10　漢字符号の変換

植村　俊亮　（電子技術総合研究所）

0. はじめに

各種の漢字入出力機器の開発が進むにつれて，電子計算機で漢字かなまじり文を直接（ローマ字などに書きかえずに）処理することがさかんになりつつある．しかし漢字をふくめた標準的な符号（コード）系がまだ確立されていないので，各人，各機種，各社が独自の符号を使っているのが現状である．電子技術総合研究所で言語処理の研究のためになんらかの形で使っている漢字符号系が5種にもなったので，相互の間で自由に符号変換を行なえるプログラムを開発した．このプログラムは18の漢字符号系間の相互変換にまで拡張しうる汎用性のある変換ルーチンである．

1. 漢字符号系

「漢字符号」というのは厳密には正しい表現でない．漢字入出力装置はふつう漢字のみならず，ひらがな，カタカナ，英数字，特殊記号などをも処理する．各装置で漢字以外の文字がしめる割合は，全字数の10％ないし20％にのぼるが，漢字符号系の一部としてあつかう場合には，すでにJIS規格などがある文字についてもあらたに独自の符号を割り当てて，全体として1つの符号系を構成させる．したがってこれは一般的に「文字符号」系，あるいは「国際字符号」系とでも呼ぶべきものである．この変換プログラムも，そのような文字符号系間の変換を行なう．

2. 変換処理

2.1　5つの入出力装置とその符号系

今回の実験で直接とり扱った装置と文字符号系は下記の5つであった．

略称	装置	備考
漢テレ1	新興製作所製漢字テレタイプライタ	電子技術総合研究所　言語処理研究用（符号は独自のものを指定）
漢テレ2	〃　〃	〃
国語研符号	沖電気製漢字テレタイプライタ	国立国語研究所（新聞語い調査に使用）
沖ディスプレイ	沖電気製漢字ディスプレイ装置	おなじく電総研で研究用に使用（符号は既存のものをそのまま）
学研符号†	JEM 電算植字編集システム	

† 学研符号は紙テープによる入力符号（鍵盤）とプリンタ用符号とが異なるが，ここでは入力符号のみを扱った．

これらの符号系のほとんどは8ビットの符号を2回組み合わせた16ビットで文字1字を表現している（6ビット2回の12ビットの系もある）．各ビットが実際にどう使われているかを表に示す．

8ビット2列

略称	6ビット×2	2ビット×2
漢テレ1	JIS情報交換用符号のサブセットが2列††	使用せず
漢テレ2	ホレリス符号が2列	オ7(15)ビットは原則として使用せず オ8(16)ビットはパリティチェック用
国語研符号	ホレリス符号が2列	オ7(15)ビットはパリティチェック用 オ8(16)ビットは使用せず
沖ディスプレイ	JIS情報交換用符号のサブセットが2列††	オ7(15)ビットはけたずれ判定用 オ8(16)ビットはパリティチェック用
学研符号	特殊構成の6ビット符号が2列	オ7(15)，オ8(16)ビットとも，けたずれ判定用

†† まったく異なるサブセット

したがって，どの系でも文字の識別そのものをつかさどるのは16ビット中の12ビットにすぎない．学研符号にだけは，さらにシフト符号があるが，今回はこれを考慮しなかった．

2.2　変換アルゴリズム

入力符号は12ビットに圧縮（特定の4ビットを削除するだけ）して，それをそのまま変換表の探索のキーとして使う．出力符号は16ビットそのままとする．

2.3　変換プログラム

入力符号と出力符号の対応表はカードで与える．1文字に1枚のカードを対応させて，この文字に対する各系の符号を順に穿孔する．2重穿孔などのわずらわしさをさけ，またカード内容のラインプリンタへのEP刷を保証するために，カード上では文字符号を16進表記する．符号変換表作成ルーチンがこのカードを読み込んで，各系の間の変換表を順に作成し，磁気テープに記録する．利用者のプログラムから呼ぶ変換ルーチンが用意されており，このルーチンは，最初に呼ばれたときに磁気テープから必要な変換表を読み込んで，その表を使って変換を行なう．2度目以降は変換だけを行なう．どちらのルーチンもCOBOLでプログラムされている．

穿孔カード上の4ケタの欄が1つの符号系に対応しているので，18の符号系の相互変換にまで拡張可能である（4×18＝72，残りの8ケタはカード識別用に使用）．そのほかのいくつかの機能については次項でふれる．

3. 符号変換の問題点

いくつかの予期された（あるいは予期しない）問題が発生したので，それを列記する．

（1）対応する文字がない場合．一方の系に含まれている文字がもう一方の系には含まれてないので，入力符号は正しくても，出力符号に変換できないことがある．漢字であれば，よみがなに置きかえるのがもっとも望ましいが，ギリシャ文字や特殊記号になると，よみがなでは解決できない．このプログラムでは，対応する文字がない場合には原則として二（ゲタ）に置きかえ，もとの文字符号をラインプリンタにEP刷する．置きかえたい文字は変換表作成時に自由に指定できる．

（2）対応する文字が複数個ある場合．まったくおなじ英字の組を2組含む符号系があった．このような系では1つの文字が複数個の符号をもつことになり，変換の立場からみれば，もとの系のある文字が結果の系では複数個の文字に対応してしまうことになる．また，たて書きとよこ書きとを考慮した系があって，ある系ではよこ書き用の句点（。）だけが含まれるのに，べつの系ではよこ書き用の句点とたて書き用の句点（．）との両方を含むような例もある．このように対応する文字が複数個ある場合は，本プログラムでは変換表作成時にどちらか一方に制限する．

（3）対応する文字があるかどうか明確でない場合．たとえば3点リーダ（…）が4点リーダ（‥）になっている符号系があった．同じ字で書体が異なることもある．またある系では，20）のように，かっこ付きの文字が1字として扱われており，これは対応する系では3字としてなら表現（EP刷）可能であっても，1字分の符号を対応させることはできなかった．

（4）機能符号．文字符号系中に印刷上のいろいろの指定（ポイントや字間の指定）を機能符号として含んでいることがある．これらは個々の装置によって処理されるべきもので，変換プログラムの範囲外と考えられる．ただし復改符号や空白符号などの簡単な機能符号は変換する．

（5）入力不正符号の処理．変換を行なっていると，もとの系では使われていないはずの符号がなんらかの事情でたまたま現われることがある．このときは入力データの検査が必要になる．本プログラムでは，入力不正符号もなにか1字の符号に置きかえて変換され，もとの不正符号がラインプリンタにEP刷される．不正符号をどの符号に変換するかを変換表作成時に自由に指定できる．

4. まとめ

この変換プログラムは現在言語処理研究室の研究に活用されている．漢字を含めた一般的な文字符号の標準化が切望される．そのような標準がひろく普及したとき，このプログラムの寿命がおわる．標準化にあたっては，前述のような符号構成の原則とのかね合いをどこまで考慮するかが問題であろう．