Optical Character Recognition for Handwritten Syriac Text

Presented by: Ameer Hameed Majeed

Supervisor: Dr. Hossein Hassani

Research ID: 10942

Date: 23/5/2024

University of
KURDISTAN
Hewlêr

Table of Contents

- Introduction
- Syriac Language – An Overview
- Methodology
- Experiments and Results
- Conclusion
- Q&A

Introduction

What is OCR?

Optical character recognition (OCR) is a computational technique that is used to recognize text from scanned documents and digital images.

Types of OCR:

- Online OCR: automatic recognition while text is being written.
- Offline OCR: recognition is performed on handwritten or typed documents.

Problem Statement

According to UNESCO World Atlas of Languages (2010), most Aramaic dialects that use the Syriac alphabet are classified as **endangered**.

Example: Iraqi Federal Government recognized Syriac as an official language in its constitution after 2005 (Iraqi Constitution, 2005).

There is an evident shortage in digital services and academic research in the field of artificial intelligence.

Objectives

- To build an OCR model for Syriac in order to recognize handwritten text
- To create custom dataset that contains handwritten samples of Syriac sentences
- To assist in digitizing the language and preserving it from extinction

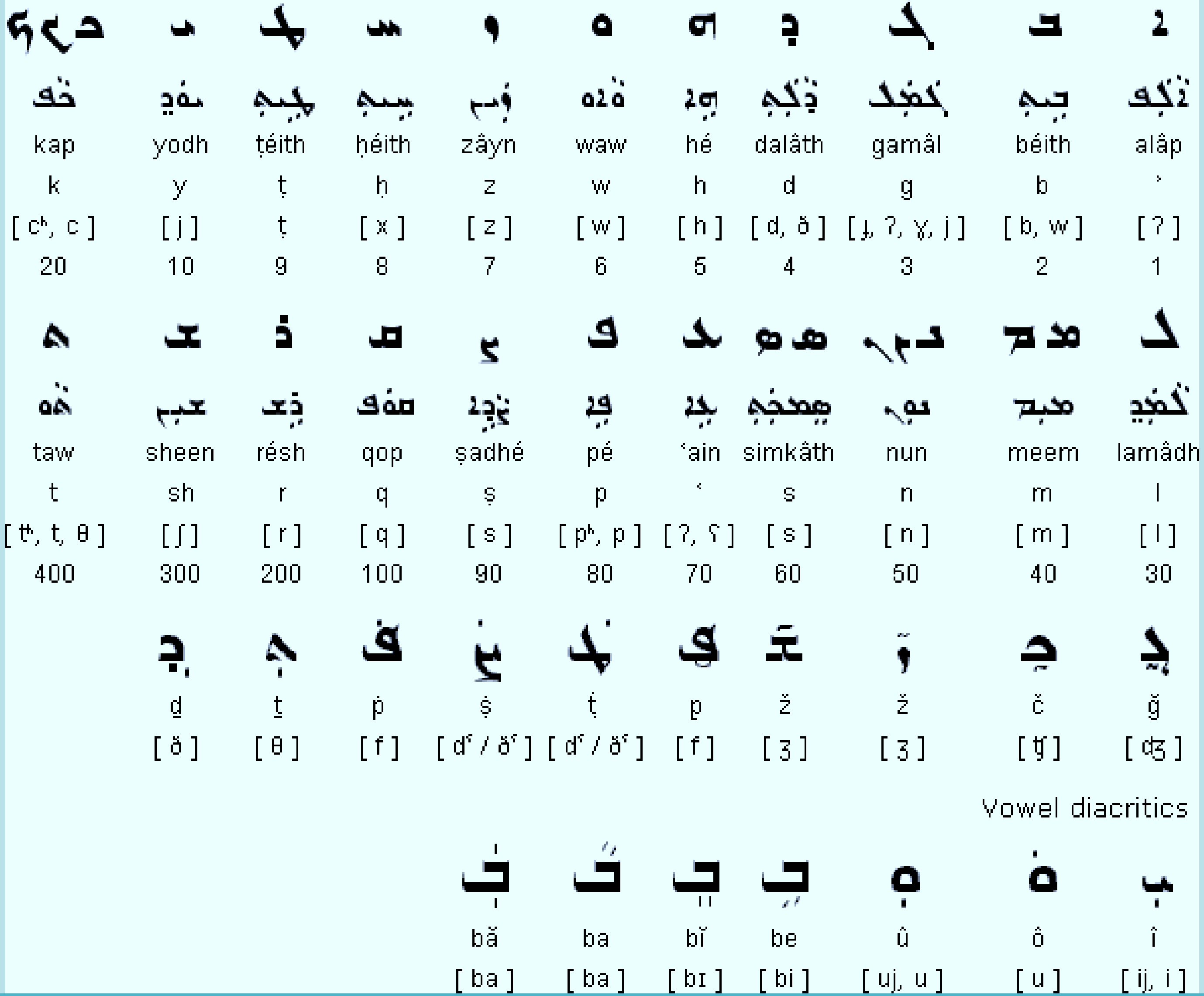
Syriac Language – An Overview

Syriac Language or Dialect?

Syriac is a dialect of the Aramaic language from the greater family of Semitic languages which is said to have originated in or around Edessa (Butts, 2011).

The Syriac alphabet consist of 22 letters and is written from right to left in a cursive style (Ackroyd & Evans, 1970).

There are three main writing systems in Syriac, namely Estrangela, East Syriac(Madnḥāyā), and West Syriac (Serṭ ā).

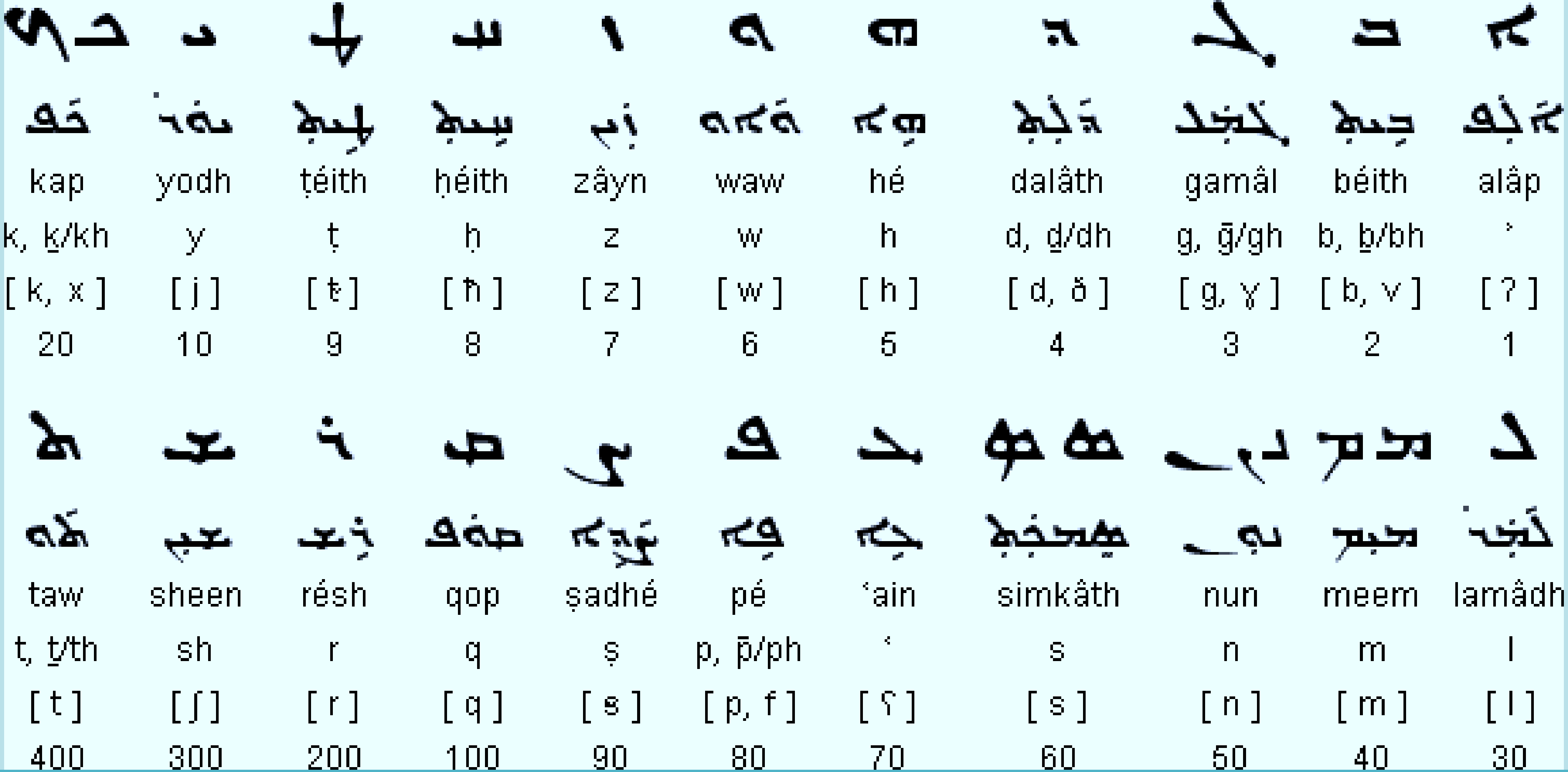


Letter	Name	Transliteration	IPA	Value
ܐ	alâp	ʾ	[ʔ]	1
ܒ	béith	b	[b, w]	2
ܓ	gamâl	g	[ɟ, ʔ, ɣ, j]	3
ܕ	dalâth	d	[d, ð]	4
ܗ	hé	h	[h]	5
ܘ	waw	w	[w]	6
ܙ	zâyn	z	[z]	7
ܚ	ḥéith	ḥ	[x]	8
ܛ	ṭéith	ṭ	ṭ	9
ܝ	yodh	y	[j]	10
ܟ	kap	k	[cʰ, c]	20
ܠ	lamâdh	l	[l]	30
ܡ	meem	m	[m]	40
ܢ	nun	n	[n]	50
ܣ	simkâth	s	[s]	60
ܥ	ʿain	ʿ	[ʔ, ʕ]	70
ܦ	pé	p	[pʰ, p]	80
ܨ	ṣadhé	ṣ	[s]	90
ܩ	qop	q	[q]	100
ܪ	résh	r	[r]	200
ܫ	sheen	sh	[ʃ]	300
ܬ	taw	t	[tʰ, t, θ]	400
ܓ̰		ğ	[dʒ]	
ܟ̰		č	[tʃ]	
ܙ̰		ž	[ʒ]	
ܫ̰		ž	[ʒ]	
ܦ̮		p̣	[f]	
ܛ̇		ṭ̇	[dˤ / ðˤ]	
ܨ̇		ṩ	[dˤ / ðˤ]	
ܦ̄		ṗ	[f]	
ܬ݂		ṯ	[θ]	
ܕ݂		ḏ	[ð]	

Vowel diacritics

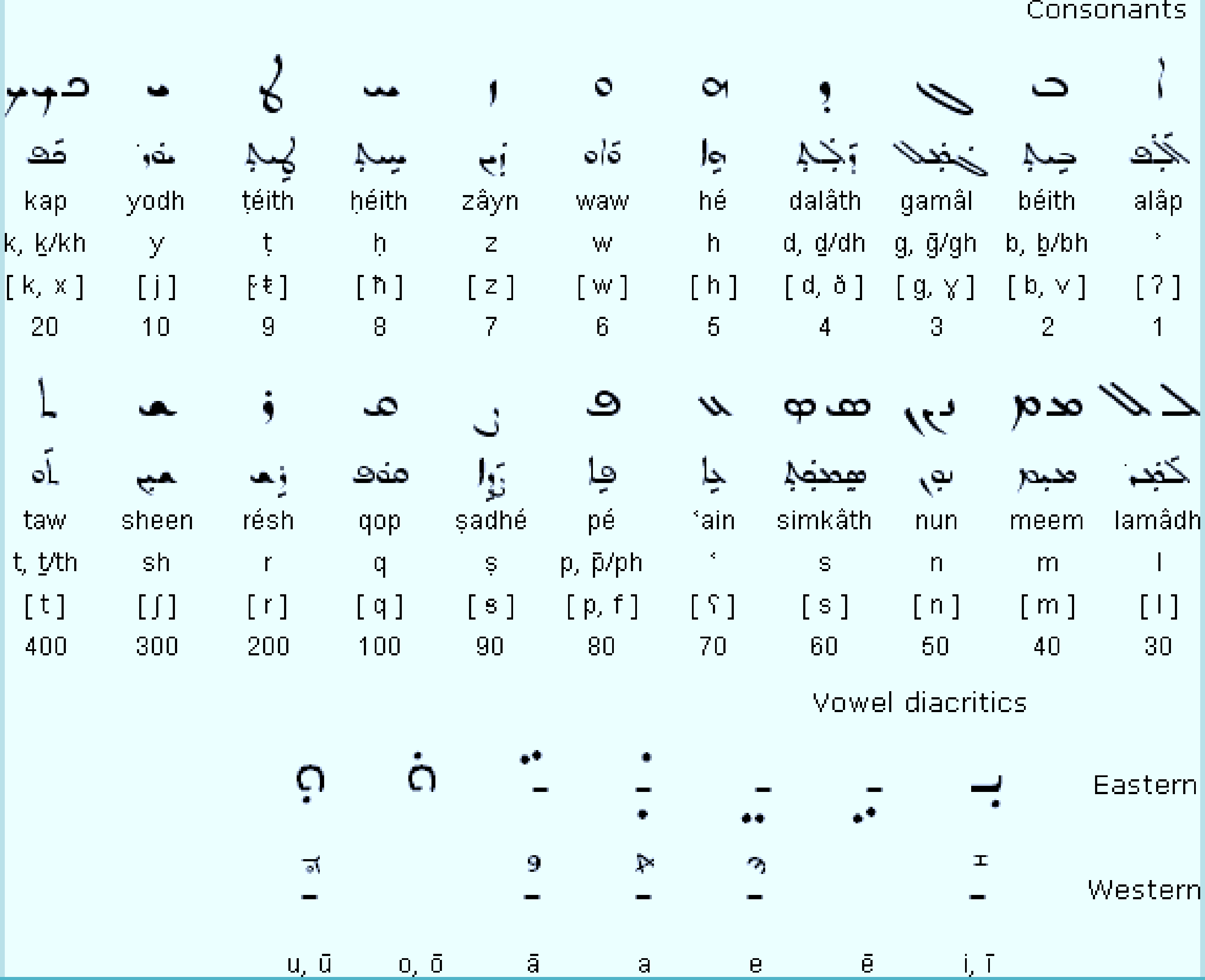
Letter	Transliteration	IPA
ܒܝ	î	[ij, i]
ܘܿ	ô	[u]
ܘܼ	û	[uj, u]
ܒܸ	be	[bi]
ܒܹ	bĭ	[bɪ]
ܒܵ	ba	[ba]
ܒܲ	bă	[ba]

Figure 2.1: East Syriac (Madnḥāyā) Script (Omniglot, 2023)



ܟ ܟ	ܝ	ܛ	ܚ	ܙ	ܘ	ܗ	ܕ	ܓ	ܒ	ܐ
ܟܦ	ܝܘܕ	ܛܝܬ	ܚܝܬ	ܙܝܢ	ܘܐܘ	ܗܐ	ܕܠܬ	ܓܡܠ	ܒܝܬ	ܐܠܦ
kap	yodh	ṭéith	ḥéith	zâyn	waw	hé	dalâth	gamâl	béith	alâp
k, k̲/kh	y	ṭ	ḥ	z	w	h	d, d̲/dh	g, ḡ/gh	b, b̲/bh	ʾ
[k, x]	[j]	[tˁ]	[ħ]	[z]	[w]	[h]	[d, ð]	[g, ɣ]	[b, v]	[ʔ]
20	10	9	8	7	6	5	4	3	2	1
ܬ	ܫ	ܪ	ܩ	ܨ	ܦ	ܥ	ܣ ܣ	ܢ ܢ	ܡ ܡ	ܠ
ܬܘ	ܫܝܢ	ܪܝܫ	ܩܘܦ	ܨܕܐ	ܦܐ	ܥܐ	ܣܡܟܬ	ܢܘܢ	ܡܝܡ	ܠܡܕ
taw	sheen	résh	qop	ṣadhé	pé	ʿain	simkâth	nun	meem	lamâdh
t, t̲/th	sh	r	q	ṣ	p, p̄/ph	ʿ	s	n	m	l
[t]	[ʃ]	[r]	[q]	[sˁ]	[p, f]	[ʕ]	[s]	[n]	[m]	[l]
400	300	200	100	90	80	70	60	50	40	30

Figure 2.2: Estrangela Script (Omniglot, 2023)



Consonants

ܟ	ܝ	ܛ	ܚ	ܙ	ܘ	ܗ	ܕ	ܓ	ܒ	ܐ
ܟܦ	ܝܘܕ	ܛܝܬ	ܚܝܬ	ܙܝܢ	ܘܘ	ܗܐ	ܕܠܬ	ܓܡܠ	ܒܝܬ	ܐܠܦ
kap	yodh	ṭéith	ḥéith	zâyn	waw	hé	dalâth	gamâl	béith	alâp
k, ḵ/kh	y	ṭ	ḥ	z	w	h	d, ḏ/dh	g, ḡ/gh	b, ḇ/bh	ʾ
[k, x]	[j]	[tˤ]	[ħ]	[z]	[w]	[h]	[d, ð]	[g, ɣ]	[b, v]	[ʔ]
20	10	9	8	7	6	5	4	3	2	1

ܬ	ܫ	ܪ	ܩ	ܨ	ܦ	ܥ	ܣ	ܢ	ܡ	ܠ
ܬܘ	ܫܝܢ	ܪܝܫ	ܩܘܦ	ܨܕܐ	ܦܐ	ܥܐ	ܣܡܟܬ	ܢܘܢ	ܡܝܡ	ܠܡܕ
taw	sheen	résh	qop	ṣadhé	pé	ʿain	simkâth	nun	meem	lamâdh
t, ṯ/th	sh	r	q	ṣ	p, p̄/ph	ʿ	s	n	m	l
[t]	[ʃ]	[r]	[q]	[sˤ]	[p, f]	[ʕ]	[s]	[n]	[m]	[l]
400	300	200	100	90	80	70	60	50	40	30

Vowel diacritics

ܘܼ	ܘܿ	◌ܵ	◌ܲ	◌ܸ	◌ܹ	ܒܝܼ	Eastern
◌ܽ		◌ܳ	◌ܰ	◌ܶ		◌ܺ	Western
u, ū	o, ō	ā	a	e	ē	i, ī	

Figure 2.3: West Syriac (Serṭ ā) Script (Omniglot, 2023)

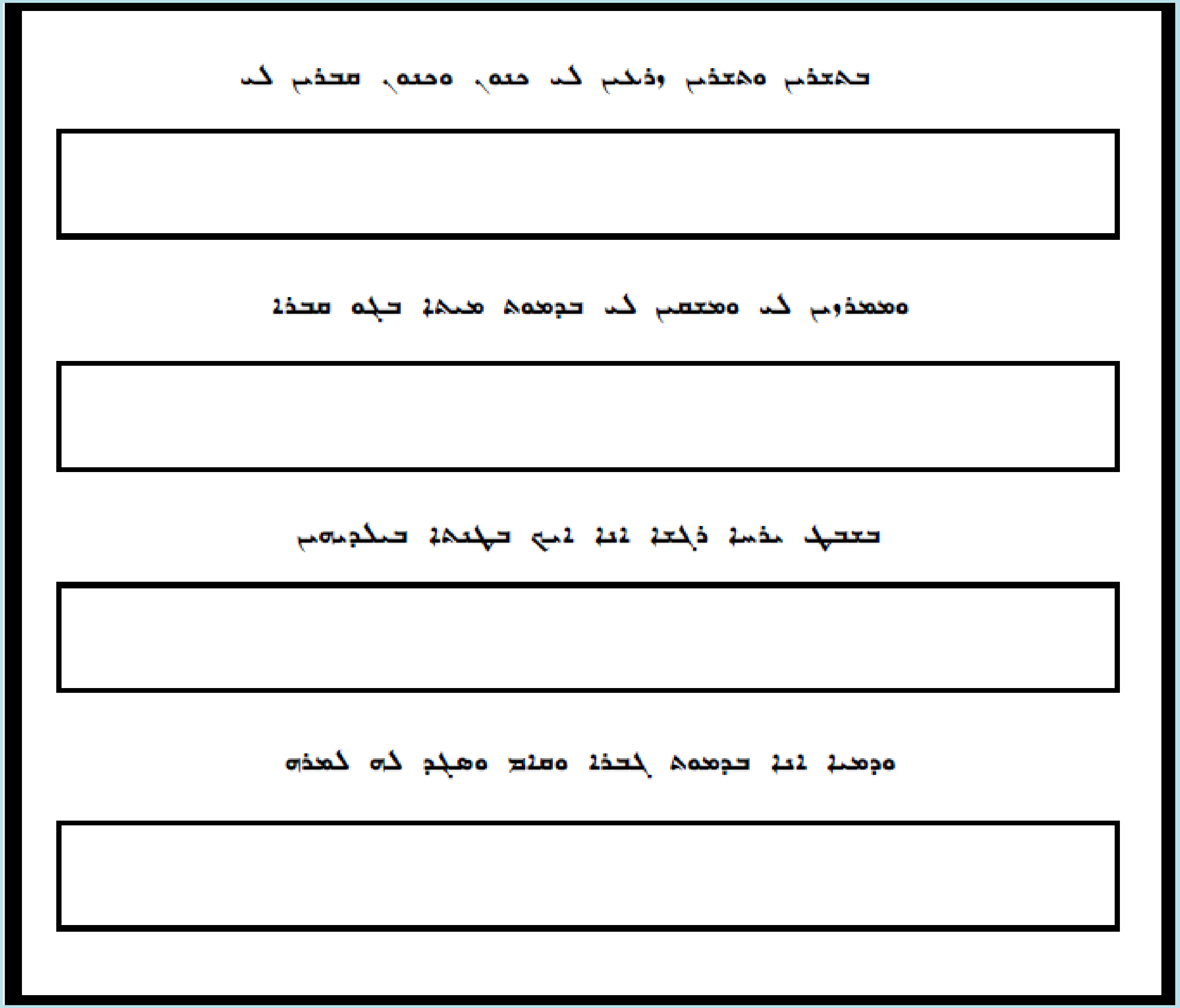
Methodology

Data Collection

A handwritten Syriac dataset will be collected from scratch.

The dataset template form will consists of multiple sentences with bounding boxes beneath them for input.

A pilot will be performed within university students who are capable of reading and writing in Syriac.



[illegible]

[illegible]

[illegible]

[illegible]

Figure 3.1: Sample of a page from the dataset template form

Preprocessing

The bounding box of each sentence will be extracted from the form paper, including it’s content.

Sentence images will be converted to grayscale and binarization will be required through using a thresholding algorithm(Gonzalez & Woods, 2018).

$$g(x, y) = \begin{cases} 1 & f(x, y) > T \\ 0 & f(x, y) \leq T \end{cases}$$

T = threshold value

Tesseract-OCR

Tesseract is an **open-source OCR engine** that was originally developed at Hewlett-Packard (HP) between 1985 and 1994 (Tesseract-OCR, 2024).

From 2006 until November 2018 it was developed by Google.

It’s currently being maintained by community contributors.

Tesseract supports and recognizes more than 100 languages, and can be trained to detect other languages.

Tesseract-OCR (cont.)

Tesseract uses **long short-term memory (LSTM)** neural network architecture.

Tesseract is equipped with trained language-specific OCR models which can be **fine-tuned** to recognize samples of newer fonts and writing styles.

> *“When the target dataset is significantly smaller than the base dataset, transfer learning can be a powerful tool to enable training a large target network without overfitting.” (Yosinki et al., 2014)*

Evaluation

Character error rate (CER):

S = number of substitutions,

D = number of deletions,

I = number of insertions,

N = total number of characters

$$CER = \frac{S + D + I}{N} \times 100$$

Experiments and Results

KHAMIS Dataset

This newly collected dataset is based on a poem of Khamis bar Qardahe, who was a 13th century East Syriac poet and priest (Mengozzi, 2011).

It consists of **624 handwritten image samples** of 20 different sentences (verses) of the **East Syriac script**.

Each sentence image is accompanied with a text file containing it’s ground-truth value.

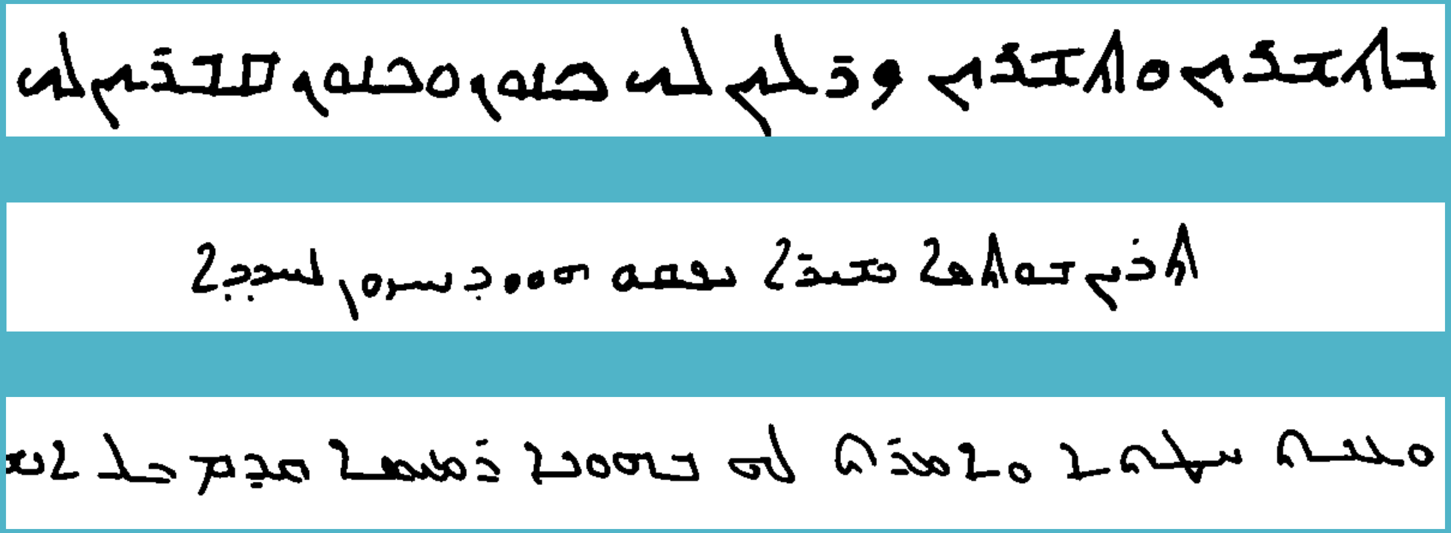


Figure 4.1: Three extracted image samples from KHAMIS Dataset

Training and Evaluating Model

The models have been trained using Tesseract 5’s tesstrain training tool on a Lenovo Thinkpad X1 Yoga, Intel Core i7-8550U CPU and 16 GB RAM

Model	Train/Eval Split	Character Error Rate (Training)	Character Error Rate (Evaluation)
Model 1: esyr	90/10	1.610%	9.864%
Model 2: esyr_lesstrain	80/20	1.402%	8.963%
Model 3: esyr_short	70/30	1.097%	10.498%

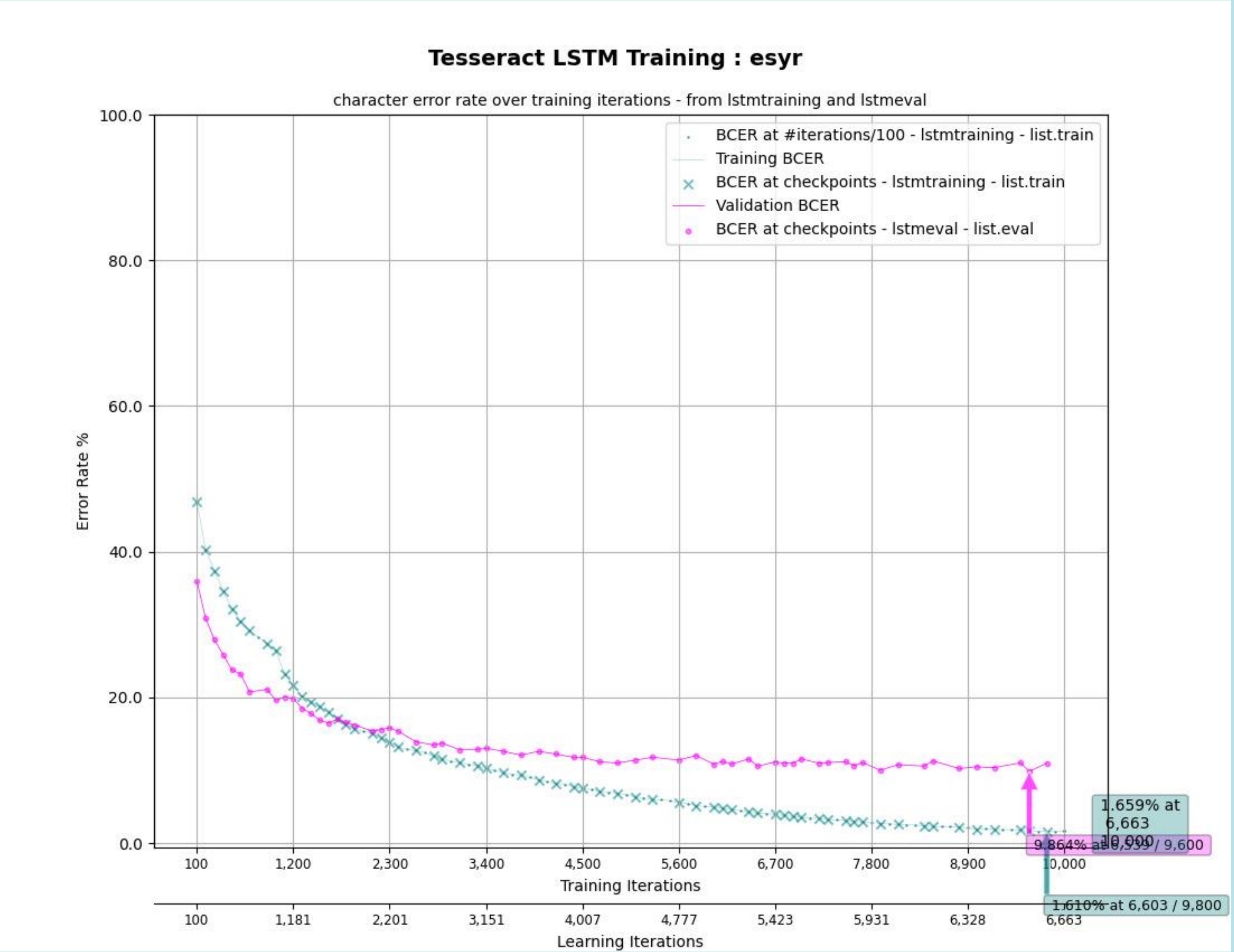


Figure 4.2: Tesseract LSTM Training - esyr

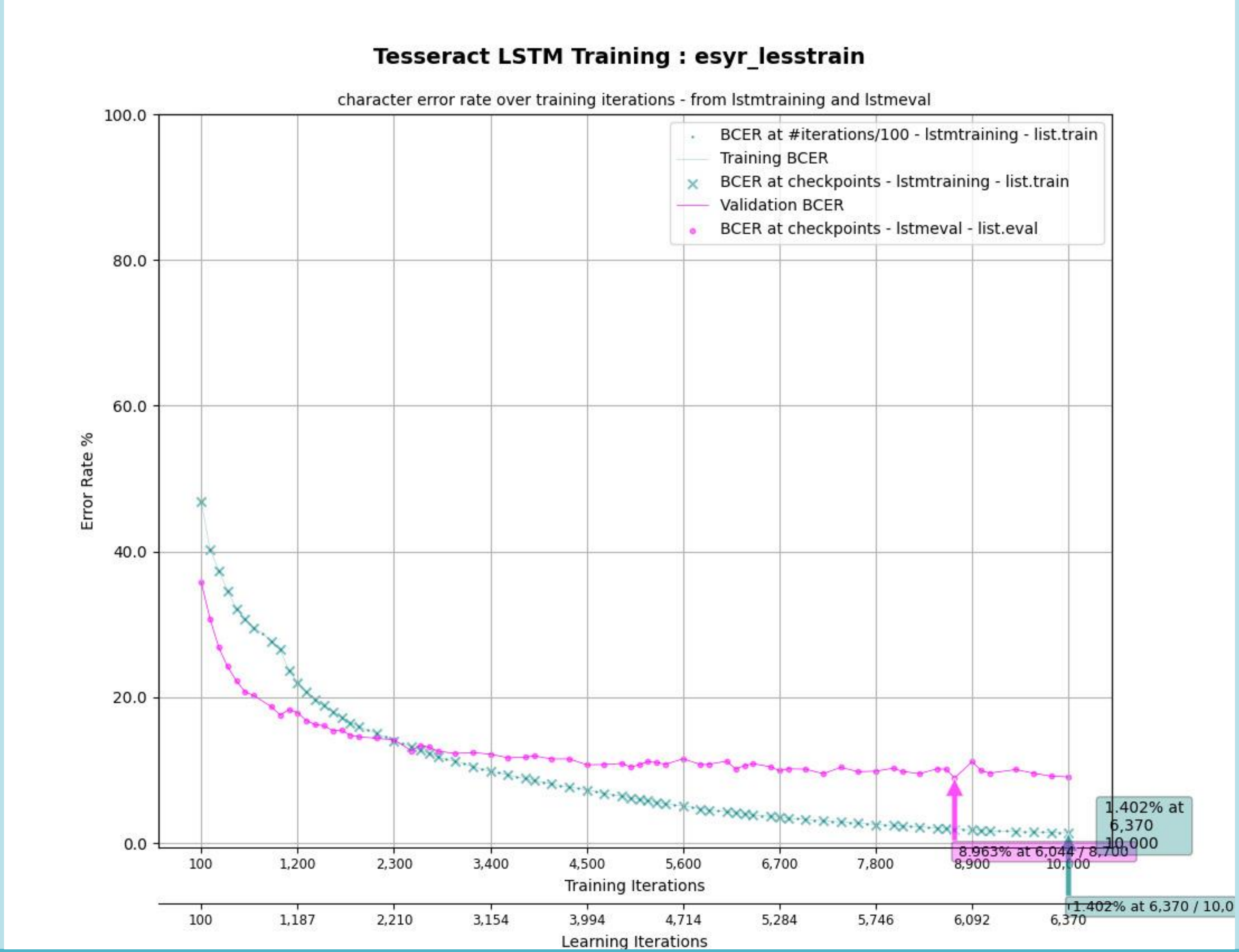


Figure 4.3: Tesseract LSTM Training - esyr_lesstrain

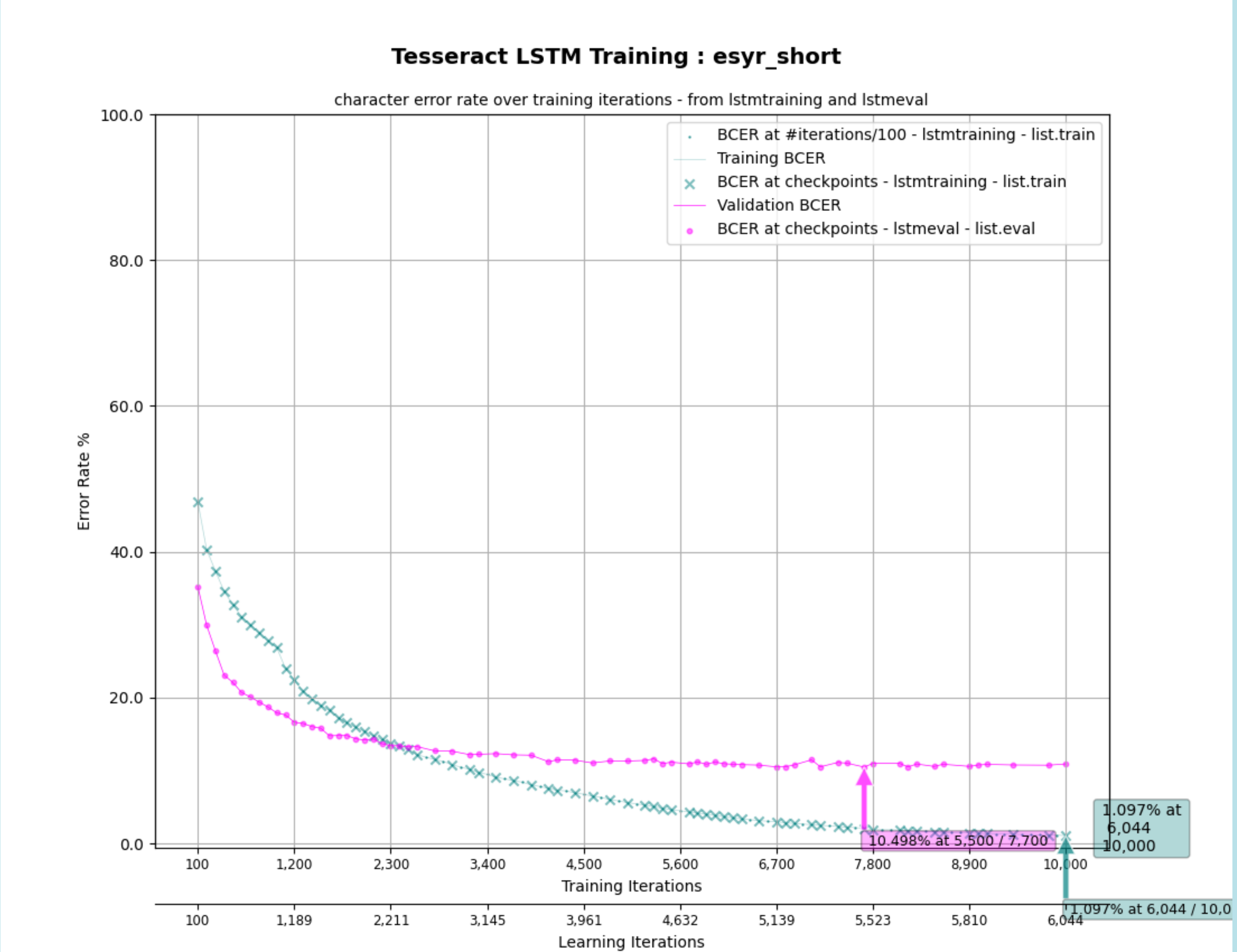


Figure 4.4: Tesseract LSTM Training - esyr_short

Testing the Model

The performance of the model will be assessed on a test dataset that contains 12 sentence images and one paragraph image (13 samples in total).

The mean of the character error rate (CER) and word error rate (WER) of all samples will be calculated.

$$\overline{X} = \frac{\sum X_i}{N}$$

X_i = CER/WER of sample
N = No. of Samples

Testing the Model (cont.)

Model	Character Error Rate (Test)	Word Error Rate (Test)
Default Model: syr	63.48%	136.84%
Model 1: esyr	32.29%	75.39%
Model 2: esyr_lesstrain	31.12%	75.78%
Model 3: esyr_short	32.0%	78.13%

Conclusion

Future Work

- More data collection
- Recognition of other writing systems: (Estrangela and West Syriac)
- Diacritics to be included
- Experiment with different algorithms and training parameters

References

UNESCO, 2010. World atlas of languages.
Available at: https://en.wal.unesco.org/discover/languages?text=aramaic

of Iraq, T. R., 2005. Constitution project.
Available at: https://www.constituteproject.org/constitution/Iraq_2005

Butts, A. M., 2011. Syriac language.
Available at: https://gedsh.bethmardutho.org/Syriac-Language

Ackroyd, P. R. & Evans, C. F., eds, 1970. The Cambridge History of the Bible: Volume 1, From the Beginning to Jerome, Cambridge University Press.

Omniglot, 2023. Syriac alphabet.
Available at: https://www.omniglot.com/writing/syriac.htm

Gonzalez, R. & Woods, R., 2018. Digital Image Processing, Pearson.

Tesseract-OCR, 2024. Tesseract-ocr.
Available at: https://github.com/tesseract-ocr/tesseract

Yosinski, J., Clune, J., Bengio, Y. & Lipson, H., 2014. How transferable are features in deep neural networks?.

Mengozzi, A., 2011. Khamis bar qardahe.
Available at: https://gedsh.bethmardutho.org/Khamis-bar-Qardahe

Thank you and ܒܣܝܡܐ

Any Questions?