Translating Old Persian cuneiform by artificial intelligence (AI)

Author: Shaghayegh Rahmani (Melanee)
Email: melaneepython@gmail.com
GitHub: https://github.com/Electronic-Old-Persian-Library
ORCID ID: 0009-0003-7317-639X
Iran, Tehran

Abstract
Old Persian cuneiform, the script used in ancient Persian inscriptions, maintains significant historical value. Optical Character Recognition (OCR) and Natural Language Processing (NLP), offer a promising approach to translating these ancient scripts. The paper discusses the development of artificial intelligence (AI) models to translate Old Persian cuneiform inscriptions to modern languages for the first time in the world.
Previous efforts using the Tesseract model were inadequate for handwritten cuneiform, prompting the development of a new model, "easyocr old persian."
Future work includes enhancing image pre-processing techniques, training the model with more data, and developing an NLP model for better translation. The author proposed applying these methods to other ancient languages and also using AI to reconstruct fragmented texts by matching broken tablets or inscriptions. The project emphasized the potential of AI in archaeology and historical linguistics, aiming to preserve and understand ancient cultures through technology.

Keywords: Old Persian, translating, cuneiform, artificial intelligence, OCR, NLP, LLM

1. Introduction

Old Persian cuneiform is a semi-alphabetic cuneiform script that was the primary script for the Old Persian language. Texts written in this
cuneiform were found in Persepolis, Susa, Hamadan, Armenia, and along the Suez Canal. They were mostly inscriptions from the time period of Darius the Great and his son Xerxes the Great kings of Achaemenid Empire (Kent, 1953).
The Achaemenid Persian empire was the largest that the ancient world had seen, extending from Anatolia and Egypt across western Asia to northern India and Central Asia. Its formation began in 550 B.C.

Being able to read and decipher ancient scripts has always been of much interest and importance for human-being. Many valuable historical secrets were revealed by archaeologists through the time, which made us aware of the culture and civilization of our ancestors.

In this regard, artificial intelligence (AI) can play an important role in translating ancient scripts. AI is the simulation of human intelligence processes by machines, especially computer systems. The pursuit of human-like intelligence must be in part an empirical science related to psychology, involving observations and hypotheses about actual human behavior and thought processes; a rationalist approach (Russell et al., 2016) .
Optical Character Recognition (OCR) is a technology that converts different types of documents, such as scanned paper documents, PDF files, or images captured by a digital camera, into digital text (Schantz, 1982). OCR leverages AI to enhance its ability to accurately recognize and convert different fonts and handwriting into readable data. AI, particularly through machine learning and neural networks, improves OCR's pattern recognition, accuracy, and adaptability.
OCR models combine traditional image processing techniques with advanced deep learning architectures to accurately detect and recognize text from images. This architecture involves two key components: text detection and text recognition models, which work together to ensure precise text extraction (see image 1). In this paper, for the "yolo_cnn_old_persian" model, I have utilized the "YOLOv8" model for text detection and a convolutional neural network (CNN) model for text recognition.

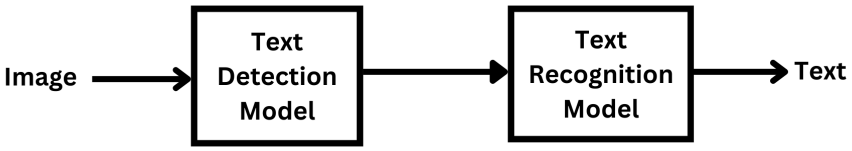

Image 1. A simple OCR architecture

Preserving the history of my country, Iran, is a highly important and valuable endeavor. I often think that if no one in the world can translate the Old Persian language, how will the next generation translate new found inscriptions or tablets? Therefore, this concern has inspired me to develop a new AI model aimed at keeping this ancient language alive forever. To the best of my knowledge, I am pioneering the development of AI models for the Old Persian language on this scale globally (see the GitHub organisation of my project).

2. Previous research

In June 2024, Professor Enrique Jiménez and his team at Ludwig Maximilian University (LMU) developed cutting-edge artificial intelligence and computer vision techniques to analyze and translate Babylonian and Akkadian inscriptions (Cobanoglu et al., 2024). They have developed several tools and algorithms to facilitate this process, which are available in their GitHub repositories and electronic babylonian library (eBL). My project is inspired by the eBL project. Professor Enrique Jiménezhas has developed models for Babylonian cuneiform but I am going to develop my models for Old Persian cuneiform.

In November 2017, Seyed Muhammad Hossein Mousavi and Vyacheslav Lyashenko developed an OCR model for Old Persian cuneiform using the Tesseract engine. However, the training code and primary dataset were not made available (Mousavi et al., 2017), prompting me to develop a new OCR model from scratch. Additionally, the Tesseract model is not well-suited for handwritten scripts like Old Persian cuneiform, being more appropriate for computer fonts. To address this, I created a new OCR model based on the EasyOCR model, which utilizes PyTorch neural networks, making it more effective for recognizing handwritten scripts (Rosebrock et al., 2020).

3. Proposed approach

3. 1 Primary data

The primary raw data is collected from all over the world. For instance, the British museum collection, cuneiform digital library Initiative (CDLI), Livius, Norman Sharp's book (Sharp, 1970), my personal photography from the national museum of Iran and Takht-e-Jamshid (Persepolis).

3. 2 Data pipeline

In the initial stage, Old Persian cuneiform will be converted into English transcription text using an OCR model. In the second stage, this English transcription will serve as an input for a natural language processing (NLP) or large language model (LLM), which will translate it into modern languages (see image 2). Essentially, the NLP model functions as a machine translation tool.

Image 2. Data pipeline for this paper

3. 3 Developing OCR models

I have developed three OCR models in this project:
- yolo_cnn_old_persian
- tessearct_old_persian
- easyocr_old_persian

The “easyocr_old_persian” model has been developed from scratch and trained using 37 images of the last 12 lines of the great Darius's DPd inscription from Persepolis. This model translates Old Persian cuneiform into English transcription. The model is available in the Old-Persian-Cuneiform-OCR repository on GitHub, which is part of the Electronic Old Persian Library (EOPL) organization.

The "tessearct_old_persian" model, developed by S. Muhammad Hossein Mousavi, is a pre-trained OCR model. I have implemented code in Python to evaluate this pre-trained model. It is designed to work exclusively with black and white images using computer fonts, as it lacks any image pre-processing capabilities. In contrast, the "easyocr_old_persian" model includes its own image pre-processing code, making it more versatile.

The "yolo_cnn_old_persian" model is also in progress and not yet complete.

Currently, my models are still under development. In the future, I plan to train the "easyocr_old_persian" model with a larger dataset of real images to improve its accuracy.

4. Results and discussion

4. 1 Stage 1: Evaluating the “easyocr_old_persian” model

To evaluate the performance of the “easyocr_old_persian” model, I input image 3 into this model:

Input:

Image 3. Input cuneiform to “easyocr_old_persian” model

```
import easyocr
reader=easyocr.Reader(['op'],recog_network='custom_example')

text = ''
results=reader.readtext('/content/Old-Persian-Cuneiform-OCR/other/
custom images/DPd.png')
for result in results:
 text = text + result[1] +  ' '

text = text[:-1]
print(text)
```

Output:

dahaiy : aitavabis : dayamiy : ma : a : daya : jayatuv : haya : Darayabis pabis : daya : aitaum : da : daha : datuv : xsa : sa : haina : ha : hainya : da : aiy : aita : ba: dam : a : daya : dam : dahyam : ba: dam dadyaiy : ditaiy : datyatuv :

4. 2 Stage 2: Translating the English transcription

I did translate that English transcription to modern languages by a large language model (Chat-GPT, version 4).

4. 2. 1 Translate to modern Persian:

"دهائی که به دست من داریوش شاه ساخته شده است.
این فرمان من است، که بر پایه اهورامزدا است.
باشد که اهورامزدا از من و پادشاهی‌ام محافظت کند.
این است آنچه که من انجام داده‌ام؛
این سخن من است، که بر پایه اهورامزدا است.
باشد که اهورامزدا از این کار من محافظت کند."

4. 2. 2 Translate to modern English:

*"*Dahaiy *made by me, Dariush Shah.*
This is my decree, based on Ahura Mazda.
May Ahuramazda protect me and my kingdom.
This is what I have done;
This is my word, based on Ahura Mazda.
May Ahuramazda protect me from this work."

In the future, I plan to develop my own NLP or LLM model for translating English transcriptions, rather than using ChatGPT
To see more translations, check the notebooks of this project on GitHub.

4.3 Training results

Training results for “easyocr_old_persian” model are show in table 1 and image 4:

Step	num_iter	valInterval	Confidence score	Execution time(min)
1	300	200	0.0032	5
2	3000	2000	0.5	15
3	50000	10000	0.45	102

Table 1. Training results for “easyocr_old_persian” model

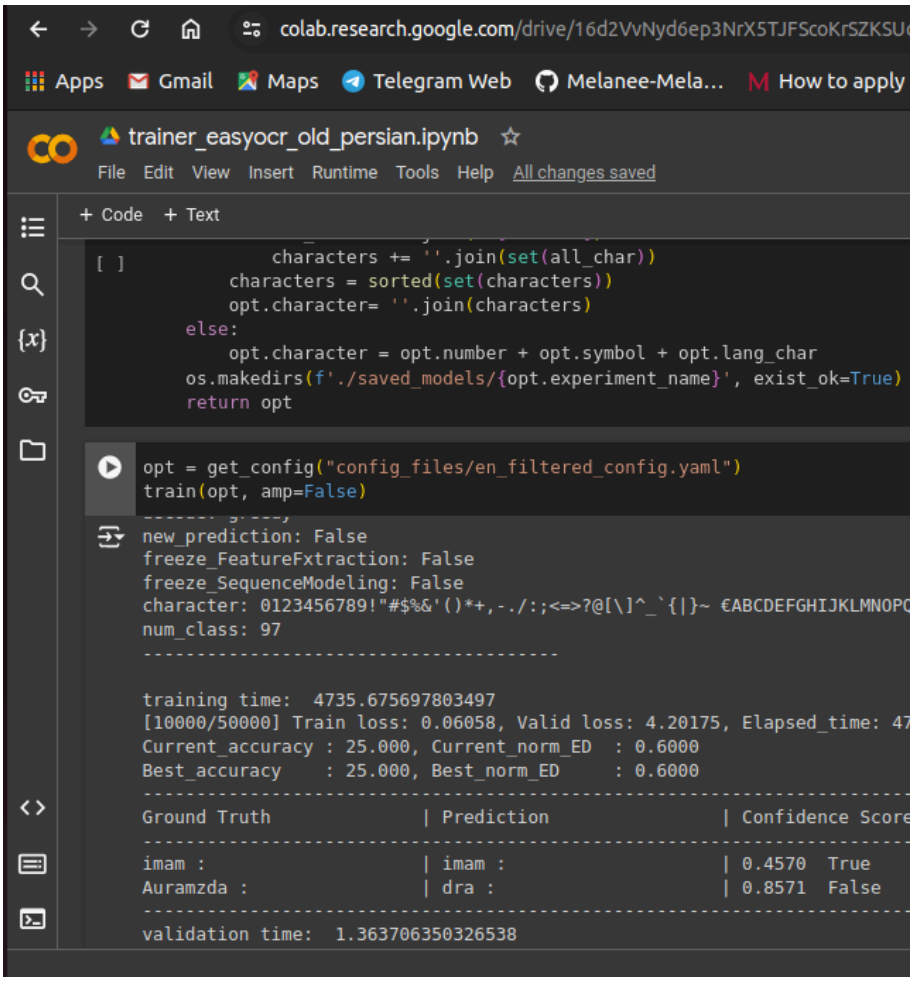

Image 4. Training results for “easyocr_old_persian” model

All my models have been developed on Google Colab notebook infrastructure (see table 2):

Operating system	version	GPU	CPU	RAM
Linux, Ubuntu	22.04 LTS	T4GPU	-	16G

Table 2. Infrastructure of my Google Colab notebook

5. Future work

- Developing image pre-processing code for the “yolo_cnn_old_persian” model to convert real images to black and white (see image 5) and omit background noise. Also, pre-processing techniques such as grayscale conversion, noise reduction, binarization, adding salt and pepper, dilation, fuzzy edge detection (Mousavi et al., 2017) are very helpful to observe unclear inscriptions or tablets.

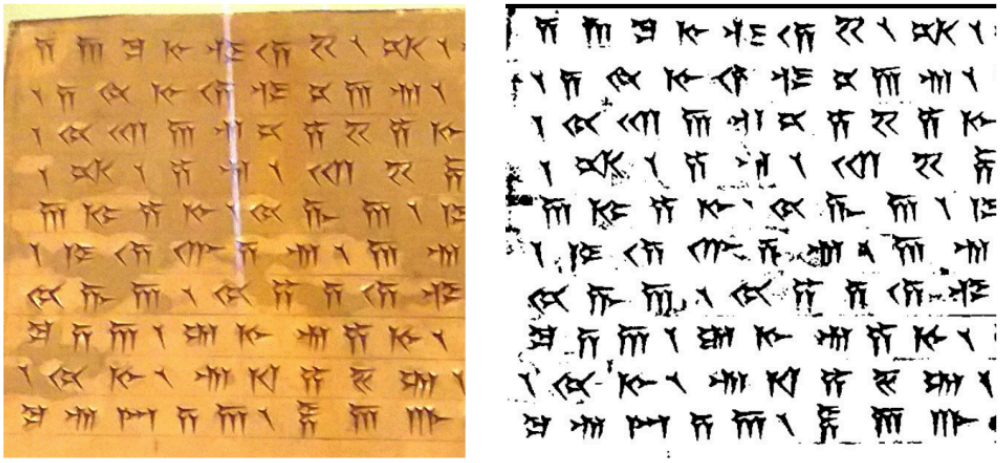

Image 5. Image pre-processing on a real inscription

- Training “easyocr old persian” model with more real inscriptions
- Developing NLP or LLM to translate English transcriptions to modern languages

6. Future Research

- The "easyocr_old_persian" model has the potential to be implemented for various other ancient languages, such as Akkadian, Pahlavi, Avestan, Sogdian and so on. This paper is very initiative and can open a new world of research for each ancient language.

- By translating Old Persian cuneiform in trilingual inscriptions (see image 6) or tablets, we can translate Babylonian and Elamite cuneiform because they have same meaning.

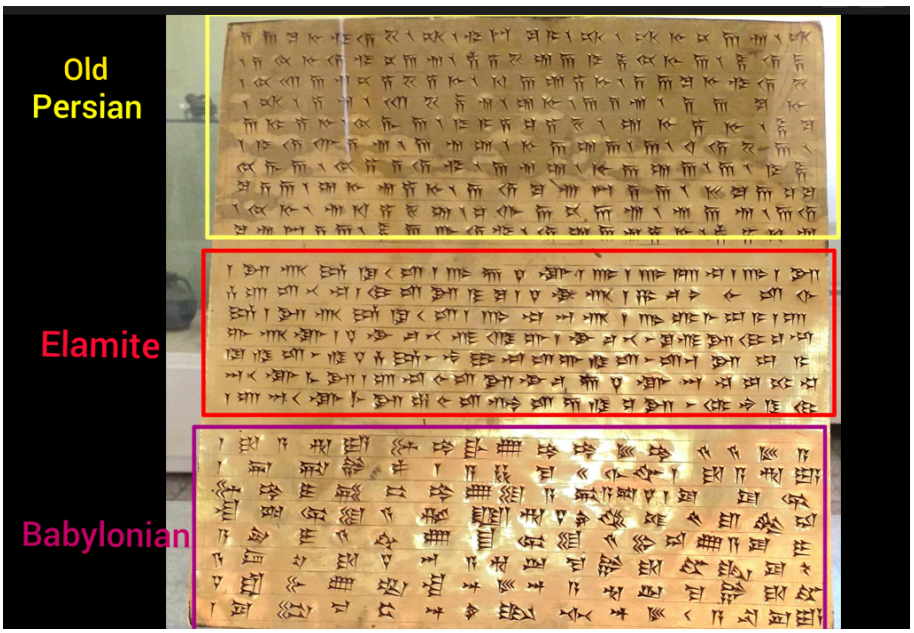

Image 6. Trilingual gold plate of king Darius, photo is taken from national museum of Iran

- Matching fragmented tablets or inscriptions to reconstruct, we can employ OCR technology to transform cuneiform signs from images into machine-readable text. Subsequently, we can leverage Prof. Enrique Jiménez's NLP project to apply algorithms that detect and match segments from various tablets or inscriptions, aiding the reconstruction of fragmented texts (see image 7). These tools collectively enable the identification and matching of text fragments by comparing pixel patterns and character sequences, allowing researchers to piece together ancient manuscripts from scattered fragments like a text puzzle (see image 8).

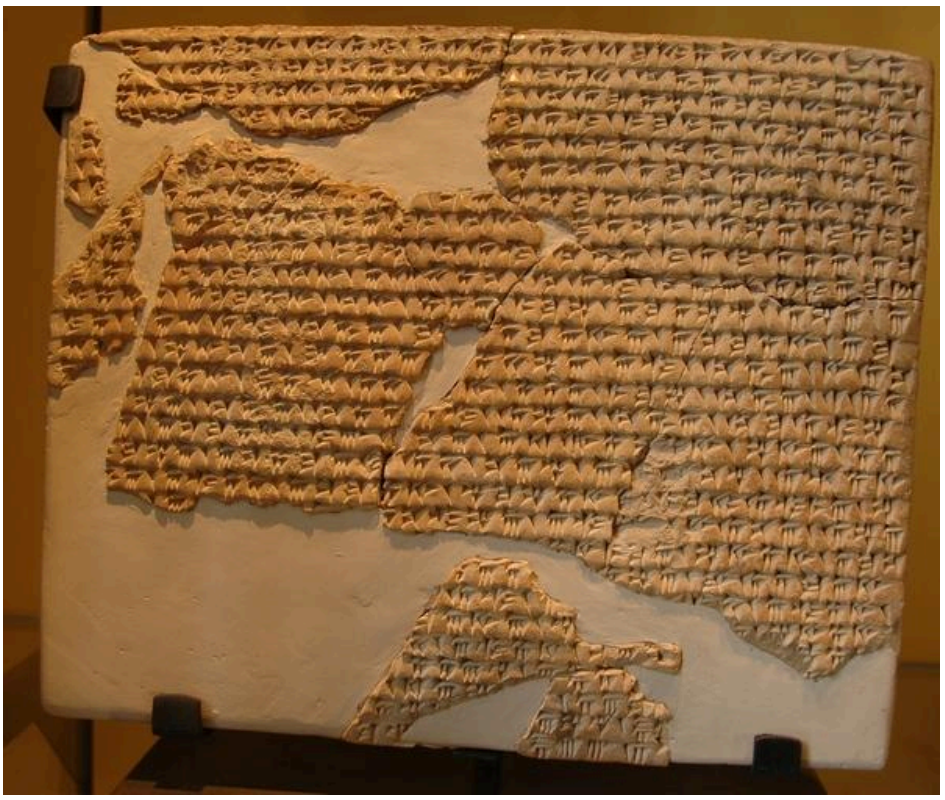

Image 7. Photo is from Apadana Castle Shush, DSf inscription

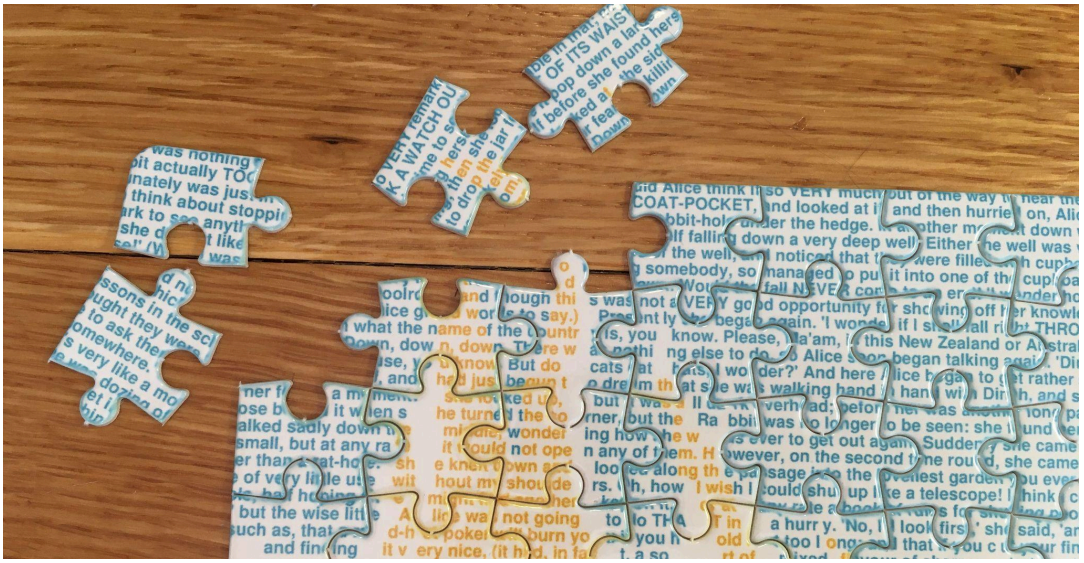

Image 8. Fragmented tablets or inscriptions can be matched like a simple text puzzle

7. Conclusion

The development of AI-driven OCR models for Old Persian cuneiform represents a significant advancement in the preservation and understanding of ancient languages. The "easyocr_old_persian" model, with its integrated image pre-processing capabilities, offers a more versatile and accurate solution compared to previous models like "tessearct_old_persian." Future enhancements, including training with larger datasets and developing custom NLP models, promise further improvements in accuracy and usability.

Moreover, the ongoing development of the "yolo_cnn_old_persian" model and the planned applications of these technologies to other ancient scripts highlight the broader implications of this research. By leveraging AI, we can reconstruct fragmented texts and gain deeper insights into ancient civilizations. This project not only preserves the linguistic heritage of Old Persian but also paves the way for similar advancements in the study of other ancient languages, significantly contributing to the fields of archaeology and historical linguistics.

8. Acknowledgements

I am grateful to Prof. Dr. Enrique Jiménez for encouraging me to continue this project. I would like to thank Dr. Elham Salehi for her kind assistance in publishing this paper.

9. References

Cobanoglu, Yunus, Sáenz, Luis, Khait, Ilya and Jiménez, Enrique. "Sign detection for cuneiform tablets" it - Information Technology, 2024. https://doi.org/10.1515/itit-2024-0028

Gutherz, G., Gordin, S., Sáenz, L., Levy, O., and Berant, J. (2023). Translating Akkadian to English With neural machine translation. PNAS Nexus 2.
https://doi.org/10.1093/PNASNEXUS/PGAD096

Kent, R. G. (1953). *Old Persian: Grammar. Texts. Lexicon* (Vol. 33). American Oriental Society.

Mostofi, F., & Khashman, A. (2014, October). Intelligent recognition of ancient Persian cuneiform characters. In *International Conference on Neural Computation Theory and Applications* (Vol. 2, pp. 119-123). SCITEPRESS.
https://doi.org/10.5220/0005035401190123

Mousavi, S. M. H., & Lyashenko, V. (2017, November). Extracting old persian cuneiform font out of noisy images (handwritten or inscription). In *2017 10th Iranian Conference on Machine Vision and Image Processing (MVIP)* (pp. 241-246). IEEE. https://doi.org/10.1109/IranianMVIP.2017.8342358

Rosebrock, A., Thanki, A., Paul, S., & Haase, J. (2020). OCR with OpenCV, Tesseract and Python.

Russell, S. J., & Norvig, P. (2016). *Artificial intelligence: a modern approach*. Pearson.

Schantz, H. F. (1982). *History of OCR, optical character recognition*. Recognition Technologies Users Association.

Sharp, R. N. (1970). *The inscriptions in Old Persian cuneiform of the Achæmenian emperors*. Central Council of the Celebration of the 25th Century of the Foundation of the Iranian Empire.