

平成 30 年 6 月 25 日現在

機関番号：62618

研究種目：基盤研究(C) (一般)

研究期間：2014～2017

課題番号：26370555

研究課題名(和文)大規模方言分布データの計量的分析方法の開発

研究課題名(英文)Development of quantitative methods of analyzing large-scale dialectal distribution data

研究代表者

熊谷 康雄 (KUMAGAI, Yasuo)

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：30215016

交付決定額(研究期間全体)：(直接経費) 3,600,000円

研究成果の概要(和文)：『日本言語地図』(LAJ)は日本の方言の地理的分布を一望できる基礎資料であり、昭和30年代に全国2400地点で調査し、作成された。LAJを使った全国的視野の計量的研究がなされたが、計算機データがなく、県別集計による研究であった。構築を進めている『日本言語地図』データベースは、2400地点のデータで、47都道府県の県別集計の約51倍の地理的な解像度である。全国的な方言分布データでは最大規模の詳細なデータを生かした観察、分析に有効な方法の開発を目的とし、方言分布形成に重要な道路網や人口分布などの言語外的情報(調査時およびそれ以前)をデータ化し、これを用いた計量的な観察・分析、方法の開発を行った。

研究成果の概要(英文)：Linguistic Atlas of Japan (LAJ), whose survey was conducted from 1957 to 1965, is one of the basic research materials capturing the geographical distribution of Japanese dialects in 2400 localities. While nationwide quantitative studies have been conducted on LAJ, most of them used calculations based on the 47-prefecture-unit. We have been developing LAJDB (Linguistic Atlas of Japan Database) to enable us to use the data of 2400 localities on computers. The nationwide "high resolution" patterns of 2400 localities is approximately 51 times the resolution of 47 prefectures. In this study, in order to explore the potential of large scale LAJ data in advancing the quantitative approaches, we prepared the data incorporating important extra-linguistic factors, such as historical road network and time series population distribution, and made some preliminary quantitative observations of LAJ data by utilizing the extra-linguistic factors.

研究分野：計量的方言研究、言語地図のデータベース化とこれを用いた方言分布研究

キーワード：日本言語地図 データベース 方言分布 言語接触 言語外情報 電子地図 計量的分析 現象解析

## 1. 研究開始当初の背景

日本の方言の全国的な地理的分布を展望できる基礎的な資料のひとつに『日本言語地図』(全6巻, 国立国語研究所 1966-1974, 以下, LAJ)がある。LAJは, 1957年から1965年にかけて日本全国2400地点で調査したもので, 言語地理学的調査方法による日本ではじめての本格的な言語地図である。昭和30年代の日本の方言分布を捉えた基礎資料として広く研究に用いられてきた。LAJには方言分布の計量的な研究の資料としても研究の蓄積がある。しかし, LAJは日本全国をカバーする言語地図(方言地図)では最大規模の2400の地点数を持つが, 全国を対象とした研究の多くは, 個々の研究者が地図を読み取って手作業で作成した県別の集計データによるものがほとんどであった。LAJの方言分布に多変量解析などの計量的手法を用いた研究の代表的なものとして井上史雄の研究(井上 2001 他)があるが, これも主としてLAJに現れた標準語形を手作業で県別に集計した頻度数をデータとしてなされている。LAJのコンピュータデータの欠如が, 全国規模の方言分布の計量的研究の制約になっていた面は大きかった。

我々が構築を進めている『日本言語地図』データベース(以下, LAJDB)は, LAJの原資料である調査の回答を記入した原カードを画像データ, LAJ上の語形の分布情報を文字データとしてデータベース化するものである。LAJDBを使えば, 2400地点の地点別データによる詳細な分布を対象とした研究が可能になる。

海外では, 方言分布を計量的に扱う研究としては, Hans Goebel の Dialectometry (Goebel 2006 他)などの蓄積があり, また, 多変量解析や, 近年では, 自動分類や画像処理の技術の応用による分布パターンの研究(例えば, Rumpf et al. 2010)など, 新たな試みが行われているが, これらの研究のデータはLAJDBの2400地点ほどの地点数, 地点密度を持たないものが多い。Goebelの代表的な研究はフランス言語地図(ALF)によるものであるが, 地点数は640地点程度である。面積地点密度は, LAJが6.5に対して, ALFは1.2である(国立国語研究所 1966)。

構築中のLAJDBのデータの一部を用いた予備的な観察で, これまでの計量的研究では, 大まかな県別の分布しか見ることができないでいた標準語形の全国分布や, 実態がほとんど知られていなかった事象に, 詳細で興味深い観察ができた(熊谷 2013a,b)。これにより, 構築中のLAJDBを用い, 全国的な方言分布データとしては最大規模の2400地点のデータを生かすことのできる新たな分析方法の開発を現実的に進めることができる見通しが得られ, LAJDBの分析の可能性を掘り起こす研究方法の開発を進めたいと考えた。

## 2. 研究の目的

方言分布の計量的研究では, LAJを用いて, 方言区画を求めたり, 方言の伝播の有り様を探ったりする研究の蓄積があるが, 全国的な視野で行われた研究のほとんどは, 県別の集計データをもとに行われている。構築を進めているLAJDBでは, LAJの2400地点の地点別の回答データが計算機で扱える。これは47都道府県の県別集計の約51倍の地理的な「解像度」を持つと言える(図1)。全国的な方言分布データでは最大規模の2400地点のデータを生かし, 県別データによる計量的分析では見えなかった方言分布, 方言伝播の詳細な観察と分析をするために有効な新たな方法の開発を目的とする。

## 3. 研究の方法

目的とする2400地点のデータを生かした詳細な方言の分布, 伝播の様態を分析するための方法を開発するために, 文献調査, データ整備, 探索的な分析を並行して, 相互に関係付けながら研究を進める。

(1) 文献調査: 方言の分布や区画を扱う計量的な諸手法に関わる文献調査や事例的な分析などを通して, 計量的な諸手法の方法論的な検討を行う。

(2) データ整備: LAJDBの分布データを分析するために有効な言語外的情報を検討し, データ化する。並行して, 有用な公開データの情報を収集する。

(3) 分析と方法の開発: 上記(2)のデータを活用し, 探索的な分析を試行し, これを通して, LAJDBの大規模方言分布データの計量的な分析方法の開発を進める。

## 4. 研究成果

### (1) 文献調査

地理情報, 地誌, 歴史, 交通・交通史, 人口統計などや, これらを用いた分析に関わる文献, 資料など, 関連領域も含めた文献調査を進め, データ整備や探索的分析と方法の開発に反映させた。過去の国勢調査, 統計資料, 統計地図, 交通史, 地誌などに関しては, 分析に関わる基礎的な資料として, できる限り幅広く調査を行った。データ整備した統計資料や統計地図などはこれらの文献調査に基づく。

### (2) データの整備

地理的分布に関する計量的な観察・分析方法についての検討や試行と並行して, 方言分布データの分析に有用な, 基盤的な言語外的情報に重点を置いてデータ化を進めた。

LAJの解釈に重要な明治大正期の道路網を詳細に記載している歴史的な地形図の電子化, 昭和30年代のLAJの調査時期を含む人口データの入力整備や, 人口の分布を示す人口地図の電子化など, 方言の分布が形成される上で重要で基礎的な情報のデータ整備に

力点を置いた（なお、これら古い時期の資料には公開データのないものが多い）。なお、利用可能な公開データの情報収集も継続的に行った。以下のデータを作成した。

#### LAJ 参考図の電子地図

LAJ の分布図の解釈のために編集された参考図の、方言分布データとの重ね合わせの観察における精度を向上させるために、印刷版原本の LAJ 参考図との照合精度を上げたベクトル形式データ（シェープファイル、AI ファイル）の参考図電子地図として（a）「参考図 V 近世道路図（明治 20 年前後）」、（b）「参考図 II 地勢図」、（c）「参考図 VI 近世藩領図（1664 年、寛文 4 年）」を作成した。

#### 歴史的地図の電子地図

LAJ の参考図 5 の「近世道路図（明治 20 年前後）」の基礎となっている「輯製 20 万分 1 図」を電子地図化（GeoTIFF）した。「近世道路図」は、LAJ の被調査者たちの育った時代、ないしは、養育者達の活躍した時代の陸上交通網の概要を示すもの（参考図 5、凡例中の注記）として作成された。この「近世道路図」が原図として参照しているものが、「輯製 20 万分 1 図」であり、明治になって日本全国を初めておおった地形図である。道路網を始めとし、江戸から明治にかけての地理的情報の電子化資料として「輯製 20 万分 1 図」（全 153 枚）の電子地図（GeoTIFF）を作成した。

人口データ：国勢調査による時系列の市区町村別人口データ（1980 年の市区町村の境域による 1920 年-1980 年の国勢調査時系列データ）

市区町村別の人口分布を分析に利用できるようにするためのデータの入力、整備を行った。人口データの時系列の比較ができるように、1980 年の市区町村の境域によって編集された統計局（1985 年）および東洋経済新報社編（1985）を典拠とした（両者は基本的に同じだが、一方にあって他方にはないものがあり、両者を典拠とした）。男女合計、男女別、年齢区分（3 区分）別の時系列の人口データを入力、整備し、基本的な人口データとともに、年齢と性別という基本的な属性も扱えるように整備した。（なお、国勢調査のデータとして公開されている機械可読な人口データは比較的近年のものであり、過去に遡ったものは見いだせず、また、時系列での比較が可能なデータでもあるので、上記の典拠を用いてデータの入力を行った。）

人口地図の電子地図：国勢調査による市区町村別人口データによる人口分布図

市区町村境界および市区町村役場の位置情報のデータを整備し、1980 年境域による市区町村別の国勢調査による人口分布の電子地図（シェープファイル形式）を作成した。

市区町村の境域による集計単位の編成基

準になっている 1980 年までの市区町村数の変遷（1920 年から 1980 年まで）を示すデータも入力した。

#### 国勢調査人口地図の電子地図

国勢調査人口地図（昭和 25 年、30 年、35 年、45 年）の画像ファイルを作成し、その中から、「昭和 25 年 人口分布（1:2000000）」、「昭和 30 年 市町村別人口分布人口密度（1:800000）」、「昭和 35 年 市区町村別人口分布図（1:1000000）」の電子地図（GeoTIFF 形式）を作成した。

#### （3）データの観察・分析、方法の開発

方言分布データ分析の基盤としての交通網や人口分布などの言語外的情報のデータ整備や公開データの情報収集と相互に関連しながら、これらのデータを取り込んだ LAJDB のデータの観察と分析を行い、方法的な探索を行った。LAJDB による全国 2400 地点の LAJ の方言分布データと整備導入した基盤的な言語外的データによる、データを中心とした、データ解析的、現象発見的なアプローチ（データを中心に据えたアプローチ）による大規模方言分布データを分析する方法を探索した。

LAJ の分布データ（LAJDB）を元に計算される標準語形の頻度分布、地点間の言語的な類似度の分布や類似度による相互関係のネットワークと上記の言語外的な情報やその他の関連データとの相互関係の観察・分析を行い、方法的な考察を行った。全国 2400 地点の詳細な分布データとこれに対応できる精度の言語外的情報（道路網、人口分布など）のデータ化により、県別データの観察では見えなかった現象の発見と詳細な観察、分析が可能となった。市区町村別時系列人口データ、人口地図、道路網等の電子地図、市区町村境域の変遷データなどの情報を用い、要因の時系列の変化の観点（人口、市区町村境域）や交通史等の歴史的観点も参照し、観察・検討を進めることができた。

#### 調査地点における標準語形の出現頻度の分布

地点別の共通語形の頻度分布（図 1）の観察では、言語外的な情報として、道路網、地誌的情報、人口分布、交通史などの情報を導入した事例観察を行い、地点別の頻度情報がこれらの言語外的情報と有意な関連を見ることができた。標準語形の分布と合わせて、語の併用回答の地理的分布、新語、標準語として意識された語形の分布などの観察も同様に有意なパターンが観察できた。

#### 方言分布に基づく言語的類似関係の地理的模様との関連を観察

分布パターンや相互関係を示すネットワーク表示による観察、分析を行った。

- ・調査地点間の言語的類似度(方言類似度)の分布の観察(図2)
- ・調査地点間の言語的類似度(方言類似度)で測った関係のネットワーク(図3, 4)

人口分布や交通網(図5)と、方言分布に基づく言語的類似関係の地理的模様との関連を観察し、言語接触や伝播の現象に関わる観察を行った。人口分布とネットワーク法(類似度の分布(図2)やNT-1(r)(図3)やドロネーネット表示(図4)によるネットワーク表示)による方言類似度表示を突き合わせ、これらと人口分布や交通網との関連に、明瞭で有意な相互関係を観察できた。

#### 人口分布, 市区町村境域の変遷との関係の観察

人口データとして入力した1920年から1980年の時系列データは1980年時点の市区町村の境域によって集計されたデータであるが、明治以降、市区町村の大規模な統廃合が行われてきたため、市区町村の境界の変遷も合わせた観察を進めた。LAJの方言分布データから計算される地点間の類似関係と交通網や人口分布データ等の言語外データとの間のより詳細な関係性を観察することができた。また、変遷や変動の中でも、安定的に観察される全体的パターンが重要な意味を持つ見通しが得られた。

#### 空間的な統計手法との比較

近年、地理的な分布を扱う空間統計の手法を利用した研究が行われている(例えば, Rumpf et al. 2010)が、空間統計的な手法による分析の特性を探るための事例的な探索として、LAJDBの方言分布データに対してカーネル密度推定を用いる手法を試行した。また、比較のために、クラスター分析などの多変量解析の適用も試行し、上記の観察との比較によって、方法ごとの特性の検討を行った。本研究のアプローチで、より詳細な現象観察の可能性が得られた。

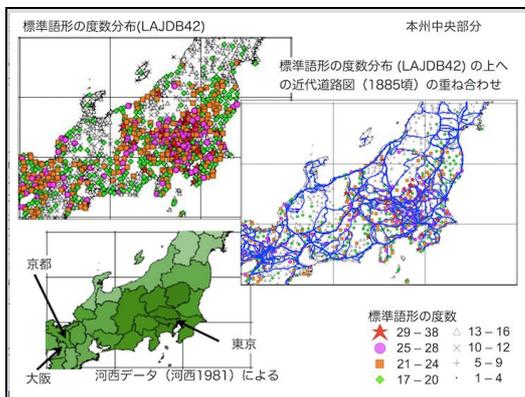


図1 解像度の違いの例(熊谷 2017:42)  
(上の左下が県別集計による本州中央部の表示であり、左上が地点単位によるもの)

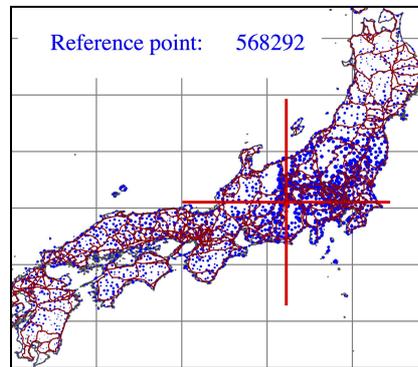


図2 言語的類似度(方言類似度)の分布と近代道路網(赤)との重ね合わせの例。十字は参照の中心点(熊谷 2017:44)

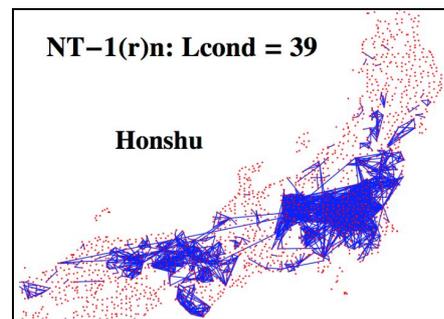


図3 地点間の言語的類似度(方言類似度)の関係のネットワーク NT-1(r)n の例。類似度 Lcond 以上が結ばれる。(Kumagai 2016:356)

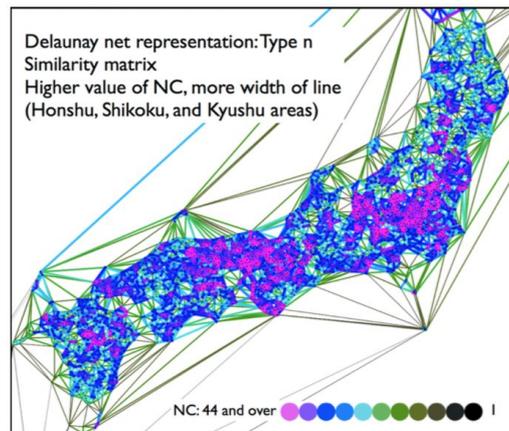


図4 調査地点間の言語的類似度(方言類似度)で測った関係のネットワーク(Delaunay net)の例(Kumagai 2016:354)

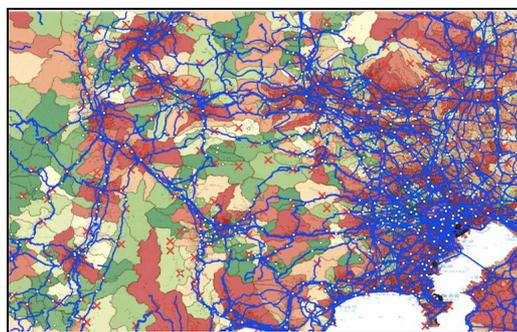


図5 人口分布図(1965, 人口: ドット), 市区町村役場(1985, 白丸), 調査地点(×印), 道路網(熊谷 2017:45)

LAJの「高解像度」の全国的な分布データを生かすことのできる詳細で全国的な言語外的データを用意し、両者を結合させた観察、分析により、以前の県別単位のデータの分析などの集計の括りを大きくした観察・分析では浮き上がってこない事象を捉えることができる。人口や交通網、交通史、地形図、地誌、歴史的データ等の方言分布の形成に関わる基盤的なデータを取り込んだ、統合的なデータの作成と観察を中心としたアプローチが有効であった。また、同時に、異なった出自のデータを結合、関連させて統合的に分析する上での課題、問題点も捉えられた。データの用意の仕方やその扱い方も含めた、さらなる検討、開発の必要性がある。

海外では言語地図の方言分布データを用いた計量的な研究が盛んであるが、LAJに比べた小規模なデータによるものであり、LAJのような全国規模で、地理的な背景も異なる広い地域にわたる、大規模で詳細な言語地図のデータによるものは見当たらない。この利点を生かし、今後、さらにデータの整備と分析を進めることにより、方言分布とその背後にある要因とのダイナミックな関係を、詳細なデータに基づいて、追求することができる。

#### <引用文献>

井上史雄(2001)『計量的方言区画論』東京：明治書院

熊谷康雄(2013a)『『日本言語地図』のデータベース化と計量的分析：併用現象、標準語形の分布と交通網、方言類似度の観察』『大規模方言データの多角的分析 成果報告書：言語地図と方言談話資料』pp.111-128.

熊谷康雄(2013b)『『日本言語地図』のデータベース化が開く新たな研究』、『国語研プロジェクトレビュー』4(1): pp.1-9.

国立国語研究所(1966)『日本言語地図』第1集解説：方法。東京：大蔵省印刷局

国立国語研究所(1966-1974)『日本言語地図』(全6巻)東京：大蔵省印刷局,(縮刷版1981-1985)

総理府統計局(1985)『昭和55年10月1日の境域による各回国勢調査時の市区町村別人口：大正9年～昭和55年』

東洋経済新報社編(1985)『国勢調査集大成 人口統計総覧』東京：東洋経済新報社。

陸地測量部・清水靖夫(1983)『幕末明治日本国勢地図 初版縮刷二十万分一図集成』東京：柏書房

Goebel, H. (2006) Recent Advances in Salzburg Dialectometry. *Literary and Linguistic Computing* 21(4): pp.411-435.

Rumpf J, S Pickl, S Elspasz, W Koenig, V Schmidt. (2010) Quantification and Statistical Analysis of Structural Similarities in Dialectological Area-Class Maps. *Dialectologia et Geolinguistica* 18: pp.73-97.

#### 5. 主な発表論文等

(研究代表者、研究分担者及び連携研究者には下線)

[雑誌論文](計2件)

熊谷康雄(2017)『日本言語地図』と『日本言語地図』データベース：データベース化(LAJDB)による多角的分析に向けて『方言の研究』3号: pp.29-51. 査読有  
Yasuo Kumagai (2016) Developing Linguistic Atlas of Japan Database and advancing analysis of geographical distributions of dialects. In: Marie-Hélène Côté, Remco Knooihuizen & John Nerbonne (eds.) The future of dialects: Selected papers from Methods in Dialectology XV. Berlin: Language Science Press: pp.333-362. Language Variation Vol.1. 査読有  
DOI:10.17169/langsci.b81.159

[学会発表](計2件)

Yasuo Kumagai. A quantitative observation of the relation among population distributions, road networks, and dialect similarities. Methods in Dialectology XVI. 2017.  
Yasuo Kumagai. Developing Linguistic Atlas of Japan Database and advancing analysis of geographical distributions of dialects. Methods in Dialectology XV. 2014.

#### 6. 研究組織

(1)研究代表者

熊谷康雄(KUMAGAI, Yasuo)

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：30215016