日本語文入力用カタカナ語検出規則とオンライン国語辞典の一分析†*

木　村　泉††

単語わかち書きされたかな書きの日本語文において，ひらがな語とカタカナ語の区別が前もって与えられていないとするとき，規則(1)長音符を含む語はカタカナ語，(2)ンまたは小さいッの直後以外にパ行音があらわれたらカタカナ語，および(3)小さいァィゥェォを含む語はカタカナ語，の以上三つを利用すると，ひらがなとカタカナの切りわけという問題のかなりの部分を利用者による積極的指定なく処理でき，これらの規則にあてはまらない場合の処置も容易であることはすでに報告した．本文では三省堂「新明解国語辞典第二版」の見出し項目を全数検査することによってこれら3規則の性能を評価するとともに，類似の規則を追加することによって性能の向上をはかることができないか調べる．おもな結果は次の通り．(a)上記3規則には事実上例外がない．すなわちひらがな語を誤ってカタカナ語と判定することははっきり特殊とわかる場合を除いてはほとんど存在しない．(b)3規則でカバーされるカタカナ語の割合は 60% 近い．(c)ション，ックスなど外来語によくある語尾に着目する方法は思いがけない例外があって上記3規則ほど具合がよくない．

1. 背　景

三省堂「新明解国語辞典第二版」[1]から見出し語を拾い出し，その中のカタカナ語およびひらがな語**に，それぞれどのような特徴があるか調べてみたので報告したい．そうした動機は次の通りである．

筆者はかねて，日本語技術文書の作成を助ける機械的支援はいかにあるべきか，という問題に興味をもっており，その一環としてわかち書きのひらがなカタカナ英語まじり文を利用したらどうかと考えている．実際，ローマ字風入力に基づいてそのような形式の日本語文を打ち出すシステムを試作してみたことは，すでに報告した[2]．

そのシステムでは，入力はローマ字風*** に表記されており，したがって出力結果においてどこをひらがなにし，どこをカタカナにするかを指定してやる必要がある．たとえば大文字でカタカナ，小文字でひらがなをあらわすなどの案もあり得るが，(たとえロック式のシフトを使うにしても) 全部のカタカナ語について2ストロークの余分な打鍵が必要になり，きわめてわずらわしい．もっと少ない打鍵数ですっきり行かないものか？

そこで綴りが一目みてひらがな語ではあり得ないような場合はカタカナ語として扱い，カタカナ語であることが必ずしも明らかでない語についてのみ，適当なシフトコードを語の先頭につける（その効果は語の終りのスペース，その他の区切り記号をもって打ち切られるものとする）という手を考えた．ただしカタカナとひらがなの両方を含む語（以下**混成語**と称する）については境目に適当なシフトコードを入れるものとし，一般には，一つの綴りはひらがなカタカナのどちらかだけから成る程度にこまかくわかち書きするものとした****．

さきに報告したこころみでは，次の三つの規則をもちいた．

(1) **長音則**——長音符「ー」を含む語はカタカナ語．

(2) **パ行則**——「ン」または小さい「ッ」の直後以外のところにパ行の文字を含む語はカタカナ語．

(3) **アイウエオ則**——小さい「ァィゥェォ」を含む語はカタカナ語．

これらの規則はおぼえやすく，その割合には強力で，なかなか使いやすいものといえたが，もともと直観的にみつけ出されたものであり，なんら理論的根拠はなく，これでたしかによいものかどうかは気になる

---

† Analyzing a Machine-Readable Japanese Dictionary for Evaluating a Set of Rules That Detect Imported Western Words by IZUMI KIMURA (Tokyo Institute of Technology and Carnegie-Mellon University).

†† 東京工業大学理学部・カーネギーメロン大学計算機科学科

* この論文の「日本語技術文書用入力設計と国語辞典の一分析」と題する原型は第22回プログラミングシンポジウム報告集 pp. 9-17 (1981) に掲載された．

** かな書きするとすればそれぞれカタカナおよびひらがなで書かれる語．後者は漢語を含む．

*** おおむねヘボン式に準じ，「私は」は「watakushi ha」と書くなど，かな表記を基本にとる．そのようなものを選んだ一つの理由は外来語の表記を軽快にしたかったことにある．

**** このように文献2)の話はローマ字風表記に基づいているが，以下では原入力はカタカナわかち書きであると考えていただいてもさしつかえない．

点であった．そこでこれらの規則が実際にどのくらいのパフォーマンスを示すものかを調べ，あわせてこれらの規則に新項目を追加していっそう強力にすることができないか検討しようと思い立った．

そのためには，実際の文書作成作業にこれらの規則を大々的に使ってようすを見る，というのがおそらくもっとも筋目正しい行きかたであろうが，国語辞典にとりあげられている語を片っ端からあてはめてみることによって，これらの規則の地位を明らかにする，というのもまた一つのおもしろい方向と考えた．以下本文で報告するのは，そういう方向の研究の一結果である．

辞書に出ている語のうちには使用頻度の高いものも低いものもあり，それらを一律に勘定するのは問題といえば問題であるが，一方では辞書は普遍性を重んじて編集されるものであり，それなりの意味をもったデータが得られるのではないかと期待される．

なお，上記入力方式[2]については，シフトコードを綴りの先頭でなくむしろ末尾につけた方がよいのではないかとの疑問も出よう．そしてそれは一理ある考えかたである．「カタカナ語これこれ」といわされるよりは「あっと，これカタカナね」という方が一般利用者にとってはずっと楽，という可能性もある．ただし混成語の扱いとのからみから，入力設計の全体としての調和をとることが若干むずかしくなる．

それともう一つ，上のやりかたではたとえば「パフォーマンス　を」というようなわかち書きを前提とすることになるが，これはポピュラーなわかち書き方式である文節わかち書きと比べると，ずっとこまかいので好ましくないのではないか，との疑問も出るかも知れない．これについては，文節わかち書きの節末から助詞等を切り落す程度のことは，もし望むなら安価に自動化できると思われるので，その方向で考えてはどうであろうか？　ただし筆者の個人的感覚では，漢字かなまじり文において漢字とかなの間に存在する自然な境界をかなわかち書きで表現しようとすれば，文節わかち書きは粗すぎ，むしろ格助詞を切って書く方式の方が自然のような気がする．これらの点についてのくわしい議論は，システムの全体としての使われかたに依存するので別の機会にゆずりたい．

2. 辞書ファイルの概要

ここで，本文の研究に使用した辞書ファイルについて若干の心おぼえを記しておこう．原ファイルは電子技術総合研究所で開発された．それを筆者の手もとで一応 JIS 漢字コード[3]に変換した上でもちいた．ここで利用したのは辞書本文に対応する部分だけであり，漢字に関する解説を記した部分（「漢語の造語成分」の解説）は利用しなかった．利用した部分の総行数は 114,962行であり，その中には 58,439個の主見出し項目が含まれており，うち 1,812 項目に最重要語，3,434 項目に重要語の表示が与えられていた．

以上から今回は見出し語のみ拾い出した．上記 58,439 個の主見出し項目のほか，複合語項目の一部と派生語項目全部を拾った．前者は総数 7,736 項目あったうち，表記が漢字を含むものを除き，1,508 項目を取り出した．（たとえば主見出し項目「アール　エイチ」のもとには子見出しとして「――因子」および「――マイナス」が与えられていたが，前者には読みが与えられておらず，手もとの計算設備では処理があまりにもめんどうなので割愛し，後者のみ「アール　エイチ　マイナス」という形で拾った．）派生語項目（後者）は全部で 1,657 項目あったが，実はそれらは次の 5 種類しかなかった．

「――がる」の形	106 項目	計 1,657 項目
「――げ」　の形	260 項目	
「――さ」　の形	1,278 項目	
「――そう」の形	1 項目	
「――み」　の形	12 項目	

取り出した項目の総数は 61,544 語である．なお主見出し項目の中には一部「漢語の造語成分」への参照が含まれている．たとえば見出し「あ」のもとに拾われた項目のうちの一つに「亜・阿・唖・鴉――→〔漢語の造語成分〕」なるものが含まれている．ただし前記のように，これらの漢語の使われかたを説明した解説部分は今回は利用しなかった．

当然ながら若干のデータ誤りがあった．たとえば「けしきばむ」が「けしきばむ」となっていた，などのような原ファイルの打鍵過程に起因すると思われるものがあった一方，「四面楚歌」の楚が区切り記号に化けてしまった，というような JIS コードへの変換中に発生したと見られる誤りもあった．気づいた綴り誤りについては適宜補正したが，見落しも多少ないとはいいきれない．ただし大勢には影響ないものと信ずる．

項目の処理順（したがって以下に示す各種の処理結果の配列順序）は，諸般の事情により不同であり，たまたま原ファイルに配列されていた順序（おおむね重要語，一般語の順にそれぞれアイウエオ順），ローマ

字風表記のアルファベット順などが混在していることを前もっておことわりしておく．また上例において，「アール」と「エイチ」の間に空白があけてあるのは，もとの辞書において語の構造を明らかにするために置かれている空白をそのまま受け継いだものである．空白のほか，原著では活用語尾を語幹と区切るのに中点（・）が使われている．以下の語例では（一部を除き）これらをそのまま残してある．

3. 既成3規則の評価

手はじめに，抽出した見出し項目計 61,544 語（前章）をひらがな語，カタカナ語，混成語の三つに分類し，それらを第1章の規則（1）～（3）にあてはめてみた．その結果を**表1**に示す．

ただし一般語，重要語とあるのは，原ファイルがその2区分にわかれている（重要語ファイルには原著での最重要語も含む）のを個別に処理してみたものである．重要語という中には重要語の子複合項目，および派生語項目も含めてある．混成語には「あか　ゲット」，「あか　チン」などがある．

百分率は一般語，重要語，総計のそれぞれについて「たて」に計算してある．表から，重要語ではカタカナ語が少なく，ひらがな語が多め，という当然の傾向があることがわかる．例外ひらがな語とあるのは規則（1）～（3）のどれかに該当しながら実はひらがな語であるもの，例外カタカナ語とは規則（1）～（3）のどれにも該当しないカタカナ語を指す．例外混成語とはそのひらがな成分またはカタカナ成分（「あか　ゲット」における「あか」または「ゲット」など）に例外語（例外ひらがな語，または例外カタカナ語）が一つ以上含まれているような混成語を指す．「あか　ゲット」はこの部類に属する．

表1　既成3規則による判定状況

Table 1　Performance of the three rules.

区　分	一般語	重要語	計
総件数	54,750(100.0%)	6,794(100.0%)	61,544(100.0%)
ひらがな語	49,282(90.0)	6,473(95.3)	55,755(90.6)
カタカナ語	5,162(9.4)	296(4.4)	5,458(8.9)
混成語	306(0.6)	25(0.4)	331(0.5)
例外ひらがな語	111(0.2)	0(0.0)	111(0.2)
例外カタカナ語	2,204(4.0)	100(1.5)	2,304(3.7)
例外混成語	222(0.4)	17(0.3)	239(0.4)
例外語計	2,537(4.6)	117(1.7)	2,654(4.3)

例外ひらがな語がきわめて少ないことは特筆に値する．実際，それらは**図1**に示すだけ（111語）しかない．このうち，「ぽっくり」が二つ出ているのは，「木履」という意味のぽっくりと，「ぽっくり死んだ」というときのぽっくりの両方がリストアップされているものである．「ちぇえ」は「ちぇえ，かたじけねぇ」のように使うもの，「ちぇえすと」は鹿児島方言の由，「ごっつぁん」は周知のすもうことばであり，これら

ちぇえ　ちぇえすと　ちぇっ　がた ぴし　ごっつぁん　まる ぽちゃ　にこ ぽん　ぱあ　ぱちくり　ぱちんこ　ぱち ぱち　ぱち つか・せる　ぱく ぱく　ぱくり　ぱく・る　ぱく つ・く　ぱぱっ と　ぱっぱ と　ぱら ぱら　ぱらり と　ぱら つ・く　ぱり ぱり　ぱりっ と

ぱさ ぱさ　ぱたり と　ぱたっ と　ぱっちり　ぱったり　ぱっ と　ぺ　ぺちゃ くちゃ　ぺちゃんこ　ぺい ぺい　ぺこん と　ぺこ ぺこ　ぺんぺん ぐさ　ぺら ぺら　ぺろ ぺろ　ぺろり　ぺしゃん こ　ぺたん と　ぺた ぺた　ぺてん　ぺったり　ぴちゃ ぴちゃ　ぴち ぴち

ぴちっ と　ぴい ぴい　ぴか どん　ぴか いち　ぴか ぴか　ぴかり と　ぴかっ と　ぴくぴく　ぴくり　ぴん ぴん　ぴん しゃん　ぴん と　ぴり ぴり　ぴしゃり　ぴし ぴし　ぴたり　ぴっちり　ぴったり　ぴちゃ ぴちゃ　ぴち　ぴ・い　ぴい　ぴか

ぱかん と　ぱか ぱか　ぱかっ と　ぱっかり　ぱっきり　ぱっくり　ぱっくり　ぱん ぴき　ぱん こつ　ぱん ぱん　ぱん つく　ぱっぱ　ぱろ ぱろ　ぱろり　ぱた ぱた　ぱたり　ぱっち　ぱっちり　ぱつん　ぱつねん と　ぱつ ぱつ　ぱつり　ぱつり ぱつり

ぱっ と　ぱっつり　ぱうっ と　ぷい と　ぷか ぷか　ぷん ぷん　ぷん と　ぷりん ぷりん　ぷり ぷり　ぷっ と　ぴょい と　ぴょこん と　ぴょこ ぴょこ　ぴょん ぴょん　ぴょん と　すぱ すぱ　すぱっ と　たんぱぱ　てき ぱき

図1　例外ひらがな語

Fig. 1　The exceptional hiragana words.

は特殊なものであるからシフトコードを要求しても問題あるまい．残りの大部分は擬声語であり，カタカナで書いてもそうおかしくはなく，また「ぱち　つか・せる」のように全部カタカナで書いては異様であるようなものも，適宜区切って「パチ　つかせる」のように書けばうまく行く．ぜひこれらをひらがなで書きたい利用者にも「パ行を含む擬声語には気をつけてください．」とだけいっておけばすむ．

ちょっと意外の感のあるのは「たんぽぽ」である．きわめて一般的なひらがな語でありながらパ行則によってカタカナ語と判定される．語源上実は擬声語からきた，というような特殊な事情があるのだろうか？ともかくひらがな語に関する限り，ことは意外にうまく行っているといってよかろう．

残念ながらカタカナ語については，カタカナ語でありながらそうと認識されないものが非常に多い．文献 2) の建前では，これらについてはシフトコードで処理することになるわけである．三つの規則による個別的なカタカナ語識別状況をまとめて，**表2**に示す．

一般語，重要語の別については表1におけると同様である．累積長音則などとあるのは規則（1）～（3）によってカタカナ語と判定されるものの総数を示し，重複を含む．個々の規則の効き具合をもっとくわしく分析したものが「(1)のみ」の行以下である．(1)～(3)はこの順に効力が高いので，それらを順に適用して行ったとき，あらたにどれだけのカタカナ語が認識されることになるかを，最後の3行に示す．例によって百分率は「たて」に計算してある．

われわれの規則の効きは総計でみたとき 57.8% しかない．もっとも，重要語についてのみ見たときには 66.2% と，若干よくなる．規則（3）――アイウエオ則―― の効きが比較的小さいのは，ひとつには文献 1) が小さい「ァィゥェォ」をあまり使わない傾向をもっていることにもよる．たとえば項目として「アイディア」をたてずに「アイデア」をたてている．

なお，この解析をしてみて，規則（3）の「ァィゥェォ」のほかに小さい「ヮ」があることに気づいた．「クヮルテット」のように使われ，はっきりカタカナ語としての特徴をあらわしているが，実はわずか1項目（そのクヮルテット）にあらわれるのみであった．

4. 例外カタカナ語の語尾分析

さて，以上の3規則は，例外が事実上存在しないという意味ではきわめて優秀であるが，カタカナ語の認識率が 2/3 にも満たず，実用上多少の不満がある．もう一つ二つうまい規則を追加して，認識率をたとえば 80% まで高めることはできないものであろうか？

表 2　規則（1）～（3）の効き具合
Table 2　Detailed analysis of how the rules perform.

区　分	一般語	重要語	計
カタカナ語総数	5,162(100.0%)	296(100.0%)	5,458(100.0%)
認識カタカナ語数	2,958(57.3)	196(66.2)	3,154(57.8)
例外カタカナ語数	2,204(42.7)	100(33.8)	2,304(42.2)
累積長音則（1）	2,343(45.4)	167(56.4)	2,510(46.0)
累積パ行則（2）	773(15.0)	67(22.6)	840(15.4)
累積アイウエオ則(3)	469(9.1)	5(1.7)	474(8.7)
(1) のみ	1,778(34.4)	125(42.2)	1,903(34.9)
(2) のみ	387(7.5)	28(9.5)	415(7.6)
(3) のみ	202(3.9)	1(0.3)	203(3.7)
(1), (2)	324(6.3)	38(12.8)	362(6.6)
(2), (3)	26(0.5)	0(0.0)	26(0.5)
(3), (1)	205(4.0)	3(1.0)	208(3.8)
(1), (2), (3)	36(0.7)	1(0.3)	37(0.7)
長音則（1）	2,343(45.4)	167(56.4)	2,510(46.0)
それ以外のパ行則(2)	413(8.0)	28(9.5)	441(8.1)
それ以外のアイウエオ則（3）	202(3.9)	1(0.3)	203(3.7)

実際にこれらの規則を使ってすぐ目につくのは，認識されないカタカナ語の中に—shon, —ingu, —izumu などのような「外来語くさい」語尾をもつものが多いことである[2]．そこでこの種の語尾に着目した新規則を追加してはどうかとの考えが浮ぶ．たとえば上の3種の語尾をもつ語をカタカナ語として扱うことにすると，ナレーション，パッキング，ディレッタンティズムなどのようにすでに規則（1）～（3）でカバーされている語も多いので，性能の飛躍的向上は必ずしも望めないかも知れないが，マンション，キング，マゾヒズムなどのように，これで助かる例も少なくないので，ひょっとするとうまく行くかも知れず，この方向はたしかに検討の価値がある．

一つの検討方法として，例外カタカナ語にあらわれる語尾綴りを集め，それらのうちから，多数のカタカナ語にあらわれ，ひらがな語の語尾綴りとしてはあらわれるにしても稀にしかあらわれないようなものを拾い出すという手法が考えられる．そこで，例外カタカナ語全部について，ローマ字風表記の語尾の 2～6 文字をとり，その出現頻度を調べ，同じ語尾がひらがな語にあらわれる頻度とつき合わせてみた．

結果の一部を**図2**に示す．ひらがな語件数 20 件以内の項目のみ集め，例外カタカナ項目数の降順で整列した結果である．(実際の処理は例外カタカナ語数5以上の項目のみとっておこなった.)たとえば最初の部

```
kku     108   19    ..... / いっ く / がっ く / きくの せっく / けっ く / ご せっく / こっ
                      く / さっ く / しっ く / じっちゅう はっく / じゅうちゅう はっく / せっ く / ぜっ く /
                      とっく / はっ く / はつ ぜっく / ひなの せっく / ほっ く / ものの せっく / ろっ く /
ngu      78   15    ..... / あん ぐ / がん ぐ / がん ぐ / けん ぐ / ごん ぐ / さん ぐ /
                      しん ぐ / せん ぐ / そうしん ぐ / つり てんぐ / てん ぐ / ねん ぐ / ぶん ぐ / ぼん ぐ /
                      みん ぐ /
ingu     70    3    ..... / しん ぐ / そうしん ぐ / みん ぐ /
suto     68    3    ..... / づと / つと / つ と /
etto     55    4    ..... / しれっ と / すけ っと / でれっと / べっ と /
ia       46    1    ..... / せい あ /
zumu     43   12    ..... / うち しず・む / くれ なず・む / くろ ず・む / たたず・む /
                      なき しず・む / なず・む / ふし しず・む / **しず・む / *すず・む / **はず・む /
                      はなし・がはずむ / ひず・む /
ikku     42    2    ..... / いっ く / しっ く /
izumu    42    5    ..... / うち しず・む / なき しず・む / ふし しず・む / **しず・む /
                      ひず・む /
nsu      38   13    ..... / かん す / きん す / ぎん す / ござん・す / ごんす / さい
                      ざんす / さん・す / しゃん・す / せん す / ちゃ だんす / どん す / よう だんす / *たん す /
isuto    35    0    .....
akku     32    5    ..... / がっ く / さっ く / じっちゅう はっく / じゅうちゅう はっく /
                      はっ く /
iumu     30    0    .....
kusu     30   13    ..... / いい つく・す / おし かく・す / くす / しゅく・す / たち
                      つく・す / て つく・す / なめ つく・す / ぬり かく・す / ひしかく・す / **かく・す / *
                      じゅく・す / *つく・す / **なく・す /
kkusu    26    0    .....
```

図 2 例外カタカナ語に多い語尾

Fig. 2 Common endings in exceptional katakana words.

分は（ヘボン風）ローマ字表記した場合に kku で終る例外カタカナ語が108語，同じく kku で終るひらがな語が 19 語あったことを示す．またみつかったひらがな語全部が示してある．それらは一句，学区，菊の節句，結句，五節句，刻苦，作句，疾駆，十中八九，十中八九，節句，絶句，疾っく，八苦，初節句，雛の節句，発句，桃の節句，六区の計 19 語である．また少し下の方を見ると，語尾をもう1字前までとって kku の代りに ikku を考えた場合，例外カタカナ語は42語と半減するが，ひらがな語も一句と疾駆のわずか2語と，ほぼ10分の1になってしまうことがわかる．語例の先頭の＊は重要語，＊＊は最重要語をあらわす．

上記の発想に基づいて新しい規則を作るとすれば，これらの語尾綴り項目のうちから例外カタカナ語を多くもち，ひらがな語はなるべく少ないようなものを適当に拾い出し，「これこれの綴りで終るものはカタカナ語．ただしこれこれという例外あり」という形にまとめるのが自然である．特にひらがな語の数はごく少ないことが望ましい．既成3規則には非常に使いやすい，安心して使える，という感じがあったが，これは主として例外ひらがな語が図1に示すようにごく少なかったということに起因すると考えられるからである．具体的な選びかたは無数に考えられるがまぎれ込むひらがな語数が1語増すことは少なくともカバーされる例外カタカナ語数が 10 語増すことに匹敵する，と判断される．そこで語尾群の代表例として，$n=1, 2, 3, \cdots$ に対し，

「例外カタカナ語 $10n$ 語以上，ひらがな語 $n-1$ 語以下のものを全部拾う」

という形のものを考えてみることによって上述の方向でどのくらいのところまで行けそうかの見当をつけてみよう．実際，そのいずれかの段階で十分強力な判別規則が得られればよし，そうでなかったとすれば，少なくとも個別に語尾を拾い集めるという考えかたではうまく行きそうもないと結論することができよう．そう考えたとき，**表3**に示すような結果が得られる．

ただし表中の $n=4$ 項にかかげられている三つのひらがな語「つと」（語尾綴り suto に対応）はそれぞれ

ナットウをつとにくるむ
このことはつとに明らかである
彼女はつと立ち上った

のように使われるものである．一つ目はツトでもおかしくなく，三つ目は「つ と」とわかち書きすれば大丈夫であるが二つ目はこまる．これらは（ヘボン風の）ローマ字表記を基礎にとったために入ってきたもので，ストで終るもの，というとらえかたをすれば例外なし，ということになるが，その代り ingu でキングもリングも代表させる，というまとめかたはできなくなる．

こうまとめてみると，外来語尾に着目するという本章の考えかたは，ある程度の効果はあるが思ったほどすばらしい結果はもたらさない，ということがわかる．実際表3の規則群は，第1に総認識率があまり高くないことで，また第2に規則として必ずしもおぼえやすくないことで不満である．また認識率の高いものは複雑でおぼえにくく，おぼえやすいものは認識率がいっそう貧弱，という当然の傾向がある．

たとえば $n=3$ の規則などは中ではおぼえやすいものといってよいがカバーされる語数がわずか 153 語

で，アイウエオ則にもはるかに劣る．一方 $n=1$ のものは，認識率の点では（60% を 80% にするほどの力はないものの）パ行則にせまり，必ずしもわるくないが，おぼえなければならない語尾綴りが21種もあり，表3でも別記しなければならなかったほどであり，また別記の内容をよく見てみると利用者にたとえば「ria と nia があって bia がないのはなぜか」というような疑問を起させるような傾向があって，ますますおぼえにくい．（実は bia はひらがな項目0であるが，例外カタカナ項目が6個しかないのでリストアップされていない．）

もちろん表3のまとめかたには，いろいろなバリアントが考えられる．たとえば表3の段階1にかわって，「例外カタカナ語数20以上，ひらがな語数0」というものを考えると

「isuto, iumu, kkusu, shon, ringu で終るものはカタカナ語」

という規則が得られる．これならばまずはおぼえやすく，かつ何となく必然性を感じさせる面もあってなかなかよいが，カバーされる語数が141語（全カタカナ語の 2.6%）しかない．このように表3の構成が唯一の可能性ではないが，この方向で行ったときの全体的傾向は表3に示すところでほぼつきていると考えられる．

なお本章のはじめに例としてかかげた shon, ingu, izumu から成る組は例外カタカナ語を 24+70+42=136 しかカバーせず，しかもひらがな語を 0+3+5=8 語も持つので，あまり魅力的とはいいがたい．ション，イング，イズムなどはまさに外来語を象徴するような綴りであり，それがこの程度の成績にしか結びつかないというのは意外の感もあるが，たぶんこれは意外に感ずる方が錯覚なのであろう．上掲の既成3規則は，いずれも日本語の音韻に起源をもつ規則である．たとえば長音「ー」は本来日本語になかったものである．これに対しション，イングなどはたまたま外国から入ってきたためにそこにある，といったおもむきのものである．それがもともとの日本語に偶然含まれていなかったという保証はない．実際，寝具，民具などの意外な例外があって足をすくわれる．このように本来日本語（中国伝来のものを含む）にあらわれない音韻の組み合わせに着目する規則は安定性があるが，外来語に多い語形に着目する規則は不安定となるおそれがある，といえる．これはこの種の規則を作って行く上での基本的注意点として興味深い．

5. 検　討

以上のように，既成3規則（1）～（3）はおぼえやすくてその割にきわめて有効であり，まずは文句なしといってよいが，あとに続くものが問題である．前章の結果からみて，なるべく音韻に着目した方が賢明と思われる．実際そういう目で表3の結果を見なおしてみると，可能性の芽はたしかにある．たとえば，ia の46例，ea の16例，oa の12例をまとめると「井蛙」といういやな例外はあるものの，それだけで例外カタカナ語74語がカバーされる．一方，「イアやエアやオアはひらがな語ならイヤやエヤやオワになるのだから」という音韻に基づいた直観的説明ができる．このような可能性をさぐることは将来の課題である．

表3　外来語尾に着目した新規則の可能性

Table 3　Prospective rules based on foreign endings.

n	条件		あらたに拾われる語尾綴り（かっこ内例外カタカナ語）	持ち込まれるひらがな語	規則	認識例外カタカナ語数（百分比はカタカナ語全体との比率）
	例外カタカナ語数	ひらがな数語				
1	10 以上	0	［別記］		［別記］の21種の綴りで終る語はカタカナ語	347 語 (6.4%)
2	20 以上	1以下	ia (46) oppu (23)	井蛙 六腑	ia, shon, isuto, ringu, iumu, oppu, kkusu で終る語はカタカナ語，ただし，井蛙，六腑は例外	207 語 (3.8%)
3	30 以上	2以下	ikku (42)	一句，疾駆	ia, isuto, iumu, ikku で終る語はカタカナ語，ただし，井蛙，一句，疾駆は例外	153 語 (2.8%)
4	40 以上	3以下	suto (68) ingu (70)	つと，つと，つと {寝具，装身具，民具	ia, suto, iugu, ikku で終る語はカタカナ語，ただし，井蛙，つと，つと，つと，寝具，装身具，民具，一句，疾駆は例外	226 語 (4.1%)
5	50 以上	4以下	etto (66)	{しれっと，助っ人 でれっと，別途	etto, suto, ingu で終る語はカタカナ語，ただし，つと，つと，つと，寝具，装身具，民具，しれっと，助っ人，でれっと，別途は例外	204 語 (3.7%)
6	60 以上	5以下	以下新規項目なし			

［別記］ ea (16), nia (12), ria (15), oa (12), shon (24), ddo (18), mento (10), netto (12), auto (13), asuto (13), esuto (12), isuto (35), kingu (13), ringu (23), chikku (14), rikku (10), ukku (10), iumu (30), rizumu (18), resu (11), kkusu (26).

あ,痢,ア　合,相,愛**,間,藍,アイ　愛す,アイス

味**,鰺,アジ　圧砕,アッサイ　あど,アド

あな,穴**,アナ　亜麻,尼,蜑,阿娟,アマ

糠蝦,網**,アミ　安置,アンチ　暗譜,アンプ

い,矣,夷,医,易,威,胃*,異,意,緯,井,亥,寝,闇,イ

幾等**,イクラ　銀杏,医長,胃腸*,異朝,移牒,移調,イ調

図 3　カタカナ語とひらがな語の対立

Fig. 3　Collision between katakana and hiragana words.

もう一つの有力な可能性は adaptive katakanization とでもいうべき方法である．たとえば「システム」というカタカナ語が一度出て，シフトコードによってそれがカタカナ語であると知れたら，以後は特にシフトコードをつけないでも引き続きそれをカタカナ語として扱うようにするのである．この方法は規則(1)〜(3)だけによったときと比べれば計算資源を食うが，利用者から見てのわかりやすさという点ではほぼ同等と見られ，有望な方向といえる．

なお，前章のようにして新しい規則をいろいろ新設して行ったとしても，おのずと限界があり，意味に立ち入った処理をしないかぎりどこからか先はシフトコードにたよらざるを得ない．というのは「ほて・る」と「ホテル」のように，ひらがな語でもありながら同時にカタカナ語でもあるような綴りも存在するからである．そのほか，ひらがな語と混成語，およびカタカナ語と混成語の組み合わせもあり得る．そのような衝突がどれほど起るか，というのも興味深い問題である．「アジア」または「アメリカ」の略としての「ア」と感動詞の「あ」のような，ばかばかしいものを含めて，その総数は219組あった．図3にその種の対立の例を示す．このことが実際にどの程度問題になるかは，もっとくわしい検討（たとえば重要語「あい」と眼をあらわす「アイ」の対立などは，後者はめったと使われまいから，というので無視してよいかどうか，など）をしなければにわかに判断しがたいが，ともかくこの数字も一つの目安としては役立つかも知れない．

謝辞　この研究の主要なアイディアは著者のカーネギーメロン大学計算機科学科訪問中に得られた．本文はそれを帰国後発展させ，まとめたものである．快適かつ刺激的な研究環境に身を置く機会を与えられた同学科の関係者各位に深謝したい．また辞書ファイルの利用についてお世話になった電子技術総合研究所淵一博氏，著作権者山田忠雄主幹，および三省堂倉島節尚氏に謝意を表する．

参考文献

1) 金田一京助ほか編：新明解国語辞典第二版，三省堂，東京 (1972).
2) Izumi Kimura: Cheap Production of Japanese Documents—an Experiment in Programming Methodology, Carnegie-Mellon University, Department of Computer Science, CMU-CS-78-130 (1978)，および木村　泉：ゼロックス・グラフィックス・プリンタに日本語を教えた話，第20回プログラミングシンポジウム報告集, pp.139-156 (1979).
3) 日本工業規格：情報交換用漢字符号系，JIS-C 6226-1978，日本規格協会，東京 (1978).

(昭和56年5月21日受付)
(昭和56年10月7日採録)