情報処理学会論文誌 Vol.59 No.2 332–340 (Feb. 2018)

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

守岡 知彦[1,a)]

受付日 2017年5月10日, 採録日 2017年11月7日

概要：項書き換え系を用いた漢字の包摂規準の形式化手法を提案する．漢字の包摂規準は本質的に木構造のパターンに対する書き換えとして記述されているため，項書き換え系における書き換え規則として表現することは容易である．また，完備化アルゴリズムを用いることで計算機にとってより扱いやすい形に変換することができる．しかしながら，包摂除外をはじめとする包摂規準の例外や定義の不完全さ，符号化された漢字レパートリの不斉一さといった問題を扱うためには文字単位の包摂関係と包摂規準に基づく部品間の包摂関係の双方でダブルチェック可能な手法が望ましい．そこで，完備な包摂記述という概念をあわせて提案する．

キーワード：漢字，包摂規準，IDS，項書き換え系

An Attempt to Formalize Unification Rules of Chinese Characters Based on Term Rewriting System

Tomohiko Morioka[1,a)]

Received: May 10, 2017, Accepted: November 7, 2017

Abstract: We propose a formalization method of unification rules of Chinese characters based on term rewriting system. Since unification rules of Chinese characters are essentially described as rewriting for tree structures, it is easy to express as a set of rewrite rules in a term rewriting system. By using completion algorithms, the term rewriting system can be transformed into a more manageable form for the computer. However, in order to deal with problems of unification rules and coded character set such as exception of unification rules, incompleteness of definitions, asymmetry of unification coverage, it is desirable to use a double checkable method to compare unifiablity of characters and components based on unification rules. Therefore, we also propose a concept of description of complete subsumption.

Keywords: Chinese characters, unification rules, IDS, term rewriting system

1. はじめに

符号化された文字の指示対象が明確であること（任意の漢字が符号化可能か判定可能であること．符号化可能である場合にどのように符号化されるか（一意に）決定可能であること．任意の 2 つの符号化文字の同値性が決定可能であること）は計算機における文字処理や電子テキストのセマンティクスを確定するうえできわめて重要であるが，漢字においてはその数の多さやその使われ方・解釈の歴史的・地域的変遷などによって必ずしも自明ではなく，検索やデータ処理の際の挙動の揺れの要因となったり，ある字形が既存の符号化文字に包摂されているかが分かりにくいためにいたずらに外字を作成してしまったり，あるいは，UCS [1] における重複したり包摂域が重なった符号化文字の増加といった結果を招いてきた．文字符号の標準化においてこうした事態は望ましいことではないので，符号化された漢字の指示対象を明確化するための努力が講じられてきた．このためには漢字の性質を適切に表現した計算可能なモデルを作成する必要があり，東アジアの文字資料を対象とした人文情報学における重要な課題の 1 つであったといえる．

[1] 京都大学人文科学研究所
Institute for Research in Humanities, Kyoto University, Kyoto 606–8265, Japan
[a)] tomo@zinbun.kyoto-u.ac.jp

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]．

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るが，漢字がどういう部品の組合せからなっているかを示した機械可読な情報（漢字構造記述）が存在すれば，包摂規準を機械可読化することによって，（包摂規準の集合によって規定される包摂ポリシに基づく）漢字の同値性を検証することが可能になると考えられる．実際，文献 [2] では「CHISE 漢字構造情報データベース」[3] を用いた IDS 形式の漢字構造記述の正規化アルゴリズムとそれに基づく漢字の同一性のチェック手法を提案しており，UCS に追加提案される漢字の標準化作業を行っている IRG（Ideographic Rapporteur Group; ISO/IEC JTC1/SC2/WG2/IRG）における新規提案漢字の重複チェックに用いられている．このチェッカは大変有用なものであるといえるが，ヒューリスティックな手法に基づいていること，音符や意符といった意味的・機能的部品を無視し，『見た目』に基づく分解と IDS 文字列の記述のみに基づいて処理すること，字体や抽象文字といった包摂粒度の概念なしにアルゴリズムが定義されていることなどから，すべての重複が検出できること（完全性）や重複と判定されたものが本当に重複していること（健全性）は保証されておらず，重複候補を提案するだけのアルゴリズムとなっている．歴史的な漢字字形のバリエーションの豊富さやその変容・生成メカニズムの複雑さを考えれば，完全なチェッカを作ることは容易ではないといえるし，また，その前提となる漢字構造記述や包摂規準の集合を完全に記述しきることも難しい問題である．しかしながら，何が書けていて何が書けていないのか，何ができて何ができないのかを情報科学的にきちんと把握することが重要であるといえ，このためには経験則の理論化・形式化を行っていく必要があるといえる．

この問題を解決する 1 つの方法は数理論理学の知見を応用して，形式言語としての包摂規準の性質を明らかにすることであろう．すなわち，完全性・健全性の存在する論理体系の上で包摂規準や重複チェック手続きを論理式として記述することによって，こうしたものが情報科学的にどのようなもの・ことを扱っているかを厳密に把握するわけである．包摂規準は部品（パターン）の同値性を示したものであるので，等式論理を用いることは自然であり，等式論理に基づく計算モデルである項書き換え系も包摂規準の定義に近く親和性が高いと考えられる．そこで，本論文では，包摂規準の集合を項書き換え系における書き換え規則に変換することを試みた．

本論文で取り組んだもう 1 つの問題は包摂規準と包摂粒度の関係の明確化である．そもそも，包摂規準の集合は字体と（複数の字体を包摂した）抽象文字の関係を与えるものであるが，その定義において陽に包摂粒度の概念は現れない．本論文では，抽象文字とそれに包摂される字体（あるいは，字形），および，それらの漢字構造記述間の関係を項書き換え系に基づいて扱うことで，包摂規準および項書き換え系における完備化の観点から包摂粒度の概念を導出することを試みた．

2. 項書き換え系と完備化

項書き換え系（Term Rewriting System; TRS）は等式論理に基づく計算モデルの 1 つで，理論計算機科学の一分野として研究されており，定理自動証明や関数型・論理型プログラミング言語，代数的仕様記述といった分野で用いられている [4]．項書き換え系は左辺を右辺に書き換える規則（書き換え規則）の集合として定義される．この書き換え規則は等式に方向性を付けたものといえ，等式の世界と計算の世界を自然な形で結び付けることが可能となる [5]．

項書き換え系では与えられた項に対して書き換え規則を適用して書き換えることによって計算を表現する．そして，もうこれ以上書き換え規則を適用することができない状態に至ったときの項を『正規形』（normal form）と呼ぶ．

一般に項に適用可能な書き換え規則および書き換え可能な場所は複数存在しうるので，そのいずれを選ぶかは書き換え戦略に依存する．一般に正規形は書き換え戦略に依存して異なった結果になりうるが，どのような書き換え戦略をとったとしても正規形が一意に決まるとき，項書き換え系は『合流性』（confluent）を持っているという．また，任意の項 a，b，c が $a \to b$，$a \to c$ のように書き換えられるときに，b と c に書き換え規則を複数回適用して同じ項を得られる場合，『弱合流性』（weakly confluent）を持っているという．

$a \to b$，$b \to c$，$c \to a$ のように永遠に書き換え続けられるような規則の集合を考えることができるため，項書き換え系は一般に停止性は保証されていないが，ある項書き換え系が必ず停止するとき，その項書き換え系は『停止性』（terminating）を持っているという．また，停止性と合流性を満たしている場合，『完備』（complete）であるという．

項書き換え系が停止性と弱合流性を持つ場合，合流性を持つことが証明されているため，停止性と弱合流性を満た

[*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．

せば完備であるといえる．

ある項が書き換え可能なときに，その書き換え対象となる部分の部分項も書き換え可能なとき，この 2 つの書き換えを『危険対』（critical pairs）と呼ぶ．項書き換え系が弱合流性を持つための必要十分条件はすべての危険対に対してその片割れどうしに書き換え規則を複数回適用して同じ項を得られることである．よって，同じ項に収束しない危険対を洗い出して，それらが同じものになるような書き換え規則を付け足してやれば項書き換え系に合流性をもたらすことができるといえる．これが完備化手続きの考え方である．

完備化手続きとしては，Knuth-Bendix 完備化手続き [6] やいくつかのアルゴリズムが提案されている*2．

3. 漢字構造記述

多くの漢字は偏と旁などの部品の組合せによって構成されている．こうした漢字の部品の組み合わせ方のことを『漢字構造』と呼ぶことにする [7]．漢字構造の表現法としていくつかの形式が提案され利用されてきたが [3]，現在では，ISO/IEC 10646 [1] の一部として標準化された Ideographic Description Sequence（IDS）という形式が普及している．

漢字構造は部品の組み合わせ方を示す演算子と部品からなる構文木で表現できる．IDS は演算子として IDC（Ideographic Description Characters），部品として UCS に収録された漢字，部品用文字，および IDS を用いたものであるが，部品としてそれ以外のものを用いることも原理的には可能である．

IDS は Lisp の S 式と同様な前置記法の一種であり，IDC の後に後置される部品の数が IDC の種類ごとに決まっているために括弧なしで表現できる．IDS は IDC を関数記号，部品として使用される漢字・部品用文字を定数とした項と見なすことができる．

4. 包摂規準とは

一般に文字符号の符号位置はある範囲の字形の集合を表現しているが，その包摂範囲は包摂規準と呼ばれるルールの集合によって表現される．

こうした包摂の概念は ISO/IEC 10646-1:1993 で理念的に示されていたが*3，JIS X 0208:1997 では実際に網羅的な包摂規準の集合を示し，各符号位置のセマンティクスを明確化した．

包摂規準は，通常，同一視される部品字体（字形）を列挙する形で表現される．また，その包摂規準の例外となる適用除外文字が列挙される．この 2 つが包摂規準を定義する要素といえるが*4，このほかに適用文字例も示されている．

同一視される部品字体（字形）は，実際には，字体（字形）だけでなく，漢字構造のパターンで示されるものもある．

JIS X 0208:1997 および JIS X 0213 の包摂規準（以下では，『JIS の包摂規準』と呼ぶことにする）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．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 [8] も同様にすでに符号化された漢字に対して遡及的に適用されるものではない*5．また，JIS の包摂規準はそれぞれの符号化文字集合に収録された漢字に対して網羅的に説明可能なものになっている（少なくとも，そうなることを意図して作られている）が，IWDS 1 は現状そうなっていない．これは，CHISE 漢字構造情報データベース*6が公開され，IDS を用いた機械的なチェックが可能になる前は，手作業だけで多量の提案漢字をチェックしていたために，各符号位置の包摂範囲を斉一な包摂規準に基づいたものにすることができなかったからだといえる．JIS X 0208 も本質的には同様な問題をかかえていたが，6,000 字程度と数が少なく，JIS X 0208:1997 の包摂規準はそれまでの JIS X 0208 の変遷を後づけで説明したものとはいえ，29 個の例外的な包摂規準（過去の規格との互換性を維持するための包摂規準）を立てるだけで済んでいる．一方，UCS の場合，IDS を用いた機械的なチェックが可能になる前に約 7 万字の符号化を行っており，その結果，包摂すべきではない例示字形を持つ漢字を誤って過剰に包摂（統合）してしまったと思われるものや重複していると思われるものが少なくない数発生しており，このすべてを網羅的に説明可能な包摂規準を定義することは容易ではないといえる．

1 つの漢字（字体）に複数の包摂規準を適用することは原理的に可能であり，また，実際，同時に複数の包摂規準を適用することで包摂範囲を説明しているが，JIS の包摂規準では 1 つの部分字体に複数の包摂規準を順次適用することを禁止している．たとえば，「滋」（=⿰氵⿱艹幺）は JIS X 0213 の 1-28-02（滋）の例示字形に類似しており，この抽象文字に包摂されることが期待され，もし，JIS 包摂 83 を適用して玆と茲を包摂したうえで，茲の草冠に JIS 包摂 82 を適用して 3 画の草冠「艹」と 4 画の草冠「⺾」を同一視できれば，この字形を 1-28-02 に包摂することができるわけであるが，これでは部分字体「茲」に複数の包摂規準

*2 文献 [4] の 7.6 節でその概略が述べられている．
*3 ISO/IEC 10646:2000 以降では附属書 S で示されている．
*4 この 2 つの要素を用いて，ある符号位置に対応する例示字体（字形）に対し，それが適用除外文字でなければ，その例示字体（字形）中の部品を同一視される他の字体（字形）に置き換えたものもその符号位置に対応するものとするわけである．
*5 UCS に提案された漢字がすでに符号化されたものと重複していないかをチェックするための目安として用いられている．
*6 http://git.chise.org/gitweb/?p=chise/ids.git;a=tree

を順次適用したことになり，少なくとも，この規定を形式的に解釈すれば包摂できないことになる．複数の包摂規準の順次適用の禁止は停止性の問題を回避するための制約といえるが[*7]，互いに重なる部分を持つ包摂規準が多数存在し，その場合の優先順位も特に示されていないため，包摂規準の適用をいたずらに難しくしているきらいがあり，包摂範囲の明示という観点では問題があるといえる．また，次章のように包摂規準を書き換え規則と見なしたときに，項書き換え系としては適用可能な書き換え規則が適用できないケースが生じるため，そのままでは完備な項書き換え系にすることが困難であるといえる[*8]．

5. 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[*9]．

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（1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）．こうしたことを考慮して，ここでは後者の方法をとることにする．

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 → $J_{179'}$
- ㇒ → $J_{179'}$

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．

6. 包摂規準の完備化

包摂規準を書き換え規則と見なして項書き換え系を構成した場合，一般には停止性や合流性が保証されておらず，実際，JISの包摂規準やIWDS 1はそのままでは合流性が存在しない．そこで，包摂規準の完備化を行う．

たとえば，JISの包摂規準の場合，

54 ヨ 彐

55 彐 ヨ

167 彐 ⺕

168 ⺕ 彐

があり，これを

ヨ → J_{54} , 彐 → J_{54}

彐 → J_{55} , ヨ → J_{55}

彐 → J_{167} , ⺕ → J_{167}

⺕ → J_{168} , 彐 → J_{168}

という書き換え規則にした場合，たとえば，「録」＝「□金□彐氺」はその部分字体「彐」の J_{55} への書き換えと J_{167} への書き換えがともに適用可能であり，合流性を満たさない．

そこでこの書き換え規則の集合を完備化する．このケースの場合，問題となるのは

(1) ヨ → J_{54} / J_{55}

(2) 彐 → J_{54} / J_{168}

(3) 彐 → J_{55} / J_{167}

(4) ⺕ → J_{167} / J_{168}

---

*7 最左最外戦略をとれば停止性を持たない項書き換え系においても無限ループに陥ることはないが，計算効率は良くない．

*8 複数の包摂規準の順次適用の禁止を適用した場合と等価な包摂範囲を実現するような書き換え規則の集合を構成することも考えられるが，本論文ではこの問題については扱わない．

*9 文献 [2] ではCHISE漢字構造情報データベース [3] 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チェック手法を提案している．

という 4 つの危険対である．(1) を解消するには J_{54} と J_{55} から同じ項に書き換える規則を追加すればよい．たとえば，

$J_{54} \to J_{54+55}$, $J_{55} \to J_{54+55}$

という 2 つの書き換え規則を足すことにする．同様に，(2) を解消するために，

$J_{54} \to J_{54+168}$, $J_{168} \to J_{54+168}$

を足し，(3) を解消するために

$J_{55} \to J_{55+167}$, $J_{167} \to J_{55+167}$

を足し，(4) を解消するために

$J_{167} \to J_{167+168}$, $J_{168} \to J_{167+168}$

を足すと，今度は

- $J_{54} \to J_{54+55}$ / J_{54+168}
- $J_{55} \to J_{54+55}$ / J_{55+167}
- $J_{167} \to J_{55+167}$ / $J_{167+168}$
- $J_{168} \to J_{54+168}$ / $J_{167+168}$

という危険対が生じるので同様に書き換え規則を足していくと最終的にすべてを同じ抽象部品 $J_{54+55+167+168}$ に書き換えることで完備化が達成される．すなわち，

(1) ヨ → $J_{54+55+167+168}$
(2) ⺕ → $J_{54+55+167+168}$
(3) ⺔ → $J_{54+55+167+168}$
(4) 彑 → $J_{54+55+167+168}$

となる．

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わけである．

ちなみに，JIS X 0208:1997 で規定された 186 個[*10]の包摂規準のうち，部品未満の筆画パターンである連番 102, 127 を除く 184 個の包摂規準を書き換え規則化して完備化を試みたところ，連番 12 と 22，連番 37 と 38，連番 52 と 53，連番 54 と 55 と 167 と 168，連番 61 と 62，連番 67 と 70，連番 77 と 78，連番 111 と 153 と 154，連番 171 と 172 の組が縮退し，184 個の包摂規準が 172 個の抽象部品に縮退した．

あるいは，

82 艹 艹 卝

83 兹 茲 玆

の場合，

兹 = ⿱丷⿰幺幺
茲 = ⿱艹⿰幺幺
玆 = ⿰玄玄

であるので，完備化すれば，

- 艹 → J_{82} , 艹 → J_{82} , 卝 → J_{82}
- ⿱丷⿰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「⿱卝⿰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：大般涅槃經巻十一 (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よいといえる．

7. 文字と部品の包摂関係の対応

7.1 完備な包摂記述

包摂規準の集合から完備化された項書き換え系を構成したとき，左辺に変数を含まない書き換え規則は字体粒度の部品を抽象文字粒度の部品に書き換えるものと見なすことができる．たとえば，6 章で述べた完備化された書き換え規則の集合

- ヨ → $J_{54+55+167+168}$
- ⺕ → $J_{54+55+167+168}$
- ⺔ → $J_{54+55+167+168}$
- 彑 → $J_{54+55+167+168}$

は字体粒度の部品「ヨ」「⺕」「⺔」「彑」を抽象文字粒度の部品 $J_{54+55+167+168}$ に書き換える項書き換え系である．言い換えれば，これは部品の包摂関係を表現したものといえ，部品の包摂関係の情報があればそこから書き換え規則を生成できることを意味している．左辺に変数を含む場合，場所に依存した派生部品と見なして場所に関する情報を部品オブジェクトの文字素性（character feature）[9] として表現すれば同様の扱いが可能である．

このように，字体粒度の部品と抽象文字粒度の部品という部品の包摂関係が存在するとき，その部品を含む漢字字体の漢字構造記述は項書き換え系によって抽象文字粒度の部品を含む正規形が導かれる．たとえば，

- 艹 → J_{82} , 艹 → J_{82} , 卝 → J_{82}
- 少 → J_{136} , 少 → J_{136}

があるとき，「蔵」（=⿱艹⿰⿱少隹戈）と「蔵」（=⿱艹⿰⿱少隹戈）の正規形は⿱ J_{82} ⿰⿱ J_{136} 隹戈となる．

一方，その漢字字体に対応する抽象文字にも漢字構造記述があるとするならば，それは漢字字体の漢字構造記述から導かれた正規形と一致していることが期待される（ただし，ここでは IDS を漢字に組み上げたものは等価であるとする．たとえば，「⿰イヒ」と「化」は等価と見なす）．

たとえば，抽象文字〈蔵〉（U+229F5）の漢字構造記述を⿱ J_{82}〈戯〉とする．ここで，〈戯〉は U+229F5 の抽象文

*10 連番は 185 までだが，正誤表で 152-1 が追加されている．なお，この 152-1 は JIS X 0213 では連番 186 となっている．

字で，その漢字構造記述を⿰〈雀〉戈とする．ここで，〈雀〉は U+96C0 の抽象文字で，その抽象文字粒度の漢字構造記述を⿱ J_{136} 隹とする．すると，抽象文字〈蕨〉(U+229F5) の漢字構造記述は

⿱ J_{82}〈戳〉
= ⿱ J_{82} ⿰〈雀〉戈
= ⿱ J_{82} ⿰⿱ J_{136} 隹戈

のように展開することができる．よって，前述の「蕨」と「蕨」の正規形と一致する．

ここで，文字間の包摂関係を書き換え規則と見なして，包摂規準の集合から構成された項書き換え系にその書き換え規則を付け足したときの完備化の問題を考える．

たとえば，字体「蕨」「蕨」とこの 2 つの字体を包摂する抽象文字〈蕨〉の間に包摂関係が存在し，その包摂関係の存在があらかじめ分かっているならば，そのこと自体から書き換え規則

- 蕨 → 〈蕨〉
- 蕨 → 〈蕨〉

が得られる．この文字単位の包摂関係から導出された書き換え規則の集合と包摂規準（言い換えれば，部品単位の包摂関係）から導き出された完備化された書き換え規則の集合

- ⺾ → J_{82} , ⺾ → J_{82} , ⺾ → J_{82}
- 少 → J_{136} , 少 → J_{136}

を混ぜたとき，字体の漢字構造記述に包摂規準に基づく書き換え規則の集合を適用した結果得られた正規形と，字体と抽象文字の包摂関係に基づく書き換え規則の集合を適用した結果得られた正規形が一致しているため，弱合流性があることから，これらを混ぜた書き換え規則の集合も完備化されていることが分かる．これに，さらに抽象文字〈戳〉や抽象文字〈雀〉とそれらに包摂される字体との包摂関係に基づく書き換え規則の集合を入れたとしても同様である．

このように，ある抽象文字の漢字構造記述が対応する字体の漢字構造記述を包摂規準の集合に対応する完備化された項書き換え系で処理した結果得られた正規形と等価な場合，その抽象文字と対応する字体は完備な包摂記述を持つということにする．

7.2 包摂規準の適用除外の問題

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．UCS 統合漢字における元規格分離や別字源の文字と判断された結果符号位置が分離されたものも同様である．ただし，字源や音価，字義といった形状以外の要素で分離されたものはここでは扱わない．

包摂規準の適用除外（以下では，包摂除外とする）が必要となるのは典型的には包摂規準的には包摂可能な形状の似た異体字を別の符号位置に分離してしまったことに起因しており，本来想定された抽象文字粒度よりも細かい粒度の抽象文字が生じてしまっているのだと考えることができる（あるいは，逆に，本来分離すべきものを過剰に包摂してしまったケースでは，本来想定された抽象文字粒度よりも粗い粒度の抽象文字が生じてしまっているのだと考えることができる）．つまり，包摂除外は単一の包摂規準の一様な適用がなされていないということであり，言い換えれば，文字ごとに異なる包摂規準が適用されているということである．

たとえば，「巻」(= ⿱⺷己）と「巻」(= ⿱⺷巳）の場合を考える．JIS 包摂規準より

- ⺷ → J_{14} , ⺷ → J_{14}
- 己 → J_{67+70} , 巳 → J_{67+70} , 巳 → J_{67+70}

であるから，この両者の正規形はともに ⿱ $J_{14}J_{67+70}$ となるが，JIS X 0208 には 20-12 (巻) と 50-43 (巻) の 2 つの符号位置が存在し，UCS でもそれぞれ U+5DFB と U+5377 に分離されている．よって，このそれぞれに抽象文字〈巻〉と〈巻〉が存在することになる．もし，互いの包摂範囲が重ならないようにするならば，それぞれの包摂範囲を狭めることによって ⿱ $J_{14}J_{67+70}$ の包摂範囲を分割する必要がある．たとえば，〈巻〉と〈巻〉の弁別のポイントを「己/巳」と「巳」の差異だと解釈すれば，JIS 包摂 67 の適用を除外した書き換え規則

- ⺷ → J_{14} , ⺷ → J_{14}
- 己 → J_{70} , 巳 → J_{70}

に対応する部品を用い，抽象文字〈巻〉と抽象文字〈巻〉の漢字構造記述をそれぞれ ⿱ $J_{14}J_{70}$ と ⿱ J_{14} 巳のように書くことができる．これらは字体「巻」「巻」に包摂除外がなかった場合の正規形 ⿱ $J_{14}J_{67+70}$ と異なっており，完備な包摂記述になっていない．

こうした場合に，包摂規準の集合から仮想的に想定される抽象文字を設け，その仮想的な抽象文字と現実の符号位置に対応する抽象文字の包摂関係を記述すれば，その仮想的な抽象文字はそれに包摂される字体との間で完備な包摂記述を実現することが可能である．一方，現実の符号位置に対応する抽象文字は包摂規準の集合から想定される抽象文字ではないので，抽象文字粒度と字体粒度の中間に位置する粒度の文字オブジェクトとして扱い，同様にこの中間的粒度の部品オブジェクトを設けて，対応する抽象文字粒度および字体粒度の部品との包摂関係を記述すれば，その中間的粒度に対応する包摂規準の集合における完備な包摂記述を持たせることができる．

たとえば，包摂除外がなかった場合，字体「巻」「巻」は

正規形⿰ $J_{14}J_{67+70}$ に対応する仮想的な抽象文字〈巻/巻〉との間で完備な包摂記述を満たす．この仮想的な抽象文字〈巻/巻〉は現実の符号位置に対応する抽象文字〈巻〉・〈巻〉を包摂するので，この包摂関係に基づく書き換え規則

- ⿰ J_{14} J_{70} → ⿰ J_{14} J_{67+70}
- ⿰ J_{14} ⿱ → ⿰ J_{14} J_{67+70}

を追加すれば，弱合流性が満たされる．

このことは，包摂規準の集合の観点から見て現実の符号位置に対応する抽象文字〈巻〉・〈巻〉の包摂粒度が抽象文字粒度よりも細かく字体粒度よりも粗いことを示している．そして，JIS 包摂 67 の適用を除外した書き換え規則の集合はこの粒度に対応したものといえ，J_{70} はこの粒度の部品と見なすことができる（なお，JIS 包摂 67 より J_{70} は抽象文字粒度の部品 J_{67+70} に包摂される）．そして，この書き換え規則の集合を適用すると「巻」= ⿰⺍己の正規形は⿰ $J_{14}J_{70}$ となる．また，「巻」= ⿰⺍⿱己の正規形は⿰ J_{14} ⿱ となる．これらは，それぞれ，〈巻〉・〈巻〉の漢字構造記述と一致するため，この粒度の書き換え規則の集合における完備な包摂記述が満たされる．

8. 関連研究

文献 [2] では「CHISE 漢字構造情報データベース」[3] を用いた IDS 形式の漢字構造記述の正規化アルゴリズムとそれに基づく漢字の同一性のチェック手法が提案されている．この手法は『見た目』に基づく分割を行った IDS 形式の漢字情報記述を正規化して重複チェックするものである．ここで，『見た目』に基づく分割というのは漢字の字源や音符・意符といった部品の機能や部品の生産性といった観点は無視し，見掛け上の部品の構成だけに着目して，たとえば，「幹」という字であれば，「⿰ 卓⿰ へ干」のような構造として解釈することである（字源的に考えれば「⿰軋干」のように解釈される）．

IDS の正規化は，部品の正規化（統合漢字と部品文字や，統合漢字内に文字単独用と部品用が重複して存在する場合や，包摂規準に基づいて似た形の部品をどれかに寄せる処理）と囲み系 IDC の除去処理，縦横の並びの平滑化処理からなる．囲み系 IDC の除去処理は「⿴」「⿵」「⿶」を「⿰」に，「⿷」「⿸」を「⿱」に，「⿻」は「⿻ xy」を「⿰ yx」にする処理である．縦横の並びの平滑化処理は，囲み系 IDC の除去処理を行った後の IDS に対して，縦または横に連続する部品列をまとめる処理である．これにより 3 個以上の部品が縦または横に連続している場合の分割点の差異を除去することができる．

この手法は『見た目』に基づく分割を行った漢字構造記述を対象にしているため，JIS 包摂規準のようにそうした仮定をおいていない包摂規準の集合はそのままでは適用できず，事実上，IWDS 1 専用になっているといえる．IWDS 1 は完備化されていないため，部品の正規化をナイーブに行うと無限ループに陥るが，より複雑な（画数の多い）文字に置き換えたり，逆方向への書き換えを避けたりするといった工夫を用いることでこの問題を回避している．また，「⿻」はうまく扱えない[*11]．

項書き換え系という観点で見た場合，書き換え系の停止性とその処理系の停止性は別問題であり，最左最外戦略のような停止性を持たない項書き換え系においても無限ループに陥ることがない書き換え戦略も存在する．また，本論文で提案する手法のように，書き換え規則の集合を完備化することも考えられる．

本手法では『見た目』に基づく分割を行った IDS を前提にしており，漢字構造記述の変換は扱っていないが，字源的・機能的な分割を行った IDS を『見た目』に基づく分割を行った IDS に変換する書き換え規則を記述することも考えられる [10]．項書き換え系を用いた場合，字源的・機能的な分割を行った IDS を直接扱うことも可能であり，より一般的な漢字構造記述を機械処理することが可能であるといえる（この方が本手法が苦手とする，『見た目』に基づく分割を行った結果部品が離れ離れになってしまうケースに対処しやすくなると考えられる）．しかしながら，異体字生成パターンには字源的・機能的な部品の変化もある一方で，『見た目』に基づく分割を行ったときに見掛け上現れる部品が組織的に変化する例もあるため，理想的には可能な分割パターンすべてに対して包摂規準の適用が可能であることが望ましいといえる．項書き換え系という枠組みはこうした問題にも対処可能なものであるといえるが，実際にすべての異なる分割パターン間の変換を実現する書き換え規則を記述するのは容易ではないといえる．よって，部品ごとの『見た目』に基づく分割と字源的・機能的な分割の関係（書き換え規則）の蓄積を中心に，これら以外の分割パターンを生成するような書き換え規則を求めることが重要であるといえる．この観点で見た場合，本手法における平滑化処理はこの一種と見なすことができる．いずれにせよ，項書き換え系という観点で分析することにより，本手法の利点と問題点をより的確に把握することが可能となると考えられる．

9. 今後の課題

9.1 別字源の部品の扱い

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まった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

[*11] これはどのような手法であっても仕方がないことだとはいえる．

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な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きたものや，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．

10. おわりに

項書き換え系を用いた包摂規準の形式化の試みについて述べた．漢字は字種数が多いうえ，同一字種内での異体字も存在しうるため，しばしば数万から数十万種類にも及ぶ膨大なセットを対象にした整理作業を行う必要が生じ，人手で行うには労力・コスト的に大変であり，また，作業ミスも多発しやすいため，機械的なチェックを導入することは不可避であるといえる．しかしながら，漢字における文字の等価性や符号化文字の包摂範囲といったその基礎的な定義自体が高度な人文知を要求する難しい問題であったために，その形式化が進まず，機械処理の恩恵を受けにくかったといえる．IDS の策定と CHISE 漢字構造情報データベースの開発によって，検索や分析を行うことが可能になったが，各符号位置の包摂範囲の曖昧性の問題を解決するためには包摂規準を計算可能なものにして機械処理することが望ましい．そこで，すでに様々な理論的蓄積があり実装も行われている項書き換え系の分野の知見を用いて検討を行った．

漢字の包摂規準は本質的に木構造のパターンに対する書き換えとして記述されているため，項書き換え系における書き換え規則として表現することは容易である．また，完備化アルゴリズムを用いることで計算機にとってより扱いやすい形に変換することができる．しかしながら，これは同時に包摂規準の集合のバグ（包摂すべきものをちゃんと包摂できているか，分離すべきものをちゃんと分離できているかという満たすべき性質を満たせていないこと）に対してどう対処するかという新たな問題を生じさせる．

包摂規準が複数の字体（例示字形）を包摂した既存の文字符号ないしはその運用実態から帰納的に構成されていることを考えれば，文字単位の包摂関係と包摂規準の双方からダブルチェックできることが望ましい．本論文では，文字単位の包摂関係を，包摂規準に基づく部品単位の包摂関係に基づく項書き換え系に加えた完備化された項書き換え系を用いて，完備な包摂記述という概念を導入し，このようなダブルチェックの仕組みの形式化を試みた．これは現在 CHISE 文字オントロジ [9] で導入を進めている『多粒度漢字構造モデル』[11] を項書き換え系の観点から形式的に説明したものと見なすことができる．完備な包摂記述は証明図の一種と見なすことができ，これにより安定した包摂情報の記述が可能になると考えられる．

本論文では cognate/non-cognate の問題を扱うことができなかった．この問題を実質的に扱うためには字音の情報が必要であるといえ，漢字構造および包摂規準という漢字の抽象形状を対象とした枠組みでは限界があるからである．字音の情報を含めた形式化は今後の課題としたい．

参考文献

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

[2] 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，pp.105–119 (2006).

[3] 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 (2002)，入手先 ⟨https://www.ipa.go.jp/files/000005566.pdf⟩.

[4] Baader, F. and Nipkow, T.: *Term Rewriting and All That*, paperback edition, Cambridge University Press (1999).

[5] 外山芳人：完備化による等式証明，人工知能学会誌，Vol.16, No.5, pp.668–674 (1997).

[6] Knuth, D.E. and Bendix, P.B.: Simple Word Problems in Universal Algebras, *Computational Problems in Abstract Algebra*, Leech, J. (Ed.), pp.263–297, Pergamon Press (1970).

[7] 守岡知彦：CHISE 漢字構造情報データベース，東洋学へのコンピューター利用第 17 回研究セミナー，pp.93–103 (2006).

[8] IRG Working Document Series, available from ⟨http://appsrv.cse.cuhk.edu.hk/~irg/irgwds.html⟩.

[9] Morioka, T.: Multiple-policy Character Annotation based on CHISE, *Journal of the Japanese Association for Digital Humanities*, Vol.1, No.1, pp.86–106 (2015).

[10] 守岡知彦：多粒度漢字構造情報のための包摂規準機械可読化の試み，東洋学へのコンピューター利用第 24 回研究セミナー，pp.91–102 (2013).

[11] 守岡知彦：CHISE による HNG データ収録の試み，漢字字体史研究二：字体と漢字情報，高田智和，馬場　基，横山詔一（編），石塚晴通（監修），pp.185–203, 勉誠出版 (2016).

守岡 知彦 （正会員）

1969 年生．1999 年北陸先端科学技術大学院大学情報科学研究科博士後期課程修了，博士（情報科学）．1999 年電子技術総合研究所 COE 特別研究員．2000 年京都大学人文科学研究所附属漢字情報研究センター助手．2009 年同所附属東アジア人文情報学研究センター助教．漢字文献を中心とした人文情報学の研究に従事．2007 年度山下記念研究賞受賞．