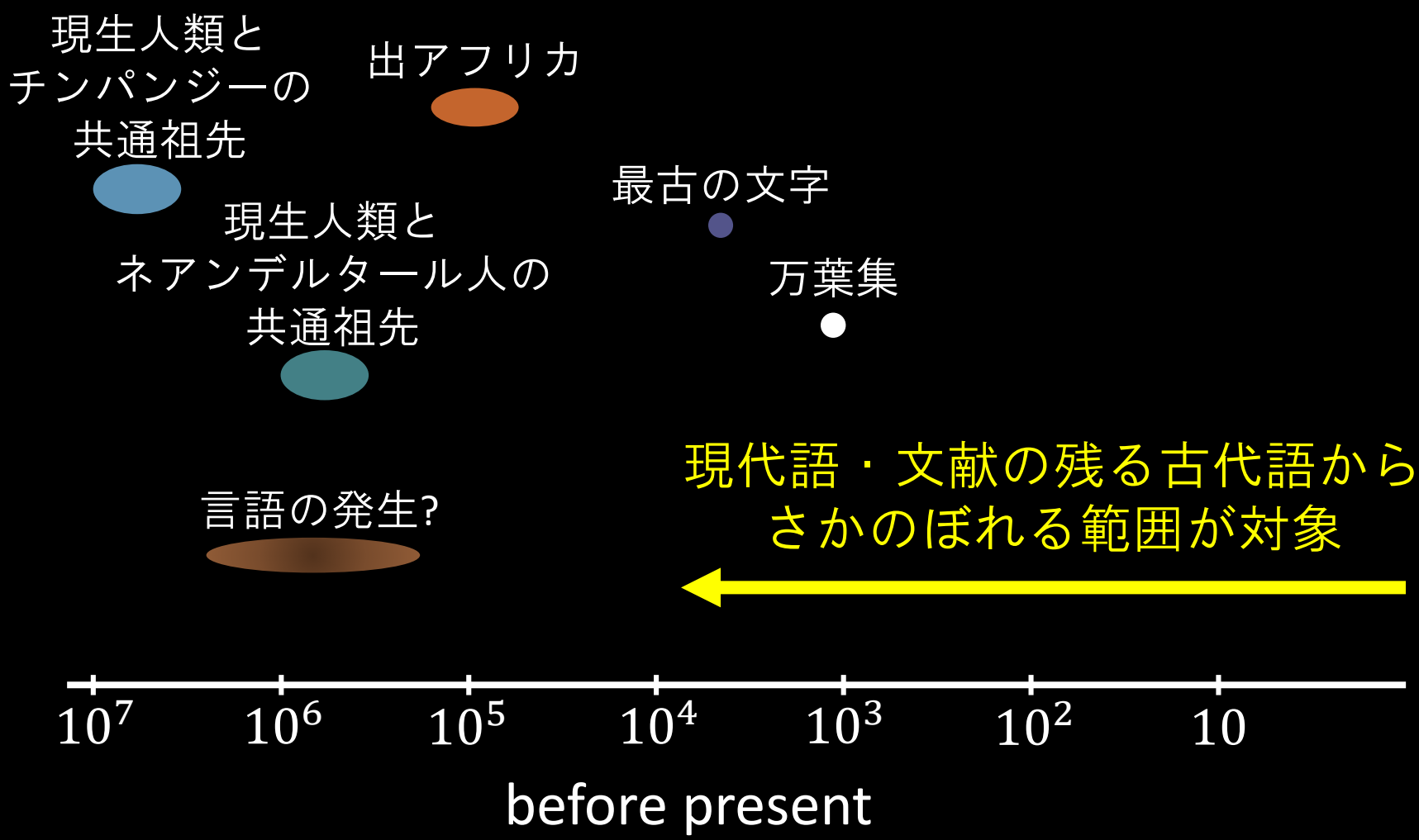
言語進化史の統計的研究

京都大学
村脇 有吾

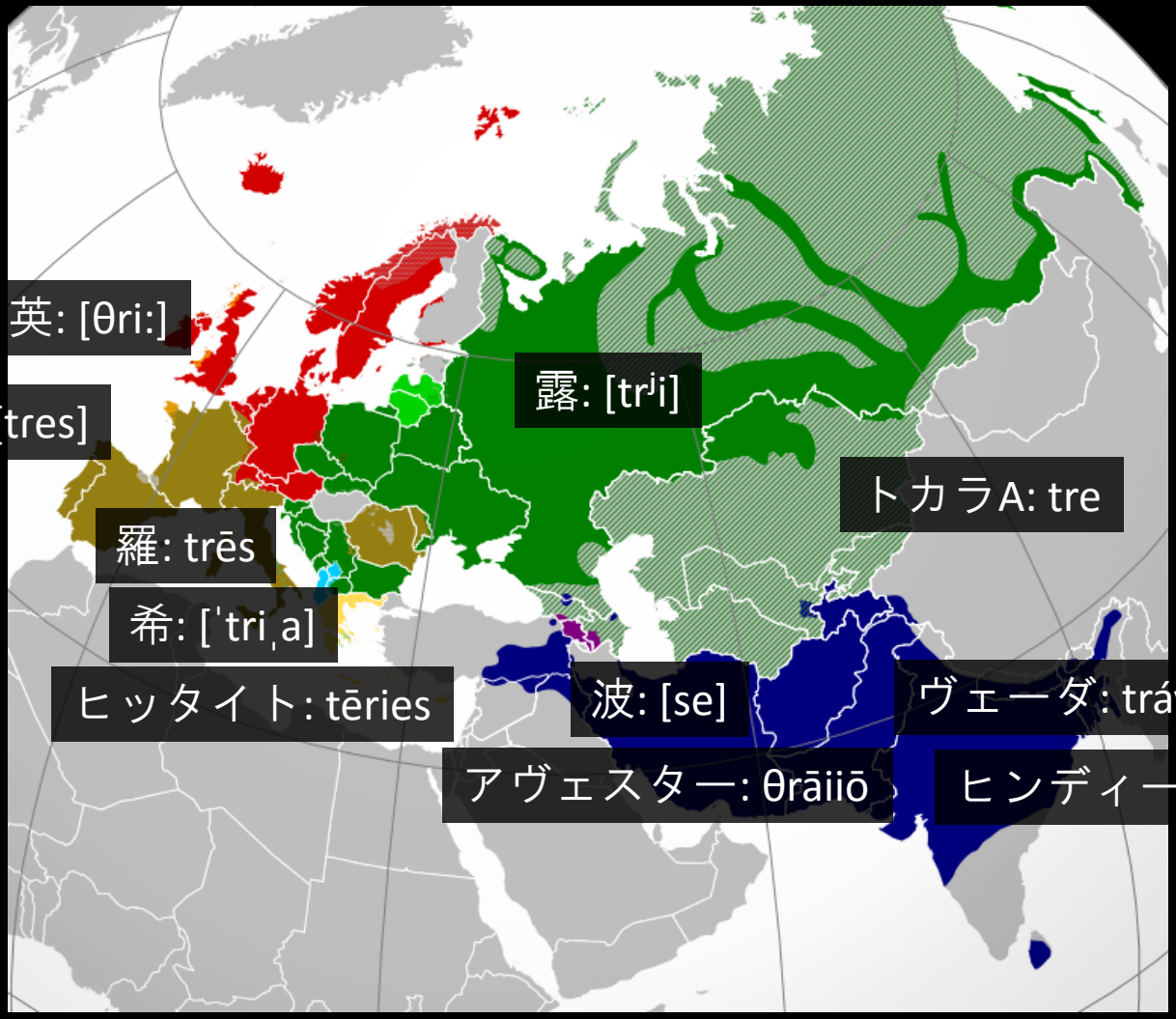
言語処理学会 第22回年次大会 チュートリアル資料 (2016年3月)

今日の話の対象範囲


現生人類と
チンパンジーの
共通祖先
出アフリカ
最古の文字
現生人類と
ネアンデルタール人の
共通祖先
万葉集
言語の発生?
現代語・文献の残る古代語から
さかのぼれる範囲が対象
10^7
10^6
10^5
10^4
10^3
10^2
10
before present

インド・ヨーロッパ (印欧) 語族

three



Sources:
Cognates: IELex
Map:
- Author: Alphathon
- License: CC BY-SA 3.0

インド・ヨーロッパ (印欧) 語族

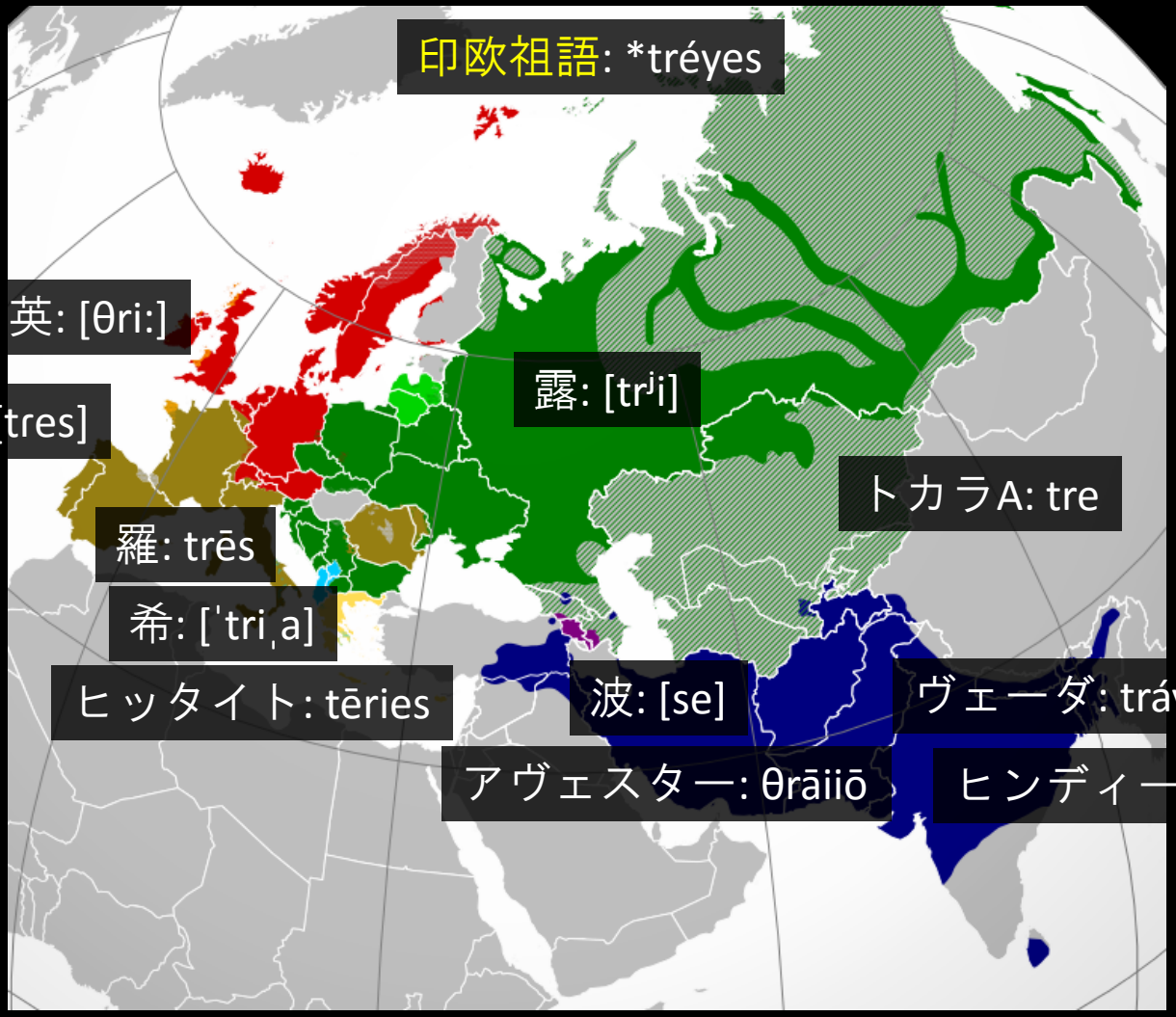
The *Sanscrit* language, whatever be its antiquity, is of a wonderful structure; more perfect than the *Greek*, more copious than the *Latin*, and more exquisitely refined than either, yet bearing to both of them a stronger affinity, both in the roots of verbs and the forms of grammar, than could possibly have been produced by accident; so strong indeed, that no philologer could examine them all three, without believing them to have sprung from some common source, which, perhaps, no longer exists;

Sir William Jones
(1746-1794)

Discourses delivered before the Asiatic Society: and miscellaneous papers, on the religion, poetry, literature, etc., of the nations of India (1824[1786])

インド・ヨーロッパ (印欧) 語族

three

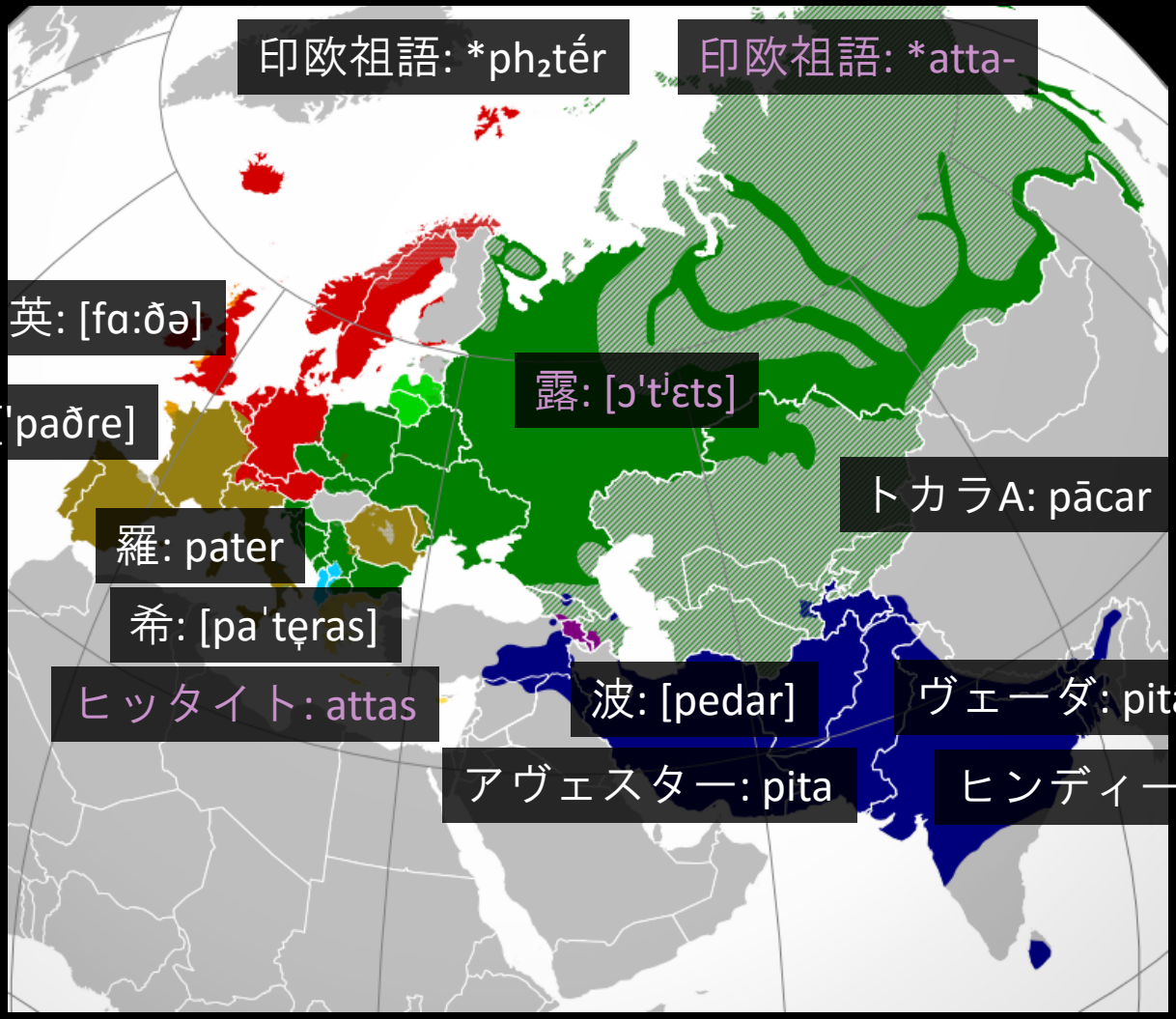


*は理論的な再構形を表す

Sources:
Cognates: IELex
Map:
- Author: Alphathon
- License: CC BY-SA 3.0

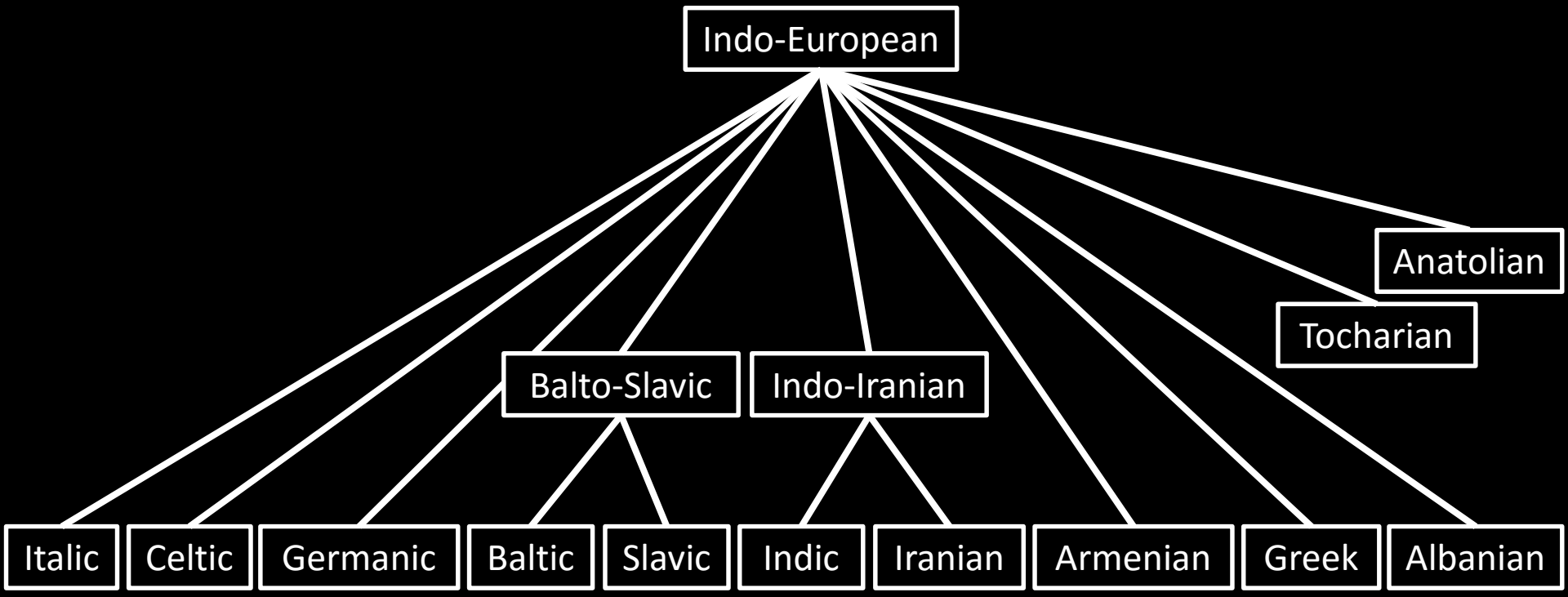
インド・ヨーロッパ (印欧) 語族

father



Sources:
Cognates: IELex
Map:
- Author: Alphathon
- License: CC BY-SA 3.0

インド・ヨーロッパ (印欧) 語族



※上位の分類には未確定部分が多い

印欧祖語の年代と故地 (Urheimat)

>200年間の研究で

- 系統樹はできた
  - 上位の分類はまだまだ怪しいけど
- 祖語の語形も復元できた
  - 細部については議論が絶えないけど

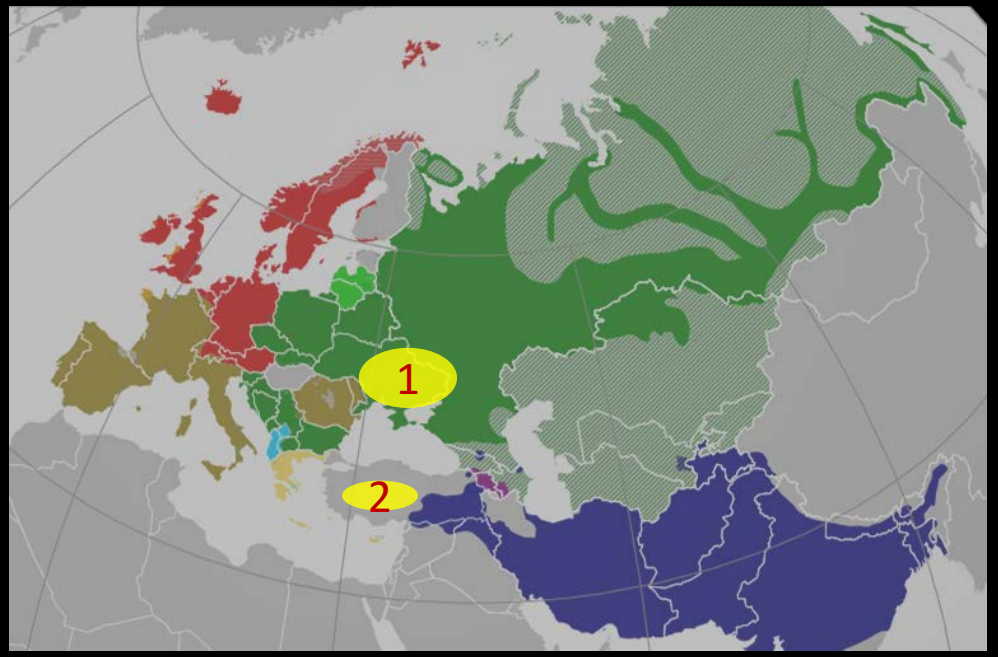
しかし祖語が

- いつ
- どこで

話されていたかは、従来手法では解けない

印欧祖語の年代と故地 (Urheimat)

1. クルガン仮説
   - 5,000-6,000年前
   - 黒海周辺のステップ
   - 遊牧民の軍事的征服
2. アナトリア仮説
   - 8,000-9,500年前
   - アナトリア
   - 農耕とともに拡大
     - Renfrew (考古学者) の農耕・言語同時伝播モデル



考古学の間接的手がかりだけでなく、言語データから直接的に年代と故地を推定したい

Source:
Map:
- Author: Alphathon
- License: CC BY-SA 3.0

計算機を使えば、年代 (と故地) を言語データから統計的に推定できる

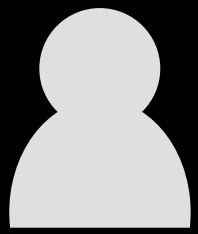
[Bouckaert+, Science 2012]

A copy of Figure S1 from Supplementary Materials for *Mapping the Origins and Expansion of the Indo-European Language Family* was removed.

From Bouckaert et al. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. Science 337(6097). Reprinted with permission from AAAS.

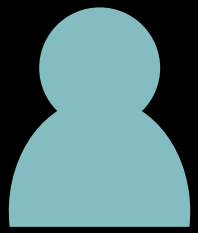
FAQ

そういう研究があるんですね。
この分野を何て呼ぶんですか?



某氏

Computational linguisticsです!!!
文字通り!



私

ACL 2016 has the goal of a broad technical program. Thus, ACL 2016 invites papers in the following categories:

- Applications/tools
- Empirical/data-driven approaches (submissions reporting negative results of sensible experiments are also welcome)
- Resources and evaluation
- Theoretical
- Surveys

Relevant topics for the conference include, but are not limited to, the following areas (in alphabetical order):

- Cognitive modeling and psycholinguistics
- Dialog and interactive systems
- Discourse and pragmatics
- Document analysis including text categorization, topic models, and retrieval
- Generation
- Information extraction, text mining, and question answering
- Machine learning
- Machine translation
- Multilinguality
- Phonology, morphology, and word segmentation
- Resources and evaluation
- Semantics
- Sentiment analysis and opinion mining
- Social media
- Speech
- Summarization
- Tagging, chunking, syntax, and parsing
- Vision, robots, and other grounding

Association for ***Computational Linguistics***!!!

http://acl2016.org/index.php?article_id=9

ACL系学会での発表例

- Grzegorz Kondrakのグループによる同源語自動発見 (NAACL2000, CoNLL2005, NAACL2009, ほか)
- Berkeley NLP Groupによる祖語の語形の自動再構と同源語自動発見 (EMNLP2007, NAACL2009, ACL2010, EMNLP2011, (PNAS 2013))
- Hal Daumé IIIによる言語類型論と地域言語学の統計モデル化 (ACL2007, NAACL2009)
- 私の類型論の研究 (NAACL2015, NAACL2016)

4. 申込の際には，口頭発表，ポスター発表，テーマセッションでの発表など，発表の種別を問わず，発表の「該当分野」として，以下の項目（小項目）から3種類選択していただきます．発表に関係が深いと思われる順に3種類選択してください．例えば，「語彙・辞書」であれば「B-1」を選択してください．指定して頂いた該当分野は大会プログラムの決定の際に重要な情報となります．不適切な分野を不適切な順序で選ぶと発表者にとって不本意なセッションでの発表が割当てられる可能性があります．その点を考慮して，該当分野はなるべく発表の実態に合ったものを選ぶようにして下さい．

A. 言語学・言語分析

（1）音声・音韻 （2）語彙・形態論 （3）統語論 （4）意味論
（5）語用論 （6）計量・コーパス言語学 （7）心理言語学
（8）認知言語学 （9）社会言語学 （10）対照言語学

B. 基盤技術・言語資源

（1） 語彙・辞書 （2） 形態素解析 （3） 構文解析
（4） 意味解析 （5） 談話解析（6） 固有表現解析
（7） 生成 （8） 言語資源・コーパス （9）アノテーション
（10）含意関係・言い換え （11）知識獲得 （12）文書分類
（13）機械学習 （14）マルチモーダル

C. 応用技術

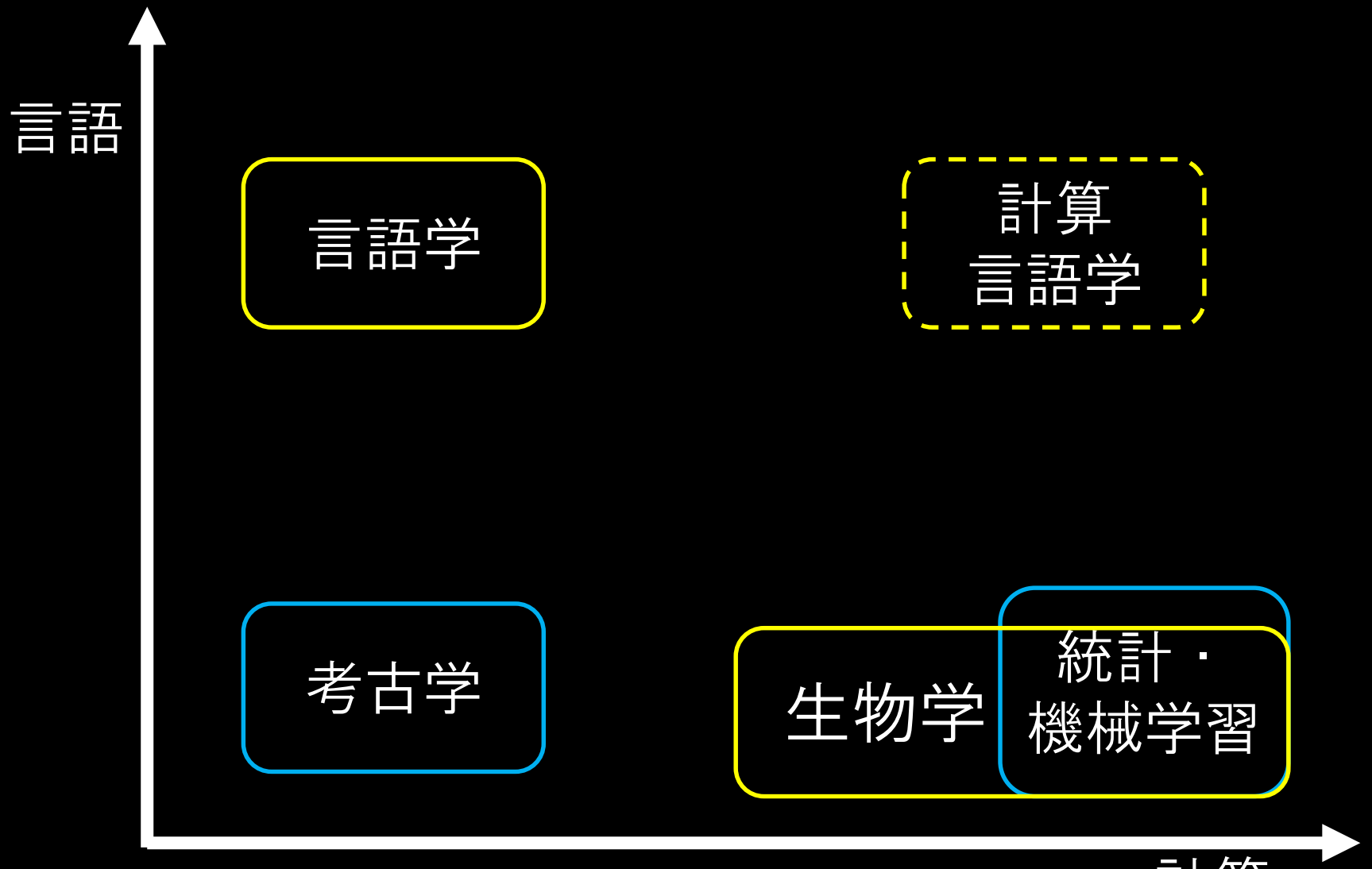
（1）機械翻訳 （2）情報検索 （3）対話 （4）要約
（5）情報抽出 （6）質問応答 （7）Web応用
（8）テキストマイニング （9） 評判・感情解析
（10）音声言語処理 （11）教育応用

D. その他 (　　)

テーマセッション

http://www.anlp.jp/nlp2016/submit.html

では誰が研究しているのか



言語処理研究者にとって いまが参入の機会

- 既存研究は以下のいずれかが欠ける傾向
  - (生物ではなく) 言語現象の理解
  - 言語向けの統計モデルの開発
- 統計的研究に不可欠な計算機可読な言語資源が整備されつつあるが、統計的分析が追いついていない (後述)

今日の目標

- この分野を認知する
- Bayes系統モデルがどういうものかを (なんとなく) 理解する
- 公開されているソフトウェアを使えば、とりあえず系統推定ができることを知る
- この分野に参入する気になる!

今日のおはなし

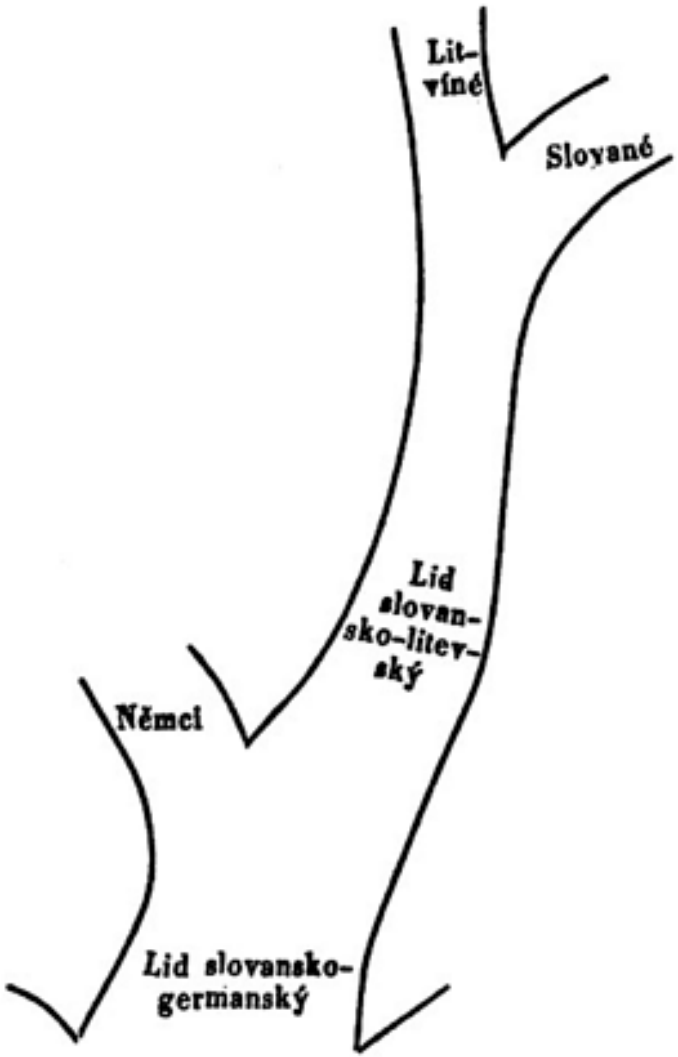
- 音法則からBayes系統モデルまで
- Bayes系統モデルのソフトウェア
- 言語資源
- 発展的な話題

今日のおはなし

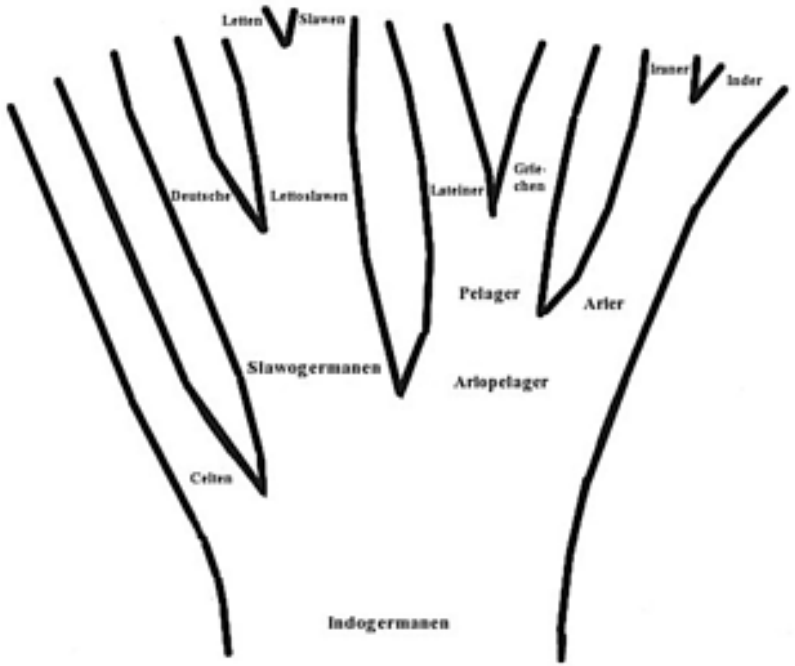
- 音法則からBayes系統モデルまで
- Bayes系統モデルのソフトウェア
- 言語資源
- 発展的な話題

音法則からBayes系統モデルまで

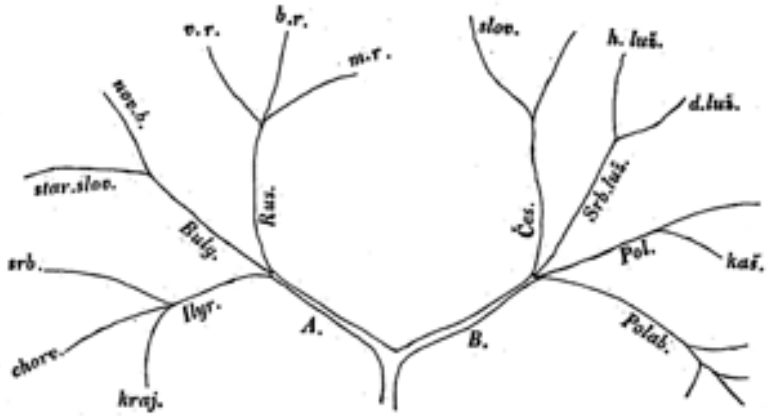
- 比較言語学のはじまり [Jones, 1786]
- 進化論と印欧語族の系統樹 [Schleicher, 1853]
- 青年文法学派 (19世紀後半) による音法則の確立
- 言語年代学 (1950年代) とその後の停滞
- 分子生物学由来のBayes系統モデルの導入 (2000年代-)



A) August Schleicher, 1853 [21]



B) August Schleicher 1853 [22]



C) František Ladislav Čelakovský 1853 [24]

Taken from [List+, 2013]

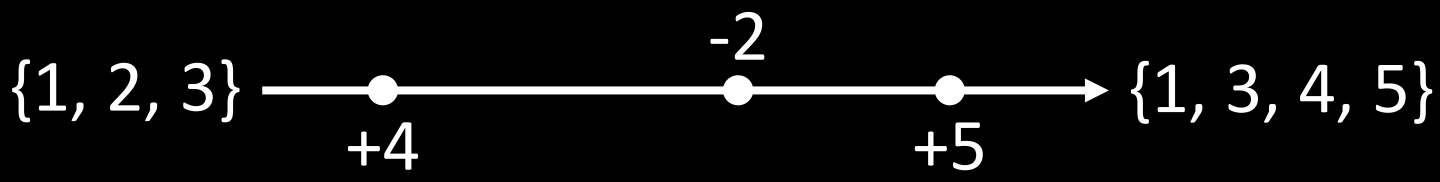
音法則からBayes系統モデルまで

1. 基本的な概念
2. 音法則に基づく祖語再構
3. 言語年代学
4. 確率モデルへ
5. Bayes系統モデル

音法則からBayes系統モデルまで

1. 基本的な概念
2. 音法則に基づく祖語再構
3. 言語年代学
4. 確率モデルへ
5. Bayes系統モデル

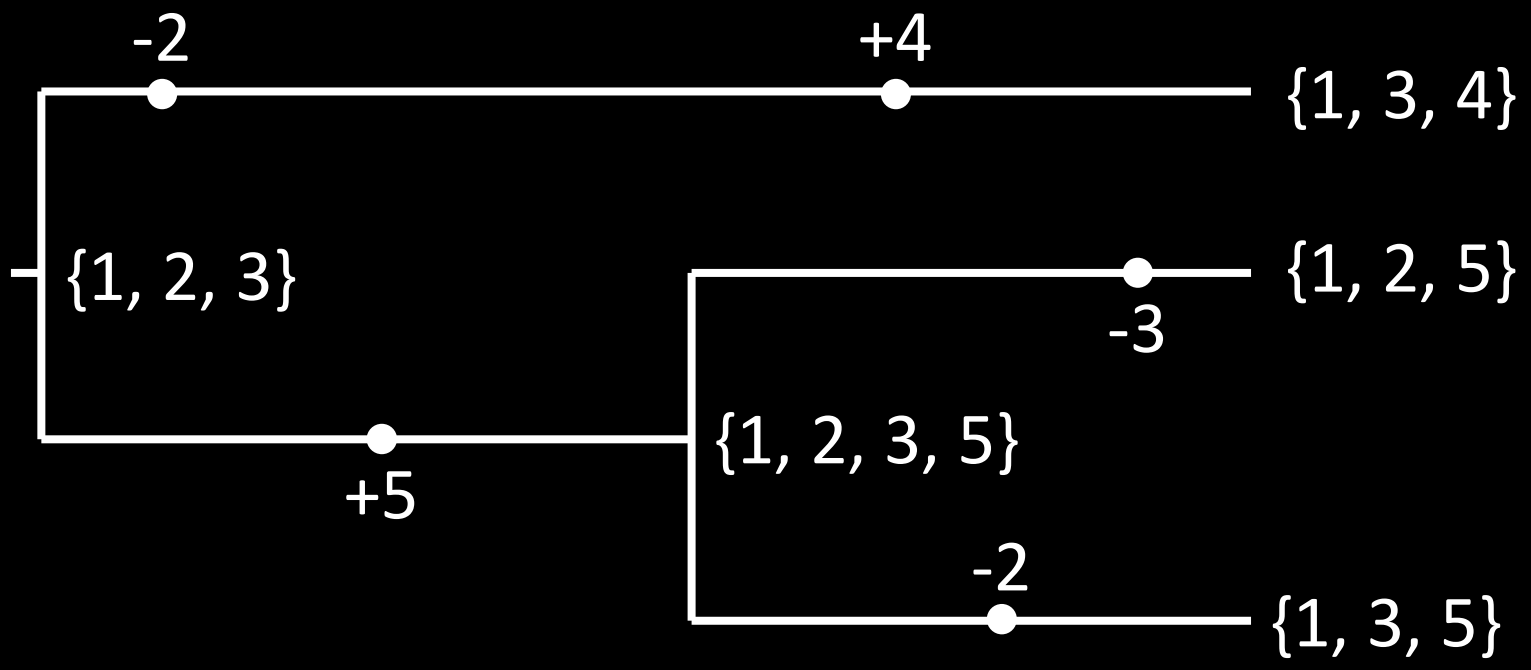
進化=変化を伴う由来 (descent with modification)



※特徴の例: ID化された単語

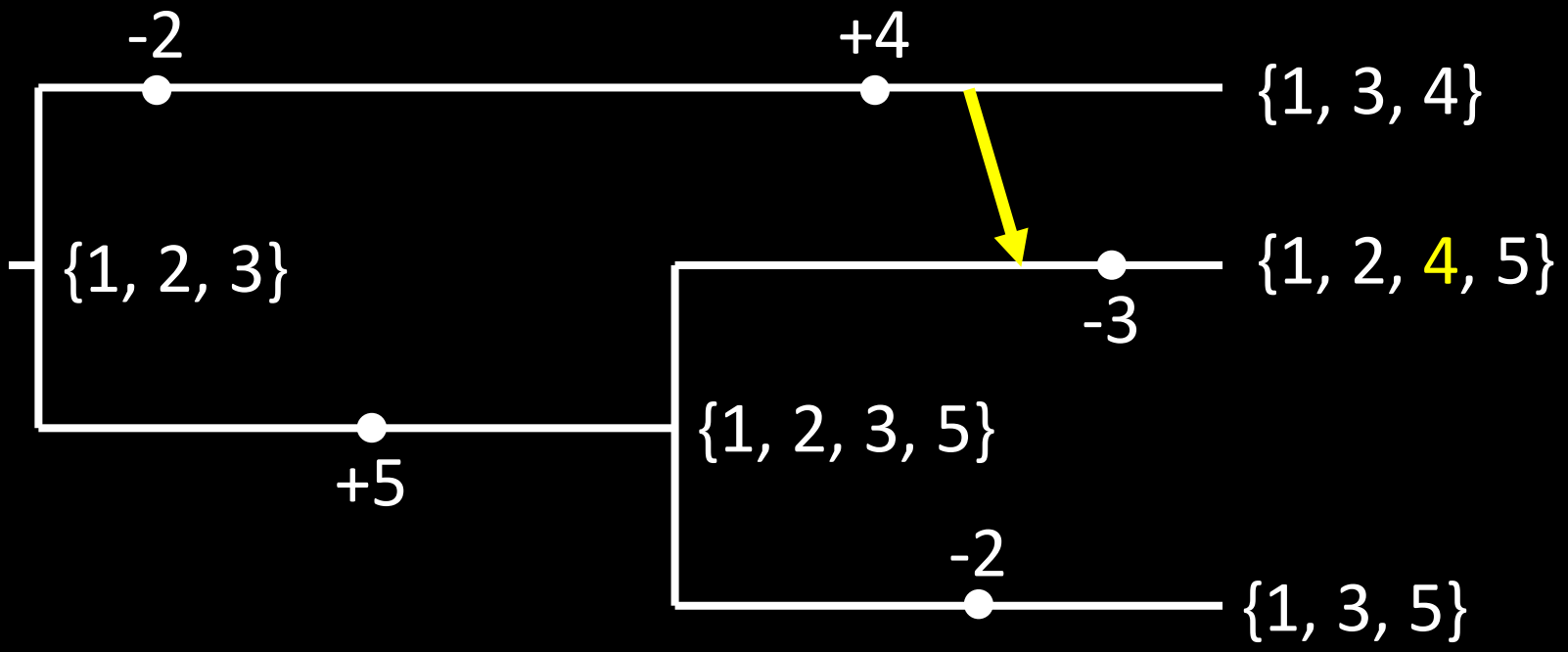
- 親から子へと途切れなく特徴が受け継がえる
- ただし、特徴は不変ではなく、次第に変化する
- 進化≠進歩 (価値判断を含まない)

系統樹 (phylogenetic tree)



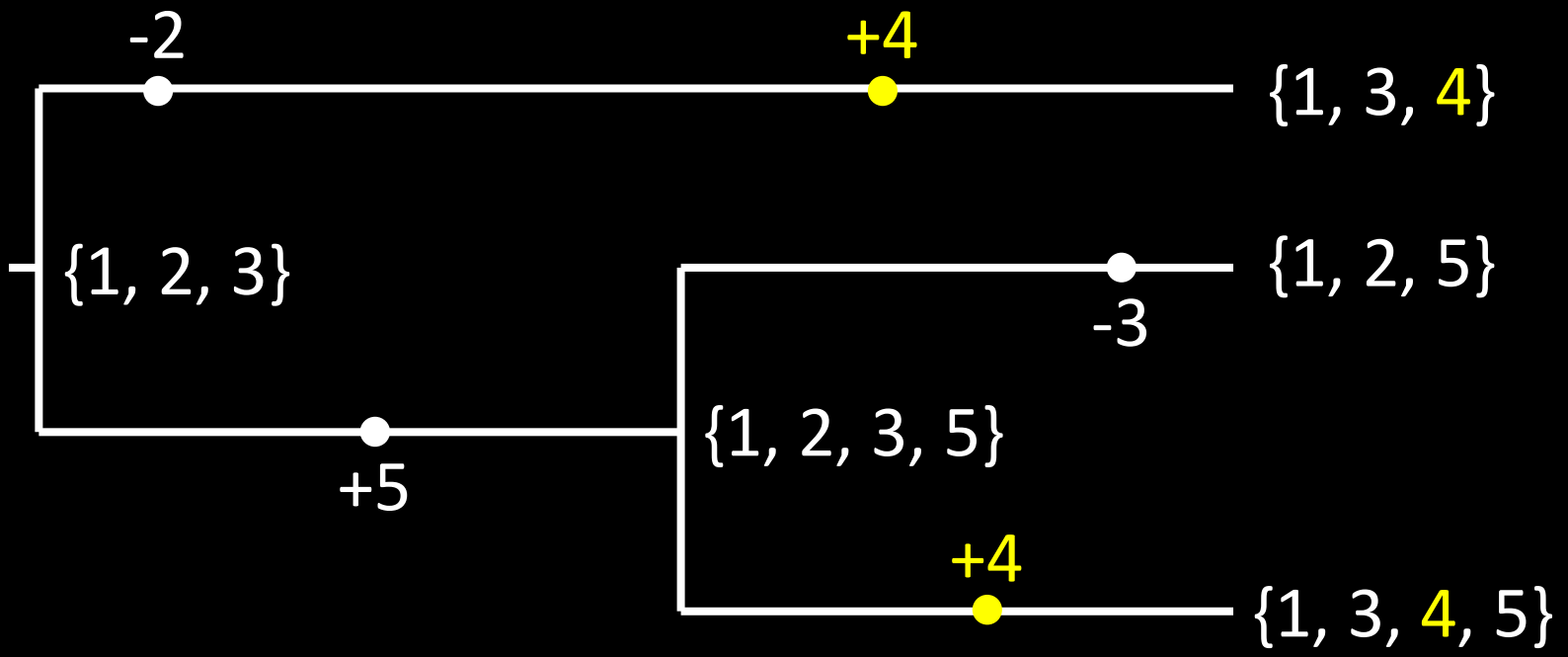
- 変化の速度がほぼ一定と仮定すると、言語ペアが多くの特徴を共有 ⇔ 比較的新しい共通祖先
- 素朴には、現代語群を距離に基づいてクラスタリングすれば系統樹相当のものが得られる

接触 (contact) による変化



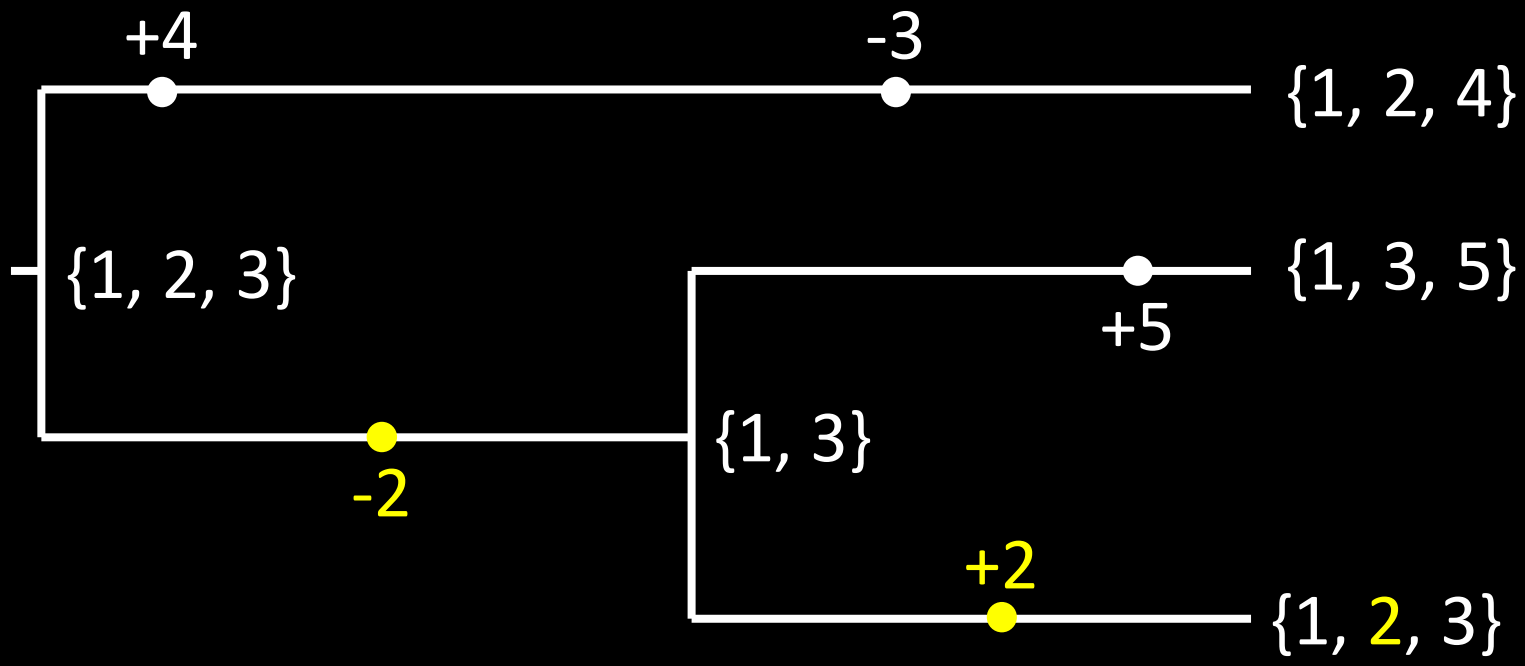
- 分岐後の接触は系統樹の仮定に反する
  - 水平伝播 (horizontal/lateral transmission) ともよばれる
    - cf 遺伝子の水平伝播 (horizontal gene transfer)
  - 伝播 (diffusion) も似た概念

成因的相同 (homoplasy)



- 同じ (似た) 特徴が独立に発生すること
  - 収斂 (convergence) とも
  - ≒平行進化 (parallel development)
- 系統推定の際にノイズになる

復帰突然変異 (back mutation)



- 一度死んだ特徴が復活すること
- これも系統推定の際にノイズになる

語彙に基づく手法

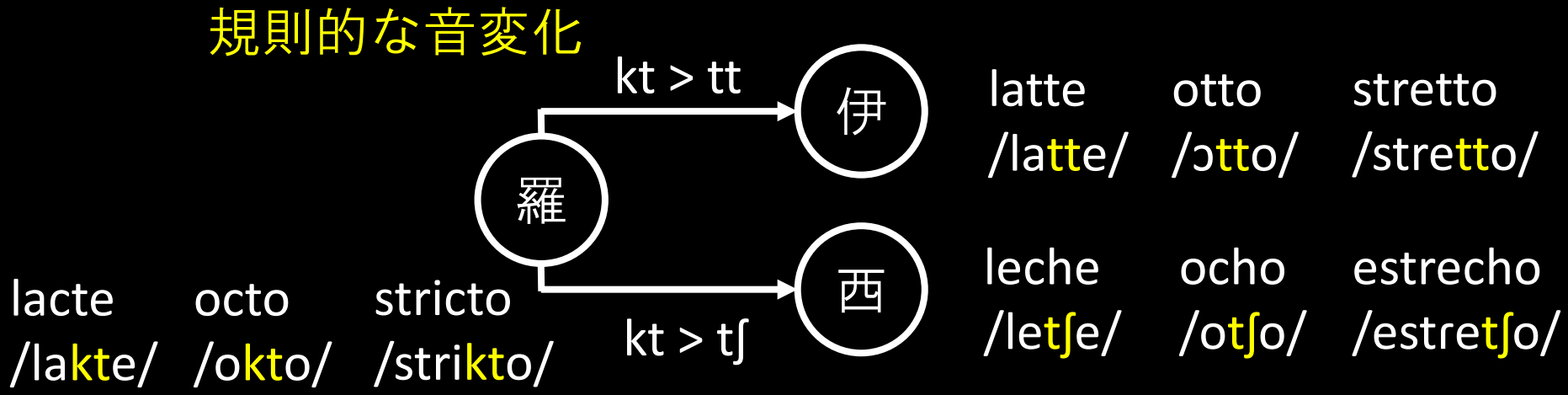
- 音法則、言語年代学、Bayes系統モデル (のほとんど) は特徴として語彙を用いる
- 記号の恣意性
  - 犬という概念と dog という音の結びつきに必然性はない
  - 同じ語が偶然複数回生まれる可能性は低い
    - homoplasy, back mutationは起こりにくい
- 接触 (借用) は起こりえる
- 語形自体は時間とともに変化する

Ferdinand Saussure
(1857-1913)

音法則からBayes系統モデルまで

1. 基本的な概念
2. 音法則に基づく祖語再構
3. 言語年代学
4. 確率モデルへ
5. Bayes系統モデル

音法則に基づく同源語 (cognate) 認定



言語ペアの持つ語が同一特徴 (同源語) である

⇔ 偶然や借用で説明できないほど類似

cf. 名前とname, 骨とbone

⇔ 規則的な音対応が存在

祖語再構はアート

1. 通言語的傾向
   - 起こりやすい変化
     - 不変化: *X* > *X*
     - 弱化: p > ɸ > h > Ø (zero), s > h
     - 有声音間の有声化: p > b / vocalic _ vocalic
   - 起こりにくい変化
     - k > a, a > k
     - 弱化の反対: h > p
2. 体系の自然さ
   - 5母音体系なら/a/の出現頻度は30~40%が普通で、あまりに少ないと不自然
3. 内的再構
   - 交替現象: k ~ g, s ~ z, t ~ d ⇒ *p ~ b

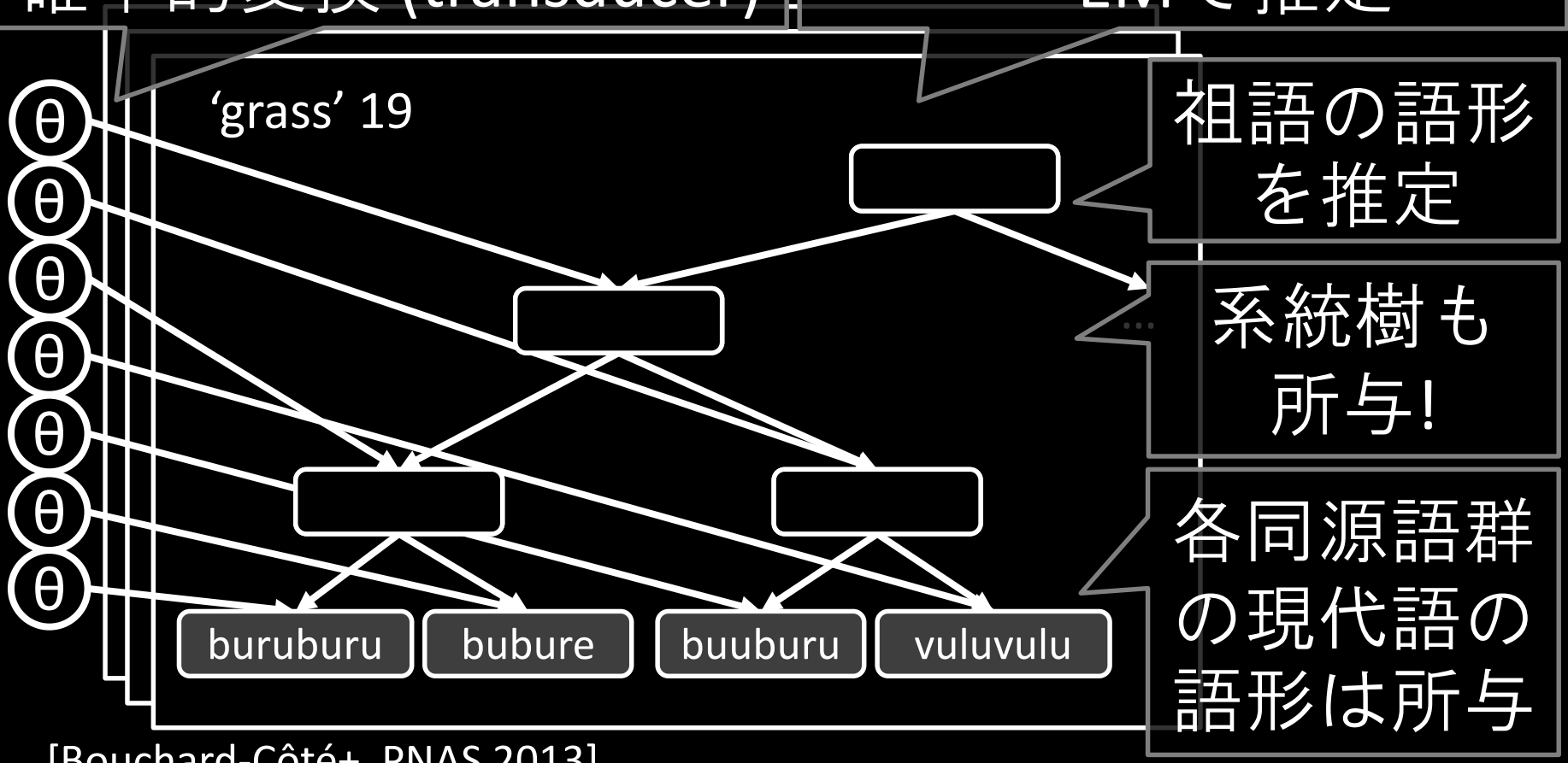
祖語再構の手続き

1. 対象の言語群について、同源語 (cognate) 候補を収集
2. 規則的な音対応を確立
   - 借用や偶然の一致を排除
   - 例外を個別に説明
3. 共通祖語を再構

統計モデルによる祖語再構 1/2

文字列から文字列への確率的変換 (transducer)

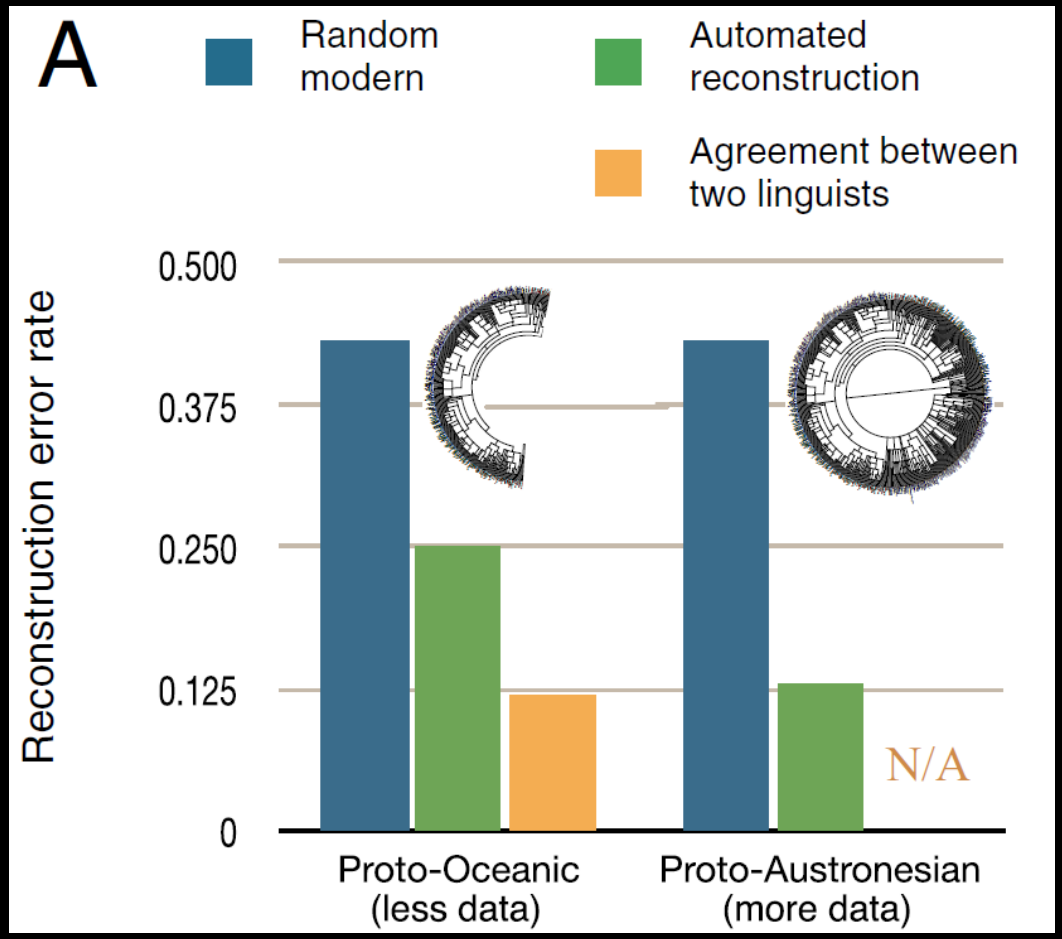
潜在変数とパラメータ
⇒ EMで推定



[Bouchard-Côté+, PNAS 2013]

統計モデルによる祖語再構 2/2

- オーストロネシア語族 (659言語)
- 言語学者の再構した語形にかなり近い
- 言語に関する新たな知見は?
- オーストロネシア語族以外は?



[Bouchard-Côté+, PNAS 2013]

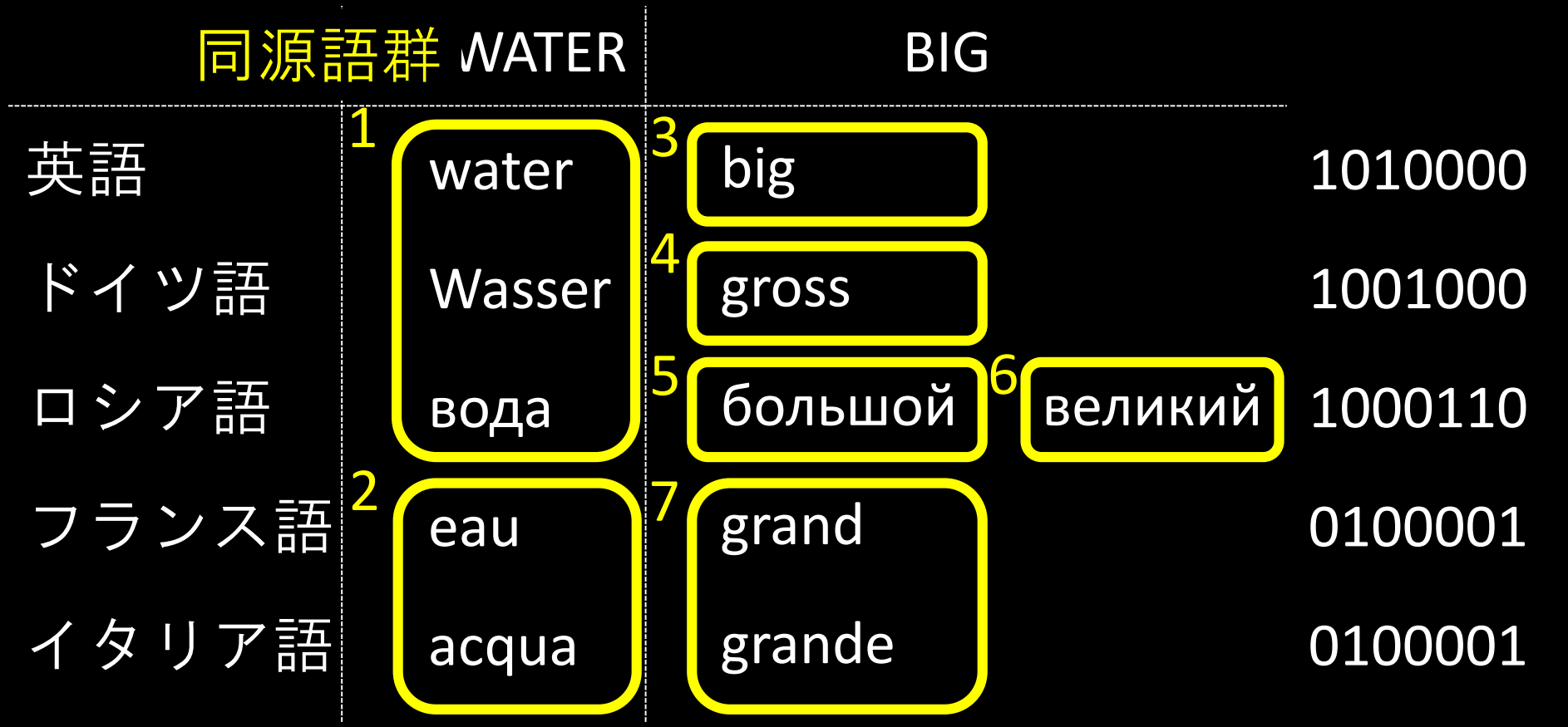
音法則からBayes系統モデルまで

1. 基本的な概念
2. 音法則に基づく祖語再構
3. 言語年代学
4. 確率モデルへ
5. Bayes系統モデル

基礎語彙 (basic vocabulary)

- どんな言語でもそれを表す言葉がありそうな基本的な概念100-200項目
  - ○ WATER, BIG, EYE, STAR, …
  - × SNOW, MILLION, PAPER, …
- 借用されにくく、比較的変化しにくい (と期待される)
  - 英語の一般語彙は50%が借用だが、基礎語彙に限ると6% [Swadesh, 1951]

語彙のバイナリ表現

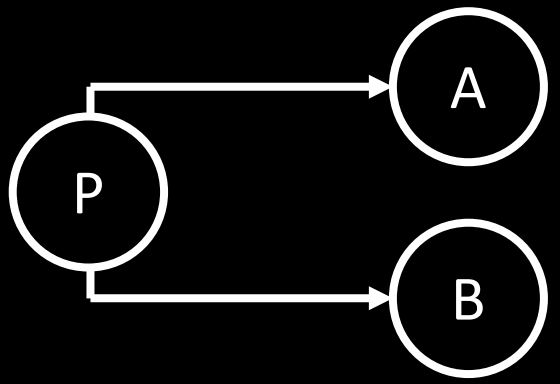

同源語群 WATER
BIG
英語
ドイツ語
ロシア語
フランス語
イタリア語
1
water
Wasser
вода
2
eau
acqua
3
big
4
gross
5
большой
6
великий
7
grand
grande
1010000
1001000
1000110
0100001
0100001

言語年代学 (glottochronology)

- 人類学・言語学のアメリカ・インディアン分類問題 [Sapir, 1921]
  - 音法則によらず、語彙の異同で分類したい
- 考古学での放射性炭素年代測定 [Libby, 1946]
- 祖語の年代推定への応用 [Swadesh, 1948,1951]
  - インド・ヨーロッパ語族から基礎語彙の残存率を求め、アメリカ・インディアンの言語に適用
- 日本語方言、アイヌ語方言への適用 [服部, 1954][服部+, 1960]
- 生物学の分子時計 (molecular clock) 仮説 [Zuckerkandl+, 1965] よりも早い!

言語年代学

$$t = \frac{\log c}{2 \log r}$$



- t: 祖語Pの年代 (単位: 千年)
- c: A, Bの基礎語彙共有率
- r: 基礎語彙の残存率 (200項目で0.81)

年代	1K	2K	3K	4K	5K	6K	7K	8K	9K
共有率 (r=.81)	.66	.43	.28	.19	.12	.08	.05	.03	.02

言語年代学への批判

- 基礎語彙の残存率が一定という仮定がなりたたない
  - 古ノルド語からアイスランド語への残存率は >0.95 [Bergsland+, 1962]
- 同系言語からの借用は区別が難しい
- 基礎語彙の中でも語によって安定性が異なるのでは?

言語年代学のその後

- 言語学者の激しい批判を受けて言語年代学は衰退
- より一般には語彙統計学 (lexicostatistics) の研究が (細々と) 続けられた
- 後発の分子生物学由来のモデルで置き換えられた (2000年代-)
- 手法自体は廃れたが、作成された語彙データベースは、Bayes系統モデルでも使われている

音法則からBayes系統モデルまで

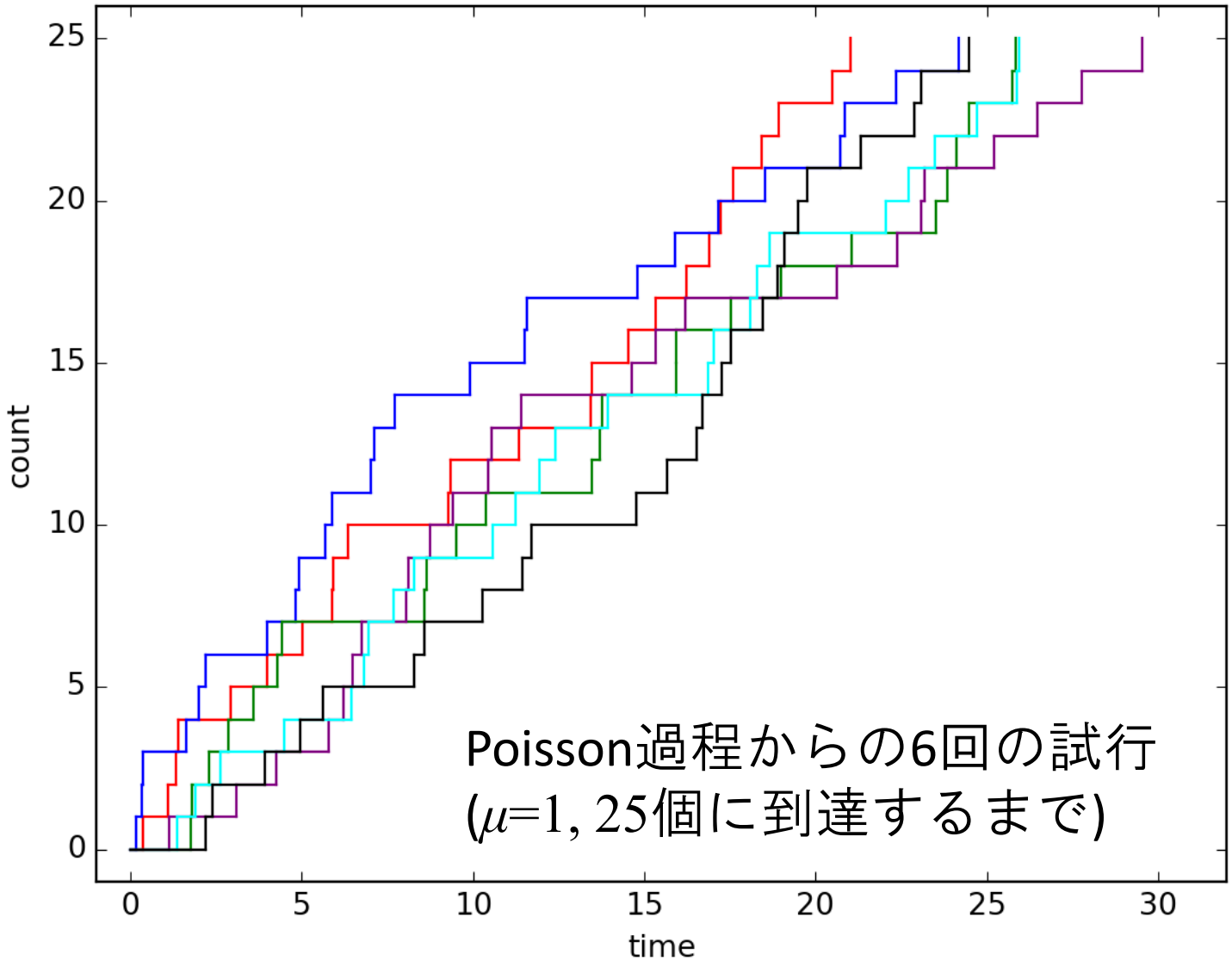
1. 基本的な概念
2. 音法則に基づく祖語再構
3. 言語年代学
4. 確率モデルへ
5. Bayes系統モデル

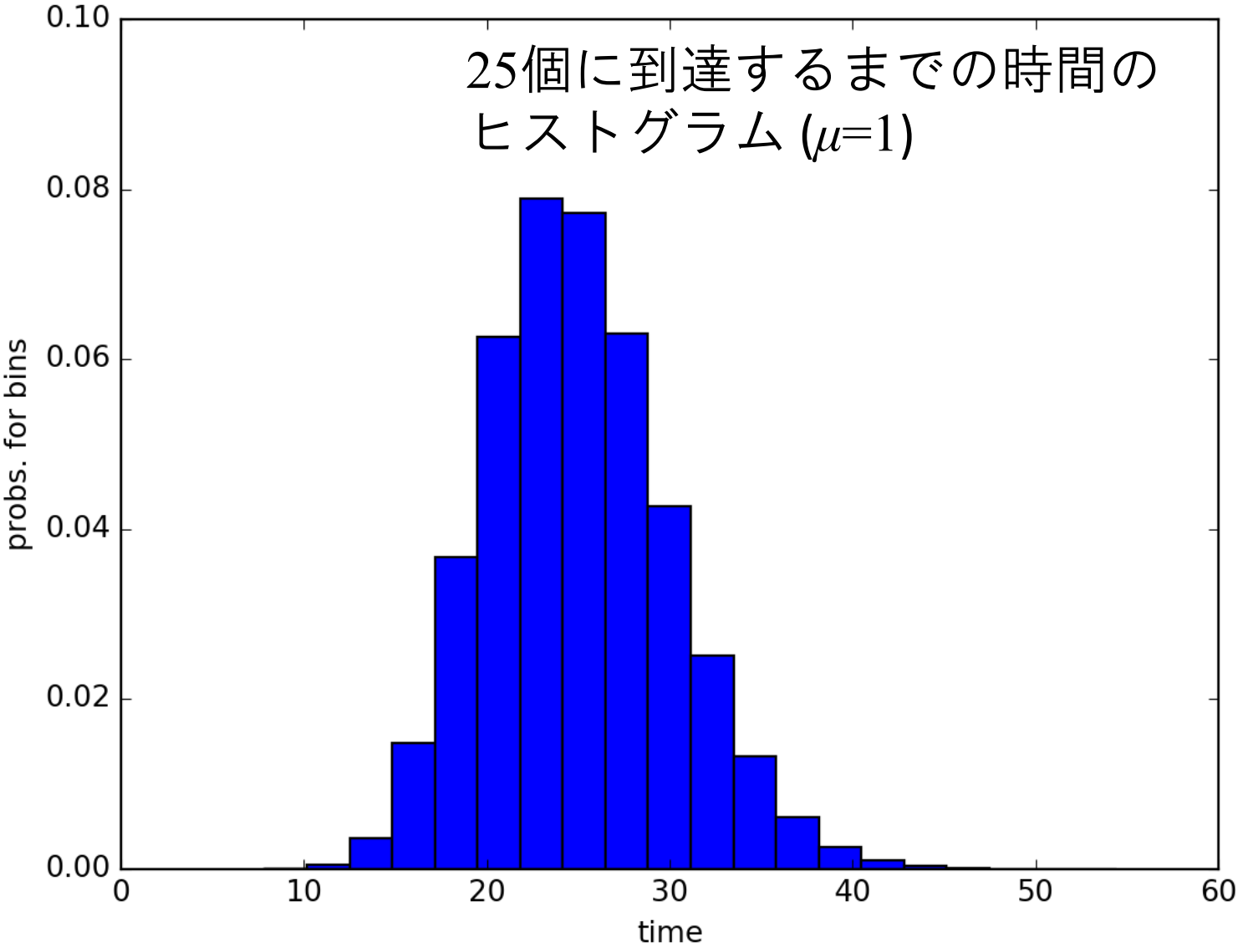
言語変化の確率モデル化

- 変化はメトロノームのように一定間隔で起きるわけではない
- 変化の間隔には確率的ゆらぎがあり、仮に一定だとしても、それは確率分布のパラメータでは?

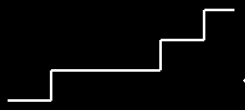
Poisson過程 (process)

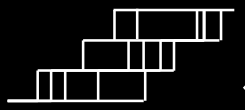
- 独立に発生する事象 (e.g. 言語の変化) を数える確率モデル
- X_t: 時間幅 $[0, t]$ で起きた変化の数
  $X_t \sim \mathrm{Poisson}(\mu t)$
  $X_t - X_s \sim \mathrm{Poisson}\big(\mu(t-s)\big)$
  μ: 変化率
- 2つの連続した事象の間隔は$\mathrm{Exp}(\mu)$に従う


25
20
15
10
5
0
count
0
5
10
15
20
25
30
time
Poisson過程からの6回の試行
(μ=1, 25個に到達するまで)

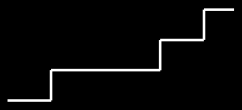

25個に到達するまでの時間の
ヒストグラム (μ=1)
0.10
0.08
0.06
0.04
0.02
0.00
probs. for bins
0
10
20
30
40
50
60
time

言語変化の確率モデル化

- 言語年代学の手法:
  モデル(言語ペア; r) = $time$
- 確率モデル:　　※時間は入力の一部
  モデル$_1$( ; μ) = $score$　(確率)

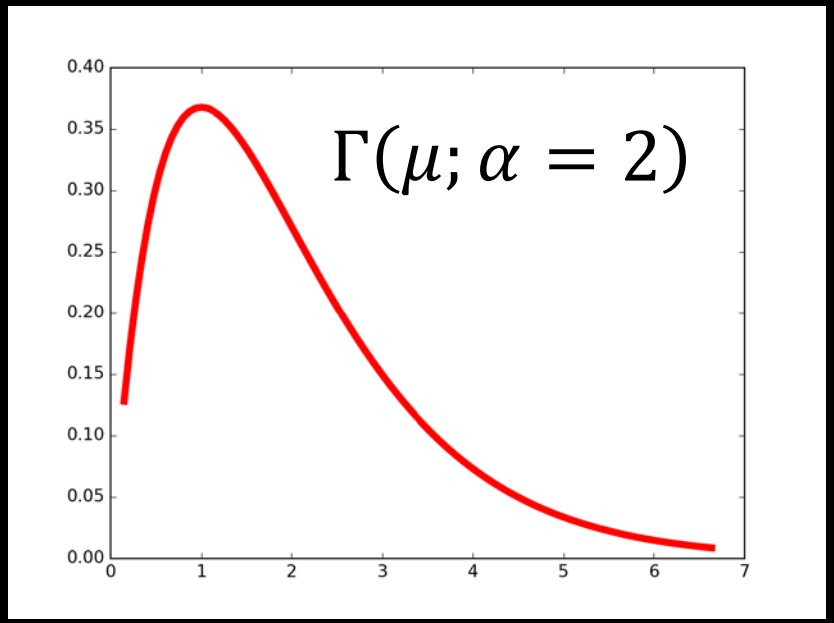
  モデル$_2$( ; μ) = $\int$ モデル$_1$( ; μ)
  = $score$　(確率)

Bayesモデル

- モデル$_{\text{Bayes}}$( , μ ; α)
  ∝モデル$_1$( | μ) prior(μ ; α)

ハイパーパラメータ

モデルのパラメータに事前分布を与える
= スコアを与える

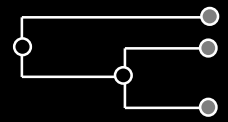


音法則からBayes系統モデルまで

1. 基本的な概念
2. 音法則に基づく祖語再構
3. 言語年代学
4. 確率モデルへ
5. Bayes系統モデル

Bayes系統モデル

1. モデル設計
   - 自然な系統樹に高いスコア (確率) を与えるような

$$\text{モデル}\left(\ \text{  }\ , \underline{\boldsymbol{\theta}}; \underline{\boldsymbol{\alpha}}\right) = score$$

パラメータ (確率変数)　ハイパーパラメータ

   を設計
2. 推論
   - 与えられたモデルのもとで高いスコアを返すような系統樹+パラメータを探す
   - Markov chain Monte Carlo (MCMC) サンプリングを用いる

Bayes系統モデルの設計

- 系統樹とパラメータを構成要素 (部分モデル) に分解して採点
  - $\log(\mathrm{P}_{\text{部分モデルA}} \times \mathrm{P}_{\text{部分モデルB}})$
    $= \log(\mathrm{P}_{\text{部分モデルA}}) + \log(\mathrm{P}_{\text{部分モデルB}})$
    なので、対数化するとスコアは足し算になる
- 部分モデルやその組み合わせには多数の変種がある
  - 現在も活発に研究されている

Bayes系統モデルの設計

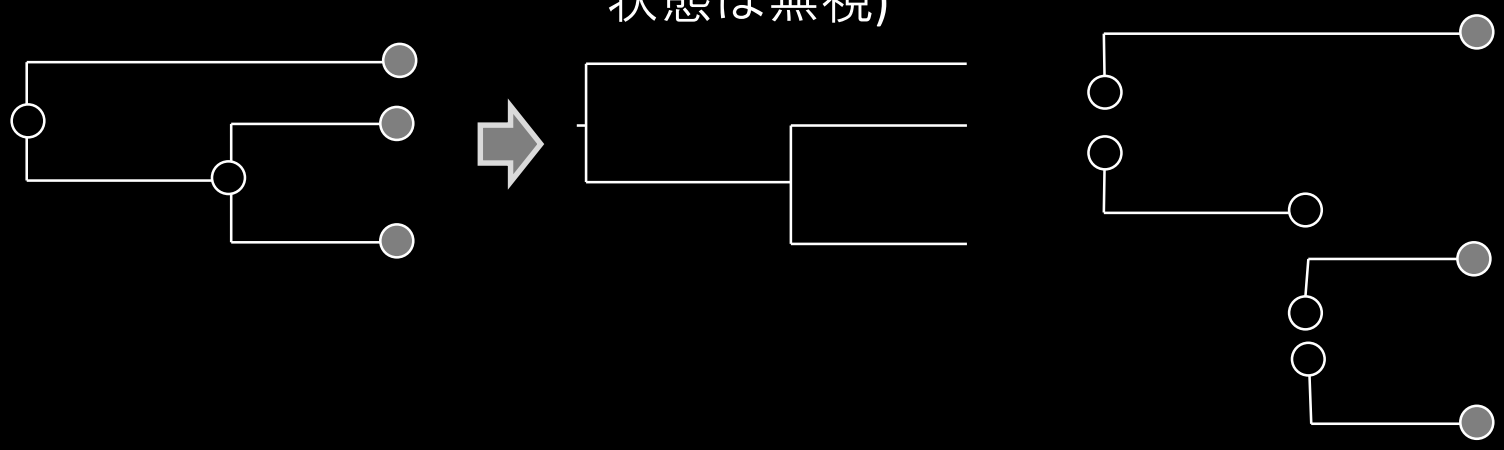
- 部分モデルは大きく3つ (他にも各種事前分布)
  1. 木モデル
  2. 置換 (substitution) モデル
  3. 時計 (clock) モデル

3. 置換の速度 (時計)

2. 親から子への遷移時のノードの各要素の置換

1. 木 (ノードの状態は無視)

系統樹

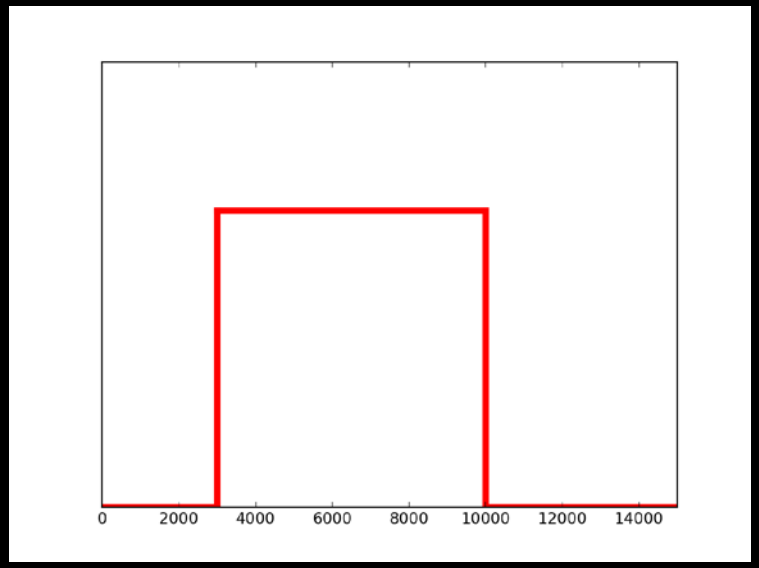


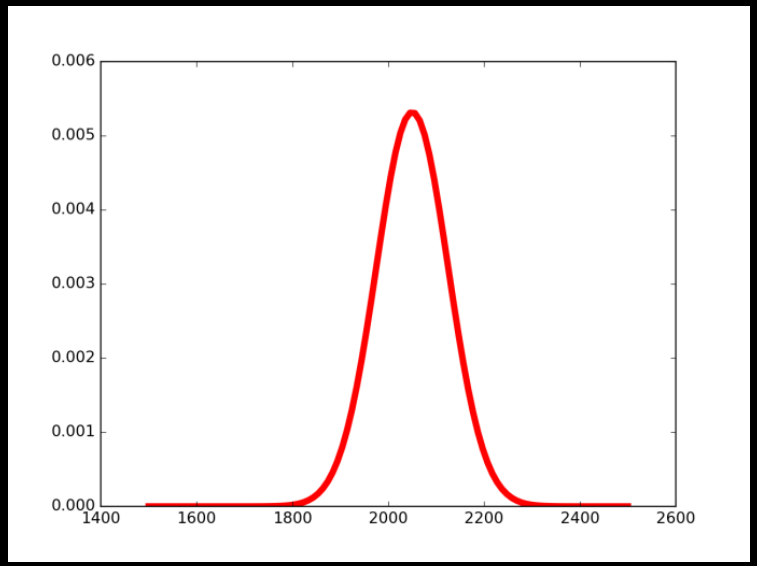
木モデル

- 時間付きの木 (の骨組み) を採点
- モデルの例
  - Yule Process
  - Birth-death model
  - Bayesian skyline model
  - …
- 言語系統樹における意味?

年代較正 (calibration)

- 内部ノード (e.g. インド・イラン祖語) や葉ノード (e.g. ラテン語) の年代はおおよそ既知
  - 生物の場合は化石の年代など
- こうしたソフトな制約を事前分布としてモデルに組み込む
- 推論時には、これらのソフトな制約を満たすように変化率が推定される





置換 (substitution) モデル

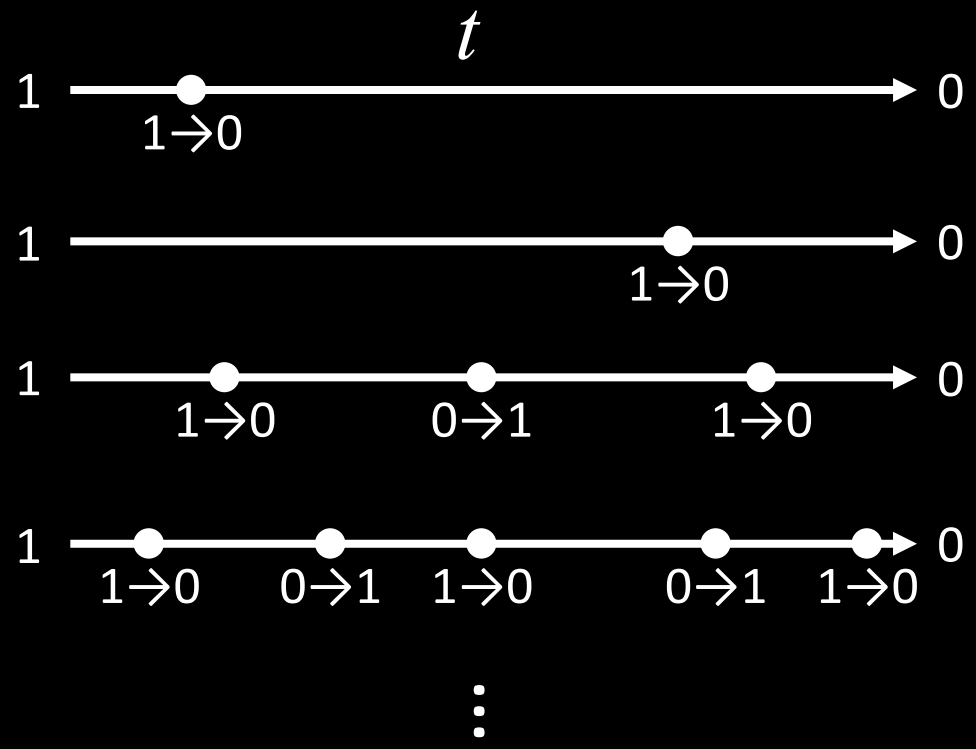
遷移確率: $P(x = j|\pi(x) = i, t) = \exp(tQ)_{i,j}$

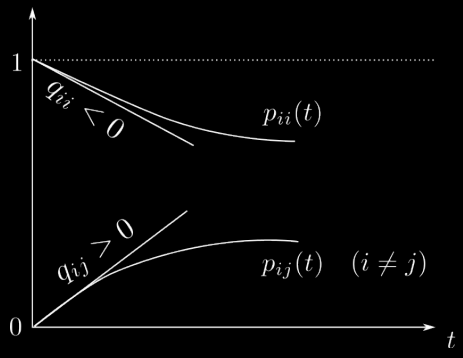
$$Q = \begin{pmatrix} * & \pi_C & \pi_A & \pi_G \\ \pi_T & * & \pi_A & \pi_G \\ \pi_T & \pi_C & * & \pi_G \\ \pi_T & \pi_C & \pi_A & * \end{pmatrix} \qquad Q = \begin{pmatrix} * & \alpha \\ \beta & * \end{pmatrix}$$

連続時間マルコフモデル

$P(x = 0|\pi(x) = 1, t) = \exp(tQ)_{1,0}$

1から始まり、時間 t 後に0となっている
すべての遷移を積分したもの



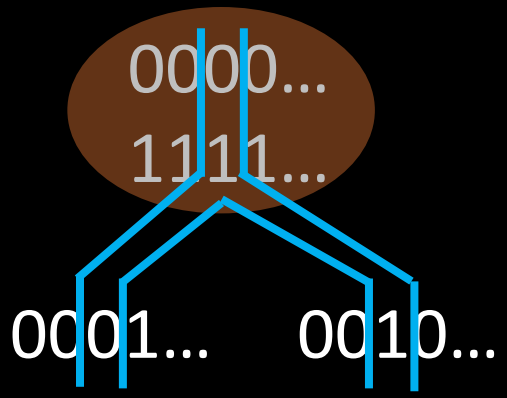


言語の場合、back-mutation (語彙の復活)は不自然。復活のないモデルとして、確率的Dolloモデルが提案されている [Nicholls+, 2008]

配布版にtypo

内部状態の積分消去

- すべての内部ノードの状態候補を一度に考慮する
  - 陽に内部ノードの状態を持つよりも推論が効率的になる
- 動的計画法により効率的に解ける [Felsenstein, 1981]



時計 (clock) モデル

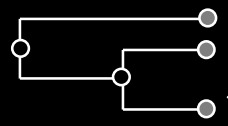
- 厳密時計 (strict clock): 系統樹全体で同じ遷移率
- 緩和時計 (relaxed clock): 系統樹中の場所ごとに異なる遷移率

$$P(x = j|\pi(x) = i, t, k) = \exp(t r_k Q)_{i,j}$$

  - 枝 k ごとに異なる係数 r_k をかける
  - r_kの確率分布の設計に様々な変種がある
    - 分子生物学でもモデルが提案され始めたのは1990年代後半からで、決定版がまだない

推論 (inference)

- 目的: モデルのもとで高いスコアを返すような系統樹 (+パラメータ) を探す

  モデル( , $\boldsymbol{\theta}; \boldsymbol{\alpha}$) = $score$
- 課題: 系統樹を決めるとスコアが返ってくるが、良いスコアを返す系統樹は (解析的には) 求められない
- 解決策: スコアが (だいたい) 上がるように、少しずつ系統樹を変更していく

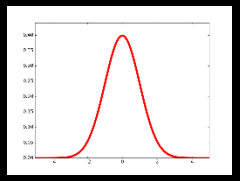
推論: MCMCサンプリング

1. 適当に初期系統樹 (+パラメータ) を決める
2. 系統樹の一部をランダムに変更した新しい系統樹を提案
3. ある確率にしたがって
   - 採択: 提案された系統樹を採用
   - 棄却: もとの系統樹を採用
4. 2-3をひたすら繰り返すと、次第にスコアが高い系統樹となる

サンプリング

- 確率分布からサンプルを得る手続き

3, 4, 5, 3, 1, 4, 6, …



1.78, -0.32, 0.27, -1.05, -0.11, …

$(\quad, \boldsymbol{\theta}), (\quad, \boldsymbol{\theta}), (\quad, \boldsymbol{\theta}), \ldots$

Source:
Dice:
- Author: Steaphan Greene
- License: CC BY-SA 3.0

サンプリング: 補助変数法

- $p(x)$から直接サンプリングするのが難しい場合に、同時分布$p(x,u)$からのサンプル$(x_1,u_1),(x_2,u_2),\cdots$を得る
- uを無視してxだけに注目すれば、$p(x)$からのサンプルになっている

- 例: 系統樹のサンプルから根の年代の分布を得る

Bayes系統モデルによる印欧祖語の年代推定 (再掲)

※Bouckaert+ (2012) は
祖語の位置も同時推定

[Bouckaert+, Science 2012]

A copy of Figure S1 from Supplementary Materials for *Mapping the Origins and Expansion of the Indo-European Language Family* was removed.

From Bouckaert et al. 2012. Mapping the Origins and Expansion of the Indo-European Language Family. Science 337(6097). Reprinted with permission from AAAS.

今日のおはなし

- 音法則からBayes系統モデルまで
- Bayes系統モデルのソフトウェア
- 言語資源
- 発展的な話題

実行するだけなら
系統推定は難しくない

- Bayes系統モデルの実装はそれなりに複雑
- ソフトウェア・パッケージ化されている
  - 生物系の研究では、モデルを作ってソフトウェア化する人と、それを利用してデータを分析する人が分離する傾向にある
- GUIでモデル設定を決めれば、(細部まで理解していなくても) 実行できる
- モデル設定を含むデータは論文の補助資料として公開されていることが多いので、まずは追試から始めればよい

Bayes系統モデルのソフトウェア例

- BEAST / BEAST2
- MrBayes
- PhyloBayes
- MCMCTree
- DPPDiv
- Multidivtime
- …

その他のソフトウェア

系統樹が与えられたもとで、他の情報 (e.g. 祖語の状態) を推定するソフトウェア

- BayesTraits
- Mesquite (描画やその他の機能も)

Bayesではない従来の系統推定

- SplitsTree (距離ベースの手法)

NEXUS
(生物系の共通フォーマット)

```
#nexus
BEGIN DATA;
DIMENSIONS ntax=19 nchar=350;
FORMAT
                datatype=standard
                symbols="01"
                missing=?
                gap=-
                interleave=NO
;
MATRIX

Yakumo                          10011101001010100010011111011111111101110100101011111110111111
11110110010010101010110000111111101110100100001110110111010100110011110010101100111111
00111110111110110100100100101001010010111111011111101001100000001101110101001111100001
00011111000100010000111101111111111111110011111011001111110100100010010010101010110100
00101001001101100101111010000101001^MOshamambe                 10011101001O??1000100111110
10111101101110100101011111101111111110101010010101010101000111111101110100100001101111
01110101001100111100101011001111???111110111110101100100100101001010010111111101111110
10011000000011010101010011110100100011111000100010000111101111111111111110011111001001
11111010010001001001010101???10000101001000101100101111010000101001^MHorobetsu
        10011101001110100010011111010111011011101001010111111011111111101010100101010101
01001001111111011101001000011101101110101001010111100101011001111010111111011111011010
01001001010010100111111110111111010010100000011011010101011111100100011111100100000100
01111011111111111111001111100100111111010010001001001110111010010000100101001101100010
11110110000011001^MHiratori    10011100100110100010011111011111101101110100101011111101111
11111101100100101010101010001111111011101001000011101101110101001010111100101011001111
```

ヘッダ
19言語、各言語350要素
欠損値は “?”

データブロック
各言語のデータ

Data Source:
Lee & Hasegawa (2013)

Newickの木
(これも生物系の共通フォーマット)

系統樹が枝の長さ付きの二分木として表現されている

系統樹の初期値や推定結果のフォーマットとして利用

((Soya:677.9837811542398,((Asahikawa:284.0551009118723,Nayoro:284.0551009118723):204.84575442059077,((Samani:302.37881916856713,(Bihoro:188.05996265030115,(Obihiro:150.60102259767746,Kushiro:150.60102259767746):37.45894005262369):114.31885651826599):153.43733621849776,((((Niikappu:93.41464071381677,Nukkibetsu:93.41464071381677):34.05188504173179,Hiratori:127.46652575554856):131.61755013591392,Horobetsu:259.0840758914625):83.35489289078737,(Yakumo:111.24099186141609,Oshamambe:111.24099186141609):231.19797692083375):113.37718660481505):33.08469994539814):189.08292582177 67):909.7742355049573,((Ochiho:325.41631213730045,((Raichishka:224.95428536409622,Shiraura:224.95428536409622):62.589987065565 2,Nairo:287.5442724296614):37.87203970763903):37.83126605465577,(Tarantomari:216.23752826690372,Maoka:216.23752826690372):147.0100499250525):1224.5104384672409)

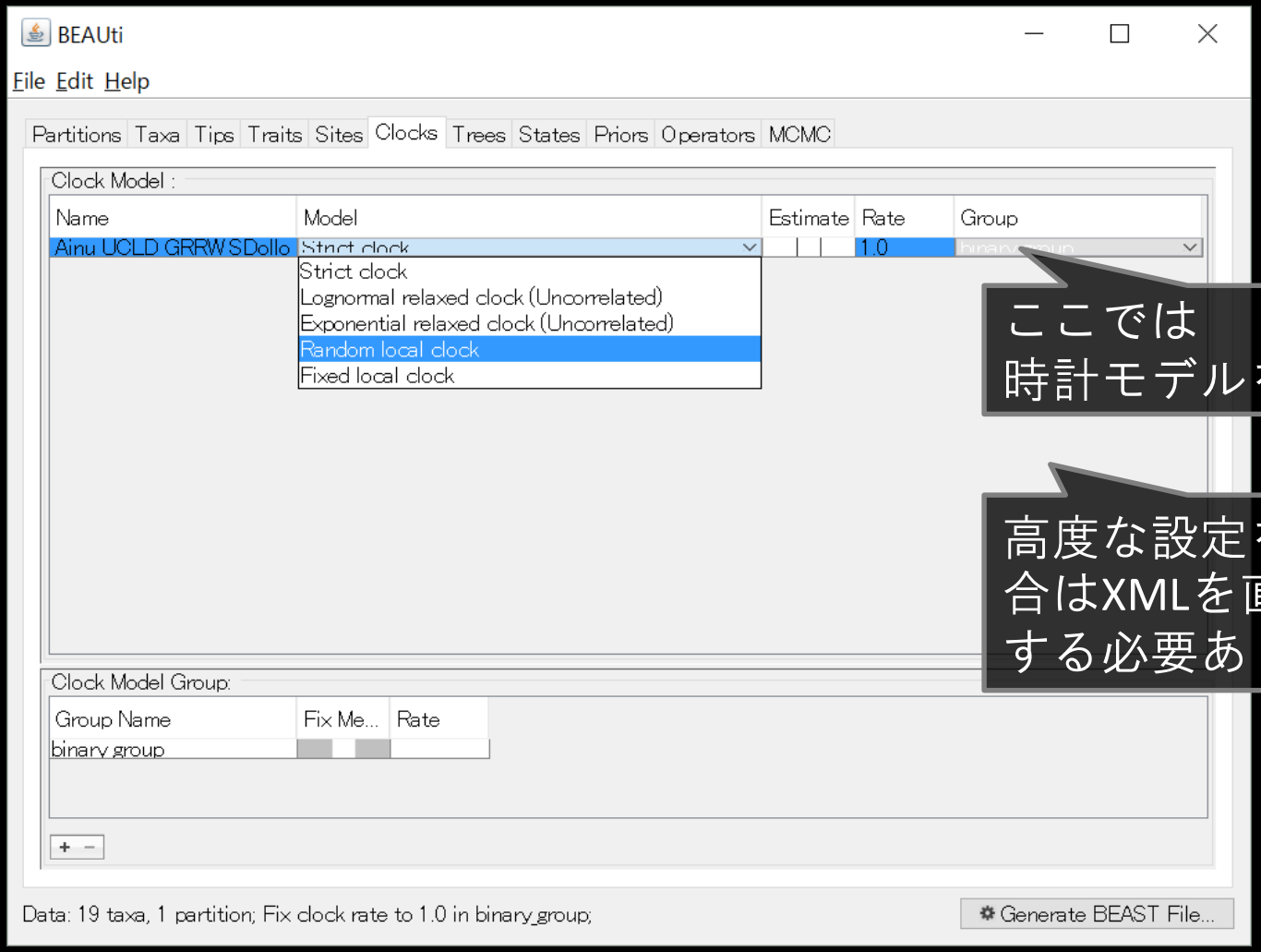
Data Source:
Lee & Hasegawa (2013)

BEASTの設定ファイル (XML)

```
        <!-- Generate a tree model                                    -->
        <treeModel id="treeModel">
                <coalescentTree idref="startingTree"/>
                <rootHeight>
                        <parameter id="treeModel.rootHeight"/>
                </rootHeight>
                <nodeHeights internalNodes="true">
                        <parameter id="treeModel.internalNodeHeights"/>
                </nodeHeights>
                <nodeHeights internalNodes="true" rootNode="true">
                        <parameter id="treeModel.allInternalNodeHeights"/>
                </nodeHeights>
    <nodeTraits name="trait" rootNode="true" internalNodes="false" leafNodes="false" traitDimensio
n="2">
                        <parameter id="rootTrait"/>
                </nodeTraits>
                <nodeTraits name="trait" rootNode="true" internalNodes="true" leafNod
es="false" traitDimension="2">
                        <parameter id="internalTraits"/>
                </nodeTraits>
                <nodeTraits name="trait" rootNode="true" internalNodes="true" leafNod
es="true" traitDimension="2">
                        <parameter id="allTraits"/>
                </nodeTraits>
        </treeModel>
```

Data Source:
Lee & Hasegawa (2013)

BEASTのモデル設定はGUIでできる



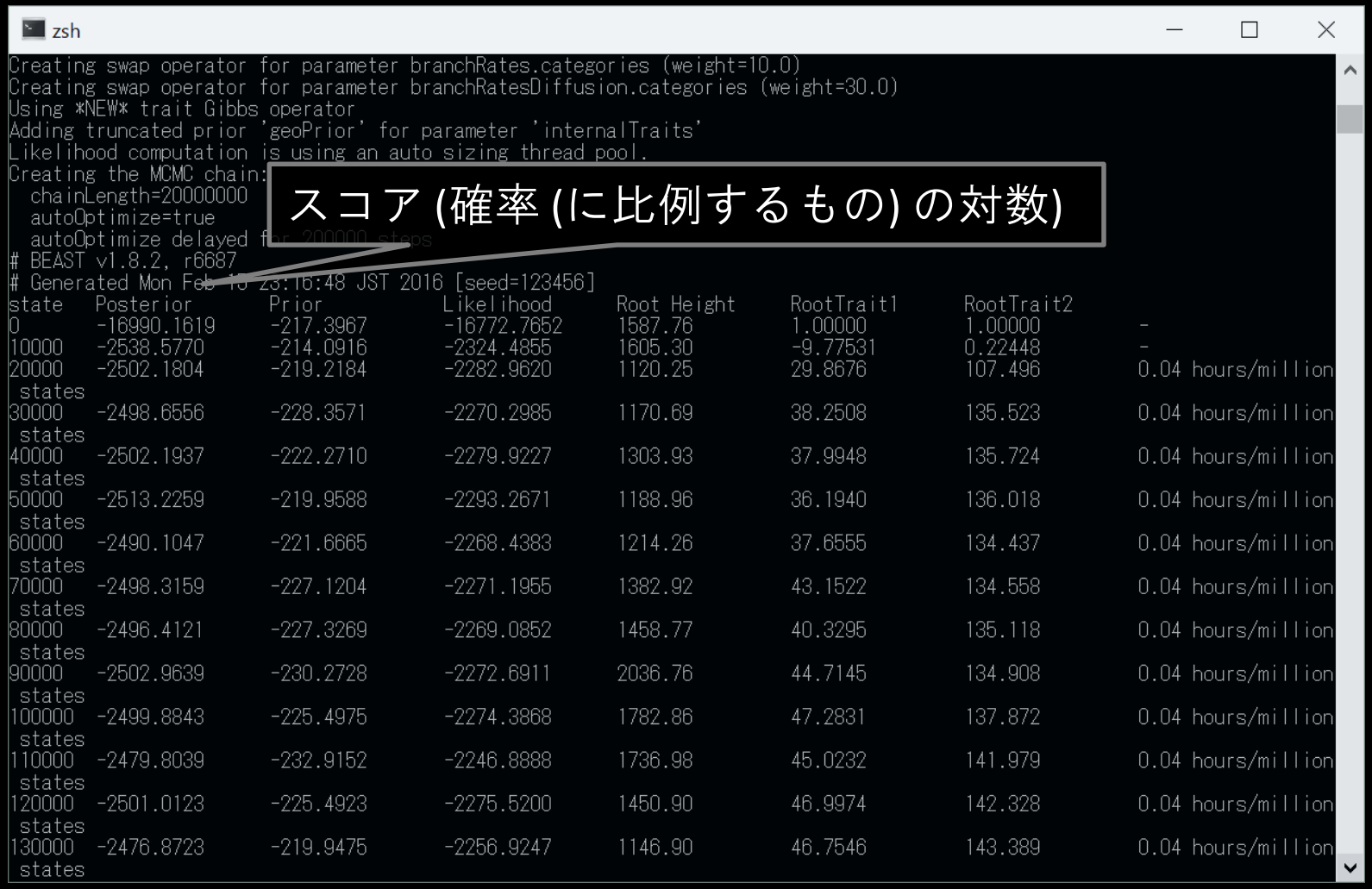
ここでは
時計モデルを選択中

高度な設定を行う場合はXMLを直接編集する必要あり

BEASTの実行: 起動

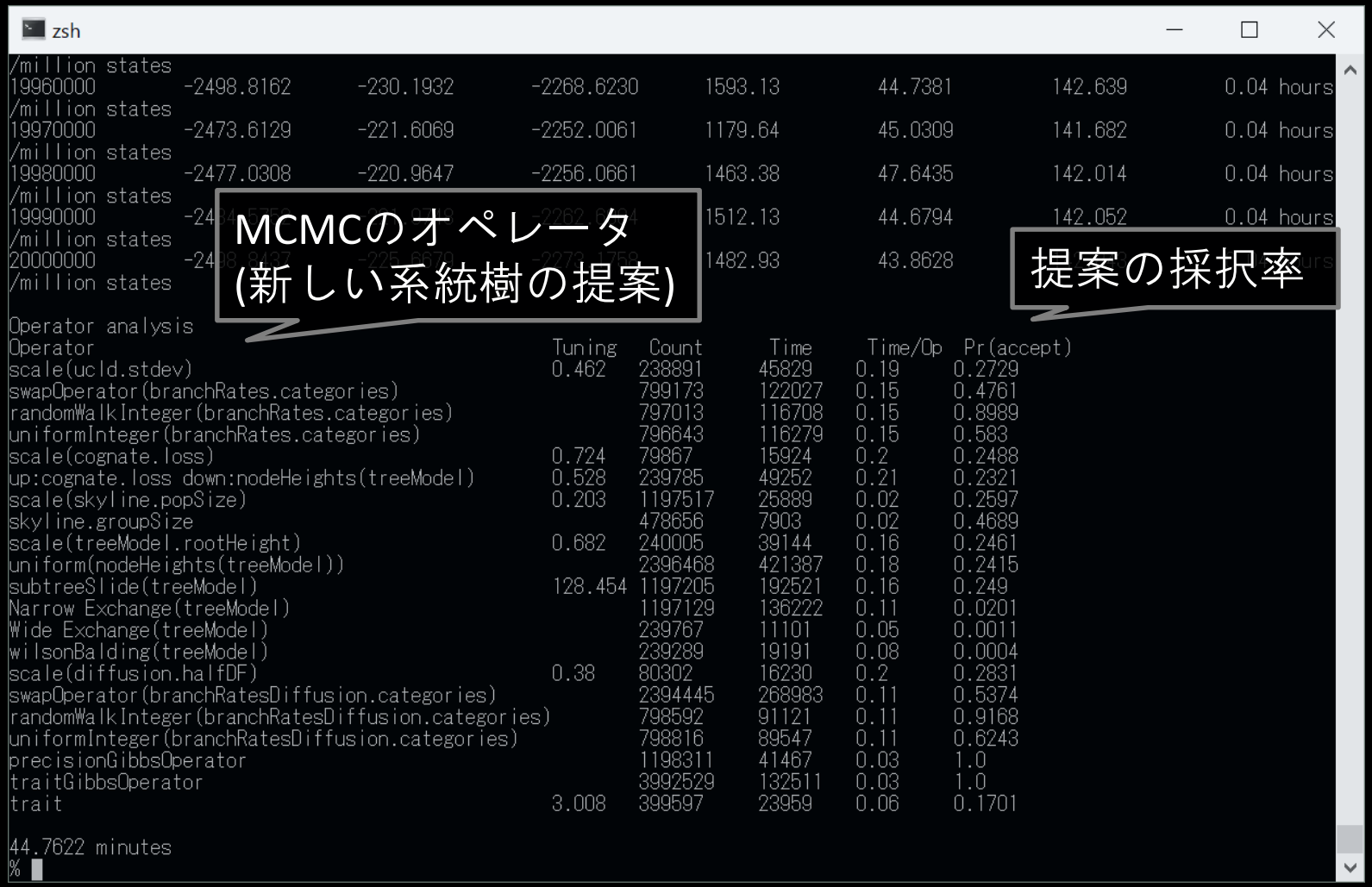
```
zsh
% java -jar /home/murawaki/local/BEAST¥ v1.8.2/lib/beast.jar -seed 123456 -overwrite -beagle_off Ainu_UCLD_GRRW_SDollo.xml


                  BEAST v1.8.2, 2002-2015
       Bayesian Evolutionary Analysis Sampling Trees
                  Designed and developed by
   Alexei J. Drummond, Andrew Rambaut and Marc A. Suchard

              Department of Computer Science
                  University of Auckland
                 alexei@cs.auckland.ac.nz

            Institute of Evolutionary Biology
                 University of Edinburgh
                    a.rambaut@ed.ac.uk

             David Geffen School of Medicine
          University of California, Los Angeles
                    msuchard@ucla.edu

               Downloads, Help & Resources:
                       http://beast.bio.ed.ac.uk

Source code distributed under the GNU Lesser General Public License:
               http://code.google.com/p/beast-mcmc

                     BEAST developers:
       Alex Alekseyenko, Guy Baele, Trevor Bedford, Filip Bielejec, Erik Bloomquist, Matthew Hall,
       Joseph Heled, Sebastian Hoehna, Denise Kuehnert, Philippe Lemey, Wai Lok Sibon Li,
       Gerton Lunter, Sidney Markowitz, Vladimir Minin, Michael Defoin Platel,
               Oliver Pybus, Chieh-Hsi Wu, Walter Xie

                        Thanks to:
       Roald Forsberg, Beth Shapiro and Korbinian Strimmer


Random number seed: 123456
```

BEASTの実行: MCMCの途中経過




```
Creating swap operator for parameter branchRates.categories (weight=10.0)
Creating swap operator for parameter branchRatesDiffusion.categories (weight=30.0)
Using *NEW* trait Gibbs operator
Adding truncated prior 'geoPrior' for parameter 'internalTraits'
Likelihood computation is using an auto sizing thread pool.
Creating the MCMC chain:
  chainLength=20000000
  autoOptimize=true
  autoOptimize delayed for 200000 steps
# BEAST v1.8.2, r6687
# Generated Mon Feb 15 23:16:48 JST 2016 [seed=123456]
state   Posterior       Prior          Likelihood     Root Height     RootTrait1      RootTrait2
0       -16990.1619     -217.3967      -16772.7652    1587.76         1.00000         1.00000         -
10000   -2538.5770      -214.0916      -2324.4855     1605.30         -9.77531        0.22448         -
20000   -2502.1804      -219.2184      -2282.9620     1120.25         29.8676         107.496         0.04 hours/million
 states
30000   -2498.6556      -228.3571      -2270.2985     1170.69         38.2508         135.523         0.04 hours/million
 states
40000   -2502.1937      -222.2710      -2279.9227     1303.93         37.9948         135.724         0.04 hours/million
 states
50000   -2513.2259      -219.9588      -2293.2671     1188.96         36.1940         136.018         0.04 hours/million
 states
60000   -2490.1047      -221.6665      -2268.4383     1214.26         37.6555         134.437         0.04 hours/million
 states
70000   -2498.3159      -227.1204      -2271.1955     1382.92         43.1522         134.558         0.04 hours/million
 states
80000   -2496.4121      -227.3269      -2269.0852     1458.77         40.3295         135.118         0.04 hours/million
 states
90000   -2502.9639      -230.2728      -2272.6911     2036.76         44.7145         134.908         0.04 hours/million
 states
100000  -2499.8843      -225.4975      -2274.3868     1782.86         47.2831         137.872         0.04 hours/million
 states
110000  -2479.8039      -232.9152      -2246.8888     1736.98         45.0232         141.979         0.04 hours/million
 states
120000  -2501.0123      -225.4923      -2275.5200     1450.90         46.9974         142.328         0.04 hours/million
 states
130000  -2476.8723      -219.9475      -2256.9247     1146.90         46.7546         143.389         0.04 hours/million
 states
```

BEASTの実行: 終了


zsh
/million states
19960000 -2498.8162 -230.1932 -2268.6230 1593.13 44.7381 142.639 0.04 hours
/million states
19970000 -2473.6129 -221.6069 -2252.0061 1179.64 45.0309 141.682 0.04 hours
/million states
19980000 -2477.0308 -220.9647 -2256.0661 1463.38 47.6435 142.014 0.04 hours
/million states
19990000 1512.13 44.6794 142.052 0.04 hours
/million states
20000000 1482.93 43.8628
/million states
MCMCのオペレータ
(新しい系統樹の提案)
提案の採択率
Operator analysis
Operator Tuning Count Time Time/Op Pr(accept)
scale(ucld.stdev) 0.462 238891 45829 0.19 0.2729
swapOperator(branchRates.categories) 799173 122027 0.15 0.4761
randomWalkInteger(branchRates.categories) 797013 116708 0.15 0.8989
uniformInteger(branchRates.categories) 796643 116279 0.15 0.583
scale(cognate.loss) 0.724 79867 15924 0.2 0.2488
up:cognate.loss down:nodeHeights(treeModel) 0.528 239785 49252 0.21 0.2321
scale(skyline.popSize) 0.203 1197517 25889 0.02 0.2597
skyline.groupSize 478656 7903 0.02 0.4689
scale(treeModel.rootHeight) 0.682 240005 39144 0.16 0.2461
uniform(nodeHeights(treeModel)) 2396468 421387 0.18 0.2415
subtreeSlide(treeModel) 128.454 1197205 192521 0.16 0.249
Narrow Exchange(treeModel) 1197129 136222 0.11 0.0201
Wide Exchange(treeModel) 239767 11101 0.05 0.0011
wilsonBalding(treeModel) 239289 19191 0.08 0.0004
scale(diffusion.halfDF) 0.38 80302 16230 0.2 0.2831
swapOperator(branchRatesDiffusion.categories) 2394445 268983 0.11 0.5374
randomWalkInteger(branchRatesDiffusion.categories) 798592 91121 0.11 0.9168
uniformInteger(branchRatesDiffusion.categories) 798816 89547 0.11 0.6243
precisionGibbsOperator 1198311 41467 0.03 1.0
traitGibbsOperator 3992529 132511 0.03 1.0
trait 3.008 399597 23959 0.06 0.1701
44.7622 minutes
%

その他のツール

- Tracer: MCMCのログを解析
  - 収束の判定など
- TreeAnnotator: 複数のサンプルを1つの系統樹に要約
  - 要約手法は最大系統群信頼度木 (maximum clade credibility tree) など
- FigTree: 系統樹を描画

今日のおはなし

- 音法則からBayes系統モデルまで
- Bayes系統モデルのソフトウェア
- 言語資源
- 発展的な話題

言語資源

- 統計的研究には計算機可読な言語資源 (データベース) が不可欠
- 言語資源の作成は超高コスト
  - ある言語を追加するには、その言語の専門知識が不可欠
- 近年、言語資源を組織的に作成して共有する例が増えている
  - 特にMax Planck Institute for Evolutionary Anthropology / the Science of Human History / Psycholinguistics
- 言語資源を作れなくても研究に参入できる!

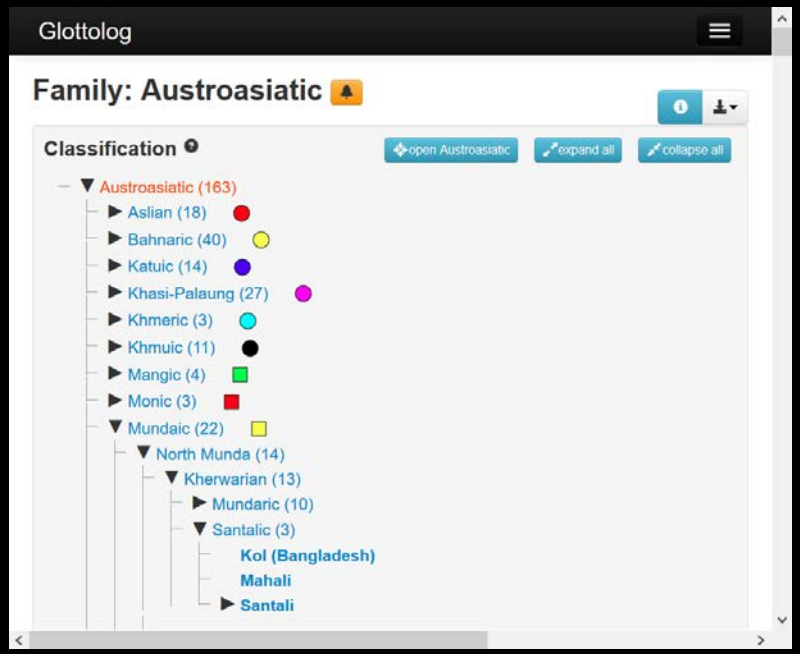
語彙 (同源語) データベース

- IELex: インド・ヨーロッパ語族
  - Isidore Dyenの語彙統計学の遺産がベース
- Austronesian Basic Vocabulary Database
- Bantu Basic Vocabulary Database
- Trans-New Guinea
- (論文の補助資料)
- Automated Similarity Judgment Program

言語学者による (年代なし) 系統樹

系統推定の定量評価に使える

- Glottolog (おすすめ)
  - 開発が盛ん
  - Newickフォーマットで系統樹を配布
- Ethnologue
  - 計算機可読ではない
- WALS (後述)
  - Family, Genusの2段階だけ



Source:
Screenshot:
http://glottolog.org/resource/languoid/id/aust1305

言語コード

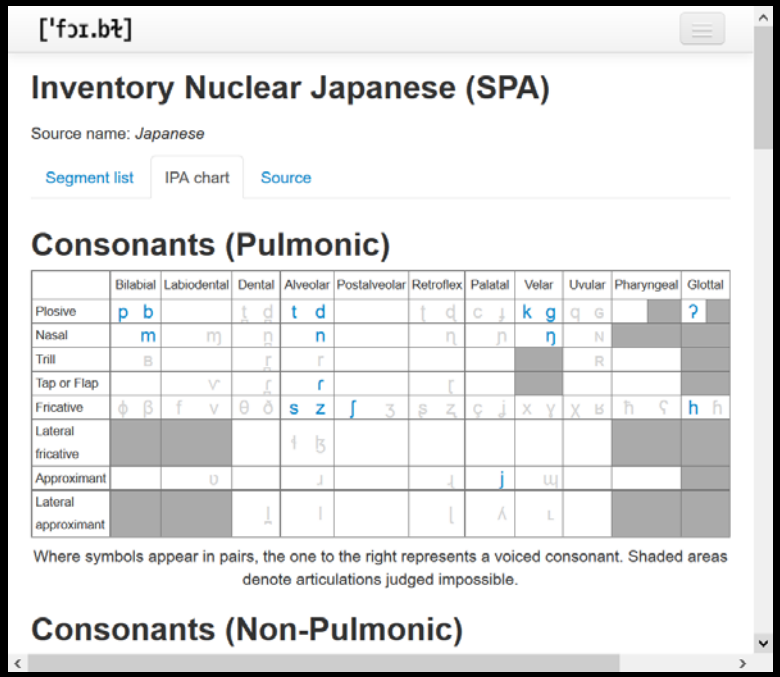
複数の資源を統合する際に、言語の対応付けのために必要

- Glottocode (おすすめ)
  - 8文字のコード (e.g. nucl1643)
  - Glottologで使用
  - 方言レベルでも割り当てられている
  - ISO 639-3へのマッピングあり
- ISO 639-3 language code
  - 3文字のコード (e.g. jpn)
  - Ethnologueに対応
- WALS code
  - 3文字のコード (e.g. jpn)
  - WALSで使用
  - ISO 639-3へのマッピングあり

音素目録 (phonological inventory) データベース

- PHOIBLE
  - 既存資源 (UPSID, SPA, etc) の統合と独自追加

- そもそも音素を通言語的に一貫性をもって比較するのは難しい
  - 同じ日本語でも、認定された音素数は UPSIDで20個、SPAで40個



['fɔɪ.bɬ]

Inventory Nuclear Japanese (SPA)

Source name: *Japanese*

Segment list | IPA chart | Source

Consonants (Pulmonic)

	Bilabial	Labiodental	Dental	Alveolar	Postalveolar	Retroflex	Palatal	Velar	Uvular	Pharyngeal	Glottal
Plosive	p b			t d				k g			ʔ
Nasal	m			n				ŋ			
Trill											
Tap or Flap				ɾ							
Fricative	ɸ β	f v	θ ð	s z	ʃ ʒ	ʂ ʐ	ç ʝ	x ɣ	χ ʁ	ħ ʕ	h ɦ
Lateral fricative				ɬ ɮ							
Approximant		ʋ		ɹ		ɻ	j	ɰ			
Lateral approximant				l		ɭ	ʎ	ʟ			

Where symbols appear in pairs, the one to the right represents a voiced consonant. Shaded areas denote articulations judged impossible.

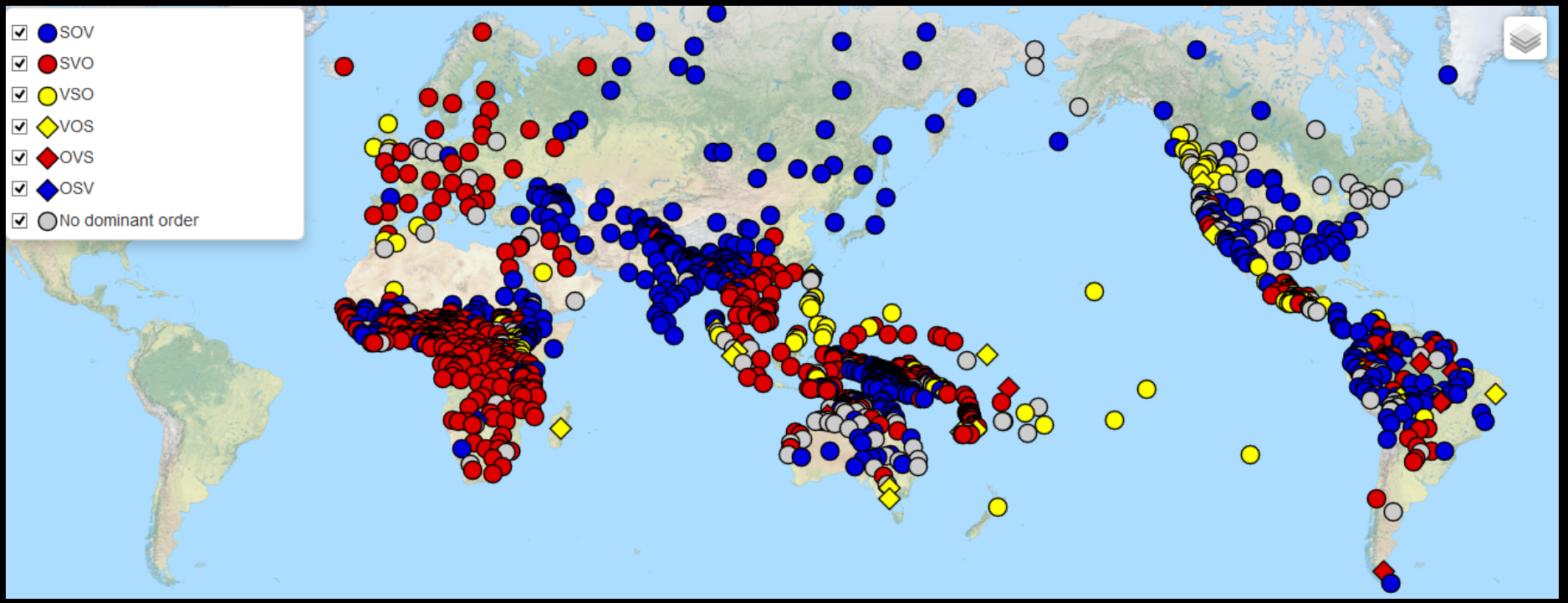
Consonants (Non-Pulmonic)

Source:
Screenshot:
http://phoible.org/inventories/view/197#tipa

言語類型論 (Linguistic Typology)

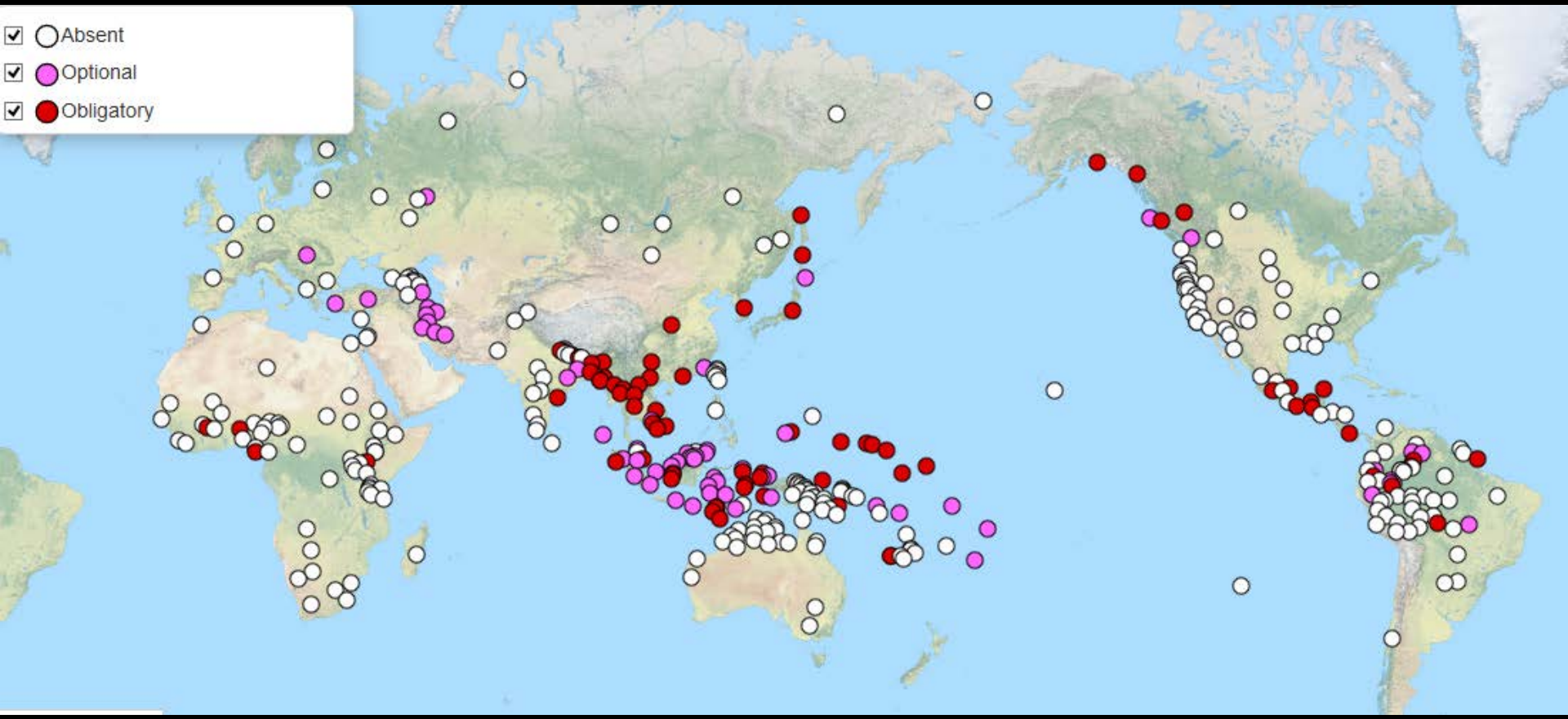
- 世界の言語を類型によって分類
  - 語順、助数詞の有無、声調の有無, etc
- World Atlas of Language Structures (WALS)
  - 2,679言語
  - 192種類の特徴量

Feature 81A:
Order of Subject, Object and Verb



Source:
Screenshot:
http://wals.info/feature/81A

Feature 55A: Numeral Classifiers (助数詞を使うか)



Source:
Screenshot:
http://wals.info/feature/55A

その他の言語資源

- Atlas of Pidgin and Creole Language Structures (APiCS)
  - 類型論、音素目録、社会言語学的特徴量
- World Loan Word Database (WOLD)
- AfBo: 接辞の借用
- Concepticon: 基礎語彙リストのリスト

今日のおはなし

- 音法則からBayes系統モデルまで
- Bayes系統モデルのソフトウェア
- 言語資源
- 発展的な話題

発展的な話題

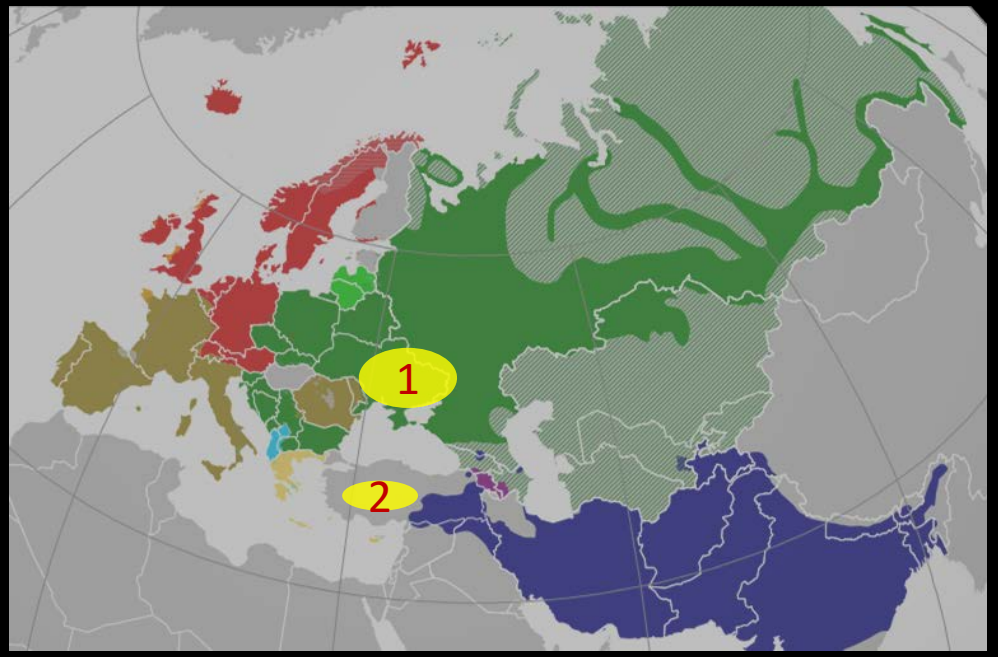
- 印欧祖語の年代論争の続報
- 言語接触の影響
- 方言同士の関係
- 日本語の起源と類型論

発展的な話題

- 印欧祖語の年代論争の続報
- 言語接触の影響
- 方言同士の関係
- 日本語の起源と類型論

印欧祖語の年代と故地 (再掲)

1. クルガン仮説
   - 5,000-6,000年前
   - 黒海周辺のステップ
   - 遊牧民の軍事的征服
2. アナトリア仮説
   - 8,000-9,500年前
   - アナトリア
   - 農耕とともに拡大



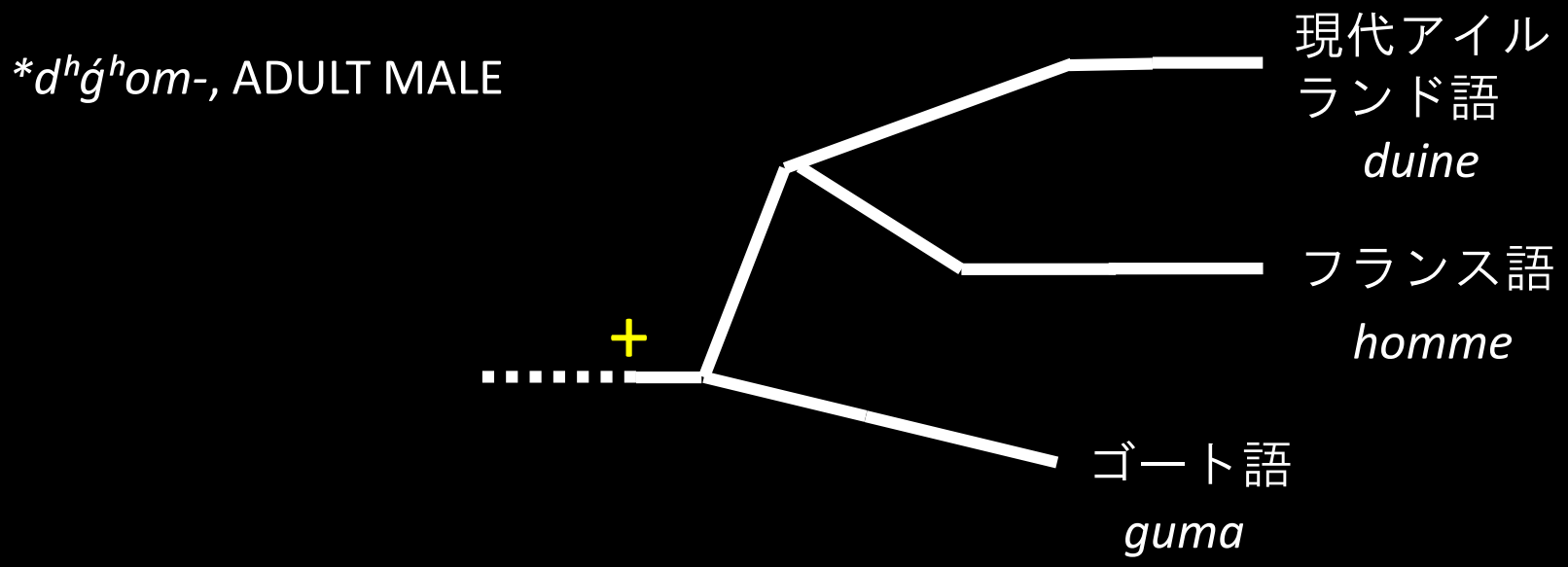
- Bouckaert+ (2012) が支持するアナトリア仮説は言語学者の間では評判が悪い
- もしクルガン仮説が正しいとすると、Bayes系統モデルの何がまずかったのか?

Source:
Map:
- Author: Alphathon
- License: CC BY-SA 3.0

意味変化によるhomoplasy 1/2

[Chang+, 2015]

- homoplasyが無視できないほど頻出
  - IELEXのロマンス諸語の基礎語彙の8.1%
- 同じ意味変化が独立に起きている



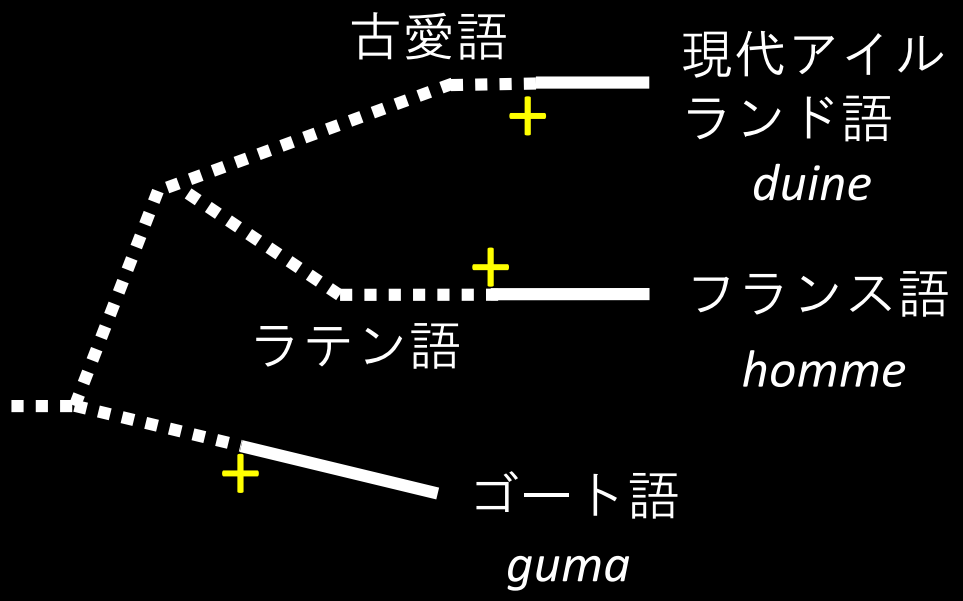
意味変化によるhomoplasy 2/2

[Chang+, 2015]

- 提案手法: 古代語を制約として使う
- 結果: 印欧祖語の年代は6,500年前に繰り上がり、ステップ説に近づいた

*$d^hǵ^hom$-, ADULT MALE

PERSON → ADULT MALE
の意味変化が独立に発生
- ラテン語: *homo*, PERSON
- 古愛語: *duine*, PERSON



発展的な話題

- 印欧祖語の年代論争の続報
- 言語接触の影響
- 方言同士の関係
- 日本語の起源と類型論

系統樹は理想化にすぎない

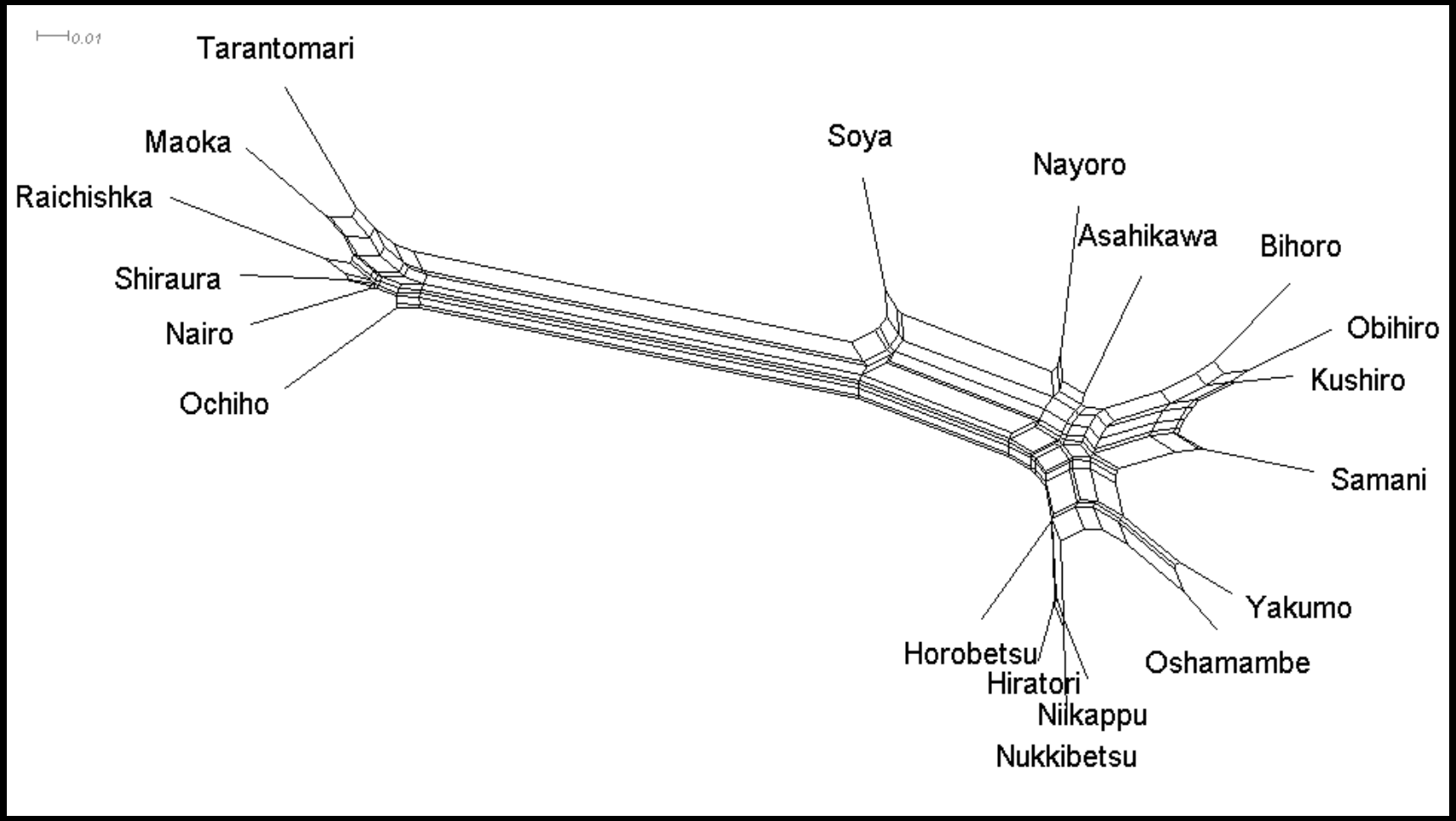
- 言語学では、木モデルに従わない、接触に基づく現象が昔から研究されてきた
- 系統樹が縦の (vertical) 伝達だとすると、接触は横の (horizontal) 伝達
- 文化人類学におけるphylogenesis (縦) vs. ethnogenesis (横) 論争とも類似

接触に基づく現象の例

- 語彙・文法の借用
- 方言 (非常に近い言語) 群の相互作用
- 地域言語学 (areal linguistics)
  - e.g. バルカン言語連合
- ピジン・クレオール

NeighborNetによる分析 1/2

[Bryant+, 2004]



Data Source:
Lee & Hasegawa (2013)

NeighborNetによる分析 2/2

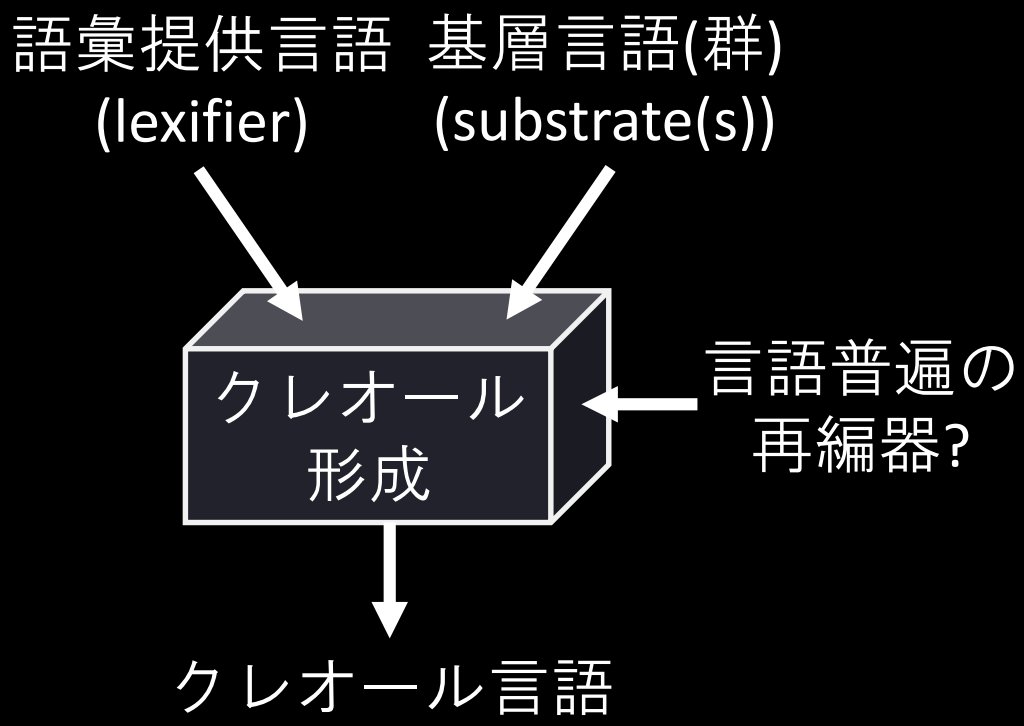
[Bryant+, 2004]

- 距離ベースのボトムアップ・クラスタリング
  - 無根木 (unrooted tree)
- 複数の木を統合し、矛盾する情報を菱型で可視化

- 実装としてSplitsTreeがよく使われる

クレオール形成の混合モデルによるモデル化

3/10 (木) D-5 言語学・言語分析(2)
10:00-10:20 で発表予定



- 分岐を繰り返す系統樹とは反対に、言語が複数のソースを持つ
- 混合モデルが向いている
  - LDAに似たモデル
  - 分子生物学のBayesモデル (Structure) により似ている

[Murawaki, 2016]
(to appear)

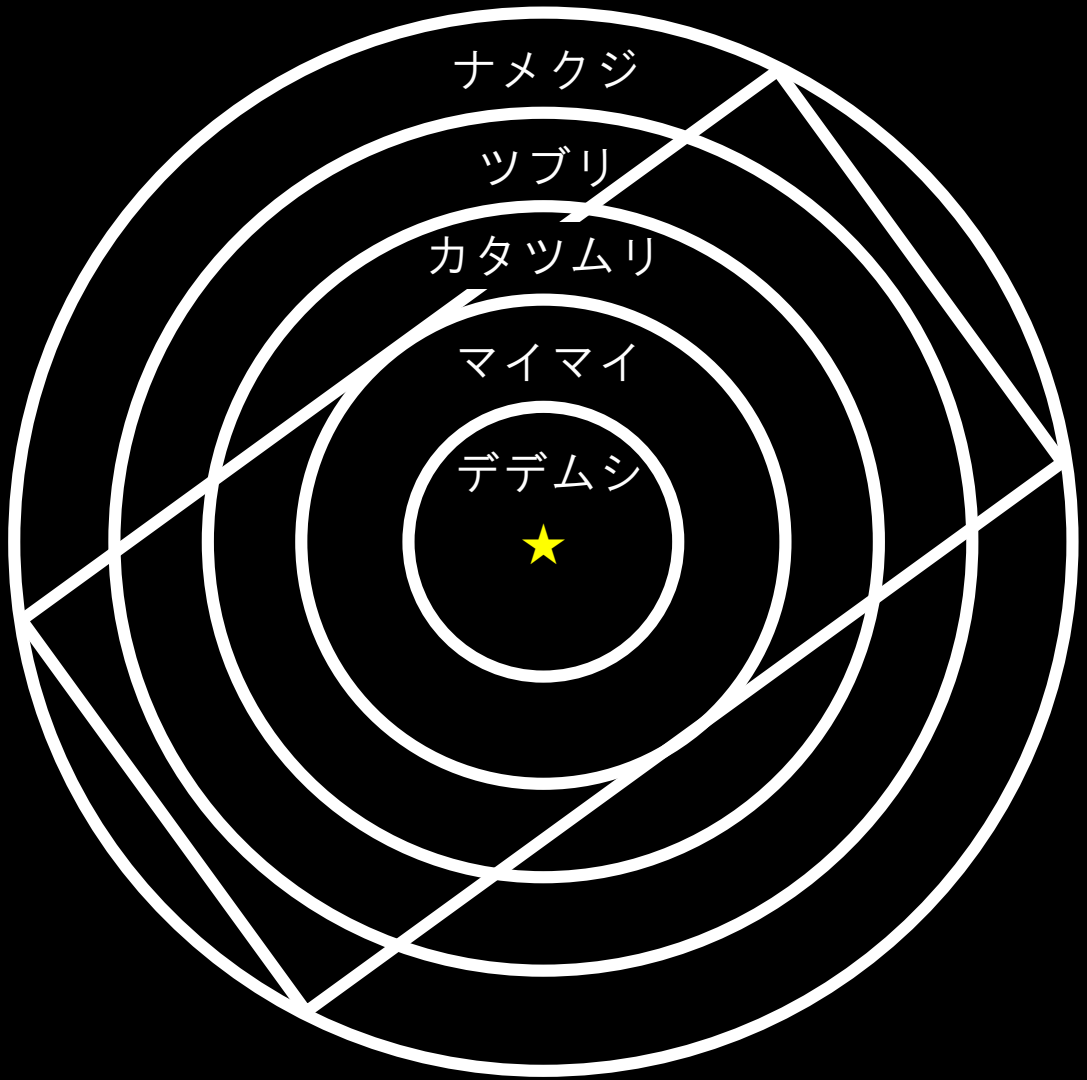
発展的な話題

- 印欧祖語の年代論争の続報
- 言語接触の影響
- 方言同士の関係
- 日本語の起源と類型論

方言同士の関係

- 恒常的な接触の影響により、系統モデルは適さないと思われる
- 伝統的な方言区画論も、現代語の特徴に基づくクラスタリングであり、歴史的変化を表す系統樹という観念は希薄
- 拡散 (diffusion) の (非統計的) モデル
  - 引力モデル (gravity model) [Trudgill, 1974]
  - 方言周圏論 [柳田, 1930]
    - シミュレーションモデル [Lizana+, 2011]

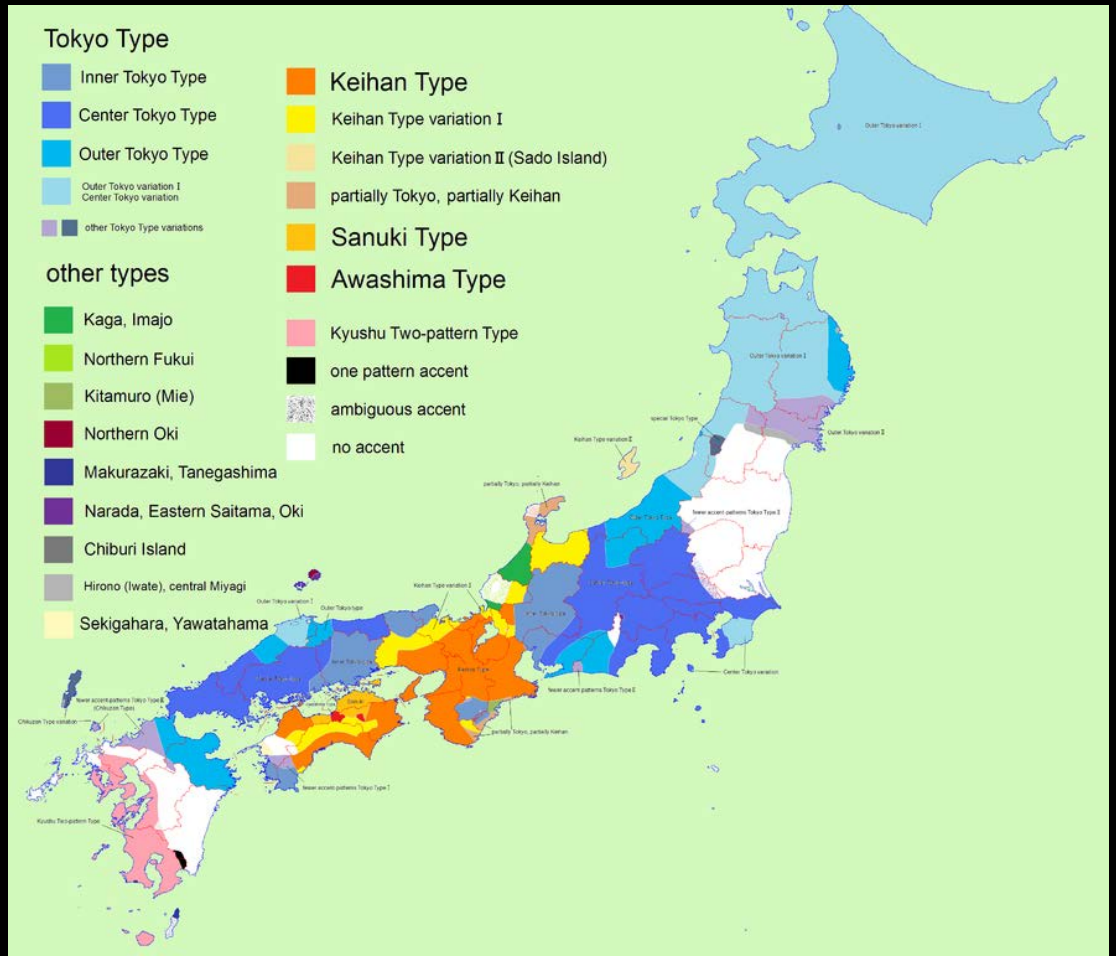
方言周圏論 [柳田, 1930]



- 中央で生まれた語が周辺に伝播
- 結果として古語は周縁に残存

- 定量的分析?

アクセント体系の系統樹

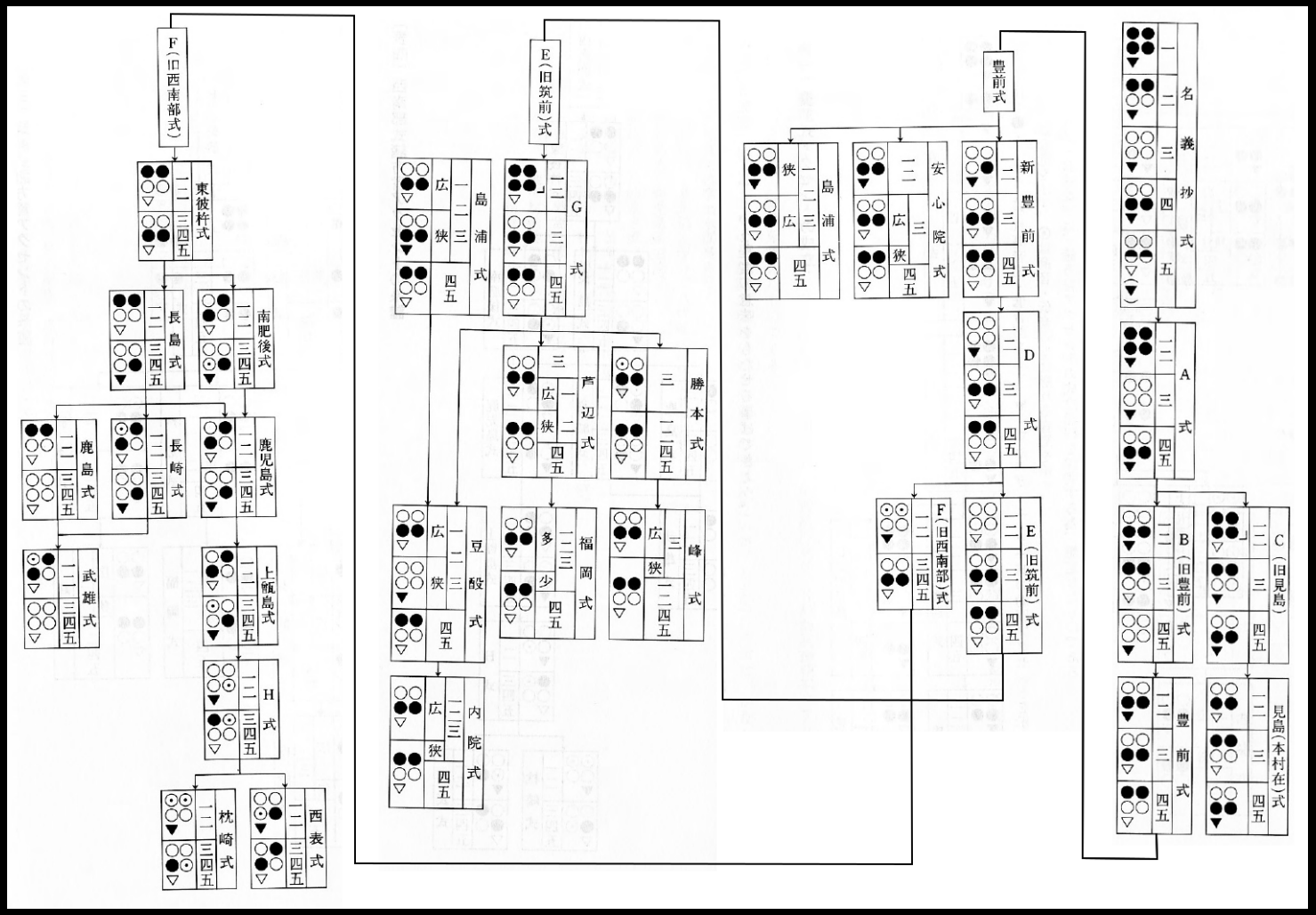


- アクセント体系は地域差が非常に大きい
- 体系なので、語彙と違って借用に強い
- 言語学者が系統樹を作った例はあるが、統計モデルはまだ

Source:
Map:
- Author: Henlly
- License: CC BY-SA 3.0

アクセント体系の系統樹

[奥村, 1990]



Source:
奥村三雄. 1990. 九州諸方言アクセントの系譜. 方言国語史研究.
pp.215-218の4図を合成

発展的な話題

- 印欧祖語の年代論争の続報
- 言語接触の影響
- 方言同士の関係
- 日本語の起源と類型論

日本語の起源、同系言語は?

- 朝鮮語 [Aston, 1879][金澤, 1910][Martin, 1966]
- アルタイ語族 [Miller, 1971]
- ノストラ語族 [Starostin, 1989]
- ユーラシア語族 [Greenberg, 2000]
- オーストロネシア語族 [川本, 1980][Benedict, 1990]
- タミル語 (ドラヴィダ語族) [大野, 1980]
- レプチャ語 [安田, 1955]
- 高句麗地名 [新村, 1916]

代表的な文献
必ずしも初出ではない

語彙に基づく手法は成功していない

- >100年の研究にもかかわらず、日本語と他の言語との間で信頼できる同源語群が確立できていない [Vovin, 2010]
  - 仮に同系言語が見つかったとしても、祖語の年代は相当さかのぼりそう [服部, 1999[1956]]
- 同源語群がなければ、上述のBayes統計モデルは適用しようがない

類型論に基づく系統推定

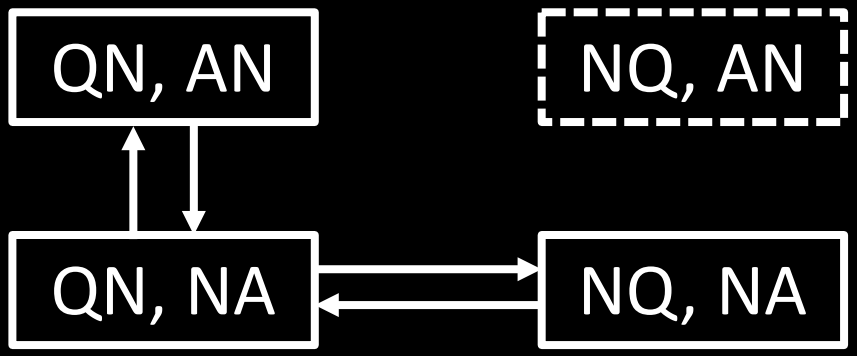
肯定的な結果 [Dunn+, 2005][Longobardi+, 2009] とやや否定的な結果 [Greenhill+, 2010][Dunn+, 2011] が混在

- Pros
  - 任意の言語対を比較できる
  - 語彙よりも歴史的に安定した特徴がありそう [Nichols, 1992][松本, 2007]
- Cons
  - homoplasyだらけ
    - SVO語順は歴史上何度も誕生している
  - back mutationもあり得る
  - 接触による変化 (areal linguistics) も知られている
  - 各特徴の変化の予測可能性が未知数

類型論の特徴間の依存関係

[Greenberg, 1978]

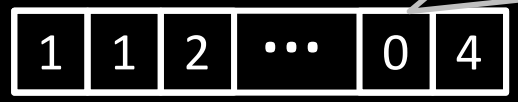
QN: 数詞 + 名詞 語順
AN: 形容詞 + 名詞 語順
NQ, NAは逆の語順



- 特定の特徴の組み合わせを持つ言語がない/非常に少ない
- 特徴が独立に変化するのではなく、依存関係を持つことを利用すれば、変化の経路を絞り込めるのでは?

特徴の依存関係に基づく言語の自然さ判定 [Murawaki, 2015]

- $f(x;\theta) = d \in [0,1]$
  - x: 言語候補
  - d: xの自然さ

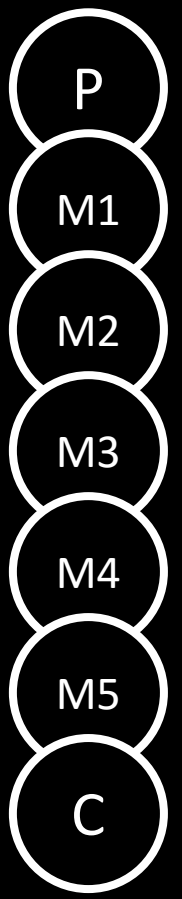


Feature 81A
Order of SOV
- 0: SOV
- 1: SVO
- 2: VSO
- ⋮

- 実在の言語のdを引き上げ、それ以外のxのdを引き下げるようにθを訓練する
  - 実在の言語によく現れる特徴の組み合わせに高いスコアを、そうでない組み合わせに低いスコアを与える

自然な変化の経路

- ある言語Pから別の言語Cへの変化を考える
- PとCは言語として自然 ($f(x;\theta)$が大)
- PとCの中間状態M1, M2, … も言語として自然であるはず
  - 中間状態も人間が話していたはずだから
- PからCへの経路が絞り込めるはず



まとめ

- 不確実性・連続値を含む問題には、計算機を用いた統計的手法が適している
- 近年は分子生物学由来の手法が言語に適用されてきた
- 言語資源の整備が進んでいる一方、適切な統計モデルが開発されていない現象がまだまだ残っている
- 一緒にこの分野で研究しましょう!

文献案内

- Nichols and Warnow. 2008. Tutorial on Computational Linguistic Phylogeny. Language and Linguistics Compass, 2(5).
  - 言語研究者向けの丁寧なチュートリアル
  - 少し古い
  - Bayes系統モデルの中身の説明はほとんどない
- Drummond and Bouckaert. 2015. Bayesian Evolutionary Analysis with BEAST.
  - BEAST作者によるモデルやプログラムの解説本
  - 言語の話はない
  - 上級者向け
- 村脇. 2016. 言語変化と系統への統計的アプローチ. 統計数理, 64(2). (*to appear*)
  - 今日の話とたいだい同じ内容 (になる予定)