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　３ 年　 ６ 月　 ２ 日現在

機関番号：１４３０１
研究種目：若手研究
研究期間：2018～2020
課題番号：１８Ｋ１８１０４

研究課題名（和文）計算的取り組みによる言語の歴史的変化の解明

研究課題名（英文）Computational approaches to understanding historical changes of languages

研究代表者
村脇　有吾（Murawaki, Yugo）

京都大学・情報学研究科・講師

研究者番号：７０６１６６０６

交付決定額（研究期間全体）：（直接経費）　　3,200,000円

研究成果の概要（和文）：言語群を比較することで、それらの言語がどのように変化してきたかを解明するための計算集約的統計的手法の開発に取り組んだ。当初の対象は世界中の諸言語であり、時間規模は数百年、数千年、比較のための手がかりは主語・目的語・動詞(SOV) の基本語順や、声調の有無といった構造的特徴であった。その後、同じ手法がわずかな修正をともなうだけで、語彙的特徴を手がかとりした方言群の比較に適用できることを発見した。

研究成果の学術的意義や社会的意義
言語の歴史的変化を推定するには祖先を共有する他の言語との比較が強力な手段であるが、日本語の場合は近縁関係が明らかな他の言語が存在せず、この手法が適用できなかった。本研究では構造的特徴がどのように変化するかを確率的にモデル化した。この手法は、日本語がどのように変化してきたのか、また今後どのように変化していくのかを推定するのに役立つ。さらに、ほぼ同じ手法が方言群に適用できることを発見した。日本語の歴史を考える上では、上記の理由から、日本語内部の変異を明らかにすることが残された有効な手段の1つである。今後はこの点を追究していきたい。

研究成果の概要（英文）：We have developed computer-intensive statistical methods that, by comparing a group of languages, uncover how they have changed over time. Our initial target was languages around the world, with the time scale of hundreds and thousands of years and with structural features as comparanda. Structural features include the basic word order of subject, object and verb (SOV) and the presence and absence of tone. We later found that, with small modifications, the same model can be applied to a group of dialects, with lexical features as comparanda.

研究分野：計算言語学

キーワード：ベイズ統計

様　式　Ｃ－１９、Ｆ－１９－１、Ｚ－１９（共通）

１．研究開始当初の背景

我々が話す言語はどのように変化してきたのだろうか。そもそもどうして言語は時間とともに変化するのだろうか。これらの疑問、特に前者に取り組んできた分野として、文献学や歴史比較言語学が挙げられる。文献学の射程は当然ながら文字記録がある言語・年代に限られる。一方、歴史比較言語学は複数の言語を比較し、それらの共通祖先を再構することで、19 世紀以来大きな成功をおさめてきた。しかし、この方法が有効なのは祖先を共有する諸言語が比較できる場合のみである。日本語は系統不明であり、その歴史を解明する手段としては比較手法の有効性は限定的である。

研究代表者は、若手研究(B)「統計的手法による日本語諸方言の系統樹推定」(2014-2017 年度)に取り組むなかで、言語の構造的特徴が特に有望な分析対象であることに気づいた。構造的特徴とは、主語・目的語・動詞 (SOV) の基本語順や、声調の有無といった特徴であり、言語類型論とよばれる言語学の一分野で研究が進められてきた。歴史言語学において主な比較対象であった語彙とは異なり、構造的特徴は、その性質上、任意の言語対を比較可能とする。したがって、構造的特徴の歴史的変化に関する性質を明らかにすれば、日本語の歴史についても何らかの示唆が得られる可能性がある。

研究代表者は、上記研究を進める際、人手による論証に頼ってきたこれまでの主流研究とは異なり、計算集約的な統計モデルを採用した。人手による論証において大きな障害となっていたのは、現在得られる手がかりからでは過去の状態を確実には復元できないという不確実性の問題である。一方、特にベイズ統計に基づく手法は、個別には不確実な証拠を積み上げて全体として確からしい仮説を探索するのに向いている。構造的特徴の振る舞いは語彙的特徴と比べて不確実性が高く、この取り組みにふさわしい分析対象である。

２．研究の目的

上記の研究背景に基づき、研究代表者のこれまでの研究を発展させることで、世界の諸言語が、数百年、数千年の規模でどのように変化してきたかを部分的に解明することが研究目的である。研究代表者は、この目的に対して、ベイズ統計に基づく計算集約的手法が有効であると考えており、具体的なモデルの開発が中心的な位置を占める。

３．研究の方法

(1) データ　　計算集約的手法を使うには、まず計算機可読な言語データが必要となる。言語データの作成そのものは本研究の範囲外であり、主に既存の大規模データベースである WALS (World Atlas of Language Structures, http://wals.info) を利用した。WALS は世界の様々な言語の構造的特徴を言語学者がコード化したもので、行を言語、列を構造的特徴とする行列とみなせる。言語数は約 2,700、特徴数は約 100（いずれも前処理後）と大規模である。ただし、行列の要素の 70%以上が欠損値であり、その扱いは難しい。この行列に加えて、各言語の地理的位置やその他のメタ情報が付与されている。WALS の他に、Glottolog (https://glottolog.org/) という言語カタログも利用した。Glottolog が各言語に与える ID (glottocode) は WALS 収録の各言語にも付与されており、2 つのデータベースの相互接続が容易である。Glottolog は、歴史比較言語学の成果に基づく（つまり主に人手による語彙の比較により得られた）系統樹を収録しており、本研究ではこれも活用した。

さらに、本研究の期間中に、WALS と似たデータベースとして、スイス・チューリヒ大学の Balthasar Bickel らが開発している AUTOTYP が公開されていることに気づいた。そこで、WALS と AUTOTYP の 2 つを用いて統計モデルの検証を行った。AUTOTYP も欠損値が大半を占めるデータベースであり、統計的には扱いづらいデータである。ただ、世界中の言語を同じ基準で比較することの困難さを考慮すれば、欠損値の多さはやむを得ないものであり、統計モデルの工夫で凌ぐしかない。

(2) 系統樹に基づく時間変化のモデル化　　WALS も AUTOTYP も基本的には現代語のデータベースである。そして世界の言語の圧倒的多数は文字記録を持たない。では、どうすれば時間変化について推論できるのだろうか？その鍵は系統樹である。系統樹は近縁関係が語彙的証拠により明らかになっている言語群の歴史的関係を木で表現したものであり、根を含む内部ノードは共通祖先を表す。先祖を共有する 2 つの言語がある特徴について別の値を持つならば、系統樹上のどこかで少なくとも 1 回は変化が起きたはずである。1 回だけで 2 回や 3 回ではないのかや、その変化がどの時点で起きたかなどを特定するのは難しいため、人手による論証は困難である。一方、計算集約的な統計モデルは、計算資源に物を言わせて、近似的であって網羅的ではないものの、様々な可能性を検討する能力を備えている。

変化をモデル化するには言語の状態についてのモデルも必要である。各言語は構造的特徴の列（順番には意味はないが便宜的に並べたもの）で表現できる。各特徴はカテゴリカル変数とみなせる。そうすると、特徴の異なり数だけの状態を持つマルコフ連鎖モデルを導入するには自

然な発想である。各言語はある状態から別の状態に時間とともに遷移していく。通常のマルコフ連鎖モデルは離散時間を考えるが、言語変化に対して離散時間を設定するのは自然ではないので、連続時間マルコフ連鎖モデルを用いる。

連続時間を導入すると、系統樹のノードに対しても時刻（年代）を対応付ける必要が生じる。しかし、言語学者が人手で作成した系統樹は時刻を伴わない。インド・ヨーロッパ語族やオーストロネシア語族などに対しては、21 世紀に入ってから、語彙的特徴を用いた系統樹全体の年代推定が行われるようになっているが、そうした語族はごく一部の例外に過ぎない。そこで、文献的証拠や考古学的証拠を手がかりに、一部のノードのおおよその時刻を事前分布という形で与えることにし、それらノードの具体的な時刻や残りのノードの時刻は統計的に推定することにした。

(3) 言語の潜在表現の導出　　実は、一部の従来研究においても、語彙的手がかりを元に作った時刻付き系統樹に沿った構造的特徴の時間変化の推定は行われてきた。本研究の核心は、それらの研究とは異なり、言語を潜在空間に写像した上で、その空間上で時間変化を推定することである。

潜在空間を用いる動機は、構造的特徴間に依存関係があるという言語類型論では古くから知られている事実である。言語類型論研究者の Joseph Greenberg は、含意的普遍性とよばれる関係を幾つか提示している。例えば、名詞・数詞(NQ) 語順をとる言語は名詞・形容詞 (NA) 語順をとりやすい。しかし、上述の素朴な時間変化のモデルは特徴間の独立性を仮定する。したがって、名詞・数詞 (NQ)、形容詞・名詞 (AN) という不自然な組み合わせを持つ祖先を推定するおそれがある。Greenberg は 2 つの特徴間に通言語的に良く成り立つ関係を調べたが、より多くの特徴同士が有機的に依存関係を持つ可能性が高い。それに応じて、歴史的にも、多くの特徴が連鎖的に変化する可能性を考慮すべきであろう。

本研究で提案した潜在表現は、特徴間の依存関係に対処するためのものである。言語の表層特徴列の背後には独立成分 (パラメータとよぶ) の列があり、パラメータからの確率的生成により表層特徴列を観測していると仮定した上で潜在表現を推定する。このような表現学習は深層学習の研究動向を知っていれば自然な発想であるが、歴史言語学への応用は新しい。

ここまでのアイデアは実のところ上記の若手研究(B)の期間中に思いついていたことであるが、このアイデアを頑健な統計モデルとして結実させるのは予想以上に困難であった。その難しさは、上述の通り、言語データベースの大半が欠損値であることに起因する。表現学習にはニューラルネットを使うのが自然な選択であるが、表現力の高すぎるニューラルネットでは欠損値への対応が困難であった。そこで、表現力を制限しつつ、追加の手がかりをモデルに組み込んだ。すなわち、系統的に近い言語や地理的に近くに位置する言語は潜在空間上でも同じパラメータの値を取りやすいという仮定である。その結果として大掛かりなベイズモデルができあがった。モデルに対応する推論手続きの開発も挑戦的な課題であったが、二重の近似を盛り込むことで何とか実現した。

追加の仮定を盛り込んだ副産物として、潜在表現は 2 値表現となった。つまり離散状態であり、上述の連続時間マルコフ連鎖モデルを用いた時間変化のモデル化がそのまま流用できるようになった。すなわち、(1) 世界中の言語を一旦潜在空間に写像し、(2) 潜在空間上で各パラメータの時間変化のモデルを推定し、(3) 必要に応じて元の表層特徴に戻して変化を分析した。

４．研究成果

(1) 潜在表現の評価　　上述の通り、構造的特徴の列で表現される各言語の背後に潜在パラメータ列を仮定し、その具体的な値を確率的に推論した。得られた結果はどの程度良い表現なのだろうか？研究成果の信頼性を担保する上で、潜在表現の良さを適切に評価することは避けて通れない。しかし、潜在表現には正解が存在しないため、どのように評価すれば良いかは自明ではない。

本研究では欠損値の復元精度を評価指標として採用した。何度も強調してきたように、分析に用いた言語データベースは欠損値が大半を占めるが、残りの観測値の一部をわざと隠した上でモデルに復元させ、その正解率を測った。その結果、ベースライン手法を大きく上回る精度を達成することを明らかにした。また、系統的・地理的に近い言語が同じ値を取りやすいという追加の手がかりが精度向上に貢献することも明らかになった。とはいえ、一種の多数決を後処理として行った上で、精度が 70-80%程度である。通言語的に成り立つ傾向について議論するには十分な精度であるが、個別の系統関係の議論に説得力をもたせるには十分とは言えないというが研究代表者の感触である。

(2) 言語の時間変化の分析　　得られた潜在表現は構造的特徴同士の依存関係をモデル化したものである。例えば、日本語と北米先住民のラコタ語はいずれも基本語順が SOV (主語・目的語・動詞) だが、その他の特徴は異なるところが多い。例えば、情報構造という観点から見ると、日本語は旧情報を先に出してそれに新情報を結びつけるという戦略を取るが、ラコタ語は反対に新情報を先に出して旧情報を結びつけるという戦略を取る。これに応じて、主語や目的語が旧情

報の場合には動詞に後置されることがラコタ語においては自然である。一方、日本語においては動詞を最後に置くのは強い制約であり、逸脱は稀である。特徴間の依存関係を考慮すれば、日本語とラコタ語では基本語順の歴史的安定性が異なるのではないかと推測できる。

この仮説を検証するために、推定されたモデルを用いて 2 つの言語の未来の状態をシミュレートした。その結果、(1) 日本語の方がラコタ語よりも SOV 語順が安定的に維持されやすく、(2) ラコタ語は no dominant order（特に支配的な語順がない状態）に比較的移行しやすいという結果を得た。実際、ラコタ語を含むスー諸語との系統関係が取り沙汰されるカド諸語やイロコイ諸語に属する言語は WALS では no dominant order とされており、今回得られた結果と整合的である。

(3) 方言群の解析への応用可能性

ここまでで述べた通り、本研究における主要な対象は世界の言語であり、その構造的特徴であった。しかし、そのデータを分析するために提案した統計的モデルは、方言群の主に語彙的な特徴の解析にも転用できることに気づいた。この発見は研究代表者自身にとっても意外なものであり、統計モデルを用いた抽象化の醍醐味と言える。しかしこのことに気づいたのは研究期間の最末期であり、期間内に成果を結実させるまでには至らなかった。残された課題は後継研究において解決したい。

５．主な発表論文等

〔雑誌論文〕　計5件（うち査読付論文　3件／うち国際共著　0件／うちオープンアクセス　3件）

１．著者名 Yugo Murawaki	４．巻 45(2)
２．論文標題 Bayesian Learning of Latent Representations of Language Structures	５．発行年 2020年
３．雑誌名 Computational Linguistics	６．最初と最後の頁 228
掲載論文のDOI（デジタルオブジェクト識別子） 10.1162/coli_a_00346	査読の有無 有
オープンアクセス 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 －

１．著者名 村脇 有吾	４．巻 2020年3月
２．論文標題 基本語順の歴史的変化の数理モデル	５．発行年 2020年
３．雑誌名 数学セミナー	６．最初と最後の頁 36-40
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

１．著者名 Yugo Murawaki	４．巻 45(2)
２．論文標題 Bayesian Learning of Latent Representations of Language Structures	５．発行年 2019年
３．雑誌名 Computational Linguistics	６．最初と最後の頁 印刷中
掲載論文のDOI（デジタルオブジェクト識別子） 10.1162/COLI_a_00346	査読の有無 有
オープンアクセス 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 －

１．著者名 Yugo Murawaki	４．巻 -
２．論文標題 Analyzing Correlated Evolution of Multiple Features Using Latent Representations	５．発行年 2018年
３．雑誌名 Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing	６．最初と最後の頁 4382
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 有
オープンアクセス 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 －

1．著者名 村脇 有吾	4．巻 7
2．論文標題 言語系統論への計算的アプローチの可能性	5．発行年 2018年
3．雑誌名 歴史言語学	6．最初と最後の頁 77-91
掲載論文のDOI（デジタルオブジェクト識別子） なし	査読の有無 無
オープンアクセス オープンアクセスではない、又はオープンアクセスが困難	国際共著 －

〔学会発表〕　計4件（うち招待講演　0件／うち国際学会　0件）

1．発表者名 Yugo Murawaki
2．発表標題 Relaxing the Tree Constraint
3．学会等名 Fijian Languages Symposium
4．発表年 2020年

1．発表者名 村脇 有吾
2．発表標題 方言群の時空間解析にむけて：フィジー語を例に
3．学会等名 新学術領域・ヤポネシアゲノム・言語班2018年度第2回研究集会
4．発表年 2019年

1．発表者名 村脇 有吾
2．発表標題 潜在表現を用いた言語変化の通時的分析
3．学会等名 京都大学第13回 ICT イノベーション
4．発表年 2019年

１．発表者名
Yugo Murawaki

２．発表標題
Toward Spatio-Temporal Analysis of Dialects of Fijian

３．学会等名
Fijian Languages, Maps and Beyond: An Interim Report of the Fijian Language GIS (Geographic Information System) Project

４．発表年
2019年

〔図書〕　計0件

〔産業財産権〕

〔その他〕

lattyp
https://github.com/murawaki/lattyp/

６．研究組織

	氏名 （ローマ字氏名） （研究者番号）	所属研究機関・部局・職 （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