『現代日本語書き言葉均衡コーパス』に見る語表記の量的分布 —品詞，レジスター，頻度との関係—

山崎　誠
国立国語研究所
yamazaki@ninjal.ac.jp

概要

『現代日本語書き言葉均衡コーパス』を使って，現在の日本語の語表記の種類（書字形の数）の量的分布を調査した結果を報告する．これまで国立国語研究所では語彙調査とともに用字調査を行い，漢字表や表記のゆれなどの報告書を刊行してきた．書き言葉コーパスとして代表的な『現代日本語書き言葉均衡コーパス』は 2011 年にリリースされたが，今のところこれを使った総合的な用字調査が行われていない．本発表では，語表記の種類に着目し，その量的実態を報告するものである．本発表では，全体の概要を示すとともに，誤表記の種類と品詞，レジスター，出現頻度との間に一定の傾向が見られたことを確認した．

1 はじめに

国立国語研究所では 1948 年の創立以来，書き言葉の実態調査である語彙調査を実施してきたが，それと同時に用字調査も行い，漢字表や表記のゆれなどの報告書を刊行してきた．例えば，『総合雑誌の用語』（1960）や『現代新聞の漢字』（1976）『現代表記のゆれ』（1983）などである．書き言葉コーパスとして代表的な『現代日本語書き言葉均衡コーパス』（以降，BCCWJ）が公開されて久しいが，まだこれを使った漢字表や表記の調査は行われていない．

2 先行研究

個々の語の表記の計量調査ではなく，コーパス全体を調査したものは少ない．書き言葉の総合的調査である，国立国語研究所(1963)[2]は漢字の計量調査の報告であるが，表記についての言及はない．国立国語研究所(2006: 32-33)[3]には，調査対象となった，雑誌に出現した語について，主に語種を中心とした以下のような分析を行っている．

出現した和語 11530 語について，そのうちの 29.3%である 3378 語についてゆれが見られた一方，漢語 15214 語について，そのうち 4.16%に揺れが見られ，和語の方がゆれの出現率が高かった．また，語の表記形の数が増えるにつれて語数が減っていくことも示された（以上，筆者による要約）．国立国語研究所(2006)[3]は，BCCWJ と語の認定の仕方がやや異なるため，厳密な比較はできないが，これらは大まかな傾向として参考になるものである．

3 データと方法

3.1 データ

本発表で使用したデータは，BCCWJ Ver.1.1(2015)[DVD 版]である．その DISC4OT に納められている TSV_SUW_OT フォルダの TSV ファイルを利用した．OT というのは，コーパス構築にあたって，NumTrans[1]という数字変換処理を施していない，Original Text という意味である．NumTrans を施した本文は，アラビア数字が漢数字に置き換えられてしまうので，表記の調査としては，NumTrans 前の Original Text がふさわしいと考えた．

3.2 語の同定

語の表記を数えるには，語の同定と表記の同定の 2 つの方法を決定しなければならない．BCCWJ をはじめとする国立国語研究所のコーパスは UniDic という形態素解析用の辞書を用いて分析されているため，UniDic の形態論情報を用いて語の同定と表記の同定を行うのが妥当である．語の同定は，形態論情報である（1）語彙素 ID を使う，（2）語彙素，語彙素細分類，語彙素読み，品詞の組で語を使う，の 2 つが考えられる．しかし，次に述べる理由で以下

[1] NumTrans については，[1]の P.100 以降を参照のこと．

のような変則的な方法を採用した．

以下は，語彙素[2]，語彙素読み，語彙素 ID，品詞の 4 つを用いて語を同定し，その表記をまとめたものである．表記形に続けて括弧内に頻度を示した．

「あいす(4),あいする(7),愛す(1300),愛する(5010),愛せる(135)」

これを見ると，活用の型が異なる「愛する」（サ変）と「愛す」（五段活用）が同じ語の範囲になっているほか，可能動詞形の「愛せる」も同じ語の範囲と扱われていることが分かる．これらを同じ語と考えると，「あいす」と「あいする」は同じように平仮名で書かれているにも関わらず，表記が違うということになる．これは表記の違いというより語形の違いをカウントしていることになるので，これらは別語と扱うほうが望ましいだろう．そこで，基本的には，語彙素，語彙素読み，語彙素 ID，品詞の 5 つで語を同定するが，動詞に限っては，語彙素読みを語形で置き換えたものを使った．それによって，この例は, 以下のような 3 語に分かれることになる．

・あいす(4),愛す(1300)
・あいする(7),愛する(5010)
・愛せる(135)

なお，長単位を使えばこの問題は起きないが，名詞に長い複合語が出てくるため，今回はできるだけ短い言語単位での実態の把握を優先した．この調整により，動詞の 1 語あたりの書字形数の平均は 2.90 から 2.32 に下がった．

3.3 表記の同定

表記の同定は，形態論情報である書字形を使う．書字形は, 原文での出現形を表記の点で保持するが，活用形の違いを捨象した終止形で示した形である[3]．

3.4 除外した語

以下の語は, 本稿の趣旨に合わないため除外した．

・語彙素が NULL のもの[4]　298410 語
・伏せ字を示す■[5]　65487 語

[2] 語彙素細分類は使用していない．語彙素 ID があればその代わりになるからである．
[3] 書字形出現形や原文文字列使うと活用形の違いも表記の違いにカウントされるので，採用しなかった．
[4] これらの語の品詞欄は，「URL，web 誤脱，カタカナ文，ローマ字文，名詞-固有名詞-人名-名，名詞-数詞，名詞-普通名詞-サ変可能, 名詞-普通名詞-一般, 形状詞-一般, 新規未知語，方言，未知語，漢文，英単語，言いよどみ，記号-文字」となっている．

4 結果

4.1 全体について

2 節で述べた語の数え方により, BCCWJ の短単位全体で異なり 190,373 語，延べ 124,256,025 語が得られた. 1 語あたりの書字形数の分布を図 1 に示した．1 語あたりの書字形数の平均（算術平均）は，1.509 であった．また，最小値，第 1 四分位数，中央値，第 3 四分位数はいずれも, 1, 最大値[6]は 81 であった．

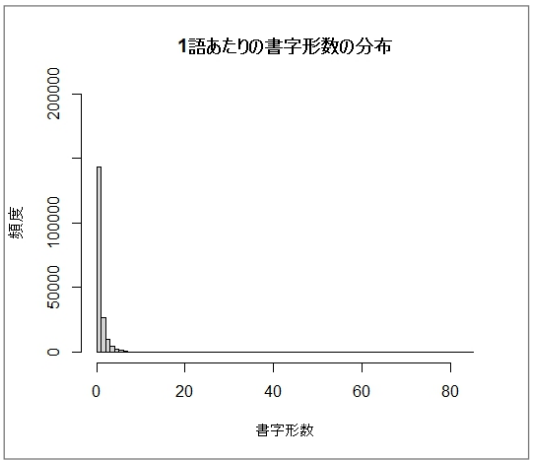

図 1　1 語あたりの書字形数の分布

4.2 品詞との関係

図 2 は, 品詞別に 1 語あたりの書字形数の平均（以降，平均書字形数という）を示したものである．

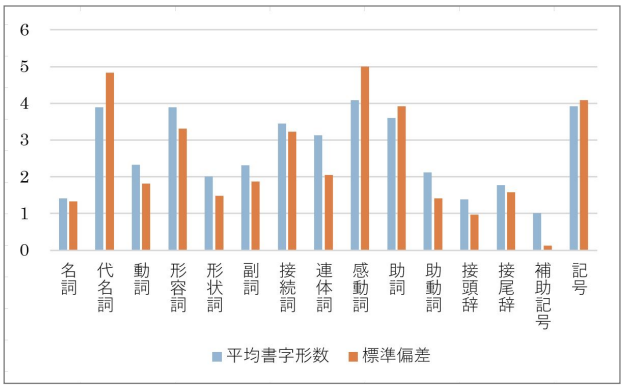

図 2　品詞と平均書字形数

[5] 新聞やブログなどで個人情報保護のため, 伏せ字にしたものがあり，出現形を■で置き換えている．なお，本来の記号である■は，除外していない．
[6] 最大値は固有名詞（人名）の「こうじ」であった．なお，1 語あたりの書字形数が 37 以上の 21 語はすべて人名であった．

平均書字形数が多い品詞は，感動詞(4.081)，記号(3.926)，代名詞(3.987)，形容詞(3.986)，助詞(3.603)である．感動詞の平均書字形数が多いのは例えば，「ああ」に以下のような 36 個のバリエーションが見られるように，臨時的な語形が多いからと推測される．

ぁぁぁ(3),ぁああ(25),あ〜(814),あ〜あ(62),あ〜っ(17),あぁ(845),ああ(7980),あああぁ(4),あああ(641),あああ〜(16),あああー(3),あああっ(216),あああー(62),あゝ(181),あァ(28),あア(1),あー(517),あーあ(166),あーっ(92),あーア(1),アゝ(1),アァ(24),アア(61),アアア(1),アー(22),アーア(5),アーッ(19),吁嗟(1),嗚〜呼(3),嗚乎(1),嗚呼(84),嗟(2),嗟呼(3),噫(21),臆(7),嗚呼(5)

4.3 レジスターとの関係

図 3 は，BCCWJ を構成するレジスター[7]ごとに平均書字形数と標準偏差を見たものである．

LB（図書館書籍）と PB（出版書籍）が平均書字形数 1.4 以上で大きい．これはこれらのレジスターに相対的に多様な表記が用いられていることを示すものであろう．一方，平均書字形数が小さいのは順に，OL（法律,1.04），OM(国会会議録,1.10)，PN（新聞,1.12），OW（白書,1.13），OT（教科書,0.18）となっており，公的な媒体において表記のバリエーションが少ないことが分かる．バリエーションの小さいレジスターは，以下のように，標準偏差も小さい．OL（法律，0.21），OL（国会会議録，0.42）OW（白書，0.43），OT（教科書，0.53），OP（広報紙，0.58）．これは，これらの媒体の表記が安定していることを示していると考えられる．

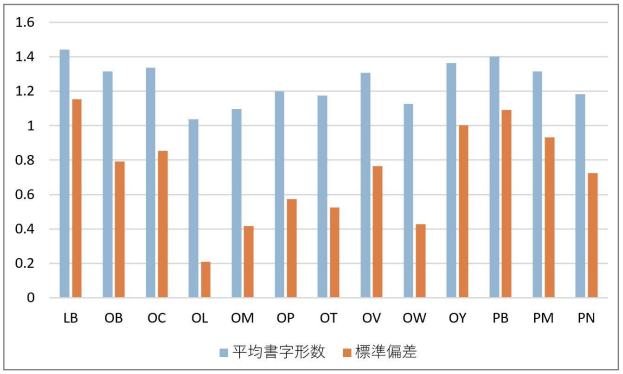

図 3　レジスターと平均書字形数

[7] レジスターの略号は以下のとおり．LB(図書館書籍)，OB(ベストセラー)，OC(Yahoo!知恵袋)，OL(法律)，OM(国会会議録)，OP(報紙)，OT(教科書)，OV(韻文)，OW(白書)，OY(Yahoo!ブログ)，PB(出版書籍)，PM(雑誌)，PN(新聞)

4.4 頻度との関係

表 1 は，固有名詞を除く，書字形数が 22 以上の語の形態論情報と出現頻度を示したものである．表 1 を見ると，書字形数が多い語には，頻度も多いものが多いようである．そこで，語の出現頻度と書字形数の関係を見たのが図 4 である．図 4 は，出現頻度を 1000 刻みでその範囲に属する語の平均書字形数を求めたものである．

表 1　書字形数の多い語

rank	語彙素	品詞	語彙素ID	頻度	書字形数
1	ああ	感動詞	67	11934	36
2	何	代名詞	27920	169085	34
3	コウ	記号	11823	671	31
4	ずっと	副詞	19640	12340	29
5	一寸	副詞	24199	29800	28
6	あはは	感動詞	915	748	28
7	ほほほ	感動詞	248901	193	28
8	おお	感動詞	4507	2273	26
9	ショウ	記号	430	513	26
10	婆	名詞	30509	5707	25
11	はあ	感動詞	30476	2600	25
12	良く	副詞	39182	42307	24
13	暗い	形容詞	10405	5277	24
14	シ	記号	14924	296	23
15	て	助詞	24874	3493117	22
16	良い	形容詞	38988	198994	22
17	さん	接尾辞	14495	169978	22
18	爺	名詞	17702	4096	22
19	いや	感動詞	2513	1712	22
20	くくく	感動詞	177320	152	22

図 4　出現頻度と平均書字形数

図 4 からは，途中で増加が止まるような箇所もあ

るが，大局的にみると，語の出現頻度が高くなるにつれて，平均書字形数も多くなる様子が見て取れる．

4.5 品詞とレジスターの関係

図 5 は，品詞とレジスターを変数としてコレスポンデンス分析を行ったものである[8]．

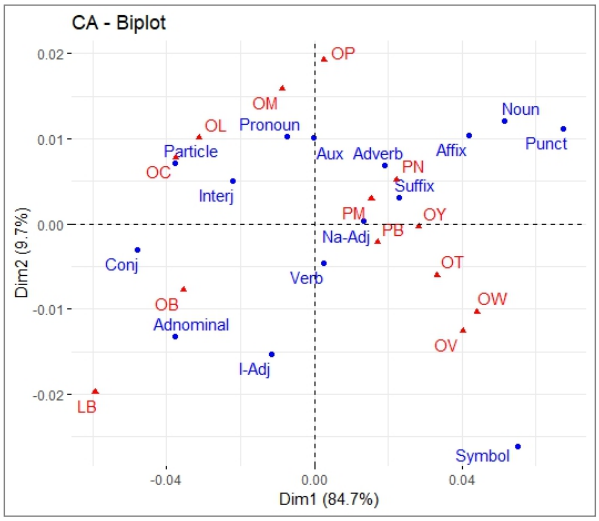

図 5　コレスポンデンス分析の結果

図 5 から X 軸（Dim1）で全体の約 85%が説明できることが分かる．この軸はレジスターでは，OW（白書）と LB(図書館書籍)を対極とし，品詞では句読点と感動詞を対極とすることから，書字形のバリエーションの多寡に対応する軸であると思われる．ただし，同じ書籍であっても LB(図書館書籍)と PB(出版書籍)との布置はかなり違っていることについては説明が難しい．

5 おわりに

本発表では，BCCWJ を使って現代日本語の語表記のバリエーションを計量的に記述した．今回分析できなかった観点としては，先行研究にあった，和語，漢語外来語などの語種との関係が挙げられる．先行研究では，和語よりも漢語の方が表記が安定していることが示されていたが，表記のバリエーションが多い漢語や逆にバリエーションが少ない和語も考えられることから，どのような条件がそこに働いているか，解明する必要がある．また，課題として語の認定のしかたがある．今回は 3.2 節に述べたように，語彙素，語彙素読み，語彙素 ID，品詞の 4 つで語を同定．動詞に限って，語彙素読みを語形で置き換えるという方法だったが，例えば，名詞「バイオリン」は，「バイオリン(293),ヴァイオリン(415),ヴィオロン(4)」のような分布になっており，語形の異なる「ヴィオロン」も同じ語の範囲に入っている．では，全ての品詞で語形によって語を認定するとよいかというと，その場合，「ハロウィン(130),ハロウィーン(69),ハロウイーン(4),ハローイン(1)」がすべて別語になってしまい，表記のバリエーションを見る目的に適さなくなってしまう．活用語における活用の型の違いや可能動詞を別語とするくらいがよいのかもしれない．その辺の検討を経て，BCCWJ の表記一覧を同漢字表とともに国立国語研究所の学術情報リポジトリで公開する予定である．

また，今回は品詞を大分類のレベルで分けているが，これを中分類や小分類のレベルにまで分けたときにどのような違いが見られるかという点も検討課題である[9]．

謝辞

データとして利用した BCCWJ は，国立国語研究所のプロジェクト及び文部科学省科学研究費補助金特定領域研究「代表性を有する大規模日本語書き言葉コーパスの構築：21 世紀の日本語研究の基盤整備」（平成 18～22 年度，領域代表者：前川喜久雄）による補助を得て構築したものである．

参考文献

1. 国立国語研究所コーパス開発センター,『現代日本語書き言葉均衡コーパス』利用の手引第 1.1 版, 2015 https://clrd.ninjal.ac.jp/bccwj/doc.html
2. 国立国語研究所，『現代雑誌九十種の用字用語：第 2 分冊漢字表』，1963，秀英出版. http://doi.org/10.15084/00001234
3. 国立国語研究所，『現代雑誌の表記 ：1994 年発行 70 誌』，2006. https://doi.org/10.15084/00001286

[8] R の devtools，ggplot2，factoextra，FactoMineR のライブラリを使用した．

[9] 例えば人名では，「名詞-固有名詞-人名-姓」は平均書字形数が 1.44 に対し，「名詞-固有名詞-人名-名」が 2.43 と開きがある．