Visualization Pipeline for CDLI Accounting Corpora

Logan Born

Introduction

The CDLI hosts digitized transcriptions of a wide variety of ancient administrative and accounting corpora, including documents in scripts such as proto-Elamite, proto-cuneiform, and Sumerian cuneiform. Many of these documents contain numeric information encoded using sign names such as 2(N01) or 1(gesz2@c); significant time and domain expertise is required to convert these sign strings into modern arabic notation, and the corpora are large enough that manual analysis may be insufficient to reveal large scale trends in the numeric data.

This proposal outlines a suite of tools to assist researchers by visualizing information related to the numeric content of CDLI corpora, such as the distribution of values associated with different counted objects, the distribution of different number systems, and correlations between number systems and counted objects. These tools would present the information hosted by the CDLI in a more accessible format, by converting transcribed cuneiform numerals into interactive graphics with arabic numerals. This will open the data to a wider audience by reducing the need for domain expertise to interpret the data. Experts will also benefit from seeing summary statistics about the corpora and graphs highlighting trends which may not be apparent from manual inspection of the texts.

Deliverables

The main deliverable for this project will be an online interface (ideally built to be included in the new CDLI framework) where users can explore accounting corpora hosted by the CDLI. The user will first specify a corpus or document, for example by selecting a language from a dropdown menu or by typing a CDLI P-number into a search bar. The page will then display visualizations of the numeric content of the specified document(s). See **Related Work** for discussion of a similar system.

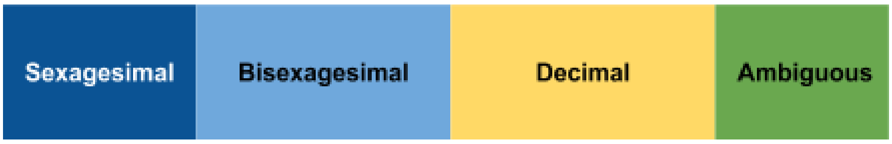
During the community bonding period I hope to work with CDLI staff to determine exactly which information will be most useful to display to researchers. The remainder of this section outlines some ideas to use as a starting point for discussion; based on the duration of the project I anticipate completing two of these suggestions during the Summer of Code, with the remainder as options for future work. The figures accompanying these suggestions display placeholder data and are meant only to be illustrative. The placeholder data is based on proto-Elamite, but the intent is to also support proto-cuneiform and Sumerian cuneiform, with the option of adding support for further languages after the main project is completed.

Searching by Numeric Content

A simple but potentially useful feature is the ability to search for texts based on their numeric content, for instance, to find all texts which count between 1000 and 2000 units of a certain item using a particular number system. Combined with the option to filter out tablets with a seal or tablets containing a given search string, this would make a simple but powerful addition to the CDLI's existing search functionality.

Number Systems

If a script employs multiple number systems, the relative frequency of each can be summarized as a stacked bar, where the width of each segment corresponds either to the number of times that number system is used, or to the total number of items counted in that system:

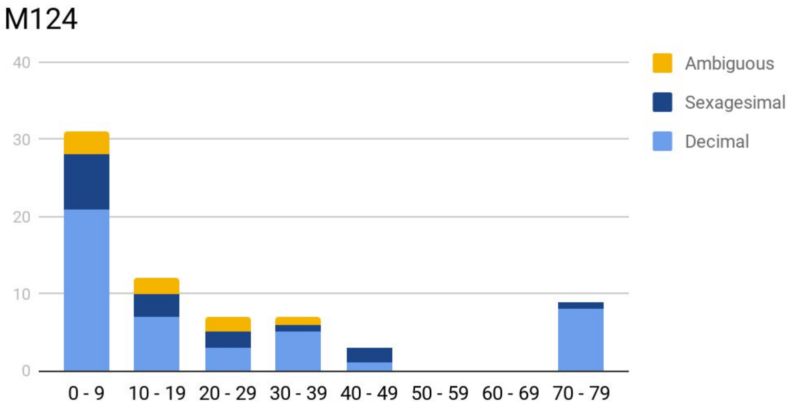


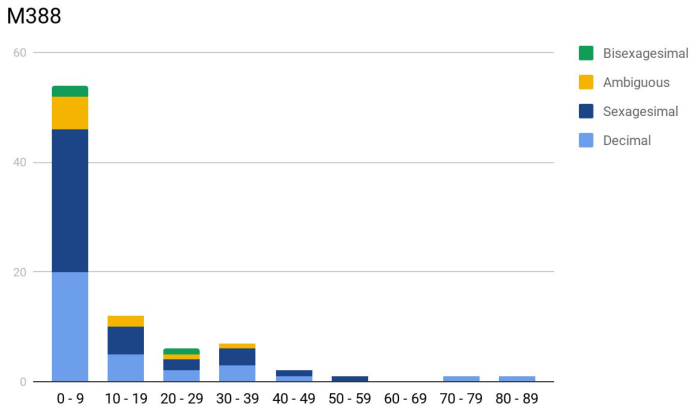
Since different number systems are used to record different kinds of objects (Englund 2011), this gives a general impression of the prevalence of different classes of objects within the corpus.

If the user selects a specific number system (for instance, by clicking one of the segments in the stacked bar), they will receive a summary of all sign-strings associated with that number system. For undeciphered languages such as proto-Elamite, this will simply comprise the set of all sign-strings occurring in the entry preceding the number. For deciphered languages, where text is segmented into words, we can retrieve the words or strings of words which are adjacent to the count. Ideally, as automated part-of-speech tagging becomes available for languages such as Sumerian, the data could be filtered to include only the nouns, but present limitations mean the data would have to include an unfiltered list of all adjacent words. This may include a list of the most common items counted with that system, with the option to sort these items by average count, frequency of occurrence with the chosen system, and overall frequency of occurrence. These options permit exploring different facets of the data, for example to find objects which are rare in general but common with a certain number system. The exact format for presenting this information will depend on usability tests and feedback from prospective users. Options include a simple table, a histogram showing word frequencies, and a bar or line plot showing differences between a word's overall frequency and its frequency in a particular numeric context.

Object Count Histograms

Given a (user-supplied) sign or word, a histogram will display how many times that word is counted with a given value. Optionally, stacked bars show how frequently the chosen string occurs in association with each number system. Side-by-side displays allow comparison of different words. Additional details, such as measures of variability and central tendency, may be printed beside the main display.

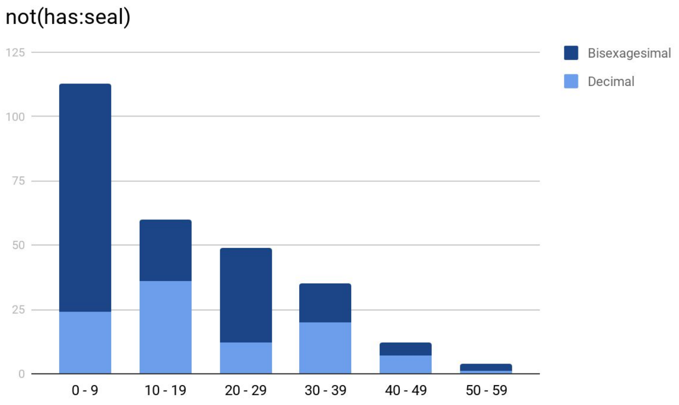


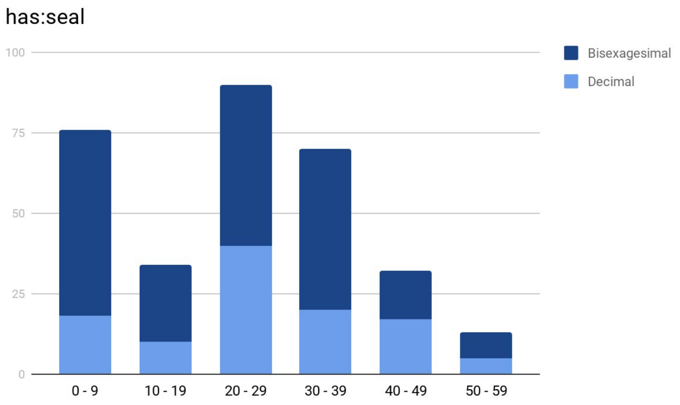


This display can reveal whether the counts associated with a given word have a multimodal distribution; whether a word tends to be counted with one or more number systems; and whether a word is usually associated with large or small quantities.

A “Find Similar Objects” button will allow users to see a list of other words which have a similar distribution. This may assist in finding groups of items with similar economic or administrative functions. Users can specify whether they wish to take the number system into account when computing similarity (so that, for example, two words with identical distributions will only be considered “similar” if they are also counted with the same number system).

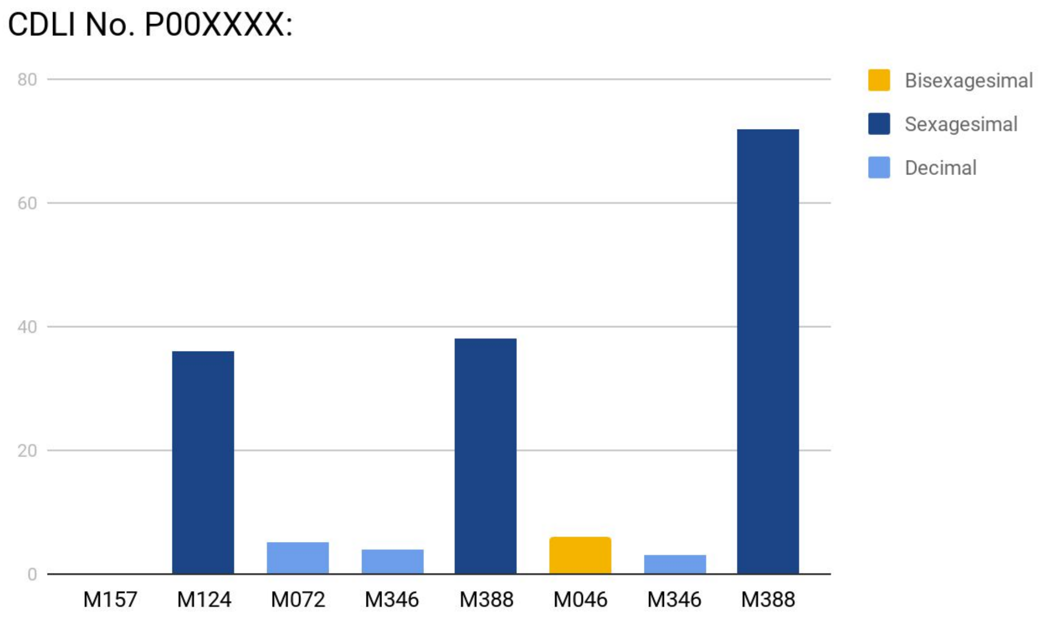
Rather than displaying a histogram for a single sign or word, the user may choose to see a histogram of all counts in a corpus. By comparing different subsets of the corpus in this way, the user may be able to identify differences between document classes or genres; for example, that sealed tablets tend to count larger volumes of items than unsealed tablets:





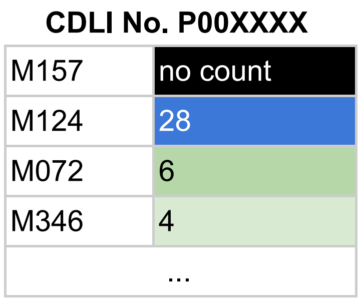
Document Overviews

To visualize an individual document, each entry or case may be displayed as a bar with height proportional to the counted value. Colors may be added to show different number systems.



This can reveal trends in the structure of a document, such as the presence of large-valued summaries within or at the end of a document; alternations between objects counted with different number systems; and whether a document employs multiple number systems.

A possible alternate view is a spreadsheet display where numerals are colored according to the number system to which they belong, and the intensity of the color reflects the size of the count:



CDLI No. P00XXXX

M157	no count
M124	28
M072	6
M346	4
...	

Implementation Overview

I propose to implement this project as a static HTML page using the D3 Javascript library to render the visualizations. To ensure fast performance and limit the processing load on viewers' web browsers, most information will be precomputed and hosted as JSON files. This has the added benefit of allowing other parties to directly access the JSON files if they wish to reuse the data from this project for other purposes. The JSON will most likely be produced by Python scripts; a cronjob or other automated process can be set up to recompute the JSON when the contents of the corpora are updated.

Proposed Timeline

Community Bonding Period		Agree on desired features. Acquire corpora or subsets of corpora to use for development. Determine method for sharing progress (e.g. as part of the CDLI framework git repository).
Week 1	18–22 May	Complete sign-value mapping for each number system in each corpus. Data cleaning (already completed for PE and PC administrative corpora).
Week 2	25–29 May	Finish data cleaning. Data preprocessing: compute values needed for the visualizations (e.g. counts, means, variance, etc.). Store and host online as JSON files. Provide scripts to recompute data when corpus updates.
Week 3	1–5 June	Finish data preprocessing.
Week 4	8–12 June	Implement first feature/visualization (e.g. interface to search by numeric values).
Week 5	15–19 June	Implement first feature/visualization (e.g. interface to search by numeric values). Phase 1 evaluation.
Week 6	22–26 June	Feedback and refinements to first feature.
Week 7	29 June–3 July	Implement second feature/visualization (e.g. object count histograms).
Week 8	6–10 July	Implement second feature/visualization (e.g. object count histograms).
Week 9	13–17 July	Feedback and refinements to second feature. Phase 2 evaluation.
Week 10	20–24 July	Test and ensure cross-browser compatibility; visual polish.
Week 11	27–31 July	Finish testing and polishing; complete user documentation or demo.
Week 12	3–7 Aug	Buffer time in case of delays.
Week 13	10–14 Aug	Final Submission

Related Work

I have previously explored visualization techniques for the non-numeric component of the CDLI's proto-Elamite corpus. A prototype of these visualizations is available online at https://mrlogarithm.github.io/767-vis-project/index.html (best viewed in Firefox). The system outlined in this proposal will function similarly to this previous work, with the addition of a more polished user interface, documentation or a guided demo to help new users, and an increased focus on cross-browser compatibility.

About Me

Why am I a good fit?

- Prior experience developing visualizations for CDLI corpora, using similar tools and techniques to those required for this project
- Strong technical background and deep interest in the CDLI's work
- Prior experience working with CDLI data
- Prior experience working with Assyriologists affiliated with the CDLI
- Have already performed data cleaning on some of the data required for this project

Personal Information

Email	loborn@sfu.ca
Website	https://mrlogarithm.github.io/about-me/about.html

Skills

Formal training in machine learning, natural language processing, and information visualization at a graduate school level. Proficiency with Python for data processing, HTML for online interface design, and Javascript/D3 for interactive data visualization.

Education

Ph.D.	Simon Fraser University Computing Science Ongoing
M.Sc.	Simon Fraser University Computing Science 2018
B.Sc.	University of Calgary Computer Science (Linguistics minor) 2016

References

Englund, Robert K. 2011. Accounting in proto-cuneiform. In *The Oxford Handbook of Cuneiform Culture,* Karen Radner and Eleanor Robson, eds. Oxford University Press.