NINJAL OPEN HOUSE 2022

「現代日本語書き言葉均衡コーパス」の漢字と表記

山崎　誠（国立国語研究所 研究系/言語資源開発センター）

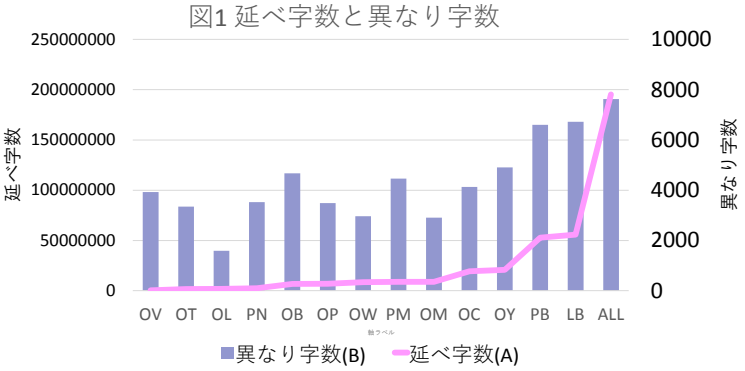
1.はじめに

- 国立国語研究所では語彙調査にともない、用字調査を行ってきた。その成果は、『総合雑誌の用字』(1960)や『現代新聞の漢字』(1976)などとして報告されてきた。
- 書き言葉の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』(BCCWJ)が公開されて久しいが、まだ用字に関する総合的な調査結果は発表されていない。
- 本発表ではBCCWJに見られる、用字（字数、字種、表記）について概況を報告するものである。

- レジスターの略号
- 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）

2.延べ字数と異なり字数

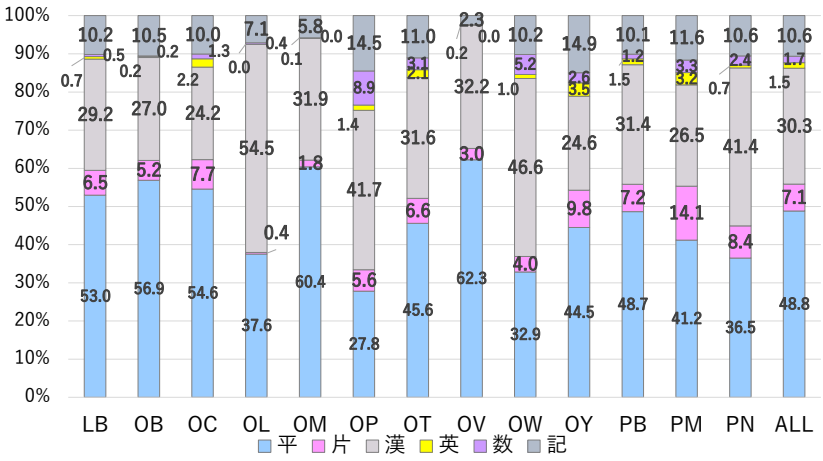
図1 延べ字数と異なり字数



- 図１(上)は、延べ字数の順に、横軸にレジスターを並べたものである。全体的に、異なり字数は延べ字数に比例しているようであるが、OL(法律）、OP（広報紙）、OW(白書）、OM(国会会議録）は全体的な傾向に反するようである

3.レジスターごとの字種の割合

図2 レジスターごとの字種の割合



- 図2(左)は、レジスターごとの字種（平仮名、片仮名、漢字、英字（ギリシャ文字、キリル文字を含む）、数字、記号の割合である。
- 平仮名：OM（国会会議録）、OV(韻文）に多く、OL（法律）、OP（広報紙）、OW（白書）、PN（新聞）に少ない。
- 片仮名：PM(雑誌）に多く、OL(法律）、OM（国会会議録）に少ない。
- 漢字：OL、OP、OW、PNに多く、OC(知恵袋）、OY(ブログ）に少ない。

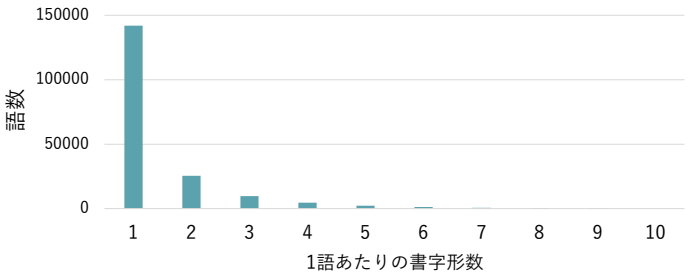
4. 漢字の頻度（上位10字）

表1　レジスター別の漢字頻度表数

順位	全体	LB	OB	OC	OL	OM	OP	OT	OV	OW	OY	PB	PM	PN
1	人	人	人	人	第	一	日	地	日	年	日	一	人	日
2	一	一	一	思	定	大	月	生	人	業	人	人	大	人
3	日	大	大	方	条	国	時	人	一	国	今	的	日	年
4	年	日	言	出	項	思	会	大	夜	等	出	大	年	一
5	大	年	日	日	十	十	市	分	見	地	大	分	一	大
6	分	出	出	分	一	員	人	物	花	事	一	年	本	国
7	出	中	見	事	二	方	年	国	子	者	時	日	分	会
8	行	分	子	言	者	年	合	化	風	的	見	生	中	十
9	中	見	分	一	規	的	所	図	水	行	年	行	出	中
10	生	本	思	時	法	私	場	子	来	人	中	出	上	本

5. 1語あたりの書字形数

図3 1語あたりの書字形の分布



6. 品詞による違い

表2　品詞別の平均書字形数

品詞	書字形数
感動詞	4.07
形容詞	3.86
代名詞	3.83
助詞	3.36
接続詞	3.31
連体詞	3.04
動詞	2.90

品詞	書字形数
副詞	2.31
助動詞	2.07
形状詞	2.01
接尾辞	1.76
名詞	1.41
接頭辞	1.39
補助記号	1.01

- 表2(上)は、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。
- 図4(左)は、語の出現頻度が高くなると書字形数が増えることを示唆している。

7. 語の出現頻度と書字形数

図4　出現頻度と平均書字形数