Proto-Elamite and Proto-Cuneiform Datasets

Logan Born
loborn@sfu.ca
Simon Fraser University
School of Computing Science

Abstract

This article serves as documentation for the datasets hosted at https://www.github.com/sfu-natlang/pe-pc-datasets.

1 ngrams

1.1 Proto-Elamite

1.1.1 Methodology

The proto-Elamite sign and n-gram frequencies were derived from the proto-Elamite corpus hosted by the CDLI. As of September 2019, these texts are available here.

Transcription annotations (#, ?, !, [and], etc) were stripped from the corpus. Minor edits were performed to bring the data in line with CDLI transliteration conventions, for example by inserting a missing space after the comma in strings like ",[...]". These errata have been supplied to CDLI and will be reflected in future versions of their data.

The corpus includes signs which have been corrected by the transliterator; these are marked by angle brackets, as in <M288>. These corrections reflect the annotator's understanding of the text, but are not actually present on the physical tablet. Sign and n-gram frequencies were computed in two ways: once with corrections included, and once with corrections omitted to better reflect the actual tablet content.

Sign variants add another layer of complexity, as there is some degree of interpretation required to distinguish a malformed glyph from a legitimate variant. Similarly to the compound glyphs, we consider a setting where variants are counted exactly as written in the transcription, and one where variants are merged together, so that for example M131∼d and M131∼e are both counted as instances of M131.

1.1.2 Structure of the data

The sign and n-gram frequencies are provided as a JSON-formatted dictionary. The JSON file is structured as follows:

```
{"corrections omitted":
  "variants merged":
    "M131": 50,
    "M388 M066": 26,
    ...
  "variants separate":
    ...
"corrections included":
    "variants merged":
      ...
    "variants separate":
      ...
}
```

1.2 Proto-Cuneiform

1.2.1 Methodology

The proto-cuneiform sign and n-gram frequencies were derived from the administrative texts hosted by the CDLI. As of September 2019, these texts are available here (for the Uruk III period) and here (for Uruk IV).

Transcription annotations (#, ?, !, [and], etc) were stripped from the corpus, as were annotations marking subcases. Where the data did not follow CDLI conventions, minor corrections were performed, for example by repairing |LAGAB∼bx(HIx**N04**)| to |LAGAB∼bx(HIx**1(N04)**)|.

The proto-cuneiform corpus contains both transcription-level and transliteration-level details. For the purpose of counting sign use, transcriptions are more relevant than transliterations, as they record sign *names* rather than sign *readings*. Signs recorded as complex graphemes (as in |NUN∼a+EN∼a|) are already in a suitable format. For signs with a proposed reading attached, as in NERGAL∼x(KISZ), we have removed the reading, so that NERGAL∼x(KISZ) is counted simply as an instance of KISZ. Signs such as ABGAL, however (actually written with the signs NUN and ME) still need to be reverted to transcriptions to better reflect actual sign usage. While the overall number of signs affected by this distinction is small, an updated version of the data with these transliterations reverted to transcriptions may be released in the future.

Depending on the application, it may be relevant either to consider compound glyphs as atomic units, or to decompose them into their component parts. For this reason we provide frequencies for two different settings: one where compounds are counted as their own sign, and one where compounds are split apart, so that for example |EZEN∼a+KI| is counted as an instance of EZEN∼a and an instance of KI.

As in proto-Elamite, we consider a setting where variants are counted exactly as written in the transcription, and one where variants are merged together, so that for example EN∼a and EN∼b are both counted as instances of EN.

Proto-cuneiform is not written in a linear order; signs are rather grouped into "cases". The linear ordering of the transcriptions reflects transcribers' best guesses at the reading order of the signs. While a systematic transcription system has not been established, transcriptions are usually based on the reading order of signs in later cuneiform and an interpretation of the visual ordering of signs on the tablet. For this reason, we consider two approaches to counting n-grams. In the ordered setting, we count n-grams in the transcription as one would in any other linguistic data. In the unordered setting, we ignore the linear order of signs within a transcribed case, and instead consider every n-sign subset[1] of a case to be a valid n-gram.

1.2.2 Structure of the data

The sign and n-gram frequencies are provided as a JSON-formatted dictionary. The JSON file is structured as follows:

```
{"admin iii":
  "compounds split":
    "variants merged":
      "ordered":
        "SANGA": 484,
        "AN SIG": 3,
        ...
      "unordered":
        ...
    "variants separate":
      "ordered":
        ...
      "unordered":
        ...
  "compounds together":
    "variants merged":
      "ordered":
        ...
      "unordered":
        ...
    "variants separate":
      "ordered":
        ...
      "unordered":
        ...
"admin iv":
  ...
}
```

[1] Technically not a set but a multiset, to allow for repeated signs in a case.