漢字部品記述における複数ドメイン導入の試み

守岡 知彦 （京都大学）

1 はじめに

漢字は長い年月にわたって使われ、また、東アジアの広い地域において使われてきたため、その形状（グリフ）は時代や地域、書記媒体、用途などに応じてさまざまな書体や字体を生み出しながら変化してきた。こうした形状の差異はデザイン差や字体差（異体字関係)、あるいは、文字の差異として固定されることもあり、また、本来別字だったものが似た形状で書かれるようになって区別が付きにくくなる（別字衝突）こともある。翻刻などにおいて、漢字の形状変化や異体字関係、別字衝突の情報は漢字を同定する上で重要であり、UCS 統合漢字 [1] のような巨大な漢字集合を整理する上でもこうした情報が必要となってきているといえる。

漢字の同定作業を行う上で、対象となる漢字がどういう部品の組合せからなっており、その部品がどういう漢字（部品）に対応するかを把握することは重要なプロセスのひとつであるといえる。例えば、「肌」は「月」と「几」という部品を左右に並べた形に見えるが、この「月」は〈月〉(moon) ではなく〈肉〉に対応するものである。一方、「朧」の「月」は「肉」ではなく〈月〉(moon) に対応するものである。このような部品の変形・衝突は複数の部品を組み合わせた複合部品でも起こり得る。例えば、「癶」や「𠂉」は別起源の部品であるが、これらの部品を持った漢字ではそれぞれの左側が「𠂇」「𠂆」「夕」などに変形しやすく、また、右側はそれぞれ「𠂊」「𠂈」（又）などに変形しやすく、部品の衝突が起こるとともにその形状はこうした左右の部品のバリエーションの順列組合せや筆遣いに起因した線の連続化や省略・変形（これらを総称して『筆法的変形』と呼ぶことにする）によってさまざまな形に書かれ得る（表 1)。

漢字の符号化では字体の包摂規準によって複数の部品字体を同一視するという手法が用いられているが、こうした変化の幾つかは通常の包摂規準ではカバーできないものといえる。しかしながら、こうしたものをカバーするために、例えば、「月」と「夕」と「𠂇」を同一視するための包摂規準を追加すると衝突するケースが増えてしまい問題である。こうした問題を解決するために、CHISE [3] 文字オントロジーに対し、字体の包摂規準に基づく抽象部品に加えて、楷書における筆法的変形に基づく抽象部品を追加するとともに、従来の字体の包摂規準に基づく抽象部品も字源的・機能的な抽象部品（字源的もしくは機能的に同じことを以下では『同根』(cognate) と呼ぶことにする）と見掛け上の抽象部品を区別する試みを行っている。本稿ではこの試みについて概説する。

2 文字と部品の包摂関係と包摂規準

ある漢字字形がどのような文字を表現したものであるかを判断するには、文字の置かれた文脈等からその文字自体を直接的に同定する方法と、その漢字がどのような部品の組合せからなるかを見出すことで同定する方法がある。漢字字形を字体や字種等のカテゴリーに分類する場合も、文字単位での is-a 関係と文字が持つ部品単位での is-a 関係が考えられる（なお、以下では、こうした is-a 関係を「包摂関係」と呼ぶことにする)。

文字単位に文字を同定・分類するには文脈情報等からその文字が何であるかが判明できなければならないが、一般にはこれは成り立たず、形状に関する情報だけから同定・分類可能であることが望ましい。もし、文字単位での包摂関係と文字が持つ部品単位での包摂関係が矛盾しなければ、部品単位の包摂関係に基づいて文字の包摂関係を定義できるはずである。これが漢字字体の包摂規準の原理である。

多くの漢字は複数の部品の組合せからなっているが、漢字を部品の組合せとしてとらえた時に似た形の部品を同一視するためのルールを決めれば、比較

字種	业・业		ヲ乀		夕乀				夕又	夕又	夕乀	収
發	𭛁*91	發*7	發*19					發*36				
祭			祭*14	祭*92	𫞴*93	祭*94	祭*95	(祭)		祭*18		
際			際*10		際*7			際*11		際*18		
察			察		察*7				察*28	察*18		
登			登*19		登*33							
登										登		
鐙			鐙							𨭕*96		
癸			癸*30		癸*33							
葵			葵*26							𬝢*97	葵*98	
登									登			登

[7] HNG:賢劫經卷二 (正倉院本) [10] HNG:妙法蓮華經卷五 (今西本) [11] HNG:妙法蓮華經卷三 (守屋本)
[14] HNG:漢書楊雄傳 (上野本) [18] HNG:開成石經論語 [19] HNG:開成石經周易
[26] HNG:齊民要術卷五 (高山寺本) [28] HNG:華嚴孔目 (高山寺本) [30] HNG:法藏和尚傳 (高山寺本)
[33] HNG:日本書紀 卷二十四 (岩崎本) [36] HNG:日本書紀 卷二十四 (兼右本) [91] U+2D6C1 [92] 登記統一文字 01065000
[93] U+2B7B4 [94] 戸籍統一文字 276070 [95] 戸籍統一文字 276370 [96] U+28B55 [97] U+2C762 [98] U+28B55

表 1 「癶・癶」のバリエーション

的少数のルールの組合せによって多数の漢字を対象とした符号化文字の包摂範囲の定義が可能である。ここで、このルールのことを『包摂規準』と呼ぶ。*1

漢字字体の包摂規準はある漢字字形がどういう字義・字音の文字であるかの同定はできないが、その字形の持つ（字体粒度の）部品の組合せパターン（漢字構造）から対応する（複数の字体を包摂した）抽象文字を決めることができる。

3 包摂関係の拡張

CHISE には超抽象文字（字種）—抽象文字—字体—抽象字形—字形といった包摂粒度の概念があり、部品においても文字と同じ包摂粒度の概念が用いられている。そして、対応する荒い包摂粒度の部品と細かい包摂粒度の部品の間には部品間の包摂関係が存在する。[8]

この包摂関係に対して、同根な包摂関係と筆法的変形に基づく見掛け上の包摂関係を区別するために、後者を示すために `component` というドメイン [3] を設けた。同根な抽象部品と字体粒度の部品の包摂関係には従来のドメインなしの包摂関係素性 `<-denotational` を用い、同根でない見掛け上の抽象部品と字体粒度の部品の包摂関係にはドメイン `component` を付けた包摂関係素性 `<-denotational@component` を用いることにした。

表 2 に見掛け上の包摂関係の例を示す。例えば、「遥/遙」に見られるように「⺤」と「夕」は交換可能であり、「䍃」も「䍃」も U+4343 に対応する抽象文字に包摂される。しかしながら、「⺤」(爪) と「夕」(肉) は文字単体としては別字であるので、この両者の差異を捨象した抽象部品〈⺤/夕〉との間に見掛け上の包摂関係を示す素性 `<-denotational@component` を用い、

〈⺤/夕〉 `->denotational@component` ⺤
〈⺤/夕〉 `->denotational@component` 夕

という関係を記述する。

ただ、実際には、「⺤」は「⺥」や「爫」を包摂する抽象部品〈⺤/⺥/爫〉であり、この抽象部品と「⺤」「⺥」「爫」の間には

〈⺤/⺥/爫〉 `->denotational` ⺤
〈⺤/⺥/爫〉 `->denotational` ⺥
〈⺤/⺥/爫〉 `->denotational` 爫

という包摂関係が存在する。

*1 JIS X 0208/0213 では字形の細かなデザイン差を捨象した字体を対象にどう包摂するかを定めるようにしているので、このことを強調して『字体の包摂規準』と呼ぶ。

〈厂/厂〉	→	厂 , 厂
〈灬/火〉	→	灬 , 火
〈卜/卜〉	→	卜 , 卜
〈爫/夕〉	→	〈爫/爫/爫〉 , 夕/夕
〈歹/歹〉	→	〈歹/歹〉 , 〈歹/歹/歹〉
〈丐/丐〉	→	丐 , 丐
〈广/广〉	→	广 , 广
〈匆/匆〉	→	匆 , 匆
〈谷/谷〉	→	谷 , 谷
〈囱/囱/囱〉	→	囱 , 囱 , 囱
〈⿰/⿰/⿰〉	→	⿰ , ⿰, 〈⿰/⿰〉
〈畾/畾〉	→	畾 , 畾
〈尿/尿〉	→	尿 , 尿
〈⿱/⿱〉	→	〈⿱〉 , 〈⿱〉
〈溥/溥〉	→	〈溥/溥/溥〉 , 溥

但し、「→」は `->denotational@component` を示す。

表 2　見掛け上の包摂関係の例

よって、抽象部品〈爫/夕〉からはこの 2 つの関係素性を使った包摂関係のグラフが構成されることになる。

4　抽象文字粒度 ID 素性の拡張

従来、CHISE では抽象文字粒度よりも荒い包摂粒度として超抽象文字粒度が存在し、BUCS [5] に基づく ID 素性 `==>ucs@bucs` を設けていた。

BUCS は字源を問わず、現代の日本・中国・台湾・韓国等での文字の使われ方に基づいて同値性を定めたものといえ、歴史的には異なる文字として扱われてきたものを同一視したり、逆に、歴史的には同字根であったものを分離している場合がある。漢字構造記述における部品を整理する場合、こうした文字単位での現代の用法よりも部品としての（歴史的な用法も含めた）挙動の方が重要であるといえる。また、ここで扱う問題は、部品として同根、ないしは、同根な文字の筆法的変形であるといえ、基本的には字体粒度の差異を対象としたものと見ることができる。よって、同字根ないしは同字根の文字の部品の筆法的変形を扱うための（UCS 統合漢字に対応する抽象文字オブジェクトとは別の）抽象文字粒度のオブジェクトが記述できれば良いといえ、このためにはこれらを指示するための ID 素性が必要となる。

4.1　包摂規準に基づく抽象文字粒度の ID 素性

包摂規準に基づく抽象文字粒度のオブジェクト（ID 素性）としては、従来、JIS X 0208:1997/0213 の各符号位置の包摂範囲に基づくものとして、

`=>jis-x0208`　JIS X 0208 の抽象文字を示す ID 素性。JIS X 0208:1997 と JIS X 0213:2000/2004 で包摂範囲に変化がないものを示す。

`=>jis-x0208@1997`　JIS X 0208:1997 の抽象文字を示す ID 素性。JIS X 0208:1997 の包摂規準に基づく。

`=>jis-x0213-1`　JIS X 0213 第 1 面の抽象文字を示す ID 素性。JIS X 0213:2000/2004 で包摂範囲に変化がないものを示す。

`=>jis-x0213-1@2000`　JIS X 0213:2000 の第 1 面の抽象文字を示す ID 素性。JIS X 0213:2000 の包摂規準に基づく。

`=>jis-x0213-1@2004`　JIS X 0213:2004 の第 1 面の抽象文字を示す ID 素性。JIS X 0213:2004 の包摂規準に基づく。

`=>jis-x0213-2`　JIS X 0213 第 2 面の抽象文字を示す ID 素性。

を設けていたが、UCS 統合漢字をカバーするには不十分であり、また、漢字符号化の規準となる UCS 統合漢字自体の包摂範囲を適切に記述するためには、UCS の包摂規準を定義・形式化する必要があるといえる。

そこで、そのベースとして、UCS 統合漢字における事実上の包摂規準と考えられる IWDS-1 [*2] [2] を採用し、これに対応する抽象部品を示す抽象文字粒度の ID 素性として `=>iwds-1` を設けた。

この素性は値として自然数をとる。IWDS-1 の番号が自然数の場合、その値を用いる。

また、27a のように枝番が付いている場合、1027 のように枝番の a を 1, b を 2, ... とした時の数に 1000 をかけたものを主番号に足すことで整数値化する。

また、複数の包摂規準を結合する場合、値の小さ

[*2] IRG Working Document Series (IWDS) 1: List of UCV (Unifiable Component Variations) of Ideograhs

いものから 3 桁毎の自然数を繋げて表現することにした。例えば、連番 54 と連番 56 を結合したものは 54056 となる。また、連番 55 と連番 346 を結合したものは 55346 となる。

このように構成した素性値と素性名 `=>iwds-1` を用い、素性対

```
(=>iwds-1 .  素性値)
```

で IWDS-1 に対応する抽象部品を表現（指示）することができる。

例えば、連番 37 の場合、対応する抽象部品を

```
(=>iwds-1 .  37)
```

という素性対で示すことができる。[*3] また、連番 132a に対応する抽象部品は素性対

```
(=>iwds-1 .  1132)
```

で示すことができる。[*4] また、連番 55 と連番 346 の結合に対応する抽象部品は素性対

```
(=>iwds-1 .  55346)
```

で示すことができる。[*5]

4.2 UCS 統合漢字に対応する ID 素性

4.2.1 `=>ucs@iwds-1` 素性

UCS 統合漢字では複数の符号位置に分離されているが IWDS-1 的には包摂可能なもの[*6]に対して、その分離された符号位置に対応する複数の抽象文字を包摂する抽象文字オブジェクトを記述するための ID 素性として `=>ucs@iwds-1` を設けた。

これは IWDS-1 の抽象部品や IWDS-1 から演繹すると包摂されることになる複数の抽象文字を包摂する抽象文字を指示するためのものである。

この ID 素性を用いることにより、UCS では異なる複数の符号位置があるが IWDS-1 的には包摂されるものに対し、包摂分離されたものの中から一つ代表を取り出し、その符号位置を用いて、抽象文字粒度の文字オブジェクトを構成することができる。

例えば、〈靑〉と〈青〉は文字単体としては U+9751 と U+9752 に分離されているが、IWDS-1:319 により部品としては両者は包摂される。そこで、U+9751 の符号位置と ID 素性 `=>ucs@iwds-1` を用いて、〈靑〉と〈青〉を包摂した抽象部品〈靑/青〉を素性対

```
(=>ucs@iwds-1 .  #x9751)
```

として表現（指示）することができる。[*7]

例えば、「高」と「髙」は U+9AD8 と U+9AD9 に分離されているが、もし単純に IWDS-1:316 を適用すれば包摂可能だったはずである。このケースの場合も同様に、この両者を包摂する仮想的な抽象文字 〈高/髙〉 を素性対

```
(=>ucs@iwds-1 .  #x9AD8)
```

で表現（指示）することができる。[*8]

但し、IWDS-1 の抽象部品の内、単純な線画からなり、非同根な部品として使われやすいものに関しては 4.2.3 節で述べる `=>ucs@component` 素性を用いることにした。

例えば、「𠂇」(U+20087) と「十」(U+5341) は IWDS-1:37 により部品としては両者は包摂されるが、文字単体としては同根でなく、また、非常に単純な線画であり、多様な使われ方が予想されるため、`=>ucs@iwds-1` 素性ではなく、`=>ucs@component` 素性を用いて素性対

```
(=>ucs@component .  #x5341)
```

で表現（指示）する。[*9]

[*3] なお、これは XEmacs CHISE の S 式による表現であり、CHISE-wiki [6] (EsT [7]) ではこの素性対は `http://www.chise.org/est/view/character/a.iwds-1=37` という URL に対応する。

[*4] CHISE-wiki では `http://www.chise.org/est/view/character/a.iwds-1=1132` という URL に対応する。

[*5] CHISE-wiki では `http://www.chise.org/est/view/character/a.iwds-1=55346` という URL に対応する。

[*6] 元規格分離が適用されたものや non-cognate と判断されたもの、また、拡張漢字 B 等での作業ミスと思われるものもある。

[*7] CHISE-wiki では `http://www.chise.org/est/view/character/a.ucs@iwds-1=0x9751` という URL に対応する。

[*8] CHISE-wiki では `http://www.chise.org/est/view/character/a.ucs@iwds-1=0x9AD8` という URL に対応する。

[*9] CHISE-wiki では `http://www.chise.org/est/view/character/a.ucs@component=0x5341` という URL に対応する。

4.2.2 =>ucs@cognate 素性

IWDS-1 では包摂されず UCS 統合漢字において符号位置が分離されているが同根な抽象文字（部品）を表現するために =>ucs@cognate を設けた。

例えば、「酉」(U+9149) と「酉」(U+2E815) は同根であるので、両者を包摂した仮想的な抽象文字〈酉/酉〉 を素性対

```
(=>ucs@cognate .  #x9149)
```

で表現（指示）する。[*10]

4.2.3 =>ucs@component 素性

IWDS-1 では明示的に包摂されないが UCS 統合漢字において包摂例がある（あるいは、IVS で指示されるグリフで IWDS-1 的にはその基底文字に包摂できないものの、差異が軽微で IWDS-1 を拡張しても差し支えないと考えられる）部品を示すための抽象文字粒度の ID 素性として =>ucs@component を設けた。IWDS-1 に含まれない同根でない見掛け上の抽象部品の多くはこの素性対を持つ抽象部品オブジェクトで表現できる。

例えば、「畾」(U+55A6) と「畾」(U+5D52) は別字であるが、形状が類似しており、部品としては混同して使われる。例えば、U+27B0C の例示字形には両方の部品を用いたものが存在している（図 1）。

図 1　U+27B0C の例示字形

よって、U+27B0C の抽象文字粒度の漢字構造記述を行う場合、「畾」と「畾」を包摂した抽象部品〈畾/畾〉が必要となるが、これは素性対

```
(=>ucs@component .  #x5D52)
```

で表現（指示）することができる。[*11]

[*10] CHISE-wiki では http://www.chise.org/est/view/character/a.ucs@cognate=0x9149 という URL に対応する。

[*11] CHISE-wiki では http://www.chise.org/est/view/character/a.ucs@component=0x5D52 という URL に対応する。

4.3 外字部品の表現

UCS では表現できないものに対しては、主に Big5-CDP[*12] と GT-K[*13] および GlyphWiki [4] のグリフに対応した抽象部品オブジェクトによって表現することにした（表 3 に抽象部品オブジェクト用 ID 素性名の一覧を載せる）。

GlyphWiki ではさまざまなグリフ名の形式が用いられているが、現在の所、抽象部品オブジェクトの表現には uHHHH(H)-itaiji-ddd と cdp-HHHH-itaiji-ddd という形式のグリフ名のみを用いるようにしている。

これらは同根でない抽象部品に限らず、同根な抽象部品もこの方式で表現し、同根かどうかは 3 節で述べた包摂関係素性の種類によって表現する。

5 おわりに

漢字字体の歴史的変遷を考慮した場合、漢字構造記述における部品の整理には意符や音符のような字源的・機能的な単位での対応関係と筆法的変形に起因する変異の双方を扱う必要があるといえる。そのため、ここでは包摂関係と抽象部品（文字）の双方において両者を区別可能にするためのドメインを導入した。

包摂関係にドメインを導入し見掛け上の包摂関係を同根な包摂関係と分離したことにより、字源・機能的な観点での分類と筆法的変形に基づく分類を混在させることができた。また、抽象文字・部品の記述においても UCS の符号位置に対して、IWDS-1 に基づくもの (iwds-1), 同根なもの (cognate), 見掛け上の部品 (component) という異なるドメインを導入することにより、包摂規準と字源・機能的な観点と字形用例の整理という 3 種類の視点を包含した記述が可能になったといえる。

参考文献

[1] International Organization for Standardization (ISO). *Information technology — Universal Coded Character Set (UCS)*, 2014 年 9 月. ISO/IEC 10646:2014.

[*12] 台湾中央研究院 CDP 外字。CHISE での定義に基づく。

[*13] GT 書体の部品文字セット

表 3　抽象部品オブジェクト用 ID 素性名

素性名	素性値	説明
=>jis-x0208	16 進 (94 × 94)	JIS X 0208（共通部分）(4.1 節)
=>jis-x0208@1997	16 進 (94 × 94)	JIS X 0208:1997（4.1 節）
=>jis-x0213-1	16 進 (94 × 94)	JIS X 0213 第 1 面（4.1 節）
=>jis-x0213-1@2000	16 進 (94 × 94)	JIS X 0213:2000 第 1 面（4.1 節）
=>jis-x0213-1@2004	16 進 (94 × 94)	JIS X 0213:2004 第 1 面（4.1 節）
=>jis-x0213-2	16 進 (94 × 94)	JIS X 0213 第 2 面（4.1 節）
=>iwds-1	10 進	IWDS-1（4.1 節）
=>ucs@iwds-1	16 進 (UCS)	IWDS-1 に基づく UCS 抽象文字（4.2.1 節）
=>ucs@cognate	16 進 (UCS)	同根部品を統合した UCS 抽象文字（4.2.2 節）
=>ucs@component	16 進 (UCS)	筆法的変形を統合した UCS 抽象文字（4.2.3 節）
=>ucs-itaiji-001	16 進 (UCS)	GlyphWiki の uHHHH(H)-itaiji-001
=>ucs-itaiji-002	16 進 (UCS)	GlyphWiki の uHHHH(H)-itaiji-002
=>ucs-itaiji-003	16 進 (UCS)	GlyphWiki の uHHHH(H)-itaiji-003
=>ucs-itaiji-004	16 進 (UCS)	GlyphWiki の uHHHH(H)-itaiji-004
=>ucs-itaiji-005	16 進 (UCS)	GlyphWiki の uHHHH(H)-itaiji-005
=>ucs-itaiji-006	16 進 (UCS)	GlyphWiki の uHHHH(H)-itaiji-006
=>ucs-itaiji-007	16 進 (UCS)	GlyphWiki の uHHHH(H)-itaiji-007
=>ucs-itaiji-009	16 進 (UCS)	GlyphWiki の uHHHH(H)-itaiji-009
=>big5-cdp	16 進 (Big5 PUA)	台湾中央研究院 CDP 外字
=>big5-cdp-itaiji-001	16 進 (Big5 PUA)	GlyphWiki の cdp-HHHH-itaiji-001
=>mj	10 進	文字情報基盤
=>gt	10 進	GT
=>gt-k	10 進	GT 部品文字セット
=>daikanwa	10 進	大漢和番号
=>daikanwa/ho	10 進	大漢和番号（補巻）
=>cbeta	10 進	CBETA 外字
=>ruimoku-v6	16 進 (UCS PUA)	東洋学文献類目現行外字

[2] IRG Working Document Series. http://appsrv.cse.cuhk.edu.hk/~irg/irgwds.html.

[3] Tomohiko Morioka. Multiple-policy character annotation based on CHISE. *Journal of the Japanese Association for Digital Humanities*, Vol. 1, No. 1, pp. 86–106, 2015 年 11 月.

[4] 上地宏一. GlyphWiki. http://glyphwiki.org/wiki/GlyphWiki.

[5] 情報処理学会. 符号化文字基本集合 Basic Subset of Coded Character Sets, 2002 年. 情報処理学会 試行標準 IPSJ-TS 0005:2002.

[6] 守岡知彦. CHISE のセマンティック Wiki 化の試み. 情処研報, Vol. 2010-CH-87, No. 8, pp. 1–8, 2010 年 7 月.

[7] 守岡知彦. Wiki 的手法に基づく構造化データの編集について. 人文科学とコンピュータシンポジウム論文集 —人文工学の可能性〜異分野融合による「実質化」の方法〜, 情報処理学会シンポジウムシリーズ, 第 2010 巻, pp. 33–40. 情報処理学会, 情報処理学会, 2010 年 12 月.

[8] 守岡知彦. CHISE による HNG データ収録の試み. 石塚晴通監修, 高田智和, 馬場基, 横山詔一 (編), 漢字字体史研究 二— 字体と漢字情報, pp. 185–203. 勉誠出版, 2016 年 11 月月.