自然会話音声データの電子化について

武田雅一

磯野英治

はじめに

東京外国語大学 COE 言語教育学班談話グループでは，2005 年度に行われた COE 活動の一環として宇佐美研究室に蓄積，保存されてきた「話し言葉データ」の整備を行った。これらのデータを，COE プロジェクトとしての会話分析に活用したり，一般公開コーパスである「BTS による多言語話し言葉コーパス」を構築したりするためである。本報告では蓄積会話データ整備の目的，データ整備の設備，整備作業の工程とその問題点，また蓄積会話データの整備の必要性などを中心にまとめる。

1. データ整備の目的

宇佐美研究室にはこれまで,卒業論文・修士論文のデータを中心として，様々な自然会話データが蓄積されている(以下，蓄積会話データ)が，これらはカセットテープ，MDなど様々な媒体で保存されている。COEプロジェクトにおいては，この「話し言葉データ」を効率的に会話分析や「BTSによる多言語話し言葉コーパス」構築に活用するため，作業環境の整備に力を注ぐ必要があった。そのため様々な媒体で保存されている「話し言葉データ」を電子化しデータを共有化することによって，会話分析や「BTSによる多言語話し言葉コーパス」構築の際に必要な「基本的な文字化の原則（BTSJ)」による文字化作業を円滑に行えるようにした。

2. 設備面について

今回の蓄積会話データ整備では，カセットテープ，MD に収録されている「話し言葉データ」の電子化を行った。使用した機器を以下に挙げる。

- パソコン
- 外付けハードディスク
- オーディオキャプチャー（カセットテープ，MD を PC に取り込むための電子ファイル変換キャプチャー）
- オーディオプレイヤー（カセットテープ，MD 対応のもの）

電子化の方法はパソコン, オーディオキャプチャー, オーディオプレイヤーをそれぞれ接続し，

カセットテープ, MD をオーディオプレイヤーで再生してオーディオキャプチャーに取り込むという方法である。電子化の際, 基本的なオーディオキャプチャーの操作はパソコン上で行えるため，電子化の進行状況の確認や音声の品質管理などはパソコンで行った。

電子化した音声データは, WAVE ファイルという形式の音声ファイルになる。これはカセットテープや MD をそのまま電子化したものであるため容量が大きく，保存，管理には不向きであった。そのためこの WAVE ファイルを圧縮，変換するソフトを使用し，電子データとして扱い易いレベルまで加工（エンコード）する必要があった。今回の蓄積会話データ整備では，簡易的で無料配布されている音声ファイル変換ソフト「午後のこ〜だ（http://www.marinecat.net/mct_top.htm）」を使用し, WAVE ファイルを MP3 ファイルに変換した。

こうして出来上がった電子ファイルは，MP3 ファイル形式に圧縮しても，パソコンの本体に蓄積すると容量がかさみパソコン本来の機能に支障が出るので, 外付けハードディスク内に保存した。また, 完成した電子ファイルは, 随時ディレクトリを立ち上げて電子ファイルへの変換が成されていることを確認した。

3. 蓄積データ整備を振り返って

データ整備をするにあたっての進め方を振り返ってみると, まず原データとして存在するたくさんのデータの保存の現状を検証することから開始した。全体像を掴むためにも作業の各工程を考察して, 作業時間がかかる部分, すなわち別媒体への録音時間に相当する部分は一斉作業ができないのでこれらの作業を抽出してその段取りを特別に分けた。さらに全体像を把握するために作業管理表を作成し，進捗状況を随時掌握できるようにした。

電子化の対象となったデータは，17 時間分の会話である。実際に研究に使用されたり文字化されたりした場面は一部分であっても, 談話の全体像を捉えるためには収録されている全ての会話をスクリプト化するので，録音時間分のデータを取り込んだ。但し，CD-R のように媒体変換済みのものは, パソコン内に取り込むときには電子データのフォルダーをそのままコピーできたものもあった。各媒体に適応した作業環境を常駐していないので, 各機材を配置して随時変換できる環境を準備するためだけでも数十分かかった。電子媒体として作成された内容をいつでも簡単に検索し，取り出せるようにするために，パソコン内の整理に数十分かかった。

困難な点，もしくは問題となった点は以下の通りである。

第一に特にカセットテープを電子化する際は, 作業時間の見積もりがなかなかたてられなかったことである。個々の媒体の最大録音時間が様々なので，全テープを流す必要があり，内容の確認と検証に手間取った。

第二にオーディオキャプチャーなどの作業環境を随時設定しなければならなかったので, 設定と撤収の時間も累計すればかなり時間を費やしたことになる。

第三に実際作成した電子データを活用する際には, 音声データ, 画像データともに容量が大きいために, 現在一般に普及されている通信回線を使用したデータの伝送ができなかったので，データの受け渡し方法はやはり物理的媒体での授受が必要となった。

また現状として, 各会話データを収録する人たちの中には録音媒体としてまだアナログ式のテープが用いられることが多い。そこで文字化作業を行う際に, 繰り返し何度も同じ場面を聞き返すことが必然的に生じる。このときに繰り返される巻き戻し行為によって原データとしてのテープの劣化は避けられない。長時間録音ができるテープはそのままの状態で保存していると将来, 音質や媒体そのものの疲弊度が目立ち, 最終的に当初収録したときと比べて同様の性能を維持しているとはいいがたいものである。将来も保存していくものについては, 原データをデジタルの媒体に一旦コピーしてから、パソコン内に移し変えることも必要である。

4. 蓄積データの管理の必要性

一般的にデータバンクとして, 蓄積されたデータをできるだけ多くの人が汎用的に使えるように共有化する方策を考えると, 電子化することは必然的である。データの電子化にかかる作業の効率性を追究するならば, 音声・画像データの収録の際の媒体はできるだけ統一した方がよい。今回のデータ整備の経験からオーディオキャプチャーという中継機器を経由しないで直接パソコンに取り込めるようなUSB接続タイプの機材を導入することが推奨できる。またデータを複数の媒体で保管, 保存せざるを得ない場合は, 逐次電子化できる環境を整理することが必要である。そして教室活動の中でその手法をルール化して研究者各自が体得していくことも研究の意義に関わるものである。

原データは当初の状態のまま保存するものである。しかしこれを電子媒体にて作成したものがあると, データの一元的管理と共有化がしやすくなる。そうするとデータの貸出時の授受が簡略される。またデータを使用する人にとっても, 各自がパソコン内に取り込んで操作できるので電子化されたデータの方が, 文字化作業やデータの受け渡し事務の効率アップにつながり, 利便性が高くなる。

おわりに

本稿では, 自然会話分析や「BTS による多言語話し言葉コーパス」構築に必要不可欠な音声データを, どのように扱いやすいかたちにしていくかを共有性, 汎用性の観点から述べた。今回の蓄積会話データ整備では, 音声データを整備し随時視聴, 閲覧できる環境にすることによって, 文字化作業への行程を円滑に行うことができた。

音声データを電子化, 整備する際に使用する電子機器は, 日進月歩でその技術が進歩している。私たちは常に新しい技術に関心を寄せ, それぞれの研究に相応しい環境を追求していくことが必要であろう。