BCCWJ-MEG：日本語脳磁図データの構築

杉本 侑嗣 [1] 吉田 遼 [2] 鄭 嫣婷 [3]
菅野 彰剛 [3] 小泉 政利 [3] 大関 洋平 [2]
[1] 大阪大学 [2] 東京大学 [3] 東北大学
sugimoto.yushi.hmt@osaka-u.ac.jp {yoshiryo0617,oseki}@g.ecc.u-tokyo.ac.jp
{jeong,akitake.kanno.c8,koizumi}@tohoku.ac.jp

概要

近年の自然言語処理の成功を受け、神経科学の分野でも自然言語処理との融合的アプローチが急速に発展している。このアプローチにおいては、言語モデルの評価に応用可能な「自然な刺激文に基づく脳活動データセット」が必要不可欠であるが、2025年現在、日本語を対象としたものは存在していない。本研究では、日本語を対象とした新たな脳磁図 (Magnetoencephalography, MEG) データセットである BCCWJ-MEG を構築する。BCCWJ-MEG は、41 名の日本語母語話者が、新聞記事 20 件（229 文・1642 文節）を読んでいる際の MEG データを含む。さらに、我々は、ケーススタディとして、アーキテクチャの異なる複数の言語モデルを BCCWJ-MEG を用いて評価し、(i) 自己注意あり言語モデルが自己注意なし言語モデルよりも脳活動をよりよく説明すること、および (ii) 脳活動をよりよく説明するモデルが工学的性能が高いとは限らないということを確かめた。

1 はじめに

近年の自然言語処理の成功を受け、大規模言語モデルに代表されるようなディープニューラルネットワーク (DNN) に基づく言語モデルモデルがさまざまな分野で応用されている。神経科学の分野でも、自然言語処理との融合的アプローチが発展してきており、例えば、言語モデルの内部表現や算出確率と人間の脳活動を対照することで、その共通部分や差異を洗い出し、それに基づいて人間の文処理メカニズムに対する理解を深めるという研究手法が取られている [1, 2, 3]。

このアプローチにおいては、言語モデルの評価に応用可能な「自然な刺激文に基づく脳活動データセット」が必要不可欠であるが、2025 年現在、OpenNeuro[1)] などで一般に公開されている日本語を対象にした脳活動データセットは存在していない。先行研究は主に英語を対象とした脳活動データにより取り組まれているが [4, 5, 6]、英語など特定の言語を用いた研究によって得られた知見が、他の言語でも同様の結果が確認されるかは自明ではない。言語の普遍性や多様性を理解する上でも、英語と異なる特徴を持つ言語を対象とした脳活動のデータセットの構築は重要である。また、現在公開されている脳データセットは fMRI データが多くを占めているが、fMRI は空間分解度が高いが時間分解度が低いことが知られており、言語の時間的な処理のプロセスを知るうえでは、時間分解度の高い脳データセットが必要である。

そこで本研究では、日本語を対象とした新たな脳磁図 (Magnetoencephalography, MEG) データセットである BCCWJ-MEG を構築する。BCCWJ-MEG は、41 名の日本語母語話者が、新聞記事 20 件（229 文・1642 文節）を読んでいる際の MEG データを含む。日本語は語順などの観点で英語と大きく異なる特徴を持つ言語である上、MEG は空間分解度のみならず時間分解度が高いため、先行研究で主に用いられてきた英語を対象とした fMRI データのみでは得られない知見を提供することが期待できる。さらに、我々は、ケーススタディとして、アーキテクチャの異なる複数の言語モデルを BCCWJ-MEG を用いて評価し、(i) 自己注意あり言語モデルが自己注意なし言語モデルよりも脳活動をよりよく説明すること、および (ii) 脳活動をよりよく説明するモデルが工学的性能が高いとは限らないということを確かめた。

1） https://openneuro.org/

2 実験

2.1 MEG 実験手続き

実験参加者 本研究の MEG 実験には 41 名の日本語母語話者（全員右利きの健常成人、女性 17 名、平均年齢 21.56 歳、SD=2.598) が参加した。参加者は全員正常な視力（矯正を含む）を有していた。

刺激文と課題 本実験で使用した刺激文は、「現代日本語書き言葉均衡コーパス」[7] に収められている新聞記事 20 件（229 文・1642 文節）である。MEG 実験では、BCCWJ-EEG [8] と同様に新聞記事を一件ずつ、文節ごとに提示した。また各記事の提示は PsychoPy [9, 10] を使用し、Rapid Serial Visual Presentation によって、500ms ごとに分節を提示し、各文節提示後は、500ms の間、何も表示されないよう設定した。各記事の終わりには、記事内容の理解度を問うための質問が提示された。本実験では、200 チャンネルの全頭型システム（MEG vision PQA160C-RO; Richo、Tokyo、Japan) を用い、参加者が MEG 装置の中で記事を各文節ごとに読んでいる際の MEG データを収集した。全参加者から収集された MEG データは計 67322 文節に相当する。

前処理 オンラインでは、1000Hz でデータ収集したが、オフライン処理として 200Hz までダウンサンプリング処理を行い、さらに、0.1-40Hz の間でバンドパスフィルタを施した。その後、独立成分分析 (ICA) により、瞬きなど、文処理とは関係のないアーチファクトを除去した。-100ms から 1000ms でエポックス化した後、absolute threshold（2e-12）を用いてエポックごとのノイズ除去を自動で行った。また-100ms から 0ms の間の活動をベースライン補正として使用した。

2.2 MEG データのモデリング

言語モデルの算出確率を、自然な刺激文に基づく脳活動データと対照する実験 [11, 12] のケーススタディを行った。人間のオンライン文処理には予測処理が伴っており、文脈から文節の予測が難しい時には、処理の負担が高くなり、脳活動の負荷が大きくなる一方で、予測がしやすい文節の処理の負担は低くなり、脳活動の負担が軽減されると言われている。先行研究 [13, 14] では、言語モデルの算出確率に基づき導出されるサプライザル（-log p(文節 | 文脈)）を、この予測難易度の定量指標として用い、実際の脳活動と対照することで言語モデルの予測と人間の予測の共通部分や差分が明らかにされてきた。本研究のケーススタディでも、アーキテクチャの異なる複数の言語モデルから導出されるサプライザルと、BCCWJ-MEG の脳活動を比較することで、人間のオンライン文処理の認知モデルとして妥当なアーキテクチャを探求する。

2.3 言語モデル

本研究では、自己注意と統語的構成という 2 つのアーキテクチャの有無に着目して検証を行った。実験に用いた言語モデルを表 1 に示す。

表 1 本研究で評価する言語モデル

	統語的構成なし	統語的構成あり
自己注意なし	LSTM	RNNG
自己注意あり	Transformer	CAG

LSTM RNN に基づく、自己注意なしの、統語的教示のない言語モデル [15]。

Transformer Transformer に基づく自己注意をもつ言語モデル [16]。

RNNG RNN に基づく、自己注意を持たないが、統語的構成を持つ（統語的教示を統合された）言語モデル (Recurrent Neural Network Grammar, RNNG; [17])。文処理に関連する脳波・fMRI データの説明力が高いことが知られている [13, 18]。

CAG Transformer に基づく、自己注意を持ち、かつ統語的構成性も持つ言語モデル (Composition Attention Grammar, CAG; [19])。

2.4 MEG データ解析

MEG 解析は前処理も含め MNE-Python (v1.9.0) [20]、Eelbrain (v0.40.0) [21] と Rstudio (v4.4.1) [22] を使用した。解析にあたって、前処理で問題のあった 6 データを除き、35 人分の前処理済みの MEG データ（計 52349 文節）を統計解析に使用した。

Spaito-temporal clustering tests まず、各言語モデルから算出されたサプライザルが文処理中のどの脳活動領域・時間幅を予測するかを spatio-temporal clustering tests で探索的に検証し、結果を後の評価で機能的関心領域として使用する。ここでの解析に使用する領域として、文処理における fMRI 先行研究 [23] で報告されている関連領域の MNI 座標を使用して、関心領域 (Region of Interests, ROIs) を設定した (図 1 を参照)。この ROIs を使用

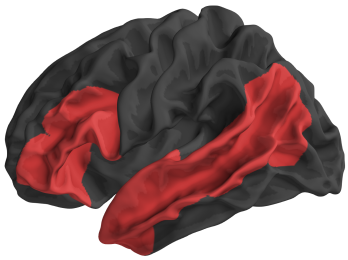

図 1 Spatio-temporal cluster tests で使用した ROIs (赤色で表示した部分)

して、刺激提示後の 300ms-800ms を time window として spatio-temporal clustering tests を行った。spatio-temporal clustering tests では、まずベースラインとなる説明変数を用いて dSPM [24] で信号源推定した MEG 信号推定値を予測する回帰モデルを構築し、次に、本研究で評価する言語モデルに基づく説明変数を一つずつ追加した (詳細は、[25] などを参照)。

評価方法: PPP と PPL の関係 spatio-temporal clustering tests で得られた結果を機能的関心領域 (fROI、図 2 を参照) とし、425-615ms の時間幅での fROI 内の脳活動を平均化した脳活動をモデル化したベースライン回帰モデルを準備する。この回帰モデルに本研究で評価する各言語モデルのサプライザルを説明変数として加えた際の対数尤度の増量分 (Δ LogLik) を評価する。先行研究ではこの増量分を、Psychometric Predictive Power (PPP) と呼び [26]、追加した言語モデルのサプライザルがどの程度 MEG データの説明に寄与しているかを検証する。fROI 内脳活動のベースライン回帰モデルには以下の線形混合モデルを用いる。

$$fROI \sim word\_length + word\_freq + sentid + sentpos + (1|subject\_number) + (1|section\_number)$$

数値型の変数はすべて中心化を行った。また、モデル化した際に 3 標準偏差を超えるデータポイントは除去した。さらに先行研究にならい、各言語モデルから算出される Perplexity (PPL) と PPP との相関を検証する。検証にあたって、無相関検定を行った。同様の解析を fROI とは独立に設定した文処理に関わる関心領域 (ATL、PTL、IFG、付録 B の図 4 を参照) に関しても行った。

3 結果と考察

Spatio-temporal clustering tests Spatio-temporal clustering tests を行い、CAG のサプライザルを用いた結果を図 2 に示す (他のモデルのサプライザルを用いた結果は付録の表 3 を参照)。

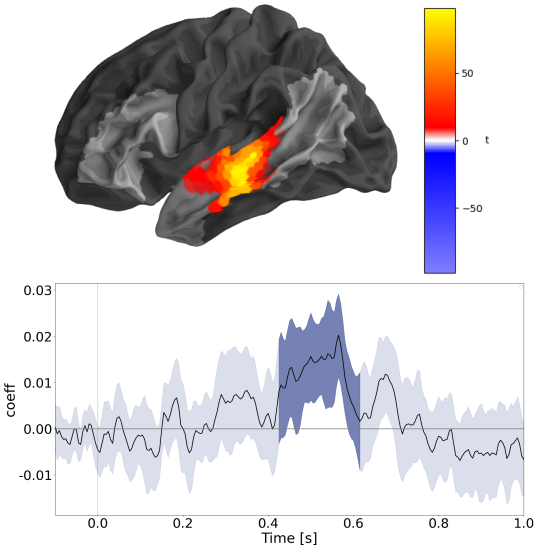

図 2 CAG のサプライザルを用いた spatio-temporal clustering tests の結果。紺色の領域は、95%信用区間を示し、より濃い部分は統計的に有意となった時空間クラスター部分を示す。

結果、統計的に有意な時空間クラスターが 425ms-615ms の時間幅、中側頭回付近で確認された (p=0.049*)。

PPP と PPL それぞれの (f)ROI ごとに算出 PPP と PPL の相関関係をプロットした結果を図 3 に示す。各モデルのサプライザルは、各シードごとに算出された。PPL が良いモデルほど PPP の値が低いという強い逆相関が確認でき、ヒトの読み時間における文処理の先行研究で報告されている結果 [27]、視覚に関する DNN モデル [28] で報告されている結果と概ね一致する結果となった。

また、PPP の有意差を検定するため、(機能的) 関心領域における MEG データを従属変数としてネストしたモデル比較も行い、表 2 に結果を示した。各サプライザルの値は、各シードの平均を使用した。

自己注意あり vs. 自己注意なし モデル比較の結果、fROI では自己注意ありのモデルが自己注意のないモデルよりも優れていることが確かめられた (LSTM < Transformer, RNNG <CAG)。この結果は先行研究 [29] の結果と合致し、先行するシーケンシャルな文字列・あるいは統語構造の情報に対して、選択的な注意を行うことが、自己注意のないモデルよりも脳活動を説明することを示している。fROI とは独立に設定した関心領域 (ATL,PTL,IFG) では、Transformer が全ての領域で LSTM より優れているが、CAG は RNNG に対して統計的に有意には至

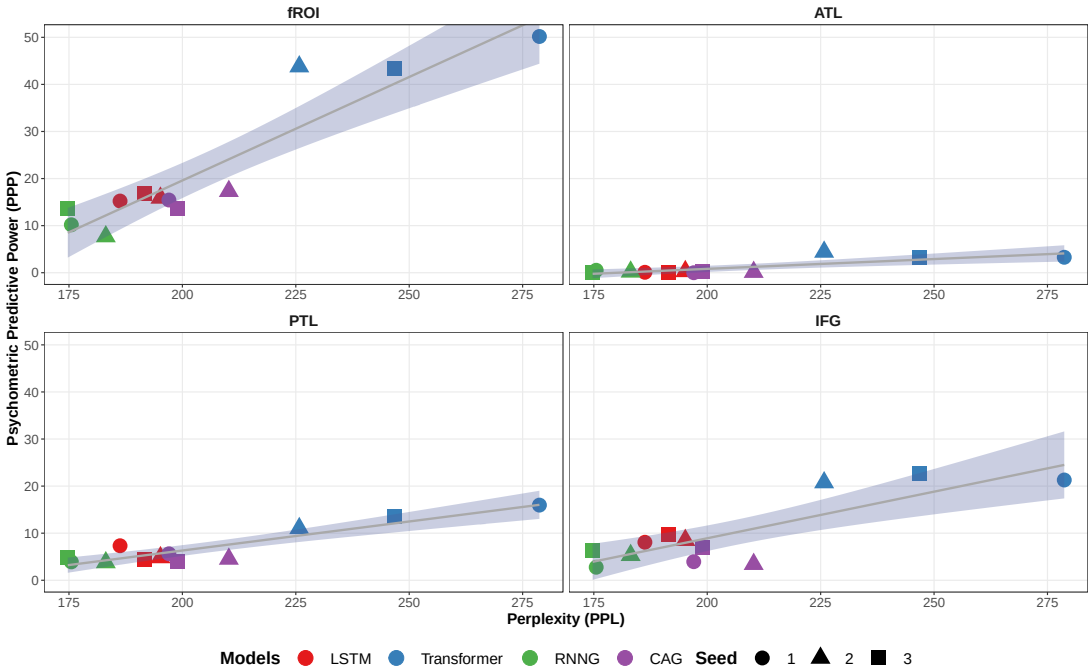

図 3　機能的関心領域 (fROI) と関心領域 (ATL, PTL, IFG) における PPP と PPL の相関関係。相関係数は各領域以下の通り。 fROI: 0.9286655 (p<0.0001), ATL: 0.801 (p=0.00173), PTL: 0.917 (p<0.0001), IFG: 0.841 (p=0.000614)。紺色の領域は、95%信用区間を示す。

表 2　ネストしたモデル比較の結果。Bonferroni 法により有意水準 $\alpha = 0.003125$ で検定を行った。

	fROI			ATL			PTL			IFG		
	χ^2	df	p	χ^2	df	p	χ^2	df	p	χ^2	df	p
自己注意なし < 自己注意あり												
LSTM < Transformer	51.602	1	**<0.00001**	15.9497	1	**<0.00001**	12.5879	1	**<0.00001**	20.3538	1	**<0.00001**
RNNG < CAG	9.4252	1	**0.00214**	0.3965	1	0.5289	0.8362	1	0.36050	0.0795	1	0.77802
統語的構成なし < 統語的構成あり												
LSTM < RNNG	0.3077	1	0.5791	1.0227	1	0.3119	0.0002	1	0.99010	0.7872	1	0.374936
Transformer < CAG	4.9444	1	0.02618	6.4014	1	0.011403	1.3694	1	0.241917	6.0351	1	0.01402

らなかった。

統語的構成性あり vs. 統語的構成性なし　統語的構成のあるモデルとないモデルでの比較の結果、統計的に有意な結果は示されなかった。先行研究では、サプライザル以外の指標を用いた検証もしており、例えば、Brennan et al. 2020 [18] では、RNNG の指標として distance が統語処理に関わる複数の脳部位のデータを説明することが確認されているため、今回扱ったサプライザルでは扱えない統語的な情報がこれらの指標によって評価される可能性がある。

4　おわりに

近年、自然言語処理と神経科学の融合的アプローチにより、脳データセットの需要が高まっているが、日本語を対象とした脳データセットは一般公開の形で利用可能なものが存在していない。そのため、今回新たに日本語を対象にした MEG データセットを構築し、複数の異なるアーキテクチャの言語モデルの評価を行った。結果、自己注意ありモデルは自己注意なしモデルよりも文処理に関わる脳活動をより良く説明することが確認されたが、脳活動をよりよく説明するモデルが工学的性能が高いとは限らないということが確かめられた。今後、BCCWJ-MEG を一般公開し、利用されることを期待したい。

謝辞

MEG データの収集していただいた江村玲氏および東北大学院生の院生の方々、また実験に参加していただいた東北大学の学生の方々に感謝します。本研究は、JSPS 科研費 JP22K18437、JP24H00085、JP24H00087、JST さきがけ JPMJPR21C2 の支援を受けたものです。

参考文献

[1] Charlotte Caucheteux and Jean-Rémi King. Brains and algorithms partially converge in natural language processing. **Communications Biology**, Vol. 5, No. 1, p. 134, 2022.

[2] Martin Schrimpf, Idan Asher Blank, Greta Tuckute, Carina Kauf, Eghbal A. Hosseini, Nancy 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**, Vol. 118, No. 45, p. e2105646118, 2021.

[3] Sreejan Kumar, Theodore R. Sumers, Takateru Yamakoshi, Ariel Goldstein, Uri Hasson, Kenneth A. Norman, Thomas L. Griffiths, Robert D. Hawkins, and Samuel A. Nastase. Shared functional specialization in transformer-based language models and the human brain. **Nature Communications**, Vol. 15, No. 1, p. 5523, 2024.

[4] Shohini Bhattasali, Jonathan Brennan, Wen-Ming Luh, Berta Franzluebbers, and John Hale. The alice datasets: fMRI & EEG observations of natural language comprehension. In **Proceedings of the Twelfth Language Resources and Evaluation Conference**, pp. 120–125, Marseille, France, May 2020. European Language Resources Association.

[5] Jixing Li, Shohini Bhattasali, Shulin Zhang, Berta Franzluebbers, Wen-Ming Luh, R. Nathan Spreng, Jonathan R. Brennan, Yiming Yang, Christophe Pallier, and John T. Hale. Le petit prince multilingual naturalistic fmri corpus. **Scientific Data**, Vol. 9, No. 1, p. 530, 2022.

[6] Amanda LeBel, Lauren Wagner, Shailee Jain, Aneesh Adhikari-Desai, Bhavin Gupta, Allyson Morgenthal, Jerry Tang, Lixiang Xu, and Alexander G. Huth. A natural language fmri dataset for voxelwise encoding models. **Scientific Data**, Vol. 10, No. 1, p. 555, 2023.

[7] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu De. Balanced corpus of contemporary written japanese. **Language Resrouces & Evaluation**, Vol. 48, pp. 345–371, 2014.

[8] Yohei Oseki and Masayuki Asahara. Design of BCCWJ-EEG: Balanced corpus with human electroencephalography. In **Proceedings of the 12th Language Resources and Evaluation Conference**, pp. 189–194, Marseille, France, May 2020. European Language Resources Association.

[9] Jonathan W. Peirce. Psychopy–psychophysics software in python. **Journal of neuroscience methods**, Vol. 162, No. 1-2, pp. 8–13, 2007.

[10] Jonathan W. Peirce. Generating stimuli for neuroscience using psychopy. **Frontiers in Neuroinformatics**, Vol. 2:10, , 2009.

[11] Jonathan Brennan. Naturalistic sentence comprehension in the brain. **Language and Linguistics Compass**, Vol. 10, No. 7, pp. 299–313, 2016.

[12] John T. Hale, Luca Campanelli, Jixing Li, Shohini Bhattasali, Christophe Pallier, and Jonathan R. Brennan. Neurocomputational models of language processing. **Annual Review of Linguistics**, Vol. 8, No. Volume 8, 2022, pp. 427–446, 2022.

[13] John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. Finding syntax in human encephalography with beam search. In **Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 2727–2736, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[14] Cory Shain, Idan Asher Blank, Marten van Schijndel, William Schuler, and Evelina Fedorenko. fmri reveals language-specific predictive coding during naturalistic sentence comprehension. **Neuropsychologia**, Vol. 138, p. 107307, 2020.

[15] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-Term Memory. **Neural Computation**, Vol. 9, No. 8, pp. 1735–1780, 11 1997.

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

[17] Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent neural network grammars. In Kevin Knight, Ani Nenkova, and Owen Rambow, editors, **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 199–209, San Diego, California, June 2016. Association for Computational Linguistics.

[18] Jonathan R. Brennan, Chris Dyer, Adhiguna Kuncoro, and John T. Hale. Localizing syntactic predictions using recurrent neural network grammars. **Neuropsychologia**, Vol. 146, p. 107479, 2020.

[19] Ryo Yoshida and Yohei Oseki. Composition, attention, or both? In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 5822–5834, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[20] Alexandre Gramfort, Martin Luessi, Eric Larson, Denis A. Engemann, Daniel Strohmeier, Christian Brodbeck, Roman Goj, Mainak Jas, Teon Brooks, Lauri Parkkonen, and Matti Hämäläinen. Meg and eeg data analysis with mne-python. **Frontiers in Neuroscience**, Vol. 7, , 2013.

[21] Christian Brodbeck, Proloy Das, Marlies Gillis, Joshua P Kulasingham, Shohini Bhattasali, Phoebe Gaston, Philip Resnik, and Jonathan Z Simon. Eelbrain, a python toolkit for time-continuous analysis with temporal response functions. **eLife**, Vol. 12, p. e85012, nov 2023.

[22] R Core Team. **R: A Language and Environment for Statistical Computing**. R Foundation for Statistical Computing, Vienna, Austria, 2021.

[23] Christophe Pallier, Anne-Dominique Devauchelle, and Stanislas Dehaene. Cortical representation of the constituent structure of sentences. **Proceedings of the National Academy of Sciences**, Vol. 108, No. 6, pp. 2522–2527, 2011.

[24] Anders M. Dale, Arthur K. Liu, Bruce R. Fischl, Randy L. Buckner, John W. Belliveau, Jeffrey D. Lewine, and Eric Halgren. Dynamic statistical parametric mapping: Combining fmri and meg for high-resolution imaging of cortical activity. **Neuron**, Vol. 26, No. 1, pp. 55–67, 2000.

[25] Eric Maris and Robert Oostenveld. Nonparametric statistical testing of eeg- and meg-data. **Journal of Neuroscience Methods**, Vol. 164, No. 1, pp. 177–190, 2007.

[26] Stefan L. Frank and Rens Bod. Insensitivity of the human sentence-processing system to hierarchical structure. **Psychological Science**, Vol. 22, No. 6, pp. 829–834, 2011.

[27] Tatsuki Kuribayashi, Yohei Oseki, and Timothy Baldwin. Psychometric predictive power of large language models. In Kevin Duh, Helena Gomez, and Steven Bethard, editors, **Findings of the Association for Computational Linguistics: NAACL 2024**, pp. 1983–2005, Mexico City, Mexico, June 2024. Association for Computational Linguistics.

[28] Soma Nonaka, Kei Majima, Shuntaro C. Aoki, and Yukiyasu Kamitani. Brain hierarchy score: Which deep neural networks are hierarchically brain-like? **iScience**, Vol. 24, No. 9, p. 103013, 2021.

[29] Danny Merkx and Stefan L. Frank. Human sentence processing: Recurrence or attention? In Emmanuele Chersoni, Nora Hollenstein, Cassandra Jacobs, Yohei Oseki, Laurent Prévot, and Enrico Santus, editors, **Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics**, pp. 12–22, Online, June 2021. Association for Computational Linguistics.

A Spatio-temporal clustering tests の全結果

表 3 各モデルごとの spatio-temporal clustering tests の結果。

Models	Time window	Regions	$p_{cluster}$
LSTM	415 - 610ms	ATL-lh	p=0.0469
	505 - 710ms	IFGtri-lh	p=0.033
Transformer	300 - 780ms	ATL-lh	p=0.0001
	420 - 800ms	IFGorb-lh	p=0.0046
RNNG	Not significant		
CAG	425 - 615ms	ATL-lh	p=0.049

B 関心領域

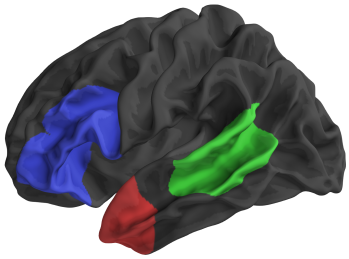

図 4 関心領域解析で使用した ROIs (青色: Inferior Frontal Gyrus (IFG), 赤色: Anterior Temporal Lobe (ATL), 緑色: Posterior Temporal Lobe (PTL))。時間幅は先行研究及び 3 などを考慮して以下の通りに設定した: ATL (300-500ms)、PTL (400-600ms)、IFG (500-700ms)