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

平成　３０　年　　６　月　２２　日現在

機関番号：　１４３０１
研究種目：　若手研究(B)
研究期間：　2014～2017
課題番号：　２６７３０１２２

研究課題名（和文）統計的手法による日本語諸方言の系統樹推定

研究課題名（英文）Statistical phylogenetic inference of the Japonic language family

研究代表者
村脇　有吾（Murawaki, Yugo）

京都大学・情報学研究科・助教

研究者番号：７０６１６６０６

交付決定額（研究期間全体）：（直接経費）　　2,800,000円

研究成果の概要（和文）：諸言語の系統的な関係を解明するための計算集約的な統計手法を開発した。この問題は長年言語学者が人手によって取り組んできたが、過去に復元する問題は本質的に不確実であり、統計的推論が適していると考えている。成果は多岐に及ぶが、特に言語類型論の特徴列を潜在空間に写像するベイズ統計の手法は、複数の特徴が連鎖的に変化し得るという類型論的特徴の特性を捉えることを可能にしたという点で重要である。

研究成果の概要（英文）：We developed compute-intensive statistical methods for uncovering phylogenetic relationships between languages. While this problem has long been tackled by linguists through manual labor, we argue that the inherent uncertainty can best be handled by statistical inference. Among key achievements is a Bayesian induction of latent variables from a sequence of features of linguistic typology because it has the potential of tracing holistic structural changes of language involving multiple features.

研究分野：　計算言語学

キーワード：　ベイズ統計学　言語系統論　言語類型論　基礎語彙

１．研究開始当初の背景

諸言語が数千年のオーダでどのように変化してきたかを研究する分野は歴史言語学あるいは歴史比較言語学とよばれ、いわゆる文系の言語学者が長年取り組んできた。しかし、Gray & Atkinson (2003)以降、計算集約的な統計モデルを適用する事例が増えていた。こうした統計モデルはもともと遺伝子などの生物データを分析するために開発されたもので、研究代表者の専門である自然言語処理分野ではほとんど認知されていなかった。

研究代表者は生物学由来の研究にいち早く気づき、さらに、そうした研究には限界があることを認識した。従来の統計的手法は広い意味での語彙的手がかり (基礎語彙の同源語) を用いるという点で伝統的な歴史比較言語学と共通する。これは、歴史比較言語学では100年以上の取り組みにもかかわらず未解決となっている日本語の系統の問題には適用できないことを意味する。また、生物に対応するものがない言語の特性はモデル化されない傾向があることに気づいた。

２．研究の目的

本研究課題の申請時においては、以下の 2 つの目的を考えていた。

(1) 日本語諸方言の系統: 日本語と他の言語との関係だけでなく、日本語内部の系統についても未解決であることに研究代表者は着目した。この問題に対しては、Lee & Hasegawa (2011)が先行して取り組んでいたが、語彙を用いた彼らのモデルが問題を抱えているのは研究代表者の目には明らかであった。したがって、この先行研究の問題点を立証するとともに、これに代わる手法を開発することを第 1 の目的とした。

(2) ベイズ統計に基づく手法の開発: Gray & Atkinson (2003)やそれ以降の研究ではベイズ統計に基づく確率的系統樹モデルを言語データに適用しており、系統樹のように複雑な潜在構造を持つ問題に対するベイズ統計の有効性を実証してきた。系統樹モデルとはまったく独立に、研究代表者の出身分野である自然言語処理においても、文書のトピックモデルや教師なし単語分割といったタスクにおいてベイズ統計が威力を発揮することが示されてきた。こうした成果を応用することで、生物学由来の手法を補完するモデルを開発することを第 2 の目的とした。

３．研究の方法

上述の目的と対応する形で、主に 2 つの観点から研究に取り組むことを当初は考えていた。

(1) 水平伝播のモデル化: 系統樹モデルは親から子へと縦に特徴が継承されるモデルであり、言語同士の横の接触によって特徴が変化するという水平伝播の影響を無視している。Lee and Hasegawa (2011) の問題を解明する鍵として、また、より一般に系統樹モデルを補完するために、水平伝播のモデルを開発することを考えた。

(2) 手がかりとしてのアクセント体系の利用: 京都と東京の方言について、「水」や「食べる」のような基礎語彙から違いを探すのは難しいが、アクセントが異なることは誰が聞いても明らかである。例えば、「橋」のアクセントは京都では HL (高低)、東京では LH (低高) と異なる。一方、「箸」はそれぞれ LH、HL であり、ちょうど逆転している。アクセントを手がかりとした史的変化の研究は、金田一をはじめとする言語学者の研究があり、そうした基盤のうえに統計的手法を開発することで、定量的で客観的な議論を実現することを考えた。従来研究の成果としてアクセントデータが紙媒体では公開されていたが、これを計算機可読な形で電子化する必要があった。また、アクセントの音声的な実現の背後には音韻的な表現があり、近年の研究はその重要性を示唆している。このような複雑な現象を統計的に扱うために一次近似を考案する必要があった。

４．研究成果

基礎語彙の同源語に代わる手がかりとしてアクセント体系を用いるという方向性ではめだった成果は得られなかった。しかし、別の手がかりとして言語類型論の構造的特徴を用いる手法が予想外の成功をおさめた。

(1) 水平伝播のモデル化（雑誌論文⑦、学会発表⑪、⑭）：そもそも系統樹モデルが成功をおさめてきた理由の一つは、モデルの自由度を抑えて推論を実現することである。時間をさかのぼるとともに不確実性が高まるが、同時に諸言語を共通祖先に合流させることで、推定すべきパラメータも減らしている。これに対し、水平伝播はモデルの自由度が高く、与えられた現代の語彙データに対応する自然な過去のシナリオは、系統樹の場合以上に絞り込まれていない。したがって、現在のデータが生成された過程を復元するという逆問題を解くのは難しい。そこでシミュレーションを用いた探索的解法を提案した。まず、水平伝播を通じて諸言語の特徴が時間とともに変化するようにシミュレーションモデルを設計する。次に、このモデルにデータを生成させ、それを系統樹モデルに与える。このとき、系統

樹モデルがどのように騙されるかを探索的に調べる。Lee and Hasegawa (2011)に見られる不自然な部分木の一部は、本研究で検討したシナリオにより説明できる可能性を示唆した。

(2) 垂直継承と水平伝播の同時モデル化（雑誌論文①、②、⑤、学会発表②）：上述の通り、水平伝播のモデル化は難しいが、系統樹を推定するのではなく、既に与えられているという設定においては、垂直継承と水平伝播を同時にモデル化できることを示した。具体的には、垂直継承と水平伝播それぞれの相対的な強さを垂直安定性と水平伝播性という 2 つのスカラー値で表し、それらをデータから推定した。提案モデルは自己ロジスティックモデルとよばれる一群のモデルの拡張であり、系統的な、あるいは地理的な言語間の関係を隣接グラフにより近似的に表現する。雑誌論文⑤では尤度最大化による推論を行ったが、雑誌論文①ではモデルをベイズ化し、より頑健な推論を実現した。さらに雑誌論文②では、自己ロジスティックモデルを後述の潜在表現の導出に応用した。

(3) 混合モデルの利用（雑誌論文⑥、学会発表⑧）：自然言語処理における文書のトピックモデルは、観測された文書が複数のトピックの確率的混合であると仮定する。これと同様に、クレオール言語は語彙提供言語、基層言語、言語普遍な再編器の混合であることが示唆されており、同様の混合モデルが適用できることを示した。丁度 APiCS (Atlas of Pidgin and Creole Language Structures)というクレオール言語を含む諸言語の類型論的特徴のデータベースが 2013 年に公開され、これが WALS (World Atlas of Language Structures)という大規模な類型論データベースと比較可能な形で設計されていたことから、定量的な実験が可能となった。

(4) 構造的(類型論的)特徴の潜在表現の導出（雑誌論文②、⑧、学会発表②、③、⑤、⑥、⑫）：言語間の系統的関係を解明するための手がかりとして従来用いられてきたのは広い意味での語彙的手がかりであった。一方、言語類型論という言語学の一分野があり、言語間の構造的特徴を比較してきた。Nichols (1992)のように構造的特徴から言語間の深い関係を探る試みもあったが、主流とはほど遠い状態にあった。その大きな理由は、語彙的手がかりは人手で扱いやすいのに対し、構造的特徴は不確実性が高く、人間による推論が難しいことであると研究代表者は考えた。そして、計算集約的な統計手法であればこの問題を克服できる可能性があるという見込みのもと研究を進めたところ、当初の想定以上の成果を得た。

構造的特徴の重要な特性として、特徴間に依存関係が見られることが挙げられる。語彙的特徴は独立性(「水」を表す語は「食べる」を表す語と独立に変化する）を仮定しても大きな問題とならない。一方の構造的特徴は、例えば動詞が目的語に先行する（VO 語順)なら、名詞を修飾する関係節が後置される（NRel 語順)を傾向があることが知られている。この関係を VO→NRel と表記すると、RelN→OV もよく成り立つことが知られている。この 2 つから、VO 語順かつ RelN 語順という組み合わせが珍しく、VO 語順かつ NRel 語順から OV 語順かつ RelN 語順に変化するとき、VO 語順かつ RelN 語順を経由する可能性が低いことを示唆する。しかし、従来の系統樹モデルは特徴間の独立性を仮定するため、誤って不自然な祖語を再構するおそれがあった。

そこで、表層特徴列からなる各言語を潜在変数列に写像するという解法を提案した。表層表現と潜在表現をつなぐ重み行列が表層表現における依存関係を捉えるようにする。そうすると、潜在表現を変化させると、一般に複数の表層特徴が変化する。したがって、潜在空間において系統推論を行えば、表層表現における依存関係が暗黙のうちに扱える。

なお、潜在変数をパラメータとよぶことにした。ベイズ統計におけるインディアンビュッフェ過程では潜在変数は「特徴」（feature）とよばれるが、この用語は言語類型論では観測変数を指すのに用いられている。そこで、生成文法から「パラメータ」という用語を借りることにした。

雑誌論文⑧ではニューラルネットの自己符号化器を用いたモデルを提案したが、その後の研究により、欠損値が大半を占める WALS のデータに対して頑健でないことが判明した。文献②ではベイズ統計のモデルを提案し、頑健性を向上させた。さらに、特徴間の依存関係に加えて、言語間の依存関係を捉えるために自己ロジスティックモデルをモデルに組み込んだ。

欠損値を復元させる実験において提案手法はベースライン手法を大きく上回る精度を達成した。また、垂直安定性と水平伝播性の観点から表層特徴列と潜在パラメータ列を比較すると、両者が全体として同等の安定性・伝播性を持つことが示された。これらの結果は、導出された潜在表現が表層に見られるパターンを適切に捉えていることを示唆する。

導出された潜在表現を系統樹モデルに適用することで、複数の特徴が連動する変化が定量的に捉えられることが期待される。この点に関しては、本研究課題の期間内においては予備的な調査しかできなかった。引き続き研究を進めたい。

(5) 学際的研究の促進（雑誌論文③、④、学会発表④、⑨、⑬）：通常の意味での研究成果とは異なるが、学際的研究の促進に結果として尽力することになった。
上述のように、従来いわゆる文系の言語学者が取り組んできた問題に対し、生物学由来の統計的手法が導入されてきたという経緯があり、研究代表者の出身分野である自然言語処理ではこうした研究はほとんど認知されていない。また、高度な統計手法はほとんどの言語学者にとって手に負えない代物である。
このような背景があるなか、自然言語処理（学会発表⑨、⑬）や人工知能（雑誌論文④）、統計（雑誌論文③）に加え、歴史言語学(学会発表④)の分野において、本研究課題やその背景にある研究動向を紹介する機会に恵まれた。今後は、これらの機会を通じて得たつながりを基盤として、学際的な共同研究を進めたい。

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 8 件）

①Yugo Murawaki、Kenji Yamauchi、
A Statistical Model for the Joint Inference of Vertical Stability and Horizontal Diffusibility of Typological Features、Journal of Language Evolution、査読有、 Vol. 3、No. 1、2018、pp. 13-25
DOI: 10.1093/jole/lzx022

② Yugo Murawaki 、 Diachrony-aware Induction of Binary Latent Representations from Typological Features、Proceedings of the 8th International Joint Conference on Natural Language Processing、査読有、2017、pp. 451-461
http://aclweb.org/anthology/I17-1046
http://anthology.aclweb.org/attachments/I/I17/I17-1046.Notes.pdf

③村脇 有吾、言語変化と系統への統計的アプローチ、統計数理，査読有(招待)、Vol. 64、No. 2、2016、pp. 161-178
http://www.ism.ac.jp/editsec/toukei/abstract/64-2j.html#161

④村脇 有吾、言語系統解明のための計算的取り組み、人工知能、査読無(招待)、Vol. 31、No. 6、2016、pp. 780-786
http://id.nii.ac.jp/1004/00008650/

⑤ Kenji Yamauchi 、 Yugo Murawaki 、 Contrasting Vertical and Horizontal Transmission of Typological Features 、 Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers、査読有、2016、pp. 836-846
http://aclweb.org/anthology/C/C16/C16-1080.pdf

⑥Yugo Murawaki、Statistical Modeling of Creole Genesis、Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies、査読有、2016、pp. 1329-1339
DOI: 10.18653/v1/N16-1158

⑦ Yugo Murawaki 、 Spatial Structure of Evolutionary Models of Dialects in Contact、PLOS ONE、査読有、Vol. 10、No. 7、2015、e0134335
DOI: 10.1371/journal.pone.0134335

⑧ Yugo Murawaki 、 Continuous Space Representations of Linguistic Typology and their Application to Phylogenetic Inference 、 Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies、査読有、2015、pp. 324-334
DOI: 10.3115/v1/N15-1036

〔学会発表〕（計 14 件）

①時武 孝介、村脇 有吾 他、ガウス埋め込みに基づく単語の意味の史的変化分析、言語処理学会 第 24 回年次大会、2018

②村脇 有吾、特徴間の依存関係を考慮した基本語順の史的変化の分析、言語処理学会第 24 回年次大会、2018

③村脇 有吾、潜在表現に基づく言語構造の史的変化の分析、機構間連携・文理融合プロジェクト「言語における系統・変異・多様性とその数理」シンポジウム、2018

④村脇 有吾、言語系統論への計算的アプローチの可能性、日本歴史言語学会 2017 年大会 公開シンポジウム 言語系統論の過去(これまで) と未来 (これから)、2017

⑤村脇 有吾、言語類型論の特徴からの潜在表現の獲得とその歴史的変化の分析への応用、機構間連携・文理融合プロジェクト「言語における系統・変異・多様性とその数理」研究発表会、2017

⑥村脇 有吾、言語の構造的特徴はなぜ、どのように変化するのか、NLP 若手の会 (YANS) 第 12 回シンポジウム、2017

⑦村脇 有吾、言語類型論的特徴からの潜在的 2 値パラメータの獲得、言語処理学会 第

23 回年次大会、2017

⑧村脇 有吾、クレオール形成に対する混合モデル、言語処理学会 第 22 回年次大会、2016

⑨村脇 有吾、言語進化史の統計的研究. 言語処理学会第 22 回年次大会 チュートリアル、2016

⑩村脇 有吾、言語変化と系統への統計的アプローチ、国立国語研究所・統計数理研究所合同研究集会「統計的言語研究の現在」、2015

⑪村脇 有吾、語彙拡散の空間構造モデル、統計数理研究所共同研究集会「社会物理学の現代的課題」、2015

⑫村脇 有吾、言語類型の連続空間表現とその系統推定への応用、言語処理学会 第 21 回年次大会、2015

⑬村脇 有吾、諸言語の歴史的変化に対する数理的取り組み、情報処理学会 第 220 回自然言語処理研究会 招待講演、2015

⑭村脇 有吾、方言群の語彙は系統樹をなすか、NLP 若手の会 (YANS) 第 9 回シンポジウム、2014

〔図書〕（計 0 件）

〔産業財産権〕

○出願状況（計 0 件）

○取得状況（計 0 件）

〔その他〕
ホームページ等
語彙の空間伝播のシミュレーションのソースコード
https://github.com/murawaki/lexwave
クレオール形成の混合モデルのソースコード
https://github.com/murawaki/creole-mixture
言語の特徴の垂直安定性と水平伝播性を推定する自己ロジスティックモデルのソースコード
https://github.com/murawaki/bayes-autologistic
言語の特徴列の潜在表現を導出するモデルのソースコード
https://github.com/murawaki/latent-typology

６．研究組織
(1)研究代表者
村脇 有吾 （MURAWAKI, Yugo）
京都大学・大学院情報学研究科・助教
研究者番号：７０６１６６０６