短歌を読む際の情動に関する脳活動の解析

佐藤 杏奈 [1]　近添淳一 [2]　船井正太郎 [2]
持橋大地 [3]　鹿野 豊 [4]　浅原正幸 [5]　磯 暁 [6]　小林 一郎 [1]
[1] お茶の水女子大学 理学部 情報科学科　[2] 株式会社アラヤ
[3] 統計数理研究所　[4] 群馬大学　[5] 国立国語研究所　[6] 高エネルギー加速器研究機構
{g1920519, koba}@is.ocha.ac.jp　{chikazoe_junichi, funai_shotaro}@araya.org
daichi@ism.ac.jp　yshikano@gunma-u.ac.jp　masayu-a@ninjal.ac.jp　iso@post.kek.jp

概要

本研究では，詩や短歌といった言語芸術が誘起する情動がヒト脳内においてどのように表現されるのかを調査する．とくに本研究では短歌を取り上げ，fMRI 内で提示した短歌を読んだ際の被験者の脳活動を観測し，BERT や GPT などの汎用言語モデルを通じて抽出された言語特徴量を入力として脳内状態を推定する符号化モデルを構築し脳内状態を推定した．被験者が「詩的である」,「詩的でない」と感じるのはヒト脳内でどのような処理が行われていることに起因しているのかについて調査を行った結果，GPT を用いた際に, 情動と深く関係しているとされる大脳辺縁系の一部である島皮質，中層では帯状回，高層では眼窩前頭皮質のハブ性が詩的と感じているとき強くなることが確認できた．

1　はじめに

近年では，磁気共鳴装置 (fMRI) や脳磁図 (MEG) などの非侵襲的な脳機能計測技術の発展と，深層学習に代表される機械学習技術の高度化により，ヒト脳内の情報処理プロセスの解明や定量的理解を行う研究が盛んになっている [1]．論理的な情報処理能力と共に，情動に関する情報処理能力はヒトの知能を解明する上で重要であり，本研究ではヒト脳内の情動に関する処理プロセスの解明を目的とし，統計モデルを通じた調査を行う．

2　関連研究

近年，脳神経科学において脳内に生起する意味表象が知覚カテゴリとしてどのように大脳皮質上に構成されるかについても明らかにされてきている [2, 3]．とくに自然言語の単語の意味をベクトルとして表現する word2vec [4] が出現して以来，ヒト脳に与えられる言語刺激の特徴量が自然言語処理技術の汎用言語モデルを用いて表現され，脳内状態の推定に利用されるようになっている [5, 6, 7]．Schrimpf ら [5, 6] は，43 個の様々な汎用言語モデルを用いて言語刺激を特徴量として表現し，ヒト脳活動との対応関係を調査した．Caucheteux ら [7] は脳内状態を表現する特徴量として性能が良いとされる GPT-2 [8] について詳細な調査をしている．

ヒト脳活動における情動を対象に解析したものとして，Kim ら [9] は音楽刺激下の被験者の脳活動を fMRI で観測したデータから感情反応を推定し，人間の感情処理に関連する特定部位の反応が優位になったことを確認している．Koide-Majima ら [10] は，動画視聴時の脳活動を fMRI を用いて計測し，その刺激となった動画クリップに対して注釈づけされた 80 個の感情表現との相関をとり，脳内における感情の分布を調べている．また，Satpute ら [11] は，言語が情動にどのような役割を担っているかについて神経科学の文献をレビューすることによって調査をしており，言葉の意味処理に関わる脳領域は，とくに情動に関する表現内容を保持するという考えを示唆する証拠が得られたことを報告している．

このようなアプローチに対して，本研究では，自然言語を処理する深層学習モデルを作業モデルとして利用し，様々な刺激下のヒト脳内に直接の反応として表出される感情を扱うのではなく，言語芸術である短歌によって誘起される感性や感覚に近いヒト脳内情動活動を対象に調査を行う．

3　言語刺激下の脳活動解析

3.1　実験概要

本研究の概要を図 1 に示す．言語モデルの中間層から抽出した表現ベクトルを特徴量として使用

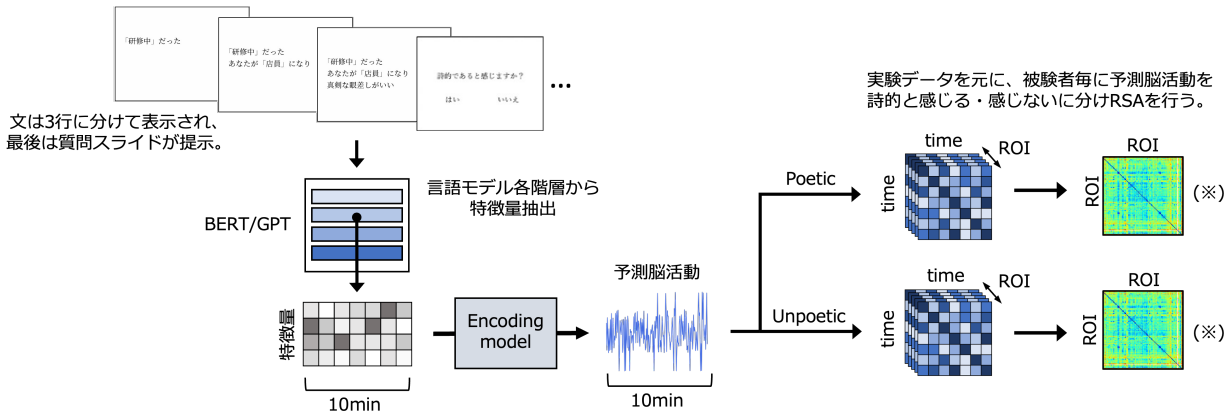

図 1: 詩的と感じるまたは感じないときの脳の振る舞いの調査の流れ.

し，符号化モデル (3.3 節参照) を用いて脳活動を推定する．深層学習モデルの各階層が表現できる情報がどのように遷移するかを調べるため，各階層から符号化モデルを構築し．予測された脳活動を用いて RSA(3.4 節参照) を行い，関心領域 (ROI) の振る舞いを表す，ROI × ROI の RDM(図 1(※)，3.4 節にて後述) を作成する．

3.2 被験者実験

本研究の被検者は 18 歳から 34 歳までの日本人 32 名で，脳画像データは fMRI により取得した．実験では，『現代日本語書き言葉均衡コーパス』(BCCWJ) に含まれる短歌と, 同じ 31 文字の平文各 150 文を fMRI 内で提示し，それが詩的と感じるかどうかを手元のボタンで解答させた．短歌または平文は 3 行に分けて，それぞれ 3 秒間提示された後，「詩的であると感じますか？」と記載された質問スライドが 3 秒間提示される (図 1 参照)．1 試行の所要時間は 12 秒であり，50 試行を 1 セッションとし，6 セッションを行った．文は各セッション内で参加者ごとにランダムな順に提示され，各セッションは短歌 25 首，平文 25 文で構成される．脳画像データは，FreeSurfer[1)] を用いて解析を行い，それにより抽出された大脳皮質のボクセルのみを分析に用いた．また，脳領域区分アトラスには FreeSurfer で提供される Destrieux Atlas [12] を用いた．

3.3 符号化モデル

本研究における符号化モデル (Encoding model) の構築手法は，Naselaris ら [13] によるものを採用した．符号化モデルの構築方法として，ヒト脳への刺激となるデータから抽出した特徴量と刺激下の脳活動状態を線形回帰し，計測脳活動パターンと予測脳活動パターンが近づくように重みを学習する．一般的に線形回帰にはリッジ回帰が適用され，回帰係数を観察することでボクセルに対する振る舞いを観察することなどが可能となる．

3.4 表象類似解析 (RSA)

表象類似解析 (Representational Similarity Analysis, RSA) は，Kriegeskorte ら [14, 15] によって提案された手法である．RSA を用いることにより，比較対象となる二つのパターンの非類似性は，それらの距離に対応するものとして表現される．全てのペアの表現距離 (または非類似度) を測定することで，Representational Dissimilarity Matrix(RDM) を作成でき，また，RDM 同士の相関を計算し様々な表現を比較することが可能になる．本研究では非類似度として相関距離 (1 – ピアソン相関係数) の値を使用する．

4 実験

4.1 符号化モデル作成

ヒト脳に与えられた刺激の特徴量から脳内状態を推定するために，符号化モデルを構築する．1 枚のスライドが提示されている間，そこに表示されている文で埋めて言語データを構築し，脳データと対応づけた．質問スライドが表示されている間は，直前に表示されていた短歌もしくは平文に，「詩的であると感じますか？」という文を加え，言語データとした．

言語モデルの中間層から抽出した言語特徴量を説

1） https://surfer.nmr.mgh.harvard.edu/

明変数として脳活動データを予測する符号化モデルには，線形リッジ回帰を適用した．その際，神経活動に伴う血流の増加の反応時間を考慮し，fMRI で観測された脳活動データをその 3〜9TR[2]前の該当区間全ての言語特徴量から予測する形で回帰を行う．全 6 セッションのうち，1 セッションをテストデータ，残り 5 セッションを訓練データとし，各セッションの脳活動データを予測した．また，訓練データについて 5 分割交差検証を行い，平均の相関係数が最も良くなる正則化項 α を $[10^1, 10^6]$ における 15 個の値の中から採用した．

4.2 実験設定

本研究では，言語特徴量を表現する言語モデルとして BERT [16] と GPT [17] を使用する．事前学習モデルとして，BERT_{large}[3]と GPT_{1b}[4](共に 24 層) を採用し，以下の 2 ステップで Multi-step fine-tuning を行った．

短歌の学習　それぞれ，BCCWJ に含まれる実験で使用していない短歌 3571 首 (訓練データ 90%) で fine-tuning を行った．学習タスクはそれぞれ事前学習として採用されている，BERT は Masked Language Modeling(15%)，GPT は Causal Language Modeling を使用した．

詩的感覚の有無の分類　被験者毎に，実験に使用した短歌平文 300 文を，被験者が fMRI 実験で詩的と感じる・感じないと回答した結果に合わせてラベル付けをし，huggingface の transformers[5]で提供されている，`BertForSequenceClassification` と `GPT2ForSequenceClassification` を使用し fine-tuning を行った．この時，符号化モデル構築時に訓練データとなる 5 セッションで提示された 250 文を訓練データとした．

また，解析では被験者間で平均をとるため，符号化モデルの精度が低い被験者のデータがノイズになってしまうことを鑑み，(1) 短歌もしくは平文，詩的と感じる・感じないでラベル付けをしたとき，マクロ F_1 スコア >0.85 であり，(2) 事前実験 (fine-tuning をしていない事前学習モデルから抽出した特徴量を使用) での予測脳活動が，実験で計測された脳活動に有意な正の相関がある ($p < 0.01$) 被験者 7 名のみを解析の対象とした．

2) 1TR=0.75 秒
3) https://huggingface.co/cl-tohoku/bert-large-japanese
4) https://huggingface.co/rinna/japanese-gpt-1b
5) https://github.com/huggingface/transformers

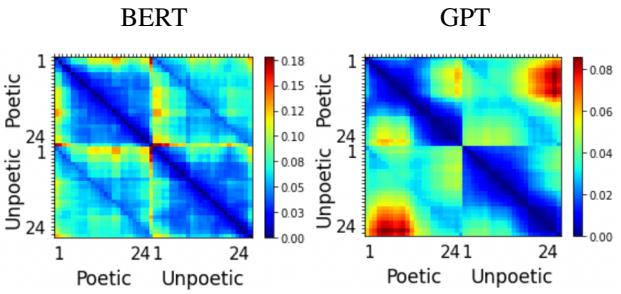

図 2: 中間層×中間層の RDM.

4.3 実験結果

言語モデル各階層の予測脳活動を詩的と感じる (Poetic) もしくは感じない (Unpoetic) に分け，それぞれで，ROI 毎に時間×時間の RDM を作成する．本研究では．実測脳活動と予測脳活動に相関がないという帰無仮説のもと．False discovery rate(q<0.05) 補正を適用する．帰無仮説が棄却され．かつ．実測脳活動との相関が正となったボクセルのみを使用して RSA を行う．

ROI 毎に作成された RDM を比較することで ROI × ROI の RDM(図 1(※)) を作成する．その RDM を被験者間で平均をとったのち，各階層で詩的と感じるもしくは感じないで作成された (階層数× 2) 個の RDM を比較することで，層×層の RDM を作成した (図 2).

さらに上記 2 つの RDM を 1 つにした RDM を UMAP [18] を用いて 3 次元に次元圧縮をし，可視化する (図 3).

次に，他の様々な ROI と類似した情報表現を持つ，脳内情報のハブとなっている ROI を見つけるため，PageRank アルゴリズム [19] を用いて解析を行う．ROI 間の非類似度を表している RDM(図 1(※)) を，(1 – 非類似度) とすることで相関行列を作成し，この相関行列を遷移確率行列として PageRank アル

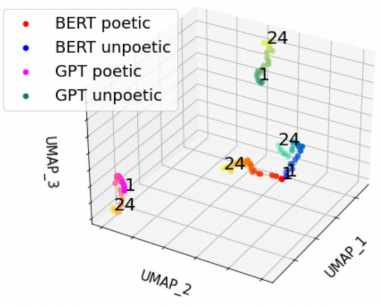

図 3: RDM の次元圧縮結果.

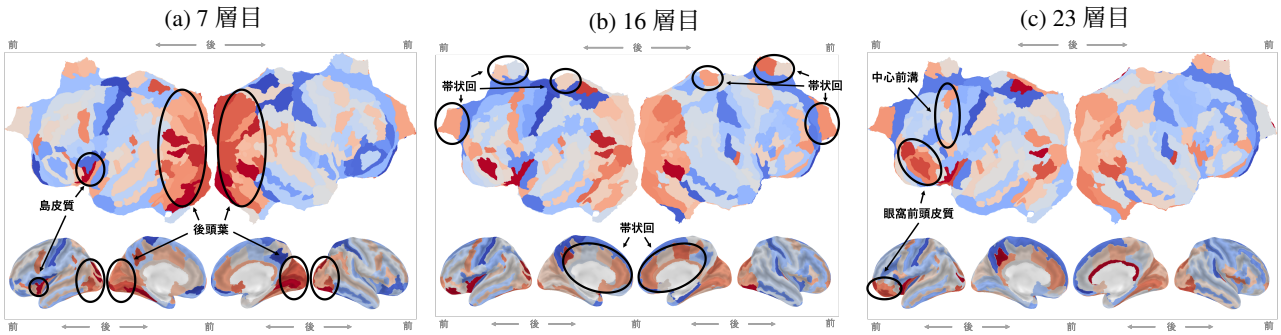

図 4: GPT 各階層における PageRank 値の差分 (詩的と感じる (赤) vs. 詩的と感じない (青)).

ゴリズムを適用する．具体的に，B_{P_i} は相関行列 P_j から遷移する可能性のある相関行列の集合とし，相関行列 P_i における PageRank である $r(P_i)$ をべき乗法で以下のように求める．

$$r_{k+1}(P_i) = \sum_{P_j \in B_{P_i}} \frac{r_k(P_j)}{|P_j|}$$

各階層，詩的と感じる・感じないの各 RDM に対して上記 PageRank アルゴリズムを適用し，$(r(P_{poetic_i}) - r(P_{unpoetic_i}))$ として差分を取ったものを図 4 に示す．ここでは，詩的と感じる・感じないの非類似性が高かった GPT の結果のみを示す．赤い部分が，減算後値が大きくなった, つまり詩的と感じる時により PageRank 値が高かった脳領域であり．青い部分が詩的と感じない時により PageRank 値が高かった脳領域である．

4.4 考察

図 2 において，BERT では各階層同士で Poetic に対する Unpoetic の非類似度が他層に比べ低くなっている一方で，GPT ではその値の低下は小さい．よって，GPT は詩的と感じているときと感じていないときでは．各階層でより異なる情報表現を持っていると言える．

図 3 において，BERT と GPT どちらも徐々に情報が遷移していることから，深層学習モデルの各階層が説明できる脳活動は階層が進むに従って変化していることがわかる．

図 4 では，低層，中層，高層として GPT の 7，16，23 層目での PageRank 値の差分を可視化している．他の階層や BERT での結果については一部付録 6.3 節で示している．ほとんどの層において，後頭葉 (図 4a 参照) で差分後の値が正，つまり，詩的と感じているときの方がより，後頭葉周辺のハブ性が強くなっていることがわかった．この値は低層の方が大きく，10 層目あたりをピークに層が進むにつれて小さくなっていった．また，殆どの層において左島皮質，左中心前溝 (図 4a, 4c 参照) 周辺の値が大きくなった．島皮質は情動と密接な関係があるとされる大脳辺縁系の一部であり，BERT の高層でも同様にハブ性が強くなることが確認できた (付録 6.3 節参照).

中層では，帯状回周辺で値が大きくなる様子が見られた．帯状回は，大脳辺縁系の各部位を結びつける役割があり，感情の形成と処理に関わりを持つ領域と言われている．詩と散文を読んだときの脳活動を調査した先行研究 [20] では，文の情動性が上がるにつれて活性化した脳領域として，帯状回が挙げられている．また，高層では，情動の処理に重要な役割を果たしているとされている [21]，左眼窩前頭皮質 (OFC) が赤くなる様子が見られた．

5 おわりに

本研究では，詩的であるという情動がヒト脳内においてどのように表現されるのかを言語モデルの中間層から予測した脳活動に対して，RSA を用いて調査を行った．PageRank を用いて ROI のハブ性を調べた結果，GPT において情動と深く関係しているとされる大脳辺縁系の一部である島皮質，中層では帯状回，高層では眼窩前頭皮質のハブ性が詩的と感じているとき強くなることが確認できた．一方で，後頭葉や中心前溝など，情動と関係しているとされていない領域でもハブ性が見られた．

今後は，解析対象の被験者を増やした時の結果の考察を行うつもりである．また，短歌を読んだときのヒト脳内と深層学習モデルでの階層処理についての調査，考察も進めていきたい．

参考文献

[1] Daniel L. K. Yamins, Ha Hong, Charles F. Cadieu, Ethan A. Solomon, and James J. Seibert, Darren an d DiCarlo. Performance-optimized hierarchical models predict neural responses in higher visual cortex. *Proceedings of the National Academy of Sciences*, Vol. 111, No. 23, pp. 8619–8624, 2014.

[2] Alexander G Huth, Shinji Nishimoto, An T Vu, and Jack L Gallant. A continuous semantic space describes the representation of thousands of object and action categories across the human brain. *Neuron*, Vol. 76, No. 6, pp. 1210–1224, Dec 2012.

[3] Alexander G Huth, Wendy A de Heer, Thomas L Griffiths, and Jack L Theunissen, Frédéric E andG̃allant. Natural speech reveals the semantic maps that tile human cerebral cortex. *Nature*, Vol. 532, No. 7600, pp. 453–458, Apr 2016.

[4] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. Vol. 26, , 2013.

[5] Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. Artificial neural networks accurately predict language processing in the brain. *bioRxiv*, 2020.

[6] Martin Schrimpf, Idan A Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy G Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative modeling converges on predictive processing. *Proceedings of the National Academy of Sciences (PNAS)*, 2021.

[7] Charlotte Caucheteux, Alexandre Gramfort, and Jean-Rémi King. Gpt-2's activations predict the degree of semantic comprehension in the human brain. *bioRxiv*, 2021.

[8] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. 2018.

[9] Hyun-Chul Kim, Peter A. Bandettini, and Jong-Hwan Lee. Deep neural network predicts emotional responses of the human brain from functional magnetic resonance imaging. *NeuroImage*, Vol. 186, pp. 607–627, 2019.

[10] Naoko Koide-Majima, Tomoya Nakai, and Shinji Nishimoto. Distinct dimensions of emotion in the human brain and their representation on the cortical surface. *NeuroImage*, Vol. 222, p. 117258, 2020.

[11] Ajay B. Satpute and Kristen A. Lindquist. At the neural intersection between language and emotion. *Affective Science*, Vol. 2, No. 2, pp. 207–220, 2021.

[12] Christophe Destrieux, Bruce Fischl, Anders M. Dale, and Eric Halgren. Automatic parcellation of human cortical gyri and sulci using standard anatomical nomenclature. *NeuroImage*, Vol. 53, No. 1, pp. 1–15, 2010.

[13] Thomas Naselaris, Kendrick N. Kay, Shinji Nishimoto, and Jack L. Gallant. Encoding and decoding in fmri. *NeuroImage*, Vol. 56, No. 2, pp. 400–410, May 2011.

[14] Nikolaus Kriegeskorte, Marieke Mur, and Peter Bandettini. Representational similarity analysis - connecting the branches of systems neuroscience. *Frontiers in Systems Neuroscience*, Vol. 2, p. 4, 2008.

[15] Nikolaus Kriegeskorte and Rogier A Kievit. Representational geometry: integrating cognition, computation, and the brain. *Trends Cogn Sci*, Vol. 17, No. 8, pp. 401–412, Aug 2013.

[16] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. 2018. cite arxiv:1810.04805Comment: 13 pages.

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

[18] Leland McInnes, John Healy, and James Melville. Umap: Uniform manifold approximation and projection for dimension reduction, 2018.

[19] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank Citation Ranking: Bringing Order to the Web. Technical report, Stanford Digital Library Technologies Project, 1998.

[20] Adam Zeman, Fraser Milton, Alicia Smith, and Rick Rylance. By heart an fmri study of brain activation by poetry and prose. *Journal of Consciousness Studies*, Vol. 20, No. 9-10, pp. 9–10, 2013.

[21] Junichi Chikazoe, Daniel H Lee, Nikolaus Kriegeskorte, and Adam K Anderson. Population coding of affect across stimuli, modalities and individuals. *Nature Neuroscience*, Vol. 17, No. 8, pp. 1114–1122, 2014.

6　付録 (Appendix)

6.1　実験データ

3.2 節にて述べた対象被験者は，男性 15 名，女性 17 名の合計 32 名である．脳画像データの撮像は，生理学研究所に設置された 3.0 テスラの MRI(シーメンス社製) を用いて取得された．解剖学的スキャンの撮像条件は，TR0.75 秒，ボクセルサイズ 2.0mm × 2.0mm × 2.0mm である．

6.2　非類似度の可視化

図 2 では，BERT と GPT において，詩的と感じる・感じないときの各階層の非類似性を RDM で示した．このときの Poetic に対する Unpoetic の非類似度を示す (図 5).

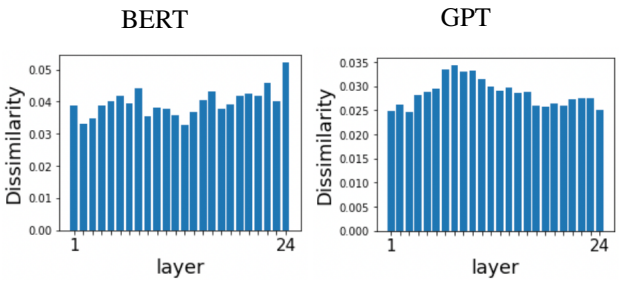

図 5: Poetic-Unpoetic 間の各階層における非類似度.

BERT と GPT2 どちらでも，8 層目あたりで非類似度があがり，その後は BERT は徐々に上がり，GPT では下がる様子が見られた．

図 3 の異なる角度からの可視化結果を示す．

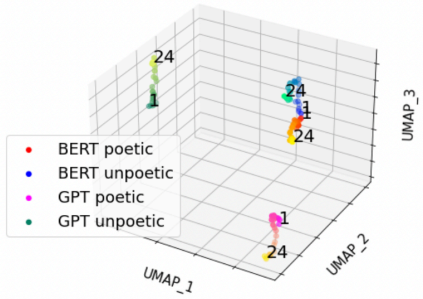

図 6: RDM の次元圧縮結果 (図 3 別角度).

6.3　PageRank の差分結果

4.3 節で述べられているように，PageRank アルゴリズムを用いてハブ性を調査した．本文に掲載できなかった BERT と GPT の一部階層について，(詩的と感じる – 詩的と感じない) の差分結果を示す．

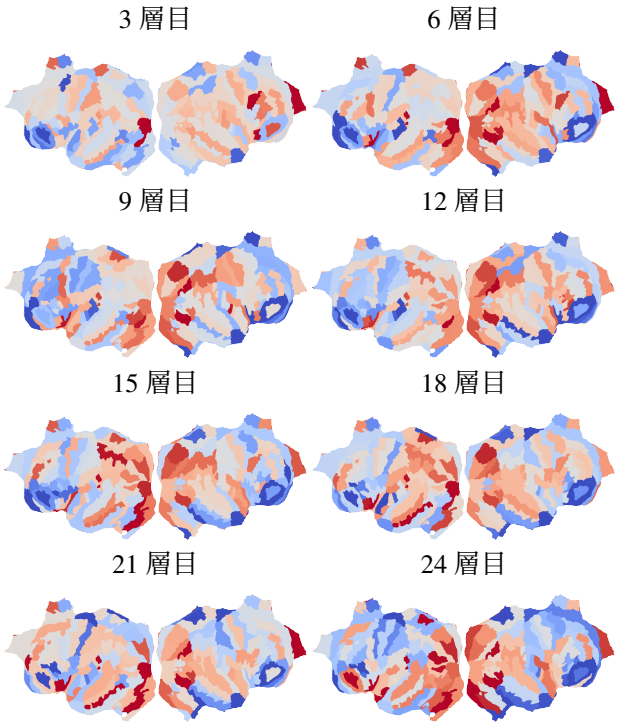

図 7: BERT 各階層における PageRank 値の差分.

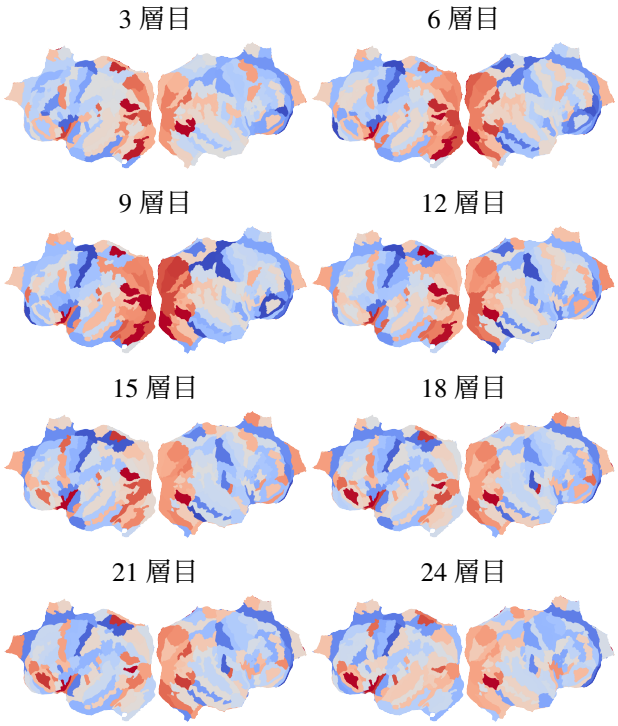

図 8: GPT 各階層における PageRank 値の差分.