

科学研究費助成事業 研究成果報告書

平成 28 年 6 月 21 日現在

機関番号：62618

研究種目：基盤研究(B) (一般)

研究期間：2013～2015

課題番号：25284087

研究課題名(和文) 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

研究課題名(英文) Constructing a speech corpus of Japanese dialects and a research on the dialects using corpora

研究代表者

木部 暢子(Kibe, Nobuko)

大学共同利用機関法人人間文化研究機構国立国語研究所・言語変異研究領域・教授

研究者番号：30192016

交付決定額(研究期間全体)：(直接経費) 13,000,000円

研究成果の概要(和文)：本研究の目的は、日本語諸方言を同じ条件で横断的に検索する「方言コーパス」を構築すること、及びそれを使った新しい方言研究の方法を提示することである。そのために、まず6地点の方言データを使用してデータ整備の方法を検討した。その結果、方言テキストと共通語訳テキストを平行に整備し、共通語から方言形と方言音声を検索する方法でコーパスを構築する方法が有効であることがわかった。次に、27方言のデータによる「方言コーパス試作版」を作成し、格助詞の使用に関する分析を行った。その結果、文脈に即した分析が可能な点で、コーパスを使った研究は方言研究に新しい展開をもたらす可能性が高いことを示した。

研究成果の概要(英文)：The purpose of this study is to construct a corpus of Japanese dialects in which various dialects are searchable in the same manner, and to offer new methods for research on dialects using the corpus. First, we examined methods of how to organize the data, using data of dialects from six regions. We figured out that it is effective to construct a corpus where texts of dialects and their translation in standard Japanese are displayed in parallel and also texts and sounds of dialects are searchable in standard Japanese. Second, we built a trial version of the corpus of Japanese dialect, which contains dialect data from 27 regions, and then conducted an analysis of uses of case particles in the corpus. Based on the findings from the analysis, we presented a new analysis of dialects with reference to contexts, and thus, we claim that such a corpus-based study will advance the research on dialects further.

研究分野：人文学

キーワード：方言コーパス 方言談話資料 パラレルコーパス 方言テキストデータ 方言音声 日本のふるさとことば集成 格助詞の地域差 「です」の使用の地域差

1. 研究開始当初の背景

近年、言語研究の分野では、実際の使用例を網羅的に集積した言語コーパスの構築とそれを使った言語研究の重要性が指摘されている。日本語に関しては、「日本語話し言葉コーパス (CSJ)」や「現代日本語書き言葉均衡コーパス (BCCWJ)」、「日本語歴史コーパス (CHJ)」(いずれも国立国語研究所)が整備され、それを利用した日本語研究が成果を上げている(文献②、⑦、⑪、⑬)。一方、方言に関しては、これまで方言談話資料がいくつか作られているものの(文献⑤、⑥、⑨)、コーパスの整備が遅れている。特に音声付きの方言コーパスは、日本にはまだない。

面接聞き取り調査で得られるデータとコーパスにより得られるデータは、それぞれ性質を異にしており、相補い合って研究の質を高める関係にある。コーパスを使った新たな手法を導入することにより、これまで見過ごされていた方言の特徴が見えて来る可能性があり、方言研究は大きく進展すると予測される。

また、方言が急速に衰退しつつある現在、大量のデータを処理する方言コーパスは、方言の記録・保存・公開の方法としても有効である。その意味で方言コーパスの構築の必要性は高い。

2. 研究の目的

(1) 第1の目的は、「方言コーパス」構築のための方法を確定することである。その際の問題点は、形態素辞書がほとんど完備されていない諸方言において、コーパスをどのように構築するかである。これには2つの方法が考えられる。1つは、諸方言の形態素辞書を作成し、その後、検索システムを構築する方法、2つは、パラレルコーパスの考え方を取り入れ、共通語訳を利用して検索システムを構築する方法である。前者に関しては、方言ごとに形態素辞書を作成しなければならず、現時点では現実的ではない。後者に関しては、近年、言語間のパラレルコーパスが作られるようになり、その方法を応用することが可能である。以上のような状況を考慮しながら方言コーパス構築の方法を検討し、確定する。

(2) 第2の目的は、「方言コーパス」を使った言語分析のモデルを提示することである。コーパスを使った方言分析の長所は、一定量のデータから無作為に用例を抽出することができる点、文脈情報を取り入れた言語分析ができる点等にある。一方、話題や話し手、聞き手、会話の場面等により使用語彙に偏りが出るといった注意点もある。これらの注意点を考慮しつつ、方言コーパスの導入により、どのような分析が可能になるのかについて考察し、そのモデルを示す。

3. 研究の方法

(1) 「方言コーパス」の概要：「方言コーパ

ス」とは、コンピュータの画面に検索条件を入力すると、それに対応する方言形を含む方言テキストと共通語訳テキストが表示され、方言音声を再生することもできるシステムのことである(図1)。本コーパスでは、方言データとして国立国語研究所が所蔵する「各地方言収集緊急調査」のデータを使用する。これは、1977年から1985年にかけて文化庁が行った事業による方言音声データで、原資料は全都道府県224地点、1地点につき10~30時間の談話を収録した約4,000本のカセットテープよりなる。その一部は『方言音声談話資料 日本のふるさとことば集成』(佐藤亮一・井上文子編、2001~2008、国書刊行会)として公開されている。本研究ではこの公開データ(以下「ことば集成」と呼ぶ)を使用した。将来的には未公開データを取り込む予定である。

313-000-1	北海道	C	ウチラフ ソト ウチヤ	私たちは 外。うちは	▶
8-002-1	弘前	C	ウエダケンタネスー [笑]	私のようなね [笑]	▶
25-000-1	秋田	B	オラハダテ アノ	私だって あの	▶
195-000-1	茨城	B	ワカンネーガ' アダシラ。	わからないが 私達は。	▶
176-000-1	群馬	B	ワタシガ ジューロクシチダネ。	私が 16、7[歳]だね。	▶
45-006-1	大阪	E	ワタシラヤツラ	私らだったら	▶
178-001-1	北九州	B	ワタシガ オモ オモイダシテモ。	私が × × 思い出しても。	▶
251-000-1	大分	C	アー ワタシヤ スル。	ああ 私は する。	▶
246-000-1	鹿児島	C	アタヤ ミダド。	私は 見たよ。	▶
104-000-1	沖縄	A	アー ソー ソー ワヌー	ああ そう そう 私	▶

図1 検索結果画面イメージ

(2) 「方言コーパス」構築の方法と問題点の検討：本研究では、パラレルコーパスの手法で「方言コーパス」を構築する。サンプルデータとして「ことば集成」のうちの青森県弘前市、東京都台東区、石川県羽咋郡押川、大阪市、広島市、福岡県北九州市の6地点、各30分のデータを使用し、具体的なデータ整備の方法と問題点の検討を行った。主な内容は以下のとおりである。

①方言テキスト、共通語訳テキストの整備：「ことば集成」は方言音声、その文字化テキスト、共通語訳からなる。データはすでにチェック済みであるが、「ことば集成」では方言テキストを共通語テキストに置き換える方向でデータの整備が行われている。これに対し、「方言コーパス」では、共通語から方言形を検出するので、その方向でデータを点検し直す必要がある。この作業が大きな課題となる。

②方言テキストと音声のリンク付け：「方言コーパス」では、検索画面上で方言音声が再生できるようにするため、テキストデータに音声データの時間情報を付与し、両者をリンクさせる必要がある。本研究では最初、手作業でこの作業を行ったが(図2)、研究機関後半にこの作業を音声認識ソフトで自動的に行うことができないかについて検討した。

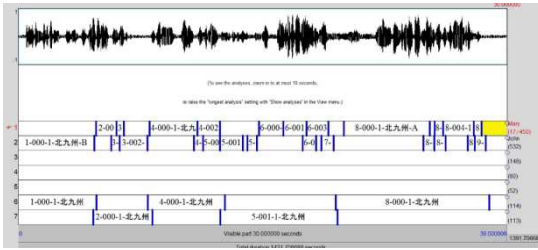


図2 音声データと方言テキストデータのリンク付け

(3)「方言コーパス試作版」の作成：上記6地点のデータを使って、検索システムを構築した。その後、地点数を増やし、27地点のデータによる「方言コーパス試作版」を作成した（試作版のため非公開）。

(4)コーパスを使った方言分析：「方言コーパス試作版」を使って、方言の分析を行った。主な内容は、①コーパスを使った方言分析の長所と注意点の検討、②格助詞（共通語の「が」「を」「に」等に当たる表現）の地域差の分析、③丁寧の助動詞「です」の使用の地域差の分析である。①は第39回九州方言研究会（2015年1月11日、熊本大学）で発表した。②は合同シンポジウム「日本語の Aspekte・ヴォイス・格」（2015年8月23日、国立国語研究所）、及び日本方言研究会第101回研究発表会（2015年10月30日、パルトピア山口）で発表した。③はNINJAL 合同シンポジウム「正しい日本語」ってなに？ーコーパスに見る日本語のバリエーション（2015年9月3日、国立国語研究所）で発表した。

4. 研究成果

(1)「方言コーパス試作版」の作成：パラレルコーパスの方法による「方言コーパス試作版」（27方言のデータによるコーパス）を作成した。「3. 研究の方法」の(2)で述べた2つの問題については、次のような対応を行った。

①共通語から方言形を検出するためのデータの整備：方言と共通語は1対1で対応しない。例えば、西日本方言の「タベヨル」と「タベトル」は、共通語の「食べている」に対応する。また、青森方言の「アズマシー」は、共通語形を対応させるのが困難である。このような中で共通語から方言形を検索するためにどのようなテキスト整備を行えばよいか。まだ、試行錯誤の段階であるが、現時点では検索条件の共通語に最も近い方言形を検出し、検索結果に関しては、利用目的に合わせて利用者が吟味を行うのが最良の方法であると考えられる。上の例でいうと、「食べている」で西日本方言の「タベヨル」と「タベトル」を検出し、また、「気持ちいい」で青森方言の「アズマシー」を検出する。検索結果に関しては、前後の文脈を考慮して利用者が吟味した上で

利用するというものである。そのためには、検索条件の共通語と検索結果の方言形が必ずしも1対1で対応するものではない、ということを利用者に周知しておく必要がある。また、品詞検索、意味分野検索、文法カテゴリー検索等、いろいろな検索ができるようにしておくことも重要である。

②方言テキストと音声のリンク付け：基本的に手作業でこの作業を行った。しかし、データが大量になるとを考え、音声認識ソフトを使ってこの作業を効率化する方法を検討した。その結果、東京都台東区、北九州市等、共通語に比較的近い方言ではある程度、自動認識が可能であることがわかった。今後、自動化できる部分は自動化し、作業を効率よく行う可能性がある。

(2)コーパスを使った言語分析ー長所と注意点：

①話題に左右されやすい現象の検索には、注意が必要である。個別の単語の出現状況はその典型で、例えば、「雨」にあたる単語は、話題によって一度も出現しないこともある。何度か出現することもある。

②アспектも話題に左右されやすい。例えば、自然談話では昔の思い出を語る場面が多くなる。その場合、過去の習慣を表すアспект（「昔はよく～していた」等）の出現頻度が高くなり、逆に発話時の動作の継続を表すアспект（「今ごはんを食べている」等）の出現が低くなる。

③敬語の研究には「方言コーパス」が有効に働く。自然談話では、敬語表現が頻繁に現れるため、コーパスを使えば敬語のデータが文脈付きで多数収集できる。一方、敬語の使用は話し手や聞き手の属性（男女、年齢、社会的立場等）、両者の関係（年齢の上下や社会的立場の上下、親疎等）、及び話題等に大きく左右されるので、敬語の地域差を見るときには注意しなければならない。例えば、A地域でaという敬語が使われ、B地域でbという敬語が使われている場合、aとbを単純に比較して敬語の地域差を論ずることはできない。A地域の談話とB地域の談話とでは、話し手や聞き手の属性、両者の関係、話題等が異なっており、それらを考慮した上で比較する必要がある。

人称代名詞も同じで、話し手、聞き手、両者の関係、場面等を考慮せずに人称代名詞だけをコーパスで検索して、地域差を論ずることはできない。4(1)①で述べたように、検索結果の利用にあたっては、利用者が検索結果を吟味した上で用例を使用するというルールを確立することが重要である。

④格標示形式（共通語の「が」「を」「に」等）のように文の基本的な要素は、出現頻度が高

く、話し手や聞き手の属性、話題等に左右されることが比較的少ないため、コーパスを使った言語分析が有効に働く。ただし、焦点化、主語や目的語の属性等により使用形式が変わる場合があるので、ここでも前後の文脈を十分吟味する必要がある(文献③、④、⑩)。

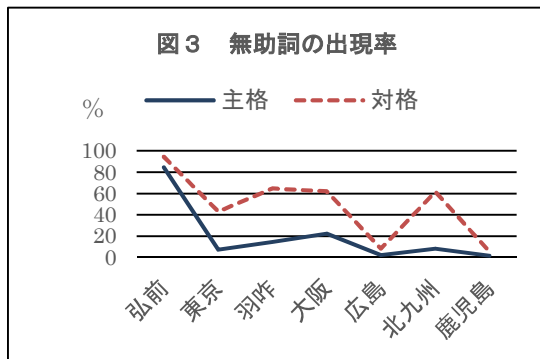
以上のように、コーパスの利用に際しては、大なり小なり、話題、話し手や聞き手の属性、関係性、前後の文脈等に注意を払う必要がある。

(2) コーパスを使った言語分析—格助詞の地域差

「方言コーパス試作版」を使って主格助詞(共通語の「が」に当たる表現)、対格助詞(共通語の「を」に当たる表現)の地域差について考察した。特に、無助詞(助詞を使わずに主格や対格を標示する現象、例えば、「太郎 ごはん 食べた」等)の出現状況に注目し、地域差を分析した。その結果、以下のような地域差があることが明らかになった。

- ・青森県弘前市方言では、主格、対格ともに無助詞が基本である。
- ・広島市方言、鹿児島市方言では、主格を「ガ」、対格を「オ」で標示するのが基本である。
- ・東京都台東区方言、石川県羽咋方言、大阪市方言、北九州方言では、主格を「ガ」、対格を無助詞または「オ」で標示する。無助詞の出現の度合いを高い順に並べると、以下ようになる。

弘前>羽咋・大阪・北九州>東京>広島 = 鹿児島(図3)



- ・無助詞の出現には、名詞句と動詞との距離、焦点の位置、名詞句の意味の特定性等が要因として働いている(文献①、⑫、⑧)。例えば、対格名詞句が動詞に隣接する場合、大阪市方言では、対格は無助詞が基本だが(オモテ ミテマス[表(を)見えています])、隣接しない場合は助詞「オ」が現れることが多い(オーサカベンオー モー バッカリベッタリ ヤルノー[大阪弁を もう ばかりべったり やるのを])。

(3) コーパスを使った言語分析—丁寧の助動詞「です」の使用の地域差

「動詞句+です(行っただです)」、「形容詞句+です(ないです)」の使用の地域差の分

析を行った。その結果を以下に示す。

- ・「方言コーパス試作版」の限りでは、「動詞句+です」、「形容詞句+です」は、福島県より北には用例がない。
- ・茨城県水戸市方言では、「ナイデスカ」、「ナカタデスカ」のようにもっぱら形容詞に「デスカ」が付き、同意要求に使われる。また、千葉県長生町方言では、形容詞「ネー(ない)」に「デス」が続いた例がある。
- ・「動詞+デス」は石川県より西で見られる。石川県羽咋方言では、過去の出来事を説明的に語る場合に、「動詞過去形+デス(〜トオモタデスワ)」が使われる。
- ・大阪市方言、広島市方言では「動詞+マシタ+デス」が多く見られる(イーマシタデス[言いました])。ただし、「動詞+マス+デス」(*イーマスデス)は用例がない。
- ・九州方言(宮崎市方言を除く)では「動詞活用形+デス」「形容詞活用形+デス」が多用され、「デス」が丁寧を表す接辞形式として生産性を持って使われている(オモースネ[思いますね]、ワカランダスヨ[分かりませんよ]、アマカデスネ[甘いですね])。
- ・動詞句では「過去形+デス」(「思っただです」)の方が「終止形+デス」(「思うです」)よりも分布が広く、形容詞では「終止形+デス」(「ないです」)の方が「過去形+デス」(「なかったです」)よりも分布が広い(図4)。ただし、この差に話題等の文脈情報がどの程度影響しているかについては、まだ考察しておらず、今後の課題である。

図4 27地点の「動詞+デス」「形容詞+デス」の数

	動詞				形容詞		
	終止 行く デス	過去 行った デス	否定 行かない デス	否定過去 行かなかった デス	マス 行きました デス	終止 ないデ ス	過去 なかつ たデス
茨城		1	3	2		1	
千葉		1				3	
羽咋		3				2	
大阪		1	1		11	2	1
広島		3	1	1	2	4	
北九州	1	4	1		1	3	
長崎	14	11	4	1		2	
熊本	3	2	2				
宮崎						2	1

<引用文献>

- ①阿部貴人(2009)「対話における無助詞化の地域差—東京・大阪・津軽方言の対照から」『月刊言語』38-4, pp. 40-46
- ②小磯花絵編(2015)『講座日本語コーパス3』朝倉書店
- ③小西いずみ(2015)「広島方言の対格表示—談話資料による軽量的把握—」『国語教育研究』56, pp. 13-24.
- ④佐々木冠(1998)「水海道方言の対格—有生対格と無生対格の統語論—」『日本語科学』4, pp. 99-120
- ⑤佐藤亮一・井上文子編(2001~2008)『方言音声談話資料 日本ふるさとことば集成』国書刊行会

- ⑥杉藤美代子代表(1989～1992) 科研費特定領域研究「日本語音声」研究成果CD
- ⑦田野村忠温編(2014)『講座日本語コーパス6』朝倉書店
- ⑧玉懸元(2002)「仙台市方言における格助詞相当『ドゴ』の用法」『国語学会 2002 年度秋季大会予稿集』pp. 127-132.
- ⑨日本放送協会(1959～1972)『全国方言資料』
- ⑩日高水穂(2000)「秋田方言の文法」秋田県教育委員会編『秋田のことば』無明舎出版.
- ⑪前川喜久雄編(2013)『講座日本語コーパス1』朝倉書店
- ⑫松田謙次郎(2000)「東京方言格助詞『を』の使用に関わる言語的諸要因の数量的検証」『国語学』51-1, pp. 61-76.
- ⑬山崎誠編(2014)『講座日本語コーパス2』朝倉書店

5. 主な発表論文等

〔雑誌論文〕(計12件)

①日高水穂、近畿中央部方言におけるシテイル相当形式の動態—現在形と過去形の非対称現象をめぐって—、国文学(関西大学国文学会)、査読有、No. 100、2016、pp. 85-99.

②田附敏尚、青森県五所川原市方言の文末形式「デバ」について、Theoretical and Applied Linguistics at Kobe Shoin、査読無、No. 19、2016、pp. 123-135.

③木部暢子、鹿児島方言辞典—遊戯の部—、国語国文薩摩路、査読無、No. 59、2015、pp. 1-7.

④五十嵐陽介・松浦年男、天草諸方言のアクセント資料の提示と新しいアプローチに基づいた西南部九州諸方言の系統分析の試み、九州大学言語学論集(KUPL)、査読有、No. 35、2015、pp. 71-110.
http://www2.lit.kyushu-u.ac.jp/~linguist/kupl/doc/kupl35_Igarashi.pdf

⑤木部暢子、鹿児島方言の「イッ」と「イタッ」—テキストを使った方言研究の実践—、西日本国語国文学、査読有、No. 1、2014、pp. 1-14.

⑥木部暢子、奄美喜界島方言の親族語彙—お父さん・お母さん・お爺さん・お婆さん—、国語研プロジェクトレビュー、査読無、Vol. 5、No. 2、2014、pp. 57-67.
<http://www.ninjal.ac.jp/publication/review/0502/>

⑦Nobuko Kibe and Kaori Ototake、Regional Differences in the Usage of 'Yes' and 'No' in Response to Negative Interrogatives in Japanese”、Second

International Conference on Asian Geolinguistics、査読無、2014、pp. 222-227.

〔学会発表〕(計18件)

①木部暢子、対格助詞ゼロの地域差—方言コーパスの可能性—、日本方言研究会第101回研究発表会、2016年10月30日、パルトピア山口(山口県・山口市)。

②木部暢子、方言コーパスの構築と利用、第39回九州方言研究会、2015年1月11日、熊本大学(熊本県・熊本市)。

③木部暢子、方言コーパスの構築と日本語研究、合同研究発表会「コーパスに見る日本語のバリエーション—会話・方言・学習者・歴史コーパスから—」、2014年12月6日、国立国語研究所(東京都・立川市)。

④五十嵐陽介、句末境界音調のピッチレンジ制御に関わる要因、日本音声学会第328回研究例会、2013年12月7日、日本大学(東京都)。

⑤日高水穂、愛知県の昔話の「語りの型」の特徴、中部日本・日本語学研究会、2013年11月2日、刈谷市総合文化センター(愛知県刈谷市)。

⑥新田哲夫、越前海岸のN型アクセント、第27回日本音声学会全国大会公開シンポジウム N型アクセントの諸相、2013年9月29日、金沢大学(石川県・金沢市)。

⑦木部暢子、鹿児島方言の「イテ」と「イタテ」、九州大学国語国文学会、2013年6月8日、九州大学(福岡県・福岡市)。

⑧日高水穂、昔話の「語りの型」の地域差—文末表現を中心に—、日本語学会 2013 年度春季大会、2013年6月2日、大阪大学(大阪府・豊中市)。

⑨木部暢子・中山俊秀・下地理則・大槻知世・かりまたしげひさ、日本語学会ワークショップ テキストを使った方言研究から見えてくること—危機方言の調査と記述—、日本語学会春季大会、2013年6月1日、大阪大学(大阪府・豊中市)。

〔図書〕(計4件)

①木部暢子、岩波書店、そうだったんだ日本語—じゃつで方言なおもしとか—、2013、199.

6. 研究組織

(1) 研究代表者

木部 暢子 (KIBE, Nobuko)

大学共同利用機関法人人間文化研究機構
国立国語研究所・言語変異研究領域・教授
研究者番号：30192016

(2)研究分担者

新田 哲夫 (NITTA, Tetsuo)
金沢大学・歴史言語文化学系・教授
研究者番号：90172725

日高 水穂 (HIDAKA, Mizuho)
関西大学・文学部・教授
研究者番号：80292358

五十嵐 陽介 (IGARASHI, Yosuke)
一橋大学大学院・社会学研究科・准教授
研究者番号：00549008

井上 文子 (INOUE, Humiko)
大学共同利用機関法人人間文化研究機構
国立国語研究所・言語変異研究領域・准教授
研究者番号：90263186

三井 はるみ (MITSUI, Harumi)
大学共同利用機関法人人間文化研究機構
国立国語研究所・言語変異研究領域・助教
研究者番号：50219672

田附 敏尚 (TATSUKI, Toshihisa)
神戸樟蔭女子学院大学・文学部・講師
研究者番号：90645813

(3)連携研究者

前川 喜久雄 (MAEKAWA, Kikuo)
大学共同利用機関法人人間文化研究機構
国立国語研究所・音声言語研究領域・教授
研究者番号：20173693