予測の正確な言語モデルがヒトらしいとは限らない

栗林 樹生 [1,2]　大関 洋平 [3,4]　伊藤 拓海 [1,2]　吉田 遼 [3]　浅原 正幸 [5]　乾 健太郎 [1,4]
[1] 東北大学　[2]Langsmith 株式会社　[3] 東京大学　[4] 理化学研究所　[5] 国立国語研究所
{kuribayashi, t-ito, inui}@ecei.tohoku.ac.jp
{oseki, yoshiryo0617}@g.ecc.u-tokyo.ac.jp, masayu-a@ninjal.ac.jp

1 はじめに

文章を読んでいると，ある部分はスラスラ読めたり，特定の箇所ではつっかえたりする．このような逐次的な読みやすさはどのように計算されるのだろうか．本研究では，ヒトの漸進的な文処理の計算モデルについて洞察を得ることを目指す．

近年では，単語や文のサプライザル（$-\log p($ 単語や文 $\mid$ 先行文脈 $)$）が読みやすさを決める主要な要因であるとする**サプライザル理論** [1, 2] が支持されている．ヒトは先読みをしながら文章を読んでおり，予想と異なる情報が出現する（サプライザルが大きくなる）と処理負荷が高まる（例えば，読み時間が長くなる）という説である．本理論はサプライザルをどのようなモデルで計算するかについては中立的であり，ヒトの読み活動をうまくモデリングできるようなサプライザルを算出するモデルを探求することで，構成論的なアプローチによりヒトの文処理に対して洞察を得てきた [3, 4, 5, 6, 7, 8]．本研究でも，モデルから得られたサプライザルがヒトの読み活動をどれほどモデリングできるかを「ヒトらしさ」の指標とし，ヒトらしいモデルを探求する．

最近ではサプライザル理論に基づいた実験から，**パープレキシティ（PPL）の低い言語モデルほどヒトらしい**という報告がされてきた [3, 4, 9, 10, 11]．本研究ではこの報告の一般性について再検証し，例えば自然言語処理分野で行われている訓練データ量やパラメータ数を増やすといった改良 [12] の先に認知科学が目指すヒトらしいモデルがあるのか，両分野が向かう先の関係について示唆を与えることを目指す．図 1 に示すとおり言語モデルの予測の正確さ（PPL の低さ）はヒトらしさを必ずしも含意しない．また既存研究では英語の限られたコーパス上でのみ議論が行われており，様々な言語や現象において予測の正確さ（PPL の低さ）とヒトらしさの間の関係が成り立つかは定かでない．

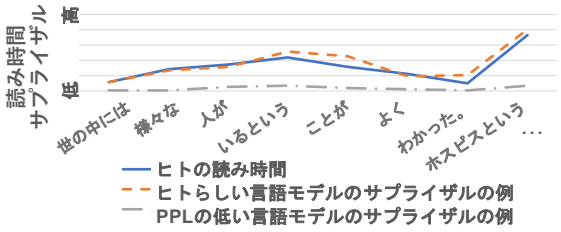

図 1　PPL の低い言語モデルとヒトらしい言語モデルの要略．縦軸はヒトの読み時間と言語モデルのサプライザル．

省略，語順などの観点で英語と大きく異なる日本語に焦点を当てて PPL とヒトらしさの関係について検証した．実験結果より，PPL が低い言語モデルほどヒトらしいという英語で観察された関係は日本語では必ずしも成立せず，本関係は**言語横断的な一般性を欠く観察である**と一旦結論づけた．また乖離の傾向について，言語の階層的な構造を明示的に教師として与えていない言語モデルが，ヒトの読み活動よりも統語構造に過剰に敏感なサプライザルを計算するという観察が得られた．具体的には，主辞後置言語の読み時間で典型的に観察される統語構造に関する効果 [13] の観点で，言語モデルがヒトよりも強くバイアスを受けていた．本乖離について，日本語と英語の違いに結びつけて議論する．本研究で BCCWJ-EyeTrack に対して付与した 111 の言語モデルによるサプライザルデータ（BCCWJ-UniSeqLM）は公開する[1)]．

2 関連研究

ヒトの漸進的な文処理モデルの解明に向けて，認知科学，心理言語学の分野ではヒトの読み活動（読み時間や脳活動データ）が長らく分析され

1)　https://github.com/kuribayashi4/BCCWJ-UniSeqLM.
BERT などの双方向言語モデル（bidirectional）や，RNNG などの階層構造を明示的に考慮する言語モデル（hierarchical）と対比させて “unidirectional sequential LM” と総称する．

てきた [14, 13, 1, 2, 3, 4, 15, 16]．読み時間については，特に主辞先行言語（英語など）と主辞後置言語（日本語など）間で異なる傾向が報告されており [13, 17, 18, 19]，特定の言語で報告された観察（本研究の文脈では，PPL∝ヒトらしさ）が異なる言語で成り立つかは自明でない．これまでも，主辞先行言語で提案されたワーキングメモリによる読み時間の説明 [14] について，主辞後置言語におけるヒトの傾向をうまく説明できないことが指摘され [13, 17]，後の anti-locality theory やサプライザル理論に繋がった [13, 1, 2]．

3 実験設定: 言語モデル

サブワード[2)]を入力単位とした文章レベルの left-to-right 言語モデルを用いる．長さ N の文 $w_{1:N}$ における長さ T の文節 $b = w_{k:k+T}$ $(1 < k \leq k+T \leq N)$[3)] のサプライザルを以下のように計算する：

$$\begin{aligned}\text{surprisal}(b) &= -\log_2 p(w_k, \cdots, w_{k+T} | w_1, \cdots, w_{k-1}) \\ &= -\sum_{i=k}^{k+T} \log_2 p(w_i | w_1, \cdots, w_{i-1}) \ .\end{aligned}$$

言語モデルの種類: パラメータ数の異なる 2 種類の Transformer 言語モデル（400M パラメータの Trans-L と 55M パラメータの Trans-S）と LSTM ベースの言語モデルについて，学習データ量（1.4G, 140M, 14M サブワード）とパラメータアップデート回数（100K, 10K, 1K, 0.1K）を変えて学習し，さらにそれぞれの設定について 3 つの異なるランダムシード[4)]でモデルを学習した（$3 \times 3 \times 4 \times 3 = 108$ モデル）．学習データは新聞記事と日本語 Wikipedia から成る．さらに，3 グラム，4 グラム，5 グラム言語モデル[5)]も加え，計 111 の設定について分析した．

4 実験 1: PPL とヒトらしさ

各言語モデルについて，PPL とヒトらしさの関係を調べる．あらかじめベースライン特徴量で読み時間をモデリングし，ある言語モデルから得られたサプライザルを固定因子として追加した際にモデリング性能がどれほど上昇するかで言語モデルのヒトらしさを評価する．具体的には，既存研究に従い，サプライザルを考慮する前後における読み時間データの文節平均対数尤度の変化 ΔLogLik を報告する．ΔLogLik が大きいほど，その言語モデルがヒトらしいことを示す．読み時間（RT）のモデリングは以下の式で行う：

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

各特徴量の詳細は付録に示す．(1|x) は x をランダム切片として組み込むことを指す．ベースラインモデルでは式 1 の説明変数からサプライザル（surprisal）を除く．既存研究に従い [22]，視線走査法により計測された対数注視時間（first pass time）をモデリングした．予測モデルとして一般化線形混合モデル（GLMM）を用いた．

既存研究 [18] に従い読み時間がゼロ秒である文節は除いた．また，読み時間について 3 標準偏差を超える文節についても除外した．最終的に 13,148 のデータポイントを用いた．

4.1 結果

図 2 に結果を示す．各プロットは各言語モデルに対応する．X 軸が言語モデルの PPL（対数スケール），Y 軸が言語モデルの読み時間モデリング能力であり，モデリング能力が高いことと本研究でいう「ヒトらしい」ことは同義である．グラフが左肩上がりの場合は PPL が低いほどヒトらしいと言える．Goodman ら [9] は，言語モデルの PPL が低くなるほど読み時間モデリング能力が向上し，両者には直線的な関係があることを報告した．一方で本実験結果では，言語モデルの PPL と読み時間モデリング能力の間のピアソン相関係数は-0.21 であり，さらに PPL が 0 から 1000 の範囲では，PPL の低い言語モデルほど**ヒトらしくない**という正の相関が見られた（ピアソン相関係数 0.19）．これらの結果から一旦，PPL の低い言語モデルほどヒトらしいという知見は言語横断的な一般性を欠くと結論づける．

アーキテクチャ，データ量，アップデート回数別の読み時間モデリング能力の平均を表 1 に示す．スコアが高いほどヒトの読み時間のモデリングに貢献するサプライザルを計算することを意味する．一般的な言語モデルの性能に対する傾向に反し，N-gram

2) mecab [20] と unidic で国語研短単位に分割した後，バイト対符号化 [21] によってさらに分割した（character coverage=0.9995, vocab size=100000）．
3) BOS トークン (w_1) の存在から，文節は w_2 以降から始まる．
4) 140M，14M サブワードのデータで学習するモデルについては，データのサンプリングも変えている．
5) 学習データ（1.4G サブワード）をすべて用い https://github.com/kpu/kenlm でモデルを作成した．

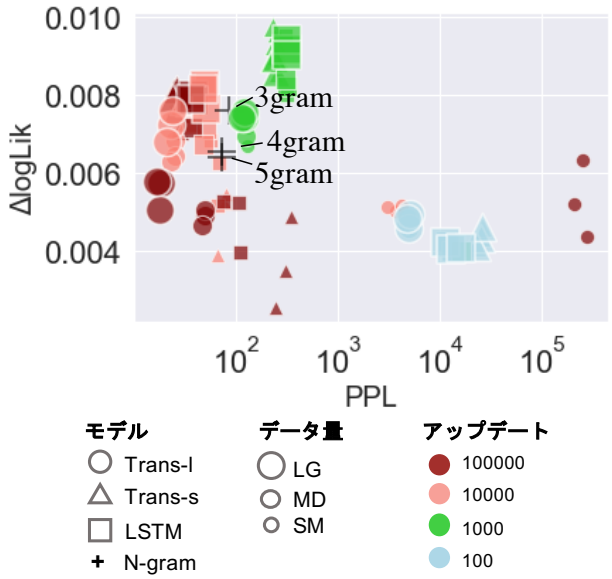

図 2　PPL と読み時間モデリング能力の関係.

表 1　左からアーキテクチャ，学習データ量，アップデート回数ごとの読み時間モデリング能力の比較.

	ΔLogLik		ΔLogLik		ΔLogLik
TRANS-L	0.0059	1.4G	**0.0069**	100K	0.0061
TRANS-S	0.0067	140M	0.0063	10K	0.0066
LSTM	0.0059	14M	0.0056	1K	**0.0075**
N-GRAM	**0.0068**			0.1K	0.0045

言語モデルやアップデートが比較的少なめ（1000 回）のモデルのモデリング性能が高いことが分かった．N-gram 言語モデルが比較的ヒトらしいことは英語における既存研究でも報告されている [9]．5 節では，各言語モデルがもつ性質について分析をし，どのような観点で言語モデルがヒトの読み活動データから逸脱していくか調べる．

5　実験 2: 統語情報とサプライザル

読み時間が統語範疇や統語構造の影響を受けているという知見に従い [18, 23]，本研究では統語範疇と統語構造の 2 つの観点から各言語モデルの性質を調べる．具体的に本研究では，各言語モデルが計算するサプライザルの統語範疇と統語構造に対する敏感さを測定する．

言語モデルが各文節に対して計算するサプライザルを擬似的な読み時間とみなし，既存研究 [18, 23] と同様にモデリングを行う．具体的には 4 節と同様，はじめに各言語モデルが計算するサプライザルをベースライン特徴量（式 1 の `surprisal` を除く説明変数）[6] によってモデリングする（ベースライン

6）　スクリーン上の提示位置などといったヒトに対する測定で特有に生じる要因については，素性の設計を一部変更している．詳しくは付録に記載.

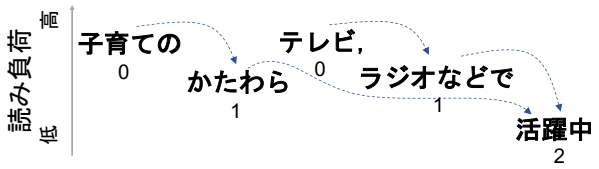

図 3　Anti-locality 効果の要略．矢印は係り受け関係を，各文節の下の数字は先行文脈に存在する係り元文節の数を表し，この値が大きいほど読み負荷（Y 軸方向の位置）が小さくなるとされている.

モデル）．続いて統語範疇や統語構造に関する特徴量を追加したときのモデリング性能の具合いである（文節平均対数尤度の変化）ΔLogLik を調べる．

統語範疇に対する敏感さ　BCCWJ 中の各文節は用の類，相の類，体の類，その他，分類不能のいずれかのカテゴリに分類されている．ベースラインモデルに対して各文節がどのカテゴリに属すかという説明変数を加え，サプライザルに対する対数尤度がどれほど変化するかを測定する．

統語構造に対する敏感さ　日本語を含む主辞後置言語では，単語や文節の読みやすさが文内の先行文脈に出現する係り元の要素数の影響を受ける anti-locality 効果が報告されている [13, 17]．ある要素の係り元が前方にたくさん存在するほど，その要素を予測する手がかりが多くなり，読み負荷が下がるという仮説である．図 3 に anti-locality 効果の概略を示す．anti-locality 理論に基づくと，例えば係り元が存在しない図 3 中の「テレビ」よりも，直接係り元が 2 つ存在する「活躍中」の方が読み負荷が低いとされる．

各言語モデルが統語バイアスをどれほど強く有するかを検証するため，ベースラインモデルに対して先行文脈に存在する係り元の数を説明変数として追加し，対数尤度がどれほど変化するかを測定する．

5.1　結果

各言語モデルが計算するサプライザルについて，統語範疇と統語構造（anti-locality 効果）に対する敏感さを測定した結果が図 4 である．各プロットが各言語モデルに対応し，縦軸がそれぞれの特徴量に対する敏感度である．図 4 左では X 軸を各言語モデルの読み時間モデリング能力，図 4 右では X 軸を PPL としている．まず図 4 左より，読み時間予測性能の高いモデルほど統語範疇に対して強く敏感であることが分かる（ピアソン相関係数で 0.85）．また図 4

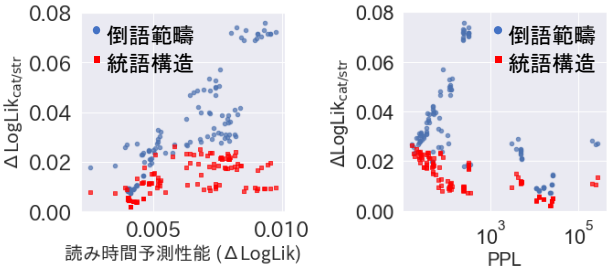

図 4　言語モデルの PPL，ヒトらしさ，統語に対する敏感さの関係．

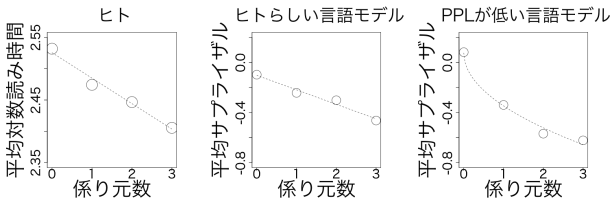

図 5　ヒトの読み時間，最もヒトらしい言語モデルのサプライザル，PPL の低い言語モデルのサプライザルにおける，anti-locality 効果の違い．

右より，ある程度 PPL の高い（およそ PPL200）言語モデルは統語範疇に対して敏感であるのに対し，さらに PPL が下がっていくと統語構造に対して敏感になることが観察され，PPL と統語構造に対する敏感さの間には順位相関係数-0.79 の相関が見られた．統語範疇から統語構造への敏感さの変化や，目的関数として構造を明示的に与えていないにも関わらず言語モデルが統語構造に敏感になっていく様子は言語獲得の観点からも興味深い．

これらの観察から，日本語話者は統語構造よりも統語範疇に近いレベルで先読みをしており，PPL の低い言語モデルはより高度な統語構造のレベルで文を処理していることで乖離が生じている可能性があげられる．分析では，PPL の低い言語モデルがどのように統語構造に対して敏感であるかを調査する．

5.2　分析

統語構造の観点におけるヒトと言語モデルの乖離を詳細に分析する．図 5 に，ある文節に対する先行文脈に存在する係り元の数と平均読み時間・サプライザルを示す．左からヒトの対数読み時間，読み時間モデリング性能の高い言語モデルが算出するサプライザル，PPL の最も低い言語モデルが算出するサプライザルである．ヒトの対数読み時間や読み時間予測性能の高い言語モデルでは直線的な anti-locality 効果が観察されるのに対し，PPL の低い言語モデルでは，anti-locality 効果が曲線的であった．本曲線に対する一つの解釈として，PPL 最小化の目的関数のもと言語モデルを訓練していくと，文内に係り元のない文節（以降，DEP_0）の予測が係り元の存在する文節と比べて相対的に苦手になり，DEP_0 に対して不当に高いサプライザルを計算してしまっている可能性をあげる．なお，英語ではこのような係り元の数という観点における乖離は観察されなかった（付録）．

DEP_0 の予測については文内に統語的な手がかりが存在せず，例えば先行文脈を考慮した自然な主題展開といった談話レベルの処理が必要になると考えられる．近年，談話レベルの文章展開において言語モデルとヒトの間に乖離があることが報告されており [24]，更に日本語では顕在性の高い話題などは省略されることから，文章に表出されている情報のみで談話的な手がかりをモデリングすることが難しいと考えられる．一方でヒトは例えば省略された話題などを補って文章を読んでいると考えられ，結果的に文を超えた予測が必要になる DEP_0 における読み負荷 (読み時間とサプライザル) の乖離という形で，ヒトと言語モデルの違いが観察されているという解釈をあげる．

6　おわりに

本研究では PPL の低い言語モデルほどヒトらしいという知見に対して既存研究とは異なる観察を提示し，分野が目指すモデルの間の関係に示唆を与えた．具体的には，言語モデルの予測が正確になる（PPL が低くなる）につれて，ヒトの文処理モデルに近づくとは**限らない**ということを示した．既存研究と異なる結果が得られた原因については，本稿であげた解釈の他にも読み時間測定時のノイズやサプライザルを通して評価すること（サプライザル理論）の妥当性といった様々な要因が考えられるため，引き続き調査をすすめる必要がある．一つの方向性として，日本語以外の言語についても検証を進め，言語のどのような性質が乖離に結びつくのか分析していきたい．

謝辞．　本研究は JSPS 科研費 JP20J22697，19H04990 の助成を受けたものです．また，国立国語研究所共同研究プロジェクト「大規模コーパスを利用した言語処理の計算心理言語学的研究」の支援を受けたものです．

参考文献

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

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

[3]Stefan L. Frank and Rens Bod. Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Journal of Psychological Science*, Vol. 22, No. 6, pp. 829–834, 2011.

[4]Victoria Fossum and Roger Levy. Sequential vs. Hierarchical Syntactic Models of Human Incremental Sentence Processing. In *Proceedings of CMCL*, pp. 61–69, Montréal, Canada, 6 2012. Association for Computational Linguistics.

[5]John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R. Brennan. Finding syntax in human encephalography with beam search. In *Proceedings of ACL*, pp. 2727–2736, 2018.

[6]Tal Linzen. What can linguistics and deep learning contribute to each other? *Journal of Language*, Vol. 95, No. 1, pp. 99–108, 2019.

[7]Danny Merkx and Stefan L. Frank. Comparing Transformers and RNNs on predicting human sentence processing data. *arXiv preprint arXiv:2005.09471,2020*, 2020.

[8]Martin Schrimpf, Idan Blank, Greta Tuckute, Carina Kauf, Eghbal A Hosseini, Nancy Kanwisher, Joshua Tenenbaum, and Evelina Fedorenko. The neural architecture of language: Integrative reverse-engineering converges on a model for predictive processing. *bioRxiv*, 2020.

[9]Adam Goodkind and Klinton Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of CMCL2018*, pp. 10–18, 2018.

[10]Christoph Aurnhammer and Stefan Frank. Comparing gated and simple recurrent neural network architectures as models of human sentence processing. In *Proceedings of the 41st Annual Conference of the Cognitive Science Society*, pp. 112–118, 2019.

[11]Ethan Gotlieb Wilcox, Jon Gauthier, Jennifer Hu, Peng Qian, and Roger Levy. On the Predictive Power of Neural Language Models for Human Real-Time Comprehension Behavior. In *Proceedings of CogSci*, pp. 1707–1713, 2020.

[12]Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[13]Lars Konieczny. Locality and parsing complexity. *Journal of Psycholinguistic Research*, Vol. 29, No. 6, pp. 627–645, 2000.

[14]Edward Gibson. Linguistic complexity: Locality of syntactic dependencies. *Journal of Cognition*, Vol. 68, No. 1, pp. 1–76, 1998.

[15]Stefan L Frank, Leun J Otten, Giulia Galli, and Gabriella Vigliocco. The ERP response to the amount of information conveyed by words in sentences. *Brain and language*, Vol. 140, pp. 1–11, 2015.

[16]Jonathan R Brennan, Edward P Stabler, Sarah E Van Wagenen, Wen-Ming Luh, and John T Hale. Abstract linguistic structure correlates with temporal activity during naturalistic comprehension. *Brain and language*, Vol. 157, pp. 81–94, 2016.

[17]Shodai Uchida, E Miyamoto, Yuki Hirose, Yuki Kobayashi, and Takane Ito. An ERP study of parsing and memory load in Japanese sentence processing – A comparison between left-corner parsing and the Dependency Locality Theory. *Technical report of IEICE. Thought and language*, Vol. 114, pp. 101–106, 2014.

[18]Masayuki Asahara, Hajime Ono, and Edson T Miyamoto. Reading-Time Annotations for "Balanced Corpus of Contemporary Written Japanese". In *Proceedings of COLING*, pp. 684–694, 2016.

[19]Masayuki Asahara. Between reading time and clause boundaries in Japanese - wrap-up effect in a head-final language. In *Proceedings of PACLIC*, pp. 19–27, 2018.

[20]Taku Kudo. Mecab: Yet another part-of-speech and morphological analyzer. *http://mecab. sourceforge. jp*, 2006.

[21]Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. In *Proceedings of ACL*, pp. 1715–1725, 2016.

[22]Nathaniel J. Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Journal of Cognition*, Vol. 128, No. 3, pp. 302–319, 2013.

[23]Masayuki Asahara and Sachi Kato. Between Reading Time and Syntactic / Semantic Categories. In *Proceedings of IJCNLP*, pp. 404–412, 2017.

[24]Shiva Upadhye, Leon Bergen, and Andrew Kehler. Predicting Reference: What do Language Models Learn about Discourse Models? In *Proceedings of EMNLP2020*, pp. 977–982, Online, 11 2020. Association for Computational Linguistics.

[25]Alan Kennedy, Robin Hill, and Joël Pynte. The dundee corpus. In *Proceedings of the 12th European conference on eye movement*, 2003.

[26]Maria Barrett, Željko Agi, and Anders Søgaard. The Dundee Treebank. In *Fourteenth International Workshop on Treebanks and Linguistic Theories*, pp. 242–248, 2015.

表 2　本実験で用いた素性.

変数名	型	記述
surprisal	実数	サプライザル
RT	実数	読み時間 (ms)
article	カテゴリカル	記事番号
screenN	整数	画面呈示順
lineN	整数	何行目か
segmentN	整数	画面右から文節目か
sentN	整数	何文目か
tokenN	整数	文内で何文節目か
length	整数	文節の文字数
freq	実数	文節の頻度 (構成するサブワードの頻度の幾何平均)
is_first	ブーリアン	行内最左要素
is_last	ブーリアン	行内最右要素
is_second_last	ブーリアン	行内右から 2 番目の要素
BOS	ブーリアン	文頭要素
EOS	ブーリアン	文末の要素
pre_EOS	ブーリアン	文末から 2 番目の要素
subj	カテゴリカル	被験者番号
WLSPLUWA	カテゴリカル	倒語範疇
anti_locality	整数	先行分脈における掛かり元要素数

A　素性

本実験で用いた素性を表 2 に示す．なおヒトに対する読み時間測定の際文末に改行を加えているため，例えば is_first は「文頭である」か「文が長く画面上で改行が生じた際の改行後の文節」で発火する素性である．(is_first, is_last, is_second_last) について，画面上での改行を考慮しない素性が (BOS, EOS, pre_EOS) であり，ヒトの読み時間のモデリングの際には前者を，言語モデルのサプライザルのモデリングの際には後者を用いた．(lineN, segmentN) と (sentN, tokenN) についても，画面上での改行を考慮するかしないかの違いである．画面呈示順など読み時間測定に関わる素性は，既存研究でより詳細に記述されている [18].

B　指標

B.1　ヒトらしさ

4 節で報告した ΔLogLik について詳細に記述する．4 節中式 1 でモデリングした際の読み時間データに対する対数尤度を $\mathrm{LogLik}_{\mathrm{SP}}$ とする．また，以下の式で読み時間をモデリングした際の読み時間データに対する対数尤度を $\mathrm{LogLik}_{\mathrm{BASE}}$ とする：

$$\begin{aligned}\log(\mathrm{RT}) \sim\ & \mathtt{freq} + \mathtt{length} + \mathtt{prev\_freq} + \mathtt{prev\_length} + \mathtt{is\_first} \\ & + \mathtt{is\_last} + \mathtt{is\_second\_last} + \mathtt{screenN} + \mathtt{lineN} \\ & + \mathtt{segmentN} + (1|\mathtt{article}) + (1|\mathtt{subj})\ .\end{aligned} \tag{2}$$

ΔLogLik は以下のように計算される：

$$\Delta\mathrm{LogLik} = \frac{\mathrm{LogLik}_{\mathrm{SP}} - \mathrm{LogLik}_{\mathrm{BASE}}}{\text{データ全体の文節数}}\ . \tag{3}$$

本値をそれぞれの言語モデルから得られたサプライザルを用いて求め，各言語モデルのヒトらしさとした.

B.2　統語情報に対する敏感さ

以下の式でサプライザルをモデリングした際のサプライザルに対する対数尤度を $\mathrm{LogLik}^{\mathrm{surprisal}}_{\mathrm{BASE}}$ とおき，これらの説明変数に加えて統語情報を追加したときの対数尤度変化を報告した：

$$\begin{aligned}\mathtt{surprisal} \sim\ & \mathtt{freq} + \mathtt{length} + \mathtt{prev\_freq} + \mathtt{prev\_length} + \mathtt{BOS} \\ & + \mathtt{EOS} + \mathtt{pre\_EOS} + \mathtt{sentN} + \mathtt{tokenN} + (1|\mathtt{article})\ .\end{aligned} \tag{4}$$

図 6　ヒトの読み時間，最もヒトらしい言語モデルのサプライザル，PPL の低い言語モデルのサプライザルにおける，anti-locality 効果の違い.

B.2.1　倒語範疇

以下の式でサプライザルをモデリングした際のサプライザルに対する対数尤度を $\mathrm{LogLik}^{\mathrm{surprisal}}_{\mathrm{CATEGORY}}$ とおく：

$$\begin{aligned}\mathtt{surprisal} \sim\ & \mathtt{WLSPLUWA} + \mathtt{freq} + \mathtt{length} + \mathtt{prev\_freq} \\ & + \mathtt{prev\_length} + \mathtt{BOS} + \mathtt{EOS} + \mathtt{pre\_EOS} + \mathtt{sentN} \\ & + \mathtt{tokenN} + (1|\mathtt{article})\ .\end{aligned} \tag{5}$$

倒語範疇に対する敏感さを以下のように求める：

$$\frac{\mathrm{LogLik}^{\mathrm{surprisal}}_{\mathrm{CATEGORY}} - \mathrm{LogLik}^{\mathrm{surprisal}}_{\mathrm{BASE}}}{\text{データ全体の文節数}}\ . \tag{6}$$

B.2.2　統語構造（anti-locality 効果）

以下の式でサプライザルをモデリングした際のサプライザルに対する対数尤度を $\mathrm{LogLik}^{\mathrm{surprisal}}_{\mathrm{STRUCTURE}}$ とおく：

$$\begin{aligned}\mathtt{surprisal} \sim\ & \mathtt{anti\_locality} + \mathtt{freq} + \mathtt{length} + \mathtt{prev\_freq} \\ & + \mathtt{prev\_length} + \mathtt{BOS} + \mathtt{EOS} + \mathtt{pre\_EOS} + \mathtt{sentN} \\ & + \mathtt{tokenN} + (1|\mathtt{article})\ .\end{aligned} \tag{7}$$

倒語構造に対する敏感さを以下のように求める.

$$\frac{\mathrm{LogLik}^{\mathrm{surprisal}}_{\mathrm{STRUCTURE}} - \mathrm{LogLik}^{\mathrm{surprisal}}_{\mathrm{BASE}}}{\text{データ全体の文節数}} \tag{8}$$

C　英語における anti-locality 効果

英語言語モデルについても，5 節の分析と同様に anti-locality 効果の観点から PPL の低い言語モデルがヒトから逸脱していくかを調査する．Dundee コーパス [25, 26] を用い first pass time をヒトの読み時間としている．既存研究が公開しているサプライザルデータ[7] を用い，PPL の低い言語モデルとして GPT-2 の結果を，また PPL の低下による乖離を調査したいため，比較として比較的 PPL の高い 5-gram 言語モデルの結果を用いた．英語で逸脱が見られなかった場合は，英語と日本語の何らかの差異（例えば主辞先行言語であるか主辞後置言語であるか）において特有の乖離であることが示唆される．英語では，ある単語について先行文脈に存在する直接係り先および係り元の数を anti-locality 値とし，この値と読み時間，サプライザルの関係を調べた．図 6 に結果を示す．まず，anti-locality 値が大きくなるほど読み時間やサプライザルが小さくなるといった単純な傾向は見られない．これは，anti-locality 効果が主辞後置言語で観察されるという既存の知見と一致する．図 6 右（PPL の低い言語モデル）とヒトを比較しても顕著な乖離は観察されない．従って，少なくとも英語と日本語の比較という観点では，5 節で観察された乖離は日本語特有のものであった.

7)　https://github.com/wilcoxeg/neural-networks-read-times