Bayesian Modeling of the Process of Dialect Standardization

ベイズモデルによる方言音声共通語化過程の分析

前川 喜久雄（国立国語研究所 音声言語研究領域・コーパス開発センター）†

Bayesian Modeling of the Process of Dialect Standardization

Kikuo Maekawa (National Institute for Japanese Language and Linguistics)

要旨

国立国語研究所が山形県鶴岡市で収集した共通語化調査データのうち第 1〜3 回調査の音声項目データを用いて、方言音声共通語化過程の統計モデルを構築した。既に報告した第 1 回調査データと同様、第 2 回・第 3 回調査データも二項分布に基づくロジスティック回帰モデルを適用するには分散が大きすぎる（過分散状態）。そのため、ベルヌーイ分布の成功確率が種々の要因によって変動するベイズモデルを考案した。7 種のモデルの性能を F 値・平均予測誤差・WAIC の三者で評価した結果、回帰直線の切片が話者と語彙の要因によって変動し、傾きが語彙の要因によって変動するモデルが最良モデルとなった。このモデルの F 値は 0.95 に達しており、強い説明力を有している。さらにこのモデルにおける話者の個体性情報を「性別・言語形成地域・教育歴」の情報で置換したモデルを評価したところ、第 2・第 3 回調査データについては、最良モデルとほぼ同等の性能を発揮するものの第 1 回調査については性能がかなり低下することが判明した。

１．はじめに

国立国語研究所が 1950 年以来、ほぼ 20 年間隔で 4 回実施してきた山形県鶴岡市における社会言語学的調査（以下鶴岡調査と呼ぶ）は、方言の共通語化過程をリアルタイムで追跡したデータとして有名である（国語研 1953, 1974, 2007）。2017 年 5 月には鶴岡調査データのうち第 1〜3 回調査の音韻項目（36 項目）のデータベースが一般公開された。現在、このデータはエラー修正を施した最新版が公開されており、誰でも自由に利用することができるオープンデータとなっている（関連 URL 参照）。

以下では、一般公開されたデータを用いて、鶴岡における共通語化を統計的にモデル化することを試みる。鶴岡調査データのうち第 1 回調査の音声項目については既に報告済みであるが（前川 2017）、本稿では、第 2 回・第 3 回調査データの音声項目に分析を施し、3 回の調査で把握された共通語化プロセスの異同について議論する。

２．第１回調査音声項目の分析結果

最初に前川(2017)（以下では前報と呼ぶことにする）の成果をまとめておく。前報における成果のひとつは、第 1 回調査音韻項目データは、二値データ（方言 vs 共通語）ではあっても、二項分布には従っていないことの発見であった。音声項目は 36 個の質問からなり、質問に対する被調査者（話者）の回答は原則として共通語（0）か方言か(1)の二値データとして記録されている。しかしこれを成功確率 p、試行回数 N=36 の二項分布から生成されたデータとみなして、話者の年齢による回帰モデルを構築しても予測精度は F 値で 0.57 程度にしか達しない。その原因はデータに観察される分散が理論値 N*p*(1-p)に較べて大きすぎる（過分散）ことにあり、話者の年代別にみると、最小でも 7 倍以上、最大で 10 倍近い分散が観察された。このように顕著な過分散が生じる原因は、p の値が年齢以外の様々な要因

† kikuo@ninjal.ac.jp

の影響を受けて変動するためと考えられた。

前報では、そのようなデータに対応しうるモデルとして、個々の調査項目の成功（共通語形）と失敗（方言形）をベルヌーイ分布で予測する回帰モデルを作成した。説明変数としては、話者の年齢に加えて、音韻クラス、調査項目、話者の個体差をとりあげ、これらの変数が回帰式の切片ないし傾きに影響を及ぼすか否かを様々に組み合わせたモデルを比較検討した。最終的に選択されたのは、話者ごと・調査項目ごとに切片が変化し調査項目ごとに傾きが変化するモデルであり、F 値、平均予測誤差、WAIC のいずれに関しても最良の値を示した。このモデルの F 値は 0.823 であった。

3．今回の分析

今回の分析では、第 2 回・第 3 回調査データの音韻項目に対して前報と同様の分析を施し、その結果を比較検討する。また前回の検討では対象外とした社会的属性の影響を検討するために、話者の個体差の要因を、話者の性別、言語形成地域、教育歴の 3 要因でどの程度まで代替できるかという問題もあわせて検討する。ただしその前に第 2 回・第 3 回調査データの性格を第 1 回と比較する形で把握しておく必要がある。

3．1　第 2 回・第 3 回調査データの過分散指数

今回分析対象とするデータは、鶴岡市民を母集団として無作為抽出されたデータであり、鶴岡調査関係の文献では経年調査データと呼ばれるタイプのデータである。図 1 に第 1〜3 回の調査データの生年代ごとの分布状態をバイオリンプロットで示す。縦軸が共通語化得点（0-36）、横軸が 10 年刻みの生年代（1，2…は 10 代、20 代を表す）である。また図中の丸印は平均値を示している。

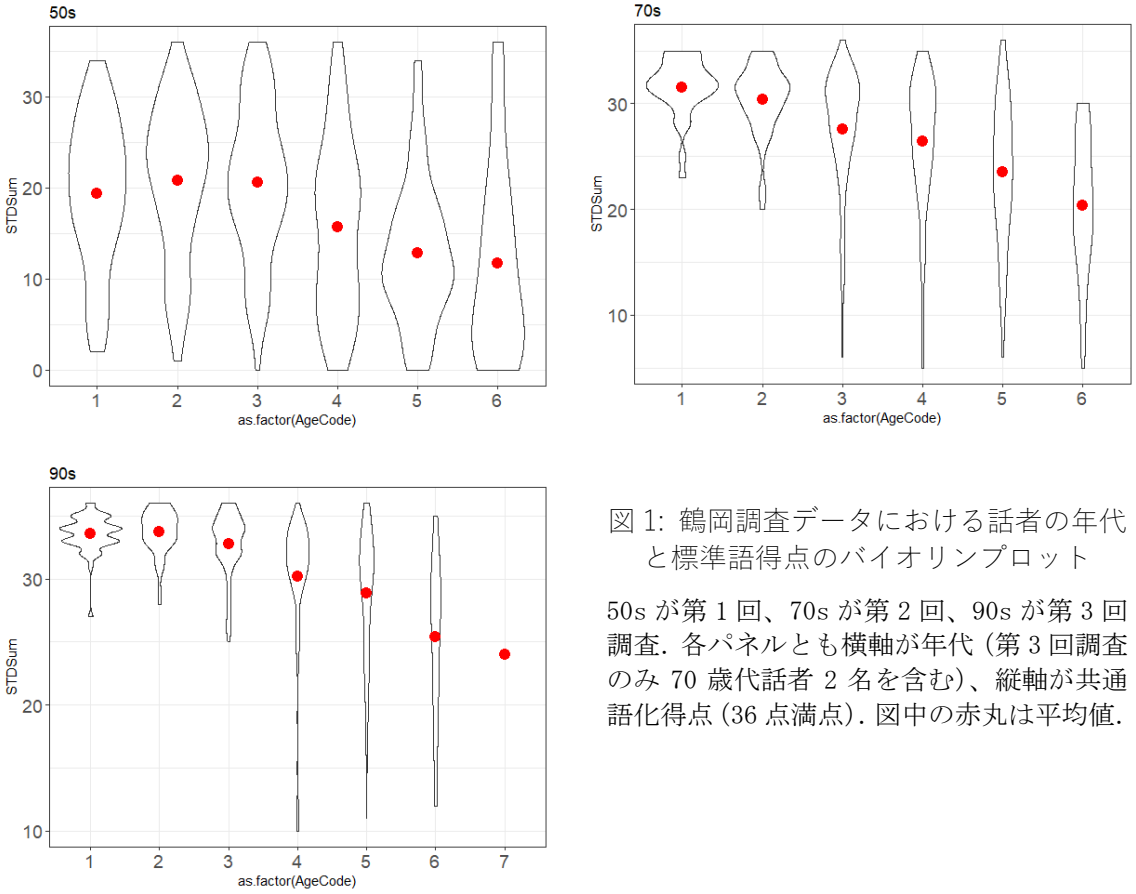

図 1: 鶴岡調査データにおける話者の年代と標準語得点のバイオリンプロット

50s が第 1 回、70s が第 2 回、90s が第 3 回調査．各パネルとも横軸が年代（第 3 回調査のみ 70 歳代話者 2 名を含む）、縦軸が共通語化得点（36 点満点）．図中の赤丸は平均値．

第 1 回調査の分布に比較すると、第 2・第 3 回調査データの分布は特に若年層において分布域を収斂させる傾向をみせてはいるものの、まだ多くの年代において縦軸の全域にわたってサンプル（つまり話者）が分布している。

表 1 は図 1 と対応させる形でデータの過分散指数を計算した結果である。過分散指数は、観察された分散÷分散の理論値で計算される値であり、分散の理論値は N×p×(1-p)で計算する。試行回数 N は 36 であり、成功確率 p の値には観察されたデータから計算した確率を充てた。表 1 をみると過分散状態を完全に脱しているのは第 3 回調査の 10 代だけであることがわかる。

表 1. 調査別・年代別の過分散指数と話者数

	第 1 回調査		第 2 回調査		第 3 回調査	
年代	過分散指数	話者数	過分散指数	話者数	過分散指数	話者数
10 代	7.54	57	1.95	31	0.99	45
20 代	7.94	95	2.31	68	1.45	52
30 代	8.16	131	5.33	99	2.22	86
40 代	10.07	97	5.53	86	5.84	74
50 代	7.33	7	5.60	70	4.80	74
60 代	14.11	39	4.80	47	5.02	*68

*70 代 2 名を含む

３．２　音韻クラスと調査項目

データが過分散状態に陥る原因は共通語化の成功確率 p が何らかの要因で変動することにある。同一年代に属する話者の個体差が大きな要因であるが、それ以外に、前報で確認できたように、音韻のクラスやさらには同一クラス内の調査語彙（調査項目）ごとに p が変動している可能性もある。この点を確認したのが図 2 である。

図 2 の各パネルは鶴岡調査音韻項目を構成する 7 種の音韻クラス（「アクセント」「中舌化」「i と e」「唇音化」「口蓋化」「前鼻音化」「有声化」）別に、各クラスに含まれる調査項目の平均共通語化率の調査毎の推移を比較している。全体的傾向としては第 1 回から第 2 回へ、また第 2 回から第 3 回へと進むにつれ、平均共通語化率が上昇する。しかし、第 3 回調査の段階においても、音韻クラス間の格差は解消されていない。「i と e」「唇音化」「口蓋化」「前鼻音化」「有声化」では、共通語化の平均値が 0.9 前後に達しているが、「中舌化」では 0.8 程度、「アクセント」では 0.5 以下である。

クラス内での調査語彙間の差も、全体としては調査が進むにつれて減少している。しかし、アクセントのように調査項目間の差が拡大したクラスもあることは注目に値する。

結論として、第 2 回・第 3 回調査データのモデリングにおいても、すべての調査項目に対して同一の共通語化確率を当てはめる（つまり二項分布を想定する）ことには無理がある。そこで前報と同じく、共通語化得点を予測するのではなく、ひとつひとつの調査項目における回答が共通語形（1）か非共通語形（0）かを予測するベルヌーイ分布に基づく回帰モデルを採用し、共通語化の成功確率 p に音韻クラスや調査項目が影響を及ぼすモデルを考案することにした。この場合、予測すべきサンプルの数は理論上は表 1 の話者数を調査項目数(36)倍したものになるが、話者によって無回答などの欠損値が存在することがあるので、若干減少する。

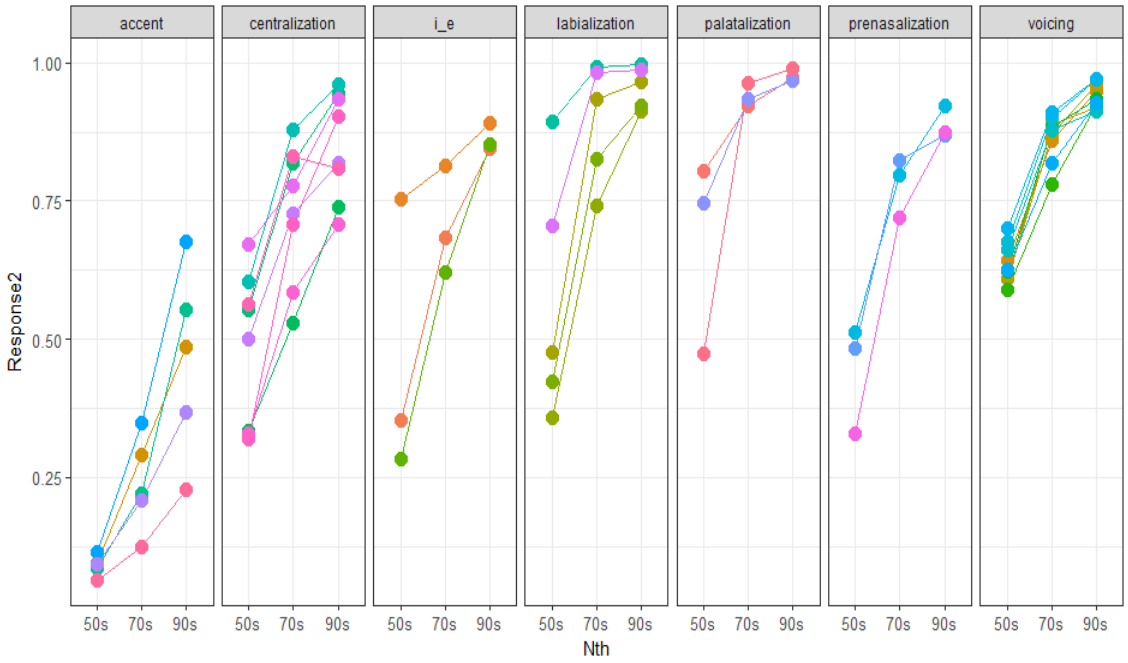

図 2: 音韻クラスおよび調査語彙による平均共通語化率の調査毎の推移
各パネルが音韻クラス（左から「アクセント」「中舌化」「i と e」「唇音化」「口蓋化」「鼻音化」「有声化」）の別、パネル内の線が調査項目の別を示す.

4．統計モデリング

4．1　ベイズ回帰モデルによる予測

前節の分析結果を踏まえ、今回の分析でも前報と同様、複数のベイズ回帰モデルを考案して比較することにした。本節で検討の対象とする 7 種のモデルの特徴を表 2 にまとめた。Stan 言語による回帰モデルのプログラム（後掲する図 3-5 参照）では、i 番目のサンプルに関するベルヌーイ分布の成功確率 q[i]が inv_logit()関数によって決定される（図 3 の 21 行、図 4 の 31 行、図 5 の 35 行参照）。その際、inv_logi()関数の引数は、`a[i]+ b[i]*Age[i]`のような話者の年齢（`Age[]`）の一次式の形をしている。表 2 で切片、傾きと呼んでいるのは、この一次式の切片（`a[]`）と傾き（`b[]`）のことである。

表 2 のモデル 1～4 は前報におけるモデル 1～4 と同一であり、表 2 のモデル 6, 7 はそれぞれ前報におけるモデル 5, 6 と同一である。表 2 のモデル 5 は話者の個体差単独での影響をより詳しく検討するために、今回新たに追加したモデルである。

これらのモデルのパラメータはベイズ統計の手法によって推定し、モデルの実装には Stan 言語を利用した。図 3 に表 2 のモデル 5 にあたる回帰モデルのベイズ推定用 Stan プログラムを示す。これ以外のモデルの推定に用いた Stan プログラムは前報に掲載されているので参照されたい。

図 3 も含め、今回利用したベイズモデルでは推定すべきパラメータのレンジを指定しているだけで、事前分布は指定していない。このような場合、Stan は無情報事前分布として一様分布ないし非常に大きな分散をもつ正規分布を採用する。Stan プログラムは R 言語からライブラリ Rstan を介して実行した。MCMC による事後分布のシミュレーションの実行条件に関しては、チェイン数を 3 に固定し、iteration は 2000～4000、warmup は 1000～2000、

thinning は 1〜3 の範囲で、モデルの収束状態を観察しながら調整した。収束の判定条件としては rhat < 1.1 を採用した。これは Stan による分析で広く採用されている判定条件である（松浦 2016）

表 2：7 種の回帰モデル

モデル	特　徴
1	一次式の切片も傾きも一定のモデル（ベースライン）
2	音韻クラスごとに一次式の切片と傾きの両方が変化するモデル
3	語彙ごとに一次式の切片と傾きの両方が変化するモデル
4	話者ごとに一次式の切片だけが変化するモデル。傾きは固定
5	話者ごとに一次式の切片と傾きの両方が変化するモデル（2018.01.19 に追加）
6	話者ごとに一次式の切片が変化し語彙ごとに傾きが変動するモデル
7	話者ごと・語彙ごとに切片が変化し語彙ごとに傾きが変化するモデル

```
01:  #BernLogitRegHie4_waic.stan by KM
02:  #話者によって切片と傾きが変化する Bernoulli 階層ロジスティック回帰
03:  data {
04:    int I;                          //データ総数
05:    int Nsbj;                       //話者数
06:    int<lower=14,upper=70> Age[I];  //話者の年齢
07:    int<lower=0, upper=1> Y[I];     // i 番目のデータの共通語化状態(1 ないし 0)
08:    int<lower=1, upper=Nsbj> Subject[I];   // 話者の個体識別番号
09:  }
10: parameters {
11:  real<lower=-5, upper=5> as[Nsbj];
12:  real<lower=-0.2, upper=0.2> bs[Nsbj];
13: }
14: transformed parameters {
15:   real a[I];
16:   real b[I];
17:   real<lower=0, upper=1> q[I];
18:   for (i in 1:I) {
19:     a[i] = as[Subject[i]];
20:     b[i] = bs[Subject[i]];
21:     q[i] = inv_logit(a[i] + b[i]*Age[i]);
22:   }
23:  }
24: model {
25: for (i in 1:I){
26:   Y[i] ~ bernoulli(q[i]);
27:   }
28: }
29: generated quantities {
30:   real y_pred[I];
31:   real log_lik[I];
32:   for (i in 1:I){
33:     y_pred[i] = bernoulli_rng(q[i]);
34:     log_lik[i] = bernoulli_log(Y[i], q[i]);  //WAIC のために対数尤度を保存
35:   }
36: }
```

図 3: ベイズ推定による回帰分析の Stan プログラム（表 2 のモデル 5）

モデルによる予測の評価指標も前報と同様、平均予測誤差・F 値・WAIC の 3 種を用いる。平均予測誤差はモデルによって予測された 36×話者数個のサンプルの値（0 か 1）と実際の

観測値との差の絶対値の平均である。平均予測誤差は理解しやすいが、予測すべきデータが均等に分布していない（例えば大部分が 0 で 1 が僅かであるなど）場合、適切な指標となりがたい。

F 値はこの欠点を補うために考案された指標であり、モデルの予測値の適合率(例えば 1 と予測したもののうち実際に 1 であったものの割合）と再現率(例えば実際に 1 であったもののうち 1 と予測されたものの割合）の調和平均として定義される。

最後に WAIC は AIC を非正規分布に基づくモデルにも適用できるよう拡張した新しい情報量基準である。WAIC による判定は交差検定(cross validation)と漸近等価であると考えられている（関連 URL 参照）。

表 3: 第 1 回調査データの予測

モデル	平均予測誤差	F 値	WAIC
1（ベースライン）	0.422	0.568	24121
2（音韻クラス～切片・傾き）	0.419	0.676	21503
3（語彙～切片・傾き）	0.298	0.708	20318
4（話者～切片）	0.284	0.714	20154
5（話者～切片・傾き）	0.283	0.716	20156
6（話者～切片、語彙～傾き）	0.179	0.821	15267
7（話者・語彙～切片、語彙～傾き）	0.175	0.823	14684

表 4: 第 2 回調査データの予測

モデル	平均予測誤差	F 値	WAIC
1（ベースライン）	0.261	0.850	15923
2（音韻クラス～切片・傾き）	0.185	0.882	12463
3（語彙～切片・傾き）	0.178	0.885	11886
4（話者～切片）	0.230	0.858	14982
5（話者～切片・傾き）	0.229	0.858	14984
6（話者～切片、語彙～傾き）	0.143	0.907	10591
7（話者・語彙～切片、語彙～傾き）	0.131	0.914	9993

表 5: 第 3 回調査データの予測

モデル	平均予測誤差	F 値	WAIC
1（ベースライン）	0.151	0.919	11428
2（音韻クラス～切片・傾き）	0.128	0.928	8720
3（語彙～切片・傾き）	0.120	0.932	8318
4（話者～切片）	0.141	0.922	10908
5（話者～切片・傾き）	0.142	0.921	10899
6（話者～切片、語彙～傾き）	0.093	0.947	7440
7（話者・語彙～切片、語彙～傾き）	0.088	0.949	6960

表 2 の各回帰モデルのパラメータを図 3 のようなベイズモデルで推定し、その予測性能を評価した結果を調査別に表 3-5 として示す。どの調査においても、またどの評価指標に関しても、モデル 7 が最良と判定されている。[1] この結果は、前報における第 1 回調査データ

[1] 平均予測誤差は小さいほど、F 値は 1.0 に近いほど、そして WAIC は小さいほどモデルの性能が良いと判断する。

のモデリングの結果とも一致している。

次に調査間の差に注目すると、第 1 回よりも第 2 回、そして第 2 回よりも第 3 回調査データの方が、より高い精度で予測できていることがわかる。このような結果は、表 1 に示した過分散指標から予想されるところではあるのだが、第 3 回調査データ（表 5）の場合は、ベースラインモデルの F 値が既に 0.9 を上回っており、最良モデルと判定されたモデル 7 以外のモデルも実質的に高い性能を発揮していることが注目される。

４．２　話者の個人差の性別・出生地・教育歴による代替

表 3-5 の結果は、共通語化の要因として、話者の個体差が語彙の個体差（調査項目の差）と同程度に予測に貢献していることを示していた。[2] 本研究で話者の個体差をモデルに含める場合、M 名の話者は共通語化に関して M 通りの異なる状態をとりうると考えている（図 3 のプログラムでは、19-21 行でその関係が実装されている）。しかしこのように文字通りの意味での個体差を想定することは、鶴岡調査を含む従来の社会言語学の分析では稀であり、年齢・性別・出身地・教育等の諸属性の集合によって話者の個体差を代替していることが多い。以下ではこのような代替の有効性を鶴岡調査のデータを用いて定量的に検討する。話者の年齢以外の属性として、性別・言語形成地域・教育歴の影響を評価する。

この目的のために、新たに 2 個の回帰モデルを考案した。表 2 のモデル 6（話者が切片にだけ影響し、語彙が傾きにだけ影響するモデル）をベースラインとして、話者の個体差（純粋な個体差）を「話者の性別＋言語形成地域」の組み合わせで代替するモデル（これをモデル 8 とする）と、「話者の性別＋言語形成地域＋教育歴」で代替するモデル（モデル 9）の 2 種類である。

これらの回帰モデルの Stan プログラムを図 4, 5 として示す。図 4 の 31 行では、一次式の切片を int1[i]と int2[i]の二つの分離しており、前者が話者の性別（図 4 の 28 行参照）、後者が言語形成地域（同 29 行参照）による影響を被るモデルとなっている。同じく図 5 では一次式の切片を 3 個に分割し（図 5 の 35 行）、それぞれが性別、言語形成地域、教育歴による影響を被るモデル（同 31-33 行）となっている。

```
01: # BernLogitRegHie10_waic.stan
02: # 話者(Subject2)を性別と言語形成地情報でどれだけ代替できるかの検討
03: # 話者に替えて、性別と言語形成期によって切片が変化し、
04: # Item2 をハイパーパラメータとして傾きが変化する Bernoulli ロジスティック回帰
05: # by KM  2018.01.25
06: data {
07:   int I;
08:   int Nitm;
09:   int Nsbj;
10:   int <lower=1, upper=2> Sex[I];
11:   int <lower=1, upper=3> BP[I];
12:   int<lower=1, upper=36> Item[I];
13:   int<lower=14,upper=70> Age[I];
14:   int<lower=0, upper=1> Y[I];
15:   int<lower=1, upper=Nsbj> Subject[I];
16: }
17: parameters {
18:   real<lower=1,upper=2> asx[2];
19:   real<lower=-1,upper=2> abp[3];
```

[2] ベースライン（モデル 1）とモデル 3 ないしモデル 5 の相違を比較せよ。

```
20:   real<lower=-0.2,upper=0.1> bs[Nitm];
21: }
22: transformed parameters {
23: real<lower=-2, upper=2> int1[I];
24: real<lower=-2, upper=2> int2[I];
25: real<lower=-0.5, upper=0.5> b[I];
26: real<lower=0, upper=1> q[I];
27: for (i in 1:I) {
28:   int1[i] = asx[Sex[i]];
29:   int2[i] = abp[BP[i]];
30:   b[i] = bs[Item[i]];
31:   q[i] = inv_logit(int1[i] + int2[i] + b[i]*Age[i]);
32: }
33: }
34: model {
35: for (i in 1:I){
36:   Y[i] ~ bernoulli(q[i]);
37:   }
38: }
39: generated quantities {
40:   real y_pred[I];
41:   real log_lik[I];
42:   for (i in 1:I){
43:     y_pred[i] = bernoulli_rng(q[i]);
44:     log_lik[i] = bernoulli_log(Y[i], q[i]);
45:   }
46: }
```

図 4: モデル 8 の Stan プログラム

```
01: # BernLogitRegHie11_waic.stan
02: # 話者に替えて、性別と言語形成期と教育歴をハイパーパラメータとして
03: # 切片が変化し、Item2 をハイパーパラメータとして傾きが変化する
04: # Bernoulli 階層ロジスティック回帰
05: # by KM  2018.01.25
06: data {
07:   int I;
08:   int Nitm;
09:   int Nsbj;
10:   int <lower=1, upper=2> Sex[I];
11:   int <lower=1, upper=3> BP[I];
12:   int <lower=1, upper=9> Edu[I];
13:   int<lower=1, upper=36> Item[I];
14:   int<lower=14,upper=70> Age[I];
15:   int<lower=0, upper=1> Y[I];
16:   int<lower=1, upper=Nsbj> Subject[I];
17: }
18: parameters {
19:   real<lower=1,upper=2> asx[2];
20:   real<lower=-1,upper=2> abp[3];
21:   real<lower=-2,upper=2> aed[8];
22:   real<lower=-0.2,upper=0.1> bs[Nitm];
23: }
24: transformed parameters {
25:   real<lower=-2, upper=2> int1[I];
26:   real<lower=-2, upper=2> int2[I];
27:   real<lower=-2, upper=2> int3[i];
28:   real<lower=-0.5, upper=0.5> b[I];
29:   real<lower=0, upper=1> q[I];
30:   for (i in 1:I) {
31:     int1[i] = asx[Sex[i]];
```

```
32:     int2[i] = abp[BP[i]];
33:     int3[i] = aed[Edu[i]];
34:     b[i] = bs[Item[i]];
35:     q[i] = inv_logit(int1[i] + int2[i] + int3[i] + b[i]*Age[i]);
36:   }
37: }
38: model {
39: for (i in 1:I){
40:   Y[i] ~ bernoulli(q[i]);
41:   }
42: }
43: generated quantities {
44:   real y_pred[I];
45:   real log_lik[I];
46:   for (i in 1:I){
47:     y_pred[i] = bernoulli_rng(q[i]);
48:     log_lik[i] = bernoulli_log(Y[i], q[i]);
49:   }
50: }
```

図 5: モデル 9 の Stan プログラム

表 6-8 に各属性に関する話者数の分布を示す。

表 6: 話者の性別の分布

	第 1 回調査	第 2 回調査	第 3 回調査
男	212	228	317
女	284	280	396

表 7: 話者の言語形成地域の分布

	第 1 回調査	第 2 回調査	第 3 回調査
鶴岡市	338	318	456
山形県内	117	139	178
山形県外	38	49	75
不明	3	2	4

表 8: 話者の教育歴の分布

	第 1 回調査	第 2 回調査	第 3 回調査
小学校	145	61	37
中学校	184	194	211
高校	83	197	275
専門学校	8	22	37
旧制高校	5	1	8
大学	12	23	66
その他	31	6	70
無し	24	1	--
無回答	4	3	9

表 9 にモデル 8, 9 による予測精度の評価結果を示す。表 9 からは、話者の純粋な個体差を話者の社会言語学的な属性で代替したモデルの予測精度はベースラインに劣ることが分かる。つまり、社会言語学的な属性は話者の個体性を十全には把握できていない。

ただし、代替による劣化の程度は調査によって異なっている。図 6 は表 9 の F 値だけを

調査間で比較したグラフである。第 1 回調査データにおいては、モデル 8, 9 はベースラインであるモデル 6 から顕著に劣化しているが、第 2 回調査では劣化幅が減少し、第 3 回調査では一層減少していることがわかる。

表 9: モデル 6, 8, 9 による予測の精度

調査	モデル	予測誤差	F 値	WAIC
第 1 回	6	0.179	0.821	15267
	8	0.289	0.724	19571
	9	0.268	0.737	18805
第 2 回	6	0.143	0.907	10591
	8	0.184	0.884	12043
	9	0.178	0.887	11665
第 3 回	6	0.093	0.947	7440
	8	0.122	0.931	8477
	9	0.120	0.932	8262

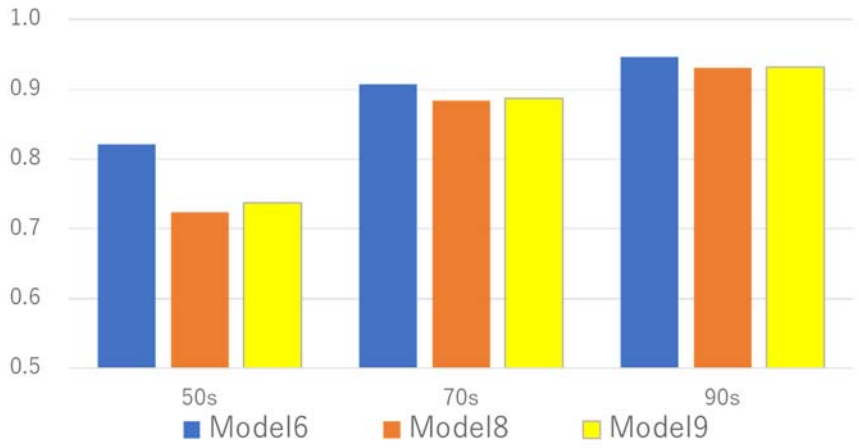

図 6: モデル 6, 8, 9 の F 値の 3 回の調査を通しての比較

5. 議論と結論

今回の分析では、まず第 2 回・第 3 回調査データも、二項分布によって生成されたとみなすと、過分散状態に陥ること、従って、前報と同じベルヌーイ分布に基づく回帰モデルでの分析が望ましいことを確認した。

その方針に従って種々の回帰モデルを比較検討したところ、前報で最良モデルとして採用したモデルが第 2 回・
第 3 回調査データにおいても最良の予測をもたらすことが判明した。「話者ごと・語彙ごとに切片が変化し語彙ごとに傾きが変化する」モデルである。

一方、第 1 回調査データと第 2 回・第 3 回調査データの間には明らかな差も認められた。最も単純なベースライン回帰モデル（切片も傾きも固定で、話者の年齢だけを変数とするモデル）による予測の F 値は、第 1 回調査では 0.57 にとどまっているのに対して、第 2 回調査では 0.85 を、また第 3 回調査では 0.9 を上回っている（表 4, 5 参照）。これからわかるように、第 1 回調査に比べると第 2 回・第 3 回調査では年齢による予測の有効性が大きく向上している。前報では、話者の年齢が言語変化（共通語化）の要因として最も重要であると

いう社会言語学上の仮定は無批判には受け入れられないと考えたが、共通語化が進展した段階ではこの仮定はある程度まで有効になることが確認できた。

また、年齢以外の話者の個体差を、性別・言語形成地域・教育歴で代替したモデル 9 による予測結果をみても、第 1 回調査データと第 2 回・第 3 回調査データとの間には質的な相違が認められた。第 1 回調査では代替によって顕著な劣化が生じるが、第 2 回・第 3 回では劣化の幅が小さい。この点においても、従来の社会言語学的分析は共通語化がある程度まで進展したデータに対してよりよく適合しているといえる。

しかし、以上を換言すれば、共通語化の初期状態は、従来の社会言語学的分析手段によっては十分に把握しきれないということである。初期段階にある言語変化の分析においては、話者と調査対象語彙の個体差に関して格別の配慮が必要となることを今回の分析は示している、というのが本稿の結論である。

本研究の次のステップとしては、今回は独立して分析した第 1 回から第 3 回までのデータを統一的に分析するための統計モデルの開発が挙げられる。

謝　辞

鶴岡調査の話者と調査者、ならびに鶴岡調査データベースの公開に尽力された国立国語研究所の関係者に感謝します。本研究は国立国語研究所コーパス開発センターの共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」(2016-2021 年度)の成果です。

文　献

国立国語研究所.『地域社会の言語生活―鶴岡市における実態調査―』（国立国語研究所報告 5）秀英出版, 1953. doi/10.15084/00001214

国立国語研究所.『地域社会の言語生活―鶴岡市における 20 年前との比較―』（国立国語研究所報告 52）秀英出版, 1974. doi/10.15084/00001251

国立国語研究所.『地域社会の言語生活―鶴岡における 20 年間隔 3 回の継続調査―』国立国語研究所, 2007.

前川喜久雄「鶴岡市共通語化調査データの確率論的再検討」言語資源活用ワークショップ 2017 発表論文集, pp.163-180, 2017.09.06. doi/10.15084/00001517

松浦健太郎『Stan と R でベイズ統計モデリング』共立出版, 2016.

関連 URL

鶴岡調査データベース：http://www2.ninjal.ac.jp/longitudinal/tsuruoka.html

WAIC：http://watanabe-www.math.dis.titech.ac.jp/users/swatanab/waic2011.html