方言群の語彙は系統樹をなすか

村脇 有吾
九州大学大学院システム情報科学研究院
murawaki@ait.kyushu-u.ac.jp

概 要

計算モデルによる系統推定が自然言語に盛んに適用されるようになっている。この取り組みに対する根本的な疑問は、対象に対して系統モデルが本当に適切かである。恒常的に接触のある方言群の語彙については、系統モデルが適さない可能性が従来から指摘されている。そこで、本稿は、系統ではなく接触によって方言群が語彙を共有するモデルを提案する。接触モデルを用いてシミュレーションを行い、得られた人工データを系統モデルに解釈させる。その結果、現実の方言群から推定された系統樹に見られる不自然な振る舞いが再現できた。これは、現実の方言群がシミュレーションの想定にしたがって形成された可能性および接触モデルの有用性を示唆する。

1 はじめに

進化という点で、生物と言語の類似性は早くから認識されていた。ダーウィンの『種の起源』[10] とアウグスト・シュライヒャーによるインド・ヨーロッパ (印欧) 語族の系統樹 [37] は、いずれも 19 世紀半ばに発表されている。この問題に対する計算的取り組みは生物において先行し [14]、その成果が言語に応用されてきた。さらに、言語に限らず、写本 [3] や民話 [36]、さらには物質文化 [42]、宗教 [28] のような文化要素にまで適用先は拡大している。言語においては、例えば、印欧語族 [15, 4] やオーストロネシア語族 [17]、バントゥー諸語 [19, 9] の発達過程の解明に用いられている。

系統モデルの仮定するところでは、諸言語は共通の祖先から特徴を受け継ぐ。ただし、木にそって変化、すなわち特徴の誕生、消滅が起きる。ある時点で分岐が起きると、2 つの複製が作られ、その後 2 つの言語は別々に変化する。つまり、系統モデルは、2 つの言語に共通する特徴を共通祖語によって説明する。

系統モデルに対する根本的な疑問は、木が本当に対象に対して適切かである。言語学においては、系統モデルと競合する波モデルが、シュライヒャーによる系統モデルの発表のすぐ後に、ヨハネス・シュミットによって提示されている [21]。このモデルでは、新たな特徴は変化の中心から周辺に向かって波のように伝播する。つまり、波モデルは言語間に共通する特徴を接触によって説明する。計算的取り組みにおける系統モデルの隆盛とは裏腹に、言語学においては、依然として両モデルは補完関係にあると説明されている [6, 25]。

木を土台としつつ接触を扱う試みはこれまでもなされてきた。ここでは、接触による影響は水平伝播と呼ばれる。生物においては、原核生物の進化に遺伝子の水平伝播が重要な役割を果たすことが知られていたが、真核生物の進化にも影響していることが明らかになりつつある [1]。言語においては、例えば、印欧語族の系統樹におけるゲルマン語派の特異な振る舞いを系統上やや離れたにイタリア・ケルト語派との接触で説明する試みがある [35]。

計算的取り組みにおいては、系統モデルの水平伝播に対する頑健性が主張されてきた [31, 18, 8, 2]。ここでは、系統樹に水平伝播を追加した人工データを作成し、これに系統モデルを適用したところ、元の系統樹がある程度正しく復元されることが示されている。

木ではなく、水平伝播を含むグラフを観測データから推定する試み [29, 43] もあるが、一般に難しい。このグラフは木よりも自由度が高いためである。与えられた参照木に水平伝播の辺を追加するモデルも提案されている。当初の対象は原核生物だったが、言語にも適用されている [30]。

こうした計算的取り組みは、例え水平伝播を考慮したとしても、木を土台としてきたが、そもそもこの仮定は本当に適切だろうか。これまでの適用事例の多くは、印欧語族やオーストロネシア語族などの大語族であった。より規模の小さい方言群 [1] は異なる性質を持つ可能性がある。実際、恒常的な接触下にあり、高い相互理解性を維持する方言群 (方言鎖) を系統モデルで扱うのは難しいという指摘がたびたびなされている [29, 32, 16]。その一方で、日本語諸方言やアイヌ語諸方言のような方言群に対しても、系統モデルが実際に適用されてしまっている [26, 27]。推定された系統樹のうち、日本語諸方言については、他の言語的証拠との不整合が指摘されている [45]。しかし、計算的観点からみたモデルの妥当性はいまだ追究されていない。

系統モデルの方言群への適用にどのような問題があ

[1] 言語と方言の間に言語学的に本質的な差はない。本稿で言う方言とは単に系統的に近い言語群である。

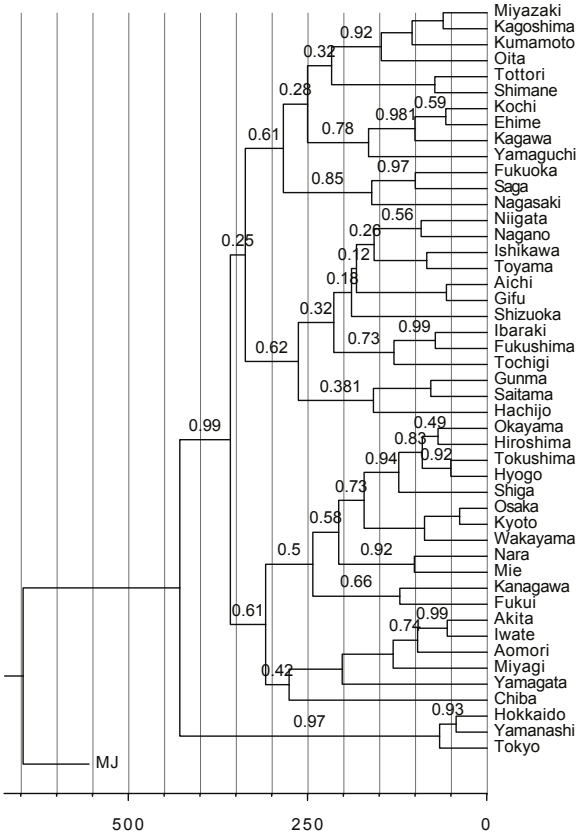


図 1: 日本語本土諸方言の系統樹の推定結果

るか見るために、例として日本語諸方言を取り上げる。図 1 は、日本語諸方言の系統樹のうち、琉球諸方言を除き、本土のみを抜粋したものである。横軸は現在から何年前かを示す。他の表記は 4 節で説明する。言語データは文献 [26] の補助資料を用いた。本土諸方言の系統推定は不安定であり、文献 [26] の系統樹とやや異なるが、全体としては同様の結果を得た。

この系統樹は様々な問題を抱えているが、ここで注目するのは、ここ 500 年、特にここ 250 年で起きた頻繁な分岐により本土諸方言が形成されたとされていることである。このような分岐を説明する大規模移住等の歴史事象を我々は知らない [2]。北海道を除くすべての地点にここ 500 年間日本語話者が定住していたことは確実である。本土諸方言が分岐により形成されたと考えるより、古くから (少なくともここ 500 年は) 別々に存在し、頻繁な接触を通じて現在の姿になったと考える方が自然である。

本稿では、系統モデルではなく、接触により方言群の説明を試みる。方言群が系統樹に基づく分岐によって出現するのではなく、最初からわかれており、恒常的な接触によって方言間で語彙が共有されるようになると仮定する。そのために、接触下にある方言群をネットワークで表現し、ネットワーク上を語彙が伝播するモデルを提案する。このモデルを用いて、様々なネットワークトポロジーを対象にシミュレーションを行う。そして、得られた結果を系統モデルがどのように解釈するかを調べ、実データの系統樹に見られる不自然な振る舞いの説明を試みる。

[2] もし系統樹上の分岐が歴史事象と対応づかないなら、推定された祖語の分岐年代に基づく日本語起源論は意味をなさない。

2 関連研究

語彙や文法的特徴などの言語的特徴の地理的分布は、これまでも調査が行われてきた [34]。こうした調査で得られるのはある時点でのスナップショットであり、歴史的変化は直接はわからない。しかし、いくつかの特異な分布については過去の状態が推測できる。例えば、ある特徴が地理的に離れた複数の地点で観測された場合、かつてはより広く分布していたと考えられる。この現象は、日本では、柳田國男の『蝸牛考』[50] により方言周圏論として知られている。

言語の動的な変化は方言平準化 (dialect levelling) の研究が調査してきた。ここでは、接触により方言間の差異が消滅していく過程が追跡されている [24, 25]。ただし、方言平準化の主な研究対象は音声的な変異である。系統モデルに使われる基礎語彙については、一般に、非常に近い方言間で借用関係を同定するのが困難である。また、時間の面でも、研究者が追跡調査したり、世代間の差が観測できるのは数十年のオーダにすぎない。基礎語彙は変化が起きにくいと仮定されており、このような短期間での変化は期待できない。

長期的かつ系統モデルに反した言語の変化については、言語連合が知られている [7, 11]。これは、接触下にある言語群が、系統関係のあるなしに関わらず、似た特徴を持つようになる現象を指す。ただし、言語連合の研究における関心は、語順などの類型論的特徴である。本稿が対象とするような方言群はもとから互いに似通っており、類型論的特徴にほとんど差がない。

進化において接触を扱う分野には、進化ゲームや、それに空間構造を導入した進化グラフがある [33]。これらの分野では、人口集団中の各個体がある時点でどの戦略を選ぶかをモデル化する。そのために、あらかじめ決められた数 (典型的には 2 つ) の戦略と、それに対する利得行列、さらに進化グラフの場合にはネットワークトポロジーを与える。これらの分野での主な関心は、動態を特徴づけるパラメータ (例えば、特徴が人口集団全体に広がる確率) に向けられている。本稿が扱う方言群の語彙はゲームの戦略とみなせる。ただし、次々に生まれる新語を扱いたいので、その異なり数は事前に決められない。

ネットワーク上を特徴が伝播する現象は口コミ (word of mouth) と呼ばれており、語 (word) そのものの伝播を研究対象としない手はない。ただし、従来研究は、マーケティングや世論形成を対象としてきた [23, 44, 12]。このような背景から、主な関心は特徴の広がりの最大化に向けられている。これまでに様々なシミュレーション用のモデルが提案されているが、語彙の伝播に適用するには不適切な仮定が少なくない。例えば、一度生まれた特徴が拡散する過程を考えるのみで、消滅をモデル化しない場合が多い。これに対して、語彙の場合は消滅も重要な現象である。また、一つの特徴のみを対象とする場合が多い。一方、語彙については、同じ概念を表す語を複数保持するのは経済的ではなく、それらは競合関係にある。ある語の消滅は、同じ意味を表す別の語の隆盛により説明するのが自然である。

3 材料と方法

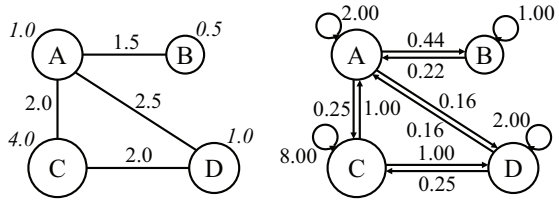
3.1 基礎語彙

系統モデルにおいて、各言語は特徴の集合で表現される。特徴として何を使うかに選択肢があるが、ここでは基礎語彙を想定する。基礎語彙は言語年代学の研究から生まれた基本概念の一覧である [40, 41]。この一覧中の概念を表す語は歴史的に変化しにくいと仮定されている。

基礎語彙データベースを作るには、対象となる各言語について、該当する概念をどのような語で表すかを調査する。例えば、warm という概念を表す語として、青森方言で「アッタゲ」、東京方言で「アタタカイ」、大阪方言で「ヌクイ」、福岡方言で「ヌッカ」が採取されたとする。このうち、青森方言と東京方言は語源を同じくする同源語 (cognate) を用いている。同様にして、大阪方言と福岡方言も同源語を用いている。

対象言語がある同源語を用いる場合に 1、用いない場合に 0 を立てる。ただし、ある概念について一つの言語から複数の語が採取され得るとする。すべての同源語についてこの操作を行い、結果を集約すると、各言語は $10010\cdots$ のようなバイナリ列で表現できる。また、言語間の距離を 0/1 の不一致率で定義できる。従来、系統モデルはこのようなバイナリ列に対して適用されてきた。

このようなバイナリ列は、概念を共有する同源語のまとまりを無視する。上記の例では、「アタタカイ」系の語と「ヌクイ」系の語が同じ warm という概念を表すという情報は失われている。本稿では、両者の競合関係が重要と考え、概念を単位としたモデル化を行う。



(a) (b)

図 2: ノードのネットワーク

3.2 モデル

方言群を接触により説明するためのシミュレーション用のモデルを提案する。従来の系統モデルは連続時間を扱ってきたが、ここでは簡単のために離散時間モデルを採用する。モデルはマルコフ性を持ち、時刻 t における状態は時刻 $t-1$ の状態のみに依存する。

各方言に対応するノードを用意する。ノードの状態は、与えられた概念を表す語の集合である。時刻 $t \geq 0$ においてノード i が持つ語の集合を $V_i^t \neq \emptyset$ とする。基礎語彙は、その意味を表す何らかの語がなければ困ると仮定し、少なくとも 1 語を持つという制約を与えている。

ノードはエージェントとして自律的に振る舞い、確率的に次の状態を決める。ノードの状態選択に影響するのは他のノードとの接触である。これを表現するために、ノードのネットワークを用意する。図 2 にネットワークの例を示す。辺が引かれたノードの対は接触下にある。ネットワークトポロジーは原則的に時間に対して不変とする。ただし、必要に応じて変更することも可能である。

ノードの影響力は人口と距離により制御される。図 2a に示す無向グラフにおいて、ノード、辺に付された数値はそれぞれ人口と距離を表す。ノード i の人口を p_i、ノード i, j 間の距離を $d_{ij} = d_{ji}$ とする。ノード i に対するノード j の影響力を次のように定める。

$$w_{ij} = \begin{cases} p_j \,/\, (d_{ij} \times d_{ij}) & \text{if } j \in N_i, \\ s\, p_i & \text{if } j = i, \\ 0 & \text{otherwise} \end{cases}$$

ここで、N_i はノード i との間に辺が引かれているノードの集合である。人口が大きいほど影響力が大きく、距離が離れているほど影響力が小さい。パラメータ s は自身の影響力を制御する。図 2b の有向グラフは影響力 w_{ij} を示している。この例では $s = 2.0$ とした。

ある時刻 $t > 0$ におけるノード i の状態は、時刻 $t-1$ におけるネットワークの状態に基づき確率的に決定される。$t = 0$ におけるすべてのノードの状態はあらかじめ決めておく。簡単のため、今回はすべてのノードが同じ語を 1 つだけ持つとする。

時刻 t においてノード i が取り得る語の集合を C_i^t とすると、$V_i^t \in \mathcal{P}(C_i^t) \setminus \emptyset$ である。つまり、C_i^t の冪集合から空集合を除いたものである。V_i^t はこれらの候補から以下に定める確率分布にしたがって選ばれる。

C_i^t は、時刻 $t-1$ において自身を含む隣接ノードのいずれかで用いられていた語の集合に新語を加えたものとする。

$$C_i^t = \{\,\text{newword}\,\} \cup \bigcup_{j \in \{i\} \cup N_i} V_j^{t-1}$$

新語は 1 ステップで各ノードにつき最大で 1 語しか誕生しないと仮定している。また、新語は誕生時点では 1 ノードでしか使われない。なお、一度すべてのノードから失われた語は二度と復活しない。

$v \in \mathcal{P}(C_i^t) \setminus \emptyset$ の確率を定義するために、まずは各語 $c \in C_i^t$ に以下のスコアを与える。

$$\text{wscore}(c) = \begin{cases} \sum_{j \in M_{i,c}^t} w_{ij} / |V_j^{t-1}| & \text{if } c \neq \text{newword}, \\ b\, p_i & \text{otherwise} \end{cases}$$

ここで $M_{i,c}^t = \left\{ j \,|\, j \in \{i\} \cup N_i \wedge c \in V_j^{t-1} \right\}$、つまり自身を含む隣接ノードであり、かつ時刻 $t-1$ において語 c を保持していたものの集合である。自身を含む多くの隣接ノードで使われている語ほど高いスコアを得る。パラメータ b は新語の誕生しやすさを制御する。

次に、$V_i^t = v$ としたとき、この語集合のスコアを

$$\text{sscore}(v) = \sum_{c \in v} \left(\text{wscore}(c) - d^{|v|} \right)$$

とする。語コスト $d > 1$ は語数の多い候補ほどペナルティがかかるように働く。最後に、$V_i^t = v$ となる確率を次のように定義する。

$$\frac{\exp(\text{sscore}(v))}{\sum_{v' \in \mathcal{P}(C_i^t) \setminus \emptyset} \exp(\text{sscore}(v'))}$$

この確率分布にしたがって V_i^t を選ぶという操作をすべてのノード i に対して行うと、1 ステップが終了する。

このシミュレーションは、あらかじめ決められたステップ数 T だけを行う。今回はすべての場合に $T = 1,000$ とした。こうして得られるのは、1 つの概念に関する語のネットワーク上の分布である。この操作を基礎語彙として定義された概念の数だけ繰り返し、結果をバイナリ化することで、各言語の特徴集合が得られる。概念の数は、言語年代学でよく用いられるスワデシュリスト [40, 41] にしたがい 100 とする。

最後に、このモデルの妥当性について議論する。同じ概念を表す語の競合関係については、与えられた概念に丁度あてはまる語があると想定している。一方、概念が細分化されおり、総称が存在しない場合がある。例えば、人間に対して言う場合と、動物に対して言う場合で別の語を使うことがある。こうした場合は、語は競合せず棲み分けていると思われる。対象外の言語からの借用は、語の誕生として問題なく解釈できる。ただし、同じ語の借用が複数回起こり得る。その場合は、同じ語は一度しか誕生しないという仮定に反する。語の誕生は、文字通り誕生するだけでなく、既存の語の意味の変化によっても起こり得る。そうした意味の変化も並行的に起きる可能性がある。以上、モデルの仮定に反する様々な場合を想定したが、既存の基礎語彙データを見る限り、ひとまず例外として無視できると考えている。

3.3 ネットワークトポロジー

シミュレーションに使うネットワークトポロジーとして、Grid、Star、TwoStars、Bottleneck、Colony の 5 種類を採用する。それらを図 3 に示す。影響力は人口と距離により示している。

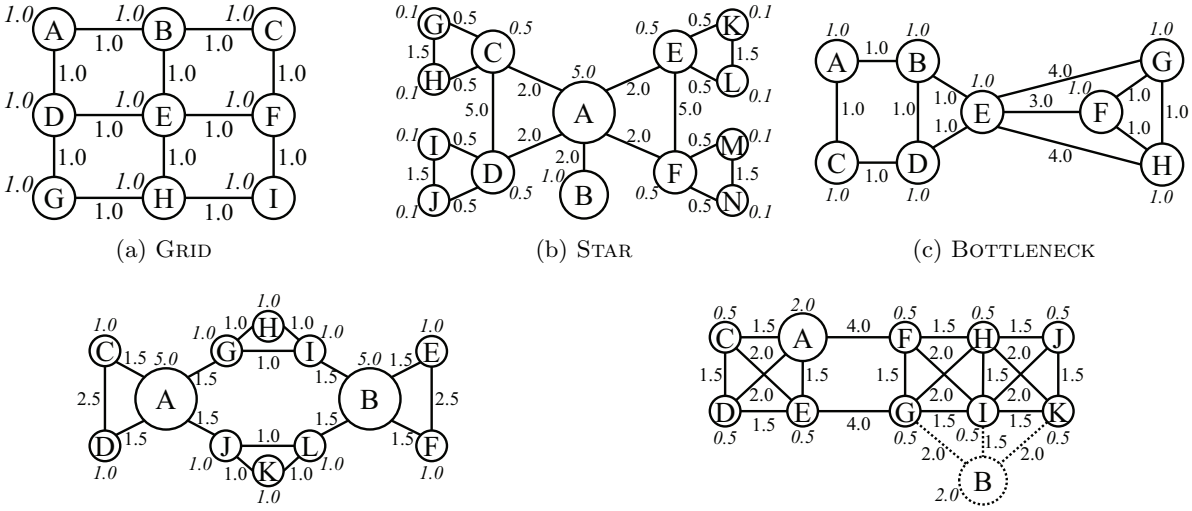
Grid (図 3a) は同規模のノードを格子状に配置している。これを空間的に木らしくない構造の例とする。パラメータは $s = 4.0$、$b = 1.0$、$d = 3.0$ とする。

Star (図 3b) では、人口の大きなノードが中心に位置し、人口の小さなノードがまわりに配置されている。これは政治的文化的中心と地方との関係を想定している。Grid と異なり、このトポロジーは空間的に木らしい。いずれのノードも他のノードとそれほど離れておらず、系統モデルにおける分岐として解釈できそうな断絶は存在しない。これを時間的な木である系統モデルがどのように解釈するかが焦点となる。パラメータは $s = 2.0$、$b = 0.1$、$d = 3.0$ とする。

Bottleneck (図 3c) は 2 つの方言群の間に隘路を設けている。Star と異なり、系統モデルの観点からは、シミュレーション開始時点で 2 つに分岐したと解釈できる。注意すべきは 2 つが完全に断絶しているのではなく、多少の接触があることである。これが結果にどのような影響をもたらすかに関心がある。パラメータは $s = 4.0$、$b = 0.1$、$d = 3.0$ とする。

TwoStars (図 3d) は Star の変種で、中心が 2 つ存在する。2 つの中心に挟まれたノードがどうなるかに関心がある。パラメータは $s = 2.0$、$b = 1.0$、$d = 3.0$ とする。

Colony (図 3e) では途中でネットワークトポロジーを変更する。1,000 ステップのうち、最初の 750 ステップは、ノード B (破線部) 抜きで実行し、残り 250 ステップでは B 込みで実行する。移行時には A の状態を



(a) Grid (b) Star (c) Bottleneck

(d) TwoStars (e) Colony

図 3: シミュレーションに用いるネットワークトポロジー

B に複製する。これは系統モデルにおける分岐に対応する。このトポロジーでは方言群は大きく東西にわかれており、B は西から東への殖民として解釈できる。これがどのような結果をもたらすかに興味がある。パラメータは $s = 4.0$、$b = 0.1$、$d = 3.0$ とする。

3.4 分析

シミュレーションにより得られた各言語の特徴集合が木らしいかを分析する。分析手順は [16] を参考にして、木らしさを表す指標として δ スコア [20] を用いる。この指標は 0 以上 1 以下の実数であり、0 に近いほど木らしい。この指標の基礎となるのは 4 つ組ノード x, y, u, v である。x, y の距離を d_{xy} とし、$d_{xy|uv} = d_{xy} + d_{uv}$ とする。さらに、一般性を失うことなく、$d_{xy|uv} \leq d_{xu|yv} \leq d_{xv|yu}$ としたうえで、δ を以下のように定義する。

$$\delta = \frac{d_{xv|yu} - d_{xu|yv}}{d_{xv|yu} - d_{xy|uv}}$$

ただし $d_{xy|uv} = d_{xu|yv} = d_{xv|yu}$ のときは $\delta = 0$ とする。別の説明をすると、ノード間の距離は図 4 に示すように分解できる。ここで、箱は木としての不整合性を表している。$l_1 = d_{xv|yu} - d_{xy|uv}$、$l_2 = d_{xv|yu} - d_{xu|yv}$ であり、つまり $\delta = l_2/l_1$ である。$l_2 = 0$ のとき完全に木らしい。この計算をすべての 4 つ組に対して行い、平均を取ったものが δ スコアである。残念ながら、複数の δ スコアを比較する検定手法は知られていない [16]。

次に、諸言語の関係を NeighborNet [5] を用いて可視化する。NeighborNet は距離に基づく凝集型クラスタリングを行うが、木ではなくネットワークを構築する。木らしからぬ不整合性は箱状に表現される。実装としては SplitsTree4 (バージョン 4.13.1)[3] を用いる。

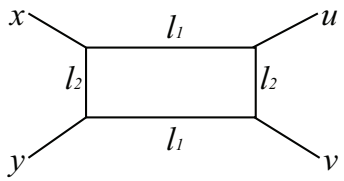


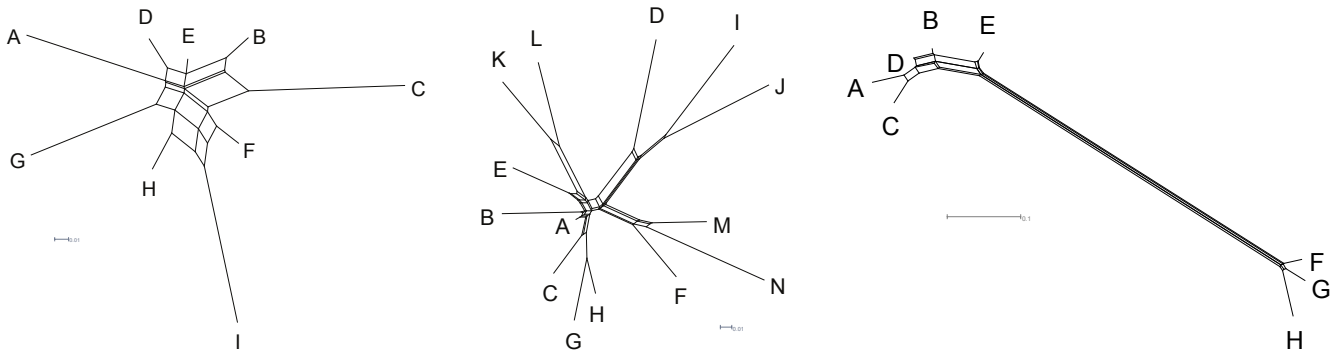
図 4: 4 つ組ノード

最後に、各言語の特徴集合を系統モデルがどのように解釈するかを調べる。系統樹作成には、ベイズ統計に基づく生成モデルを用い、事後分布からサンプルを得る。特徴のモデルとしては、確率的ドロモデルを採用する。このモデルでは特徴の誕生は木全体で一度しか起きない。また、ある枝で一度死んだ特徴が子孫で復活することはない。シミュレーション用モデルでは接触により特徴が復活し得ることに注意を要する。ドロモデルの死亡率の事前分布は平均 10^{-5} の指数分布とする。木の事前分布は Bayesian skyline plot を用いる。また、絶対年代の推定は行わないため、単純な厳密時計モデルを採用する。Markov chain Monte Carlo サンプリングは 20,000,000 ステップ行う。このうち最初の 100,000 ステップをならし期間とし、以後 1,000 ステップごとにサンプルを得る。実装としては BEAST 1.8.0[4] を用いる。

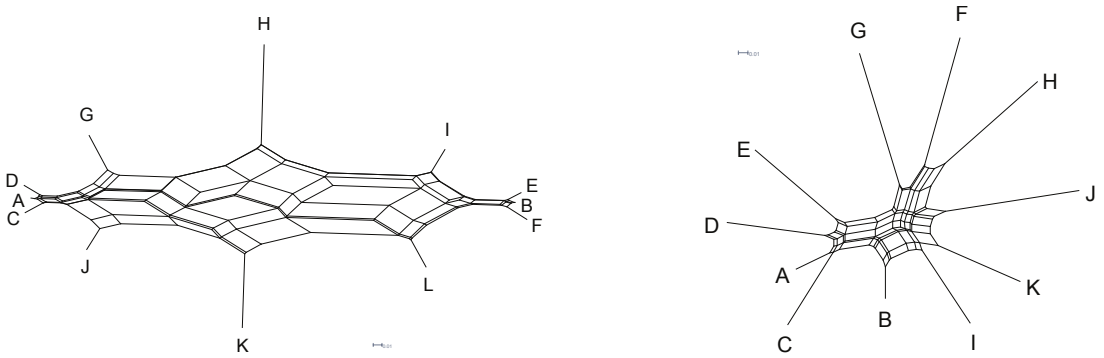
4 結果

NeighborNet を図 5 に、系統樹を図 6 に示す。系統樹は最大系統群信頼度木 (maximum clade credibility

[3] http://www.splitstree.org/
[4] https://code.google.com/p/mcmc/

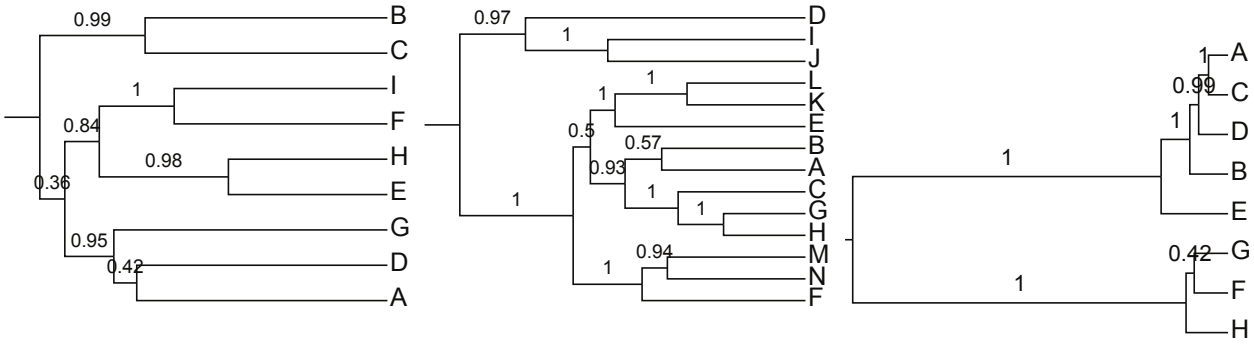


(a) Grid (b) Star (c) Bottleneck

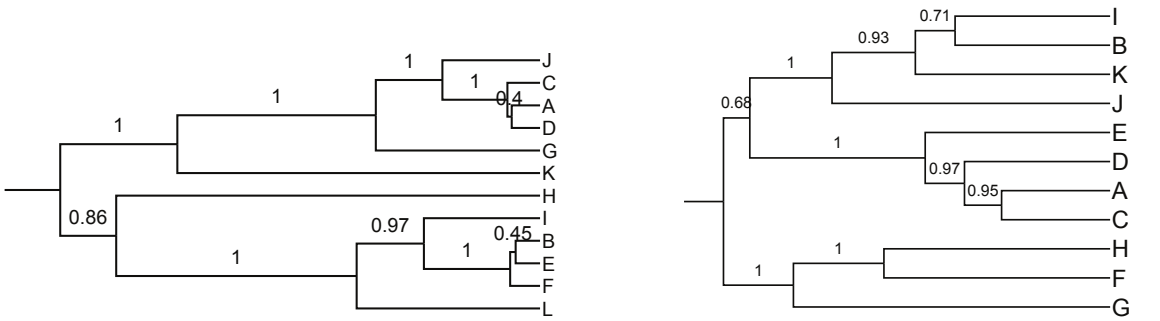


(d) TwoStars (e) Colony

図 5: NeighborNet の結果



(a) Grid (b) Star (c) Bottleneck



(d) TwoStars (e) Colony

図 6: 推定された系統樹

tree) である。枝に付した数値は該当ノードが事後分布において支持された割合を示す。比較のために既存文献で報告されている δ スコアを見ると、アイヌ語諸方言で 0.25 [27]、印欧語族は全体で 0.22、ゲルマン語派で 0.23、オーストリネシア語族のポリネシア諸語で 0.41 [16] である。日本語諸方言 [26] の δ スコアは報告されていないが、補足資料をもとに計算したところ、全体で 0.35、現代本土諸方言に限定すると 0.39 を得た。

Grid の δ スコアは 0.43 であり、予想通りあまり木らしくない。NeighborNet (図 5a) を見ると、木としての不整合性が確認できる。ネットワークトポロジー上の中央の E が中央に位置し、接触の少ない四隅が離れている。系統樹 (図 6a) では根に近い系統群の推定結果が不安定である。

Star の δ スコアは 0.28 であり、ある程度木らしい。NeighborNet (図 5b) では、ネットワークトポロジー上の中心たる A が中心にきており、空間構造が再現されている。これは、木らしさに、時間的木らしさだけではなく、空間的木らしさが貢献することを示唆している。系統樹 (図 6b) においては、空間的に離れた方言群がより早く分岐したことになっている。空間構造が時間構造に読みかえられている。

Bottleneck の δ スコアは 0.15 であり、非常に木らしい。NeighborNet (図 5c) を見ると、2 つの方言群が再現されている。期待通り、E が FGH の系統群に相対的に近い。系統樹 (図 6c) においても 2 つの方言群が再現されている。ここでも E が方言群内の外れ値となっている。系統モデル自体は凝集型クラスタリングではないが、それと同じように中間的状態を持つノードのクラスタリングを後回しにしていることが確認できる。同じ現象がアイヌ語諸方言の系統樹 [27] でも起きているのではないかと推測する。この系統樹では、北海道最北端の宗谷方言が北海道方言における外れ値となっている。これは時間構造よりも、サハリンに近いという空間構造により説明できるのではないだろうか。文献 [27] は祖語の位置も同時推定し、アイヌ語話者が北海道北部に起源を持つと主張している。しかし、根の近くで分岐したことになっている宗谷方言が、北海道祖語の位置を北寄りにし、ひいては北海道・サハリン共通祖語の位置を北寄りに推定するのに貢献した可能性がある。

TwoStars の δ スコアは 0.19 であり、木らしい。NeighborNet (図 5d) では一見不整合のようだが、箱の 2 辺の比は 0 寄りである。2 つ中心から等距離にある K と H はどっちつかずとなっている。系統樹 (図 6d) は 2 つの方言群を推定している。K と H はネットワークトポロジーにおいては対称であるが、それぞれ別の方言群にクラスタリングされている。

Colony の δ スコアは 0.33 であり、あまり木らしくない。NeighborNet (図 5e) で見ると A と B はそれほど離れていないが、系統樹 (図 6e) では、早々に分岐したことになっている。IBKJEDAC の系統群の支持が 0.68 であることから推測されるように、最初の分岐が A と B を隔てるサンプルも少なくない。つまり、多くのノードが、最初から存在するにも関わらず、A と B の分岐以降に起きた分岐により形成されたと解釈されている。同様の現象が起きていると思われるのが図 1 の日本語諸方言である。現代語における最初の分岐が京都と東京を隔てている。また、語彙の分布を見ると、750 ステップの時点では方言群が東西に分化していたが、その後 B が周辺の東方言を西方言の語彙で上書きしている。系統樹上も東西の区別がぼやけている。同様にして、図 1 は日本語の東西対立の同定に部分的に失敗している。

5 議論

接触に基づくモデルのシミュレーション結果を系統モデルに解釈させた結果、現実の方言群から推定された系統樹に見られるのに似た不自然な振る舞いが確認できた。もちろん、この結果をもって、現実の方言群がシミュレーションの想定にしたがって形成されたとは断定できない。本手法は発見的であり、あくまで可能性の一つを示したにすぎない。別の仮説がよりうまく実データを説明するかもしれない。

本稿で行ったようなシミュレーションは、どのような仕組みで語彙変化が起きるかという疑問に答えるための新たな方法論である。例えば、言語学では、借用が起きる要因の一つとして威信が挙げられている [6] が、こうした仮説をシミュレーション用モデルに組み込むことで、その妥当性を検証できる。

系統モデルでは、絶対年代の校正やノードの地理位置 (緯度・経度) の同時推定が行われている。一方、提案した接触モデルは、パラメータ (人口、距離、ステップ数等) と現実との関係が明らかでない。絶対年代や地理位置、人口変動などの実データをモデルに盛り込むには、この対応づけが欠かせない。

現在のモデルでは、一部の語を長期的に生き残らせるには、背後に膨大な数の短命の語が必要となる。図 7 に Star における語の生存期間を示す。これは、語の誕生から消滅 (すべてのノードから消える) までのステップ数ごとに語を集計したものである。ただし、初期状態の語、および終了時点で生き残っていた語は

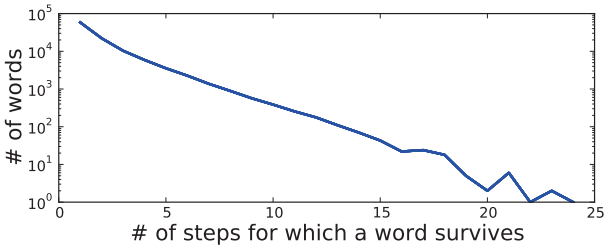


図 7: STAR における語の生存期間

含まない。この図から、短期間に消滅する語が少なくないことが確認できる。新語が誕生しても、隣接ノードの影響ですぐに古い語が復活しやすいからである。

現在のモデルでは、人口の大きなノードほど新語が生まれにくい。確率分布の構成に指数関数の正 (または 0 に近い負の) 定義域を使うため、大ノードほど歪んだ分布になる。そのため、スコアが相対的に低い新語に対してより低い確率が与えられる。その代わり、ひとたび大ノードで新語が生まれたら、その影響力により周辺に広がりやすい。この性質は妥当ではないかと考えている。琉球諸方言内の語彙のばらつきが本土諸方言よりもはるかに大きいことから、人口が小さく、周辺との接触が少ないほど語彙の変化が進むのではないかと推測できる。しかし、その一方で、アイスランド語のように、話者が少ないが極めて保守的な例も知られている [22]。現在の仮定には実証的な裏付けが必要である。

現在のモデルは、実際には複数の話者が担う各言語を一つのノードで近似し、言語間の非対称性を人口パラメータで制御している。実データにおいても、通常一つの集団から単一の言語データを採取する。一方、生物の遺伝子の場合、データの単位は個人であり、一つの集団から複数のサンプルを得る。そして、集団内のばらつきは大きな手がかりとなる [46, 13]。言語には遺伝子ほど集団内でのばらつきは期待できない。言語は通信手段であり、話者間で語彙の共有が必要だからである。言語データの採取自体が高コストであることも考慮すると、個人単位でのデータ採取は割にあわない。しかし、シミュレーションであれば容易に実現できる。今後の方向性として、各言語の状態を観測状態と分離したうえで、各言語自体を複数の話者からなるネットワークで表現するという拡張が考えられる。こうすれば、短命の語を大量に観測することなく語彙変化を起こせる可能性がある。また、通婚関係や世代交代などを組み込み、モデルをより現実的に設計できるかもしれない。

既存の実データの多くは 1 つのスナップショットであるが、仮説を効果的に検証するには、時系列データが望ましい。また、充分な分解能があれば、シミュレーションではなく、実データを説明するモデルパラメータの推定も夢ではない。基礎語彙は安定的であり変化を追跡しにくいが、現代の新語であれば、大規模な時系列コーパスを用いれば、伝播過程を明らかにできるかもしれない。ただし、変化の定着を確認するには複数世代にまたがるデータが必要だろう。

図 1 では、比較対象の現実の方言群として、日本語諸方言のうち本土のみに着目し、琉球諸方言を除外した。琉球諸方言は離島に位置し、恒常的な接触という前提が成り立たない可能性があるからである。しかし、琉球諸方言の語彙も接触という点で興味深い。琉球諸方言の分岐自体は、規則的な音対応に着目する伝統的な比較手法に基づき、上代語 (8 世紀の畿内) 以前と考えられている [38, 39]。こうした研究では見過ごされがちだが、本土諸方言と琉球諸方言の共通語彙には、本土では上代語以降に登場するものが少なからず確認でき、それにはいくつかの文法化した語も含まれる [47, 48, 49]。そうした語は、系統上の分岐以降に琉球諸方言が本土諸方言から借用したとみられる。琉球諸方言の成立過程にはいまだに謎が多いが、これを解明するうえで、本稿で提案したような接触に基づくモデルが役立つかもしれない。

6 おわりに

系統モデルは言語群のマクロな動態に対する近似として、語彙データに盛んに適用されてきた。一方、小規模な方言群においては、系統モデルを特徴づける分岐のあとを見つけるのは難しい。この問題に対する本稿の提案は、系統モデルではなく、むしろ進化グラフやマーケティングのモデルに近い、接触に基づく語彙伝播モデルを用いることである。

本稿では、シミュレーション用モデルを設計し、その人工データを系統モデルがどのように解釈するかを調べた。これにより、現実の方言群から推定された系統樹に見られる不自然な振る舞いの説明を試みた。

参考文献

[1] Jan O. Andersson. Lateral gene transfer in eukaryotes. *Cellular and Molecular Life Sciences CMLS*, Vol. 62, No. 11, pp. 1182–1197, 2005.

[2] François Barbançon, Tandy Warnow, Steven N. Evans, Donald Ringe, and Luay Nakhleh. An experimental study comparing linguistic phylogenetic reconstruction methods. *Diachronica*, Vol. 30, No. 2, pp. 143–170, 2013.

[3] Adrian C. Barbrook, Christopher J. Howe, Norman Blake, and Peter Robinson. The phylogeny of the Canterbury Tales. *Nature*, Vol. 394, p. 839, 1998.

[4] Remco Bouckaert, Philippe Lemey, Michael Dunn, Simon J. Greenhill, Alexander V. Alekseyenko, Alexei J. Drummond, Russell D. Gray, Marc A. Suchard, and Quentin D. Atkinson. Mapping the origins and expansion of the Indo-European language family. *Science*, Vol. 337, No. 6097, pp. 957–960, 2012.

[5] David Bryant and Vincent Moulton. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, Vol. 21, No. 2, pp. 255–265, 2004.

[6] Lyle Campbell. *Historical Linguistics: An Introduction (2nd edition)*. Edinburgh University Pres, 2004.

[7] Lyle Campbell. Areal linguistics: A closer scrutiny. In Yaron Matras, April McMahon, and Nigel Vincent, editors, *Linguistic Areas: Convergence in Historical and Typological Perspective*, pp. 1–31. Palgrave Macmillan, 2006.

[8] Thomas E. Currie, Simon J. Greenhill, and Ruth Mace. Is horizontal transmission really a problem for phylogenetic comparative methods? a simulation study using continuous cultural traits. *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 365, No. 1559, pp. 3903–3912, 2010.

[9] Thomas E. Currie, Andrew Meade, Myrtille Guillon, and Ruth Mace. Cultural phylogeography of the Bantu languages of sub-Saharan Africa. *Proceedings of the Royal Society B: Biological Sciences*, Vol. 280, No. 1762, 2013.

[10] Charles Darwin. *The Origin of Species by Means of Natural Selection or, the Preservation of Favored Races in the Struggle for Life*. John Murray, 1859.

[11] Hal Daumé III. Non-parametric Bayesian areal linguistics. In *Proc. of HLT-NAACL*, pp. 593–601, 2009.

[12] Sebastiano A. Delre, Wander Jager, Tammo H. A. Bijmolt, and Marco A. Janssen. Will it spread or not? the effects of social influences and network topology on innovation diffusion. *Journal of Product Innovation Management*, Vol. 27, No. 2, pp. 267–282, 2010.

[13] Laurent Excoffier, Peter E. Smouse, and Joseph M. Quattro. Analysis of molecular variance inferred from metric distances among DNA haplotypes: Application to human mitochondrial DNA restriction data. *Genetics*, Vol. 131, No. 2, pp. 479–491, 1992.

[14] Walter M. Fitch and Emanuel Margoliash. Construction of phylogenetic trees. *Science*, Vol. 155, No. 760, pp. 279–284, 1967.

[15] Russell D. Gray and Quentin D. Atkinson. Language-tree divergence times support the Anatolian theory of Indo-European origin. *Nature*, Vol. 426, No. 6965, pp. 435–439, 2003.

[16] Russell D. Gray, David Bryant, and Simon J. Greenhill. On the shape and fabric of human history. *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 365, No. 1559, pp. 3923–3933, 2010.

[17] Russell D. Gray and Fiona M. Jordan. Language trees support the express-train sequence of Austronesian expansion. *Nature*, Vol. 405, No. 6790, pp. 1052–1055, 2000.

[18] Simon J. Greenhill, Thomas E. Currie, and Russell D. Gray. Does horizontal transmission invalidate cultural phylogenies? *Proc. of the Royal Society B: Biological Sciences*, Vol. 276, No. 1665, pp. 2299–2306, 2009.

[19] Clare Janaki Holden. Bantu language trees reflect the spread of farming across sub-Saharan Africa: a maximum-parsimony analysis. *Proceedings of the Royal Society of London. Series B: Biological Sciences*, Vol. 269, No. 1493, pp. 793–799, 2002.

[20] B. R. Holland, K. T. Huber, A. Dress, and V. Moulton. δ plots: A tool for analyzing phylogenetic distance data. *Molecular Biology and Evolution*, Vol. 19, No. 12, pp. 2051–2059, 2002.

[21] Schmidt Johannes. *Die Verwandtschaftsverhältnisse der indogermanischen Sprachen*. Hermann Böhlau, 1872. (in German).

[22] Stefán Karlsson. *The Icelandic Language*. Viking Society for Northern Research, 2004.

[23] David Kempe, Jon Kleinberg, and Éva Tardos. Maximizing the spread of influence through a social network. In *Proc. of KDD*, pp. 137–146, 2003.

[24] Paul Kerswill. Dialect levelling and geographical diffusion in British English. In David Britain and Jenny Cheshire, editors, *Social dialectology: in honour of Peter Trudgill*, pp. 223–243. John Benjamins Publishing, 2003.

[25] William Labov. Transmission and diffusion. *Language*, Vol. 83, No. 2, pp. 344–387, 2007.

[26] Sean Lee and Toshikazu Hasegawa. Bayesian phylogenetic analysis supports an agricultural origin of Japonic languages. *Proceedings of the Royal Society B: Biological Sciences*, Vol. 278, No. 1725, pp. 3662–3669, 2011.

[27] Sean Lee and Toshikazu Hasegawa. Evolution of the Ainu language in space and time. *PLoS ONE*, Vol. 8, No. 4, p. e62243, 2013.

[28] Luke J. Matthews. The recognition signal hypothesis for the adaptive evolution of religion. *Human Nature: An Interdisciplinary Biosocial Perspective*, Vol. 23, No. 2, pp. 218–249, 2012.

[29] Luay Nakhleh, Don Ringe, and Tandy Warnow. Perfect phylogenetic networks: A new methodology for reconstructing the evolutionary history of natural languages. *Language*, pp. 382–420, 2005.

[30] Shijulal Nelson-Sathi, Johann-Mattis List, Hans Geisler, Heiner Fangerau, Russell D. Gray, William Martin, and Tal Dagan. Networks uncover hidden lexical borrowing in Indo-European language evolution. *Proceedings of the Royal Society B: Biological Sciences*, 2010.

[31] Geoff K. Nicholls and Russell D. Gray. 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)*, Vol. 70, No. 3, pp. 545–566, 2008.

[32] Johanna Nichols and Tandy Warnow. Tutorial on computational linguistic phylogeny. *Language and Linguistics Compass*, Vol. 2, No. 5, pp. 760–820, 2008.

[33] Martin A. Nowak, Corina E. Tarnita, and Tibor Antal. Evolutionary dynamics in structured populations. *Philosophical Transactions of the Royal Society B: Biological Sciences*, Vol. 365, No. 1537, pp. 19–30, 2010.

[34] Takuichiro Onishi. Analyzing dialectological distributions of Japanese. *Dialectologia*, Vol. 2010, No. Special Issue I, pp. 123–135, 2010.

[35] Don Ringe, Tandy Warnow, and Ann Taylor. Indo-European and computational cladistics. *Transactions of the Philological Society*, Vol. 100, No. 1, pp. 59–129, 2002.

[36] Robert M. Ross, Simon J. Greenhill, and Quentin D. Atkinson. Population structure and cultural geography of a folktale in Europe. *Proceedings of the Royal Society B: Biological Sciences*, Vol. 280, No. 1756, 2013.

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

[38] Leon A. Serafim. The use of Ryukyuan in understanding Japanese language history. In Bjarke Frellesvig and John Whitman, editors, *Proto-Japanese: Issues and Prospects*, pp. 79–99. John Benjamins Publishing, 2008.

[39] Moriyo Shimabukuro. *The Accentual History of the Japanese and Ryukyuan Languages: A Reconstruction*. Global Oriental, 2007.

[40] Morris Swadesh. Lexicostatistic dating of prehistoric ethnic contacts. *Proc. of American Philosophical Society*, Vol. 96, pp. 452–463, 1952.

[41] Morris Swadesh. *The Origin and Diversification of Language*. Aldine Atherton, 1971.

[42] Ilya Tëmkin and Niles Eldredge. Phylogenetics and material cultural evolution. *Current Anthropology*, Vol. 48, No. 1, pp. 146–154, 2007.

[43] Tandy Warnow, Steven N. Evans, Donald Ringe, and Luay Nakhleh. A stochastic model of language evolution that incorporates homoplasy and borrowing. In Peter Forster and Colin Renfrew, editors, *Phylogenetic Methods and the Prehistory of Languages*, pp. 75–90. McDonald Institute for Archaeological Research, 2006.

[44] Duncan J. Watts and Peter Sheridan Dodds. Influentials, networks, and public opinion formation. *Journal of Consumer Research*, Vol. 34, No. 4, pp. 441–458, 2007.

[45] John Whitman. Northeast Asian linguistic ecology and the advent of rice agriculture in Korea and Japan. *Rice*, Vol. 4, No. 3-4, pp. 149–158, 2011.

[46] Sewall Wright. Isolation by distance. *Genetics*, Vol. 28, No. 2, p. 114, 1943.

[47] 亀井孝. 日本語系討論のみち, 琉球方言の史的地位, pp. 91–114. 吉川弘文館, 1973.

[48] 中本正智. 琉球語彙史の研究, 琉球語の成立と時代層: 日本語の原像をさぐる, pp. 15–27. 三一書房, 1983.

[49] 荻野千砂子. 八重山地方の授受動詞タボールンと中世語「給はる」: 敬意優先の授受動詞体系. 日本語の研究, Vol. 7, No. 4, pp. 39–54, 2011.

[50] 柳田國男. 蝸牛考. 刀江書院, 1930.