連続的な動作と思考による時間的文脈の設計：モデルベース対話デザイン

Designing temporal context through continuous action and thought: Model-based dialogue design

佐々木康佑[1]　西川純平[1]　白砂大[1]　森田純哉[1]
Kosuke Sasaki[1]　Jumpei Nishikawa[1]　Masaru Shirasuna[1]　Junya Morita[1]

[1] 静岡大学　[1] Shizuoka University

Abstract: 対話では，動作や思考の連続が文脈効果を生み，後続の行動や認知に影響を与える．近年の深層学習に基づく動作生成アプローチでは，時間的文脈における人間の内部処理を近似することは困難である．本研究では，対話における認知プロセスをモデル化することで発話のカテゴリを連続させ，単語の大きさを相対的に比較するジェスチャーを通じて思考と動作の連続性を表現した．これにより時間的文脈を設計することを目指した．実験ではこの連続性が対話の評価に与える影響を検証し，人間との自然なインタラクションの実現への貢献を示した．

1　はじめに

人間は時間の中で生きている．時間の中で，自身の内部状態，あるいは環境の状態は，相互に影響しながら，変化を続ける．特に，自己の時間と相手の時間が重なったときに，“対話” が生じ，社会の形成が導かれる [1]．対話は，言語と身体という 2 つのチャネルを介し，それぞれのチャネルで生まれた文脈が時間の中で意味的に混ざり合うことで成立する．

対話のなかでの言語と身体のチャンネルの関係は，co-speech gesture の形で表れる．例えば，人間は，目の前のモノに手を伸ばしたり掴んだりするような身体動作を通じてモノの物理的特性を探索する．これにより情報が概念化され，発話の生成が容易となる [2]．このように，時間的な文脈において動作や思考は深く結びついており，動作や思考が連続することで後の動作や思考に影響を及ぼす [3]．この動作と思考のダイナミクスが「生きている」という状態を生む．

また，マクニールはこのような “連続性” を成長点理論で提唱した [4, 5]．“成長点” とは，後に動作として表れる心的なイメージと発話として表れる言語が結合された心的表象の単位である．成長点理論ではイメージと言語は相互作用を通じてジェスチャーと発話へ成長するとされる．このような動作と思考のダイナミクスの交差が人間の対話の本質である．

対話の本質を追求することは，人間との自然なインタラクションを実現する人工物を開発する上で非常に重要である．近年，こうした人工物を開発するアプローチとして，深層学習ベースの動作生成が主流である [6]．こういったボトムアップなアプローチでは，人間の行動データから一般的な特徴を抽出することができるため，自然な動作や発話を生成することが容易である．しかし，著者らは，対話における動作と思考によって時間的文脈を構築する人工物を，ボトムアップなアプローチのみによって検討することは困難と考える．なぜなら，深層学習は他者との対話を通じた時間的文脈の中で生じる人間の内部処理を十分に再現することができないためである．深層学習は知覚や動作，あるいは発話など，外界と接する表層的な人間の活動には強力に働くが，その際の認知プロセスや心的表象などの内部状態に関する仮定をおくことが困難である．そのため，人間にとって，自然と感じられるインタラクションを設計することには限界がある．

そこで，本研究では，動作と思考の連続性を表現するモデルを提示し，そのモデルに基づくことで人工物間でのインタラクションによって生じる時間的文脈を設計する．実験課題として対話のシンプルな形式である「しりとり」を利用し，2 体のロボットの動作と発話を操作する．上記の理論的背景に基づくと，連続的な対話における身体動作は，他者の存在する対話において相手の身体動作に影響を受けることが考えられる．例えば，人間が，他者の発話において出現した単語よりも大きい対象を示す際には，実際の単語の指示対象と比較して誇張した身体動作で表現する．また，連続的な対話では，発話される単語のカテゴリが思考に影響を与えることが考えられる．例えば，人間の行うしりとりにおいては，しりとりが進むにつれ，前に出現した単語と同一カテゴリの単語の回答が多くなるとい

う文脈効果を考える．

本研究の目的は，動作と思考の連続性を表現することにより構築された時間的文脈が，しりとりや動作の自然さやロボットへの心的状態の帰属に影響を及ぼすのか検討することである．動作の連続性は大規模言語モデル（LLM）の Llama3[7] を用いて得られた大きさに基づいてロボットの関節を操作することで表現する．思考の連続性は，認知アーキテクチャの 1 つである，Adaptive Control of Thought-Rational (ACT-R)[8] を用いて西川らにより構築されたしりとりモデル [9] を改変することで表現する．

本研究のリサーチクエスチョンは以下の通りである．

1. 思考の連続性は対話の印象にどのような影響を与えるのか？
2. 動作の連続性は対話の印象にどのような影響を与えるのか？
3. 動作と思考の連続性の一致は対話の印象にどのような影響を与えるのか？

2 関連研究

2.1 言語のグラウンディング

言語を理解して利用するためにはグラウンディングが必要である．グラウンディングとは，単体では “意味” を有さない言語表現と何らかの対象が結びつくことにより “意味” が生じることである．グラウンディングは物理に接地するものもあれば，社会に接地するものもある [10].

物理に接地するグラウンディング（シンボルグラウンディング）は身体性との関係で検討されてきた．そもそも言語は身体と結びついて獲得される．Hawkins は各概念の背後には概念に関する知識を内包する連続的な空間が存在し，言語などを介在することで身体動作に変換されるという座標系理論を提案した [11]．Pinker は，そのような時空間に埋め込まれた知識構造をベースとして思考が形成され，思考によって言語が形成されるとした [12]．また，今井によると，人間はシンボルグラウンディング [13] によって記号（言語）を身体的に理解して文化的な文脈に接地させることで新たな記号を創発していく [14].

こういった理論背景に基づく，ジェスチャーに関する研究も盛んである．Lakoff によると，人間の語彙はメタファーの利用，つまり過去の類似の経験に基づいて新しい状況にラベル（単語）を割り当てることによって構築されてきた [15]．このような身体的な経験からの概念へのマッピングを “プライマリーメタファー” と呼ぶ．そして，人間の発話においてはプライマリーメタファーが身体動作として現れる [16, 17]．筆者らは，Lakoff によるプライマリーメタファーの定義を参照して，単語分散表現から抽出した大きさに関する数量的意味からロボットの co-speech ジェスチャーを生成した [18, 19, 20]．これにより人間と同様な記号と数量の変換機構を構築し，身体を介した言語のグラウンディングを検討した．

社会に接地するグラウンディングに関する研究としては，コモングラウンドの形成に関する研究などがある．コモングラウンドとは，インタラクションへの参加者間で共通基盤を徐々に構築し，維持するプロセスのことである [21]．そういった共通基盤の形成には，自身の目的を他者が理解していると相互に信じて協力し合うことが必要とされる [22]．人とエージェント，もしくはエージェント間でのコモングラウンドの形成はこれまで数多く検討されてきた．Tolzin らは，人間のジェスチャー情報を利用するなど，言語および非言語の文脈情報を活用することで人とエージェント間の共通基盤の構築を容易にするフレームワークを提案した [23]．また，referential game のような 2 者間のインタラクションによって共通の意味が創発することを示す研究もある [24, 25].

これまで紹介してきたように，物理と社会に接地するグラウンディングに関する研究はそれぞれ多く存在する．しかし，両者の研究の統合は十分にはなされていない．

2.2 LLM の能力

LLM は，大量のテキストデータによる広範な事前学習を通じて，特定のタスクのために設計されたモデルよりも高度な言語理解を実現する．従来より利用されてきた，word2vec [26] などの小規模な単語分散表現と比較して学習データが非常に多く，文脈情報を考慮した意味推論が可能であり，自然言語処理タスクにおいてより高い精度を持つことが示されている [27]．また，Binz らは GPT-3 [28] の意思決定能力や因果推論を，一般的な認知心理課題により検証し，人間と同等の能力を持つことを示した [29].

こういった LLM は，旧来の記号的な人工知能技術とは異なり，人間が有する定量的な意味を高精度に表現する可能性を示唆する．Xu らは，LLM に人間の発話を入力することで，人間にとって自然なジェスチャーを生成することが可能であるとした [30]．また，Hensel らは，発話文とジェスチャーの組み合わせを少量提示することで，未知の文脈に対しても LLM が適切なジェスチャーを選択することを示した [31]．Sumanathilaca らは，従来手法と比較して LLM が語義曖昧性解消に大きく貢献することを示した [32].

上述した研究の結果は，LLM 上の分散表現に人間が身体的に獲得してきた言語から数量的意味が含有される可能性を示しており，記号と数量を人間の感覚にあわせて変換する有効な機構と言える．この考えに基づき，本研究では LLM を利用することで，対話時に発話される単語の大きさをジェスチャーに反映させる手法を採用する．

2.3 認知モデルと認知アーキテクチャ

認知科学の領域では，人の認知を理解・予測する方法として認知モデリングというアプローチがとられる．認知モデリングでは，人の認知に関するメカニズムや認知プロセスを近似することで認知モデルを構築し，認知モデルの振る舞いや内部状態から人の認知プロセス，内部状態を推定する．例えば，人間の意思決定に関する認知モデリング研究として Gonzalez らによるもの [33] がある．この研究は,「人間は蓄積した対話の記憶や意思決定の状況などを基に学習する」という Instance-based learning theory (IBLT) に基づき，状況に対して想起された過去の事例を参照することで意思決定を行う認知モデルを構築し，人間のデータと近似することを示した．

こういった認知モデルを開発するための基盤として，認知アーキテクチャがある．認知モデルは，モデル構築者による仮定の自由度が高い．認知アーキテクチャは，認知モデリングの方法を共有し，共通の基盤の上での議論を可能にするためのプラットフォームである．認知アーキテクチャという共通基盤に個人の認知特性やタスク固有の知識を加えることで人間の認知プロセスのモデル化が実現される．本研究では，対話における人間の認知プロセスをモデル化するために認知アーキテクチャを用いる．これまで多くの認知アーキテクチャが開発されてきたが，ここでは代表的な認知アーキテクチャである ACT–R を利用する．ACT–R は思考や記憶に関する心理学実験の知見に基づき，複数のモジュールを持つプロダクションシステムとして実装されている．ACT–R の詳細な解説は [34, 35, 36] を参照されたい．

ACT–R の記憶プロセスは，各記憶に付与される活性値によって制御される．この活性値は個々の記憶に付与される数値であり，いくつかの要素の加算として得られるものである．活性値の項のひとつであるベースレベル活性値は，記憶の学習と忘却を表現する．西川らは，しりとりを行うモデルを構築し，ベースレベル活性値の調整によって，言語発達過程の音の処理の誤りやその抑制を表現した [9]．もう一つの項である活性化拡散は，文脈の効果をモデル化する際に重要な概念であり，対話の連続性を再現することが可能である．例えば Dix らは，web サイトから取得する情報の選択に活性化拡散を適用し，文脈に沿った情報取得が可能となる認知モデルを構築した [37].

本研究では，西川らによって構築されたしりとりモデル [9] を基盤とし，シンプルな対話の形式としてしりとりを扱う．そして，しりとりモデルに活性化拡散パラメータを追加し，操作することで発話される単語のカテゴリを連続させ，対話における思考の連続性を表現する．これにより，人間の対話の特性とも言える時間的文脈を設計することができるかを確かめる．

3 提案手法

図 1 に提案手法の概観を示す．しりとりモデルの活性化拡散パラメータを制御し，出現する単語のカテゴリを連続させることで思考の連続性を表現する．そして，Llama3 モデルを用いて前の単語の大きさと相対的に比較することで動作の連続性を表現する．

3.1 動作の連続性

人間は様々な数量的イメージに基づいてジェスチャーを生成するが，本研究では対象の大きさを表現するジェスチャーに着目する．大きさに関する数量的意味は多くの研究で利用されているものであり [38, 39]，本研究でも大きさに着目することで，様々な数量的意味を反映した人工物の開発へ向けた第一歩とする．

筆者らの先行研究 [20] では，代表的な単語分散表現である word2vec を用い，単語の大きさに関する数量的意味をロボットの腕と関節のパラメータに対応づけることで記号接地した人工物の作成を試みた．このアプローチにより，数量的意味と身体的イメージの対応づけが人間に自然な印象を与える可能性が確認されたものの，その効果は限定的であった．これは，小規模な単語分散表現を基盤とした手法では，人間と同様の記号と数量の変換機構を十分に再現できなかったことに起因すると考えられる．したがって，本研究では Meta 社の提供する LLM である Llama3 の 8B モデルを活用し，より精度の高い大きさに関する数量的意味を抽出する．

3.1.1 相対参照

提案手法では，単語の大きさを，前に出現した単語と比較することで動作の連続性を表現する．本研究ではこれを “相対参照” とする．なお，今回利用した Llama3 の 8B モデルは日本語に対応しておらず，正確な回答を得るためにプロンプトは英語で構成された．プロンプトには，事前知識として先行研究 [18] において得ら

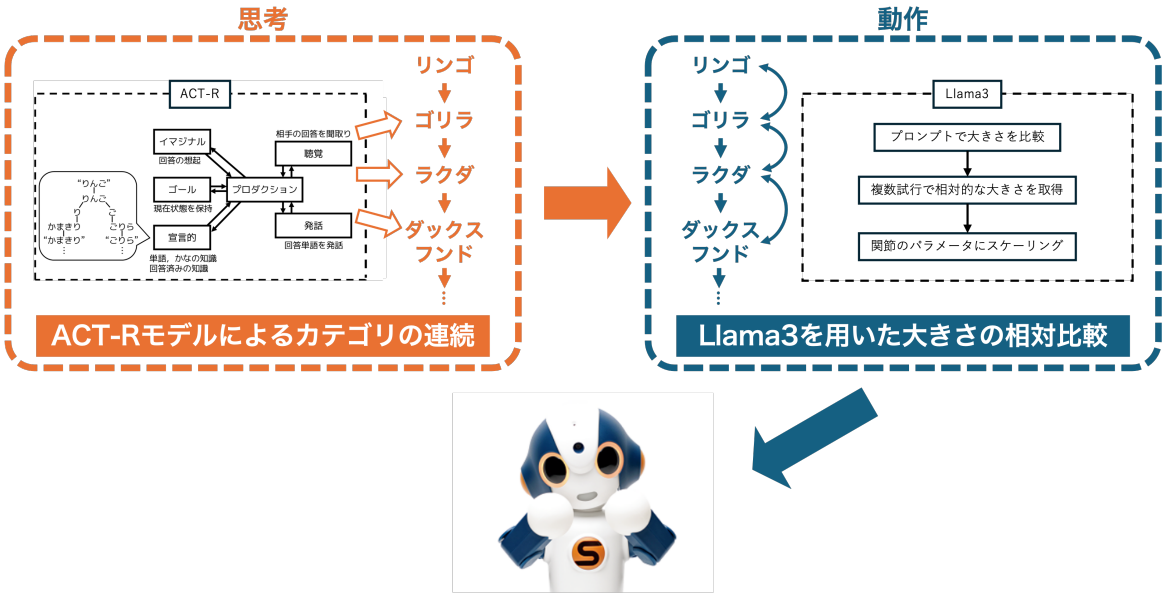

図 1: 提案手法の流れ

れた人間の認識を提示する（例えば,「ゾウはアリよりも大きい」）．そして,「提示された事前知識を参照して入力される 2 つの単語の大きさを比較して欲しい」という指示を与える．入力された 2 つの単語は,「後の単語は前の単語よりも大きいか？」という質問で比較され,「はい」か「いいえ」で回答される．回答はダミー変数に変換され,「はい」と回答された場合には 1,「いいえ」と回答された場合には 0 を対応づける．そして，各単語を相対参照により 5 回評価し，その平均値を最終的な大きさとする．

3.1.2 絶対参照

時間的文脈を含む相対参照に対する対照条件として，単語の大きさを独立して絶対的に判別する “絶対参照” を定義する．なお，絶対参照においてもプロンプトには，事前知識として先行研究において得られた人間の認識を提示する．絶対参照では単語を相対的に比較するのではなく，大きい単語と小さい単語をそれぞれリストとして提示する．各単語の大きさは 1 10 の 9 段階で絶対的に評価される（10 が最大，1 が最小を示す）．そして，各単語を絶対参照により 5 回評価し，その平均値を最終的な大きさとする．

実験では，動作要因として絶対参照と相対参照を設定し，それぞれで生成したロボットのジェスチャーを比較することにより 1 つ目のリサーチクエスチョンを検証する．ジェスチャーの生成手順に関する詳細は 4.2.1 節で示す．

3.2 思考の連続性

先行研究 [9] において，西川らにより ACT–R [8] を用いたしりとりのモデルがすでに実装されている．このモデルは ACT–R のパラメータを調整することでその振る舞いを設定することができる．本研究では，しりとり中に出現する単語の身近さと，対話の文脈に相当する単語の連想にまつわるパラメータによって思考の連続性を表現する．

3.2.1 モデルの構成

本研究で構築したモデルの概観を図 2 に示す．このモデルには，個人に対応するエージェント（破線で囲まれた範囲）が含まれ，交互に単語を回答することでしりとりを繋げる． エージェント中のボックスは ACT–R の各モジュールに対応する．以下では，このような ACT–R のモジュール構造のなかで，しりとり単語系列を特徴づけると考えられる宣言的記憶の設定と，その検索に関するパラメータを示す．

3.2.2 モデルの宣言的記憶

ACT–R の宣言的モジュールを用いて，しりとりの遂行に必要な記憶をモデル化する．ACT–R の宣言的モジュールにおいて，記憶はチャンクと呼ばれる構成要素からなる．本モデルでは，各単語の発音（音のパターン）と表記をつなぐ「単語チャンク」，各モーラの発音

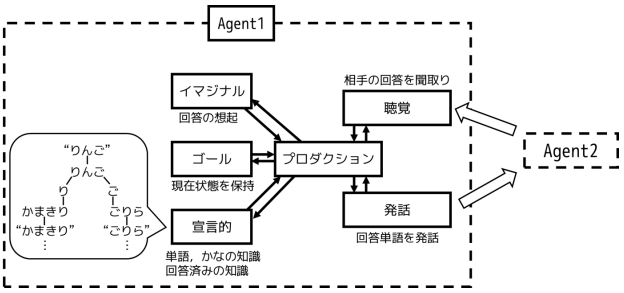

図 2: モデル概観

と表記を表す「モーラチャンク」，単語を構成するモーラを示す「単語–モーラの関連チャンク」の 3 種類を持つ．宣言的記憶では，これら 3 種類のチャンクがネットワークとして接続される．図 2 の宣言的モジュールから出る吹き出しはこの構造を示している．エージェントはこのネットワークをたどることで，前の単語とモーラを介してつながる答えの単語を検索する．

3.2.3 モデルのパラメータ

ACT–R のモジュールは，数値的なパラメータによって調整される[1]．本研究においては，宣言的記憶を検索することに関わるパラメータである活性値，

$$A_i = B_i + S_i + \epsilon_i \tag{1}$$

が重要となる．活性値は学習と忘却（B_i, ベースレベル活性値と呼ばれる），文脈（S_i, 活性化拡散と呼ばれる）などに対応する複数の項と，各検索のタイミングで付与されるノイズの加算として定義される．本研究においては，しりとりの単語系列の特徴として，使用される単語の身近さを表現するためにベースレベル活性値，対話中の連想を表現するために活性化拡散に注目する．

単語の身近さを表現するベースレベル活性値は以下の式により設定される．

$$B_i = \ln\left(\frac{n}{1-d}\right) - d \times L + \beta_i \tag{2}$$

B_i はチャンク i のベースレベル活性値を表す．ベースレベル活性値は，そのチャンクが参照された回数 n, 初めてチャンクが参照されてからの経過時間 L，減衰率 d およびオフセットの数値 β_i から算出される．d と β_i は，それぞれひとつの数値によって設定されるパラメータであり全てのチャンクに対して共通の影響を及ぼす．チャンク（単語）ごとの身近さは n および L を調整することで可能となる．

[1]詳細は ACT–R のマニュアル [40] を参照

式 1 の活性化拡散 (S_i) は，他のモジュールが現在保持しているチャンクからの影響を表し，文脈の効果に対応する．

$$S_i = \sum_k \sum_j W_{kj} S_{ji} \tag{3}$$

k は，活性値が拡散するようパラメータによって設定されたモジュールであり，j は活性化ソースと呼ばれる，k が保持するチャンクと検索キューが共通して持つ値である．チャンク i がその要素として値 j を持つとき，モジュール k に設定された活性化量を活性化ソース j の数で割った値が W_{kj} となる．$S_j i$ は活性化ソース j とチャンク i への関連付けの強さであり，式 4 で表される．

$$S_{ji} = S - \ln\left(\frac{1 + slots_j}{slotsof_{ji}}\right) \tag{4}$$

ここで，S は連想強度の最大値を設定するパラメータである．第 2 項は Fan 効果 [41] と呼ばれ，宣言的記憶全体に含まれる値の数 $slots_j$ と，値 j をもつチャンク i のスロット数 $solotsof_{ji}$ から計算される．

これらの要素を項として算出される活性値は，そのチャンクの想起確率，および想起に要する時間に影響する（活性値の高いチャンクほどより多くの頻度で素早く想起される）．さらに，活性値が閾値よりも低いチャンクは，長時間の検索の試行が行われた後に失敗する．

3.2.4 思考の連続性の表現

上述した設定を持つ 2 体のモデルを用いて思考の連続性を表現する．モデルが持つ語彙には，幼児・児童の連想語彙表 [42] に含まれる反応語の 2,436 語を利用した．この表は，参加者が刺激語（たとえば「動物」）に対して知っていることばを 40 分間にできるだけたくさん回答するという課題から作成された．参加者は 3 歳児，4 歳児，5 歳児，6 歳児，小学 1 年生，小学 2 年生，小学 3 年生，小学 4 年生，成人に区別された．

この語彙に対応するチャンクのそれぞれに対して，単語の身近さの表現のために，ベースレベル活性値を設定する．連想語彙表（「動物」に対して 46 人の 3 歳児が「象」を回答したなどといった情報）を元に，式 2 の L と n に代入する値を計算する．本モデルでは成人を想定して L を設定した（たとえば，3 歳児の回答した単語は，$L = 21 - 3$ 年間 $= 567,648,000$ 秒）．さらに，単語 w に対する n（n_w）は

$$n_w = \sum_i (N_{wi} \times 365) \tag{5}$$

表 1: しりとりの生成例

単語	カテゴリ
リンゴ	果物
ゴリラ	動物
ラクダ	動物
ダックスフンド	動物
ドラ猫	動物
コウモリ	動物
リス	動物
スカンク	動物
熊	動物
マンモス	動物
水牛	動物
馬	動物
マムシ	動物
シェパード	動物
ドレス	衣服
スカート	衣服
トレーニングウェア	衣服
アンサンブル	衣服

にあてはめた．ここで，i は参加者の年齢である．N_{wi} は，単語 w を回答した年齢 i の参加者の人数を示す．さらに 1 人が 1 日に 1 度その単語を参照したという仮定に基づき 365 をかけた．

また，しりとり中の単語のカテゴリの一致を表現するために，活性化拡散を導入する．モデルが持つ語彙のそれぞれに，幼児の連想語彙表 [42] における刺激語をカテゴリとして紐づけた．しりとりにおける一つ前の単語のカテゴリをゴールモジュールに配置することで，単語検索の際に単語の意味に関する連想，あるいは文脈の効果が見られることを狙う．本研究では，式 1 の B_i と S_i の影響が同程度になるよう調整し，式 4 の S を 35 とした．これらの設定に基づいて生成したしりとりの一例を表 1 に示す．

表 1 を参照すると，果物カテゴリの後に動物カテゴリの単語が続き，最後に衣服カテゴリの単語が出現していることが分かる．本研究では上述した設定に基づいてしりとりを生成する，“活性化拡散あり” 条件で思考の連続性を表現する．対照群として活性化拡散をオフにした “活性化拡散なし” 条件を設定し，2 つの条件を思考要因とする．思考要因内の 2 つの水準を比較することにより，2 つ目のリサーチクエスチョンを検証する．また，3.1 節で定義した動作の連続性に基づき，“相対参照” 条件によりジェスチャーを，“活性化拡散あり” 条件によりしりとりを生成したときの効果を検討することで 3 つ目のリサーチクエスチョンを検証する．

4 実験

4.1 目的

本実験では，思考要因（活性化拡散あり vs. 活性化拡散なし）×動作要因（相対参照 vs. 絶対参照）の 4 条件下で参加者の印象を比較した．そして，実験の結果から 1 節で説明したリサーチクエスチョンを検討する．この目的に基づき，動画ごとにアンケートを収集した．アンケートでは，2 体のロボットがジェスチャーを交えて行うしりとりの自然さや心的状態の帰属に関する印象の評価を実施した．

4.2 方法

4.2.1 材料

Llama モデルを用いて算出した単語の大きさに応じたジェスチャーの生成手順を以下に示す．

1. **最大・最小動作の設定**
   3.1 節で算出した大きさと，身体の各部位から構成される姿勢の対応付けを行う．そのために，最小の姿勢と最大の姿勢を定義する．“相対参照” では，最小の姿勢と最大の姿勢の範囲をさらに 5 つの範囲に分割する．分割された 1 つの範囲（最小の単語は 0，最大の単語は 1）を基準として，各単語の大きさを 0 から 1 の範囲に配置する．“絶対参照” では，この姿勢（最小の単語は 1，最大の単語は 10）を基準として，各単語の大きさを 1 から 10 の範囲に配置する．

2. **各関節のパラメータ計算**
   姿勢を構成する各関節の関節角度に上記のスケーリングを適用する．なお，今回は “絶対参照” と “相対参照” の最大値と最小値を均一にするため，各しりとり系列で動作量を定義する．すなわち，各しりとり系列において最大の単語を最大の姿勢に，最小の単語を最小の姿勢に対応づける．

3. **ジェスチャーの生成**
   ステップ 2 で得られた値に基づいてジェスチャーを生成する．なお，ジェスチャーは発話と同時に行われるものとする．また，各ロボットは自身が前に発話した単語のジェスチャーを保持し，その位置から次に発話する単語のジェスチャーを行う．そのため，“相対参照” では，ステップ 1 で分割された 1 つの範囲に発話する単語を当てはめ，自身が前に発話した単語に対応する各部位の位置から対応する大きさだけ動作する．

上記の手順を具体化するために，Vstone[2] 社の小型コミュニケーションロボット Sota を用いた．Sota の体の動きは 9 つの関節（胴体 1 つ、首 3 つ、肩 2 つ、腕 2 つの関節）で制御されている．これらの関節の角度と速度を制御することで，Sota は様々な動きを生み出すことができる．また，Sota には発話機能があり，ジェ

[2] https://www.vstone.co.jp/english/index.html

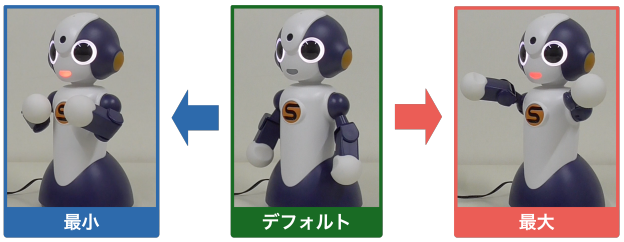

図 3: Sota のジェスチャー

スチャーを行いながら任意の言葉を話すことが可能である．

本研究では，Sota の腕と肩のパラメータを Llama3 モデルの出力した値に応じて制御することでジェスチャーを生成した．ジェスチャーの生成例を図 3 に示す．しりとり開始時における Sota の姿勢は，図の中心に示されており，肩を下げ，腕を少し曲げている状態を設定した．この状態から腕と肩の関節のパラメータが変更され，単語の大きさに対応するジェスチャーが生成される．なお，上記の生成手順で説明した，最小の姿勢と最大の姿勢は，Sota の右手と左手の距離が最小となる位置と最大となる位置に従って定義される．

実験には「リンゴ」から始まるしりとりの系列を使用した.「しりとりは“しりとり”の語尾である“り”から始まる」と仮定し，“り”から始まる単語の中で最も高いベースレベル活性値を持つ単語である“リンゴ”を開始単語として設定した．

しりとりの系列は活性化拡散あり・なし双方で 3 つずつ作成し，思考要因の水準が同じである場合は，動作要因の両水準で同じしりとりの系列を使用した．また，今回の実験ではチャンクの検索失敗，既出語の回答，語尾「ん」の単語の回答という終了条件以外にも 3 分の時間制限を設け，時間に達したしりとりはその時点で終了させた．実験では，しりとりを 10 回試行したうち 2 分以上継続したしりとりの中で最初から 3 つの系列を使用した．

4.2.2 実験デザイン

参加者は動画を観察し，動画ごとにアンケートに回答した．アンケートは 3 つのセクションに分かれており，設問は表 2 の通りであった．程度の評定には 7 段階リッカート尺度を用い，1（最小値）を「全く当てはまらない」または「不自然」とし，7（最大値）を「非常に当てはまる」または「自然」とした．また，注意散漫な参加者の回答を除外するために，追加で「行動が一貫している」というダミー質問が設けられ，回答画面の教示で回答する値が指示された．

ロボットの印象評価は，参加者がロボットに対して心的状態を帰属したかどうかを調べるために設けられた．印象評価項目は，対象への心の帰属（エージェンシーの帰属）を扱った先行研究から流用した [43, 44, 45]．なお,「一体感がある」と「活発である」という項目に関しては，複数ロボットの印象評価を扱った先行研究で扱われた項目である [46, 47]．2 体のロボットの対話の印象を評価するために，これらを設定した．

4.2.3 参加者と手順

本実験にはクラウドソーシングサイト Lancers のタスク方式にて募集した 300 名が参加した．ここで参加者を 300 名としたのは，予備実験の結果を踏まえ，本実験においては約 300 名分のデータが適切な検定力を持つと推定されたためである．それぞれの参加者は，依頼画面に表示された説明に同意した上で実験に参加した．実験参加者には，実験協力への報酬として Lancers のシステム手数料を除き，手取り（税抜）100 円を支払った．

まず，参加者は Lancers の依頼概要に記載されたリンクから専用のサイトを訪問した．サイトでは，まず評価方法や，動画の視聴に関わる注意事項（動画の表示から 3 分以上の時間をおいた後に，回答の送信が可能になるなど）に関する注意点を説明し，最後に注意すべき内容の確認クイズと動画の再生チェックを実施した．参加者はクイズに正常に回答した後，Lancers のユーザ ID を入力して回答画面に移った．

回答画面には,「リンゴ」から始まる 3 つのしりとり系列それぞれで作成された 4 つの条件の動画，計 12 本の動画のうち 4 つの条件からそれぞれ 1 本，計 4 本がそれぞれ別のページに配置された（提示される条件の順序と動画の種類はカウンターバランス）．動画は YouTube にアップロードしたものを埋め込む形式で表示された[3]．

動画の表示ページにおける教示では，先述した回答時間の下限に関する注意点の説明に加えてダミー質問で回答する数字が指示された．参加者は動画の視聴中に 4.2.2 に示した 2 つの設問に回答し，4 つ目の動画の後には事後アンケートに回答した．全ての質問に回答した後表示されたアンケート ID を Lancers のタスクページ下部にあるテキストボックスに記入することで実験を終了した．

表 2: アンケートの設問

1. **ロボットの印象評価**
   ロボットの印象について，以下の各項目に対してロボットがどの程度当てはまるかを 7 段階で直感的に評価してください．
   (a) 行動が複雑である
   (b) 行動が規則的である
   (c) 行動が予測できる
   (d) 行動が意図的である
   (e) 知的である
   (f) 心がある
   (g) 一体感がある
   (h) 活発である

2. **しりとりと動作の自然さ評価**
   ご覧いただいた動画において，しりとりの単語系列とロボットの動作はそれぞれどの程度自然だと感じましたか．7 段階で直感的に評価してください．
   (a) ロボット単体が発する単語
   (b) ロボット単体による姿勢あるいは動作
   (c) 2 体のロボットが発する単語と単語のつながり
   (d) 2 体のロボットによる姿勢と姿勢のつながり

5 結果

5.1 取得データ

応募された 300 名のうち 70 名は一部回答が欠損していたもしくはログを正常に取得できなかった．また，57 名はアンケート中に設定されたダミー質問に誤って回答した．そのため，残りの 173 名の回答から自然さの評定と心的状態の帰属に関する評定の平均値を算出した．

それぞれの評定値の平均を図 4 と図 5 に示す．エラーバーは標準誤差を表す．これらの値を利用して冒頭で述べた 3 つのリサーチクエスチョンを検討する．分析には，対応のある 2 要因分散分析（ANOVA）[動作要因（活性化拡散あり vs. 活性化拡散なし）×思考要因（相対参照 vs. 絶対参照）] を各評定項目ごとに実施した．なお，有意水準は 0.05 とした．交互作用が有意となった場合は，単純主効果の検定を実施した．グラフ中にて得られた主効果は，対応する項目間, あるいは複数の項目をまたぐグループ間の差を結ぶ直線として示している．

5.2 思考の連続性

1 つ目のリサーチクエスチョンを検討するため，思考要因の主効果に着目した．心的状態の帰属に関する評定において，活性化拡散の効果を観察できた項目は，「行動が意図的である」 $(F(1,171) = 3.95, p = 0.05)$,「知的である」 $(F(1,171) = 15.08, p < 0.01)$,「心がある」 $(F(1,171) = 14.34, p < 0.01)$,「活発である」 $(F(1,171) = 4.24, p = 0.04)$ であった．これらの項目で活性化拡散ありが活性化拡散なしを上回ったことから，思考に時間的文脈をいれることで，より意図が感じられ，知的な印象を与える対話を設計できる可能性が示された．

[3]https://youtube.com/playlist?list=PLWFbVU0ku1AjrQLDzFq_9BEFcIc-BmIQJ&si=RNrBXDodNDDK3vOL

5.3 動作の連続性

2 つ目のリサーチクエスチョンを検討するため，動作要因の主効果に着目した．相対評価が絶対評価を上回った項目は,「行動が予測できる」 $(F(1,171) = 23.73, p < 0.01)$,「行動が規則的である」 $(F(1,171) = 26.79, p < 0.01)$ であった．一方で,「行動が複雑である」 $(F(1,171) = 64.71, p < 0.01)$,「行動が意図的である」 $(F(1,171) = 9.63, p < 0.01)$,「知的である」 $(F(1,171) = 10.47, p < 0.01)$,「心がある」 $(F(1,171) = 30.32, p < 0.01)$,「活発である」 $(F(1,171) = 114.57, p < 0.01)$ などの項目では，絶対参照が相対参照を上回った．これらの結果，動作における連続性は予測や規則性を高めるものの，他のエージェンシー認知に関する項目については，ネガティブな印象となった．6 節にて，この点に関して考察する．

5.4 思考と動作の連続性の一致

3 つ目のリサーチクエスチョンを検討するため，動作要因と思考要因の交互作用に着目する．交互作用は

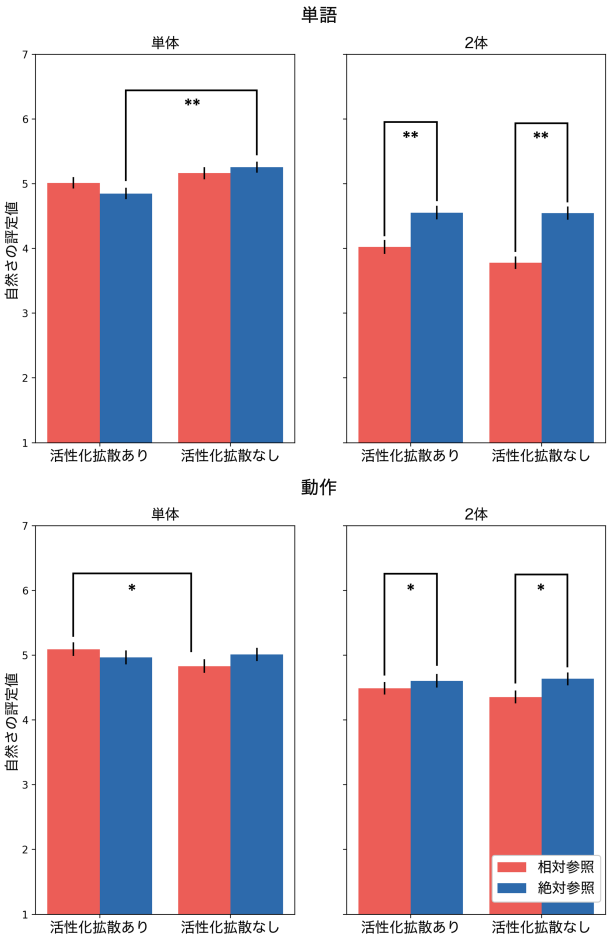

図 4: 自然さの評定の平均値（エラーバーは標準誤差）

単体単語 ($F(1,171) = 4.76, p = 0.03$)，単体動作 ($F(1,171) = 5.13, p = 0.02$) おいて確かめられた．単体単語においては，絶対参照において活性化拡散が不自然な印象を与えている ($F(1,171) = 18.49, p = 0.03$) こと，相対参照においてそのような不自然さは認められないことが示される．また単体動作においては，相対参照にて活性化拡散ありが活性化拡散なしに対して自然な印象となっている ($F(1,171) = 4.31, p = 0.04$) ことから，活性化拡散が加わることで，動作の自然さが向上することが示される．

6 考察

本研究は，1 節で設定された 3 つのリサーチクエスチョンを検討した．1 つ目のリサーチクエスチョン（思考の連続性は対話の印象にどのような影響を与えるのか？）については，思考要因の主効果に着目することで回答した．本実験の結果から，活性化拡散によるカテゴリの連続が，対話における思考の連続性を表現し得ることが確かめられた．

2 つ目のリサーチクエスチョン（動作の連続性は対話の印象にどのような影響を与えるのか？）については，動作要因の主効果に着目することで回答した．本実験の結果では，相対的な大きさの比較が規則性，予測可能性，一体感の喚起に繋がることが確かめられた．この点から，相対参照における 2 体のロボットの動作の大きさの連続性を参加者が認識できたことが示される．しかし，他の印象評価の結果からは，そのような動作の連続から冒頭の議論で述べたような「生きている」という感覚が生じたとは言えない結果となった．

本研究において相対参照が，エージェンシー知覚の項目にネガティブな影響を及ぼした理由は，スケーリングに起因する可能性がある．本研究におけるスケーリングは，結果として，絶対参照の動作量が相対参照を上回るものとなっていた．このことが「行動が複雑である」や「行動が活発である」などの項目における結果を引き起こしたと考えられる．さらに，こういった動作量の大きさは，ジェスチャーとしての自然さやエージェンシー知覚にも影響することが著者らの先行研究でも確かめられている [20]．そのため,「知的である」や「心がある」などの項目で相対参照が絶対参照を下回ることとなったと解釈できる．

3 つ目のリサーチクエスチョン（動作と思考の連続性の一致は対話の印象にどのような影響を与えるのか？）については，思考要因と動作要因の交互作用に着目することで回答した．本実験の結果では，単体のロボットが発する単語は活性化拡散と相対参照という思考の連続と動作の連続が組み合わさることによって，より自然になることが確かめられた．この結果より，マクニールなどが指摘するようなジェスチャー生成における思考と動作の不可分性が示唆される．

なお，絶対参照における単語の自然さにおいて，活性化拡散ありが，活性化拡散なしを上回った原因として，式 1 における S_i が加わることで，B_i の効果が相対的に低下したことが考えられる．S_i が加わることで，利用頻度が相対的に低い単語が発話される確率が高まる．つまり，子どもを想起させる Sota の外見とは整合していない難解な単語が発話されたことで，単語の不自然さの印象が強くなった可能性がある．なお，ACT-R では活性値が低い単語を想起する際には，想起時間が延長される機能が含まれているが，本研究のしりとりモデルでは適切に働いていなかった．

上記のような活性化拡散が加わることによる対話の不自然さは，相対参照によって減じられる．本研究において得られた思考と動作の相乗効果は，大きさの比較を異なるカテゴリ（例えば文房具と果物）で行うことが困難であることに由来すると考えられる．大きさなどの定量的な意味の比較は，通常は同一カテゴリ内で行われるものであり，カテゴリをまたいで行われる

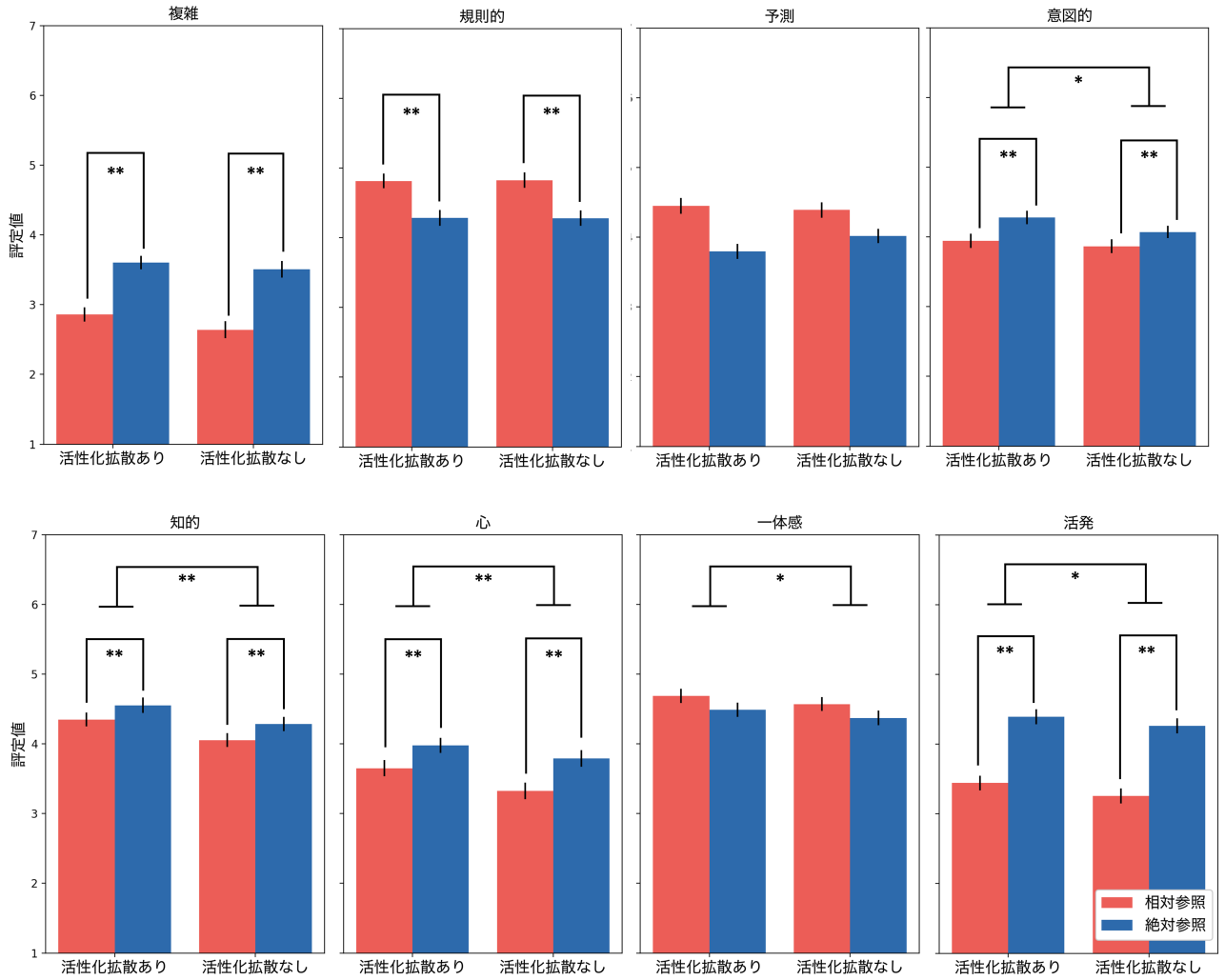

図 5: 心的状態の帰属に関する評定の平均値（エラーバーは標準誤差）

ことは稀である．そのため，絶対参照や相対参照に関わらず，異なるカテゴリの間での大きさの連続は不自然な印象が生じ，カテゴリの連続する発話系列において，相対的に自然さが高まったと考える．

7　結論

本研究では，対話を時間的文脈の中で生じる人間の活動と捉え，思考と身体の異なるチャンネルにおける連続性を備えたロボット間のインタラクションをデザインした．結果として，それぞれのチャンネルにおける連続性が，狙った対話の文脈を観察者に生起させることが確認された．さらに，それぞれのチャンネルの具体的な相乗効果（カテゴリの連続による自然な大きさの比較）を見出すことにも成功した．対話における時間的文脈の設計は，自然なコミュニケーションの実現において重要である．よって，本研究は Human-Agent Interaction（HAI）や Human-Robot Interaction（HRI）分野における理論的・発展的実践に寄与するものと考える．

本研究にはいくつかの限界がある．対話の自然さ，あるいは対話を通したエージェンシーの生起には，多様な要因が関与する．そのため本研究において導入した活性化拡散や相対参照のような仕組みに付随して生じる負の要因（単語の馴染み深さの毀損，動作量の低減）を低減する工夫が必要となる．このような限界を乗り越えるために，使用するロボットとジェスチャーの表現形式を改善することが考えられる．本研究では最小と最大の姿勢を，小型コミュニケーションロボット Sota の右手と左手の距離に基づいて定義した．しかし，これらの姿勢が人間にとって小さいもしくは大きいことを表現するかどうかは不明であり，動作の連続性の評価が良くなかった原因である可能性も否定できない．したがって，人間が小さいもしくは大きいと認識するジェ

スチャーに関するさらなる検討が求められる．

また，しりとりモデルの語彙についても改善が必要である．本研究では連想語彙表の刺激語をカテゴリとして代用することで思考の連続性を表現した．しかし，回答者に幼児が含まれていることもあり本来の単語のカテゴリ情報であるとは言えない．例えば，連想語彙表では「赤い」という形容詞が「楽器」という刺激語で回答されていた．回答者は 1 名であったが，活性化拡散の効果により，実験で使用したしりとり系列でも出現していた．したがって，単語とその単語の所属カテゴリが紐づいたデータベースを利用することで，活性化拡散の効果を加えたしりとりに出現する単語の自然さは向上する可能性がある．

以上のような問題への対処に加え，大きさ以外の数量的意味（速さや丸さ等）に基づいたジェスチャーの生成を検討することで，より人間の対話に近い人工物による対話の実現を目指す．本研究のアプローチは，身体と言語に関する認知科学的な理論を分解する試みである．この基礎研究を通じ，最終的には，時間的文脈を設計する人間の対話に基づいた，人間とのシームレスなインタラクションが可能な高度な人工物の開発に貢献すると考えている．

参考文献

[1] 細川英雄. 対話をデザインする―伝わるとはどういうことか. 筑摩書房, 2019.

[2] 喜多壮太郎. ひとはなぜジェスチャーをするのか. 認知科学, Vol. 7, No. 1, pp. 9–21, 2000.

[3] 池上高志. 動きが生命をつくる: 生命と意識への構成論的アプローチ. 青土社, 2007.

[4] David McNeill. Hand and mind: What gestures reveal about thought. Chicago: Chicago University Press, 1992.

[5] David McNeill. Growth points in thinking-for-speaking. Language and Gestures/Cambridge University Press, 2000.

[6] Youngwoo Yoon, Woo-Ri Ko, Minsu Jang, Jaeyeon Lee, Jaehong Kim, and Geehyuk Lee. Robots learn social skills: End-to-end learning of co-speech gesture generation for humanoid robots. In 2019 International Conference on Robotics and Automation (ICRA), pp. 4303–4309. IEEE, 2019.

[7] Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

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

[9] Jumpei Nishikawa and Junya Morita. Cognitive model of phonological awareness focusing on errors and formation process through shiritori. Advanced Robotics, Vol. 36, No. 5-6, pp. 318–331, 2022.

[10] Mary Lou Maher, Dan Ventura, and Brian Magerko. The grounding problem: An approach to the integration of cognitive and generative models. In Proceedings of the AAAI Symposium Series, Vol. 2, pp. 320–325, 2023.

[11] Jeff Hawkins. A thousand brains: A new theory of intelligence. Basic Books, New York, 2021.

[12] Steven Pinker. The stuff of thought: Language as a window into human nature. Penguin, London, 2007.

[13] Stevan Harnad. The symbol grounding problem. Physica D: Nonlinear Phenomena, Vol. 42, No. 1-3, pp. 335–346, 1990.

[14] 今井むつみ, 佐治伸郎, 山崎由美子, 浅野倫子, 渡邊敦司, 大槻美佳, 松井智子, 喜多壮太郎, 安西祐一郎, 岡田浩之, 橋本敬, 増田貴彦. 言語と身体性. 岩波書店, 2014.

[15] George Lakoff and Mark Johnson. Metaphors we live by. University of Chicago press, 2008.

[16] Carolyn Saund, Haley Matuszak, Anna Weinstein, and Stacy Marsella. Motion and meaning: Data-driven analyses of the relationship between gesture and communicative semantics. In Proceedings of the 10th International Conference on Human-Agent Interaction, pp. 227–235, 2022.

[17] Mingtong Li, Suzanne Aussems, and Sotaro Kita. Is adults' ability to interpret iconicity shared between the spoken and gestural modalities? In Proceedings of the Annual Meeting of the Cognitive Science Society, 第 46 巻, 2023.

[18] 佐々木康佑, 西川純平, 森田純哉. 単語分散表現を用いた概念の数量的意味獲得. 2023 年度第 40 回

日本認知科学会大会論文集, pp. 511–514. 日本認知科学会, 2023.

[19] 佐々木康佑, 西川純平, 森田純哉. 文脈に応じた synset 選択に基づく単語の定量的意味抽出の検討. 2024 年度第 41 回日本認知科学会大会論文集, pp. 437–440. 日本認知科学会, 2024.

[20] Kosuke Sasaki, Jumpei Nishikawa, and Junya Morita. Evaluation of co-speech gestures grounded in word-distributed representation. Frontiers in Robotics and AI, Vol. 11, p. 1362463, 2024.

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

[22] Herbert H Clark and Edward F Schaefer. Contributing to discourse. Cognitive science, Vol. 13, No. 2, pp. 259–294, 1989.

[23] Antonia Tolzin and Andreas Janson. Mechanisms of common ground in human-agent interaction: A systematic review of conversational agent research. In HICSS, pp. 342–351, 2023.

[24] Gautier Dagan, Dieuwke Hupkes, and Elia Bruni. Co-evolution of language and agents in referential games. In Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, pp. 2993–3004. Association for Computational Linguistics, 2021.

[25] Roberto Dessì, Diane Bouchacourt, Davide Crepaldi, and Marco Baroni. Focus on what's informative and ignore what's not: Communication strategies in a referential game. arXiv preprint arXiv:1911.01892, 2019.

[26] Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. A neural probabilistic language model. Advances in neural information processing systems, Vol. 13, , 2000.

[27] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. arXiv preprint arXiv:1301.3781, 2013.

[28] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. Advances in neural information processing systems, Vol. 33, pp. 1877–1901, 2020.

[29] Marcel Binz and Eric Schulz. Using cognitive psychology to understand gpt-3. Proceedings of the National Academy of Sciences, Vol. 120, No. 6, p. e2218523120, 2023.

[30] Chenghao Xu, Guangtao Lyu, Jiexi Yan, Muli Yang, and Cheng Deng. Llm knows body language, too: Translating speech voices into human gestures. In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pp. 5004–5013, 2024.

[31] Laura Birka Hensel, Nutchanon Yongsatianchot, Parisa Torshizi, Elena Minucci, and Stacy Marsella. Large language models in textual analysis for gesture selection. In Proceedings of the 25th International Conference on Multimodal Interaction, pp. 378–387, 2023.

[32] Deshan Sumanathilaka, Nicholas Micallef, and Julian Hough. Assessing gpt's potential for word sense disambiguation: A quantitative evaluation on prompt engineering techniques. In 2024 IEEE 15th Control and System Graduate Research Colloquium (ICSGRC), pp. 204–209. IEEE, 2024.

[33] Cleotilde Gonzalez, Javier F Lerch, and Christian Lebiere. Instance-based learning in dynamic decision making. Cognitive Science, Vol. 27, No. 4, pp. 591–635, 2003.

[34] 森田純哉. 機械学習時代における認知的学習モデルの役割 – ACT-R による学習モデルの事例と支援システムへの搭載–. 人工知能, Vol. 35, No. 2, pp. 223–232, 2020.

[35] 寺尾敦. 認知アーキテクチャの理論による脳の構造と機能の解明. 電子情報通信学会誌, Vol. 98, No. 12, pp. 1083–1090, 2015.

[36] Frank E. Ritter, Farnaz Tehranchi, and Jacob D. Oury. Act-r: A cognitive architecture for modeling cognition. WIREs Cognitive Science, Vol. 10, No. 3, p. e1488, 2019.

[37] Alan Dix, Akrivi Katifori, Giorgos Lepouras, Costas Vassilakis, and Nadeem Shabir. Spreading activation over ontology-based resources: from personal context to web scale reasoning. International Journal of Semantic Computing, Vol. 4, No. 01, pp. 59–102, 2010.

[38] Joseph Edward Grady. Foundations of meaning: Primary metaphors and primary scenes. University of California, Berkeley, 1997.

[39] Debora S Herold, Lynne C Nygaard, Kelly A Chicos, and Laura L Namy. The developing role of prosody in novel word interpretation. Journal of Experimental Child Psychology, Vol. 108, No. 2, pp. 229–241, 2011.

[40] Dan Bothell. ACT-R 7.21+ reference manual, 2020.

[41] John Robert Anderson. Retrieval of propositional information from long-term memory. Cognitive Psychology, Vol. 6, No. 4, pp. 451–474, 1974.

[42] 国立国語研究所. 幼児・児童の連想語彙表. 東京書籍, 1981.

[43] Kazunori Terada and Seiji Yamada. Mind-reading and behavior-reading against agents with and without anthropomorphic features in a competitive situation. Frontiers in Psychology, Vol. 8, p. 1071, 2017.

[44] 子安増生, 龍輪飛鳥. 運動図形に対する心的状態の付与に及ぼす図形の種類と運動パターンの効果. 京都大学大学院教育学研究科紀要, Vol. 50, pp. 1–21, 2004.

[45] Carey K Morewedge, Jesse Preston, and Daniel M Wegner. Timescale bias in the attribution of mind. Journal of personality and social psychology, Vol. 93, No. 1, p. 1, 2007.

[46] 一條剛志, 棟方渚, 小野哲雄. 複数ロボットの対話の活性度を用いた注意誘導システムの研究. HAI シンポジウム 2015, pp. G–17, 2015.

[47] 水丸和樹, 坂本大介, 小野哲雄. 複数ロボットの発話の重なりによって創発する空間の知覚. HAI シンポジウム 2017, pp. G–15, 2017.