大規模言語モデルにおける方言生成過程の内部機序分析

平川 稜真， 坂井 優介， 上垣外 英剛， 渡辺 太郎
奈良先端科学技術大学院大学 (NAIST)
hirakawa.ryoma.hq9@naist.ac.jp
{sakai.yusuke.sr9,kamigaito.h,taro}@is.naist.jp

概要

大規模言語モデル（LLM）は，方言や文体といった言語内のスタイル差を高精度に生成できる一方，それらの差異がモデル内部で，どのように表現や制御されるかなどの解明は十分ではない．本研究では，LLM における方言生成過程を分析し，同一言語内に存在する文体差（スタイル差）が内部表現でどのように扱われているかを明らかにする．関西方言と東北方言を対象に，複数の言語モデルを用い，Logit Lens により層ごとの方言検出数およびトークン確率の推移を比較した．その結果，いずれのモデルにおいても，方言的特徴は中間層で初めて顕在化した後に一時的に減少し，最終層に向けて再び強化されることがわかった．一方トークン確率は層の深さとともに一貫して上昇するという共通の層構造を示した．これらの結果は LLM 内部では，文体や話法を処理する中間層と，内容に基づく確信度を形成する上位層が分離して存在することを示唆する．

1 はじめに

近年，大規模言語モデル（LLM）の発展により，翻訳や要約，対話生成など多様な自然言語処理タスクで高精度な出力が得られるようになり，多言語間の翻訳や意味対応においても優れた性能を示している．一方で，自然言語には言語間の差異だけでなく，同一言語内における話し方や文体，さらには地域に根ざした方言といった多様性が存在する．しかし，このような言語内のスタイル差を，LLM がどの程度区別し，内部的に理解・表現できているかの内部機構解明は十分に明らかにされていない．さらに，LLM の出力がどのような内部過程を経て生成されるかは依然不透明であり，モデルの透明性・信頼性を高めるためには内部機序の分析が重要な課題となる．この課題意識のもと，近年では LLM の内部表現に着目した研究が進展している．Wendler ら [1] は，多言語モデルが非英語入力に対しても英語に強くバイアスされた内部表現をピボットとして用いる傾向を示した．また，Zhong ら [2] は，日本語特化型モデルの内部表現を分析し，日本語内部における言語的特徴の潜在軸（latent language）の存在を示している．しかし，これらの研究は主として言語間の差異に焦点を当てたものであり，同一言語内の文体や方言といったスタイル差に対して，モデルがどのように内部的に対応しているかについての分析は十分に行われていない．

そこで本研究では，日本語における標準語と方言（関西方言・東北方言）の差異に着目し，LLM がそれらをどのように内部的に表現・生成しているかを分析する．具体的には，Logit Lens [3] により層ごとの出力分布を可視化し，形態素解析および n-gram 照合により方言的語彙・語尾の出現層を定量的に分析する．

実験では，日本語特化 3 モデル（LLM-jp-3-3.7B-Instruct3，Sarashina-2.2-3B-Instruct-v0.1，Swallow-7B-Instruct-v0.1）を用い，関西方言および東北方言の生成過程を比較した．その結果，いずれのモデル・方言においても，中間層付近で検出数が増加したのちに減少し，最終層で再び増加するという共通傾向が確認された．また，トークン確率は層が深くなるにつれて上昇する傾向を示した．これらの結果から，内容ではなく文体・スタイルといった言語的多様性を考慮する層が LLM 内部に存在し，最終層は中間層で形成された方向性を踏まえて出力を確定している可能性が示唆される．本研究は，同一言語内のスタイル差を層構造に基づき可視化・定量化する分析枠組みを提示し，スタイル制御の透明性向上に寄与することを目指す．

2 関連研究

大規模言語モデルの内部機序分析 LLM の内部表現理解に関しては多角的な研究が行われている．

Wendler ら [1] は，多言語モデルが非英語入力に対しても英語を内部のピボットとして用いることを示し，言語間での内部表現の偏りを指摘した．尾崎ら [4] は，プロンプト設計と出力の関係から内部挙動を分析し，主に出力制御の観点で評価を行った．さらに，Huan ら [5] は，LLM が意図的に虚偽出力を生成しうる状況を分析し，自己整合性および倫理的信頼性の課題を提起した．これらの多くは Logit Lens[3] を用いて層ごとの出力分布を観測し，内部表現の遷移を可視化している．

日本語特化モデルと内部言語　日本語を対象とした内部表現分析も進んでいる．Zhong ら [2] は，日本語特化型 LLM の内部表現を分析し，日本語モデル内部における言語的特徴の潜在軸の存在を検証した．一方で，焦点は主に言語間差（英語–日本語）や翻訳方向にあり，同一言語内のスタイル・方言の多様性に踏み込んだ分析は少ない．

3　分析手法と実験設定

3.1　方言生成過程の内部分析

先行研究 [1, 2, 4, 5] は，多言語処理や信頼性評価の観点から LLM の内部表現を明らかにしてきたが，同一言語内の文体・方言といったスタイル差が内部でどのように形成・統合されるかは未解明である．本研究はこの点に着目し，標準語と関西方言および東北方言の違いを対象として，層構造における方言的特徴の出現傾向を Logit Lens により可視化・定量化することで，スタイル表現の内部メカニズムを明らかにする．

3.2　方言辞書の作成

今回の辞書作成において関西弁コーパス [6]，東北地方民話コーパス [7]，名大会話コーパス [8]，日本語話題別会話コーパス [9] を用いた．

(1) サブワード頻度辞書の作成　各 LLM に付属するトークナイザを用いて，コーパスごとにサブワード単位で出現頻度を集計した．方言側は関西弁コーパス・東北地方民話コーパス，参照側は名大会話コーパス・日本語話題別会話コーパスを使用し，モデルごとに独立したサブワード頻度辞書を構築した．

(2) 差分抽出　方言側と参照側でサブワード出現頻度を比較し，参照側に出現しないサブワードを方言特有語彙候補として抽出した．各サブワードに対し次式で差分を計算し，正の値をもつサブワードを候補とした：

$$\Delta f(\text{piece}) = f_{\text{dialect}}(\text{piece}) - f_{\text{reference}}(\text{piece}) \quad (1)$$

閾値や正規化処理は行わず，単純差分ベースで参照側に出現しないサブワードを方言特有候補とした．

(3) 機能語フィルタ　Bond らの [10] による日本語計算言語学の整理を踏まえ，助詞・接続語・動詞語尾等の機能語に着目した．MeCab+NAIST-jdic で品詞タグを付与し，該当語を方言辞書として残した．正規化処理（全角/半角統一，記号除去）後，1〜3-gram を構成し，関西・東北方言辞書を得た．

3.3　検出対象テキストとプロンプト設計

各方言コーパスから 1 文単位でサンプリングを行い，方言辞書に一致する文のみを対象とした．年齢層などを表す話者属性はシステムプロンプトに反映し，発話者のスタイルを保持した入力形式とした．

プロンプトは全タスクで共通雛形を用い，方言名・話者属性・文脈情報を明示する形式とした．全文は付録 A に掲載している．

3.4　検出ロジック

方言的特徴の検出は教師強制下で行い，参照トークン位置 t における各層 l の出力を評価した．Logit Lens [3] により，層 l・位置 t の隠れ状態 $h_{l,t}$ を出力空間へ線形写像し，

$$p_{l,t} = \text{softmax}(h_{l,t} W_{\text{out}} + b) \quad (2)$$

として各トークンの確率分布を得た．

得られた分布から top-3 のトークン候補を抽出し，文字列に復元した上で，直前の入力文と連結してサブワード単位の n-gram 文脈を構成した．この文脈を事前に作成した方言辞書と照合し，一致が確認された場合には，層 l・時刻 t・一致 n-gram・確率値・順位（top-k）を記録した．

層構造上での方言特徴の出現傾向を定量化するため，指標を二つ定義する．まず，層 l における辞書一致トークンの出現回数を表す層別検出数 D_l は

$$D_l = \sum_t \mathbf{1}\{\text{match}_{l,t}\} \quad (3)$$

とし，一致トークンの平均確率を示す検出時確率 $\bar{P}_l$ は

$$\bar{P}_l = \frac{1}{D_l} \sum_{t:\text{match}} p_{l,t}(y_t) \quad (4)$$

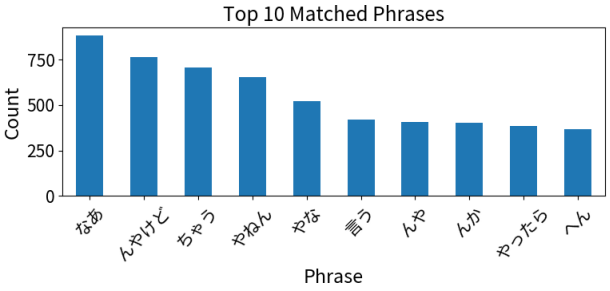

図 1: Sarashina-3B における関西方言の上位検出語彙

で定義する．ここで，l は Transformer の層番号，t は生成系列におけるトークン位置，$p_{l,t}(y_t)$ は層 l の Logit Lens 出力における一致トークン y_t の確率値を表す．これら 2 指標により，各層における方言的特徴の活性度と確信度を定量的に評価した．

3.5 使用モデルとトークナイザ

使用モデルは LLM-jp-3-3.7B-Instruct3, sarashina2.2-3B-instruct-v0.1，Swallow-7B-instruct-v0.1．各モデルに付属するトークナイザを辞書構築・検出の両段階で一貫して利用した．

4 実験結果

4.1 モデル出力中における方言語彙の分布

3 つのモデル（LLM-jp3-3.7B, Sarashina-3B, Swallow-7B）の生成出力に方言辞書照合を行い，上位 10 件の方言語彙を比較した．代表例として Sarashina-3B の関西方言結果を図 1 に示す．他モデルは付録 B に記載している．関西方言では，全モデルで「ねん」「へん」「やろ」「なあ」などの語尾表現が高頻度であった．Sarashina-3B は「んやけど」「ちゃう」など多様な語彙を含む一方，LLM-jp3-3.7B と Swallow-7B では上位語が比較的集中する傾向が見られた．

東北方言では，LLM-jp3-3.7B と Sarashina-3B が「ご」「ど」「べ」「ぐ」などの短音節要素を多く含んだのに対し，Swallow-7B では「ぇ」「ぁ」「ぃ」など音韻断片が上位となり，語彙単位よりも音韻的再現が顕著であった．

4.2 層別検出傾向と確率推移

教師強制下で得た生成列に対し，方言辞書との一致に基づく層別検出数と出力確率（top-1〜top-3）の推移を分析した．代表として Sarashina-3B の結果を図 2 に示す．他のモデルの結果は付録 C に掲載している．

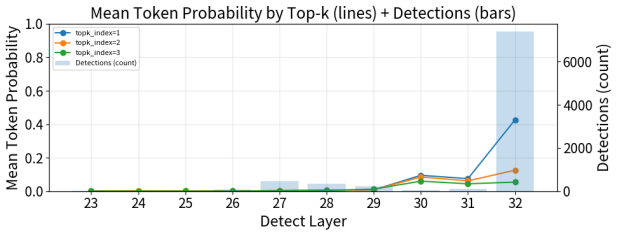

図 2: 関西—Sarashina-3B：層別検出数（棒）と平均トークン確率（折れ線）の推移

方言特徴の初出層　全てのモデルにおいて，方言特徴は例えば 23 層あるいは 19 層といった中間層で初めて検出され，最終層では初出しなかった．これは，方言語彙が出力直前ではなく，中層段階で候補として浮上することを示唆する．

検出数のピーク　全てのモデルにおいて，中間層付近で検出数の局所的なピークが観測されたが，その後に最終層へ向かってどのように検出数が推移するかにはモデル間で違いが見られた．具体的には，中間層で一度活性化した後，最終層に向けて再び検出数が強まる二段階的な推移を示すモデルと，中間層での活性化の後，中後層で一度調整が行われ，最終層で再び増幅される三段階的な推移を示すモデルとに大別された．

平均トークン確率の推移　top-1 確率は概ね層の進行に伴い上昇し，top-2/3 は大きく変動しなかった．これは，中層で複数候補が活性化し，後段で単一語へ収束する過程を示す．

考察：段階処理仮説と層機能　検出数と確率の分布が異なる推移を示したことから，モデル内部における方言語彙処理は

(a) 候補活性化 → (b) 候補間競合 → (c) 確信度形成

という段階的構造に基づいて進行していると考えられる．中層では「話法・語尾などスタイル候補の活性化」，後層では「候補の絞り込みと再配分」，最終層では「確信度の調整と出力確定」が主に行われている可能性が高い．このような層的役割分担は，意味内容の合成過程とスタイル選択過程が内部的に部分分離していることを示唆しており，LLM におけるスタイル制御の基盤構造を説明するものである．

4.3 標準語コーパスにおける層別挙動の比較

方言タスクで観測された層別挙動が方言固有の現象であるか，あるいはより一般的な文体処理機構に由来するものであるかを明確化するため，標準語コーパスに対しても同一手続きによる層別分析を

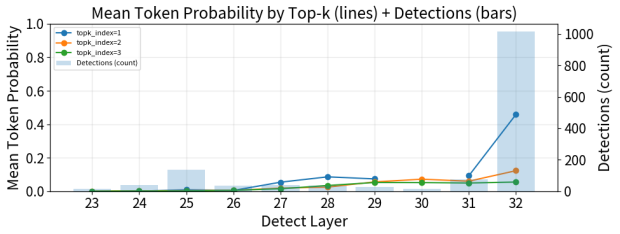

図 3: 標準語コーパスに対する Sarashina-3B の層別検出数と平均トークン確率

行った．具体的には，日本語話題別会話コーパスの文を対象とし，教師強制下で生成したトークン列に対して，方言タスクと同一の特徴語辞書および検出ロジックを適用した．

代表例として Sarashina-3B の結果を図 3 に示す．他モデルの結果は付録 D に掲載している．いずれのモデルにおいても，(i) 中間層付近で特徴語検出数が局所的に増加すること，(ii) トークン確率が層の進行に伴って単調に上昇し，最終層で最大化することが確認された．

このことから，文体候補の中間層活性化と上位層での収束は，特定方言に依存しない一般的構造であると考えられる．

4.4 Perplexity による方言コーパスの難易度分析

文体差とモデル予測難易度の関係を評価するため，LLM-jp3-3.7B-Instruct3 を用いて標準語・関西方言・東北方言コーパスの perplexity を算出した．LLM-jp3 は，公開日本語モデルの中でも学習データと事前学習設定が比較的明確であり，文体差に基づく perplexity 比較の基準として適しているため，本分析では同モデルを採用した．各コーパスに対する基本統計量の詳細は付録 E に示す．

平均値に基づく比較では，関西方言が最も低く，標準語，東北方言の順に perplexity が高くなる傾向が確認された．そこで，最も値が大きかった東北方言について，GPT-5.1 を用いて文全体を標準語へ書き換え，同様の計測を行った．その結果，平均値はむしろ悪化したものの，四分位統計に着目すると大部分の文で perplexity が改善しており，特に中央値および第 1・第 3 四分位では標準語コーパスよりも低い値を示した．これは，一部の文が極端に高い perplexity を示し，平均値を大きく押し上げているためであると考えられる．

以上の結果から，東北方言コーパスの高 perplexity は文体全体に起因するものではなく，「低頻度で方言性の強い語彙・表現」が局所的に難易度を高めている可能性が高い．この点は，関西方言の perplexity が比較的低いという結果とも整合的であり，学習データ中で比較的高頻度に現れる関西系表現がモデルにとって予測しやすい一方，東北方言のごく一部の低頻度表現が予測難度を著しく上昇させている構造を示唆する．

5 おわりに

本研究では，大規模言語モデル（LLM）が同一言語内の文体差，とりわけ方言にどのように対応しているかを明らかにするため，関西方言・東北方言を対象に内部機序分析を行った方言サブワード辞書と Logit Lens を用いた層別分析の結果，方言的特徴は最終層ではなく中間層で初出し，後層で再調整されるという段階的構造を示すことが分かった．さらに，標準語コーパスに同様の分析を適用したところ，中間層での文体候補の活性化と後層での確信度形成という過程が共通して観測され，文体選択機構が特定の方言に依存しない一般的な内部構造である可能性が示唆された．また，perplexity に基づく難易度分析から，低頻度で方言性の強い語彙・表現が局所的に予測難度を高めていることが明らかとなった．この結果は，学習データ中における方言表現の頻度分布がモデルの予測容易性に影響することを示している．

一方，本研究は二方言に限定された分析に基づいており，層別検出数や確率推移といった限られた指標のみを用いている点に加え，学習データ中の方言頻度を直接評価できていないという制約がある．特に，Web crawl データ等を用いた大規模コーパスにおける方言頻度の推定は，方言の予測容易性と内部表現との関係をより厳密に検証するうえで重要な課題である．今後は，こうした方言頻度の外部分析と本研究で示した層別挙動を組み合わせることで，方言的特徴を担うニューロン群の同定や文化的知識に対応した内部構造の分析へと拡張し，文体処理の理解につながる枠組みの構築を目指す．

参考文献

[1]Chris Wendler, Veniamin Veselovsky, Giovanni Monea, and Robert West. Do llamas work in English? on the latent language of multilingual transformers. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 15366–15394, Bangkok, Thailand, August 2024. Association for Computational Linguistics.

[2]Chengzhi Zhong, Fei Cheng, Qianying Liu, Junfeng Jiang, Zhen Wan, Chenhui Chu, Yugo Murawaki, and Sadao Kurohashi. What language do japanese-specialized large language models think in? 言語処理学会第 31 回年次大会発表論文集, pp. 2618–2623, March 2025.

[3]nostalgebraist. interpreting GPT: the logit lens. LessWrong, August 2020.

[4]尾崎慎太郎, 平岡達也, 大竹啓永, 大内啓樹, 渡辺太郎, 宮尾祐介, 大関洋平, 高木優. 大規模言語モデルにおけるペルソナの役割と内部動作の理解. 言語処理学会 第 31 回年次大会 発表論文集, pp. 2648–2653, March 2025.

[5]Haoran Huan, Mihir Prabhudesai, Mengning Wu, Shantanu Jaiswal, and Deepak Pathak. Can LLMs lie? investigation beyond hallucination, 2025. arXiv:2509.03518.

[6]ヘファナンケビン. 関西弁コーパスの紹介. 総合政策研究, Vol. 41, pp. 157–164, 2012.

[7]高道慎之介, 丹治尚子, 佐伯高明, 森松亜依, 庄司潤子, 佐藤照一, 猿渡洋. 東北方言昔話に関する歴史的音声コーパスと機械学習ベース自動音声復元の試み. じんもんこん 2022, December 2022.

[8]藤村逸子, 大曽美恵子, 大島ディヴィッド義和. 会話コーパスの構築によるコミュニケーション研究. 藤村逸子, 滝沢直宏（編）, 言語研究の技法：データの収集と分析, pp. 43–72. ひつじ書房, 2011.

[9]中俣尚己. 日本語話題別会話コーパス：j-tocc 解説資料, 2021.

[10]Francis Bond and Timothy Baldwin. Introduction to japanese computational linguistics. In Francis Bond, Timothy Baldwin, Kentaro Inui, Shun Ishizaki, Hiroshi Nakagawa, and Akira Shimazu, editors, *Readings in Japanese Natural Language Processing*, pp. 1–28. CSLI Publications, 2016.

A　使用プロンプト例

本研究で使用したプロンプト設定例を示す．各プロンプトは，方言設定を含むシステムプロンプト（system），入力文（input），および生成出力（output）から構成される．

表 1: プロンプト例（ID: 3）

System: あなたは大阪や神戸の都市圏に住んでいる 70 歳〜79 歳の女性であり、常に自然な関西弁で発話します。
Input: ああ、ほぼじゃあ体弱かったんか。
Output: だからまあ、お父ちゃん子やったと思うわおばあちゃんはな。

B　モデル出力中の上位検出語彙（補足）

本文では Sarashina-3B の結果を代表例として示した．本節では，他モデルにおける関西方言の上位検出語彙を補足として示す．

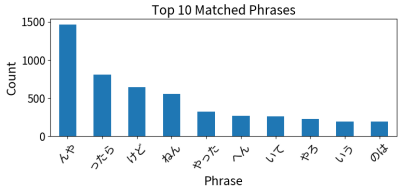

図 4: 関西—LLM-jp3：上位検出語彙

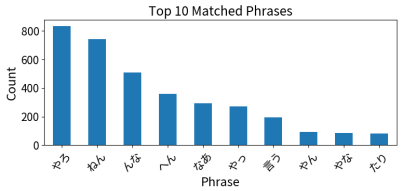

図 5: 関西—Swallow：上位検出語彙

C　層別検出数および確率推移（補足）

本文では Sarashina-3B を代表例として示した．本節では，他モデルにおける関西方言の層別検出挙動を補足として示す．

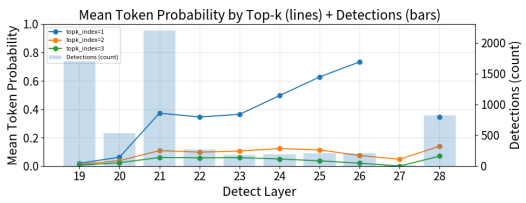

図 6: 関西—LLM-jp3：層別検出数と平均トークン確率

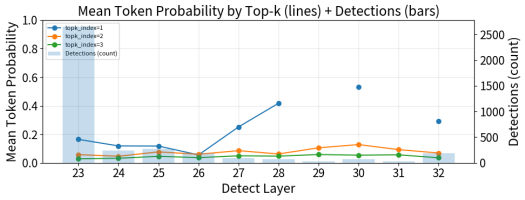

図 7: 関西—Swallow：層別検出数と平均トークン確率

D　標準語コーパスに対する層別挙動（補足）

本文では Sarashina-3B の結果を代表例として示した．本節では，標準語コーパスに対する LLM-jp3 の層別挙動を補足として示す．

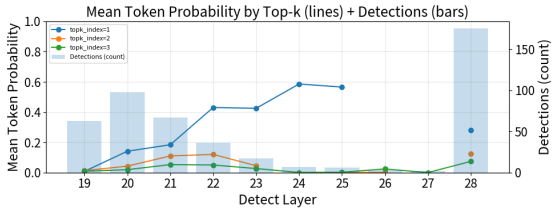

図 8: 標準語—LLM-jp3：層別検出数と平均トークン確率

E　Perplexity 統計量

本研究で用いた標準語・関西方言・東北方言コーパス，および東北方言コーパスを標準語へ書き換えたコーパスに対する LLM-jp3-3.7B の perplexity の基本統計量を表 2 に示す．

表 2: 各コーパスに対する LLM-jp3-3.7B の perplexity 統計量

指標	標準語	関西方言	東北方言	東北方言 → 標準語
件数（count）	1500	3545	649	649
平均（mean）	478.9	300.4	641.1	914.4
中央値（median）	204.8	124.9	293.2	133.9
第 1 四分位（Q1）	92.9	59.2	152.2	62.4
第 3 四分位（Q3）	464.3	272.0	625.5	344.1
最小値（min）	2.19	3.95	5.96	7.75
最大値（max）	21682.1	17715.1	16670.0	180887.2