

カテゴリ形成課題を用いた二者間記号創発実験の メトロポリス・ヘイスティングス法に基づくダイナミクス検証 Dynamics validation based on the Metropolis-Hastings method of bilateral symbol emergence experiments using category formation tasks

奥村 亮太[†], 萩原 良信[†], 谷口 彰[†], 谷口 忠大[†]

Okumura Ryota, Hagiwara Yosinobu, Taniguchi Akira, Taniguchi Tadahiro

[†] 立命館大学

Ritsumeikan University

okumura.ryota@em.ci.ritsume.ac.jp

概要

エージェントがそれぞれ独立にカテゴリを形成し、形成したカテゴリに紐付けられたサインを確率的に受容と棄却を決定する名付けゲームによって共有するコミュニケーションを行う状況を考える。この際、受容確率がメトロポリス・ヘイスティングス法により決まる確率に従うなら、二者間の記号創発が分散的なベイズ推論と等価になるということが知られている。もし人間の記号創発もこのようなダイナミクスに従っていることが明らかになれば、記号創発に関して新しい洞察を得ることが出来る。そこで本研究では実際の人間に名付けゲームを行わせ、人間の受容・棄却行動を観測し、メトロポリス・ヘイスティングス法に基づく受容確率と比較することで、人間の記号論が前述の記号創発モデルによって説明可能かどうかを検証した。その結果、検定により、メトロポリス・ヘイスティングス法に基づく確率に依存せずに一定の確率で受け入れの判断をしているという帰無仮説が棄却され、人間はある程度前述のモデルで計算された受容確率に依存してサインの受容を行っていることが示唆された。

キーワード：記号創発，実験記号論，カテゴリ形成

1. はじめに

人間の記号や言語を用いたコミュニケーションは記号に関する創発システムである記号創発システムの上で成り立っていると考えられる [1]。人間のコミュニケーションのダイナミクスを解明するには、この記号創発システム全体を表す計算論モデルを作り、実際にその計算論モデルに人間が従って行動しているのかを検証することが重要であると考えられる。記号創発システムを表現する計算論モデルの研究には様々なものがある。Steels の名付けゲーム (naming game) を通じて遠隔に存在するロボット間で対象を表すサインを共

有させた先駆的な研究 [2] や Deep Learning に基づくマルチエージェントを用いた研究が多数ある [3,4]。これらの研究に対して、萩原らの各エージェントがそれぞれ独立にカテゴリを形成し、形成したカテゴリに紐付けられたサインを名付けゲームによるコミュニケーションによって共有させた研究がある [5]。

本研究では萩原らのモデルに着目する [5]。以降、このモデルを Inter-DM と呼ぶ。しかし、その名付けゲームのようなコミュニケーションを実際の人間が行なっているのかはわかっていない。このモデルでは名付けゲーム自体が、マルコフ連鎖モンテカルロ法的一种であるメトロポリス・ヘイスティングス法と等価であることが数学的に示されており、コミュニケーションに参加する全エージェントの感覚情報に基づくベイズ推論として記号創発が意味づけられる [5]。しかし、この理論での、各エージェントがメトロポリス・ヘイスティングス法の理論に基づいて計算される受容確率に従い他者のサインを受け入れるかどうかを決定する [6]、という仮定が、実際に人間において成立するかどうかは明らかでない。

そこで我々は、被験者実験を行い、全体として推論した受容確率が高い方が、実際に名付けを受容した割合が高くなる傾向が観察された [7]。具体的には、Inter-DM の名付けゲームと同等の実験を設計した。Inter-DM の理論に基づいて推論した名付けに対する人間の受容確率と、実際に名付けを受け入れた割合をヒストグラムにより可視化した。その結果、全体として推論した受容確率が高い方が、実際に名付けを受容した割合が高い傾向があることを確認した。しかし、その名付けゲームの理論が人間の行動を説明できているかどうかは検証できていなかった。

そこで本論文では、モデルにより推論された受容確率が高い場合に、人間がより高い確率で名付けを採

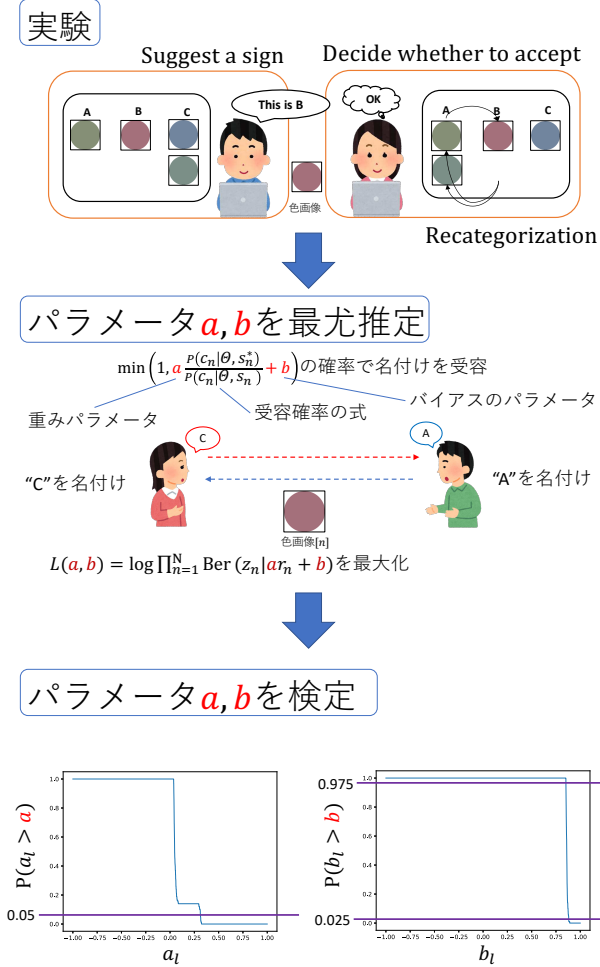


図 1 本研究の全体の流れの図。

択するかを統計的な手法により解析する。このためにモデルが捉えきれない要素も考慮するために、推論した受容確率 r に重み a を掛けバイアス b を足した $p = ar + b$ により人間が受容・棄却を判断しているという理論を考える。その上で、人間がメトロポリス・ヘイスティングス法によって計算される受容確率を用いて名付けを受容しているかを検証する。図 1 に本研究全体の流れを示す。

本稿では、2 章で分析方法について述べ、3 章で実験方法について述べ、4 章で結果について述べ、5 章でまとめる。

2. 準備

2.1 モデルベースデータ解析

本論文では萩原らの Inter-DM を改変した Inter-GM を用いる。Inter-DM はエージェントが Dirichlet mixture を用いて個々にカテゴリを形成し、コミュニケーションによって形成したカテゴリに紐付けられた

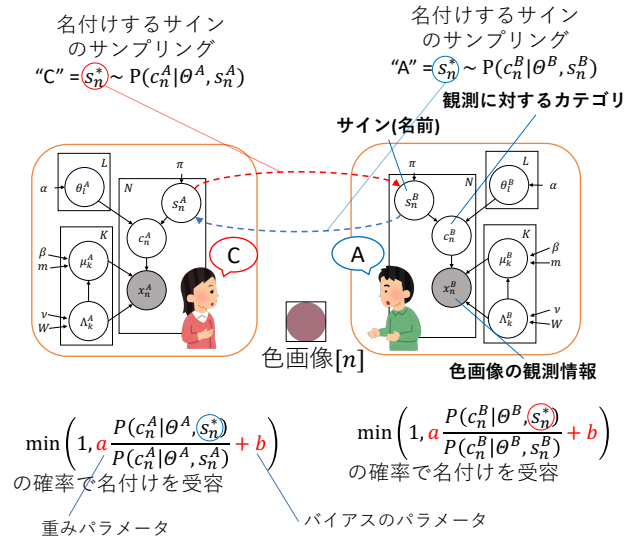


図 2 本研究の被験者実験と名付けゲームを対応させた概要図。 x_n^A, x_n^B は観測の色画像である。 c_n^A, c_n^B は被験者の色画像の分類結果である。被験者の色画像への名付けがサイン s_n のサンプリングであり、メトロポリス・ヘイスティングス法によりサインを共有する。赤と青の点線は名付けによるサインの伝達を表す。

サインを共有する。Inter-DM の Dirichlet mixture によってカテゴリ形成していた部分を単純に Gaussian mixture に置き換えたものが Inter-GM である。

図 2 は被験者実験と名付けゲームを対応させた概要図である。これは 2 体のエージェントによるそれぞれのカテゴリ化とサインの推論を統合した確率的生成モデルになっている（詳しくは [5] 参照）。 x_n が観測情報であり、 c_n は x_n が分類されるカテゴリである。 s_n は x_n に対するサインである。 s_n^* は提示された x_n に対して相手にサンプリングしたサインである。サイン s_n を推論する際にエージェント A, エージェント B が交互にサイン s_n をサンプリングし合い、相手のサインをメトロポリス・ヘイスティングス法の受容確率 $\min \left(1, \frac{P(c_n | \Theta, s_n^*)}{P(c_n | \Theta, s_n)} \right)$ に従って受け入れる。 Θ は図 2 の θ_l を L 個まとめたものである。この推論仮定そのものが、確率的な名付けゲームに対応しており、記号創発システムの計算論モデルとみなすことができる。

人間がメトロポリス・ヘイスティングス法の受容確率に従って相手の提案を受容しているのかを調べるため、人間の受容確率と実際の受容割合を評価指標として用いる。被験者の分類結果 c とその名前 s を使用し

て、パラメータ Θ を推論し、メトロポリスヘイスティングス法の受容確率 $\min\left(1, \frac{P(c_n|\Theta, s_n^*)}{P(c_n|\Theta, s_n)}\right)$ を計算し、その値を人間の受容確率とする。 n はデータの番号であり、 s_n^* は相手の提案である。

2.2 検定

2.2.1 パラメータ推定

推論した受容確率よりも多様な要因でずれることを説明するために重みを掛けて、バイアスのパラメータを足して判断していると考え、それらのパラメータを推定する。 z_n は n 回目の名付けに関して相手の名付けを受け入れたかどうかの変数であり、受け入れた場合は 1 をとり、拒否した場合は 0 をとる。 r_n は 2.1 で述べた方法で推論した被験者の受容確率である。名付けを受け入れるという事象が確率 r_n で起こると考えれば、これらを次のようにベルヌーイ分布に当てはめ、全体の名付けの数 N だけ掛け合わせることで尤度を計算することができる。

$$L = \prod_{n=1}^N \text{Ber}(z_n|r_n)$$

検定では図 2 のようにパラメータ a, b を含めた受容確率を使用する。パラメータ a, b を追加し、対数をとったものを対数尤度とし、

$$L(a, b) = \log \prod_{n=1}^N \text{Ber}(z_n|ar_n + b)$$

とする。このパラメータ a, b を最尤推定により決定する。解析的に解くのは困難であったため、勾配法を用いて a, b の最尤推定を行った。

2.2.2 サンプルに基づく検定

人間が推論された受容確率をある程度使用しているのか、一定の確率で受容しているだけなのかを調べるために、2.2.1 で推定したパラメータ a, b について検定する。帰無仮説 H_0 と対立仮説 H_1 は以下である。

- H_0 : 使用している受容確率が一定である。
- H_1 : 使用している受容確率が一定でない、つまり推論された受容確率を人間はある程度使用している。

推定された統計量 a, b に関して検定を行う場合、これが従う分布を解析的に求めることが困難であるため、 t 検定等の検定を行うことができない。帰無仮説 H_0

に対応する分布として r に依存しない（つまり $a = 0$ ）のベルヌーイ分布 $\text{Ber}(\bar{z}|\bar{b})$ を仮定する。なお $\bar{b}$ は被験者全員の行動より決定する。

$\text{Ber}(\bar{z}|\bar{b})$ を用いてサンプリングを行い、統計量 a, b のリストを得る、その a, b の累積分布関数を作成して検定を行う。以下が統計量 a, b のリストを得る手順である。

1. 実験結果から被験者全員の全試行を通した受容割合 $\bar{b} = \frac{1}{N} \sum_{n=1}^N z_n$ を算出する。
2. ベルヌーイ分布のパラメータを $\bar{b}$ とし、各回の名付けの受容・棄却をベルヌーイ分布 $\text{Ber}(\bar{z}|\bar{b})$ からサンプリングする。

$$\bar{z}_n \sim \text{Ber}(\bar{z}|\bar{b}) \dots n \in \{1, \dots, N\}$$

3. 2 の結果 $\{\bar{z}_n\}$ から 2.2.1 と同様に最尤推定でパラメータ a, b を求め、これらを統計量のリストに追加する。
4. 2 と 3 を十分な数 L 回行う（本論文では $L = 1000$ ）。

N はコミュニケーション回数の 45、データセットの数の 2、被験者の数の 12 を掛け合わせた数であり、 $N = 45 * 2 * 12 = 1080$ である。

得られた統計量 a のリスト $a_1, a_2, \dots, a_L$ から累積分布関数 $P_a(a) = \frac{1}{L} \sum_{l=1}^L f(a_l, a)$ を計算する。また同様に統計量 b のリスト $b_1, b_2, \dots, b_L$ から累積分布関数 $P_b(b) = \frac{1}{L} \sum_{l=1}^L f(b_l, b)$ を計算する。ただし、

$$f(x, y) = \begin{cases} 1, & x \geq y \\ 0, & x < y \end{cases}$$

重みパラメータ a は非負の値であり帰無仮説が $a = 0$ であるため片側検定を行う。バイアスパラメータ b は両側検定を行う。具体的には次のように行う。 $P_a(a) \leq 0.05$ であれば帰無仮説 H_0 が棄却される。およびに $P_b(b) \geq 0.975$ もしくは $P_b(b) \leq 0.025$ であれば帰無仮説 H_0 が棄却される。

3. 被験者実験

3.1 実験方法

実験は二人一組のペアで行う。同じペアの二人をそれぞれ被験者 A、被験者 B と呼ぶ。二人の被験者は図 3 の流れに従い、PC を使って実験する。

図 4 は実験用のアプリケーションの操作画面である。図 4 の (1) は、はじめに行ってもらうカテゴリ分類の画面であり、(2) は名付け側、(3) は聞き手側の画面である。詳しい手順は以下に示す。

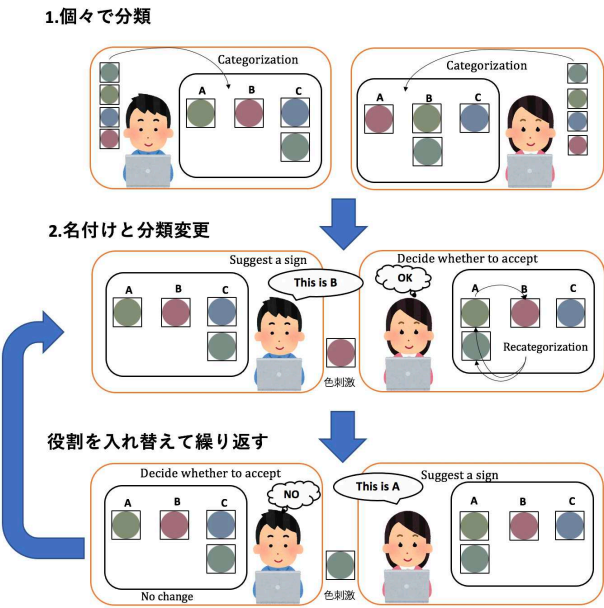


図 3 被験者実験の流れ

- ① **カテゴリ分類**：被験者 A, B それぞれに 15 個の画像を 5 つ以下の A~E のラベルのついたカテゴリに分類してもらう。
- ② **サインの受理/棄却**：① で使用した画像を一つ表示し、被験者 A が名付け側として名前を付け、被験者 B が聞き手側として名付けを受け入れるか拒否するかを判断し、ボタンを押す。
- ③ **分類とサインの更新**：聞き手側の被験者 B は相手の名付けを受け、分類結果変える場合は自分の分類結果を変更する。
- ④ 被験者 A, B で役割を入れ替えて②, ③を行う
- ⑤ ②~④をデータの数分、15 回行う
- ⑥ ④を 3 回繰り返す

手順①, ②でお互いに名付けの提案をし、受け入れるかの判断をすることがコミュニケーションである。一度のコミュニケーションを手順②終了時とし、その際に毎回記録する。分類結果はいつでも変更できるが、聞き手側の際に相手の提案を受け入れるか選んだ後に変更する場合はするように促すような文章が表示される。分類とコミュニケーションは Python で開発したアプリケーションを用いて通信しながら別々の PC で行う。使用する画像は手順③の役割の入れ替え前と後で同じ画像であると、先に聞き手だった被験者 B が被験者 A の名付けの影響を強く受けると考えられたため、画像を表示する順番はランダムとした。

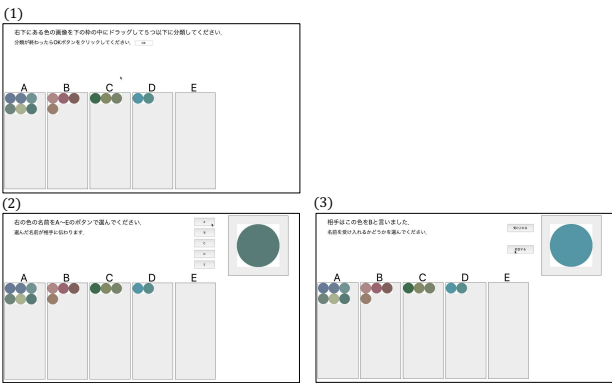


図 4 実験用アプリケーションの実際の実験中の操作画面。(1) カテゴリ分類の画面、(2) 名付け側操作画面、(3) 聞き手側操作画面

3.2 被験者とデータセット

実験の被験者は立命館大学の生徒を 2 人 1 組で 6 組分、計 12 名募集した。12 人のうち 1 人は色覚異常を持っていた。

色空間は人間の知覚する心理的距離を正確に表している CIE- $L^*U^*V^*$ 空間を使用する [8]。CIE- $L^*U^*V^*$ 空間では L^* で明るさ、 U^*V^* で色相が決まる。使用する色の画像の詳細は以下に示す。

- 1. Python の画像処理ライブラリである Pillow (PIL) を使用し、色のついた円形の画像を作成する。
- 2. 3 つの三次元ガウス分布から L^*, U^*, V^* をそれぞれサンプリングする。
- 3. データセットは難易度によるコミュニケーションの違いを見るため、分類が難しいデータセット 1 と分類が簡単なデータセット 2 の 2 つ用意する。
- 4. 2 人に見せる画像は共通で、1 つのデータセットにつき 15 個用意する。
- 5. どのガウス分布からサンプリングするかは一様乱数を使用して一様に決める。

表 1 はデータセットを三次元ガウス分布から生成した際の各パラメータである。

図 5 がデータセット 1(難しい)、データセット 2(簡単)の画像である。

4. 結果と考察

まず、推論した受容確率と実際の受容状況の結果について見ていく。図 6 の上部のグラフはある被験者のコミュニケーション毎の受容確率と相手の提案を受容したかどうかのグラフである。コミュニケーション毎の受容確率と相手の提案を受容したかどうかのグラフ

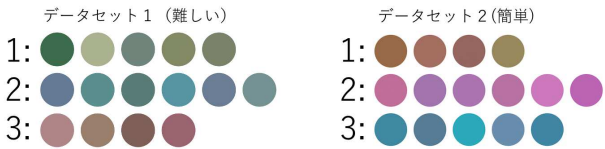


図 5 被験者実験で使用する画像. 左がデータセット 1(難しい) で右がデータセット 2(簡単) である. 同じ行にあるものが同じガウス分布から生成された画像であり, 数字はガウス分布の番号である.

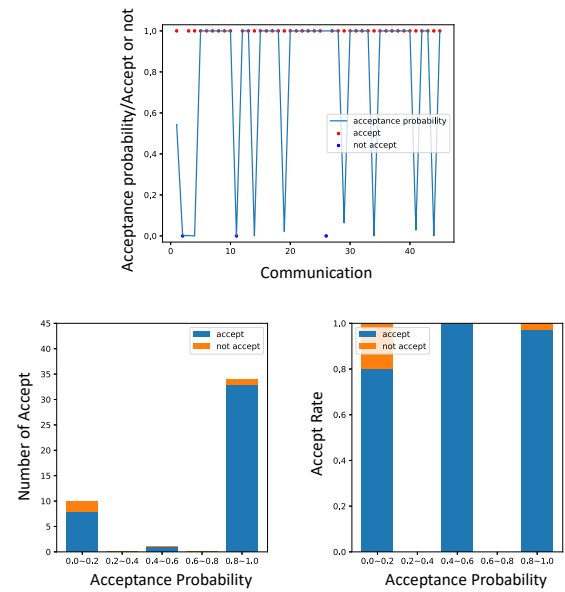


図 6 上部: データセット 1(難しい) での一人の被験者のコミュニケーション毎の推論した受容確率と相手の提案を受容したかどうかのグラフ, 下部: (左) 推論した受容確率ごとの受容した数のグラフ, (右) 推論した受容確率ごとの実際に受容した割合のグラフ

では提案を受容した場合を 1, 拒否した場合を 0 としてプロットしている. 図 6 の下部は同じ被験者の推論した受容確率ごとの受容した数と実際に受容した割合のグラフである. 推論した受容確率が 0 に近いような場合でも受容することがあることがわかる. 相手の提案を受容しやすい傾向が人間の場合には存在すると考えられる. 図 7 は全被験者の受容確率ごとの受容した数のグラフ (左), 受容確率ごとの実際に受容した割合のグラフと, 推定した重み a , バイアス b を使用した $y = ar + b$ のグラフ (右) である. 図 7 を見ると [7] の 4 実験での結果と同じく推論した受容確率が 0.2 ~ 0.8 になることが少なかったが, 全体としては推論した受容確率が高い方が実際に人間が受容する割合が高いことが確認できる. 推論したパラメータ a, b を用い

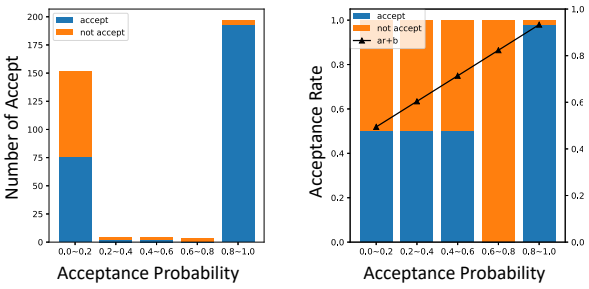


図 7 全被験者の受容確率ごとの受容した数のグラフ (左), 推論した受容確率ごとの実際に受容した割合のグラフと, 推定した重み a , バイアス b を使用した $y = ar + b$ のグラフ (右).

た $ar + b$ のグラフを見ると推論した受容確率がデータの多い 0.2 以下と 0.8 より大きい時に寄ったグラフに見える.

次に検定の結果について見ていく. データセット 1, データセット 2, 両方のデータセットで推定したパラメータが表 2 である. 両方のデータセット, 全被験者のデータを用いての重みパラメータ a とバイアスパラメータ b を最尤推定で求めるとそれぞれ $a = 0.5484, b = 0.4395$ となり, 各データセットでの結果と似たものとなった. 両方のデータセット, 全被験者のデータを用いて求めたパラメータを用いて 2.2.2 の検定を行った. その結果 a の上側確率 $P_a(a) = 0.00 \leq 0.05$, b の上側確率 $P_b(b) = 1.00 \geq 0.975$ となった. よって帰無仮説は棄却され, 推論された受容確率を人間はある程度使用していることが示唆された.

5. まとめ

本研究では, ペアで名付けゲームを行う記号創発実験を行い, Inter-DM の名付けゲームの理論に基づいて推論した名付けに対する人間の受容確率と, 実際に名付けを受け入れた割合を用いて比較した. 具体的にはモデルから推論された受容確率を線形関数をパラメータとしたベルヌーイ分布により人間の受容・棄却行動をフィッティングし, 検定を行った. その結果, 人間は一定の確率で受け入れの判断しているという帰無仮説を棄却し, 推論された受容確率をある使用して判断していることが示唆された.

謝辞

本研究は JSPS 科研費 JP21H04904 および JP17H06379 の助成を受けたものです.

文献

[1] 谷口忠大. 心を知るための人工知能-認知科学としての 記号創発ロボティクス-. 共立出版株式会社, 2020.

[2] Luc Steels. *The Talking Heads experiment: Origins of words and meanings*. Language Science Press, Berlin, 2015.

[3] Katrina Evtimova, Andrew Drozdov, Douwe Kiela, and Kyunghyun Cho. Emergent communication in a multi-modal, multi-step referential game. In *International Conference on Learning Representations*, 2018.

[4] Diane Bouchacourt and Marco Baroni. Miss tools and mr fruit: Emergent communication in agents learning about object affordances. In *Proceedings of the 57th annual meeting of the association for computational linguistics (ACL)*, pp. 3909–3918, 2019.

[5] Yoshinobu Hagiwara, Hiroyoshi Kobayashi, Akira Taniguchi, and Tadahiro Taniguchi. Symbol emergence as an interpersonal multimodal categorization. *Frontiers in Robotics and AI*, Vol. 6, p. 134, 2019.

[6] Tadahiro Taniguchi, Yuto Yoshida, Akira Taniguchi, and Yoshinobu Hagiwara. Emergent communication through metropolis-hastings naming game with deep generative models. *arXiv preprint arXiv:2205.12392*, 2022.

[7] 奥村亮太, 萩原良信, 谷口彰, 谷口忠大. 複数の人間における記号創発実験のメトロポリス・ヘイスティングス法に基づくダイナミクス検証. 人工知能学会全国大会論文集, Vol. JSAI2022, pp. 3L3GS804–3L3GS804, 2022.

[8] Luc Steels, Tony Belpaeme, et al. Coordinating perceptually grounded categories through language: A case study for colour. *Behavioral and brain sciences*, Vol. 28, No. 4, pp. 469–488, 2005.

表 1 各データセット生成時の各パラメータ. μ_k は k 番目の三次元ガウス分布の平均ベクトルである. Σ は分散共分散行列であり, 3つの三次元ガウス分布で共通である.

データセット 1(難しい)			データセット 2(簡単)		
μ_1		$\begin{pmatrix} 60 \\ -10 \\ 20 \end{pmatrix}$			$\begin{pmatrix} 60 \\ 30 \\ 30 \end{pmatrix}$
μ_2		$\begin{pmatrix} 60 \\ -20 \\ -10 \end{pmatrix}$			$\begin{pmatrix} 60 \\ 30 \\ -30 \end{pmatrix}$
μ_3		$\begin{pmatrix} 60 \\ 20 \\ 10 \end{pmatrix}$			$\begin{pmatrix} 60 \\ -30 \\ -30 \end{pmatrix}$
Σ		$\begin{pmatrix} 5^2 & 0 & 0 \\ 0 & 9^2 & 0 \\ 0 & 0 & 9^2 \end{pmatrix}$			$\begin{pmatrix} 5^2 & 0 & 0 \\ 0 & 10^2 & 0 \\ 0 & 0 & 10^2 \end{pmatrix}$

表 2 各データセットに全被験者のデータで推定したパラメータ a, b

	a	b
データセット 1(難しい)	0.5528	0.4387
データセット 2 (簡単)	0.5434	0.4408
両方のデータセット	0.5484	0.4395