# かなべた文の単語分割アルゴリズムの一方式

沼田泰之、藤田克彦、林　大川、山内佐敏
（（株）リコー・技術本部）

## １　はじめに

日本語の入力作業において、最も大きな問題の一つが多数の漢字の存在であったことは言うまでもない。特に、専任のオペレータではない一般の人間にとっては、数千にもおよぶ文字の中から希望の文字を探して入力するということは、事実上不可能に近いことであった。

かな漢字変換技術はそのような背景のもとに産み出され、これにより、一般の人間にも日本語の入力が従来よりはるかに容易に行なえるようになったのである。

しかし、現在の日本語ワード・プロセッサで主流をなしている方式は、単語ないし文節といった文法上の単位、概念を意識しながら入力作業を進めなければならないという問題点を残している。何が理想的な日本語入力法であるか、の議論は別にしても、より人手の介入の少なくてすむシステムのほうが望ましいという点については異論はないであろう。すでに、そのような観点から、べた書きされたかな文字列に対するかな漢字変換方式の提案もいくつかなされている。[1~3]

そうした中で、我々は入力に追随して逐次変換結果の出力ができるべた文のかな漢字変換方式を開発した。この方式は、日本語文章中において漢字表記される単語の多くがいわゆる漢語であることに着目し、それら漢語の読みであるかな文字列の性質を利用した、高速かつ精度の良いかな漢字変換方式である。本稿では、特にその中の単語区切りの方式について述べる。

## ２　日本語文章の特徴

図１に、日本語の単語分類を示す。

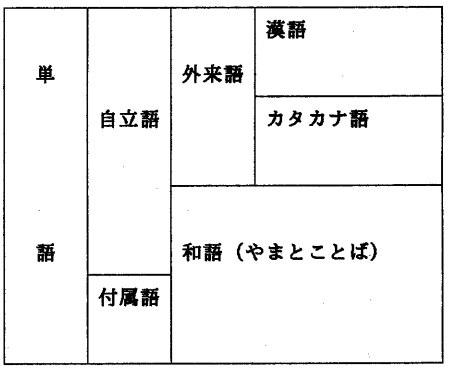


図１　日本語の単語の分類

まず、単語は自立語と付属語とに分類できる。ここで、付属語は一般にひらがな表記される。自立語は、さらに和語と外来語に分類できる。和語には、動詞、形容詞などが多いが、これらは必ずしも漢字表記される必要はない。一方、外来語には漢語とカタカナ語とがあるが、かな漢字変換は、かな表記されたものを、適切に漢字に変換するこ

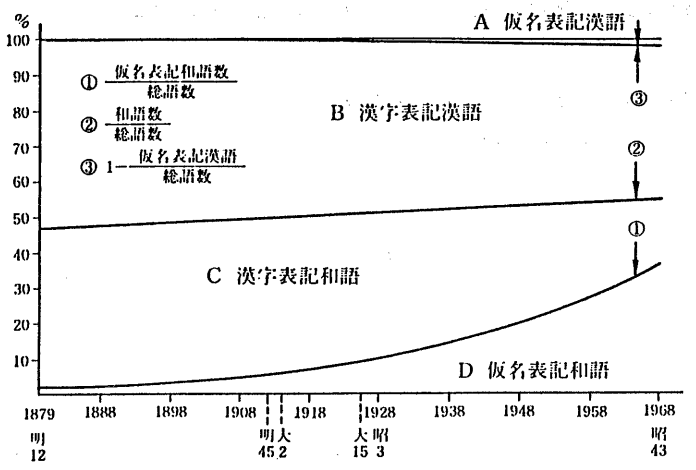


図２　語種別の語表記率の変遷[4]

とが目的であるから、カタカナ語はカタカナで入力することを原則とすれば、処理の対象としては漢語の方に注目すればよい。

次に、図２が示すのは、語種別にみた漢字表記率の変遷である。全体として漢字による表記の割合は低下している傾向は明らかだが、漢字で表記されている単語に占める漢語の比率は、高くなっていることがわかる。

筆者らの事務文に対する調査においても、漢字表記された単語の８割弱が漢語であるという結果が得られた。このことは、漢字で表記されるものの多くが漢語であることを示している。

以上の事から、かな漢字変換の対象として、漢語が極めて重要であることがわかる。

## ３　漢語を特徴づける漢字音

かな漢字変換において重要な対象であるこの漢語を特徴づけるものは、何か。これは、明らかに漢字音（音訓でいう音）である。

この漢字音には、かなで表記すると１字のもの、２字のもの、３字のものがある。

新明解国語辞典[5]には、造語成分の読みとして漢字音が２７２種が記載されているが、そのうち２字以上のものが２２４種で、１字のものが４８種ある。さらに２字以上のものについて、促音化したり半濁音化したものを考慮すると、現実には上記の他に１１３種が漢字音として存在しうることになる。

つまり２字、３字の漢字音が３３７種、１字の漢字音が４８種ということである。

漢語の読みは、これら漢字音によって構成されているわけである。

ここで、以下での説明の便宜上、改めて漢字音を定義しておくことにする。

漢字音－－上記漢字音のうち、かなで２字以上である３３７種。

そして、それ以外のもの、すなわち上記３３７種の漢字音の構成要素になっていないものは、一般のかなとみなすことにする。このように定義すると

例　開始（かいし）

においては、「かい」が漢字音、「し」がかなということになる。

表１に、上の定義による漢字音の例を示す。

| | | | | |
|---|---|---|---|---|
| じゅう | じゅく | じゅつ | じゅっ | じゅん |
| じょう | じょく | じょっ | | |
| すい | すう | すん | | |
| ずい | ずう | ずん | | |
| せい | せき | せつ | せっ | せん |
| ぜい | ぜき | ぜつ | ぜっ | ぜん |
| そう | そく | そつ | そっ | そん |

表１　漢字音の例（上記定義による）

## 4　漢字音による単語の分類

以下の議論のため、漢字音とかなとの組合せによる単語の分類を行なう。

以下に示したのが、単語の形態的分類である。

ＰＡＴＴＥＲＮ１　漢字音＋漢字音

例　開校（カイ＋コウ）、設定（セッ＋テイ）

ＰＡＴＴＥＲＮ２　漢字音＋かな

例　要素（ヨウ＋そ）、新規（シン＋き）

ＰＡＴＴＥＲＮ３　かな＋漢字音

例　期間（き＋カン）、事実（じ＋ジツ）

ＰＡＴＴＥＲＮ４　かな＋かな（＋……）

例　緑（み＋ど＋り）、川（か＋わ）

ＰＡＴＴＥＲＮ５　かな

例　胃（い）、絵（え）

ＰＡＴＴＥＲＮ６　漢字音

例　法（ホウ）、感（カン）

ＰＡＴＴＥＲＮ７　その他

例　私（わ＋タク＋し）

考え（カン＋が＋え）

注　カタカナは漢字音を示す。

この分類にしたがって、筆者らが作成した辞書中の単語の比率を算出したものが表１である。

| 分類 | 語数 | 比率（％） |
|---|---|---|
| ＰＡＴＴＥＲＮ１ | １２３２４ | ３２．１ |
| ＰＡＴＴＥＲＮ２ | ３３５７ | ８．８ |
| ＰＡＴＴＥＲＮ３ | ３３４０ | ８．７ |
| ＰＡＴＴＥＲＮ４ | １４３９６ | ３７．５ |
| ＰＡＴＴＥＲＮ５ | ６６８ | １．７ |
| ＰＡＴＴＥＲＮ６ | １３０５ | ３．４ |
| ＰＡＴＴＥＲＮ７ | ３１１５ | ８．１ |
| 合計 | ３８５０５ | １００．０ |

表１　単語の分類別比率

## 5　漢字音の利用

漢字音に関しては、次に示すような経験的事実がある。

１　漢字音は漢語の構成要素として、多く用いられる。

２　漢字音が２個連続して漢語を構成することが多い。

３　漢字音と直後のかなとで漢語を構成することが多い。

４　かなと直後の漢字音とで漢語を構成することが多い。

これらの漢字音の特徴を利用して、あらかじめ、辞書検索の前段階において、単語の切目を推測することを考える。

そのためには、入力される区切りの対象となるかな文字列に対しても、漢字音による特徴づけを行なう必要がある。

区切りの対象例文

じしょのけんさくにあたり

のような入力について、漢字音を単位として区切りを行なってみると次のようになる。

処理結果

じ／ショ／の／ケン／サク／に／あ／た／り

ここで、カタカナで表記されている部分が上述の漢字音である。これによれば、漢語相当部分が推定できると考えられる。

そこで、区切り対象となる入力文字列についても、その先頭からのかな文字の並びによって、次のように分類を行なっておくことにする。

［１］ＴＹＰＥ１　漢字音＋漢字音

例　ケン／シュツ

［2］ＴＹＰＥ２　漢字音＋かな

例　カン／じ

［3］ＴＹＰＥ３　かな＋漢字音

例　か／ノウ

［4］ＴＹＰＥ４　ＴＹＰＥ１，２，３以外

例　こ／の

注　カタカナは、漢字音を示す。

今後は、区切り対象文字列の属性を、上記ＴＹＰＥ１，ＴＹＰＥ２，ＴＹＰＥ３，ＴＹＰＥ４と略称する。

これら文字列は、単語との関連でいえば、それぞれ次のような特徴を有する。

ＴＹＰＥ１

ＰＡＴＴＥＲＮ１かＰＡＴＴＥＲＮ６の単語が存在している、つまり抽出される可能性が高い。

ＴＹＰＥ２

ＰＡＴＴＥＲＮ２かＰＡＴＴＥＲＮ６の単語が存在している可能性が高い。

ＴＹＰＥ３

ＰＡＴＴＥＲＮ３かＰＡＴＴＥＲＮ５の単語が存在している可能性が高い。

ＴＹＰＥ４

ＰＡＴＴＥＲＮ４かＰＡＴＴＥＲＮ５の単語が存在している可能性が高い。

このことを用いると、区切り対象文字列に対する辞書検索を限定することができる。

## ６　付属語と漢字音

表３は、自立語と、それに後続する付属語の接続の頻度を、事務文２２７例を対象として調査した結果である。

| ＮＯ | 自立語の品詞 | 付属語 | 出現頻度 | ％出現率 | 累積％出現率 |
|---|---|---|---|---|---|
| 1 | 名詞 | の | １２１７ | １６．７ | １６．７ |
| 2 | 名詞 | に | ８９１ | １２．２ | ２８．９ |
| 3 | 名詞 | を | ６２８ | ８．６ | ３７．５ |
| 4 | サ名 | の | ６２３ | ８．６ | ４６．１ |
| 5 | サ名 | を | ４２０ | ５．８ | ５１．９ |
| 6 | サ名 | に | ４０９ | ５．６ | ５７．５ |
| ： | ： | ： | ： | ： | ： |
| １４ | 名詞 | から | １０９ | １．５ | ８０．２ |
| １５ | サ名 | が | １０５ | １．４ | ８１．６ |
| ： | ： | ： | ： | ： | ： |
| ２３ | サ名 | する | ５５ | ０．７ | ９０．３ |
| ２４ | サ名 | から | ５２ | ０．７ | ９１．０ |
| ： | ： | ： | ： | ： | ： |
| １０７ | サ名 | ずつ | １ | ０．０ | １００ |

表３　自立語と付属語との接続頻度

自立語と付属語の組合せ１０７種中、上位１４位までで８０％を占め、２３位で９０％を越えている。

この結果より、次のことが明らかになる。

名詞の直後には「の」「に」などの助詞が現れやすい、また、サ変名詞の直後にはサ変動詞の、「さ」「し」が現れやすいことがわかる。

つまり、直前に抽出され、認識された単語の品詞が名詞で、かつ、それに続く文字列の先頭が「の」「に」などであれば、それは助詞である可能性が相当高く、同様にサ変名詞直後の「さ」「し」などはサ変動詞である可能性が高いといえる。加えて、経験的事実より、名詞の直後に用言が接続するケースは稀である。

しかしながら、複合語の存在を考慮すると、

「の」や「し」については、複合語の後半部を構成する名詞（あるいは接辞）の先頭文字である可能性と、付属語の可能性とを並列に検討する必要がある。

ここで、判断の材料として漢字音を利用することができる。

たとえば、直前に名詞が抽出された状態で、そこでの文字列の先頭文字が上記の付属語相当文字の類であったとする。漢字音による前処理を行なった結果、その文字列がＴＹＰＥ１あるいはＴＹＰＥ２であれば、その先頭文字は漢語の一部である可能性が高く、複合語のケースと考えられるようになる。

一方、ＴＹＰＥ４の場合は、少なくとも漢語の一部ではないといえる。また、ＴＹＰＥ３の場合は、可能性としては、付属語と漢語の先頭文字としての両方が考えられる。ここでは、複合語の出現より付属語の出現のほうが頻度的に大きいと考えて、付属語と考えることにした。

以下に、この考えかたに基づいた単語区切りアルゴリズムを示す。

## 7　単語区切りアルゴリズム

### 7．1　アルゴリズムの概要

図３は、単語区切り対象文字列から単語区切りを１つ見出すまでのアルゴリズムをブロック図として示したものである。

単語区切りのプロセスは、大別して次の２つとなる。

（ａ）一般単語の読みの抽出（被検索文字列の設定－辞書検索－最適候補選択）

（ｂ）特徴付属語の抽出

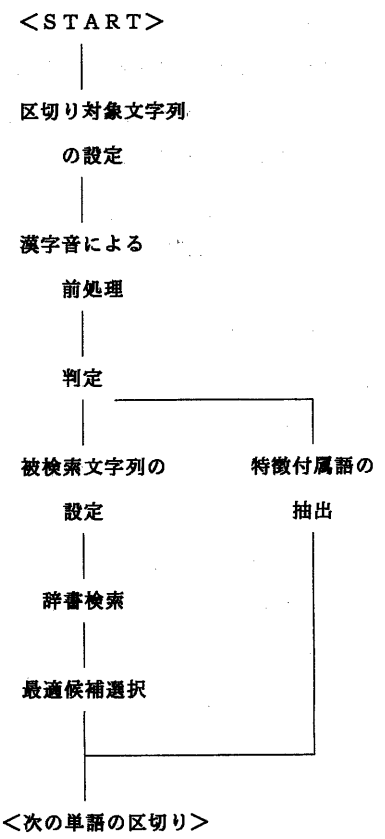


図３　アルゴリズム概要

ただし、（ｂ）から抽出される単語は、後述する一部の付属語に限られる。また、ここでは、通常の辞書検索は行なわれない。

図中の［判定］処理において、次の３項目がチェックされて、すべての条件を満たすときのみ、（ｂ）のプロセスが適用される。

チェック項目

１）漢字音による前処理の結果、そこでの区切り対象文字列の属性がＴＹＰＥ３，ＴＹＰＥ４である。

２）直前に抽出された単語の読みの品詞が名

詞がサ変名詞、形容動詞語幹のいずれかである。

3）区切り対象文字列の先頭文字が、設定した付属語（サ変動詞語尾、形容動詞語尾等を含む）と一致する。

以下に、個々の処理を説明する。

7．2　一般単語の読みの抽出

7．2．1　被検索文字列の設定と辞書検索

かなでべた書きされた区切り対象文字列に対し、漢字音による区切りを実行する。その先頭からの漢字音とかなの並び（ＴＹＰＥ）により、辞書検索の対象となる被検索文字列を限定する。

これにより辞書へのアクセス回数を減少させることができる。

7．2．2　最適候補の選択

上記の被検索文字列に対する辞書検索において、複数の単語の読みが得られる場合がある。解析を続行するためには、単語の読みの評価を行ない、一つに候補を絞らなければならない。

従来、評価法としては最長一致法（読みの長さを基準とする）、最尤評価法（読みの長さと使用頻度を組み合わせて評価する）などが知られている。

本方式では、品詞間の接続の重みを重視し、従来よりもやや細かい品詞分類を採用した。評価は、この品詞間の接続の重みと読み単位の評価値との和の形となっている。

なお、辞書検索で候補となる単語の読みが見出されない場合は、辞書検索の限定を解除して、検索を実行し、それでも候補が見出されない場合は、直前の解析に戻る、いわゆるバック・トラックを採用している。もちろん無制限なバック・トラックを防ぐために評価値の累積による制御を行なっている。

7．3　特徴付属語の抽出

漢字音をさらに有効に利用するために、上述のように一部の付属語の抽出において、区切り対象文字列のＴＹＰＥによる処理を行なっている。

8　単語区切りの具体例

以下に、具体例に即して処理の過程を説明する。

入力文字列の例

（1）さぎょうのうりつがいい

（2）さぎょうのうきがきまる

（3）さぎょうのこうかがひくい

（4）さぎょうのきぼがおおきい

いずれの文例においても、文頭から「さぎょう」という単語の読みが抽出されたとすると、残る文字列は次のようになる。

（1）のうりつがいい

（2）のうきがきまる

（3）のこうかがひくい

（4）のきぼがおおきい

「さぎょう」という単語の読みは名詞またはサ変名詞であり、その直後に位置する「の」は助詞の「の」である可能性が高いことになる。しかし、その反面、名詞連続、すなわち複合語を形成する名詞の一部である可能性もある。従来の漢字音を用いないアルゴリズムでは、辞書検索を実行するまでは、いずれの「の」であるか、まったく判断の方法がなかった。しかし、漢字音が漢語を構成する成分であるという特徴を利用しＴＹＰＥによ

る処理の切替を行なうことにより、辞書を検索する前にかなり妥当な判別ができる。その様子を次に示す。

（1）ノウ／リツ／が／い／い　　ＴＹＰＥ１
（2）ノウ／き／が／き／ま／る　ＴＹＰＥ２
（3）の／コウ／か／が／ひ／く／い
　　ＴＹＰＥ３
（4）の／き／ぼ／が／お／お／き／い
　　ＴＹＰＥ４

これらに対する妥当な区切りは次のようになる。

（1）のうりつ（名詞）／が
（2）のうき（名詞）／が
（3）の／こうか（名詞またはサ名）／が
（4）の／きぼ（名詞）／が

つまり、これらは漢字音による前処理の結果、漢字音の一部であれば、漢語の先頭文字であり、逆に漢字音の一部でなければ付属語であるという考えかたに即した例となっている。

この処理の特徴は、被検索文字列の設定、辞書検索、最適候補選択といった一連の処理を不要としながらも一部の頻出する付属語を的確に抽出することができる点にある。その結果、後続の文字列の処理へ即座にと移ることができる。

また、この方式は最長一致法などで問題となる「読みの長い単語が常に優先される」傾向への解決策の一つともなっている。

## 7　本方式の性能

市販の文例集[6]にある事務文２１文例を対象に、汎用機上で行なった本方式による単語区切りの実験結果では、単語数５１６７語に対し区切りを誤った単語数は２０２で、単語区切り率は９６．１％であった。

ここで、単語区切り率とは、同音異表記語への誤りは無視した、単語区切りの成功率のことであり、次に示す式で定義している。この数値は、同音語の学習が可能な単語の率でもある。

単語区切り率＝（正しく区切られた単語の数／総単語数）＊１００

次に、上記の漢字音処理を行なわない場合と行なった場合とでの効率の差を比較するために行なった辞書検索回数の調査によれば、単語の候補数は、漢字音処理を行なった場合、行なわない場合よりも約４０％削減できた。

## 9　おわりに

単語区切り率、辞書検索回数から考えて、本方式は十分な性能を有すると考えられる。

しかし、漢字音による処理により、かえって誤解析を生じる例もみられたので、今後はそれらに対して辞書で個別に対応するなどの方法により、さらに変換性能を向上させる方向へ向けて、試行するつもりである。

参考文献

1　牧野ほか：べた書き文の分かち書きと仮名漢字変換

情報処理，ｖｏｌ２０，Ｎｏ．４（１９７９年）

2　内田ほか：自由入力形式のカナ漢字変換

情報処理学会ＮＬ研資料２７－３（１９８１年）

3　吉村ほか：最長一致法と文節数最小法について

情報処理学会ＡＩ研資料２４－１（１９８２年）

4　森岡健二：「近代語の生成－明治期語彙編」

（明治書院、１９６９年）

ただし、本文中の図は、林大監修「図説日本語」（角川書店、

１９８２年）より再録した。

5　金田一ほか：「新明解国語辞典（第一版）」（三省堂、１９７２年）

6　安田賀計編：「企業経営文例全書」（ぎょうせい、１９７８年）