アーカイブデータを利用した言語研究とその応用可能性

鈴木成典（国際基督教大学大学院アーツ・サイエンス研究科、日本学術振興会）[1]
鎌野慈人（ストーニーブルック大学大学院）
坂本誓（国際基督教大学 RA）
鎌倉欧亮（国際基督教大学教養学部）
Seunghun Lee（国際基督教大学, ヴェンダ大学）
Yu Yan（立命館大学）
Jeremy Perkins（会津大学）
五十嵐陽介（国立国語研究所）

Linguistic Research using Archive Data and its Application

Michinori Suzuki (International Christian University Graduate School & Japan Society for the Promotion of Science)
Shigeto Kamano (Stony Brook University Graduate School)
Chikau Sakamoto (International Christian University RA)
Osuke Kamakura (International Christian University, College of Liberal Arts)
Seunghun Lee (International Christian University & University of Venda)
Yu Yan (Ritsumeikan University)
Jeremy Perkins (University of Aizu)
Yosuke Igarashi (National Institute for Japanese Language and Linguistics)

要旨

本稿は、国立国語研究所の共同利用型共同研究で利用可能である豊富な音声データベースに対して、我々が実際に行っている音声データの処理方法を紹介する。使用したデータベースは大規模な録音実験に基づくものであり、1000 人以上の話者の録音が存在する一方で、方言ごとに刺激リストが異なっている。そこで、初めに各方言の録音を確認し、セクション・刺激・話者が識別可能なアーカイブ ID を作成した。その際、同一方言内の話者間でも刺激の順番や繰り返しの回数にばらつきが確認されたが、Praat スクリプトや Excel を用いることで対応を可能とした。様々な方言話者を対象に産出実験を実施する際、必ずしも刺激リストの語彙や順番の通りに録音されないため、本処理方法を用いることでアーカイブデータを利用した今後の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（B）において利用可能な、過去の研究プロジェクトにて収集された豊富なアーカイブデータを取り上げ、これを用いた言語資源の整理とメタデータの設計、及び音声データの処理方法に関し、工夫した点について紹介する。共同利用型共同研究は、国立国語研究所の保有する多種多様な研究資料や言語資源等を使用して研究を行うことができる制度であり、著者らは、「国立国語研究所研究資料室収蔵資料」の「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・アーカイブされた録音音声データを用いて日本語の有声性に関する研究を

[1] michinorisuzuki19[at]gmail.com

実施している。本データベースは様々な日本語方言を対象にした大規模な録音実験に基づいて作られたものであり、1100 人以上の話者の豊富な録音データを含んでいる。同時に、アーカイブされている音声データは各被験者につき録音実験全体の音声ファイルのみであるため、まず初めに各方言の録音を聴き、どのような刺激の種類でセクションが分かれているかを確認することで刺激の ID 作成を試みた。しかしながら、方言間で録音されているセクションの種類や、同一セクション内での刺激リストなど、複数の点においてばらつきが散見されており、同一の刺激 ID 作成は困難であった。方言内においても話者ごとに刺激の順番や繰り返しの回数が異なっていた。そのため、録音をもとに Section ID、Word ID、繰り返しの回数、Speaker ID が識別可能な刺激 ID を効率的に作成する方法を考案した。本データベースのように、様々な方言及び年代の話者を対象とした録音実験を実施する場合、必ずしも一つの刺激リスト通りの刺激や順番で録音できない可能性がある。そのような場合でも、本稿で紹介する手順を用いることで、識別可能な刺激 ID を付与することで、様々なトピックの言語研究を行うことが可能となると考えられる。

２．使用しているデータベースについて

本節では、国立国語研究所の研究資料室収蔵資料「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」の実験録音データ全体について説明を行う。

２.１　被験者

「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」は、そのプロジェクト名の通り、日本語の韻律的特徴とその教育についての研究プロジェクトであり、日本語諸方言における様々な年代の話者の録音データが含まれている。各方言と各年代における話者の数を表 1 に示す。

表 1「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」プロジェクトにおいて収集された方言の種類、ファイル数、及び録音時間

方言	ファイル数	録音時間（時間:分:秒）
札幌	127	78:44:45
弘前	99	85:56:21
仙台	76	49:11:05
新潟	99	105:22:46
浅草	87	55:34:51
五日市	54	45:00:03
名古屋	83	60:41:20
富山	69	35:48:44
大阪	70	35:30:30
高知	77	51:49:09
広島	77	43:36:54
福岡	61	25:45:01
鹿児島	76	50:31:56
那覇	53	49:16:21
その他	120(43 都道府県)+148(琉球諸語辞典:奄美、石垣、今帰仁)	340:11:12
計：14 方言+その他	計：1376	計：1113:00:58

表 1 が示すように、本データベースには 1100 時間以上の豊富なデータが含まれていることが分かる。

２.２　刺激

本データベースに含まれる音声ファイルには、（1）に示したような刺激の種類（セクション）が存在する。

(1)　刺激の種類

a.　名詞
b.　動詞とその活用
c.　形容詞とその活用
d.　文章
e.　童話（桃太郎）の朗読
f.　天気概況
g.　五十音・数字（1 から 9 までと四桁の数字）
h.　固有名詞（地名・人名）
i.　会話

これらの刺激の種類は、オリジナルの研究プロジェクトの資料でも確認できた。しかし同時に、方言間で刺激のセクションが異なっていることや、刺激の重複や形式の一致により同一セクション内と考えられる部分でも異なる刺激が録音されていることがあると確認された。

２.３　実験手順

実際の録音では、実験者が被験者に渡す紙に書かれた単語や文章の読み上げや、絵を見せてその名前を言ってもらうという方法を取っていた。また、被験者の言い間違えや周囲の騒音が入った際には再度言い直しを依頼していた。

３.　データ整理

ここまで、使用したデータベースの説明を行なったが、本節ではデータ整理及び処理の際に困難であった点とその解決方法について述べる。

３.１　Speaker ID

アーカイブされていたエクセルにおいて、各音声ファイルに「通し番号」と「ファイル名（DAT ID）」が付けられており、これを参考に Speaker ID をつけていたが、1 つのファイルに複数の話者が録音されていることや、複数の音声ファイルに同じファイル名がついていることがわかったため、新たに話者が識別可能な Speaker ID を作成した。整理後の各方言での話者の人数は表 2 の通りである。

表 2「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」プロジェクトにおいて収集された話者の人数（方言ごと）[2]

	高齢層	壮年層	若年層	中学生	小学生
札幌	27	29	31	20	20
弘前	14	19	18	24	24
仙台	13	10	11	20	21
新潟	28	18	13	20	18
浅草	31	26	2	30	0
五日市	32	20	2	30	0
名古屋	12	14	12	25	20
富山	10	6	13	20	20
大阪	10	6	14	20	20
高知	12	13	12	20	20
広島	9	11	12	23	22
福岡	10	10	10	17	14
鹿児島	10	14	12	20	20
那覇	12	4	7	15	15
計	230	200	169	304	234

３.２　Section ID

2.2 節でも述べた通り、方言間での刺激のセクションのばらつき、及び方言内での順番や繰り返しの回数のばらつきが存在していた。そのため、複数の話者に対し単一の刺激リストをもとに識別可能な Stimuli ID を作成し、各刺激に対しラベリングを行うというデータ処理の手順をそのまま適用することができなかった。

そこで、まず全ての方言の録音を若干名ずつ確認し、「どの方言にどのセクションが含まれているのか」を確認した。その結果、「ほぼ全ての方言データに共通するもの」「一部の方言データにしか存在しないもの」「セクション自体は他の方言と共通しているものの、実際の刺激が異なるもの」が存在していた。これらのセクションに対し、「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」の研究資料を参考に Section ID を付与した。

３.３　Stimuli ID

方言ごとにセクションにバリエーションがあったものの、少なくとも同一方言データの同一セクションでは、刺激自体は共通であり、差があるのは繰り返しの回数のみであった。そのため、以下（2）の手順で Stimuli ID の作成を行なった。

(2)　本データベースに対する刺激 ID 作成の手順

a. まず一人の被験者の録音を参考に Excel に刺激のリスト（①）を書き出す
b. 音声分析ソフトウェア Praat (Boersma & Weenink, 2023) を使用し、各話者のデータに対して Praat スクリプトを使用した刺激間の境界の配置
c. Praat 上で各刺激の単語を Textgrid に手入力

[2] 表 1 に記載されているデータに加え、43 都道府県から 120 名分の録音データや琉球諸語のデータも存在する。

d. 別の Praat スクリプトを使用して入力した単語を取り出し（テキストファイル形式）、Excel（①と同じファイルの別シート）へ貼り付け
e. Speaker ID、繰り返しの回数、貼り付けたも刺激の番号を入力
f. VLOOKUP 関数で①を参照して word ID、ひらがな、英語訳を付与
g. CONCATENATE 関数で「Section ID － word ID － 繰り返しの回数 － speaker ID」を組み合わせる（参考：図 1）

A	B	C	D	E	F	G	I	J	K
ItemID	Speak	Word-SectionID	Section	Japane	wordID	repetit	Hiragana	English	Recording
A1-W002-1-KOC027	KOC027	A1-W002	A1	藤	W002	1	ふじ	wisteria	1
A1-W002-2-KOC027	KOC027	A1-W002	A1	藤	W002	2	ふじ	wisteria	2
A1-W002-3-KOC027	KOC027	A1-W002	A1	藤	W002	3	ふじ	wisteria	3
A1-W003-1-KOC027	KOC027	A1-W003	A1	鈴	W003	1	すず	bell	4
A1-W003-2-KOC027	KOC027	A1-W003	A1	鈴	W003	2	すず	bell	5
A1-W003-3-KOC027	KOC027	A1-W003	A1	鈴	W003	3	すず	bell	6
A1-W004-1-KOC027	KOC027	A1-W004	A1	地図	W004	1	ちず	map	7
A1-W004-2-KOC027	KOC027	A1-W004	A1	地図	W004	2	ちず	map	8
A1-W004-3-KOC027	KOC027	A1-W004	A1	地図	W004	3	ちず	map	9
A1-W005-1-KOC027	KOC027	A1-W005	A1	巻き寿司	W005	1	まきずし	Sushi rolls	10
A1-W005-2-KOC027	KOC027	A1-W005	A1	巻き寿司	W005	2	まきずし	Sushi rolls	11
A1-W006-1-KOC027	KOC027	A1-W006	A1	三日月	W006	1	みかづき	crescent moon	12
A1-W006-2-KOC027	KOC027	A1-W006	A1	三日月	W006	2	みかづき	crescent moon	13
A1-W006-3-KOC027	KOC027	A1-W006	A1	三日月	W006	3	みかづき	crescent moon	14
A1-W007-1-KOC027	KOC027	A1-W007	A1	頭巾	W007	1	ずきん	hood	15
A1-W007-2-KOC027	KOC027	A1-W007	A1	頭巾	W007	2	ずきん	hood	16
A1-W007-3-KOC027	KOC027	A1-W007	A1	頭巾	W007	3	ずきん	hood	17
A1-W008-1-KOC027	KOC027	A1-W008	A1	缶詰	W008	1	かんづめ	canning	18
A1-W008-2-KOC027	KOC027	A1-W008	A1	缶詰	W008	2	かんづめ	canning	19
A1-W008-3-KOC027	KOC027	A1-W008	A1	缶詰	W008	3	かんづめ	canning	20
A1-W009-1-KOC027	KOC027	A1-W009	A1	火事	W009	1	かじ	fire	21
A1-W009-2-KOC027	KOC027	A1-W009	A1	火事	W009	2	かじ	fire	22
A1-W009-3-KOC027	KOC027	A1-W009	A1	火事	W009	3	かじ	fire	23
A1-W010-1-KOC027	KOC027	A1-W010	A1	舵	W010	1	かじ	rudder	24
A1-W010-2-KOC027	KOC027	A1-W010	A1	舵	W010	2	かじ	rudder	25
A1-W010-3-KOC027	KOC027	A1-W010	A1	舵	W010	3	かじ	rudder	26

図 1：刺激 ID を作成した Excel ファイルの一例

作成した刺激 ID（及びひらがなと英語訳も）を Praat スクリプトでラベリングし、また別の Praat スクリプトを使用して個別刺激ファイルへの切り分け（chopping）を行う。その後、切り分けられた個別刺激ファイルに対しアノテーションを行った。この点に関しては、言語資源ワークショップ 2022 で発表したデータ処理方法（Suzuki, Igarashi, and Lee, 2022）を参照されたい。

４．おわりに

パンデミックを機に、オリジナルの実験データだけでなく、過去の研究で収集されたアーカイブデータなどの言語資源を利用した研究が以前よりも大きく注目されることとなった。しかし、アーカイブデータは過去の一時点における研究手法や概念に基づいて収集されたものであるため、必ずしも現在ほど再現可能性が高くなるような手順で行われていない可能性がある。また、そもそも産出実験を実施する際、必ずしも刺激リストの通りに録音できないことや、言い直しなどで刺激の順番が前後する可能性がある。このような一見処理しづらいアーカイブデータであっても、本稿で紹介した整理手順を応用することで再現性を高め、今後様々なトピックの言語研究へ活用が可能であると考えている。

また、本稿で紹介した手順は過去のアーカイブデータのみならず、今後の実験においても重要だと考えられる。識別可能な刺激 ID とともに個別刺激ファイルをアーカイブすることで、再現性を保ちつつ広く言語研究に応用可能なデータベースを作成することができるためである。

謝　辞

本稿は、国立国語研究所の共同利用型共同研究（B）プロジェクト「東北・東京方言における有声性の対立への音響指標の影響」（研究代表者 : 鈴木成典）の研究成果である。また、本研究は JSPS 科研費 23KJ1921 の助成を受けたものである。

文　献

Boersma, Paul & Weenink, David (2023). Praat: doing phonetics by computer [Computer program]. Version 6.3.14, retrieved 4 August 2023 from http://www.praat.org/

鈴木成典, 五十嵐陽介, & 李勝勲. (2023). NINJAL データベースを活用した言語研究の実施について. In 言語資源ワークショップ発表論文集= Proceedings of Language Resources Workshop (Vol. 1, pp. 79-82). 国立国語研究所.