Oriental TEX: A new direction in scholarly complex-script typesetting

Project by Idris Samawi Hamid

Preservation of much ancient scholarship of non-Western civilization exists only through unedited manuscripts. The manuscripts themselves are often not readily accessible, and the ability to make available accurate and culturally authentic typeset copies is restricted. Particularly for documents in Arabic script, typesetting is hampered by the lack of complete sets of vowel markings and diacritics, crucial for understanding the meaning of these texts.

For scholars working on critical editions of documents in Latin script, TEX is the tool of choice. Dr. Idris Samawi Hamid of the Colorado State University Department of Philosophy has received a grant to provide a comparable tool for Arabic scholars.

Dr. Hamid's own fields of specialty include Islamic philosophy, metaphysics, and cosmology. He has prepared critical editions of two major works of Arabic scholarship:

- *Shaykh Aḥmad Aḥsāʾīʼs* Observations in Wisdom: *Critical Edition, Translation, Notes, and Glossary*
- *The Mystical Theology of Muḥammad ʿAlī Shaḥʾābādī: A Critical Edition of* Rashahạtu ãl-Biḥar*, with Notes and Glossary*

Both works have been accepted for publication; what remains is to prepare proper typeset copy ready for printing. This is the impetus for the project and the proposal.

In February 2006, Dr. Hamid submitted a proposal to CSU's Integrated Research Projects Program, requesting a grant to develop and implement extensions to TEX to provide such software. The result of this work will be called Oriental TEX. The proposal was accepted by the program, and work began in May 2006. Taco Hoekwater is the principal programmer for the project.

Development includes a number of components:

- Extending the data structures of LuaTEX and Aleph to handle non-Latin languages and UTF-8 input files;
- Implementing two levels of right-to-left machinery: a global component for handling page elements, and a local component for switching direction in text without disrupting the typesetting process;
- Implementing dynamic ligaturing to accommodate the multiple, contextually-dependent shapes of Arabic letters, and the vertical shifting characteristic of particular written Arabic dialects; this will be based on concepts already present in Aleph;
- Creating control languages to handle conversion from the input stream to internal character representation, and to manage the context-driven glyph selection; again, many of the basic concepts are present in Aleph, but require adaptation to separate the two distinct processes;
- Developing OpenType font support, enabling use of fonts larger than 256 characters, and providing the mechanism to use such fonts;
- Adding extensions for critical editions, including a line-numbering engine and improved footnote handling;
- Quality control, involving extensive testing to assure that the goals of the development have been met;
- Documentation, a basic reference that is good enough for a skilled macro package programmer to learn how to take advantage of the special features provided by Oriental TEX; the goal is to provide user-interface macros that are easy to use, and suitable for typesetting of critical editions.

At TUG 2006, Taco presented a report on the status of the project. (This was particularly appropriate to the venue of the conference, in Marrakesh, Morocco, where other scholars and students reported on their own projects in Arabic typesetting.) As of early November 2006, the first stage was complete; this included support for full Unicode input, and merging of Aleph and pdfTEX.

The second stage, comprising support of OpenType fonts and PDF output from Aleph, is essentially complete as of this writing.

The third stage, expected by the time of BachoTEX, includes completion of hypenation patterns, character (case-mapping) tables, end-of-sentence discovery, handling of minimal word size, line justification, and the ligature table. Also in the third phase are the decoupling of characters and glyphs, with separate nodes for Unicode characters, fonts, and glyphs in fonts, as well as revamped paragraph breaking routines and a dynamic font interpolation engine.

The fourth and final phase of the TEX extensions, expected by the time of the TUG San Diego meeting, will see two-dimensional line typesetting of Arabic script and improved font handling, completion of the line-numbering engine, and improved footnote handling.

The generous support of the CSU Integrated Research Projects Program is gratefully acknowledged.

◇ Reported by Barbara Beeton