日本語文書処理　6－3
（1986. 5. 14）

カナ漢字変換用辞書を利用した日本語雑誌タイトルの分かち書き

桧垣泰彦*　　池田宏明*　　堀込静香**
* 千葉大学工学部　　** 千葉大学附属図書館

日本語で記述されている雑誌のタイトルを分かち書きする一つの方法として，ここではデータの先頭から順に最長一致の原則に従って，漢字表現と仮名読みを頼りに辞書と照合するという非常に簡単な方法で分かち書きを行うことを試みた．分かち書きに利用する辞書はシステムのかな漢字変換用の辞書をもとに作成した．その結果，意味的に正しく分かち書きされなかったものも含めると約８５パーセントのものが一応分かち書きされた．全体の約５０パーセントのものについては人手による修正を加えずに利用できた．残りの約半数のものについては，後で人手による修正を必要としたが，人手による分かち書き作業の前処理として利用することを前提とすれば一応の効果が得られた．

A Simple Method of Wakachigaki for Japanese Magazine's Title Using Dictionary for Kana-Kanji Translation

Yasuhiko HIGAKI*,　Hiroaki IKEDA*,　Shizuka HORIGOME**
* Faculty of Engineering, Chiba University　** Chiba University Library

'Wakachigaki' in Japanese means to separate Japanese text into some words. Japanese text style is ordinarily not separated by space like English text. So it is necessary to separate text into words to make searching index for information retrieval system. A very simple method of this wakachigaki using dictionary of Kana-Kanji translation is described. This method is not perfect, but it is useful sometimes because of it's simplicity. This is a case study of applying this method to Japanese Magazine's title.

１．はじめに

筆者らは本学附属図書館の所蔵する和文・欧文学術雑誌の所在を研究室のＴＳＳ端末から検索するための学術雑誌所在検索システムを構築中である．雑誌タイトルから検索する場合に，欧文では単語単位にインデクスすることが容易に行えるが，和文の場合はデータがあらかじめ適当に分かち書きされて入力されていない場合には，インデクスの作成は容易ではない．開発中の検索システムの和文データは学内に配布する「学術雑誌総合目録」を印刷するために作成されたデータであったため，分かち書きされてはおらず，これをインデクスするのに都合の良いように適当な方法で分かち書きする必要を生じた．

現在では既に日本語の文書を分かち書きすることは勿論，日本語文書から検索のためのキーワードを自動抽出する研究も行われている[1]-[2]が，これらを今回の例のような１回限りの処理のために応用できる迄にはユーティリティが普及しておらず簡単には利用できない．

このようなコンピュータによる完全自動化された分かち書きでは無く，人手による分かち書き作業の補助としてコンピュータを利用することを考えた場合，もっと単純で簡単な方法で分かち書きを行っても，全てを人手に頼るのに比べれば労力を大幅に節約できるはずである．

ここでは，このように人手による分かち書き作業の前処理とすることを前提として，できるかぎり簡単な方法で日本語の分かち書きを行う方法を示すと同時に，和文学術雑誌タイトルの分かち書きに適用した場合どの程度の結果を得ることができるかを示す．

２．テストデータ

分かち書きする対象は「千葉大学学術雑誌総合目録」を印刷するためにＭＴ上に作成されたデータ（６４１８タイトル分）のタイトル表記部分である．このデータには千葉大学附属図書館の所蔵する学術雑誌の雑誌名（タイトル）や出版社，雑誌の所在（配架先）と所蔵データなどが記録されている．各雑誌のタイトルは漢字混じりによる表現とカナ読みによる表現とで記録されている．カナ読みは分かち書きされてはいない．

例：漢字まじりの表現 → 情報処理学会論文誌

カナ読み → ｼﾞｮｳﾎｳｼｮﾘｶﾞｯｶｲﾛﾝﾌﾞﾝｼ

これらのデータはコンピュータにより処理することを一応意識してはいるが，そのデータには統一のとれていないものが多い．ｼﾞｮｳﾎｳ が ｼﾞﾖｳﾎｳ となっているものや，かなのばし（ー）がマイナス記号（－）ではいっているものなどがあった．

３．分かち書きの方法

現在では人手の介在しない事を目指した分かち書きも実現されているが，ここでは，システムのかな漢字変換用の辞書を利用して，最も単純な方法で分かち書きを行った．その方法について説明する．

「情報処理学会論文誌」を分かち書きする例で説明する．図１に示すような漢字による表現とカナによる表現をもとに，図２の分かち書き用辞書を参照しながら分かち書きを行う．辞書には漢字による表現と，カナによる表現が登録されている．

漢字表現データの１文字目から順に辞書をひきながら，漢字表現もカナ読みも一致しているものを捜す．このとき分かち書きの結果として図３に示すようにいくつか考えられるが図３(b)の様にできる限り長く一致するもの（最長一致）を採用するようにする．

４．辞書の構造

分かち書きのための辞書は，漢字表現の語句をキーとして高速にひくことができなければならない．そこで，ＶＳＡＭのキー順データセット（ＫＳＤＳ）を使用した．図４に示すように漢字表現のために４０バイト，カナ読みのために４０バイトのフィールドをそれぞれ用意し１レコード８０バイトとした．そしてこの８０バイトを全てキーとして定義した．また，語句の後ろの余った空白の部分はFF_{16}で埋めた．このようにすれば図４に示したように「情報処理」の

様な複合語が「情報」の様な単語より先に並ぶことになり，先に述べた最長一致による分かち書きを実現できることになる．

図１．サンプルデータ

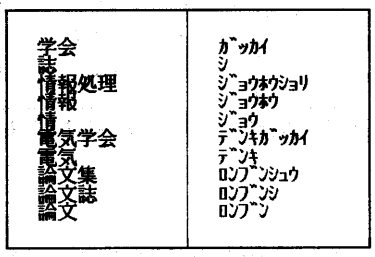

図２．分かち書き用辞書

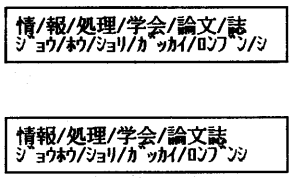

図３．分かち書きの結果

辞書をひくには，例えばＰＬ／Ｉの場合には，ＶＳＡＭの総称キー（ＧＥＮＫＥＹ）オプションを使用して

```
dcl dict file record input keyed env(vsam,genkey);
dcl 01 rec, 02 kanji char(40), 02 yomi char(40);
read file(dict) into(rec) key('情報');
read file(dict) into(rec);
```

のようにすればよい．図４の場合，初めのkeyオプション付きのread文で「情報処理学会」が得られ２つめの順次読みのread文で次の「情報処理」が得られる．もう一度順次読みのread文で読むと「情報」が得られる．

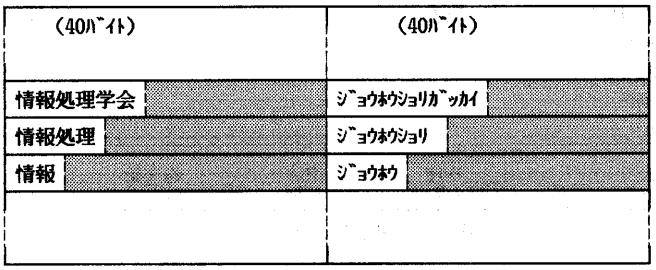

図４．分かち書き用辞書の構造

５．辞書の作成

辞書の作成には，システムのカナ漢字変換用辞書[3]（登録語数139,258語）を使用した．この辞書は主にシステムのカナ漢字変換に使われるもので，一般名詞の他，地名や人名などの固有名詞が登録されている．この辞書をもとに４で述べた分かち書き用辞書を作成した．

(1) カナ漢字変換用の辞書からカナ読み部分と漢字表現部分を取り出し分かち書き用辞書を作成する．このとき，カナ漢字変換用の辞書は図５に示すように同じ読みで複数の表現が考えられる場合には「☆」で区切られて入っているのでこれを分割する必要がある．また，漢字表現からひけるように漢字表現の部分を第一キーとなるようにする．

(2) テストデータは「ｼｮﾘ」と「ｼｮﾘ」のように表記に統一性が無いので， これに対応するため「ｼｮﾘ」を読みだして「ｼｮﾘ」に変換して登録し両方どちらでもひくことのできる様にする．

(3) カナ漢字変換用辞書に登録されていない接頭語，接尾語，かな，アルファベット，記号を追加登録する．

以上の作業を行った結果，完成した分かち書き用辞書の語数は１８５，４０５語になった．また，その大きさはレコード長８０バイトの固定長形式で約１５Ｍバイトとなった．

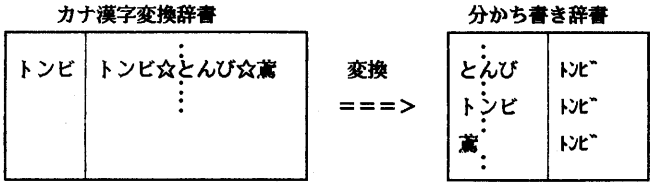

図5．カナ漢字変換辞書から分かち書き辞書への変換

6．分かち書きプログラム

5．で述べた分かち書き用辞書を使って分かち書きを行うプログラムについて説明する．分かち書きは最長一致で行うが，ここでは簡単のため次のような方法でこれを実現した．

初めに，タイトル全部をキーとして辞書をひく．辞書にその語が有ればそのカナ読みを比較する．もし，その語が辞書にないか読みが異なる場合には一文字短くして同様に辞書をひく．これを繰り返し漢字表現もカナ読みも一致した時点を分かち書きすべき部分であると判断する．漢字表現とカナ読みについて処理の終わった部分を除いた部分についてこの操作を繰り返す．もし一文字になってその語が辞書に無かったり，読みの一致するものが無い場合には分かち書き不能ということになる．この過程をを図6に「情報処理学会誌」の例で示す．これは，分かち書き出来た例である．

以上のような原理に基づきＰＬ／Ｉにより分かち書きプログラムをコーディングした．

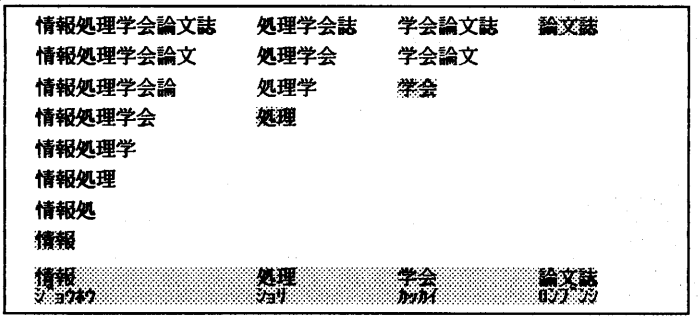

図6．分かち書きの過程

7．実験結果

以上の方法で和文雑誌６４１８タイトルのテストデータについて分かち書きを試みた結果，次ぎのような結果が得られた．

テストデータ数	６４１８件	
分かち書きできたもの	５４２５件	（８４．５%）
分かち書きできなっかたもの	９９３件	（１５．５%）
分かち書きに要したＣＰＵタイム	７４７秒	
（一件当りの平均ＣＰＵタイム	０．１１６秒）	

ここで分かち書きできたとは，６．に述べた原理で分かち書き処理を行ったとき全ての語が辞書に有り処理を終了した時点で漢字表現データもカナ読みデータも空になった（全て処理できた）事を意味する．後で触れるようにこれらの中には意味的には必ずしも適当では無い分かち書きがなされたものも含まれている．付録に分かち書き出来たもののリストの一部を掲げる．

分かち書きに要したＣＰＵタイムは６４１８件のテストデータ全てを処理するのに要した時間で千葉大学情報処理センターのＨＩＴＡＣ　Ｍ－１８０での処理時間である．

８．考察

付録に示すように一応分かち書き処理できたものの中にも適切な位置で分かち書きされていないものも含まれている．

一般には漢字で表現するものを仮名で表現したものや，英語等の表記が一部に含まれているものはうまく分かち書き出来ない．「あ.お.も.り」，「アイ.デ.ア」，「A.u.d.i.o.l.o.g.y」等がその例である．辞書に「あ」，「お」などの仮名や「デ」，「ア」などのカタカナ，「A」，「u」などのアルファベットも登録したため一応は辞書との照合はうまくいっているがこれらは人手による修正を行わなければ使用できない．

「技術家.庭.教育」の例は「技術家」という語が辞書に有ったため最長一致の原則によりこのようになってしまったものである．この様な簡単な方法を取る限りこれを「技術.家庭.教育」と分けるのは難しい．

かな漢字変換用の辞書を利用したため一般に接尾語がうまく処理されていない．かな漢字変換では接尾語を特別扱いするため辞書には登録されていない．分かち書き用辞書を作成する時点でこれを追加したが，「高分子.論文.集」などの様に接尾語の部分が独立してしまう結果となった．

正しく分かち書きするのに必要な語が辞書になく，「医.用器.材.研究.所.報告」のように分かち書きされてしまった例もある．この場合は「医用」という語が辞書に有ればうまく分かち書きできたはずである．この「用」もかな漢字変換時には接尾語扱いになっているものと思われる．

以上のように一応分かち書きできたものの中にも人手による修正を必要とするものが含まれており，全く人手を加えずに利用できるのは初めのデータの内，約５０パーセントであった．

次ぎに，分かち書きの過程において辞書に無い語に出会うなどのため，分かち書き出来なっかたものの原因について触れる．「口腔」，「医事」，「房総」などの語が辞書に無く，これらを含むデータは分かち書き出来ていない．

「文学論叢」の様な場合，最長一致に従って「文学論.叢」に分けようとするため「文学」や「論叢」が辞書に有っても「叢」を捜してしまい，分かち書きに失敗している．

「０　Ｐｌｕｓ　Ｅ」というタイトルの読みとして「ｵｰﾌﾟﾗｽｲｰ」を与えていたり，「ｾﾝﾀｰ」ではなく「ｾﾝﾀｱ」としているなど読みの与え方が適切ではないものが，変換できなっかったもののうち２０パーセント程度あった．

９．まとめ

ここでは，前方から順に最長一致で辞書との照合を行う非常に単純なアルゴリズムによって，和文雑誌のタイトルを漢字表現とカナ読みを頼りに分かち書きする事を試みた．その結果，約８０パーセントが一応分かち書きでき，全体の５０パーセントのものについては人手を加えなくても利用できる事が判った．人手による分かち書き作業の前処理とするのであれば利用価値のある結果を得ることができた．

謝辞：　日頃から御指導御鞭達を頂いている山本博教授に感謝致します．

参考文献

[1] 斉藤,野寄：日本語文解析によるキーワード抽出，電子通信学会技術研究報告，AL81-45,1981.7

[2] 吉村,日高,吉田：日本語科学技術文における専門用語の自動抽出システム,情報処理学会論文誌,Vol.27,No.1,
pp.33-40(1986)

[3] HITACﾏﾆｭｱﾙ：カナ漢字変換辞書，資料番号 8080-2-056

付録　分かち書きできたものの例（不適当に分かち書きされたものも含む）

1	あゆみ.	ｱﾕﾐ/
2	アメリカ.法.	ｱﾒﾘｶ/ﾎｳ/
3	エレクトロニクス.ダイジェスト.	ｴﾚｸﾄﾛﾆｸｽ/ﾀﾞｲｼﾞｪｽﾄ/
4	医学.の.あゆみ.	ｲｶﾞｸ/ﾉ/ｱﾕﾐ/
5	医学.輯録.	ｲｶﾞｸ/ｼｭｳﾛｸ/
6	医学.図書館.	ｲｶﾞｸ/ﾄｼｮｶﾝ/
7	医薬.ジャーナル.	ｲﾔｸ/ｼﾞｬｰﾅﾙ/
8	育種.研究.	ｲｸｼｭ/ｹﾝｷｭｳ/
9	茨城県.気象.月報.	ｲﾊﾞﾗｷｹﾝ/ｷｼｮｳ/ｹﾞﾂﾎﾟｳ/
10	英文学.研究.	ｴｲﾌﾞﾝｶﾞｸ/ｹﾝｷｭｳ/
11	科学.技術.研究.調査.報告.	ｶｶﾞｸ/ｷﾞｼﾞｭﾂ/ｹﾝｷｭｳ/ﾁｮｳｻ/ﾎｳｺｸ/
12	科学.測器.	ｶｶﾞｸ/ｿｯｷ/
13	基礎.工.	ｷｿ/ｺｳ/
14	季刊.レオロジー.	ｷｶﾝ/ﾚｵﾛｼﾞｰ/
15	郷土.趣味.	ｷｮｳﾄﾞ/ｼｭﾐ/
16	金融.統計.月報.	ｷﾝﾕｳ/ﾄｳｹｲ/ｹﾞﾂﾎﾟｳ/
17	計測.自動制御.学会.論文.集.	ｹｲｿｸ/ｼﾞﾄﾞｳｾｲｷﾞｮ/ｶﾞｯｶｲ/ﾛﾝﾌﾞﾝ/ｼｭｳ
18	結核.研究.の.進歩.	ｹﾂｶｸ/ｹﾝｷｭｳ/ﾉ/ｼﾝﾎﾟ/
19	月刊.児童劇.	ｹﾞﾂｶﾝ/ｼﾞﾄﾞｳｹﾞｷ/
20	月刊.実践.障害児.教育.	ｹﾞﾂｶﾝ/ｼﾞｯｾﾝ/ｼｮｳｶﾞｲｼﾞ/ｷｮｳｲｸ/
21	月刊.労働.調査.時報.	ｹﾞﾂｶﾝ/ﾛｳﾄﾞｳ/ﾁｮｳｻ/ｼﾞﾎｳ/
22	建築.ニュース.	ｹﾝﾁｸ/ﾆｭｰｽ/
23	現代.数学.	ｹﾞﾝﾀﾞｲ/ｽｳｶﾞｸ/
24	個人.企業.経済.調査.年報.	ｺｼﾞﾝ/ｷｷﾞｮｳ/ｹｲｻﾞｲ/ﾁｮｳｻ/ﾈﾝﾎﾟｳ/
25	呼吸.と.循環.	ｺｷｭｳ/ﾄ/ｼﾞｭﾝｶﾝ/
26	光合成.関係.文献.抄録.集.	ｺｳｺﾞｳｾｲ/ｶﾝｹｲ/ﾌﾞﾝｹﾝ/ｼｮｳﾛｸ/ｼｭｳ/
27	高分子.化学.	ｺｳﾌﾞﾝｼ/ｶｶﾞｸ/
28	国際.科学.情報.	ｺｸｻｲ/ｶｶﾞｸ/ｼﾞｮｳﾎｳ/
29	自転車.生産.技術.研究.報告.書.	ｼﾞﾃﾝｼｬ/ｾｲｻﾝ/ｷﾞｼﾞｭﾂ/ｹﾝｷｭｳ/ﾎｳｺｸ/ｼｮ/
30	実地.医学.	ｼﾞｯﾁ/ｲｶﾞｸ/
31	授業.研究.	ｼﾞｭｷﾞｮｳ/ｹﾝｷｭｳ/
32	上方.はな.し.	ｶﾐｶﾞﾀ/ﾊﾅ/ｼ/
33	情報.処理.学会.論文誌.	ｼﾞｮｳﾎｳ/ｼｮﾘ/ｶﾞｯｶｲ/ﾛﾝﾌﾞﾝｼ/
34	情報.処理.研究.	ｼﾞｮｳﾎｳ/ｼｮﾘ/ｹﾝｷｭｳ/
35	新し.い.算数.研究.	ｱﾀﾗｼ/ｲ/ｻﾝｽｳ/ｹﾝｷｭｳ/
36	青森県.立.中央.病院.医.誌.	ｱｵﾓﾘｹﾝ/ﾘﾂ/ﾁｭｳｵｳ/ﾋﾞｮｳｲﾝ/ｲ/ｼ/
37	大学.資料.	ﾀﾞｲｶﾞｸ/ｼﾘｮｳ/
38	大審院.刑事.判決.抄録.	ﾀﾞｲｼﾝｲﾝ/ｹｲｼﾞ/ﾊﾝｹﾂ/ｼｮｳﾛｸ/
39	地域.開発.	ﾁｲｷ/ｶｲﾊﾂ/
40	電気.計算.	ﾃﾞﾝｷ/ｹｲｻﾝ/