

Reading & deciphering ancient writing systems with AI

 openaccessgovernment.org

7 December 2022



Shai Gordin, Senior Lecturer at Digital Pasts Lab, Ariel University in Israel, provides intriguing insights about reading & deciphering ancient writing systems using AI

Can computers read ancient language systems better than experts? And why would we want AI to do that? How exactly do computational models read texts?

Computational methods to study ancient texts & scripts

There are multiple benefits of using computational methods for studying ancient texts and scripts, like Egyptian hieroglyphs or Mesopotamian cuneiform writing. Imagine all ancient texts were available with an effective search engine. You can search for specific words or sign sequences, or even general keywords. The benefits of searchable databases for texts in hieroglyphic, cuneiform, or any of the alphabetic scripts of the ancient world are already well established, with websites like Trismegistos, the Cuneiform Digital Library Initiative, Hethitologie Portal Mainz, or the Comprehensive Aramaic Lexicon Project, to name a few.

Another benefit is consistency in the work process. Often historical research involves reading substantial amounts of texts and extracting relevant information. In other words, it is a task of meticulous close reading, later synthesised by an expert. Doing this process digitally helps to maintain consistency and accuracy. In fact, saving pieces of data in digital format is not at all different from traditional methods of card-file catalogues. Instead of storing small papers with pertinent information in specific boxes, they are stored digitally in files that are like cabinets: their structure, which is easy for a computer to process, is like folders within folders, which can easily take you to the information you are looking for. The important difference is that when these folders and cards are digital, you can easily rearrange them, visualise, explore them and reuse them in ways that were not possible before.

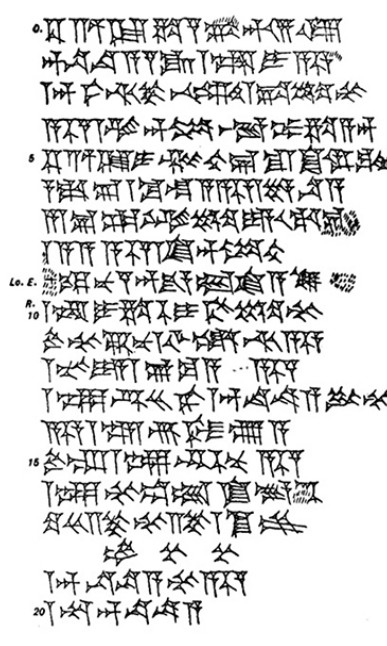
	<p>Obv. 1 1/3 2 1/2 gin kù-babbar nig-ga ^d15 ù ^dna-na-a šá šu¹¹ ^den-i a-šú šá ^{1d}dù-ù-kám ina muh-hi ¹ta-li-mu a-šú šá ¹gi-^damar-utu ina iti gan kù-babbar a 5 1/3 2 1/2 gin i-nam-dim é-su ki-sub-ba-a' šá da é ¹su-la-a a-šú šá ¹si-na-a a ¹é-sag-gil-li-a u da é ¹a-a a-šú šá ¹ki-^damar-utu-din Lo.E. [maš]-ka-nu šá ^dgašan šá unug^{ki} a-di-[j] Rev. 10 ^{1d}en-i kù-babbar-šú i-šal-li-mu lú mu-kin-mu ¹ši-rik-ti a-šú šá ¹numun-ia ¹kal-ba-a a-šú šá ^{1d}nà-re-man-ni ^{1d}na-na-a-šeš-mu a-šú šá ^{1d}in-nin-re-šu-ù-a 15 lú umbisag ^{1d}nà-en-šú-mu a-šú šá ^{1d}nà-mu-gin unug^{ki} iti kin u-22 kam mu 2-kam ¹ku-raš lugal kur-kur ^{1d}na-na-a-mu a-šú šá 20 ¹ir-^dna-na-a</p>	<p>(1)22.5 Shekels of silver, from the property of Ištar (2)and Nanaya, issued by Bēl-nā'id, son of (3)Bānītī-ereš, are owed by Talimu, (4-5)son of Mušallim-Marduk. He will give the aforementioned silver, 22.5 Shekels, in the Month of Kislimu (9th month). The unbuilt plot for his house, (6-8)which neighbours the house of Šulāya, son of Iddināya of the Esaggilāya Family and the house of Aplāya, son of Itti-Marduk-balātu, (9-10)serves as pledge of the Lady of Uruk, until he will pay Bēl-nā'id his silver. (11-14)Witnesses: Širikti, son of Zerīya. Kalbāya, son of Nabū-rēmanni. Nanāya-aḥu-iddin, son of Innin-rešūa. (15-18)Scribe: Nabū-belšunu, son of Nabū-šumu-ukīn. Uruk, the month of Ulūlu (6th month), 22nd day, 2nd (regnal) year of Cyrus, king of all the lands. (19-20)(additional witness:) Nanāya-iddin, son of Arad-Nanāya.</p>
--	--	--

Figure 1: From Left to Right, the original cuneiform line art, the transliteration, and the translation of the Achaemenid period Babylonian text Yale Oriental Series 7 11.

Computational text analysis is far from simply translating analogue methods to digital ones. Using computational methods with ancient texts and scripts can open new avenues of research that were completely impossible before. Computers understand text very differently from humans. A computer often needs to translate words into numbers or vectors, and then it can analyse lexical similarities and meanings using mathematical formulae. While this practice is completely foreign to traditional textual analysis, it opens an opportunity to view texts from a new perspective. This is just one example of many possibilities. Moving in a digital direction also provides opportunities for interdisciplinary research on an unprecedented scale. Think of the card-file catalogue example from above. When it is analogue, these boxes are of use to a limited number of scholars with access, and they cannot be connected to catalogues of other scholars from the same or adjacent fields. But if they are digital, these files can be shared, compared, and joined together through the semantic web.

Natural language processing (NLP) models

While digital corpora of ancient texts are constantly growing, many are still far from full digitisation. Reading ancient documents is difficult, the decipherment takes time, the size of the corpus is substantial and may include several ancient languages. Natural language processing (NLP) models can aid these tasks. NLP is a subfield in computer sciences and linguistics, that studies and develops methods for computers to understand and analyse human languages. The best models are often neural networks, advanced machine learning (ML) models whose mechanism attempts to imitate the workings of the human mind, ergo their name.

But while a human does not need to see millions of images of cats and dogs to differentiate between the two, neural networks do. This is not a problem when training models on the English language, for example. It is easy to extract tens of millions of words in English from the internet. Many ancient languages, however, fall under the category of low-resource languages – these have a limited number of examples, and some of them, like Akkadian, also have more complex morpho-syntactic structures. The upside of some text genres, even historical ones, is that many are formulaic and repetitive, unlike modern languages which are highly variable in nature. This counteracts, to a certain extent, the sparseness of available texts.

One of the common tasks in NLP is restoring masked words in a sentence – a task which, in practicality, is equivalent to restoring broken passages in ancient texts. Based on the complete sentences which the model has seen, it can predict how to restore fragmentary sentences. The more examples it will see, the better the results are. My research team at the Digital Pasts Lab, as well as another group from the Hebrew University, have shown the effectiveness of NLP models for restoring fragmentary cuneiform tablets written in Akkadian.

A related NLP task is predicting sequences of words or in the case of cuneiform, which can also be signs. By using Neo-Assyrian royal inscriptions and their equivalent Unicode cuneiform glyphs, it was possible to effectively train a model to provide transliteration and segmentation of cuneiform signs. This was performed using neural networks, as



Figure 2: Fragmentary obverse of the house sale contract YBC 7424. Courtesy of the Yale Babylonian Collection. Image credit: Klaus Wagensonner (Yale University, New Haven, CT).

before, and statistical models, which look at the frequency in which certain sign readings follow each other. The above tasks could be performed, by training models on existing digital texts. The greatest challenge, however, lies in optical character recognition (OCR), the visual identification of cuneiform signs; be that from drawings of cuneiform tablets, 2D images or 3D scans.

The future of reading ancient texts and scripts

The success of this kind of research allows us a glimpse into what will reading ancient texts and scripts look like, if computers will be able to perform all tasks, from visual recognition to transliteration and segmentation, even translation. Can we then say that the models are reading texts, like a scholar? Not yet. It is important to remember, that even when the models perform well, they do not perform perfectly. There is always a margin of error. Besides the fact, that there is not necessarily an ultimate truth when reading and reconstructing ancient texts – some issues are open to interpretation.

The scholar, then, needs to guide the models, correct them and analyse their results from a humanistic perspective. The combination of human and machine-based approaches will offer fresh perspectives on classical problems, as well as raise new avenues for research. Lastly, digitally curating ancient texts in this fashion, will allow smaller or marginalised fields of research, like studies of Indigenous and ancient civilisations, to join the global community of knowledge and make this specialist field more present and impactful in historical research at large.

Please Note: This is a Commercial Profile



This work is licensed under a [Creative Commons Attribution-NonCommercial-NoDerivatives 4.0 International License](https://creativecommons.org/licenses/by-nc-nd/4.0/).