Technical Reports

L2/25-266

Proposed Update Unicode Technical Report #56

UNICODE® CUNEIFORM SIGN LISTS

Version	2 (draft 11)
Editors	Robin Leroy (eggrobin@unicode.org)
Date	2025-09-16
This Version	https://www.unicode.org/reports/tr56/tr56-4.html
Previous Version	https://www.unicode.org/reports/tr56/tr56-3.html
Latest Version	https://www.unicode.org/reports/tr56/
Latest Proposed Update	https://www.unicode.org/reports/tr56/proposed.html
Revision	4

Summary

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

Status

*This is a **draft** document which may be updated, replaced, or superseded by other documents at any time. Publication does not imply endorsement by the Unicode Consortium. This is not a stable document; it is inappropriate to cite this document as other than a work in progress.*

> ***A Unicode Technical Report (UTR)** contains informative material. Conformance to the Unicode Standard does not imply conformance to any UTR. Other specifications, however, are free to make normative references to a UTR.*

Please submit corrigenda and other comments with the online reporting form [Feedback]. Related information that is useful in understanding this document is found in the References. For the latest version of the Unicode Standard see [Unicode]. For a list of current Unicode Technical Reports see [Reports]. For more information about versions of the Unicode Standard, see [Versions].

Contents

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 repertoire 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.

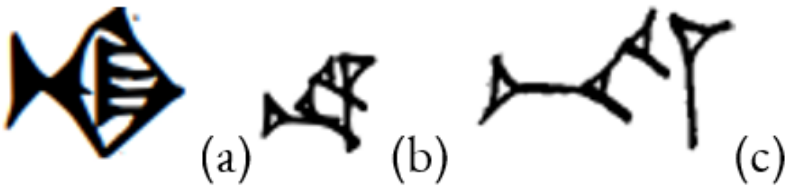
Examples of such sign lists include [aBZL], [BAU], [ELLes], [HZL] [KWU], [LAK], [MÉA], [MZL], [PTACE], [RÉC], [RSP], [ŠL], and [ZATU]. Notably, [ŠL] and [MÉA] use the same numbering; however, the other sign lists have different numbering schemes.

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. This is illustrated in Figure 1, which shows glyphs given in [MÉA] for the sign NA 𒀭 in three styles:

- Old Babylonian lapidary (a)
- Old Babylonian cursive (b)
- Neo-Assyrian (c)

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 𒀭.



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.

2.1.1 Transliteration

Review Note: The changes to this section have not yet been reviewed by the UTC, but are included for public review.

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 𒅁𒁺𒌑 and 𒅁𒁺𒌅, respectively. Note that while they share the phonetic value /u/, the signs U$_2$ 𒌑 and U 𒌅 are not stylistic variants of each other: they have distinct sets of values and meanings; for instance, 𒌑 means "grass" and 𒌅 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 normally be automatically generated from transliterated corpora. The reverse is not true; for instance, the sign 𒄿 might be transliterated *aš*, *ina*, or *dil*, depending on context.

There are occasional exceptions where a typical transliteration does not suffice to determine the cuneiform text. An example is the Eblaite version of the sign DIRI; DIRI is normally the sequence 𒋗𒄭 SI.A, but is written 𒄭𒋗 A.SI in Ebla instead, while still being transliterated diri or dirig in the literature on Ebla. When generating cuneiform from transliterations, either information about the provenience of the text should be taken into account to disambiguate these cases, or the transliterations should be adjusted to disambiguate. For instance, the Oracc Digital Corpus of Cuneiform Lexical Texts uses the transliteration dirig(A.SI) to unambiguously represent Eblaite dirig.

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

2.1.2 Numerals

Review Note: This section has not yet been reviewed by the UTC, but is included for public review.

The transliteration of numbers is less standardized. Transliterations that merely record the numeric value without also indicating the type of sign used cannot generally be used to automatically produce cuneiform text: in such a transliteration, 𒄿 and 𒁹 could both be transliterated as "1".

Other transliterations record the type of numeral, often together with an interpretation as part of a metrological system. For instance, in [ATF], 𒁹 could be transliterated as 1(barig) if it is a volume measure, or as 1(diš) if it is a count; 𒄿 could be transliterated as as 1(iku) as an area measure, or as 1(aš) as a count. These transliterations can be used to automatically produce cuneiform text. However, conventions differ as to whether the actual numeric value or only the multiplicity of the sign is recorded in the transliteration: [ATF] uses "1(u) 5(aš)" to transliterate 15 written 𒌋𒐗, whereas other systems use "10(U) 5(AŠ)". For corpora where the sexagesimal place value system is dominant, in particular in the first millennium, [ATF] allows for the sexagesimal places to be written in a so-called diš-less notation, wherein 1 implicitly represents 1(diš) 𒁹. Each sexagesimal place is a decimal number

in the range 1–60, which corresponds to one or two cuneiform signs : 10 represents 𒌋, and 32 represents the sequence 𒐼𒐹. Note that even in corpora that use diš-less notation, other types of numerals are transliterated in a qualified form, so that the type of numeric sign used remains unambiguous: the same text may have 15 ANŠE for 𒌋𒐹𒁹 (15 donkey-loads) and 1(u) 5(aš) GUN for 𒌋𒁾𒄩𒐹 (15 talents). See the *Metrology* page in [ATF]. Implementers should document what conventions they expect for numeric transliterations.

> **Note:** The Numeric_Value property of cuneiform signs corresponds to the multiplicity of the sign, rather than the numeric value represented, which depends on the metrological system. The sign U 𒌋 thus has Numeric_Value=1, rather than Numeric_Value=10. See *Cuneiform Numerals* in *Section 11.1.2, Cuneiform Numbers and Punctuation*, of [Unicode].

An additional complication when producing cuneiform text from transliterations of numeric expressions is that some variant stacking patterns for cuneiform numerals are separately encoded, even though they are rarely marked in transliteration. For instance, a transliteration 4(diš) can correspond to either U+12409 𒐉 or U+1243C 𒑜; likewise 7(diš), 8(diš), and 9(diš) can correspond to either 𒐝, 𒐞, 𒐟, or to 𒑜, 𒑝, 𒑞. The stacking pattern used primarily depends on the period and style; the style with rows of at most three wedges is more common in the Neo-Assyrian period, the style with two rows is more common in the Ur III period. When automatically generating cuneiform text from transliterations of Neo-Assyrian texts, 4(diš) should therefore generally be taken to correspond to 𒑜 rather than 𒐉.

There are some corpora where a contrast is recorded in transliteration between the 𒑜 and 𒐉 families of stacking patterns; these co-occur in some Ur III texts where the 𒑜 family is used in scratch calculations and the 𒐉 family is used in results. In that case, the 𒑜 family is transliterated as a variant, thus 4(diš@v) in [ATF]. This convention is reflected in [OSL], as well as in the character names: U+1243C 𒑜 is CUNEIFORM NUMERIC SIGN FOUR VARIANT FORM LIMMU, whereas U+12409 𒐉 is plain CUNEIFORM NUMERIC SIGN FOUR DISH.

The main reason for the disunification of stacking patterns, which would normally be a stylistic distinction, is the representability of sign lists that distinguish them, but otherwise present all signs in a consistent style; in particular, [MZL], whose cuneiform text is in Neo-Assyrian style, assigns different sign list numbers and sometimes different values to the variant stacking patterns: 𒑜 is number 860 with the value limmu, and 𒐉 is number 852 with the value limmu$_5$. Since that need does not extend to earlier periods, the stacking patterns used in the Early Dynastic period are not separately encoded, and the default versions of numeric signs should be used in these periods. For instance, the character U+12399 𒐼 should be used for Early Dynastic 2(u), even though the two stylus impressions are normally stacked vertically rather than horizontally in Early Dynastic tablets: the character U+12399 has the glyph 𒀸 in the Early Dynastic font [OFS-RSP].

2.2 Sequences

> Review Note: The changes to this section have not yet been reviewed by the UTC, but are included for public review.

Some signs can be analysed in most ~~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$_2$ 𒄑 is ~~always~~ written like the sign SAL 𒄑 followed by the sign KUR 𒆳, even as these signs change across styles; the sign DIRI 𒂗 is ~~always~~ written as SI 𒋗 followed by A 𒀀.

In cases where a sign can be analysed as a sequence both in the third millennium and in the Neo-Assyrian style, that sign is normally ~~Such signs are~~ not separately encoded; the corresponding sequence~~s~~ should be used to represent this ~~these~~ abstract character~~s~~. If the analysis as a sequence is applicable only in the third millennium, but not in Neo-Assyrian, or only in Neo-Assyrian, but not in the third millennium, the character is generally encoded atomically; examples of both are given in *Section 2.3.1, Mergers and Splits of Sequences*. See also items 2 and 5 in [Principles], and *Complex and Compound Signs* in *Section 11.1, Sumero-Akkadian*, of [Unicode]. An exception is made for signs that were taught as basic syllables as part of the early scribal curriculum, such as those in the sign

exercises Syllable Alphabets A and B (known to the scribes by their incipits 𒈨𒈨 ME-ME and 𒀀𒀀 A-A) or 𒌅𒋫𒋾 (TU-TA-TI); these basic syllables are then used later in the curriculum to describe pronunciations of more complex signs in sign lists such as Aa or Ea. The basic syllables have been encoded atomically, and should not be represented as sequences. For instance, according to the other encoding principles, the sign 𒅇 U$_3$ could be represented as the sequence 𒅆𒁲 IGI.DIB, or the sign 𒊻 UZ as 𒊺𒄷 ŠE.ḪU, but they are atomically encoded. See also item 4 in [Changes]. Note that the sequences can appear in cuneiform text when they are not read as the basic syllables:

Cuneiform	Transliteration	Translation	Representation of underlined text
𒉭𒊻𒄷	nunuz uz^{mušen}	duck eggs	𒊻 UZ
𒄿𒍪𒊻𒍪	*i-zu-uz-zu*	they will divide	
𒑏𒐈𒋡𒊺𒄷𒊺	1(ban$_2$) 3(diš) sila$_3$ še mušen niga	1 ban 3 sila (~13 l) of barley for the fattened birds	𒊺 ŠE followed by 𒄷 ḪU=MUŠEN
𒁕𒊺𒄷𒌝	da-še-ḫu-um	(a name)	
𒊭𒆷 𒄿𒉺𒀸𒊺𒄷	*ša la i-pa-aš-še-ḫu*	that cannot be soothed	

In all styles of cuneiform some signs that are analysed as sequences diverge in appearance from their components. Fonts targeting specific styles should include ligatures for these sequences as appropriate. This is discussed in *Section 2.6, Ligatures*.

> **Note:** While signs encoded as sequences are generally signs that originated as sequences, this is not always the case; some sequences are reanalyses that are not consistent with the earlier forms of the sign. For example, the sign 𒄘𒃼 IDIGNA, the name of the river Tigris, is encoded as the sequence GU$_2$.GAR$_3$, and the related sign 𒈦𒄘𒃼 DALLA, meaning “bright” or “fierce”, as MAŠ.IDIGNA=MAŠ.GU$_2$.GAR$_3$; this analysis is only applicable starting in the late third millennium: the glyph for Early Dynastic IIIb 𒈦𒄘𒃼 DALLA does not have a recognizable 𒃼 GAR$_3$, as illustrated here by the font [OFS-RSP].

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$_2$ 𒄰, which splits from DIN 𒁷 and merges with ḪI 𒄭. The source of the hand copy shown is given in each cell of the table.

Figure 2. Mergers and splits of 𒊹, 𒄭, 𒄰, and 𒁷.

	Early Dynastic IIIa	Ur III	Old Assyrian	Middle Assyrian
𒊹 ŠAR$_2$	 [P010576]	 [P142296]		 [P281820]

◇ HI	[P225950]	[P142296]	[P360975]	[P282017]
TI₂		[P142296]	[P360975]	[P282017]
DIN	[P225950]	[P103303]		[P282017]

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 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.

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

Review Note: The changes to this section have not yet been reviewed by the UTC, but are included for public review.

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~~ When such a sign has a distinctive appearance throughout the third millennium or in the Neo-Assyrian style, it is generally not considered as a sequence~~s~~ as described in *Section 2.2, Sequences*, and is ~~are~~ separately encoded. The special treatment of the Neo-Assyrian style is due to its status as the index form in most classical reference works. Fonts catering to more cursive styles may need to include many ligatures, as described in *Section 2.6, Ligatures*.

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

> **Note:** As in the single-character case, the term *split* refers to the divergence of the visual representations of two fixed character sequences, here and . That term does not refer to the phenomenon of a sign becoming a sequence of signs; indeed instead arose by two pre-existing signs coalescing into one.

Figure 3. The sequence *me-eš* and the sign MEŠ on the Neo-Assyrian prism [P422664].

As an example of a merger, the sign , whose Sumerian readings include šeš₂ “to anoint” and še₈ “to weep”, initially looks distinct from the sequence of unrelated signs SIKI.LAM , the first of which

means “hair” and the latter a kind of tree; this is the case in the reference glyphs. However, in later styles, the sign ŠEŠ₂ has the same appearance as the sequence SIKI.LAM .

> **Note:** The term *merger* refers to the convergence of the visual representations of two fixed character sequences, here and . As far as the scribes were concerned, the sign had broken up into a sequence of signs.

While the diachronic character identity used for the cuneiform encoding generally matches the understanding scribes had of character identity in their own script, there are discrepancies as scribes were not aware of mergers long past, let alone future splits. For example, some lexical texts describe explicitly the sign ŠEŠ₂ as being made up of the sequence , see [P467315.r.i.22].

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.

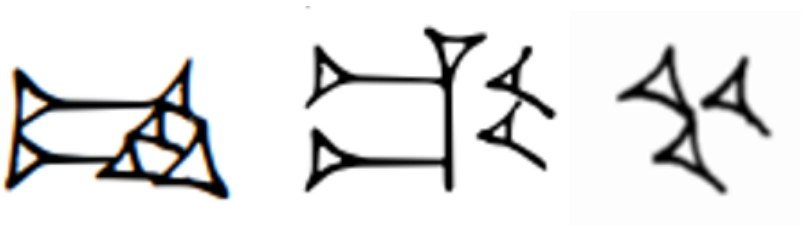
However, because many signs are merged in the Neo-Assyrian style, this was an impractical choice for the reference glyphs in the code charts; instead these reference glyphs are primarily in an Ur III style, where most signs are distinct; where a sign is unattested in the Ur III period, or where signs 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 ŠAR₂ ● is in an Early Dynastic style, because that sign merges with ḪI ◇ by the Ur III period; the reference glyph for TI₂ is in a style that is Old Assyrian or newer, because it has not yet split from DIN 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 is described as GUD×KUR (× , meaning inscribed inside), 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 , GUD , and KUR 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 CUNEIFORM SIGN NE SHESHIG (left) instead appears like NE×PAP ×. For comparison, the appearance of the sign NE on the same artifact is shown on the right.

Figure 5. The signs BIL₂ and NE on the stele of Hammurapi [P249253].

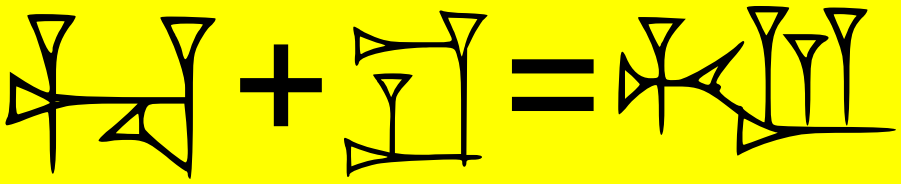
See also item 8 in [Principles].

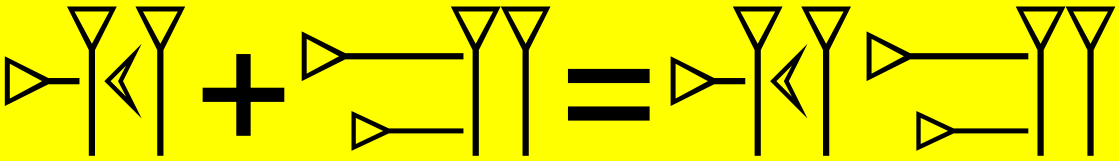
2.6 Ligatures

Review Note: This section has not yet been reviewed by the UTC, but is included for public review.

All styles of cuneiform require ligatures for some character sequences in order to properly capture the appearance of compound signs. As the analysis of signs as sequences takes into account their appearance in the Neo-Assyrian style, that style requires fewer ligatures. For example, the sign U_5 𒌋𒅅, whose meanings include “to ride”, is encoded as the sequence ḪU.SI. In some Early Dynastic styles and in the Neo-Assyrian style, no ligature is needed for this sign. However, in the style of Old Babylonian literary texts, a ligature should be used to capture the appearance of the U_5 sign. This is illustrated in Figure 6, which shows the sequence 𒌋𒅅 as displayed in an Old Babylonian literary font [OBF] and a Neo-Assyrian font [OFS-NAO].

Figure 6. The text 𒌋+𒅅=𒌋𒅅 shown with two cuneiform fonts.

[OBF] 

[OFS-NAO] 

The same ligatures that occur within a sign encoded as a sequence can also occur when that sequence corresponds to multiple signs. For instance, in the Hellenistic period, the sign 𒅅 DIRI is ligated, but that same ligature is used in occurrences that are read si-a; in the Ur III period, the sequence 𒌑𒀊 um-me is typically ligated as 𒌑. Note that while some transliterations use a single value for these sign sequences, such as sa_5 for for si-a or eme_2 for um-me, this practice is neither consistent nor strongly correlated with ligation.

Even the Neo-Assyrian style requires a few ligatures. Some are classically analysed as ligatures between separate signs, such as the very frequent 𒀸+𒋗=𒀸 *aš-šur*. Others are analysed as compound signs, such as 𒁀+𒌓=𒁀 *dul*(U.TUG_2), or variably transliterated as sequences or single signs, such as 𒉌 nenni, often transliterated BUL.BUL, where BUL is 𒉌.

In order to prevent a ligature between two signs, U+200C ZERO WIDTH NON-JOINER can be used; see *Non-joiner* in *Section 23.2.2, Cursive Connection and Ligatures*, of [Unicode]. When generating cuneiform text from transliterations, a zero width non-joiner should be inserted only where the transliteration marks an exceptional lack of joining. Since many ligatures occur not only within compound signs, but also between signs that are separately transliterated without the ligation being marked in the transliteration, it is not advisable to systematically prevent ligatures wherever the transliteration indicates a sign boundary with a hyphen or a dot.

Ligatures can occasionally occur across signs that are analyzed as being part of separate words; for instance, in Early Dynastic IIIb Ɲirsu, illustrated here by the font [OFS-RSP], the signs 𒊩 SAƝ and 𒄑 ƝAL$_2$ are ligated in 𒃻 𒊩𒄑 gur saŋ ŋal$_2$, a unit of volume. While, for searchability, it is generally preferable to separate words when generating cuneiform text, if interword ligatures are desired, the space between ligated words should be suppressed.

2.6.1 Discretionary Ligatures

Review Note: The changes to this section have not yet been reviewed by the UTC, but are included for public review.

On occasion, some sequences of signs may be combined in a ligature for stylistic effect, without that ligature being used systematically. This is illustrated in Figure 7, where the signs 𒀭 and 𒌓 are ligated on the inscription on the left, but not on the inscription on the right, even though the inscriptions are in consistent styles which could be expected to be covered by the same font. Such ligatures are not usually distinguished in transliteration from the corresponding sequences, so that both inscriptions would be transliterated ᵈsuen or ᵈEN.ZU; they do not carry distinct semantics. They are not separately encoded; it is left to the font to display these if desired, possibly based on the presence of a zero-width joiner; see *Joiner* ~~*Cursive Connection and Ligatures*~~ in *Section 23.2 2, Cursive Connection and Ligatures*~~*Layout Controls*~~, of [Unicode], and item 2 in [Principles]. When one needs to convey the ligature in transliteration, a plus sign is used, thus ᵈ⁺EN.ZU for the ligated example in Figure 7. When converting transliteration to cuneiform plain text, such a plus sign should be mapped to U+200D ZERO WIDTH JOINER.

Figure 7. The name of the god Sîn, 𒀭𒌓𒇻.

[P226934]	[P232275]
	

3 The Oracc Sign List

The Oracc Sign List [OSL] (formerly 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 Oracc Sign List is available as the machine-readable file https://github.com/oracc/osl/blob/master/00lib/osl.asl. A specification of the structure of that file may be found at [ASL].

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

For example, the signs 𒁹 BALAG and 𒁾 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 OSL correspond to a Unicode code point. Some signs are encoded as sequences, as described in Section *Section 2.2, Sequences*; the OSL documents the appropriate sequence. Other signs have no documented encoding. Some of them may be candidates for encoding; however, as the OSL 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 OSL, 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$_2$ 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). The grave accent is equivalent to subscript 3, and še$_3$ and eš$_2$ are values of the same sign, so the misreading DUB×ŠÈ was encoded as DUB×EŠ$_2$.

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$_3$]; the sign appearing on these tablets should have been read GUM×ŠE.

References

[aBZL]	Catherine Mittermayer. *Altbabylonische Zeichenliste der sumerisch-literarische Texte*. 2006.
[ASL]	Steve Tinney. “ASL/OSL File Format”. *Oracc Sign List.* The OSL Project, 2024. http://oracc.org/osl/asloslfileformat/
[ATF]	Steve Tinney & Eleanor Robson. “Working with ATF to edit texts”. *Oracc: The Open Richly Annotated Cuneiform Corpus*. http://oracc.org/doc/help/editinginatf/index.html
[BAU]	Eric Burrows, *Archaic Texts* (Ur Excavations Texts 2; London 1935)
[Changes]	Steve Tinney, Rationale for changes to N2664R. UTC document L2/04-080.
[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.
[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.
[OBF]	Corvin R. Ziegeler, *Old Babylonian Freie*, Version 2.0.0. November 2024. http://dx.doi.org/10.17169/refubium-44983 https://github.com/crzfub/OB-Freie/releases/tag/v.2.0.0
[OFS-NAO]	Steve Tinney, *Oracc NA Outline*. 2008. http://oracc.org/osl/OraccCuneiformFonts/ofs-nao/index.html
[OFS-RSP]	Steve Tinney, *Oracc RSP*. 2025. http://oracc.org/osl/OraccCuneiformFonts/ofs-rsp/index.html
[OSL]	Niek Veldhuis, Steve Tinney, et al. “Oracc Sign List”. *Oracc: The Open Richly Annotated Cuneiform Corpus.* http://oracc.org/osl/
[P010576]	“CDLI Lexical 000014, Ex. 013 & 000027, Ex. 14 Artifact Entry.” 2001. Cuneiform Digital Library Initiative (CDLI). December 4, 2001. https://cdli.earth/P010576

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

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

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

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

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

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

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

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

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

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

[P467315.r.i.22] Niek Veldhuis, et al. YOS 01, 53, reverse i 22. “Digital Corpus of Cuneiform Lexical Texts”. *Oracc: The Open Richly Annotated Cuneiform Corpus.*
http://oracc.org/dcclt/P467315.210

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

[PTACE] Amalia Catagnoti, “La paleografia dei testi dell’amministrazione e della cancelleria di Ebla”. *Quaderni di Semitistica* 30. 2010.

[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, Asmus Freytag, Sara Manasterska, Roozbeh Pournader, Erica Scarpa, Steve Tinney, Niek Veldhuis, Ken Whistler, Ben Yang, Corvin Ziegeler.

Modifications

The following summarizes modifications from the previous revision of this document.

Revision 4

- Section 2.1.1, Transliteration: Added a discussion of cases where usual transliterations are not sufficient to determine the cuneiform text.
- Added Section 2.1.2, Numerals: A discussion of practices in numeric transliteration, the disunification of stacking patterns, and the implications for generating cuneiform text.
- Section 2.2, Sequences: Significantly reworded to better reflect the nuances of the encoding model.
- Section 2.3.1, Mergers and Splits of Sequences: Reworded to take ligatures into account.
- Added Section 2.6, Ligatures: A discussion of non-discretionary ligatures.
- Section 2.6.1, Discretionary Ligatures: Added a recommendation to map transliteration + to ZWJ.

~~**Revision 3**~~

- ~~Publication of first approved version.~~

~~**Revision 2**~~

- ~~Advanced from Proposed Draft to Draft Unicode Technical Report.~~
- ~~Addressed feedback from the Editorial Committee.~~
- ~~Added an example of a sign-sequence merger and a note on scribal understanding of character identity.~~
- ~~Updated the references to OGSL to reflect its renaming to OSL.~~
- ~~Added a reference to PTACE.~~

~~**Revision 1**~~

- ~~Initial version following proposal L2/23-071 to the UTC.~~
- ~~L2/23-186: Added a section on discretionary ligatures.~~
- ~~L2/23-229:~~
  - ~~Rewrote Section 3 to reflect changes to the OGSL and its documentation.~~
  - ~~Clarified that glyphs may exhibit structures different from the ones described by the name.~~
  - ~~Clarified implications for fonts and input methods.~~
  - ~~Added some rationale for the encoding model and elaborated on the analogy with other large scripts.~~