国立国語研究所学術情報リポジトリ

NINJAL データベースを活用した言語研究の実施について

著者	鈴木 成典, 五十嵐 陽介, 李 勝勲
雑誌名	言語資源ワークショップ発表論文集
巻	1
ページ	79-82
発行年	2023
URL	http://doi.org/10.15084/00003726

NINJAL データベースを活用した言語研究の実施について

鈴木　成典（国際基督教大学大学院）†
五十嵐　陽介（国立国語研究所）
李　勝勲（国際基督教大学）

Linguistic Research Using NINJAL Database

Michinori Suzuki (Graduate School of International Christian University)
Yosuke Igarashi (National Institute for Japanese Language and Linguistics)
Seunghun J. Lee (International Christian University)

要旨

本稿では、国立国語研究所の共同利用型共同研究（登録型）で利用可能な豊富なデータベース（研究資料室収蔵資料：fo0245）を用いたデータ処理方法について紹介する。今回利用したデータベースでは多くのデータへのアクセスが可能な反面、実験全体の録音ファイルのみが利用可能であったため、初めに実験の録音音声をもとに刺激のメタデータを構築し、単語や実験内でのセクション、話者、繰り返しの有無をもとに研究後にも識別可能な刺激のアーカイブ ID を作成した。次に、Praat スクリプトを用いて録音全体における刺激間の境界の配置や各刺激への ID の付与、録音音声全体の個別刺激ファイルへの分割を行うことで、分析対象に対するアノテーションの保存を半自動的に可能とした。本研究手順により、録音データベースを用いたより効率的な研究が可能となるだろう。また、処理したデータをアーカイブすることで将来の様々な研究に役立てることができるだろう。

１．はじめに

本稿では、国立国語研究所の共同利用型共同研究（登録型）で利用可能な豊富なデータベースを用いたデータ処理方法について紹介する。共同利用型共同研究（登録型）は、国立国語研究所が所有する多種多様な研究資料・言語資源・分析装置を活用して研究を行うことができる制度である。著者らの行なった研究では、「国立国語研究所研究資料室収蔵資料」（主に 1950 年代から 2000 年代初頭にかけて行われた 250 以上の研究プロジェクトの研究資料）にある、「fo0245：日本語音声における韻律的特徴の実態とその教育に関する総合的研究」において収集された録音音声データを用いて分析を行った。データ処理の過程で録音データに含まれる各刺激のアーカイブ ID を作成し、この ID をもとに各刺激を個別の音声及びテキストグリッドファイルへの分割を行ったため、本研究以降に同一データを活用して研究を実施する際に役立て、より効率的に研究を行うことが可能である。2022 年 8 月現在においてもコロナ禍は依然として続いており、新たに実験を行うことが簡単ではないため、共同利用型共同研究（登録型）の利用や識別可能な ID とともに処理したデータのアーカイブを行うことで、このような社会情勢化においても研究を促進することが可能となると考えている。

† Email address: g239704k@icu.ac.jp

２．使用したデータベースについて

著者は、1990 年代初頭に実施された「fo0245：日本語音声における韻律的特徴の実態とその教育に関する総合的研究」の実験録音データを使用して日本語の有声性の対立に関する研究を行なった。本節ではこのデータ全体に関する説明を行う。

２.１　被験者

被験者は日本の各地方（北海道・東北・関東・中部・近畿・中国・四国・九州・沖縄）に在住の日本語方言話者であり、年齢も小学生、中学生、若年層、壮年層、中年層、高年層（老年層）と多様な被験者による録音が存在する。

２.２　刺激

録音実験を行なったオリジナルの研究プロジェクトは日本語の韻律をトピックとしており、実験録音データは（1）に示したような様々な種類の刺激を含んでいた。

（1）　刺激の種類

a. 五十音
b. 名詞
c. 動詞とその活用
d. 形容詞とその活用
e. 文章
f. 童話（桃太郎）の朗読
g. 数字（1 から 9 までと四桁の数字）

これらの刺激は地域や被験者に応じて刺激やセクションの順番が異なっている場合があり、その差異は名詞の種類や名詞に格助詞「が」を付けた繰り返しの有無などに見られた。例を挙げると、後述するように東北方言と東京方言の録音実験は異なる刺激リストを用いて行われており、東北方言の録音データにおいては繰り返しがあったものの、東京方言の録音データには繰り返しはなかった。

２.３　実験手順

録音実験では、まず実験者が被験者に名前や生年月日の確認などを行い、例文を読み上げてもらい練習をして実験の手順について説明した後に、被験者が紙に印刷されている刺激を一つずつ読み上げていくという形で行われた。実験の途中で外を走る車の音などの外部音が入った場合や、録音が不明瞭な場合などは実験者の判断により被験者に再度の読み上げをお願いしていた。

ここまで第二節では利用した「fo0245 日本語音声における韻律的特徴の実態とその教育に関する総合的研究」のデータベース全体の説明を行なったが、次節ではこの中から著者の行った研究で焦点を当てたデータの処理方法について詳説する。

３．データ処理

本節ではデータの処理方法について順を追って詳しく説明する。著者は日本語方言間での有声性の対立に焦点を当てており、方言間での音声的な差異を研究するため、東北方言と東京方言を分析した。分析対象とした話者はデータベース上の「高齢層」であり、それぞれ

の方言から 10 名のデータに対して分析を行なった。被験者の男女比は、東北方言では女性のデータが存在しなかったため全員男性であったのに対し、東京方言では 5:5 であった。該当するデータの被験者全員分の年齢は正確には判明しなかったが、分析した録音実験は 1991 年に実施されており、少なくとも 20 名中 12 名が 1910 年代から 1930 年生まれのため、「高齢層」に分類されている被験者は実験実施時点で 60 歳以上であったと推察される。

また、本データベース上には録音リストが存在するものの、刺激のリストを見つけることができなかったため、まず初めに実験音声をもとに刺激リストを作成した。刺激は前節で述べたように東北方言と東京方言間で名詞や文章が異なっており、東北方言の録音実験では、ある名詞を発話した後に格助詞「が」を付けた繰り返しが存在した。刺激のリストを作った後に、各刺激に対して「“単語 ID”-“繰り返し番号”-“実験内のセクション ID”-“被験者 ID”」となるように、本研究が終わった後においても識別可能なアーカイブ ID を作成した。被験者 ID はデータベース上で各音声ファイルに割り振られていた番号を使用した。例えば、「リスト上の一つ目の単語(W001)」で「セクション番号 S1」、「被験者 ID が 001(JPD001)」だった場合は“W001-S1-JPD001”となり、繰り返しがある場合には最後に“W001-1-S1-JPD001”のように繰り返し番号を付けた。単語 ID と繰り返し ID を頭の部分に配置した理由は、のちの処理で個別の音声ファイルへと分割される際に、フォルダ内で同じ刺激のファイルをまとめて表示させるためである。

録音音声データの処理は、Praat (Boersma and Weenink, 2022)を用いて行なった。本研究では名詞に焦点を当て、まず録音音声ファイルの中から(1b)や(1e)のセクションを切り取った。その後 Praat スクリプトを用いてセクション内の各刺激の境界を配置し、さらにその境界をもとにして各刺激に対して上で説明したアーカイブ ID を与えた。また、各刺激をアーカイブ ID がファイル名となる個別の音声ファイルへと分割し、保存した。ここまでの処理が終わった後、個別の音声ファイルに対して手動でアノテーションを行なった。これらの Praat 上での処理は、Praat スクリプトを用いることでアノテーションを除き半自動的に行った。

アノテーションが完了した後は、個別の音声ファイルとアノテーションを行なったテキストグリッドファイルの入っているフォルダに対し Praat スクリプトを使用してアノテーションを行なった箇所の情報を抽出した。その後、抽出されてできたテキストファイルをエクセルから開き、統計分析ソフトウェアの R (R Core Team, 2021)を用いて分析できるように加工した後に csv ファイルとして保存した。

４．おわりに

本稿では国立国語研究所の「共同利用型共同研究（登録型）」を用いて実施した研究手法を説明した。今回の研究で実施した処理過程の中でも、アーカイブ ID の使用が重要な点だと考える。一つの研究の後でも識別可能なアーカイブ ID をそれぞれの刺激に対し与えることで、将来同じデータベースを用いて研究を行う際に役立てることができ、また先行研究の結果の再現性の確保にもつながるだろう。特に、現在の社会情勢を念頭に置くと、まだ新たな実験を対面で実施することは必ずしも簡単ではないため、データベースを活用した研究の重要性がさらに高まると考えられる。その際にオリジナルのデータベースに加え、処理されたデータが識別可能な ID とともにアーカイブされていたとしたら、仮に研究トピックが異なっていたとしても、将来行われる研究の大きな促進につながると考えている。

謝　辞

本稿は、国立国語研究所の共同利用型共同研究プロジェクト「日本語の有声性の対立への複数の音響指標の影響」（研究代表者：李勝勲）の研究成果である。また、第一著者は国際基督教大学より「国際基督教大学博士研究員（A 種）」の助成を受けている。

参　考　文　献

Boersma, P. and Weenink, D. (2022). “Praat: doing phonetics by computer [computer program] (version 6.2.14)”.

R Core Team (2021). R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing, Vienna, Austria.