W-2-1

なぜいま言語データの救出が必要か

加藤　重広（北海道大学）

1. 記述言語学における音声データ

音声の記録（録音）が可能になったのは 19 世紀後半であり，一世紀半ほどの歴史しかない。しかし，蝋管（1877 年〜）やレコード盤（1894 年〜）では現地調査での録音には適さない。1940 年のリール・トゥ・リール方式の磁気テープ（いわゆるオープンリール）も高額で大きな機械であることから必ずしも現地調査には適していなかった。1951 年には Sony のデンスケが発売されるが，当初は放送局などの業務用であり，民間に普及するのは 1960 年代だという。言語研究に多く利用されるようになったのもこの時期だろう。1962 年にカセットテープ（いわゆるコンパクトカセット）がオランダのフィリップス社によって開発されると，オープンリールに徐々にとって変わる。しかし，当時のカセットテープの録音機はポータブルではあるものの，それほど小型化は進んでおらず，乾電池とテープを持って現地調査に赴くのは大変だったようだ。カセットテープが廉価になり入手しやすくなった 1970 年代には現地調査する研究者の多くが音声を録音したデータを残すようになったが，この時期は現地調査への研究費が充実し始めた時期でもあったようだ。

カセットテープなどのアナログ形式からデジタル形式での変化は 1990 年代から始まり，いまではカセットテープは主流ではなくなった[1]。IC レコーダー(digital voice recorder)は 1990 年代半ばから発売されているが，2000 年代以降一気に普及し，いまは音声録音の主流になっている（PC やタブレット，スマホでも同等のことができる）。以上の状況を踏まえると，言語音等の記録のために磁気テープが使用されていた期間は長く見積もっても半世紀程度で，特に 1970 年ごろから 1990 年代半ばまでの四半世紀が録音媒体としての磁気テープが主役だった時期だと言えるだろう。この時期に言語・方言の記述的研究を行い，音声データを所有している研究者は 1930 年代以降の生まれで 1990 年代以降に退職したと推定できる。

これらのアナログデータは，本来の形では，保存状況の影響を強く受け，時間経過とともに音質などが劣化することがある。また，再生には専用の機材[2]が必要であり，詳しいメタデータを欠き，データがユニットに細分化されていないことが多い。デジタルデータにしなければ利用しにくくなるだけでなく，失われてしまうことが考えられる。

---

[1] 2000 年には消費者物価指数調査の対象品目から外れている。主なデジタル形式の媒体としては，コンパクトディスク(1982 年)，DAT（1987 年），MD（1992 年）などがある。

[2] 標準的な機材ではノーマルポジションやハイポジションは再生できるが，メタルテープは対応していない場合がある。

2. データの形状

もちろん言語調査のデータは，音声データだけではない。調査時にフィールドでとったノートなどの筆記録もある。これらはいわば紙媒体であるが，ノートやルーズリーフ，カード類など多岐にわたる形状が見られ，ノートにはメモ[3]や別葉の頁が貼り付けてあるものもある。近年ドキュメントスキャナーが普及しているが，個別に確認してスキャンしなければ必要な情報が欠落してしまう可能性があり，扱いには注意を要する。ほかに，写真（ネガやポジ）や，動画（アナログビデオ・デジタルビデオ）もあり，さらには現地で入手したとみられる物的資料が含まれていることもある。ノート類は，概ね手書きであるため，スキャンして PDF あるいは画像の形式でデータ化しても，それを検索できる形にするには処理が必要である（人間がおこなっても AI がおこなっても，解読のプロセスが必要であり，ときには誤読・誤入力の可能性がつきまとう。加えて経費を要する）。

例えば，30 年程度の間に集めたデータは，どの程度フィールド調査をおこなったかによって量の多寡に相当なばらつきがあるが，段ボール箱で 10 箱以上にはなるだろう。退職後，自宅で保管しておいても，研究者自身が物故すれば遺族は処置に困ることになる。遺族が価値を理解しており，大学や研究機関に連絡してきても，適切な部署で対応しなければ，通常は引き取ってもらうことも難しい。結果として，この種の事業を担う制度やシステムが必要であり，研究者自身が健在のうちにそのシステムが適切に対応する以外にこの種のデータを救う方法はない。

いま現在在職中の言語研究者はほぼ独力でデジタル化が可能だと想定しても，60 歳台後半から上の世代の研究者の中には，アナログデータをそのまま保持している方が一定数存在する。もちろん，そのすべてがデータのデジタル化を望んでいるわけではなく，①すでに弟子など他研究者がデジタル化を済ませたケースもあり，②そもそも自分のデータを他人に見られたくないというケースもある。しかし，デジタル化して残せるなら協力したいが，その方法がわからないというケースも少なくないと推測している。

60 分や 90 分のテープをデジタルデータに変換しても，形態素・単語ごとに切り出す作業，また，それをノート類のデータと照合して紐付ける作業が必要になる。個々の形態素の紐付けをおこなわなくとも，データ群の紐付けをしてメタデータを作成する必要がある。メタデータの作成時には，研究者自身に確認してもらわねばならないこともある。単語の音声リストであれば，対応する一覧（簡易表記や意味）があるはずであり，例文の読み上げ音声も対応する文の一覧があるはずである。いわゆるテキスト（民話や伝承の独話や対話など）はすぐに文字に起こせない可能性があるので，対応する文字データが完成していないケースが考えられる[4]。

---

[3] フィールドノートは，分析上気づいた事や野外調査の予定，旅程計画，日記に相当する記録や個人的なメモを含む場合もあれば，必要最小限の言語学的な情報しか含まないこともある。

[4] 音声資料の録音スタイルも研究者によりさまざまである。言い間違いや読み間違いがあっても止めずにそのまま録音し続けたり，途中で別の話が始まっても中断することなく続けたりする場合もあれば，誤りや言い直し，ノイズを避けるため，いちいち録音し直すこともある。読み上げリストと対応する音声だけ

もちろんデータのなかには，整理・分析が終わり，論文や資料・報告書として公刊されているものもある。しかし，公刊された成果がすべてのデータを網羅的に含んでいることは少ない。

ここで想定しているデータは，研究者が収集作成した資料（＝第一次資料）である。母語話者の音声を録音したカセットテープは，「研究資料」であるが，研究者の分析や判断を含まないので「第一次資料」であり，「音素を決めるためのメモや音素一覧を含むフィールドノート」は「第二次資料」である。しかし，「一次性」に関する区分は大まかな区分としては有効であるが，研究者の分析や判断がどの程度関与するかは容易に決められない（下記の表 1, 2, 3 とあわせて，加藤.2020 を参照）。

研究資料	研究成果を作成するための資料	第一次資料	研究者の分析・判断を含まない資料
		第二次資料	研究者の分析・判断を含む資料
研究成果	論文や報告書など成果として公刊されたもの		

（表 1）データ形式の区分

Himmelman(2012) では，原データ(raw data)と主要データ(primary data)と構造データ(structural data)にわけ，構造データを二次データ(secondary data)とも呼んでいる。上記で筆者が第一次資料と呼んだものが「原データ」におおむね相当し，第二次資料と呼んだものが「主要データ」に相当するようだ。

	歴史言語学データ	現代のデータ	メタ言語的技能に基づくデータ
原データ	碑文	音声記録・画像記録	個別の反応や評価の記録
主要データ	翻刻	フィールドノート・対訳付きテキスト	実験の統計分析結果・フィールドノート
構造データ	音法則	グロスつき例文の記述・辞書・データベース	処理速度・頻度データ

（表 2・Himmelman.2012 によるデータタイプ区分）

研究者が生み出すもの・データとしては上記のようなものがあるが，管理上，不可欠の情報・データがある。それはデータに関するデータ，**メタデータ**(meta data)である。ここで言うメタデータとは，データの一覧・データの種類や形状・テータ間の対応関係・データの属性（収集時・収集地・提供者とその属性）などを広く含む。研究者個人がデータを保持している状況では，メタデータが不完全でもあってもみずからの記憶で補うことができるかもしれないが，第三者が保持する場合は，管理上，できるだけ精密なメタデータが必要になる。この種のメタデータを作成することが保存管理作業では不可欠となる。

があれば混乱はないが，言い直しがある場合，いずれが言い間違いなのかあとから第三者が見た場合判然としなくなる可能性はある。

3. データにかかわる権利

研究者が収集・作成するデータは誰のものだろうか。カセットテープやフィールドノートといった所有物に関する所有権は無論研究者に所属する。ではその内容に関する権利はどうだろうか。例えば，学会が刊行する雑誌に掲載された論文を例にすると，一般的に，論文の内容に関する権利は著者が，学会誌という印刷物の版面の権利は出版社か学会が，公衆送信権は学会が有している。しかし，研究者個人の作成・収集したデータとなると複雑で，難しい。

研究者個人が，分析や論考の途次に得た着想やアイディアを記したメモであれば，その内容もメモそのものも研究者個人のものだろう。では，ある言語を調査するために現地を訪れた研究者が母語話者に依頼して発音してもらった音声データはどうだろうか。協力してくれた母語話者は研究者個人の研究のために発音や吹き込みをしているので，研究者個人がその研究で分析に用いる場合は，権利侵害は発生しないと考えられる。しかし，研究者がその音声データをなんらかの方法で公開するとしたらどうだろう。研究者と協力者が詳細な契約書を取り交わしているケースは少ないと思われる。例えば，母語話者が提供した音声を学会といった聴衆が特定される（所属会員しか参加しない）場で実例として聞かせるケースはどうだろうか。実際にそういう事例はあり，調査時に学会発表で聞かせることを考えていれば，言語研究者は協力者に口頭でそのように使うことに許可を得ることは可能だ。しかし，それを書面で「許諾範囲の特定」として定めているケースは少ないのではなかろうか。

学会発表で研究者が経済的利益を得ることはあまりないので，著作隣接権たる実演に対して対価が生じることも考えにくい[5]。また，研究者と調査協力者に個人的な信頼関係が成立しているうちは，問題になりにくい。問題が生じるのは，その一方または両方が物故してしまった場合，その権利と義務を当人が管理できない場合だろう。

データ種・活用形態	権利者	権利形態
フィールドノート（メモ類含む）	研究者	著作権
音声データ	インフォーマント	著作隣接権（実演）
	研究者	著作隣接権（編集・作成）
動画	インフォーマント	著作隣接権（実演）・肖像権
	研究者	著作隣接権（編集・作成）
一般公開	学会・研究者	公衆送信権

（表 3・データ種と権利の関係）

著作権と著作隣接権には人格権が認められる。通例，これらが学術研究にのみ用いられる

[5] Eckert(2014)は，協力者に対価を払うことについて触れている。協力関係が長く続けば言語知識の提供とその対価だけにとどまらず，他の知識の提供や移動の便宜から生活物資の提供にまで及び，ビジネスか友人関係か境目が決められないような対人距離になってしまうと言う（Eckert.2014:23）。

のであれば，著作人格権や著作隣接人格権は侵害されないであろう。また，研究を成立させる資源という観点から考えると，ある言語に関する研究データを得るには，研究者個人の資質とエフォート，研究協力者の資質とエフォート，研究経費が必要だということになる。

4. なぜいま「発掘」するのか

ここまでに述べたように，発掘すべきデータは 1990 年前後から 2015 年前後に退職した言語研究者(RLR)が保有するアナログデータである。①所在と状況の確認，②デジタル化の有無の確認，③提供の意思確認，④デジタル化，⑤メタデータ（言語データの中身について一覧表のようなもの）の作成，⑥アーカイブ化（利活用を含む）といった手順を想定しているが，①は個人的なネットワークだけでは限界があり，また，②〜⑤（特に④）には研究経費の措置が必要である。⑤には，ある程度当該言語についての言語学的知識を有する者（研究者が望ましい）の協力があるとよい。⑥には，永続性のある機関か組織が責任を持って受け入れ，管理するしくみが必要になる。電子データの保存だけであれば大きな負担ではないが，長期的に保管し，利活用する場合には，国策としてでも対応するべきだと考える。このほかに，大学等に寄贈されたアナログデータが眠っていることもあり，それも対象に含めることができる。

データの提供を受けると言っても，電子化が済めば RLR に最終的にデータは返却することになるが，返却ではなく処分することを求められるケースもある。電子化されたデータは提供者たる RLR にも提供される。

ここで述べた内容はいわば一般論であり，総論である。記述言語学的なデータは，いわば現地調査の数だけ存在し，それも研究者ごとに異なり，調査地や対象言語ごとに個別の事情がある。ある言語ではあまり問題にならない形態論の問題が別の言語では非常に複雑で記述上最も手間を要することも想定され，同様の事態が音韻論や統語論でも生じうる。この課題に対処するには，制度設計を一般論にしたがってトップダウン的におこなっても，個別言語の記述の事情に合致するよう柔軟な対処ができる体制が必要になる。われわれの発掘作業はまた端緒についたばかりであり，理解と支援に加えて情報提供をお願いしたい。

参考文献

加藤重広(2020) 「言語データの継承と保存に関する課題について」『北海道大学文学研究院紀要』161，35-49

Eckert, Penelope (2014) Ethics in linguistic research, in *Research Methods in Linguistics,* Edited by Robert J. Podesva and Devyani Sharma, Cambridge: Cambridge University Press, 11-26

Himmelman, Nikolau P. (2012) Linguistic Data Types and the Interface between Language Documentation and Description, *Language Documentatio & Conservation*, Vol.6, 187-207