認知的妥当性の高い言語モデルの文体的妥当性

松尾陽平 [1]　中西（大野）義典 [2]
[1] 同志社大学 大学院文化情報学研究科　[2] 同志社大学 文化情報学部
ctmk0012@mail4.doshisha.ac.jp　ynakanis@mail.doshisha.ac.jp

概要

様々な言語モデルが生成するコーパスと実際の人間の日本語書き言葉コーパスとを計量文体学の観点から比較し，認知的妥当性の高い言語モデルがもつ文体特徴を明らかにする．特に，認知的妥当性が向上すると近年指摘されている注意機構や統語構造に着目し，これらの構造的バイアスを利用した言語モデルが文体特徴に与える影響を分析した．調べた言語モデルの中では，注意機構・統語構造を共に利用する Composition Attention Grammar が実際の人間と最も類似した文体特徴をもつことが分かった．さらに，読点の直前に出現する語の分布や，後置詞を出現順に並べた列の bigram のみでは両コーパスの識別が困難になる結果が得られた．

1　はじめに

近年のニューラル言語モデルによる翻訳や要約，文章生成の性能の向上は目覚ましく [1–3]，人間にしかできないと考えられてきたタスクを自動的にできるようにすることは工学的にも重要である．しかしながら，言語モデルが有用であるのは，自然言語処理タスクの性能が向上するからというだけではない．生身の人間の言語活動がどのようになされているかを理学的な観点から明らかにするためにも有用であると考えられる．

計算心理言語学では，言語モデルの認知的妥当性をサプライザル理論に基づいて評価する研究が盛んに行われている [4, 5]．言語モデルの認知的妥当性の評価は，人間が文章を読みながら頭の中で処理するときの認知負荷を視線や脳波として計測し，こうして得られる生体情報と，言語モデルから算出される情報量とを比較することにより行う．経験的に人間は，それまでの文脈からはその位置に出現することが予想しづらい語句を目にしたときに驚きを感じ，それにより視線が停留することがある．サプライザル理論は，言語モデルが感じる「驚き」としてサプライザルという量を導入する．サプライザルは自己情報量とも呼ばれる．サプライザル理論では，人間が文章を読むときの語句ごとの視線停留時間を説明するときに，サプライザルの説明変数としての寄与が大きいことをもって，言語モデルの認知的妥当性が高いとする．

言語モデルの認知的妥当性の研究が進んでいる一方で，人間の言語活動は読む・書く・聞く・話すなど多岐に渡り，認知的妥当性の研究は読みという一側面を論じているにすぎない．書きに着目した最近の研究として，計量文体学では，GPT-3.5 および GPT-4 が生成した日本語の文章と人間が執筆した和文学術論文中の文章とを比較し，トピックに依存しないとされる代表的な文体特徴量により両者を高い精度で識別可能であることが示されている [6]．しかしながら，これらの言語モデルはその詳細が公開されていないため認知的妥当性を評価することが難しく，書きについて得られた知見を読みと関連させて議論するには不十分である．

本研究では，計量文体学の方法を用いて，認知的妥当性が評価されている様々な言語モデルが生成するコーパスと実際の人間の書き言葉コーパスとを比較する．特に，注意機構や統語構造が言語モデルの認知的妥当性を向上させるという最近の計算心理言語学の知見を踏まえ，これらの構造的バイアスの有無が言語モデルの文体的妥当性に与える影響を調べる [7]．読み書き双方の観点から人間と言語モデルとを比較し，高い認知的・文体的妥当性をもつ言語モデルの構造を解明することを通じて，人間の言語活動を複眼的に理解することを目指す．

2　方法

本研究で文体的妥当性を評価する言語モデルは，Long Short-Term Memory（LSTM），Transformer，Recurrent Neural Network Grammar（RNNG），Composition Attention Grammar（CAG）の 4 つである．これらはモデルが利用する構造的バイアスに差があ

表 1　構造的バイアスによる言語モデルの分類

	注意機構無し	注意機構有り
統語構造無し	LSTM	Transformer
統語構造有り	RNNG	CAG

る．表 1 にまとめる通り，LSTM は注意機構も統語構造も利用しない最も素朴なモデルであり [8]，Transformer は注意機構を利用するモデル [2]，RNNG は統語構造を利用するモデル [9]，CAG は注意機構・統語構造を共に利用するモデルとして位置付けられる [10]．

読みに関する認知的妥当性の観点からこれらのモデルを整理すると，まず，注意機構を利用する Transformer が LSTM より高い認知的妥当性を示すことが報告されている [11]．また，統語構造を利用する RNNG の認知的妥当性の高さも示されている [12]．さらに最近では，注意機構と統語構造を共に利用する CAG が，それぞれを単独で利用する Transformer や RNNG と比べて，高い認知的妥当性を示すことが報告されている [7]．

言語モデルの学習は『NINJAL Parsed Corpus of Modern Japanese』（NPCMJ）を用いて行う [13]．これは統語解析情報が付いた日本語の書き言葉コーパス（一部，話し言葉を含む）であり，特に統語構造を利用した RNNG や CAG の学習に有用である．言語モデルの実装には，LSTM は Gulordava et al. (2018) [14]，RNNG は Noji and Oseki (2021) [15]，Transformer は Huggingface の Transformers パッケージ [16]，CAG は Yoshida and Oseki (2022) [10] を用いた．

学習済みの言語モデルを用いて生成した文章の文体特徴量を算出する．日本語の文章に関する著者推定の研究において，内容やトピックによらず文体の特徴を捉えられるとして有効な「読点の位置」，「機能語の割合」，「品詞の bigram」，「後置詞の bigram」に着目する [17]．読点の位置とは，文中の読点の直前に出現する語の頻度分布とする．機能語の割合とは名詞，動詞，形容詞などを除く機能語が文中に出現する割合のことである．品詞の bigram とは，文中の各語に品詞タグを付与して得られる品詞タグ列における bigram である．後置詞の bigram とは，文中の助詞，助動詞を出現順に並べて得られる後置詞列における bigram である．なお，形態素解析には MeCab を用い，UniDic に基づく品詞体系を採用する．

4 種類の文体特徴量の観点から，言語モデルが生成するコーパスと人間が執筆した書き言葉コーパスとを比較し，言語モデルの文体的妥当性を評価する．人間の書き言葉コーパスとして，各言語モデルの学習にも用いた NPCMJ を用いる．まず，各言語モデルのコーパスと人間のコーパスとを，文体特徴量を説明変数として，ランダムフォレストにより識別したときの正解率を調べる．次に，文体特徴量に基づく主成分分析を実施し，各コーパスを低次元空間において可視化するとともに，言語モデルの構造的バイアスが文体的妥当性に与える影響を調べる．

3　結果と考察

4 つの言語モデル（LSTM，Transformer，RNNG，CAG）からそれぞれ 8000 文を生成し，NPCMJ を構成する約 90000 文と比較する．まず，各文から抽出した文体特徴量に基づいて，ランダムフォレストにより識別したときの正解率を表 2 に示す．4 種類の文体特徴量全てを用いて識別したときの正解率は，いずれの言語モデルでも 100%に近い値を示した．全ての文体特徴量を用いたとき，文の著者が人間であるか言語モデルであるかを特定するのは容易であり，絶対的な意味での文体的妥当性はそれほど高くないことが分かった．

言語モデル間の相対的な文体的妥当性を比較することも重要である．それぞれの文体特徴量を 1 種類ずつ用いたときの識別正解率の結果によると，機能語の割合や品詞の bigram では言語モデル間の優劣は見られなかったが，読点の位置および後置詞の bigram では特徴的な結果が見られた．読点の位置では，識別正解率が高い順に LSTM，Transformer，RNNG，CAG である．特に，CAG の識別正解率は 50%程度まで低下しており，読点の位置に関する文体的妥当性は，他の言語モデルに比べて高いということが分かる．後置詞の bigram では，識別正解率が高い順に Transformer，LSTM，RNNG，CAG である．後置詞の bigram についても CAG の文体的妥当性は他のモデルより高い．

次に，4 種類全ての文体特徴量に対する主成分分析の結果に基づいて，各コーパスを低次元空間に射影した結果を図 1 に示す．主成分得点空間上で，灰色の点で示される NPCMJ の文が分布する領域と，言語モデルが生成する文が分布する領域との重なりを比較する．ピンク色の点で示される CAG の文が

表 2　文体特徴量に基づく言語モデルが生成するコーパスの NPCMJ との識別正解率．単位は%．

	全ての特徴量	読点の位置	機能語の割合	品詞の bigram	後置詞の bigram
LSTM	100.0	93.8	98.6	100.0	93.8
Transformer	99.4	82.9	94.7	100.0	99.3
RNNG	100.0	80.6	99.2	100.0	92.7
CAG	99.6	51.1	95.2	99.3	79.6

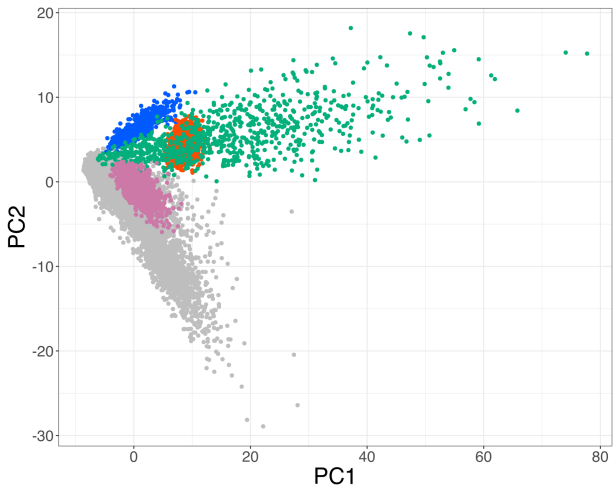

図 1　コーパス中の各文を主成分得点空間に射影した結果．青が LSTM，緑が Transformer，赤が RNNG，ピンク色が CAG であり，灰色が NPCMJ である．4 種類全ての文体特徴量についての第 1 主成分（PC1），第 2 主成分（PC2）をそれぞれ横軸，縦軸とする．

分布する領域との重なりが最も大きく，NPCMJ の領域の一部として含まれるような位置関係にある．ランダムフォレストによる識別正解率の結果から示唆される相対的な文体的妥当性の高さに整合するものであると考えられる．

各言語モデルの構造的バイアスの観点から考える．読点の位置および後置詞の bigram に関する識別正解率についていえば，統語構造を利用する RNNG や CAG が，利用しない LSTM や Transformer に比べて低い識別正解率（高い文体的妥当性）を示している．このことは，最近の計算心理言語学の研究で指摘されている [7]，認知的妥当性の高さに影響を与える構造的バイアスは注意機構よりもむしろ統語構造であるということと関連しうる．しかしながら，読点の位置については注意機構のみを利用する Transformer と統語構造のみを利用する RNNG とが僅差であり，また，後置詞の bigram については CAG を除く 3 つのモデルが拮抗していることからも，文体的妥当性については詳細な分析が必要である．

図 2(a)–(e) に各コーパスを構成する各文の主成分得点に関する箱ひげ図を示す．これまでの結果から文体的妥当性が相対的に高い CAG は，第 5 主成分についていえば他のモデルよりも人間（NPCMJ）との類似度が低いが，第 1 主成分から第 4 主成分までの範囲についていえば，最も人間に近い文体特徴をもつ言語モデルの一つであるといえる．注意機構と統語構造との違いを詳細に分析する上で，第 4 主成分に関する結果が興味深い．第 1 主成分から第 3 主成分まででは，Transformer と RNNG は人間から見て同じ向きに離れており，注意機構と統語構造それぞれが文体に与える影響を分離して考えることは難しいが，第 4 主成分まで調べて初めて，Transformer と RNNG が人間から見て反対の向きに離れるという結果を得た．

第 4 主成分に対する各文体特徴量の負荷量を図 3 に示す．第 4 主成分は，局所的な名詞の繋げ方を表していると考えられる．絶対値の大きい正の負荷量をもつのは，品詞の bigram における「助詞（連体化）-格助詞」や，後置詞の bigram における「の-の」や「を-の」などの内容語を繋げる役割をもつ文体特徴量である．また，絶対値の大きい負の負荷量をもつのは，品詞の bigram における「名詞（一般）-名詞（サ変接続）」のような複合語に関する役割をもつ文体特徴量である．以上のことから，第 4 主成分得点が高いモデルは，助詞を用いて名詞と名詞とをつなぐ表現が多く，その一方で，第 4 主成分得点が低いモデルは，複合名詞として名詞を繋げる表現が多いことが示唆される．これらに基づいて，構造的バイアスが言語モデルの文体に与える影響を議論していくことが期待される．

4　おわりに

本研究では，様々な言語モデルが生成するコーパスと実際の人間の日本語書き言葉コーパスとを計量文体学の観点から比較し，認知的妥当性の高い言語

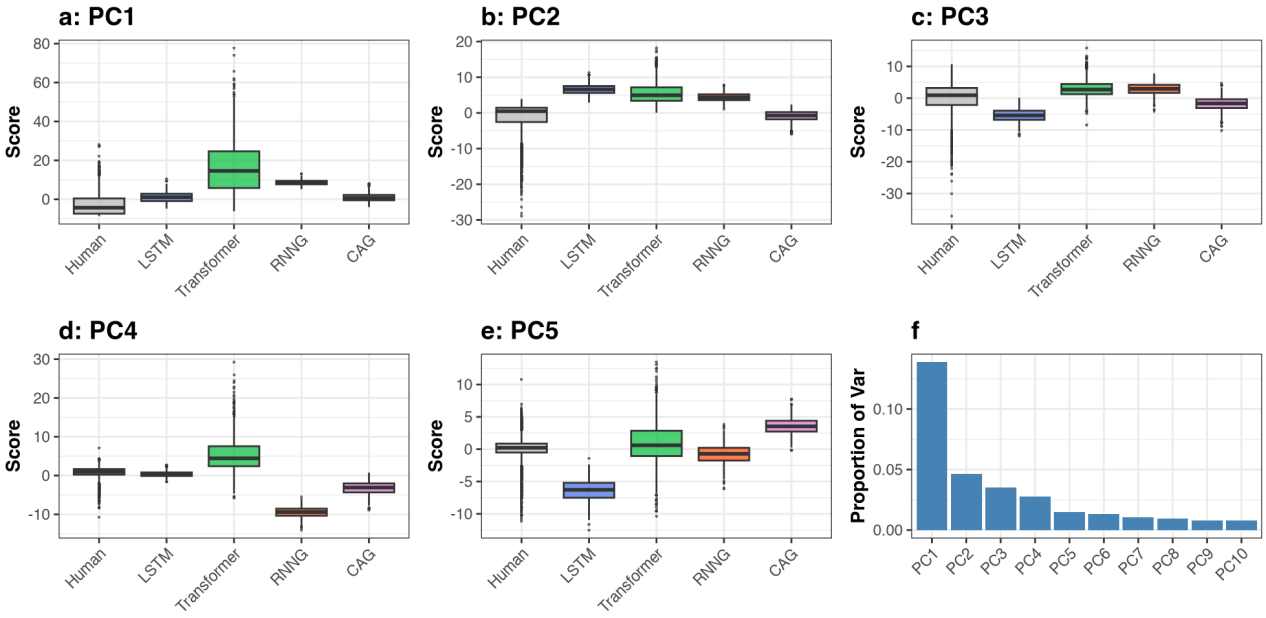

図 2　(a)–(e) コーパス中の各文の主成分得点に関する箱ひげ図．(a) から (e) までが，順に，第 1 主成分（PC1）から第 5 主成分（PC5）までを示す．(f) 主成分分析における各成分の寄与率．

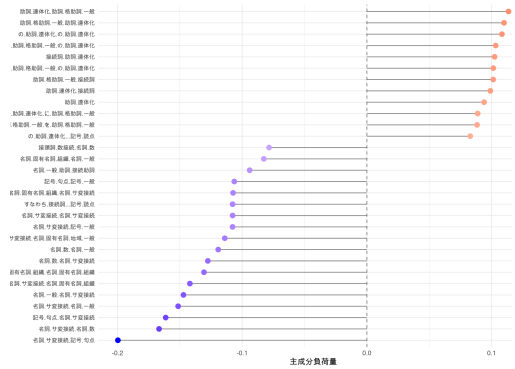

図 3　各文体特徴量の第 4 主成分に対する負荷量

モデルがもつ文体特徴を明らかにした．その結果，調べた言語モデルの中では，注意機構・統語構造を共に利用する CAG が実際の人間と最も類似した文体特徴をもつことが分かった．さらに，読点の直前に出現する語の分布や，後置詞を出現順に並べた列の bigram のみでは両コーパスの識別が困難になる結果が得られた．また，主成分分析の結果を詳細に調べることにより，注意機構と統語構造それぞれが文体に与える影響を分析した．

参考文献

[1] Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In Lluís Màrquez, Chris Callison-Burch, and Jian Su, editors, **Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing**, pp. 379–389, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[2] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**, 2017.

[3] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. Technical report, OpenAI, 2019.

[4] John Hale. A probabilistic earley parser as a psycholingusitic model. In **Proceedings of the second meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies**, 2001.

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

[6] Wataru Zaitsu and Mingzhe Jin. Distinguishing chatgpt (-3.5,-4)-generated and human-written papers through japanese stylometric analysis. **PLoS One**, Vol. 18, No. 8, p. e0288453, 2023.

[7] Ryo Yoshida, Yushi Sugimoto, and Yohei Oseki. Investigating psychometric predictive power of syntactic attention. In **Proceedings of the 29th Conference on Computational Natural Language Learning**, pp. 407–418, 2025.

[8] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. **Neural computation**, Vol. 9, No. 8, pp. 1735–1780, 1997.

[9] Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent neural network grammars. In **Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies**, pp. 199–209. ACL, 2016.

[10] Ryo Yoshida and Yohei Oseki. Composition, attention, or both? In **Findings of the Association for Computational Linguistics: EMNLP 2022**, pp. 5822–5834, 2022.

[11] Danny Merkx and Stefan L Frank. Human sentence processing: Recurrence or attention? In **Proceedings of the workshop on cognitive modeling and computational linguistics**, pp. 12–22, 2021.

[12] 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, 2018.

[13] 国立国語研究所. 『NINJAL Parsed Corpus of Modern Japanese』（バージョン 1.0）, 2016. https://npcmj.ninjal.ac.jp/interfaces/（2024 年 9 月 3 日確認）.

[14] Kristina Gulordava, Piotr Bojanowski, Edouard Grave, Tal Linzen, and Marco Baroni. Colorless green recurrent networks dream hierarchically. In **Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)**, pp. 1195–1205, 2018.

[15] Hiroshi Noji and Yohei Oseki. Effective batching for recurrent neural network grammars. In **Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021**, pp. 4340–4352, 2021.

[16] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, et al. Transformers: State-of-the-art natural language processing. In **Proceedings of the 2020 conference on empirical methods in natural language processing: system demonstrations**, pp. 38–45, 2020.

[17] Wataru Zaitsu. Tekisuto mainingu niyoru hisshashikibetsu no seikakusei narabini hantei tetsuduki no hyoujunka [accuracy and standardized judgment procedures for author identification]. **Jpn. J. Behaviormetrics**, Vol. 45, No. 1, p. 39, 2018.