数理的手法を用いた日本語の系統に関する考察

小橋 昌明　　田中 久美子
東京大学工学部計数工学科　　東京大学大学院情報理工学系研究科
kobashi@cl.ci.i.u-tokyo.ac.jp　　kumiko@i.u-tokyo.ac.jp

1　はじめに

系統関係を表現した図を系統樹 (Phylogenetic tree) という。昨今では生物種の進化の系統樹を作る研究が行われ、その過程で多くの数理的な生成手法が提案された。その手法を応用して、これまでに言語のデータから系統樹が作られてきた。従来は、比較的系統関係が明確な言語に対して系統樹を生成していた。例えば、Gray & Atkinson は、インド・ヨーロッパ語族のデータを用いてコンピュータで計算し、分岐年代を推定している [3]。また、Gray らはオーストロネシア語族の系統樹を生成し、語族の拡散と停止について論じている [4]。Rexová らはアフリカ中南部に分布するバントゥー語群について分析している [9]。ここに挙げた他にも、これらの語族に関しては多くの研究がなされている。また、江原は WALS のデータを用いて日本周辺の言語を多次元尺度構成法で分析している [13]。

本研究では、日本語と関連する近隣の諸言語との系統関係を、既存の仮説に基づいて考察した。言語の構造的特徴量のデータベースである WALS (The World Atlas of Language Structures) [5] などから、関連する言語について 100 を越える特徴量を抽出した。近隣結合法・最大節約法・ベイズ法の 3 つの手法で系統樹を生成し、その結果を比較検討する。

2　現在の仮説

日本語の起源の仮説は主に 3 つある。一つ目は、アルタイ語族の一つだという、服部四郎 [15] らの説である。二つ目は、オーストロネシア語族との関連を指摘する説である。中でも松本克己 [14] の説は、日本語とオーストロネシア語族の他にも、中国語、ミャオ・ヤオ語族、タイ・カダイ語族、オーストロアジア語族などを包含する「環太平洋言語圏」というグループを考える比較的新しい説であり、今回はこれを取り上げる。三つ目は、タミル語を祖先とするという、大野晋の説である。これらの学説はいずれも、定説として広く受け入れられたわけではない。この 3 つの他にも、多くの仮説がある。韓国語と同系だとする説も古くから存在する。

今回は、近隣の諸言語の中でも比較的広範囲にわたる日本語の系統に関する上の 3 つの仮説に特に着目し、それらの妥当性を数理的手法に基づいて考察する。

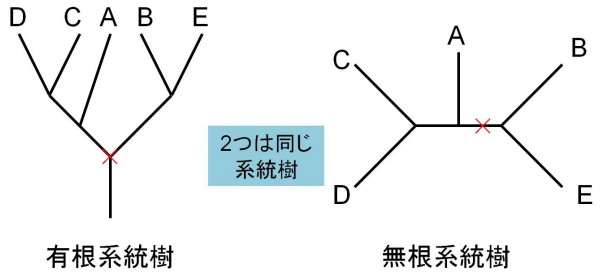

図 1: 有根系統樹と無根系統樹

3　有根系統樹と無根系統樹

系統樹は、樹根の有無によって有根系統樹 (rooted tree) と無根系統樹 (unrooted tree) に分類される。例えば図 1 のように、同一の系統関係を両方の系統樹で表すことができる。図 1 で、右の木の赤い×印の箇所が最も時代が古いと仮定して根を導入すると左の木になる。根の導入位置は一意でないので、一つの無根系統樹に複数の有根系統樹が対応する。有根系統樹には時間の前後という情報が含まれているが、無根系統樹には含まれていない。殆どの系統樹作成法は根を特定できないので、有根系統樹を考える際には根の導入を別に行う必要がある。

今回は仮説との兼ね合いで、無根系統樹を考える。その理由は以下の通りである。松本は「日本語の発祥が伝統的な比較言語学では手の届かない遠い過去に遡る」とし、「このような古い言語圏ないし言語層から受け継いだ構造的特徴の共有は、(中略) 比較言語学的な意味での同系性とはもちろん異なり、それよりもっと根の深いところで諸言語を結びつけるような “類縁

性” に根ざしている」と述べている [14]。

有根系統樹で表せるのは、もともと単一であった言語が分岐し現在の姿になったという状況のみである。日本語の場合には、言語の類縁関係を表すには無根系統樹のほうが適切である。類縁性に基づく仮説を可能性の一つとして考える以上、無根系統樹を用いて本研究を行う必要がある。

4 系統樹の生成とその手法

系統樹は、言語や生物といった対象の集合の各要素に対して特徴量のベクトルを抽出し、それを用いて生成されるものである。生成の手法は多くあるが、複数の異なる方法で同じ結論が得られたなら、その結論はより確実性が高いと考えられる。今回は 3 つの手法で系統樹を生成した。以下にそれぞれの方法を詳述する。

4.1 手法 1：近隣結合法

近隣結合法 (Neighbor Joining) は、Saitou & Nei が 1987 年に提唱した古典的な方法である [11]。短い計算時間で系統樹を生成できるのが特徴である。

言語間の距離を何らかの形で定義して並べたものを距離行列という。近隣結合法は、距離行列から系統樹を作成する代表的な方法である。今回は、2 つの要素の特徴量のうち異なるものの割合を、2 つの要素間の距離と定義した。

近隣結合法は全ての要素が 1 点で連結された星状の系統樹から始める。そこから各要素を結合するが、判定基準として系統樹の枝長の総和を用いる。ある時点の系統樹に対して、任意の 2 つの要素の組み合わせについて、その 2 つを結合したときの系統樹の枝長の総和を計算する。この値が最小になるような 2 つの要素を結合する。この操作 1 回につき要素数は 1 減少するので、この操作を繰り返して要素数が 3 となったときに終了とする。

4.2 手法 2：最大節約法

最大節約法 (Maximum Parsimony) は、値の変化の回数が最も少なくなるような系統樹を出力する手法であり、古くから使われている [1]。

系統樹内部の点での各素性の値を仮定すると、系統樹上で値が変化する回数を計算することができる。ある一つの系統樹を定めたとき、その上での変化回数の最小値とそれを与える素性値は、動的計画法を用いて同時に計算できる [12]。変化の回数が最小になるような系統樹が、最大節約法による出力である。

ただし、有根の二分岐系統樹の場合、n 個の要素に対する可能な系統樹の個数を $f(n)$ とすると

$$f(n) = 3 \times 5 \times 7 \times \cdots \times (2n-3) = \frac{(2n-3)!}{2^{n-1}(n-1)!}$$

となる。この値は n について急激に増加し、$f(20) = 8.20 \times 10^{21}, f(50) = 2.75 \times 10^{76}$ となる。

そのため、すべての木を調べ上げて、「ある判定基準について最適な木」を決定するのは、n が小さい時でない限り実用上不可能である。n がそれほど大きくないときは分枝限定法を使って探索時間を短縮できるが、それにも限界がある。n が大きい時は heuristic search を使って最適解を探索するが、得られた解が大域的最適解とは限らないことは注意が必要である。

4.3 手法 3：ベイズ法

ベイズ法 (Bayesian method) は比較的新しく提唱された方法であり、明示的に進化のモデルを考えて計算を行うのが特徴の、 [8] などで提案された手法である。ただし大量の計算を行うので、一般に他の手法より時間がかかる。

データ D、仮説 (系統樹など)H に対して、

$$Prob(H|D) = \frac{Prob(H)Prob(D|H)}{\sum_H Prob(H)Prob(D|H)}$$

が成り立つ (ベイズの定理)。右辺の分母を解析的に計算することは事実上不可能である。しかし、実際にはマルコフ連鎖モンテカルロ法 (Malkov Cain Monte Carlo、MCMC) を使い、Metropolis-Hastings MCMC ではそれぞれの仮説の事後確率の比

$$\frac{Prob(H_1|D)}{Prob(H_2|D)} = \frac{Prob(H_1)Prob(D|H_1)}{Prob(H_2)Prob(D|H_2)}$$

を問題にしているため、この項を計算することなく事後確率を計算できる。

MCMC では多数の系統樹をサンプリングし、それらの情報から最終的な 1 つの系統樹を作る。Metropolis-Hastings MCMC では、ある状態の系統樹 T_n から規則に従って次の系統樹 T_n^* を作る。次の状態の系統樹 T_{n+1} は、候補 T_n^* を受理するか棄却するかで変わり、受理した場合 $T_{n+1} = T_n^*$、棄却した場合 $T_{n+1} = T_n$ となる。

データの解析には、定常分布に達したあとの木を使う。作った系統樹を全てデータの解析に使うと、標本ごとの独立性が無くなるので、一定の間隔で木を採取し、データの解析に使う。

5　用いたデータ・特徴量

今回、データとして主に WALS を用いた [5]。これは、言語の構造的特徴に関するデータベースであり、母音の数、受動態の有無、動詞と目的語の順序など、全 142 項目の特徴について言語ごとにその値が記されている。

ただし、全ての項目が埋まっているわけではなく、殆どデータがないような言語も存在する。また、日本語との系統関係が殆ど無いと思われる言語を選んでもあまり意味が無い。これらの点を考慮し、以下のように言語を選択した。

日本の近くの孤立言語 (日本語と同様、どの語族に属するが判明していない言語) には、韓国語、アイヌ語、ギリヤーク語 (ニヴフ語) がある。この 3 つは考察の対象とする。第 2 節を踏まえ、以下の言語を考察する。

(1) アルタイ語はツングース諸語・モンゴル諸語・テュルク諸語の 3 つに分かれる。ツングース諸語、モンゴル諸語の中で WALS データが一番充実しているのは、それぞれエベンキ (エウェンキー) 語、モンゴル語である。この 2 つを対象に入れる。

(2) オーストロネシア語族の中で主な言語であり、WALS データも充実しているインドネシア語、タガログ語、マオリ語を対象に入れる。その他、日本の近くの大言語で、「環太平洋言語圏」の要素である中国語を入れる。

(3) タミル語、及びタミル語と同じドラヴィダ語族南部ドラヴィダ語派に属するカンナダ語を対象に入れる。

以上をまとめると、選んだ言語は右上の表のようになる。

追加データとして、松本 [14] からのデータも用いた。同書 pp. 188〜191 の表をもとにデータを作成した。素性は流音のタイプ、形容詞のタイプなど全 10 項目である。松本によれば、これらの素性は「手近な語彙項目や表面的な形態・統語構造ではなく、言語のもっと内奥に潜みしかもそれぞれの言語の基本的な骨組みを決定づけるような言語特質、話し手の認知の在り方や言語によるそのカテゴリゼーション、言語のいわば遺伝子型に相当するような形質である」 [14]。

なお、これらの項目のうち、3 つ以下の言語でのみ値が入っているものは削除し、WALS と松本のデータで重複する特徴量は重複を削除した。素性の数は全部で 144 である。ただしどの言語についても、全ての素性に値が入っているわけではない。実際に値が埋まっている素性の数は表の通りである。

言語名	語族	素性数
日本語 Japanese	孤立	136
韓国語 Korean	孤立	134
ギリヤーク語 Nivkh	孤立	118
アイヌ語 Ainu	孤立	124
エベンキ語 Evenki	アルタイ	137
モンゴル語 Khalkha	アルタイ	134
インドネシア語 Indonesian	環太平洋	141
タガログ語 Tagalog	環太平洋	132
マオリ語 Maori	環太平洋	134
中国語 Mandarin	環太平洋	136
タミル語 Tamil	ドラヴィダ	79
カンナダ語 kannada	ドラヴィダ	134

6　結果

用いたソフトウェア：近隣結合法には SplitsTree4 [7]、最大節約法には phylip 3.69 [2]、ベイズ法には MrBayes 3.1.2 [6] [10] をそれぞれ用いた。得られた系統樹の表示には、SplitsTree4 を用いた。系統樹を次頁の図 2〜図 4 に示す。

MrBayes については、Metropolis-coupled MCMC を用い、100000 回の繰り返しを行った。20 回ごとに木を抽出し、5001 個の木が得られたが、始めの 501 個を burn-in(定常状態に達する前の状態) として棄却し、残りの 4500 個を採用した。2 つの MCMC を並行して行なったので、解析対象となった系統樹は 9000 個である。それらから 最終的な 1 つの系統樹を計算し出力した。

この他、いくつか条件を変えて実験を行ったが、結果に大きな変化は見られなかった。下記にその条件を示す。

・MCMC については、「各素性が等速度で変化すると仮定したモデルの場合」「各素性の変化速度がガンマ分布に従うと仮定したモデルの場合」の 2 つの場合で系統樹を生成した。図示したのは前者である。

・WALS のみのデータから系統樹を生成した。

3 つの手法で生成した系統樹は類似している。同じ語族の言語は系統樹の近い位置に出現する。これは、それぞれの手法の妥当性をある程度示していると言える。また、どの手法でも日本語に最も近いのは韓国語になった。

日本語の系統について、どの仮説が妥当か考えるために、それぞれの系統樹で日本語からの距離を計算する。各仮説に対応する 2 つないし 4 つの言語までの距離の平均を以下に示す。それぞれの手法で距離の定め方は異なるので、異なる手法の間で距離を比較するのは無意味である。

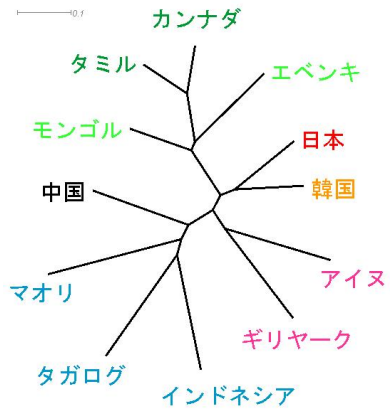

図 2: 近隣結合法

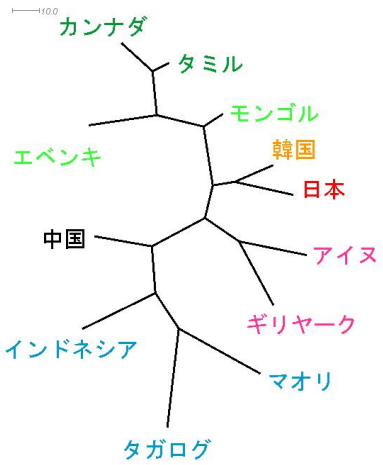

図 3: 最大節約法

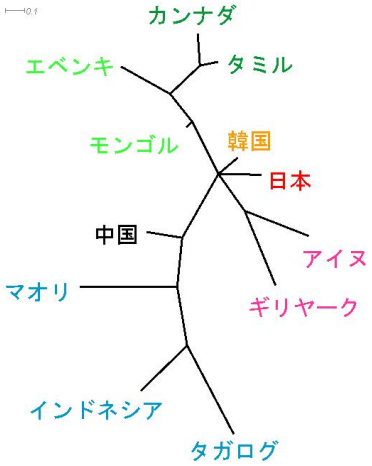

図 4: ベイズ法

手法	アルタイ	環太平洋	ドラヴィダ
近隣結合法	**0.423**	0.510	0.465
最大節約法	**76.6**	114.6	95.6
ベイズ法	**0.770**	1.319	1.039

どの手法についてもアルタイ語族が一番距離が短く、ついでドラヴィダ語族、環太平洋言語圏となっている。

7 おわりに

この結果からは、3 つの仮説の中では日本語がアルタイ語族の系統であることを 3 つの手法が共に示唆している。松本のデータは日本語と環太平洋言語圏との類似性を指摘するものであり、それを加味してもなおアルタイ語族が最も近い語族という結果になった。また、3 つの異なる手法で同様の結果が得られたことも確実性を高めたと言える。

とはいえ、この結果は言語間の親近性を示すものであり、当然ながらこれだけから日本語がアルタイ語族の一員であると断定はできない。実際にアルタイ語族の一つであるというためには、祖語の存在とそこからの分岐を説明する必要がある。また、アルタイ語族が語族であるのか、語族とはどうあるべきかについても再検討の余地があるだろう。これについてはさらなる研究を行う必要がある。

参考文献

[1] AWF Edwards and LL Cavalli-Sforza. The reconstruction of evolution.(Abstr.) Heredity 18: 553. *and Annals of Human Genetics*, Vol. 27, pp. 104–105, 1963.

[2] J. Felsenstein. PHYLIP-phylogeny inference package (version 3.2). *Cladistics*, Vol. 5, No. 1, pp. 164–166, 1989.

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

[4] R.D. Gray, A.J. Drummond, and S.J. Greenhill. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *Science*, Vol. 323, No. 5913, p. 479, 2009.

[5] Haspelmath, Martin & Dryer, Matthew S. & Gil, David & Comrie, and Bernard. *The World Atlas of Language Structures Online*. Available online at `http://wals.info/`.

[6] J.P. Huelsenbeck and F. Ronquist. MRBAYES: Bayesian inference of phylogenetic trees. *Bioinformatics*, Vol. 17, No. 8, pp. 754–755, 2001.

[7] D.H. Huson and D. Bryant. Application of phylogenetic networks in evolutionary studies. *Molecular biology and evolution*, Vol. 23, No. 2, p. 254, 2006. software available from `www.splitstree.org`.

[8] B. Mau, M.A. Newton, and B. Larget. Bayesian phylogenetic inference via Markov chain Monte Carlo methods. *Biometrics*, Vol. 55, No. 1, pp. 1–12, 1999.

[9] K. Rexová, Y. Bastin, and D. Frynta. Cladistic analysis of Bantu languages: a new tree based on combined lexical and grammatical data. *Naturwissenschaften*, Vol. 93, No. 4, pp. 189–194, 2006.

[10] F. Ronquist and J.P. Huelsenbeck. MrBayes 3: Bayesian phylogenetic inference under mixed models. *Bioinformatics*, Vol. 19, No. 12, p. 1572, 2003.

[11] N. Saitou and M. Nei. The neighbor-joining method: a new method for reconstructing phylogenetic trees. *Molecular biology and evolution*, Vol. 4, No. 4, p. 406, 1987.

[12] D. Sankoff. Minimal mutation trees of sequences. *SIAM Journal on Applied Mathematics*, Vol. 28, No. 1, pp. 35–42, 1975.

[13] 江原暉将. WALS データを用いた日本周辺言語の分析. 言語処理学会 第 15 回年次大会 発表論文集, 2009.

[14] 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.

[15] 服部四郎. 日本語の系統. 岩波書店, 1959.