# 工学的性能と人間らしさの関係はトークン分割に依存する

三輪敬太　吉田遼　大関洋平
東京大学
{miwakeita, yoshiryo0617, oseki}@g.ecc.u-tokyo.ac.jp

## 概要

言語モデルがトークンに与える情報量と人間の文の読み時間との関連が近年の研究で示されており，この観点から言語モデルの「人間らしさ」を評価できる．本研究では情報量や言語モデルの挙動が文のトークン分割単位に深く依存することに着目し，分割単位の異なる言語モデルにおいて工学的性能（パープレキシティ）に対する人間らしさの関係の違いを実験的に調査した．日本語の短単位と長単位が大きく異なる挙動を示すことを報告し，このような差異の原因を議論する．

## 1　はじめに

サプライザル理論 [1] では，人間の文の読み時間の線型な予測変数として情報量（サプライザル）の有効性が示されている．トークンの情報量は文脈 $\mathbf{c}$ におけるトークン $t$ の確率を用いて $I = -\log_2 P(t \mid \mathbf{c})$ と定義される．サプライザル理論の妥当性は多くの言語での実験で示されている [2, 3, 4]．情報量は言語モデルを用いて計算できるが，言語モデルは様々な要素に依存して異なる挙動を示す．そこで，サプライザル理論の観点から，言語モデルの「人間らしさ」をモデルの情報量が人間の読み時間をどれだけ反映するかによって評価できる．こうした研究では言語モデルの工学的性能（パープレキシティ，PPL）と人間らしさの関係が注目されており，初期の英語の研究では PPL が小さいほど人間らしさも向上すると主張された [5, 6]．しかし，[7] は日本語でこの関係の破綻を報告し，大規模言語モデルを用いた研究では英語でも対応の破綻が報告されている [4, 8]．

情報量や言語モデルの挙動は言語のトークン分割単位に深く依存する．例えば「言語学」トークンを持つ分割単位では持たないものに比べて「言語」トークンの頻度が下がり，情報量が大きくなる．自己回帰モデルでは分割単位が小さいと予測すべき情報が少ないため，大きい分割単位と比べて簡単なタスクで訓練していることになる．さらに，例えば文字単位の言語モデルでは次単語の先頭の文字ではなく単語内部の次の文字を予測するような点が多く，次単語予測の学習の重みが相対的に下がる．

サプライザル理論でも分割単位について議論がなされており，文字単位の言語モデルの妥当性がしばしば主張されている [9, 10] ほか，人間の文処理モデルとしての妥当性の観点から，サブワードの利用の是非にも関心が持たれている [11]．一方で工学的性能と人間らしさの観点では，トークン分割の影響がまだ議論されていない．特に分割単位の非自明な日本語ではその影響が重要である．

本研究では，日本語においてトークン分割単位の異なるニューラル言語モデルを構築し，その工学的性能と人間らしさの関係を調べ，分割単位の影響を検証した．長単位・短単位に顕著な違いが見られ，長単位の言語モデルでは工学的性能と人間らしさの関係の破綻が抑制されていることを報告する．

## 2　予備実験

予備実験では国立国語研究所の長単位（Long unit word, LUW），短単位 (Short unit word, SUW) と 1 文字ずつに分割する文字単位 (Character, CHR) で言語モデルを構築し，モデルの学習過程での工学的性能と人間らしさの関係の傾向を調べる．語彙規模が大きい長単位や短単位を用いるため未知語率や語彙数の統制が難しい．予備実験で大域的な傾向を確認し，本実験でこれを厳密に検証する．

### 2.1　方法

**人間らしさ**　読み時間推定精度への言語モデルで計算する情報量の貢献で評価する [5, 6, 7]．[7] で利用された線形混合効果モデルを用いて読み時間の推定を実施し，情報量を考慮する推定モデルとしないモデル（ベースラインモデル）の間の尤度比の対数 ΔLogLik がより大きいと，情報量が人間の読み時間を相対的により反映すると見做せる．ベースライン

モデルに (1) を，比較対象のモデルとして (1) に推定したい文節を含む前 3 文節分の情報量を独立にプレディクタに加えたものを構築し，この 2 モデル間の ΔLogLik をデータ数で割った値を人間らしさの指標とする．プレディクタの詳細は付録 B に示す．

$$\begin{aligned}\text{time} \sim\ & \text{is\_first} + \text{is\_last} + \text{length} * \text{count\_ave\_kika} \\ & + \text{length\_prev} * \text{count\_ave\_kika\_prev} + \text{space} \\ & + \text{articleN} + \text{sessionN} + \text{screenN} + \text{lineN} \\ & + \text{segmentN} + (1 \mid \text{article}) + (1 \mid \text{subj})\end{aligned} \tag{1}$$

time には [7] と同様，BCCWJ-EyeTrack[12] の First pass time (FPT) を利用した．ここでは 24 名の被験者の読み時間データが文節[1] ごとに付与されている．いずれかの分割単位で未知語トークンを含む文節，読み時間が 0 の文節，本文以外に含まれる文節を除外した．[13, 14] に従い，まずベースラインモデルをフィッティングし，推定誤差が $3\sigma$ を上回る点を除外し両モデルを改めてフィッティングする．

**工学的性能**　推論データの PPL で評価する．分割単位の異なるモデルで比較するためトークン単位ではなく文節単位 PPL[2] を用いる．

**言語モデル**　GPT-2 (Generative Pre-trained Transformer 2）[15] を用いる．近年の研究では事前学習モデルを用いることが一般的だが，分割単位による差異を調べるため，全ての分割単位でモデルを初期状態から学習した．学習や言語モデルの詳細な設定は付録 C に示す．

学習データとして，現代日本語書き言葉均衡コーパス（BCCWJ）[16] を利用した．検証データとして新聞データ 7,030 文（約 10%）と，推論データを学習データから除いた．推論で文脈とする最大 3 文でモデルのコンテキストサイズ（512 トークン）を超えないよう，170 文字以上の文を含む記事を除去した．最終的に 20,931 記事，3,074,317 文を利用した．全体で 3 エポック学習する．

**分割単位**　長単位と短単位は [17] でその分割方針が示されている．長単位は文節の認定のもと，文節を「内容語＋機能語」に分割することで設定され，最長で文節 1 つ分となる．短単位は「最小単位」に基づいて設定され，最短で最小単位 1 つ分となる．3 種類の分割の比較を表 1に示した．

1）　文節の情報量 $I$ は文節を構成するそれぞれのトークン $t$ の情報量 $I_t$ の和として，$I = \sum_t I_t = -\sum_t \log_2 P(t \mid \mathbf{c}_t)$ で計算する．$\mathbf{c}_t$ は文脈である．

2）　文節単位の PPL は $\text{PPL}_s = \left(\prod_{S,\mathbf{c}} \frac{1}{P(S|\mathbf{c})}\right)^{\frac{1}{n}}$ として計算する．$S = t_1, \ldots t_m$ は文節，$n$ は文節数である．$P(S \mid \mathbf{c}) = \prod_{i=1}^{m} P(t_i \mid \mathbf{c}_t)$ で計算する．

**表 1**　それぞれの分割単位の比較

| | LUW | SUW | CHR |
|---|---|---|---|
| 延べ語数 | 50.36M | 60.38M | 95.65M |
| 異なり語数 | 1,496,168 | 297,541 | 7,046 |
| 語彙数 | 136,958 | 90,787 | 5,074 |
| 未知語率 | 4.1% | 0.6% | 0.01% |

BCCWJ は長単位と短単位でアノテーションされている．文字単位は長単位の語を分割して得た．学習データ中の頻度が 11 以下の語を未知語として扱った．数字は桁数別のトークンで置換し，未知語のうち英単語とカタカナ語では文字数ごとに異なる未知語トークンを当てた．推論データは新聞データに由来するため，同ドメインの新聞データにおける頻度が 3 以上の語は必ず語彙に含めた．

### 2.2　結果

文節単位 PPL と ΔLogLik の関係が図 1の左の 3 つである．長単位モデルでのみ，PPL が小さくなる（工学的性能が上がる）と ΔLogLik が単調に大きくなる（人間らしくなる）．文字単位モデルでは PPL と ΔLogLik の間に自明な関係が見られず，短単位モデルでは PPL と ΔLogLik の関係が山型を成す．

## 3　実験

予備実験ではトークン分割単位によって工学的性能と人間らしさの関係が大きく異なることが示された．長単位は直線的な対応関係を示すが，短単位と文字単位ではそのような関係が見られない．

しかし，予備実験は定量的特徴の統制が不十分であり，長単位の未知語率は短単位の約 6.8 倍である．近年 [18] が導入した欠損文脈サプライザル理論では，不完全な文脈に対する情報量を用いることでより人間らしい挙動が観察されている [19]．未知語により文脈が欠落した長単位のモデルの特徴的な挙動は，欠損文脈サプライザル理論的な効果の可能性がある．また，長単位の言語モデルは短単位の言語モデルの 1.5 倍の語彙を持っており，同一サイズの空間により多くの語彙を埋め込むことの影響も考えられる．そこで，本実験では長単位と短単位の挙動の違いについてより厳密な検証を行う．

### 3.1　方法

言語モデルのトークン分割単位に Byte Pair Encoding（BPE）によるサブワード分割を利用する．

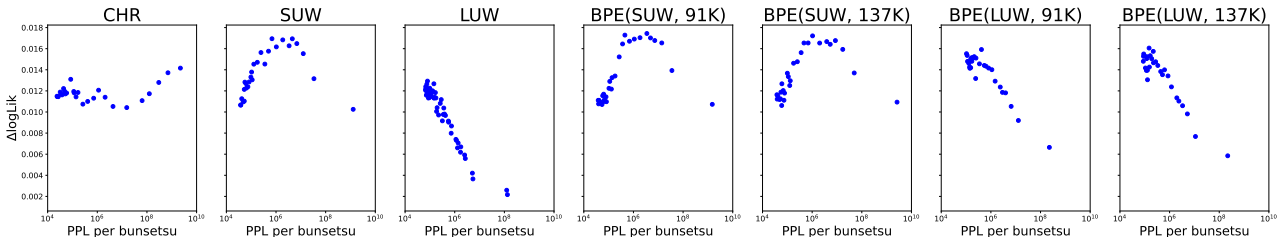


**図 1**　予備実験の 3 種類の言語モデル（CHR,SUW,LUW）と本実験の 4 種類の言語モデル（BPE(SUW/LUW,91K/137K)）での文節単位の PPL の対数（横軸）と推定性能（縦軸）の関係．長単位系のモデルでは直線的な関係が見られる．

SentencePiece[20] を利用し，両単位に類似した BPE 分割モデルとして，事前分割制約（LUW/SUW）と語彙数（137K/91K）の異なる 4 種類のサブワード単位（表 2）を構築する．未知語率は全て 0.1%に統制した．LUW と BPE(LUW, 137K) の挙動の比較から未知語率の影響を，また 4 種類のモデルの比較から語彙数の影響を，それぞれ検証できる．

**表 2**　4 種類のサブワード単位モデル．

| | BPE (SUW,91K) | BPE (LUW,91K) | BPE (SUW,137K) | BPE (LUW,137K) |
|---|---|---|---|---|
| 延べ語数 | 61.40M | 56.18M | 60.97M | 55.10M |
| 事前分割 | **短単位** | **長単位** | **短単位** | **長単位** |
| 語彙数 | **90,787** | | **136,958** | |

予備実験と同様に，工学的性能を推論データの文節単位の PPL によって評価し，人間らしさを ΔLogLik で評価する．学習方法やデータの取り扱いは予備実験と同様のため，新たに構築する言語モデルの評価は予備実験で得た結果と比較できる．

## 3.2　結果

構築した 4 つのサブワード単位モデルの工学的性能と人間らしさの関係は図 1の右の 4 つである．

短単位を事前分割とする 2 つのモデルでは短単位モデルに類似の挙動が，長単位を事前分割とする 2 つのモデルでは長単位モデルに類似の挙動が示された．この結果は長単位による事前分割のみが長単位的関係の再現に必要であり，語彙数の違いによる影響は見られないことを示すものと解釈できる．また，サブワード単位モデルは全て未知語率が 0.1%と低いにもかかわらず，事前分割が長単位の 2 モデルでは直線的な関係が示されており，未知語率の影響は見られないと考えられる．

# 4　議論

予備実験及び本実験では，工学的性能と人間らしさの関係は分割単位に強く依存していることが示された．結果は長単位と長単位による事前分割モデルでのみ [5] が主張した直線的関係が現れることを示した．こうした差異が現れる原因を議論する．

**超人的な挙動**　[8, 21] では PPL が極度に小さい範囲で言語モデルが損失を下げるため表現をそのまま覚え，情報量を過剰に低く見積ることによる関係の破綻が報告された．しかし，超人的な挙動と呼ばれるこの振る舞いは，パラメタ数が 10 億を超える大規模言語モデルによる研究で見られるもので，本研究の短単位モデルには当てはまらない．

**UID 仮説**　今回の結果はむしろ，[7] の議論に沿って理解し得る．彼らは日本語で直線的関係の破綻を報告し，自然言語の情報の分布が一様だと主張するいう情報密度一様性仮説（UID 仮説）[22] の観点からこれを考察した．[7] による検証では，英語では一文中の情報量が一様な傾向があるが，日本語では文の終わりにかけて情報量が小さくなる．これは日本語の文節あたりの情報量が一様ではなく UID 仮説への適合性が弱いことを意味する．言語モデルの学習は次トークンの予測確率を最大化する過程であり，伴って理想的にはトークンあたりの情報量が一様になる．このことから彼らは学習が対象言語が UID 仮説を満たすことを期待して進むとし，トーク

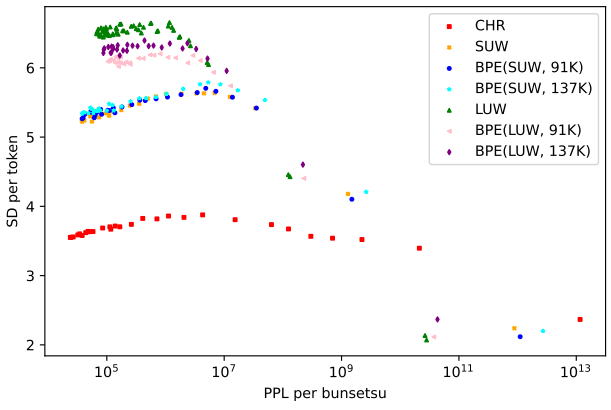


**図 2**　各言語モデルの文節単位 PPL（横軸）と言語モデルのトークン単位の SD（縦軸）の関係．短単位系モデルは山型を示すが，長単位系モデルは大きな値を維持する．

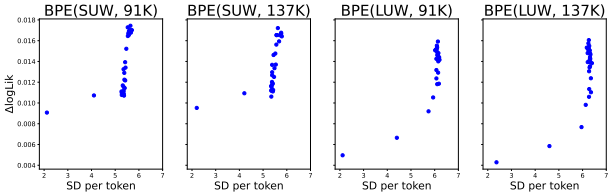


**図 3**　各分割単位による言語モデルのトークン単位の SD（横軸）と ΔLogLik（縦軸）の関係．どのモデルでも右肩上がりの関係が見られる．予備実験の長単位・短単位モデルも同様の挙動を示す．

ン情報量の過剰な一様化が，文節情報量が一様ではない日本語で破綻を引き起こしたと議論した．これに基づき，本研究では短単位で過剰な一様化が発生し，長単位では発生しなかったと考えられる．

そこで，言語モデルにとっての情報量の一様性を，推論データに対して言語モデルが与えるトークンあたりの情報量の標準偏差（SD）で評価する．より小さい SD はより一様であることを示す．図 2 に PPL と SD の関係を示した．長単位系のモデルでは PPL が小さい範囲でも SD が高い値を保つのに対し，短単位系のモデルでは PPL と SD に山型の関係が見られる．したがって，[7] の議論の通り，短単位系のモデルでは過剰な一様化が発生し，長単位系のモデルではこれが発生していないと考えられる．実際，図 3 のとおり SD と ΔLogLik は長単位・短単位の両方で右肩上がりの関係を示す．このことは過剰な一様化が人間らしさを悪化させるという [7] を裏付けるものである．

以上から，長単位モデルで何らかの理由で過剰な一様化（SD の低下）が起こりづらいことが，人間らしさの対応関係を低 PPL 領域で維持することに繋がったと解釈できる．

**統語的な教示**　このような一様化の抑制につながった長単位の特徴として，長単位の統語的な分割単位としての側面を挙げる．BCCWJ において長単位は「構文的な機能に着目した」単位とされており [17]，短単位に比べて統語的な特徴をよく反映すると考えられている．

本研究で構築した言語モデルは品詞を考慮しないため，品詞体系の違いに基づく議論は当てはまらない．しかし長単位では「自立部＋付属部」の形で文節を分割するため，自立語のあとは短単位に比べて高い割合で機能語が出現する．このため機能語の予測は短単位に比べて容易になる．この長単位分割の特徴のため，長単位によるモデルは助詞によって標示される項構造など統語的側面により依存した予測

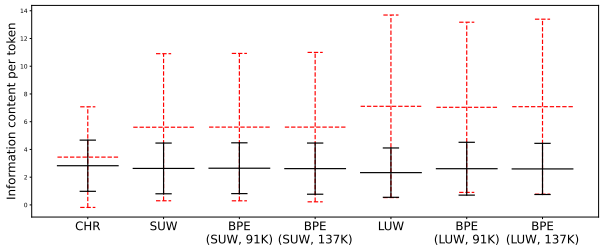


**図 4**　3 エポック学習したそれぞれのモデルで推論データでの 1 文字の助詞のトークンの情報量の分布 (黒い実線) と全体の情報量の分布（赤い破線）．長線が平均値で，短線は 1$\sigma$ の範囲を示す．

を行うと考えられる．実際，図 4 のとおり，長単位系モデルでは助詞トークン[3)]の情報量が全体の情報量に比べて小さい値となる傾向がある．

一方で長単位では複合語が 1 トークンになるため，その情報量は短単位に比べて平均的に大きくなる．これにより，長単位モデルでは統語的に予測しやすい機能語の情報量が小さく，複合語など自立語の情報量は大きくなり，情報量の一様化が起こりにくいと考えられる．本解釈は更なる検証を要するが，長単位分割の統語的な側面によって情報量の過剰な一様化が抑えられ，工学的性能と人間らしさの対応関係の破綻を抑制した可能性がある．

## 5　おわりに

本研究ではトークン分割の観点から工学的性能と人間らしさの関係を検証し，長単位の特徴的な挙動を発見した．サブワード単位を用いて詳細な分析を行い，これが未知語率や語彙数ではなく分割自体の性質に由来することを示した．この結果はトークン分割がこれまで考えられていた以上に大きな影響を持っている可能性を示唆する．

本研究では長単位の統語的な分割単位としての側面がトークンあたり情報量の過剰な一様化を抑制し，これによって短単位で生じるような工学的性能と人間らしさの対応関係の破綻を抑えたとする解釈を提示した．今後の研究ではこのような長単位モデルの挙動がどの程度広く成り立つのかという観点や，長単位モデルの統語的な性能はどうなっているのか，また応用的にどのような利点があるかなど，さらなる詳細な分析を行いたい．

3)　文字単位で 1 トークンとなっている格助詞「がでにへを」，係助詞「はもや」，連体助詞「の」のトークンを助詞トークンと考えて利用した．

## 謝辞

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] John Hale. A Probabilistic Earley Parser as a Psycholinguistic Model. In **Second Meeting of the North American Chapter of the Association for Computational Linguistics**, 2001.

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

[3] 栗林樹生, 大関洋平, 伊藤拓海, 吉田遼, 浅原正幸, 乾健太郎. 日本語の読みやすさに対する情報量に基づいた統一的な解釈. 言語処理学会年次大会発表論文集, March 2021.

[4] Ethan G Wilcox, Tiago Pimentel, Clara Meister, Ryan Cotterell, and Roger P Levy. Testing the predictions of surprisal theory in 11 languages. **Transactions of the Association for Computational Linguistics**, Vol. 11, pp. 1451–1470, 2023.

[5] Adam Goodkind and Klinton Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In **Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)**, pp. 10–18, Salt Lake City, Utah, January 2018. Association for Computational Linguistics.

[6] 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, June 2020.

[7] 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, 2021. Association for Computational Linguistics.

[8] Byung-Doh Oh and William Schuler. Why does surprisal from larger transformer-based language models provide a poorer fit to human reading times? **Transactions of the Association for Computational Linguistics**, Vol. 11, pp. 336–350, 2023.

[9] Michael Hahn, Frank Keller, Yonatan Bisk, and Yonatan Belinkov. Character-based Surprisal as a Model of Reading Difficulty in the Presence of Errors. Preprint, PsyArXiv, May 2019.

[10] Byung-Doh Oh, Christian Clark, and William Schuler. Surprisal Estimators for Human Reading Times Need Character Models. 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. 3746–3757, Online, August 2021. Association for Computational Linguistics.

[11] Sathvik Nair and Philip Resnik. Words, Subwords, and Morphemes: What Really Matters in the Surprisal-Reading Time Relationship?, October 2023.

[12] 浅原正幸, 小野創, 宮本エジソン正. BCCWJ-EyeTrack. 言語研究, Vol. 156, pp. 67–96, 2019.

[13] R. Harald Baayen and Petar Milin. Analyzing reaction times. **International Journal of Psychological Research**, Vol. 3, No. 2, p. 12–28, Dec. 2010.

[14] 新井学, Douglas Roland. 言語理解研究における眼球運動データ及び読み時間データの統計分析. 統計数理, Vol. 64, No. 2, pp. 201–231, 2016.

[15] Alec Radford, Jeff Wu, Rewon Child, D. Luan, Dario Amodei, and Ilya Sutskever. Language Models are Unsupervised Multitask Learners. 2019.

[16] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written Japanese. **Language Resources and Evaluation**, Vol. 48, No. 2, pp. 345–371, June 2014.

[17] コーパス開発センター国立国語研究所, National Institute for Japanese Language, Center for Corpus Development Linguistics. 『現代日本語書き言葉均衡コーパス』利用の手引 第 1.0 版. Technical report, apr 2015. 『現代日本語書き言葉均衡コーパス』利用の手引 第 1.0 版, application/pdf, 国立国語研究所, National Institute for Japanese Language and Linguistics.

[18] Richard Futrell, Edward Gibson, and Roger P Levy. Lossy-context surprisal: An information-theoretic model of memory effects in sentence processing. **Cognitive science**, Vol. 44, No. 3, p. e12814, 2020.

[19] Tatsuki Kuribayashi, Yohei Oseki, Ana Brassard, and Kentaro Inui. Context limitations make neural language models more human-like. **arXiv preprint arXiv:2205.11463**, 2022.

[20] Taku Kudo and John Richardson. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In **Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations**, pp. 66–71, Brussels, Belgium, January 2018. Association for Computational Linguistics.

[21] Julius Steuer, Marius Mosbach, and Dietrich Klakow. Large GPT-like models are bad babies: A closer look at the relationship between linguistic competence and psycholinguistic measures. In Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors, **Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning**, pp. 142–157, Singapore, December 2023. Association for Computational Linguistics.

[22] Dmitriy Genzel and Eugene Charniak. Entropy rate constancy in text. In **Proceedings of the 40th annual meeting of the Association for Computational Linguistics**, pp. 199–206, 2002.

[23] Masayuki Asahara, Hajime Ono, and Edson T. Miyamoto. Reading-Time Annotations for "Balanced Corpus of Contemporary Written Japanese". In **Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers**, pp. 684–694, Osaka, Japan, February 2016. The COLING 2016 Organizing Committee.

## A　分割の例

本研究で扱ったそれぞれの分割単位による分割の例を表 3に示す．

**表 3**　それぞれの分割単位による分割例

| | |
|---|---|
| CHR | 国 / 会 / で / 外 / 国 / 人 / 参 / 政 / 権 / が / 議 / 論 / さ / れ / ま / す |
| SUW | 国会 / で / 外国 / 人 / 参政 / 権 / が / 議論 / さ / れ / ます |
| BPE(SUW, 137K) | 国会 / で / 外国 / 人 / 参政 / 権 / が / 議論 / さ / れ / ます |
| BPE(SUW, 91K) | 国会 / で / 外国 / 人 / 参政 / 権 / が / 議論 / さ / れ / ます |
| LUW | 国会 / で / 外国人参政権 / が / 議論さ / れ / ます |
| BPE(LUW, 137K) | 国会 / で / 外国人 / 参政権 / が / 議論さ / れ / ます |
| BPE(LUW, 91K) | 国会 / で / 外国人 / 参 / 政権 / が / 議論さ / れ / ます |

## B　プレディクタ

読み時間の推定に利用したプレディクタを表 4に示す．各々のプレディクタの詳細は [23] を参照せよ．

**表 4**　それぞれのプレディクタの定義

| | |
|---|---|
| time | 読み時間 |
| length | 文字数 |
| count_ave_kika | 文節に含まれる単語頻度の幾何平均 |
| space | 文節間の空有の有無 |
| articleN | 記事呈示順 |
| sessionN | セッション呈示順 |
| screenN | 画面呈示順 |
| lineN | 行呈示順 |
| segmentN | 文節呈示順 |
| is_first | 行内で最も左にあるか否か |
| is_last | 行内で最も右にあるか否か |
| (1 \| article) | 記事 ID（ランダム効果の切片） |
| (1 \| subj) | 実験協力者 ID（ランダム効果の切片） |

## C　言語モデルのハイパーパラメタ

[7] に倣った上で，学習に関係する部分を変更した．分割単位を変更するため，語彙数を変更した．また，学習に NVIDIA RTX™ A5000 を 1 つ使用したため，パラメタを 24GB のメモリサイズに合わせ調整した．

**表 5**　長単位及び短単位のモデルのハイパーパラメタ

| モジュール | パラメタ | 値 |
|---|---|---|
| GPT-2 | context_size | 512 |
| | n_embed | 384 |
| | n_layer | 8 |
| | n_head | 6 |
| | ffn_dim | 2,048 |
| | dropout | 0.1 |
| | init_params | $\sigma = 0.2$ の切断正規分布からサンプル |
| | attn_pdrop | 0.1 |
| Optimizer | type | AdamW |
| | $lr$ | 2e-4 |
| | betas | (0.9, 0.98) |
| | weight_decay | 0.01 |
| | $\varepsilon$ | 1e-6 |
| Scheduler | scheduler | linear |
| | warm up steps | 40,000 |
| Training | batch_size | 12 |
| | step/epoch | 252,134 |
| | epoch | 3 |