サプライザルと読み時間からみた日本語リーダビリティの分析

王簫影　ホドシチェク ボル
大阪大学　人文学研究科
u989796h@ecs.osaka-u.ac.jp
hodoscek.bor.hmt@osaka-u.ac.jp

概要

リーダビリティ測定は，テキストの測定値を付与することを目的とし，対象言語を問わず自然言語処理分野で長年研究されてきた課題である．近年では，テキストから算出可能な言語的特徴や心理言語学の指標と読み時間との関連が示され，これらを用いて人間の読解プロセスとの対応関係を比較できる．本研究では，日本語母語話者の読み時間データを用い，従来の日本語リーダビリティ計算式及び心理言語学由来の指標を文・段落・談話レベルで比較し，読み時間との相関に基づき平均サプライザル指標の有効性と限界を実証的に検証した．

1　はじめに

リーダビリティ測定システムの研究は，初期の言語特徴で算出する重回帰公式に続いて機械学習の予測モデリングから，近年の大規模言語モデルを用いた難易度判断まで，様々なジャンルのテキストにおいて手法の多様性を示している [1]．また，それらの測定システムは手法に関わらずモデルの訓練に使用されているテキスト材料に依存し，対象者の属性を考慮した上での利用を用いる傾向にあるところに一般化の改善が求められている [2, 3, 4]．

2　関連研究

英語では，心理言語学由来の指標を用いて読み時間とリーダビリティを結びつける研究が進められてきた [5, 6, 7]．それと共に人間の文処理過程を解明するために様々な読み時間データベースが整備され，最初は特定の言語処理プロセスの有無及び前後順序を特定するための繊細なコントロールでデザインされた文を読む時のデータが構築され，その後より読解の実態が再現できる自然文を読む時のデータも作成されてきた [8]．さらに，自然言語処理の研究の急速な進歩と一緒に，多様なメタデータが追加されている読み時間データベースを用いた心理言語学の文処理理論の検証も挙げられる [9, 10]．テキスト材料をベースにした直接の読み時間，眼球運動パターンの予測のほか，読解タスク，読者のリーディングレベルや，テキストのリーダビリティパターンを判定する分類器やプレディクターの作成を目的とした研究も盛んに行われている [7, 11, 12, 13]．

一方，Dundee Corpus や Natural Stories Corpus につぎ，日本語の自然文を読む時の日本語母語話者の読み時間データのコーパス BCCWJ-EyeTrack[14] が整備されたことで，日本語を読んで理解する際の読み時間をモデリングする研究がコーパス開発された後続々と発表され，コーパス作成者からの読み時間に影響を与えるテキストの特徴を網羅的に実証した研究が挙げられる [15, 16, 17, 18]．これらの日本語の読み時間データを用いた研究のモデリングは注視順の文節データポイントを集約せず，焦点である言語情報アノテーションを重ね合わせながら，対数読み時間を用いた [19]．その後，日本語における英語起源の文処理仮説や言語モデルの挙動の検証もされつつあり，異なる言語システムに因んだ読み時間データまたは言語モデルの振る舞いの異同が観測される [20, 21, 22, 23]．しかし読み時間データを用いた文の読みやすさと直接関連しているリーダビリティ測定システムとの研究は殆ど存在しない．

最近，成人英語母語話者のアイトラックコーパスを使用した心理言語学由来の指標と従来のリーダビリティ測定システムとのどちらがよりテキストの難易度変化ゆえの読み時間指標の変化を捉えているかの検証が提示され，他言語や他属性の話者への発展が求められている [24]．本研究では，このような心理言語学由来の指標と従来のリーダビリティ測定システムの計算出力の読み時間との相関を，文・段落・文章まで集約した後の成人日本語母語話者の読み時間データのコーパスを用いて検証を行う．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 1: 実験 A にて使用したデータポイント

BCCWJ-EyeTrack			本研究	
文節数	文数	画面数	行数	画面数
1,643	218	71	250	68

3 実験 A

人間の文処理理論と行動データを繋ぐ Linking hypothesis（橋渡し仮説）としてサプライザル理論 [9, 10] が読み時間モデリング研究において広範に応用されている．文の次にくる単語や文節の予測しやすさをその文脈における条件付き確率の負の対数として定式化され，読み時間などの行動データとの相関が注目されている．

今までの日本語読み時間データ及びサプライザルを用いた研究は，それらの要素を被説明変数に位置し，異なるアーキテクチャーやパラメーターの言語モデルからのサプライザルがどれくらい読み時間との相関を遂げ，それを人間らしさとし，言語モデル中心の検証が行われている [20, 21, 22, 23]．

本研究では，言語モデルのサプライザル計算力から人間らしさに基づいた選別のほか，[24, 25, 26] のようにトーケンごとのサプライザルの各単位までの総和を指標とし，テキストベースのリーダビリティ中心の評価を比較した．また，リーダビリティを算出するための最小単位『文』まで集約することにより，対数時間の代わりに各読み時間指標をそのままの時間データを用いた．

3.1 方法

読み時間データベース 実験 A で使用していた読み時間データベース BCCWJ-EyeTrack[14] は視線計測コーパスであり，『現代日本語書き言葉均衡コーパス』（以下 BCCWJ）[27] のコアデータの PN サブコーパスの 20 篇の新聞記事サンプルに対し，五行ずつ 24 名の日本語母語話者に呈示され，文節ごとの読み時間が付与されている．六つの読み時間指標を全部採用し，[14] に則り，本文ではない文節データポイントを削除した．パラグラフ境界を screenN の情報により同じスクリーンに提示されている文のかたまりとして定義した．さらに，文節間に全角空白の有無と，それに関わらず，3 通りのデータに分けて相関を図った．本文のみの文節データポイントを文・段落まで集約した後 [14] との比較を表 1 に示している．

リーダビリティ指標 読み時間データの本文を用いて計算した従来リーダビリティ指標は: ARI(Automated Readability Index)，jReadability 公式 (以下 Lee-formula) と柴崎の難易度計算公式 (以下 Shibasaki-formula) の 3 つである．

ARI は，一単語当たりの平均文字数と一文当たりの平均単語数を用いて英語文書のリーダビリティを計算するための式である．[1] 算出するスコアは 1 から 13 で，高ければ高いほどテキストが読み難くなる．

Lee-formula は jReadability の日本語学習者のためのリーダビリティ公式 [3] で，[2] 文章単位の読みやすさスコアを平均文長，漢語率，和語率，動詞率と助詞率を用いて計算する．算出されたスコアは 0.5 から 6.4 の範囲に収まって，高ければ高いほどテキストが読みやすくなる．

Shibasaki-formula は [4] で紹介された日本語リーダビリティ計算式で，[3] 文書の総文字数に対するひらがなの割合と一文の平均述語数から，テキスト難易度に相当する学年が算出される．日本語国語教科書を用いて小・中学校の文章難易度学年判定が目的なため，計算で得た値は 1 から 12 までの範囲に収まる．

心理言語学指標 心理言語学由来の指標は平均サプライザル (Surprisal)，語彙密度 (Lexical density)，平均頻度，平均長さである．

Lexical density は語彙密度で，書き言葉を研究対象とする場合は，いくつかの方法で計算されるようになっている．本研究における語彙密度の計算項目に関して，内容語は名詞，副詞，形容詞，動詞，形状詞を集計対象とし，機能語は助詞と助動詞を集計対象としてカウントする．[4]

Suprisal は Suprisal theory[9, 10] に由来で，今までのコンテクストであるトーケンが出現することがどれくらい予想がつくかという文脈における次に来る内容の予測性を数式で表している．ここでは Minicons[5] [28] の sequence_score でファインチューニングなし GPT2[6] を使用して一文・一パラグラフのサプライザルを算出した．また，日本語における

1) $ARI = 4.71 \times \frac{x_{文字数}}{x_{単語数}} + 0.5 \times x_{一文あたりの平均単語数} - 21.43$
2) 文章の読みやすさ $= -0.056 \times x_{平均文長} - 0.126 \times x_{漢語率} - 0.042 \times x_{和語率} - 0.145 \times x_{動詞率} - 0.044 \times x_{助詞率} + 11.724$
3) 学年 $= 14.016 - 0.145 \times x_{ひらがな率} + 0.587 \times x_{平均述語数}$
4) $Lexicaldensity = \frac{x_{機能語数}}{x_{内容語数}}$
5) https://github.com/kanishkamisra/minicons
6) https://huggingface.co/rinna/japanese-gpt2-medium

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

語彙単位の曖昧性を考慮しながら英語の先行研究と比較するために，文やパラグラフのサプライザル値をそのまま使うのではなく，英語研究で一般的なトーケン正規化に加え，文節単位での正規化を事前仮説として設定した．式は下に参照:

$$\text{surp\_avg} = \frac{1}{|\text{Seq}|} \sum_{w \in \text{Seq}} -\log_2\big(p(w \mid \text{context})\big) \qquad (1)$$

更に，先行研究と同じパッケージ `Wordfreq`[7] を使って単語頻度をカウントし，可視化した．

3.2 結果

文まで集約した文節間に全角空白なしの六つの読み時間指標との相関を図 1 で示している．計算している読み時間指標とリーダビリティ指標の異同を踏まえ，今回使用した日本語母語話者の行動データを分析するに不適切なリーダビリティ指標 ARI と，どのリーダビリティ指標との相関も低い読み時間指標：First Fixation time と Second Pass time があり，後に行う追加分析で除外する必要性があると示されている．段落まで集約した相関や文節間空白あり・全体の文と段落ベースの相関図は付録に参照．

英語の先行研究 [24] と同じく文単位・パラグラフ単位に関わらず文節数で正規化した平均サプライザルが読み時間指標との相関が一番高い．また，伝統的公式が読み時間との相関を捉えている項目がありつつ，完全に捉えていない項目もある．ARI は英語・日本語両方算出可能でありつつ両方の行動データとの相関が低い．

一方，文単位・パラグラフ単位に関わらず，トーケンで正規化した平均サプライザルより，文節で正規化した平均サプライザルの方が日本語母語話者の読解プロセスを捉えている．英語と対等な短単位ではなく，日本語母語話者にとってより長い語彙単位をベースにしたオンライン読解プロセスの方が自然だということを示している．

さらに shibasaki-formula と lee-formula の読み時間との相関の差から分かるように，文節で正規化した平均サプライザルの次に相関が高い指標は Shibasaki-formula で，従来指標の説明で述べた通り，Shibasaki-formula の計算式に一番ウェイトの高いの言語項目が「述語数」で，文節・長単位語彙ユニットとの重なりが多く，日本語母語話者のフォーカスを捉えていると言える．その反面，日本語学習者にまつわるテキスト資料で訓練された学習者のための

7) https://github.com/rspeer/wordfreq

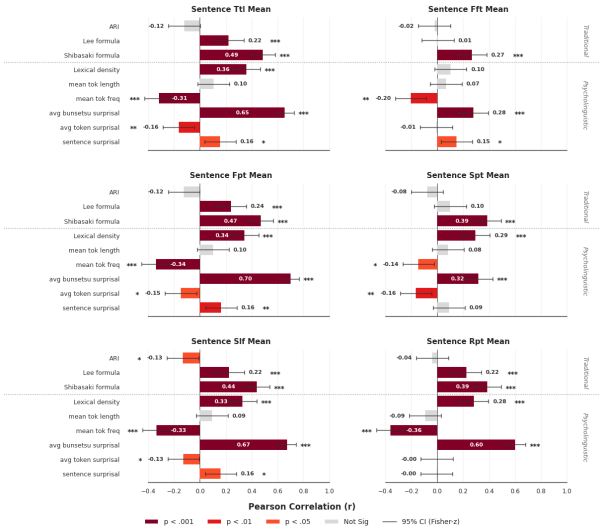


図 1: 文節間空白なし文単位の各指標と読み時間との相ピアソン相関．エラーバーは Fisher の z 変換に基づく 95%信頼区間を示す．

リーダビリティ計算式 Lee-formula が日本語母語話者の行動データとの相関が低く，式の訓練データと対象母集団の間の相違に起因すると考えられる．

本研究では [24] に倣い，指標間の対応関係を明示的に比較する目的から，相関分析を主とした．個人差や文書差を考慮した混合効果モデルによる検証は今後の課題とする．

4 実験 B

[24] では難易度に変化がある段落ペアーからなるテキスト材料を対象に収集した眼球運動データを使用しているため，文・段落レベルの難易度の差に因んだ読み時間または各リーダビリティ指標の差が算出可能になっている．既存の日本語読み時間データベースから明確なテキスト難易度アノテーションが付与されているコーパスがないため，BCCWJ-EyeTrack の他，学年情報が追加されている教科書を収集対象としている BCCWJ-SPR2[29] の OT レジスターデータを用い，難易度変化と共に，文章レベルの読み時間，リーダビリティ指標または心理言語学由来指標の変化を明らかにした．

4.1 方法

読み時間データベース 実験 B で使用していた読み時間データベース BCCWJ-SPR2[29] は自己ペース読文法による読み時間コーパスであり，BCCWJ[27] の OT レジスターの総計 37 篇のサンプルに対し，200 名の日本語母語話者の，文節ごとの読み時間が

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

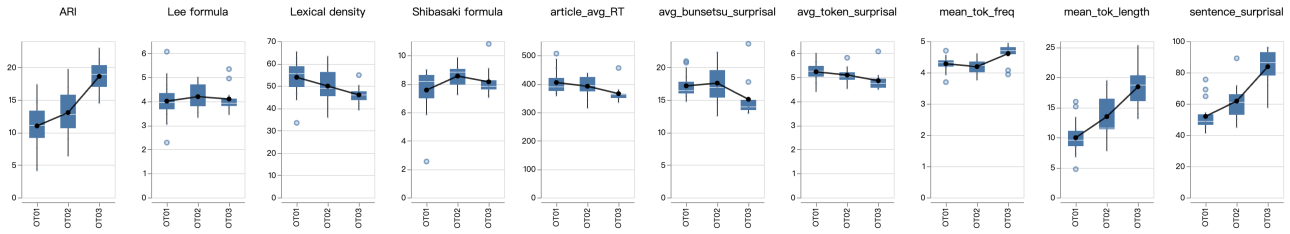


図 2: 各指標の学年国語教科書における分布．小（OT01）・中（OT02）・高校（OT03）のサンプル数は 17:9:11．記事間の差異は Kruskal–Wallis 検定を用いて評価した．

表 2: 実験 B で使用したデータポイント

BCCWJ-SPR2/OT			本研究	
文節数	文数	サンプル数	文数	サンプル数
50,606	9,521	38	9,351	37

付与されている．作成者が GitHub に挙げている文節・文のテキスト情報がマスクされているデータ[8]と，BCCWJ-NT データを重ね合わせながら復元した．テキスト難易度アノテーションの代表できる学年情報をサンプル ID により復元した．本文の文データポイントを文章（サンプル）まで集約した後 [29] との比較を表 2 に示す．

4.2 結果

文章（サンプル）まで集約した小（OT01）・中（OT02）・高校（OT03）レベルの国語サンプルについて，各指標の分布を図 2 に示す．レベル別にサンプル単位の平均読み時間の分布を見ると，小・中レベルは外れ値を除けば大きく重なり合う一方で，高校レベルでは分布の位置が異なる傾向が観察される．小・中レベルにおいて平均読み時間が相対的に長くなる傾向は，難易度ラベルと整合しない傾向になっているが，被験者が成人読者であることを考慮すると，高校レベルの国語教科書に対する馴染みの影響が示唆される．

区別度の観点からは，平均トーケン長において学年間の分布差が比較的大きく，学年が上がるにつれて国語教科書により長い語が含まれる傾向が確認できる．同様の傾向は文の平均サプライザルにも見られ，より長いテキストを対象とする場合，文節やトーケン単位まで正規化すると情報量が減少し，区別が困難になる可能性が示唆される．さらに従来のリーダビリティ公式の値の分布からは，国語教科書サンプルの難易度が中学レベル付近に集中し，テキスト材料自体に顕著な難易度幅が存在しないことが推測される．この点については，学年メタデータとの対応関係を付録に示す．

5 議論

平均文節サプライザルの有効性に限界がある　英語を対象とした先行研究と同じく視線計測コーパスの読み時間データにおいては，平均単位サプライザルが今回用いた指標の中に一番読み時間との相関が高かった．しかし，テキスト長の増加につれその有効性が失われ，サプライザルをリーダビリティ測定システムに取り込む場合には，平均以外の集約方法と比較した上での実装が求められている．

リーダビリティ測定システム拡張への示唆　日本語学習者向けの計算式による算出値は，母語話者の読み時間データとの相関が相対的に低いことが確認された．この結果は，既存のリーダビリティ指標が読者背景の違いを十分にモデル化できていない可能性を示している．また，多様なレジスターにまたがるテキストにおいて，L1・L2 読者の行動データを考慮することの重要性を示す結果である．

6 おわりに

日本語母語話者の行動データにおける読み時間との相関を英語先行研究との比較分析により，平均サプライザルの説明有効性は英語以外の他言語への拡張が可能となりつつ，自然言語処理の研究手法と心理言語学実験データとの合同で今までのリーダビリティ測定システムのパフォマンスをより一層改善できる可能性を示唆している．

また，学年の代わりにテキスト相対難易度のメタ情報付き日本語コーパスと日本語学習者の行動データの不足から，文書難易度付き日本語パラレルコーパスを用いた視線走査法と自己ペース読文法実験の必要性も示されている．

8） https://github.com/masayu-a/BCCWJ-SPR2

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は，次世代研究者挑戦的研究プログラムの助成を受けたものです．

参考文献

[1] Isabel Cachola, Daniel Khashabi, and Mark Dredze. Evaluating the evaluators: Are readability metrics good measures of readability? **arXiv [cs.CL]**, 2025.

[2] 艷萍楊, 賀津雄玉岡, 張, Yanping Yang, Katsuo Tamaoka, Jingyi Zhang. 中国人日本語学習者の文法能力は作文の語彙特性にどう影響するか. ことばの科学, Vol. 33, pp. 147–165, 12 2019.

[3] 在鎬李. Bccwj に含まれる学校教科書コーパスの計量的分析. 計量国語学, Vol. 32, No. 3, pp. 147–162, 2019.

[4] 秀子柴崎, 信一郎原. 12 学年を難易尺度とする日本語リーダビリティー判定式. 計量国語学 = Mathematical linguistics : 計量国語学会機関誌, Vol. 27, No. 6, pp. 215–232, 2010.

[5] Daniel R Hittleman. Seeking a psycholinguistic definition of readability. **The Reading Teacher**, Vol. 26, No. 8, pp. 783–789, 1973.

[6] Sowmya Vajjala and Detmar Meurers. On improving the accuracy of readability classification using insights from second language acquisition. **Proceedings of the Seventh Workshop on Building Educational Applications Using NLP**, pp. 163–173, June 2012.

[7] David M Howcroft and Vera Demberg. Psycholinguistic models of sentence processing improve sentence readability ranking. In **Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers**, Stroudsburg, PA, USA, 2017. Association for Computational Linguistics.

[8] Charles Clifton, Jr, Adrian Staub, and Keith Rayner. Eye movements in reading words and sentences. In **Eye Movements**, pp. 341–371. Elsevier, 2007.

[9] 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.

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

[11] Daniel Wiechmann, Yu Qiao, Elma Kerz, and Justus Mattern. Measuring the impact of (psycho-)linguistic and readability features and their spill over effects on the prediction of eye movement patterns. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 5276–5290, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics.

[12] Nora Hollenstein, Marius Tröndle, Martyna Plomecka, Samuel Kiegeland, Yilmazcan Özyurt, Lena A Jäger, and Nicolas Langer. Reading task classification using EEG and eye-tracking data. **arXiv [cs.CL]**, December 2021.

[13] Nora Hollenstein, Itziar Gonzalez-Dios, Lisa Beinborn, and Lena Jäger. Patterns of text readability in human and predicted eye movements. In Michael Zock, Emmanuele Chersoni, Yu-Yin Hsu, and Enrico Santus, editors, **Proceedings of the Workshop on Cognitive Aspects of the Lexicon**, pp. 1–15, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics.

[14] 浅原正幸, 小野創, 宮本 エジソン正. BCCWJ-EyeTrack ——『現代日本語書き言葉均衡コーパス』に対する読み時間付与とその分析——. 言語研究 (Gengo Kenkyu), Vol. 156, pp. 67–96, 2019.

[15] Masayuki Asahara. Between reading time and information structure. In Rachel Edita Roxas, editor, **Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation**, pp. 15–24. The National University (Phillippines), November 2017.

[16] Masayuki Asahara and Sachi Kato. Between reading time and syntactic/semantic categories. In Greg Kondrak and Taro Watanabe, editors, **Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)**, pp. 404–412, Taipei, Taiwan, November 2017. Asian Federation of Natural Language Processing.

[17] 浅原正幸, 小野創, 宮本エジソン正. 『現代日本語書き言葉均衡コーパス』の読み時間とその被験者属性. 言語処理学会第 23 回年次大会 発表論文集, pp. 473–476, 2017. P6-5.

[18] Masayuki Asahara. Between reading time and clause boundaries in japanese - wrap-up effect in a head-final language. In Stephen Politzer-Ahles, Yu-Yin Hsu, Chu-Ren Huang, and Yao Yao, editors, **Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation**, Hong Kong, dec 2018. Association for Computational Linguistics.

[19] 浅原正幸. テキストの読みやすさについて. 言語処理学会第 25 回年次大会 発表論文集, pp. 245–248, 2019. C3-1.

[20] 栗林樹生, 大関洋平, 伊藤拓海, 吉田遼, 浅原正幸, 乾健太郎. 予測の正確な言語モデルがヒトらしいとは限らない. 言語処理学会第 27 回年次大会 発表論文集, pp. 267–272, 2021. C2-3.

[21] 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. 言語処理学会第 27 回年次大会 発表論文集, pp. 723–728, 2021. P4-1.

[22] 吉田遼, 能地宏, 大関洋平. 再帰的ニューラルネットワーク文法による人間の文処理のモデリング. 言語処理学会第 27 回年次大会 発表論文集, pp. 273–278, 2021. C2-4.

[23] 三輪敬太, 吉田遼, 大関洋平. 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. 言語処理学会第 30 回年次大会 発表論文集, pp. 1908–1913, 2024. D7-6.

[24] Keren Gruteke Klein, Shachar Frenkel, Omer Shubi, and Yevgeni Berzak. Eye tracking based cognitive evaluation of automatic readability assessment measures. **arXiv [cs.CL]**, 2025.

[25] 田村鴻希, 土井惟成, 西田直人, Junjie Chen, 谷中瞳. サプライザルを利用した日本語の流暢性フィルタリングの試み. 言語処理学会第 29 回年次大会 発表論文集, pp. 2111–2116, 2023. A9-4.

[26] 山下陽一郎, 原田宥都, 大関洋平. 早押しクイズにおける超次単語予測の認知モデリング. 言語処理学会第 30 回年次大会 発表論文集, pp. 1892–1896, 2024. D7-3.

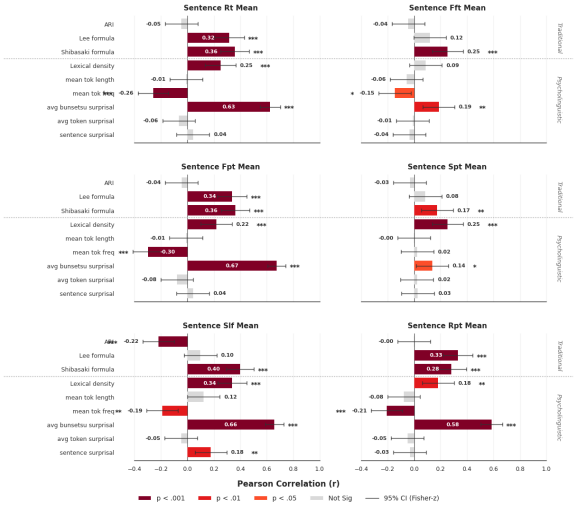
[27] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written japanese. **Language resources and evaluation**, Vol. 48, No. 2, pp. 345–371, 2014.

[28] Kanishka Misra. minicons: Enabling flexible behavioral and representational analyses of transformer language models. **arXiv preprint arXiv:2203.13112**, 2022.

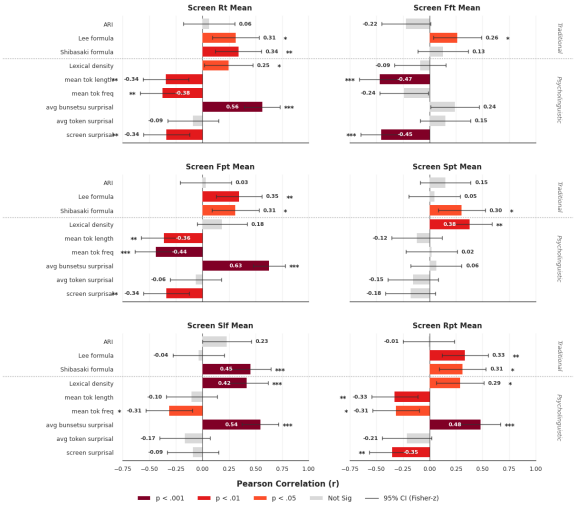
[29] Masayuki Asahara. Reading time and vocabulary rating in the Japanese language: Large-scale Japanese reading time data collection using crowdsourcing. In Nicoletta Calzolari, Frédéric Béchet, Philippe Blache, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Jan Odijk, and Stelios Piperidis, editors, **Proceedings of the Thirteenth Language Resources and Evaluation Conference**, pp. 5178–5187, Marseille, France, June 2022. European Language Resources Association.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

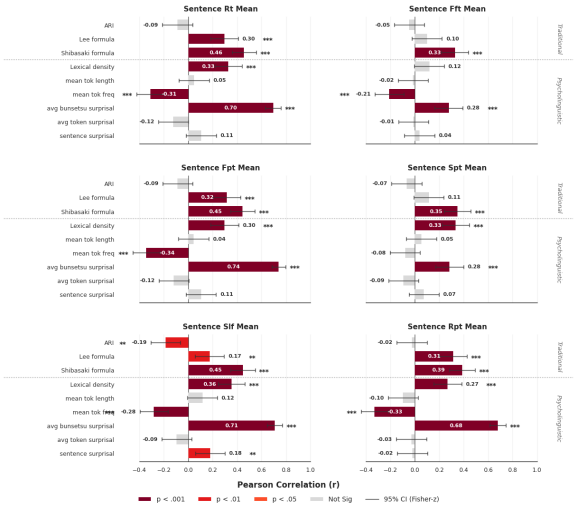
A　各指標と読み時間との相関 (a-e) と学年情報の分布 (f)



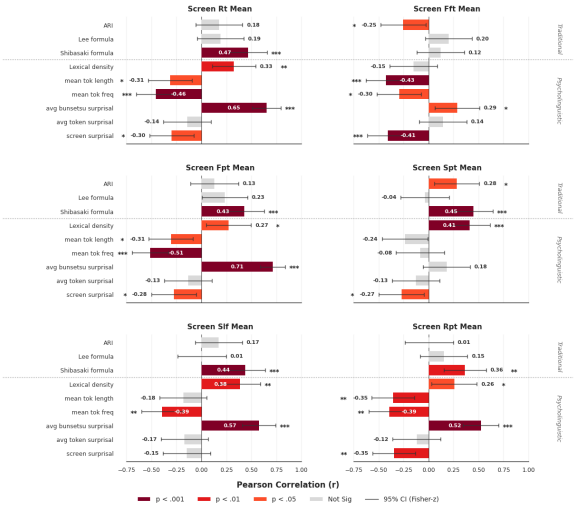
(a) 文節間空白あるセンテンスベースの相関



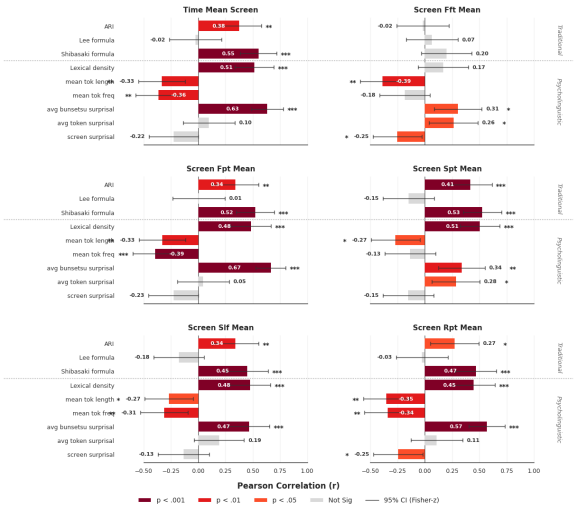
(b) 文節間空白あるパラグラフベースの相関



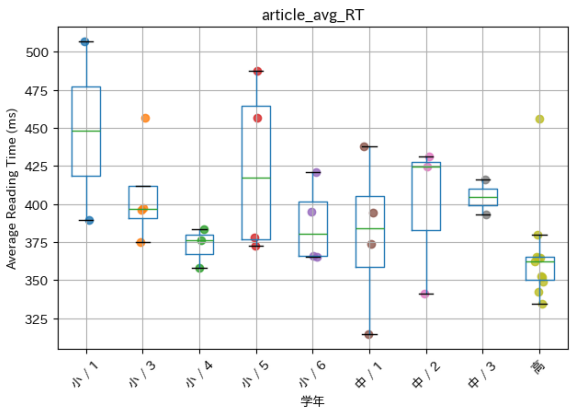
(c) 全体センテンスベースの相関



(d) 全体パラグラフベースの相関



(e) 文節間空白なしパラグラフベースの相関



(f) 同一学年内における教材処理負荷のばらつき

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).