「電子メール」とその競合的同義語の選択に関わるメカニズムの分析（1）

福田　薫

1．はじめに

1980年代後半から90年代にかけて，パソコンが普及し，インターネットを利用できる人口が急速に拡大したことに伴って，パソコンによるメッセージの送受信，いわゆる「電子メール」の利用が爆発的に増加した．さらに，2000年以降は携帯電話の普及が進んで，いわゆる「携帯メール」の利用が増加している．

このような社会情勢の変化を反映して，「電子メール」という新語が使われ始め，その後，使用頻度が急増し，一般社会に幅広く浸透し，定着し始めた．岩波書店発行の『広辞苑』は日本語の代表的な辞書の一つである．「電子メール」という新語が，『広辞苑』に収録されたのは1991年発行の第4版からである(注1)．この事実は，おおむね，その間の事情に対応していると推測される．第4版において，すでに「電子メール」の略語としての「メール」が見出し語に立てられており，「電子メール」の異形態としての「イー・メール」にも言及がある．

一般に，文法や語法の変化は長い時間をかけて徐々に推移する．語の生長と消失も同様に長い時間を要する．本稿では，「電子メール」と「イー・メール」「メール」というIT関連用語の変異形を取り上げ，その使用推移を調査と分析の対象とする．上でも述べたように，これらはいずれも比較的最近使われ出した語であり，比較的短期間の間に生長と減衰を示し，その選択的競合は現在も進行中である．本稿の目的は，新聞データベースを用いて「電子メール」およびその変異形の使用頻度の推移を計測し，定量的な観点から分析を行い，多変量解析の手法を用いて観察されるデータの説明を試みることである．より具体的には，多項ロジットモデルによるロジスティック回帰分析を用いて，競合する語彙の選択の際に働いている説明要因を特定することである．その結果として，対象語が用いられる年代とジャンルおよび対象語全体の頻度は重要な説明要因として働いているのに対し，対象語使用者である書き手の性別，年齢という要因は説明要因として働いていないことを主張する．

以下，第2節では，調査対象と分析の方法について記述する．第3節では，ロジスティック回帰分析の結果を提示し，その意味合いを検討する．第4節はまとめと今後の課題を述べる．

2．分析の目的と方法

この節では，分析の目的と調査対象の分析方法について簡単に述べる．

2.1　目的

本稿の調査項目は「電子メール」「Eメール」「メール」および「メエル」「メイル」などの語である．これらの意味内容は基本的に同じであり，したがってほとんどの場合互いに置き換え可能であるので，異形態の関係にあると見なしうる．形態的な観点から見ると，「電子メール」や「Eメール」は2つの要素からなる複合名詞であるのに対し，「メール」は単純語である．この意味では，「メール」は基の複合語の第1要素が省略されてできてきた短縮形であると見なしうる(注2)．

これらの語彙はどのように使い分けられるのだろうか．言い換えると，互いに代用可能な，競合する語彙の選択はどのような要因によって決定されるのだろうか．そのメカニズムを定量的に検討するのが本稿の目的である．この現象にはおそらく複数の要因が複雑に絡み合っていることが予想される．したがって，本稿では，応答変数(response variable)として「電子メール」「Eメール」「メール」という3つのカテゴリからなる多項ロジスティック回帰分析(polytomous logistic regression analysis)を用い，説明要因(predictor)として対象語彙の使用者の性別と年齢，対象語彙が使用されるジャンルおよび使用年代，および対象語全体の頻度を想定することにする(注3)．

2.2　調査対象

対象語彙項目を調査するためのデータベースとして『朝日新聞』の1984年から2011年までの記事を対象とすることにした．書き手の性別と年齢に関する情報を得る必要があるため，データベース中の記事のうち，読者からの投書欄，特に「声」と「ひととき」に掲載された記事を調査対象に定めた(注4)．2011年11月に，朝日新聞データベース版「聞蔵」を利用して調査項目の検索を行ない，その結果をテキスト加工してデータベース化をおこなった．最終的に，「声」3,117件，「ひととき」2,075件，合計5,129件の使用例が統計分析の対象となった(注5)．下の表1は，ジャンルごとに，対象語彙の使用頻度を使用者の性別で分割した表である．

表1　ジャンルと性別と対象語彙の3元分割表

「ひととき」	性別			「声」	性別		
	女性	男性	合計		女性	男性	合計
電子メール	68	0	68	電子メール	95	63	158
Eメール	41	0	41	Eメール	88	68	156
メール	1,937	29	1,966	メール	1,877	926	2,803
合　計	2,046	29	2,075	合　計	2,060	1,057	3,117

2.3　分析の方法

ロジスティック回帰分析は，一つのカテゴリカル変数を目的変数とし，その目的変数をその他の変数で説明する形のモデルを使って分析する手法である(藤井 2010: 86)．Agresti(2007: 99)によれば，「成功」と「失敗」など2値の応答変数のモデル化を行う際にロジスティック回帰分析はもっとも一般的なモデルである．たとえば，2値応答Yと計量値の説明変数Xのモデルにおいて，Xがxの値のときの成功確率を$\pi(x)$とするとき，$\pi(x)$は二項分布に従う．ロジスティック回帰モデルはこの確率$\pi(x)$を(1)式のようにロジット変換を行い，そのロジットとxの間に線形的な関係が成立するモデルである．

$$\mathrm{logit}(\pi(x)) = \log(\pi(x)/(1-\pi(x))) = \alpha + \beta x \qquad (1)$$

(1)の式から$\pi(x)$はxのS字型関数として，$\beta > 0$のときには増加を，$\beta < 0$のときには減少を示す．(1)のロジット関数がS字型「成長曲線」を表すことから，経済データの解析，たとえば携帯電話やインターネットの普及率の解析に用いるのに適している（金 2004）．普及率が大きくなり，飽和状態に近づくとその伸び率は小さくなり，普及率が100%を超えることはない．言語の変化もまた，いわゆるS字型の成長曲線を描くように推移することがよく知られている．Aitchison(1991)はその推移を“slow, quick, quick, slow”と形容している．すなわち，言語変化はその初期段階では緩やかに始まり，次の段階ではその変化が当該言語内で急速に普及，浸透していく．したがって，ロジスティック回帰分析は2値の応答確率を扱うのに望ましい性質をもっているので，言語変化の有効な解析法となることが期待されている(注6)．

ロジスティック回帰分析は元来医療研究の分野において2値応答に対する分析法として開発された（丹後，他 1996: 1）．現在では，応答が3つ以上のカテゴリである場合も扱えるように，多項ロジスティック回帰分析という拡張された手法が開発されている．本研究では「電子メール」「Eメール」「メール」という3語を応答カテゴリとして扱うので，この分析法を用いることにする．前述したように，「メール」は他の2語に対する短縮形と見なすことも可能であるから，応答カテゴリの間に順序性を認めることもできる．しかし，「電子メール」と「Eメール」の間には短縮の関係を想定できないため，順序性は部分的に過ぎない．したがって，本稿では応答カテゴリ間の順序性を利用したモデル，たとえば比例オッズモデルなどの分析手法を用いることはしない．

順序関係を想定しない多項ロジスティック回帰分析を，解析環境R（バージョンはR 2.13.1）の下で，主にVGAMパッケージ（バージョン0.7-7）で提供されているvglm関数を用いて行った(注7)．多項ロジスティック回帰分析において，任意に選択した一つの応答カテゴリをベースラインカテゴリとして，そのカテゴリに対する他のカテゴ

リのオッズ比の対数をロジットとする．最後のカテゴリ (J) をベースラインカテゴリとした場合，J-1 個のロジット式が構成される (Agresti 2002: 174).

$$\log(\pi_j / \pi_J) = \alpha_j + \beta_{jxj}, \qquad j=1, \cdots, J\text{-}1 \qquad (2)$$

本稿では，VGAM パッケージの仕様に従い，ベースラインカテゴリとして「メール」を選択している．

3．結果と考察

前節で述べたように，『朝日新聞』データベースを検索して，投書欄における使用例 5,129 件のデータを対象としてロジスティック回帰分析を行った．この節では，分析の結果を提示し，その結果の意味するところを検討していくことにする．3.1 節では説明変数ごとに単変量解析を行い，それらが説明的な要因であるかどうかを検討する．3.2 節では，多変量解析を行い，当てはまりの良いモデルの構築を試み，そのモデルからの予測を試みる．

3.1　単変量解析

ここでは，「電子メール」とその同義語の語彙選択に関わる説明要因の候補として，対象語が使用される文脈のタイプ（ジャンル），対象語を使用する人（書き手）の性別や年齢，対象語が使用された年代，および使用年における対象語全体の総頻度，という５つの変数を個別に取り上げる．そして，浜田 (1999) に従い，ロジスティック回帰モデルに当該変数のみを入れて，それが説明的であるかどうかを判定することにする．

はじめに，対象語が使用される記事のタイプ，具体的には「ひととき」欄と「声」欄の違いが対象語の選択に影響しているかどうかを調べることにする．表２は，２つの投書欄における対象語彙の使用頻度と相対頻度（列パーセント）を示している．

表２　投書欄ごとの対象語彙の頻度表

	「ひととき」	「声」	合計
電子メール	68 (3.3)	158 (5.1)	226 (4.4)
Ｅメール	41 (2.0)	156 (5.0)	197 (3.8)
メール	1,966 (94.7)	2,803 (89.9)	4,769 (93.0)
合　計	2,075 (100.0)	3,117 (100.0)	5,192 (100.0)

χ^2 乗検定により分布の等質性を検定すると，χ^2=42.3 となり，有意水準 0.1% で有意である．残差を検討すると，「声」欄において「電子メール」や「Ｅメール」の使用

率が高く，「ひととき」欄において「メール」の使用率が高いことがわかる．

VGAM パッケージの vglm 関数において，対象語彙を応答変数，投書欄の別を説明変数として多項ロジットモデルを構築した．説明変数に対する回帰係数の値は，0.488 と 0.981 であり，ともに有意であった．これらの値はいずれも正であることから，「メール」に対する「電子メール」および「Ｅメール」のオッズ比は，「ひととき」欄よりも「声」欄において有意に高いことを示している．また，説明変数をひとつも持たない，切片のみのモデルを構築し，このモデルとの間で尤度比検定を行ったところ，χ^2=45.6，自由度 2 で，有意であった（正確な p 値 <0.001）．このことは，「ひととき」欄と「声」欄の違い，すなわち，文脈タイプ（ジャンル）は説明的な要因であることを示している．

次に，書き手の性別が説明的かどうかを検討する．表３は全データを性別と対象語彙で分割した頻度表である．

表３　性別による対象語彙の頻度

	女性	男性	合計
電子メール	163 (4.0)	63 (5.8)	226 (4.4)
Ｅメール	129 (3.1)	68 (6.3)	197 (3.8)
メール	3,814 (92.9)	955 (87.9)	4,769 (93.0)
合　計	4,106 (100.0)	1,086 (100.0)	5,192 (100.0)

表３の数値に基づいて「電子メール」と「メール」の性別によるオッズ比を計算すると 1.54，同様に，「Ｅメール」と「メール」の性別によるオッズ比は 2.11 となる．すなわち，「メール」に対して「電子メール」を使用するオッズは男性が女性に比べて 1.54 倍高く，同様に「メール」に対して「Ｅメール」を使用するオッズは男性が女性に比べて 2.11 倍高いと解釈される．性別を説明変数としてロジスティック回帰モデルに投入すると，性別の回帰係数はそれぞれ 0.434 と 0.744 であり，通常のオッズ比に指数換算すると，それぞれ 1.54 と 2.11 となる．これらの係数に対するワルド統計量を算出し，対応する p 値を求めると，いずれも有意であった．このことから，書き手の性別もまた説明要因として働くと考えられる．

次に，書き手の年齢が対象語彙の選択に影響を及ぼすかどうかを検討してみる．新聞に掲載された投稿者の年齢は 6 歳から 92 歳にわたり，平均は 49.1 歳であった (注8)．下の図１は，投稿者の年齢層を４歳から 10 歳刻みで９つのカテゴリに分け，対象語彙の使用比率を図示したものである．

図１　対象語彙の年齢層別使用比率

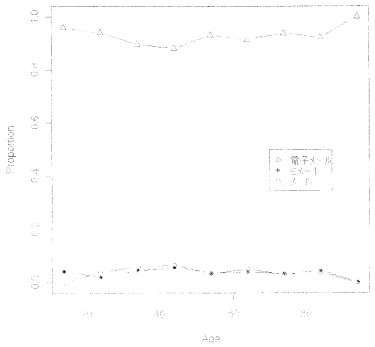


図２　対象語彙の年代別使用比率

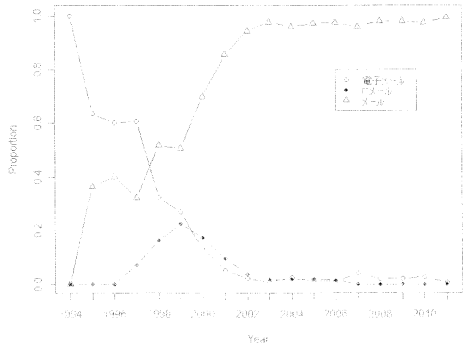


図１から，どの語の使用比率も年齢に応じて大きな変化を示さないことが見て取れる．実際，書き手の年齢を説明変数として回帰モデルを構築したところ，年齢に対する回帰係数の値は-0.0067, 0.0020 と 0 に近い値となった．回帰係数のワルド統計量に対応するｐ値はいずれも有意でなかった．尤度比検定を用いて，すべての回帰係数の値が 0 であるという帰無仮説を検定したところ，χ^2=3.45，自由度２で，ここでも有意とならなかった（正確なｐ値は 0.178）．このことから，書き手の年齢は説明要因として働いていないと考えられる．

次に，対象語彙の使用年代が語彙選択の説明要因となるかどうかを検討してみる．『朝日新聞』データベースは 1984 年から検索可能であるが，対象語彙が一般読者からの投書欄「ひととき」や「声」に登場するのは 1994 年からである．(3) や (4) はその使用例である．

(3) a. 例えば，売りたい人が掲示板に書き込み，買いたい人がその人に電子メールで連絡すれば，裏ビデオや下着などが，未成年でも簡単に購入できてしまう環境なのです．(1994/7/8，「声」，男性，43 歳)
　　b. コンピューターをつなぎ，文字をやりとりする電子メールや音や絵も出るマルチメディア通信も出来るインターネットは世界で三千万人に利用されている，という．(1994/8/6，「声」，男性，65 歳)
(4) a. アメリカに出張中の夫と電子メールを二，三通送受信しました．国際電話もダイヤル直通でかかるのですが，インターネットで電子メール交換をやってみたかったんです．(1995/12/7，「ひととき」，女性，39 歳)
　　b. 毎朝私はパソコンの電子メールボックスを開く．「あなたあてのメールが届いています」というメッセージを見るのは，郵便箱をのぞくのとは違った楽しみがある．(1996/7/29，「ひととき」，女性，62 歳)

図２は「電子メール」およびその同義語の年代別の使用比率をグラフに図示したものである．図から明らかなように，年代によって使用比率が大きく変化している．90 年代中盤までは「電子メール」が優勢であったが，90 年代後半からは「メール」の方がより多く用いられるようになった．「Ｅメール」は 90 年代後半までは増加傾向にあったがその後減少に転じた．2000 年以降は「メール」が圧倒的優勢となり，他の２語の比率は極めて小さくなってきている．

上で述べたように，『朝日新聞』投書欄における対象語彙の初出例は 1994 年にあるそこで 1994 年を起点とし，そこからの経過年数を説明変数としてロジットモデルに組み込むことにした．その結果，年代に対する回帰係数はそれぞれ-0.411 と -0.464 で，負の値が得られた．これは，「メール」に対して「電子メール」や「Ｅメール」が選択される比が年代が進行するにつれて減少することを意味している．すなわち，対数オッズ比を指数変換すると，１年単位で 0.66 倍，0.63 倍に減少し，10 年では 100 分の１程度に減少するという意味である．したがって，対象語彙の選択に対して，年代は極めて強い影響力を及ぼしていると考えらえる．

最後に，対象語彙の使用頻度が説明的かどうかを検討することにしよう．Bybee and Thompson(1997)，Bybee and Scheibman(1999)，Bybee(2003) や Krug(1998) などの文法化現象の研究の中で，高頻度で（共起）出現する要素ほど縮約効果を受けることが指摘されている．頻度が縮約の要因であるとすれば，同義語群の全体としての使用頻度が多くなるほど，より単純な形式を持つ短縮語形が好まれると予想される．そこで対象語彙全体の総頻度を使用年ごとに集計した．下の表４は，対象語彙の頻度を，年代順ではなく，集計した総頻度の順に並べ替えたものである．

表４　総頻度の昇順と対象語彙の頻度表

	4回 (1994)	11回 (1995)	15回 (1996)	28回 (1997)	31回 (1998)	93回 (1999)	292回 (2008)	297回 (2009)	304回 (2011)
電子メール	4	7	9	17	10	25	6	6	2
Ｅメール	0	0	0	2	5	21	0	0	0
メール	0	4	6	9	16	47	286	291	302
	336回 (2010)	340回 (2006)	342回 (2005)	368回 (2007)	413回 (2000)	455回 (2001)	518回 (2004)	634回 (2002)	711回 (2003)
電子メール	9	4	4	15	54	23	12	13	6
Ｅメール	0	5	6	0	71	43	9	23	12
メール	327	331	332	353	288	389	497	598	693

表４のデータに基づいて，一般化マンテル検定 (extended Mantel test) を行った[注9]．この検定は，I×J 表の一方の変数が順序カテゴリである場合，その順序性を考慮したスコアを与えることによって統計量を算出し，それが自由度 I－１のカイ２乗分布で近似できることを利用して，列変数と行変数の独立性を検定するものである．検定の結果，χ^2=114.07 自由度２で，高度に有意であった ($p<0.0001$)．すなわち，対象語彙の総頻度は語彙の選択に関連があることが示された．

これまでと同様に，対象語彙の年代別総頻度を説明変数としてロジスティック回帰分析を行ったところ，説明変数に対する回帰係数として-0.0060 と-0.0013 が得られ，いずれも有意であった．これらの回帰係数は，対象語彙の年代総頻度が１増加するごとに，「メール」に対する「電子メール」と「Ｅメール」のオッズ比がそれぞれ 0.994 倍，0.998 倍に減少することを示している．「メール」に対する「電子メール」のオッズ比は，総頻度が 100，200，400 増加すると，それぞれ，約半分，約３分の１，約 10 分の１に減少すると予測する．図３は，年代頻度だけを説明変数とするロジスティック回帰モデルによって予測される確率曲線に，実際に観測された比率をプロットしたものである[注10]．

図３　対象語彙の予測確率と観測比率

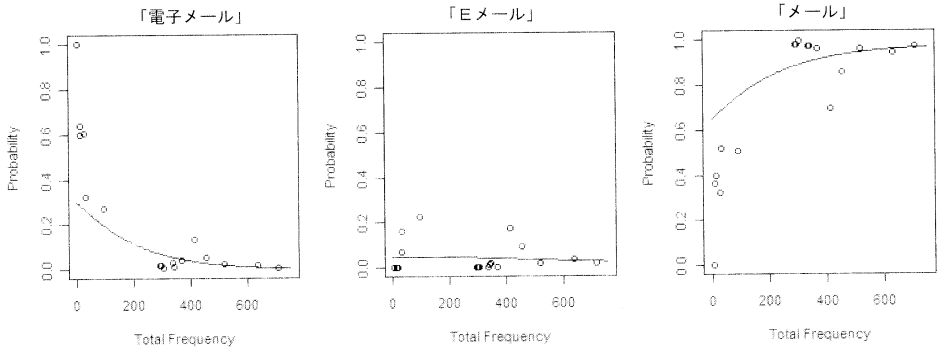


図３を見ると，頻度 200 以下の場合に実測比率と予測確率の乖離が大きいことがわかる．これはまさに低頻度であるために，重みづけが足りないからである．年代別の総頻度を見ると，1994 年から 2003 年までは急激に増加したものの，その後減少に転じ，最近数年は横ばいが続いている．ピークの 2003 年を境にして，2000 年と 2007 年は総頻度ではそれほど変わらないけれども，対象語彙の比率は大きく異なっている．一度上限値近くまで到達すると，その比率が固定化される「高止まり」効果とでも呼ぶべき現象が観察される．このことは，対象語彙の選択を説明するには総頻度だけでは不十分であることを示していると考えられる．

ここでは説明変数の候補の単変量解析をおこなった．その結果，対象語彙が使用される文脈のタイプ，書き手の性別，使用される年代，年代別の総頻度などの変数が，対象語彙の選択に影響していることが示された．これに対して，書き手の年齢は説明的な要因として働いていないことが示された．

3.2　多変量解析

この節では，対象語彙の選択を説明するための多変量解析を行う．すなわち，複数の説明変数の候補を同時にモデル化して多重ロジットモデルを構築する．ロジスティック回帰分析では交絡因子の候補をもモデルに取り込んで，その変数の影響を調整することができる．複数の変数を同時に調整して解析を行い，調整した後も有意に関連する変数は，当該の応答カテゴリの選択に独立に関与する説明要因であると見なすことができる（浜田 2004）．

はじめに，3.1 節で検討した５変数すべてを投入したモデルを構築した．この主効果モデルと切片のみのモデルとの尤度比検定の結果は高度に有意であるので，主効果モデルの回帰係数は全体として説明力があると言える．表６は，５つの変数に関して，3.1 節で行った単変量解析の結果得られた回帰係数の値（粗対数オッズ比）と，主効果モデルの回帰係数の値（調整済み対数オッズ比），ワルド統計量および調整済みオッズ比の 95% 信頼区間を示している．

表６　粗オッズ比と調整済みオッズ比の比較表

	β推定量（単変量解析）	β推定量（主効果モデル）	ワルド統計量	P値	95%信頼区間（下限，上限）
文脈タイプ1	0.488	0.663	3.60	<0.001	(1.35, 2.78)
文脈タイプ2	0.982	1.277	6.32	<0.001	(2.41, 5.33)
性別1	0.434	0.593	1.60	0.12	(0.93, 1.93)
性別2	0.744	0.926	1.90	0.06	(0.99, 2.03)
年齢1	-0.007	-0.005	- 0.04	0.97	(0.99, 1.01)
年齢2	0.002	0.009	3.45	<0.001	(1.01, 1.03)
年代1	-0.411	-0.290	-12.94	<0.001	(0.71, 0.77)
年代2	-0.464	-0.505	-12.57	<0.001	(0.54, 0.64)
総頻度1	-0.006	-0.005	-10.52	<0.001	(0.995, 0.996)
総頻度2	-0.0013	-0.0004	- 0.45	0.65	(0.999, 1.001)

表６の粗（対数）オッズ比と調整済み（対数）オッズ比を比較すると，調整の前後で数値の符号が変わるほどの変化，すなわち，増加と減少に関して効果が逆転するほど

大きな変化を示すものはない．このことから５変数の中に交絡因子は含まれていないと判断できる．

表６には，主効果モデルに投入された変数の回帰係数に対するワルド統計量およびそのｐ値が示されている．他の変数の影響を調整した後もこのｐ値が小さい変数は有意性が高く，対象語彙の選択と強く関連している要因と考えらえる．逆にｐ値の大きい変数は当該の現象にあまり貢献をしていないと判断して，モデルから順次取り除いていくことにする．はじめに，主効果モデルから性別変数を除去することができる．次に，残る４変数を組み込んだモデルを構築すると，年齢変数の有意性の低いので，それを除外した．３変数からなるモデルでは，どの変数も有意であるため，これ以上変数を減少することはできないと判断した．表７は，最終的なモデルにおける各変数のパラメータの値を示している．

表７　最終の多項ロジットモデルのパラメータ

	β推定量	ワルド統計量	P値	オッズ比	95%信頼区間（下限，上限）
切片1	1.139	4.97	<0.001	3.12	(1.99, 4.89)
切片2	0.521	1.77	0.077	1.68	(0.95, 3.00)
文脈タイプ1	0.754	4.57	<0.001	2.13	(1.54, 2.94)
文脈タイプ2	1.305	7.06	<0.001	3.69	(2.57, 5.30)
年代1	-0.300	-13.04	<0.001	0.74	(0.71, 0.77)
年代2	-0.509	-12.38	<0.001	0.60	(0.55, 0.65)
総頻度1	-0.0046	-10.61	<0.001	0.995	(0.995, 0.996)
総頻度2	-0.0004	- 0.83	0.408	0.999	(0.999, 1.001)

このモデル中のジャンル変数は「ひととき」か「声」か，どちらかの値しかとらない．モデル中の説明変数がすべて２値カテゴリであれば，ワルド検定のｐ値が小さい要因ほど，強い影響力を与えていると見なすことができる（浜田 2004）．しかし，この最終モデルの年代変数や総頻度変数は連続変数であり，さらにスケールも違うため，ｐ値を基準にして有意性を単純に比較することはできない．

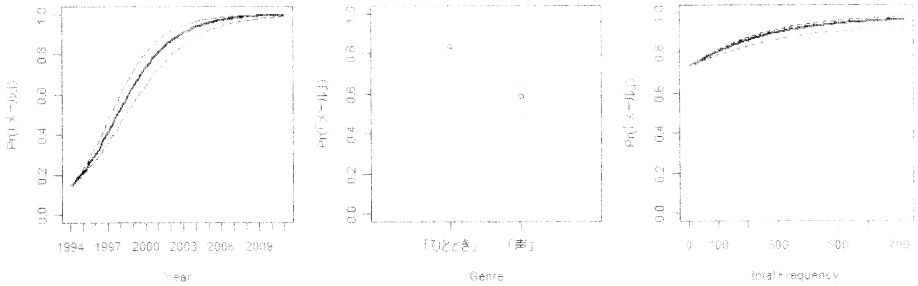
最終のロジットモデル式は (5) のように表される．

$$\eta 1 = 1.139 + 0.755*\text{Genre} - 0.300*\text{Year} - 0.005*\text{TotalFreq} \quad (5a)$$

$$\eta 2 = 0.521 + 1.305*\text{Genre} - 0.509*\text{Year} - 0.0004*\text{TotalFreq} \quad (5b)$$

下の図４は，応答として「メール」が選択される際に，最終モデルに含まれる説明変数の部分効果の大きさを 95% 信頼区間とともに示したものである．部分効果とは，Baayen(2008: 175) によると，モデル中の他の説明変数の値としてその変数のデフォルト値あるいは平均値を用いて一定にしたときに，当該の説明変数の値によって決定される応答確率である[注11]．

図４　「メール」の応答確率に対する各変数の部分効果



ロジスティック回帰分析は，他の回帰分析と同様に，構築されたモデルに基づいた予測が可能である点が長所の一つである．最後に，ロジット関数 (5) を使って，10 年後における対象語彙の応答確率を予測してみよう．対象語彙の総頻度が今後 10 年間 445 で変わらないと仮定すると，「声」欄において「電子メール」「E メール」「メール」が選択される確率はそれぞれ 0.02%，0.006%，99.98% と予測される．

4．まとめ

本稿では，「電子メール」とその同義語を対象語彙として取り上げ，新聞データベースを対象にして，変異形の使用頻度の推移を調査した．５千件を超える使用例データを対象に多項ロジスティック回帰分析の手法を適用することによって，競合する語彙の選択をする際に働いている要因の特定を試みた．５つの説明要因の候補となる変数に対して単変量および多変量解析を行った．その結果として，「電子メール」とその同義語の使用選択に関して，書き手の性別と年齢は有効な説明要因とならないけれども，対象語彙の使用年代，対象語彙が使用されるジャンルおよび対象語彙全体の頻度が当該現象と強く関連していることを実証した．この分析結果は，非言語的な要因よりも，文脈や頻度などの言語的要因が当該現象にとって本質的であることを示している．近年，Bybee や Krug らの文法化研究の中で頻度要因が言語変化において重要な役割を果たすことが指摘されてきている．「電子メール」と「メール」が同義語であると同時に短縮の関係でもあることを考慮すると，頻度効果が語彙選択の領域においても重要な要因として働くことが確認されたことには意義がある．

本稿で提案したモデルが観測データによく当てはまり，当該現象の今後の推移予測を可能にする限りにおいて，本研究は言語変化のメカニズム研究においてこの種のアプローチ方法の妥当性を示したことになるだろう．ただし，本稿ではいくつかの課題が未検討のまま残されている．その一つは，対象語彙全体の年代総頻度のような広域的に作用する頻度要因ばかりでなく，同一テキスト内における対象語彙の繰り返しによって生じる，いわゆる局所的な頻度要因の効果を測定することである．この課題については稿を改めて考察してみたい．

注

本研究の一部は，田嶋毓堂語彙研究会からの平成22年度研究補助金の支援を受けている．ここに感謝の意を表する次第である．

[1] 2008年11月発行の最新版の第6版によると，「電子メール」とは「コンピュータ同士が，ネットワークを通して文書・画像などの情報を伝達・蓄積する通信システム．また，そのシステムにより交換されるメッセージ．インターネットの利用形態の一つ．」と定義されている．「メール」の項には「電子メール」の略語である旨が記載されている．「イー・メール」の項にも同様の記載がある．

[2] 英語においても，短縮化の現象が観察される．「電子メール」はelectronic mailであるが，その後e-mailのように第1要素を簡略化した複合語が使われ始め，最近ではハイフンがつかない異形態も一般的になってきている．

[3] Fowler and Housum(1987)は，当該語彙が先行文脈に生じていて既知の情報となっていると，話し手によって語形が短縮される傾向があることを実証している．書きことばにおいても同様の傾向が観察されることが予想される．すなわち，同一文脈内で対象語が繰り返し使われるときに，出現頻度が多くなるほど短縮形が使用される傾向があると予想される．これについては，本稿ではこれ以上議論せず，次の機会に詳しく分析することにする．

[4] 書き手の性別については，投稿者の名前と職業から推定した．「薫」「千尋」など，判別不可能なケース（54件）はその後の分析から除外した．

[5] 今回，「メールアドレス」「メールマガジン」「メール友だち」「写真メール」は分析対象に含まれているが，これらに対応する「メルアド」「メル友」「メルマガ」「写メ」などの例を除いている．これらの短縮関係にある同義語の使用選択については，別の機会に検討することにしたい．

[6] 真田(2008)では，言語変化のメカニズムを説明するためにロジスティック回帰分析を利用した研究事例が紹介されている．

[7] VGAMパッケージはThomas W. Yee氏によって開発，提供されている．同氏のURL (http://www.stat.auckland.ac.nz/~yee)には，VGAMパッケージに関する詳細な情報を記した文書が公開されている．ここに記して感謝の意を表したい．

[8] これらの数値はもちろん，対象語彙が使用される時点での書き手の年齢である．同じ年齢の書き手であっても，時代によって社会的環境が異なるのであるから，その影響の受け方をどのように数値化すべきかという課題がある（真田 2008: 9）．

[9] 藤井(2010: 64)に，一般化マンテル検定を行うためのRスクリプトが掲載されている．ここに記して感謝したい．本文の結果は，スコアとしてc(0, 1, 2, 3, 3, 4, 5, 5, 5, 6, 6, 6, 6, 7, 8, 9, 10, 11)を与え，RスクリプトextendedM.test.Rを実行して得られた．

[10] 各応答カテゴリの予測確率は，ロジット関数の逆関数であるロジスティック関数から推定される．多項ロジットモデルの場合のロジスティック関数は次のように定式化される．

$$\pi_j = e^{(\alpha_j + \beta_j x)} / \sum_j e^{(\alpha_j + \beta_j x)}, \quad j = 1, \ldots, J \quad \text{(Agresti 2007: 176)}$$

[11] 部分効果を算出するために，ここでは，ジャンル(Genre=1)変数として「書き」(Genre=1)，年代として2003年(Year=10)，総頻度として平均値445(Total Freq=445)の値を(5)に代入している．

参考文献

Agresti, Alan (2007) *An Introduction to Categorical Data Analysis*, 2nd edition. Hoboken, NJ: John Wiley & Sons.

Aitchison, Jean. 1991. *Language Change: Progress or Decay?* 2nd edition. Cambridge: Cambridge University Press.

Baayen, R. H. (2008) *Analyzing Linguistic Data: An Introduction to Statistics Using R*. Cambridge: Cambridge University Press.

Bybee, Joan, and Sandra Thompson (1997) 'Three frequency effects in syntax.' *Berkeley Linguistics Society* 23, 65-85.

Bybee, Joan, and Joanne Scheibman (1999) 'The effect of usage on degrees of constituency: The reduction of don't in American English.' *Linguistics* 37: 575-596.

Bybee, Joan (2003) 'Mechanisms of change in grammaticization: The role of frequency.' In *Handbook of Historical Linguistics*, ed. by R. Janda and B. Joseph, 602-623. Oxford: Blackwell.

藤井　良宣 (2010)『Rで学ぶデータサイエンス1　カテゴリカルデータ解析』．東京：共立出版．

浜田　知久馬 (1999)『学会・論文発表のための統計学－統計パッケージを誤用しないために』．東京：真興交易医書出版部．

金　明哲 (2004)「フリーソフトによるデータ解析・マイニング第15回　Rと一般化線形モデル」．*Estrela* 83, No. 127 (2004年10月)．

Krug, Manfred (1998) 'String frequency: A cognitive motivating factor in coalescence, language processing and linguistic change.' *Journal of English Linguistics* 26, 286-320.

R Development Core Team (2011) *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing*, Vienna, Austria.

真田　治子 (2008)「言語変化のＳ字カーブ―解析手法の比較とその適用事例―」,『埼玉学園大学紀要（人間科学部)』8, 1-11.

丹後　俊郎, 山岡　和枝, 高木　晴良 (1996)『ロジスティック回帰分析－ SAS を利用した統計解析の実際－』, 東京：朝倉出版.