テキストの読みやすさについて

浅原 正幸 *
国立国語研究所

1. はじめに

テキストの読みやすさ・難易度を眼球運動に基づく読み時間の調査により定量的に評価することを試みる。過去の研究では、ヒューリスティックな特徴量（何年生で習う漢字か、日本語教育語彙表、係り受けの深さなど）に基づき、頻度主義的な考え方に基づく高次関数へのあてはめや識別学習を行う傾向がある。これらは、特徴量の選定が内省や他言語に適用された先行研究に基づくものであり、証拠の定量化が弱い。これに対して、『現代日本語書き言葉均衡コーパス』の新聞サンプルに対する日本語成人母語話者の眼球運動計測に基づいた読み時間データ (BCCWJ-EyeTrack)(Asahara et al. 2016) と、各種言語情報アノテーションを重ね合わせたうえで、ベイジアン線形混合モデル（一部、頻度主義的な一般化線形混合モデル）に基づく線形回帰を行い、被験者の揺れやサンプルの揺れをランダム効果として考慮しながらモデル化し、**どのようなテキストの特徴が読み時間に影響を与えるか**について実証研究を行ったので報告する。特徴ごとの読み時間の差異から、テキストの読みやすさが何によって変わるのかを検討する。

2. 読み時間に影響を与えるテキストの特徴

2.1 読み時間とレイアウト情報

読み時間に影響を与えるものとしてレイアウト情報がある。読み時間を評価する際には、等幅フォントで刺激文を横書きで呈示するが、その呈示文の両端では、復帰改行のような眼球移動が行われる。Asahara et al. (2016) によると、文節単位の読み時間は行内の、最左要素が短く、右から 2 番目の要素が長い。また、実験環境への慣れや実験が進むにつれて、読み時間が短くなる傾向がある。読み時間の分析を行ううえで、これらの要因を考慮する必要がある。

テキストを単純に読みやすくするために、文節間に半角空白を入れることが考えられる。BCCWJ-EyeTrack の調査によると、**文節境界に空白を入れたほうが、入れないときよりも移動距離が長くなるにも関わらず、読み時間が短くなる傾向**が確認されている。

2.2 読み時間と被験者特性

浅原ほか (2017) は、被験者特性として、被験者の記憶力テスト結果と語彙力テストを要因に入れた統計分析を行っている。記憶力テストとしてリーディングスパンテスト (苧坂 2002) を、語彙力テストとして NTT の語彙数判定テスト (天野・近藤 1999) を用いた。

記憶力がある群は、最初に 1 回文字列をなぞる速度は速いものの、複数回読む傾向があり、全体の読み時間は変わらなかった。また、**語彙力がある群は読み時間が長くなる傾向**が見られた。これは一部の人の期待する結果に反するが、一つの理由として、語彙力がある群は、テキスト中に出てくる語句の連接の予測する空間（連接語句や可能な語義）が大きく、そのことがかえって読み時間を遅くしている可能性がある。この点については、2.8 節で再度検討する。

* masayu-a@ninjal.ac.jp

2.3 読み時間と係り受け

BCCWJ-EyeTrack には、あらかじめ文節係り受けアノテーション BCCWJ-DepPara (Asahara and Matsumoto 2016) による、文節単位の係り受けの数が付与されている。統計分析結果から、**係り受けの数が多いほど読み時間が短くなる傾向**が明らかになった。

また、英語において句構造木の高さ（深さ）がリーダビリティに影響を与えるということから、日本語において非終端記号を持たない bilexical な文節係り受け木を用いて、係り受けの深さがリーダビリティに影響を与えるという研究が多々ある。これらに対して、本分析結果は、係り受けの数が、後続要素の予測に効果的で、読みを促進するという結果を示す。

2.4 読み時間と節境界

Asahara (2018a) は、節境界アノテーション BCCWJ-ToriClause(Matsumoto et al. 2018) を重ね合わせ、節末や節分類情報による読み時間の影響を調査した。

英語においては、wrap-up effect (Warren et al. 2009) と呼ばれる、節末でそれまでの部分木を統合する処理が行われ、読み時間が長くなるという主張もある(1)。

日本語においては、係り受けの数の効果を考慮したうえでも、**並列節以外の節末において読み時間が短くなる傾向**がみられた。また、並列節においても、wrap-up effect のように読み時間が長くなるという傾向は確認できなかった。日本語のような主辞後置言語においては、従属節が基本的に主節に先行し、これらが予測に効くために、節末で読み時間が短くなる傾向があるのではないかと考える。節の種類においては、次のような傾向がみられた。名詞修飾節においては、「関係節ウチの関係」が「関係節ソトの関係」よりも読み時間が短くなった。補足節においては、名詞節が引用節よりも読み時間が短くなった。副詞節においては、付帯状況よりも因果関係のほうが読み時間が短くなる傾向がみられた。

一部のリーダビリティ評価において、英語で言われている条件節（仮定節）を特徴量として導入するものがあるが、日本語においては、頻度 5 ではあるが、読み時間を短くする傾向が確認されている。

2.5 読み時間と述語項構造・外界照応

Asahara (2018b) は、述語項構造・共参照アノテーション BCCWJ-PAS (植田ほか 2015) を重ね合わせることで、述語項構造・共参照情報が読み時間にどのような影響を与えるかについて調査した。

英語と異なり、日本語においてはガ格も含めて必須格が省略される傾向にあり、省略された名詞句要素をゼロ代名詞と呼ぶ。先に述べた係り受けの数は、ゼロ代名詞はカウントされていないが、基本的に旧情報であるために読み時間を促進することが予測できる。

分析結果から、**主語のゼロ代名詞が外界 2 人称を指す場合に、述語要素で 2 度目以降の読み時間が短くなる傾向**がみられた。そのほか、主語が外界照応である場合には、読み時間が若干短くなる傾向がみられた。

2.6 読み時間と統語・意味分類

Asahara and Kato (2017) は、分類語彙表番号アノテーション BCCWJ-WLSP (Kato et al. 2018) と重ね合わせることで、語句の統語・意味分類との対照比較を行った。

統計分析の結果、**統語分類は、用の類 < 相の類 < 体の類 で読み時間が長くなる傾向**がみられた。また、**意味分類は、関係の部門 < 主体の部門 ∼ 活動の部門 ∼ 生産物の部門 ∼ 自然の部門のように、関係の部門だけ読み時間が短くなる傾向**がみられた。用の類は、項の数が多いために予測が効き、読み時間が短くなる傾向があると考える。関係の部門は、「ＡとＢの関係」のように複数の要素に関連する情報を持つために、文脈に内在する関連する情報が予測に効く傾向があると考える。

(1) 一方、本研究の発表時に、中国語でも wrap-up effect は存在しないという意見も聞かれた。

2.7 読み時間と名詞句の情報の状態

Asahara (2017), 浅原 (2018a) は、BCCWJ-Infostr Miyauchi et al. (2017) を重ね合わせることで名詞句の情報の状態が読み時間に与える影響について調査した。

BCCWJ-Infostr では、名詞句の情報の新旧に関して、文脈中に既出か未出かという書き手視点の情報状態 (information status) と、読み手が既知か想定可能 (ブリッジング) か未知かという共有性の 2 つの観点の情報が付与されている。読み時間においては、**読み手の観点の共有性に関して、既知 < 想定可能 < 未知 の順で読み時間が長くなる傾向**がみられた。また、不定名詞句より定名詞句のほうが読み時間が長い、有生名詞句より無生名詞句のほうが読み時間が長いという傾向もみられた。

2.8 読み時間と語彙の統計的ふるまい

Hale (2001) は、頻度が文処理過程に影響を与えると言及し、漸進的な文処理の困難さについて情報量基準に基づいたモデルを Surprisal theory として定式化している。この Surprisal に基づく日本語の読み時間の分析が求められている。

しかしながら、日本語においては、心理言語学で行われる読み時間を評価する単位と、コーパス言語学で行われる頻度を評価する単位に齟齬があり、この分析を難しくしていた。具体的には、前者においては一般的に統語的な基本単位である文節が用いられるが、後者においては斉一な単位である短い語（国語研短単位など）が用いられる。

浅原 (2018b) は、この齟齬を吸収するために、単語埋め込み (Mikolov et al. 2013) の利用を提案した。単語埋め込みは前後文脈に基づき構成することにより、単語の置き換え可能性を低次元の実数値ベクトル表現によりモデル化する。このうち skip-gram モデルは加法構成性を持つと言われ、句を構成する単語のベクトルの線形和が、句の置き換え可能性をモデル化できる (Mikolov et al. 2013)。

日本語の単語埋め込みとして、『国語研日本語ウェブコーパス』(NWJC)(Asahara et al. 2014) から fastText (Bojanowski et al. 2017) により構成した NWJC2vec (Asahara 2018c) を用いた。ベイジアン線形混合モデルに基づく統計分析の結果、文節内の形態素の頻度の幾何平均と読み時間が逆相関の関係があるほか、Skip-gram モデルに基づく単語埋め込みのノルムと隣接文節間のコサイン類似度が、読み時間を予測する因子となりうることが分かった。**前者のノルムが連接する文節の多様性を、後者の隣接文節間のコサイン類似度が隣接確率をモデル化する**ことが分かった。語彙の連接多様性が読み時間に影響を与えることは、語彙力があるものが多様な連接を想起するがために読み時間が長くなるという傾向の一つの仮説となりうる。今後、より詳細な調査が求められる。

前節までの分析は、言語学的知識を持った者が時間をかけてアノテーションを行う必要があった。しかしながら、本節の単語埋め込みによるモデルは、形態素解析を行うことにより、文節単位の読み時間をモデル化する特徴量を展開することができる。これにより、任意の入力テキストに対して、読み時間の自動推定器を線形式により構成することもできる。

3. おわりに

本稿では、読み時間と観点からテキストの読みやすさに影響を与える要因について、ここ数年の調査で明らかにしたことをまとめた。テキストの読みやすさについては、国語教育や日本語教育の文脈で検討する前に、日本語母語話者がどのようにテキストを読むのかを明らかにする必要がある。本研究は読み時間とアノテーションとの対照により、日本語母語話者の行動特性について明らかにした。本研究の工学応用における重要な点として、以下があげられる：(1) 眼球運動という人間の行動記録から言語の受容過程を直接的に評価した；(2) 文書要約においては、テキストを短くするだけでなく、テキストの読み時間を短くすることが求められるが、眼球運動により直接評価できる；(3) 本調査は、一般化線形

混合モデルやベイジアン線形混合モデルに基づく、線形式によりモデル化するもので、どの要因が読み時間に影響を与えるのかについて説明しやすい。

今後の展開として、読み時間データの拡充 (森山ほか 2019) があげられる。

謝　辞

本研究は国立国語研究所コーパス開発センター共同研究プロジェクトおよび科研費 JP17H00917, JP18H05521, 18K18519 によるものです。

文　献

M. Asahara, H. Ono, and E. T. Miyamoto (2016). “Reading-Time Annotations for ”Balanced Corpus of Contemporary Written Japanese”.” *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pp. 684–694.

浅原正幸・小野創・宮本 エジソン正 (2017).「『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. 473–477.

苧坂満里子 (編) (2002).『ワーキングメモリー脳のメモ帳』　新曜社.

天野成昭・近藤公久 (編) (1999).『単語親密度』　三省堂　1 巻 NTT データベースシリーズ日本語の語彙特性.

M. Asahara, and Y. Matsumoto (2016). “BCCWJ-DepPara: A Syntactic Annotation Treebank on the ‘Balanced Corpus of Contemporary Written Japanese’.” *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pp. 49–58.

M. Asahara (2018a). “Between Reading Time and Clause Boundaries in Japanese – Wrap-up Effect in a Head-Final Language.” *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32)*.

S. Matsumoto, M. Asahara, and S. Arita (2018). “Japanese Clause Classification Annotation on the ‘Balanced Corpus of Contemporary Written Japanese’.” *Proceedings of the 13th Workshop on Asian Language Resources (ALR13)*, pp. 1–8.

T. Warren, S. J. White, and E. D. Reichle (2009). “Investigating the causes of wrap-up effects: Evidence from eye movements and E-Z Reader.” *Cognition*, 111:1, pp. 132–137.

M. Asahara (2018b). “Between Reading Time and Zero Exophora in Japanese.” *READ2018: International Interdisciplinary Symposium on Reading Experience and Analysis of Documents*, pp. 34–36.

植田禎子・飯田龍・浅原正幸・松本裕治・徳永健伸 (2015).「『現代日本語書き言葉均衡コーパス』に対する述語項構造・共参照関係アノテーション」　第 8 回コーパス日本語学ワークショップ予稿集, pp. 205–214.

M. Asahara, and S. Kato (2017). “Between Reading Time and Syntactic/Semantic Categories.” *Proceedings of the 8th International Conference on Natural Language Processing (IJCNLP-2017)*, pp. 404–412.

S. Kato, M. Asahara, and M. Yamazaki (2018). “Annotation of ’Word List by Semantic Principles’ Labels for the Balanced Corpus of Contemporary Written Japanese.” *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32)*.

M. Asahara (2017). “Between Reading Time and Information Structure.” *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation (PACLIC 31)*.

浅原正幸 (2018a).「名詞句の情報の状態と読み時間について」　自然言語処理, 25:5, pp. to appear.

T. Miyauchi, M. Asahara, N. Nakagawa, and S. Kato (2017). “Information-Structure Annotation of the ‘Balanced Corpus of Contemporary Written Japanese’.” *Proceedings of International Conference of the Pacific Association for Computational Linguistics (PACLING 2017)*, pp. 155–165.

J. Hale (2001). “A probabilistic earley parser as a psycholinguistic model.” *Proceedings of the second conference of the North American chapter of the association for computational linguistics* Vol. 2., pp. 159–166.

浅原正幸 (2018b).「単語埋め込みに基づくサプライザルのモデル化」　日本言語学会第 157 回予稿集, pp. 82–87.

T. Mikolov, K. Chen, G. Corrado, and J. Dean (2013). “Efficient Estimation of Word Representations in Vector Space.” *International Conference on Learning Representations*.

T. Mikolov, I. Sutskever, K. Chen, G. Corrado, and J. Dean (2013). “Distributed Representations of Words and Phrases and their Compositionalyty.” *Advances in Neural Information Processing Systems*, pp. 3111–3119.

M. Asahara, K. Maekawa, M. Imada, S. Kato, and H. Konishi (2014). “Archiving and Analysing Techniques of the Ultra-large-scale Web-based Corpus Project of NINJAL, Japan.” *Alexandria: The Journal of National and International Library and Information Issues*, 25:1–2, pp. 129–148.

P. Bojanowski, E. Grave, A. Joulin, and T. Mikolov (2017). “Enriching Word Vectors with Subword Information.” *Transactions of the Association for Computational Linguistics*, 5, pp. 135–146.

M. Asahara (2018c). “NWJC2Vec: Word embedding dataset from ‘NINJAL Web Japanese Corpus’.” *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, 24:2, pp. 7–25.

森山奈々美・荻原亜彩美・近藤森音・浅原正幸・相澤彰子 (2019).「BCCWJ-EyeTrack2: 書籍と教科書データに対する読み時間付与」　言語処理学会第 25 回発表論文集, pp. to appear.