Cuneiform Digital Library Bulletin 2021:6
<http://www.cdli.ucla.edu/pubs/cdlb/2021/cdlb2021_006.html >

ISSN 1540-8760
Version: 9 December 2021

A Quantitative Analysis of Proto-Cuneiform Sign Use in Archaic Tribute

Logan Born
Simon Fraser University

Kathryn Kelley
University of Toronto

§1. Introduction

§1.1 The proto-cuneiform texts, which represent the earliest phases of writing in Mesopotamia, form one of the most poorly understood of any cuneiform corpus. Among the problems for scholars working on proto-cuneiform is the difficulty of devising testable hypotheses about sign meanings. In this endeavor, approaches which consider the full corpus may support more sound argumentation than those which consider signs in isolation.

§1.2 R. K. Englund (1998: pp. 70–71) advocated an approach which pays attention to sign use, and provided a list of the top sixty most frequently attested cuneiform signs; references to this list have appeared periodically in subsequent literature and have been used to make suggestions about features of Late Uruk society. Since Englund's list, hundreds of proto-cuneiform texts from private or formerly private collections have been added to the known corpus. This note is therefore intended to update and supersede that list, offering a full sign frequency list based upon the corpus as known to the CDLI in 2020. In addition, we describe the tool used to create this list, which is more robust and flexible for exploring sign use than the CDLI interface. This tool is available online (see section 3) and is suited for further research on proto-cuneiform sign use from a number of angles.[1] Finally, we describe a case study which demonstrates how counts derived from this tool may be useful for answering questions about the proto-cuneiform corpus.

§2. Difficulties in counting proto-cuneiform sign use

§2.1 At present, there are a few traditional print tools available for studying sign use in proto-cuneiform, beyond the counts provided by Englund. The transliteration conventions used by the CDLI and adopted here are derived from the signlist of Green and Nissen (1987),[2] even though some sign names may be re-assessed with further research (see review by Steinkeller 1995b).[3] Green and Nissen also count the number of times a sign occurs in combination with other signs in a proto-cuneiform case, according to the data available at the time. The counts in Green and Nissen reflect a smaller known corpus in comparison to Englund. For example, they cite 464 uses of EN in administrative texts, and a couple dozen further lexical uses, while Englund cites 996 uses of the form EN_a alone (the most common); and we identify 1947 uses of EN using our default search criteria (described in section 5, Methodology), of which 1702 instances are EN_a.

---

[1] For example, to support work along the lines of that in Gabriel (2020), who drew wider conclusions regarding the nature of the lexical corpus based upon the extent of overlap between sign use in the proto-cuneiform lexical and administrative texts, using data derived from sign statistics according to ATU 2 (the signlist) and ATU 3 (the lexical corpus).

[2] The CDLI transliteration conventions are described at: http://oracc.museum.upenn.edu/doc/help/editinginatf/cdliatf/index.html and http://oracc.museum.upenn.edu/doc/help/editinginatf/primer/

[3] Sign counts resulting from the current tool are expected to include collations and sign names that may be updated in the future, and sign frequency numbers are therefore intended to be used as approximations rather than infallible counts. The CDLI transliterations are considered a reliable (if conventional) standard.

§2.2 A more recent alternative to these print publications is the CDLI search interface, which can report how many instances of a sign exist in an up-to-date edition of the proto-cuneiform corpus. This search, while useful for certain kinds of informal checks, is not well suited to handling complex queries about sign use. This is especially true for searches involving data on more than one sign in combination. Firstly, this is because proto-cuneiform signs are not inscribed on tablets in a strictly linear order, meaning there is only an artificial connection between the transliteration of a text and the actual arrangement of signs on the clay. Although users can already search for multi-sign strings using the CDLI search, the results will only include tablets where those signs are transliterated in the same order the user typed in their search. It is not possible to determine how often a set of signs occur in an entry together ignoring their transcribed order.

§2.3 Other types of questions, regarding features such as variant use and sign co-occurrence, are similarly challenging when conducted through the CDLI, often requiring a comparison of multiple searches. Furthermore, careful selection of search parameters is required to ensure accurate data are returned. If one searches by text period alone (e.g. "Uruk", to obtain all Uruk V, IV, and III texts), the results will also include composite lexical texts, which are "imaginary" and therefore should not be included in sign frequency counts. Tools such as the Nino-cunei Python library by Johnson and Roorda (2018) offer a powerful alternative to the CDLI search. The present work seeks to offer an intermediate between the CDLI search and the work of Johnson and Roorda: while our code is available to technically-oriented users, we also provide a simple query interface which does not require any programming ability on the part of the user, as well as an updated static list of proto-cuneiform sign frequencies with carefully defined parameters.

§3. A new tool for studying proto-cuneiform sign use

§3.1 The tool presented in this work seeks to resolve the challenges outlined in the preceding section by offering a manipulatable sign-count tool accessible to researchers who do not have knowledge of Python or other scripting languages.

§3.2 We provide our proto-cuneiform sign data in two formats to suit different workflows. First, an online interface displays sign frequencies in either list or histogram form, with the ability to filter the data along multiple parameters. Users are able to specify text period, genre (administrative or lexical), provenience, variant and compound labelling choices, and ordering within sign combinations. These options afford a degree of control over the data and flexibility in comparing sign frequencies. All options are clearly conveyed using interface elements such as sliders and checkboxes, so the user does not need to intuit which search terms might be appropriate (as in the CDLI search) nor to consult API documentation to learn about available functionalities (as in Nino-cunei).

§3.3 Our interface can be accessed at this link.[4] The tool includes the option to count data from three genres drawn from the CDLI data: "Administrative", "Lexical", or "Other" (school, legal, uncertain), or from any combination of these genres. "Other" is a marginal category including only 9 texts.

§3.4 We also provide a static list of sign frequencies[5] for offline use. This list contains the frequencies of single signs across all proveniences and genres from the Uruk III and Uruk IV periods, following the "default" settings discussed in section 5 (sign variants merged, compounds split apart).

§3.5 For consistency, all data, both static and interactive, is produced by the same code. Publications which use this code or the resulting frequencies are free to cite either this paper and/or the code itself. The code which generates the results in this paper is available at this versioned Zenodo DOI;[6] in case significant numbers of new tablets are ever published and we update this work to incorporate the new data, this DOI[7] will always

[4] https://mybinder.org/v2/gh/MrLogarithm/pe-pc-datasets-interface/master?filepath=api%2Fpc%20sign%20counts.ipynb

[5] https://raw.githubusercontent.com/MrLogarithm/pe-pc-datasets-interface/master/static/pc_frequencies_static.json

[6] https://doi.org/10.5281/zenodo.4062226

[7] https://doi.org/10.5281/zenodo.3858116

point to the version of the code which is most up-to-date.

§3.6 Englund (1998) presented the most common proto-cuneiform signs drawn from the known corpus as transliterated at that date. Figure 1 provides a quick comparison to Englund's list to highlight how the corpus has evolved since his publication. We choose parameters that we believe best resemble Englund's counting process: alphabetical variants and compound graphemes are retained as unique items, and all script phases (V–III), genres, and sites were considered. Increased counts in comparison to Englund 1998 should therefore be attributable to the increase in transliterated text available from newly published tablets. Finally, unlike Englund's original list, we have chosen not to provide "translations" for the signs, since it is clear that proto-cuneiform regularly used signs in multiple ways. Presenting a translation like "reed" for GI could be misleading, since GI may also be used to express an administrative function.

Sign	Count	Updated count	Sign	Count	Updated count
EN_a	996	1702	AB_2	202	347
$ŠE_a$	496	1044	TUR	197	381
BA	495	778	DUG_c	196	224
AN	485	983	IB_a	195	305
NUN_a	456	645	$UNUG_a$	190	262
PAP_a	409	845	NE_a	186	323
SAL	388	781	SI	183	385
GI	368	645	DUG_a	181	249
$SANGA_a$	365	713	ḪI	180	334
GAL_a	353	1164	SUḪUR	179	302
E_{2a}	335	571	KU_{6a}	176	376
UDU_a	330	572	TE	162	305
ŠU	298	614	GA_a	155	255
U_4	286	479	$ERIM_a$	153	217
TUG_{2a}	268	330	MA	151	195
BAR	265	423	KU_{3a}	146	199
BU_a	265	562	ZATU753	132	146
$ŠITA_{a1}$	252	334	SU_a	131	277
A	250	572	$APIN_a$	115	226
AB_a	242	401	MAŠ	115	220
$ŠU_2$	238	297	GAN_2	135	267
DU	237	478	KUR_a	114	239
PA_a	236	426	DA_a	113	239
KI_a	229	538	MUŠEN	111	278
SAG	224	417	GU_4	110	199
ME_a	223	429	ŠUBUR	108	299
GU_7	220	313	ZATU752	106	129
MUS_{3a}	219	289	$ŠE_3$	106	171
GAR	212	346	NI_a	105	277
NAM_2	209	515	SIG_{2b}	104	158

Table 1: Most common proto-cuneiform signs after Englund 1998: 70–71 and updated counts from our tool using comparable parameters

§3.7 Comparing our list against Englund 1998 reveals general consistencies, although the updated list includes a much higher number of tokens, partly attributable to the influx of transliterated tablets since Englund's 1998 publication; in places it also demonstrates a slightly different ordering of signs by frequency. In the interest of providing a comparable counting method to Englund, table 1 differs from our static list of sign frequencies where we find it preferable to count signs with

alphabetic variants merged. The updated count of EN in that list refers to all variants (EN_a, EN_b, etc.) and not specifically to EN_a as in table 1 and Englund's original list. Counts for all individual variants remain available via our online interface.

§4. Basic Data

§4.1 Our basic data consist of all Uruk-period tablets available on the CDLI as of 19 May 2020.[8] This dataset comprises 6,726 artifacts, but only 6,267 of these have a transliteration, and only 5,274 of those have at least one sign which is actually readable (i.e., not transliterated as a break [...] or a broken sign X).

§4.2 The Uruk V period data include a number of texts from contexts in Iran that may be considered "pre-proto-Elamite" (this is due to the manner of labelling texts in the CDLI database). These include 170 tablets from Uruk V Susa and other Iranian sites that precede and straddle the Uruk IV-III proto-cuneiform and proto-Elamite worlds,[9] and whose numerical signs are transliterated using the same naming conventions but different line organization. The cultural affiliation of the script in these texts is not always clear and they are not labeled consistently, which makes it impractical to exclude them from our data without also excluding the rest of the Uruk V texts. For this reason, the token counts in the following paragraph include a small number of numeric notations from these texts as well as four unique proto-Elamite signs. Unless otherwise stated, for applications of the tool presented in this paper, we exclude all tablets from the Uruk V period to avoid any such complications.

§4.3 With this in mind, our total corpus contains 52,943 readable tokens from non-numerical entries, which may be grouped under 1990 different sign names. Merging sign variants reduces the number of sign names to 1299. If we choose to split compound graphemes into their component parts, the count rises to 56,504 tokens with only 1261 possible sign names. Merging sign variants in this setting further reduces the number of sign names to 705.[10]

§5. Methodology

§5.1 A number of choices impact how signs and collocations are counted by our tool. Aside from routine data-cleaning tasks like removing the annotations which mark damaged or uncertain signs,[11] a few special cases present themselves. In the rare cases that the transliterations have indicated a scribal error and provided a modern "correction", we remove the correction in favor of the original writing, so that for example APIN!(KASKAL) reverts to KASKAL. Where a reading has been proposed alongside the original transliteration, we likewise remove the reading and keep the original writing, so that for example $ENLIL_x(EN_c.NUN_a)$ becomes $EN_c.NUN_a$.

§5.2 We count the frequency of each sign after making these adjustments. We do not count X or [...] as their own signs, but X exists as a component of some compound graphemes. Neither do we harmonize the transcription of compound signs: for example, the signs transliterated as $URU+1(N_{58})$ and $URUx1(N_{58})$ remain distinct in our data. We chose not to harmonize these spellings because +, x, ., and can represent different kinds of ligature.

§5.3 We consider four basic ways of counting signs. In the basic case, signs are grouped according to their transliterations, after the data cleaning described above. This means that EN_a is counted separately from EN_b, and that $EN_c.NUN_a$ is counted as a unique sign.

§5.4 In the "split" setting, compound signs are first broken into their constituent parts, which are then counted separately, so that $EN_c.NUN_a$ is counted as one instance of EN_c and one instance of NUN_a. This process is limited to signs which are explicitly transliterated as compounds: when the signlist by Green and Nissen has as-

[8] Corpus retrieved via CDLI search for Period "Uruk" and Object Type "tablet". This excludes composite texts, which have an Object Type of "other".

[9] For example, one text from Godin Tepe is transliterated with a proto-cuneiform sign name, and another with a proto-Elamite sign name.

[10] Since numerical signs appear in compound graphemes, this count of unique signs includes 42 numerical signs as well as the four proto-Elamite signs mentioned above. This means that our current count for non-numerical signs in proto-cuneiform script phases IV–III is 659 unique signs when counted in the merge variants / split compounds mode.

[11] Damaged signs are retained in the data on the rationale that the transliterations have produced very likely identifications of these signs.

signed a single name to a recognisable "frozen" sign combination such as ENDIB (made up of EN+ME+MU), the constituent parts are not provided in the transliteration, and therefore cannot be separated out in our data. We also consider a "merged" setting, where variant signs such as EN_a and EN_b are both counted as instances of a basic sign EN. Finally, we consider a case where compounds are split up and variants are merged, which we take as our default setting. The list of sign frequencies reported in the appendix to this paper are taken from this setting.

§5.5 Each transliteration is associated with a collection of metadata specifying information such as the text's genre, period, and provenience. Using this metadata, it is possible to filter the corpus prior to computing the sign frequencies. For example, we can limit the data to documents from the administrative genre (as opposed to the lexical genre), or from a specific period or periods. The counts reported in this work are taken from all origins and collections (provenience) and both administrative and lexical genres, but limited to the Uruk III and IV periods. In addition to frequencies for individual signs, we provide frequencies for combinations of signs (collocations). The user can specify the length of sequences they wish to see data about; by default we show sign frequency data for single signs.

§5.6 Proto-cuneiform tablets are visually divided into "cases", each of which contains one or more signs. Signs are not typically ordered in a linear way within a case, but may instead be subject to an as-yet poorly understood spatial grammar. This means that the linear order of signs in modern transliteration is artificial, or at best represents educated guesses on behalf of the specialist about sign hierarchies and groupings within a case. Because of this, the current tool allows users to choose to ignore sign order when determining how often a set of signs occur together; of course, this mode will also present sets of sign combinations that are not necessarily meaningful as direct proto-cuneiform collocations, but are nonetheless useful for certain quantitative approaches (see below). The default setting is to preserve the order in which signs are recorded in the CDLI transliteration.

§5.7 Ignoring sign order in this way introduces a peculiarity in the frequency of some sign sequences. While ignoring sign order is useful for finding collocations which are obscured by the linear nature of the transliteration, when a line contains multiple instances of the same sign this method of counting can be problematic. For instance, ignoring sign order, the combination EN_a KID_a occurs twice in EN_a EN_a KID_a although there is only a single instance of KID_a. Even worse, consider the transliteration of the first case of ATU 3, pl. 082, W 13982 (CDLI number P000022): $ZATU707_a$ $ZATU686_a$ $ZATU707_a$ $ZATU707_a$ KA_a E_{2a} E_{2a} KA_a $LAGAB_b$ $ZATU707_a$

§5.8 Ignoring sign order, this line contains 12 instances of the combination $ZATU707_a$ E_{2a} KA_a, and as a result this combination of signs seems much more prevalent in the corpus than a collocation like DUG_a E_{2a} $KAŠ_a$, which occurs only 4 times. In fact, $ZATU707_a$, E_{2a}, and KA_a only occur together in this case, while DUG_a, E_{2a}, and $KAŠ_a$ occur together in four cases across four different tablets. For this reason, we provide the option to count the number of cases (lines) that contain a sign combination, rather than the raw number of times those signs occur in combination. Together with the option to ignore word order this gives a more accurate view of the frequencies of these sign combinations.

§6. A case study in using proto-cuneiform sign frequency data

§6.1 We now consider a research context which demonstrates potential uses for the frequency data. The content of the proto-cuneiform composition Tribute (Civil 2013 and previous literature) has been debated for several decades. It is categorized as a lexical text, and is attested in multiple fragmentary witnesses; however, its use of uneven numerical notations and its diverse, often opaque contents mark it out as "different" from lexical texts such as Vessels (Englund and Nissen 1993: 29–32 / composite P471683). Is Tribute indeed the first attempt at recording a literary narrative (Englund 1998: 99; Civil 2013) or cultic knowledge (Westenholz 1998)? Or, can it be explained primarily as a practical scribal exercise displaying sign use with close parallels in the administrative corpus (Veldhuis 2006)? While previous scholars have relied on select observations to posit the distance between Tribute and administrative practice, the tool presented here allows us to explore the closeness of sign use in Tribute to the known corpus using computational techniques.

§6.2 Tribute can be divided into discernible sec-

tions whose relationships to one another are not well understood. Following a short header (cases 1–2), the first two thirds of the text present a "list" of consumables (28 items) that is immediately repeated case after case (altogether cases 3–58). The composition then shifts to a new topic, perhaps marsh plants and products associated with them (cases 59–82), followed by crop and field designations in the final ten cases which also make up the first part of a lexical composition known separately as Plant (cases 83–94).

§6.3 Sign frequencies

§6.3.1 Taking the first third of the text (including the header but ignoring the list's verbatim repetition) and the final third separately, we have computed the frequency of each sign in the wider proto-cuneiform administrative corpus of the Uruk III and IV periods using the default sign counting settings described above.

§6.3.2 To begin, we observe that signs from the first third of the text are moderately less well attested in the administrative corpus than signs from the final third of the text, only occurring about two-thirds as many times. Four signs—all from the opening cases of the composition—are not attested administratively (ABRIG, GAZI, $MUNU_3$, and HALUB), although ABRIG could be split into its components NUN, ME, and DU which are commonly known in administrative texts separately.[12] Three further signs (6.7%) attested in the first part of Tribute are very rare: they are in the bottom 50% (of c. 700 signs) in the sign frequency list, with fewer than 10 administrative attestations. In the final third of Tribute, 2 signs (5.1%) are likewise ranked in the bottom 50%.

§6.3.3 Tribute likewise contains some very common signs. 6 signs (13.3%) in the first part of the text are among the top 10 most frequent signs in the administrative corpus, while 7 signs (17.9%) from the final third are. If this is expanded to the top 25 most frequent signs, then this rises to 9 signs (20.0%) in the first part of Tribute, and 9 signs (28.2%) in the final third. Considering the composition as a whole, only 20 of the top 50 most common administrative signs appear, including 8 of the top 10 most frequent (or, 9 out of 10 if the NUN element from ABRIG is considered separately). In short, while almost all of the exceptionally common signs appear in this text, less than half of the top 50 most common signs overall appear, and a number of rare and unattested signs are mixed in. The overall picture is that neither section demonstrates outstandingly common or uncommon sign use.

§6.4 Collocation frequencies

§6.4.1 Single sign frequencies can only tell us so much. We can also consider sign combinations, and it is from this perspective that we observe a greater distance from the administrative corpus and a greater difference between the beginning and end sections of Tribute.

§6.4.2 To count bigram collocations in this section, we find it preferable to merge variants (in this case, to increase the likelihood of finding administrative parallels), keep compound signs intact, ignore sign ordering, and to count how many cases a collocation occurs in, as opposed to how many instances of that collocation occur overall (see section 5). For example, this last choice reduces the number of appearances of the collocation SAL SAL by about two-thirds, which reflects the administrative reality much more closely. "Bigrams" below is used to refer to any two signs that appear anywhere in the same case.

§6.4.3 Therefore, in this section, all possible sign pairings within a case are considered, due to the current deficiencies in understanding sign "ordering" within proto-cuneiform cases. Our bigram analysis does not limit itself to discrete "words", but instead explores sign use at the case level. While this study therefore does not address the important and poorly understood phenomenon of spatial organization of signs within proto-cuneiform cases, it takes advantage of the dataset in its existing form to draw out significant observations on sign distribution at the case and text level. By understanding sign use at this broader level we may be better equipped to return to individual cases and traditional Assyriological observation.

§6.4.4 Most entries in Tribute include either only one or two signs, although the first two cases of the text contain four signs each (or, splitting ABRIG, six signs in the second case). The header entries are (in composite):

[12] We note that UZ_a in case 6 may be comparable to instances in the administrative corpus where the constituent signs of UZ are transliterated as MUSZEN SZE_a (e.g. W 09579, cm / P001335).

[13] ABRIG is written NUN.ME.DU and these signs may also be read as ABGAL(=NUN.ME) DU. Note that a

1. U_4 KI SAG AD_a
2. U_4 AD_a ḪAL ABRIG[13]

§6.4.5 These first two cases are not attested verbatim in other proto-cuneiform administrative texts. Neither even do many of the possible two-sign subsets occur in the wider corpus.

§6.4.6 The following sign pairings in Tribute 1–2 are never attested administratively:

ABRIG AD

ABRIG ḪAL

ABRIG U_4

AD ḪAL

AD SAG

AD U_4

§6.4.7 While a few sign pairings in Tribute 1–2 are attested administratively, though not very frequently:

KI SAG (1 time)

HAL U_4 (1 time)

AD KI (5 times)

SAG U_4 (5 times)

KI U_4 (10 times)

§6.4.8 To this we can add that one manuscript has NAM_2 instead of AD[14] in case 1, with a clear placement of SAG inside NAM_2 that probably indicates a meaningful unit; there is one possible attestation of SAG NAM_2 in the proto-cuneiform corpus, in badly damaged fragment W 22090,4 / P004486.

§6.4.9 The meaning of these beginning entries of Tribute has been debated, with much of the discussion relying on their form as appearing in later period manuscripts (see Kelley 2021 for comments and previous literature). Further traditional analysis of these lines is not the objective of this study. However, we demonstrate that very few sign pairs can be cross-examined in light of the proto-cuneiform administrative corpus in order to strengthen hypotheses regarding their meaning.

§6.4.10 In the remainder of Tribute, around half of the possible sign combinations in cases never appear together in any case in an administrative text.

Sign combination	**Case number in Tribute**
MAŠ SI_4[15]	7 / 35
AB_2 KAL	10 / 38
AMAR GA	11 / 39
DA ZATU297[16] / $PEŠ_2$	17 / 45
$ŠA_3$ UB	29 / 57
BALA MUŠEN	26 / 54
A BALA	26 / 54
KIŠIK U_2	61
GI ZI	65
GI $ŠE_3$	66

proto-cuneiform administrative text, MSVO 1, 145 (P005212) includes both ABGAL and DU in a case, following MSVO 1 / CDLI transliteration. ABGAL does not otherwise appear in other administrative sign pairings with the signs from Tribute line 2 (U_4, AD, or ḪAL), nor does DU, excepting only one combination with U_4, and only if the combination for ADAB is deconstructed in CUSAS 21, 29 r I: DU ADAB (=U_4.NUN!).

[14] Following the reading by Englund and Nissen: 1993, 112. The sign form is not very easy to distinguish, as it is interrupted by SAG.

[15] We note a perhaps comparable correspondence between MAŠ and SI_4 in consecutive cases in the lexical Vessels (composite 36. DUG_b+MAŠ / 37. DUG_b+SI_4).

[16] Following Steinkeller (2004). The sign in our tool remains for now as KIŠ, following the CDLI corpus and Nissen and Englund 1993:113. At least one manuscript (W 20258, 4 / P000243) shows the sign to be $PEŠ_2$ "mouse", a sign otherwise unattested in proto-cuneiform following the current literature. Civil 2013: 27–28 reads $PEŠ_2$ for later Tribute manuscripts and leaves the reference for proto-cuneiform manuscripts as ZATU297 (Green and Nissen clearly include the sign form in W 20258,4 among the other forms, all under the sign name KIŠ).

ZI $ŠE_3$	66
PIRIG ZATU718	67
EN ZI	69
GAR UB	71
GIŠ SAR	74
APIN BIR_3	76
TAR ZATU751	78
DI ZATU751	80
AD EN	83
BAD SUG	91
SAR ŠAGAN	93
NAGAR ŠAGAN	94

Table 2: Sign combinations from Tribute 3–94 that never appear in administrative texts

§6.4.11 Some of the combinations in this list would not be expected to represent realistic, meaningful combinations in proto-cuneiform as simple pairs, such as AD EN (case 83) in which the whole sign series is EN GAN_2 AD, with the combination EN GAN_2 clearly being a meaningful unit known from multiple proto-cuneiform accounts. That AD and EN never appear together in an administrative case, however, is perhaps a useful observation. Other combinations from the list above, while not known in proto-cuneiform administrative tablets, may find parallels in later cuneiform.

§6.4.12 A difference in bigram use between the first and final part of Tribute is apparent. Of the fewer number of cases with more than one sign in cases 3–58, only two (AN GIR_2, A MUŠEN) appear at least once in an administrative case together. The remainder of the sign combinations that are identifiable in administrative texts derive from the final section of Tribute (cases 59–94).

Sign combination	**Case number in Tribute**	**Number of administrative attestations**
E_2 ZATU718	67	1
IŠ SAL	70	1
GAR SAG	71	1
AL GI	72	1
NE SAL	82	1
KI SAG	84 (and 1)	1
UR UR	89	1
BA KI	92	1
AN GIR_2	16	2
EN $ŠE_3$	68 / 69	2
A GIŠ	74	4
A SAR	74	4
KI KI	60	5
SAG UB	71	5
A MUŠEN	26 / 54	6
E_2 PIRIG	67	9
AN KI	85	10
EN GAN_2	83	11
SAL SAL	88	11

Table 3: Sign combinations from Tribute 3–94 with administrative parallel

§6.4.13 Using the tool presented in this paper, we can offer some context for the frequency of these bigrams, through comparison with the most common bigrams attested across the proto-cuneiform administrative record. The parameters selected on the interface are shown in Figure 1. Figure 2 presents the top 50 most common sign co-occurrences in proto-cuneiform as can be downloaded in histogram form with our tool.

§6.4.14 Of 11,611 bigrams drawn from cases in proto-cuneiform administrative texts (Uruk III/IV) using the above parameters, the top fifty (within the top 1%) are identified 35 or more times. Some of these combinations make up parts of identifiable place names (e.g. KU_6 RAD) or officials' titles (e.g. KAB NAM_2). Collocation frequency drops steadily to a very long tail of bigrams with only a few attestations: 16% of all bigrams occur only twice, and about 40% only once. A remaining 550,000 potential combinations of signs from the signlist are not administratively attested. These are only the very first steps in considering case-level sign use in proto-cuneiform, and without ready parallels we cannot yet fully interpret this data.

Figure 1: The sign count interface, with the settings selected for producing the results in figure 2

§6.4.15 However, for the purposes of the current discussion, we can observe that none of the 412 most common bigrams appear in Tribute. Following the 2-case header, 70% (7 out of 10) of bigrams from the first part of Tribute are never administratively attested, along with 23% (11 out of 43) from the final part. Considering that there are 11,611 different bigrams attested in the administrative record, it may appear quite remarkable that such percentages of Tribute bigrams should be entirely unattested. However, recall that most proto-cuneiform bigrams are relatively rare, so that the most common pairs from Tribute (EN GAN_2 and SAL SAL), which appear only 11 times administratively, are actually within the top 4% of the most common bigrams overall. Those bigrams with 5 or more occurrences are in the top 12%.

§6.4.16 Overall, Tribute does not display outstandingly common administrative collocations, and altogether 56% of unique bigrams in the composition don't have a single administrative parallel. However, the text does include a small number of bigrams of moderate frequency in the administrative corpus. The general picture is thus of a tablet with some connection to a broader administrative reality but which also employs a significant amount of material with no known administrative parallel.

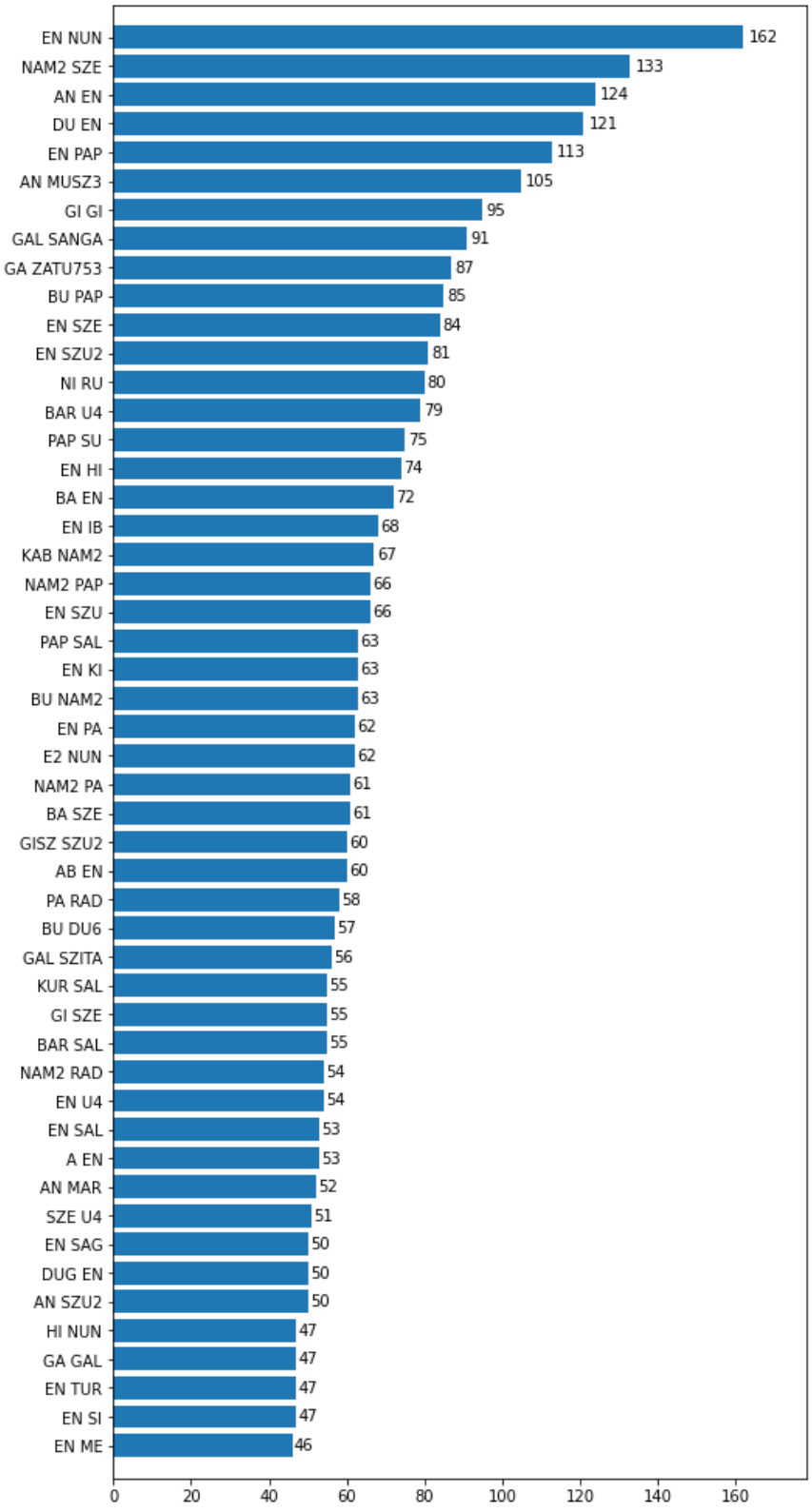

Figure 2: The 50 most common sign co-occurrences in proto-cuneiform cases

§6.5 Case ordering and tablet-wide sign use

§6.5.1 We can also examine the relationships between signs in neighbouring cases or across entire sections of the composition, in order to establish the extent to which particular administrative tablets may deal with similar sets of signs as those presented in Tribute. We compare sign combinations in the two main parts of Tribute to the administrative corpus in two ways below. Because the first and final parts of Tribute present easily observable differences in sign use, we can anchor our interpretation of the closeness to administrative material by comparing each section separately.

§6.5.2 First, we draw signs from immediately neighbouring cases in Tribute and identify how often those two signs are common to an administrative tablet, regardless of where they appear in that tablet (that is, the two signs need not be in the same case or neighbouring cases). The aim of this search is to consider how closely the ordering of entries in Tribute may reflect administrative realities: were "items" recorded next to each other in Tribute also typically recorded in particular types of administrative accounts?

§6.5.3 On average, pairs drawn from neighbouring cases in the first part of Tribute are attested in 15 administrative texts, whereas pairs drawn similarly from the latter part of Tribute are attested in 87 administrative texts. As we are considering averages, the difference is not due to the total number of signs in each section. This difference can be interpreted to mean that the structure and content of the latter third of the text are more frequently documented in surviving administrative texts, while the first third either reflects rarer administrative subgenres or uses signs and orders entries in ways foreign to the attested administrative subgenres.

§6.5.4 It is challenging to determine how to assess the statistical significance of these results. Both sections contain some otherwise uncommon signs, such as AD and IŠ, indicating some meaningful coherence across the composition and implying that the two sections are not fully independent. We make the simplifying assumption that two sections are conditionally independent given that they occur in the same document. We believe this assumption is justified as there appear to be legitimate differences in content and structure across the two parts of Tribute. This conditional independence assumption mirrors the approach taken in past work such as by Gu et al. (2018), who add a latent variable z to the conditioning context of a translation model and assume that words within a sentence are conditionally independent given this z. Under this assumption, we are able to apply Welch's unequal variances t-test to determine the significance of the difference in collocation frequencies across the two sections, and we find that the difference is highly significant at $p=0.0005$.

§6.5.5 Next, we expand the scope of our comparison to encompass any two signs drawn from anywhere within the same section (beginning vs. end) of Tribute. We again count how many administrative tablets contain both of these signs. This method was devised with the aim of investigating longer-distance relationships between signs in Tribute, to learn how well each section reflects the content of individual administrative tablets.

§6.5.6 For the first part of Tribute, 9 administrative tablets on average contain a given pair, while for the final section of Tribute, an average of 61 administrative tablets are identified per pair. This difference is highly significant ($p \ll 0.0001$) under the independence assumptions given above. Again, the first and the final part of the text seem to resemble administrative genres to different degrees, with the first part showing less similarity to known sign use. This is consistent with our interpretation of the preceding result, and strengthens the notion that the beginning of the text does not strongly reflect known administrative practices or may be a progression of disparate and less well attested subgenres.

§6.5.7 If we compare the neighbouring case query and the section-wide query, the first part of Tribute sees a 60% reduction in the average number of administrative tablets containing a given sign pair, from 15 tablets in neighbouring Tribute cases down to 9 section-wide. Since signs in neighboring cases are much more likely to reflect known administrative pairings than signs from more distant cases, this suggests the ordering of cases is somewhat reflective of broader administrative practices. However, since the section as a whole does not contain many known administrative pairings, any administrative structure must be limited to sequences of neighboring entries, and the section as a whole does not seem to function as a coherent administrative composition.

§6.5.8 There is a less dramatic change for the final section of Tribute (87 down to 61 administrative tablet parallels on average, a reduction of about a quarter), which we might hazard to interpret as suggesting that the final third of the text somewhat more consistently draws on the vocabulary of a particular administrative genre or related genres. Judging by the vocabulary, the genres apparently relate to marsh resources and certain types of field management. By comparison, the first third of the text might be said to combine vocabulary that was typically used in quite different genres, at least in relation to the known administrative corpus.

§6.5.9 Putting these results together, it appears that the latter third of Tribute uses signs in ways that are significantly more typical of the known administrative corpus. That is not to say that sign use in the first half is atypical, it is just less typical. A hypothesis guided by the above data might be as follows: the first part of the list could represent consumables that were in actual practice given in tribute or as temple offerings; however, our findings would indicate that scribal practice rarely dictated the recording of many of these offerings together on a single administrative tablet (whether as a record of offering or otherwise). To take a typical example of this: we have no administrative tablets recording both pigeons (TU_b) and fat-tailed sheep ($GUKKAL_c$), although these signs appear in adjacent cases in Tribute. Nor, taking the next case of Tribute (KAL_a[17] AB_2 "high-quality(?) cow") do we find any administrative tablets recording both fat-tailed sheep and cows of any kind. Beyond following no known administrative logic, neither does the progression from pigeon to sheep to cow seem easy to understand: it appears too irregular for a basic animal classification, and (to our eyes) the sign shapes do not appear to be the guiding principle, for example, by training a scribe for specific strokes or patterns. However, all three items could plausibly represent temple offerings, lending intuitive support to the conventional label "Tribute".

§6.5.10 In the first part of Tribute, the neighbouring case study suggested a somewhat closer relationship to administrative material than did the section-wide study. Thus the text may include some clusters of cases with meaningful ordering, even if the section as a whole does not reflect a particular administrative genre. A unique section within the first part of Tribute stands out in this respect: cases 20–25 (AB_2 / GU_4 / U_8 / UTUA / UD_{5a} / $MAŠ_2$) present a pattern of "adult female" followed by "adult male" for cattle, sheep, and goats respectively and records the ratio of 10:1 female to male, which reflects the prominent place of females in herding accounts of Mesopotamia. However, such recognizable general logic has not yet emerged for most of the other case sequences in this section of Tribute.

§6.5.11 The particular combination of signs for adult sheep and goats, female and male, in cases 22–25 finds only a very basic parallel in the vocabulary of the known animal husbandry accounts as studied by Green (1980), the latter of which regularly include the unique sets of signs for young animals which are not mentioned in Tribute. However, such accounts do maintain the positional primacy of females, as do the neighbouring proto-Elamite accounts (Dahl 2005). Proto-cuneiform account W 09578c / P001235, on the other hand, parallels Tribute by including male and female adult sheep and goat counts while excluding young, but with a complete reversal of the key parameters: males are more numerous, and appear before females; and goats appear before sheep. The very brief "sketch" of small livestock terms presented in the first part of Tribute may therefore conceivably be interpreted as an administrative training exercise related to broadly understood herd management principles but perhaps not closely related to (as yet known) specific account types.

§6.5.12 Yet the hypothesis that the first part of Tribute functioned as a general training guide lacks, at present, the ability to meaningfully explain the majority of case progressions, such as "pigeon" / "fat-tailed-sheep" / "'quality' cow". The importance of the composition to proto-cuneiform scribes (it is the most commonly attested of the lexical genre), and its persistence among scribal communities into the Old Babylonian period over a thousand years later, suggests to us that the choice of signs and case ordering demonstrated in the composition are likely to have held a cultural or educational logic—that is, a sort of narrative guiding the progression, particularly since graphical form of signs does not seem to be a significant guiding principle. Given

[17] The Tribute manuscripts use forms of ZATU281 with a crescent below, which appears to be the forerunner of UET 326 "KAL". Compare Steinkeller 1995a fn. 32.

the data described above, we propose that such a narrative—which may, we hypothesize, have been an idealised enumeration of cultic offerings—may not have significantly overlapped with scribal administrative practice, or that the relevant administrative tablets have not survived in robust numbers. Our analysis also suggests, however, that the final part of the text may be more fruitfully compared with the known administrative corpus to help clarify its contents.

§6.6 Summary

§6.6.1 Each of the analyses above indicates to us that while Tribute may riff on particular administrative terminology, it is unlikely to have been designed primarily with the intent to train scribes in the most frequent signs of the writing system, nor in practicing the most common sign combinations or case sequences. This is particularly true for much of the first part of the text, while the final part appears as a whole to have a closer relationship with well-attested collocations and certain administrative genres. We admit that this analysis has depended upon the existing corpus of around 6,700 proto-cuneiform texts, and that our current understanding of the diversity of genres across the proto-cuneiform corpus is poor: significant numbers of new texts presenting survivals of new genres could alter the picture. If Tribute were known to reflect genuine areas of proto-cuneiform administrative practice, it could then be used as a point of comparison to the currently known corpus, to indicate the original existence of some areas of scribal administration for which examples have not yet been found. However, that is probably an over-optimistic assessment, given the remaining uncertainties about the type of cultural knowledge represented in Tribute. While perhaps producing more questions than answers, this discussion has advanced our understanding of the extent to which data from the known proto-cuneiform corpus can be used to explore the design of Tribute.

§7. Conclusions

§7.1 The sign frequencies presented here are a snapshot in time of the CDLI corpus and should be cited as such, since further proto-cuneiform tablets may be excavated and added to the published corpus, and substantial revisions to the proto-cuneiform signlist and accompanying transliterations may eventually be undertaken. Our tool can be updated in the future to accommodate any large influx of data, alongside archiving of the older dataset.

§7.2 The difficulties presented in this paper surrounding how best to count signs are not unique to proto-cuneiform: similar issues around the handling of complex graphemes and sign variants arise in the contemporary proto-Elamite script, and similar tools for investigating proto-Elamite are described in Born et al. (2019).

§7.3 The case study demonstrates one way in which enhanced control over searching the data may be relevant to research questions current in proto-cuneiform scholarship. We also hope it may contribute to the formulation of new questions regarding the nature of the proto-cuneiform script and the contents of the known corpus.

BIBLIOGRAPHY

Born, L. (2020). *MrLogarithm/pe-pc-datasets-interface: Proto-cuneiform sign count utilities (Version v1.0.2)*. URL: https://doi.org/10.5281/zenodo.4062226.

Born, L. et al. (2019). "Sign Clustering and Topic Extraction in Proto-Elamite". In: *Proceedings of the 3rd Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*. URL: https://www.aclweb.org/anthology/W19-2516.pdf.

Civil, M. (2013). "Remarks on AD-GI$_4$ (A.K.A. "Archaic Word List C" or "Tribute")". *Journal of Cuneiform Studies* 65, 13–68.

Dahl, J. (2005). "Animal Husbandry in Susa During the Proto-Elamite Period". *Studi micenei ed egeo-anatolici* 47, 81–134.

Englund, R. K. (1998). "Texts from the Late Uruk Period". In: *Mesopotamien: Späturuk-Zeit und Frühdynastische Zeit (Orbis Biblicus et Orientalis 160/1)*. Ed. by P. Attinger and M. Wäfler. Fribourg, Switzerland / Göttingen, 15–217.

Englund, R. K. and J.–P. Grégoire (1991). *The Proto-Cuneiform Texts from Jemdet Nasr. Vol. 1: Copies, Transliterations and Glossary*. Berlin: Mann.

Englund, R. K. and H. J. Nissen (1993). *Die lexikalischen Listen der archaischen Texte aus Uruk*. ATU 3. Berlin.

Gabriel, G. (2020). "Die archaischen Listen aus Uruk und die proto-keilschriftliche frontier. Überlegungen zu Funktion und Genese des ältesten lexikalischen Corpus". *Journal of Ancient Near Eastern History* 7(1), 1–24.

Green, M. and H. J. Nissen (1987). *Zeichenliste der Archaischen Texte aus Uruk*. ATU 2. Berlin.

Green, M. W. (1980). "Animal Husbandry at Uruk in the Archaic Period". *Journal of Near Eastern Studies* 39/1, 1–35.

Gu, J. et al. (2018). "Non-autoregressive neural machine translation". *International Conference on Learning Representations 2018*. URL: https://openreview.net/pdf?id=B1l8BtlCb.

Kelley, K. (2021). "More Than a Woman: On Proto-cuneiform SAL and the Archaic "Tribute List", in: Current Research in Early Mesopotamian Studies, Paris 2019". In: *Current Research in Early Mesopotamian Studies. Workshop Organized at the 65th Rencontre Assyriologique Internationale*. Dubsar 21.

Roorda, D. and C. Johnson (2018). *Nino-cunei/uruk: Renamed release binaries (1.2)*. URL: https://doi.org/10.5281/zenodo.1482791.

Steinkeller, P. (1995a). "Ceremonial Threshing in the Ancient Near East. Part II. Threshing Implements in Ancient Mesopotamia: Cuneiform sources". *Iraq* 52, 15–23.

– (1995b). "Review of Green and Nissen Zeichenliste der Archaischen Texte aus Uruk (1987)". *Bibliotheca Orientalis* 52, 690–713.

– (2004). "Studies in Third Millennium Paleography, 4: Sign KIŠ". *Zeitschrift für Assyriologie und vorderasiatische Archäologie* 94, 175–185.

Veldhuis, N. (2006). "How Did They Learn Cuneiform? Tribute/Word List C as an Elementary Exercise". In: *Approaches to Sumerian Literature Studies in Honour of Stip (H. L. J. Vanstiphout)*. Ed. by T. Abusch et al. Cuneiform Monographs 35. Leiden, 181–200.

Westenholz, J. G. (1998). "Thoughts on Esoteric Knowledge and Secret Lore". In: *Intellectual Life in the Ancient Near East. Papers presented at the 43rd Rencontre Assyriologique International Prague, July 1–5, 1996*. Ed. by J. Prosecký. Prague, 451–462.