解説

世界の文字と文字符号
（前編）

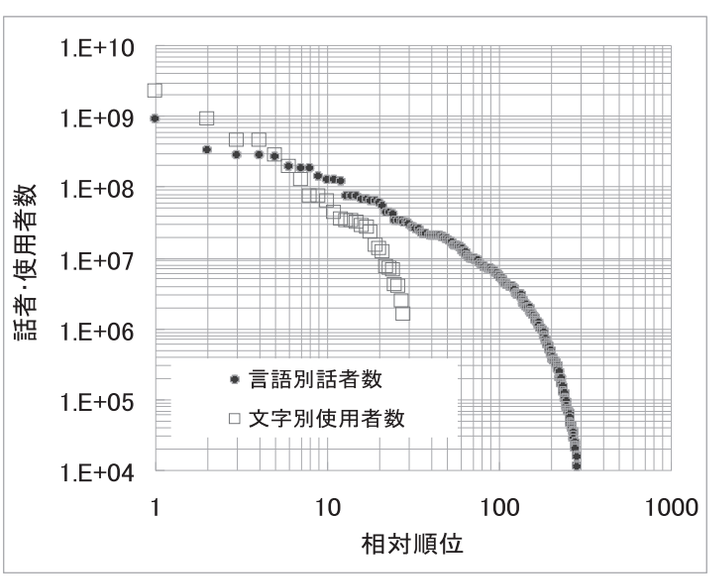
三上 喜貴
長岡技術科学大学
mikami@kjs.nagaokaut.ac.jp

本稿は，前編・後編の 2 回にわたり，世界の多様な文字の符号化の歩みを振り返り，その到達点としての国際符号化文字集合 ISO/IEC 10646 の意義となお残る課題について述べる．前編では世界の言語と文字体系を概観するとともに，ラテン文字に代表されるアルファベットとアラビア文字に代表される単子音文字について解説する．

◦ 世界に言語はいくつあるのか ◦

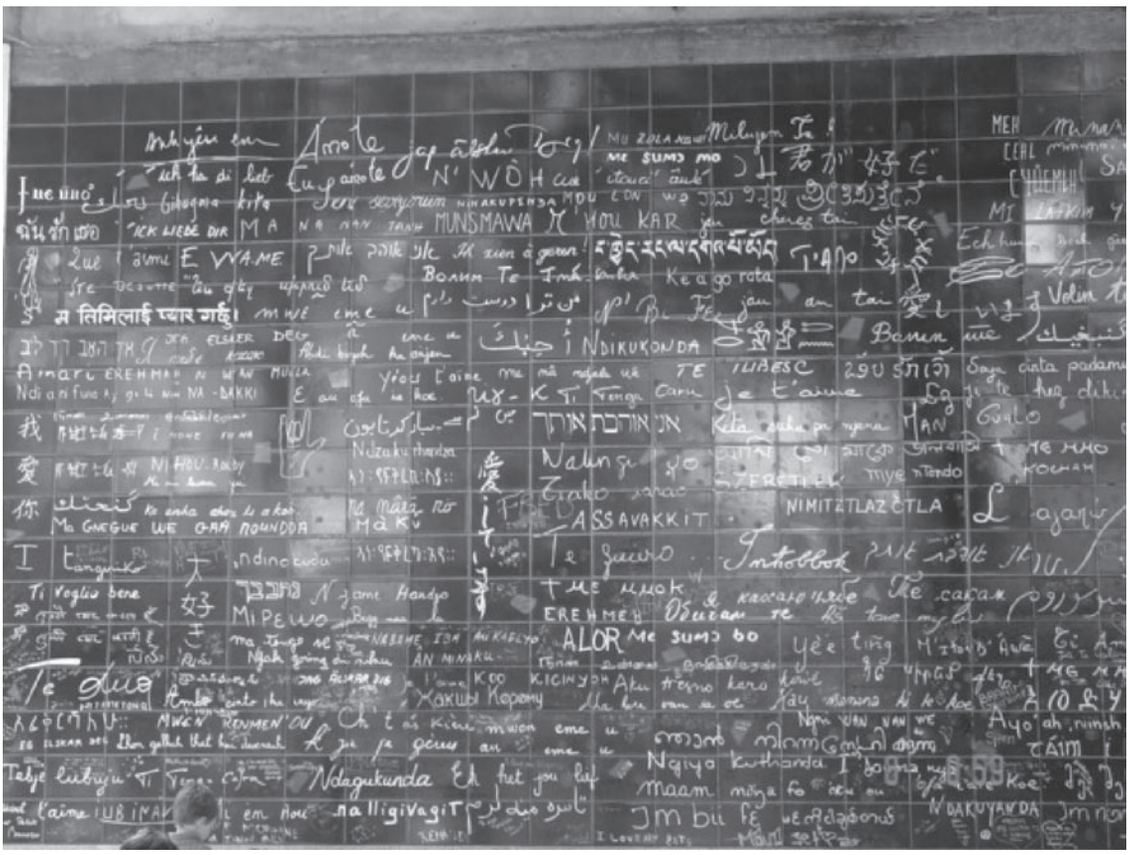
世界には一体いくつの言語があるのか．世界の詳細な言語分布地図を明らかにした労作である『世界民族言語地図』は約 6,500 言語を，この分野でしばしば引用される言語カタログ "Ethnologue" 第 15 版は 6,912 言語を収録している．毎年 2 月 21 日を国際母国語記念日と決めたユネスコの決議文も「地球上で話されている 6,000 余りの母国語を称えるために云々」と述べている．歴史上最も多くの言語に翻訳された文書は聖書であろうが，最初の聖書英訳を行った John Wycliffe の名に因むウィクリフ聖書翻訳協会は，「部分訳を含めると聖書はすでに 2,212 言語に翻訳されており，2025 年までには 6,800 言語への翻訳を完成させる計画である」と述べている．言語の総数は 6,000 ～ 7,000 と考えてよかろう．

ここで，話者数の大小を基準にして言語の順位付けを行い，順位と話者数の関係をグラフにすると Zipf の法則のような関係が現れる．話者数最大の言語である中国語の話者数を 100 とすると，2 位の英語が 36，第 10 位に位置する日本語の話者数は 14，20 位のトルコ語は 6.6，30 位のスワヒリ語は 3.3，50 位のウズベク語は 2.0，100 位のトルクメン語が 0.6 となり，両対数グラフ上でほぼ直線上に並ぶ（**図 -1**）．しかしこの辺りから Zipf の法則からの乖離が著しくなり，話者数は急速に小さくなる．話者数の



注）話者人口は世界人権宣言のサイトに掲載された数値による．文字別使用者数は話者数を使用文字に従って集計した．

図 -1　言語・文字の相対順位と話者・使用者人口



場所）Jehan Rictus Square, Paris.　2005 年 6 月　Wunna Ko Ko 撮影.

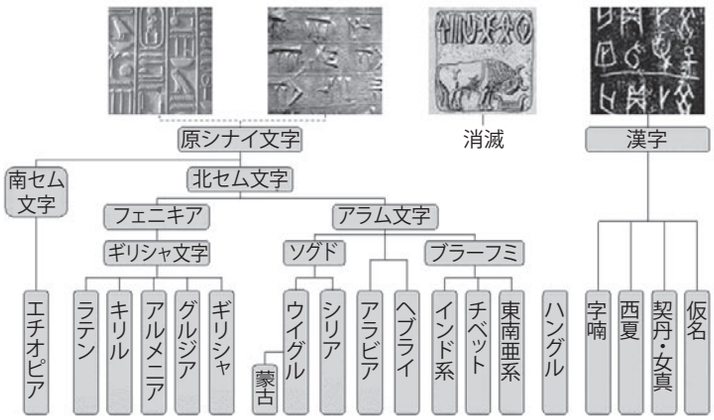
図 -2　世界の文字で表現された「愛しています」

少ない言語の中には絶滅に瀕する言語も増えており，ユネスコの絶滅言語に関する報告書[5]によれば，世界中で，平均して 2 週間に 1 つのペースで言語が死滅しているという．同報告書は，世界中の言語のうちの半分はパプア・ニューギニア（言語数 832），インドネシア（同 731），ナイジェリア（同 515），インド（同 400），メキシコ（同 295），カメルーン（同 286），オーストラリア（同 268），ブラジル（同 234）の 8 カ国に集中しており，その多くは書かれた言語を持たないともいう．こうした事情を考慮すると，文字符号を論ずる文脈からすれば，対象となる言語総数は 6,000 よりかなり小さい．

たとえば，国連高等人権弁務官のサイト[3]をみると『世界人権宣言』（Universal Declaration of Human Rights）の各国語訳が閲覧できるが，本稿執筆時点で翻訳されているのは 329 言語であり，言語コード ISO 639 が識別しているのも約 440 余りの言語である．筆者らは，科学技術振興機構が実施する社会技術研究システムの公募プロジェクトとして，「言語天文台」と称するネットワーク上の言語活動を観測する調査プロジェクトを進めている[6]．その問題意識は，ネットワーク上の言語活動は話者数の分布と見合ったバランスのとれた姿になっているのか否かを確認するとともに，使用されている文字符号の実態を確認しようということであり，当面の観測対象としているのは，世界人権宣言の翻訳対象となっている 300 余りの言語である．このプロジェクトのことについては後編でふれる．

◦ 文字体系はいくつあるのか ◦

では言語を写す文字体系の種類はいくつあるのであろうか．ラテン文字，キリル文字やアラビア文字のように多数の言語で共用される文字も多く，また，文字を持たない言語も多いから，異なる文字体系の総数は言語の総数よりもはるかに小さい．Florian Coulmas の編集した "Encyclopedia of Writing Systems" は約 300 の文字体系を収録しているが，その多くは歴史上の文字である．また文字体系に関するコードである ISO 15924 Codes for the representation of names of scripts には本稿執筆時点で 100 種類余りの文字体系が登録されており，国際符号化文字集合 ISO/IEC 10646 Universal multiple-octet coded character set（UCS）のトップページである基本多言語面には 50 余りの文字体系が収録されているが，このうち各国公用語で現在使用されている文字体系の総数は 28 種類である（残りは康熙部首，数学記号，国際音標記号など）．トップはラテン文字で使用者数が約 22 億人，そして，漢字，キリル文字，アラビア文字，ベン



出典)「世界の文字の図典」, 吉川弘文館より, ただし一部省略して簡素化

図-3　世界の文字の系統樹

ガル文字，デーヴァナーガリー文字，日本語の仮名文字，ハングル文字，テルグ文字，タミル文字が上位 10 位までの文字体系である．言語別の話者数分布と同様にして文字体系別の使用者数分布をみると，この 28 種類の使用者数人口と順位の関係は Zipf 法則よりも傾斜のきつい次数が -2 程度の Power-Law 曲線に従う．

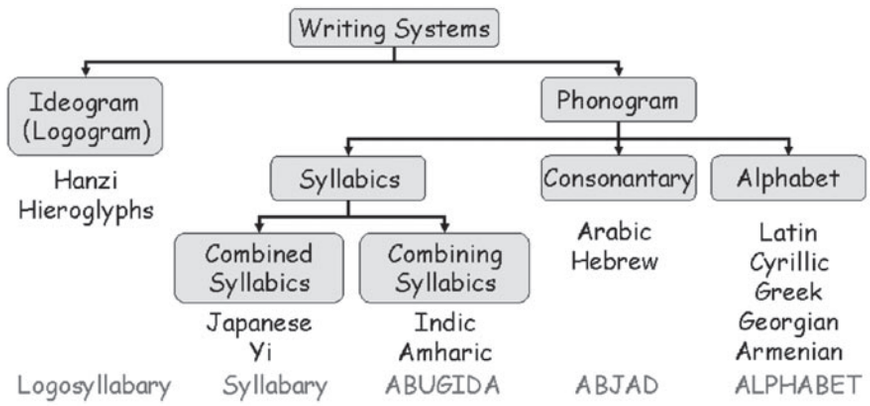
ここできわめて大雑把に世界の使用文字を概観すれば，ヨーロッパ大陸ではラテン文字，キリル文字，ギリシャ文字の 3 つの文字体系が使われており，コーカサス地方にグルジア文字とアルメニア文字がある．アメリカ大陸は南北あわせておおむねラテン文字で足りる．アフリカ大陸の北半分はアラビア文字圏であり，エチオピアにはアムハラ文字もあるが，南半分の言語はラテン文字をベースとした文字体系で表記される．しかし，アジア地域──中近東から中央アジア，南アジア，東南アジア，東アジアへと連なるユーラシア大陸アジア部──には，国境を越えればまちがいなく文字が変わり，さらに国境の内側ですら多数の文字が共存するというほどの多様な文字世界がある．

◦ 文字の系統樹と分類 ◦

歴史的に見ると，四大文明の生み出した文字を出発点として，実に多様な文字体系が生み出されてきた．インダス文明の残した未解読の文字を別として，中国の黄河文明の残した文字からは漢字，仮名をはじめ，契丹・女真・西夏・字喃などの歴史上の文字が生まれ，メソポタミアやエジプトで生まれた文字からは，複雑な混交の歴史を経て，現代の各種アルファベット，アラビア文字やインド系文字へと連なる文字が生まれてきた（**図 -3**）．

一方，文字をその構成原理に着目して整理すれば，文字が表意的であるか表音的であるかによって，表意文字（ideogram）と表音文字（phonogram）とに大きく二分できる．さらに，文字が表す単位が単語・音節・音素という階層のどこに位置するかという基準によって，表語文字（logogram），音節文字（syllabics），アルファベット（alphabet）と分類することもできる．表音文字は音節文字とアルファベットに分かれ，また，表意文字のうち，意味を表す最小の単位が語であるものは表語文字に対応する．表語・音節・アルファベットの三分法は文字符号化を論ずる技術文書にもしばしば登場する．ISO/IEC 10646 は，技術的な規格文書の常として「文字」（規格文書では一般に "character" と呼ばれるが，これは日常用語としての「文字」よりもはるかに広い概念である）そのものの定義について「データの構成，制御または表現に用いる要素の集合の構成単位」という無機的な定義を与えているだけであり，文字の分類についての説明には立ち入っていないが，暗黙裡には三分法が仮定されており，文字の名称には，"letter"，"syllable (or syllabics)"，"ideograph" のいずれかの接尾名を付している．Latin Capital Letter A，Ethiopic Syllable HA，CJK Unified Ideographs の如くである．

しかし，今日における文字符号開発と利用の問題点を分析する上では，これをもう少し細分化してみる必要がある．すなわち，ラテン文字，キリル文字，ギリシャ文字のように母音文字，子音文字からなる単音文字の体系としての alphabet，アラビア文字やヘブライ文字のような子音字だけで綴る単子音文字体系（consonantary あるいは alphabet の命名法に倣ってアラビア文字の最初の 3 文字から abjad），子音字を中心に母音の音価を変更するための母音記号が結合し，また複雑な多重子音を作ることの多いインド系文字（combining syllabics あるいは同様の性質を持つエチオピア文字の最初の 4 文字の名前から abugida），そして仮名文字や中国の彝文字



出典）Peter T. Daniels, William Bright (ed.): The World's Writing Systems, Oxford University Press, 1996.

図-4　文字体系の分類図

のような表意文字由来の音節文字（単に syllabary あるいは表音的な単位に分解できない音節文字という意味で combined syllabics と命名）および漢字という 5 分類を行い（**図 -4**），以下においては，各文字体系の特色を要約しながら，その文字符号開発の歴史と利用の現状を見ていくことにする．

◦ アルファベットと補助記号 ◦

アルファベットは，表音文字の体系としては最も単純な構造であり，わずか 30 前後の小さな文字集合によって無数の単語を綴る．かつて，新井白石は，屋久島に渡来したイタリア人イエズス会士シドッチ（シローテ）に対する尋問記録ともいうべき『西洋紀聞』（1715 年）の中でラテンアルファベットに触れて「其字母僅に二十余字一切の音を貫けり，文省き，義広くして，其妙天下に遺音なし」とその合理性に驚きを示したという．文字集合の大きさはラテン文字で 26 文字，キリル文字で 33 文字，ギリシャ文字で 24 文字である．

しかし，わずか数種類のアルファベットによって世界中の多くの言語が記述されているということは，これを使用するほとんどの言語にとってアルファベットが元来「借りてきた文字」であるということを意味している．そして，借り物の文字を自らの言語表記にふさわしいものへと磨いていく過程で，文字・記号の追加や綴字法の確立といった表音能力の拡張が行われてきた．フェニキア人の作った 22 文字のアルファベットにギリシャ人が《Y》や《Ω》を追加し，ローマ人は《G》を追加した．今日のラテンアルファベットはローマ時代のアルファベットの中からさらに《J》や《U》を分化させて最終的に 26 文字としたものである．そして，さらにこれがさまざまなヨーロッパ言語に使用される過程で《Ä，Ç，Ö，Ñ》などの補助記号付き文字や《Æ，Œ》などの二重文字が登場した．現時点における拡張ラテン文字の総数は，欧州標準化機構 CEN が定めた Multilingual European Subset-1（ISO/IEC 10646 のラテン文字部分集合）によれば，ユーロ通貨記号《€》まで含めて 335 字である．

このようにして多種多様に分化した拡張ラテン文字集合をいかにして符号化するかという課題は，文字符号に関する最初の国際規格である ISO 646 の制定当初からの難題であった．ラテン文字使用国の枠内に限られていたとはいえ，それは文字符号多言語化の第一歩であった．1960 年代初めの ISO 文書を紐解くと，欧州および中南米の各国語で使用されていた 36 カ国，80 種類に及ぶタイプライタの印字文字集合を調べ上げ，ASCII の 26 文字に対して各国語ごとにどのような拡張が必要であるかをまとめたイタリア代表の労作がある．この文書は，すべての補助記号付き文字に独自の符号を割り当てる "brute force solution" は非実際的であるので国別に特化した符号表とせざるを得ないが，それでも，できる限り共通性を高めるために，多くの言語で使用される補助記号（diacritical mark）を選び，これをバックスペースと組み合わせて重ね打ちすることにより補助記号付き文字を表現するという方策を提言した．結果として ISO 646:1967 では最大 10 文字分のナショナルユース領域が留保され，同時に，diaeresis《 ¨ 》，grave accent《 ˋ 》，circumflex accent《 ˆ 》などの頻出補助記号が収録された．これらの記号は，それぞれ通常は quotation mark，apostrophe，upward arrow として用いられるが，バックスペースと組み合わせて用いられるときにのみ補助記号として解釈されるという多義的解釈を導入することによって辛うじて符号位置が確保されたものである．また，符号表の切り替えのためにはエスケープシーケンスによる方法が工夫され，ISO 2022 Character code

structure and extension techniques が制定された．なお，このエスケープシーケンスの国際登録簿 ISO-IR は，現在本会の情報規格調査会がその管理運営を担当している[7]．

8 ビット表が出現して利用可能な符号位置が倍増すると，ISO/IEC 6937 Coded graphic character set for text communication と ISO/IEC 8859 8-bit single-byte coded graphic character set という 2 つのアプローチが生み出された．前者の ISO/IEC 6937 は，ノンスペーシングの補助記号を導入し，補助記号付き文字を [D]L という符号列で表現した（D はノンスペーシング補助記号，L はアルファベットの基本文字，[ ] は省略可であること示す）．基本文字よりも補助記号が先行するのは，当時の印字装置が機械式タイプライタと同じメカニズムで動作していたからである．また，この規格は補助記号付き文字を含む文字集合全体をレパートリ（repatoire）という形でも規定し，合成の許される文字の種類を制限した．一方，後者の ISO/IEC 8859 シリーズでは，バックスペースもノンスペーシング文字も共に排除し，すべての文字を合成済み文字として表現する道を選択した．しかしながら ISO/IEC 6937 のレパートリが 332 文字という大きさを持つことからも明らかなように，補助記号付き文字を含むすべてのラテン文字集合を 8 ビット表に収容することは不可能であるから，適用地域を適当な地理的範囲に限定することによって文字集合サイズを制約条件内に収め，Latin-1（主として西欧），Latin-2（主として東欧）等の一連のシリーズ規格として制定された．

しかし，ヨーロッパ以外の地でラテン文字を利用するベトナムなどはこの動きから取り残された．声調が 6 種類もあるベトナム語表記に用いられるクオックグー・アルファベットは，すべてを合成済み文字で表現しようとすると通常の ASCII 文字集合に加えて 134 文字の追加が必要となり，このことは 8 ビット表の制約条件下でベトナム語の文字符号開発者を大いに悩ませ，コード乱立の原因となった．同様のことはキリル文字の拡張においても見られた．モンゴルは通常のキリル文字集合に対して 2 つの追加文字を必要とするが，やはり蚊帳の外に置かれ永く文字符号の混乱に悩まされた．ベトナムやモンゴルでこれらの問題が解消されたのは，ようやく ISO/IEC 10646 の登場によってである．

ISO/IEC 10646 では，アルファベットに関する世界中の需要に応えるために必要なすべての補助記号を収録した．その際，合成済み文字による表現と，ノンスペーシングの補助記号を用いた合成列としての表現とが混在することを避けるために，Unicode コンソーシアムなどの実装においては補助記号付き文字を用いた合成列（decomposed form）を正規化表現とするというルールが設けられた．また，基本文字と補助記号の順序も，印字メカニズムの制約などから解放されたより論理的な順序となった（**表 -1**）．

符号化方式	補助記号付き文字の符号化表現
ISO/IEC 646	L BS D　（BS はバックスペース）
ISO/IEC 6937	[D] L　（[ ] 内は省略される場合もある）
ISO/IEC 8859 シリーズ	すべてを合成済み文字として表現
ISO/IEC 10646 の正規化表現	L { D }　（{ } 内は 0 回以上の繰り返し）

表 -1　ラテン文字に関するさまざまな符号化方式

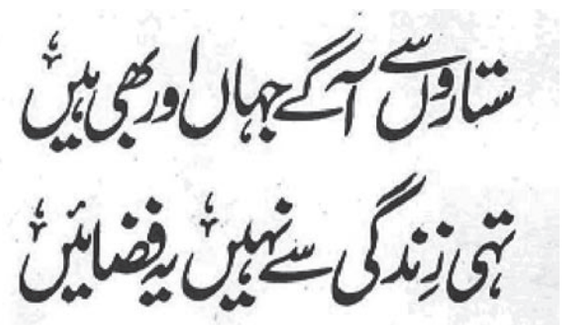


図 -5　ウルドゥー語の流麗なるテキスト

◦ 単子音文字 ◦

単子音文字は子音文字のみによって語が綴られる文字体系であり，ltd. と書いて limited と読ませるのに似ている．アラビア文字の場合には 28 の子音文字によって，ヘブライ文字の場合には 22 文字によって語を綴る．シリア文字，モルジブのターナ文字などもこのカテゴリーに属する．これらの文字は右から左へという特徴的な書記方向を持っているが，符号化された文字列は右書き・左書きといった具体性は持っておらず，先頭から末尾へという抽象的・論理的順序を持っているだけである．書記方向の異なる文字系が混在する場合には面倒な問題となるが，単独で扱う場合には大きな問題ではない．文字符号という観点からみるとき，むしろ，同一文字に複数の表示形が存在するという点が障害となる．

アラビア文字のように連綿として書かれる文字の場合，文字の書き終わりの位置から次の文字の起筆位置へとつなげる線や，語頭・語末の装飾的な運筆，リガチャ（合字形）形成などがつきものである．タイプライタでこれを表現しようとするときには必要な活字を準備するほかない．標準的なアラビア語タイプライタは 65 字程度のアラビア文字（数字は除く）を備えているが，これは，表示形の美しさについて妥協を行った結果であり，期待する表示形を完全な形で表記しようとすれば，これをはるかに上回る多種類の活字を用意しなければならない．

<table>
<tr><td></td><td>活字印刷機
タイプライタ</td><td>ラインプリンタ
ドットプリンタ</td><td>Postscript
True Type</td><td>Character-Glyph
Model</td></tr>
<tr><td>文字サイズ</td><td rowspan="4">活字</td><td>印字カセット切替</td><td>サイズ指定</td><td></td></tr>
<tr><td>字形デザイン</td><td>印字カセット切替</td><td>フォント指定</td><td></td></tr>
<tr><td>グリフ</td><td colspan="2" rowspan="2">文字符号</td><td>グリフ</td></tr>
<tr><td>文字</td><td></td></tr>
</table>

図-6　活字から文字符号へ

アラビア文字を用いて表記するパキスタンのウルドゥー語は流麗なる書体をもって記述されるため，新聞の原稿などですら，比較的近年に至るまで専門の書家の手書き原稿に頼らざるを得なかったようであるが，1980 年代にウルドゥー語の DTP 処理が始まったとき，この流麗な書体を再現するため，著名書家の書体を元に 18,000 ものリガチャに関するディジタルデータが作成されたという．これは，おおむね単語を単位としたリガチャを揃えたことに相当しよう．

しかしながら，こうした表示形の相違は，連綿と書かれるテキストを人為的な単位で区切ったときに生じる見かけ上の相違である，と解釈すべきであろう．日本語の場合でも，連綿と書かれた草書体をコンピュータで出力しようとすれば，アラビア文字と同様に，1 つのひらがなに対して複数のグリフを用意しなければならない．草書仮名の書家たちの間では「連綿字典」が用いられているが，これは前後のかな文字とのつなぎをどのように運筆するかの指針を与えるものであり，この指針に完全に従おうとすれば，ひらがなの場合にも相当数の表示形を用意しなければならないだろう．このような意味で，複数表示形の問題は単子音文字に固有の問題というよりも，むしろ連綿と書かれる文字系に共通の問題であるといえる（ちなみに，同じ単子音文字であっても連綿と書かれることのないヘブライ文字に関してはこの問題はない）．

しかし，検索や編集の都合を考えれば，同一文字の異なる表示形に対して異なる符号を割り当てたときの不便は明らかであり，符号と表示形とを区別することのメリットは大きい．このために導入されたのが，符号とグリフの分離という原則である．この原則が "Character-Glyph Model" という明確な形で定式化されたのは ISO/IEC TR 15285: 1998 An operational model for characters and glyphs であるが，実質上，この機能が大きな意味を持つに至ったのはアラビア文字の処理系だと考えている．後に ISO/IEC 10646 に継承されることとなるアラビア語圏の共通規格 ASMO 449:1982 7-bit Coded Arabic character set for information interchange は表示形の相違を区別せずに同一の文字として扱っており，出力するときには制御文字によってフォントの切り替えを行った．制御文字に関する国際規格である ISO/IEC 6429 Control functions for coded character sets はその初版である 1983 年版に Select Graphic Rendition（SGR）という制御文字が導入されており，10 種類のフォントからの選択に備えた．1988 年に発行された ISO/IEC 6429 第 2 版では，Select Alternative Presentation Variants（SAPV）という制御文字が導入され，アラビア文字出力固有のパラメータとして，独立形，語頭形，語中形，語尾形の選択等が規定された．この国際規格の原型となった European Computer Manufacturer's Association の制御文字規格 ECMA-48 第 4 版（1986）では，この制御文字は同じく SAPV と呼ばれているが，その正式名称は Select Arabic Presentation Variants であった．

活字印刷の時代から情報処理の時代へという変化の過程で，大きさ，字形デザインといった具象性が 1 つずつ剥ぎ取られていき，符号とグリフの分離によって，文字符号はついに特定の図形的表現と結びつかない抽象的存在となったのである（**図 -6**）．同時に，リガチャなどに関する厳しい要求は，すべてレンダリングソフトウェアへの負荷となることとなった．

参考文献と参考 URL

1）小林龍生，安岡孝一，戸村　哲，三上喜貴編：インターネット時代の文字コード，共立出版 (2001).
2）三上喜貴：文字符号の歴史 – アジア編，共立出版 (2002).
3）世界人権宣言の多国語訳：http://www.unhchr.ch/udhr/navigate/alpha.htm
4）Daniels, P. T.：The World Writing Systems, Oxford University Press (1996).
5）Wurm, S. A. (ed.): Atlas of the World's Languages in Danger of Disappearing 2nd edition, UNESCO (2001).
6）言語天文台サイト：http://www.language-observatory.org
7）国際登録簿 ISO-IR: http://www.itscj.ipsj.or.jp/ISO-IR/

（平成 17 年 7 月 13 日受付）