『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 研究系/言語資源開発センター）

1.はじめに

- 『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が、2011年にリリースされた『現代日本語書き言葉均衡コーパス』は、今のところ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。語表記の数と品詞、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。

2.先行研究

- 国立国語研究所(2006: 32-33)での指摘
- 表記のゆれの割合（異なり語数）：和語29.3%、漢語4.16%
- 語の表記形の数が増えるにつれて語数が減っていく。

国立国語研究所(2006)現代雑誌の表記 : 1994 年発行70 誌　https://doi.org/10.15084/00001286

3.データと方法

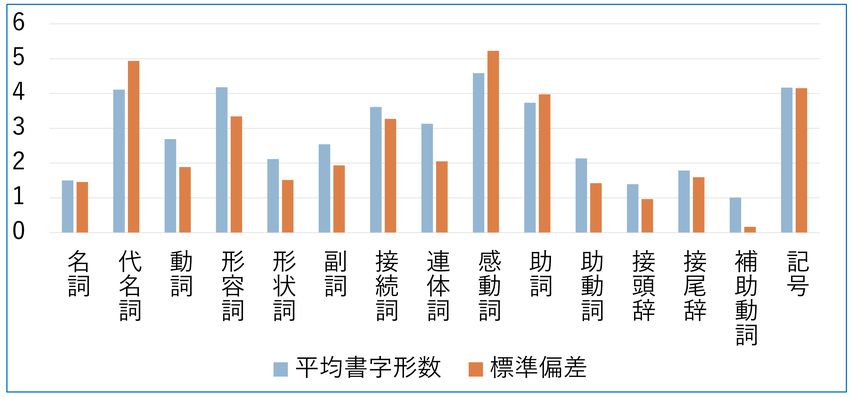
- BCCWJ Ver.1.1(2015) [DVD版]を使用。
- NumTrans（数字変換処理）を施していないOTのデータを使用した。
- 語の同定の方法は，語彙素，語彙素読み，語彙素ID，品詞の4つで語を同定する。

- ただし、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。そうしないと、「あいす、あいする、愛す、愛する、愛せる」が同じ語の表記のゆれとして扱われることになり、違和感がある。語彙素読みの代わりに語形を使えばこの例は、「あいす、愛す」「あいする、愛する」「愛せる」の3語の扱いになり、表記のゆれの測定という観点から妥当な分割になる。
- 表記の同定は，形態論情報である書字形を使った。
- 除外した語：語彙素がNULLのもの（298410語）および伏せ字を示す■（65487語）

4. 結果

- BCCWJの短単位全体で異なり190,373語（頻度2以上155,986語）、延べ124,256,025語（頻度2以上124,221,639語）が得られた。
- 1語あたりの書字形数の平均は1.621であった。また、最小値、第1四分位数、中央値は1、第3四分位数は2、最大値 は81であった。
- データは近いうちにリポジトリで公開の予定。

5. 品詞との関係

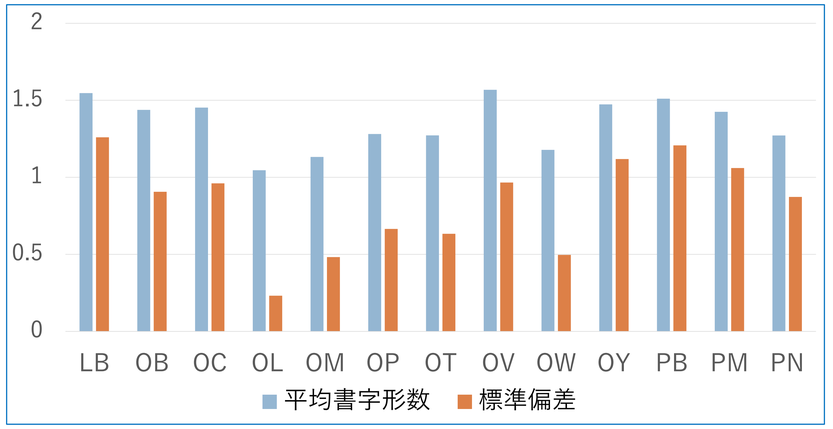


- 品詞により書字形数が異なることを示唆している。書字形数が一番多いのは人名の「コウジ」で81個であった。ただし、「名詞-固有名詞-人名-名」の平均書字形数は2.43でさほど高いわけではない。山崎(2022)とほぼ同じ結果。

1.1 お詫び

- 本ポスターの分析は、予稿集で行ったものと以下の点で異なっています。
- 予稿集では、語の頻度が1のものを集計に含めていましたが、表記の「ゆれ」を見る観点からは、頻度1は含めないほうがよいと判断し、本ポスターでは語の頻度2以上を対象としています。
- 数値に変動がありましたが、分布や傾向はだいたい同じで、結論への影響はありませんでした。

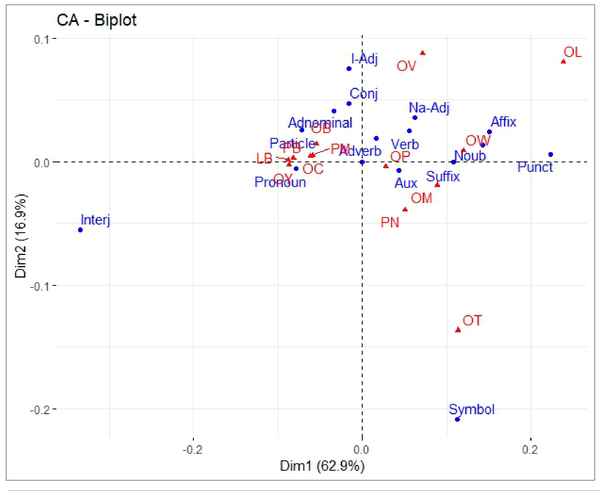
6. レジスターとの関係



- いちばん平均書字形数が大きいのはOV(韻文)、次いでLB(図書館書籍)、PB(出版書籍)、OY(Yahoo!ブログ)、OC（Yahoo!知恵袋）と続く。個人の特徴が出やすいものが平均書字形数が大きい傾向がある。
- 一方、公的なレジスター（OL(法律)、OM(国会会議録)、OW(白書)、OT(教科書)、OP(広報紙)）において平均書字形数が少ない。平均書字形数が少ないレジスターは標準偏差も小さい。

- 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（広報紙）　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）

8 .品詞とレジスターの関係



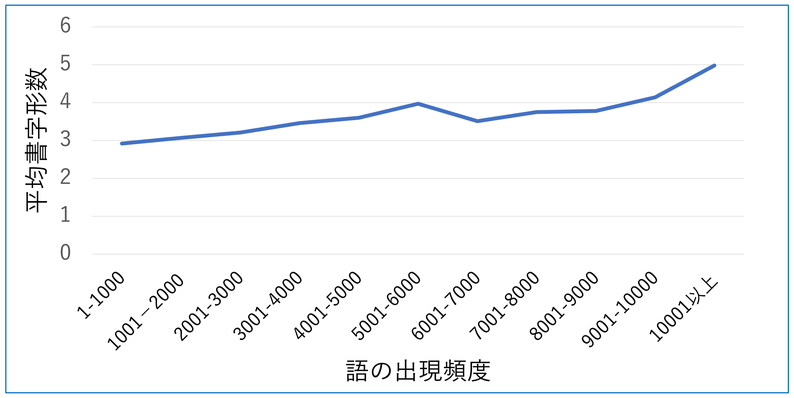
- 平均書字形数によるコレスポンデンス分析の結果
- 横軸は平均書字形数の多寡を示している。

- お詫び2
- 予稿集では、OL(法律)の感動詞の値を間違って算出していました。
- こちらの図が正しいものです。

7 .頻度との関係

書字形数の多い語（固有名詞を除く）

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



- 出現頻度とのゆるやかな相関が見られる。山崎(2022)とほぼ同じ結果。

9 .課題

- 表記のゆれの計り方を要検討。語彙素，語彙素読み，語彙素ID，品詞の4つで語を同定すると、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)」のような分布になり，語形の異なる「ヴィオロン」も同じ語の範囲に入る（が、これでよいか）。

参考文献
国立国語研究所(2006)『現代雑誌の表記：1994 年発行70 誌』https://doi.org/10.15084/00001286
山崎誠(2022)　「現代日本語書き言葉均衡コーパス」の漢字と表記、NINJAL OPENHOUSE 2022, https://www2.ninjal.ac.jp/openhouse/2022/pdf/27.pdf