Vol.2021-CH-126 No.7
2021/5/22

消滅危機言語の辞書データベースの構築と公開：「鳩間方言 音声語彙データベース」、「うちなーぐち活用辞典 テキストデータベース」の事例報告

中川 奈津子[1,a)]　加治工 真市[2]　宮良 信詳[3,b)]

概要：本発表では消滅の危機にある日本の言語の辞書を構築・公開したプロセスを報告する。1 つ目は八重山諸島の鳩間島で加治工（第 2 著者）が約 60 年にわたって収集・記録した南琉球八重山鳩間方言の辞典、2 つ目は沖縄本島で話されている沖縄語で宮良（第 3 著者）が母語話者に確認を取りつつ完成させた沖縄語の用例を集めた辞典である。Microsoft Word で作成された非構造化テキストを中川（第 1 著者）が Python スクリプトと手作業で構造化したテキストデータベース（CC ライセンス）、それを LaTeX 形式、HTML 形式に変換した PDF 版、ウェブ版を公開した。またこの変換スクリプトも公開予定である。オープンサイエンスを志向し、コミュニティに研究成果を還元するためにすべてをオンラインで公開したが、言語調査の協力者であり理解者であり支持者でもある高齢者には特にアクセスが難しいことが多いため、書籍も印刷・製本し配布した。本発表ではこの事の利点と限界点も論じる。

Construction and publication of databases of endangered languages: Cases of *The Audio Database of Hatoma Lexicon* and *Text Database of the Practical Use of Okinawan*

1. はじめに

本稿では国立国語研究所で公開した消滅危機言語の辞書のテキストデータベース化、ウェブサイト作成、書籍の印刷と配布の流れを記す。まずテキストデータを一次資料としてクリエイティブ・コモンズ・ライセンスで公開し、これを元にウェブサイト（HTML 形式）や書籍（ここでは LaTeX 形式）に変換する手順を確立することで、研究用にも一般閲覧用にも使えるデータ構築を目指す。また、それぞれの引用方法を明確にして研究業績として引用されるようにした。

事例として本稿では、「鳩間方言 音声語彙データベース」(加治工・中川, 2021)、「うちなーぐち活用辞典 テキストデータベース」(宮良・中川, 2021) の処理過程とその形式変換の作成過程を述べる。前者は約 18000 項目、35000 例文、後者は約 2700 項目、6500 例文を含む。それぞれのデータベースは LaTeX 形式に変換して『鳩間方言辞典』(加治工, 2020)、『うちなーぐち活用辞典』(宮良, 2021) として書籍の形態で出版されている。（国立国語研究所 言語変異研究領域を出版者として、印刷・製本し、無料配布している。）「鳩間方言 音声語彙データベース」はさらに HTML 形式に変換し、音声とともにウェブ公開されている。[*1]

鳩間方言（ぱとぅまむに）は、沖縄県に属する八重山諸島の 1 つ鳩間島で話されている。鳩間島は西表島の北にある小さな島で人口は約 50 人であるが、[*2] 島外にも出身者がいるので鳩間の人というアイデンティティを持つ人は 50 人よりもっと多いといえる。日本語との相互理解性はほぼないので「鳩間語」と呼んでも良いが、慣習に従って「方言」という表現を用いている。「鳩間方言」を意味する「ぱ

[1] 国立国語研究所 (National Institute for Japanese Language and Linguistics) 190-8561 東京都立川市緑町 10-2
[2] 沖縄県立芸術大学 (Okinawa Prefectural University of Arts) 903-8602 沖縄県那覇市首里当蔵町 1-4
[3] 琉球大学 (University of the Ryukyus) 903-0213 沖縄県中頭郡西原町字千原 1 番地
[a)] nakagawanatuko@gmail.com
[b)] s7miyara@gmail.com

[*1] https://www2.ninjal.ac.jp/hatoma
[*2] https://www.town.taketomi.lg.jp/about/hatoma/, 2021/4/14 最終アクセス

とぅまむに」は直訳すると「鳩間物言い」であり、「鳩間言葉」というような意味合いである (加治工, 2020, p. 1305)。

沖縄語（うちなーぐち）は、沖縄本島の特に首里・那覇で話されている。宮良 (2021) で「沖縄語」という表現が用いられているので、本稿でもこれをそのまま踏襲する。「うちなーぐち」を直訳すると「沖縄口」で、「沖縄語、沖縄ことば」という意味合いであり、「やまとぅぐち」（日本語)、「とーぬ　くち」（中国語、「唐の口」）などと対比される (宮良, 2021, pp. 94–95)。

鳩間方言も沖縄語も、日常的に使用しているのは高齢者が多く、[*3] このまま何もしなければ消滅してしまう言語である。今からこの言葉を学習したい人や研究者らのために、これらの言語における単語の意味や用法が詳細に記述された両データベースは貴重な資料である。どちらの言語も弥生時代末期あるいは古墳時代に日本語祖語から分岐し、別々の言語変化を経て今に至っている (Pellard, 2015)。述語が文の最後に現れること、後置詞を用いること、名詞の前にそれを修飾する形容詞が現れることなど、基本的な類型論的特徴は日本語と似通っている。

2. 元ファイルからテキスト DB へ

第 2 著者と第 3 著者が作成した元ファイルはどちらも Microsoft Word の doc(x) 形式であったため、これをテキストファイルにコピー&ペーストして作業を開始した。『うちなーぐち活用辞典』のほうはかなりきれいに構造をとりだしやすかったため、ここではより大変だった『鳩間方言辞典』を例に説明するが、基本的に同じような手順とスクリプトで作業し、見出し語、発音記号、品詞、意味記述、例文という語学辞書の基本となる構成要素を取り出して構造化することを目標にした。『鳩間方言辞典』では、多くの辞書と同じように以下の要素が、この順番で記述されていた。(4) はあるときとないときがあり、現時点ではうまく取り出せておらず (5) の一部になっている。(5), (6) はこの順番で任意に繰り返しがあり、(6) 自体も繰り返しがあることがある（図 1参照)。

(1) 見出し語
(2) [　] で囲まれた発音記号
(3) 品詞（(　) で囲まれていることが多く、1 種類に限らず / で区切られていることがある）
(4) （　）で囲まれた意味分類
(5) 意味記述（丸囲み数字、カッコ囲み数字などが用いられている）
(6) 例文（カタカナ例文、発音記号例文、(　) に囲まれた

[*3] 70 歳以上の方言使用者の割合が最も高く、65.2%が方言を主にあるいは共通語と同じぐらい使っている。10 代で方言を使っているのは 9.2%のみである (沖縄県, 2021, p. 5)。(https://www.pref.okinawa.jp/site/bunka-sports/bunka/2019-shimakutuba-ishikichosa.html, 2021/4/22 最終アクセス)

「アー¬クン[「?a:¬kuN]（自動)。(助動)
①いる（居る)。あるく。ア「ラ¬クン[?a「ra¬kuN]（歩く）の転訛した形。¬ミサレーティ 「アー¬クン[¬misare:ti 「?a:¬kuN]（元気でいる＜暮らしている＞)。「ワー¬マナー 「アーク¬ワ[「wa:¬ mana: 「?a:ku¬wa]（君は何処に居るのか)。¬カナー「アー¬クンケン 「サンガリ パッ¬タヤー[¬kana: 「?a:¬kuNken 「saNgari pat¬taja:]（あそこに居るところを引っ張られていったよ)。¬アイブ ¬ントンナー「アー¬キ ¬ミサカー 「アー¬ク ¬クトー ナ「ルン¬ドゥ 「アーカン¬バルマ「シ¬ナー[¬?aibu ¬tonna: 「?a:¬ki ¬misaka: 「?a:¬kuku。to: na「run¬du 「?a:kam¬baru ma「ʃi¬na:]（あんな所に居てよければ居ることは出来るが、居ないほうが良いよねえ)。¬ウナー 「アー¬ケー ¬ミサムヌ[¬?una:「?a:¬ke:¬misamunu]（そこに居れば良いのに)。「アー¬キバ[「?a:¬kiba]（居れよ)。②助動。～ている。ア「サビアー¬クン[?a「sabi?a:¬kuN]（遊んでいる。遊びまわっている＜遊びあるく＞)。「ヨータビアー¬クン[「jo:tabi?a:¬kuN]（よたよたとふらついている＜あるく＞。千鳥足で歩く）

図 1　整形前の非構造化テキスト (Microsoft Word 形式)

例文の意味。文末は句点「。」で終わっていることが多い。）

カッコが閉じられていない、全角と半角が入り乱れている、句点がないなど小さな問題はあったが、基本的に忠実にこの順序と表記法を守って書かれており、そうでなければある程度自動でデータベース形式にするのは不可能であった。

第 1 著者が、上述のパターンを取り出す Python スクリプトを書き、スプレッドシートに貼り付け、1 行ずつ目で見て間違いがないか確認する、という作業を繰り返すことで、ある程度構造化されたデータベースの形式に変換した。この作業に最も時間がかかったので、辞書を作成するのは Word などの文書作成ソフトではなく、表計算ソフトのほうが向いていると思われる。また、方言をひらがなで書くかカタカナで書くかは議論の余地があるが、この作業をする上では方言が必ずカタカナで書かれている（正確には「ふ」以外）ということは自動化の重要な手がかりとして用いることができた。もし方言も日本語と同じくひらがなで書かれていたら、自動化はもっと困難であったと予測される。1 行 1 項目にできた時点で、それぞれの項目に自動で ID を付与した。この構造化のチェック過程で発見した、第 1 著者の裁量で修正できない箇所は、項目 ID と問題の箇所を別ファイルにメモしておき、後でその ID の項目のみを LaTeX 形式に変換・印刷して手書きで校正原稿をやりとりした。

自動処理にはいくつか問題があった。まず、(6) における 1 つの例文が 1 文とは限らず、1 つの例文中に「。」が入っている場合や、(　) 中の例文日本語訳の中にさらに (　) で説明がある場合など、例文の区切れ目の判定が困難であったため、これも 1 行ずつ目で確認して 1 つの例文中の「。」をすべて別記号に置き換え、(　) の中にさらに (　) が入っている場合は別の括弧に置き換えた。ルビは漢字の後に半角カタカナで書かれており、これも正規表現検索を駆使しつつ、1 つ 1 つ確認しながらルビの箇所を同定してタグで囲んだ。また漢文中に返り点「レ」「一」「二」な

どが現れていることがあり、発見したものには返り点タグを付けたが、見逃しもあるだろう。(LaTeX で返り点を下付き文字として表示しようとしたが、実現できなかった。)[ ] の中には基本的には発音記号が入っているため発音記号のタグを付けたが、たまに日本語が入っていることもあるため別の括弧に置き換えた。さらにアクセント記号としてカギカッコ(「)、あるいは別の記号が使われている箇所があったので、これも正規表現を駆使してなるべく修正した。(カタカナや発音記号に囲まれている 「 などの条件で検索した。)

番号付与の問題もあり、基本的に丸囲み数字は意味別の記述、( ) で囲まれた数字は品詞別の記述に対応しているものの、一貫性がない箇所もあったので上記の方針に基づいて修正した。番号が付いているものは他にも、民謡の歌番号(1 番の歌詞、2 番の歌詞... など)、詳細な説明があるときの分類、など多岐にわたっており、これも目で見て気づく範囲内で別のタグを付けて区別した。タグの一覧は加治工・中川 (2021) 付属の readme ファイルに記載されている。

ある項目の品詞は 2 つ以上のこともあり、その場合は「(他動/自動)」のように / で区切られていたが、その後の記述の順番と対応していない、そもそも意味番号が分かれていない場合などがあり、第 2 著者に確認しつつ意味番号を分け、意味記述の順番を揃えた。この結果、データベースとしては品詞別に意味記述と例を取り出すことができた。また、品詞一覧も作成できたので、品詞名の過不足や一貫していないところを発見するなど、校正にも役立った。

以上をまとめると、自動処理のための作業は、辞書における要素の役割を一貫して明確に区切るタグ付けの作業であったといえる。第 1 著者が辞典の著者である第 2, 3 著者にいちいち確認して修正すると膨大な時間と手間がかかるため、括弧や数字などの本文の記述内容に寄与しない部分、文や項目の順序入れ替え、明らかな誤植と思われる箇所に関しては第 1 著者の裁量で修正して良いとの許可を得て修正した。(バージョン管理を行いつつ修正しているので、元ファイルと修正箇所の違いはいつでも比較可能な状態ではある。)自動処理にあたっては、最初にコピーしたファイルの行数や文字数と違いがないか、違いがあるとして説明できる違いなのかどうかを慎重に考えながら作業した。

また、『鳩間方言辞典』の場合は、構造化作業と並行して見出し語と例文のリストをそれぞれ作り、第 2 著者の読み上げを録音して音声データベースも構築した。見出し語の音声のみテキストデータベースの ID と関連付けて公開済み、[*4] 例文音声も録音済みで切り出し作業中である。

2 つのデータベースはテキストファイル形式で国立国語研究所のリポジトリから公開した。『鳩間方言 音声語彙データベース』は CC BY-SA、『うちなーぐち活用辞典テキストデータベース』は CC BY-ND のライセンスをそれぞれの著者と協議して付与した。

3. テキスト DB から書籍・ウェブサイトへ

テキストデータベースが完成し、見出し語、発音記号、品詞、意味記述、例文を区別して取り出せるようになったあと、新たな Python スクリプトによりこのデータベースを LaTeX と HTML の形式に変換した。並行して校正作業を進めていたが、常に元のテキストデータベースを修正し、LaTeX と HTML ファイルを直接修正することはしなかった。これにより常にデータベースが最新の状態であることを前提に作業ができた。

この辞書のソースファイル自体は、それぞれの辞書に特化しているため公開していないが、より一般的な目的で使えそうなソースを、オンラインで LaTeX ソースがコンパイルできる Overleaf のテンプレートとして公開している。[*5] 同時に、セルを適切に埋めていけば自動的に上述のテンプレートに合わせたソースコードが出力される Microsoft Excel ファイルも作成・公開し、使い方の講習も行った(日本学術振興会 DC/東京外国語大学 加藤幹治氏主催)。[*6]Overleaf のテンプレートは、日本語、音声記号、アルファベットにしか対応できていなかったが、このテンプレートを使って中国語の簡体字やタイ文字を入力したいという要望があり、個別に対応している。また、Excel の関数を使ってソースコードを出力するというアイディアに着想を得て、ウェブサイトを作った人もいる。

テキストデータベースから HTML ファイル群が生成できる Python スクリプトも公開予定である。

4. テキストデータベースの利点と問題点

本節ではオープンライセンスで公開する方言辞典のテキストデータベースの利点と問題点について論じる。

4.1 先行事例

書籍として出版されている方言辞典は数多くあるが、デジタルで公開されているものはあまりない。デジタルで公開されているものも多くは HTML 形式のみで、ライセンスが不明である。「しまむに宝箱」[*7]、「危機言語データベース」 [*8] などは日本の方言を知り、音声も聞くことができる貴重なウェブサイトはあるが、著者らの知る限り HTML 版しか公開されておらず、データを一括ダウンロードできないため、検索性が下がり、統計分析も困難になる。また、

[*4] https://doi.org/10.5281/zenodo.4560935

[*5] https://www.overleaf.com/latex/templates/language-dictionary-template/hvtyccvfbyrw

[*6] Excel ファイルの入手、講習の録画は以下のウェブサイトから: https://sites.google.com/view/fieldworkertech/home

[*7] https://www.erabumuni.com/

[*8] http://kikigengo.ninjal.ac.jp/

後者は複数のデータ作成者がいるはずだが、誰がどの部分にどのような責任で貢献しているのか明示されていない。ライセンスが明示されていないので、例えば自主的に整形したデータがあっても、許可なしで公開することはできない。「琉球語音声データベース」[*9] は新サイトの準備に向けて閉鎖中であるが、旧サイトは上述したような問題点があった。非研究者による方言の語彙に関するウェブサイトも数多く存在するが、同様の問題を抱えている。[*10]

先駆的なものは、『沖縄語辞典』(国立国語研究所, 2001a) をもとにした『沖縄語辞典 データ集』(国立国語研究所, 2001b) である。底本である『沖縄語辞典』も PDF で無料公開されており、データ集は CC ライセンスが付与されている。Excel 形式で一覧性が高く、列ごとの意味も明確で、検索も容易である。ただし著者、引用方法などが不明確なところに問題点が残るが、それらは底本に準ずると考えて良いだろう。フランス国立科学研究センターの Pangloss[*11] は HTML 版（と PDF 版）がオープンアクセスで公開されており、著者が明確で引用方法も明示されている。The UCLA Phonetics Lab Archive[*12] は語彙というよりは音声に興味があるアーカイブであるが、世界の多くの言語の語彙集が音声ファイルとともにクリエイティブ・コモンズ・ライセンスで公開されており、引用方法も明確である。しかし個々の語彙リストの著者は “UCLA students” とだけ書かれていて明らかでない場合もある。いずれにせよ今回調べた限りでは、オープンアクセスではあるが HTML 形式のみ公開されており、正規表現検索などを行おうとすると HTML ファイルを加工して使うことになり、やや手間がかかる。形態素分析など自然言語処理が盛んに行われている大言語ではテキスト形式のオープンな辞書はかなり充実しており、例えば日本語では UniDic (伝ほか, 2007; 小木曽・中村, 2014) が CC BY-NC-SA, GPL/LGPL/BSD のライセンスで公開されている。[*13] コプト語の辞書 Coptic Dictonary Online (the Koptische/Coptic Electronic Language and Alliance) は、ウェブサイトに加えて TEI 準拠の XML データが CC BY-SA のライセンスで公開されており (Burns et al., 2019)、汎用性が高い。

本稿では、どのような言語データも基本的には、UniDic やコプト語の辞書のようにオープンなテキストデータを一次資料としてリポジトリに公開して readme ファイルに引用方法を明示し、それを元に書籍版フォーマット（例えば LaTeX 形式）やウェブサイト（HTML 形式）に変換することを提案する。これにより、多様な層に見やすく、しかも研究者もデータを分析しやすいデータの両方を公開することができる。以下の節では、この手順の利点と問題点を議論する。

4.2 利点

一次資料としてテキスト形式で辞書を公開することで、まずは研究者が使いやすいという利点があげられる。適切なエディタを使えば正規表現を使って検索でき、例えば特定の音素配列のパターンを取り出したり、アクセントのパターンを見たり、研究の可能性が広がる。ある表現の用例を、見出し語の直下にある例文だけでなく、辞書全体から探すこともできる。日本語による用例検索もでき、逆引き検索（意味による項目の検索）の代替として使えるので語学学習用にも便利である。また、テキスト形式のファイルは機械で処理しやすく、様々な形式に変換して使うことができる。

作成者側の視点としては、ファイル管理が楽になることがあげられる。古いバージョンとの差分を見ることも容易で、ワードプロセッサで作成したときと異なり見た目を整える労力をなくすことができ、意図しない編集記号が挿入されることもない。様々な OS で同じように扱うことができるのも便利である。さらに、1 つの元ファイルを管理し、それだけに変更を加えるというルールを徹底し、その他の形式は元ファイルから自動生成することで、修正の手間も大幅に削減できる。Python などのオープンなソフトを用いてファイルを生成・変換しており、LaTeX もフリーで組版できるので限られた資金で作ることができる。[*14]

4.3 問題点

リポジトリ公開のテキストデータの問題点として、機械の操作が苦手だったりインターネット接続できないとアクセスできない点があげられる。特に、方言調査の貴重な協力者、理解者、支持者である高齢者がアクセスしづらいのは大きな問題であり、誰でも自由にどこからでも閲覧できるというオープンアクセスの利点が生かされない。インターネットに接続された機械でのみアクセスできるデータはまだまだ多くの人を排除してしまっている。したがって、書籍版を印刷して配布することは必須である。この利点と問題点については次節で議論する。

素人がウェブサイトを作ることの問題点もある。いつでも自分の意志でウェブサイトの更新が可能であり、コンテンツ追加・機能追加が随時できるという点は大きな利点であるが、セキュリティ対策も自分で行わなければならず、ありうる危険に備えて対処できるとは言い難い。危険を回避するためには JavaScript などの使用をやめれば良いが、見

[*9] 旧URL(休止中): http://ryukyu-lang.lib.u-ryukyu.ac.jp/
[*10] ジャパンナレッジに入っている『日本方言大辞典』(小学館辞典編集部, 1989) は便利であるが、有料サービスなのでここでは議論しない。
[*11] https://pangloss.cnrs.fr/
[*12] http://archive.phonetics.ucla.edu/
[*13] https://unidic.ninjal.ac.jp/
[*14] 第 1 著者の人件費はかかっているが、元ファイルだけ作って組版とウェブ作成を外注するよりは安上がりである。

出し語クリックでの音声再生、アコーディオンなど、HTMLのみで実現するのは困難な機能がある。ユニバーサルデザインやウェブアクセシビリティ規格（JIS X 8341）をどこまで達成できるのかという懸念もある。Mother Tongues (Littell et al., 2017)[*15] のような洗練されたオンライン辞書作成のコードが公開されているので、それを応用すべきかもしれない。しかしセキュリティ対策は自分で行わなければならないだろう。

テキストデータベース自体の本来的な問題ではないが、国語研リポジトリの通信環境の制限により、誰でもダウンロードできる音声ファイル一式は外部サイトに置いた。[*16] 公開箇所がバラバラになってしまい、関連性が不明確になりやすいので、気をつけなければならない。

5. 書籍印刷・配布の利点と問題点

一次資料であるテキストデータベースを LaTeX に変換し、印刷・製本して希望者に無料配布した。本節ではこのことの利点と問題点を論じる。

5.1 利点

書籍版の利点としてまず、多くの人に親しみやすいということがあげられる。デジタルデータを扱ったほうが便利な研究者でも、書籍版も欲しいと望む声が多かった。自分の地域の方言が立派な装丁の書籍になれば誇らしい気持ちになるし、目につくので学習機会も増えるかもしれない。さらに、データを公開しただけはニュースにならなさそうだが、書籍出版はニュースになる。『鳩間方言辞典』は地元沖縄の新聞である沖縄タイムスと琉球新報、そして八重山毎日新聞に取り上げられ、これにまつわるコラムも掲載された。[*17] 『うちなーぐち活用辞典』は出版されたばかりなのでまだ新聞には載っていないが、興味を示している新聞記者がいると聞いた。第 2 著者（加治工）は琉球新報活動賞（出版活動部門）を受賞したが、[*18] これも書籍を出版してこその成果であると考えられる。

研究成果は無料で公開する原則のもと、印刷物も無料で配布したので、高価な辞典を買わない人たちにも手にとる機会が増えたのではないかと考えている。

5.2 問題点

一方で、このような形で書籍を配布する様々な問題点も明らかになった。出版社を通して出版しないことで、書店に書籍が流通せず、関係者と直接・間接の知り合いでなければ本を入手できない状態になってしまった。[*19] 沖縄県には（他の多くの地域と同じように）地元の情報を敏感に察知して情報を発信している出版社、地元関連書籍を積極的に集めている書店が数多く存在する。このような地元出版社や書店と連携できればさらに良かったのではないかと考えている。実際、地元出版社から出版することも視野に入れて議論もしていたが、やはり出版社独自の予算で印刷するのはかなり厳しいとのことだった。出版助成を獲得するなど、書籍を出版・販売しても良い予算で、地元出版社から本を出し、それを書店に流通させるのが理想であると思われる。あるいは印刷費と送料のみを書籍代金としたクラウドファンディング（実質はほぼ事前予約）を実施しても良かったかもしれない。しかし多くのクラウドファンディングはインターネットを通じて行われており、アカウント作成、メールアドレス登録などの作業が必要である。これもまたインターネットに接続された機器を持たない人たちの排除につながってしまうのは悩ましい問題である。

書籍は物理的な存在なので、大量に印刷すると場所をとること、在庫管理、発送などの手間がかかることも大きな問題であるが、出版社を通して出版するとこれは解決する。ただし沖縄の出版社の 1 つに問い合わせたところ、内地（日本本土）への流通システム確保が難しいかもしれないとのことで、出版社に丸投げすればすべて解決というわけでもなさそうである。Amazon などでオンデマンド印刷する方法もあるかもしれない。オンデマンド印刷は頁数に上限があるので、頁数の多い辞書の類はやむを得ず分冊することになるだろう。

紙媒体の本来的な問題として、音声を再生できない、数に限りがある、印刷したらそれだけお金がかかるなどの問題もある。このような問題はどこからどのように出版しても解決しないので、やはり機関リポジトリに公開された電子データはこの点でも重要な役割を果たす。

プロフェッショナルな校正を活用できるというのも出版社を通した書籍出版の利点である一方で、筆者らの知る限り校正原稿を紙でやり取りする習慣が根強く、出版社がオリジナルのデジタルデータを持っている場合、出版社側で編集されてしまい著者側にオリジナルのデータが残らない可能性もある。完成原稿が PDF 形式、あるいは有料ソフトでなければ開かないような形式に変換されてしまうと、その後の扱いに苦労するので、元データは必ず著者側も確保できるように交渉する必要があるだろう。

6. おわりに

本稿では『鳩間方言 音声語彙データベース』、『うちなーぐち活用辞典テキストデータベース』の事例をもとに、テキスト形式で構造化された一次資料をオープンライセンスで公開し、書籍版やウェブ版はこの一次資料から自動変換

[*15] https://mothertongues.org/
[*16] https://doi.org/10.5281/zenodo.4560935
[*17] 2020/7/21 沖縄タイムス、2020/1/8 琉球新報など
[*18] https://ryukyushimpo.jp/news/entry-1296828.html, 2021/4/15 最終アクセス
[*19] 国立国語研究所に電話やメールなどで問い合わせれば書籍がもらえるが、問い合わせる人はかなり限られていると思われる。

して作成する手順を提案した。これにより、引用方法が明らかになり、研究にも教育にも利用しやすいデータになると主張した。

今後は、Coptic Dictionary Online の TEI 版を見習って TEI 形式の辞書データも作り、より一般的な語学辞書としての構造化を目指し、書籍版、ウェブ版の変換スクリプトも一般的に使えるようにしていきたい。また、データベースの作成者、構造化した作業者、エンコーダーは著者になれるのかどうか等、言語学の領域内部で確立していく必要のある慣習が数多く存在する。それに関しても今後議論して合意を形成していければと考えている。

Thanyehténhas Brinklow et al. (2019) が指摘する通り、テクノロジーの利点は成果物よりもそのプロセスを重視することで得られる。データ公開やその他言語に関わるテクノロジーによって、地域の人々の言葉が自分たちのものであることが阻害されてはならない。歴史的経緯を踏まえ、地域の目標に寄り添うことで、地域のためにもなり研究者も使えるデータ公開方法を模索していきたい。

謝辞 本研究は、国立国語研究所共同研究プロジェクト「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」(代表: 木部暢子)、「博物館・展示を活用した最先端研究の可視化・高度化事業」、科研費「日琉語族の語順の変異とその相関変数の解明」(JP21H00352, 代表：中川奈津子) の助成を受けたものです。本稿執筆に当たり、大向一輝氏、永崎研宣氏、宮川創氏をはじめ人文情報学勉強会の方々に、また五十嵐陽介氏ら NINJAL サロン参加者の方々にアドバイスを多数受けました。データ公開やウェブサイト作成の方針などに関しては籠宮隆之氏に、Windows での Python スクリプトの実行に関しては大村舞氏、浅原正幸氏に相談に乗っていただきました。全員の名前を記すことは出来ませんが、感謝します。

参考文献

Burns, Dylan Michael, Frank Feder, Katrin John, and Maxim Kupreyev (2019) “Comprehensive Coptic Lexicon: Including Loanwords from Ancient Greek v 1,” https://doi.org/10.17169/refubium-2333.

伝康晴・小木曽智信・小椋秀樹・山田篤・峯松信明・内元清貴・小磯花絵 (2007) 「コーパス日本語学のための言語資源：形態素解析用電子化辞書の開発とその応用」,『日本語科学』, 22, pp. 101-123.

加治工真市 (2020)『鳩間方言辞典』, 国立国語研究所言語変異研究領域, 東京, http://doi.org/10.15084/00002991.

加治工真市・中川奈津子 (2021) 「鳩間方言音声語彙データベース」, http://doi.org/10.15084/00003209.

国立国語研究所 (2001a) 『沖縄語辞典』, 財務省印刷局, 東京.

——— (2001b) 「沖縄語辞典データ集」, URL：https://mmsrv.ninjal.ac.jp/okinawago/.

the Koptische/Coptic Electronic Language and Literature International Alliance“Coptic Dictionary Online,” https://coptic-dictionary.org/.

Littell, Patrick, Aidan Pine, and Henry Davis (2017) “Waldayu and Waldayu Mobile: Modern digital dictionary interfaces for endangered languages,” in *Proceedings of the 2nd Workshop on the Use of Computational Methods in the Study of Endangered Languages*, pp. 141–150, Honolulu: Association for Computational Linguistics, March, DOI: 10.18653/v1/W17-0119.

宮良信詳・中川奈津子 (2021) 「うちなーぐち活用辞典テキストデータベース」, http://doi.org/10.15084/00003211.

小木曽智信・中村壮範 (2014) 「『現代日本語書き言葉均衡コーパス』形態論情報アノテーション支援システムの設計・実装・運用」,『自然言語処理』, 21 (2), pp. 301-332.

沖縄県文化観光スポーツ部文化振興課部 (2021) 「令和元年度しまくとぅば県民意識調査」, Technical report, 沖縄県, 沖縄.

Pellard, Thomas (2015) “The linguistic archeology of the Ryukyu Islands,” in Heinrich, Patrick, Shinsho Miyara, and Michinori Shimoji eds. *Handbook of the Ryukyuan Languages*, Berlin: Mouton de Gruyter, pp. 13–37.

小学館辞典編集部（編） (1989)『日本方言大辞典』, 小学館, 東京.

Thanyehténhas Brinklow, Nathan, Patrick Littell, Delaney Lothian, Aidan Pine, and Heather Souter (2019) “Indigenous Language Technologies & Language Reclamation in Canada,” in *Proceedings of the 1st International Conference on Language Technologies for All*, pp. 402–406, Paris, France: European Language Resources Association (ELRA), December.

宮良信詳 (2021)『うちなーぐち活用辞典』, 国立国語研究所言語変異研究領域, 東京, http://doi.org/10.15084/00003210.