古文書データベースにおける日本語データ処理の諸問題

Some Problems with Japanese Data Processing in the Ancient Document Database

高井正三
Shoso Takai
富山大学総合情報基盤センター
Information Technology Center, Toyama University
〒930-8555 富山市五福 3190
3190 Gofuku, Toyama 930-8555 Japan
takai@cns.toyama-u.ac.jp

概　要

1994 年以降，日本現存朝鮮古書を対象とする古文書の書誌情報データベースを構築してきた．今回このデータベースの冊子体目録を作成するに当たってそのマスター・データを分類して並び替える必要がでてきた．入力された日本語データ処理を行うに際して，コード変換，文字化け，国語辞書用ソーティング，外字処理，縦書き印刷，双行印刷など，解決しなければならない諸問題が発生した．本稿ではこれらの諸問題に対するいくつかの解決策と今後の課題を提案する．

キーワード

日本語コード　日本語データ処理　国語辞書用ソーティング　外字処理　コード変換　古文書データベース

１．はじめに

筆者は富山大学人文学部教授の藤本幸夫氏が 30 年の歳月をかけて，1910 年以前の朝鮮古書を対象に全国から収集してきた「日本現存朝鮮古書」に関する 28 項目にわたる書誌情報約 15,000 件を，データベース化し，インターネット上に公開するため，データ入力，データベース・システムの開発に 1994 年から取り組んできた．

この時点まで入力されたデータは 2,682 件 3.74MB，字数にして約 200 万字，全体の 95%以上は漢字が占めている．このデータは当初，MIFES エディタ，一太郎日本語ワード・プロセッサ，ATOK 日本語 IME を使って入力され，後に秀丸エディタで編集，追加が行われた．IME は後継版 ATOK を使用してきた．

今回，平成 17 年度科学研究費補助金（研究成果公開促進費）学術図書を申請するに当たって，冊子体目録の見本版を作成した．書誌情報の分類，LaTeX ソースの生成，索引作成など，日本語データ処理では，コード変換をはじめ，国語辞書用ソーティング等，様々な問題に遭遇した．ここではその問題解決の方法をいくつか提案する．

２．既存データの漢字データ処理

２．１　PL/I から Perl へ

データベースのマスター・データは，当初 IBM PC で入力し，Mainframe System 上に転送して，IBM 漢字コードに変換し，PL/I 言語で処理していたが，Mainfram System から UNIX サーバに替わって，文字列の処理に便利な Perl 言語を使用するようになった．これに伴って日本語データは IBM 漢字コードからシフト JIS(SJIS)コードに代わり，データ処理には EUC (Extended Unix Code)を使用するようになった．

日本語データ処理では文字列の照合，置換，分類などが主な処理となるので，インデックス(index)関数やサブストリング(substr)関数が使用できる言語が必要で，旧来は PL/I が，現在は Perl，C 等のプログラミング言語が使われるのが一般的である．

２．２　文字化け

（１）メタ文字

Perl では，EUC データは問題なく処理できるが SJIS コードでは，漢字の中にメタ文字（バック・スラッシュ付きの制御文字），特にバック・スラッシュ「＼(日本では￥を割り当て)16 進表記 0x5C」と同じコードを有する漢字があり，これが Perl データ処理の不具合を引き起こしている．「ー（0x815C)」「能(0x945C)」「表(0x955C)」などの文字を扱うときは，そのコードが(¥)を含む文字であることに注意が必要である．以後のデータ処理が暴走する．図１の SJIS

を Perl 処理し，pLaTeX2e の¥item[･･･]を生成出力した結果（図２）である．

,"表紙左肩墨書「嘉林世稿」。該書は徐宗泰序･「雲江遺稿」･「竹陰遺稿」を収める上冊一冊のみを存し、「近水軒遺稿」･「附録」を収める下冊を失っている。刻工名は東大本と一致しない所がある。それには、筆者の見落し、版心部が磨耗して刻工名が見えない、後の補刻等の原因が考えられるが、調査カードのままに記した。",

図１　SJIS の本文（CSV 形式の一部）

\item[註　記　]表紙左圭紙左肩紙左肩昧左肩墨左肩墨掌肩墨書肩墨書→墨書「墨書「穎書「嘉書「嘉卵「嘉林「嘉林訊嘉林世嘉林世催林世稿林世稿・世稿」世稿」△稿」。稿」。各」。該」。該竣。該書。該書・該書は該書は什書は徐書は徐書は徐宗は徐宗堆徐宗泰徐宗泰徐宗泰序宗泰序　泰序・泰序・・序・「序・「汚・「雲・「雲孔「雲江「雲江・雲江遺雲江遺浩江遺稿江遺稿‐遺稿」･･･める上める上罪る上冊る上冊磯上冊一上冊一阪冊一・・・

図２　pLaTeX2e の¥item[･･･]生成結果

（２）コード変換による文字化け

SJIS コードのデータ図３を nkf 変換，秀丸で EUC コード変換したものを次に示す．図４が nkf（ネットワーク用漢字変換フィルター）による変換結果，図５は秀丸エディタによる変換結果である．

０２朝鮮【N柳琴／リュウキン】編　朝鮮【N李徳懋／リトクボウ】･･･
１１上部下向黒魚尾　下部長方形黒釘　魚尾下「韓客巾衍集」
２２毎表紙題簽下部「【Ｉ清愼勤／セイシンキン】」（二．二×一．一　陰刻朱印）･･･
２４表紙左肩･･･附す。巻一首題下「李評青／潘評朱」。

図３　SJIS オリジナル

０２朝鮮【N柳琴／リュウキン】編　朝鮮【N李蛇譟織螢肇・椒Ａ曄レ・・・
１１上部下向　・竑・__蕊・絞・・・・・ゝ竑・次峪攀匐吁ьI検・
２２毎表紙題簽下部「【Ｉ　へ洞弋織札ぅ轡鵐⑤鵝曄廖米鵝テ鵝澎譟グ譟￣・鐙覿・・・・・
２４表紙左肩･･･附す。巻一首題下「李評　堯厨・昭襦廖・・・・

図４　nkf による EUC コード変換結果

０２朝鮮【N柳琴／リュウキン】編　朝鮮【N李博ヨ譟織螢肇・椒Ａ曄レ・・・１１上部下向螺ゥ竑・__蕊・絞・・・・ゝ竑・次峪攀匐吁ьI検・２２毎表紙題簽下部「【Ｉ浮ヨ洞弋織札ぅ轡鵐⑤鵝曄廖米鵝テ鵝澎譟グ譟￣・鐙覿・・・・・２４表紙左肩･･･附す。巻一首題下「李評爺汾~・昭襦廖・・・・

図５　秀丸エディタ V4.18 の EUC 変換結果

JIS 区点 89-92 相当の SJIS コード(MS-IME 漢字３)が，EUC コードの SS3(Single Shift3)の３バイト・コードに対応しているためか，変換後に文字化けが発生する．文字列の中から JIS 第１，第２水準，または置換文字列@99999 に置換してから，SJIS→EUC 変換をすべきであったが，既存データの量が膨大で，総てのシステム外字，IBM 拡張漢字を調査して変換する時間的余裕が無かった．

よく現れた文字化けの漢字を図６に示すが，右側の旧字体に割り当てられた SJIS 外字コードは，EUC コードに変換ができなかった．そこで予め対応が解る旧字体のシステム外字域の漢字は左の新字体で置換してからデータ処理を行った．

とく	徳 0x93BF	←	德	SJIS 外字	0xFABA
せい	青 0x90C2	←	靑	SJIS 外字	0xFBF2
せい	清 0x90B4	←	淸	SJIS 外字	0xFB43
こく	黒 0x8D95	←	黑	SJIS 外字	0xFCB4

図６　EUC コード変換できない漢字の例

漢字データの入力については，データベース構築時に一括して旧字体に変換するので，入力は常用漢字とすることしたが，外字登録が 700 字以上になってコード表参照に時間が掛かり，ATOK IME でも拡張漢字が入力できるようになったので，結局これらのシステム外字域の漢字が入力されてしまった．マスター・データの中に相当量のシステム外字等があったので，EUC 変換をあきらめて，SJIS のままデータ処理を余儀なくされた．

３．国語辞書用ソーティング

文書目録の入力データは，最近入力のものまでをすべてデータベースに取り込み，これに分類コードを分類表（漢籍の分類）に従って付与し，各分類内では，国別，次いで書名のアイウエオ順に並び替え，同版の場合は所蔵者の北から南へ並べる．したがって分類の順序は以下の通りとなる．

１）文書の分類コード順（漢籍の分類コード順）
２）国別順（韓国，中国，日本，・・・の順）
３）書名のアイウエオ順
４）所蔵者の北から南への順（所蔵者コード昇順＝小→大の順）

この時，書名については，後の索引にも使うので，国語辞書の様にソーティングして欲しいとのことで，これがまた大変であった．辞書用ソートのパラメータがあったが，われわれの要求を満たしてはいなかった．そこで次の例題に示すように，書名，撰者についてはカタカナ表記を取り出し，その濁点，半濁点を辞去した書名，撰者を別フィールドに生成し，このフィールドを第１の分類キーに，本来の書名，撰者フィールドを第２の分類キーに設定してソーティングすると国語辞書用の分類になる．

この変換テーブルを作るときもカタカナの「ソ」がメタ文字を含むので，16 進データとして定義する必要があった．

図７と図８はソーティングの例である．

とうさ	とうさ	踏査
とうざ	とうさ	当座
とうし	とうし	投資
とうじ	とうし	冬至
どうさ	とうさ	動作
どうざ	とうさ	同座
どうし	とうし	同士
どうじ	とうし	同時

図７　第１フィールドによるソーティング

とうさ	とうさ	踏査
とうざ	とうさ	当座
どうさ	とうさ	動作
どうざ	とうさ	同座
とうし	とうし	投資
とうじ	とうし	冬至
どうし	とうし	同士
どうじ	とうし	同時

図８　国語辞書用ソーティング結果

４．外字処理

データの入力ルールは次の通りである．

（１）SJIS に無い文字

・旧字体を用いてもよいが，JIS 第２水準までとする．

・「余」，「芸（ウン）」等，旧字体に変換されては困る漢字がある時は，原稿下部余白に明記しておく．

・入力できない漢字のうち，京都大学漢字典に記載されているものは，記号「@」と「康熙辞典コード番号（５桁）」を連結して入れる．

・「康熙辞典コード」にない漢字は，記号「@」の後ろに，朝鮮固有外字として 60000 台からの連番で数値を入れる．この連番は，共通の外字管理表を参照して，新規のものは新しい番号を発行する．

現在まで SJIS に無い文字は 1,079 字を登録し，コード入力を行っている．

（２）朝鮮固有外字

朝鮮固有外字（異体字を含む）等は@60000 台から順に外字原簿に登録し，@連番コードで入力する．現在まで 114 字を登録した．

（３）ハングル

ハングルは古文書の中ではそれほど多くないので@70000 台から入力し，最終的な印刷ではハングル文字に置換する．入力はアレハ・ハングルというアプリケーションを使う．現在まで 183 字を登録した．

（４）記号類

刻手名等を表す記号または記号に似た文字は，特殊外字として@90000 台から登録し，@連番コードを入力する．現在まで 57 字を登録した．これは殆ど記号に近く，「外字・TrueType フォント エディタ」TTEdit を使用して作成した．

５．冊子体目録の作成

５．１　縦書き印刷

古文書の書誌情報を冊子体目録や Web 上での表示を縦書き印刷にしたいのは漢籍研究者の常道であるらしい．

本学人文学部の小助川教授に，縦書きでかつ双行印刷などが可能なツールを教えてもらった．大阪大学の金水敏氏が作成した，「訓点資料（漢文訓読文）」用の LaTeX Style ファイル（マクロ・パッケージ）が使えるとのことで，これを使って冊子体目録を作成することにした．図９はこの現物の一部でページ数 1,234 になり，索引を入れて約 1,300 ページの冊子体目録（２分冊）を完成した．

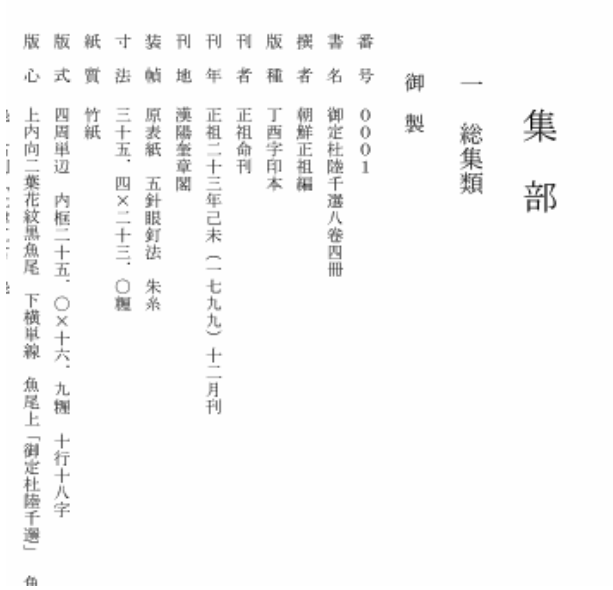

図９　pLaTeX2e による縦書き印刷

５．２　双行印刷

縦１行に小字で２列の文字を入れる双行は，この訓点マクロの¥sougyou{XX}{YY}を使用した．片側のみの場合は¥sougyou{XX}{　}の右の項を空白とした．

５．３　連数字

連数字（縦中横）の数字印刷には，LaTeX の連数字マクロを利用したマクロ¥Rsj を導入した．

```
¥def¥Rsj#1{¥rensuji*[c]{#1}}
```

５．４　索引の作成

索引は書名と撰者で，LaTeX の索引を取ってはみたが，これを国語辞書用にソーティングするのが大変で，結局は別名でファイルを作り，目録の分類同様に国語辞書用ソーティングを行って，文献の番号を付与する方法を採用し，索引を作成した（図 10，図 11）．

５．５　その他の記法

「＝＝」という表記は，「･･･」の縦書き表記に置換する．実際は pLaTeX の「$ ¥cdots $」に置換することとした．

【ウ】

迂齋集　0272,0273,0274,0275
雨念斎詩鈔　0277,0278
・・・
雲巖逸藁　0281
雲江遺稿　0032,0033
雲岡集　0282
雲谷集　0627,0628,0629,0631,0632,0633

図 10　書名索引（アイウエオ順）

【オ】

王安石　0108,0109,0110,0111,0357,0358,
0359,0360,0361,2146
王欽臣　0227,0228,0229,0230,0231,0232
王鴻儒　0901
王靜　0909
王世貞　0083,2606,2607
王鈍　0027
翁方綱　2184

図 11　撰者索引（アイウエオ順）

６．各国 IME による連続漢字入力の方法

効率的な日本語入力方法，特に連続した漢字の入力を効率的に行う IME の登場が待たれるが, 本場の台湾（繁体字）や中国（簡体字）では，それぞれの国の特徴ある IME が用意されている.

（１）日本

・ATOK･･･ジャストシステムが一太郎用に発展させてきた日本固有の IME で結構使いやすい. SJIS, JIS, Unicode もサポートされている.

・MS IME 2003･･･SJIS と Unicode をサポートする Microsoft の IME であるが, 辞書がこなれていない.

（２）韓国

Microsoft IME 2003 が国際版の Office に付いてくる．一般的には McCune Reischauer 方式のローマ字入力で漢字変換するのがベターなようだ.

（３）台湾（繁体字）

・Microsoft New Phonetic 2002a（音声：読み）

これは発音を頼りに該当する漢字を選択入力する IME である.

・Microsoft New ChangJie IME（部首合成）

台湾ではキーボードに漢字の部首を当てはめて, これを合成する方法で入力し,妥当な文字に変換していく IME で，スピードは抜群である.

・櫻花（繁体字＋日本語＋ひらがな，カタカナ等）

これは台湾中央研究院計算中心で紹介された IME だが，日本語の他，ひらがな，カタカナも入力できるようで,「櫻花（さくら）輸入法」と言っている.

（４）中国（簡体字）

・Microsoft Pinyin IME

Pinyin のローマ字入力する IME で該当する同音異義語を表示し，選択する方法である.

７．今後の課題

OS レベルでは Mac OS X，Linux で Unicode が標準化され，Application ソフトウェアの対応が待たれている．Windows でも MS-SJIS 以外に Unicode が提供され，Office 系ソフトウェアで MS 明朝，MS ゴシックのフォントがサポートされるようになった．古文書のデータベース化における旧字体漢字の使用環境が整ってきた.

また, Flash や Java で Unicode の符号化形式 UTF-8 がサポートされているので，今後はセキュリティに堅牢な Java をベースに Web データベース・システムを構築していくのが Best Solution になるように思われる．現在 Java システムの開発に挑戦中である.

謝辞

筆者等がサービスしている日本現存朝鮮古書データベースの構築に当たって支援を受けた同僚の布村紀男助教授. 朝鮮古刊本総合目録の作成に当たって, 分類方法や記法，索引の採取など，本データベースに全体に関する情報の提供を受けた人文学部の藤本教授．これらのデータ入力に献身的な努力をしてくれた越野, 洲崎, 葉山, 木戸, 竹澤の女性スタッフ. ここに記して感謝の意を表する.

参考文献

[1]The Unicode Standard Version 4.0, The Unicode Consortium, Addison-Wesley, 2004

[2]高井正三，朝鮮古書データベース蓄積・提供用旧字体および朝鮮固有外字の整備に関する研究, 平成 10 年度～平成 13 年度科学研究費補助金（基盤研究(C)(2)：課題番号 10680401）研究成果報告書, 1-142，2002

[3]高井正三，布村 紀男，日本現存朝鮮古書データベース・システムの構築, 学術情報処理研究, No.5, 87-90，2001

[4]高井正三，布村 紀男，日本現存朝鮮古書データベース・システムの構築方法，情報科学フォーラム 2003(FIT2003)論文集，D-26，57-58，2003

[5]高井正三，Unicode4.0 解説，富山大学総合情報基盤センター広報，Vol.2，96-104，2005

[6]UNIX と X Window の日本語環境－日本語入力・表示から印刷まで－, 林秀幸著, ISBN4-526-03769-9, 日刊工業新聞社，1995

[7]Unicode 標準入門，トニー・グラハム著，乾和志・海老塚徹訳，関口正裕監修，ISBN4-7981-0030-7，翔泳社，2001

[8]漢字文献情報処理研究，漢字文献情報処理研究会編集，第５号，ISBN4-87220-083-7，好文出版，2004