特　集

言語系統解明のための計算的取り組み

Computational Approaches to Understanding Linguistic Phylogenies

村脇 有吾　京都大学
Yugo Murawaki　Kyoto University
murawaki@i.kyoto-u.ac.jp, http://murawaki.org/

keywords: linguistic phylogeny, cognates, linguistic typology, linguistic universals

1. は じ め に

日本語の起源はどうなっているのか？　別の言い方をすれば，日本語と他の言語は歴史的にどのような関係にあるのか？　この問題に対する現代科学の挑戦は優に 100 年を超えるにもかかわらず，いまだに確実な答えは得られていない [Vovin 10]．本稿もこれに対する直接の答えを用意できているわけではないが，この問題への取り組みとして，計算機を用いた統計的手法が前途有望であることを示したい．

「人工知能と歴史」というお題で言語史を取り上げることに対して意外に思われるかもしれない．実際，言語史の研究は，長らく言語学者が人手で行う分野であった（2 章）．しかし，2000 年頃から，計算機を用いた統計的手法によって新たな成果が次々と報告されている（3 章）．計算的取り組みが一定の成功を収めている理由の一つとして，人間が苦手とする推論のなかに，計算機であれば扱えるものがあることがあげられる．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を得意とする反面，年代のような数量を含む問題や，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．近年の研究は，こうした問題であっても，計算機であれば（近似的に）解ける場合があることを実証してきた．

人工知能といえば昨今は過剰な期待が集まりがちだが，実際のところ，計算的取り組みは，データ，手法の両面で伝統的な言語学なしでは成り立たない．統計的手法を適用するためにはある程度まとまったデータが不可欠だが，そうしたデータはもっぱら言語学者が人手で作成している．手法面では，現在主流の統計的手法は，手掛かりとして語彙を用いるという点で言語学の伝統を引き継いでいる．まさにその語彙的手掛かりが欠けていることから，この手法によって日本語とその他の言語との関係を明らかにできそうにない．

そこで，本稿では，語彙に代わる手掛かりとして言語類型論の特徴に着目する（4 章）．類型論は世界中の言語を特徴（類型）によって分類する分野で，その研究は語彙に基づく手法と同じくらい古くから見られる．しかし，類型論的特徴は言語群の関係を決定するための手掛かりとはみなされてこなかった．その大きな理由は，語彙と比べて手掛かりとして不確実性が高く，人手による論証が難しかったことだと筆者は考えている．そして，計算機を用いた統計的手法であれば，類型論的特徴が歴史的にどのような振舞いをするか明らかにし，系統推定に応用できるのではないかと見込んでいる（5 章）．まだまた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．

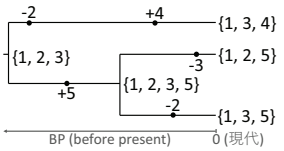


図 1　系統樹の例

2. 歴史比較言語学

言語史の解明に取り組む言語学の下位分野は歴史言語学と呼ばれる．言語史を解明するための最初の手続きは，文献を可能な限りさかのぼり，記録上最古の状態を明らかにすることである．しかし，文字記録には限りがあり，有史以前 *1 の状態を知るには別の手法が必要となる．そこで歴史言語学が採用した手法は，複数の言語を比較し，それらが共通祖先（祖語）から生じたことを立証するというものである．そのため，この分野を特に歴史比較言語学とも呼ぶ．

複数の言語の歴史的関係は通常木構造により要約され，これを系統樹と呼ぶ [Schleicher 53]．例を図 1 に示す．ここで，右端の葉が子孫（現代語や十分な記録の残る古代語），左端の根が共通祖語であり，その間の経路が歴史的変遷を表す．系統樹を特徴付けるのは分岐である．二つの言語は，分岐前は完全に同一であり，分岐後は独立に進化する．進化とは，親から子へと途切れなく状態が引き継がれるが，完全に複製されるのではなく，しだい

*1　本稿では先史時代を含めて広く歴史と呼ぶ．

に変化が蓄積する現象を指す．図 1 では言語の状態を 2 値特徴の集合で表している．例えば，共通祖語の状態は特徴 1,2,3 で表される．2 値特徴の変化は誕生（例えば図 1 の枝上の +4）と死亡（−2）のいずれかである．

いま，葉の言語群が与えられたとき，系統樹を復元したいとする．言語の場合，日本語がそうであるように，一般にはほかの言語との関係は不明である．注目する言語群が祖語を共有することを証明するには，それらが祖語から特徴を引き継いでいることを示せばよい．図 1 の場合，すべての葉が特徴 1 を引き継いでいる．同様に，分岐の前後関係を示すには，変化の共有の有無を示せばよい．例えば，+5 という変化は，下二つの葉のみが経験していることから，上の葉との分岐のほうが古いと推定できる．ここでは簡単のために数個の特徴で例示したが，実際には百のオーダの特徴を用いる．

語彙は系統推定に適した特性をもっている．例えば，ある特徴が系統樹上で複数回誕生する可能性が排除できる．語は恣意的な記号であり，DOG という意味と「いぬ」という音の結び付きに必然性はないことから，DOG を意味する「いぬ」という語が独立に複数回発生する可能性は極めて低い．したがって，語彙的特徴を共有する言語群は原則的に共通祖語をもつといえる．ただし，系統樹は分岐後の独立進化を仮定するが，実際には言語同士の接触による借用が起こり得る．また，語そのものは引き継いでも，語形は時間とともに変化するため，ある言語対がもつ語が同一特徴（同源語）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も除外しなければならない．

歴史比較言語学では，同源語の判定に音法則と呼ばれる現象を用いる．音法則は 19 世紀後半に青年文法学派と呼ばれる言語学者の集団が確立したもので，歴史的な音変化は例外なく規則的に起き，結果として，ある言語対がもつ同源語の語形には規則的な音対応が見られる（借用語の場合は対応が乱れる）．例えば，古代日本語と与那国方言の語形を比較すると，*wata* と *bata*（綿），*wodori* と *budui*（踊り）のように，語頭において *w* と *b* の規則的な対応が見られる．同様にして，*o* と *u* の対応，与那国方言における母音間の *r* の脱落も規則的である．したがって，これらの語は同源語であり，両言語が祖語を共有することが示される．

長年にわたる歴史比較言語学の研究は，世界のさまざまな言語群の系統関係を解明してきた．インド，イランからヨーロッパにかけて広がるインド・ヨーロッパ（印欧）語族が代表的な例で，ほかにも，台湾から島嶼部東南アジア，太平洋に広範囲に分布するオーストロネシア語族や，オーストラリアの大半の地域で話されていたパマ・ニュンガン語族などがあげられる．一方，日本語の場合，近隣の諸言語，なかでも朝鮮語や，ツングース諸語，モンゴル諸語など（いわゆるアルタイ諸語）との関係を立証しようという努力が長く続けられてきたが，十分な数の同源語が見つかっていない [Vovin 10]．また，既知の語族をさらにまとめる試みとして，ノストラティック大語族やアメリンド大語族などが提案されているが，広い支持は得られていない．

3. 語彙に基づく統計的手法

伝統的な比較手法は言語間の系統関係の確立や，系統樹の復元を可能とするが，祖語がいつ話されていたかは推定できない．人間は年代のような数量を推論するのは苦手であり，統計的手法の出番となる．

祖語の年代推定方法として，言語年代学と呼ばれる統計的手法が 1940 年代末から 50 年代を中心に研究された [Swadesh 52]．言語年代学は，考古学における放射性炭素年代測定に触発されたもので，語彙的特徴が時間とともに一定割合で失われるという仮定のもと，言語対からそれらの共通祖語の年代を推定する．この先駆的な試みは言語学者の間で評判が悪く，激しい批判にさらされた [Bergsland 62]．結果として，言語データに対する統計的手法の研究は長く停滞してしまった．しかし，この間に収集された語彙のデータベースが現在の統計的研究にも利用されており，研究史上の重要性は強調しておきたい．

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統計的手法は，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 *2．分子生物学においても，当初は素朴な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 01]．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あげられる．

[Gray 03] はベイズ系統モデルを印欧祖語の年代推定問題に適用し，欧米において大きな話題となった *3．欧米人にとって，彼らの祖先である印欧祖語の話者がいつどこにいたかは関心の的であり，これまでに数多くの仮説が提案されてきた．[Gray 03] はそのなかでも有力な 2 つの仮説——クルガン仮説とアナトリア仮説——にしぼって議論する．クルガン仮説は，考古学的証拠をもとに，5,000–6,000 年前の黒海周辺のステップを故地とし，遊牧民の軍事的征服により印欧語族が広がったとする．対する

*2 分子生物学において統計的手法の基盤となる分子時計仮説 [Zuckerkandl 65] は，実は言語年代学よりも後発である．

*3 [Gray 03] 以前に生物学由来の統計的系統モデルを言語データに適用した事例として [Gray 00] がある．この研究は，オーストロネシア語族の拡散過程に関する 2 つの仮説——急行列車モデルと生い茂った土手モデル——を統計的系統モデルにより検証し，前者を支持する．ただし，この研究で用いた系統モデルはベイズ以前のクラスタリング手法であり，年代推定は行っていない．急行列車モデルを支持する根拠は，得られた系統樹のトポロジーである．

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）を故地とし，農耕と言語の同時拡散を想定する．[Gray 03] は，ベイズ系統モデルから得られた祖語の年代をもとに，アナトリア仮説を支持する．Gray らはその後も一貫してアナトリア仮説を支持し，2012 年には年代と同時に故地も推定することで同仮説を補強している [Bouckaert 12].

一方，印欧語族の研究者の間では，アナトリア仮説はほとんど支持されていない [Pereltsvaig 15]．言語学者の間では統計モデルに対する懐疑論が根強いが，Gray らによるアナトリア仮説の推進がその一因となっている．ただし，ベイズ系統モデルはデータに関して多くの仮定をおいているが，その一部を見直すことでクルガン仮説寄りの年代が推定されたという報告 [Chang 15] もあり，この論争はしばらく続きそうである．

紙面の都合でベイズ系統モデルの説明は省略するが（詳細は [Drummond 15] を参照されたい），類型論的特徴との比較の都合上，語彙的特徴について簡単に見ておく．ベイズ系統モデルが用いる語彙的特徴は基礎語彙と呼ばれ，もともと言語年代学の研究で設計されたものである．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（100–200 項目）である．例えば，WATER，BIG，EYE などが該当する．語彙的特徴のデータベースを作るには，まず各言語から基礎語彙にあたる語を収集する．次に，言語間の比較を行い，伝統的な比較手法によって各項目の語を同源語に仕分ける．例えば，WATER を表す英語の *water*，ドイツ語の *Wasser* は同源語であり，フランス語の *eau*，イタリア語の *acqua* は別の同源語に分類される．各言語は各同源語をもつか否かという 2 値特徴で表現される．すべての基礎語彙についての結果を並べることで，各言語は 00101100$\cdots$ のような 2 値特徴の列 *4 として表現される．

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．実際，ある調査によると，英語の一般語彙のおよそ 50%が借用語だが，基礎語彙に限ると 6%にすぎない [Swadesh 52]．変化への抵抗性については，基礎語彙のなかでも項目によってばらつきがあると報告されている [Greenhill 10]．一部の項目は 1 万年規模の極端な保守性をもつという主張もある [Pagel 13]．ただし，この研究で用いられた語彙データは，言語学者の間で支持されていない大語族仮説に基づいている．一般には，言語年代学に基づく大雑把な推定により，語彙に基づく手法でさかのぼれるのは 6,000–7,000 年程度が限度だといわれる [Nichols 11].

語彙に基づく統計的手法では，日本語と他の言語との関係を明らかにすることはできない．日本語と他の言語との間でいまだに十分な数の同源語が特定できないことから，逆説的に，共通祖語の年代は相当さかのぼるのではないかと推測できる [服部 99].

4. 言 語 類 型 論

語彙にかわる手掛かりとして，筆者は言語類型論の特徴に着目している．類型論自体の歴史は古く，例えば，19 世紀の言語学者は，形態的特徴は循環的に変化すると考えた [Croft 02]．世界の言語は，孤立語，膠着語，屈折語に分類できる．孤立語は中国語やベトナム語のようにほとんど語形変化を起こさず，不変化詞が文法標識の役割を担う．不変化詞が内容語に従属的になり，接辞として内容語に規則的に連結するようになると膠着語と呼ばれる．日本語やトルコ語が該当する．さらに，接辞が融合し，文法範疇との対応が不明瞭になると，ラテン語やロシア語のように屈折語と呼ばれる．最後に英語のように屈折が摩耗することで孤立語に戻る，という仮説である．

古典的な類型論は特定の特徴のみに着目していたが，系統推定に用いるために，語彙の場合と同様に，複数の特徴を集め，各言語の状態を特徴の列で表現する．そうした特徴の例として，基本語順，助数詞の有無，声調の有無が挙げられる．類型論的特徴は一般には多値であり，例えば基本語順特徴は，SOV，SVO，VSO，VOS，OVS，OSV，優勢な語順なしの 7 種類の値のいずれかをとる．

類型論的特徴のデータ整備には言語学の高度な知識が必要となる．例えば，動詞のように自明に思える概念であっても，世界の言語のなかには悩ましい事例がある．かつては言語学者が個別にデータを収集していた [角田 91] が，現在は組織的収集の成果として World Atlas of Language Structures（WALS）がマックス・プランク進化人類学研究所から公開されている [Haspelmath 05]．2016 年現在，WALS は 2,679 の言語，192 の特徴を収録しており，計算的取り組みの研究基盤が整備されている *5.

系統推定における類型論的特徴の利点として，語彙的特徴とは異なり，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候補として，朝鮮語を含むアルタイ諸語が注目されてきたが，この根拠となったのは，「語頭に *r* 音が立たない」，「*have* 型の所有動詞をもたない」といった類型論上の類似であった [松本 07].

しかし，歴史比較言語学では，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．上述のアルタイ的とされた日本語の特徴についても，実は世界的にありふれており，アルタイ諸語に固有のものではないことが指摘されている [松本 07]．類型論的特徴を系統推定に用いる試みもいくつかあるが，語彙に基づく手法と比べて圧倒的に少ない [Tsunoda 95, Dunn 05, Longobardi 09].

*4 分子生物学で系統推定に主に塩基配列を用いるのに合わせて，言語データも列で表現するが，順番に意味はない．

*5 ただし，数個の特徴しか記述されていない言語が多く，言語と特徴の組のうち，85%以上が欠損値．

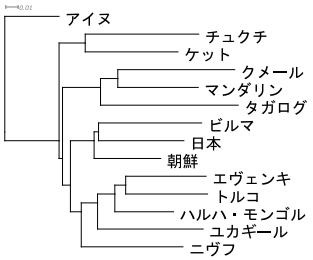


図 2 類型論的特徴に基づく日本語と近隣の言語の系統樹の例

類型論的特徴の大きな欠点は，ある言語対が特徴を共有していたとしても，その特徴を共通祖語から引き継いだとは限らないことである．形態的特徴の循環的変化の例が示すように，同じ変化が歴史上無関係に複数回起こり得るし，一度失われた特徴が復活する可能性もある．つまり，類型論的特徴は語彙的特徴よりも不確実性が高く，人手による論証になじまない．しかし，計算機を用いた統計的手法であれば扱える可能性は残っている．

5. 類型論に基づく統計的手法

上述の通り，類型論的特徴を使えば任意の言語対が比較できるため，適当なクラスタリング手法を用いることで系統樹相当のものをつくること自体は可能である．試みにつくった日本語と近隣の言語の系統樹の図 2 に示す [*6]．しかし，このような系統樹は何の説得力ももたない．類型論的特徴の振る舞いがモデルの仮定と合っているか不明だからである．類型論的特徴の性格については課題が山積みだが，本稿では，接触の影響，歴史的安定性，特徴間の依存関係の 3 点にしぼって議論する．

5·1 接触の影響

そもそも言語群が系統樹をなすという仮定は妥当だろうか？ 語彙的特徴の場合，系統樹の仮定に反する現象として借用語があったが，言語類型論の研究でも，地域的特徴と呼ばれる接触の影響に多くの議論が費やされてきた．

接触による類型論上の現象として言語連合が知られている．言語連合とは，同じ地域で話されており，系統上の親疎にかかわらず，多くの地域的特徴を共有している言語群を指す．有名なバルカン言語連合は，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からなり，属格と与格の融合や定冠詞の後置といった特徴を共有する．ほかにも，大陸部東南アジアでは，モン・ミエン語族，クラ・ダイ語族，オーストロアジア語族のように明確な系統関係のない言語群が，声調や音節構造などの特徴を共有している．

[Daumé III 09] は，言語連合を組み込んだベイズ系統モデルを提案している．一般に，接触を考慮しつつ系統推定を行うのは，モデルの自由度が高すぎるため難しい．[Daumé III 09] は，個々の言語を（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みなし，言語連合を時間不変なクラスタとして近似することで，自由度を抑えた推論を実現した．人手による既存の系統樹との比較により，言語連合を考慮するほうが良い系統樹が得られると報告している．また，系統樹と言語連合との混合比を見ると，類型論的特徴の種類によって，系統寄りのものと地域寄りのものがあることがわかる．この結果から推測すると，系統寄りの特徴に適切な重みをかければ，言語連合を明示的にモデル化せずとも高精度な系統推定ができるかもしれない．

より極端な接触現象であり，それゆえにモデル化しやすいものとして，筆者はクレオール形成を調査した [Murawaki 16]．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言語（ピジン）がまず生まれ，その後ピジン話者の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付けるという説と，語彙提供言語（社会文化的に優勢な言語で，クレオールの語彙の大半がこれに由来）や基層言語（社会的に劣勢な言語）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，クレオール形成は，(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，クレオールの各言語について混合比 $\theta = (\theta_{\mathrm{L}}, \theta_{\mathrm{S}}, \theta_{\mathrm{R}})$ を導入する．この混合比に従って 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，上記の仮説が検証できる．

このモデルをクレオールの類型論データベース [Michaelis 13] に適用したところ，混合比において構造再編が高い割合を占めた．また，クレオールのもつ諸特徴から語彙提供言語や基層言語の影響を差し引くことで，クレオールらしい特徴の値が抽出できる．これを日本語の特徴の値と比較したところ，日本語はクレオール的ではないという結果を得た．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 [川本 90] が，日本語の場合，少なくとも近い過去にクレオール形成はなかったと推測できる [*7]．もちろんクレオール形成よりも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，

*6 WALS の特徴のうち被覆率の高い 154 個を 1-of-K 法で 2 値化した．クラスタリングは距離に基づく近隣結合法 [Saitou 87].

*7 例えば，助数詞の欠如はクレオールらしい特徴だが，日本語は助数詞を用いる．しかし，上代日本語の段階では助数詞は発達途上であったと考えられている [Vovin 05].

その検証は今後の課題として残っている．

5·2 歴史的安定性

系統推定に用いる語彙的特徴は，語彙の中でも歴史的安定性の比較的高い基礎語彙であったのに対して，類型論のデータベースは，歴史的安定性を特に考慮せず，類型論の研究者達がそれぞれの興味に従って収集したものである．WALS における極端な例として，*tea* の特徴があげられる [Haspelmath 05]．これは，各言語が茶を表すのに，*cha* 系の語を用いるか，*te* 系の語を用いるか，あるいはその他かを分類したもので，主に 16 世紀以降の歴史的過程を反映するにすぎない．深い系統関係を解明するには，歴史的安定性の高い特徴が欠かせない．

[Greenhill 10] は，語彙的特徴から得られた年代つき系統樹に類型論的特徴を当てはめ，系統モデルの変化率パラメータを特徴ごとに求めた．系統樹として印欧語族とオーストロネシア語族の 2 種類を用いたところ，大半の類型論的特徴は語彙的特徴と同程度の変化率だが，一部の特徴は語彙よりも変化しにくいという結果を得た．ただし，接触の影響を考慮していないことに注意を要する．

何人かの言語学者は，明示的に系統推定を行うことなく，特徴の現在の分布から歴史的安定性を推測しようとしてきた [Nichols 92, Nichols 95, Parkvall 08, Wichmann 09]．[Dediu 13] はそうした手法を同一データを用いて比較している．この手法をもとに，いくつかの類型論的特徴は 1 万年のオーダの深い言語史を反映しているという大胆な主張もなされている [Nichols 92, 松本 07]．

現在の分布を用いる動機は，もしある特徴が安定的ならば，言語群内で同じ値が広く共有されるはずだというものである．[Nichols 92] は，あらかじめ指定された各言語群内で特徴量の最頻値を探し，次にその値を取らない言語の割合を数え，最後にその割合の言語群間の平均を求める．もし言語群が系統に基づくなら，この値は系統上（縦）の不安定性の指標とみなせる．同様にして，言語群を地域的に定義すると，この値は横の不安定性（接触による伝播のしにくさ）を表す．[Parkvall 08] はより複雑な計算式を用いて系統的一貫性 C_{FAM} と地域的一貫性 C_{ARE} を定義する．さらに，最終的な安定性指標を $S = C_{\mathrm{FAM}}/C_{\mathrm{ARE}}$ と定義するが，なぜ割り算を行うのか判然としない．[Wichmann 09] も同様の指標を提案するが，縦の安定性のみを考える．この種の指標は操作的に定義され，理論的な裏付けが乏しいという欠点がある．

実は，文化人類学においても，縦・横のいずれによって特徴が伝承されるかが論争の的になっている．[Towner 12] は，自己ロジスティックモデルによって両伝承形態を単一モデルに取り込み，それらの相対的な重みを推定する．彼らは極端な多様性で知られる北アメリカ西部の先住民社会の分析にこの手法を適用し，ほとんどの特徴の説明に両伝承形態が必要だと主張する．このモデルをほぼそのまま用いることで，類型論的特徴についても伝承形態の定量的評価が可能である [Yamauchi 16]．

5·3 特徴間の依存関係

ここまでは，各特徴が独立に変化すると仮定してきた．語彙的特徴の場合，同じ意味を表す同源語同士は競合関係にあり，独立性の仮定に反すると思われるが，シミュレーション実験により，この仮定が系統推定に大きな影響を与えないことが示されている [Atkinson 05]．異なる意味の特徴同士であれば，影響はさらに軽微なはずである．では類型論的特徴の場合はどうだろうか？

実のところ，類型論的特徴の間には複雑な依存関係があり，それこそが類型論研究の焦点の一つであり続けている．[Greenberg 63] は世界の言語で普遍的に成り立つ法則をいくつか提示した．例えば，前置詞を用いる言語では属格はほとんど常に名詞に後続するが，後置詞を用いる言語では反対に先行する．このような依存関係を無視して系統推論を行うと，言語として不自然な祖語を推定するおそれがあり，ひいては系統樹への悪影響も懸念される．

そこで，筆者は，類型論的特徴を直接扱うのではなく，依存関係を反映した潜在表現に変換したうえで系統推定を行うことを提案した [Murawaki 15]．具体的な表現として，自己符号化器による連続空間表現（実数列）を採用した．しかし，離散的特徴と異なり，連続空間表現上では変化の方向性をモデル化しづらいという欠点がある．また，WALS を扱ううえで避けて通れない欠損値に対する頑健性も課題として残っている．

ところで，そもそもなぜこのような依存関係があるのだろうか？ これに対する説明には大きく 2 種類ある．機能的説明は，特徴の値のある組み合わせが認知的に処理しやすいといったように，人間の生得的な言語能力に理由を求める．[Baker 02] は，生成文法の立場から，特徴の値の組み合わせを決定する潜在的な 2 値パラメータ群の存在を主張する．しかし,「普遍」規則にしばしばまれな例外が発見されることから，そのような決定的なパラメータの存在は首肯しがたい [Evans 09]．もう一つの通時的説明は，一般性のある歴史的変化が原因だと考える [Heine 07]．例えば，日本語の格助詞「へ」は名詞「辺」が文法化したものだが，もしこのような名詞から後置詞への歴史的変化に一般性があれば，後置詞と属格前置との関係が説明できる．この議論をもう少し整理すれば，系統推定にも応用できるかもしれない．

6. ま と め

本稿では，計算機を用いた統計的手法により言語系統の解明を目指す研究を紹介した．日本語系統論を念頭に置き，最後の希望ともいえる言語類型論の特徴を中心に議論した．現状ではまだ日本語系統論についていえることはほとんどないが，意外な実験結果の例として，類型

論的特徴の歴史的安定性を考慮して日本語と他の現代語を比較すると，朝鮮語が通常期待されるほど似ていない一方，チベット・ビルマ系言語が上位に来ることがあげられる [Murawaki 15].

言語は人間集団を特徴付ける主要な要素だが，人類史を解明するには，集団遺伝学や考古学などの諸分野の知見との整合性も求められる．とりわけ遺伝子データは，そもそも量，質ともに言語データを圧倒しているうえに，古代 DNA の解析の進展により，古代の状態の直接的復元すら実現しつつある [Reich 10]．遺伝子データにひも付けた言語の系統推定も今後の方向性として有望だろう．

◇ 参 考 文 献 ◇

[Atkinson 05] Atkinson, Q., Nicholls, G., Welch, D., and Gray, R.: From Words to Dates: Water into Wine, Mathemagic or Phylogenetic Inference?, *Transactions of the Philological Society*, Vol. 103, No. 2, pp. 193–219 (2005)

[Baker 02] Baker, M. C.: *The Atoms of Language: The Mind's Hidden Rules of Grammar*, Basic Books (2002)

[Bergsland 62] Bergsland, K. and Vogt, H.: On the Validity of Glottochronology, *Current Anthropology*, Vol. 3, No. 2, pp. 115–153 (1962)

[Bouckaert 12] Bouckaert, R., Lemey, P., Dunn, M., Greenhill, S. J., et al.: Mapping the Origins and Expansion of the Indo-European Language Family, *Science*, Vol. 337, No. 6097, pp. 957–960 (2012)

[Chang 15] Chang, W., Cathcart, C., Hall, D., and Garrett, A.: Ancestry-constrained Phylogenetic Analysis Supports the Indo-European Steppe Hypothesis, *Language*, Vol. 91, No. 1, pp. 194–244 (2015)

[Croft 02] Croft, W.: *Typology and Universals*, Cambridge University Press (2002)

[Daumé III 09] Daumé III, H.: Non-Parametric Bayesian Areal Linguistics, in *Proc. of NAACL-HLT*, pp. 593–601 (2009)

[Dediu 13] Dediu, D. and Cysouw, M.: Some Structural Aspects of Language Are More Stable than Others: A Comparison of Seven Methods, *PLoS ONE*, Vol. 8, No. 1, pp. 1–20 (2013)

[Drummond 15] Drummond, A. J. and Bouckaert, R. R.: *Bayesian Evolutionary Analysis with BEAST*, Cambridge University Press (2015)

[Dunn 05] Dunn, M., Terrill, A., Reesink, G., Foley, R. A., and Levinson, S. C.: Structural Phylogenetics and the Reconstruction of Ancient Language History, *Science*, Vol. 309, No. 5743, pp. 2072–2075 (2005)

[Evans 09] Evans, N. and Levinson, S. C.: The Myth of Language Universals: Language Diversity and its Importance for Cognitive Science, *Behavioral and Brain Sciences*, Vol. 32, pp. 429–492 (2009)

[Gray 00] Gray, R. D. and Jordan, F. M.: Language Trees Support the Express-train Sequence of Austronesian Expansion, *Nature*, Vol. 405, No. 6790, pp. 1052–1055 (2000)

[Gray 03] Gray, R. D. and Atkinson, Q. D.: Language-tree Divergence Times Support the Anatolian Theory of Indo-European Origin, *Nature*, Vol. 426, No. 6965, pp. 435–439 (2003)

[Greenberg 63] Greenberg, J. H. ed.: *Universals of Language*, MIT Press (1963)

[Greenhill 10] Greenhill, S. J., Atkinson, Q. D., Meade, A., and Gray, R. D.: The Shape and Tempo of Language Evolution, *Proceedings of the Royal Society B: Biological Sciences*, Vol. 277, No. 1693, pp. 2443–2450 (2010)

[Haspelmath 05] Haspelmath, M., Dryer, M., Gil, D., and Comrie, B. eds.: *The World Atlas of Language Structures*, Oxford University Press (2005)

[服部 99] 服部 四郎：日本語の系統, 岩波書店 (1999)

[Heine 07] Heine, B. and Kuteva, T.: *The Genesis of Grammar: A Reconstruction*, Oxford University Press (2007)

[Huelsenbeck 01] Huelsenbeck, J. P. and Ronquist, F.: MRBAYES: Bayesian Inference of Phylogenetic Trees, *Bioinformatics*, Vol. 17, No. 8, pp. 754–755 (2001)

[川本 90] 川本 崇雄：ピジン・クレオール化と日本語の形成, 崎山 理（編）, 日本語の形成, pp. 130–168, 三省堂 (1990)

[Longobardi 09] Longobardi, G. and Guardiano, C.: Evidence for Syntax as a Signal of Historical Relatedness, *Lingua*, Vol. 119, No. 11, pp. 1679–1706 (2009)

[Michaelis 13] Michaelis, S. M., Maurer, P., Haspelmath, M., and Huber, M. eds.: *APiCS Online*, Max Planck Institute for Evolutionary Anthropology (2013)

[Murawaki 15] Murawaki, Y.: Continuous Space Representations of Linguistic Typology and their Application to Phylogenetic Inference, in *Proc. of NAACL-HLT*, pp. 324–334 (2015)

[Murawaki 16] Murawaki, Y.: Statistical Modeling of Creole Genesis, in *Proc. of NAACL-HLT* (2016)

[Nichols 92] Nichols, J.: *Linguistic Diversity in Space and Time*, University of Chicago Press (1992)

[Nichols 95] Nichols, J.: Diachronically Stable Structural Features, in Andersen, H. ed., *Historical Linguistics 1993. Selected Papers from the 11th International Conference on Historical Linguistics, Los Angeles 16–20 August 1993*, John Benjamins Publishing Company (1995)

[Nichols 11] Nichols, J.: Monogenesis or Polygenesis: A Single Ancestral Language for All Humanity?, in Tallerman, M. and Gibson, K. R. eds., *The Oxford Handbook of Language Evolution*, pp. 558–572, Oxford University Press (2011)

[Pagel 13] Pagel, M., Atkinson, Q. D., Calude, A. S., and Meade, A.: Ultraconserved Words Point to Deep Language Ancestry across Eurasia, *PNAS*, Vol. 110, No. 21, pp. 8471–8476 (2013)

[Parkvall 08] Parkvall, M.: Which Parts of Language are the Most Stable?, *STUF-Language Typology and Universals Sprachtypologie und Universalienforschung*, Vol. 61, No. 3, pp. 234–250 (2008)

[Pereltsvaig 15] Pereltsvaig, A. and Lewis, M. W.: *The Indo-European Controversy: Facts and Fallacies in Historical Linguistics*, Cambridge University Press (2015)

[Reich 10] Reich, D., Green, R. E., Kircher, M., Krause, J., et al.: Genetic History of an Archaic Hominin Group from Denisova Cave in Siberia, *Nature*, Vol. 468, No. 7327, pp. 1053–1060 (2010)

[Saitou 87] Saitou, N. and Nei, M.: The Neighbor-joining Method: A New Method for Reconstructing Phylogenetic Trees, *Molecular Biology and Evolution*, Vol. 4, No. 4, pp. 406–425 (1987)

[Schleicher 53] Schleicher, A.: Die ersten Spaltungen des indogermanischen Urvolkes, *Allgemeine Monatsschrift für Wissenschaft und Literatur*, Vol. 3, pp. 786–787 (1853), (in German)

[Swadesh 52] Swadesh, M.: Lexicostatistic Dating of Prehistoric Ethnic Contacts, *Proceedings of American Philosophical Society*, Vol. 96, pp. 452–463 (1952)

[Towner 12] Towner, M. C., Grote, M. N., Venti, J., and Mulder, M. B.: Cultural Macroevolution on Neighbor Graphs: Vertical and Horizontal Transmission among Western North American Indian Societies, *Human Nature*, Vol. 23, No. 3, pp. 283–305 (2012)

[角田 91] 角田 太作：世界の言語と日本語, くろしお出版 (1991)

[Tsunoda 95] Tsunoda, T., Ueda, S., and Itoh, Y.: Adpositions in Word-order Typology, *Linguistics*, Vol. 33, No. 4, pp. 741–762 (1995)

[Vovin 05] Vovin, A.: *A Descriptive and Comparative Grammar of Western Old Japanese, Part 1*, Global Oriental (2005)

[Vovin 10] Vovin, A.: *Koreo-Japonica*, University of Hawai'i Press (2010)

[Wichmann 09] Wichmann, S. and Holman, E. W.: *Temporal Stability of Linguistic Typological Features*, Lincom Europa (2009)

[Yamauchi 16] Yamauchi, K. and Murawaki, Y.: Contrasting Vertical and Horizontal Transmission of Typological Features, in *Proc. of COLING* (2016), (to appear)

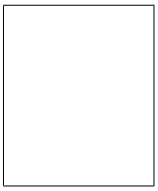
[Zuckerkandl 65] Zuckerkandl, E. and Pauling, L.: Evolutionary Divergence and Convergence in Proteins, *Evolving Genes and Proteins*, Vol. 97, pp. 97–166 (1965)

[松本 07] 松本 克己：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 (2007)

〔担当委員：××○○〕

19YY 年 MM 月 DD 日　受理

著　者　紹　介



村脇　有吾

2011 年京都大学大学院情報学研究科博士後期課程修了，博士（情報学）．同年京都大学学術情報メディアセンター特定助教，2013 年九州大学大学院システム情報科学研究院助教，2016 年京都大学大学院情報学研究科特定助教，同年助教，現在にいたる．テキスト解析および計算言語学に関する研究に従事．言語処理学会，情報処理学会各会員．