Sign Clustering and Topic Extraction in Proto-Elamite*

Logan Born[1] loborn@sfu.ca
Kate Kelley[2] kathryn.kelley@ubc.ca
Nishant Kambhatla[1] nkambhat@sfu.ca
Carolyn Chen[1] nll-contact@sfu.ca
Anoop Sarkar[1] anoop@sfu.ca

[1]Simon Fraser University
School of Computing Science

[2]University of British Columbia
Department of Classical, Near Eastern,
and Religious Studies

Abstract

We describe a first attempt at using techniques from computational linguistics to analyze the undeciphered proto-Elamite script. Using hierarchical clustering, n-gram frequencies, and LDA topic models, we both replicate results obtained by manual decipherment and reveal previously-unobserved relationships between signs. This demonstrates the utility of these techniques as an aid to manual decipherment.

1 Introduction

In the late 19th century, excavations at the ancient city of Susa in southwestern Iran began to uncover clay tablets written in an unknown script later dubbed 'proto-Elamite'. Over 1,500 tablets have since been found at Susa, and a few hundred more at sites across Iran, making it the most widespread writing system of the late 4th and early 3rd millennia BC (circa 3100–2900 BC) and the largest corpus of ancient material in an undeciphered script.[1]

Proto-Elamite (PE) is the conventional designation of this script, whose language remains unknown but was presumed by early researchers as likely to be an early form of Elamite. A number of features of the PE writing system are understood. These include tablet format and direction of writing, the numeric systems, and the ideographic associations of some non-numeric signs, predominantly those for livestock accounting, agricultural production, and possibly labor administration. Yet the significance of the majority of PE signs, the nature of those signs (syllabic, logographic, ideographic, or other) and the linguistic context(s) of the texts remain unknown. It was recognized from the outset, due to the features of the script, that all the proto-Elamite tablets were administrative records, rather than historical or literary compositions (Scheil, 1905).

Texts are written in lines from right to left, but are rotated in publication to be read from top to bottom (then left to right) following academic practice for publishing the contemporary proto-cuneiform tablets. The content of a text is divided into entries, logical units which may span more than one physical line. The entry itself is a string of non-numeric signs whose meanings are for the most part undeciphered. Each entry is followed by a numeric notation in one of several different numeric systems, which quantifies something in relation to the preceding entry. This serves to mark the division between entries. An important exception exists in what are currently understood to be 'header' entries: these can present information that appears to pertain to the text as a whole, and are followed directly by the text's first content entry with no intervening numeric notation. A digital image and line drawing of a simple PE text along with transliteration are shown in Figure 1.

Although a complete digital corpus of PE texts exists (Section 2), it has not been studied using the standard toolkit of data exploration techniques from computational linguistics. The goals of this paper are threefold. By applying a variety of computational tools, we hope to

i. promote interest in and awareness of the problems surrounding PE decipherment

ii. demonstrate the effectiveness of computational approaches by reproducing results previously obtained by manual decipherment

*We would like to thank Jacob Dahl and the anonymous reviewers for their helpful remarks. 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 last author.

[1]New PE texts have been found as recently as 2006–2007, when excavations at Tepe Sofalin near Tehran uncovered ten tablets (Dahl et al., 2012).

Proc. of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, pp. 122–132
Minneapolis, MN, USA, June 7, 2019.

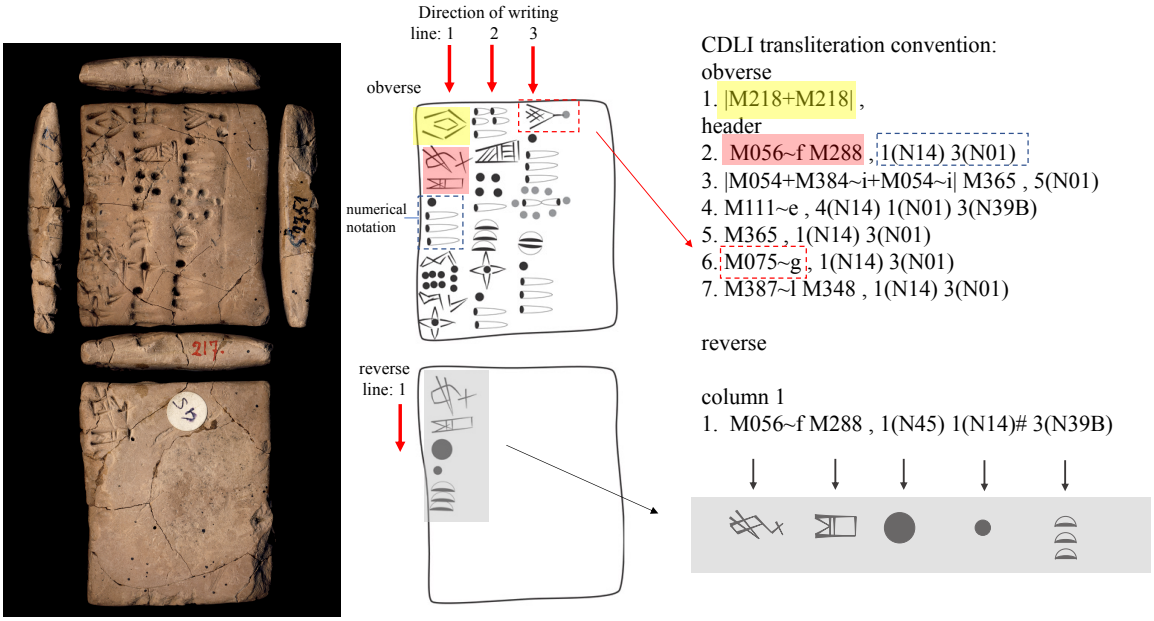

Figure 1: PE tablet *Mémoires de la Délégation en Perse* (MDP) 6, no. 217 (P008016; Scheil 1905). Digital image, line art, and transcription (called transliteration by the CDLI) from the Cuneiform Digital Library Initiative. Explanatory annotation added by current authors.

iii. highlight novel patterns in the data which may inform future decipherment attempts

We hope to show that interesting data may be extracted from the corpus even in the absence of a complete linguistic decipherment. To encourage further study in this vein, we are also releasing all data and code used in this work as part of an online suite of data exploration tools for PE.[2] Additional figures and interactive visualizations are also available as part of this toolkit.

2 Conventional Decipherment Efforts

Studies towards the decipherment of PE can be summarised by a relatively short bibliography of serious efforts (Englund, 1996).[3] Stumbling blocks to decipherment have included inaccuracies in published hand-copies of the texts, a lack of access to high-quality original images, and the associated difficulty in drawing up an accurate signlist and producing a consistently-rendered full transcription of the corpus. Members of the *Cuneiform Digital Library Initiative* (CDLI) have been remedying these deficiencies over the past two decades, and the PE script can now boast (i) a working signlist with a consistent manner of transcribing signs in ASCII, and (ii) an open-access, searchable database hosting the entire corpus in transcription, alongside digital images and/or hand-copies of almost every text.[4]

Historically, specialists of PE have operated on a working hypothesis that it may be, like later Sumerian cuneiform, to some extent a mixed system of ideographic or logographic signs alongside signs that may represent syllables. However, the level of linguistic content represented in both PE and proto-cuneiform has been called into question (Damerow, 2006), and the presence of a set of syllabic signs in PE is yet to be proven.

The strict linear organisation of signs in PE is the earliest such known to a writing system: proto-cuneiform arranged signs in various ways within cases (and sometimes subcases), and only in cuneiform from several hundred years later did scribes begin to consistently write in lines with one sign following the next. However, it is not clear to what extent the linear sign organization of PE reflects the flow of spoken language as in later writing systems.[5]

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

[3] For the most complete and up-to-date bibliography, see Dahl 2019.

[4] http://cdli.ox.ac.uk/wiki/doku.php?id=proto-elamite

[5] Dahl (2019:83): "proto-Elamite texts are organized in an

Analysis of sign and entry ordering in the texts has also revealed some tabular-like organising principles familiar from proto-cuneiform. Longer sequences of signs can often be broken down into constituent parts appearing to follow hierarchical ordering patterns apparently based upon administrative (rather than phonetic/linguistic) principles, and hierarchies can be seen across entries as well (Hawkins, 2015; Dahl et al., 2018).

Traditional linguistic decipherment efforts have not yet succeeded in identifying a linguistic context for PE, though progress has been made, for example in positing sets of syllabo-logographic signs thought to be used to write personal names (PNs). We refer to Meriggi's (1971:173–174) syllabary as shorthand for these signs, as he was the first to identify such a set and his work has since been closely imitated (Desset 2016; Dahl 2019:85). Although he called it a syllabary, Meriggi was aware that the signs might not prove to be syllabic and that object or other signs might remain mixed in.

Continued efforts to establish the organizational principles of the PE script and to isolate possible syllable sequences or PNs may be advanced by computational techniques, which can be used to evaluate hypotheses much faster than purely manual approaches. In this endeavour it is necessary to remember that although early writing encodes meaningful information, that information may or may not be linguistic (Damerow, 2006). Although it is not known why PE disappeared after a relatively short period of use, one of several possibilities is that this relates to the way it represents information, perhaps providing a poorer, less versatile encoding compared to later cuneiform with its mixed syllabo-logography.

3 Data

All data in this work are based on the PE corpus provided by the CDLI. After removing tablets which only bear unreadable or numeric signs, this dataset comprises 1399 distinct texts. Most of these are very short: the mean text length is 27 readable signs, of which only 10 are non-numeric on average. Long texts do exist, however, up to a maximum length of 724 readable signs of which 198 are non-numeric.

Our working signlist (extracted from the transcribed texts) contains 49 numeric signs and 1623 non-numeric signs. Of these, 287 are 'basic' signs, and 1087 are labeled as variants due to minor graphical differences. Sign variants are denoted by ∼, as in M006∼b, a variant of the basic sign M006. In an on-going process, analysis of the corpus aims to confirm whether sign variants are semantically distinct, or reflect purely graphical variation. Where the latter case is understood, the sign is given a numeric rather than alphabetic subscript, as in M269∼1. The remaining 249 non-numeric signs are compounds called complex graphemes which are made up of two or more signs in combination, as in |M136+M365|.

Future work is required to establish which sign variants are meaningfully distinct from their base signs; in the absence of such work, we have chosen to treat all variants as distinct until proven otherwise. Our models give interpretable results under this assumption, suggesting this is a reasonable approach. There are, however, cases where collapsing sign variants together would seem to affect our results, and we highlight these where relevant.

4 Analysis of Signs

4.1 Hierarchical Sign Clustering

Manual decipherment of PE has proceeded in part by identifying that some signs occur in largely the same contexts as other signs. This has produced groupings of signs into "owners", "objects", and other functionally related sets (Dahl, 2009). For example, M388 and M124 are known to be parallel "overseer" signs which appear in alternation with one another (Dahl et al., 2018:25).

In the same vein, we have investigated techniques for clustering signs hierarchically based on the way they occur and co-occur within texts. Our work considers three approaches to sign clustering: a neighbor-based clustering groups signs based on the number of times each other sign occurs immediately before or after that sign in the corpus; an HMM clustering groups signs based on the emission probabilities of a 10 state hidden Markov model (HMM) trained on the corpus; and a generalized Brown clustering groups signs as described in Derczynski and Chester 2016. By using three different clustering techniques, we can search for clusters which recur across all three methods to maximize the likelihood of finding those that are meaningful. This reduces the impact of noise in the data, which is especially useful

in-line structure that is more prone to language coding than proto-cuneiform..."

given the small size of the PE corpus.

4.1.1 Clustering Evaluation

We identified commonalities between our three clusterings using the following heuristic. Given a set of signs S, we found for each clustering the height of the smallest subtree containing every sign in S. If all of these subtrees were short (which we took to mean not larger than $2|S|$) then we called S a stable cluster.

In many cases, the stable clusters comprise variants of the same sign. This is the case for M157 and M157∼a, which cluster together across all techniques and are already believed to function similarly to each other, if not identically.

One very large stable cluster consists of the signs M057, M066, M096, M218, and M371. This cluster is shown as it appears in each clustering in Figure 2. These signs belong to Meriggi's proposed syllabary (Meriggi 1971, esp. pp. 173–4) and are hypothesized to represent names syllabically (or logographic-syllabically; Desset 2016:83). Desset (2016:83) likewise identified "approximately 200 different signs" from possible anthroponyms, "among which M4, M9, M66, M96, M218 and M371 must be noticed for their high frequency." Desset's list differs from our cluster by only two signs, replacing M057 with M004 and M009. M004 and M009 group with other members of the putative syllabary in each clustering, but their position is much more variable across the three techniques. For M009 at least, this may indicate multivalent use: besides its inclusion in hypothesised PNs (e.g. Meriggi 1971:173; Dahl 2019:85), it appears in various different administrative contexts that don't appear to include PNs (e.g P008206) and as an account postscript (see below here and 5.3).

All three methods group the five signs in our cluster close to other suspected syllabic signs; however, since each technique groups them with a *different* subset of the syllabary, only these five form a stable group across all three methods. This may be due simply to their frequency, or they could in fact form a distinct subgroup within the proposed syllabary; future work may yield a better understanding of possible anthroponyms by trying to identify other such subgroups.

While this discussion has focused on the stable clusters for which we can provide some interpretation, others represent groups of signs with no previously recognised relationship, such as

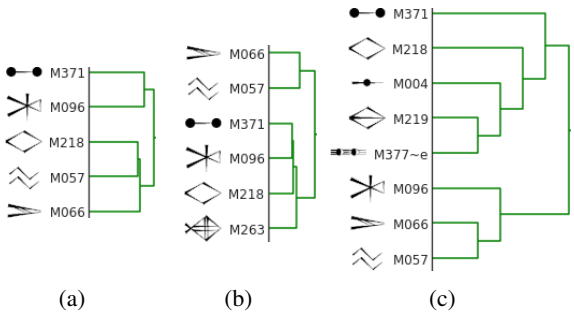

Figure 2: Detail of the (a) neighbor-based, (b) HMM, and (c) Brown clusterings showing signs possibly used in anthroponyms. M057, M066, M096, M218, and M371 are considered a stable cluster due to their proximity in all three clusterings.

M003∼b and M263∼a (Figure 3). M003(∼a/b) are "stick" signs (,) understood in some PE contexts to denote worker categories (Dahl et al., 2018); they are graphically comparable to proto-cuneiform PAP∼a-c () and PA (), the latter of which can, in later Sumerian, indicate *ugula*, a work group foreman/administrator.

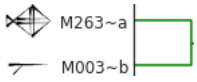

Figure 3: M003∼b clusters identically with M263∼a in all three techniques.

M263∼a is one of a series of depictions of "vessels" (), this particular variant appearing in 27 texts; notably the base sign M263 appears as a possible element in PNs (Dahl, 2019:85). Interestingly, M003∼b and M263∼a only appear together in a single text (P008727), one of a closely-related group of short texts[6] that each end in the administrative postscript M009 M003∼b or M009 M003∼c. It can also be noted that M263∼1 occurs in another text belonging to this small group.

It thus remains for future work to interpret this and the many other stable clusters resulting from our work. These additional groupings are detailed in our data exploration toolkit, along with complete dendrograms for each clustering which are too large to include in this publication.

Although we have not performed a full study of the clusterings produced when sign variants are collapsed together, a preliminary comparison

[6] Available online at https://cdli.ucla.edu/search/search_results.php?SearchMode=Text&requestFrom=Search&TextSearch=M009+M003

suggests this is worth pursuing. For instance, a new cluster of small livestock signs arises in the neighbor-based clustering, comprising M367 ("billy-goat"), M346 ("sheep"), M006 ("ram"), and M309 (possible animal byproduct). Existing clusters, such as the stable cluster of syllabic signs, appear to remain intact, but a complete comparison of the techniques in this setting is warranted.

4.2 Sign Frequency and n-Gram Counts

Sign frequency is another useful datapoint for understanding the overall content of the corpus and for building a more nuanced understanding of sign use (Dahl, 2002; Kelley, 2018). Figure 4 shows the most common PE unigrams, bigrams, and trigrams. These counts exclude n-grams containing numeric signs or broken or unreadable signs (transcribed as X or [...]); n-grams which span the boundary between entries are also excluded. Note the sharp drop-off in frequency from the most frequent signs to the rest of the signary; in fact nearly half the attested signs (745 out of 1623) occur only once. Similar results were presented in Dahl 2002.

The most common unigrams include "object" signs and signs belonging to Meriggi's syllabary. The object signs are M288 (a grain container), M388 ("person/man"), M124 (a person/worker category paralleling M388), M054 (a yoke, usually indicating a person/worker category or animal), M297 ("bread"), and M346 ("ewe"). The syllabary signs are M218, M371 (which may double as an object sign/worker category), M387 (also a numeral meaning "100"), and M066.

The n-gram counts reveal the scale at which complex sequences of information are repeated across tablets. Over 1600 strings contain at least 3 non-numeric signs. Of these, only 11 trigrams are repeated at least 5 times across the corpus; two of these end in the "grain container" sign M288 and are therefore best parsed as undeciphered bigrams followed by an object sign. Following this, 52 other trigrams are repeated three or four times across the corpus, leaving the great majority (98%) of trigrams to appear only once or twice.[7] The most frequent trigram, M377∼e M347 M371 (found 17 times per Figure 4), appears in no more than about 1.5% of the texts. Even among bigrams, the most common can only occur in up to 3.2% of texts.

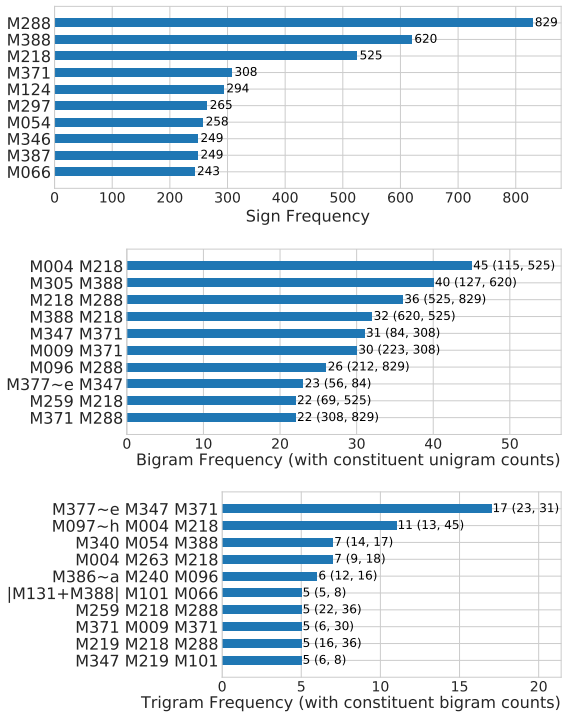

Figure 4: The 10 most frequent PE unigrams, bigrams, and trigrams (top to bottom). In parentheses are given the frequencies of the two unigrams comprising each bigram, and the two bigrams comprising each trigram: note that some frequent n grams are comprised of relatively infrequent $n-1$-grams.

External comparisons may help determine whether this is a meaningful degree of repetition, but such comparisons are not straightforward. Third millennium Sumerian or Akkadian accounting tablets are reasonable corpora to compare against, but these are available only in transliteration (using sign *readings*) while PE is transcribed (using sign *names*). This distinction makes n-gram counts from the two corpora incomparable without further work to transform the data.

Despite this, an impressionistic assessment of Ur III Sumerian administrative texts suggests that they are highly repetitious: information of wide importance to the administration (e.g. basic nouns, phrases describing administrative functions, month names, ruler names, etc.) occurs frequently. If one expects a similar pattern in the PE administrative record, our initial analysis suggests that trigrams (and perhaps bigrams) may not be a significant tactic for encoding these types of information, although unigrams might.

An n-gram analysis can also be used to be-

[7] This assumes that sign variants are meaningfully distinct, as is the working hypothesis among PE specialists. Collapsing variants together does not appreciably change these results, however, as it only increases most trigram counts by 1 or 2 instances. A similar result holds for bigram counts.

gin exploring the frequency of suspected anthroponyms within the PE corpus. Dahl (2019:85) lists frequently-attested signs (10 instances or more) with "proposed syllabic values" obtained through traditional graphotactical analysis; Figure 5 presents the frequency of the most common bigrams and trigrams limited to this signset. This list fails to include what is thought to be the most commonly attested PN, M377∼e M347 M371 mentioned above, since the middle sign, M347, is uncommon. Nonetheless the strings in this figure are more representative of possible PNs, since object signs which are understood to encode separate units of information have been weeded out. Overall we see that a small handful of 3-sign PNs are repeated at least 4 times across the corpus, but the majority appear 3 times or less. 2-sign PNs might be more frequent,[8] although some of the bigrams in the figure simply represent substrings from the trigrams. The ten most common bigrams all appear 13 or more times across the corpus, and the most frequent alone appears 45 times (M004 M218, including as part of a common trigram in Figure 5, accounting for 11 of its uses).

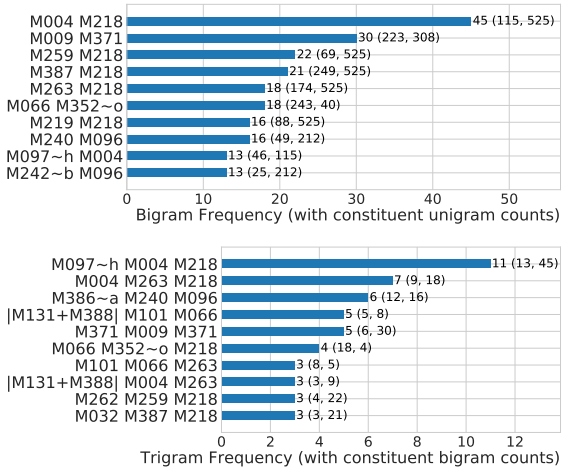

Figure 5: The 10 most frequent PE bigrams and trigrams (top to bottom), limited to signs in Dahl's (2019) syllabary. In parentheses are given the frequencies of the two unigrams comprising each bigram, and the two bigrams comprising each trigram.

Repeated n-grams, anthroponymic or otherwise, become increasingly rare for $n > 3$. No 4-gram or 5-gram appears more than 3 times; no 6-gram appears more than twice; and no 7-gram appears more than once. This low level of repetition indicates that common frequency-based linguistic decipherment methods may be ineffective on this corpus. We can, however, identify repeated strings which are similar to one another, if not exact copies, which may lead to insights about the function of certain PE signs and sign sequences. For example, the only two 6-grams which occur multiple times in the corpus differ from one another by only a single sign:

M305 M388 **M240** M097∼h M004 M218
M305 M388 **M146** M097∼h M004 M218

A further variant appears once in the corpus:

M305 M388 **M347** M097∼h M004 M218

Traditional graphotactical analysis parses the first of these strings as follows:

- Institution, household, or person class: M305
- Person class: M388
- Further designations of the individual: M240 M097∼h M004 M218

Side-by-side comparison of these 6-grams raises the question of whether the third sign in each sequence (M240, M146, and M347 respectively) is yet another classifier preceding a stable PN M097∼h M004 M218, or may reflect a PN pattern in which the first element (perhaps a logogram?) can alternate.

Although there are no repeated 7-grams or 8-grams, there are three pairs of 7-grams which differ by only a single sign, and one such pair of 8-grams. We hope that by exploring sign usage within such strings, future work will be able to identify new sign ordering principles and possibly reach a more controlled set of signs that may represent anthroponyms. Such a list would offer a better (if still slim) chance at linguistic decipherment. Our data exploration toolkit provides an interface for fuzzy string matching to facilitate further investigation of strings like these.

5 LDA Topic Model

Latent Dirichlet Allocation (LDA; Blei et al. 2003) is a topic modeling algorithm which attempts to group related words into topics and determine which topics are discussed in a given set of documents. Notably, LDA infers topical relationships solely based on rates of term co-occurrence, meaning it can run on undeciphered texts to yield information on which terms may be related. Note, however, that topics may be semantically broad, and

[8] However, according to Desset's (2016) traditional analysis of 515 hypothetical anthroponymic sequences,"250 (48.5 %) were made of 3 signs, 118 (22.9 %) of 4 signs, 83 (16.1 %) of 2 signs, 38 (7.3 %) of 5 signs, 15 (2.9 %) of 6 signs, 8 (1.5 %) of 7 signs and 3 (0.5 %) of 8 signs."

one must be careful not to infer too much about a sign's meaning simply from its appearance in a given topic. LDA differs from the other clustering techniques we have considered in that it also provides a means for grouping tablets based on the topics they discuss, which may reveal genres or other meaningful divisions of the corpus.

We induced a 10-topic LDA model over the PE corpus. We chose a small number of topics to make the task of interpreting the model more manageable; fewer topics make for fewer sets of representative signs to analyze. Furthermore, with 10 topics the model learns topics which are mostly non-overlapping (Figure 6), meaning there are few redundant topics to sort through. We note, however, that model perplexity drops sharply above 80 topics, and topic coherence peaks around 110 topics; future work may therefore do well to investigate larger models.

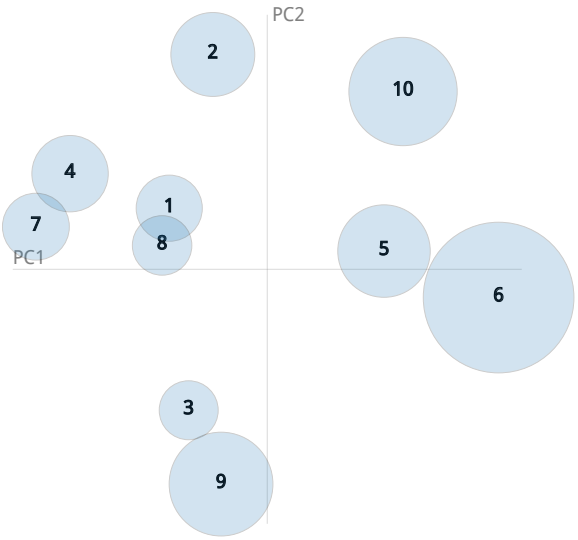

Figure 6: Intertopic distance (measured as Jensen-Shannon divergence) visualized with LDAVis (Sievert and Shirley, 2014) using two principal components (PC1 and PC2). Larger circles represent more common topics.

The following sections begin to elaborate on the topics which we can most easily interpret, although space constraints prohibit full analysis of each individual topic. Our data exploration toolkit provides additional details including information about topic stability using the stability measure introduced by Mäntylä et al. (2018).

5.1 Topic 1

The most representative signs for this topic are M376 and M056∼f. M376 has been speculated to represent either a human worker category or cattle; M056∼f is a depiction of a plow ( , comparable to the proto-cuneiform sign for plow, APIN ). This is an intriguing connection as a sign-set for bovines has not yet been identified in PE, despite the clear cultural importance of cattle suggested by PE cylinder seal depictions (Dahl, 2016). More interesting still is the fact that M376 and M056∼f never occur in the same text. Their inclusion in the same topic implies that they simply occur in the presence of similar signs (though not as direct neighbors of those signs, since they do not group together in the neighbor-based clustering). Topic modelling in this case has brought to light tendencies in the writing system that may have been intuitively grasped but would be difficult to quantify manually.

5.2 Topic 3

The signs M297∼b and M297 are both highly representative of this topic. This is interesting as the relationship between these two signs has been uncertain (Meriggi, 1971:74). M297∼b was hypothesised to indicate a "keg" by Friberg (1978). It is an "object" sign that almost always appears in the ultimate or penultimate position of sign strings; it sometimes appears in the summary line of accounts followed by numerical notations that quantify amounts of grain or liquids. Friberg suspected such texts referred to ale distributions. Ale is thought to have been a staple of the PE diet at Susa. Meriggi suggested M297 may indicate "bread", but he also included it in his syllabary; it is the 6th most common sign in PE, appearing in 145 texts, and M297∼b is the 31st most common appearing in 66 texts. Yet topic 3 is the dominant topic in only 85 texts, suggesting that the LDA model has identified a particular subset of the accounts that refer to M297 or M297∼b. Also of note is the fact that M297∼b occurs in topic 3 at a significantly higher rate than M297, despite being rarer in general—a much higher percentage of the overall uses of M297∼b appear in this topic (around 75%) than do the overall uses of M297 (less than 15%).

5.3 Topics 4 and 7

The texts included in topics 4 and 7 successfully reproduce aspects of Dahl 2005 with reference to the genres of PE livestock husbandry and slaughter texts. Dahl was able to decipher the ideographic meaning (if not phonetic realiza-

tion) of signs for female, male, young, and mature sheep and goats and some of their products, beginning with the key observation that proto-cunieform UDU ( , "mixed sheep and goats") is graphically comparable to M346 (). The most representative signs in topic 4 are M346 ("ewe") and M367 ("billy-goat").

While almost every instance of M346 is representative of topic 4, it is assigned to topic 5 in the atypical text P272825 (see 5.4). Several other typical livestock context uses of M346 belong to topic 7. Topic 7 was the most stable topic across 30 repeated runs in our topic stability evaluation. The most predictive sign for this topic is M009 (), which is also representative of topic 4 (and appeared in Section 4.1.1). The most representative texts in this topic include a few nanny-goat herding texts; many more texts in this topic have no known association with livestock or animal products, though a few (e.g. P009141 and P008407) do bear seal impressions depicting livestock.

5.4 Topic 5

The reason that the LDA model groups these 144 texts is not immediately apparent to the traditional PE specialist. An odd feature of the topic is that M388 ("person/man") is considered the most representative sign, but the most representative *text* is a simple tally of equids that never uses M388, and in fact uses few non-numerical signs overall. This may be due simply to noise in the model: M388 may be a kind of "stopword" which crops up in unrelated topics due to its high frequency. That said, an intriguing feature is that a significantly larger proportion of the texts in this topic bear a seal impression than do texts in the other topics. Seal impressions are unknown to the LDA model, and their presence suggests that it is at least possible the model has identified similarities in tablet content not easily observed through traditional analysis. The atypical "elite redistributive account" (Kelley, 2018:163) P272825, which is also sealed, is associated with this topic. This text has around 116 entries using complex sign-strings, fifteen of which include M388.

5.5 Topic 6

The ten most representative signs for topic 6 include the five of Meriggi's possible syllabic signs that grouped most stably in our clustering evaluation (see 4.1.1). Nine of the ten are also included in Meriggi's syllabary, excluding only M388, the second most representative sign in the topic. M388 has been key to the identification of possible PNs, since it tends to appear just before longer sign strings and, through a series of arguments drawing on cuneiform parallels, may function as a Personenkeil (a marker for human names; Damerow and Englund 1989; Kelley 2018:222 ff.). The texts of topic 6 are of diverse size and structure, but do tend to include many traditionally identifiable PNs.

5.6 Topic 10

This topic also confirms existing understanding of a PE administrative genre, namely that of "labor administration" (Damerow and Englund, 1989; Nissen et al., 1994). The most representative signs are the characteristic "worker category signs" described in the very long ration texts discussed by Dahl et al. (2018:24–23), and indeed all of those texts appear in this topic, in addition to a variety of other identifiable labor texts of somewhat different (but partially overlapping) content.

5.7 Remaining Topics (2, 8, and 9)

Initial assessments also suggest promising avenues of analysis for topics 2, 8, and 9. Topic 2 is heavily skewed towards M288 ("grain container"), the most common PE sign;[9] its third most representative sign (M391, possibly meaning "field") may suggest an agricultural management context for some texts in this topic. Topic 8 is strongly represented by |M195+M057|. This is an undeciphered complex grapheme, frequently occurring as a text's second sign after the "header" M157. In topic 9, the two most representative signs are M387 and M036 (possibly associated with rationing). Since the LDA model is not aware of the numeric notation between entries, it is interesting that the bisexagesimal numeric systems B# and B appear prominently in this topic, whether or not M036 (associated with those systems) appears: see particularly P009048 (the text most strongly associated with this topic) and P008619.

5.8 LDA Summary

The preceding sections confirm that the LDA model largely learns topics which traditional PE

[9] A remarkable 37.3% of the topic's probability mass is allocated to this sign, compared to just 2.5% for the second most predictive sign (M157, the "household" header sign). No other topic is so skewed: only topic 4 comes close, with 20.3% of its mass assigned to M346 ("sheep").

specialists recognise as meaningful. Our brief interpretations of the topics serve only to highlight the amount of potentially fruitful analysis that still remains to be done. It also remains to see what topics arise when sign variants are collapsed together: preliminary results suggest that topics resembling our topic 6 and topic 10 are still found, but new topics also appear which have no clear correlates in the model discussed in this paper.

6 Related Work

Meriggi (1971:173–174) conducted manual graphotactic analysis of PE (and later linear Elamite) texts, for example by noting the positions in which certain signs could appear in sign-strings. Dahl (2002) was the first to use basic computer-assisted data sorting to present information on sign frequencies, and Englund (2004:129–138) concluded his discussion of "the state of decipherment" by suggesting that the newly transliterated corpus would benefit from more intensive study of sign ordering phenomena. Apart from the use of Rapidminer[10] to perform simple data sorting in Kelley 2018, no publications have yet described any effort to apply computational approaches to the dataset.

Computational approaches to decipherment (Knight and Yamada, 1999; Knight et al., 2006), which resemble the setup typically followed by human archaeological-decipherment experts (Robinson, 2009), have been useful in several real world tasks. Snyder et al. (2010) propose an automatic decipherment technique that further improves existing methods by incorporating cognate identification and lexicon induction. When applied to Ugaritic, the model is able to correctly map 29 of 30 letters to their Hebrew counterparts. Reddy and Knight (2011) study the Voynich manuscript for its linguistic properties, and show that the letter sequences are generally more predictable than in natural languages. Following this, Hauer and Kondrak (2016) treat the text in the Voynich manuscript as anagrammed substitution ciphers, and their experiments suggest, arguably, that Hebrew is the language of the document. Hierarchical clustering has previously been used by Knight et al. (2011) to aid in the decipherment of the Copiale cipher, where it was able to identify meaningful groups such as word boundary markers as well as signs which correspond to the same plaintext symbol.

Homburg and Chiarcos (2016) report preliminary results on automatic word segmentation for Akkadian cuneiform using rule-based, dictionary based, and data-driven statistical techniques. Pagé-Perron et al. (2017) furnish an analysis of Sumerian text including morphology, parts-of-speech (POS) tagging, syntactic parsing, and machine translation using a parallel corpus. Although Sumerian and Akkadian are both geographically and chronologically close to PE, these corpora are very large (e.g. 1.5 million lines for Sumerian), and are presented in word level transliterations rather than sign-by-sign transcriptions. This makes most of these techniques inapplicable to PE. Our study is more similar in spirit to Reddy and Knight (2011), as the Voynich manuscript and PE are both undeciphered and resource-poor, making analysis especially difficult.

7 Conclusions

We have shown that methods from computational linguistics can offer valuable insights into the proto-Elamite script, and can substantially improve the toolkit available to the PE specialist. Hierarchical sign clustering replicates previous work by rediscovering meaningful groups of signs, and suggests avenues for future work by revealing similarities between yet-undeciphered signs. Analysis of n-gram frequencies highlights the level of repetition of sign strings across the corpus as a point of further research interest, and also reveals sets of similar strings worth examining in detail. LDA topic modelling has replicated previous work in identifying known text genres, but has also suggested new relationships between tablets which can be explored using more traditional analysis. The methods we have used are by no means exhaustive, and there remain many more approaches to consider in future work. Particularly in a field populated by a small handful of researchers, the faster data processing and ease of visualization offered by computational methods may significantly aid progress towards understanding this writing system. We hope that our data exploration tools will help facilitate future discoveries, which may eventually lead to a more complete decipherment of the largest undeciphered corpus from the ancient world.

[10] https://www.rapidminer.com/

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022.

Jacob L. Dahl. 2002. Proto-Elamite sign frequencies. *Cuneiform Digital Library Bulletin*, 2002/1.

Jacob L. Dahl. 2005. Animal husbandry in Susa during the proto-Elamite period. *Studi Micenei ed Egeo-Anatolici*, 47:81–134.

Jacob L. Dahl. 2009. Early writing in Iran, a reappraisal. *Iran*, 47(1):23–31.

Jacob L. Dahl. 2016. The production and storage of food in early Iran. In M.B. D'Anna, C. Jauß, and J.C. Johnson, editors, *Food and Urbanisation. Material and Textual Perspectives on Alimentary Practice in Early Mesopotamia*, volume 37, pages 45–50. Gangemi Editore.

Jacob L. Dahl. 2019. Tablettes et fragments proto-élamites / proto-Elamite tablets and fragments. *Textes Cunéiform Tomes XXXII Musée de Louvre*.

Jacob L. Dahl, Laura Hawkins, and Kate Kelley. 2018. Labor administration in proto-Elamite Iran. In Agnès Garcia-Ventura, editor, *What's in a Name? Terminology related to the Work Force and Job Categories in the Ancient Near East*, pages 15–44. Alt Orient und Altes Testament 440. Ugarit Verlag: Münster.

Jacob L. Dahl, M. Hessari, and R. Yousefi Zoshk. 2012. The proto-Elamite tablets from Tape Sofalin. *Iranian Journal of Archaeological Studies*, 2(1):57–73.

Peter Damerow. 2006. The origins of writing as a problem of historical epistemology. *Cuneiform Digital Library Journal*, 2006/1.

Peter Damerow and Robert K. Englund. 1989. *The Proto-Elamite Texts from Tepe Yahya*. Bulletin (American School of Prehistoric Research). Peabody Museum of Archaeology and Ethnology, Harvard University.

Leon Derczynski and Sean Chester. 2016. Generalised Brown clustering and roll-up feature generation. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*, AAAI 2016, pages 1533–1539. AAAI Press.

François Desset. 2016. Proto-Elamite writing in Iran. *Archéo-nil. Revue de la société pour l'étude des cultures prépharaoniques de la valée du Nil*, 26:67–104.

Robert K. Englund. 1996. The proto-elamite script. In Peter Daniels and William Bright, editors, *The world's writing systems*. Oxford University Press, Oxford, UK.

Robert K. Englund. 2004. The state of decipherment of proto-Elamite. *The First Writing: Script Invention as History and Process*, pages 100–149.

Jöran Friberg. 1978. *The Third Millennium Roots of Babylonian Mathematics I-II*. Göteborg Dept. of Mathematics, Chalmers University of Technology, Göteborg, Sweden.

Bradley Hauer and Grzegorz Kondrak. 2016. Decoding anagrammed texts written in an unknown language and script. *Transactions of the Association for Computational Linguistics*, 4:75–86.

Laura F. Hawkins. 2015. A new edition of the Proto-Elamite text MDP 17, 112. *Cuneiform Digital Library Journal*, 1.

Timo Homburg and Christian Chiarcos. 2016. Akkadian word segmentation. In *Tenth International Conference on Language Resource Evaluation (LREC 2016)*, pages 4067–4074.

Kate Kelley. 2018. *Gender, Age, and Labour Organization in the Earliest Texts from Mesopotamia and Iran (c. 3300–2900 BC).* Doctoral dissertation, University of Oxford.

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.

Kevin Knight and Kenji Yamada. 1999. A computational approach to deciphering unknown scripts. In *Proceedings of the ACL Workshop on Unsupervised Learning in Natural Language Processing*.

Mika Mäntylä, Maelick Claes, and Umar Farooq. 2018. Measuring LDA topic stability from clusters of replicated runs. In *Proceedings of the 12th ACM/IEEE International Symposium on Empirical Software Engineering and Measurement*, ESEM '18, New York, NY, USA. ACM.

Piero Meriggi. 1971. *La scrittura proto-elamica. Parte Ia: La scrittura e il contenuto dei testi*. Accademia Nazionale dei Lincei, Rome.

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.

Émilie Pagé-Perron, Maria Sukhareva, Ilya Khait, and Christian Chiarcos. 2017. Machine translation and automated analysis of the Sumerian language. In

Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature, ACL, pages 10–16. Association for Computational Linguistics.

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.

A. Robinson. 2009. *Lost Languages: The Enigma of the World's Undeciphered Scripts*. Thames & Hudson.

Jean-Vincent Scheil. 1905. Documents archaïques en écriture proto-élamite. *Mémoires de la Délégation en Perse*, 6:57–128.

Carson Sievert and Kenneth Shirley. 2014. LDAvis: A method for visualizing and interpreting topics. In *Proceedings of the Workshop on Interactive Language Learning, Visualization, and Interfaces*, pages 63–70. Association for Computational Linguistics.

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.