コーパスに基づく現代語表記のゆれの調査
― BCCWJ コアデータを資料として ―

小椋秀樹（国立国語研究所言語資源研究系）

Corpus-Based Survey of the Orthographic Variation in Contemporary Japanese: Analysis of the BCCWJ-Core

Hideki Ogura（Dept. Corpus Studies, NINJAL）

１．はじめに

音節，語など，種々の言語単位において，形式が一つに定まらず，複数の形式が許容されることがある。この複数の形式が共時的に存在する現象を「ゆれ」と呼ぶ。

語のレベルにおけるゆれには，語形やアクセントのゆれのほか，日本語においては表記のゆれが多く見られる。例えば，「俺－おれ」「さくら－サクラ」のような異なる文字体系間の対立によるゆれのほか，「付属－附属」のような異なる漢字の対立によるゆれ，「行う－行なう」のような送り仮名の違いによるゆれ等がある。また，「上げる－挙げる－揚げる」のような異字同訓も，それぞれを別語とせず同一の語と見なした場合，動詞《アゲル》に「上げる」「挙げる」「揚げる」という複数の表記が共時的に存在すると捉えられ，《アゲル》の表記のゆれとして扱うことができる。

日本語の語表記のゆれについては，これまでに次の三つの調査が行われている。

宮島達夫（1997）　：　1956 年発行雑誌 90 種の調査
国立国語研究所（1983）　：　1966 年発行朝日・毎日・読売 3 紙の調査
国立国語研究所（2006）　：　1994 年発行雑誌 70 誌の調査

しかし，これらの調査については，二つの問題点がある。1 点目は，いずれも調査対象が単一の媒体という点である。現代語表記のゆれの実態解明という面からは，複数の媒体を対象に調査を行い，語表記のゆれに媒体差があるのか明らかにする必要がある。

2 点目は，国立国語研究所（2006）の調査対象年（1994 年）から既に 18 年が経過しているという点である。1990 年代から，情報機器の急速な普及に伴って書記環境が大きく変化するとともに，それに伴う漢字使用の増加が指摘されている。その結果，語表記のゆれの実態にも変化が生じていることが予想される。そこで，より現在に近い時期における語表記のゆれを調査する必要がある。

本研究は，『現代日本語書き言葉均衡コーパス』（以下，BCCWJ）のコアデータ(1)を資料として，そこに収録された白書・新聞・雑誌・書籍・Web という五つの媒体を対象に，より現在に近い時期における語表記のゆれの実態を明らかにしようとするものである。

２．調査対象

本研究の目的は，複数の媒体を対象として，より現在に近い時期における語表記のゆれの実態を明らかにすることにある。そこで，複数の媒体のデータを収録した BCCWJ を調

(1)　コアデータの設計・構成等については，小椋秀樹・小木曽智信・小磯花絵ほか（2009）を参照。

査対象とした。ただし BCCWJ 全体を対象とするのではなく，今回はコアデータのみを対象とした。

コアデータは，以下の媒体から成り，延べ語数は，短単位で約 110 万語，長単位で約 84 万語（短単位，長単位とも記号，空白，補助記号を除く。）となっている。

出版サブコーパス　　：2001 年～ 2005 年発行の新聞，雑誌，書籍
特定目的サブコーパス：2001 年～ 2005 年発行の白書
2004 年 10 月～ 2005 年 10 月投稿の Yahoo!知恵袋
2008 年 4 月～ 2009 年 4 月投稿の Yahoo!ブログ

コアデータは，自動形態素解析をした後に，全体に対して人手による確認，修正を行ったデータで，解析精度は長単位・短単位とも約 99%以上である。

BCCWJ を対象とした語表記のゆれの調査では，BCCWJ 全体を対象とすることも考えられるが，コアデータ以外のデータの精度が約 98%と少し低いことから，調査に当たって誤解析がどの程度影響するかが気になるところである。そこで，今回は BCCWJ 全体を対象とした語表記のゆれの調査を行う前の予備調査として，より精度の高いコアデータのみを対象とすることとした。

本研究では，固定長・可変長サンプルの両方を対象とした。また，長短 2 種類のデータのうち短単位を用い，固有名詞・数詞・感動詞・助詞・助動詞・記号・補助記号を除く，いわゆる一般語を対象とした。

３．語表記のゆれの認定方法

本研究では,媒体別・語種別に計量的な観点から語表記のゆれの実態を明らかにするが,その際，どのように表記のゆれを認定し，集計するかが問題となる。以下，本研究における語表記のゆれの認定方法について述べる。

BCCWJ の形態素解析には，形態素解析辞書 UniDic(2)を用いた。UniDic では，表記や語形の違いにかかわらず，同じ語であれば，同一の見出しを与えるという方針を取り，語を階層化した形で登録している。この階層の最上位を語彙素（国語辞典の見出しに相当）と呼んでおり，この語彙素の下に語形（語形の違いを区別する層），更に語形の下に書字形（表記の違いを区別する層）という階層を設けている。

語彙素	語形	書字形
ヤハリ【矢張り】	ヤハリ（副詞）	やはり
		ヤハリ
		矢張り
	ヤッパリ（副詞）	やっぱり
		ヤッパリ
		矢っ張り

図 1：UniDic の階層構造

この階層構造は，BCCWJ の形態論情報にも反映しており，コーパス中の全ての短単位に対して語彙素・語形・書字形という階層的な見出しや品詞情報が付与されている。本研究では，この階層的な見出しを利用して表記のゆれを認定することとした。具体的には，

(2) UniDic の概要については伝康晴・小木曽智信・小椋秀樹ほか（2007）を参照。

「任意の二つの書字形が，同じ語彙素・語形・品詞を持つ場合，同じ語の表記のゆれと認める」とした。つまり，図 1 で言えば，「やはりーヤハリー矢張り」は，同じ語彙素「ヤハリ【矢張り】」，語形「ヤハリ」，品詞「副詞」を持つので，これらは語表記のゆれと認められる。同様に，「やっぱりーヤッパリー矢っ張り」も語表記のゆれと認められる。

語表記のゆれの認定に語形まで含めたのは，同一の語彙素を持つ場合に語表記のゆれと認めるとすると，語形が異なることによる表記の差異も表記のゆれとして扱うことになるからである。例えば，図 1 の書字形欄に掲げた六つの表記は全て《ヤハリ》という語の表記のゆれとなる。しかし「やはりーやっぱり」「ヤハリーヤッパリ」「矢張りー矢っ張り」の対立は，語形の違いによるものであり，本研究では除外する必要がある。

ただし，このようにして認定した語表記のゆれは，語形「ヤハリ」「ヤッパリ」における表記のゆれであり，語の表記のゆれを調査するという目的からは，問題があるという指摘もあろう。しかしながら，本研究では BCCWJ から自動で取得可能な情報を活用することとし，上記のような方法を取った。

このように，語形レベルで集計を行うため，以下に示す語数は，全て語形の数である。表 1 に媒体別の異なり語数・延べ語数を示した。本研究では，Yahoo!知恵袋と Yahoo!ブログとを併せて Web として集計することとした。

表 1：媒体別語数（異なり・延べ）

媒体	異なり	延べ
Web	12,652	93,594
書籍	12,408	105,248
雑誌	14,674	105,610
新聞	15,809	166,630
白書	6,474	120,636

語表記のゆれとして扱う範囲については，冒頭でも述べたように，異なる文字体系間の対立によるゆれ，異なる漢字の対立によるゆれ，送り仮名の違いによるゆれ等のほか，異字同訓も語表記のゆれに含める。また，公用文の表記の基準では，《サラニ》について，副詞は「更に」，接続詞は「さらに」と書き分けることとしている。本研究では，このように基準によって書き分けられているものも，語表記のゆれとして扱う。

４．調査結果

４．１　語表記にゆれの見られる語の割合

まず，媒体別にどの程度の語に表記のゆれが見られるのかを見ていく。語表記にゆれのある語の異なり数，割合を媒体別及び語種別にまとめ，表 2 として示した。

表 2 から，Web・書籍では約 1 割の語に表記のゆれが見られることが分かる。雑誌も 9.0%と Web・書籍に近い割合である。一方，新聞（5.8%）は Web・書籍のおよそ半分程度，白書はそれよりも更に低く，3.3%となっている。このように，語表記のゆれには媒体差が見られ，今回の調査では，語表記のゆれの割合が 1 割程度の Web・書籍・雑誌と 6%以下の新聞・雑誌とに大きく分けられる。

表 2：表記にゆれのある語の割合（媒体別・語種別，異なり）

媒体	異なり	ゆれ	%	語種	異なり	ゆれ	%
Web	12,652	1,299	10.3%	和	4,640	899	19.4%
				漢	5,367	170	3.2%
				外	2,233	198	8.9%
				混	412	32	7.8%
書籍	12,408	1,343	10.8%	和	5,086	1,117	22.0%
				漢	5,852	176	3.0%
				外	1,097	17	1.5%
				混	373	33	8.8%
雑誌	14,674	1,323	9.0%	和	5,189	889	17.1%
				漢	6,622	155	2.3%
				外	2,388	248	10.4%
				混	475	31	6.5%
新聞	15,809	916	5.8%	和	4,944	670	13.6%
				漢	8,330	152	1.8%
				外	2,064	77	3.7%
				混	471	17	3.6%
白書	6,474	211	3.3%	和	1,311	173	13.2%
				漢	4,334	25	0.6%
				外	688	11	1.6%
				混	141	2	1.4%

今回調査した五つの媒体が，このように大きく二分される要因としては，その媒体に共通の表記の基準があるかないかということが考えられる。白書は，全省庁とも常用漢字表，送り仮名の付け方等の国が定めた表記の基準に基づいて書かれている。新聞は，漢字使用に関しては，各社共通の基準として，日本新聞協会が常用漢字表を基に定めた新聞漢字表がある。送り仮名等についても，各社とも国の定めた表記の基準によっている。このように，新聞・白書には共通の表記の基準があるため，語表記のゆれも低く抑えられているものと思われる。

一方，雑誌・書籍については，出版社や雑誌，著者ごとに独自の用字の方針を持っているとしても，新聞漢字表のような各社共通の基準は見られない。特に書籍は，著者個人の自由度が高いと予想される。Web（Yahoo!知恵袋・ブログ）は，どのような表記を用いるかは全く個人の自由である。このように共通の表記の基準を持たないことから，出版社や著者個人による表記の差が多くあり，その結果，語表記のゆれの割合が高くなっているものと思われる。

表 2 には，語種別の語表記のゆれの割合も示した。これを見ると，四つの語種の中で和語が最も語表記のゆれの割合が高い。Web・書籍では 2 割程度にゆれが見られ，媒体別に見た場合にゆれの割合の少なかった新聞・白書でも和語は約 13%にゆれが見られる。一方，漢語は，Web の 3.2%が最も割合が高く，最も割合の低い白書では 0.6%となっている。

外来語については，Web で 8.9%，雑誌で 10.4%と他の媒体に比べて割合が高くなっている点が注意される。これは，外来語が英字で表記されたものも加えていることによるものである。英字表記の外来語を除外して，語表記のゆれの割合を算出したところ，Web は 1.9%（異なり 2,085，ゆれ 40），雑誌は 0.9%（異なり 2,141，ゆれ 20）で，かなり低い割合となる。

４．２　表記の種類数

本節では，媒体別・語種別に，何種類の表記が，それぞれ何語見られるのか見ていくこととする。その結果を，媒体別・語種別に表 3 として示した。

媒体別・語種別のいずれにおいても，表記が 1 種類の語，つまり表記にゆれのない語が最も多く，表記の種類が多くなるほど，語数が少なくなっている。

次に媒体別に見ると，語表記のゆれの割合の高い Web・書籍・雑誌では 3 種類以上の語も比較的多く見られるのに対し，ゆれ割合の低い新聞・白書は，3 種類以上の語は少ない。特に白書は，表記が 3 種類以上の語が 22 語と，全ての媒体の中で最も少ない。

表 3：表記の種類数（媒体別・語種別，異なり）

媒体	語種	表記種類数						
		1	2	3	4	5	6	7
Web	和	3,741	717	139	36	4	1	2
	漢	5,197	141	27	2	0	0	0
	外	2,035	153	40	3	1	1	0
	混	380	26	6	0	0	0	0
書籍	和	3,969	909	172	31	3	1	1
	漢	5,676	158	16	2	0	0	0
	外	1,080	17	0	0	0	0	0
	混	340	31	2	0	0	0	0
雑誌	和	4,300	742	121	23	3	0	0
	漢	6,467	140	13	1	1	0	0
	外	2,140	208	36	4	0	0	0
	混	444	28	3	0	0	0	0
新聞	和	4,274	580	76	11	2	0	1
	漢	8,178	140	12	0	0	0	0
	外	1,987	70	7	0	0	0	0
	混	454	14	3	0	0	0	0
白書	和	1,138	155	18	0	0	0	0
	漢	4,309	21	1	3	0	0	0
	外	677	11	0	0	0	0	0
	混	139	2	0	0	0	0	0

語種別に見ると，和語には表記の種類の多い語が見られる。7 種類の語が Web に 2 語，書籍に 1 語あり，さらにゆれの少ない新聞にも 1 語ある。6 種類の語も Web，書籍にそれぞれ 1 語ずつ見られる。

6 種類の表記を持つ語と 7 種類の表記を持つ語とを，その表記とともに，表 4 として示した。

表 4：表記種類数 6，7 の和語

媒体	見出し	度数	表記（度数）
Web	トル	141	とる（58），取る（50），執る（1），採る（2），撮る（26），獲る（1），録る（1）
	ドウ	366	だう（1），ど～（3），どう（352），どぉー（1），どー（5），ドォー（2），如何（2）
	ワカル	243	わかる（123），ワカル（3），分かる（94），分る（2），判る（15），解る（6）
書籍	カワル	64	かわる（6），代る（1），代わる（3），変る（5），変わる（47），替る（1），替わる（1）
	トル	182	とる（140），取る（32），採る（2），撮る（5），獲る（2），盗る（1）
新聞	トル	150	とる（94），取る（48），執る（1），捕る（1），採る（2），摂る（3），撮る（1）

7 種類の表記が見られる語は，Web では動詞《トル》と副詞《ドウ》，書籍では動詞《カワル》，新聞では動詞《トル》である。6 種類の表記が見られる語は，Web では動詞《ワカル》，書籍では動詞《トル》である。

動詞《トル》《カワル》《ワカル》は，異字同訓の語である。常用漢字表には，《トル》を訓に持つ漢字として「採」「執」「取」「捕」の 4 字が，《カワル》を訓に持つ漢字として「換」「代」「替」「変」の 4 字が掲げられている。《トル》《カワル》ともに意味によっ

てこれら 4 字での書き分けが求められるものであり，元々表記の種類が多くなる可能性のある語と言える。それに加えて，《トル》には「撮る」「獲る」「録る」という表外訓による表記や平仮名表記があり，《カワル》には送り仮名の違いによるゆれと平仮名表記がある。その結果，表記の種類が 7 種類と最も多くなっている。

《ワカル》については，常用漢字表では《ワカル》を訓に持つ漢字として「分」のみを掲げているが，「解る」「判る」という表外訓による表記のほかに，送り仮名のゆれ，平仮名表記，片仮名表記があり，表記の種類が多くなっている。

なお，表記が 5 種類ある和語を見ると，Web では動詞《アラワレル》《カカル》《ススメル》《ツクル》，書籍では動詞《アラワス》《オサエル》《ヒク（他動詞）》であり，これらも異字同訓の語である。

動詞のうち異字同訓の語は，表内字・表外字を含めて書き分けがなされることで元々表記の種類が多く，さらにそこに送り仮名の異なる表記や平仮名表記，片仮名表記が用いられることで，更にゆれが大きくなる傾向があると考えられる。

４．３　動詞の語表記のゆれ

前節で見たように，異字同訓の動詞は，漢字の書き分けに加え，送り仮名の違い等のゆれもあり，表記の種類が多くなっていた。そこで，本節では，動詞に注目し，どのような語表記のゆれの類型があるのかを見ていく。語表記のゆれの類型については，国立国語研究所（1983）で用いられた以下の類型を用いた。なお，この調査では媒体別に集計せず，コアデータ全体でまとめて集計した。

a. 異なる漢字の対立 　例：付属－附属，会う－合う b. 送り仮名の対立 　例：行う－行なう c. 仮名遣いの対立 　例：行う－行ふ d. 外来語表記法の対立 　例：バイオリン－ヴァイオリン	e. 漢字と平仮名の対立 　例：俺－おれ，微妙－びみょう f. 漢字と片仮名の対立 　例：俺－オレ，微妙－ビミョウ g. 平仮名と片仮名の対立 　例：さくら－サクラ h. 文字と記号の対立 　例：国国－国々

コアデータ全体に動詞は異なりで 3,326 語あり，そのうち語表記にゆれのある語は，1,169 語（35.1%）である。動詞の大半は，和語に分類されるが，表 2 に示した和語のゆれの割合を上回っている。動詞は，ゆれの割合の高い語群ということができる。

語表記にゆれのある動詞について，上に示した類型に分類した結果を表 5 として示した。類型 d，h に分類されるものはなかったので，表 5 には，この二つの類型を示していない。なお，動詞の中には，複数の類型に分類されるものがある。例えば，動詞《アタル》の表記として「あたる」「当たる」「当る」の 3 種類がある場合，「当たる－当る」は，「b.送り仮名の対立」に分類され，「あたる－当たる・当る」は「e.漢字と平仮名の対立」に分類される。したがって，動詞《アタル》は類型 b と e との二つに分類されることになる。このような語があるため，各類型の異なり語数の合計が上に示したゆれのある語の異なり語数 1,169 を超えている。

表 5：動詞における語表記のゆれの類型

類型	異なり	%	延べ	%	語　例
a	238	16.5%	33,704	17.1%	暖める－温める，打ち込む－撃ち込む
b	66	4.6%	5,273	2.7%	荒す－荒らす，打ち上げる－打上げる
c	9	0.6%	20,670	10.5%	うなずける－うなづける，考える－考へる
e	1,050	72.7%	88,775	45.1%	あう－合う，あえる－和える
f	33	2.3%	10,140	5.2%	オソレイル－恐れ入る，ハネ上がる－跳ね上がる
g	48	3.3%	38,064	19.4%	するースル，いじめるーイジメる

表 5 を見ると，類型 e に属する語が最も多く，異なりで 72.7%，延べで 45.1%を占める。類型 a がそれに次ぎ，異なりで 16.5%，延べで 17.1%となっている。

次に，単純動詞・複合動詞に分けて集計した結果を表 6，表 7 として示した。

表 6：語表記のゆれの類型（単純動詞）

類型	異なり	%	延べ	%
a	185	18.6%	33,151	17.3%
b	38	3.8%	4,657	2.4%
c	9	0.9%	20,670	10.8%
e	688	69.3%	84,797	44.4%
f	29	2.9%	9,988	5.2%
g	44	4.4%	37,911	19.8%

表 7：語表記のゆれの類型（複合動詞）

類型	異なり	%	延べ	%
a	53	11.8%	553	10.1%
b	28	6.2%	616	11.3%
c	0	0.0%	0	0.0%
e	362	80.3%	3,978	73.0%
f	4	0.9%	152	2.8%
g	4	0.9%	153	2.8%

単純動詞，複合動詞共に，1 位は類型 e，2 位は類型 a となっている。複合動詞では，類型 e の割合が異なり・延べとも動詞全体（表 5），単純動詞（表 6）より高くなっている。

類型 e に属する語を見てみると，単純動詞では度数順で上位から《イル》《アル》《イウ》《ナル》《クル》が挙げられる。これらは，基本動詞であり，また複合辞の構成要素にもなっている語である(3)。複合辞の構成要素については，実質的な意味が薄れていることから，平仮名表記される傾向が見られる。《イル》等が類型 e に属することには，複合辞での使用例も関わっていると思われる。

このほか，動詞《トル》《キク》《ツクル》といった異字同訓の語や，《ツナガル》《マトメル》といった漢字表記した場合に表外漢字となる語がある。

複合動詞では，全体を平仮名表記にしたものや，前項又は後項を平仮名表記したものが見られる。例えば，「とりくむー取り組む・取組む」「くりかえすーくり返すー繰り返す・繰返す」などである。

類型 a に属する語は，基本的に異字同訓の語である。「聞くー聴くーきく」「換わるー代わるー替わるー変わるーかわる」といった常用漢字表内で書き分けるもののほか，「見るー診るー観るーみる」「取るー執るー採るー撮るー獲るー録るーとる」「言うー云うーいう」のように，表外訓・表外字（下線部）が見られるものもある。また，異字同訓については，前節で見たように，多くの場合，平仮名表記も共に用いられている。類型 a に分類されている 238 語のうち，171 語が類型 e にも分類されている。

異字同訓の語で平仮名表記も用いられる要因の一つとして，書き分けの難しさから，平仮名表記が選択されるということが考えられる。例えば，次に挙げる動詞《トル》の仮名

(3) これらの動詞を構成要素に持つ複合辞として，例えば「ている」「てある」「という」「ことになる」「てくる」が挙げられる。

表記例は，いずれも「取」で表記して問題ない例ではあるが，動詞《トル》の中心的な意味用法ではないため，どの漢字で表記するか判断に迷う面があり，平仮名表記が選択された可能性がある。

コミュニケーションをとる手段を身につけさせる（教育再生！）
首相がこうした言動をとることで（読売新聞）
防災に関しとるべき措置と地域防災計画の作成（消防白書）

なお，これらについては，中心的な意味用法ではないため，漢字表記すること自体に違和感があり，平仮名が選択されたという可能性もあろう。異字同訓の語については，どのような意味用法の時に，漢字表記が選択されるのか，また平仮名表記が選択されるのかといったことを調べていく必要がある。

５．終わりに

本研究では，BCCWJ のコアデータを対象に語表記のゆれに関する調査を行った。その結果，以下のことが明らかとなった。

(1) 語表記のゆれには媒体による差異がある。コアデータに収録した五つの媒体については，語表記のゆれの割合の高い Web・書籍・雑誌とゆれの割合の低い白書・新聞とに大きく分けられる。
(2) 語種別に見た場合，和語が最も語表記のゆれの割合が高い。
(3) 動詞は，異なりで約 35%の語に表記のゆれが見られる。類型別に見た場合，平仮名と漢字の対立によるゆれ，異なる漢字の対立によるゆれが多い。
(4) 異字同訓の動詞については，漢字の書き分けにより表記の種類が多く，さらに送り仮名の違いや平仮名表記により，更にゆれが大きくなる傾向が見られる。

今回の調査を基に，他の品詞等についても調査を進めていく必要がある。また，2 節に述べたように，今回の調査は予備調査である。BCCWJ 全体を対象として，語表記のゆれの調査を行い，現代における語表記のゆれの実態をより一層明らかにしていくことが必要である。今後の課題としたい。

謝辞 本研究は，国立国語研究所共同研究プロジェクト（基幹型）「コーパス日本語学の創成」（リーダー：前川喜久雄）による補助を得た。

参 考 文 献

小椋秀樹・小木曽智信・小磯花絵・冨士池優美・宮内佐夜香・渡部涼子・竹内ゆかり・小川志乃・小西光・原裕・中村壮範(2009)「『現代日本語書き言葉均衡コーパス』における形態論情報付与作業の進捗状況」『特定領域「日本語コーパス」平成 20 年度公開ワークショップ（研究成果報告会）予稿集』, pp.57-64.

国立国語研究所（1983）国立国語研究所報告 75『現代表記のゆれ』.

国立国語研究所（2006）国立国語研究所報告 125『現代雑誌の表記― 1994 年発行 70 誌―』.

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵（2007）「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』22, pp.101-123, 国書刊行会.

宮島達夫（1997）「雑誌九十種表記表の統計」『日本語科学』1, pp.92-104, 国書刊行会.