

# Disambiguating Numeral Sequences to Decipher Ancient Accounting Corpora

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

A numeration system encodes abstract numeric quantities as concrete strings of written characters. The numeration systems used by modern scripts tend to be precise and unambiguous, but this was not so for the ancient and partially-deciphered proto-Elamite (PE) script, where written numerals can have up to four distinct readings depending on the system that is used to read them. We consider the task of disambiguating between these readings in order to determine the values of the numeric quantities recorded in this corpus. We algorithmically extract a list of possible readings for each PE numeral notation, and contribute two disambiguation techniques based on structural properties of the original documents and classifiers learned with the bootstrapping algorithm. We also contribute a test set for evaluating disambiguation techniques, as well as a novel approach to cautious rule selection for bootstrapped classifiers. Our analysis confirms existing intuitions about this script and reveals previously-unknown correlations between tablet content and numeral magnitude. This work is crucial to understanding and deciphering PE, as the corpus is heavily accounting-focused and contains many more numeric tokens than tokens of text.

## 1 Introduction

Proto-Elamite (PE) is a partially-deciphered script unearthed at early 3rd millennium BCE sites across the Iranian plateau. This script was exclusively used to record spreadsheet-style administrative accounts, and well over half of the attested glyphs are known to be digits. Despite this abundance of numeric notations, no large-scale quantitative analysis of PE numerals has ever been undertaken, and most prior work has focused on the adjoining *text*. This is likely because the script employs multiple distinct number systems, which use partially-overlapping sets of digits and occasionally assign distinct values to identical-looking sign shapes (Figure 1). Many

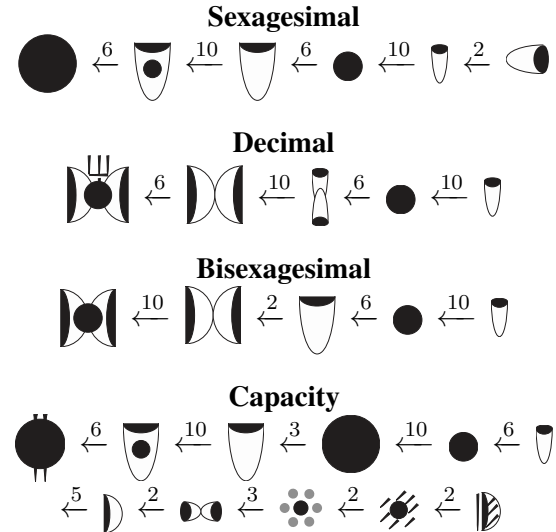


Figure 1: Relative values of digits in the main proto-Elamite number systems.  $X \xleftarrow{n} Y$  means that one  $X$  has the same value as  $n$   $Y$ s.

numerals can be read according to two or more of these systems, and represent different values depending on the system used.

Some PE signs have been given tentative readings in prior work, and on the basis of these readings there appears to be a regular relationship between the kind of object recorded and the number system used to count it (in a manner not entirely dissimilar to the measure words found across East Asian languages). Knowing which system is in use for a given numeral therefore increases the possibility of understanding what category of object is recorded in the adjoining text, and thus opens new avenues for the ongoing decipherment of this script.

In this work, we consider the task of disambiguating which systems are used in ambiguous PE numeral notations in order that the values of these numerals may be determined. We describe a simple rule-based technique to extract lists of possible readings from PE numeral notations, which allows

us to give the first large-scale survey of PE numerals since Friberg 1978 (whose manual analysis occurred at a time when fewer texts were known). We then propose two disambiguation techniques, one based on the subset-sum problem and another which uses a bootstrap classifier (Yarowsky, 1995). We describe the construction of a test set for evaluating PE numeral disambiguation models, and propose a novel approach to cautious rule selection which significantly improves the performance of a bootstrap classifier on our data. Our analysis shows how these techniques lead to a deeper understanding of this ancient and undeciphered writing system.

## 2 Data & Background

We base our analysis off the transliterated PE corpus hosted by the CDLI.<sup>1</sup> Each text in this corpus contains an optional header, followed by a series of “entries” written one per line of the transliterated file. Each entry contains a (possibly empty) span of text, a comma delimiter, and a (possibly empty) numeral notation. The transliterations use a work-in-progress signlist that reflects experts’ current understanding of the texts, but which may not exactly match the true character inventory of the underlying script. In this signlist, characters that are believed to represent text are transliterated with labels beginning in “M” (e.g. M001, after Meriggi who pioneered the study of this script), and those representing digits are labeled with “N” (e.g. N01). The notation  $n(N00)$  means that the digit  $N00$  is written down  $n$  times. Figure 2 shows an example of a tablet alongside its transliteration.

Most proto-Elamite numerals are written using one of four<sup>2</sup> number systems, which are called *decimal* (D), *sexagesimal* (S), *bisexagesimal* (B), and *capacity* (C). In spite of their names, all of these systems use mixed radices. Figure 1 shows the relative values of the digits in each of these systems, as derived through prior manual analyses (Friberg, 1978; Damerow and Englund, 1989; Dahl, 2019). Unlike Hindu-Arabic notation, where the value of a digit depends on its position, the values

<sup>1</sup>Cuneiform Digital Library Initiative, <https://cdli.mpiwg-berlin.mpg.de>; corpus downloaded 3 Oct 2022.

<sup>2</sup>Additionally, there are marginal systems (labeled B#, C#, and C”) which appear to be derived from one of the four main systems by the addition of hatch marks or boxes drawn around the digits. These systems are rare, and the extra hatching or boxing makes them trivial to identify, so we ignore them for the remainder of this work.



Text	Numeral
M056~f	1(N34) 5(N14) 1(N01) 1(N8B) = $111.5 \times N01^S$
M341 M288	7(N14) 2(N01) 3(N39B) = $44.6 \times N01^C$

Figure 2: Proto-Elamite tablet MDP 26, 177 (Scheil 1935; P008805) alongside its transliteration and converted readings for both numerals. The tablet is read right-to-left: M056~f is the sign in the top-right corner. Observe that  $111.5/44.6$  equals the 2.5:1 ratio noted in Section 4.1.2.

of proto-Elamite digits are fixed, and larger values are denoted by repeating a digit multiple times.

Note that some digits can be used with distinct values in multiple systems (e.g.  $N14 \bullet$  equals  $10 \times N01 \cup$  in S, but only  $6 \times N01 \cup$  in C): this means that it can sometimes be impossible to determine the absolute value of a numeral unless context makes clear which system it employs.

Since  $N01 \cup$  occurs as part of every number system, we use this sign as a standard unit and report values as multiples of  $N01$  whenever we convert to Hindu-Arabic notation. The S, D, and B systems are understood to represent unitless cardinal numbers, while C records *unitful* measures of volume. When it is necessary to emphasize that these systems are not commensurable, we add a superscript to denote the system in use: e.g.  $12 \times N01^C$  is a measure of volume which is not equivalent to  $12 \times N01^D$ , despite having identical magnitude.

We summarize the above points with an illustrative example. The following notation (read from right-to-left)



would be transliterated as 1(N45) 2(N14) 7(N01). This numeral must use either the S or C system, as

the large circle N45 only occurs in these systems (Figure 1). Using the readings from the S system, this notation encodes a value of

$$3627 \times N01 = (1 \times 3600 + 2 \times 10 + 7 \times 1) \times N01$$

Using the C system, it instead encodes

$$79 \times N01 = (1 \times 60 + 2 \times 6 + 7 \times 1) \times N01$$

Of these, we know that the S reading must be the correct one, since  $7(N01)$  should never occur in the C system (every 6 N01 would be bundled into an N14, so the actual notation for  $79 \times N01^C$  would be  $1(N45) 3(N14) 1(N01)$ ).

It is not clear whether these notations would have been considered ambiguous at the time they were written. It is very plausible that the original scribes would have been able to infer the correct reading for a numeral based on contextual information which is not salient to the modern reader (for example, by knowing that certain items are consistently counted with a particular number system, a possibility discussed by Englund 2004, 2011). There is equally the possibility that these notations were ambiguous even to their authors, but that this ambiguity did not interfere with their intended use, for example if the PE texts were intended for short-term use when the scribes would still recall which system they intended at the time of writing.

### 3 Methodology

#### 3.1 Automated Conversion

We extract all of the numeral notations from the transliterated corpus by using a regular expression to find every contiguous sequence of N-signs; we discard sequences which are damaged, which we identify as being immediately adjacent to a transliterated X or . . . . Algorithm 1 uses the relative values from Figure 1 to automatically extract a dictionary of possible readings for each of these numerals.

Of the 8011 intact numerals which we have extracted, there are 7954 for which this conversion returns at least one reading. Of these, only 1919 unambiguously belong to a particular number system: the remainder are ambiguous between two, three, or even all four systems (Table 1). The following sections outline two proposals for disambiguating the ambiguous cases. Section 4.1.1 discusses the 57 cases for which there is no valid reading in any system.

---

#### Algorithm 1 PE Numeral Readings

---

**Input:**  $digits = [(n_1, sign_1), \dots, (n_k, sign_k)]$   
 $\triangleright$  A list of signs and num. times each one occurs.

**Returns:** A map from number systems to possible readings for this digit list.

---

**for**  $sys \in \{S, D, B, C\}$  **do**

$value_{sys} \leftarrow 0$

**for**  $(n, sign) \in digits$  **do**

**for**  $sys \in \{S, D, B, C\}$  **do**

**if**  $sign \notin signs\_used\_by(sys)$  **then**

$value_{sys} \leftarrow \perp$

$\triangleright \perp$  means there is no valid reading in this system.

**continue**

**if**  $n > max\_count(sign, sys)$  **then**

$\triangleright max\_count$  returns the max num. of times this digit can occur before it would carry over to a higher value digit.

$value_{sys} \leftarrow \perp$

$v \leftarrow$  value of  $sign$  in  $sys$

$value_{sys} \leftarrow value_{sys} + n \times v$

$\triangleright \perp$  plus anything equals  $\perp$

**return**  $\{sys \mapsto value_{sys} \forall sys \in \{S, D, B, C\}\}$

---

#### 3.2 Subset-Sum Analysis

Our first approach to disambiguation relies on the fact that some PE documents end in a summary line, which records the total sum of the preceding entries. Although the entries themselves may be ambiguous, the sums naturally record larger amounts and are therefore more likely to use high-magnitude digits that unambiguously belong to a particular system. When a tablet records values from multiple number systems, they are summarized separately; thus if we can identify unambiguous summaries, we can infer that all of the entries which they sum must belong to the same system.

To achieve this, we filter the corpus to find texts with one or two entries on the reverse, as current understandings of the corpus suggest that these are likely to be summaries.<sup>3</sup> For each of these texts, we solve an instance of the subset-sum problem to identify whether any combination of readings from the obverse adds up to the same value as any reading of the reverse.

---

<sup>3</sup>Some transliterations include an annotation which explicitly labels a particular entry as a summary. However, not all summaries are labeled in this way, so we rely on automatic detection of summaries to expand the number of texts available for this analysis.

Possible Readings	Number of Numerals
none	57
B	19
C	1727
D	66
S	107
B or D	21
B or S	5
C or S	143
B, C, or S	185
B, D, or S	292
B, C, D, or S	5389

Table 1: Distribution of readings produced by our automated conversion. A majority of numerals in the corpus can be read using *any one* of the four number systems.

If an accurate summation is found, and any of the component terms has an unambiguous number system, we use this as evidence to disambiguate the entire text to that system. We manually evaluate this approach by confirming with domain experts whether the resulting disambiguations are correct.

### 3.3 Bootstrapping

Some of the PE numeral notations are inherently unambiguous, either because they use a digit which only occurs in a single system, or because they contain more instances of a digit than would be allowed by some systems. We propose to use these cases as seed rules to train a bootstrap classifier (Yarowsky, 1995) for disambiguation.

We choose bootstrapping because it requires only a small number of seed labels, and as seen in Table 1 some systems have few unambiguous attestations. Moreover, bootstrapping yields interpretable results which can be understood by examining the label distribution associated with each input feature. This helps to legitimize model outputs to domain experts, and to situate model predictions relative to prior manual analyses.

Table 2 lists the features used by our classifier. A numeral’s initial label distribution is uniform over every system for which our automated conversion returns a valid reading, and zero elsewhere. We use the DL-2-ML algorithm (Haffari and Sarkar, 2007; Abney, 2002), which models  $\pi_x(j)$  (the likelihood that sample  $x$  belongs to class  $j$ ) as

Feature	Description of Value(s)
TABLET	The tablet where this numeral occurs.
FIRST_SIGN	The first sign of the tablet where this numeral occurs (where we may expect to find a header).
SAME_ENTRY	Bag of signs which occur in the entry preceding this numeral.
SAME_TABLET	Bag of signs which occur anywhere on the same tablet as this numeral.
OBJECT	The sign immediately preceding this numeral (where we may expect to find a counted object).
IMPLICIT_OBJECT	The last sign in the first entry of the text where this numeral occurs (where we may expect to find an implicit object).

Table 2: Each numeral is associated with a set of features from this list, which we use to train our bootstrap classifiers.

$$\pi_x(j) \propto \prod_{f \in F_x} \theta_{fj}$$

where  $F_x$  is the set of features associated with sample  $x$ , and  $\theta_{fj}$  is a learnable parameter which measures the association between feature  $f$  and class  $j$ . We apply the “cautious” approach from Collins and Singer 1999, which limits the number of rules that can be added to the decision list at each iteration of training. Specifically, candidate rules are sorted according to the number of labeled examples that support them, and only the  $n$  with the largest support are added to the decision list.<sup>4</sup>  $n$  starts at 5 and increases by 5 each iteration.

The cautious algorithm was motivated by the observation that “the highest frequency rules [are] much ‘safer’ [than low-frequency rules], as they tend to be very accurate” (Collins and Singer, 1999). This observation does not seem to hold for our data, where many of the most frequent features barely meet the confidence threshold to be added to the decision list and, impressionistically, do not appear to be any more accurate than those with lower frequency. We therefore propose a novel approach to cautious rule selection whereby  $\theta_f$  is

<sup>4</sup>Whitney and Sarkar 2012 note that many details are omitted from the description of the cautious algorithm in Collins and Singer 1999. We follow Whitney and Sarkar in assuming that confidence thresholding is performed using unsmoothed label counts. However, we differ from their approach by selecting the top  $n$  rules overall (not the top  $n$  for each label) as this yields stronger results on our data.

only updated if the update increases  $\max_j \theta_{f,j}$ , i.e. if it increases the confidence of the label distribution associated with feature  $f$ . In this setting we visit the features in a random order each iteration, to prevent a degenerate outcome whereby endless incremental updates are made only to the first rules visited.

System	Number of Test Items
B	3
C	18
D	14
S	13

Table 3: Distribution of target classes in our numeral disambiguation test set. This set contains every instance of the B class which we were able to manually disambiguate with the help of domain experts; the other classes are kept small to maintain as balanced a distribution as possible.

Prior to training, we upsample seeds with rare labels to obtain an equal number for each class. We evaluate our classifiers on a test set which we construct by manually disambiguating some of the ambiguous notations in the corpus. We endeavoured to keep this set as balanced as possible, but some systems (particularly B) can only be confidently identified in a few texts. This means the test set cannot contain every numeral for which we know the target label, as doing so would yield too great an imbalance between classes. Appendix A describes what evidence was used to disambiguate each numeral in the set, and Table 3 summarizes the overall class distribution. All of the labels in the test set have been verified by domain experts.

## 4 Results

### 4.1 Automated Conversion

#### 4.1.1 Invalid Notations

There are 57 intact numerals for which our automated conversion does not return a valid reading according to any number system. The vast majority of these violate the bundling principles established in prior work and shown in Figure 1. For example, the notation 11(N01) in P008043 should not be attested in any system, as every system carries over to a higher digit after at most ten N01s.

Some of these illegal notations may reflect errors on the part of the original scribes. For ex-

Systems Used	Number of Tablets
C and S	12
C and D	15
C and B	4
S and D	1

Table 4: Number of tablets which unambiguously use more than one number system.

ample, in P008844, the sum of the D values on the obverse equals 9(N23) 3(N14) 3(N01), or 573; the scribe has actually written 9(N23) 7(N14) 3(N01), which violates the usual principle that 6 N14s carry over to one N23. This suggests that the scribe may have conceived of the D system as a truly decimal notation (in which case the least significant digits would be written as 7 tens and 3 ones, exactly as we find on the tablet), forgetting that N23 uses a different radix. Several of the other aberrant notations lend credence to this view, such as P008788 which apparently records 88 M367s (likely goats, usually counted with D) as 8(N14) 8(N01). These cases suggest a lack of standardisation across scribes or across documents, which is consistent with the longstanding view that the writing system never achieved a significant degree of standardisation (Dahl, 2019).

#### 4.1.2 Mixed Systems

Our automated conversion reveals numerous accounts (Table 4) which unambiguously use two different number systems (no tablets unambiguously use more than two systems). In all but one of these, the C system occurs alongside one of the “integer” systems S, D, or B. This suggests a general pattern of accounts which record capacities of goods received/disbursed from/to individual people, animals, households, or other entities also counted in whole numbers on the tablet. The text P009383 is unique in that it unambiguously uses two of the “integer” systems S and D. On close inspection, however, the original tablet is heavily abraded where the putative S notation occurs, and the sign which forces this notation to be read as S (N08) is almost entirely unreadable. Given the otherwise total absence of texts which mix integer systems, we posit that this text may contain a transliteration error and that the broken sign is not in fact an N08.

Several texts which mix the S and C systems also have other features in common. P008796, P008798,

and P008805 are exemplary of this group, which are all two-entry texts where the first entry is an S-denominated amount of M056~f  $\frac{1}{2}$  (possibly a plow), and the second is a C-denominated amount of M288. In all of these texts, there are exactly  $2.5 \times N01^S$  per  $N01^C$ : this ratio was previously identified and discussed by Damerow and Englund (1989); Englund (2004). P008791 appears to belong to this same class of texts, and given that the first entry records  $128 \times N01^S$  M056~f we should expect  $51.2 \times N01^C$  in the final entry (or  $8(N14)$   $3(N01)$   $1(N39B)$ ). We actually find  $52 \times N01^C$ , or  $8(N14)$   $4(N01)$ . Close inspection of the tablet, however, reveals that there has been a transliteration error, and that the final sign, although mostly broken, is recognizably an N39B. The text yields the expected ratio when this mistake is corrected. Errors such as these are much easier to identify when dealing in converted Arabic numerals, which modern readers can understand and manipulate more quickly and intuitively than the original PE notations.

Out of 244 signs which precede at least two unambiguous notations, only 11 occur next to notations from distinct systems. These 11 (M001, M056~f, M059, M096, M124, M218, M305, M327, M371, M387, M388) include signs which have been speculated (Dahl, 2019) to represent human laborers or overseers (M388, M124), signs with possible syllabic values (M001, M096, M218, M387, M371), and headers or account owners (M059, M305, M327). In other words, these are signs which we expect to qualify or describe an object being counted, and not to be counted themselves. It therefore appears that, while counted objects are consistently recorded using one particular number system, these qualifying signs can potentially be used to qualify objects from several different systems. This suggests a novel approach to determine the function of signs with unknown meanings, by looking at the variety of number systems they occur beside.

## 4.2 Subset-Sum Analysis

Our subset-sum analysis identifies 24 texts which, upon manual inspection, can be fully or partially disambiguated based on their summary line(s). We highlight one which is of particular interest. In P008014, all entries must use the C system in order to equal the same value as the unambiguous C summary on the reverse. The text of the summary

contains only the “grain container” sign M288, implying that the entire tablet should record amounts of M288. However, on the obverse of the tablet, M288 only occurs as the final sign of the very first entry. This suggests that the scribe has only explicitly marked the counted object in the first entry, and left it implicit in the following entries (this is in keeping with known practices from other ancient Mesopotamian accounting corpora; Nissen et al. 1993, pp. 37–38, Englund 2001). This means that there exist long-distance dependencies between the entries in some texts, which need to be accounted for if these texts are to be fully understood.

## 4.3 Bootstrapping

Figure 3 and Table 5 compare results from our two approaches to bootstrapping. The baseline, vanilla bootstrapping algorithm achieves a mediocre 0.60 F1. Recall of the B and S classes is perfect, but only half of the instances of the C class are correctly labeled. This classifier does not seem to have uncovered any clear signals to identify the D system: 3 instances of this class remain unlabeled (none of their associated features ever attained the required confidence to be added to the decision list), and those which are labeled are distributed uniformly across all of the possible classes.

Model	4-way			2-way		
	prec.	rec.	F1	prec.	rec.	F1
Baseline	0.64	0.56	0.60	0.74	0.65	0.69
Ours	<b>0.88</b>	<b>0.88</b>	<b>0.88</b>	<b>0.90</b>	<b>0.90</b>	<b>0.90</b>

Table 5: Numeral disambiguation results. In the 4-way setting, we seek to identify exactly which number system is in use for each numeral. In the 2-way setting, we only seek to distinguish C notations from everything else.

By contrast, our proposed approach to cautious rule selection yields 0.88 F1, with significantly better recall of both the C and D classes than the baseline. This suggests that frequency-based caution may be ill-suited for bootstrapping on some datasets, and that, in settings where bootstrap classifiers remain viable, it may be worthwhile to explore alternative approaches to cautious rule selection.

Note from Figure 1 that ambiguities between the S, D, and B systems primarily come from the digits  $N01$   $\cup$  and  $N14$   $\bullet$ , which have the same relative values across all three systems. S and B further overlap in the sign  $N34$   $\cup$ , which also maintains the

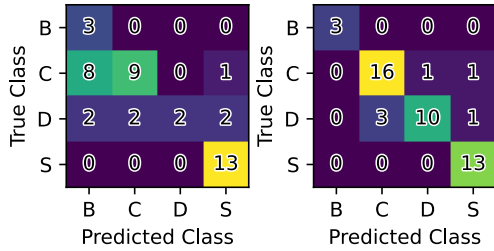


Figure 3: Confusion matrices from classifiers trained using the vanilla bootstrap algorithm (left) and our proposed variant (right).

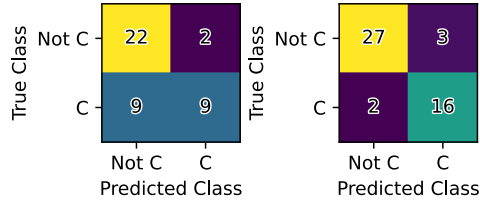


Figure 4: 2-way disambiguation results. Confusion matrices from vanilla bootstrapping (left) and our proposed variant (right).

same value across both systems. Thus many numerals which are technically ambiguous between these *systems* nonetheless have unambiguous *values*.<sup>5</sup> In settings where the absolute value of a numeral is all that matters, it is therefore usually sufficient to distinguish these systems from C without distinguishing them from one another. In this two-way setting, our model achieves 0.90 F1, versus 0.69 F1 from the vanilla baseline (Figure 4).

We emphasize that our test set only contains numerals which domain experts were able to disambiguate based on manual inspection. Easy cases may therefore be over-represented, and we expect our results to be an upper bound on the classifier’s accuracy across the whole corpus. Despite this, the results are strong enough to suggest that a majority of the ambiguous numerals in the corpus can be disambiguated with some certainty.

## 5 Analysis

In this section, we investigate how the features from our bootstrap classifier relate to known or hypothesized properties of the script.

A number of signs have been suggested as indicating types of livestock (sheep, goats, etc.); these include M346, M362, M367, and M417 (Dahl,

<sup>5</sup>Here we assume that the *absolute* value of N01 is the same across all three systems, and not just its value relative to the other digits. Current understandings suggest that this is the case, but the undeciphered nature of the script means this is technically not certain.

2005). In our analysis, all of these signs also predict the decimal system, which suggests that they were typically used to count flock sizes.

M376 has been suggested as either a "high-status human" (Dahl, 2005, p. 95) or livestock (Kelley, 2018, p. 165); it is strongly predicative of the sexigesimal system in our analysis. Other objects associated with this system include M056~f (a plow), M219, M371 and M296 (very speculatively, syllables used to write personal names), M059, M145, and M365 (“owners”, possibly persons or institutions to whom these accounts belonged), and M269~c (a vessel?). This may cast some doubt on the livestock reading for M376, in favor of the high-status human reading which is a more natural fit among the owner signs and possible personal names in this collection.

We note one text, P008212, which may exhibit a consistent ratio related to M376. P008212 alternates between entries ending in M288 and entries ending in M376 or M367~i. The magnitude of the numeral in an M288 entry is always exactly 4 times as large as the magnitude of an adjacent M376 entry, or 2 times as large as an adjacent M367~i entry. This pattern is very unlikely to be due to chance, as it holds across 44 total entries. Whatever the meaning of M376, on the basis of this text we can assert that it is associated with amounts of M288 that are exactly twice as large as those associated with M367~i. To our knowledge this ratio has not yet been noted in previous publications.

After disambiguating every numeral in the corpus to the most likely system according to our bootstrap classifier, we measure the average magnitude of counts associated with each feature. We observe that certain features accompany significantly higher or lower counts than others. Entries ending in M288 have the largest capacity magnitudes on average, while those ending in M263 are among the smallest. Both signs have been speculated to represent containers; from our results one might further speculate that M263 is a container of smaller dimension, or one that was never dealt with in bulk quantities. Entries ending in M297~b also accompany unusually large capacity measures, though the visually-similar M297 does not stand out as unusually large or small. This may point towards these signs having distinct meanings or uses despite sharing a visual resemblance.

The numerical systems of proto-Elamite have been proposed to have functional uses relating to

cultural practices in 3rd millennium south-western Iran. For instance, the capacity system (C) is suggested to be used for counting rations disbursed to households or workers (Kelley, 2018, pgs. 153-155). Among the recipients of these rations are M388 and M124, parallel “worker categories” which may represent the heads of work teams. We find that entries beginning in M388 accompany significantly larger capacity measures on average than those in M124, which may point towards M388 individuals heading teams of larger sizes or comprising workers of higher status.

## 6 Related Work

Naik et al. (2019) demonstrate that word embedding models fail to capture magnitude and numeration (i.e. the equivalence between 3 and *three*), and suggest the need for dedicated representations of numerals in NLP models. Sundararaman et al. (2020) follow up with DICE embeddings designed to explicitly capture both magnitude and numeration, and demonstrate improved results on numeracy tests introduced by Wallace et al. (2019). Spithourakis and Riedel (2018) describe a GMM-based approach to numeral embedding for language models, which also incorporates explicit representations of magnitude. These models assume that the magnitudes in question are known and must simply be encoded; they do not consider the task of *determining* magnitude from ambiguous notations.

While introducing a benchmark to test LM numeracy, Shi et al. (2022) note that numeral representations can vary across scripts; however, they assume a setting where the conversion to Hindu-Arabic notation is straightforward, and do not discuss ambiguities which may result from this conversion.

One approach to handling numeric values in word problems is to replace them with variables  $v_1, v_2, \dots$ , generate the solution as an equation in terms of these variables, and substitute the original values back to obtain a concrete solution. Wu et al. (2021) note that the choice of equation can sometimes depend on whether the original quantities were absolute values or percentages, and therefore this replacement can introduce ambiguities which make some problems unsolvable. They introduce a magnitude-aware encoding for digit sequences, and describe a numerical properties prediction mechanism which estimates whether a numeral is an integer, fraction, percentage, etc. This mirrors our

attempt to predict an underlying number system.

Berg-Kirkpatrick and Spokoyny (2020) investigate the task of predicting a numeric value given surrounding text as context. They find that models which implicitly separate a value’s mantissa from its exponent achieve better results than those which predict the value directly, and that context from large pretrained text encoders is useful even when the pretraining task was not focused on promoting numeracy. As we are dealing with an undeciphered corpus, our models are unfortunately unable to rely on pretrained embeddings for context.

Friberg (1978) is responsible for early analyses of proto-Elamite and proto-cuneiform (a related script which is also partially deciphered) which helped to establish the relative values of the digits in these corpora. Nissen et al. (1994) perform what is possibly the earliest computer-assisted analysis of bookkeeping practices in proto-cuneiform, while Damerow and Englund (1989) and Englund (2011) discuss accounting practices in proto-cuneiform and proto-Elamite and the relationships between the two.

## 7 Conclusion

We have automated the process of extracting candidate readings for ancient proto-Elamite numeral notations, and have described ambiguities in the original script which make this extraction challenging. We present two approaches for disambiguating these ambiguous notations: one exploits a common structural property of proto-Elamite accounts to look for unambiguous summations, and the other exploits the few unambiguous notations to train a bootstrap classifier. We create a test set for the disambiguation task by manually disambiguating a subset of the corpus, and we describe a novel variant of cautious rule selection which significantly improves bootstrapping performance on this test set. As a result of this work, we are able to assign confident values to a majority of the numeral notations in proto-Elamite, to identify and correct a number of transliteration errors in the proto-Elamite corpus, and to shed new light on existing hypotheses about the meanings of some signs in this partially-deciphered script. As the proto-Elamite script was fundamentally an accounting technology, we believe that this work represents a crucial step towards deepening our understanding of this ancient corpus.



## Limitations

As PE remains largely undeciphered, our results can only be evaluated on the small subset of numerals which we have manually disambiguated. As noted in the paper, these may be easier than the rest of the corpus, meaning our evaluation can only give an upper bound on model performance.

We use a feature-based classifier to give interpretable results which can more easily be shared and discussed with non-technical experts in Assyriology. A limitation of this approach is that model performance depends on the choice of input features, and features which are effective can sometimes seem arbitrary. We attempt to justify the features used by our model by explaining in Table 2 which aspects of the script each is intended to capture.

Lastly, PE numerals are just one aspect of a complex and multifarious decipherment problem. Our results alone cannot paint a complete picture of this script, and must be interpreted in relation to results from outside of computer science.

## References

- Steven P. Abney. 2002. [Bootstrapping](#). In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 360–367. ACL.
- Taylor Berg-Kirkpatrick and Daniel Spokoynny. 2020. [An empirical investigation of contextualized number prediction](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4754–4764, Online. Association for Computational Linguistics.
- Michael Collins and Yoram Singer. 1999. [Unsupervised models for named entity classification](#). In *1999 Joint SIGDAT Conference on Empirical Methods in Natural Language Processing and Very Large Corpora*.
- 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.
- Robert K. Englund. 2001. Grain accounting practices in archaic Mesopotamia. In J. Høyrup and Peter Damerow, editors, *Changing Views on Ancient Near Eastern Mathematics*, pages 1–35.
- Robert K. Englund. 2004. [The state of decipherment of proto-Elamite](#). *The First Writing: Script Invention as History and Process*, pages 100–149.
- Robert K. Englund. 2011. Accounting in proto-cuneiform. In K. Radner and E. Robson, editors, *The Oxford Handbook of Cuneiform Culture*, pages 32–50.
- Jöran Friberg. 1978. *The Third Millennium Roots of Babylonian Mathematics I-II*. Göteborg Dept. of Mathematics, Chalmers University of Technology, Göteborg, Sweden.
- Gholamreza Haffari and Anoop Sarkar. 2007. [Analysis of semi-supervised learning with the Yarowsky algorithm](#). In *Uncertainty in Artificial Intelligence (UAI 2007)*, pages 159 – 166. AUAI Press. Conference in Uncertainty in Artificial Intelligence 2007, UAI 2007 ; Conference date: 19-07-2007 Through 22-07-2007.
- 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.
- Piero Meriggi. 1971. *La scrittura proto-elamica. Parte Ia: La scrittura e il contenuto dei testi*. Accademia Nazionale dei Lincei, Rome.
- Aakanksha Naik, Abhilasha Ravichander, Carolyn Rose, and Eduard Hovy. 2019. [Exploring numeracy in word embeddings](#). In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3374–3380, Florence, Italy. Association for Computational Linguistics.
- Hans J. Nissen, Peter Damerow, and Robert K. Englund. 1994. *Archaic Bookkeeping: Writing and Techniques of Economic Administration in the Ancient Near East*. University of Chicago Press.
- H.J. Nissen, P. Damerow, and R.K. Englund. 1993. *Archaic Bookkeeping: Early Writing and Techniques of Economic Administration in the Ancient Near East*. University of Chicago Press.
- Vincent Scheil. 1935. *Texte de comptabilité Proto-Élamite (Troisième Série)*, volume 26 of *Mémoires de la Mission Archéologique de Perse*. Paris: Librairie Ernest Leroux.
- Freda Shi, Mirac Suzgun, Markus Freitag, Xuezhi Wang, Suraj Srivats, Soroush Vosoughi, Hyung Won Chung, Yi Tay, Sebastian Ruder, Denny Zhou, Dipanjan Das, and Jason Wei. 2022. [Language models are multilingual chain-of-thought reasoners](#). *CoRR*, abs/2210.03057.
- Georgios Spithourakis and Sebastian Riedel. 2018. [Numeracy for language models: Evaluating and improving their ability to predict numbers](#). In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2104–2115, Melbourne, Australia. Association for Computational Linguistics.

Dhanasekar Sundararaman, Shijing Si, Vivek Subramanian, Guoyin Wang, Devamanyu Hazarika, and Lawrence Carin. 2020. [Methods for numeracy-preserving word embeddings](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4742–4753, Online. Association for Computational Linguistics.

Eric Wallace, Yizhong Wang, Sujian Li, Sameer Singh, and Matt Gardner. 2019. [Do NLP models know numbers? probing numeracy in embeddings](#). In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 5307–5315, Hong Kong, China. Association for Computational Linguistics.

Max Whitney and Anoop Sarkar. 2012. [Bootstrapping via graph propagation](#). In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 620–628, Jeju Island, Korea. Association for Computational Linguistics.

Qinzhuo Wu, Qi Zhang, Zhongyu Wei, and Xuanjing Huang. 2021. [Math word problem solving with explicit numerical values](#). In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5859–5869, Online. Association for Computational Linguistics.

David Yarowsky. 1995. [Unsupervised word sense disambiguation rivaling supervised methods](#). In *33rd Annual Meeting of the Association for Computational Linguistics*, pages 189–196, Cambridge, Massachusetts, USA. Association for Computational Linguistics.

## A Constructing a Test Set for Numeral Disambiguation

This appendix describes a set of ambiguous numerals which we believe can be confidently disambiguated. Together, these numerals comprise the test set which we use to evaluate automated disambiguation via bootstrapping.

### A.1 Capacity Measures

**P008014** Our subset-sum analysis (Section 3.2) reveals that the sole entry on the reverse of this tablet (P008014:18:num) exactly equals the sum of the entries on the obverse, provided the whole tablet is read in the capacity system. This is consistent with the fact that four of the entries on the text are unambiguously capacity measures (P008014:11:num, P008014:14:num, P008014:15:num, and P008014:18:num; the other entries are SDBC-ambiguous). Moreover, the apparent summary line has an M288 object, and the first entry on the obverse ends in M288. This suggests that the entire text may record amounts of M288 (which is strongly associated with the capacity system), but that this object has been left implicit in all but the first and last entries.

On the basis of this evidence, we disambiguate the seven ambiguous entries in this text (P008014:6:num, P008014:7:num, P008014:8:num, P008014:9:num, P008014:10:num, P008014:12:num, P008014:15:num) to the capacity system.

### A.2 Sexagesimal Measures

**P008173** Our subset-sum analysis reveals that the first entry on the reverse (P008173:13:num), which is an unambiguous sexagesimal notation counting  $7.5 \times N01^S$  instances of M376, exactly equals the sum of the M376 entries on the obverse (P008173:5:num, P008173:7, P008173:8) if these entries are read in the sexagesimal system. On the basis of this evidence we disambiguate these entries to the sexagesimal system.

**M056~f/M288 Texts** P008798 is exemplary of a set of two-entry texts of comparable physical dimensions (approx. 43mm x 31mm x 18mm) which count M056~f in the first entry and M288 in the second. In many of these texts, the first entry is unambiguously sexagesimal and the second is unambiguously capacity; in these cases the amount of M056~f is always exactly 2.5 times the amount of M288.

If P008797:6:num is read as a sexagesimal notation, then this text follows the same pattern as these other texts, and exhibits the expected 2.5:1 ratio of M056~f to M288. On the basis of this evidence we disambiguate this entry to the sexagesimal system.

By the same argument, we also disambiguate P008791:6, P008799:6, P008800:6, P008801:6, P008802:6, P008804:4 and P008810:7 to the sexagesimal system.

### A.3 Decimal Measures

**P008179** This tablet is broken: a fracture on the obverse renders one digit partially unreadable. The broken digit has been restored as 2(N23), but we propose that the correct reading should in fact be 1(N23).

To see that this should be the case, notice that the first entry on the reverse (P008179:14:num) unambiguously equals  $852 \times N01^D$ ; this is exactly the same sum obtained by reading the obverse entries in the decimal system, provided one uses our restoration instead of the current transliteration.

This reading is consistent with the fact that some entries (P008179:8:num, P008179:9:num, P008179:14:num) are already unambiguous decimal counts. Moreover, all but one of the entries on the obverse count the same object (M388), which further points towards their likely using the same number system. On the basis of this evidence, we disambiguate P008179:10 and P008179:11 to the decimal system.

The original restoration (2(N23)) presumably arose from the observation that the lacuna is wide enough to span two signs. The last visible sign before the lacuna is M388, which only occurs at the very end of entries elsewhere in this text, implying that all of the missing signs are likely digits. Our proposed reading requires that the N23 in the lacuna occupy a slightly wider space than would be typical, or perhaps that some of the lacuna be occupied by an erasure. If the original restoration is correct, there must instead be an arithmetic error in the summary line (which in this case differs from the true sum by 1(N23)).

**P008012** Following the claim that sheep and goats are counted decimally in proto-Elamite (Englund, 2011), the entirety of this text should be expected to use the decimal system. Our subset-sum analysis confirms that the entry on the reverse (P008012:16) equals the sum of the entries on the obverse when read in this system, though in this

case that would also be true for sexagesimal and bisexagesimal readings. Notably, the summation does *not* work out if the summary is read as a capacity measure. Thus, while we may confidently say that this text does not contain capacity measures, we can only tentatively assign it to the decimal system in particular.

**P008243** Kelley (2018) observes that this is a decimally-counted roster. We follow this author in disambiguating all ambiguous numerals in this text to the decimal system.

### A.4 Bisexagesimal Measures

**P009048** This text contains a large number of unambiguous bisexagesimal notations. Of these, P009048:16:num counts the same object (M352~h) as P009048:19, on the basis of which we propose that P009048:19 is also a bisexagesimal notation. Similarly, the unambiguous entries P009048:10 and P009048:15 count the same objects (M351+X and M354, respectively) as P009048:13 and P009048:7, respectively, on the basis of which we also disambiguate these entries to B.