Using computational tools to translate ancient cuneiform writing

Project introduction:

This project aims to translate cuneiform writing into English with the help of computing tools for a quite small dataset. As considerations for the future, I can maybe try my hand at not transliterated/not translated cuneiform writing or maybe try with other ancient languages.

Statement of significance:

Cuneiform writing appeared in what is now southern Iraq around 3000 BC (before Christ), developing from pictograms, that is, "drawings" of the realities they represented while gradually transforming into the wedge-shaped signs that give the name to the writing – figure 1(Homburg et al. 2023). They appeared on various mediums, such as clay tablets and rock carvings, while also being used for various genres, such as royal decrees, letters, administrative documents (Gordin et al. 2020).

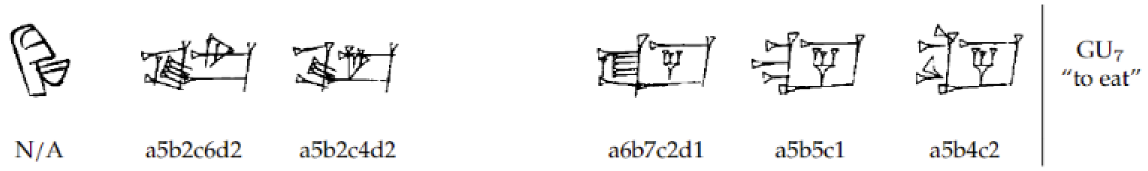


Figure 1. Adapted from Homburg et al. 2023, Table 1. It represents a concept – to eat – from a drawing connected to the reality to the left (a head with a bowl near the nose) transforming into a wedge-shaped drawing to the right.

Cuneiform writing seems to have been originally developed for Sumerian, and after that, it started to be used by another language, Akkadian, which was not related to Sumerian, although some Sumerian words remained in the Akkadian language. The Akkadian language also split later into two main dialects – Assyrian and Babylonian, which also have their dialects, such as Neo-Assyrian, Late Babylonian, all of which were written in cuneiform (Jauhiainen et al. 2019). The corpus of cuneiform writing is already huge, which still grows. These texts can provide a wealth of information regarding ancient Mesopotamian civilizations (Gordin et al.2020).

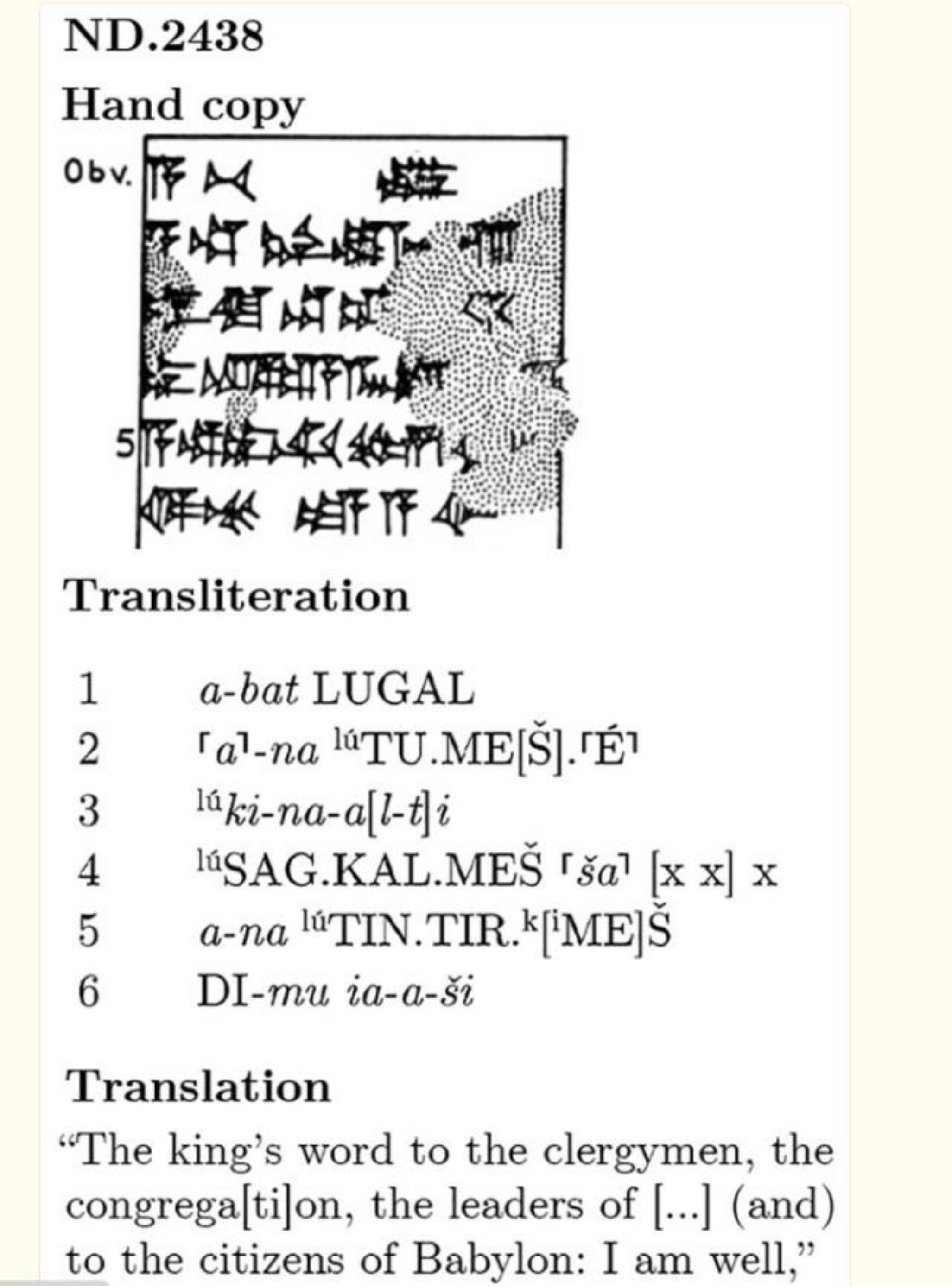
It may seem that this can be of interest only for a small group of people, such as these with a professional interest in ancient civilizations (archaeologists, for example). There is also a story, when NASA sent the Voyager spacecrafts into space, they attached to each spacecraft a phonograph record with greetings in 55 languages, two of which were Akkadian and Sumerian, and it seems that this was done for two audiences – these inhabiting Earth and these inhabiting other places than Earth (https://voyager.jpl.nasa.gov/golden-record/whats-on-the-record/greetings/).

Here comes the emphasis on large quantities of data (such as many cuneiform translations) - if more and more people come across these translations, such as when many people heard about the launching of spacecrafts, these events may make people think about their place in history, even in universe, which in turn, may lead in how they decide to live their lives.

In more practical terms, translating cuneiform writing into modern language with the help of computers will not replace the work of human scholars who translate it, but instead help them keep up with the huge corpus of cuneiform tablets which continues to grow. As will be explained below, many of the computer-done translations from my project clearly need to be refined by a human scholar.

Also, to put it bluntly, what the computer does when translating into modern languages is providing people with little to no knowledge of these ancient languages (like me), with the meaning of texts (to some extent). Even if it cannot transliterate and/or translate everything and/or does it less correctly than a scholar would, it will be a start in giving access to this knowledge important for humanity to people that are not experts in these languages. If you do not know about it, how can you love it?

Computational transliteration & translation of cuneiform writing:



ND.2438

Hand copy

Transliteration

1 *a-bat* LUGAL
2 ⸢*a*⸣-*na* ^{lú}TU.ME[Š].⸢É⸣
3 ^{lú}*ki-na-a*[*l-t*]*i*
4 ^{lú}SAG.KAL.MEŠ ⸢*ša*⸣ [x x] x
5 *a-na* ^{lú}TIN.TIR.^{k}[^{i}ME]Š
6 DI-*mu ia-a-ši*

Translation

"The king's word to the clergymen, the congrega[ti]on, the leaders of [...] (and) to the citizens of Babylon: I am well,"

Figure 2. From Gutherz. G et al. 2023, figure 10. It shows the process of translating from a tablet inscribed with cuneiform writing.

Human scholars got their hands on the tablets with the cuneiform signs (for example by photographing the tablets), then transliterated the signs (wrote them in the Latin script), and then translated the transliteration into modern languages. The transliteration step helps in gauging the context of the words, because cuneiform has words with multiple meanings (Bogacz and Mara 2022). It however, seems that now computers are capable of doing this work too. The starting point is by digitizing collections of cuneiform tablets. Quite a lot of repositories are now available online. Take as an example the Open Richly Annotated Cuneiform Corpus (ORACC), which is connected to the online materials of University of Pennsylvania and from which my dataset was taken. In this repository, the cuneiform corpora appear as transliterations and translations.

The dataset:

My dataset is created by a team of researchers which presented computational identification of languages in which cuneiform was written at the VarDial workshop (about computational work in linguistics) in 2019, taking place in the United States. They used the ORACC repository to select transliterated texts from there and created a tool, Nuolenna (available on Github), which transformed the transliterations back into cuneiform signs (Jauhiainen et al. 2019).

They chose only the lines of text written entirely in one language per line, leading to the creation of a new dataset for Sumerian and six Akkadian dialects (Old Babylonian, Middle Babylonian peripheral, Standard Babylonian, Neo-Babylonian, Late Babylonian, Neo-Assyrian). The lines were taken from various corpora of texts created for different projects. In total, the final dataset had around 139000 lines of cuneiform writing, on which they used the HeLI supervised-learning method to identify languages (Jauhiainen et al. 2019).

It seems it was more difficult for the model to identify languages on corpora of texts having a different genre that the texts it had been trained on, and also it was more difficult to assign a shorter line, or a line instead of a collection of lines, to a language, because some signs can be present in more languages. It also seems that it was tried to identify the languages using the same supervised-learning approach, but after removing duplicate lines and lines that were shorter than three characters. In short, it seems that the approach was carried with and without duplicate lines and lines shorter than three characters (Jauhiainen et al. 2019).

Although using lines shorter than three characters may seem risky, given the possibility of being assigned to the wrong language, to me it can be a way in which I can use the possible language as a point of reference when translating to English, as will be explored below.

The dataset with which I worked was a part of the larger dataset and included the first 3999 lines of cuneiform writing (this number is when containing duplicates), which, if we take the language assigned into account, were taken as belonging to the Late Babylonian language (a dialect of Akkadian). It seems that I worked with a dataset containing duplicate lines and lines shorter than three characters. I also worked with a dataset in which the duplicates are removed, but the short lines kept. I am discussing this second dataset here because a version having the same lines several times probably contributes less than having a version in which each line is different.

Finally, I decided to use the dataset from this workshop, even if it was not initially in cuneiform, because I wanted to explore how the computer can work on texts that are themselves a product of computer manipulation and also to test the transliteration and translation on texts, that although impacted by the computer manipulation, were originally from different projects than these on which my transliteration and translation were used on, so it will be interesting to look at the results. Also, the researchers working with the package I used to automatically transliterate cuneiform into the Latin alphabet also used their package on cuneiform corpus created from transliteration using Cuneify tool (more on this tool regarding my dataset below), which was then transliterated back using their package (Gordin et al. 2020), so I can even compare between which method to use if you have the transliteration and want to transform it back into cuneiform, so I can compare here two datasets created in a similar way.

Realistically, this may be viewed as a try in gauging the capabilities of computers in transliterating and translating the cuneiform texts into modern languages, by using already transliterated texts to maybe see how good is the transliteration/translation. If after the try, you are quite convinced of the usefulness of a specific method to automatically transliterate/translate text that can be compared with the work done by scholars, you can then try to use it on cuneiform writing that nobody had worked on yet.

The computer in action – the translation process:

I used the Python package "akkadian" to first transliterate these lines, after which the transliterated lines were translated into English with a language model (known as "praeclarum/cuneiform") taken from Hugging Face, which was created to translate from both Sumerian and Akkadian to English.

Both of these tools are some years old. Findings related to the "Akkadian" package were published in 2020 – Gordin et al. 2020, but last year (2023), they also published a paper related to translation into English of cuneiform also using one corpus of texts that was already used for the 2020 paper (Gutherz et al. 2023), but now for the next step, so as we can see that in just a few years, development is quick.

However, regarding the 2023 paper, it seems that they chose corpora of texts that already have transliterations and translations. Regarding the model on Hugging Face, it does not have a publication associated with it (to my knowledge, only a blog post) and unlike the 2023 paper, it seems it only translated texts not translated before (https://praeclarum.org/2023/06/09/cuneiform.html).

I did not use the tools presented in the 2023 publication to translate my cuneiform texts into English because I knew about the paper only after doing the analysis, but I can be an interesting suggestion of future work, given that it seems to use a different computational approach in doing the work than the model on Hugging Face.

Discussion:

- Nuolenna and Cuneify:

To transform the transliterated lines into cuneiform, Nuolenna was used. The context is as follows: on the ORACC website, is a tool called Cuneify, which transforms the transliterations that are in ATF encoding into lines of cuneiform writing (wedges) in Unicode encoding. However, to create this dataset they created another tool, naming it Nuolenna, to do this job. The reasons for not using Cuneify instead were that Cuneify could not be downloaded on their machines and that, quoted in full here: "it does not handle the Unicode ATF transliteration" (Jauhiainen et al. 2019: 92).

I am however, quite unsure about the second reason, and because of that, I quoted it in full and I am trying to understand what they mean. From other internet users (on Reddit, for example), it seems that Cuneify had problems with various types of transliterated words when transforming from transliterations into cuneiform, although some opinions were that it was doing a good job (https://www.reddit.com/r/Assyriology/comments/m5wlei/signpad_because_cuneify_is_not_very_good/).

I am however not sure if that (difficulty in transforming the transliterations into cuneiform) is what the second reason not to use Cuneify refers to. This can be because they do not give more details than "it does not handle" – they are very vague when describing the problems of Cuneify, which does not help in identifying the problem. I also believe that we need to analyse the opinions of internet users carefully before making a judgement.

A possible further step regarding Nuolenna and Cuneify (and Cuneifyplus) is to check if they still work now, in 2024, by creating a dataset from ORACC transliterations into cuneiform Unicode on my machine.

- HeLI:

This is a supervised-learning method, based on probability distributions, with which the creators of the dataset I am also using identified the languages in which the cuneiform was written (Jauhiainen et al. 2019). This was already used before 2019, for example in 2016, in which a text in one language is assigned to a language from a set of languages. This is based on several language models, and one model is connected to a specific word in the text to be assigned. It can be said that these models predict the language based on words from text to be assigned (Jauhiainen et al. 2016). However, one problem is the difficulty to have a model with many words – also mentioned by the researchers (although keep in mind this was 2016 and developments are fast regarding the capacity and types of data with which these models work, also increasing the speed at which they work with the data).

However, it seems that the HeLI method was still useful several years after testing it on various datasets (including the one I used), because in 2022 (admittedly, it was also two years ago), they created the HeLI-OTS tool to identify languages, based on the same method (Jauhiainen et al. 2022). Although here I did not try to identify a language, it was, for me, a point of reference, when transliterating and then translating. What I myself did with the dataset (transliterating and translating) seemed to me like the next task in a chain of tasks done automatically by the computer, after I got a sense from what ancient language I am supposed to translate into English (which can be viewed as a previous task).

However, when using a dataset which already is assigned to one or more languages through this method, as in my case it may be also worth thinking about the cases in which this approach may have wrongly assigned a text to a language and about the number of these cases with regards to the correctly identified ones from the dataset. For example, in my corpus of roughly

4000 lines of text, there may be lines that are wrongly assigned to Late Babylonian, for example the lines having less than three characters (I kept these lines to have an idea of their translation into English). What could be a way forward is the creation of a tool to identify subtle differences between these dialects/languages, and I think it may involve a human.

- akkadian:

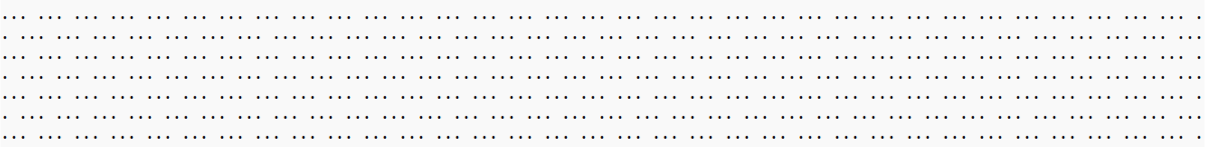
According to Gordin et al. 2020, they used this package to transliterate from other projects that were not connected to my dataset and also they seemed to use the several methods in the package – the Hidden Markov Model (HMM), Maximum-Entropy Markov Model (MEMM) and also neural networks – the Long-Short Term Memory (LSTM) and the Bidirectional Long-Short Term Memory (BiLSTM). It seems that in their case HMM had the lowest accuracy in transliterating, at around 89.5%, and BiLSTM, the highest – 96.7%. Their dataset contained around 23000 lines, so maybe the size of the dataset is connected to how each method will be (my transliterated dataset was smaller).

In my case, I only worked with the HMM to transliterate, given that, even this was several years ago, it was still reliable at 89.5% to accurately transliterate lines. A future step might to try to also use the other methods on my dataset and then to compare the accuracy between them. However, I also did not check the accuracy for HMM – maybe it has other value than 89.5%, and then try to improve the accuracy, but this can be a future step. What I aimed for in this project was to try some techniques to automatically transliterate and then translate a corpus in order to obtain some preliminary results – the translations – which were to be compared to the translations done by human scholars. Also, not least, to try to automatically transliterate texts that are from different initial transliterations.

- HuggingFace model:

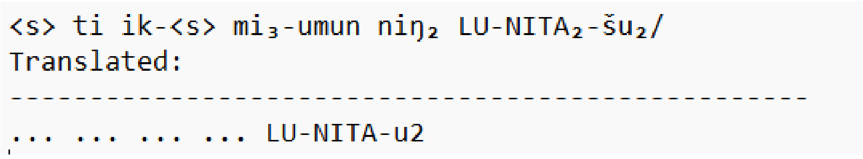
This model, the creation and training of which is explored at the blog post available at https://praeclarum.org/2023/06/09/cuneiform.html, was created by an "amateur" (not a professional) interested in cuneiform writing and who took a Large Language Model (LLM) – the T5 one, which already existed – and trained it on transliterated cuneiform. It was used on a corpus of diverse genres and from different periods, rising to a large number of 130000 already translated.

Regarding its limitations, I remarked several that were connected to the times when it probably did not understand how to translate and it appeared as: some lines of points, repeating the transliteration instead of translating or writing strange signs instead of translation – figure 3, figure 4, figure 5 – all of them in my output file.



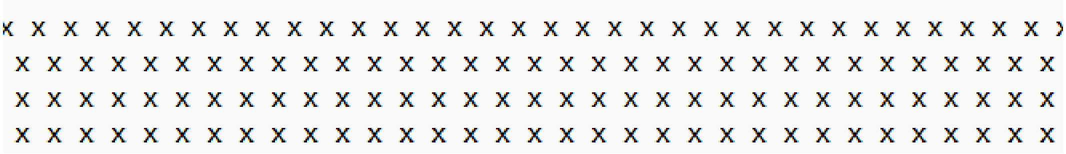
```
... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .
. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .
. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .
. ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ...
... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... ... .
```

Figure 3 – lines of points.



```
<s> ti ik-<s> mi₃-umun niŋ₂ LU-NITA₂-šu₂/
Translated:
--------------------------------------------------
... ... ... ... LU-NITA-u2
```

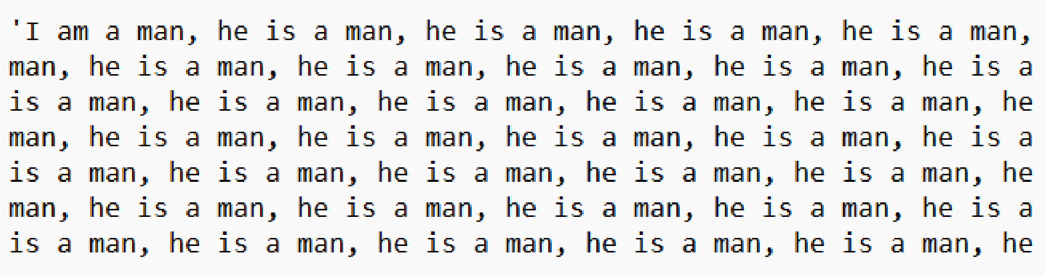
Figure 4 – also repeating part of transliteration to the right.



```
x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x
 x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x
 x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x
 x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x x
```

Figure 5 – strange signs as translation – many “x”

Another quite visible limitation appeared often – repeating a word without being necessary – figure 6.



```
'I am a man, he is a man, he is a man, he is a man, he is a man,
man, he is a man, he is a man, he is a man, he is a man, he is a
is a man, he is a man, he is a man, he is a man, he is a man, he
man, he is a man, he is a man, he is a man, he is a man, he is a
is a man, he is a man, he is a man, he is a man, he is a man, he
man, he is a man, he is a man, he is a man, he is a man, he is a
is a man, he is a man, he is a man, he is a man, he is a man, he
```

Figure 6 – repeating the translated words

It however, seems that T5 model has a preference for shorter sentences and in my translations, it seems that the shorter ones may fare better. As for comparison with human scholars translations, I am unfortunately not sure which paragraphs from the ORACC texts are these taken from (and one project from which such texts were taken seemingly does not exist now).

What it can be done is to try to associate the cuneiform lines to their transliterations in ORACC, line by line, to have an idea where to look for the human translation and then to try to compare the two.

Conclusion:

This started from my wish to try to discover how can we automatically translate cuneiform into English. Although I may have used datasets and/or methods that may not work in the future, this process of automatic translation should continue (even if datasets, methods change), because the cuneiform corpus continues to grow, but the human scholars will still need to assist with the checking of transliteration/translation.

References:

Bogacz B. and Mara H. 2022. Digital Assyriology—Advances in Visual Cuneiform Analysis. J. Comput. Cult. Herit.15, 2 : 38 (May 2022)

Gordin, Shai & Gutherz, Gai & Elazary, Ariel & Romach, Avital & Jiménez, Enrique & Berant, Jonathan & Cohen, Yoram. 2020. Reading Akkadian cuneiform using natural language processing. PLOS ONE. 15. e0240511. 10.1371/journal.pone.0240511.

Gutherz G, Gordin S, Sáenz L, Levy O, Berant J. 2023. Translating Akkadian to English with neural machine translation. PNAS Nexus. May 2:2(5):pgad096. doi: 10.1093/pnasnexus/pgad096. PMID: 37143863; PMCID: PMC10153418.

Homburg, T, Brandes B, Huber E-M, and Hedderich. M. A. 2023. From an Analog to a Digital Workflow: An Introductory Approach to Digital Editions in Assyriology. Cuneiform Digital Library Bulletin 2023 (4). https://cdli.mpiwg-berlin.mpg.de/articles/cdlb/2023-4.

https://huggingface.co/praeclarum/cuneiform the language model which I used to translate

https://praeclarum.org/2023/06/09/cuneiform.html

https://pypi.org/project/akkadian/ the Python library which I used to transliterate

https://oracc.museum.upenn.edu/ where to find the ORACC repository

https://voyager.jpl.nasa.gov/golden-record/whats-on-the-record/greetings/ ancient languages greetings on the record on the spacecraft

https://www.reddit.com/r/Assyriology/comments/m5wlei/signpad_because_cuneify_is_not_very_good/ opinions of Redditors regarding Cuneify

Jauhiainen T., Lindén K., and Jauhiainen H.. 2016. HeLI, a Word-Based Backoff Method for Language Identification. In Proceedings of the Third Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial3) :153–162, Osaka, Japan. The COLING 2016 Organizing Committee.

Jauhiainen T., Jauhiainen H., Alstola T., and Lindén K. 2019. Language and Dialect Identification of Cuneiform Texts. In Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects: 89–98, Ann Arbor, Michigan. Association for Computational Linguistics.

Jauhiainen T., Jauhiainen H., Linden K. 2022. HeLI-OTS, Off-the-shelf Language Identifier for Text. Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022): 3912–3922, Marseille.

Words of main text and citations: 2902

The files attached to the assessment are: the written report, the output file, the cuneiform font to be installed on your computer to view the signs, the Python script of the project and the original csv file with the cuneiform dataset and the languages associated with it.