L2/23-229

Proposed changes to PDUTR #56

To: SAH, UTC
From: Robin Leroy
Date: 2023-10-14

Summary and rationale

At UTC #175 when L2/23-071 *Unicode Cuneiform Sign Lists* was presented to the UTC, it was pointed out that Section 3 of UTR #56 should not specify the syntax of a file that is not maintained by the UTC, and that whatever documentation was needed to use the encoding should be added to the documentation of the OGSL project. A review note was added to the Proposed Draft reflecting that comment.

On the 10th of July I forwarded this feedback to Niek Veldhuis and Steve Tinney, the maintainers of the OGSL, and at UTC #176 I reported that work had started in earnest on not just documenting the data files but also improving their structure. On the 14th of August I met with Niek Veldhuis and Steve Tinney to discuss the new structure, the documentation, and future encoding prospects. On the 28th of August Steve Tinney posted draft documentation for the new format on the build-oracc server.

As a side effect of these changes, the link to the old OGSL documentation in PDUTR #56, revision 1, draft 2, is broken, and the documentation of the format in Section 3 refers to an obsolete format; the review note in Section 3 of PDUTR #56 should therefore be addressed as soon as possible in order to be able to get useful feedback from the public. This proposal does so.

This proposal also includes changes based on feedback from Assyriologists and from SAH members.

Additional background for proposed changes is provided in this document in orange boxes.

Proposed changes

No changes are proposed to Section 1, Introduction; it is omitted in this document.

2 Principles of Cuneiform Encoding

2.1 Cuneiform Signs

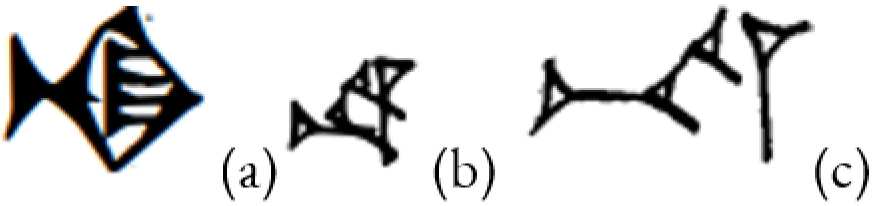
Assyriologists have published many *sign lists*, that is, classifications of the répertoire of cuneiform signs; these are numbered lists of signs, each illustrated with its glyphic range in the area and time period of interest, and often associated with a representative glyph from the Neo-Assyrian period and with the phonetic and logographic values of the sign. The sign lists play a similar role to the *sources* used in the CJKV or Tangut encodings.

Added the comparison to sources based on feedback from Ben Yang. Egyptian hieroglyphs would be another example once the new data file is added.

Examples of such sign lists include [BAU], [ELLes], [HZL] [KWU], [LAK], [MÉA], [MZL], [aBZL], [RÉC], [RSP], [ŠL], and [ZATU]. Notably, [ŠL] and [MÉA] use the same numbering; however, the other sign lists have different numbering schemes.

The glyphic range of a sign is stylistic, encompassing for instance variation between lapidary inscriptions and cursive on clay tablets, regional variation, and variation between time periods; see Figure 1. Distinct glyphs for the same sign are not used contrastively, nor do they co-occur in texts that use a consistent style. In particular, for a given sign, the various phonetic and logographic values are not distinguished by contrasting glyphs.

Figure 1. Glyphs for the sign NA 𒈾 in (a) Old Babylonian lapidary style (b) Old Babylonian cursive style (c) Neo-Assyrian style, as shown in [MÉA].



These signs are the abstract characters of the cuneiform script. See also point 5 in [ICE]. This approach makes it possible to encode texts known from multiple copies (so-called *composite texts*) that use different styles but consistent spellings, or to use encoded text to refer to the signs diachronically, as in dictionaries or sign lists covering broad timespans.

A short rationale for the encoding model seemed useful, given that many SAH members found it surprising.

2.1.1 Transliteration

No changes are proposed to this section; it is repeated in this document as it is referenced in proposed text.

Texts are often published in transliterated form; the scheme for transliteration (and for the notation of sign values) originates with Thureau-Dangin's [Syllabaire]. It uses numeric subscripts to distinguish homophones; the numbering of homophones is kept consistent across sign lists.

Note that accents can be used interchangeably with numbers (ú for u_2, ù for u_3), and additional information about the interpretation of signs is conveyed by capitalization and styling; a discussion of the specifics of assyriological transliteration is out of scope for this document.

Thanks to this numbering, a transliteration uniquely determines the sequence of signs of the original text. For example, the transliterations *ib-bu-*u_2 and *ib-bu-u* of distinct spellings of Akkadian *ibbû* "they named" are unambiguously transliterations of the sequences of signs 𒅁𒁍𒌑 and 𒅁𒁍𒌋, respectively. Note that while they share the phonetic value /u/, the signs U_2 𒌑 and U 𒌋 are not stylistic variants of each other: they have distinct sets of values and meanings; for instance, 𒌑 means "grass" and 𒌋 means the number 10, meanings that are not shared with the other sign.

This relation between transliteration and abstract characters means that encoded cuneiform texts can be automatically generated from transliterated corpora. The reverse is not true; for instance, the sign 𒀀 might be transliterated *aš*, *ina*, or *dil*, depending on context.

A machine-readable format for cuneiform transliteration exists to facilitate such automatic processing of transliterated corpora. See [ATF].

2.2 Sequences

No changes are proposed to this section; it is repeated in this document as it is referenced in proposed text.

Some signs can be analysed in all styles as a sequence of other signs written one after the other, and some sequences of signs have special values unrelated to their components; for instance, the sign $GEME_2$ 𒂗𒆠 is always written like the sign SAL 𒂗 followed by the sign KUR 𒆠, even as these signs change across styles; the sign DIRI 𒀭𒇽 is always written as SI 𒀭 followed by A 𒇽.

Such signs are not separately encoded; the corresponding sequences should be used to represent these abstract characters. See also items 2 and 5 in [Principles], and *Complex and Compound Signs* in *Section 11.1, Sumero-Akkadian*, of [Unicode].

2.3 Mergers and Splits

Some signs have distinct glyphs in the styles of earlier periods, but identical glyphs in those of later periods; such occurrences are called *mergers*. Conversely, some signs have identical glyphs in the styles of earlier periods, distinct glyphs in those of later periods; such occurrences are called *splits*.

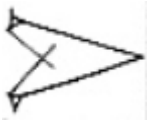
When encoding texts written in styles where the glyphs of merged or split signs are identical, the character corresponding to the correct sign value should be used, so that the encoding of a text is independent of the style in which it is written.

Figure 2 illustrates splits and mergers affecting four signs; note that a sign can be affected both by a split and a merger, as is the case of TI_2 𒋾, which splits from DIN 𒁹 and merges with ḪI 𒄿.

Figure 2. Mergers and splits of ●, 𒄿, 𒋾, and 𒁹.
The source of the hand copy shown is listed in each cell.

This diachronic approach to the encoding means that characters newly encoded to represent a contrast present in some styles may need to be supported in fonts where that contrast is absent. For instance, after the sign 𒈨 MEŠ was encoded in Unicode Version 7.0 to represent the contrast with the sequence me-eš in Neo-Assyrian styles, as illustrated in *Section 2.3.1, Mergers and Splits of Sequences*, fonts for Old Babylonian styles had to be updated to support newly encoded Akkadian texts, even though the plural marker MEŠ looks identical to the sequence of syllables me-eš in Old Babylonian.

The preceding paragraph was added based on feedback from Peter Constable, who had asked about implications for fonts.

	Early Dynastic IIIa	Ur III	Old Assyrian	Middle Assyrian
● $ŠAR_2$	 [P010576]	 [P142296]		 [P281820]
◇ ḪI	 [P225950]	 [P142296]	 [P360975]	 [P282017]
◬ TI_2		 [P142296]	 [P360975]	 [P282017]
◇ DIN	 [P225950]	 [P103303]		 [P282017]

See also item 11 in [Principles], as well as *Mergers and Splits* in *Section 11.1,* Sumero-Akkadian, of [Unicode].

2.3.1 Mergers and Splits of Sequences

No changes are proposed to this section; it is repeated in this document as it is referenced in proposed text.

A special case of mergers and splits is that of signs that look like sequences of other signs in some styles, but have a different appearance (and are sometimes even used contrastively with the corresponding sequence) in other styles. In such cases, they are not considered as sequences as described in *Section 2.2, Sequences*, and are separately encoded.

For example, the sign MEŠ 𒈪 (an Akkadian plural marker) originally looks like the sequence of syllables *me-eš* 𒈨𒋛, but their appearance diverges in Neo-Assyrian styles, as shown in Figure 3.

Figure 3. The sequence *me-eš* 𒈨𒋛 and the sign MEŠ 𒈪 on a Neo-Assyrian prism; photograph from [P422664].

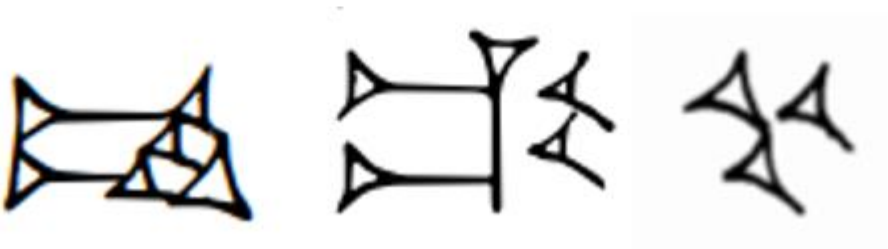
2.4 Representative Glyphs

No changes are proposed to this section; it is omitted in this document.

2.5 Sign Names

The names of the signs are generally based on a structural analysis of the signs, rather than on the common sign values; thus 𒄠 is described as GUD×KUR (𒄞 × 𒆳, meaning 𒆳 inscribed inside 𒄞), rather than AM. Note that this structural analysis may not be evident in all styles; see Figure 4.

Figure 4. Neo-Assyrian glyphs for AM 𒄠, GUD 𒄞, and KUR 𒆳 from [MÉA].



In some styles, the sign may even have a different structure from the one described by the name, as shown in Figure 5, where U+1224B 𒉋 CUNEIFORM SIGN NE SHESHIG instead appears like NE×PAP 𒉈 × 𒉽.

Figure 5. Left: the sign BIL$_2$ 𒉋 on the stele of Hammurapi [P249253]. Right: the sign NE 𒉈 on the same stele. In that style, BIL$_2$ appears as NE×PAP.

The preceding paragraph and figure were added based on feedback from Erica Scarpa.

See also item 8 in [Principles].

2.6 Discretionary Ligatures

No changes are proposed to this section except from the renumbering of the figure; it is omitted in this document.

3 The Oracc Global Sign List

The Oracc Global Sign List [OGSL] associates signs with their encoding, with their values, and with their numbers in various sign lists; it can therefore be used to automatically produce encoded versions of

transliterated texts as described in *Section 2.1.1, Transliteration*, to build input methods based on transliteration, and to look up the glyphic range of a sign in various styles.

The preceding change is based on feedback from Peter Constable, who had asked about implications for input methods.

~~3.1 Structure~~

The Oracc Global Sign List is available as the machine-readable file https://github.com/oracc/ogsl/blob/master/00lib/ogsl.asl. A ~~complete~~ specification of the structure of that file ~~the OGSL is outside the scope of this document; we merely describe how these associations are represented. Information on additional data stored in the OGSL, such as notes or deprecated values,~~ may be found at [~~G~~ASL].

The Oracc Global Sign List treats the Unicode encoding as a sign list, and establishes a concordance with the other sign lists. However, while multiple OGSL signs may share the same number in the classical sign lists, a code point corresponds to at most one OGSL sign. This is a consequence of the principles described in *Section 2.3, Mergers and Splits*.

For example, the signs 𒁄 BALAG and 𒂀 DUB$_2$ both correspond to sign number 565 in [MZL] because they merge after the Ur III period, but they are encoded separately as they are distinct in earlier styles.

Not all signs in the OGSL correspond to a Unicode code point. Some signs are encoded as sequences, as described in Section *Section 2.2, Sequences*; the OGSL documents the appropriate sequence. Other signs have no documented encoding. Some of them may be candidates for encoding; however, as the OGSL is a working dataset, others may eventually be found to be misreadings, to be duplicates or variants of already-encoded signs, or to otherwise be unencodable.

Indeed, some signs in the OGSL, including some that are encoded in Unicode, are marked as deprecated, because they are the result of errors in the classification of cuneiform signs. Some of these errors occurred as part of the encoding process. For example, the sign DUB×EŠ$_2$ 𒁿 does not exist; sign number 243 in [MZL] is named DUB×ŠE, but that was misread during encoding as DUB×ŠÈ (with a spurious grave accent, equivalent to subscript 3), where še$_3$ and eš$_2$ are values of the same sign 𒂠. Others are errors in earlier scholarship that were spotted after encoding. For example, the sign DUB×ŠE, which represents sign number 243 in [MZL], does not exist; it was listed in [MZL] based on a misreading of actual tablets in [gaz$_3$]; it should have been read GUM×ŠE.

~~This file consists of a sequence of sign and non-sign records.~~

~~Comments are indicated by the character U+0023 NUMBER SIGN (#); all characters from the number sign to the end of the line are ignored.~~

~~Lines of ogsl.asl are separated into fields by sequences of spaces or horizontal tabulations.~~

Example: ~~The following line consists of the fields `@sign` and `|GUD×KUR|`.~~

~~`@sign |GUD×KUR|`~~

~~3.2 Signs and forms A sign record begins with a line whose first field is @sign; the second field is the name of the sign according to the conventions described in *Section 2.5, Sign names*. It ends with the line @end sign.~~

~~**Example:** The following line marks the beginning of the sign record for 𒄠.~~

```
@sign |GUD×KUR|
```

~~A sign record may contain form records. Forms are variants of the signs; a form record begins with a line whose first field is @form, whose second field is the identifier of the form, which starts with U+007E TILDE (~), and whose third field is the name of the form, according to the same conventions as sign names. The form record is terminated by the line @end form, or by the beginning of an other form record or the end of the sign record.~~

~~**Example:** The following line within the sign |A.EDIN.LAL| marks the beginning of its form ~b.~~

```
@form ~b |A.EDIN.A.LAL|
```

~~A sign or a form record may have a line whose first field is @ucode. The second field then represents the encoding for that sign or form. The code points are in hexadecimal, prefixed by the letter x, and separated by U+002E FULL STOP (.).~~

~~**Examples:**~~

~~Within the record for sign |GUD×KUR|, its encoding is given as follows, where U+12120 is 𒄠.~~

```
@ucode x12120
```

~~Within the record for form |A.EDIN.A.LAL|, its encoding is given as follows, representing the sequence 𒀀𒂔𒀀𒇲.~~

```
@ucode x12000.x12094.x12000.x121F2
```

~~A sign or form may have lines whose first field is @list. The second field of such a line consists of a prefix identifying a sign list, followed by the number of that sign in that sign list.~~

~~**Example:** the sign record for 𒄠 has the following @list lines, indicating that it is sign number 124 in [LAK] and sign number 309 in [MZL].~~

```
@list LAK124
@list MZL309
```

~~A sign or form may have lines whose first field is @v. The last field of such a line is a value of the sign.~~

~~**Examples:** The sign record for 𒄠 has the following line, which indicates that it has the value am.~~

```
@v am
```

~~The sign record for 𒄭 has the following line, which indicates that it has the value bir₃; the second field indicates that the value is only used in Elamite.~~

```
@v %elx bir₃
```

~~The file ogsl.asl also contains non-signs; these are identical to signs except that they start with @nosign rather than @sign. These represent signs that do not exist, but were mistakenly catalogued in earlier sign lists or mistakenly encoded. Notes provide additional context.~~

~~**Examples:**~~

~~The character DUB×EŠ$_2$ 𒂏 was mistakenly encoded due to a misreading of MZL243 DUB×ŠE as DUB×ŠÈ (where šè and eš$_2$ are values of the same sign 𒂠).~~

~~The character DUB×ŠE in turn, which represents MZL243, does not exist; it was listed in [MZL] based on a misreading of GUM×ŠE 𒄤 in [gaz$_3$].~~

References

[ATF] Steve Tinney & Eleanor Robson. "Working with ATF to edit texts". *Oracc: The Open Richly Annotated Cuneiform Corpus*. http://oracc.museum.upenn.edu/doc/help/editinginatf/index.html

[BAU] Eric Burrows, *Archaic Texts* (Ur Excavations Texts 2; London 1935)

[ELLes] Pietro Mander, "Lista dei segni dei testi lessicali di Ebla", in *Materiali epigrafici di Ebla* 3, pp. 285-382. 1981.

[gaz$_3$] Miguel Civil, "Bloc-notes: sa-gaz$_x$(DUB×ŠE)--ak.", in *Revue d'Assyriologie et d'archéologie orientale* 60, p. 92. 1966.

[GASL] Steve Tinney. "ASL/OGSL File Format". *Oracc Global Sign List.* The OGSL Project, 2023. https://build-oracc.museum.upenn.edu/ogsl/aslogslfileformat/index.html

Review note: The above link needs to be updated to oracc.org rather than build-oracc.museum.upenn.edu once the OGSL is rebuilt.

[HZL] Christel Rüster & Erich Neu, *Hethitisches Zeichenlexikon* (Harrassowitz Verlag 1989)

[KWU] Nikolaus Schneider, *Die Keilschriftzeichen der Wirtschaftsurkunden von Ur III* (Rome 1935)

[LAK] Anton Deimel, *Liste der archaischen Keilschriftzeichen von Fara* (Wissenschaftliche Veröffentlichungen der Deutschen Orient-Gesellschaft 40; Berlin 1922)

[MÉA] René Labat, *Manuel d'épigraphie akkadienne* (6th ed. Paris 1988)

[MZL] Rykle Borger, *Mesopotamisches Zeichenlexikon* (Alter Orient und Altes Testament 305; Ugarit-Verlag 2003)

[ICE] Dean A. Snyder. "Cuneiform: From Clay Tablet to Computer". UTC document L2/00-398.

[aBZL] Catherine Mittermayer. *Altbabylonische Zeichenliste der sumerisch-literarische Texte*. 2006.

[OGSL] Niek Veldhuis, Steve Tinney, et al. “Oracc Global Sign List”. *Oracc: The Open Richly Annotated Cuneiform Corpus.*
http://oracc.museum.upenn.edu/ogsl/

[P010576] “CDLI Lexical 000014, Ex. 013 & 000027, Ex. 14 Artifact Entry.” 2001. Cuneiform Digital Library Initiative (CDLI). December 4, 2001.
https://cdli.ucla.edu/P010576

[P103303] “AUCT 1, 458 Artifact Entry.” 2001. Cuneiform Digital Library Initiative (CDLI). December 20, 2001.
https://cdli.ucla.edu/P103303

[P142296] “YOS 04, 232 Artifact Entry.” (2001) 2023. Cuneiform Digital Library Initiative (CDLI). February 1, 2023.
https://cdli.ucla.edu/P142296

[P225950] “CDLI Lexical 000010, Ex. 014 Artifact Entry.” 2003. Cuneiform Digital Library Initiative (CDLI). August 19, 2003.
https://cdli.ucla.edu/P225950

[P226934] “RIME 3/2.01.04.22, Ex. 01 Artifact Entry.” (2003) 2023. Cuneiform Digital Library Initiative (CDLI). June 14, 2023.
https://cdli.ucla.edu/P226934

[P232275] “RIME 3/1.01.07, St B Witness Artifact Entry.” (2003) 2023. Cuneiform Digital Library Initiative (CDLI). June 14, 2023.
https://cdli.ucla.edu/P232275

[P249253] “RIME 4.03.06.Add21, Ex. 01 Artifact Entry.” (2004) 2023. Cuneiform Digital Library Initiative (CDLI). June 15, 2023.
https://cdli.ucla.edu/P249253

[P281820] “BAM 3, 314 Artifact Entry.” 2005. Cuneiform Digital Library Initiative (CDLI). November 11, 2005.
https://cdli.ucla.edu/P281820

[P282017] “KAJ 002 Artifact Entry.” 2005. Cuneiform Digital Library Initiative (CDLI). November 11, 2005.
https://cdli.ucla.edu/P282017

[P360975] “AAA 1/3, 01 Artifact Entry.” 2007. Cuneiform Digital Library Initiative (CDLI). February 13, 2007.
https://cdli.ucla.edu/P360975

[P422664] “RINAP 5/1 Ashurbanipal 010, Ex. 001 Artifact Entry.” (2011) 2023. Cuneiform Digital Library Initiative (CDLI). February 1, 2023.
https://cdli.ucla.edu/P422664

[Principles] Michael Everson & Karljürgen Feuerherm. “Basic principles for the encoding of Sumero-Akkadian Cuneiform”.
UTC document L2/03-162.

[RÉC] François Thureau-Dangin, *Recherches sur l'origine de l'écriture cunéiforme* (Paris 1898)

[RSP] Yvonne Rosengarten, *Répertoire commenté des signes présargoniques sumériens de Lagash* (Paris 1967)

[ŠL] Anton Deimel, *Šumerisches Lexikon* (Rome 1925/1950)

[Syllabaire] François Thureau-Dangin, *Le Syllabaire Accadien* (Paris 1926)

[Unicode] *The Unicode Standard*
Latest version:
https://www.unicode.org/versions/latest/

[UAX38] *Unicode Standard Annex #38: Unicode Han Database (Unihan)*
Latest version:
https://www.unicode.org/reports/tr38/

[ZATU] Margret W. Green and Hans J. Nissen, *Zeichenliste der Archaischen Texte aus Uruk* (Archaische Texte aus Uruk 2; Berlin 1987)

Acknowledgements

Robin Leroy authored the bulk of the text, under direction from the Unicode Technical Committee.

Thanks also to the following people for their feedback or contributions to this document: Deborah Anderson, Peter Constable, Karljürgen Feuerherm, Erica Scarpa, Steve Tinney, Niek Veldhuis, Ben Yang.