データサイエンスが使えるようになるまで

前川喜久雄（国立国語研究所）

１．はじめに

この発表のタイトルは意図的に曖昧にしてあります。格助詞の「が」を対格と解釈すれば、私がデータサイエンスを使えるようになるまでの経緯という意味になり、主格と解釈すれば、むかしは言語研究に使えなかったデータサイエンスが最近は使えるようになったその経緯という意味になります。以下ではこの両面に触れたいと思います。筆者がこれまでに携わった言語研究の領域は、音声学と社会言語学と言語資源の開発です。これらの実践を通して 40 年ほど統計学、今流にいえばデータサイエンスとつきあってきました。もちろんその道のプロではなく、優れたユーザーでもありません。なんとか 40 年間ボロをださずにやってきたユーザーにすぎません。以下はそんな猫かぶりユーザーが 40 年間の折々に感じたことをまとめた感想文です。

２．統計学との出会い

以下の話題と関係するので、私が統計学・データサイエンスにどのように関わったかをふりかえっておきます。最初に統計学の必要性を感じたのは、修士論文で日本人がフランス語の母音をどう知覚するかという問題をあつかったため、実験心理学を勉強したことが契機でした。不確実性をともなうデータの解釈において、帰無仮説の検定が果たす役割の重要性はすぐに理解できました。当時は麻布にあった統計数理研究所の統計技術者養成講座に半年通って終了証をもらいましたが、統計学には検定論以外に広い領域があることを知ったのが成果でした。多変量解析に触れ、統計的予測という考え方があることを知りました。

博士後期課程に進学してからは当時経済学部で統計学を教えておられたM先生にほとんどマンツーマンで離散空間の確率論の指導をうけることができました。これは大きな幸運でした。博士課程の 2 年目に初めて査読論文を書きました。津軽・下北地方の母音体系と出雲地方の母音体系ではどちらもイとエの区別が曖昧だとされていますが質的な相違があるように思われたので、その差をフォルマント周波数で表現して検定しました。ただし普通に分散分析をおこなうのではなく、フォルマント空間でイとエの母音対を結ぶベクトルの角度が一様分布に従っていたら母音の音韻的対立がないという帰無仮説をたてて、これを検定する統計手法を提案しました。その結果、津軽・下北では帰無仮説は棄却されませんが、出雲では帰無仮説が棄却されました。ただし出雲の対立は東京に比べると対立が不明瞭であることもわかりました。この論文は 1984 年春に計量国語学に掲載され(前川 1984)、数年後にはその後の進展を含めて英語でまとめ直したものを水谷静夫氏（だろうと思います）の推薦でドイツの計量言語学の雑誌の特集号に掲載してもらえました(Maekawa 1989)。

1984 年に就職した鳥取大学では学生たちと方言学・社会言語学的な調査を行い、ほかに文学全集からサンプリングしたテキストを使った文体分析なども試みました。後で触れる

Varbrul(variable rule analysis)の理論に触れたのもこの時期でした。1987 年に鳥取市でランダムサンプリング調査による共通語化調査を行ったところ、国立国語研究所が山形市鶴岡市で行った有名な共通語化調査とは共通語化の進展がかなり異なることがわかりました。

1989 年には国立国語研究所に移りました。新しい職場で何をなすべきか、いろいろと考えましたが、結局、社会言語学からは撤退して音声研究に集中することにしました。そして、統計学に関しては、自分の数理的な才能の限界にかんがみて、今後はユーザーに徹することにしました。1991 年には山形県鶴岡市での第 3 回共通語化調査が始まり、私も調査に協力させてもらいました。鶴岡で調査をしてみて強く感じたのは話者の個人差の大きさでした。

1999 年からは言語資源の設計と実装が仕事の中心になりました。この後のことは比較的世間によく知られているようですから省略します。2004 年に『日本語話し言葉コーパス』を、2010 年に『現代日本語書き言葉均衡コーパス』を公開した後は、主に『日本語話し言葉コーパス』を利用した日本語の音声学的な分析にとりくんでいます。

この 40 年間で、比較的まじめに統計学・データサイエンスを勉強したのは最初の 10 年と最後の 10 年、つまり 1980 年代と 2010 年代でしたが、2010 年頃に勉強を再開したときは、20 年の間に統計学・データサイエンスが劇的に進歩したことを知り驚きました。

３．統計手法の問題点

３．１．カウントデータ

1980 年当時に私が学んだ統計学は記述統計以外には推計学（検定論）と多変量解析（重回帰分析のように説明変数群の値を与えられたときに目的変数の未知の値を予測するための統計モデル）でした。しかしこれらをそのまま言語研究のデータに利用できる事例はあまり豊富ではありませんでした。基本的な原因は、正規分布に基本をおく検定論や多変量解析では、分析対象とする変量も連続量であることを想定しているのに対して、言語データには離散量が多いことが根本的な原因ですが、さらにいくつかの特殊事情もありました。

ひとつには、一見連続量に見える言語データのなかには正規分布からは導けないものが潜んでいることがあります。例えば共通語化調査結果を百分率で表した共通語化率がその例です。共通語化率のもとになっているのは調査票で定められた複数の質問に対する被験者の回答を 1（共通語形）と 0（非共通語形）に分類して合計した値ですが、この数値には下限（0）と上限（質問項目数N）が存在しており、その範囲を超えた値はとりません。また負の値をとることもなく、整数値だけをとります。このようなデータは「上限のあるカウントデータ」と呼ばれ、二項分布に従います。上限のないカウントデータもありますが（例えばある単語が『日本語話し言葉コーパス』に記録された話し手ごとにそれぞれ何回出現したかを数えたデータ）、その場合はポワソン分布に従います。

カウントデータを統計的に解析するとき、例えば共通語化率を年齢から予測する回帰分析をおこなうとき、正規分布を仮定した回帰を行うと、共通語化率が負の値をとったり、N（あるいは 100%）を超えたりすることが生じます。1980 年当時、この問題に対する対策として提案されていた手法はいずれも対症療法的なもので、理論的な対処はできていませんでした。正確にいうと、理論的な解決策は 1970 年代に提案されていたのですが、私のよ

うな末端ユーザーが使えるまでには普及していませんでした。私以外のユーザーも似たようなもので、1980 年代には社会言語学の領域で多変量解析を活用した研究がたくさん世に出ましたが、上記の問題を真剣に考慮した研究はほとんどなかったのではないかと思います。なお、共通語化データのなかには二項分布でもあつかえないものがあります。このことには後で触れます。

３．２　質的変数と量的変数の混在

先に統計学が基本的には連続量を対象としていると書きましたが、1980 年当時でも、質的変数を扱う統計手法は存在していました。そもそも統計的検定では、連続量である変数を質的な変数でふたつないしそれ以上のグループに分け、その間で母集団における平均値に差があるかどうかを検定していますが、これは説明不要でしょう。多変量解析のなかでも、林の数量化理論（特にⅠ類とⅡ類）のように、説明変数（現象の説明に用いる変数、上の回帰分析の例でいえば年齢）として質的な変数（性別、職業、出生地等々）を用いる手法も提案されており、社会言語学では頻繁に利用されていました。

しかし実際の分析では、説明変数に量的変数（年齢、収入など）と質的変数が混在する事例が生じます。先の例と同様、当時、この問題に柔軟に対応できる手法は末端ユーザーには知られていませんでしたので、本来量的変数であるものを便宜的に質的な変数として扱ったり、逆に本来は質的である変数を数値に置き換えたりして分析していました。こういった操作には経験科学として積極的な意味は認められません。それがポジティブな結果を生むこともあったでしょうが、それは僥倖に近いものでした。

３．３．交互作用

ある変量の予測にふたつ以上の説明変数を利用する場合、説明変数間の交互作用が存在することがあります。例えば共通語化率を性別と年齢で説明する場合、年齢の効果が性別によって異なるというようなケースです。統計的な予測の精度をあげるためにはこの交互作用をモデルに組み込む必要があります。1980 年当時でも、分散分析（三群以上の平均値の差の検定問題）では交互作用項を用いることが普通でしたが、多変量解析では何故か交互作用に触れる説明が稀だったように思います。数学的な扱いとして理論的な困難はないはずなので不思議な気がします。余談になりますが、アメリカの社会言語学者たちが交互作用項を彼らの Varbrul モデルに取り入れようとしないことも不思議でした。グラフに明らかに交互作用が現れている場合でも、交互作用なしのモデルで押しとおすのです。彼らの間で広く利用されていたソフトウェアが交互作用を扱えなかったことがひとつの原因だったようですが、残念なことに、この傾向は今も完全にはなくなっていないようです。

３．４．個人差・個体差

言語研究のデータは普通個人差をともなっています。検定論でも多変量解析でもデータを提供する被験者は無作為に抽出されることが前提になっており、そのため、個人差は正規分布にしたがうと仮定して、他の誤差項と一括して処理されています。しかしコーパスを構築したり、社会言語学的調査に参加したりしていると、被験者の無作為性が必ずしも保証されていないことが気にかかってきます。特に被験者数が限られたデータの場合、目的変数の

変動の少なからぬ部分がむしろ個人差に帰着していることがあります。同種の問題は、音素の知覚実験をおこなうときに、問題の音素がどのような単語の一部として提示されるかで結果が異なるというような形でも生じます。しかし昔の多変量解析モデルで個人差・個体差を扱おうとすると、個体数Nに対してN-1個の説明変数を用いる必要があり、統計モデルの良さが損なわれる結果をまねきました。

４．統計学・データサイエンスの進歩

４．１．GLM と GLMM

前節で指摘した問題は、現代の統計学・データサイエンスではほぼ完全に解決されています。1970 年代に提案された GLM（一般化線形モデル）がその突破口となりました。GLM では正規分布だけでなく、ベルヌーイ分布、ポワソン分布、二項分布など、指数分布族と呼ばれる一連の確率分布が扱えるようになりました。その結果、上述のカウントデータもポワソン分布ないし二項分布を用いて分析することが可能になりました。特に後者は、言語研究で頻出する 0 か 1 かの二項対立を表すデータの処理に適していることから、ロジスティック回帰分析の名前で言語や音声の研究で広く利用されています。カウントデータにおけるデータの下限と上限の問題はロジスティック回帰分析を使えば解消できます。GLM の特徴としては、ほかに質的変数と量的変数が混在したモデルを扱えることがあります。これは言語研究の実務においては大変役に立つ特徴です。また、これは当然というべきですが、交互作用項を含めた分析も可能です。

上述の諸問題のうち GLM で扱えなかったのは個人差の問題でしたが、この問題は、GLM を拡張した GLMM（一般化線形混合効果モデル）によって解決されました。混合効果というのは、説明変数として固定効果とランダム効果（変量効果とも）のふたつを含んだ統計モデルを意味します。固定効果(fixed effect)は従来からの説明変数のことで、実験研究の場合であれば、実験者があらかじめ実験計画に組み込んで統制しておいた変数をさします。一方、ランダム効果はあらかじめ実験に組み込まれてはいないけれども、実際には実験データに影響を及ぼしている（あるいはその可能性が考えられる）、制御されていない変数のことをさします。例えば共通語化の社会調査データを検討したら、調査員によって共通語化率が変動しているらしい、というような場合、調査員をランダム効果に指定するのです。心理言語学的な実験では、被験者と語彙がよく変量効果に指定されます。コーパスの分析では、実験とはちがって、何が固定効果で何がランダム効果かが自明ではありませんが、そのような場合は、その研究において本質的な興味の対象である変数を固定効果とし、そうではないけれどもデータへの影響が予想される変数をランダム効果に指定することになります。

４．２．ベイズモデル

統計学・データサイエンスは、GLM・GLMM の開発によって、それ以前の多変量解析にくらべて格段に高いフレキシビリティを実現し、その結果、経験科学上無理な仮定を置くことなく、広い範囲のデータが扱えるようになりました。しかしそれでも自由度が十分でないことも起こりえます。先ほど社会調査における調査員を変数（ランダム効果）に指定するという例を挙げましたが、その効果が調査員の性別や出身地によって変化するというモデル

を考えることができます。さらに調査員の性別と被調査者の性別の組み合わせによる交互作用が存在するかもしれません。こうした例は分析に用いる変数が独自のパラメータをもっていて、その値によって変数の効果が変化するモデルを意味しています。GLMM でモデルを計算するためには最尤法という手法が用いられますが、最尤法でこのような複雑なモデルを解くことには限界があります。そこで解析的な計算ではなく、確率論的なシミュレーションによって複雑なモデルのパラメータの分布を推定しようというアプローチが用いられるようになりました。これがベイズモデリングです。ベイズ統計は確率の主観的解釈の理論として 18 世紀にまでさかのぼる歴史をもっており、現代のベイズモデリングもその理論の枠組みにそった仕組み（事前分布の指定）のなかでおこなわれますが、そのような理論的ないし哲学的な枠組みとは一応切り離して、自由度の高い統計モデルのパラメータを確率シミュレーションによって推定する技術としてドライにとらえることもできます。

４．３．利用環境の変化

現代の統計学・データサイエンスを支える重要な要素となっているのがコンピューティング環境です。1980 年代と比較するとハードウェアもソフトウェアも劇的な進化をとげています。CPU 速度や記憶容量が何万倍にも増えたこともちろん大きな変化ですが、それ以上に劇的な変化は、今世紀に入ってからは R 言語や STAN 言語のような先端的な統計ソフトウェアが無償で提供されるようになったことではないかと思います。ほんの十数年前までは、SAS のような信頼性の高い統計ソフトウェアは何十万円かのお金を払って毎年ライセンスを購入するものでした。フリーソフトの文化は統計学・データサイエンスの利用環境を抜本的に改善してくれました。そしてもうひとつ、大規模な言語資源が容易に入手できるようになったことも近年の大きな変化のひとつです。言語研究上の仮説をなにか思いついたときに、データを集める手間を経ずにとりあえず検証してみることができる環境は、これもまた 20 年前には想像することのできないものでした。

５．ひとつの実例

ここでひとつ分析の具体例を示します。国立国語研究所が山形県鶴岡市で 1950 年代、70 年代、90 年代の各初頭に実施した地域社会の共通語化に関する面接調査のデータです。被験者は住民票に基づいて無作為抽出されており、大変信頼性の高いデータです。また現時点では音声項目だけですが、データは国語研究所のホームページで公開されており、誰でも利用することができます（国立国語研究所 1953, 1974; 高田 2019）。

図 1 は 3 回の調査で得られた音韻項目共通語化率を重ね合わせたグラフです。3 本の折れ線が 50 年代、70 年代、90 年代の調査に対応しており、話者の年代ごとの平均値を結んでいます。横軸の 1,2,3,…が年代（10 代, 20 代, 30 代…）を表し、縦軸は音韻調査 36 項目のうち何項目に標準語形で回答したかの得点を示しています。最低値は 0、最大値は 36 です。平均値を示す丸印の上下にのびたエラーバーは標準誤差（標準偏差をデータ数の平方根で除した値、母集団における平均値の標準偏差の推定値）です。どの調査においても、時間の経過とともに（つまり話者が若くなるほど）共通語化が進行する傾向を読みとることができます。このグラフの変種は国内外の多くの文献で引用されており、日本の社会言語学を代

表する成果のひとつとみなされています。しかし図 1 は実はデータサイエンス的にはかなり misleading です。この図は年齢が共通語化の大きな要因であることを示唆しています。しかし、二項分布に基づく GLM で個々の話者の年齢から共通語化得点（Y 軸の値）を予測してみても、よい予測モデルは構築できません。実際に第 1 回調査のデータ（凡例の 50s）を分析してみると、平均予測誤差が 6.95 になります。満点が 36 ですから、許容しがたいほど大きな誤差です。この問題は、図 1 のデータが二項分布に従っていないことを疑わせます。二項分布のパラメータは試行における成功の確率 p と試行の回数 N ですが、分布の分散は Np(1-p)になります。つまりパラメータが決まれば分散は自動的に決まります。この点に着目して第 1 回調査のデータを年代ごとに分析してみると、データから計算される分散は、二項分布を仮定した場合の分散の理論値よりもはるかに大きく、最低でも 3.5 倍（20 代）、最大だと 9.8 倍（60 代）に達しています。このようなデータは「過分散」の状態にあると言われます。図 2 に第 1 回調査データの話者 433 名の共通語化率と話者の年齢の散布図を示しました。年齢を問わず得点の幅が非常に広いことが分かります。

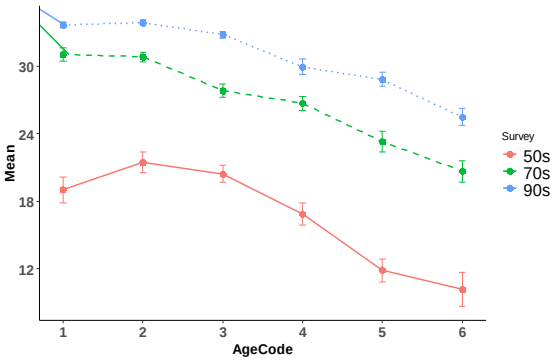

図 1. 鶴岡市における共通語化

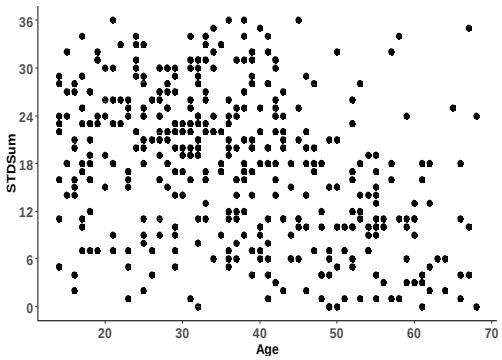

図 2. 鶴岡第 1 回調査における全被験者の年齢と共通語化得点の散布図

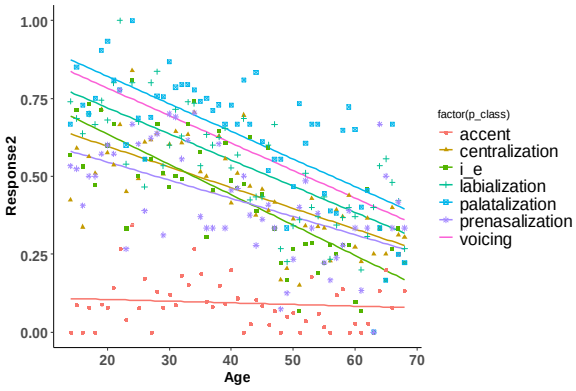

図 3. 鶴岡第 1 回調査における音韻クラス別の散布図と回帰直線

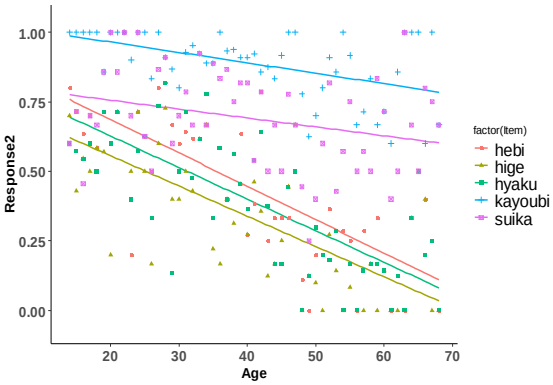

図 4. 音韻クラス「唇音化」に属する 5 項目の散布図と回帰直線

言語学的にみると、この問題は、調査項目ごとに共通語化の進み方が大きく異なっていることに起因しています。図 3 は、鶴岡調査の音韻関係 36 調査項目を「アクセント」「中舌

化」「イとエ」「唇音化」「口蓋化」「鼻音化」「有声化」の 7 音韻クラスに分類して、クラス毎に散布図と回帰直線を描いたものです。顕著な特徴はアクセントの共通語化率が年齢によらず非常に低いことですが（図の底に貼りついているのがアクセントの回帰直線)、それ以外の 6 クラスでも、年齢と共通語化率の間に相関はあるものの、同時にクラス間の差も存在していることが分かります。そして、このような差は各クラスの内部にも存在します。図 4 は「唇音化」クラスを構成する 5 個の調査項目「蛇」「髭」「百」「火曜日」「西瓜」ごとに散布図と回帰直線を描いたものです。年齢による共通語化の進み方がクラス内部でも一様ではなく、少なくともふたつのグループに分けられることがわかります。

要するに共通語化の進行には話者の年齢という社会的要因以外に多くの言語的要因が関係しているのです。そのことは第 1 回鶴岡調査の報告書（国立国語研究所 1953）でも明らかにされており、周知の事実なのですが、データサイエンスの観点からすると様々な要因に配慮した統計モデルの構築が課題になります。詳しい議論は省略しますが、図 2-4 に示されたような特性をもつデータは GLM や GLMM を用いても、うまく分析できないと考えられたので、ベルヌーイ分布によるベイズモデリングを試みました（前川 2017)。

ベルヌーイ分布というのは、コイン投げのように 1 回の試行で勝ち負けが決まる現象の分布です。これを N 回繰り返すと二項分布になるのですが、二項分布ではすべての試行で同じ確率が適用されるのに対し、ここでは話者や調査項目の特性に従って 1 回ごとに成功確率（共通語形が得られる確率）が変化するモデルを考案しました。具体的にはベルヌーイ分布の成功確率を q、それに影響する i 番目の要因を Xi としたとき q = Ai + Bi Xi の一次式で q を予測することとし、一次式の切片 Ai ないし傾き Bi の分布を MCMC という手法でシミュレートしました。

表 1. 種々のベイズモデルの評価

モデル	予測誤差	F 値
モデル 1：年齢によって切片と傾きが変化	0.420	0.571
モデル 2：音韻クラス毎に切片と傾きが変化	0.338	0.676
モデル 3：調査項目ごとに切片と傾きが変化	0.296	0.710
モデル 4：話者ごとに切片が変化	0.286	0.713
モデル 5：話者ごとに切片が、調査項目ごとに傾きが変化	0.180	0.819
モデル 6：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	0.174	0.823

詳しくは文献をご覧いただくこととして（推定用プログラムも掲載してあります)、結果を表 1 にまとめました。ここでの予測誤差は 0 か 1 かの予測についての誤差です。F 値というのは、正解における正負の分布の偏りを考慮に入れた正解率だと考えてください。F 値は 0.0-1.0 の範囲で変化し、予測が完璧であれば 1.0 となります。表 1 は話者と調査項目に配慮したベイズモデリング（モデル 6）を行うことで、年齢だけ（モデル 1）ではほぼチャンスレベルであった予測性能を F 値で 0.8 以上にまで高められたことを示しています。目的にもよりますが、予測することに意味があるといえる数字になっています。なお、モデルを複雑化させれば性能があがるのは当たり前だと考える人がいるかもしれません。そのとお

りです。そこで、WAIC というモデルの複雑さを考慮に入れた評価指標も計算してみましたが、やはりモデル 6 が最良という結論が得られました。また、ここでは第 1 回調査のデータだけに触れましたが、第 2 回、第 3 回のデータを分析した結果からも、やはりモデル 6 が最良という結論が得られています（前川 2018）。現代的なデータサイエンスの観点からすれば、鶴岡市の共通語化はモデル 6 の形で表現され、年齢は話者のもつひとつの属性として影響を及ぼしていると解釈することになります。そこが年齢だけに特化した図 1 の解釈との根本的な違いです。このような解釈をデータの裏付けをもって主張できるようになったのは、近年のデータサイエンス発展の功徳です。

6．おわりに（オープンデータについて）

現在、少なくとも私の関係している言語研究の領域では、データサイエンスないし統計科学が使える技術になりました。そして末端ユーザーも次第にデータサイエンスが使えるようになりつつあります。しかしデータサイエンスはあくまで技術です。データサイエンスがもたらす分析上のすばらしい可能性は、データがあってこそ花開きます。この点はいくら強調してもしすぎることはありません。Garbage in garbage out は不磨の金言です。

そしてデータが大切という場合、きちんとしたデータを作ることが大切なのは当然ですが、それを研究コミュニティ内で公開し共有することも、データ作りに劣らず大切です。先に紹介した鶴岡の共通語化の場合も、データが公開されているからこそ再分析が可能になったのですが、鶴岡調査データが公開されたのはごく最近（2017 年）のことです。先に述べたように、私は 1991 年の第 3 回調査に参加しましたが、実際にデータを分析するまでに 26 年待たねばなりませんでした。

現在、世界的にオープンデータ、オープンサイエンスの潮流が大きくなりつつありますが、研究者の活動に関する説明責任や科学への市民参加という側面だけでなく、実は、建設的な批判を可能にするという点で、科学そのものの健全な発達にとって、オープンデータは必須の要件なのです。もちろん、言語研究も例外ではありません。データサイエンスの発展と価値の高いデータの共有化とが手を携えてともに進むことを願いつつ筆をおきます。

文献

国立国語研究所(1953)『地域社会の言語生活 : 鶴岡における実態調査』国立国語研究所報告 5. do/10.15084/00001214

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

高田智和(2019)「鶴岡調査データベース」計量国語学, 32(2), pp. 96-102. doi.org/10.24701/mathling.32.2_96

前川喜久雄(1984)「母音の合一と混同の理論 －津軽，出雲方言を例として－」計量国語学, 14(4), pp.149-162.

前川喜久雄(2017)「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. doi/10.15084/00001517

前川喜久雄(2018)「ベイズモデルによる方言音声共通語化過程の分析」言語資源活用ワークショップ 2018 発表論文集, pp.326-336. doi/10.15084/00001667

Maekawa, K.(1989) “Statistical tests for the study of vowel merger”. *Quantitative Linguistics*, 39, pp.200-219.