『日本語諸方言コーパス』の構築について

著者	木部 暢子，佐藤 久美子，中西 太郎，中澤 光平
雑誌名	言語資源活用ワークショップ発表論文集
巻	1
ページ	57-68
発行年	2017
URL	http://doi.org/10.15084/00001458

『日本語諸方言コーパス』の構築について

木部暢子（国立国語研究所言語変異研究領域）†
佐藤久美子（国立国語研究所言語変異研究領域）
中西太郎（国立国語研究所言語変異研究領域）
中澤光平（与那国町与那国語辞典編集業務嘱託員）

For building of “Corpus of Japanese Dialects”

Nobuko Kibe (National Institute for Japanese Language and Linguistics)
Kumiko Sato (National Institute for Japanese Language and Linguistics)
Taro Nakanishi (National Institute for Japanese Language and Linguistics)
Kohei Nakazawa (Education board of Yonaguni town, Okinawa Prefecture)

要旨

『日本語諸方言コーパス（Corpus of Japanese Dialects、略称：CJD）』とは、諸方言の談話資料を横断的に検索することのできるコーパスのことで、方言に関するコーパスとしては、日本で初めてのものである。資料として、1977～1985 年に実施された文化庁の「各地方言収集緊急調査」の談話データを利用し、標準語で検索してそれに対応する方言形とそれを含む談話の一節を検出する方式でデータベースを構築している。2021 年度までに最低 75 時間（３時間×25 地点）の方言データ（音声データ、転記テキスト、標準語テキスト）を公開する予定である。本発表では、CJD の概要と特徴、構築のプロセス、及び本コーパスを使った方言研究の一例を紹介し、CJD を活用することにより、方言研究にどのような研究の方向性が開けるのか、また、活用する際にどのような注意が必要なのかについて報告する。

１．はじめに

近年、大量の言語データの整備と言語コーパスの構築が世界各国で進み、それに基づく言語研究が盛んになっている。しかし、方言に関してはこれまで、地域横断的なコーパスはもちろんのこと、一地点の方言に限定したコーパスでさえ作成されていない。このような状況を踏まえ、国立国語研究所共同研究プロジェクト「消滅危機方言の調査・保存のための総合的研究」（2010～2015 年度）、「日本の消滅危機言語・方言の記録とドキュメンテーションの作成」（2016～2021 年度）では『日本語諸方言コーパス（CJD）』の構築を行うこととし、現在、その作業を進めている。

本コーパスの特徴は、諸方言の談話を標準語で検索し、それに対応する方言形とそれを含む一定の発話単位を横断的に検索する点にある。言うまでもなく、方言と標準語は１対１で対応しない。そのため、標準語での方言検索には対応のずれの問題が生じる。しかし、各地方言の形態素辞書を作る時間と労力を考えると、すでにある日本語形態素辞書を利用して、標準語による検索を行い、並行的に方言形を検索するシステムの方がよいと判断した。また、諸方言コーパスがどのように利用されるかを考えてみると、標準語での検索システムは必須のように思われる。

資料としては、1977～1985 年に文化庁が行った「各地方言収集緊急調査」のデータを使用する。全体は、全都道府県 224 地点、１地点につき 30 時間程度の談話録音テープよりなる資料で、内容は当時 60 歳以上の地元出身者数人による自然談話である。一部は『全国方

† nkibe@ninjal.ac.jp

言談話データベース 日本のふるさとことば集成』（国書刊行会）として音声、テキスト、標準語訳が公開されているが、多くは未公開の状態である。本コーパスでは公開分、未公開分を合わせて、2021 年度までに最低 75 時間（３時間×25 地点）のデータをコーパスとして公開する。あわせて音声と方言テキスト、標準語テキストがダウンロードできるようにする予定である。

本コーパスの構築に向けて、現在、次のような手順で作業を進めている（詳細については第２節参照）。①方言音声の転記テキスト（方言テキスト）のチェック。②発話単位の認定。③方言テキストに対する時間アライメント情報の付与。④方言テキストに対応する標準語テキストのチェック。①の方言テキストと④の標準語テキストは、文化庁の事業の際に作成されたものがあり、これをもとにして、チェック作業を進めている。ただし、標準語テキストについては、全面的な見直しが必要である。前述のように、方言と標準語は１対１で対応するわけではないので、標準語テキストの付け方によっては、本来、検出されるべき方言形が検出されなかったり、検索結果が変わってきたりする可能性があるためである。標準語テキストの付け方については、現在、マニュアルの作成作業を進めており、CJD 公開の際にはマニュアルも併せて公開する予定である。②の発話単位の認定については、基本的に 0.2 秒の無音という基準で発話単位を認定している。また、話者同士の発話の重なりや相づち、フィラー、間投詞等のタグ付けも行っている。

上記の作業と並行して、本コーパスを用いた研究を試験的に行っている（詳細については第３節参照）。木部(2015)でも指摘したが、作業の中で次のような問題点が浮かび上がっている。

(a)各地の談話において、話者数、話者の属性（年齢や居住歴についてはある程度、指定があるが、男女、職業等については指定されていない）、話者同士の関係、話題等の統一が図られていない。例えば、話者同士の関係の統一が図られていないので、人称代名詞や待遇表現の地域差を単純に比較することはできない。また、話題により出現語彙に偏りがあることを十分に考慮する必要がある。

(b)標準語で検索し、方言形とそれを含む談話の一節を検索するという方法であることを念頭に置いて利用しなければならない。例えば、状態や感情には方言特有の語が使用されることが多く、秋田方言「あずましい」を標準語でどう検索するかというような問題がある。これについては、標準語テキストの問題として、コーパス構築作業の中で検討することになる。

２．日本語諸方言コーパス構築のプロセス

本節では、まず、CJD 構築のために行う作業の全体像を示した後、例を挙ながら具体的な取組みを紹介する。次に、一連の作業において特に重要となる物に関して、問題点とそれを解決するための試みを述べる。

２.１　日本語諸方言コーパス構築の流れ

本節では、CJD 構築のための一連の流れを述べる。図 1 は時間軸に沿って各作業を並べたものである。ここでは、「I. テキスト・音声の形成」にある作業に焦点を絞って詳細を述べる。

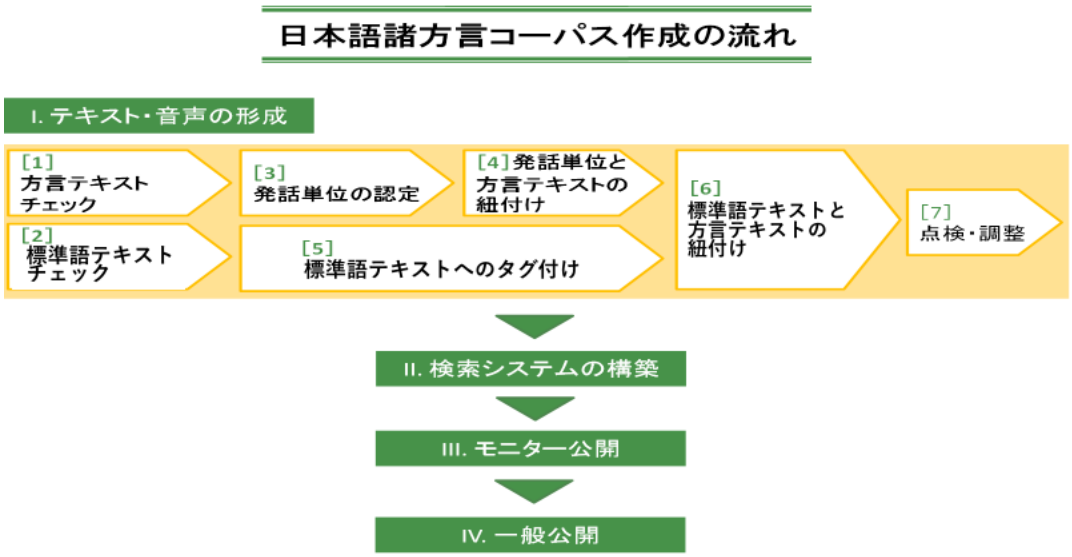

図 1. CJD 作成の流れ

［1］方言テキストチェック

1節で述べたとおり、本プロジェクトで扱う諸方言の音声データには、既に書き起こしされた「方言テキスト」と、それに対応する「標準語テキスト」がある。両テキストは複数の協力者がそれぞれの地域ごとに作成しているため、表記に揺れの問題が見られ、それらの修正が必要になる。そのために、方言間で見られる表記の揺れを統一するためのマニュアルを作成し、それに従いテキストの修正を行っている。

(1)に、方言テキストチェックの際の修正例として、東北地方を中心に見られる前鼻音の表記を示す。前鼻音を表すために地域ごとに異なる表記が用いられていたが、それらの表記を“n”に統一した。

(1)

文字化	文字化修正後
ヤンドト　シテ　カンネ	ヤｎドト　シテ　カンネ

音声の書き起こしには通常片仮名が用いられているが、必要に応じて片仮名以外の記号も用いられている。どのような記号をどのように用いるか、という点を明確にした上で、方言テキストの修正を行っている。

［2］標準語テキストチェック

標準語テキストチェックでは、標準語訳と方言が適切に対応していることを確認する。本コーパスは標準語からの検索を基本としているため、標準語テキストは、標準語としての訳の自然さではなく、方言と標準語を形態素ごとに適切に対応させることを優先している。以下に例を示す。

(2)

文字化	標準語訳
ソレオ　キカネーチャッタンダヨ	それを　きかないでしまったんだよ

標準語の訳としては自然ではないが、形態素が適切に対応するようにテキストを修正して

いる。ここまでが、「方言・標準語テキストチェック」として行う作業と、その具体例である。

［３］発話単位の認定

テキストの修正が済むと、次に、音声データの処理に進む。各地点 30 分程度の音声データを、検索上適当である単位に区切っていく。これは、表 1 に示した「発話単位認定」で行う作業である本コーパスでは方言音声が検索対象となるため、文法構造によらず、0.2 秒のポーズという音声的な基準に従って区切り目を設定している。このような基準で区切られた単位を、ここでは「発話単位」と呼ぶ。発話単位の認定作業は、プログラムと手作業で行い、この認定作業によって、音声データは発話単位に細切れにされる。

［４］発話単位と方言テキストの紐付け

続けて行うのが、この発話単位（音声）と方言テキストの紐付けである。具体的には、音声分析ソフト praat を使用し、0.2 秒のポーズで区切られた発話単位に方言テキストを貼り付けていく作業である。画面上での作業のイメージを図 2 に示す。

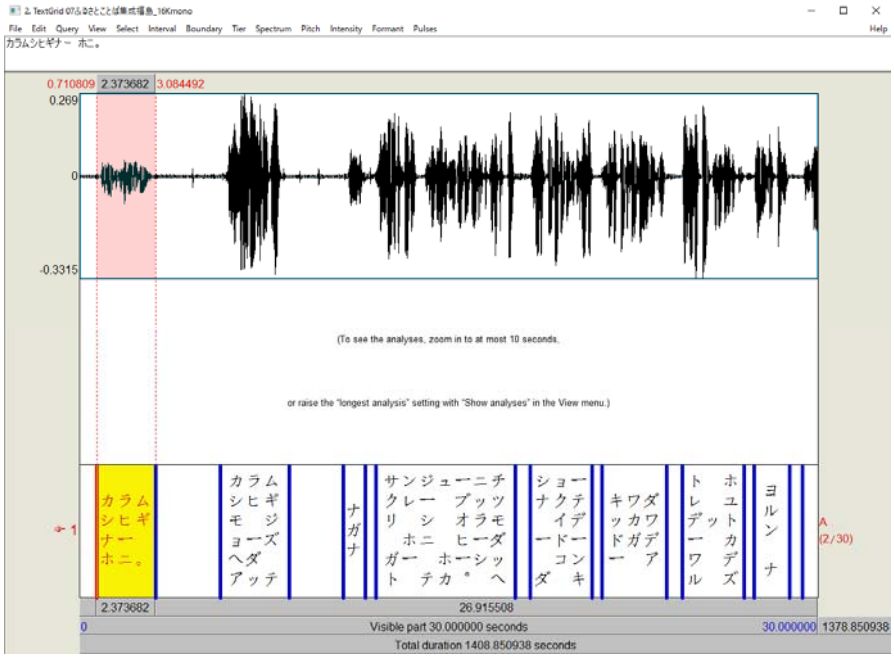

図 2. praat 上の作業イメージ図

この紐付け作業によって、コーパスの検索結果として、方言音声と方言テキストが連動して表示されることになる。紐付け作業にはいくつかの注意が必要であるが、それに関しては、２．２節で詳しく述べる。

［５］標準語テキストへのタグ付け

上記の「発話単位認定」、「音声と方言の紐付け」と平行して行う作業が、「標準語へのタグ付け」である。繰り返し述べている通り、本コーパスは標準語で方言音声・テキストを検索する仕組みとなっているため、標準語へのタグ付けはコーパス構築までの作業の中で最も重要であると言える。現時点で設定が決定しているタグ一覧を挙げる。

表 1. タグ一覧

	タグ	標準語へのタグ付けが 必要となる方言テキストの表現例
文成分の省略	［が］、［を］等	―
人称・単複	＜１・単＞　等	アッシ、オイ
終助詞	＜終助＞	ガ、ワ、ヤ
副助詞	＜副助＞	バリ、バッカイ
指小辞	＜指小辞＞	コ、メ
フィラー	＜F＞	エー、ホレ、アンタ
非言語的音	｛笑｝、｛咳｝等	―

以下では、「終助詞」と「人称・単複」のタグ付けの例を紹介する。

(3)は終助詞を表示するためのタグである。方言テキストにおいて様々形式を持つ終助詞にこのタグを付している。(4)は格標識のためのタグである。方言によっては格標識が音声的に顕在化しないことがあり、そのような場合は、方言テキストには存在しない要素をタグによって表示することになる。

(3)

文字化	標準語訳
エソエｎデ　ケサエンヤ。	急いで　ください＜終助：よ＞。

(4)

文字化	標準語訳
ヤケヒバシ　オッツケンデスヨ	焼け火箸［を］　押し付けるんです＜終助：よ＞

以上の工程で、発話単位（音声）に紐付けされた方言テキストと、タグを付された標準語テキストが揃うことになる。

［6］標準語テキストと方言テキストの紐付け

次の段階では、方言テキストと標準語テキストの紐付けが行われる。具体的に例を示すと次の通り（図 3)。

＜コーパス化作業前＞

発話No	枝番	話者	文字化テキスト	共通語訳
8	―	A	モー　ソノ　オー　オリャー　イマー　アスコラヘンニ　タキモン　カリー　イキョライ　チ　チューチ　（B　ハー　ハー）　ユー　（B　ハー）　ユーチ　ユーグライノ　コトヤッタヨ。　（B　ハー）　ナー。	もう　その　「おお　私は　今　あそこらへんに　薪［を］　刈りに　行ってるぞ」　×　と　（B　はあ　はあ）　言う　（B　はあ）　と　言うぐらいの　ことだったよ。　（B　はあ）　ねえ。

＜コーパス化作業後＞

8	000	A	モー　ソノー	もう　その
8	001	A	マー　オリャー　イマー　アスコラヘンニ　タキモン　カリー　イキョライ　チ　チューチ	「まあ　私は　今　あそこらへんに　薪［を］　刈りに　行っているぞ」　と　言って
8	002	A	ユー	言う
8	003	B	ハー	はあ
8	004	B	ハー	はあ
8	005	A	ユーチ　ユーグライノ　コト　ヤッチョローナー。	と　言って　言うぐらいの　こと　だったろうね。

図 3.『ふるさとことば集成』データのコーパス処理前後（福岡県北九州市の談話）

これによって、タグを含めた標準語を入力とし、方言音声とテキストを検索するコーパスが完成する。

［7］点検・調整

最後に、一連の作業を終えて整理された各方言における音声・方言テキスト・標準語テキストのセットの点検を行い、全体の調整を行う。様々な標準語やタグを入力として検索を行い、その結果が適切であるかどうかを確認する。地点ごとに作成されたテキストに揺れが生じている場合は、方言テキストで使用されている表記や、標準語のタグの使用に関する基準の改定や精緻化を行い、テキストの修正が必要となる。このような点検と調整を繰り返し、検索の精度を高め、方言横断的な研究に耐えうるコーパスの構築を目指す。

２.２　日本語諸方言コーパス構築作業における問題点

２.２.１　発話単位（音声）と方言テキストの紐づけ作業上の問題点

２．１節に示した通り、発話単位の認定（0.2 秒以上のポーズで区切られる有音区間を切り出す）作業は、まずプログラムによって行った。一定以上の音声の波形の振幅がある箇所とない箇所の境に自動で境界（時間情報）を入れるというプログラムである。

その結果、図 4 の矢印（➡）で示したような境界が入ったテキストグリッドが出来上がる。この境界に区切られた区間のうち、音声の波形の振幅が目立つところが有音区間であり、発話単位と認められる可能性がある区間である。

ここで「可能性がある」としたのは、後述するいくつかの理由で、この段階では正確な発話単位として切り出されていない可能性があるからである。その問題を、作業者の手で修正し（発話単位の修正）、そこに方言の文字化テキストをペースト（発話単位と方言テキストの紐づけ）していくことになる。

この作業時に問題になるのが、次のようなことである。

①背景の雑音などによる境界の過剰付与
②複数名の発話の連なり・重なりによる発話単位の結合
③単語の途中での強調された促音や、言いよどみなどによる発話単位の断絶

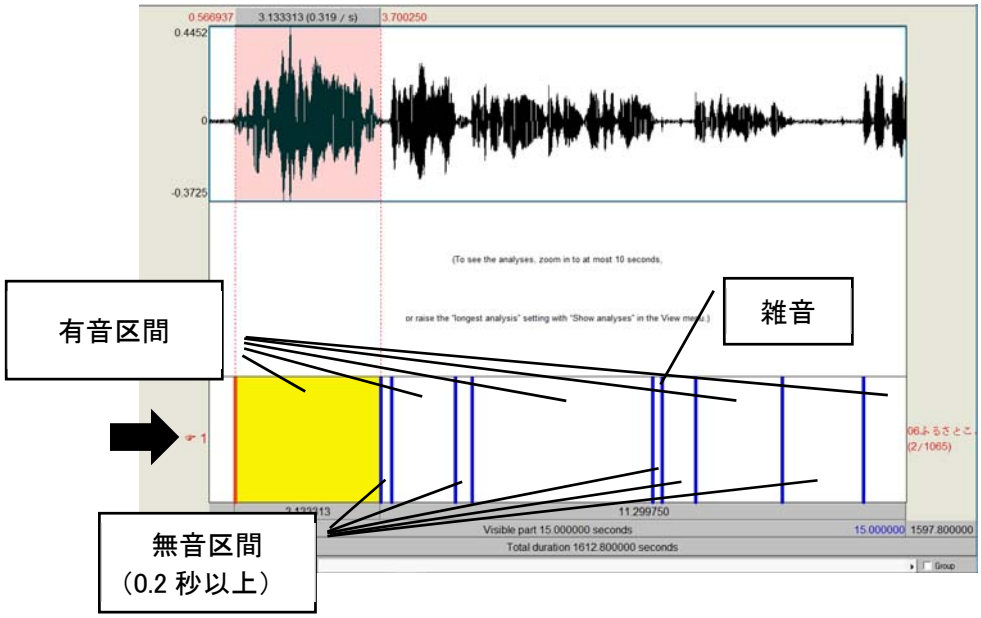

図 4．プログラムによる有音・無音区間境界付与（ふるさとことば集成：山形談話）

①については、『ふるさとことば集成』に発話とともに録音されてしまった様々なノイズ（蝉の声、電話の音、機械音など）によって、発話単位と別のところで音声波形が生じ、適切に発話単位を区切れなくなってしまうという問題である。例えば、図 4 には、前後に波形のまとまりがほぼ見られないのにも関わらず、境界が入っている部分（雑音）がある。これは録音時に入った雑音の波形が一定以上の振幅を見せたときにそれを拾ってしまった「雑音境界」である。図 4 では、一瞬の雑音であるため無視すれば問題ないが、ある程度の長さの雑音になると、そこに重なった発話が区切れなくなる。これについては、すべて作業者が音声を確認し、正確に発話単位に境界を付与する作業を行っている。

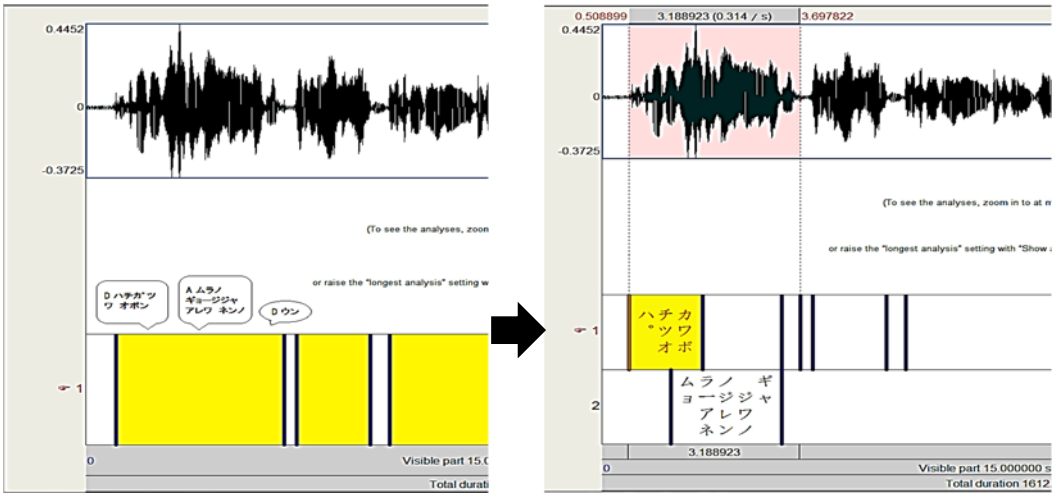

図 5a．発話の連なり・重なりによる結合　　図 5b．連なり・重なりの修正処理

②は、発話単位認定の過程で必ず処理しなければならない問題だが、複数名の発話が立て続けに（0.2 秒以内のポーズなく）なされたり、前の発話に覆いかぶさる形でなされたりすると、複数名による発話が結合された区間ができてしまうのである（図 5a）。これについては、層（Tier、図 5b の➡）を増やし、手作業で発話単位を区切ることで処理している。

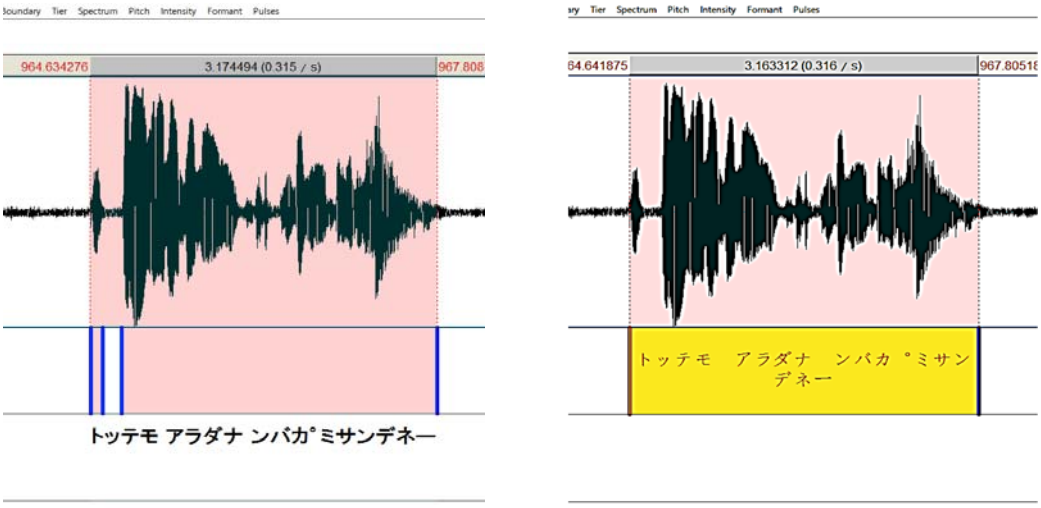

図 6a．促音による発話単位の断絶　　図 6b．促音による発話単位断絶の処理後

また、一発話単位と認めうる区間に、境界が入って断絶してしまうこともある。促音や言いよどみによる無音（無波形）区間が続く場合である。例えば、「とても」を強調して「とっても」という場合、「っ」の構えが 0.2 秒以上続くと、その間、音声波形が途切れ、境界が入って発話単位が断絶してしまうのである（図 6a）。単語の途中で言いよどみによる間ができた時も同様である。これらについて手作業で境界を消し、発話単位をつなげる処理をしている（図 6b）。ただし、この作業を行う上では、こういった途切れのうち、非語彙的な現象を、タグ付けするかどうかという点で議論がある。

２.２.２　方言テキストと標準語テキストの紐づけ作業上の問題点

CJD は、２．１節に示した通り、標準語テキストへのタグ付け作業を行う。タグの付け方次第で、コーパスは有用なものになる。例えば、西日本、特に九州では、「アンタ」「オマエ」などの形式が、発話の中の様々な箇所に頻繁に現れる。これらの形式は、標準語テキストに訳を当てる時は、直訳の方針に従って「あなた」や「おまえ」と訳すことになるが、機能としては「フィラー」として使われることが指摘されている（山本、松田 2015）。こういったものを他地域のフィラーと比較するために横断的に検索できるようにするには、「フィラー」を示すタグ「F」を用い「＜F：あなた＞」のようなタグ付けが必要になるわけである。

また、このようにタグをつけることで、各都道府県の作業担当者の違いによる標準語テキストの訳のゆれにもある程度対応できるようになる。例えば、『ふるさとことば集成』では、標準語のフィラー「ホラ」相当の形式には、全国を横断的に見渡した場合、少なくとも「ほら」という訳と、「ほれ」という訳が当てられている。これらのゆれも、＜F：ほら＞、＜F：ほれ＞のようにタグをつけることで、一括して検索することができる。

ただし、このタグの付け方に関連して、考えておくべき問題点が考えられる。

①どのような情報をタグとして採用するか
②標準語訳に対応する形式がない場合の処理をどのようにするか

①に関しては、各地の方言の特徴を踏まえて横断的に考え、先のフィラーのように、問題が生じる場合を想定し、それに対応できるタグを洗い出す作業を進めている。先に示した通り、標準語テキストの訳にゆれが生じやすいものや、直訳した標準語訳の品詞のはたらきと、運用上のはたらきにずれが生じるようなものにタグを付ける方針で設計を行っている。

また、標準語テキストに訳出するのが難しい方言形式のタグ付け処理をどのようにするかという問題もある。例えば、指小辞や副助詞、終助詞といった類いが問題となる。現時点では、対応する標準語が訳出できないものは「Z」でその存在を示し、それとともにタグ付けすることで検索を可能にする処理を施している。

(5)

文字化	標準語訳
ハヤグ**バリ**　シテ　ケサエンネ**ヤ**。	早く**＜副助：Z＞**　して　ください**＜終助：ねZ＞**。

(6)

文字化	標準語訳
ツリ**コ**サ　エク゜ダエンダゲントモ	釣り**＜指小辞：Z＞**に　行きたいんだけれども

(7)

文字化	標準語訳
オヨカ゜シェデ　クンデ　ナエ**ガワヤ**。	泳がせて　くるんで　ない**＜終助：かＺＺ＞**。

(7)のように、複合形式の場合（ガ＋ワ＋ヤ）は、対応する終助詞の単位数分を「Z」で示している。ただし、この場合も、そもそも対応する単位数を示した方がよいか、厳密に示すことができるか（融合した形式の単位認定など）という点で議論がある。

3. 日本語諸方言コーパスを用いた方言研究

本節では、CJD の検索を使って、どのような研究の方向性が開けるか、また、どのような使い方をすると不適切となるのか、具体的な検索結果を用いて、分析事例を示す。

３.１　「ホラ」相当形式に関する検索結果を用いた分析

本節では『ふるさとことば集成』を使った分析事例として、全国の「ホラ」相当形式を検索した結果を示す。なお、ここでは標準語訳の「ほら」とそのバリエーションによる検索で得られた方言形式を対象とした分析を行うため、「「ホラ」相当形式」と称した。

標準語に関する先行研究では、「ホラ」は、聞き手への注意喚起を促し、話し手と聞き手の共有知識に働きかけたり、あたかも知識などが共有可能だと表現したりするものと記述されている（大島 2001）。方言の談話展開の研究においても、情報共有喚起を促す談話マーカーとして、その使用の地域差が指摘されていたり（久木田 1990、琴 2005）、立ち上げ詞としての「ホラ」のバリエーションが記述されていたり（方言研究ゼミナール編 2006）する。だが、そもそも標準語「ホラ」の用法の全貌や、全国的な使用実態の地域差が明らかになっているとは言えない。そこで、今回『ふるさとことば集成』のデータを資料として分析を行った。

なお、『ふるさとことば集成』は、昔のことが話題に上がる会話が多いという性質から、聞き手に「ホラ」相当形式を用いて働きかける文脈が多く、用例の採取に適している。それは、話し言葉を扱った他のコーパスの検索結果との比較を見ても明らかである（表 2）。

表 2. 話し言葉のコーパスの「ホラ」検索結果

コーパスの種類	総時間（分）	「ほら」度数	出現比率
名大会話コーパス	6000	631	0.3
日常生活のことば	1058	198	0.5
ふるさとことば集成	1393	508	1.0

検索結果を抽出するにあたって、現時点ではまだデータ全体のタグ付けなどの整備が終わっていないため、検索ワードの選定の仕方が重要となる。標準語の談話研究で得られた「ホラ」のバリエーション（日本語記述文法研究会 2009、中島 2011）を参照しながら、全地点の標準語テキストのデータを形態素解析して「ホラ」と等価の形式を洗い出し、最終的に「ホラ」の音声的変異まで含めて網羅できる「ほら」「ほれ」の 2 形式で検索を行うことにした。なお、この検索ワードで検索をかけたところ、標準語テキストの訳の仕方に偏りが見られることが分かった。具体的には、静岡県の「ホラ」相当形式の検索結果のうち、40 例中 29 例が「ほら」、11 例が「ほれ」と当てられていた。そのため、検索時にこういったバリエーションへの目配りがなければ、重大な結果の相違を生んでしまう可能性がある。

今後、利用者がこういった問題に陥らないように、検索の仕方を検討する際の資料（標準語訳使用単語一覧、タグ情報一覧など）を公開し、コーパスの性質の周知を図るとともに、

検索システムの仕様の検討や研究実践を通じた利用方法の周知を行うことを検討している。

３.１.１　望ましくない分析

表 3（次頁）は、『ふるさとことば集成』の談話における「ホラ」相当形式を横断的に検索した結果である。数値は、それぞれの地域で得られた用例の度数と、何秒当たりに 1 回観察できるかという頻度を示したものである。例えば、北海道は 371 秒に 1 回「ホラ」が観察できるということを意味し、数値が小さくなるほど「ホラ」が目立つ談話ということになる。

表 3 からは、特に東北などの東日本や九州の一部の地域、さらに高知などで頻度が高く、一方、近畿を中心とした西日本にはほとんど見られないという結果が得られた。ここで、この頻度の差が「ホラ」相当形式の使用実態の地域差だと分析することについては慎重になるべきである。なぜなら、そもそもこの談話は、各地域 2 人～5 人の話者の 30 分程度の談話から導き出された結果であり、地点ごとのデータ量の少なさという問題がある上に、フィラー使用の個人差なども想定されるため、即地域差と断定するのは危険だからである。

表 3. 都道県別「ホラ」検索結果

都道府県（地域）	度数	1回／～秒	都道府県（地域）		度数	1回／～秒
北海道	6	371.0	滋賀県		0	―
青森県	45	48.6	京都府		0	―
岩手県	23	122.3	大阪府		0	―
宮城県	43	30.8	兵庫県		8	228.8
秋田県	12	129.1	奈良県		3	673.0
山形県	8	201.0	和歌山県		2	952.5
福島県	19	73.8	鳥取県		9	34.2
茨城県	4	582.5	島根県		0	―
栃木県	32	65.1	岡山県		1	1656.0
群馬県	11	215.1	広島県		0	―
埼玉県	14	162.8	山口県		0	―
千葉県	7	324.4	徳島県		10	220.5
東京都	6	348.5	香川県		0	―
神奈川県	8	260.4	愛媛県		0	―
新潟県	4	548.8	高知県		104	19.4
富山県	8	164.8	福岡県		1	1417.0
石川県	5	257.2	佐賀県		13	96.4
福井県	1	1387.0	長崎県		2	736.0
山梨県	24	66.5	熊本県		14	90.4
長野県	1	1118.0	大分県		0	―
岐阜県	0	―	宮崎県		0	―
静岡県	40	35.6	鹿児島県		19	108.9
愛知県	0	―	沖縄県A	那覇	0	―
三重県	0	―	沖縄県B	宮古島	1	720.0

３.１.２　望ましい分析

前節で「ホラ」相当形式の使用量の差について地域差とみる判断には慎重になるべきだと述べた。それならばCJDはどのような分析に有用と言えるのか。例えば、「ホラ」相当形式の分析については、次のような目的での利用が考えられる。

①各地域の「ホラ」相当形式のバリエーションの洗い出し
②日本語「ホラ」相当形式の持つ用法の幅の把握
③「ホラ」相当形式のイントネーションの比較

①に関しては、今回の検索によって、次のようなバリエーションが得られた。

(8) ホラ系（ホイ、ホエ、ホー、ホーラ、ホラ、ホラー、ホリ、ホリャ、ホレ、ホレー）、ハラ系（ハー、ハラ、ハレ）、アラ系（アラー、アリェ、アリャ、アレ、アレー、アレャ）、ソレ系（ソイ、ソラ、ソリェ、ソレ、ソレー）、オラ系（オラ、オレ、レ、レァ、レー、ロ、ロー）、オッキャ系（オッキャ、キャ、キャー）、ワヤ系（ワイ、ワヤ、ワヤー）、その他（クヤ、デ、ドー、ミナイ、メーデ）

これは、方言研究ゼミナール編（2006）を上回るバリエーション数と言える。なお、紙面の都合で割愛するが、地域ごとの使用バリエーションの差も見ることができる。

さらに用法についても、今回、従来の研究にない用法を見つけることができた。

(9) ソレン　ダンダン　アレン　ナッテクルダイナ　　アノ　ホレ　ケーケン
それが　だんだん　あれに　なってくるのだよな　あの　ほら　経験［に］

ナッテクルダイ。
なってくるのだよ。　　（『ふるさとことば集成』静岡 151、下線は筆者による）

この例は、「ソレン　ダンダン　アレン　ナッテクルダイナ」と「アレ」の適切な表現が思い出せなかったことについて、「アノ　ホレ」の部分で自分の記憶に情報喚起を促し、その結果、適切な表現としての「ケーケン」という言葉が思い浮かんだという例である。つまり、ここでの「ホレ」は、他者に注意喚起を促すようなものではなく、自分に働きかけていることを示す情報検索表示の用法と見られる。これは、内省する限り、標準語の「ホラ」でも可能で、従来の理論的な研究からは漏れていた「ホラ」の新たな用法を、今回の検索を通して見出すことができたと言える。また、このような検索で方言独自の用法が見つかれば、日本語の「ホラ」相当形式の持つ用法の広がりをつかむことにもつながる。

③に関しては分析が及ばなかったが、CJDは検索結果を通して、同じバリエーションの音声の違いを手軽に聞き比べることができることもメリットと言える。

４．おわりに

本発表では、現在、構築作業を進めている『日本語諸方言コーパス（CJD）』の概要と特徴、構築のプロセス、及び本コーパスを使った方言研究の一例を紹介し、方言コーパスの可能性と使用の際の注意点を指摘した。一般公開は2021年度の予定であるが、その前にモニター公開を行い、モニタリングの結果報告を受けてCJDをさらに改善し、一般公開へとつなげる予定である。興味のある方は、ぜひ、ご協力をお願いしたい。

謝　辞

本研究は、2010～2015 年度 国立国語研究所共同研究プロジェクト「消滅危機方言の調査・保存のための総合的研究」、2016～2021 年度 同プロジェクト「日本の消滅危機言語・方言の記録とドキュメンテーションの作成」、2013～2015 年度 科研費基盤研究(B)「方言話し言葉コーパスの構築とコーパスを使った方言分析に関する研究」（課題番号 25284087）、2016～2020 年度 科研費基盤研究(A)「日本語諸方言コーパスの構築とコーパスを使った方言研究の開拓」（課題番号 16H01933）の支援を受けて行った。

文　献

大島弘子(2001).「「ほら」の機能について」『日本語教育』108 号, pp.34-41.

木部暢子(2015).「対格助詞ゼロの地域差－方言コーパスの可能性ー」日本方言研究会第 101 回研究発表会発表原稿集

琴鍾愛(2005).　「日本語方言における談話標識の出現傾向—東京方言、大阪方言、仙台方言の比較—」『日本語の研究』1 巻 2 号，pp.1-18.

久木田恵(1990).「東京方言の談話展開の方法」『国語学』162 号，pp.1-11.

国立国語研究所(編)(2001-2008).『全国方言談話データベース 日本のふるさとことば集成』国書刊行会.

中島悦子(編)(2011).『自然談話の文法—疑問表現・応答詞・あいづち・フィラー・無助詞—』, おうふう.

日本語記述文法研究会(編)(2009).『現代日本語文法 7 第 12 部 談話 第 13 部 待遇表現』, くろしお出版.

方言研究ゼミナール(編)(2006).『日本語方言立ち上げ詞の研究』広島大学教育学部国語教育学研究室方言研究ゼミナール.

松田美香(2015).「大分と首都圏の依頼談話」—大分方言の「アンタ」「オマエ」のフィラー的使用について—」『別府大学紀要』56 号，pp.11-22.

山本空(2015).「方言談話における対称詞の使用量の地域差」『国文学』100 号, pp.482-466.