琉球冊封使関係資料情報化の課題

岩井 茂樹：京都大学人文科学研究所

夫馬進教授を代表者とする計画研究「環東シナ海地域間交流史ッッ中国福建を中心として」では，プロジェクトの一環として，使琉球録の全文データベース作成を試みた。報告者個人としても初めて取り組んだ漢籍の全文電子テキスト化の経験をもとに，現時点で，漢籍の電算処理が抱えている困難を指摘するとともに，成果の公表にさいして選択しうるさまざまな方法を紹介する。

1　歴史的資料電子化とその目的

紙に印刷されたり，書写されたりしたテキストから，いわゆる「機械可読形式」によるコンピュータ・ファイルを作るさいに，まず決定すべき事は，その目的である。紙の上のテキストが持つ豊富かつさまざまな情報のうち，どの階層の情報をコード化するか，逆から言えば，どの情報を切り落とすかは，利用目的によって決定される。

いま，二つの極端なケースを考えてみよう。書かれたテキストを，言語によるコミュニケーションの根源たる「語り」にまで還元して分析したり，利用者に提供することが目的であれば，紙の上に見える文字を直接のコード化の対象とする必要はなくなる。紙の上では，文は文字の連なりとして表現されているが，その文が意味を伝えるのは，音節の連なりとして読まれることによる。この考えによるならば，文字の字形がもつ視覚的な情報を削ぎ落としたとしても，文の意味は伝達可能である。綴られた語そのものではなく，それが指し示す発音をあらわす記号が保存され，伝達されれば，この目的のためには十分なのである。

この目的にそってテキストを電子ファイル化するというのであれば，あの厄介な漢字を追放することが可能となる。東アジアの漢字文化圏においても，ヴェトナム語はアルファベットによる表記法を採用することによって，また韓国・朝鮮語はハングル表記を徹底することによって，漢字を追放している（あるいは，漢字の追放が進行中であると言うべきか）。わが国でもカナやローマ字文によって，日本語文を表記しようという主張は，コンピュータ普及以前からあるし，中国でもピンイン方式の普及によって，日常的な言語生活から漢字を排除することが構想された事もある。同時代の文献であれば，作者と読者とが属する言語環境が近いために，音声に直接結びつく単純化された記号を媒介として意味を伝達したり，テキストを分析したりすることは可能である。

しかし歴史的な文献となると，作者と読者の言語環境の差異は大きい。古い時代の言語環境のなかにある現代人は，古い語彙や語法にもとづく文を，語りとして耳から受け取っても，その意味をただちに了解することは困難である。意味不明の語に逢着した場合，その発音をあらわす記号をキーとして辞書を引くことが可能であれば，この困難は軽減される。

東アジアの言語においても，モンゴル語や満洲語の古い文献の多くは，ウイグル文字に起源する固有のアルファベットを使って表記されているが，これらの言語にあっては，このアルファベット，ないしはそれをローマ字などに転写した表記によって語彙を配列した辞書を使うことができる。こうした言語のテキストは，語の発音と緊密に結びついた単純化された文字によって表記されているのだから，もとのテキストを構成する字母の字形や書法から切り離された記号の体系を用いて電子ファイル化することも容易である。というよりは，元来のモンゴル語や満洲語の字母（および圏点つき字母）をそのままコード化し，画面に映し出すような処理方法を採用しても，別の発音記号によって転写したものを電子ファイル化しても，伝えうる情報に決定的な差異はない。

ところが，基本的には単音節語である漢語の場合には，発音だけをキーとした辞書を利用することがほとんど不可能である。これは，漢字文化圏のなかでも，本家の中国に特有の言語文化上の困難であるのかも知れない。われわれが，その発音を通じては意味が了解できない語に逢着した場合，部首や画数といった漢字の字形情報が利用できないとなると，辞書を引いて正しい理解に到達することは，困難の度を増す。古語や僻字が頻出する古い文献であれば，困難であるというよりは，不可能に近いと言えるであろう。したがって，この発音の記号によるコード化方式を適用できるのは，他の言語の場合はいざ知らず，漢語文の場合においては，限られた範囲のテキストでしかないであろう。

上のように，綴られた語や文を，発音記号にまで還元して電子ファイル化する方法は，書かれたテキストのもつ情報を，ほぼ極限まで最小化して保存し，伝達するものと言えよう。これとは反対方向の極端な例としては，あるテキストを，異本の一つとして，他の諸本との校合の素材とする目的で電子テキスト化することなどが考えられる。

ウルス・アップ氏は，こうした目的をもって，敦煌莫高窟で発見された仏典写本のいくつかを処理した。興味深いレポートがアップ氏によって書かれている。こうした目的のためには，原文書に即した行替えや頁替えなど，テキストの外形的な構造を「印づけ mark-up」するだけでなく，テキストに附加された情報まで，印づけの規則にしたがってコード化することが有用である。アップ氏が，こうした印づけを施すべき要素として指摘するのは，以下のようなものである。

文字サイズ(双行注など付加的部分は，しばしば小さな文字で書かれる)
異体字
判読不能の文字
残存部分から推定できる文字
行間に附加された文字

塗抹，上書きなどによる訂正文字
記号によって倒置された文字
記号類
紙の破れ
貼り合わせによる紙の追加
インクの特徴

一つのテキストが，さまざまな写本や刊本で伝えられている場合，相互の照合によって異同を摘出することが，テキスト批判の出発点である。異本のそれぞれについて，上記のような印づけをおこなったうえで，こうした電子テキストの校合用に開発されたソフトによって処理させれば，さまざまな形態で校合の結果を出力することができる。正確に入力され，統一的な規則にしたがって印づけされた電子テキストを用意すれば，コンピュータ・プログラムがあっという間に校合作業をしあげてしまう。その速さ，正確さが，コンピュータ処理のメリットであるし，これに加えて，校合の対象とするテキストの選択や，出力結果の選択を，分析の目的に応じて柔軟に指示することができる[1]。

こうした付加的な情報を含んだ電子テキストは，諸異本の校合という目的にとっては，きわめて価値の高いものである。しかし，本文中に多くの印づけのための記号を含んでいるために，読むためのテキストとしては，却って読みにくいものとなる。また，テキスト中の文字列を検索するさいにも，行替えや印づけのための記号を無視して検索をおこなうような工夫が必要となる[2]。印づけなどの記号を削除して，より単純な電子テキストを派生させることは，機械的におこなうことができる。したがって，書かれたテキストから，上記のような附加的情報まで抽出してコード化した電子テキストは，情報量の少ない単純なテキストを包含しているわけである。だが，多様な印づけ作業をおこなうことは，作業量を大幅に増大させる。検索や索引作成などを目的として作る電子テキストにとっては，もっと情報量をそぎ落とした単純なテキストが望ましいであろう。

電子テキストの作成における作業量の大小を横軸にとれば，上で取り上げた二つの例は，ほぼその最小値と最大値に位置するであろう。

---

[1] ウルス・アップ「コンピュータによる敦煌文献の校合」『電子達磨』4号　1995　花園大学国際禅学研究所。

[2] 現在，ひろく使われている文字検索用のツール，たとえばgrepなどは，行を跨いだ文字列を検索できない。コンピュータ・ファイルの場合，「行」というのは，書かれたテキストにおける段落に相当する。検索対象の語が，段落を跨ぐという事態はあり得ないので，grepがこのような仕様であっても，何ら問題は生じない。ところが，紙に書かれたテキストの行は，文の途中であろうが，漢語の熟語の真ん中であろうが，字詰めの限界に達すれば機械的に改行してある。こうした紙の上における行の形態を，コンピュータの改行・復帰記号を用いて表現することが，じつは不具合の原因である。

2　電子版本の質と技術的条件

この二つの例の間にあって，より一般的な電子テキスト作成の目的とされているものは，およそ次の二つにまとめることができる。

目的1　信頼しうる新たな版本として電子テキストを作成する

目的2　検索用と割り切り，テキストの質の高さは必ずしも求めない

目的1による電子テキストは，まず，伝統的な校訂本の作成と同じだけの作業を必要とする。その上で，コンピュータ上での入力，校正作業などがこれに加わる。印刷された校訂本の作成と比較すれば，最後の印刷，製本などの力仕事を欠いているだけである。この目的1による電子テキスト作成が，専門家による注意深い作業を必要とすることは当然である。入力作業などは，本来，比較的単純な仕事であり，校訂テキスト作成者の頭を悩ますようなものではない。しかし，現時点では，これが漢字を使った電子テキスト作成者を悩ます大きな問題となっている。

作成する電子テキストが，ある特定のコンピュータでのみ使われるのであれば，そのコンピュータ環境に依存した漢字コードのセットを基盤として，そこに含まれない文字を「外字」として編号していくことが，もっとも実用的である。しかし，利用者へファイルを配布したり，ネットワーク上で電子テキストを公開するなどの場合，特定のコンピュータ環境に依存した外字方法をとることは，得策でない。作成者の入力画面に表示される外字が，利用者の画面の上に同じ字として現れることは期待できない[1]。ネットワーク上では，その外字部分だけでなく，それ以降の文字をも道連れにした文字化けを引き起こすという現象も起こっている。

では，外字を使わずにすむ方法があるかというと，これまた困難である。そもそも，過去の歴史的な文献を処理するに堪える漢字文字コードのセットが，いまだに存在しないからである。こうした目的に堪える漢字コードセットを用意するためには，そもそも過去使われたことのある漢字のすべてを確定することが必要である。一般的には，大規模な字典や辞書は，その親字として採録する漢字の種類が5万字に近い。こうした大規模な字典・辞書編纂の作業は，やはり，これによって過去使われてきた漢字を網羅することを目指していた筈である[2]。ところが，その成果を利用しつつ，過去の文献に当たってみると，なおそこに含まれて

---

[1] もちろん，利用者が作成者と同じOSを使い，電子テキストのファイルとともに外字フォントファイルを受け取って，それを画面に表示させるような設定をしておくならば，外字方式を採用してよいわけである。しかし，利用者の使うOSは多様であるので，このような方法を採用すれば，その電子テキストを利用可能なOSを制限することにつながる。

[2] 5万字前後の採録漢字数をもつ辞書類として，下記のようなものを挙げることができる。

いない漢字に逢着することがある。現在までのところ，最大の漢字数をほこる字典は，『中華字海』（中華書局 1994年9月）であり，そこには，約8万7千の漢字が採録されているという。しかし，これで歴史的文献にあらわれる漢字のすべてが採録されたとは断言できない。このように，現在までの漢字学が提供する情報の不完全さが，大規模な漢字コードセットを実現するうえでの困難の一端をなしている[1]。

こうした現状のもとで，歴史的文献を対象として，缺字の許されない電子テキストを作ろうとすれば，多かれ少なかれ，独自の追加文字セットをコード化して，これを表示するようなシステムを構築することが必要になる。臺北の中央研究院は，すでに二十五史全文データベース(約4千万字　1990年に完成)をはじめとする「古籍全文資料庫」という世界最大の漢籍電子テキストを作成している。

無償公開部分=6,000万字	二十五史，臺湾方志，臺湾档案，文心雕竜考異及注，仏経三論，新清史本紀，上古漢語語料庫
有償部分=6,000万字	十三経，諸子，古籍三十四種，大正新脩大蔵経

中央研究院では，こうした電子テキスト作成にあたり，まず，同一の底本を使い二つの作業者が並行して入力をおこない，ファイルを2本用意した。この2本のファイルをファイル比較プログラムにかけると，両ファイルの相違点として誤入力箇所が摘出される。まず，これを訂正した上で，専門家がさらに底本と対校するという厳密な方法を採用した。こうして，

『集韻』	53,525字	宋代勅撰(1037)
『康熙字典』	47,043字	清代勅撰(1716)
『康熙字典文字集覧』	49,188字	京都大学(1981)
諸橋『大漢和辞典』	49,964字	大修館書店(1960)
『中文大辞典』	49,905字	中国文化研究所(1968)
『漢語大字典』	54,960字	湖北辞書出版社・四川辞書出版社(1990)

[1] 現代の日常的な文献を対象とするかぎり，こうした大規模な漢字コードセットを用意する必要はない。現代中国文でも，たかだか3,500の漢字によって，用字の99.48%をカバーしている（国家語言委員会『現代漢語常用字頻度統計』）。日本語文の場合も，大同小異である。したがって，現在のJISコードやGBコードが採用している6千～7千文字を，常用部分として実装するというのは，正しい方法なのである。ところが，これ以外の僻字のコード化が進まず，また臺湾で規格化されつつあるCNSのような5万字レベルの大規模なコードセットによっても，歴史文献の処理にさいして缺字が生じてしまうわけである。漢字学が，しっかしりとした典拠情報をもち，かつ異体字関係を整理した漢字全覧を提供していれば，僻字部分のコード化には，さしたる困難はなかった筈である。

誤りが1万分の1以下という高い精度をもつテキストを提供している[1]。この電子テキストは，臺湾，香港で普及しているBig5の漢字コード（13,523文字を編号）を使っているが，中央研究院が独自に開発した装置を付け加えたコンピュータでないと，Big5によってコード化されていない缺字を表示することができない。1997年3月より，中央研究院はインターネット上における電子テキスト検索サービスの範囲を拡大した。検索結果の出力数に上限があるなどの制限はあるが，二十五史を含む上記無償公開部分の全文検索が，ネットワーク上で行えるようになったことは，画期的である。

中央研究院の「古籍全文資料庫」は，専門家による校訂をへた高い精度をもつテキストでありながら，インターネット上では，こうした缺字問題の制約があるため，新たな版本としてこれを生かすことが困難となっている。しかし，二十五史などは，底本とされた中華書局本それ自体の入手が容易である。このため，検索用のテキストとして，これが公開されていることは，利用者にとっては十二分の利便性を提供していると評価できよう。

目的2としてあげた検索用の電子テキストは，現状での漢字情報処理がかかえるさまざまな制約のもとで，とりあえず実用を旨として作るものと位置づけることができる。たとえば，『歴代宝案』にしても，各地の諸写本を校訂の材料とし，諸本の文字の異同まで注記した校訂本（沖縄県立図書館刊行）が，現状では依拠すべき最善の版本である。これを底本として，校注も含めたかたちで，電子テキストとして提供することができれば，上に掲げた「信頼しうる新たな版本」としての電子テキストを作るという目的に合致するものとなろう。『歴代宝案』は外交文書集であるため，漢語の文語文としては用字範囲の狭い部類に属する[2]。にもかかわらず，画面表示や印字ができない缺字を含まず，かつ利用環境に依存する外字方式をつかわない，という条件を課すならば，書物として出版されている校訂本『歴代宝案』と同等の電子テキストを作ることは，現状では困難なのである。仮に，臺湾で規格化され5万字近い漢字数をほこるCNSコードを利用したとしても，缺字が皆無にならないのが，頭の痛いところである。修辞をこらした冊封使の詩文は，公文書よりもはるかに広い用字範囲をも

---

[1] 李珊「縦横古籍新利器―中研院古籍全文資料庫」『光華』民国86年5月号(1997)。例えば，二十五史については，北京の中華書局が刊行した評点本(1959～1978)を底本としているが，この底本の誤りを正した点もあるという。なお，二重入力による入力ミス検出の技法は，京都大学人文科学研究所附属東洋学文献センターと大型計算機センターで，『東洋学文献類目』のデータ入力や，康熙字典所載漢字に，三角編号を付与する作業をおこなうさいに発案された。

[2] 『歴代宝案』第一集および，清朝の内閣大庫に蓄積された琉球関係公文書を編纂した『清代中琉関係档案続編』（中華書局　1990）の入力の過程で遭遇した缺字の一覧が，赤嶺守氏によって提供されている。既知の俗字や異体字などを「正字」に統合した結果，JIS漢字(JIS X 0208)にコード化されていない缺字は，およそ330字～340字ほどである。うち，227字については，JIS補助漢字(JIS X 0212-1990)にコード化されている。缺字の数に幅をもたせたのは，こうした手書き文書では「訛字」が珍しいことではなく，さらに校訂が加えて「訛字」を正規化すれば，缺字の数が減る可能性があるからである。

つのであり，困難はさらに大きい。

こうした現状のもとでは，目的1は将来の課題とし，暫定的に，あるいはその目標に向けた第一歩として，検索用と割り切った電子テキストを作成することもやむを得ない。上記のような漢字処理が抱える問題を一掃するようなコード体系の実用化を待つならば，歴史資料の電子テキストの形態による公開と利用は，遅々として進まぬであろう。

JIS漢字は，言うまでもなく現代日本で常用されている漢字の情報処理のためのものである。漢籍，それも近世の公文書や古典的な詩文を対象とする作業において，JIS漢字を利用するのは，そもそもお門違いなのである。用字範囲が広い漢籍を電子テキスト化するにさいし，JIS漢字コードは，もっとも多くの缺字を発生させる。しかし，臺湾で普及しているBig5など，中国語（漢語）のコードセットを使ったとしても，缺字の頻度は減少こそすれ，やはり相当の缺字は残るのである。CNSのような大規模なコードセットが使えたと仮定しても，皆無にはならない。入力作業や，日本国内における利用のしやすさを評価基準として重視すれば，現状では，JIS漢字を採用するのが，むしろ賢明な選択肢であろう[1]。どのようなコードを利用しても，JIS，Big5，GBなどの文書ファイルを，相互に変換することは容易であるし，変換テーブルさえ用意すれば，将来日の目を見るであろう大規模なコードセットを用いたシステムに移行することは可能である[2]。

採用した漢字コードセットに含まれない缺字を，他の文字と同じように画面に表示したり，印刷したりすることは（外字方式ではこれがほぼ可能である），上に述べたように特定のコンピュータ環境に依存することと引き替えでなければ実現できない。検索用とわりきった電子テキストでは，システムに依存した外字方式ではなく，何らかの代替策を施したテキスト

---

[1] 委託入力が可能であれば，Big5コードが選択さるべきである。缺字は減少するし，JISコードへの変換は機械的におこなうことができる。普通のBig5→JIS変換ツールは，JISに変換すると缺字になる字を空白などにしてしまうが，後述のアップ・ウイッテアン両氏が作成した変換ツールは，JISに変換することで缺字になってしまう箇所に，CNSコードに基づく代替記号を自動的に挿入してくれる。

[2] 各国のコードセットを，一つのファイルのなかで混在して使えるようにする多言語処理が，Unix上のmuleなどでは実現されている。そこでは，JIS，Big5，GBばかりか，CNSのような大規模なコードセットまで，フォントさえ用意すれば，混在して使うことが可能である。また，あらたな，コードセットを加えることもできる。しかし，現状ではmule以外のソフトウェアで，こうした多コード混在のファイルを処理できないという問題がある。さらに，ある漢字がどのコードセットを使って入力されているか知っていなければ，検索も難しい。JISの缺字箇所にはBig5を入力し，Big5でも缺字になる箇所にCNSを使うというようなルールで，入力をすることになるのであろうか。処理の単純さを考えれば，一つのコード平面のなかに「CJK統合漢字」を含むUnicodeの方が，用字範囲の広い漢籍の処理には適しているかも知れない（もちろん缺字は避けられない）。古い時代においては，漢字は東アジアの国際文字だったのであり，そこから分化した各国の通用漢字における微妙な字体の相違を，コードのレベルで区別しようという「多言語処理」の枠組みは，古い漢籍の処理にとって意味あるものではないし，却って複雑な処理の必要を持ち込むことになる。

を作るという選択が可能である。また，電子テキストの正確度は高ければ高いほど好ましいのであるが，これはコストと労力にはね返る目標である。検索の対象となるような述語が間違っていなければ，まあまあ用には足りるという考えのもとに，作業を進めねばならないこともあるかも知れない。電子テキストの長所は，改善を繰り返すことが簡単だという点である。しかも，ネットワークが発達しつつある現状では，利用者からの誤りの指摘を受けることも，改善された新たなファイルを送達することも，ともに容易である。最初から完璧をめざすよりも，必要に応じた漸進的な改善を見込んで，校訂未完の検索用テキストを公開することも考慮されてよい。研究用に使うには，十分な注意が必要であるが，手許において利用するテキストに騙されないために，必要に応じてよい版本を参照するのは，紙の書物の場合も同じである。

また検索用のテキストの場合，異体字はできるだけ正字に統一することが好ましい[1]。この措置によって，情報落ちが生じることは確かである。しかし，現状ではいかなる漢字コードも異体字関係を対照するしくみを内在させてはないし，検索用のツールにおいても，検索条件中の文字について，自動的にその異体字を包含した検索をおこなうことはできない。異体字関係を属性の一つとして持つようなシソーラスないしは電子漢字辞書を用意すれば，検索用のツールが自動的に異体字を包含した検索をおこなうことも容易になるし，異体字を含むテキストを自動的に正字に正規化するようなプログラムを作ることは可能である。しかし，現状ではこれも難しい[2]。大規模な漢字コードセットが利用できない現状のもとで，異体字をそのまま入力するような方針をとれば，却って缺字を増やすだけだということも考慮さるべきである。

3　テキスト入力の方法

大きな予算を割り当てることができれば，入力は外注し，専門家が底本との対校や異本を

---

[1] 「沖縄の歴史情報」プロジェクトでは，金城正篤氏の主宰する研究班「琉球・沖縄の対外関係史」の方々が，「異体字・常用漢字の＜統一字＞一覧表」を作って，テキストの正規化をおこなっている。これは，『歴代宝案』などの校訂・入力作業を通じて策定された実践的な作業方針として貴重な成果である。われわれが試みた冊封使関係資料の処理においても，これを利用させていただいた。

[2] JIS漢字内の異体字関係を自動的に処理することは，簡単なプログラムと置換リストによって行うことができる。ここで「難しい」と言うのは，JISに含まれていない漢字を含めて，異体字の正規化処理をおこなうことである。どの漢字とどの漢字が，同一字の異体関係にあると見なせるかという問題は，実は簡単ではない。それが，時代によって揺れるからである。また，刊本や抄本にあらわれる多様な文字の字形のうち，どこまでを「単なる書法の差異」とみなし，どこからが「異体」「別体」として判定さるべきかという同定作業を，徹底しておこなわない限り，過去の異体字まで含んでコード化した大規模漢字コードは編成できない。日本の古文献を対象としてこうした作業をおこなうことは大変な作業であるし，中国のそれを加えるならば，さらに大がかりなものとなる。

もちいた校訂作業をおこなうのが理想的であろう。中央研究院がおこなったような，二重入力による入力誤り検出法を採用すれば，精度を向上させることは間違いないが，これも予算との相談である。いずれにしても，専門家による対校作業の質が重視さるべきは言うまでもない。

もう一つの選択肢として，光学読み取り装置（以下ＯＣＲと略称）を使うことが考えられる。近年，パソコン上で作動するものでも，処理速度の向上は著しく，活字のテキストであれば，精度も高い。ＯＣＲの導入によって，入力の経費は省けるが，委託入力と比較すると，対校の手間が増大するし，機械読み取りの「癖」を知らなければ却って誤りを増加させる危険性もある。また，入力の底本として木版本や鈔本しか得られない場合には，いまのところそのような文字を満足に読みとるＯＣＲがないので，手入力を選択せざるを得ない。

本プロジェクトの計画研究「環東シナ海地域間交流史ッッ中国福建を中心として」(研究代表者　夫馬進教授)では，まず，使琉球録の全文データベース作成を試みた。対象となるテキストは，16世紀から18世紀の中国の官僚が，修辞の技術をつくして書いた詩文である。全文データベース作成の対象としたのは，4種類の使琉球録である。

1陳侃『使琉球録』1巻	嘉靖13年(1534)尚清冊封
2蕭崇業・謝杰撰『使琉球録』2巻 附『皇華唱和詩』1巻	万暦 7年(1579)尚永冊封
3夏子陽『使琉球録』2巻	万暦34年(1606)尚寧冊封
4徐葆光『中山伝信録』6巻	康熙58年(1719)尚敬冊封

これらの使琉球録は，臺湾銀行経済研究室が刊行した臺湾文献叢刊に収められている。この臺湾銀行の活字本は，字体は旧字体である。また，句読点も字の右下に寄せる日本式とは違い，行の中央にカンマやセミコロン，句点（まる）などを打つ。こうした特殊なテキストではあるが，適当な入力委託先が見つからなかったこともあり，試験的にＯＣＲによる入力を検討した。

底本とする活字本が，質の悪いものでは用をなさない。そこで，まず，原刊本（夏子陽のものは鈔本影印本）と対校してみた。すると，この臺湾銀行本は，専門家の手によって校訂がなされており，誤植も少ない信頼するに足るテキストであることがわかった。

ＯＣＲソフトの選択については，対象が臺湾で刊行された活字本であることを考えれば，Big5コードをつかう中国語ＯＣＲが理想的であろう。しかし，適当なソフトが見つからなかった。また，導入するＯＣＲが，日本語の文献にも利用できることを考えれば，JISコードの制約は受けるけれども，種類の豊富な日本語ＯＣＲの中から選択することも，試みる価値があると判断した。単純な日本語の読みとりについては，大抵のソフトは認識率98%〜99%に達している。文字パタン照合だけでなく，語法解析を行なうことによって，精度を高めているのだが，古典漢語文が相手では，こうした機能がつかえず，カタログの読みとり精度の数

値は参考にならない。すると，選択の基準は，可能な限り多くの漢字を読みとれるという点に絞られるといってよい。

たまたま，JIS第二水準漢字の全てを読みとることが可能なＯＣＲソフトが京都大学人文科学研究所東方部に導入されていたので，これを試用してみた。上記の臺湾銀行本を読みとらせてみると，活字字体や句読点が，日本のそれと異なっていることが，読み誤りを引き起こすなどの問題はあるものの，充分に実用的であるとの感触をえたので，これを導入した。

このＯＣＲ装置は，MY-QREADER Pro for Windows（日立マイコンシステム）というWindows上で作動するものである。文字のパタン認識を，ASAスロットに挿入する専用のロジックボードでおこなうため，高速な文字認識が可能である。なお第二水準漢字すべてを含む認識辞書は，オプションとして用意されている。頁を画像として入力するために，別にスキャナが必要である。最近は安価なものでも400dpi（1インチあたり400本の線密度）の読みとり能力があるので，選択の幅は広い。

経験を積むことによって，こうした特殊なテキストを読みとらせるには，いくつかのコツがあることが分かった。

①認識辞書への登録によって，文字認識を学習させる。

日本語ＯＣＲの場合，標準の認識辞書は，認識対象の文字と照合すべきパタンとして日本で通用している活字の字体を記憶している。いわゆる旧字体で印刷されたものを読みとらせようとすると，「社「神」などの漢字の扁が「示」になっているので，正しく照合できず，読みとり結果が不安定になる。こうした文字に遭遇するごとに，そのパタンを認識辞書に登録していくと，正しい照合結果を返すようになる。また，JISに含まれない缺字に遭遇した場合も，なんらかの代替記号を認識辞書に登録することによって，こうした代替記号を含んだ認識結果を得ることができる。

②スキャナによる頁画像の入力にさいし，行のならびが傾かぬよう注意する。

スキャン後に画面のうえで画像の傾きを補正する機能が用意されているので，これを利用することもできる。文字認識をかけるさいに，行のならびが傾いていると，縦横数文字分からなるブロックを，一つの文字として認識するなどという現象が起き易くなる。縦書きの場合には，横書きよりも，行が長くなるためか，頁画像の傾きの許容度が小さいようである。

日本語の文書でも，「当用漢字」に略字体が採用される以前の書物などを読みとらせる場合には，ＯＣＲの認識辞書の登録機能を活用することが必要になる。また，中国で使われている簡体字を登録した認識辞書を作れば，日本語ＯＣＲを使って，中国語の印刷物を対象とした読みとりをおこなうことも不可能ではない。したがって，ＯＣＲソフトを選択するさいに，認識辞書への登録機能が充実しているかという点は，重要な評価基準となる。缺字の代替記号は，英数字2文字以上であることが多い（後述のIRIZ漢字Baseでは9文字）。したがって，ある一つの文字パタンにたいし，1文字ではなく，こうした長い文字列を登録できる仕様が必要である。また，登録できるパタン数には上限があるので，上限が大きいほど，好ま

しい[1]。

市販されているＯＣＲは，日本語や英語のビジネス文書を主たる適用対象として想定している。臺湾で印刷された旧字体の古籍の読みとりが巧くいかなくて当然である。しかし，作業を重ねるうちに，上記の認識辞書の登録機能や学習機能を活用して，しだいに賢くすることは可能である。また，画面上で頁画像と読みとり結果を対照しながら校正する作業も，読みとり機構の癖や能力が分かってくると，ツボをおさえて効率があがるようになる。委託入力ができない場合や，長い書物の一部だけを研究資料として入力したいなどの場合，こうした工夫が，ＯＣＲを入力代替の有力な候補にしうる。

われわれは，使琉球録の入力に際し，校正を三度重ねることとした。初校は，ＯＣＲの画面を使っての対校である。上に述べた認識辞書への登録作業なども並行しておこなうし，画面上での作業であるので，仕事の強度はかなりのものとなる。また，作業者が，古い漢字の取り扱いに慣れていることが必要であり，誰でもできる仕事ではない。すると，スキャナによる入力作業も含めて，専門家がおこなうことになる。再校は，読みとり結果を印字したものを，底本（台湾銀行本）と対校した。この作業は，大学院生によっておこなわれた。この段階までは，底本とおなじ箇所で改行しておくことが，対校作業の効率をあげる。三校は，底本が依拠した原刊本と対校した。これは，岩井が担当したが，厳密にこの作業を行ったのは，蕭崇業・謝杰の使録のみである。臺湾銀行本は信頼すべき良質のテキストを提供しているが，活字本を作るさいの避けられない誤植が，きわめて僅かではあるが，存在する。さきにも述べたように，現在の漢字コードシステムでは，完全な電子版本を作ることは不可能であり，われわれの作る使琉球録の電子テキストも，検索用のものでしかない。しかし，良質のテキストを作ろうとするならば，原刊本にまでたち帰った対校をほどこすのが望ましい。

こうしたＯＣＲ入力による電子テキストの作成と並行して，もっとも良質のものであろうと判断される刊本（一部はその影印本）の画像入力をスキャナでおこなった[2]。現在の電子テキストにおいては，利用者のコンピュータ環境に依存せずに，缺字の字形を表示する手段がない。これを補う一つの手段として，原刊本の画像をファイル化して，CD-ROMなど大容量の記憶媒体に記録したものを，研究者に配布するという選択肢がある。また，利用者が，

---

[1] MY-QREADER Pro for Windowsでは，一つの登録用辞書には250文字ほどのパタンを登録できる。文書の性質ごとに，個別の登録用辞書を作り，それを切り替えて使うという機能は重要である。MY-QREADERでは，文字認識にさいし，二つの登録用辞書を組み合わせて使うことができるので，理論的には標準的なJIS漢字以外に，500文字のパタンを加えて文字認識させるることが可能になる。また，一文字にたいし，16字までの長さの文字列を登録することができる。

[2] マイクロフィルムに撮影したものがあれば，一括して画像ファイルに変換するサービスをおこなう業者に委託することができる。経費は掛かるが，スキャナ入力よりは効率的である。紙が変色していたり，シミなどがある木版本の場合，白黒2階調では鮮明な画像が得られないことが多い。データ量は増えるが，16階調や256階調のグレースケールの出力画像を採用することも考慮されてよい。

電子テキスト上の字句に疑義をいだいた場合，このようなCD-ROMが手許にあれば，画面の上に原刊本の頁画像を呼び出して対照することが手軽にできる。校正が不十分な電子テキストを試験的に配布する場合には，こうした便宜を提供して資料の共有をはかることができるし，意欲と必要性をもつ利用者が電子テキストを改善するのを促すことにもなろう。高速な通信手段が普及すれば，こうした原刊本の画像ファイルを，ＣＤ－ＲＯＭに焼くのではなく，ネットワーク上のサーバーに置いて，遠隔地から呼び出せるようにすることで，資源の有効利用を図ることができる。

4　缺字の取り扱い

「沖縄の歴史情報」のプロジェクトに先行して，沖縄では『歴代宝案』の校訂と電子テキスト化が進行中であった。当初，この電子テキストは，ワープロによって入力され，その外字機能によって，缺字を埋めていたようである。最終的には，赤嶺守氏が出現した缺字を整理して独自の代替記号をわりあて，れをファイルのなかに入力することとなった[1]。『歴代宝案』第一集と『清代中琉関係档案続編』の入力作業の過程で遭遇し，赤嶺氏の代替記号表のなかに編号された缺字は384字であった。うち11字は，JIS漢字のなかに対応する「正字」を見いだしうる異体字ないしは俗字であると認定され，除外された。したがってJISコードでは対処できない漢字は373種となっていた。

われわれが使琉球録の入力作業をはじめた当初，この「赤嶺外字表」は，まだ配布されおらず，利用することができなかった。岩井は，従来より，花園大学国際禅学研究所のアップ氏とウイッテアン氏が開発した「IRIZ漢字Base」[2]を利用して，CNSコードに基づく代替コードを使っていたので，ＯＣＲによる蕭崇業・謝杰『使琉球録』の入力にさいし，これによって缺字を埋めることとした。後述のように，『使琉球録』と『歴代宝案』などの公文書類とでは，用字の範囲の相違が大きいため，当初の「赤嶺外字表」を利用して缺字を埋めることは，ほとんど不可能だったからである[3]。

蕭崇業・謝杰『使琉球録』のテキストの全体量は，『歴代宝案』に比べてはるかに小さい。

---

[1] 赤嶺守氏の考案された代替記号の長所は，それが英数字2文字からなることである。この文字幅は，漢字1字分と等しいため，底本の行の並びや長さを乱すことなく，代替記号入りの電子テキストを作成することが可能になる。底本との対校を容易にする点で，この方式は優れている。

[2] 「IRIZ漢字Base」については，頁x以下でやや詳しく紹介してある。

(http://www.iijnet.or.jp/iriz/irizhtml/indexj.htm)に，詳しい紹介がる。

[3] 既出の缺字を随時編号する方式では，あらたなテキストを入力するたびに「外字表」が伸びていく訳である。その度に，外字表全体を部首順などの原則で再整列すると，入力済みのテキスト中の代替記号も，すべて変換せねばならない。この作業を避けようとすれば，追加部分については，それまでに整列されていた部分の外に附加されることになる。したがって，部首順，筆画順などの原則にもとづいて外字表を検索することが難しくなる。これが，外字表方式の難点である。

単純にファイルの大きさで比較すると，前者は後者のほぼ10分の1の総文字量である。にもかかわらず，遭遇したJIS外漢字は523字にものぼった。この523字の延べ出現回数は961回である。蕭崇業・謝杰『使琉球録』の総文字量は5万字弱であるから，およそ100文字の本文のなかに，缺字が2字ほど出現するという頻度になる[1]。

この523字と，既出の『歴代宝案』第一集と『清代中琉関係档案続編』から摘出されたJIS外の缺字ッッ当初「赤嶺外字表」に含まれていたものッッとの関係を示せば以下のようである。

『歴代宝案』第一集・『清代中琉関係档案続編』の外字と共通するもの	76字
JIS漢字のなかに「正字」を見いだしうる異体字と認定されたもの	10字
蕭崇業『使琉球録』だけに現れたもの	437字

沖縄では，『歴代宝案』第一集と『清代中琉関係档案続編』につづいて，『中山世譜』を入力が行われた。この作業のなかで新たに遭遇した缺字が，わずか65字であったのと比べると，蕭崇業・謝杰『使琉球録』の用字範囲の広さは印象的である。官庁文書の集成である『歴代宝案』『清代中琉関係档案続編』と，士大夫官僚の詩文を中心とした『使琉球録』とで，共通するJIS外の漢字が76字に過ぎないということは，文章の性質によって用字の範囲にそれぞれ偏りが強いということを示している。また，『中山世譜』が65字の追加にとどまったということは，この書物の文章が，档案のごとき行政事務にかかわる平明な文章と，ほぼ同程度の用字範囲しかもたないことを示している。ちなみに，この65字のうち，蕭崇業『使琉球録』にも見いだせるものは，5字に過ぎない。

言うまでもなく，分析の材料となる入力済みテキストの量が充分でないことも，JIS外の漢字の相互の重なりを小さく見せる要因の一つではあるのだが，現在，校訂作業の段階にある陳侃や夏子陽など明代の『使琉球録』が分析材料に加わっても，上でのべた用字範囲の偏差の傾向を打ち消すような結果にはならないという印象をもっている。

ちなみに，蕭崇業・謝杰『使琉球録』をJISとは異なった漢字コードシステムを用いて入力した場合の缺字出現率を紹介しておこう。まず，BIG5コードで入力したとすると，缺字は202字となる。これは，総字数に対し0.36%になる。約48,000字をコード化しているCNSコードを使っても，73字の缺字が出る。これは総字数に対し0.13%である。これらは，いずれもある漢字の異体字として，親字に統合できるものではない。

花園大学国際禅学研究所で入力された禅籍80点（約200万字）についての統計は，左のとおりである。

JIS缺字	16字／千字

[1] 句読点などッッ，。「」『』（）＊　[スペース]ッッを総文字数から除外すれば，缺字比率は0.3%ほど高くなり，20字／千字となる。ちなみに，蕭崇業のテキストでは，句読点などは総字数の17%ほどである。

BIG5缺字	1.8字／千字
CNS缺字	0.3字／千字

CNSのような大規模漢字コードセットが，どのように編成されているか，あるいはなにに基づいて編成されているか，知ることはできない。こうした大規模な漢字コードの編成にさいしては，旧来の『康熙字典』をはじめとする字典や辞典が，参照されるべきは勿論である。しかし，相手は，歴史のなかで蓄積されてきた膨大な中国の文献である。特に，出版文化が飛躍的に発展した明清時代の文献の用字について資料を集めることは，作業の困難さもあって，疎かにされている。禅藉に出現するCNS缺字の頻度よりも，5万字ほどの『使琉球録』のそれの方が大きいということは，こうした文献が，従来の文字学の研究対象から抜け落ちているということを示しているだろう。

日本においても，古い文献の電子テキスト化を目指して，あらたな漢字コードを編成しようとすれば，田村毅氏のように，まず，「日本語の漢字は何文字あるのか」を探らねばならないという現実がある[1]。欧米諸国では，文献の電子テキスト化が着々と進んでいる。こうした作業に着手するさいに，「フランス語を表記するのに，どれだけのアルファベットイが必要か」すら分からずに，入力を進めるということは，およそ考えられない話である。ところが，われわれは，文字の総数もわからずに，また，どの文字とどの文字とを異体字として同一視すべきかについても明確な指針のないまま，入力すべき文献に向き合っているわけである。

蕭崇業・謝杰『使琉球録』から検出されたJIS外の缺字を，「赤嶺外字表」に加えたことで，その総数は一挙に900字に近づいた。その後，さらに他の文献の入力作業が進んだ結果，「沖縄の歴史情報」プロジェクトの研究期間終了までには，1,050字ほどの編号が完了することが見込まれる。

この外字表には，不思議な漢字が散見する。『歴代宝案』の諸本が抄写を重ねて今日に伝わっていることを考えれば，そこに一般的でない俗字や訛字が紛れ込んでいて不思議ではない。こうした文字を，正規化するのも校訂作業の一環ではあるが，それがどの写本のどこに使われているかを，電子テキストのなかに情報として含めようとすれば，こうした文字も，外字表のなかに登録しておく必要がある。漢字学の視点からすれば，歴史的な異体字，俗字，訛字の出典情報を幅広く収集することは，われわれがもっと注意しておこなうべきことであろう。

電子テキストの缺字部分に，代替記号を入れておくだけでは，利用に不便である。テキストのなかで，こうした記号に遭遇するたびに，外字表を引っぱり出して，それに対応する漢字の字形を確認することも不可能ではないが，やはりこれは面倒である。外字表をめくらなくても，画面の上に字形を表示してくれる仕組みは必要であろう。漢字の字形を小さな画像

[1] 田村毅「文字文化の継承と未来」(1996年12月　http://www.um.u-tokyo.ac.jp/DM_CD/DM_TECH/KAN_PRJ/HOME.HTM#2）

として用意し，それを本文の缺字箇所に挿入するのも一つの方法である。現在のところ，これを実現する方法として，次のようなものを挙げることができる。

①花園大学国際禅学研究所の「IRIZ漢字Base」

②HTMLのインライン画像挿入タグを利用したハイパーテキスト文書

③Adobe社の提唱するPDF形式など，電子配布ファイル

①は，JISあるいはBig5の缺字を，CNSの漢字コードを利用する代替記号によって，テキストのなかに埋め込もうというものである。その代替記号は，&C0-BFEC;のように，&と;で挟まれた形式である。これは，ISOで定められたSGML（Standard Generalized Markup Language）という文書の構造を記述するための印付け言語における，外部実体参照（External Entity Reference）の形式である。外部実体とは，参照の記号と置き換わるべき文字などの実体が，外部の定義リストによって示されることである。&C0-BFEC;の例でいえば，

C0　　: CNSの第一字面において定義されている漢字であることを示す
BFEC　　: 16進表記による文字コードを示す

という二つの部分によって，これが示す漢字の字形は，CNSのコード表を定義リストとして，そこから得られることが表現されているわけである[1]。

IRIZ漢字Baseのメリットは，CNSという約48,000文字もの大規模な漢字コードセットによって，コード化された漢字を，四角号碼，ピンイン，部首・総画数などをキーとして画面上で検索することができ，候補群から目的の漢字を選択すると，簡単な操作で，それをテキストの入力位置に挿入できることである。現状では，古籍の電子テキスト化けにさいし，CNSコードを使っても，缺字がでることは避けられな。しかしCNSには，JISやBig5で缺字になる漢字の大多数がコード化されており，IRIZ漢字Baseによって効率的にそのコードを検索することができるし，また，目的の漢字が未編号である場合にも，その事実を確認するにの手間がかからない。

代替記号に対応する字形の表示について，IRIZ漢字Baseが採用したのは，MS-WordというMicrosoft社のワープロのマクロ機能と使い，代替記号の部分に，対応する漢字字形の画像を貼り付けるという方法である。MS-Wordは，画像の貼り付け方法として，テキストから独立した図形枠のなかに貼り付ける方法と，テキストの行の一部となるインライン画像として貼り付ける方法の二つを区別している。インライン画像として，普通の文字と大きさを合わせて缺字の画像を貼り付ければ，行が編集されると，画像の前後の文字と一緒に移動するので，好都合である。IRIZ漢字Baseは，電子テキストをのなかにちりばめられた缺字の代替記号の箇所に，まず対応する漢字画像を貼り付け，つぎにが代替記号そのものを「隠し文字」に属

[1] ちなみにSMGLでは，ISO/IEC 10646文字集合の特定の文字を参照する形式も定められいる。例えば，ユニコード（USC2）に含まれる漢字は，&#x4F5F のように表記する（=佟である）

性変更するという作業を，MS-Word上で自動的におこなうマクロプログラムを提供している[1]。

②は，なんらかの大規模漢字コードセットで定義されている漢字の字形を，小さな画像ファイルとして用意しておき，その画像ファイルへのリンクを示すタグをテキストの缺字出現箇所に挿入することで実現する。例えば，つぎのようになる。

而以乘船危竦我，是<IMG SRC="IMG00037.GIF">死而後已

これを，Netscapeなどのブラウザ上に呼び出すと，「是」という文字に続いて，画像ファイル"IMG00037.GIF"として用意してある「黔」の字形が表示される。

近年，インターネット上のハイパーテキストは急速に普及し，一般のワープロやエディタもこの形式の文書の編集を補助する機能を搭載するようになっている。また，これを閲覧するブラウザには，無料で配布されるものもあり，最近のコンピュータにはもれなくインストールしてあるといっても過言ではない。したがって，②の方式で缺字を表示するファイルは，ほとんどの利用者が特別なソフトを導入することなく，閲覧したり，編集を加えたりすることが可能である。

問題は，多数の缺字の画像ファイルをどうやって用意するかにある。一つ一つ，手作業で作ることも不可能ではないが，手間と時間を必要とする。それよりは，すでに提供されている資源を利用するのが賢明であろう。電子テキストをHTML形式によるハイパーテキストとして，ネットワーク上で公開するのであれば，安岡孝一氏が京都大学大型計算機センターのWebサイト上で提供しているJIS補助漢字の画像ファイルを利用することが出来る。この場合には，

孔子謂季氏，八<img src="http://www.kudpc.kyoto-u.ac.jp/cgi-bin/kanjibukuro/16dots/JIS+-1691">舞於庭

というように，缺字箇所に，安岡氏のサイト上の画像ファイルへのリンクを定義しておく[2]。一文字の画像ファイルのサイズは小さいので，通信手段が高速であれば，京大大型計算機センターから画像を呼び出して表示するのに，さほどの時間はかからない[3]。また，勝村哲也

---

[1] IRIZ漢字Baseについての詳しい解説と使用方法は，花園大学国際禅学研究所のホームペイジ（http://www.iijnet.or.jp/iriz/irizhtml/indexj.htm）を参照されたい。また，Christian Wittern's Web Space（http://www.gwdg.de/~cwitter/）でも情報がえられるし，検索画面がhttp://www.kb.oas.hist.uni-goettingen.de/kb/query.htmに用意されており，ネットワーク上から，四角号碼，ピンイン，部首・総画数などを検索キーとして漢字の情報（字形を含む）を得ることができる。

[2] この方法は，インターネット上の「倉頡計画」（http://village.infoweb.ne.jp/~fxba0016/）が採用している。見本として掲げた一文も，そこで公開されている「論語」の一文である。

[3] ネットワークから切断されたコンピュータの上で電子テキストを閲覧させるには，そのフィアルとともに，必要な缺字の画像ファイルを配布する必要がある。本文中の缺字箇所のリンクの指示も，ローカルなディレクトリ上の漢字画像ファイルを指すように変更せねばならない。

氏らが開設した「e漢字」というサイトでも，数多くの漢字のフォントが提供されており，そこから漢字画像を得ることも可能になった[1]。

③の方式は，印刷された冊子に代替する電子的な出版物の作成と閲覧のために考案されたものである。まず，ワープロやDTP（Desk-top Publishing）ソフトなどで作成した文書を，PDFという形式で保存する。このファイルを開くには，Acrobat Readerという閲覧用ソフトが必要であるが，これはadobe社が無料で配布することを認めているので，電子テキストとともに配布することも可能である。

PDF形式のファイルにしてしまえば，システムに依存する外字方式で缺字を埋めておいても，また，漢字画像を貼り付けておいても，それらはすべてPost Scriptというページ記述言語の表現形式に変換して保存される。閲覧ソフトがそれを解釈して，版面を再構成し，画面に表示するという仕組みのため，どのコンピュータの上でも，Acrobat Readerさえ使えば，まったく同じように表示される[2]。この方式は，画面の上に，印刷物と同等の品質をもって文書を表示しようとするため，画面表示や頁めくりの速度がやや遅いという問題がある反面，文字や図の割り付け，フォントの種類，大きさ，頁の余白など，作成者が用意した版面のとおりに，利用者の画面に表示される[3]。

単純なテキストファイルとして配布されたもののように，利用者が自由に編集を加えることができるという長所は失われるが，閲覧や単純な検索の用途にかぎれば，このPDF形式のような電子出版物に加工して資料を配布することも，有力な選択肢の一つとなる。

5　おわりに

漢字というはなはだ厄介な文字のゆえに，漢字文化圏における典籍の電子化と利用は，少なからぬ制約を受けてきた。明清時代の文人官僚たちの筆になる琉球冊封使関係の資料を電子テキスト化するわれわれの試みは，その制約の大きさを再確認させることとなったともいえよう。しかし，技術の進歩は急速である。「沖縄の歴史情報」のプロジェクトが出発した4年前，缺字問題解決のための方策としては，システムに依存した外字方式もやむを得ないか，とも思われた。しかし，現在ではここで紹介したような代替策を検討したり，試行したりすることが可能である。

どの方式が最善であるか，漢字コードシステム自体が変化発展しつつある今，にわかに見

---

[1] e漢字(電子漢字)ホームページのアドレスは http://www.zinbun.kyoto-u.ac.jp/~ekanji/。なお，ここで言及した「赤嶺外字表」もここで見ることができる。

[2] したがって，缺字を埋めるのに，TrueTypeのようなベクトル形式のフォントを利用することも可能である。ビットマップ形式の画像を貼り込む方式では到達不可能な，表示や印字の質の高さを追求することができる。

[3] 利用者が，Post Scriptを搭載したプリンタを用意すれば，画面に表示されるとおりの印刷結果を得ることもできる。

極めることは難しい。逆に言えば，便宜的な方策として，どれを採用しても悪くはないということであろう。遭遇した缺字を整理し，他の大規模な漢字コード中の漢字との対応関係を示すテーブルを用意しておけば，将来，より望ましい方法に乗り換えたり，別のコードシステムを使ったテキストに変換することが可能だからである。