潜在表現に基づく言語構造の史的変化の分析

京都大学
村脇 有吾

機構間連携・文理融合プロジェクト「言語における系統・変異・多様性とその数理」シンポジウム
2018年2月2日 TKP東京駅大手町カンファレンスセンター

自己紹介: 村脇 有吾

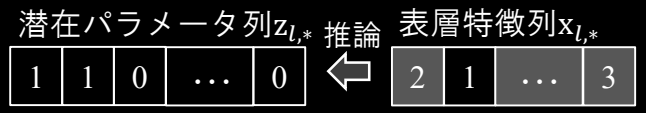
- 京都大学 大学院情報学研究科
  知能情報学専攻 助教
  工学部電気電子工学科 兼担
- 専門: 計算言語学と自然言語処理
  - 表の仕事は普通のテキスト処理
    - 単語分割、ゼロ照応解析、常識的知識の獲得ほか
  - 今日お話も裏の仕事
    - 言語の研究ですが、テキストは直接扱いません

言語構造の
潜在表現

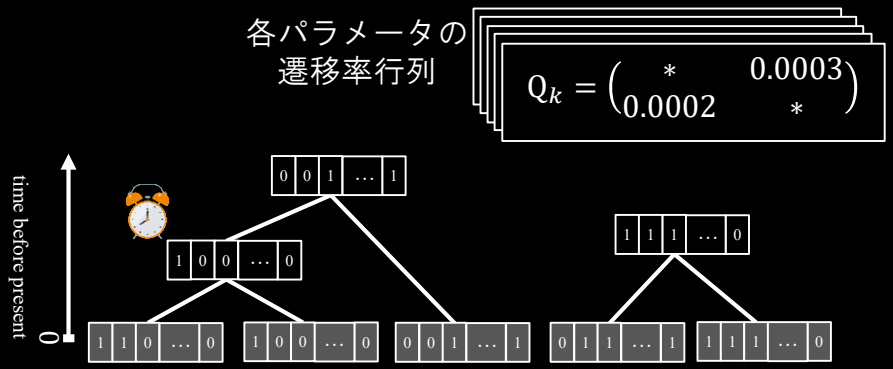
史的変化の
統計的推論

シミュレーション
による分析

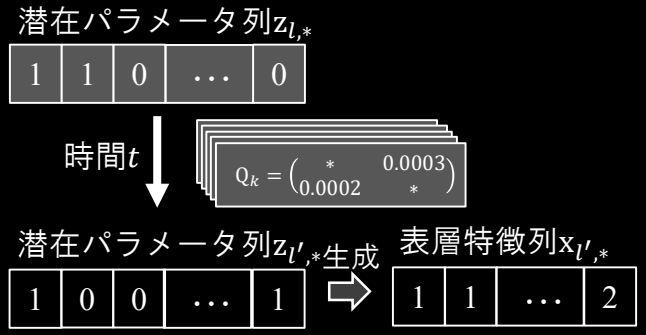
Step 1. 各言語を潜在表現に変換



Step 2. 系統樹群から各潜在パラメータの遷移率行列を推定
(内部ノードの状態、年代も同時推定)



Step 3. 遷移率行列を用いて言語の時間変化をシミュレート



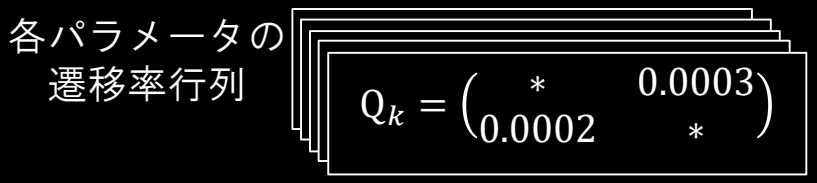
言語構造の
潜在表現

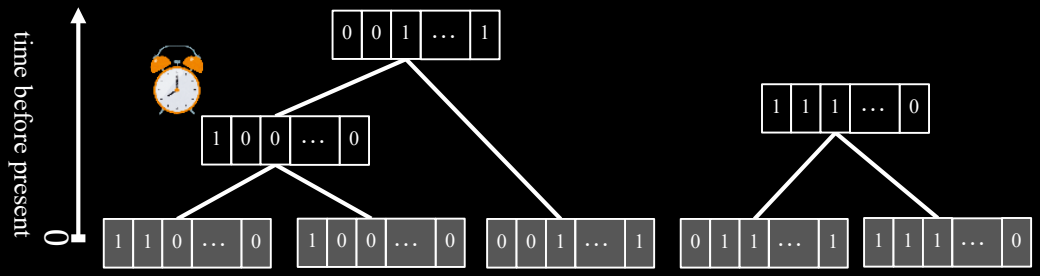
史的変化の
統計的推論

シミュレーション
による分析

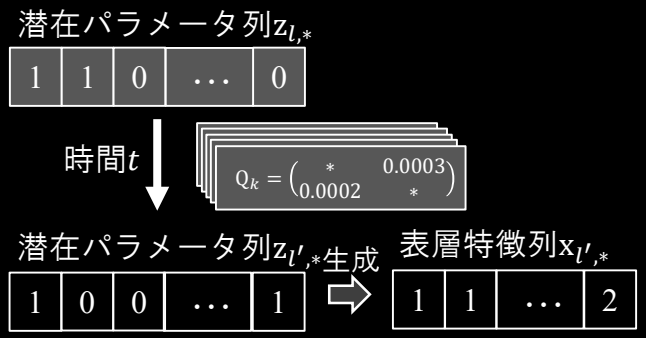
Step 1. 各言語を潜在表現に変換

Step 2. 系統樹群から各潜在パラメータの遷移率行列を推定
(内部ノードの状態、年代も同時推定)





Step 3. 遷移率行列を用いて言語の時間変化をシミュレート



基本語順: Subject, Object, Verb

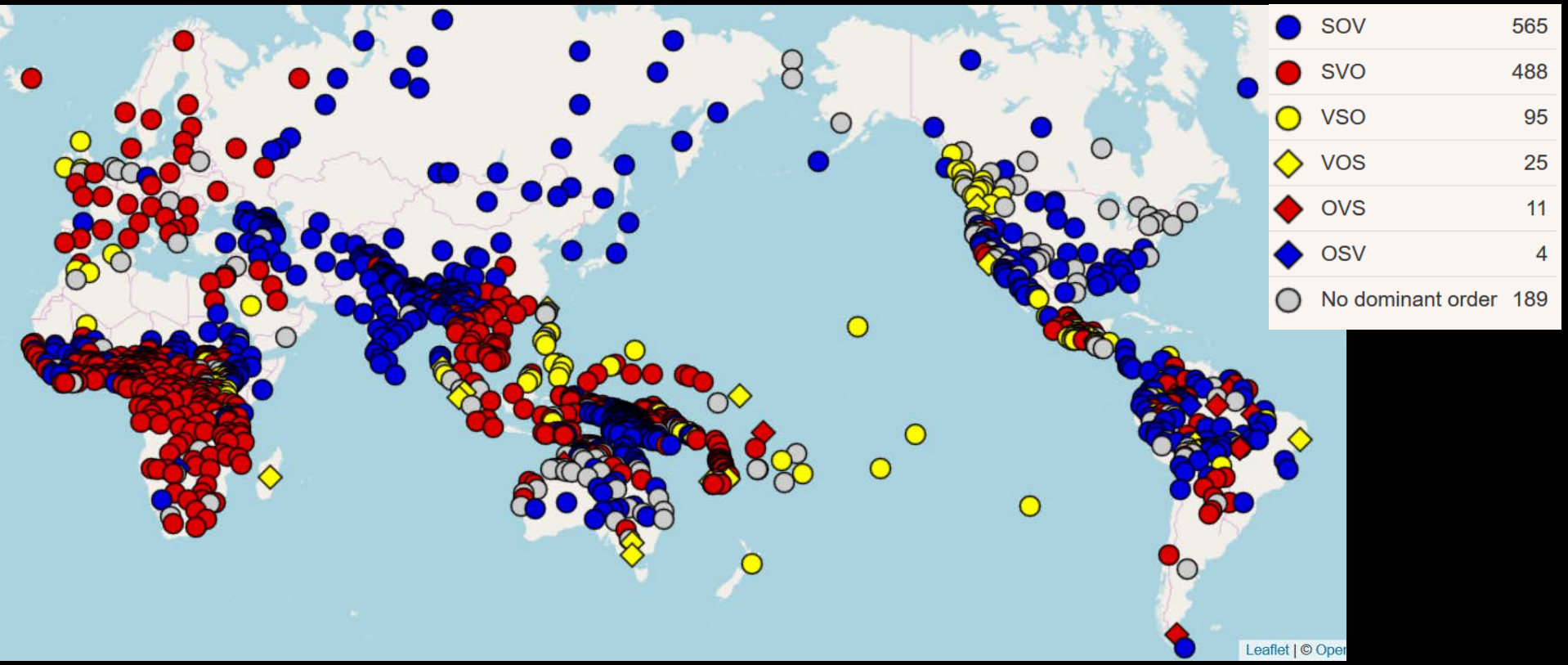
[Dryer, 2005]

SOV	日本語	*John ga*	*tegami o*	*yon-da*
		S	O	V
SVO	英語	*The dog*	*chased*	*the cat*
		S	V	O
VSO	アイルランド語	*Léann*	*na sagairt*	*na leabhair*
		V	S	O

The priests are reading the books.

基本語順: Subject, Object, Verb

[Dryer, 2005]



出典: http://wals.info/feature/81A

基本語順に関する疑問

[Maurits+, PNAS 2014]

- なぜ世界における分布がこのようになっているのか?
- どのように変化してきたのか?
- 原型言語 (それが存在したとして) はどの語順だったのか?

基本語順に関する従来説の例

[Maurits+, PNAS 2014]

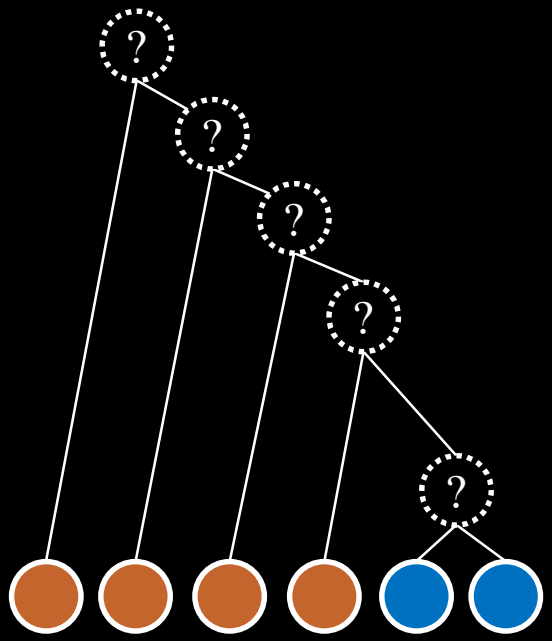
- SOVが最も高頻度なのは…
  1. 機能的に優れているから
  2. かつてより高頻度だったなごりにすぎない
     - SVOの方が機能的に優れている
     - SOV⇒SVOがSVO⇒SOVよりも多い (と推測)
       - 原型言語もSOVだったかも
- パントマイム実験でSOVが選好される
  - 原型言語はSOVだった傍証?
  - 現代人 (しかもWEIRD) を使った実験で原型言語のことが本当にわかるのか?

史的変化の推論

- 個別言語の分析
  - 歴史文献の分析
    - 中英語期にSVOへの固定化が進む
  - 内的再構
    - オーストロアジア語族のムンダ語派は(S)OV語順だが、VO語順の痕跡が残る [Donegan+, 2004]
- 言語間比較
  - 共時類型論の動態化 [Greenberg, 1969]
  - 系統学的比較法

系統学的比較法

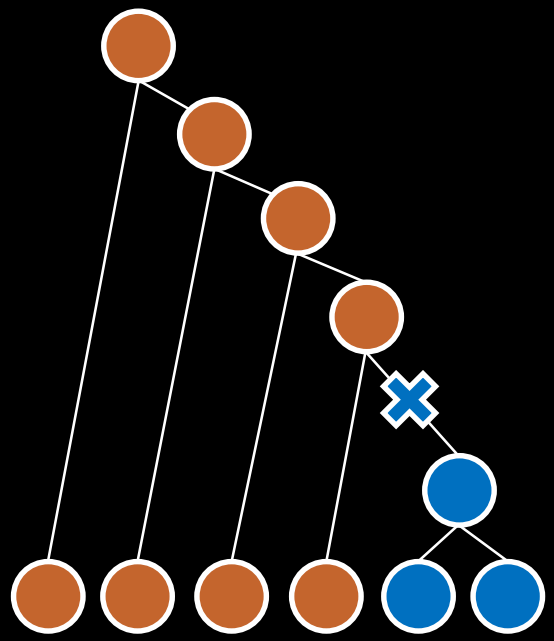
- 系統樹が (主に語彙的手がかりを用いた推論により) 既知とする
- 系統樹上のどのエッジで注目する値が出現したか推測できる (場合がある)



簡単のために2値特徴で例示するが、
多値特徴への拡張は容易

系統学的比較法

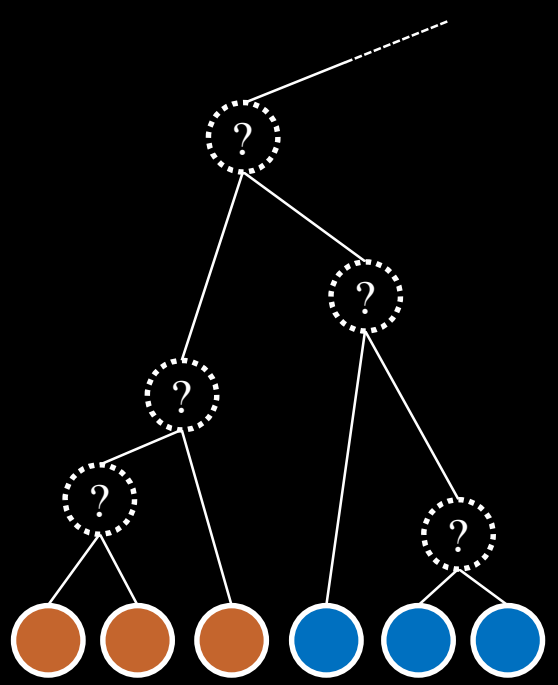
- 系統樹が (主に語彙的手がかりを用いた推論により) 既知とする
- 系統樹上のどのエッジで注目する値が出現したか推測できる (場合がある)



簡単のために2値特徴で例示するが、
多値特徴への拡張は容易

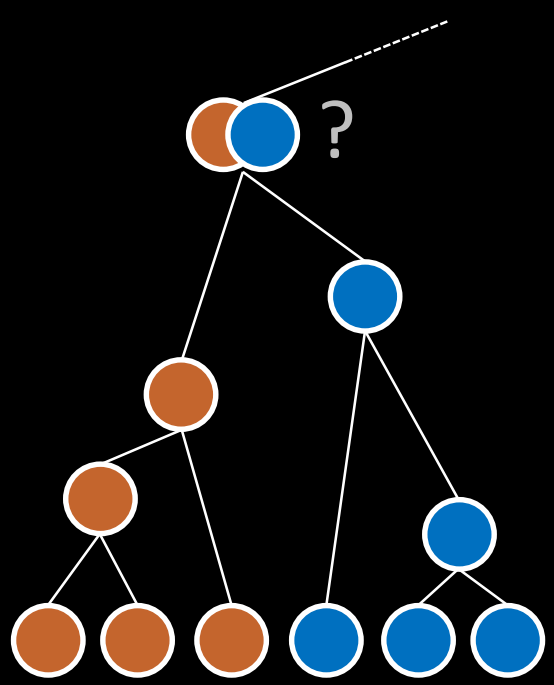
系統学的比較法

- もちろん現在得られる手がかりだけでは確信を持って決められない場合も多い
- 人間はお手上げ



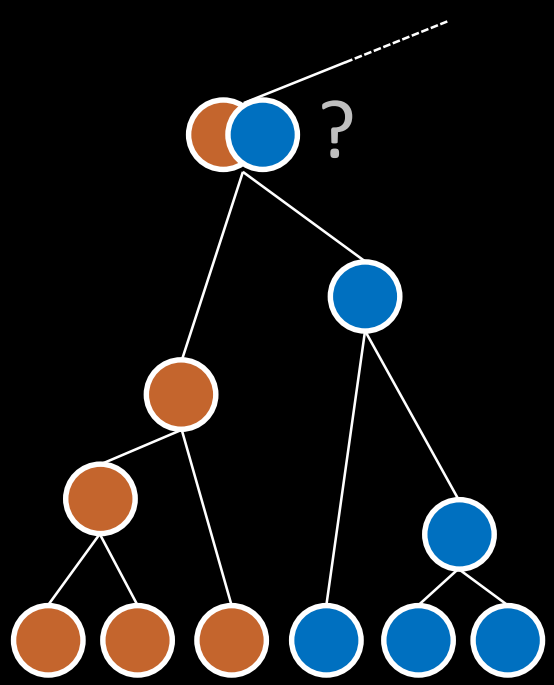
系統学的比較法

- もちろん現在得られる手がかりだけでは確信を持って決められない場合も多い
- 人間はお手上げ



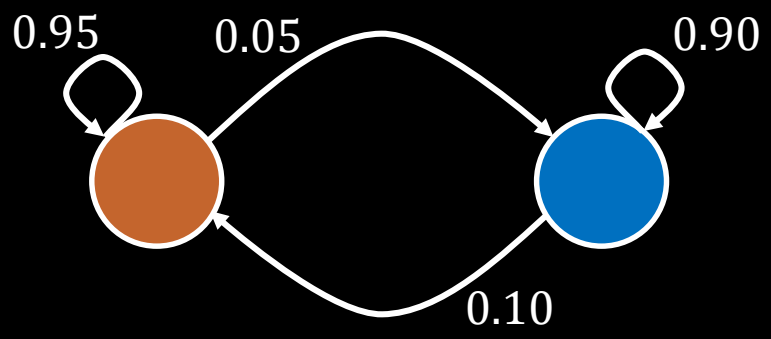
系統学的比較法

- もちろん現在得られる手がかりだけでは確信を持って決められない場合も多い
- 人間はお手上げ
- 確率的手法の出番



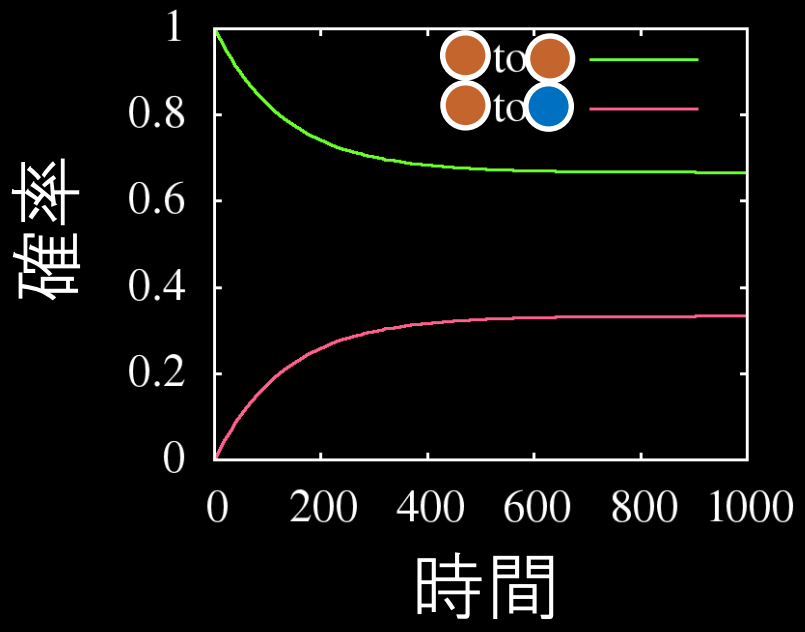
状態遷移モデル (マルコフ連鎖)

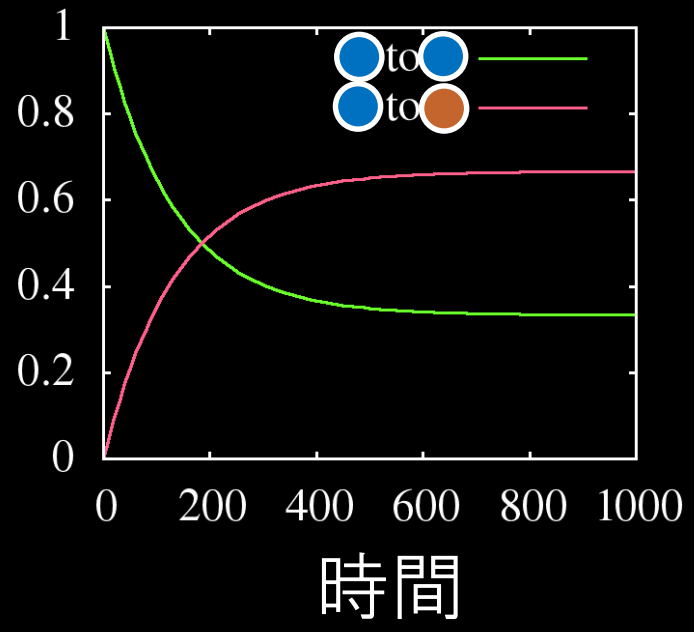
- まずは簡単のために離散時間を考える
- 時刻tである値 (状態) のとき、時刻$t+1$で取る値 (状態) の確率のモデル



連続時間マルコフ連鎖 (CTMC)

- 現在の値がaのとき時間t後に値がbになる確率: $\exp(tQ)_{a,b}$
  ただし遷移率行列$Q = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}$

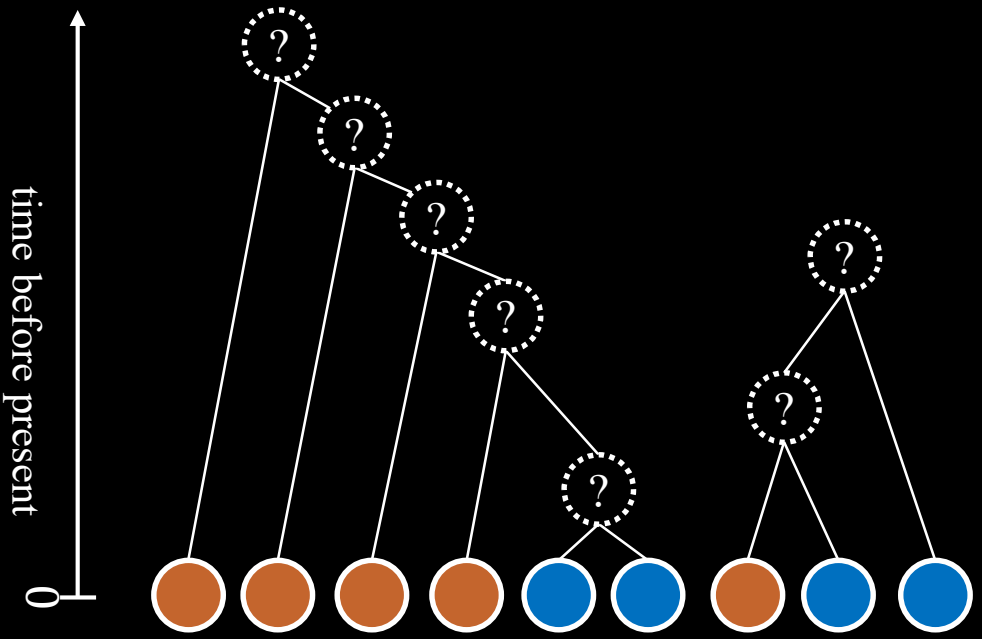




(年代つき) 系統樹群を用いた遷移率行列の推定

[Greenhill+, 2010]
[Maurits+, PNAS 2014]

- 観測データ
  - (年代つき) 系統樹 (群)
  - 葉ノードの状態
- 潜在データ
  - 遷移率行列
  - 内部ノードの状態



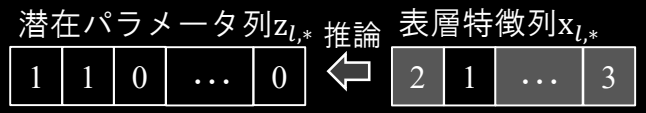
$$Q = \begin{pmatrix} -\alpha & \alpha \\ \beta & -\beta \end{pmatrix}$$

言語構造の
潜在表現

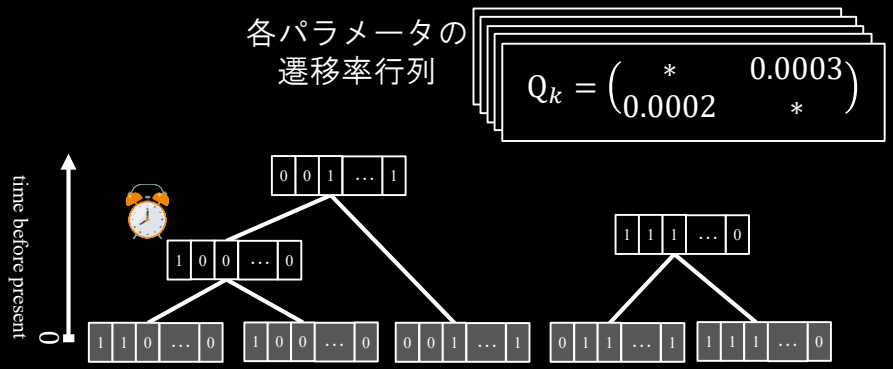
史的変化の
統計的推論

シミュレーション
による分析

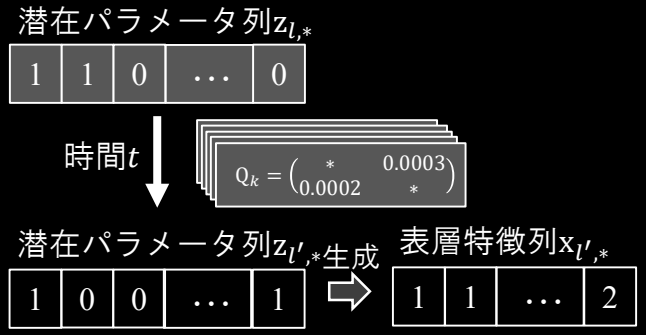
Step 1. 各言語を潜在表現に変換



Step 2. 系統樹群から各潜在パラメータの遷移率行列を推定
(内部ノードの状態、年代も同時推定)



Step 3. 遷移率行列を用いて言語の時間変化をシミュレート

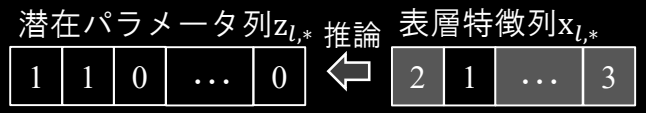


言語構造の
潜在表現

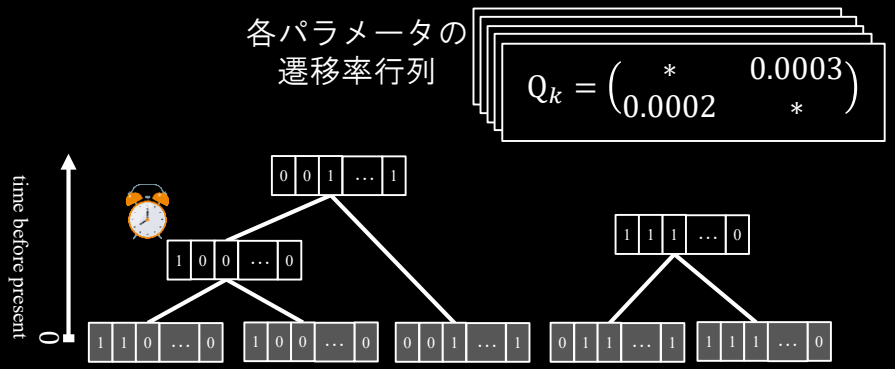
史的変化の
統計的推論

シミュレーション
による分析

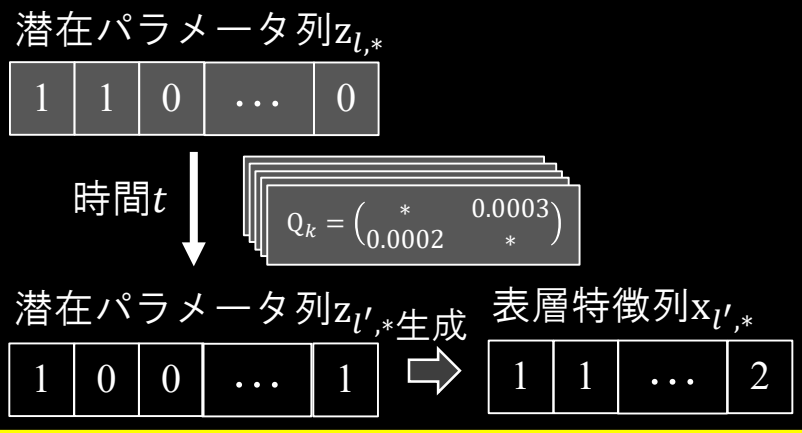
Step 1. 各言語を潜在表現に変換



Step 2. 系統樹群から各潜在パラメータの遷移率行列を推定
(内部ノードの状態、年代も同時推定)

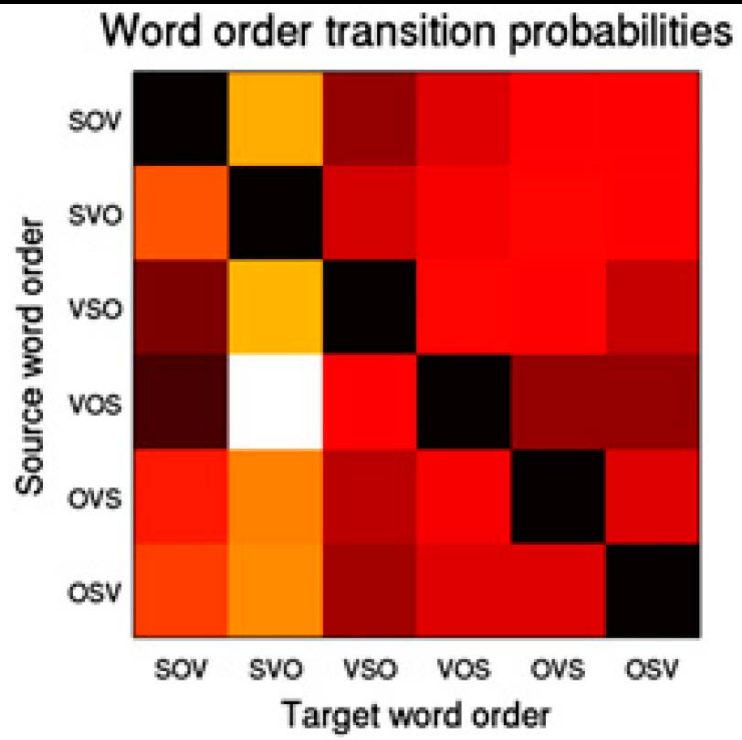


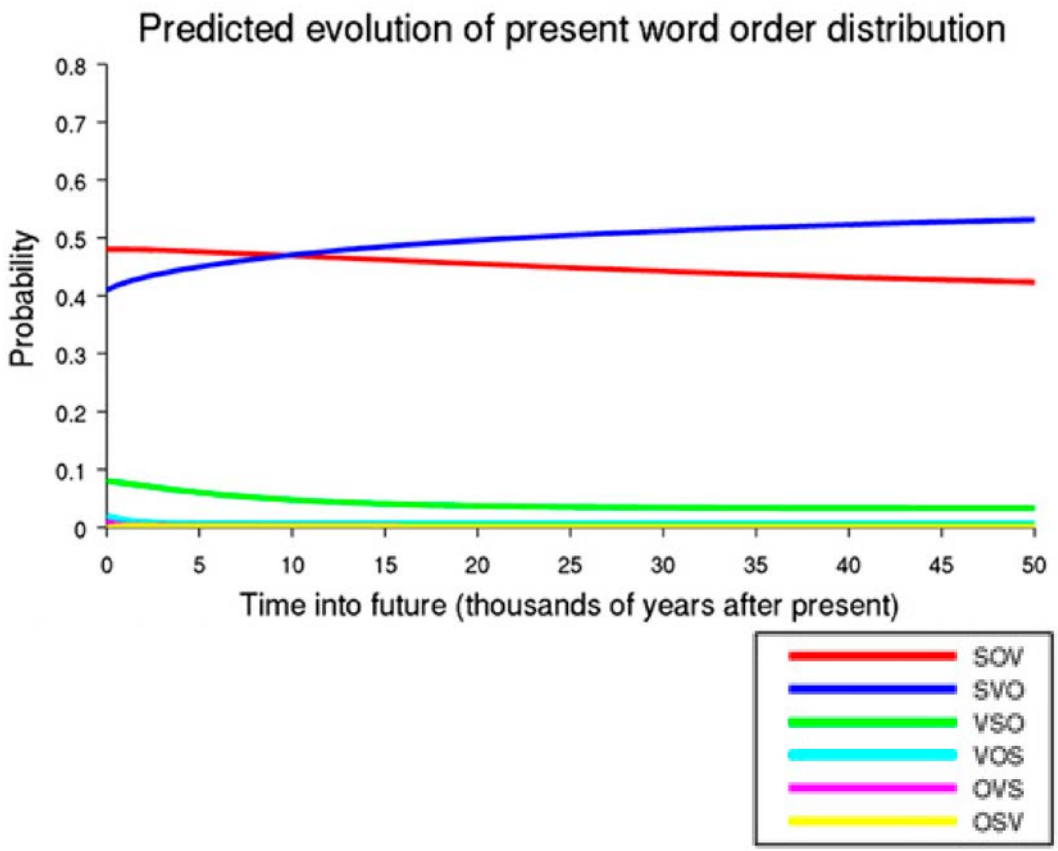
Step 3. 遷移率行列を用いて言語の時間変化をシミュレ



基本語順の遷移確率と将来変化の予測

[Maurits+, PNAS 2014]



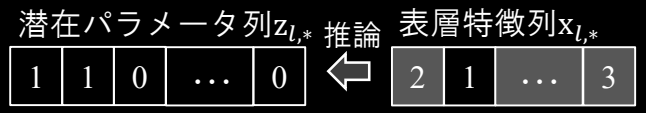


言語構造の
潜在表現

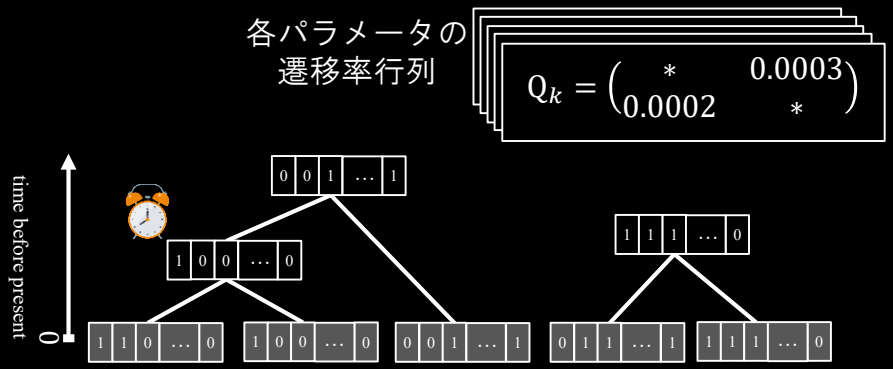
史的変化の
統計的推論

シミュレーション
による分析

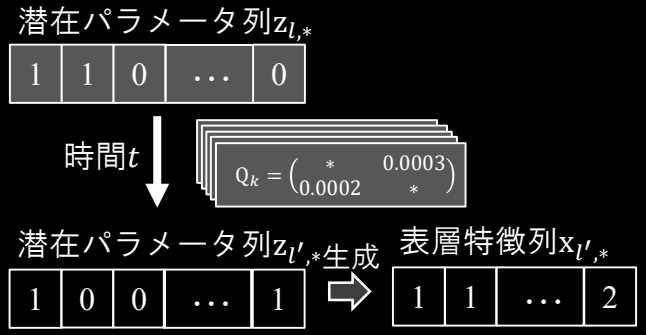
Step 1. 各言語を潜在表現に変換



Step 2. 系統樹群から各潜在パラメータの遷移率行列を推定
(内部ノードの状態、年代も同時推定)

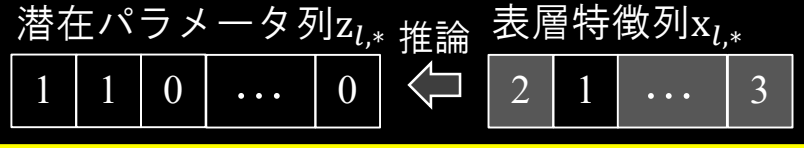


Step 3. 遷移率行列を用いて言語の時間変化をシミュレート



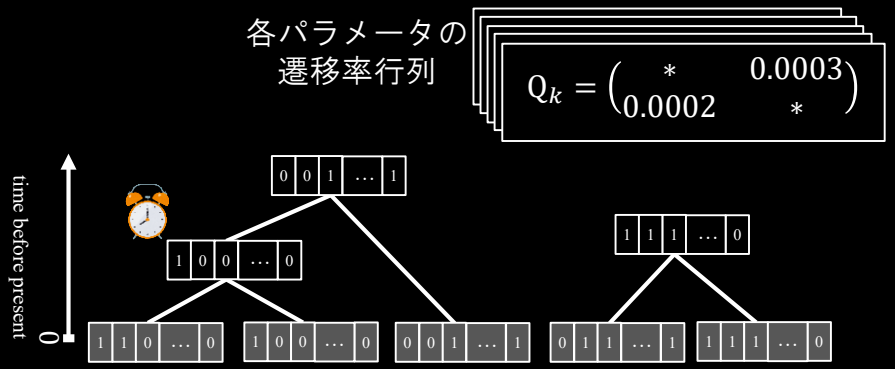
言語構造の
潜在表現

Step 1. 各言語を潜在表現に変換



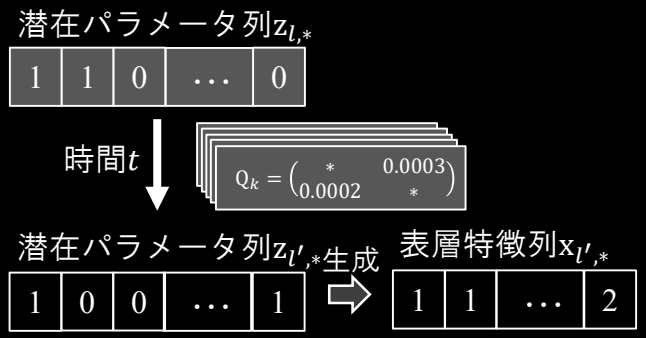
史的変化の
統計的推論

Step 2. 系統樹群から各潜在パラメータの遷移率行列を推定
(内部ノードの状態、年代も同時推定)



シミュレーション
による分析

Step 3. 遷移率行列を用いて言語の時間変化をシミュレート



特徴間の依存関係を利用した分析の精緻化

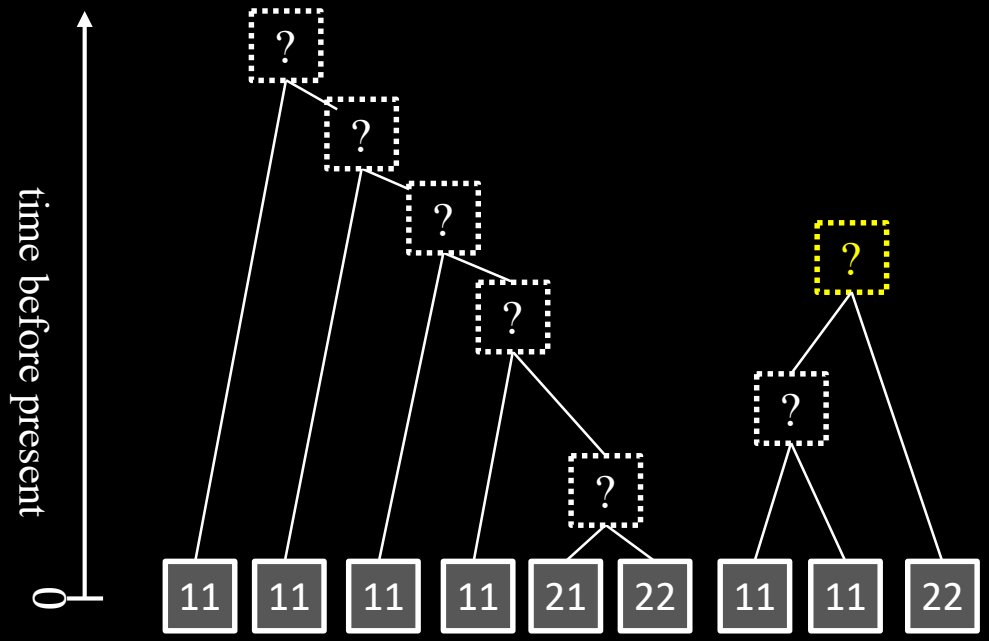
- 含意的普遍性 [Greenberg, 1963]
  - 目的語・動詞と名詞・関係節の語順の関係
    - If VO, then NRel
    - If RelN, then OV

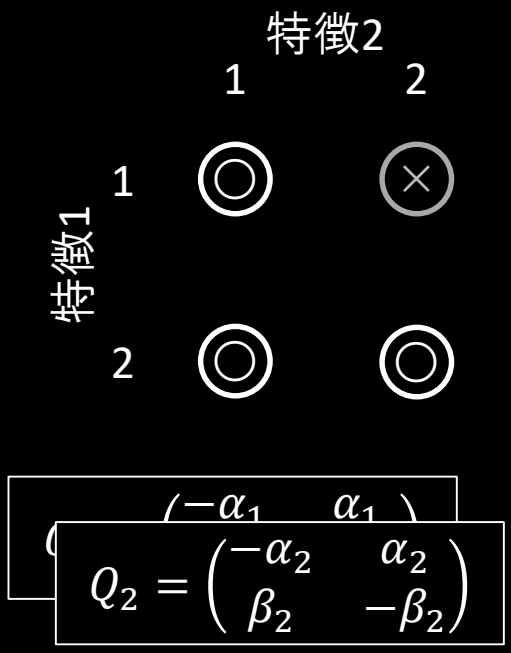
	NRel	RelN
VO	○	×
OV	○	○

- 基本語順の変化は一大変化であり、他の特徴の変化と連動しているはず
  - 英語のSVO語順への変化は、孤立語化と連動しているように見える

単純に特徴ごとに遷移率行列を用意すると独立性を仮定したことに

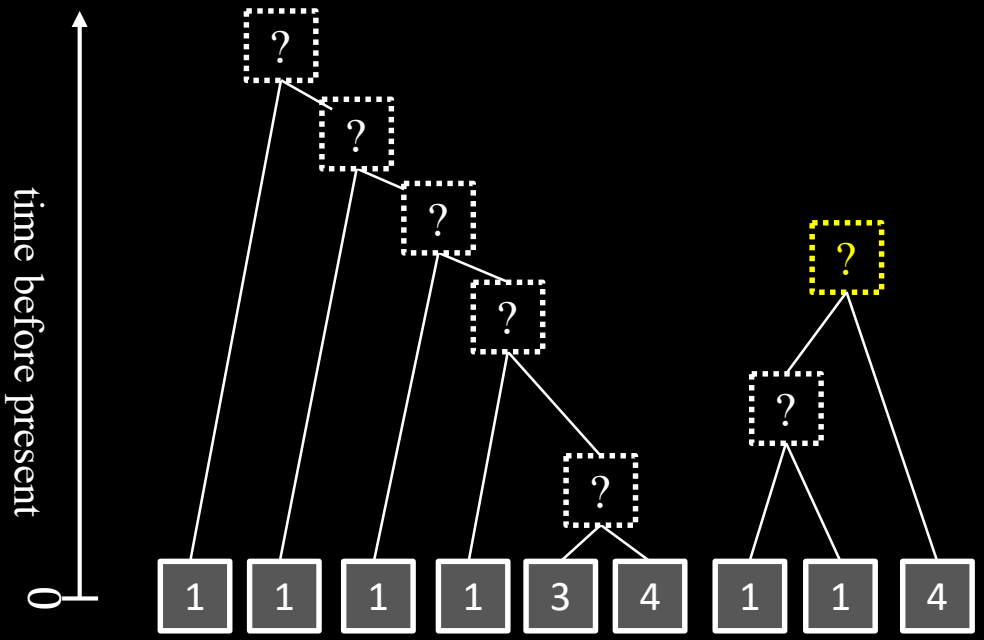
- 特徴対の値の組み合わせ12が不自然だという知識を推論に反映させられない

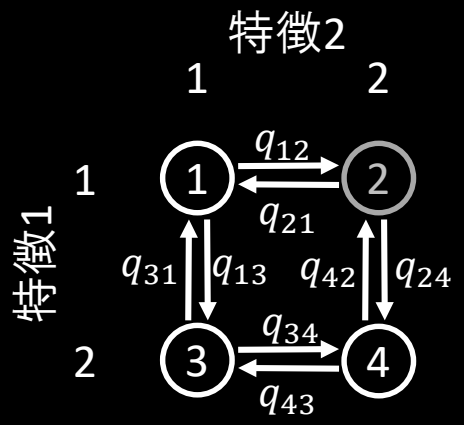




値の組み合わせの展開による依存関係モデル化 [Dunn+, Nature 2011]

- 3個以上の特徴や、多値特徴 (基本語順は6-7値) は組合せ爆発を起こして推論困難



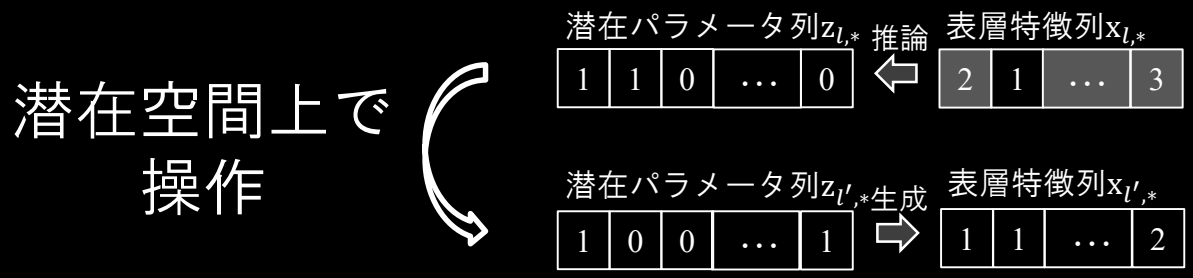


$$Q = \begin{array}{c} \\ 11 \\ 12 \\ 21 \\ 22 \end{array} \begin{array}{c} \begin{array}{cccc} 11 & 12 & 21 & 22 \end{array} \\ \begin{pmatrix} * & q_{12} & q_{13} & - \\ q_{21} & * & - & q_{24} \\ q_{31} & - & * & q_{34} \\ - & q_{42} & q_{43} & * \end{pmatrix} \end{array}$$

表層特徴列を互いに独立な潜在パラメータ列に写像

[Murawaki, IJCNLP2017]

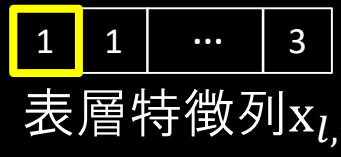
- 104個の特徴を100個の2値パラメータに再編
  - パラメータは仮定により互いに独立
- 確率的生成モデル
  - パラメータ列から特徴列への変換は生成
  - 特徴列からパラメータ列への変換は事後推論



特徴間の依存関係を捉えるのは重み行列W

[Murawaki, IJCNLP2017]

Softmax分布から
確率的に生成 ⇓



特徴間の依存関係を捉えるのは重み行列W

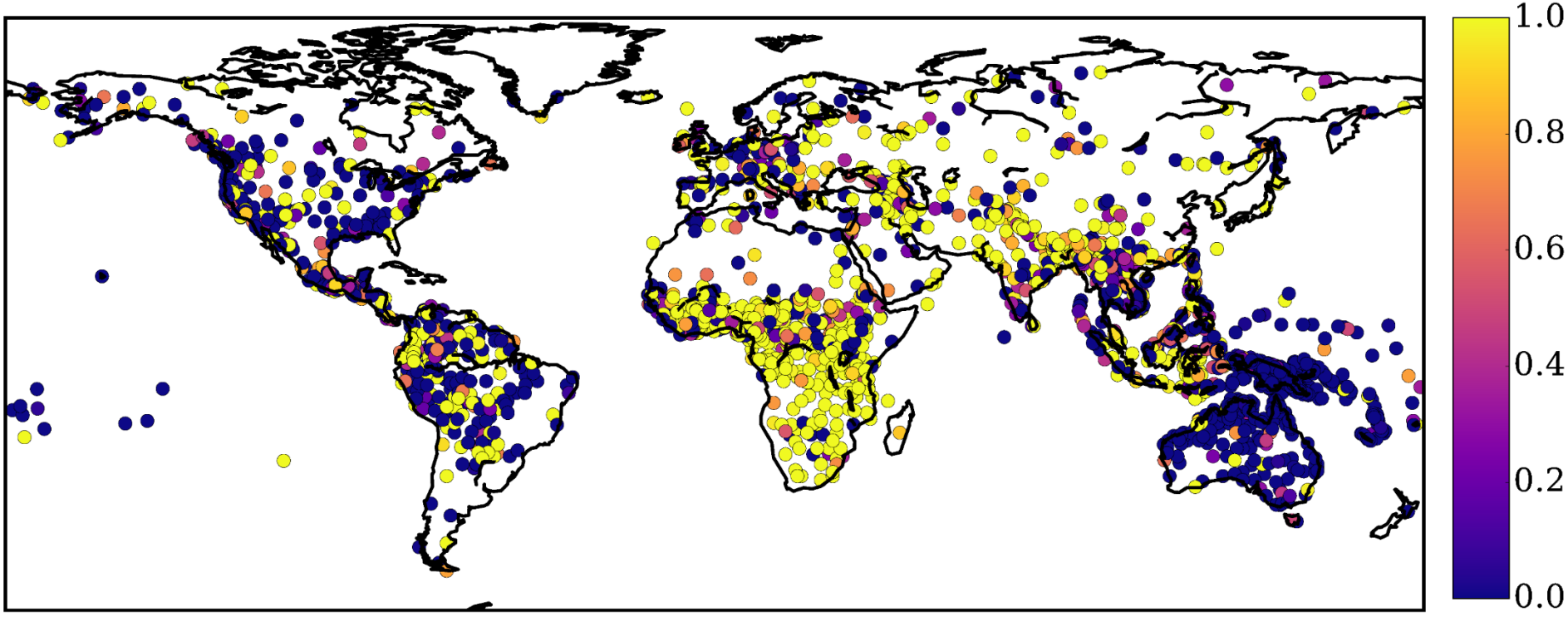
[Murawaki, IJCNLP2017]

Softmax分布から
確率的に生成 ⇓

表層特徴列の一部 (26.9%) が与えられたとき、
残りの欠損値、潜在パラメータ列、重み行列W
を事後推論

1 1 ⋯ 3
表層特徴列$x_{l,*}$

獲得されたパラメータの地理的分布の例



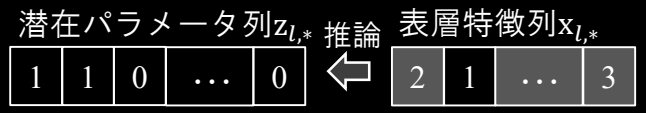
表層的特徴に見られた地理的信号を潜在的パラメータへ写像しても維持している (ように見える)

言語構造の
潜在表現

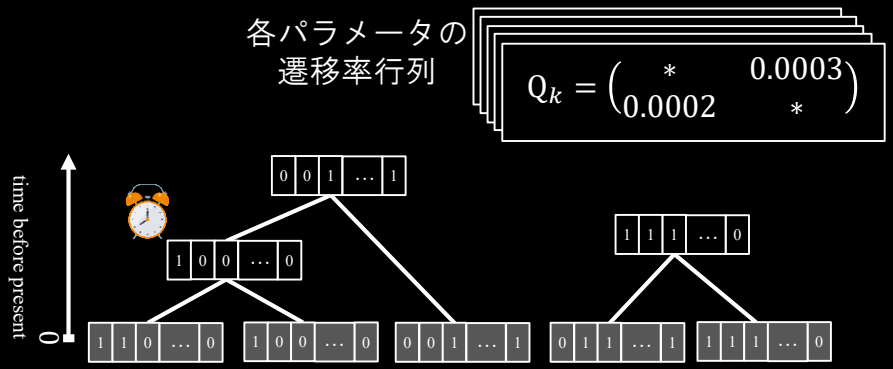
史的変化の
統計的推論

シミュレーション
による分析

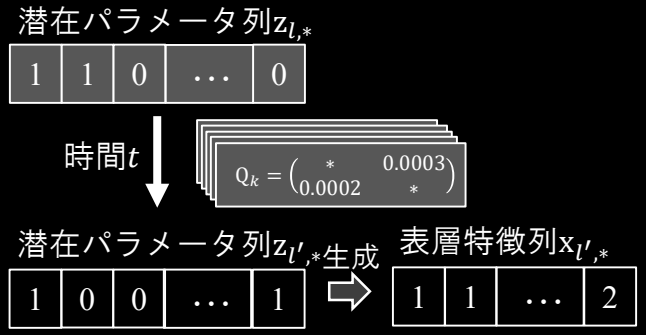
Step 1. 各言語を潜在表現に変換



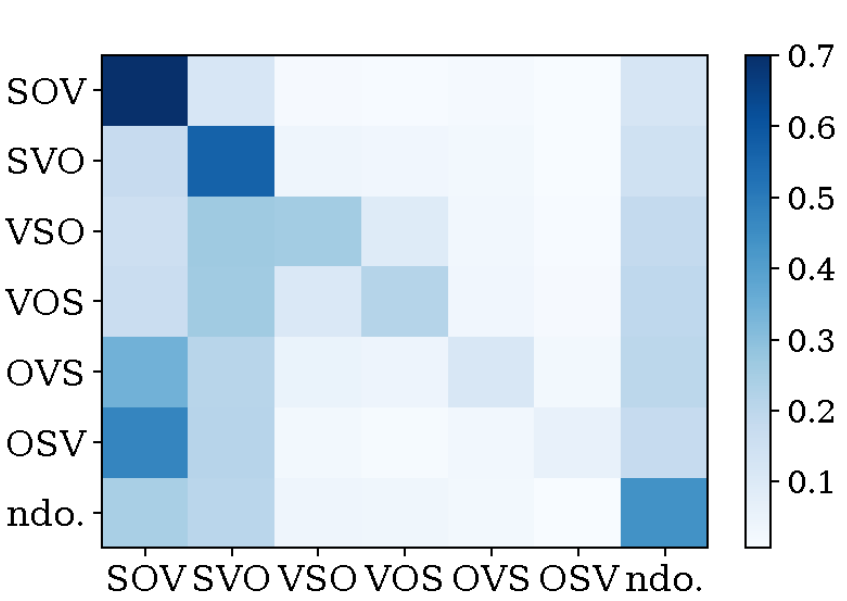
Step 2. 系統樹群から各潜在パラメータの遷移率行列を推定
(内部ノードの状態、年代も同時推定)



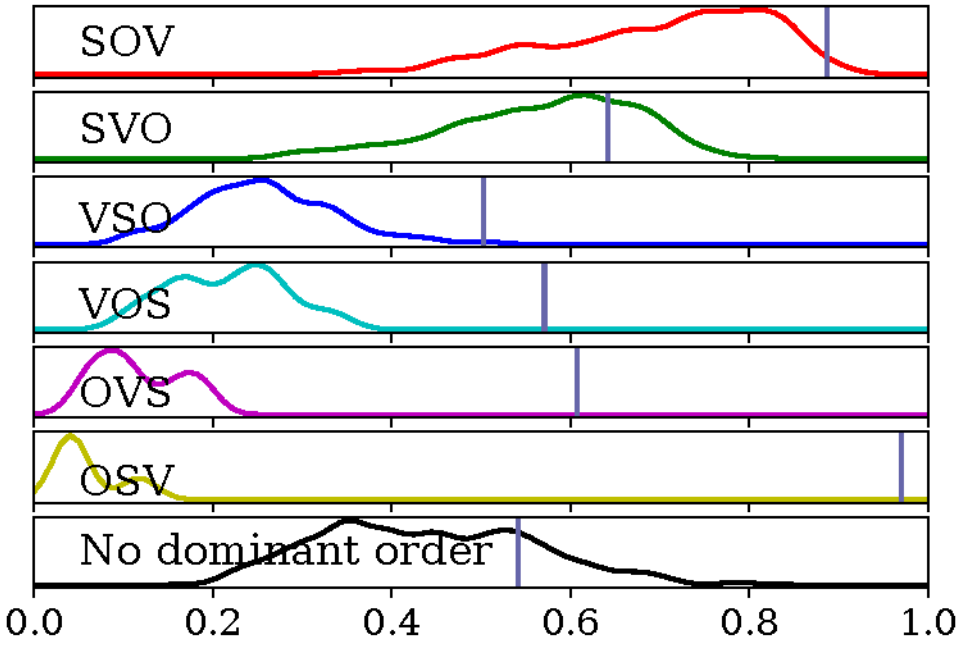
Step 3. 遷移率行列を用いて言語の時間変化をシミュレート



2千年後の基本語順の予測

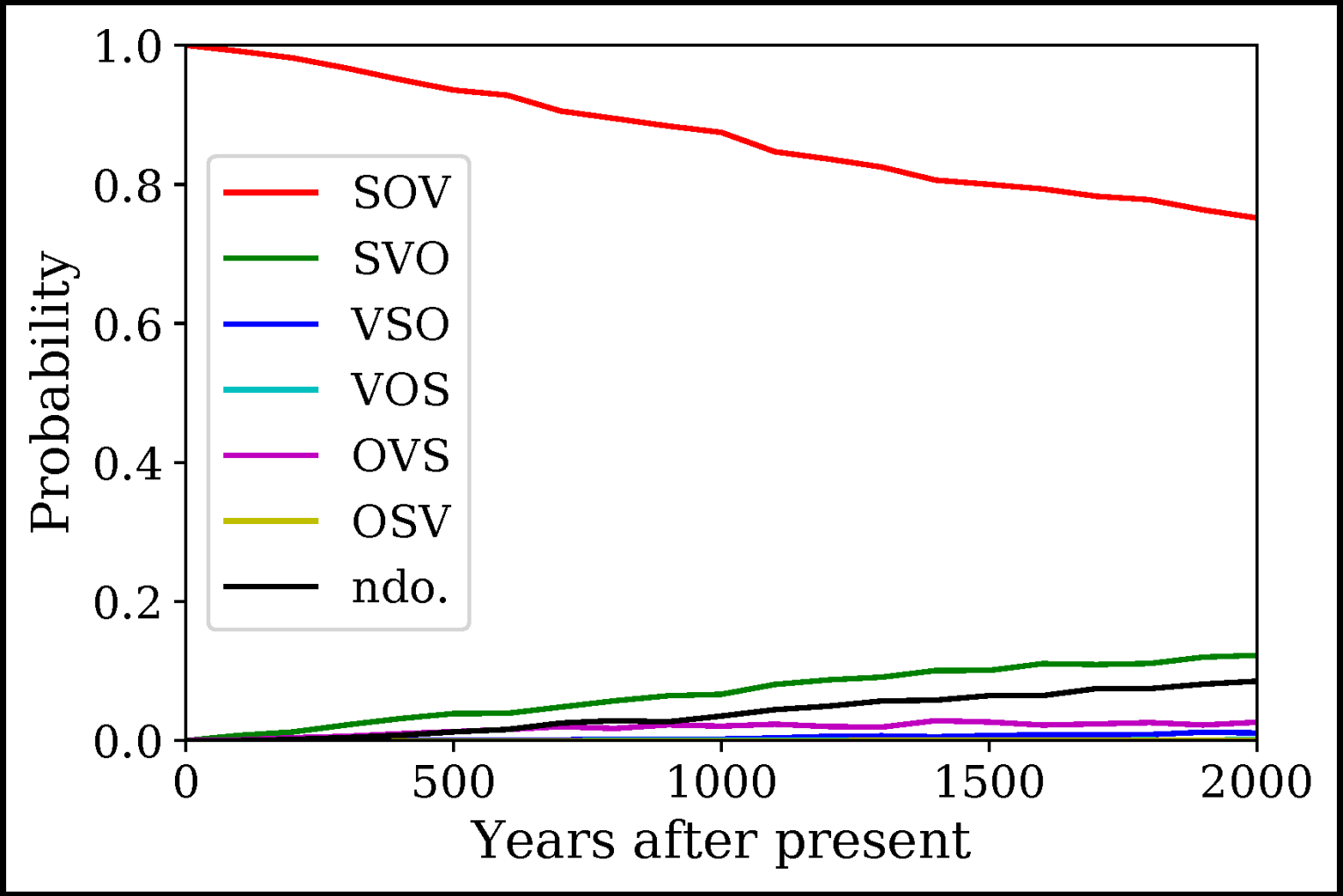


平均遷移確率



語順維持確率の
言語ごとのばらつき

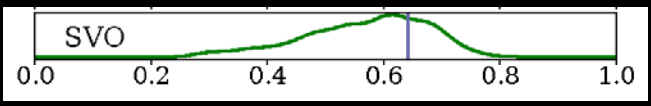
日本語の将来予測

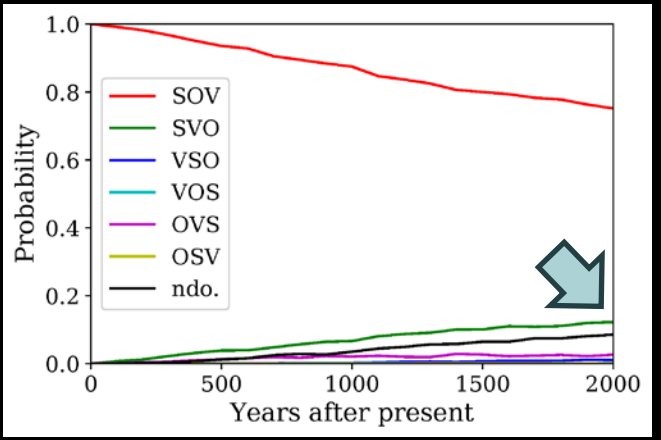

1.0
0.8
0.6
0.4
0.2
0.0
Probability
SOV
SVO
VSO
VOS
OVS
OSV
ndo.
0
500
1000
1500
2000
Years after present

回帰分析

- SVO語順を持つ言語のうち、どのような特徴を持つものが安定的?
  - 孤立語性と語順維持確率に高い相関
- 日本語が2千年後にSVO語順に変化する (12.3%) 場合、何が特徴的?
  - 格表示に接語 (「が」「を」) を使わない可能性が高い

このばらつきの説明





まとめと今後の課題

- 潜在表現への変換により言語の構造的特徴間の依存関係を捉える史的変化の分析手法を提案
- 今後の課題
  - 基本語順以外の特徴の分析
  - 祖語の推論結果の分析
  - 大語族の系統推定と世界祖語(?)の語順推定
  - 接触のモデル化 [Murawaki, NAACL2016]