

TOWARDS SYRIAC DIGITAL CORPORA: EVALUATION OF TESSERACT 4.0 FOR SYRIAC OCR

EMILY CHESLEY

PRINCETON THEOLOGICAL SEMINARY

JILLIAN MARCANTONIO

DUKE UNIVERSITY

ABIGAIL PEARSON

UNIVERSITY OF EXETER

ABSTRACT

This paper summarizes the results of an extensive test of Tesseract 4.0, an open-source Optical Character Recognition (OCR) engine with Syriac capabilities, and ascertains the current state of Syriac OCR technology. Three popular print types (S14, W64, and E22) representing the Syriac type styles Estrangela, Serto, and East Syriac were OCRed using Tesseract's two different OCR modes (Syriac Language and Syriac Script). Handwritten manuscripts were also preliminarily tested for OCR. The tests confirm that Tesseract 4.0 may be relied upon for printed Estrangela texts but should be used with caution and human revision for Serto and East Syriac printed texts. Consonantal accuracy lies around 99% for Estrangela, between

89% and 94% for Serto, and around 89% for East Syriac. Scholars may use Tesseract to OCR Estrangela texts with a high degree of confidence, but further training of the engine will be required before Serto and East Syriac texts can be smoothly OCRed. In all type styles, human revision of the OCRed text is recommended when scholars desire an exact, error-free corpus.

1. INTRODUCTION¹

Digital humanities scholars are constantly searching for automatable processes that will free up time for deeper analysis. One such process, optical character recognition (OCR), has dramatically expanded scholars' ability to search, cross-reference, and analyze large corpora. But while OCR capabilities for Western languages have been in advanced stages for some time, Syriac OCR remains in development, and Syriac scholars have not yet analyzed the accuracy of available OCR engines. To rectify said lacuna, a team of analysts carried out extensive tests on Tesseract 4.0, an open-source OCR engine with Syriac capabilities. This paper summarizes the test results in order to provide data-based recommendations for current Syriac OCR possibilities.

This project evaluated three popular print types (S14, W64, and E22) representing the Syriac type styles Estrangela, Serto, and East Syriac and tested a representative sample of each type style against both modes available for Syriac in Tesseract, Syriac Language and Syriac Script.² After briefly summarizing

¹ This research was conducted under digital humanities fellowships offered in 2018 at Beth Mardutho: The Syriac Institute. The authors wish to thank Dr. George A. Kiraz for his guidance, refining thought, and bestowal of the research opportunity. They also wish to thank the three reviewers for their incisive feedback on a previous draft.

² A brief note on terminology used throughout the paper is in order: the term "font" denotes a computer font such as the Meltho fonts. The term "print type" is used to denote the physical type that corresponds to its code in J. F. Coakley's *Typography of Syriac*. For instance, the Meltho font Serto Jerusalem is based on the print type W11B.

the history of Syriac OCR ventures, this paper details the testing methodology, presents the detailed data results, and finally highlights the most significant errors and trends across type styles. The analysts also briefly tested six handwritten manuscript pages for comparison with the printed pages; the results of these tests follow those for the printed texts.

The project's extensive tests of Tesseract 4.0 confirmed that this OCR engine may be relied upon for printed Estrangela texts but should be used with caution and human revision for Serto and East Syriac printed texts. Consonantal accuracy lies around 99% for Estrangela, between 89% and 94% for Serto, and around 89% for East Syriac. When diacritics, punctuation, and non-Syriac characters are taken into consideration, accuracy rates drop dramatically: around 95% for Estrangela, approximately 86% for Serto, and around 77% for East Syriac.³ Scholars may use Tesseract to OCR Estrangela texts with a high degree of confidence, but further training of the engine will be required before Serto and East Syriac texts may be easily OCRed. In the latter two type styles, Tesseract may still be helpful for the first step of a project, but scholars will have to carefully review and edit all texts to ensure precision and readability. Serto's OCR accuracy would likely be improved by training Tesseract on a font based on the W64 print type. Handwritten manuscripts are recognized with even less accuracy and demand extensive digital editing, so Tesseract is not recommended for use on manuscripts at this time.

In order to comprehend the tests that were conducted, it is essential to clarify at the outset a technical distinction within Tesseract's computation and the terms used to describe it. Tesseract can be invoked in two Syriac modes: Language and Script. The former makes use of language-specific training data

When the Tesseract OCR engine is invoked, it takes the "-l" (for language) command-line argument. The language value typically matches the ISO 639-2 3-letter code (e.g., "syr" for Syriac) but is in reality the name given to the trained data in Tesseract's "tessdata" subfolder.

³ As will later be explained, every type style was OCRed with two modes, hence the results averaged here. Precise calculations follow.

for OCR. The user can use this mode to recognize texts in English, German, or French, for example. The latter is script wide (e.g., the Latin script) and is not sensitive to a specific language. Having said that, according to the Tesseract user documentation, when one invokes it in Script mode, English is always added to the mix.⁴ While the Tesseract engine remains the same, its two modes OCR with slightly different tools and thus produce differing results. The tests accounted for the mode variable, and results will be reported in both modes. Since the term “script” can also refer in Syriac studies to the three Syriac scripts (Estrangela, Serto and East Syriac), the analysts have opted in this paper to use the phrase “type style” to refer to the three Syriac scripts and the capitalized word “Script” to refer to the Tesseract mode. In order to simplify terms, this paper refers to the six variable combinations of type style and mode in hyphenated form, e.g., as Estrangela-Language.

This lengthy paper contains overall results, results split by type style, analysis of the significant error trends, manuscripts results, practical tips, and concluding remarks. Reading the work in full will give the best understanding of the project and its conclusions; however, readers can still gain valuable information from focusing on certain parts of the paper. For those readers who prefer data to prose, tables and graphs are included throughout the paper and especially in Appendix 2. For scholars who have a specific Syriac text or project in mind, this paper highlights the results by type style (Estrangela, Serto, and East Syriac) and analyzes the most significant accuracy difficulties in each; these results will help those scholars understand if and how Tesseract can best assist them in their task. For those readers who are interested in handwritten texts, there is a section on manuscript results. The methodology of the project as well as the tips for usage will provide some

⁴ “TESSERACT (1) Manual Page,” GitHub, last updated July 2, 2018, <https://github.com/tesseract-ocr/tesseract/blob/master/doc/tesseract.1.asc#languages>.

practical advice for OCRing Syriac, though there are practical insights throughout the paper as well.

2. HISTORY OF SYRIAC OCR

Syriac optical character recognition (OCR) has been sought since the early 1990s. OCR is the electronic conversion of typeset or handwritten text images into machine-encoded texts; the process turns an unsearchable image of a text into a searchable text file. As one can imagine, effective OCR drastically speeds up the process of searching and analyzing large sets of text. Syriac OCR will thus allow for the creation of virtual text corpora on par with, for example, the *Thesaurus Linguae Graecae*.

The primary hindrance to creating an OCR engine for Syriac has always been its cursive nature of writing, since the script prevents precise letter differentiation.⁵ Previous approaches have centered on resolving the character connectivity problem.⁶ Syriac OCR was first tackled, to the best of the team's knowledge, by William Clocksin in the early 1990s, then a faculty member at the Computer Lab of the University of Cambridge. In collaboration with students, Clocksin developed an OCR system for hand-written Estrangela and reported high success rates of up to 100% accuracy.⁷ Separately, Elizabeth Tse and Josef Bigun from Halmstad University in Sweden developed a different system for Serto and reported around 90% accuracy.⁸ However, both

⁵ This same concern has applied to Arabic OCR.

⁶ Maxim Romanov, Matthew Thomas Miller, Sarah Bowen Savant, and Benjamin Kiessling, "Important New Developments in Arabographic Optical Character Recognition (OCR)," March 2017. <https://arxiv.org/abs/1703.09550>

⁷ William F. Clocksin and Prem Fernando, "Towards Automatic Transcription of Estrangelo Script," *Hugoye: Journal of Syriac Studies* 6, no. 2 (2003): 249–68. Clocksin has returned to the Syriac OCR problem in recent months, and his new OCR software which is in development should be evaluated in the future.

⁸ Elizabeth Tse and Josef Bigun, "A base-line character recognition for Syriac-Aramaic," 2007 *IEEE International Conference on Systems, Man and*

systems remained in trial state, and neither gained widespread use.

In 2017, Grigory Kessel noticed that Syriac PDFs uploaded to Google Drive became searchable files. Kessel correctly assumed that there must be an OCR engine running in the background and notified the Syriac studies community through a Syriac studies listserv.⁹ Indeed, Google currently manages an open-source OCR engine named Tesseract that was originally developed by Hewlett-Packard.¹⁰ Tesseract can be trained to OCR a particular language by feeding it images of that language and their corresponding transcriptions. With this training method, it learns how to recognize letters from this training data and then becomes able to recognize images of other texts in that language it has not seen before. Tesseract 4.0 already supports Syriac in all its type styles (i.e., Estrangela, Serto, and East Syriac).

The team's preliminary question, before analyzing Tesseract's output, was: how was Tesseract trained and on what sort of data? In the absence of any documentation from Tesseract's programmers, the team considered two options: either a large set of texts was transcribed and fed into Tesseract with its associated text images or an automated method was used to produce text images and their corresponding texts.¹¹ The former option seemed unlikely since the Syriac world is a small one; if the hundreds of thousands of lines that are required for thorough training had been transcribed from books by a Syriac scholar, the Syriac community would

Cybernetics (Montreal, Quebec: IEEE, 2007), 1048-1055. doi: 10.1109/ICSMC.2007.4414012.

<https://ieeexplore.ieee.org/document/4414012/authors>

⁹ Grigory Kessel, April 21, 2017, message to hugoye-list, <https://groups.yahoo.com/neo/groups/hugoye-list/conversations/messages/8069>.

¹⁰ For more information, see "Tesseract OCR," *Google Open Source*, <https://opensource.google.com/projects/tesseract>, accessed 30 July, 2018.

¹¹ "TrainingTesseract 4.00," GitHub, last revised 15 July 2018, <https://github.com/tesseract-ocr/tesseract/wiki/TrainingTesseract-4.00>, accessed 30 July, 2018.

presumably have heard about this endeavor. The team then began to investigate whether another mechanical method could have been used for training. Inverting the process, one can use a font to generate a huge amount of texts, say in a PDF format, and then convert the PDFs into images of individual pages. Since the Syriac content and the images that correspond to them are both known, the compiled data can be used to train Tesseract on which shapes represent which characters. Searching internet archives unearthed web pages which indicated that Beth Mardutho's Meltho fonts may have been used for just such a task.¹² To verify this assumption, the analysts produced a PDF from the non-standard, calligraphic Meltho font Estrangelo TurAbdin, which was based on the calligraphy of Isa Benjamin and does not appear in any print edition. Estrangelo TurAbdin is an elaborately swirled and decorated font, too unique to be recognized without intentional training, yet Tesseract had no problems recognizing the characters produced by it. This demonstrates that Tesseract was indeed trained with Beth Mardutho's Meltho fonts and not by manual transcriptions of texts.

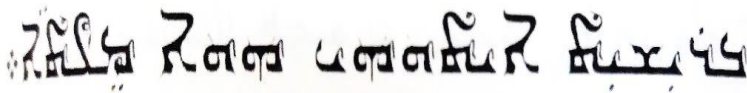


Figure 1: Example of Estrangelo TurAbdin Font¹³

Fortunately, the standard non-calligraphic Meltho fonts are based on actual print types that were used in the vast majority of Syriac publications from the late-nineteenth century onward. This suggests that Tesseract should be able to accurately recognize printed editions that are typeset in those print types that were the basis of Meltho fonts. The non-

¹² Tesseract language-specific training guide, GitHub, last revised 19 July 2018, <https://github.com/tesseract-ocr/tesseract/blob/master/src/training/language-specific.sh>.

¹³ *The Patriarchal Journal of the Syrian Orthodox Patriarchate of Antioch and All the East* 40, nos. 211-212-213 (2002), 94.

calligraphic Meltho fonts with their corresponding print types are:

1. **Estrangelo Talada.** The original type was designed by William Morley for the London-based typefoundry of William Watts in 1855, based on a 7th century manuscript from the British Library, *Add. 14640*. It was used by the University of Cambridge Press, among other presses. Books published in this type include Burkitt's *Evangelion Da-Mepharreshe* (Cambridge, 1904). This corresponds to print type Coakley S14.¹⁴
2. **Estrangelo Nisibin.** The type was designed in 1886, probably by Yūsuf Dīyār Bakrī, and was used in the Dominican Press in Mosul. This corresponds to Coakley S17.¹⁵
3. **Estrangelo Edessa.** The font is based on types provided by an Ohioan press and was probably designed after the 1954 Estrangelo Monotype font. The Monotype font was designed with the assistance of R. Draguet, and in turn is based on an 1851 type used for the Estrangelo Talada font.
4. **Serto Jerusalem.** This is the oldest existing type still in popular use in fonts. The original design dates back to a type associated with the famous diplomat and printer of Arabic, Savary de Brèves, around 1612. After changing hands over the centuries, the type was acquired by the Imprimerie Catholique in Beirut from the Imprimerie Nationale in Paris sometime in the second half of the 19th century. It was also acquired by the Syriac Orthodox press at St. Mark's Monastery in Jerusalem, from which the font was designed. This corresponds to Coakley W11B.¹⁶

¹⁴ J. F. Coakley, *The Typography of Syriac: A historical catalogue of printing types, 1537-1958* (New Castle, DE, and London: Oak Knoll Press and The British Library, 2006), 172–76.

¹⁵ Coakley, *Typography of Syriac*, 178–79.

¹⁶ Coakley, *Typography of Syriac*, 50–56.

5. **Serto Kharput.** The type was originally designed by the Swedish designer O. Tullberg and G. H. Bernstein for the Teubner type foundry and was first seen in 1853. Bernstein's edition of the Harklean version of St. John's Gospel was published with this type (Leipzig, 1854). This was one of the first Serto types to contain a sizeable amount of ligatures. This corresponds to Coakley W49.¹⁷
6. **Serto Batnan.** This type was designed by the London-based typefoundry of William Watts for Oxford University Press around 1864. It was used in printing Payne Smith's *Thesaurus Syriacus* (Oxford, 1866) and P. E. Pusey and G. H. Gwilliam's edition of the Syriac New Testament (Oxford, 1901) which is still reproduced by the United Bible Society. This corresponds to Coakley W52.¹⁸
7. **Serto Malankara.** The original type goes back to 1870 and was designed by Saint Thomas Press at Cochin. Lee's edition of the Old Testament is still reprinted from this type by the United Bible Society. Until recently, the print type was still used by the Syriac Orthodox and the Syro-Malabar presses of India. This corresponds to Coakley W54, which itself was derived from W36.¹⁹
8. **Serto Mardin.** The original type design was made in 1888 by W. Drugulin of Leipzig and overseen by Paul de Lagarde. A number of books were printed in this type including Bar Ebroyo's *Ṣemḥe* (1922). This corresponds to Coakley W61.²⁰
9. **East Syriac Adiabene.** This type was used in the Dominican Press in Mosul and by Cambridge

¹⁷ Coakley, *Typography of Syriac*, 125–28.

¹⁸ Coakley, *Typography of Syriac*, 132–135.

¹⁹ Coakley, *Typography of Syriac*, 104–106, 136–137.

²⁰ Coakley, *Typography of Syriac*, 144–46; Bar Hebraeus, *Le Livre des Splendeurs de Grégoire Barhebraeus*, ed. Alex Moberg (London: Humphrey Milford, 1922).

University Press. Its design is derived from a type designed by W. Drugulin in 1883. This corresponds to Coakley E22.²¹

In addition to these print type-based fonts, the Meltho set includes two fonts based directly on manuscripts. The Estrangelo Antioch font was designed based on *Damascus 12/21*, a manuscript copied in 1041/2 CE and housed in the Syriac Orthodox Patriarchal Library in Damascus, and Estrangelo Midyat was designed based on a 13th century manuscript, originally of the Church of Mort Shmoni in Midyat and now at Mor Gabriel Monastery in Tur Abdin.

3. METHODOLOGY

3.1 OCR Process

This project aimed to give an overview of Tesseract's ability to OCR Syriac, and thus, the three major scripts were represented. A popular print type was chosen for each of the three type styles Estrangela, Serto, and East Syriac:

Estrangela: print type S14²²

Serto: print type W64²³

East Syriac: print type E22²⁴

Syriac texts that were printed in these frequently-used print types were chosen for OCRing. The Estrangela and Serto texts

²¹ Coakley, *Typography of Syriac*, 225–28. For the history of this East Syriac type, see J. F. Coakley, "Edward Breath and the Typography of Syriac," *Harvard Library Bulletin*, New Series, vol. 6, no. 4 (1995): 4–64.

²² Išo'dad de Merv, *Commentaire d'Išo'dad de Merv Sur l'Ancien Testament I. Genèse*, ed. J.-M. Vosté and C. Van den Eynde, CSCO 126, *Scriptores Syri 67* (Louvain: Imprimerie Orientaliste L. Durbecq, 1950).

²³ Severus of Antioch, *Les Homiliae Cathedrales de Sévère d'Antioche: traduction Syriacque de Jacques d'Édesse* (Homélies LII–LVII), ed. and trans. R. Duval, *Patrologia Orientalis* 4 (Paris: Librairie de Paris, 1908).

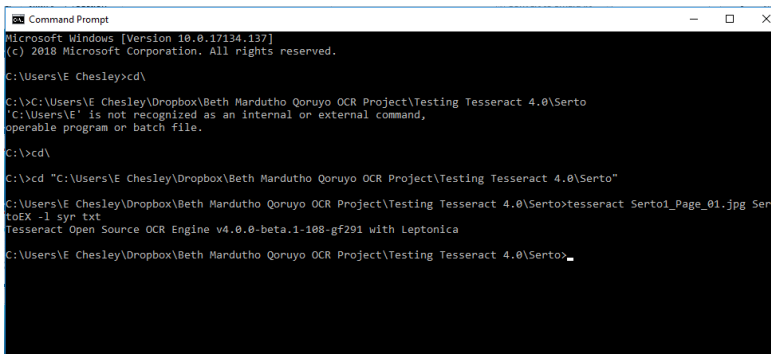
²⁴ *Acta Martyrum et Sanctorum Syriace*, vol. 1, ed. Paul Bedjan (Hildesheim: Georg Olms Verlagsbuchhandlung, 1968).

are unvocalized since the former is usually not vocalized and the latter is not generally vocalized in text editions. East Syriac, on the other hand, tends to be vocalized in many text editions. Several pages from each text were scanned at 300 dpi in black and white and cropped in order to generate images that only included the main body of the text, thus excluding marginal notes and footnotes.²⁵ Images were rotated as necessary to gain even, horizontal lines. Each page was then saved individually as an image file (.jpg or .tif), and three pages from each text were chosen for the current evaluation.²⁶

²⁵ In prior OCR tests, footnotes, headings, and marginal notes had caused difficulties with line segmentation and made the OCR results less accurate.

²⁶ In choosing a page to test, the clarity of the image and the amount of text were both considered. Pages that only included full lines of text were prioritized, as were images with the most even lines. The pages were also not pre-segmented in this OCR process; they were OCR'd as they appear in Figure 2.

Tesseract was downloaded onto two Windows laptops.²⁷ The analysts tested three images, each one page of the text, for Estrangela (S14), Serto (W64), and East Syriac (E22). Each image, in turn, was OCRed using both Syriac language modes for Tesseract (command-line arguments “-l Syr” and “-l script/Syriac” for Syriac Language and Syriac Script, respectively). Given the potential differences in outcome between these two modes, it was essential that Tesseract be tested separately with both options. Part of the project’s goal was to determine how similar or divergent the two modes are and to determine which one produces the most accurate OCR results in each Syriac type style. This double testing thus produced 18 OCRed documents, 6 for each type style, which were analyzed concerning their accuracy. The team analyzed 84 lines of Estrangela, 45 lines of Serto, and 66 lines of East Syriac in each mode.



```

Microsoft Windows [Version 10.0.17134.137]
(c) 2018 Microsoft Corporation. All rights reserved.

C:\Users\E Chesley>cd\

C:\>cd "C:\Users\E Chesley\Dropbox\Beth Mardutho Qoruyo OCR Project\Testing Tesseract 4.0\Serto"
'C:\Users\E' is not recognized as an internal or external command,
operable program or batch file.

C:\>cd\

C:\>cd "C:\Users\E Chesley\Dropbox\Beth Mardutho Qoruyo OCR Project\Testing Tesseract 4.0\Serto"

C:\Users\E Chesley\Dropbox\Beth Mardutho Qoruyo OCR Project\Testing Tesseract 4.0\Serto>tesseract Serto1_Page_01.jpg Ser
toEX -l syr txt
Tesseract Open Source OCR Engine v4.0.0-beta.1-108-gf291 with Leptonica

C:\Users\E Chesley\Dropbox\Beth Mardutho Qoruyo OCR Project\Testing Tesseract 4.0\Serto>

```

Figure 3: Tesseract Run through Command Prompt

There are multiple variables at play in this OCR process, and they were accounted for as carefully as possible

²⁷ The team decided to run Tesseract only on Windows computers as a point of consistency across the outputs, but it used more than one machine in order to check the outputs against each other. Tesseract 4.0 was downloaded from the GitHub page maintained by the Mannheim University Library (UB Mannheim): <https://github.com/UB-Mannheim/tesseract/wiki>, accessed 30 July, 2018.

throughout the test. The first is Tesseract's language code command-line arguments (which run Tesseract using Syriac Language and Syriac Script modes). A second variable is the output file kind. Tesseract can be invoked to output text (.txt files), PDF, or hOCR files,²⁸ among others.²⁹ Timing and particular computer used were additional variables, the impact of which the analysts did not initially know. In order to confirm the consistency of Tesseract, every OCR'd image was later re-OCR'd by the same analyst and with the same set of command line-arguments to confirm that identical results would be generated. Several days later, a different analyst re-OCR'd images processed by her colleagues on the other computer and input the same command-line arguments. The re-OCR process consistently produced texts that were identical in content (i.e., Syriac letters, diacritics, Roman letters, and Arabic characters) when the same image was run with the same commands. These quality control steps suggest that timing and hardware do not interfere with the Tesseract engine; these conclusions, in turn, allow the team to assume that its analysis applies broadly to

²⁸ The hOCR file type is built upon HTML. FAQ, GitHub, <https://github.com/tesseract-ocr/tesseract/wiki/FAQ>, accessed 30 July, 2018.

²⁹ One option for command-line arguments that occasionally interfered with accuracy was selecting multiple output file types at once. When executing the Tesseract command-line arguments, one must choose a file kind for output; selecting "txt" generates the OCR as a .txt file, and so forth. However, one can choose to generate multiple file kinds at once by entering all codes into the same command-line argument. For the most part, the OCR text that results from multi-output functions is identical to those generated by single-output functions. However, sometimes a multi-output function generated gibberish. For example, a page in Serto was run through Tesseract with the Script command-line argument and with an output of "txt pdf hocr," and one of the resultant lines was: "bs avwnE vrtwAUM oa airog + 𐤀𐤁𐤃𐤅 NA Kota) CApeTtoLG ®." This was by no means a common occurrence, and the team did not find the same problem repeated with other Serto images or in other type styles, which suggests that researchers need not anticipate an error like this happening frequently. That being said, running one output file kind at a time ensures the most accurate OCR of one's image.

Tesseract as an engine and not merely to individual analysts or machines.

3.2 Data Collection

After the pages were processed through Tesseract and converted to Word files, data collection began. First the team calculated data for the original images in the following categories: consonants, punctuation marks, diacritics, and non-Syriac characters (including numerals, brackets, and asterisks). To determine these numbers, several lines of each page were counted and averaged, and that average was then used to estimate character counts for the whole page.³⁰ The character count without spaces given within the OCR'd Word document was collected as well. This allowed for an overall comparison between character counts in the original image and in the resulting document.³¹ Additionally, numbers for individual consonants were counted amongst all of the original pages in order to calculate and discuss OCR error rates for specific consonants.³²

Next, the analysts began to identify and record the errors that appeared in the 18 OCR'd documents, comparing each one to its original image.³³ Any consonant, diacritic, punctuation mark, or other character that did not exactly match those in the original image were considered errors.³⁴

³⁰ One analyst counted the exact number of characters for her pages without using averages and found her counts to be similar to that of her colleagues.

³¹ This information is given in Appendix 2.

³² In other words, the team estimated *total* character counts for consonants, diacritics, punctuation, and non-Syriac characters, but it counted exact numbers of individual consonants.

³³ Despite the type style of the original image, the Word documents used by the analysts for comparison generally presented the text in each computer's default Syriac font, generally an Estrangela font. Analysts could leave the Estrangela font or change fonts to match the original image. Two analysts choose the former method, one the latter.

³⁴ Smudges or dust that appeared on the original image as if it were a character was not considered an error, per se, as the OCR correctly

These were recorded as precisely as possible. For example, if a *nun* was incorrectly recognized as a *ynd* in one word and as an *olaph* in another word, these occurrences were recorded as distinct errors. This allowed more precision in data analysis later. Over the course of the project, it is likely that each analyst reviewed her six pages more than five times, rechecking data and pulling information on specific issues. These reviews were often prompted by an unusual number of a certain type of error or a need to record some errors even more specifically than originally intended. Though tedious, the process allowed the analysts to identify the most common errors within Tesseract modes, within particular type styles, and within individual images, as well as to develop basic understandings of why these particular errors might have occurred. Having a record of all errors produced also provided a wider picture of the breadth of issues that can appear when OCRing Syriac texts.³⁵ After data collection was completed, analysis of accuracy and error trends began.

	Image #	ⲗ Missing	added ⲗ	ⲛ ⲙⲗ	ⲛ ⲙⲗ	ⲛⲗⲙ ⲙⲗ
OCR (lang)	Page 1	8				
	Page 2	2	2			
	Page 3	2		2		1
	Subtotal	12	2	2	0	1
OCR (script)	Page 1	6	1		1	1
	Page 2	1	0		1	1
	Page 3	2		1		6
	Subtotal	9	1	1	2	8
	Total	21	3	3	2	9

Figure 4: Sample Selection of Data Collected from East Syriac Spreadsheet

identified the character even though the character was not intended in the original image.

³⁵ Further on, there will be a discussion of truly odd occurrences found within these pages. Of course, this does not present an exhaustive account of what could appear in Syriac OCR pages, but it does give a basic understanding of what sorts of errors can occur.

3.3 Data Analysis

The key data point to determine was the rate of recognition accuracy. Accuracy, simply stated, is the inverse of error rate; accuracy was calculated by subtracting the error rate from one. For most data points, the error rate is understood as the total errors³⁶ (in the considered category) divided by the total characters in the original image (in the considered category).³⁷ Accuracy was computed for consonants, diacritic marks, punctuation marks, and other characters, as well as total accuracy across characters. Rates were calculated for each individual page, for each Tesseract mode of operation (Language or Script) within a Syriac type style (Estrangela, Serto, or East Syriac), and each Syriac type style as a whole (averaging results for both modes). Additionally, errors and their respective accuracy rates involving certain letters were considered, specifically when a letter in the original became a different character or disappeared in the resulting document.³⁸ The relative frequency of individual letters and their average accuracy within each combination of type style and Tesseract mode factored into the analysis. Comparisons were then made, and observations collected.

³⁶ This includes any truncated characters, which may be considered an OCR segmentation issue rather than an OCR recognition issue.

³⁷ Because accuracy was defined as total errors (including added characters) divided by characters in the original image, the accuracy rates can be negative if there are more resulting errors than characters in the original image. The characters that were added cause this problem. Some accuracy rates could not be calculated due to the lack of certain characters (e.g., non-Syriac characters) in the original.

³⁸ These consonant accuracy rates do not include truncated characters, as that is typically a segmentation issue and not a problem of recognition. They also do not include added consonants since added characters do not reflect Tesseract's ability to recognize the specific character.

Table 1: Variables Considered during Testing

Combinations of Accuracy Calculated	Subsets of Documents Analyzed	General Types of Errors Identified
Consonants Diacritics Punctuation Non-Syriac characters Total characters Individual letters (e.g., <i>mim</i> , <i>koph</i> , <i>yudh</i> , etc.) Groupings of letters (e.g., tooth letters, <i>riṣḥ</i> and <i>dolath</i> , etc.)	Individual pages/images Estrangela-Language Estrangela-Script Serto-Language Serto-Script East Syriac-Language East Syriac-Script	Addition Deletion Substitution

The analysts sought to discover which were the most predominate kinds of errors, which combinations of type styles and Tesseract modes were most accurate for which kinds of letters, and what were the overall accuracy rates scholars could expect with the current version of Tesseract. The project further sought to identify important areas for future training of the Tesseract engine. Analysis began with the most detailed descriptions of errors possible and grouped error types into increasingly broader categories. At the most narrow level, the analysts considered categories such as *ayn* becoming *héth* in East Syriac-Script, which only occurred once across all three pages. At the broadest level, the test tabulated total percentage of consonant errors compared across all six combinations of type style and Tesseract mode. The team also attempted to evaluate the impact of other factors such as pixel size and image quality on the OCR results. The section that follows details the specific results for each type style, subdivided into discussions of consonants, diacritics, punctuation, and non-Syriac characters.

4. TESSERACT TEST RESULTS

4.1 Summary of Overall Results

Table 2: Overall Accuracy Rates

	Conso- nants	Punctua- tion	Diacri- tics	Non- Syriac ³⁹	Total
Estrangela -Language	99.28%	84.62%	45.45%	4.76%	95.31%
Estrangela -Script	98.51%	80.77%	32.17%	-9.52%	93.70%
Serto- Language	93.67%	78.67%	22.52%	-33.33%	88.48%
Serto- Script	88.98%	61.33%	16.56%	-66.67%	83.28%
East Syriac- Language	89.62%	73.75%	60.69%	N/A	79.01%
East Syriac- Script	88.11%	75.00%	53.27%	N/A	75.50%

In the broadest terms, Tesseract generated the most accurate results with the Estrangela type style and using Tesseract's Language mode. The East Syriac type style paired with Script mode produced the lowest accuracy rates overall. Unsurprisingly, diacritics caused the most errors across all type styles and modes, while consonants were more easily recognized and differentiated. Punctuation and other non-Syriac characters also caused some errors across the board.

³⁹ Because accuracy was defined as total errors (including added characters) divided by characters in the original image, the accuracy rates can be negative if there are more resulting errors than characters in the original image. The characters that were added cause this problem. Some accuracy rates could not be calculated due to the lack of certain characters (e.g., non-Syriac characters) in the original.

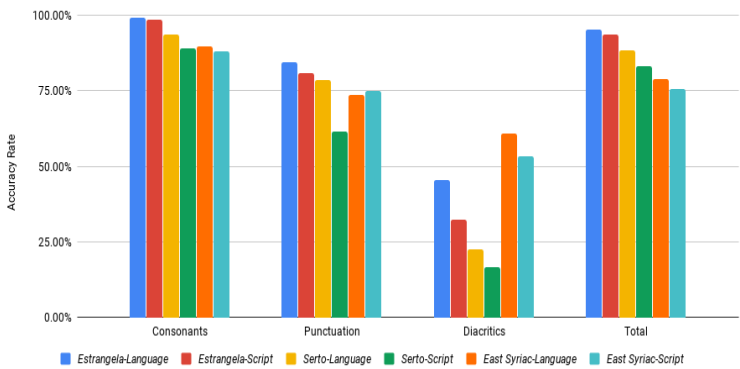


Figure 5: Overall Accuracy Rates

The analysts, each of whom examined different pages, naturally recorded different OCR errors.⁴⁰ The tested pages generated various kinds of errors, some of which were repeated across other pages and some of which were unique to one page—and sometimes unique to one instance. Though the analysts made great effort to be diligent and consistent in their individual work and as a team, human error must be acknowledged as a reality in any project that relies on human eyes. That being said, the pages studied and the resulting data were discussed and reexamined multiple times to give the team confidence in the data presented here.⁴¹ The analysts were also randomly assigned to review different variables from the other analysts’ data.

The individual data was analyzed to determine the range of the data set as well as the variety of accuracy rates within the pages. Most of the data given in this paper is based on the averages of the three analysts’ work, as this gives the best

⁴⁰ Individual results can be found within Appendix 2.

⁴¹ Within the data tables of individual data, it was observed that there was no consistent pattern that would indicate a strong distinction between the analyst’s data. For example, no one analyst consistently had the lowest or highest accuracy rates. The lack of pattern speaks to the consistent method of data collection across each of the three analysts.

overall picture of the results and compensates for human error. However, the range and standard deviation between individual data gave the team a deeper understanding of the underlying data points, and these are provided in tables in Appendix 2. For the most part, the data remained reasonably close between analysts; consonant accuracy rates only have ranges of one to four percent, while the ranges of total accuracy rates do not reach six percent. Diacritic accuracy rates, on the other hand, vary greatly between individual pages. For example, the ranges within the data for the Estrangela and Serto texts in Script mode hovered around 55%.⁴²

This section gives a general numerical summary of the test results. Tables and graphs, rather than prose, dominate this section in order to communicate clearly the most crucial pieces of data. The results of each type style are given, along with brief discussions about consonants, diacritics, punctuation, and non-Syriac characters. Further data can be found in Appendix 2.

4.2 Estrangela S14

Table 3: Overall Accuracy Rates for the Estrangela Type Style

	Conso- nants	Diacritics	Punctua- tion	Non- Syriac ⁴³	Total
Language	99.28%	45.45%	84.62%	4.76%	95.31%
Script	98.51%	32.17%	80.77%	-9.52%	93.70%
Total	98.89%	38.81%	82.69%	-2.38%	94.51%

⁴² Interestingly enough, East Syriac-Script had the most consistent diacritic rates among the analysts with a range of 1.65%.

⁴³ Because accuracy was defined as total errors (including added characters) divided by characters in the original image, the accuracy rates can be negative if there are more resulting errors than characters in the original image. The characters that were added cause this problem. Some accuracy rates could not be calculated due to the lack of certain characters (e.g., non-Syriac characters) in the original.

4.2.1 Consonants

Estrangela consonants were recognized very accurately with 99.28% of consonants OCRed correctly in Language mode and 98.51% OCRed correctly in Script mode. A total of 22 different types of consonantal error were recorded, 16 of which occurred only once or twice. This low number of occurrences for the majority of errors makes it difficult to determine their cause.

Table 4 shows the six consonantal errors which occurred three times or more. The most frequent error was word truncation, which caused three consonants to be deleted in Language mode and eight to be deleted in Script mode. The second most common error was a *yudh* being inserted into the text, which occurred twice in Language and six times in Script. These two errors were also common in Serto and East Syriac, and they are discussed in more detail in the Analysis.

Table 4: Most Frequent Consonant Errors in Estrangela⁴⁴

Type of Error	Estrangela- Language	Estrangela- Script
Word truncation	3	8
<i>Yudh</i> added	2	6
<i>Dolath-rish</i> added	2	2
<i>Yudh</i> deleted	2	2
<i>Simkath</i> added	3	0
<i>Shin</i> deleted	1	2

Table 5 gives the individual accuracy rate for each consonant in the two modes. In both modes, the majority of consonants

⁴⁴ These most frequent errors include those which occurred three or more times in total across both Language and Script modes.

Ɱ	97.62	Ɱ	98.53
Ɱ	94.00	Ɱ	92.86

4.2.2 Diacritics

The Estrangela pages chosen for testing were all unvocalized, meaning that in terms of diacritics they contained only homograph dots and *syomés*. In Language mode, only 45.45% of these diacritics were recognized correctly, and in Script mode, only 32.17% were recognized correctly. The most frequently occurring error was the diacritic homograph dot being substituted for another kind of diacritic. It was most often rendered as one of the very similar-looking dotted vowels, but was also occasionally misidentified as a Greek vowel, an Arabic diacritic, or a *syomé*. The second most common error was the homograph dot being missed or added, which happened 14 times in Language and 32 times in Script.

Table 6: All Types of Diacritic Errors in Estrangela

Type of Error	Estrangela- Language	Estrangela-Script
Homograph dot misidentified Total	48	48
As dotted vowel	40	43
As Greek vowel	5	4
As Arabic diacritic	2	1
As <i>syomé</i>	1	0
Homograph dot added or deleted	14	32
<i>Syomé</i> added or deleted	8	9
Dotted vowel added	2	5
Greek vowel added	4	2
<i>Syomé</i> misidentified	2	1

4.2.3 Punctuation

There were 78 Syriac punctuation marks on the Estrangela pages tested, with 12 errors recorded in Language mode and 14 recorded in Script. Only two types of errors were noted. The most common was the addition or omission of punctuation, which happened 12 times in Language and 11 times in Script. There were also three occurrences where a Syriac punctuation mark was recognized as an Arabic diacritic—for example, the *pāsuqā* was recognized as a *sukun*—and this error was exclusive to Script.⁴⁶ Syriac punctuation was only substituted with Arabic characters in Script mode, but diacritics were recognized as Arabic characters in both Language and Script modes.

Table 7: All Punctuation Errors in Estrangela

Type of Error	Estrangela-Language	Estrangela-Script
Punctuation added or deleted	12	11
Punctuation misidentified as an Arabic diacritic	0	3

4.2.4 Non-Syriac Characters

There were 21 non-Syriac characters in the Estrangela pages tested, consisting of asterisks and superscript numbers. None of these characters was recognized accurately by Tesseract, and the errors involving these non-Syriac characters were very inconsistent. For example, on one page there were ten superscripted Arabic digits (i.e., 1, 2, 3) identifying footnotes.⁴⁷ Nine of these numbers were rendered as other characters—one as an asterisk, three as double quotation marks, four as

⁴⁶ See, George Anton Kiraz, *Turrās Mamllā: A Grammar of the Syriac Language*, vol 1: Orthography (Piscataway, NJ: Gorgias Press, 2012), 149–50.

⁴⁷ Estrangela Page 1.

single apostrophes, and one as an exclamation point.⁴⁸ Only one numeral was recognized correctly, but Tesseract did not recognize that it was superscripted and instead placed it on the main line of text. Interestingly, two asterisks were rendered as double quotation marks, despite Tesseract evidently having an asterisk in its training module. Running English Language or Latin Script modes at the same time as Syriac did not improve the recognition of these characters, only resulting in an increased number of Syriac characters recognized as Latin characters (e.g., Estrangela *hé* recognized as Latin *m*).

Attentive readers may be perplexed by the negative percentages in this category: they are caused by Tesseract's addition of non-Syriac characters which were not present in the original images. This occurred in Script mode on only two occasions, but these two additions had a notable effect on the accuracy rate because of the small number of non-Syriac characters on the original pages.

4.3 Serto W64

Table 8: Overall Results for the Serto Type Style

	Consonants	Diacritics	Punctuation	Non-Syriac ⁴⁹	Total
Language	93.67%	22.52%	78.67%	-33.33%	88.48%
Script	88.98%	16.56%	61.33%	-66.67%	83.28%
Total	91.33%	19.54%	70.00%	-50.00%	85.88%

⁴⁸ Intriguingly, the number that was recognized as an exclamation point was the digit 8, not a 7 or a 1 as one might assume based on similarity of character shapes.

⁴⁹ Because accuracy was defined as total errors (including added characters) divided by characters in the original image, the accuracy rates can be negative if there are more resulting errors than characters in the original image. The characters that were added cause this problem. Some accuracy rates could not be calculated due to the lack of specific characters (e.g., non-Syriac characters) in the original.

4.3.1 Consonants

In Language mode, Tesseract recognized 93.67% of consonants accurately. In Script mode, 88.98% of consonants were recognized accurately. The consonantal errors identified in Serto varied widely, with over 60 different types of errors recorded. Half of these errors only occurred once or twice—again making it difficult to identify any patterns. However there were also certain errors which occurred with a particularly high frequency. Table 9 lists the 15 most commonly occurring errors and gives their frequency in both Language and Script mode.

Table 9: Top 15 Consonantal Errors in Serto

Type of Error	Serto-Language	Serto-Script
<i>Yudh</i> added	41	84
<i>Hé</i> misidentified		
Total	6	16
As ܚ	4	7
As ܚܐ	1	6
Other	1	3
<i>Ayn</i> added	2	16
<i>Yudh</i> misidentified		
Total	8	8
As ܝ	5	6
Other	3	2
<i>Héth</i> added	5	10
<i>Mim</i> misidentified	3	10
<i>Taw</i> misidentified		
Total	1	11
As ܬ	1	11
Other	0	0

<i>Shin</i> misidentified		
Total	9	2
As 𐌺	7	2
Other	2	0
<i>Dolath-rish</i> added	7	4
<i>Waw</i> added	1	10
<i>Dolath-rish</i> misidentified	6	4
<i>Olaph</i> added	2	7
<i>Héth</i> misidentified		
Total	2	7
As 𐌿	2	4
As 𐌾	0	3
Other	0	0
<i>Ṣodhé</i> misidentified		
Total	3	6
As 𐌽	3	6
Other	0	0
<i>Béth</i> added	3	4

The most common issue in Serto was the addition of extra letters, particularly the tooth letters *yudh*, *ayn*, and *héth*. This was common in both Language and Script mode but occurred twice as many times in Script mode. This is discussed in more detail in the Analysis section under *Added Tooth Letters*. *Dolath-rish*, *waw*, *olaph*, and *béth* were also incorrectly added to the text fairly frequently. *Waw* and *olaph* were inserted more times in Script mode, while *dolath-rish* was inserted more often in Language mode, and *béth* was inserted an almost equal number of times in both modes.

After the addition of consonants, misidentification of consonants was the second most common cause of errors. Certain consonants, when misidentified, were misidentified consistently. For instance, when Tesseract misidentified a *Ṣodhé*, it was always substituted with a *dolath*. Similarly, on all

the occasions that *taw* was misidentified, it was recognized as an *olaph*. Other consonantal substitutions were not completely consistent but still showed a tendency towards a particular letter. For example, *shin* was most likely to be misidentified as *phé*, which accounted for seven out of the nine substitution errors in Language and both of the substitution errors in Script. *Hé* was most frequently misidentified as a *waw* in both modes, and in Script it was also frequently mistaken for a *ṣayn* and *waw* together, which in Serto type style creates a very similar shape to *hé*. *Yudh* was mistaken for a *héth* 11 out of the 16 times it was misidentified and, likewise, all nine times *héth* was misidentified it was mistaken for one or two *yudhs*. The remaining two consonants that were frequently misidentified, *mim* and *dolath-rish*, showed more variation in their substitutions, with the 13 *mim* substitutions occurring across ten distinct error categories and the ten *dolath-rish* substitutions occurring in eight distinct error categories.

Table 10 shows the accuracy rate of each consonant in Serto in both Language and Script modes. The majority of consonants achieved over 97% accuracy in both Serto modes. However, when compared with Estrangela, Serto's numbers become less praiseworthy. Serto has an average of eight consonants performing at lower-than-97% accuracy, while Estrangela only has one that falls below this threshold. Moreover, certain consonants in Serto had a significantly low accuracy rate. *Ṣodhé*, for example, proved particularly difficult for Tesseract to recognize. Also illustrated by the table is the fact that the two modes can produce extremely varied accuracy ratings for the same letter. For example, *héth*, *taw*, and *mim* were over 10% more accurate in Language mode, whereas *shin* was 19% more accurate in Script mode.

Table 10: Consonantal Accuracy Rates in Serto

Serto-Language	Serto-Script
100	100
100	100
100	100
100	100
99.11	100
98.97	100
98.72	100
98.57	98.72
97.99	98.46
97.86	97.32
97.67	97.14
97.48	97.12
97.30	96.43
97.14	95.58
96.15	94.59
95.35	92.31
93.86	88.39
92.27	87.72
84.62	87.39
75.68	83.72
50.00	0.00

4.3.2 Diacritics

As with Estrangela, the Serto pages tested were unvocalized, meaning that the images’ only diacritics were homograph dots

and *syomés*. Of all three type styles tested, Tesseract produced the least accurate results for diacritics in Serto, with only a 22.52% accuracy rate for diacritics in Language mode and a 16.56% accuracy rate for diacritics in Script mode. These figures include a high number of diacritics that were incorrectly added by Tesseract, many of which the analysts believe were caused by stray marks in the original images.

Table 11 lists the seven different types of diacritic errors recorded for Serto and gives their frequency in Language mode and Script mode. The homograph dot was the locus of most Serto diacritic errors; it was frequently added, deleted, or substituted. When the homograph dot was misidentified it was most often substituted for a dotted vowel, as was also the case in Estrangela.

Table 11: All Types of Diacritic Errors in Serto

Type of Error	Serto- Language	Serto- Script
Homograph dot added or deleted	38	53
Homograph dot misidentified		
Total	33	34
As dotted vowel	22	25
As Greek vowel	8	9
As Arabic diacritic	3	0
<i>Syomé</i> added or deleted	14	12
<i>Syomé</i> misidentified	2	0
Dotted vowel added	18	23
Greek vowel added	12	3
Grave accent added	0	1

4.3.3 Punctuation

There were 74 punctuation marks on the Serto pages tested, and 15 errors were recorded in Language mode and 29 recorded in Script mode. The most common punctuation errors were the addition or omission of punctuation,

accounting for about 73% of punctuation errors in both Language and Script. The remaining errors involved punctuation being misidentified—usually as another Syriac punctuation mark, though in one instance, Serto-Script rendered a punctuation mark as an Arabic diacritic.

Table 12: All Punctuation Errors in Serto

Type of Error	Serto- Language	Serto- Script
Punctuation added or deleted	11	21
Punctuation misidentified as another Syriac punctuation mark	4	7
Punctuation misidentified as Arabic diacritic	0	1

4.3.4 Non-Syriac Characters

The Serto pages tested contained only three non-Syriac characters, consisting of asterisks and one superscript numeral. None of these were recognized correctly by Tesseract. Again, a minor number of added non-Syriac characters caused the negative accuracy rates—one extra non-Syriac character was inserted in Language mode and two were inserted in Script mode.

4.4 East Syriac E22

Table 13: Overall Results for the East Syriac Type Style

	Conso- nants	Diacri- tics	Punctua- tion	Non- Syriac ⁵⁰	Total
Language	89.62%	60.69%	73.75%	N/A	79.01%
Script	88.11%	53.27%	75.00%	N/A	75.50%
Total	88.87%	56.98%	74.38%	N/A	77.25%

⁵⁰ Because accuracy was defined as total errors (including added characters) divided by characters in the original image, the accuracy rates can be negative if there are more resulting errors than characters in the original image. The characters that were added cause this problem. Some

4.4.1 Consonants

Tesseract recognized 88.67% of consonants accurately with Language mode and 88.16% with Script mode. The consonantal errors were incredibly varied, with 99 different types of consonantal error recorded. Figure 6 gives a visual representation of the distribution of these errors.

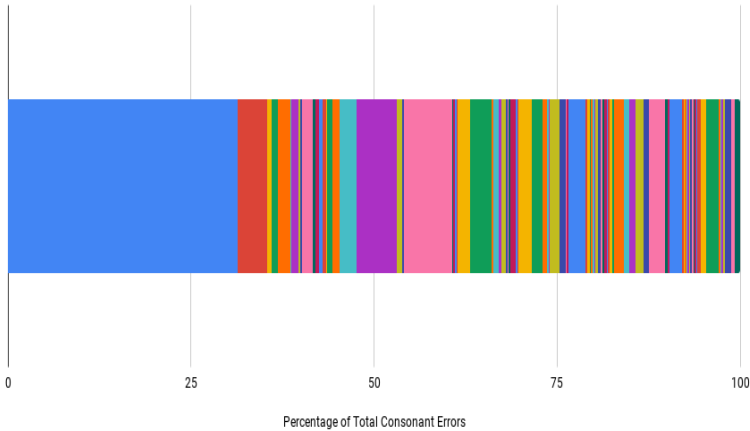


Figure 6: Distribution of Consonantal Errors in East Syriac

Since so many different categories of consonantal errors were produced in East Syriac and the majority of errors only occurred once or twice, it is difficult to discuss them all in depth and to discern patterns within the data. Table 14 lists the 12 most frequently-occurring errors and gives their frequency in both Language and Script modes. Here, specific errors have been subsumed under more general categories in order to provide a broader picture of the errors encountered.

accuracy rates could not be calculated due to the lack of certain characters (e.g., non-Syriac characters) in the original.

Table 14: The 12 Most Frequent Consonantal Errors in East Syriac

Type of Error	East Syriac- Language	East Syriac- Script
Word truncation	80	82
<i>Dolath-rish</i> misidentified		
Total	49	46
As ܐ	19	15
As ܕܕܕ	18	11
As ܕ	2	12
As ܐ	8	1
Other	2	7
<i>Nun</i> misidentified		
Total	15	24
As ܕܕܕ	4	9
As ܕ	2	4
As ܐ	2	3
Other	7	8
<i>Olap</i> deleted	12	9
<i>Zayn</i> misidentified		
Total	5	17
As ܕܕܕ	3	6
As ܐ	1	7
As ܕ	1	2
Other	0	0
<i>Olap</i> misidentified		
Total	3	18
As ܕܕܕ	1	9
As ܐ	0	5
As ܐ	2	1
Other	0	3
<i>Béth</i> misidentified		
Total	3	13
As ܕ	1	7

Other	2	6
<i>Shin</i> misidentified		
Total	8	8
As ܫ	5	4
As ܫ	2	2
Other	1	2
<i>Dolath-rish</i> deleted	6	6
<i>Téth</i> misidentified		
Total	6	6
As ܬ	6	6
Other	0	0
<i>Nun</i> deleted	6	5
<i>Héth</i> deleted	4	3

The issue that most affected Tesseract's accuracy with East Syriac text in both modes was word truncation, a problem resulting from segmentation issues. This caused Tesseract to omit 80 consonants in Language and 82 in Script. Tesseract also omitted many individual consonants from within words that had otherwise been recognized. The most commonly omitted consonants were *olaph*, *dolath-rish*, *nun*, and *héth*.

The majority of misidentification errors related to *dolath-rish*. Not only were these the most commonly misidentified consonants, they were also the consonants that others were most frequently substituted for. This issue is discussed in more detail in the Analysis section. Another frequently substituted consonant was *téth*, which was substituted for a *gomal* on every occasion that it was misidentified.

Table 15 gives the accuracy rate for each individual consonant in East Syriac. Tesseract was less than 97% accurate at recognizing the majority of consonants in both modes. The letters *dolath-rish*, *téth*, and *zayn* posed particular difficulty for Tesseract in East Syriac type style. The table also illustrates some of the large differences in accuracy of individual consonants between the two modes. *Zayn* is 50% more

accurate in Language and *béth* is 11% more accurate in Language, while *Ṣodhé* is 20% more accurate in Script.

Table 15: Consonantal Accuracy Rates in East Syriac

East Syriac- Language	East Syriac- Script
ܐ 100	ܐ 100
ܬ 100	ܬ 100
ܒ 100	ܐ 99.25
ܦ 100	ܐ 99.12
ܐ 98.87	ܬ 98.80
ܐ 98.80	ܦ 98.41
ܦ 97.62	ܐ 97.62
ܐ 97.62	ܬ 97.10
ܐ 97.55	ܐ 96.93
ܬ 96.74	ܦ 95.45
ܬ 95.65	ܐ 95.18
ܐ 94.05	ܦ 94.59
ܐ 88.95	ܬ 90.48
ܦ 86.35	ܐ 89.68
ܬ 85.71	ܦ 86.11
ܐ 84.31	ܬ 85.87
ܦ 80.56	ܐ 84.74
ܬ 80.00	ܐ 82.35
ܐܐ 75.23	ܐܐ 76.15
ܐܐ 73.91	ܐܐ 69.57
ܐ 71.43	ܐ 19.05

4.4.2 Diacritics

The original East Syriac pages tested contained *syomés*, homograph dots, dotted vowels, and Syriac oblique lines (Unicode character U+0747). Tesseract recognized the diacritics in East Syriac most accurately out of the three type styles tested, but the accuracy rate was still very low, with a 60.69% rate of accuracy in Language and a 53.17% rate of accuracy in Script.

Both the relative accuracy and the (still) high number of errors in East Syriac are likely due to the high frequency of vowel pointings in East Syriac to begin with. The three pages tested had a total of 1,239 original diacritics. The more diacritics there are and the more *kinds* of diacritics there are, the more opportunities Tesseract has to misread the image.⁵¹ By the same token, because East Syriac has so many more diacritics than the other type styles, Tesseract has far more chances to read them correctly. Tesseract can have 487 errors and 579 diacritical errors in East Syriac-Language and -Script, respectively, yet still achieve the highest percentage of diacritic accuracy of all three type styles tested.

Table 16: All Types of Diacritic Errors in East Syriac

Type of Error	East Syriac- Language	East Syriac- Script
Dotted vowel added or deleted	234	275
Dotted vowel misidentified		
Total	77	126
As equivalent Greek vowel	51	81
As non-equivalent Greek vowel	23	43
As Arabic diacritic	3	2

⁵¹ In addition, as will be discussed further in the section on Segmentation Issues, East Syriac's many diacritical marks invade the white space between lines and thus create difficulties in distinguishing words. Besides affecting Tesseract's segmentation, this may also contribute to misidentified diacritics.

substituted with other Syriac characters but on anomalous occasions were also recognized as Arabic or Latin characters.

Table 17: All Punctuation Errors in East Syriac

Type of Error	East Syriac-L anguage	East Syriac-S cript
Punctuation added or deleted	13	11
Punctuation misidentified as other Syriac character	6	5
Punctuation misidentified as Arabic character	0	1
Punctuation misidentified as Latin character	0	1

4.4.4 Non-Syriac Characters

There were no non-Syriac letters, numerals, or symbols in the original East Syriac pages, and so an accuracy rate was not calculated for this category.

With the detailed results for each type style fully recounted, this paper turns now to an analysis of prominent and unusual error trends that were observed during testing. To put it in terms of the popular metaphor, now that the individual trees have been catalogued, the following section will describe the forest.

5. ANALYSIS OF OCR ERRORS

While the OCR testing generated a long list of errors, some of which were unique, trends are still discernible in the data. The following pages detail major error trends that emerged from the Tesseract testing data. Tooth letters had statistically significant error rates across all type styles, especially in terms of addition and substitution errors. Other kinds of errors were perceived primarily in one type style; examples of this were *dolath* and *rish* in East Syriac, *mim* in Serto, and *judh* in Serto. Additional subsections consider unique errors, segmentation issues, and spacing issues as examples of what scholars can

expect to find in their OCR ventures. Along with summarizing each of these common error trends, the paper analyzes what factors from letter shape to OCR training caused these tendencies.

All three type styles differed as to their most common type of error, whether added, deleted, or substituted. While Estrangela-Script had relatively even numbers of all three, Estrangela-Language had significantly more added and deleted letters than substituted. Interestingly, for both Script and Language modes in Estrangela, Tesseract deleted the same number of letters it added. In Serto, added letters were most common in both modes, and deleted letters were least common. In East Syriac-Language, deletions were slightly more common error types than substitutions, but both far outnumbered insertions. In East Syriac-Script, additions remained the least common error type, but substitutions predominated over deletions.

Table 18: General Consonantal Errors in All Type Styles

	Additions	Deletions	Substitutions	Truncations ⁵³
Estrangela-Language	5	5	5	3
Estrangela-Script	17	17	6	8
Serto-Language	70	12	53	0
Serto-Script	146	16	73	0
East Syriac-Language	10	119	108	80
East Syriac-Script	10	115	154	82

Another important trend is immediately apparent from this information: Tesseract performed remarkably well with Estrangela under both modes. Very few consonantal errors were generated. In fact, Tesseract had a 99.28% consonantal

⁵³ In number of characters.

accuracy in Estrangela-Language, and a 98.51% consonantal accuracy in Estrangela-Script.

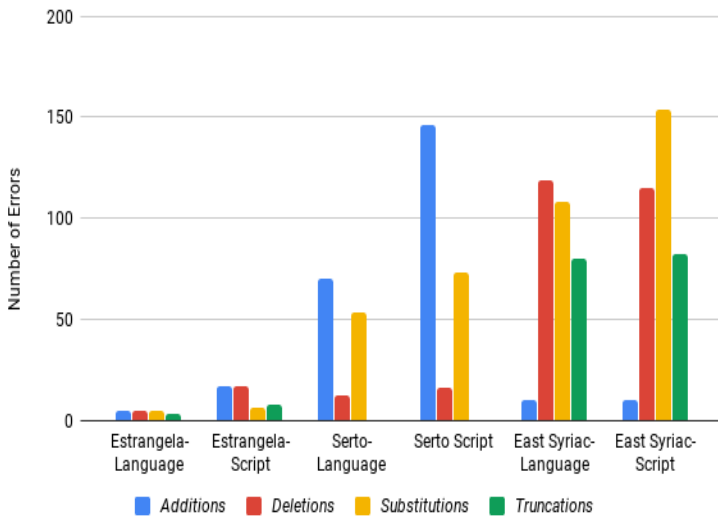


Figure 7: Most Common Error Types

Serto tended to have results at odds with the other two type styles, and this was mirrored at the underlying level of page layout. The Estrangela type style had an average of 32.2 consonants per line, and an average of 28 lines per page on the three images tested.⁵⁴ East Syriac had a similar average of 33 consonants per line, and 22 lines per page on the three images tested. Serto, on the other hand, had an average of 47.6 consonants per line—a more than 40% increase from the other type styles. Estrangela had an average of 28 lines per page on the three images tested, East Syriac had 22 lines per page, and Serto had an average of 15 lines per page. Thus, Serto differed from the other type styles both in average number of consonants per line (higher) and in average number of lines per

⁵⁴ For all three type styles, average consonants per line was calculated by choosing 5 full lines at random from the three pages (2 from one, 2 from a second, and one from a third), counting all the consonants on those lines, and taking the mean.

page (lower). Serto’s unique typographical features are paralleled in many of its Tesseract testing results. As will soon become clear, Serto exhibited markedly different results from the other type styles in several kinds of errors.

5.1 Tooth Letter Errors

Humorously referred to as “tooth letters” by Syriac instructors, *héth*, *yudh*, *ayn*, and *nun* exhibit strong similarities in all Syriac type styles.⁵⁵ All four letters are written connected to letters on either side, and they rise slightly above the line of text. Hence, they are easily confused—by human readers and by computer programs! One of the more frequent error trends in Estrangela and Serto was the rendering of these tooth letters. Tesseract frequently confused one tooth letter for another, added tooth letters (particularly *yudhs* in Serto), deleted tooth letters, or recognized other letters as tooth letters and vice versa. As Table 19 illustrates, tooth letter errors account for more than 14% of consonantal errors across all six type style-mode combinations. It ranges from 14.10% of consonantal errors in East Syriac-Language to 54.04% of consonantal errors in Serto-Script.

Table 19: Tooth Letter Errors as Percentage of Consonantal Errors in All Type Styles

	Tooth Letter Errors ⁵⁶	Consonant Errors	Percentage of Total Consonant Errors
Estrangela-Language	5	18	27.78%

⁵⁵ *Lomadh* is not considered a tooth letter despite its similarities to *ayn* because its “leg” extends higher up the page, making it more recognizable to Tesseract. Similarly, *ẓayn* is not counted as a tooth letter because it does not connect to the letters on either side, making it uniquely distinguishable.

⁵⁶ “Total Tooth Letter Errors” is defined as tooth letters being substituted with other tooth letters, tooth letters being added or deleted, and tooth letters being substituted with other non-tooth letters.

Estrangela-Script	14	37	37.84%
Serto-Language	68	135	50.37%
Serto-Script	127	235	54.04%
East Syriac-Language	32	227	14.10%
East Syriac-Script	43	259	16.60%

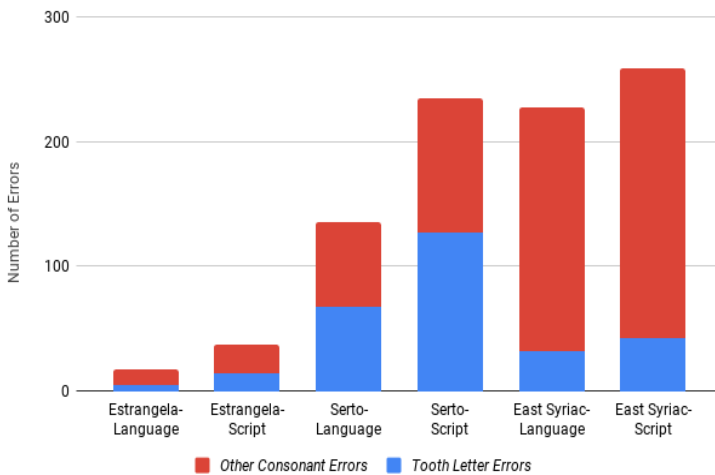


Figure 8: Tooth Letter Errors as Percentage of Consonantal Error

The kinds of tooth letter errors varied between the different type styles. By far the most common type of tooth letter error in Serto was added tooth letters. East Syriac, on the other hand, had more substitution errors, in which one tooth letter was misread as a different tooth letter, and errors of tooth letter deletion. The types of tooth letter errors were relatively even in Estrangela, with four, eight, and six total substitution, addition, and deletion errors, respectively.

As Table 20 shows, Tesseract rendered tooth letters with a range of accuracies across the six combinations of type style and mode. Both Estrangela modes had higher than 97%

accuracy rates at rendering all four tooth letters, and the two East Syriac modes ranged from 90.61% to 93.01% accurate. Particularly for Estrangela, Tesseract may be relied upon with a high degree of certainty in regard to these four letters (*héth*, *yud*, *nun*, and *ayn*). Serto, on the other hand, performed relatively poorly. Serto-Language generated an 82.96% accuracy rate, and Serto-Script was only 68.17% accurate.

Table 20: Tooth Letter Accuracy Rates in All Type Styles

	Total Tooth Letter Errors ⁵⁷	Total Tooth Letters in Original Image	Tooth Letter Accuracy Rate
Estrangela-Language	5	505	99.01%
Estrangela-Script	14	505	97.23%
Serto-Language	68	399	82.96%
Serto-Script	127	399	68.17%
East Syriac-Language	32	458	93.01%
East Syriac-Script	43	458	90.61%

⁵⁷ Defined as tooth letters being misidentified as other tooth letters or as other non-tooth letters (substitution), tooth letters being removed (deletion), and tooth letters being added from nothing (addition).

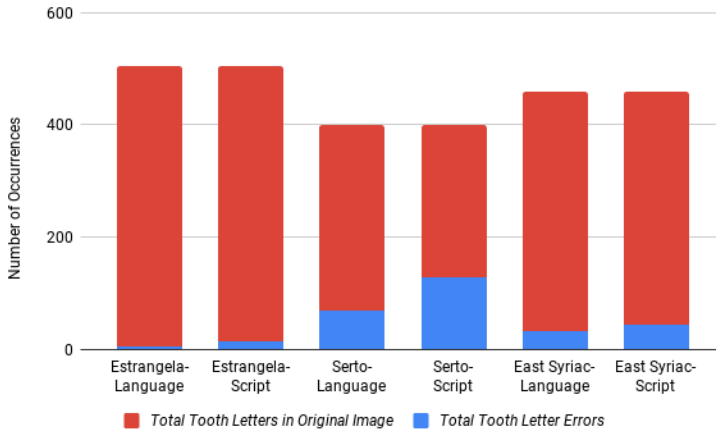


Figure 9: Tooth Letter Errors as Percentage of All Tooth Letters

Zayns are less similar to the “tooth letters,” differentiated particularly by lack of script attachment to surrounding letters. However, when a *zayn* precedes a medial *nun* or *yudh* Tesseract occasionally recognized the two letters as a *héth*. This happened once each way in East Syriac-Script.

5.1.1 Added Tooth Letters

As illustrated by Table 21, the three type styles varied in their encounters with added tooth letter errors. Surprisingly, East Syriac had the least problems with added tooth letters, and East Syriac-Language did not add any tooth letters. The OCR process only added a few tooth letters in Estrangela-Language and Estrangela-Script. In fact, Estrangela’s only tooth letter errors were eight *yudhs* added across all six pages (two in Language and six in Script).

Contributing in large measure to Serto’s poor numbers regarding tooth letters, is the frequent addition of *yudhs* in both Serto-Language and Serto-Script modes. While the reasoning behind this phenomenon is unclear, Tesseract added many *yudhs* during the OCR process. Sometimes—though not every time—these appeared to be added when a dot rested above the line of text in the original image, as if Tesseract believed a letter

was pushed out of line. The differences between Serto and the other type styles are stark: while East Syriac added only one *yudhs* across all six pages tested (the one occurrence was in Script mode) and Estrangela added eight *yudhs*, Serto inserted an astounding 41 *yudhs* in Language and 84 in Script! This accounts for about 30–35% of consonantal errors in both of Serto’s modes. Consequently, added *yudhs* also account for a large percentage of added tooth letter errors in Estrangela and Serto.

In Serto-Language and Serto-Script, however, 48 and 111 tooth letters were added, respectively. Most of its added tooth letters were *yudhs*: Serto’s many added *yudhs* made up 85.42% of added tooth letter errors in Language mode and 75.68% in Script mode. If *yudhs* are discounted, Serto only had 7 tooth letter additions in Language and 27 in Script.

Table 21: Added Tooth Letter Errors as Percentage of Total Tooth Letter Errors

	Added Tooth Letters	Total Tooth Letter Errors	Percentage of Total Tooth Letter Errors
Estrangela-Language	2	5	40.00%
Estrangela-Script	6	14	42.86%
Serto-Language	48	68	70.59%
Serto-Script	111	127	87.40%
East Syriac-Language	0	32	0.00%
East Syriac-Script	1	43	2.33%

Another unique error within Serto came when one line in both Serto-Language and Serto-Script added multiple *héths* that could not be accounted for with stray marks.⁵⁸ This error only occurred one time elsewhere on the page. The *héths* were added when there was a longer connecting line between two letters,

⁵⁸ Serto Page 3.

such as a *yudh* and a *taw*. The rest of the page did not see such a wide expanse between letters, and it is possible that Tesseract added the *héths* to make up for the added space between letters.

5.1.2 Tooth Letter Substitutions

Given the similar appearances of each tooth letter to another, the analysts anticipated prior to the testing that a high rate of substitution errors would result, wherein one tooth letter (an *ayn*, for instance) would be incorrectly recognized as a difference tooth letter (say, a *nun*) by Tesseract. In actuality, Tesseract had a very high accuracy rate in this regard. The least-accurate permutation tested, Serto-Language, still only misidentified 2.01% of tooth letters as other tooth letters. Put another way, while added and deleted tooth letters made up a high percentage of consonantal errors, confusion between letters did not significantly contribute to the error rate. Tesseract thus appears to have a high degree of accuracy in recognizing tooth letters, and evidently other factors cause the added letters.

Table 22: Substitution Errors as Percentage of Original Tooth Letters

	Tooth Letter Substitution Errors	Total Number of Tooth Letters in Original Image	Substitution Error Percentage within Total Occurrences
Estrangela- Language	1	505	0.20%
Estrangela- Script	3	505	0.59%
Serto- Language	8	399	2.01%
Serto-Script	7	399	1.75%
East Syriac- Language	11	458	2.40%
East Syriac- Script	15	458	3.28%

5.2 Errors Unique to One Type Style

Some errors are typical across all combinations of type style and mode, but the tests also generated discrepancies between type styles in which one type style had markedly different results than the other two. This difference was particularly noticeable in the cases of *rišes* and *dolaths*, *mims*, and added *yudhs*.

5.2.1 *Dolath* and *Rish* in East Syriac

Table 23: Accuracy Rates of Dolath and Rish in All Type Styles

	<i>Dolath-rish</i> Errors	Total <i>Dolaths</i> and <i>Rishes</i> in Original Image	Error Rate	Accuracy Rate
Estrangela- Language	3	359	0.84%	99.16%
Estrangela- Script	4	359	1.11%	98.89%
Serto- Language	15	208	7.21%	92.79%
Serto- Script	10	208	4.81%	95.19%
East Syriac- Language	59	218	27.06%	72.94%
East Syriac- Script	54	218	24.77%	75.23%

While close to 100% of *dolaths* and *rishes* in Estrangela and over 92% in Serto were recognized accurately in both modes, East Syriac had relatively imprecise accuracy rates for these letters: 72.94% in Language and 75.23% in Script. Close to 25% of *dolaths* and *rishes* were incorrectly identified or deleted in each East Syriac mode. This may have to do with the shape of *rishes* and *dolaths* in the East Syriac font. With their large curved shape, they look similar to *waws* and *kophs*. Indeed, Tesseract substituted a total of 35 *dolaths* and *rishes* for *waws* and a total of 14 *dolaths* and *rishes* for *kophs*. Interestingly, the misidentification did not go both ways; on no occasion was a *waw* substituted for a *dolath* or *rish*, and *kophs* were very rarely identified as a *rish* or *dolath* (only three total occurrences of this type of error in both East Syriac modes). Tesseract also had some difficulty distinguishing *dolaths* and *rishes* from each other and from the dotless *dolath-rish* consonantal stem. These were

original *olaphs*. This indicates that Tesseract’s ability to accurately recognize East Syriac text could be improved by further training focused on distinguishing *dolath-rish* from consonants with similar shapes and further training to distinguish diacritics from *dolath-rish* dots.

5.2.2 *Mim* in Serto

Mims exhibited a similar pattern, being very accurately rendered in two type styles but less-accurately so in the third. Estrangela had a nearly 100% accuracy rate in both modes, and East Syriac-Language and -Script were both about 98% accurate. However, Serto-Language was 95.8% accurate, and Serto-Script was 86.55% accurate. The different shape of *mims* in Serto likely contributes to Tesseract’s less-accurate performance here. A glance at the kinds of *mim* substitution errors across type styles supports this hypothesis: In East Syriac, *mims* were mistaken as a *qoph* (twice in Script), a *phé* (once in Language), and an *olaph* (once in Language). In Serto, by contrast, *mims* were misidentified with six (different) letters. In Serto-Language, *mims* were recognized as a *lomadh* (1), a *waw-béth* (1), and a *koph-lomadh* (1); and in Serto-Script, *mims* were recognized as a *koph*-combination (3),⁵⁹ a *béth* (2), a *waw* (1), a *lomadh* (1), a *waw-lomadh* (1), a *héth* (1), and an *ayn* (1).⁶⁰ A *mim*’s shape in Serto is similar to greater variety of letters than in the other type styles; particularly in the font W64, *mims* can look like a rounded letter followed by a tooth letter. Given that even human readers might mistake the *mim* as a different set of letters, it is perhaps unsurprising that Tesseract was occasionally confused.



Figure 11: W64 Serto *mim*

⁵⁹ Mims were recognized variously as: *koph*, *koph-lomadh*, and *koph-ayn*.

⁶⁰ Tesseract exhibited no *mim* substitution errors in Estrangela. In fact, its only *mim* error at all was one deleted *mim* in Estrangela-Script.

A second issue arises when considering the great disparity between Serto-Script and the five other Tesseract permutations. A nearly 87% accuracy rate is not unsatisfactory, but what makes this error rate so intriguing is the fact that it differs so markedly from the five other type style-mode combinations. What would make Serto *mims* particularly confusing to the Script mode but not to the Language mode? While the shape of Serto *mims* seems to be an issue for Tesseract, if shape were the only problematic variable one should expect to see it causing similar problems for Tesseract in *both* modes, not only one. Evidently, some aspect of the Script command-line argument (-l script/Syriac) contributes to this misreading.

Rarely were other letters read as *mims*. A *shin* became a *mim* once in East Syriac-Script. A *mim* was substituted for a *koph* once in Serto-Language. A *waw* was recognized as a *mim* once in both Serto-Language and Serto-Script. Estrangela did not substitute a *mim* for any letter in either mode.

Table 24: *Mim* Error Rates in All Type Styles

	<i>Mim</i> Errors	Total <i>Mims</i> (in original image)	Total <i>Mim</i>
Estrangela-Language	0	153	0.00%
Estrangela-Script	1	153	0.65%
Serto-Language	5	119	4.20%
Serto-Script	16	119	13.45%
East Syriac-Language	3	126	2.38%
East Syriac-Script	2	126	1.59%

5.2.3 *Yudhs* in Serto

Added *yudhs* have already been discussed in detail under *Added Tooth Letters*. However, it bears repeating that this particular error was unique to Serto. Tesseract added a high number of *yudhs* in Serto (41 and 84 in Language and Script modes, respectively), but the other two type styles had a miniscule number (three and eight total for Estrangela and East Syriac, respectively).

5.2.4 A Brief Look into Serto

Attentive readers will have noticed that Serto tends to have distinct error trends compared with the Estrangela and East Syriac type styles. Tooth letters, particularly *yudhs*, were added erroneously in all type styles but at a far higher rate in Serto. Tesseract misread *mims* to a higher degree in Serto. Serto had the lowest diacritic accuracy of all three type styles. While stray marks in the Serto images may have contributed to some of these OCR errors, they cannot account for all of Tesseract's mistakes. Since Serto images had essentially identical resolutions to those in Estrangela and East Syriac, pixel size cannot account for this discrepancy either.

The most likely explanation for Tesseract's unusual error patterns in Serto lies at the level of font. The print types tested in this paper were chosen based on popularity—because they were frequently-used print types in each type style. As the analysts discovered partway through the testing project, Tesseract was most assuredly trained on Meltho fonts, and thus, print types closely related to Meltho fonts will perform better with Tesseract. Print types S14 and E22, those tested here for Estrangela and East Syriac, happened to be part of the set of types used as the basis for the Meltho computer fonts. Their closeness in shape to Tesseract's training modules explains their OCR superiority in these areas. However, the Serto print type tested, W64, was not part of the set of print types upon which Meltho fonts were based, making its letter shapes further from Tesseract's baseline. By correlation,

Tesseract is less able to recognize text printed in W64. This explains much of the broadest level results between Estrangela, East Syriac, and Serto. If Tesseract were trained with a font based on W64, Tesseract's OCR accuracy with Serto print type would likely improve significantly.

5.2.5 A Note on East Syriac


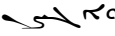





East Syriac still had the lowest average consonantal accuracy (88.87%, averaging the two modes), slightly below Serto (91.33%, averaged) and far behind Estrangela (98.89%, averaged). Yet E22 was part of the pool of print types underlying the Meltho fonts. How can its low performance be explained? Very simply, as it turns out. The shapes of letters in the East Syriac type style are more similar in shape to one another than the same letters are to each other in Serto and Estrangela type styles, which show a greater differentiation of letter design. *Dolaths* and *rishes* in particular, are uniquely shaped in the E22 print type and look very similar to *kephs*, *béths*, *waws*, *hés*, *phés*, and *qophs*. Errors with *dolath* and *rish* makeup 25.99% and 20.78% of consonantal errors in East Syriac-Language and East Syriac-Script, respectively.⁶¹ Because more letters look far more alike in East Syriac than in the other type styles, it is more difficult for Tesseract to distinguish between them.

5.3 Unusual One-Off Errors

In order to give a well-rounded picture of Tesseract's OCR results, it is necessary not only to highlight the larger trends observed within the results but also to mention some of the specific oddities that appeared in those same pages. These kinds of errors, examples of which are given below, do not seem to follow any specific trend in formation. While smudges or marks contributed to some errors throughout the texts and

⁶¹ By comparison, *dolath* and *rish* errors make up 16.67% of errors in Estrangela-Language, 10.81% in Estrangela-Script, 11.11% in Serto-Language, and 4.26% in Serto-Script.

modes, those errors are not included here. The following errors have been deemed to lack a clear or obvious explanation.⁶²

- In Estrangela-Script, the word  became .⁶³ The addition of these two letters, absent in this same page when run in Language mode, appeared at the end of a line and made the word twice as long as it appeared in the original.
- In Serto-Script, a  became Jo.⁶⁴ Interestingly, this change happened within a word ( to ). Though these Serto letters mirror the given Latin letters in shape, there is no indication as to why Tesseract decided to shift from Syriac to Latin characters here yet did not do so elsewhere on this page.
- Also in Serto-Script, the word  dropped its final letter and became OO.⁶⁵ As in the preceding example, there is no obvious reason to read the Syriac characters as Latin letters here, and even if there were, the last letter could be easily understood as a Latin letter as well and yet was not.
- In East Syriac-Script, a similar occurrence took place when  became eS.⁶⁶ These Latin letters do not map as easily onto the Syriac letters, and the analyst even ran the page through Tesseract again to confirm this result.
- In both East Syriac-Language and East Syriac-Script an Arabic *shadda* was added by Tesseract in various places, seemingly at random. Examples include one placed above a *waw* that had no diacritic and one set alongside

⁶² This list is not exhaustive. Other unusual errors have been given throughout this report, and even more happened without a specific mention here.

⁶³ Estrangela-Script Page 3.

⁶⁴ Serto-Script Page 2

⁶⁵ Serto-Script Page 3.

⁶⁶ East Syriac-Script Page 1.

the lower dot of a colon.⁶⁷ While both of these examples include a colon since the *waw* was originally followed by one, the connection between an additional *shadda* and an existing colon cannot be fully determined.


5.4 Segmentation

Before Tesseract can properly read the Syriac text it must be able to segment the page correctly into words and lines. One of the telling signs of incorrect segmentation comes from truncated words and missing lines. If Tesseract drops letters at the beginning or end of words, particularly words at the beginning or end of lines, or if entire lines are skipped during the OCR process, then Tesseract most likely did not recognize the correct beginning and ends of word and line gaps. Tesseract is trained to determine word and line divisions based on white spaces. So, if many punctuation marks fill spaces between words or many diacritics appear above and below the lines, Tesseract can erroneously conclude that there is no white space and, hence, that there are no letters to be read in those places. The test results align with this understanding of Tesseract's segmentation function. While the majority of lines were accurately recognized as such by Tesseract in both Language and Script modes, Tesseract nevertheless exhibited issues with segmentation in East Syriac, the type style containing the most diacritical marks.

Tesseract segmented Serto with 100% accuracy in both Script and Language modes. There were no truncated words or missed lines. Serto's remarkable segmentation accuracy likely arises from its wide white spaces between lines, as well as from its extra-long lines of text. The Serto images which were OCR'd contained on average 47.6 characters per line, a 45–47% increase from the average line length in the East Syriac and Estrangela images. And the tests conducted on

⁶⁷ East Syriac-Language Page 1, East Syriac-Script Page 3.

manuscripts (discussed below) gave evidence that the longer a line of text is, the better Tesseract is able to recognize it.

Tesseract performed similarly well with Estrangela, truncating only 11 characters total across all six pages of Script and Language modes. Six of those characters were a  cluster at the end of the first line of the page that dropped off in both Language and Script.⁶⁸ The remaining five were truncated from the first word in the 18th line of the same page in Script. Given the 5,541 total characters in all six pages, this produces a segmentation accuracy rate of 99.80%. This segmentation accuracy likely results from the wide, white spaces between lines in the Estrangela images tested.

With East Syriac, Tesseract truncated multiple words and missed a full line in both Script and Language for a total of 162 dropped characters. With a higher total number of characters on the page to begin with, this still generated a 97.59% accuracy rate. Given that the East Syriac text had far less white space between lines compared with the Serto and Estrangela images as a result of its many diacritical marks, the less-precise segmentation results for East Syriac are naturally explained.

5.5 Spacing

Alongside character recognition and segmentation errors, the testing detected errors in spacing between words. The OCR process often added extra paragraph breaks between lines, mostly when there was truly a paragraph break in the text. There were also additional spaces between punctuation marks and words. As those issues do not inhibit the reading or searchability of the text itself, this paper does not discuss them.

Some spacing issues, however, do impact the reading of the text. Found in several documents, there were cases of added spaces in the middle of words and deleted spaces between words.⁶⁹ The chief example of this comes from a line

⁶⁸ Estrangela Page 1.

⁶⁹ Among these cases were East Syriac-Language Page 3 (additions and deletions) and Serto-Script Page 3 (deletion).

in the East Syriac text (Figure 12). A line with few pixels between words, it was first recognized in Language mode as one word with no spaces, even before and after the colon. The consonant and diacritic errors found in this line were typical of those found throughout the East Syriac results, but the extreme lack of spacing made this line unique. The Script mode fared far better, correctly adding spaces between several of the words, though it also combined the last few words into one. The number of pixels between words cannot be the sole cause of the problem at hand; other type styles exhibited spacing problems on a smaller scale. The wide spacing between words in the Serto document did not prevent two words combining into one for no discernible reason.⁷⁰ The Estrangela text, also with wide spacing, did not encounter this particular problem in the pages studied.

It should be noted that the accuracy rates given in this paper do not include spacing errors since the project's analysis is based on characters in the original image. The aim of the project was to determine character recognition, and the analysts prioritized texts with wide spacing in order to do so. When others perform OCR in the future, texts with narrower spacing between words may encounter more instances like the one given above, where multiple words are OCR'd as one word. Texts with wide spacing, like the majority of those tested here, may also find deleted or added spaces scattered throughout with no obvious cause. While this did not occur in every page tested in this project, even the small numbers of errors on these few pages imply a high likelihood of spacing errors across large texts.

Given this, as Tesseract stands, it seems inevitable that spaces will be added or eliminated between some words during the OCR process and thus interfere with the accuracy of the searchable text. This assumption is based on the team's observation that the current version of Tesseract does not encode whether the letters in an image are initial, medial, or

⁷⁰ Serto-Script Page 3.

final. Instead, it appears that the program maps the various allographic forms of a grapheme into one Unicode value and thus does not transfer the differentiation between letter forms to the resulting document. With the tight spacing of this East Syriac line (Figure 12), the reader must be able to recognize the final forms of the letters in order to determine where words end in the initial image. The OCR'd document in Language mode, for example, does not receive the information that the first *nun* in the line is a final *nun*, resulting in a much larger word. Fonts determine a letter's form based on what follows it, and thus the second *nun* is turned into a final *nun* because of the colon. Based on this observation, future training of Tesseract should work to encode the allographic forms of the grapheme into separate Unicode values, one for each initial, medial, and final. To aid with the spacing issue itself, spaces could be required after every final letter.

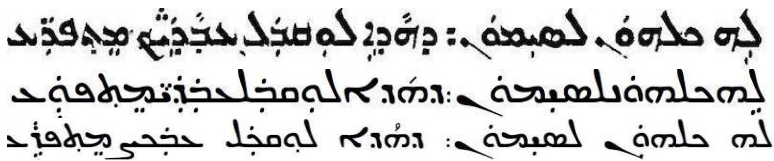


Figure 12: Comparison of East Syriac Original Image (top) with OCR Results for Language (middle) and Script (bottom)

With Estrangela, Serto, and East Syriac type styles tested in Tesseract, only one category of Syriac texts remains: handwritten manuscripts. No analysis of Tesseract's performance would be complete without an investigation into manuscripts. The analysts ran six manuscript pages through Tesseract, and their findings are outlined below.

6. MANUSCRIPTS

Following the study of Tesseract's performance on printed Syriac texts, the analysts conducted additional tests on images of Syriac manuscripts. The aim was to assess Tesseract's potential as a tool for digitally transcribing handwritten Syriac.

Three pages of *Damascus 12/21* from the Syriac Orthodox Patriarchate were chosen for testing: 72r, 159r, and 163v. *Damascus 12/21* was selected because its hand is the basis for the Meltho font Estrangela Antioch, and thus it was hypothesized that Tesseract had a high chance of being able to recognize this text accurately. Three particular pages were chosen for their relatively straight and well-spaced lines.

As in the study on printed texts, Tesseract was tested on all three pages using both language command-line arguments: Syriac Language (-l Syr) and Syriac Script (-l script/Syriac). Tesseract's output was then compared to the original image and all the consonantal errors were logged. In order to provide a point of comparison, three additional pages of Estrangela text from three different manuscripts were selected from *Hatch's Album of Dated Syriac Manuscripts* to be run through Tesseract and analyzed for errors using the same method.⁷¹ The pages selected were plates 29, 34, and 35.⁷² Again, these were chosen for their relatively straight and well-spaced lines.

6.1 Preparing the Images for Testing

Manuscripts initially proved more difficult for Tesseract to recognize than the printed texts. When the images of the manuscripts were run through Tesseract without any prior editing done to them, the OCR engine produced too little text to analyze—and in some cases no text at all. Therefore, several stages of edits were conducted on the images to make them “easier” for Tesseract to read and to derive more fruitful output files that could be analyzed in depth. The multiple edits were initially carried out on *Damascus 12/21* 72r until a result was achieved that the analysts felt was suitably successful; then the other images were edited following the same process.

⁷¹ W. H. P. Hatch, *An Album of Dated Syriac Manuscripts*, (Boston, MA: The American Academy of Arts and Sciences, 1946).

⁷² Plate 29 is London British Museum Add. Ms. 14599 fol.32; Plate 34 is Florence Biblioteca Laurenziana Plut. 1 Cod. 56 fol. 99; Plate 35 is London British Museum Add. Ms. 17152 fol. 30v.

The images of *Damascus 12/21* supplied for testing were high-quality color photographs that showed full pages of the manuscript on a black background. The text was written in Estrangela in two long and narrow columns per page, with approximately three words per column line and twenty-five lines per page. At the first stage in the editing process, the photograph of 72r was cropped of all empty space such as the background and page margins and then color-edited to black and white. Two images were created from the photograph, each containing one column of text. These two images were run through Tesseract, and both modes were tested. Despite the editing, Tesseract did not recognize any text in column 1 using either mode, and it recognized only 14 fragmentary lines in column 2. The specific 14 lines differed between Language and Script modes, but both produced 14 OCRed lines.

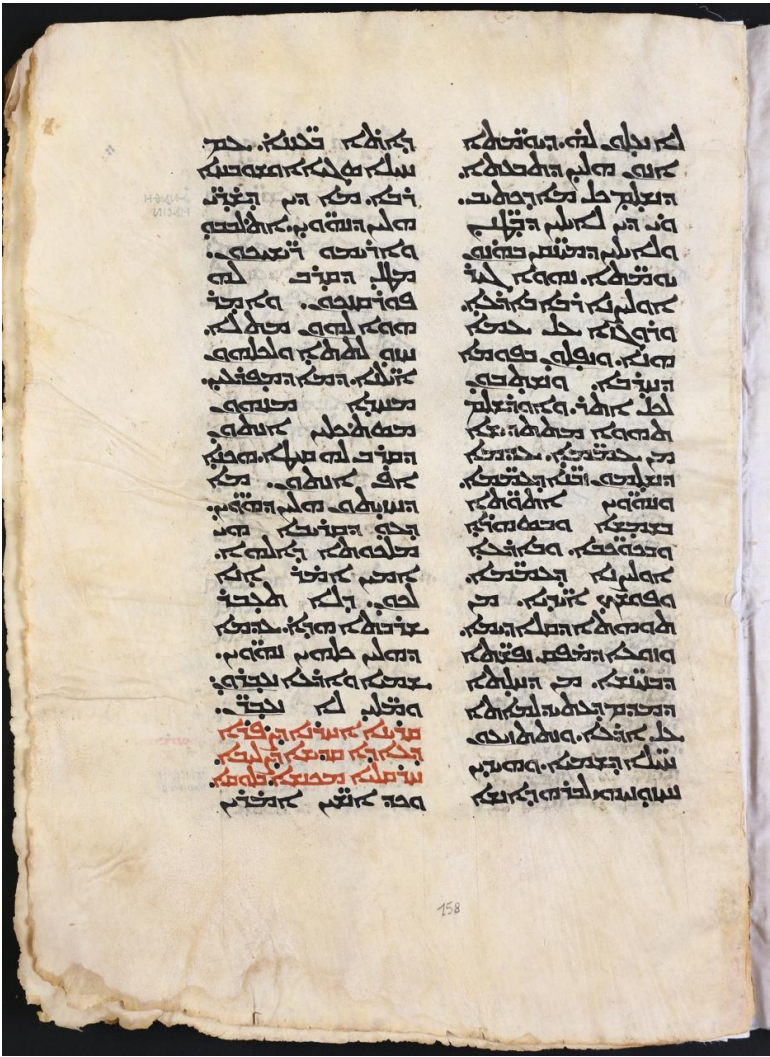


Figure 13: Sample Image of *Damascus 12/21 159r*⁷³

At the next stage of editing, line spacing was exaggerated so that no letters overlapped onto other lines. These images were tested again in both modes. A modest improvement was made

⁷³ Image courtesy of the Syriac Orthodox Patriarchate.

on the recognition of column 1, with 11 characters now recognized by Tesseract, but the results were still poor compared with the earlier tests on printed texts. Finally, the images were re-edited using a photo editor to combine both columns into a single image, elongating the lines so they fit approximately ten words per line rather than three. At this stage, more than half of the text was recognized by Tesseract and these improvements occurred in both modes.

Now that 72r had been successfully edited to produce similar accuracy rates to the printed texts, the remaining manuscript images were edited in the same way and run through Tesseract. As in previous tests, the output files were compared to the original images and the errors were recorded. The results that follow are from tests conducted on the final forms of the edited manuscript images.

6.2 Data Collection

When comparing Tesseract's output files to the original images, analysts recorded only consonantal errors in this preliminary test to provide a general idea of how well Tesseract might perform on manuscripts. The consonantal error categories were also simplified so that the errors encountered were recorded in one of four general categories—deleted, substituted, added, and truncated—which encapsulated all the varieties of consonantal errors that occurred. Consonants were considered “deleted” when the consonant in the original image was missing in the output from a word that had otherwise been recognized by Tesseract; consonants were considered “substituted” when a consonant in the original image was mistaken for a different consonant; consonants were considered “added” when a consonant appeared in the output where no corresponding consonant appeared in the original; and consonants were considered “truncated” when they were missing from the output due to partial or full lines of text being missed by Tesseract. Using this data, the analysts calculated accuracy rates for each set of pages in both modes.

6.3 Results

Table 25: Consonantal Accuracy Rates of Manuscripts

	Mode	Accuracy Rate
Plates 29, 34, and 35	Script	90.96%
Plates 29, 34, and 35	Language	86.44%
<i>Damascus 12/21</i>	Language	72.30%
<i>Damascus 12/21</i>	Script	67.26%

Table 26: Types of Consonantal Errors

	Plates 29, 34, and 35	Plates 29, 34, and 35	<i>Damascus</i> <i>12/21</i>	<i>Damascus</i> <i>12/21</i>
Mode	Script	Language	Language	Script
Deletions	52	57	87	94
Substitutions	76	97	129	112
Additions	17	28	38	53
Truncations	15	58	279	371
Total Errors	160	240	553	630

The three pages tested from *Damascus 12/21* contained a total of 1,924 consonants. There were 553 consonantal errors in Language mode and 630 consonantal errors in Script mode. With all these errors taken into account, the consonantal accuracy rate for the Language mode is 72.30% and the consonantal accuracy rate for Script is 67.26%. These results are significantly lower than the consonantal accuracy rate for the printed Estrangela texts that were tested, which were recognized with over 98% accuracy in both modes.

Tesseract generated significantly more errors of each type for *Damascus 12/21* than for the printed texts, but the low accuracy was primarily the result of the high number of consonants deleted due to line truncation. As Table 26 shows, truncated characters were by far the most frequent error, accounting for over half of the errors in both modes. As this error is possibly caused by issues with the input image, an accuracy rate was also calculated which excluded truncated characters from both the total consonant and the total error

counts. When truncated characters are removed from the calculation, the accuracy rate raises significantly to 83.34% in Language and 83.32% in Script.

Plates 29, 34, and 35 contained a total of 1,770 consonants. 240 errors were recorded in Language and 160 recorded in Script. With all these errors taken into account, Tesseract was able to correctly recognize 86.44% of consonants in Language mode and 90.96% of consonants in Script mode. These accuracy rates are still lower than those for the printed pages, but they are much higher than the accuracy ratings for *Damascus 12/21*. This increase in accuracy is attributable in part to the far-lower number of truncated characters, but the number of errors recorded for the other three categories decreased as well. When truncated characters are removed from the calculation, the consonantal accuracy rate raises slightly to 89.37% for Language and 91.74% for Script.

Table 27: Consonantal Accuracy Rate without Truncations

	Mode	Accuracy Rate
Plates 29, 34, and 35	Script	91.74%
Plates 29, 34, and 35	Language	89.37%
<i>Damascus 12/21</i>	Language	83.34%
<i>Damascus 12/21</i>	Script	83.32%

6.4 Analysis

These results indicate that Tesseract is able to recognize handwritten Estrangela with a reasonably high accuracy rate but only when the images have been heavily edited so that the manuscripts more closely resemble pages of printed texts. Even after a long editing process, Tesseract often truncates words or whole or partial lines, significantly reducing its OCR accuracy. These obstacles mean that, at present, Tesseract is probably not useful to scholars seeking a quick method for digitally transcribing manuscripts. If in the future Tesseract can be trained to recognize text in narrow columns, and if a method of minimizing line truncation is found, Tesseract could

become a useful and accurate tool for the optical character recognition of manuscripts.⁷⁴

7. TIPS FOR USAGE

This paper has outlined and analyzed a great amount of data from this project but has yet to spell out how this information may best help those who desire to use Tesseract for their scholarship. Here is a brief list of such practical tips:

- For all OCR usage, it is recommended to first crop all scanned images to eliminate all marginalia and footnotes. If there are footnotes, the superscripted footnote references within the body of the text are likely to appear as punctuation marks or numerals within the text (See Non-Syriac characters).
- Users should also be aware that any headings, especially those in a different script or language, should be specifically checked. This project did not study changes in language or styles, but it is likely that headings could cause some issues in the output.
- The sections on Usual One-Off Errors, Segmentation, and Spacing can help users conceptualize what may occur with their own texts. For each image OCR'd, the output file always inserted the line breaks as they appeared in the original image, leaving blank space on the left-hand side of the document. If the text is prose, users may want to go through the lines and eliminate the spaces. However, this should be done after the text has been checked, if line-by-line accuracy is desired.
- Ideally, the desired text would be printed in an Estrangela type style with a minimal amount of diacritics and OCR'd with using Language mode.

⁷⁴ Since completion of this study, the analysts have become aware of Transkribus—a software specifically designed for the automatic recognition of handwritten texts. Transkribus has shown good results with a variety of handwriting styles in different languages. Once trained, it may have the potential to work well on handwritten Syriac manuscripts.

These tests have established that OCRing the same page in both Language and Script modes, either simultaneously or one after another, will not substantially improve accuracy results. OCRing an already-OCR'd page will only magnify the number of errors since the second OCR run will create errors upon the first run's errors. Nor is it worthwhile to OCR the same page in both modes and then attempt to map the two onto one another to identify separate errors. Almost without exception, all of the consonantal errors present in Language mode were also created in Script mode.⁷⁵ And Script mode typically introduced additional errors beyond those of the Language mode. Thus, users can responsibly focus their OCR projects on Language mode alone.

- If looking to produce a near-perfect document, even Estrangela texts OCR'd in Language mode should be checked thoroughly. Table 30 in Appendix 2 lists the data from the three sample pages, none of which had perfect accuracy with consonants.
- The paper has focused on consonantal accuracy, but for practical use, the notes and data on diacritics should be consulted. For example, Estrangela texts OCR'd in Language mode had a significant amount of errors with homograph dots and *syomés* (see Estrangela Diacritics). Texts with minimal diacritics will have higher likelihood of success.
- If the goal is to conduct a word study, it would be best to consult Tables 39–41 in Appendix 2 and verify whether the letters within the word are likely to be OCR'd correctly.⁷⁶ Researchers might also check the

⁷⁵ The one exception being added *simkaths* in Estrangela-Language, which did not occur in Estrangela-Script.

⁷⁶ The program Voyant Tools has been brought to the analysts' attention as software that can assist with word studies by analyzing word frequency. Some researchers might find it useful after they have confirmed the accuracy of their OCR'd text corpus.

sections of this paper that cover errors common to those particular letters (See Analysis).

- If the user needs to see every single occurrence of a word or root in the text, it would be best to first create a near-perfect OCR'd text. Word truncation occurred even in the most accurate results and could cause problems in the text, beyond the letter accuracy data given in the tables below.
- On the other hand, if a more generalized understanding of a word's meaning and usage within a text is desired and the user does not require an analysis of the word's every occurrence, a full check on the OCR'd results would not be necessary.
- For example, if users desire to study a word such as ܡܚܐ in a particular text, they could check the tables in Appendix 2 below to confirm the expected accuracy rates of each letter. They would learn that, in Estrangela, each of the three letters is highly likely to be OCR'd correctly. Thus, when they utilize an appropriate find function in their word processor, most instances of ܡܚܐ should be easily identified.
- If the word in an Estrangela text includes ܐ, it may be easier to search for the letters surrounding it in the word because, as Tables 39–41 note, ܐ has the lowest accuracy rate of all consonants in that type style. Though generally inadvisable to OCR Serto or East Syriac texts, the same tables indicate that ܐ is one of the most accurate letters within those scripts. Words with a combination of more accurate consonants in Syriac or East Syriac can be studied through a search for that combination. If users want to investigate the

word ܕܠܐܬܪܝܫ in a particular East Syriac text, they would want to revisit the section of the paper dedicated to *dolath-rish*. Most likely, their search should include both ܕܠܐܬܪܝܫ and ܕܠܐܬܪܝܫ to assure that they do not miss a possible switch between the two letters.

- Beware that if users have an Arabic keyboard installed on their computers, Microsoft Word may automatically turn Arabic digits (1, 2, 3, etc.) into Arabic-Hindu numerals (١٢٣٤).

8. CONCLUSION

Several things may be concluded from the foregoing tests. First, this testing process confirmed Tesseract's reliability for OCRing printed Estrangela texts. When consonants are the reader's only concern, Tesseract 4.0 will perform at close to 99% accuracy using either mode. Since most printed Syriac texts use Estrangela, this is encouraging news for Syriac studies.⁷⁷ An electronic database of Syriac texts has been in development since the early 2000s, but the project has been hampered by the slow pace of manual transcription.⁷⁸ Using Tesseract to OCR these printed Syriac texts would dramatically speed up the process. Such OCRing should begin with Estrangela type styles. Of course, these accuracy rates are based on cropped scans in which marginalia and footnotes have been removed, so some background work will be necessary for OCRing. But given the time that is saved through digital OCR, the minor manual editing seems worthwhile for Estrangela texts.

⁷⁷ Estrangela's consonantal accuracy cannot likely be improved. According to the Tesseract programming community on GitHub, "unless you're using a very unusual font or a new language retraining Tesseract is unlikely to help." [ImproveQuality, GitHub, last updated April 21, 2018, <https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality>], accessed 30 July, 2018. Fortunately for scholars, Tesseract already generates useable OCR results for Estrangela.

⁷⁸ Digital Syriac Corpus, at syriacorporus.org, accessed 30 July, 2018.

Second, Tesseract 4.0 is not currently recommended for OCRing Serto and East Syriac. These type styles are less accurate and hover around 90% consonantal accuracy (88.11–93.4% range) in both modes. Tesseract may be useful for an initial computer-generated pass-through in these type styles, but for full readability and searchability of texts humans will need to review and edit the OCRed text themselves. While accuracy rates can be improved for these type styles, until a training method for Tesseract is developed that does not require inputting hundreds of thousands of lines of text it is unlikely that time spent retraining Tesseract's model will be time-effective. That being said, since fewer texts have been printed in East Syriac and Serto, manually transcribing these texts many prove a feasible alternative for gaining the digital texts.

Third, Tesseract's OCR accuracy with Serto can likely be improved by training Tesseract on a Serto font that is based on W64 print type. As the analysts discovered during the testing project, Tesseract was most assuredly trained on Meltho fonts, and thus, print types closely related to Meltho fonts will perform better with Tesseract. Print types S14 and E22, those tested here for Estrangela and East Syriac, happened to be part of the set of types used as the basis for Meltho fonts, while W64, that tested for Serto, was not. Much of Tesseract's unique error results for Serto are likely attributable to this discrepancy between print type and training font. If a programmer were to train Tesseract with a font based on W64, Tesseract's OCR accuracy with W64 print type would likely improve significantly.

Fourth, these tests have verified that Tesseract has the potential to recognize handwritten Estrangela texts with a good accuracy, but at present demands a laborious and time-consuming editing process to make the images readable. Even after this extensive editing process the accuracy rate varies between 67%–90%, and so it is not recommended that scholars attempt to OCR manuscripts using Tesseract at this time. Tesseract would need many programming developments

before it could become a practical, usable tool for OCRing Syriac manuscripts.

In practical advice, while variances remain when discussing individual letters, running Tesseract with the Language mode command-line argument (i.e., “-l Syr”) generates slightly more accurate results in all three type styles and is the recommended mode to use. In addition, the time spent cropping and deskewing scanned images before OCRing is well worth the investment. Without these steps, and without color-adjusting to black and white, Tesseract’s accuracy rates would drop significantly.

With a little practice, Tesseract can offer Syriac scholars a straightforward way of digitally transcribing printed Estrangela texts. If embraced by Syriac scholars, it has the potential to advance the field by improving the availability of printed Estrangela texts and opportunities to access them. As developments are made in OCR it is hoped that soon Serto and East Syriac will be recognized accurately enough to join Estrangela in this regard. The goal to OCR a high number and wide variety of printed Estrangela texts is sure to keep Syriac scholars busy in the meantime.

APPENDIX 2: DETAILED RESULTS

Individual Page Information⁸⁰

Table 29: Resolution and Size of Chosen Texts

	Estrangela	Serto	East Syriac
Page 1	300 dpi (1176 x 1814)	300 dpi (1687 x 1305)	300 dpi (1164 x 1823)
Page 2	300 dpi (1208 x 1901)	299.5 dpi ⁸¹ (1665 x 1408)	300 dpi (1170 x 1818)
Page 3	300 dpi (1182 x 1941)	300 dpi (1681 x 1459)	300 dpi (1170 x 1833)

Table 30: Individual Page Results for Estrangela-Language

	Consonants	Diacritics	Total
Page 1	98.52%	76.09%	95.53%
Page 2	99.76%	35.71%	96.13%
Page 3	99.53%	27.27%	94.31%
Mean	99.27%	46.36%	95.32%
Range	1.24%	48.81%	1.82%
Standard Deviation	0.66%	26.09%	0.92%

Table 31: Individual Page Results for Estrangela-Script

	Consonants	Diacritics	Total
Page 1	96.80%	65.22%	93.51%
Page 2	99.76%	23.81%	95.02%
Page 3	98.94%	10.91%	92.60%
Mean	98.50%	33.31%	93.71%

⁸⁰ The page numbers listed here do not reflect the page numbers in the printed books from which the images were scanned; rather they were the analysts' testing designation.

⁸¹ 299 dpi (horizontal) and 300 dpi (vertical).

Range	2.96%	54.31%	2.43%
Standard Deviation	1.53%	28.37%	1.22%

Table 32: Individual Page Results for Serto-Language

	Consonants	Diacritics	Total
Page 1	93.12%	36.36%	89.21%
Page 2	92.06%	8.93%	86.16%
Page 3	95.77%	25.49%	90.09%
Mean	93.65%	23.59%	88.49%
Range	3.70%	27.44%	3.92%
Standard Deviation	1.91%	13.82%	2.06%

Table 33: Individual Page Results for Serto-Script

	Consonants	Diacritics	Total
Page 1	88.14%	-11.36%	80.92%
Page 2	87.33%	44.64%	84.15%
Page 3	91.39%	9.80%	84.63%
Mean	88.95%	14.36%	83.24%
Range	4.07%	56.01%	3.71%
Standard Deviation	2.15%	28.28%	2.02%

Table 34: Individual Page Results for East Syriac-Language

	Consonants	Diacritics	Total
Page 1	90.18%	55.81%	76.49%
Page 2	89.05%	67.74%	82.22%
Page 3	89.65%	60.47%	78.57%

Mean	89.63%	61.34%	79.09%
Range	1.13%	11.93%	5.73%
Standard Deviation	0.56%	6.01%	2.90%

Table 35: Individual Page Results for East Syriac-Script

	Consonants	Diacritics	Total
Page 1	87.38%	54.12%	73.84%
Page 2	90.14%	53.08%	78.61%
Page 3	86.78%	52.47%	74.29%
Mean	88.10%	53.22%	75.58%
Range	3.35%	1.65%	4.77%
Standard Deviation	1.79%	0.84%	2.63%

Character Counts

Table 36: Character Counts for Estrangela

	Page 1	Page 2	Page 3	Total
Consonants	812	829	847	2488
Diacritics	46	42	55	143
Punctuation Marks	25	27	26	78
Non-Syriac Characters	11	6	4	21
Total in Original Image	894	904	932	2730
Total in Results	912	918	951	2781
Difference	-18	-14	-19	-51

Table 37: Character Counts for Serto

	Page 1	Page 2	Page 3	Total
Consonants	683	718	732	2133
Diacritics	44	56	51	151
Punctuation Marks	31	20	24	75
Non-Syriac Characters	2	1	0	3
Total in Original Image	760	795	807	2362
Total in Results	791	772	954	2517
Difference	-31	+23	-147	-155

Table 38: Character Counts for East Syriac

	Page 1	Page 2	Page 3	Total
Consonants	713	740	734	2187
Diacritics	473	341	425	1239
Punctuation Marks	22	27	31	80
Non-Syriac Characters	0	0	0	0
Total in Original Image	1208	1108	1190	3506
Total in Results	1169	1128	1125	3422
Difference	+39	-20	+65	+84

Overall Accuracy Rates

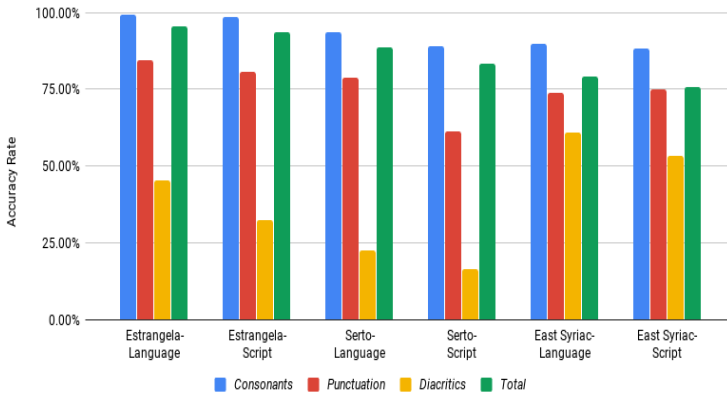


Figure 14: Overall Accuracy Rates of Type Style and Mode

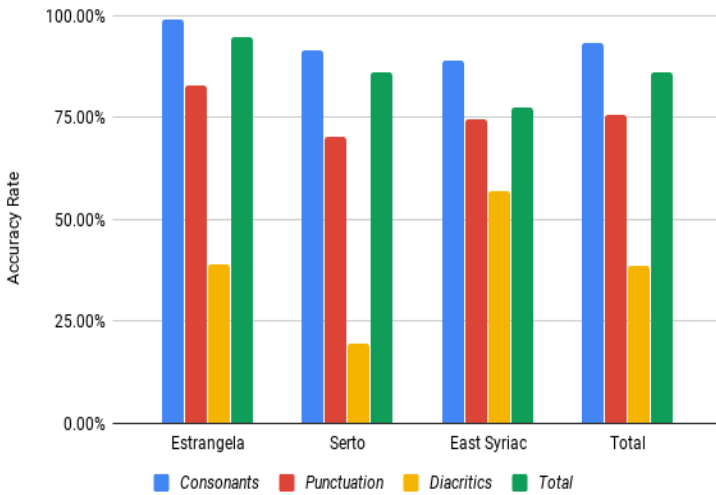


Figure 15: Overall Accuracy Rates of Type Style

Individual Letter Accuracy

Table 39: Accuracy Rates for Every Consonant in Every Type Style

	Estrangela -Language	Estrangela -Script	Serto- Language	Serto- Script	East Syriac- Language	East Syriac- Script
ܠ	100	100	98.72	98.72	94.05	89.68
ܘ	100	100	98.57	100	96.74	85.87
ܝ	100	100	100	96.43	85.71	90.48
ܝܐ	99.69	99.69	96.15	97.12	75.23	76.15
ܝܝ	100	100	93.86	87.72	100	99.12
ܝܝܐ	99.62	99.62	98.97	98.46	98.87	99.25
ܝܝܝ	100	100	84.62	92.31	71.43	19.05
ܝܝܝܐ	97.62	92.86	95.35	83.72	80.56	86.11
ܝܝܝܝ	100	100	100	100	73.91	69.57
ܝܝܝܝܐ	99.09	98.64	92.27	95.58	97.55	96.93
ܝܝܝܝܝ	100	100	97.30	100	86.36	95.45
ܝܝܝܝܝܐ	99.39	98.79	97.99	97.32	100	98.80
ܝܝܝܝܝܝ	100	99.35	97.48	87.39	97.62	98.41
ܝܝܝܝܝܝܐ	100	98.92	97.86	100	88.95	84.74
ܝܝܝܝܝܝܝܐ	94.00	100	100	100	100	100
ܝܝܝܝܝܝܝܝܐ	100	100	97.14	97.14	95.65	97.1
ܝܝܝܝܝܝܝܝܝܐ	100	100	97.67	100	97.62	97.62
ܝܝܝܝܝܝܝܝܝܝܐ	100	100	50.00	0.00	80.00	100
ܝܝܝܝܝܝܝܝܝܝܝܐ	100	100	100	100	100	94.59
ܝܝܝܝܝܝܝܝܝܝܝܝܐ	98.53	98.53	75.68	94.59	84.31	82.35
ܝܝܝܝܝܝܝܝܝܝܝܝܝܐ	100	100	99.11	88.39	98.80	95.18

Table 40: Consonantal Accuracy Rates, Arranged in Order of Letter Frequency in Syriac⁸²

	Estrangela -Language	Estrangela -Script	Serto- Language	Serto- Script	East Syriac- Language	East Syriac- Script
ܐ	100	100	98.72	98.72	94.05	89.68
ܐ	99.62	99.62	98.97	98.46	98.87	99.25
ܐ	100	98.92	97.86	100	88.95	84.74
ܐ	99.09	98.64	92.27	95.58	97.55	96.93
ܐ	99.39	98.79	97.99	97.32	100	98.80
ܐ	99.69	99.69	96.15	97.12	75.23	76.15
ܐ	100	99.35	97.48	87.39	97.62	98.41
ܐ	100	100	93.86	87.72	100	99.12
ܐ	100	100	99.11	88.39	98.80	95.18
ܐ	100	100	98.57	100	96.74	85.87
ܐ	100	100	97.30	100	86.36	95.45
ܐ	98.53	98.53	75.68	94.59	84.31	82.35
ܐ	100	100	97.14	97.14	95.65	97.10
ܐ	97.62	92.86	95.35	83.72	80.56	86.11
ܐ	100	100	97.67	100	97.62	97.62
ܐ	100	100	100	100	100	94.59
ܐ	94.00	100	100	100	100	100
ܐ	100	100	100	96.43	85.71	90.48
ܐ	100	100	100	100	73.91	69.57
ܐ	100	100	84.62	92.31	71.43	19.05
ܐ	100	100	50.00	0	80.00	100

⁸² The most common letters in Syriac, in decreasing order of frequency, are: *olaph* (13.9%), *waw* (10.1%), *nun* (9.6%), *yudh* (9.0%), *lomadh* (7.4%), *dolath* and *mim* (6.4%), *hé* (5.5%), *taw* (5.3%), *rish* (4.4%), *béth* (4.3%), *keph* (3.0%), *shin* (2.8%), *ayn* (2.6%), *héth* (2.3%), *phé* (1.5%), *qoph* (1.4%), *simkath* (1.3%), *gomal* (0.9%), *téth* (0.8%), *zayn* (0.6%), and *šodhé* (0.3%). [George Anton Kiraz, *Tūrāṣ Mamllā: A Grammar of The Syriac Language*, vol. 1: Orthography (Piscataway, NJ: Gorgias Press, 2012), 53–54.]

Table 41: Consonantal Accuracy Rates, Arranged in Descending Order of Accuracy

Estrangela- Language	Estrangela- Script	Serto- Language	Serto- Script	East Syriac- Language	East Syriac- Script
𐤀 100	𐤀 100	𐤁 100	𐤂 100	ܐ 100	ܐ 100
𐤃 100	𐤃 100	𐤄 100	𐤅 100	ܢ 100	ܢ 100
𐤆 100	𐤆 100	ܦ 100	ܥ 100	ܦ 100	ܦ 99.25
ܥ 100	ܥ 100	ܦ 100	ܥ 100	ܦ 100	ܦ 99.12
ܥ 100	ܥ 100	ܬ 99.11	ܦ 100	ܦ 98.87	ܬ 98.80
ܦ 100	ܦ 100	ܦ 98.97	ܦ 100	ܦ 98.80	ܦ 98.41
ܥ 100	ܥ 100	ܬ 98.72	ܦ 100	ܦ 97.62	ܥ 97.62
ܦ 100	ܦ 100	ܥ 98.57	ܬ 98.72	ܥ 97.62	ܦ 97.10
ܥ 100	ܥ 100	ܦ 97.99	ܦ 98.46	ܥ 97.55	ܥ 96.93
ܥ 100	ܥ 100	ܥ 97.86	ܦ 97.32	ܥ 96.74	ܥ 95.45
ܥ 100	ܥ 100	ܥ 97.67	ܥ 97.14	ܥ 95.65	ܥ 95.18
ܥ 100	ܥ 100	ܦ 97.48	ܥ 97.12	ܥ 94.05	ܥ 94.59
ܥ 100	ܥ 100	ܥ 97.30	ܥ 96.43	ܥ 88.95	ܥ 90.48
ܥ 100	ܥ 99.69	ܥ 97.14	ܥ 95.58	ܥ 86.35	ܥ 89.68
ܥ 99.69	ܥ 99.62	ܥ 96.15	ܥ 94.59	ܥ 85.71	ܥ 86.11
ܥ 99.62	ܥ 99.35	ܥ 95.35	ܥ 92.31	ܥ 84.31	ܥ 85.87
ܥ 99.39	ܥ 98.92	ܥ 93.86	ܥ 88.39	ܥ 80.56	ܥ 84.74
ܥ 99.09	ܥ 98.79	ܥ 92.27	ܥ 87.72	ܥ 80.00	ܥ 82.35
ܥ 98.53	ܥ 98.64	ܥ 84.62	ܥ 87.39	ܥ 75.23	ܥ 76.15
ܥ 97.62	ܥ 98.53	ܥ 75.68	ܥ 83.72	ܥ 73.91	ܥ 69.57
ܥ 94.00	ܥ 92.86	ܥ 50.00	ܥ 0.00	ܥ 71.43	ܥ 19.05

Table 42: Range of Letter Accuracy Results⁸³

	Maximum	Mean	One Standard Deviation	Two Standard Deviations	Minimum
Estrangela-Language					
Value	100%	99.43%	98.04%	96.66%	94.00%
Consonants with Equal or Higher Accuracy	14	16	19	20	21
Estrangela-Script					
Value	100%	99.35%	97.78%	96.20%	92.86%
Consonants with Equal or Higher Accuracy	13	16	20	21	21
Serto-Language					
Value	100%	93.83%	82.23%	70.63%	50.00%
Consonants with Equal or Higher Accuracy	4	17	19	20	21

⁸³ This table aims to show the overall distribution of consonantal accuracy in each type style and mode. The percentage values here do not necessarily represent actual accuracy rates but rather give a clearer understanding of the results through statistical analysis (specifically the mean and standard deviation). The counts are the number of consonants that have accuracy rates that are equal to or higher than the statistical value.

Serto-Script					
Value	100%	91.23%	69.75%	48.28%	0.00%
Consonants with Equal or Higher Accuracy	7	16	20	20	21
East Syriac-Language					
Value	100%	90.74%	80.99%	71.23%	71.43%
Consonants with Equal or Higher Accuracy	4	13	16	21	21
East Syriac-Script					
Value	100%	88.50%	70.54%	52.57%	19.05%
Consonants with Equal or Higher Accuracy	2	14	19	20	21
Overall Type Styles and Modes					
Value	100%	93.80%	86.34%	78.88%	71.11%
Consonants with Equal or Higher Accuracy	0	15	19	19	21

BIBLIOGRAPHY

Acta Martyrum et Sanctorum Syriace. Volume 1. Edited Paul Bedjan. Hildesheim: Georg Olms Verlagsbuchhandlung, 1968.

Antioch, Severus of. *Les Homiliae Cathedrales de Sévère d'Antioche: traduction Syriaque de Jacques d'Édesse* (Homélies LII–LVII). Edited and translated by R. Duval. *Patrologia Orientalis* 4. Paris: Librairie de Paris, 1908.

- Coakley, J. F. “Edward Breath and the Typography of Syriac.” *Harvard Library Bulletin*. New Series. Volume 6, no. 4 (1995): 4–64.
- . *The Typography of Syriac: A historical catalogue of printing types, 1537-1958*. New Castle, DE, and London: Oak Knoll Press and The British Library, 2006.
- Clocksins, William F., and Prem Fernando. “Towards Automatic Transcription of Estrangelo Script.” *Hugoye: Journal of Syriac Studies* 6, no. 2 (2003): 249–68.
- FAQ. GitHub. <https://github.com/tesseract-ocr/tesseract/wiki/FAQ>. Accessed 30 July 2018.
- Hatch, W. H. P. *An Album of Dated Syriac Manuscripts*. Boston, MA: The American Academy of Arts and Sciences, 1946.
- Hebraeus, Bar. *Le Livre des Splendeurs de Grégoire Barhebraeus*. Edited by Alex Moberg. London: Humphrey Milford, 1922.
- “ImproveQuality.” GitHub, Last updated April 21, 2018. <https://github.com/tesseract-ocr/tesseract/wiki/ImproveQuality>. Accessed 30 July 2018.
- Kessel, Grigory. 21 April 2017. Message to hugoye-list. <https://groups.yahoo.com/neo/groups/hugoye-list/conversations/messages/8069>. Accessed 30 July 2018.
- Kiraz, George Anton. *Turrās Mamllā: A Grammar of the Syriac Language*. Volume 1: Orthography. Piscataway, NJ: Gorgias Press, 2012.
- Merv, Išo‘dad de. *Commentaire d’Išo‘dad de Merv Sur l’Ancien Testament I. Genèse*. Edited J.-M. Vosté and C. Van den Eynde. CSCO 126. Scriptores Syri 67. Louvain: Imprimerie Orientaliste L. Durbecq, 1950.
- The Patriarchal Journal of the Syrian Orthodox Patriarchate of Antioch and All the East* 40, nos. 211–212–213 (2002).

- Romanov, Maxim, Matthew Thomas Miller, Sarah Bowen Savant, and Benjamin Kiessling. "Important New Developments in Arabographic Optical Character Recognition (OCR)." March 2017. <https://arxiv.org/abs/1703.09550>. Accessed 30 July 2018.
- Tesseract language-specific training guide. GitHub. Last revised 19 July 2018. <https://github.com/tesseract-ocr/tesseract/blob/master/src/training/language-specific.sh>. Accessed 30 July 2018.
- "Tesseract OCR." *Google Open Source*. <https://opensource.google.com/projects/tesseract>. Accessed 30 July 2018.
- "TESSERACT (1) Manual Page." GitHub. Last updated 2 July 2018. <https://github.com/tesseract-ocr/tesseract/blob/master/doc/tesseract.1.asc#language>s. Accessed 30 July 2018.
- "TrainingTesseract 4.00." GitHub. Last revised 15 July 2018. <https://github.com/tesseract-ocr/tesseract/wiki/TrainingTesseract-4.00>. Accessed 30 July 2018.
- Tse, Elizabeth, and Josef Bigun. "A base-line character recognition for Syriac-Aramaic." *2007 IEEE International Conference on Systems, Man and Cybernetics*. Montreal, Quebec: IEEE, 2007. Pp. 1048–1055. doi: 10.1109/ICSMC.2007.4414012. <https://ieeexplore.ieee.org/document/4414012/authors>.