方言群の時空間解析にむけて: フィジー語を例に

村脇 有吾
京都大学大学院情報学研究科
murawaki@i.kyoto-u.ac.jp

1 はじめに

方言学は、“every word has its own history” という成句に象徴されるように、諸特徴の個別的で複雑な地理的分布を重視し、方言 (言語) 全体の歴史の解明を二の次とするきらいがある [7]。この点で、言語間の比較を通じて系統関係の解明や祖語の再構を行う比較言語学との間で断絶が見られる。

こうしたなか、比較言語学の手法を方言群に適用し、その系統関係を解明しようとする試みがある [3, 8, 2]。そもそも系統関係は親から子への縦の継承に対応し、これを明らかにするためには、接触による横の拡散、普遍的変化、さらには偶然の一致を排除しなければならない。さらに、系統分類のための確実性の高い手がかりとして共通改新に着目し、共有残存を排除する。これ自体は比較言語学の確立された手続きだが、方言群は相互の近さゆえ、系統分類に用いられる特徴は結果的に人手で大きく絞り込まれることになる。しかし、排除された諸特徴も言語史を構成する一部には違いない。また、縦の継承と横の拡散の識別が本質的に難しいという点も未解決に見える。

統計的観点からは、系統樹モデルは、過去にさかのぼるほどに増す不確実性を木構造制約により抑え込んでいる。この制約を取り除き、現代という時間の 1 断面のみから過去を復元することは依然として難しい。

本研究では、時間的解析を視野に入れつつ、まずは方言群の大域的空間構造を捉えるための統計的モデルを提案する。最終的目標は方言群を解析するための一般的な手続きの確立だが、本研究ではフィジー語を題材として予備調査の結果を報告する。

2 フィジー語諸方言基礎語彙

フィジー語には明瞭な東西対立が見られるものの、等語線は交錯し、方言連続体を成している。また、共有改新 (と推定されるもの) が系統関係を反映するとは限らないと考えられている [1]。

発表者が分担者をつとめる科学研究費補助金・国際共同研究強化 (B) 「時空間を融合する: GIS と数理モデルを用いた新たな言語変化へのアプローチ」(研究代表: 菊澤律子) では、フィジー語諸方言の基礎語彙データの電子化を進めるとともに、GIS データを活用した分析を計画している。基礎語彙データは共同研究者の Paul A. Geraghty が長年にわたって収集したもので、各 communalect (以下では単に言語とよぶ) において各概念を表す語形が収録されている。

本研究ではこれらの語形に対して前処理を行った。具体的には、LingPy[4] を用い、語形の文字列比較に基づいて同源語を自動認定し、各概念について、同源語グループに番号を割り振った。その結果、97 個の概念が複数の同源グループから構成され、したがって各言語は 97 個の要素からなるベクトルで表現された。なお、提案手法が要求するのは各特徴の離散コード化だけであり、音韻的・統語的特徴も扱える。最後に座標データとの対応が現時点ではとれない言語を取り除き、106 個の言語を以降の分析に用いた。

3 事前分析

以下の手法でデータの事前分析を行った。(1) ボロノイ図を用いた等語線束の可視化、(2) 主成分分析 [7]、(3) admixture 分析 [9]。等語線束は多くの方言境界を浮かび上がらせるが、東西対立は埋もれてしまった。主成分分析と admixture 分析では東西対立が確認できたものの、その他の大域的構造を捉えがたい。

4 提案手法

提案手法は発表者が以前提案した潜在表現学習に若干の変更を加えたモデルである [5]。以前の対象は全世界の言語、特徴は言語類型論の特徴、目的は特徴間の依存関係を捉えることであり、本研究の対象とはまったく異なる。それにも関わらず、両者がほぼ同じモデルで扱えることの発見が本研究の主要な貢献である。

提案手法は複雑な確率的生成モデルであり、詳細は省略するが、概略は以下の通りである。$N = 97$ 個の表層特徴を K 個の潜在変数によって再構成する (K は事前に与える)。表層特徴、潜在変数のいずれも $L = 106$ 個の言語からなる地理的分布を持つが、この再構成は (1) 複数の特徴に共通する地理的分布をクラスタリング、(2) 各特徴の地理的分布を一般に 1 つ以上の分布に分解という 2 つの操作を同時に行う。潜在的な地理的分布を推定する際には座標データを利用することで、ある言語は近隣の言語と同じ値を取りやすいという事前知識を組み込む [6]。

予備実験の結果は、東西対立だけでなくその他の大域的構造をモデルが捉えられている可能性を示唆する。その詳細な言語学的分析は今後の課題である。

謝辞 本研究は一部 JSPS 科研費 18K18104、18KK0012 の助成を受けた。

参考文献

[1] Paul A. Geraghty. *The History of the Fijian Languages*. University of Hawai'i Press, 1983.

[2] 五十嵐陽介. 琉球語を排除した「日本語派」なる系統群は果たして成立するのか?–「九州・琉球語派」と「中央日本語派」の提唱. 国際日本文化研究センター共同研究会「日本語の起源はどのように論じられてきたか–日本言語学史の光と影」第 3 回共同研究会, 2016.

[3] ローレンスウェイン. 沖縄方言群の下位区分について. 沖縄文化, Vol. 40, No. 2, pp. 101–118, 2006.

[4] Johann-Mattis List, Mary Walworth, Simon J. Greenhill, Tiago Tresoldi, and Robert Forkel. Sequence Comparison in Computational Historical Linguistics. *Journal of Language Evolution*, Vol. 3, No. 2, pp. 130–144, 2018.

[5] Yugo Murawaki. Bayesian learning of latent representations of language structures. *Computational Linguistics*, 2019. (to appear).

[6] Yugo Murawaki and Kenji Yamauchi. A statistical model for the joint inference of vertical stability and horizontal diffusibility of typological features. *Journal of Language Evolution*, Vol. 3, No. 1, pp. 13–25, 2018.

[7] John Nerbonne and Martijn Wieling. Statistics for aggregate variationist analyses. In Charles Boberg, John Nerbonne, and Dominic Watt, editors, *The Handbook of Dialectology*, pp. 400–414. John Wiley & Sons, 2018.

[8] Thomas Pellard. Ōgami—Éléments de description d'un parler du Sud des Ryukyu. PhD thesis, Ecole des Hautes Etudes en Sciences Sociales (EHESS), 2009. (in French).

[9] Ger Reesink, Ruth Singer, and Michael Dunn. Explaining the linguistic diversity of Sahul using population models. *PLoS Biology*, Vol. 7, No. 11, 2009.