Vol. 22 No. 3 情報処理学会論文誌 May 1981

平仮名列の自動分かち書き†

田中康仁†† 古賀勝夫†††

この研究の目的は日本語の解析の最初に発生する分かち書きの問題を解決するための一手法である．分かち書きは一般に文字種（漢字，平仮名，片仮名，英字，数字，記号等）の変わり目で機械的に行われる．この時に発生する平仮名列に注目し，この性質を分析して分かち書きを行うものである．

分かち書きの方法は次のように行う．平仮名列の中には慣用的な表現が多くあるのでこれを集め辞書とし，文章中に発生する平仮名列を最長一致法により平仮名列辞書で右側より分割し，残った平仮名列はテーブルを参照することにより，語尾であるとか，その他の文法的接続検証等を行う．分かち書きされた平仮名列は対応する平仮名列辞書により，よりこまかく分割され，品詞情報を付ける．

このような方法により232万件の平仮名列を処理することにより約3万2千項目の平仮名列辞書があれば分かち書きできることがわかった．これだけの量を処理したことによって平仮名列辞書に新しく追加しなければならない件数は処理する平仮名列の 0.3%～0.4% 程度である．このような実験から平仮名列の中に繰り返し使われる平仮名列を使い分かち書きを行うことができる．分かち書き処理を行うごとに平仮名列辞書を充実させ人手に頼る部分を少なくすることができる．この実験は日本科学技術情報センターの抄録テープによって行った．

1. はじめに

日本語の研究を始めるにあたって，一番最初にぶつかる問題は，分かち書きである．ここでは，この問題の解決の一手法について述べる．

漢字列・カタカナ文字列の自動分かち書きは，用語の調査，インデクシング業務の機械化と共に進んできている．しかし平仮名列の分かち書きは，あまり研究が進んでいない．

自動分かち書きには，大別して二つの方法がある．

（1） ぼう大な辞書あるいはテーブルを用いて，処理する．

（2） 簡単な構文規則や語結合法則を用いて，プログラムによって処理する．

（1）の方法はプログラムが単純であり，わかりやすいが，処理速度に少し時間がかかる．また正確に分かち書きされる．

（2）の方法は，プログラムがむつかしく，例外処理に弱い．しかし処理速度は早い．正確さは前者に比べて劣っている．最近の分かち書きの研究を調べると（1）と（2）の統合化されたような方法が研究されている．小規模な辞書やテーブルを持ち，しかも構文解析を行い，正確にしかも処理速度を上げようとしているものがある．ここで述べる方法は（1）の方法を採用している．

さらにこの研究で用いている方法は文字種による分かち書きを基本にしている．漢字・英字・数字・カタカナ・平仮名・記号によって日本語を分割する．この方法は機械的な分割であるため，各種の誤りが発生する．

たとえば動詞・形容詞・形容動詞の語尾で発生する（起こる→起△こる）．

このほか，名詞・副詞等の一部仮名書きによる誤った分割もある（けい素→けい△素，ひ素→ひ△素）．さらに調べていくと，色々な誤りが発生する．

そこでここでは，平仮名列の特性を調べ，その特性を利用し，自動分かち書きに使うことを考える．

従来の考え方は規則によって平仮名列を分かち書きする方法が主流であったが，この研究では平仮名部分にある慣用的表現を大量に集め分かち書きに利用するものである．この考え方はすでに発表[16]しているが，資料を整備し，平仮名列辞書を充実させることができたので発表する．

この研究で取り扱う文章は，特許であるとか科学技術文献の抄録等に表現されている文章であって，一般の小説・会話文は対象にしていない．これらの文章の性質をまとめると，次のようになる．

（1） 現代文である．古文・候文等ではない．

（2） 簡潔な文章である．冗長な文章は少ない．

† Automatic Segmentation of Hiragana Strings Appearing in the Japanese Sentence by YASUHITO TANAKA (Nippon Univac Kaisha, Ltd.) and KATSUO KOGA (Resource Sharing Company).

†† 日本ユニバック(株)
††† (株)リソース

（3）「である」調である．「です」「ます」調ではない．

（4） 平叙文である．

2. 平仮名列の性質

2.1 平仮名列の種類と使用頻度

日本文の中で平仮名列はどの程度の種類があるか，また，その使用頻度について調査したものを調べた．電総研資料によると，次のようになっている．

（1） 調査対象資料　特許資料（電総研で入力）
（2） 調査文数　約3,000件
（3） 平仮名列の総数　32,000件
（4） 平仮名の種類　3,012件
（5） 頻度の高い上位1,000種で 30,000 件を占める．

この調査結果を図にあらわしてみると**図1**のようになる．平仮名列が1,000語を越えたあたりから語種が増えても累積 % はなかなか増加しないことがわかる（図1）．

このことから使用頻度の高い平仮名列をうまく分析し使用すれば，自動分かち書きに使えることがわかる．

2.2 平仮名の分類

平仮名列の分析を行うにあたって，次の3つの分類を行った．

（1） 頻度順の分類

仮名文字の使用頻度順に分類すると，助詞の“の”，“を”，“に”，“は”，“が”等が上位を占めていることがわかる．

また，上位に「び」のような語尾等の文字列が現われていることがわかる．

（2） 平仮名列を先頭文字からアイウエオ順に分類する．

平仮名列の先頭には動詞・形容詞等の語尾が残っているため，この分類からは分かち書きに有効な方法は得られなかった．

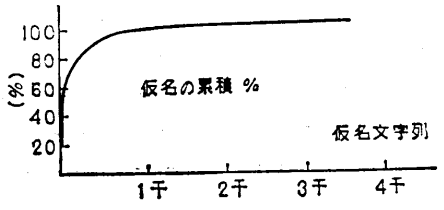

図1 仮名列の種類と異積 %

Fig. 1 Number of kana strings and its cumulative percentage.

表1 語尾順仮名文字列

Table 1 Kana-character strings sorted by end-of-word character.

仮名文字列	頻度
になる	17
のようになる	1
するようになる	1
まるようになる	1
されたことになる	1
ることになる	3

（3） 平仮名列を末尾文字からアイウエオ順に分類する．

この分類を分析してみると，かなり共通の文字列があることがわかる．この一部を表にすると**表1**のようになる．

2.3 平仮名列の分割

表1から，平仮名列の中には共通平仮名列があることがわかる．この共通平仮名文字列を抽出し，この文字列で平仮名列の分割を末尾から最長一致法で行ってみた．この内容をまとめると，次のようになる．

（1） 調査対象資料：特許資料
（2） 調査文数：約3,000件
（3） 平仮名列種類：3,012種
（4） 平仮名列総数：32,000件
（5） 平仮名列分割方法：右側最長一致法（1回だけ）
（6） 平仮名列辞書：2,768項目（分割に使用する平仮名列）
（7） 結果：738種類の平仮名列と1,128件の辞書の項目に分割される．

この実験により，処理できない文字列があった．これは頻度は少ないが，次のようなものである．

例 「りんおよびけい」

これは“けい素”の「けい」の部分が切り出されていたためである．

次に平仮名列辞書の一部（**表2**），平仮名列辞書により最長一致法で右側より分割した状況（**表3**），表3の右側に残った平仮名列の集約（**表4**），等をまとめた．

2.4 平仮名列の桁数

平仮名列が桁数別にどのように変化するかを，その種類件数と桁数別延件数で調べた．電総研 32,000 件を分析すると**表5**のようになる．桁数が増えるに従い件数が一様に減少していることがわかる．桁数別の平仮名列の種類は4～5桁が多いこともわかる．この結果はデータ件数を増やしてもほぼ同じである．

表 2 平仮名列辞書の一部
Table 2 A sample of kana-character string dictionary.

といえるであろう
かかるであろう
できるのであろう
ことができるのであろう
とするであろう
しようとするであろう
にするであろう
なくなるであろう
なければならなくなるであろう
となるであろう
ものとなるであろう
になるであろう
ようになるであろう
ことになるであろう
ことにもなるであろう
とされるであろう
がなされるであろう

表 3 平仮名列辞書により最長一致法で右側より分割した状況
Table 3 A sample result of segmentation.

る	といわれており	1
る	ことが	7
る	が	14
る	ことができるが	1
る	ことになる	3
る	ことができる	6
る	ためには	4
る	ようにされる	1
る	ことができない	1
る	ような	1
る	ようなはつきりした	1
る	ために	6
る	ための	3
る	ことの	1
る	ことによつて	2
る	ことはできなかつた	1
る	ように	4
る	と	6
る	こと	2
る	ようになつた	1
る	ことは	3
る	に	3
る	ことを	5
る	ことができ	1

3. 分かち書きの方法

平仮名列を平仮名列辞書に登録してある文字列で分割し，語尾等の文字列と平仮名列辞書にある文字列に分割することができる．平仮名列辞書の文字列はしばしば使われる平仮名列であり，慣用句のようなものである．平仮名慣用句は辞書により自動的に分割を行い，品詞付けも行うことができる．語尾等の文字列は，そのテーブルにより，語の推定や，品詞付けを行うことができる．

表 4 表3の右側に残った平仮名列の集約表
Table 4 End-of-word kana-character string.

する	1511
し	776
した	507
して	456
な	364
される	349
い	348
る	232
す	225
く	207
の	203
び	197
つ	159
に	149
り	147
させる	141
され	139
え	136
を	136
む	125

表 5 平仮名の桁数と発生頻度
Table 5 Frequency of occurrence of kana-character string.

桁数	種類	件数	件数 %
1	34	15,471	48.57
2	192	6,419	20.15
3	361	3,501	10.99
4	483	2,746	8.62
5	484	1,345	4.22
6	418	819	2.57
7	311	522	1.64
8	231	378	1.19
9	159	210	0.66
10	143	226	0.71
11	90	107	0.34
12	42	47	0.15
13	25	25	0.08
14	16	16	0.05
15	10	10	0.03
16	6	6	0.02
17	2	2	0.00
18	2	2	0.00
19	3	3	0.01
合計	3,012	31,855	100.00

例をあげて説明する.

(1) “するようになる”という平仮名が切り出される.

(2) “する△ようになる”右側最長一致により分割する.

(3) “する△よう△に△なる”平仮名辞書(平仮名慣用句)により分割する.

(4) 同時に品詞付けも行う.

する△ よう△ に△ なる
形名 助詞 動詞
→サ変動詞の語尾と推定する.

(5) “する”は語尾等の文字列のテーブルよりサ変動詞語尾とわかるので,サ変動詞を調べ,語幹を抽出し,分かち書きを行う.

以上のようにして分かち書きを行う.複数個の分かち書きの場合がある時には,それぞれの場合のテーブルをもうけて処理する.

切り出された平仮名列の右端から,品詞の推定は,可能性のある語そのもの,またはある品詞の活用グループを指定しておき,そのグループ内の単語を調べ見つけることができる.例をあげて説明する.

び:及び144件(90%),再び9件(5.62%),伸び4件(2.5%),選び2件(1.25%),運び1件(0.62%)合計160件(100%)

びは,バ行五段活用とバ行上一段活用の動詞と「及び」「再び」であることがわかる.

このように語尾のテーブル(切り出された平仮名列の左端)より,語を推定しながら確定し,分割することができる.また品詞付けも行うことができる.

以上のように処理できるもののほかに少ない件数ではあるが例外が発生する.たとえば次のようなものである.

例 けい素, ひ素, ろ過する

このため,前述の最長一致法だけでなく,このような例外を事前に除くことが必要である.

これらの方法でも処理できない例外が発生する場合がある.これらは平仮名列辞書を増やすとか,一部人手に頼る等の方法を取らざるをえない.

4. 分かち書き処理プロセス

平仮名列の分かち書きプロセスを各処理ステップごとに説明する.(図2)

(1) 正規化処理

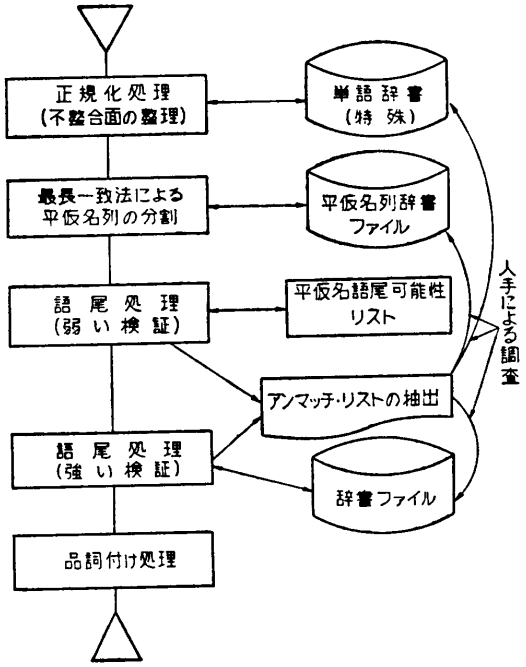

図2 平仮名列自動分かち書きプロセス

Fig. 2 Process of segmentation of kana-character string.

これは平仮名列の中に含まれる平仮名書き自立語,一部平仮名書き自立語の抽出を事前に行う.

たとえば“ねずみ”,“るつぼ”……等の平仮名書き自立語,“は握”“赤ん坊”……等の一部平仮名書き自立語を特殊単語辞書にファイルし,この中に含まれている単語か否か調べる.もしそうであれば抽出し平仮名列を事前に処理する.これにより平仮名列の不具合な面を整理する.

(2) 右側最長一致法による平仮名列の分割

約3万2千項目の平仮名列辞書により平仮名列を最長一致法で分割する.将来これは約5万項目程度に増やす予定である.

(3) 語尾処理(弱い検証)

この処理は最長一致による平仮名列の分割によって発生した文字列が,その可能性リストに含まれているか否かを調べる.もし含まれていなければアンマッチ・リストに印書し,人手により調べ平仮名辞書または特殊単語ファイルへ追加すべき用語として取り扱う.

(4) 語尾処理(強い検証)

最長一致法により切り出された平仮名列の左端を含む語がテーブルにあるか否かを調べる.

たとえば“研究する”の場合“する”により,研究という語が辞書にあるか否か調べ,さらにサ変動詞と

なっているかも検証する．もし該当する単語が無い場合は，その内容をリストし，人手によって調査し，平仮名列辞書ファイルに登録する．

弱い検証と，強い検証とに分けた理由は一度に多くの例外リストを出力するよりも段階を追って出力したほうがわかりやすいためである．また，強い検証は辞書引きをするため処理時間がかかる，このため処理時間がかからなくて検証できるものはあらかじめ処理しようという考えから弱い検証をもうけた．

（5） 品詞付けを行う

分かち書きの途中で平仮名列辞書を引くこと，接続検証，辞書ファイルの検索によってわかった情報をもとに品詞付けを行う．

5. 大量データによるテスト

前述したプロセスのすべてを実験することはできないので平仮名列を平仮名列辞書で，最長一致法を用いて分割する部分を特に実験した．また実験を行っても，小規模では処理可能でも，大規模な実験では不具合な点が発生する場合がある，そこで大量データによるテストを行った．個々の実験は 25 万件程度であったが，これを集約してみると次のようになる．

実験内容

（1） 実験に使用した資料：日本科学技術情報センター・ファイル

（2） 処理した平仮名列総数：2,325,045 件

（3） 1回の最長一致で平仮名列全体が一致した件数：1,686,169 件

（4） 1回の処理では部分的一致となり平仮名列の左端に文字列が残ったもの：638,876 件

（5） （4）の種類 7,967 件(*)

（6） 平仮名列辞書 32,771 項目

(*) 7,967 種類の文字列は平仮名書き自立語（たとえば，ねずみ，たんぱく，……）等を整理することにより約 3,000 種類程度に整理できる．

232 万件の膨大な平仮名列も 3 万 2 千項目程度の平仮名列辞書で処理できる．処理する平仮名列が増えても平仮名列辞書はその増加ほど増えない．

6. 分かち書きテストの考察

6.1 平仮名列辞書の収束について

このような実験を行ってみると平仮名列辞書が無限に拡大するのではないかという危惧があるが，これはつぎの 2 つの理由により問題にならない．

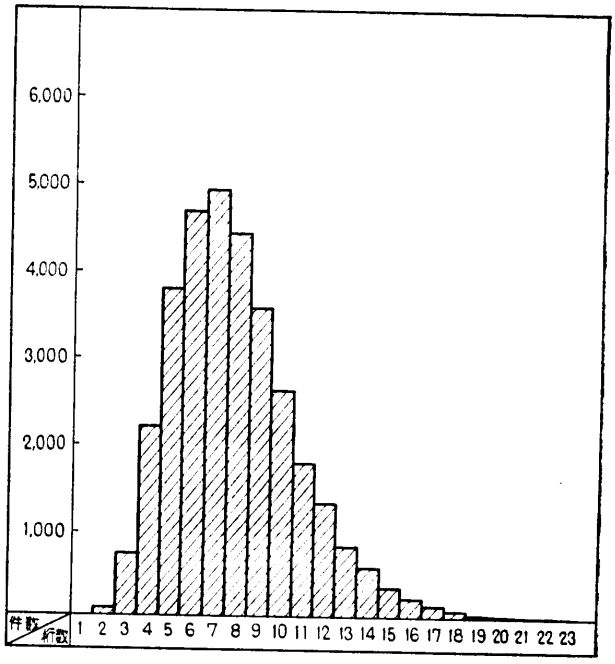

図 3 平仮名列辞書の桁数別頻度グラフ (32,771 件)

Fig. 3 Length of strings in kana-character string dictionary.

（1） 1回の実験は約 25 万件程度の処理件数であるが，その中に発生する新規に平仮名列辞書に登録しなければならない項目数が実験を行うごとに減少している．

（2） 平仮名列辞書の増加にともない桁数別の増加を棒グラフに表示すると，桁数の短いものから増加の伸びが止まってくる．またデータ中の長い桁数のものは発生頻度が少ないため平仮名列辞書の増加は急激には起らない．

平仮名列辞書に新しく追加しなければならない項目数は処理する平仮名列の 0.3%～0.4% でしかない (32,771 項目の辞書の場合)．このことから平仮名列辞書は約 4～5 万項目程度まで増やせば，ほぼ問題ないものとなる．

平仮名列辞書の桁数別項目数をグラフにまとめてみると図 3 のようになる．今後この辞書の増加にともない 7, 8, 9 桁あたりが伸びる．

今後は法律文，小説，新聞などの異なった分野でも，適用し研究したい．この平仮名列辞書の発生頻度の多い部分はほかの分野にも共通に使える．しかし，分野が異なると新しい平仮名列がかなり発生すると予想される．新しい平仮名列の種類は多いが，延件数に対する割合は少ない．

小型のコンピュータに平仮名列辞書を格納する場合

は発生頻度の少ないもの，桁数の長いものを除けば十分使える．精度はあまり落ちない．

5万項目程度のファイルを索引するには ISAM を利用すれば効率がよい．

6.2 このシステムの利用分野

この分かち書きは平仮名部分だけであるが，ほかの分野と組み合わせることにより，より利用分野が開かれる．

（1） 日本文の解析手段として，分かち書きの一手法として使える．

（2） 平仮名列辞書により，日本文の仮名漢字変換の自動分かち書きが行える（付属語を中心とした分かち書きが行える）．仮名表記でよい部分を積極的に見つけ出すこともできる．

（3） 文章の送り仮名チェック，誤字検出に利用することができる．最長一致法で平仮名列を処理するとあきらかに誤りと思われるものがしばしばみつかった．

6.3 今後の計画

平仮名列の分かち書きをさらに発展させるため次のことを考えている．

（1） 最長一致法で切り出された左端の平仮名列を整理し，接続検証のためのテーブル作成

（2） 接続検証の実施

（3） 接続検証により判定した用語の定量的分析

たとえば“研究する”という用語があると“研究”というサ変動詞の語幹がどの程度発生したか分析する．

7. おわりに

平仮名の分かち書きについて約2年あまり研究したが，今少しやり残した部分もあるが，かなり平仮名の性質，取り扱い方がわかってきた．今後この経験をもとにして，さらに研究を進めるとともに漢字の分かち書き，用語の分析などを行ってゆきたい．

最後にこの研究を指導して下さった日本科学技術情報センターの中井浩氏，茨城大学石綿教授，京都大学長尾教授に深く感謝する．

参考文献

1) 蓼沼良一，込山敏子：分かち書き一案，機械翻訳研究会 (1965, 5).
2) 込山敏子，蓼沼良一：分かち書き，第2回 Doc 研究集会 (1965).
3) 江川 清：漢字かな混り文の「自動単位分割」に関する研究，計量国語学.
4) 石綿敏雄他：単語認定プログラム，情報処理学会，CL研究委員会 (1969, 3).
5) 小野寺夏生他：化学的に有用なキーワードの検索のための自動的語分割システム，第 14 回情報科学技術研究集会発表論文集.
6) 斎藤秀紀：漢字仮名混り文のエントロピー，計量国語学 43/44 号 (1968).
7) 野村雅昭：漢字かなまじり文の文字連続，国立国語研究所報告 46 (1972).
8) 秋元啓次他：日本特許情報蓄積における分かち書きの二，三の実施例，第6回情報科学技術研究集会発表論文集.
9) 江川 清：単位分割自動化のシステムについて，計量国語学第 51 号.
10) 石綿敏雄，斎藤秀紀，木村 繁：言語単位分割自動化の研究，計量国語学第 50 号.
11) 坂本義行：文節の認定，昭和 53 年プログラミング・シンポジウム夏の大会予稿.
12) 板山和彦，荒木啓介：JICST 理工学文献ファイルの文献文標題，漢字・カナ変換用辞書作成の試み，第 13 回情報科学技術研究集会発表論文集.
13) 現代雑誌 90 種の用語用字(3)分析，国立国語研究所，秀英出版.
14) 中井，古賀，高浜，中瀬：汎用構文解析システム AUTOSEG-II について，第 14 回情報科学技術研究集会発表論文集，pp. 181-192.
15) 坂本：日本文の KWIC 特許資料，電総研.
16) 田中，古賀：日本文の自動分かち書きについて（仮名文字列の自動分かち書き），第 15 回情報科学技術研究集会発表論文集.
17) 長尾：計算機による日本語文章の解析に関する研究，昭和 53 年度研究報告書.

（昭和 54 年 11 月 2 日受付）
（昭和 55 年 12 月 18 日採録）