国立国語研究所学術情報リポジトリ

On the statistical method to analyze the writing forms of a word in contemporary Japanese

メタデータ	言語: jpn 出版者: 公開日: 2017-03-31 キーワード (Ja): キーワード (En): 作成者: 佐竹, 秀雄, SATAKE, Hideo メールアドレス: 所属:
URL	https://doi.org/10.15084/00001050

表記のゆれを測る

佐竹秀雄

1. はじめに

日本語の表記には，ひらがな・カタカナ・漢字・ローマ字・数字などが用いられている。このように数種類の文字を用いて書き表すことは，日本語の一つの特徴だと言えるだろう。しかも，これらの文字は単にその種類が多いというだけではない。お互いに関連をもっていて，同じことばを幾通りもの異なった表記形に書ける点に特色がある。そのために，日本語の表記体系は複雑なものとなっている。

このような，日本語特有の複雑な表記体系を記述することは，困難ではあろうが，それだけに重要な課題である。表記体系を記述するためには，表記法を明らかにしなければならない。つまり，日本語を表記する際，どのような場合にどのような文字（表記形）が使われるかの法則を見出さねばならない。

ところが，日本語の表記法はそれほど単純なものではない。その主な原因は表記の安定していないものが存在するところにある。たとえば，最初にあげた五種の文字の使い分けについて考えてみると，傾向らしきものを述べることはできる。しかし，それを絶対的な規則として認めることはできない。例をあげれば，概念を表す部分には漢字が用いられ，外来語や音を表す部分にはカタカナが用いられるということは，傾向としては認められる。が，それを規則と呼ぶまでには致らない。概念を表す語にしても，外来語にしても，それが書き表される表記形態は常に決まっているわけではないからである。

あることばを書き表すときに，どのような表記形態が選ばれるかについては，いくつかの要因が考えられる。個人の習慣的な書きぐせが表記形態を決定

することもあれば，それが書かれ，読まれる場面や文脈が表記に影響を及ぼすこともある。ここで言う場面・文脈には，その表記がなされる文・文章などがどのような性質のものであるかとか，その表記の前後の文字列がどのようなものであるかなどの問題を含んでいる。文・文章のあり方や前後の文字列のあり方は，読み手に与える印象や読みやすさの点で，表記の選択とかかわってくるからである。さらに，実際の具体的な場面では，もっとこまかな要因も働くであろう。そして，それらの要因の影響を受けた結果，ある条件下では常に一つの表記形態に固定されるし，また他の条件下では表記がゆれるということになるのであろう。

そこで，表記法を明らかにするためには，表記の実態とそれを生み出す要因との関係をとらえねばならない。あることばの表記にあたって，〝どのような表記形態が用いられるのか〟〝その表記は安定しているのか，ゆれているのか〟などの実態をまずとらえる。次に，それらの表記が選択され使用された要因をさぐるのである。その際，表記の実態の把握の段階では，表記のゆれが大きいものに目をつけることにする。つまり表記が安定していないものをより出すのである。そして，その分析を通してゆれの要因をさぐり出し，最後には，表記の安定したものをも含めた全体の表記の本質に迫りたいと考える。

ここに述べる論は，その第一段階である。すなわち，表記のゆれの大きいものを選び出すために，表記のゆれを数量化しようとした試みである。

2. 表記のゆれのパタン

「表記のゆれ」といっても，その内容はとらえ方によっていろいろである。一つの文字をどういう字体で書くかといういわば文字レベルの表記のゆれや，一つの語をどう表記するかという語レベルの表記のゆれから，縦書きにするか横書きにするか，手書きか印刷かといった表記行動のレベルでの表記のゆれまで，はばが広い。また観点を変えれば，個人差による個人間の表記のゆれも考えられるし，各個人における場面や文脈間の表記のゆれも考えることができる。

ところで，我々が一般にことばを書き表す場合，音声を忠実に一音一音文字化しようとはしない。何らかのまとまりに分割し，それを単位にして表記を考えると思われる。その単位となるのはほとんどの場合，意味的なまとまりを担った語もしくは語相当のものであると思われる。もちろん，分割された語の内部については，さらにその語を表記するための文字の選択がなされよう。しかし，まず第一段階で表記の単位として意識されるのは語であると考えてよいだろう。

このような理由から，ここでは，語の表記形式のゆれをとりあげることにする。すなわちある一つの語がどのような表記形式で書き表されるかを問題にするのである。一つの語が二種以上の異なる表記形式で書かれているとき，語の表記がゆれているといい，そのゆれの程度をみようというのである。

語の表記形式の種類とその数は，語によっていろいろであるが，整理すれば次のような対立のパタンとしてまとめることができる。

Ⅰ　同一文字体系内での対立によるゆれ

(a)　二種以上の異なる漢字の対立　　例（尚-猶）

(b)　二種以上の異なる字体の文字の対立　（學-学）（峰-峯）

(c)　二種以上の送りがなの対立　　（行なう-行う）

(d)　かなづかい・かな用法上の対立　　（ついに-つひに）

(e)　カタカナの外来語表記上の対立　　（コンベアー-コンベヤ）

(f)　ローマ字のヘボン式表記と日本式表記の対立　（Fuji-Huzi）

(g)　数字で桁の単位の有無の対立　　（一二三四五--一万二千三百四十五）

Ⅱ　異なる文字体系間の対立によるゆれ

漢字・ひらがな・カタカナ・ローマ字・数字などの間での相互の対立

Ⅲ　その他

(h)　くり返し記号と文字との対立　　（人々-人人）

(i)　その他の記号類と文字との対立　　（%-パーセント）

ここではまず文字体系のあり方によって大きく三つに分類した。「Ⅰ　同一

文字体系内での対立によるゆれ」については，さらに代表的な対立パタンをいくつか列挙している。「Ⅱ　異なる文字体系間の対立によるゆれ」は例を省いているが，たとえば「漢字とひらがな」「ひらがなとカタカナ」など文字体系が異なる文字どうしの間での対立によるゆれであり，具体的には（山-やま），（やま-ヤマ）などの形で現れるものをさす。「Ⅲ　その他」は，いわば文字以外の記号類と文字との対立であるが，そのうち，くり返し記号については一つ独立の項目を立てている。これはくり返し記号をくり返される文字と同じものが書かれているとみなすことも可能だからである。

また，分類のゆれの中には語以前の段階の文字としての表記にゆれがあるとみなせるものが含まれている。たとえば(b)がその典型的な例である。これは厳密には文字レベルの表記のゆれであるから，理論上ここに含めない考え方もありうる。しかし語が文字のあり方に影響を受けるのは当然で，文字表記にゆれがあれば，即座に語表記のゆれとなって現れる。そう考えて分類上は文字レベルの一部のゆれも含めてかかげた。

以上これらはいずれも基本的な対立パタンであって，それですべて十分であるというわけではない。これ以外の分類法もあるし，分類項目の過不足もあるかもしれない。分類法は表記のゆれの認定のあり方が関係するからである。語表記にゆれが認められるということは，同じ語であると判定された，二つ以上の表記において，その表記形式上に差異があると認められた場合である。したがって「語が同じであるか否か」「表記形式が異なっているかどうか」の判定基準が違えば，表記のゆれの認定も当然ながら異なってくる。

たとえば，「人が泣く」の「泣く」と「鳥が鳴く」の「鳴く」が同語であるか異語であるかの判定によって，ゆれのあり方は異なる。また「行なう」と「行う」や「ピーナッツ」と「ピーナツ」の表記上の差異を差異として認めるか，無視するかによってもゆれのあり方は異なる。これらの判定は判定する人や目的などによって違ってくるものであり，その結果，分類のあり方が変わるのは当然である。たとえば「漢字・ひらがな・カタカナの使い分け」だけを調べる

ことが目的ならば，上記の分類のⅠ・Ⅲはまったく無視してよいことになる。

要するにゆれに対する概念は人や目的によってさまざまなものがありうるのであって，ゆれの尺度を想定する場合には，それらの種々の考え方を十分に反映しうるものでなければならない。言いかえれば，ゆれに対するどのような考え方にも応用しうる尺度であるべきだということになる。

3. 表記のゆれの尺度

それでは，表記のゆれの大きさを測るにはどうすればよいか。

例をあげて考えよう。かりに次のような表記の用例が見られたとする。

『すなわち』：すなわち（70例）・即ち（30例）

『ちょうど』：ちょうど（70例）・丁度（15例）・恰度（10例）・チョウド（5例）

『すべて』　：すべて（91例）・総て（4例）・凡て（3例）・全て（2例）

これらのゆれの大きさの大小関係はどのように考えるべきか。『すなわち』と『ちょうど』を比べれば，表記形式の種類が多い『ちょうど』の方がゆれが大きいと言えよう。また『ちょうど』と『すべて』を比べれば，『ちょうど』の方が大きいと言えるだろう。『すべて』の用例の多くが「すべて」に集中し，『ちょうど』よりもゆれが小さいと思われるからである。それでは，『すなわち』と『すべて』ではどうか。

これには相反する二つの意見がありうる。一つは表記形式が多いことを理由に，『すべて』の方がゆれが大きいとする意見である。他の一つは『すべて』の「総て」「凡て」「全て」の用例数が少ないから『すなわち』のゆれが大きいとみる意見である。前者は表記形式の多少を，後者は表記の用例数の分布を問題にしているわけである。両者の意見には，どちらにも一理あり，一方だけを取りあげるわけにはいかない。

そこで，このことを一般化して考えよう。そのために，比喩的なモデルにた

とえてみる。「n 個のボールを x 個の箱に任意の入れ方で入れる」という条件を立てる。ボールは一つの語を意味し，箱は表記形式を意味している。だから n は表記の用例の総数であり，x は表記形式の種類の数である。そして，この条件に対する結果が表記のゆれに相当するわけである。

先のゆれの大きさについての二つの意見のうち，表記形式の種類の多少を問題にする立場では，ボールがいくつの箱に散らばって入っているかをみることになる。ボールの入った箱の数すなわちゆれの大きさである。この場合，ボールが一つ入った箱も，百個・千個入った箱も同じ一つに数えるわけで，その点無理が感じられる。だから，箱の中にそれぞれいくつのボールが入っているかも問題にすべきだと考えられる。すなわち，ボールが入った箱の数と，それらの箱の中のボールの数との両方を考慮に入れたうえで，ボールの散らばりぐあいを調べるべきである。そして，より多くの箱に，より平均的にボールが入っているほど，散らばりが大きいということである。

結局，表記のゆれを測るには，先程の二つの意見がそれぞれ問題にした「表記形式の種類の数」と「用例度数の分布」の両者を指標として見ることが必要である。そして，より多くの表記形式が用いられ，その用例の数の分布がより平均的に散らばっているほど，ゆれが大きいと判定するのである。これは一種の散らばり度を測定していることである。

そこで散らばりの尺度を示す分散の考え方を応用してみる。分散の基本的な考え方は，$S=\frac{1}{N}\sum(Y_i-\bar{y})^2$ で示される。実測値 Y_i に対して何らかの意味で基準となる理論値 $\bar{y}$ を求め，それと実測値 Y_i とのへだたりの平方の総和を計算し，それを度数Nで除して算出するというものである。

いま，ある語がN回出現したとする。その表記形式は G_1，G_2，……G_n のr通りあり，各表記形式の出現回数は L_1，L_2，……L_r だとする。当然 $N=\sum L_i$ であり，各表記形式の出現率 P_i は $P_i=L_i/N$ で求められ，$\sum P_i=1$ となる（表1参照）。そして分散をこの出現率で考えると，

$$C=\sum(P_i-P)^2 \quad \cdots\cdots①$$

表 1　表記のモデル

表記形式	G_1 G_2 ……G_i ……G_r	計
出現度数	L_1 L_2 ……L_i ……L_r	N
出 現 率	P_1 P_2 ……P_i ……P_r	1

と表せる。

①式で問題なのは $\bar{p}$ なる理論値である。すぐに考えられるのは平均値である。たとえば，各語において漢字表記された割合を調べる。それらの平均値を求めて，それを漢字表記がなされる理論値とするのである。しかし，これには無理がある。なぜなら，表記形式のあり方は語によってさまざまであり，常に漢字表記されるものもあれば，まったく漢字表記されることがないものもある。それらを等質のものとして扱うことは無意味だからである。

そこで，観点を変えることにする。表記のゆれが最大になる極限状態を想定してみよう。これは表記形式の数 r が無限であり，各表記形式の出現率が平等な状態，すなわち P_i がほとんどゼロに等しい状態である。この状態を基準にとり，理論値 $\bar{p}=0$ とするわけである。したがって，これによって求められるのは，表記のゆれのもっとも大きな状態からのへだたりである。ゆれが最大の状態からのへだたりが大きいことは，実際にはゆれがもっとも小さいことを意味する。よって $\bar{p}=0$ を①に代入して，

$$c'=\sum P_i^2 \quad \cdots\cdots②$$

を求めることは，語の表記形式の集中度を測ることに等しい。だから，表記が一つの形式に常に固定されていて，集中度がもっとも高い場合は，

$$c'=1^2=1$$

となる。これが c' の最大値である。そこで再びゆれの大きさを示すものとして，この1から c' を差し引いたものをとることにする。つまり，表記のゆれの尺度Sを，

$$S=1-\sum P_i^2 \quad \cdots\cdots③$$

と定めるわけである。

さきの例『すなわち』,『ちょうど』,『すべて』について演算してみると,

『すなわち』：$S=1-(0.7^2+0.3^2)=0.42$

『ちょうど』：$S=1-(0.7^2+0.15^2+0.1^2+0.05^2)=0.475$

『すべて』　：$S=1-(0.91^2+0.04^2+0.03^2+0.02^2)=0.169$

となる。ゆれの大きさの順序は『ちょうど』,『すなわち』,『すべて』である。

4. ゆれの尺度の適用

以上で，この論の主たる目的は一応達せられたわけである。が，この尺度が実際にどの程度役立つ見込みがあるか，また改良すべき点はどこかなどを明らかにしておくことも必要である。そうした意味から，具体的なデータに適用してみた。データは，『電子計算機による新聞の語彙調査(Ⅱ)(注1)』で扱われた，昭和41年の新聞記事の一部である。これを選んだのは，手近にあって，データの量が多いことと，比較的簡単に元のデータにもどって実際の表記形を確認できることという便宜的な理由からである。調査した語は，使用度数20以上の語で，形容詞（45語），形容動詞語幹（42語），副詞（67語），及び動詞（80語）である(注2)。使用度数があまり低い場合は，偶然によって起こる誤差が大きくなり結果をゆがめてしまう。それを極力抑えるために，調査対象の語を度数20以上に限ったのである。

調査の手順のあらましを簡単に述べよう。まず『電子計算機による新聞の語彙調査(Ⅱ)』に収められた「品詞別度数順短単位表」から対象にする語を選び出した。次にそれぞれの語について，用例つき用語総索引である『KWIC(注3)』を見て，品詞や表記形式の確認をしながら度数を数えた。その際，表記形式のゆれとして認めたのは，異なる文字体系間の対立と，二種以上の漢字表記の対立である。そして数えた度数をもとに先の式にあてはめて演算したわけである。

その結果を整理したものの一部を表2，表3とグラフに示す。表2の「一語あたりのゆれ」と「一表記あたりのゆれ」について説明を加えておこう。ある

表2　品詞別にみた表記のゆれ

	対象語数	延べ語数	一語あたりのゆれ	一表記あたりのゆれ
動詞	80	10151	0.218	0.177
形容詞	45	4637	0.123	0.068
形容動詞語幹	42	1412	0.119	0.107
副詞	67	4624	0.132	0.137

表3　ゆれの大きな語（S＞0.3）

＜形容詞＞

1.	堅い	.578
2.	くわしい	.482
3.	おもしろい	.477
4.	忙しい	.463
5.	楽しい	.454
6.	激しい	.361
7.	すばらしい	.349

＜形容動詞語幹＞

1.	すてき	.550
2.	まじめ	.540
3.	主	.500
4.	大変	.493
5.	たしか	.480
5.	盛ん	.480
7.	豊か	.305

＜副詞＞

1.	いっそう	.522
2.	はじめて	.515
3.	まったく	.499
4.	直ちに	.498
5.	つねに	.497
6.	さきに	.458
7.	最も	.454
8.	果たして	.453
9.	次第に	.444
10.	一番	.432
11.	とくに	.429
12.	なお	.404
13.	ときどき	.326
14.	たとえば	.324

表記のゆれの大きさによる語の分布

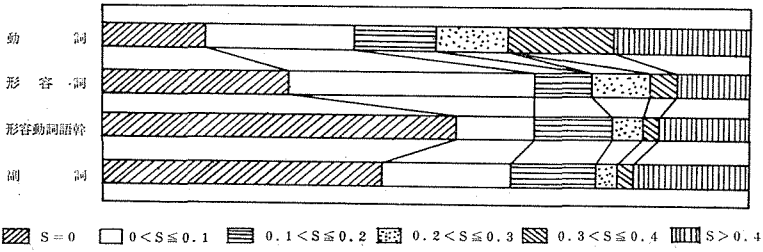

表4

語	W_1	W_2	……W_i	……W_r
出現度数	N_1	N_2	……N_i	……N_r
ゆれの大きさ	S_1	S_2	……S_i	……S_r

品詞において，語がW_1，W_2，……W_r のr通り出現したとする。すなわち異なり語数rである。そして，それらの語の出現度数はそれぞれ N_1，N_2，…N_r であり，ゆれの大きさは S_1，S_2，……S_r である。延べ語数は $\sum N_i$ となる(表4参照)。このとき，「一語あたりのゆれ」と「一表記あたりのゆれ」とは，次の式で表される値である。

$$\text{一語あたりのゆれ}=\frac{\sum S_i}{r}$$

$$\text{一表記あたりのゆれ}=\frac{\sum (S_i \cdot N_i)}{\sum N_i}$$

これらの結果からどのようなことが考えられるのだろうか。資料の性格や調査の方法の点で，まだ問題が残るので，断定的な態度はさけなければならない。よって，以下目につくことをいくつか推測として述べる。

A）表2から動詞のゆれが他より大きいことがわかる。これは，動詞が形容詞や形容動詞語幹に比べて，一つの語が含みもつ意味が多様であることと関係があるのではないか。つまり意味の差異を表記の違いで示している可能性があると思われる。

B）表2から形容詞の〝一表記あたりのゆれ〟が他に比べて小さいことが認められる。これは使用度数の高い語のゆれが小さいためである。つまり，形容詞では使用率の高い語は表記が安定していると言えよう。

C）グラフによれば，形容動詞語幹でゆれないものがかなり多い。このほとんどは漢語であり，漢字表記である。それに対し，表3のゆれの大きい形容動

詞語幹の7語のうち，和語が6語を占めている（調査対象42語のうち和語は12語）。これは漢語の表記が和語に比べて安定していることを示すものと考えられる。

D）表3で，ゆれの大きな語の中に，「面白い・素敵・真面目」のような当て字・熟字訓の表記を用いる語が見られる。また「固い・主な・盛んな」のような，昭和41年当時は当用漢字音訓表で認められていない表記が用いられている語がある。この事実は，当用漢字表・当用漢字音訓表による制約が表記に与えた影響が働いたと思われる。

以上のように，品詞・使用率・語種・表記上の制約などを表記のゆれの要因として推察することができるということは，この尺度の有効性をある程度示すものだと言えよう。ただ，使用度数が低い場合はゆれの誤差が大きくなるなどの欠点がある。しかし，大量のデータの処理によって，ゆれの大きな語と小さな語とをよりわけることは可能である。それを土台として，さらに次の研究段階である質的な分析への発展が期待できよう。

（注1）　国立国語研究所報告38『電子計算機による新聞の語彙調査（Ⅱ）』（昭和45年）

（注2）　品詞別度数順短単位表をもとに，度数20以上のほとんどすべての形容詞・形容動詞語幹・副詞を調査。動詞については度数20以上の約40％の語を調べた。

（注3）　電子技術総合研究所の植村俊亮氏によって，マイクロフィッシュ化されたもの。

この研究は，昭和51年度文部省科学研究費補助金（一般研究A）による研究の一部である。