

異体字の単純接触効果に関するロジスティック回帰分析

—コーパス 4 種と最尤推定法を用いた検討—

横山 詔一・エリク＝ロング (国立国語研究所)¹

ディスクリプタ：異体字，選好，ロジスティック回帰分析，
最尤推定法，コーパス

1. はじめに

1. 1 頻度と好意度の関係

日本の新聞で使用頻度が一番高い漢字は「日」であり，その使用率は新聞紙面に登場する漢字の 1.4%を占めることが知られている (横山・笹原・野崎・ロング，1998)。新聞をよく読む人であれば，一生を通じて「日」に接触する頻度は膨大なものになるだろう。米国の印刷物で一番よく使用される単語は「The」である (Kučera & Francis, 1967)。米国民のなかには，一生で「The」を数百万回以上は見聞きする人が珍しくないと考えられる。この種の言語接触は，どのような単純接触効果 (mere exposure effect) を引き起こすのだろうか。ここで単純接触効果とは，なじみ (親近度：familiarity) のない新奇な刺激に繰り返し接触しているだけで，その刺激に対する好意度 (favorability) が高まる現象をいう。

ある商品や政党の宣伝がマスメディア等で全国に流れた場合，その商品や政党に対する接触頻度は社会全体で高くなり，大規模な単純接触効果を生み出す可能性がある。ただし，その効果を実際に測定し，検証するのは容易ではない。商品や政党の選好 (preference) についての検討は，心理学だけではカバーできない多くの要因が複雑に絡んでいるため，経済学，経営学，政治学の守備範囲である。一方，言語における単純接触効果は，社会言語学 (sociolinguistics) や言語心理学 (linguistic psychology) の要因に狙いを絞ることができる。接触頻度と好意度の関係を正確に究明したいとき，言語は格好の刺激材料となるだろう。そのための方法論を経済学，経営学，政治学は十分には持っていない。

Zajonc (1968) は，接触頻度が好意度に影響する程度を語彙表データによって検討した。米国英語の各種辞書から「able – unable, better – worse, clean – dirty, good – bad, peace – war, life – death」などの対義語 154 ペアを抽出し，ペアのそれぞれについてどちらの単語が好きかを実験参加者に 2 肢強制選択法 (2-Alternatives Forced Choice) で尋ねた。そして Thorndike & Lorge (1944) の語彙表を用いてペアのそれぞれに使用頻度の

¹ YOKOYAMA Shoichi, Eric LONG (The National Institute for Japanese Language)—Logistic Regression Analysis of Preference for Kanji Variants: Maximum Likelihood Estimation Based on Text Corpora

データを付し、使用頻度を接触頻度の推定値とみなして選好との関係を分析した。その結果、ペアのうち実験参加者に選好される単語は使用頻度の高い方であることが示唆された。例えば able と unable のペアでは実験参加者の 100% が able を選好した。使用頻度は able が 930 で unable は 239 であり、able の使用頻度の方が高かった。Zajonc (1968) は、さらに木 (例: pine), 果物 (例: apple) などの単語を一つずつ実験参加者に呈示し、どの程度好きかを 0 ~ 6 の 7 段階評定尺度で尋ねた。その結果、単語の使用頻度と好意度の間には正の相関があることがうかがわれた。これらから、日常生活でよく使われる単語ほど好意度が高く、また好意度の高い単語ほど使用頻度が高くなるという解釈が導かれた。しかし、彼自身も認めているように、対義語の抽出方法などに問題があるほか、ペアの単語どうして文字数、発音、意味などが違っていたため、接触頻度の効果を正確に捉えていないのではないかとの疑いが残る。

1. 2 日本語で検討する意義

英語など欧米諸外国語では、どんなに工夫しても、文字数や読みが等しい刺激のペアを大量に準備することは不可能である。ところが、幸運なことに、日本語の場合は、この問題をうまく回避できる。日本語の漢字には異体字 (variant) の豊富なバリエーションが存在する。異体字とは「桧一檜」のように読みと意味は同じで字体だけが異なる文字の集合を指す。異体字を刺激材料とすれば、文字数、読み、意味が等価で、形や画数だけが異なる刺激ペアを大量に作成できる。これは日本語を材料とするメリットの一つである。英語等の諸外国の言語でも「English - eNGLISH」というように、文字数、読み、意味が等価な言語刺激のペアを作成することは不可能ではないが、ペアの一方は明らかに人工的で、現実の生活にはほとんど登場しない架空の変異 (variation) になるケースが多く、生態学的妥当性を著しく欠く。その点で、日本語における異体字は国民各層の言語生活に深く浸透しており、日常場面でペアの両方を目にする機会があるものが多いので、適切な刺激だと言えよう。

社会言語学では、異体字刺激を用いた選好の研究がすでに行われている (横山・笹原・當山, 2006)。その手法は、異体字ペアを実験参加者に呈示して字体選好課題を実施するというものである。字体選好課題とは「桧一檜」など 263 字種の新旧字体ペアを実験参加者に呈示し、携帯メールなど IT 機器で字を書く場面をイメージしたときにより使いたいと感じる方の字体を 2 肢強制選択法で直観的に選ばせる課題である。そのほか、字体親近度比較課題を実施する場合もある。その課題では、新旧字体ペア 263 ペアのうち、親近度 (なじみ: familiarity) を強く感じる方の字体を 2 肢強制選択法で判断させる。字体選好課題と字体親近度比較課題は、いずれも信頼性の高いデータを得られることが再テスト法によって示されている。

1. 3 数理モデルの先行研究

さらに、計量国語学 (mathematical linguistics) では、「一般対応法則 (the generalized matching law)」と同形の数式を使って、新聞のコーパス (corpus) で計数した漢字頻度から字体選好課題の結果を予測する試みがなされている (横山, 2006)。コーパスとは電子化された言語資料を指す。一般対応法則とは、動物の選択行動研究から Baum (1974) が導いたもので、反応比 $R1/R2$ と強化比 $r1/r2$ が式(1)のような単純な関数関係のもとで対応していることをいう。log は自然対数 (底は e), パラメータ S は反応感度, b は反

応バイアスを示す。

$$\log (R1/R2)=S \log (r1/r2)+\log b \quad \text{----- (1)}$$

反応比，強化比とは，次のような意味である。ハトやラットなどの動物がキーを押すと餌（報酬）がもらえる環境を実験室内に準備し，2つのキーを設定する。例えば，右側のキーを1，左側のキーを2としよう。1を押すと報酬が出る頻度はr1，2を押すと報酬が出る頻度はr2とする。この報酬頻度の比が強化比である。訓練（学習）を受けた動物がキーを押す反応の頻度を観察すると，1を押す反応頻度はR1，2を押す反応頻度はR2となる。この反応頻度の比が反応比である。左右キーの反応比は，強化比と，式(1)の形で対応することが数多くの研究で示されている（Woolverton & Alling, 1999; Belke & Belliveau, 2001）。

この式(1)は，生態学の分野において野生動物集団がいくつかの餌場間でどのように分布するかを説明する「理想自由分布理論（the ideal free distribution theory）」とも一脈通じる部分がある。理想自由分布理論（Fagen, 1987; 山口・伊藤, 2006）におけるモデル式も，個体分布比をR1/R2，餌の量の比をr1/r2とすれば，式(1)と一致する。

今後，自然言語における単純接触効果の研究を進めるにあたり，選好の心的プロセスを視野に入れた計量的な分析法を手にしておくことはきわめて重要であろう。横山（2006）とYokoyama & Wada（2006）は，式(1)が「ロジスティック回帰分析（logistic regression analysis）」の形になっていることを指摘した。ロジスティック回帰分析は医学統計などのほか，社会言語学のLabov（1972）によって開拓された変異理論（variation theory）の分野でも盛んに利用されている（Matsuda, 1993）。そのモデル式は式(2)になる。Zは線型の関数である。2肢強制選択で選択肢1と2のうち1を選ぶ確率をp1とおくと，2を選ぶ確率は1-p1となり，式(2)の左辺に含まれる項p1/(1-p1)を「オッズ（odds）」という。

$$\log \{p1/(1-p1)\}=Z \quad \text{----- (2)}$$

選択肢1と2の反応頻度をそれぞれR1，R2とおけば，反応の合計頻度NはR1+R2である。選択肢1を選ぶ確率p1はR1/Nで，オッズは式(3)になる。オッズの対数を「対数オッズ」あるいは「ロジット」という。

$$\begin{aligned} p1/(1-p1) &= (R1/N)/(R2/N) \\ &= R1/R2 \quad \text{----- (3)} \end{aligned}$$

式(2)のZにS log (r1/r2)+log bを代入すると，式(4)になる。p1/(1-p1)=R1/R2であるから，式(4)は式(1)と等しい。つまり，式(1)は一般対応法則や理想自由分布理論の式と同形であり，しかもロジスティック回帰分析の式にもなっている【末尾注1】。

$$\log \{p1/(1-p1)\}=S \log (r1/r2)+\log b \quad \text{----- (4)}$$

計量国語学における式(4)の検討は次のようにしてなされている。「桧—檜」のような異体字ペアのうち、日常の言語生活で旧字体に接触する頻度を r_1 、新字体に接触する頻度を r_2 とする。字体選好課題で旧字体を選択する人数を R_1 、新字体を選択する人数を R_2 とする (式(3)から $R_1/R_2 = p_1/(1-p_1)$ となる)。それぞれの字体に対して人間がどのくらい接触しているかの頻度 r_1 と r_2 については信頼に足る実測データがどこにも存在しないため、Zajonc (1968) と同じく、コーパスでの使用頻度を接触頻度の推定値とする。コーパスで計数した漢字頻度データから式(4)の説明変数 $\log(r_1/r_2)$ を求め、最小 2 乗法によって目的変数である対数オッズ $\log\{p_1/(1-p_1)\}$ を予測してみると決定係数が .65 を超えることが示された (横山, 2006)。この予測精度は、自然言語を対象にしたこの種の研究においてはかなり高いと言える。Zajonc (1968) はこれと似た考え方を繰り返し述べているが、具体的な予測式は示していない。

1. 4 本研究の目的

以上のように式(4)は理論面で興味深い性質を有するのだが、その実際の予測力はどの程度なのだろうか。横山 (2006) と Yokoyama & Wada (2006) による先行研究には 2 つの問題があった。一つは、漢字頻度をカウントするために用いたコーパスが新聞だけに限定されていたことであり、もう一つは、ロジスティック回帰分析のパラメータ推定に最小 2 乗法を用いたことである。問題の前者については、人間の文字生活は書籍等とも関係が深いので、新聞以外のジャンルのコーパスもなるべく幅広く対象に含めるのがよい。後者については、ロジスティック回帰分析は確率モデルに基づく手法なので、「最尤推定法 (maximum likelihood estimation)」を用いるべきである。

本研究は、これらの問題を解消するため、新聞、百科事典、小説の 3 つのジャンルにおける 4 種類のコーパスを用いて新旧両字体の使用頻度をカウントした。そのデータを式(4)に投入して、パラメータの S (感度) と $\log b$ (バイアス) を最尤推定法で計算した。以下、選好と選択 (choice) を区別しないで用いることがある。

2. 実験 1 : 字体選好

方法

刺激材料 刺激材料の一部を Table 1 に示す。ペアの呈示順序と新旧字体の左右位置はランダム化された。刺激項目は新旧 263 ペアで、以下の 3 つの規準にしたがって選ばれた。

(1) JIS X0208-1983 の第 1・第 2 水準に含まれる漢字で、新字体 (拡張新字体) と旧字体 (正字体) の関係にあるもの。処理が複雑になるため、JIS 漢字に含まれる異体字の中でほとんど使われないものは原則として扱わなかった。(2) MS 明朝フォントと FA 明朝フォントで各字体が表現できるもの。この基準を導入した理由は、ワープロソフトによる印字の制約を考慮したことによる。(3) 上記 2 つの規準に適合した異体字集合から、被調査者になじみがないと思われる字や、字種が多いグループの字を原則として削除した。

漢字頻度データ 異体字 263 ペアのうち、新旧両字体とも JIS X0208-1983 で表示可能な 86 ペアについて新旧字体頻度を計数した。新旧字体頻度は、次の 4 種類のコーパスによって計数した。なお、電子化テキストの字体はコーパス間で統一されていない (Long & Yokoyama, 2005)。

Table 1 新聞コーパスの旧字体頻度ロジットで排列した京都4月データの一部

ID	新旧 字体ペア	旧字体 頻度 ロジット	字体 選好 課題(%)	ID	新旧 字体ペア	旧字体 頻度 ロジット	字体 選好 課題(%)
68	尔爾	3.7297	62.5	29	狭狭	-6.8554	15.3
138	釵鐸	3.6636	58.3	54	国國	-6.8888	6.9
39	頸頸	1.9218	83.3	19	覚覚	-7.1291	0
69	迓邇	1.7918	45.8	121	錢錢	-7.1325	8.3
182	蕊藁	1.6094	33.3	90	繩繩	-7.1412	0
176	箆籠	1.3072	79.2	32	焼焼	-7.1523	1.4
122	賤賤	1.2040	72.2	7	陥陥	-7.1593	4.2
177	鼠鼠	0.0953	83.3	65	齒齒	-7.3595	0
109	俛儘	0	37.1	137	訳譯	-7.3658	4.2
174	竜龍	-0.0929	34.7	127	騷騷	-7.3902	1.4
31	堯堯	-0.4618	33.3	136	駅驛	-7.4134	5.6
92	竈竈	-0.6931	18.1	13	誉譽	-7.6958	1.4
169	遙遙	-0.9008	40.3	96	飲飲	-7.7297	2.8
52	砵礪	-0.9808	22.2	9	奥奥	-7.8598	2.8
3	壺壺	-1.0498	88.9	204	卷卷	-7.9697	13.9
111	藪藪	-1.0745	38.9	150	読讀	-8.0790	4.2
14	鶯鶯	-1.2238	58.3	66	齡齡	-8.1253	1.4
76	涛涛	-1.2252	40.3	38	経経	-8.2610	1.4
24	灌灌	-1.3863	90.3	60	贊贊	-8.3638	1.4

(中略)

(後略)

- (1) 朝日新聞データ：横山・笹原・野崎・ロング（1998）による朝日新聞コーパスの解析結果によるもの。1993年1月1日から12月31日の間に朝日新聞社東京本社管内で発行された最終版の朝刊および夕刊で、『CD-HIASK'93 朝日新聞記事データベース』（朝日新聞社，1994）の電子化テキストに基づく。漢字の延べ数は1,707万字で、異なり数は4,562であった。
- (2) 小学館百科事典データ：『スーパー・ニッポニカ 2001』（小学館，1998-2000）の電子化テキストに基づく。漢字の延べ数は2,357万字で、異なり数は7,103であった。これと以下2つのデータはLong & Yokoyama（2005）による。
- (3) 平凡社百科事典データ：『世界大百科事典』（平凡社，1998）の電子化テキストに基づく。漢字の延べ数は2,547万字で、異なり数は7,420であった。
- (4) 新潮社小説データ：『新潮文庫の100冊』のほか明治期と大正期の文豪や絶版の作品を掲載した計4枚のCD-ROMに基づく（新潮社，1995，1997a，1997b，2000）。これらのうち『新潮文庫の100冊』の作品は他のCD-ROMにも重複して収録されて

いる場合がある。その重複部分は削除して『新潮文庫の 100 冊』の収録作品を残した。漢字の延べ数は 1,698 万字で、異なり数は 6,221 であった。

手続き 実験の冒頭で「ワープロを打っている場面だけをイメージするように」と伝え、異体字のペアを実験参加者に呈示して、より使いたいと感じる方の字を選択させた。具体的な教示は次の通り。「この実験は、漢字の使われ方を調べるものです。これから、字の形は違いますが、読みと意味がまったく同じ漢字のペアをお見せします。たとえば「断」と「斷」は、同じ読みで同じ意味の漢字のペアです。もし、あなたがワープロを打っているとしたら、どちらの字を使いたい、教えてください。2つの漢字をよく見て、使いたいと感じる程度を比較し、より使いたいと思う方の字に○印をつけてください。両方とも使いたい、あるいは両方とも使いたくないと感じるペアがあるかも知れませんが、とにかく、どちらか一方の字だけに○印をつけてください。判断は、あまり深刻に悩まずに、直観的に行ってください。(以下略)」

実験参加者 異体字ペア 263 ペアの字体選好課題データは次の 3 種類のデータセットから構成されていた。

- (1) 東京データ：東京都内の女子大学生 85 名を対象に 1996 年から 1997 年にかけて収集した。
- (2) 京都 4 月データ：京都市内の立命館大学の学生 72 名（男性 20 名，女性 52 名）を対象に 1998 年 4 月に収集した。
- (3) 京都 9 月データ：先の京都選好 4 月データの実験参加者を対象に同じ実験を約半年後に実施した。実施時期は 1998 年 9 月。

結果と考察

異体字 263 ペアのそれぞれについて旧字体選好率を算出した。(100%から旧字体選好率を引き算すれば新字体選好率が得られる。) 263 ペアのうち新旧字体ともに JIS X0208-1983 で表示可能な 86 ペアを抽出した。これ以外の 177 ペアは旧字体が JIS X0208-1983 で表示できないため、以後の分析からは除外した。Table 1 に新聞コーパスから得た旧字体頻度ロジット $\log(r1/r2)$ と、京都 4 月データによる旧字体選択率のうち、86 ペアの一部を示す。

回帰分析の結果 この実験は、より使いたい方の字体を選択する課題であった。R1 を旧字体選択人数、R2 を新字体選択人数、r1 をコーパスで計数した旧字体頻度、r2 を新字体頻度として式(1)のロジスティック回帰分析を行った。具体的には、目的変数を東京データ、京都 4 月データ、京都 9 月データのそれぞれにおける R1/R2 とし、説明変数を朝日新聞、小学館百科事典、平凡社百科事典、新潮社小説のコーパスごとの r1/r2 とした。パラメータは SPSS による最尤推定法で求めた。得られたすべてのパラメータについて Wald 検定を実施し、いずれも有意 ($p < .01$) であることを確認した。

式(4)を変形すると式(5)が得られ、異体字ペアごとに旧字体選択確率 $p1$ の予測値を求めることができる。

$$p1 = 1 / \{1 + \exp[-S \log(r1/r2) - \log b]\} \quad \text{----- (5)}$$

この式(5)に小説コーパスから求めた頻度を代入し、旧字体選択確率の予測値を求めた。

それと実測値の散布図を Figure 1 に示す。

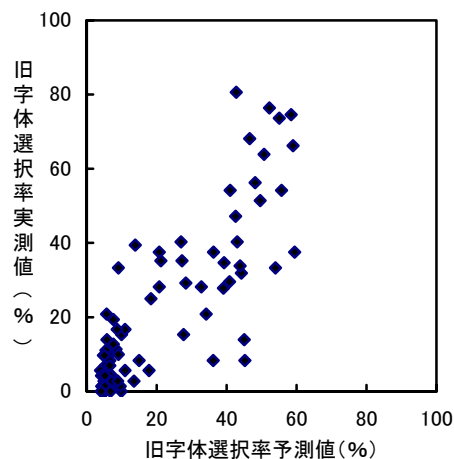


Figure 1 選好課題（京都9月）の予測値と実測値の散布図

また、Table 2 には、すべてのパラメータと決定係数を示す。予測値と実測値の積率相関を2乗した決定係数はいずれも.500を越え、最大は.697に達した（無相関検定の結果はすべて $p < .01$, $df = 84$ ）。

決定係数をコーパス間で比較してみると、小説が東京データ、京都4月データ、京都9月データのいずれにおいても高く、他の3つのコーパスより優れていた。逆に、新聞はいずれにおいても最小値であった。百科事典の2つのコーパスは、小説と新聞の中間に位置付いていた。

Table 2 選好課題のパラメータ（Sは感度、log bはバイアス）

コーパス	地域と時期								
	東京			京都4月			京都9月		
	r^2	S	log b	r^2	S	log b	r^2	S	log b
新聞	.517	0.294	-0.301	.623	0.379	-0.147	.646	0.323	-0.232
事典 a	.604	0.195	-1.075	.659	0.237	-1.148	.688	0.205	-1.082
事典 b	.590	0.165	-1.008	.670	0.203	-1.063	.657	0.171	-1.011
小説	.612	0.209	-0.957	.696	0.259	-1.010	.697	0.220	-0.959

注) 決定係数はすべて有意 ($p < .01$)。また、いずれのパラメータも Wald 検定で有意 ($p < .01$)。

以上のようにかなり高い予測精度が得られた理由として、異体字刺激を用いた利点を指摘できる。Zajonc (1968) のように対義語ペアを実験参加者に呈示して選好課題を行う場合は、刺激ペア間で文字数、読み、意味のいずれかに違いが生じてしまうため、それが攪乱要因となって単純接触効果の純粋な決定係数を求めることが極めて困難だった。本研究の異体字ペアはこの問題が生じないので、実験1で得られた決定係数約.500~.697という数

値は、言語表現の選好に及ぼす単純接触効果の影響の大きさをかなり正確に捉えていると言えるだろう。ある文字の社会における使用頻度が高い場合は、人間がその文字に接触する確率が高くなり、接触確率が高くなると単純接触効果が生じて好意度が高くなると考えられる。

3. 実験 2：字体親近度比較

実験 2 は、親近度をより強く感じる方の字体を 2 肢強制選択法で判断させる字体親近度比較課題を 263 ペアに実施した。

方法

刺激材料 実験 1 と同じ。

漢字頻度データ 実験 1 と同じ。

説明変数と目標変数の定義 実験 1 と同じ。

手続き 異体字のペアを被調査者に示して、いずれの漢字によりなじみを感じるか判断させた。実験に先立って以下の教示を与えた。「この実験は（中略、好み調査と同文）漢字のペアです。2 つの漢字をよく見て、見慣れていると感じる程度を比較し、より見慣れていると思う方の字に○印をつけてください。両方とも全く見慣れないペアがある場合は、より見慣れていると感じる方の字を選んで、○印ではなく、△印をつけてください。判断は、あまり深刻に悩まずに、直観的に行ってください。」

実験参加者 異体字ペア 263 ペアの字体親近度比較課題データは次の 3 種類のデータセットから構成されていた。

- (1) 東京親近度比較データ。東京都内の女子大学生 98 名を対象に 1996 年から 1997 年にかけて収集した。
- (2) 京都親近度比較 4 月データ。京都市内の立命館大学の学生 65 名（男性 29 名、女性 36 名）を対象に 1998 年 4 月に収集した。
- (3) 京都親近度比較 9 月データ。先の京都親近度比較 4 月データの実験参加者を対象に同じ実験を約半年後に実施した。実施時期は 1998 年 9 月。

なお、この 3 つのデータに関する実験参加者は、誰も実験 1（字体選好課題）には参加していない。

結果と考察

異体字 263 ペアのそれぞれについて旧字体選択率を算出した。実験 1 と同様、263 ペアから新旧字体ともに JIS X0208-1983 で表示可能な 86 ペアを抽出した。データは△印と○印の区別をしないでコーディングした。Table 1 に新聞コーパスの旧字体頻度ロジット $\log(r1/r2)$ のうち、86 ペアの一部を示す。

回帰分析の結果 実験 2 は、よりなじみのある方の字体を選択する課題であった。実験 1 と同様のロジスティック回帰分析を行った。目的変数を東京データ、京都 4 月データ、京都 9 月データのそれぞれにおける $R1/R2$ とし、説明変数を朝日新聞、小学館百科事典、平凡社百科事典、新潮社小説のコーパスごとの $r1/r2$ とした。パラメータは SPSS による最尤推定法で求めた。得られたすべてのパラメータについて Wald 検定を実施し、いずれも有意であることを確認した ($p < .01$)。

小説コーパスの頻度から京都 4 月データの旧字体選択率を予測した値と実測値の散布

図を Figure 2 に示す。

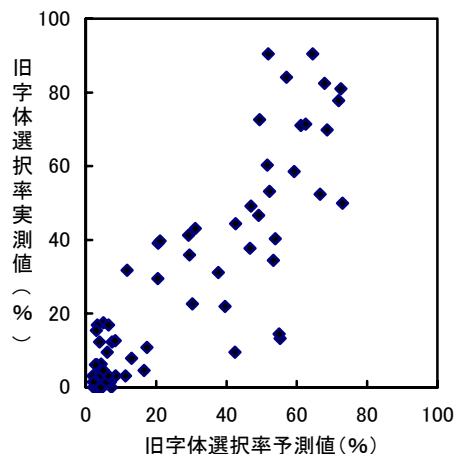


Figure 2 親近度比較課題（京都 4 月）の予測値と実測値の散布図

また、Table 3 には、すべてのパラメータと決定係数を示す。旧字体選択率の予測値と実測値の積率相関の 2 乗である決定係数はすべて .574 を越え、最大値は .785 に達した（無相関検定の結果はすべて $p < .01$, $df = 84$ ）。

コーパス間で決定係数を比較すると、小説が、東京データと京都 4 月データにおいて高く、他の 3 つのコーパスより優れていた。逆に、新聞は東京データと京都 4 月データにおいて最小値であった。百科事典の 2 つのコーパスは、小説と新聞のほぼ中間に位置付いていた。

Table 3 親近度比較課題のパラメータ（S は感度，log b はバイアス）

コーパス	地域と時期								
	東京			京都 4 月			京都 9 月		
	r^2	S	log b	r^2	S	log b	r^2	S	log b
新聞	.574	0.449	-0.018	.687	0.434	-0.165	.715	0.382	-0.016
事典 a	.715	0.307	-1.221	.769	0.275	-0.989	.755	0.241	-1.025
事典 b	.703	0.257	-1.099	.718	0.223	-0.882	.688	0.196	-0.936
小説	.753	0.349	-1.070	.785	0.300	-0.835	.750	0.257	-0.883

注) 決定係数はすべて有意 ($p < .01$)。また、いずれのパラメータも Wald 検定で有意 ($p < .01$)。

4. 総合的考察

本研究は、言語生活で生じる単純接触効果の大きさを最尤推定法によるロジスティック回帰分析で予測した。実験 1 では字体選好課題を実施した。小説コーパスにおける新旧字体間の頻度比の対数を説明変数とした場合、旧字体選択確率の予測値と実測値の決定係数は最高で .697 を示した。実験 2 では字体親近度比較課題を実施した。小説コーパスにおけ

る新旧字体間の頻度比の対数を説明変数とした場合、旧字体選択確率の予測値と実測値の決定係数は最高で.785に達した。

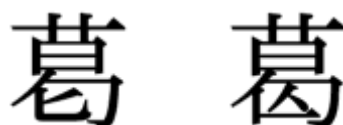
決定係数を、Table 2（実験1：字体選好課題）とTable 3（実験2：字体親近度比較課題）で比較すると、実験2の方が確かに高い。これは単純接触の効果が親近度比較課題に対してより直接的に影響しているためだと解釈するのが自然であろう。異体字の選択においては「頻度→接触確率→親近度→好意度」という流れがあるように見える。Monin (2003)は、新奇刺激で美しい顔とそうでない顔を実験参加者に呈示すると、美しい顔の親近度評定の方が高くなると報告している。これは「好意度→親近度」という流れも存在していることを示しているが、異体字ペアの場合は、そのような影響があまり強くはないのかもしれない。

コーパス間で決定係数を比較すると、実験1と2の両方において小説が高くなる傾向があった。その次に高いのは百科事典であり、新聞が一番低い傾向にあった。ただし、調査の時期や場所によって決定係数の違いは僅差になる場合があり、小説が顕著に優れているとは言えない。今回のデータだけからコーパス間で差があるとの結論を導くのは危険であろう。今後の詳しい研究を待つ必要がある。地域間で決定係数を比較すると、いずれの実験においても京都が東京より高くなる傾向があった。ただし、東京が全員女性であったのに対して、京都は男性も25%以上は参加していたことから、地域差ではなく性差が影響した可能性が考えられる。この点も今後の検討が待たれる。

京都の4月データと9月データは同じグループから得たため、約半年の期間をおいた再テスト法になっていた。決定係数ならびにパラメータの値は、約半年後でもあまり変化していないように見える。字体選好課題と字体親近度比較課題は、いずれも信頼性の高いデータを得られるとの報告がある（横山ほか、2006）。本研究は、課題の信頼性をロジスティック回帰分析の角度から再確認したと言えよう。

文字は公共財であり、言語生活を支えている。文字政策が変化すると、ある文字に対する単純接触効果にも変化が生じる可能性があるだろう。例えば、マイクロソフト社が2007年1月30日に発売したWindows OS「ビスタ（VISTA）」は、「JIS X 0213:2004」に対応した字体（字形）を搭載している。JIS X 0213:2004とは、経済産業省がJIS漢字（JIS X0213）の160字種あまりについて規格書の例示字形を2004年2月に変更したものを指す。周知のようにJIS漢字規格は文字政策の一つとして強い影響力を持っている。

當山（2006）や横山・高田・米田（2006）は、ビスタの登場が国民全体の文字生活に何らかの変化をもたらすと示唆している。ビスタ登場以前は、通常の電子メールでFigure 3(b)に示す「葛（ヒ→L+人）」を使いたい場合でも、原則としてFigure 3(a)の「葛」しか使えなかった。



(a)

(b)

Figure 3 JIS漢字で例示字形が変更された例

ところが、ビスタではこの関係が逆転し、原則として Figure 3(b)の「葛 (ヒ→L+人)」が IT 機器に標準装備され、それ以前の標準であった「葛」(Figure 3(a))は標準的な字体ではなくなる。文字政策が変化した影響はいずれ日本全国に及び、大規模な単純接触効果として観測される可能性がある。そのデータ解析に本研究の手法が有用となるだろう。

冒頭部で述べたように、ある商品や政党の宣伝がマスメディア等で全国に流れた場合、その商品や政党に対する接触頻度は社会全体で高くなり、大規模な単純接触効果を生み出す可能性があるものの、その効果を検証するのは容易ではない。しかし、文字政策の変化に伴う全国的な単純接触効果は、社会言語学や言語心理学の手法で捉えることができそうである。接触頻度と好意度の関係を正確に究明したいとき、異体字選好課題は有力な検証手段になると考えられよう。

【注 1】式(1)から式(4)までの変形については別のやり方もあるが、ここでは式(2)の Z に代入する式を明示できるように心がけた。これは Z がリンク関数という役割を果たしているからである。リンク関数の説明は紙幅などの関係でほかの機会に譲る。

謝 辞

本研究で使用した字体選好課題と字体親近度比較課題のデータは、笹原宏之氏（早稲田大学社会科学総合学術院）、當山日出夫氏（花園大学）、横山の3名が共同で採取したものである。データ使用の許諾をくださった笹原氏と當山氏に深く感謝申し上げます。

引用文献

- 朝日新聞社 (1994). CD-HIASK'93 朝日新聞記事データベース
- Baum, W. M. (1974). On two types of deviation from the matching law: Bias and undermatching. *Journal of the Experimental Analysis of Behavior*, 22, 231-242.
- Belke TW, & Belliveau J. (2001) The general matching law describes choice on concurrent variable- interval schedules of wheel-running reinforcement. *Journal of the Experimental Analysis of Behavior*, 75(3), pp.299-310.
- Fagen, R. (1987). A generalized habitat matching rule. *Evolutionary Ecology*, 1, 5-10.
- 平凡社 (1998). 世界大百科事典
- Kučera, H., & Francis, W. N. (1967). *Computational analysis of present-day American English*. Providence, RI: Brown University Press.
- Labov, W. (1972). *Sociolinguistic patterns*. Philadelphia: University of Pennsylvania Press.
- Long, E. & Yokoyama, S. (2005). Text genre and kanji frequency, *Glottometrics*, 10, 55-72
- Matsuda, K. (1993). Dissecting analogical leveling quantitatively : The case of the innovative potential suffix in Tokyo Japanese. *Language Variation and Change*, 5, 1-34.

- Monin, B. (2003). The warm glow heuristic: When liking leads to familiarity. *Journal of Personality and Social Psychology*, 85, 1035-1048.
- 新潮社 (1995). 新潮文庫の 100 冊
- 新潮社 (1997a). 新潮文庫 明治の文豪
- 新潮社 (1997b). 新潮文庫 大正の文豪
- 新潮社 (2000). 新潮文庫の絶版 100 冊
- 小学館 (1998-2000). スーパー・ニッポニカ 2001
- 當山日出夫 (2006). 「京都における「葛」と「祇」の使用実例と「JIS X 0213 : 2004」—非文献資料に基づく考察— 『情報処理学会研究報告 2006-CH-70』, 53-60
- Thorndike, E. L., & Lorge, I. (1944). *The teacher's word book of 30,000 words*. New York: Columbia University, Teachers College Press.
- Woolverton WL, & Alling K. (1999) Choice under concurrent VI schedules: comparison of behavior maintained by cocaine or food. *Psychopharmacology*, 141(1), pp.47-56.
- 山口哲生・伊藤正人 (2006). 理想自由分布理論に基づく個体分布の実験的検討——絶対報酬量と集団サイズの効果, *心理学研究*, 76, 547-553
- 横山詔一 (2006). 異体字選好における単純接触効果と一般対応法則の関係, *計量国語学*, 25, 199-214
- Yokoyama, S. & Wada, Y. (2006). A logistic regression model of variant preference in Japanese kanji: an integration of mere exposure effect and the generalized matching law. *Glottometrics*, 12, 63-74.
- 横山詔一・笹原宏之・當山日出夫 (2006). 文字コミュニケーションにおける異体字の選好と親近度: 再検査法による信頼性の検討, *社会言語科学*, 9, 16-26
- 横山詔一・高田智和・米田純子 (2006). 東京山の手と葛飾・葛西における文字生活の地域差, *人文科学とコンピュータシンポジウム*, 379-386, 情報処理学会
- 横山詔一・笹原宏之・野崎浩成・エリク＝ロング (1998). 新聞電子メディアの漢字——朝日新聞 CD-ROM による漢字頻度表, 国立国語研究所プロジェクト選書No.1, 三省堂
- Zajonc, R. B. (1968). Attitudinal effects of mere exposure. *Journal of Personality and Social Psychology*, 9, 1-27.

(2007年4月21日受付)

(日本語抄録)

調査報告

異体字の単純接触効果に関するロジスティック回帰分析

ーコーパス4種と最尤推定法を用いた検討ー

横山 詔一・エリク＝ロング (国立国語研究所)

日本の新聞で使用頻度が一番高い漢字は「日」であり、その使用率は新聞紙面に登場する漢字の1.4%を占める(横山・笹原・野崎・ロング, 1998)。新聞をよく読む人であれば、一生を通じて「日」に接触する頻度は膨大なものになるだろう。この種の言語接触は、どのような単純接触効果(mere exposure effect)を引き起こすのだろうか。ここで単純接触効果とは、なじみ(親近度:familiarity)のない新奇な刺激に繰り返し接触しているだけで、その刺激に対する好意度が高まる現象をいう。

ある商品や政党の宣伝がマスメディア等で全国に流れた場合、その商品や政党に対する接触頻度は社会全体で高くなり、大規模な単純接触効果を生み出す可能性がある。ただし、その効果を検証するのは容易ではない。商品や政党の選好(preference)は多くの要因が複雑に絡む。一方、言語における単純接触効果は、社会言語学や言語心理学の要因に狙いを絞ることができる。接触頻度と好意度の関係を正確に究明したいとき、言語は格好の刺激材料となるだろう。そのための方法論を経済学、経営学、政治学は十分には持っていない。

本研究は、新聞、百科事典、小説の3つのジャンルにおける4種類のコーパスを用いて新旧両字体の使用頻度をカウントした。そのデータから一般対応法則(理想自由分布理論)に立脚したモデルのパラメータを最尤推定法で求めた。実験1では字体選好課題を実施した。小説コーパスにおける新旧字体間の頻度比の対数を説明変数とした場合、旧字体選択確率の予測値と実測値の決定係数は最高で.697を示した。実験2では字体親近度比較課題を実施した。小説コーパスにおける新旧字体間の頻度比の対数を説明変数とした場合、旧字体選択確率の予測値と実測値の決定係数は最高で.785に達した。

文字政策が変化した影響はいずれ日本全国に及び、大規模な単純接触効果として観測される可能性がある。そのデータ解析に本研究の手法が有用となるだろう。

REPORT

Logistic Regression Analysis of Preference for Kanji variants: Maximum Likelihood Estimation Based on Text Corpora

YOKOYAMA Shoichi, Eric LONG (The National Institute for Japanese Language)

Descriptors: Kanji, variant, preference, familiarity, logistic regression analysis, maximum likelihood estimation, mere exposure effect, the generalized matching law

Abstract:

Japanese kanji characters often exhibit variants, which are often pairs of traditional and simplified forms, representing the same pronunciation and meaning. This paper focuses on familiarity with and preference for variants, which are intra-personal psychological variables, in recognition of kanji variants by educated native speakers of Japanese. We examined their correlation with frequency based on data obtained from corpora, which potentially explains familiarity and preference behaviors. The frequency data were obtained from four corpora of three genres: newspaper, encyclopedia, and literary works. Recognition of variants was measured by two tasks, preference and familiarity judgment tasks, in which the participants were asked to choose the more preferable or familiar one of the paired variants. The study also examined which of the two psychological behaviors, i.e. selection based on familiarity or preference, were more accurately estimated by the frequency data. The contribution of the frequency data to familiarity and preference behavior were analyzed and compared using logistic regression models, based on the generalized matching law. The predictive power across genres of corpora was also compared. The analyses indicated that: (1) The model incorporating frequency data efficiently and reliably explains and estimates the preference and familiarity behaviors; (2) The frequency data estimated the familiarity performance more accurately than preference performance, suggesting a more direct contribution of frequency to familiarity than to preference; and (3) The corpus of literary works best explains preference and familiarity behaviors among the three genres.