統計数理(2016)
第 64 巻 第 2 号 161–178

特集「統計的言語研究の現在」
[研究詳解]

言語変化と系統への統計的アプローチ

村脇 有吾†

(受付 2016 年 3 月 3 日；改訂 10 月 7 日；採択 10 月 7 日)

要　　旨

言語変化や諸言語の系統関係の解明といった歴史言語学の課題は，従来は言語学者が人手により取り組んできたが，21 世紀に入る前後から，計算機を用いた統計的手法を適用する事例が増えている．もともと分子生物学分野で開発され，近年言語データに適用されるようになった統計的手法は，年代のような連続値を含んでいたり，不確実性が解候補の組合せ爆発を生むなどの理由から人間が苦手としてきた問題に取り組むことを可能にしつつある．本稿の前半では，特に重要な手法である語彙を手がかりとしたベイズ系統モデルについて，歴史言語学の研究経緯を踏まえつつ，統計的な観点から解説する．ただし，語彙を手がかりとする手法は，インド・ヨーロッパ語族のような既知の語族に対しては一定の成果を上げつつあるが，日本語はモデルを適用する基盤が整っていない．そこで，本稿の後半では，日本語系統論を解決に導く可能性のある手がかりとして言語類型論の特徴に着目する取り組みについて紹介する．

キーワード：言語系統樹，歴史言語学，言語類型論，ベイズ統計．

1. はじめに

我々が話している言語が歴史的にどのような変化を経てきたか，複数の言語が歴史的にどのような関係を持つかといった問題に取り組む分野を歴史言語学とよぶ．言語は人間集団を特徴づける主要な要素であるため，千年のオーダの人類史(例えばヨーロッパにおける人類の定住過程)を解明する上で，歴史言語学は重要な役割を果たす．そのため，人類史の仮説として，歴史言語学だけでなく，集団遺伝学や考古学などの諸分野の知見と整合的なものを求める学際的研究も近年盛んである．

歴史言語学では，21 世紀に入る前後から，計算機を用いた統計的手法を適用する事例が増えている(Forster and Renfrew, 2006)．その特徴は，分子生物学分野で開発されたモデルを言語データに適用することによって主要な成果が得られていることである．そもそも，Darwin (1859)の『種の起源』と Schleicher (1853)によるインド・ヨーロッパ(印欧)語族の系統樹が 19 世紀半ばの同時期に発表されたことに象徴されるように，草創期の進化研究においては，生物と言語の類似性が意識されていた．しかし，その後の両分野の研究は目立った交流がないまま進んだ．統計的手法の適用という点では，生物学分野では 20 世紀後半に順調に研究が進展したのに対して，言語学においては人手による研究手法が主流であり続けた．こうしたなか，生物向けに開発された統計モデルが 1990 年代末から徐々に言語データに適用されはじめ，特に Gray and Atkinson (2003)が印欧祖語の年代推定にベイズ系統モデルを適用したことが大きな話題となった．これがきっかけとなり，系統モデルやその他の生物学由来の統計モデルを言語研究に導入

†京都大学大学院 情報学研究科：〒 606–8501 京都市左京区吉田本町

する事例があいついでいる．

そこで，本稿の前半では，近年の統計的言語研究の中心となっているベイズ系統モデルを紹介する．この統計モデルは，手がかりとして語彙を用いるという点で，伝統的な比較法や，先行する統計的手法である言語年代学と共通しており，実際，これらの研究成果の上に成り立っている．そのため，まずは歴史言語学の従来研究を統計という観点から整理し，その上でベイズ系統モデルを導入する．

なお，歴史言語学的課題に対する統計的取り組みについては，既に言語学者向けの丁寧なチュートリアル(Nichols and Warnow, 2008)が公表されている．しかし，このチュートリアルはベイズ系統モデルの中身についてはほとんど触れていない．一方，モデルの中身については，ベイズ系統推定ソフトウェア BEAST の開発者による網羅的な解説本(Drummond and Bouckaert, 2015)が出ている．しかし，この本は分子生物学のデータを想定しており，言語データは一切登場しない．そこで，本稿では，言語データに軸足を置いてモデルを解説したい(分子生物学における対応物には適宜関連づける)．

現在のベイズ系統モデルには，語彙に基づいた手法であるという点に限界がある．この手法が一定の成功を収めているのは，伝統的な比較法によって大まかな系統関係が既に明らかになっている言語群であり，印欧語族やオーストロネシア語族(台湾から東南アジア島嶼部，太平洋に広がる大語族)などが該当する．一方，日本語と他の言語の系統関係については，100 年以上にわたる諸研究者の尽力にもかかわらず，依然として不明のままである．したがって，このモデルを適用しようにも，必要な基盤が整っていない．

語彙に代わる手がかりとして，筆者は言語類型論の諸特徴に着目しており，本稿の後半ではこれを紹介する．類型論に関する素朴な議論は，比較法と同程度に古くから見られるが，類型論の系統推定への応用は確立されていない．その大きな理由は，語彙とくらべて，類型論の特徴が手がかりとして不確実であり，人手による論証になじまなかったことだと筆者は考えている．そして，この問題は計算機を用いた統計的手法により克服できると見込んでいる．まだまだ始めたばかりの研究だが，これまでの経過と今後の展望を述べたい．

なお，筆者は歴史言語学の体系的な教育を受けた者ではないことを断っておきたい．筆者の専門は計算言語学・自然言語処理とよばれ，この分野では，統計・機械学習を用いた研究が盛んに行われているものの，機械翻訳や質問応答といった応用処理や，それを支えるテキスト解析技術の開発が中心であり，歴史言語学的課題についてはほとんど認知されていない．それでも本稿を書くのは，筆者や本稿の読者にも参入の余地があることを訴えたいからである．発表文献に占める生物系の論文誌の多さが示すように，生物学を背景とする研究者が研究を主導する例が多いが，彼らにとって，生物向けのモデルをそのまま言語データに転用するのではなく，言語独自のモデルを開発する動機はとぼしい．一方の言語学者は，当然ながら言語現象を深く分析し，その性質を考察している．しかし，主流研究が統計とは無縁であったこともあり，統計モデルを用いて問題を解くという発想を受け入れること自体に困難が見られる(Pereltsvaig and Lewis, 2015)．生物学由来のモデルを受け入れた一部の言語学者についても，ほとんどの場合に生物データ向けの既存のソフトウェアパッケージに依存している．そのため，生物に対応物が存在しない言語現象がモデル化から取り残される傾向にある．見方を変えれば，この断絶を埋める研究にいち早く取り組むことで，大きな成果が得られると期待している．

2. 基本的な概念とデータ

2.1 進化と系統樹

言語(や生物種)はそれ自体が複雑なシステムであることから，一部の側面を特徴として取り

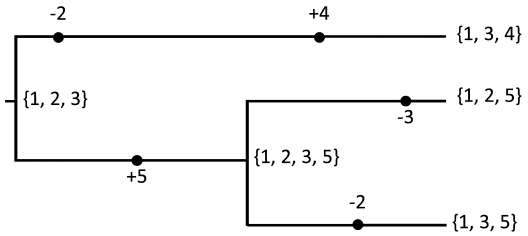

図 1．系統樹の例．

出して分析に用いる．例えば，生物種であれば，羽の有無や歯の形といった形態的特徴や，ゲノムの塩基配列，そこから取り出した STR（縦列型反復配列）や SNP（一塩基多型）などの特徴が用いられる．言語の場合は，語彙的特徴（ある語を持つか否か）が広く用いられているが，本稿ではほかに類型論的特徴にも着目する．特徴にはキリンの首の長さのような連続値もありえるが，本稿では離散値のみを扱う．

いま，言語（種）について，それが持つ諸特徴を取り出し，便宜的に $10110\cdots$ のように並べた列を状態とする．ここでは，簡単のために，ひとまず 2 値特徴を考えている．このとき，特徴に番号を振り，1 が立っている特徴のみに着目することで，同じ特徴を $\{1,3,4,\ldots\}$ のように集合的に表現することも可能である．

進化とは，親から子へと途切れなく状態が受け継がれるが，完璧に複製されるのではなく，次第に変化が蓄積する現象を指す．生物種の場合は遺伝により，言語の場合は第一言語習得により状態が継承される．なお，進化は変化を伴う由来（descent with modification）ともよばれるように，価値判断を含まず，単に変化するという現象のみに着目している．2 値特徴の場合，変化は特徴の誕生，死亡の 2 種類からなる．それぞれ $+4$，-2 のように表すとする．

系統樹は，図 1 のように，複数の言語（種）の歴史的関係を木構造により要約したものである．系統樹を特徴づけるのは分岐である．2 つの言語は，分岐前は完全に同一であり，分岐後は独立に進化するとしている．系統樹には，トポロジーのみを表した時間なし系統樹と，各ノードに年代が紐づいた時間つき系統樹がある．

通常の問題設定では，現代語や文献に記録された古代語，つまり系統樹の葉ノードの状態が観測されている．推定すべきは，観測されていない部分，つまり木のトポロジー，および祖先（祖語）の状態や年代などである．伝統的な歴史言語学では，時間なし系統樹が人手により推定されてきたが，統計的手法では，時間つき系統樹の推定を目的とすることが多い．なお，言語の場合，疎遠な言語同士の関係は一般に不明のままである．共通祖語を持つことが立証できたとき，系統関係が確立されたと言う．

葉ノードの状態が得られたとき，時間なし系統樹相当のものは素朴な手法でも作ることができる．まず，言語対に対して適当に距離が定義できる．仮に変化の速度がほぼ一定と仮定すると，言語対の距離は分岐後の時間におおよそ比例する．したがって，距離に基づく階層的クラスタリングを適用すれば木が得られる．より複雑なモデルを適用する場合でも，推定される系統樹はクラスタリング結果と大まかには似たものとなる．

ただし，系統推定のさまたげとなる厄介な現象がいくつか知られている．系統樹は分岐後の独立進化を仮定しているが，実際には，言語同士の接触により特徴が変化する場合がある．この現象は生物系の用語で水平伝播（horizontal/lateral transmission）とよばれる．言語の場合は語彙の借用が典型例として挙げられる．また，同じ特徴が系統樹上の複数箇所で誕生したり，一度死亡した特徴が復活することも考えられる．これらの現象をそれぞれ成因的相同（homoplasy），

復帰突然変異(back mutation)とよぶ．このような現象が生じうる場合，2 つの言語が同じ特徴を有していたとしても，それが祖語に由来するとは限らない．

2.2 データベースとその整備

統計的研究を行うには，準備として特徴のデータベースの整備が不可欠である．生物データ，特に集団遺伝学で用いられるゲノムデータと比較したとき，言語データの特性として，規模の小ささと高コスト性が挙げられる．

規模については，ヒトゲノムの SNP の場合，10 万のオーダの特徴が得られる(International HapMap Consortium, 2005)．個体数についても，千あるいはそれ以上のオーダで得られ，その数は今後も増え続けると見込まれる．さらに，各個体には，日本人，サルデーニャ人，ヨルバ人といった集団ラベルを割り振ることができる．つまり，日本人という集団は複数の個体によって表現され，個体間のばらつきが進化史の解明の手がかりとして利用できる．

一方，言語の場合，語彙的特徴，類型論的特徴のいずれについても，数は百のオーダにすぎない．比較可能な言語数も，語彙的特徴では十から百，類型論的特徴でも千のオーダで頭打ちである．さらに，言語は集団ごとに 1 つ採取される．言語はコミュニケーションの手段であり，発信者と受信者の双方が理解できなければならないため，集団内での大きなばらつきが期待できないからである．こうした制約から，集団遺伝学で近年発展した諸手法は言語データには適用できない場合が多い．代わって適用されるのは，ヒトとチンパンジーとの共通祖先の年代推定のようなよりマクロな比較や，進化の速度が桁違いなウイルスの系統推定に用いられてきた手法である．

言語データベース整備の高コスト性は，次世代シーケンサのような機械的手段ではなく，もっぱら言語学者が人手でデータを作成していることに起因する．一つの言語の習得だけでも何年もの時間を要するなか，複数の言語から斉一な基準にしたがってデータを採取するには高度な専門知識が欠かせない．おまけに，現状ではデータ整備の機械化の見通しは立たない．言語の話者数には著しい不均衡があり，数千とも言われる世界言語の大半は，電子化されたテキストどころか文字すら確立されていない小言語だからである．そして，そうした小言語が系統推定に重要な役割を果たすことが少なくない．統計・機械学習分野では，音声認識研究の大家 Fred Jelinek のものとされる，「言語学者をクビにするたびに音声認識器の精度が上がる」という発言が知られているが，歴史言語学では，統計的手法を用いる場合でも，依然として言語学者の貢献が欠かせない．

統計や言語処理の研究者にとって好都合なことに，データ整備という高い参入障壁は引き下げられつつある．統計的研究と並行して，2000 年代以降，言語データの整備と公開も急速に進んでいる．特に，マックス・プランク進化人類学研究所をはじめとするマックス・プランクの研究所群からは，後述の WALS (Haspelmath et al., 2005)と APiCS (Michaelis et al., 2013)のほか，Glottolog (Hammarström et al., 2016)などの有用な言語データが公開されている．データ公開による共有には副作用もある．参入障壁の引き下げは，その性質を深く理解しないままデータを扱う研究を生み出す危険がある．実際，類型論のデータベース(Haspelmath et al., 2005)の公開後，言語の特徴と人間や環境の特徴との相関を探す怪しげな研究が数多く発表されている(Roberts and Winters, 2013)．しかし，こうした問題を懸念してデータの公開を控えるよりも，通常の科学的批判を通じて淘汰を行う方が健全だと筆者は考えている．

3. 語彙に基づく系統推定の従来手法

3.1 比較法

歴史言語学の伝統的な手法や，近年のベイズ系統モデルの多くは，言語間の系統推定に語彙

的特徴を用いる．語彙に基づく系統推定の基盤は記号の恣意性である．DOG という意味と「いぬ」という音の結びつきに必然性はない．このことから，DOG を意味する「いぬ」という語が歴史上無関係に複数回発生する可能性は極めて低い．つまり，成因的相同や復帰突然変異の可能性が排除できる．ただし，接触による借用は起こりえる．また，語そのものは引き継いでいても，語形は時間とともに変化するため，ある言語対が持つ語が同一特徴か否かは自明ではない．「名前」と *name*，「骨」と *bone* のような偶然の類似や借用を排除し，祖語から引き継いだ特徴であることを立証する必要がある．そうした語を同源語(cognate)とよぶ．

比較法(英語でも単に comparative method とよばれる)では，同源語の特定に音法則を用いる．歴史的な音の変化は，例外なく規則的に起こることが知られている．結果として，ある言語対が持つ同源語の語形には規則的な音対応が見られる．こうした音対応を立証することで，偶然の類似や借用を排除し，同源語を特定できる．

言語間の系統関係が確立するためには，通常は百のオーダの同源語を特定する必要がある．ただし，重要なのは量そのものではなく，音法則を通じて特徴の特定の質を担保することである．結局のところ，比較法は，対象となる諸言語が祖語から同一特徴を受け継いでいることを示すにすぎない．なお，3 個以上の言語の系統関係については，分岐の前後関係を明らかにする必要がある．そのために，例えば，語彙ではなく個々の音変化そのものを特徴とみなし，言語間での特徴の共有を調べるといったことが行われる(Pellard, 2009)．

歴史言語学では，さらに祖語の語形の再構が試みられるが，観測できる手がかりは不完全であり，再構形は理論上の産物という側面が強い．とはいえ，この作業は語形という記号列の操作であり，人手による論証と親和性が高い．計算機を用いた統計的祖語再構も提案されているが，歴史言語学の成果を追認するにとどまっており，言語学上の新たな知見はとぼしい(Bouchard-Côté et al., 2013)．また，成功事例が報告されているのは，600 以上の言語からなる世界的にも稀な大語族(オーストロネシア語族)のみであり，小規模データに対してはうまく働かないのではないかと推測される．

また，比較法だけでは祖語の年代は推定できない．人間は年代のような連続値を直接扱うのが苦手であり，統計的手法の出番となる．

3.2 言語年代学

これに対し，言語年代学(glottochronology)は，その名の通り，言語対の祖語の年代を推定する統計的手法である．この手法は，考古学における放射性炭素年代測定に触発されて生まれたもので，生物の体内にわずかに含まれる放射性炭素が死後一定割合で減衰していくのと同様に，祖語にあった語彙的特徴も一定割合で失われると仮定する(Swadesh, 1952)．言語年代学の研究は 1940 年代末から 50 年代を中心に行われており，生物学における分子時計(molecular clock)仮説(Zuckerkandl and Pauling, 1965)に先行する．

言語年代学では，準備として基礎語彙を設定する．基礎語彙とは，どんな言語でもそれを表す言葉があるような基本的な概念(100 から 200 項目)である．例えば，WATER, BIG, EYE などが該当し，SNOW のように地域が限定される概念や，MILLION や PAPER のような文化語彙は排除される．また，基礎語彙は借用されにくく，比較的変化しにくいと仮定される．例えば，ある調査によると，英語の一般語彙はおおよそ 50%が借用語だが，基礎語彙に限ると 6%にすぎない(Swadesh, 1952)．各言語について基礎語彙を収集し，次に同源語を特定する．特定には上述の比較法が用いられる．

言語年代学における目標は，言語対 A, B について，それらの祖語 P の年代 t を推定することである．ここでの仮定は，P が持っていた基礎語彙が時間とともに一定割合で失われるというものである．A, B の基礎語彙共有率を c とすると，基礎語彙の残存率 r が与えられたとき，祖

語の年代は

$$t = \frac{\log c}{2 \log r}$$

と求まる(分母の 2 は，P から A，P から B の 2 本の枝に対応)．残存率 r は文献が豊富な言語群を用いて推定しておく．例えば，$r = 0.81$(200 項目基礎語彙で単位は千年)とすると，共有率 $c = 0.43$ のとき，$t \approx 2$(千年)が得られる．

言語年代学は歴史言語学に統計を持ち込んだ先駆的な研究だが，言語学者の間では評判が悪い．特に批判が集中したのは残存率一定という仮定である．例えば，極端な保守性で知られるアイスランド語の場合，古ノルド語と比較すると，残存率が 0.95 を超えており，0.81 という仮定からかけ離れている(Bergsland and Vogt, 1962)．

こうして言語年代学は激しい批判にさらされて衰退し，1990 年代末から 2000 年代にかけて，後発の分子生物学由来の手法で置き換えられることとなった．しかし，基礎語彙データの収集という面では貢献が大きく，近年の統計的研究も言語年代学(より一般には語彙統計学)の成果に依存している．

4. ベイズ系統モデル

言語年代学に代わって 2000 年頃から言語に適用され始めたベイズ系統モデルは，もとは分子生物学のデータを解析するために開発されたものである．分子生物学においても，当初は素朴な階層的クラスタリング手法や，確率モデルを用いる場合でも最尤推定法が用いられてきたが，1990 年代後半からベイズ系統モデルが盛んに研究されるようになった(Huelsenbeck and Ronquist, 2001)．ベイズ系統モデルの利点として，生成モデルであることから結果の解釈が容易なこと，様々な事前知識を柔軟に組み込めること，Markov chain Monte Carlo(MCMC)という理論的裏付けのある推論手法が存在することが挙げられる．

なお，ベイズ系統モデルの言語データへの応用事例として(Gray and Atkinson, 2003)が有名だが，これ以前にも(Gray and Jordan, 2000)が生物学由来の統計的系統モデルを言語データに適用している．ただし，この研究で用いた系統モデルはベイズ以前の階層的クラスタリング手法であり，年代推定は行っていない．

4.1 確率的解釈

ベイズ系統モデルは大掛かりなモデルであることから，準備として，より単純な言語変化の確率的解釈を考える．変化はメトロノームのように一定間隔で起きるのではなく，確率的ゆらぎがあると思われる．仮に一定のものがあるならば，それは確率分布のパラメータであると仮定するのが自然である．

実際のベイズ系統モデルでは，単語の誕生，死亡という 2 種類の変化をモデル化するが，簡単のために，まずは種類に関わりなく変化の回数を数えることとし，その回数と時間との関係をモデル化する．連続時間において独立に発生する離散事象を数えるモデルとしてポワソン過程(Kingman, 1993)が知られている．N_t を時間幅 $[0, t]$ で起きた独立な変化の数とすると，これはポワソン分布，$P(N_t = k) = \frac{e^{-\mu t}(\mu t)^k}{k!}$ に従う．また，2 つの連続した事象の間隔は指数分布 $e^{\mu t}$ に従う．ここではパラメータ μ を変化率とよぶことにする．

ポワソン過程からの複数の試行を図 2 に示す．変化率 μ が一定であっても，一定数の変化に要する時間に確率的ゆらぎがあることが確認できる．上述の言語年代学では，言語対を入力とし，共通祖語からの経過時間 t を点推定していた．これに対し，確率的解釈では，経過時間 t を含む一連の変化をモデルに与え，その確からしさを確率で返す．ある回数の変化が起きるのに要する時間は，点ではなく確率分布により表現される．つまり，考えられる様々な可能性に対

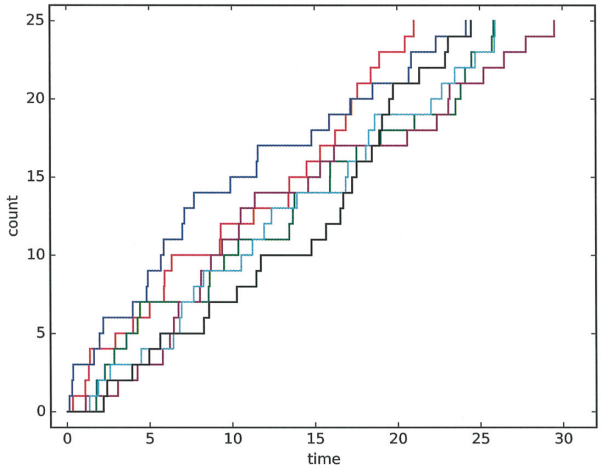

図 2. ポワソン過程からの試行 ($\mu = 1$).

して，その自然さを確率によって評価していることになる．これは，言語対にとどまらず，複数の言語を含む系統樹のモデルを設計する上で重要な性質である．系統樹は複雑な構造であり，局所的には最適でない解釈が全体的な整合性を考えると良いことがありえるが，こうした場合への柔軟な対応が可能となる．

さらに，ベイズモデルにおいては，パラメータ μ に事前分布を設定し，

$$P(\text{時間つき変化過程}, \mu; \alpha) \propto P(\text{時間つき変化過程}\ |\mu) \times P(\mu; \alpha)$$

とすることで，パラメータ割り当てにも確率を与える．後で見るように，系統樹は多くの潜在変数を含むが，パラメータに事前分布を置くことで，パラメータをその他の潜在変数と統一的に扱え，推論時に都合が良い．

掛け算は対数化すると足し算となる．例えば，時間つき変化過程の対数確率は

$$\log P(\text{時間つき変化過程}\ |\mu) + \log P(\mu; \alpha) + C$$

となり，定数項 C を無視すれば，対数確率というスコア 2 つの足し算によって時間つき変化過程の自然さが採点されていることになる．系統樹はより複雑な部分モデルの組み合わせによって構成されるが，基本は同じである．部分モデルは系統樹の構成要素の自然さをスコアによって採点するものであり，それらのスコアを合算すれば系統樹全体のスコアとなる．

4.2 モデル設計と推論

ベイズ系統モデルは，時間つき系統樹をモデル化する．上述の確率的解釈と同様に，モデルは $P(\text{時間つき系統樹}, \theta; \alpha)$ と表される．ここで，θ はモデルのパラメータ群，α はハイパーパラメータ群を表す．パラメータを含む系統樹は複雑な構造であり，いくつかの構成要素（部分モデル）に分解することで構成される．自然な系統樹に相対的に高いスコア（対数確率）を与えるようなモデルの設計が最初の目標となる．ベイズ系統モデルは図 1 のような系統樹を直接スコアで評価しており（後述のように正確には異なる），結果を素直に解釈できることが魅力の一つである．

次に，与えられたモデルのもとで，高いスコアを返すような系統樹とパラメータの組を探す．

この手続きを推論とよぶ．系統樹のうち，葉ノードの状態と年代は観測されている．さらに，いくつかの中間ノードの年代や，場合によっては部分木もモデルに与える．残りは潜在変数であり，連続値を含むことから非加算無限個の候補があるが，とにかく一通り値を割り当てれば，モデルからスコアが得られる．

4.3 部分モデル

部分モデルは現在でも活発に研究されており，その組み合わせには多数の変種がある．部分モデルは大きくは木モデル，置換モデル，時計モデルからなり，他にも各種パラメータの事前分布がある．

木モデルは，ノードの状態を無視し，時間つき木の骨組みを採点する．モデルの例として，Yule 過程，誕生・死亡モデル(birth-death model)，Bayesian skyline モデル等が知られている．しかし，生物学上の問題意識から開発されており，言語系統樹における意味はあまり明らかにされていない．

木モデルに関してむしろ重要なのは，年代較正(calibration)である．内部ノード(例えばインド・イラン祖語)や葉ノード(例えばラテン語)の年代について既知であることを他のノードの年代推定に利用する．生物の場合は主に化石の年代を用いる．年代は点で与えることも可能だが，多くの場合は正規分布のようなソフトな制約を設定する．このとき，平均から離れた年代を該当ノードに割り当てるほどスコアが減点されることになる．例えば，ラテン語の年代の事前分布を $\mathcal{N}(\mu = 2050.0, \sigma = 75.0)$ と置くと，2050BP(before present)から 75 年離れると 0.5，150 年離れると 2.0 の減点(いずれも対数値)が課される．結果として，推論時には，こうしたソフトな制約を満たすような変化率が推定される．

置換モデルは，親から子への状態変化を採点する．ポワソン過程では変化の数を数えたが，置換モデルは，それぞれの 2 値特徴の具体的な値の変化をモデル化する．図 1 のように親から子への枝のどの時刻で変化が起きたかを陽に持つことも可能だが，効率化のために始点と終点の両ノードの値のみに着目する．すなわち，ある親言語のある特徴の値が $x' = i \in \{0, 1\}$ のとき，時間 t 後の言語の特徴 x の値が j である確率 $P(x = j|x' = i, t)$ を設計したい．これは i から始まり，時間 t 後に j となるすべての遷移の確率を積分したものである．

置換モデルは，通常，連続時間マルコフ連鎖によってモデル化される．連続時間マルコフ連鎖はパラメータとして遷移率行列 Q を持つ．2 値特徴の場合，$Q = \left(\begin{smallmatrix} -\alpha & \alpha \\ \beta & -\beta \end{smallmatrix}\right)$，という 2×2 の行列で表される．ここで，$\alpha \geq 0$ は $0 \to 1$ の変化の起こりやすさ，$\beta \geq 0$ は $1 \to 0$ の変化の起こりやすさを表す．遷移率行列 Q を用いると，親から子へのある特徴の変化の確率は

$$P(x = j|x' = i, t) = \exp(tQ)_{i,j}$$

と求まる．図 3 に連続時間マルコフ連鎖の例を示す．遷移率行列 $Q = \left(\begin{smallmatrix} -0.5 & 0.5 \\ 0.25 & -0.25 \end{smallmatrix}\right)$，$i = 0$ とし，$j = 0$ の確率を実線，$j = 1$ の確率を破線で示している．$t = 0$ では初期値のままの $j = 0$ である確率が 1 だが，時間とともに $j = 1$ となる確率が上昇し，定常分布に収束している．遷移率行列の要素の大きさは，定常分布や収束の速さを制御する．

なお，内部ノードの状態を陽に持つのではなく，周辺化によりすべての状態の組み合わせを一度に考慮することも広く行われている．この周辺化は動的計画法により効率的に解けることが知られている(Felsenstein, 1981)．

置換モデルとしては，歴史的には，生物学で遺伝子(ACGT)の置換に対応する 4×4 の遷移率行列が 1960 年代末から研究されてきた．なお，連続時間マルコフ連鎖は $1 \to 0 \to 1$ のように，死亡した特徴が復活する場合も考慮するが，語彙の場合，復帰突然変異は起きないと仮定するのが自然である．この問題に対応するために，復活のないモデルとして，確率的 Dollo モデル

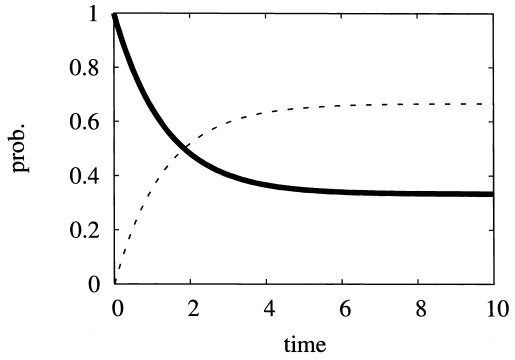

図 3. 連続時間マルコフ連鎖の例.

が提案されている(Nicholls and Gray, 2008).

これに対し, 遷移率を拡張するのが時計モデルである. 従来からの厳密時計モデルでは, 系統樹全体で同じ遷移率が用いられる. しかし, 変化の速度は一定とは限らないという問題意識が生物学においてもあり, 1990 年代以降, 変化率に柔軟性を持たせる緩和時計が研究されてきた. 緩和時計モデルの多くは, $P(x = j|x' = i, t, k) = \exp(r_k tQ)_{i,j}$ のように, 枝 k ごとに異なる係数 r_k を遷移率行列にかけることで実現される. r_k 自体も確率分布から生成されるが, 具体的なモデルは乱立しており, 評価が定まっていない.

4.4 推論

推論時には, 与えられたモデルのもとで, 観測データを入力し, 高いスコアを返すような系統樹とパラメータの組を探す. ただし, モデルの構成要素には複雑な依存関係があり, 最適解は解析的には求められない. そこで, 乱択により近似的に解を探索する. その具体的な手法として, Markov chain Monte Carlo(MCMC)法によるサンプリングが広く用いられている.

系統樹とパラメータの組のうち, 一部は観測されており, 残りは潜在変数である. 潜在変数は連続値を含むことから非加算無限個の候補があり, 離散値に限っても組合せ爆発を起こすため候補の列挙は事実上不可能である. しかし, ともかく一通り値を割り当てればモデルからスコアが得られる. また, 系統樹やパラメータを局所的に変更すると, 新たなスコアが得られる(差分に着目すれば効率的に計算できる). そこで, ひとまず潜在変数を一通り割り当て, 局所的な変更を繰り返すことで, 良い系統樹を近似的に探索するという方針をとる.

MCMC 法の一つである Metropolis-Hastings アルゴリズムでは, スコアが上がった場合はその変更を採用する. スコアが下がった場合は, 多くの場合は変更を破棄して元の割り当てを採用するが, ある確率で変更を採用する. これにより局所解からの脱出が可能となる. 適当な条件が満たされるとき, この操作を無限に繰り返すことで, 最適解にたどり着くことが保証される.

この手続きがサンプリングとよばれるのは, 確率分布からサンプルを得ているとみなせるからである. 簡単な例では, サイコロを振って $3, 4, 5, 3, \ldots$ とサイコロの目を得たとき, それらは多項分布(いわゆる categorical distribution で, ベルヌーイ分布の多項版)からのサンプルである. 連続値の場合も同様に, 正規分布から $1.78\ldots, -0.32\ldots, 0.27\ldots, \ldots$ といったサンプルが得られる. 系統モデルはこれらにくらべてはるかに複雑で, 連続値と離散値の組み合わせからなる確率分布だが, MCMC 法により得られる系統樹とパラメータの組は, この確率分布からのサンプルとみなせる.

こうして系統樹とパラメータの組の複数のサンプルが得られるが，人間が容易に解釈可能な形で要約する必要がある．多くの場合，興味の中心は共通祖語の年代であるため，サンプルから年代を集めてきてヒストグラムを作ればよい．複数の系統樹の要約方法には様々な変種があるが，なかでも最大系統群信頼度木(Heled and Bouckaert, 2013)とよばれる手法がよく用いられる．

4.5 現状と今後の展望

分子生物学由来の統計モデルの適用に対する言語学者の反応は，大半が懐疑的であるという印象を筆者は抱いている．特に一部の言語学者からは激烈な反応が発せられている(Pereltsvaig and Lewis, 2015)．こうした反応のなかでは，これまでの研究経緯を無視したような論文が，言語学者の査読を経ていない(ようにみえる)まま高名な論文誌に掲載されることへのいらだちや，人文系の予算が急速に削減されるなか，計算機を使った研究が科学メディア等でもてはやされていることへの感情的な反発が渾然一体となっている．しかし，統計的手法の研究者の目からは，データベース整備や定性的な議論の面で言語学者の協力が不可欠であることは充分すぎるほど明らかである．したがって，言語学者が説得すべき相手は，統計的手法の研究者よりも，むしろ科学メディアや研究資金提供機関であろう．また，Bouckaert et al. (2012)の著者の一人である Drummond と Pereltsvaig and Lewis (2015)の著者らのブログ上でのかみ合わない議論(http://www.geocurrents.info/cultural-geography/linguistic-geography/mismodeling-indo-european-origin-and-expansion-bouckaert-atkinson-wade-and-the-assault-on-historical-linguistics#disqus_thread)は，伝統的な歴史言語学の教育を受けた者が，統計モデルを用いて問題を解くという発想を受け入れることの難しさを痛感させるものである．これは高等教育の制度設計にかかわる問題であり，一朝一夕には解決できない．

一連の統計的研究の火付け役となった Gray and Atkinson (2003)をはじめとする多くの研究は印欧祖語を対象としている．印欧祖語の故地と年代に関してこれまでに多くの説が提案されてきたが，特に有力なのはクルガン仮説とアナトリア仮説である．クルガン仮説は，考古学的証拠をもとに，5,000–6,000 年前，黒海沿岸のステップ地帯(クルガンはこの地帯にみられる墳墓のこと)に印欧祖語の話者がおり，遊牧民の軍事的征服により各地に広がったとする．一方のアナトリア仮説は，同じく考古学的証拠をもとに，8,000–9,500 年前のアナトリア(現在のトルコのアジア側)を起源とし，農耕とともに拡散したとする．ベイズ系統モデルを用いる Gray らは，一貫してアナトリア仮説を支持している(Gray and Atkinson, 2003; Bouckaert et al., 2012)．しかし，アナトリア仮説は言語学者の間では評判が悪く，これがベイズ系統モデルへの不信感にもつながっているのではないかと筆者は推測している．

もしクルガン仮説が正しいとすると，これまでのモデルのどこに問題があったのだろうか．この問題について，Chang et al. (2015)は意味変化による成因的相同に着目している．伝統的な比較法では，意味変化にかかわらず同源語を追跡するが，言語年代学以降の統計手法では，最初に意味ごとに区切って語を収集する．しかし，ある種の意味変化は通言語的によく起こり，結果として成因的相同を生み出している．例えば，ADULT MALE を意味する現代アイルランド語の *duine*，フランス語の *homme*，ゴート語の *guma* は同源語だが，PERSON から ADULT MALE への意味変化が，古アイルランド語から現代アイルランド語，およびラテン語からフランス語にいたる過程で独立に起きた結果誕生したものである(ゴート語も同様と思われる)．しかし，系統推定を行うと，それらの言語の共通祖先にまでこの語をさかのぼらせるのがより自然な解釈となる．つまり，変化を実際よりも古い段階に持って行き，かつ，時間あたりの変化数を実際よりも低く推定してしまう．

この問題に対処するために，Chang et al. (2015)は，古代語を現代語の過去の状態として使う

というモデル変更を提案している．例えば，現代アイルランド語の ADULT MALE の意味での *duine* の値は 1 だが，古アイルランド語の状態を経るため，時間をさかのぼって他の言語と合流する前に値が 0 となる．これに対し，従来手法では古代語も葉ノード扱いしていたため，内部ノードとして現代語・古代語共通祖語が推定され，その値は 1 になる可能性が高かった．この変更の結果，祖語の年代は約 6,500 年前と推定され，Bouckaert et al. (2012)とくらべて大幅にクルガン仮説に近づいている．

Chang et al. (2015)の第一著者は，計算機科学から言語学に転じたという特異な経歴を持っている．言語現象を丁寧に分析し，それに対応する統計モデルを提案するというこの研究の取り組みは，今後の研究のあり方を示す模範例だと筆者は考えている．

Chang et al. (2015)は従来のモデルの仮定が成り立っていないことを指摘したものだが，そもそも系統モデルは多数の部分モデルの組み合わせであり，それぞれの部分モデルの背後にはデータに関する仮定がある．派手な研究成果が喧伝される一方で，数多くの仮定の検証はまだ充分になされていない．例えば，上述のアイスランド語のように極端に変化が遅い言語を含むデータに対しても緩和時計モデルであれば適合することが期待されるが，これまでの報告を読む限りでは，実際にうまくいっているか判然としない．地道な検証が必要であり，そのための道標として，歴史言語学でこれまでになされてきた議論に有用なものが少なくないと考えている．

仮定のなかで疑うべき最大のものは，系統樹そのものかもしれない．系統樹は理想化にすぎず，系統モデルでは説明できない接触に基づく言語現象が存在することは，歴史言語学においては系統樹の発明当初から認識されてきた(Schmidt, 1872)．これに応じて，系統モデルを推進する Gray らも，接触が系統推定に与える影響について多くの議論を費やしている(Greenhill et al., 2009; Gray et al., 2010)．

特に方言(非常に近い言語)群の関係を考える場合，接触の影響が無視できる範囲を超えていると考えられる．実際，伝統的な方言区画論は現代の特徴に基づく階層的クラスタリングにすぎず，それが時間変化を表す系統樹に対応するという観念は希薄である(東条, 1927)．方言群に強引に系統モデルを適用した報告(Lee and Hasegawa, 2011)もあるが，その結果得られた不可解な系統樹は，少なくとも部分的には接触の影響によって説明できると見られる(Murawaki, 2015b)．方言の語彙的特徴については，むしろ拡散を扱う非統計的モデルが提案されてきた(柳田, 1930; Trudgill, 1974)．統計モデルについては，シミュレーションモデル(Lizana et al., 2011; Murawaki, 2015b)は提案されているものの，実データを扱えるモデルと推論手続きの開発が課題として残っている．接触を考慮すると，モデルの自由度が単なる系統樹にくらべて大幅に上がるからである．

5. 言語類型論に基づく系統推定に向けて

5.1 日本語の起源と言語類型論

現在のところ，語彙的特徴を用いる統計モデルは，日本語と他の言語との系統関係の解明には利用できない．日本語系統論の研究には百年以上の蓄積があるにもかかわらず，日本語と他の言語との間で信頼できる同源語が充分に特定できていないからである(Vovin, 2010)．このことから，逆説的に，仮に同系言語が現存していたとしても，祖語の年代は相当さかのぼるのではないかと推測される(服部, 1999)．

筆者は語彙に代わる手がかりとして言語類型論に注目している(Murawaki, 2015a)．言語類型論とは，世界の諸言語を類型によって分類する分野である．類型の例には，基本語順(SVO, SOV 等)，助数詞の有無，声調の有無がある．これらの特徴を用いると，言語の状態は多値特徴の列で表現できる．

類型論の利点として，語彙とは異なり，日本語を含む任意の言語対が比較できることが挙げられる．実際，日本語と同系の言語の有力候補として，朝鮮語，さらにはツングース，モンゴル，テュルクを核とするいわゆるアルタイ諸語が挙げられてきたが，その根拠となったのは，「語頭に r 音が立たない」，「*have* 型の所有動詞を持たない」といった類型論上の類似である．また，類型論のいくつかの特徴は，語彙とくらべて歴史的に安定的だと推測される（Nichols, 1992; 松本, 2007）．

しかし，歴史言語学では，系統関係は語彙的特徴によって確立されるもので，類型論上の類似は決定的な証拠にならないという見方が支配的である．アルタイ諸語の場合も，従来指摘された特徴は実は世界的にありふれており，該当言語群に固有のものではないことが指摘されている（松本, 2007）．

進化という観点から類型論的特徴の性質を見直すと，欠点としてまず挙げられるのは，成因的相同や復帰突然変異が広範囲に起こりえることである．例えば，ある言語対の基本語順の値がいずれも SVO だとしても，その特徴を共通祖先から引き継いだとは限らない．SVO 語順は歴史上無関係に複数回誕生したと考えられるし，一度失われたとしても復活しうるからである．つまり，語彙と違って類型論的特徴は手がかりとして不確実性が高く，人手による論証になじまない．見方を変えれば，計算機を用いた統計的手法が活躍できそうな未開拓地が広がっている．まずは，データに成り立つ統計的性質を明らかにするところからはじめる必要があるだろう．

これまでにも類型論の系統推定への利用を試みる報告もいくつかあるが（Tsunoda et al., 1995; Dunn et al., 2005; Longobardi and Guardiano, 2009），語彙に基づく手法とくらべて少ない．系統推定の手がかりとしての類型論的特徴の有効性についても，肯定的な報告（Dunn et al., 2005; Longobardi and Guardiano, 2009）とやや否定的な報告（Greenhill et al., 2010; Dunn et al., 2011）が混在し，結論が出ていない．

従来，類型論に対する統計的取り組みの障害となっていたのは，データ整備に言語学の高度な知識が必要となることである．例えば，動詞のように自明に思える概念であっても，世界中の言語を収集すると，悩ましい事例が出てくる．かつては言語学者が個別にデータを収集していた（角田, 1991）が，現在は組織的なデータ収集の成果として World Atlas of Language Structures（WALS）とよばれるデータベース（Haspelmath et al., 2005）が公開されており，望むなら統計モデルの設計に専念することも可能な環境が整いつつある．

5.2 類型論的特徴の変化の経路

類型論的特徴はどのように変化するのだろうか．そもそも，語彙の交代とくらべると，例えば語順の変化がどのように起きるかは直感的に想像しづらい．基本語順が SOV から SVO に変化するのは一大変化であり，言語システム全体に複雑な影響があると思われる．

これまでの系統モデルの研究では，語彙的特徴の場合と同様に，特徴ごとに独立な置換モデルを仮定することが多い（Teh et al., 2008; Daumé III, 2009; Maurits and Griffiths, 2014）．しかし，類型論的特徴の場合，特徴間の依存関係は無視できる範囲を超えているのではないかと筆者は推測している．

実際，言語類型論の従来研究では，類型論の特徴間の依存関係が大きな関心事であった．例えば，数詞（Q）と名詞（N）の語順と形容詞（A）と名詞（N）の語順を考えると，QN，NQ と AN，NA の組み合わせにより，図 4 のように 4 通りの状態を取りえる．しかし，世界の言語を見ると，（QN, AN）型および（NQ, NA）型という一貫した語順を持つ言語が多く，（QN, NA）型の言語も存在するが，（NQ, AN）型の言語は非常に稀である．つまり，「NQ ならば NA」という含意関係がよく成り立つ．さらに歴史的変化を考え，もし（QN, AN）型から（NQ, NA）型への変化があったとすると，途中で（NQ, AN）型ではなく，（QN, NA）型を経由したと解釈するのが自然である

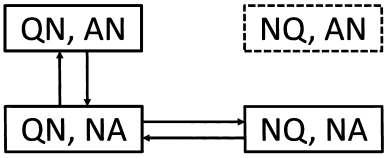

図 4. 2 つの特徴からなる状態とその遷移.

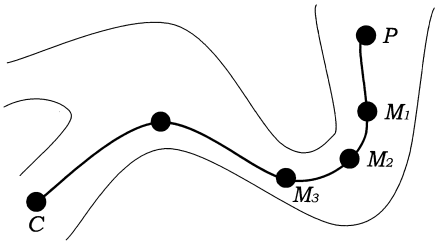

図 5. 祖語 P から言語 C への多様体(帯)上の遷移(太線).

(Greenberg, 1978). ここで, 特徴ごとに独立な変化を仮定すると, (NQ, AN)型の祖語を推定するおそれがある.

依存関係に関する統計的研究としては, 特徴対の含意関係をデータから自動発見するモデル(Daumé III and Campbell, 2007)が提案されているが, 統計が真価を発揮するのは, むしろ 3 つ以上の特徴間の関係のモデル化だと筆者は考えている. 類型論データベースに記述された特徴の数は百のオーダであり, このような組み合わせは人間の手に余るが, 計算機であれば扱える.

特徴間の依存関係を一般化して考えると, 特徴列が作る高次元空間のなかで, 言語が占めうる部分空間はごく一部だと推測される. また, 言語がある程度漸進的に変化してきたことを考慮すると, ある祖語 P から言語 C への経路について, P, C はもちろん, 途中状態 $M_1, M_2, \ldots$ も, 実際に人間が話した言語であるからには自然でなければならない. だとすると, この部分空間は相互に孤立した部分空間の集まりではなく, 図 5 のように, 局所的には近傍と繋がりあった多様体のような構造を持っていると推測できる. この多様体上であれば言語として自然, そこから外れていれば不自然とみなせる.

多値特徴列からなる状態 x の自然さを判定するモデルは, エネルギーに基づく学習(LeCun et al., 2006)と同様の枠組みで作ることができる(Murawaki, 2015a). つまり, モデルのパラメータを θ とすると, 関数 $f(x;\theta) = d \in \mathbb{R}$ が, 自然な特徴の組み合わせに高い確率を, そうでない組み合わせに低い確率を返すようにするのが目標となる. そのために適切なモデルを設計したうえで, パラメータ θ を類型論データベースに登録された実在の諸言語から学習する問題として定式化できる.

関数 f は言語学で言うところの共時的な様態を学習していることになるが, 上述の議論の通り, 歴史的変化にも応用できる. すなわち, 言語 C およびその祖語 P はもちろん, 途中状態 $M_1, M_2, \ldots$ に対しても, f が高い確率を返すはずである. また, 進化はある程度漸進的であり, M_t は M_{t-1} の近傍に位置するはずである. この 2 つの条件を考慮することで, 言語変化の経路がかなりの程度絞り込めるのではないかと見込んでいる.

5.3 言語接触の影響

類型論的特徴の場合，語彙的特徴以上に系統樹の仮定が疑われる．実際，言語類型論では，地域的特徴とよばれる接触の影響に多くの議論が費やされてきた．特徴の変化のモデル化と並行して，系統樹の妥当性も検証していく必要がある．

系統関係の親疎にかかわらず，多くの地域的特徴を共有する言語群は言語連合とよばれる．特に有名なのがバルカン言語連合であり，ギリシア語，アルバニア語，東南スラヴ諸語など，印欧語族のなかでも比較的系統的に遠い言語が含まれている．統計モデルとしては，系統推定に言語連合を組み込んだベイズモデルが提案されている(Daumé III, 2009)．このモデルでは，個々の言語は(1)系統樹に沿った進化の結果と(2)言語連合からの生成の確率的混合としてモデル化されている．ただし，モデルの自由度を抑えて推論可能とするために，言語連合は時間不変なクラスタとしてモデル化されている点が不自然である．

接触に関わる現象で，より扱いやすいものとして，筆者はクレオール形成に着目している(Murawaki, 2016)．クレオールは複数の言語の影響下で成立したとみられる一群の言語で，その多くがヨーロッパによる植民地化の影響を受けた大西洋・インド洋沿岸に分布している．クレオール形成過程は論争の絶えない課題だが，有力な仮説によると，文法が極端に単純化したピジンがまず生まれ，その後子供がピジンを母語として獲得し，その過程で複雑な意思疎通が可能なほど文法が発達することによって成立するという．しかし，この際に起きる言語普遍的な構造再編がクレオールを特徴づけているという説と，語彙提供言語や基層言語などとよばれる言語の影響が強いという説が入り乱れている．

これらの説を踏まえると，クレオールは，(1)語彙提供言語 L，(2)基層言語 S，(3)構造再編 R という 3 種類の確率的混合としてモデル化できる．具体的には，各クレオールについて混合比 $\theta = (\theta_{\mathrm{L}}, \theta_{\mathrm{S}}, \theta_{\mathrm{R}})$ を導入する．この混合比にしたがって 3 種類のいずれかの特徴を確率的に選んだ結果，各クレオールが形成されたと仮定する．この混合比をデータから推定することで，上記の仮説を検証する．つまり，集団遺伝学における Admixture 解析(Pritchard et al., 2000)や，言語処理におけるトピックモデル LDA(Latent Dirichlet Allocation)(Blei et al., 2003)に似た混合モデルによってクレオール形成が分析できる．

このモデルをクレオールの類型論データベース(Michaelis et al., 2013)に適用したところ，クレオール形成において構造再編が無視できない影響を持っているという結果を得た．また，クレオールの持つ諸特徴から語彙提供言語や基層言語の影響を差し引いたものと日本語を比較したところ，日本語はクレオール的ではないことが確認できた．日本語系統論のなかには，日本語混成言語説との関係においてクレオールに注目する議論があるが，日本語の場合，近い過去にクレオール形成はなかったと推測できる．ただし，クレオール形成よりもより穏健な言語接触が日本語の形成に影響を及ぼした可能性はあり，その解明は今後の課題として残っている．

6. おわりに

本稿では，歴史言語学的課題に対する近年の統計的研究について，これまでの研究経緯を含めて簡単に紹介するとともに，今後の方向性を議論した．歴史言語学的課題への取り組みは，長年人手による手法が主流であった．人間は記号の離散的な操作や，一歩一歩積み上げるような論証は得意だが，一方で，年代のような連続値を含む問題や，不確実性が解候補の組合せ爆発を生む問題は苦手である．こうした問題については計算機を用いる統計的手法が適しており，分子生物学由来の統計的手法がこれまでにない成果を生んできた．

しかし，主に生物データ向けのソフトウェアパッケージを転用することで研究が進んできたため，生物に対応物が存在しない言語現象がモデル化から取り残される傾向にあるように筆者

は感じている．その一方で言語資源の組織的な整備が進んでおり，統計モデルによる仮説の検証が容易に行える環境が整いつつある．したがって，言語現象を丁寧に分析し，それに対応する統計モデルを提案していけば，大きな成果が期待できる．本稿をきっかけに，こうした課題に取り組む研究者が一人でも増えれば幸いである．

参 考 文 献

Bergsland, K. and Vogt, H. (1962). On the validity of glottochronology, *Current Anthropology*, **3**(2), 115–153.

Blei, D. M., Ng, A. Y. and Jordan, M. I. (2003). Latent Dirichlet allocation, *Journal of Machine Learning Research*, **3**, 993–1022.

Bouchard-Côté, A., Hall, D., Griffiths, T. L. and Klein, D. (2013). Automated reconstruction of ancient languages using probabilistic models of sound change, *Proceedings of the National Academy of Sciences*, **110**(11), 4224–4229.

Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., Alekseyenko, A. V., Drummond, A. J., Gray, R. D., Suchard, M. A. and Atkinson, Q. D. (2012). Mapping the origins and expansion of the Indo-European language family, *Science*, **337**(6097), 957–960.

Chang, W., Cathcart, C., Hall, D. and Garrett, A. (2015). Ancestry-constrained phylogenetic analysis supports the Indo-European steppe hypothesis, *Language*, **91**(1), 194–244.

Darwin, C. (1859). *The Origin of Species by Means of Natural Selection or, the Preservation of Favored Races in the Struggle for Life*, John Murray, London.

Daumé III, H. (2009). Non-parametric Bayesian areal linguistics, *Proceedings of Human Language Technologies*: *The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 593–601.

Daumé III, H. and Campbell, L. (2007). A Bayesian model for discovering typological implications, *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, 65–72.

Drummond, A. J. and Bouckaert, R. R. (2015). *Bayesian Evolutionary Analysis with BEAST*, Cambridge University Press, Cambridge.

Dunn, M., Terrill, A., Reesink, G., Foley, R. A. and Levinson, S. C. (2005). Structural phylogenetics and the reconstruction of ancient language history, *Science*, **309**(5743), 2072–2075.

Dunn, M., Greenhill, S. J., Levinson, S. C. and Gray, R. D. (2011). Evolved structure of language shows lineage-specific trends in word-order universals, *Nature*, **473**(7345), 79–82.

Felsenstein, J. (1981). Evolutionary trees from DNA sequences: A maximum likelihood approach, *Journal of Molecular Evolution*, **17**(6), 368–376.

Forster, P. and Renfrew, C. (eds.) (2006). *Phylogenetic Methods and the Prehistory of Languages*, McDonald Institute for Archaeological Research, Cambridge.

Gray, R. D. and Atkinson, Q. D. (2003). Language-tree divergence times support the Anatolian theory of Indo-European origin, *Nature*, **426**(6965), 435–439.

Gray, R. D. and Jordan, F. M. (2000). Language trees support the express-train sequence of Austronesian expansion, *Nature*, **405**(6790), 1052–1055.

Gray, R. D., Bryant, D. and Greenhill, S. J. (2010). On the shape and fabric of human history, *Philosophical Transactions of the Royal Society B*: *Biological Sciences*, **365**(1559), 3923–3933.

Greenberg, J. H. (1978). Diachrony, synchrony and language universals, *Universals of Human Language* (ed. Joseph H. Greenberg), Volume 1, Stanford University Press, Stanford, California.

Greenhill, S. J., Currie, T. E. and Gray, R. D. (2009). Does horizontal transmission invalidate cultural

phylogenies?, *Proceedings of the Royal Society B*: *Biological Sciences*, **276**(1665), 2299–2306.

Greenhill, S. J., Atkinson, Q. D., Meade, A. and Gray, R. D. (2010). The shape and tempo of language evolution, *Proceedings of the Royal Society B*: *Biological Sciences*, **277**(1693), 2443–2450.

Hammarström, H., Forkel, R., Haspelmath, M. and Bank, S. (eds.) (2016), *Glottolog 2.7*, Max Planck Institute for the Science of Human History, Jena.

Haspelmath, M., Dryer, M., Gil, D. and Comrie, B. (eds.) (2005). *The World Atlas of Language Structures*, Oxford University Press, Oxford.

服部四郎 (1999).『日本語の系統』, 岩波書店, 東京.

Heled, J. and Bouckaert, R. R. (2013). Looking for trees in the forest: Summary tree from posterior samples, *BMC Evolutionary Biology*, **13**(1), 1–11.

Huelsenbeck, J. P. and Ronquist, F. (2001). MRBAYES: Bayesian inference of phylogenetic trees, *Bioinformatics*, **17**(8), 754–755.

International HapMap Consortium (2005). A haplotype map of the human genome, *Nature*, **437**(7063), 1299–1320.

Kingman, J. F. C. (1993). *Poisson Processes*, Oxford University Press, Oxford.

LeCun, Y., Chopra, S., Hadsell, R., Ranzato, M. and Huang, F. J. (2006). A tutorial on energy-based learning, *Predicting Structured Data* (eds. Gökhan Bakır, et al.), MIT Press, Cambridge, Massachusetts.

Lee, S. and Hasegawa, T. (2011). Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages, *Proceedings of the Royal Society B*: *Biological Sciences*, **278**(1725), 3662–3669.

Lizana, L., Mitarai, N., Sneppen, K. and Nakanishi, H. (2011). Modeling the spatial dynamics of culture spreading in the presence of cultural strongholds, *Physical Review E*, **83**(6), 066116.

Longobardi, G. and Guardiano, C. (2009). Evidence for syntax as a signal of historical relatedness, *Lingua*, **119**(11), 1679–1706.

松本克己 (2007).『世界言語のなかの日本語: 日本語系統論の新たな地平』, 三省堂, 東京.

Maurits, L. and Griffiths, T. L. (2014). Tracing the roots of syntax with Bayesian phylogenetics, *Proceedings of the National Academy of Sciences*, **111**(37), 13576–13581.

Michaelis, S. M., Maurer, P., Haspelmath, M. and Huber, M. (eds.) (2013). *APiCS Online*, Max Planck Institute for Evolutionary Anthropology, Leipzig.

Murawaki, Y. (2015a). 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*, 324–334.

Murawaki, Y. (2015b). Spatial structure of evolutionary models of dialects in contact, *PLoS ONE*, **10**(7), 1–15.

Murawaki, Y. (2016). Statistical modeling of creole genesis, *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics*: *Human Language Technologies*.

Nicholls, G. K. and Gray, R. D. (2008). Dated ancestral trees from binary trait data and their application to the diversification of languages, *Journal of the Royal Statistical Society*: *Series B* (*Statistical Methodology*), **70**(3), 545–566.

Nichols, J. (1992). *Linguistic Diversity in Space and Time*, University of Chicago Press, Chicago and London.

Nichols, J. and Warnow, T. (2008). Tutorial on computational linguistic phylogeny, *Language and Linguistics Compass*, **2**(5), 760–820.

Pellard, T. (2009). Ōgami: Éléments de description d'un parler du sud des Ryūkyū, Ph.D. Thesis, Ecole des Hautes Etudes en Sciences Sociales (EHESS) (in French).

Pereltsvaig, A. and Lewis, M. W. (2015). *The Indo-European Controversy*: *Facts and Fallacies in Historical Linguistics*, Cambridge University Press, Cambridge.

Pritchard, J. K., Stephens, M. and Donnelly, P. (2000). Inference of population structure using multilocus genotype data, *Genetics*, **155**(2), 945–959.

Roberts, S. and Winters, J. (2013). Linguistic diversity and traffic accidents: Lessons from statistical studies of cultural traits, *PLoS ONE*, **8**(8), 1–13.

Schleicher, A. (1853). Die ersten Spaltungen des indogermanischen Urvolkes, *Allgemeine Monatsschrift für Wissenschaft und Literatur*, **3**, 786–787 (in German).

Schmidt, J. (1872). *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*, Hermann Böhlau, Weimar (in German).

Swadesh, M. (1952). Lexicostatistic dating of prehistoric ethnic contacts, *Proceedings of American Philosophical Society*, **96**, 452–463.

Teh, Y. W., Daumé III, H. and Roy, D. (2008). Bayesian agglomerative clustering with coalescents, *Advances in Neural Information Processing Systems 20*, 1473–1480.

東条 操 (1927).『国語の方言区画』, 東京堂出版, 東京.

Trudgill, P. (1974). Linguistic change and diffusion: Description and explanation in sociolinguistic dialect geography, *Language in Society*, **3**, 215–246.

角田太作 (1991).『世界の言語と日本語』, くろしお出版, 東京.

Tsunoda, T., Ueda, S. and Itoh, Y. (1995). Adpositions in word-order typology, *Linguistics*, **33**(4), 741–762.

Vovin, A. (2010). *Koreo-Japonica*, University of Hawai'i Press, Honolulu.

柳田國男 (1930).『蝸牛考』, 刀江書院, 東京.

Zuckerkandl, E. and Pauling, L. (1965). Evolutionary divergence and convergence in proteins, *Evolving Genes and Proteins*, **97**, 97–166.

Statistical Approaches to Language Change and Linguistic Phylogenies

Yugo Murawaki

Graduate School of Informatics, Kyoto University

Since around the turn of the twenty-first century, there has been a growing trend to employ computer-intensive statistical methods to answer historical linguistic questions, such as language change and phylogenies of extant and documented languages. Although these questions have traditionally been addressed manually by linguists, manual analysis has limitations. Because human inference is based on logic, humans are unable to estimate continuous values (e.g., dating the common ancestor of extant languages). They are also bad at inherent uncertainty because it leads to a combinatorial explosion. Computational statistics provides powerful ways to solve these problems.

The current trend can be characterized by the fact that key results have been achieved with statistical methods originally developed in the field of molecular biology. Although historical linguistics itself has a record of adopting statistical models, the new statistical techniques have been developed largely independently of historical linguistics. Therefore their scientific foundations have yet to be fully understood by linguistic communities. We also observe that since most recent statistical studies on linguistic questions depend on ready-to-use software packages that are designed to address biological questions, linguistic phenomena that lack exact counterparts in biology tend to be left untouched.

In light of this, we first overview the new statistical models while relating them to the research history of historical linguistics. After reviewing the concept of evolution, the comparative method that exploits regular sound changes, and ill-fated glottochronology, we explain the essence of recently developed Bayesian phylogenetic models.

Since most phylogenetic models use lexical traits, they can be applied only if the group of languages in question has a sufficient number of shared lexical traits. Unfortunately, this is not the case in Japanese, and we have no choice but to seek for different kinds of traits. Later in this paper, we describe novel approaches based on typological traits, which we believe have the potential to trace the origin of the Japanese language.

Key words: Linguistic phylogeny, historical linguistics, linguistic typology, Bayesian statistics.