Vol.2010-CH-88 No.7
2010/10/30

方言調査データの XML によるデータベース化

鑓水兼貴[†]　　小西いずみ[††]　　松丸真大[†††]

本発表では，方言調査データ（言語地理学的調査データ）のデータベース化について検討する．調査票の構造を XML 形式で記述し，インフォーマント単位の XML ファイルからなるデータベースを構築することにより,方言調査データが効率的に蓄積され，データ共有や過去の調査資料との経年比較等に有用であることが見込まれる．

Creation of the Dialectological Database Using XML

Kanetaka Yarimizu[†]
Izumi Konishi[††]　Michio Matsumaru[†††]

In this presentation we propose about making the database of dialectological survey data (geolinguistic survey data). By describing the structure of the questionnaire in XML format and creating the dialectological database using XML by each informant, we can accumulate the dialectological data efficiently. It is useful for sharing the past survey data and developing comparison between survey data of various researches.
.

[†] 国立国語研究所
National Institute for Japanese Language and Linguistics
[††] 広島大学
Hiroshima University
[†††] 滋賀大学
Shiga University

1. はじめに

本研究は，方言調査データ，特に言語地理学的調査データのデータベース化について提案を行うものである．調査票を XML 形式で表し，１話者１XML ファイルという形式のデータベースを構築することにより，データ間の比較を容易にし，データの共有をしやすくすることを目的としている．

また本発表のデータベースは，大学共同利用機関法人人間文化研究機構国立国語研究所における基幹型プロジェクト「方言の形成過程解明のための全国方言調査」（以下「全国方言調査」）（リーダー：大西拓一郎）において使用予定であり，今後，過去の方言分布研究資料との対照による，分布変化理論の検証・構築にも貢献するものと考える．

2. 方言調査データの蓄積について

2.1 言語地理学的調査

言語地理学的調査データは，言語地図資料としては大量に蓄積されているものの，データベース化はあまり進んでいない．近年の調査データは基本的に電子化されているが，言語地理学的調査が隆盛だった 1970～80 年の資料は，個人にコンピュータの普及が進んでいなかった（また，性能が低く複雑なデータを扱えなかった）こともあって，電子化は限定的である[1]. 多くのデータが現在も紙の資料のままの状態にあるが，これは言語地理学的調査では，言語地図を最終報告形式とすることが通例であり，データベースの整備が遅れやすかったことも影響している．

ここでは言語地理学的調査について簡単に解説する．言語地理学的調査では，各調査地点から原則１人の「話者」と呼ばれるインフォーマントを選ぶ．地点を代表させるため，その土地で生まれ育った「生え抜き」が選ばれることが通例である．話者は，その地点の代表者として扱われるため，地点情報が非常に重要となる．

調査方式は，面接方式のアンケート調査に近い．質問票にしたがって，方言語形を話者から聞き出していく．

2.2 言語地理学的データ

最低限必要な情報は，(1)話者情報（地点）と，(2)回答情報（質問項目）と単純であり，「地点×質問項目」という地理行列のイメージで捉えられる(表 1)．そして，言語地理学においては，各質問ごとに，回答語形を分類・記号化して地図上の各地点に表示した「言語地図」によって結果を表示するのが普通である．

データはアンケート調査と同様に，調査項目ごとに結果を整理して示す必要がある

ⓒ2010 Information Processing Society of Japan

Vol.2010-CH-88 No.7
2010/10/30

ため，言語地図においても，地図ごとに回答をまとめて示すことが多い（基本的には質問項目単位だが，整理する観点によっては複数の項目を１つの地図で表わしたり，逆に１つの項目を複数の地図で表わすことはある）．

	問 1	**問 2**	**問 3**	**・・・**	**問 m**
話者 1	語形	語形	語形	・・・	語形
話者 2	語形	語形	語形	・・・	語形
：	：	：	：	：	：
話者 n	語形	語形	語形	・・・	語形

表 1　地理行列によるデータのイメージ

また，多くのアンケート調査と同様に，１つの質問項目は１つの回答には収まらず，複数回答や注が発生するため，それらの処理が問題となる．

こうした問題に対処するため，現在公開されている国立国語研究所編『方言文法全国地図(GAJ)』[2]のデータベース[3]は，１回答を１行としたデータ形式を取っている．

この形式の場合，回答をどう分類したかといった情報や，回答ごとの注記などを表すことができる(表 2)．図 1 は，GAJ データベースの一部である．TAB 区切りテキスト形式となっている．

問	**話者**	**調査時の語形**	**整理後の語形**	**・・・**	**注記**
問 1	話者 1	語形	語形	・・・	注記
問 1	話者 1	語形	語形	・・・	注記
問 1	話者 2	語形	語形	・・・	注記
：	：	：	：	：	：
問 1	話者 n	語形	語形	・・・	注記
問 2	話者 1	語形	語形	・・・	注記
：	：	：	：	：	：
問 m	話者 n	語形	語形	・・・	注記

表 2　１回答１行形式によるデータのイメージ

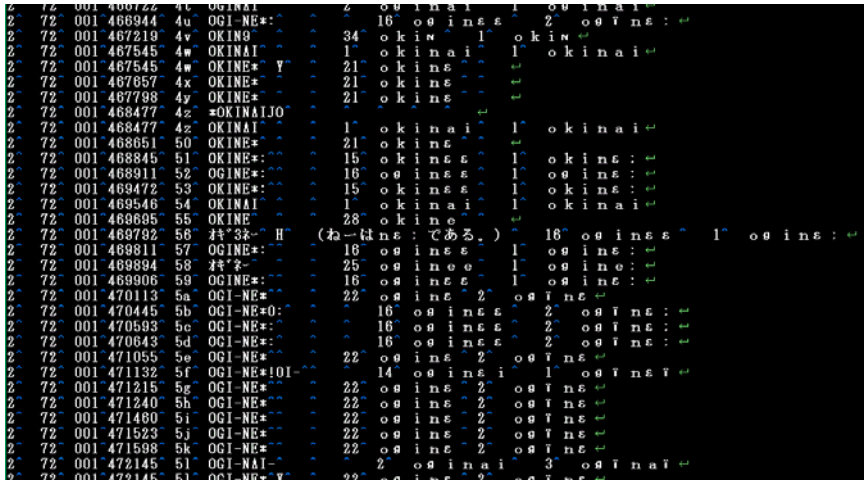


図 1　『方言文法全国地図』のデーターベース

2.3 比較に適したデータベース

異なる調査の結果同士の比較は慎重に行われなければならない．一見同じ項目のように見えても，質問の仕方は同じかどうか，回答形式は自由回答か選択式か，といった項目の問題から，その調査が誰によってどのような目的で行われたのか，また，話者はどのような属性を持っているのか，などといった調査環境や話者の問題まで，さまざまな比較すべき要素が含まれている．そのため，ある調査のある項目だけを取り上げて，そのまま結果同士を比較するのは，いささか乱暴であろう．

利用する側には，データは質問項目単位で整理されているほうが便利であろう．しかし実際の調査においては，１人の調査者が話者全員に特定の項目を質問する，といったことは不可能であり，調査は話者単位で行われる．すなわち同一項目でも，調査環境が異なっている可能性は排除できない．

また，本発表のデータベースのフォーマットを用いる予定の「全国方言調査」は，調査期間が長期にわたるため，調査期間中でもデータベースを利用できるようにすることが望まれる．話者のデータが順次追加されていく場合，項目単位でのデータベースでは全項目が追加対象となってしまい，処理が煩雑になる恐れがある．

以上から，話者単位でのデータベースを設計する必要があると考える．データの追加がしやすいだけでなく，なるべく調査時の調査票をそのままデータ化することで，調査時の状況を再現しやすくなり，分析においても適していると思われる．

ⓒ2010 Information Processing Society of Japan

Vol.2010-CH-88 No.7
2010/10/30

3. 話者単位でのデータベース

3.1 XML の採用

話者単位でのデータベースを作成するためは，実質的に調査票をそのままデータベース化できることが望ましい. そのため,「１調査票が１ファイル」となる形式とした.

ファイル形式は，構造化テキストである XML を採用した．ただし表形式のデータと比較した場合, XML 形式では, そのままでは利用できない研究者も多いと思われる. そのため, データ利用に際しては表形式への変換等のツールが不可欠になってしまう.

それでもなお表形式よりも有利と考えるのは，他の言語地理学的データと組み合わせて比較・分析を行う場合に適しているからである．表形式の場合でもデータフォーマットを規定すれば XML である必要はないが，多くの調査の場合，質問項目や内容は異なっており，共通する項目を選び出す作業は複雑になってしまう．また，共通するフォーマットを作成しようとしても空欄の多いものになり，使いにくいものになるだろう.

調査によって調査項目が大きく異なるデータの場合には，データの内部構造の規定が明示的である XML 形式は有利であると思われる．このことは他の方言調査資料のデータベース化においても指針となりうると考えられる.

3.2 データ構成

一つの調査データは，以下のような構成となる．調査全体の情報を示す 1 つのプロジェクトファイルと，調査地点分のデータファイルから構成される（図 2）. 以下，プロジェクトファイルとデータファイルの構成について述べる.

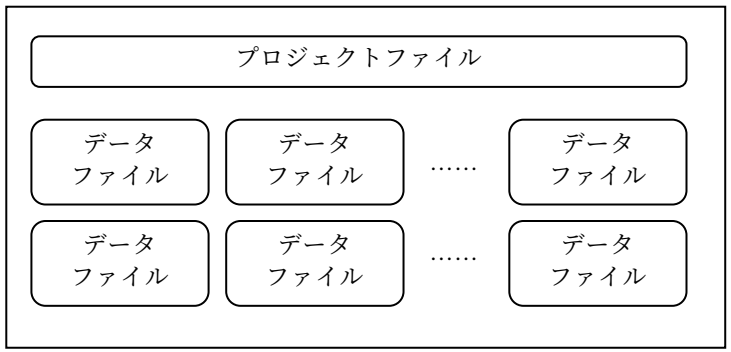


図 2　話者単位のデータベースの全体像

3.2.1 プロジェクトファイル

調査同士を比較する上で，調査全体に関する記録をしておく必要がある．これは単純に比較や集計などができる資料ではないが，どのような調査なのかを記録しておくことで，研究者が比較資料として採用するか否かの基準となる重要なものである.

こうした情報が入った調査情報全体を，ここでは「プロジェクト」とし，以下の３つの情報が入る．これらは XML 情報によって記述される.

①プロジェクト情報

調査の趣旨･調査概要などを記録するものである.「話者の選定基準」,「地点の精度」,「語形の採用原則」などの情報が入る.

②調査者情報

調査者に関する情報が入る．ただし個人情報保護を考慮したデータ化をしなければならない.

③調査票情報

調査票の情報，すなわち，実際の質問文や，調査時に提示する図などが入る.

3.2.2 データファイル

データファイルは，調査のデータ部分である．以下の３つの部分からなる.

①調査情報
②話者情報

調査におけるフェイスシートのデータが入る．そのため，個人情報保護を十分に考慮する必要がある．フェイスシートの項目の中には，いつ，誰が，どのような環境で調査したのか，といった調査そのものに関する情報も入っている．これらは XML タグでは「調査情報」として独立させ, その他の情報については話者情報として,「話者」タグの中に入れる.

③回答情報

各質問の回答は「調査データ」タグの中に入る．質問文は各個人のデータの中に入れるとサイズが大きくなるため，プロジェクトファイルの中の調査票情報にリンクさせる予定である.

以下, XML のデータ構造について例を示す（図 3）. タグ名は現時点では仮である.

ⓒ2010 Information Processing Society of Japan

```
<?xml version="1.0" encoding="UTF-8"?>
<?xml-stylesheet type="text/xsl" href="chosahyo.xsl"?>
<調査票>
    <調査情報>
        <調査者コード>○○○○</調査者コード>
        <調査開始日>○○○○年○○月○○日</調査開始日>
        <所用時間>○○時間○○分</所用時間>
        <同席者>○○○○</同席者>
        <備考>○○○○</備考>
    </調査情報>

    <話者>
        <話者コード>○○○○</話者コード>
        <話者生年>○○○○</話者生年>
        <話者生育地>○○○○</話者生育地>
        <話者最長居住地>○○○○</話者最長居住地>
        <備考>○○○○○○○○</備考>
    </話者>

    <調査データ>
        <質問項目>
                <質問番号>○○○</質問番号>
                <質問項目名>○○○<質問項目名>
                <質問文>(リンクによって示す)</質問文>

                <回答>
                        <語形>回答語形１</語形>
                        <注記 種類="話者">語形注記</注記>
                </回答>

                <回答>
                        <語形>回答語形２</語形>
                        <注記 種類="話者">語形注記</注記>
                </回答>

                <修正 種類="削除" 修正日時="○○○" 修正者="○○">
                        <回答>
                                <語形>回答語形３</語形>
                                <注記 種類="話者">○○○</注記>
                        </回答>
                </修正>

                <注記 種類="話者">
                        ○○○○○○○○
                </注記>

        </質問項目>

        <質問項目>
                :
              (略)
                :
        </質問番号>

        <注記 種類="話者">
                ○○○○○○○○
        </注記>

    </調査データ>

</調査票>
```

図 3　データファイルの例

ⓒ2010 Information Processing Society of Japan

3.3 データの注記

調査において得られる情報は回答ばかりではない．回答に対する注記，項目に対する注記，調査全体に関わる特記事項などといった，さまざまな補足情報が得られる．これらの情報は，どのレベルの情報かが階層化されているため，XML での表記に適している．これらの情報は「注記」タグで表され，言及した人が話者か調査者か同席者か，といった情報はタグ属性で分類される．

3.4 データの修正

前述のとおり，「全国方言調査」は，調査中から利用可能とするため，バージョン改訂ごとに従来のデータの部分にも修正が行われる可能性がある．そうした場合，データ利用者が混乱しないよう，リリース後のデータの修正はすべて「修正」タグによって行い，過去の情報は削除しないことにする．

「修正」タグは，「削除」と「挿入」からなり，タグ属性によって決まる．また，「修正日時」と「修正者」もタグ属性によって指定され，利用者がどの時点のデータを利用したかがわかるようになっている．

4. データベースの利用

XML データは，スタイルファイルによって調査票に近いの状態での閲覧が可能となる．しかし，この状態では不親切であり，多くの利用者がただデータを閲覧できるだけで，分析することが困難なままであると思われる．そのため，サーバー上に格納したデータを，利用者の望む形式に変換して出力するためのツールを提供することで，データベースの利用を促進する必要がある．

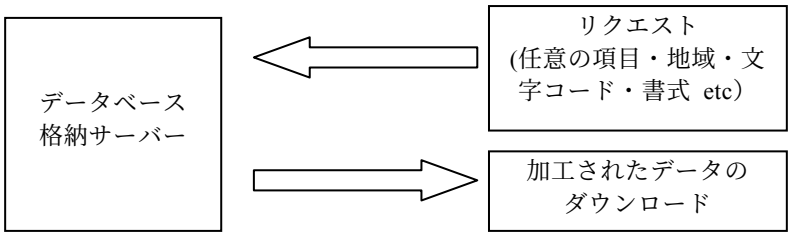


図 4　データベースの利用イメージ(1)

データは調査ごとに１つのフォルダにまとめられるが，フォルダ内は話者単位でデータ化されている．そのため複数の調査データを統合するには，データファイルを統合するだけでよい．

たとえば，調査 A のある地域と，調査 B のある地域のデータだけを取り出して新しいデータベースを作成するには，図 5 のように必要な地点の XML ファイルを取り出せばよい．両者の調査に共通する「調査項目名」があれば，データファイルの「調査項目名」タグのある項目を取り出すだけでデータの比較が可能となる．

ただし，調査項目の質問の仕方が同じかどうかといった情報は，調査 A と調査 B のプロジェクトファイルを比較して確認しなければならない．

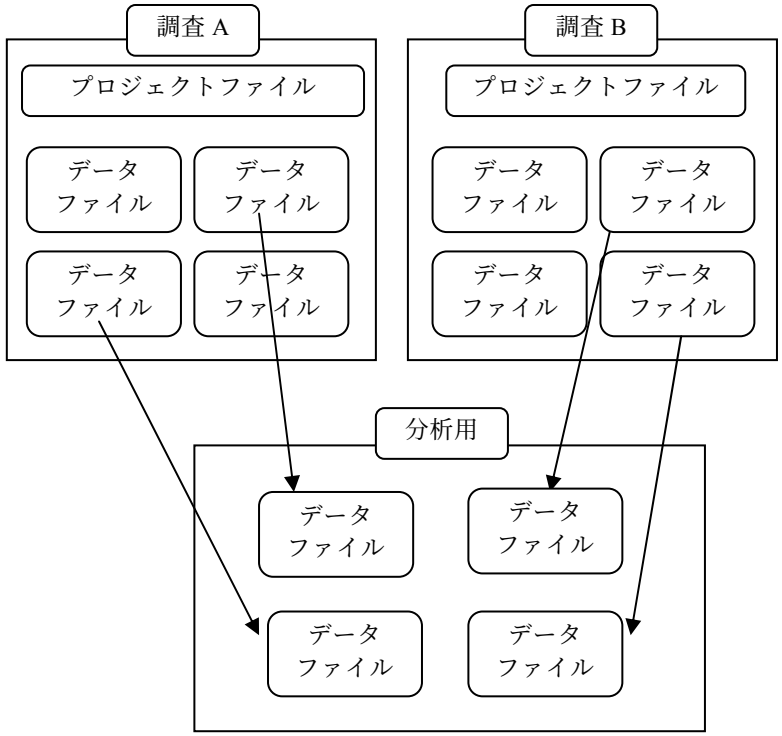


図 5　データベースの利用イメージ(2)

ⓒ2010 Information Processing Society of Japan

Vol.2010-CH-88 No.7
2010/10/30

また，データの加工というよりも，応用的なことであるが，言語地図作成や計量的分析を支援するツール群の提供も重要と思われる．

方言研究においては，言語地図は分析には必要不可欠なものであるものの，コンピュータを用いた地図作成方法は，近年ようやく[4]など平易な解説書が出てきている状態で，今だ普及過程にある状況といえる．また，既存の方言データベースの利用した計量的分析についてもあまり進んでいない状況である[5].

せっかくデータベースを提供しても，考察の手助けになるための地図作成や計量的分析の段階でつまづいてしまう恐れがある．そのため，データベースの利用においては，データの加工といった基本的な部分のみならず，応用的な部分までを「データベースの利用」として考慮にいれておく必要があるだろう．

5. おわりに

以上，言語地理学的データベースに，XML を用いた共通のフォーマットを用いることによる利点について考察した．

全国の貴重な方言調査資料の共有化を促進することによって，時間的変化・地理的変異を動的に分析するためのデータ整備がなされることが期待される．また，複数の調査を比較するような言語地図の作成や，計量的分析といった研究も促進されることが望まれる．これらは「全国方言調査」における「方言の分布形成の理論構築」という目的にも一致し，今後の方言研究に大きく貢献するであろう．

今後は，「全国方言調査」の準備調査データや，すでに公開されている「方言文法全国地図」のデータなどを本発表における XML 形式によってデータベース化し，2010 年 10 月より始まる「全国方言調査」のデータと組み合わせていく予定である．

そしてデータベースの利用の過程で出た問題などを基に，さらに XML のフォーマットの再検討や改訂を行っていきたいと思う．

参考文献

1) 福嶋 秩子, 方言地図作成の機械化, 方言地理学の課題(明治書院), pp.418-431, (2002).
2) 国立国語研究所, 方言文法全国地図 1-6,国立印刷局 (1989-2006).
3) 国立国語研究所, 『方言文法全国地図』全データ,
http://www2.ninjal.ac.jp/hogen/dp/gaj_dp_i/gaj_all/gaj_all.html.
4) 中井 精一, 社会言語学の調査と研究の技法―フィールドワークとデータ整理の基本, おうふう (2005).
5) 沢木 幹栄, 方言地図データの活用―GAJ のデータによる地点のクラスター分析, 方言地理学の課題(明治書院), pp.432-444, (2002)

ⓒ2010 Information Processing Society of Japan