漢字字体の包摂基準の衝突評価の試み

守岡 知彦

1　はじめに

文字はさまざまな形に書かれ得るがそうした字形の差異の多くは文字の区別には寄与しないため、文字の符号化においては、そうした文字の弁別という観点では意味のない差異を捨象し、同じ文字を指しているであろうものを同じカテゴリーにまとめるためのモデルや仕組みが必要であるといえる。特に、漢字は文字数が多いため、その必要性が高いといえるが、その半面、長期にわたって広い地域で使われてきたため、文字の同値性やどのような差異が意味のある差異かということが必ずしも明らかではないといえる。このことは、異体字の整理や異体字関係の確定の困難さの一因であるといえよう。

漢字は伝統的に形音義の組合せからなるものと考えられてきたが、UCS をはじめとする現在の一般的な文字符号では主に形に着目して（字音や字義の情報を捨象して）抽象的な形状を符号化対象とするアプローチがとられている。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とき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]。

しかしながら、包摂規準に基づく抽象形状の定義だけでは「大」と「犬」のように形状は似ていても字音や字義が全く異なる字を区別することができず問題である。だからといって、こうしたケースにおいて単純にこの差異を包摂しないことにした場合、「類/類」のように部品として含む場合にこの両者の差異が字音・字義の差異を生じないケースが無数に存在するため問題である。このため、UCS における漢字の統合に関する原則を説明した ISO/IEC 10646 附属書 S では、歴史的に区別されてきた類似の文字の組を *non-cognate characters*（別字源の文字）として区別することにしている。しかしながら、non-cognate かどうかという判断は漢字学の知識無しには判断できないものといえ、また、もともと別字だが歴史的に混同されてものや、もともと同字だったものが現在では別字と見なされるようになったものなど、同時・別字の解釈が国・地域によって異なるものなどをどうするかといった問題に対して non-cognate のものは分離するという原則だけでは対処できないといえる。また、字音や字義が同様な同字源・同一字種に属する形状が似た異体字で

[*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。

あっても歴史的事情から別の符号位置に分離されてきたものもあり、無知識的に単純なルールだけで過不足なく記述することはできない。実際に初めて網羅的な包摂規準の集合を明示した JIS X 0208:1997 では、すでに別々の符号位置が与えられ区別されているものを包摂するような包摂規準の適用を禁止し、こうしたケースを『適用除外』としている。

文字符号から見た場合、包摂 (unification) とは「複数の字体を区別せずに、それらに同一の区点位置[*2]を与えること」[2] であり、包摂規準はその適用除外を除けば、抽象文字の同値性やコードポイントの指示対象を定めたものといえ、文字符号にとってはその定義の根幹に位置すべき所与のものといえる。いいかえれば、包摂規準は文字符号化における品質や性能を規定するものといえ、なるべく少ない簡潔なルール（そして、なるべく少ない適用除外で）によってなるべく多数の漢字を網羅的かつ斉一に別字が衝突することなく定義されていることが望ましいといえる。そして、このような観点で包摂規準の性能評価を行うための手法や指標が必要であるといえる。

しかしながら、包摂規準は、現状、人間が見て判断することを想定して作られているといえ、従来はその理論的な評価が必ずしも容易ではなかった。しかしながら、[4] [6] により、項書き換え系 (Term Rewriting System; TRS) の理論に基づき包摂規準を等式論理に基づいて解釈することで形式化する試みが提案され、こうした手法に基づき形式化（完備化）された包摂規準の集合を用いて CHISE 漢字構造情報データベースを処理することにより、包摂規準の定量的な評価が可能となった。

包摂規準を似た形の異体字をまとめるための仕組みととらえた場合、パターン認識や自然言語処理、情報検索などと同様に、その評価指標として F 値を用いることが考えられる。しかしながら、現状、CHISE 漢字構造情報データベースの品質の問題もあり、その適合率や再現率を完全に求めることができない。このため、今回は別字衝突の問題に限定して分析を試みる。また、本来は複数の包摂規準の比較を行うべきであるが、今回は UCS 統合漢字における事実上の包摂規準といえる IRG Working Document Series (IWDS) 1: List of UCV (Unifiable Component Variations) of Ideograhs [1] のみをその対象とした。

2　包摂規準の書き換え規則化

適用除外を無視すれば、包摂規準は漢字構造を記述したもの（たとえば IDS）の構文木の部分木に対する書き換え規則と見なすことができる。そして、漢字構造記述を項と見なすと、包摂規準を用いて漢字構造記述を簡約化する項書き換え系を考えることができる。

包摂規準を項書き換え系の書き換え規則に変換する方法としては、包摂規準の中で同一視されるものとして列挙されている各パターンのうち、その 1 つを代表パターン（部品）とし、それ以外を異体パターン（部品）として、異体パターンを代表パターンに書き換える規則と見なす方法が考えられる。こうして異体パターンを含む漢字構造記述を代表パターンからなる漢字構造記述に正規化するわけである。[*3]

[*2] コードポイント

[*3] [3] では CHISE 漢字構造情報データベース [5] を用いた IDS の正規化アルゴリズムとそれによる漢字の同一性の

もう 1 つの方法としては、包摂規準に対応する抽象的なパターンを表現する項を設け、包摂規準の中で同一視されるものとして列挙されている各パターンからこの抽象パターンへの書き換え規則とする方法である。

包摂規準は形式的には同一視される部品字体（パターン）を列挙したものであり、それらの等式と見なすこともできるが、意味的には列挙された字体粒度のパターンがその包摂規準で示される抽象文字粒度のパターンに包摂されることを示したものといえ、この観点では後者の方が自然といえる。また、前者の場合、停止性のない書き換え規則が生じやすいが、後者ではその問題が起こらない（一度、書き換えた箇所は抽象字体粒度の部品（パターン）になっており、各書き換え規則の左辺は字体粒度であるため、マッチしない)。こうしたことを考えて、ここでは後者の方法を採ることにする。

たとえば、JIS の包摂規準の

1　王 壬 壬

はこの 3 つの部品字体を包摂した抽象部品を J_1 とすると

- 王 → J_1
- 壬 → J_1
- 壬 → J_1

という 3 つの書き換え規則で表現することができる。

180　

のように置かれる場所が指定されている場合、

- ⿺麥 x → $J_{180}(x)$
- ⿸麥 x → $J_{180}(x)$

のように IDC と変数を含んだ項を用いて表現することができる。

179　

の場合、

- ⿰厂⿱ xy → ⿰ J_{179} ⿱ xy
- ⿱⿰厂 xy → ⿰ J_{179} ⿱ xy

のように漢字構造の差として解釈するか、あるいは、

- 厂 → $J_{179'}$

チェック手法を提案している。

- 𠂉 → $J_{179'}$

のように部品のバリエーションとして解釈するかという問題があるが、ここでは、後者ののような部品バリエーションは字体差ではないと考え、前者として解釈することにする。

項書き換え系は書き換えの方向性を無視すれば等式を意味しているから、→を= と読み代えると、どこかで共通した部品字体を持つ包摂規準は同じ抽象部品に縮退する。しかしながら、項書き換え系では書き換え方向を一方向に限定しているため、書き換えの順番によっては異なる正規形(これ以上書き換えられない項）になってしまうことがある。そこで、本来等式としては同じ正規形になるべきものがちゃんと同じになるように足りない書き換え規則を補ってやれば良いといえる(言い替えれば、書き換え規則の集合の表現力が等式の集合と同じになるようにしてやれば良い)。これが完備化の直感的な説明である。

ただ、包摂規準を項書き換え系としてとらえた形式化は元々の包摂規準の想定とは異なる結果を導き得る。例えば、

82　艹 艹 艹
83　茲 茲 茲

の場合、

茲 = ⿱艹⿰幺幺
茲 = ⿱艹⿰幺幺
茲 = ⿰玄玄

であるので、完備化すれば、

- 艹 → J_{82} , 艹 → J_{82} , 艹 → J_{82}
- ⿱艹⿰幺幺 → J_{83}
- ⿱ J_{82} ⿰幺幺 → J_{83}
- ⿰玄玄 → J_{83}

となる。この場合、元々の JIS の包摂規準の規定では部分字体に対する複数の包摂規準の順次適用が禁止されているため、「茲 = ⿱艹⿰幺幺」や「⿱艹⿰幺幺」は包摂規準 83 に対応する抽象部品に包摂されないはずであるが、この完備化された項書き換え系では J_{83} に包摂されてしまう。しかしながら、現実には「滋」(戸籍統一文字 202160）や

 (HNG: 大般涅槃經卷十一 (S81)-532)

のような用例が存在する一方、これらの字体差を文脈自由的に別字として使い分けている例が見当たらないため、この場合、むしろ包摂された方が良いといえる。

しかしながら、IWDS-1 の場合、連番 22, 54, 55, 56, 99, 104, 117, 119, 203, 214, 221, 222, 236, 249, 250, 346 の包摂規準が等式としてつながれることにより、「几/儿/丷/八/丸/凡/卂/丌/

兀/尤/尢/九/尢」などが同一視されることになる。また、「兀」は「⿱一儿」と分解できるので、これらの上に「一」を載せた字体もこの完備化された包摂規準によって包摂される。例えば、「无」は「⿱一尢」と分解できるため、この完備化された包摂規準で包摂可能である。即ち、上に「一」を n 個載せたものも包摂できるということが判るが、これは素直には首肯しがたいかも知れない。また、同様に、「凡」は「凡」が「⿰几丶」と分解できることからこの完備化された包摂規準で包摂可能である。このことから、中に「丶」を n 個入れたものも包摂できることになる（「凡」は中が「一」の字体も包摂できるので、実際には中が「一」でも良いし、「几」と交差する形状でも良い）。即ち、上に「一」を n 個、中に「丶/一」を n 個入れたものも包摂するという結論になるがこれはやや直観に反するかも知れない。

3 IWDS-1 での衝突評価

IWDS-1 を形式化（完備化）したものを CHISE 漢字構造情報データベースに対して適用し、その全ての正規形を求め、同じ正規形になった統合漢字を調査した。その結果、漢字構造記述を持つ UCS 統合漢字 86527 文字に対し、約 1500 件の衝突を検出した。

CHISE 漢字構造情報データベースにおける記述ミスが存在する可能性があるのと、今回、「旗」のように複数の分解戦略がある例における正規化処理を行っていないため、この数は正確なものではないと考えられるがある程度の目安にはなると思われる。

この内、約 1/4 にあたる 379 件を調査した所、包摂される UCS 統合漢字の全てが異体字関係になると思われるものが 222 件、不明なものが 48 件、別字衝突を含むものが 109 件あった。

別字衝突に含まれる包摂規準の内、異体字と別字衝突の用例数が 5 以上のものを適合率の低い順に並べると

IWDS-1:0114 (且/旦) 異体字=1, 別字衝突=12 ; 適合率=7.7%
IWDS-1:83+182 (田/毋/毌/母) 異体字=2, 別字衝突=10 ; 適合率=17%
IWDS-1:0186 (柬/東) 異体字=2, 別字衝突=4 ; 適合率=33%
IWDS-1:20+21+53 (夕/⺼/月/夕) 異体字=2, 別字衝突=3 ; 適合率=40%
IWDS-1:0037 (𠂇/十) 異体字=8, 別字衝突=10 ; 適合率=44%
IWDS-1:22+54+55+56+99+104+117+119+203+214+221+222+236+249+250+346 (几/儿/丷/八/丸/凡/⺄/丌/兀/尤/尢/九/尢) 異体字=23, 別字衝突=37 ; 適合率=49%
IWDS-1:1+252 (壬/壬/王/玉) 異体字=11, 別字衝突=9 ; 適合率=55%
IWDS-1:0239 (大/犬) 異体字=5, 別字衝突=4 ; 適合率=56%
IWDS-1:0096 (屮/山) 異体字=7, 別字衝突=5 ; 適合率=58%
IWDS-1:0118 (⺼/日/曰) 異体字=7, 別字衝突=5 ; 適合率=58%
IWDS-1:0246 (豕/豕) 異体字=3, 別字衝突=2 ; 適合率=60%
IWDS-1:0152 (艹/廾) 異体字=14, 別字衝突=8 ; 適合率=64%
IWDS-1:0305 (⿱・⿴) 異体字=111, 別字衝突=57 ; 適合率=66%

IWDS-1:0307 (⿴・⿺) 異体字=178, 別字衝突=88 ; 適合率=67%
IWDS-1:120+121+122+123 (已/⺃/己/巳) 異体字=6, 別字衝突=3 ; 適合率=67%
IWDS-1:72+312 (土/⺡/土) 異体字=14, 別字衝突=7 ; 適合率=67%
IWDS-1:220+220a (⺕/⺠/⺟) 異体字=6, 別字衝突=2 ; 適合率=71%
IWDS-1:189,243,331 (免/兎/兔) 異体字=5, 別字衝突=2 ; 適合率=71%
IWDS-1:0180 (⺶/开) 異体字=4, 別字衝突=1 ; 適合率=80%
IWDS-1:0181 (并/幷) 異体字=6, 別字衝突=1 ; 適合率=86%
IWDS-1:199+355+356+357 (艮/⾁/⾋/⾒) 異体字=8, 別字衝突=1 ; 適合率=89%

となる。

まず気づくことは、IWDS-1:22+54+55+... (凡/...) や IWDS-1:120+121+122+123 (已/⺃/己/巳) のような複数の包摂規準がつながってしまい 4 つ以上の字を部品として包摂するものの適合率が直観に反してそれほど低くないことである。これは筆による書写では書き分けが難しいものはなるべく衝突しないようにしてきたからではないかと思われる。また、部品がおかれる位置の偏りを利用しているのかも知れない。

その一方で、IWDS-1:0114 (且/旦) や IWDS-1:83+182 (田/⺭/⺲/母) は適合率が低い。前者に関しては、IWDS-1 においても包摂分離されている例が多く、互換漢字が 1 例あるのみである。もしかすると「虐」のように比較的大きな単位での包摂規準に置き換えるべきかも知れない。後者では「田」と「母」を同一視してしまったことが悪影響を及ぼしていると考えられる。即ち、「母」と「毋」や「田」と「⺲」を同一視するのは問題が少なく、「母」と「⺲」を同一視するのもそれほど問題ではないといえるが、「⺲」を介して「田」と「母」がつながってしまうと問題が生じてしまうと考えられる。このことを鑑みれば、IWDS-1:0182 から「⺲」を削り、「母」と「毋」を同一視する条件を限定する（例えば、「貫」のケースに限る）ことが考えられる。

4 おわりに

包摂規準を等式として解釈した場合に生じる別字衝突の数と異体字の数を調査し、包摂規準の適合率を求めることにより包摂規準の品質評価を試みた。

今回の試みでは IWDS-1 を対象に CHISE 漢字構造情報データベースを用いて実験を行ったが、CHISE 漢字構造情報データベースの品質の問題から極めて限られた評価しか行うことができなかったが、一応の傾向性を調べることはできたといえる。

今後の課題としては、全てのデータを対象に、適合率だけでなく再現率も求め、F 値を求めることが挙げられる。また、IWDS-1 だけでなく JIS X 0213 の包摂規準に対しても評価を行い、異なる包摂規準間の比較を行うことも重要であると考えられる。

参考文献

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

[2] 7 ビット及び 8 ビットの 2 バイト情報交換用符号化漢字集合. 日本工業規格, 1997 年 1 月. JIS X 0208:1997.

[3] 川幡太一. IDS による UCS 漢字の「同一性」の判定手法. 東洋学へのコンピューター利用 第 17 回研究セミナー, 全国文献・情報センター人文社会科学学術セミナーシリーズ、京都大学学術情報メディアセンター 第 78 回研究セミナー, pp. 105–119, 2006 年 3 月.

[4] 守岡知彦. 項書き換え系を用いた漢字字体の包摂規準の形式化の試み. 情報処理学会論文誌, Vol. 59, No. 2, pp. 332–340, 2018 年 2 月.

[5] 守岡知彦, クリスティアン・ウィッテルン. 文字データベースに基づく文字オブジェクト技術の構築. 情報処理振興事業協会 平成 13 年度 成果報告集. 情報処理振興事業協会, 2002 年. http://www.ipa.go.jp/NBP/13nendo/reports/explorat/charadb/charadb.pdf.

[6] 白須裕之. 漢字構造の代数的記述についての予備的考察. 東洋学へのコンピューター利用 第 30 回研究セミナー, pp. 129–138, 2019 年 3 月.