ニューラル言語モデルの過剰な作業記憶

栗林樹生 [1,2] 大関洋平 [3,4] Ana Brassard[4,1] 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
{kuribayashi, inui}@tohoku.ac.jp oseki@g.ecc.u-tokyo.ac.jp
ana.brassard@riken.jp

概要

近年，ニューラル言語モデルの分析が盛んに行われ，ヒトの言語処理について探求する計算心理言語学の両視点を取り入れた分析も増えた．例えば言語モデルとヒトの文処理の振る舞いを直接比較する研究が行われている．本研究では，どのような計算モデルがヒトの振る舞いを忠実に再現するかという，計算心理言語学的な関心で研究を行う場合，工学的な文脈で典型的に活用される LSTM や Transformer 言語モデル [1, 2] をそのまま用いることが，必ずしも道具立てとして妥当な選択ではないことを指摘する．具体的には，ニューラル言語モデルに入力する文脈を限ることでよりヒトらしい振る舞いが促されることから，これらのモデルはヒトと比べて過剰な作業記憶 (同時に処理できる文脈情報) 容量を有していることを提示する．

1 はじめに

ニューラル言語モデルには，自然言語処理分野からの工学的な関心と，計算心理言語学からの科学的な関心が寄せられている．近年では両分野の視点を融合させた研究も行われ，自然言語処理分野において工学的成功を収めたモデルに対して，心理言語学的な視点から，例えばヒトと同様の振る舞いをするか調査するといった試みが盛んである [3, 4, 5, 6, 7]．

読み活動のモデリング: 本研究では，ヒトの言語処理活動を忠実に再現するモデルを探求することで構成論的にヒトの言語処理を理解するという認知科学・計算心理言語学の立場をとる．人工知能・認知科学分野の目標の一つは，ヒトの振る舞いを説明するモデルを作り，ヒトの理解を深めること (計算心理学) であり [8]，計算心理言語学では特にヒトの逐次的な読み活動 (読み時間など) をモデルの対象とする [9]．モデル・理解にも様々な段階が存在するが，ここでは，物理学が物体の複雑な運動を簡潔な数式

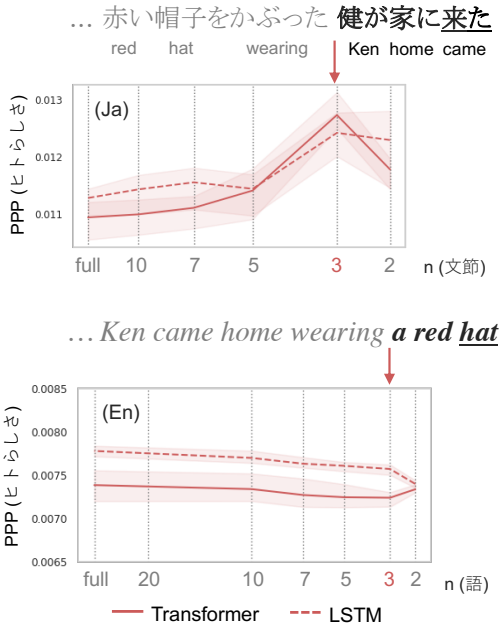

図 1 サプライザルを条件付ける文脈長 (横軸) を制御した場合の，ヒトの読み時間との傾向の近さ (縦軸) の変化．例えば n=3 は，3-gram サプライザルの結果に対応する．縦軸の値 (psychometric predictive power; PPP) が大きいほど，そのサプライザルがヒトの逐次的読み負荷をうまく説明することを表す．

で説明しようとするように，ヒトが文を一単語ずつ読む際に示す逐次的な読み負荷を簡潔に説明する理想化されたモデルに関心がある．

自然言語処理技術の発展により，これまで計算が困難であった量を近似的に求め，読み負荷のモデルとして検証することが可能になった．例えば $-\log p(\text{単語} \mid \text{長い文脈})$ はその最たる例であり，長い系列の頻度を数えて値を求めることは現実的でないが，ニューラル言語モデルを訓練することで，(少なくとも次の単語の出現を正確に予測するという観点で) 正確な値を算出できるようになった．このような発展も後押しし，言語モデルや統語解析器が推定するサプライザル ($-\log p(\text{単語} \mid \text{文脈})$) でヒトの読み負荷を説明しようとする，**予測に基づく説明**が

注目を集めた [10, 4]．ヒトは次の単語の出現を予測しながら文を読んでおり，予測を裏切られると処理負荷が上がるという説明である．また次なる関心として，どのようなモデル・アルゴリズムで計算されるサプライザルがヒトの文処理をうまく説明するかが焦点となっている．

作業記憶の乖離: このようなヒトの文処理の機序を説明する視点からは，近年の強力な言語モデルを用いて認知モデリングを行うことが必ずしも妥当な選択ではないことを指摘する．具体的には，工学的に高い性能を示す LSTM や Transformer がヒトと比べて過剰な**作業記憶** (同時に処理できる文脈情報の量) をもつことを実験的に示し，むしろ忘却が促される単純な再帰型ニューラルネットワーク [11] や n-gram モデルなどの方が，ヒトの読み活動の計算モデルとしては妥当である可能性を提示する．

近年のニューラル言語モデル (例えば Transformer [2] に基づくモデル) は数百から数千単語の文脈に同時にアクセスしながら次の単語を予測する設計であり，帰納バイアスとして非常に強力な作業記憶容量 (同時に処理できる文脈の量) を仮定している [3]．それに対して，一般にヒトの作業記憶は非常に制限されており，高々 4 つ程度の項目を同時に処理することが限度などと示唆されている [12]．例えば 10 桁程度の電話番号を一度で覚えることができるヒトはそういないだろう．ヒトの文処理もまた作業記憶の制約を受けていると考えられており，例えば長距離依存により読み負荷が増加することや [13]，関連する要素がなるべく近くに配置されるという言語の一般的な性質を説明する [14]．

実験と示唆: ヒトらしい作業記憶の模倣としてニューラル言語モデルに入力する文脈に制限を与えることで，よりヒトの文処理活動データ (視線停留時間) と近い，ヒトらしい振る舞いが促されるか，文処理活動データの説明力を損なわなかった (図 1)．このことから，近年の言語モデルがヒトよりも強力な作業記憶を持つということを主張する．

本研究では作業記憶の局所性 (遠い情報ほどアクセスが難しい) を仮定し，直近 $n-1$ 語の文脈から計算されるサプライザル (n-gram サプライザル) とヒトの読み活動の近さを分析する．日本語と英語の言語横断的な調査から，高々 **3-gram** 程度のモデルで計算されるサプライザルによりヒトの読み活動がうまく説明できる (日本語)，または文脈を限っても説明能力を損なわない (英語) ことが示された．

また，コーパス平均的に妥当な文脈長が定まらなかった英語の結果 (図 1 下) について分析すると，特定の語では長い/短い文脈を考慮したサプライザルが読み時間をうまく説明するといった，統語依存な傾向が観察された．特に動詞の読み時間について，主語が離れているほど，長い文脈で読み時間を説明できた．この傾向は日本語では観察されず，英語における主語と動詞の数の一致や，日本語における (主語でなく) 主題の卓越性といった要因が，結果の言語依存性 (図 1) と関連している可能性がある．

2 記憶に基づく説明

文処理の計算論的モデルについて，予測に基づく説明とは伝統的に対立する形で，**記憶に基づく説明**が提唱されてきた．記憶に基づく説明では，離れた文脈情報へのアクセスには負荷がかかるなどの仮説をおいている [13, 15]．言語や現象に応じて予測と記憶の説明の妥当性は異なっていた [16, 17]．本研究は，読み負荷のモデリングにおける予測と記憶の両側面の融合を試みた研究とも解釈できる．本研究で扱う不完全な文脈に基づくサプライザルは両者を融合した指標として近年提唱されてきたものの [18]，データに基づく検証は限られており，本研究はその妥当性について経験的な証拠を提供した．

また，n-gram モデルは予測と記憶の両側面を反映した典型的なモデルであるものの，特に近年はニューラル言語モデルの分析におけるベースラインとして，頻度ベースの n-gram 言語モデルが登場することが多く [10, 4, 6]，両者はニューラルネットに基づくかなど複数の軸で仮定が異なるため，サプライザルの条件として適切な文脈について分野内で厳密に調査されていなかった．本研究ではニューラル言語モデルの入力を制限し，文脈長の観点でのみ切り分けて分析を行う．

3 手法・実験設定

ヒトの逐次的な文処理負荷 (視線停留の長さ) と言語モデルが計算する指標の傾向の近さを測定する．作業記憶の局所性を踏まえ [13]，本研究では，直前 $n-1$ 語を文脈として用いる n-gram サプライザルの認知的妥当性について分析し，n を動かした際の影響を観察する．

不完全な文脈に基づくサプライザル: Left-to-right 言語モデル θ によって計算される不完全な文脈に基づくサプライザルを用いる．言語モデルはサ

ブワードを入力単位としているが，読み時間は単語 (英語) や文節 (日本語) といったより大きな単位に付与されている．読み時間データに従い，本研究では「語」は英語の単語または日本語の文節を指す．

文中の i 番目のサブワードを w_i，i 番目の語を s_i とする．各語は 1 つ以上の連続するサブワードで構成される．ある語 $s_i = [w_k, \cdots, w_l]$ の文脈 $c_{<i} = [s_0, \cdots, s_{i-1}]$ における n-gram サプライザル $I_{\text{n-gram}}$ を以下のように計算する．

$$I_{\text{n-gram}}(s_i, c_{<i}, n) = -\log p_\theta(s_i | s_{i-n+1}, \cdots, s_{i-1}) \ ,$$
$$= \sum_{j=k}^{l} -\log p_\theta(w_j | w_m, \cdots, w_{j-1}) \ .$$

ここで，w_m は，s_{i-n+1} の開始サブワードを指す．例えば日本語における 3-gram サプライザルでは，直前 2 文節を構成するサブワード系列を文脈として与えている．単語や文節のサプライザルは，確率の連鎖律より，それらを構成するサブワードのサプライザルの和として計算する．なお，通常のテキストで学習した言語モデルを用い，推論時に文脈を制限することで $I_{\text{n-gram}}$ を計算しており，学習と推論の乖離については付録 A で議論する．

評価指標: 尺度 $I_{\text{n-gram}}$ の読み時間に対する説明力を一般化線形混合モデルで測定する．具体的には，$I_{\text{n-gram}}$ とベースライン素性を説明変数とし，各語の視線停留時間を被説明変数とする回帰モデルを訓練する (付録 B)．同様に，$I_{\text{n-gram}}$ を含まないベースライン素性のみの回帰モデルも訓練し，両回帰モデルが示す対数尤度 (当てはまりの良さ) を求める．両対数尤度の差のデータポイント数平均を「ヒトらしさ」の尺度 (psychometric predictive power; PPP) とし，PPP が大きいほど $I_{\text{n-gram}}$ が読み時間をうまく説明するとみなす．この PPP を n-gram サプライザルを計算する際の文脈長を変えて比較する．具体的には，{2,3,5,7,10,20}-gram サプライザルと，文内の前方文脈すべてを用いる “full” 設定の計 7 設定を比較する．なお，本研究の主たる関心は文の統語的処理であり，文を超えた文脈は考慮しない．

読み時間データ: Dundee corpus (英語) と BCCWJ-EyeTrack (日本語) を用いて英日横断的な検証を行う [19, 20]．はずれ値などを除去し (付録 C)，Dundee corpus では 217,876 データポイント，BCCWJ-EyeTrack では 9,217 データポイントを用いる．読み時間として，first pass duration を対象とする．

言語モデル: Transformer/LSTM 言語モデルを利用した [2, 1]．詳細は付録 D に記載する．異なるシードで学習した 3 つのモデルの結果の平均と標準偏差を報告する．

4 実験: ヒトらしい文脈長

はじめに，コーパス全体の傾向として，読み時間を平均何語の直前の文脈情報によってうまく説明できるか調査する (4 節)．その後，長い文脈で条件付けることの効能を調査する (5 節)．結果から，記憶のバイアスと統語のバイアスを確認した．

4.1 結果

結果を図 1 に示す．日本語ではある程度短い文脈に条件づいたサプライザル (3-gram サプライザル) が読み時間をうまく説明でき，英語では文脈を限っても認知的妥当性が大きく変化しないことがわかった．英語での妥当性を保ちつつ，日本語での妥当性を向上させたという点で，限られた文脈に基づくサプライザルが言語横断的に妥当な尺度であることが示唆された．言い方を変えると，文内のすべての文脈をニューラル言語モデルに与えサプライザルを計算する方法では，ヒトの読み活動を説明する上で過剰な文脈を考慮していると言える．

日本語: 日本語では文節 3-gram がもっとも良い PPP を達成し，文脈を過度に用いて計算されたサプライザルはヒトの読み活動から乖離することが示された．このことは，ヒトの逐次的な統語処理の負荷が局所的な情報で説明できるというひとつの可能性を示唆する．また非常に荒い紐付けではあるが，作業記憶の限度としてしばしば挙げられる 4 ± 1 項目という数と 3 文節という数は近しい [12]．

英語: 文脈長を変化させても PPP は大きく変化しなかった (図 1 下)．一つの解釈としては，データポイントに応じてサプライザルを条件付けるべき適切な文脈長が異なる可能性があり，コーパス平均的な傾向で見ると，文脈長を変えても PPP に変化がないように見えている可能性がある．特に，作業記憶の制約に反し，長い文脈で条件付けることが適切であると観察されるケースがどのようなものであるか，次節でより詳しく分析していく．

5 分析: いつ長い文脈が有効か？

英語における遠い主語の存在が，対応する動詞の読み時間の説明に必要な文脈長を引き伸ばしている

ことが示唆された．

設定: 4 節では，様々な条件でサプライザルを計算し，視線停留時間を説明する回帰モデルの当てはまりの良さを通して分析を行った．本節では，長い/短い文脈条件で付けたサプライザルによる視線停留時間回帰モデル ($M_{\text{short}}, M_{\text{long}}$) について，読み時間への当てはまりが悪い箇所を調査し比較する．具体的には，長い文脈で条件づけたサプライザルを用いた回帰モデル M_{short} と，短い文脈で条件づけたサプライザルを用いた回帰モデル M_{long} が推定するデータポイント (語) ごとの残差 (小さいほど回帰式が視線停留時間をうまく説明できている) を観察する．あるデータポイント集合 $\mathcal{D}$ について，長い文脈を考慮した M_{long} による二乗残差平均 $r_{\text{long}}(\mathcal{D})$ が M_{short} の示す二乗残差平均 $r_{\text{short}}(\mathcal{D})$ と比べてどれほど小さくなるかを，長い文脈を考慮する有効性とみなす．具体的には，$r_{\text{short}}(\mathcal{D}) - r_{\text{long}}(\mathcal{D})$ を $\mathcal{D}$ における長い文脈の貢献度 (C) とし，$r_{\text{short}}(\mathcal{D})$ が大きく (当てはまりが悪く)，$r_{\text{long}}(\mathcal{D})$ が小さくなる (当てはまりが良い) ほど C は大きくなる．M_{short}，M_{long} として，2-gram と full の設定で当てはめた回帰モデルを用いた．また，文脈を考慮することの効果に焦点を当てるため，文の後方に出現する語に分析対象を限った．[1)]

統語依存距離と記憶: 統語的に関連する前方要素に作業記憶が影響を受けている可能性をあげる [13]．例えば，“Ken **goes** to the school.” と “Ken, who is my brother, **goes** to the school.” という文において，“**goes**” の処理の際に主語 “Ken” が意識されるのであれば，後者の文における “**goes**” の読み時間のモデリングに長い文脈が必要になると示唆される．

本分析では，主語と動詞の依存距離について分析を行う．[2)]動詞に対応するデータポイント集合を，対応する主語との依存距離 (何語離れているか) に基づいて分割し，各データポイント集合について長い文脈を考慮することの残差の減少量 C を観察する．図 2 左より，主語が離れた動詞ほど読み時間のモデリングにおいて，長い文脈を用いることの有効性が確認された．このことから，読み時間が主語と動詞の関係に強く影響を受けていることが示唆された．

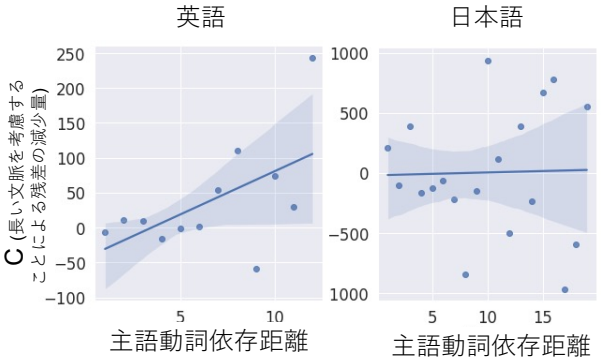

図 2 動詞の読み時間のモデリングにおける，長い文脈の貢献度 C と主語との距離の関係．

同様の分析を日本語データで行ったところ，日本語では，主語と動詞の距離と，サプライザルを条件付ける文脈長との間に影響が観察されなかった (図 2 右)．このことから，図 1 における言語依存な結果は，主語動詞間の長距離依存に対するヒトの振る舞いの違いと関連していることが示唆された．英語では，主語と動詞に数の一致が課されていること，日本語が主題卓越言語であるため主語の存在が読み手にとって意識的なものではないことなどがこの差異を生む要因として考えられる．本研究では，記憶 (局所性) と統語のバイアスを外部的に文脈を編集して取り入れたが，どのようなモデル設計，学習を行うことでこのようなヒトらしい記憶の制約を獲得させることができるかは興味深い．

6 おわりに

本研究では，言語モデルに入力する文脈長を制限した実験から，近年自然言語処理分野で用いられるニューラル言語モデルが，ヒトに比べて過剰な作業記憶を有している可能性を提示した．特に近年言語モデルが注目を集め，ヒトの言語処理との比較にも分野の関心が向いている．もしヒトの逐次的文処理を純粋に再現し，ヒトの計算論的なモデルを探求することが研究の目的であれば，LSTM や Transformer といったモデルよりも，むしろ単純な再帰型ニューラルネットワーク [11] や n-gram のようなモデルを用いることが妥当かもしれない．また，本実験は不完全な文脈に基づくサプライザルの妥当性を経験的に支持し，読み活動データにおける記憶 (局所性) と統語的バイアスの相互的な作用を示唆した．

1) 英語/日本語の文内での語のインデックスの中央値が 12,8 であったことを踏まえ，英語では文内の 13 単語目以降，日本語では 9 文節目以降を用いた．

2) なお，主語と動詞の依存に限らない全体の傾向としては，統語距離と長い文脈を考慮することの影響には強い関係が見られなかった．

謝辞

本研究は JSPS 科研費 JP20J22697，JST さきがけ JPMJPR21C2 の支援を受けたものである．

参考文献

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

[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 **Proceedings of NIPS**, pp. 5998–6008, 2017.

[3] Danny Merkx and Stefan L. Frank. Human Sentence Processing: Recurrence or Attention? In **Procceding of CMCL 2021**, 2020.

[4] 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.

[5] Ethan Wilcox, Pranali Vani, and Roger Levy. A targeted assessment of incremental processing in neural language models and humans. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 939–952, Online, August 2021. Association for Computational Linguistics.

[6] Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Ryo Yoshida, Masayuki Asahara, and Kentaro Inui. Lower perplexity is not always human-like. In **Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 5203–5217, Online, August 2021. Association for Computational Linguistics.

[7] Ryo Yoshida, Hiroshi Noji, and Yohei Oseki. Modeling human sentence processing with Left-Corner recurrent neural network grammars. In **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 2964–2973, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[8] Stuart C Shapiro. Artificial intelligence (AI). In **Encyclopedia of Computer Science**, pp. 89–93. John Wiley and Sons Ltd., GBR, January 2003.

[9] M W Crocker. Computational psycholinguistics. **Computational Linguistics and Natural Language**, 2010.

[10] 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.

[11] Jeffrey L Elman. Distributed representations, simple recurrent networks, and grammatical structure. **Mach. Learn.**, Vol. 7, No. 2, pp. 195–225, September 1991.

[12] N Cowan. The magical number 4 in short-term memory: a reconsideration of mental storage capacity. **Behav. Brain Sci.**, Vol. 24, No. 1, pp. 87–114; discussion 114–85, February 2001.

[13] Edward Gibson. The dependency locality theory: A distance-based theory of lingustic complexity. **Image, language, brain**, Vol. 2000, pp. 95–126, 2000.

[14] Edward Gibson, Richard Futrell, Steven P Piantadosi, Isabelle Dautriche, Kyle Mahowald, Leon Bergen, and Roger Levy. How efficiency shapes human language. **Trends Cogn. Sci.**, Vol. 23, No. 5, pp. 389–407, May 2019.

[15] Richard L Lewis and Shravan Vasishth. An activation-based model of sentence processing as skilled memory retrieval. **Cogn. Sci.**, Vol. 29, No. 3, pp. 375–419, May 2005.

[16] Lars Konieczny. Locality and Parsing Complexity. **Journal of Psycholinguistic Research**, Vol. 29, No. 6, pp. 627–645, 2000.

[17] Samar Husain, Shravan Vasishth, and Narayanan Srinivasan. Strong expectations cancel locality effects: evidence from hindi. **PLoS One**, Vol. 9, No. 7, p. e100986, July 2014.

[18] Richard Futrell, Edward Gibson, and Roger P. Levy. Lossy-Context Surprisal: An Information-Theoretic Model of Memory Effects in Sentence Processing. **Journal of Cognitive Science**, 2020.

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

[20] 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.

[21] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In **Proceedings of EMNLP**, pp. 66–71, 2018.

A　N-gram サプライザルの正確性

表 1　ニューラル/数え上げ n-gram 言語モデルのパープレキシティ．サブワードレベルの言語モデルで比較している．

言語	N	ニューラル	数え上げ
	2	**388**	414
Ja	3	**264**	313
	5	**200**	295
	2	**106**	150
En	3	**75.6**	82.6
	5	**60.3**	70.2

本研究では，ニューラル言語モデルはコーパスを適当なブロックで区切ったチャンクをミニバッチとして訓練している．各ミニバッチの先頭周辺では，n-gram 言語モデルのような学習が行われるが，全体としてみるとそのような訓練事例の数は限られているため，今回のようにニューラル言語モデルで n-gram サプライザルを求めた場合，学習と推論の乖離が想定される．しかしながら，サブワードレベルの n-gram サプライザルによって計算されるコーパスのパープレキシティを本実験の設定 (ニューラル) と数え上げに基づく n-gram 言語モデルで比較すると，本設定の n-gram サプライザルの方が低いことが分かった．したがって，本研究ではニューラル言語モデルが，先頭数語のみを入力した際にも適切な振る舞いをしていると一旦結論づけた．パープレキシティは読み時間が付与されたコーパスで測定した．

B　回帰モデル

読み時間（RT）のモデリングは以下の式で行った：

$$\begin{aligned}\mathtt{RT} \sim\ & \mathtt{surprisal} + \mathtt{surprisal\_prev\_1} \\ & + \mathtt{surprisal\_prev\_2} + \mathtt{freq} * \mathtt{length} \\ & + \mathtt{freq\_prev\_1} * \mathtt{length\_prev\_1} \\ & + \mathtt{screenN} + \mathtt{lineN} \\ & + \mathtt{segmentN} + (1|\mathtt{article}) + (1|\mathtt{subj})\ .\end{aligned}$$

各素性の概要は表 2 に示す．prev_1 などは，一単語まえの該当する素性を示す．前の語のサプライザル (surprisal_prev_1 と surprisal_prev_2) は，周辺の語の処理負荷が読み時間に影響を与えるスピルオーバー効果を考慮して含めた．ただし，日本語においては有意な効果が観察されなかったため除いた．

C　データの前処理

英語では以下のいずれかの条件に合致するデータポイントを，日本語では (a), (c), (e) のいずれかに合致するデータポイントを除いた．

(a)　読み時間が 0 秒であるか，3 標準偏差から外れている
(b)　句読点が含まれる
(c)　数量を表す文字・語が含まれる
(d)　句読点・数量の次の語
(e)　読み時間測定時に画面上で各行の最初に提示されている

表 2　回帰モデルで用いた素性

変数名	型	記述
surprisal	実数	サプライザル
RT	実数	読み時間 (ms)
freq	実数	語の頻度 (構成するサブワードの頻度の幾何平均)
length	整数	語の文字数
screenN	整数	画面呈示順
lineN	整数	何行目か
segmentN	整数	画面右から何文節目か
sentN	整数	何文目か
tokenN	整数	文内で何語目か
subj	カテゴリカル	被験者番号
article	カテゴリカル	記事番号

(f)　読み時間測定時に画面上で各行の最後に提示されている

(b) と (f) に関しては，日本語に適用すると「考える。」のような文末の文節がすべて除かれるため，適用しなかった．

D　言語モデルハイパーパラメータ

表 3 にニューラル言語モデルのハイパーパラメータを示す．およそ 500 万文の日本語テキスト (Wikipedia と新聞) と，400 万文の英語テキスト (Wikipedia) を学習データとして用い，10 万回アップデートした言語モデルを用いた．日本語テキストは，読み時間データの分割との一貫性を保つため，一度国語研短単位に分割した後，BPE でサブワードに分割した[3)]．

表 3　ニューラル言語モデルのハイパーパラメータ．Transformer (上) と LSTM (下)．

Fairseq model	architecture	transformer_lm_gpt
	adaptive softmax cut off	50,000, 140,000
	share-decoder-input-output-embed	True
	embed_dim	384
	ffn_embed_dim	2,048
	layers	8
	heads	6
	dropout	0.1
	attention_dropout	0.1
Optimizer	algorithm	AdamW
	learning rates	5e-4
	betas	(0.9, 0.98)
	weight decay	0.01
	clip norm	0.0
Learning rate scheduler	type	inverse_sqrt
	warmup updates	4,000
	warmup init lrarning rate	1e-7
Training	batch size	61,440 tokens
	sample-break-mode	none

Fairseq model	architecture	lstm_lm
	adaptive softmax cut off	50,000, 140,000
	share-decoder-input-output-embed	True
	embed_dim	400
	hiden_size	1,024
	layers	2
	dropout	0.1
Optimizer	algorithm	AdamW
	learning rates	1e-3
	betas	(0.9, 0.98)
	weight decay	0.01
	clip norm	0.0
Learning rate scheduler	type	inverse_sqrt
	warmup updates	4,000
	warmup init lrarning rate	1e-7
Training	batch size	20,480 tokens
	sample-break-mode	none

3)　SentencePiece [21] を用い，文字の網羅率を 0.9995，語彙数を 100,000 とした．