Vol.2010-CH-88 No.8
2010/10/30

定点経年調査による共通語化研究のための データベース構築

阿部貴人†

国語研究所が過去 40 年間・３回にわたって調査し，蓄積してきた共通語化調査の回答結果をデータベース化した。
その作成過程での技術的・内容的な検討課題とその解決方法を中心に，共通語化のメカニズム・プロセス解明に寄与するデータベースの全容を報告する。

Database architecture for research of Standardization of Dialect

Takahito ABE†

National Institute for Japanese and Linguistics investigated the longitudinal survey of the standardization of dialect in Tsuruoka city of Yamagata Prefecture in Japan, over the 40 years
Focusing on an agenda and their solutions in the process of creating the technical content of the database for standardization of dialect.

1. 目的

国立国語研究所は山形県鶴岡市で地域社会における共通語化の実態調査を経年的に実施してきた．1950 年の第 1 回調査から約 20 年間隔で 1991 年まで 3 回行われた調査は,住民基本台帳にもとづく無作為抽出サンプルを中心とした大規模なものであり，言語変化に関する調査としてはデータの質と量の両面で世界をみわたしても群を抜いている．

サンプリング調査はいずれも約 400 名前後を対象としたほか，同一人物への追跡調査（パネル調査）も行ってきた（国立国語研究所報告書，1953，1974，1994，2007）．

この過去３回の回答データを整備し，地域社会（＝サンプリング調査）と個人（＝パネル調査)の 40 年間にわたる言語変化をみることのできるデータベースを作成した．本稿は，作成にあたっての懸案事項とその対処などを交えながら，鶴岡調査回答データベースの概要を紹介する．

1.1 鶴岡市の位置

鶴岡市は，山形県の庄内地方南部に位置する．旧鶴岡藩（通称，庄内藩）の城下町で，文化・経済の中心都市として栄えた．時代小説の作家である藤沢周平の作品に登場する「海坂藩」のモデルと言われる．

過去３回の調査時の人口は，第一次調査時（1950 年）が約 96,000 人，第二次調査時（1971 年）が約 95,000 人，第三次調査時（1991）が約 100,000 人と，大きな人口変動のない，日本に点在する中小都市の典型である．

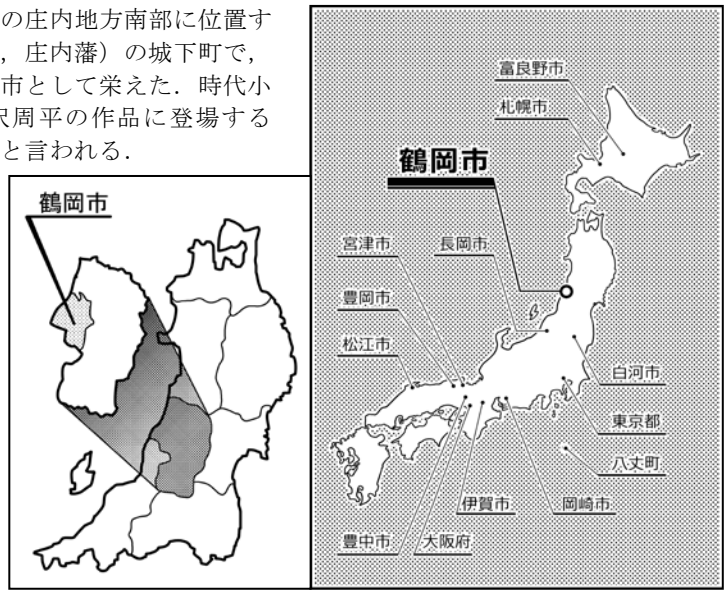

図 1 鶴岡市の位置

† 大学共同利用機関法人・国立国語研究所
National Institute for Japanese and Linguistics

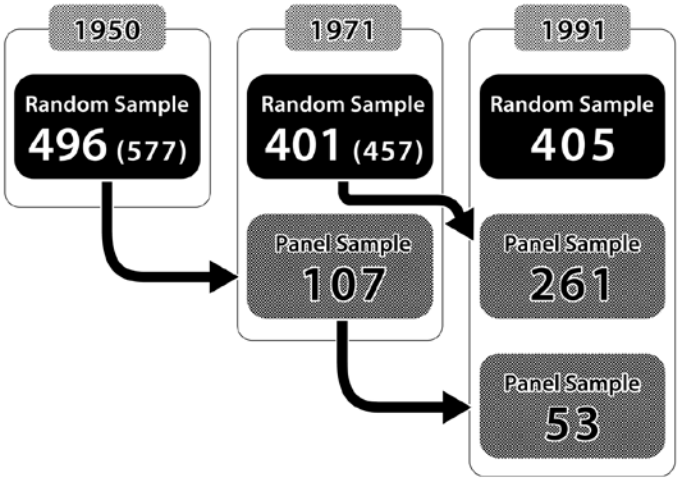

図 2 サンプリング調査とパネル調査の構成

1.2 調査デザイン：世界最初のコーホート系列法

鶴岡調査のデザインを図 2 に示す．これは，後で述べる横断法と縦断法を組み合わせた形になっており，ランダムサンプルが横断法に，パネルサンプルが縦断法にそれぞれ対応する．図 2 と同じ調査デザインは，生涯発達心理学や老年学でも採用されることがあり，「コーホート系列法（cohort sequential method）」と呼ばれている．鶴岡調査の具体的なデザインは以下のとおりである．

- 第 1 回調査：住基台帳を用いてサンプルをランダム抽出（577 名）
- 第 2 回調査：第 1 回のサンプルを追跡調査（107 名）＋新たなサンプルをランダム抽出（457 名）→計 564 名のデータ
- 第 3 回調査：第 1 回のサンプルを追跡調査（53 名）＋第 2 回のサンプルも追跡調査（261 名）＋さらに新たなサンプルをランダム抽出（405 名）→計 719 名のデータ

図 2 のようなコーホート系列法は，時代効果，加齢効果，世代効果の 3 者を分離可能な「もっとも効率的な調査法」として諸学界から高い評価を受けているが，手間がかかる．そのため，まともな実査データは世界中で，鶴岡共通語化調査，岡崎敬語調査のほかには，たった 1 つ「シアトル調査」があるのみのようである．

1.3 調査内容

鶴岡調査は，様々な調査を組み合わせることによって，共通語化現象を多角的に捉えてきた（表 1）。

その調査群の中で，過去 3 回のいずれでも実施されている中心的な調査が「共通語の調査」（表の太枠で囲んだ A，F，G，H，I）である。それらは，音声（音韻），語彙，文法項目について，調査員が調査票を用いて個別面接方式で実施する調査であった。言語的事象以外では，性別・生年といった被調査者の属性，マス・コミュニケーションとの接触度，方言や共通語に対する意識など，共通語化に関連する可能性のある種々の情報についても尋ねた。

表 1　鶴岡調査の調査構成

	調査名称	抽出	調査対象者	調査対象数（完了数）	調査方法
第一次	A.共通語の調査	○	15～69 歳	496 名	個別面接
	B.24 時間調査	×	ネイティブ	3 名	行動観察
	C.ﾊﾟｰｿﾅﾘﾃｨの調査	○	A.の Info.から	156 名	個別面接郵送留置
		×	高校生	8 名	生活記録の収集
	D.ﾏｽ･ｺﾐｭﾆｹｰｼｮﾝの調査	○	小学生の父兄世帯	1,011 世帯	個別面接
	E.学校における共通語指導状態の調査	×	小･中･高の教官	137 名	留置調査
第二次	F.共通語の調査 1（ﾗﾝﾀﾞﾑｻﾝﾌﾟﾙ調査）	○	15～69 歳	401 名	個別面接
	G.共通語の調査 2（ﾊﾟﾈﾙｻﾝﾌﾟﾙ調査）	×	A.の Info.から	107 名	個別面接
第三次	H.共通語の調査 1（ﾗﾝﾀﾞﾑｻﾝﾌﾟﾙ調査）	○	15～69 歳	405 名	個別面接
	I.共通語の調査 2（ﾊﾟﾈﾙｻﾝﾌﾟﾙ調査）	×	A･F･G の Info.から	314 名	個別面接
	J.言語生活調査 1（ﾗﾝﾀﾞﾑｻﾝﾌﾟﾙ調査）	○	H と同一 Info.	405 名	留置調査
	K.言語生活調査 2（ﾊﾟﾈﾙｻﾝﾌﾟﾙ調査）	×	I と同一 Info.	314 名	留置調査
	L.場面差調査	○	H と J の Info.から	175 名	個別面接
	M.方言記述調査	×	ネイティブ	10 名	個別面接

国立国語研究所の研究情報資料センターでは，研究資料の高度学術利用を目的に，

「共通語の調査」の回答データをリレーショナルデータベース化するプロジェクトを進めている。次節では，その整備中の回答データベースの構成と，今後のデータ公開の予定について述べる．

2. データベースの構築

現在整備中の過去３回の「共通語の調査」の回答データベースについて説明する．

データ整備の方法は時代をよく反映している．第１回調査（1950 年）のデータ整備は，手書きで記したカードを用い，それを年代別や性別といった観点で分類し，カードの数を数えて集計した．第２回調査（1971 年）のデータ整備になると，大型計算機が導入された．パンチカードを計算機に読み込ませて集計し，それをデータシートという紙に打ち出して利用した．そして，第３回調査（1991 年）のデータ整備では，既に普及が始まっていたパソコンが利用された．

改めて過去３回のデータを整備するとき，手書きのカードやパンチカードといった資料は,決して利用価値のない過去の産物ではない.データを整備するにあたっては，それらの資料がどのようにして作成されたのか，つまり，調査票からカードに回答を転記する段階で，どのような整理を行ったのかを知る必要が出てくる．大抵の回答というものは，コード化されてカードに転記されるのであり，そのコードの種類，コード化の基準といったものを知らなければ，過去の調査結果・報告と，現在のそれとがリンクしない．かつての調査研究と現在（そして未来）を結ぶ，社会調査でいうところの「受け渡し」は，その調査研究にとって重要な事柄なのである．

コードの種類や基準といったものをまとめたコーディング・マニュアルがあれば良いかというと，それだけでもいけない．各話者のそれぞれの回答がどのように整理されたかは, 結局のところ, 調査票そのものに戻って確認するしか方法はない. そして，このような確認作業は，調査が実施される度に繰り返されるのである．

そのような確認作業は，過去のデータを知るうえで必須であるとしても，できるだけ無駄を省いて, 効率化した方が良い. そこで, 今回のデータベース化にあたっては，調査票に記された回答をできるだけそのまま転記するフィールドを設けた．これにより，将来，第四次，第五次調査が実施され，過去に遡ってデータが整備される際も，調査票の原票に戻る必要は起こらないだろう．硫酸アルミニウム等の影響で酸化・劣化した紙を，恐る恐る繰りながらも「破壊」を続ける心配もない．

なお，鶴岡調査回答データベースでは，コード化した回答データに加えて，できる限りローデータも研究者に提供できるように整備している．調査研究では，何ら手を加えないローデータが電子化され，公開されることは，珍しい部類に入る．ローデータを公開する理由は，研究者が回答をそれぞれの視点でコード化できるようにするためである．

現在，鶴岡調査回答データベースは，図１のような構成で構築中である．

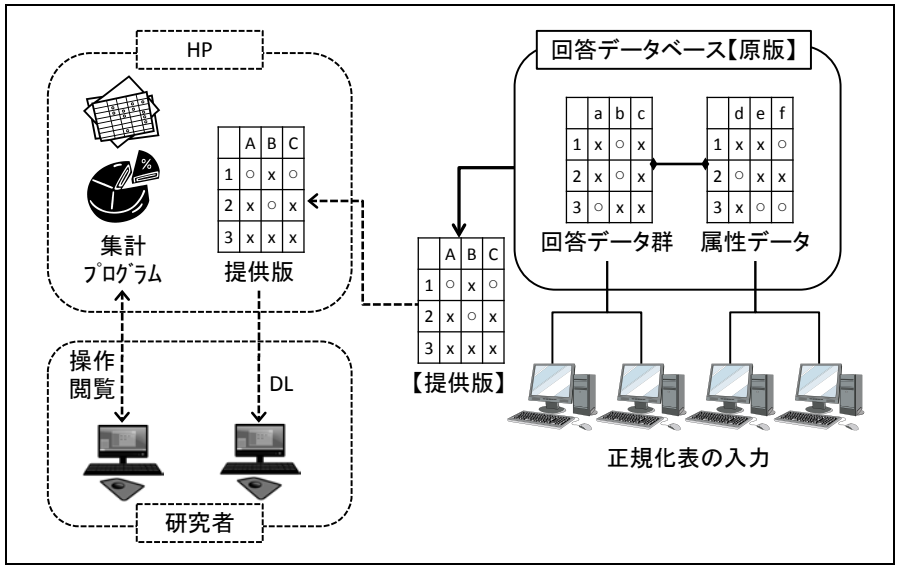

図３ 鶴岡調査回答データベースの構成

鶴岡調査回答データベースは，図１の「回答データベース【原版】」のように，シンプルで典型的なリレーショナルデータベースの構成である．すなわち，各調査項目の回答（図中では「回答データベース群」としてまとめている）と，話者の属性等のデータを関係づけた構成をとる．

以下，データベースを構成する（1）回答データ，（2）原版と提供版，（3）属性データ，（4）ホームページによるデータ公開（予定・検討中）の各要素について説明を加える．

（1）回答データ

鶴岡調査の中核的な調査項目は, 全項目数の過半を占める音声（音韻）項目である．調査項目は，鶴岡方言（および東北方言全般）に特徴的な観点（表 2）を調べることを目的に設定されている. 第１回～第３回調査では, 項目に若干の増減があるのだが, 表２の「調査語」に挙げる 36 項目は第１回～第３回調査で継続して調査されているものである．

Vol.2010-CH-88 No.8
2010/10/30

表 2　音声（音韻）に関する調査項目

分類名	調査観点	調査語
唇音性Ⅰ	合拗音 kwa の有無	スイカ，カヨウビ
唇音性Ⅱ	ハ行における両唇音の有無	ヒゲ，ヘビ，ヒャク
口蓋化	「せ」「ぜ」における口蓋化の有無	セナカ，アセ，ゼイムショ
有声化	非語頭におけるカ行・タ行の有声化の有無	クチ，ハチ，ハト，ネコ，ハタ，クツ，カキ，マツ
鼻音化	非語頭におけるザ行・ダ行・バ行の直前の入りわたり鼻音の有無	マド，スズ，オビ
中舌化Ⅰ	ウ段音における中舌化の有無	チズ，スミ，カラス，キツネ
中舌化Ⅱ	イ段音における中舌化の有無	チジ，シマ，カラシ，ウチワ
イとエⅠ	語頭の母音エにおける狭母音化の有無	エキ，エントツ
イとエⅡ	語頭の母音イにおける広母音化の有無	イキ，イト
アクセント	共通語のアクセント型の実現	セナカ，ネコ，ハタ，カラス，ウチワ

各項目は，(1) 共通語の発音（アクセント型），(2) 伝統的な方言の発音（アクセント型），(3) その他のコードを付している（その他は，その具体的な回答を注記）．

(2) 原版と提供版

調査では様々な情報を得る．例えば，住所，電話番号等の情報も知ることになる．このような情報は，次のパネル調査を計画する際に，個人を追跡するうえで極めて重要な鍵となるものであり，整備が必要である．しかし，個人情報の保護の観点から，このような情報を研究者に提供することはできない．無論，居住する番地といった情報は，言語変化の分析に活用できるものでもない．

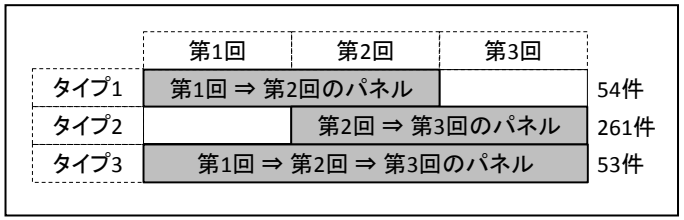

図 4 パネルサンプルのデータベースの構成

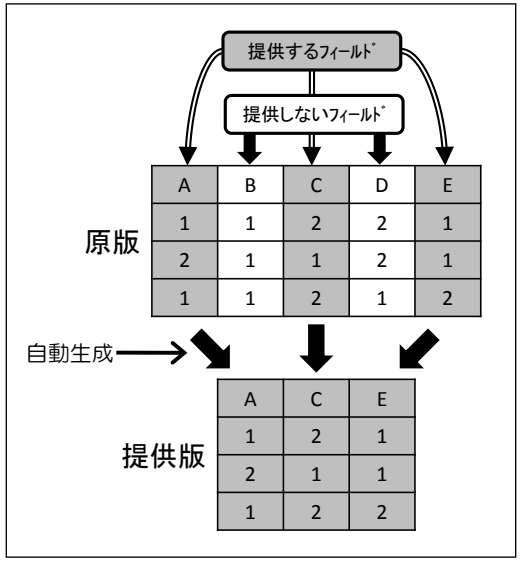

図 5 提供版生成のイメージ

そこで，調査で知り得たすべての情報をデータベース化するとともに，研究者に提供するための，情報を絞ったデータベースも必要になる．前者を【原版】と呼び，後者を【提供版】と呼んでいる．

原版はリレーショナル・データベースの形態をとるが，研究者に供する提供版は回答データと属性データをまとめた Excel のワークシートである．第 1 回・第 2 回・第 3 回の各調査のワークシートと，パネルサンプルだけを抜き出したワークシートから成る（図 4）．それらのワークシートは，研究者に提供できる／提供すべきフィールドを自動的に抜き出して作成されるようにしてある．すなわち，原版に何らかの修正が必要になった場合でも，原版上で修正をすれば，自動的にデータが更新された提供版が出来上がる仕組みとなっている（図 5）．

(3) 属性データ

研究者に提供する属性データは，以下に掲げる 12 の情報である（予定）．

(1) 被調査者 No.　(2) 性　(3) 生年
(4) 調査員名　(5) 言語形成期 1　(6) 言語形成期 2
(7) 言語形成期 3　(8) 本人の職業（職務内容）
(9) 学歴　(10) 父親の出身地　(11) 母親の出身地
(12) 配偶者の出身地

このうち，(5) ～ (7) の 3 種類の「言語形成期」について説明を加える．

言語形成期とは，言語使用者の母語（母方言）を決定する期間をという．この形成期の期間が何歳～何歳までであるかは，一般に 5 歳～12 歳とか，3 歳～15 歳までとする諸説があり，決着をみていない．今回のデータ整備では，これまでの鶴岡調査の考え方・データ整備の方針を踏襲し，言語形成期を 5 歳～13 歳までと考えることにする（報告 5 ：79-80 を参照）．

しかし，人は５歳～13 歳までに一地域で生活を営むとは限らない．５歳～８歳までの４年間は東京に居住し，９歳～13 歳までの５年間は鶴岡に居住するといったケースも考えられる．そこで，かつての鶴岡調査では，５歳～13 歳までの９年間の半数以上（すなわち５年以上）を過ごした地域によって言語形成地を決定した。今回のデータ整備においても，その方針を引き継ぐこととした．

ところで，言語形式期のコード化には，上記の鶴岡調査の方針とは別に，外住歴の観点から整備する方法もある．つまり，言語形成期に鶴岡市に住んだ／住んでいない期間は何年間か，鶴岡市を含む山形県に住んだ／住んでいない期間はどのくらいであるか等の年数をコード化するというものである．この方法は，かつて，国立国語研究所が３回にわたって調査した愛知県岡崎市における岡崎敬語調査（国立国語研究所，1958，1983，2010）で用いた．今回のデータ整備では，この方法によるデータ整備も併せて行っている．

以上をまとめると，言語形式期に関するフィールドは，以下の３つとなる。

言語形成期（５歳～13 歳）に；
（Ⅰ）最も長く（５年以上）居住した地域の地域名　〔＝言語形成期１〕
（Ⅱ）鶴岡市に居住した年数　〔＝言語形成期２〕
（Ⅲ）（鶴岡市を含む）山形県に居住した年数　〔＝言語形成期３〕

（Ⅰ）は各話者の言語形成地を特定するコードであり，（Ⅱ）（Ⅲ）は居住歴を分析に活かすためのコードである．

（4）ホームページによるデータの公開

国立国語研究所・研究情報資料センターでは，近年中にこのデータベースの公開を目指す．データベースを研究者に提供するにあたっては，ホームページを利用することを計画している．ホームページからデータベースをダウンロードするとともに，研究者の操作によって，インタラクティブに結果を表化・グラフ化するプログラムを組み込む可能性を検討している．このプログラムは統計数理研究所の「日本人の国民性調査」をはじめとした経年調査で既に運用されているものであり，実際に鶴岡調査のデータを投入した準備･検討を始めている（http://www.ism.ac.jp/kokuminsei/index.html）．

このプログラムは，メニューバーに性別，年代，第何回調査か等々のボタンを備え，それを選択することで，リアルタイムに表ないしグラフ化するものである．無論，クロス集計にも対応する．ダウンロード版を各研究者がそれぞれの視点でハンドリングして得られるような細かな分析はできないまでも，大まかな結果・傾向をつかむことのできる簡便な方法であり，利便性が高いと考えている．

このようにして作成し，公開を目指して整備しているデータベースの構築にあたっては，調査研究であるが故の様々な課題があった．次節では，その課題と対処の結果を記すこととする．

3. 構築にあたっての課題とその対処

データベース構築にあたって懸案であった課題とその対処について，①プライマリーキーの整備，②継続性のある属性データのコード化の２点から記すこととする．

3.1 プライマリーキーの整備

リレーショナル・データベースにとって，プライマリーキーは最も基本的で，重要なものである．このキーを利用することで，別々に作成された回答データと属性データとをマージすることができるからである．一般的には，連番の ID をプライマリーキーとすることが多いだろう．しかしながら，この鶴岡調査が調査研究であることに由来して，最も基本的であり重要なプライマリーキーの整備といったことから検討する必要があった．

先に述べたように，ランダムサンプリング調査は，住民基本台帳から被調査者を無作為に選び出し，情報（氏名・住所・性・生年）を転記する．この転記した名簿に ID ナンバーをふって実査に臨む．この ID は，単に連番である場合や，居住地区のコードと連番の番号を組み合わせる場合がある．そして，当然のことながら，抽出した被調査者全員が調査に応じてくれるわけではないから，調査ができた被調査者の ID ナンバーは連番にならない．連番になっていなくとも，この ID は当然ユニークなわけだから，リレーショナル・データベースのプライマリーキーとしては何ら問題ない．さて，調査ができた被調査者は，数十年後の次の調査でパネル調査の被調査者となる．このとき，そのパネルサンプルの ID はかつてのそれが利用されるのではなく，実査がしやすいように，改めて ID が与えられることになる．

以上のように，調査実施時の ID が調査毎に異なったとしても，それらをつなぐプライマリーキーとしての，データベース上の ID を付与すればよいだけのことであって，何ら問題はないようにみえる．しかし，これまでの調査研究では，調査時の ID とは異なるデータ整理上の ID が付与されることはなかった．手書きのカードや，大型計算機が利用されていた時代には，データ整理上の ID を付与することは，その整理を煩雑にするだけだったからである．そして，コンピュータが導入されてからも，その「伝統」が引き継がれてきた．恥をさらすようなことかもしれないが，今回のデータベース化で，はじめてデータをデータたらしめるために，プライマリーキーを整備したというのが現状である．

3.2 継続性のある属性データのコード化

属性などのコードが，調査を実施し，データを整備する毎にアドホックに付与され

ることになると，各々の調査結果が比較できなくなる．経年調査にとって，継続性のあるコード化は必須である．それは単にコーディング・マニュアルを書けばよいということではない．そのことを被調査者の職業のコード化を例に説明する．

一般にいう「職業」とは，所属産業と職務内容の両者を指す．鶴岡調査はその両者について質問を設けている．その両者の中でも，鶴岡調査は，研究課題の「共通語化」との関連から，職務内容に重きをおいてきた．例えば，同じ銀行に勤める者でも，窓口での接客に従事する者と，事務的・管理的な職務に従事する者との間に方言使用・共通語使用の度合いに違いがあるか否かを探るため，すなわち，他者との関わりの度合いと共通語化の相関を探るという観点からである．

職務内容を尋ねる質問項目とその集計方法は,過去３回のいずれにも独自性があり，質問の仕方や選択肢（分類枠）が一致しない．職業を取り巻くそれぞれの時代背景を反映しているためである．仮に質問の仕方や選択肢が統一されていたとしても，時代の変遷（産業の変化）によって，その職務自体が無くなったり，これまでにない新しい職務が発生することもある．経年調査である鶴岡調査とって，経年的に，統一された分類でデータを比較するために，各調査の職務内容をどのように分類するかという分類のあり方が課題の１つとなる．

もう１つ課題がある．経年調査は，研究テーマが引き継がれることはもちろんのこと，コード分類・付与などといったデータ整理作業の詳細な情報も引き継がれなければならない．すなわち，後のデータ整理担当者が，同じデータに同じコードを付与できるような，明確で透明性のある「基準」が継承されなければならないのである．それが実現されなければ，調査が実施される度にコードの解読作業が必要になる．これが分類の追試可能性の確保という２つ目の課題である．

今回のデータベース化では，この２つの課題を解決するために，総務省統計局が制定した『日本標準職業分類』（平成９年 12 月改定）を利用することとした．職業分類には，前掲の他に，厚生労働省の『労働省編職業分類』（平成 11 年 11 月改訂），国際労働機関（ILO）の『国際標準職業分類』,「社会階層と社会移動」全国調査（いわゆる SSM 調査）の職業分類などがある。それらの中で『日本標準職業分類』を選んだのは，以下の理由による．

（１）この分類が「統計調査の結果を正確で客観的に示す」ことを目的に作成されたものであり，鶴岡調査の性格・目指すところと一致すること
（２）分類の基準が公開されており，追試可能性が保証されること
（３）産業の変化に伴って分類は改訂されるのであるが，かつての分類枠と新しい分類枠の対照表が明示され，コードの再付与が簡便であること
（４）職務を検索するための Web 上のシステム（政府統計の総合窓口「e-Stat」）およびそれらをデータベース化した Excel ファイルが利用できること

上記のうち，特に（２）（３）によって，前述の２つの課題を解決できると考える．

なお,『日本標準職業分類』は平成９年 12 月改定（平成 14 年６月に一部改定）版に準拠した．

4. おわりに

以上のように，鶴岡調査の過去３回の回答データを整備し，地域社会（＝サンプリング調査）と個人（＝パネル調査）の 40 年間にわたる言語変化をみることのできるデータベースを紹介した．

言語研究にとって望まれることは，このデータベースを利用して数十年間にわたる言語変化をみつめ，優れた理論・モデルの導出されることである．そのためには，このデータベースが広く研究者に提供されなければならない．言語変化についての研究は，その他の言語事象，言語レベルの研究に比して遅れがちであるように映るのは，鶴岡調査のような規模の調査を誰しもができるわけではないことと，その回答データが利用できないことに由来するのであろう．研究の発展・展開のためにも，データの公開にむけた準備を進める．

参考文献

江川　清 (1973). 最近 20 年間の言語生活の変容－鶴岡市における共通語化について－ 言語生活 257，56-63.
井上史雄 (2000). 東北方言の変遷　秋山書店
井上史雄・江川清・佐藤亮一・米田正人 (2009). 音韻共通語化の S 字カーブ－鶴岡・山添 6 回の調査から－　計量国語学，**26**，269-289.
国立国語研究所 (1953). 国立国語研究所報告 5　地域社会の言語生活 鶴岡における実態調査 秀英出版
国立国語研究所 (1958). 国立国語研究所報告 11　敬語と敬語意識.
国立国語研究所 (1974). 国立国語研究所報告 52　地域社会の言語生活 鶴岡における 20 年前との比較　秀英出版
国立国語研究儒 (1983) 国立国語研究所報告 77　敬語と敬語意識－愛知県岡崎市における 20 年前との比較.
国立国語研究所 (1994). 国立国語研究所報告 109-1　鶴岡方言の記述的研究 第 3 次鶴岡調査報告 1　秀英出版
国立国語研究所 (2007). 国立国語研究所報告　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 (2010) 科研費報告　敬語と敬語意識―愛知県岡崎市における第三次調査.