Available online at www.qu.edu.iq/journalcm

JOURNAL OF AL-QADISIYAH FOR COMPUTER SCIENCE AND MATHEMATICS

ISSN:2521-3504(online) ISSN:2074-0204(print)

Technology in Sumerian Text Translation: A Review of Tools and Techniques

Sajad M. Shah [a], ***Karim Q. Hussein*** [b]

[a,b]*Mustansiriyah University, College of Science, Computer Science Department, Baghdad, Iraq.*

[a]*Email:(sajadmajeed36@gmail.com)*

[b]*Email:(karimzzm@yahoo.com)*

ARTICLEINFO

Article history:
Received: 11 /11/2024
Rrevised form: 7 /12/2024
Accepted : 11 /2/2025
Available online: 30 /3/2025

Keywords:

Sumerian language, Cuneiform, , NLP, Language Translation, AI, OCR, ML.

ABSTRACT

The Sumerian language, among the earliest recorded languages, originated in Mesopotamia centuries ago and was inscribed on clay tablets using "cuneiform." These artifacts are currently housed in several museums in Iraq and internationally. The translation of these inscriptions poses a challenge for scholars due to the scarcity of texts and their intricate structure, necessitating a labor-intensive manual translation process that is susceptible to errors. Nonetheless, the progression of contemporary technology, particularly in machine learning (ML), natural language processing (NLP), and optical character recognition (OCR), is ushering Sumerian text translation into a new epoch. This research examines the utilization of digital technology to analyze and interpret Sumerian culture, emphasizing successful implementations, ongoing initiatives, and prospective advancements. Integrating artificial intelligence (AI) with specialist expertise might enhance the accuracy and accessibility of Sumerian translation, therefore preserving and elucidating ancient history.

MSC..

https://doi.org/10.29304/jqcsm.2025.17.11968

1. Introduction

The Sumerian civilization is recognized as the first known civilization in southern Mesopotamia's historical region, present-day Iraq (see Figure 1) [1]. The Sumerian texts are considered some of the earliest and most complex forms of human writing, having emerged around 5,000 years ago. The primary writing medium used by the Sumerians was clay. They inscribed their writing on clay tablets, which were often in the shape of pillows, using sharpened reeds known as "pens." The inscriptions consisted of raised patterns and lines on the surface of the clay tablets. These inscriptions' concise and mostly linear shapes contributed to developing the term "cuneiform" [2].

∗Corresponding author

Email addresses:

Communicated by ‘sub etitor’

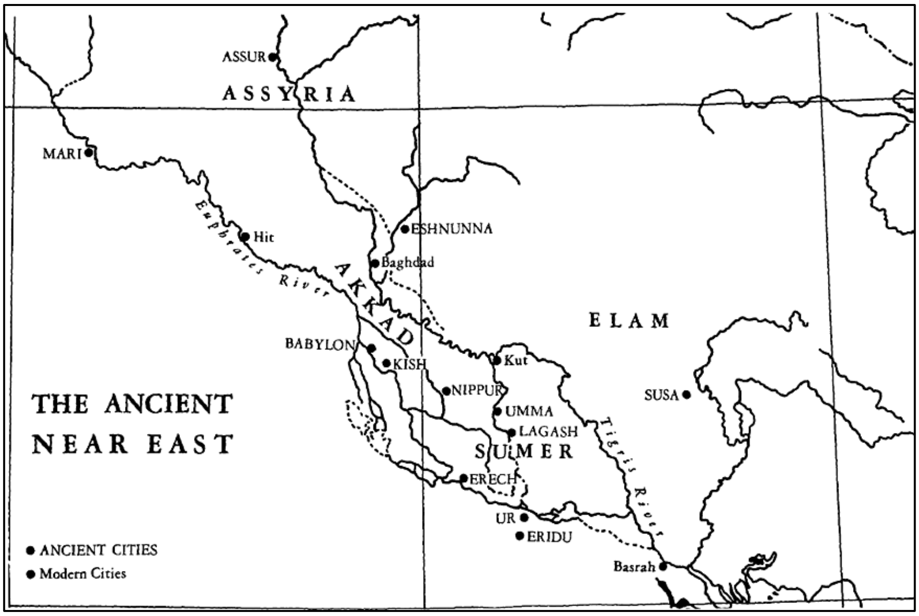


Figure 1: Representation of Sumer's location [2]

More than (100,000) tablets still exist today in various states of disintegration. They range in size from two to dozens of centimeters, see Figure (2). Typically, they hold administrative data, document historical events and commercial transactions, and narrate everyday life [3]. Most Mesopotamian tablets possess two axes: a horizontal axis for numerical information and a vertical axis for data related to various individuals or locations. The persona shown in many tablet writings helps comprehend Sumerian vocabulary and writing style [4].

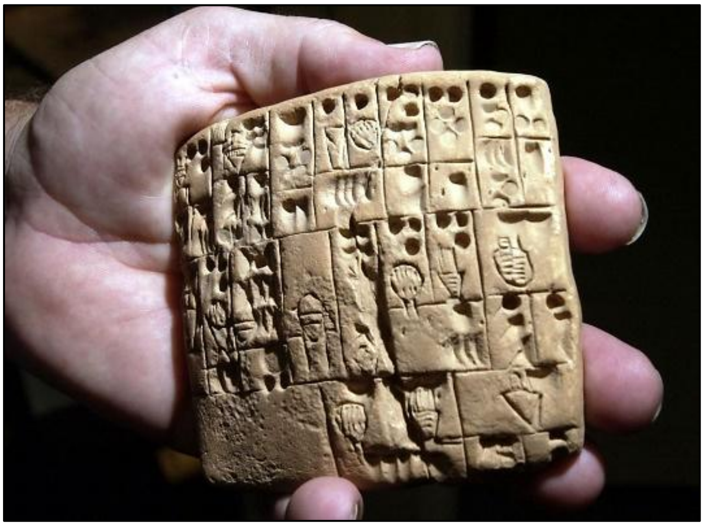


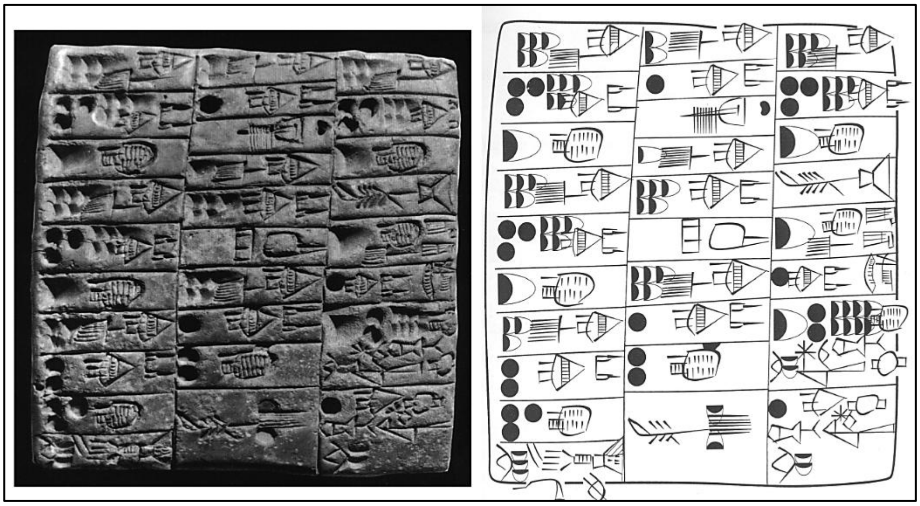
Figure 2: Sumerian clay tablet [4]

However, translating Sumerian symbols is complex due to its grammar structure, lack of complete bilingual texts, and the language's extinct nature. Traditional translation methods often relied heavily on human expertise, making the process time-consuming and error-prone [5]. In recent years, technology has become increasingly important to overcome these challenges, enabling faster and more accurate translations of Sumerian texts. This research reviews the most recent technologies used to decode and translate Sumerian tablets. With the advancement of Machine Learning and Deep Learning technologies, understanding ancient languages has become an active research area.

This paper has the following sections: in section two, the traditional approaches used in the Sumerian translation are explained; in section three, the technological advancements in Translation are discussed. In section four, challenges and constraints encountered in the translation of Sumerian texts are discussed, and finally, the conclusions are presented in section five.

2. Traditional Approaches to Sumerian Translation

The translation of Sumerian texts depended largely on epigraphers and linguists with specialized knowledge of cuneiform script. The discovery of multilingual inscriptions, such as the Behistun Inscription, bolstered early attempts to decipher cuneiform. The Behistun Inscription is a trilingual ancient carving created by Darius I around (522 BCE) in Iran, featuring texts in Old Persian, Elamite, and Akkadian (Babylonian). It was crucial in deciphering cuneiform script, much like the "Rosetta Stone" did for Egyptian hieroglyphs. The key to deciphering cuneiform was having the same text in three different languages . This allowed scholars to compare similar texts across languages, making it easier to understand the meanings of the symbols. Early attempts began with deciphering Old Persian because it was partly known through the names of well-known kings; this helped with understanding the alphabet and grammar. After understanding Old Persian, scholars compared Elamite and Akkadian texts, noticing similarities in names and places. This helped link cuneiform symbols to specific sounds., see Figure (3) [6].



(a) **(b)**

Figure 3: (a) Cuneiform tablet in Sumerian (b) A pen-and-ink style transcription of this tablet [6]

The complexity of the Sumerian language—marked by its logographic and syllabic elements—has made manual translation difficult. As a result, even today, many cuneiform tablets remain untranslated or only partially deciphered. In addition to the difficulty of translation, there is difficulty in determining the type of language because Sumerian and Akkadian languages are written in cuneiform, and to determine the type of language, we use the Cuneiform Language Identification (CLI) [7].

CLI (Cuneiform Language Identification) is a computer technique developed to recognize languages and dialects inside cuneiform writings. It was developed to manage the intricacies of ancient languages like Sumerian and

Akkadian, inscribed in cuneiform script. CLI was created during the VarDial assessment campaign to promote advancements in artificial language recognition techniques for ancient texts. The two languages, Sumerian and Akkadian, were categorized into six temporal phases: Sumerian (SUX), Old Babylonian (OLB), Middle Oceanic Babylonian (MPB), Standard Babylonian (STB), Neo-Babylonian (NEB), Late Babylonian (LTB), and Neo-Assyrian (NEA), Table (1) shows the number of lines for each language or dialect in the training set provided during the VarDial evaluation campaign. It uses advanced algorithms such as HeLI (Helsinki Language Identification Method), which achieves up to 93% F1-score in domain text identification [8].

This facilitated the swift differentiation between Sumerian and Akkadian, along with their respective dialects (e.g., Neo-Assyrian and Babylonian). It assisted scholars in analyzing extensive cuneiform datasets without requiring laborious manual effort. This will facilitate digital archiving, translation, and study of historical documents [8].

Table 1: Classes of the CLI dataset [8]

Language or Dialect	Code	Texts	Lines	Signs
Late Babylonian	LTB	671	31,893	ca. 260,000
Middle Babylonian peripheral	MPB	365	11,015	ca. 95,000
Neo-Assyrian	NE	3,570	65,932	ca. 490,000
Neo-Babylonian	NEB	1,212	19,414	ca. 200,000
Old Babylonian	OLB	527	7,605	ca. 65,000
Standard Babylonian	STB	1,661	35,633	ca. 390,000
Sumerian	SUX	5,000	107,345	ca. 400,000
Total		13,006	278,837	ca. 1,900,000

3. Technological Advancements in Translation

This section will explain many strategies applicable to translating the Sumerian language:

3.1 Digitization of cuneiform files using optical character recognition (OCR)

This section will explain many strategies applicable to translating the Sumerian language. Implementing Optical Character Recognition (OCR) is a key advancement in Sumerian translation technology [9]. Although OCR technology has been extensively employed to digitize contemporary texts, innovations have facilitated its usage for ancient writings such as cuneiform. The large volume of records, with at least 300,000 published and many more unpublished, necessitates digitalizing cuneiform texts to facilitate effective academic research. Several digitalization initiatives for cuneiform documents started around the year 2000.

For instance, several investigations employed 3D scanning of authentic clay tablets (e.g., Digital Hammurabi Project) [10]. In contrast, others focused on transcription and translation (e.g., Cuneiform Digital Library Initiative (CDLI), Open Richly Annotated Cuneiform Corpus (ORACC)) [11], [12]. To enable digitization initiatives, (semi-)automatic digitization technologies are necessary. Regrettably, existing digitalization initiatives have been executed manually. Limited research has attempted to identify a character class using handwritten reproductions of genuine tablets, grammatical analysis, and automated machine translation. Nevertheless, dependable (semi-automated digitizing methods have yet to be developed.

In 2018, Kenji Yamauchi et al. [13] focused on digitizing cuneiform documents, requiring creating an Optical Character Recognition (OCR) technique for handwritten replicas of the original materials. The research introduces the Handwritten Cuneiform Corpus (HCCC), an image collection intended to advance optical character recognition

(OCR) algorithms for cuneiform writings [14]. The objective is to automate the identification and translation of handwritten cuneiform writings to improve the efficacy of academic research. The procedure starts with identifying the 50 most often utilized cuneiform characters derived from sources like the Electronic Text Corpus of Sumerian Literature (ETCSL). Images were extracted from handwritten copies in academic collections and the Cuneiform Digital Library Initiative (CDLI). Heuristics and image processing methods (e.g., Gaussian smoothing and line detection) were synthesized to identify distinct characters. The dataset was refined to exclude missing or indistinct photos, accompanied by hand tagging and cleaning to enhance precision. Approximately 200 photos were designated for each character class to guarantee data variety. The study report indicates that (4,358) annotated photos were gathered, encompassing 26 50 designated character groups. Certain classes had over 200 photos, but others comprised around 25. The dataset has the potential to facilitate OCR algorithms to identify handwritten cuneiform. However, difficulties persist with overlapping or intricate letters. Future endeavors intend to augment the dataset, use one-shot learning methodologies for efficient categorization, and improve the precision of hieroglyph recognition throughout several historical epochs [13].

The study encounters several obstacles in digitizing handwritten cuneiform writings, mostly due to the intricacy and resemblance of cuneiform symbols, rendering them challenging to differentiate. The restricted dataset size and the variability in handwriting styles across scribes diminish the accuracy of the OCR model. Moreover, existing techniques have challenges identifying overlapping characters and clarifying ambiguity among visually like characters. Manual labeling is labor-intensive, and restricted access to the original plates constrains data variety [13].

3.2 Translating Sumerian texts using natural language processing (NLP)

Once Sumerian texts are digitized, machine learning (ML) and natural language processing (NLP) models can be employed to aid in translation. ML models, particularly those based on deep learning, are adept at recognizing patterns in data. By training on large datasets of translated Sumerian texts, these models can begin to predict the meaning of unfamiliar characters or incomplete texts [15].

In 2017, É. Pagé-Perron et al. [16] researched machine translation and automated analysis of the Sumerian language to quickly process extensive collections of Sumerian administrative writings. The project constructs a natural language processing pipeline for morphological analysis, part-of-speech tagging, syntactic parsing, and machine translation. The system employs distant supervision approaches and Neural Machine Translation (NMT) techniques to enhance translation accuracy, utilizing parallel Sumerian-English texts, see Figure (4).

The team effectively developed a system that can autonomously translate Sumerian texts, yielding encouraging preliminary results. It improved the quality of linguistic annotation and enabled information extraction, especially for prosopographic research (monitoring persons throughout texts). The system could manage intricate Sumerian morphology and syntax, enhancing academic data accessibility [16].

Primary obstacles are data sparsity resulting from the restricted quantity of annotated texts and the intricate morphology of Sumerian, an agglutinative language characterized by extensive affixation. Regional differences, code-switching with Akkadian, and varying orthographic conventions further complicate research. The absence of contemporary descendants of Sumerians complicates annotation projection due to the lack of closely similar languages for reference. Surmounting these constraints necessitates sophisticated lemmatization methods and the augmentation of the dataset via hand annotation [16].

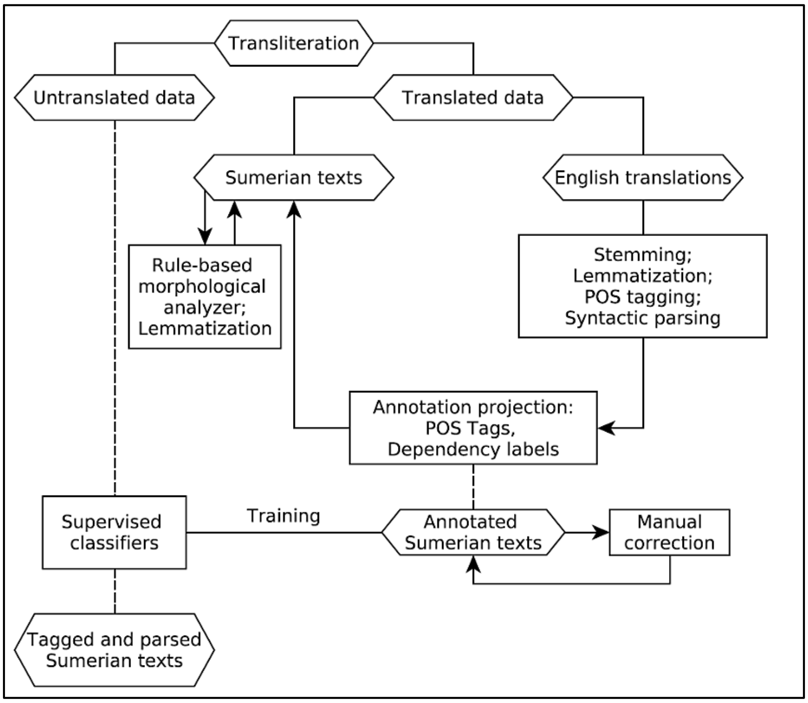

Transliteration
Untranslated data
Translated data
Sumerian texts
English translations
Rule-based morphological analyzer; Lemmatization
Stemming; Lemmatization; POS tagging; Syntactic parsing
Annotation projection: POS Tags, Dependency labels
Supervised classifiers
Training
Annotated Sumerian texts
Manual correction
Tagged and parsed Sumerian texts

Figure 4: NLP pipeline for Sumerian [16]

3.3 Cuneiform reading process based on deep learning

The use of deep learning with convolutional neural networks (CNNs) in translating Sumerian texts is a recent development in the field of natural language processing (NLP) and translation of ancient texts. CNNs were originally designed to process images and structure data such as texts written on clay tablets, which makes them effective in recognizing complex patterns within cuneiform texts. With their ability to analyze and interpret images, CNNs have greatly improved the process of optical character recognition (OCR) of Sumerian texts. This technology converts digital images of cuneiform inscriptions into processable text. This conversion helps facilitate subsequent processing steps, such as linguistic analysis and translation [17].

In his research published in 2020, T. Dencker et al. [18] aim to facilitate the process of reading cuneiform texts by developing a detector based on deep learning, which can identify and classify cuneiform signs in images. The system relies on using existing translation (which represents the content of the tablets in Latin) to generate training data, reducing the need to collect expensive annotation data.

The methodology used in his research was:

1. Weak Learning: Translations are used to locate signs in images, allowing training data to be generated without the need to collect expensive annotation data.

2. Iterative Training: The model is gradually improved by repeating the training process, where the generated signs are used to train the sign detector.

3. Web Application: A web application that allows experts to use the detector in their studies has been developed. (see Figure 5).

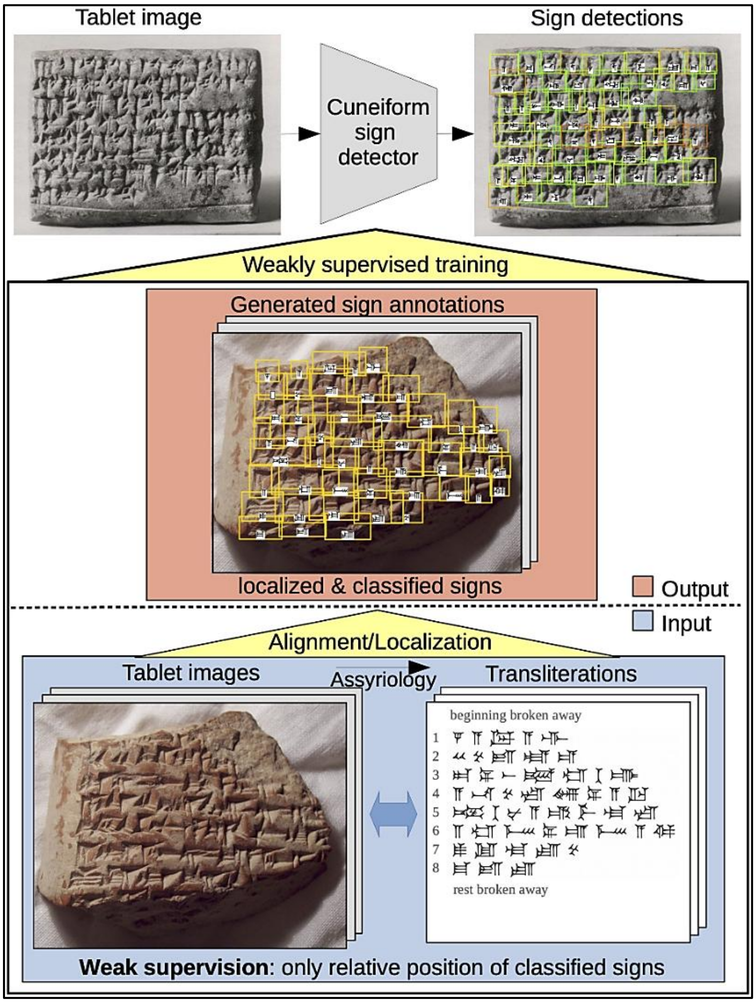


Figure 5: An overview of its approach [18]

The model attained satisfactory accuracy in identifying cuneiform signs, suggesting that alignment with transliteration might serve as an efficient method for training models without necessitating manual annotation of each symbol. The experiment revealed that employing weak supervision effectively addresses the issue of insufficient annotated data. The work encounters problems like the diversity of cuneiform forms, difficulty differentiating identical symbols, and the paucity of high-quality data that hamper the model's ability to learn effectively. The physical deterioration of clay tablets, such as erosion and damage, may impair recognition quality. The constraints of inadequate supervision, due to imprecise alignment between text and picture, may impact overall performance. The work introduces a novel method for recognizing cuneiform signs using artificial intelligence, emphasizing the significance of data and transliteration as indirect supervisory mechanisms [18].

3.4 Improving Sumerian tablet images for translation

Image enhancement techniques, such as wavelet transforms, have also been used to improve the clarity of tablet inscriptions. These methods, which enhance image contrast and detail, allow for more accurate OCR and machine translations. The improved image quality reduces the errors that arise from worn or damaged cuneiform symbols, thus improving overall translation accuracy [19].

In 2017, M. Talib [20] presented a method for extracting cuneiform characters from Sumerian texts using the Discrete Wavelet Transform (DWT) and the Split Region method. Initially, preprocessing techniques are applied to enhance and segment the tablet images. Then, DWT is used to extract relevant features, followed by applying the Split Region algorithm to isolate individual characters. The procedure can be delineated in the phases below:

Inputs: The image of the clay board is inserted.

Enhancing image: Improving image quality to eliminate noise or defects caused by the passage of time.

Segmentation: The image is divided into specific areas.

Wavelet Transformation: DWT technology is applied to the image.

Characters Extraction: After improving the image and dividing the areas, the Sumerian characters are extracted from the text.

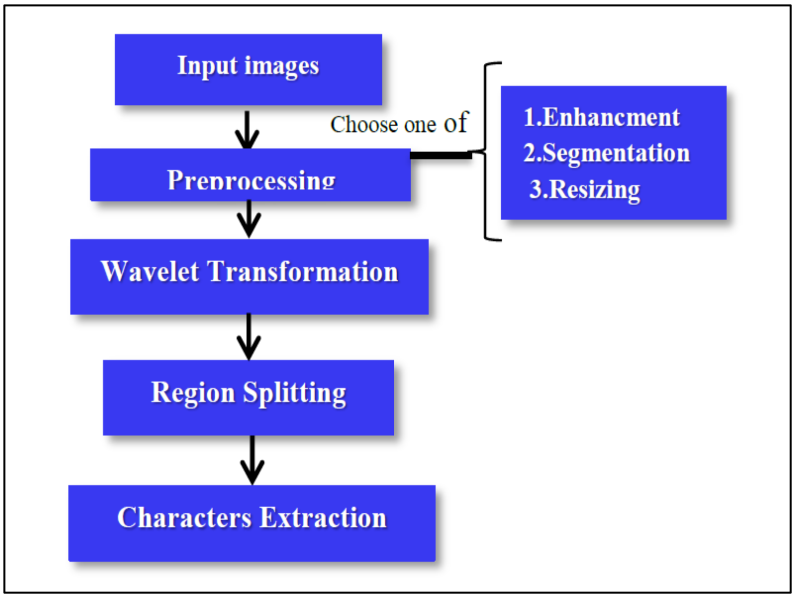


Figure 6: The Stages Involved in the Proposed Method [20]

CNN was used to classify the correct Sumerian letters extracted from cuneiform tablet images. After pre-processing and segmentation, the extracted letter regions are compared with a dataset of verified letters. CNN analyzes each extracted region, matches it with the dataset, and identifies only the correct Sumerian letters (see Figure 7) [21].

The findings show that this technique can effectively extract characters from tablet images, which benefits scholars studying cuneiform texts. The research encountered several challenges, including the intricacy of Sumerian script, the degradation of the tablets with time, and the requirement for high-resolution images for precise analysis. Additionally, constraints were identified, including dependence on particular datasets, the computing expense of wavelet transformations, and possible difficulties in generalizing the methodology to all varieties of cuneiform tablets. Notwithstanding these constraints, the research offers a significant instrument for digital preservation and automated analysis of old texts [20].

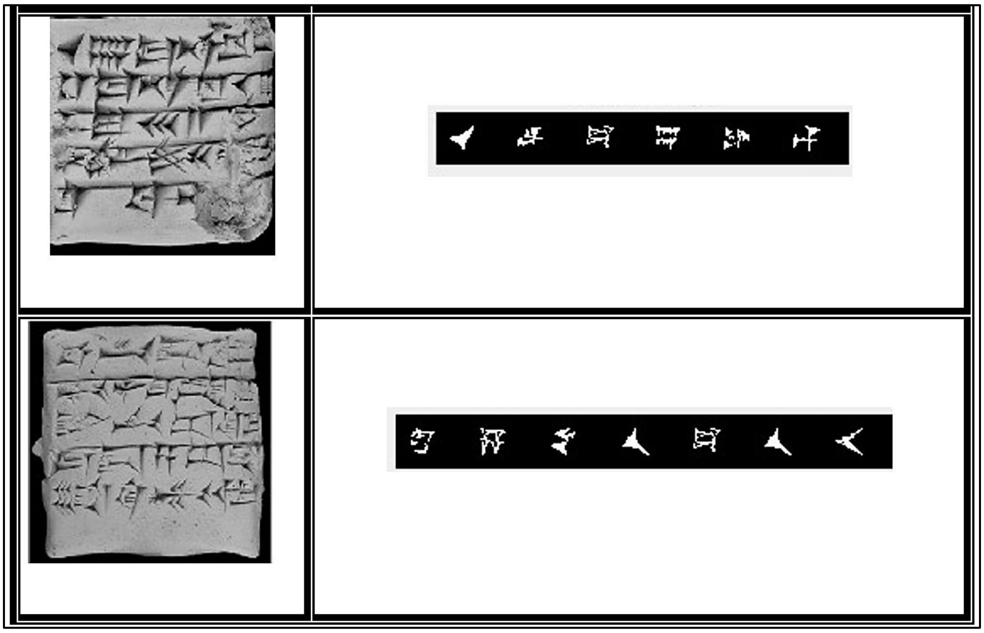


Figure 7: shows the correct resulting characters from the tablet image [20]

4. Challenges and Constraints encountered in the Translation of Sumerian Texts

A significant issue in Sumerian translation is the scarcity of bilingual or annotated literature, which hinders the successful training of AI models. In contrast to commonly spoken languages, the scarcity of translated Sumerian writing limits the availability of machine learning datasets. A significant concern is the precision of translations, as cuneiform writings are sometimes partial, damaged, or confusing, resulting in possible interpretative mistakes. Moreover, Sumerians coexisted alongside Akkadian and other ancient languages, complicating linguistic variances and contextual dependencies for AI models to analyze. Optical character recognition (OCR) for cuneiform signs continues to pose difficulty due to several uneven inscriptions and eroded characteristics [22].

It is imperative to augment databases via crowdsourcing and multidisciplinary cooperation to address these restrictions. Creating a consistent and annotated multilingual database will enhance training resources for AI models. Integrating machine learning with human experience enhances translation accuracy by merging computer capabilities with language analysis. Advanced AI methodologies, including context-aware deep learning models and cross-linguistic learning, can improve translation precision, especially by utilizing the commonalities between Sumerian and Akkadian. Enhancing OCR technology and preprocessing techniques will provide more accurate text extraction from damaged tablets [23].

5. Conclusion

The utilization of contemporary technology in the translation of Sumerian writings presents significant potential for surmounting the conventional obstacles associated with manual translation. Optical character recognition (OCR), machine learning (ML), and natural language processing (NLP) have proven their capability to enhance transcription and translation efficiency while providing improved accuracy and accessibility. Integrating AI with human experience can improve translation outcomes, guaranteeing more contextual accuracy. Human experience is vital for improving the accuracy of machine learning models in Sumerian text translation by supplying necessary linguistic, historical, and contextual insights.

Although AI models can analyze extensive datasets and identify patterns in cuneiform script, they frequently encounter difficulties with unclear, partial, or contextually dependent interpretations. Authorities in Sumerian linguistics and ancient history may enhance AI training by assembling superior datasets, rectifying inaccuracies in automated translations, and optimizing machine learning algorithms using linguistic principles and cultural subtleties.

Moreover, human experts may authenticate AI-generated translations, guaranteeing that the results conform to historical interpretations and academic agreement. Researchers may create hybrid models that reconcile automation with linguistic precision by integrating AI's computational capabilities with expert supervision. Ultimately, combining human experience with machine learning yields more accurate and dependable translations, facilitating the connection between ancient literature and contemporary comprehension.

References

[1] S. N. Kramer“*History Begins at Sumer.pdf*.” Accessed: Feb. 03, 2025. [Online]. Available: https://s3.us-west-1.wasabisys.com/luminist/EB/I-J-K/Kramer%20-%20History%20Begins%20at%20Sumer.pdf.11. Chicago: The Univ. of Chicago Press, 1956.

[2] H. J. Nissen, P. Damerow, and R. K. Englund, *Archaic bookkeeping: early writing and techniques of economic administration in the ancient Near East*. Chicago, Ill: University of Chicago Press, 1993.

[3] H. T. Lopes and I. Almeida, “The Mediterranean: The Asian and African Roots of the Cradle of Civilization,” in *Mediterranean Identities - Environment, Society, Culture*, B. Fuerst-Bjelis, Ed., InTech, 2017. doi: 10.5772/intechopen.69363.

[4] S. N. Kramer, *The Sumerians: their history, culture, and character*, 11. [print]. Chicago: The Univ. of Chicago Press, 1991.

[5] J. N. Postgate, *Languages of Iraq, ancient and modern*. London? British School of Archaeology in Iraq, 2007.

[6] S. R. Narang, M. K. Jindal, and M. Kumar, “Ancient text recognition: a review,” *Artif. Intell. Rev.*, vol. 53, no. 8, pp. 5517–5558, Dec. 2020, doi: 10.1007/s10462-020-09827-4.

[7] E. A. Saeed, A. D. Jasim, and M. A. Abdul Malik, “Cuneiform Text Dialect Identification Using Machine Learning Algorithms and Natural Language Processing (NLP),” *Iraqi J. Inf. Commun. Technol.*, vol. 7, no. 2, pp. 26–40, Dec. 2024, doi: 10.31987/ijict.7.2.265.

[8] M. Zampieri *et al.*, “A Report on the Third VarDial Evaluation Campaign,” vol. 7, no. 2, pp. 26–40, Sep. 2019.

[9] S. Gordin, M. Alper, A. Romach, L. Saenz Santos, N. Yochai, and R. Lalazar, “CuReD: Deep Learning Optical Character Recognition for Cuneiform Text Editions and Legacy Materials,” in *Proceedings of the 1st Workshop on Machine Learning for Ancient Languages (ML4AL 2024)*, Hybrid in Bangkok, Thailand and online: Association for Computational Linguistics, 2024, pp. 130–140. doi: 10.18653/v1/2024.ml4al-1.14.

[10] C. Weiler, “The Epic of Gilgamesh”.

[11] “Home,” CDLI. Accessed: Dec. 15, 2024. [Online]. Available: https://cdli.mpiwg-berlin.mpg.de/

[12] “Oracc: The Open Richly Annotated Cuneiform Corpus.” Accessed: Dec. 15, 2024. [Online]. Available: https://oracc.museum.upenn.edu/

[13] K. Yamauchi, H. Yamamoto, and W. Mori, “Building A Handwritten Cuneiform Character Imageset”.

[14] A. M. Mutawa, M. Y. Allaho, and M. Al-Hajeri, “Machine Learning Approach for Arabic Handwritten Recognition,” *Appl. Sci.*, vol. 14, no. 19, p. 9020, Oct. 2024, doi: 10.3390/app14199020.

[15] E. A. Saeed, A. D. Jasim, and M. A. Abdul Malik, “Cuneiform Text Dialect Identification Using Machine Learning Algorithms and Natural Language Processing (NLP),” *Iraqi J. Inf. Commun. Technol.*, vol. 7, no. 2, pp. 26–40, Sep. 2024, doi: 10.31987/ijict.7.2.265.

[16] É. Pagé-Perron, M. Sukhareva, I. Khait, and C. Chiarcos, “Machine Translation and Automated Analysis of the Sumerian Language,” in *Proceedings of the Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, Vancouver, Canada: Association for Computational Linguistics, 2017, pp. 10–16. doi: 10.18653/v1/W17-2202.

[17] S. Sakib, N. Ahmed, A. J. Kabir, and H. Ahmed, “An Overview of Convolutional Neural Network: Its Architecture and Applications,” Feb. 14, 2019, *MATHEMATICS & COMPUTER SCIENCE*. doi: 10.20944/preprints201811.0546.v4.

[18] T. Dencker, P. Klinkisch, S. M. Maul, and B. Ommer, “Deep learning of cuneiform sign detection with weak supervision using transliteration alignment,” *PLOS ONE*, vol. 15, no. 12, p. e0243039, Dec. 2020, doi: 10.1371/journal.pone.0243039.

[19] P. Nagy Israel Michael, “The Importance of Using Artificial Intelligence in Image Processing Field and Printing,” vol. 0, no. 0, pp. 0–0, Sep. 2024, doi: 10.21608/jsos.2024.313524.1581.

[20] M. Talib and J. H. S, “Sumerian Character Extraction by Using Discrete Wavelet Transform and Split Region Methods,” *Kurd. J. Appl. Res.*, vol. 2, no. 3, pp. 62–65, Aug. 2017, doi: 10.24017/science.2017.3.20.

[21] P. Purwono, A. Ma’arif, W. Rahmaniar, H. I. K. Fathurrahman, A. Z. K. Frisky, and Q. M. U. Haq, “Understanding of Convolutional Neural Network (CNN): A Review,” *Int. J. Robot. Control Syst.*, vol. 2, no. 4, pp. 739–748, Jan. 2023, doi: 10.31763/ijrcs.v2i4.888.

[22] A. H. S. Hamdany, R. R. Omar-Nima, and L. H. Albak, “Translating cuneiform symbols using artificial neural network,” *TELKOMNIKA Telecommun. Comput. Electron. Control*, vol. 19, no. 2, p. 438, Apr. 2021, doi: 10.12928/telkomnika.v19i2.16134.

[23] C. Simmons, “Pioneering Deep Learning Approaches to Sumerian Cuneiform,” p. 56, Jun. 2024.