変形ルールと禁則ルールを用いた片仮名の表記ゆらぎの解消法

飯田敏幸† 中村行宏††

外来語はローマ字読みや発音を元にした読みを片仮名で表されるため，語によっては多くの表記ができてしまう．文字列から文字列への変形ルールを使った表記ゆらぎ解消方法では，変形の過程で生成される表記数が非常に膨大になる．そこで，あり得ない文字列を禁則ルールとして利用することによりこの問題を解決できることを示す．EDR 辞書インタフェースを使用して評価した結果，93個の変形ルールと4種類の禁則ルールを用いることにより，①辞書中の片仮名の登録語を1/3に削減できること，②変形ルールの適用は平均1.1回ですむこと，③変形の過程で生成される表記は最大262，平均4.9であることがわかり，本方式の有効性が立証できた．

A Method to Accept *Katakana* Variants

Toshiyuki Iida† and Yukihiro Nakamura††

A word of foreign origin is spelled in various ways in Japanese *katakana* based on its pronunciation or its *romaji* reading. To reduce vocabulary words in an electronic dictionary, whose origins are the same, rules to convert a sequence of *katakana* characters into another one would make tens of thousands variants. This paper shows that rules not to make impossible sequences can solve this porblem. Evaluation of the method with 93 conversion rules and 4 types of prohibition rules results in 1) the number of vocabulary words in *katakana* whose origins are the same can be reduced to 1/3, 2) the number of conversion rules applied averages only 1.1, and 3) the number of sequences produced in conversion process is 262 in the maxium and 4.9 on the average.

1. はじめに

近年，辞書の電子化が盛んに行われ，書籍を CD-ROM 化した辞書が利用できるようになっただけでなく，日英機械翻訳のための辞書[1]や大規模知識ベースとしての電子化辞書[2]が発表されている．書籍や CD-ROM の辞書を利用するときには，調べたい文字列と見出し語とが完全に一致していなくとも，人間が適当に変形することにより所望の見出し語を引き当てることができる．しかし，自然言語処理システムでは，人間のように適当な変形ができないために，調べたい文字列を電子化辞書から探すことができない可能性がある．辞書中の見出し語と一致しない場合には，その文字列は未知語として扱われてしまう．そこで，通常は辞書中にいくつかの異なる表記を持たせることにより，上記問題点を解決しようとしている．

外来語はそのローマ字読みを片仮名で表記したり，発音を片仮名で表記する*．このため，単語によっては非常にたくさんの表記ができてしまう．例えば，英語の“processor”は“プロセサ”，“プロセッサ”，“プロセサー”，“プロセッサー”，“プロセサア”，“プロセッサア”の6通りの表記が可能である．書籍の辞書では，通常は片仮名の表記を統一し，凡例や編集方針の中で表記法を定めている．また，官報中には，『外来語の表記』[3]をよりどころとするようにとの記述がある．このように標準的な表記を定めても，文章を書くときには必ずしも標準的な表記の片仮名を使うとは限らない．従って，不特定多数の人の書いた日本語の文書を扱うシステムにおいては，このような表記上のゆらぎに対処できる必要がある．そこで，可能なすべての表記を辞書中に持たせようとすると，辞書は非常に膨大になってしまう．単語の意味を扱う辞書では複数の単語のつながった複合語を見出し語として登録する必要があるために，このような辞書ではさらに表記数が増えてしまう．これは，複合語の表記ゆらぎの数がその複合語を構成する要素の表記ゆらぎの数の積とな

† NTT コミュニケーション科学研究所
NTT Communication Science Laboratories
†† NTT 情報通信網研究所
NTT Network Information Systems Laboratories

* 片仮名で表記される単語は外来語だけではなく，和語にもあるが（例えば，サケ），外来語に比べれば表記のゆらぎも少なく，余り問題にはならない．本論文では和語，外来語の区別は特に問題にはしていない．

るからである．例えば，“user interface” という複合語に関しては，“user” は “ユーザ”，“ユーザー”，“ユーザア” の3通り，“interface” は “インタフェース”，“インターフェース”，“インタフェイス”，“インターフェイス” の4通りの表記が可能である．さらに，単語の区切りを入れるか否かの2通りがあり，この1つの複合語に対し，24(3×4×2) 通りのバリエーションができてしまう．後述の EDR 辞書インタフェース* には，1つの英語の見出しに対応する片仮名の見出しとして最大256の表記が登録されている．このように可能なすべての表記を漏れなく列挙することも不可能である．

そこで，文字列と文字列の置き換え規則（以下，変形ルールと呼ぶ）を用いて，派生する表記を作成し，辞書中の見出し語と突き合わせる方法が提案されている[4),5)]．この方法では，元となる語の長さが短い場合には，生成される表記の数が少ないが，長さが長くなるとルールの適用回数が多くなるために，生成される表記数は非常に膨大になってしまう．また，英単語の発音記号から片仮名表記を自動的に生成し，表記のゆれを吸収する方法が提案されている[6)]．この方法では，英語以外の言語を元とする片仮名表記もあること，ローマ字読みの片仮名表記があることから，この問題の解決にはならない．

本論文では，変形ルールの適用に当たり，変形の過程であり得ない文字列が現れないようにするために禁則ルールを用いることにより，生成される表記の数を減らす方式を提案する．以下，2章では，片仮名の表記ゆらぎ解消のための考え方について，3章では変形ルールと，禁則ルールについて，4章では EDR 辞書インタフェースを用いた本方式の評価について述べる．

2. 片仮名表記

2.1 片仮名の定義

本論文では片仮名を以下のように定義する．

【定義】

片仮名 : :=JIS 漢字コード表のカタカナ（1 進数で 2521〜2576）|長音|区切り記号

長音 : :=ー

区切り記号 : :=・|=|… □

$K=k_0k_1\cdots k_n$ を長さ $n+1$ ($n\geqq 0$) の片仮名表記と呼ぶ．ただし，k_i ($0\leqq i\leqq n$) は上記片仮名の1文字である．

2.2 表記ゆらぎ解消のための考え方

片仮名表記 K_0 と K_1 とが表記のゆらぎの関係にあることは次のように定義できる．

【定義】

2つの表記が同じ単語（外来語であれば，同一のアルファベットの並び）を起源として持ち，K_0 を構成する一部の連続する文字列を有限回変形することにより，K_1 に一致させることが可能である． □

起源となる単語wに対し，可能な表記の集合を

$$\Omega(w)=\{K_0, K_1, K_2, \cdots, K_n\}$$

とする．このとき，すべての i, j ($0\leqq i\neq j\leqq n$) の組に対し K_i と K_j とは表記のゆらぎの関係にある．辞書の見出し語は少ない方がよいので，起源となる単語wに対し1つの表記 $K\in\Omega(w)$ を選び，これを辞書に登録し，K以外の $\Omega(w)$ の要素の表記は変形ルールによりKに変形することを考える．$\Omega(w)$ から辞書に登録する表記Kは変形ルールの適用回数（距離と呼ぶ）を最少にするように，

$$\sum_i d(K_i, K)=\min_j \sum_i d(K_i, K_j)$$

を満たすようなKを選ぶ．ただし，$d(K_i, K)$ は K_i からKへの距離である．以後，このようにして作られた辞書を標準辞書と呼ぶことにする．当然のことであるが，標準辞書を作るには変形ルールの充実が必須である．

3. 変形ルールと禁則ルール

3.1 変形ルール

前述の官報の『外来語の表記』の変形ルールを整理すると，**表1**に示す 64 種類のルールが抽出できる．しかし，例えば，“フロッピィ” から “プロッピイ” への変形ができないので，このルールでは不十分であることは容易に分かる．そこで，このルールを元にして，以下の一般化，連想を行うことによりルールの集約と拡張を行う．

（1） ルールの一般化

例えば，表1のルール 23，29，52 から，

R1：ァ ←→ ア

を導く．

（2） ルールからの連想

例えば，表1のルール 16 と 17 からルール 18 のほかに，

R2：ギ ←→ グィ

---

* 株式会社日本電子化辞書研究所の EDR 電子化辞書インタフェース第1版

表 1 官報による変形ルール
Table 1 Conversion rules derived from *Kanpo* (the official gazette).

No	ルール内容	No	ルール内容	No	ルール内容
1	aア ←→aー	23	クァ ←→クア	45	ド ←→ドゥ
2	iア ←→iヤ	24	クィ ←→クイ	46	ハ ←→ファ
3	iイ ←→iー	25	クェ ←→クエ	47	バ ←→ヴァ
4	iウム←→iュームム	26	クェ ←→ケ	48	ヒ ←→フィ
5	uウ ←→uー	27	クォ ←→クオ	49	ヒュ ←→フュ
6	eア ←→eヤ	28	クォ ←→コ	50	ビ ←→ヴィ
7	eエ ←→eー	29	グァ ←→グア	51	ビュ ←→ヴュ
8	oオ ←→oー	30	グァ ←→グゥ	52	ファ ←→フア
9	イ ←→ウィ	31	シア ←→シャ	53	フィ ←→フイ
10	イェ ←→エ	32	シェ ←→セ	54	フェ ←→フエ
11	ウィ ←→ウイ	33	ジ ←→ディ	55	フェ ←→ヘ
12	ウェ ←→ウエ	34	ジェ ←→ゼ	56	フォ ←→ホ
13	ウォ ←→ウオ	35	ジュ ←→デュ	57	ブ ←→ヴ
14	エイ ←→エー	36	ズ ←→ドゥ	58	ベ ←→ヴェ
15	オウ ←→オー	37	チ ←→ツィ	59	ボ ←→ヴォ
16	カ ←→クァ	38	チ ←→ティ	60	ム ←→ン
17	ガ ←→グァ	39	チュ ←→テュ	61	ンn ←→n
18	キ ←→クィ	40	ッ ←→×	62	ンm ←→m
19	キサ ←→クサ	41	ツ ←→トゥ	63	ー ←→×
20	キシ ←→クシ	42	テ ←→ティ	64	区切り記号←→×
21	キス ←→クス	43	デ ←→ディ		
22	キソ ←→クソ	44	ト ←→トゥ		

注：a，i，u，e，oはそれぞれア列，イ列，ウ列，エ列，オ列の片仮名文字を意味する．例えば，1のルールを右方向に適用することにより，“プロセサア”から“プロセサー”へ，逆に左方向に適用することにより“プロセサー”から“プロセサア”に変形できる．
×は空文字の意味である．
n，mはそれぞれナ行，マ行の片仮名文字を意味する．

のルールを連想する．ルールの一般化によって得られた R1 からさらに

R3：ィ ←→ イ

のルールを連想することにより，“フロッピィ”から“プロッピイ”への変形も可能となる．

ルールを追加していくと，ルールが重複する可能性が生じる．例えば，表1には

ジ ←→ デ

がないが，ルール 33 と 43 からこのルールを導くことができる．上記ルールを追加すれば，距離は小さくなるが，変形ルールが多くなる．そこで，いくつかのルールから導けるようなルール（重複ルールと呼ぶ）があれば，これを削除する．

3.2 禁則ルール

上記変形ルールを適用すると，ありえない文字列が生成される可能性がある．例えば，“プロセッサー”に表1のルール 40 を左向きに適用すると“プロセッッサー”が生成されてしまう．そこで，隣接する文字の並びの中でありえない組合せを禁則ルールとして持ち，禁則ルールに合致する文字列が生成された場合には，その文字列を使った以降の変形を中止するようにする．

区切り記号を1種類とすると，片仮名の種類は 88 あり，隣接する文字の組合せは 7744(=88^2) 通りになる．例えば，EDR 辞書インタフェースではこのうち 3008 通りの組合せが存在しない．この存在しない組合せを禁則ルールとして登録しても，①変形ルールで生成されない文字列がほとんどであること，②緩い禁則ルールでもかなり無駄が省けることから，以下のように直感的にありえない文字列の条件のみを選び，禁則ルールとした．

小文字 ::=ァ|ィ|ゥ|ェ|ォ|ャ|ュ|ョ|ヮ|カ|ケ

促音 ::=ッ

撥音 ::=ン

とする．長さ $n+1$ の片仮名表記 $K=k_0k_1\cdots k_n$ に対し，禁則ルールは次の4種類にまとめられる．

【禁則ルール1】

k_0＝小文字，促音，撥音，長音，区切り記号

【禁則ルール2】

k_n＝促音，区切り記号

【禁則ルール3】

k_i＝小文字 $(0<i\leqq n)$

のとき，

k_{i-1}＝capital(k_i)

ただし，capital(x) は小文字 x に対応する大文字（例えば，capital(ィ)＝イ）

である．

【禁則ルール4】

表2に示す接続関係 □

例えば，“プロセッサ”に表1のルール 40 を左向きに1回だけ適用すると，“ップロセッサ”，“プッロセッサ”，“プロッセッサ”，“プロセッッサ”，“プロセッ

表 2 隣接禁止文字
Table 2 *Katakana* characters not to be adjacent each other.

w_i \ w_{i+1}	小文字	促音	撥音	長音	区切り記号
小文字	×				
促音	×	×			
撥音	×	×	×		
長音	×			×	
区切り記号	×	×	×	×	×

注：× は隣接できないことを示す．

サッ”が得られるが，禁則ルール1と4により“ップロセッサ”と“プロセッッサ”を除くことができる．

3.3 変形ルールと禁則ルールを用いたゆらぎ解消アルゴリズム

（1） アルゴリズム

与えられた片仮名表記 K_g から標準辞書に登録されている対応する片仮名表記Kに変形するアルゴリズムを示す．

2.2節で述べたように，Kは表記ゆらぎの関係にある表記の中で，距離が最小のものを選んだので，変形を繰り返し行うのではなく，各部分文字列に対して均等に変形ルールを適用する方法をとる．このため，変形の過程で得られた文字列をスタックに格納し，すべての部分文字列に適用できる変形ルールがなくなった時点で，スタック中の次の文字列に対して変形を行う．具体的なアルゴリズムは次のとおりである．

【step 1】 K_g を調査対象の文字列Wにコピーする．
【step 2】 Wの先頭文字に着目する．
【step 3】 変形ルールの最初のルールに着目する．
【step 4】 着目した文字から始まる文字列に対し，着目した変形ルールが適用できるか調べる．
適用できなければ，step 7 に移る．
適用できれば，変形する（文字列Vとする）．
【step 5】 文字列Vが禁則ルールにマッチするか，あるいは，スタックに既に登録されているか調べる．どちらかであれば，step 7 に移る．
【step 6】 文字列Vが標準辞書にあるか調べる．
標準辞書にあれば，正常終了する．
なければ，文字列Vをスタックに登録する．
【step 7】 着目していない変形ルールがあれば，その変形ルールに着目し，step 4 に戻る．
【step 8】 着目している文字が最後尾の文字でなければ，次の文字に着目し，step 3 へ戻る．
【step 9】 スタックに登録されている文字列で調べていないものがなければ，異常終了する．
調べていないものがあれば，次の文字列をWにコピーして step 2 へ戻る． □

（2） 具体例による説明

例えば，“user”を起源とする“ユーザ”が標準辞書に登録されているものとする．以下，K_g=“ユーザア”から“ユーザ”に変形する過程を簡単に説明する．ただし，変形ルールは後述の表5を用いることとする．まず，先頭の文字“ユ”に着目する．この文字に適用できるルールは

ュ ←→ ユ
uウ ←→ u一

の2つで，“ューザア”と，“ユウザア”が生成される．最初の文字列は禁則ルール1にマッチするので，除外される．以下，着目する文字を1つずつ後ろにずらすことにより，変形ルールを1回適用して得られた“ユウザア”，“ユザア”，“ユアザア”，“ユーザー”，“ユーザァ”，“ユーザワ”の6つがスタックに入れられる．次に，この各文字列について2回目の変形ルールの適用が行われる．その結果，文字列“ユーザー”から変形ルール

一 → ×

を適用して“ユーザ”を得て，終了する．

4. 方式評価

4.1 EDR 辞書インタフェース

EDR 辞書インタフェースは日本語と英語の2種類があり，それぞれは EDR 電子化辞書[7]の内の日本語単語辞書と英語単語辞書から単語見出し，概念 id，概念見出しの3項目を抽出して作成された辞書である．単語見出しは単語の表記，概念 id は単語の表す概念の識別子，概念見出しは日本語と英語による概念の説明である．当然，1つの概念に対して単語見出しが複数あることがある．この場合，各単語見出しは互いに同義の関係にあるととらえることができる．逆に，1つの単語見出しに対して概念が複数存在することがある．この場合，その単語は複数の意味を持つことを示している．

以下，EDR 辞書インタフェースの日本語部分を単に EDR 辞書と呼ぶことにする．EDR 辞書中の概念数と見出し語数の関係，同一概念内の片仮名の見出し語数と概念数との関係をそれぞれ**表3**と**表4**に示す．同一概念に複数個の片仮名の見出し語がある場合に，必ずしもその元となる語が1つとは限らないが，ほとんどは表記のゆらぎとして登録されている．EDR 辞書では，片仮名の表記のゆらぎをかなり網羅的に集めてあり，本方式評価の材料として最適である．

表3 EDR 辞書における概念数と見出し語数の関係
Table 3 The number of concepts and vocabulary words in EDR dictionary.

	概念数	見出し語数
全体	202,849	428,411
片仮名のみ	29,854	59,527

表 4 EDR 辞書における同一概念内の片仮名見出し語数の分布
Table 4 The number of vocabulary words whose origins are the same.

同一概念内見出し語数	概念数	同一概念内見出し語数	概念数	同一概念内見出し語数	概念数
1	17,777	11	3	27	1
2	7,553	12	36	28	1
3	1,046	14	1	32	53
4	2,164	16	195	33	1
5	102	17	2	40	1
6	113	19	1	48	1
7	21	20	3	64	13
8	730	22	1	96	1
9	8	23	1	128	3
10	7	24	14	256	1

4.2 EDR 辞書の特徴

表3と表4で示した統計的な特徴のほかに，以下の特徴がある．

①区切り記号は“・”1種類である．

②バグが若干混入している（例えば，“マグロ”と“マアジ”が同一の概念に属している）．

③ゆらぎの登録に一貫性がないケースがある．例えば，英語の“weight”に対して，

Ω (weight)
＝{ウエイト，ウェイト，
ウエート，ウェート}

が期待される．当然，weightを含む単語（例えば，weight lifting）についても同様に，weightの部分には4種類のゆらぎがあるはずであるが，実際にはまちまちである．

多数の人が手作業で辞書を作成したために，②や③の問題が生じるが，辞書の内容はすべて正しいものとみなして実験を行った．

4.3 変形ルールと標準辞書

変形ルールの設定は以下の手順で行った．

【step 1】 3.1節で述べた変形ルールを利用する．

【step 2】 重複ルールがあれば，変形ルールから削除する．

【step 3】 同一概念に登録されている表記の間で変形が可能か否かをすべての組合せについて調べる．

【step 4】 変形できない組合せを調べ，変形ルールが導けるものがあれば変形ルールとして採用する．

【step 5】 追加された変形ルールがあれば step 2 に戻る． □

この結果，**表5**の変形ルールが得られた．なお，空文字との変形を行うルールは変形のバリエーションが増えてしまうので，空文字への変形しか許さないようにしている．

次に，2.2節で述べた方法を使って，変形できる

表 5 変形ルール（使用回数順）
Table 5 Conversion rules (in order of usage).

順位	ルール	使用回数	順位	ルール	使用回数
1	・ → ×	19,309	48	キシ ←→ クシ	16
2	ー → ×	14,041	49	ヒ ←→ フィ	16
3	ッ → ×	2,604	50	ョ ←→ ヨ	13
4	eー ←→ eイ	2,304	51	サウルス ←→ ザウルス	12
5	ェ ←→ エ	818	52	ビュ ←→ ヴュ	11
6	チ ←→ ティ	606	53	ガ ←→ ギャ	10
7	ィ ←→ イ	516	54	シャ ←→ シュア	10
8	iア ←→ iヤ	465	55	ッ ←→ ツ	10
9	バ ←→ ヴァ	277	56	uウ ←→ uー	10
10	ジェ ←→ ゼ	250	57	ンア ←→ ンナ	9
11	ビ ←→ ヴィ	232	58	ガ ←→ グァ	8
12	チャ ←→ チュア	230	59	ハ ←→ ファ	8
13	ンn ←→ n	227	60	チ ←→ ツィ	7
14	ジュ ←→ デュ	209	61	eエ ←→ eー	7
15	ブ ←→ ヴ	172	62	ウォ ←→ オ	6
16	デ ←→ ディ	150	63	バイ ←→ ヴィ	6
17	oウ ←→ oー	126	64	ウァ ←→ ワ	5
18	ベ ←→ ヴェ	112	65	ウィルス ←→ ビールス	5
19	フォ ←→ ホ	108	66	ズ ←→ ドゥ	5
20	イ ←→ ウィ	95	67	シェ ←→ セ	4
21	リツ ←→ レ	92	68	ド ←→ ドゥ	4
22	ツ ←→ トゥ	91	69	ナ ←→ ンア	4
23	ジ ←→ ディ	89	70	ファイ ←→ フィ	4
24	iイ ←→ iー	88	71	ゥ ←→ ウ	3
25	eア ←→ eヤ	86	72	キ ←→ クィ	3
26	aウ ←→ aー	84	73	キソ ←→ クソ	3
27	キサ ←→ クサ	75	74	ニ ←→ ンイ	3
28	iウ ←→ iュー	75	75	oオ ←→ oー	3
29	カ ←→ キャ	70	76	イェ ←→ エ	2
30	ォ ←→ オ	69	77	ギ ←→ グィ	2
31	ア ←→ ー	65	78	クォ ←→ コ	2
32	キス ←→ クス	65	79	サイ ←→ シ	2
33	ボ ←→ ヴォ	62	80	ジ ←→ ダイ	2
34	ア ←→ ワ	52	81	ハイドロ ←→ ヒドロ	2
35	ャ ←→ ヤ	50	82	ピェ ←→ ペ	2
36	ァ ←→ ア	37	83	フェ ←→ ヘ	2
37	ト ←→ トゥ	34	84	ム ←→ ン	2
38	ンm ←→ m	23	85	ュ ←→ ユ	2
39	テ ←→ ティ	22	86	キリ ←→ クリ	1
40	スム ←→ ズム	21	87	クェ ←→ ケ	1
41	oル ←→ oー	21	88	セ ←→ チェ	1
42	チュ ←→ テュ	20	89	ヒ ←→ フュ	1
43	シ ←→ シュ	19	90	ヒェ ←→ フェ	1
44	ジ ←→ ヂ	19	91	ブウ ←→ ブュー	1
45	ウェ ←→ エ	18	92	ミ ←→ ミュ	1
46	カ ←→ クァ	17	93	ゥ ←→ ワ	1
47	オ ←→ ヨ	16			

注：a，i，u，e，oはそれぞれア列，イ列，ウ列，エ列，オ列の片仮名文字，nとmはナ行，マ行の片仮名文字，×は空文字の意味である．

表 6 EDR 辞書と標準辞書の関係
Table 6 Standard dictionary.

EDR 辞書		標準辞書見出し数
概念内見出数	見出し数 (=総見出し数)	
複数	41,750	14,968
単数	17,777	17,777
合計	59,527	32,745

表 7 同一概念内の標準辞書に登録されている見出しとの距離
Table 7 The distribution of distance between a word and a vocabulary word.

距離	個数
0	32,684
1	15,264
2	7,248
3	2,996
4	1,016
5	266
6	49
7	4

表記の組合せごとに1つの表記を選ぶことにより，標準辞書を作成した．ただし，上述のように空文字からの変形ルールを禁止したことにより，Ω(w) 中に区切り記号，長音，促音を持つ表記と持たない表記の両方がある場合には，区切り記号，長音，促音を持たない表記を選ぶことにした．さらに，区切り記号を持たない表記がない場合には，区切り記号を抜いた文字列を作り，標準辞書に登録した．標準辞書の内訳を**表 6** に示す．

4.4 評価結果

EDR 辞書の中で1概念1表記の見出しはそのまま標準辞書に登録される．従って，1概念複数見出しの表記がどのくらい少なくできたかにより，本論文で提案した方式の評価ができる．表6より，93個の変形ルールを使うことにより，1概念複数見出しの表記のうち標準辞書に登録された表記は約 1/3 に削除できることが分かる．

表 7 に示した同一概念内の標準辞書に登録されている見出しとの距離の統計より，EDR 辞書中の片仮名の見出しは最大7回，平均 1.1 回の変形ルールの適用で標準辞書の見出しに変形できることが分かった．なお，区切り記号を予め抜いておくと距離の最大は5，平均は約 0.6 となる．また，標準辞書に登録されている見出しに一致するまでに生成される表記の数は最大 262，平均 4.9 であることが分かった．

EDR 辞書のすべての片仮名の見出しについて本方式を適用して，最小距離で一致した標準辞書の見出しを調べたところ，以下の結果が得られた．

①最小距離で一致した標準辞書の見出しが1つの場合，元の見出しと標準辞書の見出しは同一概念に含まれている．

②最小距離で一致した標準辞書の見出しが複数の場合，元の見出しと同一概念に含まれ標準辞書の見出しがある．この場合，元の見出し，あるいはそのゆらぎの表記が複数の意味を持っている．

以上より，本方式の有効性が確認できた．

5. おわりに

本論文では，片仮名の表記のゆらぎを変形ルールを用いることにより辞書に登録する表記を少なくする方法について提案し，その効果を EDR 辞書により確認した．変形に当たっては，途中で生成される表記のバリエーションを減らすために，禁則ルールを適用した．EDR 辞書で確認した結果，辞書を約 1/3 に削減できること，適用した変形ルールの回数は最大7，平均約 1.1 であること，変形ルールを適用して生成される表記は最大 262，平均 4.9 であることが分かった．以上より，本方式の有効性が立証できた．

本論文では片仮名のみからなる表記について言及したが，本方式は漢字と片仮名の混ざった表記（例えば，“計算機アーキテクチャ”）に対しても有効である．

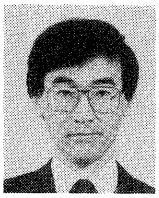
表5で示したように変形ルールの使用回数には非常に差がある．今後は，変形ルールに優先度をつけて変形のバリエーションを減らすような検討をする予定である．

参考文献

1) 池原，宮崎，横尾：日英機械翻訳のための意味解析用の知識とその分解能，情報処理学会論文誌，Vol. 34, No. 8, pp. 1692-1704 (1993).
2) 横井俊夫：日本語の情報化—その技術をめぐって—，p. 245，共立出版，東京 (1990).
3) 内閣告示第2号：外来語の表記，官報，平成3年6月28日 (1991).
4) 加藤，藤澤，大山，川口，畠山：大規模文書情報システム用テキストサーチマシンの研究，情報処理学会情報学基礎研究会，14-6 (1989).
5) 久保田，庄田，河合，玉川，杉村：カタカナ表記

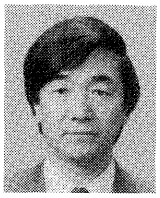
の統一方式　予備分類とグラフ比較によるカタカナ表記のゆらぎ検出法，情報処理学会自然言語処理研究会，97-16 (1993).
6) 宮内：カタカナ表記からの英語単語検索システムの実現，情報処理学会自然言語処理研究会，97-17 (1993).
7) 日本電子化辞書研究所：EDR 電子化辞書使用説明書，p. 134，日本電子化辞書研究所，東京 (1993).

(平成6年3月 3日受付)
(平成6年7月14日採録)



飯田　敏幸（正会員）

昭和26年生．昭和49年東京大学工学部計数工学科卒業．昭和51年同大学院修士課程修了．同年日本電信電話公社（現 NTT）入社．現在，NTT コミュニケーション科学研究所主幹研究員．分散データベースシステム，人工知能などの研究・開発に従事．電子情報通信学会，人工知能学会各会員．



中村　行宏（正会員）

昭和42年京都大学工学部数理工学科卒業．昭和44年同大学院修士課程修了．同年日本電信電話公社入社．同社電気通信研究所において，DIPS 論理装置の研究開発を経て，昭和56年より主に並列処理アーキテクチャを有するプロセッサの方式設計技術の研究に従事．現在，NTT 情報通信研究所高速通信処理研究部部長．平成4年4月より電気通信大学大学院情報システム学研究科客員教授を兼任．情報処理学会論文賞（平成元年度），大河内記念技術賞（平成3年度），科学技術庁長官賞（平成6年度）各受賞．著書「ULSI の効率的設計法」（共著，オーム社），「High-Level VLSI Synthesis」（共著，Kluwer Academic Publishers）等．電子情報通信学会，IEEE 各会員．