Report on Cuneiform Encoding Project

by
Rick McGowan and Debbie Anderson

1. Background to the Script (drawn from http://std.dkuug.dk/jtc1/sc2/wg2/docs/n2786.pdf)

Cuneiform has a tradition spanning 3,000 years old, attested first from ca. 3500-3200 BC, and common by 2700 BC in Mesopotamia. The initial repertoire was composed of about 700 characters (or "signs" as they are typically called). In its earliest form the script was pictographic, though some items were strictly ideographic (i.e., 'sheep' is denoted as a circle with an enclosed "+" sign). In the next period, Uruk III (3200-2900 BC), logographic usage became more widespread, and individual signs were combined with more complex designs to express other concepts: a head with a bowl beside it was used to denote 'eat' or 'drink'. The early graphs were influenced by the writing medium (primarily clay), the available writing tools (reeds, which were used to make small, wedge-shaped impressions in the clay), and the need to quickly record information for the developing urban bureaucracy of the area.

Cuneiform spread throughout the Babylonia area to Assyria, eastern Syria, southern Anatolia (Turkey), Egypt, and Elam, and the script was used for a number of different languages, including Sumerian, Akkadian, Elamite, Hittite, and Hurrian.

2. Encoding Cuneiform (drawn from comments by Cale Johnson, UCLA, and Steve Tinney, University of Pennsylvania)

a. Not everything in cuneiform was encoded. The sign repertoires were driven by well-studied and stable corpora. The graphemes drew from two sets of stable corpora, and these were merged in the encoding (Old Babylonian materials collected by Miguel Civil, University of Chicago, and Ur III corpus at the Cuneiform Digital Library Initiative, UCLA, headed by Robert Englund).

b. In order to determine what a single grapheme was, the "team" working on the proposal relied on the internal structure of the writing system (rather than historical, comparative, or conventional considerations). Hence, if a sign was written inside another sign in the Old Babylonian/Ur III period, then it was treated as a single grapheme. However, if a sign was made up of several signs that were written as a string, each grapheme would be encoded separately. These "rules of analysis" were debated extensively, but the above rule-of-thumb was finally agreed upon. In the end, a few linear strings that were broken up were added back in as a single grapheme since some scholars absolutely had to have the single graphemes.

c. Early on in the project, nearly everyone agreed that the scholarly conventions that had been used were not especially logical. As a result, primary materials had to be re-analyzed. This meant that over 100 years' worth of guesses, mistakes, and suggestions were not incorporated in the encoding.

d. It was decided that encoding would be broken up into manageable chunks. Stage One will be based on Ur III through the late periods, and will include all major contemporaneous script users. Stage Two will involve Old Akkadian and Early Dynastic, and Stage Three will be Archaic

Cuneiform. At this point, the comprehensiveness of Old Akkadian and earlier material is not deemed appropriate for encoding, given the current state of paleographic research.

3. Participation of Academic Scholars and Unicode Representatives
The success of the encoding of cuneiform as due in large part to the direct involvement of scholars in the project (particularly Steve Tinney, University of Penn., *The Sumerian Dictionary of the University of Pennsylvania*, Karljürgen Feuerherm, and members of the Cuneiform Digital Library Initiative, or CDLI, at the University of California, Los Angeles). The active participation of Unicode representatives helped to establish the basic "ground-rules" for the encoding.

The effort to encode cuneiform was partly made possible by a Johns Hopkins University staff member, who applied for funding for the gatherings and organized two gatherings of cuneiform scholars and Unicode Technical Directors at Johns Hopkins University in Baltimore, Maryland.

The first meeting, held in 2000, had three representatives from Unicode (Rick McGowan, Ken Whistler, and John Jenkins) and eight cuneiform specialists, as well as a number of student observers. The second meeting took place in 2002 and had two Unicode representatives (Rick McGowan and Ken Whistler), one seasoned Unicode proposal author (Michael Everson), and fourteen cuneiform specialists, besides a number of student auditors.

After the second meeting, the main proposal authors (Everson, Tinney, and Feuerherm) were in frequent contact electronically in order to work out additional changes to the proposal, though there was one trip for Everson to meet with Tinney in person to work on the proposal.

The overall process took about six years from the first meeting at Johns Hopkins (2000) until the proposal's final approval in Unicode 5.0 (2006). The final repertoire is made up of 879 characters and 103 numbers.

4. Funding
The first meeting at Johns Hopkins was supported by funding from various administrative units and the library at the university. The second meeting had funds from the National Science Foundation, the Society for Biblical Literature, and the Script Encoding Initiative at UC Berkeley. The proposal itself was not paid for by any one source; the three authors volunteered their time to work on it. Travel for Everson to meetings was paid for through the Society for Biblical Literature and the Script Encoding Initiative at UC Berkeley.

URLs of Projects:
Cuneiform Digital Library: http://cdli.ucla.edu/
Pennsylvania Sumerian Dictionary: http://psd.museum.upenn.edu/epsd/index.html