

方言コーパスによる徳之島方言の研究

沢 木 幹 栄

I. はじめに

この稿では、電子化されたある程度の量の言語資料（コーパス）を用いた方言研究について報告を行うとともに、それが従来の方法を補うあらたな知見をもたらす可能性について述べることにしたい。（注1）

II. 研究の発端

我々の研究チーム（注2）は徳之島方言辞典作成を目指して研究を進めているが、メンバーの一人でインフォーマントでもある岡村がアンリ・フレの『日本語二千文』をもとに『徳之島二千文』を作成した。我々はこれをコーパスとして利用することを考えた。ワープロの文書として作られたものなので、コンピューターのデータに容易に変換できたからである。

III. 『徳之島二千文』

岡村の方言は徳之島の天城町浅間である。徳之島方言は言うまでもなく琉球方言に属している。音韻的には7つの母音を持ち、子音も喉頭化音があってバラエティーが豊かである。しかし、中舌母音や喉頭化音を大文字で表記するなどの工夫をすれば、普通のワープロやコンピューターに備わっている英文字で音韻的にまぎれのないかたちで方言の音声を記述することが可能である。

この簡略な表記の原則を示す。

- ①中舌母音はそれに対応する文字の大文字で示す
- ②喉頭化音は大文字で示す
- ③声門閉鎖音は'で示す
- ④撥音はNで示す。nの喉頭化音と同じ表記になるが、位置で区別する

『徳之島二千文』（以下『二千文』）もこの表記で書かれている。しかも、すべての文にアクセント表記がつけられている。このようなことができたのは、岡村が独力で方言の記述ができる研究者でもあるからである。その意味で本研究は非常に特別なケースかもしれない。

『二千文』は『日本語二千文』をもとに作られたが、おおもとなる“Deux milles phrases”は「文の辞典」として構想されたものであり、「ひとつの言語の文法・語彙・音韻を範例的に代表し得るように選ばれた2000箇の文の目録である」（『日本語二千文』の川本茂雄の解説）。

それに対して『二千文』自体はあくまでも個々の文の着想のヒントを『日本語二千文』から得たのにすぎないが、標準語からの直訳を避け、できるかぎり自然な表現を目指した点に

において、『日本語二千年』と精神を共有していると言える。

自然な表現であることの例を二つあげると、一つは逐語訳でない文である。

1627. junaga: [tu] /katamIN [ba] [Kui] narada: [tI].

1627. 一晩中／ほとんど眠られなかった。(徳之島方言では「片目も閉じられなかった」)

ただし、このことの当然の帰結として、文節切りは意味の対応する単位でまず行った。したがって、一般的な「文節」の切り出し方から少しずれることがある。

自然な表現であることのもう一つの例は文脈の移し替えである。

1661. [na:wui] [na:] tI/ [kuN/a:] kija/joroN [ka] Ci/ [’ika:].

1661. 農閑期だから／この／秋は／与論へ／行こうよ。

(原文は 「1661. おりがあったら、この秋にあそこへ行こう。」)

もう一つ、『二千年』に特徴的なこととして、それが方言文と標準語文のセットであるということが言える。本来であれば、方言文だけで成り立つはずのものであるが、上述のように『日本語二千年』の直訳ではないために、『日本語二千年』を見ても方言文の意味が分からないことが多い。そのための説明として標準語文が存在する。言い方を変えると標準語文はそれを単独で見ると不自然である場合がある。標準語文もまた、方言文とセットで初めて意味を持つといえる。

『二千年』は文節単位での一対一の対応がほぼ保たれている標準語文と方言文がセットとなってはじめて成立するものである。

IV. コーパスとしての利用

『二千年』は一人の話者による方言データとしては屈指のものである。『日本語二千年』のデータがほぼ2000文、漢字仮名混じりで約32000字に対して、『二千年』の標準語文は約2400文、約50000字である。世の中で普通に言われるコーパスに比べれば非常に小さいが、方言資料としては非常に大きい。

方言辞典の作成のためには、できるだけ多くの語彙を拾い上げることが必要である。また、文法の記述ができていないと、動詞の見出し形としてどの形を使うのか決められないし、また、その見出し形から個々の活用形がどう導かれるかも記述できない。

したがって、我々が『二千年』をコーパスとして利用することの最大の目的は語彙の採集と文法についての見通しを得ることだと言える。『二千年』をコーパスとして利用できるようにするためにはこれをデータとして加工しなければならない。(注3)

上記の目的と我々に課せられた条件から、原資料を文節切りして、対応する方言と標準語の文節を文番号と一緒にまとめることが現実的と判断した。岡村以外のメンバーでも、標準語文と対応させれば方言文の切りわけはほぼ間違いなくできる。文節切りは文節の切れ目／を入れることで行った。

方言文と標準語文はほぼ一対一の対応をしているので、文節を切り出して文番号、方言文

節、標準語文節で一つのセットになったデータを作る。これから

(a)標準語の文節(b)標準語の文節の逆引き(c)方言の文節(d)方言の文節の逆引き

のそれぞれをキーにしてソートすることによって、4種類の資料を作ることができた。

(a)は標準語の意味の同じものが近いところに並ぶ。(b)は標準語の文節の終りが同じもの(助詞など)が近いところに並ぶ。(c)は方言の語形の同じものが近いところに並ぶ。(d)は文節の最後に来る助詞助動詞の同じものや、動詞の同じ活用形が近いところに並ぶ。

(a)の一部

開いた	'aCjI:	1473
開いてるじゃないか	'acju:sI	355
開けたまま	hE:taNma:ma	337
開けられた	hE:rattI	1474
階段が	kaidaNja	344
階段は	kaidanja	1148

(b)の一部

望みを	nuzju:mi	774
どちらを	dIN	1918
バラを	ba:ra	1477
このカメラを	kuNkame:ra	1067
子らを	KwaNkjaNba	50
彼らを	atta:	666

(c)の一部

sjera:	しましろう	1639
sjero:	なさい	79
sjero:	なさい	667
sjetaNga	したんですか	1106
sjetI	しました	1796
sjE:tI	飽きました	185
sjI:kadu	したほうが	471

(d)の一部

kutu:du	ことをぞ	1024
najuNkutu:du	なることしか	1079
arI:gadu	あれが	1600
kuihadu	濃いほうが	1355
sjI:kadu	したほうが	471

このような処理をするにあたって、データが電子化されていることは非常に大きなメリットがあった。文節切りは人手で行うが、データの形式的なチェック(文番号が正しく打たれているか、方言文と標準語文の文節の数が同じになっているかなど)や、対応した文節の組を作ること、逆引きのためのキーを生成することなど自動的にできるものはすべてプログラ

ムで処理した。最終的なソートは出来合いのプログラムを使った。手作業でこれらのことを行ったら膨大な手間が必要だったと思われる。(注3)

こうして作成した資料から何が言えるかについては以下に具体的な例に即して述べることにする

V. コーパスによって得られた知見

1. du

琉球方言に特徴的な係助詞 du は浅間方言にも存在するが、これが一つの文のなかで複数回出現することがあるだろうか。

柴田1976によれば、琉球方言のなかの宮古方言では一つの文の中で du が2回以上現れることはない。これは琉球方言全体に見られる現象のようである。

そのような文字連続を検出するコンピュータプログラムを作って実行したが、係助詞 du の重複の例はなかった。

(d)の資料で調べると係助詞の du が何例あるかが分かる。係助詞の du は文節末に現れるからである。それによると66例が確認できる。これだけの例のなかで一つの文中で du が重複する例は一つもないのである。実は徳之島方言は宮古方言にくらべて du の出現そのものがずっと少ないということもこのデータから言えることなのである。

2. 「を」

浅間方言では「を」に意味と音韻の両方に対応する助詞はないようである。それは岡村の内省によってもそうなのだが、それでは『二千文』で標準語の「を」に対してどのような形が用いられているだろうか。これは(b)の資料で簡単に調べることができる。

標準語の文節に「を」は365例あった。そのうち、方言文節では名詞のはだかの形に対応するものが大部分であった。それ以外には助詞 ja (「は」) が対応するのが15例, ba または Nba (「も」) が対応するのが4例, du が1例だった。このように『二千文』には「を」に意味と音韻の両方に対応する助詞は出てこない。

3. 助詞の異形態

(d)の資料で助詞を調べているうちに、道具格を表す sjI が sjI のほかに ssjI という形をとることがわかった。しかも, sjI は長母音, ui, ai などの二重母音, N を持つ名詞に接続するときに出る形でそれ以外のときは ssjI であることが分かった。

これは浅間方言の「一つのアクセント単位のなかに必ず heavy syllable (長母音, 二重母音, 撥音, 促音) がある」という法則(上野1977)を満たすために発達した現象である。おそらく, sjI が本来の形であって, 名詞に sjI が接続したアクセント単位のなかで heavy syllable が存在しないときに ssjI という heavy syllable を持った形になるのであろう。

そこで, この現象についてもっと詳しく調べるために岡村と対面調査を行った。岡村はこの現象に気がついていなかったが, 一音節の名詞に sjI が接続するときは kI:ssjI (「木で」), kI:sjI (「毛で」) となることを発見した。単独ではそれぞれ kI [:], [kI:] である ([:] はアクセントを表す) ので, どちらも kI:sjI となってもよさそうなものである。ところが, 実際

には最初の母音は一方では短母音で、もう一方では長母音で現れる。

このことから以下の結論が導かれる。

イ) CV の構造の単音節語は単独では必ず長母音だが、本来その語形が長母音であるとは限らない。単独のときは上の制約ゆえに語形が長母音になっているとも考えられる

ロ) ssjI が出てくるか sjI が出てくるかは上で述べた制約とアクセントによる制約を組み合わせたもので説明できることが予想される。

まさにコーパスによる調査と対面調査の協調によって得られた成果と言える。

また、(d)を調べることによって、ba, Nba (「～も」)にも同様の現象が見られることがわかった。すなわち、この二つは同じ助詞の異形態であり、前者はこれが後接する名詞に heavy syllable があるときのときに出現する形である。さらに bE:と NbE: (限定の「ばかり」にあたる), kaI と ka:cI (目的地, 方角の「へ」) ka と ka: (出発点の「から」), kara と ka:ra (同) nu と N (格助詞の「の」と「が」の両方の意味を持つ) の対でも同様の現象が確認された。

(注)

1. 本稿は第75回方言研究会研究発表会での発表「方言コーパスを利用した方言研究の可能性—『徳之島方言辞典』作成のために—」(発表者沢木幹栄・福嶋秩子・中島由美・岡村隆博)に基づき、そのうちの沢木担当分を発表では盛り込めなかった情報を含めて詳述したものである。
2. 沢木のほかに、福嶋秩子(県立新潟女子短期大学教授), 中島由美(一橋大学大学院社会学研究科教授), 岡村隆博
3. データの整備, 文節切り, 処理するためのプログラムの作成と実行はすべて澤木が行った。

プログラムはすべてawkで書かれている。プログラムの動作は概略以下の通り。一つの作業に一つのプログラムが対応し、それぞれのプログラムは非常に短い。

- ①別々のファイルになっている標準語文と方言文を一つのファイルに統合する。その際、標準語文と方言文が交代交代に並ぶようにする。
- ②対応する標準語文と方言文の文節の数が同じであるかどうかチェックする。
- ③文番号, 標準語文の文節, 方言文の文節がセットになったデータを作る。
- ④③から, 標準語文の文節を先頭の文字が最後に来るように並べ直した逆引き用のキーを付加したデータを作る
- ⑤同様に方言文の文節から逆引き用のキーを作って付加したデータを作る。
- ⑥ソート(並べ替え)を行ったデータを整形する。

参考・引用文献

Henri Frei, Le livre des deux milles phrases(Genève, Droz, 1953)

アンリ・フレ 『日本語二千字 語研選書(3)』 (早稲田大学語学教育研究所1971年)

沢木幹栄・福嶋秩子・中島由美・岡村隆博 「方言コーパスを利用した方言研究の可能性—『徳之

島方言辞典』作成のために一」(『第75回方言研究会研究発表会発表原稿集』 2001年)

柴田武 「沖縄県平良市方言の付属語 du および nu, ga について」(桜楓社『佐藤喜代治教授退官記念国語学論集』 1976年)

柴田武 『語彙論の方法』(三省堂 1988年)

東京外国語大学アジア・アフリカ言語文化研究所『アジア・アフリカ言語調査票 上』(1966年)

上野善道 「徳之島浅間方言のアクセント(1)」(岩手県国語学会国語学論集刊行会『小松代融一教授退職・島稔教授退官記念国語学論集』1977年)

上野善道 「徳之島浅間方言の活用形アクセント資料」(法政大学沖縄文化研究所『琉球の方言25号』 2001年)