『大阪樟蔭女子大学日本語研究センター報告』第9号（93～102頁、図の一部省略）

全国方言談話資料データベースの作成に向けて

国立国語研究所　井上文子
大阪樟蔭女子大学　田原広史

1．方言談話資料の現状

伝統的地域方言が全国的に消滅の危機に瀕している現在、その記録・保存の緊急性・重要性が強く認識されている。特に、自然な形に近い、まとまりをもった言語使用の実態を表すものとして、「方言談話資料」が各方面から注目を集めている。

現在、方言音声に加え、文字化・標準語訳を備えた全国規模の方言談話資料には、次の二つがある。一つは、『全国方言資料』（日本放送出版協会、1959-72 年、1966-72 年ソノシート付、1981 年カセットテープ付、1999 年 CD-ROM 版）であり、もう一つは『国立国語研究所資料集 10　方言談話資料』全 10 巻 カセットテープ付（秀英出版、1978-87 年）である。これら二つの資料については、収録量の点、あるいは地点密度の点で不満であり、さらに大量の談話データを望む声が聞かれる。また、方言談話テキストが電子化されていて、利用者個人で検索や加工が可能な形態であることを求める声も強い。

ここに、もう一つ、大規模に録音・収集されたが、資料の整備等の問題で公開されず、眠っている大量の録音テープ・文字化原稿が存在する。ここで紹介する文化庁**「各地方言収集緊急調査」**報告資料である。

2．「各地方言収集緊急調査」

1977-1985（昭和 52-60）年度に、文化庁によって「各地方言収集緊急調査」と称した方言談話収録事業がおこなわれた。目的は、「全国的に急速に変化し、失われつつある各地の方言を各都道府県において、緊急に調査し、これを記録・保存する」ことであった。実施概要は次の通りである。

①調査地点

調査地点は、各都道府県について 5 地点程度を選定するが、文化庁および地元方言研究者の意見を聞いて、各都道府県教育委員会が決定する。方言区画上、複数の区域に分かれる場合は、方言の状況が概観できるように、各区域から収録地点を選ぶ。特に、離島など、特色の認められる方言は可能な限り収録する。

②調査方法

各都道府県は、言語学、国語学、方言学の専門家から調査員を選出し、調査地点や具体的な調査方法について検討し、調査を進める。

③調査内容（一地点につき10時間程度）

1.老年男女の対話(2 時間)　　5.老年男性(目上と目下)の対話(2 時間)
2.老年男性の対話(1 時間)　　6.場面設定の対話(演技的対話をさせる)(1 時間)
3.老年女性の対話(1 時間)　　7.当該地域の民話(無理な場合は自由会話)(1 時間)
4.老年と若年の対話(2 時間)　　※ 1- 4 は 3 人の会話も可とする

④話者

その土地で生まれ育ち、よその土地に住んだことのない人（3 年以内は可とする）。年齢は原則として、老年層は収録時において 60 歳以上、若年層は 20-30 歳代とする。

⑤録音

後世に残すために、自然な方言会話を良い音質で録音することに留意する。

⑥文字化

一地点 10 時間程度の録音のうち、有効かつ適切な部分 3 時間程度を文字化する。表記は、原則としてカタカナ書きとし、方言の音声的特徴をある程度表し得るよう工夫する。対応する標準語訳をつけるとともに、場面、文脈、特徴的音声、方言形の語義・用法などについての注釈を付ける。

上記要領で、文化庁は全国の都道府県教育委員会に各地方言の収集を指示した。47 都道府県は、開始年度により、第 1 次から第 7 次に分けられ、それぞれ 3 年計画で収録を行った。収録にあたり、各教育委員会と連携して、全国各地の方言研究者が全面的に協力した結果、地域的密度、収録量、方言的内容のいずれの面からも、それまでに類を見ない高レベルのデータが得られた。

調査終了後、各教育委員会から、方言談話録音テープ及び文字化原稿が、「各地方言収集緊急調査」報告として、文化庁に提出され、永久保存されることとなった。しかし、これらの資料は、現在に至るまで、ほとんど公表されることがなかった。いくつかの教育委員会では、一部のデータを用いて独自に報告書を刊行しているが、市販されておらず入手しにくい。また、これらの報告書はすべて冊子体の印刷物であり、電子化テキストによるものはなく、録音テープを添付しているものも少数である。

その後、「各地方言収集緊急調査」報告資料は、文化庁から国立国語研究所に移管された。国立国語研究所では、これら大量の録音テープ・文字化原稿の有効利用を考え、資料を整備した上で、「全国方言談話資料データベース」として一般に公開する計画を開始した。

具体的には、平成 9 年度より佐藤亮一（東京女子大学）を代表者とし、文部省科学研究費補助金（研究成果公開促進費）<いわゆるデータベース科研>を受け、文字化資料の電子テキスト化を中心に作業をおこなってきた。平成 12 年度、すなわち本年度より音声データの編集についても取り組んでいる。

3.「全国方言談話資料データベース」の作成

3.1　データ量などの問題

「各地方言収集緊急調査」で収集された録音資料は、一地点 10 時間として約 200 地点であることから単純に計算すると 2000 時間になる。これがどんな量かは想像しにくいが、一通り聞くには、一日 24 時間ずっと流しっぱなしにして 83 日かかると言えば分かるだろうか。また、そのまま CD にすれば 1600 枚になると言えば分かるだろうか。

これらのうち、文字化されている部分を対象にするとしても、一地点 3 時間であるから 600 時間（25 日分、あるいは CD500 枚分）もあるのである。これだけのものを編集し、データベースにすることは不可能である。したがって、データベース化し、公開できる部分というのは全体のごくごく一部ということになる。

作成の目安となるのは、上でふれた『全国方言資料』（NHK）である。CD-ROM 版から収録時間を推測してみると次のようになる。

全データ量：約 2300MB（CD-ROM11 枚に格納<一枚あたり 200MB 程度）
音　　質：8bit 量子化・22.05KHz サンプリング・モノラルの WAVE 形式
　　量的には CD（16bit・44.1KHz・ステレオ）の 8 分の 1
　　1 秒あたり 22,050 バイト（フロッピー 1 枚に約 68 秒入る）
収録時間：約 1800 分（約 30 時間）… 2300MB ÷ 22,050B で秒数が求められる
地 点 数：141 地点（1 県あたり 3 地点）
表示形式：PDF 形式の文字化画像ファイルに音声ファイルをリンクしたもの

NHK のものについては、従来のカセット版の文字化資料をスキャナで画像として読み込み、PDF ファイル（Portable Document Format、Adbe 社が開発）に変換したものに、ページごとに編集した音声ファイルをリンクしておき、画面上のアイコンをクリックすることにより、そのページの音声を再生するというものである。この場合、音声の編集については、とにかくページごとに機械的に区切って音声ファイルとして保存していくので、それほどの手間はかからない。文字化部分も、もとの冊子をスキャナで読みとって画像ファイルとして保存するので簡単である。

この方法は､文字化付き音声資料の公開手段の一つの方向を示していると言えるだろう。利点は、誰にでも手軽に画面を見ながら再生できるということである。音声アイコンをクリックすれば、画面に表示された文字化資料の音声が流れるという単純な構造だからである。欠点は、特定の話者や、特定の表現が現れるものを検索すると言ったいわゆるデータベースとしての使い方ができないことである。音声に関しては、検索して聞き比べるといった必要性は少ないかもしれないが、少なくとも文字化部分については、このような操作ができないとデータベースとは言えない。

以上を考慮すると、今回のデータから作成するデータベースについては、実現可能な規模としては、おそらく一県一時間として 50 時間分くらいであろう。上記 NHK と同じ音声形式にすれば、CD-ROM 一枚に最大 8 時間納めることができる。そうすると、全体で CD-ROM10 枚程度でおさまるだろう。まず、老年層男女の対話から、1 地点 30 分を目安として選定をおこなっている。

3.2文字化部分の作業

文字化部分については、とりあえず、元の手書き文字化原稿を電子化することから始めなければならない。音声データベースとは言っても、文字化部分ならびに標準語訳がデータベースの中心となるべきであり、音声はあくまでも文字化データの付属部分ととらえた方がよい。今回の作業では、表記・形式を統一する以外はできるだけ原本を忠実に電子テキスト化している。文字化部分と標準語訳を別に扱う必要があるため、エクセルでカラムを分けて入力し、最終的に PDF 化する場合には Word を用いて作成することにした。

3.3音声部分の作業

音声については、オリジナルテープはカセットテープなので、まず、パソコンに取り込む、すなわちディジタル化することから始めなければならない。最近のパソコンは音声インターフェイスがかなり改善されており、パソコンに取り込む際の音声劣化が少なくなってきた。今回の作業の場合、特殊なインターフェイスを用いることも考えたが、最終的にソニー社製の VAIO 付属のギガポケットというビデオインターフェイスを使うことにした。

この装置はビデオを直接ハードディスクに録画するシステムで、操作するアプリケーションソフトも付属している。20GB のハードディスクに 10 時間もの画像が連続して録画できるという驚異的なものである。このシステムの音声部分を利用して編集が手軽におこなえることが分かったので利用することにした。

まず、一まとまりの談話（15 分～ 30 分程度）をカセットデッキからパソコン内に一気に録画（録音）する。この時、空いている画面部分に、ビデオ端子経由で CCD カメラを使って文字化資料を同時に録画しながら取り込みをおこなうと、その後の編集時に音声の特定がしやすい。

いったん取り込んだ後、パソコン内で再生しながら、1 ～ 2 分程度を目安に必要箇所をパソコン内で再録音する。この時にサンプリングレート、音声ファイル形式等の調整をおこなう。それをさらに文字化資料を基準に発話単位に区切り、バラバラのファイルに保存していくわけである。作業としては三段階を踏むことになるが、ディジタル化した後の、パソコン内での作業は、音質の劣化がないので何度でも編集がおこなえる。

3.4データベース化の方法

電子テキスト化した文字化資料と、ディジタル化して編集した音声ファイルとをつなぐことが、データベース化への第一歩であるが、現在筆者らが考えている方法には、次の二つがある。

1) ワープロ・エクセル等で文字化資料を作成し、それを PDF ファイル化した後、音声ファイルをリンクさせる。
2) データベースソフトを用いて、文字化資料と音声データが一体化したデータベースを作成する。

1)は NHK 方式であるが、文字化資料は画像データではなく、電子テキストから PDF 化するので、文字単位の検索は可能であるが、検索結果のみを表示・印刷することはできない。利点としては、リンクを張っているので、元の音声ファイル自体を単独でコピーしたり、分析・加工が可能であることがあげられる。

2)は田原が従来よりとっているファイルメーカーを使う方法である。この方法だと、検索が自由に出来、検索結果を音声で確認できるというメリットがある。しかし、ファイルメーカーは音声ファイルをリンクできず、すべてをデータベースファイル内に抱え込んでしまうので、それぞれの音声を単体として扱うことができない。デベロッパーエディションにより、ファイルメーカーを持っていない人でも使ってもらえるが、将来的なバージョンアップやファイルメーカー社の倒産といったことを考えると不安である。

現時点では、1)すなわち素材に近い形を提供し、2)のデータベース化については、利用者が自分の使い勝手の良い方法で加工するのが望ましいと考えている。ただし、今回の場

合、文字化資料の部分については、量的に問題がないので、テキスト形式に加えて、いくつかの形式、たとえばワード・エクセル・アクセス・ファイルメーカーなどの形式で提供するのがよいだろう。現在では、これらすべてのアプリケーションにおいて、音声を組み込むことが可能であるので、ユーザ側で必要に応じて加工することができる。

4. 今後の予定および課題

音声の編集については、これから開始するが、大量の音声資料を効率良く編集していく必要がある。自然談話は複数の人間が同時に発話していたり、声の大きさがバラバラであったりして編集作業がおこないにくい。特に今回のように発話単位で編集するとなると、NHK1999 のように機械的に切るだけでは済まず、必要箇所については重ねて前後の発話に含めるといったことも必要になるだろう。

このような作業をおこなうにあたっては、作成にかかる手間および経費と、できあがったものの学術的価値とを天秤にかけて判断することが必要と言える。常にどのような利用のされ方をするのかといったことを意識しつつ作業をおこなっていくことが、よいデータベースを作成するコツであろう。

現在の計画では、2004（平成 16）年すなわち 3 〜 4 年の間にすべての県をデータベース化することになっている。3 〜 4 年経てば、われわれを取り巻くパソコン環境がかなり変わっているかも知れず、かなり不安であるが、その時々の技術の進歩を取り入れつつ、作業を進めていくことになるだろう。

付記

この報告は、2000 年 7 月 9 日に大阪樟蔭女子大学で開催された第 16 回西日本国語国文学データベース研究会において、口頭発表した際のハンドアウトをそのまま掲載したものである。資料の最終 2 ページは、発表時にスクリーン提示したパワーポイント画面をプリントアウトしたものである。

全国方言談話資料データベースの作成に向けて

Database of Discourse in Japanese Dialect

井上文子（国立国語研究所）
田原広史（大阪樟蔭女子大学）

1.方言談話資料の現状

☆研究素材として見直されつつある方言談話資料

- 全国方言資料（NHK）
- 方言談話資料（国語研）
- 各地方言収集緊急調査（文化庁）

2.各地方言収集緊急調査（その1）

- 各都道府県において緊急に調査し、記録・保存することが目的
- 各県5地点（約200地点）
- 7場面の対話に絞り、1地点10時間の録音を収集
- 老年60歳以上、若年20～30代
- 3時間分を文字化・標準語訳

2.各地方言収集緊急調査（その2）

- 永久保存
- 公開されているのはごく一部
- 文化庁から国語研への移管を機にデータベース化を企画する
- データベース科研（H9～）
- 文字化資料の電子テキスト化
- 音声資料のディジタル化

3.全国方言談話資料データベース

3.1 データ量などの問題

- 2000時間の録音（83日分）
- 600時間分の文字化（25日分）
- どの程度が量的に妥当か？
- NHK1999との比較

  2.2GB，30時間分，141地点

  PDF形式＋音声ファイルリンク

  ただし文字化部分は画像

3.2 文字化部分の作業

- 手書き文字化・標準語訳原稿を電子テキスト化
- 表記・形式を統一
- できるだけ原本に忠実に入力
- 文字化部分と標準語訳部分を分け、エクセルで入力
- 最終的にはWordで形成

3.3 音声部分の作業(その1)

- カセットテープをパソコンに取り込む(ディジタル化)
- ソニーVAIOのギガポケット(ビデオインターフェイス)を利用
- 直接ハードディスクに談話単位(15～30分)で録画(録音)
- 画面トラックに文字化資料を同期させておく(CCDカメラで)

3.3 音声部分の作業(その2)

- パソコン内で再生しながら、音声ソフト(sp4win)を用いて再録音
- サンプリングレート、形式の調整
- 8ビット, 22.05KHz, モノラルで(NHK1999と同じ形式)
- さらに発話単位に編集して完成

3.4 データベース化の方法(その1)

二つの方法

- ワープロ・エクセル等で作成したものをPDF化し、音声ファイルをリンクさせる(PDF方式)
- データベースソフトを用いて、文字化と音声を一体化したものを作成する(データベース方式)

3.4 データベース化の方法(その2)

- PDF方式
  - 誰にでも使える
  - 個々のファイルが扱いやすい
- データベース方式
  - 検索が自由にできる
  - 色々な角度から分析できる

3.4 データベース化の方法(その3)

現時点での方針

- PDF方式とデータベース方式の両方を作成する
- 量的に両方が無理ならば、PDF方式を優先し、データベース方式のものは、音声をユーザが後で組み込む形のものにする

4. 今後の予定および課題

- 音声談話資料の編集の難しさ
- 「編集にかかる手間・費用」と「学術的価値」とのかねあい
- 常に利用のことを頭に置くこと
- 2004年までに完成の予定
- パソコン環境の変化への対応