カタカナ表記ゆれに対応した辞書引きシステム

伊藤 美咲姫 †　　佐藤 理史 ‡　　駒谷 和範 ‡

† 名古屋大学 工学部 電気電子・情報工学科　‡ 名古屋大学大学院 工学研究科

{misaki_i, ssato, komatani}@nuee.nagoya-u.ac.jp

1　はじめに

正書法が比較的緩やかなガイドラインにすぎない日本語において、語表記のゆれは、日常茶飯事の現象である。国立国語研究所がまとめた『現代表記のゆれ』[1] には、「語表記のゆれは、日本語だけに存在する現象ではない。—(中略)— しかし、—(中略)— 日本語の語表記のゆれは、量と種類の双方において、他の言語におけるそれを格段にうわまわっている。」と述べられている。

日本語テキスト処理において、表記ゆれに関わる問題は、特にカタカナ語を中心に、これまでいくたびも取り上げられてきた [2, 3, 4, 5, 6] 。しかしながら、いまだに研究が途切れていないということは、抜本的な解決がなされてないことを意味する。

本稿では、表記ゆれのなかでも特に量が豊富と考えられる外来語のカタカナ表記のゆれを対象に、辞書引きの際に生じる問題の解決策を論じる。

2　カタカナ語の異形

外来語に由来するカタカナ語は、原語の発音を、近似的にカタカナで表記すること (トランスリタレーション) によって生み出される。その過程で多くの異形が作られる。

外来語の表記に対しては、内閣告示・訓令『外国語の表記 (平成3年6月28日)』がそのガイドランを定めている。この文書では、「外来語や外国の地名・人名を書き表すのに一般的に用いる仮名 (第1表)」に加え、「原音や原つづりになるべく近く書き表そうとする場合に用いる仮名 (第2表)」が示されている。さらに、「第1表・第2表に示す仮名では書き表せないような、特別な音の書き表し方については、ここでは取決めを行わず、自由とする」との記述がある。

以上の説明から明らかなように、この文書が示すガイドラインは緩やかなもので、自由度がある。たとえば、「ヴァイオリン」の「ヴァ」は第2表で定義される仮名で、留意事項には、「一般的には『バ』と書くことができる」との記述がある。このため、「ヴァイオリン」と「バイオリン」は、どちらも、ガイドラインに沿った表記となる。

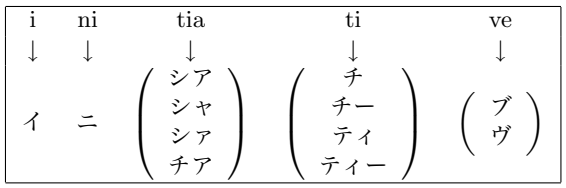

図 1: トランスリタレーションによる異形の生成

トランスリタレーションによる異形の生成は、原綴の各音素 (あるいは綴の部分) にどのようにカタカナを当てはめるかの組合せとしてモデル化できる。たとえば、“initiative” に対しては、図1のようなモデルを考えることができる。このモデルは、32種類 ($= 14 \times 4 \times 2$) の異なる異形を生成するが、そのうちの16種類は、形態素解析用辞書 UniDic-2.1.0 に収録されており、実際に存在すると考えられる。なお、ここで使われる仮名のうち、「シァ」は前述の『外国語の表記』の第1表・第2表に含まれていない仮名である。

3　語形のゆれと表記のゆれ

文献 [1] では、いわゆる表記ゆれ (異形) を、**表記のゆれ**と**語形のゆれ**の2種類に細分する立場をとっている。形態素解析用辞書 UniDic でも、この考え方が踏襲されており、表記のゆれと語形のゆれを区別できる構造を採用している。

表1に、UniDic のエントリーの構造を示す。UniDic のエントリーは、語彙素、語形、書字形という3つの階層を持つ [7]。トップ階層の**語彙素**は、語形・表記・発音の変異を考慮せず、意味・文法機能が同一であると見なし得るものに同一の見出しを与えたものである。次の階層の**語形**は、同一の語彙素に対して、形態の違いを区別したものである。最下階層の**書字形**は、同一の語形に対して、表記の違いを区別したものである。なお、それぞれの語彙素と語形には、代表形 (表記) が

表 1: UniDic のエントリーの構造

語彙素	語形	書字形
013717 イニシアチブ (initiative)	.1 イニシアチブ	.1.1 イニシアチブ
		.1.2 イニシアチヴ
	.2 イニシアチーブ	.2.1 イニシアチーブ
	.3 イニシアティブ	.3.1 イニシアティブ
		.3.2 イニシアティヴ
	.4 イニシアティーブ	.4.1 イニシアティーブ
		.4.2 イニシアティーヴ
	.5 イニシャチブ	.5.1 イニシャチブ
		.5.2 イニシャチヴ
	.6 イニシャチーブ	.6.1 イニシャチーブ
		.6.2 イニシャチーヴ
	.7 イニシャティブ	.7.1 イニシァティブ
		.7.2 イニシァティヴ
		.7.3 イニシャティブ
		.7.4 イニシャティヴ
	.8 イニチアティブ	.8.1 イニチアティブ

定義されている。表 1 では、これらを区別するために、辞書中の各要素に対して、辞書中の位置を表す ID を示してある。

このような構造下において、表記のゆれ、および、語形のゆれは、次のように説明される。

表記のゆれ 一つの語形に、複数の書字形が定義される場合、その語形には、表記のゆれが存在する。

語形のゆれ 一つの語彙素に、複数の語形が定義される場合、その語彙素には、語形のゆれが存在する。

この説明に基づき、同一語彙素に属する 2 つの書字形間に、表記のゆれ、または、語形のゆれを認める。たとえば、「イニシアチブ (013717/.1/.1.1)」と「イニシアチヴ (013717.1.2)」は、同一語形 013717.1 に属するので、表記のゆれとみなす。一方、「イニシアチヴ (013717.1.2)」と「イニシアティブ 013717.3/.3.1」は、同一語形には属しないが、同一語彙素 013717 に属するので、語形のゆれとみなす。

本研究では、辞書にこのような構造を仮定し、表記のゆれと語形のゆれを区別する。

4 表記・語形のゆれと辞書引き

一般に、辞書引きは、次のような関数としてモデル化できる。

$$\mathrm{lookup}(q, D) \quad = \quad \{e | e \in D, q \in \mathrm{spelling}(e)\} \quad (1)$$

ここで、q は辞書引き対象の語 (より正確には、文字列)、D は辞書、e は辞書のエントリー、関数 $\mathrm{spelling}(e)$ は、エントリー e のすべての書字形を返す関数を表す。つまり、この関数は、q を書字形として持つエントリー $e\,(\in D)$ をすべて返す関数である。

この関数が文字列 q に対して空集合を返すとき、次の 2 つの可能性が存在する。

1. 文字列 q は、辞書に登録されているエントリー $e\,(\in D)$ の、未登録の書字形である。
2. 文字列 q は、辞書に未登録のエントリーの書字形である。

一般に言われる、辞書引きにおける「表記ゆれ問題」とは、前者の状況を言う。(後者は、一般に、未知語問題と呼ばれる。)

この「表記ゆれ問題」に対する自明な解は、文字列 q をエントリー e の書字形として登録することである。しかしながら、事前にすべての異形 (書字形) を辞書に登録していくことは、事実上、不可能である。この事実が、この問題の本質である。

もし、辞書引き機構に、異形の生成・認定機能をもたせることができれば、すべての異形をあらかじめ辞書に登録しておく必要はない。事実、人間用の辞書には、すべての異形が網羅されているわけではないが、利用者である人間 (日本語母語話者) は、異形の生成・認定機能を持っているため、辞書引きに特に不自由しない。

ここでは、ある書字形 s に対し、その異形をすべて生成する関数 $\mathrm{variant}(s)$ を導入し、辞書引きに組み込むことを考えよう。これは、次のような関数でモデル化できる。

$$\widehat{\mathrm{lookup}}(q, D) = \{e | e \in D, s \in \mathrm{spelling}(e), q \in \mathrm{variant}(s)\} \quad (2)$$

辞書のエントリーに、前節で述べたような構造を仮定する場合、辞書 D は、語彙素の集合 L、あるいは、語形の集合 F とみなすこともできる。そこで、次のような関数を考えることができる。

$$\widehat{\mathrm{lookup}}_F(q, F) = \{f | f \in F, s \in \mathrm{spelling}(f), q \in \mathrm{variant}_S(s)\} \quad (3)$$

$$\widehat{\mathrm{lookup}}_L(q, L) = \{l | l \in L, s \in \mathrm{spelling}(l), q \in \mathrm{variant}_F(s)\} \quad (4)$$

ここで、$\mathrm{variant}_S(s)$ は、書字形 s の表記ゆれ異形をすべて生成する関数、$\mathrm{variant}_F(s)$ は、書字形 s の語形ゆれ異形をすべて生成する関数を表す。関数 (1) と上記の 2 つの関数を組み合わせることにより、次のような結果を得ることができる。

1. look_up(q, D) が空集合以外を返したとき、q は、得られた集合の要素 e の書字形である。(語彙素、語形、書字形は、すべて辞書 D に登録されている。)

2. $\widehat{\text{look\_up}}_F(q, F)$ が空集合以外を返したとき、q は、得られた集合の要素 (語形)f に属する、未登録の書字形である。(書字形は未登録であるが、語形は同定できる。)
3. $\widehat{\text{look\_up}}_L(q, L)$ が空集合以外を返したとき、q は、得られた集合の要素 (語彙素)l に属する、未登録の語形の (未登録の) 書字形である。(書字形、語形は未登録であるが、語彙素は同定できる。)
4. 上記のすべての関数が空集合を返したとき、q は、未知語 (未登録の語彙素の書字形) である。

以上のことは、2 つの異形生成関数 $\text{variant}_S(s)$ と $\text{variant}_F(s)$ を実装することで、実現できる。

5 辞書引きシステムの概要

与えられた入力文字列 q に対して、その異形を生成して、辞書の見出し語リストと照合するシステムを作成した。異形の生成には、Non-Productive Machine Transliteration (NPMT) [8] の枠組を用いる。この枠組は、入力された文字列に、部分文字列の変換規則群を適用して得られる文字列のうち、あらかじめ用意されたターゲットリストに含まれる要素を返す機能を提供する。ここで、変換規則群として、異形生成規則群を与え、ターゲットリストとして、辞書の見出し語集合 (より正確には、辞書の書字形集合) を与えることで、入力 q から、それと異形関係にある書字形を同定することが実現できる。最終的に、こうして得られた書字形を通して、それが属する語形および語彙素を同定することができる。

5.1 異形生成規則群

異形生成規則群には、表記ゆれ、および、語形のゆれを生み出す元となる、カタカナ列の変換規則を記述する。「『現代書き言葉均衡コーパス』形態論情報規程集第 3 版 (以下、規程集と略す)」[9] の同語異語判定規程には、かなり詳細に、同一語形・別語形の判定規程および語形の定め方が記述されている。たとえば、「『ヴァ⇔バ』の差異は、同一の語形とする」(すなわち、表記ゆれとみなす) と述べられているので、これに基づき、表記ゆれ規則として、「ヴァ⇔バ」を定義する。具体的な規則を定義するにあたっては、上記の規程集以外に、前述の『外来語の表記』、文献 [1]、および、複数のカタカナ語辞典を参考にした。作成した規則群の概要を表 2 に示す。規則群は、表記ゆれをカバーするための RS1 と、語形ゆれをカバーするための RS2 に分かれている。

表 2: 異形生成規則群の概要

規則タイプ		RS1	RS2	p
S	中黒「・」の有無の差異	1	–	1
L	長音符号と母音字の交替	11	–	10
R	小書きと通常文字の交替	13	–	10
H	典型的な表記ゆれ	71	–	10^2
Q	小書きの母音字の挿入	5	–	10^3
RG	「ー」「ッ」の有無の差異	1	1	10^4
G	典型的な語形のゆれ	–	91	10^5
T	複数形 (語末)	–	4	$5 \cdot 10^4$
O	その他	–	209	10^6
	合計	102	305	

個々の規則では、変換すべきカタカナ列の対に加え、規則タイプを定義する。この規則タイプは、ゆれ (差異) の種類を表すもので、現在、9 種類ある。たとえば、前述の規則「ヴァ⇔バ」の規則タイプは H で、これは、表記ゆれの典型的な規則であることを意味する。規則タイプは、最終出力を決定するために使用されるペナルティ (表 2 の p の値) の値を規定する。

5.2 ペナルティによる出力選択

一般に、入力された文字列に対して異形生成規則群を適用すると、複数の異形候補が生成される。NPMT の枠組は、これらを、ターゲットリスト (辞書の書字形リスト) に含まれるもののみに絞り込むが、それでもなお、複数の候補が残ることがある。その場合は、各候補のペナルティの値が最も小さいもの (複数の場合もある) を選択して出力する。各候補のペナルティ値には、その候補を生成するために使用された異形生成規則ペナルティの総和を用いる。

たとえば、入力「ヴァイオリン」から生成される異形候補「バイオリン」のペナルティは、「ヴァ⇔バ」のペナルティの値 10^2 である。これに対して、異形候補「バイアル」のペナルティは、「ヴァ⇔バ」(p=10^2)、「オ⇔ア」(p=10^6)、「リ⇔ル」(p=10^6)、「ン⇔ϵ」(p=10^6) のペナルティの総和 3,000,100 となる。この結果、最終的に「バイオリン」のみが出力されることになる。

6 実験

まず、実験の準備として、UniDic-2.1.0 からカタカナ外来語を抽出し、26,228 語彙素、31,826 語形、34,063 書字形からなる、カタカナ語辞書を作成した。なお、この辞書に含まれる書字形の異なりは、33,571 種類である。

次に、この辞書に含まれる書字形を、次の 3 種類に分割した。

Type-L 語彙素の代表形として採用されている書字形：25,821 種類

Type-F 語彙素の代表形には採用されていないが、語形の代表形として採用されている書字形：5,489 種類

Type-O 語彙素の代表形にも語形の代表形にも採用されていない書字形：2,261 種類

もし、システムが表記ゆれ、語形ゆれのすべてをカバーすることができれば、Type-F と Type-O の書字形は、すべて削除することができる。

6.1 実験 1

実験 1 では、システムがどの程度、表記ゆれをカバーできるのかを調べた。

システムの入力には、Type-O の書字形 2,261 件を用い、ターゲットリスト (辞書の書字形リスト) には、Type-L と Type-F の合計 31,310 件を用いる。異形生成規則群には、表記ゆれ用 RS1(102 ルール) を用いる。正解判定は次のように行なう。たとえば、入力「イニシアチヴ (013717.1.2)」に対する正解は、語形 013717.1 であるので、その代表形「イニシアチブ (013717/.1/.1.1)」をシステムが出力すれば正解、出力しなければ不正解とする。

この実験では、システムは、入力 2,261 件中の 2,247 件 (99.4%) に対して、正解を出力した。その後の調査により、14 件の不正解のうちの 10 件は、UniDic-2.1.0 の誤り (規程集と辞書との不整合) であることが判明した。以上の結果より、本システムは、規程集が定義する表記ゆれのほとんどをカバーすることができることがわかった。すなわち、本システムを使用すれば、辞書に Type-O の書字形を登録することは不要となる。

6.2 実験 2

実験 2 では、システムがどの程度、語形ゆれをカバーできるのかを調べた。

システムの入力には、Type-O と Type-F の書字形 7,750 件を用い、ターゲットリストには、Type-L の書字形 25,821 件を用いる。異形生成規則群には、RS1 と RS2 の両方を用いる。正解判定は次のように行なう。たとえば、入力「イニシャティヴ (013717.7.4)」に対する正解は、語形 013717 であるので、その代表形「イニシアチブ (013717/.1/.1.1)」をシステムが出力すれば正解、出力しなければ不正解とする。

実験結果を表 3 に示す。この結果より、本システムは、表記ゆれ・語形ゆれの大半 (86.1%) をカバーすることがわかる。

表 3: 実験結果 (実験 2)

	正解	不正解	合計
Type-O	2,149 (95.0%)	112	2,261
Type-F	4,522 (82.4%)	967	5,489
合計	6,671 (86.1%)	1,079	7,750

表 4: 不正解の例

	入力	出力	正解
(a)	オーケイ クプル	オケイ クプラ	オーケー カップル
(b)	ピッツァ ウィミン	ピッチャー ウィメンズ	ピザ ウーマン

6.3 検討

実験 1 では、ほぼすべての書字形を削除することが可能であるという結果が得られたが、語形の代表形も削除した実験 2 では、正しく語彙素を同定できない書字形が 112 件発生した。

実験 2 の不正解は、(a) 異形生成規則群により、正しい異形は生成されているが、ペナルティ値による最終判定で出力されなかったもの (あるいは、不正解の異形が同時に出力されてしまったもの) と、(b) 異形生成規則群によって正しい異形が生成されなかったもの、に分けられる。これらの例を表 4 に示す。

カタカナ語の表記のゆれは、ほとんどが規則的なゆれである。また、語形のゆれも、その大半が規則的なゆれである。その一方で、「ピッツァ/ピザ」ような語に固有のゆれも存在する。このようなゆれは、規則でカバーするのではなく、辞書に登録すべきである。

参考文献

[1] 国立国語研究所. 現代表記のゆれ. 秀英出版, 1983.

[2] 伍井啓恭, 清原良三, 鈴木克志, 太細孝. カタカナ異表記処理. 情報処理学会 全国大会講演論文集, Vol. 38, No. 1, pp. 351–352, 1989.

[3] 島津美和子, 吉村裕美子, 平川秀樹, 天野真家. カタカナ異形表記・誤記修正機能の開発・評価. 情報処理学会 全国大会講演論文集, Vol. 44, No. 3, pp. 249–250, 1992.

[4] 久保田淳市, 庄田幸恵, 河合真宏, 玉川博文, 杉村領一. 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. 情報処理学会論文誌, Vol. 35, No. 12, pp. 2745–2751, 1994.

[5] 久保村千明, 亀田弘之. 片仮名異表記処理能力を備えもつ情報検索システム. 電子情報通信学会論文誌. D-II, Vol. 86-D-II, No. 3, pp. 418–428, 2003.

[6] 服部弘幸, 関和広, 上原邦昭. 英語音韻を考慮した情報検索のための多様なカタカナ異表記生成. 情報処理学会論文誌 数理モデル化と応用, Vol. 2, No. 1, pp. 145–155, 2009.

[7] 伝康晴, 小木曽智信, 小椋秀樹. コーパス日本語学のための言語資源–形態素解析用電子化辞書の開発とその応用. 日本語科学, Vol. 22, pp. 101–123, 2007.

[8] Satoshi Sato. Non-productive machine transliteration. In *Proc. of Riao-2010*, pp. 16–19, 2010.

[9] 小椋秀樹, 小磯花絵, 冨士池優美, 宮内佐夜香, 原裕. 『現代書き言葉均衡コーパス』形態論情報規程集第 3 版. 国立国語研究所, 2010.