電子化テキストの国際的共有

電子化テキストの国際的共有—付総括—

豊島正之

キーワード：文字コード、JIS漢字、10646-1、SGML、TEI、電子化テキスト（e-text）

要　旨

平成5年度国語学会秋季大会（北海道大学）での、研究発表会分科会Cセッションの概要を紹介し、そこで取上げられた下記の分野の問題と、附随する公開の問題に就て、その後の進展を加味して要約する。

1. 文字コード
閉じた文字集合を暗黙に予期している現在のコード化文字集合は、ISO 10646の様な、漢字を含む国際共通文字コード系が制定された現在でも、漢字系の様に開いた文字集合の為には、不備である。

2. テキスト形式
SGMLの難点は、階層的テキスト構造表記しか無い点と、範囲指定・属性付与を共用する設計であり、SGMLに載ったTEIの様なテキスト交換形式も、その問題を受け継いでいる。

1　セッションの概要

以下では、「電子化テキスト」（機械可読machine readableテキスト）を簡単にe-textと言う。

従来のe-textは、特定メーカの機械、特定ソフトウェア（例えばデータベース）が前提になるなど、計算機環境（ハードウェア、基幹ソフトウェアのOS）に強く依存したものが少なくなかったがこれは、計算機環境が多言語を自由に扱うには余りに非力だった為である。計算機環境の国際化対応が一斉に進んだのは、ここ数年の事に過ぎない。

こうした環境面での国際化と、国際コンピュータネットワークthe internetの爆発的な拡大に伴い、e-textの国際的共有も現実的になって、既にe-textの公開・頒布・検索サービス（全て無償）を世界的に提供する大学・研究機関も多数現れている。

分科会C「電子化テキストの国際的共有」は、この様な世界的趨勢を踏まえて、主として日本語の電子化テキストを国際的に共有する為に、交換の為の環境整備の問題と、実践報告とで構成する事とした。

但し、internet 上の電子化テキスト共有の実践例に就ては、未だ機が熟さない為か、応募を見なかったのを遺憾とするが、日本語文献の交換で常に問題となる文字コード系(coded character set)に就ては、芝野耕司、池田証壽により、それぞれの専門的な立場からの詳細な報告が行なわれた。特に 1995 年が日本工業規格(JIS)の漢字符号系(JIS X 0208/X 0212)の改訂年度に当り、芝野パネラーがその改訂委員会主査という事もあって、熱心な質疑が行われた。又、電子化テキストの既存の交換形式の問題点に就て、家辺勝文と豊島正之が、報告を行なった。

以下は、当日のセッションでの議論と、その後の進展を含めて、電子化テキストの現在の問題を概観する。

2 文字コード系の問題

文字コード系とは

計算機やデータ通信で文字のデジタル表現として使われる「文字コード系」(コード化文字集合 coded character set)は、相異なる文字にそれぞれ別の点(コードポイント)を与えて、文字を一意に指定出来る様にした一種の文字表である。各国工業規格(eg. 日本の JIS)がそれぞれ自国分の文字コード表を制定し、ISO(International standardization organization、国際標準化機構)が、それらに登録番号を与えて国際的に管理している。これらの各種「情報交換用符号」を取り替えて使う方法に就ては、ISO 2022「ISO 7-bit and 8-bit coded character sets - Code extension techniques」が厳密に規定している。この制定経緯に就ては芝野稿に詳しい。

ISO 2022 は、アルファベットの様な文字数の少ない文字表と、漢字の様に桁違いに文字数が多い表とを、一文字あたりのビット数を可変して共存させつつ、効率よく切替える様巧妙な工夫がしてある。上記各種の「文字コード表」は、全てこのメタ規格に従って設計されているので、これらを逐次差し替えて、同一文書内に英語・ドイツ語・ヘブライ語・日本語を共存させる等も可能である。

もっとも、この ISO 2022 方式は、巧妙ではあるが、その分完全な実装が極めて困難で、数個以上のコード表を差し替えながらの多言語用運用等現実には殆ど行なわれず、国際的な文字コードの共存は、絵に描いた餅である。この為、文字コードの可変ビット長を 16 ビットに固定して世界のコード表を集大成し、その結果「差し替え」を完全に廃止して容易な運用を目指した ISO 10646-1「Universal multiple-octet coded character set, Part 1 : Architecture and BMP (Basic Multilingual Plane)」が、1993 年 5 月に制定された。

code point と glyph の分離

ISO 10646-1 BMP によって、国際共通の文字コード系という宿題は一応達成されたが、この規格が、漢字部分に就ては、中国・台湾・韓国・日本の四コード系の統合(Han-unification)を行った為、点画の小異を無視して無理やり統合していると批判する意見がある(日本電子工業振興協会(1993))。ISO 10646 の規格票は、統合漢字に就ては上記四者の字

形(フォント)を併記しているが、これを真に統合出来ていない証拠と指弾する向きもある。

しかし、本来、コード表上の位置(code point)とフォントの字形デザイン(glyph)は別の問題で、コード表が字形を決める訳では無い[注1]。現に、ラテンアルファベットで広く使われているASCIIコード表のフォントは、軽く数百はあり、ASCIIというコード表はそれらを統合(unify)した存在であるが、これに就て(フォントの豊富さが歓迎される事こそあれ)上記同様の批判が行なわれる事等は無い。

文字表と coded character set の差

従来の計算機環境のcharacter setは皆、ラテンアルファベットや平仮名の様に閉じた(要素数が固定された)文字表に基づいていたが、漢字は文字表として閉じていない点で、全く異質である。

これは、新しい漢字が今後もどんどん作字されるという意味ではなく、調査・研究の進展、精度の差によって、今まで同字として来た字を別字として区別する必要が生じるであろうという意味である。

確かに、アルファベットも、アクサン・気息記号等の補助記号(diacritical marks)を組合わせれば更に文字種が増える(cf. ア行の仮名に濁点を付ける例)。しかし、その補助記号を加える操作は、必ず別字を作り出す操作である事が了解済である。つまり、要素とその組合せの相互の弁別性は予見出来ている。これが「文字表」の最重要の特徴で、文字コードは、従来この特徴を暗黙に仮定して作られて来た。

ところが、漢字ではこの前提が全く成立しない。現行JIS X 0208/X 0212でどの程度日本語関連文献の翻刻が行えるかは池田稿に詳しいが、「JISに無い字」が存在するという事自体は、直ちにJISが欠陥文字表であるという事には結び付かない。そもそも元の文字表自体が閉じていないからである。

点画の小異を補助記号の一種と見なすとしても、アルファベットの場合と異なり、果たしてそれらの付加が相互に弁別的な別字を作り出すか否かは不明である(「土」に「、」を加える操作は「土」の異体字を生み出すだけなのに、「大」に点「、」を加えた字はもはや「大」の異体字ではない)。つまり、要素の増減が別字・異体字を作りだすか否かに就て、文字表のシステムとしての了解が無い。

従って、漢字のコード表には、従来の閉じた文字表に基づくコード表とは多少異なったアプローチが必要である。排列は、その典型例で、従来の閉じたコード表(character set)では、暗黙の内にそれを排列済のコード表(ordered character set)である事を前提にしていた。これは勿論、要素の増減が無いからである。しかし、漢字コード表では、一旦ISO 10646-1(BMP)の様に部首画数順に漢字を排列して仕舞うと、その排列順を維持したまま一字追加しようにも追加しようが無い事になる。

この事情は、ISO 10646-1に限った事ではなく、所謂「JIS漢字」であるX 0208/ 0212も同様であり、一字の増減が全く異なるコード表を生み出す。この点で、1995年に予定されているX 0208/X 0212の改訂作業は、もし同一の表の改訂という前提を崩さないのであれ

ば、字の追加・削除は全く不可能になる筈である。[注2] これに対しては、現行 JIS 規格票の不備すら改訂出来ずに固定して仕舞うという難点と、未登録の字の追加を希望する意見があるが、文字表の同一性の保存をより重視する芝野の表明に従うべきであろう。JIS X 0208(当時 C 6226)の 1978 年→ 1983 年改訂が、世上のプリンタに挙って「新 JIS」・「旧 JIS」選択のディップスイッチを用意させ、ISO 2022 に revision control sequence の新設を余儀無くさせて、世界中に多大な迷惑を掛けた愚を繰り返すべきではない。

尚、漢字仮名交じり表記の文字列の排列には、単なる漢字の排列とは別の合意(標準化)が必要で、現にその策定が行われている。[注3]

3 テキスト形式の問題

e-text のデータ形式の互換性の問題は、計算機の分野では古くから熱心に議論され、標準化も進んでいる。

家辺稿の言及する構造記述言語 SGML (ISO 8879, JIS X 4151, Goldfarb (1990)) はその成果の一つで、主として階層的(hierarchical)なデータ構造記述の互換性の確保を目的とした国際規格である。言語テキストの交換形式として 1994 年に最終版が出た TEI (Text Encoding Initiative) も、SGML の上に構築されている。

TEI は、ACH (The association for computers and the humanities), ACL (The association for computational linguistics), ALLC (The association for literary and linguistic computing) が運営委員会(steering committee)を組織した大規模なプロジェクトで、その成果は、TEI-P 1 (TEI Public draft 1, 1990)、TEI-P 2 (1992-、未完のまま放棄)、TEI-P 3 (1994) として公表された。

3.1 SGML の意味付与の限界

SGML は、言語とはいっても意味論を持たず、構造の階層関係(構造的意味)以外は、テキストに何の意味も付与しない(出来ない)。これは、SGML が「意味は構造によってのみ与えられる。要素の持つ値(即値)は単なるラベルに過ぎず、意味には無関係である。」という、構造と値(具体的には文字列)とを完全に分離する「データ抽象」モデル[注4]を基本にしている為である。つまり、SGML 内で与えられる「意味」は「位置としての意味」だけである。

この為、SGML の実際の運用に当っては、情報交換当事者間での意味に関する合意が必須で、SGML 上に構築された SGML のアプリケーションである、ハイパーテキスト(hypertext) 記述体系の HyTime (ISO 10744) や TEI は、SGML の特定記法に対する「意味付与の約束・合意」に過ぎない。

つまり、SGML の様なデータ構造記述の国際規格さえあれば自由で安全な互換性のあるテキスト交換が出来るというのは幻想に過ぎず、実際にはその運用に関する別の合意形成が必要である。

3.2 SGMLでは記述出来ない構造

値の中の構造

SGMLは、上記の様な構造と値とを分離するデータ抽象モデルを取っている為、値自体が構造を持つ様なケースが記述出来ない。

意味関係の機能不在のSGMLの中で、僅かにユーザに与えられた意味付与用のラベルがattributeで、階層的な要素(element)のそれぞれの階層で、elementにatttributeを与える事が出来る。

問題は、このattributeが単なる文字列のラベルで、その中に更に構造を作る事が不可能な点である。例えば傍訓に更に施された声点・不濁点等は容易には表現出来ない。当然、attributeの相互依存関係(hierarchyの一種)も書けない。この為、attributeの共起制限の方法も無い為、相互に矛盾したattributeを書いても、チェックの方法が無い。

従って、SGML文書には、SGMLパーザだけではなく、attributeのチェックの為のvalidatorが必須である。実際、TEI (P3) 中には、そこかしこにこのvalidatorに頼った記述があるが、汎用validatorの実装例は、未だに一つも知られていない。汎用validatorの為には、attributeの意味仕様記述が可能な言語が必要であり、SGMLは明らかにそれを満たさないから、SGMLだけに依存しているTEI (P3) の範囲でvalidatorが実現出来ないのは不思議は無い。

要素と属性

SGMLが属性(attribute)を振る事を許すのは構造中の独立の要素(ELEMENT)だけなので、これでは、要素の一部のみに付加情報が与えられている場合(eg.送り仮名・捨て仮名、部分差声、見せ消ち、部分引用)の表現方法が無い。勿論、その部分を強引に一要素として立てれば可能だが、しかしその様な便宜的な要素を乱立させると、本来の「構造記述」は殆ど有名無実になる。

SGMLモデルは、端的にいえば「要素の画定」と「属性の範囲」とは常に一致するという前提を立てており、一方、実際の(日本語)文献データでは、この前提が成立しない事が少なくないのが問題な訳である。[注5]

この他、家辺稿4.2の指摘する様に、SGMLは、上下階層的なhierarchyを成す構造しか記述出来ないので、本質的にhierarchicalではないテキストは記述のしようが無い。典型的には、掛詞・序詞、渡りゼリフ、異文注記、異なるよみの併記(「両点」)、再読等である。又、SGMLは、テキストに暗黙の内に線条性を仮定しているが、勿論これも、漢文訓読での反読や「両点」字、複数時加点、対話筆記(同時発話「クロストーク」を含む)等には通用しない。

TEIは、特にこの非線条的なテキスト表現に意を用いているが、属性として与えた順番に基づいて実質的には複数の線条系列(thread)の読み方を指定しているだけなので、前述の様に順番の欠番・重複などが生じても、チェックの方法が無いという問題が残っている。

構造記述だけでは足りない

家辺の指摘に詳しいが、そもそもテキストデータは、構造(構造的な意味、即ち構造中での位置)だけ記述していれば済むという物ではない。家辺の指摘にある様な、紙上での位置(頁組はその一つ)、効果、補入・訂正、欠損・虫損などの表現も、SGML の様な構造記述言語にはおよそ不向きである。

この他、SGML が規格から落とした問題も大きい。例えば、SGML は、文字コード系(cf. 2 節)記述に限界を抱えている[注6]。SGML で文書の互換性を保持しながら複数文字コード系に対応するのは困難で、ISO 10646-1 の様なシームレスな単一国際コード系を用いるのが実は最も簡単な解決法である筈だが、TEI-P 3 は、表記系記述仕様 (writing system declaration) という別のアプローチを取って、SGML アプリケーションとしての多言語対応を試み、結局、問題を解決しないまま終っている[注7]。

4 質の維持と権利関係

e-text の公開方法には、大別して

1.商品として (CD-ROM 等で) 売られているもの

2.巨大な集成で、テープ等で頒布されているもの (corpus)

3.anonymous FTP 等により自由頒布されているもの (以下「online text」)

がある。大雑把に言って

1.商品は、著名なテキスト (eg. Bible, Chaucer, Shakespeare) が多く、且つ質が高い。本文校訂自体も、殆ど権威ある (authentic) 校訂本を底本にしている。

2.corpus は、著名なテキストというよりは、多種多様なものを豊富に集成したものが多く、中には「校訂」とは無縁のテキストデータ (eg. 談話記録、規格本文、公文書類) もある。

3.online text にも、著名なテキストが多いが、OCR (文字読取機・読取ソフト) で入力されたもの (scanned text) が多く、中には校正が行き届いていないものもある。底本も、殆どが通行本で、権威ある校訂本を底本にしたものは珍しい。

つまり、端的に言って、online text は質が低い。これは、作成者の問題というより、次の二つの online text 特有の事情による。

1.online text は、自由に頒布可能 (フリー) でなければならない。

2.online text には、特定プラットホームの前提があってはならない。

4.1 フリーである為の制約

商品はもとより corpus も、頒布申込時に権利関係に就ての合意書に署名を求められるのが普通で、契約に基づいて頒布される。ところが、online text は黙ってコピー (FTP) して来る事が前提なので、契約が無い。従って、自由に頒布可能 (フリー) ではない online text というものは存在しない。

フリーである以上、著作権の生きている書を底本にする訳には行かない。従って、(著作

権者が許諾しない限り）現時点で最高の権威のある校訂本を底本にしたonline text等というものは、原理的に存在しようが無い。

例えば、販売されている聖書のe-textは、ヘブライ語版（Biblia hebraica Stuttugartensia）、ギリシャ語版（Nestle-Aland）、ラテン語版（Vulgata/Weber）といずれも現在最高の権威ある底本が契約に基づいてe-text化されたもので、学問的な利用に何の不安も無い。英国で近時刊行開始されたe-text出版シリーズは、Chauser、Dickens、J. Austin、Wordsworth、T. Hardy等を含むが、いずれも有数の権威あるテキストが底本である。

これらの底本は何れも著作権が生きているから、無断でonline化すれば完全に違法で、かと言って著作権者から複製許諾を得てonline textにするのも望み薄である。Shakespeareを例に取れば、フリーのe-textは、素性の良く分からない20c前半の全集本（著作権は切れている）を底本にしており、文献学的な研究にはおよそ不向きの代物である。

4.2 可搬性の為の制約

校訂本は、とかく細かな校訂注釈用の記号diacritical marksが入りがちである。しかし、アクサン、下線、ルビ、marginalia等をプレーンテキストに書込むのは困難である。

Macintosh-OS、Windowsの様に、特定の動作環境（プラットホーム）を前提にしたものを作るのなら、これらは比較的自由に出来る。しかし、internetからアクセスされるonline textは、ありとあらゆる機種からのアクセスが可能で、どの様な環境にも適応する事（可搬性）を前提にしているから、必然的に、こうした付加情報・注釈は施せない。

この点を打開する為に提唱されているのが、前節で見たSGMLのアプリケーションとしてのコーディング規格TEIであるが、当然SGML自体の問題を抱えている。

4.3 online softwareとの対比

online textの相対的な質の低さは、この様にinternetで自由に公開されるという特性に根差した本来的なものなので、中々改善が難しい。

一方、同じくinternetで自由公開のソフトウェア（以下online software）を見ると、この同じ特性が全く逆に働いている。即ち、

1. online softwareはフリーでなければならない。
   →故に、著作権者による公開を経た、オリジナルなものでなければならない。
2. online softwareは極力機種依存しないものでなければならない。
   →故に、（よいものならば）爆発的に広がり受入れられる。

つまり、online softwareの普及には標準化は不要で、全てがde facto standard（他に競合するものの僅少な事実上の規格）であり、広範な頒布を可能にしているのは、ひとえにその仕様の尊重と、質の高さである。

そもそも、現在のthe internetの根幹を支えている通信経路情報（RIP）、機関ドメイン名情報データベース（DNS/BIND）自体、どちらもフリーなソフトウェアとして実装されている約束事（プロトコル）であり、特定メーカの製品でもなければ、ISOの決めた国際規格

でもない。全世界がこれに従っているのは、単に、比肩する性能と可搬性を持つ代替品が存在しないからに過ぎない。

つまり、online text と online software のこの差をもたらしたのは、偏にオリジナリティの有無である。

オリジナルな online text

現状の欧米の online text は、OCR による scanning のみによるテキスト化が未だ少なくないが、幸い、日本では（漢字 OCR の質の低さの怪我の功名で）、複製・写真から厳密に翻刻を起こした[注8]、オリジナルで且つ良質のテキストファイルが少なからず存在する[注9]。

権利の尊重と共有

internet という情報通信サービスでは、複製・再複製は自由に許しながら著作権は主張して、同一性の保持・濫りな改変の禁止や、再複製の妨害の禁止を行おうという、所謂 GNU 式（Free software foundation）の公開が盛んで、ここでは、情報の共有が、権利の尊重と全く矛盾無く結び付けられている。自由に使ってよいという事と、自由に自分のものとしてよいという事とは異なるのであり、この点で、e-text 化、或は e-text の公開が公開即著作者の権利の喪失を招くとする漠然とした不安[注10]は、払拭されてよい。

internet 上の情報源を次々に渡り歩く gopher/WWW サービスを（無償で）提供する機関も激増し、e-text 検索を提供する所も珍しくない。こうした公開サービスは、単純なファイルコピーである FTP より遙かにホストマシンに負荷を掛けるので、かなりのハードウェア的資源が必要になり、機関全体からの後援が必要である。

現在、こうした internet 情報関連では、こと言語データに関してだけでも、日本は殆ど貰う一方の「輸入超過」である。この不均衡は、個々の研究者の個人的な努力だけで解決する問題ではなく、そろそろ「情報輸出」の為の組織的な支援活動が望ましい時期に来ている様に思われる。

引用文献

出雲朝子・豊島正之（1993）「玉塵抄と計算機 II」（文部省科学研究費報告書）
豊島正之（1994）TEI P 3 に就て（情報処理語学文学研究会会報 15）
日本電子工業振興協会（1993）「未来の文字コード体系に私たちは不安をもっています」（日本電子工業振興協会パンフレット）
Goldfarb, Charles（1990）The SGML handbook（Oxford UP）
TEI-P 3（1994）The text encoding initiative — public draft 3,
TEI-P 3 の印刷物形態（上下 2 冊、1289 ページの大冊）の入手は、〒 263 千葉市稲毛区弥生町 1-33 千葉大学文学部 土屋俊方「TEI 日本委員会」宛て。ファイル形態（約 70 ファイル、3.7 MB）の入手は、internet 上の anonymous ftp で sgml1. ox. ac. uk、ftp. ifi. uio. no から。

本稿は、文部省科学研究費による研究の一部（研究代表者出雲朝子）を含む。

注1　この点に就ては、JIS X 0208/ 0212 の改訂時に、誤解を解く為の一層明確な記述が行なわれる見込である。

注2　実際の文字コード登録作業を行う ECMA の規準も、全く同じ立場である。

注3　日本工業規格（JIS）「日本語文字列照合順番」（未刊）。

注4　データ抽象は、計算機言語の世界で一貫して追求されて来たモデルで、その発展形が「オブジェクト指向」OOP（Object Oriented Paradigm）である。

注5　具体例を含む詳細は、出雲朝子・豊島正之（1993）§2.3

注6　出雲朝子・豊島正之（1993）p. 84-。

注7　豊島正之（1994）

注8　複製・写真自体には（こと、文献の複製に関する限りは）著作権の問題が殆ど無い。これらの複製・写真は、原本に限りなく近い事が理想である筈で、原本に忠実であればある程、著作権法の言う独創的な「思想・感情の表現」とは無縁の筈である。特に高価な複製本は「原本に限りなく近い」事を標榜する事があり、これは自ら著作権の主張を放棄していると言ってよいから、安んじて底本に用いる事が出来る。

一方、他研究者による翻刻文を底本に用いるのは問題で、特に、複製が未公開で翻刻のみが存在するもの（例、名語記、訓点資料類）は、共有出来ないし、状況によっては e-text の作成自体が違法な複製と見なされる可能性がある。

注9　木越治（金沢大学）監修の一連の秋成・綾足作品 e-text が、JALLC（情報処理語学文学研究会）から頒布されている。連絡先「情報処理語学文学研究会」、e-mail：HTE 70552@PC-VAN、〒101 東京都千代田区神田神保町 3-27、共立女子大学日本文学研究室内（内田保廣）宛

木越の様な試みは、良質の online text のあるべき方向を示唆する、優れた先達と言えよう。

注10　確かに一昔前には、この様な不安を裏付ける様な事例があったのは事実であるが、「公開したものには権利主張すべきではない」等という非常識な見解は、もはや殆ど淘汰された様である。

――北海道大学文学部助教授――