国立国語研究所収蔵音源資料と所蔵音源データベース構築

高田　智和（国立国語研究所）　大石　恵輔（東京電機大学）
山口　亮（中央大学）　石本　祐一（国立国語研究所）

国立国語研究所では，方言，語彙，言語生活，日本語教育，言語コーパスといった，日本語に関する様々な調査研究において収録した音源資料が保存されている．各種音源資料の保存と再利用のため，デジタル化複製と，簡便な試聴環境を提供するため，所蔵音源データベースの提供（所内限定利用）を行っている．

Sound Materials in National Institute for Japanese Language and Linguistics (NINJAL) and Development of Sound Materials Database

Tomokazu Takada (National Institute for Japanese Language and Linguistics)
Oishi Keisuke (Tokyo Denki University)
Ryo Yamaguchi (Chuo University)
Yuichi Ishimoto (National Institute for Japanese Language and Linguistics)

The National Institute for Japanese Language and Linguistics (NINJAL) preserves sound materials collected or created by various research projects in the past 70 years. Nowadays, NINJAL digitalizes the old audio recordings for the conservation and the reuse, and develops the Sound Materials Database for the trial listening (accessible by NINJAL domain only).

１．まえがき

国立国語研究所（以下国語研）では，1948（昭和 23）年 12 月の創設以来，およそ 70 年にわたって方言，語彙，言語生活，日本語教育，言語コーパスといった，日本語に関する様々な調査研究を重ねてきた．各調査研究課題の成果は報告書や論文として公刊されてきたが，研究成果の前段階の中間生成物に相当する情報カードや集計表，さらに，研究の一次資料に相当する調査票・録音・語彙調査の雑誌原本，調査研究運営の記録である調査計画書や会議録も現存し，現在それらは国語研研究資料室に保存されている 1）2）3）．

本報告では，国語研収蔵音源資料を紹介し，保存と再利用のための所蔵音源データベース構築の現状を述べる．かつて言語資料と言えば，文字資料として保存・継承されるものであったが，近代以降の録音技術の発達によって，「音声」そのものの精緻な記録が可能となった．国語研収蔵音源も，録音技術の発達とともに，オープンリール，カセットテープ，DAT，そして，現在の電子ファイルへと，録音媒体も変遷している．また，前代の録音媒体に記録された音源は，次代の録音媒体に移し替えて保存・継承され，現在の所蔵音源データベースの構築に至っている．

２．国立国語研究所研究資料室収蔵資料

国語研研究資料室で保存している研究資料は，概ね 1 課題のもとで生産された研究資料を 1 資料群として管理し，2017 年 11 月時点で資料群は 233 である．資料群の概要記述は「国立国語研究所研究資料室収蔵資料」（http://rmr.ninjal.ac.jp/）として，2017 年 3 月から一般公開している（図 1）．


国立国語研究所 研究資料室収蔵資料
Research Materials on NINJAL
資料群一覧
List of Research Materials

図 1　研究資料室収蔵資料

また，個別の収蔵資料に対しては，目録を作成している．以下の 4 点である．

- 保存箱目録　文書資料や情報カード

- 雑誌目録　語彙調査に用いた雑誌原本
- 地図目録　言語地図や地形図
- 音源映像資料目録　オープンリールやカセットテープなど記録媒体ごと

目録は整備途上であるため，所内限定公開であったが，2017 年 5 月に雑誌目録を「中央資料庫未製本雑誌所蔵リスト」（http://rmr.ninjal.ac.jp/magazinelist.html）として一般公開を開始し，国語研 OPAC への収蔵雑誌の書誌登録も進めている．

研究資料室の収蔵資料そのものは，原則として公開であり，閲覧利用（来館利用）が可能である．研究資料室は，広く共同利用に提供するために，目録類の整備と一般公開を今後も継続していく．

３．収蔵音源資料

国語研収蔵音源資料は，音声言語の調査研究課題によって収集された録音音源である．録音の対象は，ラジオ番組（マス・メディアのことば）や講演（フォーマルな話しことば）もあるが，中核は一般個人を対象とした自然談話とインタビュー（面接調査）の録音である．以下，主要な音声言語・話しことば研究の録音資料を紹介する．

- fo0104 談話語の実態

共通語による日常談話を分析するために，1952-53 年に録音．文字起こし原稿や KWIC（keyword in context）も作成されている．報告書は『談話語の実態』（1955 年）．収蔵音源は，オープンリール 74 本，DAT 110 本．

- fo0100 待遇表現の実態：松江 24 時間調査資料から

1963 年に松江市のある市民の家庭内での一日の発話をすべて録音．文字化資料には文・文節・形態素の切れ目を付加し，コンピュータ処理に利用．報告書は『待遇表現の実態―松江 24 時間調査資料から―』（1971 年）．収蔵音源は，オープンリール 6 本，DAT 10 本

- fo0150 企業の中の敬語

企業の中での敬語意識と敬語使用を解明するため，1975-1977 年に日立製作所ほかの協力のもと面接調査を実施し，面接調査を録音．報告書は『企業の中の敬語』（1982 年）．収蔵音源は，オープンリール 34 本，カセットテープ 260 本

- fo0172 学校の中の敬語

中学生・高校生の敬語意識と敬語使用を解明するため，1989-1991 年に東京・大阪・山形の中学校・高校で面接調査を実施し，面接調査を録音．報告書は『学校の中の敬語Ⅰ―アンケート調査編―』（2002 年），『学校の中の敬語Ⅱ―面接調査編―』（2003 年）．収蔵音源は，カセットテープ 207 本．

- fo0216 敬語と敬語意識―愛知県岡崎市における第三次調査―（岡崎調査 3 回目）

敬語と敬語意識に関する定点経年調査．1953 年，1972 年に続き，2007-08 年に愛知県岡崎市で行われた第 3 回目の面接調査を録音．報告書は『敬語と敬語意識―愛知県岡崎市における第三次調査―』（2010 年）．「岡崎敬語調査データベース」を公開（http://www2.ninjal.ac.jp/longitudinal/okazaki.html）．収蔵音源は，DVD 159 枚

- fo0219 地域社会の言語生活―鶴岡市における戦後の変化―（鶴岡調査 3 回目）
  fo0214 第 4 回鶴岡市における言語調査（鶴岡調査 4 回目）

統計数理研究所との共同研究として実施した，共通語の普及に関する定点経年調査．1950 年，1971 年に続き，山形県鶴岡市で行われた 1991-92 年（第 3 回目），2011-12 年（第 4 回目）の面接調査を録音．報告書は『地域社会における言語生活―鶴岡における 20 年間隔 3 回の継続調査―』（2007 年），『第 4 回鶴岡市における言語調査　ランダムサンプリング調査の概要　資料編：第 1 分冊「音声・音韻」編』（2014 年），『第 4 回鶴岡市における言語調査　報告書　資料編：第 2 分冊「語彙・文法，言語生活項目」編』．「鶴岡調査データベース」を公開（http://www2.ninjal.ac.jp/longitudinal/tsuruoka.html）．fo0219（第 3 回目）の収蔵音源は，カセットテープ 1,179 本，DAT 1,081 本．fo0214（第 4 回目）の収蔵音源は，DVD 19 枚．

- fo0141 方言談話資料

日本各地の方言談話の調査・記録・文字化を目的とした基礎研究において，1974-75 年に録音．報告書は『方言談話資料』（1978-87 年）．収蔵音源は，オープンリール 233 本，カセットテープ 217 本，DAT 326 本．

- fo0148 方言録音文字化資料に関する研究

1977-85 年度の文化庁調査「各地方言収集緊急調査」で全国 224 地点の方言談話を録音・文字化．資料は国語研に移管され，一部を『日本のふるさとことば集成』（2001-08 年）として刊行．収蔵音源は，カセットテープ 7,493 本，DAT 13,974 本．

- fo0170 日本語学習者による日本語と母語発話の対照言語データベース

2002-04 年に，日本語教育実践・教師研究等で利用するために，日本語学習者による日本語での朗読・スピーチと，母語での同内容の朗読・スピーチを録音・文字化．「日本語学習者による，日本語・母語対照データベース」を公開（http://contr-db.ninjal.ac.jp/speech_01.html）．収蔵音源は，DAT 288 本．

- fo0204 日本語学習者会話データベース（横断調査）

2006-09 年に，日本語教育研究・言語習得研究の基礎データを得るために，日本語学習者と日本語母語話者（インタビュアー・テスター）の会話を録音・文字化．「日本語学習者会話データベース」を公開（https://nknet.ninjal.ac.jp/nknet/ndata/opi/）．収蔵音源は，MD 393 本，CD 62 枚．

- fo0209 話し言葉コーパスの言語的･パラ言語的構造の解明に基づく『話し言葉工学』の構築（CSJ）

自発的な「話し言葉」の情報処理技術の基盤を確立することを目的に，1999-2003 年に講演等を録音・文字化しコーパスを作成．報告書は『話し言葉コーパスの構築法』（2006 年）．「日本語話し言葉コーパス」を公開（2017 年 2 月から Web 上の検索アプリケーション「中納言」での利用も可能．http://pj.ninjal.ac.jp/corpus_center/csj/）．収蔵音源は，CD 348 枚，DAT 1,022 本．

４．保存と再利用のための媒体変換

国語研では，収集した音源資料の媒体変換を随時行ってきた．媒体変換の目的は，第一に音源資料の保存である．第二に，国語研の調査研究での再利用である．過去の自然談話の録音は不可能であり，古くなればなるほど資料としての貴重度が増してくる．特に，各地方言は衰退が進んでいるため，記録として録音そのものが貴重である．また，前述の「fo0100 待遇表現の実態：松江 24 時間調査資料から」や「fo0150 企業の中の敬語」の調査は，同様の調査の実施が現代では困難であるため，研究事例として貴重なものとなる．

録音保存に使用された主な媒体は，年代順に，オープンリール，カセットテープ，DAT（Digital Audio Tape）である．古い録音であれば，オリジナルのオープンリールから，カセットテープと DAT の複製が作られている．カセットテープをオリジナルとする場合では，DAT が複製である．DAT はほぼ複製物であるが，DAT がオリジナルである資料群もある．研究資料室の収蔵本数は次の通りである．

- オープンリール　約 2,900 本
- カセットテープ　約 16,000 本
- DAT　約 23,000 本

かつては DAT で複製・保存する方針であったが，現在はオリジナルのオープンリールやカセットテープから電子ファイル（wav 形式，サンプリング周波数 48 kHz，量子化ビット数 16 bit）に複製・保存する方針に転換し，研究資料室で媒体変換を進めている．DAT 作成時に電子ファイル（aiff 形式）も併せて作成したものもあり，その場合は，aiff 形式の電子ファイルを wav 形式に変換して保存する方針としている．

このほかにも，ミニディスク（MD）に録音されたものが約 800 本ある．ミニディスクは，1900-2000 年代に主に日本語教育分野の調査研究で利用された．この時期には，国語研の公開行事（「ことばフォーラム」）の録音記録でも，ミニディスクが利用されている．しかし，ミニディスクの流通状況に鑑み，現在はミニディスクによる録音資料もデジタル化複製の対象としている．

デジタル変換後の音源資料の再利用例には，前述の「fo0104 談話語の実態」の録音が「昭和話し言葉コーパス」に，「fo0148 方言録音文字化資料に関する研究」の録音が「方言コーパス」にそれぞれ素材として再利用されている．「昭和話し言葉コーパス」は「大規模日常会話コーパスに基づく話し言葉の多角的研究」,「方言コーパス」は「日本の消滅危機言語･方言の記録とドキュメンテーションの作成」のように，どちらも国語研の現在の調査研究課題の一部として，過去の録音資料の再利用によってコーパス開発が進められている．

なお，2000 年代以降の録音資料は，wav 形式等の電子ファイルで最初から作成され，CD や DVD，ハードディスクに格納して研究資料室で保存している．

５．所蔵音源データベースの構築

音源資料の研究利用を促すには利用者が適切な資料を見つけられる環境が必須であり，目録提供だけではなく聴取による資料内容の確認が求められる．しかし，資料貸出には手続き上利用申請が必要なため，これまで試聴による事前確認が難しい状況であった．また，デジタル変換により通常の PC での再生が可能となったが，各資料の音源収録時間は長いものでは 2 時間以上あり，ファイルサイズを考慮すると取り扱いに難がある．そこで，国語研研究資料室ではデジタル変換後の所蔵資料の試聴環境を提供するため，「所蔵音源データベース」の構築に取り組んでいる．

所蔵音源データベースは資料のメタデータと

音源ファイルで構成されており，資料の検索および試聴を行うことができる Web サイトとして 2017 年 3 月に運用を開始した（図 2）．

図 2　所蔵音源データベース

2017 年 11 月現在，次の 4 つの資料群の音源（488 ファイル，合計 371 時間）を配信している．

- fo0061 話しことばの文法の調査研究
- fo0104 談話語の実態
- fo0122 全国方言文法の対比研究
- fo0145 話しことばの文型

今後もデータベースへは電子ファイルへの媒体変換が済んだものから随時収録する計画であり，2017 年度中の増補として 11 資料群，約 3,000 ファイルを予定している．増補予定の資料群は次の通りである．

- fo0008 大都市における言語生活の実態調査
- fo0017 日本語教育のための言語能力の測定：日本人の知識階層における話しことばの実態
- fo0030 文字言語の学習負担についての研究
- fo0032 言語能力の発達に関する調査研究
- fo0059 国際社会における日本語の総合的研究
- fo0150 企業の中の敬語
- fo0170 日本語学習者による日本語と母語発話の対照言語データベース
- fo0172 学校の中の敬語
- fo0202 日本語教育における基本文型に関する研究
- fo0204 日本語学習者会話データベース（横断調査）
- fo0217 社会変化と言語生活の変容（鶴岡調査 2 回目）

メタデータは資料群と音源に分け，関連づけられた 2 つのテーブルとしてデータベース化した．資料群の分類は調査研究課題と 1 対 1 で対応しておらず，ひとつの調査研究課題の音源が 2 つの資料群と関連する場合がある．そのため，音源ごとに関連資料群を複数設定できる設計とし，資料群テーブルと音源テーブルとして分けることとした．テーブルの構成を表 1, 2 に示す．資料群テーブルの「資料群 ID」と音源テーブルの「資料群 ID 1」および「資料群 ID 2」が関連づけられている．また，資料群の詳細は前述の「研究資料室収蔵資料」によって管理されているため，本データベースでは必要最小限のテーブル構成としている．

表 1　資料群テーブル

フィールド	データ型	データ例
資料群 ID	char(6)	fo0061
表題	varchar(1024)	話しことばの文法の調査研究

表 2　音源テーブル

フィールド	データ型	データ例
音源 ID	varchar(64)	wa-dt00391
音源名	varchar(256)	ラジオことばの研究室(1)
備考	text	話しことば研究室資料
資料群 ID 1	char(6)	fo0061
資料群 ID 2	char(6)	
時間長	char(6)	48:16
サイズ	int(11)	46343808

これらのテーブルを Web ブラウザから参照でき，対応する音源ファイルの試聴も行える Web アプリケーションを構築した．このアプリケーションの特徴の一つとして，音源再生の機構が挙げられる．所蔵音源の中には個人情報が含まれるものがあり，無制限に公開することはできない．そのため，研究者が所蔵資料の研究利用を希望する場合は利用申請の上，研究資料室から音源ファイルを提供する措置をとっている．

また，本システムは利用申請前の試聴環境の提供を目的としており，利用者が未申請時でも音源を聴取でき，かつ音源ファイルを取得できない構成が必要となる．そこで，Web アプリケーションによる専用の再生環境を実装したシステムとし，音源ファイルのダウンロードによる入手を妨げる仕組みを導入した．

さらに，本アプリケーションで試聴できる音源には保存用の wav 形式から mp3 形式（サンプリング周波数 16 kHz，ビットレート 128 kbps）に変換したファイルを用いることで，再生時のファイル転送の負荷を軽減するとともに高品質の音

源が流出しない設計とした．なお，データ保全のために所内ネットワークからのみアクセス可能としているが，所員に限らず来館者も利用することができる．

本アプリケーションでは資料群の一覧や検索を行うことができる（図 3）．ほかにも，すべての音声ファイルの一覧表示も可能であり（図 4），複数の経路から目的の音源にたどり着けるような導線を考慮した．また，資料群一覧からは資料群 ID により「研究資料室収蔵資料」の当該資料群を閲覧することができ，資料群の音源リスト（図 5）や音源の内容（図 6）もシステム内で一貫して参照できることで目的の資料を容易に探し出すことができるようにしている．前述のように音源の再生はブラウザ上の専用プレイヤーで行っており（図 7），音源の途中からの再生にも対応するなど，音源の試聴における簡便さとデータ保護を満たした機能を備えている．

図 3　所蔵資料群一覧

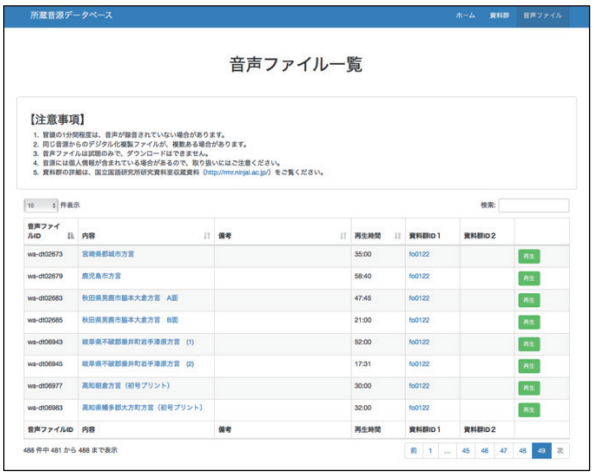

図 4 音声ファイル一覧

図 5　資料群ごとの音源リスト

図 6 音源内容

図 7 音源再生プレイヤー

本データベースの遷移を図 8 に示す．本データベースから「研究資料室収蔵資料」への参照は実装しているが，「研究資料室収蔵資料」から本データベースへの参照は未実装であり，今後の課題である．

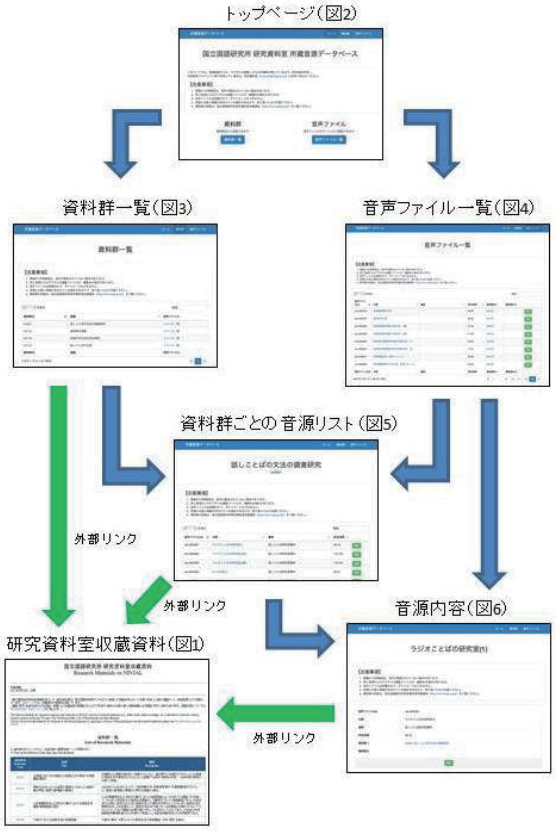

図 8 遷移図

7．あとがき

本報告では，国立国語研究所の収蔵音源資料と，収蔵音源資料のデータベース構築の取り組みについて述べた．最後に今後の課題を列挙する．

- 音源資料があることの報知

日本語学・日本語教育の研究者であっても，国語研究所に音源資料があることはあまり知られていない．資料群概要記述の一般公開だけでは不足であるため，音源資料目録の早期一般公開に努める．

- 共同研究の制度設計

現在，音源資料の提供は，国語研の共同研究プロジェクトに限られる．今後，例えばNHK 番組アーカイブス学術利用トライアルのような，国語研音源資料を用いた公募型共同研究の制度を設けることで，共同利用の促進をはかる．

- メタデータの整備

収蔵音源資料の共同利用を進めるにしても，言語研究に用いるならば，録音年代や録音場所，話者の属性（性別・年齢・出身地域・母語）などの情報が必須である．これらの情報は，保存箱の紙資料や報告書に収録されていることが多いため，それらをメタデータに増補することが有効である．また，文字化データとの連携も，すぐに研究利用ができるデータベースとしては魅力的であろう．しかし，方言音声のように取り扱いに専門的知識が必須となるものに対しては，アーカイブズ的位置づけの研究資料室だけでは，メタデータ整備に限界がある．

参考文献

1）森本祥子：EAD を用いた資料記述システムの開発について―国立国語研究所の事例，アーカイブズ学研究，No.4， pp.92-102（2006）．
2）寺島宏貴：日本語研究資料の整備と公開―国立国語研究所研究資料室の取組み，国立国語研究所論集，No.10，pp.245-263（2016）．
3）山口亮・関川雅彦： 国立国語研究所所蔵資料アクセス環境改善への取り組み，人文科学とコンピュータシンポジウム論文集人文学情報の継承と進化―ビックデータとオープンデータの潮流の中で，情報処理学会シンポジウムシリーズ Vol.2016,No.2，pp.51-56（2016）