

言語変化と系統 ~~言語類型論~~への 統計的アプローチ

九州大学
村脇 有吾

2015年9月4日国立国語研究所・統計数理研究所 合同研究集会
「統計的言語研究の現在」

自己紹介

村脇 有吾

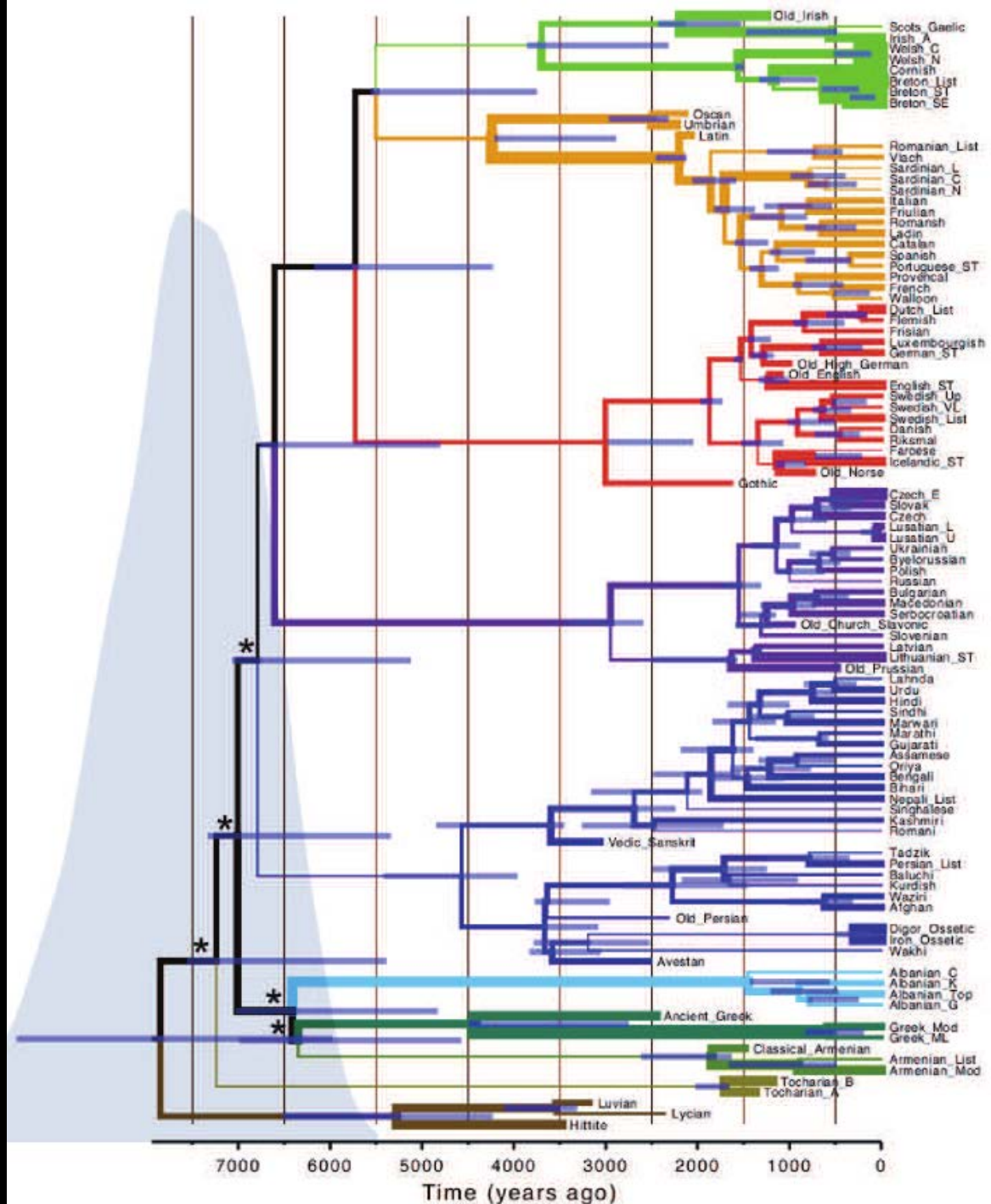
- 出身は京都大学黒橋研究室 (2011年博士)
- 九州大学助教 (2013年10月-)
- もともとは (今も) 日本語解析を研究
- 1年半前から言語の系統と変化の研究も

今日のお話

- 近年の言語系統の統計モデル
- 言語学と生物学との関わり
- 言語の性質を考慮したモデル化に向けて

Bayes統計による 系統推定

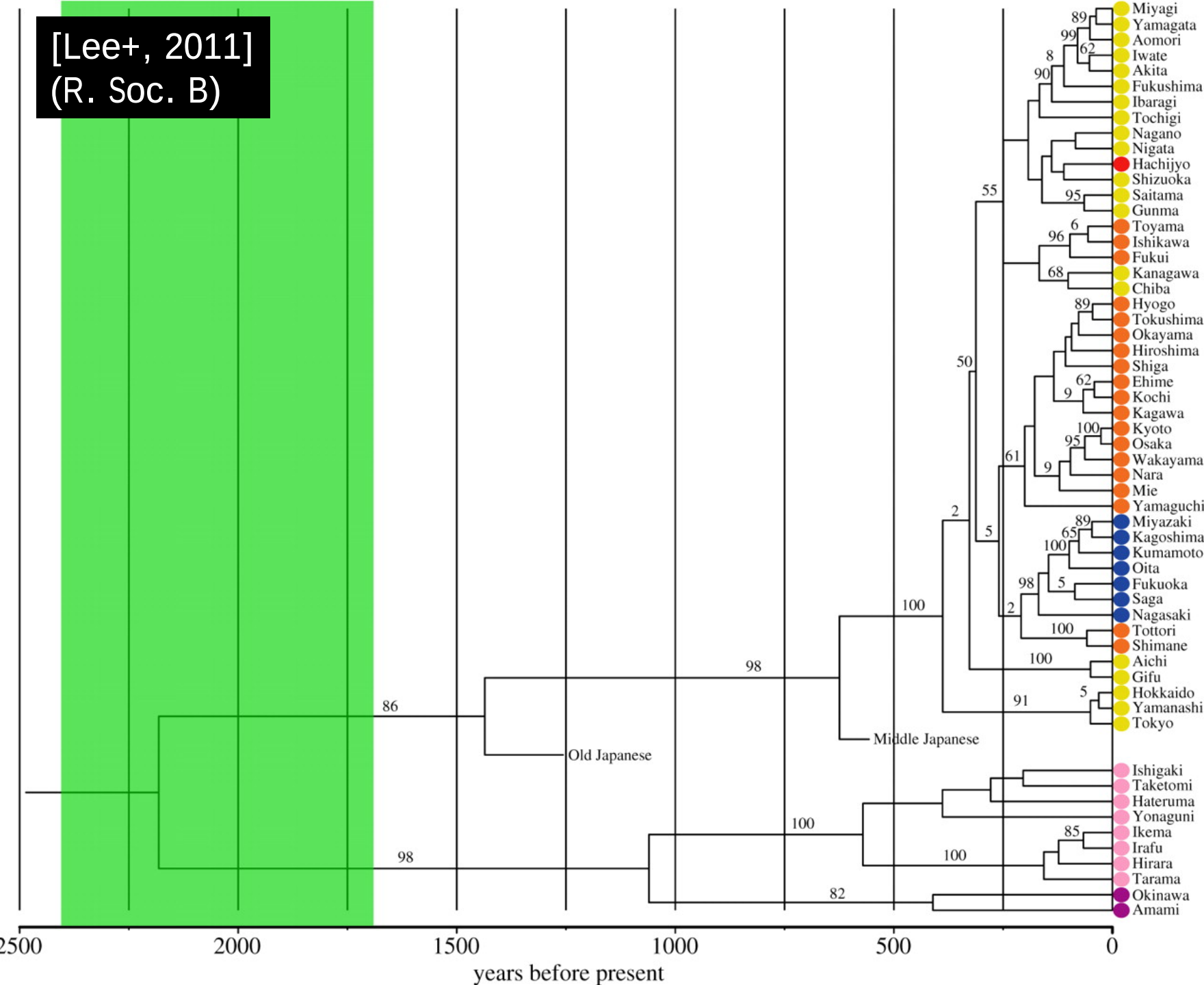
[Bouckaert+,
Science 2012]





[Bouckaert+,
Science 2012]

[Lee+, 2011]
(R. Soc. B)



もはや言語学者は不要?

もはや言語学者は不要?



Every time I fire a linguist,
the performance of the
speech recognizer goes up.

--Fred Jelinek

No. 言語学の成果に 全面的に依存している

- 基礎語彙データベース
 - [IELex](#): インド・ヨーロッパ語族
 - [Austronesian Basic Vocabulary Database](#)
 - [Automated Similarity Judgment Program](#) (ASJP)
 - 現代日本方言大辞典 [平山+, 1992-1994]
 - アイヌ語諸方言基礎語彙 [服部+, 1960]

No. 言語学の成果に 全面的に依存している

- 基礎語彙データベース
 - [IELex](#): インド・ヨーロッパ語族
 - [Austronesian Basic Vocabulary Database](#)
 - [Automated Similarity Judgment Program](#) (ASJP)
 - 現代日本方言大辞典 [平山+, 1992-1994]
 - アイヌ語諸方言基礎語彙 [服部+, 1960]
- 類型論データベース
 - [World Atlas of Language Structures](#)
- 音素体系データベース
 - [PHOIBLE](#), UPSID, etc

言語学関係者へのお願い

- 統一基準で作成された
- 中~大規模な言語データベースを
- (できれば機械可読な形で)
- 公開すれば、
- 思いもしなかったような
統計的応用が現れるかも

基礎語彙による 言語の2値ベクトル表現

	水	大きい
英語	water	big
ドイツ語	Wasser	gross
ロシア語	вода	большой великий
フランス語	eau	grand
イタリア語	acqua	grande

基礎語彙による 言語の2値ベクトル表現

同源語群 水

大きい

英語

water

big

ドイツ語

Wasser

gross

ロシア語

вода

большой

великий

フランス語

eau

grand

イタリア語

acqua

grande

基礎語彙による 言語の2値ベクトル表現

同源語群

水

大きい

英語	1	water	3	big	{1, 3}
ドイツ語		Wasser	4	gross	{1, 4}
ロシア語		вода	5	большой	{1, 5, 6}
			6	великий	
フランス語	2	eau	7	grand	{2, 7}
イタリア語		acqua		grande	{2, 7}

基礎語彙による 言語の2値ベクトル表現

同源語群

水

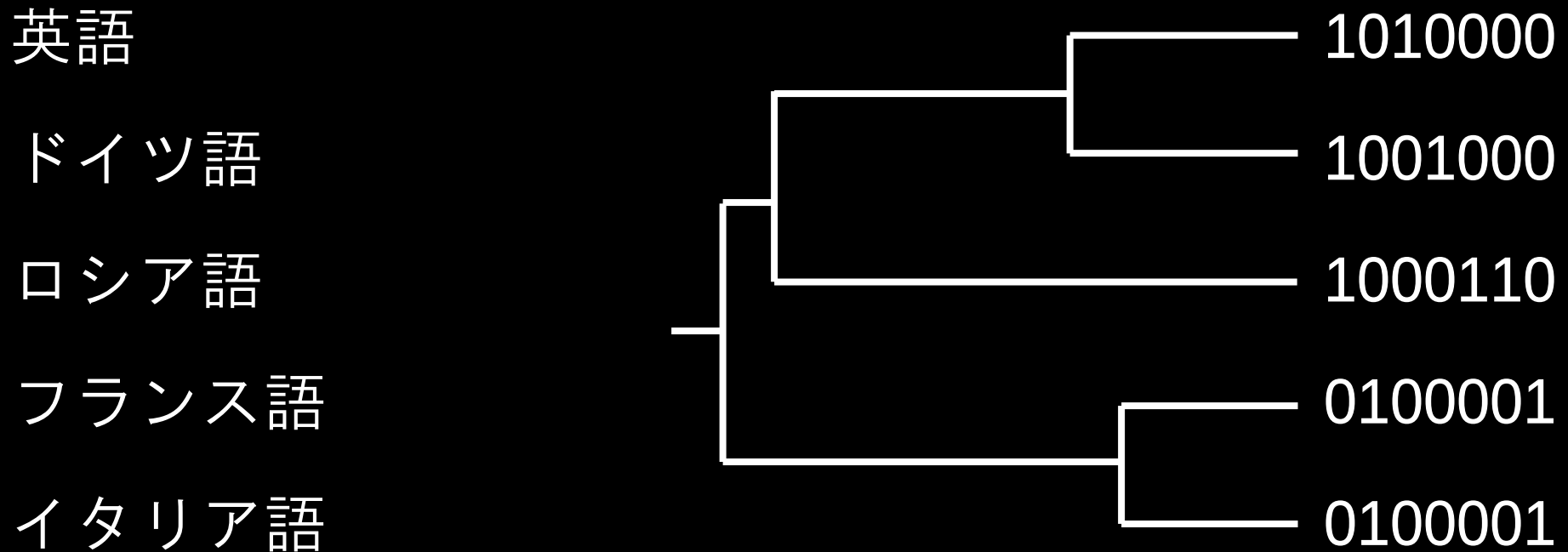
大きい

	1	3		
英語	water	big		1010000
ドイツ語	Wasser	gross		1001000
ロシア語	вода	большой	6	великий
フランス語	eau	grand		0100001
イタリア語	acqua	grande		0100001

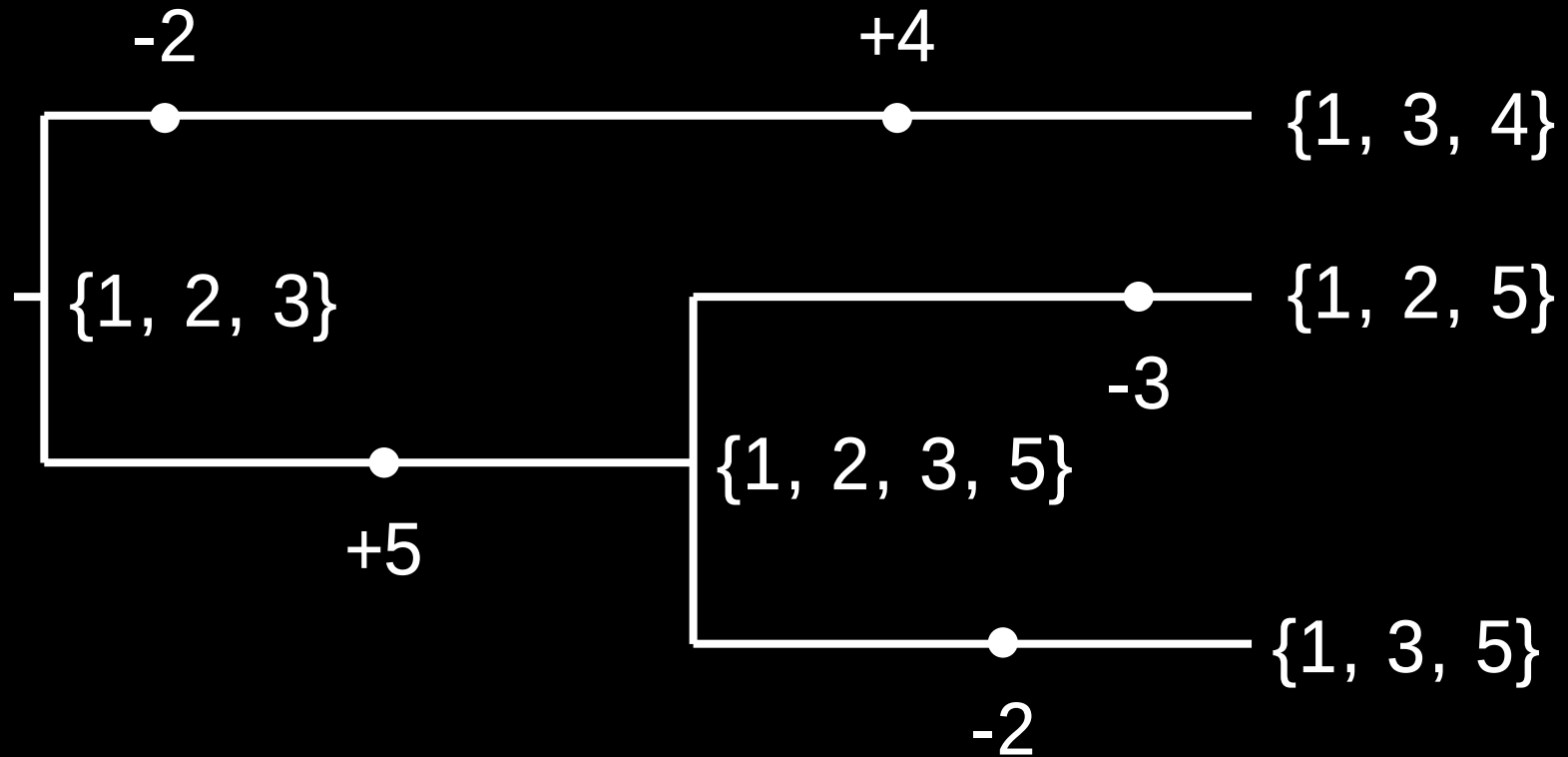
言語のクラスタリング = 系統樹?

英語	1010000
ドイツ語	1001000
ロシア語	1000110
フランス語	0100001
イタリア語	0100001

言語のクラスタリング = 系統樹?

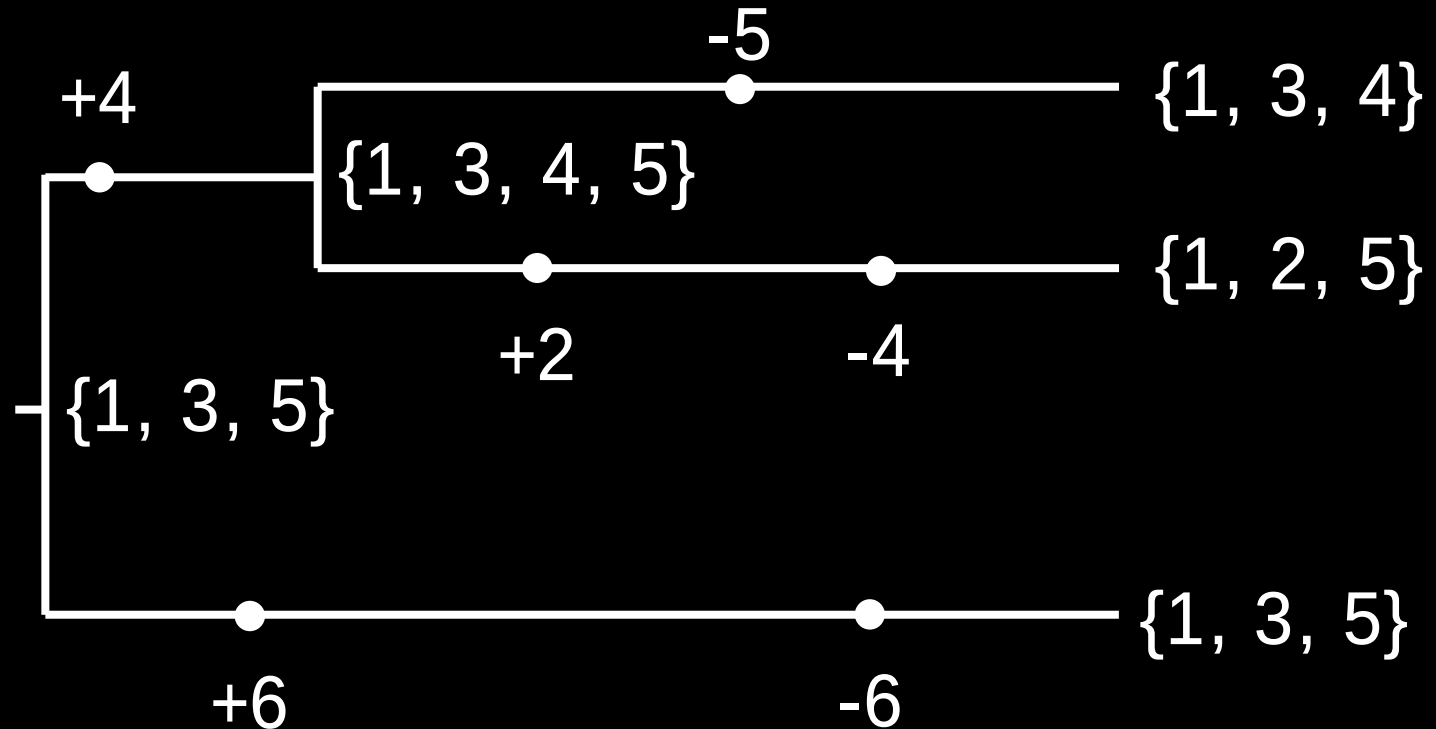


Bayes統計による系統推定



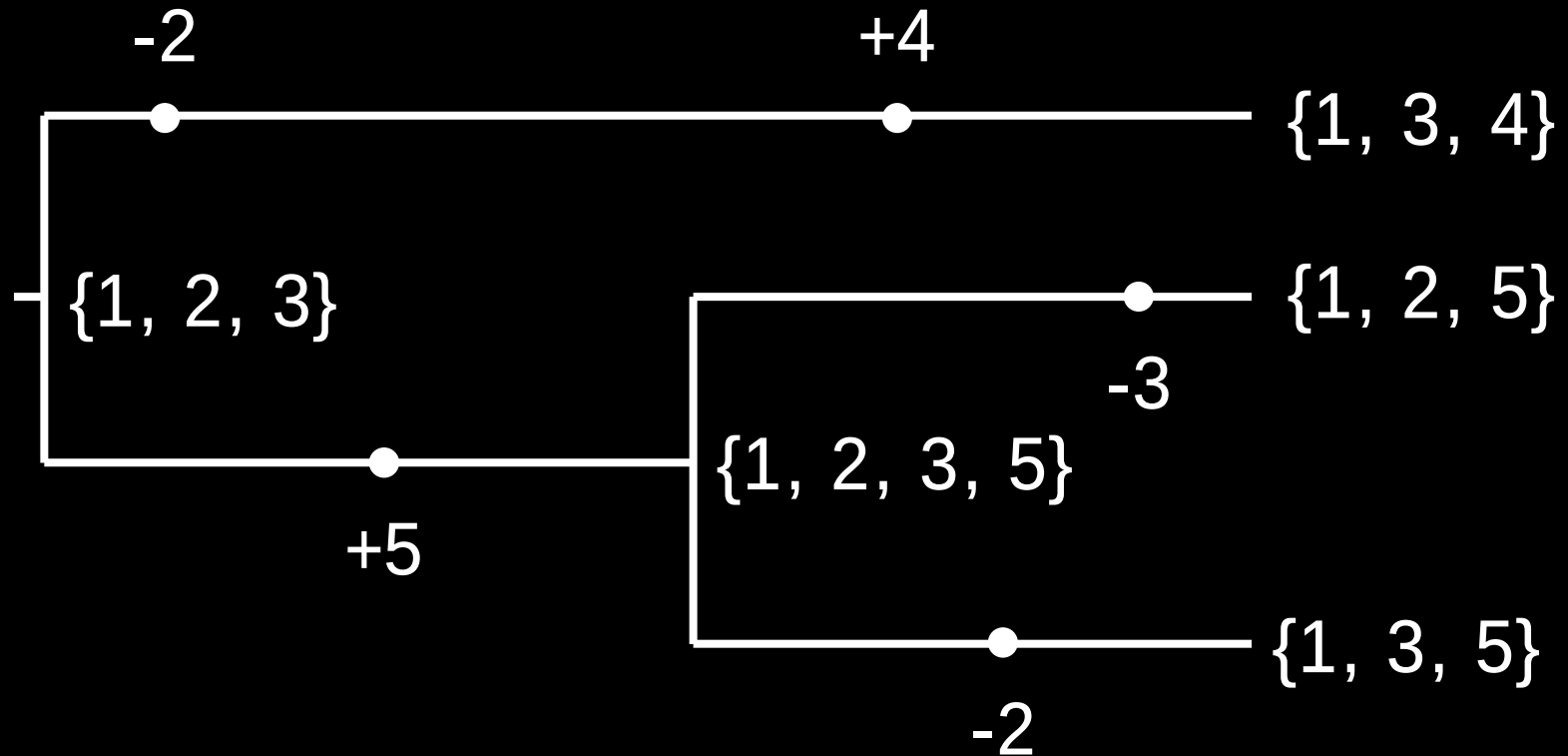
- 同源語の生死を直接モデル化
- 系統樹と同時に年代も推定

Bayes統計による系統推定



- 同源語の生死を直接モデル化
- 系統樹と同時に年代も推定

Bayes統計による系統推定



- 同源語の生死を直接モデル化
- 系統樹と同時に年代も推定

計算機の使いどころ

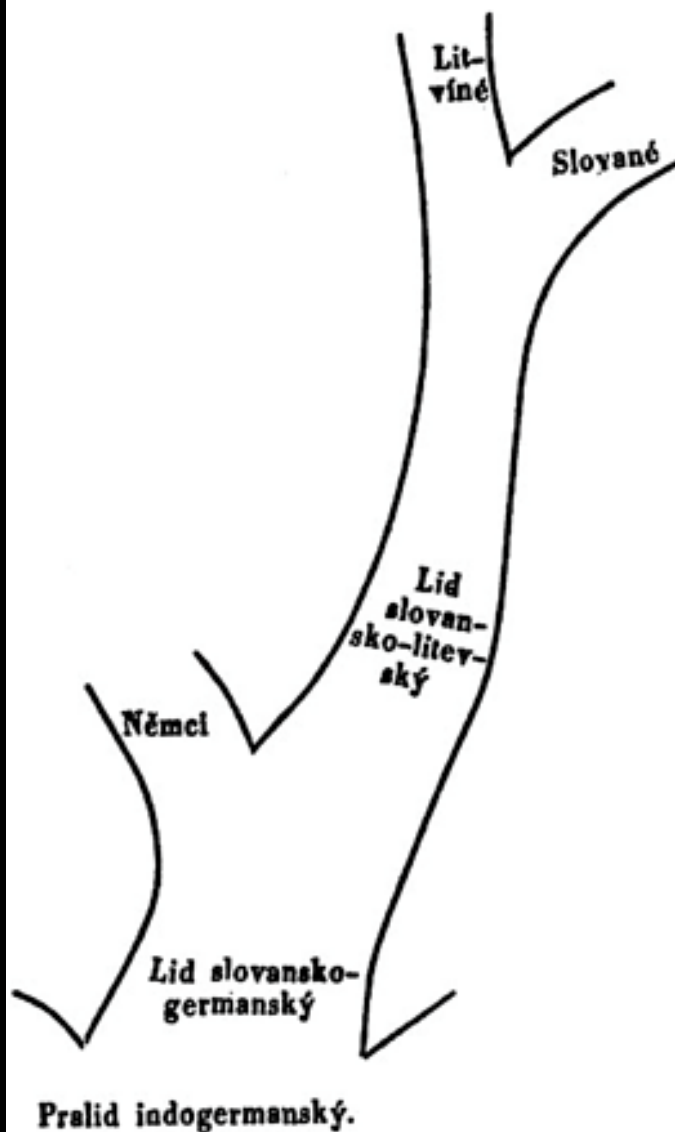
- 不確実性のあるデータ
 - 現在得られる手がかりだけでは、過去を確実に復元することはできない
 - 人間は多数の可能性を考慮しながら推論できない
 - 計算機は組み合わせ爆発に (近似的に) 対処できる
- 連続値を含む場合
 - e.g. 祖語の年代

今日のお話

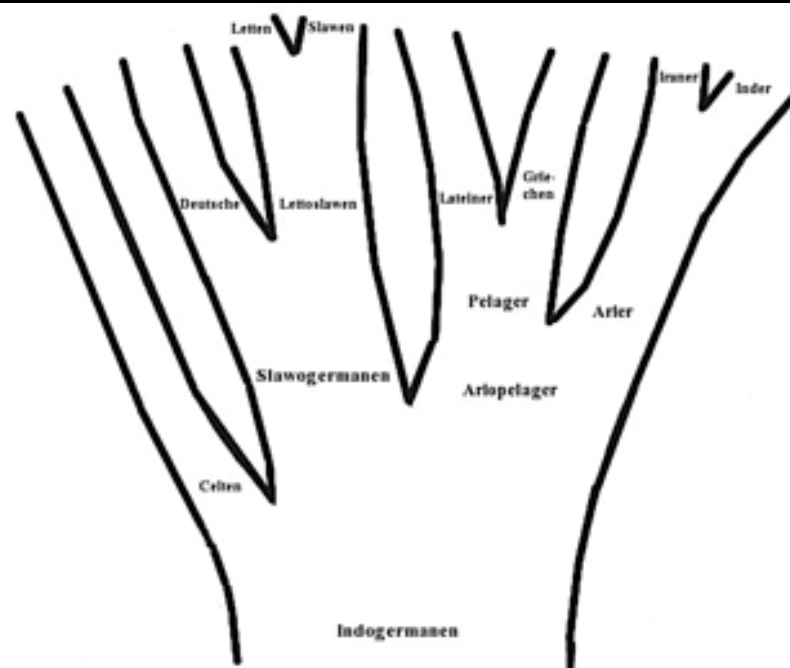
- 近年の言語系統の統計モデル
- 言語学と生物学との関わり
- 言語の性質を考慮したモデル化に向けて

歴史言語学と生物学の関わり

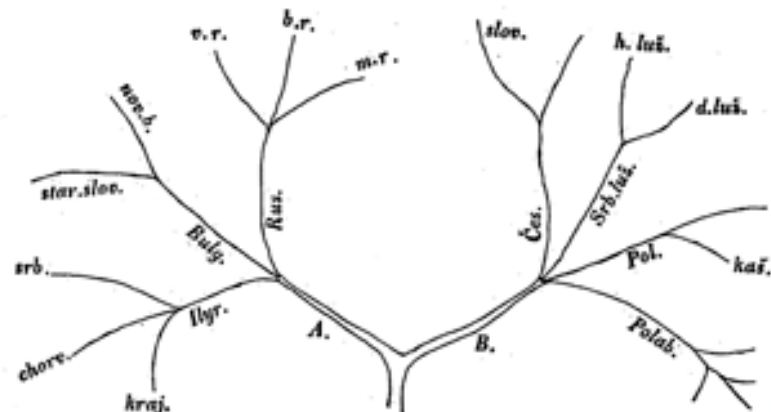
- 19世紀中頃：系統論の興隆
 - Schleicherの印欧語族系統樹
 - Darwinの進化論
- 20世紀中頃から終わり：統計モデルの登場
 - 言語年代学 [Swadesh, 1948,1951]
 - 激しい批判にさらされ統計的研究は低調に
 - 生物学の分子時計仮説 [Zuckerkandl+, 1965]
 - 順調に発展して現在にいたる
- 21世紀：生物から言語へのモデルの転用



A) August Schleicher, 1853 [21]



B) August Schleicher 1853 [22]



C) František Ladislav Čelakovský 1853 [24]

DNA = 基礎語彙?



DNA = 基礎語彙?



DNA = 基礎語彙?



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

DNA = 基礎語彙?



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

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

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

生物用モデルの転用は 長続きしない

	生物	言語
サンプルの規模	数十万 (ゲノムワイドSNP)	数百
1集団の表現	個体サンプルの集合	1サンプル
採取に必要な知識	汎用	個別言語の専門知識
過去のデータの拡 充	あり (e.g. Ötzi the Iceman)	あまり期待できない
進化	垂直 (水平伝播は例外的)	水平伝播を無視できる か怪しい
基本設計図?	YES (DNA)	NO? むしろ生物の表現型 (phenotype) に近い

今日のお話

- 近年の言語系統の統計モデル
- 言語学と生物学との関わり
- 言語の性質を考慮したモデル化に向けて

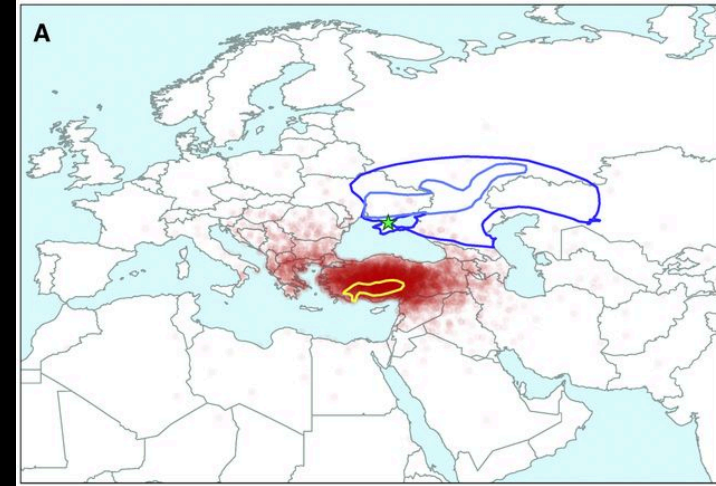
インド・ヨーロッパ祖語の 年代と故地 (Urheimat)

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
- 言語学者の広い支持

2. アナトリア仮説 (Bouckaert+はこちら)

- 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
- 批判: 印欧語アナトリア語派は祖語からかけ離れすぎ



[Bouckaert+, Science 2012]

成因的相同 (homoplasy)

[Chang+, 2015]

- 別個の変化により同じ (似た) 特徴を得る現象
- 成因的相同は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%

d^hg^hom-, ‘adult male’

現代アイル
ランド語
duine

フランス語
homme

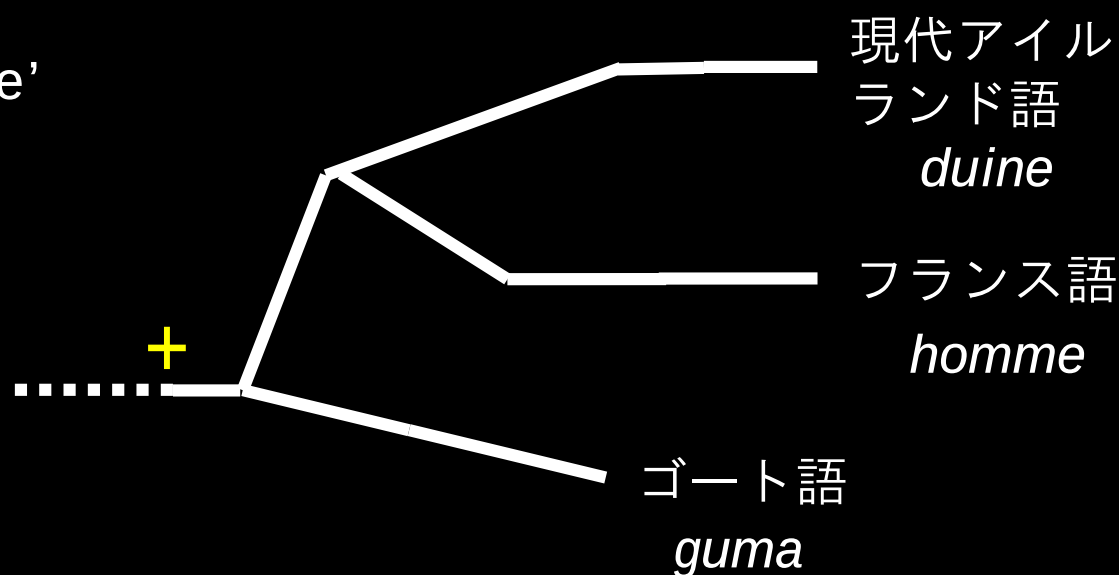
ゴート語
guma

成因的相同 (homoplasy)

[Chang+, 2015]

- 別個の変化により同じ (似た) 特徴を得る現象
- 成因的相同は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%

dhǵʰom-, 'adult male'



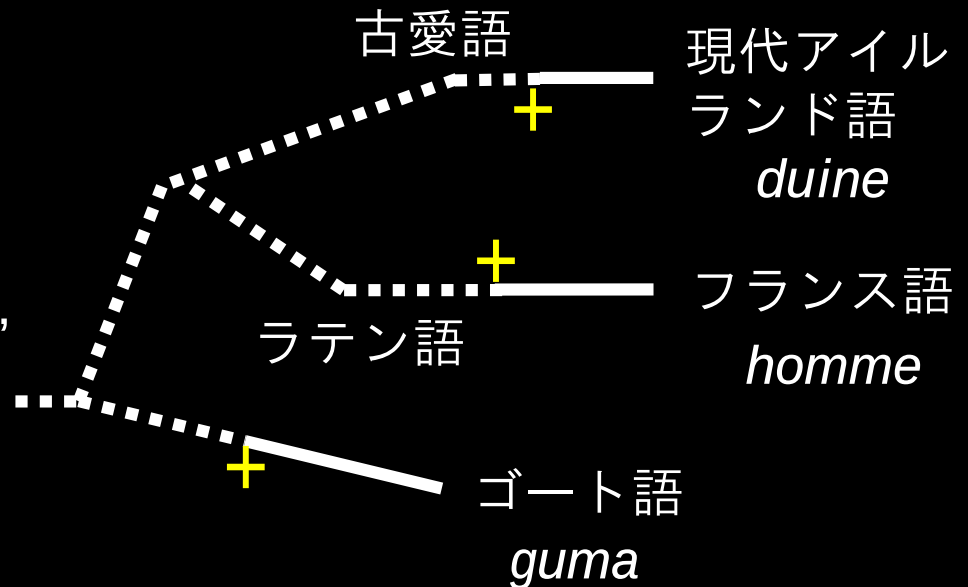
成因的相同 (homoplasy)

[Chang+, 2015]

dhǵʰom-, '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'



成因的相同 (homoplasy)

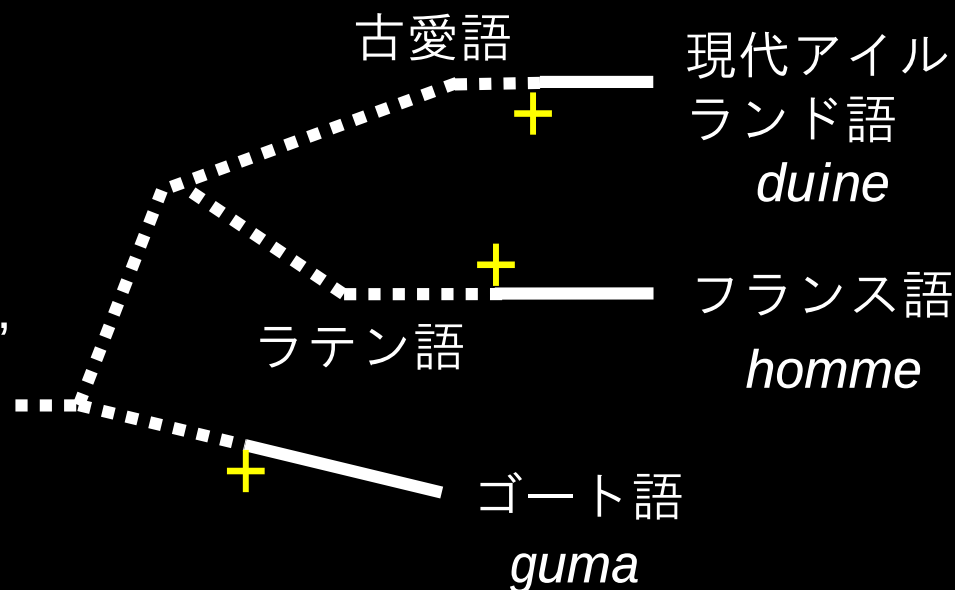
[Chang+, 2015]

- 提案手法: 古代語を制約として使う
- 結果: 祖語の年代は6,500年前となり、ステップ説に近づいた

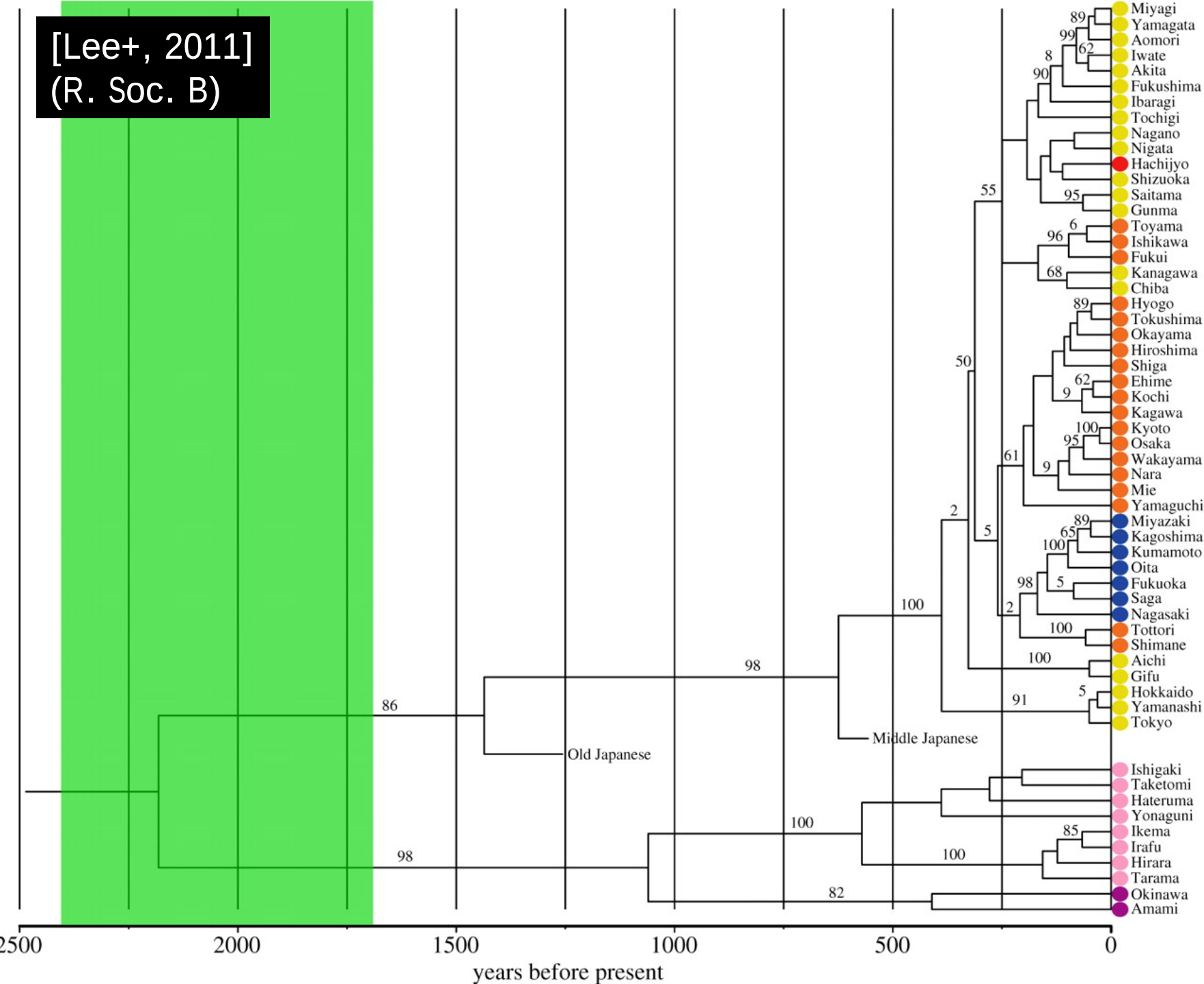
dhǵʰom-, 'adult male'

'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'

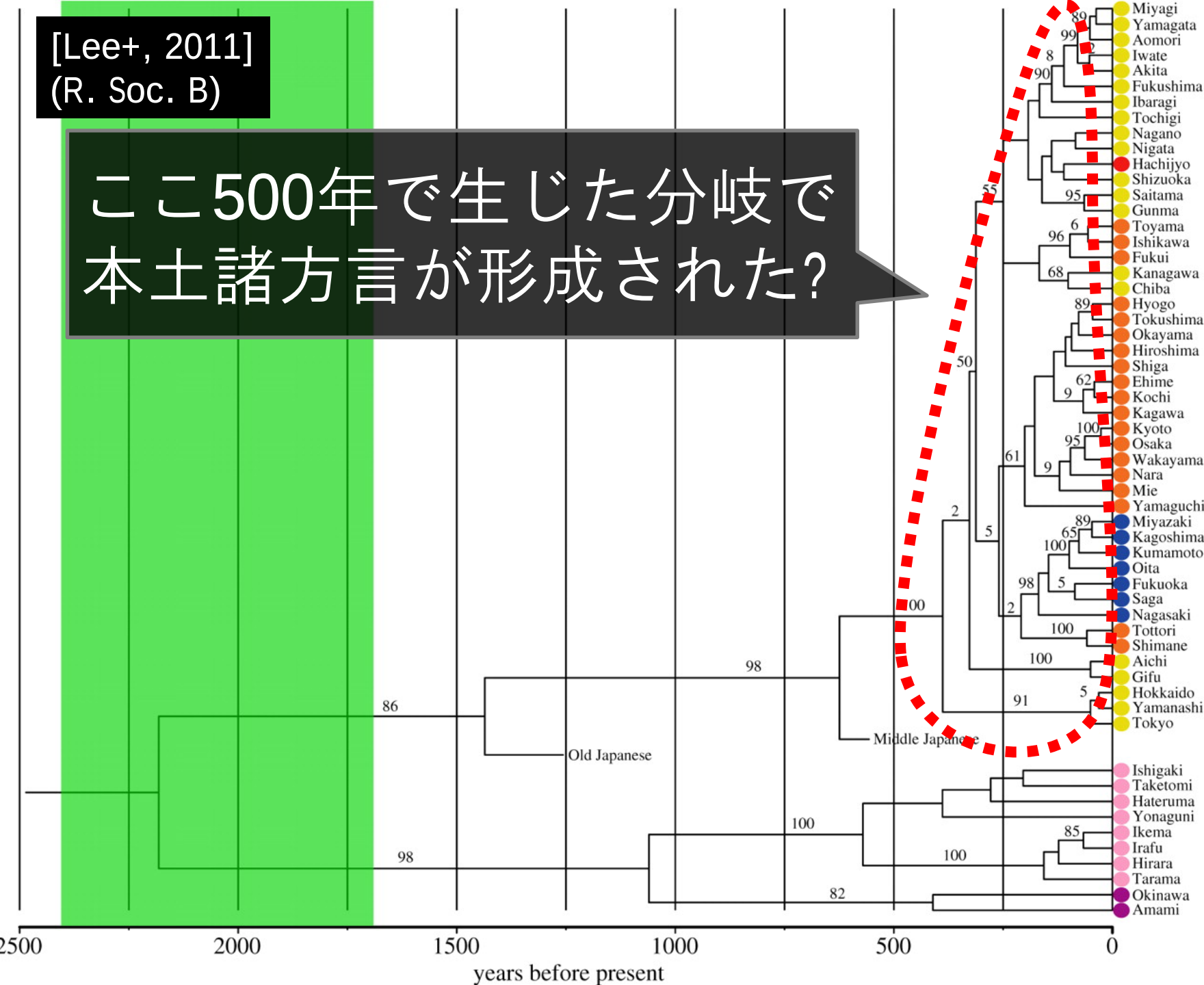


[Lee+, 2011]
(R. Soc. B)

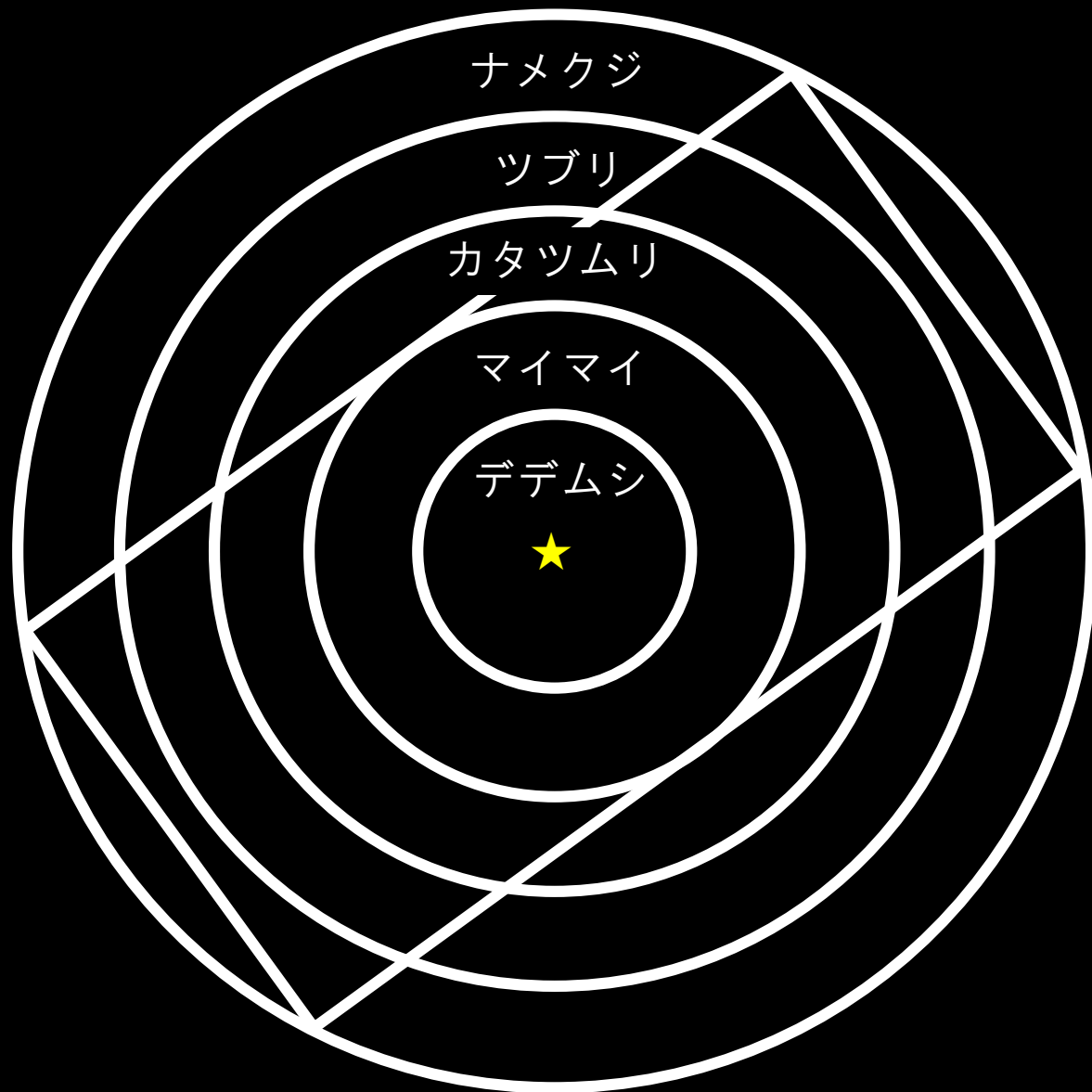


[Lee+, 2011]
(R. Soc. B)

ここ500年で生じた分岐で
本土諸方言が形成された?



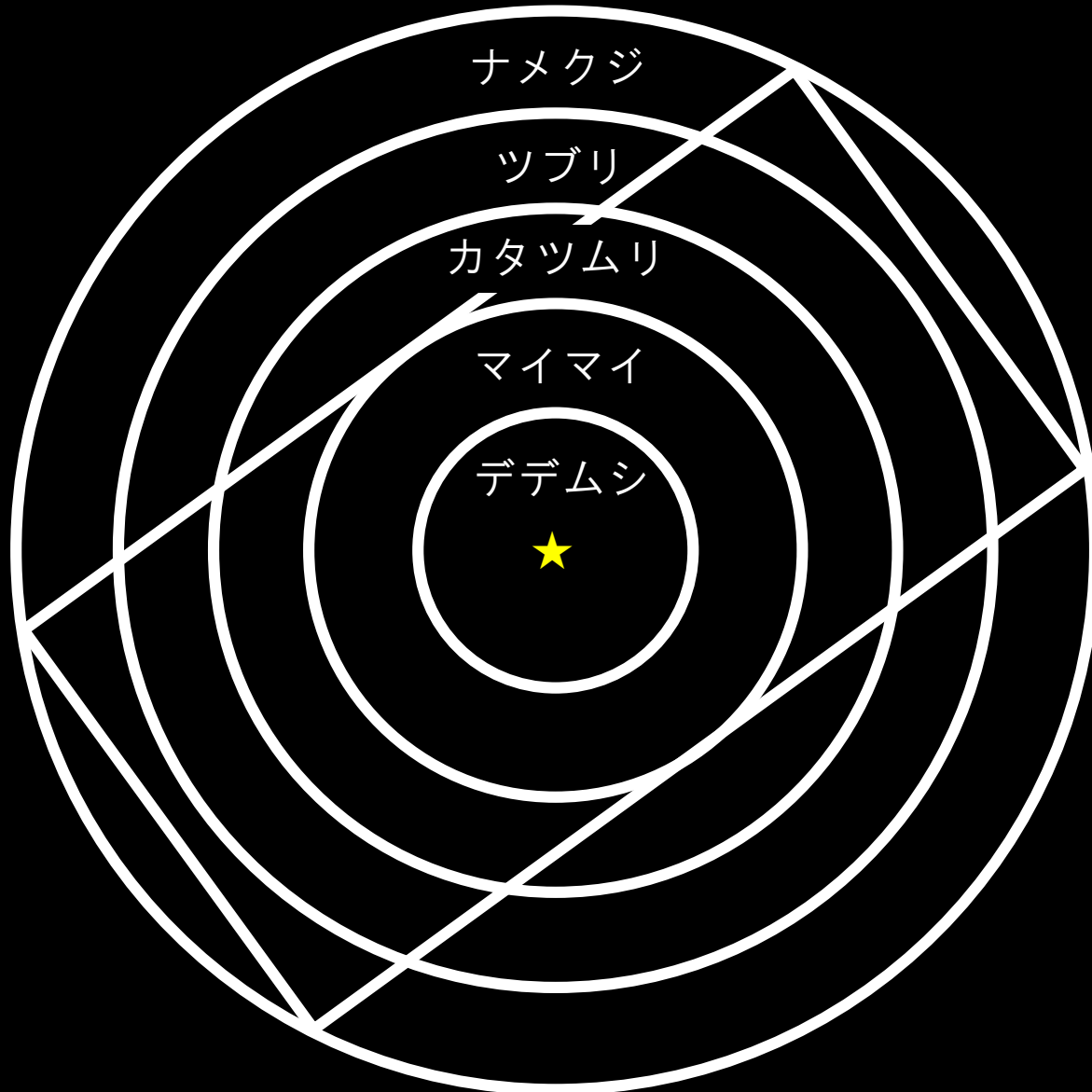
モデルとして木は妥当なのか



方言圏論

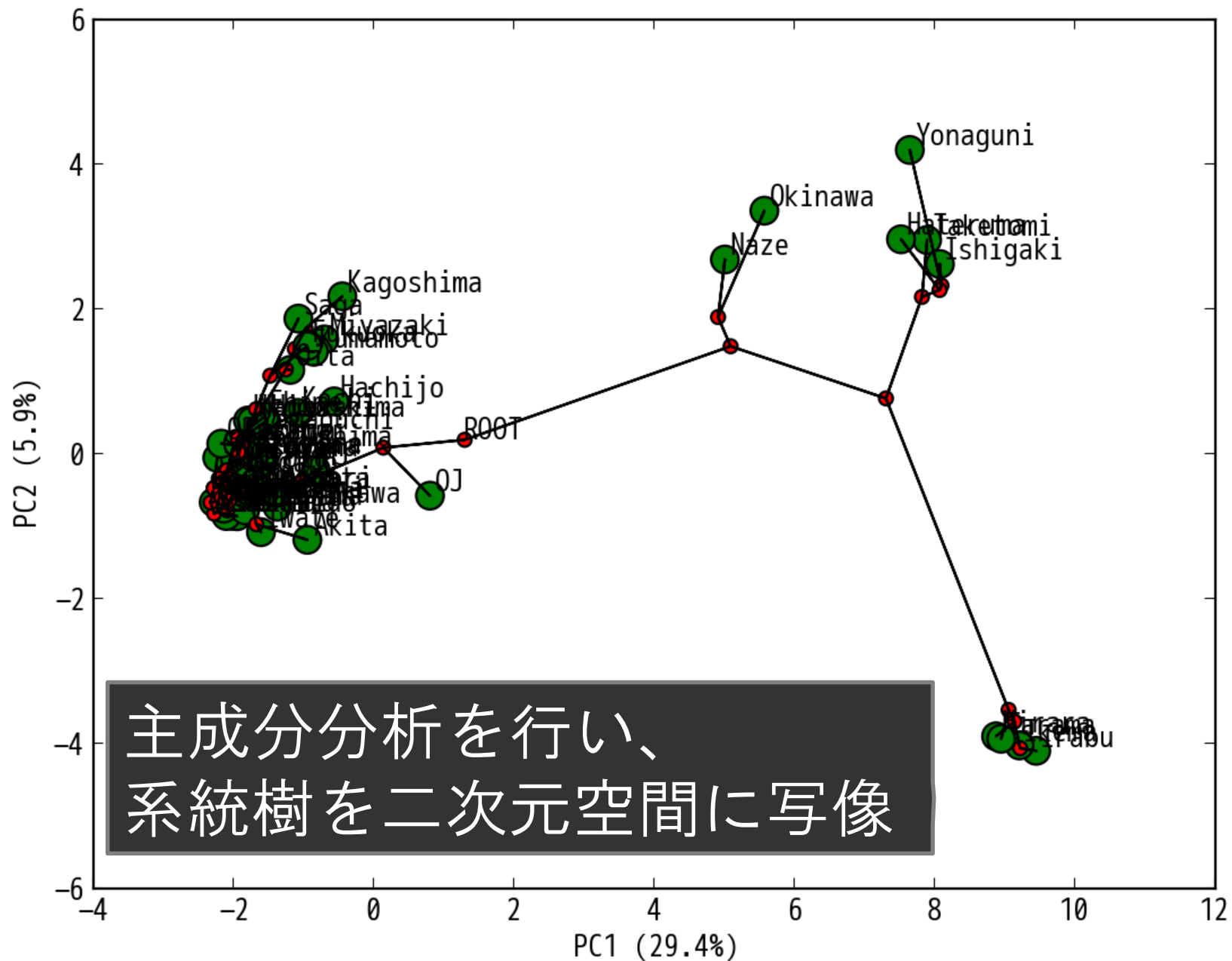
- 中央で生まれた語が周辺に伝播
- 結果として古語は周縁に残存

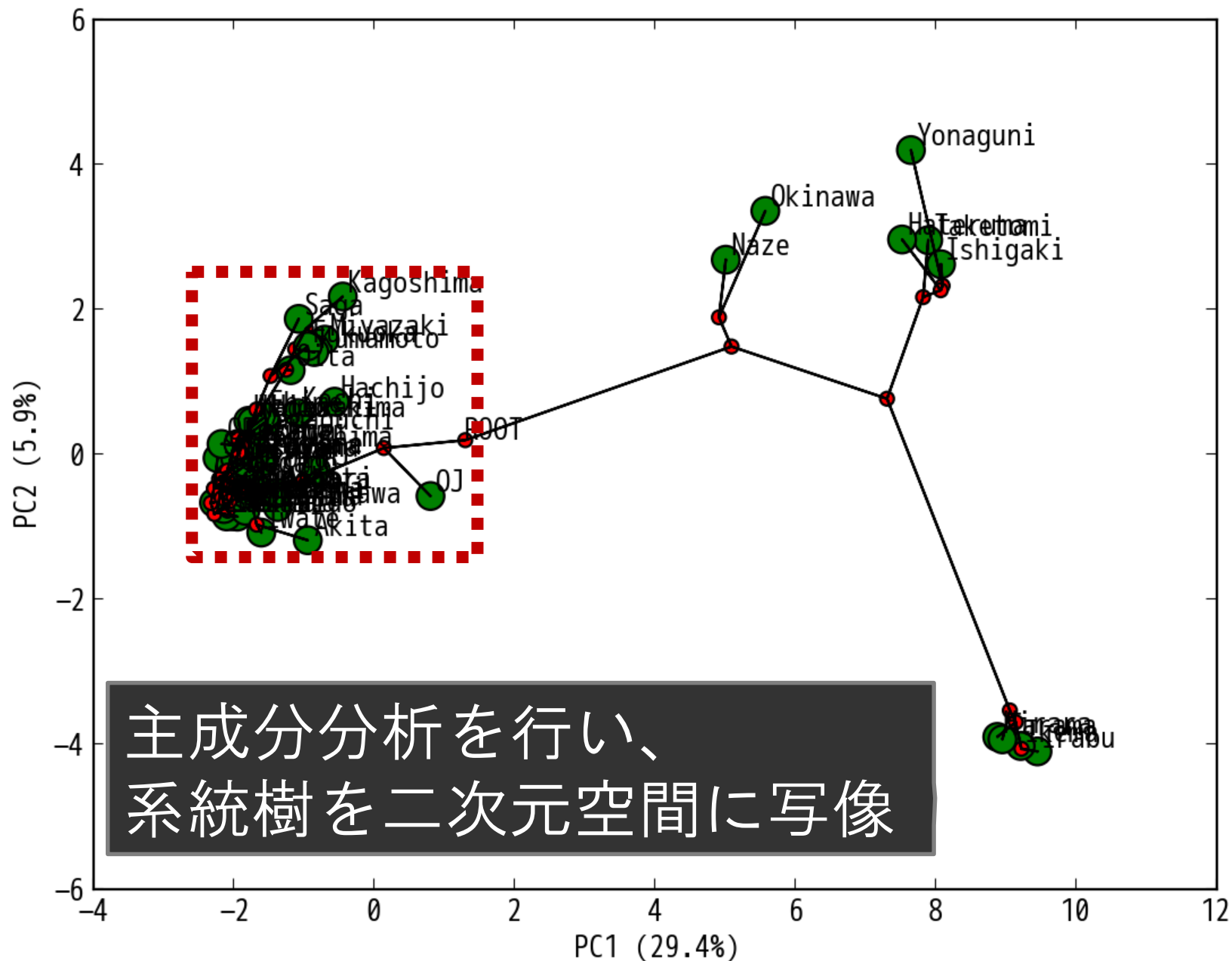
モデルとして木は妥当なのか

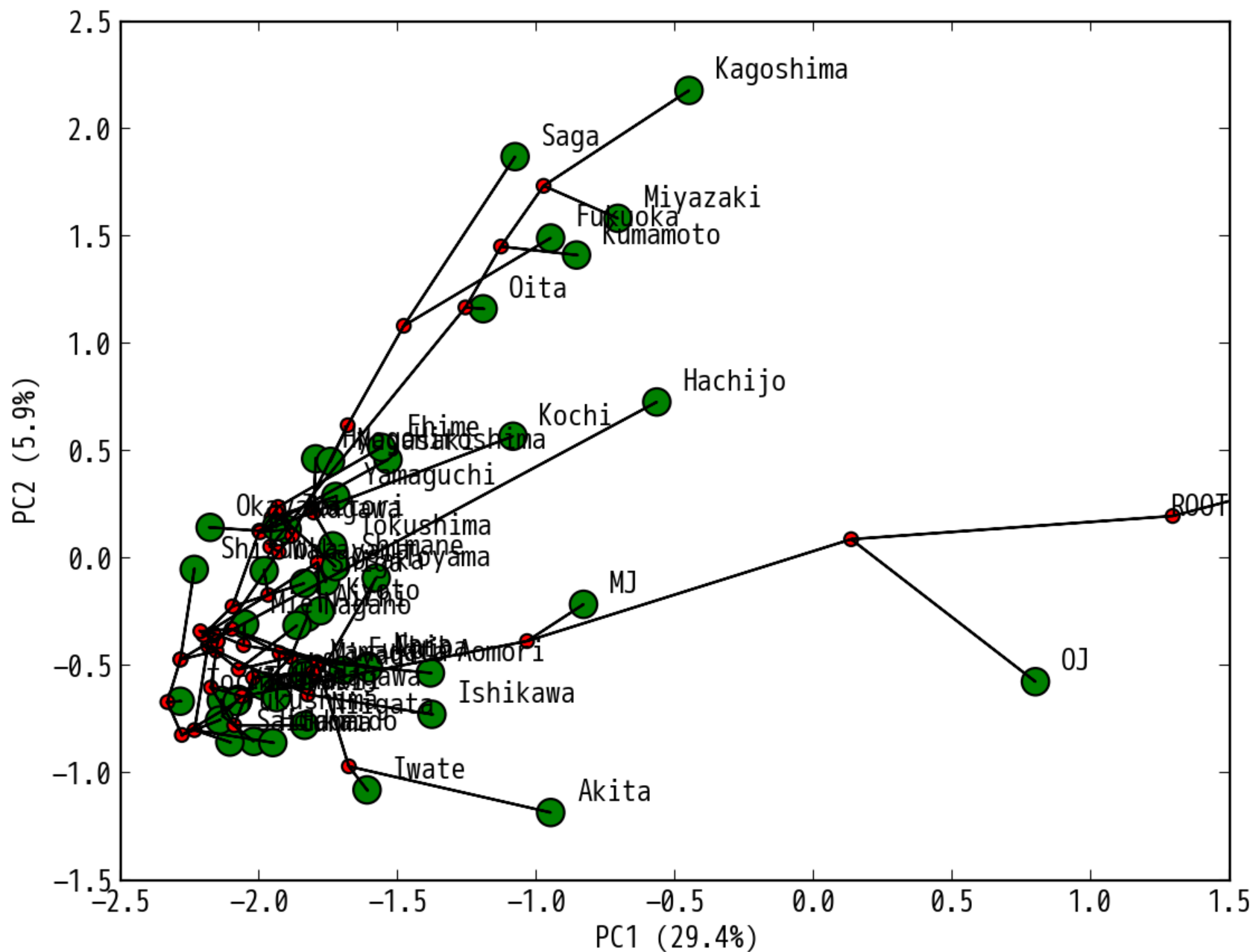


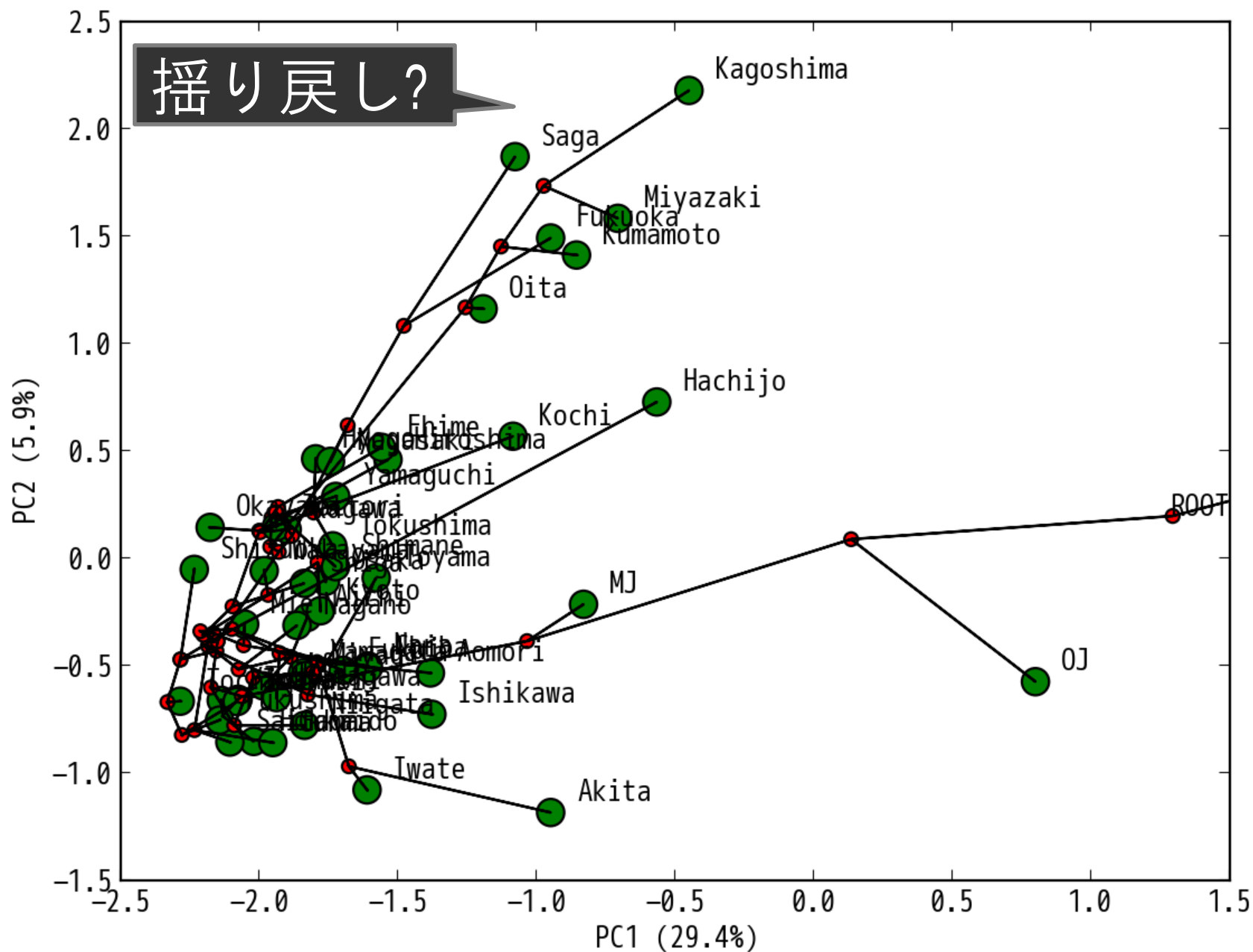
方言圏論

- 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
- 系統樹の仮定に反した水平伝播モデル
- 定量的分析の欠如



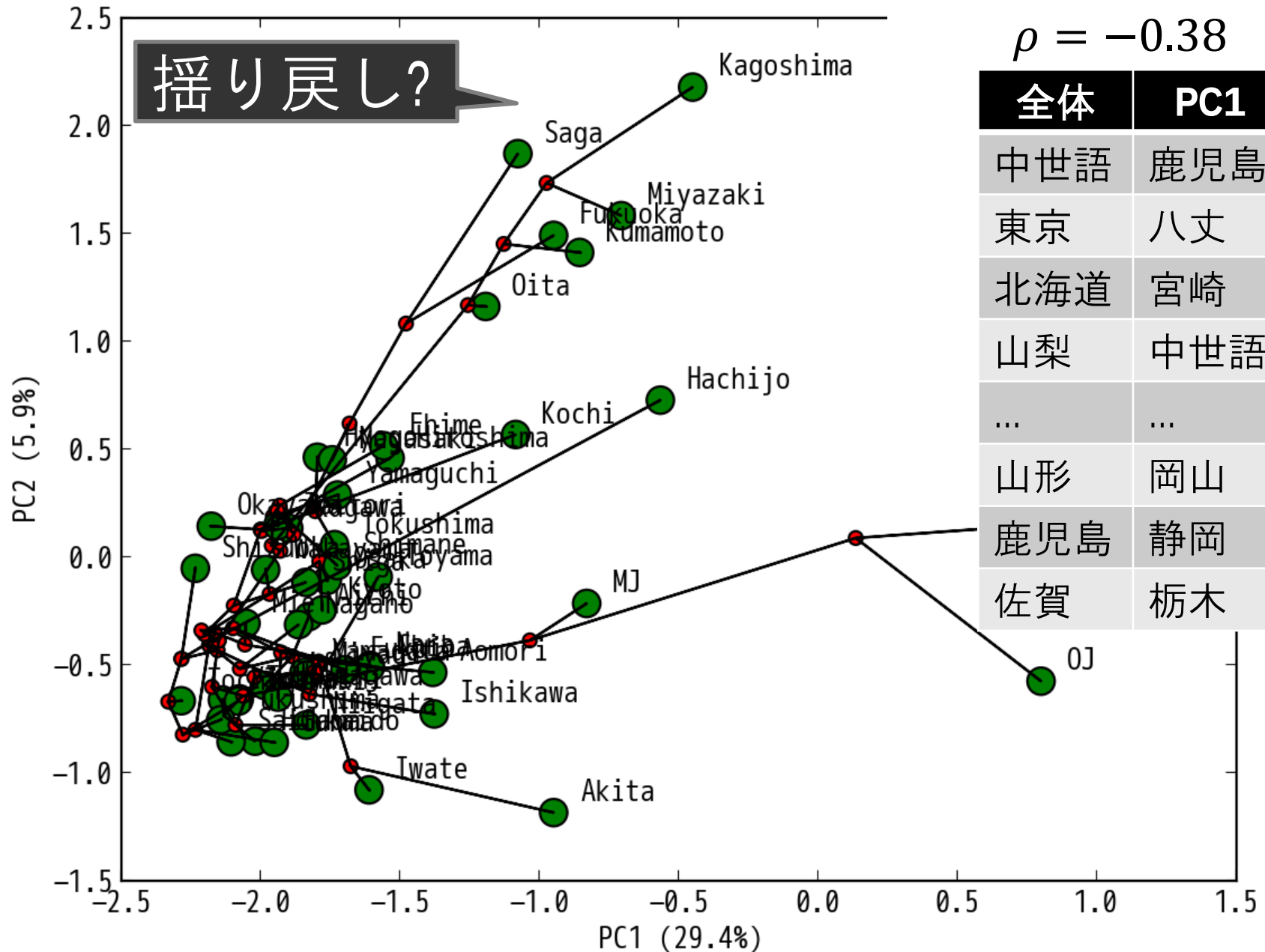






上代語との類似度順

$$\rho = -0.38$$



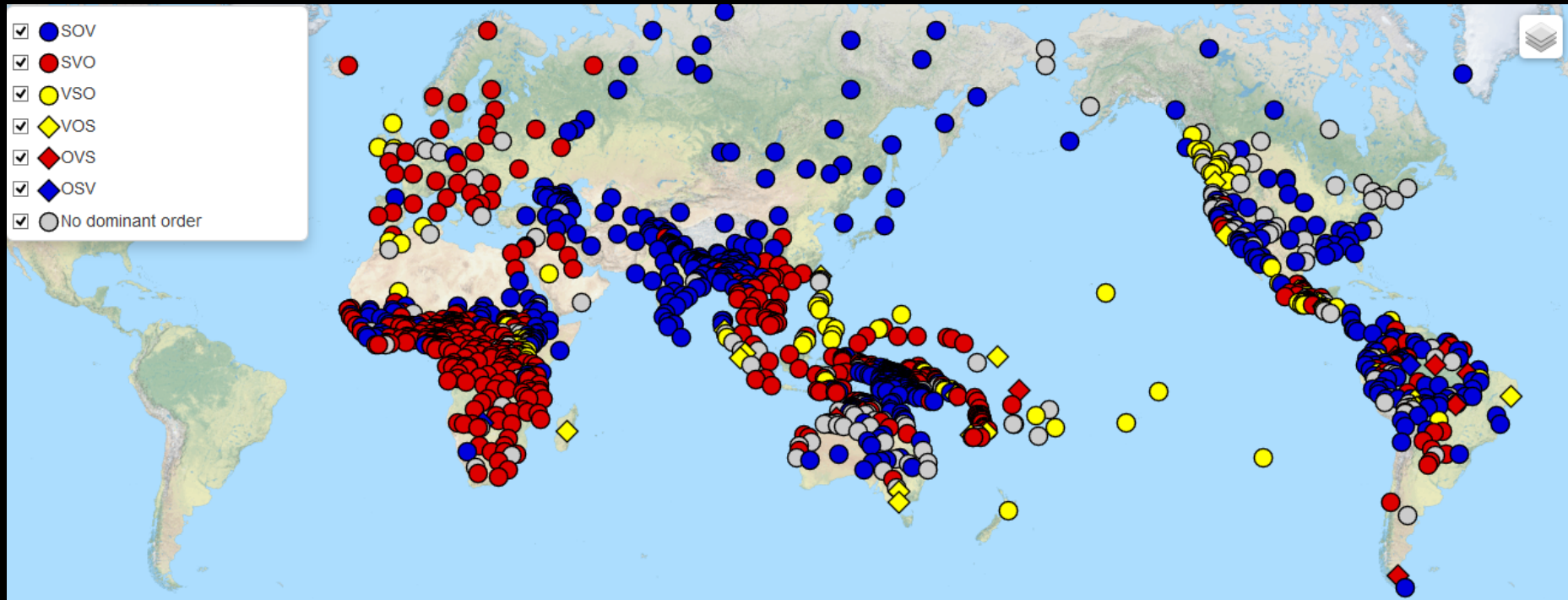
考えられる説明：周圏論 + 変化速度

- 中央部：全体の類似度 大, PC1の類似度 小
 - 変化の数は少ないが影響範囲が広い
- 周縁部：全体の類似度 小, PC1の類似度 大
 - 変化の数は多いが影響範囲が狭い

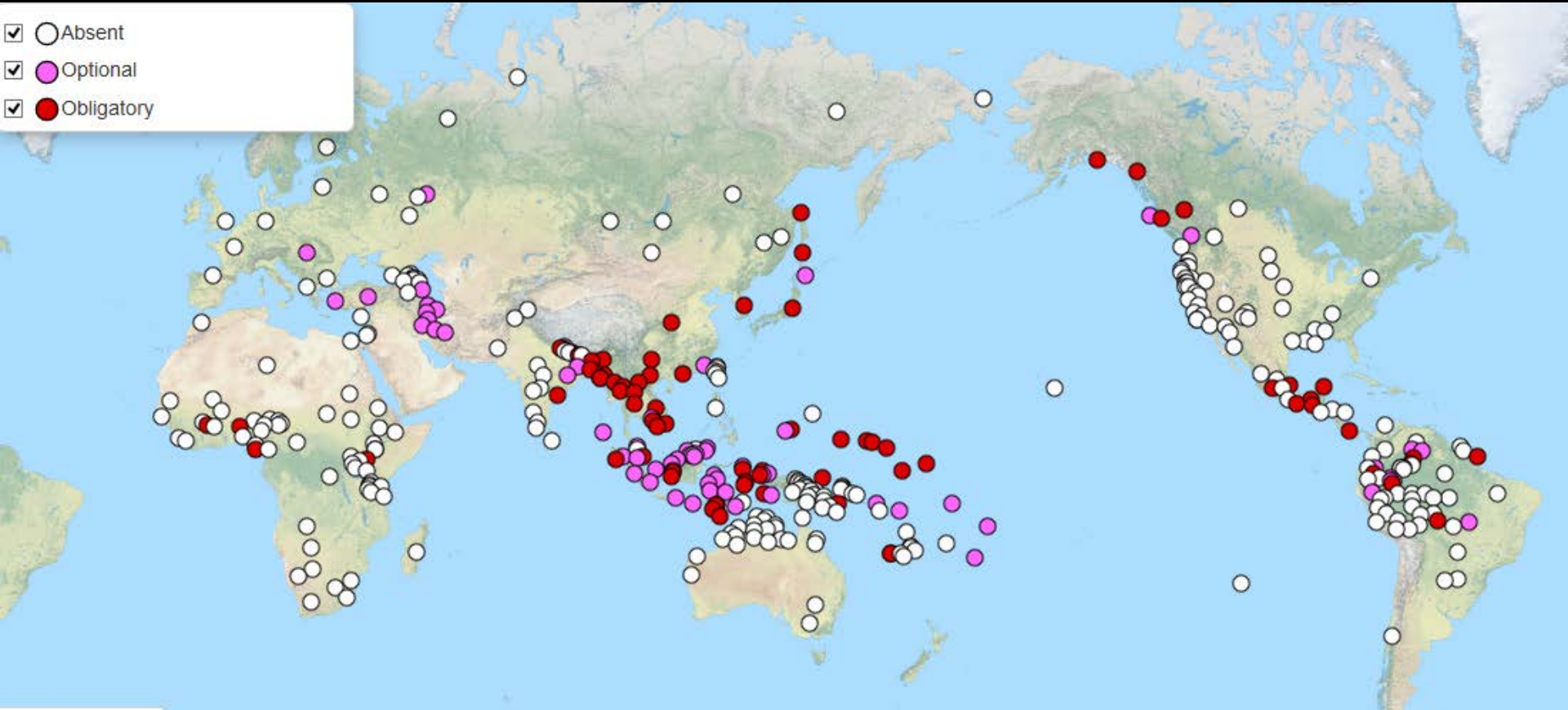
言語類型論 (Linguistic Typology)

- 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
- The World Atlas of Language Structures (WALS)
 - 2,679言語
 - 192種類の特徴量

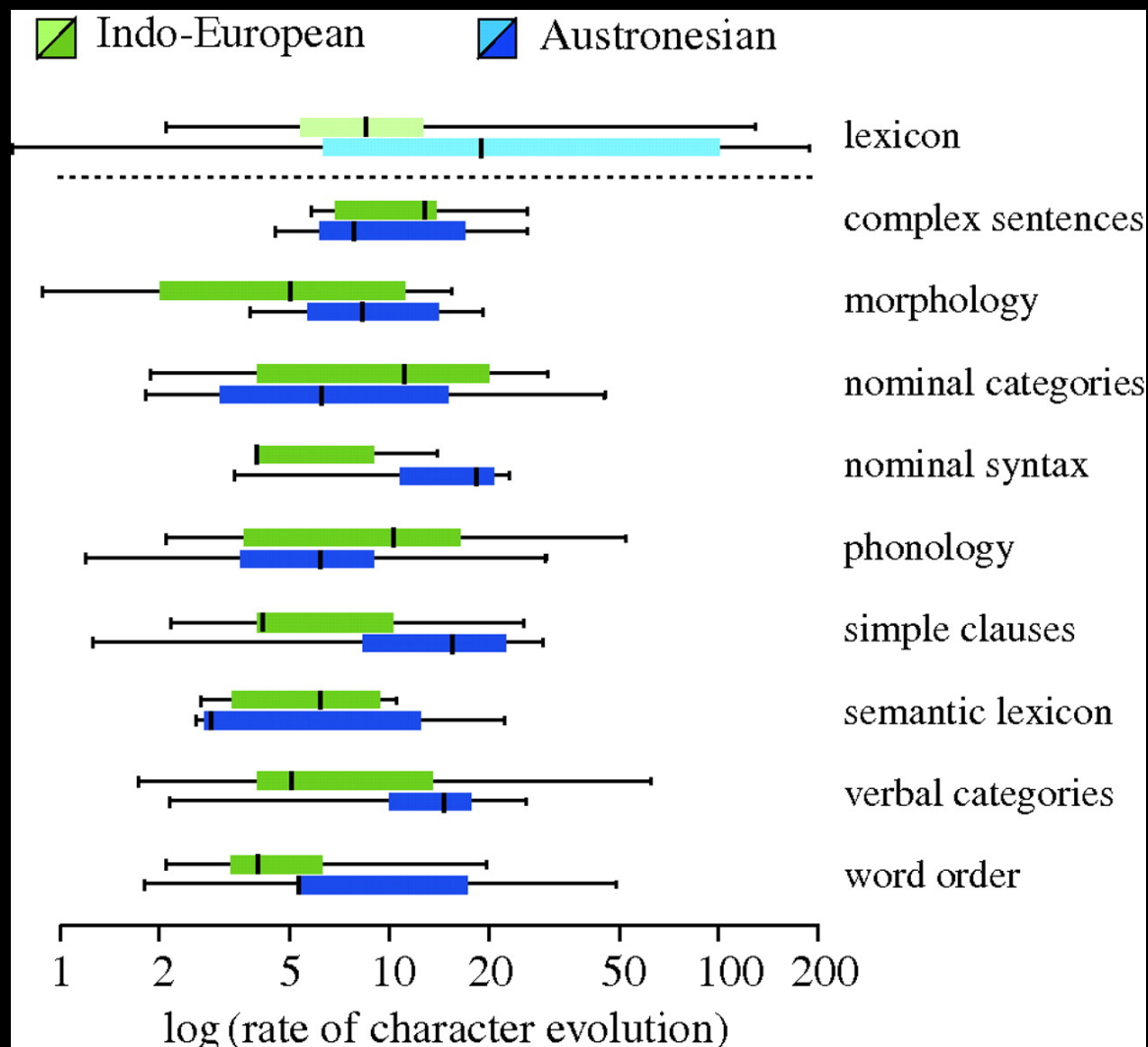
Feature 81A: Order of Subject, Object and Verb



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



言語類型の系統推定への応用

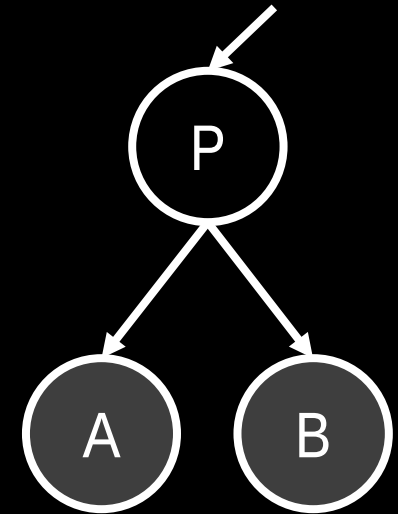


- 少ない
- 既存の語族を用いた調査では、有効性に疑義

[Greenhill+, 2010]
(R. Soc. B)

子孫を観測したとき、 祖語について何ができる？

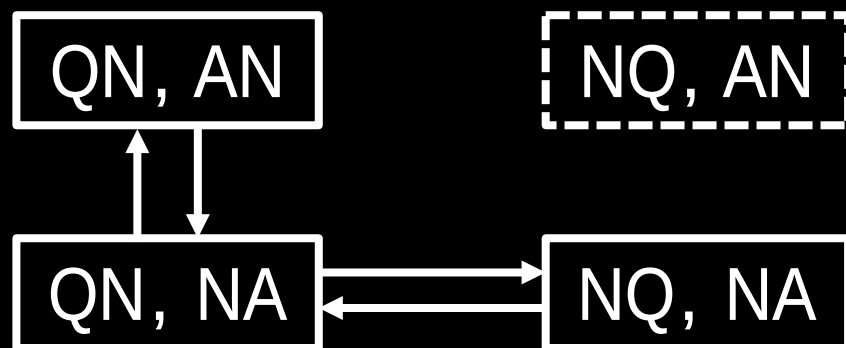
- 子孫から比較的近いはず
- 自然な言語であるはず
- もし言語に普遍的に成り立つ性質があるなら、祖語についても成り立つはず



言語類型の状態・過程モデル

[Greenberg, 1978]

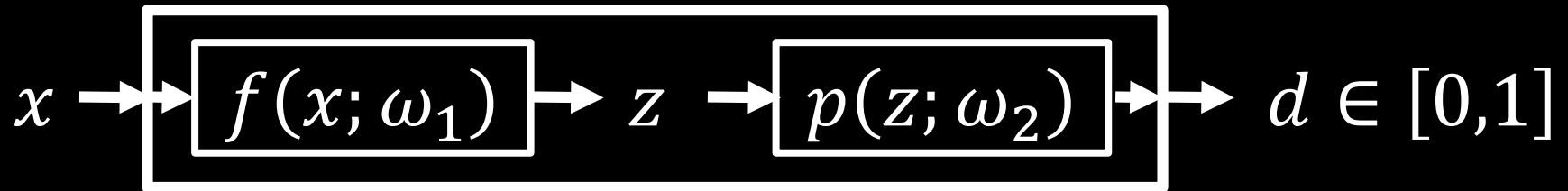
- 言語の普遍的な自然さ
 - どのような状態なら自然か
 - どのような変化なら自然か
- 特徴量間に依存



QN: 数詞 + 名詞 語順
AN: 形容詞 + 名詞 語順

言語類型の自然さ判定

[Murawaki, 2015]

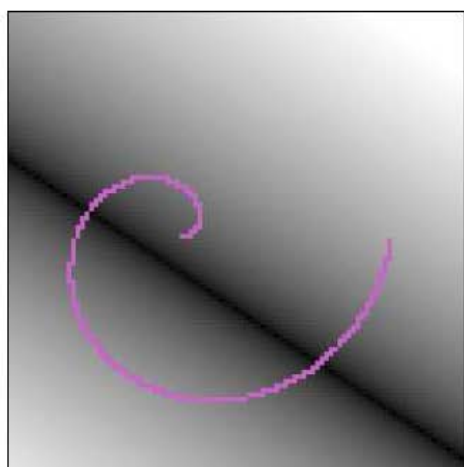


- x : 言語候補 $(0, 1, 0, 1, 0, 0, \dots, 0)$
- d : x の自然さ
- z : x の潜在表現 $(0.15, 0.98, 0.23, \dots, 0.02)$

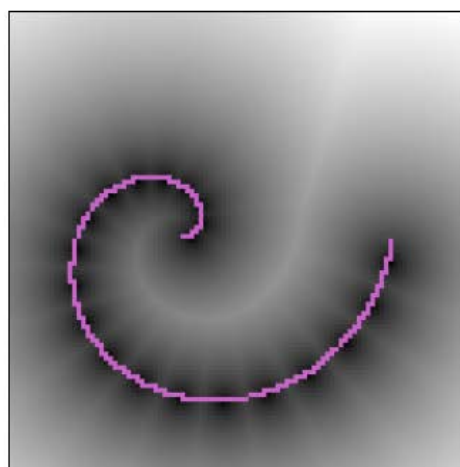
パラメータ ω_1, ω_2 を以下のように訓練

- 実在の言語 $x_p \in X$ に対する d を引き上げる
- それ以外の言語に対する d を引き下げる

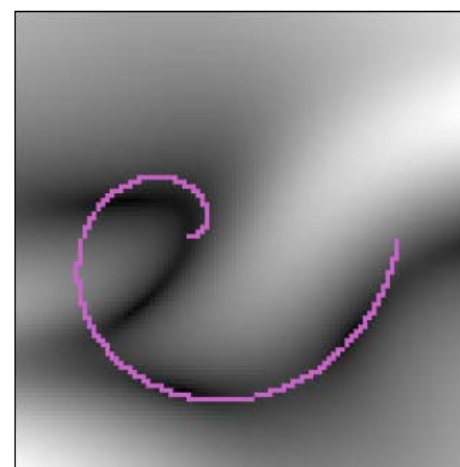
教師なし表現学習



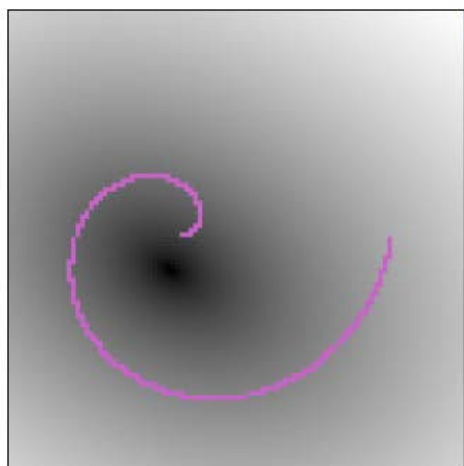
A



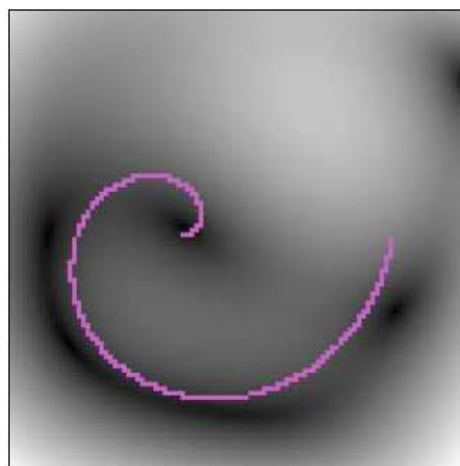
B



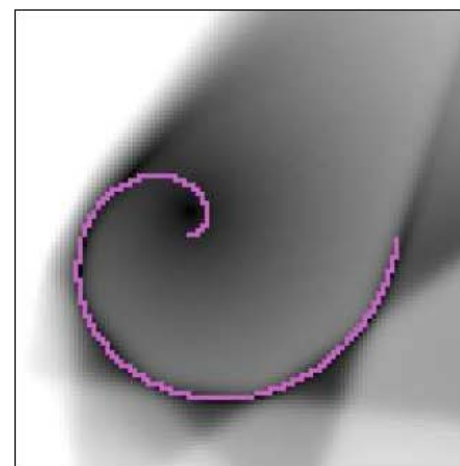
C



D



E



F

まとめ

- (機械可読な) 言語データベースが不可欠
- 不確実性・連続値は計算機にお任せ
- 生物用モデルの転用は長続きしない
- 言語を性質を考慮したモデル化が必要

