# 改定常用漢字表のモデル化

菅野 倫匡
筑波大学 理工学群 社会工学類
s1211245@sk.tsukuba.ac.jp

## 1 はじめに

2010 年に「常用漢字表」(昭和 56 年 10 月 1 日内閣告示第 1 号) は廃せられ，新たな「常用漢字表」(平成 22 年 11 月 30 日内閣告示第 2 号) が示された．以後，前者を旧表，後者を新表とする．

なお，旧表に掲げられている漢字は 1945 字種あり，旧表から 5 字種を削り，196 字種加えたものが新表である．当然，字種に限らず，音訓や語例も改められているが，本稿では両表の漢字の異同に着目する．

新表を示した文化審議会は情報機器の広汎な普及が漢字使用の実態に変化を齎したとの見方に立ち，旧表と漢字使用の実態との懸隔を解消することを目指していた (小椋 2011)．そのため旧表との差異は漢字使用の実態の変化を反映する 1 つの指標と看做し得る．

また，審議会は漢字表を漢字政策の核となるものと位置づけ (文化庁 2011)，漢字使用の実態に合わせて定期的に見直すことを決めた (小椋 2011)．このため旧表との差異について検討すること，更には漢字使用の実態について検討することは漢字政策の評価や今後の見直しに繋がるものであり，新たな意義を持つ．

しかし，この点についての検討は十分に進められているとは言えない．また，検討する際の対象や手法に疑問の残るものもある．前者の例としては情報機器の普及という背景を顧みずに 1994 年の雑誌調査における漢字の出現頻度から新表について検証した島村 (2011) があり，後者の例としては Google 検索のヒット件数は時間変動が大きいために信頼し得ないとする田野村 (2008) の指摘を顧みずにヒット件数から新表について検証した野崎・江島・梅田 (2012) がある．

このような状況を打開するために本稿では統計的なモデルによる漢字表の評価という新たな観点から新表の検証を試みる．とりわけ，追加や削除を決定づけた要因を定量的に明らかにすることは漢字使用の実態の変化やその変化と政策との関わりを明らかにするための布石となるばかりか，漢字政策の評価や見直しにも資するものである．

## 2 対象および方法

初めに審議に用いられた「漢字出現頻度数調査」について述べる．この調査は 5 つの調査から成る．その概要を表 1 に示した．なお，A∼E は答申に拠る．

表 1: 漢字出現頻度数調査の概要

| | 調査期間 | 調査対象 |
|---|---|---|
| A | ’04 年 ∼’06 年 | 864 冊分の組版データ |
| B | | A の教科書部分 |
| C | ’06 年 10 月 1 日 | 朝日新聞の紙面データ |
| D | ∼’06 年 11 月 30 日 | 読売新聞の紙面データ |
| E | ’07 年 2 月, 4 月, 6 月 | ウェブサイト |

A· · · 凸版 (3) 調査　B· · · 第 2 部調査
C· · · 朝日新聞調査　D· · · 読売新聞調査
E· · · ウェブ調査

審議では凸版 (3) 調査を基本資料とし，他の調査を補助資料と位置づけている (文化庁 2011)．審議資料 (文化審議会国語分科会漢字小委員会 2008) を見ても凸版 (3) 調査における各字の順位に続いて他の調査の順位がそれぞれ示されている．この審議資料の概要は表 2 に示した．なお，この他に表外漢字字体表，新聞常用漢字等，備考を示した列がそれぞれある．

表 2: 漢字出現頻度表 順位対照表 (Ver.1.3) の概要

| A | 漢字 | 種類 | C | D | E | B |
|---|---|---|---|---|---|---|
| 1 | 人 | 常用 | 4 | 4 | 2 | 2 |
| 2 | 一 | 常用 | 10 | 12 | 5 | 12 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 3361 | 祁 | 人 B | | | 4237 | |
| 3362 | 冤 | 第 2 | 2857 | 2784 | 2499 | 2802 |
| 3363 | 圀 | 第 2 | 3492 | 2585 | 3690 | 2558 |
| ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ | ⋮ |
| 3500 | 謄 | 常用 | 2932 | 2576 | 3336 | 2168 |

表 2 のように審議資料には 3500 字が示されている．これは凸版 (3) 調査の出現頻度の上位 3500 字を選び出した上で必要な漢字を絞り込むという審議の方針に因る (文化庁 2011)．本稿はこの方針を踏襲し，凸版 (3) 調査の出現頻度の上位 3500 字を後述するモデルに組み入れることにする．

ただし，原データを入手できなかった第 2 部調査と読売新聞調査とを除外した．また，残りの 3 つの調査全てに出現するものを採用したことから計 250 字をも対象から除いた．そのため新表と旧表とで異同のある字は，いずれの調査にも現れない 2 字 (楷・鋼) と表 2 の「祁」のように順位の欠損があることから除かれた 5 字 (唾・淫・脹・匁・勺) とが外されて 194 字である．

次に本稿で採用するロジスティック判別と呼ばれる手法について述べる．なお，この手法が新表の検証に対して適用し得ることは本稿に先駆けて横山 (2006) の示唆するところであるが，未だに検証を試みたものがないことから本稿において検証してみたい．

ロジスティック判別とは $A$ という事象と $\bar{A}$ という事象があり，A の生起確率を $p$ とし，$\bar{A}$ の生起確率を $(1-p)$ としたとき，これを説明する $i$ 個の変数を取り入れたロジスティック回帰分析によって得られた $p > 0.5$ ならば $A$ が生起したと看做し，反対に $p \leq 0.5$ ならば $\bar{A}$ が生起したと看做すことによって判別するというものである．また，ロジスティック回帰分析は次式で表されるロジスティック函数を用いた回帰分析である．なお，$\beta_0$ は定数で，$\beta_i$ は偏回帰係数である．

$$p = \frac{1}{1 + \exp\{-(\beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i)\}}$$

これを変形すると次の式になる．偏回帰係数 $\beta_i$ を指数変換すると他の説明変数が一定の場合に説明変数 $x_i$ が 1 増加したときのオッズ比 $\frac{p}{1-p}$ となることからモデルを解釈し易い．

$$\ln \frac{p}{1-p} = \beta_0 + \beta_1 x_1 + \cdots + \beta_i x_i$$

また，本稿では新表に入るという事象の生起確率を $p$ としてモデルを構築した．

## 3 結果および考察

本稿では 3 つの分析結果を 3 つのモデルという形で示したい．初めに出現頻度の調査の順位のみから成る第 1 のモデルと第 2 のモデルとを表 3 に示す．なお，3 つの調査の順位を全て盛り込むとウェブ調査の順位が有意とならなかったことから，この点については後に示す補遺において再度検討した．

表 3: 第 1 のモデルならびに第 2 のモデル

| N=3250 | 第 1 のモデル | 第 2 のモデル |
|---|---|---|
| 切片 | 7.2031463*** | 6.5083165*** |
| | (0.2588723) | (0.2294174) |
| A の順位 | -0.0017088*** | -0.0021077*** |
| | (0.0001281) | (0.0001343) |
| C の順位 | -0.0016156*** | |
| | (0.0001234) | |
| E の順位 | | -0.0008680*** |
| | | (0.0001119) |
| Nagelkerke's $R^2$ | 0.7133 | 0.6845 |
| AIC | 1830.9 | 1968.1 |

() 内は標準誤差, $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

第 3 のモデルには 3 つの調査の順位に加えて旧表にあるか否か (旧表ダミー),「表外漢字字体表」にあるか否か (表外ダミー)，旧表にないもののうち都道府県名に用いられる漢字とそれに準ずる漢字とに該当するか否か (例外ダミー) をそれぞれ 0 か 1 かを取るダミー変数として採り入れた．第 3 のモデルを表 4 に示す．

表 4: 第 3 のモデル

| N=3250 | 第 3 のモデル |
|---|---|
| 切片 | 2.4342134*** |
| | (0.3967583) |
| A の順位 | -0.0007840*** |
| | (0.0001887) |
| C の順位 | -0.0006004** |
| | (0.0001858) |
| E の順位 | -0.0008076*** |
| | (0.0001846) |
| 旧表ダミー (0-1) | 8.2501993*** |
| | (0.7480471) |
| 表外ダミー (0-1) | 1.3275896*** |
| | (0.2253047) |
| 例外ダミー (0-1) | 2.6427783* |
| | (1.1453626) |
| Nagelkerke's $R^2$ | 0.8845 |
| AIC | 875.4 |

() 内は標準誤差, $^{***}p < .001$, $^{**}p < .01$, $^{*}p < .05$

モデルを解釈するために 3 つのモデルの偏回帰係数を対数変換して表 5 に示す．第 1 のモデルの凸版 (3) 調査の順位を例に取ると他の変数が一定のとき，A の順位が 1 単位増加する，すなわち，順位が 1 つ下がると新表に入る確率が 0.9983 倍になることが判る．

表 5: 3 つのモデルにおけるオッズ比

| | オッズ比 | 95%信頼区間 | |
|---|---|---|---|
| | | 下限 | 上限 |
| 第 1 のモデル | | | |
| A の順位 | 0.9983 | 0.9980 | 0.9985 |
| C の順位 | 0.9984 | 0.9981 | 0.9986 |
| 第 2 のモデル | | | |
| A の順位 | 0.9979 | 0.9976 | 0.9982 |
| E の順位 | 0.9991 | 0.9989 | 0.9994 |
| 第 3 のモデル | | | |
| A の順位 | 0.9992 | 0.9988 | 0.9996 |
| C の順位 | 0.9994 | 0.9990 | 0.9998 |
| E の順位 | 0.9992 | 0.9988 | 0.9996 |
| 旧表ダミー | 3828.3887 | 883.6199 | 16586.9516 |
| 表外ダミー | 3.7719 | 2.4254 | 5.8661 |
| 例外ダミー | 14.0522 | 1.4886 | 132.6467 |

表 5 を見ると 3 つのモデル全てにおいて，いずれかの調査の順位が下がれば，新表に入りにくくなることが判る．このことは直感とも整合的である．ただし，どのモデルにおいてもオッズ比間の差に大きな開きがないことから調査間の重みには大きな差がないものと考えられる．このことから基本資料と補助資料との間にも新表への入り易さを決定づけるという点において大きな差はないと言える．

また，第 3 のモデルの旧表ダミーのオッズ比からは旧表に入っていることが新表に入る確率を約 3828 倍高めることが判る．すなわち，旧表にある漢字と旧表にない漢字と間には大きな差があると言える．実際の審議においても旧表にある漢字と旧表にない漢字とが分けられており，旧表にある漢字のうち凸版 (3) 調査において 2500 位以上のものは個別の検討を経ることなく，無条件に新表に入れられている (文化庁 2011)．旧表にある漢字の出現頻度がもともと高いということも勿論だが，このことからは，やはり旧表にある漢字は新表に入り易いということが裏づけられる．

「表外漢字字体表」は旧表にない漢字で，使用頻度が高く，旧表にある漢字と共に使われるものを収めている (文化庁 2011)．この表にある漢字も出現頻度がもともと高いものと考えられるが，この表にあることも，やはり審議では有利に働いたものと見られる．

今回の審議では都道府県名とそれに準ずる漢字とを新表に入れる例外とした (文化庁 2011)．オッズ比を見ると表外ダミーの約 3.8 倍よりも高く，約 14 倍であることから例外の大きさが察せられる．

次に各モデルの的中率について表 6 に示した．

表 6: 3 つのモデルの的中率

| | 全体の的中率 | 異同字の的中率 |
|---|---|---|
| 第 1 のモデル | 87.97% | 57.73% |
| 第 2 のモデル | 86.77% | 61.86% |
| 第 3 のモデル | 94.25% | 17.01% |

まず，全体の的中率について見ると，第 1 のモデルや第 2 のモデルでは 87%前後であるのに対し，ダミー変数を加えた第 3 のモデルでは約 94%である．ここで的中精度を比較するために第 1 のモデルのそれ (図 1) と第 3 のモデルのそれ (図 2) とを示す．なお，作図の際に前後左右の振動を施した．

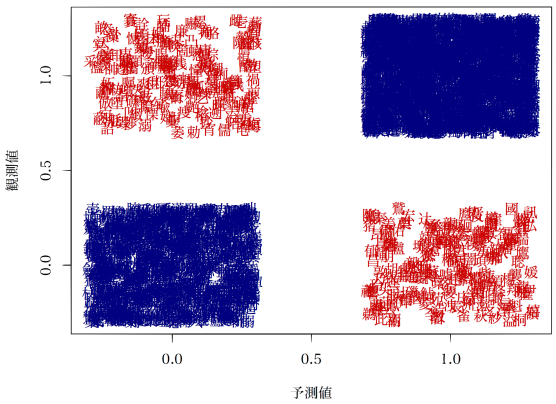


図 1: 第 1 のモデルの的中精度

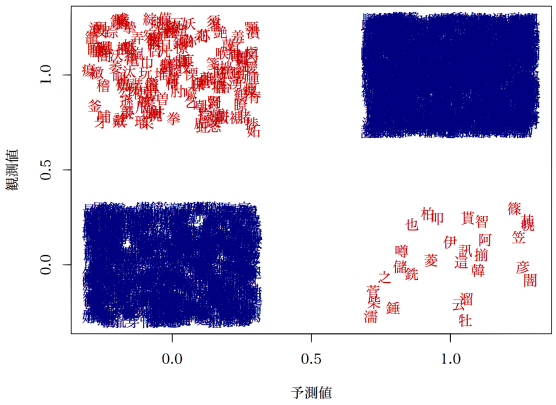


図 2: 第 3 のモデルの的中精度

両者を較べると右下の部分に大きな変化が見られる．この部分はモデルが新表に入ると予測したもののうち実際には新表に入らなかったものを示している．第 3 のモデルにおいて，この部分が減少したということはダミー変数の追加が新表に入るという予測の精度向上に寄与したということである．

しかし，このことは追加したダミー変数では新表に入らないことを決定づける要因を十分に捉えていない

ことを示唆している．また，第 1 のモデルが順位のみから成るモデルであることを鑑みれば，順位の上では新表に入らないものが実際の新表に多く盛り込まれていることが判る．

次に異同字の的中率について見ると表 6 から第 1 のモデルや第 2 のモデルでは 60%程度であるのに対し，第 3 のモデルでは 20%未満である．このことから異同字に関しては漢字の順位によって決定づけられる部分が大きいものと考えられる．ただし，第 2 のモデルにおける異同字の的中率も約 62%に留まっており，残りの部分を説明する変数が引き続き求められている．

## 4　おわりに

本稿では常用漢字表の改定のモデル化を試みてきた．検討した 3 つのモデルのうち全体の的中率が最も高いのは第 3 のモデル (94.25%) である．しかし，異同のある漢字の的中率について見ると最も高いのは第 2 のモデル (61.86%) であった．

本稿では審議資料の「漢字出現頻度数調査」の順位の一部をモデルに組み入れた．今後は出現頻度を使用することの検討に加え，この調査自体の検証も必要である．他にも字種の選定と字体の選定との分離や造語力の変数化など残る課題は枚挙に遑がない．

しかし，統計的なモデルによる漢字表の評価はまだ始まったばかりである．将来，新たに示される漢字表に向けて今後の進展が望まれる．

## 謝辞

執筆にあたり，筑波大学システム情報系の佐野幸恵助教からご指導いただいた．また，幸いにも国立国語研究所理論・構造研究系の横山詔一教授からもご指導いただく機会を得た．ここに記して感謝の意を表する．

## 補遺

初めに 3 つの調査に対し，Spearman の順位相関係数を求めた．凸版 (3) 調査と朝日新聞調査とでは 0.9145，朝日新聞調査とウェブ調査とでは 0.9169，ウェブ調査と凸版 (3) 調査とでは 0.9216 である．この結果が示す通り，3 つの調査には非常に高い相関が認められる．

このような場合には多重共線性について考える必要があり，これを確認する方法として分散拡大係数 (VIF) が知られている．また，多重共線性が認められたときは主成分分析で得られた主成分得点を変数とする方法がある．この点を踏まえ，本稿のモデルに対して VIF を求めたところ，全て 2 未満であった．このため多重共線性は問題とならないものないものと結論する．

なお，蛇足ながら，主成分得点の撒布図を図 3 に示す．第 1 主成分のみに縮約した場合の寄与率は 93.24%である．また，2 つの主成分得点を変数としたモデルの的中率は第 1 のモデルと概ね同様であったが，解釈が複雑になるために本稿では採用を見送った．

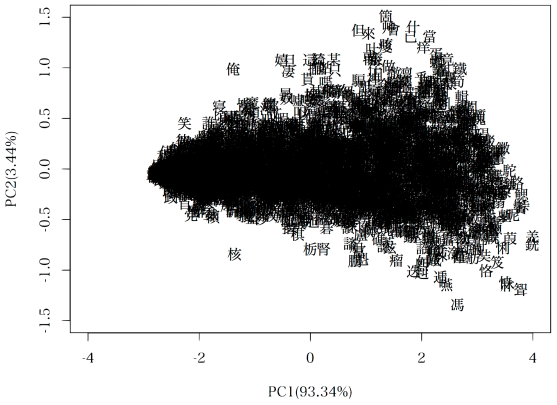


図 3: 主成分得点の撒布図

## 参考文献

[1] 小椋秀樹. (2011). 漢字使用の実態–表外訓・表外字の使用について (特集 いま, 漢字は)–(日本語社会と漢字・漢語). 国文学: 解釈と鑑賞, 76(1), 67-75.

[2] 島村直己. (2011). 改定常用漢字表を検証する. 国語教育史研究, (12), 38-41.

[3] 田野村忠温. (2008). 日本語研究の観点からのサーチエンジンの比較評価: Yahoo!と Google の比較を中心に. 計量国語学, 26(5), 147-157.

[4] 野崎浩成・江島徹郎・梅田恭子. (2012). 「常用漢字表」改訂に関する研究: 常用漢字表から削除された漢字の辞書掲載状況の分析. 愛知教育大学研究報告: 教育科学編, 61, 207-211.

[5] 文化審議会国語分科会漢字小委員会. (2008). 漢字出現頻度数順位対照表 (Ver.1.3). http://www.bunka.go.jp/seisaku/bunkashingikai/kokugo/kanji_kako/22/pdf/sanko_3.pdf.

[6] 文化庁. (2011). 新訂公用文の書き表し方の基準 (資料集). 第一法規出版.

[7] 横山詔一. (2006). 意思決定理論を援用した漢字研究 (新常用漢字表の作成に向けて)–(新常用漢字表への期待と問題点). 日本語学, 25(11), 105-114.