



Translating Akkadian Transliterations to English with Transfer Learning

Najat Nehme¹, Danielle Azar¹^a, Diana Kutsalo² and Jalal Possik³^b

¹*Department of Computer Science and Mathematics, Lebanese American University, Byblos, Lebanon*

²*École du Numérique (EDN), Faculté de Gestion, Économie et Sciences (FGES),
Université Catholique de Lille, F-59000 Lille, France*

³*ICL, Junia, Université Catholique de Lille, LITL, F-59000 Lille, France*

Keywords: Akkadian Translation, Machine Translation, Natural Language Processing, Transfer Learning, Historical Linguistics, Low-Resource Languages, Neural Machine Translation, Computational Linguistics.

Abstract: Akkadian is an ancient Semitic language with a complex cuneiform script and fragmented artifacts. These attributes make the translation of text from this language to another one very challenging. In this paper, we utilize transfer learning with a model pre-trained on multiple languages to enhance the accuracy to translate from Akkadian to English. By fine-tuning this model on a curated Akkadian-English dataset, the research aims to leverage the extensive linguistic pre-training of the model in order to adapt it to Akkadian's specificities.

1 INTRODUCTION

Translating Akkadian, an ancient Semitic language, to another language, presents unique challenges not only due to the language's complex grammar and vocabulary, but also because of the transliteration process from its original cuneiform script. Although significant progress has been made in the transliteration of Akkadian cuneiform into readable text, with models achieving high accuracy, the subsequent translation of these transliterations into contemporary languages like English remains a substantial challenge. This work aims to bridge this gap by utilizing a model that has been pre-trained in multiple languages and enhancing its ability to translate Akkadian transliterations effectively. By fine-tuning this model on a curated Akkadian-English dataset, the proposed approach leverages extensive linguistic training to adapt to the specificities of Akkadian. This approach not only helps produce more precise translations, but also aids in the interpretation and understanding of Akkadian texts, offering valuable resources for scholars studying ancient Mesopotamian cultures. This research is also aimed at offering a methodological model that can be


applied to other ancient languages that are dealing with comparable difficulties.

2 PROBLEM STATEMENT

The translation of Akkadian transliterations into English encompasses unique challenges that conventional Natural Language Processing (NLP) methods struggle to address effectively. Given the complexity of Akkadian and the nuanced nature of its transliterated texts, these challenges are twofold: First, the scarcity and fragmentation of textual data hinder the training of robust NLP models; second, the linguistic diversity and structural intricacies of Akkadian pose significant barriers to accurate and meaningful translation.

Traditional NLP models, which are often trained on large, well-annotated datasets in widely spoken modern languages, fail to accommodate the specific requirements of a low-resource, ancient language like Akkadian. Additionally, the source texts, being derived from ancient cuneiform inscriptions, often come in incomplete or reordered forms, further complicating the transliteration and subsequent

^a <https://orcid.org/0000-0002-6159-3714>

^b <https://orcid.org/0000-0002-5246-8102>

translation processes. These factors lead to potential inaccuracies and a loss of critical historical context. To overcome these significant hurdles, this work employs a transfer learning approach, leveraging a model pre-trained on a diverse corpus of languages. This method allows the model to utilize its pre-existing linguistic knowledge, which can be fine-tuned to accommodate the peculiarities of Akkadian transliterations. By adapting advanced machine learning techniques, particularly transfer learning, this research aims to enhance the model's ability to understand and translate Akkadian accurately, thereby contributing to the preservation and understanding of ancient Mesopotamian literature and civilization.

3 LITERATURE REVIEW

Recent advancements in NLP have significantly propelled the task of translating Akkadian, an ancient Semitic language. However, work on this ancient language remains scarce. Gutheisz et al. (2023) undertook a comprehensive exploration of neural machine translation (NMT) for Akkadian, employing Transformer-based NMT models to handle both transliteration-to-English (T2E) and cuneiform-to-English (C2E) translations. The proposed models achieved high BLEU4 scores of 36.52 for C2E and 37.47 for T2E, demonstrating substantial advancements in the field. Despite these achievements, the study acknowledged the need for further accuracy improvements, a gap that this project aims to address. Gordin et al. (2020) developed an innovative method for the automatic transliteration of Unicode cuneiform glyphs, achieving a 97% accuracy. This breakthrough significantly advanced the digitization of Akkadian texts, creating a foundation that this research builds upon to improve translation from transliterations to English. Lazar et al. (2021) introduced a BERT-based model aimed at completing missing text in Akkadian transliterations, achieving an 89% accuracy. This study highlights the potential of large-scale multilingual pretraining in enhancing translation models for low-resource languages like Akkadian, which is closely aligned with the methodologies employed in this work.

These studies collectively underscore the ongoing challenges in AI-driven Akkadian translation, especially with longer sequences and diverse text genres. The findings provide valuable insights into effective methodologies and highlight the need for more advanced decoding schemes and the exploration of the influence of related languages to enhance translation performance.

4 BACKGROUND

Akkadian was the administrative and cultural dominant language of ancient Mesopotamia, used extensively between 2,500 BCE and 500 BCE. As a pivotal medium for documenting a broad spectrum of societal activities—from governmental decrees to religious texts—Akkadian inscriptions on clay tablets are invaluable for scholars studying the socio-economic and political landscapes of early civilizations in the region.

4.1 Linguistic Complexity

Akkadian presents unique linguistic challenges. It is a cuneiform language i.e. it was written using a system of wedge-shaped symbols pressed into clay tablets, as depicted in Figure 1.



Figure 1: Akkadian Cuneiform Tablet: This image shows an ancient Akkadian clay tablet inscribed with cuneiform script, exemplifying the typical medium for administrative decrees and cultural records in Mesopotamia between 2,500 BCE and 500 BCE (Yale University Library, n.d.).

The script's complexity is compounded by its evolution over centuries, with hundreds of signs that vary significantly across different epochs and regions. This variability poses substantial challenges for translation and interpretation, particularly when it comes to ensuring accuracy and maintaining the integrity of the texts. Figure 1 illustrates a well-preserved example of an Akkadian clay tablet, showcasing the typical cuneiform script used for administrative decrees and cultural records in Mesopotamia between 2,500 BCE and 500 BCE.

4.2 Preservation Challenges

The physical state of Akkadian artifacts further complicates scholarly efforts. Many clay tablets have survived only in fragmented forms, with inscriptions that are often eroded or incomplete, such as the one shown in Figure 2. This deterioration not only impedes readability but also leads to gaps in the texts that require careful scholarly reconstruction, which can be speculative and prone to errors. Figure 2 displays a deteriorated Akkadian clay tablet, highlighting the preservation challenges faced by scholars. The erosion and incompleteness of such artifacts complicate the efforts needed to translate and understand fully the ancient language.



Figure 2: Fragmented Akkadian Inscription (n.d.): Displayed is a deteriorated Akkadian clay tablet, highlighting the preservation challenges faced by scholars. The erosion and incompleteness of such artifacts complicate the efforts needed to translate and understand the ancient language fully.

4.3 Computational Challenges

From a computational linguistics perspective, translating Akkadian is daunting due to the scarcity of comprehensive and coherent datasets. Unlike languages that benefit from vast amounts of digital data, Akkadian lacks extensive, well-annotated corpora necessary for training conventional NLP models. This is a significant barrier for applying advanced machine learning techniques, which typically rely on large datasets to learn from and make accurate predictions.

This study aims to address these challenges by leveraging a transfer learning approach, utilizing a preexisting advanced machine translation model trained on a diverse linguistic dataset. By fine-tuning this model on a curated set of Akkadian transliterations to English, the project seeks not only

to enhance the accuracy and fluency of translations but also to deepen the understanding of ancient Mesopotamian civilization. This research endeavor aims to advance the field of computational linguistics for low-resource languages, providing a methodological blueprint that can be adapted to other ancient languages facing similar challenges.

5 PROPOSED METHODOLOGY

This research leverages a transfer learning approach, utilizing the Helsinki-NLP's opus-*mt-ROMANCE*-en model, which was pre-trained on a corpus of multiple Romance languages such as French (fr, fr FR), Spanish (es, es ES), Portuguese (pt, pt PT), Italian (it, it IT), and Romanian (ro). These languages represent a broad linguistic base, providing the model with extensive training on various grammatical structures and vocabularies (Tiedemann & Thottingal, 2020). This strategy is intended to exploit the model's existing linguistic capabilities and adapt them to the specific requirements of translating Akkadian transliterations into English.

5.1 Data Description

The datasets used in this study are primarily sourced from the Open Richly Annotated Cuneiform Corpus (ORACC), providing a rich pool of Akkadian transliterations for the translation project. The data sets utilized include the Royal Inscriptions of the Neo-Assyrian Period (RINAP), Royal Inscriptions of Assyria Online (RIAO), Royal Inscriptions of Babylonia Online (RIBo), State Archives of Assyria Online (SAAo), and The Corpus of Suhu Online. These collections are instrumental in providing textual data that reflects a wide array of administrative, cultural, and historical aspects of ancient Mesopotamia: Royal Inscriptions of the Neo-Assyrian Period (RINAP) consists of royal decrees and annals detailing the reigns of various Assyrian kings (University of Pennsylvania, n.d., "Neo-Assyrian Period"). Royal Inscriptions of Assyria Online (RIAO) features a comprehensive collection of inscriptions from Assyrian kings, used primarily to understand the empire's military and administrative strategies (University of Pennsylvania, 2011 "RIAO Project"). Royal Inscriptions of Babylonia Online (RIBo) offers insights into Babylonian royal practices and is essential for comparative studies with Assyrian texts (University of Pennsylvania, 2011, "RIBo Project"). State Archives of Assyria Online (SAAo) includes administrative records, treaties, and

correspondence between kings and their officials, providing a deep dive into Assyria’s bureaucratic practices (University of Pennsylvania, 2015, "SAAo"). The Corpus of Suhu Online contains texts from the Suhu region, highlighting the local governance and cultural practices at the periphery of the Assyrian Empire (University of Pennsylvania, 2011, "Suhu Project").

<p>1 [aš-šur]-[SAG]-[i-ši] [ša-ak]-[ni] [AB] ŠID aš-šur šá</p> <p>2 [a-nu] [BAD] u [DİS] DINGIR.MEŠ GAL.MEŠ a-na [šu-te-šur] KUR aš-šur EN-su ib-bu-ú (...)</p> <p>3 [MAN KALA] [MAN] KIŠ MAN KUR aš-šur A mu-[tāk]-[kil-⁹musku] ŠID aš-šur A aš-šur-dan ŠID aš-šur-ma</p> <p>4 [e-nu-ma] [É] ša-¹hu-ri ša É ku-¹tál-[li] ... ša ...]</p> <p>5 [... pa]-[ni]-ia DÜ-šu i-na ri-be ša tar-[ši] ...]</p> <p>6 [...] x lu ú-ša-ak-lil ia-e-re [TA] [...]</p> <p>7 [...] šu-a-tu-nu ú-šal-ba-ru-ma e-na-hu NUN EGIR-¹ú [...]</p> <p>8 [...] [šu]-mi it-ti MU-šu lil-tu-ur i-na aš-ri šu-a-tu x [...]</p> <p>9 [ina tu-ub lib-bi ù ka-šad er-nin]-[te] [ta- bi-iš] DU.DU.MEŠ-šu mu-né-¹kir' [...]</p> <p>10 [...] DINGIR.MEŠ GAL]. [MEŠ] ez-zi-¹is' [li-kil-mu-šu-ma er-re-ta ma-ru-ul]-te li- ru-ru-šu MU-šu NUMUN-šu i-na [KUR]¹ [lu-¹hal-li-qu]</p>	<p>(1) [Aššur]-reša-iši (I), appointee [of the god Enlil, vice-regent of (the god) Aššur, (the one) whose dominion the gods Anu, Enlil, and Ea] — the great gods — [designated] for [the proper administration of Assyria (and whose priesthood they blessed), strong king], king of the world, king of Assyria; son of Mut[akkil-Nusku, vice-regent of (the god) Aššur, son of Aššur-dān (I), (who was) also vice-regent of (the god) Aššur].</p> <p>(4) [At that time], (as for) the šaḫūru-house of the hinter house [... which ..., a king who came before] me had built, in an earthquake, during the time [of Aššur-dān (I), ...] I completed (it). [...] rosettes [...].</p> <p>(7) [When the room of] those [...] becomes old and dilapidated, [may] a future ruler [renovate its dilapidated section(s) ...]. May he write my name with his name (and) [return (it)] to that place. [May the god(s) ...] guide him properly [in joy and success].</p> <p>(9) (As for) the one who removes [my inscription and my name, may ..., the great gods, glare at him] angrily [and] inflict upon him [an evil curse. May they make] his name (and) his seed [disappear] from the land.</p>
--	---

Figure 3: A snapshot from Aššur-rēša-iši I 04 [via RIAO/RIA2], a Middle Assyrian text. The achievements and religious dedication of Aššur-rēša-iši I, an Assyrian monarch, are documented in this passage from the Middle Assyrian work Aššur-rēša-iši I 04. Along with English translations, it offers Akkadian transliterations, which are phonetic representations of cuneiform symbols into Latin character. These translations help the model map Akkadian nouns and phrases to their English counterparts. References to gods, titles, deities, and official positions—such as "vice-regent of the god Ashur"—are among the text's many intricate grammatical constructions. To ensure successful translation, the model needs to be trained to manage these syntactic and cultural nuances (Open Richly Annotated Cuneiform Corpus (ORACC). (n.d.)).

5.2 Data Selection and Limitations

Due to the extensive computational resources required for training large neural translation models, this project was constrained to utilize a subset of the available Akkadian transliterations. The total dataset, originally encompassing a vast array of inscriptions, was limited to 20,000 lines selected strategically to

represent a broad spectrum of the corpus. This selection was made to ensure diversity in the training data while balancing the computational limitations encountered.

We created the subset by including samples from various historical periods and genres from the Akkadian corpus, thus providing a comprehensive cross-section reflecting the linguistic and contextual variety of the source material. This approach is critical in maintaining the integrity of the translation model’s training process under the constraints of available computational power, specifically limitations in memory and processing capacity. The selection process involved prioritizing texts that offer rich linguistic features and historical significance, ensuring that the model learns from the most informative and representative examples of the language.

5.3 Data Preprocessing

Preprocessing is essential for ensuring the Akkadian transliteration data is ready for machine translation. For this, we resorted to the following three steps in order to address challenges inherent in dealing with ancient scripts and ensure consistency across the dataset:

Normalization: Through normalization, we converted all text to a consistent format by transforming it to lowercase. This step ensures uniform representation across the dataset, facilitating more accurate processing.

Cleaning: We used regular expressions to remove extraneous elements such as non-alphanumeric characters (except hyphens, spaces, and text within curly brackets) and excessive whitespace. This step is crucial for maintaining the integrity of the original texts and preventing any non-original content from affecting the translation process.

Tokenization: We used the SentencePiece tokenizer for segmented texts, which creates a sub-word vocabulary based on substring frequency, ideal for Akkadian transliterations. This method efficiently handles unknown or rare words by splitting text into manageable tokens for neural network processing in machine translation.

By employing these preprocessing steps, the transliteration texts are optimally prepared for any further processing by neural translation models, ensuring that the linguistic nuances of Akkadian are accurately captured and translated.

5.4 Model Selection and Adaptation

The opus-mt-ROMANCE-en model, based on the Transformer architecture, is known for its robust performance in translating between European languages. This architecture supports extensive sequence-to-sequence tasks, making it ideal for handling the complex syntax of Akkadian transliterations. We adapted the model by fine-tuning it to better capture the syntactic and semantic peculiarities of Akkadian. The two key considerations we had when selecting the model were its sophisticated pre-processing capabilities which is crucial for managing non-standard text formats and the extensive linguistic diversity it was trained on, providing a solid base for transfer learning.

6 EXPERIMENTS AND RESULTS

We trained the model on Google Colab Pro using PyTorch and the Transformers library, taking advantage of its T4 GPU and High RAM settings to manage the computational demands effectively while custom data loaders optimized the handling and batching of training data. We conducted a series of experiments to evaluate the effectiveness of the Helsinki-NLP/opus-mt-ROMANCE-en model for translating Akkadian transliterations into English (Tiedemann & Thottingal, 2020). We tested the model with various hyperparameters in order to optimize the performance, focusing on different configurations of batch size, maximum sequence length, learning rate, and the number of epochs.

6.1 Data Split and Model Training

The dataset, primarily sourced from the Open Richly Annotated Cuneiform Corpus (ORACC), was divided into training, validation, and testing sets to ensure robust model evaluation and to avoid overfitting. The set of 20,000 lines of Akkadian transliterations was split into a training set consisting of 18,000 lines (90%) ensuring that the model has enough examples to learn the complex patterns of Akkadian transliterations, a validation set of 1,000 lines (5%) used to tune the hyperparameters and prevent the model from overfitting and a testing set of 1,000 lines (5%) used to evaluate the model's performance after the training phase. This step is crucial for assessing how well the model generalizes to new, unseen data and ensure that the model performs reliably when deployed in real-world scenarios. We trained the model using the AdamW optimizer with a learning

rate of 0.0001, over a total of 10 epochs. Label smoothing was employed to prevent overfitting. This technique helps in softening the confidence of the labels, which is particularly useful in cases where the dataset is not perfectly labeled or when the model needs to be robust to small perturbations in the input data. During training, the model's performance was periodically evaluated on the validation set to monitor progress and adjust training parameters if necessary.

6.2 Model Performance Evaluation

To comprehensively assess the translation quality of the Akkadian transliterations, we used the BLEU score, a widely recognized metric in the field of machine translation that evaluates the correspondence between a machine's output and that of a human (Papineni et al., 2002). The metric quantifies the quality of the machine-generated translation at the sentence level by comparing the co-occurrence of n-grams up to four words long with those of reference translations. Precision scores for each n-gram are calculated and combined using a geometric mean, supplemented by a brevity penalty to discourage overly short translations that might artificially inflate the score.

6.3 Results Analysis

We experimented with several configurations to determine which settings would yield the best translation quality and model reliability. Table 1 summarizes the obtained results.

The obtained BLEU scores show the good performance of the model with a highest value reaching 34.09 with the following model configuration:

Batch Size: 8; Maximum Sequence Length: 200; Learning Rate: 5e-5; Epochs: 10; Label Smoothing: Employed to make the model less confident in its predictions, thereby enhancing its ability to generalize from the training data to unseen data.

Table 1: Summary of experiments with different configurations.

Batch Size	Max Length	Learning Rate	Epochs	BLEU Score
8	200	5e-5	10	34.09
2	150	5e-5	5	29.89
8	200	5e-5	5	28.20
6	150	0.0001	5	21.46
32	128	0.00001	10	20.31
8	128	5e-5	15	23.93
6	150	0.001	5	14.56

This was particularly useful in managing the diversity of the Akkadian language.

The variance in results indicates the impact of parameters on model performance. The model used SentencePiece tokenization for both Akkadian transliterations and English translations, maintaining a consistent max sequence length. The AdamW optimizer with a linear learning rate scheduler (without warmup) optimized the training, and the cross-entropy loss function was used to measure learning efficiency.

7 CONCLUSION AND FUTURE WORK

This study has demonstrated the effectiveness of using a transfer learning approach to translate Akkadian transliterations into English. The best-performing model, leveraging the Helsinki-NLP/opus-mt-ROMANCE-en architecture, achieved significant translation accuracy, as evidenced by the highest BLEU, reaching nearly the same score as in the existing researches. These results underscore the model's capability to handle complex syntactic and semantic challenges presented by the Akkadian language.

Given Akkadian's complexity and resource constraints, label smoothing, a novel strategy in this setting, improves the model's capacity to generalize from training data to unseen data by reducing overfitting. Pre-trained on extensive parallel corpora from the OPUS collection, the model gains performance on low-resource languages by utilizing linguistic knowledge from many languages through transfer learning. By fine-tuning, the model can further improve its translation abilities by adjusting to the unique syntactic and semantic subtleties of Akkadian. For low-resource languages like Akkadian, where training data may be scarce or dispersed, this is especially crucial.

This work makes a substantial contribution to the field by showcasing the viability and potential of applying transfer learning and refined models for translating ancient languages like Akkadian, as well as the efficiency of transfer learning in deciphering ancient scripts. By emphasizing the value of parameter optimization and the potential of transfer learning, the work offers a strong basis for further research.

Future studies may focus on further optimizing the translation model by exploring more advanced neural network architectures and integrating larger, more diverse datasets to improve the model's linguistic understanding. Exploring the potential of

real-time translation tools and expanding the model's capabilities to include other ancient languages could also provide valuable insights and practical tools for scholars and linguists specializing in ancient texts.

REFERENCES

- Gutierrez, G., Gordin, S., Sáenz, L., Levy, O., & Berant, J. (2023). Translating Akkadian to English with neural machine translation. *PNAS Nexus*, 2(5), Article pgad096. <https://doi.org/10.1093/pnasnexus/pgad096>.
- Gordin, S., et al. (2020). Reading Akkadian cuneiform using natural language processing. *PLOS ONE*, 15(10), e0240511. <https://doi.org/10.1371/journal.pone.0240511>.
- Grayson, A. Kirk. (1987). *Assyrian Rulers of the Third and Second Millennia BC (to 1115 BC) (RIMA 1)*. Toronto: University of Toronto Press. Adapted by Jamie Novotny (2015–16) and lemmatized and updated by Nathan Morello (2016) for the Munich Open-access Cuneiform Corpus Initiative (MOCCI). Retrieved from <http://oracc.org/riao/Q005902/>.
- Lazar, K., Saret, B., Yehudai, A., Horowitz, W., Wasserman, N., & Stanovsky, G. (2021). Filling the gaps in ancient Akkadian texts: A masked language modeling approach. *arXiv*. <https://doi.org/10.48550/arxiv.2109.04513>.
- Yale University Library. (n.d.). A 3770-year-old Babylonian clay tablet written in Akkadian, containing the oldest known cooking recipes. Yale University Library Collection. Retrieved from www.library.yale.edu.
- Fragmented Akkadian Inscription. (n.d.). The Gallery, Trent Park Equestrian Centre, Eastpole Farm House, Bramley Road, Oakwood, United Kingdom; St James's Ancient Art, London, United Kingdom. Retrieved from <https://www.antiquities.co.uk>.
- Tiedemann, J., & Thottingal, S. (2020). OPUS-MT – Building open translation services for the world. *ACLWeb*. Retrieved from <https://aclanthology.org/2020.eamt-1.61>.
- University of Pennsylvania. (2011). The Royal Inscriptions of Babylonia online (RIBo) Project. Retrieved from <http://oracc.museum.upenn.edu>.
- University of Pennsylvania. (2011). The Royal Inscriptions of Assyria online (RIAo) Project. Retrieved from <http://oracc.museum.upenn.edu>.
- University of Pennsylvania. (n.d.). The Royal Inscriptions of the Neo-Assyrian Period. Retrieved from <http://oracc.museum.upenn.edu/rinap/>.
- University of Pennsylvania. (2011). Suhu: The Inscriptions of Suhu online Project. Retrieved from <http://oracc.museum.upenn.edu>.
- University of Pennsylvania. (2015). State Archives of Assyria Online - State Archives of Assyria Online (SAAO). Retrieved from <http://oracc.org/saa>.
- Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002). BLEU: A method for automatic evaluation of machine translation. Retrieved from <https://aclanthology.org/P02-1040.pdf>.