共通基盤の構築に寄与する認知機能: モジュラー生成モデルによるシミュレーション

馬場 龍之介 [1] 森田 純哉 [1] 天谷 武琉 [1] 東中 竜一郎 [2] 竹内 勇剛 [1]
[1] 静岡大学大学大学院　[2] 日本電信電話株式会社
{baba.ryunosuke.20, amaya.takeru.19}@shizuoka.ac.jp,
{j-morita,takeuchi}@inf.shizuoka.ac.jp
ryuichiro.higashinaka@ntt.com

概要

抽象的な物体を指示対象とする共通基盤の構築過程を，シミュレーションにより検討した．利用したモデルは，コミュニケーションの過程を，対象の知覚，イメージ生成，言語生成などのモジュールの組み合わせにより説明する．各モジュールの機能を検討する実験により，(1) 対象知覚の調整に比べ，言語生成の調整が，対話相手との意思疎通の成立に有効であること，(2) イメージ生成は，言語生成の多様化を促進することが明らかになった．この結果は，共通基盤構築に寄与する認知機能の役割を検討するうえで，本研究のアプローチの有効性を示す．

1　はじめに

コミュニケーションを成立させるために，参加者が共有する認知的な構成物（i.e., フレーム）を共通基盤と呼ぶ [1]．共通基盤は，会話の繰り返しにより形成され，意思疎通を円滑化する．例えば，抽象的な画像の名称を 2 者間で共有する実験課題において，参加者は画像を具体物に喩えることで表象化し，他者との相互理解の基盤に利用する [2]．ただし，こういった過程において，表面的にはコミュニケーションが成立しているように見えても，実際には相互の理解に差異が生じている場合がある．

共通基盤の形成に寄与する要因を具体的に理解するためには，その内部過程を透明化する認知モデルが必要である．過去，認知科学の分野で，記号の創発に関するモデルが多く構築されてきた．ただし，それらの研究は，現実の実験状況を，コンピュータ内の記号に置換することで表現するものが多く [3, 4]，アナログな内部表象（イメージ）の形成や豊饒な意味を含む自然言語の生成に踏み込む研究は稀であった，そこで本研究では，深層学習によって構成された生成モデルをモジュールとして組み入れたモデルを構築する．このモデルを通じて，コミュニケーションにおける人間の内部処理を具体的にモデル化し，その多様性を可視化することを長期的な目的とする．以下では，本研究において扱った課題と共通基盤の構築モデルを示し，その後に本研究の実験における具体的な目的を示す．

2　タングラム命名課題

本研究では，シミュレーション課題としてタングラム命名課題を用いる．認知科学の分野において，コミュニケーションにおける共通基盤の形成プロセスは長年にわたり検討されてきた．本課題は，そのなかでも古くから用いられてきたものの一つである [5]．この課題を用いて人対人の共通基盤構築を調べた研究では [6]，交わされた会話のパターンを，形状を細分化する「分析的 (analytic) な発話」と対象とするタングラムの形状を別の具体的な対象に喩える「全体的 (analytic) な発話」に分解した．その結果，得られた会話データのなかで，全体的な発話は 2 割，分析的な発話は 1 割を占めた．本研究では，これらの発話のうち，全体的な発話に焦点を当てたモデルの構築を行った．

3　共通基盤の形成モデル

本研究で用いるモデルは，著者らの先行研究 [7] に基づく．このモデルは，タングラム命名課題における送り手による全体的発話の生成と受け手によるその解釈のプロセスをモデル化するものである．そのプロセスは，以下のように複数の深層学習モデル（認知アーキテクチャのモジュールと仮定）による系列的なプロセスとして記述される．

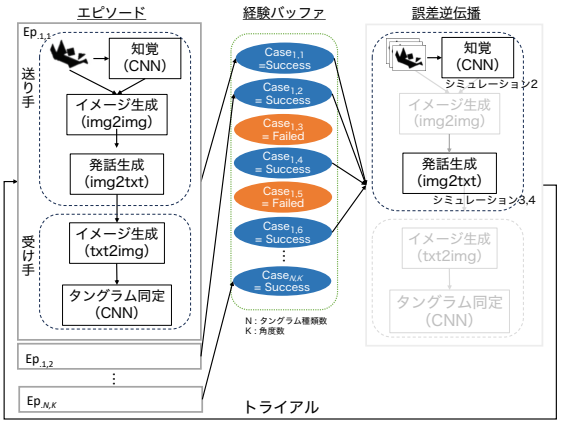

図 1: モデル全体

送り手の処理

1. **タングラムの知覚**：知覚モジュールを介し，各タングラムの形状から物体認識を行う．この処理について，本研究は，Convolutional Neural Network (CNN) を想定した実装を行う．
2. **イメージ生成**：認識された物体のイメージを詳細化する．この処理のために，Stable diffusion[8] の機能である img2img を用いる．その際に，ステップ 1 において生成された全体的なラベルを文脈情報として入力する．
3. **発話生成**：イメージ生成によって得られた画像に対して，イメージキャプショニング (Vison Encoder Decoder[9]) を適用し，詳細な言語表現を得る．

受け手の処理

1. **イメージ生成**：送り手が生成したキャプションから Stable diffusion を利用してイメージを生成する．
2. **タングラムの同定**：前のステップで生成されたイメージから，受け手はタングラム画像の特定を試みる．このプロセスは，生成された画像と観察可能なタングラムとの間の類似度計算によって実現される．画像間の類似度計算には多様な手法が考えられる．本研究では，送り手と受け手は基本的な認知モジュールを共有していると仮定し，送り手の (1) と同様の構造を持つ CNN の出力層のコサイン類似度を想定する．

学習　図 1 は，上記のフローを繰り返すことによる学習プロセスを表現している．図の 1 列目は上記の一連の流れを一つのエピソードとして表している．各エピソードは，実験にて参加者が観察している特定の角度に配置されたタングラムに適用される．結果として得られたネットワークの状態と入力のペアをバッファにプールし，成功（受け手が送り手のタングラムを同定）もしくは不成功のラベルを付与する．プール内の成功事例を用いた誤差逆伝播により，共通基盤モデルの各モジュールの重みを調整する [10]．この一連の手続きを一つのトライアルとし，複数回繰り返す．

4　実験

上記のモデルは，複数の深層学習モデルをモジュールとして組み入れ，系列的なコミュニケーションのプロセスを表現している．実験では，このプロセスの各段階を独立に操作することで，共通基盤の形成に果たす各モジュールの役割を検討した．

本研究で実施した実験（シミュレーション）の全体像は表 1 に示される．表は各シミュレーションにおいて操作したモジュール（2-6 列），および結果の概要（7, 8 列）からなる．操作モジュールに関わるセルのうち，空白 (-) は，そのモジュールがシミュレーションにおいて固定されていることを示す．また，エポック数とバッチ数の設定（例: 5，正事例数）は，そのモジュールが誤差逆伝播法により学習されたこと，中黒で接続された変数（例: 高・低）は，その変数を有する複数のモデルが設定され，比較されたことを示す．また，ランダムと表記されたセルは，そのセルにて深層学習のシード値の変動が検討されたことを示す．

表は，4 つのシミュレーションを示している．これらは段階的に実施された．シミュレーション 1 は，他のシミュレーションで設定するイメージ化のシードを探索した．シミュレーション 2 とシミュレーション 3 は，共通基盤の成立に寄与する送り手側の内部処理の変化を検討した．これらでは，送り手の出発点となる処理（知覚），出口となる処理（言語生成）を対象とし，それぞれ独立に誤差逆伝播法による学習を実施した．シミュレーション 4 はシミュレーション 3 から生じたイメージ生成の存在意義への疑念を解消する補足的なものである．

なお，いずれのシミュレーションも実験には 6 種のタングラムを用い，それらに対して，回転角度（45 度間隔）の異なる 8 種類バリエーションを設定した．すなわち 1 トライアルは 48 エピソードからなる．以下，各シミュレーションの狙いと手続き，得られた結果の詳細を示す．

表 1: 実験の全体

	Sender			Receiver		最大正解率	
	視覚 *	イメージ化	言語化 *	イメージ化	視覚化	高 †	低 †
シミュ 1	-	ランダム	-	ランダム	-	0.35	0.06
シミュ 2	5，正事例数	高・低	-	高・低	-	0.38	0.10
シミュ 3	-	高・低	5，正事例数	高・低	-	0.45	0.41
シミュ 4	なし	なし	5，正事例数	高・低	-	0.16	0.43

* モジュールを学習させる際のエポック数，バッチ数

† シミュレーション 1 で得られた正解率の高いシード値と低いシード値の 2 条件

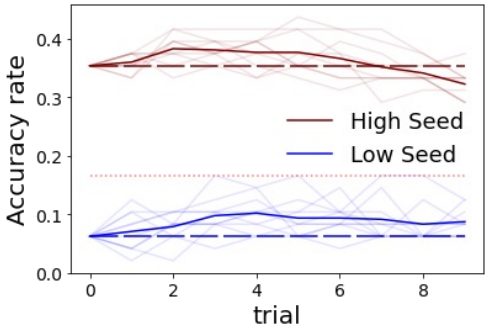

図 2: シミュレーション 2 の結果（破線は各シミュレーションの初期値，点線はチャンスレベル）

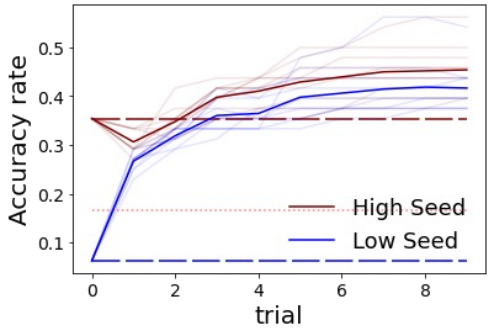

図 3: シミュレーション 3 の結果（破線は各シミュレーションの初期値，点線はチャンスレベル）

シミュレーション 1　イメージ化に利用している Dffusion モデルは，設定されるシードの初期値によって，異なるスタイルの画像を生成する．続くシミュレーションの準備段階として，そのようなスタイルの差異が，本研究の課題に及ぼす影響を検討した．

このシミュレーションでは，送り手と受け手のそれぞれのイメージ生成に利用するシード値をそれぞれランダムにサンプリングした．サンプリングの後に図 2 におけるトライアル (6 種のタングラムと 8 の角度からなる 48 エピソード) を 1 回のみ実行し，それを独立に 100 回繰り返した．(n=100)

100 回の平均は 0.18，標準偏差 0.05, 正解率の最大値が 0.35，最小が 0.06 であった．平均はチャンスレート付近となるものの，正解率にはばらつきが認められた．このばらつきを考慮するために，以後のシミュレーションでは最小値と最大値に対応するシード値をイメージ化に設定した 2 つのモデル（以後，低モデル，高モデルと表記）でシミュレーションを実施した．

シミュレーション 2　低モデルと高モデルに対して，図 1 の「知覚 (CNN)」を学習対象とするトライアルを 10 回繰り返した．さらに，これを 10 回独立に実行し，その結果を図 2 にプロットした．

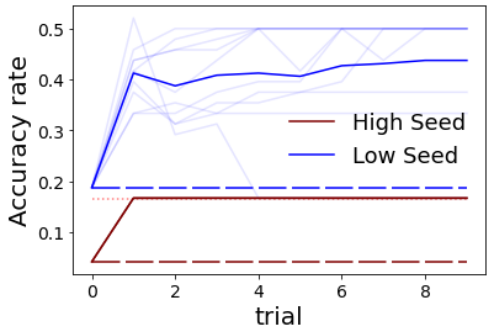

図 4: シミュレーション 4 の結果（破線は各シミュレーションの初期値，点線はチャンスレベル）

正解率の初期値が高いシード値は，トライアルが進行しても学習傾向は見られなかった．それに対して，正解率の初期値が低いシード値は，トライアルの進行に伴いわずかに正解率が上昇したがチャンスレートに届かなかった．

シミュレーション 3　低モデルと高モデルに対して，図 1 の「言語生成 (img2txt)」を学習するトライアルを 10 回繰り返した．これを 10 回独立に実行し，その結果を図 3 に示した．

両方のモデルにおいて学習傾向が示された，特に低モデルでは，0.062 と低い数値から，41%とおよそ

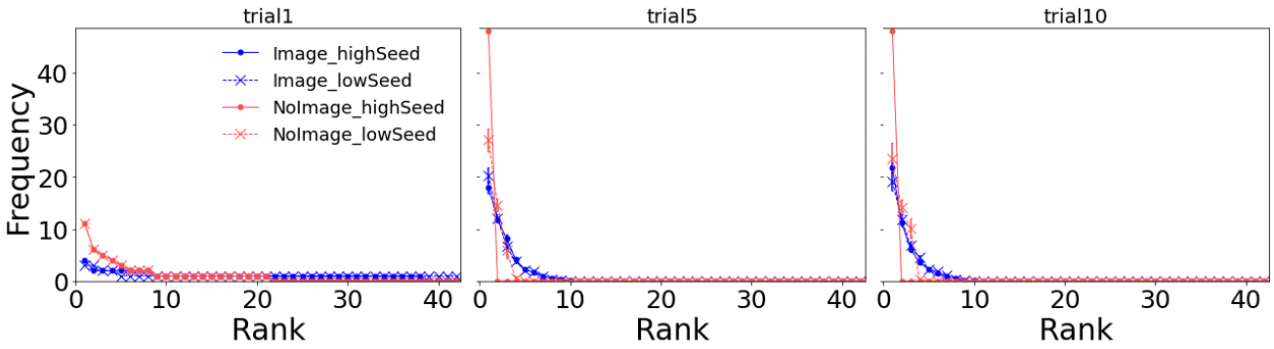

図 5: シミュレーション 3, 4 にて得られた言語表現の頻度グラフ

4 割もの改善が見られた．なお，この結果から生じる一つの疑念は，モデルにおけるイメージ化の必要である．本モジュールにおいては送り手のイメージ化のプロセスがなくても学習が可能である．

シミュレーション 4　本モデルにおけるイメージ化の影響を調査するためにイメージ化なしの条件で実験を行った．イメージ化なし条件は，送り手の知覚とイメージ化のプロセスを削除し，タングラム画像から直接，言語生成を行う．このモデルに対して，シミュレーション 3 と同様の条件（img2txt の学習）で実行した．なお，シミュレーション 4 においても低モデルと高モデルを設定する．ただし，これらのモデルにおけるシードの設定は，受け手のイメージ化のみに適用された．それぞれのモデルによる学習の結果は図 4 に示される．結果，高モデルはチャンスレベルを上回らなかったものの，低モデルにおいてチャンスレベルを上回る上昇が得られた．

この結果は，本研究のモデルが，イメージ化なしでも，タングラム命名課題の成績を向上させることを示している．このうえで，イメージ化の機能を検討するために，生成された言語表現に関する検討を実施した．図 5 は，シミュレーション 3（イメージ化あり）とシミュレーション 4（イメージ化なし）で，生成された言語表現の頻度を比較したものである．1 試行目は発話されるキャプションはどの条件においても多様性が示される．その多様なキャプションがイメージ化なしでは，極度に減少していくことが示された．

5　総合考察と結論

本研究はコミュニケーションの内部プロセスを生成モデルにより表現した認知モデルによるシミュレーションを実施した．本研究の中核となるシミュレーション 2 とシミュレーション 3 を比較することで，共通基盤の形成においては，知覚よりも言語生成を学習するほうが効果を発揮することが示された．つまり，タングラム画像を認識し，解釈しイメージ化の材料を作る箇所ではなく，何を話すべきかを学習することの重要性を示唆する [11]．このようなシミュレーション 3 の結果は，送り手の内部処理の必要性に対して疑問を投げかけるものである．そのため，イメージ生成を除いたシミュレーション 4 を実施した．結果，本研究で設定した課題の達成にはイメージ化が必ずしも必要ではないという結果が得られた．しかし実際に生成されたキャプションを見ると，固定化されイメージ化ありに比べて多様性が少なくなっていることが確認されるこのことから人間のような言語生成の多様さを導くためには，イメージ化が必要であることが示唆される．

今後，本研究のモデルに対するより包括的な検討が必要である．本研究では送り手のモジュールのみを検討した．それに対して，実際のコミュニケーションにおいては受け手の学習も行われる．そのような相互適応過程に関する検討を進めていく必要がある．

また，本研究において検討したモデルは，最大で 0.45 の平均正解率に至った．この正解率に対する評価を進める必要がある．人間の認知プロセスのシミュレーションを志向する本研究にとって，完全な一致を目指す必要はない．ただ，本研究に先立って行われた人間相手の実験 [6] で，課題後に命名が不一致となった事例はなかった．そのことを踏まえれば，現状の正解率は十分ではない．共通基盤の形成プロセスの理解を進めるためには，今後，包括的なモデルの学習を検討しつつ，より認知的に妥当で学習率を向上させる手法を検討していく必要がある．

参考文献

[1] Robert Stalnaker. Common ground. **Linguistics and philosophy**, Vol. 25, No. 5/6, pp. 701–721, 2002.

[2] Shali Wu and Boaz Keysar. The effect of information overlap on communication effectiveness. **Cognitive Science**, Vol. 31, No. 1, pp. 169–181, 2007.

[3] David Reitter and Christian Lebiere. How groups develop a specialized domain vocabulary: A cognitive multi-agent model. **Cognitive Systems Research**, Vol. 12, No. 2, pp. 175–185, 2011.

[4] 森田純哉, 金野武司, 奥田次郎, 鮫島和行, 李冠宏, 藤原正幸, 橋本敬. 協調的コミュニケーションを成立させる認知的要因-認知アーキテクチャによるシミュレーション. ヒューマンインタフェース学会論文誌, Vol. 20, No. 4, pp. 435–446, 2018.

[5] Herbert H Clark and Deanna Wilkes-Gibbs. Referring as a collaborative process. **Cognition**, Vol. 22, No. 1, pp. 1–39, 1986.

[6] Saki Sudo, Kyoshiro Asano, Koh Mitsuda, Ryuichiro Higashinaka, and Yugo Takeuchi. A speculative and tentative common ground handling for efficient composition of uncertain dialogue. In **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 3150–3157, 2022.

[7] Junya Morita, Tatsuya Yui, Takeru Amaya, Ryuichiro Higashinaka, and Yugo Takeuchi. Cognitive architecture toward common ground sharing among humans and generative ais: Trial modeling on model-model interaction in tangram naming task. In **Proceedings of the 2023 AAAI Fall Symposium on Integrating Cognitive Architectures and Generative Models**. AAAI Press, 2023.

[8] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In **Proceedings of the IEEE/CVF conference on computer vision and pattern recognition**, pp. 10684–10695, 2022.

[9] Jiahui Yu, Yuanzhong Xu, Jing Yu Koh, Thang Luong, Gunjan Baid, Zirui Wang, Vijay Vasudevan, Alexander Ku, Yinfei Yang, Burcu Karagol Ayan, et al. Scaling autoregressive models for content-rich text-to-image generation. **arXiv preprint arXiv:2206.10789**, Vol. 2, No. 3, p. 5, 2022.

[10] Maxim Lapan. **Deep Reinforcement Learning Hands-On: Apply modern RL methods, with deep Q-networks, value iteration, policy gradients, TRPO, AlphaGo Zero and more**. Packt Publishing, 6 2018.

[11] Jarrad AG Lum, Gina Conti-Ramsden, Debra Page, and Michael T Ullman. Working, declarative and procedural memory in specific language impairment. **cortex**, Vol. 48, No. 9, pp. 1138–1154, 2012.