特徴間の依存関係を考慮した基本語順の史的変化の分析

村脇 有吾
京都大学大学院情報学研究科
murawaki@i.kyoto-u.ac.jp

1 はじめに

歴史比較言語学において伝統的に利用されてきた手がかりは、同源語、その背後にある音法則、語形変化表の変化など、広い意味で語彙的なものであり、構造的特徴の史的変化は周辺的地位にとどまる。インド・ヨーロッパ語族のように資料に恵まれた語族であっても、祖語の再構は音韻論や形態論に偏りがちであり、基本語順や関係節の構造といった統語的特徴に関しては不明な点が多い。その理由として、語彙が音法則という論証に適した強力な手がかりに支えられているのに対し、構造的特徴の変化は不確実であり、本質的に人手では扱いづらいことが挙げられる。

不確実な手がかりを効果的に扱えるのは、計算集約的な統計モデルである。近年は元々生物進化の分析用に開発された確率的系統樹モデルの言語データへの転用が進んでおり、言語の構造的特徴の史的変化を分析した事例も見られる [6, 2, 11]。確率的系統樹モデルでは、言語は系統樹に沿って進化し、その状態は連続時間マルコフ連鎖 (CTMC) モデルにしたがって時間とともに確率的に変化すると仮定する。このモデルのもとで年代つき系統樹と葉ノードの状態が与えられたとき、根を含む内部ノードの状態と、CTMC の遷移率行列が推定される。この遷移率行列を応用すると、例えば、与えられた言語の特徴が一定時間の後どの値を取るかが確率的に予測できる [11]。

ここで問題となるのは、構造的特徴には依存関係があり、連鎖的な変化が起こりうることである [5]。Greenhill ら [6] は特徴間の独立性を仮定しており、全体として不自然な変化を推論しているおそれがある。Dunn ら [2] は 2 値特徴対の依存関係を分析している。彼らは、特徴対が取りうる値の組み合わせ (2 値特徴なので $2 \times 2 = 4$ 値) を展開し、4×4 の遷移率行列を推定している。この手法は組み合わせ爆発を起こすため、多値特徴や 3 個以上の特徴の依存関係を扱うのは現実的ではない。

そこで、本稿では、筆者が以前提案した言語の潜在表現 [12] を用いた分析手法を提案する (図 1)。この手法では、各言語の表層表現 (多値特徴列) の背後に 2 値パラメータ列という潜在表現を仮定する。パラメータは仮定により互いに独立であり、1 個のパラメータを変化させると依存関係にある複数の表層的特徴が変化する。この潜在空間上で史的変化を推論すれば、暗黙のうちに特徴間の依存関係が扱える。潜在表現は分析の必要に応じて表層表現に戻せばよい。

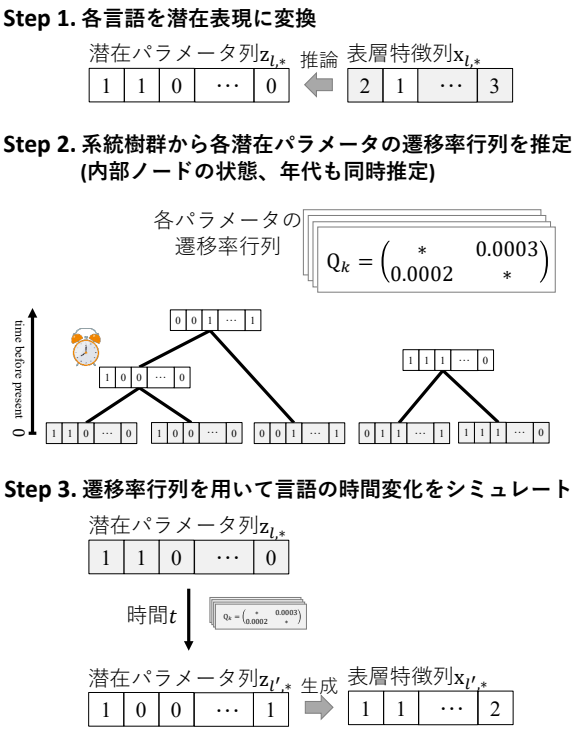

図 1: 分析手順 (灰地は観測変数、白地は潜在変数)

本稿が提案するのは一般的な分析手法だが、概念実証のための具体例として基本語順 (主語 (S)、動詞 (V)、目的語 (O) の語順) を取り上げる。世界における基本語順の分布には心理言語学的関心が寄せられている。例えば、SOV が最も一般的なのはそれが最適だからという説や、SOV から SVO への史的変化がその逆よりも多いという推測をもとにした、SVO が SOV よりも機能的に優れているという説、パントマイム実験における SOV への選好から原型言語における SOV 語順を想定する説などが入り乱れている [11]。しかし、このように基本語順のみに着目する取り組みには限界があると筆者は考えている。構文構造だけではなく、

情報構造も語順を伝達手段として用いる。形態論的に貧弱で、もっぱら語順を使って構文構造を伝える言語と、構文構造を伝える形態論的手段が豊富で、情報構造に応じて語順が柔軟に変わる言語とでは、基本語順の史的安定性も異なるかもしれない。本稿の分析は、このような推測に定量的裏付けを与える。

2 言語の潜在表現

構造的特徴のデータベースとして WALS [9] を用いる。前処理の結果、言語数 $L = 2,607$、特徴数 $N = 104$ を得る。ただし、73.1% の要素が欠損値である。特徴は一般に多値である。特徴 i の取る値の異なり数を F_i とすると、多値特徴は $(1,1),(1,2),\cdots,(1,F_1),(2,1),\cdots,(i,j),\cdots,(N,F_N)$ のように索引づけされる。これをならした索引を $1,\cdots,m,\cdots,M$ とする ($M = \sum_{i=1}^{N} F_i$)。この 2 種類の索引を関数 $f(i,j) = m$ により対応づける。

筆者の提案したモデル [12] では、言語 l の表層表現 $\mathrm{x}_{l,*} \in \mathbb{N}^N$ は潜在表現 $\mathrm{z}_{l,*} \in \{0,1\}^K$ から確率的に生成されると仮定する。K は事前に決めるパラメータ数であり、本稿では 100 とする。$\mathrm{z}_{l,*}$ と $\mathrm{x}_{l,*}$ との間は、特徴間の依存関係を捉える重み行列 $W \in \mathbb{R}^{K\times M}$ が取り持つ。まず $\tilde{\theta}^{\mathrm{T}}_{l,*} = \mathrm{z}^{\mathrm{T}}_{l,*} W$ により特徴スコア列 $\tilde{\theta}_{l,*} \in \mathbb{R}^M$ を得る。次に $\tilde{\theta}_{l,*}$ を特徴ごとに正規化することで特徴選択確率列 $\theta_{l,*} \in [0,1]^M$ を得る。

$$\theta_{l,f(i,j)} = \frac{\exp(\tilde{\theta}_{l,f(i,j)})}{\sum_{j'} \exp(\tilde{\theta}_{l,f(i,j')})}$$

最後に $P(x_{l,i}|\theta_{l,*}) = \theta_{l,f(i,x_{l,i})}$ にしたがって各表層的特徴を生成する。なお、パラメータ k の列 $\mathrm{z}_{*,k}$ は自己ロジスティックモデルから生成される。これにより系統的・地理的に近い言語が同じ値を取りやすくなる。詳細は文献 [12] を参照のこと。

推論時 (図 1 Step 1) には、観測値 $x_{l,i}$ が与えられ、Gibbs サンプリングにより、自己ロジスティックモデルを制御する変数の集合 A、重み行列 W、欠損値 $x_{l,i}$ とあわせて $\mathrm{z}_{l,*}$ を求める。本稿では焼き入れ 1,000 反復後の 1 サンプルを用いる。生成時 (図 1 Step 3 後半) には、$\mathrm{z}_{l',*}$ が与えられ、上述の生成過程にしたがって $\mathrm{x}_{l',*}$ を得る。ただし、以降の分析では、実際には $\mathrm{x}_{l',*}$ の代わりに $\theta_{l',*}$ を用いる。

3 連続時間マルコフ連鎖モデル

言語系統樹に沿ったパラメータ列の史的変化を推論することで、連続時間マルコフ連鎖 (CTMC) モデルの遷移率行列を得る (図 1 Step 2)。本稿で用いる CTMC モデルは連続時間における離散状態の変化を扱い、遷移率行列 Q_k で制御される。状態数が 2 の場合、遷移率行列は 2×2 で $Q_k = \begin{pmatrix} -\alpha_k & \alpha_k \\ \beta_k & -\beta_k \end{pmatrix}$ と表される。$\alpha_k, \beta_k > 0$ の事前分布としてガンマ分布をおく。遷移率行列 Q_k を用いると、子言語 l のパラメータ k の値 $z_{l,k}$ の確率は、親言語 $\pi(l)$ と時間 t に条件づけられ、$P(z_{l,k} = b | z_{\pi(l),k} = a, t) = \exp(tQ_k)_{a,b}$ となる。2×2 の行列指数関数は以下のように解析的に求まる: $\exp(tQ_k) = \frac{1}{\alpha_k+\beta_k} \begin{pmatrix} \beta_k + \alpha_k e^{-(\alpha_k+\beta_k)t} & \alpha_k - \alpha_k e^{-(\alpha_k+\beta_k)t} \\ \beta_k - \beta_k e^{-(\alpha_k+\beta_k)t} & \alpha_k + \beta_k e^{-(\alpha_k+\beta_k)t} \end{pmatrix}$。なお、親を持たない根の各パラメータは CTMC の定常分布から生成する。

言語系統樹として Glottolog 3.0 [8] の階層的分類を用いる。Glottolog を採用したのは (1) 系統的分類を意識し、(2) 最近の研究成果をよく反映し、(3) WALS との対応づけが容易だからである。WALS 中の言語から Glottolog の系統樹と対応づかないものを除くと 2,545 個が残る。対応する言語が得られた語族 (系統樹) の数は 309 である。ただし、そのうち 155 語族は所属する言語が 1 個 (つまり孤立語) である。各系統樹からは WALS にない葉ノードや子を 1 個しか持たない内部ノードを取り除く。

文献 [2] が資料の豊富な 4 語族だけを実験に用いたのに対し、本稿ではできる限り多くの語族を利用する。その反面、4 語族については語彙データにもとづく年代つき系統樹が得られるのに対し、Glottolog の内部ノードの年代は不明であり、木の形 (トポロジー) しか得られない。そこで、代替手段として、一部の内部ノードのおおよその年代を事前分布として与え、年代較正を行う (図 1 Step 2 の時計)。較正に用いるのは二次文献 [10, 1, 4, 11, 7] から集めた 50 点である。事前分布としては正規分布、対数正規分布や最大最小値つきの連続一様分布を用いる。

推論時には、諸系統樹のトポロジーと葉ノードの状態、較正用の年代が与えられ、根を含む内部ノードの状態と年代、各パラメータの遷移率行列 (α_k, β_k) を Gibbs サンプリングにより求める。各パラメータの値は、動的計画法 [3] を用いて系統樹単位で一度に更新する。各内部ノードの年代は、自身の親 (あれば) と最も古い子に規定される時間幅の中から slice sampling [13] によって得る。各パラメータの α_k, β_k は Hamiltonian Monte Carlo [14] により一度に更新する。焼き入れ 1,000 反復の後、10 反復間隔で 1 サンプル、合計 10 サンプルを得る。

4 基本語順の安定性のばらつき

提案手法の概念実証のための具体例として、基本語順 (WALS の特徴 81A) を分析する。まず、2,545 個の

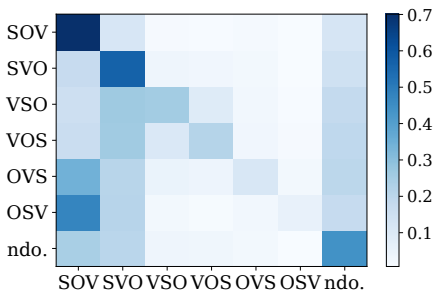

図 2: 基本語順の遷移確率 ($t = 2,000$)

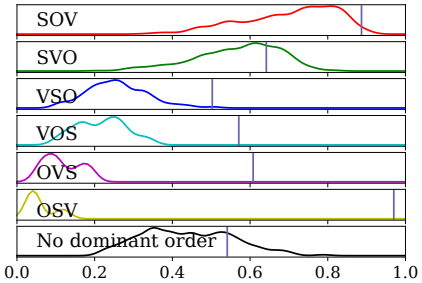

図 3: 基本語順維持確率の分布 ($t = 2,000$)

現代語のうち、基本語順が観測されている言語 1,350 個を選択する。遷移率行列を用いて、各言語の潜在表現 $z_{l,*}$ から時間 t 後の状態 $z_{l',*}$ を確率的に生成する (図 1 Step 3)。次に、重み行列 W を適用して基本語順の生成確率 $\theta_{l',f(i,j)}$ を得る。シミュレーションは各言語につき 1,000 回、これを遷移率行列 10 サンプルそれぞれについて行う。最後に言語ごとに基本語順の生成確率の算術平均を得る。

基本語順の生成確率を現在の語順ごとに集約した結果を図 2 に示す。行列の要素 (a, b) は、現在の語順が a のとき、2,000 年後の語順が b である確率を表す。現代の基本語順の頻度分布に対応するように、SOV が最も安定的で、SVO がそれに続くという結果を得た。VSO は SOV よりも SVO に変化しやすいという結果は文献 [11] と一致する。

次に、言語ごとの安定性のばらつきを図 3 に示す。各語順について、各言語が 2,000 年後に同じ基本語順を維持する確率をカーネル密度推定を用いて平滑化した。同じ基本語順であっても言語によって安定性がばらつくことが確認できる。

比較のために基本語順の遷移率行列 (7×7) を推定する。具体的には、潜在表現の遷移率行列の推論時に得られた年代つき系統樹と現代語の状態を与え、内部ノードの状態と同時に遷移率行列を推定する。この遷移率行列にもとづく基本語順維持確率を図 3 中の縦棒で示す。潜在表現にもとづく推定とくらべて高い確率が得られる傾向があり、特に OVS、OSV については乖離が著しい。表層表現から潜在表現へ、潜在表現から表層表現への 2 回の確率的変換により不確実性が増幅することは避けがたいが、その過程で極端に低頻度な基本語順を切り捨てている可能性がある。

重み	説明変数 (特徴とその値の組)
0.0489	93A Position of Interrogative Phrases in Content Questions: Not initial interrogative phrase
0.0486	86A Order of Genitive and Noun: Genitive-Noun
0.0402	85A Order of Adposition and Noun Phrase: Postpositions
0.0311	51A Position of Case Affixes: Postpositional clitics
0.0283	37A Definite Articles: No definite, but indefinite article
0.0222	26A Prefixing vs. Suffixing in Inflectional Morphology: Weakly suffixing
0.0115	37A Definite Articles: Definite affix
0.0044	26A Prefixing vs. Suffixing in Inflectional Morphology: Strongly suffixing
−0.0021	37A Definite Articles: No definite or indefinite article
−0.0175	14A Fixed Stress Locations: Initial
−0.0193	89A Order of Numeral and Noun: Numeral-Noun
−0.0291	120A Zero Copula for Predicate Nominals: Impossible

表 1: SOV 語順維持確率の回帰分析 (抜粋)

重み	説明変数 (特徴とその値の組)
0.0537	84A Order of Object, Oblique, and Verb: VOX
0.0432	26A Prefixing vs. Suffixing in Inflectional Morphology: Little affixation
0.0388	88A Order of Demonstrative and Noun: Noun-Demonstrative
0.0281	85A Order of Adposition and Noun Phrase: Prepositions
0.0200	26A Prefixing vs. Suffixing in Inflectional Morphology: Strong prefixing
0.0166	86A Order of Genitive and Noun: Noun-Genitive
0.0111	14A Fixed Stress Locations: No fixed stress
0.0067	38A Indefinite Articles: No indefinite, but definite article
0.0013	37A Definite Articles: Definite affix
−0.0079	51A Position of Case Affixes: Case suffixes
−0.0079	37A Definite Articles: Definite word distinct from demonstrative
−0.0107	86A Order of Genitive and Noun: Genitive-Noun
−0.0108	38A Indefinite Articles: No definite or indefinite article
−0.0137	14A Fixed Stress Locations: Initial

表 2: SVO 語順維持確率の回帰分析 (抜粋)

基本語順を共有する言語のうち、どういった言語がより安定的だろうか。動態を制御するすべての変数が手中にあるものの、それらの複雑な依存関係を解析するのは容易ではない。この点は、近年のニューラルネットワークにおける解釈可能性の問題と共通する。本稿では、複雑なモデルの振る舞いをより単純な別のモデルに説明させるという解法を採る。具体的には、L1 正則化つき線形回帰 (lasso) を用いる。目的変数を 2,000 年後の基本語順維持確率とし、説明変数を現代の表層表現 (2 値化特徴列) とする。正則化項の重みは 3 分割交差確認で求める。

SOV 言語と SVO 言語の回帰分析結果をそれぞれ表

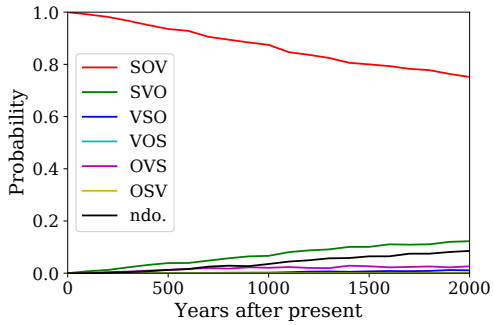

図 4: 日本語の基本語順の進化の予測

1、2 に示す。SOV 言語では主要部後置型の語順 (属格・名詞語順、後置詞)、SVO 言語では主要部前置型の語順 (名詞・指示詞語順、前置詞、名詞・属格語順) と基本語順維持確率に正の相関が見られる。ただし、SOV 言語の数詞・名詞語順は予想に反して負の重みを得た。同様に、SOV 言語は接尾辞や後置接語、SVO 言語は接頭辞との正の相関が見られる。SVO 言語について接辞の乏しさ (little affixation) が高い重みを得ていることは注目に値する。この結果は、形態論的手段によらずに構文構造を伝える言語にとって SVO 語順が適している可能性を示唆する。語順によらずに情報構造を伝える手段である冠詞に着目すると、SVO 語順の安定性と定冠詞の使用に正の相関が見られるが、その重みは小さい。SOV 語順についてははっきりした傾向は確認できなかった。

5 日本語のシミュレーション

個別言語の例として日本語を分析する。日本語の今後の基本語順のシミュレーション結果を図 4 に示す。各時間 t について、遷移率行列 10 サンプルそれぞれについて 2,000 回シミュレーションを行い、基本語順の生成確率の算術平均を求めた。日本語が 2,000 年後に SOV 語順である確率は 75.2%となり、SVO の 12.3%、支配的語順なしの 8.5%がそれに続く。これは SOV 言語の中では平均的な値である。

SOV 語順を維持する場合や SVO 語順に変化する場合、将来の日本語はどのような性格を持つだろうか。この疑問に答えるために再び回帰分析を行う。各シミュレーション結果 l' について、目的変数を注目する基本語順の選択確率とし、説明変数を特徴選択確率列 $\theta_{l',*}$ とする。ただし、基本語順そのものや、基本語順への依存が自明な特徴 [15] は説明変数から除外する。

SOV 語順を維持する場合の回帰分析結果では、おおむね現在の日本語と同じ値が高い重みを得た。ただし、特徴 116A 極性疑問文は現在の疑問不変化詞ではなく疑問動詞形態に高い重みが与えられた (沖縄の首里語は疑問動詞形態に分類されている)。SVO 語順へ変化する場合の結果を表 3 に示す。日本語は現在は「が」、「を」のような後置接語を用いるが、最上位の説明変数は、SVO 語順へ変化する場合は英語や中国語のような孤立語化を伴う可能性を示唆する。

重み	説明変数 (特徴とその値の組)
0.0977	51A Position of Case Affixes: No case affixes or adpositional clitics
0.0733	94A Order of Adverbial Subordinator and Clause: Initial subordinator word
0.0566	90A Order of Relative Clause and Noun: Noun-Relative clause
0.0556	17A Rhythm Types: No rhythmic stress
0.0383	44A Gender Distinctions in Independent Personal Pronouns: 3rd person only, but also non-singular

表 3: 日本語の SVO 語順への変化の回帰分析 (上位 5)

6 おわりに

本稿では、潜在表現を用いることで特徴間の依存関係を暗黙的に扱う史的変化の分析手法を提案した。本稿では概念実証のための例として基本語順に着目したが、他の特徴の分析も実りあるものと予想する。

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] Remco Bouckaert, et al. Mapping the origins and expansion of the Indo-European language family. *Science*, Vol. 337, No. 6097, pp. 957–960, 2012.
[2] Michael Dunn, Simon J. Greenhill, Stephen C. Levinson, and Russell D. Gray. Evolved structure of language shows lineage-specific trends in word-order universals. *Nature*, Vol. 473, No. 7345, pp. 79–82, 2011.
[3] Joseph Felsenstein. Evolutionary trees from DNA sequences: A maximum likelihood approach. *Journal of Molecular Evolution*, Vol. 17, No. 6, pp. 368–376, 1981.
[4] Russell D. Gray, Alexei J. Drummond, and Simon J. Greenhill. Language phylogenies reveal expansion pulses and pauses in Pacific settlement. *science*, Vol. 323, No. 5913, pp. 479–483, 2009.
[5] Joseph H. Greenberg. Diachrony, synchrony and language universals. In Joseph H. Greenberg, Charles A. Ferguson, and Edith A. Moravesik, editors, *Universals of human language*, Vol. 1. Stanford University Press, 1978.
[6] Simon J. Greenhill, Quentin D. Atkinson, Andrew Meade, and Russel D. Gray. The shape and tempo of language evolution. *Proc. of the Royal Society B*, Vol. 277, No. 1693, pp. 2443–2450, 2010.
[7] Rebecca Grollemund, Simon Branford, Koen Bostoen, Andrew Meade, Chris Venditti, and Mark Pagel. Bantu expansion shows that habitat alters the route and pace of human dispersals. *PNAS*, Vol. 112, No. 43, 2015.
[8] Harald Hammarström, Robert Forkel, Martin Haspelmath, and Sebastian Bank, editors. *Glottolog 3.0*. Max Planck Institute for the Science of Human History, 2017.
[9] Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. *The World Atlas of Language Structures*. Oxford University Press, 2005.
[10] Eric W. Holman, et al. Automated dating of the world's language families based on lexical similarity. *Current Anthropology*, Vol. 52, No. 6, pp. 841–875, 2011.
[11] Luke Maurits and Thomas L. Griffiths. Tracing the roots of syntax with Bayesian phylogenetics. *PNAS*, Vol. 111, No. 37, pp. 13576–13581, 2014.
[12] Yugo Murawaki. Diachrony-aware induction of binary latent representations from typological features. In *Proc. of IJCNLP*, pp. 451–461, 2017.
[13] Radford M. Neal. Slice sampling. *Annals of Statistics*, Vol. 31, No. 3, pp. 705–767, 2003.
[14] Radford M. Neal. MCMC using Hamiltonian dynamics. In Steve Brooks, Andrew Gelman, Galin L. Jones, and Xiao-Li Meng, editors, *Handbook of Markov Chain Monte Carlo*, pp. 113–162. CRC Press, 2011.
[15] Hiroya Takamura, Ryo Nagata, and Yoshifumi Kawasaki. Discriminative analysis of linguistic features for typological study. In *Proc. of LREC*, pp. 69–76, 2016.