Machine Learning and Cuneiform Transliteration: How AI Can Help Historians Uncover the Ancient Empires of the Sumerians and Akkadians by DB Christenson

Introduction to Cuneiform and its Importance

Primary sources are often touted as the most valuable form of evidence for a historian. Researchers of more contemporary history, like the world wars, have the great advantage of extensive documentation of events, people, places, and more. Assyriologists have much less to work with. Due to the chaotic political situation and climate, archaeological sites are much more degraded than those of other ancient societies, like Egypt. Administrative documents, journals, trade receipts, and more are not only destroyed or heavily damaged but available in extremely small quantities compared to other periods of human history. The numbers are so small in comparison and yet, as of 2018 almost 90% of discovered cuneiform documents have been left untranslated.[1]

Enrique Jiménez, an assyriologist at Maximillians University, notes that half of all of human history is recorded in cuneiform tablets. The lack of translations is concealing the knowledge, lives, and politics of humanity's oldest literate societies. With an easier way to transcribe cuneiform tablets into modern language, historians and assyriologists can learn much more about the functioning of the ancient civilizations spanning thousands of years and the empires of Sargon and the Ur III period specifically.

[1] Sophie Hardach, "The Key to Cracking Long-Dead Languages?," BBC Future, February 24, 2022, https://www.bbc.com/future/article/20181207-how-ai-could-help-us-with-ancient-languages-like-sumerian.

This research has been guided by the question: to what extent can machine learning and other machine models aid historians in revealing information about the Akkadian and Ur III societies and could this information lead to the cementation of these civilizations as empires? I argue that several teams of researchers have shown the efficacy of using AI for transliteration of cuneiform and their efforts will facilitate the transliteration of thousands more documents in a much quicker fashion. Additionally, because of the organization and quantity of administrative documents from the Ur III period, these new translations will give researchers much greater insights into the bureaucratic nature of these civilizations.

Challenges of Manual Cuneiform Transliteration

To prove the merit of AI models for transliteration, we must look at the complexities of cuneiform that has led manual transliteration to be a difficult and time-consuming process. Cuneiform signs were used to represent multiple languages through its existence such as Akkadian, Hittite, and Sumerian. Using most of the same signs for different languages is standard practice today like how the latin alphabet can be used for English, Spanish, French, and many more languages. Sumerian, however, is the most ancient of these languages and holds a unique trait: it is a language isolate.[2] In the Akkadian and Ur III periods, Sumerian was still the primary language recorded using cuneiform. Its status as a language isolate makes it very difficult to find proper translations for it and, because cuneiform was in early development at this phase, the signs themselves are difficult to interpret in terms of grammar and interpersonal relationships.

[2] 1. Jarle Ebeling, “Cuneiform Writing,” ETCSL, June 28, 2005, https://etcsl.orinst.ox.ac.uk/edition2/cuneiformwriting.php.

There also exists a lack of scholars qualified for the reading and translation of cuneiform documents. No exact number is given when considering multiple sources but the estimates seem to be in the low hundreds if not less.[3] This is a huge bottleneck when the corpus can span hundreds of thousands of tablets, the timely analysis of all of them is impossible with such a tiny number of capable translators. With tablets being authored by different scribes who may exhibit different hand writing styles, recovered in various states of preservation, among other complexities these tablets take a long time to process, transcribe, and annotate manually. Human transcribing is also very error-prone and inconsistent. A famous example of this is from the Sumerian King List—one of the most well-known and well-preserved works in relation to the dynasties of Sumer. Its original author is unknown but it had been copied and revised by scribes over time. Even with such a polished product, translators are unable to fully agree on some simple translations such as the time spans that each king ruled for. When a translation of a less known document is released, its translator may have a great impact on how the document is viewed which opens the door for bias in the intent, vocabulary, or personal interests of the text edition.

Machine Learning Methods in Transliteration

How can machine learning rectify some of the issues posed by human transliteration of cuneiform? Machine learning will be most useful to the field of assyriology by expediting the

[3] 1. Alison George, "How the Secrets of Ancient Cuneiform Texts Are Being Revealed by Ai," New Scientist, March 30, 2023, https://www.newscientist.com/article/mg25533981-400-how-the-secrets-of-ancient-cuneiform-texts-are-being-revealed-by-ai/.

translation process of undocumented tablets for scholars. If machines can rapidly output translations and annotations for untranslated tablets, the corpus of hundreds of thousands of unread text will begin to shrink at blinding rates. First we will look at a high-level overview of the techniques that researchers have used to solve the problem of cuneiform transliteration. Then we will examine recent projects on the topic and measure how successful AI has been in terms of accuracy and speed of translation.

Ample data is the most important first step to training a model that is capable of automatically transliterating cuneiform. The Cuneiform Digital Library Initiative is at the forefront of cataloguing all information about cuneiform tablets translated and untranslated. These standardized images and data attributes are digestible by machines and allow them to be as accurate as possible when presented with an untranslated tablet.

The process of translation for a machine is similar to the process for a human and is broken down into 3 general steps.[4]

1. Identifying each sign on the tablet and converting it to an appropriate transliteration. Because cuneiform signs can take on the value of a logographic, phonetic sound, or determinative it is important to associate each sign with all its possible meanings for a more accurate translation.

[4] 1. Shai Gordin et al., "Reading Akkadian Cuneiform Using Natural Language Processing," PloS one, October 28, 2020, https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7592802/.

2. Using natural language processing, tokenize and segment the words in the ways that make the most sense. Cuneiform symbols are often combined to give them different meanings, segmenting them and finding the most probable combinations is essential.

3. Finally translate the segmented text into a modern language.

Each step comes with its own unique problems but recent studies follow this breakdown to simplify and optimize the problem as far as possible. For example, if step 1 is not predicting the correct signs, then steps 2 and 3 become meaningless.

In 2020, University of Chicago professor Sanjay Krishnan was invited to work on the problem. The then Oriental Institute (now known as the Institute for the Study of Ancient Cultures—ISAC) was a massive data contributor for the project and, by the end of the study, professor Krishnan and the Neubauer Collegium boasted an 80% accuracy[5] for the deciphering of cuneiform signs. They and other teams understand that 100% accurate translation are close to impossible with current technologies, but perfection was never the goal with machine translated cuneiform. Susanne Paulus, the Tablet Collection Curator at the ISAC, states that even with imperfect accuracy, “If the computer could just translate or identify the highly repetitive parts and leave it to an expert to fill in the difficult place names or verbs or things that need some interpretation, that gets a lot of the work done.”

[5] 1. Rob Mitchum, “How Ai Could Help Translate the Written Language of Ancient Civilizations,” University of Chicago News, accessed May 17, 2023, https://news.uchicago.edu/story/how-ai-could-help-translate-written-language-ancient-civilizations.

Although these models are impressive, they are still extremely limited, particularly when it comes to making predictions across different collections of tablets. Different time periods and locations where the tablets were made can mean that sign values or handwriting are different. This process is particularly useful for when there exists a large amount of tablets from a centralized region and time period. This will be important later in the investigation of the Ur III empire.

Significance of Administrative Documents in Empire Building

Pivoting away from technical details for a moment, we must ask what we are trying to achieve by translating these documents. Not all historians agree that Sargon and the Ur III dynasty were true representations of empire. Because administrative documents mainly represent the bureaucratic tendencies of an empire, they are the ideal type of document to translate if we want to understand how the Ur III dynasty operated and why its operations have the potential to be "empire-esque." To find out what we should watch out for in these documents we need to first define what makes an empire an empire.

According to Howe,[6] empires require some qualifications before they can indeed be classed as empires. An empire is big in reach and diverse in population. However, all the people of the empire are not equal, there must be a hierarchy that places the conquerors at the top and the conquered at the bottom. This leads us to the next point: empires must be built through conquest, warfare, and violence. The maintenance of an empire also requires violence and the judicial

[6] 1. Stephen Howe, *Empire: A Very Short Introduction* (Oxford: Oxford Univ. Press, 2010), 15-19.

system to make it known that the rulers are not afraid of realizing this violence. The central state has ultimate authority but its surrounding states govern themselves for the most part. As long as the vassal states abide by the rulers taxes and laws they are protected by the central authority. We know that the Ur III dynasty participated in and arguably filled out each of these criteria due to the administrative record keeping of the Ur III government.

Administrative documents have given us insight into the law codes[7] of the period, correspondence between the capital and local elites of city states spread all throughout Mesopotamian region. We also know that those who were conquered were turned into slaves which exemplifies the caste like society that those under the empire lived in. The existence of administrative documents themselves also tells us about the bureaucratic efficiency of the empire. Although a majority documents only come from the central states in the Ur III dynasty, all kinds of records of economics, warfare, etc. that were important to the state were being documented and kept.

Ur III Dynasty Data as the Perfect Candidate for AI Transliteration

We now know about the limitations of the current AI models in that they lack the ability to predict well over different collections, time periods, and locations if they do not have extensive training data on them. Interestingly, across the Ur III and Akkad periods, the distribution of the number of recovered administrative documents is greatly skewed towards some specific years during the Ur III period.

[7] 1. Martha T. Roth, Piotr Michalowski, and Harry A. Hoffner, *Law Collections from Mesopotamia and Asia Minor*(Atlanta, GA: Scholars Press, 1997).

Garfinkle writes that "the sheer volume of extant texts from this era[Ur III] has convinced us not only that this was a highly organized state, but also that the central power of the state was absolute."[8] The Ur III period produced more textual sources that almost every other period of time in the BC era combined—about 90,000 documents. Additionally let us consider where all of these texts were written. Astonishingly, about 77% of these documents came from just 3 locations in the Ur III dynasty.[9] This specific collection of documents comes from a very centralized place both periodically and geographically.

Visualizing the data allows one to almost demarcate the rise, golden age, and fall of the Ur III dynasty simply by volume of documents produced during that period. A failing empire is unable to maintain the strict regiment required for record keeping. This behavior gives the impression of the life cycle of an empire.

These conditions also create the perfect storm for the introduction of AI to the problem. If scholars began annotating and translating more Ur III documents, models would use that additional training data and iterate upon itself giving better and better translations as a result. The model would be able to better pick up on the minute style differences of the period like handwriting, grammar, and more. The huge corpus of available documents also lets the model normalize its predictions over the handwriting of different scribes, different shapes in tablets.

[8] 1. Steven J. Garfinkle and Manuel Molina, From the 21st Century B.C. to the 21st Century A.D.: Proceedings of the International Conference on Neo-Sumerian Studies Held in Madrid 22-24 July 2010 (Winona Lake, IN: Eisenbrauns, 2013). 154

[9] 1. Steven J. Garfinkle and Manuel Molina, From the 21st Century B.C. to the 21st Century A.D.: Proceedings of the International Conference on Neo-Sumerian Studies Held in Madrid 22-24 July 2010 (Winona Lake, IN: Eisenbrauns, 2013). 155

Conversely, there is not too much variation in handwriting or shape because all of these scribes would have been trained in the same schools under the same people because of their geographic and temporal proximity. The model may also be able to pick up on domain specific knowledge of the Ur III dynasty from the training data it was given. It may be able to learn the format of legal or economic documents which would better its prediction capabilities.

In the circumstance of the Ur III dynasty, AI has the potential to help understand the period without any additional technological advancements. Its unique stature as a relatively centralized and short-lived empire makes it a prime candidate for analysis by AI. Uncovering the mysteries held within the remaining 50,000+ documents may also shed some light on the imperial practices that the Ur III dynasty employed and further solidify its classification as an empire—one of the first.

Conclusion

In conclusion, the application of machine learning and computer vision techniques to the transliteration of cuneiform tablets presents a transformative opportunity in historical analysis. By leveraging these technologies, we can overcome the challenges posed by human biases, ensuring greater accuracy, objectivity, and efficiency in deciphering and translating ancient texts. The scarcity of qualified translators and the limitations of manual methods have hindered our understanding of the administrative documents from the empires of Sargon of Akkad and the Ur III empire. However, AI transliteration offers a solution by providing consistent and standardized interpretations, while data-driven algorithms reduce individual biases. These advancements enable us to gain deeper insights into the political, economic, and social systems of these empires, proving their status as true empires through the identification of imperial characteristics in the administrative records. Collaboration between AI and human expertise is vital, as we enter a new era of historical research, unlocking the secrets of the past and advancing our understanding of ancient civilizations.

Works Cited

Ebeling, Jarle. "Cuneiform Writing." ETCSL, June 28, 2005. https://etcsl.orinst.ox.ac.uk/edition2/cuneiformwriting.php.

Garfinkle, Steven J., and Manuel Molina. *From the 21st Century B.C. to the 21st century A.D.: Proceedings of the International Conference on Neo-Sumerian Studies held in Madrid 22-24 July 2010*. Winona Lake, IN: Eisenbrauns, 2013.

George, Alison. "How the Secrets of Ancient Cuneiform Texts Are Being Revealed by Ai." New Scientist, March 30, 2023. https://www.newscientist.com/article/mg25533981-400-how-the-secrets-of-ancient-cuneiform-texts-are-being-revealed-by-ai/.

Gordin, Shai, Gai Gutherz, Ariel Elazary, Avital Romach, Enrique Jiménez, Jonathan Berant, and Yoram Cohen. "Reading Akkadian Cuneiform Using Natural Language Processing." PloS one, October 28, 2020. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC7592802/.

Hardach, Sophie. "The Key to Cracking Long-Dead Languages?" BBC Future, February 24, 2022. https://www.bbc.com/future/article/20181207-how-ai-could-help-us-with-ancient-languages-like-sumerian.

Howe, Stephen. *Empire: A very short introduction*. Oxford: Oxford Univ. Press, 2010.

Mitchum, Rob. "How Ai Could Help Translate the Written Language of Ancient Civilizations." University of Chicago News. Accessed May 17, 2023. https://news.uchicago.edu/story/how-ai-could-help-translate-written-language-ancient-civilizations.

Roth, Martha T., Piotr Michalowski, and Harry A. Hoffner. *Law collections from Mesopotamia and Asia minor*. Atlanta, GA: Scholars Press, 1997.