W-2-2

デジタル化とアーカイブ化の実相：コイサンと琉球の事例

加藤幹治 (東京外国語大学大学院・日本学術振興会)
jiateng.ganzhi@gmail.com

1　はじめに

発表者は、JSPS 科研費 JP18H00661 「研究職を離れた言語研究者が保持する言語データの適正再資源化のための基盤確立研究」（代表: 加藤重広）のプロジェクト（以下、本プロジェクト）の一環として菅原和孝京都大学名誉教授が記録したグイ語の談話資料（以下、菅原資料）のアーカイブ化に携わってきた。詳しくは本ワークショップの 6 番目の中川・加藤・木村の発表に譲るが、総語数約 56 万語という大部の資料の約 7 割をテキストに起こし、その一部に形態素情報を付けている。大きい資料であるから、処理に時間がかかり、年度をまたげば処理にあたる人員が交替することが予想される。そのため、作業を自動化して処理する時間を減らす、作業の属人化を防ぐなどの工夫が必要であった。また、電子化したデータをどのようにして活用するかも課題である。本発表ではそうしたアーカイブ化作業の中で得られた知見を紹介する。また、発表者はフィールドワークによって琉球語族奄美語徳之島方言のデータを独自に収集している。本プロジェクトによって得られた知見の徳之島方言に対する適用事例も若干紹介する。

2　資料の電子化

菅原資料をアーカイブ化する手順を簡単に述べる。

まずフィールドノートをスキャンして pdf 化する。次に、pdf を人力で電子テキストに起こす[1]。電子テキスト化した資料はそのままではただの文字の羅列なので、文に訳を付け、句を形態素に分割し、それぞれの形態素に意味・機能を付す。本プロジェクトでは、音声や動画に注釈をつけるアプリケーションである ELAN を用いて形態素境界などの情報を付与した。談話が ELAN に入力された状態から §3.2で述べる環境を用いて資料として提示できる状態にする。

以下では、資料の電子化の際に用いた環境について述べる。

2.1　専用キーボード作成による入力の効率化

菅原資料には、国際音声記号 (IPA) の文字が多く含まれる（例えば、吸着音 ǀ ǁ ǂ !)。これらの IPA の記号は通常の打鍵では入力できない。SIL International の開発したキーボードを利用すれば IPA が入力できるが、ほぼ全ての記号を網羅してしまっていることが仇になり、一部の記号の打鍵が非常に複雑になってしまっている。外部のウェブサイトや IPA 入力用アプリケーションを利用して入力することもできるが、クリックやコピー・ペーストの操作を要し、入力に時間がかかる。これらを用いると書き起こしに時間を消費してしまう。したがって、グイ語の属するカラハリ盆地言語帯 (Kalahari Basin Area: KBA) で用いる記号を不足なく直感的なキー配列で直接入力できるよう、KBA 専用のキーボードを作成した。図1はシフトキーを押下した状態の Mac 版 KBA キーボードの配列である。無声口蓋垂摩擦音 χ は x と似ているので Shift + x で入力できるようにするなど、直感的に入力できるようになっている。

[1] 大部分はアルバイトの学生による人力の入力であるが、一部、次に述べる Transkribus で処理を行った

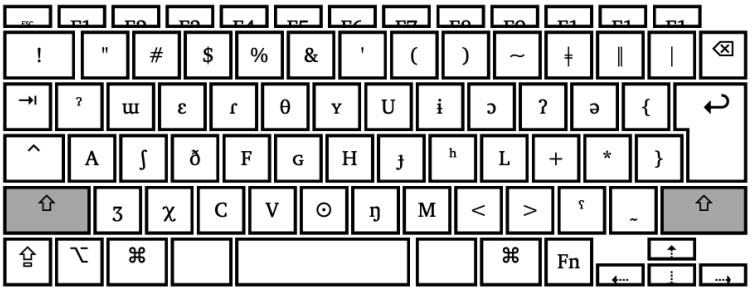

図1　KBA キーボードの配列（シフト押下）

2.2　Transkribus による手書きテキスト認識

Transkribus (`https://readcoop.eu/transkribus/`) は、簡単な GUI によって、機械学習に関する知識なしに手書きテキスト認識のモデルを作成できるツールである。Transkribus を用いて、菅原資料を自動で認識するためのモデルを作成した。このモデルを用いることで手書き文字を自動で読み取る事ができ、手動の入力よりも早く文字を起こすことができる。同一の書き手による資料が大量にある場合、Transkribus を用いることで、効率的にテキストの電子化が行える。

本プロジェクトで作成したデータは、現時点では菅原氏による資料のみが学習の対象となっているため、モデルは菅原氏の筆跡に適合した状態である。今後なんらかのグイ語手書きテキストが現れた際、今のままではそのテキストを読み取ることができない。今後資料のバリエーションが増えれば、それを学習データとしてグイ語テキスト一般の読み取りモデルを作ることができるかもしれない。

2.3　ELAN の活用

ELAN では辞書情報があればテキストに自動でアノテーションを行うことができ、本プロジェクトではこの機能を用いて形態素情報の付与を行っている。細かく注釈層を分けることで、§3.2で後述する Python による処理を行いやすくしている。発表者は、本プロジェクトで作成した ELAN テンプレートを用いて徳之島方言の書き起こしを行っている。その成果には加藤（印刷中）などがある。グイ語は比較的孤立的で ELAN によるアノテーションが容易だが、屈折の激しい言語（例えばラテン語）では容易でなく、どのような言語でも同様に処理できるようになるかは今後の課題である。

3　電子化資料の利用

電子化した菅原資料の研究への利用について述べる。

3.1　音頻度分析アプリケーション

本プロジェクトの一貫として、音頻度分析 web アプリケーション (Phoneme Frequency Analyzer: PFA。`https://phoneme-frequency-analyzer.herokuapp.com/`) を作成した（加藤 2020）。PFA はブラウザ上で動くツールであり、アクセスすることで誰でも使用することができる。Python の Flask フレームワークを用いて作成し、Heroku というサーバーで公開した。

音韻類型論では、世界の言語における音の出現の仕方の傾向や、個別言語における傾向と世界の言語の傾向が問題になる (cf. Everett 2018)。(i) 量が十分に多い、(ii) 音声表記か音韻表記か、あるいはそれらに近い表記である、(iii) 類型的な特異性をもつ、という条件を満たす資料があれば、それは音韻類型論に新たな知見を加えうるという点で重要な資料である。我々がアーカイビングを行っている菅原資料は、これらの条件を満たす資料である。現時点での電子化済み語数から想定すると、総語数は約 56 万語になると予想される。また、菅原氏による表記はほぼ一貫しており、少し置換を行えば音の生起頻度を計測するのに適した表記になる。さらに、グイ語は非常に大きな子音目録を持っていること、吸着音という特異な音類を持っていること、語根における音類の配列に不均衡が見られることなどが知られており（中川 2021）、類型論的に稀な性質を持つ。

このように音韻類型論の資料として適切なグイ語資料の音素頻度を計測するため、PFA を作成した。使用者は、分析対象のテキストと音ラベルの一覧を入力することで、対象のテキストに各音が何回生起したかを出力として得る。例えば、/watashi/ というテキストに対して /w/, /a/, /t/, /sh/, /i/ をラベルとして与えると、/a/ は 2 回、/sh/ は 1 回……という出力が得られる。PFA は以下のような特徴を持つ：各ラベルに用いられる文字が重複している場合、その重複を排除して頻度を計測できる。例えば、「刺し身」/sashimi/ というテキストに対して /s/, /a/, /sh/, /i/, /m/ というラベルを与えることを考える。単純に頻度を計測した場合、テキスト中に s の文字は 2 回生起するので、/s/ の生起頻度は 2 回と判定される。ところが、このテキストでは 2 つ目の s は/sh/の一部として判定することが望まれるので、単純に検索してヒットした数を返すのではなく、重複して数えた分を減じた数を返す必要がある。したがって、PFA では適切にラベルを与えれば重複を考慮して生起頻度を出力するようにした。次に、PFA 複数のラベルを一つの音にまとめることができる。例えば、先程例に挙げた「刺し身」/sashimi/ の分析では、/s/ と /sh/ を別のラベルと分析するようにしたが、これらをひとまとめにして計測することができる。すなわち、このテキストに対して /s または sh/, /a/, /i/, /m/ というラベルを与えることで[2]、/s または sh/ が 2 回、/i/ が 2 回……という出力を得る。

残念ながら PFA を用いた研究で公開されているものはまだないが、各言語の研究者が自身で得たか、あるいは「発掘」したデータの分析をする上で役立つことが期待される。

3.2 資料の公開のための文書処理環境の作成

書き起こした資料は、将来的に形態素情報や訳を付けて公開する必要がある。56 万語という大量のデータを処理する必要があるので、形態素情報の付加や公開資料の形式作成はできるだけ自動化することが求められる。本プロジェクトでは、ELAN、Python、TEX という三つのツールを用いてこれらをできる限り自動化する。

菅原資料が一旦 ELAN に入力されることは既に述べた。ELAN には注釈結果を様々な形（例えば、タブ区切りテキスト）で出力する機能が備わっている。しかし、ELAN 標準の出力は言語資料として論文の形で提示するために十分に整理された形式であるとは言えない。また、仮に ELAN で出力されたものを Microsoft Word で提示するとしても、適切な形式へ整形するのには膨大な時間がかかる。また、一旦結果を出力した後に注釈を変更した場合、その都度出力を行って、結果を論文の形へ再度整形する必要がある。

これらの欠点を克服するため、本プロジェクトでは、ELAN のファイル (`.eaf`) をプログラミング言語 Python と文書組版ツール TEX によって直接処理し、適切に整形された言語資料（`.pdf`）を

[2] 実際の PFA の記法では、中括弧 {} で複数のラベルを括ることで同じラベルとして計測する。

自動で[3]出力する環境を作成した。これによって、ELAN のアノテーションの変更の反映が自動的に資料に反映され、形態素に対するグロスの割付（インターリニアグロス）も自動で適切に整形される。作成した環境の技術的な詳細は省くが、大まかな流れは以下のようである：`.eaf` ファイルは xml 形式で書かれているので、これを Python の xml 処理ライブラリで読み、形態素・グロス・訳・時間情報などを抽出する。抽出したデータを Python で TEX の形式 (`.tex`) へ整形し、それを組版して資料の pdf ファイルを作成する。

同様の環境を用いて、以下の論文が作成された：加藤・大野・中川（2021 a, b, c, d：すべてグイ語）と加藤（印刷中：徳之島方言）。一旦環境を用意するとアノテーションと資料の提示が短時間で行えるようになるので、その他の言語においてもこれらの論文のような資料を手早く提示できるようになることが期待される。

4 おわりに

本発表では、菅原資料のアーカイブ化によって得た知見と、その応用例を紹介した。

参考文献

Everett, Caleb (2018) “The similar rates of occurrence of consonants across the world’s languages: A quantitative analysis of phonetically transcribed word lists”. *Language Sciences* 69: 125–135.

加藤幹治 (2020) “Phoneme frequency analyzer” `https://phoneme-frequency-analyzer.herokuapp.com/` (最終アクセス日 2021 年 10 月 14 日)

加藤幹治（印刷中）「奄美語徳之島伊仙町方言のモノローグ談話資料 ―「ハマウリとミーバマクマシと天照大神」の話―」『Asian and African Languages and Linguistics』16.

加藤幹治・大野仁美・中川裕（2021a）「グイ語資料：受動表現」『語学研究所論集』25: 335–341.

加藤幹治・大野仁美・中川裕（2021b）「グイ語資料：アスペクト」『語学研究所論集』25: 343–352.

加藤幹治・大野仁美・中川裕（2021c）「グイ語資料：モダリティ」『語学研究所論集』25: 353–360.

加藤幹治・大野仁美・中川裕（2021d）「グイ語資料：ヴォイスとその周辺」『語学研究所論集』25: 361–369.

中川裕（2021）「多数のクリック子音をもつ言語は音韻体系をどう組織化するか：“コイサン” 諸語の子音・母音・音素配列」日本音声学会第 35 回大会特別講演, 2021 年 9 月 25 日, オンライン.

[3] 自動とは言っても、論文として提出する場合には資料と言語の概要を書いたりする必要があるが、それらはもちろん執筆者が自ら行う必要がある。自動で行われるのは、あくまで資料部分の整形である。