L2/23-071

Unicode Cuneiform Sign Lists

To: SAH, UTC
From: Robin Leroy 𒊩𒌆
Date: 2023-03-24

We propose that the UTC authorize a Proposed Draft Unicode Technical Report based on this document.

A. Rationale

Character identity in the cuneiform script (specifically: Sumero-Akkadian Cuneiform, Xsux) is complex; it is impractical to ascertain it solely from the code charts. The abstract character répertoire is also vast, and includes code point sequences as well as individual code points; all of these have many names beyond those given in the code charts.

Assyriologists who use the encoded characters therefore depend on ancillary data for character identity; this is however not mentioned by the standard, and has led to confusion.

For instance:

1. the documentation of the Open Richly Annotated Cuneiform Corpus (Oracc) mentions the impracticality of the code charts: http://oracc.museum.upenn.edu/saao/knpp/cuneiformrevealed/aboutcuneiform/computercuneiform/index.html (see section *Unicode Cuneiform*);
2. some fonts in the Neo-Assyrian style get some mergers wrong: for instance, we are aware of a font created by an Assyriologist which has ≠ = , whereas the correspondence to sign lists given by the Oracc Global Sign List would give = ≠ , which is consistent with the evolution of those signs;
3. users have complained about the difficulty of using the code charts, especially for Neo-Assyrian texts, see, *e.g.*, https://www.unicode.org/mail-arch/unicode-ml/y2007-m06/0123.html.
   - Ken Whistler had suggested that a UTR be created to clarify this; unfortunately no action was taken: https://www.unicode.org/mail-arch/unicode-ml/y2007-m06/0129.html.

By clarifying that the encoding is *designed* to be used in conjunction with ancillary data, and by describing how Oracc provides the necessary data as part of the Oracc Global Sign List, we aim to alleviate that confusion, and facilitate the use of the encoded cuneiform script. It should be noted that the OGSL is the *de facto* authority on the use of the cuneiform script: it is maintained by the authors of the encoding proposals, and the tools that produce encoded cuneiform are based on it.

In addition, we hope to paint a clearer picture of cuneiform character identity within Unicode's own documentation.

Finally, by recognizing the role of the OGSL in establishing cuneiform sign identity, we hope to provide a basis for future proposals to address character identity issues raised by the OGSL project.

B. Proposed text

Unicode Cuneiform Sign Lists

Summary This document outlines the need for ancillary data in the use of the Sumero-Akkadian Cuneiform script, and describes how the Oracc Global Sign List provides that data.

Status *This is a **draft** document which may be updated, replaced, or superseded by other documents at any time.*

*This document is **not** a Proposed Draft Unicode Technical Report authorized by the UTC; it is a proposal presented to the UTC.*

Whoever cites this document as other than a work in progress, may Nabû and Šamaš cause his standards to be withdrawn!

1. Introduction

The Unicode Standard formally establishes the character identity of cuneiform signs by means of their names and representative glyphs in the code charts; see D2 in *Section 3.3, Semantics*, in [Unicode].

However, while the identity of abstract characters is well-established in the cuneiform script, the abstract characters are not usually referred to by standardized names, and the glyphic ranges of the abstract characters are vast and overlapping.

In practice, implementations of the script require an association of sequences of code points with entries in the classical sign lists that establish abstract character identity, and with the sign values which provide the usual names of these signs. Similar reliance on ancillary data may be found in other large scripts; see for instance *Unicode Standard Annex #38, Unicode Han Database (Unihan)* [UAX38].

This document briefly discusses the approach to the complexities of cuneiform sign identity taken by the encoding; it then describes the sign list maintained by the Open Richly Annotated Cuneiform Project (Oracc) which provides the ancillary data necessary to the effective use of the encoded script.

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.

Examples of such sign lists include [BAU], [ELLES], [HZL] [KWU], [LAK], [MEA], [MZL], [OBZL], [REC], [RSP], [SLLHA], and [ZATU].

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.

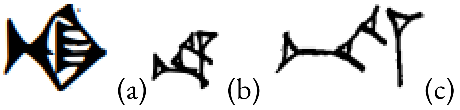


Figure 1: Glyphs for the sign NA 𒈾 in (a) Old Babylonian lapidary style (b) Old Babylonian cursive style (c) Neo-Assyrian style, as shown in [MEA].

These signs are the abstract characters of the cuneiform script. See also point 5 in [ICE].

2.1.1 Transliteration

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-u₂* 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 𒀸 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

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 readings unrelated to their components; for instance, the sign $GEME_2$ 𒊩𒆳 is always written like the sign SAL 𒊩 followed by the sign KUR 𒆳, even as these signs change across styles; the sign DIRI 𒋛𒀀 is always written as SI 𒋛 followed by A 𒀀.

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$_2$ 𒄱, which splits from DIN 𒁷 and merges with ḪI 𒄭.

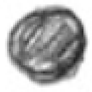
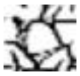
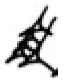
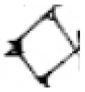
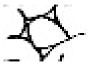
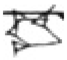
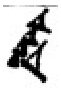
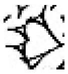
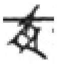
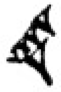
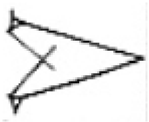
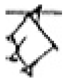
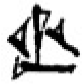
	Early Dynastic IIIa	Ur III	Old Assyrian	Middle Assyrian
𒊹 ŠAR$_2$	 [P010576]	 [P142296]		 [P281820]
𒄭 ḪI	 [P225950]	 [P142296]	 [P360975]	 [P282017]
𒄱 TI$_2$		 [P142296]	 [P360975]	 [P282017]
𒁷 DIN	 [P225950]	 [P103303]		 [P282017]

Figure 2: Mergers and splits of 𒊹, 𒄭, 𒄱, and 𒁷. The source of the hand copy shown is listed in each cell.

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

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 𒎌 (an Akkadian plural marker) originally looks like the sequence of syllables *me-eš* 𒈨𒌍, but their appearance diverges in Neo-Assyrian styles, as shown in Figure 3.

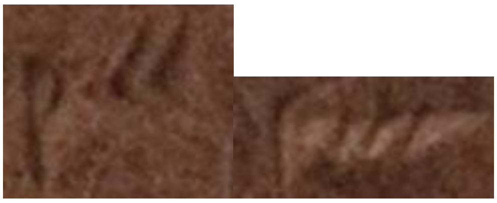


Figure 3: The sequence *me-eš* 𒈨𒌍 and the sign MEŠ 𒈨𒌍 on a Neo-Assyrian prism; photograph from [P422664].

2.4 Representative glyphs

As mentioned in *Section 2.1, Cuneiform Signs*, sign lists typically use a Neo-Assyrian style for their reference glyphs, even when illustrating a different style.

However, because many signs are merged in the Neo-Assyrian style, this was an impractical choice for the reference glyphs in the code charts; instead these reference glyphs are primarily in an Ur III style, where most signs are distinct; where a sign is unattested in the Ur III period, or where signs appear identical in the Ur III period, a different style was chosen for the sake of distinctiveness of the reference glyphs. For example, the reference glyph for ŠÁR 𒊹 is in an Early Dynastic style, because that sign merges with ḪI 𒄭 by the Ur III period; the reference glyph for 𒆰 is in a style that is Old Assyrian or newer, because it has not yet split from DIN 𒁷 in the Ur III period.

See also item 7 in [Principles], as well as *Fonts* in *Section 11.1, Sumero-Akkadian*, of [Unicode].

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.

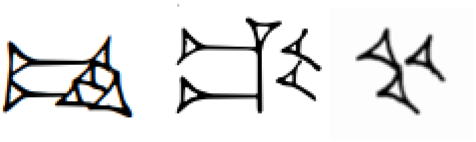


Figure 4: Neo-Assyrian glyphs for AM 𒄠, GUD 𒄞, and KUR 𒆳 from [MEA].

See also item 8 in [Principles].

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 produce encoded versions of transliterated texts, and to look up the glyphic range of a sign in various styles.

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 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 [GSL].

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
> ```

3.3 Lists

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. The abbreviations used in the reference section are the same as the prefixes used by OGSL.

> **Example:** the sign record for 𒂔 has the following @list lines, indicating that it is sign number 124 in [LAK] and sign number 309 in [MZL].
>
> `@list LAK124`
>
> `@list MZL309`

3.4 Values

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 𒂔 has the following line, which indicates that it has the value *am*.
>
> `@v    am`
>
> The sign record for 𒎙 has the following line, which indicates that it has the value *bir*$_3$; the second field indicates that the value is only used in Elamite.
>
> `@v    %elx bir₃`

3.5 Non-signs

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$_2$ 𒁿 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 𒄣 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.

[GSL] Steve Tinney. “GSL Source Format”. *Oracc: The Open Richly Annotated Cuneiform Corpus*.
https://github.com/oracc/ogsl/blob/master/00web/asl/index.html

[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)

[MEA] 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.

[OBZL] 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.

[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.

[REC] 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)

[SLLHA] Anton Deimel. *Šumerisches Lexikon. 1. Šumerische, akkadische und hethitische Lautwerte nach Keilschriftzeichen und Alphabet.* (Rome 1947)

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

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

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

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