少数言語のデジタルアーカイブ：PhoPhoNO と BantuDArc

李 勝勲（国際基督教大学）†
倉部 慶太（AA 研）
品川 大輔（AA 研）

Digital archives of understudied languages: PhoPhoNO and BantuDArc

Seunghun J. Lee (International Christian University)
Keita Kurabe (ILCAA)
Daisuke Shinagawa (ILCAA)

要旨

大言語を対象とした様々なデジタルアーカイブに基づく研究が進展する一方で、少数言語を対象としたデジタルアーカイブの構築とその利活用はまだ充分に進んでいるとはいいがたい。本稿では少数言語を中心に著者らが構築したデジタルアーカイブを紹介し、少数言語を対象としたアーカイブ化に関して議論する。一つ目はチベット・ビルマ系の5言語に関する資料を公開するアーカイブサイト 'PhoPhoNO'、もう一つはバントゥ系の5言語の資料をアーカイブ化したサイト 'Bantu Language Digital Archive (BantuDArc)' である。各サイトは言語に関するメタデータ、地図、そして言語資源から構成される。音声資料を含む個別のデータ項目には固有のIDが付与され、申請によってアクセスを認められれば、利用者はそれらデータを研究資源として利活用することができる。

１．はじめに

本稿は少数言語を対象にした二つのアーカイブサイトを紹介することを目的とする。少数言語を対象とする大規模デジタルアーカイブとして国際的に広く認知されているサイトには、イギリスの危機言語アーカイブ（ELAR: Endangered Languages Archive）、ドイツの DOBES アーカイブ（Documentation of Endangered Languages）、オーストラリアの PARADISEC（Pacific and Regional Archive for Digital Sources in Endangered Cultures）などが挙げられるが、それ以外にも近年では南アフリカの SADiLaR（South African Centre for Digital Language Resources; 南部アフリカ諸語を対象とする）やアメリカ・テキサス州の AILLA（Archive of the Indigenous Languages in Latin America; ラテン・アメリカの少数言語を対象とする）など、特定の地域や言語群を対象としたアーカイブも構築されている。日本では国立国語研究所が提供する日本語を対象とした例えば「危機言語データベース」などのアーカイブや東京外国語大学アジア・アフリカ言語文化研究所（AA 研）の情報資源利用研究センター（IRC）が管理・維持するアジア・アフリカ地域を中心とした言語文化に関するアーカイブがある。

アーカイブの構築に際しては、保存されているデータの質だけではなく、ユーザビリティの観点からデータの整理形式についても配慮する必要がある。さらに、データの共有範囲や

† seunghun@icu.ac.jp

データそのものに関する質的な説明も求められる。以下、二節では南アジアと東南アジアの言語をアーカイブする PhoPhoNO の、三節ではアフリカの 5 つのバントゥ諸語の言語資料アーカイブである BantuDArc の、データ構造とその意義について紹介する。

２．南アジア・東南アジア言語のアーカイブ（PhoPhoNO）

２.１　背景

このアーカイブのもととなるデータは、日本学術振興会国際共同研究事業（スイスとの国際共同研究プログラム） Phonetics, Phonology and New Orthographies in Roman and Indigenous Script: Helping Native Language Communities in the Himalayas (PhoPhoNO) によって収集された。2017 年から 2020 年までのプロジェクト期間中、インド・シッキム州のデンゾン語（Drenjongke）、ネパールのタマン語（Tamang）、ブータンのゾンカ語（Dzongkha）、ミャンマーのビルマ語（Burmese）など 4 つのチベット・ビルマ系言語、およびミャンマーで話されるモン・クメール系のモン語（Mon）の言語データを収集した。

本アーカイブは、それらフィールドワークによって収集した一次資料の音声データをアーカイブ化したものである。元のデータはそれぞれのプロジェクトの目的に従って処理されている。そこで、アーカイブ構築に際しては、アーカイブ化に適したデータ処理を行った。なお、録音に参加した調査協力者からはアーカイブ関連の承認を事前に取得済であった。アーカイブ化には AA 研の情報資源利用研究センター（IRC）から 2019 年度と 2020 年度に支援を受け、PhoPhoNO プロジェクトによって収集された音声データを処理することで、デジタルアーカイブを構築した（https://phophono.aa-ken.jp、図 1）。グーグルフォームによる申請によって研究目的の利活用であることが承認されれば、ファイル・サーバへのリンクが送信されるシステムになっている。

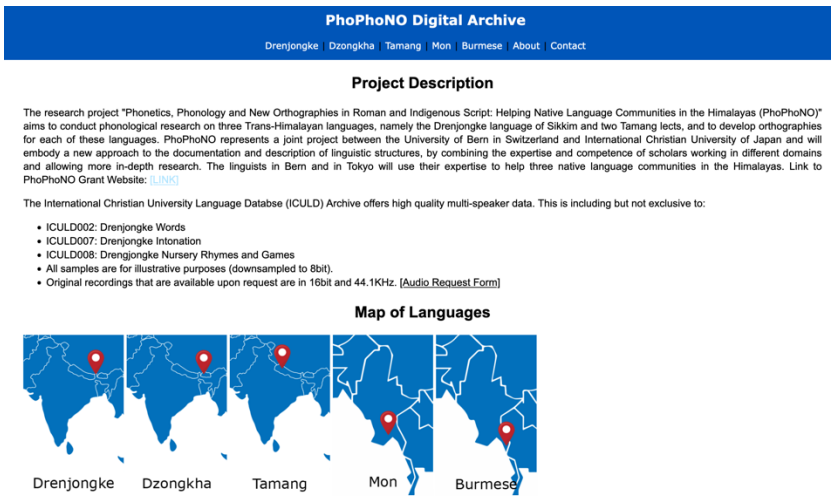

図 1 PhoPhoNO サイトのメインページ

表 1　PhoPhoNO アーカイブ内のデータベース

言語	ICULD	サイズ	長さ	内容
デンゾン語	1~9	2836.8 MB	314.7 分	音声・テキスト・プロット
ゾンカ語	10~11	99.7 MB	16.5 分	音声・テキスト・プロット
タマン語	12~15	9340MB	309.75 分	音声・テキスト・プロット
ビルマ語	32	5890MB	74.9 分	音声・テキスト・EGG
モン語	31	345MB	9.28 分	音声・テキスト・プロット

２.２　PhoPhoNO アーカイブの構造

このアーカイブは表 1 に示されている 5 つの言語のデータベース（ICU Language Database, ICULD）から構成されている。デンゾン語の内容については Baldoria et al. (2020a) および Kunzang Namgyal & Lee (to appear, 2022) に詳しく述べているので、ここでは他の 4 つの言語のアーカイブを中心に説明する。また Baldoria et al. (2020b) では 4 つの言語のトークンの確認ができる。

ブータンの公用語であるゾンカ語のデータベースには一人の話者が発話したゾンカ語の単語と音節（syllabary）の録音（ICULD-0010）と、ゾンカ語の個別の分節音を示すための録音（ICULD-0011）が収録されている。タマン語のデータベースには多数の話者による東部方言（ICULD-0012 と-0013）および西部方言（ICULD-0014 と-0015）の録音が収録されている。これによって、二つの方言の母音の発音又は単語の発音の相違を対照的に把握することが可能になる。ビルマ語のデータベース（ICULD-0032）には声調（tone）と発声（phonation）に関する特徴を示すための録音があり、唯一のモン・クメール系言語であるモン語のデータベース（ICULD-0031）には発声に関する諸特徴を検証するためのデータが収録されている。

２.３　PhoPhoNO アーカイブの意義

デンゾン語の場合、これまで研究利用が可能な音声データ自体が存在しなかったため、PhoPhoNO によって初めて、電子的に利用可能な言語資源のアーカイブが構築されたことになる。ゾンカ語、タマン語、ビルマ語、モン語についても、国際的に電子データの公開が多かれ少なかれ進められているものの、PhoPhoNO アーカイブは独自資料に基づくアーカイブであり、また既存データを補完する役割も果たす。将来はこのデータベースの資料を利用した分析論文が出版されることが期待される。

３．バントゥ系言語のアーカイブ：Bantu Language Digital Archive (BantuDArc)

３.１　背景

バントゥ諸語は日本における長期の研究蓄積を有する言語群の一つである。BantuDArc はその研究成果の一部としての音声資料を一箇所に集約することを目的とする。6 つの南部バントゥ諸語の形態統語論調査によって収集され、2021 年に公開された音声資料を含むアーカイブ（Bantu Microvariation Digital Archive; https://renelda.aa-ken.jp/about.html）、およびツォンガ語を含む統語構造の声調に対する影響に関するプロジェクト（ECPPT: Effects of Syntactic Constituency on Phonology and Phonetics of Tone" Digital Archive; https://ecppt.aa-ken.jp/sub.html）のアーカイブを作成した経験に基づき、2022 年に構築された。PhoPhoNO アーカイブ同様、BantuDArc も AA 研の IRC からの支援を受けて、データのアーカイブ化作業およびウェブサイト構築が可能になった。

BantuDArc は当初から収録言語を拡張することを目的に構築されている。現在は南アフリカのツォンガ語（S53）と東アフリカのルヮ語（E621A）、ベンデ語（F12）、ルンディ語（JD62）、さらにスワヒリ語（G42）の録音ファイルを申請ベースによってダウンロードすることが可能である。

３.２　BantuDArc アーカイブの構造

BantuDArc のデータベース構成は表 2 のとおりである。Praat で生成された各音声ファイルのプロット（具体的には spectrogram と pitch track）はウェブサイトから制限なしでダウンロードすることができる。音声ファイルに関しては、PhoPhoNO 同様、研究目的であることが申請によって確認できる場合のみアクセスが可能になる。データベースの内容は Johnson

et al. (2022)で確認出来る。

表 2　BantuDArc アーカイブ内のデータベース

言語	ICULD	サイズ	長さ	内容
ツォンガ語	41	724.2 MB	186.41 分	音声・テキスト・プロット
ルヮ語	44	229.3 MB	49.66 分	音声・テキスト・プロット
ベンデ語	42~43	966.2　MB	199.8 分	音声・テキスト・プロット
ルンディ語	46	137.7MB	31.57 分	音声・テキスト・プロット
スワヒリ語	45	333.2MB	96.51 分	音声・テキスト・プロット

ツォンガ語のデータベース（ICULD-0041）には一人の話者による発話がトークン化されて収録されている。多様なデータの内容に関する形態統語論レベルの記述情報は Lee et al. (2022) に整理されている。ルヮ語（ICULD-0044）には時制や相をパラメターとする多様で複雑な動詞の変化形が網羅的に録音されている。ベンデ語は 2 つのデータベースから構成されている。ReNeLDA プロジェクトのために収集した形態統語論に関するデータ（ICULD0042）に加え、ICULD-0043 にはさまざまな語形式が記録されている。具体的には、名詞の場合は単数形と複数形、また形容詞や指示代名詞によって修飾される構造が、動詞の場合は人称や名詞クラスなどの文法範疇に基づく活用パラダイムが録音されている。ルンディ語（ICULD−0046）は語の活用形や形態統語論的分析を目的とした例文データを収録し、ICULD-0045 にはスワヒリ語の付加詞（particle）のイントネーションを分析するために収集された発話の録音が保存されている。

３.３　BantuDArc アーカイブの意義

バントゥ諸語の記述的資料は多く蓄積されているが、バントゥ諸語に特化した録音資料をダウンロード出来るアーカイブはほとんど存在しない。そのような状況をふまえ BantuDArc はあらゆるバントゥ諸語を対象とし、研究者と現地コミュニティの人々の双方が言語資源にアクセスできる場を提供している。

現在、ツォンガ語に関しては Lee et al. (2022) に、ルンディ語に関しては Shinagawa et al. (2022) に、それぞれ収録データに基づく言語学的な分析結果が提示されている。これらの成果は、母語話者の著者との共同出版としての意義も併せ持つ。

４．おわりに

本稿では二つのアーカイブ PhoPhoNO および BantuDArc を紹介した。研究成果の一部としての録音ファイルを基盤に構築されたこれらウェブサイトは、より広範囲の研究者のみならず母語話者コミュニティに対する資源的貢献となることをも意図して構築されている。

謝　辞

本研究は東京外国語大学アジア・アフリカ言語文化研究所（AA 研）の情報資源利用研究センター（IRC）の支援を端緒とするものであり、本稿は科研費の国際共同研究加速化基金 (B)「Microvariation in Bantu languages of South Africa: Building theories from typology data」及び AA 研共同利用・共同研究課題「通言語的観点からみた音声類型論（PhonTyp）」の支援を受けて行われたものである。

文　献

Baldoria, Yukki, Audrey Lai, Hannah Lee & Tomoko Monou (eds.) (2020a) *ICU Working Papers in Linguistics 11: ICU Language Database Series 1: PhoPhoNO Digital Archive 1*. Tokyo, Japan: ICU LingLab.

Baldoria, Yukki, Audrey Lai & Rachel Liu (eds.) (2020b) *ICU Working Papers in Linguistics 12: ICU Language Database Series 2: PhoPhoNO Digital Archive 2*. Tokyo, Japan: ICU LingLab.

Johnson, Kiara, Sayaka Manabe, Chrisanne M. Tuaño. (eds.). (2022) *International Christian University Working Papers in Linguistics 21: Bantu Language Digital Archive*. Tokyo, Japan: IUC LingLab.

Kunzang Namgyal & Seunghun J. Lee (2022, to appear) Drenjongke (Bhutia) databases: a collaborative project with language experts. Ms.

Lee, Seunghun J., Babane Morris Thembhani & Madala Crous Hlungwani. (2022) *Aspects of Xitsonga Grammar*. Tokyo, Japan: ILCAA.

Shinagwa, Daisuke, Yuko Abe, Seunghun J. Lee & Chérubin Mugisha (2022) *Selected topics of Kirundi Grammar: A micro-typological perspective*. Tokyo, Japan: ILCAA.

関連 URL

PhoPhoNO アーカイブ	https://phophono.aa-ken.jp/
BantuDArc アーカイブ	http://BantuDArc.aa-ken.jp/