

**"2nd Iranian Unicode Meeting"**

TITUS Project, Frankfurt/M., Germany,  
November 22-23, 2001 (cf.  
<http://titus.fkidg1.uni-frankfurt.de/unicode/iranian/tagung.htm>)

## Meeting Report

On Thursday and Friday November 22-23 the group of Iranianists who had already met on November 4, 2000 in Paris, again held a meeting in order to keep track of the work done so far with respect to the encoding of ancient Iranian scripts. Not all of the Paris participants were present in Frankfurt, but other specialists joined the group (see list of participants below).

The results of the Frankfurt meeting can be summarized as follows:

1) Bactrian: The participants agreed that for the proper encoding of Bactrian, only a small addition is necessary for Unicode block U+0370-03ff ("Greek and Coptic"), viz. for both the upper case and the lower case representatives of the symbol usually transcribed with a "thorn-" like letter. The Latin thorn letters should not be used for this in order for an adequate rendering to be possible. The encoding need is illustrated by the material provided by Nicholas Sims-Williams (cf. [http://www.gengo.l.u-tokyo.ac.jp/~hkum/images/bact\\_b09.jpg](http://www.gengo.l.u-tokyo.ac.jp/~hkum/images/bact_b09.jpg)).

2) Christian Sogdian written in Syriac script(s): With the exception of three characters (<f>, <zh>, <x>), all characters necessary for the encoding are present in Unicode block U+0700-074f (Syriac). For the three characters mentioned, additional code positions should be provided within this block; illustrative material is available from F.W.K. Mueller's and N.S.-W.'s editions. Additionally, there is a strong need, both for the encoding of Sogdian and of Syriac itself, of adding a code position for the diacritic siyame (the substitutional usage of U+0308, Latin diaeresis, is impossible because of the inherent directionality problems).

3) Manichaean script: The participants agreed that for Manichaean script, a mapping onto the Syriac block (U+0700-074f) should be envisaged as the best solution. As against the inventory provided for Syriac, there is a need of at least two additional character code positions, viz. for <zh>, which is also necessary for Christian Sogdian, and <dh>. As to the other peculiarities of the Manichaean script, Desmond Durkin-Meisterernst provided a thorough analysis, and the following possible solutions were discussed in detail: For the character representing gh (gamma), a mapping onto Syriac-Garshuni (U+0714) might be envisageable. Depending on the space available for an extension of the existing Syriac block, two solutions were discussed for the variants of characters that are marked by one-dot or two-dot diacritics (e.g., <k:~> vs. <k>, <x> vs. <k>, <q:~> vs. <q>): If there is no space to provide a code point for each of these variants, at least three diacritics should be added to the Syriac block, the two-dot diacritic

(not to be confused with Syriac siyame, cf. above) also serving to distinguish <ud> from <w> or <ush> from <sh>. For the distinction of other variants of letter forms such as the variants of <t>, the introduction of a "variant marker" would be a feasible solution. For the notation of a dotless <d> which might also represent <r> in unclear contexts, the code position U+0716 is usable. The variant marker could also be used to distinguish a subscript <y> appearing in some texts. There is no special encoding need for Manichaean ligatures (<cy>, <cn>, but also <dhdh>), the correct rendering of which would have to be provided by font engines. As against the existing punctuation marks of the Syriac block, there is at least a need for one mark consisting of a dot in a circle as well as of a line filler. Desmond Durkin-Meisterernst will provide a more detailed draft of an encoding proposal for the Manichaean script.

4) Sogdian script: For the Sogdian script (as well as the Uyghur script), two possible encoding strategies were discussed. While the soundest solution would consist in providing a code block of its own, a mapping onto the existing Unicode block of Mongolian (U+1800-18af) would be historically also adequate, given that the latter script developed from it. Whether or not such a mapping could be feasible depends on the question whether there are prescriptions as to the directionality (top-down) of the Mongolian characters. For Sogdian (as well as Uyghur), the directionality right-to-left is required instead. Further investigation is necessary as to this question.

5) Old Persian cuneiform script: The encoding proposal provided by Michael Everson was discussed with respect both to its completeness and to the designations of characters. As against this proposal, the addition of three signs representing abbreviations was underlined (AURAMAZDAA variant, AURAMAZDAHA genitive, DAHYU variant). The necessity of providing a code position for a hook-shaped variant of the SEPARATOR sign was denied. Further investigation is necessary as to whether there is a need for a code position of a <YA> variant and of a sign representing the number <3> in the shape of three wedges arranged vertically. Of the character designations used in Everson's proposal, three at least should be named differently, viz. <JA> and <JI> (instead of "GHA" and "GHI") and <ÇA> (instead of <THRA>).

Guenter Schweiger will provide an improved version of the encoding proposal.

6) Avestan and Pahlavi scripts: On the basis of Michael Everson's encoding proposal for Avestan, the participants reconsidered the question whether a unified encoding of Avestan and Pahlavi would be feasible. Given that the existing proposal for Avestan contains several signs that are also used in Pahlavi environments (esp. punctuation marks) or are even restricted to Pahlavi contexts (the numeric symbols), this solution was regarded as preferable, all the more since it would also account for the necessity of present day usage of Pahlavi script in the Parsee community which has been underlined by members of this community after the Paris meeting. As a result of this discussion, a proposal for a unified block covering both Avestan and Pahlavi will be worked out in the following

months. As against Everson's proposal for Avestan, the need of encoding ligatures (<SHCA>, <SHTA>, <SHSHA>) and elements of ligatures ("variant <TTA>") was denied, provided that the rendering of sequences of the given characters as ligatures can be determined by using a ligature marker. The need of encoding duplicate characters (<"noninitial <YA>" and "noninitial <VA>") was also denied. Further investigation into manuscripts is necessary to decide whether there is an encoding need for (seeming) glyph variants such as "variant <DHA>" or "variant <BA>". The encoding needs for Pahlavi were thoroughly discussed, esp. with respect to the question of encoding graphemic units rather than "surface" characters. On the basis of C.-M. Bunz's lecture on "Creating a Critical Edition in Digital Form", there is good evidence that such an encoding would be feasible, with the result that only a very small set of code positions would be necessary to cover all appearances of characters in Pahlavi words (cf. the similar approach provided by the authors of the article on Pahlavi in "Alphabete und Schriftzeichen des Abend- und Morgenlandes, 2., von Fachwissenschaftlern überarbeitete und erweiterte Auflage, Berlin, Bundesdruckerei, 1969 p. 52-55 [prepared with the collaboration of Erich Bruns, Georg Hincha, Johannes Meyer-Ingwersen, Gernot Windfuhr]").

7) Arabic script used for Iranian languages: The participants discussed two points concerning further encoding needs, viz. a) characters missing for the encoding of languages such as Balochi (a preliminary list was provided by Agnes Korn) and b) the necessity of encoding archigraphemic units. On the basis of detailed considerations reported by Thomas Milo and Gerd-Rüdiger Puin, the participants agreed that the encodability of archigraphemic units (esp. consonant characters without dotting) is necessary esp. for a scientific handling of older textual data (cf. <http://titus.fkidl.uni-frankfurt.de/armazi/gelati/dmsas/dmarabz.htm#dmarab4> for a specimen of an Arabic inscription of the 12th century).

A subsequent meeting of this group of Iranianists is planned for summer 2002 in Prague. The final results of the considerations will be published in a special volume dedicated to the question of encoding Iranian languages and scripts, issued by Enigma Corporation, Prague. We would like to thank Petr Vavrousek for this offer.

List of participants:

Carl-Martin Bunz, M.A., Research Assistant, Institute of Comparative Indo-European Linguistics and Indo-Iranian Studies, Dept. 3.5, University of Saarland, Germany; Unicode initiative of TITUS project

Dr. Desmond Durkin-Meisterernst, Turfanforschung, Berlin-Brandenburgische Akademie der Wissenschaften, Germany (formerly Manichaean Dictionary Project - Ancient India and Iran Trust, Cambridge, Great Britain)

Dr. Ralf Gehrke, Hochschulrechenzentrum der Johann Wolfgang Goethe-Universität Frankfurt am Main, Germany, and research project "Vorislamische alttürkische Texte: elektronisches Corpus", Department of Turcology at Frankfurt University

Prof. Dr. Jost Gippert, Professor of General Comparative Indo-European Linguistics, Leader of TITUS project, University of Frankfurt/Main, Germany

Thomas Milo, President, DecoType, Amsterdam

Claudius Naumann, Institut für Iranistik, Freie Universität Berlin, Germany

Dr. habil. Irina Nevskaya, research project "Vorislamische alttürkische Texte: elektronisches Corpus", Department of Turcology at Frankfurt University

Dr. Guenter Schweiger, Taimering, Germany (formerly Institute of Indo-European Linguistics, University of Rgeenburg, Germany)

Prof. Nicholas Sims-Williams, FBA, Professor of Iranian and Central Asian Studies; Old and Middle Iranian languages; Indo-European comparative philology; History of Central Asia; School of Oriental and African Studies (SOAS), London, Great Britain

Prof. Dr. Petr Vavrousek, Ústav starého Predního východu, Charles University, Prague

-----  
Carl-Martin Bunz, M.A.  
Institute of Comparative Indo-European  
Linguistics and Indo-Iranian Studies  
Dept. 3.5  
University of Saarland  
P.O. Box 15 11 50  
D-66041 Saarbruecken  
Germany

Tel. +49-(0)681-302-2744 /-2304  
Fax +49-(0)681-302-4234  
-----