計量国語学第二十六巻第三号〔Mathematical Linguistics, Vol.26 No.3〕2007 年

論文

多変量 S 字カーブによる言語変化の解析

―仮想方言データのシミュレーション―

横山　詔一（国立国語研究所）・真田　治子（埼玉学園大学）[1]

デイスクリプタ：言語変化，S 字カーブ，ロジスティック回帰モデル，
グロットグラム，最尤推定法

１．はじめに

言語変化は一般に「遅→速→速→遅」という S 字カーブの過程をたどって伝播するとされる(Aitchison,1991)。その例として，カナダ Golden Horseshoe 地域における英語の動詞「sneak」の過去形について Chambers(2006)が報告した「sneaked→snuck」の語形交替に関するデータを図１(１)に示す。【末尾注１】横軸は年齢(80 歳代，70 歳代，60 歳代など)，縦軸は「snuck」を使う人の百分率である。図１(１)のグラフは，年齢が若くなるにつれて「snuck」を使う人の割合が高くなることを示している。

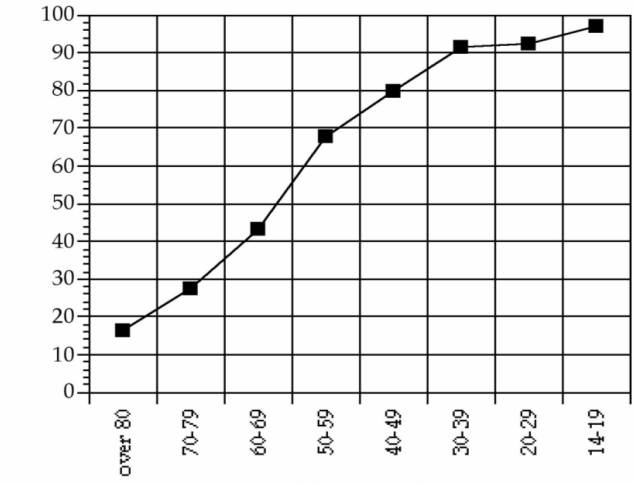

図１(１) カナダ英語の語形交替の例
(Chambers,2006 から引用)

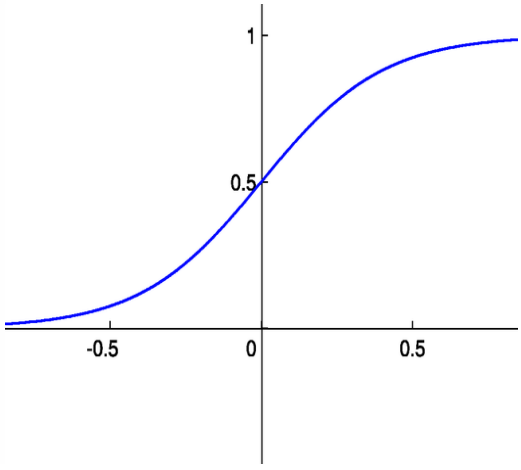

図１(２) ロジスティック曲線の例
(縦軸上限値 1.0，下限値 0.0)

[1] YOKOYAMA Shoichi*, SANADA Haruko** (*The National Institute for Japanese Language, **Saitama Gakuen University)― Multiple Logistic Regression Analysis for Formulating a Change in Language―

人口の増加過程，ミジンコの個体数の増殖過程，新商品の普及過程などもS字カーブを描き，ロジスティック関数(logistic function)がよく当てはまることが知られている。ロジスティック曲線（シグモイド曲線）の例を図1(2)に示す。図1の(1)と(2)は似た形状であることから，年齢を「見かけ上の時間(apparent time)」だと仮定することにより，言語変化のS字カーブもロジスティック関数で近似できると考えられる。

井上(2000)は，言語変化を40年以上の期間にわたって「実時間(real time)」でとらえた先駆的研究を行っている。井上は国立国語研究所による山形県鶴岡市における3回にわたる共通語化の調査データ，およびその近郊櫛引町山添付近での井上自身による2回の調査データを綿密に分析し，共通語化の過程がS字カーブになることを実証した。言語変化のS字カーブについて，これほど大規模な研究は世界を見わたしても例がない。

以上のように，言語変化の記述研究は着実に厚みを増しているが，言語変化のS字カーブを解析する手法に関する研究はきわめて少ない。そこで，本研究は「多重ロジスティック回帰モデル（Multiple logistic regression model)」による言語変化の多変量解析法を提唱する。以下では，グロットグラムの説明からスタートして，S字カーブに関する理論の概観を行い，単回帰分析の実例をふまえて多変量解析の実例へと論を進める。

(1)グロットグラムについて

言語はふつう時間の経過とともに変化すると考えられるが，変化の痕跡を地域における伝播の結果から読みとることもできる。例えば，表1のように縦軸に年齢層，横軸に地点をとる。AからDまでの地点は鉄道によって線上に並んでいるものとする。地点は空間の変数であり，このような表を「グロットグラム(glottogram)」という。これは地域差と時間差を同時に観察できるため，言語変化の効率的な分析が可能である（鑓水，2006)。

表1に，各地点で各年齢層の回答者が1名ずつのデータを示す。これは仮想データである。ある語形を使うと回答した場合は「○」を，それ以外は「△」を付している。

表1　仮想グロットグラム

年齢	A地点	B地点	C地点	D地点
85	△	△	△	△
75	△	△	△	○
65	△	△	○	○
55	△	○	○	○
45	○	○	○	○
35	○	○	○	○
25	○	○	○	○
15	○	○	○	○

この例では「○」と「△」の境界線が単調な右肩上がりになっており，D 地点で生じた変化が C 地点→B 地点→A 地点という順番で伝播したと推測できる。グロットグラムは言語変化に関するデータであることから，次に述べるような手法で S 字カーブを当てはめることができる。

（２）S 字カーブに関する理論の概観

ヨーロッパやロシア周辺諸国では，言語変化の S 字カーブについて Altmann, Buttlar, Rott & Strauss(1983)の近似カーブがよく知られている。その式は，p を「比率」，t を「時間」，K と A を定数として

$$p=1/\{1+A\cdot\exp(-K\cdot t)\} \qquad 〔1〕$$

これを変形して

$$\log\{(1/p)-1\}=\log A-K\cdot t \qquad 〔2〕$$

式〔1〕や式〔2〕が描くカーブはロジスティック関数そのものであり，人口学や生物学などの分野では「密度効果(density effect)」のモデル式として教科書にもしばしば登場する。密度効果とは，単位面積あたりの個体数が環境の限界をこえて増殖すると増加へのマイナスの影響が生じることを指す。密度が増加するとエサが不足したり病気が蔓延したりして，増殖が止まったり，減少が始まったりする。密度効果の一番単純な式として，密度効果が完全に密度に比例する場合を微分方程式に仕立て，それを解くと式〔1〕が得られる。この式は 1845 年にベルギーのフェルフュルストが発見し，1919 年に米国の Perl が人口を調査していて同じ式に到達した（伊藤，1994）。Altmann ら(1983)の微分方程式も基本的には密度効果と同じアイデアに立脚しているが，言語変化の S 字カーブについて数理モデルを明示したものはこれ以外に見当たらないし，言語研究の分野でこれに肩を並べる新手法の提言はなされておらず，言語変化の時系列分析をヨーロッパ各国の研究者がこの方法で行ってきたという実績を有する。Altmann 学派は，式〔2〕に最小２乗法を適用して A と K を求める。日本においても真田(2002)が辞書における学術漢語の消長を，橋本(2006)が新聞における外来語の増加過程をこれに類する方法で検討している。

しかし，Altmann 学派の方法は次の３つの点でさらに発展させるべき余地がある。

１つ目は，p が 0.0 あるいは 1.0 のとき計算不能に陥ってしまうことである。例えば，先に示した図１(１)で 20 歳代と 30 歳代の snuck 使用率が 100%だったと仮定してみよう。その場合，Altmann 学派の方法では式〔2〕の左辺が計算できなくなるため，20 歳代と 30 歳代のデータを捨てざるを得ないのだが，計算不能だからといって p が 0.0 や 1.0 のデータを無視するのは分析手法の都合にあわせた恣意的な判断に過ぎない。言語変化のデータを時系列で解析するとき，これは大きな壁になる。

２つ目は，変数が「時間」だけに制限されていることである。グロットグラムのように「年齢（時間)」と「空間(距離)」を同時にまとめて解析するとき，Altmann 学派の方法はそのままでは使えない。多変量の S 字カーブを扱える手法が必要である。

３つ目は，表１のような名義尺度のカテゴリカルなデータを扱えないことである。Altmann 学派の方法は目的変数が比率だけに限定されているため，目的変数が「○か△」といった２値データのものにはまったく対応できない。

以上の問題をすべて解決できるのが「最尤推定法(maximum likelihood estimation)」を用いたロジスティック回帰モデルである。本研究はそれを言語データの時系列解析に導入する。ロジスティック回帰モデルは，p を「比率（確率)」，Z を $a1\cdot X1+a2\cdot X2+a3\cdot X3+\cdots+b$ というような重回帰式に類する線型結合，対数の底を e として，式〔3〕で示される。

$$\log\{p/(1-p)\}=Z \qquad 〔3〕$$

以下，$p/(1-p)$ をオッズ(odds)，$\log\{p/(1-p)\}$ をロジット(logit)という。ロジスティック回帰モデルは，左辺がロジット，右辺が重回帰式の形をなす。式〔3〕を変形すると

$$p=1/\{1+\exp(-Z)\} \qquad 〔4〕$$

ここで，$X1$ を「時間」，$a1$ と b を定数，$Z=a1\cdot X1+b$ とおくと式〔3〕と式〔4〕は

$$\log\{p/(1-p)\}=a1\cdot X1+b \qquad 〔5〕$$

$$p=1/\{1+\exp[-(a1\cdot X1+b)]\} \qquad 〔6〕$$

このロジスティック回帰モデルは世界中の医学研究できわめて盛んに利用され，その有用性についての評価は不動の地位を築いている。Altmann ら(1983)の式〔2〕もロジスティック回帰モデルの式〔3〕に帰着することが横山・真田(2007)によって証明されている。すなわち，式〔2〕はさらに変形すると式〔2.a〕になる。

$$\log\{p/(1-p)\}=K\cdot t-\log A \qquad 〔2.a〕$$

社会言語学の分野でも 1960 年代に米国の Labov(1972)が言語データの解析にこの手法を積極的に導入したという歴史がある。日本の言語研究にロジスティック回帰モデルを導入する試みは，Labov の薫陶を受けた Hibiya(1988)と Matsuda(1993)を嚆矢として，横山(2006)や Yokoyama & Wada(2006)などのほか，南部(2007）の詳細な論考がある。

しかし，いずれの研究においても言語変化の多変量 S 字カーブをロジスティック回帰モデルで解析するという発想はまったくなかった。言語変化の S 字カーブに関する研究をさらに飛躍させ，とかく記述研究にとどまりがちな現状から脱皮するには，確率論の尤度に着目したロジスティック回帰モデルを新たな分析法として手中におさめる必要があろう。言語変化の動向を数量的に予測する研究に一歩踏み出すために「確率予測の道具立て」を積極的に取り入れる努力が，記述研究や探索研究にくわえて仮説検証研究のさらなる活性化に寄与するものと思われる。以下，ロジスティック回帰モデルにより言語変化を多変量 S 字カーブで解析する方法の実例を示す。

２．最小２乗法でも対応できる単回帰分析の例

先に図１(１)で示した Chambers(2006)に近い数値の仮想データを作成し，ロジスティ

ック回帰モデルで分析した。分析に用いたデータは【末尾注２】付表１に示す距離「0」の snuck 使用率であった。説明変数は年齢(t または $X1$)で単回帰分析である。年齢は 80 歳代以上を「85」，70 歳代を「75」，というように各年代の中央の値とした。

推定された近似式は，最小２乗法で $\log\{p/(1-p)\}=-0.075\cdot t+4.769$，最尤推定法で $\log\{p/(1-p)\}=-0.075\cdot X1+4.747$ となった。両者で推定値に大きな開きはない。最尤推定法による p の予測値は $p=1/\{1+\exp(0.075\cdot X1-4.747)\}$ で求められる。得られた p の予測値を図２に示す（以後のグラフでは p を 100 倍したパーセントで表示)。

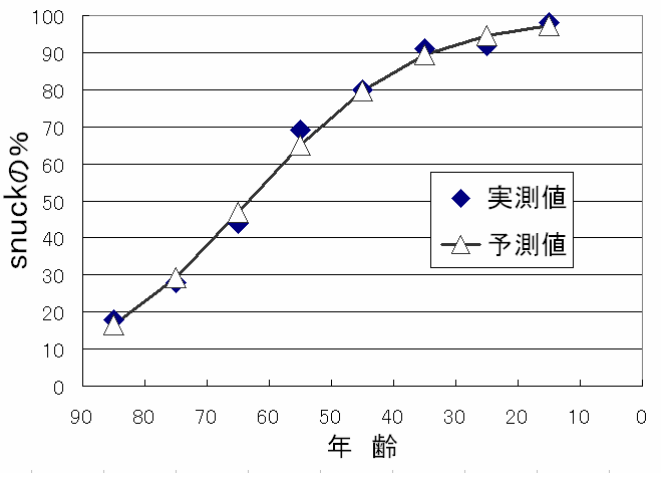

図２ 最小２乗法でも対応できる例（100%のデータを含まない）

pの予測値と実測値の相関係数を２乗したもの，すなわち決定係数（説明率）は$r^2=0.994$であった。フィッティングの精度は極めて高い。

ちなみに，先に言及した真田(2002)による学術漢語の消長データや，橋本(2006)の新聞における外来語の増加過程データを横山・真田（2007）が最尤推定法で解析したところ，最小２乗法による推定値とほぼ同じ値が得られることが確認されている。

３．最尤推定法であれば対応できる単回帰分析の例

先に述べたように図１(１）の Chambers(2006)のデータで 20 歳代と 30 歳代の snuck 使用率が 100%だったと仮定してみよう。Altmann 学派の最小２乗法による方法では p が 0.0 や 1.0 になると式〔2〕の左辺が計算できなくなる。20 歳代と 30 歳代のデータを捨てざるを得ない。

統計学において解析手法の制約からデータを捨てることは避けなくてはならない。この問題を回避する便法として，p が 0.0 や 1.0 のデータを「経験ロジット(empirical logit)」という数値に変換する方法も提案されてはいるが，その変換式を使う理論的な根拠は希薄である(Collett,2003)。p が 0.0 や 1.0 のデータを扱えない解析手法は，それが依拠する理

論の限界を露呈していると言えよう。最尤推定法であれば何の問題もなくパラメータが推定できる。最尤推定法の理論は 20 世紀前半には確立されていたが，膨大な数値計算を必要とするため，コンピュータの普及を待って表舞台に登場するようになった。

最尤推定法で求めた近似式は $\log\{p/(1-p)\}=-0.092\cdot X1+5.961$ となった。p の予測値は $p=1/\{1+\exp(0.092\cdot X1-5.961)\}$ で求められる。p の予測値を図３に示す。

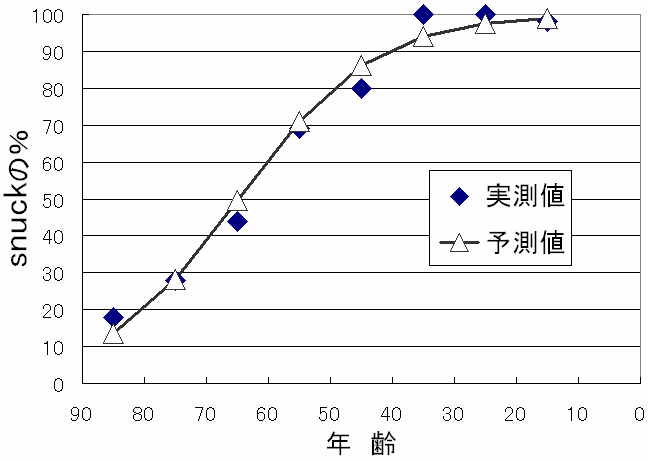

図３ 最尤推定法であれば対応できる例（100%のデータを含む）

pの予測値と実測値の相関係数を２乗したもの，すなわち決定係数（説明率）は$r^2=0.983$であった。フィッティングの精度は極めて高い。

４．多変量解析に展開した例

言語変化の S 字カーブに関するこれまでの研究は，変数が「時間」だけに限られていた。しかし，言語変化に影響する要因は時間軸だけではなく，多くの変数がからみあっていると容易に想像できる。その証左として，社会言語学などでは時間のほかに，先に述べた「空間(距離)」や「性差」なども言語変化を究明する際の重要な要因とされている。方言学は地域間の変異を追究する。単回帰分析の範囲にとどまるのではなく，重回帰分析のような多変量解析に発展させることにより，解析の射程を飛躍的に伸ばせると期待できるが，これまでの手法はそのような発想を持っていない。

ロジスティック回帰モデルは，右辺が重回帰式に類する線型結合であり，多変量解析を念頭においた手法である。説明変数は名義尺度でもよい。そのデータ解析における有用性を仮想データで検証してみよう。

図１(１)の Chambers(2006)のデータにおいて，60 歳代および 50 歳代と，他の年齢層とは回答において何らかの「場面差（心理状態の差)」があったと仮定する。例えば，調査員の熟練度や言語学的知識の多寡などによって，調査参加者が無意識のうちに普段とは違った特殊な心理状態で回答する場合があるとしよう。その影響で，60 歳代と 50 歳代の snuck

使用率は図１(１)の数値より 30%から 50%も高くなった状況を考えてみよう。そのような仮想データを表２に示す。年齢は 80 歳代以上を「85」，70 歳代を「75」，60 歳代を「65」，というように各年代の中央の値で数値化した。場面差は普段の心理状態を「0」，特別な心理状態を「1」というように示した。

表２　「snuck」データの多変量解析に用いた仮想データ（使用率は%）

年齢	場面差	snuck 使用率
85	0	18
75	0	28
65	1	95
55	1	95
45	0	80
35	0	91
25	0	92
15	0	98

目的変数は snuck 使用率，説明変数の *X1* は「年齢（時間）」，*X2* を「場面差（特殊な心理状態の有無）」，*a1*，*a2*，*b* を定数，$Z = a1 \cdot X1 + a2 \cdot X2 + b$ とした。

最尤推定法でパラメータを推定した結果，$a1 = -0.076$，$a2 = 0.970$，$b = 4.782$ となった。*p* の予測値は $p = 1/\{1+\exp(0.076 \cdot X1 - 0.970 \cdot X2 - 4.782)\}$で求められる。この式から得た *p* の予測値を図４(１)に示す。比較のため，図４(２)に最小２乗法による単回帰分析の結果を示す。フィッティング精度の違いは一目瞭然であろう。

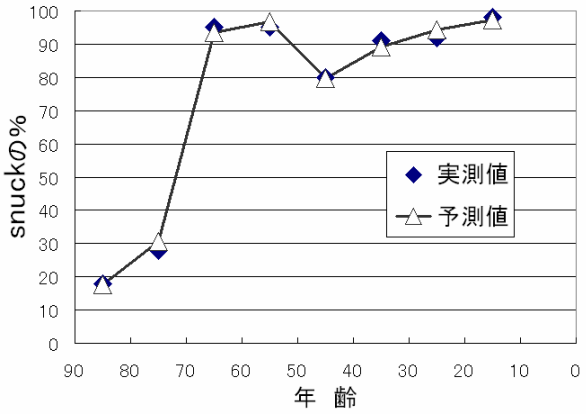

図４（１）　S 字カーブの多変量解析

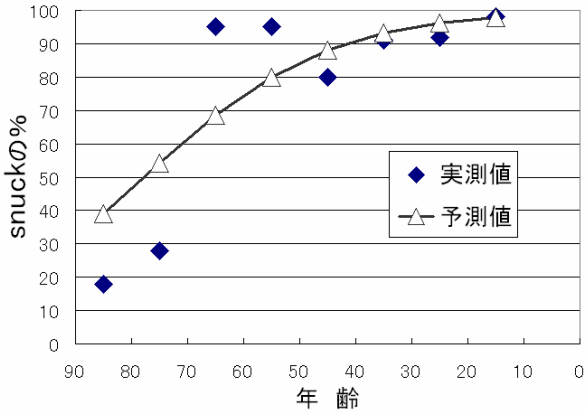

図４（２）　S 字カーブの単回帰分析

このように，従来の単回帰分析による S 字カーブの適用では同時には分析しえなかった

り，ノイズとみなされたりした複数の属性を，ロジスティック回帰モデルは多変量S字カーブで解析することが可能になる。社会言語学の場合は性差，職業，地域指標などの変数をまとめて解析できるほか，古典作品の分析では作品のジャンル，書き手の性別，語種などさまざまな属性を組み込んだ言語変化の検討に道がひらかれるであろう。

５．グロットグラムの新たな解析手法

(１)カテゴリカルなデータの場合

先に示した表１を再掲する。このようなデータの多変量解析は，これまでもっぱら「林の数量化理論Ⅲ類」で行われてきたが，本研究はロジスティック回帰モデルで解析してみる。このデータは各地点１名から得たものであり，目的変数が名義尺度すなわちカテゴリカルなデータである。

表１　仮想グロットグラム　再掲

年齢	A地点	B地点	C地点	D地点
85	△	△	△	△
75	△	△	△	○
65	△	△	○	○
55	△	○	○	○
45	○	○	○	○
35	○	○	○	○
25	○	○	○	○
15	○	○	○	○

目的変数の「○」を１,「△」を０とコーディングし，説明変数の *X1* を「年齢（時間)」，*X2* を「A地点からの距離」，*a1*，*a2*，*a3*，*b* を定数，$Z = a1 \cdot X1 + a2 \cdot X2 + b$ とおいて解析した。A地点からの距離は，A地点をゼロとして隣接地点の値に10を加算した。

最尤推定法でパラメータを推定した結果，$a1 = -3.449$，$a2 = 3.444$，$b = 172.453$ となった。p の予測値は $p = 1/\{1+\exp(3.449 \cdot X1 - 3.444 \cdot X2 - 172.453)\}$ で求められる。

この式に時間とA地点からの距離を投入し，p が0.5以上であれば「○」に，0.5未満であれば「△」と判別した。これは「判別分析」や「林の数量化理論Ⅱ類」と似ているが，説明変数が多変量正規分布しない場合でも問題なく使える等の長所があるため，医学分野では病気の診断などによく用いられる。

分析の結果，表１と同じ表が再現できることが確認された。この手法を使えば，例えば *X2* の「A地点からの距離」を変化させることでB地点とC地点の中間地点での回答を予測できるなどの利点がある。グロットグラムの実査では，複数のセルでデータが欠けてし

まう場合が珍しくないと聞くが，この解析法を工夫することで欠落データを補完するのにも役立つだろう。【末尾注３】採取できたデータから予測式を立てれば，データ欠落セルの年齢や地点を予測式に投入して回答を推定できる。「〇」と「△」の境界線にデータ欠落セルがあるとき，そのまま林の数量化理論Ⅲ類で分析するよりは，この手法を応用して「〇」もしくは「△」を補完した方が明解な解析結果を得られる場合もあるだろう。

（２）比率データで撹乱要因を含む場合

ここでは，回答者が各セルに複数存在して比率が求まり，かつ調査時の情報が得られている場合について検討する。B 地点の 70 歳代のほか，C 地点の 70 歳代と 60 歳代が「特殊な心理状態」で回答したと仮定する。その場合の snuck 使用率を表 3 に示す。A 地点の数値は図 2 と同じで，【末尾注２】付表 1 の距離「0」の snuck 使用率である。それ以外は A 地点から D 地点に向けて隣接地点の値に 10 を加算して作った。

表 3 のデータは B 地点の 70 歳代，C 地点の 70 歳代と 60 歳代の数値が高くなっているため，グロットグラムとしてはきれいな傾向を示していないように見えてしまう。そのことをさらに分かりやすく示すため，使用率が 70%以上のセルに「〇」を，それ以外のセルに「△」を付した。

表３　仮想パーセント・グロットグラム（数値は%）

年齢	A 地点	B 地点	C 地点	D 地点
85	18 △	28 △	38 △	48 △
75	28 △	75 〇	90 〇	58 △
65	44 △	54 △	95 〇	74 〇
55	69 △	79 〇	89 〇	99 〇
45	80 〇	90 〇	100 〇	100 〇
35	91 〇	100 〇	100 〇	100 〇
25	92 〇	100 〇	100 〇	100 〇
15	98 〇	100 〇	100 〇	100 〇

これを視察すると，やはり「〇」と「△」の境界線が単調な右肩上がり（もしくは左肩上がり）にはならないことが分かる。表 3 の数値をグラフ化したのが図 5（1）である。B 地点の 70 歳代，C 地点の 70 歳代と 60 歳代の数値が S 字カーブから大きく外れているため，あまりきれいな傾向を示していない。このようなデータをそのまま林の数量化理論Ⅲ類などに投入すると，明解に解釈できる布置が得られない，あるいは本来の年齢（時間）と空間（距離）の次元がそれ以外の撹乱要因に邪魔されてうまく抽出できない，などの不都合が生じやすくなると予想される。

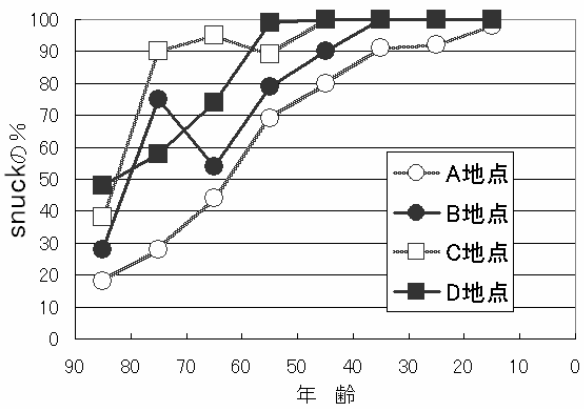

図５（１） 表３の２次元グラフ

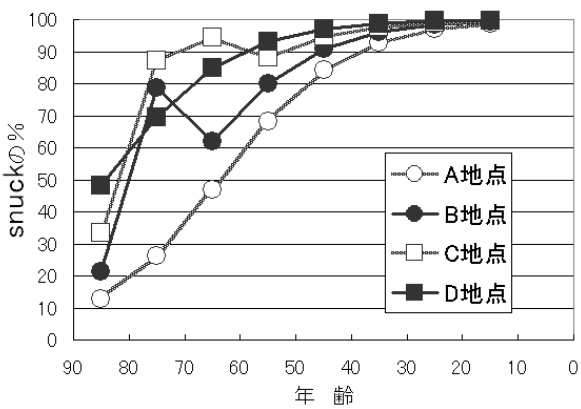

図５（２） 多変量解析による予測値

X1 を「年齢（時間）」，*X2* を「A 地点からの距離」，*X3* を「特殊な心理状態の有無」，*a1*，*a2*，*a3*，*b* を定数，$Z = a1 \cdot X1 + a2 \cdot X2 + a3 \cdot X3 + b$ とおき，最尤推定法でパラメータを推定した。その結果，$a1 = -0.090$，$a2 = 0.062$，$a3 = 1.713$，$b = 5.719$ となった。*p* の予測値は $p = 1/\{1 + \exp(0.090 \cdot X1 - 0.062 \cdot X2 - 1.713 \cdot X3 - 5.719)\}$ で求められる。この式から得た *p* の予測値を図５(２)に示す。これは図５(１)とよく一致する。

(３)比率データから撹乱要因を除外した場合

年齢（時間）と地点（空間）の効果をうまく抽出するには *X3* の「特殊な心理状態の有無」を撹乱要因とみなして，その影響を除外する必要がある。B 地点 70 歳代，C 地点 70 歳代と 60 歳代の３セルが「特殊な心理状態」で回答した。これらが他の回答者と同じように「普通の心理状態」で回答した場合を予測するには前述の予測式の *X3* にゼロを代入すればよい。その結果を表４に示す。表３と比較するため，使用率が 70%以上のセルに「○」を，それ以外のセルに「△」を付した。

表４ 撹乱要因を除外した仮想グロットグラム（数値は%）

年齢	A 地点	B 地点	C 地点	D 地点
85	18 △	28 △	38 △	48 △
75	28 △	40 △	55 △	58 △
65	44 △	54 △	75 ○	74 ○
55	69 △	79 ○	89 ○	99 ○
45	80 ○	90 ○	100 ○	100 ○
35	91 ○	100 ○	100 ○	100 ○
25	92 ○	100 ○	100 ○	100 ○
15	98 ○	100 ○	100 ○	100 ○

表４は「○」と「△」の境界線が単調な右肩上がりになっている。この結果から，D 地点で生じた変化が C 地点→B 地点→A 地点という順番で伝播したことが明確に読み取れる。生データをそのまま示した表３では，場面差（心理状態の差）という撹乱要因によって伝播の起点が C 地点もしくは B 地点であるように見えてしまうが,そのような解釈は妥当ではない。グロットグラムの結果を解釈する際は,「時間と空間だけの効果」を抽出することが肝要である。分析に投入したデータの書式は【末尾注２】を参照されたい。

生データにロジスティック回帰モデルの解析を行い，得られた予測式を利用してデータ欠落セルの数値を補完するとか，撹乱要因を除外した数値を得るといった処理を試みる必要があろう。その後で数量化理論Ⅲ類などを適用すれば，きわめて明解な布置が得られ，次元の解釈も格段に容易になる場合が増えると期待される。つまり，本研究は，１段目の解析でロジスティック回帰モデルによる撹乱要因のフィルタリングを行い，しかる後に２段目の解析（例えば数量化理論Ⅲ類など）に持ち込むという新たな手法を提唱する。おそらく，これまでに蓄積された膨大なグロットグラムのデータをこの新手法で解析し直してみると，撹乱要因の陰に隠れていた真の伝播過程が少なからず発掘されるものと考える。

表４の数値をグラフ化したのが図６である。この結果からも，D 地点で生じた変化が C 地点→B 地点→A 地点という順番で伝播したことが分かる。さらに時間軸における変化はいずれの地点においても基本的に S 字カーブを描いていることが理解できよう。

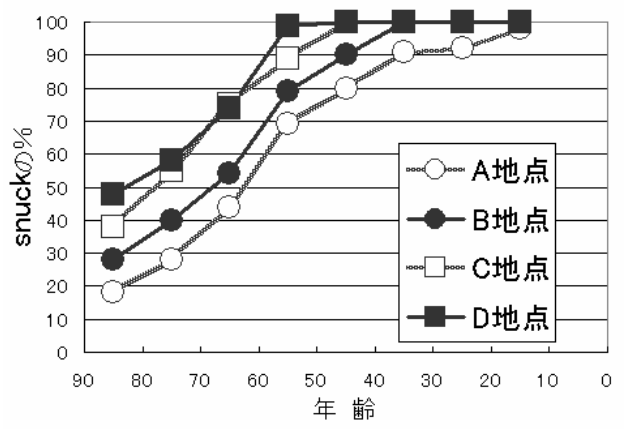

図６　撹乱要因を除外した２次元グラフ

年齢（時間）と地点（空間）の２つの説明変数によるグロットグラムデータをロジスティック回帰モデルで分析し，多変量 S 字カーブに当てはめた３次元グラフを図７に示す。これはロジスティック回帰モデルで求めた理論値（予測値）を３次元で示したものになっている。理論式（予測式）は $p=1/\{1+\exp(0.090\cdot X1-0.062\cdot X2-5.719)\}$であった。撹乱要因 *X3*「特殊な心理状態の有無」は除外した。この図を回転させると，年齢の側から見

ても地点の側から見てもS字カーブを描いていることがみてとれる。井上(2000)は「時間」と「文体差」がもたらす言語変化の多変量 S 字カーブを3次元グラフで示し,「言語変化の水槽モデル」と呼んだ。【末尾注4】図7も水槽モデルと似た形状になっている。

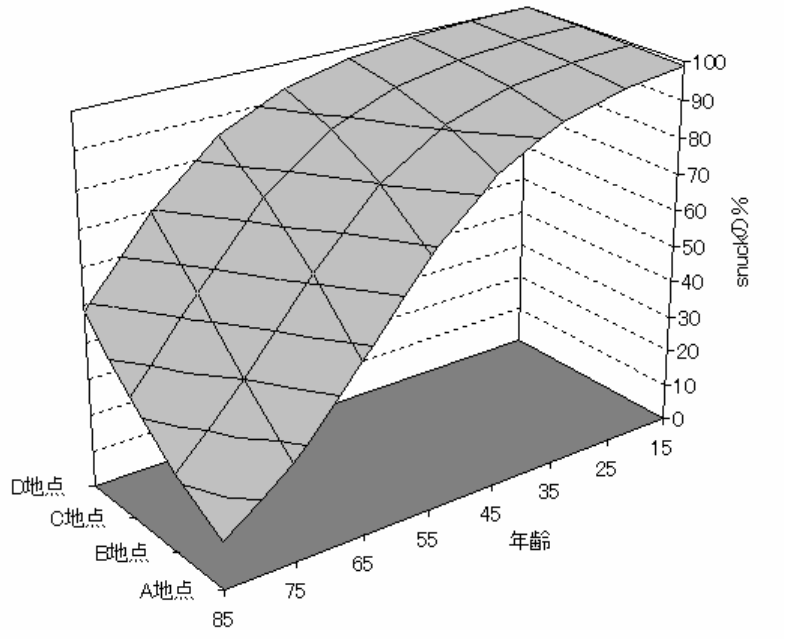

図7　多変量S字カーブの理論式（予測式）による3次元グラフ

6．まとめ：経年調査への応用を目指して

本研究は,言語変化を多変量S字カーブによって説明する手法を提案した。具体的には,理想的なグロットグラムに多変量S字カーブを当てはめ,解析手法の妥当性と有用性を確認した。これは,実査データの解析に向けた模擬実験（シミュレーション）になったと考えられる。

2007年から国立国語研究所は愛知県岡崎市における敬語の調査を開始した。これは岡崎市をフィールドとする敬語調査の第3回目であり,1953年（昭和28年),1972年（昭和47年）に続く過去約50年間以上にわたる実時間での経年調査になる。第1回調査から社会は大きく変わっている。見かけ上の時間や実時間の効果に加えて,さまざまな場面差の要因が言語変化に及ぼす影響を明らかにするために,新たな分析手法が求められている。本研究がその候補の1つになることを期待する。

【末尾注1】ここに示すデータは,2006年10月20日に国立国語研究所で開催された講演会でChambers氏自身が配布した資料の一部による。よってChambers(2006)とした。その配付資料によると,このデータの出典はChambers(1998)となっているがChambers(2006)とは数値が微妙に一致しないことを確認した。本研究は仮想データを対象としたシミュレーションを

目的の１つとしたので，Chambers(1998)ならびに Chambers(2006)のグラフになるべく近い仮の数値を解析に用いた。

【末尾注２】入力データの形式は付表１のとおりであった。ロジスティック回帰モデルによる解析は SPSS によった。

付表１　３変数のロジスティック回帰モデルに投入したデータ

年齢	距離	場面差	snuck 使用率	予測値
85	0	0	18	12.7
75	0	0	28	26.3
65	0	0	44	46.7
55	0	0	69	68.3
45	0	0	80	84.1
35	0	0	91	92.9
25	0	0	92	97.0
15	0	0	98	98.7
85	10	0	28	21.2
75	10	1	75	78.6
65	10	0	54	62.0
55	10	0	79	80.0
45	10	0	90	90.8
35	10	0	100	96.0
25	10	0	100	98.4
15	10	0	100	99.3
85	20	0	38	33.4
75	20	1	90	87.2
65	20	1	95	94.4
55	20	0	89	88.1
45	20	0	100	94.8
35	20	0	100	97.8
25	20	0	100	99.1
15	20	0	100	99.6
85	30	0	48	48.2
75	30	0	58	69.6
65	30	0	74	84.9
55	30	0	99	93.3
45	30	0	100	97.1
35	30	0	100	98.8
25	30	0	100	99.5
15	30	0	100	99.8

【末尾注３】データ欠落セルの補完法について概略を示す。たとえば，表１のグロットグラムにおいて，A 地点（距離「0」の地点）で年齢「55」のデータが欠落していたとする。それを除外した全ケースのデータをロジスティック回帰モデルの解析に投入し，得られた予測式に年齢「55」と A 地点を示す距離「0」を代入して A 地点の年齢「55」について補完値を求める。この場合，説明変数は距離と年齢である。この方法でパラメータを推定すると，表１ですべてのデータがそろっている場合の解析結果とほとんど違いがなく，補完値も（仮想の）実測値ときわめて近い値になる。この方法は簡便で予測精度が高いことから，グロットグラムを使った方言研究の実用面に寄与できる部分があると考える。このほかにも，さまざまな補完法を考えることができる。補完法についてのより詳細かつ具体的な検討は別稿をなす予定である。

【末尾注４】井上(2000)の水槽モデルは共通語化の進行プロセスを大局的にとらえる目的で考案されたものである。時間と文体差（改まりの程度）を説明変数，共通語の普及率を目的変数とする多変量 S 字カーブ（多次元 S 字曲面）の特徴を的確に視覚化している。水槽モデルを統計学の観点からみると，ロジスティック回帰モデルを分かりやすく説明するのに役立つ。

謝　辞

井上史雄先生（明海大学）と鑓水兼貴氏（国立国語研究所）にグロットグラムや S 字カーブについて，松田謙次郎先生（神戸松蔭女子学院大学）にはロジスティック回帰モデルについて多くのご教示を賜った。ここに記して深く感謝申し上げる。

引用文献

Aitchison, J. (1991) *Language change: progress or decay?* 2nd ed. Cambridge: Cambridge University Press.

Altmann, G., von Buttlar, H., Rott, W., & Strauss, U. (1983). A Law of Change in Language. *Historical Linguistics.* (*Quantitative Linguistics.* Vol.18). Bochum: Studienverlag Dr. N. Brockmeyer, 104-115.

Chambers, J. K. (1998). Social embedding of changes in progress. *Journal of English Linguistics*, **26**, 5-36.

Chambers, J. K. (2006). 国立国語研究所講演会資料，私信

Collett, D. (2003). *Modelling Binary Data*. London, United Kingdom: Chapman and Hall.

橋本和佳 (2006).「Logistic 曲線による外来語増加過程のモデル化－大正から平成までの社説を用いて－」『計量国語学』, **25**, 293-308

Hibiya, J. (1988). *A quantitative study of Tokyo Japanese.* Doctoral Dissertation, Dept of Linguistics, University of Pennsylvania.

井上史雄（2000).『東北方言の変遷』，秋山書店
伊藤嘉昭（1994).『生態学と社会：経済・社会系学生のための生態学入門』，東海大学出版会
Labov, W. (1972). *Sociolinguistic patterns.* Philadelphia: University of Pennsylvania Press.
Matsuda, K. (1993). Dissecting analogical leveling quantitatively: The case of the innovative potential suffix in Tokyo Japanese. *Language Variation and Change*, **5**, 1-34.
南部智史(2007).「定量的分析に基づく「が／の」交替再考」『言語研究』，**131**，115-149
真田治子（2002).『近代日本語における学術用語の成立と定着』，絢文社
鑓水兼貴（2006).「東北・北海道における方言文法の共通語化過程」Linguistic Informatics**Ⅵ**, Linguistic Informatics and Spoken Language Corpora－Contributions of Linguistics, Applied Linguistics, Computer Sciences－, Graduate School of Area and Culture Studies, Tokyo University of Foreign Studies, 365-381
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
横山詔一・真田治子（2007).「フィールド言語学にロジスティック回帰分析は寄与しうるか」『情報処理学会研究報告・人文科学とコンピュータ』，2007-CH-73，9-16.
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.

（2007 年 10 月 24 日受付）

日本語要約

論文
多変量S字カーブによる言語変化の解析　－仮想方言データのシミュレーション－

横山詔一（国立国語研究所）・真田治子（埼玉学園大学）

デイスクリプタ：言語変化，S字カーブ，ロジスティック回帰モデル，グロットグラム，最尤推定法

言語変化は一般に「遅→速→速→遅」というS字カーブの過程をたどって伝播するとされる。その記述研究は厚みを増しているが，言語変化のS字カーブに関する理論研究はきわめて少ない。そこで，本研究は仮想方言データ（グロットグラム）を使って，多重ロジスティック回帰モデル（multiple logistic regression model）で解析する新たな手法を提唱し，言語の時系列変化と地理的変化についてシミュレーションを行った。多重ロジスティック回帰モデルは，pを比率（確率），Zを$a1\cdot X1+a2\cdot X2+a3\cdot X3+\cdots+b$というような重回帰式に類する線型結合，対数の底を$e$として，式〔１〕で示される。

$$\log\{p/(1-p)\}=Z \qquad 〔１〕$$

$p/(1-p)$をオッズ，$\log\{p/(1-p)\}$をロジットという。ロジスティック回帰モデルは，左辺がロジット，右辺が重回帰式の形をなす。式〔１〕は変形すると式〔２〕となる。

$$p=1/\{1+\exp(-Z)\} \qquad 〔２〕$$

グロットグラムに代表される方言データの解析では，この多重ロジスティック回帰モデルを用いて$X1$，$X2$，$X3\cdots$に年齢，地点，場面差などの変数をあてはめることで，名義尺度を含む複数の属性を同時に分析したり，撹乱要因を除外して分析したりすることができる。また実データでよく見られるデータの一部欠損にも，高い精度の予測が期待できる。さらに一般的な文献調査でも語彙量・語種比率の変化のほかに，作品のジャンルや書き手の性別などの情報を組み込んだ，広い範囲の言語変化の分析に応用が可能である。

Abstract

Paper

Multiple logistic regression analysis for formulating a change in language

YOKOYAMA Shoichi*, SANADA Haruko**

*The National Institute for Japanese Language
**Saitama Gakuen University

Descriptors : language change, S shape curve, logistic regression, glottogram, maximum likelihood estimation

Many studies show that a process of the language change follows the S shape curve. There are few theoretical studies on the S shape curve of the language change whereas many empirical studies have been published.

This study proposes a new method to apply a multiple logistic regression model for dialectological data, so called glottogram, showing the process of analysis with virtual cases.

A multiple logistic regression model is given as

$$\log\{ p/(1-p) \} = Z, \quad (1)$$

where p is probability, Z is a linear combination shown in the form $Z = a1 \cdot X1 + a2 \cdot X2 + a3 \cdot X3 + \cdots + b$, and log is the logarithm to base e. The equation (1) can be transformed into equation (2) as

$$p = 1/\{1 + \exp(-Z)\}. \quad (2)$$

The multiple logistic regression model can be applied for an analysis of the glottogram and other dialectological data considering the factors like age, the point of observations, the case of situations, and other factors denoting them as the variables $X1$, $X2$, $X3$, etc.

Employing this model, it is possible to analyze two or more factors of dialectological data including those with a nominal scale within one equation, to analyze data excluding disruptors, and to estimate future trend with a high precision even if observed data are not complete, a situation which often faces us. For investigations of written materials it is also possible to analyze a change of vocabulary size, a change of

ratios of word types, or diverse language change with factors such as genre of the text or a gender of the author of the text.