1. Introduction (Overhead: **Figure 1**)

At the last *Rencontre Assyriologique Internationale* (Paris, July 2000), I presented a proposal by the *Computer Representation of Cuneiform* (CRC) project outlining fundamental principles whereby cuneiform documents might be most usefully electronically encoded for the purposes of text processing and analysis: in other words, a ground plan for a cuneiform character code capable of facilitating the exchange of cuneiform information, *as cuneiform*, across temporal, geographic, and script-tradition boundaries.

The proposal was for the most part received with enthusiasm, and we were encouraged to move forward. Over the next few months, CRC engaged in active dialogue with other interested parties, and we were invited, along with representatives of the Unicode Consortium and several prominent assyriologists, to participate in a two (2) day conference dedicated to the subject at Johns Hopkins University, Baltimore, in early July, under the heading *Initiative for Cuneiform Encoding* (ICE). By the end of the conference, we had reached agreement on all essential matters, and within the week, a precis of the key issues was prepared and submitted to the Unicode Consortium for official consideration. I am pleased to report that this precis was accepted in principle at a meeting of the Unicode Technical Committee, and that the Cuneiform Character Code, under the auspices of Unicode, is now an official work in progress.

2. Encoding Scope (Overhead: **Figure 2**)

The goal of the encoding process is to facilitate the entering, processing, manipulating, and analysing of cuneiform text using all the sophistication presently available for modern scripts. This would include searching and sorting, dictionaries, spelling checkers, email, and so on.

In order to meet the needs of all prospective users, the encoding will rely as a starting point upon an architecture based upon the mainstream Sumero–Akkadian development of the script, with account taken where necessary of any and all specific concerns pertinent to other significant script traditions.

At the present time, the lower temporal boundary of the encoding is still under discussion, in that the repercussions of the inclusion or exclusion of archaic stages of the script have not been sufficiently investigated. One issue may suffice to make the point: in round figures, 600 signs (not 'characters') belonging to the NeoAssyrian script period reflect a shrinkage from 900 in the UR III script period, and this in turns represents a significant shrinkage from the archaic period. The inclusion of the archaic period script, while desirable from a comprehensiveness point of view, may well be undesirable in light of the added complexity which would ensue. The effect on the encoding of accommodating shrinkage will be illustrated by example further on, and the value of keeping it to a minimum will, I hope, then become self-evident.

3. Of Signs and Characters (Overhead: **Figures 3, 4**)

In general, 'signs' in the sense commonly understood by cuneiformists and 'characters' in the sense commonly understood by linguists or computer scientists are very roughly speaking equivalent. However, the needs of the encoding system require that 'characters' conform rigorously to certain principles which one may or may not envisage when speaking of 'signs'. Specifically:

a. 'characters' are defined abstractly without recourse to the external appearance they display in this or that instance;

b. all other things being equal, ‘characters’ are differentiated according to their function within the script.

As one can see from **Figure 3**, any given character can be represented by a variety of shapes depending upon temporal and geographic locus. For this reason, shapes used to represent a character in one instance may represent a completely different character in another instance (*cf.* the before-last instances of [UD] and [ERIN$_2$] in **Figure 3**). In certain cases, substantial variation in shape (*cf.* the appearance of the characters [UD] and [AN], encoded as [4] and [10] respectively in **Figure 4**) may occur even within a single document. This sort of variation is not reflected in a character code, which tracks identity alone. If desired, however, such differences could be tracked by an external mechanism such as markup tagging.

4. Other Potential Sign/Character Divergences (Overhead: **Figures 5, 6, 7, 8**)

There are at least (3) other areas in which the analysis of text from a traditional ‘sign’-based perspective may differ from that of a ‘character’-based perspective:

a. ligatures (**Figure 5**);

b. (apparent?) alternate sign choices (**Figure 6**);

c. compounds (**Figure 7**).

a. Ligatures. There are a number of ‘signs’ in the traditional understanding which result from the ligature of two or more other ‘signs’, in a manner similar to the modern ‘o/e’ ligature or in some type, ‘f/l’ ligature; **Figure 5** lists a few of these. The key issue here is that the underlying information in the text is the same, whether or not a writer chose to use a ligature at this or that point in time or not. Because the purpose of a character code is to reflect underlying textual data, rather than its visual presentation, the encoding should enable this. The proposed approach would consistently encode the underlying components, but indicate potential ligature situations by a special operator known as a *Zero-width Joiner*. Given an appropriate supporting font, application software would then be able *either* to display the component characters as they generally appear in isolation from one another *or* the special ligature form globally, or on a case by case basis, at the user's option.

b. (Apparent alternate sign choices) is a slightly more difficult matter. In the top portion of **Figure 6**, we find the sign traditionally identified as LAḪ$_4$, which appears structurally as two DU signs, one above the other. This could either be considered a compound, a type of ligature of DU with itself, or a sign in its own right. In the figure, I have supposed, for argument's sake, that the latter view has been chosen. In that instance, apparent sequences of two DU signs serving the function generally represented by LAḪ$_4$ are probably better interpreted as the same character written differently, perhaps due to space restrictions, parallel to what a modern writer might do when approaching the right margin of a sheet of paper, and opting to write up along the edge. Because this variation in representing the character is systemic and relates to character identity, and in that respects differs from the incidental variations in appearance which are purely arbitrary or developmental, one might wish to encode such differences as subtypes of a given character (*cf.* **Figure 6**). Similarly, systemic abbreviations such as the third example would be encoded as subtypes of the same base character, and *not* as a different character (in this case, [DU]).

In the bottom portion of **Figure 6**, we have an example of what are usually considered to be distinct signs, but which in this case are related by means of the script feature known as *gunû*: in essence, a *gunûfied* sign is created by adding some hash marks to the base sign; the effect or interpretation varies from case to case, and is not all that well understood.

The issue here is that frequently, one sign may be used in a context where the other is normative, and in many cases, due to the prevailing writing practice, it is not in reality possible to differentiate between the two. Thus, 'Ur' is variously written as ŠEŠ-AB or ŠEŠ-UNUG, as the case may be. Here again, it may be preferable to encode a common base character (in this case, [AB]), with some other special code indicating whether or not *gunûfication* is present. This parallels to some extent the similarities and differences in extended Latin script between the character [A] and, say [Ä]. In the same way that many word processors can search text and treat accents and other diacritics as significant or not, the proposed approach to *gunû* could be used to differentiate [AB] from [AB]-*gunû* or not.

c. Compounds. By compounds, I mean here those signs which are formed by combining together two or more base signs with some non-trivial effect on significance. For the purposes of this presentation, I will consider two types of compounds, which I will name *sequential compounds* and *composite compounds*.

Sequential compounds are, as the names implies, compounds of signs occurring in a sequence, but with a function different or not obviously related to the same sequence of signs interpreted individually. In **Figure 7**, the top line shows a sequence of three signs, in which each functions as an individual item within the text in the usual manner; no special encoding is necessary here. On the second line, we have a situation where the same two signs IGI and RU occur in a context where they are generally thought of as a single sign PÀD, whose functional significance is not related in any simple fashion to the functional significance of IGI in combination with the functional significance of RU. Nevertheless, within many texts, the two components of PÀD which in other contexts are interpreted as IGI and RU, respectively, behave on the physical level as though they were entirely independent entities. For this reason, it may be preferable to encode them as separate entities joined by an appropriate operator (shown as an equals sign in **Figure 7**) similar to the *Zero-width Joiner* mentioned above, though of course not identical with it, since ligation in the conventional sense is not involved.

Composite compounds are compounds of signs produced by treating one sign as a base or container, and then affixing or infixing another single sign or sign compound. There are two ways to deal with these, each with advantages and disadvantages; it may be that the most useful encoding scheme will use one or the other approach on a case by case basis (*cf.* **Figure 8**).

Encoding composite compounds in terms of their component characters joined by a suitable operator (represented by x in **Figure 8**) means that such characters could presumably be decomposed as desired (in a manner similar to ligatures), which could be handy for didactic purposes, and that they would likely be sorted implicitly according to the sort order specified for the base character followed by the affixed or infixed characters; in other words, they would naturally fall into their 'conventional' sort sequence without the need for additional support. On the other hand, in the case of some composite compounds, the infixed or affixed modifiers may be difficult or impossible to ascertain, and it may well be that due to the development of individual sign shapes over time, the 'natural' sort order alluded to above may not in fact be the most desirable.

5. Mergers and Splits (Overhead: **Figures 9, 10**)

The issue of shrinkage was alluded to briefly above in connection with the question of whether or not to include the archaic period within the present encoding scheme. Shrinkage arises for a number of reasons. In some cases, a sign (or for our purposes, character) simply falls into disuse; this poses no significant problems, since texts beyond this point will simply fail to reflect such a character. A more important cause of shrinkage is the phenomenon of sign merger.

Mergers occur when two or more signs, over the course of their development, come to assume a sufficiently similar shape that they are no longer practically differentiable, though they clearly once were. **Figure 9** shows several instances of sign mergers.

The only straightforward way of dealing with the fact of mergers is to define as many characters as there are differentiable items at the widest point, and to encode accordingly across the board. This poses no problem for display purposes, since a font used for texts beyond the merger point need only have the same visual symbol assigned to each slot. There is an implication for searching and sorting, however. Without additional software support, search and sort algorithms will not recognize differing usages of a sign as identical, since they are encoded differently. The solution here is presumably to offer users a switchable ‘profile’ according to which such algorithms can be informed to treat [gur_8] as identical to [ku_4] and [tu], for example.

A similar but rarer phenomenon known as splitting also occurs. In essence, a split is for all intents and purposes the reverse operation to a merger: in this case, a single sign, which may have been expressed visually in a number of ways, develops in such a way that the originally equivalent visual expressions come to assume separate functional significance. A clear case of splitting occurs in the sign TA, whose Middle Assyrian expression TA* comes to be used in NeoAssyrian royal inscriptions as a logogramme for the preposition /issu/, ‘from’. The proposed solution here is more or less a mirror image of the solution for splits, in that multiple characters must be provided throughout the system; however, the ‘new’ character simply does not occur except in texts from the period where a functional distinction exists.

6. Sort Order

In conclusion, a brief note on the question of sort order.

It has been the traditional practice of cuneiformists to arrange signs into lists organized by the order of wedges making up the signs, according to an agreed upon procedure. In general, the form of signs from the NeoAssyrian period has been chosen, due to the lesser variation in wedge orientation and the simplicity in ordering which this accommodates.

Two points should be made here with regards to this practice. First of all, given the high shrinkage from the earlier periods to the later, on the order of 30% or more, and given the fact that the encoding must accommodate the broadest number of characters in existence at any point within the scope of the encoding, producing a comprehensive list in terms of the traditional NeoAssyrian arrangement is not only impractical, but also does not serve the end of communicating the most information concerning the character code and otherwise. The general consensus at present seems to be that the UR III period, or at the latest the earliest portion of the Old Babylonian period are most likely to furnish a reasonably if not absolutely complete basis for organizing a comprehensive overview of the character repertoire. At the same time, given the structural nature of the signs at those periods, and the complexity involved in defining a simple order based upon wedge arrangement, some other criterion such as organization into semantic subsystems similar to that used by Rosengarten may prove preferable.

On the other hand, unlike previous computer character codes, a Unicode based character code does allow for multiple sort tables. This means that it will be possible to offer users several choices of sort order, tailored to particular needs, including the traditional one according to which most existing resources are organized. In this fashion, we expect that users of the new cuneiform encoding will be able to have their cake, and eat it too.

Cuneiform Encoding Organizational Chart

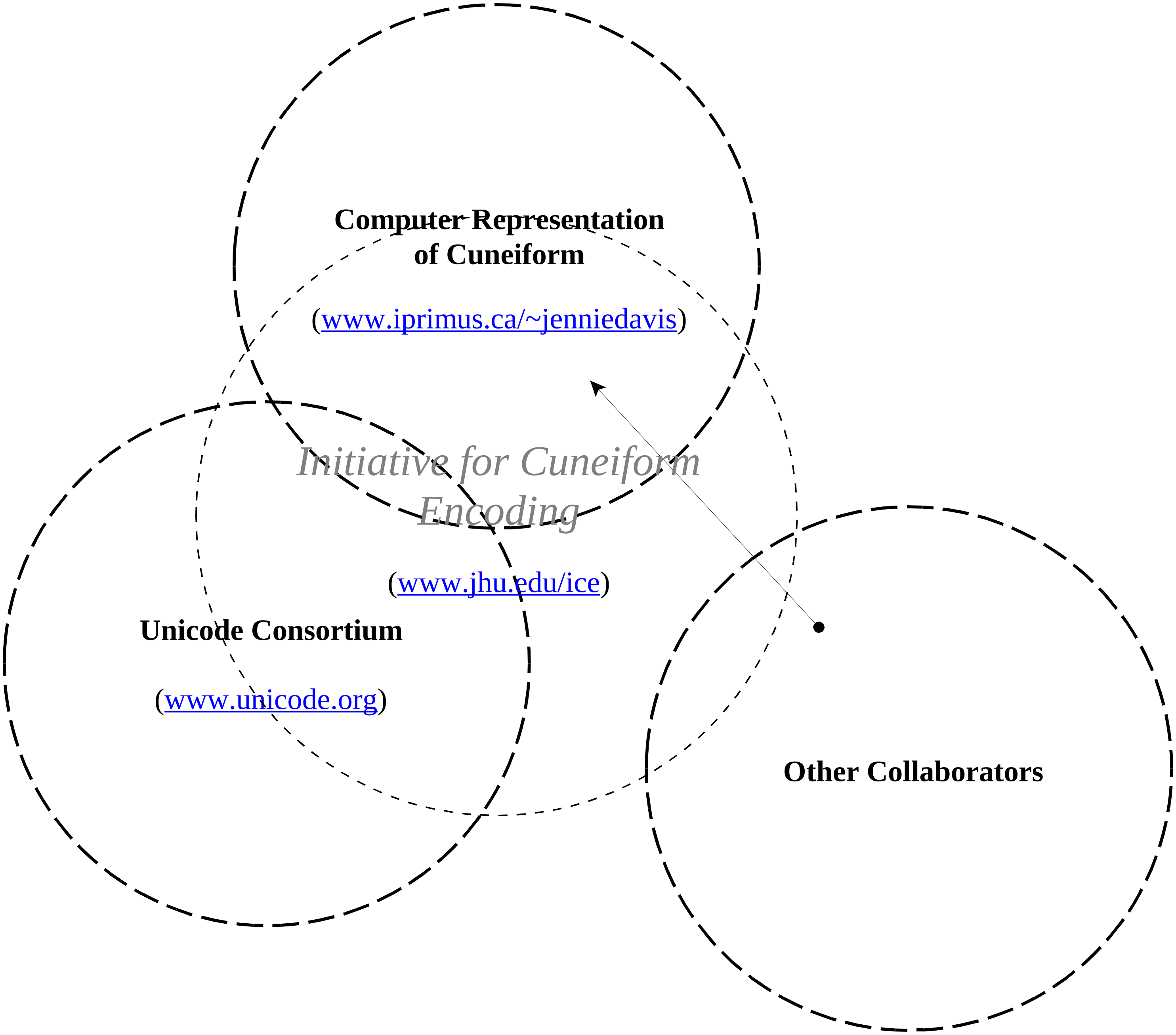
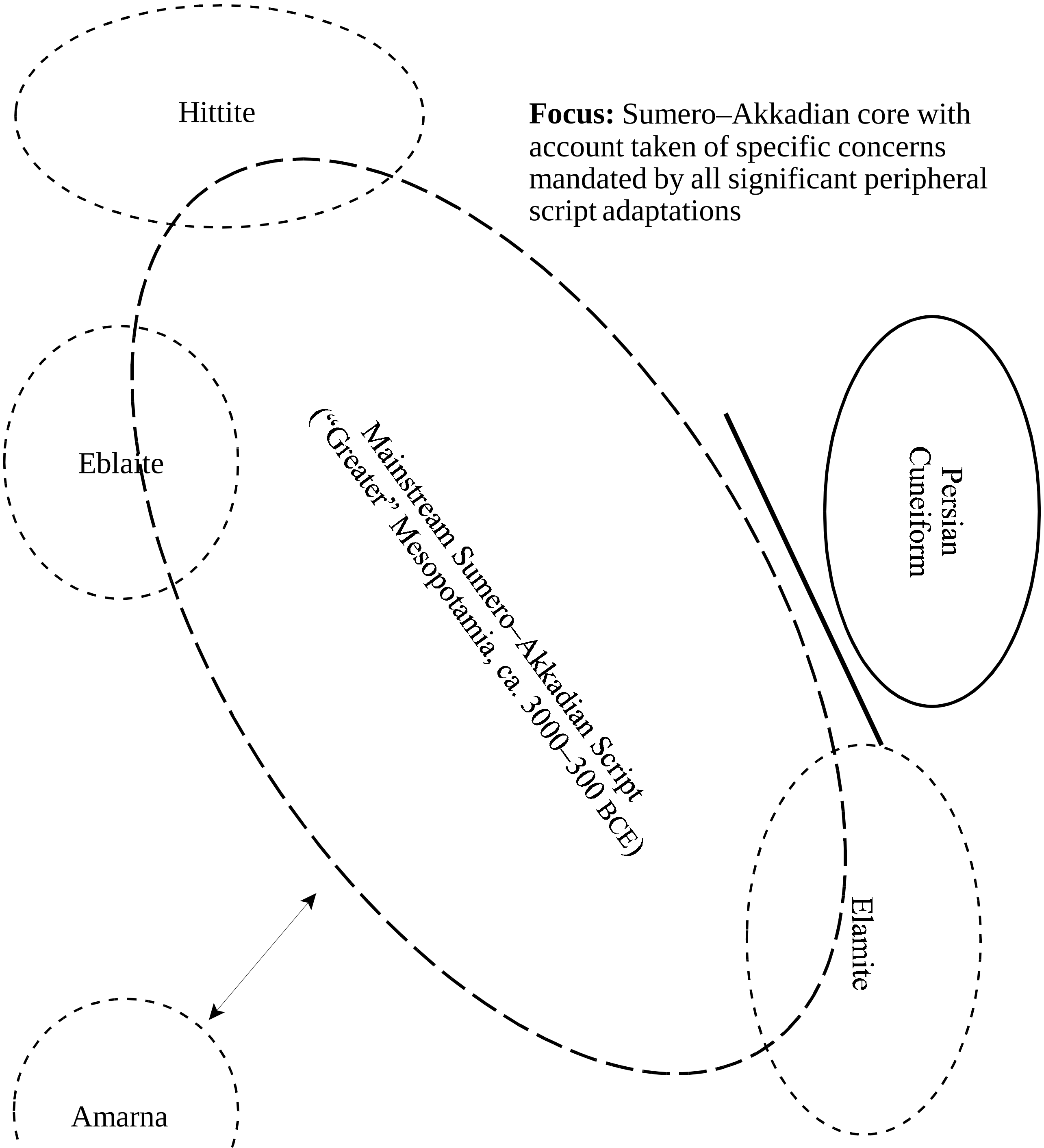

Computer Representation of Cuneiform
(www.iprimus.ca/~jenniedavis)
Initiative for Cuneiform Encoding
(www.jhu.edu/ice)
Unicode Consortium
(www.unicode.org)
Other Collaborators

Figure 1

Scope of Cuneiform Encoding


Hittite
Eblaite
Mainstream Sumero–Akkadian Script
(“Greater” Mesopotamia, ca. 3000–300 BCE)
Persian
Cuneiform
Elamite
Amarna

Focus: Sumero–Akkadian core with account taken of specific concerns mandated by all significant peripheral script adaptations

N.B.: diagramme illustrative of underlying encoding principle, and not intended as an exhaustive statement

Figure 2

Signs and Characters

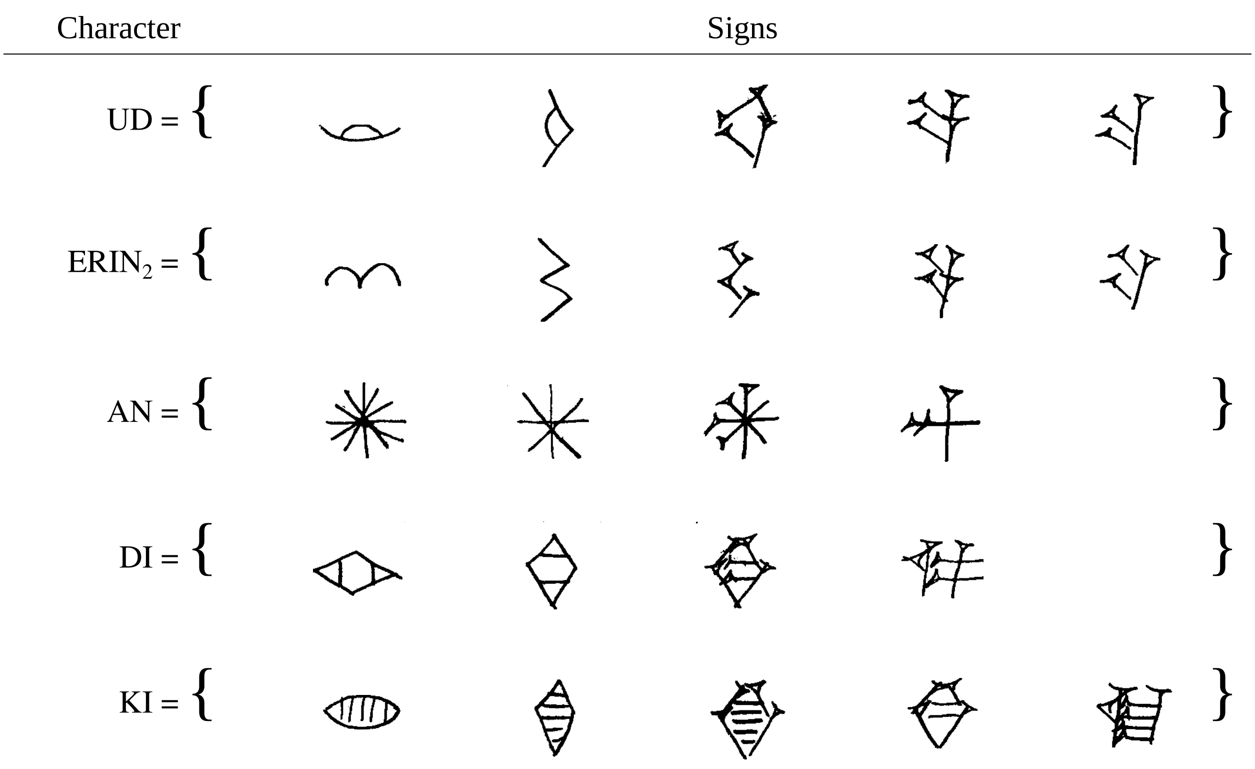
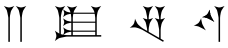
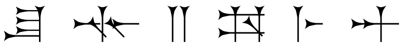
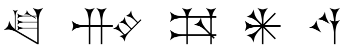
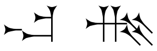


Figure 3

Original Document **(Form Alone: Signs)**	**Traditional Transliteration** **(Sign & Semantic Value)**	**Encoded Text** **(Identity: Characters)**
	2 gín kù-babbar	[1] [2] [3] [4]
	šu-ti-a *iš-me*–an	[5] [6] [7] [8] [9] [10]
	ki *gi-mil*–^{d}utu	[11] [12] [8] [10] [4]
	ba-zi	[13] [14]

Translation: (RE:) 2 *šiqil* silver— receipt taken by Išme–Ilī; it was disbursed by Gimil–Šamaš.

Figure 4

Sign/Character Divergences (I)

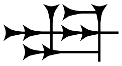
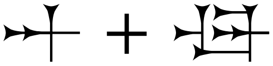
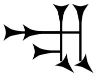
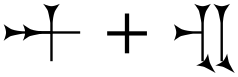
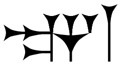
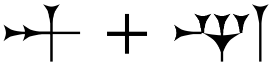
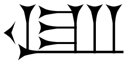
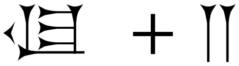
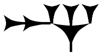
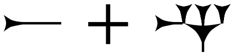
Ligated Sign	Decomposition Into Signs	Character Representation
		[AN] + [AG]
		[AN] + [EN]
		[AN] + [MÙŠ]
		[KI] + [MIN]
		[AŠ] + [ŠUR]

Figure 5

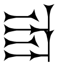
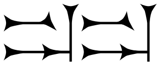
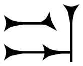
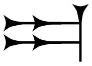
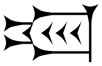
Text	‘Sign’-based Transliteration	Character Representation
	laḫ$_4$	[LAḪ$_4$]<1>
	laḫ$_5$ (=DU-DU)	[LAḪ$_4$]<2>
	laḫ$_6$ (=DU)	[LAḪ$_4$]<3>
	ab	[AB]
	unug	[AB]-gunû

Figure 6

Sign/Character Divergences (II)

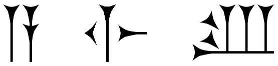
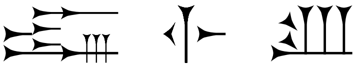
Sign Sequence	Character Representation
	[A] - [IGI] - [RU], *i.e. a-ši-ru* (‘inspector’)
	[IN] - [IGI]=[RU] *i.e. in-pà* (‘s/he swore’)

Figure 7

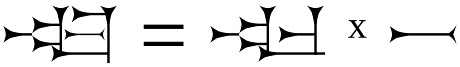
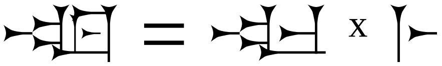
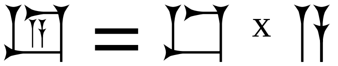
Composite and Components	Character Representations
	[BÀD] or [KA] x [BAD]
	[EME] or [KA] x [ME]
	[SUG] or [LAGAB] x [A]

Figure 8

Mergers and Splits

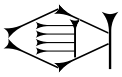
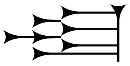
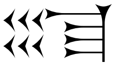
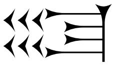
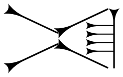
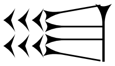
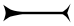
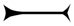
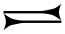
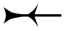
Original Signs	Character Representation	Merge As
	[gur$_8$]	
	[ku$_4$]	
	[ku$_4$]	
	[ku$_4$]	
	[tu]	
	[bad]	
	[idim]	
	[eše$_3$]	

Figure 9

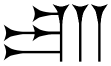
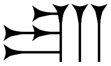
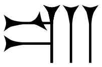
Original Sign	Splits Into	Character Representation
		[TA]
		[TA*]

Figure 10