言語類型の連続空間表現とその系統推定への応用

九州大学
村脇 有吾

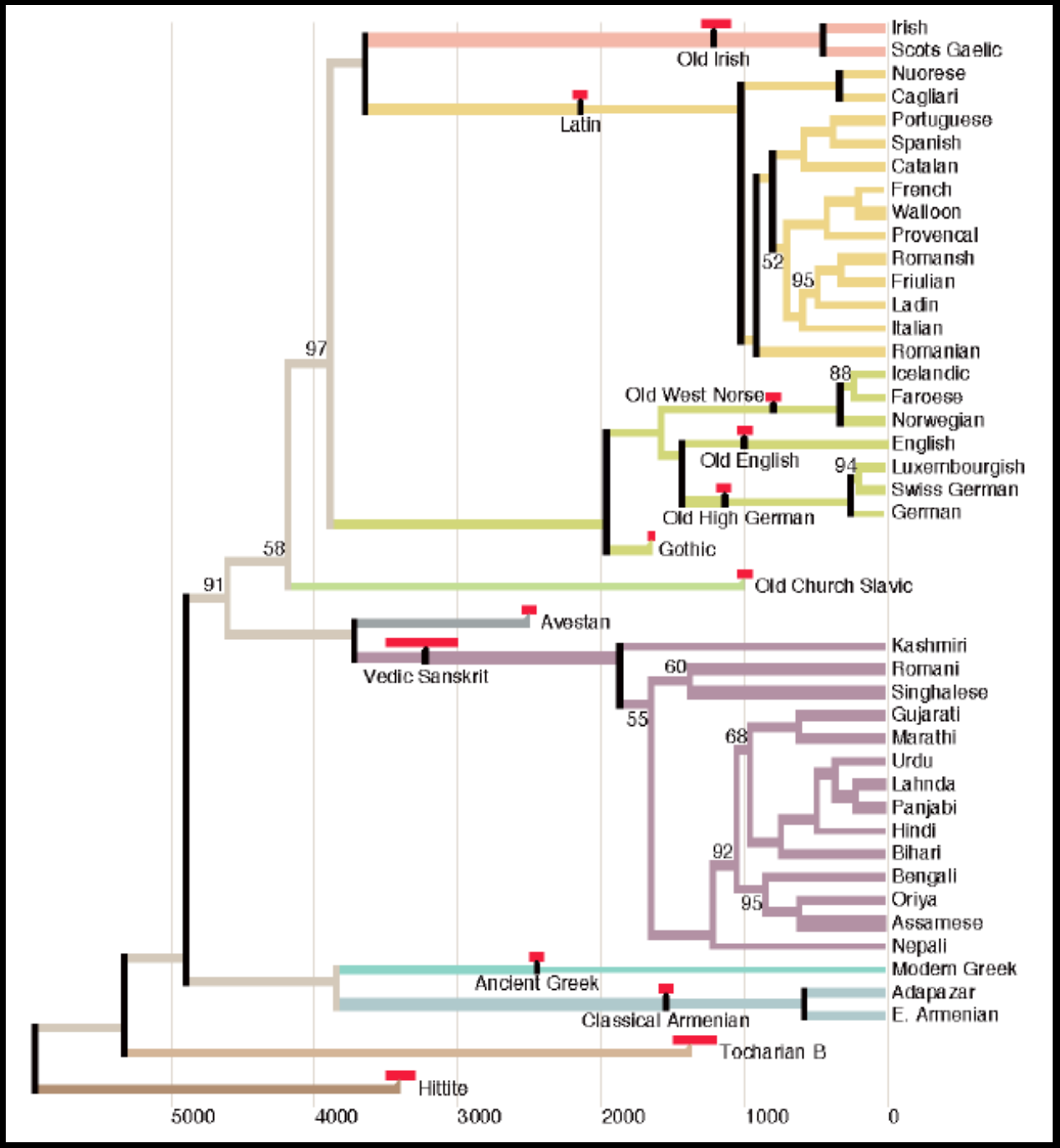
2015年3月17日 NLP2015@京大

今日のお話し

- 言語類型論は日本語系統論の最後の希望
- 言語普遍性に基づく自然な祖語の推定
- 単一起源仮説に基づく世界系統樹の推定

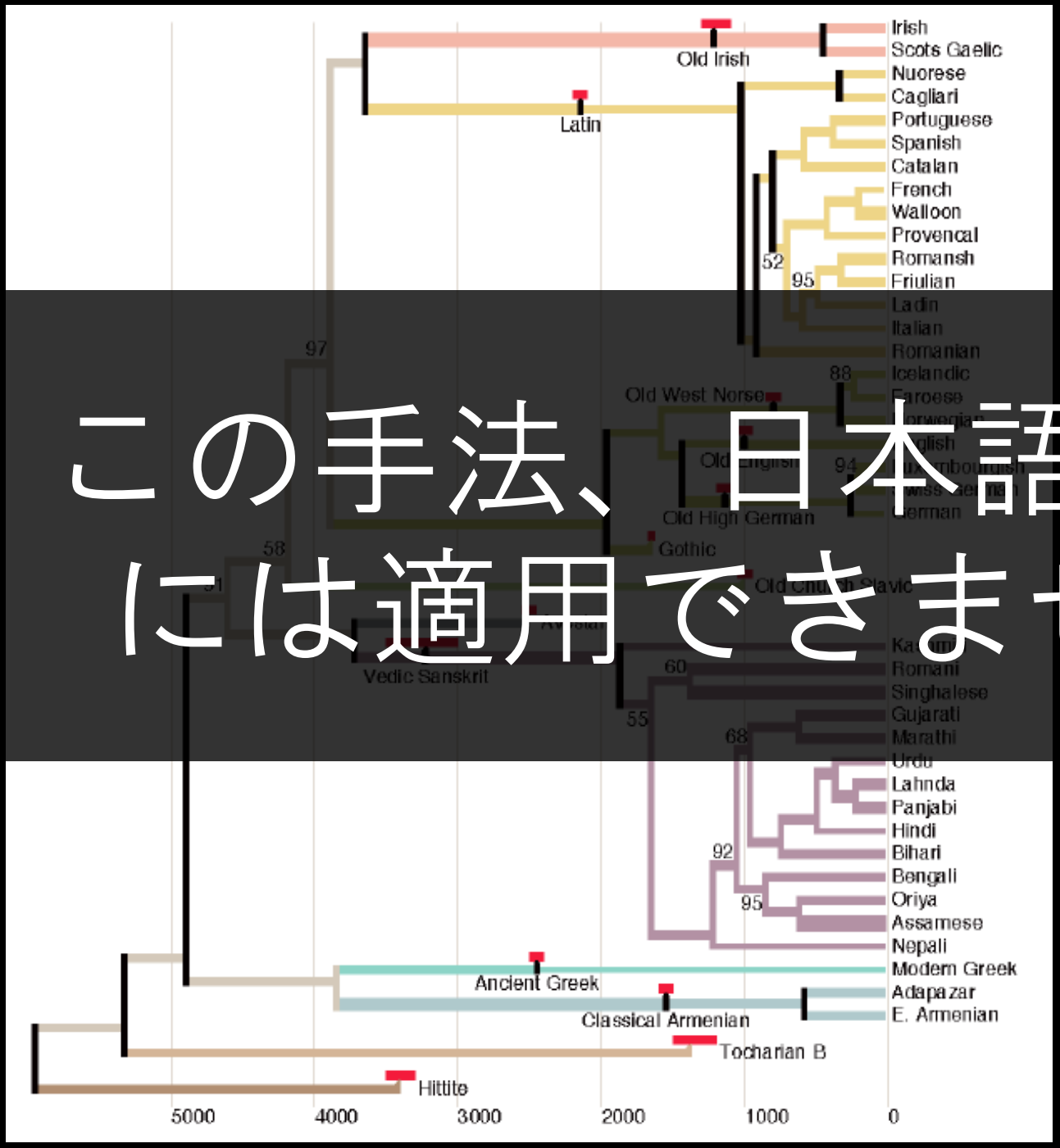
今日のお話し

- 言語類型論は日本語系統論の最後の希望
- 言語普遍性に基づく自然な祖語の推定
- 単一起源仮説に基づく世界系統樹の推定



インド・ヨーロッパ語族の系統研究における計算モデルの隆盛

[Chang+, 2015]



この手法、日本語系統論には適用できません!!!

インド・ヨーロッパ語族の系統研究における計算モデルの隆盛

[Chang+, 2015]

なぜ既存の計算モデルは日本語系統論に適用できないか

- そもそも他の言語とまともに比較できない
  - 既存手法は語源が同じ語彙の共有を前提とするが、そうした語彙が得られない
- 仮に共通祖語が存在しても、語彙に基づく手法ではたどれないほど古いのでは?
  - 言語年代学の推定では、日本語と朝鮮語の共通祖先は楽観的に見ても6,700年前 [服部, 1999]

類型論は日本語系統論最後の希望

- そもそも他の言語とまともに比較できない
  - 既存手法は語源が同じ語彙の共有を前提とするが、そうした語彙が得られない
- 仮に共通祖語が存在しても、語彙に基づく手法ではたどれないほど古いのでは?
  - 言語年代学の推定では、日本語と朝鮮語の共通祖先は楽観的に見ても6,700年前 [服部, 1999]

類型論は日本語系統論最後の希望

- そもそも他の言語とまともに比較できない
  ⇒どんな言語同士でも比較できる
- 仮に共通祖語が存在しても、語彙に基づく手法ではたどれないほど古いのでは?
  - 言語年代学の推定では、日本語と朝鮮語の共通祖先は楽観的に見ても6,700年前 [服部, 1999]

類型論は日本語系統論最後の希望

- そもそも他の言語とまともに比較できない
  ⇒どんな言語同士でも比較できる

- 仮に共通祖語が存在しても、語彙に基づく手法ではたどれないほど古いのでは?
  ⇒言類型論の特徴量は語彙よりずっと変化が遅いから、遠い過去までたどれる・・・かも

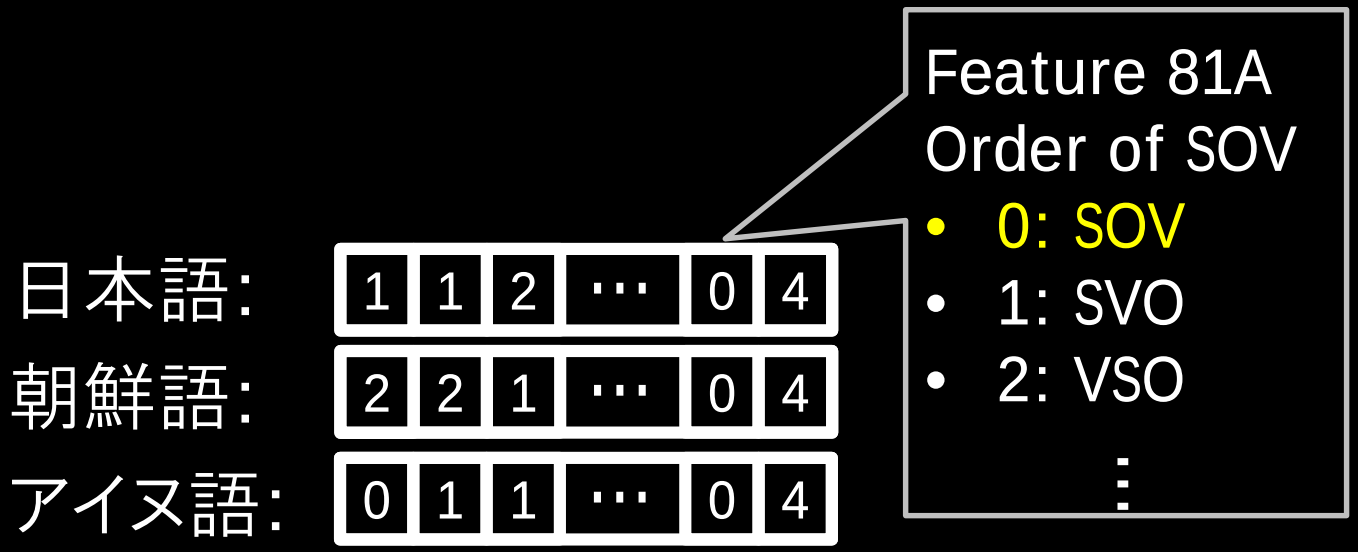
どんな言語同士でも比較できる

日本語: 1 1 2 … 0 4

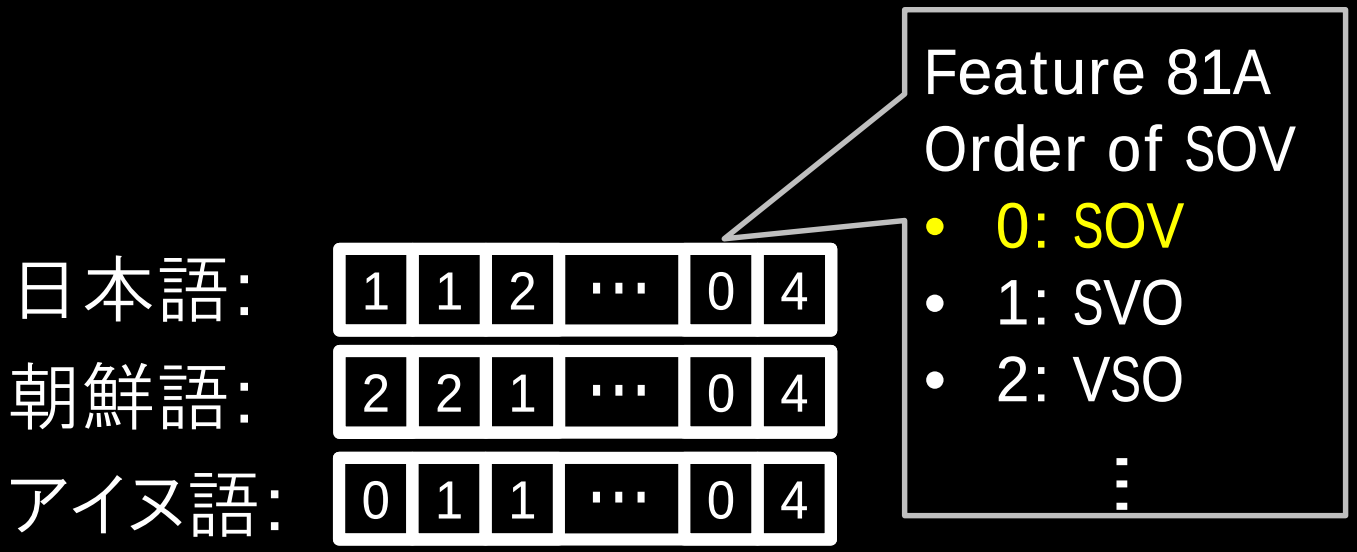
朝鮮語: 2 2 1 … 0 4

アイヌ語: 0 1 1 … 0 4

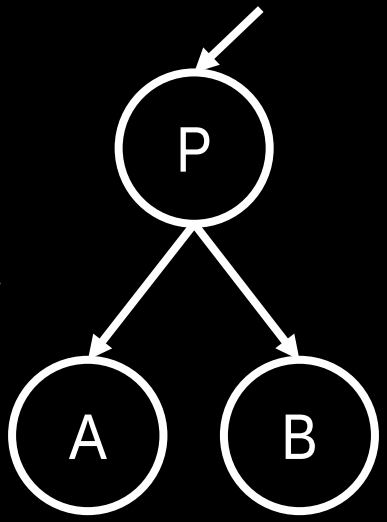
どんな言語同士でも比較できる



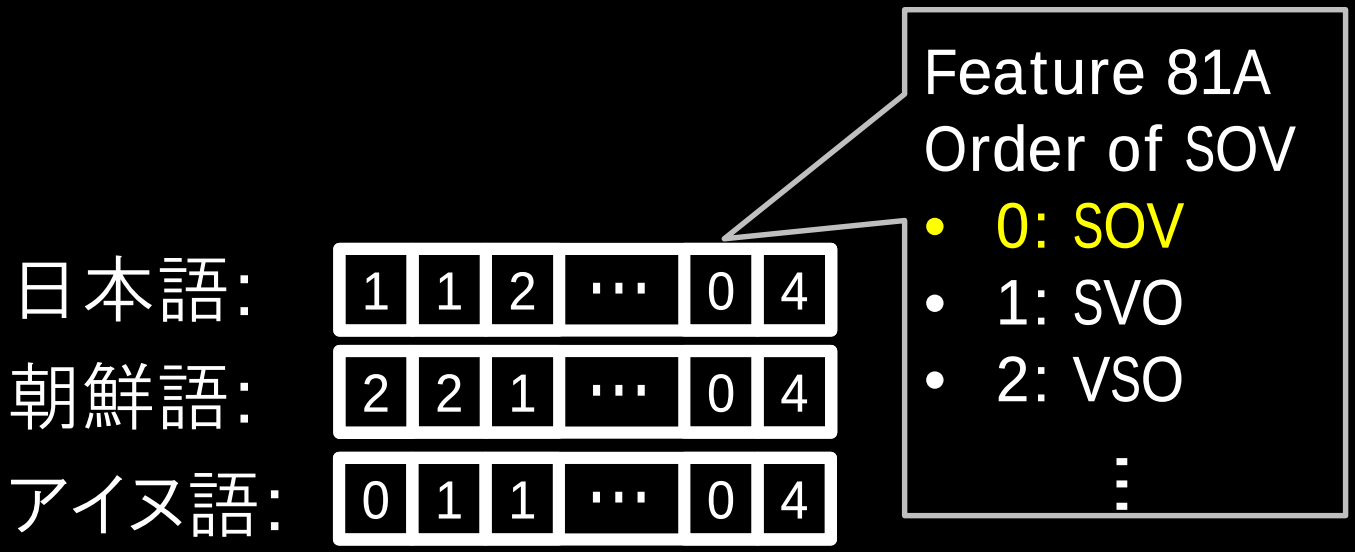
どんな言語同士でも比較できる



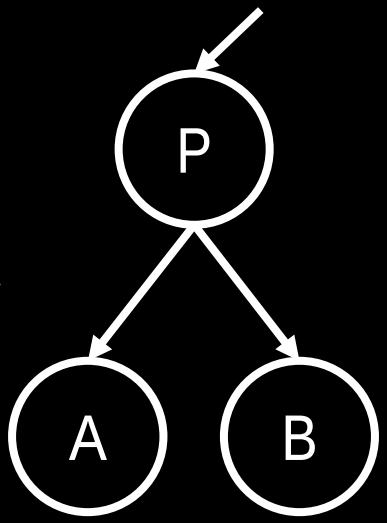
- 系統上近い → ベクトルの距離が近い

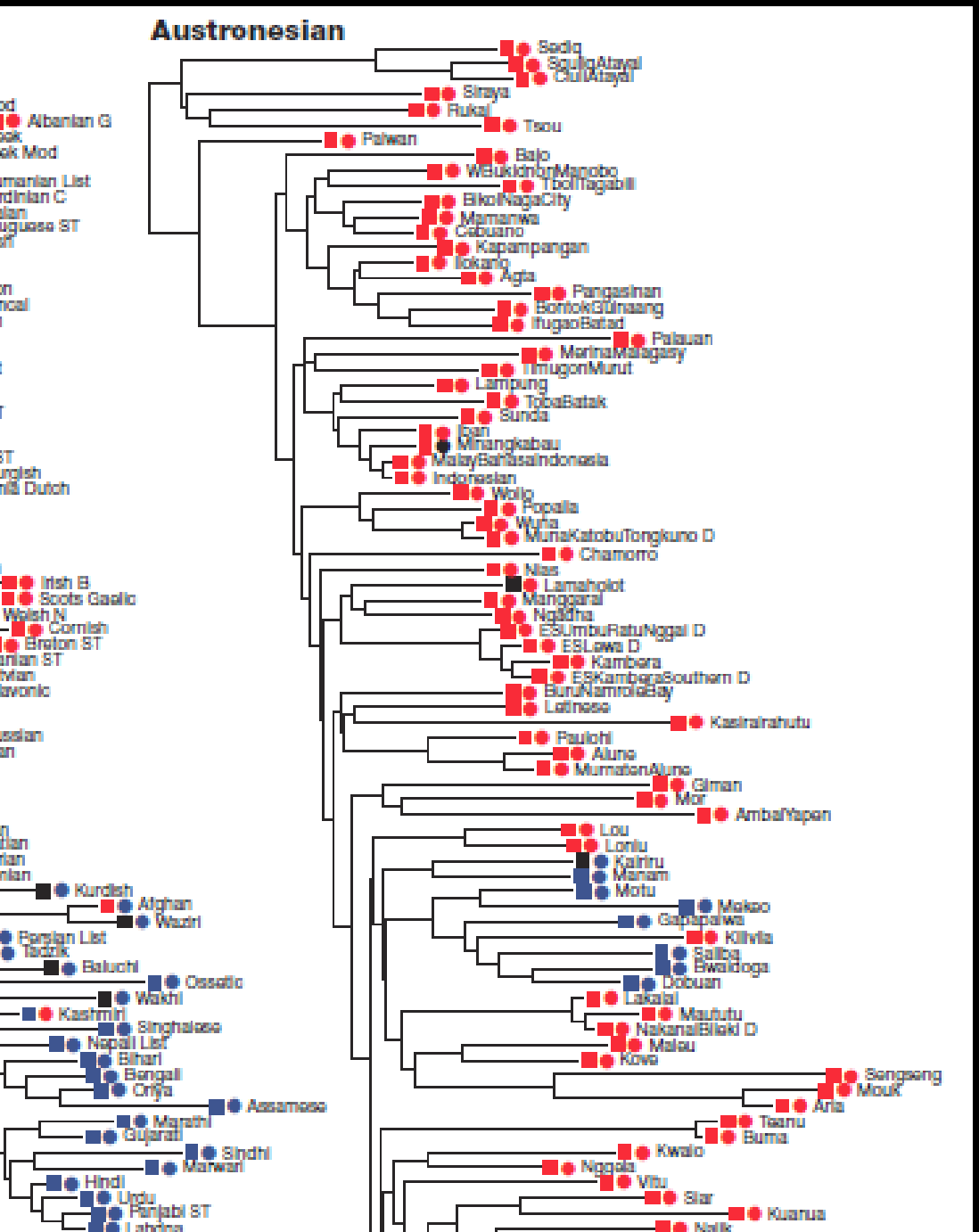


どんな言語同士でも比較できる



- 系統上近い $\overset{?}{\leftrightarrows}$ ベクトルの距離が近い

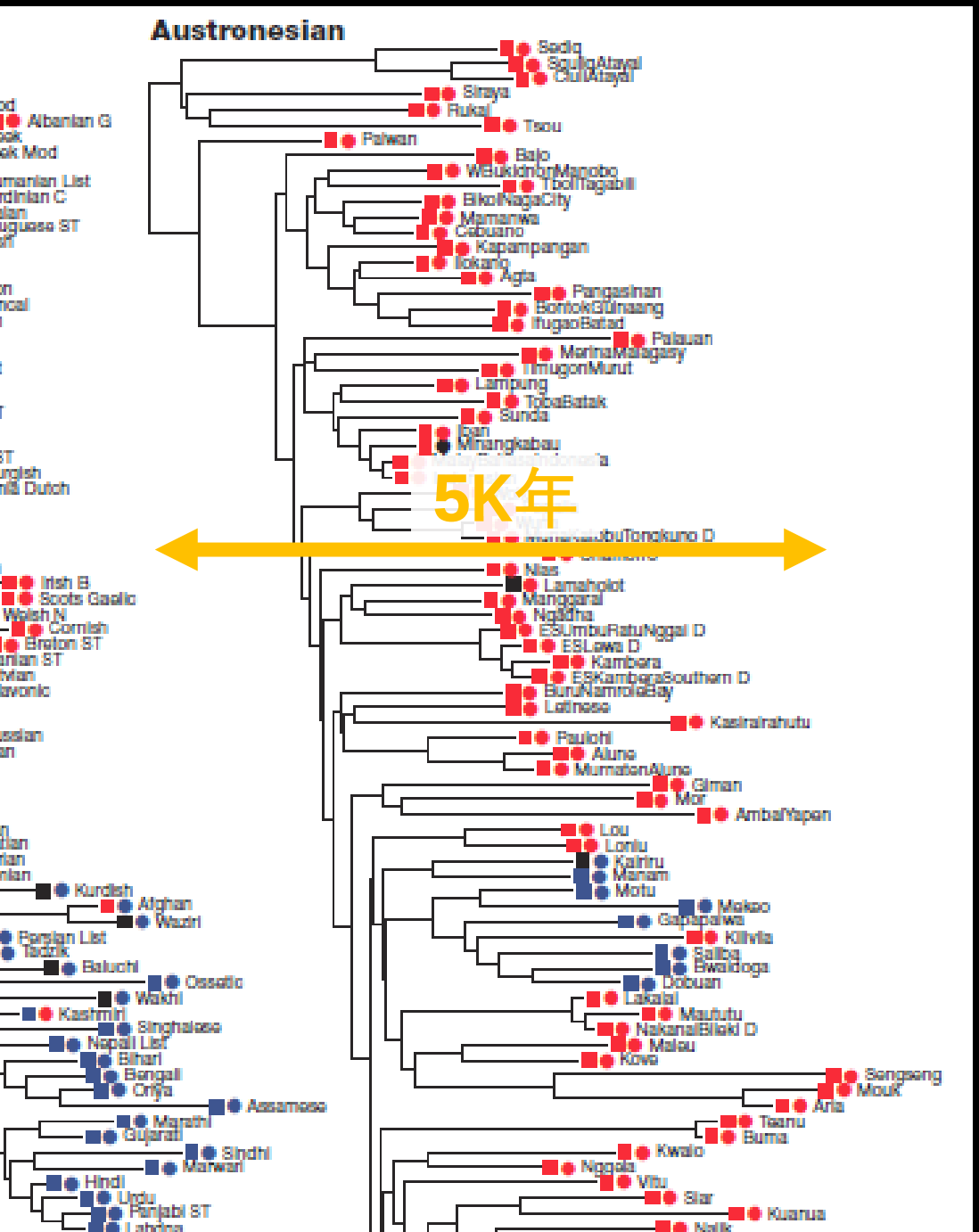




類型論の特徴量の変化は長期的

もし特徴量変化のふるまいが統計的に予測可能なら、遠い過去までたどれる・・・が、実際に予測可能かは未知数

[Dunn+, Nature 2011]



類型論の特徴量の変化は長期的

もし特徴量変化のふるまいが統計的に予測可能なら、遠い過去までたどれる・・・が、実際に予測可能かは未知数

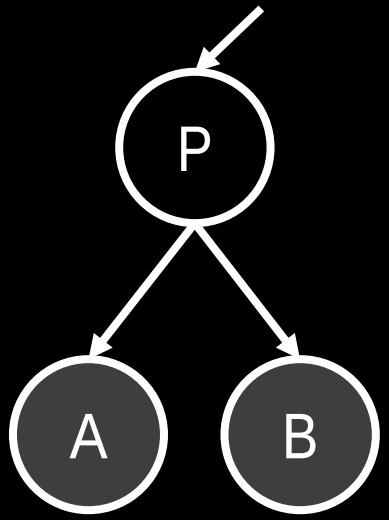
[Dunn+, Nature 2011]

今日のお話し

- 言語類型論は日本語系統論の最後の希望
- 言語普遍性に基づく自然な祖語の推定
- 単一起源仮説に基づく世界系統樹の推定

子孫を観測したとき、祖語について何が言える?

- 実際のところ、特徴量がどう変化するのかわからない
- 子孫から比較的近いはず
- 自然な言語であるはず



言語の自然さと普遍性

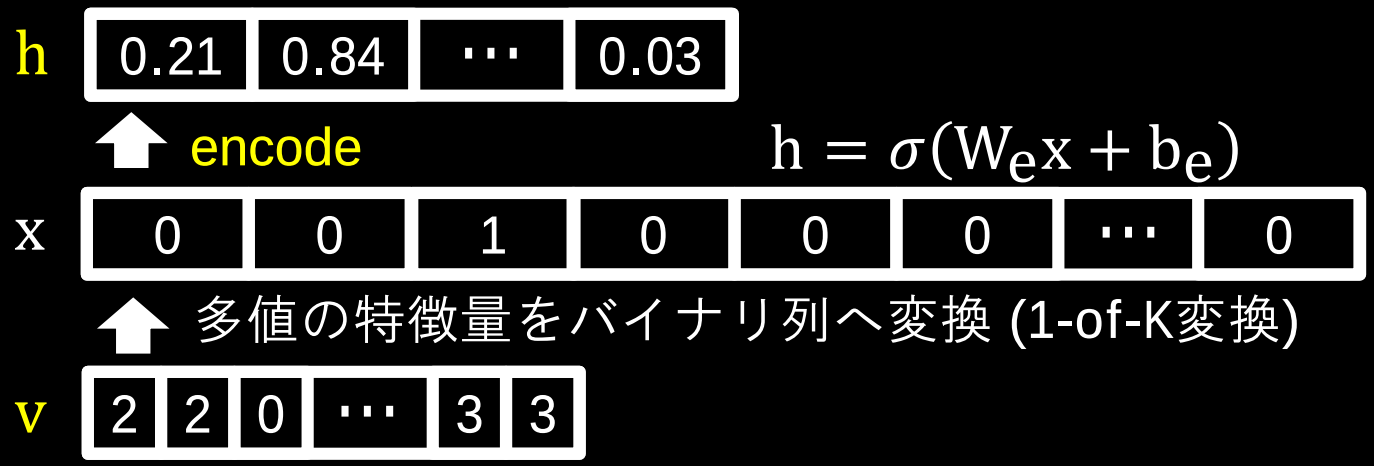
- 示唆的普遍性 (implicational universal)
  [Greenberg, 1963][Daumé III, ACL 2007]
  - OV ⊃ 後置詞型，VO ⊃ 前置詞型
  - 後置詞型 ⊃ 属格節-名詞の語順

言語の自然さと普遍性

- 示唆的普遍性 (implicational universal)
  [Greenberg, 1963][Daumé III, ACL 2007]
  - OV ⊃ 後置詞型，VO ⊃ 前置詞型
  - 後置詞型 ⊃ 属格節-名詞の語順

- 特徴量間に依存があるなら、行列をかけて変換すれば良いのでは?
- 言語 (特徴量ベクトル) の自然さを現代語から教師なしで学習し、祖語に適用すれば良いのでは?

提案手法: 類型論の連続空間表現



提案手法: 類型論の連続空間表現

自然さ判定

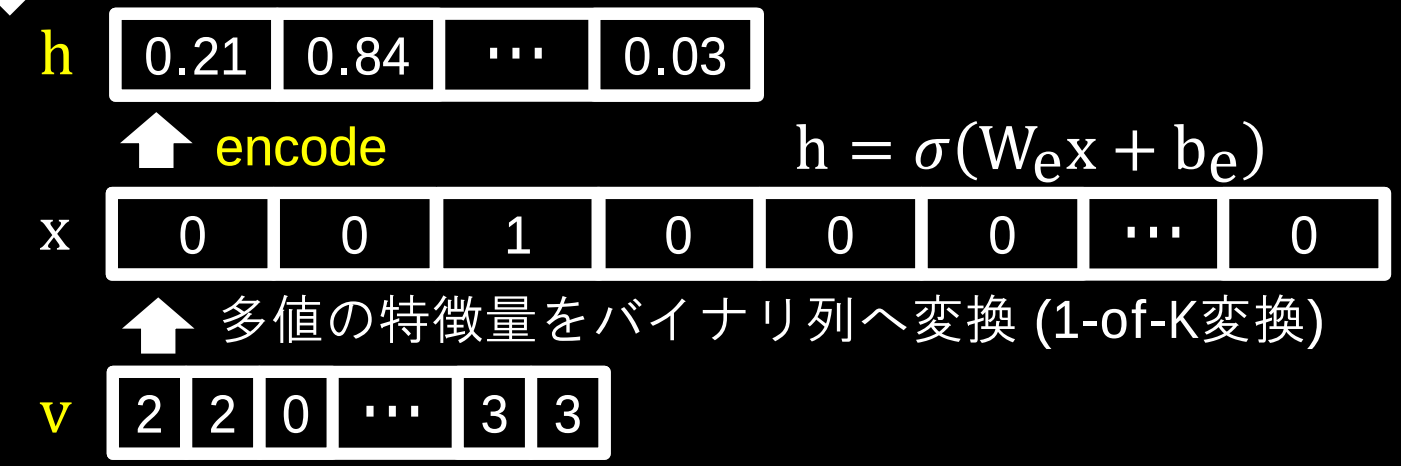
$$P(x) = \frac{\exp(f(h))}{\sum_{x'} \exp(f(h'))}$$

提案手法: 類型論の連続空間表現

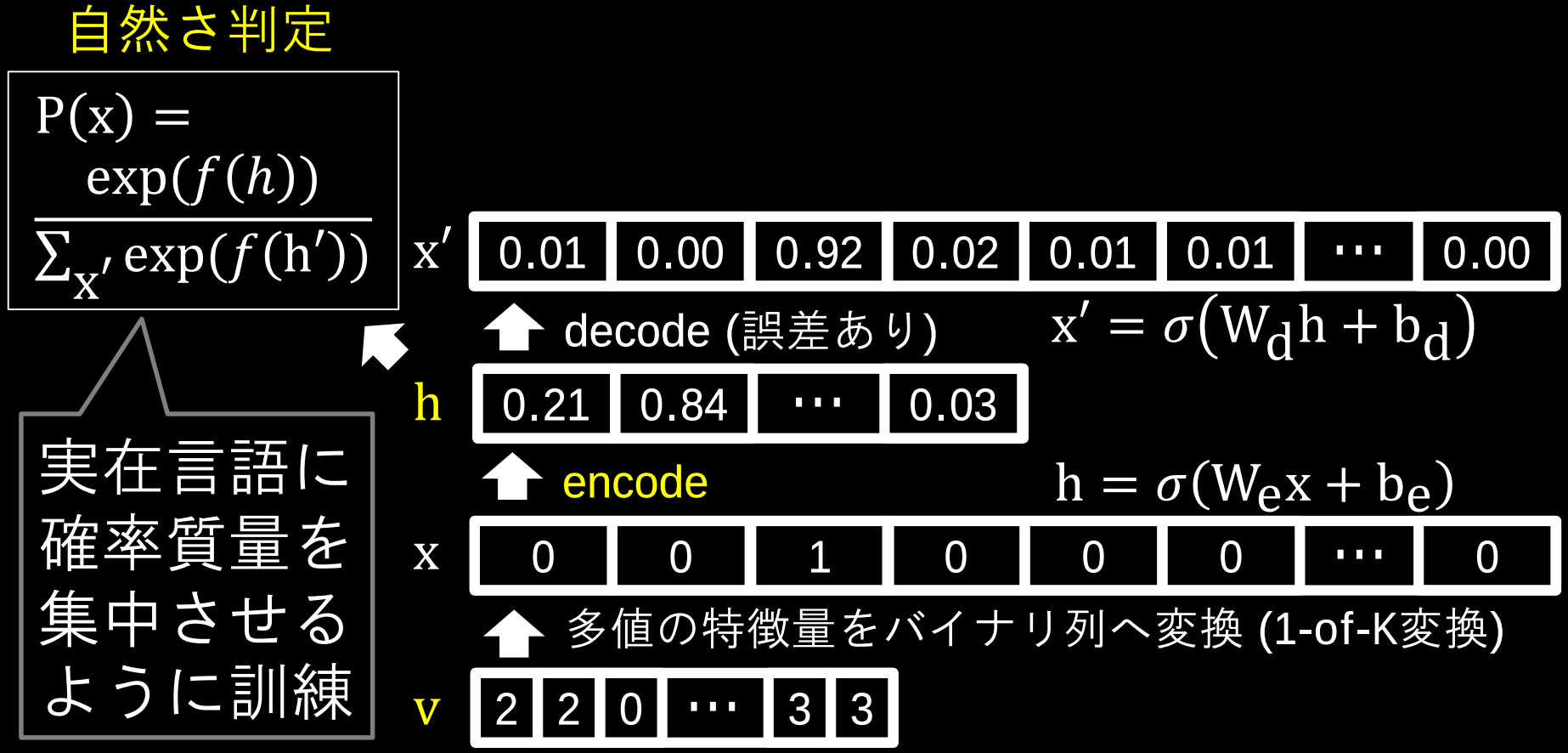
自然さ判定

$$P(x) = \frac{\exp(f(h))}{\sum_{x'} \exp(f(h'))}$$

実在言語に確率質量を集中させるように訓練



提案手法: 類型論の連続空間表現



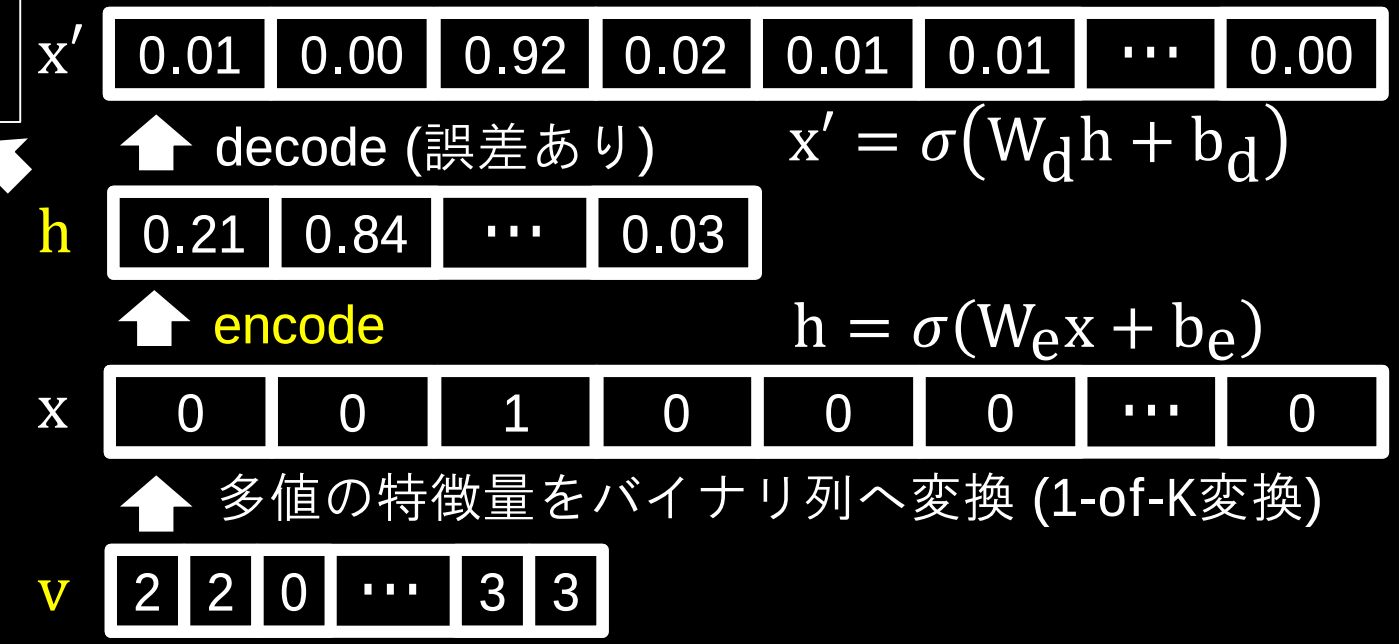
提案手法: 類型論の連続空間表現

連続空間表現が元の情報を保つように訓練

自然さ判定

$$P(x) = \frac{\exp(f(h))}{\sum_{x'} \exp(f(h'))}$$

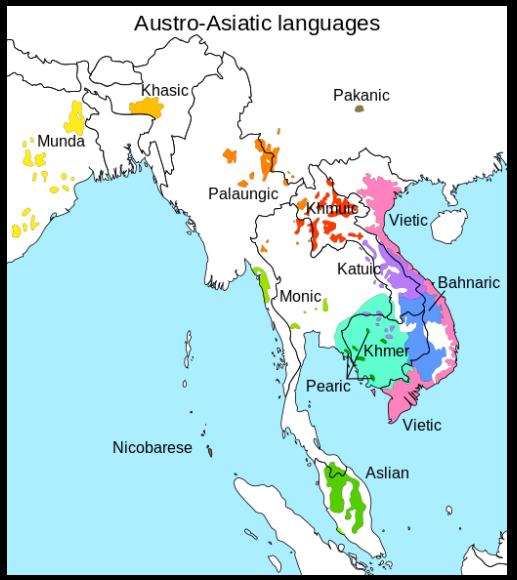
実在言語に確率質量を集中させるように訓練



提案手法: 類型論の連続空間表現

極端な例: オーストロアジア語族のムンダ諸語とモン・クメール諸語

	ムンダ	モン・クメール
文法	統合的	分析的
語順	主辞後置	主辞前置
	OV	VO
	後置詞	前置詞
接辞	接頭辞/接中辞, 接尾辞	接頭辞/接中辞 孤立的

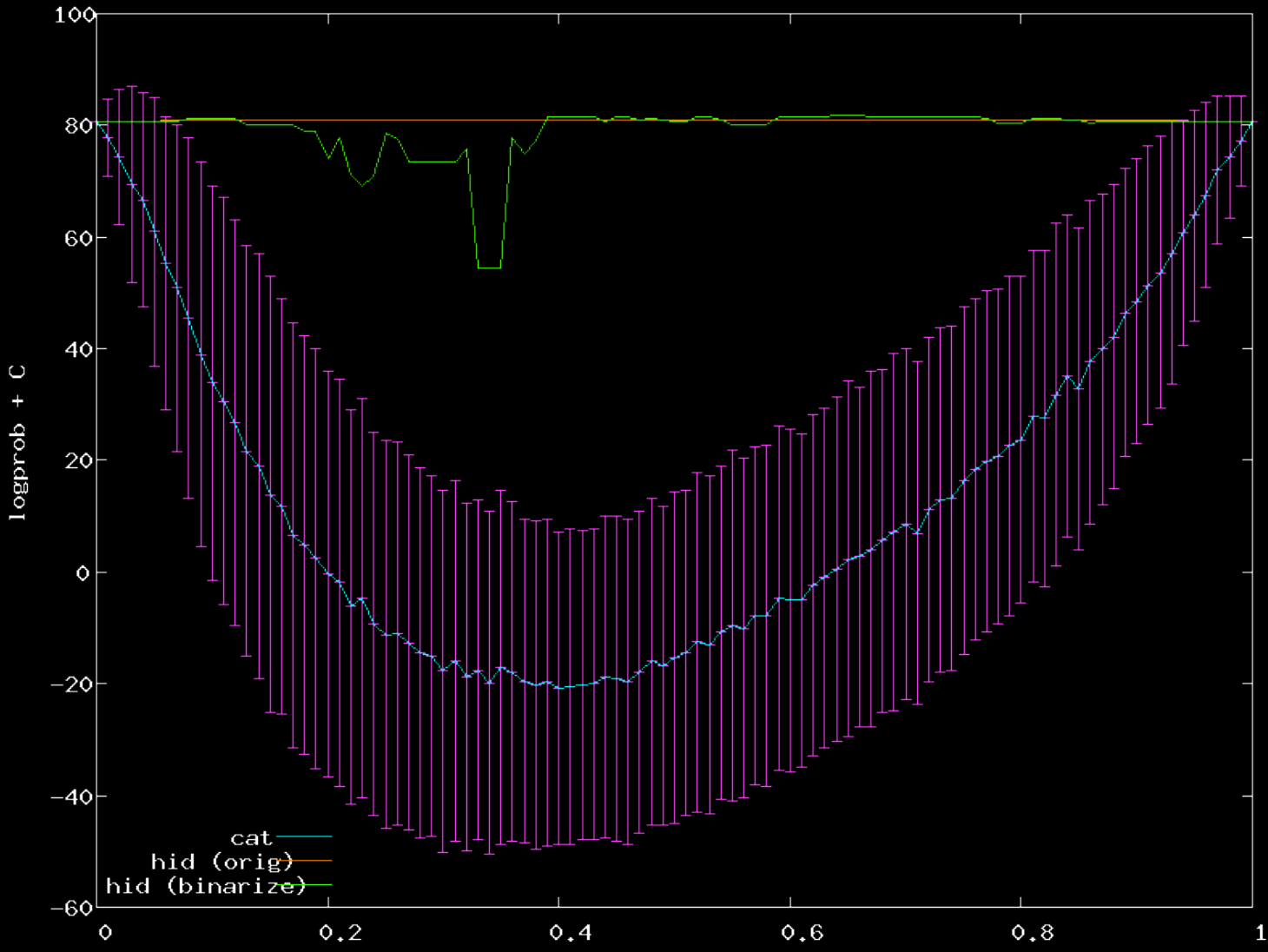


ソラ語	*ɑnin*	*dɔŋ- ɲɛn*	*dɑrəj*	*-ən*	*ə-*	*tiy*	*-ben*	*idsɨm*	*-tɛ*	*ted*
(ムンダ)	he/she	OBJ-	me	rice	-ART	INF-	give	–INF	want	-3PR	not

クメール語	*kŏət*	*ʔət*	*cɑŋ*	*ʔaoy*	*bay*	*kŋom*
	he/she	not	want	give	rice	me

[Donegan+, 2004]

言語としての自然さ

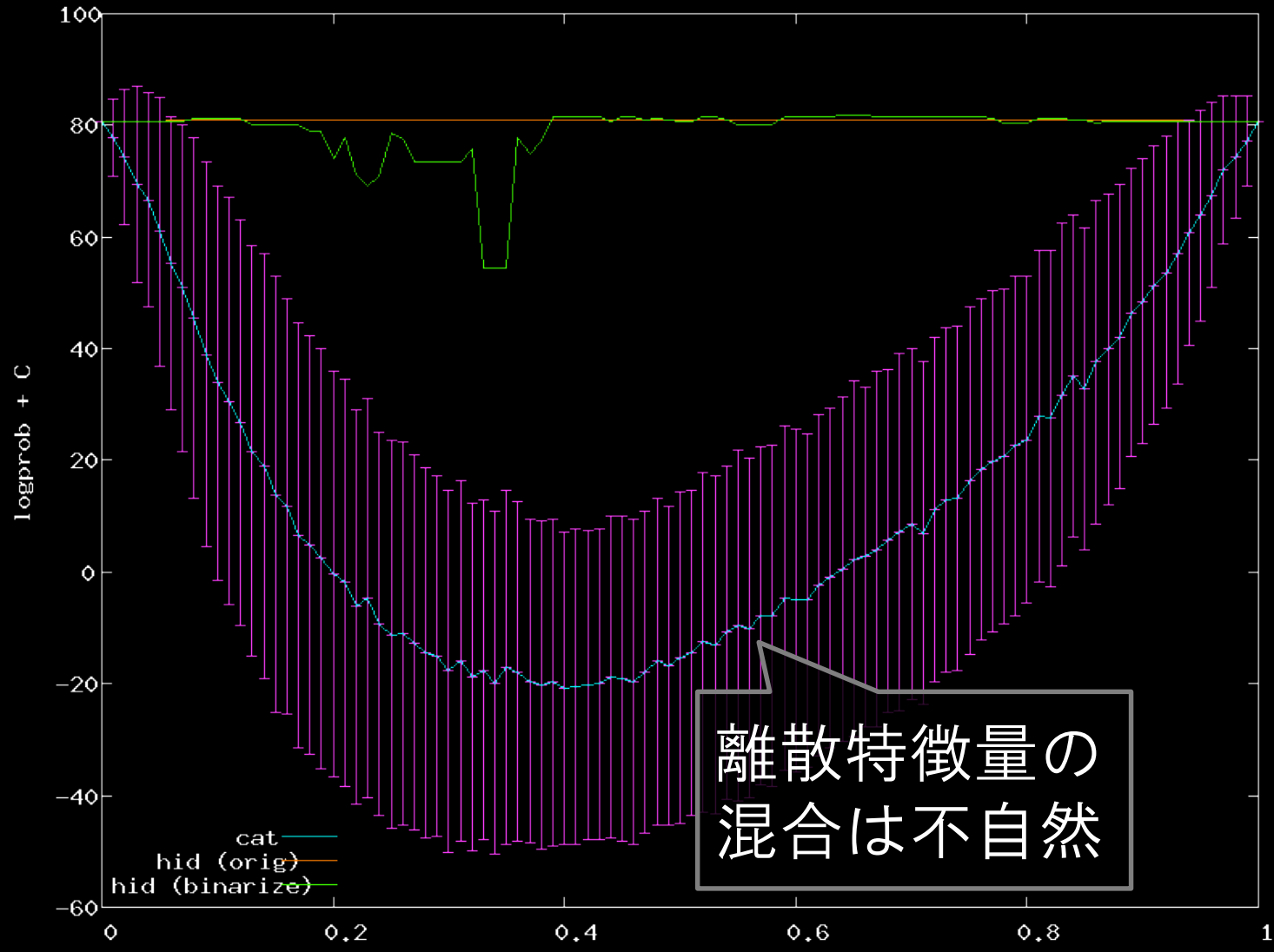


ムンダリ語
(ムンダ)

混合比

クメール語

言語としての自然さ

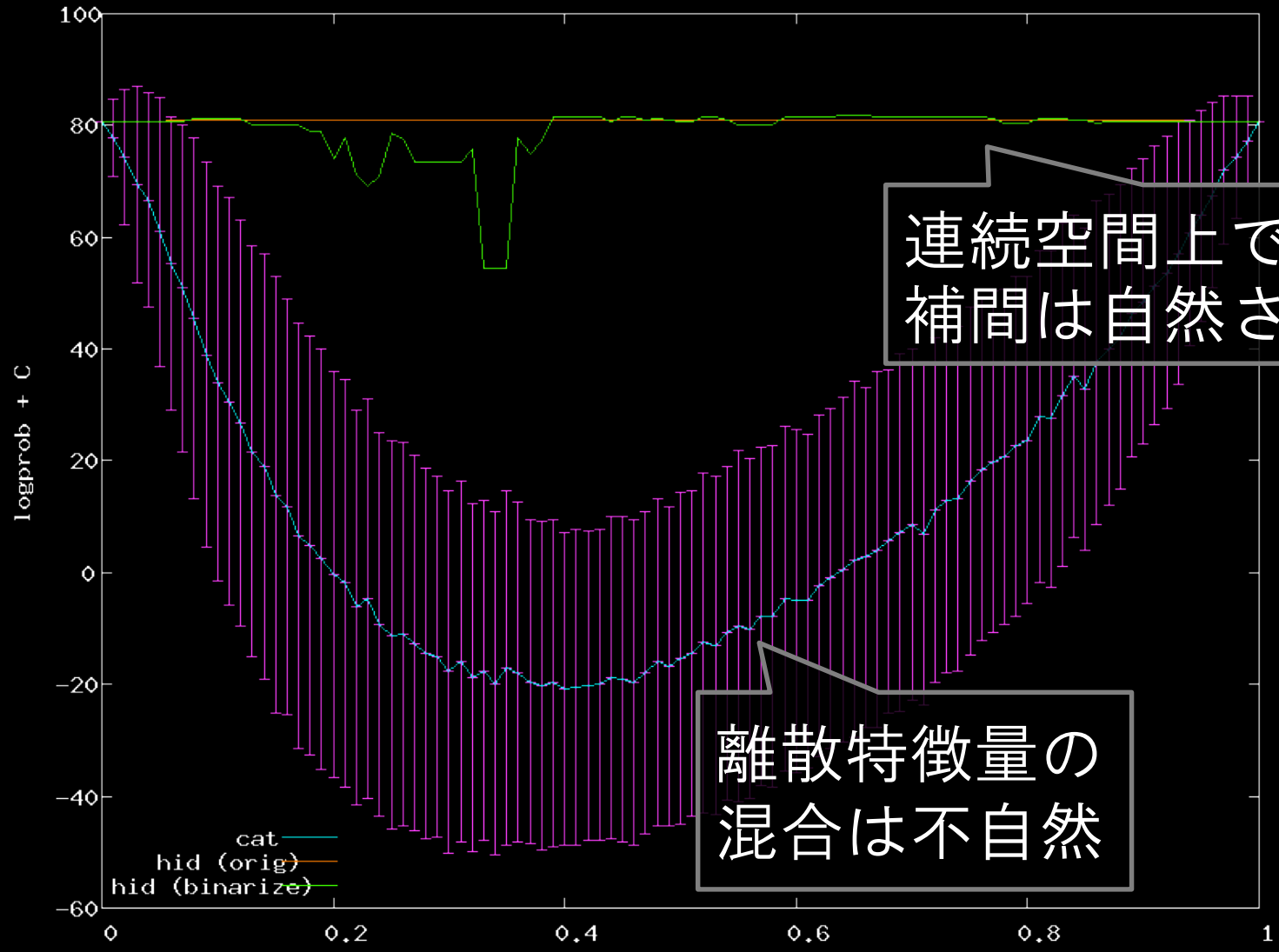


ムンダリ語
(ムンダ)

混合比

クメール語

言語としての自然さ



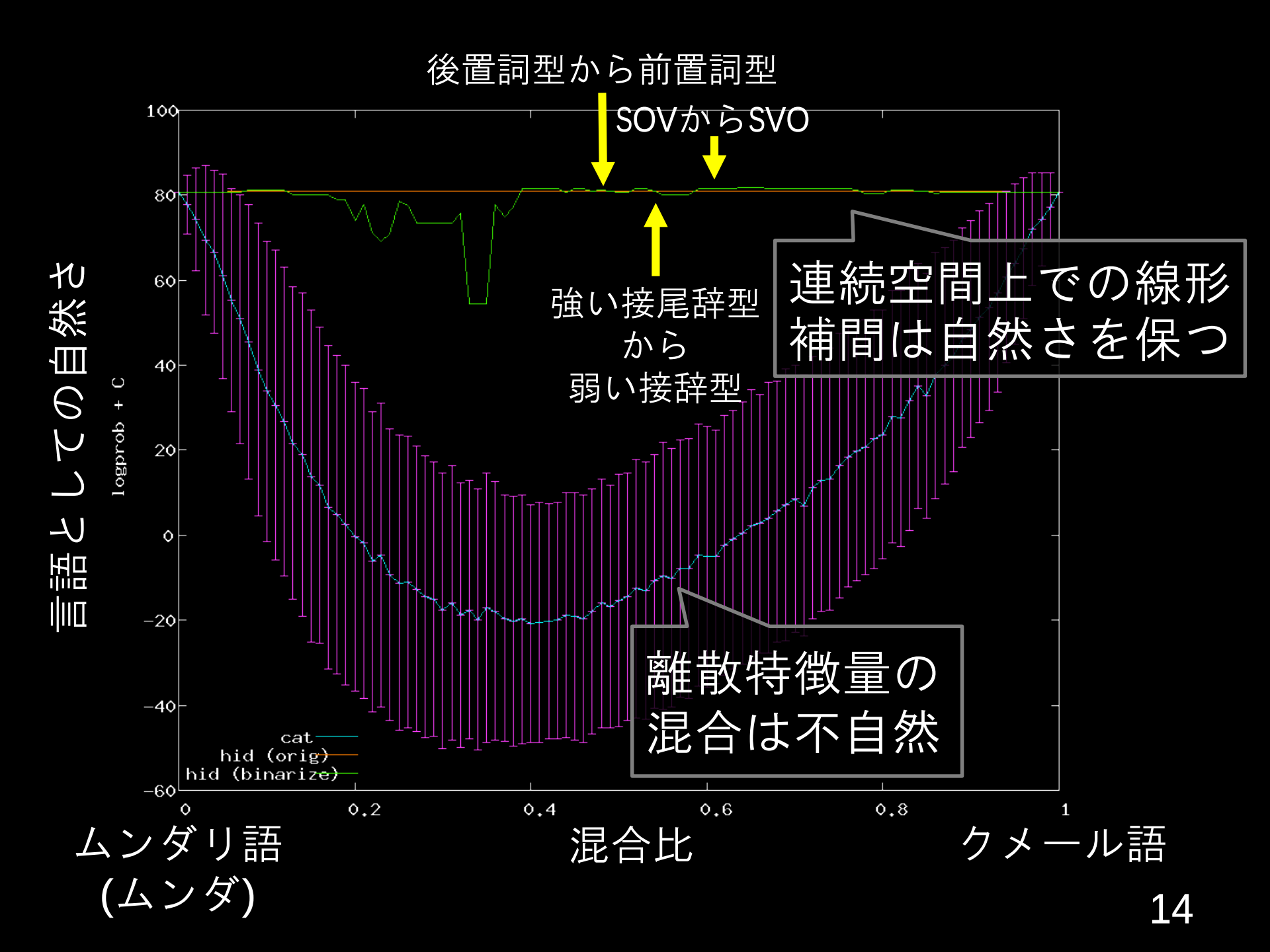
連続空間上での線形補間は自然さを保つ

離散特徴量の混合は不自然

ムンダリ語 (ムンダ)

混合比

クメール語


後置詞型から前置詞型
SOVからSVO
強い接尾辞型
から
弱い接辞型
連続空間上での線形
補間は自然さを保つ
離散特徴量の
混合は不自然
言語としての自然さ
logprob + C
100
80
60
40
20
0
−20
−40
−60
0
0.2
0.4
0.6
0.8
1
cat
hid (orig)
hid (binarize)
ムンダリ語
(ムンダ)
混合比
クメール語

今日のお話し

- 言語類型論は日本語系統論の最後の希望
- 言語普遍性に基づく自然な祖語の推定
- 単一起源仮説に基づく世界系統樹の推定

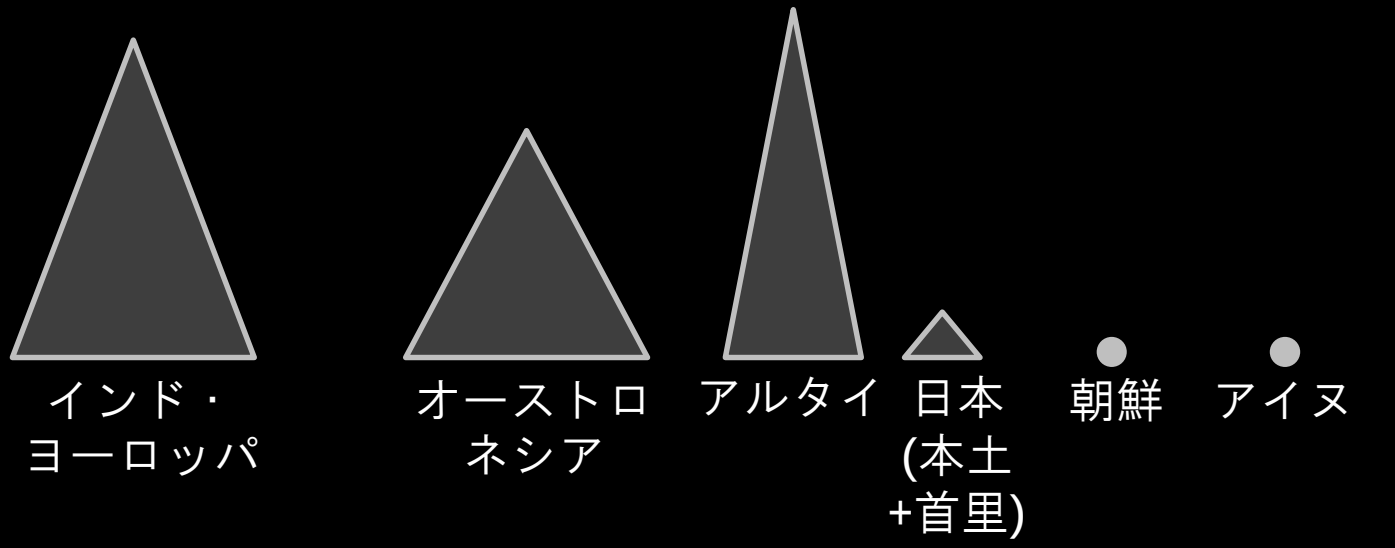
単一起源仮説に基づく世界系統樹の推定

- 単一起源 (monogenesis) 仮説
  - 世界の現代の諸言語は単一の祖語に由来?
- 類型論の特徴量のように長期的変化なら世界系統樹も推定可能?
- 注意
  - とりあえずやってみただけ
  - 検証しないといけない仮定・使っていない手がかりが多い
  - 現在の結果が実際の進化の過程を示しているかは極めて怪しい

特徴量ベクトルは系統を自然に反映しているか?

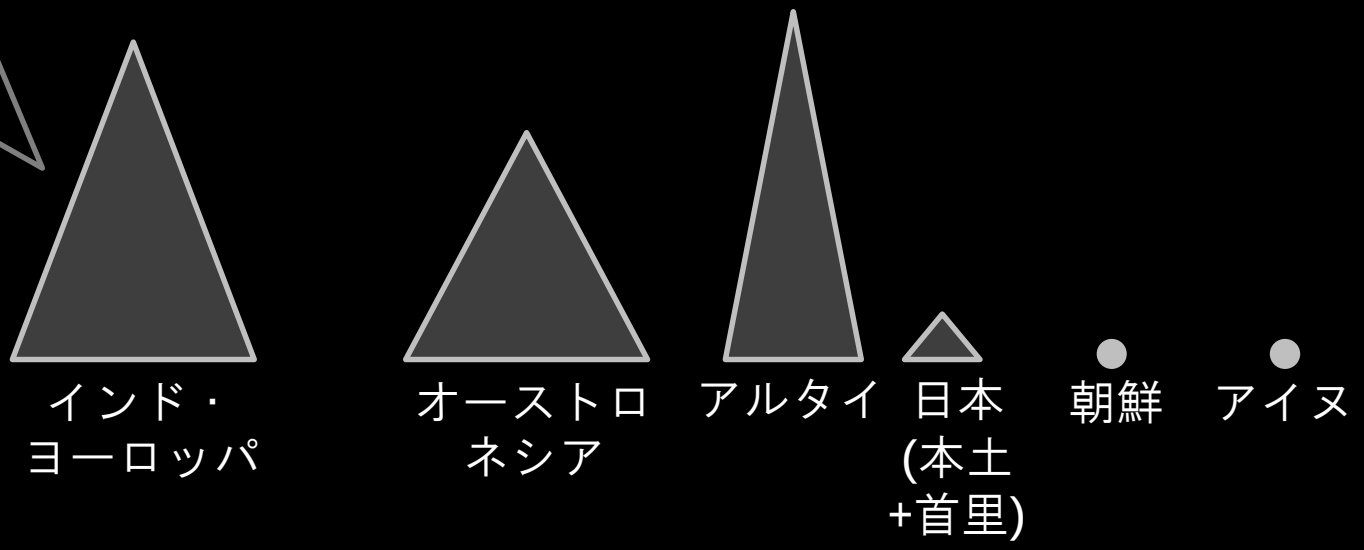
- 必ずしもそうではない
  - 言語連合: 系統に反する特徴量の変化
    [Trubetzkoy, 1923][Aikhenvald+, 2001][Daumé III, NAACL 2009]
- 提案手法: 既知の系統樹を教師に使う
  - 系統上の安定性を学習
  - 世界中の語族を使うことで、スパース性の緩和を期待

連続空間上での世界系統樹の推定

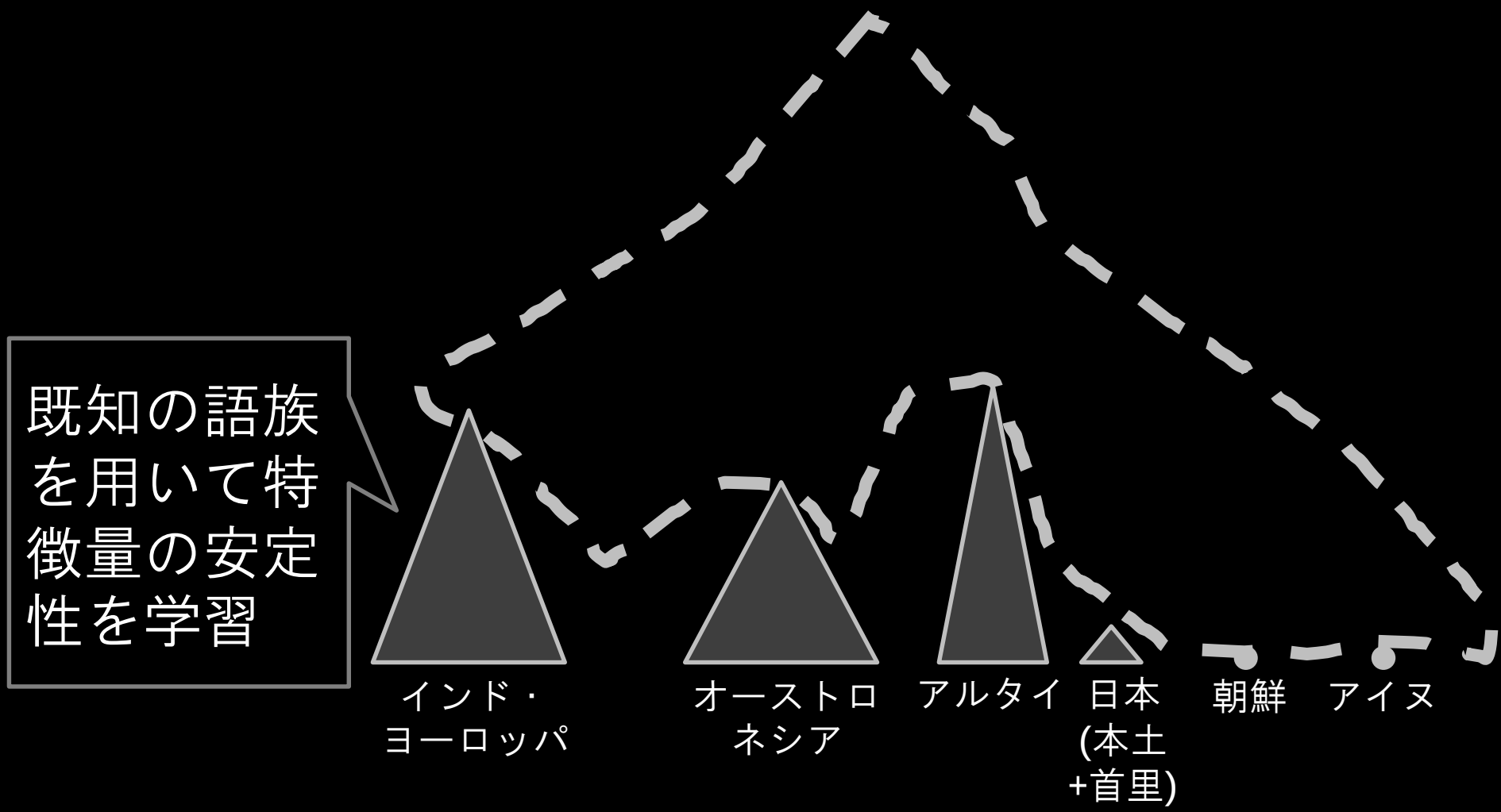


連続空間上での世界系統樹の推定

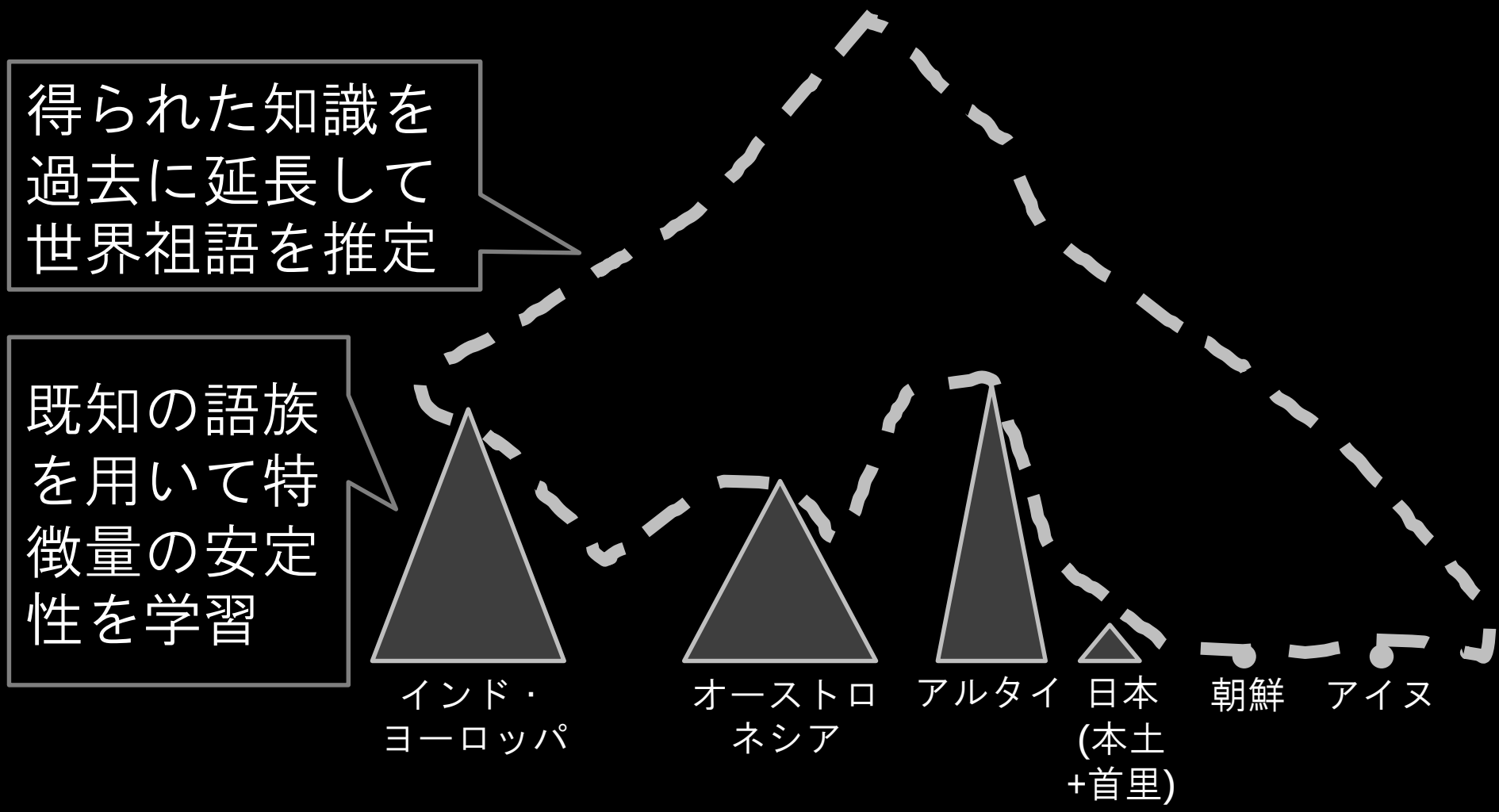
既知の語族を用いて特徴量の安定性を学習



連続空間上での世界系統樹の推定



連続空間上での世界系統樹の推定



祖語の自然さと特徴量の安定性を考慮した世界の系統樹

連続空間上で安定性を考慮した日本語と他の現代語との距離

安定性を考慮した距離			離散特徴量上の距離 (log不一致率)		
1.	jpn	76	1. jpn		0.0
2.	ryu Japonic	-33	2. kxv	Dravidian	-0.394
3.	khk Altaic->Mon.	-198	3. grt	ST->Tib.Bur.	-0.397
4.	lep ST->Tib.-Bur.	-202	4. ggo	Dravidian	-0.403
5.	chv Altaic->Tur.	-209	5. lez	NC->EC->Legzi.	-0.409
6.	mvf Altaic->Mon.	-213	6. chv	Altaic->Tur.	-0.431
7.	bxm Altaic->Mon.	-217	7. huu	Witotoan	-0.436
8.	der ST->Tib.-Bur.	-221	8. khk	Altaic->Mon.	-0.453
9.	uum Altaic->Tur.	-228	9. ryu	Japinic	-0.460
10.	huu Witotoan	-229	10. mal	Dravidian	-0.461
159.	kor （孤立）	-281	26. kor	（孤立）	-0.515
399.	ain （孤立）	-390	61. ain	（孤立）	-0.576

まとめと今後の課題

- ヨーロッパ系の人が印欧語研究でフィーバーしている間に日本で類型論を研究すれば天下がとれるかも
- 課題は山積み
  - より良い欠損値推定
  - 変化の方向性
  - 変化の速度
  - 年代推定
  - 地理位置との統合的推論