様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９　（共通）

# 科学研究費助成事業　　研究成果報告書

令和　６ 年　 ６ 月　 ５ 日現在

機関番号：１４３０１
研究種目：基盤研究(C)（一般）
研究期間：2021～2023
課題番号：２１Ｋ１２０２９

研究課題名（和文）文化進化の分析のための分岐と伝播の統合的モデル化

研究課題名（英文）Integrated Modeling of Branching and Horizontal Transfer for the Analysis of Cultural Evolution

研究代表者
村脇　有吾（Murawaki, Yugo）

京都大学・情報学研究科・准教授

研究者番号：７０６１６６０６

交付決定額（研究期間全体）：（直接経費）　　3,300,000円

研究成果の概要（和文）：人間諸集団が担う広義の文化、特に言語が歴史的にどのように変化してきたか（文化進化）を解明するための統計的モデルの開発に取り組んだ。文化進化のモデルには、(1) 分岐の繰り返しをともなう縦の継承と (2) 波状の横の伝播の大きく2つが知られているが、両者を統合的に扱う統計的モデルの開発が主要な研究目標として取り組んだ。統計的モデルは一応の完成を見たものの、実データから言語学的に意味のある結果を引き出すという点では課題が残った。一方、この開発の途上で、縦の継承のみを仮定する系統樹モデルにその仮定を満たさないデータを与えた場合に起きる問題を可視化する手法の解析を進め、大きな成果を得た。

研究成果の学術的意義や社会的意義
人間諸集団の歴史を解明するうえで、 それらの集団が話す言語は重要な手がかりとなる。本研究は言語進化に関する2つの従来モデルの統合に取り組むとともに、このモデルを適用するうえでの最大の問題は不確実性であると考え、統計的に取り組むことに焦点をおいた。この問題に伝統的に取り組んできた歴史言語学は言語同士のマクロな関係の解明については大きな成果を上げてきた一方で、例えば日本語内部の多様性（特に本土方言内部の関係）のようなミクロな言語関係については課題が残っていた。本研究の成果はこうした残された課題の解決につながる可能性がある。

研究成果の概要（英文）：I worked on developing statistical models to elucidate how culture in a broad sense, including language, has historically changed among human groups (cultural evolution). There are two main known models of cultural evolution: (1) vertical transmission characterized by repeated branching and (2) wave-like horizontal transmission. My primary research objective was to develop a statistical model that integrates both. While I achieved a preliminary completion of the statistical model, challenges remained in deriving linguistically meaningful results from real data. However, during the development process, I made significant progress in analyzing and visualizing problems that arise when a phylogenetic model is suppled with data that does not satisfy the assumption of vertial transmission.

研究分野：計算言語学

キーワード：文化進化　 分岐　水平伝播　ベイズ統計　言語

様　式　Ｃ－１９、Ｆ－１９－１（共通）

１．研究開始当初の背景

同じ日本語のなかでも京都方言、東京方言、鹿児島方言には様々な違いがある。それらの違いはどのような歴史的変化を経て生じたのだろうか。少し一般化すると、どのようにすれば方言群の歴史的変化を現代の観測データから復元できるだろうか。さらに一般化して、民話の類型や織物の文様といった人間諸集団が担う文化の諸要素も同じ手法で分析できないだろうか。本研究は、これらの難問に対して、少なくとも部分的な答えを見つけ出すことを目的としていた。

方言群のように、それを担う諸集団が密な接触を維持しているデータに対しては、長い間伝播に基づく説明が行われてきた。日本では柳田國男の『蝸牛考』(1930) やテレビ番組『探偵!ナイトスクープ』の 1990--91 年の企画「アホ・バカ分布図」が有名である。

こうした流れとは独立に、歴史比較言語学という分野が確立されており、インド・ヨーロッパ語族やオーストロネシア語族に代表されるように、言語同士のマクロな歴史的関係を木（系統樹）という形を採用することで明らかにしてきた。歴史比較言語学は系統分類に関して進化生物学と部分的に方法論を共有しており、この科学的に洗練された方法論は分岐学とよばれている。進化生物学側では計算集約的な統計的モデルの研究が先行して発展し、2000 年前後からはその成果の言語への転用が進んだ。

分岐学的手法の従来の適用範囲はマクロな言語関係であったが、近年はその対象を方言学が扱ってきたようなミクロな言語関係に拡大する動きが見られる。こうした研究はもっぱら言語学者による人手による論証によって進められており、一定程度の成果はあがっているもの、方法論的な限界に直面しているように見える。人手による論証は、確実性の高いと思われる証拠を積み上げることで行われる。しかし、方言データの大部分は接触による影響によって形作られていると想定され、分岐に基づくと思われるそのうちの一部を確信をもって抽出するのは本質的に困難である。

２．研究の目的

本研究の目的は、歴史比較言語学（分岐学）の方法論を基軸としつつ、接触の影響を取り込んだモデルを確立することである。上述のように、方言データは分岐と接触の両方の影響を受けて形作られており、後者の影響の方が圧倒的に大きいと想定される以上、両者を同時にモデル化することは不可欠である。そのためにはデータを特徴づける不確実性の高さを克服しなければならない。ここに計算集約的な統計的モデルの出番があると期待できる。

数理モデルという点では、従来の文化進化の分析は、ほとんどが分岐に基づく統計的系統樹モデルを用いて行われてきた。本研究のように、系統樹が未知という設定で両者を統合的に扱う研究は少なく、あったとしても分岐の影響が接触の影響を圧倒しているという仮定が置かれてきた。

３．研究の方法

（１）潜在的地理的分布に基づく木の導出

分岐と伝播の両者を統合的に扱う統計モデルの開発は挑戦的な課題である。いずれのモデルも 19 世紀半ばには提案されており、それぞれ言語学の下位分野で研究が進んできた。数理モデルの面では、進化生物学の分野で、分岐に基づく系統樹モデルを土台に計算集約的なベイズ統計のモデルが開発され、21 世紀に入ってから言語データに適用されるに至っている。しかし、分岐と伝播の両者を統合的に扱う必要性は認識されていながら、具体的な統計的モデルは未だ出現していない。正攻法は系統樹モデルに伝播を表す横の枝を追加することだが、あまりにも自由度が高いため、確からしい仮説を効果的に絞り込むことが現実的には不可能である。

本研究の提案の核心は、分岐と伝播がいずれも地理的分布を生み出すという共通性に着目することである。伝播が一定の地理的分布を示すのは自明だが、分岐も地理的分布と関連付けられる。系統樹中のある枝で起きた変化は、その子孫の言語群からなる地理的分布に関連付けられるからである。ある時点で起きた変化はその後起きた変化によって上書きされることがあり、元の分布が直接観測できるとは限らない。そこで、現代語に見られる諸形質の地理的分布の背後には潜在的地理的分布があり、それらが確率的に混合した結果を観測していると仮定する。そして、観測データから潜在的地理的分布を導出するための統計的モデルと推論手続きを開発した。

潜在的地理的分布というアイデアを拡張し、一部の分布同士に分岐学的制約を課すと木が得られる。つまり、一部の地理的分布群を木構造に基づいて組織したとき、(1) 中間ノードは 2 個の子を持ち、(2) 親の分布は 2 個の子に分割されるという条件を満たさなければならない。逆に言えば、現実の諸特徴がなす地理的分布の背後に潜在的な地理的分布を仮定したとき、それらが適切な包含関係にあるならば、分岐に基づく変化に由来するのではないかと期待するのである

統計的モデルは、実効性のある推論手続きがともなうことではじめて意味を持つ。従来の統計的系統樹モデルは、枝を系統樹内の別の場所にランダムに付け替えるという確率的操作を繰り返すことでより良い仮説を探索していた。しかし、提案モデルの場合は、親子が密に結合しており、枝の付け替えの計算コストが大きいだけでなく、局所解から抜け出す可能性も低い。

そこで、枝を付け替えることなく推論できるようなモデルの開発に取り組んだ。具体的には連続緩和とよばれる技法を用いて木を有向非巡回グラフに拡張しつつ、木構造から逸脱するほどペナルティがかかるようなモデルを開発した。このモデルでは、推論が進むにつれグラフの中から木が徐々に浮かび上がってくる（図 1）。

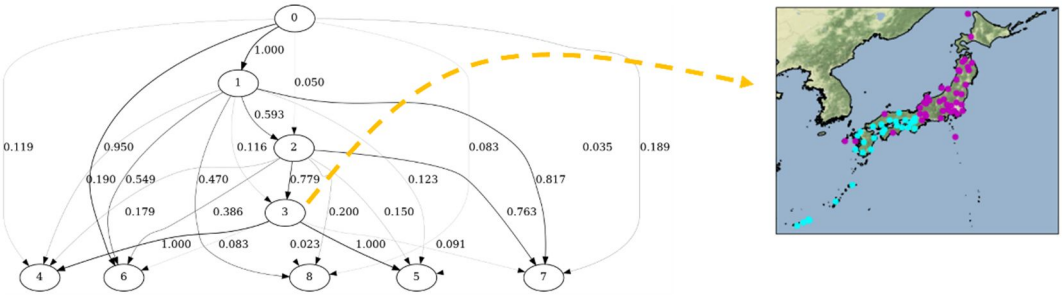


**図 1 日本語方言データへの提案モデルの適用例**

（2）系統樹モデルにおける異常検出

文化進化の研究においては（統計的）系統樹モデルが依然として支配的であるが、ここまでの議論で見たように、モデルの仮定に反する接触の影響が無視できない場合が少なくないと思われる。いくつかの従来研究はそのような場合でも系統樹モデルを適用していた。このような事例が不適切であることを効果的に検出することで、分岐と伝播の両者を統合的に扱う統計モデルの必要性をアピールできると考えた。

本研究では、ベイズ系統樹モデルが推論によって再構する系統樹のノード状態に主成分分析を適用し、系統樹を低次元空間に線形写像することで異常を可視化した。分岐に基づく変化だけが累積する場合は、先祖から子孫への遷移を考えたとき、子孫は先祖との類似性を（ほぼ）単調に下げるはずである。したがって、先祖から子孫への経路は（比較的解釈しやすい）第 1 主成分に関して（ほぼ）一方向に進むはずである。そうではなく大きな揺り戻しが発生している場合は、データがモデルの仮定に大きく違反していると判断できる。

４．研究成果

（1）潜在的地理的分布に基づく木の導出

潜在的地理的分布という新たな概念を普及さえるために、主に言語学者と一般読者を想定した分担執筆書籍でその概要を説明した。

木の導出に関しては統計モデルの実装にまでこぎ着けた。図 1 に示すように、実データを用いた検討も進めたが、グラフからは明確な形では木が浮かび上がってこなかった。おそらく「親の分布は 2 個の子に分割される」という条件が今回扱ったデータに対してはあまり適切でなかったのではないかと思われる。集団が 2 つに分裂するとき、典型的には一方は故地にとどまり、もう一方が新天地に移民するという形をとると思われるが、後者の側で相対的に速い変化が起きると想定される。要するに、2 個の子供のうち一方の存在はデータによって比較的支持されるが、それと相補分布をなすようなもう片方の存在はデータからあまり支持されないのではないかと考えられる。今後はこの点を引き続き検証したい。

（2）系統樹モデルにおける異常検出

実のところ、このアイデア自体は研究期間より前の 2015 年には思いついていたが、実証実験が十分に行えていなかった。今回、比較的新しい調査報告のデータも使った検証実験を行い、その有効性を実証した。この成果は査読付き国際会議で発表した。

提案手法のキモは、系統樹モデルが変化をマルコフ的に（記憶を持たずに）モデル化しているため、系統樹全体の大域構造を調べれば、モデルの仮定への違反が浮き彫りになるというアイデアである。系統樹モデルには様々な変種があるものの、基本的にはいずれについてもこのアイデアが適用できると予想していた。しかし、擬似ドロモデルとよばれる状態変化モデルは、先祖が失った特徴を子孫が再び獲得することがないという制約を課しており、このためにモデルの仮定への違反が浮き彫りになりにくいという知見を得た。ただし、今回検証に用いた調査報告は類型論のデータを対象としており、そもそも擬似ドロモデルは適切なモデル選択ではないと考えられる。このように、広い観点で見れば、この場合についても提案手法の適用には意義があったと言える。

５．主な発表論文等

〔雑誌論文〕　計1件（うち査読付論文　1件／うち国際共著　0件／うちオープンアクセス　1件）

| | |
|---|---|
| １．著者名<br>Yugo Murawaki | ４．巻<br>－ |
| ２．論文標題<br>Principal Component Analysis as a Sanity Check for Bayesian Phylolinguistic Reconstruction | ５．発行年<br>2024年 |
| ３．雑誌名<br>Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024) | ６．最初と最後の頁<br>12999－13013 |
| 掲載論文のDOI（デジタルオブジェクト識別子）<br>なし | 査読の有無<br>有 |
| オープンアクセス<br>オープンアクセスとしている（また、その予定である） | 国際共著<br>－ |

〔学会発表〕　計0件

〔図書〕　計1件

| | |
|---|---|
| １．著者名<br>林 由華、衣畑 智秀、木部 暢子 | ４．発行年<br>2021年 |
| ２．出版社<br>開拓社 | ５．総ページ数<br>316 |
| ３．書名<br>フィールドと文献からみる日琉諸語の系統と歴史 | |

〔産業財産権〕

〔その他〕

－

６．研究組織

| | 氏名<br>（ローマ字氏名）<br>（研究者番号） | 所属研究機関・部局・職<br>（機関番号） | 備考 |
|---|---|---|---|

７．科研費を使用して開催した国際研究集会

〔国際研究集会〕　計0件

８．本研究に関連して実施した国際共同研究の実施状況

| 共同研究相手国 | 相手方研究機関 |
|---|---|