「歴代宝案」の情報化と漢字処理

柴山 守：大阪市立大学学術総合情報センター

『歴代宝案』本文テキストのデータベース化に際して、ＪＩＳ第１水準、第２水準に含まれない外字をどのように扱うかはテキストの表示、検索、印刷、流通において極めて重要な問題である。まず、第一に外字を如何に処理するかを考える前に、本文テキストに含まれる外字が異体字なのか、旧字なのか、あるいは本文テキストの誤りで修正するか否か、他の正字に置き換え可能なのか等を決めなければ成らない。これには種々の考え方が存在し、この校訂作業そのものが研究であると言える。

本文テキストに含まれる字形をともなった外字が決定されると次に外字処理が問題になる。この問題を解決するために、以下に示すような手法が考えられる。

(1) 必要となる外字のすべてについて、外字エディタを用いて作成する。基本ソフトウェア（OS）に依存した登録外字数の最大値に注意しなければならない。

(2) 外字の１字ごとに識別子を付加し、本文テキストには識別子を用いて入力する。識別子と字形の対応テーブルで正しい字形が得られる。アルファベットと数字の組み合わせによって２バイトで表現する「赤嶺コード」が例である。

(3) 外字の１字ごとに字形のイメージを本文テキストに張り付ける。花園大学国際禅学研究所のウルスアップ氏、クリスティアン・ウイッテアン氏による IRIZ 漢字ベースの方式で、MS Word 上では外字イメージを本文テキストに張り付け、コード "&C3-5043;" のようなコードを付加する。通常このコードを非表示にしておく。

(4) 外字の１字ごとにその外字を形成する部分品を "{"や"}"で囲み、本文テキスト中に埋め込む。例えば "{金+平}" のように表し、本文テキストに埋め込む。

外字処理は、主に以上のような方式が考えられる。パーソナルコンピュータの日本語シフト JIS を基準にプラットホームを考えると、ＪＩＳ第１水準、第２水準に含まれない多くの外字を作成しなければならない。第一集の校正作業では、約５００字が必要になることが判った。したがって、作成が必要となる字数から Windows 版の最大外字登録可能数1,800 字に収まると判断され、また今後の流通などを考え、前述の (1) の方式を採用することとし、Windows95 をプラットホームとしてテキスト処理できるように外字フォントファイルを配布できるよう工夫した。外字パターンの作成には IRIZ 漢字ベースを使用している。

この外字ファイルを作成する手順は、つぎのとおりである。

(a) IRIZ 漢字ベースを用いて、ＪＩＳ第１水準、第２水準に含まれないフォントを抽出する。

(b) 抽出したフォント（.bmp ファイル）を Windows3.1 版―外字フォントファイルに登録する。登録は、できる限り手間を省くため別に作成したプログラムを使用した。外字エディタ上で、クリップボードからの [張り付け] を使用しても登録できる。

(c) Windows3.1 版―外字フォントファイル "USERFONT.FON" をフロッピーディスクに複写する。

(d) 上記の "USERFONT.FON" ファイルを配布し、Windows95 日本語版上で標準提供されている外字エディタを使ってロードする。　以上の操作で、校正された本文テキストの外字は、Windows95 版 のアプリケー

ション上で表示される。

IRIZ 漢字ベースは、Windows3.1 や Windows95 版上で動作する台湾政府の CNS コード（約 48,000 字）をベースに部首や総画数、四角号馬でフォントを検索することができるクリスティアン・ウイッテアン氏作成のソフトウェアである。検索結果には、Unicode、JIS 等のコードや読み等の属性情報も表示される。

漢字ベースのフォントの替わりに使用される代替記号は、つぎのような形式である。

図１に示す "chan4" のフォントは、コード "&C3-5043;" で表される。最初の "&" と最後の ";" は、コードの始端と終端を表す区切記号であり、始端の次に表される "C"が CNS コードであることを示す。引き続く "3" などの英数字は分類指標である。続く "-" の後に１６進数で表される４桁の数値列が漢字コードである。

（例）	"&C0-425E;"	３桁目	0	―	Big-5 コード　　コード：A440-C67E, C940-F9D5
	"&C4-223A;"		3-4	―	CNS レベル 3-7　　コード：2121-7C51
	"&CY-2134;"		X,Y	―	X：IRIZ 予約、　Y：一般利用可
	"&U-4E00;"	２桁目	U	―	Unicode

図１に示すフォントの検索結果は、[Copy]ボタンをクリックすることにより、MS Word (Ver.6) の編集画面上に張り付けられる。これは、漢字ベースの中で提供されている漢字ツールに CEF2BMP マクロが用意されている。このマクロにより、文字パターンはビットマップ・ファイルに変換される。

図１　　IRIZ 漢字ベースによるフォント検索例（部首番号 032：「土」）

図１において、フォントの検索結果のウインドウ上で表示されている字形をダブルクリックすることにより、文字パターンがクリップボードに複写される。これを他の図形処理アプリケーションや Windows に標準提供されている外字エディタに[張り付け]ることによって、文字パターンが抽出できる。

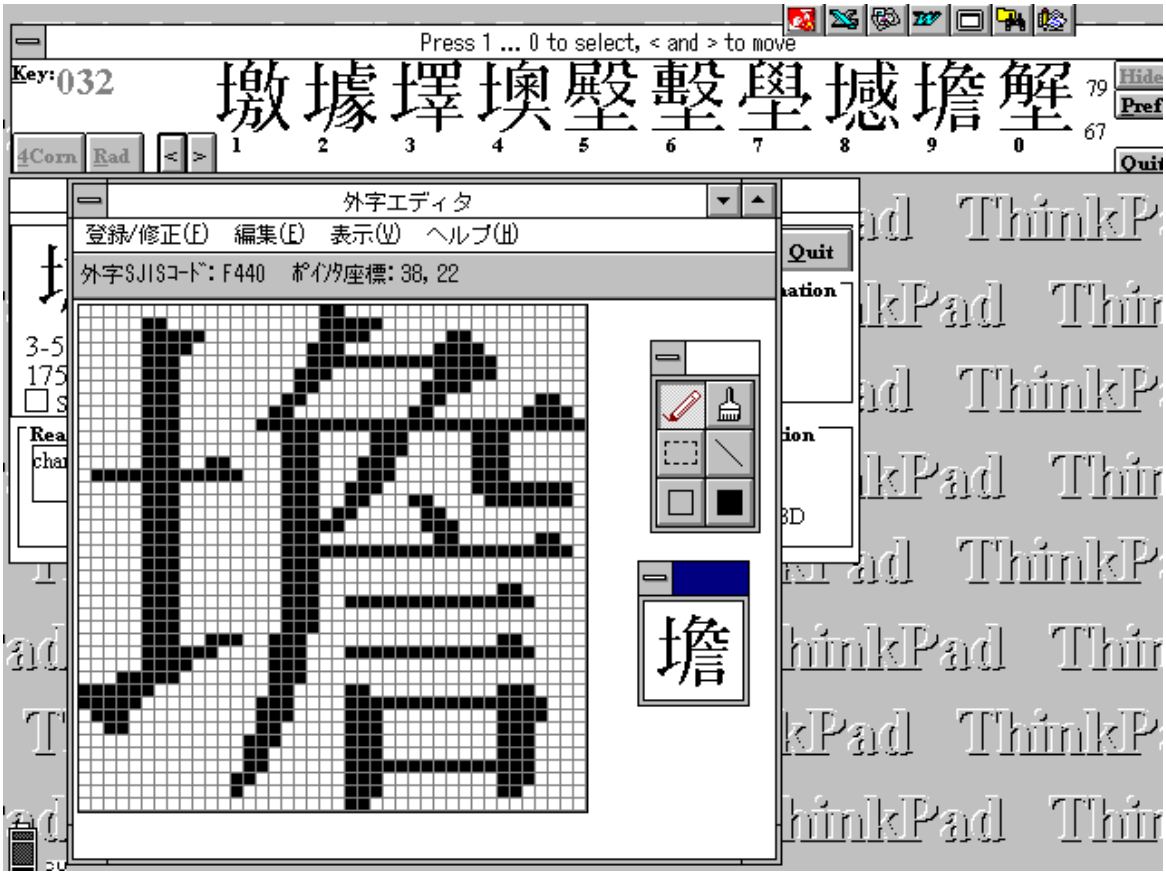

図２　IRIZ 漢字ベースの文字パターンを外字エディタ（Windows3.1）に張り付け

図２において、外字エディタ上にペーストされた文字パターンの状況を示し、シフト JIS コード "F440" の文字パターンとして編集している例である。

この文字パターンの登録作業を必要な外字すべてについて行い、保存する。この後、ディレクトリ C:¥Windows の下にあるフォント・ファイル "USERFONT.FON" を FD ディスク等に取り出す。なお、本登録作業は、Windows3.1 上で行うことである。