Universal Multiple-Octet Coded Character Set
International Organization for Standardization
Organisation internationale de normalisation

L2/04-048

Doc Type: Working Group Document
Title: Proposal to Encode Cuneiform Ideographic Descriptors
Source: Dean A. Snyder, Johns Hopkins University
Status: Individual Contribution
Date: 2004-01-30

A. Administrative

1. Title:
Proposal to Encode Cuneiform Ideographic Descriptors in the SMP of the UCS
2. Requester's name:
Dean A. Snyder
3. Requester type (Member body/Liaison/Individual contribution):
Individual Contribution
4. Submission date:
2004-01-30
5. Requester's reference (if applicable):
6. Choose one of the following:
This is a complete proposal:
Yes
More information will be provided later:
No

B. Technical - General

1. Choose one of the following:
a. This proposal is for a new script (set of characters):
No
Proposed name of script:
. b. The proposal is for addition of character(s) to an existing block:
Yes
Name of the existing block:
Cuneiform and Cuneiform Numbers.
2. Number of characters in proposal:
14
3. Proposed category (select one from below - see section 2.2 of P&P document):
Category F - Archaic Hieroglyphic or Ideographic
4. Proposed Level of Implementation (1, 2 or 3) (see Annex K in P&P document):
Level 1
Is a rationale provided for the choice?
Yes
If Yes, reference:
Characters are ordinary spacing characters.
5. Is a repertoire including character names provided?

Yes

a. If YES, are the names in accordance with the "character naming guidelines" in Annex L of P&P document?

Yes

b. Are the character shapes attached in a legible form suitable for review?

Yes

6. Who will provide the appropriate computerized font (ordered preference: True Type, or PostScript format) for publishing the standard?

Either I will do this or these characters could be added to Steve Tinney's Cuneiform Classic Font.

If available now, identify source(s) for the font (include address, e-mail, ftp-site, etc.) and indicate the tools used:

If I do the font, I will create a TrueType font with Fontographer.

Dean A. Snyder
24 West Railroad Avenue
Shrewsbury, Pennsylvania, USA 17361
dean.snyder@jhu.edu

(See proposal N2698 for information about Steve Tinney.)

7. References:

a. Are references (to other character sets, dictionaries, descriptive texts etc.) provided?

Yes

b. Are published examples of use (such as samples from newspapers, magazines, or other sources) of proposed characters attached?

Yes

8. Special encoding issues:

Does the proposal address other aspects of character data processing (if applicable) such as input, presentation, sorting, searching, indexing, transliteration etc. (if yes please enclose information)?

Yes.

9. Additional Information:

Submitters are invited to provide any additional information about Properties of the proposed Character(s) or Script that will assist in correct understanding of and correct linguistic processing of the proposed character(s) or script. Examples of such properties are: Casing information, Numeric information, Currency information, Display behaviour information such as line breaks, widths etc., Combining behaviour, Spacing behaviour, Directional behaviour, Default Collation behaviour, relevance in Mark Up contexts, Compatibility equivalence and other Unicode normalization related information. See the Unicode standard at http://www.unicode.org for such information on other scripts. Also see http://www.unicode.org/Public/UNIDATA/UCD.html and associated Unicode Technical Reports for information needed for consideration by the Unicode Technical Committee for inclusion in the Unicode Standard. 1 Form number: N2652-F (Original 1994-10-14; Revised 1995-01, 1995-04, 1996-04, 1996-08, 1999-03, 2001-05, 2001-09, 2003-11)

C. Technical - Justification

1. Has this proposal for addition of character(s) been submitted before?

No

If YES explain

2. Has contact been made to members of the user community (for example: National Body, user groups of the script or characters, other experts, etc.)?

Yes

If YES, with whom?

Deborah Anderson, Visiting Scholar, Linguistics, Univ. of California at Berkeley

Lloyd Anderson, Linguist , Font Vendor, Ecological Linguistics
Richard Averbeck, Prof. Old Testament & Semitic Languages, Trinity International University
Robert Black, PhD Candidate, Near Eastern Studies, Johns Hopkins
Rykle Borger, Seminar für Keilschriftforschung, Goettingen University
Giorgio Buccellati, Prof. Emeritus, Department of Near Eastern Languages & Cultures, Department of History, Director of the Institute of Archaeology's Mesopotamian Laboratory, UCLA
Carl-Martin Bunz, M.A., Indo-European Linguist, University of Saarland, Germany
Miguel Civil, Emeritus Professor of Sumerology, Oriental Institute, Univ. Chicago, Editor, Materials for the Sumerian Lexicon
Jerrold Cooper, Prof. of Assyriology & Sumerian, Johns Hopkins
Robin Cover, SGML/XML, Oasis
T. R. Davis, Lecturer in Bibliography & Palaeography, Univ. of Birmingham, England
Patrick Durusau, Dir. Research & Development, Society of Biblical Literature, Emory Univ.
Robert Englund, Prof. of Assyriology & Sumerian, UCLA
Michael Everson, Font Vendor, Everson Typography, Ireland
Karljürgen Feuerherm, PhD in Akkadian, Univ. of Toronto
Madeleine Fitzgerald, Visiting Assistant Prof., Department of Near Eastern Languages & Cultures, UCLA, NSF Digital Libraries Initiative Postdoctoral Fellow for the Cuneiform Digital Library Initiative
Eckart Frahm, Assistant Prof, Assyriology, Department of Near Eastern Languages & Civilizations, Yale Univ.
Gene Gragg, Prof. of Near Eastern Languages & Linguistics, Oriental Institute, Univ. of Chicago
William Hallo, Prof. Emeritus, Department of Near Eastern Languages & Civilizations, Yale Univ.
Edwin Hart, Senior Computing Staff, Applied Physics Laboratory, Johns Hopkins
Harry A. Hoffner, Jr., The John A. Wilson Professor of Hittitology Emeritus, Co-editor, Hittite Dictionary, Oriental Institute, Univ. Chicago
Hermann Hunger, Prof. Assyriology, Altsemitische Philologie und Orientalische Archäologie, Institut für Orientalistik, Universität Wien
John Jenkins, System Software Engineer, Apple, Unicode Technical Director
Cale Johnson, PhD Candidate, Department of Near Eastern Languages & Cultures, UCLA, Cuneiform Digital Library Initiative staff
Charles Jones, Research Associate - Bibliographer, Oriental Institute, Univ. of Chicago
Alasdair Livingstone, Reader in Assyriology, Univ. of Birmingham, England
John McGinnis, PhD Cambridge, England
Rick McGowan, Vice President, Unicode
Piotr Michalowski, Prof.f Ancient Near Eastern Languages & Civilizations, Department of Near Eastern Studies, Univ. Michigan, Editor-in-chief, Journal of Cuneiform Studies
David Owen, Prof. of Ancient Near Eastern & Judaic Studies, Cornell Univ.
Gerfrid Müller, Institut für Altertumswissenschaften, Universität Würzburg
Simo Parpola, Prof. of Assyriology, Univ. of Helsinki
Philip Payne, Font Vendor, Linguist's Software
Gonzalo Rubio, Asst. Prof of Assyriology, Ohio State Univ.
Eric Smith, Graduate Student, Dept. of Linguistics, Univ. Toronto
Dean A. Snyder, Assistant Research Scholar, Manager, Digital Hammurabi, Johns Hopkins University
Matthew Stolper, Prof. of Assyriology, Oriental Institute, Univ. Chicago
Jonathan Taylor, an editor for Electronic Text Corpus of Sumerian Literature, Oriental Institute, Univ. Oxford
Steve Tinney, Associate Prof. of Assyriology & Sumerian, Editor, Pennsylvania Sumerian Dictionary, Univ. of Pennsylvania
Niek Veldhuis. Assistant Prof. Assyriology, Department of Near Eastern Studies, UC Berkeley
Lee Watkins, Jr., Director, Center for Scholarly Resources, Director, Digital Hammurabi, Johns Hopkins
Bruce Wells, PhD Near Eastern Studies, Johns Hopkins
Kenneth Whistler, Software Engineer, Sybase, Unicode Technical Director, Managing Editor, The Unicode

Standard
Christopher Woods, Assistant Prof. Assyriology, Oriental Institute, Univ. Chicago
If YES, available relevant documents:
Archives of discussions on cuneiform@unicode.org and unicode@unicode.org can be downloaded. Also much of the discussion was in person at conferences and via telephone.
3. Information on the user community for the proposed characters (for example: size, demographics, information technology use, or publishing use) is included?
The international cuneiform scholarly community can be numbered in the hundreds. Members are found on every continent.
Reference:
Common knowledge in the scholarly community.
4. The context of use for the proposed characters (type of use; common or rare)
Rare
Reference:
See N2698.
5. Are the proposed characters in current use by the user community?
No
If YES, where? Reference:
6. After giving due considerations to the principles in the P&P document must the proposed characters be entirely in the BMP?
No
If YES, is a rationale provided?
If YES, reference:
7. Should the proposed characters be kept together in a contiguous range (rather than being scattered)?
Yes
8. Can any of the proposed characters be considered a presentation form of an existing character or character sequence?
No
If YES, is a rationale for its inclusion provided?
If YES, reference:
9. Can any of the proposed characters be encoded using a composed character sequence of either existing characters or other proposed characters?
No
If YES, is a rationale for its inclusion provided?
If YES, reference:
10. Can any of the proposed character(s) be considered to be similar (in appearance or function) to an existing character?
No
If YES, is a rationale for its inclusion provided?
If YES, reference:
11. Does the proposal include use of combining characters and/or use of composite sequences?
No
If YES, is a rationale for such use provided?
If YES, reference:
Is a list of composite sequences and their corresponding glyph images (graphic symbols) provided?
If YES, reference:
12. Does the proposal contain characters with any special properties such as control function or similar semantics?
No
If YES, describe in detail (include attachment if necessary)

13. Does the proposal contain any Ideographic compatibility character(s)?
No
If YES, is the equivalent corresponding unified ideographic character(s) identified?
If YES, reference:

D. Proposal

1. Introduction

Taking the large view of the Sumero-Akkadian cuneiform script system, we observe that it experienced two major phases - an earlier phase marked by more extensive use of logograms with a concomitant higher productivity in new sign generation for new words, and a later phase marked by a greater use of the same signs as syllabograms with markedly less productivity in the generation of new signs. One could therefore call the earlier phase dynamic and the later phase static, with regards to new sign generation.

The earlier scribes formed new signs for new semantics in two ways - they either created completely new signs or they modified existing ones. Such modifications include sign rotation, flipping, curving, infixing, etc. Fourteen such modifications of cuneiform signs have been identified and provide the basis for this proposal.

An international group of approximately fifty scholars, the Initiative for Cuneiform Encoding, has, for over four years, been working on a proposal to encode cuneiform in ISO10646/Unicode. Early on, the decision was made, for good reasons, to do a static encoding of cuneiform. This choice drove, in turn, the decision to limit the encoding to the basic and modified signs occurring in the later (URIII and on) periods, where the script is much better understood than in the earlier periods. The Everson, Feuerherm, Tinney proposal, N2698, “Revised proposal to encode the Cuneiform script in the SMP of the UCS” is the result of that process and those decisions. N2698 gives us an excellent and simple to implement encoding for later cuneiform with, however, the unfortunate side effect that that part of early cuneiform unencoded in N2698 is now left with no standard encoding.

This proposal attempts to remedy that situation by proposing 14 cuneiform ideographic descriptors.

A quick check of the signs in N2698 shows that of the 985 proposed signs, almost three-fourths of them are actually modified versions of the remaining one fourth, which are basic, or simple, signs. It is theoretically possible to dynamically generate in software all 985 signs from the 280 encoded basic signs plus the 14 proposed cuneiform sign modifiers, following, in modern software, much the same model the ancient scribes employed. But complexities in operating system, font, and application design, combined with market realities (the modern cuneiform user community is numbered in the hundreds) all argue against taking such an approach now.

This has left us with proposing a model similar to that adopted for Han ideographic descriptors.

2. Encoding Issues

Due to the relative dearth of knowledge about early cuneiform script, both new analyses of known signs and discoveries of new signs are being made all the time as research continues. And even though we foresee adding new signs to the 10646/Unicode, at no point will we accept that the entire ancient cuneiform sign repertoire has been encoded. Moreover the acceptance of new sign interpretations by the scholarly community can take years.

The 14 characters proposed for the Cuneiform Ideographic Description block (see Figure 1) provide a mechanism for the standard interchange of text that must reference unencoded cuneiform signs.

The basic graphic design shared by all the cuneiform ideographic descriptors in this proposal was chosen for two reasons:

1) It clearly associates the character with the cuneiform script.

2) It provides the basis for the unambiguous graphic representation of all modifications proposed.

Unencoded signs can be described using these characters plus encoded cuneiform signs; the reader can then create a mental picture of the ideographs from the description. This process is different from a formal encoding of cuneiform signs. There is no canonical description of unencoded signs; there is no semantic content assigned to described signs; there is no equivalence defined for described signs. As with the Han Ideographic Descriptors, conceptually, cuneiform ideograph descriptions are more akin to the English phrase, “an ‘e’ with an acute accent on it,” than to the character sequence <U+006E, U+0301>.

Support for the characters in the Cuneiform Ideographic Description block does not require the rendering engine to recreate the graphic appearance of the described character, although sophisticated ones may choose to do so. Note also that many of the cuneiform signs that users might represent using the Cuneiform Ideographic Description characters are expected to be formally encoded in future versions of the Unicode Standard.

3. Syntax

The syntax for Cuneiform Ideographic Descriptions Sequences includes the following rules:

1 The 14 Cuneiform Ideographic Descriptors can only be used to describe modifications to encoded cuneiform signs - U+12000 - U+123FF.

2 The descriptors are divided into 3 major groups - decorators, orienters, and positioners.

3 All descriptors follow the encoded cuneiform sign whose modification they describe.

4 Decorators and orienters are unary descriptors - they describe modifications for only one cuneiform sign at a time.

5 The positioners are binary descriptors - - they describe positional relationships between two cuneiform signs.

6 The unary descriptors can occur in any order.

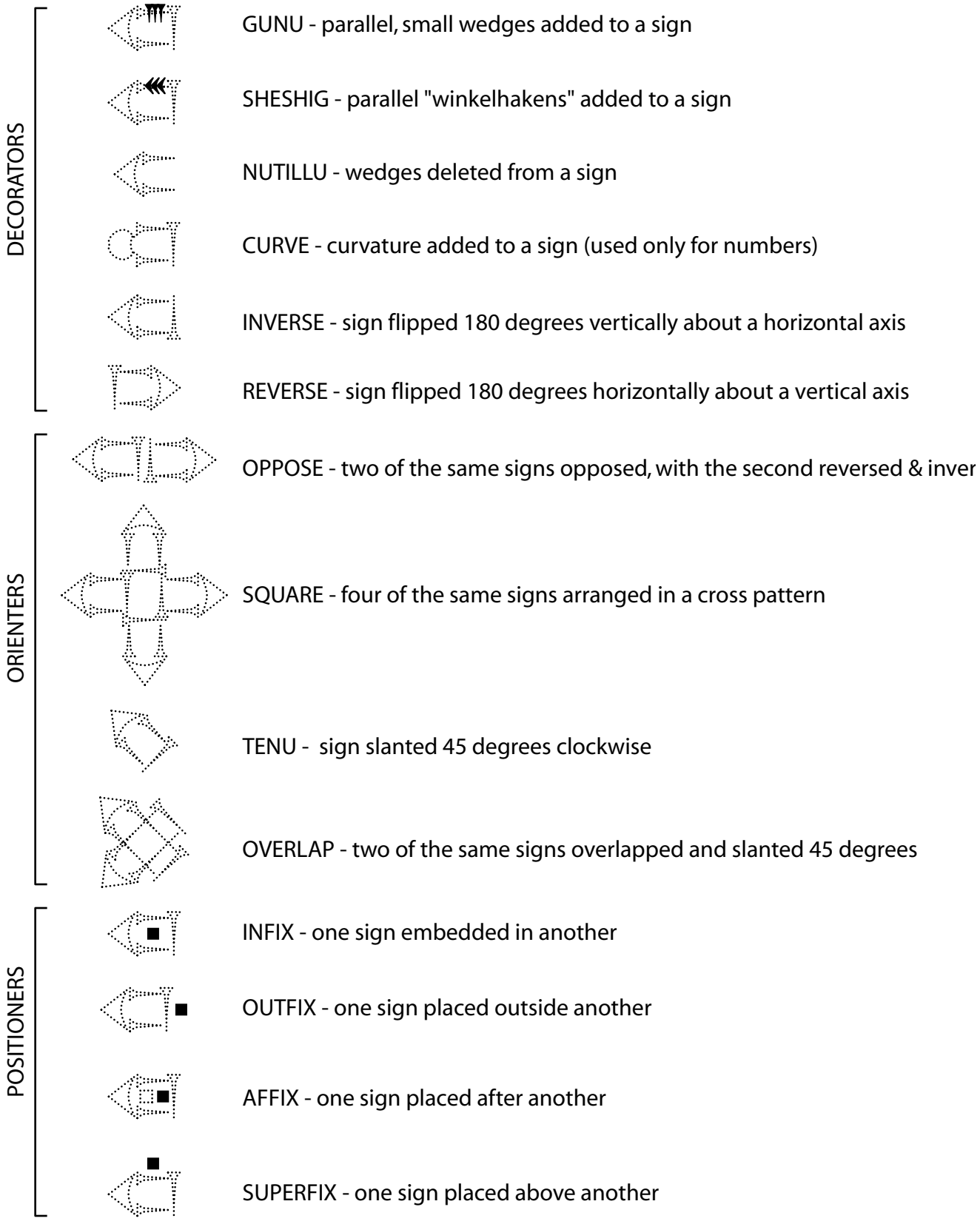
7 Any combination of decorators is permissable - they can all be applied to the same encoded cuneiform sign.

8 Of the decorators, only GUNU and SHESHIG can occur more than once after any given encoded cuneiform sign.

9 Only one orienter can be applied to an encoded cuneiform sign at a time.

10 Except for the OUTFIX descriptor (see below), only one positioner can be applied to an encoded cuneiform sign at a time.

Figure 1. Cuneiform Ideographic Descriptors



11 The binary descriptors follow the cuneiform sign they modify but after all unary operators associated with the same cuneiform sign.

12 Multiple binary descriptors and encoded cuneiform signs can be chained to form a single unencoded cuneiform sign.

13 The INFIX descriptor increases the level of sign embedding and indicates that the following encoded cuneiform sign is to be positioned inside the preceding encoded cuneiform sign. This allows for nested levels of infixing.

14 The OUTFIX descriptor decreases the level of sign embedding and indicates that the following encoded cuneiform sign is to be positioned outside the preceding encoded cuneiform sign.

15 The AFFIX descriptor indicates that the following encoded cuneiform sign is to be positioned after, and at the same level of nesting as, the preceding encoded cuneiform sign.

16 No more than one binary descriptor can occur between two encoded cuneiform signs, except for the OUTFIX descriptor which can occur multiple times contiguously in order to back out from multiply embedded levels created by multiple INFIX descriptors.

17 A binary descriptor must be followed immediately by an encoded cuneiform sign, with the exception that multiple OUTFIX descriptors can occur contiguously as outlined above.

18 The maximum allowable number of characters in any given Cuneiform Ideographic Description Sequence is sixteen.

Any character sequence not conforming to this syntax is not a Cuneiform Ideographic Description Sequence.

4. Usage

The utility of Cuneiform Ideographic Description Sequences depends on the fact that approximately 75% of cuneiform signs can be broken down into smaller pieces that are themselves independent cuneiform signs, and also because it is more likely that new compound and complex cuneiform signs will be discovered in early cuneiform rather than new basic signs. Therefore it is expected that the vast majority of unencoded cuneiform signs can be encoded using the encoded signs plus the proposed sign modifiers. (But we have made no attempt in this proposal to deal with cuneiform number description.)

A user wishing to represent an unencoded cuneiform sign will need to analyze its structure to determine how to describe it using a Cuneiform Ideographic Description Sequence. Although a given cuneiform sign can be described in various ways given the syntax outlined above, typically the shortest possible Cuneiform Ideographic Description Sequence is to be preferred. The length constraint allows random access into a string of ideographs to have well-defined limits. Only a small number of characters need to be scanned backward to determine whether those characters are part of a Cuneiform Ideographic Description Sequence.

Even though many unencoded cuneiform signs can be described in more than one way using this syntax, we do not define equivalence for two Cuneiform Ideographic Description Sequences that are not identical. In particular, Cuneiform Ideographic Description Sequences are not to be used to provide alternative graphic representations of encoded ideographs; searching, collation, and other content based text operations would then fail.

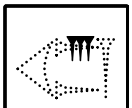
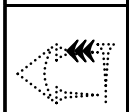
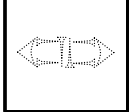
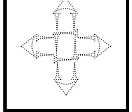
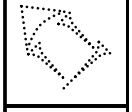
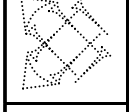
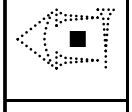
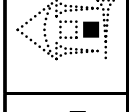
Cuneiform Ideographic Description characters are visible characters. They are not to be treated as control char-

acters. The sequence U+12000 U+1240A U+12043 typically would have a distinct appearance from U+12002, except in more sophisticated software that chooses to render them the same. An implementation may render a valid Cuneiform Ideographic Description Sequence either by rendering the individual characters separately or by parsing the Cuneiform Ideographic Description Sequence and drawing the sign so described. In the latter case, the Cuneiform Ideographic Description Sequence should be treated as a ligature of the individual characters for purposes of hit testing, cursor movement, and other user interface operations.

Cuneiform Ideographic Description characters are not combining characters, and there is no requirement that they affect character or word boundaries. Thus U+12000 U+1240A U+12043 may be treated as a sequence of three characters or even three words. Implementations of the Unicode Standard may choose to parse Cuneiform Ideographic Description Sequences when calculating word and character boundaries, but such a decision will, of course, make the algorithms involved significantly more complicated and slower.

Cuneiform Ideographic Description Characters

These are visibly displayed graphic characters, not invisible composition controls.

1240		
0		12400 CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER GUNU
1		12401 CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER SHESHIG
2		12402 CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER NUTILLU
3		12403 CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER CURVE
4		12404 CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER INVERSE
5		12405 CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER REVERSE
6		12406 CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER OPPOSE
7		12407 CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER SQUARE
8		12408 CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER TENU
9		12409 CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER OVERLAP
A		1240A CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER INFIX
B		1240B CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER OUTFIX
C		1240C CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER AFFIX
D		1240D CUNEIFORM IDEOGRAPHIC DESCRIPTION CHARACTER SUPERFIX