# WRITING IN EARLY MESOPOTAMIA: EXPANDING THE DATABASE

Christopher Woods and Massimo Maiocchi

## 1. Introduction

The Writing in Early Mesopotamia (WEM) project endeavors to provide a comprehensive description of how the technology of cuneiform writing represented language. The project investigates early cuneiform from the perspective of both language — how sound and meaning are systematically expressed diachronically and synchronically — and semiotics — the graphic organization and history of the symbols that comprise the system. The scope of the project is the cuneiform written record from the invention of writing in the late fourth millennium BC (ca. 3300 BC) through the Old Babylonian period (ca. 1600 BC). While Sumerian writing is at the center of the project — Sumerian being in all likelihood the language for which writing was invented in Mesopotamia — the adaptation of the script to express Semitic (Akkadian and Eblaite) and the long-term interplay between these writing systems are major concerns. At the core of the project is an extensive database of spellings that is of central importance to the description of the writing system. The collection of this data is an endeavor that requires the systemic review, categorization, and analysis of thousands of texts written between the invention of writing and the end of the Old Babylonian period.

During the last year, the database of textual variants at the heart of the project has been vastly expanded. The work has proceeded on two fronts. On the one hand, we continued encoding morphological data belonging to the group of literary compositions known to modern scholars as the Decad — these were the first ten literary texts apprentice scribes would encounter in the scribal curriculum in the early second millennium BC. On the other hand, we implemented new features in order to perform advanced queries and produce summary reports, which provide synthetic overviews of the relevant data. The new additions present simple yet powerful tools to evaluate characteristics of the writing system, such as the tendency of ancient scribes to use logographic versus syllabic writings (see below §2), as well as the strategies to write phonemic clusters. The beta version of the database, including a selection of compositions, is expected to be freely available online by the end of the year, hosted on a dedicated server at the Oriental Institute. A Filemaker Server platform will allow users to remotely browse the database without the need of installing software.

## 2. The encoding of texts

In order to produce results that elucidate the writing system and Sumerian morphology, standard transliterations must be supplemented to include linguistic information. The goal here is to allow for grammatical queries. Although such encoding necessarily relies on a given understanding of particular linguistic features, care has been taken to create a flexible system, capable of producing results while minimizing assumptions about the features that are being investigated.

To facilitate the encoding process, a check-box system has been implemented that reduces the time needed for this operation, while minimizing the risk of typos. The same layout is used to perform queries, enabling the user to put multiple check-marks on the lexical and/or morphological entries of interest. As Sumerian is structurally agglutinative with a relatively complex verbal morphology, the available check-boxes are relatively numerous, yet the encoding process is readily mastered by users with a basic knowledge of Sumerian grammar. At present, roughly 13,000 lexical items have been encoded. The individual entries are first categorized on the basis of broadly defined lexical classes (verbs, adjectives, nouns, pronouns, numerals, etc.), and then linked to morphological information contained in dedicated fields. Importantly, this system makes it possible to search for morphological variants, so that, for example, a query regarding ablative infixes will produce the complete list of this morpheme in all of its attested written forms. The ability to generate results of this kind is necessary in order to produce a comprehensive description of the writing system and its historical development.

## 3. New features

After running some preliminary queries on an early version of the database, we concluded that it would be useful to extend its functionality and features. As the morphological encoding operates on the graphic level, interesting information can be retrieved from analyzing individual signs. For instance, one goal of our research is to investigate the degree of logography embedded in the writing system from a diachronic perspective. In order to do this, it is necessary to tokenize the transliterations, splitting every word into its constituent signs. This was achieved through use of a Perl script, which automatically populates a table of signs. Every sign therefore receives an appropriate identification number, as well as a word ID as a foreign key (an identification number shared between the sign and word tables), which is required to relate the two. Specific additions and enhancements to the database are given below.

### *3.1. Sign typology*

The database recognizes three basic sign types within the writing system: logograms, syllabograms, and determinatives. This terminological distinction is rather loose, but is nevertheless adequate for our purposes, while facilitating the encoding process (note, for instance, that morphograms are not distinguished in the encoding because of the complexities involved). Syllabograms are further subdivided into syllabograms proper and syllabograms used in phonetic spellings of logograms, allowing the database to produce results that reflect this meaningful distinction. Additionally, the database distinguishes semantic and phonetic determinatives, and those signs that occur in personal names. Finally, a sign can be encoded as unclear (when a sign is physically present on a tablet, but its reading is difficult), or unknown (used for broken signs). A provisional chart showing the relative distribution of signs for the composition Shulgi A is shown in figure 1. Such charts, generated for a broad spectrum of texts, will reveal how the proportion of logograms and syllabograms varies through space and time.

### 3.2. Matrixes

The term matrix refers to the conventional display of literary compositions in such a way that only the composite text is given in full, while the individual texts are rendered using a series of conventions that reflect meaningful differences between exemplars. Matrices filter out redundant noise, facilitate the identification of variant spellings, which maybe regionally or diachronically motivated, at a glance. An example, again from the composition Shulgi A, is shown in figure 2. At present, matrixes are still a work in progress, but we expect to make progress in this area within the next couple of months.

### 3.3. Variations in lexical texts

Besides literary variants present in the Decad, the database now includes variants occurring in other corpora as well. Our focus in this area, to date, has been the variations in spelling exhibited in lexical lists. Although the significance of these variants for the project cannot be overstated, the encoding of all the entries occurring in the lexicographical tradition would represent an enormous undertaking. As a provisional measure, we have implemented this feature of the database using the entries provided by the ePSD (electronic Pennsylvania Sumerian Dictionary, http://psd.museum.upenn.edu/), providing links for checking references and readings. The IDs connected to the individual entries of the ePSD have been maintained in the new section of the WEM database. As is the case of other online repositories of transliterations, the information present in the ePSD although readily available requires further encoding to be of use to the WEM project. Therefore, two different sets of encoding protocols have been implemented: a first one to spot variations in the rendering of certain phonemes, and a second one to identify variations in the rendering of syllables.

#### 3.3.1. Phonographic variations

The addition of phonographic and syllabographic variation tables represents a second set of data, which has not yet been related to the table of literary variants. Future implementations of the WEM database may link the data of both tables to the primary database.

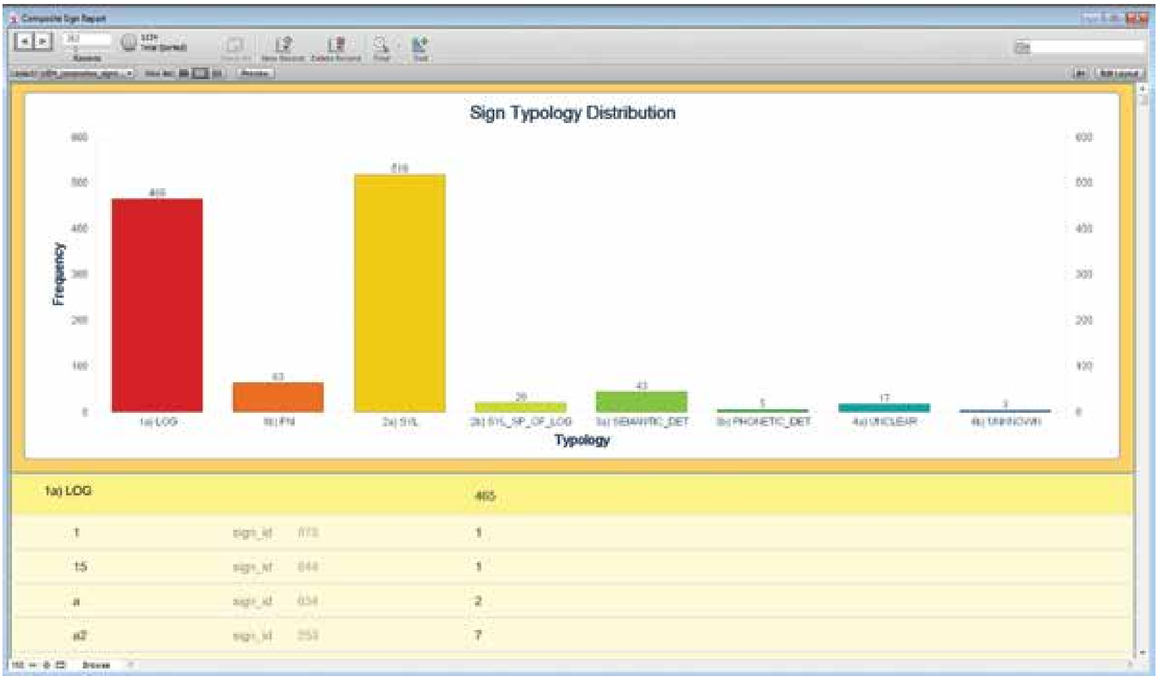
The attestation of phonographic variation is often dependent upon the way we transliterate texts, which is typically conventional rather than reflecting the actual phonetic structure of a given word. The term phonographic is therefore used to stress the fact that the possible underlying phonetic structure of a given word is inherently embedded within a logo-syllabic system, which is still poorly understood. In other words, the phonemes that we are trying to isolate belong to syllables and are represented in writing by a sign or group of signs, the phonetic structure of which is often uncertain. With this caveat in mind, it is nevertheless of interest to consider the typology and distribution of phonographic variations. The data have been entered in a layout populated by a Perl script that is also responsible for generating Unicode cuneiform graphs for the individual signs composing the lexemes. This feature is especially useful for quickly isolating variations in spelling that are merely graphic (see fig. 3); distributions of this kind of variation are best appreciated in chart view (see fig. 4). Note, for instance, that the most common variation is the alternation a ~ Ø, which is explainable in most cases in terms of the presence or absence of the subjunctive morpheme -a in the written representation. The second most common is the a ~ u variation, which is, on the other hand, largely explainable in terms of loanwords from Akkadian. Other variations are less frequent, such as the much-discussed b ~ g alternation, which requires a more detailed description than can be given here.

### 3.3.2. Syllabographic variations

The variation in the use of certain sequences of syllables to express lexemes stands on more certain ground. The encoding uses conventions that renders each syllable for a given lexeme as either V, CV, VC, CVC, etc. (V = vowel, C = consonant, see fig. 5). This allows for queries about the syllabic structures of specific lexemes as well as the distribution of syllable-types across morphemes. Of particular interest are the variations within a syllable-type, such as the case of a CVC sign alternating with a CV-CV sequence, where the second syllable represents /C/, as in the case of the term gun$_3$ alternating with gu-nu, or in the case of more phonologically complex variations such as šeg$_5$-šeg$_5$/ši-ši-ig, bir/bi-bi-re, etc. The chart in figure 6 displays the relative distribution of the primary syllabographic variations encoded to date. One notes the above mentioned CVV/CV-CV variation, which is typologically common and deserving of an in-depth study.

### 3.3.3. Graphemic sign list

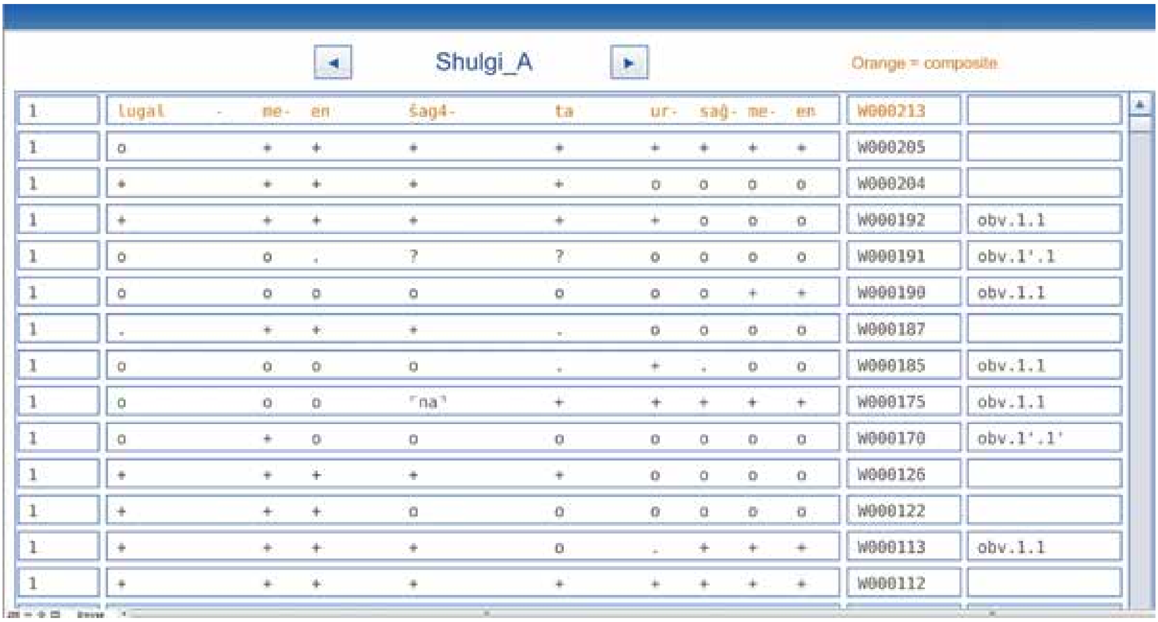
Still in its beta phase, this feature has been implemented with the goal of better understanding sign usage. The graphemic sign list is generated by a Perl script that is responsible for rendering standard transliterations as a sequence of sign names, breaking composite signs (known as diri compounds) into their relative components (see fig. 7). For instance, the graph NE is recognized not only for its representation of individual morphemes, but also as an element in diri compounds, such as ERIM$_2$ = NE.RU. The frequency of a graph as a constituent in these compounds is also given. This is the first step towards a better understanding of diri compounds, which will comprise a major avenue of research for the WEM project.



1a) LOG 465

| | | | |
|---|---|---|---|
| 1 | sign_id | 073 | 1 |
| 15 | sign_id | 044 | 1 |
| a | sign_id | 034 | 2 |
| a2 | sign_id | 253 | 7 |

Figure 1. The distribution of logograms and syllabograms in the composition Shulgi A

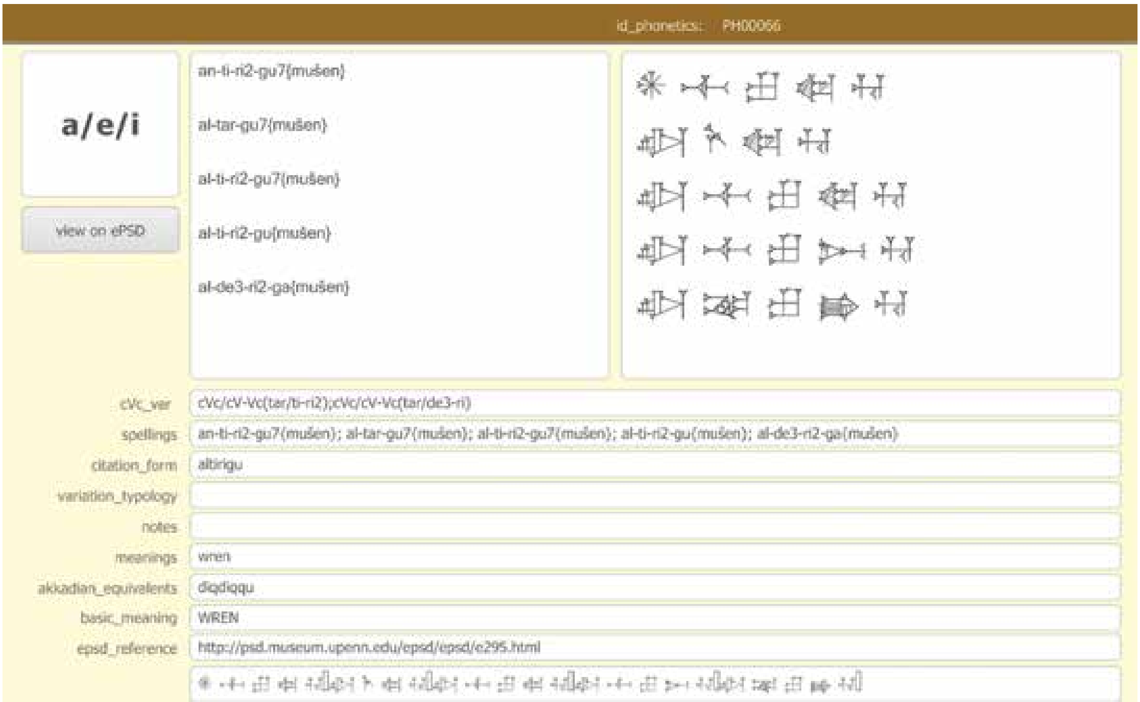


Shulgi_A

Orange = composite

| | | | | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|
| 1 | lugal | - | me- | en | šag4- | ta | ur- | saĝ- | me- | en | W000213 | |
| 1 | o | | + | + | + | + | + | + | + | + | W000205 | |
| 1 | + | | + | + | + | + | o | o | o | o | W000204 | |
| 1 | + | | + | + | + | + | + | o | o | o | W000192 | obv.1.1 |
| 1 | o | | o | . | ? | ? | o | o | o | o | W000191 | obv.1'.1 |
| 1 | o | | o | o | o | o | o | o | + | + | W000190 | obv.1.1 |
| 1 | . | | + | + | + | . | o | o | o | o | W000187 | |
| 1 | o | | o | o | o | . | + | . | o | o | W000185 | obv.1.1 |
| 1 | o | | o | o | ⌜na⌝ | + | + | + | + | + | W000175 | obv.1.1 |
| 1 | o | | + | o | o | o | o | o | o | o | W000170 | obv.1'.1' |
| 1 | + | | + | + | + | + | o | o | o | o | W000126 | |
| 1 | + | | + | + | o | o | o | o | o | o | W000122 | |
| 1 | + | | + | + | + | o | . | + | + | + | W000113 | obv.1.1 |
| 1 | + | | + | + | + | + | + | + | + | + | W000112 | |

Figure 2. Matrix view



id_phonetics: PH00066

a/e/i

view on ePSD

an-ti-ri2-gu7(mušen)

al-tar-gu7(mušen)

al-ti-ri2-gu7(mušen)

al-ti-ri2-gu(mušen)

al-de3-ri2-ga(mušen)

| | |
|---|---|
| cVc_ver | cVc/cV-Vc(tar/ti-ri2);cVc/cV-Vc(tar/de3-ri) |
| spellings | an-ti-ri2-gu7(mušen); al-tar-gu7(mušen); al-ti-ri2-gu7(mušen); al-ti-ri2-gu(mušen); al-de3-ri2-ga(mušen) |
| citation_form | altirigu |
| variation_typology | |
| notes | |
| meanings | wren |
| akkadian_equivalents | diqdiqqu |
| basic_meaning | WREN |
| epsd_reference | http://psd.museum.upenn.edu/epsd/epsd/e295.html |

Figure 3. Phonographic variants

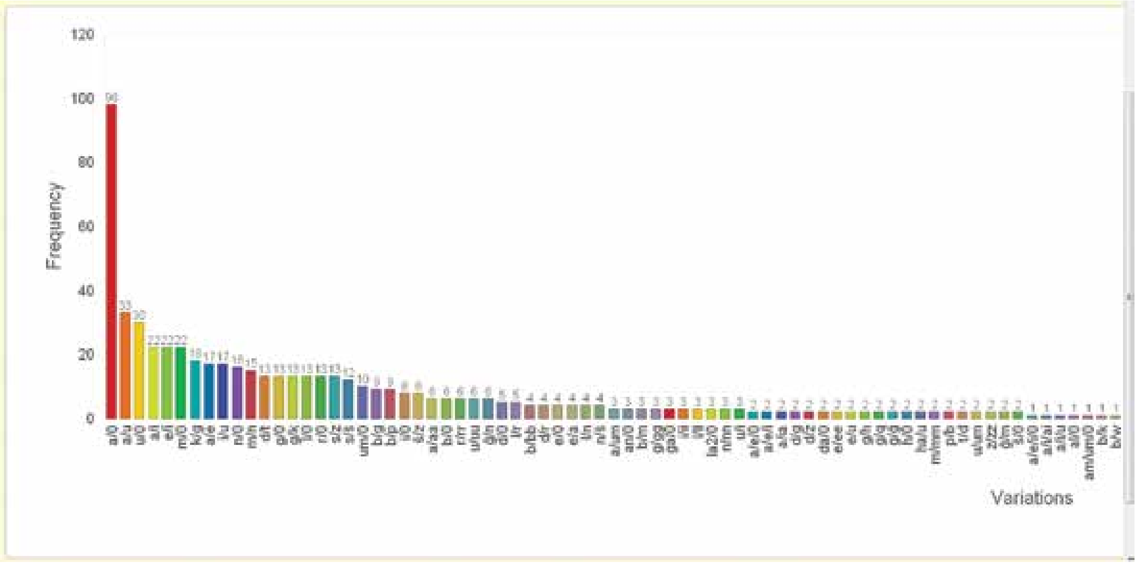
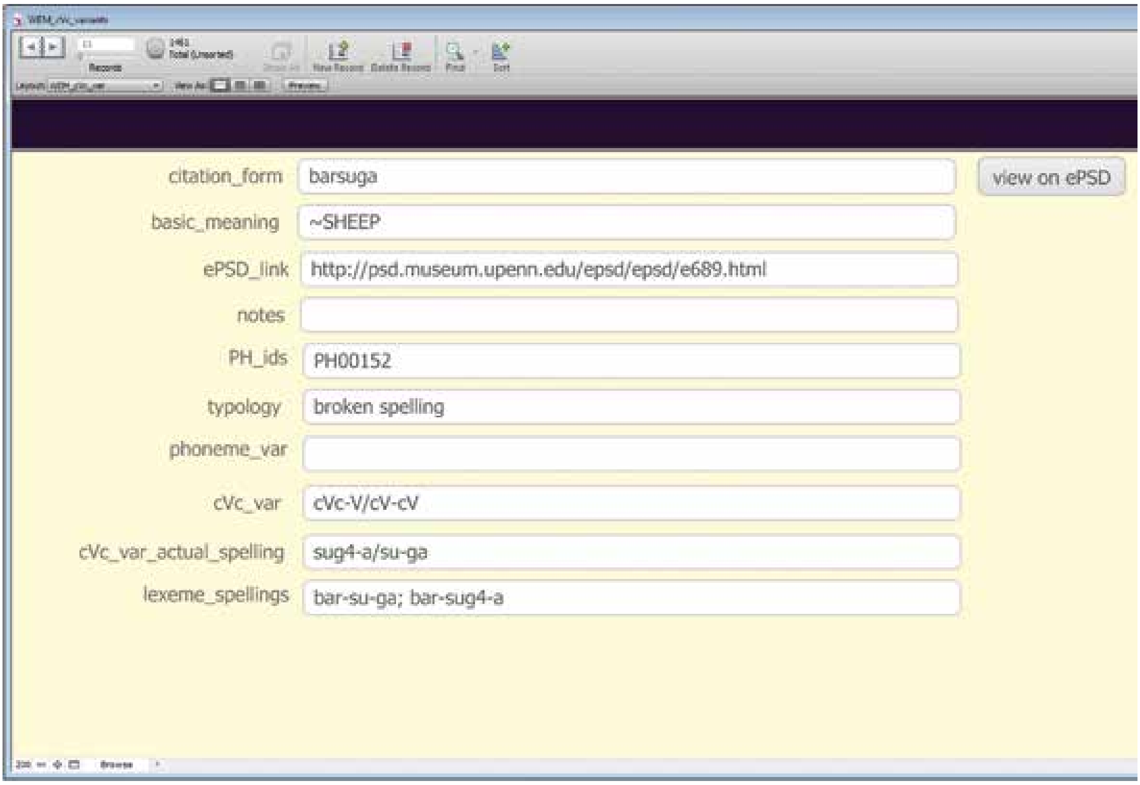
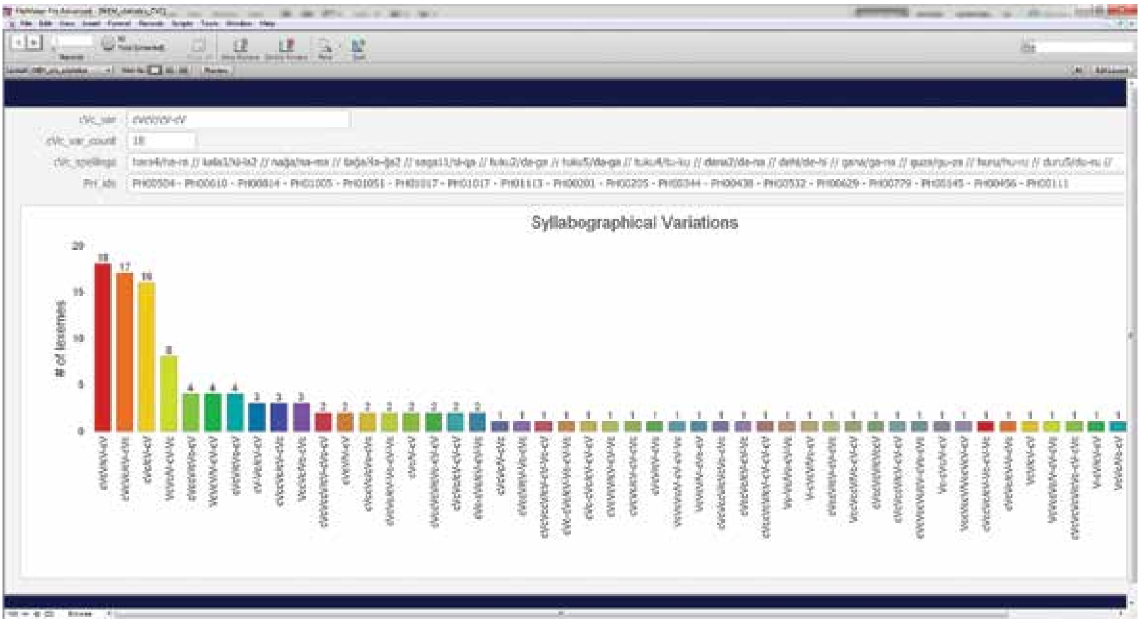


*Figure 4. Chart of phonographic variations*



| | |
|---|---|
| citation_form | barsuga |
| basic_meaning | ~SHEEP |
| ePSD_link | http://psd.museum.upenn.edu/epsd/epsd/e689.html |
| notes | |
| PH_ids | PH00152 |
| typology | broken spelling |
| phoneme_var | |
| cVc_var | cVc-V/cV-cV |
| cVc_var_actual_spelling | sug4-a/su-ga |
| lexeme_spellings | bar-su-ga; bar-sug4-a |

*Figure 5. Syllabographic table*



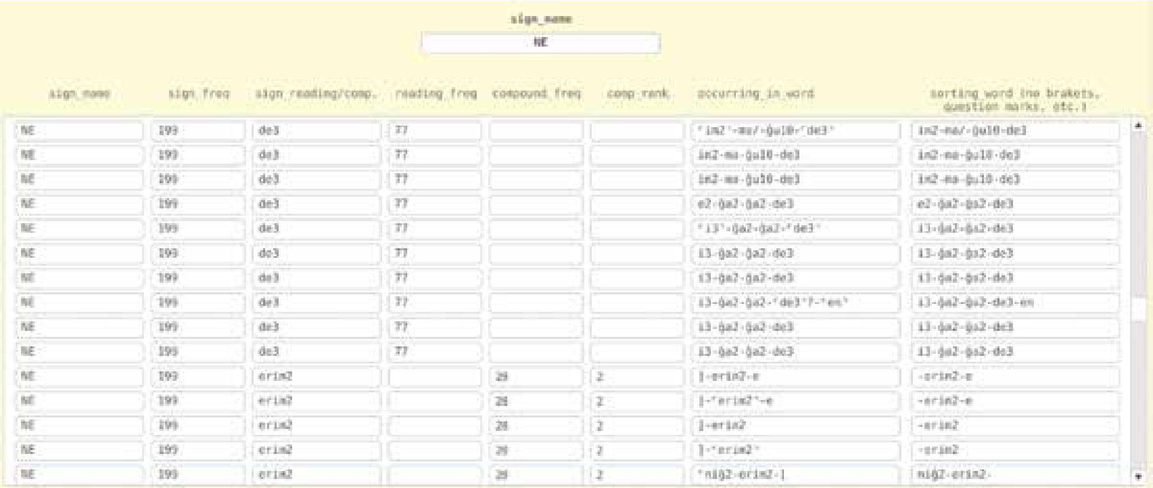
cVc_var: cvcvcvcv-cV

cVc_var_count: 18

cVc_spellings: hara4/ha-ra // kala1/ki-la2 // naĝa/na-ma // taĝa/ta-ĝa2 // saga11/si-ga // tuku2/da-ga // tuku5/da-ga // tuku4/tu-ku // dana2/da-na // dehi/de-hi // gana/ga-na // guza/gu-za // huru/hu-ru // duru5/du-ru //

PH_ids: PH00504 - PH00610 - PH00814 - PH01005 - PH01051 - PH01017 - PH01017 - PH01113 - PH00201 - PH00205 - PH00344 - PH00438 - PH00532 - PH00629 - PH00779 - PH00145 - PH00456 - PH00111

*Figure 6. Chart showing syllabographic variations*



sign_name: NE

| sign_name | sign_freq | sign_reading/comp. | reading_freq | compound_freq | comp_rank | occurring_in_word | sorting_word (no brakets, question marks, etc.) |
|---|---|---|---|---|---|---|---|
| NE | 199 | de3 | 77 | | | ˹im2˺-ma/-ĝu10-˹de3˺ | im2-ma/-ĝu10-de3 |
| NE | 199 | de3 | 77 | | | im2-ma-ĝu10-de3 | im2-ma-ĝu10-de3 |
| NE | 199 | de3 | 77 | | | im2-ma-ĝu10-de3 | im2-ma-ĝu10-de3 |
| NE | 199 | de3 | 77 | | | e2-ĝa2-ĝa2-de3 | e2-ĝa2-ĝa2-de3 |
| NE | 199 | de3 | 77 | | | ˹i3˺-ĝa2-ĝa2-˹de3˺ | i3-ĝa2-ĝa2-de3 |
| NE | 199 | de3 | 77 | | | i3-ĝa2-ĝa2-de3 | i3-ĝa2-ĝa2-de3 |
| NE | 199 | de3 | 77 | | | i3-ĝa2-ĝa2-de3 | i3-ĝa2-ĝa2-de3 |
| NE | 199 | de3 | 77 | | | i3-ĝa2-ĝa2-˹de3˺?-˹en˺ | i3-ĝa2-ĝa2-de3-en |
| NE | 199 | de3 | 77 | | | i3-ĝa2-ĝa2-de3 | i3-ĝa2-ĝa2-de3 |
| NE | 199 | de3 | 77 | | | i3-ĝa2-ĝa2-de3 | i3-ĝa2-ĝa2-de3 |
| NE | 199 | erim2 | | 28 | 2 | ]-erim2-e | -erim2-e |
| NE | 199 | erim2 | | 28 | 2 | ]-˹erim2˺-e | -erim2-e |
| NE | 199 | erim2 | | 28 | 2 | ]-erim2 | -erim2 |
| NE | 199 | erim2 | | 28 | 2 | ]-˹erim2˺ | -erim2 |
| NE | 199 | erim2 | | 28 | 2 | ˹niĝ2-erim2-] | niĝ2-erim2- |

*Figure 7. Graphemic sign list (beta)*