村脇 有吾. 基本語順の歴史的変化の数理モデル. 数学セミナー, 2020 年 3 月号, pp. 36–40. Author Manuscript

基本語順の歴史的変化の数理モデル

村脇 有吾
京都大学大学院情報学研究科

英語の他動詞文 “John bought a book” を日本語に訳すと「ジョンが本を買った」となり，動詞と目的語の語順が入れ替わります．このように言語によって語順が異なることを不思議に思ったことはないでしょうか？ある言語における主語 (S)，目的語 (O)，動詞 (V) の標準的な順番を**基本語順**とよぶことにすると，日本語の基本語順は SOV，英語は SVO です．S，O，V の組合せにはほかに VSO，VOS，OVS，OSV があり，合計で 6 種類ですが，実はそれらすべてが世界で確認されています．例えば，英語の発祥地，ブリテン島の西部で話されているウェールズ語は VSO です．さらに，英語とウェールズ語は遠い共通祖先から枝分かれしたことが知られています．祖先を共有する 2 つの言語が異なる基本語順をとるということは，歴史上少なくとも 1 回は語順が変化したはずです．基本語順はなぜ，どのように変化するのでしょうか？こうした疑問に答えるための足がかりとして，歴史的変化の数理モデルをできるだけ精緻に作る試みを紹介します．

1　連続時間マルコフ連鎖

まずは簡単なモデルとして，**連続時間マルコフ連鎖**を導入します．6 種類の基本語順に対応する 6 個の状態からなる状態空間を考えます．各言語はこの状態空間上を時間とともに確率的に推移するものとみなせます．語順の変化は連続時間軸上の点におけるある状態から別の状態へのジャンプとします．

このような有限状態の連続時間マルコフ連鎖は**推移率行列** Q により特徴づけられます．基本語順の場合，Q のサイズは 6×6 です．その要素 q_{ij} をパラメータ $\lambda_{ij} > 0$ $(i \neq j)$ を用いて

$$q_{ij} = \begin{cases} -\sum_{j' \neq i} \lambda_{ij'} & (i = j) \\ \lambda_{ij} & (i \neq j) \end{cases}$$

とすると，状態 i からはじまり，時間 $t > 0$ が経過したあとに状態 j をとる確率は，行列乗を用いて $\exp(tQ)_{ij}$ と表されることが知られています．あとはパラメータ λ_{ij} の具体的な値をデータから推定すればひとまず完成です．

2　系統樹を用いたパラメータ推定

基本語順のデータとしては，言語学者のドライヤー (Dryer) が収集した 1377 言語のデータ [1] が有名です．ドライヤーは，いずれの語順も支配的ではない (no dominant order) という第 7 の特殊な語順を設定していますが，ひとまずこれを無視して 6 種類の語順を対象とします．問題は，収録言語はいずれも現代語であり，連続時間軸上の推移が観測されていないことです．そもそも基本語順の変化が文献で追える言語は数えるほどしかなく，パラメータ推定には不充分です．

そこで活躍するのが**系統樹**です．系統樹は，冒頭の英語とウェールズ語の例のように，祖先を共有していることが各種の証拠からわかっている諸言語について，それらの歴史的関係を木構造で表したものです．図 1(a) (次ページ) の系統樹では，葉ノード C, D, E は現代語であり，その状態 (基本語順) は観測されています．根を含む内部ノード A, B は現代語の祖先にあたり，その状態は観測されていません．しかし，系統関係を考慮すれば，図 1(b) のような経

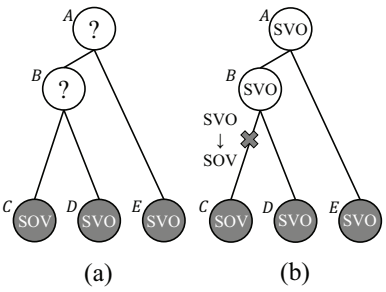

図 1

緯があった可能性が高いと推測できます．過去の状態の推定と，その確率を特徴づけるパラメータ λ_{ij} の推定は鶏と卵の関係ですが，それらの確からしい値を絞り込む手がかりとして系統樹は効果的です．誌面の都合で詳細は省略しますが，このアイデアを応用することで，両者の同時推定が行えます．

具体的なパラメータ推定結果はモーリツ (Maurits) ら [2] が報告しています．それによると，SOV から SVO への変化がその逆よりも起こりやすい，VSO からの変化先としては SVO の方が SOV よりも一般的といった知見が得られたとのことです．

3　基本語順が同じでも内実は多様

筆者はモーリツらの報告に不服でした．基本語順の変化は巨大な変化であり，何の前触れもなく突然起きるはずがありません．この「前触れ」を具体化してモデルに組み込めないでしょうか？

そのヒントは冒頭の日英の例文にあります．日本語では S と O の順番を入れ替えて「本をジョンが買った」とできますが，英語ではできません．日本語の方が語順が柔軟といえます．

語順の柔軟性の点で興味深いのが北米先住民の言語です．言語学者のミスーン (Mithun) [3] によると，スー諸語はかなり厳密に SOV 語順を守る一方，カド諸語は SOV から逸脱して S や O を V の後ろに置く頻度が高く，イロコイ諸語にいたっては no dominant order です．これらの言語で何が違うかというと，カド諸語やイロコイ諸語では，語順が S，O，V のような**統語構造**の代わりに**情報構造**を表す傾向が見られるという点です．情報構造は文に情報を格納する方式のことで，なかでも重要な区分が，聞き手にとって情報が既知 (旧情報) か未知 (新情報) かです．世界の言語の大半は，文を旧情報から始めてそれに新情報を結びつけるという戦略をとりますが，これらの言語は反対に，新情報から旧情報へと並べます．つまり，S や O が旧情報の場合に V に後置されていたのです．そして，語順を情報構造の表示に使う代わりに，語順によらずに統語構造を明らかにする手段が発達しています．具体的には，動詞につく人称接辞を強化したり，名詞抱合 (日本語の「腰掛ける」のような複合語形成をずっと生産的にした，動詞が名詞を取り込む現象) を多用することで統語的役割を特定すべき名詞句自体を減らしたりといった方法です．

やや専門的になりましたが，要するに，基本語順の変化は言語のその他の特徴と連動していそうです．歴史的変化のモデルを精緻化するには，基本語順以外の特徴も同時に考慮する必要があります．

4　特徴ベクトルのモデル化の難しさ

ドライヤーの基本語順データは，実は World Atlas of Language Structures (WALS) というデータベースの一部です．WALS は名前の通り世界地図による可視化機能を備えており，オンライン (`https://wals.info/`) で一般公開されています．ぜひ一度ご覧になってください．

WALS は言語の構造的特徴 192 個をカテゴリカルにコード化しています．基本語順特徴はその 1 つです．今回は，さまざまな理由から取り扱いが難しい特徴を前処理で取り除き，$N = 152$ 個の特徴を採用します．N 個の特徴のうち，どれが基本語順と特に関係が深いかは事前に決めがたいので，すべてモデルに組み込むことにします．これらの特徴を適当な順番で並べると，1 つの言語が N 個の要素からなる**特徴ベクトル**で表現されることになります．

特徴 $i \in \{1, \cdots, N\}$ がとる値の異なり数を F_i とします．例えば基本語順の場合は $F_i = 6 + 1 = 7$ で

す (ここからは特殊語順 no dominant order を復活させます)．これらの値に便宜的に $1, \cdots, F_i$ のいずれかの数値を割り振ります．

では，特徴ベクトルの歴史的変化はどのようにモデル化すればよいでしょうか？安直な方法は，N 個の特徴それぞれに対して推移率行列を用意し，それらのパラメータを系統樹を用いて同時推定するというものです．しかし，この方法は特徴間の独立性を仮定しています．基本語順がその他の特徴と相互作用するという性質を考慮したいのに，それができていません．

第 2 の策は，特徴の組合せに対して推移率行列を用意するというものです．特徴 i_1, i_2 の組合せに対応する推移率行列のサイズは $(F_{i_1} F_{i_2}) \times (F_{i_1} F_{i_2})$ です．こうすれば，特徴 i_1 のある値と特徴 i_2 のある値の組合せが安定である，あるいは反対に不安定で別の組合せに変化しやすいといった性質を捉えられます．しかし，この方法では同時に考慮する特徴の数を増やすのが困難です．N 個の特徴すべてを考慮すると，推移率行列のサイズは $(\prod_{i=1}^{N} F_i) \times (\prod_{i=1}^{N} F_i)$ となり，現実的にはパラメータ推定は不可能です．

5　表現学習

ここで筆者の手法の登場です [5, 4]．アイデアは簡単です．特徴ベクトルがそのままでは扱いにくいのなら，扱いやすい別の表現に変換してしまえばよいのです．観測される特徴ベクトルに対応する，直接観測されることのない潜在表現を自分で作るということです．こうした発想は昔からあり，主成分分析もその一例といえますが，2010 年代に深層学習が流行りだすと，**表現学習**というキャッチーな (？) 名前がつけられて再び脚光を浴びています．

まずは潜在表現に求める要件を挙げていきます．図 2 のように，特徴ベクトルから潜在表現への変換 (エンコード) と潜在表現から特徴ベクトルへの変換 (デコード) が行える必要があります．特に，ある特徴ベクトルをエンコードし，続けてデコードした場合，もとの特徴ベクトルが復元されなければなりません．次に，潜在表現も K 個の要素からなるベクトルとします．特徴ベクトルの多様性を表現するには，潜在表現にも複数の要素が必要となります．潜在ベクトルは，各要素を独立に操作しても安全なような，特徴同士のもつれを解きほぐした (disentangled) 表現である必要があります．最後に，潜在ベクトルの要素は連続値でも離散値でも構いませんが，今回は離散の 2 値変数を採用します．そうすれば，これまで見てきた手法—連続時間マルコフ連鎖と系統樹を用いたパラメータ推定—が使い回せます．同じ手法を特徴ベクトルではなく潜在ベクトルに適用するだけです．

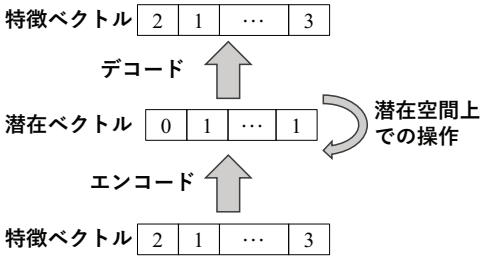

図 2

6　潜在表現に基づく歴史的変化の分析

上記の要件を満たすモデルの作り方はいろいろ考えられます．深層学習時代の流行はエンコードとデコードをそれぞれ別のニューラルネットで実現するものです．しかし，普通のニューラルネットは WALS のデータに適用するには自由度が高すぎます．というのも，2600 を超える収録言語全体で見ると，特徴ベクトルの要素の 80%以上が欠損値です．欠損値が多いのは無理もないことで，ある言語のある特徴の値を決めるには，当該言語の調査報告を読み解き，当該特徴の設計理論に基づいて再解釈する必要があります．その大変さが想像できるかと思います．

筆者は試行錯誤の末，デコードは確率的生成モデルで，エンコードはそれに対応する確率的推論を実装することで実現しました [5]．エンコードとデコードを単一のモデル (パラメータ集合) で実現すること

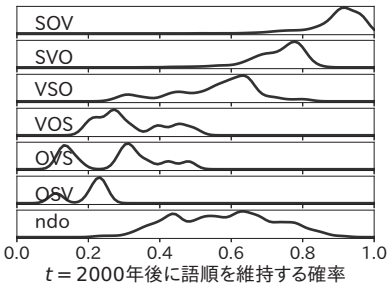

図 3

で，モデルの自由度を抑えています．また，欠損値補完の精度向上のための工夫として，系統的・地理的に近い言語同士は同じ値をとりやすいという事前の期待をモデルに組み込みました．こうして世界中の言語を潜在ベクトルに変換したのち，潜在ベクトルの K 個の要素それぞれに対して推移率行列を用意し，それらのパラメータを系統樹を用いて同時推定しました [4]．得られたパラメータを用いると，例えば時間 $t > 0$ が経過したあとの潜在ベクトルの状態をシミュレートできます．潜在ベクトルをデコードすれば対応する特徴ベクトルが得られ，基本語順がどう変化したかが確認できます．

図 3 はシミュレーション結果の 1 つです ($K = 100$)．$t = 2000$ 年後も同じ基本語順を維持する確率を基本語順ごとに集約しました (ndo は no dominant order)．同じ基本語順でも，その他の特徴の値によって安定性がばらつきました．日本語は約 94%で, SOV 言語の最頻値よりも少し高い確率です．

個別言語のシミュレーション例として，日本語と，同じく SOV 語順のスー諸語のラコタ語を図 4 に示します．日本語と比較するとラコタ語の SOV 語順は不安定で，変化先として no dominant order が目立ちます．

日本語が現在の SOV から SVO に変わる場合に，どのような特徴が連動するかも調べてみました．日本語は現在は「が」,「を」のような文法要素を内容語に後続させて統語構造を示しますが，SVO に変化する場合は，こうした文法要素を使わない，まるで英語のような言語になりやすいという結果を得ました．

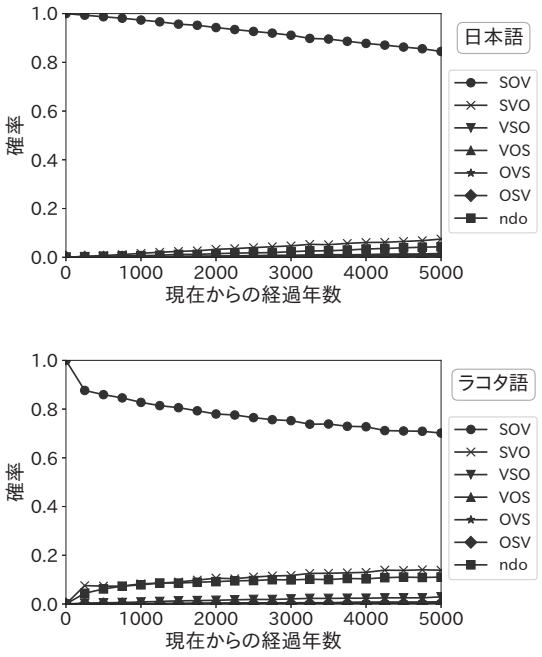

図 4

7 自然な言語の分布

複数の特徴が連動することは，言語学者のグリーンバーグ (Greenberg) が 1960 年代に本格的に議論して以来，言語学ではずっと注目されてきた現象です．しかし，言語学者は人手で論証するため，一度に分析できる特徴の組合せは高々 2 つでした．筆者は，数理的道具を導入すれば N 個の特徴の組合せを同時に扱えることを示しました．

考えてみると，筆者のモデルは特徴ベクトルのなかで基本語順特徴を特別扱いしているわけではありません．別の特徴に着目しても同じ手順で分析できますが，さらに議論を一般化できないでしょうか？

各言語は，特徴ベクトルに対応する N 次元の空間上の点で表現されます．重要なのは，この空間上で言語が一様には分布しておらず，ある点は言語として自然であり，別の点は不自然だといえそうなことです．しかも，言語が時間とともに変化することは，この空間に対する局所的制約とみなせます．すなわち，言語の漸進的な変化は空間上に経路をなし，こ

の経路上のすべての点が自然でなければなりません．これらを考慮したとき，この空間がどのような大域的な構造を持つといえるかは興味深い問題です．

参考文献

[1] Matthew S. Dryer. Order of subject, object and verb. In Matthew S. Dryer and Martin Haspelmath, editors, *The World Atlas of Language Structures Online*. Max Planck Institute for Evolutionary Anthropology, 2013.

[2] Luke Maurits and Thomas L. Griffiths. Tracing the roots of syntax with Bayesian phylogenetics. *Proceedings of the National Academy of Sciences*, Vol. 111, No. 37, pp. 13576–13581, 2014.

[3] Marianne Mithun. Morphological and prosodic forces shaping word order. In Pamela A. Downing and Michael Noonan, editors, *Word Order in Discourse*, pp. 387–423. John Benjamins Publishing, 1995.

[4] Yugo Murawaki. Analyzing correlated evolution of multiple features using latent representations. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 4371–4382, 2018.

[5] Yugo Murawaki. Bayesian learning of latent representations of language structures. *Computational Linguistics*, Vol. 45, No. 2, pp. 199–228, 2019.