抽象図形への命名を介したコモングラウンド形成モデルの構想

森田 純哉 [1] 由井 達也 [1] 天谷 武琉 [1]
光田 航 [2] 東中 竜一郎 [2] 竹内 勇剛 [1]
[1] 静岡大学情報学部　[2] 日本電信電話株式会社
{j-morita,takeuchi}@inf.shizuoka.ac.jp
{yui.tatsuya.20, amaya.takeru.19}@shizuoka.ac.jp
{koh.mitsuda.td,ryuichiro.higashinaka.tp}@hco.ntt.co.jp

概要

コミュニケーションにおいて，送り手が発した記号は，受け手の有する認知的な枠組みを介して復元される．つまりコミュニケーションを効率化するためには，送り手と受け手の間で認知的な枠組みをすり合わせていくプロセスが必要である．本稿では，そのような共有された枠組みをコモングラウンドと呼び，タングラム命名課題を対象としたコモングラウンド形成の認知モデルを提案する．特に，本稿ではモデルの背景，設計，実装例を示し，今後の課題を議論する．

1　はじめに

コミュニケーションにおけるコモングラウンドの必要は様々な研究者によって指摘されている [1, 2]．送り手が発した記号の意味は，受け手の有する知識によって復元される．よって，送り手と受け手の間で共通の認知的枠組みが存在しない場合，意図の伝達に膨大なコストが必要となる．共通の認知的枠組みが存在することで，多義的な記号の意味が絞り込まれ，簡素な表現による素早いコミュニケーションが可能になる．そして，そのようなコモングラウンドを構成する共有知識や信念は，通常は暗黙的なものであり，その全てを記号的に表現することは困難である．

本稿では，コモングラウンドの形成がどのように計算機上でモデル化できるかを議論する．近年の言語処理の研究では，Transformer [3] をベースとした Bidirectional Encoder Representations from Transformers (BERT) [4] や Generative Pretrained Transformer (GPT) [5] などの手法が全盛である．深層学習によって調整された大規模なパラメータは，人間の言語利用の背後にある暗黙的な知識をよく表現する．しかし，それらのパラメータは，人間が課題遂行において保持する目標や時間的な文脈を陽には表現しない．また，人間が言語で表現する経験は多様なモダリティから形成されるため，コモングラウンドの形成には，複数モダリティの経験を統合する必要がある．

こういった複数のモジュールを統合し，問題解決に至る一連のプロセスをモデル化するための考え方として，認知アーキテクチャが存在する．既存の認知アーキテクチャとして，Soar [6] や ACT-R [7] は有名である．これらの認知アーキテクチャは，現状では記号的な表現に大きく依存しており，スケールの問題を抱える．実世界のコミュニケーションを説明するモデルを構築するためには，深層学習などで構築された分散表現をベースにした認知アーキテクチャが必要である．この考えから，本稿は，深層学習で構成されたモジュールを，認知アーキテクチャへ統合し，コモングラウンド形成のプロセスを明示化することを志向する．

2　タングラム命名課題

コモングラウンドの形成過程を検討する課題として，タングラムと呼ばれる抽象図形を用いたコミュニケーション課題が検討されている [1]．タングラムは，“タン” と呼ばれる平面図形を組み合わせることで構築される．タングラムを具体物のシルエットとみなすことで，多様な解釈が生成される．この解釈の生成は，その際の知覚者が有する認知的枠組みによって変化する．よって，言語表現のみにより，発信者の指示するタングラムを受け手が同定するためには，コモングラウンドの共有が必要になる．

タングラムを題材とした対話のデータとして，本研究では須藤ら [8] の研究に焦点を当てる．須藤らの実験において，セッションに参加した 2 者は，6 つのタングラムの命名について合意することを目指

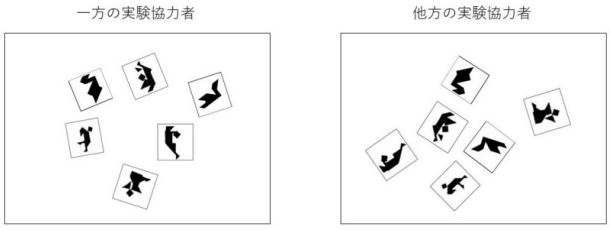

図 1　タングラム命名課題における配置の一例.

した．以降，この実験課題をタングラム命名課題と呼ぶ．図 1 は，タングラム命名課題において，実験の参加者が観察するタングラムセットの例を示している．両者は同一のタングラムセットを提示されるものの，その配置や角度は異なっている．タングラム命名課題において，参加者は互いの画面が見えず，音声のみで課題を遂行することが求められた．

須藤らは，タングラム命名課題における発話を，全体的発話と分析的発話の観点から分析した．全体的発話は，タングラムの形状を具体物に喩える表現（例：“塔っぽいやつ”，“枝がでている木”，“サボテンとか”）であり，分析的発話はタングラムを構成する平面図形への言及（例：“小さい三角形が 2 つ”，“左右に四角と三角が出てる”）である．須藤らのデータでは，実験全体を通して全体的発話が部分的発話よりも多く，セッションが経過するに従い，その差が拡大していくことを示されている．

3　タングラム命名モデル

タングラム命名課題で得られたデータをシミュレーションするモデルの構想を示す．まず，認知アーキテクチャとして考えた際に必要なモジュールについて議論し，続いて一部の処理に焦点を当てた詳細なモデルを検討する．最後に現時点で得られている予備的な実行の結果を示す．

3.1　モジュールの構成

図 2 はタングラム命名課題に関与すると想定されるモジュールを統合したものである．一般的に認知アーキテクチャは，視覚，聴覚，運動などの入出力に関するモジュール，ゴールや記憶を保持する内的なモジュール，そしてそれらを結合する中枢モジュール（ワーキングメモリやルールエンジン）から構成される [9]．図 2 では上部に内的モジュール，下部に入出力に関するモジュールが配置され，中央のプロダクションによってそれらが統合される[1)]．

1)　ここでの統合は数ある認知アーキテクチャのなかでも ACT-R [7] をベースとしている．

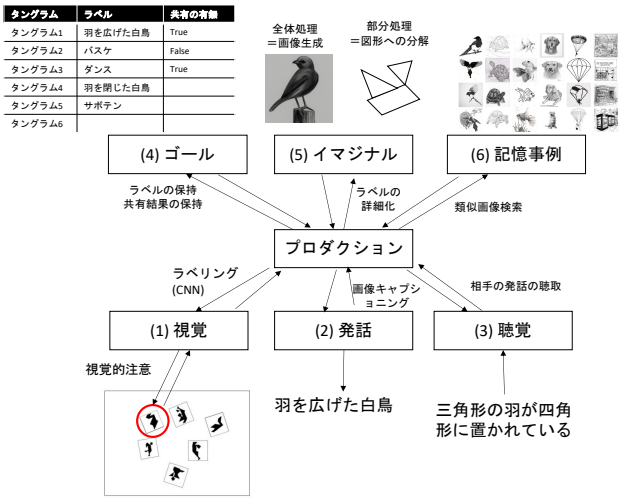

図 2　タングラム命名モデルのモジュール構成.

以下，各モジュールの役割を簡単に記載する．

1. 視覚モジュール：外界のタングラムに注意を向け，タングラムを認識する．
2. 発話モジュール：個々のタングラムを区別する言語ラベルを生成し他者へ伝達する．
3. 聴覚モジュール：他者が生成した言語ラベルを聴取する．
4. ゴールモジュール：個々のタングラムに対して付与された言語ラベルを保持する．また各ラベルに対する相手との合意の有無を管理する．
5. イマジナルモジュール：言語ラベルとタングラムを結びつけるためのイメージ操作を行う．須藤らのデータに即せば，以下の 2 つの処理が想定できる．
   - 部分処理：タングラムを幾何図形に分解し，対話相手がタングラムを特定できる言語表現を生成する．
   - 全体処理：対話相手から聴取した言語ラベルをもとにイメージを生成する．
6. 記憶事例：一般的な言語ラベルと画像の組み合わせからなるデータを有する．このデータを活用することでモデルはタングラムに対してラベルを付与し，イメージからラベルを生成する．

3.2　送受信パイプライン

図 2 のアーキテクチャの全体像は大きいものであり，部分に分割したモデルの構築が有効である．特に本稿では，須藤らの分類のうち，全体処理に関する発話の送受を詳細化することを試みる．すなわち，送り手がタングラムからイメージを解釈し，そ

のイメージを言語化し，受け手が言語化されたイメージを復元するまでのプロセスである．以下に各プロセスの概要を示す．

送り手の処理

1. **タングラムの知覚**：視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．視覚モジュールの実装としては，Convolutional Neural Network (CNN) を想定する．しかし，タングラム図形は多義性を意図して構成されおり，通常の CNN による物体認識に困難が生じる可能性が存在する．特に一般物体認識の CNN モデルに対しては，テクスチャバイアス（全体的な形状ではなく画像表面のテクスチャに影響された認識をおこなうこと）の存在が報告されている [10]．このバイアスは，テクスチャに分類の有効な手がかりが存在しないタングラム命名課題において，クリティカルな影響を生じさせると考えられる．よって，本研究では，テクスチャバイアスを生じさせないデータセットとして．ImageNet Sketch [11] を用い，白黒画像から 1000 のラベルを分類する CNN を新規に学習した．
2. **イメージ検索**：タングラムの認識から物体の特徴を詳細化する発話を生成するためには，ラベルのイメージの具体化が必要である．この処理のために，本研究では過去に獲得した記憶からのイメージの復元を考える．本研究のモデルの場合，上記の CNN の学習に利用した画像データを過去の記憶として想定できる．この画像データから，CNN が出力したラベルを付与する画像セットを絞り込む[2)]．そして，絞り込まれた画像からタングラムの形状と最も類似する画像を選択する．画像の類似の計算方法として，本研究では，AKAZE によって抽出される特徴点マッチング [12] を用いる．これにより，タングラムと形状が最もマッチする画像が検索され，ラベル付けの根拠がイメージとして具体化されると想定する．
3. **発話生成**：イメージ検索によって得られた画像は，図 2 のアーキテクチャのイマジナルモジュールに一旦格納される．そして，イマジナルモジュール内の画像に対して，発話モジュールがイメージキャプショニングを適用することで詳細な言語的なラベルを得る．本研究ではこの処理に Vision transformer (ViT) [13]) と GPT-2 [5] をもちいた学習済み Vison Encoder Decoder モデル [14] を利用する．

受け手の処理

1. **イメージ生成**：送り手の生成した言語ラベルは，受け手の聴覚モジュールに格納される．受け手モデルは格納された言語ラベルからイメージ生成を行う．言語ラベルからのテキスト生成には，Stable Diffusion [15] を利用する．画像生成モデルにテキストを入力する際，送り手のイメージ生成 (ImageNet Sketch) の利用に合わせ，受け取った言語ラベルの先頭に，“a monochrome sketch of” という文字列を付与する．
2. **タングラムの同定**：AKAZE による特徴点マッチングを利用することで，生成された画像と類似するタングラム画像を同定し，結果をゴールモジュールに格納する．なお，この際，タングラムを 45 度単位で回転させ，合計で 48 (6 × 8) のタングラムのイメージとの比較を行う．

3.3 実行結果

既に述べたように，上記のパイプラインは，タングラム命名課題における処理の一部である．実際の課題においては，上記のパイプラインが他の処理と組み合わされ，反復的に適用されることで送り手と受け手の認知的枠組みがすり合わされていく．

本稿では，そのようなプロセスの初期状態を示すために，3.2 のパイプラインの実行結果を検討した．図 3 はその一例である．送り手が注目したタングラムに対し，Black Swan というラベルが出力された．そして，そのラベルと形状の類似している学習事例（首が右方向に向けられた黒鳥）が検索され，その画像に対して “a black and white bird standing on top of a white bird” というキャプションが生成された．このキャプションに対し，受け手は，黒い鳥の上に白い鳥が乗ったイメージを生成した（白い鳥と黒い鳥の関係がキャプションとは逆転している）．この画像に対して同定されたタングラムは，生成されたイメージに含まれる 2 つの鳥のうち，上側の鳥と対応付けられるものと見ることができる（右側にくちばしがあり，下部に足が配置される）．

上記のように本研究で提案するパイプラインは，抽象図形から言語的ラベルの生成，言語的ラベルから抽象図形の同定に至るまでのプロセスを可視化す

2) ImageNet Sketch は各ラベルに対して 50 枚の画像を用意している．

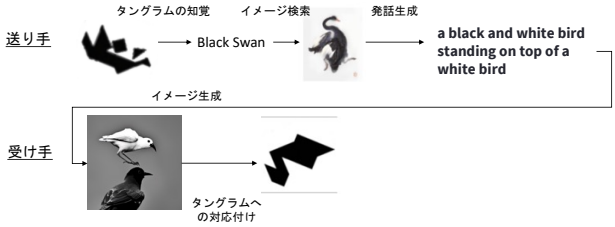

図 3　実行例.

ることに成功している．そして，それら一つ一つのステップは，既存の画像処理ライブラリや学習済み深層学習モデルを用いたものであり，一定の妥当性を有しているとみなせる．ただし，これらの処理には程度の差はあれノイズが混入し，処理がまとめられた際には，送り手が観察したタングラムと受け手が同定したタングラムが食い違うものとなった．すなわち，この事例は，タングラム命名課題におけるミスコミュニケーションの一例をモデル化したものとなっている．

上記のようなミスコミュニケーションが，提案したパイプラインにおいてどの程度の頻度で生じるのかを検討した．この検討において，送り手は，角度を 8 段階に変更した 6 種類のタングラムを観察し，それぞれに対して言語ラベルを生成した．生成された合計 48 の言語ラベルに対し，受け手が同定したタングラムとの対応をカウントした．図 4 は送り手が観察したタングラムと受け手が同定したタングラムの間での混同行列を示す．表中の数値から正解率を計算したところ 0.166 (8/48) となり，6 クラス分類のチャンスレートと等しいものとなった

このように今回の実装は，標準で送り手と受け手の認識の齟齬を生じさせるものとなった．もともとタングラム命名課題は，タングラムの多義性を利用し，コミュニケーションの開始時に十分なコモングラウンドが存在しない状況を形成することを意図したものである．その意味で，本研究は，この課題の前提を，複数の深層学習モデルの統合によってデモンストレートしたものとみなせる．

4　おわりに

本稿では，コモングラウンド形成の背後にある認知プロセスをシミュレーションする計算機モデルの構想を示した．なお，タングラム命名課題を深層学習によってモデル化した事例は既に存在する [16]．この先行研究に対して，本研究は，認知アーキテクチャに基づくモジュールの仮定に基づくことに特徴

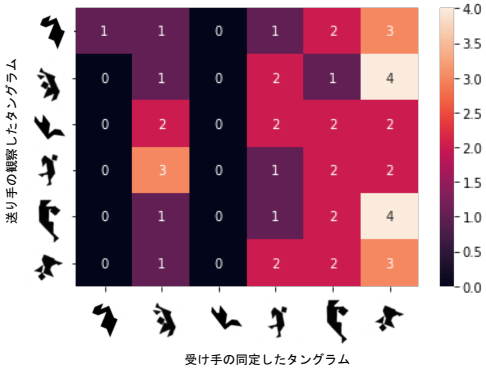

図 4　送り手と受け手の混同行列.

がある．深層学習モデルに埋め込まれたサブシンボリックな知識構造を利用することで，タングラムの知覚から言語ラベルの生成，言語ラベルの聴取からタングラムの同定までの処理を試作した．試作した実行結果を検討することで，タングラム命名課題における送り手と受け手のミスコミュニケーションを表現した．

本稿において示された多義性を解消する手段は複数考えられる．まず，本研究で扱った複数の深層学習モデルは異なるデータセットから構成された．これらの整合を図ることで，送り手と受け手の一致率が向上する可能性がある．また，本研究の受け手モデルが同定したタングラムは，図 4 に示されるように一部（6 列目）に偏っている．このようなタングラムの同定に関する競合を対話相手とのインタラクションのなかで解消する仕組みが必要である．こういったノイズを含む状況でのコモングラウンドの調整は，対話理解を報酬とした強化学習によって解決できる可能性があり [17]，対話理解の報酬を発生させるためには，本研究ではモデルの詳細化ができなかった部分処理などのプロセスも必要になる．全体処理的な発話から受け手が部分処理を返すことで，両者の認識の一致を確かめることができるようになると考える．先述の先行研究 [16] は事前学習されたモデルをタングラム課題のデータでファインチューニングすること，送り手のメッセージに部分処理の表現を付与することで，同定率が向上することを示している．

今後，上記の検討を進めることで，本稿で示したフレームワークの完成を目指す．そして，フレームワークに基づくモデルが完成することで，人間のコモングラウンド形成に対する理解が成し遂げられ，同時に人とコモングラウンドを共有する人工物の構築がなされると考える [18]．

参考文献

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

[2] David R Traum. A computational theory of grounding in natural language conversation. Technical report, Rochester Univ NY Dept of Computer Science, 1994.

[3] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **NIPS'17: Proceedings of the 31st International Conference on Neural Information Processing Systems**, 2017.

[4] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. **arXiv preprint arXiv:1810.04805**, 2018.

[5] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[6] John E Laird. **The Soar cognitive architecture**. MIT press, 2019.

[7] John R Anderson. **How can the human mind occur in the physical universe?** Oxford University Press, 2007.

[8] 須藤早喜, 浅野恭四郎, 光田航, 東中竜一郎, 竹内勇剛. 推測的かつ暫定的な対話による共通基盤の形成過程. 電子情報通信学会和文論文誌 (D), Vol. 106, .

[9] John E Laird, Christian Lebiere, and Paul S Rosenbloom. A standard model of the mind: Toward a common computational framework across artificial intelligence, cognitive science, neuroscience, and robotics. **AI Magazine**, Vol. 38, No. 4, pp. 13–26, 2017.

[10] Robert Geirhos, Patricia Rubisch, Claudio Michaelis, Matthias Bethge, Felix A Wichmann, and Wieland Brendel. Imagenet-trained cnns are biased towards texture; increasing shape bias improves accuracy and robustness. **arXiv preprint arXiv:1811.12231**, 2018.

[11] Haohan Wang, Songwei Ge, Zachary Lipton, and Eric P Xing. Learning robust global representations by penalizing local predictive power. In **Advances in Neural Information Processing Systems**, pp. 10506–10518, 2019.

[12] Pablo F Alcantarilla and T Solutions. Fast explicit diffusion for accelerated features in nonlinear scale spaces. **IEEE Trans. Patt. Anal. Mach. Intell**, Vol. 34, No. 7, pp. 1281–1298, 2011.

[13] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. **arXiv preprint arXiv:2010.11929**, 2020.

[14] Ankur Kumar. The illustrated image captioning using transformers. **ankur3107.github.io**, 2022.

[15] Robin Rombach, Andreas Blattmann, Dominik Lorenz, Patrick Esser, and Björn Ommer. High-resolution image synthesis with latent diffusion models. In **2022 IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)**, 2022.

[16] Anya Ji, Noriyuki Kojima, Noah Rush, Alane Suhr, Wai Keen Vong, Robert D. Hawkins, and Yoav Artzi. Abstract visual reasoning with tangram shapes. In **Proceedings of the Conference on Empirical Methods in Natural Language Processing**, 2022.

[17] Atsumoto Ohashi and Ryuichiro Higashinaka. Adaptive natural language generation for task-oriented dialogue via reinforcement learning. **Proceedings of the 29th International Conference on Computational Linguistics**, p. 242–252, 2022.

[18] Roberto Dessì, Eugene Kharitonov, and Baroni Marco. Interpretable agent communication from scratch (with a generic visual processor emerging on the side). **Advances in Neural Information Processing Systems**, Vol. 34, pp. 26937–26949, 2021.