※この原稿は,「国語研の窓」第 36 号（2008 年 7 月 1 日発行）の「創立 60 周年に寄せて」の原稿に加筆されたものです。

「方言文法全国地図」データの電子化

沢木幹栄
（信州大学人文学部教授）

＜データの電子化とは＞

「方言文法全国地図」（以下 GAJ と略）のもとになったデータはすべて電子化されているがこれだけの規模の言語地図では世界的に見てもほかに例がないと思われる。ここでは，その経緯について簡単に述べたい。

「日本言語地図」（LAJ）のときもそうだったが，調査票にしたがって調査した結果はまず調査票に記入されそのあとで所定のカードに転記される。地方研究員からの報告はこのカードを提出することで行われる。言いかえれば，GAJ の全データはカードの形でまず存在しているのである。

1984 年から始まり 1988 年ごろまで続いた作業（当時はコンピューター入力と言っていた）はカードに記入された情報すべてをそのまま入力することをめざした。その目的は地図の自動作成とデータベース化だったが，そのどちらも第 1 集刊行時には達成できなかった。簡単に言えば当時のハードとソフトでは無理だったのだ。当時私の所属していた研究室（言語変化研究部第一研究室）のハードディスクは 7MB（7GB の間違いではない）の容量しかなく，PC のメモリーは 64K バイトだった。また，今のように便利なデータベースソフトもなかった。データベースのプログラムを自前で書きかけたが，PC のスピードも容量も不足していた。

入力の原則は前述の通り,「カードに記入されている情報をそのまま電子化する」ということだった。自分以外の調査者の報告は時として賛同しかねる点があったりするのだが，だからと言ってそこで手を加えてしまうのは越権行為であり,してはならないことである。報告されたことに対しては全く解釈を行わず，書かれている通りに入力することにした。

データの整備は地図集が順次刊行されている間も続き，最終的に現在の形になるが，私が関わったのはデータ整備より最初の入力作業の部分だった。

＜作業の流れ＞

入力そのものは業者に外注した。納品は最初はパンチカードで，その後は大型計算機用の磁気テープだったこともあるが，8 インチのフロッピーディスクが一番多かった。そのどれも今では探し回っても見付けられない媒体である。調査項目の地点ごとの回答は音声記号で記録されているのだが，音声記号を英数字記号の組み合わせに置き換えて入力することにした。例を挙げると，スモールキャピタルの N は N9 になる。直接の担当は言語変化研究部第一研究室（変化一研）の私と白沢宏枝さん（元所員）だったが，その仕事はカードに鉛筆で記号化の仕方を書きいれることだった。入力業者には前処理をしたカードを

渡し，業者は納品時にそのカードを返却するという流れになっていた。もちろん，データができたらそれを校正する作業も必須だ。

今考えると入力作業を一気呵成にしなかったのは問題があったかもしれない。大きな作業量だったために短期間で行う勇気がなかったのだが，このような作業を何年も行うと作業の原則にぶれが生じたり，いろいろな問題が起きてくる。この間入力業者も何社か関わったので後述のような問題もあった。

＜時代の制約＞

世の中のあらゆるものがそうであるように，我々の入力作業も当時の状況から来る制約から逃れることはできなかった。まず，音声記号の置き換え規則がそうである。変化一研で使っていたのは 8 ビットではあったが,いわゆるパソコンで,大文字も小文字も使えた。しかし，入力業者は当時の業務用の主流であった大型計算機用の仕事が普通であり，そこで使われる EBCDIC という記号体系でデータを作成する。EBCDIC では大文字しか使えない。仮に大文字と小文字が同時に使える状況だったら，置き換え規則はかなり単純化され，分かりやすくなっていたはずだ。また，入力業者によって記号体系に微妙な点で違いがあることも悩みの種だった。

業者の納入したデータは大型計算機用なので，変換プログラムを使ってパソコンで使えるように変換した。

大型計算機のデータの入出力の標準になっていたのは 80 桁のパンチカードで，納品が磁気テープであってもフロッピーであっても，80 桁が基本だった。そこで，80 桁に収まるようにデータの構造を設計した。一つの質問に対して回答語形が二つまでだったら対応できるように，項目番号に 5 桁，第一答に 35 桁を与え，その語形情報（「古い」,「よく使う」など）に 5 桁，第二答に 34 桁，語形情報に 5 桁，地点番号に 6 桁を与えた。しかし，これも，固定長であるがための苦肉の策で，語形が例外的に長い場合や，回答語形の数が 3 以上のときは特別な処理をしなければならなかった。

語形情報は「古」とか「多」と記されているものは対応する 1 字のアルファベットあるいは記号を与え，文の形で記されていてなおかつ定義されている記号に還元できないものは C を与えた。このようにすると大抵の場合，少ない桁数でも足りるのである。

＜データの公開と活用＞

入力量は文字数にして数百万字（公開されているデータで数えると語形データだけで 480 万字ほど）だった。あとで考えると恐ろしいようなデータ量である。確かに入力は外注だったが，前処理も校正も作業量が膨大だった。

GAJ の第 1 集が刊行されるときにすべての回答を印刷して公開しようということになった。このころには，自分で縦 24，横 16 ドットで字形を作ってオリジナルの文字を印刷できるようなプリンターが市販されていた。音声記号を印刷するのは問題なくできる。

問題はプログラムで，私が作成したのだがこれが出版直前になってもなかなか完成しない。ぎりぎりまで待ってもらってやっと動くものができた。まさに冷や汗ものだった。出版に間に合わなかったら，データの公開はあきらめようという話まで出ていたのである。仮定の話になるが，第 1 集に資料一覧がつかなかったら，その後の第 6 集までずっとデー

タが公開されないままだったろう。そうなったら，せっかく入力したデータはきちんと校正されることもなく，研究に利用される機会も資格も与えられずに忘れ去られていたはずである。

その後，GAJ のすべての地図に対してはそのデータを印刷したものが資料一覧としてつけられるようになった。したがって，全データをほぼ記録された通りの形で見ることができる。

第 3 集の資料一覧（この資料一覧は第 2 集と第 3 集のデータ）からはエプソンのレーザープリンターを使って印刷した。第 2 集の編集作業のときに，資料一覧の校正刷りに当たるようなものを製本して地図作成の補助にすることにした。第 2 集以後は地図作成に関わっていないので実際にどうだったかは分からないが，これは役に立っただろうと思う。また，製本したものをカードと対照することで入力されたデータの間違いを発見するということも多々あったに違いない。

音声記号として印刷するまえは，記号の組み合わせで作られたデータをもとの表記と照らしあわせて校正していたが，これは記号の組み合わせと音声記号の対応がすべて頭のなかに入っている白沢さんほかのごく少数の人しかできないことだったし，当然見落としもあった。もとの表記と同じ形で印刷することで校正が容易になったのである。

資料一覧を出版するときは，地図作成の過程で校正され追加情報（質問文に対する回答として適切かどうかなど）の入ったデータを使ってレーザープリンターで印刷する。これを最初から最後まで読む。これが最後の校正になった。定義されていない記号の組み合わせは音声記号に変換されないで無意味な記号の組み合わせとして印刷されるので，ただ字面を追っていくだけで間違いを見つけることができるのだ。

第 2 集以後はこのような手順で校正を行っている。

また，地図作成が終わったデータはさらに情報が付加され，整備されて直接いろいろな研究の材料として利用することができるようになった。所期の目的が達成されてうれしい限りである。