言語類型の連続空間表現とその系統推定への応用

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

1 はじめに

日本語系統論において、類型論は最後の希望と言っても過言ではない。第一に、類型論が扱うのは語順や助数詞の有無といった特徴量であり、任意の言語対が比較できる。一方、伝統的な比較言語学や、基礎語彙を統計的に処理する言語年代学や計算系統学は、いずれも対象言語群の同源語の共有を前提とするが、日本語と他の言語との間で信頼できる同源語が充分に見つかる見込みは薄く、出発点にすら立てない [10]。第二に、類型論は千年、あるいは 1 万年のオーダの長期的変化を捉えられるかもしれない。信頼できる同源語集合の欠如は、日本語と他の言語との共通祖語の年代が大きくさかのぼることを示唆する [11] が、語順を含む類型論の特徴量の多くはまれにしか変化しない。

しかし、類型論による系統推定の可能性は未知数である。類型論の多種多様な特徴量が系統に沿ってどのように変化するか明らかでないからである。一方、比較言語学が依拠する音法則は青年文法学派の時代に確立されている。言語年代学や計算系統学も、語彙の生死という直感的に理解しやすい現象を扱う。

系統推定を目的として類型論を計算的に扱った従来研究は少なく、しかもデータの特性を無視している。各言語が特徴量の多値ベクトルで表現できるため、階層的クラスタリング自体は容易に行える。しかし、その結果が何を意味するか説明できない。Teh ら [9] が提案する系統樹は時間的な生成モデルであり、進化の過程を直接的に説明する。しかし、彼らは、多値ベクトルを 1-of-K 方式 (K 値の要素を K 個の 2 値要素に展開) で 2 値ベクトルに変換した上で、要素の独立性を仮定する。したがって、推定された祖語が 1-of-K 制約に反した論理的にありえない言語となり得る。

実は、ベクトル要素間の依存関係は、1-of-K 制約にとどまらない、言語的に重要な特性である。そもそも類型論の主要な目的の一つは普遍性の探究であり、特徴量間で一般に成り立つ関係が議論されてきた。例えば、OV 語順は形容詞-名詞の語順を示唆する [3]。とすれば、OV 語順が変化した場合に形容詞-名詞の語順はどうなるかという疑問がわく。この問題を考えるためにオーストロアジア語族を例にとる。表 1 に示すムンダ諸語とモン・クメール諸語は、系統関係が確立されているにもかかわらず、正反対と言って良いほど類型論的特徴が異なる。祖語はよりモン・クメール的だったと考えられることから、ムンダ諸語側での地滑り的変化が推定される [2]。この例を一般化すると、類型論の特徴量の歴史的変化は不自然な中間状態を避けるように起きるという仮説が得られる。

	ムンダ諸語	モン・クメール諸語
文法	総合的	分析的
語順	主辞後置, OV, 後置詞的	主辞前置, VO, 前置詞的
接辞	接頭・接中辞, 接尾辞	接頭・接中辞, 孤立的
融合	膠着的	融合的

表 1: オーストロアジア語族のムンダ諸語とモン・クメール諸語の類型論的比較 ([2] の表 1 から抜粋)

この仮説を系統推定に組み込むために、言語類型の連続空間表現を提案する。要素間の依存関係は、特徴量ベクトルに行列をかけ、多次元連続空間に写像することで捉えられる。さらに、連続空間上で言語の自然さを判定する。言語の自然さは、現代語を特徴量ベクトルの他のとり得る値から識別することで学習する。現代語から導き出された自然さが古代語にも成り立つと仮定すると (斉一性)、これは類型論的に自然な祖語を推定するのに役立つ。

この連続空間表現を系統推定に応用する。現代の諸言語の単一起源を仮定し、既知の語族群の上に二分木を構築することで世界言語の系統を推定する。多次元連続空間の各次元の系統上の安定性は既知の語族群から間接的に学習する。そして、推定された系統樹、世界祖語の特徴量、日本語の位置付け等を分析する。

2 データ

類型論のデータベースとして World Atlas of Language Structures (WALS) [4] を用いた。2014 年時点で、2,679 言語、192 種類の特徴量を収める。ただし、言語・特徴量ペアの被覆率は 15%以下である。WALS の提供する各語族の系統樹は 2 階層しかないため、代わりに Ethnologue [7] の階層的分類を用いた。両者の対応付けには ISO 639-3 言語コードを用いた。また、手話、混合言語、クレオール、未分類は削除した。子

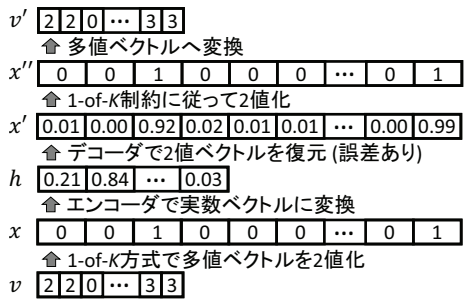


図 1: 各言語の特徴量の様々な表現

が 1 個しかない中間ノードは省き、孤立語は独立した語族として扱った。Ethnologue の分類には異論も少なくないが、特に変更は加えずそのまま用いた。

被覆率の特に低い言語、特徴量は削除した。系統推定には、内部で少なくとも 1 つの言語に被覆される特徴量が 30%以上となる 110 語族を用いた。また、少なくとも 10%の言語を被覆する 98 特徴量 (2 値化すると $d_0 = 539$ 次元) のみを残した。言語の自然さの学習には特徴量の 30%以上を被覆する 887 言語を用いた。

欠損値は R パッケージ missMDA [6] を用いて補完した。10 分割交差確認により既知の特徴量の予測精度を調べたところ、64.6%を得た。これはランダム (22.4%)、最頻値 (28.1%) を大きく上回っており、要素間の依存関係が確認できる。

3 言語の自然さ判定

図 1 に各言語の特徴量の表現を示す。$\boldsymbol{v}$ を多値特徴量ベクトルとすると、まず $\boldsymbol{v}$ を 2 値化し $\boldsymbol{x} \in \{0,1\}^{d_0}$ を得る。次に $\boldsymbol{h} = s(\boldsymbol{W}_{\mathrm{e}}\boldsymbol{x} + \boldsymbol{b}_{\mathrm{e}})$ により $\boldsymbol{x}$ を d_1 次元の実数ベクトル $\boldsymbol{h} \in [0,1]^{d_1}$ に写像する。ここで s はシグモイド関数、行列 $\boldsymbol{W}_{\mathrm{e}}$ およびベクトル $\boldsymbol{b}_{\mathrm{e}}$ は推定すべき重みである。パラメータの訓練は、このエンコーダに対応するデコーダ $\boldsymbol{x}' = s(\boldsymbol{W}_{\mathrm{d}}\boldsymbol{h} + \boldsymbol{b}_{\mathrm{d}})$ を用意し、クロスエントロピー最小化を目的関数 L_{AE} として行う。この訓練により、写像先でも元の情報が維持される。なお、$\boldsymbol{W}_{\mathrm{d}} = \boldsymbol{W}_{\mathrm{e}}^T$ とひも付ける。

言語の自然さは確率 $p(\boldsymbol{x})$ により判定する。

$$\begin{aligned} p(\boldsymbol{x}) &= \frac{\exp(\boldsymbol{W}_{\mathrm{s}}^T \boldsymbol{g})}{\sum_{\boldsymbol{x}'} \exp(\boldsymbol{W}_{\mathrm{s}}^T \boldsymbol{g}')} \\ \boldsymbol{g} &= s(\boldsymbol{W}_{\mathrm{l}}\boldsymbol{h} + \boldsymbol{b}_{\mathrm{l}}) \end{aligned}$$

$\boldsymbol{g} \in [0,1]^{d_2}$ で、$\boldsymbol{W}_{\mathrm{s}}$ はベクトル。$\boldsymbol{x}$ としてとり得る値は 2^{d_0} 通りあるが、その一部のみが自然であり、残りは不自然だったり、そもそも 1-of-K 制約に違反する。そこで、訓練事例の対数確率最大化を目的関数とする。オートエンコーダと自然さ判定を組み合わせると、最大化すべき最終的な目的関数は $\sum_{i=1}^{N}(-L_{\mathrm{AE}}(\boldsymbol{x}_i, \boldsymbol{x}_i') + C \log p(\boldsymbol{x}_i))$ となる。ここで、N は訓練事例数、C はある定数。訓練事例の欠損値は補完されているとする。訓練アルゴリズムとして、勾配法に基づく AdaGrad を採用する。なお、$\log p(\boldsymbol{x})$ の偏微分は計算困難なため、contrastive divergence [5] により近似する。

4 系統推定

4.1 系統樹モデル

系統推定を多次元連続空間上で行う。基本戦略は、通常のクラスタリングと同様に、空間上で近い言語群同士を結合していくことだが、2 つの点で改良を施す。第一に、自然さ判定を組み込むことにより、類型論的に自然な祖語を推定する。第二に、連続空間の各次元について、系統に沿った歴史的変化の安定性を考慮する。そのために、連続空間上での親から子への移動をガウス分布でモデル化する。いま、第 k 次元における親ノードの値を h_k^{P} とすると、子の値は平均を h_k^{P}、精度 (分散の逆数) を λ_k とするガウス分布に従う。つまり、親から離れるほど小さな確率を得る。安定性は精度により制御される。λ_k の事前分布としてガンマ分布 $(\alpha = \beta = 0.1)$ を置き、共役性を利用して λ_k を積分消去すると、事後ハイパーパラメータ $\alpha_n = \alpha + n/2$ および $\beta_n = \beta + \frac{1}{2}\sum_{i=1}^{n} m_i^2$ を得る。ここで n はサンプル数、m_i は i 番目のノードとその親との距離 $h_k - h_k^{\mathrm{P}}$ である。事後予測確率は t 分布 $p_k(h_k | h_k^{\mathrm{P}}, M_{\mathrm{hist}}, \alpha, \beta) = t_{2\alpha_n}(h_k | h_k^{\mathrm{P}}, \sigma^2 = \beta_n/\alpha_n)$ となる、ここで M_{hist} は、α、β およびそれまでに観測した距離の組である。親から子への移動確率 $p_{\mathrm{MOVE}}(\boldsymbol{h} | \boldsymbol{h}^{\mathrm{P}}, M_{\mathrm{hist}})$ は各次元の移動確率の積である。根の値は一様分布から得る。

系統樹の確率 $\boldsymbol{\tau}$ は 2 つの表現、すなわち (1) 各ノードの自然さの確率および (2) ノード間の移動確率の幾何平均とする。

$$\Big(\prod_{\boldsymbol{h} \in \mathrm{nodes}(\boldsymbol{\tau})} p(\boldsymbol{h}) \times \mathrm{Uniform}(\boldsymbol{root}) \times \prod_{(\boldsymbol{h}, \boldsymbol{h}^{\mathrm{P}}) \in \mathrm{edges}(\boldsymbol{\tau})} p_{\mathrm{MOVE}}(\boldsymbol{h} | \boldsymbol{h}^{\mathrm{P}}, M_{\mathrm{hist}}) \Big)^{1/2}$$

ここで、$p(\boldsymbol{h})$ は自然さ判定の確率 ($\boldsymbol{h}$ が対応する 2 値ベクトル $\boldsymbol{x}$ を持つとする)、$\mathrm{nodes}(\boldsymbol{\tau})$ は $\boldsymbol{\tau}$ のノードの集合、$\mathrm{edges}(\boldsymbol{\tau})$ は $\boldsymbol{\tau}$ の辺の集合とする。M_{hist} は実際にはノードを観測するたびに更新される。

4.2 推論

観測データは (1) 各葉ノードの特徴量ベクトル (ただし欠損値を除く)、および (2) 各語族の木のトポロジーからなる。推定すべきは、(1) 葉ノードの欠損値、(2) 根を含む内部ノード、および (3) 語族より上の木のトポロジーからなる。推論の際、葉ノードは多値ベクトル $\boldsymbol{v}$、内部ノードは実数ベクトル $\boldsymbol{h}$ を一次表現と

する。ただし、自然さは $\boldsymbol{x}$ に対して定義されるため、$\boldsymbol{h}$ はデコーダを用いて適宜近傍の $\boldsymbol{v}'$ に変換する。

推論は Markov chain Monte Carlo 法により行う。紙面の都合上、概略のみを述べる。葉ノードの各欠損値は Gibbs サンプリング、内部ノードの各次元の値はガウス分布を提案分布とする Metropolis 法で確率的に選択する。また、可動性を高めるため、内部ノード全体に対しては、親あるいは子のいずれかとの線形補間を提案分布とする Metropolis-Hastings 法も適用する。木のトポロジーには、Li ら [8] の手法に基づき局所的に木を変更する Metropolis-Hastings 法を適用する。

初期値として、葉ノードの値は欠損値補完を用いる。内部ノードの値は子の値の (非欠損値率による重み付き) 平均とする。木のトポロジーは距離に基づく凝集型クラスタリングにより構築する。

5　世界系統樹の推定実験

自然さ判定のパラメータの次元は $d_1 = 100$、$d_2 = 10$ とした。推論は 3,000 回の焼き入れのあと、10 回間隔で 100 サンプルを得た。これを乱数の種を変えて 3 回行い、合計 300 サンプルを得た。

図 2 に世界の系統樹の推定結果を示す。3 文字のラベルは ISO 639-3 言語コードを表す。表示の都合上、語族以下の系統は省略した。複数のサンプルの要約には最大系統群信頼度木 (maximum clade credibility tree) を用いた。枝に付した数値は該当系統群がサンプルにおいて支持された割合を表す。全体的に低い支持率が示すように、推定結果は非常に不安定である。図では日本語族 (Japonic。今回のデータでは本土方言と沖縄の首里方言からなる) はアルタイ語族と兄弟関係にあるが、支持率は 13%にすぎない。世界全体としては、(サンプル間で出入りがあるため支持率に反映されていないが) 大きく 3 つの集団が観測できる。1 つ目はアフロ・アジア語族、ナイル・サハラ語族、ニジェール・コンゴ語族、タイ・カダイ語族、オーストロネシア語族を含み、図の中央部を占める。2 つ目は図の左側で、ウラル語族、セイリシュ語族を含み、アイヌ語とブルシャスキー語はここに属している。最後は図の右側で、日本語、アルタイ語族、ドラヴィダ語族、北コーカサス語族を含む。アメリカ大陸やパプア・ニューギニアの言語は 3 集団に分散した。

連続空間表現を分析するために、2 つの言語の混ぜあわせを試した。上述のムンダ諸語とモン・クメール諸語の場合を図 3 に示す。横軸は左端がムンダリ語、右端がクメール語で、その間が両者の混合である。縦軸は自然さ判定の対数確率 + 定数を表す。多値特徴量の混合 (赤) では、指定の混合率で確率的にムンダリ

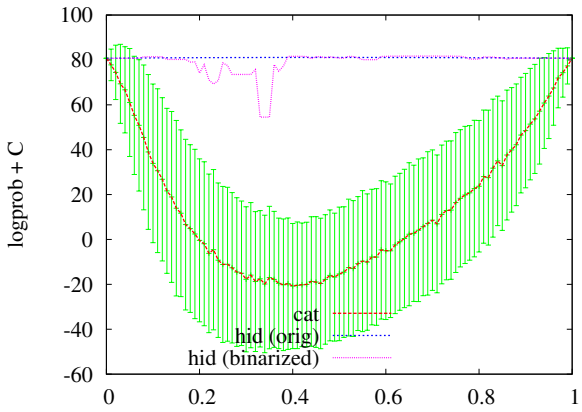


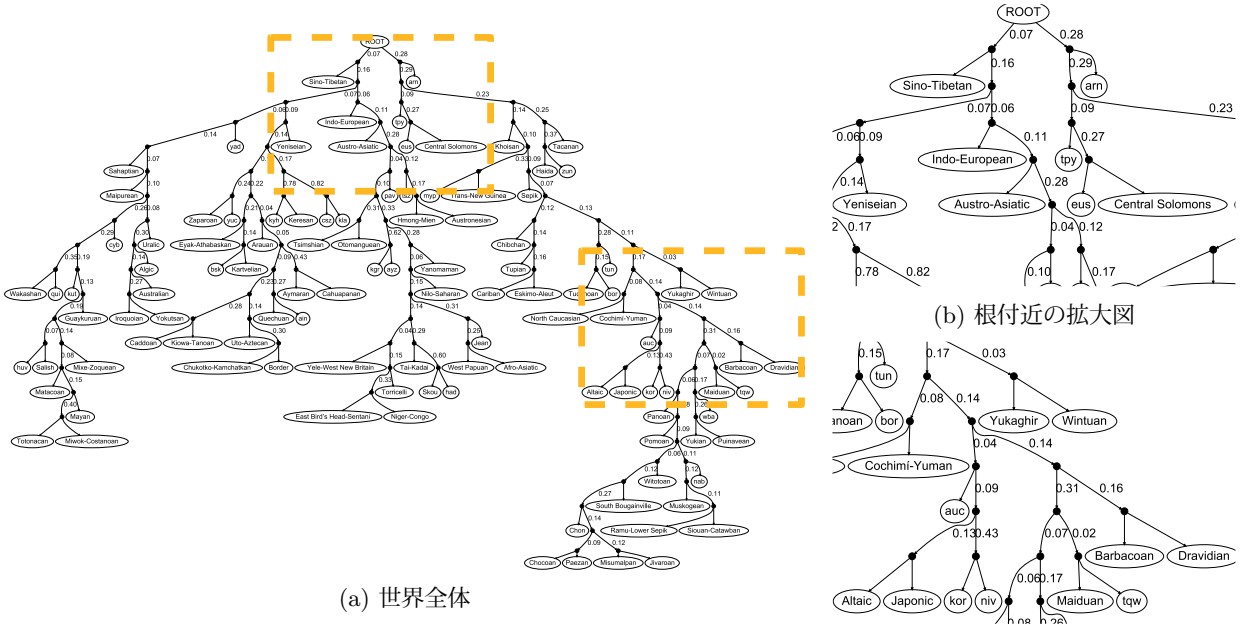
図 3: ムンダリ語 (ムンダ諸語) とクメール語 (モン・クメール諸語) の混合と自然さとの関係

特徴量	支持率/値
Consonant Inventories	95 Moderately small 5 Small
Vowel Quality Inventories	87 Average (5-6) 12 Large (7-14) 0 Small (2-4)
Syllable Structure	99 Moderately complex 1 Complex
Coding of Nominal Plurality	97 Plural suffix 1 Plural word 1 No plural 0 Mixed morph. plural 0 Plural clitic
Ord. of Numeral and Noun	58 Noun-Numeral 42 Numeral-Noun
Ord. of Subject, Object and Verb	53 SOV 46 SVO 1 No dominant order
Ord. of Adposition and Noun Phrase	87 Postpositions 13 Prepositions
Ord. of Adjective and Noun	85 Noun-Adjective 15 Adjective-Noun

表 2: 世界祖語の特徴量

語の特徴量の値をクメール語のそれで置き換えた。この操作を各混合率に対して 1,000 回行い、平均と標準偏差 (エラーバー) を示した。両言語の間で谷が生じており、混合が不自然と判定されてしまっていることがわかる。一方、連続空間表現 (ピンク) では、両言語の線形補間を行い、デコーダを用いて近傍の $\boldsymbol{v}'$ に変換した。若干のゆらぎがあるものの、自然さを保ったままでの遷移が実現されている。この結果は、元の多値特徴量を用いて系統推定を行うと類型論的に不自然な祖語を推定しがちである一方、連続空間表現を用いると自然な祖語を推定しやすいことを示唆する。特徴量変化の詳細を見ると、0.46–0.47 で後置詞型から前置詞型へ、0.53–0.54 で強い接尾辞型から弱い接辞型へ、0.60–0.61 で SOV から SVO へ変化していた。

表 2 に世界祖語の主要な特徴量の値とサンプルにおける支持率を示す。音韻体系は平均的である。文法的には日本語と同じ SOV 語順・後置詞型が優勢だが、一方でチベット語のような後置修飾が支持されている。



(a) 世界全体

(b) 根付近の拡大図

(c) 日本語族周辺の拡大図

図 2: 推定された世界の系統樹

順位	言語	分類	対数確率
1	(日本)	日本	76.8
2	首里	日本	-32.7
3	ハルハ	アルタイ > モンゴル	-198.4
4	レプチャ	シナ・チベット > チベット・ビルマ	-201.6
5	チュヴァシ	アルタイ > テュルク	-208.8
6	オルドス	アルタイ > モンゴル	-213.5
7	ブリヤート	アルタイ > モンゴル	-216.9
8	デウリ	シナ・チベット > チベット・ビルマ	-221.0
9	ウルム	アルタイ > テュルク	-227.6
10	ウィトト	ウィトト	-229.1
159	朝鮮	(孤立)	-281.2
399	アイヌ	(孤立)	-389.5

表 3: 日本語と他の現代語との近さ

表 3 に日本語と他の現代語との近さを示す。ここで、近さは各サンプルから計算された p_{MOVE} の幾何平均であり、連続空間上で系統上の安定性を考慮している。この順位と多値ベクトルの一致率による順位との Spearman の順位相関係数は 0.76 であり、かなりの変動がある。系統上の安定性を考慮すると、朝鮮語やアイヌ語が日本語とそれほど近くない一方、ヒマラヤ以南のチベット・ビルマ系言語が上位に来ることは興味深い。シナ・チベット語族の系統樹は、インド東北部の小言語の調査次第で根本的に書き換えられる可能性があり [1]、今後の研究の進展が期待される。順位の最下位層はオーストロネシア語族のマレー・ポリネシア語派とナイル・サハラ語族が占めた。

6 おわりに

系統推定における類型論の有効性は未知数である。本稿では、研究の第一歩として、言語的に自然な祖語の推定に取り組んだ。オーストロアジア語族を用いた小実験は、提案する連続空間上では、点 (ノード) だけでなく線 (変化の経路) も言語的自然さを保つ可能性を示唆するが、これはさらなる裏付けが必要である。今後は、年代や地理位置をモデルに組み込むとともに、変化の速さや方向性も検討したい。それと同時に、系統論のみならず、通時類型論全般の研究を進展させるために、記録に残る古代語のデータで類型論データベースを拡充させる必要性を訴えたい。

謝辞

本研究は一部 JSPS 科研費 26730122 の助成を受けた。

参考文献

[1] Roger Blench and Mark W. Post. Rethinking Sino-Tibetan phylogeny from the perspective of North East Indian languages. In Nathan Hill and Tom Owen-Smith, editors, *Trans-Himalayan Linguistics*, pp. 71–104. De Gruyter, 2013.

[2] Patricia Donegan and David Stampe. Rhythm and the synthetic drift of Munda. In Rajendra Singh, editor, *The Yearbook of South Asian Languages and Linguistics*, pp. 3–36. Mouton de Gruyter, 2004.

[3] Joseph H. Greenberg, editor. *Universals of language* . MIT press, 1963.

[4] Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. *The World Atlas of Language Structures*. Oxford University Press, 2005.

[5] Geoffrey E. Hinton. Training products of experts by minimizing contrastive divergence. *Neural Computation*, Vol. 14, No. 8, pp. 1771–1800, 2002.

[6] Julie Josse, Marie Chavent, Benot Liquet, and François Husson. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification*, Vol. 29, No. 1, pp. 91–116, 2012.

[7] M. Paul Lewis, Gary F. Simons, and Charles D. Fennig, editors. *Ethnologue: Languages of the World, 17th Edition*. SIL International, 2014.

[8] Shuying Li, Dennis K. Pearl, and Hani Doss. Phylogenetic tree construction using Markov chain Monte Carlo. *Journal of the American Statistical Association*, Vol. 95, No. 450, pp. 493–508, 2000.

[9] Yee Whye Teh, Hal Daumé III, and Daniel Roy. Bayesian agglomerative clustering with coalescents. In *NIPS*, pp. 1473–1480, 2008.

[10] Alexander Vovin. *Koreo-Japonica*. University of Hawai'i Press, 2010.

[11] 服部四郎. 日本語の系統. 岩波書店, 1999.