べた書きかな漢字変換の一方式について

平塚 良治　八田 敏　津田 貴
（沖電気工業株式会社）

1 はじめに

最近、計算機が対象とする業務が拡大し、それに伴って計算機が扱うデータも、従来の数値データに加えて、日本語文字データ，イメージデータ等へと拡大して来ている。

日本語の文章を扱う処理では、欧米の26文字と違って、数万種類もの文字を対象としなければならず、その入力方法を容易にすることが、非常に重要な課題である。

現状の日本語ワードプロセッサでは入力方法として、全文字配列型のタブレット方式やかな漢字変換方式，カナツータッチ方式などが提案されているが、初心者にも楽に使え、かつ適当なスピードで入力が可能であるという点から、この頃では、かな漢字変換方式が中心となって来ている。

かな漢字変換方式にはいくつかのレベルがあるが、それぞれのレベルの分かち書きの例を図－1に示す。

日本語ワードプロセッサでは文節分かち書きレベルのものが中心となっているが、利用者にとって最も負担が少ないのは、かな漢字変換用の分かち書きを一切行わず、文章のよみの通りに入力するレベルのものであろう。

このような入力方法は、一般にべた書きかな漢字変換入力と呼ばれている。

さて、べた書き入力のかな漢字変換における課題は次の2点である。

(1) 文節への分割
べた書き文に対して、どのような方法で文節に区切るか。

(2) 同音語の選択
複数個の同音語に対して、どのような方法で最適な同音語を選ぶか。

本文では、主として課題(1)の解決策について述べる。今まで課題(1)については、いくつかの方法が提案されているが[1][2]、それらの多くは、文節を構成する文字の長さを中心にした解析であり、文節の基本である助詞，助動詞を中心にしたものではなかった。

我々は以前に、文献[3]において、文節長による解析では正しく処理されない部分を、助詞の可能性をもつ文字を使って、訂正する考えを提案した。

今回はその考え方をより発展させ、べた書き文が与えられたとき、慣用的な付属語の連接語や助詞の可能性のある文字列を中心に解析し、その結果を利用して文節を決定する方法を提案する。

以下、第2章では概略処理の流れ、第3章では助詞「の」の解析方法とその結果について、第4章，第5章で本方式の特徴，問題点を述べる。

レベル	名　称	例　文
1	べた書き入力	ぎじゅつかいはつはちょうきてきなしやでおこなわれる。
2	文節分かち書き入力	ぎじゅつ゛かいはつは゛ちょうきてきな゛しやで゛おこなわれる。
3	漢字部指定入力	゛ぎじゅつ゛かいはつ゛は゛ちょうき゛゛てき゛な゛しや゛で゛おこ゛なわれる。

図－1　かな漢字変換の入力レベル（入力例）

2 文節への分割の方式

日本語の文章は文節単位に構成されており、実際に日本語ワードプロセッサなどで行われているかな漢字変換でも、文節単位に区切る方法が中心となっている。

べた書き文のかな漢字変換では、文字列に対して、句読点以外には区切り符号がないので、いかにして、べた書き文を文節単位に区切るかが重要な問題である。

そのための方法として、一文内の文節数を最小になるように文を区切る方法[2]や、連続する2文節を構成する文字の長さの最も長いものを選ぶ方法[1]等も有効であるが、変換対象となったべた書き文の中の助詞とか助動詞等の日本語の基本的な単語[4]を利用して、文節へ分割するのが一番自然な方法であろう。

一般的な文節の定義は、次のとうりである。

(1) 自立語

(2) 自立語 ＋ 付属語

ところで、実際の文節がどのような構造から成っているかを調べるため、文節の定義における自立語をさらに分析していき、活用のある語，活用のない語に着目して文節を構成する単語で整理すると、表－1のようになる。

表－1 文節末の単語

項番	文節末の単語	品詞	備考
1	変化する単語	動詞 形容詞 形容動詞 助動詞	未然形 連用形 終止形 連体形 仮定形 命令形
2	変化しない単語	名詞 連体詞 副詞 感動詞 接続詞 助詞	複合語もあり
3	記号，句読点、	－	「，」，『，』，。，、

このような分類について、実際の文節の頻度について調査したのが表－2である。

この例は、朝日新聞社の「天声人語」[5]から、約2000文節程度を取り出して、分析した結果である。

表－2 文節(末)の頻度

順位	文節末の内容	%
1	助詞	54.4
2	記号・句読点、	17.5
3	名詞(複合語)	11.4
4	連体形	9.7
5	連体詞	3.0
6	副詞	3.0
7	その他*1 (感動詞，連用形 終止形 など)	1.0

表－2からわかるように、文節末を構成する単語として、助詞が最も多く、複合語、連体形の活用語尾の順に多い。

このような単語をさきに述べた基本的な単語として、べた書き文を解析すれば、より正しく文節へ分割できると思われる。

そこで、文節への分割処理としては、助詞と連体形を中心に分割を行うこととし、その前後に慣用的な付属語による特殊文字列処理と副詞，連体詞，複合語等の活用しない語を辞書引きによって確認する処理を追加した。

＊1：終止形や連用形が少ないのはそれらの場合は、その直後に句読点があるためで、大部分は記号，句読点でカウントされている。

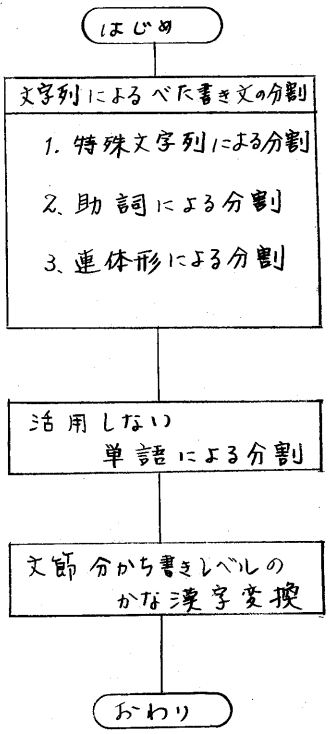
以上、図－2に、本方式の概略フローチャートを示し、以下にその内容を簡単に説明する。

2.1 特殊文字列による分割

べた書きの文字列の中に、ある特殊な文字列が見つけられ、その文字列が分野とか記述する人によらず、共通に現われる文字列であれば、それらについては、あらかじめ文節の区切りを明確にしておけば、その文字列をサーチするだけで、文節の区切りがわかる。

このような文字列としては、助詞などの文字列のように汎用性をもたなくても、慣用的に用いられる言葉がよく、同じような文字列があったら、まちがいなくそれと認識できるものがよい。

表－3に、それらの特殊文字列の一部を示す。



図－2 概略フローチャート

2.2 助詞による分割

第3章に詳細を示す。

2.3 活用語尾による分割

連体形や連用形，仮定形等すべての変化について考えられるが、表－2における頻度面から、連体形を例として説明する。

一般に、連体形を構成する活用語尾として、次の6種類が考えられる。

(1) ウ段の文字（5段活用）
(2) Xる（X：い段又はえ段の文字 い段なら上一段活用，え段なら下一段活用）
(3) くる，する（カ変，サ変活用）
(4) い（形容詞型活用）
(5) な（形容動詞型活用）
(6) た等（助動詞の活用）

助詞による解析をするとき、上記(1)～(6)と同様の文字が見つかったら、その直前の文字列を調べ、連体形であることを確認する処理である。首尾よく連体形ならば、そこは文節末である可能性が非常に高いことを示す。

2.4 活用しない語による分割

助詞や活用語尾による分割が終った後、文節の先頭文字から辞書引きを行ない、連体詞，副詞，複合語をみつける処理である。

2.5 文節レベルのかな漢字変換

2.1～2.4までの処理により、すべての文節の区切りが明確になったので図－1に示した文節レベルのかな漢字変換をそのまま利用できる。

表－3 主要特殊文字列

について	している	となって
にかんして	において	という

3 助詞による分割

べた書き文の中で、助詞の可能性をもつ文字について、それらが本当に助詞なのか、又は単語の一部なのかを決定する処理が、助詞による文節への分割処理である。

そこで、助詞の中でも頻度が最も高い「の」について、例として説明する。

3.1 「の」の可能性

べた書き文の中に「の」が現われた場合、それらの「の」は大旨次の6つのうちのいずれかである。

(1) 自立語の一部

この場合は、その単語は自立語であり、その単語の前で文節の区切りになっている。

(2) 格助詞の「の」

「の」が付属語として現われる場合で、「の」の直後が文節の区切りになる。

(3) 助詞の連接

「の」を含む複数文字で、助詞を構成したり、助詞同士の連接になったりする場合で、通常「の」の直前は連体形である。文節の区切りとして最後の助詞までとか、最後より一個前の助詞まで等が考えられ、一意には決らない。

(4) 助詞「のみ」の場合

(3)と同様であるが、「のみ」は名詞、格助詞「で」「に」等にも接続する。

(5) 終助詞

通常話し言葉の中に現われ、「の」「のよ」等が文末に起る。

(6) 死のう

ナ行五段動詞「死ぬ」の未然形でこの単語しかない。

実際的には、(5),(6)はほとんど現われず、(4)も数える程度しか現われない。従って一般には(1)~(3)の区別がつけばよい。

3.2 「の」の解析処理

「の」について解析しなければならない項目は、次の2つである。

(i) 「の」を含む単語であるかどうか。
(ii) 助詞連接の場合に文節の最終文字を認識すること。

これらを解析するために、次のアステップの処理を実行する。

(1) 「の」の直前,直後で字種が変わる。字種として、ひらがな,カタカナ,数字,英字等がある。この場合は「の」の直後で文節の区切りとする。

(2) 」,』の直前に「の」,「のよ」がある場合は、」,』を文節の区切りとする。

(3) 「の」を含む単語について

- 名詞なら、体言接続の助詞又は、助動詞「だ」「らしい」「です」のいずれかが、その単語の直後にあるかどうかチェックする。
- 連体詞なら、その単語の直後に次のいずれかがあるかどうかチェックする。
  - * 名詞又は副詞
  - * 動詞なら連用形,連体形
  - * 形容詞,形容動詞なら連用形又は連体形
- 人名の場合は、その直後に、さん,氏,選手等の人の名前を補助するものがあるかどうかチェックする。もしなければ、さらに名詞の場合と同様のチェックを行う。

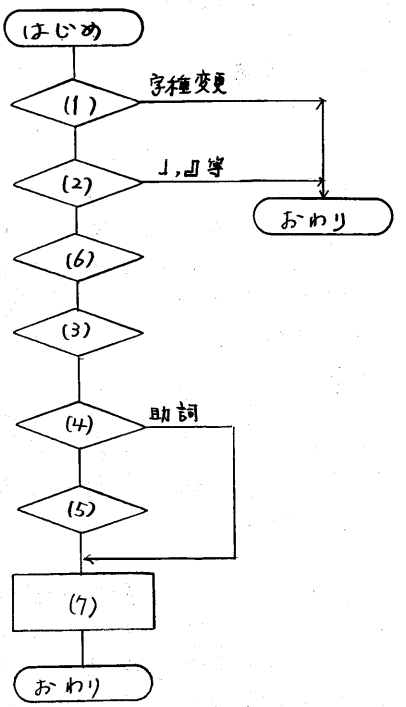
・地名の場合も、人名の場合と同様の処理を行う。
・用言の場合は活用形，活用語尾，接続する付属語等をチェックし、もし一致しなければ「の」を含む用言はないものと考える。

(4) 「の」を含む助詞又は「の」と他の助詞との連接の場合は「の」の直前の接続形が連体形かどうかをチェックする。さらに「の」に連接している助詞の可能性をもつ文字より後にある文字列の接続形態をチェックする。（チェック内容は（6）と同様）

(5) 「の」の直後に助動詞「だ」「らしい」「ようだ」「です」があったら、「の」は文節の区切りとはならず、「の」の後に現われる助詞によって文節の区切りが決定できるよう情報を設定しておく。

(6) 「の」の直後の接続について
「の」の直後の文字列が次のいずれであるかどうかチェックする。
・名詞又は副詞
・動詞なら連体形，連用形
・形容詞，形容動詞なら連体形，連用形

(7) (1)～(6)の処理の結果を表－4に従って処理し、文節の区切りを決定する。(1)～(7)の処理手順を図－3に示す。



図－3 「の」の解析処理

以上のアルゴリズムの評価として、天声人語から 約1000個程度の「の」をとりだして適用した結果、約98%正しく解析できることが判明した。

分類	Xで始まる単語の有無(6)	(3) 「の」	「……の」	「…のX…」X_1	「…のX…」Y_1
1	の X Y_2	「の」の直後	「の」の直後	「の」の直後	通常はY_1の直後と考えられるが、Y_1の処理後に決定する
2	の X X_2	「の」の直後	「の」の直後	X の長さによる	
3	の X —	なし	「の」の直後	「X_1」の直前	
4	「の」の直前が連体形	Yの直後で区切れる可能性あり			

(4)のX　(4)のY

X ：Xで始まる単語あり
X — ：Xで始まる単語なし

Y, Y_i：助詞の可能性のある文字
X, X_i：助詞の可能性のない文字

表－4 結果判定表(一部)

4. 本方式の特徴

本方式の特徴として、次の3点をあげることができる。

(1) 本方式では、べた書き文の解析において、日本語の最も基礎となる単語……助詞や活用語尾を中心に解析しており、その点、解析の根拠が明確である。そのため、解析の途中で接続が一致しないなどの矛盾が生じたり、辞書にない未定義語があっても、その文節を無視するだけでよく、他の文節には悪影響を与えない。

(2) 本方式では、日本語に特有な文字列を中心に解析し、それらについての知識情報を操作するという簡単な方式なので、注目すべき文字列の追加，変更が容易に行え、システムの発展によって得たノウハウも容易にシステムに反映できる。

(3) 本方式では、文節に区切る処理をプリエディット処理として独立させることが可能なので、現状保有している文節単位のかな漢字変換のプログラムをそのまま利用できる。

5. 今後の課題

今後の課題として、次の2つを早急に行う必要がある。

(1) 計算量についてモデルを作成し評価すること。

(2) 本方式では辞書の容量が既存の他のシステムより3倍程度多く必要と考えられるので、何らかの形の圧縮や新しいサーチ手法を考慮する必要がある。

最後に、このような研究の機会を与えて下さったのみならず、いろいろ有益なヒントを与えて下さった 当社 椎野部長に感謝します。

参考文献

(1) 牧野他：べた書き文の分かち書きと仮名漢字変換 ―二文節最長一致法による分かち書き―
情報処理学会論文誌，Vol 20，No 4，1979

(2) 吉村他：文節数最小法を用いたべた書き日本語文の形態素解析
情報処理学会論文誌，Vol 24，No 1，1983

(3) 坂本他：前処理を導入したカナ漢字変換
情報処理学会第26回全国大会講演論文集(I)，2H-2

(4) 林：新聞語彙調査の概略と語彙分析法試案
国立国語研究所報告 31，1968

(5) 朝日新聞：天声人語 '80 冬の号
原書房 1981