「ことだま」文書処理システムの文節わかち書き仮名漢字変換†

藤崎哲之助†† 大河内正明†† 諸橋正幸††

近年，カナ漢字変換の手法を用いた日本文の入力方式が種々提案されており，それに基づくワード・プロセッサも実用機が出現している．しかし，カナ漢字変換は，少なくとも現在の技術水準では，変換誤りの発生することを避けることができない．したがって，その意味で漢テレ方式，その他の漢字直接打鍵，および，英文の入力方式と本質的に異なり，入力系全体における人間の介入およびその負担を人間工学的側面より考慮し，設計を行う必要がある．また，特に打鍵者への負担が少ないといわれる文節わかち書き方式においては，表記のゆれ，同音語，言葉の学習，誤変換の訂正などにおいて多くの重要な問題が発生する．本稿では，著者らの試作システムに基づき，仮名漢字変換方式の設計に当っての人間工学的側面からの考慮を述べるとともに，特に文節わかち書き方式における問題点，その解決案を議論する．

1. まえがき

近年，日本語に対する文書処理の要求が高まり，入力・出力・処理に関する数多くの提案がなされてきているが，特に，日本文の入力においては，仮名漢字変換の手法を用いる入力方式が，実用化されているものも含めて多く提案されている[1]．

仮名漢字変換方式の入力法は，タブレット方式などのいわゆるSIGHT方式の漢字入力方式に比べて，タッチ・タイピング[2]による高速度打鍵が可能であり，また，連想コード方式，漢テレ方式に比べて，ある程度の入力速度を習得するに要する時間が少なくてすむ，などの利点があると言われている．したがって，将来のオフィスオートメーションで，管理者・担当者などの文書打鍵の非専門家が文書の入力を行うことにも適したものとして期待されている[3]．

しかし，その反面，タブレット方式，連想コード方式，漢テレ方式などの漢字直接入力法に対し，計算機による仮名文字列より漢字仮名混じり表記への変換の過程が含まれる漢字の間接入力方式であるため，計算機による変換の効率と精度の問題，変換に誤りが生ずることが避けられないための人間の介入の問題などが生ずる．

また，鍵盤からの入力においても，仮名文字列をべた打ちした場合[4]はそこに生ずるあいまいさの程度が大きいため現時点では実用上困難がある．(「キノウハイシャヘイッタ」は「昨日歯医者へ行った.」，「昨日は医者へ行った.」のいずれにも解釈可能である.) したがって，入力においては，文章の読み方を与えると同時に，なんらかの補助的情報を与えるのが通常となっている．この補助的な情報としては，字種の指定を与える字種指定方式，空白を適当な単位（漢字単位，言葉単位，文節単位など）ごとに挿入するわかち書き方式がある[5]．しかし，この補助的な情報はあくまで人工的な規則に基づくものであるため，なんらかの意味で打鍵者の負担となり，入力時の効率と，変換処理への補助の程度のかね合いに考慮すべき問題がある．特にこれらの問題は，漢字の直接入力法では生じないものだけに，これらの問題への対処が重要であって，仮名漢字変換に基づく文書の入力方式の成功の重要な要因となる．

筆者らは，日本語文書の作成・編集・検索・索引作成などを一貫して行う日本語文書処理システム「ことだま」の研究試作を進めてきており[6]，その一部として，文節わかち書きの仮名漢字変換を利用した日本語エディタを開発した．

本論文では，その経験に基づき，仮名漢字変換における入力系全体としての問題を明らかにし，それらの個々の問題に対して人間工学的な立場からの考察，および，それに基づく「ことだま」の日本語エディタにおける設計方針を紹介し，その評価を行う．また，特に文節に基づくわかち書き方式を仮名漢字変換の入力として実現する際に生ずる困難，それらの解決法についても紹介を行う．

† Kana to Kanji Conversion Text Input of KOTODAMA Document System by TETSUNOSUKE FUJISAKI, MASAAKI OKOCHI and MASAYUKI MOROHASHI (Tokyo Scientific Center, IBM Japan, Ltd.).

†† 日本アイ・ビー・エム(株)東京サイエンティフィック・センター

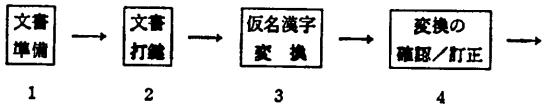


図 1 文書の入力過程
Fig. 1 Document creation steps by Kana Kanji conversion methods.

2. 仮名漢字変換を利用する入力方式の問題点

日本語の語彙には限りがなく，特に複合語が容易に形成される傾向があり，また，同音語も多く存在するので，現在の仮名漢字変換技術では，完全な変換を常に要求することはできない．たとえば，「アザブカラミタ，イイクラマデノ ヤケイハ」は「麻布から見た，飯倉までの夜景．」，「麻布から三田，飯倉までの夜景．」のいずれにも変換可能であり，たとえ意味の解析を行っても，常に正しい変換が行われることを期待することはできない．したがって，仮名漢字変換を基本とする文書入力システムを利用した文書の作成は，図1に示すような4つの過程を経ることとなる．

ところが，図1における4つの過程のうち最後の2つの過程は英文の入力方式，漢字の直接入力方式には不要の過程である．その意味で，仮名漢字変換を用いる文章の入力方式の設計に当っては，従来の英文入力方式（英文タイプライタ）や漢字の直接入力方式よりもさらに多くの人間工学的な配慮を必要とする．この配慮をより具体化すると次の4つの要請としてまとめることができる．

要請1：鍵盤からの入力に習熟を要せず，かつ高速度の打鍵が可能であること．また，操作者の仮名入力の熟練度に応じて，適切な入力方式が選べること．

要請2：変換の正答率が十分に高く，4の過程がなるべく変換結果の確認だけで済むこと．

要請3：仮に変換誤りが生じ，4の過程で変換誤りの訂正の必要が生じても，その訂正が簡単な操作で迅速に行えること．変換誤りに対して，再度の仮名入力を必要とするのでは困る．

要請4：システムが学習の機能を持ち，操作者もしくは文書に依存する語彙を長期的に習得し，短期的には同じような変換誤りを繰り返さないこと．

これらの要請を満たす設計により初めて仮名漢字変換を利用する文書の入力方式は実用的なものとなる．

3.「ことだま」の仮名漢字変換

3.1 鍵盤入力への配慮

文節に基づくわかち書き

文節を基本とするわかち書き方式は，字単位，言葉単位のわかち書き方式，字種を指定する方式などに比べて，打鍵数が少なく，入力のリズムを保ちやすく，タッチ・タイピングによる高速度の打鍵を可能にする利点があるので，これを基本方式とした．

ただ，文法とは形式的なものであり，あいまいさを排除するのが目的であるので，文法に忠実に文節を基本とするわかち書き方式を実現すると，わかち書きに自由度がなくなり，また操作員の感覚にそぐわないわかち書きを強要する場合も生ずる．これではかえって操作員への負担となるので，わかち書きに自由度を持たせるとともに，操作者の感覚にそぐわない点を直す工夫をする必要がある．このための工夫として，意味的に補助的な役割を果たす言葉の取扱いを工夫している．たとえば

ニホンゴノ ショリヲ スルト イウ コトハ は文節の従来の定義に基づく入力であるが，「スル（動詞）」，「イウ（動詞）」，「コト（形式名詞）」，などの言葉は，「日本語」，「処理」などが，文書の内容に関連した意味を持つのに反し，文書を表現する上で補助的に用いられているに過ぎない．そこでこれらの補助的な役割を果たすための言葉は，文法の規定するそれぞれの品詞として扱うと同時に，一種の付属語としても扱われるように付属語を拡大している．その結果，それらの言葉は，自立語（名詞や動詞など）としても，付属語としても扱われ，文節に基づくわかち書きにおいては次のいずれかのように文節の中に埋め込まれてわかち書きされても構わない．

ニホンゴノ ショリヲスルトイウコトハ，
ニホンゴノ ショリヲ スルトイウコトハ，
ニホンゴノ ショリヲ スルトイウ コトハ，
ニホンゴノ ショリヲ スルト イウ コトハ，

この配慮により，わかち書きの目安が文章上の意味を持つ主要な言葉の位置と一致して分かりやすいと同時に，多少のわかち書きのゆれに対しても対応ができるようになる．さらに，わかち書きの誤りによる変換率の低下が避けられることになる．

多様なわかち書き方式

また仮名書き方式において，カタカナによる表記，ローマ字による表記のいずれをも許しており，さらに

ブンショニュウリョクハ ジュウヨウナ カダイデアル
(a) わかち書き入力
「ブンショニュウリョク」ハ「ジュウヨウ」ナ「カダイ」デアル
(b) 字種指定入力
ブンショニュウリョクハ「ジュウヨウ」ナ カダイデアル
(c) わかち書きと字種指定の混在した入力
BUNSHONYUURYOKUHA JUUYOUNA KADAIDEARU
(d) ローマ字表記による入力

図 2 種々の入力形式

Fig. 2 Variety of input.

適宜それらを使い分けることも可能としている．仮名書き入力文のわかち書き規則においても，上記の文節に基づくわかち書き方式，字種指定記号を用いるわかち書き方式の両者を許しており，さらには，同一行内でそれらを混在させることも許している．**図2**に仮名書き入力のさまざまな例を示す．

文節に基づくわかち書き方式に字種指定方式のわかち書きを混ぜることは意味がないようにも思えるが，

「玉葱は 玉ネギ タマネギとも書かれる.」のように初めから字種が明らかな意味を持つ文章の入力の便宜を計るものである．これにより，変換後に表記を変える二重手間が避けられる．

このような，入力方式の自由度は，従来日本語においては用いられることのなかった口述録音機を利用することを可能とした．通常の片仮名タイプライタで口述録音機を用いてタイプをしても，出力が仮名文であったため利用価値がなかったが，仮名漢字変換を用いる文章の入力方式が今後普及するに伴い，口述録音機を利用する文書の作成もふえると考えられる．その際は，字種の指定*を入力文に含む方式は適切ではなく，このような自由度の高い文節に基づくわかち書き方式が有効である．

3.2 変換率を高めるための配慮

同音語選択

日本語には同音語が多く存在するので仮名漢字変換の過程における読みからの辞書引き当てにおいて，複数個の候補が得られることがある．これらの候補から一つを選びだす技術が問題となる．このような場合，まず文脈から規定される品詞に基づき同音語の選択を行う．すなわち，後続の付属語列から規定される品詞制限がある場合，同音語の中でもそれに合致するものだけが引き当ての対象となる．字種指定によるわかち書き入力を行っている場合も，漢字指定部分に後続するひらがな指定部分との品詞接続条件を考慮する[7]．

* 字種指定の記号は，「カタカナ」，「ひらがな」，「英小文字」，「英大文字」，「辞書引き」に対して用意されている．

文節の仮名入力	付属語列からの品詞条件	
シンチョウ	———	身長，慎重，新調，……
シンチョウシタ	サ変動詞	新調した，伸長した，……
シンチョウデアル	形容動詞	慎重である，

図 3 品詞による同音語の選択

Fig. 3 Homonym selection from local context.

図3に「シンチョウ」の同音語の選択が後続する付属語列との関係により正しく行われる例を示す．この構文的な繋りによっても候補が一つに絞れない場合（図3の上2例）には，さらに同音語間の相対使用頻度による選択を行う．

このような同音語の選択は現在発表されている実用システムの多くで行われているが，文節に基づくわかち書き方式を採用した場合には問題がある．すなわち，入力が文節の単位でわかち書きされる場合，文節中の自立語部分の後端が明示されないので，同音語によるあいまいさに加えて，自立語，付属語部分の境界のあいまいさも加わる．たとえば，「クロウトハ イツモ」における「クロウトハ」は「クロウ（苦労）トハ」，「クロウト（玄人）ハ」のいずれにも分解可能である．このような場合に，自立語の長いものを優先して採用し（自立語の最長一致の優先），長さの等しい自立語間でのみ，言葉の使用頻度に基づく同音語の選択を行うのが普通であるが，それでは，いつでも「玄人は」が選ばれ，「苦労とは」が最初の候補として選ばれることはない．**図4**に同様な問題を引き起こす同音語の例を示す．

(派，歯)は⟷母　　返る⟷蛙
取り　⟷鳥　　聞き⟷機器，危機
火とは⟷人は⟷一葉　　気休，期待⟷来たい

図 4 文節わかち書きでの同音語

Fig. 4 Word groups having homonym relationship.

また，使用に伴う言葉の学習にも限界がある．すなわち，言葉の使用頻度の学習は辞書内の頻度情報を更新することで行うので，この方式を取った際には，言語の学習は，自立語最長一致の優先には影響を与えることはできない．したがって，短い言葉に対する学習の効果は長い言葉に比べて少ないこととなる．

これを解決するために，「クロウト（玄人）」と「クロウ（苦労）」，「メンカ（綿花）」と「メン（面）」のように，読みの長さの異なる言葉でも，読みの先頭からの部分文字列（それぞれクロウとメン）が同一で，残りの部分の読み（それぞれトとカ）が正しい付属語列

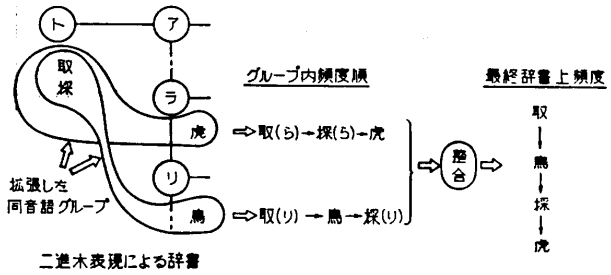


図 5 拡張された同音語間頻度と辞書上頻度
Fig. 5 Word usage frequency on the dictionary and the homonym groups.

の先頭になるもの，すなわち，上記の問題を引き起こす可能性のある言葉をグループ化し，(**図5**) 各グループ内の言葉間に相対頻度を与えている．辞書の各言葉の頻度は，このグループ内の相対頻度に基づき与えている．ただ，グループ間の相対頻度は互いに矛盾する場合もありえるので，辞書上の頻度に展開する際にはその面での考慮も必要である．このように，従来の同音語を拡張し，(上記のグループ) そこでの言葉の使用頻度を考慮しているので，同音語の選択でも，自立語の最長一致を優先せず，長さの異なる言葉も含めて，同音語の選択を言葉の使用頻度だけで行うことが可能となった(**図6**)．

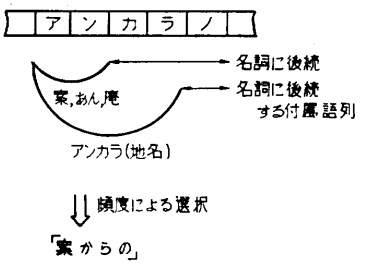


図 6 拡張した同音語内選択
Fig. 6 Homonym selection.

入力時のわかち書きの負担を軽減するために付属語を拡大解釈することを前に述べたが，それにも，この同音語の拡張した扱いは重要な役割を果たしている．すなわち，付属語扱いの言葉を次々と導入すると，さまざまなあいまいさが生ずるが，この同音語の拡張した扱いがそれらのあいまいさを解決する．たとえば敬称としての「さん」を付属語として扱うことにより，「タロウサンカラ (太郎さんから) ハナコサンヘノ (花子さんへの)」などが扱えるようになるが，「フジサンヘ ノボッタ，ケイサンヲ ハジメタ」などに対して，「藤さんへ登った」，「圭さんを始めた」のように変換される可能性もある．「藤」と「富士山」，「圭」と「計算」を同音語として扱い，頻度の比較をすることにより初めて，付属語を拡大することの危険を避けることができる．

未登録語処理

複合語，造語などの存在により日本語の語彙に限りはない．複合語処理では，辞書の探索が不成功に終わった言葉に対して，辞書中の，より短い言葉の繋ぎ合わせとしてその言葉を引き当てることを試みる．「疾病＋予防＋対策」，「健康＋管理＋室」などは，その例である．ただ単純に，入力列を複数個の部分文字列に分解し，それぞれに対して辞書引きを行い，辞書引きが成功したものを繋ぎ合わせるのが従来の方式であった．しかしこの方式では，むだな辞書引きが数多く行われるため，変換の速度を遅くし，かつ，変換速度への配慮から辞書の探索範囲を限らざるをえなかったので，変換の質も悪かった．

本システムでは，複合語が2文字の漢字基本単語と1文字の接辞の漢字の組み合わせであることが多いこと．また，複合語を構成する基本語は多くの場合漢語であり，したがって漢字が音で読まれること，の2つを利用した，より効果的な複合語処理を行っている．具体的には，入力仮名列から，漢字の音読み表を用いて，漢字文字の切れ目を推定し，その後に，複合語の構成パターンを利用したゴール指向の辞書探索を行っている．探索は，複合語のパターンとして確率の高いものから順次行われる[8]．したがって，長い複合語も，効率よく分解，変換される．たとえば，「コクレンアンゼンホショウリジカイ」は漢字の音読み表により，「コク・レン・アン・ゼン・ホ・ショウ・リ・ジ・カイ」のように漢字単位に分解され，2文字単語を尊重して「コクレン」，「アンゼン」，「ホショウ」，「リジ」への辞書探索と，「カイ」に対する接辞用漢字の範囲での辞書探索が行われる．したがって，長い複合語にもかかわらず，この例の場合には5回の辞書探索が行われるだけである．

数字列を含む「第1回，第2回，第3回，……」などの数詞句表現も，それらを別々の言葉とみなせば無限となる．数詞句処理は，パターン・マッチングを利用して，変換を行う．たとえば，「6ガツ30ニチヨリ」に対しては，辞書探索に先立ち数字が特殊記号で置き換えられる (&ガツ&ニチヨリ)．この置き換えにより，辞書中の数詞パターン (&ガツ&ニチ，&月

&日）が言葉として引き当てられ，数字の置き換えにより（6月30日より）のように変換される．このように数詞句は通常の言葉と同一に扱われるので，処理は単純であり，かつパターンとして，まとまった単位で扱われるため変換の精度もよい．また，通常の言葉と同様に，使用に伴う頻度の自動更新，新しい数詞句パターンの自動追加なども行われる．

変換不能語の取扱い

以上の過程を経ても変換されない場合は，付属語列の解析の結果を利用して，自立語として考えられる部分をカタカナで，付属語として考えられる部分をひらがなで表示する．たとえば，「アインシュタインニヨル ソウタイセイリロンノ……」にたいして「アインシュタイン」が辞書に未登録でも，「アインシュタインによる」のように表示されることとなる．

これは，日本語においては外来語が多く，表記に特にゆれが大きいため，辞書にそれらすべてが登録されているのが期待できないために行われている工夫である．これにより，「イタリー，イタリア，イタリヤ，エディター，エディタ」などが仮に辞書になくとも，変換は正しく行われる．

3.3 変換誤りの訂正を容易とするための配慮について

仮名漢字変換における変換誤りを完全になくすことはできないので，変換誤りが仮に生じたときに，それらを意図されたものに直す手続きの容易さ，簡便さが重要である．その操作が複雑であったり，変換誤り部分に対する，再度の仮名文字の入力を要したり，変換誤り部分が文字単位でしか直せないのであっては，変換率がいくら良くても，訂正作業が苦痛となり，仮名漢字変換方式での文書入力は実用的でなくなってしまう．

この点を考慮して，変換誤りの訂正を容易に，完全に行わせることに多くの配慮を行っている．具体的には変換誤りの位置にカーソルを合わせプログラム機能キーを押すことが，変換誤りの訂正の基本動作である．この操作によりカーソルが置かれた位置の言葉単位に訂正が行われる．このため，訂正の操作を行ったときのカーソル位置から，スクリーン上のその位置に表示されている文節の情報がすべて得られるようなデータ構造を用いている．得られる情報は，入力時のわかち書きの単位，仮名文字列，付属語列解析結果，自立語の候補などである（図7）．このようなデータ構造の実現により，同音語の選択誤りに加えて，入力時の打ち誤り，わかち書きの誤り，表記上の誤り，自立語付属語分離時の誤りなどの起こり得るすべての変換誤りを以下に示す3種の基本操作で修正することが可能となる．

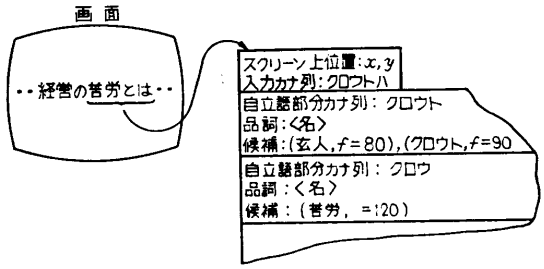


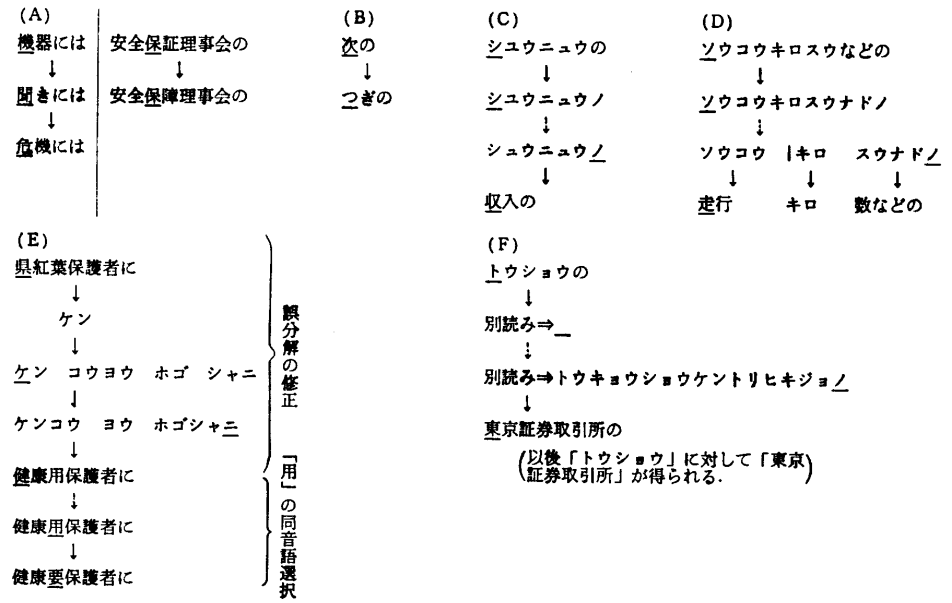
図7 文節情報のデータ構造
Fig. 7 Data structure for a phrase.

同音語内選択：従来のシステムに採用されているように，同音語を逐次もしくは一括表示し，訂正を行うための機能である（図8のA）．特に「玄人は，苦労とは」のような拡張された意味での同音語も，ここで同音語として扱われる．

表記変更：「次の，つぎの」，「さらに，更に」，「もちろん，勿論」などをいずれの表記で書くかは，文書の作成者の好みともいえる．文節に基づくわかち書き方式では，各々の言葉の字種に対する指定を行わないので，これらの問題が深刻となる．

表記変更の訂正機能は，このような異なった表記間の変更を1操作で行うための機能である（図8のB）．これも，前述の文節ごとのデータ構造の保持により可能となっている．

入力仮名文字列の再現と編集：誤り位置にカーソルを置き，このプログラム機能キーを押すことにより，その誤りを含む文節の入力仮名文字列が前述のデータ構造から得られて，変更可能な状態で表示される．それに対して，直接仮名文字列の変更を行い，再変換を行わせることができる．これは，入力における打ち誤り，わかち書きの誤りなどの訂正を容易にしている（図8のC）．また，言葉が未登録で変換されず，外来語と推定されカタカナ表示された場合に，入力仮名文字列を表示し，より短い単位にそこでわかち書きをすることもできる（図8のD）．字種の誤りに対しては，字種を指定する記号をその場面で挿入することも可能である（図8のD）．複合語分解においても，計算機により誤った分解が行われる場合があるが，この機能により，計算機の行った分解を容易に訂正することができる（図8のE）．この訂正機能の一種として，あ



(注) ・↓, ↓はそれぞれ計算機, 操作者の行うことを示している.
・(D)の「|」はカタカナ指定用制御文字.
・(E)は2つの操作を組み合わせた例 (入力仮名文字の再現と編集を行い, その後同音語選択を行っている.)
・(F)は(E)とは別の機能キーを用いている.

図 8 種々の訂正操作例
Fig. 8 Examples of corrections

て字の機能もある (図8のF). このあて字の機能を使えば,「IBM−−>日本・アイ・ビー・エム株式会社」,「トウショウ−−−>東京証券取引所」のような, 略語を登録することもできる.

3.4 言葉の学習に対する配慮

変換に用いられる言葉はすべて辞書により管理されるが, 使用者ごと, 文書の種類ごとの使用語彙のゆれ, 表記のゆれ, さらには, 用語の選択の局所性に対処するため, 3段階の辞書構成を行っている.

主記憶辞書は, そのセッション内で行われた変換誤りの訂正操作により得られる情報を蓄積する主記憶内の辞書で, 同じ変換誤りを再び繰り返さないために存在する. そのため, 変換誤りの訂正操作により新規に作られた言葉, 頻度が更新されるべき言葉などに対して, 読み, 正書, 品詞などの情報が, 訂正が行われた文節のデータ構造から抽出され, ここに蓄積される. ユーザ辞書は二次記憶装置上の辞書で, その使用者, その分野の文書に特異な語彙および同音語間頻度を与えている. セッションの開始時に使用者はこのユーザ辞書を選択することができる. 基本辞書はやはり二次記憶装置上の辞書で, すべての使用者に共通の語彙および同音語間頻度を格納している. これらの辞書のレコードは, 読み, 正書, 品詞, 使用頻度などを含む.

ユーザ辞書と基本辞書はセッション中には参照のみが行われるが, セッション終了時には, その間に蓄積された主記憶辞書内の情報がユーザ辞書に吐き出される. ユーザ辞書にすでに存在する言葉においては頻度の更新が, ユーザ辞書に存在しない言葉に対しては, 新規の挿入が行われる. したがって, ユーザ辞書には使用に伴い, 使用環境を反映した語彙が蓄積されるとともに, 同音語間の使用頻度も効果的に反映される.

特に新しい言葉が辞書に登録される際に, 文節全体を言葉として登録するのでなく, 文節中の自立語部分だけを切り出し, さらにその品詞を推定しなければ効果的な学習を行うことはできない. たとえば,「ゲキシャシタノデ……」の入力に対しては「激写 (ゲキシャ)」を名詞もしくはサ変動詞として登録することにより初めて, 以後の「ゲキシャトハ, ゲキシャシナカッタノガ」などの用法にも適用可能となる. 文節に基づくわかち書きを採用した場合には, 自立語の後端が明示されないので, これは一般に容易ではない. そこで, 入力時に与えられた読み「ゲキシャシタノデ」と最終的な表現「激写したので」の比較により,「したので」が付属語部分であるとの推定を行い, それから,「激写 (ゲキシャ)」が名詞かサ変動詞のいずれかであることを予想している. したがって, 何らかの訂

正が行われた文節から自立語が適切に切り出され，推定された品詞情報とともに辞書に登録されるので学習の自動化が可能となっており，学習の効果も大きい．

特に，変更の操作により入力時の1文節が分解されても（図8のE）分割の過程がデータ構造上に木構造で保存されるので，それを利用して，入力時に与えた分かち書きの単位での学習も行われる．図8のEの例では，「健康要保護者」が新語（名詞）として登録されるとともに，「ヨウ（要）」に対する頻度の更新も行われる．

4.「ことだま」の日本語エディタ

4.1 運用環境

前章までに述べた配慮に基づき実現された「ことだま」の日本語エディタは，IBM システム/370，VM/CMS，IBM 3278 漢字ディスプレイの組み合わせの下で1980年9月以来稼働している．それは，中／大型計算機の時分割環境のもとで実現されて，文書の入力・編集の処理を行い，それ以外の文書処理サービス・プログラムとファイルを介して結び付けられている．したがって，このエディタにより作成された文書の校正出力を行うこと，文書の質を高めるための各種の文脈索引の出力を行うことなどが可能となっている[9]．また，作成され，蓄積された文書が構築するデータベースの管理検索用の各種の索引作成，データベース内の文書よりの自動キーワード抽出およびそれらを利用したオンラインの検索なども実現されている．

4.2 辞　書

辞書はユーザ辞書，共通辞書ともにB—木（IBM/VSAM）[10]上に実現されている．現在，基本辞書には，名詞，動詞，形容詞などの一般単語類が約6万語，姓，名，住所，企業名などの固有名詞類が約4万語含まれており，その他の漢字類，数詞のパターンなどを含めて約11万語が収納されている．

特に，文節わかち書きを意識しているので表記のゆれに対応するための配慮も行っている．たとえば，「組立，組み立て，組立て」などが異なる辞書上の言葉として頻度とともに収納されている．

4.3 評　価*

評価1：わかち書き（表1）

仮名漢字変換方式の文書入力を行うに当って，手書の文書をみながら打鍵者が頭の中でわかち書きをすることの効率への影響が問題である．この作業が負担となるなら，タッチ・タイピングの利点が損なわれ打鍵の効率が下がることになる．表は英文タイプの経験者に3日間程度のカタカナタイプの練習を行わせ，その後に上記のデータを計測したものである．さらに，口述録音機を用いた場合も計測を行った．この結果，カタカナ打鍵の速度に対して，文節わかち書きをしながらの打鍵速度がそれほど低下していないことが示されている．特に，口述録音機の利用（文節わかち書きをして通常の速度で録音してある．）ではその利用の有効性が示されている．カタカナ鍵盤はJIS配列を利用しているが，打鍵速度は期待したものに至っていない．これは特に対象が科学技術論文であったため文中に特殊記号，数字，英字が多く，鍵盤上でのシフト

* これらの測定は，操作者，対象文書，分野によって変わり得る．

表1 わかち書き入力の速度評価

Table 1 Typing speed with segmenting.

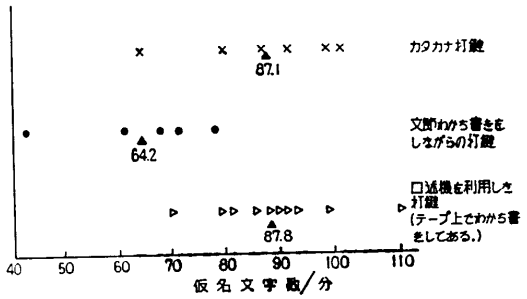


表2 初期変換率

Table 2 Initial hit ratio.

	文献数	総漢字数	変換率% (a)	変換率% (b)	(a)—(b) %
新聞社説*	3	3,539	86.7	82.1	4.6
社内規則**	3	3,197	88.8	85.9	2.9
情報処理論文***	2	3,569	82.0	79.4	2.6
IBM 社内雑誌****	2	3,477	93.3	90.6	2.7
総平均	10	13,782	87.7	84.5	3.2

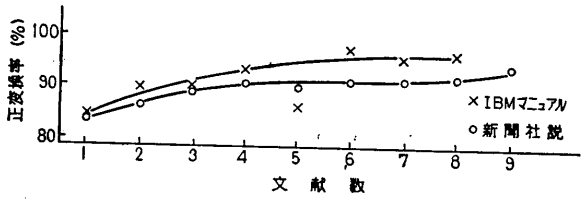
* ・日本経済新聞1981年2月17日社説「代替エネ法案と民間活力」全文.
・日本新済新聞1981年2月18日社説「最上昇に転じた米金利」全文.
・読売新聞1981年2月18日社説「園田特使は中東外交に新風を」全文.

** ・IBM 社内規則，費用精算について，一部.
・IBM 福利と厚生，厚生年金住宅融資制度，全文.
・IBM 福利と厚生，海外留学制度，全文.

*** ・情報処理学会誌 Vol. 20, No. 11, 1979,「計算機複合体のオペレーティングシステム」pp. 1001-1002, 1，2章のみ.
・日経エレクトロニクス 1979, 12-10 (No. 227),「IBM 4331/4341プロセッサの高密度実装」pp. 191-192,「システム全体の実装」の節.

**** ・IBM マニュアル N: GH 20-0765, IBM システム/370 IMS/VS 概説書，1章の1部.
・IBM アクセス No. 4, 1979, IMS-ADF〔適用業務開発機能〕について，全文.

表 3 学習による変換率の推移
Table 3 Learning Curve.



(4段) が繁雑であったためと思われる.

評価2：初期変換率 (表2)

4つの分野から計10種の文書 (1300漢字文字平均, 5.3漢字文字／1文節) に対する, 文節単位の変換率を測定した. (1文字でも誤っていればその文節を誤りとして数える.) ただし, このエディタの使用の初期状態を測定するため, それぞれの文書は空のユーザ辞書で変換を行った. 特に文節わかち書きであるため, 表記のゆれが問題であることが示されている. (a)は, (特に↔とくに, 従って↔したがって) のような表記の多少のゆれを許容したときの変換率であり, (b)はそれらを誤りとした場合の変換率である.

評価3：学習の効果 (表3)

学習の効果を測定するため, 2種のグループの文書 (新聞社説〔a〕, IBM マニュアル〔b〕に対して, 変換率の変化を測定した. それぞれ2つのグループに対する測定は空のユーザ辞書から始めている.

2つの測定とも, 言葉の習得と, 使用頻度が更新されることにより, 急速に変換率が向上していることが示されている. 2つの測定では, 新聞の社説の場合の収束が悪い. これは, 新聞の社説が科学, 政治, 金融などの多岐にわたる一方, マニュアル類は, 計算機に内容が限定され, 語彙が少ないことを反映していると考えられる.

5. おわりに

近年の計算機技術の発展により, オフィスオートメーションの実現も近い将来に期待されているが, そこでの文書の入力の主要な方式として仮名漢字変換が重要な位置を占めている.

本研究では, 特に, 仮名漢字変換の利点を生かすため, 打鍵者の負担の少ない文節に基づくわかち書き方式を採用し, そこで生じる問題を明らかにした. また, それらの問題の解決を示した. その結果, 文節に基づくわかち書き方式の仮名漢字変換が実用の域に至ったと考える.

なお, 本研究は, 文書の入力だけでなく, 他の文書処理サービス (校正出力, 索引作成, キーワード抽出, 検索など) を一貫システムとして実現する日本語文書処理システム「ことだま」の一部として行った. したがって, 本稿で紹介された仮名漢字変換にも, それらとの結び付きを意識した配慮が含まれている. しかし, 紙面の制限よりその部分についてはすべて割愛した.

最後に, 「ことだま」エディタの設計・開発・評価において, 多くの方々からの助言, 協力をいただいたことを感謝する. 特に, 実働化, 辞書の整備において, 大深悦子氏, 戸沢義夫氏, 間下浩之氏 (以上IBM) の協力に感謝する.

また, 辞書の作成にあたっては, 三省堂のご厚意により, 新明解国語辞典を利用させていただいたことを感謝する.

参考文献

1) 森, 河田：かな漢字変換, 情報処理, Vol. 20, No. 10 (1979).
2) 山田：日本語テキスト入力法の人間工学的比較, 日本語情報処理シンポジウム報告集, 情報処理学会 (1978).
3) Office Automation in Japan, JIPDEC Report, Winter, pp. 1-19 (1980).
4) 牧野, 木澤：べた書き文の分かち書きと仮名漢字変換, 情報処理学会論文誌, Vol. 20, No. 4 (1979).
5) 牧野, 勝部, 木澤：カナ漢字変換の一方法, 情報処理, Vol. 18, No. 7 (1977).
6) 藤崎, 大河内, 諸橋, 戸沢：日本語文書処理システム「ことだま」, IBM 東京サイエンティフィック・センター・レポート, N：G 318-1512.
7) 大河内, 藤崎, 諸橋：仮名漢字変換のための文法解析, 情報処理学会計算言語学研究会資料 CL 25-4 (1981).
8) 野村：複次結合語の構造, 国立国語研究報告 49, 電子計算機による国語研究V (1973).
9) 藤崎, 諸橋, 大河内：日本語文書処理システム「ことだま」, 情報処理学会第22回全国大会 (1981).
10) Planning for Enhanced VSAM under OS/VS, IBM Manual, GC 26-3842.

(昭和56年4月1日受付)
(昭和56年6月16日採録)