in only one aligned line, and short quotations from case [3] will lead to only short consecutive spans of aligned lines. We have therefore found it effective to keep only alignment results with at least three consecutive aligned lines.[10]

To return to our categories of successful alignments: we meanwhile treat cases [1] and [2] as true positives, using the alignment ratio and the total of consecutive aligned lines as key indicators. In the first case, the identification can be a direct candidate for automatic cataloguing. In the second case, the identification might be an indirect candidate for automatic cataloguing, identifying the text but not the specific recension.[11]

5.3. Failed alignments

Alignment failure can have very different reasons.

1. Poor preservation of the surface (e.g. damage through ink corrosion, fragmentary state).
2. Poor preservation of the writing traces (e.g. water damage, smear or shine-through).
3. Imaging problem.
4. Poor layout segmentation leading to a bad representation of the overall text in the recognition (missing lines or parts of lines, missing columns or parts of columns, or erroneously joining separate lines).
5. Poor text recognition.
6. No matching text in the textual database.

Among these cases, only case [6] is a true negative, indicating that the manuscript did not contain any texts from the corpus of e-texts used for alignment. In the remainder of cases, the pipeline may be returning false negatives. Whether or not the manuscript contains a work from the e-text corpus cannot be automatically determined so long as these other factors are impeding potential alignments. Further elaboration of the six cases and methods for addressing them follows below.

Cases 1-2: poor preservation of the writing or page surface. These cases lead to poor performance further along our workflow: fragmentary or water-damaged folios lead to poor segmentation results (since our models expect undamaged columns of text); poor visibility of the text leads to poor automated transcription results. There is unfortunately little to be done

[10] Handling quotations or parallels of longer length is a challenge in the present version of the experiment, but can be addressed in future work by running full manuscripts through the pipeline rather than isolated folios. This would allow for the use of a similar "max_cluster" metric, albeit on the folio-level rather than the line-level – such a metric could help to disambiguate when a manuscript contains a text in full or only a (long) quotation of it.

[11] There are some cases where the two highest ranked alignment suggestions have equal evaluation scores – in such cases it requires closer research to determine which one is the better identification.

at this point in time for images of a manuscript whose text has been washed away because it fell into the Nile. Therefore, for cases [1] and [2] we primarily seek to flag these problem images so that they can be filtered out.[12]

Poor segmentation results can be identified through the segmentation data in the ALTO XML files. The segmentation models used in this experiment identify sets of lines contained in regions (columns). A simple evaluation of the line segmentation quality for a given column is the quotient of the mean line length and the column's average width. A low value implies broken or too short lines.

For poor transcription results, one possibility for flagging problem cases can be found in the reported confidence of the recognition network used to produce the HTR transcriptions. For 142 images, the average recognition confidence for that page is below 96.25%. In most cases (123 images, i.e. 86.6%), this indicated a problem like water damage, smeared ink, gold ink on dark background (Vat. sir. 622), complex page layout (e.g. a table), fragmentary preservation, or shine-through. Vice versa, for the 973 cases with a recognition confidence higher than 96.25%, visual inspection revealed that only 197 (20.3%) presented a problem such as water damage, low contrast or fragmentary state.

We can also get an indication of poor transcription results by checking the ratio of the recognized characters in a line to the segmented line length. We have found low values tend to correspond to water damage, smeared ink, or gold ink on dark folios.

Case 3: imaging problems. In the present experiment, we avoided imaging problems by selecting test images only from the high-quality color digitizations in the Vatican Library. In future work, microfilm or other greyscale / black and white images in a massive corpus of digitized manuscripts could be easily flagged by checking the color mode of each image or a random sampling of a small number of pixels. Further automatic evaluation of imaging problems requires further research.

Cases 4-5: poor recognition performance by the segmentation or transcription models. We have discussed methods for detecting poor segmentation or transcription above. Where cases [1-3] are ruled out as potential reasons for this performance, we are likely dealing instead with a manuscript for which our models perform poorly.

These segmentation and transcription models will of course achieve variable results on new manuscripts. The Passim parameters used here are selected to be tolerant for noisy HTR results, but poor enough performance will still lead to failed alignments. One option for remedying poor HTR results is using a tool like ACDC to refine models so that they may better handle the manuscripts with subpar results.

Case 6: no aligned text found. This itself could be caused by one of several reasons, depending on whether the image contains:

a. no text,

[12] Such damage, however, does not always lead to failed alignments: visual inspection showed that the current segmentation and transcription models in fact dealt with some of these difficult cases remarkably well.

b. text not written in the Syriac script,
c. Syriac script used to write another language (i.e., Garshuni, see § 2.1), or
d. a Syriac text not present in the e-texts used for alignment.

Our experiment has already manually excluded case [6a] from the images in our dataset. Such filtering could also be achieved automatically by flagging images with no (or only minor) segmentation results.

In our dataset, we identified examples of case [6b] as occurring in Vat. sir. 23 (left columns are Arabic), Vat. sir. 20.2 (fol. 164r has Arabic in the same column as Syriac), and Vat. sir. 51.1 (fol. 7r is in Latin). To distinguish case [6b] from case [6d], we can take advantage again of the reported confidence of the recognition network since the images that contain non-Syriac scripts have a lower reported confidence. Unlike cases [1] and [2], however, case [6b] does not result in lines with unusually few characters, and so when we have poor reported confidence but normal line character lengths, this can indicate the presence of a non-Syriac script.

Our experiment additionally seeks to flag case [6c] because the Simtho database only contains Syriac texts, not Garshuni ones. At this point, we did not have the capacities to train a dedicated BERT model and therefore relied on simpler NLP methods, i.e. bigram distribution. We calculated the 200 most common bigrams in Simtho to represent all styles and periods of Syriac in general. Then we calculated the distribution of these 200 bigrams for each automatically recognized page and measured the Euclidean distance to the Simtho bigram vector. We then used the automatic transcriptions of these Garshuni pages to calculate a bigram-vector for these 200 most common bigrams also for Garshuni. There were 82 images in our dataset which contained Garshuni text; leveraging the calculated Euclidean distances allowed us to successfully flag 76 of them.

When the above methods rule cases [6a, b, c] to be unlikely, then we are dealing with case [6d], and so have successfully located a manuscript Syriac text not present in the input e-text corpus (here, the Simtho database of Syriac).

6. Conclusion

The experiments of this paper have highlighted the variety of challenges that need to be surmounted when attempting corpus-level automated indexing or detection of absent texts through text alignment methods. Through the features discovered in the evaluation of the test dataset that lead to false positives or false negatives, we have been able to strategize and test methods for mitigating these challenges.

Overall, this small experiment on the randomized set of folios sampled from the Vatican Library collection of digitized Syriac manuscripts has shown that the proposed workflow has promise. The pipeline achieved great success in aligning e-texts to noisy HTR in the test dataset, and the results can be used to determine candidates for text identifications of the contents of those images. Additionally, even in this small slice of the Vatican Syriac manuscript collection we were able to identify three Syriac texts which were absent from the Simtho database. With procedures put in place to handle the challenges described in § 5.3, such a workflow could be put to the fuller corpus. To leverage this workflow on a full-sized dataset, however, we will want to optimize our pipeline to function in a more time-efficient manner.

The case study detailed in the present paper has covered Syriac corpora, but the broad strokes of the workflow are not language-specific. Given suitable e-text repositories, digitized manuscript collections, and preliminary ATR models, this workflow is broadly applicable to a wide range of language traditions.

Acknowledgments

Funded by the European Union (ERC, MiDRASH, Project No. 101071829). Views and opinions expressed are however those of the author(s) only and do not necessarily reflect those of the European Union or the European Research Council Executive Agency. Neither the European Union nor the granting authority can be held responsible for them.

References

[1] E. Chammas, C. Mokbel, and L. Likforman-Sulem. "Handwriting Recognition of Historical Documents with Few Labeled Data". In: *International Workshop on Document Analysis Systems (DAS)*. 2018, pp. 43–48. URL: 10.1109/das.2018.15.

[2] R. Cordell and D. Smith. *Viral Texts: Mapping Networks of Reprinting in 19th-Century Newspapers and Magazines*. 2022.

[3] A. Desreuxmaux and F. Briquel-Chatonnet. *Répertoire des Bibliothèques et des Catalogues de Manuscrits Syriaques*. Paris: Editions du Centre National de la Recherche Scientifique, 1991.

[4] D. S. B. Ezra, B. Brown-DeVost, N. Dershowitz, A. Pechorin, and B. Kiessling. "Transcription Alignment for Highly Fragmentary Historical Manuscripts: The Dead Sea Scrolls". In: *17th International Conference on Frontiers in Handwriting Recognition (ICFHR), Dortmund, Germany, 2020*. 2020, pp. 361–366. URL: 10.1109/icfhr2020.2020.00072.

[5] S. Gogawale, L. Bambaci, B. Kurar-Barakat, D. Vasyutinsky-Shapira, D. Stökl Ben Ezra, and N. Dershowitz. "NetLay: Layout Classification Dataset for Enhancing Layout Analysis". In: *Magazén: International Journal for Digital and Public Humanities* (2024).

[6] C. Griffin. "Syriac Manuscripts from the Egyptian Desert". In: *The Newsletter of the Neal A. Maxwell Institute for Religious Scholarship* 31.1 (2011). URL: https://scholarsarchive.byu.edu/insights/vol31/iss1/4.

[7] K. Heal. "Vatican Syriac Manuscripts: Volume 1." In: *Hugoye: Journal of Syriac Studies* 8.1 (2018), pp. 115–122. URL: https://hugoye.bethmardutho.org/article/hv8n1prheal2.

[8] M. Ibn Khedher, H. Jmila, and M. A. El-Yacoubi. "Automatic Processing of Historical Arabic Documents: A Comprehensive Survey". In: *Pattern Recognition* 100 (2020), p. 107144. DOI: 10.1016/j.patcog.2019.107144. URL: https://www.sciencedirect.com/science/article/pii/S0031320319304455.

[9] B. Kiessling. "Kraken: A Universal Text Recognizer for the Humanities". In: *Digital Humanities (DH)* (2019).

[10] G. Kiraz. “Garshunography: Terminology and Some Formal Properties of Writing One Language in the Script of Another”. In: *Scripts beyond Borders: A Survey of Allographic Traditions in the Euro-Mediterranean World*. Ed. by T. P. Johannes den Heijer Andrea Schmidt. Peeters, 2014, pp. 65–74.

[11] T. de Reuse and I. Fujinaga. “Robust Transcript Alignment on Medieval Chant Manuscripts”. In: *Proceedings of the 2nd International Workshop on Reading Music Systems, Delft, the Netherlands, November 2, 2019*. 2019, pp. 21–26.

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

[13] E. J. Wilson. “The Digitizing of Selected Syriac MSS in the Vatican Apostolic Library”. In: *Hugoye: Journal of Syriac Studies* 3.2 (2000 [2010]), pp. 282–285. URL: https://hugoye.bethmardutho.org/article/hv3n2crdigitizingsyrmss.

7. Appendices

We ran our pipeline on HPC cluster equipped with 128 CPUs (see Table 3 for the full hardware specifications). We measured the execution time as the experiment proceeded through each step. As can be seen in Table 4, the optimizations of Passim were clear: detecting text reuse between a 16-million token corpus (Simtho) and a 200-thousand token corpus (automatic transcriptions) took only three and a half minutes on our hardware. The following steps, on the other hand, were more costly, requiring 25 minutes for our current small-scale dataset. Since in future work we will want to apply this pipeline to a much larger dataset, we will be exploring optimizations to improve the execution time of these steps.

Listing 1
List of Passim arguments used. An complete explanation of these arguments can be found in Passim documentation (see note 1).

```
--master local[125]
--executor-memory 200G
--driver-memory 30G
seriatim
--docwise
--floating-ngrams
--fields ref
--filterpairs 'ref = 1 AND ref2 = 0'
--all-pairs
--complete-lines
-n 7
/TABA/data/processed/json_for_passim/passim_input.json
/TABA/data/processed/passim_output
```

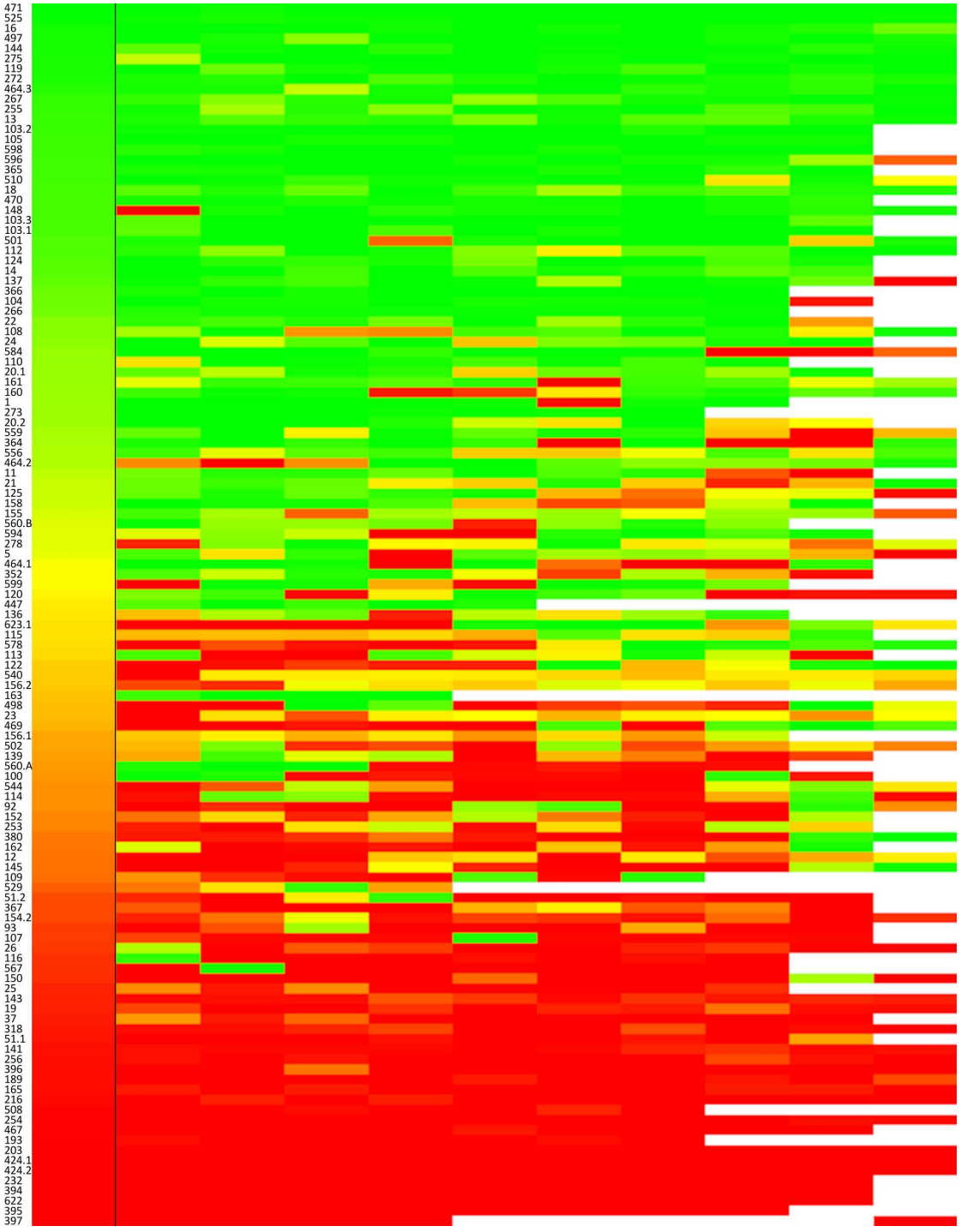


Figure 2: Alignment ratios by manuscript image.
Rows = manuscripts, with numbers on the left to indicate the shelfmark, e.g. 103.3 = Vat. sir. 103.pt.3.
Leftmost column = average alignment success for the test folios in the manuscript.
Subsequent columns = alignment ratios for each folio image. Green = Alignment successful. Red = Alignment unsuccessful. Yellow and orange in between. White = no image in test dataset.

Table 3
List of hardware specifications from the HPC cluster.

Attribute	Details
Architecture	x86_64
CPU(s)	128
Cores per socket	32
Socket(s)	2
Model name	Intel(R) Xeon(R) Gold 6338
CPU MHz	2.00 GHz
L3 cache	96 MiB
Memory	251 G

Table 4
Execution time breakdown of pipeline steps

Step	task	execution time
1	Text preparation	0:00:08.827338
2	Text-reuse data production	0:03:27.621587
3	Alignment data extraction	0:25:01.316454
4	Metrics generation	0:00:03.381940