

# Sequence Models for Document Structure Identification in an Undeciphered Script

Logan Born<sup>1</sup>

loborn@sfu.ca

M. Willis Monroe<sup>2</sup>

willis.monroe@ubc.ca

Kathryn Kelley<sup>3</sup>

kathrynerin.kelley@unibo.it

Anoop Sarkar<sup>1</sup>

anoop@cs.sfu.ca

<sup>1</sup>Simon Fraser University  
School of Computing Science

<sup>2</sup>University of British Columbia  
Department of Philosophy

<sup>3</sup>Università di Bologna  
Dipartimento di Filologia Classica e Italianistica

## Abstract


This work describes the first thorough analysis of “header” signs in proto-Elamite, an undeciphered script from 3100-2900 BCE. Headers are a category of signs which have been provisionally identified through painstaking manual analysis of this script by domain experts. We use unsupervised neural and statistical sequence modeling techniques to provide new and independent evidence for the existence of headers, without supervision from domain experts. Having affirmed the existence of headers as a legitimate structural feature, we next arrive at a richer understanding of their possible meaning and purpose by (i) examining which features predict their presence; (ii) identifying correlations between these features and other document properties; and (iii) examining cases where these features predict the presence of a header in texts where domain experts do not expect one (or *vice versa*). We provide more concrete processes for labeling headers in this corpus and a clearer justification for existing intuitions about document structure in proto-Elamite.

## 1 Introduction

Proto-Elamite (PE) is a largely undeciphered script of the Early Bronze Age, inscribed on clay tablets unearthed in Iran. PE shares certain features, most notably its number systems, with another ancient script called proto-cuneiform: these similarities have allowed for a partial decipherment which informs current understandings of the texts as administrative accounts recording amounts of various goods and personnel. Figure 1 shows a typical text with annotations explaining how it is divided into columns and entries. Dahl (2019) gives a thorough survey of PE from an archaeological perspective; Born et al. (2019) introduce the corpus to technical audiences and describe initial computer-assisted exploratory analyses.

Specialists have hypothesized that PE texts frequently begin with a “header”, that is, a sign

(or string of signs) which “qualifies all transactions recorded in a text” by specifying an institution or owner in charge of the associated account (Damerow and Englund, 1989, 14-16). This understanding of headers depends in part on the claim that they correspond to visually demarcated “colophons” in proto-cuneiform accounts (Englund 2004, 144; Damerow and Englund 1989, 15), which are however also largely undeciphered and therefore not certain to consistently convey ownership information.

Some (but not all) of the signs that occur at the beginning texts have been tentatively labeled as headers by domain specialists. This labeling is recorded using comments in the transliteration of the texts; no explicit list of header signs has been published. The clearest example of this putative category is the ubiquitous sign M157 , which occurs at the start of fully one-fifth of all PE accounts. Most header signs, including M157, may also appear elsewhere in a text with uncertain function.

In light of modern scholarship’s very partial understanding of the PE corpus, there does not seem to be proof beyond reasonable doubt that headers record ownership, much less that *all* headers do so. Moreover, headers have thus far been identified through manual analysis which has not been fully documented in any publication, and some of the experts who originally identified this category are no longer alive. Thus the criteria for identifying headers are opaque and the question of their existence is a matter of qualitative judgement.

In this work, we combine computer-aided analysis with domain expertise to undertake the first focused study of headers in PE. We use statistical and neural sequence models to show that headers are a genuine structural phenomenon of PE. We independently replicate manual annotations from past work with high accuracy, and our models also identify and allow us to correct a number of annotation mistakes. Based on our results, we argue

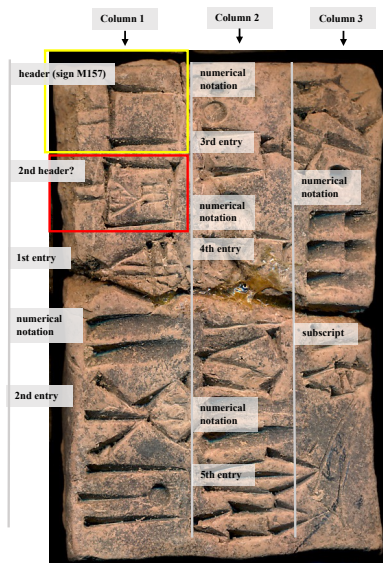


Figure 1: PE tablet *Scheil 1935* (MDP 26), no. 009 (P008697; *Scheil 1935*) with the "header" sign M157 highlighted in yellow. Clay, 6.6 x 4.2 x 1.7 cm. Reverse not shown. Image adapted from the Cuneiform Digital Library Initiative.

against the conventional understanding that headers nearly always span only one sign. In conjunction with this, we show that signs in the first and second positions of a text predict distinct information, suggesting they have distinct functions.

## 2 Data

The vast majority of PE texts have been made publicly available online by the *Cuneiform Digital Library Initiative* (CDLI<sup>1</sup>) in the form of (i) high-quality digital images, (ii) hand-drawn line-art, and (iii) ASCII transliterations representing experts' current understanding of each text. The transliterations employ a working signlist comprising numbered signs from M001 through M521 (*M* acknowledging the original standardisation of the signlist by scholar P. Meriggi). Tilde annotations (as in M001~a) represent possible variants of a sign; experts remain agnostic as to whether these variants are distinct characters. Texts are identified by "P" numbers assigned by the CDLI (Figure 1 caption). *Born et al. (2019)* publish a cleaned snapshot of the transliterated corpus together with a library of tools for interacting with the texts: this snapshot contains roughly 26k readable tokens across 1467 transliterated texts. The data for the present work is a version of this snapshot where transliteration

<sup>1</sup>[cdli.ucla.edu](http://cdli.ucla.edu)

errors (e.g. transliterations that do not accurately reflect a tablet's photograph) were fixed by domain experts.<sup>2</sup>

## 3 Methodology

The CDLI transliterations include rich annotations, including which signs (if any) are understood to comprise a text's header. We propose to train two unsupervised sequence models on the PE corpus and assess whether these models suggest any internal structure at the beginning of these texts. We also aim to evaluate whether and to what extent the features learned by these models can be used to recover the expert annotations, to establish whether these models are identifying the *same* structure posited by experts. We aim to arrive at a richer understanding of the meaning and purpose of headers by (i) examining which features are useful for predicting this category; (ii) finding correlations between these features and other document properties; and (iii) identifying why the models disagree with (or fail to recover) the human labeling if such disagreements occur. We hope to provide new and quantifiable evidence that headers are a real structural phenomenon in PE, and to concretely justify why any given may have or not have a header. Overall, we seek to *assess* and *understand* the human labeling rather than to indiscriminately replicate it.

### 3.1 Hidden Markov Model

Hidden Markov models (HMMs; *Cave and Neuwirth 1980*) have become a standard tool for unsupervised analysis of other undeciphered text corpora. We fit a 15-state HMM to our corpus; this number of states was chosen to slightly exceed the number of different sign categories which can be informally speculated to occur in PE (most saliently, headers, qualifiers, counted objects, syllables, owners, numerals, and subscripts). We train ten models from random initializations, using complete tablets as input sequences; we keep the model which assigns the highest likelihood to the corpus. For each tablet, we compute the optimal state sequence according to this model using Viterbi decoding. We hypothesize that, if headers exist, their existence will be reflected in the HMM by a state which only occurs at the very beginning of some texts. If such a state does not exist, it may mean that headers are not a salient structural feature of the corpus; if such

<sup>2</sup>Updated transliterations and annotated data are available at <https://github.com/sfu-natlang/pe-headers>.

a state exists, but is not associated with texts where human annotators believe there to be a header, it may imply that current understandings of headers somehow fail to reflect the true distribution of this structure.

### 3.2 Transformer

We also train an autoregressive Transformer (Vaswani et al., 2017) language model from a random initialization using the vanilla fairseq recipe.<sup>3</sup> Neural architectures such as the Transformer offer significantly greater inferential power than statistical models like the HMM, though the large amounts of data required for training can make them unsuitable for extremely low-resource archaeological data. For the present work, we are purely interested in using our models as analytic devices (i.e. feature extractors), and we neither require nor expect them to generalize. For this reason, we proceed with training a Transformer language model as a more powerful alternative to the HMM, with full knowledge that it will overfit to our low-resource corpus.

Under the hypothesis that headers convey information which is relevant to the interpretation of a tablet as a whole, we predict that the LM will attend to the beginning of a tablet on all or most time steps if that tablet has a header. In texts without a header, the beginning of the document will contain no such special information, and thus should not be expected to receive stronger attention than any other part of the text. Thus, if headers are a legitimate structural phenomenon, we should observe two classes of text which are differentiated by the average amount of attention paid to their initial signs.

Formally, let  $z_{i,j}$  denote the self-attention score for token  $t_i$  at time step  $j$ , and for a sequence of length  $L$  let  $n_i = L - i - 1$  denote the number of tokens following  $t_i$ . Then  $\tilde{z}_i = \frac{1}{n_i} \sum_{j>i} \frac{z_{i,j}}{\max_k z_{k,j}}$  is the average self-attention paid to  $t_i$  by the rest of the document. This is essentially the mean of the self-attention scores for  $t_i$  across all following time steps (which would be  $\frac{1}{n_i} \sum_{j>i} z_{i,j}$ ), except that we have normalized the scores at each time step so that the largest is always 1 (this controls for text length, as the true mean tends to zero as text length increases). For a given text and indices  $m$  and  $n$ , let  $\tilde{\mathbf{z}}_{m,n} = [\tilde{z}_m, \tilde{z}_{m+1}, \dots, \tilde{z}_{n-1}]$  denote the average attention paid to tokens  $t_m$  through  $t_{n-1}$

<sup>3</sup>[github.com/facebookresearch/fairseq](https://github.com/facebookresearch/fairseq)

The first numeral of a text gives an upper bound on the length of that text’s header, if it has one. Hereinafter, let  $n$  stand for the number of signs which precede the first numeral of a given text (each tablet thus has its own  $n$ ). We hypothesize that each text’s  $\tilde{\mathbf{z}}_{0,n}$  will capture information about whether that text has a header, and therefore (if headers are a real structural phenomenon) that a logistic regression over  $\tilde{\mathbf{z}}_{0,n}$  should be able to accurately predict which texts human experts have annotated as having a header. Later sections of the text should be less predictive; thus, as a baseline, a logistic regression over  $\tilde{\mathbf{z}}_{10,20}$  (or, equivalently, any other span of signs believed to lie outside the putative header) should *not* be able to predict the expert annotations.

### 3.3 Training

We train the HMM and Transformer LM on sequences of sign names, where each sequence spans a single document. We omit all annotations, such as those marking damaged signs: this reduces the vocabulary size and makes the distribution for most signs less sparse. We set aside 200 tablets (out of 1399 total) for the Transformer to use as a validation set for its language modeling task.

As we are interested in tablet headers, we only evaluate our models on texts where the beginning is substantially intact. If a text’s transliteration contains the comment "beginning broken", if there is a prime ' in the first line number of the transliteration, or if the first sign is X or [...], we omit that tablet from our analysis. After pruning we are left with 795 documents.

We construct the mean attention vectors  $\tilde{\mathbf{z}}_{0,n}$  and  $\tilde{\mathbf{z}}_{10,20}$  for each text in the pruned corpus (where  $n$  differs for each text, according to how many signs that text has before its first numeral). We zero-pad the  $\tilde{\mathbf{z}}_{0,n}$  vectors to the length of the longest, and train two logistic regressions to predict whether human experts annotated a text as having a header: the first is trained on the set of (padded) text-initial vectors and the second on the set of text-internal vectors. In both settings, the most accurate model is selected using 10-fold cross validation.

## 4 Experimental Results

### 4.1 Hidden Markov Model

Encouragingly, the Viterbi sequences from our HMM exhibit a heavily skewed state distribution at the beginning of tablets. In particular, 55% of

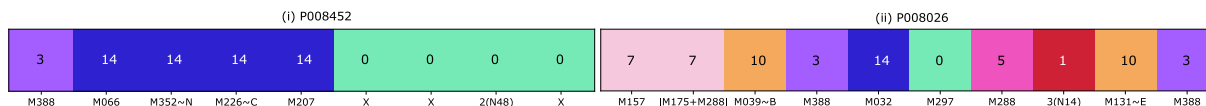


Figure 2: Illustration of state sequences learned by the HMM. The observed sequence of sign names is shown on the  $x$ -axis (truncated to at most 10 signs); the numbers in the cells report the states in the Viterbi sequence. Each color represents a distinct state. (i) HMM state 3 does not suggest the presence of a header, though one is present in the expert annotations; (ii) HMM state 7 suggests the presence of a header, which is present in the expert annotations.

all texts begin in state 7; a significant majority of these cases (76%, or 42% of texts overall) *only* exhibit state 7 on the very first sign. A mere 7 tablets (0.8% of the total) exhibit this state on or after the 4th sign. Thus state 7 is strongly localized to the beginning of tablets.

The fact that the model learns such a strongly localized state suggests that some documents *do* have a discernible internal structure and that the beginning of these texts is measurably distinct from what follows. This is fully consistent with the hypothesis that headers exist as a structural phenomenon within PE.

The contingency table in Table 1 allows us to assess whether HMM state 7 captures the same information as the headers identified by human annotators. We observe that state 7 recovers the human labels with high precision (0.93) but low recall (0.67), for an overall accuracy of 0.70. This could imply that the HMM has failed to recover some crucial feature that human annotators used to identify headers; that human annotators have proposed headers in some contexts where no header truly exists; or that the HMM states distinguish between finer categories than are encompassed by the specialist’s monolithic header annotation.

Initial HMM State	Expert Annotation		$\Sigma$
	Header	No Header	
State 7	410	30	440
Other	205	150	355
$\Sigma$	615	180	795

Table 1: Contingency table comparing the initial state of a tablet’s Viterbi sequence against the presence of a header annotation in the tablet metadata.

Examining the state sequences from some sample texts (Figure 2) helps in comparing these possibilities. In sequence (i), human annotators identified M388 as a header, whereas the HMM places this sign in state 3 rather than the putative “header” state 7. M388 is a very common sign in the body of

tablets, and has been identified as having a unique distribution in prior work (Kelley, 2018). The HMM clearly recognizes this, and learns a state (3) which is almost exclusively used for M388. Most instances of M388 are followed by so-called “syllabic” signs, which the model appears to identify using state 14 (as seen in both sequences of Figure 2). The M388 in sequence (i) is followed by syllabic signs and looks like other typical examples of this sign, making it unclear why a header was identified here by human annotators (especially given that other tablet-initial M388s are not labeled as headers in the expert annotations). This tablet also contains some unreadable signs (denoted by X), which appear to confound the HMM in most texts where they occur. The model typically predicts state 0 whenever it observes an X, and continues to predict state 0 for every subsequent sign, even when that sign is common and receives a more interpretable state in other contexts. We see this behaviour in sequence (i), where the model remains in state 0 even when seeing the intact numeral sign 2(N48). Thus, although the presence of a header in this text may in fact be questionable, the fact that the model falls into this failure state calls into question the validity of the Viterbi sequence, and suggests that HMMs may lack the power to completely and accurately model this corpus.

## 4.2 Transformer

The logistic regression trained on  $\tilde{z}_{0,n}$  is able to predict whether human annotators identified a header in a text with 92% accuracy. By contrast, the model trained on  $\tilde{z}_{10,20}$  only achieves 77% accuracy, which is the same score achieved by simply predicting the majority class.

Thus the Transformer’s behaviour at the beginning of a text *is* predictive of expert opinions about the presence of a header in that text, but these behaviours do not persist into later parts of the document. As we had hypothesized, the model attends much more strongly to the beginning of texts where

	Expert Annotation			Initial HMM State		
	Header	No Header	$\Sigma$	State 7	Other	$\Sigma$
<b>LR Predicts Header</b>	596	44	640	421	219	640
<b>LR Predicts No Header</b>	19	136	155	19	136	155
$\Sigma$	615	180	795	440	355	795

Table 2: Contingency table comparing predictions from a logistic regression over  $\tilde{\mathbf{z}}_{0,n}$  against (left) the presence of a header in the tablet metadata, and (right) the initial state of the Viterbi sequence.

experts believe there to be a header: Figure 3 illustrates this using heatmaps of  $\tilde{\mathbf{z}}_{0,6}$  from two texts, one of which is annotated as having a header and the other of which is not.



Figure 3: Heatmap of  $\tilde{\mathbf{z}}_{0,n}$  (mean attention over signs before the first numeral, truncated to length 6) for two tablets, one with a human-labeled header (left) and one without (right). Darker cells indicate stronger attention.

Table 2 compares the predictions from the regression over  $\tilde{\mathbf{z}}_{0,n}$  against the expert annotations and the initial HMM states. The regression achieves significantly better recall (0.97) than the initial state of the HMM, which suggests that the HMM may have failed to identify a header in many texts where one does in fact exist.

## 5 Analysis

Our results are fully consistent with the prevailing assumption that the beginnings of certain PE tablets exhibit some degree of internal structure. This is suggested by the existence of an HMM state which is strongly localized to the beginning of tablets, but which does not occur at the beginning of every text as a generic “start” state. Further evidence is seen in the behaviour of the Transformer, where in certain texts the model pays more attention than usual to early tokens. On their own, these features merely confirm that some internal structure is present, but do not tell us what that structure may represent. In this section we interpret our models’ predictions in order to understand what factors may have motivated the original human annotations, and what features may be exploited to understand headers’ meanings.

### 5.1 Inter-Annotator Agreement

Table 3 reports inter-annotator agreement between our three approaches to labeling headers (expert annotations [Expert], initial HMM state [HMM], and

logistic regression over Transformer self-attention scores [LR]). We report Cohen’s  $\kappa$  (Cohen, 1960), where 1 (resp. -1) implies perfect agreement (resp. disagreement) and 0 implies no more agreement than expected if labels were assigned at random. The purpose of this comparison is not to evaluate the models’ accuracy (since it is not known that the expert labels reflect the ground truth) but rather to assess whether all three techniques recover similar information.

All techniques agree more than expected by chance. The most common disagreement comes from the HMM, which in 205 cases does not assign state 7 to a sign labeled as a header by human annotators. In 188 of these cases, the regression over Transformer attention *does* recover the human annotation, suggesting that these simply reflect the limited power of the HMM and its aforementioned susceptibility to noise from damaged contexts. Supporting this interpretation, most of these texts offer comparatively little context on which the HMM can base its decision: the majority contain unreadable signs, rare or hapax signs, or are very short. In fact, it is possible to predict whether the HMM will agree with the human annotation with better than chance accuracy simply by knowing whether the second sign of a tablet is intact, which suggests that the HMM is severely hampered by the fragmentary nature of the corpus.

	Expert	HMM	LR
Expert	1.0	0.372	0.766
HMM	0.372	1.0	0.362
LR	0.766	0.362	1.0

Table 3: Agreement (Cohen’s  $\kappa$ ) between human and model annotations.

Tablets that are damaged also impact the Transformer’s ability to recognize headers. There are 19 texts where the logistic regression fails to predict the presence of a header in the expert annotations, 17 of which are also disputed by the HMM. In all of

these texts, either the document contains only a single readable sign, or some early signs are damaged to the point of being unreadable.

Much more interesting are the cases where the regression proposes a novel header. This occurs in 44 texts, 13 of which also begin in state 7 according to the HMM.<sup>4</sup> Encouragingly, we find among this collection 25 texts<sup>5</sup> where the manner of transliteration indicates that experts have recognised a header but did not mark this according to the usual convention. If we correct the annotation of these texts, we find that the regression’s accuracy rises to 95% and  $\kappa$  to 0.849.

This leaves 18 cases where the regression proposes headers which are truly novel. Several of these texts are substantially intact and contain signs which are generally common and well-understood. M393~g in P008621 appears to the specialist a plausible header, since some other variants of M393 are so labeled, although other variants in second position are not marked as headers either by experts or the models (P009486; P009209). However, M362 (P009075) typically understood as a “counted object” sign (perhaps a nanny-goat, [Dahl 2005](#)) and the related |M362+M005| (P008294) challenge the conventional expectation that headers are distinct from counted objects. M489 is unique to P009526: [Damerow and Englund \(1989\)](#) keep open the possibility that M489 could be a header, but express some skepticism given that it also marks the summary line on the reverse (signs in the summary are usually expected to be counted objects). Specialists have not thoroughly fleshed out the distinction between “counted object” and “institution” signs, but believe that headers typically comprise the latter. The predictions from our models suggest that it may be worth considering whether “counted object” signs can also occur in some headers.

## 5.2 Multi-Sign Headers

Two-sign headers are a very marginal category in the expert annotations, occurring only five times.<sup>6</sup> By contrast, in 119 texts the Viterbi sequence stays in state 7 until the second sign, and in 33 texts it stays in state 7 until the third sign. The prevalence

<sup>4</sup>One of these texts, P008329, must be omitted as it has a damaged first sign. This was not removed during our data cleaning as the damage was not transliterated following the usual convention.

<sup>5</sup>Listed in Appendix B

<sup>6</sup>Listed in Appendix B

of long headers is one of the most significant points of divergence between the human labels and HMM states.

To predict the presence of a header with the Transformer, we perform a regression over all of  $\tilde{z}_{0,n}$  and therefore do not identify an explicit boundary where the header ends. However, by examining the coefficients from this regression, we can see that the outcome depends mainly on the attention paid to the second through fourth signs of the tablet, with mean attention to the second sign being most predictive overall. This suggests that the Transformer, like the HMM, has identified relevant structures beyond the first sign of a tablet.

In fact,  $\tilde{z}_1$  (the mean attention paid to the second sign of a tablet) is, by itself, sufficient to predict the presence of a header with the same accuracy as the entire  $\tilde{z}_{0,n}$ . Mean attention to the first sign ( $\tilde{z}_0$ ) gives the same accuracy as predicting the majority class (77%), suggesting that the first sign may be less relevant than the second to the rest of the document, despite it being the near-exclusive focus of past examinations of PE headers. We return to this discussion in Section 5.3, where we further explore the role of the second sign of a tablet.

In the expert annotations, most two-sign headers involve compounds of M327, generally followed by another sign which can also occur as a header on its own. This pattern recurs in the multi-sign headers identified by the HMM, and is expanded to cover more combinations of M327 compounds with a following sign. Notably, the HMM also introduces a new kind of multi-sign header not found in the human-labeled data, comprising M157 plus a following sign. An example of this is found at the beginning of the tablet shown in Figure 1, where the HMM replicates the manually-identified header but expands it to cover the first two signs of the text.

## 5.3 Cramér’s V

Cramér’s V ([Cramér and Goldstine, 1946](#)) measures relationships between pairs of categorical variables; it ranges from 0 to 1, where 0 signifies that the variables are unassociated and 1 denotes that they are perfectly associated. This section uses Cramér’s V (with the bias correction due to [Bergsma 2013](#)) to look for correlations which may have implications for the interpretation of headers. Our interpretation of V values follows the guidelines given by [Cohen \(1988\)](#).

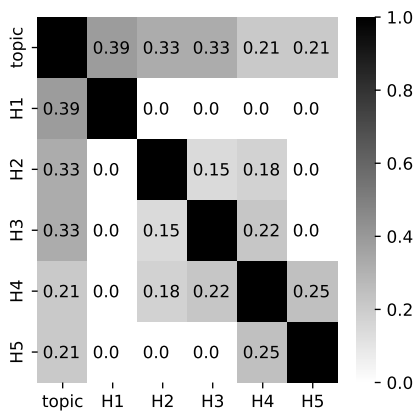


Figure 4: Heatmap showing the strength of the association (measured with Cramér’s  $V$ ) between the first five signs of a tablet and that tablet’s main topic according to an LDA model.

We begin by assessing whether and to what extent header information determines the content of a tablet. Let  $H_n$  be a categorical variable denoting the name of the  $n$ th sign in a tablet, and let *topic* denote the topic with which a tablet is most strongly associated according to the PE topic model produced by Born et al. 2019. Figure 4 depicts Cramér’s  $V$  between all of  $H_n$  and *topic* for  $1 \leq n \leq 5$ .

*topic* has a strong to moderate association with all of the  $H_n$  features; its strongest relation is  $V = 0.39$  with  $H_1$ , implying that the first sign of a tablet strongly predicts the genre of the following text.  $V$  drops monotonically for later signs, implying that genre-defining information is primarily localized to the beginning of a text.

Most of the  $H_n$  features exhibit moderate to weak associations with one another, however there is no association at all ( $V = 0$ ) between  $H_1$  and  $H_n$  for  $n > 1$ . This would suggest that the first sign of a tablet is somehow disjoint from the rest of the text, and though this sign may predict the overall topic of the following text, it does *not* predict exactly which signs will immediately follow. This has implications for the interpretation of multi-sign headers, as it suggests that they may not comprise a unified whole (such as a two-sign-long word) so much as a concatenation of distinct signs with complementary roles.

To assess this further, we introduce a variable *long\_header* which is True just in case the HMM proposes the existence of a multi-sign header.  $H_1$

has no association ( $V = 0$ ) with *long\_header*, meaning the first sign of a text does not predict whether the HMM will identify the presence of a multi-sign header. By contrast,  $H_2$  and  $H_3$  have a very strong association to *long\_header* ( $V = 0.54$  and  $0.61$ ), and  $H_4$  is only slightly weaker ( $V = 0.43$ ).  $H_5$  also has no association. The lack of association with  $H_1$  further suggests that multi-sign headers are not variants or refinements of whatever sign occurs in the first position, and are rather concatenations of disjoint pieces of information.

Some texts bear one or more *seal* impressions; PE seals depict objects and animals and their use on tablets records extra-textual information related to administrative practice. A *subscript* (see Figure 1) is a string of signs which occurs at the very end of some tablets, after the final numeral. Subscripts are unique in PE in that they are not directly followed by a numeral.

We introduce a categorical variable representing whether a text has a seal (resp. subscript), and another representing *which* seal (resp. subscript) is present.  $H_1$  does not determine whether a text is sealed ( $V = 0$ ); however, it does predict *which* particular seal was used ( $V = 0.39$ ). Intriguingly,  $H_2$  shows the opposite pattern, and weakly determines whether a text is sealed ( $V = 0.14$ ) but *not* which seal was used ( $V = 0$ ). A similar pattern holds for subscripts, where  $H_2$  predicts the presence of a subscript ( $V = 0.27$ ) and  $H_1$  does not ( $V = 0$ ), though in this case both  $H_1$  and  $H_2$  predict the text of the subscript ( $V = 0.30, 0.33$  resp.).

$H_1$  is strongly predictive of another variable *prov*, which records a text’s provenience ( $V = 0.56$ ), which could support theories that headers relate to activities undertaken at particular locales. Given that  $H_1$  also correlates with seal impressions, it is possible that the first sign of a tablet may convey information about where the tablet was sealed (and thus, likely, where it was written).

In sum, we have seen that the first sign of a tablet predicts extra-textual information such as provenience and choice of seal impression, but fails to predict textual information such as the signs that occur near to itself or the presence of a subscript. By contrast, the second sign of a tablet predicts textual content such as adjacent signs and the presence and content of a subscript, as well as some extra-textual content such as the presence of a seal. The first sign thus appears to look “outward” at the administrative context surrounding a text, whereas

the second looks “inward” at the text itself.

#### 5.4 Compositionality in Header Signs

Complex graphemes (CGs) are cases where one sign appears to be written inside of or otherwise ligatured with another. CGs are common near the beginning of PE tablets, and many of the human-annotated headers are CGs themselves or participate in the construction of CGs in other contexts.

Born et al. (2021) measure additive compositionality in PE sign embeddings learned by a variety of contextual embedding models. They show that certain CGs tend to receive *compositional* embeddings which are close to the sum of the embeddings of the signs used in their construction. A similar pattern has been observed (Mikolov et al., 2013; Salehi et al., 2015; Cordeiro et al., 2016) in modern languages where phrasal representations are often close to the sum of their parts, but only when the phrase is semantically compositional. Embeddings for idiomatic phrases are less likely to receive compositional embeddings. Born et al. exploit these patterns to divide the set of CGs into two groups: those which are probably semantically compositional (and thus may be understood if their parts are deciphered) and those which are idiomatic and likely to pose a greater challenge for decipherment.

We hypothesize that there may be some relation between a CG’s compositionality and its tendency to occur in headers. To test this, for every CG  $|X+Y|$ , we measure the cosine similarity between the embedding for  $|X+Y|$  and the sum of the embeddings for  $X$  and  $Y$  using the embeddings from Born et al. 2021’s best performing model.<sup>7</sup> Table 4 shows the average similarity for CGs occurring in headers (as identified by any of our three approaches), and for CGs in non-initial position.<sup>8</sup> We perform the averaging both over tokens (so that a CG occurring at the beginning of multiple tablets is included in the average multiple times) and over types (so that each CG is included in the average at most once).

CGs occurring in headers (according to any of the three possible labelings) are on average more

<sup>7</sup>Born et al. (2021) demonstrate that their embeddings reflect experts’ understandings of signs and exhibit interpretable patterns of compositionality. We use their embeddings because the same has not yet been shown for embeddings from the Transformer model in this work.

<sup>8</sup>Since the Transformer does not identify an explicit boundary to the header, we only count a CG as being part of the header when it is the first sign of the tablet. If long headers really exist, it is possible that some CGs which are not the first sign of the tablet should still be counted as part of a header.

compositional than CGs occurring in the body of a text. The difference is not significant when averaged over types, but is highly significant when averaged over tokens ( $p \ll 0.01$ , Mann-Whitney U). This likely reflects the fact that (i) the more frequent a CG is in tablet-initial position, the more compositional it is<sup>9</sup>, and (ii) there are many fewer types than tokens, so those samples are too small to show significance.

Mean compositionality is lower for the expert annotations than for the other approaches, but the difference is not significant. This difference is mainly a consequence of the broken and fragmentary tablets where our models fail to identify a header that is present in the human annotations. Many of these tablets begin with a CG, and many of these CGs are non-compositional, possibly because they occur in short and fragmented contexts and therefore receive poor quality embeddings.

The apparent overlap between headers and more compositional CGs on the one hand, and non-headers and less compositional CGs on the other, increases our confidence that CGs can be partitioned into measurably distinct groups and should not necessarily be conceived of or analyzed as a monolithic category.

## 6 Related Work

HMMs have a storied pedigree in the field of decipherment, being first used (under codename PTAH) by members of the NSA to analyze the Voynich MS (D’Imperio, 1979). As this work was originally classified, most HMM-based approaches to decipherment instead trace back to Knight et al. 2006 who demonstrate the effectiveness of HMMs on a range of unsupervised decipherment tasks, and whose framework is adopted or used as a baseline in a significant volume of later work (Ravi and Knight 2009; Snyder et al. 2010; Knight et al. 2011; Reddy and Knight 2011; Berg-Kirkpatrick and Klein 2013; Kim and Snyder 2013 *inter alia*). We are not aware of any work which has employed Transformers or other neural architectures as feature extractors for a comparable, unsupervised analysis of undeciphered text.

<sup>9</sup>There is no significant correlation between sign frequency and compositionality in general; this trend only (weakly) applies to tablet-initial signs.



	Tablet-Initial CGs			Non-Initial CGs
	Expert	HMM	LR	
Avg. cos over tokens	<b>0.682</b>	<b>0.693</b>	<b>0.683</b>	0.565
Avg. cos over types	0.608	0.648	0.616	0.593

Table 4: Mean compositionality of complex graphemes found in expert-annotated headers (Expert), in headers identified using HMM state 7 (HMM), in headers predicted by logistic regression over Transformer self-attention (LR), and in non-initial positions. Bolded values differ significantly from the rightmost column.

## 7 Conclusion

This work offers the first and most exhaustive assessment of proto-Elamite headers in order to inform the ongoing decipherment of this ancient script.

We have demonstrated that two distinct unsupervised sequence modeling techniques exhibit unique behaviours at the beginning of some proto-Elamite texts. These behaviours are consistent with, and offer independent evidence in support of, the prevailing hypothesis that these documents begin with a header.

The features recovered by these models predict with up to 95% accuracy whether experts understand a text to contain a header. This inspires confidence that the expert labels have been applied according to a consistent logic and following salient structural features of the text. Our error analysis has also allowed us to identify and emend 25 mistakes in the expert annotations, expanding the total number of headers in the corpus by nearly 4% and reducing the amount of noise in a low-resource dataset where small errors may have an outsize effect.

We have demonstrated that there are measurable differences between the contextual embeddings learned for signs labeled as headers versus those in other contexts, reaffirming that these signs are somehow functionally distinct from the rest of the script.

Using self-attention scores from a Transformer language model, we have demonstrated that the *second* sign of a text predicts the presence of a header more accurately than the first sign; we have also shown that state sequences from an HMM suggest that many more multi-sign headers exist than were previously assumed. On the basis of these results we have argued against the conventional understanding that header information is localized to a single sign, and suggest that headers may commonly span two or even three signs in some texts.

Finally, we identify correlations between sign

usage at the beginning of a text and other features such as genre, seal impressions, and the presence of a subscript. These correlations suggest that the first sign of a text captures more extra-textual information than later signs, and that if multi-sign headers exist, their two (or more) constituent signs likely convey distinct kinds of information.

## Acknowledgements

We extend our sincere thanks to the reviewers for their many helpful comments and suggestions, and to Jacob Dahl whose tireless efforts have made this work possible. This research was partially supported by the Natural Sciences and Engineering Research Council of Canada grants NSERC RGPIN-2018-06437 and RGPAS-2018-522574 and a Department of National Defence (DND) and NSERC grant DGDND-2018-00025 to the fourth author, and by an NSERC award CGSD3-547773-2020 to the first author.

## Limitations

Although we introduce techniques which may generalize to other settings (e.g., the use of Transformer attention as an analytic device for decipherment), our conclusions focus on a singularly unique writing system and thus have very narrow applicability.

Due to the undeciphered nature of our data, we cannot compare against any ground truth. Thus, while we are confident in the patterns we have identified, our interpretations of these patterns cannot be definitively evaluated until this script is better understood.

## Ethics Statement

As noted under Limitations, undeciphered data admit many interpretations which cannot always be evaluated against a ground truth. In light of this, a responsible analysis of such data must be informed by domain experts; without their involvement, any analysis is liable to be groundless speculation. The

authors of this work include Assyriologists and proto-Elamite specialists, who have been involved at every step in order to ground these results as much as possible in the existing scholarship.

## References

- Taylor Berg-Kirkpatrick and Dan Klein. 2013. Decipherment with a million random restarts. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 874–878.
- Wicher Bergsma. 2013. A bias-correction for cramér’s *v* and tschuprow’s *t*. *Journal of the Korean Statistical Society*, 42(3):323–328.
- Logan Born, Kate Kelley, Nishant Kambhatla, Carolyn Chen, and Anoop 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*, pages 122–132, Minneapolis, USA. Association for Computational Linguistics.
- Logan Born, Kathryn Kelley, M. Willis Monroe, and Anoop 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*, pages 4136–4146, Online. Association for Computational Linguistics.
- Robert L. Cave and Lee P. Neuwirth. 1980. Hidden markov models for english.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1):37–46.
- Jacob Cohen. 1988. *Statistical Power Analysis for the Behavioral Sciences*, 2nd edition. Routledge.
- Silvio Cordeiro, Carlos Ramisch, Marco Idiart, and Aline 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)*, pages 1986–1997, Berlin, Germany. Association for Computational Linguistics.
- Harald Cramér and Herman H. Goldstine. 1946. *Mathematical methods of statistics*. Princeton: Princeton University Press.
- Jacob L. Dahl. 2005. Animal husbandry in Susa during the proto-Elamite period. *Studi Micenei ed Egeo-Anatolici*, 47:81–134.
- Jacob L. Dahl. 2019. *Tablettes et fragments Proto-élamites / Proto-Elamite Tablets and Fragments*. Département des Antiquités Orientales.
- Peter Damerow and Robert K. Englund. 1989. *The proto-Elamite texts from Tepe Yahya*, volume 39 of *American School of Prehistoric Research: Bulletin*. Peabody Museum, Cambridge, Massachusetts.
- Mary D’Imperio. 1979. An application of PTAH to the Voynich Manuscript (U). In *National Security Agency Technical Journal*, volume 24, pages 65–91.
- Robert K. Englund. 2004. The state of decipherment of proto-Elamite. *The First Writing: Script Invention as History and Process*, pages 100–149.
- Kathryn Kelley. 2018. *Gender, Age, and Labour Organization in the Earliest Texts from Mesopotamia and Iran (c. 3300–2900 BC)*. Doctoral dissertation, University of Oxford.
- Young-Bum Kim and Benjamin Snyder. 2013. Unsupervised consonant-vowel prediction over hundreds of languages. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1527–1536, Sofia, Bulgaria. Association for Computational Linguistics.
- Kevin Knight, Beáta Megyesi, and Christiane Schaefer. 2011. The Copiale cipher. In *Proceedings of the 4th Workshop on Building and Using Comparable Corpora: Comparable Corpora and the Web*, pages 2–9. Association for Computational Linguistics.
- Kevin Knight, Anish Nair, Nishit Rathod, and Kenji Yamada. 2006. Unsupervised analysis for decipherment problems. In *Proceedings of the COLING/ACL on Main Conference Poster Sessions, COLING-ACL ’06*, pages 499–506, Stroudsburg, PA, USA. Association for Computational Linguistics.
- Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey 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, NIPS’13*, page 3111–3119, Red Hook, NY, USA. Curran Associates Inc.
- Sujith Ravi and Kevin Knight. 2009. Learning phoneme mappings for transliteration without parallel data. In *Proceedings of human language technologies: The 2009 annual conference of the north american chapter of the association for computational linguistics*, pages 37–45.
- Sravana Reddy and Kevin Knight. 2011. What we know about the Voynich manuscript. In *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities*, pages 78–86.
- Bahar Salehi, Paul Cook, and Timothy 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*, pages 977–983, Denver, Colorado. Association for Computational Linguistics.

Vincent Scheil. 1935. *Texte de comptabilité Proto-Élamite (Troisième Série)*, volume 26 of *Mémoires de la Mission Archéologique de Perse*. Paris: Librairie Ernest Leroux.

Benjamin Snyder, Regina Barzilay, and Kevin Knight. 2010. [A statistical model for lost language decipherment](#). In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1048–1057. Association for Computational Linguistics.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. [Attention is all you need](#). In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 5998–6008.

## A Reproducibility Details

We implement our HMM using the `hmmlearn` package for Python. We train our Transformer LM following the instructions at [https://github.com/facebookresearch/fairseq/blob/main/examples/language\\_model/README.md](https://github.com/facebookresearch/fairseq/blob/main/examples/language_model/README.md). Our corpus is small, and this model trains on a single GTX 1070 for approximately one hour.

Our revisions to the corpus from [Born et al. 2019](#) are available at <https://github.com/sfu-natlang/pe-headers>. We also include a csv listing the expert labels and the predictions from our models.

We preprocess the data by removing all comments and annotations (lines beginning in \$, &, or #) and deleting the , character which marks entry boundaries (entries are logical units delimited by explicit numeral notations). We remove annotations marking damage and corrected signs (characters matching the regular expression `[\[\]<>#?!]`). We delete newlines from each text and compile the corpus into a file with one complete tablet per line. We shuffle the lines of this file and set aside 200 tablets as a validation set. The Transformer LM is trained directly on the data at this stage, tokenized on spaces only (we do not use a subword tokenizer). Before training the HMM, we circumfix beginning- and end-of-sequence tokens `<bos>` and `<eos>` to each line.

The embeddings which we use to evaluate compositionality are available upon request to the authors of [Born et al. 2021](#).

## B Lists of Texts

This section summarizes which texts belong to certain categories identified in the body of the paper. Texts are identified by the P-number assigned by the CDLI.

Texts with an implicit header, for which we have corrected the transliteration:

P008020, P008251, P008255, P008311, P008365, P008463, P008641, P008845, P008850, P008853, P008878, P008880, P009051, P009053, P009055, P009060, P009094, P009126, P009320, P009422, P009441, P009461, P009469, P393079, P393080

Human-labeled two-sign headers:

P009524, P008220, P008258, P008281, P008702