

日立標準漢字と漢字コード

Hitachi's Standard KANJI Characters and Their Code

近年、漢字情報処理に寄せるユーザーの関心と期待は極めて大きい。

コンピュータシステムの中で漢字を処理するためには、字種を選定すること、コード体系を決めること及び個々の漢字がもつ情報をライブラリとして管理することが必要である。本稿では、今回日立製作所が開発した日立漢字情報処理システムでの漢字字種、字形、コード体系及び漢字情報を収容した漢字辞書ライブラリについて述べる。

これらは、日立漢字情報処理システムの基礎になるもので、ますます拡大する漢字処理の分野で多様なニーズに対応していくことができるものとする。

福来友康* *Fukurai Tomoyasu*

及川 巖* *Oikawa Iwao*

馬場徳夫* *Baba Tokuo*

松岡 潤** *Matsuoka Hiroshi*

1 緒 言

漢字をコンピュータで処理するためには、個々の漢字の文字図形、個々の漢字がもつ属性をコンピュータシステムの中に取り入れる必要がある。

本稿は、日立漢字情報処理システムでの

- (1) 文字種と文字字形の表現
 - (2) 文字コードとコンピュータシステム内での取扱い
 - (3) 漢字辞書ライブラリ
- について述べるものである。

漢字コードは、JISコードと容易に変換できるような体系を採用しており、ユーザーに提供する漢字辞書ライブラリも約7,000字を収容し豊富な属性をもっている。

漢字辞書ライブラリは、システムの中で使用する漢字及びその他記号に関する文字パターンと属性が収容されていて、漢字情報処理の基本になるものである。

2 日立標準文字種

2.1 漢字の字種

漢字をコンピュータで処理する場合、まず問題となるのは膨大な漢字のうちどれだけのものを選定、準備しなければならないかということであろう。字種が多いこと、字種の統一が困難なことが漢字情報処理を難しくしている。これについては、実際にどれだけの文字種が必要かという調査が過去幾つか実施されている。一般には、一つのアプリケーションの分野について見れば3,000～4,000字もあれば十分と言われているが¹⁾、各アプリケーション分野に共通な文字種は2,000字²⁾と少ない。特に人名、地名などの固有名詞を扱う分野では、特殊なものを除き、8,000字程度は必要とされている。各分野に使用される文字種をすべてひろい上げると8,000～10,000字程度になると言われている。このように多くの字種を扱わなければならないが、漢字機器に収容する字数は、そのシステムのコストや処理能力に密接に関連する。各分野でよく使用される漢字を選び出し標準と定め、その他の文字については、システム運用上必要に応じて追加収容することにより使用できるように、システムと漢字機器を設計しておくのが良いと考えられる。

2.2 日立標準文字種

日立漢字情報処理システムの標準提供文字種は、将来の漢

字情報処理での文字種の標準化を考慮して、日本工業規格(JIS)³⁾で選定した文字種に合わせた。

漢字機器に収容する文字を選定する場合、

- (1) 選定が適切なものであり、幅広いユーザー各層に受け入れられるものであること。
 - (2) 字種を選定しても、それが字種の限定にならないような、融通性のあるシステムにすること。
 - (3) 用途に応じて文字種の選択が、段階的に行なえるようにすること。
- などを考慮しておく必要がある。

JISで選定された文字種は、メーカー及びユーザーの漢字表、各種の漢字使用頻度調査結果に基づく漢字表をもとに選択されているので、一般性があると考えられる。

JISでは、使用頻度の高い文字群として、第1水準文字(非漢字453字、漢字2,965字)、次いで頻度の高い文字群として第2水準文字(漢字3,384字)を選定している。

日立製作所の漢字システムは、前記JIS選定文字のほかに、JISでは定めていない漢字システム固有の機能を実現するための記号類を追加し、日立標準文字種とした。

追加した文字種は、書式印刷文字と機能キャラクタ記号である。書式印刷文字とは、罫線、表などの書式を印刷するための特殊な図形文字である(例:□田国田)。機能キャラクタ記号とは、後の3.4で述べる機能キャラクタを印刷して表現するために図形化した文字である。また、字種全体を用途に応じて段階的に選べるようにレベルIとレベルIIとに大別した。日立標準文字種を図1に示す。日立標準文字種以外に、更に必要になる文字については、容易に追加ができるように機器及びシステムの仕様を定めた。

2.3 字形の表現

2.3.1 書 体

漢字の書体には、明朝体、ゴシック体、清朝体、隸書体、正楷書体など、多くの種類があり、また特定の書体についても活字メーカーにより差異がある。日立漢字システムの書体は、標準書体として広く用いられている明朝体を採用した。

2.3.2 ドットによるデザイン

コンピュータで漢字を出力する場合、個々の文字図形をドットの集合で表現することが多い。これは各ドットをコンピ

* 日立製作所神奈川工場 ** 日立製作所システム開発研究所

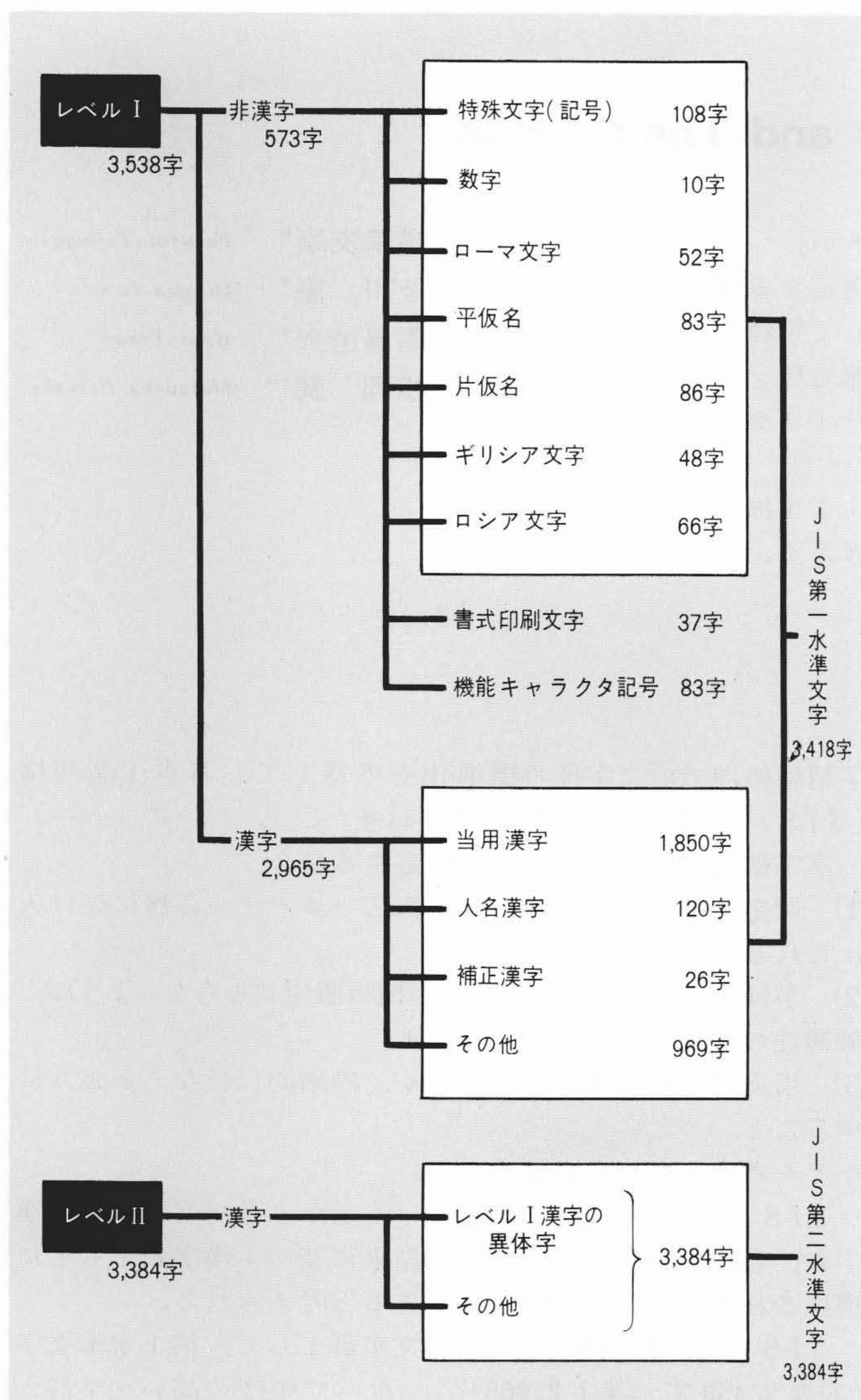


図1 日立標準文字種 日立標準文字種は、レベルⅠ、レベルⅡを合わせて6,922字である。

ユータの記憶単位，すなわちビットと対応させて処理することが便利であるからであり，特に文字の追加，修正及び外字処理（装置が所有していない文字を処理すること。本特集論文「漢字処理システムのハードウェア」参照）を可能とするのに有効な手段である。

文字図形をドットの集合で表現する場合に、一つの文字を何ドットで表現するかを定めなければならない。多くの文字を書体まで区別し鮮明に表わすためには、ドットの数が多いことが望ましいが、ドット数を増すことは、ドットを収容するメモリ容量が増え、ドットを取り出し文字を再現するハードウェアに要求される処理速度、複雑さが増すことになる。日立標準文字種をドットデザインした文字の印字例を図2に示す。

3 日立標準漢字コード

3.1 漢字コードの特異性

漢字処理では、多くの文字種を取り扱う際に、書体の指定、文字の大きさの指定、文字間隔の指定など、文字種以外に必要な付帯条件が従来の英・数字、片仮名処理に比べて多い。これはコンピュータ処理で出力した漢字の印刷物に対しても、通常の印刷物に近いものが要求されるからである。したがって、漢字システムの中では、個々の記号、英・数字、片仮名、漢字など（これらを図形文字と称する）を識別する膨大なコード群と、これら図形文字を取り扱う上で必要になる前述の付帯条件を区別するコード群（機能キャラクタと称する）を取り扱う必要がある。図形文字コードは、2バイトで表わされるのが普通であるが、コードそのものは統一されていない。機能キャラクタについては、種類及びコード体系のいずれもがまちまちであり、メーカー、ユーザー固有のものが用いられている。

3.2 日立標準漢字コードの特長

日立標準漢字コードの設定に際しては、以下に述べる点を考慮した。

- (1) コードの決定は、ハードウェア、ソフトウェアの仕様に

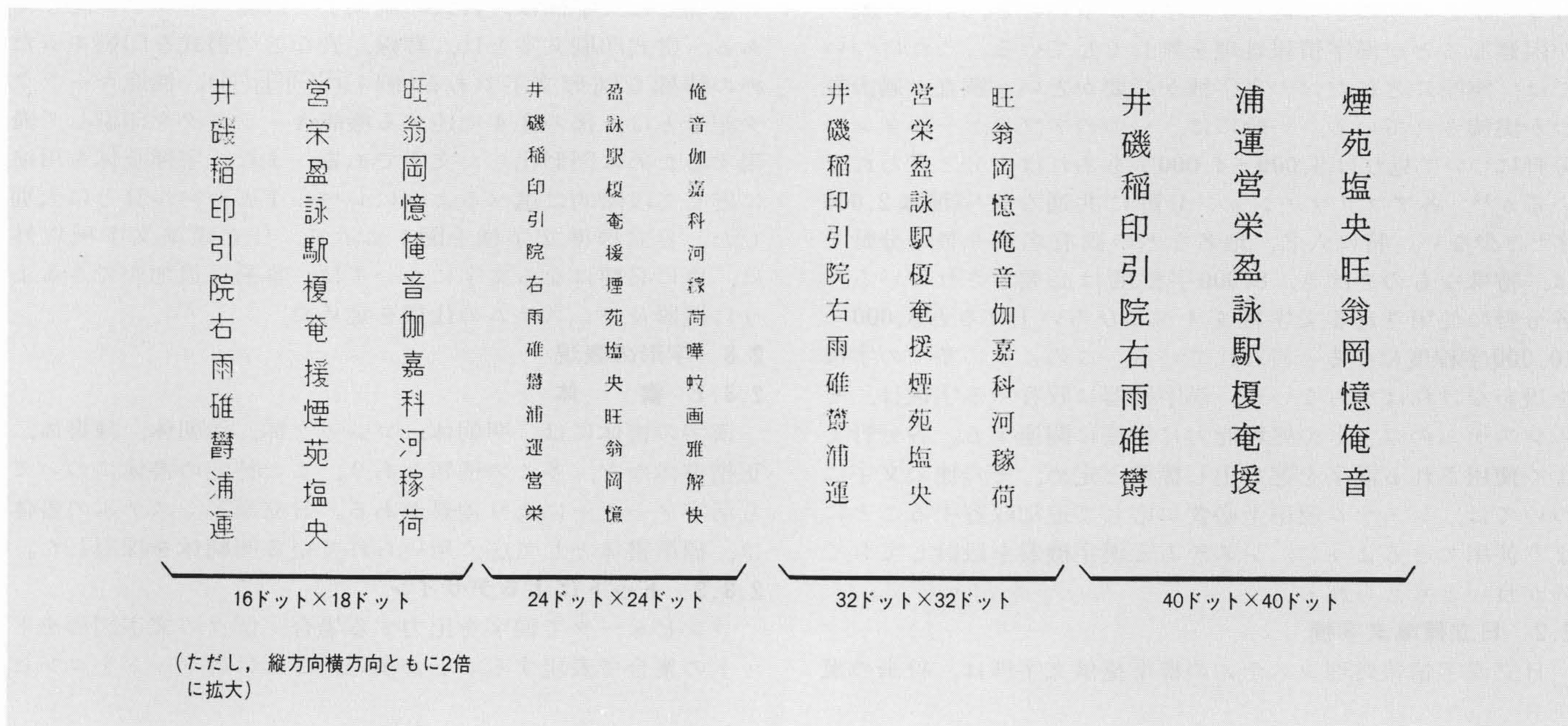


図 2 ドットによる文字の表現 各種のドット数の文字を、H-8195 漢字プリンタで印字したものを示す。

密接にかかわるので、将来の発展、拡張に対処できること。

(2) ユーザーが導入した漢字システム相互の間で、将来情報交換が行なわれることを想定し、このため必要になるコードの標準化に対処できること。

具体的には次に述べるようにした。

(a) 図形文字コードについては、JISに準拠して定める。

(b) 漢字をEBCDIKコード系(従来のコンピュータコード系)の中で取り扱うため、図形文字コードは、EBCDIKコード系の機能キャラクタコード域を避けて割り当てる。このため、JISコードとの間ではコード変換が必要になるが、簡単な交換処理で可能となるように定める。

(c) 機能キャラクタについてもEBCDIKコード系の中で取り扱うため、EBCDIKコード系の機能キャラクタコード域内に定める。ただし、多数の漢字用機能キャラクタのすべてを収容することは不可能なので、EBCDIKコード系の機能キャラクタコード領域中に、機能キャラクタ拡張コードを二つ新設しこのコードと図形文字コード領域のコードとを組み合わせるコード体系とした(漢字用機能キャラクタのJISは制定されていない)。

3.3 図形文字

3.3.1 コード領域上の配列

図形文字のコード領域上への配列については、JISに従い、次に述べるようにした。

- (1) 特殊文字……括弧、単位など同系統の記号を集めた配列。
- (2) 数字、英字……数字、英大文字、英小文字の順に配列。
- (3) 平仮名、片仮名……通常の国語辞典の順(小文字→清音→濁音→半濁音)
- (4) ロシア文字、ギリシア文字……通常の配列順
- (5) レベルI漢字……代表音による50音順。同音訓の場合には、音→訓。音又は訓が同じ場合は画数順。
- (6) レベルII漢字……部首別画数順、同画数の場合には、代表音の五十音順に配列。

レベルIとレベルII漢字の配列法は異なるが、使用頻度の高い文字(レベルI漢字)は音順、使用頻度の低い文字(レベルII漢字)は、字形による配列が検索に便利であると言われている。

3.3.2 JISコードとの対応

JISコードと日立漢字コードとは、 2^7 ビットによって対応づけている。JISコードの第1バイト、第2バイトの 2^7 ビットを“1”とすると日立標準漢字コードが得られ、逆の操作により、日立標準漢字コードからJISコードが得られる。図形文字のコード領域上の関連を図3に示す。

3.4 機能キャラクタ

3.4.1 機能キャラクタの設定

漢字情報処理では、従来の英・数字及び仮名処理に比べ多種の機能キャラクタの必要性が出てくる。出力用ハードウェアを制御する機能キャラクタとしては、文字の大きさ指定、書体指定、媒体上に印字又は表示される位置の制御などが必要になる。この中で位置の制御については、1文字単位に制御できる方式が柔軟性に富むが、これでは印字や表示の速度が遅くなる。

今回開発した漢字プリンタの位置の制御については、漢字プリンタの印字速度を確保するために、行単位に行なうことにした。すなわち、用紙送り方向の位置制御は行単位に行ない、1行内での行方向の位置制御は、文字単位に行なえるようにした。

簡易な文章の出力編集まで可能にすることを前提に、日立

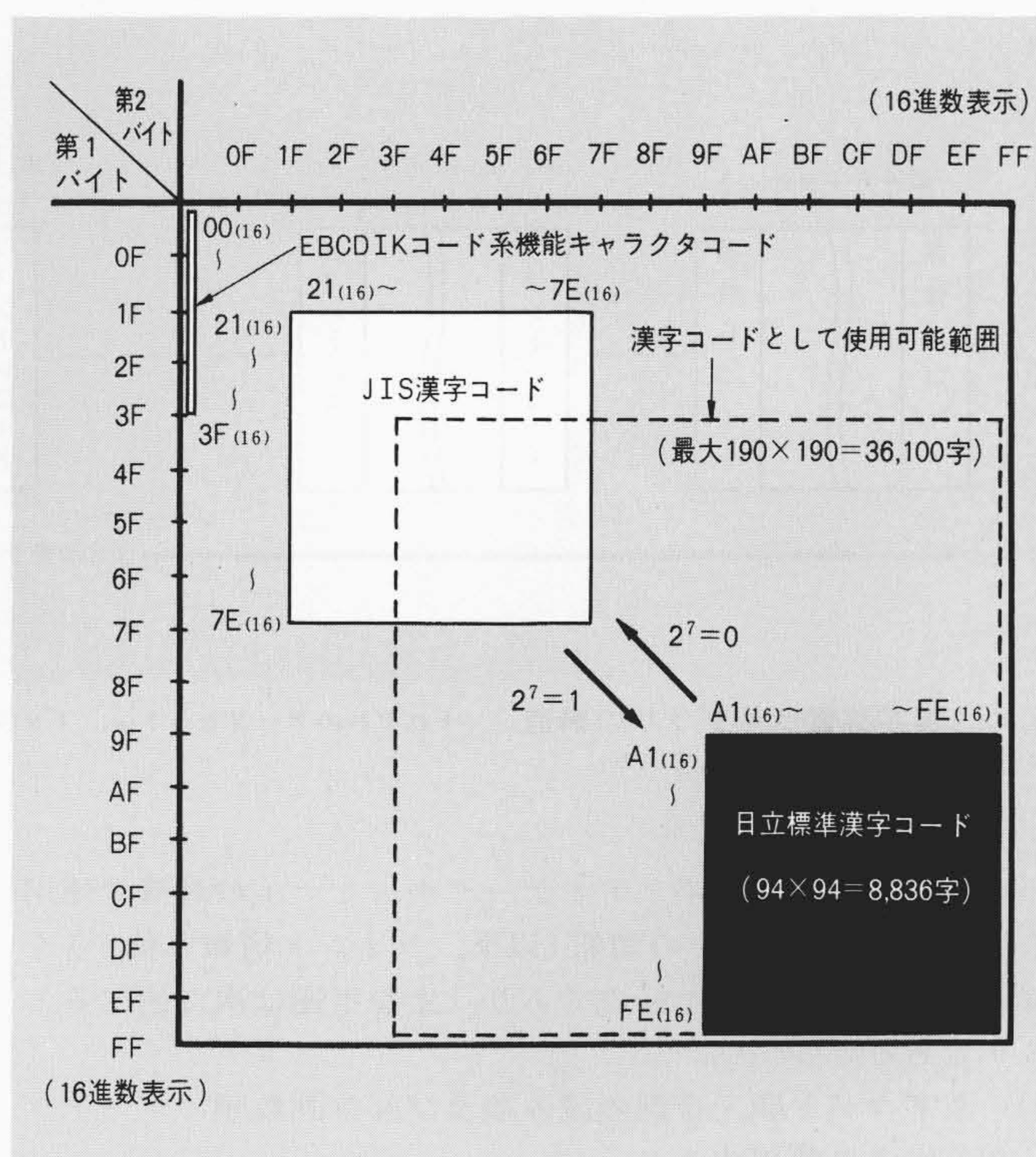


図3 日立標準漢字コード領域 漢字コードの第1バイト及び第2バイトを、それぞれ2桁の16進数(00₍₁₆₎からFF₍₁₆₎まで)で表わしたコード領域上での日立標準漢字コードの位置を示す。JIS漢字コードの 2^7 ビットを1にすると日立標準漢字コードとなる。

の漢字システムで今回設定した機能キャラクタを大別すると次に述べるようになる。

- (1) 漢字処理と、英・数字及び仮名処理(EBCDIK)の切り分け

例：漢字コード開始、EBCDIK開始

- (2) 漢字機器に対する指示

ハードウェア機能を指示するもの。

例：書体、文字サイズ、文字間隔などの指定

- (3) ソフトウェアに対する指示

ソフトウェアが行なう編集処理のため指示するもの。

例：柱文開始、柱文終了、ルビ開始、ルビ終了

- (4) 入力補助、その他

例：盤外字、パラメータ数(①, ②, ……⑦)

3.4.2 コードの割付

漢字をEBCDIKコード体系の中で取り扱うために、機能キャラクタ拡張コードはEBCDIKコード系の機能キャラクタコード領域の中に定めた。

0A₍₁₆₎、1A₍₁₆₎の二つのコードを割り付け、第2バイトで詳細機能を表わすようにした。拡張性、指示機能の柔軟性の確保のため、1A₍₁₆₎はパラメータ付き機能キャラクタとし、パラメータを第3バイト以下に置き、更に細かな機能指示はパラメータで行なうことにした。

4 漢字辞書ライブラリ

4.1 漢字情報処理のための漢字辞書ライブラリ

コードブックの作成のために、漢字を音訓の読みの順、あるいは部首・画数の順で並べる処理をコンピュータにより行なうためには、漢字のコードに加えて音訓の読み、部首及び画数の情報をデータ化したテーブル(辞書)が必要である。

日立製作所が開発した漢字辞書ライブラリは、漢字の音訓の読み、部首・画数などの情報(以下、漢字属性情報と称す)

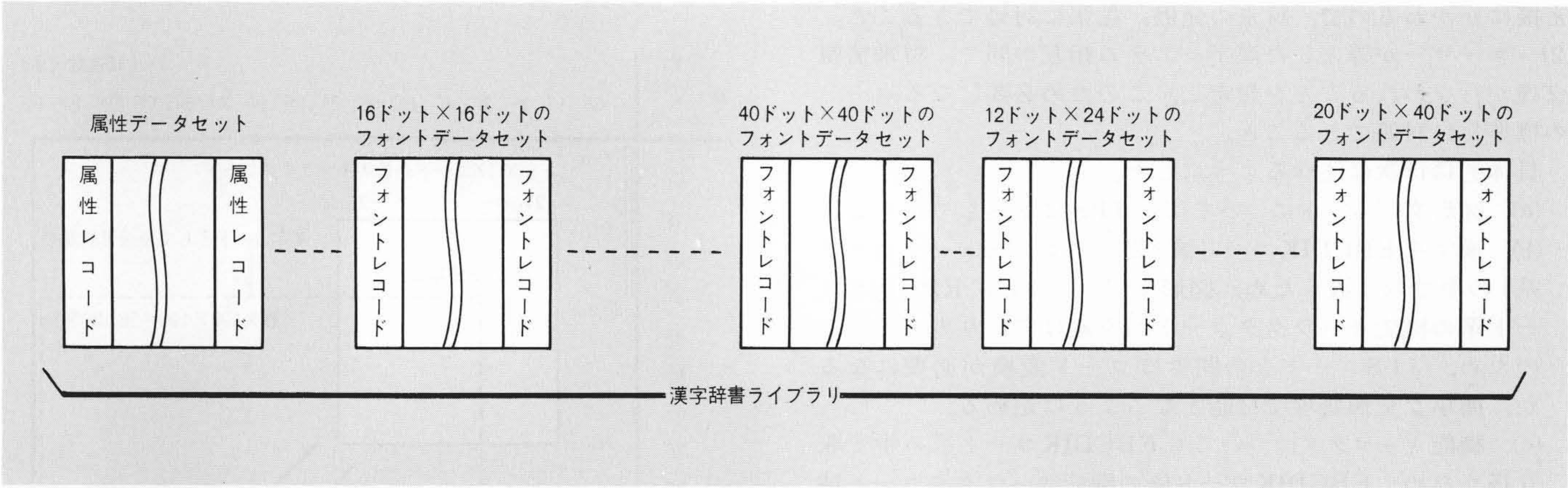


図4 漢字辞書ライブラリの構造 それぞれのデータセットは、1文字単位のレコード群から成り、コードの上昇順に並べられている。

と漢字プリンタ、漢字ビデオデータターミナルの装置で使用されるドットパターンの情報(以下、フォント情報と称す)を磁気テープに収容したものであり、主な用途は次に述べるとおりである。

- (1) 漢字コード順、音訓の読み順及び部首画数順のコードブックの作成に使用する。
- (2) 漢字プリンタ、漢字ビデオデータターミナルの外字用のフォントファイルとして使用する。
- (3) H-8196漢字プリンタでサポートする書式オーバーレイの文字パターンライブラリとして使用する。
- (4) JIS漢字コードテーブルの作成に使用する。

4.2 漢字辞書ライブラリの概要

収容されている漢字属性情報(属性データセットと称す)を表1に、漢字フォント情報(フォントデータセットと称す)を表2に、また漢字辞書ライブラリの構造をH-8195漢字プリンタ用フォントを例として図4にそれぞれ示す。漢字辞書ライブラリの主な特長は次に述べるとおりである。

- (1) 「漢字ライブラリ保守」ユーティリティを使用して、カード入力によりユーザー固有の属性情報及びフォント情報を追加することができ、ユーザー固有の漢字辞書ライブラリへ拡張することができる。
- (2) フォントデータセットを、ドット数、書体などの種別ごとに独立なデータセットとし、かつ1文字単位のレコード群構造としているので、システムに最適な外字処理用の漢字辞

表2 漢字フォント情報 漢字フォント情報は、フォントパターン、フォントパターンの識別情報(使用装置、ドット数など)、及びコード(全角文字用及び野線用は日立標準漢字コード、その他は日立標準EBCDIKコード)から成る。

装置名	ドット数(横×縦)	説明
H-8195漢字プリンタ	16×16, 20×20, 24×24, 32×32, 40×40	全角文字用
	12×24, 16×32, 20×40	半角文字用
H-8196漢字プリンタ	32×32	全角文字用
	16×32	半角文字用
	24×40	OCR用
	32×64	野線用
T560/40漢字ビデオデータターミナル	16×18	全角文字用
書式オーバーレイ用	16×18, 24×24, 32×32, 40×40	全角文字用
	8×18, 12×24, 16×32, 20×40	半角文字用

注：OCR＝光学文字読取装置

書ライブラリを容易に作成保守することが可能である。

5 結 言

本稿では、今回開発した日立漢字情報処理システムの文字の種類、文字字形の表現、文字コード及びそれらを登録した漢字辞書ライブラリについて述べた。

今回開発した漢字処理上の諸方式は、日立漢字情報処理システムの基礎になるもので、今後の発展に十分対処できるものである。

漢字をコンピュータで処理したいという要求は、今後ますます強くなるものと考えられる。日立製作所は、ユーザーのニーズに立脚した使いやすい漢字情報処理システムの発展のため、今後とも努力する考えである。

参考文献

- 1) 高橋、広田：計算機による日本語の処理、情報処理、10, No. 5, p.304 (昭44-9)
- 2) 日本情報処理開発センター、情報交換のための漢字符号の標準化に関する報告書、p.11 (昭50-3)
- 3) JIS C 6226情報交換用漢字符号系(昭53-1)

表1 漢字属性情報 収容文字種は、日立標準文字種である。

	収容情報	説明
漢字属性情報	日立標準漢字コード	JISと共通する文字種は、2 ⁰ ～2 ⁶ ビットがJIS7単位コードと同じ
	字典の検字番号	株式会社角川書店発行の「新字源」検字番号
	字種	JISによる図形文字の種別
	部首	—
	画数	—
	読み	音・訓の読み
	辞書の作成保守のための情報	—
	予備	ユーザー拡張用