社団法人　電子情報通信学会
THE INSTITUTE OF ELECTRONICS,
INFORMATION AND COMMUNICATION ENGINEERS

信学技報
TECHNICAL REPORT OF IEICE.

後続ひらがな列に基づく語の活用型推定

桑江　常則† 　佐藤　理史† 　藤田　篤†

† 名古屋大学 大学院工学研究科　〒 464-8603 名古屋市千種区不老町
E-mail: †kuwae@sslab.nuee.nagoya-u.ac.jp †{ssato,fujita}@nuee.nagoya-u.ac.jp

あらまし　新語や転用・派生語を収集して辞書を作成する場合，その語の品詞と活用型を認定する必要がある．動詞や形容詞は活用型ごとに異なる活用語尾を持つとともに，名詞の直後には格助詞が現れやすい．このような理由により，品詞や活用型の推定には，語幹に後続するひらがな列が強力な手がかりとなる．本稿では，このことを利用し，与えられた語幹の活用型を自動的に推定する手法を提案する．本手法では，あらかじめ，活用型ごとに特徴的な後続最長ひらがな列の出現傾向をコーパスから学習しておく．活用型の推定時には，その活用型の用例の出現回数を仮定する．そして，出現が期待される特徴的な後続最長ひらがな列がどの程度観察されるかに基づいて，当該活用型をとるか否かを認定する．本稿では，既知語を用いた本手法の性能評価，およびカタカナ文字列に対する活用型の推定実験の結果について報告する．

キーワード　活用型，後続ひらがな列，新語，転用（転換），派生

Identifying Conjugation Types of Japanese Words Based on Succeeding Hiragana Characters

Tsunenori KUWAE†, Satoshi SATO†, and Atsushi FUJITA†

† Graduate School of Engineering, Nagoya University　Furo-cho, Chikusa-ku, Nagoya, 464-8603, Japan
E-mail: †kuwae@sslab.nuee.nagoya-u.ac.jp †{ssato,fujita}@nuee.nagoya-u.ac.jp

Abstract　In order to compile a dictionary of new words, identification of their syntactic categories and conjugation types is necessary. In Japanese language, Hiragana characters that appear just after a Gokan (word stem) are very strong clues for conjugation identification, because each conjugation type has a different set of conjugation forms with Hiragana suffixes. This paper presents a method of identifying conjugation types of Japanese words based on their succeeding Hiragana characters and describes two empirical experiments: performance evaluation using existing words as pseudo held-out, and conjugation identification of unknown Katakana words.

Key words　conjugation type, succeeding Hiragana characters, new word, conversion, derivation

1　はじめに

言語の経時変化の一つに，新語の出現がある．たとえば,「ハズい」,「キモい」,「コピる」などのいわゆるカタカナ用言は，若者を中心にかなり定着してきている．一方で,「ちゃち（な）」のような形容動詞が,「ちゃちい」,「ちゃちくない」のような形容詞型活用形で使用される例も見られるようになってきた．これは，誤用とみなすことができるが，そのような使用が定着したのであれば，転用あるいは派生とみなさざるをえない．このような変化も，言語の経時変化の一形態である．

上記のような新語や転用・派生語を収集して辞書を作成する場合，それらの語の品詞や活用型を認定することが必要となる．人間用の辞書の場合は，品詞によって見出し形が定まるため，品詞の認定が不可欠である．たとえば,「ちゃち」に形容詞を認めるのであれば,「ちゃちい」という見出しを立てることになる．一方，形態素解析用の辞書の場合は，活用する語に対しては，品詞だけでなく活用型の情報が不可欠である．

このような品詞・活用型推定を自動化するために，本稿では，与えられた語幹に対して，その活用型を推定する方法を提案する．提案手法は，既存の福島ら [1] の方法と同様に，語幹に後続するひらがな列に着目するが，その手法は大きく異なる．

以下，まず 2 節で，活用型の認定と後続ひらがな列の利用について議論する．3 節では，提案手法について説明する．4 節では，既知語を用いた提案手法の性能評価について，5 節ではカタカナ用言の活用型推定結果について述べる．最後に，6 節でまとめを述べる．

2 活用型の認定と後続ひらがな列の利用

2.1 活用型の認定

我々は,「語幹 x に活用型 c の活用語尾が観察される」ことをもって,「語幹 x が活用型 c をとる」と認定する．たとえば,「イナタいサウンド」,「イナタくてかっこいい」,「あのギターはイナタかった」のような用例が観察された場合,「イナタい」は形容詞型の活用型をとると認定する．

2.2 問 題 点

上記の方針に沿って，活用型認定を自動化する場合,

(1) 語幹 x の活用語尾（の分布）を観察する

(2) 活用型 c の活用語尾が十分観察されるならば，x が活用型 c をとると認定する

という手順をふむことになる．しかしながら，ここで，いくつかの問題が発生する．1 つ目の問題は活用語尾の認定の難しさで，それは次のような要因による．

- 活用語尾の終了点が不明（活用体系に依存する）
  例:「買ったのは」,「買わない」
- 活用語尾と助動詞・助詞とを明確に区別できない
  例:「リアルだ」と「パソコンだ」,「確実に」と「彼に」
- 活用語尾と派生語尾を明確に区別できない
  例:「イナタい」と「イナタさ」,「太い」と「太る」

2 つ目の問題は，ある語幹が複数の活用型をとる場合，我々が観察できるのは，複数の活用型の活用語尾が混在した分布のみであるという問題（**ブレンド問題**）である．これは，下記のような要因による．

- 複数の活用型の混在（語幹がたまたま同じである別の語の活用語尾が混在）
  例:「エグい」と「エグる」,「鳴る」と「鳴く」
- 転用・派生による混在
  例:「アホ（名詞）」
  →「アホな（形容動詞）」→「アホい（形容詞）」

もし支配的な活用型のみを推定したいのであれば，ブレンド問題はそれほど深刻な問題ではない．しかし，転用・派生によって出現する活用型も推定したいのであれば，非支配的な活用型も発見する必要があり，ブレンド問題を解決する必要がある．

2.3 方 針

上記の問題をふまえ，本稿で提案する手法では，次に示す方針で活用型を推定する．

(1) 活用語尾を厳格に認定するのではなく，後続最長ひらがな列を利用する（観察する）．

(2) ある活用型の用例が N 回出現していると仮定した場合に，出現すると期待される後続最長ひらがな列がどの程度観察されるかによって活用型を推定する．N を小さくすることにより，非支配的な活用型も認定できるようにすることを目指す．

3 提 案 手 法

本節では，後続ひらがな列に基づいて活用型を推定する手法について述べる．

3.1 各活用型に特徴的な後続ひらがな列

活用型ごとに頻出する後続ひらがな列は異なる．我々は，ある活用型 c に対して他の活用型よりも出現しやすい後続ひらがな列 h を **c の特徴ひらがな列**とし，各活用型を認定する際の手がかりとする．各活用型 c の特徴ひらがな列の集合 $H(c)$ は次の手順で決定する．

(1) コーパスから，各活用型の語の語幹に後続する最長ひらがな列 h を抽出する．その際，ひらがなが後続しない場合は $h = \phi$，語幹の直後に読点が続く場合(注1)は $h =$“、”とする．

(2) 活用型 c ごとの特徴ひらがな列の集合 $H(c)$ を次式で決定する．

$$H(c) = \{h \mid P(h|c) - P(h|\bar{c}) \geqq 0.001\}$$

ここで，$P(h|c)$，$P(h|\bar{c})$ はそれぞれ次のことを表す．

$P(h|c)$ = “h が活用型 c の後続最長ひらがな列として現れる確率”

$P(h|\bar{c})$ = “h が c 以外の活用型の後続最長ひらがな列として現れる確率”

活用型を推定する際は，上記の $H(c)$，$P(h|c)$，$P(h|\bar{c})$ を用いる．

3.2 活用型推定

いま，語幹 x に対して特徴ひらがな列 h が，$freq(x,h)$ 回観察されたとしよう．

ここで，語幹 x がある活用型で出現する回数を N と仮定する．

(1) もしその活用型が c であれば，h の期待される出現回数 $u(c,h,N)$ は，次式で与えられる（小数点以下は切り捨て）．

$$u(c,h,N) = \lfloor N \cdot P(h|c) \rfloor$$

(2) もしその活用型が c 以外の場合，h の期待される出現回数 $l(c,h,N)$ を，次式で見積もる（小数点以下は切り上げ）．

$$l(c,h,N) = \lceil N \cdot P(h|\bar{c}) \rceil$$

次に，実際に観察された $v = freq(x,h)$ が，l から u のどこに位置するかを次式で求める．

$$e(c,h,N,v) = \max\Big(\min\big(u(c,h,N), v\big) - l(c,h,N), 0\Big)$$

e が大きいほど活用型 c をとることを支持し，e が小さい場合は支持しない．

このような考え方に基づき，語幹 x が活用型 c をとることのスコアを次式で定義する．

$$score(x,c,N) = \frac{\sum_{h \in H(c)} e\big(c,h,N,freq(x,h)\big)}{\sum_{h \in H(c)} \big(u\,(c,h,N) - l\,(c,h,N)\big)}$$

このスコアの値域は [0,1] である．最終的に，上記のスコアが閾値 th_s 以上となった場合に，語幹 x に活用型 c を認定する．

(注1)：一段動詞の連用形は「受け入れ、...」のように読点を後続しやすいため．

(a) 公園/を/通っ/て/帰る → ⟨ 通,って,五段・ラ行 ⟩
(b) 電話/を/通し/た/会話 → ⟨ 通,した,五段・サ行 ⟩
(c) 塾/に/通う/子供 → ⟨ 通,う,五段・ワ行促音便 ⟩
(d) 通/と/見える/人達 → ⟨ 通,と,名詞-一般 ⟩

図 1 語幹「通」を持つ語からの正解データの作成

4 既知語を用いた性能評価

本節では，提案手法の客観的評価について述べる．具体的には，コーパス中に実際に出現した既知語の活用型を真の活用型とみなし，提案手法によるそれらの推定性能を評価する．

4.1 問 題 設 定

まず，与えられたコーパスから，次の 3 つ組で表される正解データを作成する．

⟨ 語幹,後続ひらがな列の頻度つきリスト,活用型 ⟩

具体的には，与えられたコーパスを形態素解析し，基本形を参照して語幹と後続ひらがな列の境界を付与する．そして，次の条件を満たす語幹を推定対象として抽出する．

- 漢字で始まる語幹
- 推定対象とする活用型（後述）のいずれかがコーパスに 1 回以上出現

ここで，表層上で語幹が一致する，図 1 のような語の集合について考える．図中の (a)「通 [とお]（る）」と (b)「通 [とお]（す）」は語幹の読みが同じである別の語である．このような現象は，カタカナ用言にも存在する．一方，これらと (c)「通 [かよ]（う）」や (d)「通 [つう]」では語幹の読みが異なる．カタカナ用言にはこのような語幹の読みの曖昧性はない．一般に新語の語幹の境界および読みは自明でないので，提案手法の評価にあたっては，(a)〜(d) の表現は，すべて同じ語幹「通」の出現とみなすことにした．すなわち，(a)〜(d) すべての後続ひらがなが混在したものから活用型を推定する．

福島ら [1] はカタカナ文字列を語幹とする用言のみを収集の対象としていたが，我々の手法はカタカナ用言以外にも適用できる．また，既知語，カタカナ語ともに,「する」を後続してサ変動詞として振る舞う語幹が比較的多数観察される．これらをふまえ，次の 15 種類の活用型を推定対象とする．

動詞・一段，動詞・五段・各行 11 種(注2)，
形容詞，形容動詞，サ変名詞

4.2 データセット

各活用型の特徴ひらがな列は，毎日新聞 2005 年度版を形態素解析器 ChaSen(注3) によって解析した結果から定めた．新聞記事をコーパスとして使う理由は次の 2 つである．

- 各活用型について，規範的用法から逸脱した用例が少ないと見込まれる
- 形態素解析器のモデルが新聞記事を用いて推定されており，語幹および後続ひらがな列を抽出する際に誤りが生じにくいと考えられる

(注2)：カ行イ音便，カ行促音便，ガ行，サ行，タ行，ナ行，バ行，マ行，ラ行，ワ行ウ音便，ワ行促音便．
(注3)：http://chasen-legacy.sourceforge.jp/

表 1 毎日新聞 2005 年版から推定した各活用型の特徴ひらがな列

活用型 c	出現数	異なり語数	$\lvert H(c)\rvert$
一段	550,069	2,766	69
五段・カ行イ音便	101,488	505	63
五段・カ行促音便	8,355	4	49
五段・ガ行	13,618	91	9
五段・サ行	150,585	1,120	65
五段・タ行	37,561	131	41
五段・ナ行	2,113	4	15
五段・バ行	23,900	66	32
五段・マ行	70,022	516	58
五段・ラ行	294,062	1,048	60
五段・ワ行ウ音便	423	13	12
五段・ワ行促音便	175,315	453	66
形容詞	165,764	890	54
形容動詞	278,219	2,060	26
サ変名詞	2,269,758	7,486	23
一般名詞	4,421,102	33,137	-
固有名詞	1,853,691	42,187	-
合計	10,416,045	92,837	-

後続ひらがな列の抽出にあたっては，特殊な活用語尾を持つ「ない」,「ある」,「する」,「なる」を除外して条件つき確率 $P(h|c)$，$P(h|\bar{c})$ を求め，全体で 100 回以上出現した 1,821 種類の後続ひらがな列のみを $H(c)$ の決定に用いた．なお，当該活用型以外の活用型 $\bar{c}$ として，一般名詞，固有名詞も用いた．コーパス中，各活用型と解析された形態素数，語の異なり数，およびそれらに基づいて認定された特徴ひらがな列の異なり数 $|H(c)|$ を表 1 に示す．

一方，推定対象とする既知語データは，毎日新聞 2004 年版を用いて作成した．ChaSen によって形態素解析した結果から 4.1 項の条件を満たす語幹を取り出した．そのうち 50 回以上出現した 5,286 種類の語幹を推定の対象とする．それぞれの語幹に対してある活用型の用例が出現した回数を M とするとき，$M \geqq th_M$ なる活用型を正解（推定すべき活用型）とみなす．

4.3 活用型の推定結果

本実験では，次の 3 つのパラメタを設定する必要がある．

- 検出の際に仮定する出現頻度 N
- スコアに対する閾値 th_s
- 正解とみなす活用型の出現頻度 M に対する閾値 th_M

th_M および N については，32，64，128，256，512，1024 の 6 種類，th_s を 0.1 から 1.0 まで 0.1 刻みの値を検討対象とした．

N を変化させることにより，活用型認定に使用する特徴ひらがな列の集合 $H(c)$ が変化する．すべての活用型に対して，この集合は 1 つ以上の要素を持つ必要がある．$N = 32$ では $|H(\text{サ変名詞})| = 0$ となるため，$N \geqq 64$ となることが不可欠である．一方，非支配的な活用型を見つけるためには，N は小さいほど良い．以上により，$N = 64$ を採用する．この場合の各活用型の特徴ひらがな列を表 2 に示す．

活用型の真の出現頻度よりも低い出現頻度を仮定するのが自然である．ゆえに，$th_M = 128$ とする．

$N = 64$，$th_M = 128$ のもとで th_s を変化させたときの再現率，精度，F 値を表 3 に示す．th_s は，再現率と精度のバランスを決めるパラメタであり，実験結果からも，値を大きくするほど高精度で低再現率，小さくするほど高再現率で低精度とな

表 2　$N = 64$ の場合の各活用型の全特徴ひらがな列
各ひらがな列のうち，太字は複数の活用型に共通，
“(∗)” 付きはスコア $\geqq 0.7$ に必須，下線はそれのみでスコア $\geqq 0.7$ になりうる．

活用型	特徴ひらがな列（重みの降順）
一段	た (∗)，**る**，て，、(読点)，ている，ません
五段・カ行イ音便	**く** (∗)，いた，**き**，いて，かれた
五段・カ行促音便	**った** (∗)，**く** (∗)，かない，**き**
五段・ガ行	ぐ (∗)，ぎ，いで，いだ，いでいる，ぐため
五段・サ行	す (∗)，**した** (∗)，し，して，している
五段・タ行	つ (∗)，ち，**った**，**って**，っている，たない
五段・ナ行	**んだ** (∗)，ぬ
五段・バ行	ぶ (∗)，**んだ**，び，ばれた，ばれる，**んで**
五段・マ行	む (∗)，**んだ**，み，**んで**，んでいる，まれた，まれる
五段・ラ行	**った** (∗)，**る** (∗)，り，ると，**って**
五段・ワ行ウ音便	**う** (∗)，いの，**い**
五段・ワ行促音便	**う** (∗)，**った**，**い**，**って**，われた
形容詞	**い** (∗)，**く**
形容動詞	な (∗)，に (∗)
サ変名詞	**した** (∗)，を，する

表 3　$N = 64$，$M = 128$ における再現率，精度，F 値
T_H: 正しい活用型の集合，T_S: 認定した活用型の集合，
T_{HS}: 正しく認定できた活用型の集合．

th	$\lvert T_H \rvert$	$\lvert T_S \rvert$	$\lvert T_{HS} \rvert$	再現率	精度	F 値
1.0	3,413	2,007	1,798	0.527	0.896	0.663
0.9	3,413	2,573	2,268	0.665	0.881	0.758
0.8	3,413	2,887	2,370	0.694	0.821	0.752
0.7	3,413	3,812	2,705	0.793	0.710	0.749
0.6	3,413	4,682	2,782	0.815	0.594	0.687

表 4　$N = 64$，$th_s = 0.7$ による各 M の範囲の再現率
T_H: 正しい活用型の集合，T_{HS}: 正しく認定できた活用型の集合．

M の範囲	区間			累積		
	$\lvert T_H \rvert$	$\lvert T_{HS} \rvert$	再現率	$\lvert T_H \rvert$	$\lvert T_{HS} \rvert$	再現率
$M \leqq 1,024$	890	815	0.916	890	815	0.916
$512 \leqq M < 1,024$	606	511	0.843	1,496	1,326	0.887
$256 \leqq M < 512$	829	655	0.790	2,325	1,981	0.852
$128 \leqq M < 256$	1,088	724	0.665	3,413	2,705	0.793
$64 \leqq M < 128$	1,193	584	0.490	4,606	3,289	0.714
$32 \leqq M < 64$	561	164	0.292	5,167	3,453	0.668

ることが確認された．$0.7 \leqq th_s \leqq 0.9$ においては，F 値がおよそ 0.75 ほどで，安定した性能が得られた．

4.4 考　　察

この節では，$th_M = 128$，$N = 64$，$th_s = 0.7$ としたときの，推定もれ，推定誤りについて考察する．

4.4.1 推定もれの分析

$N = 64$，$th_s = 0.7$ としたときの，M の範囲ごとの再現率を表 4 に示す．この表から，実際の出現頻度が高いほど推定もれが少ないことが分かる．提案手法では，適当な出現回数 N を仮定し，特徴ひらがな列の期待値と頻度に基づいて活用型を認定するため，この結果は予測通りである．ただし，M が N より大きい場合でも，M が小さくなるほど急激に再現率が低下している．このことから，十分高い頻度で出現する活用型でも，後続ひらがな列が偏っている場合があると推測される．

$th_M = 128$ に対する推定もれ 708 件の活用型ごとの分布を表 5 に示す．推定もれの大部分を占めたのが，サ変名詞と形容動詞である．これらは，そもそもは名詞であるが，サ変動詞や形容動詞としても用いられる可能性があるために，一般名詞と区別されている(注4)ものである．したがって，用言としての用例が観察されなければ，認定できなくても当然である．

サ変名詞の認定には「した」を 2 度以上後続する必要があるが，推定もれの 535 件中 460 件は 1 度以下しか後続していなかったため，サ変名詞と認定されなかった．ちなみにこの中には，「見舞い」，「集め」，「扱い」，「届け出」，「勝ち越し」，「取り締まり」など，和語の連用形が名詞として使われるものも 15 件含まれていた．

形容動詞の認定には「な」，「に」の両方がある程度観察される必要がある．認定できなかった語幹は，次に示すように，観察される後続ひらがなに偏りがあるものであった．

- 「な」のみを後続: 「貴重」，「重大」，「巨大」，「著名」など
- 「に」のみを後続: 「個別」，「不振」，「大量」など
- 「だ」のみを後続: 「必至」，「未定」，「不服」など

その他の推定もれについては，以下に例を示すに留める．

- 一段: 「許せ」，「動け」，「作れ」，「隠せ」などの可能動詞
- 五段・サ行: 「知ら」，「聞か」，「立た」，「悩ま」など，使役の一段動詞の「せる」が「す」に縮退したもの
- 五段・ラ行: 「殺 [や]」，「隠 [こも]」，「与 [あずか]」など．「強/殺/容疑」，「隠/避/容疑」，「与/四死球」と形態素解析された結果，やむなくこのような活用型が割り当てられていたが，実際はそのような活用型は現れていなかった．

今回の実験データには，4.1 項で取り上げた「通」のような，異なる複数の活用型に対応する語幹が 50 件(注5)存在した．これらのうち，33 件については，すべての活用型を認定することができた．したがって，2.2 項で述べたブレンド問題に対して，提案手法はある程度うまく作用したと言える．残りの 17 件（すべて 2 種類の活用型を持つ語幹）の内訳は，2 件がすべての活用型を推定できず（サ変名詞と五段・サ行），頻度が高い方の活用型の推定もれは 4 件，頻度が低い方の活用型の推定もれは 11 件であった．

4.4.2 推定誤りの分析

$th_M = 128$ に対する推定誤りの分布を表 6 に示す．誤って推定された 1,107 件の活用型のうち，773 件は実際にそのような活用型の用例が存在していた．したがって，推定したこと自体は誤りではない．一方，残る 334 件は，一度もその活用型の用例が現れていなかったのにその活用型と認定された，重大な推定誤りである．このような誤りを生じた主な原因は下記の 2 つである．いずれも，複数の活用型に共通の特徴ひらがな列（表 2 中の “(∗)”）の存在によるものである．

表 5　活用型ごとの推定もれの数と率（$th_M = 128$，$N = 64$）

活用型	$\lvert T_H \rvert$	推定もれ	
		$th_s = 0.7$	$th_s = 0.1$
一段	366	38 (10%)	2 (1%)
五段・カ行イ音便	68	2 (3%)	0 (0%)
五段・カ行促音便	1	0 (0%)	0 (0%)
五段・ガ行	14	0 (0%)	0 (0%)
五段・サ行	151	27 (18%)	5 (3%)
五段・タ行	20	2 (10%)	0 (0%)
五段・ナ行	1	0 (0%)	0 (0%)
五段・バ行	14	1 (7%)	0 (0%)
五段・マ行	80	3 (4%)	0 (0%)
五段・ラ行	192	22 (11%)	7 (4%)
五段・ワ行ウ音便	0	0 (0%)	0 (0%)
五段・ワ行促音便	90	5 (6%)	1 (1%)
形容詞	108	1 (1%)	1 (1%)
形容動詞	282	72 (26%)	8 (3%)
サ変名詞	2,026	535 (26%)	106 (5%)
合計	3,413	708 (20.7%)	130 (3.8%)

(注4)：ChaSen で採用している IPA 品詞体系では，それぞれ『名詞-サ変接続』，『名詞-形容動詞語幹』．

(注5)：$M \geqq 128$ なる活用型．46 件が 2 種類，4 件が 3 種類の活用型を持つ．

表 6　活用型ごとの推定誤りの数と率（$N = 64$，$th_s = 0.7$）

活用型	$\|T_S\|$	推定誤り $th_M = 128$	$th_M = 1$
一段	433	105 (24%)	1 (0%)
五段・カ行イ音便	90	24 (27%)	0 (0%)
五段・カ行促音便	6	5 (83%)	4 (67%)
五段・ガ行	16	2 (13%)	0 (0%)
五段・サ行	158	34 (22%)	0 (0%)
五段・タ行	24	6 (25%)	0 (0%)
五段・ナ行	114	113 (99%)	113 (99%)
五段・バ行	14	1 (7%)	0 (0%)
五段・マ行	106	29 (27%)	0 (0%)
五段・ラ行	238	68 (29%)	0 (0%)
五段・ワ行ウ音便	92	92 (100%)	87 (95%)
五段・ワ行促音便	97	12 (12%)	0 (0%)
形容詞	215	108 (50%)	65 (30%)
形容動詞	307	97 (32%)	3 (1%)
名詞-サ変接続	1,902	411 (22%)	61 (3%)
合計	3,812	1,107 (29.0%)	334 (8.8%)

第 1 の原因は，他の活用型にも共通かつ，それのみでスコアが 0.7 以上となるような特徴ひらがな列である．次に示すように，3 種類の活用型について多くの推定誤りを生じた．

- 五段・ナ行（113 件）：「んだ」が支配的
  - 五段・マ行（97 件）：「組」,「進」,「含」,「読」など
  - 五段・バ行（15 件）：「選」,「学」,「飛」,「呼」など
  - 一段（1 件）：盛 [も]（盛 [さか] んだの解析誤り）
- 五段・ワ行ウ音便（87 件）：「う」が支配的
  - 五段・ワ行促音便（87 件）：
    「会 [あ]」,「行 [おこな]」,「言」,「使」,「問」など
- 形容詞（64 件）：「い」が支配的
  - 五段・ワ行促音便（64 件）：
    「会 [あ]」,「行 [おこな]」,「言」,「使」,「問」など

第 2 の原因は，観察できるひらがな列のブレンド問題である．複数の活用型を持つ語幹が，各活用語の特徴ひらがな列を持つことにより，誤って，本来持たないはずの別の活用型に認定してしまう場合があった．

- サ変名詞（61 件）：「した」,「を」,「する」
  - 五段・サ行（40 件，一段と 3 件重複）：
    「面」,「愛」,「訳」,「題」,「害」など
    - 五段・サ行から「した」
    - サ変・－スルから「する」
    - 名詞から「を」
  - 一段（13 件，五段・サ行と 3 件重複）：複合語の後件となった場合にサ変名詞として機能する語幹など
    「出 [で]」,「負け」,「離れ」,「投げ」,「付け」など
  - その他（11 件）：名詞-接尾-助数詞でもある語幹など
    「戦 [せん]」,「勝 [しょう]」など
- 五段・カ行促音便（4 件）：「った」,「く」,「かない」,「き」
  - 「鳴」,「割 [わ，さ]」：
    - 五段・ラ行から「った」
    - 五段・カ行イ音便から「く」,「き」,「かない」
  - 「弱」,「太」：
    - 五段・ラ行から「った」
    - 形容詞から「く」

4.5　課題と対策

複数の活用型の後続ひらがな列が混在したものしか観察できないことが原因で，

- 用例が少ない活用型を推定できない
- 実際に出現していない活用型であると推定してしまう

という問題が生じる．この問題をいかに軽減するかが，活用型推定の中心的課題である．

前者の問題については，個々の活用型の特徴ひらがな列のみを参照することで，特定のひらがな列しか後続しない語幹でない限りは，ある程度低頻度の活用型でも推定できることが確認できた．一方，後者の問題については，

- 支配的な特徴ひらがな列がある場合に，それを共有する他の活用型の語をその活用型に認定してしまう
- 複数の活用型それぞれの後続ひらがな列（特徴的でないものも含む）が混ざることで，本来は存在しない活用型と推定されてしまう

という悪影響を避ける必要がある．

この問題に対して，今後は，次の 4 つの対策を検討する．

データの大規模化: 五段・ワ行ウ音便，五段・ナ行は，他の活用型に比べて出現数が格段に少ない（表 1）．したがって，より大規模なコーパスに基づいて特徴ひらがな列，およびそれらの重みを定めることで，支配的な特徴ひらがな列の影響を小さくできる可能性がある．

柔軟なパラメタ選択: 各パラメタを固定した上で提案手法の特性を分析したが，語幹そのものの出現頻度に応じて N を，活用型ごとに th_s を変えても良い．

階層的活用型認定: 共通の特徴ひらがな列を持つ複数の活用型をまとめて扱い，まずはその活用型クラスとしての認定，その後，共通の特徴ひらがな列以外の情報に基づいて各活用型の認定を行う．活用型クラスの設定にあたっては，まずは後続ひらがな列全体の相関を観察する必要がある．

活用型間の関係を捉える規則の導入: 論文冒頭の「ちゃち」の転用の例や，派生語の研究成果 [2] に示されるように，既存の語の用法の変化には規則性がある．また，4.4.2 項の「鳴」の例のように，どの活用型の組を持つ場合にどの活用型が誤推定されるかについても規則性がある．提案手法は個々の活用型を独立にしか見ないので，活用型間のこのような関係を捉えることで，誤りを削減できると期待できる．

5　カタカナ文字列の活用型推定

「ハズい」,「キモい」,「コピる」のようなカタカナ用言は，多くの場合，活用語尾はひらがなで記述される．しかし，一方で,「イタイ」のように活用語尾もカタカナで記述されるものも存在する．

ここでは，活用語尾はひらがなで記述されることを仮定し，カタカナ文字列に対し 4 節で用いた 15 種類の活用型の有無を推定する．

5.1　データセット

カタカナ文字列およびそれらの後続ひらがな列は，河原ら [3] のウェブコーパス(注6)から，正規表現によって抽出した．抽出し

(注6)：日本語のウェブページから収集された約 4.6 億文．

表 7　出力したカタカナの活用型の内訳
“(†)” 付きは福島ら [1] が対象としていた活用型である．
また，正しいと思われる例がなかった活用型には “△” を記してある．

活用型 c	獲得数	例
形容動詞 (†)	6,056	アブラギッシュな，アフォーな，カッペな
サ変名詞	4,865	クラチェンする，デゴルジュマンする
五段・サ行	231	ゴマカす，メザす，フカす，ボカす，カマす
形容詞 (†)	203	イナタい，モサい，シャバい，イモい
五段・ラ行 (†)	166	モニョる，ホゲる，クホる，ネグる，バビる
一段 (†)	157	イケテる，カシメる，バックレる，ガレる
五段・ナ行	44	△
五段・ワ行ウ音便	39	△
五段・ワ行促音便	34	トモナう，チガう，ニオう
五段・カ行イ音便 (†)	28	ワメく，イタダく，ホザく，シバく，サビく
五段・カ行促音便	18	△
五段・タ行	16	モつ，ブつ，マつ
五段・マ行	15	タレコむ，ツカむ，リキむ
五段・バ行	1	トぶ
合計	11,873	

た全 6,711,001 種類のカタカナ文字列のうち，50 回以上出現した 218,415 種類のカタカナ文字列を活用型推定の対象とする．

5.2　カタカナ用言の出力例

4 節の実験によって定めたパラメタ（$N = 64$，$th_s = 0.7$）を用いて推定された活用型は，9,935 語幹に対する 11,873 件であった．活用型ごとの獲得数および例を表 7 に示す．この表に示すように，本実験において，我々が普段用いているカタカナ用言が多数見つかった．また,「イナタ（い）」,「クラチェン（する）」のように，出現回数の少ない（それぞれ 84 回，77 回）カタカナ用言も認定することができた．

福島らは，表 7 の “(†)” の活用型を対象として，3,188 語のカタカナ用言を獲得した．同じ活用型に対して我々が認定した語は 6,610 語であり，前節の 70%という推定精度を考慮しても，福島らに比べて大規模なカタカナ用言を獲得できたといえる．

出力されたカタカナ用言は 4 つに分類することができる．

(a) 既知語のカタカナ表記
例:「アブナい」(危ない),「クサる」(腐る),「タイセツな」(大切な)

(b) (a) 以外の異表記: 表記の誤りなのか意図的に表記を変えたものが広まっている語であるかはわからない
例:「ダウソロードする」(ダウンロードする),「トンヅラする」(トンズラする),「クローバルな」(グローバルな)

(c) 真の未知語: 既知語の異表記でない語
- 外国語由来の語
例:「サチる」(saturate),「ウェルドする」(weld)
- 専門用語
例:「サビく」(波止のサビキ釣りで竿を上下に動かす),「ラグる」(オンラインゲームでタイムラグが生じること)
- その他
例:「タコい」,「アブラギッシュな」

(d) 推定誤りと思われるもの: “?” を記すとともに，括弧内に誤って推定した活用型を示す
例:「? バラシる」(一段),「? バイトす」(五段・サ行),「? グロく」(五段・カ行促音便)

複数の活用型の語幹として出力されたカタカナ文字列は 1,497 件であった．例を以下に示す．ただし，推定誤りと思われる語に対しては “?” を記してある．

イキ　「イキな」(形容動詞),「イキる」(五段・ラ行),「イキる」(一段),「イキする」(サ変名詞)

オモ　「オモな」(形容動詞),「オモう」(五段・ワ行促音便),「? オモう」(五段・ワ行ウ音便),「オモい」(形容詞)

カタ　「カタい」(形容詞),「カタる」(五段・ラ行)

ラグ　「ラグる」(五段・ラ行),「ラグな」(形容動詞),「ラグい」(形容詞)

アプ　「アプする」(サ変名詞),「アプる」(五段・ラ行)

タコ　「タコな」(形容動詞),「タコる」(五段・ラ行),「タコい」(形容詞)

認定したカタカナ用言を定量的に評価するには，推定された活用型が正しいかどうかを判断する必要がある．しかしながら，対象が新語であり，信頼しうるリファレンスが存在しないため，その判断が難しい．

現時点で観察されている傾向は以下の通りである．

- 既知語に対する活用型推定の場合と同様,「五段・ナ行」,「五段・ワ行ウ音便」,「五段・カ行促音便」などの特定の活用型の推定誤りが多い．
- 高頻度の語幹ほど推定誤りが多い．大規模なコーパスを用いたため顕現した特徴である．

6　おわりに

本稿では，語幹の後続ひらがな列の分布が活用型ごとに異なることを利用して，与えられた語幹の活用型を推定する手法を提案した．既知語を用いた性能評価の結果，活用型ごとに特徴的な後続最長ひらがな列のみに着目することで，支配的な活用型だけでなく，非支配的な活用型もある程度認定できることが確認できた．一方，カタカナ文字列に対する活用語推定の実験では，10,000 件以上のカタカナ用言を獲得できた．既知語に対する約 70%という精度から推定できる正しいカタカナ用言の数は約 7,000 件であり，既存の福島らの研究で得られた数を凌駕している．

今後は，4.5 項で述べた各手法を実装して，ブレンド問題による推定誤りの解消を目指すとともに，カタカナ用言に対する活用型の推定精度を定量的に評価する手法を確立したい．

文　献

[1] 福島健一, 鍛治伸裕, 喜連川優. 機械学習を用いたカタカナ用言の獲得. 言語処理学会第 13 回年次大会発表論文集, pp. 815–818, 2007.

[2] 加藤直樹, 藤田篤, 佐藤理史. 派生語辞書第一版の編纂. 言語処理学会第 14 回年次大会発表論文集, pp. 1053–1056, 2008.

[3] 河原大輔, 黒橋禎夫. 高性能計算環境を用いた Web からの大規模格フレーム構築. 情報処理学会研究報告, NL-171-12, pp. 67–73, 2006.