福井大学 工学部 研究報告 第50巻 第2号 2002年9月
Mem. Fac. Eng. Fukui Univ., Vol. 50, No. 2 (September 2002)

日本語学習者のためのローマ字表記によるカタカナ語からの英単語検索

諏訪 いずみ* 小高 知宏** 小倉 久和**

Transliteration of Katakana into English Based on the Romanization of Japanese for Learners of Japanese

Izumi SUWA*, Tomohiro ODAKA** and Hisakazu OGURA**

(Received August 7, 2002)

It is difficult for learners of Japanese to understand meanings of katakana words, because pronunciation of loan words written by katakana differs from its original English pronunciation. This paper introduces a support system for understanding of katakana words using transliteration katakana into English based on the romanization of Japanese. Furthermore, effectiveness of the system was evaluated.

Key Words : Transliteration, Katakana, Romanization of Japanese

1. はじめに

日本語を母語としない者にとって，カタカナ語の意味を理解するのは困難がともなう．その一因として，カタカナ語の音がもとになった単語の発音と異なっていることが挙げられる．一方，専門用語等でのカタカナ語の使用頻度は高く，日常的にも，カタカナ語が使用される機会が多くなっており，外来語辞書にないものも増えている．現在使用されているカタカナ語の約80%は英語起源であるといわれ，カタカナ語は，カタカナ英語と言いかえてもよいほどである[1].

一般的に，英語をカタカナ表示する場，二つの方法がある．一つは英語の発音をもとに，最も近い音を割り当てるという方法である．しかし，日本語は英語と比較して母音が少なく子音も異なるため，正確に英語の音を写すことができない．もう一つは，本来の英語発音ではなく，英語表記のローマ字読みを割り当てる方法である．実際には，これらの混用によって，カタカナ語を生成している．このため，カタカナ語の音は，もとの単語の音とかなり異なるのが普通であり，複数の読みが存在することも珍しくない．これらのことが，日本語を母語としない者のカタカナ語の理解を困難にしている．

このように，カタカナ語において英語起源のカタカナ語からもとの英単語を検索するシステムは，カタカナ語理解を支援すると思われる．そこで，日本語学習者を想定しての入力しやすさ，システムとしての使いやすさ，簡潔性などを考慮した支援システムの検討を行った．前論文[2]では，検討結果をもとに，ローマ字表記からもとの英単語を検索してカタカナ語理解を支援するシステムを試作し，その有効性を評価した．本論文では，前論文で述べられなかったシステムの詳細について報告する．

2．ローマ字表記による検索

従来の方法では，検索にカタカナ表記を使用したものが多い[3],[4]．カタカナ表記は日本人にとってはなじみやすいものである．しかし，本研究では，日本語学習者を想定していることと，子音と母音が明示的に表示されるローマ字表記の特徴により，入力にローマ字表記を採用した．

* 工学研究科システム設計工学専攻
** 知能システム工学科
* System Design Engineering Course, Graduate School of Engineering
** Dept. of Human and Artificial Intelligent Systems

ローマ字以外の文字で書かれたものをローマ字で表記することは国際的な理解のために一般に行われており，日本語のローマ字表記に関する国際規格もある．また，日本語の学習時に，読みをローマ字で表記することも一般に行われている．したがって，日本語を母語としない使用者には，カタカナでの入力よりも，ローマ字表記による入力のほうがなじみやすいと思われる [5].

さらに，日本語を入力する場合ローマ字入力が標準であるシステムが増えており，システムによってはカタカナでの入力をするために，日本語の入力システムが必要なものもある．ローマ字表記を用いると，日本語の入力システムは基本的には必要なく，システムが簡便になる．

また，ローマ字表記の特徴として，子音と母音が明示的に表記されるということがある．これにより，子音と母音を分けて処理をすることができ，二重母音の処理等の変換を効率よく行うことができる．その結果，カタカナ表記から直接検索する場合よりも，変換のための規則数が少なく済み，効率よく検索をすることができる．

ローマ字表記にはヘボン式と訓令式の二つがあるが，ここでの入力はヘボン式を基本とし訓令式でも可とした．ヘボン式は日本語のローマ字表記と英語式の発音の関係を意識してつくられている．そのため英語起源のカタカナ語をローマ字表記する場合，ヘボン式のほうが入力しやすく処理もしやすいからである．国際規格としては，ISO3602:1989があるが，訓令式を基礎としているため採用しなかった．ヘボン式表記の詳細については，文献 [5] によるものを使用した．これは，英国規格を基礎としている．文献 [5] にない表記については，英国規格を参考とした．ローマ字表記とカタカナ表記の相互変換は容易であり，検討したシステムでは，入力されたローマ字表記について入力文字列の確認のため，別途変換を行ってカタカナ表記も表示した．

3 システム

本システムの構成を図 1 に示す．ローマ字表記で入力されたカタカナ語は，「ローマ字表記－英語変換規則」を適用するための「中間表記」に変換される．「変換規則処理系」では，中間表記に対して先頭から順に変換規則を適用し，候補となる英単語を英語辞書から検索していく．カタカナ語 “フィジックス” に対応する英単語の検索を例として，システム各部における処理，及び表記の詳細について以下に述べる．

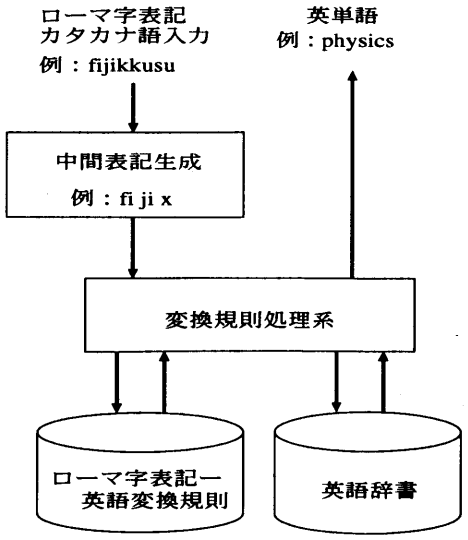

図 1: システムの構成

3.1 ローマ字表記入力

入力表記についてはヘボン式を基本としているが，キーボードからの入力であることから，以下の点で手書きの場合の表記と異なる表記を用いた．

長音に関しては，ローマ字表記の標準的な表記法では入力が困難なので，母音字の連続（“aa” 等）で表記するようにした．撥音についてはすべて “nn” とした．促音については，ヘボン式では，“ c” の場合のみ “cc” ではなく “tc” とするが，すべて直後の子音字の連続（“タッチ・tacchi” 等）とした．

ヘボン式といっても，表記の詳細については，異表記が存在するので，詳細については，文献 [5] によるものを基本表記とした．“チ” は “chi”，“ジ” は “ji”，“ズ” は “zu” のみとし，“ティ” を “ti”，“ディ” を “di”，“ドゥ” を “du” とした．基本表記と競合しない範囲で，訓令式の表記も認識するようにした．たとえば，“シ” は “ shi” と “si” のどちらでも入力可能である．これら注意の必要な表記に関しては，指針として使用時に表示するようにしている．

検索語 “フィジックス” はローマ字表記で “fijikkusu” と入力する．

3.2 中間表記生成

「中間表記生成」では，入力されたローマ字表記をローマ字表記－英語変換規則を適用するための中間表記に変換する．中間表記は，“so fu to” のように変換規則を適用するまとまりごとにスペースで区切られている．スペースで区切られた中間表記の1単位は，基本的には日本語の1音節に対応している．ただし，二重母音，促音，及び二音節をひとまとまりとする表記については，ローマ字表記－英語変換規則を効率よく適用できるように音節とは異なる区切りの中間表記に変換する．

二重母音とは，ローマ字表記で “ai”,“ei” など，異なる母音字の連続で表記されるものである．大部分は英語発音での二重母音に対応する．中間表記に変換する際には，母音部分をまとめて，“kai” → “k ai” のようにする．英語での二重母音ではないが，“ea” もこの分類にいれた．

促音については，二つの中間表記を生成している．一つは入力確認用のカタカナ表示を生成するための表記で，促音を “xtu” とし，たとえば “hatto(ハット)” → “ha xtu to” としている．もう一つは，実際にローマ字表記－英語変換規則を適用する表記で，促音を示す “xtu” を削除し，“ha to” としている．これは，日本語学習者にとって促音の有無の判定が難しいので，促音を入力しなくても候補を提示できるようにするためである．促音の問題については，考察において詳しく述べる．

二音節をひとまとまりとする表記は，“ku su”，“ku shi” など，主に元の英単語の綴が “x” を含むものである．これらは，中間表現として “x”，“xi” など空白を含まない綴に変換される．

入力例の “fijikkusu” は，促音の処理と二音節をひとまとまりとする表記の処理の結果，中間表記 “fi ji x” に変換される．

3.3 ローマ字表記—英語変換規則

「ローマ字表記－英語変換規則」は，スペースで区切られた中間表記の一単位とその一単位に対応する英語の文字列の対である．規則表に記述された規則数は 266 である．これはヘボン式ローマ字表記に訓令式表記を加えたものが 127，それぞれの長音表記が 127，二重母音に関する変換規則などの特殊な変換規則が 12 である．単語の語頭と語尾にはそこにだけ現れる綴が存在するので，変換効率を上げるため，変換規則表は語頭用，語の中間部用，語尾用の3種を用意した．

表 1: 変換規則

a	(a,u,o,er,e,re)
aa	(ar,are,er,ear,ur)
ka	(cha,ca,cou,co,cu,ka)
kaa	(ker,car,cker,cur,cor,cal)
fi	(fe,fi,phi,phy)
fii	(fee,fea,fie)
ji	(di,ge,gi,ji,si,zi)
jii	(ge,gy,jee,sy)
zi	(di,ge,gi,ji,si,zi)
zii	(ge,gy,jee,sy)
x	(x,cs)

変換規則の作成は人手により行った．まず，外来語辞書などを参考に，ローマ字表記の一音節と対応すると思われる英語綴を抽出した．これを最初の変換規則として，本システムでローマ字表記カタカナ語に対応する英単語の検索を行い，不足する変換規則を追加していった．変換規則の抽出については，変換速度や精度を考慮し，よく使用される変換規則に含まれる綴の数が極端に多くならないよう考慮した．そのため，英語の音節とは異なる分割をしたものもある．たとえば，“a-ja-su-to（アジャスト)” の場合，もとの英単語の音節としては，“ad-just” であるが，“a” に対応する綴を増やさないために，変換規則抽出のための分割としては，“a-dju-s-t” とした．

表 1 に変換規則の例を示す．左側が中間表記の一単位，右側が対応する英語綴である．“fijikkusu(フィジックス)” の中間表記である “fi ji x” の “fi”，“ji”，“x” に対する変換規則が含まれている．

3.4 英語辞書

「英語辞書」は，検索用インデックスと英単語の組である．検索用インデックスは,「変換規則処理系」で行う検索のための英語綴である．検索を容易にするために，rr，ll など文字の重複を検索に問題がない範囲で一文字にしてある．実際の辞書はフリーで使用することのできるものを前述のように加工して用いた．語数は約 24 万語である．表 2 に英語辞書の例を示す．

表 2: 英語辞書

answer,answer
borow,borrow
cofee,coffee
cuting,cutting
folow,follow
physics,physics
succes,success

3.5 変換規則処理系

「変換規則処理系」では，中間表記に対して，先頭から区切りごとに順次ローマ字表記－英語変換規則を適用して英語辞書を検索し，候補となる英単語を絞っていく．図 2 に “fijikkusu(フィジックス)” の中間表記 “fi ji x” に対する検索アルゴリズムの適用例を示す．

まず，最初の変換規則に一致する検索用インデックスを持つ英単語を辞書から抽出する．この集合内の英単語の検索用インデックスから検索に使用した規則の綴を削除する．次のローマ字表記－英語変換規則を新しい検索用インデックスに適用し，対応する綴を持った英単語だけを残す．集合内の検索用インデックスと英単語の組に対し，この手続をくり返し適用する．すべての変換規則の適用が終了した時点で，検索用インデックスが空であるものが，完全一致で候補となる英単語である．完全一致するものがなかった場合，検索に失敗する直前に集合内に残っていた英単語を候補として出力する．ただし，ローマ字表記での音節数が 2 以下のものについては，変換規則から生成される単語綴すべてについて完全一致検索を行う．たとえば，“puuru(プール)” の場合，変換規則 puu(poo,pu), ru(le,l,re) から生成することのできる全ての綴，poole，pool，poore，pule，pul，pore について完全一致検索を行う．これは，音節数の少ない単語では，検索に失敗した場合に多数の不適な候補を出力する可能性があるからである．

本システムで，実際にカタカナ語 “フィジックス” を検索した結果を図 3 に示す．ローマ字表記で “fijikkusu(フィジックス)” と入力すると，カタカナ表示とともに，“フィジックス” のもとの英単語の候補を表示する．

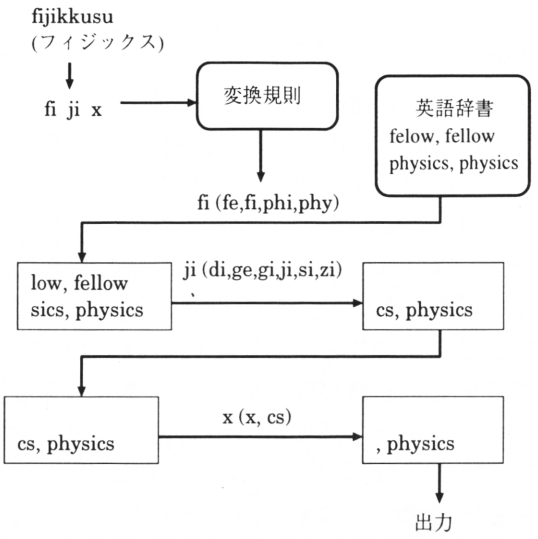

図 2: 検索アルゴリズム

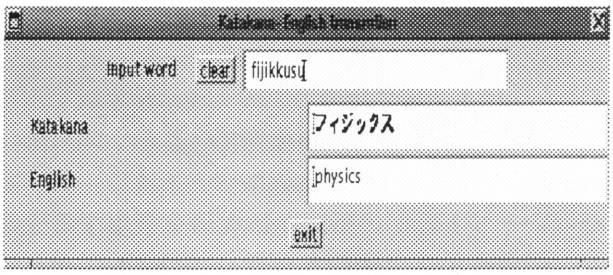

図 3: 検索例

4 評 価

和製英語，短縮語，空白やハイフンで区切られた複合語，固有名詞ではない英語起源のカタカナ語について変換率を評価した．システム入力となるカタカナ語は，テキストベースのフリーの和英辞書 EDICT のカタカナ語見出しのものを使用した．カタカナ語見出し 12233 語のうち，上記条件を充たすものは 7119 語あり，その中から 1463 語を無作為抽出した．この抽出したカタカナ語について，カタカナ表記をローマ字表記に自動変換して入力として使用した．評価した結果を表 3 に示す．また，評価結果の各項目の例を表 4 にしめす．

正しい候補とは，評価に使用した辞書に記載されている元の英単語と検索結果が一致するものをいう．正しい候補を含むとは，正しい候補以外に，似た音の単語が出力されるものをいう．正しい候補のみは全体の 53.1 %，正しい候補を含むものは 24.7 %であり，77.8 %において，正しい候補を提示できた．正しい候補の一部とは，検索語の派生

表 3: 変換率評価結果

全語数	1463 語	—
正しい候補のみ	777 語	53.1 %
正しい候補を含む	362 語	24.7 %
正しい候補の一部	108 語	7.4 %
不適な候補	204 語	13.9 %

語や，複合語である場合にはその一部などをいい，7.4 %を占めた．不適な候補とは，明らかに間違った候補や，検索が失敗した時点で 20 以上候補が残ったもの，候補が存在しなかったものである．

さらに，日本語を母語としない人の評価を得るため，研究室に在籍する中国人留学生に実際に使用してもらった．評価としては，カタカナでの入力よりもローマ字での入力のほうが，使いやすいということであった．これは，日本語の読みを学習する際にローマ字表記を用いるからだそうである．また，中国語でも読みを表記する場合にピンインというアルファベット表記を使用するので，ローマ字表記はなじみやすいとのことである．

中国語を母語とする人の場合，判別が難しい音として，促音（例：“ハット” か “ハト” か）がある．促音については，このシステムでは促音なしでも正しい単語を候補としてあげるので，使いやすいという評価を受けた．複数の候補については，各候補に順位づけや使用可能性を表示してもらえるとわかりやすいとのことであった．

5 考 察

入力表記にローマ字表記を用いた簡潔なシステムで 78 %程度の単語について正しい候補が得られた．また，7 %については，正しい候補は得られなかったが，検索語の派生語や，複合語である場合にはその一部など，候補を類推できるような結果が得られた．短い検索語については，外来語辞書にあるものと同時に，類似の発音をする他の単語が出力される場合が多い．明らかに誤った候補が出力された場合の原因としては，以下のような問題が考えられる．

(1) 促音の問題

短い検索語について，正しい候補とともに明らかに誤った候補が出力される主な原因として，促音の問題がある．促音は，英語をカタカナで表記する場合，日本語としての語調の問題で，閉鎖音，破裂音，摩擦音の前で，かつ単母音の後に挿入される場合が多い．しかし，日本語学習者には，促音が入るか入らないかの区別が困難である．また，綴の対応による変換規則では，変換規則として促音を考慮する必要はないと考えられる．そこで，ここでは促音表示を取り除いたものに変換規則を適用している．促音の削除は，長い綴では問題ないが，二音節程度の短い綴では，不要な候補を挙げる原因となる（例:hatto → hat, hot, hut, hate）．二音節程度の単語では，促音が入るのは最後が子音字で終わるケースが多い．また，促音が省かれた場合（例：hato）でも，促音がはいるケースかどうか判定は可能である．これにより，例にあげた場合では，“hate” が除外される．このように，促音の問題は，入力パターンによる選別規則を作成し，候補の単語を選別することで解決できると思われる．

表 4: 評価結果の各項目における例

例 1：正しい候補のみ

カタカナ語	dainamikusu(ダイナミクス)
和英辞書 EDICT	dynamics
検索結果	dynamics

例 2：正しい候補を含む

カタカナ語	hatto(ハット)
和英辞書 EDICT	hat
検索結果	hat,hut,hot,hate

例 3：正しい候補の一部

カタカナ語	aisoreeta(アイソレータ)
和英辞書 EDICT	isolater
検索結果	isolate

例 4：不適な候補

カタカナ語	arukooru(アルコール)
和英辞書 EDICT	alcohol
検索結果	alcalde,alcaldia

(2) 変換規則の問題

不適な候補については，原因の一つとして変換

規則に必要な綴が存在しなかった場合がある．本システムの場合，変換規則に綴を記載しておけば，検索可能である．変換規則の抽出は，人手で行っているので，現時点では，まだ，洩れている綴が存在する．稀にしか出現しない綴に関しては，あえて記載しなかった．稀にしか出現しない綴もすべて記載しておくと，変換に使用する綴の数が多くなり，検索に時間がかかるようになるからである．綴をどこまで記載するかは，変換速度を考え決定する必要がある．

(3) 辞書の問題

不適な候補のもう一つの原因としては，正しい候補を生成する規則がありながら，英語辞書に単語が存在しなかった場合がある．正しい候補の一部を出力する場合も辞書に適切な単語がなかったことによる．このケースに当てはまる単語には専門用語が多い．従って，検索に使用するテキストベースの英語辞書を選ぶことにより，解決できると思われる．本システムの手法は，カタカナ語に対応する英単語がテキストとして存在すれば，高い確率で候補を得ることができる．従って，辞書に登録されていない専門用語などについては，辞書の代わりに対応分野の英論文テキストを直接検索し，候補を得るようにすることが可能であると思われる．

これまでの研究では，カタカナ語から英語表記の候補を推定する方法として，カタカナ表記をもとにしたものが多い．カタカナ表記から直接原表記の候補を推定する方法 [3] や，発音記号をもとに，カタカナ表記と英語表記を対応づける方法 [4] が提案されている．しかし，前者のような方法は，規則数が多くなる（900 程度）という問題があり，後者の方法では，システムが複雑になる．ローマ字表記の場合，子音と母音が明示的に表記されるので，カタカナ表記をした場合よりも二重母音の処理等の変換を効率よく行うことができ，変換のための規則数が少なくてよい．われわれのシステムでの基本的な規則数は 270 程度である．また，発音記号表記については，カタカナ表記では，本来の英語発音ではなく，英語表記のローマ字読みを割り当てる場合もある．ローマ字読みからできたと思われるカタカナ表記に対しては，発音記号を用いる方法では，対応が困難である．従って，発音記号を用いるメリットは必ずしもないと思われる．候補中に正解を含む場合の変換率は，文献 [3] が 79.4 %，文献 [4] が 82.3 % である．評価方法がそれぞれ多少異なるため，単純な比較はできないが，我々の提案システムも，ほぼ同程度の変換率となっている．

ローマ字表記されたカタカナ語を用いたものとしては，カタカナ語と対応する英単語の組を自動収集する研究が，Brill ら [6] によって報告されている．これは，ローマ字表記されたカタカナ語をもとの英単語のミススペルとみなして対応を学習させ，検索エンジンで収集したログからカタカナ語と対応する英単語の組を自動収集するものである．任意のカタカナ語の変換率としては 70 %以下である．もともとカタカナ語は英単語の音表記の性格が強いので，綴としての対応には限界があると思われる．本論文では，基本的な変換は音節単位であり，変換規則としては我々の手法のほうが精度が高いと思われる．

6 今後の課題

現時点では，簡素な GUI インターフェイスからの検索か，テキストベースのコマンドラインからの検索となっている．実際に広く使用してもらうために，web での使用を検討し，CGI を用いたインターフェイスを構築中である．また，検索方法を改善し，より速く検索できるようにすることも検討している．

和英辞書 EDICT に含まれるカタカナ語で，英語起源と思われる語で固有名詞でも短縮語でもないものは 10095 語である．この内の約 30 % にあたる 2976 が空白やハイフンを含む複合語である．カタカナ語を検索する場合，こういった複合語を検索できることも必要であろう．

空白やハイフンを含む複合語を多数収録する英語辞書があれば，このシステムで検索可能である．しかし，今回使用した和英辞書 EDICT に含まれるこのタイプの複合語について一部を調べたところ，個々の単語は英語でありながら，複合語としては，和製英語と思われるものの比率が高かった．そのため，複合語と思われるカタカナ語を検索できるためには，単語の切れ目を検出し，前後の単語を別々に検索するなどの工夫が必要である．

7 おわりに

検索にローマ字表記を用いることにより，実装・使用が簡単なシステムを実現し，日本語学習者のカタカナ語理解の支援として有用であることを示した．また，このシステムは，カナ発音から英単語を検索する辞書として利用することもでき，英語学習の支援としても使用できると思われる [7].

謝 辞

本研究を行うにあたり，工学部知能システム工学科の黒岩丈介教官と高橋勇教官に多くの助言をいただいた．また，大学院工学研究科情報工学専攻の高建斌氏と工学部知能システム工学科研究生の潘維国氏には中国人留学生としてシステムの評価に協力いただいた．以上の方々に厚くお礼申し上げる．

参考文献

[1] 野角幸子，日本社会にあふれるカタカナ語，新風舎，1998.

[2] 諏訪いずみ，西野順二，小高知宏，小倉久和，“日本語学習者のためのローマ字表記に基づいたカタカナ語からの英単語検索の試み”，電子情報通信学会論文誌．(in press)

[3] 野美山浩，“カタカナ外来語の表記の揺れの解消”，情報処理学会第 41 回全国大会，3 分冊，pp.191 − 192，1990.

[4] 宮内忠信，“カタカナ表記からの英単語検索システムの実現”，情報処理学会自然言語処理研究会報告，no.97，pp.119 − 126，1993.

[5] 外国人のためのローマ字英和・和英辞典，三省堂，1999.

[6] E.Brill，G.Kacmarcik and C.Brockett，“Automatically Harvesting Katakana-English Term Pairs from Search Engine Query Logs”，Proc. the Sixth Natural Language Processing Pacific Rim Symposium，pp.393 − 399，Tokyo，Japan，November，2001.

[7] カナ発音現代英和・和英辞典，三省堂，1999.

[8] K.Knight and J.Graehl，“Machine Transliteration”，Association for Computational Linguistics，vol.24，no.4，pp.599 − 612，1997.