3Q−4

カタカナ異形表記・誤記修正機能の開発・評価

島津 美和子、 吉村 裕美子、 平川 秀樹、 天野 真家
(株) 東芝 総合研究所

1．はじめに

日本語における外来語、外国語の氾濫が言われるようになって久しいが、その傾向は強まる一方である。同時に、カタカナ表記される外来語が日本語文書中に占める比率は急増している。自然言語処理システムは、こういった日本語の変化に即対応しなければならない。しかし、カタカナ語はひらがなや漢字と違い、統一がとれておらず、一語に対して表記法(異なり語)が複数存在するという大きな問題がある。日英翻訳システム(MT)や日本語OCRでは、辞書中にある語しか処理対象にならず、辞書中に異なり語がないとき、処理精度の低下をもたらすことがある。

そこで、我々は異なり語を自動生成し、辞書を利用することにより、カタカナ語の処理を容易にするために、カタカナ語の異表記生成・誤記(誤字)修正機能を開発・評価した。

2．異表記現象

カタカナ外来語は、従来のシステムでは以下の三つのタイプの表記のゆれにより、それに相当する原語が辞書に入っていても正しく検出されず、また関連語が入っていてもその情報が参照されず、未知語と見なされていた。

(1) 異形表記 (例) リスポンス：レスポンス、パターン：パタン

厳密な意味での表記のゆれである。これには次の要因がある。第一に、カタカナ外来語は原語の発音に近付こうとして絶えず修正を迫られ、しかも原語と日本語は音韻体系が違う以上、正しい発音になり得ない。従って、外来語の表記は不安定になる。一方、外来語はその綴り字に引きずられて、ローマ字読みをし、本来の発音とは異なった形で導入され、慣用化することもある。また、国語審議会の動向に見られるように標準的表記方法は厳密には規定されていない。

(2) 活用または語形成の規則による関連語 (例) コード：コーディング、ポライト：アンポライト

活用または語形成の規則によって、原語一語に対し複数の関連語がある。英語では、動詞には現在分詞形/動名詞形、過去形、過去分詞形があり、名詞は一般に単数形と複数形があり、また接辞を付加すると品詞変換や、微妙な意味の変化が起こる。これらの派生語はカタカナに書き換えられた後、日本語の中に取り入れられている。カタカナ表記されると、日本語だけに注目している限り、原語では一目瞭然だった関連性が見落とされやすい。カタカナ語においてもこれらの関連をそのまま保持させれば、これも一種のゆれと見ることができる。

(3) 誤記、誤入力、誤読、誤認識 (例) ポイインタ：ポインタ、アセンブラ：アセンブラ

キーボード入力とOCR入力によるものとがある。前者では、一文字の抜け、余分な文字の追加、文字の取換え、隣接文字の反転が大半を占めることが実証されている。これはカタカナ語特有の現象ではなく、ひらがな表記の語や英単語にも見られる。他方、OCR入力では、形の類似した文字の間で誤認識が起こる。

本稿では、出現頻度の最も高いタイプ(1)と(2)を対象の中心にすえ、(3)は今後の課題とする。

3．異表記分析

各種の国語・外来語辞典の他、国語審議会編「外来語の表記」を参考にし、各カタカナ語につき可能な表記をすべて分析した。これにより得られた58の規則を言語理論(特に音韻論)に裏打ちされた各々の性質に着目して次の7つに分類した。

1.長音 (例) ー→null　2.音の脱落 (例) ッ→null
3.音の交替 (例) キ→ク　4.小文字・大文字 (例) ァ→ア
5.発音 (例) 濁音→清音　6.語形変化 (例) 単数→複数
7.原語の発音に近付けようとして生じた異形 (例) ファ→ハ

次に、規則を構造化するため、各規則とその逆の規則(計116)を、無条件か条件付きで成立するかで区別した。どのような環境の下で個々の規則が成立するかを細かく分析したところ、無条件に成立するものは22のみであった。このように、単なる書き換えで済むものはごく限られているため、その他の94の規則に対しては厳しい条件の付与が重要となる。

外来語を原語に戻すと、可能な表記法の規則の条件付けが容易になる。一方、日本語の表記からは一意に英語綴りを復元できない。しかし、本機能は、日本語の枠組で閉じているので、原語を参照できない。例えば、上の2.で「ッ」を削除するのは容易だが、逆に「ッ」を挿入するのは困難である。この制約を克服するためには、日本語と英語の音韻体系の違いを考慮しながら、ゆれのある箇所の文字(列)の位置や全体の文字数、文字の段と行、語中の他の文字に注目して厳しい適用条件をつけたり、ある規則と規則は自動的に適用されないという背反関係を用いたり、規則の適応順序を規定する枠組が必要である。同時に、上の制約からすべてのケースをカバーすることが不可能な場合、例えば、「ッ」の挿入では語末が「シュ」、「クス」、「キス」でその直前に「ッ」がないのものにのみ限定するといったように、パターン的に処理するための枠組も必要である。

4．異表記派生規則

4.1 特徴

前章で述べた問題を具体的に解決するため、図1に示すような異表記派生規則を独自に開発した。この規則は、ある文字列が与えられると、構文的・意味的に同じ機能をする別の文字列(タイプ(1))さらに、構文的・意味的には異なるが、関連する文字列を生成する(タイプ(2))。

本規則の特徴は以下の3つである。

1.個々の表記のゆれ・書き誤りに対応する規則(個別現象派生規則)と、これを構造化し個々の規則の適用を制御する規則(派生組み合わせ規則)とを分けて記述している。
2.条件記述が自由にできる。

An Autocorrection Function for Katakana Variants

Miwako SHIMAZU, Yumiko YOSHIMURA, Hideki HIRAKAWA, Shinya AMANO
TOSHIBA Corporation

3.派生文字列の品詞(構文的役割)を限定したり、派生文字列と入力文字列との関係を付加情報として生成するなどのアクション機能を持っている。

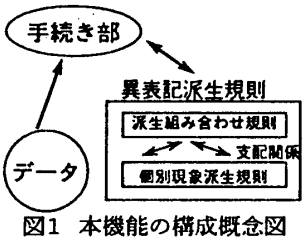


図1 本機能の構成概念図

4.2 個別現象派生規則

個別現象派生規則は、一規則が一(部分)文字列の派生を表しており、派生のバリエーションの種類だけ知識的に蓄えられている。規則間には羅列順序による優先関係はなく、まったく独立の規則群からなる。下は、その1例である。

```
1{?}[段=(あ|い)]+2{ー}[E]→;                長音の脱落
                        (例) コンピューター:コンピュータ
1{*}+2{ッ}+3{*}→1+3;                       促音便の脱落
                        (例) マトリックス:マトリクス
1{*}+2{?}[2≠ッ]+3{クス|キス|シュ}[E]→1+2+{ッ}+3;
                                            促音便の挿入
                        (例) マトリクス:マトリックス
```

「→」の左辺がマッチング文字列、右辺が書き換え文字列である。(「?」は任意の1文字、「*」は任意の文字列を示すワイルドカードである。){}に添えられた数字が左辺と右辺との対応を示す。よって、左辺にあるが右辺にない数字が指す文字列は削除され、右辺にのみあるものは新たに挿入される文字列であり、両辺にあるものは文字の置換を示している。[]内は文字列の適用条件を記述する。(Bは語頭、Eは語末を表す。)[]内に分析過程で得られた条件を豊富に盛り込んだ。本機能は現時点で、言語学の理論を用いて精緻化された個別現象派生規則を約150蓄えている。

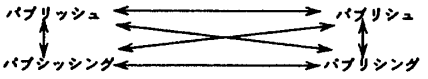
4.3 派生組み合わせ規則

派生組み合わせ規則は、個々の個別現象派生規則に対する適用の是非・優先付けなどの判断を行う。つまり、一つの文字列から別の文字列を生成するに当たって、個別現象派生規則のどの規則を使用するか、どの規則とどの規則を組み合わせるか、その際の適用順序はどうするかを支配する。これにより、容認不可な文字列の出力を抑え、さらに一般性の高い候補から出力する。

「パブリシュ」から「パブリッシング」を派生する単純化した規則の記述例を挙げる。

```
(1)S[文字数>3]→$TSU,$ING;
(2)$ING→<1>(品詞 サ変);
(3)$TSU→<2>;
(以降、個別現象派生規則)
<1>1{シュ}[E]→1{シング};
<2>1{?}[段=い&1≠イ]+2{シュ}[E]→1+{ッ}+2;
```

本規則は拡張文脈自由型文法の枠組を取る。左辺に初期記号Sを持つ規則の数だけ単語としての文字列の派生を行う。複数のS規則をうまく条件付けながら順序付けることで、確からしいものから生成を行うように制御できる。右辺の項目の順序は、最も下位のレベル(個別現象派生規則を適用するレベル)での順序を反映する。また、両辺の各項に適用条件を付すことができるので、自由に構造化した規則を記述できる。このようにして、入力文字列が「パブリシュ」、「パブリッシュ」、「パブリシング」、「パブリッシング」のいずれであっても他の3つの派生形を出力できる。



4.4 アクション機能

アクション機能は、MT利用の際、効果を発する。これは規則の右辺の()内で示す。上の(2)では、品詞をサ変名詞に限定することを意味する。他にも、接頭・接尾辞を伴う派生語の対応に用いられる。例えば、「ポライト(polite)」は辞書中に登録されているが、「アンポライト」はない場合、次のような否定接頭辞の「アン」の異形表記規則を適用して、「アンポライト」から派生元の語である「ポライト」を導き出す。そして、辞書から提供される「ポライト」の翻訳規則に、否定の意味を加えれば良い。このようにして、「アンポライト」の処理には、「ポライト」の情報を利用できる。

```
S→$HITEI;
$HITEI→<3>(否定 yes);
<3>1{アン}[B]+2{*}→2;
```

5.規則の評価

形として纏まったプログラムを実際使用し、規則の有効性を具体的に把握するため、以下の3つの観点から数値化を試みた。ここでテストのために用いたデータは技術文書からカタカナ語(計290)を抽出したものである。

(1)出るべき候補が出ているか。全単語につき、「出るべき候補のうち本機能で出力された候補の数/出るべき候補の数」という分数値を算出し、その平均値を取った(A)。

(2)明らかに排除すべき文字列が候補として挙げられているか。「明らかに不適な候補の数の合計/本修正機能で出力された候補の数の合計」という分数値を算出した(B)。

(3)優先度の高いものほど先に挙げられているか。まず、第一候補が最初に出たかどうかに注目した(C)。次に、上から5つの候補と6番目以降の候補を見比べ、6番目以降の候補にこれらの5つの候補よりも確からしさが高い候補が含まれているかを見た(D)。

実験の最終結果は以下の通りである。

A=94.02%　　B=34.20%　　C=96.55%　　D=2.76%

6.まとめ

表記が不統一であるカタカナ語のゆれを吸収する目的で作成された、入力カタカナ文字列から可能な異形表記を出力するカタカナ異表記生成・誤記(誤字)修正機能について述べた。規則を組み立て、条件付けを厳しくすることにより、基本形から変化形、変化形から基本形の双方向に、かなり高い精度で異表記を出力することができ、有効性を証明できた。タイプ(3)の関連では、本機能はカタカナに限らず漢字、ひらがな、カタカナまじりの文字列にも応用でき、今後校正支援やカタカナ以外にも対応するOCR向けに研究を進めていく。

7.参考文献

1.遠藤織枝、「外来語の表記」(武田良明編、『講座 日本語と日本語教育第8巻 日本語の文字・表記(上)』)、明治書院、1989
2.国立国語研究所、『日本語教育指導参考書16外来語の形成とその教育』、大蔵省印刷局、1990
3.文化庁、「国語審議会答申 外来語の表記」、1991