日本語の点字情報に関する計算機処理（2）
―仮名点字の仮名漢字変換―

坂本義行　（電子技術総合研究所）

目次

0. はしがき

点字と普通文字の世界を結ぶ過程に計算機を導入することにより、正確で、高能率、経済的な処理体系、点字情報処理の実用化が待たれている。とくに、自動代筆と自動点訳のシステム開発が求められている。

視覚障害者が、直接点字タイプライタを打鍵することにより、点字紙上に点字モニタが得られると同時に、穿孔された8単位の紙テープが得られる。これを計算機への入力とし、プログラムにより普通文字へ変換し、出力する方法を自動代筆と呼ぶ。

点字の世界でも、普通文字の世界と同様に、日本と欧米とでは、その表記法が異なっており、日本語が漢字、片仮名、平仮名といった多種の文字を使用している。しかし先天的な視覚障害者に漢字を教えることは非常に困難であるため、日本独自のBrailleの6点点字に仮名を当てはめた仮名点字表記法が一般に用いられている。既報の「日本語の点字情報に関する計算機処理（1）」において、長谷川氏[1]考案の漢字点字符号系を用いる自動代筆システムについて述べたが、汎用性の面から、本システムは、仮名点字を普通文字の漢字仮名混じり文に変換するシステムについて述べる。

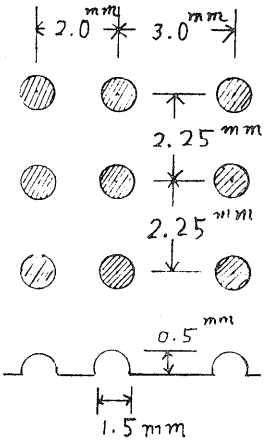
ここで採用した仮名点字体系は、できる限り日本点字表記法ならびに「点字毎日」の表記法に準ずるが、54年度に改定予定の表記法及び表記法との矛盾する点は修正を加えたものである。

「点字毎日」の3部に当たる約10万字を標本として入力し、onlineのTTP端末上に処理結果を出力するシステムを開発した。なお、入力を効率的に行なうため、汎用の複合用語に対する速記処理機能が付加されている。

1　点字符号体系

1.1　ブライユの点字符号

現在世界で使用されている6点式点字（第1図）は、1825年頃フランスの視覚障害者ルイ：ブライユ（Louis Braille）により考案されたもので、縦3点、横2点の6点（マス）で構成される63種の符号にアルファベットを割り当て、数字、大文字、特殊記号等はこの符号の組み合せで表わす。この符号の連結（concatenated）表現が習慣化しているため、後述の漢字への拡張において心理的に大きな障害とならない。



第1図　点字規格

1.2　仮名点字符号

日本においては、1890年に官立東京盲学校（筑波大学雛司ヶ谷分校）の教官、石川倉次氏により、ブライユの点字に仮名を当てはめたものが使われるようになった。石川氏は、ブライユの点字配列表と仮名の五十音を巧みに調和させ、原則的で、理解しやすく、指の触覚で読みやすい点字体系をつくりあげた。以来、日本の点字体系は仮名表記となっている。日本のような漢字文化の中で、視覚障害者だけが漢字をもたなかったことに問題もある。少なくとも後天的な失明者には漢字が理解できる。

1.3　総合点字の漢字符号

自動代筆、自動点訳を行なう場合、普通の文字体系に対応した点字体系があると、1対1の符号変換ができ、完全な自動代筆、自動点訳が可能となる。そこで長谷川点夫氏（筑波大学　雛司ヶ谷分校）は、現行の仮名点字体系を基本とし、音節を中心に3または4マスの組み合わせで漢字を符号化することにより、日本語を表現する方法を考案した。

漢字は、8種類の漢字符を前置して表わす。このうち、7種類は3マスで、1種類は4マスで表わす。

表記法として、視覚的な特徴である部首の組み合わせを用いることも考えられるが、視覚障害者の使用を前提とすると、音や訓などヨミと直結する要素をもとにする方が良いと考えられる。この方法は、視覚障害者がすでに持っている仮名点字体系の構成要素を利用できる点、符号の連結表現が習慣化している点から有利である。従来の漢字符号体系と異なる点は、漢字のヨミにおける末尾の音節に着目した点、指の認識能力を考慮した連結表現といった点である。第1表にその分類を示す。

第1表　　漢字の分類表

		音節による分類項目	個数（当用漢字）	例
A	1	1音節からなる	360	過，季，苦，呼，左，系，粗
	2	第2音節がイまたはウとなる	491	会，空，計，公，再，青，相
	3	〃　キ　〃　ク　〃	174	作，式，石，足，沢，竹，的
	4	〃　ツとなる	102	割，詰，屈，穴，骨，札，室
	5	〃　ン　〃	382	安，飲，雲，円，温，慣，金
	6	拗音を含む	317	者，主，処，邪，受，女，小
B	7	2音節からなる訓	12	竿，涼，箱，姫，坪，株，刈
	8	3　〃	10	扱，娘，芝，届，卸，峠，畑

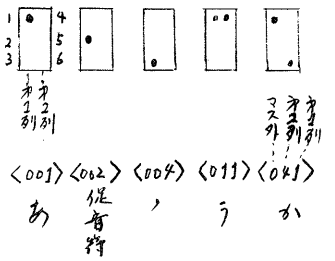
A. 音を中心とする漢字：第1マスは漢字符、第2マスは音の第1音節を表わす仮名が入り、第3マスには同音の漢字を区別する仮名、漢字に訓がある場合には、その第1音節を当て、訓がない場合にはその漢字と結合して熟語をつくる漢字の第1音節の音を割り当てる。

B. 訓を中心とする漢字：第1マスの漢字符とそれに連結する2または3マスの訓を表わす仮名より成る。

以上の方法により、約1万字の点字を構成することが可能である。しかし例外として、当用漢字の中で約10%の漢字では、同音の衝突が起るため、特定の符号が割り当てられており、訓のみを有する1字訓の漢字は、Aの1に割り当てられている。

2. 仮名点字表記法

本符号系は、仮名、数字、アルファベット、特殊記号、及び「点字毎日」で用いられている点字パターンを利用した模様などは、そのまゝ図形パターンとして含む。以上の符号は1個または複数個のマスによって表現され、紙テープ上第1ビットから第6ビットで表現される。第7、第8ビットのみで、点字の改頁、改行、間隔といった点字機能符号として表現する。点字の1マスおよび紙テープの1フレームを8進数3桁で表わすこととする。付表（A）にその符号表を示す。



第2図 各点の表記

2.1 仮名点字の特徴

1) 仮名、数字、アルファベット（大、小）、特殊記号からなる。
2) 特殊な表記文字　「は」→「わ」、「へ」→「え」、「ゆうびん」→「ゆ－びん」
3) 分かち書きは小学校低学年の教科書のものに最も近く、原則として文節分かち書きであるが、複合名詞等では例外が多い。

3. 仮名漢字変換の方法

ローマ字または仮名で日本語テキストを入力し、何らかの方法で漢字仮名混じり文に変換することは、言語学的分析を必要とし、将来音声的な分析面からも広く発展する可能性をもっている。

仮名漢字変換を自動化するには、仮名文をある単位（語、文節）に区切る手順と同音異義語の中から適正な1語を選ぶ効果的な手続きを開発せねばならない。

その試みは、1966年頃から　九大：栗原氏等によって始められ、種々の方法が考案され、紹介されてきたが、現在までのところ実用化にはいたっていないといえる。

仮名漢字変換で問題となる点を項目として次に列挙してみる。

1) 処理対象－各々処理対象を決めて、語彙分析を行ない辞書を作成する。
2) 分かち書き－1, 2の例を除いては、すべて人間による文節レベルの分かち書き

3) 辞書 — 数十～数万の自立語辞書（見出し、漢字、構文情報）と接続表（付属語中心）
4) 同音異義語 — 判定不能の場合は、並記または原文を出力
5) 意味処理 — 非常に簡単な意味分類等による判定を行なっているものがある.
6) 変換率 — 一般に70%以上、専門語の辞書を用いることにより80%以上の変換率を得たものもある.
7) 後処理 — 出力された後で、追加表示選択により正しい文にする機能を付加しているものもある.

開発されたシステムの特徴をみると、
1) 九大グループ — もっとも早い、文節分かち書き、辞書5万語
2) NHK — ニュースを対象、辞書6千5百、意味分析を行なう、変換率77.5%（90.1% 第1順位）
3) 九芸工大 — 辞書5万、係り受け（名詞と名詞の接続）変換結果不明
4) 沖電気 — 辞書（汎用15.000 私用）、通信電文、ローマ字入力、語単位変換と熟語単位変換の相違、変換率87%
5) 国語研 — 読み仮名方式、文字の左右を照合、処理速度2万字/時間
6) 日本情報処理開発センタ — 固有名詞辞書1万7千、医学分野、分かち書きにファンクションキーを挿入、変換率71%（82%、特に固有名詞辞書を拡張）
7) 通研 — 多文節にわたる構文解析、辞書3万語、漢字処理、変換率84%（文字単位93%）、変換速度2.5文節/秒
8) 阪大（木沢他）— 自立語付属語分かち書き、変換率90%
9) 電総研1（植村）— 辞書7万（新聞用語）
10) 電総研2（坂本）— 辞書8千（新聞用語）、接続表（左右の文字）、後処理英字新聞、英字分かち書き（文節）.

以上 非常に簡単にいくつかの方法について、特徴をあげてみたが、網羅的ではない.

4 変換辞書

仮名漢字変換において最も重要なのは、その変換辞書の構成である、処理対象が英字新聞である点から、下記の2種類の辞書について比較を行なった.
1) 国研報告87「電子計算機による新聞の語彙調査」所載の簡易五十音順長単位表（頻度6以上、約11.000語）
2) 新明解国語辞典（三省堂）約80.000語

この比較表を第2表に示す.

この表からも解かるように、2)の資料がはるかに多くの語彙を有しているが、1)に比べて、殆んどの見出しが、2個以上の同音異義語を有している. しかも同音異義語中で、一般には非常に出現頻度の低いと思われる語彙を多く含んでいることがわかる. 処理対象が英字新聞である点、又後述するように、厳密な構文分析および意味解析を行なわない点から、2)を辞書として採用することは同音異

義語の判別を困難にし、変換効率がむしろ低下することが考えられる。

表2表　辞書の語彙比較

見出し	① 新聞の語彙		② 新明解国語辞典	
コトバノ	言葉	の	言葉	の
ウエダケデ	上	だけで	① 上　② 飢え	だけで
モノゴトヲ	ものごとを		物事	を
セツメイシ	説明	し	説明	し
ジッタイヲ	実態	を	① 実体　② 実態	を
リカイシテ	理解	して	① 理会　② 理解	して
イナイ	い	ない	い（居，射，鋳）	ない
コトヲ	① こと　② 琴	を	① 異　② 事　③ 琴　④ 古都	を
イウ	いう		言	う
ヨウデス	よう	です	よう	です

本システムでは、1)の資料を基準に、以下のような修正を施したものを、変換用の自立語辞書とした。

1) 自立語辞書　　約 8,000 語

見出し　基本として新聞の語彙（長単位）　約 11,000 語

修正項目

a) 数詞は原則として除き、短単位表中の助数詞を加えた。

b) 付属語は抽出して、付属語辞書に登録した。
ただし、表字分から書き法に従い、形式名詞、補助動詞、助動詞（ようだ）等は残した。又接尾辞の特殊なものは、両方の辞書に登録した。

c) 固有名詞

i) 地名　行政区画名は一旦全部除いた上、都道府県名、全国の市、東京の区を加える。国名、外国の地名、駅名、鉄道路線名等は、表中のものをそのまま残す。

ii) 人名　一旦全部除いた上で、姓についてのみ、頻度順リストを用いて上位 1,000 を改めて收録する。

iii) 組織名　企業名略称は除き、他は残す。

iv) その他　映画、テレビ番組、書物の題名と思しきものを除く。

d) 略語
又字を見なければ意味の分らないような略語は除く。
例、細面、新同　自光、自現

e) 接頭辞つきの語
特に慣用性の高いもの（例、お客様、お母さん）を除いて、接頭辞はずしたもので登録した。

文法情報　前接情報（47）、後接情報（40）

表記　漢字仮名混じり、平仮名、片仮名

配列　同音異義語は頻度順に配列した。

2) 付属語辞書　　見出し、表記（平仮名）、文法情報　　173 語

3) 接続表　　前接情報（49）× 後接情報（40）のクロス・リファレンス・テーブル

5. 仮名漢字の自動代筆システム

5.1 入力テキストの作成

「漢字毎日」3部、約10万字について漢字シートを作成した。この資料作成の段階では、両面漢字の光学的な自動漢字読取り装置がなく、人手によって片面漢字シートを作成し、これを自動読取り装置にかけ、紙テープを作成した。校正を行なったものを入力テキストとした。

5.2 変換用入力テキストの作成

紙テープを読み込み、ファイル上に蓄積した後、FONT 4000 コードの平仮名表記によるテキストに変換した。さらに漢字用の特殊記号処理を施したものを変換処理への入力テキストとした。テキスト中、句点(。)、読点(、)、括弧等は、独立の文節として分離されていないため、その分割処理を行なう必要があり、前処理として、特殊記号処理を行なった。

5.3 変換処理

各文節単位に自立語辞書の見出しとマッチングをとり、全ての解をスタックする。スタックの上から順に取り出し、文節の残余のストリングを切り出す。このストリングに付属語辞書サーチ、接続情報によるチェックを行ない、文節末まで分析できたら結果をスタックする。途中で行きづまったら引き返し処理を行ない、スタックが空になるまで続ける。解が1つもなければ、原形表示とする。辞書の見出しは複合語が含まれているため、処理は必ずしも入力文節を単位に行われず、2つ以上の文節にまたがることもあるが、常に文節頭から始まって文節末で終る。

複合語を先頭とする解析が合格したときは、以後の処理を打ちきる。

5.4 出力表示

変換された普通文字は、FONT 4000 にてユーザファイル上に記憶されると同時にTSS画像端末上に、漢字かな混じり文で出力される。同音異義語の処理は全語出力又は頻度が最大のもの1語のみを出力するかの2種類の選択が可能である。また、入力テキストを検索し、要求した文のみを変換し、出力することも可能である。

6. 実験と結果

変換処理の実験は、全体の結果はまだ得られていないが、452文節について行なった結果、誤り文節92個が得られ、正解率としては80%程度である。出力例として、選択モードと全数モードについて第2図に示す。

誤り文節の内訳は、第3表のような種類である。表中、漢字新聞に出現すると思われる用語（◎印を付した項目）を専門の辞書として登録することにより、90%に近い正解率が得られることがわかる。しかし一般的な用語と思われる「弁論」、「郵便」、「番号」等は新聞語彙の統計表上では、出現頻度6以下であり、異なり総数は21万個にもなり、「弁論」は順位で2万以上である。すなわち、辞書が急激に大きなものとなる。同時に、同音異義語を増加することとなる。むしろ、処理するテキストの分野を限定することにより、辞書検索を容易にすべきだと思われる。

《00001》
てんじ　まいにち
展示　毎日
《00002》
だい2878ごー　しょーわ　53ねん　7がつ　ここのか　（　にちよーび　）　ていか　1ぶ　200えん
第2878号　昭和　53年　7月　9日　（　日曜日　）　定価　1部　200円
《00003》
はんとし　5200えん　はっこーじょ
半年　5200円　?????
《00004》
おーさかし　きたく　どーじま　1ちょーめ　6ばん　20
大阪市　北区　?????　1丁目　6番　20
《00005》
ゆーびん　ばんごー　530　まいにち　しんぶんしゃ　（　ふりかえ
?????　?????　530　毎日　新聞社　（　振替
《00006》
おーさか　450　）
大阪　450　）
《00007》
でんわ　（　06　）　343の　1121
電話　（　06　）　?????　1121
《00008》
2121　。　（　てんまい　きじの　もだん　てんきいわ　おことわりします　）　。
2121　。　（　?????　記事の　?????　?????　?????　）　。
《00009》
ーこたー　こんしゅーの　おもな　ないよー
?????　今週の　主な　内容

つぎつぎと　次々と
ぜんこく　全国
たいかいえ　大会へ
《00011》
2　もーじんの　どくしょけん　かくほに　しん
2　2
もーじんの　?????
どくしょけん　?????
かくほに　確保に
しんていあん　?????
……　……
4　4
こーちょーかい　?????
とくべついが　?????
ちゅーかん　中間
ほーこく　報告

a）全数出力

b）選択出力

第2図　変換出力例

第3表　誤った文節の種類

種類	延べ数	異なり数	種類	延べ数	異なり数
点字用語	23	18	外来語	1	1
特定分野の用語	39	36	複合語	2	2
固有名詞	7	6	接頭辞	1	1
文節でない	6	4	助詞	1	1
一般用語	6	5	その他	3	2
省略語	1	1			
選択誤り	2	2	合計	92	77

最後に、点字情報について、多くの御指導、貴重な資料をたまわった長谷川良夫先生、また本研究の機会を与えて下された石井治ソフトウエア部長、鳥居宏次言語処理研究室長、計算の作成およびプログラム作成を行なったIBSの木村睦子さん、松村氏、ならびに東京理科大学の寺文宇君に感謝いたします。

参考文献

(1) 長谷川良夫　「視覚障害者に必要な点字情報処理」　昭和52年電気4学会連合大会　1977. 10
(2) 坂本義行　「日本語の点字情報に関する計算機処理(1)」
－Braille 符号と漢字の変換処理－　計算言語学研究会資料　12－2　1977. 10

付表（A） 総合点字の仮名符号表

機能符号			
点字改頁 〈100〉	ファイル開始 〈030〉〈042〉	特殊機能 〈040〉	
点字改行 〈200〉	ファイル終了 〈030〉〈046〉	{〈004〉〈100〉〈200〉〈300〉}	
点字間隔 〈300〉			

仮名符号

清音符

あ 〈001〉	い 〈003〉	う 〈011〉	え 〈013〉	お 〈012〉
か 〈041〉	き 〈043〉	く 〈051〉	け 〈053〉	こ 〈052〉
さ 〈061〉	し 〈063〉	す 〈071〉	せ 〈073〉	そ 〈072〉
た 〈025〉	ち 〈027〉	つ 〈035〉	て 〈037〉	と 〈036〉
な 〈005〉	に 〈007〉	ぬ 〈015〉	ね 〈017〉	の 〈016〉
は 〈045〉	ひ 〈047〉	ふ 〈055〉	へ 〈057〉	ほ 〈056〉
ま 〈065〉	み 〈067〉	む 〈075〉	め 〈077〉	も 〈076〉
や 〈014〉		ゆ 〈054〉		よ 〈034〉
ら 〈021〉	り 〈023〉	る 〈031〉	れ 〈033〉	ろ 〈032〉
わ 〈004〉	ゐ 〈020〉〈006〉	ゑ 〈020〉〈026〉	を 〈024〉	ん 〈064〉

濁音符 〈020〉〈（清音符）〉

半濁音符 〈040〉〈（ハ行清音符）〉

拗半濁音符 ぴゃ 〈050〉〈045〉 ぴゅ 〈050〉〈055〉 ぴょ 〈050〉〈056〉

仮名符号

拗音符 〈010〉{〈(1)〉|〈(2)〉}

1	きゃ 〈041〉	しゃ 〈061〉	ちゃ 〈025〉	にゃ 〈005〉*	ひゃ 〈045〉
	きゅ 〈051〉	しゅ 〈071〉	ちゅ 〈035〉	にゅ 〈015〉*	ひゅ 〈055〉
	きょ 〈052〉	しょ 〈072〉	ちょ 〈036〉	にょ 〈016〉*	ひょ 〈056〉
2	みゃ 〈065〉	りゃ 〈021〉			
	みゅ 〈075〉	りゅ 〈031〉			
	みょ 〈076〉	りょ 〈032〉			

拗濁音符 〈030〉〈(1)〉 （但し * は除外）

小文字符

や 〈030〉〈014〉	よ 〈030〉〈034〉	い 〈030〉〈003〉	え 〈030〉〈013〉
ゆ 〈030〉〈054〉	あ 〈030〉〈001〉	う 〈030〉〈011〉	わ 〈030〉〈004〉
か 〈040〉〈040〉〈041〉	が 〈040〉〈020〉〈041〉		

外来音符

うぃ 〈042〉〈003〉	つぁ 〈042〉〈025〉	ふぁ 〈042〉〈045〉	ゔぁ 〈062〉〈045〉	
うぇ 〈042〉〈013〉	つぃ 〈042〉〈027〉	ふぃ 〈042〉〈047〉	ゔぃ 〈062〉〈047〉	
うぉ 〈042〉〈012〉	つぇ 〈042〉〈037〉	ふぇ 〈042〉〈057〉	ゔぇ 〈062〉〈057〉	
くぁ 〈042〉〈041〉	つぉ 〈042〉〈036〉	ふぉ 〈042〉〈056〉	ゔぉ 〈062〉〈056〉	
ぐぁ 〈062〉〈041〉	しぇ 〈010〉〈073〉	てぃ 〈040〉〈027〉	とぅ 〈040〉〈035〉	
てゅ 〈050〉〈035〉	じぇ 〈030〉〈073〉	でぃ 〈030〉〈027〉	どぅ 〈062〉〈035〉	
ゔ 〈020〉〈011〉	でゅ 〈070〉〈035〉	でぇ 〈010〉〈037〉		
ちぇ 〈010〉〈027〉	いぇ 〈010〉〈013〉			

促音符 〈002〉|〈002〉〈002〉

算用数字符号

〈（算用数字開始シフト符号）〉〈(#)〉〈（算用数字終了シフト符号）〉

開始シフト符号 〈074〉　終了シフト符号 (#) 以外の符号

# 1 〈001〉	4 〈031〉	7 〈033〉	0 〈032〉	
# 2 〈003〉	5 〈021〉	8 〈023〉	小数点 〈002〉	
# 3 〈011〉	6 〈013〉	9 〈012〉	桁区切り 〈004〉	

仮名符号

特殊記号

句点	〈062〉〈(3)〉	3	〈100〉
読点	〈060〉〈(3)〉		〈200〉
疑問符	〈042〉〈(3)〉		〈200〉〈200〉
感嘆符	〈026〉〈(3)〉		〈300〉
中点	〈020〉{〈100〉|〈200〉|〈300〉}		
繰り返し符号(清)	〈040〉〈046〉〈002〉		
繰り返し符号(濁)	〈060〉〈046〉〈002〉		
対比符号	〈020〉〈002〉		
開かぎ括弧	〈060〉〈044〉〈004〉		
閉かぎ括弧	〈040〉〈044〉〈006〉		
開丸括弧	〈066〉		
閉丸括弧	〈020〉〈066〉		
長音符	〈022〉		
改行符号	〈030〉〈062〉		
改頁符号	〈030〉〈024〉		

消去(句点に後置した部分)	〈030〉〈062〉
消去(読点 〃)	〈030〉〈066〉
消去(改頁 〃)	〈030〉〈006〉
消去(改行 〃)	〈030〉〈064〉
消去(五字改頁 〃)	〈030〉〈007〉
消去(五字改行 〃)	〈030〉〈005〉
抹消	〈377〉|〈077〉〈077〉〈077〉

英文字符号

シフト: 〈(英字開始シフト符号)〉〈(5)〉〈(英字終了シフト符号)〉
開始シフト符号 〈046〉|〈060〉 終了シフト符号 〈064〉

a 〈001〉	f 〈013〉	k 〈005〉	p 〈017〉	u 〈045〉	z 〈065〉	
b 〈003〉	g 〈033〉	l 〈007〉	q 〈037〉	v 〈047〉		
c 〈011〉	h 〈023〉	m 〈015〉	r 〈027〉	w 〈072〉		
d 〈031〉	i 〈012〉	n 〈035〉	s 〈016〉	x 〈055〉		
e 〈021〉	j 〈032〉	o 〈025〉	t 〈036〉	y 〈075〉		

大文字: 1文字: 〈040〉を前置　連続: 〈040〉〈040〉〈小文字〉〈小文字以外〉

特殊記号	. 〈062〉	; 〈006〉	? 〈046〉	) 〈020〉〈066〉	“ 〈040〉〈046〉
	, 〈002〉	: 〈022〉	- 〈044〉	‘ 〈020〉〈046〉	” 〈040〉〈064〉
	’ 〈004〉	! 〈026〉	(〈066〉	’ 〈020〉〈064〉	