Sara Brumfield*

The Key to the City: Using Digital Tools to Understand Tablet Provenience

https://doi.org/10.1515/janeh-2018-0012

Abstract: Assyriologists have a variety of methods available to assign unprovenienced materials with educated certainty to its ancient site. The occurrence of specific toponyms and month names as well as the detailed study of prosopography, paleography, orthography, lexicography, tablet shape, format and sealing practices assist specialists in reconstructing the ancient context of a specific object. Now, with the fluorescence of technology, new digital tools are being developed and refined that may contribute to the complex process of provenience assignment. Text mining, the practice of deriving information from blocks of text using pattern recognition or trend analysis, has already been applied to corpora ranging from Shakespeare to Twitter.[1] With the ability to search for statistically significant correlations in large blocks of text following user-defined criteria and rules, statistical methods, here accessed via text mining software, have significant potential for revealing new levels of data in cuneiform texts.

Keywords: digital humanities, Old Akkadian, provenience, text mining

Text mining tools, and digital tools more generally, are built upon researchers' expert insights into the data. In preparing ancient texts for modern analysis, a researcher still must make certain interpretive choices before text mining can be applied. In this sense, some of the problems of assigning provenience remain.[2] For example, should variant orthographies be retained to detect potentially meaningful variation or combined in a single lemma to mitigate individual scribal predilections or abbreviations? Are differences in unqualified names regional variations or indicative of different individuals? These are a few of

1 For an example of previous text mining analysis on cuneiform sources, see, ENEA's TIGRIS Virtual Lab (http://www.afs.enea.it/project/tigris/indexOpen.php)

2 Other issues arise when a tablet's script and shape do not correlate to the same time period or when two different writing styles exist within one text (Maiocchi 2015: 81–82; Yang 1989: 39). These problems are not as easily addressed by text mining methodologies as orthography, lexicography, prosopography, month names and specific geographic terms.

***Corresponding author: Sara Brumfield,** Independent Researcher, Washington, DC, USA, E-mail: brumfield@ucla.edu

the decisions Assyriologists must make before applying any digital tools to a corpus, and the outcome of these decisions affects the results. The methodology proposed here does not argue for a specific approach but instead lends statistical confidence to certain assertions by assessing the probability that certain observed similarities (or differences) are not due to mere chance.

1 Statistics, corpus linguistics and keywords

Using statistical methods in the comparison of two corpora has been growing in the field of linguistics, specifically corpus linguistics, a methodology of applied linguistics. For example, in this methodology at least one of the two corpora involved in comparisons is a large, standardized bank of English words, such as the British National Corpus (BNC), American National Corpus (ANC) or the Collins Birmingham University International Language Database (COBUILD). These data sets are typically comprised of millions of words collected in the 1980s and 1990s from a broad cross-section of materials. With these large language repositories, researchers have applied various statistical algorithms to better understand language distribution and usage. Linguists compare newspapers, student papers, emails, internet sites, government papers and myriad other contemporary sources to answer questions about significant differences in how people use language. This type of analysis relies upon keywords, that is a lexical item with unusually higher or lower frequency in either the reference or test corpus, and as such is called keyword analysis. These keywords are typically then a starting point for further inquiry into defining a genre or register, identifying communication styles of specific groups/contexts, isolating trends in language usage over time, etc. However, these questions are guided by the contemporary English corpora they draw from. For Assyriology, the application of keyword analysis expands beyond inquiries into sociolinguistics to identifying meaningful similarities or differences between text corpora that can also help address questions about provenience.

The methodology presented here tests the unprovenienced "Diyala" administrative texts[3] from the Classical Sargonic period[4] against administrative texts

3 Administrative includes all business documents, including legal texts in order to best approximate the genre of the unprovenienced "Diyala" texts, particularly those published in I. J. Gelb's *Old Akkadian Inscriptions in Chicago Natural History Museum; Texts of Legal and Business Interest* (*OAIC*). I appreciate W. Sommerfeld sharing with me his notes for his re-edition of the *OAIC* texts; many of his improved readings are included in the data here.

4 Classical Sargonic defined by changes in and standardization of paleography and orthography, alterations to tablet shape and layout (including the possible rotation of the tablet

from surrounding Diyala sites (Ešnunna, Tutub, Tell Suleimah) as well as the nearby northern site of Kiš from the Classical period.[5] Similar Old Akkadian administrative texts from Girsu are used as the control group in order to assess levels of (dis)similarities between sites based on personal names, terminology, language, toponyms, titles, commodities, etc.[6]

Keyword analysis alone is insufficient to assign provenience, but, when coupled with well-defined data from other proven techniques, can help elucidate extant analysis. The approach is exploratory rather than focused—it generates hypotheses instead of being guided by them (Gabrielatos 2018: 227). The results presented herein should not be taken as a definitive answer, but as one of several methods of evaluation that can be combined with more traditional methods. As a relatively new tool to Assyriology both the technical and theoretical aspects are outlined below.

2 Keyword analysis

To assist linguists and philologists with statistical analysis of texts there are several software packages available, but AntConc was selected for this study specifically for its multiplatform capabilities, user-friendly interface, *gratis* price tag and its ability to process transliterated, non-English texts.[7] While corpus linguistics has grown beyond English-based corpora, working with non-English texts presents unique challenges in keyword analysis. Fortunately, with the standardized formatting of texts on the Cuneiform Digital Library Initiative (CDLI), corpus creation is relatively easy in

orientation, for which see, Studevent-Hickman 2007: 494–499). The widespread standardization of metrology (and possibly the menology) system in conjunction with the systemization of volume and capacity measures are also key indicators of the Classical Sargonic period, assigned to the reigns of (latter) Narām-Suen and Šar-kali-šarrī.

5 Sippar, Isin and Mugdan (umm el-Jīr) are excluded here due to the small number of texts.

6 It is important to note, however, that contemporary archives from the same site may contain considerably different vocabulary, content, phrases and personal names, leading to the false conclusion that their provenience is unrelated. Therefore, it is generally suggested in the methodology here to include larger, site-wide data sets that "average" out archival differences as well as pairing keyword analysis with already established methods for understanding provenience.

7 This software has been developed by Laurence Anthony, Professor of Applied Linguistics at Waseda University (Japan) and Director of the Center for English Language Education in Science and Engineering. Version 3.4.4 was used for the provenience test presented here and is available for download here: http://www.laurenceanthony.net/software/antconc/.

Assyriology. In comparing such texts, consistency in spelling, size and genre are crucial (Rayson et al. 2004: 1–2).[8]

For Assyriological texts there is often variation in preferred sign readings, which requires some attention before any keyword analysis. Since the word is the unit of analysis, it is mandatory to standardize sign readings that comprise all words. The precise approach adopted is not as important as the consistency of the system. For the Old Akkadian texts, there are two main conflicting approaches to transliteration—that of I. J. Gelb and that of W. von Soden.[9] With the Classical Akkadian texts included here, where known, the appropriate value (voiced, voiceless or emphatic)[10] was transliterated. This permits the etymological clarity of von Soden's system but relies on Gelb's consistency for uncertain readings.[11]

For the transliterated data, non-meaningful elements, such as metadata, headers, line numbers, tags, language shift markers and commentary, should be stripped from blocks of texts (downloadable directly from CDLI or created

8 The sampling method must also be controlled for, but with such small corpora, I have elected to include all texts in a given genre from each site included in this study. In ancient texts there will always be issues regarding the chance of discovery and preservation, which cannot be controlled for, especially to the degree it can in modern languages.

9 A thorough evaluation of the advantages and disadvantages of each of these two transliteration systems is given in Sommerfeld (1999: 24–25). While I. J. Gelb prefers the basic reading as defined by the first value given to the sign in the Neo-Assyrian period, more than a millennium after the end of the Old Akkadian period (Gelb 1970b: 534), W. von Soden proffers a linguistic interpretation through his rendering of the signs in a close approximation to the actual, estimated pronunciation (e. g. GA is *ga, ka*$_3$ or *qa*$_2$). This approach results in a plethora of diacritics and reconstructed, hypothetical phonemes, which lends itself to unnecessary confusion (Hasselbach 2005: 24–25; Sommerfeld 1999: 24; Westenholz 1996: 119–120). Each approach entails its own set of interpretive issues, which cannot be resolved here.

10 In the case of the PI sign, the variation is between labial and glide.

11 This compromise follows the approach of Sommerfeld (1999: 25). The Old Akkadian sibilants present a particular problem, especially in light of the confusion between the signs SA and ŠA by Old Akkadian scribes (Westenholz 1996: 120). The number and nature of Old Akkadian sibilants is an undecided matter, yielding several contradictory paradigms. The exact relationship between the sign name, later Old Babylonian pronunciation and Old Akkadian pronunciation is murky (Faber 1981, 1985; Hasselbach 2005: 95–96). Gelb, von Soden, Hasselbach and Sommerfeld have each offered additional reconstructions for the sibilant system in the Old Akkadian texts, but again, the issue here is more of consistency than phonological reality. Therefore, when transliterating the sibilants in the Old Akkadian corpus, minimal interpretation is preferred, leaving the basic reading of the sign unaltered. However, there are cases of clear and continuous etymology, particularly in verbal forms, where an interpretation is made (e. g. *si*$_2$ instead of *zi* in *na-si*$_2$*-iḫ* [*nasāḫum* "to tear out"] and *ip-lu-si*$_2$ [*palāsum* "to look at"]). Because this is not a study of Old Akkadian linguistics or phonology, adherence to the overall tradition of the writing system is preferred.

manually). It is particularly important for transliterations of ancient texts to address broken, questionable or emended readings. For example, x-readings and altered "!"-readings for signs were emended to their intended contextual reading (e. g. $tum_x \rightarrow tum_2$; zu!(SU) → zu) in order to recover the intended lexeme independent of orthographic variations.[12] The reordered sign sequences marked with ":" were normalized and connected with the standard sign connector "-". Other elements may be erased or altered, depending on the research question and parameters. In this case, since the issue of provenience relies upon levels of (dis)similarity in words within a corpus, all quantities were erased from the text files. And, in order to remove broken passages, AntConc possesses a Stop Wordlist feature that allows the user to define "words" that should not be included in the analysis, such as x-i_3-li_2, ...-dingir.[13]

An example of this transformation is presented below with the original ATF on the left and the cleaned version on the right[14]:

&P212832 = BIN 08, 288	
#atf: lang sux	
@tablet	
@obverse	
1. 1(gesz2@c) la2 2(asz@c) gurusz#	gurusz
2. lu2 gub-ba-a#	lu2 gub-ba-a
3. 1(gesz2@c) 2(asz2@c) ki szu-ix(ASZ3) er2-du8	ki szu-i er2-du8
4. 5(asz@c) ki uz-ga	ki uz-ga
5. 2(u@c) 1(asz@c) ki gesz-i3	ki gesz-i3
6. 5(asz@c) ki gu4 niga	ki gu4 niga
7. 2(asz@c) ur-{d}inanna	ur-{d}inanna
8. 1(asz@c) i3-du8 e2-ansze	i3-du8 e2-ansze
9. 1(asz@c) i3-du8 tum-x	i3-du8 tum-x
10. [n] 1(asz@c) e2 [...]	e2 ...
@reverse	... in ...

12 These orthographic variations may also contribute to discussions about tablet provenience, however, this can also be done on a case by case basis after the keyword analysis. After all, the keywords often indicate areas for deeper analysis, devoid of innate explanatory power—they simply are more or less frequent in one of the two compared corpora. The researcher must then apply contextual information to interpret the lexical pattern identified through keywords.

13 Breaks in the text are maintained in order to preserve distance between words on the original text. This proves necessary in follow-up analyses where the context of certain keywords needs to be accurately represented.

14 All transliterations were downloaded from CDLI (https://cdli.ucla.edu), where 1,820 Old Akkadian transliterations were contributed by the author.

1. [...] in [...]	sze ...
2. 1(u@c) sze [...]	tu-ra
3. 6(asz@c) tu-ra#	ba-usz2
4. 4(disz) ba-usz2	
$ blank space	|SZU + LAGAB| gurusz
5. |SZU + LAGAB| 3(gesz2@c) 1(u@c) la2 1(asz@c) gurusz	su-bir-x
6. su-birx(|SZIMxNIG2|)-x	

This cleaned data set can now be run through an algorithm to create a ranked raw frequency list of all words within the corpus. However, there may be instances when grammatical variation is not as crucial as other variables; for example, in deciding provenience, it is not particularly relevant whether a verb is in the third person masculine plural or singular. Therefore, words should be lemmatized so that variants can be counted as the same "word."[15] All these measures are taken to try to ensure accurate raw frequency counts for each word in the two corpora, guided by the nature of the specific research question, provenience in this case.

There is no standard statistical method currently for keyword analysis, however each has its own strengths and weaknesses suited to specific purposes (Pojanapunya and Watson Todd 2016: 2). However, all approaches to keyword analysis are based on the frequency of each word in both the reference and test corpus compared against the total number of words in each corpus in order to determine if any difference in relative frequency is due to chance. There are generally two different statistical tests for determining a word's keyness:

15 In this study, all lemmatization was done manually. For example, im-hur, im-hu-ra, im-hu-ru, im-hur-ra and tam_2-hur may all be counted as "maḫārum" or whatever lemma the user defines. This lemmatization extends to titles and toponyms (e. g. $ARAD_2$ and ARAD; $azlag_2$, $azlag_3$, $azlag_4$; dumu-me, dumu, dumu-dumu, $umma^{ki}$, $umma^{ki}$-ta, etc.) with more certainty than to personal names. Variation in personal name spelling often cannot, with any degree of certainty, be correlated to the same individual because of the potentially high degree of homonymy. Therefore, small differences in spelling could reflect crucial distinctions in pronunciation or mere regional orthographic conventions. It is difficult to ascertain either way without context. Therefore, personal names should be left unlemmatized unless a demonstrated correlation has already been established through other analysis (e. g. *šar-ru*-GI, *šar-um*-GI, *šar-rum*$_2$-GI are, from clear context, variant spellings of Šarru-kēn; see Westenholz 1999: 34; Kienast and Sommerfeld 1994: 62–64). Each user must carefully consider which words to lemmatize and how that will affect the outcome of the test. For this first pass at the data, I have chosen to lemmatize all observed variations in verbs, offices, terminology, grammar particles and toponyms in order to focus on broad trends in the data. Follow-up research should focus on more refined nuances in the data.

significance tests (based on probability statistics) and effect-size tests. Significance test statistics calculate the probability of a frequency difference (i. e. the confidence we can have in the difference being not random) while effect-size statistics focus on the size of the difference between a word's frequency in the two compared corpora. For the purposes of assigning provenience, this study is more interested in whether the frequency of a word between two corpora can be deemed statistically significant than measuring the size of the frequency difference. Therefore, the probability statistics are preferred.

Probability statistics calculates a p-value that indicates the probability that the difference in a word's frequency in two different corpora is due to chance. Essentially the smaller the p-value, the more likely the result is not due to chance. The threshold for assigning statistical significance is arbitrary, however the standard threshold for corpus linguistics is $p = 0.01$ (Gabrielatos and Marchi 2012; Gabrielatos 2018: 241).

There are two common probability statistics that produce very similar results in keyword analysis: log-likelihood and chi-square (Pojanapunya and Watson Todd 2016: 13).[16] For several reasons, the log-likelihood is preferred. Chi-square becomes unreliable for low frequency words (those occurring fewer than five times) and in small corpora (those with fewer than 50,000 words) (Dunning 1993; Rayson et al. 2004: 3). This has serious implications for results from many Assyriological corpora, including the Old Akkadian texts studied here. Conversely, the log-likelihood is sensitive to corpus size, making results inconsistent across corpora of different sizes (Pojanapunya and Watson Todd 2016: 28). However, this can be controlled for if the level, or threshold for identifying statistical significance, is raised to 0.01 % (= 15.13 critical value) (Rayson et al. 2004: 8).[17]

16 The log-likelihood is an algorithm that measures the probability that a set of data would occur naturally. The algorithm is defined as: $L(\theta|x) = P(x|\theta)$ where the likelihood (L) of the specific parameters (θ) is determined by the outcome(s) (x). This is equal to the probability (P) of the observed outcomes (x) given the specific parameters (θ).

Chi-square is defined as: $\chi^2 = \sum (O-E)^2/E$ where O is the observed frequency and E is the expected frequency.

New statistical methods are continually being developed and refined, however, have not yet been widely tested (e. g. Gabrielatos and Marchi 2012; Lijffijt et al. 2014; Pojanapunya and Watson Todd 2016). Therefore, until these new algorithms are tested in a variety of corpora to demonstrate their consistency, it is best to adhere to those models that have already been thoroughly vetted in corpus linguistics.

17 This means that there is a 0.01 % chance that we would obtain a similar (or larger) statistically significant result when there is no real, genuine difference (Gabrielatos 2018: 231).

Table 1 shows the correlation between the probability of word's frequency in the two compared corpora and its statistical significance. The p-value corresponds to a critical value that is used as a measure of keyness for each word. A word's keyness quantifies its uniqueness in either the test corpus (positive values) or reference corpus (negative values). Negative values indicate a high frequency in the reference corpus but a lack of corresponding frequency in the test corpus. Contrariwise, positive values rank words that are more common in the test corpus than the reference corpus. The closer the keyness measurement (critical value) is to zero, the more likely differences in word frequency are merely due to chance; and the fewer words in the keyword list, the more similar the two corpora. Each word (including lemmatized words) generates its own keyness measurement (critical value). And words whose keyness values are above the threshold for statistical significance (15.13) are deemed interesting, important or worthy of further analysis. The interpretation of the results departs from the statistical method into the realm of contextual analysis.

Table 1: Significance Values.

Percentile	Level	p-Value	Critical Value/Keyness
95	5 %	< 0.05	3.84
99	1 %	< 0.01	6.63
99.9	0.1 %	< 0.001	10.83
99.99	0.01 %	< 0.0001	15.13

3 The text corpora

The unprovenienced "Diyala" texts serve as the test corpus against which reference corpora from Ešnunna, Tutub, Tell Suleimah, Kiš and Girsu are compared to detect levels of lexical (dis)similarity. The data set is circumscribed by time period and genre in order to control for variation, insofar as is possible with such texts.

The 85 unprovenienced tablets attributed to the "Diyala" corpus here include 51 of the 53 tablets published by Gelb in *OAIC*, which he generally attributed to the Diyala but without detailed information on their exact provenience.[18] This data set is rounded out by a medley of additional texts that have

18 The letters *OAIC* 52–53 are omitted based on their genre.

been published in various editions and journals, collected here as a single corpus.[19] Some tablets originally assigned a generic "Diyala" provenience have been associated, with some confidence, to Ešnunna through traditional techniques and are, therefore, not included in the "Diyala" corpus here.[20]

Using traditional methods, Gelb, P. Steinkeller and J. N. Postgate, and B. Kienast and K. Volk have linked some of these unprovenienced texts with Ešnunna.[21] Generally, the presence of personal names from excavated or secure Ešnunna texts,[22] specific geographic labels or the use of the deity Tišpak, the city deity of Ešnunna, is invoked as evidence for a tablet's origin at Ešnunna.

Additionally, Gelb cites the use of specific vocabulary such as *šibšum* and *kušurrā'im* (see Table 2). However, despite the similarity in the word choice, the orthography of the same term varies between the excavated Ešnunna texts and

19 *AuOr* 9, 6–9 (MM 526, 697, 560, 937); *CUSAS* 13, 161; *JCS* 26, 7; *JCS* 35, 168, 1 (AIA 4); *MAD* 4, 2–9; *MC* 4, 51; *MVN* 3, 27, 38, 57, 60, 65, 78–80, 83, 102 (= *RA* 74, p. 179), 111; *MVN* 9, 192–194; *SAKF* 2; *UCP* 9/2, 76, 83, 89.

P. Steinkeller's suggestion that various texts published in *MVN* 3 could be attributed to the Diyala region appears to be based on their linguistic affiliation. Of the fourteen Old Akkadian texts written in the Akkadian language, he posits eleven could potentially be from the Diyala area (Steinkeller 1982: 366). This association is almost exclusively based on the appearance of personal names in the *MVN* 3 texts that are popular in the Diyala region. The duplicate account of *MVN* 3, 57, *MAD* 4, 16, possessed no accompanying provenience information in the Louvre catalogue and was left tentatively unassigned by Gelb, although he suggested Nippur in place of Umma as the tablet's origin (Gelb 1970a: xviii).

The Louvre texts in *MAD* 4 are formally unprovenienced, but Gelb remarks that the internal museum catalog lists "de Tell Asmar?" for this lot of tablets acquired in 1923, prior to formal excavations by the Oriental Institute (Gelb 1970a: viii).

20 The list of excluded unexcavated "Diyala" texts is: *AuOr* 9, 4–5 (MM 401 = *OrNS* 51, p. 362; MM 497 = *AnOr* 7, 372); *JCS* 26, 8; *JCS* 28 227 (NBC 10,920); *MAD* 1, 270–336; *OrNs* 51, p. 355; *MC* 4, 50 (= *OIP* 104, 245).

21 Westenholz argues that given the extensive looting from the robber hole at Ešnunna prior to the Oriental Institute's excavations in the 1930s, all purchased tablets originating from the generically defined "Diyala region" should be attributed to Ešnunna (Westenholz 1984: 19, fn. 4). A useful observation, but one that must be combined with additional evidence to make the assignment conclusive.

22 Assigning certain personal names to Ešnunna instead of other Diyala sites is an inexact science. Only general observations, such as the popular use of Utu and Mama in personal names at Ešnunna compared to Nārum, Dagān and Suen at Tutub, can be maintained. Also, the movement of persons between sites makes the direct correlation between a person and a city circumstantial. It is only one of several pieces of evidence used to make the case for tablet provenience.

those subsequently attributed to the site, leaving the correlation tenuous.[23] Therefore, the texts originally assigned to Ešnunna based on this lexical evidence by Gelb are included in the "Diyala" corpus here.

Table 2: Common Vocabulary Between the Ešnunna and "Diyala" Texts.

Lexeme	Orthography	Text
šibšum	*ši-ib-ši-im*	*MAD* 1, 2 (excavated from Ešnunna)
	si-ib-su-um	*MAD* 1, 35 (excavated from Ešnunna)
	si-ib-šum	*MAD* 4, 3
	si-ib-šum	*MAD* 4, 9
kusurrā'im	ku_8-*sur-ra-im*	*MAD* 1, 179 (excavated from Ešnunna)
	ku_8-su_4-*ra-im*	*MAD* 4, 4
	ku_8-*su-ra-im*	*OAIC* 4

The collection of tablets published in *MAD* 1 as nos. 270–336 were assigned an Ešnunna provenience by Gelb based on "[b]oth the information given by the dealer from whom the collection was purchased and the internal evidence culled from the tablets," specifically the co-occurrence of personal names with excavated Ešnunna tablets (Gelb 1952: xi). There are some additional interrelations within this corpus that help assign specific texts to a provenience. The text *AuOr* 9, 5 mentions *i-da*-dingir šabra e_2 ("chief administrator of the household"), who is also mentioned with full title in *MAD* 1, 322, a text confidently associated with Ešnunna. The text *AuOr* 9, 5 also mentions *u-ṣi-um* gal-sukkal dingir ("chief sukkal of the deity"), who is present with this same qualifier in *JCS* 28, 227 (NBC 10,920). A sealing was excavated from Ešnunna with his cylinder seal impression: *u-ṣi-um* gal-sukkal ^{d}tišpak,[24] demonstrating the reasonable provenience of these two texts to Ešnunna. However, these contextual elements have only limited implications for the remainder of the *AuOr* 9 texts, which were purchased by P. B. Ubach in Iraq between 1922–1923, possibly in separate lots (Molina 1991: 137).[25]

23 This could be due to a number of causes: different scribal traditions, temporal distance between exemplars, register (official vs. vernacular pronunciation or spelling) to name a few obvious choices.

24 *OIP* 72, no. 593 (As.32:711b). Find spot was given as J 19:48, Houses IVb, which places it in close association with *MAD* 1, 177–179 and 181. *MAD* 1, 178 discusses slaughtered animals for the deity Ninbarre, which accords well with the seal of the temple official.

25 The personal names mentioned in the remaining *AuOr* texts correspond more closely with those known from Ešnunna.

Steinkeller and Postgate have demonstrated the assignment of *OrNS* 51, p. 355 to Ešnunna and that text's close relationship to *AuOr* 9, 4. In a subsequent publication Steinkeller also illustrated the connection between *MC* 4, 50, *JCS* 26, 8 and *MAD* 1, 336 through various land sale transactions of Dabālum (Steinkeller and Postgate 1992: 88–89).[26] The internal coherence of these four texts supports an Ešnunna provenience for all four texts given *MAD* 1, 336's probable origin from the site.

This medley of unexcavated texts is excluded from the "Diyala" corpus here based on the confluence of evidence suggesting their likely provenience from Ešnunna. This process of defining an unexcavated corpus illustrates the human influence in the statistical outcome and is a part of the process that should continue to be refined and improved. To a point, the human element is unavoidable, however, understanding user-created biases is the first step toward reconciling them.

4 Keyness criteria for estimating provenience

The keyness value may assist in determining how similar the unprovenienced "Diyala" texts are with each of the other five corpora from Ešnunna, Tutub, Tell Suleimah, Kiš and Girsu (see Table 3 for an overview of the corpora).

Table 3: Corpora Size.

Site	Number of Tablets	Number of Words
"Diyala"[27]	85	732
Eshnunna	196	966
Tutub	66	521
Tell Suleimah	47	430
Kiš	70	512
Girsu	721	2,296

26 Steinkeller's inclusion of *OAIC* 2 in this group is problematic, and, therefore, omitted here, since the personal name is written *da-bi-lum*, which Gelb claims is a short form of *i-da-bi*$_2$*-i*$_3$*-li* (Gelb 1955: 192). Furthermore, there are no personal names in *OAIC* 2 that clarify its origin. In short, the text does not share the internal coherence of the other four texts.

27 It is ideal to circumscribe the test corpus of "Diyala" texts to small lots purchased or acquired together, however, for the statistical analysis to garner significant weight, a corpus must be large. Therefore, testing data sets of three or twenty tablets would not yield as reliable a result when compared against the larger corpora of excavated Mesopotamia sites.

The unit of analysis is the word, which is defined by lexemes. While certain words have been assigned a lemma, this process is driven by lexical meanings, assigning all observed forms to the same meaning. In the text files, this is denoted by white spaces, which are used to demarcate the boundaries of a given word. This process of defining and delimiting words is subjective and certainly influences the outcome of textual analysis. For this reason, my lemma list, stop wordlist, transliteration files and their generated raw frequency lists are available online so that Assyriologists may access, critique and help improve the user-defined criteria of this methodology.[28]

5 "Diyala" texts and Ešnunna

Table 4 below represents those words that demonstrate a 99.99 % statistical significance (15.13 critical value) for uniqueness between the two data sets. The positive values reflect those words appearing in the "Diyala" texts (test corpus) atypically more frequently than the Ešnunna texts (reference corpus). Conversely, the negative keywords, located at the bottom of the table, represent those keywords that occur atypically more frequently in the Ešnunna texts. Based on the frequency of words in the "Diyala" corpus, the software algorithm expected to find similar frequencies for similar words, however there were some significant deviations.

The unprovenienced "Diyala" corpus differentiates itself from the Ešnunna corpus along several lines. First, there is an increased preference for the Akkadian language in the "Diyala" texts (e. g. *i-di$_3$-in*, *iš-te$_4$*, *a-na*, ARAD$_2$-*su*). Given the preponderance of Akkadian texts in the north compared to southern sites at this time, the "Diyala" texts appear to contain an abnormal number of Akkadian words altogether.

Second, the nature of the economy addressed in each corpus is slightly different. This is an expected deviation given that sites produce goods that are locally viable and/or profitable–the marsh cities producing more fish and reeds, those near irrigated fields produce more grains, etc. But even internally, different archives within the same city may, and often do, focus on different goods. Therefore, the economic differences between the two corpora are not

28 The cleaned atf and word lists for each site are available at < https://zenodo.org/record/1401502#.W4cXSl5Ki00 > as well as the lemma list and stopword list used with these transliteration files.

immediately interesting. Despite being "uninteresting," this type of expected result lends confidence to the statistical analysis.

However, it is important to note that the "Diyala" texts do not utilize the non-Akkadian gur saĝ-ĝal on par with the Ešnunna texts. The Akkadian gur-measure was introduced under the Akkadian kings and was often associated with imperial/royal goods.[29] As Foster succinctly concludes about the gur Agade, "only matters for royal accountability were accounted for by the royal standard" (Foster 2016: 49 fn. 86). However, the majority of references of Agade in the Diyala texts is to the toponym, not the capacity measure, suggesting a geographic proximity or other affinity, not necessarily one of metrology. But, coupled with the relative high amount of Akkadian in the "Diyala" texts compared to Ešnunna, a closer relationship with the Akkadian/imperial milieu is posited.

Table 4: Unprovenienced "Diyala" Texts Compared to the Ešnunna Corpus.

Keyness Rank	Raw Frequency	Keyness Value	Term	Translation
1	61	38.905	*a-na*	To/for
2	29	33.478	tug_2	Garment
3	13	30.697	dabin	Semolina
4	11	25.975	iri	City
5	11	25.975	*warassuni*	Personal Name
6	26	24.432	$abba_2$	Witness/elder
7	10	23.613	gi-nu-nu	Personal name
8	9	21.252	*iddin*	He gave (it)
9	8	18.891	$a\text{-}ra_2$	(n) times
10	23	16.756	e_2	House
11	7	16.529	$a\text{-}ga\text{-}de_3^{ki}$	Agade
12	7	16.529	geš-šid	--
13	7	16.529	gu_7	Eat
14	7	16.529	*ište*	With
15	9	15.484	ugula	Foreman
1 (Negative)	3	25.018	gur saĝ-ĝal$_2$	Capacity Measure
2 (Negative)	96	18.646	še	Barley
3 (Negative)	94	15.138	mu	Year

29 This suggestion was made by Cripps (2010: 15), although both B. R. Foster and M. A. Powell have suggested categorically distinct uses of the Akkadian (imperial) gur and the gur saĝ-ĝal (Foster 1982: 24; Powell 1987/90: 497). Through collocate analysis in my dissertation, I suggested that the gur Agade was more likely to render finished and fine goods, items befitting an imperial appetite (Brumfield 2013: 195–199).

6 “Diyala” texts and Tutub

Similar to the Ešnunna texts, the “Diyala” texts contain more Akkadian than the Tutub corpus, although at a diminished rate (see Table 5 below). The instances of clear Akkadian are more comparable between Tutub and the “Diyala” texts. The specific Akkadian terms that are significantly more prevalent in the “Diyala” texts are linked with differences in economy. The test corpus contains more documents that focus on the exchange rate of grains and silver, while the reference corpus has a more pastoral focus with goats and sheep. Again, this is an expected result, however, the keyness values for these terms are larger than those of the “Diyala”-Ešnunna comparison. This implies that the differences between the “Diyala” and Tutub texts are more pronounced and even less likely to be due to chance.

In the Tutub texts the unclear, yet increased, use of PAP is noted by AntConc as well as the texts’ increased use of patronymics. While the PAP phenomenon is limited and poorly understood, the presence or absence of patronymics is not indicative of any specific site—only more abbreviated texts.

Although the texts from Tutub align in vocabulary with the “Diyala” texts, the differences that are observed are much stronger than those between the “Diyala” and Ešnunna. This might suggest the “Diyala” texts to be an archive at Tutub, dealing with other aspects of the economy or that the “Diyala” texts were written from a more official/imperial perspective at Ešnunna, hence the increase in Akkadian terms, mentions of the city of Agade and limited use of the gur saĝ-ĝal.

Table 5: Unprovenienced “Diyala” Texts Compared to the Tutub Corpus.

Keyness Rank	Raw Frequency	Keyness Value	Term	Translation
1	96	75.719	še	Barley
2	75	58.980	gur	Capacity measure
3	61	35.850	*a-na*	To/for
4	25	25.151	ku_3-babbar	Silver
5	32	24.320	gin_2	~8.33 g
6	13	19.137	GAN_2	Field
7	13	19.137	gi	Reed
8	13	19.137	im	--
9	11	16.193	*warassuni*	Personal Name
10	26	15.231	$abba_2$	Witness/elder
1 (Negative)	5	81.965	$maš_2$	Goat
2 (Negative)	17	81.889	PAP	--
3 (Negative)	16	48.041	udu	Sheep
4 (Negative)	49	29.556	dumu	Child

7 "Diyala" texts and Tell Suleimah

The statistically significant deviations in the Tell Suleimah corpus are comparatively small in contrast to the sites of Tutub and Ešnunna (see Table 6 below). The prevalence of Akkadian among the Tell Suleimah texts undoubtedly contributes to this level of similarity. Additionally, there are no significant differences in the personal names between the two sites, which could indicate a similar cultural background.

The difference between these two corpora is similar in quality to that of the "Diyala" texts and Ešnunna and Tutub: the difference being one of economy. The different types of grains attributed to each corpus may be due to distinct periods during the agricultural cycle since milled products such as semolina and flours can only be processed after the harvest of barley and emmer wheat.

Again, it is easier to place the "Diyala" texts as part of the Tell Suleimah corpus than of the other two Diyala sites based on the overall similarities in their lexemes.

Table 6: Unprovenienced "Diyala" Texts Compared to the Tell Suleimah Corpus.

Keyness Rank	Raw Frequency	Keyness Value	Term	Translation
1	26	26.639	$abba_2$	Witness/elder
2	17	22.043	PAP	--
3	29	19.148	tug_2	Garment
4	13	16.857	dabin	Semolina
1 (Negative)	11	79.279	*in*	In
2 (Negative)	96	51.729	še	Barley
3 (Negative)	1	21.479	ziz_2	Emmer

8 "Diyala" texts and Kiš

When compared with the northern site of Kiš, the similarities between the language, commodities, metrology and personal names is even more striking (see Table 7 below). There are few distinctions between the corpora, and their distribution suggests that the "Diyala" texts could be a subset of the Kiš texts. Based on the higher number of negative results, the words in the "Diyala" texts fit in with the Kiš texts, excepting the use of PAP. However, the Kiš corpora has elements not as prominent in the "Diyala" corpus, suggesting that the Kiš corpus is perhaps broader in content (although not size).

The Akkadian of the "Diyala" texts appears to be most at home among the Kiš corpus, with no statistically significant deviations in Akkadian usage. And similar to Tell Suleimah, there were no significant differences in personal names between those people appearing in the Kiš texts and those in the "Diyala" corpus. Again, this may suggest a shared culture background, if naming practices are in fact similar in nature.

Table 7: Unprovenienced "Diyala" Texts Compared to the Kiš Corpus.

Keyness Rank	Raw Frequency	Keyness Value	Term	Translation
1	17	18.690	PAP	--
1 (Negative)	1	37.299	$uš_2$	Dead
2 (Negative)	49	29.226	dumu	Child
3 (Negative)	9	25.556	ugula	Overseer

9 "Diyala" texts and Girsu

With Girsu as the control group, it is not surprising that there are enormous deviations between these two corpora in onomastics, commodities, resources, metrology and linguistic affiliation (see Table 8 below). The strength of these results helps situate the unprovenienced texts attributed to the Diyala closer to the corpora of Kiš, Tell Suleimah, Ešnunna and Tutub. There are a relatively high number of words in the "Diyala" corpus that are more unique compared to Girsu. However, given the size of the Girsu corpus, fewer words appear unique for the southern corpus.

Again, there is a relatively high proportion of Akkadian words in the "Diyala" texts compared to Girsu—an expected result. The commodities, implications of the local economies, deviate in similar ways as above—again, an expected difference between sites situated in different ecologies.

Interestingly, there are several "banana" personal names (gi-nu-nu, a-ša-ša, a-li-li, i-bi_2-bi_2, i_3-lu-lu) not as well represented in the southern corpus as in the "Diyala" texts. The linguistic affiliation of this name type remains obscure, but may suggest regional naming preferences by the Sargonic period.[30] Certain Semitic names (e. g. Ummi-Eštar, Nabi'um, Bēlī) are expectedly more prevalent

30 Biggs (1967: 56, fn. 3) argued against Edzard's (1960: 243, fn. 10) suggestion that these "banana" name types had a Semitic affiliation. Although Sommerfeld does argue for assigning some "banana" names to Akkadian with extreme caution (Sommerfeld 1999: 26), most names remain analyzed as neither Semitic nor Sumerian.

in the "Diyala" corpus than the Girsu texts, in accordance with general observations about linguistic distribution during this period.[31]

The overall mismatch between these two corpora is expected given the geographic (and thus, ecological, economic, linguistic and cultural) distance of these two sites.

Table 8: Unprovenienced "Diyala" Texts Compared to the Girsu Corpus.

Keyness Rank	Raw Frequency	Keyness Measurement	Term	Translation
1	61	170.854	*a-na*	To/for
2	26	99.754	$abba_2$	Elder
3	55	86.436	*šu*	Of
4	96	72.022	še	Barley
5	17	63.185	PAP	--
6	75	52.602	gur	Capacity Measure
7	11	45.709	$eš_2$-gid_2	Length Measure
8	11	45.709	*warassuni*	Personal Name
9	12	43.082	zu-zu	Personal Name
10	10	41.554	gi-nu-nu	Personal Name
11	23	37.959	u_3	And
12	9	37.399	*iddin*	He gave (it)
13	15	34.751	sa_{10}	Exchange
14	7	29.088	a-ga-de_3^{ki}	Agade
15	7	29.088	geš-šid	--
16	7	29.088	*ište*	With
17	25	28.685	ku_3-babbar	Silver
18	6	24.932	a-ti-e	Personal Name
19	6	24.932	bur	Area Measure
20	6	24.932	*ma-šum*	Personal Name
21	6	24.932	šu-um	--
22	6	24.932	*um-mi-*$eš_{18}$*-dar*	Personal Name
23	8	23.770	*na-*bi_2*-um*	Personal Name
24	5	20.777	*imhur*	He received (it)
25	5	20.777	*su-ni-tum*	Personal Name
26	11	20.592	iri	City
27	13	20.373	im	--
28	4	16.622	*a-dam-u*	Personal Name
29	4	16.622	a-li-li	Personal Name
30	4	16.622	a-ša-ša	Personal Name
31	4	16.622	*be-*li_2	Personal Name

(*continued*)

31 It is unclear if dingir-kal is intended to be read as *ilu-dan*.

Table 8: (*continued*)

Keyness Rank	Raw Frequency	Keyness Measurement	Term	Translation
32	4	16.622	dingir-*na-ṣi*$_2$-*ir*	Personal Name
33	4	16.622	i-bi$_2$-bi$_2$	Personal Name
34	4	16.622	*la*	Negation
35	4	16.622	^{ĝeš}šubur	Chariot
36	6	16.470	*en-ma*	Thus
37	5	15.638	dingir-kal	Personal Name
38	5	15.638	i$_3$-lu-lu	Personal Name
1 (Negative)	5	26.188	maš$_2$	Goat
2 (Negative)	1	25.182	ma$_2$	Boat
3 (Negative)	3	22.613	lu$_2$	Man
4 (Negative)	16	19.304	|ŠU+LAGAB|	Total
5 (Negative)	2	16.065	zi$_3$	Flour

10 Conclusions

There are several possible interpretations of the data, assigning the "Diyala" corpus to a different archive at Tutub, or to a more official archive at Ešnunna, or to an archive written at a different time of year at Tell Suleimah. But the quantified data suggest that despite the possibilities, it is more probable that the "Diyala" texts come from the northern site of Kiš (or one of similar quality). Given both the number of statistically significant words and the size of the keyness values, keyword analysis claims that the lack of dissimilarity between Kiš and the "Diyala" texts are the strongest and least likely to be due to chance. However, this is not the definitive solution to the problem of provenience, but rather a methodology for identifying new areas of inquiry and providing a new vantage on old data. Often, as is also the case here, the results lead to further questions or deeper analysis. The keyword analysis highlights potentially interesting results, but it is still the responsibility of the researcher to assess if and to what degree certain keywords are meaningful. Especially given the unpredictability and irregularity of preservation and discovery, this technique should be paired with the subjective characteristics identified by specialists, such as orthography, paleography, tablet shape and grammatical variation in order to determine probable provenience. Specific differences in language, economy and personal names are each a pathway for deeper exploration of the texts identified by keyword analysis.

In general, however, new methods for determining tablet provenience are particularly relevant with the increase in the number of unexcavated tablets entering collections. Keyword analysis is suited to corpus comparison for larger numbers of cuneiform tablets, complementing the more individual level of analysis of paleography, prosopography, tablet shape, etc. It is important to be able to analyze not just the individual tablet but perhaps an entire lot of tablets to average out quirks, anomalies and random or arbitrary features. As a relatively new methodology, it is my hope that through collaboration this can be refined and improved upon to become a useful tool for Assyriology.

References

Anthony, Laurence. 2014. AntConc (Version 3.4.4) [Computer Software]. Tokyo, Japan: Waseda University. Available from http://www.laurenceanthony.net/

Biggs, Robert D. 1967. Semitic Names in the Fara Period. *OrNS* 36/1: 55–66.

Brumfield, Sara. 2013. *Imperial Methods: Using Text Mining and Social Network Analysis to Detect Regional Strategies in the Akkadian Empire*. Ph.D. Dissertation. University of California at Los Angeles.

Cripps, Eric. 2010. *Sargonic and Presargonic Texts in the World Museum Liverpool*. BAR International Series, 2135. Oxford: Archaeopress.

Cuneiform Digital Library Initiative. Accessed September 25, 2018. https://cdli.ucla.edu.

Dunning, Ted. 1993. Accurate Methods for the Statistics of Surprise and Coincidence. *Computational Linguistics* 19/1: 61–74.

Edzard, Dietz O. 1960. Sumerer Und Semiten in Der Frühen Geschichte Mesopotamiens. Pp. 241–258 in *Aspects Du Contact Suméro-Akkadien*, ed. Edmond Sollberger. Genava, N.S: IX Rencontre Assyriologique Internationale, 8, 1960.

ENEA. "Tigris Virtual Lab." ENEA Grid Project. Accessed September 25, 2018. http://www.afs.enea.it/project/tigris/indexOpen.php

Faber, Alice. 1981. Phonetic Reconstruction. *Glossia* 15: 233–262.

———. 1985. Akkadian Evidence for Proto-Semitic Affricates. *JCS* 37: 101–107.

Foster, Benjamin R. 1982. Administration and Use of Institutional Land in Sargonic Sumer. Mesopotamia Copenhagen Studies in Assyriology Volume 9. Copenhagen: Akademisk Forlag.

———. 2016. *The Age of Agade: Inventing Empire in Ancient Mesopotamia*. New York: Routledge.

Gabrielatos, Costas. 2018. Keyness Analysis: Nature, Metrics and Techniques. Pp. 225–258 in *Corpus Approaches to Discourse: A Critical Review*, eds. C. Taylor, and A. Marchi. Oxford: Routledge.

Gabrielatos, Costas, and Anna Marchi. 2012. Keyness: Appropriate Metrics and Practical Issues. Paper presented at Critical Approaches to Discourse Studies. University of Bologna. September 13–14, 2012. https://repository.edgehill.ac.uk/4196/1/Gabrielatos&Marchi-Keyness-CADS2012.pdf (Accessed February 18, 2017).

Gelb, Ignace J. 1952. *Sargonic Texts from the Diyala Region*. MAD 1. Chicago: University of Chicago Press.

———. 1955. *Old Akkadian Inscriptions in Chicago Natural History Museum; Texts of Legal and Business Interest*. Fieldiana: Anthropology 44/2. Chicago: Chicago Natural History Museum.

———. 1970a. *Sargonic Texts in the Louvre Museum*. MAD 4. Chicago: University of Chicago Press.

———. 1970b. Comments on the Akkadian Syllabary. *OrNS* 39/1: 516–546.
Hasselbach, Rebecca. 2005. *Sargonic Akkadian: A Historical and Comparative Study of the Syllabic Texts*. Wiesbaden: Harrasowitz.
Kienast, Burkhart, and Walther Sommerfeld. 1994. *Glossar Zu Den Altakkadischen Königsinschriften*. FAOS 8. Stuttgart: Verlag.
Lijffijt, Jefrey, Terttu Nevalainen, Tanja Säily, Panagiotis Papapetrou, Kai Puolamäki, and Heikki Mannila. 2014. Significance Testing of Word Frequencies in Corpora. *Digital Scholarship in Humanities*. Advance online publication. doi: 10.1093/llc/fqu064.
Maiocchi, Massimo. 2015. From Stylus to Sign: A Sketch of Old Akkadian Palaeography. Pp. 71–88 in *Current Research in Cuneiform Palaeography: Proceedings of the Workshop Organised at the 60th Rencontre Assyriologique Internationale, Warsaw 2014*, eds. E. Devecchi, G. G. W. Müller, and J. Mynářová. Gladbeck, Germany: Pewe-Verlag.
Molina, Manuel. 1991. Tablillas Sargonicas Del Museu De Monstserrat, Barcelona. *AuOr* 9: 137–153.
Pojanapunya, Punjaporn, and Richard Watson Todd. 2016. Log-Likelihood and Odds Ratio: Keyness Statistics for Different Purposes of Keyword Analysis. *Corpus Linguistics and Linguistic Theory*. Advance online publication. doi: 10.1515/cllt-2015-0030.
Powell, Marvin A. 1987/90. Maße Und Gewichte. Pp. 457–517 in *RlA 7*, ed. Dietz Otto Edzard. Berlin: Walter de Gruyter.
Rayson, Paul, Damon Berridge, and Brian Francis. 2004. Extending the Cochran Rule for the Comparison of Word Frequencies between Corpora. Pp. 926–936 in *Proceedings of the 7th International Conference on Statistical Analysis of Textual Data [JADT]*, eds. G. Purnelle, C. Fairon, and A. Dister. Louvain-la-Neuve: UCL Presses universitaires de Louvain. http://ucrel.lancs.ac.uk/people/paul/publications/rbf04_jadt.pdf.
Sommerfeld, Walther. 1999. *Die Texte Der Akkade-Zeit 1, Das Diyala-Gebiet: Tutub*. IMGULA 3/1. Münster: Rhema-Verlag.
Steinkeller, Piotr. 1982. Two Sargonic Sale Documents Concerning Women. *OrNS* 51/3: 355–369.
Steinkeller, Piotr, and John N. Postgate. 1992. *Third-Millennium Legal and Administrative Texts in the Iraq Museum, Baghdad*. MC 4. Winona Lake, IN: Eisenbrauns.
Studevent-Hickman, Benjamin. 2007. The Ninety-Degree Rotation of the Cuneiform Script. Pp. 485–513 in *Ancient Near Eastern Art in Context: Studies in Honor of Irene J. Winter*, eds. J. Cheng, and M. H. Feldman. Leiden: Brill.
Westenholz, Aage. 1984. The Sargonic Period. Pp. 17–30 in *Circulation of Goods in Non-Palatial Contexts in the Ancient Near East*, ed A. Archi. Rome: Edizioni dell'Ateneo.
———. 1996. Review: Frayne, Douglas R., The Royal Inscriptions of Mesopotamia, Early Periods, Vol 2 (2334–2113 BC). *BiOr* 52/1–2: 116–123.
———. 1999. Teil I: The Old Akkadian Period History and Culture. Pp. 17–117 in *Mesopotamien Akkade-Zeit Und Ur III-Zeit*, eds. W. Sallaberger, and A. Westenholz. OBO 160/3. Göttingen: Vandenhoeck & Ruprecht.
Yang, Zhi. 1989. *Sargonic Inscriptions from Adab*. PPAC 1. Changchun: Institute for the History of Ancient Civilizations.