『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

浅原正幸　　小野創　　宮本エジソン正
人間文化研究機構　　津田塾大学　　筑波大学
国立国語研究所　　学芸学部　　人文社会系

1　はじめに

我々は、日本人母語話者の読み時間データとして『現代日本語書き言葉均衡コーパス』(以下 BCCWJ)[4] に対して読み時間を付与した BCCWJ-EyeTrack [3] を構築し公開した [1]。同データには、視線走査法による読み時間が付与されている。被験者属性として、生年代（5 年刻み）・年齢（5 歳刻み）・性別・最終学歴・専門分野・視力矯正の有無・言語形成地・父親出身地・母親出身地の情報が付与されているほか、日本語リーディングスパンテスト [6] および語彙数判定テスト [1] の結果が付与されている。

本稿ではこの読み時間データを用いて、被験者属性ごとの読み時間の傾向について分析する。文献 [3] では、被験者をランダム効果として統計処理を行ったが、本稿では、リーディングスパン得点と語彙数テスト結果により被験者属性を代表させて、短期記憶と語彙知識量がそれぞれ読み時間にどのような影響を与えるのかについて検討する。また、以前の分析では、読み時間の実時間を回帰分析を行ったが、今回は対数時間を用いて [2] 統計分析を行う。

2　評価手法

2.1　読み時間の取得方法

読み時間の収集方法として視線走査法を用いる。視線走査法は、被験者がディスプレイ画面上のどの文字を注視しているのかを取得する視線走査装置を用いて、視線注視箇所と注視時間を計測する手法である。自己ペース読文法と異なり、読み戻しなどのより自然な読み時間を取得することができる。視線走査装置として SR Research 社の EyeLink 1000 シリーズ（タワーマウント）を用い、基本的には被験者の右目の情報を取得した。時間解像度は 1000Hz で、ミリ秒単位のデータが収集可能である。

読み時間を収集する対象は、『現代日本語書き言葉均衡コーパス』(Balanced Corpus of Contemporary Written Japanese: BCCWJ)[4] のコアデータの新聞記事データ (PN サンプル) の一部とする。

呈示する基本単位として BCCWJ に付与されている国語研文節単位を用いる。文節境界に半角スペースを空けたものと空けていないものの 2 種類をそれぞれの手法について実施した。実験は新聞記事 21 件を 4 つのデータに分割し、被験者は同じ記事を 2 回読まないような設定で、視線走査法を 2 回行う。

表 1 に呈示したテキスト量について示す。被験者はこのうちのサンプル 2 種について、文節単位にスペースを入れたものと入れていないものを読む。1 記事読むごとに内容確認の Yes/No Question を入れ、内容を理解しているか確認した。正答率は 99% (238/240) であった [3]。

表 1: Data sizes

Sample	文節数	文数	画面数
A	470	66	19
B	455	67	21
C	355	44	16
D	363	41	15

視線走査法で取得したデータは文字の半角単位に Start Fixation Time（注視開始時刻）と End Fixation Time（注視終了時刻）と Fixation Time（注視時間）を得る。このデータを国語研文節単位でグループ化しなおし、注視順データを得る。注視順データを集計して、テキスト生起順データに加工する。テキスト生起順データは以下の 5 種類を国語研文節単位を注視範囲として作成する。

- First Fixation Time (FFT)
- First-Pass Time (FPT)

[1] http://chunagon.ninjal.ac.jp/ より BCCWJ DVD 保持者のみに配布。

[2] 正に歪んだ読み時間が正規分布に近づき、外れ値である可能性がある長い読み時間の影響が小さくなる。

[3] 同様の設定で自己ペース読文法のデータも同じ被験者から収集した。正答率は 78%(187/240) であった

- Regression Path Time (RPT)
- Second-Pass Time (SPT)
- Total Time (TOTAL)

説明のために図 1 の例を用いる。

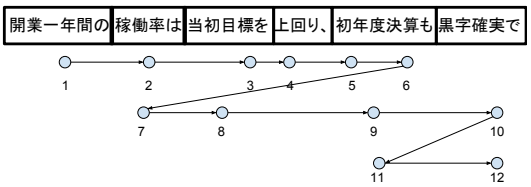

図 1: Example of fixations

First Fixation Time (FFT) は注視範囲に 1 回目に視線が停留した注視時間である。例中の「初年度決算も」の FFT は 5 の注視時間となる。

First-Pass Time (FPT) は、注視範囲に 1 回目に視線が停留し注視範囲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。

Regression Path Time (RPT) は、注視範囲に 1 回目に視線が停留し、注視範囲に再度停留して次に右切片から出るまでの総注視時間である。左側に戻る場合には再度注視範囲に戻るまで合算する。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。

Second-Pass Time (SPT) は、注視範囲に 1 回視線が停留し、注視範囲から出たあと、2 回目以降に注視範囲に停留する総注視時間である。例中の「初年度決算も」の RPT は 9, 11 の注視時間の合計である。尚、FPT + SPT が次に説明する Total Time になる。

Total Time (TOTAL) は注視範囲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 RPT は 5, 6, 9, 11 の注視時間の合計である。

2.2 被験者属性

被験者の語彙数をはかるために、語彙数推定テスト [1] を実施した。語彙数推定テストは心理実験により推定された単語親密度 [5] に基づいて、構成されたものである。50 単語を文字刺激で呈示して、各単語を知っているかどうかをマークシート形式で回答してもらう。知っている単語集合から被験者の語彙数を推定する。評価には Web [4] に公開されている指標を 1000 で割ったものを利用した。

また、被験者の記憶力をはかるために、日本語リーディングスパンテスト [6] を実施した。リーディングスパンテストとは、1 か所だけ下線が引いてある例文を 1 文ずつ被験者に呈示して、音読させるなどしながら下線部を記憶させる。複数文を呈示後に、隠した状態で呈示順に下線が引いてある部分を再生させて正答率をはかることにより、ワーキングメモリ容量を推定するテストである。評価にはオリジナルのスパン得点を用いた。

2.3 データ形式

表 2 にデータ形式について示す。

表 2: Data format

列名	データ型	摘要
surface	factor	出現書字形
length	int	文字数
time	int	読み時間
logtime	num	読み時間 (常用対数)
measure	factor	読み時間の種類
sample	factor	サンプル名
article	factor	記事情報
metadata_orig	factor	文書構造タグ
metadata	factor	メタデータ
space	factor	文節境界空白の有無
subj	factor	被験者 ID
rspan	num	リーディングスパンテスト得点
voc	num	語彙数テスト結果
dependent	int	係り受け関係
setorder	factor	文節境界空白の呈示順
sessionN	int	セッション順
articleN	int	記事呈示順
screenN	int	画面呈示順
lineN	int	行呈示順
segmentN	int	文節呈示順
is_first	factor	最左要素
is_last	factor	最右要素
is_second_last	factor	右から 2 番目の要素

surface は呈示した出現書字形 (文節単位)、length は出現書字形の文字数である。

time は読み時間、logtime は読み時間の常用対数である。measure は読み時間の種類 (FFT など) を表す。

sample は BCCWJ におけるサンプル名、article は記事を一意に決める識別子 [5] である。metadata_orig は BCCWJ に付与された文書構造タグ、metadata は記事タイトルなどの情報を人手で再付与したものである。space は、呈示時に文節単位に半角を入れたか否かの情報である。

[4] http://www.kecl.ntt.co.jp/icl/lirg/resources/goitokusei/goi-test.html

[5] BCCWJ の 1 サンプル中に複数記事含まれるため。

被験者の情報として、被験者の識別子 (subj)・リーディングスパンテスト得点 (rspan)・語彙数テスト結果 (voc) をもつ。

dependent は当該文節に係る文節の数 [2] である。

その他、被験者に対するデータの呈示順の情報として、文節境界空白の呈示順 (setorder)・セッション順 (sessionN)・記事呈示順 (articleN)・画面呈示順 (screenN)・行呈示順 (lineN)・文節呈示順 (segmentN) の情報をもつ。さらに、画面両端の視線のふるまいを特別視するために最左要素 is_first・最右要素 is_last・右から 2 番目の要素 is_second_last の情報を付与する。

2.4 統計分析

データの前処理として、metadata が {authorsData, caption, listItem, profile, titleBlock} のものを除外した。さらに視線走査実験結果の 0 (fixation がない対象) のデータポイントを除外した。この時点でのデータポイント数は SELF が 17628 件、FFT・FPT・RPT・TOTAL が 13232 件、SPT が 4769 件である。

分析は常用対数時間に対して線形混合モデルに基づいて行い、最初に一度モデル化したうえで、標準偏差 ± 3.0 を超えるデータポイントを除外した。article をランダム効果として、次のような式 [6] に基づき分析を行った。

```
logtime ~
  space * sessionN + length + dependent + is_first +
  is_last + is_second_last + articleN + screenN +
  lineN + segmentN + rspan + voc + (1 | article)
```

ていない。

3 結果と考察

FFT(表 3)・FPT(表 4)・RPT(表 5)・SPT(表 6)・TOTAL(表 7) の結果を示す。t value の絶対値が 1.96 以上のものを有意差ありとする。

まず、FFT を除いて、文字数が多い文節ほど注視時間が長くなる傾向が確認された (length)。文節単位に空白を入れたほうが、視線を移動する距離が長くても読み時間が短くなる傾向が確認された (space=TRUE)。係り受けの数 (dependent) が多くなるほど読み時間が短くなる傾向がある。

[6] なお、ランダム切片に対する係数の組み合わせの検討は、モデル選択に時間がかかっており間に合わなかった。今後報告する。

表 3: Parameters of the linear mixed model for the first fixation time (FFT) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.308	0.006	332.7
length	-0.002	0.002	-1.2
space=TRUE	-0.005	0.004	-1.4
rspan	-0.009	0.002	**-4.4**
voc	0.001	0.002	0.7
dependent	-0.009	0.002	**-4.3**
sessionN	-0.022	0.002	**-8.1**
articleN	-0.005	0.004	-1.2
screenN	-0.002	0.002	-0.9
lineN	-0.009	0.002	**-4.3**
segmentN	0.003	0.001	**3.3**
is_first=TRUE	0.020	0.006	**3.3**
is_last=TRUE	-0.008	0.006	-1.3
is_second_last=TRUE	0.000	0.005	0.1
space=TRUE:sessionN	0.045	0.004	**11.0**

表 4: Parameters of the linear mixed model for the first pass time (FPT) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.542	0.009	254.74
length	0.135	0.002	**48.83**
space=TRUE	-0.013	0.005	**-2.51**
rspan	-0.016	0.002	**-6.15**
voc	0.014	0.002	**5.17**
dependent	-0.032	0.002	**-10.96**
sessionN	-0.046	0.003	**-12.98**
articleN	-0.004	0.006	-0.76
screenN	-0.016	0.003	**-4.77**
lineN	-0.016	0.002	**-5.88**
segmentN	-0.003	0.001	**-2.68**
is_first=TRUE	0.091	0.008	**11.27**
is_last=TRUE	0.011	0.008	1.32
is_second_last=TRUE	0.035	0.007	**4.78**
space=TRUE:sessionN	0.064	0.005	**12.26**

次に実験が進むにつれて、全体的に読み時間が短くなる傾向がある (sessionN, screenN, lineN, segmentN)。画面の横方向の両端については、FFT, FPT, TOTAL が is_first と is_second_last で読み時間が長くなる一方、SPT が is_first と is_last で読み時間が短くなること傾向にある。RPT は is_first と is_last と is_second_last で長くなる。

以上は既報 [3] のものと同じ傾向である。

表 8 に、読み時間と被験者属性の関係をまとめたものを示す。表中 “-” は負の有意差 (読み時間が短い) が、“+” は正の有意差 (読み時間が長い) が確認されたことを表し、“0” は有意差がなかったことを表す。

まず、リーディングスパン得点が高い群が、FFT・FPT・RPT の読み時間が短い一方、SPT の読み時間が長い傾向あり、全体の読み時間 (TOTAL) としては有意差がないことがわかった。このことから、短期記

表 5: Parameters of the linear mixed model for the regression path time (RPT) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.610	0.011	234.63
length	0.114	0.003	**34.99**
space=TRUE	-0.015	0.006	**-2.49**
rspan	-0.014	0.003	**-4.59**
voc	0.016	0.003	**5.07**
dependent	-0.026	0.003	**-7.52**
sessionN	-0.055	0.004	**-13.17**
articleN	-0.006	0.007	-0.99
screenN	-0.014	0.004	**-3.63**
lineN	-0.005	0.003	-1.80
segmentN	-0.011	0.001	**-6.93**
is_first=TRUE	0.030	0.009	**3.12**
is_last=TRUE	0.085	0.010	**8.43**
is_second_last=TRUE	0.047	0.008	**5.32**
space=TRUE:sessionN	0.069	0.006	**11.11**

表 6: Parameters of the linear mixed model for the second pass time (SPT) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.483	0.013	177.69
length	0.020	0.004	**4.36**
space=TRUE	-0.047	0.009	**-5.29**
rspan	0.025	0.004	**5.27**
voc	0.033	0.004	**6.82**
dependent	-0.020	0.005	**-3.77**
sessionN	-0.036	0.006	**-5.89**
articleN	-0.002	0.007	-0.34
screenN	-0.013	0.005	**-2.31**
lineN	-0.019	0.004	**-4.10**
segmentN	-0.009	0.002	**-3.98**
is_first=TRUE	-0.028	0.013	**-2.16**
is_last=TRUE	-0.042	0.016	**-2.62**
is_second_last=TRUE	-0.003	0.012	-0.32
space=TRUE:sessionN	0.064	0.009	**6.81**

憶の能力が高い被験者は、1 回の処理速度が速い一方、何度も読み返すために全体としての読み時間が変わらないということがわかった。

また、語彙数テスト結果が高い群は FFT を除いて、読み時間が長い傾向にあることがわかった。

4 おわりに

本研究では、読み時間と被験者属性の対照分析を行った．結果、リーディングスパンテストの成績が良い群が、文節を 1 回読む際の読み時間が短い一方、複数回読む傾向にあり、全体としては成績が劣る群と読み時間に差がないということがわかった。語彙数判定テストの成績が良い群は、First Fixation Time (FFT) 以外のすべての読み時間が長い傾向にあることがわかった。

表 7: Parameters of the linear mixed model for the total time (TOTAL) (logtime)

	Estimate	Std. Error	t value
(Intercept)	2.682	0.010	247.88
length	0.131	0.002	**45.57**
space=TRUE	-0.027	0.005	**-5.14**
rspan	-0.001	0.002	-0.32
voc	0.026	0.002	**9.27**
dependent	-0.034	0.003	**-11.43**
sessionN	-0.055	0.003	**-15.08**
articleN	-0.001	0.007	-0.12
screenN	-0.024	0.003	**-6.83**
lineN	-0.017	0.002	**-6.13**
segmentN	-0.010	0.001	**-7.26**
is_first=TRUE	0.069	0.008	**8.22**
is_last=TRUE	-0.011	0.008	-1.31
is_second_last=TRUE	0.035	0.007	**4.57**
space=TRUE:sessionN	0.074	0.005	**13.60**

表 8: Summary: reading time and attribute of subjects

Fixed Effect	FFT	FPT	SPT	RPT	TOTAL
rspan	-	-	+	-	0
voc	0	+	+	+	+

謝辞

本研究は JSPS 科研費 基盤 (B) 25284083 「言語コーパスに対する読文時間付与とその利用」の助成を受けたものである。

参考文献

[1] S. Amano and T. Kondo. Estimation of mental lexicon size with word familiarity database. In *Proceedings of International Conference on Spoken Language Processing*, Vol. 5, pp. 2119–2122, 1998.

[2] Masayuki Asahara and Yuji Matsumoto. Bccwj-deppara: A syntactic annotation treebank on the 'Balanced Corpus of Contemporary Written Japanese'. In *Proceedings of the 12th Workshop on Asian Langauge Resources (ALR12)*, pp. 49–58, 2016.

[3] 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, 2016.

[4] 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, pp. 345–371, 2014.

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

[6] 苧坂満里子（編）. ワーキングメモリー脳のメモ帳. 新曜社, 2002.