統語的構成や自己注意を持つ言語モデルは「人間らしい」のか？

吉田遼　　大関洋平
東京大学
{yoshiryo0617, oseki}@g.ecc.u-tokyo.ac.jp

概要

本研究では、統語的構成と自己注意をアーキテクチャに持つ/持たない言語モデルの、人間の眼球運動・脳波のモデリングの精度を評価することで、それぞれの構成要素の「人間らしさ」を統一的に検証する。具体的には、それぞれの構成要素を持つ/持たない 4 つの統語的言語モデルと、自己注意を持つ/持たない 2 つの統語的教示なしベースライン言語モデルを、first pass reading time と P600 振幅で評価する。結果、統語的構成と自己注意は共に眼球運動はよく予測するが、脳波を上手く予測するのは統語的構成のみであった。これは、統語的構成は間接的な認知データの説明力に留まらず人間のオンライン言語処理における高次処理に対応するが、自己注意は高次処理とは乖離している可能性を示唆する。

1　はじめに

近年の自然言語処理で用いられる大規模事前学習可能な言語モデルは、アーキテクチャ自体も「人間らしい」のだろうか。これらの言語モデルは、様々なタスクで人間を超える精度を達成している (e.g., [1]) が、言語獲得のデータ効率の悪さなど、人間との乖離もしばしば指摘されている [2]。これに対比されるように、「人間らしい」アーキテクチャであるとして挙げられる言語モデルの代表が、再帰的ニューラルネットワーク文法 (Recurrent Neural Network Grammar, RNNG; [3]) である。RNNG は、統語的教示が統合された再帰的ニューラルネットワーク (Recurrent Neural Network, RNN; [4]) であり、その確率的予測が、人間のオフライン言語処理を反映する容認性判断 [5, 6, 7] や、オンライン言語処理を反映する眼球運動・脳活動データ [8, 9, 10] と高い一致を示すことが知られている。ここで、RNNG の最大の特徴は、句構造を階層的に一つのベクトルへと構成する**統語的構成**であり、統語的構成なしでは RNNG の性能が損なわれることが知られている [6, 8]。

一方で、近年の自然言語処理では、Transformer [11] が RNN を様々なタスクで上回り [12]、過去の情報に選択的に注意できる**自己注意**の自己回帰に対する優位性が示されている。より近年では、Transformer が、単語毎提示の読み時間や脳波に限れば RNN よりも上手く予測できること [13] から、認知モデリングの文脈においても自己注意の「人間らしさ」が主張され始めている。これらを受け、先行研究 [14] では、RNNG の統語的構成と Transformer の自己注意を統合したモデルである構成注意文法 (Composition Attention Grammar, CAG) が提案され、人間のオフライン容認性判断に対して RNNG や Transformer よりも高い一致率を示すことが明らかにされた。しかし、この統語的構成と自己注意を併せ持つモデルが、人間のオンライン言語処理とも高い一致率を示すかどうかは、未だ明らかにされていない。

さらに、これらの認知モデリングの先行研究では、RNN ベースのモデルにおける統語的構成の効果と、統語的教示を持たないモデルにおける自己注意の効果が、それぞれ独立に検証されるに留まり、統語的構成・自己注意それぞれの要素が、他方の要素の影響を受けずに、普遍的に認知的妥当性が高いかどうか、は統一的に検証されてきていない。

そこで、本研究では、統語的構成と自己注意をアーキテクチャに持つ/持たない言語モデルの、人間の眼球運動・脳波のモデリングの精度を評価することで、それぞれの構成要素の「人間らしさ」を統一的に検証する。具体的には、それぞれの構成要素を持つ/持たない 4 つの統語的言語モデルと、自己注意を持つ/持たない 2 つの統語的教示なしベースライン言語モデルを、first pass reading time と P600 振幅で評価する。

	統語的教示なし（ベースライン）	統語的教示あり	
		統語的構成あり	統語的構成なし
自己注意なし	LSTM	ActionLSTM	RNNG
自己注意あり	Transformer	PLM	CAG

表 1　本研究で評価する、統語的構成と自己注意をアーキテクチャに持つ/持たない 4 つの統語的言語モデルと、自己注意を持つ/持たない 2 つの統語的教示なしベースライン言語モデル。

2　実験

2.1　オンライン文処理のモデリング

人間のオンライン言語処理は予測処理を伴い、各単語が文脈に基づいた予測が難しい時には、処理負荷が向上し、視線停留時間が上昇したり、事象関連電位が励起されたりすると言われる。サプライザル理論 [15, 16] は、この「予測の難しさ」をサプライザル $-\log p($ 単語 $|$ 文脈 $)$ として定式化した。認知モデリングの分野では、この値が言語モデルの予測確率と人間の眼球運動や脳波などとの橋渡し仮説として用いられ、どのような言語モデルが「人間らしい」予測をするのかが、アーキテクチャ・学習データなどの観点から検証されてきた (e.g., [17])。本研究でも、サプライザル理論を橋渡し仮説として用い、統語的構成と自己注意をアーキテクチャに持つ/持たない言語モデルの認知的な妥当性を検証する。

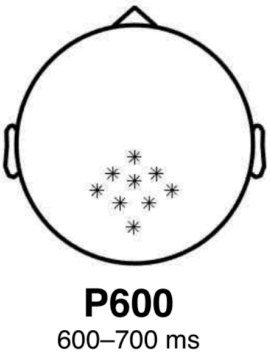

図 1　本研究で検証する P600（後頭部電極の、各単語の最初の視線停留から 600–700ms の間の振幅平均）のトポグラフィ。電極と時間窓の選択は先行研究 [8] を踏襲した。図は同先行研究より抜粋。

2.2　言語モデル

本研究で評価する、統語的構成と自己注意をアーキテクチャに持つ/持たない 4 つの統語的言語モデルと、自己注意を持つ/持たない 2 つの統語的教示なしベースライン言語モデルを、表 1 に示した。全ての言語モデルは、先行研究 [14] により Brown Laboratory for Linguistic Information Processing 1987-89 Corpus Release 1（BLLIP、LG、約 1.8M 文）[18] で学習されたパラメータ数約 16.6M のモデルを用いた。各モデルは 3 つの異なるシードで学習されている。

LSTM　RNN ベースの、自己注意を持たない、純正言語モデル [19]。自己注意なしのモデルの、統語的教示なしベースラインである。

ActionLSTM　RNN ベースの、自己注意を持たない、統語的教示を統合された言語モデル [20]。ただし、統語的構成は行わない。

RNNG　RNN ベースの、自己注意を持たない、統語的教示を統合された言語モデル [3]。統語的構成を行う。

Transformer　Transformer ベースの、自己注意を持つ、純正言語モデル [21]。自己注意ありのモデルの、統語的教示なしベースラインである。

PLM　Transformer ベースの、自己注意を持つ、統語的教示を統合された言語モデル [22]。ただし、統語的構成は行わない。

CAG　Transformer ベースの、自己注意を持つ、統語的教示を統合された言語モデル [14]。統語的構成を行う。

2.3　眼球運動・脳波データ

英語母語話者 12 人分の、眼球運動と脳波の同時計測データである、Zurich Cognitive Language Processing Corpus (ZuCo; [23]) を用いた。先行研究 [24] を踏襲し、自然な読みのデータの 700 文を用いる。

眼球運動データ　先行研究を踏襲し、first pass reading time を用いた。付録 A に示す前処理を施し、160,603 中 93,782 のデータポイント（単語）が統計分析の対象となった。

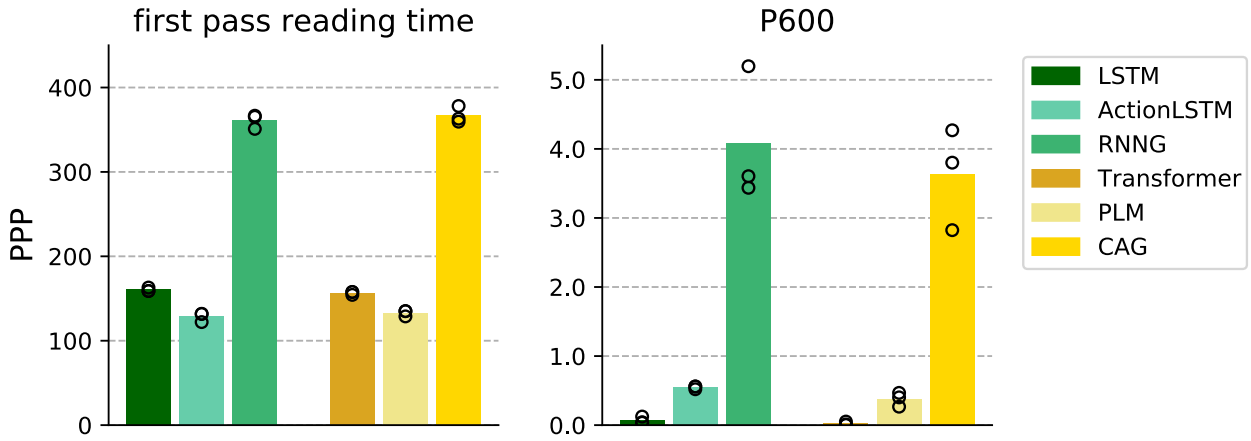

図 2　統語的構成と自己注意をアーキテクチャに持つ/持たない 4 つの統語的言語モデルと、自己注意を持つ/持たない 2 つの統語的教示なしベースライン言語モデルの、心理学的予測精度 (PPP) の結果。縦軸が PPP を、横軸が各言語モデルを表す。棒グラフはシードの異なる 3 つのモデルの PPP の平均値を表し、それぞれの点は各シードの PPP を表す。

脳波データ　統語的処理に関連する [25] とされる、P600 振幅（後頭部電極の、各単語の最初の視線停留から 600–700ms の間の振幅平均）を用いた。電極と時間窓の選択は先行研究 [8] を踏襲した（図 1）。付録 B に示す前処理を施し、160,603 中 86,070 のデータポイント（単語）が統計分析の対象となった。

2.4　評価指標

各言語モデルの予測確率 $p($ 単語 | 文脈 $)$ に基づくサプライザル $-\log p($ 単語 | 文脈 $)$ の、眼球運動・脳波のモデリング精度を評価する。評価指標としては、心理学的予測精度 (Psychometric Predictive Power, PPP)：眼球運動・脳波をモデル化するベースライン回帰モデルに、各言語モデルのサプライザルを加えた際の対数尤度の増加分 (Δ LogLik) を用いる。[1)] 先行研究 [26, 10] を踏襲し、眼球運動 (ET) 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：[2)]

$$\begin{aligned}\log(\texttt{ET}) \sim\ & \texttt{length} + \texttt{prev\_length} \\ & + \texttt{freq} + \texttt{prev\_freq} \\ & + \texttt{is\_first} + \texttt{is\_last} + \texttt{is\_second\_last} \\ & + \texttt{lineN} + \texttt{segmentN} + (1|\texttt{subj}).\end{aligned}$$

また、先行研究 [13, 8] を考慮し、脳波 (EEG) 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：

$$\begin{aligned}\log(\texttt{EEG}) \sim\ & \texttt{sent\_order} + \texttt{word\_order} \\ & + \texttt{baseline\_activity} + (1|\texttt{subj}).\end{aligned}$$

各特徴量の詳細は付録 C に記す。数値型の特徴量は全て中心化を行なった。最初に一度モデル化した上で 3 標準偏差を超えるデータポイントを除外した。これにより、眼球運動、脳波のそれぞれで、92,246、及び 85,628、のデータポイントが最終的な分析対象となった。また、各言語モデルの PPP (Δ LogLik) の差が有意であるかどうかについては、各言語モデルで最も性能が良かったシードの結果について、ネストしたモデル比較を行い確かめた：比較対象となる言語モデルのサプライザルを共に含んだ回帰モデルから、着目している言語モデルのサプライザルを除いたときの逸脱度の増加分を χ^2 検定 ($p \leq 0.05$) により検定した。

3　結果と考察

統語的構成と自己注意をアーキテクチャに持つ/持たない 4 つの統語的言語モデルと、自己注意を持つ/持たない 2 つの統語的教示なしベースライン言語モデルの、PPP の結果を、図 2 に示した。縦軸が PPP を、横軸が各言語モデルを表す。棒グラフはシードの異なる 3 つのモデルの PPP の平均値を表し、それぞれの点は各シードの PPP を表す。また、PPP の有意差を検定するための、ネストしたモデル比較の結果を、表 2 に示した。統語的教示/統語的構成/自己注意を持つ言語モデルが、持たない言語モデルより PPP が高いかどうかを、その他の構成要素を持つ/持たない条件別に比較している。

1)　眼球運動のモデリングでは wrap-up 効果を捉えるために直前単語のサプライザルも共に加える (e.g., [13])。
2)　先行研究のベースライン説明変数のうち、本研究において有意でない ($p > 0.05$) ものについては除いた。

	first pass reading time			P600		
	χ^2	df	p	χ^2	df	p
統語的教示なし < 統語的教示あり：						
自己注意なし (LSTM<ActionLSTM)	13.823	2	**0.0009961**	4.6621	1	0.03084
自己注意あり (Transformer<PLM)	0.8069	2	0.668	3.2497	1	0.07144
統語的構成なし < 統語的構成あり：						
自己注意なし (LSTM+ActionLSTM<RNNG)	423.27	2	<**0.0001**	14.676	1	**0.0001277**
自己注意あり (Transformer+PLM<CAG)	453.86	2	<**0.0001**	11.588	1	**0.000664**
自己注意なし < 自己注意あり：						
統語的教示なし (LSTM<Transformer)	6.7846	2	**0.03363**	0.2135	1	0.644
統語的構成なし (ActionLSTM<PLM)	8.9226	2	**0.01155**	0.1564	1	0.6925
統語的構成あり (RNNG<CAG)	31.353	2	<**0.0001**	0.386	1	0.5344

表 2　心理学的予測精度 (PPP) の有意差を検定するための、ネストしたモデル比較の結果。統語的教示/統語的構成/自己注意を持つ言語モデルが、持たない言語モデルより PPP が高いかどうかを、その他の構成要素を持つ/持たない条件別に比較している。

眼球運動　First pass reading time の PPP の結果と、ネストしたモデル比較の結果は、それぞれ、図 2、表 2 の左に示されている。また、事前にベースライン回帰モデルとのネストしたモデル比較を行い、全ての言語モデルの PPP が有意であり、眼球運動のモデリングに効果があることを確かめた。図 2 より、統語的構成と自己注意の両者を持つ CAG が最も高い PPP を達成しており、統語的構成と自己注意の両者が、眼球運動という、人間のオンライン文処理から得られる間接的認知データの説明に、一定程度有効であることが示唆されている。さらに、表 2 より、統語的構成・自己注意の両者が、条件間で普遍的に有効であることが統計的に確かめられた。ベースライン言語モデルとの比較による、統語的教示の有効性については、条件間で統一した結論は得られていないが、これは統語的教示を与えることは必要十分ではなく、統語的構成のようなそれを扱うアーキテクチャを先天的にモデルに組み込むことが重要であることを示唆している [6, 8]。

脳波　P600 振幅の PPP の結果と、ネストしたモデル比較の結果は、それぞれ、図 2、表 2 の右に示されている。事前にベースライン回帰モデルとのネストしたモデル比較を行い、統語的構成を持つ RNNG と CAG の PPP のみが有意であり、その他の言語モデルは脳波のモデリングに効果がないことが確かめられた。図 2 より、それら RNNG と CAG 間では、RNNG がより高い PPP を達成しており、統語的構成は脳波に反映される高次処理に対応する可能性が高い (cf. [8]) が、自己注意はそれらの高次処理には対応しない可能性が高いことが示唆されている。さらに、表 2 より、統語的構成は条件間で普遍的に有効であることが統計的に確かめられた。自己注意や、統語的教示については、いずれの条件でも有効性が確かめられなかった。

総合考察　本研究の結果を総合すると、以下のような示唆が得られる。まず、文法的構成は、眼球運動のような間接的な認知データの説明力に留まらず、人間のオンライン文処理における高次処理に対応する可能性が高い。一方で、自己注意は、眼球運動に対する説明力はあるが、オンライン文処理における高次処理とは対応していない可能性がある。また、統語的教示自体は眼球運動・脳波の予測に必ずしも有効でなく、認知的妥当性の向上にはそれを扱うアーキテクチャ (i.e., 統語的構成) が必要である。

4　おわりに

本研究では、統語的構成と自己注意をアーキテクチャに持つ/持たない言語モデルの、人間の眼球運動・脳波のモデリングの精度を評価することで、それぞれの構成要素の「人間らしさ」を統一的に検証した。具体的には、それぞれの構成要素を持つ/持たない 4 つの統語的言語モデルと、自己注意を持つ/持たない 2 つの統語的教示なしベースライン言語モデルを、first pass reading time と P600 振幅で評価した。結果、統語的構成と自己注意は共に眼球運動はよく予測したが、脳波を上手く予測するのは統語的構成のみであった。これは、統語的構成は間接的な認知データの説明力に留まらず人間のオンライン言語処理における高次処理に対応するが、自己注意は高次処理とは乖離している可能性を示唆する。

謝辞

本研究は本研究は、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, pp. 1–10, February 2022.

[2] Tal Linzen. How Can We Accelerate Progress Towards Human-like Linguistic Generalization? In **ACL 2020**, pp. 5210–5217, Online, July 2020. Association for Computational Linguistics.

[3] Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent Neural Network Grammars. In **NAACL 2016**, pp. 199–209, San Diego, California, June 2016. Association for Computational Linguistics.

[4] Jeffrey L. Elman. Finding structure in time. **Cognitive Science**, Vol. 14, No. 2, pp. 179–211, April 1990.

[5] Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better. In **ACL 2018**, pp. 1426–1436, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[6] Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. Structural Supervision Improves Learning of Non-Local Grammatical Dependencies. In **NAACL 2019**, pp. 3302–3312, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[7] Jennifer Hu, Jon Gauthier, Peng Qian, Ethan Wilcox, and Roger Levy. A Systematic Assessment of Syntactic Generalization in Neural Language Models. In **ACL 2020**, pp. 1725–1744, Online, July 2020. Association for Computational Linguistics.

[8] John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. Finding syntax in human encephalography with beam search. In **ACL 20018**, pp. 2727–2736, Melbourne, Australia, July 2018. Association for Computational Linguistics.

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

[10] Ryo Yoshida, Hiroshi Noji, and Yohei Oseki. Modeling Human Sentence Processing with Left-Corner Recurrent Neural Network Grammars. In **EMNLP 2021**, pp. 2964–2973, Online and Punta Cana, Dominican Republic, January 2021. Association for Computational Linguistics.

[11] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is All you Need. In **Advances in Neural Information Processing Systems**, Vol. 30. Curran Associates, Inc., 2017.

[12] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In **NAACL 2019**, pp. 4171–4186, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[13] Danny Merkx and Stefan L. Frank. Human Sentence Processing: Recurrence or Attention? In **CMCL 2021**, pp. 12–22, Online, June 2021. Association for Computational Linguistics.

[14] Ryo Yoshida and Yohei Oseki. Composition, Attention, or Both? In **Findings of EMNLP 2022**, Online and Abu Dhabi, the United Arab Emirates, December 2022. Association for Computational Linguistics.

[15] John Hale. A Probabilistic Earley Parser as a Psycholinguistic Model. In **NAACL 2001**, pp. 159–166, 2001.

[16] Roger Levy. Expectation-based syntactic comprehension. **Cognition**, Vol. 106, No. 3, pp. 1126–1177, March 2008.

[17] Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. Lower Perplexity is Not Always Human-Like. In **ACL-IJCNLP 2021**, pp. 5203–5217, Online, August 2021. Association for Computational Linguistics.

[18] Eugene Charniak, Don Blaheta, Niyu Ge, Keith Hall, John Hale, and Mark Johnson. BLLIP 1987-89 WSJ Corpus Release 1, 2000.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. **Neural computation**, Vol. 9, No. 8, pp. 1735–80, December 1997.

[20] Transition-Based Dependency Parsing with Stack Long Short-Term Memory. In **ACL-IJCNLP 2015**, pp. 334–343, Beijing, China, July 2015. Association for Computational Linguistics.

[21] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving Language Understanding by Generative Pre-Training. p. 12, 2018.

[22] Peng Qian, Tahira Naseem, Roger Levy, and Ramón Fernandez Astudillo. Structural Guidance for Transformer Language Models. In **ACL-IJCNLP 2021**, pp. 3735–3745, Online, August 2021. Association for Computational Linguistics.

[23] Nora Hollenstein, Jonathan Rotsztejn, Marius Troendle, Andreas Pedroni, Ce Zhang, and Nicolas Langer. ZuCo, a simultaneous EEG and eye-tracking resource for natural sentence reading. **Scientific Data**, Vol. 5, No. 1, p. 180291, December 2018.

[24] Nora Hollenstein, Federico Pirovano, Ce Zhang, Lena Jäger, and Lisa Beinborn. Multilingual Language Models Predict Human Reading Behavior. In **NAACL 2021**, pp. 106–123, Online, June 2021. Association for Computational Linguistics.

[25] Lee Osterhout and Phillip J Holcomb. Event-related brain potentials elicited by syntactic anomaly. **Journal of Memory and Language**, Vol. 31, No. 6, pp. 785–806, December 1992.

[26] Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto. Reading-Time Annotations for "Balanced Corpus of Contemporary Written Japanese". In **COLING 2016**, pp. 684–694, Osaka, Japan, December 2016. The COLING 2016 Organizing Committee.

[27] Stephen Merity, Caiming Xiong, James Bradbury, and Richard Socher. Pointer Sentinel Mixture Models. In **ICLR 2017**, 2017.

変数名	型	概要
length	int	単語の文字数
prev_length	int	直前単語の文字数
freq	num	単語頻度の対数
prev_freq	num	直前単語頻度の対数
is_first	factor	行内最左要素
is_last	factor	行内最右要素
is_second_last	factor	行内最右から二番目の要素
lineN	int	画面内提示順
segmentN	int	行内提示順
subj	factor	被験者 ID
sent_order	int	文提示順
word_order	int	文内単語提示順
baseline_activity	num	視線停留後 0–100ms の平均振幅

表 3　本研究で用いた特徴量。

A　眼球運動データの前処理

眼球運動データについては、データセットの提案論文 [23] で施されている前処理に加えて、先行研究 [10] を踏襲し、(i) 視線停留がゼロの単語、(ii) 大規模コーパス (Wikitext-2, [27]) の語彙に含まれない単語、(iii) 語彙に含まれない語に後続する単語、を除いた。

B　脳波データの前処理

脳波データについては、データセットの提案論文 [23] で施されている前処理に加えて、先行研究 [8] を踏襲し、40Hz のローパスフィルタを適用し、さらに、平均再参照を行なった。また、視線停留後 0–100ms の平均振幅に対してベースライン補正を行なった。また、 (i) 視線停留がゼロの単語、(ii) 大規模コーパス (Wikitext-2) の語彙に含まれない単語、(iii) 文の最初及び最後に現れる単語、を除いた。

P600 の電極位置については、先行研究 [8] とはモンタージュが異なるため、厳密に同じではないが、後頭部電極を用いた。

C　回帰モデルの特徴量

本研究のベースライン回帰モデルに用いた特徴量を、表 3 に示した。頻度情報は大規模コーパス (Wikitext-2) より算出した。