DHQ: Digital Humanities Quarterly
2018
Volume 12 Number 1

Computational Models for Analyzing Data Collected from Reconstructed Cuneiform Syllabaries

Laura F. Hawkins <laura_hawkins_at_fas_dot_harvard_dot_edu>, Harvard University

Abstract

This study used three interdependent techniques to help understand the use and distribution of syllabic values of the cuneiform signs during the second half of the third millennium and early second millennium BCE. The results suggest that, during this period, cuneiform syllabaries were variable. That variation can further inform us about the regional, temporal, and dialectical contexts in which they existed. The addition of this research to the wider literature on the early adaptation of cuneiform enhances the field's understanding of how cuneiform syllabic values began to emerge and spread across the ancient Near East, and demonstrates how computational methods of analysis can be applied to research questions in humanities subjects.

1. Background

Cuneiform is the earliest writing system attested in history, emerging towards the end of the fourth millennium B.C. in a region of ancient Mesopotamia that corresponds with the southern part of modern-day Iraq. The population first associated with cuneiform spoke an isolate language called Sumerian. However, Semitic-speakers who co-inhabited the same region quickly adapted the script to write their own language, Akkadian. Akkadian is the earliest attested language of the Semitic family, which includes better-known languages such as Arabic, Hebrew, and Aramaic. 1

The signs that made up this writing system consisted of four types: logograms (sometimes called ideograms), which represented whole words or ideas; syllabograms, which represented whole syllables such as *tu, ta,* or *ti*; determinatives, which were a set of signs that indicated, in limited circumstances, the category a word belonged to (i.e., gods, trees, precious stones, personal names, etc.); and numerical signs, which belonged to either a decimal or sexagesimal counting system (both systems were used simultaneously in ancient Mesopotamia). Each sign in these categories was crafted by impressing one or more wedge-shaped markings with the aid of a stylus, usually made from a reed cut to have a triangular tip, into the still soft surface of clay tablets (Figure 1). These clay tablets naturally dried and hardened in the arid and hot climate, and as a result they survived in great numbers. This script survived for over 3000 years from around 3300 BC to 70 AD and became the hallmark of Mesopotamian history and culture. 2

The spread of Mesopotamian culture from southern Iraq into neighboring regions, including northern Iraq, Turkey, and Syria, coincided with the development of the cuneiform writing system, during the late fourth millennium and first half of the third millennium BC. During the next few hundred years, this writing system was adopted by the populations outside of Mesopotamia that were interacting regularly with Sumer and Babylonia primarily through trade. As the cuneiform system developed and spread, it was adapted to express non-Sumerian words, including personal names. In general, Akkadian and other languages adapted the cuneiform script by means of the same rebus principle used by Sumerian cuneiform to express grammatical affixes [Cooper 1996, 45]. 3

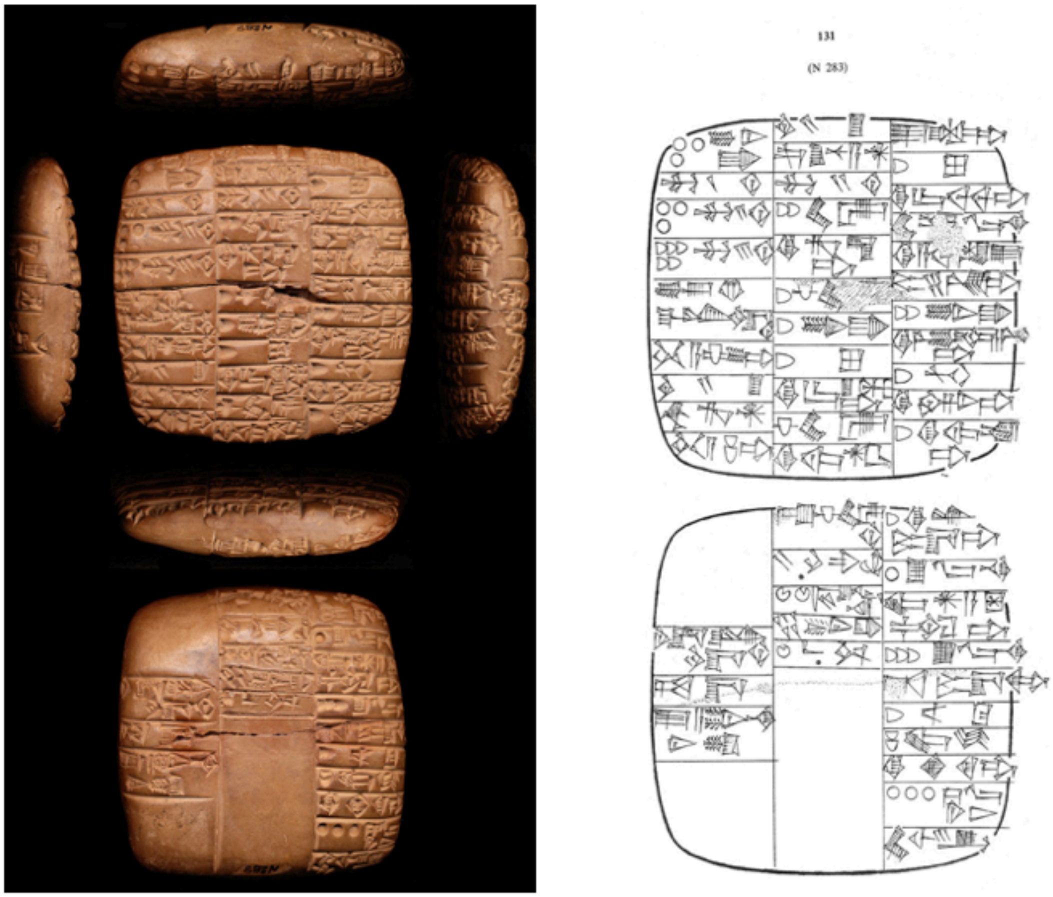

Figure 1. An example of a cuneiform tablet (OSP 1, 131; left) and the line drawing of that tablet (right). Image courtesy of the Cuneiform Digital Library Initiative (http://cdli.ucla.edu).

2. Introduction

In this study[1], interdisciplinary techniques were used to answer questions about how the cuneiform script, the earliest-attested writing system used between 3300 BCE and 100 CE in the ancient Near East, was adapted by Semitic-speaking peoples across Mesopotamia and Syria. The earliest, but scarce, evidence of cuneiform signs being used syllabically to write Semitic words and proper nouns comes from around 2600-2500 BCE. Between 2350 and 1800 BCE, there is an increase in the development and use of signs with syllabic values across Mesopotamia and Syria. These syllabic values (together called "syllabaries") continue to develop until standardizations of cuneiform signs and their values begin to be enforced around 1800 BCE, which essentially ends any major variability in the script within specific periods. This provides us with a period of almost 600 years, spanning the second half of the third millennium and early second millennium BCE, during which there is a wealth of textual data documenting the first full adaptation of the cuneiform script to syllabically write Semitic words and proper nouns. I investigated differences in the values of cuneiform signs used to write Semitic words excavated from nine sites that produced cuneiform texts during the late third millennium and early second millennium BCE in order to understand the extent to which any variation occurred. This analysis of the variation could then inform other questions about dialect differences, educational practices, and power dynamics across the region. 4

Individual studies of the signs attested with a syllabic value among the selected Syrian corpora form the basis of this investigation. These studies aim to provide a clear, consistent, and complete description of syllabic value attestations in Syria and Mesopotamia, and the information they provide is concatenated into a table of reconstructed syllabaries for each site. Together, these reconstructed syllabaries form a pan-Mesopotamian dataset informed by the most current knowledge of syllabic values attested at these sites during the roughly 600-year period being examined. 5

After curating a dataset of the syllabic values attested at each site, I analyzed the data using three complementary 6 computational methods: *phylogenetic estimation*, *hierarchical clustering*, and *principal component analysis* (PCA). These tools organized the data into visual hierarchies and identified the principal drivers behind the variation observed in the data. The results strongly support the conclusion that geography and time are the most significant factors affecting syllabic value observations across Mesopotamia, which indicated that during this period scribal communities adapted the cuneiform script's syllabic system independently and continued to use their local syllabaries despite the differences that any individual scribe encountered through his interactions with other communities across the region. The results of the computational analyses then suggested directions of further inquiry: I examined the linguistic environments of the syllabic values that most directly influenced the variation found in the dataset[2]. This component of the research provided evidence for greater dialectical variation across the geographic region than was previously assumed.

In summary, this research uses a series of three interdependent techniques to determine and understand the use and distribution of syllabic values within the cuneiform writing system during the second half of the third millennium BC and early second millennium BC. The results suggest that during this period cuneiform syllabaries are variable, and that variation can further inform us about the regional, temporal, and dialectical contexts in which they existed. The addition of this research to the wider literature on the early adaptation of cuneiform enhances the field's understanding of how cuneiform syllabic values began to develop and emerge across the ancient Near East, and demonstrates how computational methods of analysis can be applied to research questions in humanities subjects. 7

3. Sites Examined

The aim of this study was to examine the spread and adaptation of syllabic signs used to write Semitic words and proper nouns, so the sites initially considered for examination must have produced texts wholly or partially written in a Semitic language or dialect. In the third millennium, there were 19 such sites, and ten of these have been chosen for inclusion in this study due to the number of relevant tablets attested at each site and easy access (either electronic or physical) to the tablets: Ebla, Mari, Nabada, Tuttul, Adab, Kish, Tutub, Eshnunna, Assur, and Gasur[3]. Sites were excluded from this study if the entire collection of tablets is housed in Syria, Iraq, or Turkey, or if I was not granted access to the collections housed in museums Europe. 8

By examining these texts, the syllabaries attested at each of these sites were reconstructed (except Ebla, see below) by examining the published photos or transliterations of the texts of each site, and by examining a few of the tablets in person that have not yet been sufficiently published. I have therefore relied on a combination of analog data, digital data,[4] and in-person examination of texts for the collation of the syllabaries of the sites included in this study. 9

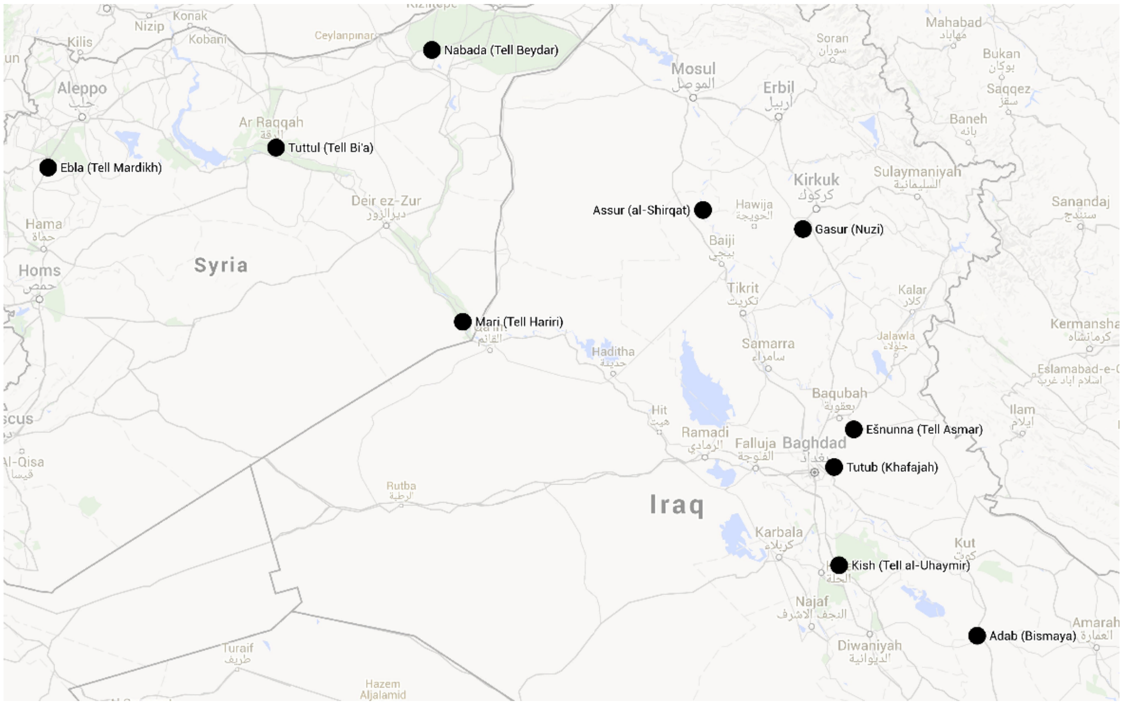

Figure 2. The ten sites included in this study are Ebla, Mari, Nabada, Tuttul, Adab, Kish, Eshnunna, Tutub, Assur, and Gasur.

The reconstructed syllabaries of Adab, Kish, Assur, and Gasur were collected using texts published on the Cuneiform Digital Library Initiative's database. The syllabary from Tutub was reconstructed using texts available both in the Cuneiform Digital Library Initiative (CDLI) database and Sommerfeld's (1999) "Die Texte der Akkade-Zeit. 1. das Dijala-Gebiet: Tutub." The data from Eshnunna was collected from texts published in Gelb's (1952) "Sargonic Texts from the Diyala Region" (MAD 1, nos. 1-195, 270-336.) and from the CDLI database. Many of these tablets are fragmentary and retain few discernible lines of text. 10

For the Tuttul and Nabada corpora, I relied on the works of Krebernik [Krebernik 2001] and Ismail [Ismail et al. 1996] respectively, as well as well-structured and collated digital data on the CDLI for both. Because of the size of the Mari and Ebla corpora, limitations had to be imposed on the type of texts examined in the reconstruction of the syllabaries of these two sites. For Mari, I have chosen to include only the texts published by Limet [Limet 1976] and for Ebla I only consider the already-published syllabaries of the lexical texts [Krebernik 1982/3, 178–236] [Conti 1990, 3–60] and the texts published on the Ebla Digital Archives.[5] 11

The table below outlines the sites that are examined in this study. It includes the periods examined, the number of texts used for the collection of the data, and the genres[6] attested among the texts included (plus the numbers within each genre in parentheses). For all sites examined, the majority of the texts are administrative in genre, with a small number being epistolary, literary, or lexical texts. 12

Site	Region	Period(s)	No. of Texts	Genres (no.)
Ebla	Syria	Old Akkadian[7]	ca. 7000	Lexical[8]; Administrative Letter
Mari	Syria	Ur III / Shakkanakku[9]	463	Administrative (463)
Nabada	Syria	Old Akkadian	223	Administrative (222); Legal (1)
Tuttul	Syria	Early Old Babylonian[10]	54	Administrative (51); Letter (2); Uncertain (1)
Adab	S. Mes. [11]	Old Akkadian, Ur III	1946, 130[12]	Administrative (1854, 102)[13]; Letter (27, 1); Royal/monumental (25, 21); Legal (22, 3); Uncertain (16, 0); Lexical (1, 0); Mathematical (1, 0); School (1, 2)
Eshnunna	S. Mes.	Old Akkadian	261	Administrative; Uncertain (26); School (8); Letter (6)l Literary (1)
Kish	S. Mes.	Old Akkadian	80	Administrative (68); Letter (5); Royal/monumental (3); Votive (2); Lexical (1); Literary (1)
Tutub	S. Mes.	Old Akkadian	73	Administrative (65); Royal/monumental (7); Legal (1)
Assur	N. Mes. [14]	Old Akkadian	20	Royal/monumental (7); Lexical (6); Administrative (4); School (3)
Gasur	N. Mes.	Old Akkadian	220	Administrative (190); Lexical (15); Letter (9); Legal (2); School (2); Mathematical (1); Uncertain (1)

Table 1. The ten sites examined for this study.

4. Methodology

4.1 Computational methods of analysis: a three-step approach

13 Three different methods of analysis were used to visualize this dataset. These three methods are phylogenetic estimation, hierarchical clustering, and principal component analysis. Phylogenetic estimation was included because it is a common method used in the study of language and manuscript evolution [Platnick and Cameron 1977] [Atkinson and Gray 2005] and the results could therefore be more easily compared with previous studies. However, because of the novelty of using phylogenetic systematics to understand writing system evolution, and because of the relative inflexibility of phylogenetic programs to filter out relevant data from noise, two other computational methods, programmed in R within the IDE RStudio, were used to test the results of the phylogenetic estimation model. All three types of analysis were conducted on an Apple laptop. The second two techniques help manipulate, filter, and visualize the data in different ways; they organize the data according to similarities and differences, and can isolate key data points that influence the results. These three methods will be discussed further below.

4.2 Unfiltered and filtered datasets

14 An initial examination of the dataset reveals that there are two particular aspects of the data that are skewing the preliminary results: the lack of sufficient data from Assur, and the presence of a large number of signs that only occur at Ebla. The insufficient data from Assur is due to my lack of access to museums in which the corpus is housed. Ebla’s larger number of syllabic sign values may be due to the different methods of determining syllabic values used by me and by the team that published the Ebla syllabary, to dialectical variation, or due to the larger number of non-administrative texts. In order, therefore, to avoid these issues and to avoid possible human error in the syllabaries I reconstructed, all hapax values, or syllabic values at each site that are attested at only one site, were removed in the filtered dataset.

15 Table 2 below shows the number of syllabic values at each site that occur at at least two sites; Assur is a clear outlier with only 6 of these signs attested. Because there is not enough data for Assur to be informative, this site will be removed from further analysis.

Ebla	Mari	Nabada	Tuttul	Adab	Eshnunna	Kish	Tutub	**Assur**	Gasur
128	100	70	105	78	60	105	108	**6**	90

Table 2. The number of syllabic values at each site that occur at at least one other site.

16 Table 3 shows the number of hapax syllabic values. The number of hapax syllabic values at Ebla is much higher than at the other sites examined, and were subsequently filtered out of the dataset for the secondary analysis.

Ebla	Mari	Nabada	Tuttul	Adab	Eshnunna	Kish	Tutub	Assur	Gasur
34	10	4	11	9	1	6	5	0	5

Table 3. The number of hapax syllabic values, or syllabic values that occur at only one site, attested at each site.

17 The three methods of analysis described below were applied to both the initial unfiltered dataset – which includes hapax syllabic values (or signs that occur at only one site), ubiquitous syllabic values (or syllabic values that occur at all sites), and the syllabic values from Assur – as well as the filtered dataset which excludes the previous three features of the data. The results of both the unfiltered and filtered datasets will be presented for each method below.

5. Phylogenetic Estimation

18 An analysis based on phylogenetic estimation[15] can scientifically test our hypotheses about the nature of the adaptation of cuneiform to write Semitic language(s) across Mesopotamia and Syria. The primary strength of phylogenetic analysis is its ability to reconstruct tree- or network-like relationships across time; because the spread and adaptation of cuneiform must have necessarily occurred over a period of time – even a relatively short one – these methods can provide interesting insights into the nature of the script’s spread and help us determine which sites cluster or diverge. Phylogenetic methods are advantageous because the data input and methods used are always transparent, and therefore the results should be repeatable.[16]

19 Phylogenetic methods, which were originally developed by evolutionary biologists for the analysis of trait inheritance and gene expression in particular phyla or species over time, have been used with increasing frequency in the field of linguistics for at least twenty years [Nichols and Warnow 2008, 760]. These methods have been employed in the analysis and historical reconstruction of Indo-European [Gray 2003], African [Marten 2006, 43–55], and Semitic [Kitchen et al. 2009] languages families; in the reconstruction of language and dialect relationships [Nakhleh et al. 2005a]; and to reconstruct manuscript evolution [Barbrook et al. 1998]. Although many of these studies have produced promising results, this application of phylogenetic inference techniques to reconstruct writing system evolution is still relatively new and untested [Skelton 2008]. The application of phylogenetic methods to this type of data is therefore unique, and the optimal search criteria and program settings have not yet been established. In this study, I used an optimization criterion called maximum parsimony (described further below).

5.1 Experimental Method

20 Phylogenetic analyses can employ a number of different methods of searching for and evaluating phylogenetic trees [Swofford et al. 1996, 478–93]. These methods tend to be either algorithm-based or optimality-criterion-based. Algorithm-based methods rely on algorithms to search for trees and to determine which tree is the correct one, and have the advantage of short computation time. Optimality-criterion-based methods, on the other hand, use different criteria for determining which tree is the best — called the optimality criterion — to find the tree in the first place — or the search strategy. This method is advantageous because the use of two different criteria for searching for the trees

and for determining which one is best makes it easier to determine the likelihood that the tree produced is the correct tree [Swofford et al. 1996, 408–9]. Optimality-criterion-based methods were used for this analysis because the relatively small dataset (compared to datasets of millions of data points in genomic studies) does not make computation time a concern.

Based on her work on Linear B, Skelton determined that the most appropriate optimality criterion for a phylogenetic analysis of a writing system is maximum parsimony, and I therefore consider it here as well [Skelton 2008, 170]. Maximum parsimony has been used in the study of several language families, including the Bantu language family on a variety of datasets, with [Bastin 1983], [Holden 2002], and [Holden et al. 2005] using only lexical data, and Rexova et al. (2006) using grammatical data. Others have used lexical data [Nakhleh et al. 2005a] [Rexovà et al. 2003] and a combination of types of data [Nakleh et al. 2005b] to study Indo-European languages. Maximum parsimony has also been used by Cysouw et al. to study Mixe-Zoquean [Cysouw et al. 2006]. 21

Maximum parsimony is an optimization problem that aims to produce a tree in which the minimum number of character state changes occurs; using this optimization problem follows the assumption that the path of least resistance would be not to adapt or create new syllabic values. It uses a simple model of character state change, which assumes that each change is equally likely to occur as any other change. While this assumption may not be correct for any given dataset, it is usually not possible to estimate actual probabilities of character change, and so an assumption of equal probabilities is necessary and appropriate; other optimality criteria, such as maximum likelihood, require explicit models of evolutionary change, which is not possible in this case. 22

Maximum parsimony analyses can allow characters with missing entries; for example, maximum parsimony encounters no problem when a language under study has no word for a given semantic slot. Since the data used here contains characters that are not always present in each taxon, this feature of maximum parsimony is necessary. 23

With maximum parsimony, the phylogenetic estimation program creates a series of possible trees and then assigns each one a “tree length,” which is the sum of the weights of the character state changes that occurred on all branches of the tree. Maximum parsimony considers the tree with the shortest tree length – or fewest character state changes – to be the best solution [Skelton 2008, 171].[17] 24

It is not uncommon for a maximum parsimony method to produce a number of trees with the same tree length. In this situation, it is possible to run an algorithm that produces a consensus tree. The algorithm examines the trees of equal tree length in order to determine which feature(s) the trees all share, or which a majority of the trees share. The usual features that are relevant in this determination are the splits, or bipartitions, on the leaf set induced by the edges of each tree; in other words, a consensus tree can be a tree that has exactly all the splits that each of the input trees have [Nichols and Warnow 2008, 770–3].[18] Since the phylogenetic estimation method produced only one tree in this case, this consensus method was not needed. 25

Another technique for estimating the support values (or the likelihood values) for a single tree, or for the branches of a given tree, is a statistical re-sampling technique called bootstrapping. Bootstrapping creates new, random datasets using characters from the original data matrix and runs them through the same set of parameters. This technique can be used simply to estimate support values for the edges of a tree, where the support values are the fraction of times that that particular bipartition appears in the random bootstrap trees; it can also be used as input data in a consensus method, like those described above, which would then be annotated with the support value estimates [Nichols and Warnow 2008, 773]. This technique dictates that a high support value for a particular bipartition increases the likelihood that that bipartition is accurate. This technique will be applied to the data for both the phylogenetic analysis and the hierarchical clustering. 26

5.2 Taxa

Taxa are the independent variables, or more basic entities being studied. In biology, species or gene sequences are often used as taxa; this study uses sites[19] as taxa. Using individual tablets as taxa is not an option because there is 27 not enough data on most tablets for this to be viable. Using scribal hands as taxa (as were used in [Skelton 2008]) is also not an option since hands have so far not been established, or even suggested, for these corpora [Biggs 1973, 39].

5.3 Characters

Characters are the dependent variables of a study. In biology, molecular or phenotype data are often used as characters; this study uses syllabic values of signs used in Semitic words and proper nouns as characters. The character states are either 1 or 0, indicating presence or absence (respectively) of that particular character in a taxon. 28

Using syllabic values as the characters in this study is not without its problems. While in most words or personal names it is clear what the syllabic value is meant to be, there are some cases where it is unclear. There is also the issue of human error, either in my own transliterations or on the transliterations and sign lists compiled by other scholars that have been relied upon. 29

5.4 Program and Settings

There are a number of programs available that perform phylogenetic analyses, each of which specializes in specific methods for estimating phylogenies. The program PAUP*, version 4.0a146 for Macintosh [Swofford 2001], specializes in parsimony methods, and so was used to analyze this dataset. Other programs such as TNT [Goloboff et al. 2003], Mesquite [Maddison and Maddison 2001], or PHYLIP [Felsenstein 2005] could also have been used to estimate the phylogeny of the data through maximum parsimony. 30

The data matrix that was imported into PAUP* consists of 326 total characters (or sign values); 116 of these are determined to be parsimony-uninformative, and 200 are parsimony-informative. Because of the relatively small dataset, an exhaustive search was used to generate the optimal tree (see Figure 3). A bootstrapping resampling method was then run using a full heuristic search strategy; the resulting consensus tree retained groups with a frequency of greater than 50 percent (see Figure 4). For both trees, no outgroup was defined so the trees are arbitrarily rooted at the first taxon (Ebla).[20] 31

5.5 Results

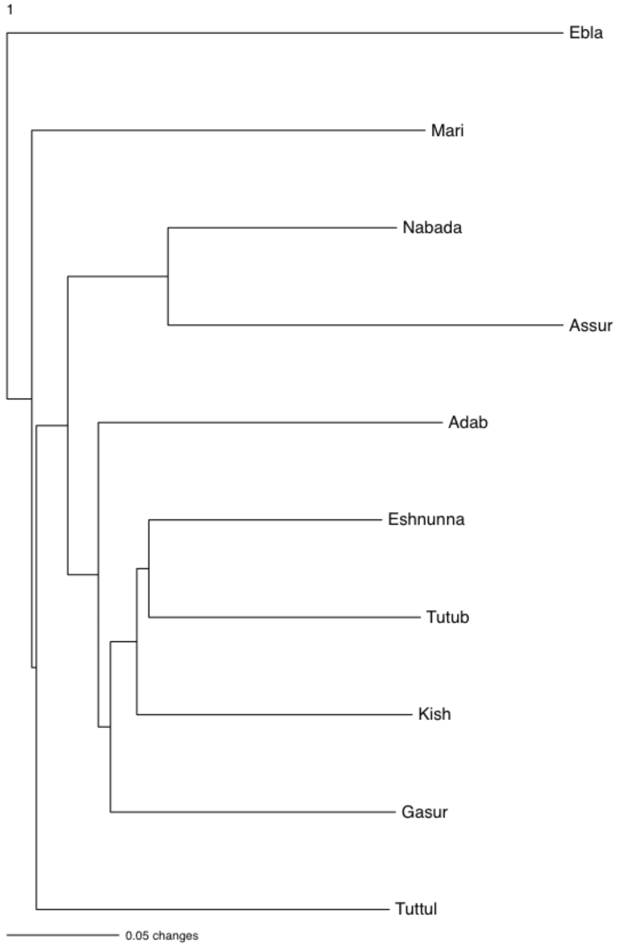

Figure 3. The first tree resulting from the phylogenetic estimation of the unfiltered dataset using maximum parsimony in the program PAUP*.

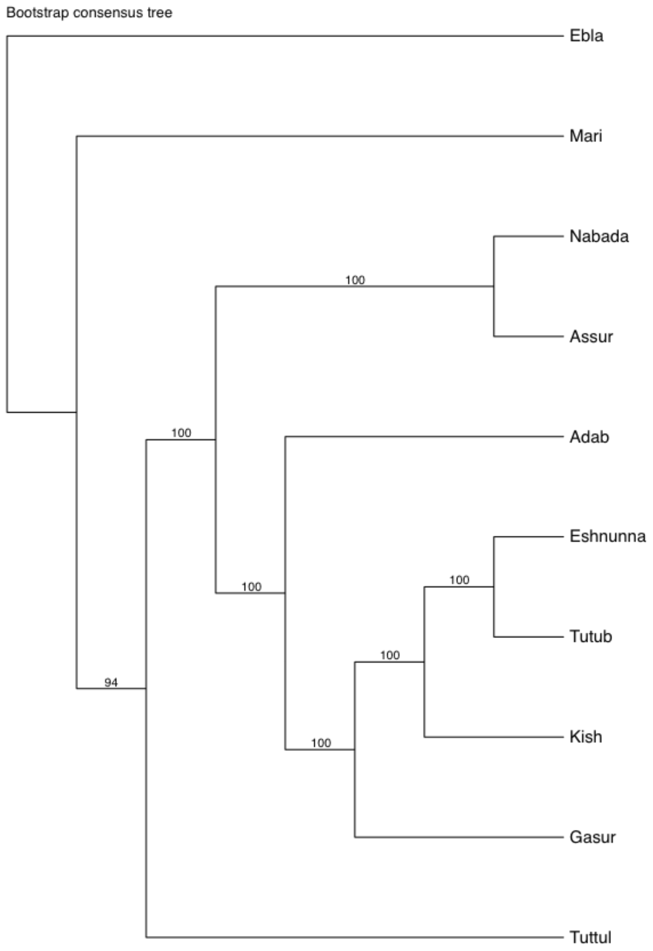

Figure 4. The consensus tree with p-values generated from a bootstrap resampling method on the unfiltered dataset.

The trees resulting from this analysis will represent similarities and differences – not the interdependence or genealogy – in the data, and could reflect three different realities and therefore support three different conclusions about the syllabaries. The possible results could be: 32

1. That the syllabaries exactly mirror the geography of the sites. This would indicate a relationship between the syllabaries that was based purely on geographic proximity of the sites. In other words, this tree would support the conclusion that there was an organic spread of the development and use of syllabic values from site to site.
2. That the trees mirror the geography of the sites to a certain extent. This would indicate that geography was perhaps one factor in how similar the syllabaries examined are. In other words, this tree would suggest that sites nearer to each other were more influenced by each other's syllabaries and sites further away from each other developed "genetic mutations" or independent changes in their syllabaries, but that this was not the sole influencing factor.
3. That the trees reflect geography in no way. This would indicate that the differences observed in the syllabaries must be attributed to another cause or causes.

The tree in Figure 4 shows the tree resulting from the phylogenetic analysis on the unfiltered dataset. The tree in Figure 5 shows the consensus tree with p-values created using a bootstrapping re-sampling technique, also on the unfiltered dataset. 33

The implications of these results would suggest either that these adaptations occurred either gradually across the region 34 as the technological innovation spread from site to site, or that there are other contributing factors, such as local scribal innovation (perhaps based on dialect differences between the sites).

A few preliminary observations can be made based on these results. First, it is notable that Ebla is an outlier among the taxa. The Mari and Tuttul syllabaries seem to be more similar to the rest of the cohort; Nabada and Assur are clustered together. Finally, the phylogenetic estimation groups together the syllabaries of Eshnunna, Kish, Tutub, and Gasur. Aside from the slightly closer relationship between the Mari and the Tuttul syllabaries, which could theoretically be explained by the geographic proximity of these two sites, these results are unexpected and not immediately explainable. Upon closer examination, the grouping of Nabada and Assur together can be explained by the relatively low number of syllabic values attested at each site; although the reasons for this differ for each. As explained above (Section 4.2), insufficient data was available from Assur which impacted the number of syllabic sign values in the reconstructed syllabary. At Nabada, at the other hand, the scribes writing these tablets seem to have used significantly fewer syllabic signs compared to logographic signs. This is an interesting and unexpected result[21]. The other patterns observed in these results are more difficult to explain.

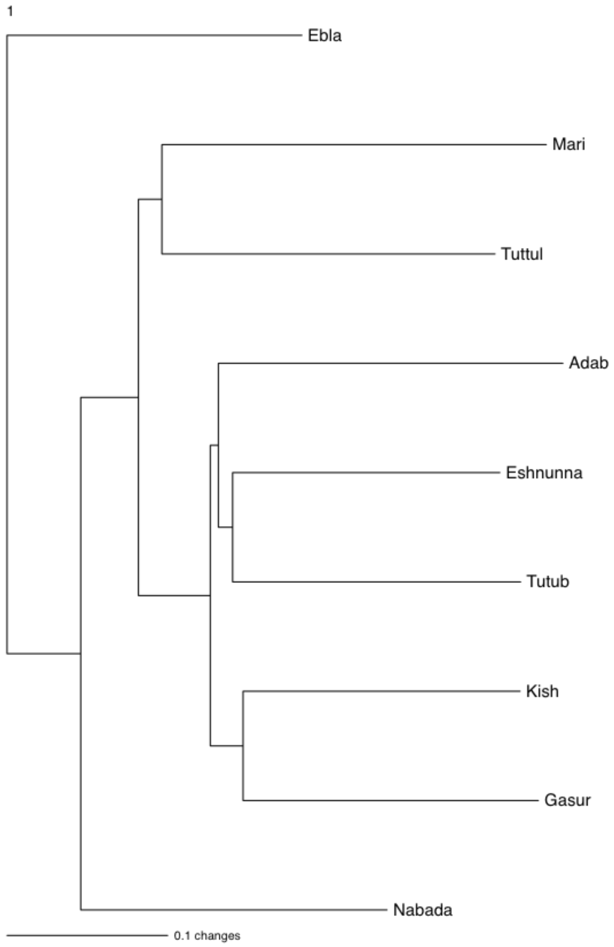

Figure 5. The tree resulting from the phylogenetic estimation of the filtered dataset using maximum parsimony in the program PAUP*.

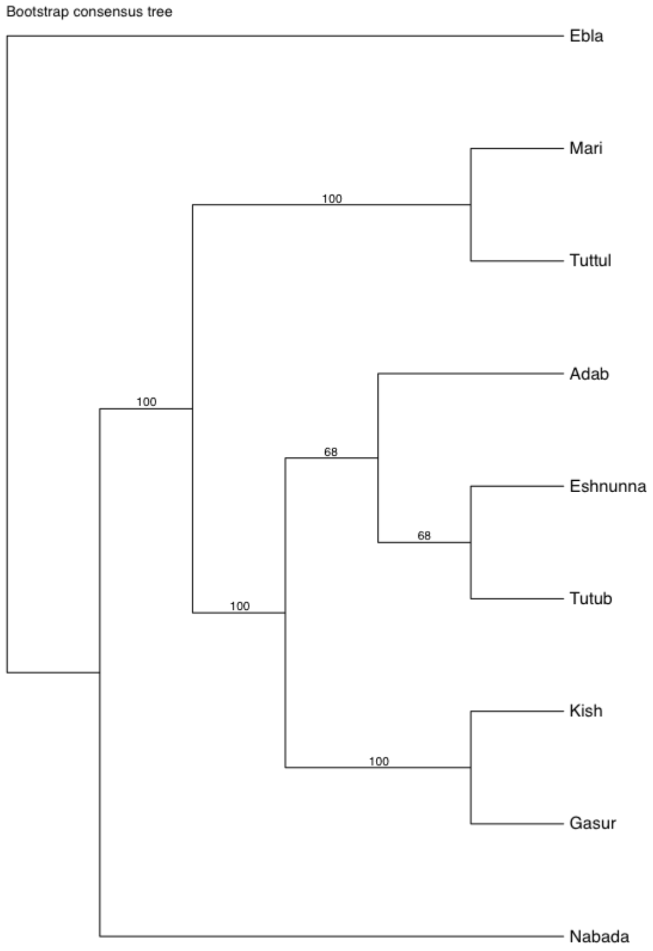

Figure 6. The consensus tree resulting from the bootstrapping resampling method on the filtered dataset.

Figures 5 and 6 show the results of the phylogenetic analysis and the bootstrapping re-sampling technique on the filtered dataset. Most interesting is that filtering hapax syllabic values and ubiquitous syllabic values did not affect the results. The bootstrap consensus tree of the filtered dataset reveals that Ebla and Nabada are most distant from the other sites, and closely similar to each other, and that Mari and Tuttul are similar with a p-value of 50%. According to this model, there are no strong associations between the Mesopotamian sites except for Kish and Gasur, whose sub-cluster has a p-value of 52% (see the section on hierarchical clustering below for an explanation of the relevance of p-values) 36

Based on this initial examination, it appears that the geography of the sites themselves is mirrored to a certain extent in the phylogenetic analysis. This supports one hypothesis that, if significant variation is attested within the syllabaries of these sites, that variation would be correlated with the geographic situation of the sites. In other words, local variation and innovation in syllabic sign values existed and permeated nearby cities. These localized sign values remained the most prominent feature of each site’s syllabary. In terms of practical relevance for modern scholars, this indicates that the composition of the syllabary of a corpus of cuneiform texts can determine where those tablets originated. 37

Clustering techniques such as phylogenetic analysis are useful for determining general tendencies within a dataset by finding the natural clusterings of that given dataset. However, since most clustering algorithms create clusters regardless of any inherent cluster structure in the dataset, other methods are required to externally validate the results. Hierarchical clustering and principal component analysis were therefore used to verify the results of the phylogenetic analysis and to determine other factors contributing to the observed variation in the syllabaries. 38

6. Hierarchival Clustering

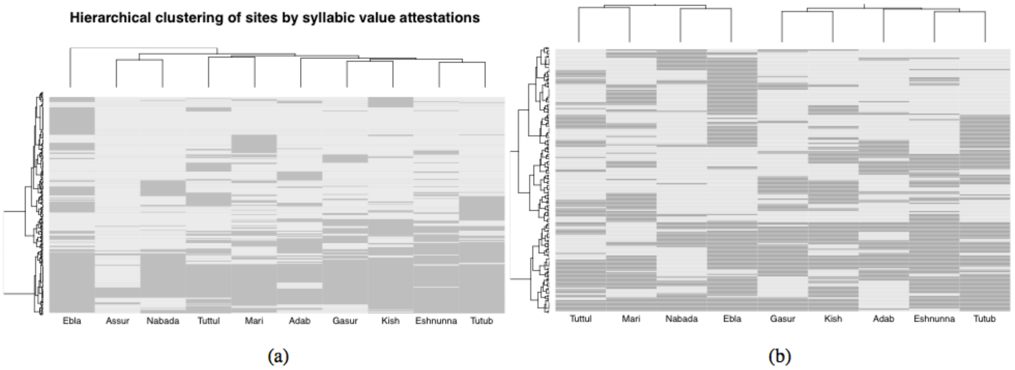

Figure 7. The hierarchical clustering on the unfiltered (a) and filtered (b) datasets show different results. The hierarchical clustering of the filtered dataset indicates a stronger connection between syllabary and geography that the hierarchical clustering of the unfiltered dataset.

39 Hierarchical clustering[22] is a visual clustering technique that organizes data into dendrograms; in this case, the technique was used to organize sites and syllabic values according to similarities and differences. Because the data collected and used in this study is straightforward, binary data, using different clustering algorithms does not produce different results. This method is in many ways similar to phylogenetic analysis, but uses different algorithms and relies on different underlying assumptions about the dataset. If these two techniques produce the same results, that is a good sign that the branchings and clusterings observed are present in the dataset and not a coincidence of the particular algorithm used.

40 Figures 7a and 7b show the results of the hierarchical clustering (using a Manhattan distance function and a Ward clustering metric) on the filtered and unfiltered datasets. The light grey indicates absence of a syllabic value at a site; dark grey indicates presence. The trees along the upper x-axes of the graphs shows how the sites cluster or diverge; the trees along the left-hand y-axes of the graphs shows the resulting clusterings of the syllabic values. It is the clusterings of the syllabic values that determine how the sites will cluster; because of the focus on how the syllabaries of these sites compare and relate to one another, the focus in this work is on the trees and clusterings of the sites themselves.

41 Compared to the results of the analysis on the unfiltered data, the results of the filtered data (to exclude Assur, any values that occur at only one site, or hapax syllabic values, and syllabic values that occur at all sites, or ubiquitous syllabic values) appear different. The results of this analysis indicate that Eshnunna, Adab, Gasur, Tutub and Kish are distinct from Nabada, Ebla, Mari, and Tuttul, the four sites in Syria. This suggests that there is a stronger geographically driven pattern in the data than the phylogenetic estimation originally seemed to indicate.

42 A bootstrap resampling technique was run on the results of the hierarchical clustering on the filtered dataset (Figure 8) using the package pvclust with a Manhattan distance method and a Ward D2 cluster method (the same methods used in the hierarchical clustering). This package provides two types of p-values: AU (Approximately Unbiased) p-value and BP (Bootstrap Probability) value. AU p-value, which is computed by multiscale bootstrap resampling, is a better approximation to unbiased p-value than BP value computed by normal bootstrap resampling [Suzuki and Shimodaira 2014, 4]. The clusters in the tree that group the Syrian syllabaries and the Mesopotamian syllabaries together are strongly supported by the data, having AU p-values of 93% and 94%, respectively.[23] The partitions grouping Mari and Tuttul together and Ebla and Nabada are supported by the data, though less strongly, with AU p-values of 83% and 86% respectively. The grouping of the Mesopotamian syllabaries is less strongly supported by the data (the Kish and Gasur cluster has an AU p-value of 69%; the Adab, Eshnunna, and Tutub cluster has an AU p-value of 71%; and the Eshnunna and Tutub cluster has an AU p-value of 73%). Overall, these support values suggest that the results of this

analysis are supported by the data, and suggest very close affinities between the Syrian syllabaries of Ebla, Nabada, Mari, and Tuttul. The data also strongly support a grouping of the Mesopotamian syllabaries.

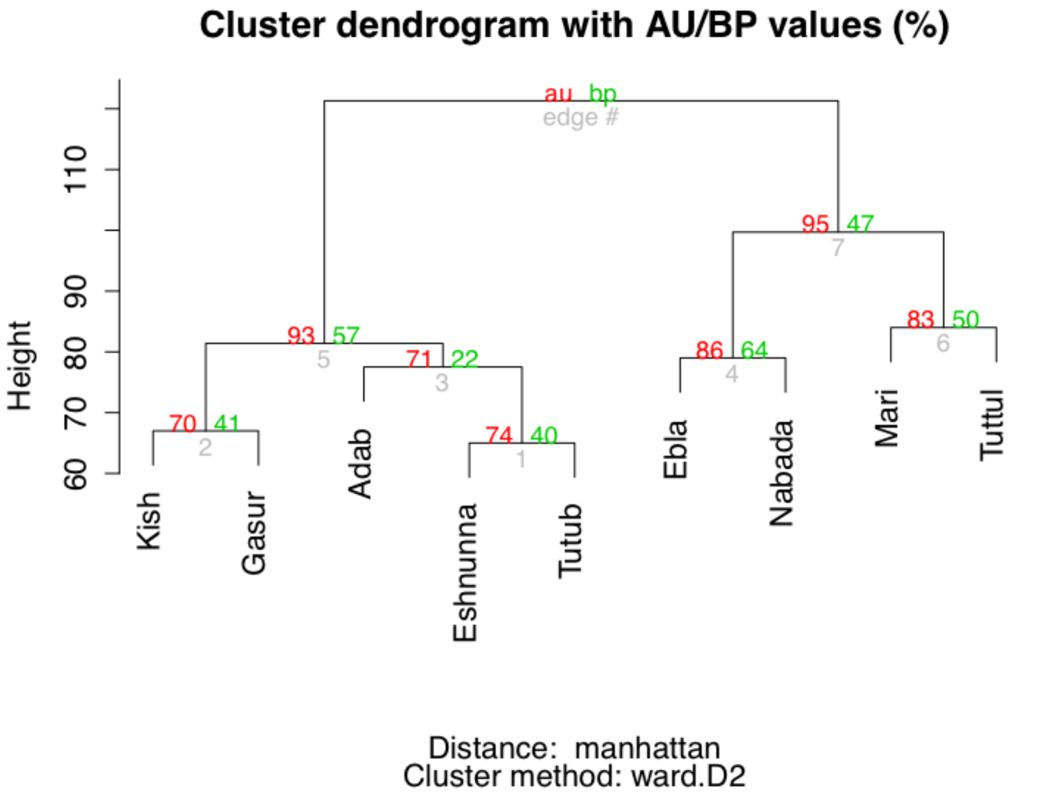

Figure 8. The AU and BP *p*-values support the results of the hierarchical clustering on the filtered dataset.

Principal component analysis can be used to visualize these patterns of clustering and to determine which syllabic values are driving these results. 43

7. Principal Component Analysis

Principal component analysis (PCA)[24] is a powerful yet simple tool for analyzing multivariate data [Wold et al. 1987, 37]. It uses linear algebra to take a very high-dimensional space (in this case, a 326-dimensional space because of the 326 data points, or syllabic values) and projects it into a two-dimensional space. An analysis using PCA always begins with a data table whose rows are termed “objects” and whose columns are termed “variables.” By analyzing the data points within this data table, one can accomplish one or several common goals: data reduction or simplification; data modeling; outlier detection; variable selection; classification; prediction; or unmixing [Wold et al. 1987, 38]. The primary goals of using PCA against the dataset in this study are to determine outliers and select variables. To address these goals, PCA rotates the axes of the data table in order to find the two axes that represent the most variation, or principal components, within the data. In other words, PCA finds the set of variables that explain the most variance found in the dataset. These variables, or syllabic values in this case, can then be further examined using complementary techniques if desired. 44

7.1 PCA on the unfiltered dataset: the number of syllabic values attested at each site is driving the observed variation

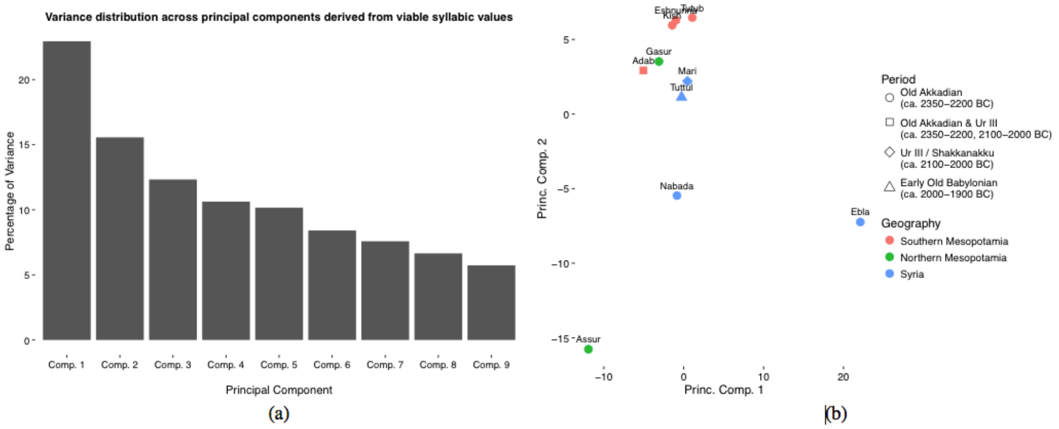

Figure 9. (a) The variance distribution across principal components derived from the unfiltered dataset shows that 25% of the total variation can be attributed to the first principal component. (b) The first principal component, which can be attributed to the number of syllabic values attested at each site, is driving the observed variation: the large number of hapax syllabic values attested at Ebla (the first principal component) and the lack of sufficient data at Assur (the second principal component) are driving the observed variation.

The principal component analysis on the unfiltered dataset (Figure 9a) shows that approximately 25% of the variation in this dataset can be attributed to the first principal component; this component accounts for significantly more variation than the subsequent principal components, which account for between 15% and 5% of variation in the dataset. By plotting the first principal component against the second principal component (Figure 9b), it becomes clear that the first principal component can be characterised by Ebla and Assur being outliers on either end of the spectrum. This suggests that the primary factor driving the results of the principal component analysis on the unfiltered dataset is the number of syllabic values attested at each site within the dataset; further, the large number of hapax syllabic values attested at Ebla and the lack of sufficient data at Assur are the significant factors influencing the first principal component.

7.2 PCA on the filtered dataset: geographic, temporal, and unknown variation drive the observed variation

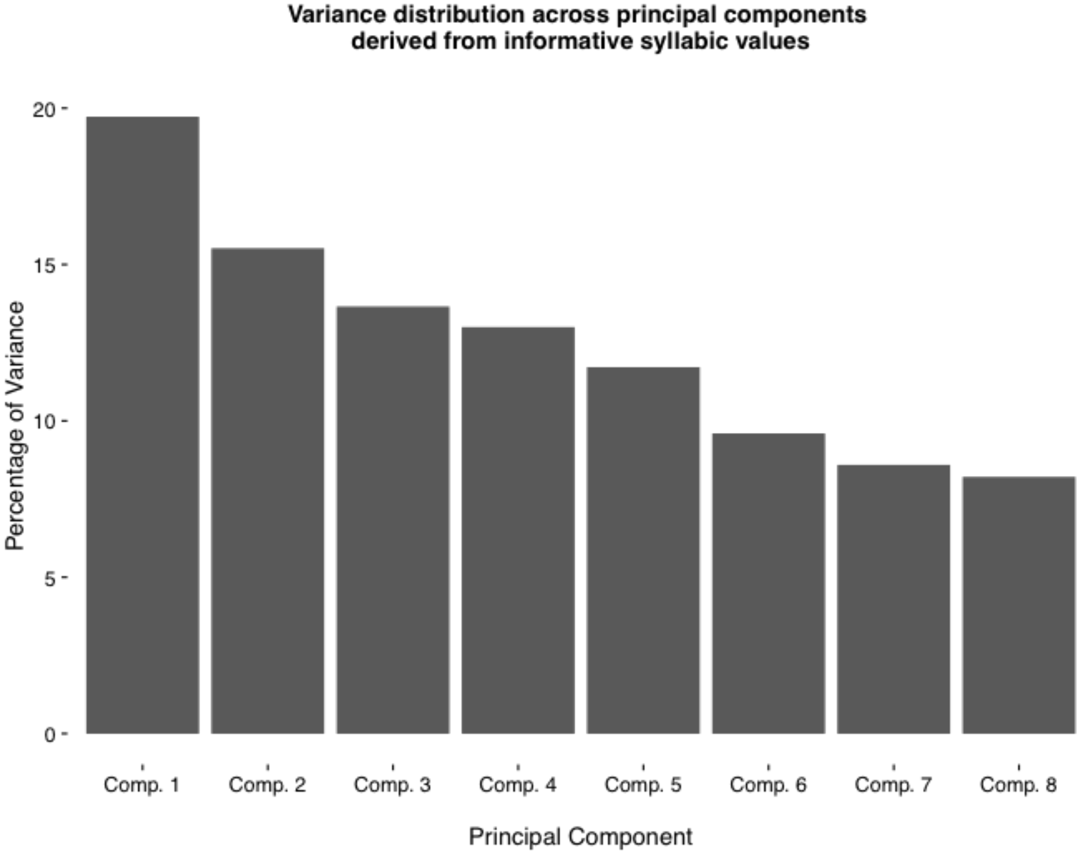

Figure 10. The variance distribution across principal components derived from the filtered dataset shows that when excluding hapax syllabic values, ubiquitous syllabic values, and the data from Assur the data is more complex.

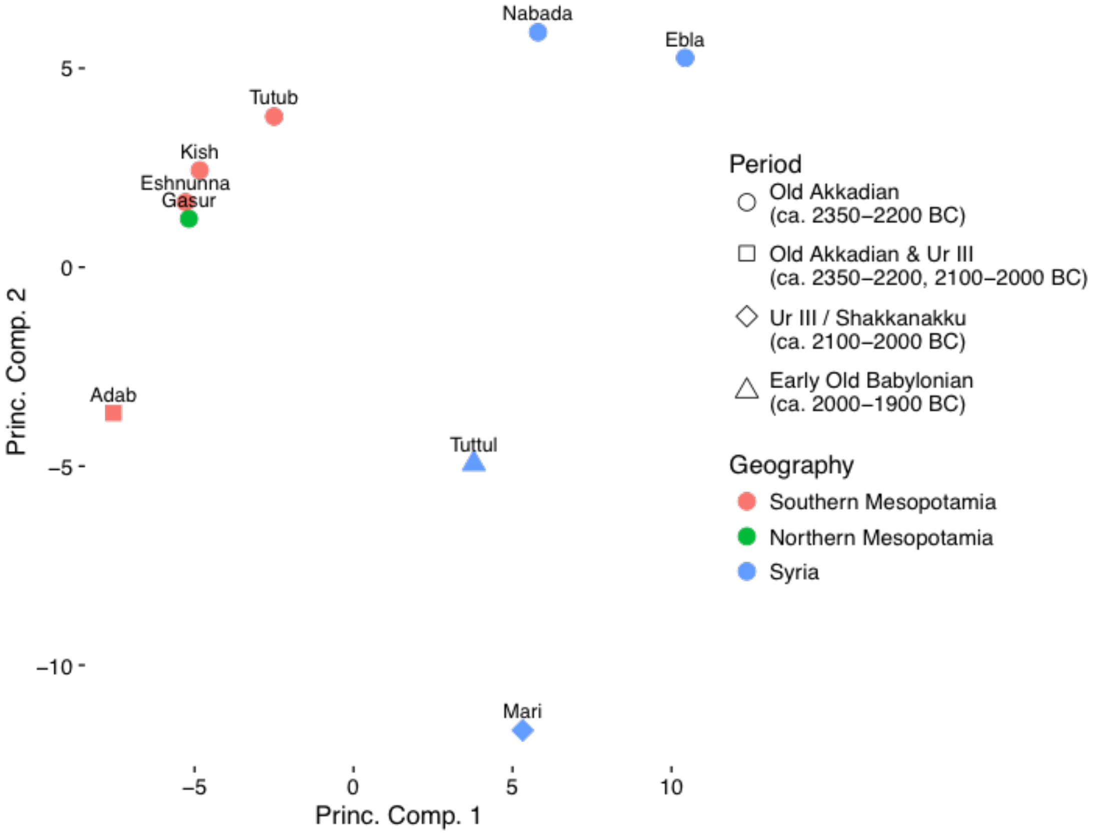

Figure 11. The plot graph comparing the first and second principal components of the filtered dataset shows that the first principal component can be attributed to geographic variation in the sites examined.

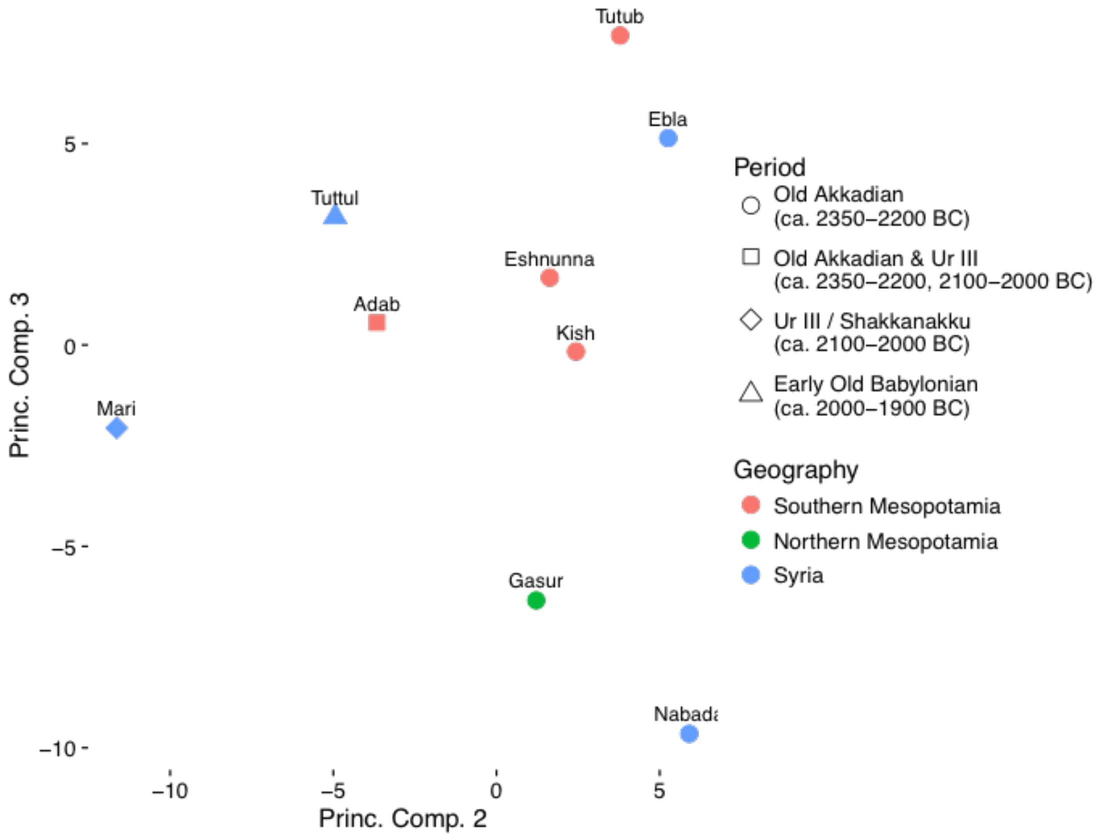

Figure 12. The plot graph comparing the second and third principal components of the filtered dataset shows that the second principal component can be attributed to temporal variation in the corpora of the sites examined.

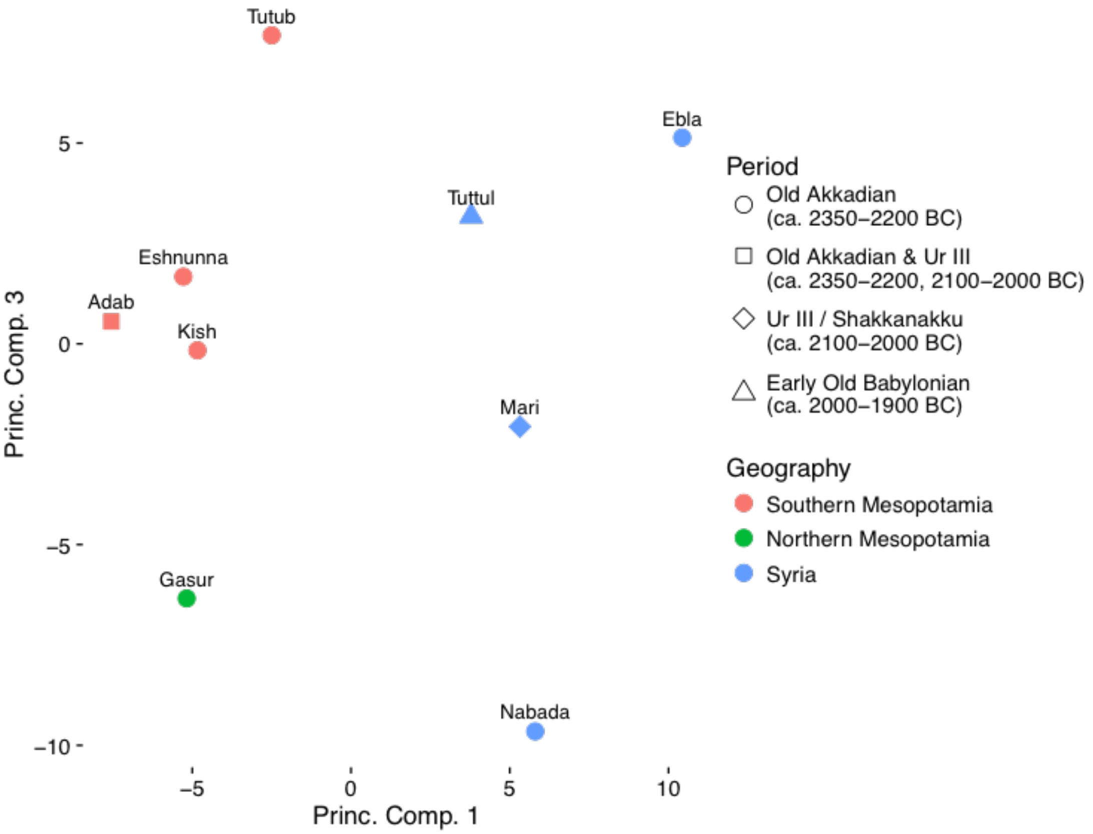

Figure 13. The plot graph comparing the first and third principal components of the filtered dataset shows that there is no apparent pattern to the clustering observed in the third principal component.

The graph in Figure 10 shows what percentage of the total amount of variation can be attributed to each main component, or sets of variables, of the filtered dataset (the nature of the first three principal components will be described further below). The results of the principal component analysis on the filtered dataset reveal that the first three components account for nearly half of the total observed variance; only those three principal components will be examined further. Plotting these principal components can provide a clearer picture of the patterns: the first two principal components can be attributed to geographic and temporal variation between the datasets. 46

The results of the principal component analysis on the filtered dataset reveal that geographic variation is the most significant factor contributing to the observed variation. This can be clearly observed in the plot graph in Figure 11: the first principal component is plotted along the x-axis of the graph, and shows that Adab, Gasur, Eshnunna, Kish, and Tutub (the sites in Mesopotamia, plotted in red for southern sites and green for the northern site) cluster together on the left-hand side, while Mari, Tuttul, Nabada, and Ebla (the sites in Syria, plotted in blue) cluster together on the right-hand side. These results mirror the trend observed in the heatmap in Figure 7b and seem to indicate that geography can explain the first principal component, which accounts for almost 20% of the variation observed in the filtered dataset. 47

The second principal component (Figure 12, plotted along the x-axis), which accounts for 15% of the variation observed in the filtered dataset, shows that Mari and Tuttul, and to a lesser extent Adab, are outliers, while showing the other sites cluster together. This pattern can be attributed to the temporal variation in the corpora examined: the syllabaries of most of the sites examined (Ebla, Nabada, Kish, Eshnunna, Tutub, Assur, and Gasur) are derived from texts dating from the Old Akkadian period (ca. 2350-2200 BC), designated with a circular plot. The texts from Adab included in this study date from both the Old Akkadian period and from the Ur III period (ca. 2100-2000 BC), designated with a square plot; from Mari, they date to the Ur III (or Shakkanakku) period, designated with a diamond-shaped plot; and from Tuttul, they date from the Early Old Babylonian period (ca. 2000-1900 BC), designated with a triangular plot. 48

The third principal component (Figure 12, plotted along the y-axis) displays no apparent pattern. Nabada and Gasur lie 49 in opposition to Tutub, while the other sites are distributed evenly between them. This principal component is included as a point of comparison, and demonstrates that random similarity in the data can contribute to the results of this type of analysis; it is only with further inspection that the significance of the results of PCA can be verified and qualified.

7.3 Summary

These three analytical tools help us visualize the data in different ways. They organize the data according to similarities and differences, and can isolate key data points that influence the results. These techniques can prompt us to think about our data in new ways, and helps us interpret the results based on our knowledge of the historical and linguistic realities of the period. 50

The results of the analyses on the unfiltered, or original, dataset suggest that the driving factor behind the variation in the unfiltered dataset is the number of syllabic values attested at each site. The principal component analysis on this dataset reveals that Ebla and Assur are outliers on either end of the spectrum within the first principal component: the large number of hapax syllabic values attested at Ebla and the lack of sufficient data at Assur are the significant factors influencing the first principal component. A closer examination of the hapax syllabic values reveals that these variables may be indicators of dialectical variation.[25] 51

The results of these analyses on the filtered dataset suggest that geographic, temporal, and random variation are driving the observed variation within the filtered dataset. The relevant syllabic values from the Syrian corpora identified through the principal component analysis can then be examined further: the lexical items that were collected in the sign studies of the Syrian corpora are used to further interpret the results of the computational analyses using traditional linguistic and text-analysis techniques. 52

The application of statistical and computational models to this dataset has demonstrated that a close examination of syllabaries can reveal new insights and confirm previous assumptions about the nature of the relationships between sites that use this syllabic writing system. While similarities between syllabaries during particular periods or within particular regions may have been assumed to exist, this methodology proves that these trends are both clearly present and strongly supported by the data. 53

8. Interpretation of the results of the computational analysis

The results of the computational models on the unfiltered dataset indicate that the primary and secondary driving factors (the first and second principal components) can be attributed to geography and time. The interpretation of the third principal component is less clear, but is presented here as a point of comparison with the first two principal components. The syllabic values that most significantly influence the variation in the data, as identified through the first three principal components, will be presented and discussed here. 54

Tables 4-6 outline the most influential syllabic values of the top three principal components and provides their respective loadings, or weights. In multivariate space — or within datasets that have multiple variables — the correlation between the principal components and the original variables is called the component loadings. The component loadings are indicative of how much of the total variation can be attributed to a given variable; in other words, the higher the component loading is, the more important that variable is for that component. For this reason, only syllabic values with loadings greater than a particular threshold are considered further. The threshold varies for each principal component, and is determined visually based on the graphs in Figures 14, 16, and 18 (for the first, second, and third principal components, respectively). 55

Syl. Value	Loading
kam	1.983368897
ḫir	1.962428466
ba [4]	1.962428466
qi [2]	1.962428466
kun	1.692715905
ag	1.612331814
tim [x] (DIN)	1.612331814
tur [2]	1.545653809
tu [3]	1.545653809
il	1.529594896
ku [8]	1.529594896
dab [6]	1.431202981
tap [x]	1.431202981
u [9]	1.431202981
sum	1.431202981
kun [3]	1.431202981
ib [2]	1.431202981
re [2]	1.347629658
par [2]	1.347629658
gur	1.347629658
qur	1.347629658
šim	1.347629658
kak	1.347629658
iz	1.347629658
ḫar	1.347629658
bar	1.302612779
tar [2]	1.302612779
nim	1.203189376
zum	1.203189376
su [4]	1.166967631
ši [2]	1.103759203

Table 4. Principal Component 1. The syllabic values that are further examined based on the loading ranges outlined in Figures 14, 16, and 18.

Syl. Value	Loading
su	2.722681481
lul	2.722681481
ib	2.722681481
bi [2]	2.348397624
ṣil [2]	2.348397624
u [3]	2.348397624
mi	2.025140491
ar	2.025140491
sa	2.000191641
dar	2.000191641
pum	1.895816769
pu	1.80880675
iš [11]	1.80880675
kab	1.645361752
se [11]	1.619356098
re	1.568075837
un	1.568075837
qu [2]	1.56399021
num	1.484426073
la [2]	1.388553884
uz	1.342110283
de [3]	1.328955355
er	1.229854594

Table 5. Principal Component 2. The syllabic values that are further examined based on the loading ranges outlined in Figures 14, 16, and 18.

Syl. Value	Loading
gan [2]	2.65175942
kum	2.65175942
gu	2.408786917
su [2]	2.222928872
ri	2.222928872
wi	2.222928872
lik	2.104672973
iḫ	2.062260202
iq	1.986903485
ul	1.986903485
ad	1.948006674
ap	1.922014923
šu [11]	1.703013331
nun	1.703013331
ša [10]	1.703013331
kur	1.703013331
we	1.703013331
pa [2]	1.69043491
al	1.69043491
u	1.69043491
ki	1.69043491
ṣa	1.69043491
sar	1.642244435
sal [4]	1.422929132
bir [5]	1.422929132
ub	1.422929132
ši [2]	1.408110193
ut	1.375395679
sa [3]	1.375395679

Table 6. Principal Component 3. The syllabic values that are further examined based on the loading ranges outlined in Figures 14, 16, and 18.

8.1 Geographic variation: the primary explanation of variation in the data

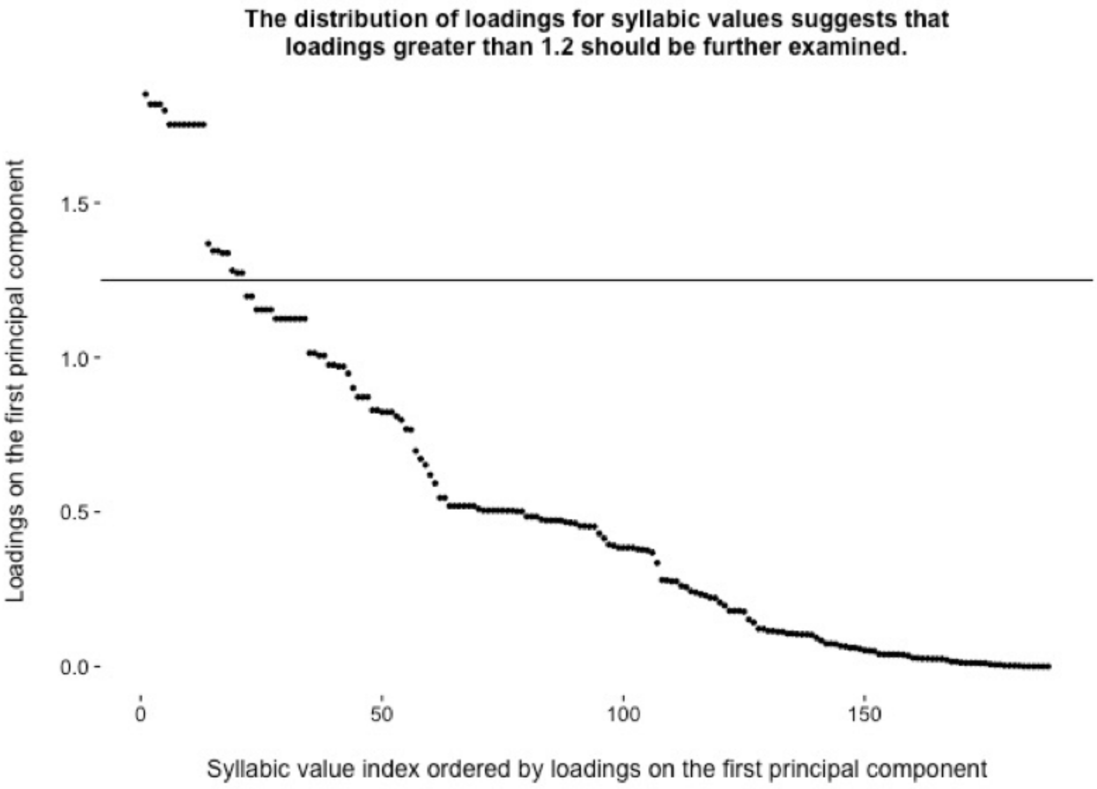

Figure 14. The distribution of loadings for syllabic values on the first principal component (see Tables 4–6 for the list of syllabic values and their respective loading).

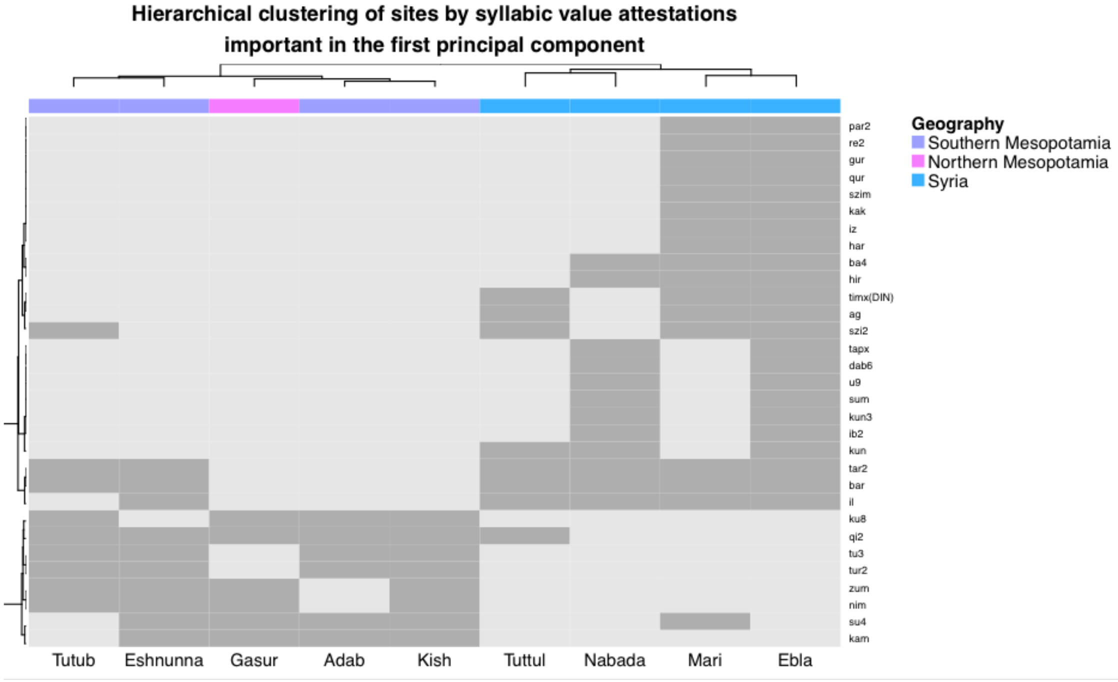

Figure 15. The 31 variables with loadings greater than 1.2 that inform the first principal component.

The results of the principal component analysis on the filtered dataset reveal that geographic variation is the most significant factor contributing to the observed variation. This can be clearly observed in the plot graph in Figure 11: the first principal component is plotted along the x-axis of the graph, and shows that Adab, Gasur, Eshnunna, Kish, and Tutub (the sites in Mesopotamia, plotted in red for southern sites and green for the northern site) cluster together on the left-hand side, while Mari, Tuttul, Nabada, and Ebla (the sites in Syria, plotted in blue) cluster together on the right-hand side.

8.2 Temporal variation: the secondary explanation of variation in the data

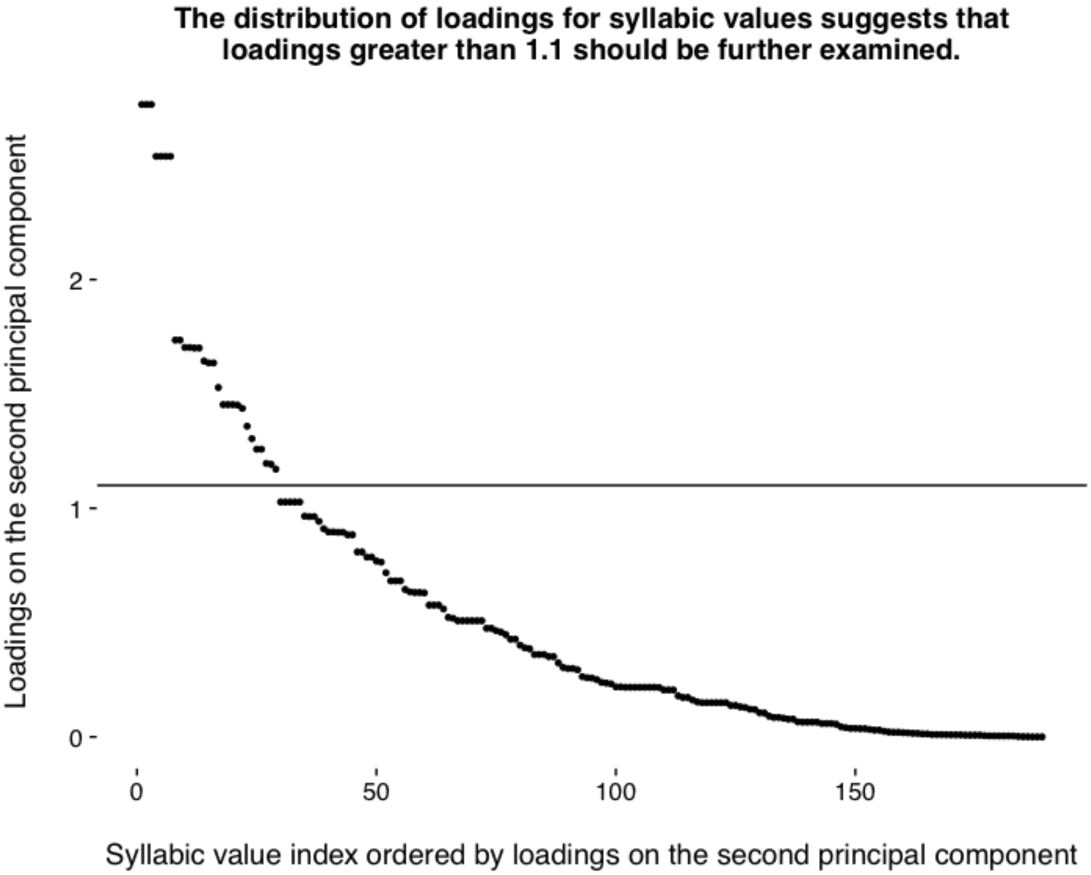

Figure 16. The distribution of loadings for syllabic values on the second principal component (see Tables 4–6 for the list of syllabic values and their respective loading).

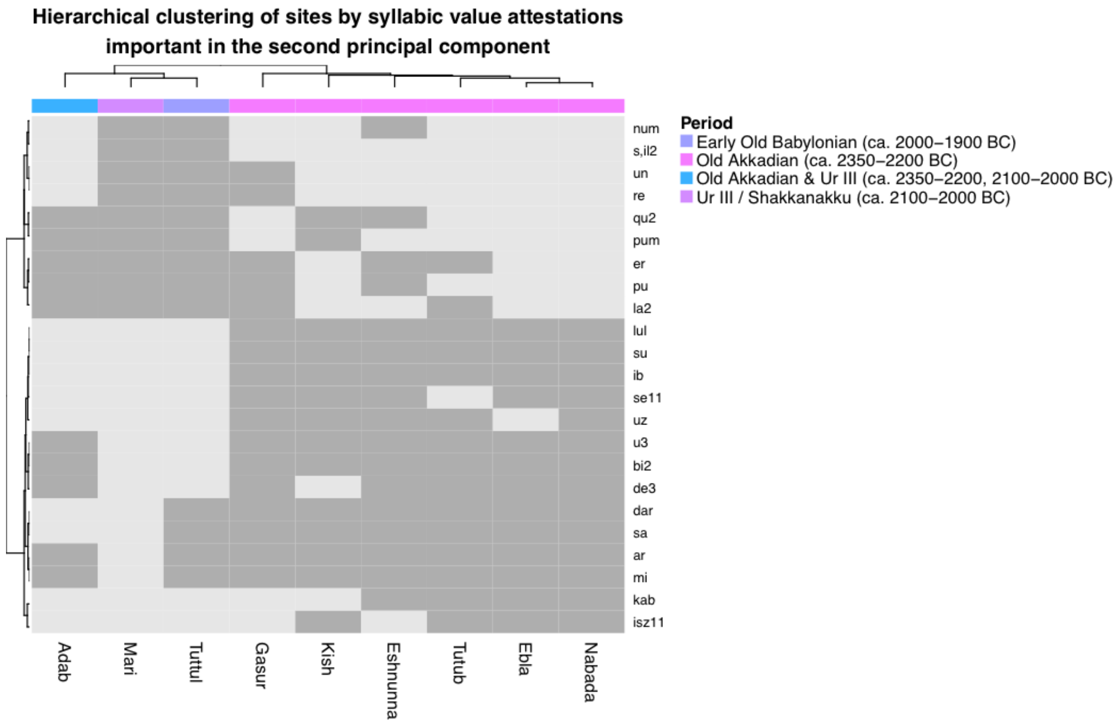

Figure 17. The 23 variables with loadings greater than 1.1 that inform the second principal component.

The second principal component (Figure 12, plotted along the x-axis), which accounts for 15% of the variation observed in the filtered dataset, shows that Mari and Tuttul, and to a lesser extent Adab, are outliers, while shows the other sites cluster together. This pattern can be attributed to the temporal variation in the corpora examined: the syllabaries of most of the sites examined (Ebla, Nabada, Kish, Eshnunna, Tutub, Assur, and Gasur) are derived from texts dating from the Old Akkadian period (ca. 2350-2200 BC), designated with a circular plot. The texts from Adab included in this study date from both the Old Akkadian period and from the Ur III period (ca. 2100-2000 BC), designated with a square plot; from Mari, they date to the Ur III (or Shakkanakku) period, designated with a diamond-shaped plot; and from Tuttul, they date from the Early Old Babylonian period (ca. 2000-1900 BC), designated with a triangular plot.

8.3 Indeterminable variation: the third explanation of variation in the dataset

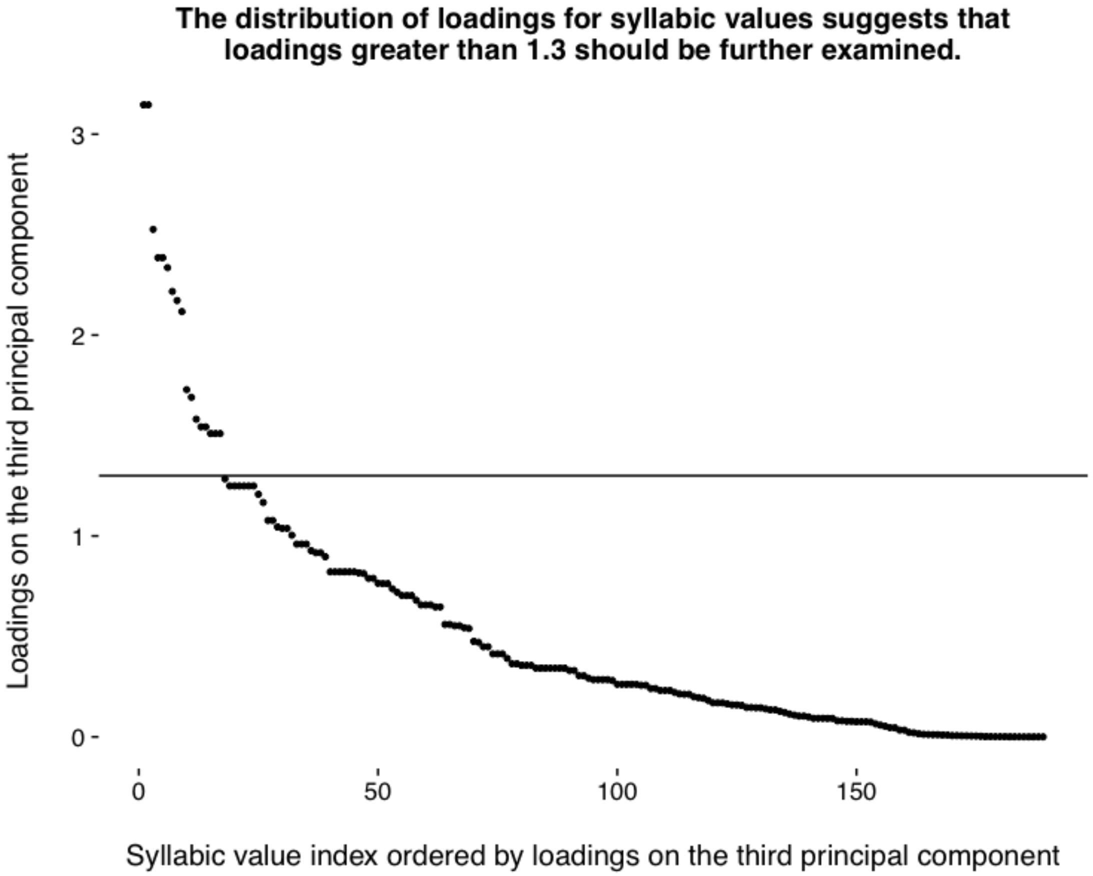

Figure 18. The distribution of loadings for syllabic values on the third principal component (see Tables 4–6 for the list of syllabic values and their respective loading).

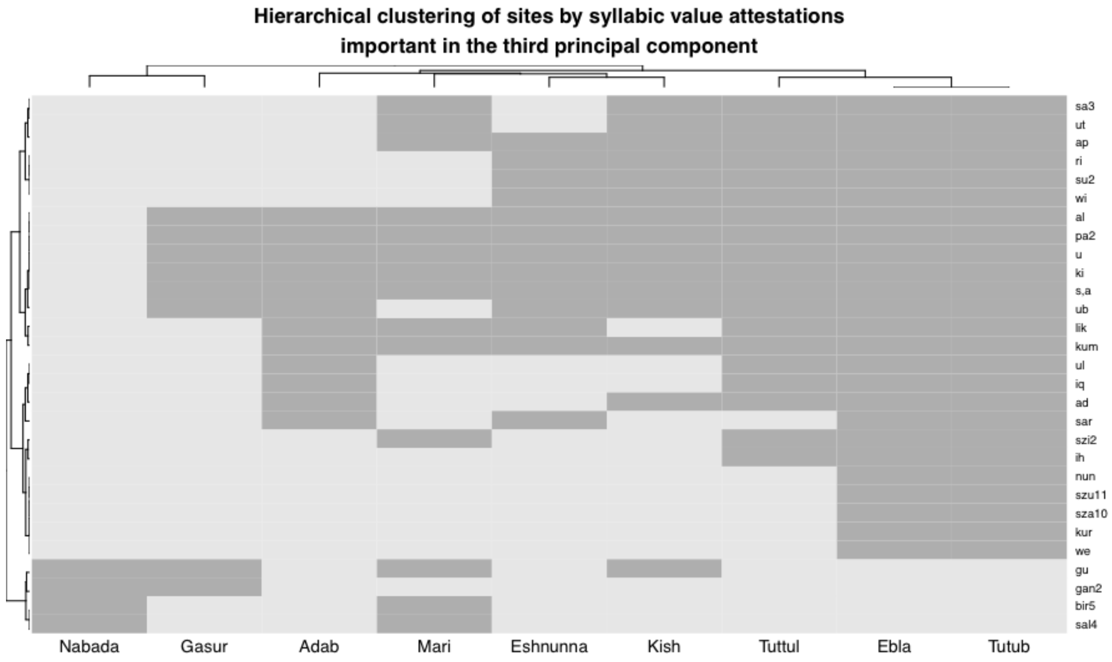

Figure 19. The 29 variables with loadings greater than 1.3 that inform the third principal component.

The third principal component (Figures 12 and 13, plotted along the y-axes) displays no readily apparent pattern. Nabada and Gasur lie in opposition to Tutub, while the other sites are distributed evenly between them. This principal component is included as a point of comparison, and demonstrates that random similarity in the data can contribute to the results of this type of analysis; it is only with further inspection that the significance of the results of PCA can be verified and qualified. 58

9. Summary of results

A close study of the syllabaries used at individual sites produces an informative dataset; when analyzed using a number of complementary techniques, this dataset can reveal new insights and confirm previous assumptions about the nature of the relationships between sites that use the syllabic sub-system within the cuneiform writing system. While similarities between syllabaries during particular periods or within particular regions may have been assumed to exist, this methodology proves that these trends are both clearly present and strongly supported by the data. 59

The three computational techniques applied to this data set produced similar general results: (1) Ebla is an outlier; (2) the Mesopotamian sites tend to cluster together more closely; and (3) the Syrian sites cluster. The concurrence in the results of these methods strongly suggests that there is indeed variation between the syllabaries of the ten sites examined, and that that variation is not simply random. 60

The results of the principal component analysis on the unfiltered dataset suggest that the driving factor behind the variation in the unfiltered dataset is the number of syllabic values attested at each site. The principal component analysis on this dataset reveals that Ebla and Assur are outliers on either end of the spectrum within the first principal component: the large number of hapax syllabic values attested at Ebla and the lack of sufficient data at Assur are the significant factors influencing the first principal component. The results of the principal component analysis on the filtered dataset suggest that geographic, temporal, and undeterminable variation are driving the observed variation within the filtered dataset. 61

By describing and comparing the most important syllabic values behind the principal components, it became apparent that dialectical variation is likely another driving factor behind the variation observed[26]. This factor is most significantly observed in the unfiltered dataset due to the relatively large number of hapax syllabic values attested at Syrian sites. 62

No similar studies into the variation of sign values used at individual sites has been conducted for any region or period in Mesopotamia, so it was not expected that such significant variation could be observed among the syllabic values used at these sites. However, it is not unexpected that this variation is associated with geography and time period, although the strength of the association is indeed unanticipated and has implications for the utility of these methods in other studies of script and dialect variation in both ancient Mesopotamia and other ancient and modern societies. 63

10. Conclusions

10.1 Methodology

The use of complementary techniques, such as digital and computational methods, to aid in the study of early scripts and cultures is a relatively new enterprise. Near Eastern archaeologists have been using complementary methods for some time, but most scholars of ancient texts and languages have only recently begun to appreciate the utility of more advanced computational modeling to help visualize and interpret data. As it stands, there are several projects that aim to digitize significant numbers of cuneiform texts for such analysis, and this first step of database compilation is indeed necessary before further analysis can be conducted. Now that these databases are in advanced stages of completion, the data can be visualized and interpreted using more advanced computational modeling in order to answer questions, which require data too extensive to be answered by examining these texts by hand, about the language, culture, economy, and history of early complex societies. Perhaps most relevantly, these methods will also be readily applicable to other fields in the humanities and social sciences. 64

In this study, computational techniques were used to understand and visualize how the ancient Mesopotamian 65 cuneiform script spread and was adapted across modern-day Iraq and Syria in the late third millennium BCE. A dataset was curated of cuneiform sign values used to write Semitic words written on tablets excavated from key geographic sites and used computational techniques, including dimensionality reduction and clustering methods, to achieve a broad view analysis of the curated data. These results directed further inquiry, in which philological approaches were used to examine and describe the linguistic environments in which each significant value was attested.

This investigation into early cuneiform syllabaries has demonstrated the advantages of using computational methods of analysis on humanities data. It also highlighted the relative strengths and weaknesses of different methods of analysis on this type of data: the results of the hierarchical clustering and PCA were able to provide a more detailed and accurate analysis of the data than the phylogenetic estimation method was. The results of this study help in understanding the broad view of cuneiform script spread and adaptation; however, there are ways such a study could be more rigorous. For example, it would be preferable to use the encoded cuneiform signs as the input data instead of using reconstructed syllabic sign values, since the step of interpreting each sign value may introduce additional human error. Currently, however, there is not a readily usable infrastructure to do this. The curation of a digital database of cuneiform signs, encoded in Unicode, along with all possible sign values for each region and time period would be beneficial to this and other types of studies. Ongoing efforts by the Unicode consortium have so far developed fonts for the following script forms: Proto-Cuneiform (late fourth millennium BC), Early Dynastic cuneiform (first half of the third millennium BC), and Neo-Assyrian cuneiform (first half of the first millennium BC), but regional or site-specific variations are not yet included. 66

10.2 Assyriological Implications

The advancement of digital tools to answer questions in the humanities is indeed an exciting prospect for the study of the ancient Near East, a field whose progress has long been hampered by the destruction of cultural heritage and political turmoil in Iraq and Syria. With the use of new methods and technologies, advancement in the field is not only possible, but will likely produce fruitful and interesting results that will bring Assyriology into the forefront of digital humanities scholarship. 67

This study attempts to give a broader view of the spread and adaptation of cuneiform across a large geographic area as opposed to examining each site individually as an isolated case. Furthermore, a complete, comprehensive syllabary allows us to determine whether an experimental phase in the writing system occurred at each site. The presence of this sort of experimentation helps us ascertain whether the cuneiform script was fully adopted by each city in Syria, inclusive of all its orthographic tendencies, or whether each city’s writing system underwent a phase of experimentation to create a slightly different result. Preliminary evidence suggests that there were clear deviations in the syllabaries of Syria from normative Mesopotamian cuneiform, which indicates that scribes experimented with the writing system more during this period than during later periods. Precisely what this can tell us about the linguistic nature of third millennium Syria should be examined through further study. 68

The adaptation and use of cuneiform in Syria provides an interesting case study for examining how people interacted with logosyllabic and syllabic writing systems. The preliminary evidence suggests that in the third millennium, particularly at sites farther away from the control of the Mesopotamian core cities, scribes were more innovative in their adaptation and use of syllabic values; there are clear deviations in the Akkadian syllabaries of Syria from normative Mesopotamian cuneiform, and also inconsistencies in the sign values and number of signs used syllabically across each of the sites investigated. This suggests that, while there must have been a more prescriptive educational approach to learning the cuneiform signs themselves and their Sumerian values, during this time period a prescribed method of writing and adapting cuneiform to write Akkadian was not included in the scribal curriculum. This lead to different adaptations of the cuneiform script and to slight variations in the number, types, and values of syllabic signs used in the individual syllabaries from sites in Syria. 69

11. Future Applications

This preliminary investigation into early cuneiform syllabaries has demonstrated the strengths of data mining and computational methods of analysis. These methodologies can be readily expanded and adapted to related and un- 70 related areas of research within the field of Assyriology; four such areas are outlined below.

11.1 A more comprehensive investigation into third millennium Akkadian

This project could readily be expanded to include the corpora from all sites that produced cuneiform texts in the third millennium. This would have comprised 19 sites in total, including the ten that compose that data sources for this thesis. The additional nine sites include: Nagar, Shehna / Shubat-Enlil, Umma, Shuruppak, Abu Salabikh, Nippur, Girsu, Umm al-Jir, and Susa. As was explained in the introduction of this thesis, many of these sites were unable to be included as a part of the current work for a number of practical reasons: (1) inaccessibility to the tablets due to their location in museums in Syria, Iraq, or Turkey; (2) a lack of published photographs, hand drawings, or transliterations and transcriptions of the inaccessible texts; and (3) irrelevance of the texts to this study of the adaptation of cuneiform to write Semitic languages (in the case of Susa). Going forward, it would be possible to expand and update this syllabary to include the orthography of each sign, either in all cases, or, more likely at first, just in the cases where they obviously appear to differ drastically. In the process of its expansion, the syllabary can then be digitized to create a free and easily accessible database. 71

11.2 Applications of this methodology to all East Semitic dialects

Another application of this methodology would be to examine the development of the East Semitic dialects across the entire history of the cuneiform script[27]. The wealth of written sources left behind by the cuneiform cultures comprise a unique and comprehensive data set through which we can understand the history and development of these dialects. These developments and relationships have been determined largely through a combination of textual analysis and the comparative method [Hetzron 1969] [Hetzron 1976] [Faber 1997] [Huehnergard 2011] (Hetzron 1974; Hetzron 1976; Faber 1997; Huehnergard 2011), but using computational methods to analyze the relevant data can provide new insights into the evolution and spread of these dialects[28]. This study would likely examine the following East Semitic dialects: 72

- Eblaite (ca. 2350-2250 BC)
- Old Akkadian (ca. 2350-2200 BC)
- Ur III Akkadian (ca. 2100-2000 BC)
- Old Assyrian (ca. 1950-1850 BC)
- Old Babylonian (ca. 2000-1600 BC)
- Middle Assyrian (ca. 1400-1000 BC)
- Middle Babylonian (ca. 1400-1100 BC)
- Neo-Assyrian (ca. 911-612 BC)
- Neo-Babylonian (ca. 626-539 BC)

11.3 Comparing computational methods to find the optimal approach

The methodology used in this thesis has significant potential to be informative and relevant not only for this project, but for other Assyriological research projects as well. A thorough grasp on phylogenetic analysis programs such as MacClade, Mesquite, and PAUP* in addition to the programming languages R, Python, and Perl can enable researchers to build custom code and pipelines specifically for the analysis of syllabaries, orthographies, or other aspects of the cuneiform script. In terms of future research, this will be particularly useful in identifying the definitive syllabaries and sign lists for early phases of the cuneiform script – such as Uruk and Ur III sign lists – as well as for little-understood relatives to the cuneiform writing system, such as Proto-Elamite. 73

11.4 Applications to the problem of texts with no known provenance

The methodology used here has the potential to aid in providing provenance to looted or misplaced tablets. To do this, a much larger dataset is required. In particular, a large collection of digitized texts with adequate encoding of the class of each sign (i.e., syllabic, logographic, determinative, etc.). This dataset can rely largely on the texts published on the CDLI database; computational methods could then be employed to identify the class of the majority of signs attested in 74 most time periods and genres. Once this dataset is collected, principal component analysis can aid in the identification of the distinctive signs or combinations of signs that, when present, suggest a specific provenance.

12. Input Data

12.1 Phylogenetic Estimation

a. Unfiltered data

```
#NEXUS
Begin taxa;
    Dimensions ntax=10;
    taxlabels
        Ebla
        Mari
        Nabada
        Tuttul
        Adab
        Eshnunna
        Kish
        Tutub
        Assur
        Gasur
    ;
End;
Begin data;
    Dimensions ntax=10 nchar=319;
    Format datatype=standard;
    Matrix
        Ebla  1 0 0 1 1 1 1 1 1 1 1 1 1 0 1 1 1 0 0 1 1 1 0 0 1 1 1 1
        0 0 1 1 0 1 0 0 1 1 1 1 0 0 0 0 1 1 1 1 1 1 1 1 1 1 1 1 1 1 1
        0 1 1 1 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 0 1 1 1 1 0 1 1 1 1 1 0
        1 1 0 0 0 0 0 1 1 1 1 1 1 0 1 1 0 1 1 1 0 1 0 0 0 0 1 1 1 1 0
        0 1 0 1 1 1 0 0 0 1 0 1 0 1 1 0 1 1 1 0 0 1 1 1 1 0 1 0 0 1 1
        0 0 0 1 1 0 1 1 0 1 1 0 1 0 0 1 1 1 1 0 1 1 1 1 1 1 0 0 1 0 1
        1 0 1 1 0 0 1 1 0 1 1 0 0 0 1 1 0 1 1 1 0 0 0 0 0 1 1 1 1 1 1
        1 1 0 1 1 1 1 1 1 1 0 0 1 0 0 0 1 1 1 1 0 1 1 1 1 1 0 1 0 1 0
        1 1 0 1 0 1 1 1 1 1 0 1 0 0 0 1 1 1 1 1 1 1 1 0 1 1 0 0 1 1 0
        0 0 1 1 0 1 0 1 0 1 0 1 1 0 0 0 0 0 0 1 0 0 1 0 0 1 0 1 1 1 0
        0 1 0 1 1 1 1 1 0 0 0 0

        Mari  0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 1 0 1 0 0 0 1 1 0 0 1 1
        0 1 0 0 1 1 0 0 1 0 0 1 0 0 1 1 1 1 0 1 1 0 0 1 0 1 1 1 0 1 1
        1 0 0 0 0 1 1 1 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 1 0 1 0 0 0 0 1
        0 0 0 0 0 0 1 0 1 1 1 1 1 1 1 1 1 1 1 1 0 0 1 1 0 0 0 0 1 0 0
        0 1 0 1 0 1 1 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0 0 0 1 0
        0 0 0 1 0 0 0 1 0 0 1 0 1 0 0 1 1 1 1 0 1 0 1 1 1 0 1 0 1 0 1
        0 0 1 1 1 0 1 1 1 1 1 0 0 0 0 1 0 1 0 1 1 0 1 1 0 1 0 1 1 0 1
        0 1 0 0 1 1 0 0 0 1 1 0 0 1 1 0 1 0 0 1 0 0 0 1 0 0 0 1 0 1 0
        0 0 0 1 0 1 1 0 1 0 1 0 1 0 1 0 0 0 1 1 0 0 0 0 1 1 0 1 0 1 1
        1 1 0 1 1 0 0 1 1 1 0 1 0 0 0 0 1 0 0 1 1 1 1 0 0 1 1 0 0 1 0
        1 1 0 1 1 1 1 0 0 0 0 0

        Nabada  1 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 1 1 1 0 1 0 0 0 1 1 1
        0 0 0 1 0 0 1 0 0 1 0 0 1 0 0 0 1 1 1 0 1 1 1 0 0 0 1 0 1 1 1
        0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 1 0 0 0 0
        0 1 0 1 0 0 0 0 1 1 0 0 1 0 0 1 1 0 1 1 0 0 0 0 0 0 0 1 1 1 0
        0 0 1 0 1 0 1 0 0 0 1 0 1 0 1 1 0 1 1 1 0 1 0 0 0 0 0 1 0 0 1
        1 0 0 0 1 1 0 1 0 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 1 0 0 0 0 0
        0 0 0 0 1 0 0 1 0 0 1 1 0 0 0 0 0 0 1 0 1 0 1 0 0 0 1 0 1 1 0
        1 0 0 0 0 1 1 1 0 0 1 0 1 0 0 0 0 0 0 1 0 0 0 0 1 0 0 0 1 0 0
        0 0 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 1 1 1 0 0 1 0 0 0 0 0
```

```
0 0 0 0 1 0 1 1 1 0 0 0 1 0 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0 0
0 0 1 0 1 0 0 1 1 0 0 0 0

Tuttul  1 0 0 0 0 1 1 1 1 1 0 0 1 1 1 0 1 0 1 1 0 0 1 1 0 0 1
0 0 0 1 0 0 1 1 0 1 0 0 1 1 1 0 0 1 1 0 1 1 0 0 0 0 1 0 1 1 1
0 0 0 1 0 1 1 1 0 0 0 0 1 1 0 0 1 0 0 0 0 0 1 1 0 1 1 0 0 0 0
0 1 0 0 0 0 0 0 1 1 0 0 1 1 0 1 0 0 0 1 1 0 0 0 0 0 1 0 0 1 0
0 1 1 1 1 1 1 0 0 0 1 0 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 0 1
1 0 0 0 1 1 0 0 0 0 0 1 0 1 1 1 0 0 1 1 0 1 0 1 0 0 1 0 0 0 0
1 1 1 0 1 1 0 1 1 0 1 1 0 0 1 1 0 0 1 0 1 0 0 0 0 0 1 0 1 1 0
1 0 0 0 0 1 1 0 0 0 1 1 0 0 0 0 1 0 0 1 1 0 0 1 0 0 0 1 1 0 1
1 1 1 0 1 1 0 0 0 1 1 1 0 0 0 1 0 0 1 0 0 1 0 1 1 1 1 1 1 0 1
0 0 1 0 1 1 0 1 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 1 1
0 0 1 0 1 1 1 1 0 0 0 0 0

Adab  1 1 1 0 0 1 1 1 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 1 1 1 0
0 0 1 0 0 0 1 1 0 0 0 1 0 0 0 0 0 1 0 0 1 1 0 0 0 0 0 1 0 1 0
0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0
0 0 0 0 0 1 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 1
0 1 0 1 1 1 0 0 1 1 1 0 0 0 0 0 1 1 1 1 0 0 0 0 1 0 0 1 1 0 1
1 1 0 1 0 0 0 1 0 0 1 0 1 0 1 0 0 1 1 0 1 0 1 1 0 0 1 0 0 0 1
1 0 0 1 0 1 0 1 1 1 1 0 0 1 0 1 1 1 1 1 1 0 0 1 0 1 0 1 1 0 1
0 1 0 1 1 1 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 1 0
0 0 1 1 0 0 1 0 1 1 0 0 0 0 0 1 0 1 0 1 1 1 0 0 0 1 1 0 0 0 0
0 1 0 1 0 0 0 1 1 1 0 1 0 0 0 0 0 0 0 0 0 0 1 1 0 1 1 0 1 1 0
0 1 1 0 1 0 0 0 0 0 0 1

Eshnunna  0 0 0 0 0 1 1 0 1 0 1 0 0 0 0 0 1 0 1 0 0 0 0 0 1 1
1 0 1 1 1 0 1 1 1 1 1 0 0 1 0 0 0 0 1 1 0 1 1 1 0 0 0 1 0 1 0
1 0 0 0 0 1 0 1 1 0 0 1 0 1 0 0 0 0 0 1 0 0 0 1 0 1 0 1 0 0 0
0 0 1 0 0 0 0 1 0 1 1 0 0 1 0 1 1 0 0 1 0 1 0 0 0 0 0 0 0 0 1
0 0 0 1 0 0 0 1 0 1 1 1 0 0 0 0 0 0 1 1 1 0 0 0 0 0 1 0 0 0 0
1 1 1 0 1 1 1 0 0 1 0 0 1 0 1 0 1 0 0 1 1 0 1 0 1 1 0 1 1 1 0
0 1 1 0 1 1 0 0 1 1 1 1 1 0 0 1 0 1 0 1 1 0 1 0 0 0 0 1 1 1 1
0 1 1 1 0 1 1 1 1 0 0 1 1 1 0 0 0 0 1 0 1 0 0 0 1 1 0 0 0 1 0
1 0 0 0 1 1 0 0 1 0 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0 0 1 1 1 1 0
0 0 1 0 0 1 0 1 1 1 1 0 1 1 0 0 1 0 0 1 0 0 1 0 0 0 1 1 1 0 1
1 1 0 1 0 1 1 0 1 1 0 0 0 0

Kish  0 0 0 0 0 1 1 1 1 0 1 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1 1 0
0 1 1 0 0 1 0 0 1 0 0 1 0 0 0 0 1 1 0 1 1 1 0 0 0 0 0 1 0 1 1
0 0 1 0 0 1 1 0 1 0 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 0 0
1 0 0 1 1 1 0 1 0 0 0 1 0 1 1 1 0 1 1 1 0 0 0 0 0 0 0 0 1 0 1
0 1 0 0 1 1 1 0 0 1 0 0 1 0 0 0 1 1 0 1 0 0 0 0 1 1 0 0 0 0 1
1 0 1 1 0 0 0 1 0 0 1 1 1 0 0 0 0 1 1 0 1 0 1 0 0 0 0 1 0 1 1
1 1 0 1 0 0 1 1 1 1 1 1 1 1 1 1 0 1 0 1 0 0 0 1 1 1 0 1 1 0 1
0 0 0 1 1 1 1 0 0 1 0 1 0 0 0 0 1 0 0 1 0 0 1 0 1 0 0 1 0 1 0
0 0 1 1 0 0 1 1 1 1 0 1 1 1 0 0 1 0 0 1 1 1 0 0 1 1 1 1 0 0 0
0 0 0 1 1 1 1 1 1 0 1 1 0 0 1 0 0 1 0 1 0 1 0 1 0 1 1 0 0 0 0
0 1 0 1 1 1 1 1 1 0 0 0

Tutub 1 0 0 0 0 1 1 1 1 1 1 1 0 0 0 0 1 0 0 0 0 0 1 0 0 0 1 0
0 1 1 0 0 1 1 0 1 0 0 1 0 0 0 0 1 1 0 1 1 1 0 0 0 1 0 1 0 1 0
0 0 1 1 1 1 0 1 0 1 1 1 0 0 0 0 1 1 0 0 0 1 0 1 0 1 0 0 0 0 0
1 1 0 0 0 1 0 1 1 0 0 1 1 0 1 0 0 1 1 1 1 0 1 0 1 0 0 0 1 0 0
0 1 0 1 1 1 0 1 0 1 1 0 0 0 0 1 1 1 1 1 0 0 0 0 1 1 0 0 0 0 1
1 0 1 1 0 0 0 1 1 0 1 0 1 0 0 0 0 1 1 1 1 0 1 1 0 0 1 0 0 0 1
1 0 0 1 0 0 1 1 1 1 1 0 0 1 0 1 0 1 1 1 1 0 0 1 0 1 0 1 1 0 1
1 1 0 0 1 1 1 1 0 1 0 1 0 0 0 0 0 0 0 1 1 0 1 1 0 1 0 1 0 0 1
0 1 0 1 0 0 1 1 1 1 0 0 1 1 0 0 1 1 1 1 1 1 0 1 1 1 1 0 0 0 0
0 1 0 1 0 1 1 1 0 1 0 1 0 0 1 1 0 1 0 0 0 0 0 1 1 1 1 0 0 1 1
0 1 0 1 1 1 1 0 0 0 0 0

Assur 0 0 0 0 0 1 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0
0 0 1 0 0 1 0 0 0 0 0 1 0 0 0 0 0 1 0 1 1 0 0 0 0 0 0 1 0 0 0
```

```
        0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
        1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
        0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
        0 0 0 1 0 0 0 1 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 0 0
        0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1
        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0
        0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 0 0 0
        0 1 0 0 0 0 0 0 0 0 0 0

        Gasur 0 0 0 1 0 1 1 1 0 1 1 0 0 0 0 0 1 0 0 0 0 0 0 0 0 1 1 0
        0 1 1 0 0 1 1 0 1 0 0 1 0 0 0 0 1 1 0 1 1 1 0 1 0 0 0 1 0 1 1
        0 0 0 1 0 1 1 0 0 1 0 0 1 0 0 0 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0
        1 0 1 0 1 0 0 1 0 0 0 1 0 1 1 0 0 1 1 0 0 0 1 0 0 0 0 0 1 0 0
        0 1 0 0 0 1 0 1 0 1 0 0 0 0 0 0 1 1 1 0 0 0 0 0 0 0 0 0 0 0 0
        0 0 0 1 0 0 0 1 1 0 1 0 1 0 0 0 0 1 1 0 1 0 1 1 0 0 0 1 0 1 1
        1 1 0 1 1 0 0 1 0 1 0 0 0 0 0 1 0 1 0 1 0 0 0 1 1 1 0 1 1 0 1
        1 0 1 0 1 1 1 0 0 1 1 1 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 1 1 1 1
        1 0 1 1 0 0 1 0 1 1 0 0 0 1 0 1 0 0 1 0 1 1 0 0 0 1 1 0 0 0 0
        0 1 0 1 1 1 0 1 0 1 1 1 0 0 1 0 1 0 0 0 0 1 0 1 0 1 0 0 1 0 0
        0 1 0 1 1 0 1 1 0 1 1 0

        ;
End;
```

b. Filtered data

```
\#NEXUS
Begin taxa;
    Dimensions ntax=9;
    taxlabels
        Ebla
        Mari
        Nabada
        Tuttul
        Adab
        Eshnunna
        Kish
        Tutub
        Gasur
    ;
End;
Begin data;
    Dimensions ntax=9 nchar=188;
    Format datatype=standard;
    Matrix
        Ebla  1 1 1 1 1 1 1 1 1 0 1 1 0 0 1 1 1 0 0 0 0 1 0 1 1 1 1 1
        1 1 0 1 1 1 1 1 1 1 1 1 1 0 0 1 1 1 1 1 1 1 0 0 0 1 1 1 1 1 0
        1 1 1 1 0 1 1 0 1 1 0 0 0 0 1 1 1 0 1 1 0 1 0 1 1 1 0 0 1 1 1
        0 0 1 1 0 1 1 0 0 1 0 1 1 0 1 0 1 0 1 0 1 1 1 1 0 0 0 1 1 1 1
        1 1 0 0 1 1 1 1 1 1 1 1 0 1 1 0 1 1 1 1 1 0 1 0 0 0 1 1 1 1 1
        1 1 1 0 1 1 0 0 1 0 0 0 1 0 0 1 0 0 0 0 1 0 0 1 0 0 1 0 1 1 0
        1 1 1 1 1

        Mari  0 0 1 0 0 0 0 0 0 1 0 0 1 1 0 0 1 1 1 0 0 1 1 1 0 1 1 1
        0 1 1 0 0 0 1 1 1 0 0 0 0 1 0 0 0 1 1 1 0 0 0 0 0 0 1 1 1 1 1
        1 1 1 1 1 0 0 0 1 0 1 0 0 0 1 0 0 0 0 0 0 1 1 0 1 0 0 0 0 0 1
        0 0 1 1 1 1 0 1 0 1 0 1 0 0 1 1 1 1 1 0 0 1 0 1 1 1 0 0 0 1 0
        0 0 1 0 1 0 1 0 1 0 0 1 0 0 0 0 1 1 0 1 0 1 0 1 0 1 0 0 0 1 1
        0 0 0 0 1 1 0 1 1 1 1 1 0 0 1 1 0 0 1 0 1 1 1 1 0 0 1 1 0 1 0
        1 1 1 1 0
        Nabada  1 0 0 0 0 1 0 0 0 1 0 1 0 0 1 1 0 0 0 0 0 1 1 1 0 0 1
        0 1 0 1 1 0 0 1 0 0 0 0 0 0 0 0 0 1 1 0 0 1 0 1 0 0 1 1 0 0 0
        0 1 1 1 0 0 1 1 0 1 0 0 0 0 0 1 1 1 0 1 1 0 0 0 1 1 1 0 0 1 1
        0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 0 1 0 1 0 0 0 0 1 0 0 0 0 0 0
        0 1 0 0 1 0 1 0 0 1 0 0 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 1 0 1
        1 1 1 0 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 1 0 0 0
        0 1 0 0 1 1

        Tuttul  1 0 1 1 1 0 0 1 1 1 1 0 1 1 0 0 0 0 0 1 0 1 0 1 0 0 1
        0 1 0 0 1 0 1 1 1 0 0 0 0 1 1 1 0 0 1 1 0 1 0 0 0 0 1 1 0 0 1
        0 0 0 1 1 0 0 0 0 1 1 0 0 0 0 0 0 0 1 0 0 0 1 0 0 1 1 0 0 1 0
        0 0 1 0 0 1 0 1 0 0 0 0 1 1 1 0 1 1 0 1 1 1 0 0 1 0 0 0 0 0 0
        0 0 0 1 0 0 1 1 1 0 0 0 1 1 1 1 0 0 0 0 1 1 1 0 0 0 1 0 0 1 0
        0 1 0 1 1 1 1 1 1 1 0 1 1 0 1 1 1 0 0 0 0 0 0 0 0 0 1 0 0 1 1
        0 1 1 1 1 0

        Adab  1 0 1 0 0 0 0 0 0 0 0 1 0 1 1 1 0 0 0 1 1 0 0 0 1 0 0 0
        0 0 0 0 1 1 0 0 0 0 0 0 0 0 1 0 0 0 0 0 0 0 0 0 1 0 0 0 0 0 1
        0 0 0 0 0 0 0 1 1 1 0 0 1 1 0 0 0 0 1 1 1 1 0 0 0 1 1 0 0 0 1
        0 1 0 0 1 0 0 1 0 0 0 1 1 0 0 0 0 1 1 1 0 1 1 1 1 1 0 0 0 1 1
        0 0 1 0 0 0 0 0 1 1 0 1 0 0 0 1 0 1 0 1 1 0 0 0 0 0 1 0 1 0 1
        1 1 0 0 0 1 1 0 0 0 1 0 0 0 1 1 0 0 0 0 0 0 0 1 1 0 1 1 1 1 0
        0 1 0 0 0

        Eshnunna  0 0 1 1 0 1 0 0 0 1 0 0 0 0 1 1 0 1 1 1 1 1 0 1 1 0
        1 0 0 0 0 0 1 0 1 1 0 0 1 0 1 0 0 0 1 1 0 1 1 0 0 0 1 1 1 0 0
        0 1 0 1 0 1 0 0 0 0 0 0 0 1 1 0 0 0 0 0 1 1 0 1 0 0 1 1 1 1 1
        0 1 0 1 0 0 1 0 1 1 1 0 0 1 1 0 1 0 1 1 1 1 0 1 1 0 1 0 0 1 1
        1 1 1 0 1 1 1 1 0 1 1 0 0 1 0 0 0 1 0 1 0 1 1 0 0 1 1 1 1 0 0
        1 1 1 1 0 0 1 1 1 1 0 1 0 0 1 1 1 0 1 1 0 1 0 1 0 0 0 1 1 1 1
        1 1 1 1 0 1 1

        Kish  0 0 1 1 0 1 0 0 0 0 0 0 0 1 0 1 0 1 0 0 0 1 0 1 1 0 0 0
        0 1 0 1 0 0 1 1 0 1 0 1 1 0 0 0 0 1 0 1 1 0 0 1 1 1 0 0 0 0 1
        1 1 1 1 0 0 0 1 0 1 1 0 0 0 0 0 0 0 1 0 1 1 1 0 0 1 1 1 0 0 1
        0 0 0 0 0 0 0 0 1 0 1 1 1 1 0 0 1 1 1 1 1 1 0 1 0 1 1 0 0 0 1
        1 0 0 1 1 0 1 1 0 1 0 1 0 0 0 1 0 1 1 1 1 0 1 1 1 0 0 1 0 0 1
        1 1 0 0 1 1 1 1 0 0 0 1 1 1 1 0 1 1 0 1 1 0 1 0 1 0 1 1 0 0 0
        1 1 1 1 1

        Tutub 1 0 1 1 1 1 1 0 0 0 0 0 1 0 0 0 0 1 0 1 0 1 0 1 1 0 1 0
        0 0 0 1 1 1 1 0 1 0 1 1 1 0 0 1 1 1 0 1 1 1 0 0 1 1 1 0 0 1 0
        0 1 1 1 1 0 0 0 1 1 0 1 0 1 0 0 0 1 1 1 1 1 1 0 0 1 1 1 0 0 1
        1 0 0 0 1 0 0 1 0 0 0 1 1 0 0 0 1 1 1 1 0 1 1 1 1 1 0 0 1 1 0
        1 1 0 1 0 0 1 1 1 0 1 0 1 0 1 0 0 1 1 1 1 0 0 1 1 0 0 1 1 1 1
        1 1 0 1 1 1 1 0 0 0 1 0 1 1 0 1 0 1 0 1 0 0 0 0 1 1 1 1 0 1 1
        1 1 1 1 0

        Gasur 0 1 1 0 1 1 0 0 0 0 0 0 0 0 0 1 0 1 0 1 0 1 0 1 1 1 0 0
        0 1 0 0 1 0 1 1 0 0 1 0 0 1 0 0 0 1 0 0 1 0 1 1 0 1 0 0 0 0 1
        0 1 1 0 1 0 0 0 0 0 0 1 0 0 0 0 0 0 1 1 0 0 0 0 0 0 0 0 0 0 1
        1 0 0 0 1 0 0 0 1 0 1 1 1 1 0 1 0 0 0 0 0 1 0 1 0 1 1 0 1 0 0
        1 0 1 1 1 1 0 0 1 0 0 1 1 1 0 1 0 1 0 1 1 0 0 0 1 0 1 0 0 1 0
        1 1 0 0 0 1 1 0 0 0 1 1 1 0 0 1 1 1 1 0 0 0 1 0 1 0 1 0 1 0 0
        1 1 0 1 1

    ;
End;
\pagebreak
```

12.2 RStudio

```
    \begin{verbatim}

    library("NMF")
    library("FactoMineR")
    library("data.table")
    library("ggplot2")
    library("cowplot")
    library("ggdendro")
    library("pvclust")
    #data.df<-data.frame(syl_matrix_forR_090416)[1:319,5:14]
    data.df<-data.frame(syl_matrix_forR_090416)[1:319,4:14]
    rownames(data.df)<-data.df$Sign.Value
    data.df<-data.df[,2:11]
    data.df<-data.df[rowSums(data.df)>0,]

    #unfiltered w/ Assur
    colnames(data.df)<-gsub("Esznunna","Eshnunna",colnames(data.df))
    rclust<-hclust(dist(data.df,method="manhattan"), method="ward.D2")
    cclust<-hclust(dist(t(data.df),method="manhattan"), method="ward.D2")
    aheatmap(data.df,color=‘grey:2’, Rowv=rclust,breaks=c(-0.05,0.5,1.05),
    labRow=NULL,main="Hierarchical clustering of sites by syllabic value
    attestations \n", legend=FALSE, fontsize=14, cexCol = 0.8)

    #unfiltered data w/ Assur
    forpca<-t(data.df)
    answer<-PCA(forpca,ncp=10,graph=FALSE)
    pc.eig.df<-data.frame(answer$eig)
    ggplot(data=pc.eig.df[c(1:9),], aes(x=gsub("comp","Comp.",
        rownames(pc.eig.df[c(1:9),])), y=percentage.of.variance)) +
    geom_bar(stat="identity" +xlab("\nPrincipal Component")
        +ylab("Percentage of Variance") +ggtitle("Variance distribution
        across principal components derived from viable syllabic values")
        coord.rs.df<-data.frame(answer$ind$coord)
        gsub("Esznunna","E?nunna",row.names(coord.rs.df))
        Geography<-factor(c("Syria", "Syria", "Syria", "Syria", "Southern
        Mesopotamia", "Southern Mesopotamia", "Southern Mesopotamia",
        "Southern Mesopotamia", "Northern Mesopotamia", "Northern Mesopotamia"),
        levels=c("Southern Mesopotamia", "Northern Mesopotamia", "Syria"))
        Period<-factor(c("Old Akkadian\n(ca. 2350-2200 BC)", "\nUr III/
        Shakkanakku\n(ca. 2100-2000 BC)", "Old Akkadian\n(ca. 2350-2200 BC)",
        "\nEarly Old Babylonian\n(ca. 2000-1900 BC)\n", "\nOld Akkadian & Ur
        III\n(ca. 2350-2200, 2100-2000 BC)", "Old Akkadian \n(ca. 2350-2200
BC)",
        "Old Akkadian\n(ca. 2350-2200 BC)", "Old Akkadian\n(ca. 2350-2200 BC)",
"Old
        Akkadian\n(ca. 2350-2200 BC)", "Old Akkadian\n(ca. 2350-2200 BC)"),
        levels=c("Old Akkadian\n(ca. 2350-2200 BC)", "\nOld Akkadian & Ur
        III\n(ca. 2350-2200, 2100-2000 BC)", "\nUr III / Shakkanakku \n(ca.
        2100-2000 BC)", "\n Early Old Babylonian \n(ca. 2000-1900 BC)\n"))
    ggplot(data=coord.rs.df, aes(x=Dim.1, y=Dim.2))+geom_point(aes(shape=Period,
        fill=Geography, color=Geography),size=4)+scale_shape_manual
        (values=c(21,22,23,24))+xlab("Princ. Comp. 1")+ylab("Princ. Comp. 2")
        +geom_text(label=gsub("Esznunna","E?nunna", row.names(coord.rs.df)),
        nudge_y=0.7)

    #filtering out hapax signs and allsites signs
    data2.df<-data.df[rowSums(data.df)>1 & rowSums(data.df)<9,-9]
    colnames(data2.df)<-gsub("Esznunna","Eshnunna",colnames(data2.df))
    colnames(data.df)<-gsub("Esznunna","Eshnunna",colnames(data.df))

    #convert table into final table for thesis for hapax signs
    data.onesite.df<-data.df[rowSums(data.df)==1,]
    data.onesite.df$Site<-"DUMMY"
    data.onesite.df[data.onesite.df$Ebla==1,]$Site<-"Ebla"
    data.onesite.df[data.onesite.df$Mari==1,]$Site<-"Mari"
```

```
  data.onesite.df[data.onesite.df$Nabada==1,]$Site<-"Nabada"
  data.onesite.df[data.onesite.df$Tuttul==1,]$Site<-"Tuttul"
  data.onesite.df[data.onesite.df$Adab==1,]$Site<-"Adab"
  data.onesite.df[data.onesite.df$E?nunna==1,]$Site<-"E?nunna"
  data.onesite.df[data.onesite.df$Kish==1,]$Site<-"Kish"
  data.onesite.df[data.onesite.df$Tutub==1,]$Site<-"Tutub"
  #data.onesite.df[data.onesite.df$Assur==1,]$Site<-"Assur"
  data.onesite.df[data.onesite.df$Gasur==1,]$Site<-"Gasur"
  final.onesite.dt<-data.table(data.onesite.df,
  final.onesite.dt<-final.onesite.dt[order(Site)]
  keep.rownames=TRUE)[,.(rn, Site)] write.table(final.onesite.dt,
  file="hapax_signs.xls", quote=FALSE, sep="\t", row.names=FALSE)

  #table of signs that occur at all sites
  data.allsites.df<-data.df[rowSums(data.df[,-9])==9,-9]
  write.table(rownames(data.allsites.df), file="allsites_signs.xls",
    quote=FALSE, sep="\t", row.names=FALSE)

  #table of filtered data
  write.table(rownames(data2.df),file="data_filtered.xls", quote=FALSE,
    sep="\t", row.names=TRUE, col.names=TRUE)
  write.table(data2.df,file="data_filtered.xls",sep="\t")

  #hierarchical clustering
  colnames(data2.df)<-gsub("Esznunna","Eshnunna",colnames(data2.df))
  rclust<-hclust(dist(data2.df,method="manhattan"), method="ward.D2")
  cclust<-hclust(dist(t(data2.df),method="manhattan"), method="ward.D2")
  aheatmap(data2.df,color=‘grey:2’, Rowv=rclust,breaks=c(-0.05,0.5,1.05),
    labRow=NULL, legend=FALSE, fontsize=14, cexCol = 0.8)
  result <-pvclust(data2.df, method.dist="manhattan",
    method.hclust="ward.D2", nboot=10000)
  plot(result)

  #PCA
  colnames(data2.df)<-gsub("E?nunna","Eshnunna",colnames(data2.df))
  forpca<-t(data2.df)
  answer<-PCA(forpca,ncp=3,graph=FALSE)
  pc.eig.df<-data.frame(answer$eig)
  ggplot(data=pc.eig.df[c(1:8),], aes(x=gsub("comp","Comp.",
    rownames(pc.eig.df[c(1:8),])), y=percentage.of.variance)) +
    geom_bar(stat="identity")+xlab("\nPrincipal Component") +
    ylab("Percentage of Variance")+ggtitle("Variance distribution across
    principal components \n derived from informative syllabic values \n")
  coord.rs.df<-data.frame(answer$ind$coord)
  #meta.pca.df<-meta.df[rownames(coord.rs.df),]

  #PC1
  gsub("Esznunna","Eshnunna", row.names(coord.rs.df))
  Geography<-factor(c("Syria", "Syria", "Syria", "Syria", "Southern
    Mesopotamia", "Southern Mesopotamia", "Southern Mesopotamia", "Southern
    Mesopotamia", "Northern Mesopotamia"), levels=c("Southern Mesopotamia",
    "Northern Mesopotamia", "Syria"))
  Period<-factor(c("Old Akkadian\n(ca. 2350-2200 BC)", "\nUr III /
    Shakkanakku \n(ca. 2100-2000 BC)", "Old Akkadian\n(ca. 2350-2200 BC)",
    "\nEarly Old Babylonian\n(ca. 2000-1900 BC)\n", "\nOld Akkadian &
    Ur III \n(ca. 2350-2200, 2100-2000 BC)", "Old Akkadian\n(ca. 2350-2200
BC)",
    "Old Akkadian \n(ca. 2350-2200 BC)", "Old Akkadian\n(ca. 2350-2200 BC)",
    "Old Akkadian\n(ca. 2350-2200 BC)"), levels=c("Old Akkadian \n(ca. 2350-
2200 BC)",
    "\nOld Akkadian & Ur III\n(ca. 2350-2200, 2100-2000 BC)", "\nUr III /
    Shakkanakku\n(ca. 2100-2000 BC)", "\nEarly Old Babylonian \n(ca. 2000-
1900 BC) \n"))
  ggplot(data=coord.rs.df, aes(x=Dim.1, y=Dim.2))+geom_point(aes
    (shape=Period, fill=Geography, color=Geography),size=4)
```

```
        +scale_shape_manual(values=c(21,22,23,24))+xlab("Princ. Comp.
        1")+ylab("Princ. Comp. 2")+geom_text(label=gsub ("Esznunna","Eshnunna",
        row.names(coord.rs.df)),nudge_y=0.5)
    ggplot(data=coord.rs.df, aes(x=Dim.2, y=Dim.3))+geom_point(aes
        (shape=Period, fill=Geography, color=Geography),size=4)
        +scale_shape_manual(values=c(21,22,23,24))+xlab(
        "Princ. Comp. 2")+ylab("Princ. Comp. 3+geom_text(label=gsub
        ("Esznunna","Eshnunna", row.names(coord.rs.df)), nudge_y=0.5)
    ggplot(data=coord.rs.df, aes(x=Dim.1, y=Dim.3))+geom_point(aes(shape=Period,
        fill=Geography, color=Geography),size=4)+scale_shape_manual
        (values=c(21,22,23,24))+ xlab("Princ. Comp. 1")+ylab("Princ. Comp.
        3"+geom_text(label=gsub("Esznunna","Eshnunna", row.names (coord.rs.df)),
        nudge_y=0.5)

    #Extract attestations that define principal components 1-3
    signs<-data.table(answer$var$contrib, keep.rownames = TRUE)

    #dim1 excel table and visualization
    dim1<-signs[order(-abs(Dim.1))][,1:2,with=FALSE]
    ggplot(data=dim1, aes(x=seq(from=1, to=188,by=1),y=Dim.1))
        +geom_point(size=1, color="black")+xlab("\nSyllabic value index ordered
by
        loadings on the first principal component")+ylab("Loadings on the first
        principal component\n")+ggtitle("The distribution of loadings for
syllabic
        values suggests that\nloadings greater than 1.2 should be further
        examined.") +geom_hline(yintercept = 1.25)
    dim1<-dim1[dim1$Dim.1>1.1,]
    data2.dim1.df<-data2.df[dim1$rn,]
    colnames(data2.dim1.df)<-gsub("Esznunna","Eshnunna",colnames(data2.dim1.df))
    rclust<-hclust(dist(data2.dim1.df,method="manhattan"), method="ward.D2")
    cclust<-hclust(dist(t(data2.dim1.df),method="manhattan"), method="ward.D2")
    anngeo<-list(Geography=Geography)
    aheatmap(data2.dim1.df,color=‘grey:2’, Rowv=rclust,breaks=c(-0.05,0.5,1.05),
        annCol = anngeo, legend=FALSE,main="Hierarchical clustering of sites by
        syllabic value attestations\nimportant in the first principal
component",
        fontsize=10,treeheight=10, cexCol = 1, cexRow=2) write.table(dim1,
        file="dim1_signloadings.xls", quote=FALSE, sep="\t", row.names=FALSE)

    #dim2 excel table and visualization
    dim2<-signs[order(-abs(Dim.2))][,c(1,3),with=FALSE]
    ggplot(data=dim2, aes(x=seq(from=1, to=188,by=1),y=Dim.2)) + geom_point(
        size=1, color="black")+xlab("\nSyllabic value index ordered by loadings
        on the second principal component")+ylab("Loadings on the second
principal
        component\n")+ggtitle("The distribution of loadings for syllabic values
        suggests that\nloadings greater than 1.1 should be further
examined.")+geom_hline(yintercept = 1.1)
    dim2<-dim2[dim2$Dim.2>1.1,]
    data2.dim2.df<-data2.df[dim2$rn,]
    colnames(data2.dim2.df)<-gsub("Esznunna","Eshnunna",colnames(data2.dim2.df))
    rclust<-hclust(dist(data2.dim2.df,method="manhattan"), method="ward.D2")
    cclust<-hclust(dist(t(data2.dim2.df),method="manhattan"), method="ward.D2")
    annperiod<-list(Period=gsub("^ ","",gsub("\n"," ", Period)))
    aheatmap(data2.dim2.df,color=‘grey:2’, Rowv=rclust,breaks=c(-0.05,0.5,1.05),
        annCol=annperiod, legend=FALSE,main="Hierarchical clustering of sites by
        syllabic value attestations\nimportant in the second principal
component",
        fontsize=10,treeheight=10, cexCol = 1, cexRow=3) write.table(dim2,
        file="dim2_signloadings.xls", quote=FALSE, sep="\t", row.names=FALSE)


    #dim3 excel table and visualization
    dim3<-signs[order(-abs(Dim.3))][,c(1,4),with=FALSE]
    ggplot(data=dim3, aes(x=seq(from=1, to=188,by=1),y=Dim.3))
        +geom_point(size=1, color="black")+xlab("\nSyllabic value index ordered
        by loadings on the third principal component")+ylab("Loadings on the
third
        principal component\n")+ggtitle("The distribution of loadings for
syllabic
        values suggests that\nloadings greater than 1.3 should be further
        examined."+geom_hline(yintercept = 1.3)
    dim3<-dim3[dim3$Dim.3>1.3,]
    data2.dim3.df<-data2.df[dim3$rn,]
    colnames(data2.dim3.df)<-gsub("Esznunna","Eshnunna",colnames(data2.dim3.df))
    rclust<-hclust(dist(data2.dim3.df,method="manhattan"), method="ward.D2")
    cclust<-hclust(dist(t(data2.dim3.df),method="manhattan"), method="ward.D2")
    aheatmap(data2.dim3.df,color=‘grey:2’, Rowv=rclust,breaks=c(-0.05,0.5,1.05),
        legend=FALSE,main="HierarchicaL clustering of sites by syllabic value
attestations
        \nimportant in the third principal component", fontsize=10,
treeheight=10,
        cexCol = 1, cexRow=2)
    write.table(dim3, file="dim3_signloadings.xls", quote=FALSE, sep="\t",
row.names=FALSE)
```

Notes

[1] This article is based on certain aspects of my doctoral research completed at the University of Oxford in 2016.

[2] For the results of this analysis, see the forthcoming publication of my doctoral thesis (Hawkins *in preparation*).

[3] Not included in this study are the following nine sites: Nagar, Sheḫna / Shubat-Enlil, Umma, Shuruppak, Abu Salabikh, Nippur, Girsu, Umm al-Jir, and Susa.

[4] This was obtained from the Cuneiform Digital Library Initiative (www.cdli.ucla.edu), unless otherwise stated.

[5] http://virgo.unive.it/eblaonline/cgi-bin/home.cgi

[6] These genres are based on those provided in the Cuneiform Digital Library Initiative Database.

[7] ca. 2350-2200 BC

[8] The majority of the texts from Ebla used here are lexical and administrative. As an attempt to address the possibility that the larger number of lexical texts from Ebla could skew the results of the analysis, the data sets were filtered to exclude any *hapax* sign values, or values that occurred at only one site (with the majority of these *hapax* signs coming from Ebla). For a further explanation of the filetered and unfiltered datasets, see Section 4.2 below. For more about the numbers of texts attested within each genre, see the following resources: Krebernik (1982-1983), Conti (1990), and the Ebla Digital Archives.

[9] ca. 2100-2000 BC

[10] ca. 2000-1900 BC

[11] Southern Mesopotamia.

[12] These are the numbers of texts from the Old Akkadian and Ur III periods, respectively.

[13] These are the numbers of texts within each genre from the Old Akkadian and Ur III periods, respectively.

[14] Northern Mesopotamia.

[15] For an introduction to phylogenetic techniques, see [Nichols and Warnow 2008].

[16] See Section 12 below for the code used for the phylogenetic analysis, hierarchical clustering, and principal component analysis.

[17] See [Swofford et al. 1996, 415–24] for a full discussion of different parsimony methods.

[18] Other types of consensus trees are majority consensus and greedy consensus. According to [Nichols and Warnow 2008, 773], “the majority consensus tree contains those edges whose corresponding bipartitions appear in strictly more than half of the input trees, and the greedy consensus tree is formed by computing the majority consensus and then refining the tree by adding bipartitions from the input trees. By construction, the strict consensus tree is the least resolved, the greedy consensus tree is the most resolved, and the majority consensus is in between these two trees with respect to resolution. Also, however, the greedy consensus tree refines the majority consensus tree, and the majority consensus tree refines the strict consensus.”

[19] Ebla, Mari, Nabada, Tuttul, Adab, Eshnunna, Kish, Tutub, Assur, and Gasur.

[20] Identifying an outgroup is desirable but not necessary. For more about outgroups and ingroups, see [Swofford et al. 1996].

[21] See Hawkins *in preparation* for a further discussion of this finding [Hawkins forthcoming].

[22] For more about clustering techniques, see L. Kaufman and P. J. Rousseeuw 1990 [Kaufman and Rousseeuw 1990].

[23] Clusters with a p-value of 95% or higher are considered to be very strong, while p-values higher than 90% are considered strong [Suzuki and Shimodaira 2014, 6].

[24] For more about principal component analysis, see [Jolliffe 2002] [Wold et al. 1987].

[25] This is explored further in a forthcoming publication [Hawkins forthcoming].

[26] This is explored further in the forthcoming publication of my doctoral dissertation [Hawkins forthcoming]).

[27] One of the two prominent languages recorded in these documents was Akkadian – which is, in fact, not a single language but an umbrella term for a series of closely related Semitic dialects that were written and spoken primarily in the regions of modern-day Iraq, Syria, and Turkey from around 2500 BC until 70 AD [Cooper 1996, 37]. The Akkadian dialects, along with their relative Eblaite, form the entirety of the eastern branch of the Semitic language family.

[28] The data of this study would consist of the Swadesh 200-word list [Swadesh 1955] (Swadesh 1955).

Works Cited

Atkinson and Gray 2005 Atkinson, Q. D. and Gray, R. D. “Curious parallels and curious connections–phylogenetic thinking in biology and historical linguistics.” *Systematic Biology* 54, no. 4 (2005): 513-526.

Barbrook et al. 1998 Barbrook, A. C. et al. “The phylogeny of the Canterbury Tales.” *Nature* 394 (1998): 839.

Bastin 1983 Bastin, Y. *La finale verbale-ide et l’imbrication en bantou*. Tervuren: Musée Royal de l’Afrique Centrale (1983).

Biggs 1973 Biggs, R. D. “On Regional Cuneiform Handwritings in Third Millennium Mesopotamia.” *Orientalia* 42 (1973): 39–46.

Conti 1990 Conti, G. “Il sillabario della quarta fonte della lista lessicale bilingue Eblaita.” *Miscellanea Eblaitica 3*. Ed. by P. Fronzaroli. Vol. 3. Quaderni di Semistica 17 . Firenze: Dipartimento di Linguistica Università di Firenze (1990).

Cooper 1996 Cooper, J. S. “Sumerian and Akkadian.” *The world’s writing systems*. Ed. P. T. Daniels and W. Bright. Oxford University Press (1996).

Cysouw et al. 2006 Cysouw, M., Wichmann, S., and Kamholz, D. “A critique of the separation base method for genealogical subgrouping with data from Mixe-Zoquean.” *Journal of Quantitative Linguistics* 13, no. 2-3 (2006): 225–264.

Faber 1997 Faber, A. “Genetic Subgrouping of the Semitic Languages.” *The Semitic Languages*. Ed. by R. Hetzron. London: Routledge (1997): 3–15.

Felsenstein 2005 Felsenstein, J. *PHYLIP (phylogeny inference package)*. Department of Genome Sciences, University of Washington. Distributed by author (2005).

Goloboff et al. 2003 Goloboff, P., Farris, J., and Nixon, K. “TNT: Tree Analysis Using New Technology.” Program and documentation available from the authors (2003).

Gray 2003 Gray, R. D. and Atkinson, Q. D. “Language-Tree Divergence Times Support the Anatolian Theory of Indo-European Origin.” *Nature*. 426, 435-439 (2003).

Hawkins forthcoming Hawkins, L. F. *The Adaptation of Cuneiform to Write Semitic: an examination of the use of syllabic sign values in late third and early second millennium Mesopotamia and Syria*. Wilbour Studies in Egyptology and Assyriology: Lockwood Press (*in preparation*).

Hetzron 1969 Hetzron, R. “La division des langues sémitiques.” *Actes du premier congrès international de linguistique sémitique et chamito-sémitique*, Paris 16-19 juillet, 1969. Ed. by A. Caquot and D. Cohen. Paris: Mouton (1974): 181–194.

Hetzron 1976 Hetzron, R. “Two principles of genetic reconstruction.” *Lingua* 38.2 (1976): 89–108.

Holden 2002 Holden, C. J. “Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis.” *Proceedings of the Royal Society of London B: Biological Sciences* 269, no. 1493 (2002): 793–799.

Holden et al. 2005 Holden, C. J., Meade, A., and Pagel, M. “Comparison of maximum parsimony and Bayesian Bantu language trees.” *The Evolution of Cultural Diversity: a phylogenetic approach*. Ed. by R. Mace, C. J. Holden, and S. Shennan. London, UK: University College London Press (2005): 53–66.

Huehnergard 2011 Huehnergard, J. *A Grammar of Akkadian.* 3rd ed. Scholars Press: Atlanta, GA (2011).

Ismail et al. 1996 Ismail, F. et al. *Administrative Documents from Tell Beydar (Seasons 1993-1995)*. Subartu 2. Turnhout: Brepols (1996).

Jolliffe 2002 Jolliffe, I. *Principal Component Analysis*. John Wiley and Sons, Ltd (2002).

Kaufman and Rousseeuw 1990 Kaufman, L. and P. J. Rousseeuw. *Finding Groups in Data: an Introduction to Cluster Analysis.* Wiley, 1990.

Kitchen et al. 2009 Kitchen, A., Ehret, C., Assefa, S., and Mulligan, C. J. “Bayesian Phylogenetic Analysis of Semitic Languages Identifies an Early Bronze Age Origin of Semitic in the Near East.” *Proceedings of the Royal Society of London B: Biological Sciences, 276* (2009): 2703-2710.

Krebernik 1982/3 Krebernik, M. “Zu Syllabar und Orthographie der lexikalischen Texte aus Ebla. Teil 1; Teil 2 (Glossar).” *Zeitschrift für Assyriologie und Vorderasiatische Archäologie* 72-73 (1982): 178–236; (1983): 1–47.

Krebernik 2001 Krebernik, M. “Ausgrabungen in Tall Bi’a/Tuttul - II: Die altorientalischen Schriftfunde.”” WVDOG 100 (2001).

Limet 1976 Limet, H. *Textes administratifs de l’époque des Šakkanakku*. Archives royales de Mari, Transcriptions et traductions XIX. = ARMT; ersch. 1977. Paris: Geuthner (1976).

Maddison and Maddison 2001 Maddison, W. P. and Maddison, D. R. *Mesquite: a modular system for evolutionary analysis*. http://mesquite.biosci.arizona.edu/mesquite/mesquite.html (2001).

Marten 2006 Marten, L. “Bantu classification, Bantu trees and phylogenetic methods.”” In: Foster, Peter and Renfrew, Colin, (eds.), *Phylogenetic Methods and the Prehistory of Languages. Cambridge: McDonald Institute for Archaeological Research* (2006): 43-55.

Nakhleh et al. 2005a Nakhleh, L., Ringe, D., and Warnow, T. “Perfect phylogenetic networks: a new methodology for reconstructing the evolutionary history of natural languages.” *Language* 81 (2005): 382–420.

Nakleh et al. 2005b Nakhleh, L. et al. “A comparison of phylogenetic reconstruction methods on an Indo-European dataset.” *Transactions of the Philological Society* 103, no. 2 (2005): 171–192.

Nichols and Warnow 2008 Nichols, J. and Warnow, T. “Tutorial on computational linguistic phylogeny.” *Language and Linguistics Compass* 2, no. 5 (2008): 760–820.

Platnick and Cameron 1977 Platnick, N. I. and Cameron, H. D. “Cladistic methods in textual, linguistics, and phylogenetic analysis.” *Systematic Biology* 26, no. 4 (1977): 380-385.

Rexovà et al. 2003 Rexovà, K., Frynta, D., and Zrzavỳ, J. “Cladistic analysis of languages: Indo-European classification based on lexicostatistical data.” *Cladistics*19, no. 2 (2003): 120–127.

Rexovà et al. 2006 Rexovà, K., Bastin, Y., and Frynta, D. “Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data.” *Naturwissenschaften* 93, no. 4 (2006): 189–194.

Skelton 2008 Skelton, C. “Methods of using phylogenetic systematics to reconstruct the history of the Linear B script.” *Archaeometry* 50, no. 1 (2008): 158–176.

Suzuki and Shimodaira 2014 Suzuki, R and Shimodaira, H. *pvclust: An R package for hierarchical clustering with p-values*. Tech. rep. Division of Mathematical Science, Graduate School of Engineering Science, Osaka University (2014).

Swadesh 1955 Swadesh, M. “Towards greater accuracy in lexicostatistic dating.” *International Journal of American Linguistics* 21.2 (1955): 121–137.

Swofford 2001 Swofford, D. L. “PAUP*: Phylogenetic analysis using parismony (and other methods).” *4.0.b5* (2001).

Swofford et al. 1996 Swofford, D. et al. “Phylogenetic inference.” *Molecular Systematics*. Ed. by D. M. Hillis, C. Moritz, and B. K. Mable. 2nd ed. Vol. 2. Sinauer: Sunderland, MA (1996): 407–514.

Wold et al. 1987 Wold, S., Esbensen, K., and Geladi, P. “Principal Component Analysis.” *Chemometrics and intelligent laboratory systems* 2, no. 2-3 (1987): 37–52.