The Extreme of Typographic Complexity: Character Set Issues Relating to Computerization of The Eastern Han Chinese Lexicon《說文解字》 *Shuowenjiezi*

Richard S. COOK
STEDT Project, Linguistics Department
University of California, Berkeley
<rscook@socrates.berkeley.edu>
<http://stedt.berkeley.edu/>
2001/03/26/9:32

• Synopsis

The present study is concerned with character-set issues relating to computerization of one of the most important and most typographically complex Chinese texts,《說文解字》 *Shuowenjiezi* (SW). The title of the SW lexicon has been translated as 'Interpreting the Ancient Pictographs, Analyzing the Semantic-Phonetic Compounds' (COOK 1996). This Eastern Han Dynasty (121 A.D.) text was the first attempt at a systematic componential analysis of all of the characters in the complex Chinese writing system. With regard to this text, this paper addresses the following four topics, listed here, and briefly described below:

◊ 1.) The SW text — its history, character and importance.
◊ 2.) The character forms — their styles and components.
◊ 3.) The font — the character set and production process.
◊ 4.) Encoding Standards — mappings and missing characters.

This study begins with a brief introduction to the SW text, including its basic history, general characteristics, and overall importance to linguists, paleographers, epigraphers, and classicists. In particular, the linguistic importance of computerization of this text is emphasized.

The character forms found in the text are then discussed, with reference to both stylistic and componential issues. Special emphasis is given to the relationship between the text's componential analyses and the actual items of the character set. The issue of natural (extrapolated) extensions to the character set is mentioned.

Next, the 11,246 character font developed to capture this text is introduced. This is a CIDFont with Type 1 outlines. The rigors of the font production process are described, including hardware, software and indexing issues. Demonstration will be given of the typographic and lexicographic database systems employed in and resulting from the production process.

Finally and most prominently, encoding issues are addressed. Primary focus is given to mappings of the text-based character set to both Big-5 and Unicode standards. In this regard, mapping and missing character issues are discussed with illustrative examples.

The Extreme of Typographic Complexity: Character Set Issues Relating to Computerization of The Eastern Han Chinese Lexicon《說文解字》 *Shuowenjiezi*

複雜印刷之極致：東漢《說文解字》字庫之電腦化

〔美〕曲理察著
語言學部，漢藏同源辭典
柏克萊美國加州大學
<rscook@socrates.berkeley.edu>
<http://stedt.berkeley.edu/>

提要

本文的主題在於探討《說文解字》全文電腦化之程序。《說文》不僅是中國字書中的經典之作，也是中文印刷史上最複雜的書籍之一。這本東漢時期所完成的字書，是第一本對於複雜的漢字系統，做系統化結構分析的書籍。本文將依序討論下列重點：

◇ (1) 《說文》：歷史、特質及其重要性。
◇ (2) 字形：風格及其結構。
◇ (3) 字體程序：字庫及編寫過程。
◇ (4) 編碼標準：對應及未能對應之文字。

以上四點，可再簡單說明如下：

首先，本文對《說文》做簡短的介紹，包括歷史背景、特徵，及其對語言學家、經學家、古字學家、及碑文學家的重要性。特別著重於《說文》的電腦化，對語言學研究的重要性。

其次，本文從風格及結構討論《說文》中的字形。重點放在該書各種方塊字的結構分析和實際字庫之間的關聯，同時探討日後字庫之自然擴充。

接著本文將介紹以 Type 1 字型表示的一組字碼 (CID) 字體，對應到《說文》的一萬一仟二百四十六個字的編寫程序。我將詳細說明字體程序之製作過程，包含軟體、硬體、及索引編製等議題；並配合展示製作過程中用到的列印資料庫，及因此得出的詞典資料庫。

最後，本文將討論最重要的編碼議題。著重在如何將《說文》中的字庫對應到大五碼 (Big5) 及統一碼 (Unicode) 的標準。我將以實例說明對應及尚未對應的字例。

• CONTENTS

◊ 0.) INTRODUCTORY COMMENTS.

Consulting any of the great modern Chinese character dictionaries, be it the 清 Qing Dynasty《康熙字典》*Kang Xi Zidian,*[1] or the mainland Chinese《漢語大字典》*Hanyu Da Zidian*,[2] one will notice at the beginning of many entries citation of a text called《説文解字》*Shuowenjiezi* (SW). This is the most obvious testament of the primacy accorded the SW text in the Chinese lexicographic tradition.

When Chinese lexicographers embark upon discussion of the meaning of a Chinese character, the first order of business is to consult and quote SW. And it is not mere hyperbole to say that the SW text is the mother of all Chinese character dictionaries. This text has in fact spawned and nurtured countless hours of lexicographic work. It serves as the starting point, the basis for, and the framework of character analyses, and it is one of the touchstones against which all character analyses are judged.[3]

The text, carefully yet imperfectly transmitted to us through nearly 2,000 years of history, attained the status of Classic by virtue of the beauty and interest which readers through the centuries have found in its pages. By virtue of the great care with which it was constructed, its complexity, and as a result of the insightfulness and thoroughness of its analyses, this text is for modern and historical readers an "immortal masterpiece".[4]

But statements such as this do not begin to do the work descriptive justice. Only when one sits down and looks into its pages does one begin to appreciate the immense effort that went into its production. SW reflects not only the effort of its first author — nor is it the combined effort of his son and their family of assistants that we are appreciating — nor is it even the united effort of all the generations of publishers, readers and annotators who have perpetuated this text's existence.

Beyond all that, it is the fact that this text looks backward unflinchingly into the darkest depths of Chinese history, into all those past generations of human lives and into the recesses of spoken and written human communication, and then looks forward again, offering answers. The answers which it offers are answers to the simplest questions, fundamental questions such as "What is this a picture of?" And yet in attempting to address superficially easy questions, its author undertook the solution of many more ambitious questions. Even beyond looking at questions such as "What does this word mean?", which semanticists know to be complex questions indeed, SW looks deep into the soul of human existence. This is the attraction which SW has held for so many people, and this is the mystery which draws new readers to the text, generation after generation.

---

[1]Cf. ZHANG Yushu 長玉書 (1716).

[2]XU Liyi 許力以 (1993).

[3]Cf. for example《金文詁林》*Jinwen Gu Lin.* This expansive multi-volume treatment of bronze epigraphy uses SW as the framework for pulling together the etymological analyses of a great many scholars. For an extended example, see the bibliographical entry for 周法高 ZHOU Fagao (1981) in Cook (1996).

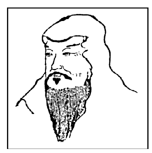
[4]Cf. e.g. the lavish praise given by ZHU Minshen, p. 7, typical of the attitude of many writers.

◊ 1.) THE SW TEXT — ITS HISTORY, CHARACTER AND IMPORTANCE.

1.1.) HISTORY AND CHARACTER

This first section of this study seeks to outline some of the fundamental issues surrounding the creation, transmission and significance of the different versions of the SW text, specifically as these issues relate to the choices made with regard to computerization.

The title of the《說文解字》*Shuo Wen Jie Zi* dictionary (which is often abbreviated as simply《說文》) may be translated roughly as *Interpreting the Ancient Pictographs, Analyzing the Semantic-Phonetic Complexes* (see the《說文解字・敍》，段玉裁 p. 754.1).[5] The SW's 東漢 Eastern Han Dynasty author 許慎 XU Shen, also called 叔重 Shuzhong,[6] lived from c.58 to c.147 A.D., to perhaps 89 years of age. He was a native of the city known in his day as 汝南召陵 Runanzhaoling, which today is called 河南郾城東 Yanchengdong, in Henan Province.[7]



許慎

XU Shen's work on SW is said to have begun in the reign of the Eastern Han Emperor 和帝 He Di, 和帝永元十二年 (100 A.D.). After two decades of work, XU Shen was approximately sixty-three years old and in ill health when in 安帝建光元年 (121 A.D.) he instructed his son 許沖 XU Chong to present the finished text to the Emperor 安帝 An Di.[8]

The amount of work involved in the composition of SW may be understood with a glance at some raw statistics. As SW itself records (in the original counts preserved in the traditional texts), the dictionary originally analyzed 9,353 正 main entries and 1,163 重 variant writings. That original text was reported to be in total 133,441 characters long (including both head entries and definitions).[9]

[5]This is the translation ventured in my 1996 monograph.

[6]This 字 *zi* "style" is a kind of nickname, often a literal interpretation of the formal name; 叔重 in this case is *Shuzhong* lit. perhaps 'uncle serious' (重 'serious' glossing 慎 'careful'), rather than *Shuchong* 'uncle variant'.

[7]郾城東 Yanchengdong (~33˚N~113˚E) is ~100 Km south of the capital of Henan, 鄭州 Zhengzhou.

[8]XU Shen therefore seems to have lived another 26 years after presentation of SW to the Emperor. For general discussions of the life and work of 許慎，see for example 陸宗達 (1981), or MILLER (1953,1977).

[9]By the time of the Song Dynasty these numbers had grown to 9,431 and 1,279, respectively, while the total for the entire text is reported to have somehow been reduced to 122,699 characters. The difference between these two couints is, perhaps not coincidentally, the approximate number of MV entries. This may indicate that MV characters are included in the larger count, but not in the smaller. For my counts for the various texts, see below.

No copy of the original SW text is known to exist in the modern day. Details of the textual transmission in the 800-year interval between Eastern Han and Tang times remain obscure, to say the least. It is clear however that the numerous citations of SW occurring in many texts dated to the second, third and sixth centuries indicate the broad and profound pre-Tang influence of SW. The earliest known SW texts are reported to be fragments of hand-copied Tang texts. These include portions of the 口部 'mouth radical' section (of which 16 entries remain) and 木部 'tree radical' section (~216 entries). The 'tree radical' fragments, for example, are said to be commonly recognized as remnants of an 820 A.D. Tang text.[10]

Extant "complete" versions of the text[11] trace their lineage to editions published in 宋 Song times, some 100 years after the Tang fragments, and more than 850 years after the time of XU Shen. Specifically, when modern people speak of *Shuo Wen* they are referring to versions of the text based upon the edition dated to 雍熙北宋太宗三年十一月， the eleventh month of the third year of the Northern Song Emperor Tai Zong, which is to say, roughly the end of 987 A.D.

This edition, produced in the school of 南唐徐鉉 the Southern Tang classicist XU Xuan (916-991), is sometimes referred to as《新校定說文解字》*Xin Jiaoding - Shuo Wen Jie Zi* . It is also known by several other titles, including the following:《說文解字・校定》*Shuo Wen Jie Zi - Jiaoding* ; 《說文解字・新附》*Shuo Wen Jie Zi - Xin Fu* 'SW with new appendices';《說文解字十五卷》*Shuo Wen Jie Zi - Shiwu Juan* 'SW in fifteen volumes'. The term 校定 *jiaoding* in the first two of these titles indicates the role of XU Xuan (and perhaps also the roles of his collaborators), relative to the 選 Xuan 'compiler' role of XU Shen. The term 校定 *jiaoding* (in modern Chinese written "校訂" with a different second character) tells us that they established the SW text with reference to the extant traditional texts, whatever those extant traditional texts may have been.

Modern editions of the XU Xuan text are available in versions based on two Qing Dynasty editions. The first of these was produced in 1809 by 孫星衍 SUN Xingyan. Sixty-four years later SUN's edition, criticized as being poorly formatted, served as the basis for the beautiful block-print edition produced by 陳昌治 CHEN Changzhi (1873).[12]

CHEN's version of the XU Xuan text is printed in vertical columns, and reads from right-to-left top-to-bottom. Each head-entry in the dictionary begins with the Small-Seal character (discussed in Section 2). The "primitive" (i.e. perceived original) XU Shen text (with its synonym and paronomastic[13] glosses, componential and phonological analyses, and occasional classical and other usage citations) is printed in larger type immediately below the Seal form. The commentary

---

[10]See ZHAO Pingan (1999[2000]:108*ff*), and the references therein, e.g. ZHOU Zumo (1966).

[11]Given the discontinuity of the tradition, and the variation among the extant texts, one cannot really speak of a "complete" XU Shen SW text, and hence the "shudder quotes".

[12]CHEN's edition is available today in reduced (縮印本) format, with two pages condensed into one, (328 pp., from the original of ~656 pp.), printed in Hong Kong under the title of《說文解字・附檢字》*Shuo Wen Jie Zi- Fu Jianzi* 'SW with appended indices'. See p. 328 for notes on the production of this CHEN's edition. Complete bibliographical information on this work is given in the References in Appendix 6 under 徐鉉 XU Xuan.

[13]Paronomastic definitions involve punning.

of XU Xuan's school appears next, in smaller type-face, briefly elucidating textual, semantic and phonological issues. At the end of each chapter (and placed slightly lower than the top line, so as to signify that they are additions) are appended those extra characters which XU Xuan's school determined *ought* to have been included in the original text.[14]

In addition to editions based upon that of XU Xuan, there is extant an important edition produced by XU Xuan's younger brother 徐鍇 XU Kai (920-974). Entitled《說文解字・繫傳》*Shuowen Jiezi - Ji Zhuan* 'SW with appended traditional commentary', this text is important for a number of reasons. First, it seems relatively conservative in some regards, and as such provides a means to judge the extent to which XU Xuan and his cohorts may have altered the received text. Secondly, it is clear that the two brothers XU produced their two versions of the text separately in order to address the differences in opinion which existed between them. This filial disagreement — perhaps even sibling rivalry — provides us with rich textual information on the state and sources of the received text on which they both labored. Furthermore, the XU Kai text, which is esteemed as the first 注本 'annotated edition' of SW, gives more extensive commentary than XU Xuan's, and is analytically superior to the XU Xuan text in many regards, though both texts have unique strengths.[15]

In the Qing Dynasty two versions of the SW were produced which constitute the principal modern annotated editions. The first of these is《說文解字・注》*Shuowen Jiezi - Zhu* (SWJZZ) 'SWJZ-Annotated', by 段玉裁 DUAN Yucai (1735-1815). The second is《說文解字・義證》*Shuowen Jiezi - Yi Zheng* 'SWJZ-Hermeneutics', by 桂馥 GUI Fu (1736-1805). These two works, produced on similar principles, copiously annotate a perceived "primitive" XU Shen text on the basis of the extant texts, with detailed references to and citations of the classical texts.

Of these two Qing editions, 段玉裁 DUAN Yucai's text with its extensive and extremely detailed commentary is often given priority today. DUAN was a master of the ancient classical traditions, "one of the most influential of the Qing phonologists"[16], simultaneously an innovative phonologist, semanticist and philologist. DUAN's textual reconstructions are squarely based upon his systematic phonological treatment of the script, and so his SW text may be said to have ushered in the modern era of historical Chinese phonology and semantics.

If limitations of DUAN's work are discussed, these will relate to details of his reconstructed phonological systems[17], or to the fact that he did not apparently make use of epigraphic evidence.[18] One might also say that his treatment of the SW text often seems heavy-handed if one

---

[14]Hence, the name "xin fu" given above.

[15]Please see COOK(1996, Section 06, p. 46 *ff*.) for discussion of the great value of the XU Kai text.

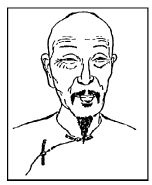
[16]These are the final words of the sketch given in BAXTER (160-2). Cf. also p. 139 in that same text.

[17]In comparison with systems of modern phonologists such as DONG Tonghe, LI Fang-kuei, BAXTER, or YU Nae-wing. Cp. the tables reproduced in YU Nae-wing's appendices (p.78*ff*) with BAXTER's Table 4.4 on p. 161.

[18]In this latter regard one might compare GUI Fu's use of Shang and Zhou epigraphic evidence, which was however very limited. If GUI's treatment of the received texts is more restrained than DUAN's, GUI does offer many valuable classical references and unique insights. For GUI Fu's text, please see the References.

does not have ready access to the sources upon which his conclusions were based. In numerous places in his sprawling textual commentary DUAN makes notes such as "This is wrong. Text A says B, text Y says Z. Now [my text] is right." He has thus altered the received text accordingly. In other cases he will modify the text in one place based upon his comments in another place, only tersely noting something such as "Wrong. Now right.", and it is left to the reader to track down the justification.[19] Or his conclusions made in one place may be inconsistently applied elsewhere, perhaps an indication of wavering opinion.

If there are imperfections in DUAN Yucai's work, such discussion must however be tempered with recognition of the scope of his astonishingly massive and complex project. From initial planning in 1776 to the first printing in 1815, work on SWJZZ occupied the last years of DUAN's life. Nearly 19 years were spent on the first draft alone, and another 13 years were given to revisions.[20] During this long span of time, he was performing an internal reconstruction of the text, endeavoring to remain true to the primary historical sources of the received tradition. This devotion to SW and command of the classical texts is marvelous by any standard.



段玉裁

A contemporary of DUAN Yucai, the classicist and historical phonologist 王念孫 WANG Niansun (1744-1832) wrote of DUAN's SWJZZ: “蓋千七百年來無此作矣。” "Nearly one thousand seven hundred years in the making, the world has never seen a work such as this!"[21]

In my youth when I first began to study classical Chinese, I was fortunate enough to have been given an edition of the《說文解字・注》*Shuowen Jiezi - Zhu* of 段玉裁 DUAN Yucai. This work had and continues to have a profound influence on me. In subsequent sections of the present study more will become apparent of the character of this edition of SW, and of the esteem in which this particular SW text is held.

1.2.) IMPORTANCE OF THE TEXT

SW's acknowledged status as "Classic" should not be taken to imply that any edition of the text is perfect. On the contrary, the various editions are clearly imperfect in certain regards, and generally regarded to be imperfect in various other respects. The more closely one compares the different texts the more apparent the imperfections become. In its transmission through the centuries the text has, in ways that may never even be known, suffered injuries that can never be repaired. In the relentless passage of time, changes have been wrought on XU Shen's original text

[19]Please see Section 3 ("The Early Planning Stages") below for further comments on DUAN's SW work.

[20]See the publisher's preface to the 上海 edition of 1989.

[21]Quoted from the foreword (p.1) [my trans.]; cf. the publisher's preface to the Shanghai (1991) edition.

as the result of interpretation, misunderstanding, and error. And even disregarding the ineluctable decay of time, the exquisite complexity of the subject matter alone makes error seem inevitable.[22] Commentators often disagree among themselves upon the meaning of passages of the text. And even where they agree on its meaning, modern etymological work demonstrates the text to be flawed in some cases in its assessment of the problems which it sought to treat.

Nevertheless, with these limitations firmly in mind, the work as we know it is undeniably important for many concrete reasons. It is important as the first surviving systematic treatment of the graphical and semantic nuances of the intricate Chinese script. It is important as the epitome (at the time of XU Shen) of some 1500 years of scribal traditions in China. And it is important as an encyclopedic if terse artifact crystallizing the received Eastern Han science and worldview. If one wishes to explore and utilize that science, one would do well to learn what XU Shen learned.

Depending on a reader's particular focus, it may be said that archeologists, astronomers, bibliophiles, biologists, botanists, chemists, classicists, epigraphers, historians, lexicographers, mathematicians, paleographers, phonologists, physicists, semanticists ... indeed, scientists of all ilks and specialties may all be interested in different aspects of the information which SW contains. And yet at heart all of these different perspectives share a common interest in the meaning of the text, and the meaning of the language which it analyzes.

It may be judged from things said in the Introductory Comments, and also from things just said about the work of DUAN Yucai, that when I speak of the importance of the SW text I am speaking from a linguistic perspective. It is in terms of interpretation of different aspects of Chinese words, be they words spoken today, written in books, words inscribed on bronze or earthenware vessels or on bone fragments, that I and many others[23] believe SW to have its greatest utility. And in so much as archeology is an unfolding science, we can not now fully appreciate what the future utility of this text may grow to be.

To one with acquaintance with the Chinese writing system, it must be clear that, if one would navigate this enormous mountain of complex data with any facility, computerization of this text is an important thing. And yet, given the complexities of the textual tradition just hinted at, it must also be clear that computerization of this text cannot be limited to the computerization of a single edition.[24] Likewise, the different versions of the text contain within them other complexities which further complicate the matter of computerization, to which we now turn our consideration.

---

[22]For further discussion of these issues, please refer to COOK (1996), a thorough introduction to the fiendish details of SW hermeneutics.

[23]See footnote 3.

[24]This issue is resumed in Section 3 below.

◊ 2.) THE CHARACTER FORMS — THEIR STYLES AND COMPONENTS.

2.1.) TYPOLOGY OF STYLES

Some aspects of the textual tradition and value of the text having been introduced above, we may now turn to demonstration and discussion of the character forms found in the text. The pioneering Swedish Sinologist Bernhard KARLGREN (1923) may here be quoted:

> The epoch making work of 許慎 Xu Shen is so much the more valuable as it was published only three centuries after 李斯 [LI Si (c. 280-c. 208), one of the scholars responsible for the 小篆 Small-Seal standardization of the Chinese script under 秦始皇 the Qin Emperor] and as therefore an unbroken tradition must have continued to the time of 許 Xu about the interpretation of most of the characters. [p.3 note 1; comments in square brackets are mine.]

These "小篆 Small-Seal" characters that Karlgren refers to here are those mentioned above in description of CHEN's version of the XU Xuan text. In the XU Xuan text, as in the annotated texts of DUAN Yucai and GUI Fu, 小篆 *Xiaozhuan* 'Small-Seal' (or simply "Seal" for short) characters appear in large-size at the top of each head entry.[25] The translation "Small-Seal" is an overly literal one: less literal translations might be "Lesser Seal" or even "Younger Seal", as 小 'small' in this case indicates relative stature, this stature being a factor of relative age. The comparison being made is between the 大篆 *Dazhuan* "Great-Seal" forms, said to date to Western Zhou times (~ 850 BC), and the Xiaozhuan characters appearing in Qin times (~200 BC).[26]

Thus, in general it may be said that 小篆 Xiaozhuan 'Small-Seal' characters are the fundamental unit of SW, standing at the head of every entry in the lexicon. In the present study these are sometimes simply referred to as "Main" (M) forms. In the current editions of the Song text, directly above the Main form stands a smaller-size Song Dynasty 楷書 *Kaishu* 'square script' version of this character, for ease of reference.[27] The SW definitions are likewise set in Kaishu.

Interspersed with the Kaishu of a definition one will occasionally encounter other forms, neither Kaishu nor Xiaozhuan. These other forms are collectively referred to as 重文 *Chongwen* 'additional variant characters'. In the present study these are sometimes simply referred to as "Variant" (V) forms. The 重文 Chongwen class of characters encompasses several different historical forms. Primary within the Chongwen class are 籒文 *Zhouwen* and 古文 *Guwen*. In addition, several Chongwen forms are associated with other historical styles occurring more rarely in the text. (The complete list of these may be sought in the computerized concordance.)

The 籒文 Zhouwen characters are sometimes said to be named after an historical person of Western Zhou times, 史留 (籒) Shi Zhou, their inventor. The variant characters in this style are equated by some with the 大篆 Dazhuan "Great-Seal" characters mentioned above. It appears however that XU Shen may have had access to a large collection of 籒文 Zhouwen / 大篆 Dazhuan characters at the time of SW's composition, and that these may have actually appeared as head entries in some cases, perhaps because their style was insufficiently distinct from the Xiaozhuan.

---

[25] A recent treatment of the Small-Seal script style is 趙平安 ZHAO Pingan (1999).

[26] See 祝敏申 ZHU Minshen (1999), for a good English introduction to issues surrounding typology of historical Chinese characters. The majority of this published doctoral thesis is in Chinese, but the English sections are well-written and contain good summaries of recent archeological findings.

[27] The two characters right here 楷書 are in this Kaishu style, which is the common modern style.

The term 古文 Guwen 'Ancient character' on the other hand constitutes not only a stylistic character class said to be representative of Confucian (Spring and Fall Period) writing; it also refers to characters of more indeterminate antiquity, which may in fact relate to vestiges of the earliest inscriptional writing such as that encountered on bronze vessels.

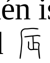
In addition to the Guwen and Zhouwen forms, a third important character type within the Chongwen is the 或體 *Huoti* 'Variant'. These forms, introduced in the text by the word 或 *huo*, are simply variant Small Seal writings of the head entry under which they occur.

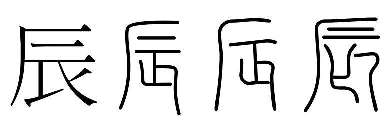
2.2.) COMPONENTS

Several other stylistic issues need be addressed, issues becoming apparent in the mapping work (Section 4) and with regard to treatment of the SW text's componential analyses.

As was said above, the SW analyses appear in the Song texts in Kaishu printed style. This however presents a number of difficulties for the typographer, who is therefore required to invent Kaishu forms to represent character components which may not otherwise occur independently. Such components may not in fact have separate head entries in SW, and may not have ever been productive, living independent lives in the script.[28] Such relatively bound graphical elements must however be catalogued by the typographer and the lexicographer.

Similarly, where SW posits a componential analysis of a character, it may then become possible for one to compose an etymologically "correct" character (according to a particular interpretation of SW). Such a character may not be otherwise attested anywhere in the history of the script. In the commentary text of DUAN Yucai, for example, DUAN's Kaishu is systematically written according to his etymologizations, sometimes component-based, and other times on the basis of whole forms. Such Kaishu forms are anachronistic and rarely if ever seen outside of DUAN's work, and would constitute examples of what I am referring to when I speak of "natural extrapolations".

An example of another type not specifically acknowledged by DUAN is here drawn from my 1996 study. In that monograph the SW's explanation of the character 辰 chén is demonstrated at length and in some detail. In short, the SW texts give two MV forms  and  of this character, and yet on the basis of SW's own componential analyses it is possible to reconstruct yet a third and otherwise unattested form * (as indicated by the preceding asterisk) synthesized from the identified components. DUAN did not actually ever write this form himself, as far as I can tell, and yet it follows logically from his reconstruction of the original text. The four forms below would therefore be typological variants of different classes, the fourth being an example of a member of the naturally extrapolated or reconstructed class.



[U+8fb0]=(GB b3bd)=(Big5 a8b0)=(Kangxi Radical 161)

For evaluation of this particular DUAN Yucai SW analysis, please see that 1996 study.

[28] The terms bound and free "grapheme" might also be used here; cf. "morpheme".

◊ 3.) THE FONT — THE CHARACTER SET AND PRODUCTION PROCESS.

3.1.) THE EARLY PLANNING STAGES

This leads us to consideration of the question of how to computerize such an intricate text which exists in minutely variant versions. One could, for example, begin with one of the Qing editions of the Song Dynasty XU Xuan text, going through the entire text, inputting character by character. Among the problems with this technique:

- 1.) Aside from stylistic considerations, no existing Chinese Kaishu encoding was adequate for such a task;
- 2.) There was certainly no suitable Xiaozhuan or Chongwen font;
- 3.) In the computerization process, tiny variations in the text will be encountered, and pose considerable difficulty to the typographer faced with the task of determining which variations constitute distinctive variations in the character forms.

The typographer would have to be a DUAN Yucai to sort it all out!

This realization in fact dictated my course of action. It was decided that, although DUAN Yucai had made significant changes to the Song text, the work which consumed the last decades of his life would serve as the best possible starting point for computerization of this text.

And the reasons for this are these: DUAN Yucai had, in the course of his years of work, sought to reconstruct what I have previously referred to as the "primitive" Eastern Han Dynasty text of XU Shen. In the process of so doing, DUAN put his received texts through a painstaking process of regularization. He weighed what a given text said in one place against what it said in other places. He weighed the different texts against each other. And he weighed these texts against the classical tradition. He corrected what he saw as inconsistencies, endeavoring to make the analytical system of SW entirely internally consistent.

Interspersed with his reconstructed text was DUAN Yucai's detailed commentary. This commentary documents his changes and the variation in the received editions. It contains semantic and phonological notes, and it contains classical references, both in support of his opinions and exemplifying difficult questions to which he had no answer.

Granted, DUAN Yucai may have made mistakes in places, just as XU Shen and XU Shen's sources no doubt did. But DUAN's presumption was that, throughout XU Shen's composition of the whole text, XU Shen was consistently aware of what he perceived to be correct and the truth, and that textual intrusions in the transmission had muddled the text's reflection of that awareness.

Scanning-in DUAN's entire text in graphical format would provide electronic access to the entire wealth of information in that text. Of course, in the early stages of the project, this access would be limited. But as work progressed, the image files would serve as the foundation for the eventual indexing of the entire work. This seemed to be a goal of unimaginable value.

The text I would decide to use as my basis would be the condensed 1989 Shanghai edition.

This version of 867 pages would greatly reduce the amount of scanning, as there are 3,468 plates bisected in the 1,734 pages of the original text. After computerization of this text, I might eventually proceed to computerization of other known texts, mapping them all to each other.

3.2.) COURSE OF ACTION

Having settled upon a base text, next the course of action needed to be determined. In 1994 personal computing still involved considerable expense for relatively limited power. In the three years during which my 1996 [1997] study was being produced, I explored the technological problems, and experimented with different scanners, scanning variously sized originals at various resolutions. Eventually, I determined the scanner and resolution which best met my needs, and purchased the hardware necessary for the production work. This hardware is described in the Hardware and Software Notes in Appendix 3.

An extended initial test of the hardware and software resulted in scans of the pages in the appendix of DUAN Yucai's text tabulating the 540 部首 Bushou 'Classifiers'. The appendix having been scanned-in at 400 ppi (optical resolution) and archived in graphical format (TIFF), the 540 character forms were then extracted and assigned a three digit file name according to the zero-padded Bushou number, 001-540. Each extracted character form would serve as a template for the hand-drawing of the character outlines in Fontographer.

Automatic generation of the character outlines based on such templates had yielded poor results, due to the reduced size, complex lines and indistinct boundaries of the originals. Although a set of larger originals was available for scanning (沙青岩 SHA Qingyan, 1980), the characters and text in that photolithographic SW edition had been rearranged in a quasi-Kangxi order, edited and augmented with Chongwen forms not actually in SW. That text was deemed to be unreliable and of secondary value in comparison with DUAN Yucai's text.[29]

Instead, the character outlines were produced by stroking a single line up the middle of the extracted template character. This line was then automatically optimized ("Clean-up Paths..."), and automatically expanded ("Expand Stroke...") to produce the uniform stroke width of 34 em units and regular line-end curvature. One result of this work was a set of three separate TrueType fonts collectively entitled "SWJZZXiaoZhuanBuShou". These fonts, which appeared as an index appended to my 1996 study, appear also in the present study as Appendix 1 (pp. 28-29). The other result of this work was that I now had a firm methodology.

3.3.) FONT AND INDEX PRODUCTION

With completion of that 1996 study, I then turned to the task of scanning in DUAN Yucai's text in its entirety. Each page of the bound text was positioned on the flatbed scanner, and scanned in black & white at 400 ppi (optical resolution). Each resulting full-page scan was assigned its three digit file name according to the zero-padded page number, and stored in TIFF format, with

[29]Oddly, in 2000 when I finally encountered another 2-byte font purporting to be of the entire SW, its outlines turned out to have been produced automatically on the basis of this 沙青岩 SHA Ch‘ing-yen (1980) text which I had earlier rejected. Produced at 北京師範大學 Beijing Normal University, this font contains errors which I have corrected in my copy. This font is mentioned below in Section 4.

each page occupying approximately 10MB of disk space. Scanning was accomplished over the course of nearly one full year, scanning at least a few pages most every day. The original text-scans were then archived in multiple redundant copies, both local and remote.

Having already demonstrated a feasible method in production of my Bushou fonts, I next set to employing this method for the entire text, and began the task of extracting all the Seal and Chongwen character forms. In this process each extracted form was assigned a unique index number of the type ABC.XYZ based on its position in the text, by 3-digit zero-padded page number (ABC), quadrant number (X), character number within the quadrant (Y), and whether the form was a main or variant (Z). Main entries were numbered with "0", incrementing this by 1 for each variant. Thus, the index number of every main entry ends in "0", while the index number of every variant entry is greater than or equal to one. This indexing scheme serves not only to locate each character in DUAN Yucai's text, but also numerically annotates the broad main/variant (MV) typological distinction. This aspect of the index proves useful in several regards discussed in Section 4 below.

Counts of the forms were made at intervals and stored in a database, all for the purpose of tracking the extraction process and safeguarding against error. My character counts in each chapter were compared with the counts given by DUAN and with the counts in the XU Xuan text. Notes were made of the counts and discrepancies.

When the forms had finally all been extracted, the counts were verified and reverified. The extracted forms were then imported into Fontographer, where they would serve as templates. The character outlines were produced as above by stroking a single line up the middle of the extracted template character. This line was then automatically expanded to produce the uniform stroke width of 34 em units in the 1000 em unit square, with regular line-end curvature. This expanded stroke was then hand-modified as the different Chongwen styles required.

The actual drawing of the outlines over the course of a year consumed roughly forty-nine eight-hour work days (roughly 400 hours), as it required a full day to produce one 221-character TrueType font. When all 49 of the separate TrueType subfonts were finished, the resulting database was verified and reverified over several months.

I now set about learning how to combine these subfonts into a single double-byte font. My initial inquiries had proven that this would be an expensive proposition with existing technology. In 1994, Altsys Corporation, producers of Fontographer at that time, offered to license the software to me for a mere $64,000. This was a kind offer, but rather beyond project budget at the time.

It was not until 1998 that I began beta-testing FontLab's new "MacComposer" software (as it was then called). Among my fellow beta-testers was Ken Lunde, whom I met on-line on July 7, 1998. It was Ken in turn who put me in direct contact with Alex Simagin, one of the programmers behind the "Composer" software project. Ken helped me to write my first CMap's, taught me the ins-and-outs of getting Composer to do what I wanted, and helped me to produce my first successful big CIDFont. This was a double-byte version of the 540 Bushou character set, the production of which was described above.[30]

[30] See Appendix 1.

3.4.) THE SWJZZ FONT

After this long process, it was not until January of 1999 that I generated my first double-byte versions of the complete character set.[31] This font contained 11,341 characters, and was named SWJZZ. My SWJZZ font is a Type 9 Font, which is to say, it is a CIDFont with Type 1 character descriptions, also known as a CIDFontType 0.[32] The version which I am currently using, generated on January 29, 2001, is an SFNT wrapped Apple Macintosh CIDFont, with embedded CMap.

The total of 11,341 characters in this font includes also the 7-bit roman characters. The total number of SW Seal and Variants (MV) is 11,246. Of these 11,246 characters, the last 540 are the Classifiers mentioned above which were produced in the initial testing. The 540 Classifiers were actually drawn twice, once according to the forms in DUAN's appendix, and again according to the forms in the main text. Minor variations between some of these character forms had been noted, and so it was determined that simply reusing the forms from the appendix would not be appropriate. The MV characters from the body of DUAN's text therefore total 10,706.

The first of these 10,706 characters occupies CID 1, with CMap position A140. The last of these occupies CID 10706, with CMap position E55D. The ordering follows that of the DUAN Yucai text. The Appended list of Classifiers comes next, in the CID range 10707-11246 (inclusive; CMap E55E-E8C4). The 7-bit roman characters, added as an afterthought, are found at CID's 11247-11341.

SWJZZ's CMap, entitled "RSCook-B5-H", encodes the characters in a Big-5 double-byte style, in chunks of 157 characters in each first byte plane. The second bytes are in two non-contiguous blocks, containing 63 and 94 characters respectively. Appendix 4 is an excerpt from my collapsed CMap, omitting the central thousands of CID's. Please note again that the ordering is that of the DUAN Yucai text. Mapping of this character set to existing standards is discussed in Section 4 below.

Successful PDF embedding of this font requires that the font not be activated by ATM. With this version of the font in the MacOS "System Folder:Font" directory, Acrobat Distiller 4.0 will embed the font in PDF documents. Note however that a copy of the custom CMap file must go into Distiller's private CMap folder: "Distiller:Data:psdisk:Resource:CMap:". In order to view such documents, Adobe Acrobat Reader version 4 or higher is required.

3.5.) THE SWJZZ DATABASES

Resulting from the font production process were detailed databases indexing the data in several ways. As each character had been initially assigned a unique index number based on its position in the DUAN Yucai text, it was now possible to link the SWJZZ font to the actual image files. This involved converting the original large TIFF images to the relatively compact JPG

[31]For more effusive acknowledgement of my gratitude to Ken Lunde, please see the Acknowledgements.

[32]"Technically speaking, a CID-keyed font is a Type 0 (composite) font with FMapType 9." (Lunde 1999:280,288)

format, for display purposes. Thus, the original 10MB TIFF's became 400K JPG's, highly legible in a web-browser.

In addition to the materials specific to my work on SWJZZ, several other data sources have come into my possession over the years. These include phonological data (Modern and historical, including HYPY and GSR[33] transcription data, Big5-based《廣韻》*Guangyun* indices), and graphemic/phonetic componential analyses in both Big5 and GB encodings.[34]

Most fortunately, I was provided with a Big5 version of the text《說文解字・附檢字》*Shuowen Jiezi - Fu Jianzi* (SWJZ-FJZ) already mentioned above. Big5 being wholly inadequate for the encoding of that text, the inputter did however input most everything that she could, with "##" and "**" inserted for missing characters.[35] This inputting work served as the basis for my work to map my character set to both Big5 and Unicode (this subject is resumed below in Section 4), and also provided yet another means to verify the counts of my original extraction work.

Finally, during a trip to Taiwan to 中央研院資訊所文獻處理實驗室 (Academia Sinica's Information Technology Institute) in the Fall of 2000 I came into possession of the SW font mentioned in a footnote above. Produced at 北京師範大學 (Beijing Normal University), this font's outlines were produced automatically on the basis of this 沙青岩 SHA Ch‘ing-yen (1980) text which I had earlier rejected. In addition to the fact that the outlines are of very rough quality, this font is rife with errors of various types (including, but not limited to, missing, misformed and misplaced characters), which have since been corrected in my copy. My revision of the BNU font does however provide yet a third means of checking my earlier work, and most importantly, it now provides me with forms, however imperfect, for the entire SWJZ-FJZ text, including those excluded by DUAN. Due to the limitations of the text upon which it is based (e.g. several misdrawn and numerous altered forms), this BNU font cannot however be used as anything more than a stop-gap until such a time as a more perfect SWJZ-FJZ font is produced.

The combination of these various resources makes for a potent relational SW database computing environment, which really must be seen to be appreciated. Characters may be found by numerous criteria, including *pronunciation* (ancient and modern), *meaning*, *stroke count*, *character component*, *encoding value*. A concordance of the SW texts, mapping the head-entries and glosses to one another has also been built.[36] But before the relational system came into being, a considerable amount of effort went into mapping, and it is to this subject that we must now turn.

---

[33]See the list of ABBREVIATIONS in Appendix 5.

[34]This latter data was provided to me by Ross PATERSON <rap@doc.ic.ac.uk>.

[35]I am grateful to Mr. Richard SEARS for granting me permission to use this Big5 text. It was SEARS who contracted the inputter for this work, a person named Ms. Ann WU. Ms. WU reportedly was engaged full-time for four months on this typing project.

[36]And adding such wonderful commercial software as Tom Bishop's Wenlin to the mix simply makes many things all that easier. Please see the acknowledgements.

◊ 4.) ENCODING STANDARDS — MAPPINGS AND MISSING CHARACTERS.

4.1.) BIG5 ENCODING AND A BIG5 SW TEXT

Primary focus in this Section is given to mappings of the text-based character set to both Big-5 and Unicode standards. Mapping, typological and missing character issues are discussed with illustrative examples. As this aspect of the work is more of a work in progress than other aspects of the computerization, the present discussion should be understood to encompass not only work which has already been completed, but the plan for future work as well.

As mentioned above, the work of mapping my character set to Big5 began on the basis of a Big5 XU Xuan text input by Ms. Ann WU and provided to me by Mr. Richard SEARS (WU/SEARS text). Despite the many characters missing from Big5 encoding, this text has proven to be a great help in accomplishing the mapping. Furthermore, as a result of this mapping, I now have built limited concordances (in relational database form) of both the DUAN and the XU Xuan texts. Below are tabulated the top 120 items from the frequency (descending) counts on the raw Big5 text:

Freq:	**Zi**	**Freq:**	**Zi**	**Freq:**	**Zi**	**Freq:**	**Zi**
12052	從	382	手	247	子	165	東
9364	也	377	臣	244	名	164	入
8694	聲	367	口	235	竹	162	南
6172	**	357	等	234	魚	159	無
2795	##	355	馬	233	書	158	玉
2560	曰	352	女	232	秋	158	見
1911	之	350	鉉	230	作	156	與
999	一	345	所	228	春	156	同
958	文	344	心	227	相	155	生
947	水	340	亦	227	字	155	此
821	若	331	形	226	籀	153	長
797	讀	326	出	226	邑	153	禮
780	木	291	而	222	傳	153	故
744	人	288	謂	222	王	147	高
728	艸	282	金	220	足	146	易
720	古	282	衣	219	今	146	白
673	屬	280	日	213	在	144	牛
647	省	276	目	212	火	143	如
624	或	276	糸	205	石	142	气
611	皆	275	中	204	小	141	俗
559	凡	274	山	196	二	140	方
551	以	274	車	195	非	137	說
501	為	273	周	189	貌	137	三
492	有	269	其	188	虫	135	走
488	象	262	鳥	186	地	134	十
475	者	257	肉	185	犬	134	器
451	詩	253	上	181	食	132	可
413	言	252	行	177	是	130	陽
408	不	248	土	174	禾	130	天
404	大	247	下	167	門	127	物

The WU/SEARS Big5 text had 11,106 lines.[37] In the first stages of work I verified and corrected the structure of the Big5 code, single-spaced the Big5 characters (for concordancing purposes), generated frequency statistics (all in Perl), and then imported the entire text into a FileMaker Pro 4 (Chinese) database. I then added serial numbers to the entries. Below I extract 21 lines from this initial work on the raw Big5 text, missing-character placeholders and all:

SN	**MV**	**Gloss**
1658	許	聽 也 從 言 午 聲
1659	諾	** 也 從 言 若 聲
1660	##	以 言 對 也 從 言 ** 聲
1661	讎	猶 ** 也 從 言 雔 聲
1662	諸	辯 也 從 言 者 聲
1663	詩	志 也 從 言 寺 聲
1664	詩	古 文 詩 省
1665	讖	驗 也 從 言 ** 聲
1666	諷	誦 也 從 言 風 聲
1667	誦	諷 也 從 言 甬 聲
1668	讀	誦 書 也 從 言 賣 聲
1669	##	快 也 從 言 從 中
1670	訓	說 教 也 從 言 川 聲
1671	誨	曉 教 也 從 言 每 聲
1672	譔	專 教 也 從 言 巽 聲
1673	譬	諭 也 從 言 辟 聲
1674	謜	徐 語 也 從 言 原 聲 孟 子 日 故 謜 謜 而 來
1675	##	早 知 也 從 言 央 聲
1676	諭	告 也 從 言 俞 聲
1677	詖	辯 論 也 古 文 以 為 頗 字 從 言 皮 聲
1678	諄	告 曉 之 熟 也 從 言 ** 聲 讀 若 庵

The field name (column head) abbreviations used above are as follows: "SN" is 'Serial Number'; "MV" is 'Main or Variant'; and "Gloss" is the 'definition'. I sometimes refer to a "Main" form (正文) as "head entry" or "main entry", and the "Variant" form (重文) may occasionally be termed the "sub-entry" of a given main entry.[38] An "entry" is therefore either Main or Variant; the main entry is mandatory, and there are zero or more sub-entries. The main entry is that form (for the most part perceived as an actual Xiaozhuan form) which accounts for the position of the character within the Bushou classificational system of the dictionary. As mentioned previously, a sub-entry is largely one of two types of Chongwen (Guwen and Zhouti), subordinate to the main entry, deriving its position in the dictionary from that of the main entry immediately preceding it. Within the sub-entry class the ordering and in some cases the Chongwen class determination have been regularized in DUAN's text.

It is evident from the above extract that there are lacunae represented by both "##" (in MV) and "**" (in Gloss). This may serve to illustrate the great inadequacy of the Big5 encoding's

[37]It should in fact have contained 11,108 lines, as mapping work later revealed.

[38]See Section 3 "Font and Index Production". See also, Section 2 "Typology Styles".

~13,000 characters[39] for this kind of text. In fact, out of a total of 128,823 double-byte character strings, 2,795 of these were "##", and 6,172 were "**".

Thus, it can be seen that the native typist had been unable (or unwilling) to find Big5 equivalents for 2,795 of the MV entries, and in typing the Gloss data, Big5 had failed in 6,172 instances. Putting this another way, in typing 128,823 characters, she had found only 119,862 of those in Big5, which is to say that a total of 8,961 were unfound. The unfound percentage relative to the total text size is thus (8967/128823*100=) 6.96%.

Frequency statistics at this point may serve to illustrate the degree of unsuitability of the Big5 character set for this task, despite the typographical errors which the typist made (exemplified and discussed briefly below). As it turns out from the frequency statistics (excerpted above), the raw text of 128,823 double-byte characters contained a total of 7,531 different double-byte strings (including the two "##" and "**" missing-character placeholders). In terms of my reference Big5 font (ALSL) of 13,053 characters, this represents in this single text a usage of (7531-2/13053*100=) 57.7% of the entire Big5 character set. From this one can see that in the large Big5 character set, which is usually adequate for most modern standard Chinese uses, 42.3% or ~5521 characters would have come into the script in the ~1,000 years since the Song dynasty.

In the above extract SN's 1663 and 1664 are in boldface to call attention to repetition of the character "詩" in both those MV entries. These forms should in fact be Xiaozhuan and Chongwen (i.e. Main and Variant), respectively. In the Big5 text, however, out of sheer necessity, a given group of Main and Variant entries has in most cases been assigned the same Big5 character code. Of course, this Main/Variant failure is a typological one, and (aside from the broad two-fold 5401/7652 distinction Big5 makes on the basis of character rareness) Big5 is not a typological encoding. It may be said therefore that the above statistics do not in fact represent all of the typist's problems with Big5 at all.

At this point an analogy may perhaps serve to clarify the problem. Imagine, if you will, the frustration which an English typist might feel if, in typing in a manuscript, for example the *OED*,[40] certain of the English alphabetic characters required for the job were simply not available on any keyboard. Based on the above statistics we cannot calculate for this analogy exactly how many of the letters might be missing. Without exact knowledge of the values lumped together with placeholders this cannot be said. These statistics will come later. If we judge only by the number of main entries for which there are substituted placeholders, this turns out to be a significant part of the character set.

As it turns out, the number of main entry "##" placeholders in my corrected version of that Big5 text is 2,527. Relative to an ALSL font augmented with the missing characters, it may be said without regard to variants that this character set is (2527/(13053+2527)*100=) 16.2% deficient. If the 26 lower-case letters of the English alphabet were so deficient, we might have keyboards

[39]My Big5 reference font is Apple LiSung Light (ALSL), with exactly 13,053 Chinese characters. The raw WU/SEARS text's Big5 font (unknown) betrayed very very minor differences, which have been annotated.

[40]*Oxford English Dictionary.*

without the five (relatively low-frequency) letters k, j, z, x and q.[41] And if we may imperfectly liken English case to SW variants, the situation is quite worse. Nearly none of the SW variant forms have Big5 equivalents, and so by this analogy we could exclude all English upper-case characters as well. Our hypothetical typist would certainly produce an unusually strange looking electronic *OED*. In fact, the abbreviated name of the *OED* itself might come out as only "***"!

The next steps in this mapping adventure involved passes through the entire 11,108 lines[42] of text, comparing and seeking to match the Big5 code to the items of my SWJZZ character set. At present, three such passes through the data have been made. In the course of this work the issues involved grew clearer as my notes in the relational databases grew. This initial work included the following corrections to the WU/SEARS text, made with reference to the printed work from which the inputting was done. Some of the major issues addressed are as follows:

1.) addition of two missing entries and glosses;
1.) correction of entry ordering lapses;
2.) correction of MV typographic errors;
3.) correction of Gloss typographic errors;
4.) correction of mistaken uses of "##" placeholder;
5.) correction of mistaken uses of "**" placeholder.

Among the typographical errors committed by the typist, the following are among the most interesting. These all involve confusion of one form with another rather similar form. The form on the left below is the erroneous form, followed by the corrected form on the right:

[U+985e]	類	纇	[U+7e87]
[U+6bcd]	母	毋	[U+6bcb]
[U+5e02]	市	巿	[U+5dff]
[U+672b]	末	未	[U+672a]
[U+6cab]	沫	沬	[U+6cac]
[U+80c4]	胄	胃	[U+80c3]
[U+6c69]	汩	汨	[U+6c68]
[U+63a5]	接	椄	[U+6904]
[U+68fc]	棼	棽	[U+68fd]
[U+8411]	萑	雈	[U+96c8]
[U+6f8c]	澌	凘	[U+51d8]
[U+7c5a]	籚	蘆	[U+8606]
[U+8d0f]	贏	羸	[U+7fb8]
[U+8e94]	躔	纏	[U+7e8f]

These errors and others like them may seem maddeningly minuscule, and incredibly difficult to identify in a body of 128,823 characters. But this is the level of detail required for clean text production.

[41]These letter frequencies are perled from an e-version of Shakespeare's comedies, histories, and tragedies.

[42]This is the corrected total for this edition of the XU Xuan text.

4.2) TEXTUAL AND CHARACTER MAPPINGS

With regard to the mapping, it should be clearly stated that this method in fact involves two separate mappings:

- 1.) mapping of the DUAN Yucai text to the XU Xuan text;
- 2.) mapping of the Main and Variant forms (MV) to Big5.

In the three pass-throughs mentioned previously, both of these mappings were being performed simultaneously.

Between the Qing DUAN Yucai and Song XU Xuan SW texts (neglecting for the moment the XU Kai text), there are several types of incongruities.

Most broadly, the Song Dynasty text of XU Xuan contains 11,108 entries, while that of the Qing text of DUAN contains exactly 10,706 (9,428 main, 1,278 var.). The DUAN character counts in this discussion are based on the indexing performed in production of my SWJZZ font of the DUAN text, accomplished in work spanning several years (see Section 03). As far as can be determined after the numerous checks and rechecks performed during these years, the figures "10706 (9428 main, 1278 var.)" are *exact* for the DUAN text. The total for the XU Xuan text is based upon the initial comparison of my counts for DUAN's text with his own and with the XU Xuan counts. All these counts were subsequently checked against the corrected Big5 text, and underwent further verification in correction of the BNU font. A final level of verification of the XU Xuan MV counts will occur when a font for that text is finally produced.

The difference between the DUAN and XU Xuan texts being (11108-10706=) 402 entries, these 402 character entries may be put in two groups.

The first group is that which contains by far the majority of the 402, and these are characters which DUAN deemed clear Song intrusions (appendages) to a Han Dynasty "XU Shen original". Such characters are found appended to the end of each "chapter" (bushou section) of the Song SW, arranged by XU Xuan & co. in a manner intended to show that they are additions to the received text. This is one case in which the Song revisers clearly sought to preserve the received text.

The second group (among these 402) includes those characters not clearly appended (largely Chongwen character variants, but also including main entries), but which DUAN deemed textual intrusions, and so excluded on the basis of that determination.

Thus, there are 402 entries in XU Xuan's text which have no match in DUAN's text. And yet, when attempting to match the actual *character set* of DUAN's entries with the character set of XU Xuan's text, another kind of incongruity presents itself, and this is with regard to the actual *character forms* in both those texts.

Parallel to his revision of the main entry glosses (the short definitions following each entry) DUAN also effected an assessment and revision of the character forms, both of the Xiaozhuan and of the Chongwen. In some cases he applied a standardization of the form found in one part of the text to forms found elsewhere (usually on phonological grounds). In other instances he changed the form slightly to reflect his total understanding of the meaning, based on no forms in SW textual evidence, though usually based to some extent on the SW glosses. In other cases his evaluation of Chongwen was simply a reassessment of which form was Guwen and which was Zhouti, a reordering and relabeling of the forms.

Disregarding "superficial" graphical differences, such as those which might be seen as simply scribal peculiarities (differences in handwriting, writing implement, or printing technique), we may distinguish true graphical variants. *True graphical variants* (TV) are parallel forms which occur in *both* texts, and which differ in the *number* of strokes, the character *components*, or the *arrangement* of character components.[43] True graphical variants may usually be viewed as "parallel" as determined by the gloss and context, though there are rare cases in which Duan reverses the glosses on adjacent characters, and also slightly alters the writing of one character.[44]

Now, this matter of "true graphical variants" is often complicated by the systematicity of Duan's changes. In fact, Duan's revision to the writing of character x usually ripples through to revisions of characters in which x is a component. This is most clearly a problem where x is a radical (bushou), as in the case of 白 bai2 'white' [363.41]. The change to this character (based on the textual gloss analysis) is propagated through the whole text. A set of forms for the Xu Xuan text must therefore be employed, although the differences are systematic, and hence predictable.[45]

At a first level in this project, I chose to treat such systematic variations (SV) as negligible, focusing instead on those variations which reflect unique (non-systematic) peculiarities in the graphical forms. Thus, a configuration of lines in Duan's forms which regularly corresponds to a configuration of lines in the Song text is not viewed as a "true graphical variant".

My corrected version of the BNU typeface is a step towards production of a version of the Xu Xuan text similar to that produced for the Duan text. The indexing of SWJZ-FJZ has been completed, as has the scanning of that text into high-resolution color graphical image format. Although it is my hope to eventually produce complete parallel electronic versions of the major editions of the Duan, Xu Xuan and Xu Kai texts, I am focusing at present on creating complete parallel versions of the Duan and Xu Xuan (SWJZ-FJZ) texts. These include Duan's perceived "Xu Shen Original", and will eventually contain a version of that text set entirely in the CID Type 1

---

[43]Cf. the section on "Ideograph Features" in the *Unicode Standard* 3.0, p. 265.

[44]Note that stroke counting may be applied to Xiaozhuan and Chongwen forms just as it may be applied to Kaishu (Song square script) forms. Stroke counts for "equivalent" Kaishu and Xiaozhuan/Chongwen forms may not, of course, correspond, and yet a systematic method of counting strokes of Xiaozhuan/Chongwen characters may be employed consistently, for classificational purposes, just as such a method is employed with Kaishu. The method developed in my work on SW is transparent, in contrast to Kaishu systems in which "stroke" does not necessarily equal "line segment". A stroke in Kaishu counts is often a calligraphic unit, equivalent to "line formed with a single, largely uninterrupted hand motion", "flourish". Thus, a "square" character like 囗 wéi 'surround' has four sides, and yet three strokes, since the upper and right lines comprise a single stroke. In terms of the kinds of lines encountered in Xiaozhuan/Chongwen characters, this is never a problem, since a square shape, were it found to occur, would be said to have four lines. *Lines* in Xiaozhuan/Chongwen forms (I will not refer to them as strokes) have very little ambiguity: each has a clear beginning and a clear end, and where it does not, as in the case of the circle, it is said to have just a single line. The system of counting lines employed in this electronic SW will be transparent to the reader, as long as confusion with Kaishu stroke counting methods is avoided.

[45]With regard to systematic and hence predictable variations between two characters, as with those that occur between many Simplified and Traditional characters, it would be sensible to specify a codepoint which signifies "perform regular transformation x on character y". By this method a large portion of the simplified Han character set could be treated as stylistic variation rather than encoded separately. But this point may be lost in the midst of the mammoth work which has been done attempting to encode the entire open-ended Chinese character set.

typefaces developed to represent all of DUAN's forms. As mentioned, each entry in these databases is linked to high resolution graphical images of the complete DUAN and XU Xuan texts.

This subsection may be concluded with summary of the considerations being made with regard to three different directional matchings, A, B and C. These notes are extracted from notes made during the first three mapping passes, and if nothing else may provide the reader with a feel for some of the considerations involved in that process. [Xuan here = XU Xuan.]

§ A.) Considering matches in the direction of **DUAN->Xuan**, there are cases in which DUAN's character form:

- 1.) and gloss are identical to Xuan;
- 2.) is identical, and yet the gloss is either slightly or greatly different from Xuan;
- 3.) is slightly or greatly different, and yet the gloss is the same as Xuan;
- 4.) is slightly or greatly different, and the gloss is slightly or greatly different from Xuan;
- 5.) is not in Xuan at all.

§ B.) Considering matches in the direction of **Xuan->DUAN**, there are cases in which XU Xuan's character form:

- 1.) and gloss are identical to DUAN;
- 2.) is identical, and yet the gloss is either slightly or greatly different from DUAN;
- 3.) is slightly or greatly different, and yet the gloss is the same as Xuan;
- 4.) is slightly or greatly different, and the gloss is slightly or greatly different from Xuan;
- 5.) is not in DUAN at all.

§ C.) Considering the overall bidirectional **DUAN<->Xuan** matching, 6 items must be marked:

- 1.) character form and gloss are identical in both;
- 2.) character form is identical, and yet the gloss is either slightly or greatly different;
- 3.) character form is slightly or greatly different, and yet the gloss is the same;
- 4.) character form is slightly or greatly different; gloss is slightly or greatly different;
- 5.) character form is not in Xuan;
- 6.) character form is not in DUAN.

To establish the character set relative to the Xuan text, we were at first interested in marking character form variations, and neglected gloss variations. Thus, the following items were relevant:

Xuan character form is
- 1.) a match (with DUAN)
- 2.) a true variant (with DUAN)
- 3.) absent (from DUAN)

Under item 1.) matches with DUAN were marked (with match code) in the first pass through, while items 2.) & 3.) were unmarked (empty match code), and at first not distinguished from one another. In subsequent passes, as the distinction between true graphical variants (TV) and systematic variants (SV) became more clear, 2 degrees of variation in item 2.) came to be marked. At present the "Variancy" field is empty for a match, "1" for TV and "2" for SV.

4.3) MAPPING TO BIG5

The Big5 mapping performed in the three passes mentioned above was accomplished in a relational database environment employing four related databases. The first of these was then and is still today my primary copy. This primary database contains all of the DUAN and XU Xuan indices and relations to the graphical images of both texts, as well as relations to other indices, including Unicode Unihan data. At present this database goes by the name of 說文解字電子版, although it is called simply "DUAN" in the table below. The second of these was used primarily in the first two passes; as Big5 forms were matched that #2 database gradually lost its utility. Tabulated below are brief descriptions of these databases:

SW Big5 Mapping Databases		
#	**Name**	**Description**
1	DUAN	The SWJZZ MV character set (sequential DUAN-ordered list of types)
2	Big5 E	Distinct Big5 characters occurring in MV Entries (Xuan-ordered)
3	Big5 G	Distinct Big5 characters occurring in MV Glosses
4	Big5 EG	Distinct Big5 characters occurring in MV Entries and Glosses

In working between these four databases, the correspondences between forms were sought and then set via relations. Work in these databases also involved exploring the differences between the Big5 character set of the MV entries field, and the Big5 character set of the Gloss field. This was mentioned in passing in Section 2 under the sub-heading *Components*, and may be elaborated here.

It should not be surprising that some MV entry characters might not occur in the Gloss field. It may however be surprising to find Big5 characters occurring in the Gloss field which do not however occur in the MV field. It was surprising to me at any rate when I first discovered that, for example, the character 免 does not have an MV entry in the XU Xuan text at all. It is surprising that this character, such a productive grapheme in forms such as those exhibited below should have been omitted. DUAN Yucai in fact added it to his text, with some speculation as to how it could have been missed. Out of 15 Big5 characters with this component, 10 occur in SW:

Unicode	HYPY	Big5	Rad	Phon	Str	SW	English Gloss
[U+514d]	mian3	免	10		7		to avoid
[U+52c9]	mian3	勉	19	免	9		exhort
[U+5195]	mian3	冕	13E	免	11		royal crown
[U+7d7b]	mian3	絻	120	免	13		cap
[U+665a]	wan3	晚	72	免	11		evening, late
[U+8f13]	wan3	輓	159	免	14		draw, pull, send funeral ode
[U+4fdb]	fu2, mian3	俛	9	免	9		incline, bow; bow the head
[U+6d7c]	mei3	浼	85	免	10		pollute; ask a favor of
[U+774c]	wan3	睌	109	免	12		look at
[U+9794]	man2, wan3	鞔	177	免	16		upper (of shoe)
[U+9bb8]	man3, mian3	鮸	195	免	18		slate cod croaker (rice fish)
[U+5a29]	mian3, wan3	娩	38	免	10		give birth to a child
[U+633d]	wan3	挽	64	免	10		draw, pull, send funeral ode
[U+6097]	man2, men3	悗	61	免	10		
[U+8115]	wan3	脘	130	免	11		

4.4) MAPPING TO UNICODE

Although links to the high resolution graphical images of the complete DUAN and XU Xuan texts compensate to some extent for the lack of adequate Kaishu fonts for encoded representation, work remains to be done on the production of an adequate Kaishu font. In this regard, those image files play a crucial part, serving as the basis for future inputting and proofing against the printed source text. Manual input of DUAN's lengthy commentary is at this point not envisaged, as improvements in encoding and OCR technology will no doubt make this easier in the very near future.

My Kaishu fonts for the complete HYDZD character set (cf. the acknowledgments to Prof. 謝清俊 C.C. Hsieh at the end of this paper) will eventually provide a means of encoding and representing most every character in all three principal SW texts. Where this character set fails, it will be augmented. Furthermore, the mapping of both MV and Gloss data to Unicode will be accomplished more fully employing these fonts in combination with the IRG's HYDZD mapping data.

The IRG's work on *CJK Unified Ideographs Extension B* is to be commended for the large number of previously unencoded characters which it places. The primary Chinese sources which are most exciting are the HYDZD and Kangxi additions. As mentioned above at the beginning of Section 1, the SW texts (including DUAN Yucai's edition) occupy an important position in both of these character dictionaries, and so their treatment of SW's character set should be very good indeed.

One may note that Extension B (42,711 ideographs) includes not simply the characters in the 7 main volumes of HYDZD, but also those in its 44 page Appendix at the end of Volume 8. A most pertinent example is Extension B's [U-00020B82]. This character occurs on the first page of the HYDZD Volume 8 Appendix, and does not occur in the Kangxi dictionary. In fact, this character originates in DUAN's change of the Seal character for 厲 [446.420] to (the reference here uses the SWJZZ indexing system described above in Section 3, *Font and Index Production*; see Cook 1996:104 for discussion of this character form).

◊ 5.) CONCLUSIONS

5.1.) TEXT-BASED TYPOLOGICAL ENCODING

At present most of the mapping of the DUAN Yucai SWJZZ character set to Unicode has been accomplished on the basis of and in accordance with the Big5 mapping work described above. Even though the current and forthcoming Unicode Unihan character sets are by far superior to Big5, it must be clear from elements of the previous discussion that many of the matching and typological problems which plague the Big5 mapping will carry through to any other Chinese encoding which is not adequately typologized.

It is my belief that adequate historical typologization of the Chinese script can only be accomplished with reference to specific texts and inscriptions, and this is what I mean when I refer to a "Text-based" or "Source-based" encoding. These encodings may also be termed "Contextual" in that they seek to document the historical context whence the glyph usage derives. The written sources of ancient Chinese are many and varied, and each offers stylistic peculiarities and mapping challenges. Oracle Bone Inscriptions, Bronze Inscriptions, Stone and Earthenware Inscriptions ... all of these contain vital historical information which only a typological system can address.

A typologization adequate for Chinese purposes would be rather simple in some respects. First, it should characterize aspects of GLYPH SHAPE, and second, it should characterize aspects of GLYPH USAGE. One may however complicate and broaden this scheme in many ways, for example with script-specific issues of componential encoding. I believe that the future of Chinese encoding lies in a typologized text-based componential encoding, and that no other method can be as successful. It seems also that distinctions adequate for the handling of the Chinese script must have implications for the encoding of other scripts. The following is an outline summarizing more generally applicable elements of some of the databases mentioned above.

- **Fields of TYPE 1, relating to GLYPH SHAPE:**

SHAPE:CLASS: Type [main,variant];
main [shape class(es),variants:list];
variant [main class,variant class];
valence [isolate,combining:list(compounds)].
element [elemental,compound:list(components)];
SHAPE:VALUE: Type [outline,bitmap];
outline [lines,curves,metrics,hinting,kerning];
bitmap [size,points,color];
render [size,position,rotation,orientation,style];

- **Fields of TYPE 2, relating to GLYPH USAGE:**

USAGE:CLASS: Type [print (or inscription),image,encoded];
print [bibliographic citation];
image [jpg,gif...;link];
encoding [codepoint(s)];
location [page.line];
stats [instance.frequency];
status in source [active,defunct...].
USAGE:VALUE: Type [orthographic,phonologic,phonetic,morphologic,semantic,syntactic...]
analysis [source data:description,variant class...].

• LIST OF APPENDICES

APPENDIX 1: TABLE OF THE 540 SWJZZ 部首 BUSHOU 'CLASSIFIERS'

001一001.11, 002二001.41, 003示002.31, 004三009.21, 005王009.31, 006玉010.11, 007珏019.41, 008气020.12, 009士020.21, 010丨020.41, 011屮021.31, 012艸022.22, 013蓐047.41, 014茻047.43, 015小048.31, 016八048.41, 017釆050.11, 018半050.23, 019牛050.32, 020犛053.21, 021告053.41, 022口054.11, 023凵062.31, 024吅062.41, 025哭063.21, 026走063.31, 027止067.31, 028癶068.24, 029步068.31, 030此068.41, 031正069.31, 032是069.33, 033辵070.11, 034彳076.11, 035廴077.34, 036㢟077.44, 037行078.11, 038齒078.33, 039牙080.41, 040足081.12, 041疋084.44, 042品085.21, 043龠085.24, 044冊085.41, 045㗊086.31, 046舌086.43, 047干087.13, 048𧮫087.22, 049只087.41, 050㕯088.11, 051句088.14, 052丩088.31, 053古088.41, 054十088.43, 055卅089.31, 056言089.33, 057誩102.13, 058音102.23, 059䇂102.41, 060丵103.11, 061菐103.31, 062𠬞103.41, 063𠬜104.43, 064共105.12, 065異105.21, 066舁105.31, 067𦥑105.41, 068䢅105.43, 069爨106.12, 070革107.11, 071鬲111.13, 072䰜112.11, 073爪113.21, 074丮113.32, 075鬥114.14, 076又114.41, 077𠂇116.41, 078史116.43, 079支117.11, 080𦘒117.21, 081聿117.24, 082畫117.41, 083隶117.43, 084臤118.21, 085臣118.33, 086殳118.42, 087殺120.31, 088𠘧120.42, 089寸121.12, 090皮122.11, 091㼱122.21, 092攴122.31, 093教127.11, 094卜127.21, 095用128.21, 096爻128.32, 097㸚128.41, 098𥄕129.31, 099目129.43, 100䀠135.43, 101眉136.13, 102盾136.22, 103自136.33, 104白136.42, 105鼻137.31, 106皕137.41, 107習138.11, 108羽138.13, 109隹141.11, 110奞144.21, 111雈144.24, 112𦫳144.41, 113𥄎145.11, 114羊145.22, 115羴147.31, 116瞿147.33, 117雔147.42, 118雥148.13, 119鳥148.16, 120烏157.15, 121𠦒158.11, 122冓158.31, 123幺158.41, 124𢆶158.43, 125叀159.12, 126玄159.31, 127予159.41, 128放160.13, 129𠬪160.21, 130𣦼161.12, 131歺161.31, 132死164.21, 133冎164.32, 134骨164.42, 135肉167.21, 136筋178.11, 137刀178.14, 138刃183.21, 139㓞183.24, 140丯183.33, 141耒183.42, 142角184.41, 143竹189.11, 144箕199.21, 145丌199.31, 146左200.32, 147工201.11, 148㠭201.31, 149巫201.33, 150甘202.11, 151曰202.31, 152乃202.33, 153丂203.21, 154可203.32, 155兮204.11, 156号204.15, 157亏204.31, 158旨204.33, 159喜205.12, 160壴205.23, 161鼓206.11, 162豈206.43, 163豆207.22, 164豊208.12, 165豐208.21, 166䖒208.42, 167虍209.13, 168虎210.21, 169虤211.31, 170皿211.34, 171𠙴213.31, 172去213.32, 173血213.41, 174丶214.43, 175丹215.31, 176青215.41, 177井216.11, 178皀216.31, 179鬯217.11, 180食218.21, 181亼222.42, 182會223.21, 183倉223.41, 184入224.11, 185缶224.31, 186矢226.13, 187高227.33, 188冂228.11, 189𩫖228.32, 190京229.11, 191亯229.13, 192㫗229.41, 193畗230.11, 194㐭230.21, 195嗇230.42, 196來231.12, 197麥231.32, 198夊232.34, 199舛234.11, 200舜234.21, 201韋234.31, 202弟236.31, 203夂237.11, 204久237.23, 205桀237.31, 206木238.31, 207東271.21, 208林271.23, 209才272.21, 210叒272.31, 211之272.42, 212帀273.11, 213出273.21, 214𣎵273.31, 215生274.13, 216乇274.31, 217𠂹274.32, 218𠌶274.41, 219華275.11, 220𥝌275.13, 221稽275.31, 222巢275.41, 223桼276.11, 224束276.22, 225㯻276.33, 226囗276.44, 227員279.11, 228貝279.21, 229邑283.12, 230𨛜300.41, 231日302.11, 232旦308.31, 233倝308.41, 234㫃308.44, 235冥312.31, 236晶312.41, 237月313.22, 238有314.12, 239朙314.23, 240囧314.41, 241夕315.21, 242多316.12, 243毌316.23, 244𢎘316.32, 245𣐺317.12, 246𠧪317.22, 247齊317.41, 248朿318.11, 249片318.21, 250鼎319.12, 251克320.11, 252彔320.21, 253禾320.22, 254秝329.21, 255黍329.32, 256香330.23, 257米330.32, 258毇334.11, 259臼334.21, 260凶334.42, 261朩335.31, 262𣏟335.41, 263麻336.12, 264尗336.24, 265耑336.32, 266韭336.41, 267瓜337.14, 268瓠337.35, 269宀337.41, 270宮342.41, 271呂343.11, 272穴343.32, 273㝱347.23, 274疒348.14, 275冖353.12, 276𠔼353.32,

277 353.44, 278 354.33, 279 355.11, 280 357.11, 281 357.21, 282 362.32,
283 363.31, 284 363.41, 285 364.21, 286 364.23, 287 365.11, 288 384.21,
289 384.42, 290 386.11, 291 386.21, 292 386.31, 293 386.33, 294 387.21,
295 387.31, 296 388.11, 297 388.13, 298 388.24, 299 388.32, 300 388.41,
301 398.11, 302 398.13, 303 398.44, 304 399.31, 305 399.33, 306 401.31,
307 402.11, 308 402.32, 309 403.11, 310 404.23, 311 404.41, 312 405.31,
313 405.42, 314 406.12, 315 406.41, 316 406.43, 317 407.12, 318 407.31,
319 410.14, 320 410.22, 321 414.14, 322 414.22, 323 414.33, 324 415.31,
325 422.23, 326 422.32, 327 423.12, 328 423.13, 329 423.41, 330 424.11,
331 424.23, 332 425.13, 333 425.23, 334 425.41, 335 429.31, 336 429.41,
337 430.21, 338 430.31, 339 431.31, 340 431.41, 341 432.21, 342 432.31,
343 432.42, 344 434.11, 345 434.31, 346 434.42, 347 436.31, 348 436.41,
349 437.11, 350 437.31, 351 441.34, 352 442.11, 353 442.32, 354 446.24,
355 448.22, 356 448.34, 357 448.42, 358 453.21, 359 453.41, 360 454.12,
361 454.21, 362 454.31, 363 456.21, 364 456.41, 365 457.11, 366 457.21,
367 458.41, 368 459.11, 369 459.31, 370 460.31, 371 469.35, 372 470.12,
373 472.21, 374 472.23, 375 472.34, 376 473.12, 377 473.21, 378 478.21,
379 478.31, 380 479.35, 381 479.41, 382 480.12, 383 487.11, 384 487.41,
385 490.31, 386 490.41, 387 491.11, 388 491.32, 389 492.21, 390 493.41,
391 494.11, 392 494.21, 393 494.41, 394 495.12, 395 495.41, 396 496.11,
397 496.21, 398 497.11, 399 497.22, 400 497.42, 401 498.31, 402 498.42,
403 499.32, 404 500.11, 405 501.12, 406 501.21, 407 501.32, 408 501.42,
409 515.34, 410 516.11, 411 567.31, 412 567.41, 413 568.11, 414 568.21,
415 568.32, 416 569.31, 417 569.41, 418 569.43, 419 570.12, 420 570.22,
421 570.42, 422 571.41, 423 575.11, 424 575.21, 425 582.11, 426 582.21,
427 582.31, 428 582.42, 429 583.11, 430 583.23, 431 584.11, 432 584.31,
433 584.41, 434 585.31, 435 586.11, 436 586.22, 437 586.41, 438 587.21,
439 591.14, 440 593.21, 441 593.41, 442 611.22, 443 612.11, 444 626.31,
445 627.11, 446 627.21, 447 627.33, 448 627.42, 449 628.11, 450 628.32,
451 628.44, 452 632.31, 453 632.41, 454 633.31, 455 633.41, 456 634.12,
457 634.21, 458 635.11, 459 635.41, 460 637.22, 461 637.33, 462 638.12,
463 639.42, 464 642.11, 465 642.21, 466 642.33, 467 643.31, 468 662.32,
469 663.11, 470 663.21, 471 663.22, 472 674.31, 473 676.21, 474 677.21,
475 678.21, 476 678.31, 477 679.11, 478 680.21, 479 681.11, 480 682.11,
481 694.11, 482 694.21, 483 694.31, 484 694.41, 485 698.12, 486 698.21,
487 698.35, 488 699.11, 489 701.31, 490 702.11, 491 715.13, 492 715.21,
493 715.31, 494 716.12, 495 716.41, 496 717.34, 497 719.31, 498 720.12,
499 730.33, 500 731.11, 501 737.11, 502 737.21, 503 737.32, 504 737.41,
505 738.12, 506 738.21, 507 738.23, 508 738.31, 509 738.32, 510 738.41,
511 739.11, 512 739.41, 513 740.11, 514 740.21, 515 740.41, 516 740.42,
517 741.11, 518 741.13, 519 741.31, 520 741.33, 521 741.41, 522 742.21,
523 742.23, 524 742.31, 525 742.41, 526 743.42, 527 744.12, 528 744.22,
529 744.41, 530 745.12, 531 745.21, 532 745.31, 533 745.42, 534 746.21,
535 746.31, 536 746.32, 537 747.13, 538 752.11, 539 752.21, 540 752.31.

APPENDIX 2: SAMPLE OF THE SWJZZ CHARACTER SET [745.310]

Translation of the above appears in (COOK 1996, Section 01).

APPENDIX 3: HARDWARE AND SOFTWARE NOTES

A variety of hardware and software has been employed over the course of this project. The list below documents some of the more important pieces of equipment used during the six year period. It is hoped that aside from documenting the progress made in personal computing (25MHz > 250MHz in four years), this list may also provide people seeking to do similar work with some idea of some of the resources available. Please see Section 3 above for discussion of the manner in which I came to choose and use some of this equipment.

>=1994:

• Affinity Microsystems *Tempo EZ* (macro utility) [this utility served to semi-automate the scanning and character template extraction and placement processes]
• Altsys *Fontographer*, version 3.5. (typographical design software) [this software was used to produce the character outlines and subfonts]
• Apple *LaserWriter Select 360* Laser Printer
• Apple Macintosh *Performa LC475*, (25 MHz 680LC40 CPU, 4->36MB RAM, 160MB HD)
• Apple Macintosh System 7 *Chinese Language Kit*, (Simplified and Traditional)
• *FileMaker Pro* 2.1v3 (database software)
• Hewlett Packard *DeskWriterC*, 300 dpi InkJet Printer
• MicroFrontier *ColorIt!* version 2.3.3 (graphical image processing software) [this was my primary image processing and scanning software]
• Microsoft *Word* 5.1a (word processing software)
• Microsoft *Works* 3.0 (database software)
• Quantum 4GB Q4000 external fixed media drive (SCSI)
• SyQuest 270MB removable media drive (SCSI)
• UMAX *Vista-S8*, 400ppi optical resolution (flatbed color scanner)
• Wenlin Institute 文林 *Wenlin.* (electronic Chinese dictionary and encoding, by Tom Bishop)
• *Word Perfect* 3.0 (2-byte savvy word processing software)

>=1998:

• Apple Macintosh MacOS 9 (soon to be X!)
• Apple Macintosh PowerBook G3 Series, (250MHz G3 processor, 288MB RAM,4GB IBM HD)
• Binary Software *KeyQuencer* 2.5.5 (by Alessandro Levi Montalcini)
• Corel *Word Perfect* 3.54 (2-byte savvy word processing software)
• *FileMaker Pro* 4.0v3 (relational database software)
• *FileMaker Pro Chinese* 4.0 (2-byte savvy relational database software, discontinued product)
• *FontLab* 3.1.2 (typographical design software)
• FontLab *MacComposer* betas (currently at 2.0b4) [big CIDFont software]
• Hewlett Packard *LaserJet 6MP*, 600 dpi Laser Printer
• Hewlett Packard *ScanJet 5300Cxi*, 1200ppi optical resolution (36-bit flatbed color scanner)
• IBM *Travelstar* 12GB IDE external fixed media hard drive (PCMCIA)
• *MacPerl* 5.20r4 (by Larry Wall & co.!)
• Macromedia *Fontographer*, version 4.1.4. (typographical design software)
• Seagate 4GB external fixed media drive (SCSI)
• Trans Tex Software *TexEdit Plus* 4.1 (2-byte savvy styled text editor, by Tom Bender)
• Yamaha CRW4416S external 650MB CDR removable media drive (SCSI)

APPENDIX 4: THE SWJZZ CMAP (THOUSANDS OF CID'S OMITTED)

```
%!PS-Adobe-3.0 Resource-CMap
%%DocumentNeededResources: ProcSet (CIDInit)
%%IncludeResource: ProcSet (CIDInit)
%%BeginResource: CMap (RSCook-B5-H)
%%Title: (RSCook-B5-H RSCook BigFive1 0)
%%MSIdentification 1
%%CreationDate: 2000-11-11 11:11:11
%%Version: 1.111
%%Copyright: 1998-2000 Richard Sterling Cook.
%%Copyright: All Rights Reserved.
%%EndComments
/CIDInit /ProcSet findresource begin
12 dict begin
begincmap
/CIDSystemInfo 3 dict dup begin
  /Registry (RSCook) def
  /Ordering (BigFive1) def
  /Supplement 0 def
end def
/CMapName /RSCook-B5-H def
/CMapVersion 1.000 def
/CMapType 1 def
/XUID [11 25 1962] def
/UIDOffset 0 def
/WMode 0 def
2 begincodespacerange
<00>   <80>
<A140> <FEFE>
endcodespacerange
100 begincidrange
<20> <7e> 11247
<a140> <a17e> 1
<a1a1> <a1fe> 64
<a240> <a27e> 158
<a2a1> <a2fe> 221
<a340> <a37e> 315
<a3a1> <a3fe> 378    [<-Thousands of CID's Omitted.]
<e640> <e67e> 10834
<e6a1> <e6fe> 10897
<e740> <e77e> 10991
<e7a1> <e7fe> 11054
<e840> <e87e> 11148
<e8a1> <e8c4> 11211
endcidrange
endcmap
CMapName currentdict /CMap defineresource pop
end
end
%%EndResource
%%EOF
```

Appendix 5: Abbreviations and Glossary

ALSL: Apple LiSung Light, my reference Big5 font (cf. CLK below).
ATM: Adobe Type Manager; <http://www.adobe.com/>.
Big5: A common Traditional Chinese double-byte encoding standard (cf. Lunde 1999:171).
Bushou: 部首 Chinese 'lexical classifiers', a.k.a. "Radicals". See Appendix 1.
Chongwen: 重文 Chinese 'historical character stylistic variant'. See Section 2 above.
CIDFont: Character I. D. Font. Cf. Lunde (1999, p.288*ff*).
CLK: Apple Computer's *Chinese Language Kit*, a MacOS 9 install option (see Appendix 3).
CMap: Character Map. Cf. Lunde (1999, p.290*ff*).
FMP: FileMaker Pro (see Appendix 3).
FOG: Macromedia Fontographer (see Appendix 3).
GIF, JPG, **JPEG**: Digital graphical image file formats.
GSR: Cf. Karlgren in the References.
Han: 漢 'Chinese dynasty names; 西漢 Western Han dynasty (206 B.C.-24 A.D.), 東漢 Eastern Han dynasty (25-220).
HY: Harvard Yan Jing (Yen-Ching) Library, <http://hcl.harvard.edu/harvard-yenching/>.
HYDZD:《漢語大字典》*Hanyu Da Zidian.* Character lexicon. See 許力以 Xu Liyi (1993).
HYPY: 漢語拼音 Hanyupinyin romanization of Modern Standard (Beijing) Chinese.
Jinwen: 金文 Chinese 'bronze inscription character'. Cf., e.g. 周法高 Zhou Fagao (1981).
Kaishu: 楷書 Chinese 'square script', Song Dynasty calligraphic style, stylistic basis for modern Chinese typographic styles.
Kangxi:《康熙》 Character lexicon. Cf. 長玉書 Zhang Yushu (1716).
LTBA: Journal *Linguistics of the Tibeto-Burman Area*, <http://stedt.berkeley.edu/ltba/>.
MacOS: Apple Macintosh Operating System. <http://www.apple.com/>.
MV: Main/Variant (cf. Section 4).
PDF: Portable Document Format; <http://www.adobe.com/>.
PPI: pixels per inch, computer monitor resolution.
Qing: 清 'Chinese dynasty name', (1644-1911).
Seal: 篆 Chinese character style. Cf. Xiaozhuan (see Section 2 above).
Shang: 商 'Chinese dynasty name', (~16th cent. B.C.-~1045 B.C.)
Song: 宋 'Chinese dynasty names'; 北宋 Northern (960-1127) and 南宋 Southern (1127-1279).
Songti: 宋體 Chinese calligraphic style, cf. Kaishu.
STEDT: Sino-Tibetan Etymological Dictionary and Thesaurus, <http://stedt.berkeley.edu/>.
SW:《說文》。*Shuowen.* (Short for SWJZ.)
SWJZ:《說文解字》。*Shuowen Jiezi* (cf. References); also, the name of my BNU-based font.
SWJZ-FJZ:《說文解字・附檢字》。*Shuowen Jiezi - Fu Jianzi* (cf. Appendix 6).
SWJZZ:《說文解字注》。*Shuowen Jiezi - Zhu* (cf. References); also the name of my font.
TIFF: Tagged Image File Format. A graphical image file format.
TTF: TrueType Font, a scalable font format, developed jointly by Apple and Microsoft.
Type 1: Adobe Systems' format for describing scalable fonts.
UCBEAL: UC Berkeley, East Asian Library, <http://www.lib.berkeley.edu/EAL/>.
Xiaozhuan: 小篆 Chinese 'Small-Seal character' (see Section 2 above).
Xu Kai: 徐鍇 (920-974) 'Author of a Song version of SW', (cf. Appendix 6).
Xu Shen: 許慎 (58?-147?) 'Eastern Han author of SW', (cf. Appendix 6).
Xu Xuan: 徐鉉 (916-991) 'Author of a Song version of SW', (cf. Appendix 6).
Zhou: 周 'Chinese dynasty names'; 西周 Western Zhou dynasty (~1027-~771 B.C.), 東周 Eastern Zhou dynasty (~770-~256 B.C.)

APPENDIX 6: REFERENCES (SELECTED)

BAXTER, William Hubbard 白一平
1992 *A Handbook of Old Chinese Phonology*. Berlin, New York: Mouton de Gruyter, 1992 (Trends in Linguistics: Studies and Monographs; 64). ISBN 3 11-012342-x.

BISHOP, Thomas Eugene 畢曉普
2001 文林 *Wenlin.* Portland: Wenlin Institute. <http://www.wenlin.com/>.

COOK, Richard Sterling 曲理察
1995 *The Etymology of Chinese 辰 Chén.*《‘辰’字的原始義》。 Monograph Volume of the Biannual Journal *Linguistics of the Tibeto-Burman Area* (LTBA), James A. Matisoff, editor. Volume 18.2, 278 pp., including abstract, end notes, indices, editor's preface, and cumulative LTBA index. <http://stedt.berkeley.edu/ltba/>.

DING Fubao 丁福保
1959 《説文解字・詁林及補遺》。*Shuowen Jiezi - Gu Lin ji Buyi.* [SWJZ-Assemblage of Multiple Versions, and Addenda.] 丁福保編纂。臺北：臺灣商務印書館，1959. HY: R5099 12432(v.1-12); Cf. UCB EAST: 5092.1032; EAL PL1281.H83 T5 1977 RR.

DUAN Yucai 段玉裁 (1735-1815)
1815 《説文解字・注》。*Shuowen Jiezi - Zhu.* [SWJZ-Annotated; Qing recension.]〔東漢〕許慎著〔清〕段玉裁注。上海：上海古籍出版社，1989. ISBN 7-5325-0487-5/H.6.
1815 《説文解字・注》。臺北市：洪葉文化，1998. [W/ red MV.] ISBN 957-8424-34-5.

GE Benyi 葛本儀
1993 《實用中國語言學詞典》。*Shiyong Zhongguo Yuyanxue Cidian.* [Pragmatic Dictionary of Chinese Linguistics.] 葛本儀主編。青島：青島出版社。 ISBN 7-5436-0773-5/H.6.

GUI Fu 桂馥 (1736-1805)
1805 《説文解字・義證》。*Shuowen Jiezi - Yi Zheng.* [SWJZ-Hermeneutics.] 清桂馥選。濟南：齊魯書社出版發行，1987. ISBN 7-5333-0061-0/H.4.

HSU, James C.H. (XU Jinxiong) 許進雄
1996 *The Written Word in Ancient China*. Hong Kong: TAN Hok seng.

ISO/IEC
2000 *Information technology - Universal Multiple-Octet Coded Character Set (UCS) - Part1: Architecture and Basic Multilingual Plane.* Reference Number ISO/IEC 10646-1 :2000(E). Second Edition, 2000-09-15. <http://www.iso.ch>.

ISO/IEC JTC1/SC2/WG2/IRG (Ideographic Rapporteur Group)
2000 *CJK Unified Ideographs Extension B,* for ISO/IEC FCD(R2) 10646-2:2000, Ordered by Kangxi Dictionary. 2000.08.22. <http://www.cse.cuhk.edu.hk/~irg/>.

JENKINS, John H. 井作恒
1999 "New Ideographs in Unicode 3.0 and Beyond". Paper presented at the 15th International Unicode Conference. San Jose, California.

KARLGREN, Bernhard 高本漢
2000 *Grammata Serica Recensa Electronica.* Complete relational database: indices, syllable canon, & images of the original text. Prepared for the STEDT Project by Richard Cook; indices after Tor Ulving (Univ. Göteborg, 1997). Berkeley: University of California.

1957 *Grammata Serica Recensa*. First published by The Museum of Far Eastern Antiquities, Stockholm Bulletin No. 29, Stockholm, 1957. Reprinted by Elanders Boktrycker Aktiebolag, Kungsbacka, 1972. HY: PL1201K341957x.

1954 *Compendium of Phonetics in Ancient and Archaic Chinese*. First published by The Museum of Far Eastern Antiquities, Stockholm Bulletin No. 26, Stockholm, 1954. Reprinted by SMC Publishing Inc., Taipei, Taiwan, R.O.C. 1992. ISBN 957-638-123-1.

1950 *The Book of Odes: Chinese Text, Transcription and Translation*. Stockholm: The Museum of Far Eastern Antiquities, 1950. HY: (W) 439 43.

1940 《中國音韻學研究》，高本漢著：趙元任、李方桂合釋。臺北：商務印書館出版，1940 [1962]. HY: RR R 5120 0253.56. An "amiable 'retouche'" of: *Etudes sur la phonologie chinoise*, par Bernhard Karlgren. [Upsala, K. W. Appelberg]; Leyde, E.-J. Brill; 1915-1926. HY: (W) PL1201.K32

1923 *Analytic Dictionary of Chinese and Sino-Japanese.* (ADCS) First published by the Libraire Orientaliste Paul Geuthner, Paris, 1923. Reprinted by Dover Publications Inc., New York, 1991. ISBN 0-486-26887-X.

LI Fang-kuei (LI Fanggui) 李方桂

1971 《上古音研究》，李方桂。北京：商務印書館，新華書店。HY: W9220 1281. 清華學報 n.s. 9, 1971 [1982], 1-61. Translated by Gilbert L. MATTOS as "Studies on Archaic Chinese", LI Fang-kuei, *Monumenta Serica* 31 (1974-5), pp. 219-87. HY: 5121 4404.

LI Zhenhua & ZHOU Changji 李珍華、周長楫 (LI&ZHOU)

1993 《漢字古今音表》。〔美〕李珍華、周長楫編撰。北京：中華書局。 HY: 5120 4414.

LU Zongda 陸宗達

1981 《說文解字・通論》。*Shuo wen jie zi - tong lun.* [General Survey of SW.] 北京：北京出版社, 新華書店發行所發行。UCB EAL PL1281.H83 L8.

LUNDE, Ken 小林劍

1999 *CJKV Information Processing*. Beijing, Cambridge: O'Reilly. ISBN: 1-56592-224-7.

1999 "Acrobat 4.0 Adds CJKV Features: Embedding multibyte fonts in PDF files eases cross-platform use of documents". *Multilingual Computing and Technology*, Volume 10, Issue 6. <http://www.multilingual.com/>.

LUO Zhufeng 羅竹風

1997 《漢語大詞典》。*Hanyu Da Cidian*. [Chinese Lexicon, 3 Vol., 7923pp.] 羅竹風主編。上海:漢語大詞典出板社。ISBN: 7-5432-0014-7/H.15.

MILLER, Roy Andrew

1977 "The Wu-Ching I-I 〔《五經異義》〕of Hsü Shen〔許慎〕." *Monumenta Serica*, 33 (1977-8) pp. 1-21. HY: 2105.23674.

1953 Problems in the study of *Shuo-wen chieh-tzu*. Ann Arbor: University Microfilms International, 1978. Photocopy of Ph.D. Thesis typescript (353pp.)- Columbia University. UCB East Asian 5099.8477.

NEEDHAM, Joseph 李約瑟

1954- *Science and Civilization in China*. Joseph Needham with Wang Ling. Cambridge University Press, 1985. Volume 5: Chemistry and Chemical Technology, Part 1: Paper and Printing (by Tsien Tsuen-hsuin). HY: Ref (W)DS721.N39.

SHA Qingyan 沙青岩
1980 《說文大字典》。*Shuowen Da Zidian*. [SW rearranged, modified, w/ large-size head-entry characters; scanned at 師範學院 (BNU) to automate production of the outlines for their SW font.] 天津：天津市美術出版社。 UCB EAL PL1281.H83 S43.

Unicode Consortium 統一碼團
2000 *The Unicode Standard, Version 3.0*. Reading, Massachusetts: Addison-Wesley.

XIA Zhengnong 夏征農
1992 《辭海》。*Ci Hai.* [Encyclopedic Lexicon.] 夏征農主編。上海：上海辭書出版社，1992. ISBN 7-5326-0135-8/Z.12.

XU Liyi 許力以等（漢語大字典工作委員會）
1990 《漢語大字典》。*Hanyu Da Zidian.* [Etymological Character Lexicon, 8 vol.] 許力以主任，徐中舒主編。武漢：四川辭書出版社，湖北辭書出版社。 ISBN 7-5403-0030-2/H.16. (Also in a 縮印本 'condensed edition' of 1 vol., ISBN 7-80543-239-2/H.63.)

XU Kai 徐鍇 (920-974)
<974 《說文解字・繫傳》。*Shuowen Jiezi - Ji Zhuan.*〔東漢〕許慎著〔南唐〕徐鍇選。北京：中華書局，1987. ISBN 7-101-00060-6/H.7.

XU Shen 許慎 (58?-147?)
121 《說文解字》。*Shuowen Jiezi.*〔東漢〕許慎著。 [Lost: see XU Kai and XU Xuan.]

XU Xuan 徐鉉 (916-991)
c.987 《說文解字・附檢字》。*Shuowen Jiezi - Fu Jianzi.* [Common modern Hong Kong edition, with appended indices; in hardcover and paperback; 陳昌治 (1873).]〔東漢〕許慎著〔南唐、宋〕徐鉉校定。香港：中華書局，1989. ISBN 962-231-208-X.
c.987 《說文解字》。*Shuowen Jiezi.* [Another common modern edition, without appended indices; this edition seems to be that of 孫星衍 (1809), upon which 陳昌治 (1873) was based.] 中國書店出版社，據商務印書館版本影印，1989. ISBN: 7-80568-039-6/H.2.

YU Nae-wing (YU Naiyong) 余迺永
1993 《新校互註・宋本廣韻》。*Xin Jiao Hu Zhu - Song Ben Guang Yun* ['A New Revision of the Sung Edition of the Kuang-yun Rhyming Dictionary'.] 香港：香港中文大學。 Hong Kong: Xiang Gang Zhong Wen Da Xue. (1 vol., ~900pp., hardcover, with indices and English appendix.) ISBN 962-201-413-5.

ZHANG Yushu 長玉書
1716 《新修康熙字典》。*Xin Xiu Kang Xi Zidian.*〔清〕長玉書等總閱，凌紹雯等纂修，高樹藩重修。臺北：啟業書局印行，1978.

ZHAO Pingan 趙平安
1999 〈《說文解字》小篆研究〉。*SWJZ Xiao Zhuan Yanjiu* [Research on the Small-Seal Characters of *SWJZ*.] 廣西教育出版社。 ISBN: 7-5435-2749-9/H.81.

ZHOU Fagao 周法高
1981 《金文詁林補》周法高編。香港中文大學。 8 Volumes. HY: 2105.6 7230.85.
1974 《金文詁林》周法高編。香港中文大學。 16 Volumes. HY: R2105.6 7230.8(1-16).

ZHU Minshen 祝敏申
1999 《說文解字》與中國古文字學。*SWJZ yu Zhong Guo Guwenzixue* ["*SWJZ: The Dawn of Studies of the Anc. Chars.*"] 上海：復旦大學出版社。ISBN 7-309-02050-2/H.333.

• ACKNOWLEDGMENTS

First, I would like to thank Ken LUNDE <lunde@adobe.com> for the boundless patience and detailed technical assistance which he has offered over many years. Without Ken's help the CIDFonts employed in this work could not have been produced. It was Ken who introduced me to Alex SIMAGIN <al@legion.ru>, the FontLab programmer *par excellence* who produced the "MacComposer" program (as it was provisionally named) which produced the double-byte CIDFonts based on my Type 1 outlines. Ken taught me to write my first CMaps, answered my endless questions, and was responsible for much debugging, including the tracking down of the elusive "binary data size" a.k.a. "embed PDF" bug. Dirk MEYER <dmeyer@adobe.com>, Ken's CJKV typographical cohort, should also be acknowledged for his invaluable guidance and support.

My brother Michael <Michael.Cook@cisco.com> was and is an endless source of the highest quality technical advice. Even if he weren't my brother I should love him for his astounding programming knowledge and the infinite patience he exhibited in answering my Perl questions.

Another computer scientist to whom I owe a debt of thanks is Dr. Sheila GREIBACH <greibach@cs.ucla.edu>. Thanks to her for first setting me on the right relational database track, and for her suggestions of mapping approaches.

I should acknowledge the help of Prof. 謝清俊 C.C. HSIEH <hsieh@sinica.edu.tw> and 莊德明 Derming JUANG <derming@gate.sinica.edu.tw> of Academia Sinica, who provided me with copies of the BNU SW font described above. They are to be thanked also for providing me with copies of their massive HYDZD fonts and databases.

Richard SEARS <searsr@eng.sun.com> provided me with the Big5 SW data input by Ms. Ann WU, and I am very thankful to them both.

The debt I owe to 畢曉普 Tom BISHOP <wenlin@wenlin.com> is perhaps less obvious, and yet Tom's 文林 *Wenlin* software <http://www.wenlin.com/> makes my life so much easier in so many ways that it would be the gravest affront to omit recognition of his hard work. For Chinese encoding work, there is simply no better tool than *Wenlin*.

For initial help with translation of the abstract into Chinese, I am indebted to 林蕙珊 LIN Huishan of 國立清華大學語言學研究所 the National Tsing Hua University Graduate Institute of Linguistics. And for her kind assistance in the editing of the Chinese abstract and proof-reading of the final draft, I am grateful to 林英津 Prof. LIN Ying-chin of 中央研究院語言學研究所籌備處 Academia Sinica's Preparatory Institute of Linguistics.

I would also like to thank 胡侃 Ken WHISTLER <kenw@sybase.com>, 抹香庵 Rick MCGOWAN <rick@unicode.org> and 周伯楷 Joe BECKER <Joseph.Becker@pahv.xerox.com>, as it was at their kind suggestion that I undertook to prepare this presentation for IUC-18.

Suggestions on translation of technical terms were provided by the following members of the Unicode email list <unicode@unicode.org>: Thomas CHAN <thomas@atlas.datexx.com>, Tom EMERSON <tree@basistech.com>, Jenny PAN <Jenny.Pan@usa.xerox.com>, and Jungshik SHIN <jshin@pantheon.yale.edu>.

The writing of this paper was supported in part by grants from:

• The National Science Foundation (NSF), Division of Behavioral & Cognitive Sciences, Linguistics, Grant No. BCS-9904950;

• The National Endowment for the Humanities (NEH), Preservation and Access, Grant No. PA-23353-99.

Finally, and most importantly, the writing and presentation of this paper would not have been possible without the kind support and insightful perspective of Prof. James Alan MATISOFF.