2007－DD－59 (7)
2007／1／26

漢字符号化のアポリア

異体字問題解決の糸口を求めて

ジャストシステムデジタル文化研究所
小林龍生

表意文字としての漢字を情報交換のために符号化することには、アーキテクチャの上からも、実際の符号化実務の上からもさまざまな困難と矛盾が伴う。

なかでも、日本社会において人名・地名に日常的に用いられる異体字の扱いは、行政実務の観点からも、住民感情の観点からも、行政の電子化を妨げる大きな要因となっている。

本稿では、長年にわたり符号化文字集合の標準化活動に携わってきた経験と、日本を代表する仮名漢字変換システムの開発に関わってきた経験を踏まえ、解決の糸口を複数異体字の統合の可能性ではなく、複数異体字を区別する現実の必要性に即して考える。

前半では、Universal Multiple-Octet Coded Character Set(UCS) ISO/IEC 10646:2003の定義を、可能な限り厳密に読み解くことにより、符号化文字集合内部で論理的に操作できる問題と、人間と機械の接面で生じるアポリアの切り分けを試みる。

後半では、このアポリアを漢字異体字に即した現象面から瞥見し、妥協可能な解決策への方向性を探る。

The "aporia" of encoding Han Characters

An aproach to the Han Variations problem

JUSTSYSTEMS Digital Culture Research Center
Tatsuo KOBAYASHI

The encoding of Han-Characters is a problematic issue to be solved, not because of technological reasons but also social and cultural reasons. Especially, Han variations used for human and place names in Japan are the most difficult barrier for e-governmentalization.

This paper discusses possibility of effective and practical solution for this issue based on the author's experience on international standardization activities and Japanese language processing software.

1. 情報交換用符号化文字集合のアポリア

1.1 符号化文字集合のターミノロジー

議論の前提として、ISO/IEC 10646:2003 の適用範囲と主立った用語の定義を見瞥見する。

［**1 Scope**］ ISO/IEC 10646 specifies the Universal Multiple-Octet Coded Character Set (UCS). It is applicable to the representation, transmission, interchange, processing, storage, input, and presentation of the written form of the languages of the world as well as of additional symbols.

［1.適用範囲］ この規格は，国際符号化文字集合を規定する。この規格は，世界の言語（用字）を書き表した形（表記形）及び記号の表現・伝送・交換・処理・蓄積・入力・表示に適用できる。

[4.6 Character] A member of a set of elements used for the organization, control, or representation of data.

[4.6　文字] データの構成、制御又は表現に用いる要素の集合の構成単位。

[4.8 Coded character] A character together with its coded representation.

[4.8　符号化文字] 符号化表現をもつ文字。

[4.9 Coded character set] A set of unambiguous rules that establishes a character set and the relationship between the characters of the set and their coded representation.

[4.10 符号化文字集合] 文字集合及びその集合の文字と符号化表現との間の関係を定めるあいまいさのない規則の集合。

[4.10 Code table] A table showing the characters allocated to the octets in a code.

[4.10　符号表] 符号の一つ以上のオクテットに割り当てられた文字群を示す表。

情報交換用符号化文字集合について論じるとき、われわれに与えられた情報は、本質的には、これだけである。

これらの情報をもう少しパラフレーズすると、下記のようになろう。

情報交換用符号化文字集合とは、世界の言語（用字）を書き表した形（表記形）及び記号の表現・伝送・交換・処理・蓄積・入力・表示などに用いる要素と、それに曖昧性なしに一対一に対応するビット列の組の集合である。

そして、文字とは、世界の言語（用字）を書き表した形（表記形）及び記号の表現・伝送・交換・処理・蓄積・入力・表示などに用いる要素そのものである。

上記の定義から、以下のことがらが、演繹的に導かれる。

すなわち、

・ある文字や記号の集合が確定すれば、符号化方式の如何を問わず、その文字集合の個々の要素に対応するビット列が曖昧性なく一つだけ定まる。

・ある文字集合に対して、複数の符号化方式が併存し、送信装置と受信装置の符号化方式が異なるとき、当然のことながら、送信装置と受信装置との間での正確な情報交換は保証

されない。

注：現実には、まれに、文字集合の一つの要素に対して、複数のビット列が対応することがあるが、これは、符号化作業の瑕疵である。
また、逆に、しばしば「文字集合の複数の要素に対して一つのビット列が対応する」という主張が見受けられるが、情報交換用符号化文字集合には、そのような主張を受け入れる余地はない。あるビット列に対応する文字は、公理的に一つに限られる。

以下の議論では、文字とは、上記 UCS の定義の意味で用いる。また、文字とそれに対応するビット列が、曖昧性なく一対一に対応することから、文字とビット列とを区別することなく論じる。

1.2 符号化すべき文字集合のアポリアとは何か

上記を前提とすると、情報交換用符号化文字集合が内包するさまざまな問題は、すなわち、符号化すべき文字集合そのものの問題であることは自明である。

もう一度、適応範囲の文言を見てみよう。

[1 Scope]

ISO/IEC 10646 specifies the Universal Multiple-Octet Coded Character Set (UCS). It is applicable to the representation, transmission, interchange, processing, storage, input, and presentation of the written form of the languages of the world as well as of additional symbols.

[1. 適用範囲]

この規格は，国際符号化文字集合を規定する。この規格は，世界の言語（用字）を書き表した形（表記形）及び記号の表現・伝送・交換・処理・蓄積・入力・表示に適用できる。

ここで言及されていることを、再度パラフレーズすると、

- **符号化文字集合が対象とするのは、世界の言語を書き表した形（表記系）である**
- **表記系の表現・伝達・交換・処理・蓄積・入力・表現などの《操作》に用いる**

という二点に集約できる。前者、《表記系》を対象とする、ということは、通常、言語学の対象となる音声言語は対象とせず、あくまでもその書き表された形のみを対象とする、ということである。

そして、後者の［さまざまな《操作》に用いる］という部分に、アポリアが内包されている。

ここで、《操作》をさらに二つのグループに分ける。

- **グループ1：《表現》《入力》《表示》**

・ グループ２：《伝達》《交換》《蓄積》

グループ１は、広い意味での機械処理とそれにかかわる人間との接面に介在する。

グループ２は、機械処理それ自体の内部処理である。グループ２の《操作》は、ビット列を対象として、曖昧性なく行われなければならない。これは、情報処理技術もしくは情報通信技術に対する根源的な要請、すなわち、この要請を否定すると情報通信技術の社会的存立基盤自体が否定されてしまうものであるあり、本稿では立ち入らない。

問題は、グループ１にある。《入力》については、機械と人間との接面にさまざまな要素が介在し、議論が煩瑣になるうえ、本稿の主題である符号化文字との関係では議論が発散するので、やはり本稿では論じず、議論を、《表現》と《表示》に絞る。

《表現》とは、何らかの言語の表記系を文字の列に置き換えること、と考えて差し支えない。

また、**《表示》とは、文字の列を何らかの言語の表記系に置き換えること**、と考えて差し支えない。

では、この言語の表記系とそれに対応する文字の列との対応関係は、どのようなものだろうか。

じつは、ここにこそ、曖昧性、もしくは、一意に解決することが不可能なアポリアの介在する余地が存在する。

この関係は、自然言語と、それが指し示す対象世界との対応関係、言語世界を構成する要素全体と、その指し示す世界を構成する要素全体との関係に類似した、互いが互いを支え合う構造と同様の構造を持っており、言語と世界の関係が持つ困難さと同質の困難さを持っている。符号化文字は対象世界との関係性の中に、言語世界が持つアポリアを共有しているのである。

2.1 対象としての文字インスタンス

以後、《言語の表記系》の具体化された存在、新聞や書籍の個々のページ、手紙やメモなどを構成する個々の単位を文字インスタンスと呼ぶことにする。（※じつは、この文字インスタンスをどう切り出すか、という部分にも、さまざまな問題が内包されているが、この点についても本稿では議論しない）

直感的に、ある文字に対応する文字インスタンスが無限に存在し、ある範囲の揺らぎの中でさまざまな変位をもつことが了解できるであろう。

この関係にもまた、具体個物と抽象概念との対応関係と類似したやっかいな問題を内包していることもまた自明である。この問題の複雑さを垣間見るためには、西欧中世以来の個物と概念とを巡る普遍論争に言及するだけで十分であろう。

本稿の後半では、文字インスタンスと文字との対応関係に関し、議論を本稿のテーマである異体字問題に絞って、具体的に論じる。

2.2 漢字コードのアポリア:異体字問題とは何か

そもそも異体字とは何か。

文字インスタンスには、その範囲を活字化されたものに限っても、意味や音価が同じでも、さまざまなレベルで視覚的な形の変位がある。

タイポグラフィーや文字コードの世界では、《書体差》、《字体差》、《デザイン差》などといった使い分けがされている。

書体差とは、

《龍》と《龍》

のような、フォントデザインの差を指す。

字体差とは、

《崎》と《埼》

のような、字の骨組みを構成する要素間に明らかな相違がある場合に用いる。

デザイン差とは、

《  》と《  》

のように、活字設計上に微細な表現の差を指す。

一般には、書体差、字体差、は直交する概念であり、デザイン差は、それらに比べて微細な形態の相違であるという了解では一致しているが、それぞれの境界をどこにおくかについては、発言者によって各者各様である。

一般的な符号化文字集合は、ある変位を持った字体をひとまとめにして一つの符号位置を付与している。すなわち、一つの文字には、字体のレベルでもある範囲の変位がありうる。

本稿では、これら、《書体差》、《字体差》、《デザイン差》の違いについては、一切言及せず、デザイン差も含め、字の形の相違が視認できるものは、すべて《字形差》と呼ぶ。

以下に、異体字関係にあると一般に認識されている文字インスタンスの例を挙げる。

このなかには、文字として区別して符号化されているものも区別せずに一つの文字として符号化されているものもある。

《高》と《髙》

《吉》と《𠮷》

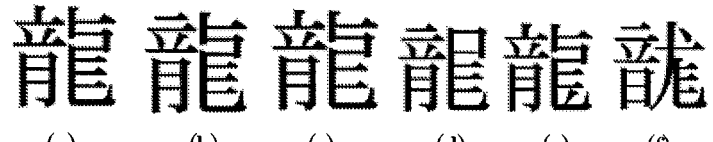
《崎》《碕》《嵜》

《辺》《邊》《邉》

《富》と《冨》

以下に挙げるのは、《龍》(U+9F8D)に相当する文字インスタンスの例である。

上段には、経済産業省が主管し、総務省、法務省の協力によって、住民基本台帳や戸籍の電子化の際に用いられる可能性のある字形の整理を進めている汎用電子情報交換環境整備プログラムで収集した字形セットに含まれる文字インスタンスを掲げてある。下段の左端は常用漢字体（略体字ではなく古字)、次が中国の簡体字。残りは、さまざまなフォントデザインの異なるインスタンスである。



(a) (b) (c) (d) (e) (f)

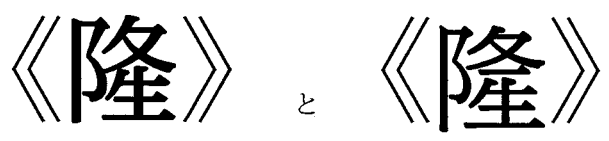
(g) (h) (i) (j) (k) (l)

これらの文字インスタンスを、改めて符号化することを考える。

両極端は、これらを、すべて「想像上の動物の一種」を表す概念に対応するものとして、同一の符号を与える場合と、これらの字形の相違を重視して、すべて異なる符号を与える場合である。(本稿で、各インスタンスの下に付した(a)～(l)の記号も、文字符号の要件を満たしている)

UCS では、これらのうち、(a)、(g)、(b)は区別して符号化しており、(i)、(j)、(k)、(l)は(a)の書体差とみなし、文字としては全く区別していない。

しかし、たとえば



のように、UCS では基本的に同一視（Unify）されていながら、常用漢字表では異なる字体として区別して扱われている場合もある。

(a)と(g)の関係は、一般の社会生活において、しばしば交換可能な形で用いられる場合があり、新聞や書籍等で、《芥川竜之介》《芥川龍之介》、《橋本龍太郎》《橋本竜太郎》といった形で、新聞社や出版社によって、しばしば表記の仕方に差異が見られる。

《檜山》と《桧山》、《淺松》と《浅松》のように、戸籍上の表記にかかわらず、常用漢字体とそれに対応するいわゆる康熙字典体を、本人が混同して用いる場合もしばしば見受けられる。

一方、(a)と(e)に関しては、現在の UCS では完全に同一視し、符号化文字として区別して標準化される可能性も非常に低い。しかし、現実には名刺などの表記に好んで(e)の字形を用いる人も多く、名刺印刷を主とする印刷業者などは、必ずと言っていいほど、方法の如何を問わず、(e)の字形での印刷を可能にする手段を提供している。

3.1 異体字のアポリアをのりこえrうために

以上、現状を瞥見しただけでも、無限に存在する文字インスタンスをどのような基準で有限小数個の文字に対応づけることが、いかに多様な可能性を内包しているかが了解できる。万人を納得させる唯一の基準が存在し得ないという了解を前提として、では、どのような現実解がありうるか。

本稿の残された紙幅では、その可能性についての私見を述べる。

まず、**《すべての文字インスタンスを有限個の文字に対応づける万人が納得する規則はない》**という共通の理解が必要である。

その上で、**《ある文字集合が規定されるとき、その文字集合の要素間の関係それ自体が個々**

の文字インスタンスと要素文字との対応規則を構成する。また、その規則はその文字集合を共有する文化集団の中での暗黙の了解に基づく》と考えられる。
さらに、《ある文字集合の中に、何らかの意味で相互に類似した複数個の要素文字の可能性があるとき、それらを独立の文字として扱うためには、その集合の利用局面において、弁別して扱う蓋然性の高い要求に関する合意》が必要である。
また、その際、《要素文字の不用意な区別は、使用用途によっては利便性を損なう》という了解が必要である。

トートロジーとなるが、《表記系の表現・伝達・交換・処理・蓄積・入力・表現などの《操作》の際、弁別する必要のない文字インスタンスに対して別個の文字を対応づける必要はない》のである。

再び、《龍》の例に戻ると、(a)と(b)と弁別して符号化することを要求するときには、これらが区別して用いられる状況について、説得性の高い論拠を示すことが必要なのである。

3.2 符号化文字のアポリアふたたび

じつは、本稿の立論自体が、あるパラドックスを内包している。

(a)と(b)とを区別して符号化する必要がないことを記述するためには、(a)と(b)を区別して扱うことが不可欠なのである。この事実を敷衍すると、《あらゆる文字インスタンスは、それぞれの差異を記述するためには、独立した文字と対応づけられなければならない》という結論に達する。文字の符号化のアポリアは尽きることがない。

参考文献

[1]ISO/IEC International Standard
International Standard 10646
ISO/IEC 10646 1st Edition
Information technology — Universal Multiple-Octet
Coded Character Set (UCS) —
Architecture and Basic Multilingual Plane
Supplementary Planes
[2]国際符号化文字集合(UCS)——第一部：体系及び基本多言語面
JIS X 0221-1:2001(ISO/IEC 10646-1:2000)
※ [2]は、[1]の旧版の翻訳 JIS である。[1]およびその Amendment 1、Amendment 2 を反映した翻訳 JIS は、翻訳作業が終了し、2007 年の早い時期に出版される予定である。
[3]常用漢字表（内閣告示第一号 昭和五十六年十月一日）
[4] 「汎用電子情報交換環境整備プログラム」4 年間の成果
http://www.itscj.ipsj.or.jp/jp/mojidb.html