Title: **Proposal to add DYC dictionary mappings to Unihan**
Type: Individual Contribution
Action: For UTC consideration
Source: Richard Cook
Date: 2019-02-15

- ***DYC dictionary source***

段玉裁 Duàn Yùcái (DYC) compiled an influential Qīng dynasty edition (c. 1815) of the Eastern Hàn Chinese dictionary《說文解字》*Shuō Wén Jiě Zì* (SW; 1st c.). SW is the first major standardization of the elements of Chinese writing, and is known today mainly from two 10th c. texts, one with 10,724 (Táng) and the other with 11,108 (Sòng) lexical entries. DYC's Qīng dynasty edition with 10,706 entries collates and annotates the received material, excludes forms appended in the Sòng edition, and is a primary lexical source underlying character identifications in later dictionaries. For example,《漢語大字典》*Hànyǔ Dà Zìdiǎn* (HDZ; 1986-1990) with 54,729 entries (including many variants) served as a primary print source for development of *CJK UI Extension B* (Unicode 3.1.0). HDZ mappings are represented today in the *Unicode Character Database* (UCD) by four Unihan properties: *kIRGHanyuDaZidian*, *kHanYu*, *kHanyuPinyin*, *kHDZRadBreak*. Each of the HDZ entries from the SW subset begins with either a reference to SW (often citing DYC), or else with a cross-reference to such an entry. In common modern editions of DYC's text, each of the 10,706 head entries corresponds to exactly one Seal form, also represented by modern Sòng-style characters appearing on each page of the text, in the header, body, and in appended indexes. These Sòng-style characters include the common and variant forms catalogued in HDZ.

- ***DYC dictionary mappings***

Because of HDZ's important role in determining the current UCS CJK repertory, and because of SW's prominence in HDZ character identifications, DYC's Song-style head-entry characters are (for the most part) all encoded in UCS. But encoded CJK character variants complicate the mappings: there are often multiple CJK code points available for the same abstract character, sometimes following the Seal forms stroke-for stroke. My inventory of the characters occurring in SW editions began in the 1980s, with results published in 2003 and in HDZ and SBGY Unihan properties. Since the publication of *CJK UI Extension B* (2001) my set of CJK mappings for the DYC head entries has been refined using a mechanism developed to handle encoded CJK variants, and updated for subsequent CJK extensions. It is this data which I propose to contribute now to the UCD.

• *Example Property Values*

The following lines (out of ~40,000) exemplify property syntax.

```
U+3976  kDYC  A1-510.441 B0-505.110 D0-505.110 #㥶
U+3977  kDYC  B1-503.410 #㥷
U+3978  kDYC  B0-512.440 B0-591.121 #㥸
U+3979  kDYC  B3-604.110 #㥹
U+397B  kDYC  C0-514.140 #㥻
U+397D  kDYC  B3-281.440 D0-502.120 D3-475.150 #㥽
U+397F  kDYC  B3-369.310 #㥿
U+3981  kDYC  A0-515.240 #㦁
U+3982  kDYC  B0-358.310 #㦂
U+3985  kDYC  B0-504.210 #㦅
U+2279D kDYC  B2-505.110 #𢞝
U+22888 kDYC  B0-510.441 #𢢈
U+22911 kDYC  B3-505.110 #𢤑
U+2295B kDYC  A0-505.110 D0-510.441 #𢥛
U+2295C kDYC  B1-505.110 #𢥜
```

...

Column 1 : USV (Unicode Scalar Value, Unihan CJK subset code point);
Column 2 : property name; edition DYC = 段玉裁 *Duàn Yùcái* (1735-1815);
Column 3 : property value, if > 1 entry, ASCII string-sorted and SPACE-delimited;
Comment# : representative glyph of USV, not for published data.

• *DYC Mapping Table (example lines)*

CIDN	DYC.QMV	PUAH	A	B	N	C?D!E#F (see p. 5-7)
7357	505.110	101cbc		𢥛	1	㥶𢥜𢞝𢤑!㥶愆僁#
7468	510.440	101d2b		愆	1	[illegible]#
7469	510.441	101d2c		㥶	3	𢢈!𢥛#[illegible]
7470	510.442	101d2d		僁	1	諐[illegible]!保#

HDZ (U+3976):

㥶  說文·心部

《説文》:“㥶，實也。从心，塞省聲。《虞書》曰:‘剛而㥶。’”朱駿聲通訓定聲:“經傳皆以塞為之。”

（一）sè 《廣韻》蘇則切，入德心。職部。

同“塞”。1.充实；充满。《説文·心部》:“㥶，實也。《虞書》曰:‘剛而㥶。’”朱駿聲通訓定聲:“經傳皆以塞為之。”按:今本《書·皋陶謨》作“剛而塞”。《集韻·代韻》:“㥶，實也。通作塞。”2.安定。《廣雅·釋詁一》:“㥶，安也。”王念孫疏證:“《方言》:‘猒、塞，安也。’郭璞注云:‘物足則定。’猒與懕通，塞與㥶通。”《龍龕手鑑·心部》:“㥶，心安也。”

（二）qiān 《集韻》丘虔切，平仙溪。

同“愆”。过错；超过。《玉篇·心部》:“㥶”，同“愆”。《集韻·僊韻》:“愆，《説文》:‘過也。’或从㥶省。”

HDZ (U+2295B):

𢥛 同“㥶”。《康熙字典·心部》:“𢥛，《正字通》:‘同㥶’。”

• ***Property Value Entry Syntax***

Each entry has the form **AO-DYC.QMV**, matching the following regex (PCRE):

```
/^[A-D][0-9A-Za-z]-[0-9]{3}\.[1-4][1-9AB][0-5]$/
```

AO- : 3 characters = 2-letter prefix for *Type* (**A**) and *Offset* (**O**), + hyphen (**-**):
A : 1 letter **[A-D]**, for the *Type* of match (Unihan 11.0.0):
A : best match (per stroke);
B : proper match (non-best, proper variant);
C : questionable match (possible/probable);
D : improper match (confusable, DYC lexical source separation).
O : 1 letter **[0-9A-Za-z]**, for the *Offset* (per Type **A**):
A=[**A**] : **O**=[**01**] : 0 = good; 1= bad (~46 new CJKUI);
A=[**B-D**] : **O**=[**0-9A-Za-z**] : offset per Type **A**.

DYC.QMV : 7-character mapping =
DYC. : 4 characters, zero-padded 3-digit page reference, + 1 dot (**.**);
Q : 1 digit [1-4], *Quadrant* on the DYC page;
M : 1 letter [1-9AB], offset in Q of *Main* Seal form;
V : 1 digit [0-5], Main = 0, *Variant* > 0.

The **AO-** prefix is an addition to the **DYC.QMV** syntax described in Cook(2003).

• ***Use-Cases***

The proposed new kDYC property is well-suited for inclusion in Unihan, for the very same reasons that other primary lexical (dictionary) mapping sources are included. Dictionary mappings determine character identities, and guide developer and end-user usage. Those who wish to find the entry for a CJK character in DYC's text will be able to query this data, using CJK code points specific to their locale, and will easily find the corresponding DYC entry (or entries), in order to look up the relevant pages.

In addition to developer and end-user usage applications, this proposed new property also provides resources for on-going Seal Script encoding development. As described in the introduction of this document, Seal Script repertories differ among the various source texts. Production of a unified Seal Script encoding requires interoperability of existing implementations, to identify duplicates within each source, and to map character identities across sources. Current Unihan property data is well-augmented by inclusion of the proposed new property.

• ***References***

《說文解字·電子版》*Shuō Wén Jiě Zì — Diànzǐ Bǎn:*
Digital Recension of the Eastern Hàn Chinese Grammaticon.
Doctoral Dissertation, UC Berkeley, Dept. of Linguistics (2003);
STEDT Monograph #9, in 4 vols. (ISBN 0-944613-48-9; Dec. 2009);
<http://linguistics.berkeley.edu/~rscook/html/writing.html#EHC>
Scholarly Editions and Translations award, NEH (2010-2013)
[Part of 文林 *Wénlín 4.X <http://www.wenlin.com>; excerpts below, p. 5-7.*]

《說文解字·注》*Shuō Wén Jiě Zì — Zhù* (DYC)
〔東漢〕許慎著〔清〕段玉裁注。
上海 (瑞金二路 272 號): 上海古籍出版社, 1981.
[1988, 1989, 1998 (9th printing)]. ISBN: 7-5325-0487 5/H.6.
[Corrected/pointed reproduction based on the 經韻樓藏版
Jīng Yùn Lóu original produced by the Duàn family, 1813-1815.]

《漢語大字典》。許力以主任，徐中舒主編，（漢語大字典工作委員會）。
武漢：四川辭書出版社，湖北辭書出版社,1986-1990.
Hànyǔ Dà Zìdiǎn [HDZ; 'Great Chinese Character Dictionary' (in 8 Volumes)].
Xú Zhōngshū (Editor in Chief). Wuhan, Hubei Province (PRC):
Hubei and Sichuan Dictionary Publishing Collectives, 1986-1990.
ISBN: 7-5403-0030-2/H.16.

UAX#38 : UNICODE HAN DATABASE (UNIHAN)
<https://www.unicode.org/reports/tr38/>
[See kHanYu, kHanyuPinyin, kHDZRadBreak, kSBGY.]

UAX#44 : U-SOURCE IDEOGRAPHS
<https://www.unicode.org/reports/tr45/#Section22>
[UAX#44 DYC *Source Tag* mappings are clarified by this kDYC property.]

Unicode® 11.0.0 : Appendix F - Documentation of CJK Strokes
<http://www.unicode.org/versions/Unicode11.0.0/appF.pdf>
[See *<https://wenlin.com/cdl>*.]

2.6.3 The SWJZZ Unicode 4.0 Variant Mapping Table

Once the entire text had been typeset using the proprietary encoding of my SWJZZ font, the next task was to translate this proprietary encoding into a non-proprietary one. As I became acquainted with the IRG work under way on Ext. B, it was clear that the Unicode Standard was the only choice.[61] The great suitability of Unicode for the task I had in mind related primarily to the mapping tables developed by the IRG for Ext. B, and in particular to the 《漢語大字典》 *Hanyu Da Zidian* (HDZ) data. A major Chinese lexical source in eight volumes, HDZ had been in my home library since 1993, and I had often dreamed of having it in electronic form.[62] My work revising the IRG's HDZ mapping data, currently available through the Unicode Consortium,[63] is discussed in Chapter 3. In the current section, that work is taken for granted as we discuss the development of the primary mapping table for relating my proprietary encoding to the international standard.

2.6.3.1 Structure of the SWJZZ mapping table

The primary copy of my SWJZZ Unicode 4.0 Variant Mapping Table is a UTF-8 text document of 11,246 lines. Initially created using a combination of tools, it is now edited only using *Wenlin* 3.x software. The field structure of this mapping table is tabulated below in Table 2-36.

Table 2-36. SWJZZ Unicode Variant Mapping Table: Field Structure

CIDN	DYC	.QMV	PUAH	A	B	N	C	?D	!E	#F

61. See above, Section 2.5.1. Unicode 3.0 became available in September of 1999 [see *Unicode Consortium*, 2000, in the *Bibliography*]. The first draft of Ext. B that I saw was dated 8/22/2000. This data went public in March 2002 with Unicode 3.2. The public release of Unicode 4.0 (for CJK purposes a minor release compared to the 3.2 release) came on April 18, 2003.

62. See *Cook 1996a* for an example of my early use of HDZ.

63. See also *Cook 2001e* in the Bibliography.

The abbreviations used in Table 2-36 are explained in Table 2-37 below.

Table 2-37. SWJZZ Unicode Variant Mapping Table: Notes on Table 2-36

Abbr.	*Explanatory Notes*
CIDN	CID Number. The CIDN is a unique numeric identifier for each MV form. See the CMap in Appendix 11.2
DYC	zero-padded 3 digit page reference to DYC:A
.QMV	Q ia page quadrant in DYC:A; M is MV, and V == 0 for M and > 0 for V
PUAH	6-digit Unicode Supplemental Plane Private Use Area codepoint of an MV form, in Hexadecimal notation, beginning with [U+100000]
A	single MV form (SWJZZ font)
B	single Unicode 4.0 reference glyph, current best available match; this is the form appearing in the "UB" field in the printed SWJZ-DB concordance
N	Numeric index from 1..3, quantifying degree of MV <=> B match: "1" => perfect stroke-for-stroke correspondence; "2" imperfect correspondence; "3" non-correspondence
C	list of proper VarClass members. If 1st element (a.k.a "first non-best match") is B5 && B is non-B5, B5 form appears in SWJZ-DB concordance's "B5" field
?D	questionable form, possibly member of C or E, sometimes also derived forms (those modified with semantic determiner)
!E	list of improper (erroneous) VarClass members (not a proper variant of B/C elements, possibly confused with or used erroneously for one or more B/C elements)
#F	list of 0 or more PUA forms; if N == 2 or 3, the unencoded B form is stored as the first item in this list (for future encoding); other forms in this field may be unencoded (a) variants or (b) components (including compositional variants)

2.6.3.2 SWJZZ Unicode Variant Mapping Table: Example Lines

In Table 2-38 below are eight lines from the SWJZZ Unicode Variant Mapping Table exemplifying the distinctions mentioned in the *Explanatory Notes* given in Table 2-37 above.

Table 2-38. SWJZZ Unicode Variant Mapping Table: Lines 256..263

CIDN	DYC.QMV	PUA	MV	B	N	C?D!E#F
256	021.310	1000ff		屮	1	屯少十艸!山
257	021.410	100100		屯	1	聖坉?囤沌忳純#屯
258	021.420	100101		每	1	每嵒苺毎
259	022.110	100102		毒	2	毒嵒隼晝!毒晝晝蛀#毒
260	022.111	100103		薊	1	薊
261	022.120	100104		岑	1	岑
262	022.121	100105		芬	1	---
263	022.130	100106		先	1	先!先#先先兂

Underlying the forms appearing above in Table 2-38 are the Unicode Scalar Values listed in Table 2-39 below.

Table 2-39. USV's for the forms in Table 2-38

CIDN	DYC.QMV	PUA	MV	B	N	C?D!E#F
256	021.310	1000ff	[U+1000ff]	[U+5c6e]	1	[U+21cfe] [U+2f878] [U+5341] [U+8278] ! [U+5c71]
257	021.410	100100	[U+100100]	[U+5c6f]	1	[U+2126b] [U+5749] ? [U+56e4] [U+6c8c] [U+5ff3] [U+7d14] # [U+e277]
258	021.420	100101	[U+100101]	[U+21d0b]	1	[U+6bcf] [U+21d15] [U+23aed] [U+6bce]
259	022.110	100102	[U+100102]	[U+21d1b]	2	[U+6bd2] [U+21e46] [U+24bdf] [U+27249] ! [U+23af4] [U+45af] [U+27276] [U+86c0] # [U+e214]
260	022.111	100103	[U+100103]	[U+25cae]	1	[U+26e15]
261	022.120	100104	[U+100104]	[U+37a3]	1	[U+21d73]
262	022.121	100105	[U+100105]	[U+82ac]	1	---
263	022.130	100106	[U+100106]	[U+21d06]	1	[U+5725] ! [U+5148] # [U+e0b7] [U+e222] [U+e2a6]