BCCWJ-EyeTrack-2: 書籍と教科書の読み時間データ

森山 奈々美　荻原 亜彩美　近藤 森音　浅原 正幸 *　相澤 彰子
国立国語研究所　津田塾大学　東京大学　国立国語研究所　国立情報学研究所

1. はじめに

テキストの読みやすさ（リーダビリティー・レジビリティ）の研究の基礎データとして、均衡コーパスに対する読み時間データが求められている。Asahara et al. (2016) は『現代日本語書き言葉均衡コーパス』(BCCWJ) (Maekawa et al. 2014) の新聞記事を刺激文とした、日本語母語話者 24 名分の読み時間データを収集した。しかしながら、BCCWJ の新聞記事データは商用利用が認められていない。加えて、テキストの読み時間データとして、ほかの分野のデータも求められている。そこで、BCCWJ-EyeTrack の拡充のため、書籍 (PB) と教科書 (OT) に対する読み時間データを構築した。本稿では、読み時間データの設計について説明するとともに、BCCWJ-EyeTrack-2 の基礎統計を示す。

2. 設計

表 1　読み時間データの設計

	(Asahara et al. 2016) BCCWJ-EyeTrack	本研究 BCCWJ-EyeTrack-2	
実験協力者数	24 人	26 人（今後拡充予定）	
実験方法	自己ペース読文法と視線走査法 文節単位に空白の挿入を行う・行わない	視線走査法のみ 文節単位に空白の挿入を行わない	
集計方法	テキスト順 FFT, FPT, SPT, RPT, TOTAL	テキスト順 FPT, TOTAL ＋視線停留順	
レジスタ	PN (新聞記事)	PB (書籍)	OT (教科書)
刺激文章数	20 記事	24 サンプル	8 サンプル
最大文字数（行内）	50 文字	50 文字	50 文字
最大行数（画面内）	5 行	9 行	9 行
データポイント数	19716	23454	39130

本節では、BCCWJ-EyeTrack-2 の設計について述べる。表 1 に BCCWJ-EyeTrack との対比を示す。

まず実験協力者は 20 歳以上の日本語母語話者で、26 人分のデータを収集した。本データは今後も拡張予定である。実験方法は BCCWJ-EyeTrack では、自己ペース読文法と視線走査法の 2 条件×文節単位に半角空白を入れるか否かの 2 条件の 4 条件による実験方法であった。BCCWJ-EyeTrack-2 では、より自然な環境のデータを多く収集するために、視線走査法のみを文節単位に空白を入れない環境で実施した。BCCWJ-EyeTrack では、利用者が扱いやすいよう視線停留順ではなくテキスト順に整形したデータを配布した。

テキスト順の集計方法として、文節単位の FFT (First Fixation Time), FPT (First Path Time), SPT (Second Path Time), RPT (Regression Path Time), TOTAL (Total Time) の 5 種類のデータ

* masayu-a@ninjal.ac.jp

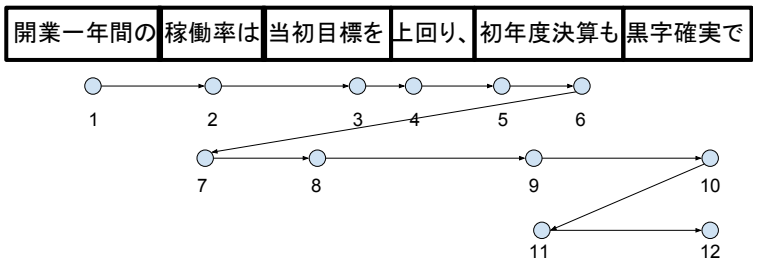

図 1　視線走査順の例

を準備していた。説明のために図 1 の例を用いる。図中 1–12 の数字が視線走査順を表す。FFT はその注視領域に初めて視線が停留した際の注視時間である。例中の「初年度決算も」の FFT は 5 の注視時間となる。FPT は、注視領域に初めて視線が停留し、その後注視領域から出るまでの総注視時間である。出る方向は右方向でも左方向でも構わない。例中の「初年度決算も」の FPT は 5, 6 の注視時間の合計である。SPT は、注視領域に初めて視線が停留し、注視領域から出たあと、2 回目以降に注視領域に停留する総注視時間である。例中の「初年度決算も」の SPT は 9, 11 の注視時間の合計である。尚、FPT と SPT の合計が後に説明する Total Time になる。RPT は、注視領域に初めて視線が停留し、その後領域の右側の境界を超えて次の領域に出るまでの総注視時間である。視線が領域の左側の境界を超えて戻った場合の注視時間も、元の注視領域の RPT として合算する。例中の「初年度決算も」の RPT は 5, 6, 7, 8, 9 の注視時間の合計である。左側に戻り再度注視領域に停留しない場合も合算する。つまり、「初年度決算も」に対する 9 の視線停留がない場合の RPT は 5, 6, 7, 8 の注視時間の合計となる。TOTAL は注視領域に視線が停留する総注視時間である。例中「初年度決算も」の TOTAL は 5, 6, 9, 11 の注視時間の合計である。SPT の扱いについて、研究者コミュニティで定義を共有していないなどの問題があったために、BCCWJ-EyeTrack-2 では、FPT と TOTAL の 2 種類のテキスト順のデータと、視線停留順のデータを準備する。利用者は、自分の定義する集計方法で視点停留順のデータからテキスト順に整形することができる。

図 2 に呈示画面と眼球運動情報例を示す。薄い水色の円が視線停留を示し、大きさがミリ秒を表す。黄土色の矢印が対象物に視線を向けるサッケード運動で、青色の矢印が実際の視線の移動である。赤色の矢印は瞬きを表す。集計は水色の円が示す視線停留に対して行う。

BCCWJ-EyeTrack は、視線走査装置の空間解像度の観点から、行内の最大文字数を 50 文字、最大行数を 5 行にしていたが、BCCWJ-EyeTrack-2 では、行内の最大文字数を 50 文字、最大行数を 9 行とした。図 2 の呈示画面のように、1 行 53 文字までのグリッド（図中黄色の線）を 9 行構築し、この範囲に視線が停留するか否か、グリッド毎に集計する。

視線走査は、急速眼球運動解析装置 EyeLink 1000 PLUS を用いる。時間解像度は 1000Hz であり、1 ミリ秒単位で眼球運動を測定することができる。

呈示文書は、新聞ではなく書籍 (PB) 24 サンプルと教科書 (OT) データ 8 サンプルを用いた。書籍は、アノテーションの優先順位が最上位の A である 25 サンプルから選んだ。残り 1 サンプルについても今後収集する。教科書データは、国語科（高等学校の「古典 A/B」を除く）から、小学校 3 サンプル・中学校 2 サンプル・高等学校 3 サンプルを用いた。呈示文書のその他の統計情報は 3 節の表 2 に示す。

実験協力者は、事前に言語背景情報（生年代、年齢、学歴、言語形成地など）などをアンケートで収

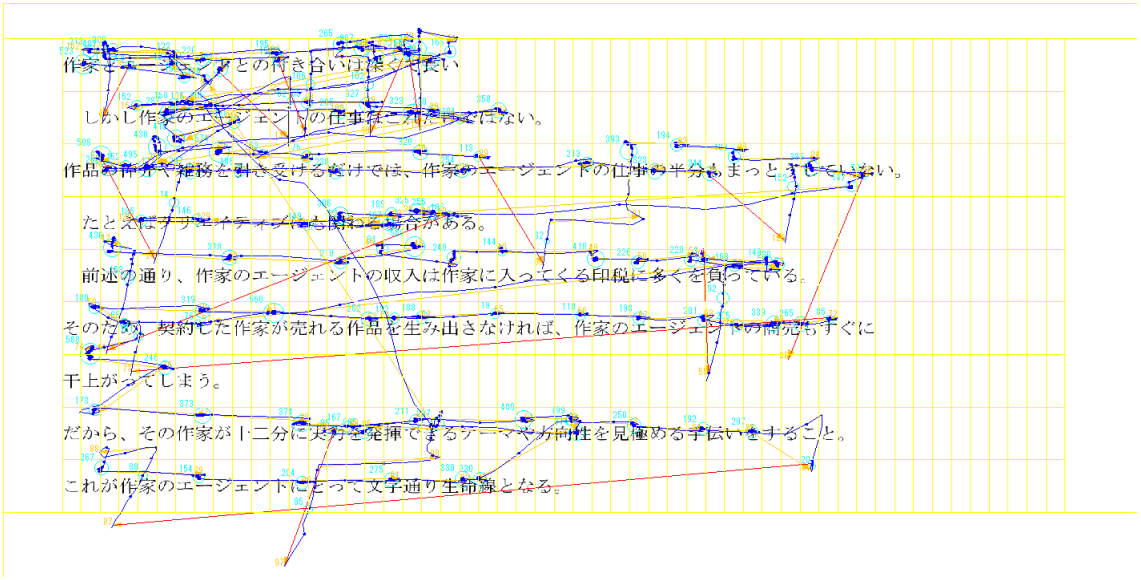

図 2　呈示画面と眼球運動情報例

集するほか、語彙数判定テストと記憶力テストを行った。また、視線走査実験を行うたびに、内容確認のための Yes-No クエスチョンを課題として課した。

全実験協力者のデータポイント数（テキスト順）は、書籍が 23454、教科書が 39130 であった。

3.　基礎統計

本節では、データの基礎統計について示す。

表 2 に呈示文書の基礎統計を示す。サンプル単位の文数は、書籍 (PB) と教科書 (OT) とでほとんど変わらない (PB: 102.6, OT: 98.8) が、文単位の文節数については書籍が教科書よりも多い (PB: 8.46, OT:5.71)。同様にサンプル単位の文節数についても書籍が教科書よりも多い (PB: 868.7, OT:564.8) 傾向にある。文字数においては、文節単位の文字数について、書籍のほうが教科書よりも若干多い (PB: 4.24, OT:3.89) 傾向にある。

表 2　呈示文書の基礎統計

レジスタ (サンプル数)	文数	文数 ／サンプル	文節数	文節数 ／サンプル	文節数 ／文	文字数	文字数 ／サンプル	文字数 ／文	文字数 ／文節数
PB (24)	2463	102.6	20849	868.7	8.46	88512	3688	35.9	4.24
OT (8)	2373	98.8	13556	564.8	5.71	52708	2196	22.2	3.89

表 3 に読み時間の基礎統計を示す。視線停留順は、視線停留ごとに集計している。通常テキストを読む場合、1 回の視線停留は平均約 210-215 ms で、書籍と教科書で差がない。第一四分位が約 143-148ms、第三四分位が約 264-267ms と、±60ms であり、これらを識別して認識するためには相応の時間解像度を必要とする。

テキスト順は、文節単位を注視範囲として集計している。読み時間の平均について、集計方法 FPT と TOTAL ともに、若干書籍のほうが教科書よりも長い (PB(FPT): 260ms, OT(FPT): 236ms, PB(TOTAL): 354ms, OT(TOTAL): 308ms) が、これは文節単位の文字数に比例する注視領域の面積とほぼ比例する。

表 3　読み時間の基礎統計 (単位 ミリ秒)

	Min.	1Q.	Median	Mean	3Q.	Max
PB 視線停留順	2	143	200	210	264	2349
OT 視線停留順	2	148	202	215	267	1610
PB テキスト順 (FPT)	0	0	207	260	348	11824
OT テキスト順 (FPT)	0	0	190	236	323	9620
PB テキスト順 (TOTAL)	0	0	248	354	487	12075
OT テキスト順 (TOTAL)	0	0	217	308	427	12707

4. おわりに

本稿では、書籍と教科書を対象とした読み時間データ BCCWJ-EyeTrack-2 の構築方法と基礎統計について示した。本データは `github.com/masayu-a/BCCWJ-EyeTrack-2` にて公開している。BCCWJ-EyeTrack は新聞記事データであったために、商用には利用できなかった。有償版 BCCWJ を購入している方であれば、商用利用可能である。今後、被験者の数を増やして引き続き拡充していくほか、各種アノテーションを重ね合わせて統計分析を進める。

謝　辞

本研究の一部は情報・システム研究機構の機構間連携・文理融合プロジェクト調査研究 (FS) 『わかりやすい情報伝達の実現に向けた言語認知機構の解明とその工学的応用』によるものです。また、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, JP18K18519 の支援を受けております。

文　献

Masayuki Asahara, Hajime Ono, and Edson 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.

Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den (2014). “Balanced Corpus of Contemporary Written Japanese.” *Language Resources and Evaluation*, 48, pp. 345–371.