Cognitive Studies, **26**(2), 219-230. (June 2019)

●研究論文●

読み時間と統語・意味分類

浅原 正幸・加藤 祥

This article presents the contrastive analysis between reading time and syntactic/semantic categories in Japanese. We overlaid the reading time annotation BCCWJ-EyeTrack and a syntactic/semantic category information annotation on the 'Balanced Corpus of Contemporary Written Japanese'. Statistical analysis based on a mixed linear model showed that verbal phrase tends to be shorter reading time than adjective, adverbial phrases or nominal phrases in the syntactic categories. Relation phrases are also shorter than other phrases in the semantic categories. The results suggest that the number of preceding phrases associated with the input facilitates the reading process, leading to the reduction in the gaze duration.

Keywords: readability（リーダビリティ）, thesaurus（シソーラス）, corpus（コーパス）, reading time（読み時間）, eye tracking（眼球運動測定）

1. はじめに

従来の文処理研究は，仮説を立てたうえで適切な例文を作成し，作例に対する被験者の読み時間を検証する確認的データ分析 (Confirmatory Data Analysis) により進められてきた．以下では『現代日本語書き言葉均衡コーパス』(Maekawa, Yamazaki, Ogiso, Maruyama, Ogura, Kashino, Koiso, Yamaguchi, Tanaka, & Den, 2014) (以下 BCCWJ) に対する読み時間アノテーション BCCWJ-EyeTrack (Asahara, Ono, & Miyamoto, 2016) と分類語彙表番号アノテーション (加藤・浅原・山崎, 2017) を重ね合わせ，探索的データ分析 (Exploratory Data Analysis) により，統語・意味分類が読み時間に与える影響について検討を試みる．

読み時間アノテーションは，BCCWJ 新聞記事コアデータ 21 記事を刺激として，日本語母語話者 24 人分の読み時間を収集したものである．自己ペース読文法に基づく SELF データと，視線走査法を単語出現順に集計しなおした FFT (First Fixation)・FPT (First-Pass)・SPT (Second-Pass)・RPT (Regresion Path)・TOTAL データの 6 種類からなる．

分類語彙表番号アノテーションは，BCCWJ の短単位と長単位に対して，語義の曖昧性を人手で解消しながら国立国語研究所で整備されている分類語彙表 (国立国語研究所, 1964, 2004) の分類番号を付与したものである．文節単位に写像（文節最右要素もしくは文節に含まれる要素）したうえで分析する．

この 2 つのデータの重ね合わせを行い，被験者と呈示サンプルをランダム効果とした，線形混合モデルによる対照比較を行った．結果，項を取りうる統語分類（用の類：主に動詞）や複数の変数を取りうる意味分類（関係）を表す語の読み時間が短くなる現象が観察され，同一文中には陽に出現しない先行要素が後置要素を予測するモデルを支持する結果が得られた．

以下 2 節では，利用するデータである BCCWJ-EyeTrack と BCCWJ に対する分類語彙表アノテーションについて説明する．3 節では統計分析手法について説明する．4 節では結果と考察を示す．5 節にまとめと今後の研究の方向性について示す．な

Between Reading Time and Syntactic/Semantic Categories, by Masayuki Asahara (National Institute for Japanese Language and Linguistics, Japan) and Sachi Kato (National Institute for Japanese Language and Linguistics, Japan).

表 1 データ形式

列名	データ型	摘要
surface	factor	出現書字形
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
length	int	文字数
space	factor	文節境界空白の有無
subj	factor	実験協力者 ID
setorder	factor	文節境界空白の呈示順
dependent	int	係り受け関係
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 つ目の要素

お，本稿の分析ではベイジアン線形混合モデルを用いるが，参考のために一般化線形混合モデルの結果を付録に示す.

2. 利用するデータ

2.1 BCCWJ

利用するデータは『現代日本語書き言葉均衡コーパス』(BCCWJ) (Maekawa et al., 2014) とそれに対する各種アノテーションである．ここでは BCCWJ について説明する．BCCWJ は，現代日本語の書き言葉を適切なサンプリング手法で集積した均衡コーパスである．BCCWJ のコアデータは人手による形態論情報（短単位・長単位・文節境界）が付与されている.

本研究ではコアデータ中の新聞記事サンプル (PN_core) を用いる．新聞記事サンプルの一部には，以下に述べる読み時間の情報アノテーション (BCCWJ-EyeTrack) と分類語彙表番号アノテーションが付与されている.

2.2 BCCWJ-EyeTrack

BCCWJ-EyeTrack (Asahara et al., 2016)(表 1) は，BCCWJ の新聞記事サンプルに自己ペース読文法と視線走査法により，日本語母語話者 24 人分の読み時間を付与したものである．以下，データの詳細について説明する.

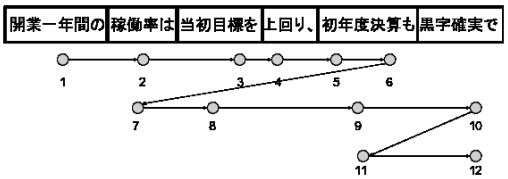

図 1 読み時間の集計方法

自己ペース読文法は，他の文節をマスクしたうえで 1 文節単位を逐次的に呈示する読み時間測定手法である．読み戻しができないため，文節単位の読み時間がそのままデータとなる．このデータを SELF と呼ぶ.

視線走査法で取得したオリジナルのデータから文字の半角単位に Start Fixation Time（注視開始時刻）と End Fixation Time（注視終了時刻）と Fixation Duration（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 種類からなる.

- First Fixation Time (FFT)
- First-Pass Time (FPT)
- Second-Pass Time (SPT)
- Regression Path Time (RPT)
- Total Time (TOTAL)

図 1 の例を用いて説明する.

First Fixation Time (FFT) は，注視範囲に視線がはじめて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，注視範囲に視線がはじめて停留し，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.

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 SPT は 9, 11 の注視時間の合計である．なお，FPT+SPT が以下で説明する Total Time になる.

Regression Path Time (RPT) は，注視範囲に視線が 1 回目に停留し，注視範囲に再度停留して次に

表 2 分類番号の構造「この」(分類番号: 3.1010)

類	部門	中項目	分類項目
相 (3)	関係 (.1)	真偽 (.10)	こそあど (.1010)

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．左側に戻っても，再度注視範囲に停留しない場合は合算しない．例中「上回り,」の RPT は 4 の注視時間である．

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 TOTAL は 5, 6, 9, 11 の注視時間の合計である．

これらの読み時間情報 (`time`, `logtime`) に対して，出現書字形 (`surface`)・記事情報 (`sample`, `article`)・文書構造 (`metadata_orig`, `metadata`) のほか，出現書字形文字数 (`length`)，文節単位の空白の有無 (`space`)，実験協力者 ID (`subj`)，係る文節数 (`dependent`)，実験協力者ごとの呈示順序 (`sessionN`, `setorder`, `articleN`, `screenN`, `lineN`, `segmentN`)，画面水平方向の位置 (`is_first`, `is_last`, `is_second_first`) を付与したデータを分析に用いる．係る文節数は BCCWJ-DepPara (Asahara & Matsumoto, 2016) のものを用いた．

2.3 BCCWJ に対する分類語彙表番号アノテーション

『分類語彙表』(国立国語研究所, 1964) は「語を意味によって分類・整理したシソーラス（類義語集）」である．初版はおよそ 33,000 語を収録していたが,『分類語彙表–増補改訂版–』(国立国語研究所, 2004) は区切り文字を含めて 101,070 件からなる．本研究では分類語彙表増補改訂版の CSV データ[1)]を用いる．

『分類語彙表』は表 2 に示す分類番号を用いて，単語の分類項目の体系的位置づけを行う．分類番号は，1–4 の最初の 1 桁が「類」と呼ばれ，品詞（統語分類）を表す．1 が名詞の仲間である体の類を，2 が動詞の仲間である用の類を，3 が形容詞・形容動詞・副詞・連体詞などの仲間である相の類を，4 が接続詞・感動詞などのその他の類を表す．ピリオドをはさんで 4 桁からなる数値が意味分類を表す．意味分類のうち 1 桁目は「部門」と呼ばれ，.1 が抽象的関係（関係）を，.2 が人間活動の主体（主体）を，.3 が精神および行為（活動）を，.4 が生産物および用具（生産物）を，.5 が自然物および自然現象（自然）を表す．また，意味分類のうち 2 桁目までを「中項目」と呼び，4 桁目までを「分類項目」と呼ぶ．さらに，増補改訂版より分類項目の下位分類として,「段落」が定義されている．

BCCWJ のコアデータの一部に対して『分類語彙表』の分類番号を付与する作業が進められている (加藤他, 2017)．分類語彙表を人手で UniDic の語彙素番号に対応させたデータ (近藤・田中, 2017) により同データの短単位と長単位の両方について，可能な分類番号を枚挙し，人手で語義の曖昧性解消を行うとともに，未定義の部分に追加して分類番号を付与する．

本分析には BCCWJ に対する分類語彙表番号アノテーションデータの長単位データに基づき，文節内最右自立語の分類番号を分析対象とする．統語分類として「類」(wlsp_syn と呼ぶ）を用い，意味分類として「部門」(wlsp_sem と呼ぶ）を用いて，統計分析を行う．

3. 統計処理

読み時間のモデリング手法として階層ベイズモデル (Bayesian Linear Mixed Model) (Sorensen, Hohenstein, & Vasishth, 2016) を用いる．言語研究でよく用いられる被験者と呈示サンプルなど 2 つ以上のランダム因子を含むようなモデルにおいては，最尤推定に基づく頻度主義的な手法（線形混合モデル）では適合させたうえで収束させることが難しい．階層ベイズモデルでは，尤度に比例する確率分布からのランダムサンプリングを行うことで，より直接的にパラメータの確率分布を推定できる．さらに頻度主義的な一般化線形混合モデルと異なり，帰無仮説をパラメータの全ての組み合わせについて有意差を検討する必要もない．

分析対象は，記事中の本文（タイトル以外の部分）に出現する文節のみとする．具体的には，新聞記事上のレイアウト情報 `metadata` が authorsData（著者情報), caption（キャプション), listItem（箇条書き), profile（著者背景情報), titleBlock（タイトル）のものを削除した．基本的には，新聞紙面の割付における本文のみを分析対象とする．詳しい `metadata` のラベルの意味については BCCWJ の

1) `https://pj.ninjal.ac.jp/corpus_center/archive.html#bunruidb`

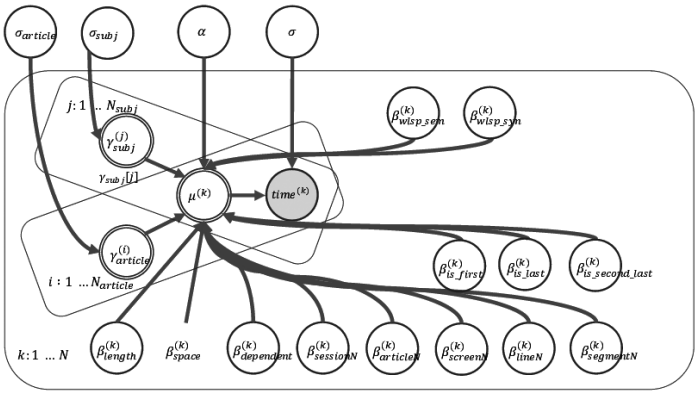

図 2 ベイジアン線形混合モデルのグラフィカルモデル

マニュアルを参照されたい．モデリングは，視線走査法 (FFT, FPT, RPT, TOTAL) の 4 種類の指標について行った．

図 2 に階層ベイズモデルのグラフィカルモデルを示す．プレート k がデータポイントに対するイテレーションを表し，プレート i がランダム因子である記事に対するイテレーション，プレート j がランダム因子である被験者に対するイテレーションである．

$\texttt{time}^{(k)}$ を，データポイント $k \in 1, \ldots, N$ の読み時間とし，(Rouder, 2005) にならい，対数正規分布によりモデル化する：

$$time^{(k)} \sim Lognormal(\mu^k, \sigma), \qquad (1)$$

(1) 式で σ が対数正規分布の分散，μ^k が次の線形式で表現される平均を表す：

$$\begin{aligned}\mu^k = \alpha &+ \beta^{(k)}_{\texttt{length}} + \beta^{(k)}_{\texttt{space}} + \beta^{(k)}_{\texttt{dependent}} + \beta^{(k)}_{\texttt{sessionN}} \\ &+ \beta^{(k)}_{\texttt{articleN}} + \beta^{(k)}_{\texttt{screenN}} + \beta^{(k)}_{\texttt{lineN}} + \beta^{(k)}_{\texttt{segmentN}} \\ &+ \beta^{(k)}_{\texttt{is\_first}} + \beta^{(k)}_{\texttt{is\_last}} + \beta^{(k)}_{\texttt{is\_second\_last}} \\ &+ \beta^{(k)}_{\texttt{wlsp\_syn\_*}} + \beta^{(k)}_{\texttt{wlsp\_sem\_*}} + \beta^{(k)}_{\texttt{is\_second\_last}} \\ &+ \gamma^{(i)}_{\texttt{article}} + \gamma^{(j)}_{\texttt{subj}} \qquad (2)\end{aligned}$$

(2) 式で，α は線形式の切片，$\beta^{(k)}_f$ はデータポイント k に対する固定因子 $f \in \{\texttt{length}, \texttt{space}, \texttt{dependent}, \texttt{sessionN}, \texttt{articleN}, \texttt{screenN}, \texttt{lineN}, \texttt{segmentN}, \texttt{is\_first}, \texttt{is\_last}, \texttt{is\_second\_last}, \texttt{wlsp\_syn\_*}, \texttt{wlsp\_sem\_*}\}$ の傾きを表す．$\gamma^{(i)}_{\texttt{article}}$ はランダム因子である記事 $i \in 1, \ldots, N_{\texttt{article}}$ の事前分布，$\gamma^{(j)}_{\texttt{subj}}$ はランダム因子である被験者 $j \in 1, \ldots, N_{\texttt{subj}}$ の事前分布であり，次の (3), (4) 式のように定義する：

$$\gamma^{(i)}_{\texttt{article}} \sim Normal(0, \sigma_{\texttt{article}}), \qquad (3)$$

$$\gamma^{(j)}_{\texttt{subj}} \sim Normal(0, \sigma_{\texttt{subj}}). \qquad (4)$$

(3), (4) 式で定義する正規分布の平均を 0 とする．また正規分布の分散をハイパーパラメータ $\sigma_{\texttt{article}}, \sigma_{\texttt{subj}}$ として推定する．

以下，各固定因子 $f \in \{\texttt{length}, \texttt{space}, \texttt{dependent}, \texttt{sessionN}, \texttt{articleN}, \texttt{screenN}, \texttt{lineN}, \texttt{segmentN}, \texttt{is\_first}, \texttt{is\_last}, \texttt{is\_second\_last}\}$ の意味について説明する．

length は，呈示している文節の文字長であり，視線が停留する面積に相当する．space は，呈示時に文節間に半角空白を入れたか否かを表し，半角空白の挿入が読み時間にどのような影響を与えるかを調査する．dependent は，当該文節に係る文節の数であり，anti-locality 現象を調べる固定因子である．1 文が複数行にわたって呈示する場合は，行を越えて係る構造を許して数える．sessionN, articleN, screenN, lineN, segmentN は呈示順であり，実験が進むにつれて被験者が慣れてくる影響を調査する．is_first，is_last，is_second_last は，1 行中の最左要素，最右要素，右から 2 番目の要素を意味し，画面上のレイアウトによる影響を調査する．最後に，wlsp_syn_* が統語分類「類」を表すカテゴリデータ，wlsp_sem_* が意味分類「部門」を表すカテゴリデータであり，本研究の分析対象である．

視線走査法の読み時間のデータポイントのうち，

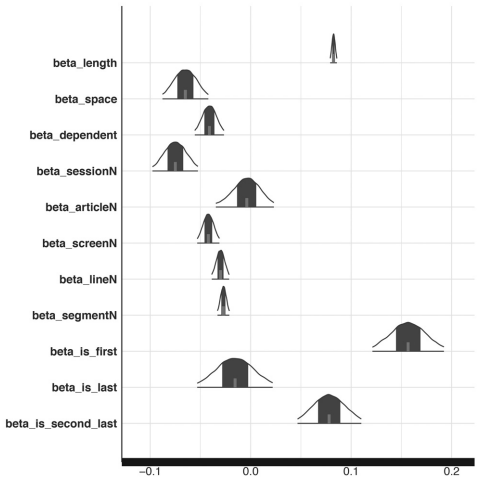

図 3 Total Time の分類語彙表番号以外の要素の係数

ゼロ秒のものは視線が停留していないということで分析データから排除した．Clifton (Clifton, Staub, & Rayner, 2007) のように SPT についてはゼロ秒を排除しない研究者が存在する．彼らの手法では，本来欠損値として扱うべきゼロ秒を 0 の値を割り当てる処理を行っている．一方，計測された読み時間については，半正定値ではなく正定値を取る前提に基づき，対数読み時間に対して推定する慣習もあり，近年では対数読み時間を評価することが一般的である (Fossum & Levy, 2012; Luong, O'Donnell, & Goodman, 2015)．対数読み時間を利用すると，モデル化する際に正定値が担保されるだけでなく，より正規分布に適合し，外れ値の影響が小さくなるという利点がある (Gelman & Hill, 2006) が，ゼロ秒を考慮することはできない．今回利用するベイジアン線形混合モデルも対数正規分布に基づく分析 (Sorensen et al., 2016) を行うために，視線走査法の場合にゼロ秒を排除して分析を行った．しかしながら，この扱いについて研究者コミュニティで合意できていないという指摘があったために，本稿ではゼロ秒を含む SPT の結果は省略する．

モデリングには Rstan を用いる．すべてのモデルについて，Rhat の値が 1.1 未満であり，収束していることを確認した．

表 3 Total Time の分類語彙表番号以外の要素の係数

Parameter	mean	sd	se_mean
α	5.715	0.143	0.002
β_length	0.082	0.002	0.000
β_space	−0.065	0.012	0.000
β_dependent	−0.041	0.007	0.000
β_sessionN	−0.075	0.012	0.000
β_articleN	−0.004	0.015	0.000
β_screenN	−0.042	0.006	0.000
β_lineN	−0.030	0.004	0.000
β_segmentN	−0.027	0.003	0.000
β_is_first	0.157	0.018	0.000
β_is_last	−0.015	0.019	0.000
β_is_second_last	0.078	0.016	0.000

4. 結果と考察

4.1 結　果

まず，分類語彙表番号以外の情報について推定した結果を，Total Time のみについて表 3 と図 3 に示す．表中 mean が事後確率平均，sd が事後標準偏差，se_mean が標準誤差である．図は，95%信用区間のカーネル密度推定を示す．表中 mean の差と sd の関係から，0.0 との差，もしくは，2 因子の差を検討する．図においては，0.0 との位置から各係数が読み時間を促進するのか（負の値），抑制するのか（正の値）を判断する．

一般に，文節長 (beta_length) が長くなればなるほど読み時間が長くなる．これは，文節の長さに応じて，表示面積が大きくなり，視線が停留する確率が線形に高くなるためだと考える．視線走査法においては，空白ありのほうが読み時間が短くなる (beta_space)．単純に読み時間を短くするという観点でレジビリティを上げるには，文節間に空白を入れたほうがよい．係り受けでは，多くの係り受けがある文節ほど読み時間が短くなる (beta_dependent)．これは後に述べる Anti-locality 現象 (Konieczny, 2000) の追認である．呈示順 (sessionN, articleN, screenN, lineN, segmentN) は，基本的に進めば進むほど読み時間が短くなる．これは実験協力者が実験に慣れてきた効果であると考える．レイアウト情報 (is_first, is_last, is_second_last) は，折り返しの視線移動に基づく影響を勘案するものである．最左要素 (is_first) と右から 2 番目の要素 (is_second_last) で

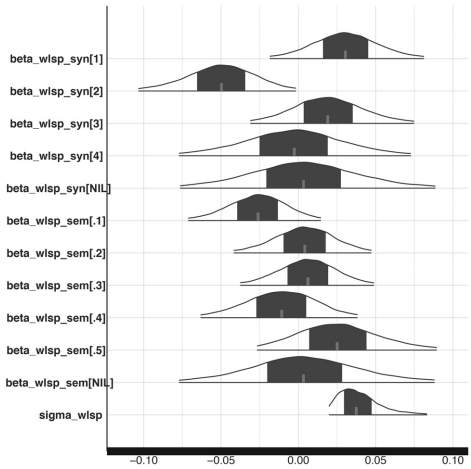

図 4 First Fixation Time (FFT) の係数

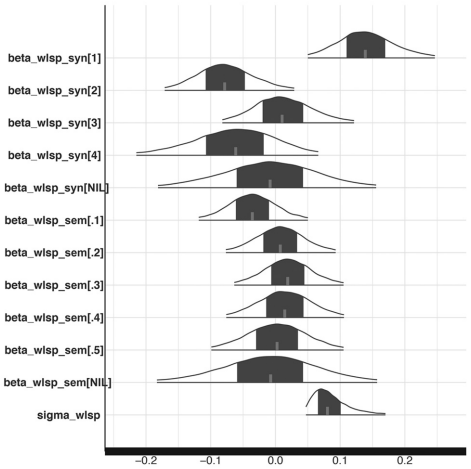

図 5 First Pass Time (FPT) の係数

表 4 First Fixation Time (FFT) の係数

Parameter	mean	sd	se_mean
β_syn[1] (体)	0.032	0.042	0.001
β_syn[2] (用)	−0.049	0.042	0.001
β_syn[3] (相)	0.021	0.043	0.001
β_syn[4] (他)	−0.002	0.049	0.001
β_syn[NIL]	0.006	0.066	0.002
β_sem[.1] (関係)	−0.026	0.028	0.001
β_sem[.2] (主体)	0.004	0.029	0.001
β_sem[.3] (活動)	0.007	0.028	0.001
β_sem[.4] (生産物)	−0.011	0.031	0.001
β_sem[.5] (自然物)	0.027	0.035	0.001
β_sem[NIL]	0.005	0.052	0.001

表 5 First-Pass Time (FPT) の係数

Parameter	mean	sd	se_mean
β_syn[1] (体)	0.141	0.050	0.001
β_syn[2] (用)	−0.076	0.050	0.001
β_syn[3] (相)	0.013	0.051	0.001
β_syn[4] (他)	−0.065	0.071	0.001
β_syn[NIL]	−0.009	0.084	0.001
β_sem[1] (関係)	−0.035	0.042	0.000
β_sem[2] (主体)	0.008	0.043	0.000
β_sem[3] (活動)	0.019	0.043	0.000
β_sem[4] (生産物)	0.015	0.046	0.000
β_sem[5] (自然物)	0.003	0.051	0.000
β_sem[NIL]	−0.009	0.084	0.001

は読み時間が長くなる傾向にある．これらは (Asahara et al., 2016) と同様の結果である．

次に First Fixation Time の結果を図 4 と表 4 に示す．統語分類においては，体の類 [1] と用の類 [2] の差が大きい (0.081)．意味分類においては，関係の部門 [.1] は主体の部門 [.2]・活動の部門 [.3]・自然物の部門 [.5] と比して，読みを促進する傾向にある (mean の差: 0.030-0.053, sd: 0.028-0.035)．

図 5 と表 5 に First Pass Time の結果を示す．統語分類においては，体の類 [1] は 0.0 と比して強く読みを抑制する傾向があることがわかる (mean: 0.141, sd: 0.050)．用の類 [2] は 0.0 と比して読みを促進する傾向がある (mean: −0.076, sd: 0.050)．意味分類においては，関係の部門 [.1] は，主体の部門 [.2]・活動の部門 [.3]・生産物の部門 [.4] と比して，読みを促進する傾向にある (mean の差: 0.043-0.053, sd: 0.042-0.046)．

図 6 と表 6 に Regression Path Time の結果を示す．統語分類においては，体の類 [1] は 0.0 と比して読みを強く抑制する傾向がある (mean: 0.132, sd: 0.051)．用の類 [2] は 0.0 と比して読みを促進する傾向がある (mean: −0.065, sd: 0.051)．意味分類においては，関係の部門 [.1] は活動の部門 [.3]・生産物の部門 [.4]・自然物の部門 [.5] と比して読みを若干促進する傾向にある (mean の差: 0.041-0.048, sd: 0.042-0.054)．

図 7 と表 7 に Total Time の結果を示す．統語分類においては，体の類 [1] は 0.0 と比して読みを強

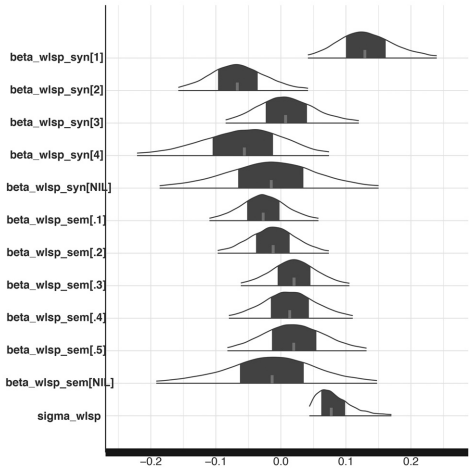

図 6 Regression Path Time (RPT) の係数

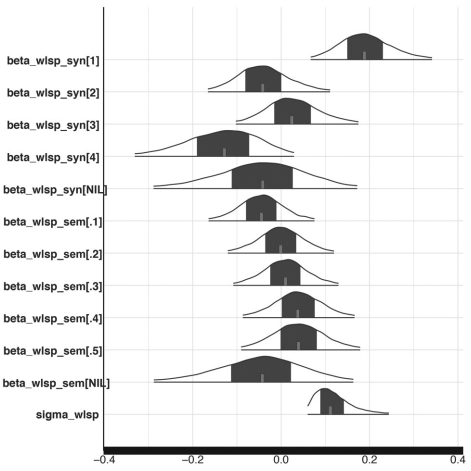

図 7 Total Time (TOTAL) の係数

表 6 Regression Path Time (RPT) の係数

Parameter	mean	sd	se_mean
β_syn[1] (体)	0.132	0.051	0.001
β_syn[2] (用)	−0.065	0.051	0.001
β_syn[3] (相)	0.010	0.053	0.001
β_syn[4] (他)	−0.061	0.074	0.001
β_syn[NIL]	−0.016	0.084	0.001
β_sem[.1] (関係)	−0.027	0.042	0.000
β_sem[.2] (主体)	−0.012	0.043	0.000
β_sem[.3] (活動)	0.020	0.042	0.000
β_sem[.4] (生産物)	0.014	0.048	0.000
β_sem[.5] (自然物)	0.021	0.054	0.000
β_sem[NIL]	−0.015	0.084	0.001

表 7 Total Time (Total) の係数

Parameter	mean	sd	se_mean
β_syn[1] (体)	0.193	0.072	0.001
β_syn[2] (用)	−0.038	0.072	0.001
β_syn[3] (相)	0.028	0.073	0.001
β_syn[4] (他)	−0.134	0.093	0.001
β_syn[NIL]	−0.044	0.141	0.002
β_sem[.1] (関係)	−0.044	0.067	0.001
β_sem[.2] (主体)	−0.000	0.068	0.001
β_sem[.3] (活動)	0.011	0.067	0.001
β_sem[.4] (生産物)	0.039	0.071	0.001
β_sem[.5] (自然物)	0.041	0.075	0.001
β_sem[NIL]	−0.048	0.118	0.001

く抑制する傾向がある (mean: 0.193, sd: 0.072). 他の類 [4] は 0.0 と比して読みを促進する傾向がある (mean: −0.134, sd: 0.093). 意味分類においては，関係の部門 [.1] は生産物の部門 [.4] や自然物の部門 [.5] と比して読みを促進する傾向にある (mean の差: 0.083-0.085, sd: 0.065-0.075).

統語分類の係数を 0.0 との比較によりまとめると表 8 のようになる．表中 ++ は mean が 2sd 以上大きいことを意味し，+ は mean が 1sd 以上大きいことを意味する．また − は mean が 1sd 以上小さいことを意味する．読み時間にもよるが体の類が読み時間が長くなり，用の類が読み時間が短くなる傾向にある．

意味分類の係数を関係 [.1] と他のラベルとの比較によりまとめると表 9 のようになる．表中 + は mean の差が 1sd 以上大きいことを意味する．読み時間にもよるが，関係 [.1] の読み時間が，他のラベルよりも短い傾向がわかる．

表 8 統語分類の係数 (0.0 との比較)

類	FFT	FPT	RPT	TOTAL
体 [1]	0	++	++	++
用 [2]	−	−	−	0
相 [3]	0	0	0	0
他 [4]	0	0	0	−

表 9 意味分類の係数 (関係 [.1] との比較)

部門	FFT	FPT	RPT	TOTAL
主体 [.2]	+	+	0	0
活動 [.3]	+	+	+	0
生産物 [.4]	0	+	0	+
自然物 [.5]	+	0	+	+

4.2 考 察

Anti-locality は，先行文脈に係り元文節（単語）が多い要素ほど読み時間が短くなるという現象であり主に二重目的語構文における動詞述語や埋め込み節の入れ子の読み時間について報告されてきた（ドイツ語 (Konieczny, 2000; Konieczny & Döring, 2003; Levy & Keller, 2013)，日本語 (Uchida, Miyamoto, Hirose, Kobayashi, & Ito, 2014)，ヒンディー語 (Vasishth & Lewis, 2006; Husain, Vasishth, & Srinivasan, 2015)）．このような読み時間の短縮は，主辞後置言語において，係り元が多い要素を読むのに負荷がかかるという予測 (Gibson, 2008) や，後続する主辞の処理コストは先行する係り受けや同一指示の影響を受けないという予測 (Nakatani & Gibson, 2010) などの，ワーキングメモリモデルによって説明できない現象であった．直接目的語と間接目的語の二つの係り元要素が先行する動詞述語のほうが，直接目的語のみ係り元要素が先行する動詞述語より読み時間が短い．しかし，この結果は anti-locality を示す必要条件であるが，十分条件ではない．

Asahara, Ono & Miyamoto (2016) では，均衡コーパスと係り受けアノテーションを用い，より一般化した設定で anti-locality 現象を調査した．BCCWJ-DepPara (Asahara & Matsumoto, 2016) と読み時間データの重ね合わせから，係る文節数が多い文節ほど読み時間が短くなることを報告した．ここで，文節係り受け関係は，ガ格・ヲ格・ニ格などの主語・目的語要素と述語間の関係だけでなく，デ格・カラ格ほかの格助詞と述語の関係，ハ・モなどの係助詞と述語の関係，名詞句に対する連体修飾関係，副詞などの用言に対する連用修飾関係なども含む．既存の研究においては，二重目的語構文などに限定して分析が進められていたが，より一般化した設定による分析がなされている．

本研究では，統語分類において，体の類＞相の類＞用の類の順で読み時間が短くなる傾向が確認された．体の類は一般に「モノ」などを表す名詞の仲間で，動詞や形容詞などの述語の項になりうる一方，修飾詞や項を取りうる句は限定的である．また判定詞などを取って名詞述語文を構成するが，相の類は形容詞・形容動詞・副詞・連体詞の仲間で，ガ格を取り述語になるものと修飾詞にのみなりうるものがある．用の類は動詞の仲間で，一般に節末などに出現し，項を取ることが多い．

係り受けと統語分類に関連する分析結果は，先行要素記憶の負荷に基づくワーキングメモリモデルよりも，先行要素が後置要素を予測するモデルのほうが妥当であることを示唆している．言い換えると，修飾詞になったり，項を取ったりする用の類のほうが，予測されやすいということを反映している．係り受けを重ねたうえでも，この差が出ているのは，日本語において項が省略されていることによるものと考える．

心理言語実験で従来行われてきた単文に基づく実験では，全ての可能な項（ガ・ヲ・ニ・デほか）のバリエーションを統制して分析することが多い．しかしながら，日本語においては，ガ格も含めて省略可能であり，本実験のような自然なテキストに対する分析においては，先行文脈で言及されている項名詞句が省略されるゼロ代名詞が多く出現する．ここでいう項は，二重目的格のみならず，斜格-述語間関係や連体修飾・連用修飾のようなものを含む．

また，意味分類においては，関係 [.1] が他の要素よりも読み時間の短縮される傾向がみられた．関係 [.1] は 2 変数を取りうる要素でありかつ，その変数が先行文脈で言及されている場合には同一文中にあまり出現しないことを反映している．

5. おわりに

本研究では，テキストの読み時間の傾向を分析するために，均衡コーパスに対する読み時間情報と分類語彙表番号アノテーションの対照比較を行った．結果，統語分類に対しては，体の類＞相の類＞用の類の順に読み時間が短くなる傾向がみられた．また意味分類（部門）に対しては，抽象的関係が他の意味分類より，読み時間が長くなる傾向がみられた．

日本語の述語項構造関係は，既存の作例による実験においては，不自然にガ格・ヲ格・ニ格が埋まっていることが多い．しかしながら，『現代日本語書き

言葉均衡コーパス』(BCCWJ) コアデータ 57,225 文のうち，ガ格・ヲ格・ニ格が全てそろった例は 584 例しか見られない．日本語において，ガ格・ヲ格・ニ格全てが必須格ではなく，先行文脈で対象が言及されており文脈から当該名詞句が項になることがわかっている場合には，陽に (overtly) テキスト中に表出せず，ゼロ代名詞として省略される傾向にある．このゼロ代名詞は「係る文節数」として係数されないために，用の類が項が多く，体の類が項が少ないという統語分類が持つ非顕在項の多寡が読み時間に影響を与えることが考えられる．Asahara (2018b) は，この点を解決するためにゼロ代名詞の影響に関する分析を行っている．

また，意味分類においても，主体・生産物・自然のようにあるモノの集合（クラス）の中の 1 インスタンスという 1 変数関数や，関係・活動のように複数のモノ（もしくはコト）をもちうる 2 変数以上の関数が考えられる．この意味分類の変数も，先行文脈から自明な場合には陽にテキスト中に表出しない．特に活動は述語項でテキスト中に表現される一方，関係についてはテキスト中にあまり表現されない傾向がある．このことが関係の読み時間を早くする原因となっていると考える．

Asahara (2018a) では，読み時間と節境界アノテーション (Matsumoto, Asahara, & Arita, 2018) の対照を行い，節末で読み時間が短くなる傾向を報告している．一般に，項や修飾詞をとる用の類は節末に出現するため，今回の結果と親和性がある．浅原 (2018b) では，読み時間と情報構造アノテーション (Miyauchi, Asahara, Nakagawa, & Kato, 2017) の対照比較を行い，ブリッジングなどにより想定可能情報や共有情報と比べて，情報の受容者側にとっての非共有情報の読み時間が長くなることを報告している．項や修飾詞を取りうる句は想定可能情報や共有情報になりやすいことと相関があると考える．

最後に日本語の読み時間の分析において，コーパス中の頻度に基づく分析の困難さについて述べる．本研究のように日本語の読み時間の分析においては，文節単位の分析が一般的であり，自己ペース読文法などの呈示単位も文節単位がほとんどである．一方，日本語の文節の表層文字列の頻度を直接計算し，適切にモデル化する規模のコーパスは存在しない．例えば，BCCWJ EyeTrack には「池田弘子先生 (75) は」「公営地下鉄二十六路線の」「全国同和食肉事業協同組合連合会（全同食）が」などの文節が出現するが，これらは 258 億語規模の『国語研日本語ウェブコーパス』をもってしても，この文字列がそのままコーパス中に出現することはない．一方，コーパス言語学においては斉一な単位として，より短い単位（例えば「国語研短単位」）により頻度を計数することが多い．これらの頻度の合計や平均を用いることも考えられるが，文字列の包含関係をどのように計算するかという問題がある．実際に，短単位に基づく頻度の合計と平均（の対数値）によるモデル化を検討したところ，よい結果が得られなかった．この問題を解決するため，単語埋め込み (word embeddings) の加法構成性を用いたモデルが提案されている (浅原, 2018a).

文 献

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

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

Asahara, M., & Matsumoto, Y. (2016). BCCWJ-DepPara: A Syntactic Annotation Treebank on the 'Balanced Corpus of Contemporary Written Japanese'. *Proceedings of the 12th Workshop on Asian Langauge Resources (ALR12)*, 49–58.

Asahara, M., Ono, H., & Miyamoto, E. T. (2016). Reading-Time Annotations for 'Balanced Corpus of Contemporary Written Japanese'. *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, 684–694.

Asahara, M., & Matsumoto, Y. (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)*, 49–58.

浅原 正幸 (2018a). 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』, 82–87.

浅原 正幸 (2018b). 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), 527–554.

Baayen, R. H. (2008). *Analyzing Linguistic Data: A practical Introduction to Statistics using R*. Cambridge University Press.

Clifton, C. J., Staub, A., & Rayner, K. (2007). *Eye movements: A window on mind and brain*, chap. Eye movements in reading words and sentences, 341–372. Amsterdam: Elsevier.

Fossum, V., & Levy, R. (2012). Sequential vs. hierarchical syntactic models of human incremental sentence processing. *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics*.

Gelman, A., & Hill, J. (2006). *Data Analysis Using Regression and Multilevel/Hierarchical Models*. Cambridge University Press.

Gibson, E. (2008). Linguistic complexity: Locality of syntactic dependencies. *Cognition*, **68**, 1–76.

Husain, S., Vasishth, S., & Srinivasan, N. (2015). Integration and prediction difficulty in Hndi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research*, **8** (2).

加藤 祥・浅原 正幸・山崎 誠 (2017). 『現代日本語書き言葉均衡コーパス』に対する分類語彙表番号アノテーション. 『第 23 回言語処理学会年次大会発表論文集』, 306–309.

近藤 明日子・田中 牧郎 (2017). 分類語彙表・UniDic 見出し対応表の構築 — コーパスへの網羅的・系統的な語義情報付与を目指して —. 『第 23 回言語処理学会年次大会発表論文集』, 90–93.

国立国語研究所 (編) (1964). 『分類語彙表』. 秀英出版.

国立国語研究所 (編) (2004). 『分類語彙表 –増補改訂版–』. 大日本図書.

Konieczny, L. (2000). Locality and Parsing Complexity. *Journal of Psycholinguistic Research*, **29** (6), 627–645.

Konieczny, L., & Döring, P. (2003). Anticipation of clause-final heads. Evidence from eye-tracking and SRNs. *Proceedings of the 4th International Conference on Cognitive Science*.

Levy, R., & Keller, F. (2013). Expectation and locality effects in German verb-final structures. *Journal of Memory and Language*, **68** (2), 199–202.

Luong, M.-T., O'Donnell, T. J., & Goodman, N. D. (2015). Evaluating Models of Computation and Storage in Human Sentence Processing. *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*, 14–21.

Maekawa, K., Yamazaki, M., Ogiso, T., Maruyama, T., Ogura, H., Kashino, W., Koiso, H., Yamaguchi, M., Tanaka, M., & Den, Y. (2014). Balanced Corpus of Contemporary Written Japanese. *Language Resources and Evaluation*, **48**, 345–371.

Matsumoto, S., Asahara, M., & Arita, S. (2018). Japanese clause classification annotation on 'Balanced Corpus of Contemporary Written Japanese'. *Proceedings of the 13th Workshop on Asian Language Resources*, 1–8.

Miyauchi, T., Asahara, M., Nakagawa, N., & Kato, S. (2017). Information-Structure annotation of the "Balanced Corpus of Contemporary Written Japanese". *Proceedings of of the 15th International Conference of the Pacific Association for Computational Linguistics (PACLING 2017)*, 166–175.

Nakatani, K., & Gibson, E. (2010). An on-line study of Japanese nesting complexity. *Cognitive Science*, **34** (1), 94–112.

Rouder, J. N. (2005). Are unshifted distributional models appropriate for response time. *Psychometrika*, **70**, 377–381.

Sorensen, T., Hohenstein, S., & Vasishth, S. (2016). Bayesian linear mixed models using Stan: a tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology*, **12**, 175–200.

Uchida, S., Miyamoto, E. T., Hirose, Y., Kobayashi, Y., & Ito, T. (2014). An ERP Study of Parsing and Memory Load in Japanese Sentence Processing — A Comparison Between Left-Corner Parsing and the Dependency Locality Theory —. *Proceedings of the Thought and Language/ the Mental Architechture of Processing and Learning of Language 2014*.

Vasishth, S., & Lewis, R. L. (2006). Argument-head distance and processing complexity: explaining both locality and antilocality effects. *Language*, **82** (4), 767–794.

(Received 25 Feb. 2018)
(Accepted 7 Jan. 2019)

浅原 正幸
2003 年奈良先端科学技術大学院大学情報科学研究博士後期課程修了．2004 年より同大学助教．2012 年より国立国語研究所コーパス開発センター特任准教授．現在同教授．博士（工学）．

加藤 祥 (正会員)
2008 年早稲田大学大学院文学研究科博士後期課程日本語日本文化専攻単位取得退学．2011 年神戸大学大学院人文学研究科博士課程後期社会動態専攻修了．2012 年より国立国語研究所コーパス開発センタープロジェクト PD フェロー．現在同プロジェクト非常勤研究員．博士（文学）．

付 録

A. 一般化線形混合モデルによる分析

以下では，ベイズ手法に慣れない研究者向けに一般化線形混合モデルによる分析結果を付記する．

A.1 分析手法

まず，対象は BCCWJ-EyeTrack の全データとする．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．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．`subj` と `article` をランダム切片として，次のような式に基づき分析を行った．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う (Baayen, 2008)．モデリングには R の lme4 パッケージを用いた．ゼロ秒のデータポイントを含めない対数線形モデルを用いることにより，定義域の正定値を担保する．ゼロ秒のデータポイントを含めた実時間を使うと外れ値が多くなることが言及されており，実際に一般化線形混合モデルにおいて，(a) ゼロ秒のデータポイントを含めた線形モデル，(b) ゼロ秒のデータポイントを含めない線形モデル，(c) ゼロ秒のデータポイントを含めない対数線形モデルを比較したところ，(c) のモデルのあてはまりが最もよかった．例えば，SPT は全体の 64% がゼロ秒であり，モデルがゼロ秒のデータポイントの影響を大きく受ける．既存の心理言語学の分析で SPT でゼロ秒を含めていたのは，不自然な例文を数多く呈示するために，ゼロ秒のデータポイントが少なかったからであろう[2)]．本研究のように，実データに出現した例文に基づく自然な実験設定では，ゼロ秒を排除するのが一般的になると考え，BCCWJ-EyeTrack を用いたすべての研究はこれにならっている．

一般化線形混合分析モデルにおいては，AIC に基づく forward selection 法によりモデル選択を行った．その過程で，傾きを考慮したモデル・交互作用・ランダム切片に対する係数の組み合わせについても検討したが，収束しなかったり，あてはまりが悪かったりしたために，5 種類の読み時間全てについて収束し，あてはまりがよかった以下の線形式を用いた．

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

A.2 結 果

表 10, 11 に分析結果を示す．

視線走査法においては FFT 以外のものについて，空白ありのほうが読み時間が短くなる．単純に読み時間を短くするという観点でリーダビリティを上げるには，文節間に空白を入れたほうがよい．文節長は FFT 以外について，長くなればなるほど読み時間が長くなる．これは，文節の長さに応じて，表示面積が大きくなり，視線が停留する確率が線形に高くなるためだと考える．係り受けでは FFT 以外について，多くの係り受けがある文節ほど読み時間が短くなる．これは本文に述べた Anti-locality 現象 (Konieczny, 2000) の追認である．レイアウト情報 (is_first, is_last, is_second_last) は，折り返しの視線移動に基づく影響を勘案するものである．最左要素 (is_first) に関しては最右要素やその隣の要素 (is_last, is_second_last) では FPT, RPT, Total などで読み時間が長くなる傾向にある．呈示順 (sessionN, articleN, screenN, lineN, segmentN) は，基本的に進めば進むほど読み時間が短くなる．これは実験協力者が実験に慣れてきた効果であると考える．

次に分類語彙表の類（統語分類）について確認する．全ての読み時間指標について，用の類

2) 今後，先行研究で行われている他のデータのゼロ秒のデータポイントを含めない分析については，オープンデータに基づく再分析が行われるべきである．

表 10 線形混合モデルに基づく分析結果 1

	Dependent variable:	
	logtime	
	FFT	FPT
space=True	−0.006	−0.017***
空白あり	(0.004)	(0.005)
length	−0.003	0.135***
文節長	(0.002)	(0.003)
dependent	−0.003	−0.016***
係り受け	(0.002)	(0.003)
is_first	0.019***	0.090***
最左要素	(0.006)	(0.008)
is_last	−0.009	0.014*
最右要素	(0.006)	(0.008)
is_second_last	−0.001	0.034***
右から 2 番目の要素	(0.006)	(0.007)
sessionN	−0.022	−0.041*
セッション順	(0.016)	(0.024)
articleN	−0.004	−0.005
記事順	(0.004)	(0.007)
screenN	−0.004	−0.018***
画面順	(0.003)	(0.003)
lineN	−0.010***	−0.018***
行番号	(0.002)	(0.003)
segmentN	0.003***	−0.005***
セグメント番号	(0.001)	(0.001)
wlsp_syn[2]	−0.038***	−0.096***
用の類	(0.006)	(0.007)
wlsp_syn[3]	−0.003	−0.056***
相の類	(0.008)	(0.010)
wlsp_syn[4]	−0.020	−0.127***
その他の類	(0.033)	(0.040)
wlsp_syn[NIL]	0.020	−0.075
未登録語	(0.061)	(0.076)
wlsp_sem[.2]	0.014**	0.018**
主体	(0.006)	(0.007)
wlsp_sem[.3]	0.015***	0.024***
活動	(0.005)	(0.006)
wlsp_sem[.4]	0.005	0.022*
生産物	(0.010)	(0.013)
wlsp_sem[.5]	0.034**	0.017
自然	(0.015)	(0.019)
space1:sessionN	0.044	0.059
	(0.031)	(0.049)
Constant	2.299***	2.532***
	(0.017)	(0.026)
Observations	13,232	13,232
Note:	*p<0.1; **p<0.05; ***p<0.01	

表 11 線形混合モデルに基づく分析結果 2

	Dependent variable:	
	logtime	
	RPT	TOTAL
space=True	−0.018***	−0.029***
空白あり	(0.006)	(0.005)
length	0.115***	0.130***
文節長	(0.003)	(0.003)
dependent	−0.012***	−0.018***
係り受け	(0.004)	(0.003)
is_first	0.030***	0.069***
最左要素	(0.009)	(0.008)
is_last	0.088***	−0.007
最右要素	(0.010)	(0.008)
is_second_last	0.045***	0.034***
右から 2 番目の要素	(0.008)	(0.007)
sessionN	−0.049*	−0.047*
セッション順	(0.025)	(0.024)
articleN	−0.007	−0.001
記事順	(0.007)	(0.008)
screenN	−0.017***	−0.025***
画面順	(0.004)	(0.003)
lineN	−0.007**	−0.018***
行番号	(0.003)	(0.003)
segmentN	−0.013***	−0.012***
セグメント番号	(0.002)	(0.001)
wlsp_syn[2]	−0.088***	−0.101***
用の類	(0.009)	(0.008)
wlsp_syn[3]	−0.054***	−0.071***
相の類	(0.012)	(0.010)
wlsp_syn[4]	−0.137***	−0.189***
その他の類	(0.049)	(0.042)
WLSPLUWAF	−0.109	−0.160**
未登録語	(0.092)	(0.079)
wlsp_sem[.2]	0.005	0.018**
主体	(0.009)	(0.008)
wlsp_sem[.3]	0.021***	0.023***
活動	(0.007)	(0.006)
wlsp_sem[.4]	0.018	0.037***
生産物	(0.015)	(0.013)
wlsp_sem[.5]	0.024	0.040**
自然	(0.023)	(0.020)
space1:sessionN	0.061	0.061
	(0.050)	(0.048)
Constant	2.603***	2.672***
	(0.027)	(0.026)
Observations	13,232	13,232
Note:	*p<0.1; **p<0.05; ***p<0.01	

(wlsp_syn[2]) は体の類 (wlsp_syn[1]) に対して，有意に読み時間が短くなる傾向がみられた．また相の類 (wlsp_syn[3]) は FFT 以外で体の類に対して有意に読み時間が短く，用の類に対して有意に読み時間が長い傾向がみられた．

最後に分類語彙表の部門 (意味分類) について確認する．抽象的関係 (wlsp_sem[.1]) に対して主体 (wlsp_sem[.2])・活動 (wlsp_sem[.3])・生産物 (wlsp_sem[.4]) が FFT, TOTAL に関して読み時間が長い傾向がみられた．