Web 版「全国方言文法辞典」の構築に向けて
ー調査データの報告システム開発についてー

林　良雄・日高　水穂

１．背景

現代語の文法研究の成果を日本語諸方言の対照研究に応用する試みは幾つかの研究がおこなわれており，一連の成果を上げている（沼田・野田編 2003 など）。これらの研究は現代語（標準語）に限定せず地理的変異（方言差）および歴史的変異（時代差）を考慮に入れた日本語文法の全体像を把握することを目的としている。そのため，大量の調査データを必要とする。

これまでの方言研究の中でも膨大な調査データの蓄積があるが，必ずしも文法項目を網羅するものではなく，文法の全体像を把握するためには不十分である。また，誰にでも容易に活用できる形になっているわけではなく，他分野での利用が十分になされているとは言えない。

そこで，著者の一人である日高が中心となって「方言文法研究会」を 2001 年に立ち上げ，日本語諸方言の文法を総合的に記述する『全国方言文法辞典』の編纂を目的に，既存の方言文献資料をデータベース化する作業および，現地調査による方言文法の記述的研究を進めている。例えば『日本方言大辞典』（徳川宗賢監修，小学館）から助詞，助動詞，接辞類の記述を抜粋しデータベース化する作業や各種の方言談話資料類から用例をピックアップする作業を行った。また，現地調査のための統一的な調査項目の選定を行い，一部の項目については，すでにその調査報告を公開している（方言文法研究会編 2010）。

２．本研究の目的

前述の方言文法研究会では，現在，記述の出発点を助詞類に定め，共通調査項目の作成と要地方言の選定を進めている。この部分が確定すれば複数の調査員が担当の地域の調査を行い，データを蓄積することになる。最終的にはデータをまとめて『全国方言文法辞典《助詞編》』の冊子版の刊行及びウェッブ版の公開を行うことを目標としている。この研究では異なる調査員が行った現地調査の結果を利用して，辞典項目の記述を行う。したがって，調査データは，誰でも利用できる形，すなわちデータベース化する必要がある。通常の調査研究において調査データはエクセルなどの個人的なデータファイルとして記録され，それを交換することによりまとめていく。最終的な結果は論文や報告書などの紙ベースで出され，電子化されたものは出されない，あるいは単純なＷｅｂのページとして公開

されることとなる。

そうすると他の研究者の利用は紙ベースのものを再度コンピュータに入力し直すか，Ｗｅｂのページから必要な部分をコピーすることによって利用することになり，誰にでも容易にということにはならない。また，網羅的な研究を行うことも困難となる。

そこで，この研究ではＷｅｂを利用した調査報告システムを構築することとした。各地で調査を行った調査員がＷｅｂサーバーにアクセスし，調査結果報告のフォームに調査結果を入力する。入力されたデータは即時にデータベース化されていくシステムである。

これには次のような利点がある。

(1)入力した時点で電子データとして扱うことができる。

(2)どの調査地点からでも即時に報告が行え，従って報告の為にどこかで会議を開く必要がない。

(3)調査報告がどのようになっているかを逐次全員見ることができでるようにすることにより，調査に参加している者全員がデータを共有できる。

(4)入力時の形式を統一することにより，記号や記述の形式を統一することができる。

このシステムで調査結果を蓄積したデータベースからデータを抽出することにより辞典の編纂およびＷｅｂの公開を行うことが最終目標である。

３．現状とシステムの要件について

本研究は科研費（基盤研究(B)「日本語諸方言の文法を総合的に記述する『全国方言文法辞典』の作成とウェブ版の構築」課題番号 21320086）の配分を受けて 2009 年度から 5 カ年で実施することになっている。初年度である 2009 年度は，Ｗｅｂのシステム構築を行った。既に入力が必要なデータ項目の洗い出しをほぼ終わり，評価用のデータベースと調査報告を入力するためのシステムの試作を行っている。以下にその概要を示す。

本研究で行う方言調査の場合，直接話者から聞き取りを行う。従って表１のような基本情報が必要となる。

現在準備している共通調査項目は，条件表現と逆接表現（ケレドモ・ノニ類，テモ類）である。条件表現は 68 項目，逆接表現はケレドモ・ノニ類 25 項目，テモ類 46 項目の調査例文が設定されている。話者情報を含む基本情報を入力したのち，これらの調査したデータの入力を行うことになる。調査員が担当する話者の情報のみを入力，確認するために図１のようにその調査員が入力した話者のみのリストから選択するようにしている。

一つの調査項目については調査例文に対する方言訳と文法性判断（○：自然，×：不自然，？：やや不自然），それに対する注記を 1 セットのデータとしている。ただし，一つの例文に対して複数の言い回しが存在する場合があるので 5 セットを一つの例文に用意する（図 2）。

現在，この調査報告を行うまでのテストシステムを構築して，実際に報告する調査員に試用してもらい調整を行っている。

表１．基本データ項目

	データ項目	説　明
地点情報	地点名	地点の名称
	調査地点	都道府県，市区町村は必須
	地点概要	自由記述
話者情報	話者氏名	必須
	話者性別	必須
	話者生年月日	西暦年は必須
	調査時満年齢	必須
	現住所	都道府県，市区町村は必須
	電話番号	省略可
	出身地※	現住所と異なる場合
	外往歴	出身地と異なる地域に居住した経験がある場合
	その他	自由記述
調査概要	調査者	ログイン情報から
	同席者	調査者以外に調査に立ち会った人
	調査場所	調査を行った場所
	調査日時	
	その他	自由記述

※言語形成期（6～12 歳）に主に過ごした土地。

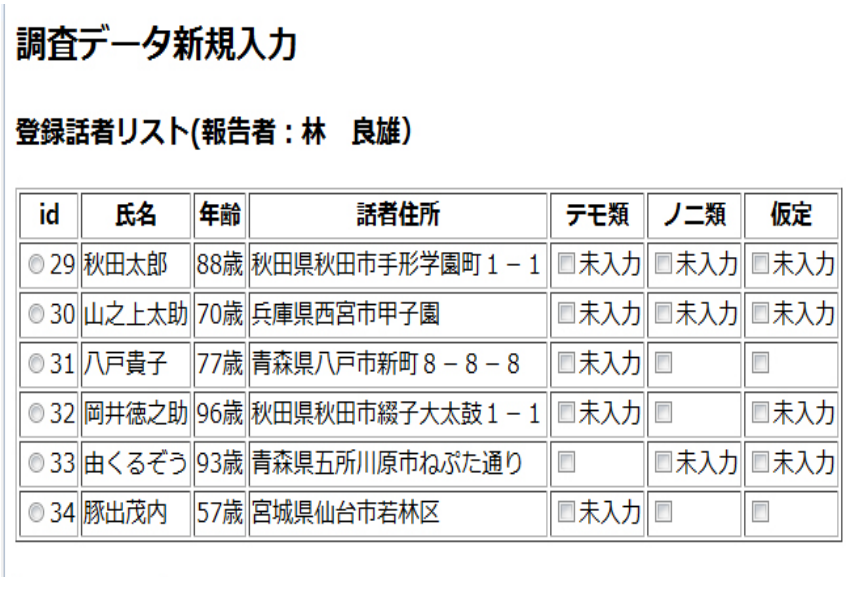

図１．登録話者リストの画面

図 2. 調査報告入力画面

４．検索機能

データの入力作業が進み，データが蓄積されれば次にそのデータをどのように活用するかが問題となる。その活用法についてはこれからの議論になるが，検索する機能は不可欠と思われる。そこで，その検討を行うために検索機能を試作した。

検索する項目として様々なものが考えられるが，利用が多いのは調査地点と調査結果の語句の一部による絞り込みであると考えられる。調査結果の語句については多数のカテゴリと質問項目があるが，それらの質問項目全てを一度に対象にすることはなく，一つの質問項目に絞って検索するのが一般的であろう。そこで，この二項目についてフリーワードで検索ができるようにした（図 3）。

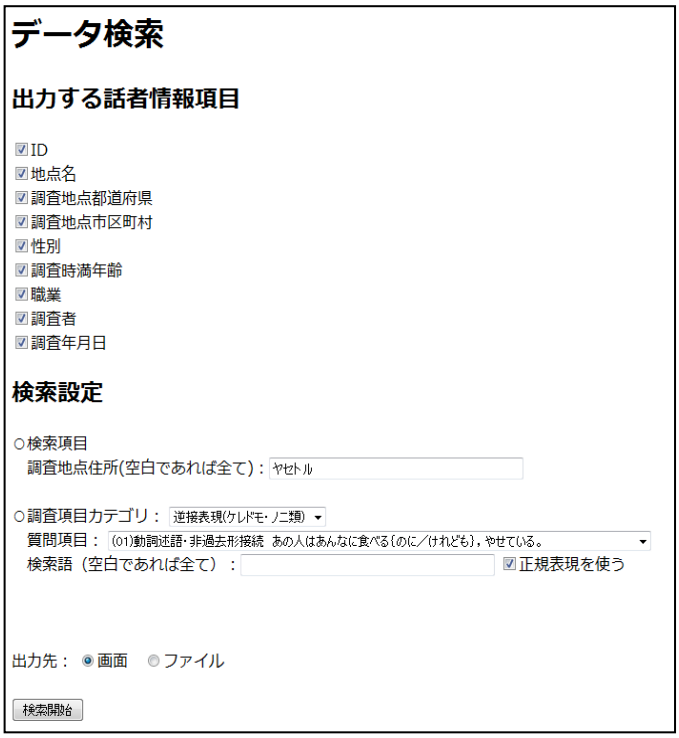

図 3. 検索メニュー画面

フリーワードは検索語が入っている例文を全て抽出するが，その際，検索語の前後の文字列との関係は考慮していない。しかし，検索語のあとに特定の文字数が入っているものが必要な場合もあろう。また，検索語が文の途中にあるのか文末にあるのかなど検索語の位置関係をみなければならないこともある。その際には正規表現と呼ばれる，文字列の集合を一つの文字列で表現する方法を使う。例えば＊は 0 文字以上の文字列を表す。“正規表現”のボックスにチェックを入れることにより，この正規表現が使えるようになっている。

出力するデータについては ID，地点名，調査地点都道府県，調査地点市区町村，（話者）性別，（話者）調査時満年齢，（話者）職業，調査者，調査年月日の基本情報の他，質問項目の例文を出力する。この例文は一つの質問項目について最大 5 つある。このようにかなり多数の情報を出力する必要があるので，画面にすべて出力すると見づらくなる。また検索する際にはデータを利用することがメインとなるので画面に出力するよりデータファイルとして出力する用途のほうが多いと思われる。

そこで出力は画面とファイルどちらかを選択できるようにし，画面出力については例文のみを表示し，ファイルには例文に付属する情報，注記までデータが出力されるようにしている。

この試作版についてはまだデザインや検索項目に関する検討が十分ではなく，改善の余地がある。また，このプロジェクトがどのような形でデータを利用するかにも依存するので，検索機能の検討を引き続きおこなっていく。

５．今後の予定

2010 年度から実際の調査に入ることになっているので，このシステムの本格的な利用は当該年度からとなる。この段階では情報の共有を行っていく必要がある。従って調査報告のデータが蓄積されたデータベースを検索するシステムを早期に構築する必要がある。また，新しい調査項目も追加されることになっているので，報告のシステムも更新していくことになる。

最終的には調査結果の公開を目指す。どのように公開するかは今後の研究会の議論となるところであるが，話者の音声情報を入れたものも含めて公開し，全国方言文法辞典のＷｅｂ版とすることが考えられている。

付記　本稿は，情報処理学会第 72 回全国大会（2010 年 3 月 9-11 日・東京大学）の発表予稿集原稿に加筆・修正を施したものである。

参考文献

沼田善子・野田尚史編（2003）『日本語のとりたて―現代語と歴史的変化・地理的変異―』くろしお出版

方言文法研究会編（2010）『全国方言文法辞典資料集(1)原因・理由表現』科学研究費補助金研究成果報告書　〈方言文法研究会編「全国方言文法辞典データベース(Web版)」http://hougen.sakura.ne.jp/〉