Kathryn Kelley, Logan Born, M. Willis Monroe, Anoop Sarkar

Image-aware language modeling for Proto-Elamite

(doi: 10.1418/105965)

Lingue e linguaggio (ISSN 1720-9331)
Fascicolo 2, luglio-dicembre 2022

IMAGE-AWARE LANGUAGE MODELING FOR PROTO-ELAMITE

KATHRYN KELLEY LOGAN BORN
M. WILLIS MONROE ANOOP SARKAR

ABSTRACT: Computational methods and machine learning have been applied to problems in the study of Proto-Elamite writing in recent years, with results which confirm that the available digital format of the corpus and the work-in-progress signlist can support meaningful computational analysis. However, the specialist's system for transliterating the texts nonetheless determines the ways that algorithms have been learning sign relationships. This paper therefore explores how image-aware language modeling can address difficulties arising from biases introduced into the data during transliteration. This approach can help us to better understand the system of signs and to revise the proto-Elamite signlist, a fundamental task of the decipherment project.

KEYWORDS: Proto-Elamite, machine learning, computational decipherment, natural language processing, neural networks.

1. INTRODUCTION[1]

Of the remaining undeciphered ancient scripts known today, proto-Elamite is both the earliest (c. 3100-2900 BC) and one of the largest corpora, represented by upwards of 1,600 inscribed clay tablets. All these tablets come from sites in Iran, with over 1,400 texts and fragments excavated at the ancient city of Susa. "Proto-Elamite" conventionally describes the writing system, as well as an archaeological period and aspects of shared material culture (c. 3300-2900 BC)[2] exhibited at sites widely distributed across Iran (Abdi & Miller 2003). The first tablets found at Susa were published already in 1900 (Scheil 1900) and by the mid-20th century most Susa tablets now known had been published in image or drawing, along with some structural analysis, commentary, and lists of signs. However, limited access to quality images or accurate copies of the texts

[1] All proto-Elamite sign images in this paper are the work of J.L. Dahl, and sign numbers reference the working signlist available to download at `https://cdli.ox.ac.uk/wiki/proto-elamite`. Proto-Elamite texts are referred to by primary publication and CDLI "P" number (MDP 6, 213 / P008013).

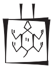
[2] Proto-Elamite script appeared only in the latter part of the more broadly defined archaeological horizon (Helwing 2013: 331).

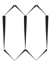
has been a factor in the delayed progress of decipherment (Dahl 2019: 68-69). Work by members of the *Cuneiform Digital Library Initiative* (CDLI) over recent decades has sought to remedy this through improved imaging and consistent digital encoding of the data.

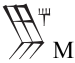
Although it is said that proto-Elamite is undeciphered, many features of the system are understood, as well as the approximate meanings of some signs. This knowledge is based on evidence including proto-cuneiform parallels, text structures, and sometimes pictographic referents. Proto-Elamite is an accounting system, tracking goods and personnel, probably with heavy emphasis on the distribution of barley, agricultural labor, and to a lesser extent animal husbandry. Texts can be divided into units (called an "entry", consisting of one or more non-numeric signs) that is followed by a numeric notation from one of the several deciphered numeric systems, most of which have parallel in neighbouring proto-cuneiform accounts (Friberg 1978; Englund 2004:107-18). Features of the accounts known as "headers" "summary lines" and "subscripts" can also be identified, if not yet thoroughly understood. A set of numeric signs are employed alongside a much larger variety of non-numeric signs (see below) that might include "object" and "owner" categories (Dahl 2019: 74) of a semasiographic/ideographic or logographic nature.

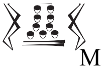
The use of a certain set of signs for their sound values, usually conceived as syllabically rendering personal names (perhaps in logo-syllabic spellings), has also been posited. Dahl has made no investigation of the types of sound values that may be represented, and the term "syllabary" remains used in a general sense. More recently, Desset *et al.* (2022) have suggested that newly proposed "vocalic, consonantal, and syllabic values" of later Linear Elamite signs may reflect "similar or identical phonemic values" in proto-Elamite, thus implying at least the possibility of a very early set of consonantal signs in addition to CV (but not VC) and V-type syllables. Meriggi's proposed syllabary of around 50 signs (1971: 173-4) did not stray quite so far from the contemporary developments in cuneiform by suggesting the possibility, based on the number of signs, that syllables of the CV, V and VC type may have existed (§458). The existence of this syllabary is not proven, however, and some recognisable sign ordering principles – such as the proposed "object final" tendency, and the ordering hierarchy of hypothesized worker signs at the tablet-level, as outlined for example in Hawkins 2015 – likely reflect bookkeeping structures as opposed to features of language.

A category of glyphs called complex graphemes consist of two or more signs combined, either one inside the other or placed adjacently, overlapping, or framing:


M157+M340


M296+M296


M59+M38


M54+M384+M54$_l$

Apart from the identification of any specific language(s) that may be (perhaps very partially) coded in the script, other major outstanding questions of interest in relation to a technical understanding of the corpus include the overall nature of the system of signs, its level of standardization, strategies for identifying humans, and the range of products and activities being tracked in the texts. With respect to the system of signs, some specific questions of interesting include: *What are the different functional sign categories and how do they work together to convey information? Which signs (if any) have more than one value or sign function, and through what mechanisms were multiple values arrived at? What is the significance of graphical variation on otherwise similar sign shapes?* Answers to these questions would inform our broader understanding of the history of the development of writing systems. An example of a specific content-related question is *What types of information does the account "header" of a text contain, and is there more than one type of header?* A more holistic "decipherment" would furthermore entail better understanding of the relationship between this technology – including the context for writing a tablet and the lifecycle of a tablet – and social and economic activity in proto-Elamite societies.

2. THE SIGNLIST CHALLENGE

A crucial step in the decipherment of a writing system is establishing a signlist that can confidently group all the tokens (individual attestations of a sign) – which may appear in minor or major graphical variations – into a given sign, and distinguishes these from other signs. This was accomplished, for example, for Linear B – another accounting corpus of roughly comparable size to proto-Elamite – by E. Bennett in the mid-20th century (Judson 2020 / Bennett Jr 1947). That script turned out to have 87 syllabic signs and around 170 ideograms (Judson 2020: 15) in a known corpus of around 6,000 texts representing c. 60,000 tokens, excluding numeric notations and word dividers.[3]

By comparison proto-Elamite has around a quarter the number of texts (c. 1700; c. 1500 of which are available in transliteration and include read-

[3] F. Aurora, personal communication, drawn from corpus presented in Aurora 2015.

able non-numeric signs) and contains c. 29,000 tokens. The data available for linguistic decipherment, however, are more limited than this initially suggests, since nearly half the tokens are numeric (c. 12,000) as opposed to non-numeric (c. 17,000). The length of proto-Elamite tablets is charted in Figure 1; most texts have fewer than 72 signs, but the longest known text contains 724 signs (198 of which are non-numeric). Thus, within a modest-sized corpus, the writing system displays a relatively high level of complexity that presents challenges – and perhaps opportunities – for the prospect of some forms of decipherment.

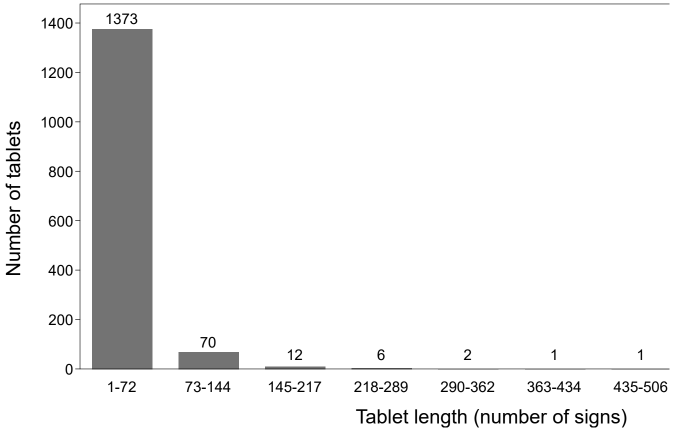
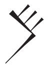
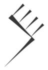


FIGURE 1: LENGTH OF PROTO-ELAMITE TABLETS BY NUMBER OF SIGNS.

A striking difference between the Linear B and the proto-Elamite corpus, which both speaks to decipherment prospects and informs strategies, is the current estimate for the number of signs. The CDLI corpus transliterations, using Dahl's working signlist, identifies 1623 distinct non-numeric signs by "sign names" labeled using the formula M1, M2 (etc.) that are adapted from Meriggi (1974a). This count includes a category of 249 complex graphemes as distinct signs, although virtually all of them are made up of two or more signs that are attested independently. It also includes large numbers of signs differentiated from their primary "M" number by a system of alphabetic subscripts (e.g. M1$_{b}$, presented in digital transliteration as M1∼b) which are presumed to be distinct signs until otherwise proven. A handful of signs receive additional numeric subscripts:

 M39$_{c}$  M39$_{c1}$

which indicates confidence that they represent graphical but not semantic variation. Work in determining which alphabetic subscripts might be merged is ongoing; the current signlist has already been reduced from an estimated 1,900 signs cited in Dahl (2002). If complex graphemes are considered as their component parts, the current list of signs is counted as fewer than 1400 signs, with the expectation that many signs might eventually be merged. On the other hand, Dahl (2019) has more recently suggested that "proto-Elamite was never standardized and every new discovery of texts will add to the numbers of signs and sign forms" (2019: 71).

There are several considerations in assessing this large number of signs. The first is that historical comparison to later cuneiform, as well as cross-cultural comparison, reveal that writing systems such as logo-syllabic Sumerian and Akkadian cuneiform can certainly have a full repertoire of upwards of 500 different signs. However, it has been argued that the two early writing systems of the Near East, proto-cuneiform and proto-Elamite, represent intermediaries between prehistoric accounting technologies and later cuneiform and other scripts adapted more fully to record spoken language (Damerow 2006). In addition, proto-Elamite represents a relatively short-lived technology, and it may take time along with consistency in cultural institutions to evolve a certain level of standardization in sign inventory.[4] These circumstances suggests that the number of graphically and functionally distinct signs in these two systems may be expected to reflect different properties from later incarnations of logo-syllabic cuneiform or other well-established and widely used "true" writing systems.

Dahl (2002) estimated that a more accurate sign count of proto-Elamite is likely to be comparable to that currently assessed for proto-cuneiform (c. 770 signs)[5] once alphabetic subscript variants (or other signs) in proto-Elamite are merged as appropriate. However, more comparison between the two scripts should be conducted to verify this posited similarity. A survey of Zeichenliste der archaischen Texte aus Uruk (Green & Nissen 1987) alongside the proto-Elamite signlist leaves the impression that proto-Elamite scribes engaged in a higher level of minor to major graphical variation on a given sign shape (Figure 2).[6]

[4] Proto-Elamite lasted long enough to evidence evolution across script phases: see (Dahl 2019: 64) for comment on "earlier" and "later" forms of signs. Note also the hypothesis presented in (Dahl 2009: 24) that "in the latest phase of its use proto-Elamite developed a syllabary where the values and graphical shapes were not yet standardised".

[5] As with proto-Elamite, similarly reduced from an interim assessment of around 1900 proto-cuneiform signs and variants (Dahl 2002: fn. 11).

[6] Comparison is complicated by the differing presentations of signs in the two lists; proto-cuneiform signs are presented in alphabetical order by applied Sumerian sign names, and the

Similarly, the mechanism of creating new signs in the form of complex graphemes appears generally more productive in proto-Elamite.

In some cases, it may be hypothesized that variation in basic sign shapes, such as additions of strokes on the depictions of vessels, may serve to indicate differences in the products they represent. In proto-Elamite one can observe cases in which signs are presented with minor variation within a single text in apparently meaningful ways, indicating that minor graphical variation on a basic sign form may be a specifically useful feature in some proto-Elamite contexts.[7] However, graphical elaboration on a basic sign shape might represent other phenomena as well, including simple scribal flourish or lack of established standardization.

Damerow (2006) discussed the "range of conventionalization" in the proto-cuneiform signary, noting that 530 signs at the time of his counting were attested only once, and a further 610 fewer than 10 times, 370 signs up to 100 times, and finally leaving a smaller core of 104 signs attested more than 100 times. He suggests that this "core" could "be flexibly complemented by modifications of existing signs or by the creation of new signs that were used only in specific contexts, and that never developed into standardized signs of cuneiform writing" (2006: 6). Dahl (2002) observed similar broad patterns in proto-Elamite sign distribution: by recent count[8] the number of proto-Elamite singleton signs is 740, or a little under half of the current signlist. Damerow points to the importance of the immediate social environment of the writing of particular groups of tablets for the creation of one-off or rare signs, and Dahl similarly links the large number of singletons to "ad hoc" sign creation in proto-Elamite (Dahl 2002, 2019: 71). High levels of standardization are thus not a given in these related "proto-writing" systems.[9]

"Reading" a proto-writing tablet may thus require high levels of knowledge about social context that is largely unavailable to us, especially since most tablets lack detailed archaeological context information. It thus becomes a question of considerable interest whether even moderately common signs

construction methodology for grouping tokens into signs involved reliance on later cuneiform lexical texts and signlists (which however may or may not always provide an accurate measure for proto-cuneiform signs and sign use – see e.g. Kelley 2021).

[7] A simple example may be MDP 17, 186 / P008348 in which M346, M6 (a possible graphical simplification of M346), and M346$_a$ are consecutively counted. Another example is MDP 6, 369, in which M264 and M263 variants (both depictions of vessels) appear in consecutive entries. Co-occurrence of variants within and across texts is worthy of further study.

[8] Obtained at `https://mrlogarithm.github.io/pe-pc-datasets-interface/pe.html`.

[9] See also Farmer *et al.* 2004:39 describing Harappan and Vinča symbols as possible "non-linguistic signs" with "standardization of a small core of signs [and] the inclusion beyond that core of hundreds of unique or rare symbols".

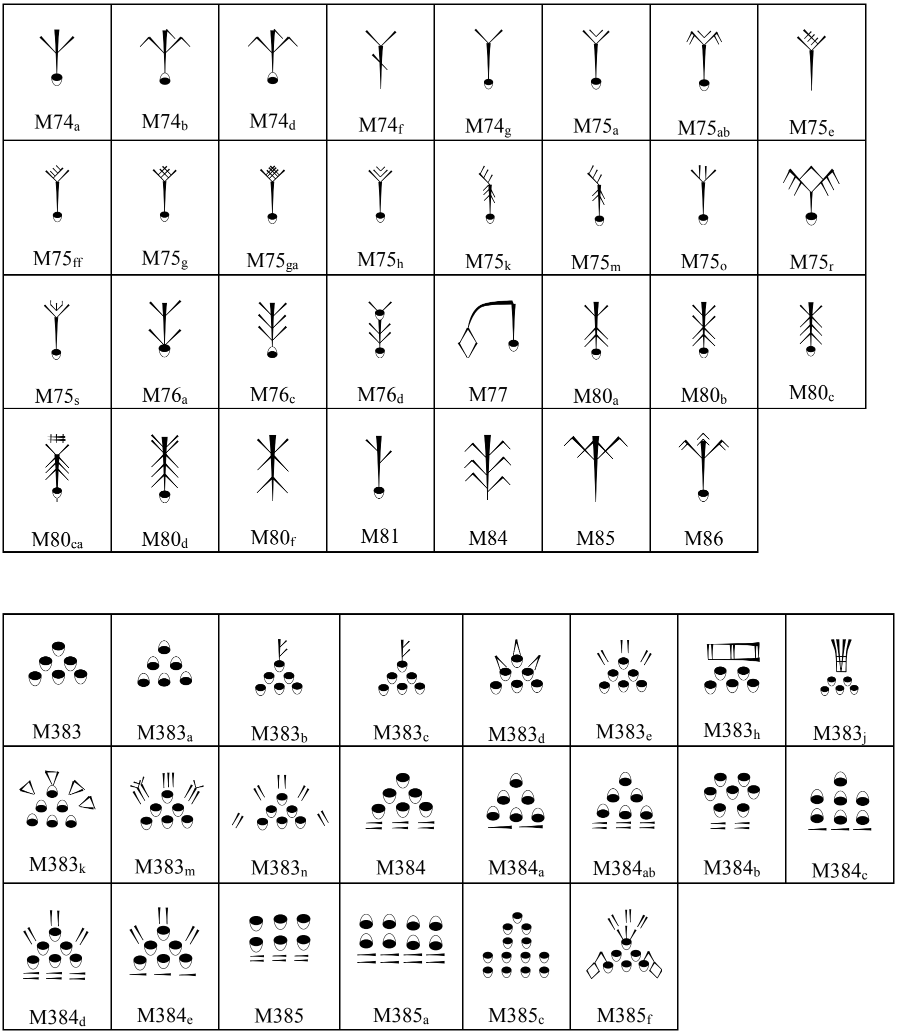
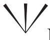
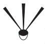
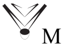


FIGURE 2: SELECTIONS OF GRAPHICALLY SIMILAR SIGNS FROM DAHL'S WORKING LIST. ABOVE: M74-M86 BELOW: M383-M385.

or indeed a core set of common signs themselves were sufficiently subject to standardized rules of use such that either conventional scholarly observation or powerful computational methods could reveal a text's meaning through recognising patterns. Promisingly, some recent research (outlined further below) using computational methods and machine learning confirms a certain level of systematization within proto-Elamite that extends beyond the more easily manually observable tendencies of sign use, and which may mean our understanding of the texts can still be improved in important ways, whether or not semantic contents can be precisely identified. A problem in the construction of the proto-Elamite signlist arises when one or more signs appear to be graphical intermediaries between two other signs, making it difficult to decide whether they should be kept distinct or assigned to one or another sign label. Consider the following:

 M72  M072$_a$  M73$_a$

Meriggi (1974a: 5) commented on the difficulty he had in grouping tokens and ordering signs in his list, which although certainly a vast improvement upon earlier publications (note Mecquenem 1949, which listed over 5,000 "signs"), retained problems: chiedo indulgenza pei segni 66-68, le cui varianti proteiformi fanno disperare "I ask for indulgence for the signs 66-68, whose protean variants make us desperate" and with respect to the broader ordering of signs by shape across the list, un tentativo di migiorarlo convincerà subito che si finisce sempre nell'arbitrario "An attempt to improve it will immediately convince you that you always end up in arbitration". Such apparently intractable situations of sign fluidity are encountered frequently when running through either the list of Meriggi or Dahl, and although Dahl's list is again a great improvement, many difficulties remain, which are compounded by likely errors in the transfer of damaged or poorly visible signs from clay (or older published line art) to modern line art. Although Dahl's signlist builds on c. 400 signs (with up to dozens of variants each) presented in Meriggi (1974a), Dahl's further contextual analysis has led to substantial deviations between the two lists. For a typical example, Dahl was able to collapse over a dozen forms of Meriggi's signs 66 and 67 (or remove probably misplaced token assignments) into M66 and three alphabetic variants, removing M67 as a label altogether.

3. COMPUTATIONAL METHODS AND MACHINE LEARNING: THE FIRST STEPS

The complexity of the data lends itself to advanced computer-assisted methods of analysis. R. Englund noted two decades ago (2004:127):

> The proto-Elamite texts do contain sign sequences which are distinctly longer than the average of those from Mesopotamia. The texts are therefore more likely to contain language-based syntactical information [...] statistical analysis of text transliterations should point

toward meaningful sign combinations of a fixed sign sequence which could reflect speech.

The extent and manner of linguistic information in proto-Elamite remains to be seen. Fortunately, the CDLI has paved the way for advanced computational study of the script by providing images, catalogue data, and transliterations of nearly all known tablets, implemented using a work-in-progress signlist constructed by J.L. Dahl. Below some first attempts at computational analysis on proto-Elamite are described.

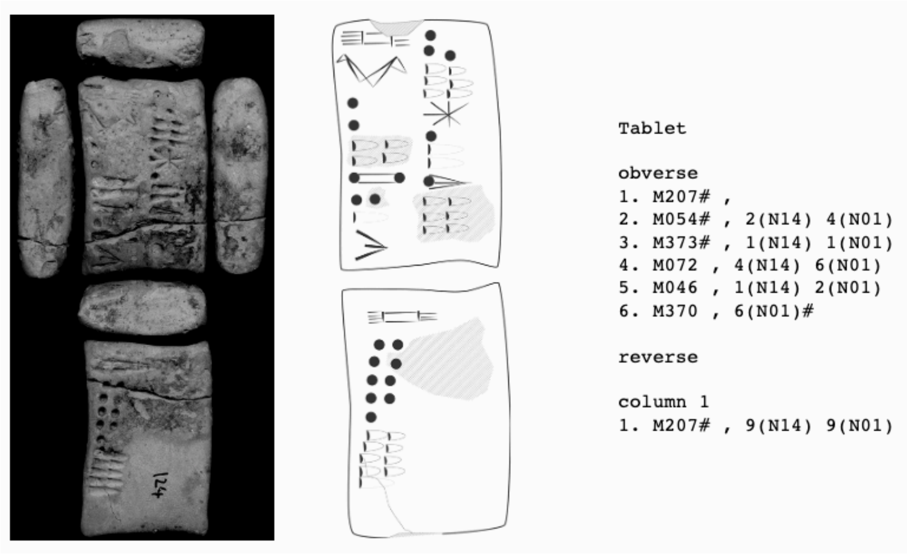
3.1 Data cleaning and improved searching

Having the digital transliterations provided by the CDLI already makes rudimentary hypothesis checking possible. Born *et al.* (2019) established better data-searching capabilities, supporting complex queries, incorporating metadata and specifying elaborate filters. This has opened the way to a wide variety of research questions. The process involved "cleaning" the CDLI data of some notational inconsistencies, and importing it into a Jupyter notebook. Jupyter is a platform that accommodates different programming languages, including Python. In this format, it is possible to conduct queries that incorporate more metadata and that look at entire texts rather than isolated entries, allowing for more sophisticated analyses. This notebook is available on Github,[10] an online platform for hosting code and for version controlling. Version controlling with a complete history of changes is a particularly useful feature for work on the proto-Elamite corpus, because proto-Elamite specialists use a signlist that is still being revised and a corpus that may be updated in the future. For any work that is published using the database on Github, a record is maintained of the corpus in the state in which that work had been done.

Improved control over the data leads to greater understanding of the corpus. For example, since the overall nature of the system of signs is not yet understood, information about the relative frequencies of different hypothesized sign types should be useful. We might ask, how frequently used across the corpus are the proposed syllabic signs (Dahl 2019:85)? Approximately half of the texts and fragments (with available transliteration) do not contain a sign from the proposed syllabary. Most texts which contain syllabic signs contain relatively few of them: less than 10% of signs in the average text are from the proposed frequent syllabary signs. Many other such overview questions can be efficiently answered.

[10] `https://github.com/sfu-natlang/pe-decipher-toolkit`.

[11] ASCII TRANSLITERATION FORMAT. See `http://oracc.museum.upenn.edu/doc/help/editinginatf/cdliatf/index.html`.



```
Tablet

obverse
1. M207# ,
2. M054# , 2(N14) 4(N01)
3. M373# , 1(N14) 1(N01)
4. M072  , 4(N14) 6(N01)
5. M046  , 1(N14) 2(N01)
6. M370  , 6(N01)#

reverse

column 1
1. M207# , 9(N14) 9(N01)
```

Figure 3: CDLI archival image and line art of tablet SE 124 / P009441, Couvent saint-Etienne, Jerusalem (left). ATF[11]encoding of text (right).

3.2 N-grams in proto-Elamite

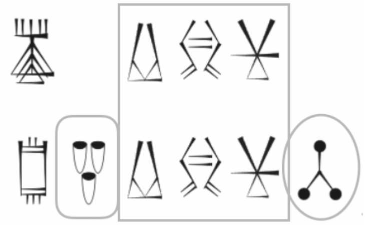
The study of sign ordering is central to traditional decipherment efforts. The linear ordering of signs in proto-Elamite is a pointed innovation in comparison to the irregular placement of signs within cases and the tabular format of the contemporaneous proto-cuneiform writing system. Proto-cuneiform has been characterized as mostly semasiographic, or non-glottographic – conveying meaning without the necessary intermediary of specific sound values. This estimation is made partly based on the lack of consistently linear sequences of signs. While proto-Elamite shares some fundamental features of the proto-cuneiform bookkeeping system, its strategy to arrange signs in strictly linear sequences has been offered as evidence that it may contain the earliest known syllabary.

While the length of proto-Elamite sign strings, which easily exceeds those in the neighbouring proto-cuneiform system (Englund 2004: 127; Dahl 2009: 24) is relatively promising for decipherment prospects, the low frequency of repeated strings across the corpus is less so. No sequence of more than 3 signs is repeated more than three times.[12] Around ten trigrams are repeated at least five times, around 50 appear three or four times, and the remaining 98% appear only once or twice (Born *et al.* 2019: 126). Low repetition may present

[12] See `https://mrlogarithm.github.io/pe-pc-datasets-interface/pe.html`.

a challenge to traditional decipherment methods that rely on the recognition of common nouns or grammatical features that can be linked to a known language. Proto-Elamite n-grams can be explored at a tool built by L. Born at: https://mrlogarithm.github.io/pe-pc-datasets-interface/pe.html

Using the Python notebook more advanced queries can be made, for example, what trigrams or above (4-grams, 5-grams) appear at Susa and at another site? The answer is only one, the sequence M110_a M242_b M96 which appears in two texts from Susa (MDP 6, 213 / P008013; MDP 26S, 5008 / P009255) and one from Tal-i Malyan (Kadmos 24, 7 / P009466). The trigram appears in differing contexts, compare Kadmos 24,7 (above) and MDP 6, 213 (below):



The Susa sequence ends in a count of “90” M376, an object sign for which the meaning “cattle” has been considered. Preceding the trigram, which itself may be a personal name, differing signs might (by hypothesis) further identify the person’s group affiliations or status in each of the texts. Notably, if alphabetic subscript variants are merged, two further examples of this trigram can be identified at Susa. However, even with variants merged across the signlist, no other trigrams appear at both Susa and another site. A separate trigram, M388 M218 M219 occurs at both Tal-i Malyan and Tepe Yahya, even with variants kept separate (Kadmos 24,7 and Tepe Yahya 11 / P009535). This means that the same text from Tal-i Malyan shares one trigram with a text from Susa and another with a text from Tepe Yahya. Without this one text there would be no examples of any trigrams shared across sites.

3.3 Topic Modeling

To understand sign use in such a complex accounting system it is useful to consider text genre and the break-down of distinct text types across the corpus. How much overlap is there between different texts in terms of sign use? Are there significant correlations between certain groups of object signs that may help us to propose approximate meanings for some especially poorly understood signs? Past work on proto-Elamite has identified a few account types, some more and some less sharply defined. Two of the most clearly defined groups are worker rosters and work group texts (variously organized) with or without grain distributions (Englund & Damerow 1989; Dahl *et al.* 2018: 21-24) and animal husbandry texts (Dahl 2005). Text groups have also been identified by other distinguishing features, without their content being understood, such as the presence of similar first entries and subscripts (Dahl 2012: 6). Beyond these observations, successful text group study thus far has been limited,[13] certainly because of the complexity of the data. Born *et al.* (2019) therefore turned to machine learning to attempt a picture of text genre across the corpus.

Latent Dirichlet Allocation (LDA) is a tool known for example in the field of Natural Language Processing. Our assumptions in using this model are that every text is a mixture of one or more "topics", and every topic is characterized by a set of representative signs. The hypothesis is that these topics might approximately align with different administrative genres or accounting contexts. The model learns topics using a Bayesian probability model, probabilistically assigning tablets to topics based on the signs that occur in the tablets. It learns based upon how often signs co-occur with one another, so signs which commonly occur together tend to occur in the same topic. It computes two measures: the degree to which each topic is present in each text, and the degree to which each sign is present in each topic. Multiplying these scores together tells you approximately how common each sign is in each text.

The results were relatively interpretable. The topics are identified by number, 1-10.[14] The location of the circles indicates their similarity to other topics, and the size indicates the number of texts included. Reasonable numbers of tablets assigned to a given topic mapped onto identifiable accounting genres presented in figure 4. Although, looking through the topics also raises questions. Instead of one topic for animal husbandry, there appear to be two distinct topics with some overlap (4 and 7). Proposed syllabic signs are most strongly represented in the large topic 10, although the group includes many apparently very different texts. Topic 2 is most represented by the grain distribution sign M288 which is however the most common sign across the entire writing system; to a lesser extent, topic 2 is defined by the "field" sign M391, the "yoke" sign M54, and some signs that typically appear in headers M157, as well as the visually similar M325 and M195. Topic 8 is strongly defined by the particular header sequence M157 M195+M57, raising the possibility that

[13] Meriggi's (1971 / 1974b) categorization of texts by products such as "timber" sign M32; see 1971: 88-99) are speculative as they are based entirely on graphical form. However, M32 does appear as representative sign of topic 1, along with M56 (a plow) and M376.

[14] Future work could try setting the number of topics in the corpus higher: ten topics were initially selected after comparing the results of a few different numbers.

distinct activities of a particular "household" or economic unit are represented within this group. Interestingly, three of the most representative signs for topic 9 are M387, M387$_{a}$ and M387$_{ef}$, although the model was not aware of the visual relationship between those signs and their labels (all alphabetic subscript variants were treated as distinct). Sign M36, associated with "rationing", and vessel signs M263 and M263$_{a}$ were also among the representative signs of topic 9, and bi-sexagesimal numeric systems similarly posited to be associated with rationing activity are well-represented (Born *et al.* 2019: 129). Perhaps notably topic 9 is close to topic 3, which is strongly defined by M297$_{b}$ and M297, the latter which Friberg (1978) suggests might stand for a "keg" of beer (or according to Meriggi, bread).

This type of work may be most useful for hypothesis generation. Overall, topic modeling has contributed to building confidence that this tool of Natural Language Processing can successfully reveal meaningful patterns in the proto-Elamite dataset in its current state of transliteration.

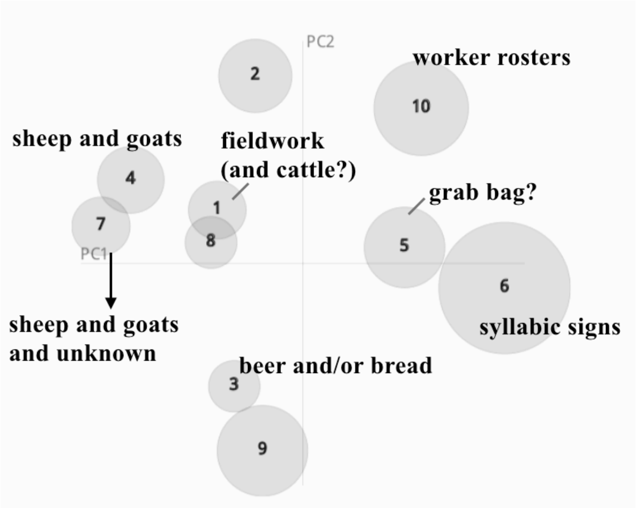


FIGURE 4: THE PROTO-ELAMITE CORPUS DIVIDED INTO 10 "TOPICS".

3.4 Sign Clustering

"Clustering" refers to a set of unsupervised machine learning techniques which attempt to group items that have similar features to one another. Born *et al.* (2019) measured similarity between two signs by counting how often those signs occur in similar contexts. Context here means the signs before and after a given sign within a line of text (1-5 signs), as well as in some methods accounting for more distant context. Because of the rarity of signs, multiple methods were compared in order to improve confidence in results based on a paucity of information, especially since in an undeciphered script native speakers cannot be asked to explain the meaning of a given sign cluster.

The first method is called the neighbor clustering and is used to group together signs based on their immediate neighbors. For every sign in the sign-list, it counts how many times each other sign occurs within 5 characters to the left or the right (often called a "window"). Signs were grouped according to how similar the resulting counts are. The second method uses a hidden Markov model (HMM), which is defined by two probabilities: the first defines how likely it is to move from one state to another, and the second defines how likely it is to generate a sign from a state. By identifying the most likely state for each input sequence of signs we can produce a cluster of signs that depends on contextual information. A third method called Brown clustering (Brown *et al.* 1992) groups together signs using a mathematical measure called mutual information which identifies a pair of signs and measures whether the signs are more likely to be seen together or whether they are more likely to be seen with other signs. It produces a hierarchical clustering where each sign starts off in its own cluster and then signs are merged depending on whether the loss in global mutual information is minimized.

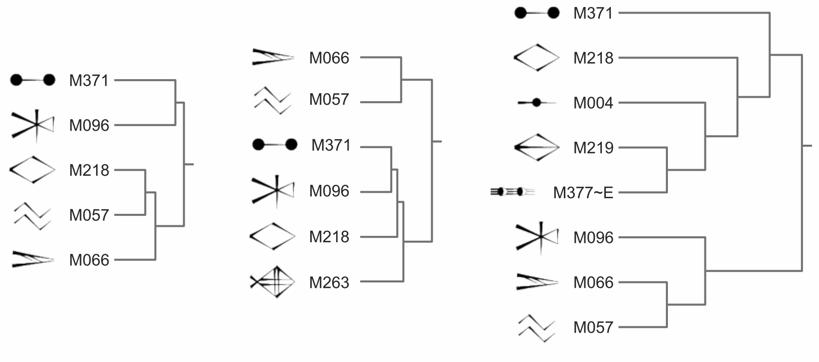
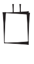
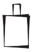


FIGURE 5: EXAMPLE OF STABLE CLUSTERS ACROSS THREE METHODS. NEIGHBOUR-BASED (LEFT) HMM (CENTER) AND BROWN (RIGHT). REPRODUCED FROM BORN *et al.* 2019 FIGURE 2.

Figure 5 presents some of the results in the form of dendrograms, where signs connected by short branches are more similar to one another than signs connected by taller branches. Each figure shows a "stable" cluster, that is one which recurs across all three methods. Some of the stable clusters turn out to group the most commonly proposed syllabic signs.

Clustering can contribute to making the case for merging of certain proto-Elamite signs with and without alphabetic subscripts (or indeed, visually similar signs with entirely different numbers), a major task in the decipherment process. For example, all three clustering methods closely relate:

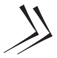
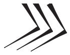
M157  and M157$_a$ 

does indicate their uses are very similar. These types of associations confirm that the clustering methods are picking up on meaningful contextual information, since alphabetic subscripts are treated as entirely distinct labels by the algorithms, and these clustering methods do not use the image of the sign. This method has not yet been extensively used to suggest signlist revisions. Other methods for clustering signs are discussed in section 4.1–4.2.

3.5 Limitations and next steps

The above methods have relied on transliteration of the corpus provided on the CDLI using the working signlist of J.L. Dahl and in large part transliterated by Dahl himself. The interpretability of results described in 3.3–3.5 show that the labeling system and transliterations are sufficient for representing some underlying patterns in the proto-Elamite system, and can support computational methods to help understand the writing system better. They also confirm that the underlying system itself contains sufficient data and sufficient inherent regularities to be amenable to these methods of analysis: this may not be assumed for a corpus of early inscriptions. However, both the signlist and the transliterations remain a work in progress; due to the level of minor variation in sign rendering by scribes and to the often-damaged state of tablets, it is easy to make counter-arguments for many of the sign names and transliteration choices. Some transliterations on the CDLI were prepared based upon poor line-art, and more recent CDLI images – such as those added by work in the collection of the National Museum of Iran – have revealed the need for transliteration updates. This means that, assuming not all labels are correct (as is to be expected for a large corpus in an undeciphered script), some possible relational information between signs is lost on the algorithms. The fact that Dahl's list remains agnostic about the relationship between all signs with letter subscripts is fortunate from a computational perspective, as it is possible to automate the merging of signs in a transliterated corpus, while the separation of one label into distinct sign labels (as would likely be needed if working from Meriggi's 1974a sign labels) requires manual re-transliteration of the corpus, or at the very least, supervision of any machine learning-based revision choices.

In summary, the current signlist makes decisions about labeling certain similar-looking signs as distinct, for example:

M43  M44 

These decisions are then fed to the computational models via the transliterated corpus. Moreover, the existing sign names bias our analyses towards replicating judgments from prior work. It would be more promising if a model could recover these earlier intuitions unaware of the existing labels.

4. IMAGE LANGUAGE MODELING

For the reasons given at the end of the previous section, we therefore turn to methods which can make more decisions of their own regarding sign relationships. We achieve this by developing models which incorporate the same visual cues used by annotators to label signs, rather than (or in addition to) incorporating the labels themselves. These models allow for more independent replication of existing results, and facilitate fact-checking our own results by showing that they reliably recur across models trained on different modalities of the data.

Ideally, this work would proceed by analyzing photographs of the original artifacts or reliable line art. A variety of issues preclude this at present: automatically extracting individual glyphs from images of non-inked 3D impressions on tablets is technically challenging (Bogacz & Mara 2022; Dencker *et al.* 2020), while extracting them manually is very laborious; some tablets have not been drawn or photographed; some are damaged and no longer show the full text known from earlier hand copies. For these reasons, the following discussion is based off an analysis of the digital sign image files drawn by J.L. Dahl. These depict the idealized form of each sign in the working sign list. As a result they lose some detail compared to an actual image of a sign on a tablet, but they nonetheless capture the most important visual features, which are otherwise obscured in computer analysis that uses only the transliterated sign names.

4.1 Dense Embeddings

Modern approaches to machine learning make extensive use of embeddings, which are lists of real numbers (tens, hundreds or thousands of them) representing discrete symbols. Intuitively, suppose one wishes to represent the English word “bank”. One may imagine counting how many times “bank” occurs next to every other word in some corpus: pairings like “river bank” or “investment bank” will be common, others like “slimy bank” will be rare, and yet others like “llama bank” may be entirely unattested. Consider a representation where all the co-occurrence counts for the word “bank” are used to denote this word. This representation can be analyzed mathematically as a vector. In practice, additional mathematical operations are performed to these counts before they are used, both to reduce the length of the lists produced, and to derive other measures which may be more informative than raw co-occurrence.

A more sophisticated technique involves the use of a language model to produce word embeddings. Informally, given a prompt like “I met a traveler from an antique ____”, a language model “fills in the blank” by suggesting continuations like “store”, “market”, or “land”. Formally, it achieves this by applying a mathematical function to embeddings of the words in the prompt, in order to derive a probability distribution over every possible continuation. Given a large corpus of text, one can construct many such prompts where the next word is known. These can be used together with a technique from calculus known as stochastic gradient descent to construct embeddings which allow the model to make the correct predictions as often as possible.

Embeddings derived through this approaches capture useful information about the contexts in which words occur. By extension, an embedding is commonly taken to encode the “meaning” of the word it represents, following an oft-quoted remark by Firth (1957): “You shall know a word by the company it keeps.” Embedding techniques also generalize well across very different domains: for example, language models have been productively applied to genomic analysis Ji *et al.* (2021). These properties make them a promising tool for decipherment, as they can convert an unreadable script to a quantitative format for which a host of analytic techniques already exist. In practice, a language model’s ability to predict the following sign is rarely useful to our analytic goals: the embeddings are of sole interest, and the sign prediction task is simply a means of producing embeddings which capture useful contextual information. Born *et al.* (2021) describe a novel type of language model which can be applied to images rather than (or in addition to) words.[15] This model is applied to sequences of proto-Elamite signs and produces a sign embedding

[15] See Born *et al.* (2021) for complete technical details. In brief, the model input is a series of sign name-plus-image pairs. Sign names are embedded from a one-hot encoding using a standard dense embedding layer. The images are passed through a stack of convolutional layers, and the output from the topmost convolution is max-pooled, flattened, projected to a lower dimension by a fully connected layer, and concatenated to the corresponding word embedding. A standard LSTM language model is trained (to predict sign names only, not images) over the concatenated word-plus-image embeddings. Either the sign name or the image can be omitted by zeroing out the corresponding half of the embedding.

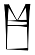
to represent each sign in the corpus, much as word embeddings can represent English words. These embeddings capture a holistic view of sign appearance and context; the underlying mathematical formalism allows this model to take broad contextual cues into account, so that a sign's embedding reflects not only its own context, but also the contexts surrounding visually similar signs. Crucially, this allows sharing of information across visually-similar variants, some of which, owing to their rarity, would be replaced by a generic UNK (or "unknown") label in order to support the functioning of text-only models, excepting when text-only models are run on a version of the signlist in which all signs with the same M-number are merged (expected to be an oversimplification). Thus image-aware models see a more accurate representation of the corpus than other models, and should therefore be expected to learn more representative embeddings. Born *et al.* (2021) also describe an image recognition model; this is another mathematical function which predicts the name of a proto-Elamite sign given an image of that sign.

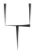
This model is trained using the same gradient descent technique as a language model. However, it only sees signs in isolation, and thus produces embeddings which depend exclusively on appearance, and have no relation to the contexts in which a sign is used. These models may be used in parallel to sign embeddings from a traditional embedding technique called GloVe (Pennington *et al.* 2014), which understands contextual signals but lacks awareness of sign shape.

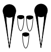
Embeddings derived from samples of modern writing typically correlate with human judgments about semantic similarity: words used in subjectively similar contexts receive numerically similar embeddings. This fact can be visualized using principal component analysis, which converts high-dimensional data (represented by many numeric quantities) into lower-dimensional data (represented by fewer numeric quantities): this converts each embedding to a point in two-dimensional (2d) space, in a way that preserves the most significant kinds of variation between embeddings. In general, points which are close together in the real embedding space remain close together in this simplified 2d representation. Nearby points in the resulting figure thus typically represent pairs of similar words. The following section explores the concept of dense embeddings applied to proto-Elamite signs through examination of the 2d embedding space.

4.2 Clustering using sign embeddings

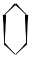
This work draws on sign embeddings computed using models that consider different types of information. Text-only language models (A) were given Dahl sign names and corpus context, but no access to images of the signs; An image recognition model (B) was given only sign images (no contextual information or Dahl labels); a text-and-image language model (C) had access to all three kinds of information; and finally, an image-aware language model (D) had access to the images and corpus context but not the Dahl labels. Figure 6 shows an excerpt from the plot produced by applying the visualization technique described in the previous section to the embeddings produced by the image language model (hereafter image LM) (D), in this case using labels that have merged Dahl's alphabetic subscript variants. Similar shapes are clustered together, but contextual features remain important, as these embeddings do not purely reflect appearance. This can be seen by comparison to an image recognition model (B) in which context is not used. In an image recognition model, all of the elongated hexagons (M304, M305, M296, M297, M298) form a single cluster; in the image LM, various lozenge shapes are set apart (only 296 and 298 appear in Figure 6). Examining these signs in context, we see that M304 and M305 occur predominantly before M388. The signs M296 and M298 never occur in this position, instead following M388. These contextual differences may have influenced the image LM to retain a separation between the signs despite their clear visual similarities. Figure 6 also shows that the "grain container" M288 and its variants cluster closely to the M36 series of signs, which contain numeric notations of capacity. Reassuringly, both sets of signs are known to indicate capacity measures. The location of the M370 complex graphemes including M370+M388+M370 at an outer extremity of the plot may represent the limited and distinct set of texts that use these signs (Kelley 2018: 400-409), particularly since some other graphically similar signs are not located anywhere nearby. These groupings help increase our confidence in previously identified sign correlation, and to establish that computational models developed for modern languages do in fact generalize to this ancient accounting system which might not directly encode significant amounts of linguistic information. The signs referenced above are:

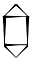
 M288

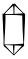
 M36

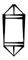
 M370+M388+M370

M388

 M296

 M298

 M304

 M305

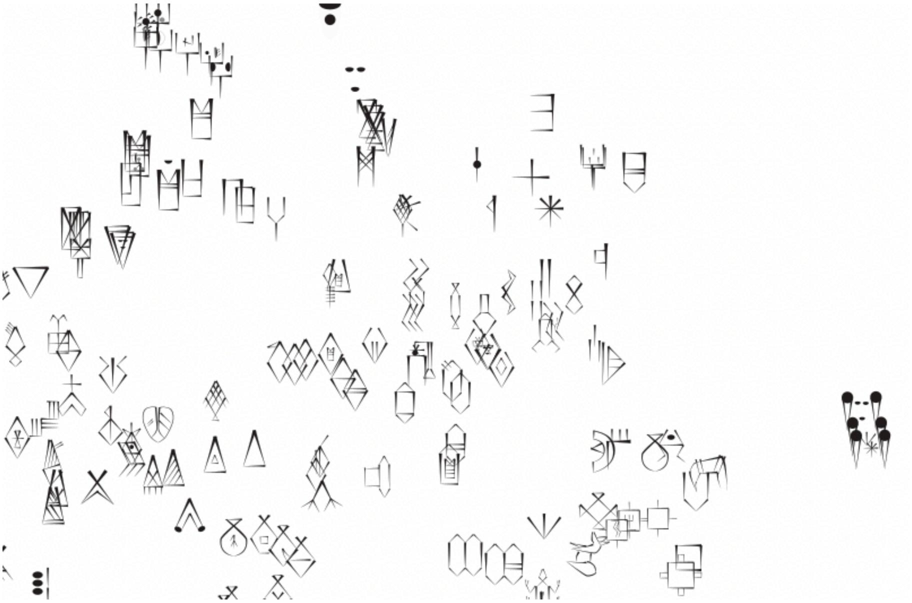


FIGURE 6: DETAIL FROM PCA PLOT OF IMAGE LM 64-DIMENSIONAL EMBEDDINGS WITH SIGN VARIANTS MERGED.

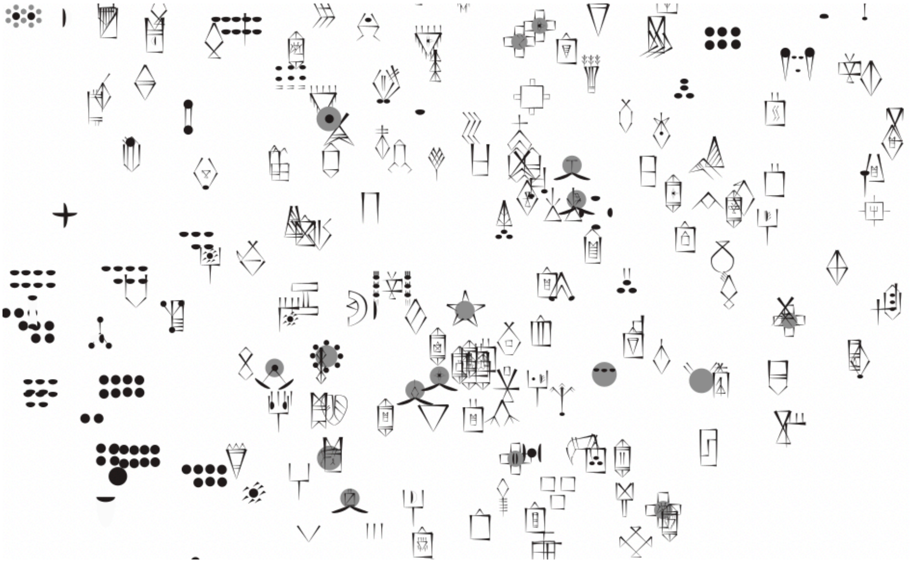


FIGURE 7: DETAIL FROM PCA PLOT OF GLOVE EMBEDDINGS.

Figure 7 shows the plot based on the image-agnostic GloVe model (A). One can observe multiple cases where this model has managed to cluster visually-similar signs despite having no awareness of their appearance. For example, distinct cruciform signs appear near one another in several places and several M305 complex graphemes appear in a tight cluster. Nearby signs should be expected to exhibit some degree of semantic or topical similarity: the fact that many neighboring signs also exhibit visual similarities confirms that there are correlations between sign shape and meaning.

Certain caveats must be kept in mind when interpreting these diagrams. First, proximity does not imply synonymy so much as it implies topical similarity, or use in similar contexts. In other words, proximity is not a smoking-gun for proposing proto-Elamite sign mergers but it can certainly inform on whether signs are more similar or more distinct in their use; because of this difficulty in confirming similarity vs. synonymy, this method may in fact be most effectively applied to make arguments about retaining distinctions between some otherwise visually-similar signs. A few other caveats must be considered. The rarer a word is, the less meaningful its position is expected to be. And if a sign is used in two distinct ways its representation reflects both of its uses. These are nonetheless useful visualizations for assessing embeddings when these caveats are properly taken into account.

Comparing embeddings obtained from different modalities demonstrates that both context and appearance can be used to group signs in meaningful ways. By assessing how signs group together, one may speculate about semantic and topical relations, and develop hypotheses which would have taken much longer to arise via traditional manual analysis.

4.3 Vector arithmetic

It is possible to interpret a sign embedding as a set of coordinates identifying some point in space, and thus to represent the embedding itself as an arrow (or vector) directed at this point in space. The relation between two signs can be represented by adding or subtracting their vectors together (by placing them end-to-end and drawing a new arrow from the beginning of the first to the end of the last). This is a standard analytic technique which has been shown to capture semantic and morphosyntactic information: in English, for example, the offset from a comparative adjective to its superlative may be roughly the same for many different adjectives. Born *et al.* (2021) exploited this kind of geometric measure to examine the relationship between the components of complex graphemes. In embeddings for English and other modern languages, it has been observed (Mikolov *et al.* 2013; Salehi *et al.* 2015; Cordeiro *et al.* 2016) that multi-word phrases receive embeddings which are close to the sum of the embeddings of the words in the phrase. This occurs more often in compositional phrases (whose meanings literally combine the meanings of their parts, as in black dog) than in non-compositional phrases (whose meanings may be more idiomatic, as in black swan). If complex graphemes are semantically compositional, they might therefore be hypothesized to receive embeddings which equal (approximately) the sum of the embeddings of their component signs. If their embeddings do not combine in this way, it may instead suggest that complex graphemes have more idiomatic meanings, which do not literally combine the meaning of their parts. Born *et al.* (2021) assess how close each complex grapheme is to the sum of its parts as a first step in assessing whether these signs may be compositional. This and other tasks were also used to compare the several different models described in section 4.2, to determine how the different types of information available to each model affect the representations which they learn.

According to the text-only model (A), some complex graphemes are found to approximately equal the sum of their parts. Crucially, significantly more complex graphemes are close to the sum of their parts than are close to the sum of two randomly-sampled signs (Fisher's exact test, $p = 0.04$).[16] Although there are few such signs (merely 13 in model A), this would seem to confirm that some complex graphemes are more closely related to their components than they are to other signs, and may therefore be speculated to have compositional meanings.

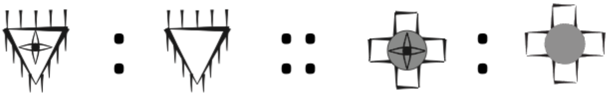
Models B, C, and D all have access to visual information, and all three produce a greater number of compositional representations than model A (totalling 38, 51, and 61 respectively). This is expected, as embeddings from these models reflect sign shape, and should therefore capture patterns of graphical composition. However, models which have access to sign context (C and D) produce even more compositional representations[17] than those without (B). This means that graphical features cannot account for all of the increase relative to model A: to some extent, it must be due to patterns of sign use, and should thus reflect semantic rather than purely graphical composition. Across all models, complex graphemes are significantly closer to the sum of their parts than to the sum of randomly-sampled sign pairs (by Fisher's exact test), but the difference is much more significant in C and D ($p \approx$ 1e-31) than in B ($p \approx$ 1e-21). Contextual information therefore allows for the recovery of more

[16] Concretely, the number of complex graphemes which are within the k nearest neighbors of the sum of their parts exceeds the number which are within the k nearest neighbors of the sum of two signs sampled uniformly at random. See Born et al. 2021 for discussion of the threshold k at which the difference is significant.

[17] As in model A, complex graphemes in models B, C and D are significantly more likely to equal the sum of their parts than the sum of a randomly-sampled sign pair (Fisher's exact test, $p < 0.05$).

confident relations between signs, further hinting that the compositionality recovered by these models is at least partly semantic. Those complex graphemes which appear compositional in context-aware models but not the image-only model are particularly likely to reflect instances of truly semantic composition.

Thus, by comparing these four approaches in aggregate, Born *et al.* (2021) find good evidence for some complex graphemes to have compositional meanings, beyond the number expected based on visual similarities alone. This by itself is a useful result, demonstrating that there are regular relationships between signs which have previously gone unobserved by specialists. The semantic content of complex graphemes has only received passing comment in specialist literature thus far, but this work now suggests the possibility of understanding these signs based on their relation to their constituent parts. Curiously, the image-aware model that did not have access to the Dahl sign names (D) produced more compositional embeddings than the model with access to sign names (C). If model D most accurately captures proto-Elamite semantics (as seems likely, given that it includes the fewest biases from modern annotators, and that its inputs most closely match the original tablets) this would indicate that compositionality can be recovered most effectively from a combination of context cues and images, while the Dahl sign labels to a certain extent add confusion. The same profile of results (D>C>B>A) holds when considering how many analogies these models are able to recover between complex graphemes. Model D produces the greatest number of analogical formulae such as:



The above example is one of 521 analogies that are offered by the image-aware model which appear by visual inspection to suggest a consistent contribution by the component parts of the CGs.[18] That same study also looked into geometric representations of embeddings using pairing consistency scores (PCS; Fournier *et al.* 2020). If an inner sign contributes a stable meaning to different outer signs, it would be expected that the vectors representing this transformation are parallel in the embedding space. PCS was calculated for vectors representing every possible complex grapheme component, and used to compute the overall average PCS of the inner and outer parts of a com-

[18] At k=15, which is estimated to be within the threshold of significance. The number in which such visually confirmable analogies were offered with K=1 is 69 (Born *et al.* 2021 table 5).

plex grapheme. Overall, there is a clear trend across all model types (A, B, C, D) that the inner part of a complex grapheme exhibits higher PCS than the outer part, or in other words that the impact of the inner part of a complex grapheme's representation is more consistent and predictable than that of the outer part. The difference is statistically significant even in text-only models (A; Mann-Whitney U test, $p < 0.05$), meaning that it must reflect some difference in use rather than simply a difference in appearance. It appears, therefore, that inner signs may contribute roughly the same meaning across many different complex graphemes. Note that this can hold even for complex graphemes which are not semantically compositional: although the inner part contributes a consistent meaning, it need not be the same meaning the sign would have when used by itself.

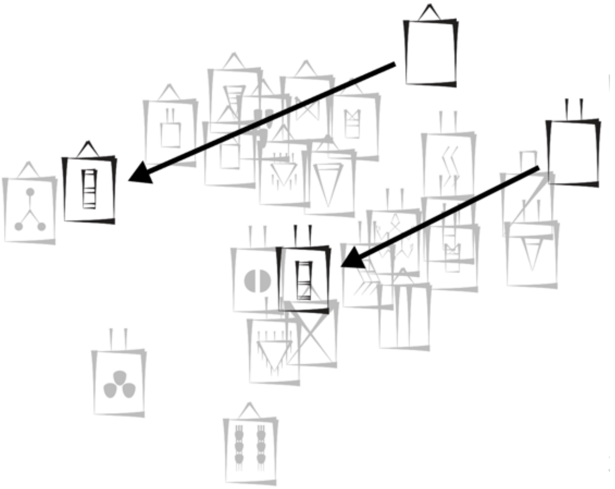


FIGURE 8: ILLUSTRATION OF PAIRING CONSISTENCY IN PROTO-ELAMITE, IDENTIFIED BY AN IMAGE LM (D).

In sum, text-only language models suggest that complex graphemes may be semantically compositional in proto-Elamite, as they position these signs closer to the sum of their parts than would be expected by chance. Image-aware language models offer yet stronger evidence, by showing that contextual information increases the number of compositional signs beyond what could be recovered from appearance alone. The use of models which balance appearance and context has thus permitted deeper and more meaningful analysis than would have been possible by considering either dimension in isolation.

If it is correct that many complex graphemes are semantically compositional, as these models have begun to indicate, the interesting consequence arises that sign names can function as "distractors", reducing the degree of compositionality recovered by the name-aware model C relative to the name-agnostic D. This suggests that some refinement of the sign-list may be warranted in order to better capture the true underlying set of glyphs in the corpus.

5. CREATING AND ASSESSING SIGNLISTS

The image-aware approaches in the previous section suggest directions for the interpretation of individual signs. However, the assumption that sign-names used by these methods are the ground truth is unwarranted. There are also other challenges: working with the existing token names in the corpus using a multimodal image-based language model is not easy. Some of the low frequency sign-names do not yet have corresponding graphical representations. Hapax signs or low-frequency signs frustrate the statistical modeling. The context is fragmentary and there are broken material objects. Despite these problems, coding distinctions between tokens is a fruitful area of research, potentially providing a more refined signlist using the following steps:

1. Create a signlist from scratch using the sign images themselves and their local context. This automated signlist is then compared with the existing working signlist from the CDLI to validate that an automated method can recover the bulk of the decisions already present in the working signlist.
2. Revise and improve the existing signlist using the validated automated methods for signlist creation.
3. Compare the automated signlist with an extremely simplified signlist which uses only graphical similarity without any contextual information, to show that the automated signlist does in fact balance the use of graphical similarity and contextual information.

5.1 Automated signlist construction

Automated signlist construction is achieved by combining a language model with a novel sign labeling component, which predicts which sign each glyph in the corpus is likely to represent. Rather than predicting the next true sign name based on the preceding signs, the language model is then trained to predict the name assigned by the labeling component, given the names assigned to the preceding signs by this component.[19] The labeling component may rela-

[19] A summary of technical details, including the model architecture and training objective, is available at https://github.com/MrLogarithm/inscribe-model-details. In brief, the sign labeling

bel individual tokens into entirely different groupings; this sometimes includes merging together signs which were distinct in the original transliterations, or proposing new signs by splitting existing ones into finer-grained subtypes. Constraints compel the labeling component to merge any pair of signs which look visually similar; additional constraints discourage mergers between signs which look dissimilar, and between signs which occur in very distinct contexts. The language model is used to detect signs in distinct contexts, as they make the distribution of signs harder to predict and cause a concomitant loss in language modeling accuracy. The model must find an assignment of names to glyphs which satisfies all of these (sometimes conflicting) constraints to the greatest extent possible. The signlist construction model recovers a substantial degree of the distinctions encoded in the working signlist. This result becomes apparent by measuring the homogeneity and completeness of the constructed list against the working sign names. Homogeneity is a measure of how similar two sets of labels are to one another, where low homogeneity implies that a single label from one labeling subsumes multiple distinct labels from the other. Completeness measures how many labels from one set are subsumed by the other: low completeness means that labels from one labeling end up divided across multiple labels in the other labeling. The auto-generated signlist when compared with the working sign-list scored highly in both measures. For example, it has the option to relabel an individual instance of a sign (that is, one token) without relabeling the other instances of that sign – yet it almost never does so. Rather it applies identical changes to (nearly) all tokens of a particular sign as already grouped by Dahl. This can be seen from the completeness score, 0.99 on a scale from 0 to 1, where the closer the score is to 1, the more the model maintains Dahl's signs as coherent groups. The results suggest that the automated signlist construction method is recovering some of the decision-making process behind the working signlist.

This is an interim step in overall analysis. The auto-generated signlist is only usable as a point of comparison, and does not reflect a replacement or emendation of the working signlist. Here we offer only a first step in the production of more accurate methods for signlist analysis. We do not expect this model to deal well with low-frequency signs (for which contextual clues are limited) or to capture the most subtle distinctions between signs. Partially for

component is a dense embedding layer with softmax activation. This embeds a token id (unique for each token in the corpus) into a probability distribution over s classes, each representing a distinct "sign name". These distributions (which approximately simplify to one-hot vectors when the labeling component's predictions are confident) are input to a standard LSTM language model. This model is trained to predict the class distribution produced by the labeling component for the following token.

these reasons the model was capped at producing 300 total signs. There are three reasons for choosing this number of expected signs: 1) on the hypothesis that some sign mergers within the Dahl list are appropriate, we want the model to group Dahl signs together and not just trivially recreate the working signlist; 2) the number of signs with variants collapsed is roughly around 300 signs; and 3) in preliminary tests, the model tended not to use a large number of signs even when many were available: no configuration produced more than around 450 distinct signs, outside of trivial cases where the model made no token merges at all. Despite overall success when comparing the auto-generated signlist and the working signlist there is not perfect agreement, and manual analysis may step in to examine the differences between the lists–this remains for future work.

5.2 Revising the working signlist

An alternative application of the signlist construction model is to initialize it using the working sign names, and allow it to make adjustments to these labels. If, as seems likely, the existing sign names are substantially correct, this approach reduces the analytic burden on the model by providing it with a reasonable starting point. Where the model ultimately deviates from this starting point, it provides insights into the working signlist.

When applied in this manner, the model overwhelmingly prefers to merge existing signs together (reflected by a low homogeneity score relative to the working signlist, 0.75) rather than to propose any finer-grained distinctions (reflected by a high completeness score, 0.99). The 300-sign cap forces the model to perform some mergers; thus the mere presence of mergers is not by itself proof that the Dahl list is over-precise. More important is the fact that the model produces even fewer signs than required by the cap, and the fact that the model rarely splits Dahl's signs apart. As will be seen in the next section, it is possible for the model to split signs apart while still reducing the overall number of signs, for example by merging some instances of a sign A into a sign B, and the remainder into another sign C. This does rarely ever happens in the current setting,[20] suggesting that any shades of meaning within a given Dahl sign are less salient than some distinctions between Dahl signs. Encouragingly, many of the candidate mergers occur between signs which share evident visual and contextual similarities, such as M304 and M305.

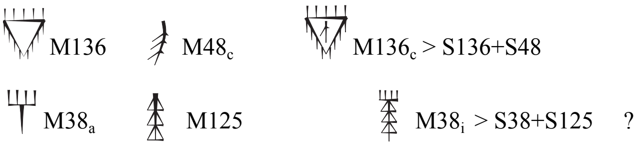
[20] There are 14 signs which the model does split across multiple classes, but even in 10 of those cases the model clusters nearly all instances of the sign together and just moves one or two tokens to another group. M5, M370$_b$+M388, M387$_c$, and M72 are the only signs which have a substantial number of tokens (>3) in multiple clusters.

5.3 Limits of graphical simplification

To provide another point of reference for understanding the role of visual information in the sign repertoire, a modified signlist was produced with the aim to test an extreme "simplification" hypothesis, namely that most of the apparent graphical variations between signs can be outright ignored. This hypothesis is expected to overly simplify the data, and thus an automated revision of this list is expected to reverse many of the changes which went into its construction, and recover a list closer to that of Dahl. This would rule out what we suspect to be an excessively simplified approach to decipherment.

If the graphical variations between signs are in fact mostly meaningless, then signs from Dahl's list could be aggressively merged based on visual resemblance. We thus attempted to merge as many signs as possible according to reasonable graphical criteria, such as discounting single versus doubled lines, and "outlined" versus "stick" renderings (compare M201 and M195). Resemblances were established from digital sign images, though the original tablets were also referenced to confirm graphical forms where possible. Contextual clues were only superficially considered to rule out the most blatantly incorrect merges. The process aimed to eliminate not only signs marked with alphabetic subscripts (which could be removed automatically) but to identify and merge graphically similar signs whose labels do not reflect their similarity in a way that permits automated merging. The intensity of the resulting merges can be illustrated by referring back to Figure 2. In the simplified signlist, all signs between 74_a and 91 were re-labeled S81.

This new list also introduced 82 new compound graphemes, by analyzing existing signs as compounds of a smaller set of "basic" signs. Some of these are made with confidence while others are more speculative:



The resulting list contains 558 signs, including the many newly introduced complex graphemes. When the components of complex graphemes are split apart so that compounds are not counted as distinct signs, and when removing a few dozen signs that were marked with a "?" to indicate that the label was highly speculative based on visual examination of the evidence, the reduced signlist reaches 243 signs. This can be compared to Dahl's 1623 non-numeric signs, 568 with alphabetical subscripts merged, or 356 with compounds split apart and variants merged. Thus the "basic" number of signs was reduced by over 100 from 356 to 243. The signlist construction model is applied to this list in the same way it was applied to revise the working signlist. The output reveals a tendency for the model to split apart signs which had been manually merged (reflected by a low completeness score), restoring distinctions that had been present in the original working signlist. This behaviour was not observed in either of the preceding settings, and it supports the hypothesis that methodologies based on purely visual resemblances are liable to obscure informative contextual cues and produce an excessively simplistic and less accurate representation of the corpus.

5.4 Summary

The relative conformity of both the automated list and the Dahl list, by comparison to the simplified list, indicates that those lists are likely capturing important contextual information. The results of 5.1–5.2, in combination with those presented in section 4, suggest that the current proto-Elamite signlist encodes some distinctions which should be collapsed together; in some cases, the distinctions may simply be more fine-grained than necessary to support a broad understanding of the script and the texts. These distinctions may still be relevant to palaeographic and other specialist studies of the script, but the evidence suggests they are not relevant to understanding the most basic relations between signs as an aid to decipherment. Care must still be taken to incorporate contextual features when assessing which signs in particular might be merged together, as the evidence in 5.3 does not support a naive reduction based on general appearance.

The model discussed in this section bears a conceptual similarity to recent work by Corazza *et al.* (2022), who consider the task of clustering signs in Cypro-Minoan. However, their technique considers a fixed window of two signs from the surrounding context, whereas our language-modeling approach has the potential to incorporate arbitrarily many of the preceding signs. Proto-Elamite texts are slightly longer, on average, than those written in Cypro-Minoan, meaning there is more reason to consider a broad context window when analyzing proto-Elamite text. Cypro-Minoan also uses a "divider" token to separate syllable sequences from one another. This allows Corazza *et al.* to guarantee that their context windows never span multiple distinct syllable sequences: their windows only contain information which is directly relevant to the interpretation of the current sign. Proto-Elamite does not use such a divider, meaning that even a narrow context window may cross word or phrase boundaries (and, by extension, may contain information which is not directly relevant to the interpretation of the current sign). For this reason, constrained context windows are likely to confer relatively less benefit to analyses of proto-Elamite than Cypro-Minoan.

6. CONCLUSIONS

Computational methods devised for modern languages are demonstrably applicable to the study of ancient writing, including proto-Elamite, a relatively modest-sized accounting system with potentially weak representation of spoken language. This and other recent work demonstrate that such methods can replicate and expand on results from prior manual work on proto-Elamite. These methods are possible because of the current digital state of the corpus as presented by the Cuneiform Digital Library Initiative and the sign files of J.L. Dahl, which have proven sufficiently accurate representations of the underlying structures of proto-Elamite to support convincing analysis.

Proto-Elamite provides a relatively unusual application for many of the methods described above precisely because it is not yet deciphered, and the explanatory power of these and similar methods in the context of a decipherment challenge is just beginning to be explored. Furthermore, to avoid circularity in reasoning, since the sign names were developed based on specialist intuitions about sign use, it is more compelling if results can be recovered without using the sign names. This had led to the work presented in sections 4 and 5. One of the most important outcomes is the discovery of and reliance on image-aware language models in combination with traditional text-based language models. Analysis derived from these combined methods in an undeciphered setting can increases analytical capabilities.

This discussion highlights the possibility of devising different levels of analysis within the proto-Elamite signlist. It is expected that there is a trade-off between a full and accurate list of all variation in graphemes, and a list that is sufficiently simplified to represent the important “rules” of the writing system and supports n-gram analysis through a less sparse dataset. It is possible that minor graphical variation in proto-Elamite often correlates to minor ideographic differences, such as variations in products or group or personal identities. This phenomenon may exist alongside graphical variation which is not (directly) semantically meaningful. Possibilities for the latter include chronological variation, differently trained scribal hands such as by geographic location, or even scribal freedom to add flourish to a sign out of motivation other than semantic signalling. The comparison of signlists in section 5 indicates that a mix of contextual and graphical features of the texts can guide merges and token re-assignments. Results also indicated that Dahl’s existing list is more suited for analysis of texts at present than is a smaller list which assumed graphical similarity can be more heavily relied upon to merge signs. Finding the balance between reducing and retaining distinct signs in the signlist remains one of the most interesting questions related to this writing system and to the nature and diversity of early writing inventions more broadly.

REFERENCES

Abdi, K. & N. Miller (2003). From écriture to civilization: changing paradigms of Proto-Elamite archaeology. In *Yeki bud yeki nabud: Essays on the Archaeology of Iran in Honor of William M. Sumner*, 140–51. Los Angeles: Cotsen Institute of Archaeology, UCLA.

Aurora, F. (2015). DĀMOS (Database of Mycenaean at Oslo). Annotating a fragmentarily attested language. In Pedro A. Fuertes-Olivera et al. (eds.), Current Work in Corpus Linguistics: Working with Traditionally-conceived Corpora and Beyond. Selected Papers from the 7th International Conference on Corpus Linguistics (CILC2015). *Procedia-Social and Behavioral Sciences*, 198. 21–31. `https://doi.org/10.1016/j.sbspro.2015.07.415`.

Bennett Jr, E.L. (1947). *The Minoan linear script from Pylos.* Ph.D. thesis. University of Cincinnati.

Bogacz, B. & H. Mara (2022). Digital Assyriology: Advances in Visual Cuneiform Analysis. *Journal on Computing and Cultural Heritage (JOCCH)*, 15(2). 38:1–38:22. `https://doi.org/10.1145/3491239`.

Born, L., K. Kelley, N. Kambhatla, C. Chen & A. Sarkar (2019). Sign clustering and topic extraction in proto-elamite. In *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. Minneapolis, USA: Association for Computational Linguistics, 122–132. `https://doi.org/10.18653/v1/W19-2516`.

Born, L., K. Kelley, M.W. Monroe & A. Sarkar (2021). Compositionality of Complex Graphemes in the Undeciphered Proto-Elamite Script using Image and Text Embedding Models. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*. 4136–4146.

Brown, P.F., V.J. Della Pietra, P.V. Desouza, J.C. Lai & R.L. Mercer (1992). Class-based *n*-gram models of natural language. *Computational linguistics*, 18(4). 467–480.

Corazza, M., F. Tamburini, M. Valério & S. Ferrara (2022). Unsupervised deep learning supports reclassification of Bronze age cypriot writing system. *PloS one*, 17(7). e0269544. `https://doi.org/10.1371/journal.pone.0269544`.

Cordeiro, S., C. Ramisch, M. Idiart & A. Villavicencio (2016). Predicting the compositionality of nominal compounds: Giving word embeddings a hard time. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin: Association for Computational Linguistics, 1986–1997.

Dahl, J.L. (2002). Proto-Elamite sign frequencies. *Cuneiform Digital Library Journal*, 2002(1).

Dahl, J.L. (2005). Animal Husbandry in Susa During the Proto-Elamite Period. *Studi Micenei ed Egeo-Anatolici*, 47. 81–134.

Dahl, J.L. (2009). Early Writing in Iran, a Reappraisal. *Iran*, 47(1). 23–31.

Dahl, J.L. (2012). The Marks of Early Writing. *Iran*, 50(1). 1–11.

Dahl, J.L. (2019). Tablettes et fragments protoélamites/proto-elamite tablets and fragments. *Textes Cunéiform Tomes XXXII Musée de Louvre*.

Dahl, J.L., L.F. Hawkins & K. Kelley (2018). Labor administration in proto-Elamite Iran. In A. Garcia-Ventura (ed.) *What's in a name? Terminology related to the Work Force and Job Categories in the Ancient Near East. (Alter Orient und Altes Testament Band 440)*. 15–44.

Damerow, P. (2006). The origins of writing as a problem of historical epistemology. *Cuneiform Digital Library Journal*. `http://cdli.ucla.edu/pubs/cdlj/2006/cdlj2006_001.html`.

Dencker, T., P. Klinkisch, S.M. Maul & B. Ommer (2020). Deep Learning of Cuneiform Sign Detection with Weak Supervision Using Transliteration Alignment. *PLOS ONE*, 15(12). e0243039. `https://doi.org/10.1371/journal.pone.0243039`.

Desset, F., K. Tabibzadeh, M. Kervran, G.P. Basello & G. Marchesi (2022). The Decipherment of Linear Elamite Writing. *Zeitschrift für Assyriologie und vorderasiatische Archäologie*, 112(1). 11–60.

Englund, R.K. (2004). The State of Decipherment of Proto-Elamite. In S.D. Houston *et al.* (eds.) *The first writing: Script invention as history and process*, 100–149. Cambridge: Cambridge University Press.

Englund, R.K. & P. Damerow (1989). *The Proto-Elamite Texts from Tepe Yahya*. Peabody Museum of Archaeology and Ethnology, Harvard Univ.

Farmer, S., R. Sproat & M. Witzel (2004). The Collapse of the Indus-Script Thesis: The Myth of a Literate Harappan Civilization. *Electronic Journal of Vedic Studies*, 11(2). 19–57.

Firth, J.R. (1957). A Synopsis of Linguistic Theory. In *Studies In Linguistic Analysis*, 1–32. Oxford: Basil Blackwell.

Fournier, L., E. Dupoux & E. Dunbar (2020). Analogies minus analogy test: measuring regularities in word embeddings. In *Proceedings of the 24th Conference on Computational Natural Language Learning*. Online: Association for Computational Linguistics., 365–375.

Friberg, J. (1979). The Third Millenium Roots of Babylonian Mathematics: A Method for the Decipherment, through Mathematical and Metrological Analysis, of Proto-Sumerian and Proto-Elamite Semi-Pictographic Inscriptions. *Göteborg: Dept. of Mathematics, Chalmers University of Technology.*

Green, M.W. & H.J. Nissen (1987). *Zeichenliste der archaischen Texte aus Uruk*, volume 2. Mann.

Hawkins, L.F. (2015). A New Edition of the Proto-Elamite Text MDP 17, 112. *Cuneiform Digital Library Journal*, 1.

Helwing, B. (2013). Some thoughts on the mode of culture change in the fourth millennium BC Iranian highlands. In C.A. Petrie (ed.) *Ancient Iran and its neighbours: local developments and long-range interactions in the fourth millennium BC. (British Institute of Persian Studies Archaeological Monographs Series III)*, 235–269. Oxford: Oxbow Books.

Ji, Y., Z. Zhou, H. Liu & R.V. Davuluri (2021). DNABERT: pre-trained Bidirectional Encoder Representations from Transformers model for DNA-language in genome. *Bioinformatics*, 37(15). 2112–2120. `https://doi.org/10.1093/bioinformatics/btab083`.

Judson, A.P. (2020). *The undeciphered signs of Linear B: Interpretation and scribal practices (Cambridge Classical Studies).* Cambridge: Cambridge University Press.

Kelley, K. (2018). *Gender, age, and labour organization in the earliest texts from Mesopotamia and Iran (c. 3300-2900 BC).* Ph.D. thesis, University of Oxford. `https://ora.ox.ac.uk/objects/uuid:afa3362e-1182-43aa-a2b9-d675bd8c585a`.

Kelley, K. (2021). More Than a Woman: On Proto-cuneiform SAL and the Archaic 'Tribute List'. In A. Bramanti, N.L. Kraus & P. Notizia (eds.) *Current Research in Early Mesopotamian Studies*, 9–44. Münster: Zaphon.

de Mecquenem, R. (1949). *Épigraphie proto-élamite, Mémoires de la Mission Archéologique en Iran.* Paris: Presses universitaires de France.

Meriggi, P. (1971/1974a-b). La scrittura proto-elamica/1 la scrittura e il contenuto dei testi. parte i^{a} la scrittura e il contenuto dei testi / parte ii^{a} catalogo dei segni / parte iii^{a} testi. *La scrittura proto-elamica*.

Mikolov, T., I. Sutskever, K. Chen, G. Corrado & J. Dean (2013). Distributed Representations of Words and Phrases and Their Compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*. Red Hook, NY, USA: Curran Associates Inc., 3111–3119.

Pennington, J., R. Socher & C.D. Manning (2014). GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. Doha, Qatar: Association for Computational Linguistics, 1532–1543.

Salehi, B., P. Cook & T. Baldwin (2015). A word embedding approach to predicting the compositionality of multiword expressions. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Denver, Colorado: Association for Computational Linguistics, 977–983.

Scheil, V. (1900). *Textes élamites-sémitiques: 1.[-5] série (Mémoires de La Délégation En Perse 2 = MDP 2)*, volume 2. Paris: E. Leroux.

Kathryn Kelley
University of Toronto, Department of Near & Middle Eastern Civilizations
4 Bancroft Avenue, Toronto
Canada
e-mail: kathryn.kelley@utoronto.ca
https://orcid.org/0000-0003-2886-1606

Logan Born
Simon Fraser University, School of Computing Science
8888 University Drive, Burnaby, Vancouver
Canada
e-mail: logan.born@sfu.ca

M. Willis Monroe
University of British Columbia, Department of Philosophy
Buchanan Tower 628, Vancouver
Canada
e-mail: willis.monroe@ubc.ca
https://orcid.org/0000-0002-9126-2991

Anoop Sarkar
Simon Fraser University, School of Computing Science
8888 University Drive, Burnaby, Vancouver
Canada
e-mail: anoop@sfu.ca
https://orcid.org/0000-0002-4795-9361