Abstract

《說文解字・電子版》
Shuo Wen Jie Zi — Dianzi Ban :
Digital Recension of the
Eastern Han Chinese Grammaticon

by

Richard Sterling Cook, Jr.

Doctor of Philosophy in Linguistics

University of California, Berkeley

Professor James A. Matisoff, Chair

Access to information for the study of historical Chinese language is impeded by the difficulties inherent in the ancient and modern Chinese scripts. The work presented in this dissertation seeks to address these issues within the framework of a digital image-based computerized system developed to identify and index distinctive features used in rendering modern script entities,[1] and to relate these features to those to be found in earlier script forms. With this mechanism, it becomes possible to give precise digital representation of the script entities and of the texts within which they occur. Representation of Chinese texts using a Chinese Character Description Language (CDL) enables precise intra-textual and inter-textual mappings, and greatly facilitates information exchange between old texts.

Digitization of the *Eastern Han Chinese Grammaticon*[2]《說文解字》*Shuo Wen Jie Zi* (SWJZ, c. 121 AD) of 許慎 *Xu Shen* has been accomplished using this system. Based upon digital images of the *Qing* Dynasty recension (1815) of 段玉裁 *Duan Yucai* (DYC), the new electronic text, 《說文解字・電子版》*Shuo Wen Jie Zi — Dianzi Ban* (SWJZ-DB), seeks to align itself with the general principles evident in DYC's work. The main principle (manifest in the domains of meaning, speech, and writing) is that only characters which are defined in the text may be employed in its definitions.

[1] *Cf*. Bishop (2003) <http://www.wenlin.com/>.

[2] The term *grammaticon* is here used to convey the Chinese distinction between 字典 *zidian* 'character dictionary', and 辭典 *cidian* 'lexicon, word (polysyllable) dictionary'. The name SWJZ might be translated as 'Interpreting the Simplex Signs, Analyzing the Complex Characters'. The purpose of SW was to catalogue, classify, and analyze all of the distinctive elements of the then nearly 1600 year-old Chinese writing system.

A concordance of the text has been produced,[3] using an indexing scheme designed for print publication.[4] In addition, a number of other corpora were digitized (in varying degrees) for the purpose of accessing material relevant to the interpretation of this text. By means of character-variant tables, Medieval Chinese readings[5] with precise references are given for all entries. Modern Chinese readings are provided for many characters, and Old Chinese readings are also included, where available in the source.[6] All characters have numerous external references,[7] brought into the corpus via the character-variant tables. These tables themselves, built using the CDL and contributing to the developing international Unicode Standard, provide permanent longterm access to all of this data.[8]

This work presents general theoretical issues framed in the context of specific application to the very difficult task of giving useful digital representation to important texts. General historical background is covered, with special attention to DYC's knowledge of the textual transmission.[9] This is followed by excursus on Chinese character classes and their componential analyses, traditional and modern. A quantitative method is outlined for establishing the distance among graphic entities, based upon specific editions of traditional texts and modern orthographic standards. Construction of intra- and inter-textual mappings is described with five specific case studies. The details of the present recension are examined, with the primary intent of documenting in detail all changes to the received text. Phonological study presented here looks at the representation of the pronunciation of Chinese characters in terms of a continuum of embeddedness. Tree structures are given for 925 Root Phonetic Classes, based upon the SW componential analyses. Statistical data, a Glossary of terminology, and numerous indexes are also included.[10]

[3] The SWJZ-DB text contains a total of 119,663 tokens of 10,706 types (108,957 gloss characters, of 6,713 types). Each type has one or more associated Unicode Scalar Values and stroke-based CDL descriptions.

[4] A 3-letter code appears under all gloss characters, permitting easy access to its entry within the text.

[5] After《新校互註・宋本廣韻》*Xin Jiao Huzhu — Song Ben Guang Yun* (余迺永 *Yu Nae-wing*, 2000).

[6] Karlgren (1957).

[7] *E.g.* 《説文詁林》*Shuo Wen Gu Lin* (丁福保 *Ding Fubao et al.*); 《漢語大字典》 *Hanyu Da Zidian* (許力以 *Xu Liyi et al.*)

[8] The Unicode Consortium <http://www.unicode.org/>.

[9] Translation of DYC's《汲古閣説文訂序》(1799) is given.

[10] Look for the data online at <http://stedt.berkeley.edu/>.