クレオール形成に対する混合モデル

村脇 有吾
京都大学大学院情報学研究科
murawaki@i.kyoto-u.ac.jp

1 はじめに

諸言語の進化史は通常系統樹によって要約されるが、言語間の接触を考慮しない木モデルの限界は研究の初期段階から認識されてきた。本稿では、木モデルに従わない極端な言語現象として、クレオール形成を取り上げる。クレオールは、典型的には社会文化的に優勢な上層言語といくつかの基層言語の強い接触を通じて形成される。上層言語は語彙提供言語 (lexifier) ともよばれ、クレオールの語彙の大半が上層言語に由来する。にもかかわらず、興味深いことに、クレオールと語彙提供言語の文法は大きく異なることが多い。

クレオールの大半はヨーロッパによる植民地化の影響を受けた大西洋・インド洋沿岸に分布するが、日本語ともまったく無関係ではない。日本語ベースのクレオールとして、台湾宜蘭県のアタヤル人が話す宜蘭クレオールが知られている。また、ポリワーノフ以来の日本語混成言語説との関係において、クレオールや関連するピジンが注目される場合がある。

クレオール形成過程は論争の絶えない課題である。本稿は、Bakker ら [1, 4] にならい、定量的にこの課題に迫る。彼らの整理によると、クレオール形成に対する立場は (1) 上層説、(2) 基層説、(3) プール説、(4) 普遍説の 4 つに大別できる。最初の 2 つはそれぞれ語彙提供言語、基層言語の影響を重視する。プール説は語彙提供言語と基層言語からなるプールからの特徴量の選択を想定する。普遍説は、言語普遍的な構造再編がクレオールと特徴づけると考える。これらの仮説を検証するために、Bakker らは NeighborNet [2] という距離に基づくクラスタリング手法を類型論データに適用した。その結果、図 1 のように、クレオールがクラスタを形成し、語彙提供言語、基層言語、その他の非クレオールから区別された。彼らはこの結果を普遍説を支持するものと解釈している。

しかし、NeighborNet はクレオール形成のモデルとして適切ではない。NeighborNet は、木としての矛盾を網状に表現するものの、基本的には木モデルである。複数の言語の関与が想定されるクレオール形成には、図 2 に示すように正反対のモデルである混合モデルがよりふさわしい。そこで、本稿では、混合モデルをクレオール・非クレオールの類型論データに適用し、上述の仮説群を検証する。

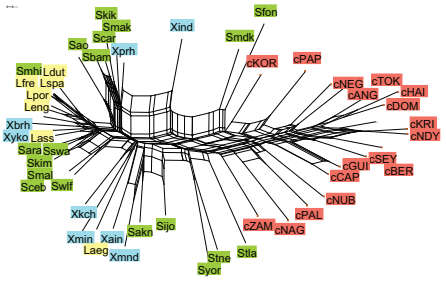
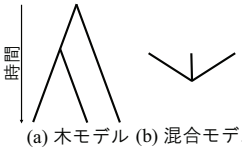


図 1: NeighborNet 分析 (赤: クレオール、黄: 語彙提供言語、緑: 基層言語、青: その他の非クレオール) により得られたクレオールのクラスタ



(a) 木モデル (b) 混合モデル

図 2: 木モデルと混合モデルの比較

2 クレオール形成の混合モデル

2.1 基本的な考え方

混合モデルでは、クレオールは (1) 語彙提供言語、(2) 基層言語、(3) 再編器の 3 つからなる混合元の確率的混合であると仮定する。この仮定のもとでの興味の中心は、それらの混合比である。このモデルが通常の混合モデルと異なる点は、混合先 (クレオール) だけでなく、語彙提供言語と基層言語という混合元も観測されていることである。残る再編器は推論が必要であり、その結果も関心の対象である。

このモデルは上述の仮説群を検証できるように設計されている。推論された混合比において、もし語彙提供言語あるいは基層言語が圧倒的であれば、それぞれ上層説、基層説が支持される。同様に、語彙提供言語と基層言語がある程度の割合を占め、再編器の割合が小さければ、プール説が支持される。再編器には注意が必要で、その割合の大きさはただちに普遍説を含意しない。再編器は他のいずれによっても説明されない特徴量の寄せ集めであり、普遍説が成り立つには、再編器の特徴量群にある種の一貫性がなければならない。

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

2.2 Mono モデル

具体的な混合モデルとして、Mono と Fact の 2 種類を導入する。各言語は J 個の多値特徴量の列で表現される。クレオール i の j 番目の特徴量 $x_{i,j}$ について、潜在割り当て変数 $z_{i,j} \in (\mathrm{L}, \mathrm{S}, \mathrm{R})$ は、$x_{i,j}$ が語彙提供言語 (L)、基層言語 (S)、再編器 (R) のいずれに由来するかを決める。ここで、クレオール i に対応する語彙提供言語、基層言語はあらかじめ与えられている。もし、混合元が語彙提供言語であれば、クレオール i はその特徴量の値 $y_{i,j,\mathrm{L}}$ を複製する (δ 関数により確率分布として扱う)。基層言語の場合 ($y_{i,j,\mathrm{S}}$) も同様である。残る再編器は多値分布の列であり、それらの事前分布は Dirichlet 分布である。

割り当て変数 $z_{i,j}$ は、混合比 $\theta_i = (\theta_{i,\mathrm{L}}, \theta_{i,\mathrm{S}}, \theta_{i,\mathrm{R}})$ をパラメータとする多値分布から生成される。簡単のため基層言語を 1 個としているが、複数個の場合に拡張するのは難しくない。多値分布には Dirichlet 事前分布を置く。

まとめると、Mono の生成過程は以下の通りである。

1. 再編器の特徴量の各型 $j \in \{1, \cdots, J\}$ について:
   1.1 対称 Dirichlet 分布から多値分布を生成: $\phi_j \sim \mathrm{Dir}(\beta)$
2. 各クレオール $i \in \{1, \cdots, N\}$ について:
   2.1 対称 Dirichlet 分布から混合比を生成: $\theta_i \sim \mathrm{Dir}(\alpha)$
   2.2 特徴量の各型 $j \in \{1, \cdots, J\}$ について:
      i. 割り当てを生成: $z_{i,j} \sim \mathrm{Categorical}(\theta_i)$
      ii. 以下の確率に従い特徴量の値を生成

$$x_{i,j} \sim \begin{cases} \delta(y_{i,j,\mathrm{L}}) & \text{if } z_{i,j} = \mathrm{L} \\ \delta(y_{i,j,\mathrm{S}}) & \text{if } z_{i,j} = \mathrm{S} \\ \mathrm{Categorical}(\phi_j) & \text{if } z_{i,j} = \mathrm{R} \end{cases}$$

共役性を利用し、ϕ_j と θ_i は積分消去する。残る $z_{i,j}$ は Gibbs サンプリングにより推論する。ここで、$z_{i,j}$ の条件付き確率は以下に比例する。

$$\begin{cases} \left(\alpha + c_{i,\mathrm{L}}^{-(i,j)}\right) \mathrm{I}(x_{i,j} = y_{i,j,\mathrm{L}}) & \text{if } z_{i,j} = \mathrm{L} \\ \left(\alpha + c_{i,\mathrm{S}}^{-(i,j)}\right) \mathrm{I}(x_{i,j} = y_{i,j,\mathrm{S}}) & \text{if } z_{i,j} = \mathrm{S} \\ \left(\alpha + c_{i,\mathrm{R}}^{-(i,j)}\right) \frac{\beta + c_{\mathrm{R},j,x_{i,j}}^{-(i,j)}}{B_j + c_{\mathrm{R},j,*}^{-(i,j)}} & \text{if } z_{i,j} = \mathrm{R} \end{cases} \quad (1)$$

ここで、$B_j = \sum \beta$、$c_{i,k}^{-(i,j)}$ はクレオール i の割り当て変数のうち値が k であるもの ($z_{i,j}$ を除く) の数、$c_{\mathrm{R},j,l}^{-(i,j)}$ は再編器に由来し、特徴量の型が j で値が l である特徴量 ($x_{i,j}$ を除く) の数を表す。Dirichlet 分布の α および β は不動点反復により求める。

2.3 Fact モデル

Fact は Mono の拡張であり、クレオールごとの混合比に加えて、特徴量ごとの選択選好を考慮する。

モデル		混合元 L	混合元 S	混合元 R
Mono		5.6%	5.6%	88.8%
Fact	全体	11.8%	5.8%	82.4%
	特徴量因子	15.1%	7.1%	77.8%
	クレオール因子	22.4%	21.9%	55.7%

表 1: 割り当て変数 (混合比) の推論結果

これを実現するために、混合比を対数空間において特徴量因子とクレオール因子に分解する [5]。すなわち、Fact において、クレオール i の j 番目の特徴量の混合比 $\theta_{i,j} = (\theta_{i,j,\mathrm{L}}, \theta_{i,j,\mathrm{S}}, \theta_{i,j,\mathrm{R}})$ は以下の通りである。

$$\theta_{i,j,k} = \frac{\exp(m_{j,k} + n_{i,k})}{\sum_k \exp(m_{j,k} + n_{i,k})} \quad (2)$$

ここで、$m_{j,k}$ は、特徴量の型 j の因子、$n_{i,k}$ はクレオール i の因子である。$m_{j,k}$ と $n_{i,k}$ の両者に対して Laplace 分布を事前分布とし、極端な値を抑制する。まとめると、Fact の生成過程は以下の通りである:

1. 再編器の特徴量の各型 $j \in \{1, \cdots, J\}$ について:
   1.1 $\phi_j \sim \mathrm{Dir}(\beta)$
   1.2 各混合元 $k \in \{\mathrm{L}, \mathrm{S}, \mathrm{R}\}$ について:
      i. $m_{j,k} \sim \mathrm{Laplace}(0, \gamma)$
2. 各クレオール $i \in \{1, \cdots, N\}$ について:
   2.1 各混合元 $k \in \{\mathrm{L}, \mathrm{S}, \mathrm{R}\}$ について:
      i. $n_{i,k} \sim \mathrm{Laplace}(0, \gamma)$
   2.2 各特徴量の各型 $j \in \{1, \cdots, J\}$ について:
      i. 式 (2) により、$m_{j,k}$ と $n_{i,k}$ を正規化して $\theta_{i,j}$ を得る
      ii. $z_{i,j} \sim \mathrm{Categorical}(\theta_{i,j})$
      iii. Mono と同様に $x_{i,j}$ を生成

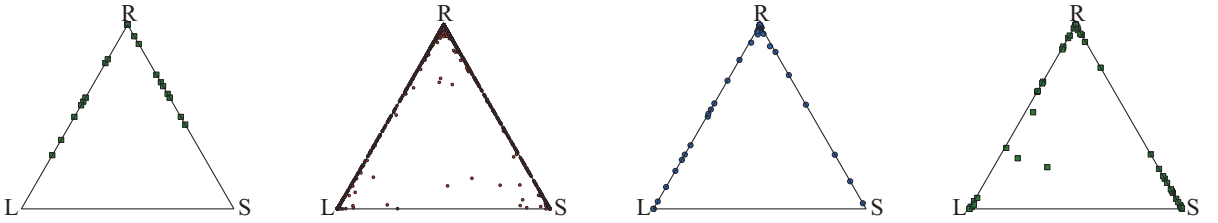
ϕ_j は積分消去できるが、$\theta_{i,j}$ に共役性は成り立たない。そこで、正規分布を提案分布とする Metropolis アルゴリズムにより $m_{j,k}$ と $n_{i,k}$ をサンプリングする。$z_{i,j}$ の Gibbs サンプリングは、式 (1) の第 1 項 $\alpha + c_{i,k}^{-(i,j)}$ を $\theta_{i,j,k}$ で置き換えるほかは Mono と同じである。また、$\gamma = 10$ とする。

3 実験

3.1 データと前処理

クレオールの類型論の情報源として、Atlas of Pidgin and Creole Language Structures (APiCS) [8] を用いた。APiCS は World Atlas of Language Structures (WALS) [6] のピジン・クレオール版といえ、文献 [1] のデータセットより大きい。ただし、APiCS はどの言語がクレオールかを明示しない (実際に論争がある)。そこで、本稿では、APiCS の社会言語学的特徴量 Ongoing creolization of pidgins の値が Not applicable (because the language is not a pidgin) あるいは Widespread であるものをひとまずクレオールとみなした。

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.



(a) Mono (b) Fact 全体 (c) Fact 特徴量因子 (d) Fact クレオール因子

図 3: 混合比の単体への写像

割合	特徴量の型	特徴量の値	日本語
90.3%	Numeral Classifiers	Absent	×
85.9%	Order of Numeral and Noun	Numeral-Noun	○
84.7%	Negative Morphemes	Negative particle	×
84.5%	Inclusive/Exclusive Distinction in Independent Pronouns	No inclusive/exclusive	○
81.9%	Predicative Possession	'Have'	×
81.8%	Applicative Constructions	No applicative construction	○
74.3%	Negative Indefinite Pronouns and Predicate Negation	Predicate negation also present	○
73.8%	Gender Distinctions in Independent Personal Pronouns	No gender distinctions	×
73.0%	Alignment of Case Marking of Full Noun Phrases	Neutral	×
70.5%	Occurrence of Nominal Plurality	All nouns, always optional	×

表 2: Fact において再編器に由来する特徴量上位 10 個

APiCS の類型論的特徴量は独自に定義されたものだが、130 種類のうち 48 種類は WALS の特徴量への対応が記述されている。これらの特徴量を用い、WALS からの非クレオールと APiCS からのクレオールを統合した。欠損値は R パッケージ missMDA [7] による多重対応分析で補完した。

提案手法では、各クレオールに対して語彙提供言語および基層言語を 1 つずつ事前に指定する必要がある。しかし、APiCS は、特に基層言語に関する記述が曖昧である (実際必ずしも明らかではない)。そこで、本稿では、いくつかの現代語をこれらの言語の代替として選んだ。この選択には再考の余地がある。

3.2 結果

表 1 に割り当て変数の推論結果を示す。これは混合比に相当する。ただし、Fact の特徴量因子については、$\tilde{\phi}_j = (\tilde{\phi}_{j,\mathrm{L}}, \tilde{\phi}_{j,\mathrm{S}}, \tilde{\phi}_{j,\mathrm{R}})$ (ここで $\tilde{\phi}_{j,k} = \frac{\exp(m_{j,k})}{\sum_k \exp(m_{j,k})}$) を用いた。同様にして、クレオール因子としては $\tilde{\theta}_i = (\tilde{\theta}_{j,\mathrm{L}}, \tilde{\theta}_{j,\mathrm{S}}, \tilde{\theta}_{j,\mathrm{R}})$ (ここで $\tilde{\theta}_{i,k} = \frac{\exp(n_{i,k})}{\sum_k \exp(n_{i,k})}$) を用いた。焼きなまし 5,000 反復ののち、100 反復ごとに 50 サンプルを集め、それらの算術平均を求めた。Mono、Fact (全体) のいずれについても、再編器に由来する特徴量が大半を占めた。語彙提供言語がそれに続き、基層言語の影響がもっとも小さい。この結果は、上層説、基層説、プール説への反証と解釈できる。

Mono と Fact (全体) を比較すると、大勢は同じだが、Fact は語彙提供言語の割合がやや多く、その分再編器の割合が少ない。因子に分解すると、よりなだらかな混合比となった。

図 3 に混合比の単体上への写像を示す。10,000 反復後の 1 サンプルを用いている。図 3(a) は Mono におけるクレオールの混合比 J 個を事後予測分布のパラメータ $\tilde{\theta}_i = (\frac{\alpha+c_{i,\mathrm{L}}}{Z}, \frac{\alpha+c_{i,\mathrm{S}}}{Z}, \frac{\alpha+c_{i,\mathrm{R}}}{Z})$ (ここで正規化項 $Z = \sum_k \alpha + c_{i,k}$) を用いて示している。図 3(b) は、Fact の $J \times N$ 個の混合比 $\theta_{i,j}$ を示している。Mono では再編器よりの部分に点が集中したが、Fact ではより広範囲に散らばっている。ただし、点は縁に集中し、また語彙提供言語と基層言語だけの混合はほぼ見られない。図 3(c-d) はそれぞれ J 個の特徴量因子、N 個のクレオール因子を示している (単体への写像方法は表 1 と同じ)。Mono と比較したときの特徴量因子の効果は図 3(d) で確認できる。Mono では再編器と語彙提供言語 (または基層言語) の中間に位置していたクレオールが、Mono では語彙提供言語 (または基層言語) 寄りの極端な値をとることが可能になっている。ここでも語彙提供言語と基層言語だけの混合は見られない。

表 2 に再編器に由来する特徴量上位 10 個を示す。表 1 と同様にして 50 サンプルを集め、それらの算術平均を求めた。ここで、特徴量の型と値 (j, l) の割合は、$|\{(i \,|\, x_{i,j} = l, z_{i,j} = \mathrm{R}\}| \,/\, N$ と定義される。つまり、クレオールの特徴量に対して、語彙提供言語と基層言語の影響を補正したものである。これらは (統計的) 普遍性とみなしうる。ただし、再編普遍性 (restructuring universal) であるかは明らかではない。この疑問に応えるには、言語普遍性を種類別に分析する必要がある。最右列は日本語の特徴量の値と一致するかを示している。この結果から、日本語は非クレオール的であり、近い過去にクレオール形成はなかったと推測できる。

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.

クレオールに顕著な特徴として、しばしば SVO 語順が挙げられる。実際、APiCS では、76 言語中 61 言語が排他的に SVO 語順で、他の語順も取り得る場合を含めると 71 言語におよぶ。しかし、この値の割合は 67.3%にとどまった。SVO が語彙提供言語の語順でもあることが理由の一つだが、同時にデータ表現の影響も考えられる。APiCS は、一部の特徴量について、「SVO かつ SVO」のように複数の値を持つことを認めているが、WALS はこれを認めていない。そのため、APiCS から WALS の特徴量に変換する際に、一部のクレオールには別の値 No dominant order が付与されており、結果として SVO の影響が過小評価されている。

3.3 議論

本稿で提示した混合モデルは、モデルを通してクレオール形成過程を理解するための最初の一歩に過ぎない。データの面では、基層言語の代替、欠損値、複数の値を持つ特徴量の扱いなどに改善の余地がある。モデルの面では、提案手法は混合元を 1 段階で混合したが、クレオール研究においては多段階におよぶ形成過程が議論されている。漸進説とよばれる仮説によると、クレオール形成はピジン化とクレオール化の少なくとも 2 段階に分けられる。ピジンは、共通言語を持たない話者同士がその場しのぎの対話を行う際に発生する言語で、この際に起きる文法の極端な単純化がピジン化とよばれる。クレオール化は、ピジンを子供が母語として獲得し、複雑な意思疎通が可能なほど文法が発達する現象を指す。この仮説に従うなら、ピジン化とクレオール化の 2 段階をモデル化する必要がある。また、クレオール形成後も上層言語の影響を受け続けると、脱クレオール化が起きるという説もある。さらには、ピジン自体にもいくつかの発達段階があり、そのいずれからもクレオールが発生し得るという説もある [9]。こうした様々な説も、モデルを通じて検証できるかもしれない。

最後に、木モデルを用いる従来の系統研究に対する本結果の影響を検討する。進化は変化を伴う由来 (descent with modification) ともよばれ、親から子への途切れない継承を仮定する。これに対し、本結果はいずれの言語でもないもの (再編器) が強い影響を及ぼす場合があることを示しており、進化の仮定に反している。しかし、本結果が示唆するように再編器に言語普遍性があれば、文献 [3] と同様の手法によって、木モデルに再編器を組み込むことは難しくない。そもそもクレオール形成が一般的な現象なのか、ヨーロッパ植民地主義が産んだ例外的現象なのかにも議論の余地があり、クレオール発生条件の研究 [10] の深化が期待される。

4 おわりに

本稿では、クレオール形成過程の定量的解明のための手法として混合モデルを提案した。混合モデルは、先行研究が用いた木モデルよりも、クレオール形成のモデルとしてふさわしい。実験により、語彙提供言語、基層言語のいずれでもない再編器がクレオール形成に大きな影響を及ぼしていること、および再編器がいくつかの統計的普遍性を持つことが示された。

近年の言語系統研究における統計的手法は、計算生物学の影響を強く受けている。実際、先行研究が用いた NeighborNet は、当該分野で開発されたソフトウェアである。しかし、従来研究はパッケージ化されたソフトウェアに依存してきたため、生物に完全に対応する現象が存在しない (もしくはまだモデル化されていない) 言語現象は手付かずのまま残される傾向にある。混合モデル自体も、集団遺伝学における DNA 解析において、LDA に似たモデルが必ずと言ってよいほど適用されている。しかし、もちろん、語彙提供言語と基層言語のような一部の混合元が観測されている場合を想定したソフトウェアは提供されておらず、本稿で独自に開発する必要があった。こうした部分に、言語処理研究者の参入の余地があると考えている。

謝辞 本研究は一部 JSPS 科研費 26730122 の助成を受けた。

参考文献

[1] Peter Bakker, Aymeric Daval-Markussen, Mikael Parkvall, and Ingo Plag. Creoles are typologically distinct from non-creoles. *Journal of Pidgin and Creole Languages*, Vol. 26, No. 1, pp. 5–42, 2011.

[2] David Bryant and Vincent Moulton. Neighbor-Net: An agglomerative method for the construction of phylogenetic networks. *Molecular Biology and Evolution*, Vol. 21, No. 2, pp. 255–265, 2004.

[3] Hal Daumé III. Non-parametric Bayesian areal linguistics. In *HLT-NAACL*, pp. 593–601, 2009.

[4] Aymeric Daval-Markussen and Peter Bakker. Explorations in creole research with phylogenetic tools. In *Proc. of LINGVIS & UNCLH*, pp. 89–97, 2012.

[5] Jacob Eisenstein, Amr Ahmed, and Eric P. Xing. Sparse additive generative models of text. In *Proc. of ICML*, pp. 1041–1048, 2011.

[6] Martin Haspelmath, Matthew Dryer, David Gil, and Bernard Comrie, editors. *The World Atlas of Language Structures*. Oxford University Press, 2005.

[7] Julie Josse, Marie Chavent, Benot Liquet, and François Husson. Handling missing values with regularized iterative multiple correspondence analysis. *Journal of Classification*, Vol. 29, No. 1, pp. 91–116, 2012.

[8] Susanne Maria Michaelis, Philippe Maurer, Martin Haspelmath, and Magnus Huber, editors. *APiCS Online*. Max Planck Institute for Evolutionary Anthropology, 2013.

[9] Peter Mühlhäusler. *Pidgin and Creole Linguistics: Expanded and revised Edition*. University of Westminster Press, 1997.

[10] Francesca Tria, Vito D.P. Servedio, Salikoko S. Mufwene, and Vittorio Loreto. Modeling the emergence of contact languages. *PLoS ONE*, Vol. 10, No. 4, p. e0120771, 04 2015.

Copyright(C) 2016 The Association for Natural Language Processing.
All Rights Reserved.