方言コーパス研究発表会
平成30年9月6日　国立国語研究所

奄美方言コーパスの作り方（配布資料）

沢木幹栄

発表の概要

１．方言のコーパスを日本で最初に作った
　このことについて間違いはない。

２．コーパスとは何か
　この資料末尾の「＊徳之島コーパスの例」を参照のこと。XMLというデータ構造を使っている。XMLは情報の付加が自由にできる形式である。
　発表者は日本語解析ソフトmecabを用いてコーパスを作った経験があり、コーパスを使えば語彙統計をとったり、特定の名詞句の構造の出現数を調べたりすることができることを知っていた。方言コーパスを作れば統語論的情報を容易に収集できると考えていた。

３．コーパス作りの過程
　岡村隆博さんによる「徳之島二千文」の作成がすべての出発点である。これを文節によって切り分けることから KWIC を作成した（『徳之島方言二千文辞典　改訂版』の DVD）。これによって、多くのことが分かった。KWIC は非常に効率的に情報を集めることができる。しかし、その一方で限界を感じたのでコーパスづくりを目指した。
　徳之島方言文を品詞情報を付加しつつ切り分けたものをプログラムにかけるとコーパスのスケルトンができる。これから品詞ごとのリストを作り、それを修正するということを繰り返して最終的に必要な情報を付加したコーパスを得た。
　動詞は活用を行うプログラムを作らないと形態素リストと実現形との対応がつかない。詳細は資料末尾の「＊プログラムに落とし込むうえでの動詞活用の問題」を参照。
　プログラムが完成した。入力のチェック、入力の半自動化、文法の精緻化に役立つ。
　コーパス作成の過程で形態素の語形確定、文法の精緻化、対面調査で見逃した語形の発見などが可能になる。これは対面調査による形態論の補完である。

４．共同研究のお誘い
　興味を持った方には発表者から共同研究をお願いしたい。

この研究は平成18年度文部科学省科学研究費（基盤研究B）「徳之島方言辞典語彙編の作成のための研究」（課題番号18320066）ならびに平成23年度文部科学省科学研究費（基盤研究B）「奄美方言データベース作成のための研究」（課題番号23320095）の助成を受けた。

＊「徳之島二千文」について

Frei,Henri “Le Livre de deux mille phrases”（フランス語二千文） Genève 1953

アンリ・フレ『日本語二千文』　早稲田大学語学教育研究所　1971
『日本語二千文』は「フランス語（ないし英語）の 2,000 の文を当該言語（この場合は日本語）に能う限り自然な表現によって翻訳するという方法で文例が収録されたのである」と「解説」にある。また、この 2,000 の文は「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」。
　徳之島浅間方言の話者で現地の研究者でもある岡村隆博さんがワープロを駆使して『日本語二千文』を徳之島方言に翻訳したテキストを作った。これが「徳之島二千文」である。なお、「徳之島二千文」は公刊された形では存在していない。

＊徳之島コーパスの例

kara:zIkacI	tI:danu	tI:tuNda:	という原文から下のようなタグ付きのデータを作
頭に	太陽が	照っているよ	る。改行は見やすくするためにつけたもの。

1 <Tok:m ps=noun id=874 >kara:zI</Tok>
<Tok:m ps=post_p id=33 type=1>kacI</Tok>
<Tok:m ps=noun id=1339 >tI:da</Tok>
<Tok:m ps=post_p id=42 type=1>nu</Tok>
<Tok:m ps=verb id=288　conjug=2 aux=00100000>tI:tuN</Tok>
<Tok:m ps=post_p id=5 type=1>da:</Tok>

＊プログラムに落とし込むうえでの動詞活用の問題

　動詞は単独で出てくるより、助動詞がついた形で出てくることが多い。しかも助動詞が複合する。動詞が単独で活用するのではなく、動詞と助動詞が組み合わさったものが活用する。たとえば、ヨムに対してヨマナイ、ヨンダという活用形を提示するだけで終わらない。実際はナイ、ダは助動詞なので、これが活用したらどうなるかも考えなくてはならない。
　また、助動詞は使役、受け身・可能、テイル、テオク、丁寧、否定、過去、疑問がこの順番に接続する。もちろん、これらの助動詞は使われることもあるし、使われないこともある。助動詞として考慮しなければならないのはこれだけである。徳之島方言では推量も様態も助動詞ではなく活用しない助詞が担っている。
　なお、テイル、テオク、疑問は普通に言う助動詞ではないが、しかし、これらは動詞に接続すると全体の形を変えて融合してしまう。このような融合してしまうものも助動詞と同じ扱いにすると処理が簡単になるのでそのようにしている。「疑問」を助動詞として扱うのもおなじ理由。単純疑問の終助詞は jI だ

が、CjI:(来た)に対して CjE:（来たか）のように融合を起こす。

このような助動詞と動詞をどうやってコーパスで情報付けをするかについては、2 通りのやり方が考えらる。

一つは動詞と助動詞に切り分けてそれぞれに対して情報付けをするというやり方。もう一つは動詞と助動詞が一つになったものを単語と考えてそれに対して情報付けをし、無理に切り分けたりしないというやり方。

最初の動詞と助動詞を分けるやり方は動詞と助動詞で融合が起きると不可能。実際融合は頻繁に起きている。

それでもどうしても分けたい場合は、ヨマレル　を　ヨム　と　ラレルに分解するということになる。ヨマレルが実現形、ヨム　と　ラレル　は基底形（underlying form）となるが、このコーパスで実現形と基底形を混在させるのは難しい。

第二の方法では動詞と助動詞が一体になったものはそれ以上切り分けることはない。したがって、テキストに使われた形がコーパスでそのまま使われる。助動詞がどう接続しているかは独立した情報となる。したがって情報として与えるのは動詞の識別番号、助動詞情報、活用情報ということになる。

助動詞情報はさきほどの連接の順番に対応する 8 桁の数字で与える。一番左の桁は使役の助動詞の有無で 0 か 1 になる。次の桁は受け身・可能の有無でやはり 0 か 1 というようにすると 8 桁の数字でどの助動詞があるかが表せる。

このプログラムが正しければ、識別番号、助動詞情報、活用情報から実際の語形が導き出せるはずである。逆に言うと助動詞情報、活用情報が間違っていれば実際の語形と違うものが出てくるので、こうした付加情報が正しいかどうかのチェックができる。今までであれば、人間が付加した情報を別の人間がチェックするのだが、そうではなくコンピューターを使ったチェックができる。

実際に、このプログラムを動かしてテキストに現れた形（実現形）とプログラムが生成した形が違うものを発見した。食い違いを発見したあとに福嶋さんの記述の通りに自分の頭で考えたものもプログラムが生成した形と同じだった。ということは、記述のほうが修正を要するということになる。

このプログラムを少し変えると、考えられるすべての助動詞の組み合わせの活用形を生成することができる。全部で４００あまりになりますが、それをテキストに出てきた実際の形と照合して一致したときの助動詞情報、活用情報がその形の助動詞情報、活用情報ということになる。

実際は候補が複数ある可能性があるので、そのつど人間が判断する必要があるが、これだけで助動詞情報、活用情報の入力作業が大幅に省力化できる。

参考文献

アンリ・フレ（1971）『日本語二千文』　早稲田大学語学研究所

沢木幹栄（2003）「方言コーパスによる徳之島方言の研究」　『人文科学論集第 37 号＜人間情報学科編＞（信州大学人文学部）』

岡村隆博、沢木幹栄、中島由美、福嶋秩子、菊池聡（2009）　『徳之島方言二千文辞典　改訂版』　徳之島方言の会（科学研究費報告書）

沢木幹栄（2018）「方言コーパス作成とその意義」『信州大学人文科学論集第 5 号』（信州大学人文学部）