漢字構造変換の試み

守岡 知彦 （京都大学 人文科学研究所）

多くの漢字は複数の部品の組合せからなるため、その構造の機械可読記述は漢字の検索や処理において有用な情報である。特に、なるべく音価や意味的カテゴリーを保持した部品（機能的部品）を使って構造を記述した「機能的漢字構造」は漢字のオントロジーを記述する上で有用であるが、検索においては見掛け上の部品（皮相部品）や見掛け上の漢字構造（皮相漢字構造）も必要になることがある。そのための一つの方法は機能的漢字構造と皮相漢字構造の双方を記述することであるが、もし機能的漢字構造を皮相漢字構造に機械的に変換できれば機能的漢字構造だけを記述すれば良くなりデータ作成やデータ管理のコストを削減する上で効果的であるといえる。そこで、実際に機能的漢字構造を皮相漢字構造に変換するための書き換えシステムを試作した。本稿ではその概要について述べる。

An Attempt to Convert Structural Descriptions of Chinese Characters

MORIOKA, Tomohiko (Institute for Research in Humanities, Kyoto University)

Since many Chinese characters consist of combinations of multiple components, the machine-readable description of their structure is useful information in Chinese character retrieval and processing. In particular, the “functional structure” of Chinese character, which describes the structure using “functional components” that retain as much as possible the phonetic value and/or semantic categories, is useful in describing the ontology of Chinese characters, however apparent components and apparent structures are also necessary for retrieval of Chinese characters. A simple way to deal with this is to describe both the functional structure and the apparent structure, however if we can convert the functional structure to the apparent structure automatically, we can describe only the functional structure. If this is possible, it can be said that this is effective in reducing the cost of data creation and data management. Therefore, we developed a rewriting system to convert a functional structure into the apparent structure. This paper describes the outline of the method.

1 はじめに

多くの漢字は複数の部品の組合せからなる。例えば、

林=⿰木木　　雲=⿱雨云　　広=⿸广厶

といった具合である。このような複数の部品の組合せからなる漢字の構造（部品をどのように組み合わせるか）のことを「漢字構造」[4] といい、その記述を「漢字構造記述」という。[5] 漢字構造記述の標準形式としては ISO/IEC 10646 [1] で定義された Ideographic Description Sequence (IDS) 形式が普及している。

上記の例では部品や漢字構造は自明だが、「旗」の場合、⿰方𠂉其 と書くか ⿸㫄其 と書くかの曖昧性がある。同様に、「嬴」の場合も ⿱吂⿲月女丮 と書くか ⿸𣎆女 と書くかの曖昧性がある。

ここで、「旗」の部品〈㫄〉は単体字「㫃」の意味的カテゴリーを保持した部品（意符）であり、

部品〈其〉は単体字「其」の音価的カテゴリーを保持した部品（声符）であるのに対し、見掛け上の部品〈方〉,〈⾅〉は元となった漢字とのつながりを持たない。このように、元となった漢字の音義等の情報の一部を保持した部品のことを「機能的部品」といい、機能的部品の組合せで構成される漢字構造を「機能的漢字構造」と呼ぶことにする。また、「旗」における〈方〉や〈⾅〉、あるいは、「嬴」における〈⿲月女卂〉のように元になった漢字が存在しなかったり同形の漢字が存在したとしてもそれとの音義等の機能的なつながりが存在しないような見掛け上の部品のことを「皮相部品」と呼び、「皮相部品」の組合せで構成される見掛け上の漢字構造のことを「皮相漢字構造」と呼ぶことにする。

漢字のオントロジーを記述するという観点では機能的漢字構造を記述する方が望ましいといえるが、検索においては皮相部品も利用できる方が望ましいといえる。そのための一つの方法は機能的漢字構造と皮相漢字構造の双方を記述することであるが、もし機能的漢字構造を皮相漢字構造に機械的に変換できれば機能的漢字構造だけを記述すれば良くなりデータ作成やデータ管理のコストを削減する上で効果的であるといえる。そこで、実際に機能的漢字構造を皮相漢字構造に変換するための書き換えシステムを試作した。本稿ではその概要について述べる。

2 機能的漢字構造と部品の生産性

もし全ての漢字の字源が明らかであればその字源に従って部品や構造を確定し記述することができるだろうが、実際には現在 ISO/IEC 10646 に収録されている統合漢字の多数の漢字の『正体』を明らかにすることは必ずしも容易ではない。漢字の音義が不明なこともあり、字源知識がなくてもある程度妥当な構造記述が可能で、かつ、字源が明らかな場合には字源的漢字構造になるべく一致するような漢字構造記述が（可能な限り機械的に）行えることが望ましい。

こうした漢字構造記述を行う上でのヒントとなるものは部品の生産性（部品の造字力）である。[*1] そもそも多くの漢字が部品の組合せでできているのは比較的少数[*2]の部品で多数の漢字を表現するためだといえ、ある部品を持つ漢字が多い程部品らしさが高いということができる。なお、この観点からみた機能的漢字構造や機能的部品、及び、それらの確からしさの計算については別稿で述べる予定である。[*3]

3 漢字構造の書き換え

漢字構造記述の表現形式として IDS の枠組を用いることにする。IDS は部品の組み合わせ方を示す IDC (Ideographic Description Characters) と呼ばれる特殊な文字の後に２つもしくは３つの部品を並べた形式で、Lisp の S 式と同様な前置記法の一種である。但し、IDC の後に後置される部品の数は 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 では部品として漢字・部品文字と IDS を用いたものが利用可能であり、部品として IDS を用いることにより入れ子構造を表現することができる。

なお、ISO/IEC 10646 での定義では漢字部

[*1] 漢字の部品は必ず意符か声符のどちらかという訳ではなく、両者を兼ね備える場合もあれば、「月」のように複数字源のものが合流していてある漢字字形を見ただけではそのどれであるかがすぐには判別できず、その形状自体に意味があるようになったものもある。[3] では漢字が部品（字符）になる際の『記号化』という表意・表音作用を薄めることで音価や字義が異なるより多くの漢字を表現可能にする作用を想定している。部品の生産性という概念はこれを漢字構造記述のデータセットというある種のコーパスに対する統計的性質という視点からいわば逆向きに見たものとみなすことができる。

[*2] 少なくとも、部品の数は漢字の数よりも少ないはずである。実際、CHISE 文字オントロジーの場合、収録文字数約 40 万オブジェクト（字種・字体・字形等各粒度のオブジェクトの延べ数）に対して部品数は１万オブジェクト程度である。

[*3] “Viewpoints on the Structural Description of Chinese Characters”; cf. https://grafematik2020.sciencesconf.org/

品文字として使えるのは UCS に収録された漢字と部品用文字だけであるが、原理的にはそれ以外のものを用いることも可能であり、実際、CHISE 文字オントロジー [2] では UCS に収録されていない漢字・部品や UCS の抽象文字と異なる包摂粒度・包摂範囲を表現するために UCS 非収録文字も併用している。また、IDC に関しても拡張を行っている。[6] ここではこのような拡張を行った IDS （これを「拡張 IDS」と呼ぶことにする）を対象とするが、本稿での議論は通常の IDS でも適用可能である。

このように（拡張）IDS は木構造形式の一種であり、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。そして、[5] で述べたように、項書き換え系 (Term Rewriting System; TRS) における書き換え規則を用いて（拡張）IDS の構文木（の部分木）に対する書き換えを定式化することができる。[5] では包摂規準の書き換え規則化について議論したが、本稿では機能的漢字構造を皮相漢字構造に変換するための書き換え規則を導入する。

網羅的なリストはまだできていないが、経験則に基づき次のような書き換えを導入した。*4

⿸⿰ $LRB \rightarrow$ ⿰ L ⿱ RB (111)

例：旗　⿸㫃其 → ⿰方𠂉其

⿸⿲ <亻><丨>$RB \rightarrow$ ⿰ <亻丨> ⿱ RB (112)

例：修　⿸攸彡 → ⿰亻丨⿱夂彡

⿸⿱ $ATR \rightarrow$ ⿱ A ⿸ TR (121)

但し、T は垂れ。

$T = \{$ ⟨ナ⟩, ⟨厂⟩, ⟨丆⟩, ⟨戸⟩, ⟨⺁⟩, ⟨严⟩, ⟨鳫/鴈⟩, ⟨屮⟩, ⟨覀⟩, ⿸ $xy\}$

例：厳　⿸𫩏敢 → ⿱⺍厰

⿸⿱ $A\overline{T}R \rightarrow$ ⿱ A ⿸$\overline{T}R$ (122)

但し、$\overline{T} = \{T$ 以外 $\}$

例：寝　⿸穿𠬶 → ⿱宀⿸丬𠬶

⿸⿸ E<彳>$R \rightarrow$ ⿸ E ⿰ <彳>R (131)

例：屧　⿸屎枼 → ⿸尸⿰彳枼

⿸⿸ $EAB \rightarrow$ ⿸ E ⿱ AB (132)

但し、$A = \{$ ⟨彳⟩ 以外 $\}$

例：暦　⿸麻日 → ⿸厂替

⿹⿰ $LRB \rightarrow$ ⿰⿱ LBR (210)

例：穀　⿹㱿木 → ⿰𥞤殳

⿻⿺ $LRA \rightarrow$ ⿺⿱ ALR (310)

例：𪓰　⿻豩业 → ⿺⿱业豕生

⿻⿱ $ABL \rightarrow$ ⿱⿰ LAB (320)

例：染　⿻杂氵 → ⿱氿木

⿻⿰ $LRA \rightarrow$ ⿰⿱ ALR (330)

例：敫　⿻放白 → ⿰㬱攵

⿴⿱ $AEM \rightarrow$ ⿱ A ⿴ EM (411)

但し、 E は構え。

$E = \{$ ⟨囗⟩, ⟨彑⟩, ⟨凷⟩, ⟨西⟩$\}$

例：鹵　⿴占メ → ⿱⺊図

⿴⿱ A<凵>$M \rightarrow$ ⿱ A ⿶ <凵>M (412)

例：画　⿴⿱一凵由 → ⿱一⿶凵由

⿴⿱ <冂>$BM \rightarrow$ ⿱⿵ <冂>MB (413)

例：𠀇　⿴𠔃日 → ⿱⿵冂日六

⿴⿱ $ABM \rightarrow$ ⿱ A ⿱ MB (414)

*4 一部、略記のため、項から項への書き換え規則として記述していないが、変数になっている所を定数として展開することで、変数の制約として記述している部分も項から項への書き換え規則に展開できる。

但し、A = { 〈冂〉 以外 }
かつ B = { E と〈凵〉 以外 }
例：橐 ⿴㯻石 → ⿱声豕

⿴⿳ A<口>BC → ⿱⿱ A ⿴ <口>CB (415)

例：櫜 ⿴橐缶 → ⿱𣝣木

⿴⿳ A<冖>BC → ⿱⿱⿱ A<冖>CB (416)

例：櫜 ⿴㯻缶 → ⿱⿱⿱亩冖缶木

⿴⿳ A<丿丿>BC → ⿱ A ⿸⿱ C<丿丿>B (417)

例：帬 ⿴帬⊠ → ⿱尹⿸⿱⊠丿丿巾

⿴⿰ AAC → ⿲ ACA (421)

例：楙 ⿰林矛 → ⿲木矛木

⿴⿰ LRC → ⿰ L ⿰ CR (422)

例：衡 ⿴行奐 → ⿰彳奐亍

⿶⿱⿰ LRBA → ⿱⿲ LARB (511)

例：鬻 ⿶鬻束 → ⿱粥鬲

⿶⿳⿰ LRMBA → ⿳⿲ LARMC (512)

例：鬱 ⿶鬱女 → ⿳⿲木女木罒尌

⿶⿴⿰ LRCA → ⿲ L ⿱ ACR (530)

例：鬻 ⿶鬻古 → ⿲弓⿱古鬲弓

⿵⿱ AKC → ⿱ A ⿵ KCR (601)

但し、K は構え。
K = { 〈門〉,〈鬥〉}
例：鬮 ⿵鬥蟲 → ⿱亠鬮

⿵⿱ A ⿰ LRC → ⿱ A ⿲ LCR (611)

例：嬴 ⿵𣎆女 → ⿱吂⿲月女凡

⿵⿳ AM ⿰ LRC → ⿳ AM ⿲ LCR (612)

例：嬴 ⿵𣎆女 → ⿳亾口⿲月女凡

⿵⿲ AMRC → ⿲ A ⿱ MCR (620)

例：徵 ⿵微言 → ⿲彳訔攵

⿷⿺ NAC → ⿺ N ⿱ AC (710)

例：亾 ⿷亾一 → ⿺乚亼

⿺⿱乚 BC → ⿱⿺乚 CB (811)

例：𥹋 ⿺矣米 → ⿱⿺乚米大

⿺⿱ LBC → ⿱⿺ LCB (813)

但し、B は繞（にょう）以外。
B = { 〈灬〉,〈日〉,〈曰〉,〈心〉,〈壬〉,〈壬〉,〈王〉,〈玉〉}
例：望 ⿺𦣠月 → ⿱朚王

4 実験結果

前節で提案した書き換え規則を用いて、CHISE 漢字構造情報データベースを取り込んだ状態の CHISE 文字オントロジー中の機能的漢字構造 2231 件に対し、式 111〜813 を用いて皮相漢字構造を生成した。このうち 2227 件に対しては適切な変換が行われた。

適切に変換できなかったのは次のものである：

111 • 拜：⿸拜乀 → ⿰龵⿱手乀
「⿰龵⿸手乀」となるべきであるが、そもそも「⿸拜乀」が「拜」の機能的漢字構造として妥当かどうかという問題がある。

122 • 㑭：⿸⿱亼中丶 → ⿱亼⿰中丶
「⿱亼⿸中丶」となるべきであるが、前項と同様に機能的漢字構造の妥当性の問題がある。

210 • 疆：⿺彊土 → ⿰⿱弓土畺
「⿰⿺弓土畺」となるべきである。「土」が意符で「彊」が声符であり、この機能的漢字構造は妥当である。

規則番号	適用件数	成功	失敗
111	341	340	1
112	41	41	0
121	138	138	0
122	71	70	1
131	9	9	0
132	781	781	0
210	122	121	1
310	2	2	0
320	17	17	0
330	9	9	0
411	30	30	0
412	3	3	0
413	1	1	0
414	256	255	1
415	6	6	0
416	5	5	0
417	3	3	0
421	14	14	0
422	164	164	0
511	85	85	0
512	6	6	0
530	7	7	0
601	7	7	0
611	76	76	0
612	10	10	0
620	6	6	0
710	5	5	0
811	2	2	0
813	14	14	0

表 1　書き換え規則の適用結果

414　● 午: ⿻午二 → ⿱𠂉⿱二十

今の所、UCS には「𠂉」が収録されていないため、UCS に基づく狭義の IDS では適切な皮相漢字構造を決定しづらい。なお、「午」は「年」の異体字であるが、この現在の字体は説文小篆の「秊」や甲骨文字の「禾」とは別字源のものと考えられ、「午」も「二」も純粋な記号部品[*5]である。即ち、これらは声符でも意符でもないが、部品の生産性に着目した場合、十分な機能性を有しているといえる。

これらの例は「弓」や「十」、「中」のように、部品の矩形領域中に小さな部品を配置可能な空間があり、その上下左右に「丶」や「二」のような小さな部品を配置した場合に食い込まれる形になることによる。特に、「十」や「中」のように複数の領域に配置可能な形状の場合、この手法では決定困難だと思われる。同様に、⿻ を含んだ複合部品が現れる場合も問題が生じ得る。このような場合、IDC を拡張するなどしてどういう交差関係になっているかを記述するためのヒントを示す必要があるといえる。なお、今回はそうした複合部品の内部構造を皮相漢字構造とすることによって問題を回避した。

5　おわりに

前節で提案した書き換え規則を用いて、CHISE 漢字構造情報データベースを取り込んだ状態の CHISE 文字オントロジー中の機能的漢字構造 2231 件に対し、式 111〜813 を用いて皮相漢字構造を生成した。このうち 2227 件に対しては適切な変換が行われた。

今回は機能的漢字構造から皮相漢字構造への変換を試みたが、皮相漢字構造から機能的漢字構造への変換や、任意の漢字構造から機能的漢字構造への変換についても考える必要があろう。

また、機能的漢字構造から皮相漢字構造への書き換え規則に限定しても、今回提案したものはこれまでに著者が発見した書き換えパターンの内、CHISE 漢字構造情報データベースに収録された文字での適用例が存在するものである。

[*5] 裘錫圭のいうところの「記号字」

潜在的にはまだ未発見の書き換えパターンが存在し得る。

また、縦もしくは横に連続して複数の部品が並ぶパターンに関しては今回は取り扱わなかった。こうしたケースの場合、より身近な（≒生産性の高い）部品の組合せにする方が判りやすいといえるが、検索に際しては可能な全パターンを生成することも考慮に値するだろう。この問題に関しては今後の課題としたい。未知の書き換え規則を発見するという観点でも、分割可能な全パターンから部品の生産力の総和を最大化するようなものを見つけ出すという『機能的漢字構造解析』という問題系を提案したい。

今回提案した書き換え規則を逆に用いた場合、項書き換え系の観点では完備な書き換え系を構成可能といえ、なんらかの正規形が得られるといえるが、その書き換え結果の妥当性は保証できない。意味のある機能的漢字構造を得るためにはこの部品の生産性の観点に基づいた機能的漢字構造解析が必要になると思われる。

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.

[2] 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.

[3] 浅原達郎. 漢字の字符, 1996 年 12 月. 戎肆庵読裘記之一.

[4] 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, 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. 93–103, 2006 年 3 月.

[5] 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.

[6] 守岡知彦. 内容アドレッシングを用いた多粒度漢字構造情報表現の試み. 情報処理学会論文誌, Vol. 61, No. 2, pp. 171–178, 2020 年 2 月.