言語研究（*Gengo Kenkyu*）**156**: 67–96（2019）　　doi: 10.11435/gengo.156.0_67

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――

浅原正幸　小野創　宮本エジソン正

国立国語研究所　津田塾大学　はこだて未来大学

【要旨】 Kennedy et al.（2003）は，英語・フランス語の新聞社説を呈示サンプルとした母語話者の読み時間データを Dundee Eye-Tracking Corpus として構築し，公開している。一方，日本語で同様なデータは整備されていない。日本語においてはわかち書きの問題があり，心理言語実験においてどのように文を呈示するかがあまり共有されておらず，呈示方法間の実証的な比較が求められている。我々は『現代日本語書き言葉均衡コーパス』（Maekawa et al. 2014）の一部に対して視線走査法と自己ペース読文法を用いた読み時間付与を行った。24人の日本語母語話者を実験協力者とし，2手法に対して，文節単位の半角空白ありと半角空白なしの2種類のデータを収集した。その結果，半角空白ありの方が読み時間が短くなる現象を確認した。また，係り受けアノテーションとの重ね合わせの結果，係り受けの数が多い文節ほど読み時間が短くなる現象を確認した*。

キーワード： 均衡コーパス，読み時間，視線走査法，自己ペース読文法

1. はじめに

20世紀半ばより，英語・フランス語のようなヨーロッパ言語を対象に，人間の文処理機構を解明するための心理言語実験に基づくデータが集積されてきた。計算機上で読み時間が容易に計測できるようになり，内省による作例に対する読み時間を心理統計学的な計測値とした研究が行われてきた。心理言語学における伝統的なアプローチは，特定の仮説を検証するため，人手で作成された刺激文を使って心理統計学的な計測値を集積する手法が一般的であった。この手法は特定の構造や現象に対する仮説の正当性を評価する局所的な手法であるが，より広範囲の現象を被覆した仮説を比較することは難しい。検証したい仮説が細分化されて検証の一般化が失われていく中で，非常に混みいった文構造を用いた作例に基づく心理言語実験が問題視され，より自然な刺激文に基づく心理言語研究の重要性が言及されている（Futrell et al. 2018）。一方，コーパス言語学の分野では，適切にサンプリングすることにより，分野横断的により自然な例文の集積が行われ，言語コーパスに対して，形態・統語・意味情報を表現するアノテーション付与が各機関で進められている。

* 本研究は科研費 25284083，17H00917，18H05521 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の成果物です。実験環境作成に協力していただいた先生方，実験補助をしていただいた方々，実験に参加していただいた方々に感謝いたします。2名の査読者による詳細なコメントに敬意と感謝を表します。

ここでコーパス言語学の手法と心理言語学の手法を結びつけることで，言語理解過程解明を目的とする再利用性のある日本語の言語資源の構築を考える。具体的にはさまざまなアノテーションが施された『現代日本語書き言葉均衡コーパス』（BCCWJ）（Maekawa et al. 2014）に対して読み時間を付与する。

同様の研究として，Kennedy et al.（2003）の Dundee Eye-Tracking Corpus がある。英語とフランス語を対象言語とし，それぞれ 10 人の母語話者を実験協力者として，新聞社説記事 20 本に対する視線走査情報を記録し，研究用途に一次情報が公開されている。Dundee Eye-Tracking Corpus は特定の言語現象分析を目的としていない。このような共有されたデータを用いることにより，心理言語学におけるさまざまな仮説の客観的な検証が可能になっている（Smith and Levy 2013，Pynte et al. 2008，van Schijndel and Schuler 2013，他多数）。例えば，Demberg and Keller（2008）は，Gibson（1998）による Dependency Locality Theory（DLT）の要素統合の負荷（integration cost）と Hale（2001）によるサプライザル理論とを Dundee Eye-Tracking Corpus 上の読み時間データを用いて検証を行い，任意の単語に対する読み時間の予測についてはサプライザル理論の方が優れていることを示した。また，Roland et al.（2012）は Demberg and Keller（2007）が行った Dundee Eye-Tracking Corpus を用いた関係節の読み時間の分析が，限られたデータポイントに基づいてゆがめられていることを再検証により証明している。このように，公開され共有されたデータを用いることにより，先行研究の結果を再検証するという取り組みも可能になっている。

また，Dundee Eye-Tracking Corpus にさまざまなアノテーションを付与し，アノテーションと読み時間の対照分析を行う試みも進められている。Pynte and Kennedy（2007）は，単語クラスの分布が異なることに基づいて，読み時間が変化することを示した。Frank（2009）は，Dundee Eye-Tracking Corpus に人手で品詞情報を付与しサプライザル理論の検証を行った。Barrett, Agic and Søgaard（2015）は，Dundee Eye-Tracking Corpus に人手で Universal Dependencies 基準（Universal Dependencies 2014）の品詞情報や単語係り受け情報を付与した。読み時間のふるまいから，品詞や係り受け情報を推定するような研究も進められている（Barrett and Søgaard 2015a, b）。Seminck and Amilli（2018）は同データに人手で共参照情報を付与し，Seminck and Amilli（2017）は同データと共参照情報を用いて共参照情報が読み時間に与える影響を分析した。

他にも，読み時間が付与された英語のデータとして，アマチュアの小説に基づく UCL Corpus（Frank et al. 2013）や既存の自然文を編集した Natural Stories Corpus（Futrell et al. 2018）がある。ヒンディー語（Husain et al. 2015）や中国語のデータ（Yan et al. 2010）が整備されているほか，作例に基づくが，ドイツ語のデータ（Kliegl et al. 2006）も整備されている。

このような人間の文処理研究の共通基盤となり，作例によらないコーパスに基づくデータの整備が日本語においても求められている。本研究では，適切に言語の生産実態をサンプリングしたコーパス母集団に対して，心理言語学の手法に基づき多

人数の読み時間情報を被覆するようにアノテーションを行うことで，日本語の文処理研究に資する言語資源を構築する。その際，読み時間の付与手法として，移動窓方式に基づく自己ペース読文法（Just et al. 1982）と視線走査法の二種類の方法を採用する。

本稿では上記目的を達成するために構築した BCCWJ-EyeTrack コーパスの仕様と基礎統計について述べる。言語背景情報・記憶力・語彙力を評価した日本語母語話者 24 人を対象とし，文節間の半角空白あり・なしなどで対照比較できるデータを構築した。

本データの研究への有効性を示すために 2 つの分析を行った（詳細は第 5 節を参照のこと）。1 つは，文節と文節の間に空白を挿入することの読み時間に対する影響である。松田（2001）と Sainio et al.（2007）は，漢字･ひらがな･カタカナといった日本語の表記を変えるとともに，文節と文節の間に空白を入れることによる読み時間の変化を調査した。彼らの研究では，全角の空白が用いられていたことが論文中の例文から推察される。これは，視線走査結果の分析時に半角文字単位に視線停留位置を集計するのが技術的に煩雑であることがあげられる。本研究においては，より自然な呈示手法であると考えられる半角空白を入れた調査を行う。同様の調査がタイ語でも行われている（Winskel et al. 2009）。

もう 1 つは，任意の要素に対する係り受けの数が増えると読み時間が短くなるという anti-locality 効果（Konieczny 2000）についての分析である。係り受けアノテーション BCCWJ-DepPara（浅原・松本 2018）との重ね合わせにより，係り受けの数と読み時間の関係が検証可能になる。日本語における係り受けの数と読み時間の関係について，Gibson（1998）による DLT の要素統合の負荷と Hale（2001）によるサプライザル理論の影響がどの程度存在するのかについて検証する。サプライザル理論のモデル化は，Levy（2008）の統語構造（PCFG）に基づく期待値に基づく手法や，Smith and Levy（2008）の trigram 文脈に基づく期待値に基づく手法などがある。BCCWJ には，短単位・長単位・文節単位・文単位の 4 つのレベルの境界情報が付与されている。サプライザル理論に基づく anti-locality 効果の検証においては予測確率を推定する単位の定義が必要になるが，4 つのレベルの境界情報のうち BCCWJ-DepPara の統語アノテーションの基本単位である文節単位を用い，係り受けの数により期待値のモデル化を行う。

従来，特定の検証したい仮説に基づき，統制した作例に基づく小規模な実験は，頻度主義的な統計手法により検証されることが多かった。特定の構造や現象に対する少ない要因を評価するためには十分であったが，複合的なパラメータを評価するためには頻度主義的な統計手法には限界があった。このため，従来の頻度主義に基づく仮説の評価は複雑な要因の統制が困難であるために極端に単純化されてきた。本研究では，文節境界空白の有無，レイアウト情報，係り受けの数など複合的な要因を同時に分析するために，ベイズ主義的な階層ベイズモデルを用いた（Sorensen et al. 2016）。なお，頻度主義的な分析と対照可能にするために，一般化線形混合モ

デルによる分析結果を付録に示す。

結果，文節間に空白を入れたほうが読み時間が短くなる傾向と，先行文脈に係り元文節が多い要素ほど読み時間が短くなる傾向が確認された。この得られた結果に基づき，後に述べる anti-locality 現象に関する先行研究との相違について議論する。

本稿の構成は以下の通りである。2 節では収集した実験協力者の言語背景情報について示す。3 節では読み時間の収集方法について，4 節では収集した読み時間の言語資源化手法について述べる。5 節に収集したデータの統計分析とそこから得られた知見について示す。最後にまとめと今後の研究の方向性について示す。

2. 実験協力者の言語背景情報

実験協力者は 18 歳以上の日本語母語話者であった。実験協力者の言語運用能力によって読文時間は変化することが考えられる。そこで言語背景情報を得るためのアンケート・語彙数推定テスト・記憶力テストなどを読文時間評価前に行った。

実験協力者は，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[1]・言語形成地[2]・父親出身地・母親出身地の情報を取得した。

実験協力者の語彙数を評価するために，語彙数判定テスト（Amano and Kondo 1998）を実施した。語彙数判定テストは心理実験により推定された単語親密度（天野・近藤 1999）に基づいて構成された日本語語彙数推定テストである。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った。知っている単語集合から実験協力者の語彙数を推定した。

また，実験協力者の記憶力を評価するために，日本語リーディングスパンテスト（苧坂 2002）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 文ずつ実験協力者に呈示して，音読させるなどしながら下線部を記憶させた複数文を呈示したのちに，それらの文を隠した状態で，下線が引かれていた部分を呈示順に再生させて，その再生の正答率を評価す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。データとしてオリジナルのスパン得点を記録した。

3. 刺激文と読み時間の収集方法

本節では読み時間の収集方法について説明する。

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）（Maekawa et al. 2014）のコアデータの新聞記事データ（PN サンプル）の一部とした。コーパスアノテーションの分野では，オープンサイエンスに向けてできる限り同じテキストに様々な情報を付与するという取り組みが進められている。対象の記事は，研究者コミュニティで共有

[1] 裸眼・ソフトコンタクトレンズ・眼鏡の方のみが実験に参加した。
[2] 15 歳までに住んでいた場所を都道府県単位・年単位で記述。

されているアノテーションの優先順位[3]に基づいて選択した。これにより，係り受け（浅原・松本 2018）・節境界（Matsumoto et al. 2018）・分類語彙表番号（加藤他 2019）・情報構造（宮内他 2018）・述語項構造および共参照（植田他 2015）・否定の焦点（松吉 2014）のアノテーションとの重ね合わせに基づく分析が可能になる。Dundee Eye-Tracking Corpus においても Dundee Treebank（Barrett, Agic and Søgaard 2015）など品詞・係り受け・共参照の整備が進んでいるが，本研究の BCCWJ-EyeTrack のようにはコーパス言語学的な統語・意味・談話レベルの情報は重畳的に付与されていない。

読み時間データの収集方法として，自己ペース読文法と視線走査法を用いた。

自己ペース読文法は，キーボード入力に基づき，逐次的にまた非累積的に文字列を表示し，実験協力者のペースで文を読む課題である。図 1 に課題の画面例を示す。最初，コンピューター画面上には，文の長さを表すアンダーバーが表示されている。被験者がスペースキーを押すごとに，刺激文の始めから 1 文節（もしくは 1 単語）ずつ表示され，直前に表示されていた文節はアンダーバーに戻る。文節が表示されてから，次にボタンを押すまでの時間が，その文節の読解時間としてミリ秒単位で記録される。英語においては視線走査で得られる読み時間と非累積移動窓方式の自己ペース読文法に相関があることが知られており（Just et al. 1982），安価な機器で読み時間を取得することができる。刺激の呈示方法として移動窓方式を用いた。自己ペース読文法を実施するソフトウェアとして Linger[4] を用いた。

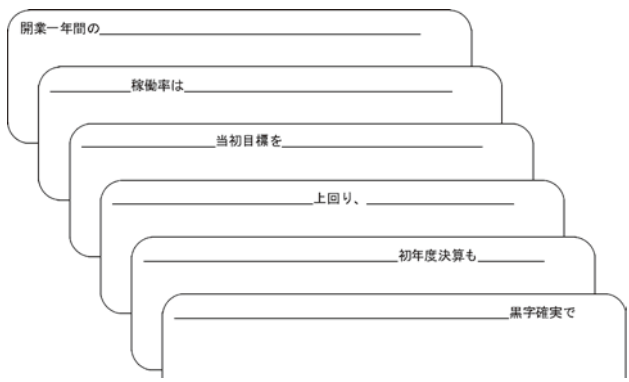

図 1　移動窓方式による自己ペース読文法

視線走査法は，実験協力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

[3] https://github.com/masayu-a/BCCWJ-ANNOTATION-ORDER
[4] http://tedlab.mit.edu/~dr/Linger/

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）[5] を用い，基本的には実験協力者の右目の情報を取得した[6]。時間解像度は 1000 Hz で，ミリ秒単位のデータが収集可能である。視線走査法においては刺激となるテキストは等幅フォント（MS 明朝 24 ポイント）を用いて，横書きで 1 画面に最大 5 行を 21.5 インチのディスプレイ [7] に呈示した。横方向には全角で最大 53 文字を呈示し，後述のとおり文節境界に半角空白を入れた場合には，最大全角 53 文字を超えないようにした単位で折り返し，1 画面に 5 行まで表示した。文境界には必ず改行を入れた。視線走査装置の上下方向の誤差を吸収するために，各行は 3 行分の空行を追加して呈示した。実験協力者はあご台に顔を固定した状態で，ハーフミラー越しに画面を見るという姿勢で，課題に取り組んだ。自己ペース読文法では，ハーフミラーつきのあご台を用いない以外は同条件で実験を行った。

自己ペース読文法では，テキスト文字列を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た。また文節境界に半角空白を入れた条件と空白を入れていない条件の 2 つの条件を用意し，読み時間を計測した。実験は新聞記事 20 件を 5–6 件ずつ A，B，C，D の 4 つのユニットに分割し，視線走査法による計測を 2 セッション実施したのちに，自己ペース読文法による計測を 2 セッション実施した[8]。実験協力者は各新聞記事20件を一度だけ読む。各ユニットの文節数，文数，画面数を表 1 に示す。1 件の新聞記事を読み終わり，次の新聞記事が始まる際には，必ず画面を改めた。実験協力者は 3 人ずつ 8 つのグループに分け，表 2 のように実験を行った。全実験協力者は視線走査法を行ったのちに，自己ペース読文法を行った。視線走査法は準備に時間がかかる一方，自己ペース読文法は準備に時間がかからないという理由とともに，順序を入れ替えて実験を行うと，分析の要因が 1 つ増えることになるため，今回は 2 つの順序を固定した。課題の順序による影響の評価は今後の課題として検討する。

表 1　それぞれの記事ユニットに含まれる文節数，文数，画面数

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

[5] タワーマウントとは，ハーフミラー越しで撮影する機材。http://sr-research.jp/products/tower_mount/

[6] 基本的に利き手の側の目のデータを収集したが，実験協力者全員が右利きであった。

[7] EIZO FlexScan EV2116W（解像度 1920x1080）をあご台から 50 cm の位置に設置。

[8] サンプル C が 5 件のつもりであったが，このうち 1 件が連続する同じトピックの記事 2 件であったため，6 件となる。この 2 件の間には改ページがあったために別の記事として扱う。

表 2　実験計画：各被験者グループにおける記事ユニット・課題・文節境界の空白の有無の対応関係

グループ	視線走査法				自己ペース読文法			
	記事ユニット	文節境界空白	記事ユニット	文節境界空白	記事ユニット	文節境界空白	記事ユニット	文節境界空白
ア	A	無	B	有	C	無	D	有
イ	A	有	B	無	C	有	D	無
ウ	C	無	D	有	A	無	B	有
エ	C	有	D	無	A	有	B	無
オ	B	無	A	有	D	無	C	有
カ	B	有	A	無	D	有	C	無
キ	D	無	C	有	B	無	A	有
ク	D	有	C	無	B	有	A	無

4. 読み時間の言語資源化

4.1. 読み時間の集計作業

自己ペース読文法で取得したデータは，取得時に語句が文節単位に呈示され，読み戻しができないために，文節単位の読み時間がそのままデータとなる。視線走査法で取得したオリジナルの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）を得た[9]。このデータを国語研文節単位でグループ化しなおしたものを注視順データと呼ぶ。この注視順データを，視線走査法を用いた読み時間計測で標準的に用いられている，以下の 5 つの計測時間データ（measures）に加工した（van Gompel et al. 2007）。これらは国語研文節単位を注視領域として作成した。

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

説明のために図 2 の例を用いる。図中 1–12 の数字が視線走査順を表す。

First Fixation Time（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。

First-Pass Time（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。

[9] 文節境界に半角空白を入れるために，半角単位の注視箇所をグループ化した。

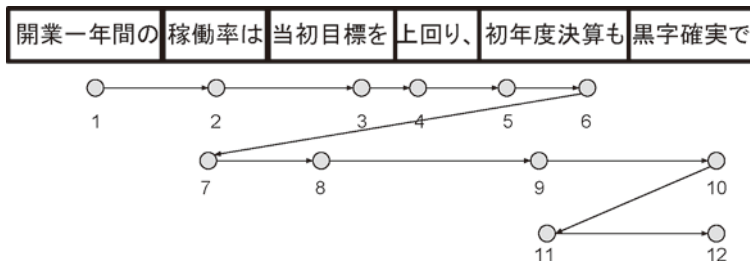

図 2　視線走査順の例

Second-Pass Time（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。SPT においては 2 回目以降に視線が停留していないデータポイントを認めない。SPT に似た指標として Rereading time がある（Vasishth and Drenhaus 2011）。Vasishth らの Rereading time は 2 回目以降に視線が停留していないデータポイントについて 0 ms としてデータに含める。2 回目以降の視線停留のデータの取り扱いについて，SPT として扱うか Rereading time として扱うかについては依然論争がある（Patterson and Drummer 2016）。重要な点として，我々の SPT の定義は，他の読み時間指標により一意に計算できるものではない点がある。この SPT の分析については後述する。

Regression Path Time（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 Time（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。

テキスト生起順データにおいて，サッケード（跳躍眼球運動）の時間は集計しない。これらの時間情報を各種情報とともに CSV 形式に整形して公開する。公開データにおいては，平均読み時間や標準偏差などを用いたトリミングなどの時間情報削除処理は実施していない。

また，被験者が記事をきちんと読んでいるか確認するために，各記事を読んだ後に，Yes/No で解答できる簡単な内容理解課題を課した。視線走査法の内容理解課題の正解率は 99.2%（238/240）で，自己ペース読文法の内容理解課題の正解率 77.9%（187/240）より有意に高かった（$p < 0.001$）。視線走査法は一画面の間は自由に再読することができる一方，自己ペース読文法は，読み戻しが許されず，複数

の画面に記事が続く場合など，内容を記憶している負荷が高かったことがうかがえる。なお読み時間の分析時には，全データポイントを利用した。

次小節に CSV 形式の公開データの概要について示す。

4.2. 公開データの形式

公開データは，時間情報を元テキストの情報・実験協力者の情報などとともに読み時間の種類ごとの CSV 形式のデータとして公開する。表 3 にデータ形式を示す。

表 3　データフォーマット

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（対数）
measure	factor	読み時間の種類
length	int	文字数
space	factor	文節境界空白の有無
dependent	int	係り受け関係
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
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
bunsetsuN	int	文節呈示順
is_first	bool	行内最左要素
is_last	bool	行内最右要素
is_second_last	bool	行内右から 2 番目の要素
subj	factor	実験協力者 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

出現書字形（surface: factor）は実験協力者に呈示した文字列である。国語研文節単位に区分されており，全角空白は除去した。

読み時間（time: int）は各実験で得た時間情報である。自己ペース読文法の場合は実験協力者がその文節を見ていた時間である。視線走査法の場合は前小節で示した First Fixation Time（FFT），First-Pass Time（FPT），Second-Pass Time（SPT），Regression Path Time（RPT），Total Time（Total）の 5 種類のいずれかである。単位はミリ秒とする。読み時間の種類（measure: factor）として{ ‘SelfPaced’，‘EyeTrack:FFT’，‘EyeTrack:FPT’，‘EyeTrack:SPT’，‘EyeTrack:RPT’，

‘EyeTrack:Total’ } を定義する。尚，配布データは読み時間の種類ごとに 1 ファイル作成する。対数読み時間（logtime: num）は time の常用対数をとったものである。

文字数（length: int）は，呈示した文節の出現書字形 surface を構成する文字の数である。注視対象の面積に相当する。文節境界の有無（space: factor）は呈示した画面に文節境界に半角スペースがある（‘1’）かない（‘0’）かを表す。係り受け関係（dependent: int）は当該文節に係る文節数。文節係り受けは人手で付与したもの（Asahara and Matsumoto 2016）を重ね合わせた。図 3 に係り受け関係の例を示す。図の各文節の下の数字が当該文節に係る文節数に相当する。

記事に関するデータとして sample，article，metadata_orig，metadata の 4 つを整備した。サンプル名（sample: factor）は，セッションごとに準備した記事ユニットで {A, B, C, D} からなる。前述の通り各ユニットは新聞記事 5–6 件から構成されている。記事情報（article: factor）は，記事単位の一意な識別子で，BCCWJ のアノテーション優先順位・BCCWJ 内サンプル ID・記事番号をアンダースコアで連結したものとする。文書構造タグ（metadata_orig: factor）は BCCWJ 内文書構造タグで，BCCWJ の XML の ancestor axis にあるタグ情報をスラッシュで連結したものである。メタデータ（metadata: factor）は前述の metadata_orig から記事の特性のみを抽出したものである。{authorsData（著者情報），caption（キャプション），listItem（リスト），profile（プロフィール），titleBlock（タイトル領域），未定義 } のいずれかであり，BCCWJ 内の文書構造タグの誤り・欠落を人手で修正したものである。

次に記事や画面の呈示順の情報について説明する。セッション順（sessionN: int）は実験法ごとに文節境界空白有と文節境界空白無の 2 種類のセッションの順序を表す。記事呈示順（articleN: int）はセッションごとの記事の呈示順（1–6）を表す。画面呈示順（screenN: int）は複数の画面にわたる記事があり，記事ごとの画面呈示順を表す。行呈示順（lineN: int）は画面ごとの行呈示順（1–5）であり，画面上の垂直方向の位置を表す。文節呈示順（bunsetsuN: int）は行ごとの文節呈示順である，画面上の水平方向の位置を表す。これらの呈示順情報により画面推移上の一意な識別が可能である。

また，文頭の文節は常に係り受けの数が 0 であり，文末の文節は係り受けの数が多い傾向にある。また，画面レイアウト上，最左要素・最右要素・右から 2 番目の要素は眼球運動中に「復帰改行」の操作の影響がある。この問題を扱うために，レイアウト情報として，最左要素（is_first: bool）・最右要素（is_last: bool）・右から 2 番目の要素（is_second_last: bool）を固定要因とする。sample_screen は，画面に対する一意な識別子である。

実験協力者 ID（subj: factor）は実験協力者を表示する一意な識別子である。実験協力者の特性として 2 つの情報を持つ。1 つはリーディングスパンテスト得点（rspan: num）であり，1.5-5.0 の 0.5 刻みの値を持つ。もう 1 つは語彙数テストの結果（voc: num）であり，オリジナルの結果を 1000 語で割ったもの（37.1–61.8）である。

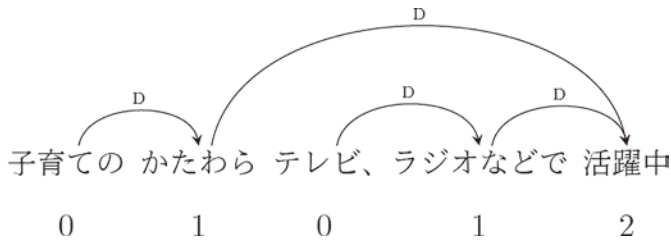

図 3　係り受けアノテーションと係り受けの数

表 4 に読み時間の基礎統計（平均・標準偏差・四分位数）を示す。単位はすべてミリ秒である。視線走査法の場合にはゼロ秒（注視されていない文節）は排除して集計した。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 and Levy 2012，Luong et al. 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 and 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う。

表 4　読み時間の代表値（ミリ秒）

Self-paced	平均	標準偏差	Min.	1st Qu.	Median	3rd Qu.	Max.
	699	506	62	415	550	798	9454
Eye Tracking (excl. 0)							
First Fixation Time	235	142	12	162	219	292	1700
First-Pass Time	475	497	14	205	321	548	7340
Second-Pass Time	330	253	20	173	258	418	2553
Regression Path Time	698	1013	19	235	391	745	21577
Total Time	597	589	18	247	416	721	8397

本研究では，SPT の結果については基礎統計のみ提示し，統計分析結果を提示しない。SPT に関しては，2 回目以降に視線停留が行われなかった場合のデータの扱いについて，0 の値を割り当てて扱うか，0 の値を排除して扱うか研究者の間で議論が未だ収束していない（Patterson and Drummer 2016）。例えば，Clifton et al.（2007）は 2 回目以降に視線停留が行われないデータポイントを排除せず，それを SPT と呼んで扱うことを主張している。一方，Vasishth and Drenhaus（2011）は 0 の値を割り当てて扱うものを Rereading time と呼び，排除して扱うものを SPT として区別し，そのような SPT を扱うべきとしている。

本稿においては，0 の値を割り当てる Rereading time は分析対象とせず，0 の値

を排除した SPT を分析対象とする後者の立場をとる。なぜならば，2 回目以降に視線停留が行われなかった部分に 0 の値を割り当てて扱う Rereading time の線形式に関しては，rereading time = TOTAL − FPT が常に成り立つので，Rereading time の効果についてはそれぞれの係数から導出できる。そのため，TOTAL と FPT に関して分析がなされている場合には，Rereading time をわざわざ分析する必要はないからである。一方，0 の値を排除した SPT においては，対数読み時間のモデル化における 0 の値の問題を回避できるほか，本来欠損値に対して 0 の値を割り当てるという overspecified の問題を回避できる。SPT の結果は責任著者に問い合わせることによって得られる。

図 4 に読み時間（左図）と対数読み時間（右図）の五数要約箱ひげ図を示す。一般に読み時間の分布は左図のように外れ値の影響が大きくなるため，対数読み時間で分析される。定義域が正数である対数読み時間で分析することにより，モデル作成時に正定値が担保される。また，箱ひげ図から対数読み時間のほうがより正規分布に適合し，外れ値の影響が小さくなる傾向という利点があること（Gelman and Hill 2006）が確認できる。

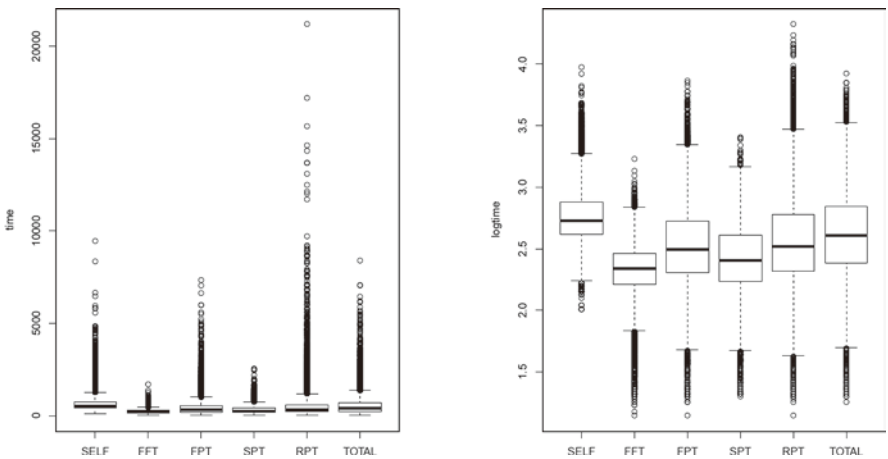

図 4　読み時間（左）と対数読み時間（右）の箱ひげ図

5. 分析

本コーパスを用いた分析例として，日本語における文節間の空白の表示と anti-locality 現象（Konieczny 2000）について調査を行った。

5.1. 分析対象：日本語における空白

言語における空白の表示が読み時間にどのような影響を与えるかは，心理言語実

験などの設定において，重要な問題である。具体的には空白の表示が，視線の着点と周辺視野にどのような影響を与えるかが重要である。英語においては，Rayner et al.（1998）が単語境界に空白を入れることにより単語認識を促進するとともに，空白が視線の着点を誘導する効果を持つことを示している。タイ語は，単語や句の境界に空白が示されない言語であるが，アメリカ在住のタイ語母語話者にスペース入りでテキストを呈示したところ，読み時間が短くなる傾向が見られた（Kohsom and Gobet 1997）。しかし，その後の調査で，タイに住むタイ語母語話者を対象にした実験によると停留時間や Total Reading Time は空白による読み時間の短縮は見られるが，全体の読み時間が空白の呈示により遅くなることが報告されている（Winskel et al. 2009）。

日本語においては，ひらがなのみで表記されたテキストについて，全角空白を文節間に入れたほうが読み時間が短くなる傾向が見られた。しかしながら，漢字かなまじり文のテキストについては，この傾向が見られなかった（Sainio et al. 2007，松田 2001）。この先行研究の問題点として，（1）半角空白ではなく全角空白を入れたために周辺視野で次の単語を読む効果がなくなったこと，（2）呈示文数が少ないために普段見慣れない空白入りの文章に慣れる時間が短かったことがあげられる。

本研究では，日本語において文節間に半角空白を入れることにより，周辺視野がどのように読み時間に影響をするのかを調査する。自己ペース読文法では周辺視野の効果が得られないが，視線走査法では周辺視野の効果が得られることを鑑み，分析を行う。

5.2. 分析対象：Anti-locality

Anti-locality 現象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る。この現象は，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 and Döring 2003，Levy and Keller 2013，日本語：Uchida et al. 2014，ヒンディー語：Vasishth and Lewis 2006，Husain et al. 2014）。

このような読み時間の短縮は，主辞後置言語において，係り元要素が多い要素を読むのに負荷がかかるという予測（Gibson 1998）や，後続する主辞の処理コストは先行文脈の数の影響を受けないという予測（Nakatani and Gibson 2010）などの，ワーキングメモリモデルでは説明できない現象であった。

この現象はサプライザル理論（Hale 2001，Levy 2008）の説明と親和性がある。

今回，均衡コーパスと係り受けアノテーション BCCWJ-DepPara（Asahara and Matsumoto 2016）とを用いることにより，日本語において特定の品詞によらない設定で anti-locality 現象を調査する。

5.3. モデリング手法

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 et al. 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することができる。

記事中の本文（タイトル以外の部分）に出現する文節のみを対象とする。具体的には metadata が authorsData, caption, listItem, profile, titleBlock のものを削除した。モデリングは，自己ペース読文法（SELF）・視線走査法（FFT，FPT，RPT，TOTAL）の 5 種類の指標について行った。

$time^{(k)}$ を，データポイント $k \in 1,\cdots,N^{(k)}$ 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：

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

（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：

$$\mu^k = \alpha + \beta^{(k)}_{length} + \beta^{(k)}_{space} + \beta^{(k)}_{dependent} + \beta^{(k)}_{sessionN} + \beta^{(k)}_{articleN} + \beta^{(k)}_{screenN} + \beta^{(k)}_{lineN} + \beta^{(k)}_{segmentN} + \beta^{(k)}_{is\_first} + \beta^{(k)}_{is\_last} + \beta^{(k)}_{is\_second\_last} + \gamma^{(i)}_{article} + \gamma^{(j)}_{subj} \qquad (2)$$

（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 ∈ {length, space, dependent, sessionN, articleN, screenN, lineN, segmentN, is_first, is_last, is_second_last} の傾きを表す。$\gamma^{(i)}_{article}$ はランダム因子である記事 i ∈ $1,\cdots N_{article}$ の正規分布，$\gamma^{(j)}_{subj}$ はランダム因子である記事 j ∈ $1,\cdots N_{\text{subj}}$ の正規分布であり，次の（3），（4）式のように定義する：

$$\gamma^{(i)}_{article} \sim Normal\left(0, \sigma_{article}\right), \qquad (3)$$
$$\gamma^{(i)}_{subj} \sim Normal\left(0, \sigma_{subj}\right), \qquad (4)$$

（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 σ_{article}, σ_{subj} として推定する。

以下，各固定因子 f ∈ {length, space, dependent, sessionN, articleN, screenN, lineN, segmentN, is_first, is_last, 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，上に述べた 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。視線走査法の読み時間のデータポイントのうち，ゼロミリ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。モデリングには RStan を用いる。

5.4. 結果

表 5 に分析の要約を示す。+ が読み時間が長くなることを表す。− が読み時間が短くなることを表す。0 が差がないことを表す。なお，ベイズ主義的な手法においては帰無仮説を立てないために，有意（significant）であるという説明は適さない。強い証拠（strong evidence）があるという記載が正しいが，ここでは差があるかないかのみを議論する。表 6，7，8，9，10 に各読み時間指標のモデリングの結果を表す。なお，表 6，7，8，9，10 中，EAP 推定量（Expected A Posteriori）の差が事後標準偏差（sd）の 2 倍を超える場合に強い証拠となる差がありと認定する。対応する図 5，6，7，8，9 の推定された係数の分布（95% 信用区間）と 0.0 との重なりの有無により判断できる。n_eff は有効サンプルサイズ，se_mean は EAP の標準偏差を表す。すべてのモデルで収束判定指標 Rhat が 1.1 以下であることを確認した。表 5 に結果の要約を示す。

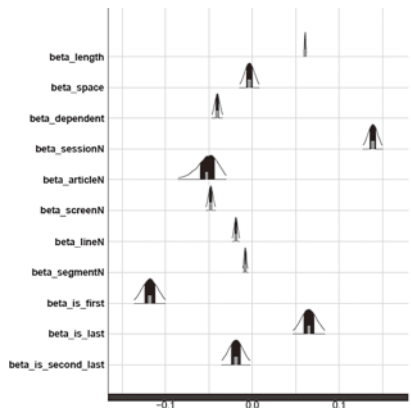

図 5　自己ペース読文法（SELF）の係数

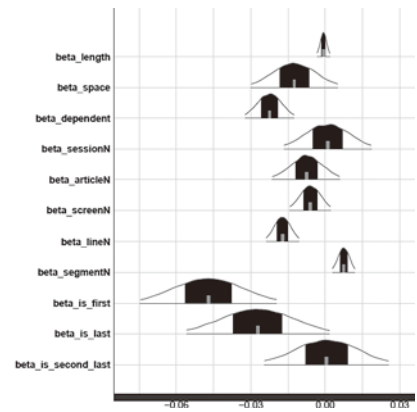

図 6　First Fixation Time（FFT）の係数

まず，文節長（length）は，FFT（最初の停留のみ）以外で読み時間が長くなる。これは視線が停留する確率は面積に対して比例することから自然な結果だと言える。

呈示順（sessionN，articleN，screenN，lineN，segmentN）は全体として，基本的に進むにつれて被験者が慣れていく傾向がみられた。記事呈示順（articleN）の効果が出なかったのは，その従属関係から $\gamma^{(i)}_{article}$（記事に対するランダム因子）により吸収されたのではないかと考える。

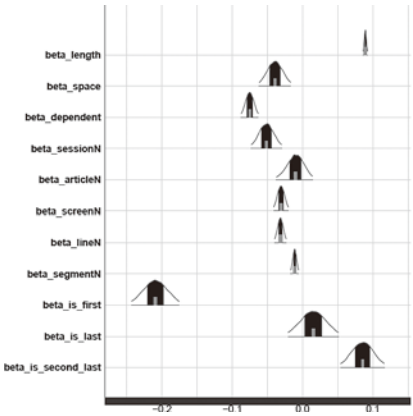

図 7　First Pass Time（FPT）の係数

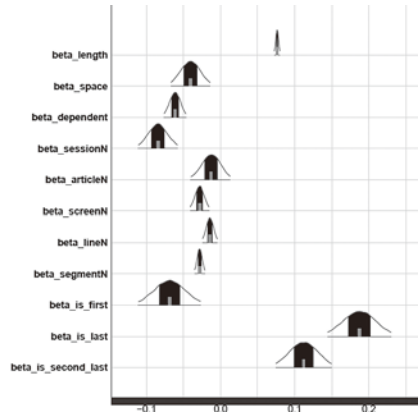

図 8　Regression Path Time（RPT）の係数

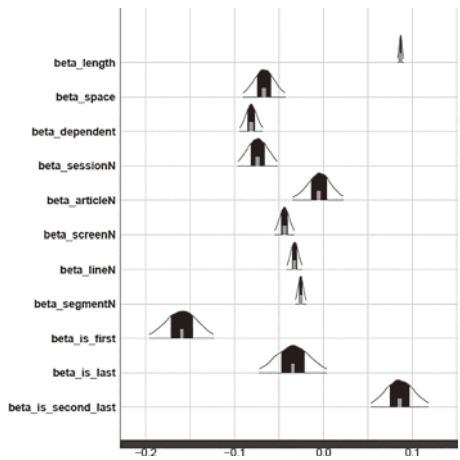

図 9　Total Time（TOTAL）の係数

レイアウト情報（is_first，is_last，is_second_last）は画面上のレイアウトによる影響を調査する。最左要素は 1 文が 1 行で呈示される場合には，係り受けの数が 0 になるために統計処理において特別扱いする必要がある。さらに右から左に眼球運動を復帰改行することも考慮する必要がある。右側要素においては，周辺視野のことを考えて最右要素と右から 2 番目の要素を検討する必要がある。最左要素（is_first）で読み時間が短くなる現象が見られた。最左要素においては復帰改行の影響が強いことが考えられる。また，SELF，RPT について最右要素（is_last）で，FPT，RPT，TOTAL について右から 2 番目の要素（is_second_last）で，読み時間が長くなる現象がみられた。RPT では次の行に眼球移動するまで読み時間を集積

するために読み時間が長くなる傾向にある。

表 5　ベイジアン線形混合モデルの結果の要約

	SELF	FFT	FPT	RPT	TOTAL
length	+	0	+	+	+
space	0	0	−	−	−
dependent	−	−	−	−	−
sessionN	+	0	−	−	−
articleN	−	0	0	0	0
screenN	−	0	−	−	−
lineN	−	−	−	−	−
segment	−	+	−	−	−
is_first=True	−	−	−	−	−
is_last=True	+	0	0	+	0
is_second_last=True	−	0	+	+	+

表 6　ベイジアン線形混合モデルの結果（自己ペース読文法：SELF）

Parameter	Rhat	n_eff	mean	sd	se_mean	2.50%	50%	97.50%
alpha	1.000	3893	6.402	0.077	0.001	6.255	6.400	6.559
beta_length	1.000	20000	0.061	0.001	0.000	0.059	0.061	0.063
beta_space	1.000	20000	−0.003	0.006	0.000	−0.015	−0.003	0.008
beta_dependent	1.000	20000	−0.040	0.003	0.000	−0.047	−0.040	−0.034
beta_sessionN	1.000	20000	0.139	0.006	0.000	0.128	0.139	0.150
beta_articleN	1.000	6356	−0.053	0.014	0.000	−0.085	−0.051	−0.030
beta_screenN	1.000	20000	−0.048	0.003	0.000	−0.053	−0.048	−0.043
beta_lineN	1.000	20000	−0.019	0.002	0.000	−0.023	−0.019	−0.015
beta_segmentN	1.000	20000	−0.008	0.001	0.000	−0.011	−0.008	−0.005
beta_is_first	1.000	20000	−0.118	0.009	0.000	−0.136	−0.118	−0.100
beta_is_last	1.000	20000	0.065	0.009	0.000	0.047	0.065	0.083
beta_is_second_last	1.000	20000	−0.019	0.008	0.000	−0.035	−0.019	−0.003
sigma	1.000	20000	0.371	0.002	0.000	0.368	0.372	0.375
sigma_article	1.000	7347	0.077	0.019	0.000	0.050	0.074	0.126
sigma_subj	1.000	20000	0.258	0.041	0.000	0.192	0.253	0.352
log-posterior	1.000	6093	8702.299	5.586	0.072	8690.323	8702.665	8712.153

以下，5.1 節で示した空白に関する係数について確認する。空白（space）は，自己ペース読文法では影響がない（図 5・表 6）一方，FFT を除く視線走査法で読み時間が短くなる効果が観察された。

文節間に空白を入れることで，文処理を行う単位が明確になり，被験者の負担が減った可能性がある。しかしこれは，先行研究（松田 2001，Sainio et al. 2007）と異なる結果であった。この違いは先行研究の境界分割単位・空白の大きさ・実験の

表 7　ベイジアン線形混合モデルの結果（First Fixation Time: FFT）

Parameter	Rhat	n_eff	mean	sd	se_mean	2.50%	50%	97.50%
alpha	1.001	4375	5.529	0.063	0.001	5.403	5.529	5.651
beta_length	1.000	20000	−0.001	0.001	0.000	−0.003	−0.001	0.002
beta_space	1.000	20000	−0.012	0.009	0.000	−0.030	−0.013	0.005
beta_dependent	1.000	20000	−0.022	0.005	0.000	−0.032	−0.022	−0.013
beta_sessionN	1.000	20000	0.001	0.009	0.000	−0.017	0.001	0.019
beta_articleN	1.000	8037	−0.008	0.007	0.000	−0.021	−0.008	0.006
beta_screenN	1.000	20000	−0.006	0.004	0.000	−0.014	−0.006	0.002
beta_lineN	1.000	20000	−0.017	0.003	0.000	−0.024	−0.017	−0.011
beta_segmentN	1.000	20000	0.007	0.002	0.000	0.003	0.007	0.012
beta_is_first	1.000	18531	−0.047	0.014	0.000	−0.075	−0.047	−0.020
beta_is_last	1.000	16878	−0.027	0.015	0.000	−0.056	−0.027	0.002
beta_is_second_last	1.000	20000	0.001	0.013	0.000	−0.025	0.001	0.026
sigma	1.000	20000	0.503	0.003	0.000	0.497	0.503	0.509
sigma_article	1.000	8956	0.038	0.010	0.000	0.022	0.036	0.060
sigma_subj	1.000	14496	0.195	0.032	0.000	0.145	0.191	0.269
log-posterior	1.001	6240	2563.044	5.701	0.072	2551	2563.374	2573.255

規模が異なることに由来するのではないかと考える。先行研究においては全角空白を入れていた可能性があり，周辺視野で隣接文節を読むことがより難しくなっている。本研究でも，自己ペース読文法では効果が出ていない。さらに先行研究では実験の規模が小さく（松田 2001：被験者 4 人・28 文字× 62 文，Sainio et al. 2007：被験者 16 人・60 語× 4 テキスト），本研究の規模（被験者 24 人・新聞記事 24 記事）の結果のほうが信頼性が高い。通常の呈示手法では，適切な境界分割に半角空白を入れることにより，視線の停留箇所を制御しながら，周辺視野で右要素を読むことにより，日本語が読みやすくなる可能性が示唆された。本データにはサッケードの情報や視線の停留位置（文字内のオフセット値）は含まれていないために，決定的な分析ではない。今後，サッケードや視線の停留位置のデータを構築し，眼球運動パターンをより詳細に分析する。

最後に 5.2 節で議論した anti-locality 現象について確認する。以上のテキスト呈示の物理的な観点に対する因子とは別に，係り受けの数（dependent）が多いほど読み時間が短くなるという結果（表 5）から，anti-locality 現象が確認できた。これは先行研究で観察された現象を大規模なコーパスを用いて，二重目的語構文や多重入れ子埋め込み節構文ではない，より一般的な環境で確認できたことを意味する。Anti-locality 現象に対する 1 つの批判は，これは新しい効果の観測ではなく，単により多くの文脈が理解を促進しているだけであるというものである。これに対する反論が 2 つ考えられる。1 つは，Uchida et al.（2014）の結果においては，二重目的格構文の間接目的語名詞句の格表示の変更（「に」→「の」）により，動詞の読み時間が遅くなることである。「に」による表示が次に続く述語を制限する役目を果た

表 8　ベイジアン線形混合モデルの結果（First Pass Time: FPT）

Parameter	Rhat	n_eff	mean	sd	se_mean	2.50%	50%	97.50%
alpha	1.000	5942	5.976	0.093	0.001	5.795	5.976	6.162
beta_length	1.000	20000	0.089	0.002	0.000	0.086	0.089	0.092
beta_space	1.000	20000	−0.039	0.011	0.000	−0.062	−0.039	−0.017
beta_dependent	1.000	20000	−0.075	0.006	0.000	−0.087	−0.075	−0.063
beta_sessionN	1.000	20000	−0.051	0.011	0.000	−0.073	−0.051	−0.030
beta_articleN	1.000	9592	−0.010	0.013	0.000	−0.037	−0.010	0.014
beta_screenN	1.000	20000	−0.031	0.005	0.000	−0.041	−0.031	−0.020
beta_lineN	1.000	20000	−0.032	0.004	0.000	−0.040	−0.032	−0.023
beta_segmentN	1.000	20000	−0.011	0.003	0.000	−0.017	−0.011	−0.006
beta_is_first	1.000	20000	−0.209	0.017	0.000	−0.243	−0.209	−0.175
beta_is_last	1.000	16638	0.015	0.018	0.000	−0.021	0.015	0.051
beta_is_second_last	1.000	18132	0.085	0.016	0.000	0.054	0.085	0.116
sigma	1.000	20000	0.627	0.004	0.000	0.619	0.627	0.634
sigma_article	1.000	10248	0.080	0.017	0.000	0.053	0.077	0.120
sigma_subj	1.000	20000	0.302	0.048	0.000	0.226	0.296	0.412
log-posterior	1.000	6731	−379.435	5.666	0.069	−391.372	−379.111	−369.377

していることが示唆される。もう 1 つは，もし anti-locality が文脈に基づく読み速度の促進の特別な場合であるとするならば，どのようにこの促進が行われているのかを特定する必要がある。これこそが，促進が関与するものを量化することによって，先行文脈を含めたさまざまな効果を統一的な枠組みでモデル化するという，サプライザル理論の論点であろう。

表 9　ベイジアン線形混合モデルの結果（Regression Path Time: RPT）

Parameter	Rhat	n_eff	mean	sd	se_mean	2.50%	50%	97.50%
alpha	1.000	5281	5.726	0.100	0.001	5.535	5.724	5.927
beta_length	1.000	20000	0.076	0.002	0.000	0.072	0.076	0.080
beta_space	1.000	20000	−0.041	0.014	0.000	−0.067	−0.041	−0.014
beta_dependent	1.000	20000	−0.061	0.008	0.000	−0.076	−0.061	−0.047
beta_sessionN	1.000	20000	−0.085	0.013	0.000	−0.111	−0.085	−0.058
beta_articleN	1.000	9316	−0.013	0.013	0.000	−0.040	−0.013	0.013
beta_screenN	1.000	20000	−0.028	0.007	0.000	−0.041	−0.028	−0.016
beta_lineN	1.000	20000	−0.014	0.005	0.000	−0.024	−0.014	−0.005
beta_segmentN	1.000	20000	−0.028	0.004	0.000	−0.035	−0.028	−0.021
beta_is_first	1.000	20000	−0.069	0.021	0.000	−0.111	−0.069	−0.027
beta_is_last	1.000	17151	0.187	0.022	0.000	0.144	0.187	0.230
beta_is_second_last	1.000	20000	0.112	0.019	0.000	0.074	0.112	0.150
sigma	1.000	20000	0.759	0.005	0.000	0.750	0.759	0.768
sigma_article	1.000	10663	0.081	0.018	0.000	0.052	0.078	0.122
sigma_subj	1.000	20000	0.309	0.049	0.000	0.230	0.304	0.421
log-posterior	1.000	6705	−2909.52	5.536	0.068	−2921.31	−2909.27	−2899.54

表 10　ベイジアン線形混合モデルの結果（Total Time: Total）

Parameter	Rhat	n_eff	mean	sd	se_mean	2.50%	50%	97.50%
alpha	1.000	6627	6.350	0.095	0.001	6.165	6.350	6.538
beta_length	1.000	20000	0.086	0.002	0.000	0.083	0.086	0.090
beta_space	1.000	20000	−0.067	0.012	0.000	−0.090	−0.067	−0.044
beta_dependent	1.000	20000	−0.081	0.007	0.000	−0.094	−0.081	−0.069
beta_sessionN	1.000	20000	−0.074	0.012	0.000	−0.097	−0.074	−0.052
beta_articleN	1.000	9109	−0.005	0.014	0.000	−0.035	−0.005	0.022
beta_screenN	1.000	20000	−0.044	0.006	0.000	−0.055	−0.044	−0.033
beta_lineN	1.000	20000	−0.033	0.004	0.000	−0.041	−0.033	−0.024
beta_segmentN	1.000	20000	−0.026	0.003	0.000	−0.032	−0.026	−0.020
beta_is_first	1.000	20000	−0.160	0.018	0.000	−0.196	−0.159	−0.124
beta_is_last	1.000	17862	−0.035	0.019	0.000	−0.072	−0.035	0.004
beta_is_second_last	1.000	20000	0.085	0.017	0.000	0.053	0.085	0.118
sigma	1.000	20000	0.658	0.004	0.000	0.650	0.658	0.666
sigma_article	1.000	11964	0.091	0.019	0.000	0.062	0.089	0.136
sigma_subj	1.000	20000	0.296	0.048	0.000	0.220	0.290	0.407
log-posterior	1.000	7046	−1015.49	5.476	0.065	−1027.04	−1015.22	−1005.66

6. おわりに

本研究では，24 人の実験協力者による読み時間を均衡コーパスに対して付与したデータを構築した。データは，実験協力者の言語背景情報・元テキストの情報・呈示時の位置情報・係り受け情報などを付与したうえで，2017 年 3 月に BCCWJ DVD 版購入者に頒布した [10]。

本データの有効性の検証として，文節間の空白の表示と anti-locality 現象の分析を行った。文節間の空白の表示は先行研究は全角空白によるものがほとんどであったが，本研究では，半角単位に視線停留位置を同定し，後に文節単位で再集計することによりこの問題を解決した。結果，先行研究において全角空白単位の文節間空白の表示では漢字かな交じり文で得られなかった読み時間の短縮が，半角空白単位での文節間空白の表示では観察されることが明らかになった。Anti-locality 現象の分析においては，従来二重目的語構文や埋め込み節の入れ子など，限られた構文についてのみ分析が行われてきた。本研究では，均衡コーパスに対して，複数人の読み時間を付与するとともに，文節係り受けを付与したアノテーションを重ね合わせることにより，サプライザル理論を支持する新たな結果を示した。

また，頻度主義的な分析と対照可能にするために，一般化線形混合モデルによる分析結果を付録に示す。従属性があるランダム要因などにおいて，一般化線形混合

[10] 読み時間データのライセンスは Creative Commons 表示－非営利（CC BY-NC）とするが，利用に際しては BCCWJ の契約の範囲に注意すること。

モデルとベイジアン線形混合モデルに違いがあることが見られた。

本データを用いたさまざまな分析が進んでいる。浅原他（2017）は本データに付与されている被験者属性（記憶力テスト結果・語彙数テスト結果）を固定要因とした，一般化線形混合モデルによる分析を行っており，記憶力テストの結果が高い群が，FFT・FPT・RPT の対数読み時間が短い一方，SPT の対数読み時間が長い傾向にあり，全体の対数読み時間（TOTAL）としては差がないことを報告している。また語彙数テスト結果が高い群は FFT を除いて，対数読み時間が長い傾向にあることを報告している。語彙数テスト結果が高い群が読み時間が長くなる理由については，語彙数テストの結果が高いほうが 1 つの語に対する複数の語義や用法を知っており，連接可能性を検討する探索空間が大きくなることが考えられる。今後より詳細な調査が必要である。しかしながら，彼らのモデルでは，一般化線形モデルにおいて収束したモデルを構築するために，記憶力テスト結果・語彙数テスト結果が完全に従属するランダム要因としての被験者属性を外している。ベイジアン線形混合モデルでは被験者属性の 2 変数と被験者間の関係を階層的にモデル化できるが，モデルに対する事前確率と適切なパラメータ推定方法を定義する必要がある。

浅原（2019b）は，本データに節境界アノテーションである BCCWJ-ToriBank（Matsumoto et al. 2018）を重ね合わせ，節境界における読み時間のふるまいをベイジアン線形混合モデルにより検証している。例えば，名詞修飾節において，関係節内の関係と関係節外の関係とで読み時間に差異があることを示している。

浅原・加藤（2019）は，本データに統語／意味分類情報である分類語彙表番号を悉皆付与（加藤他 2019）することで，統語分類や意味分類が読み時間にどのような影響を与えるのかベイジアン線形混合モデルにより検証している。統語分類においては，用の類（類 2.）＜相の類（類 3.）＜体の類（類 1.）の順に読み時間が短い傾向を報告している。意味分類においては，項を持ちうる関係（部門 .1）が他の意味分類に比して読み時間が短い傾向を報告している。

浅原（2018b）は，本データに情報構造アノテーションである BCCWJ-Infostr（宮内他 2018）を重ね合わせることで，情報の新旧が読み時間にどのような影響を与えるのかベイジアン線形混合モデルにより検証している。

浅原（2019a）は，述語項構造（ガ格・ヲ格・ニ格）・外界照応情報（植田他 2015）と重ね合わせを行い，ゼロ代名詞が読み時間にどのような影響を与えるのかについて，一般化線形混合モデルにより検証している。

浅原（2019c）は，大規模コーパスの頻度情報や単語埋め込みに基づくベクトル（word2vec）の情報による読み時間のモデル化手法を提案している。

また，他の種類のテキストに対する読み時間付与を行う。現在，書籍および国語教科書サンプルに対する読み時間付与を進めている（森山他 2019）。

最後に，情報処理的な応用を検討する。リーダビリティの評価や統計的言語モデルの評価など工学応用も視野に入れる。実験協力者には読み時間を得たサンプルを含むテキストに対して，要約文の作文（浅原他 2015）を依頼した。これらのデー

タと対照分析を行うことで，読み時間に基づく利用者ごとの自動要約器の開発を検討する。

付録：線形混合モデルに基づく統計分析

以下では頻度主義者向けに線形混合モデルの結果を傍論として提示するとともに，ベイジアン線形混合モデルとの比較を行う。

まず，なぜベイジアン線形混合モデルなのかについて説明する。

本稿では，従来の読み時間研究で用いられている一般化線形混合モデルではなく，ベイジアン線形混合モデルを用いた。頻度主義的な一般化線形混合モデルは，帰無仮説に基づき，仮説論証型の研究に向いた手法である。しかしながら，心理言語学研究においても，各要因の効果が統計的な差があるかのみならず，モデルがどれだけよく現象を記述できているかについて検討する必要が求められており，研究者はモデル選択に苦労している。一般化線形混合モデルでも，ランダム要因をモデルに組み込むことで，被験者差や呈示サンプル差など，モデルのあてはまりを低下させる要因を吸収することができる。しかしながら，一般化線形混合モデルは，要因同士の関係が非線形であったり，複雑な因果関係を持っている場合には，着目する要因をしぼるなどモデルを単純化することが求められる。そのうえで，要因数を減らした単純化したモデルでも推定できるように，作例を統制するということに研究者は時間をかけてきた。

一方，ベイズ主義的なベイジアン線形混合モデルでは，仮説論証型にも仮説探索型にも用いられる手法である。作例に基づく従来の読み時間の分析においては，検証すべき着目箇所の読み時間の対照比較に基づき，限られた要因で分析されてきた。本研究のような，均衡コーパスに基づく手法では，着目箇所以外も含めた，研究者が予測できない要因を含めて検討する「より一般的」な状況になり，複雑な仮説を柔軟に検証できる手法が必要になる。本稿では，手はじめとして，読み時間と係り受けアノテーションの対照を行った。同データには，節境界・分類語彙表番号・情報構造・述語項構造・共参照情報・否定の焦点などさまざまなレベルのアノテーションが付与されており，これらの複合的な要因の分析が今後求められる。このため，柔軟なモデリングが可能なベイジアン線形混合モデルを用いる。

次に，比較対象の一般化線形混合モデルの統計処理手法について示す。ベイジアン線形混合モデルと同様，本文（タイトル以外の部分）に出現する文節のみを対象とする。具体的には metadata が authorsData，caption，listItem，profile，titleBlock のものを削除した。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。分析が常用対数時間 logtime に対して，R の lme4 パッケージ[11]（Bates et al. 2015）を用いて行った。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，次式に

[11] https://cran.r-project.org/web/packages/lme4/

基づき分析を行った。全てのモデルは収束した。なお，ランダム切片に対する係数の組み合わせによるモデル選択は行っていない。

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

表 11 に線形混合モデルの結果の要約を示す。詳細については，表 12 に示す。表 11 中，かっこ書きで異なる記号が含まれているものは，ベイジアン線形混合モデルとの差異が見られた部分である（かっこ内がベイジアン線形混合モデルの結果）。特徴的なのは，is_first, is_last など，独立性が担保されないところで差異が見られた。is_first=True は，1 行中の位置であるが，単純な segmentN = 1 であるだけでなく，右から左に眼球運動する復帰改行の影響がある。さらに，is_first=True の要素は，常に係り受けの数がゼロ（dependent=0）になる。このように複数の要因と従属しているために，このような差が見られたと考える。しかしながら，空白および anti-locality 現象に関しては，基本的にはベイジアン線形混合モデルと同じ結果となった。

表 11　線形混合モデルの結果の要約（かっこ内はベイジアン線形混合モデルの結果）

	SELF	FFT	FPT	RPT	TOTAL
length	+	0	+	+	+
space	0	0	−	−	−
dependent	−	−	−	−	−
sessionN	0 (+)	0	−	−	−
articleN	−	0	0	0	0
screenN	−	0	−	−	−
linen	−	−	−	−	−
segmentN	−	+	−	−	−
is_first=True	+ (−)	+ (−)	+ (−)	+ (−)	+ (−)
is_last=True	+	− (0)	0	+	− (0)
is_second_last=True	−	0	+	+	+

表 12　線形混合モデルに基づく統計分析結果

Dependent variable: logtime					
	SELF	FFT	FPT	RPT	TOTAL
space1	−0.001	−0.006	−0.018***	−0.019***	−0.030***
	{0.002}	{0.004}	{0.005}	{0.006}	{0.005}
sessionN	−0.022	−0.022	−0.041*	−0.049*	−0.047**
	{0.021}	{0.016}	{0.024}	{0.025}	{0.024}
length	0.089***	−0.001	0.141***	0.120***	0.136***
	{0.001}	{0.002}	{0.003}	{0.003}	{0.003}
dependent	−0.018***	−0.010***	−0.033***	−0.027***	−0.035***
	{0.001}	{0.002}	{0.003}	{0.003}	{0.003}
is_first	0.051***	0.020***	0.091***	0.030***	0.069***
	{0.004}	{0.006}	{0.008}	{0.009}	{0.008}
is_last	0.028***	−0.012*	0.007	0.081***	−0.015*
	{0.004}	{0.006}	{0.008}	{0.010}	{0.008}
is_second_last	−0.008**	0.0003	0.037***	0.049***	0.037***
	{0.004}	{0.006}	{0.007}	{0.008}	{0.007}
articleN	−0.028***	−0.004	−0.005	−0.007	−0.002
	{0.006}	{0.004}	{0.007}	{0.007}	{0.008}
screenN	−0.030***	−0.004	−0.018***	−0.017***	−0.026***
	{0.002}	{0.003}	{0.003}	{0.004}	{0.003}
lineN	−0.011***	−0.010***	−0.019***	−0.009***	−0.020***
	{0.001}	{0.002}	{0.003}	{0.003}	{0.003}
segmentN	−0.004***	0.003***	−0.005***	−0.012***	−0.011***
	{0.001}	{0.001}	{0.001}	{0.002}	{0.001}
space1:sessionN	−0.016	0.044	0.059	0.061	0.061
	{0.042}	{0.031}	{0.048}	{0.049}	{0.047}
Constant	2.784***	2.305***	2.535***	2.602***	2.674***
	{0.022}	{0.017}	{0.026}	{0.027}	{0.026}
Observations	17,628	13,232	13,232	13,232	13,232
Log Likelihood	7,054.93	1,304.77	−1,626.57	−4,149.55	−2,260.45
Akaike Inf. Crit.	−14,077.85	−2,577.54	3,285.15	8,331.11	4,552.89
Bayesian Inf. Crit.	−13,953.42	−2,457.69	3,405.00	8,450.95	4,672.74

Note: *p<0.1; **p<0.05; ***p<0.01

付録：STAN のコード

```
data {
    int<lower=0> N;          // # of datapoints

    int<lower=0> N_article;   // # of articles
    int<lower=0> N_subj;      // # of subject
    int<lower=0,upper=N_article> article[N];  // article ID
    int<lower=0,upper=N_subj> subj[N];  // subject ID

    int length[N];            //
    int space[N];
    int dependent[N];

    int sessionN[N];
    int articleN[N];
    int screenN[N];
    int lineN[N];
    int segmentN[N];

    int is_first[N];
    int is_last[N];
    int is_second_last[N];

    real time[N]; // time
}

parameters {
    // intercept
    real alpha;

    // slopes for fixed effect
    real beta_length;
    real beta_space;
    real beta_dependent;

    real beta_sessionN;
    real beta_articleN;
    real beta_screenN;
```

```
    real beta_lineN;
    real beta_segmentN;

    real beta_is_first;
    real beta_is_last;
    real beta_is_second_last;

    // random effect
    vector[N_article] gamma_article; // article intercept
    vector[N_subj] gamma_subj; // subject intercept

    // standard deviation for Lognormal and Normal
    real<lower = 0> sigma; // error SD
    real<lower = 0> sigma_article; // article SD
    real<lower = 0> sigma_subj; // subj SD
}

model {
    real mu;
    // prior
    gamma_article ~ normal(0,sigma_article);
    gamma_subj ~ normal(0,sigma_subj);

    // likelihood
    for (k in 1:N) {
      mu = alpha +
        beta_length * length[k] +
        beta_space * space[k] +
        beta_dependent * dependent[k] +
        beta_sessionN * sessionN[k] +
        beta_articleN * articleN[k] +
        beta_screenN * screenN[k] +
        beta_lineN * lineN[k] +
        beta_segmentN * segmentN[k] +
        beta_is_first * is_first[k] +
        beta_is_last * is_last[k] +
        beta_is_second_last * is_second_last[k] +
        gamma_article[article[k]] + gamma_subj[subj[k]] ;
```

```
    time[k] ~ lognormal(mu,sigma);
  }
}
```

参照文献

Amano, Shigeaki and Tadahisa Kondo (1998) Estimation of mental lexicon size with word familiarity database. *Proceedings of International Conference on Spoken Language Processing* 5: 2119–2122.

天野成昭・近藤公久（編）・NTT コミュニケーション科学基礎研究所（監修）（1999）『NTT データベースシリーズ　日本語の語彙特性　第 1 巻　単語親密度』三省堂.

浅原正幸（2018）「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.

浅原正幸・松本裕治（2018）「『現代日本語書き言葉コーパス』に対する文節係り受け・並列構造アノテーション」『自然言語処理』25(4): 331–356.

浅原正幸（2019a）「読み時間と述語項構造・共参照情報について」『言語処理学会第 25 回発表論文集』: 249–252.

浅原正幸（2019b）「日本語の読み時間と節境界情報―主辞後置言語における wrap-up effect の検証」『自然言語処理』26(2), 301–327.

浅原正幸（2019c）「単語埋め込みに基づくサプライザル」『自然言語処理』26(3): 635–652.

浅原正幸・加藤祥（2019）「読み時間と統語・意味分類」『認知科学』26(2), 219–230.

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』: 473–477.

浅原正幸・杉真緒・柳野祥子（2015）「BCCWJ-SUMM：『現代日本語書き言葉均衡コーパス』を元文書とした要約文書コーパス」『第 7 回コーパス日本語学ワークショップ』: 285–292.

Bates, Douglas, Martin Mächler, Ben Bolker, Steven Walker (2015). Fitting Linear Mixed-effects Models using lme4. *Journal of Statistical Software*. 67(1): 1–48.

Barrett, Maria and Anders Søgaard (2015a) Reading behavior predicts syntactic categories. *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*: 345–349, Association for Computational Linguistics, URL: http://www.aclweb.org/anthology/K15-1038.

Barrett, Maria and Anders Søgaard (2015b) Using reading behavior to predict grammatical functions. *Proceedings of the Sixth Workshop on Cognitive Aspects of Computational Language Learning*: 1–5, Association for Computational Linguistics, URL: http://aclweb.org/anthology/W15-2401.

Barrett, Maria Jung, Zeljko Agic, and Anders Søgaard (2015) The Dundee Treebank. *Proceedings of the Fourteenth International Workshop on Treebanks and Linguistic Theories*: 242–248.

Clifton Jr., Charles, Adrian Staub, and Keith Rayner (2007) Eye movements: A window on mind and brain. *Eye movements in reading words and sentences*: 341–372, Amsterdam: Elsevier.

Demberg, Vera and Frank Keller (2007) Eye-tracking evidence for integration cost effects in corpus data. *Proceedings of the 29th meeting of the cognitive science society (CogSci-07)*: 947–952.

Demberg, Vera and Frank Keller (2008) Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition* 109(2): 193–210.

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

Frank, Stefan L. (2009) Surprisal-based comparison between a symbolic and a connectionist model of sentence processing. *31st Annual Conference of the Cognitive Science Society (CogSci 2009)*: 1139–1144.

Frank, Stefan L., Irene Fernandez Monsalve, Robin L. Thompson, and Gabriella Vigliocco (2013) Reading time data for evaluating broad-coverage models of English sentence processing. *Behaviour Research Methods* 45(4): 1182–1190.

Futrell, Richard, Edward Gibson, Harry J. Tilly, Idan Blank, Anastasia Vishnevetsky, Steven T. Piantadosi, and Evelena Fedorenko (2018) The natural stories corpus. *11th edition of the Language Resources and Evaluation Conference*: 76–82.

Gelman, Andrew and Jennifer Hill (2006) *Data analysis using regression and multi-level/hierarchical models*. New York: Cambridge University Press.

Gibson, Edward (1998) Linguistic complexity: Locality of syntactic dependencies. *Cognition* 68: 1–76.

van Gompel, Roger P. G., Martin H. Fischer, Wayne S. Murray, and Robin L. Hill (2007) Eye-movement research: An overview of current and past developments. *Eye movements: A window on mind and brain*: 1–28.

Hale, John (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* 2: 159–166.

Husain, Samar, Shravan Vasishth, and Narayanan Srinivasan (2014) Strong expectations cancel locality effects: Evidence from Hindi. *PLoS One* 9(7): 1–14.

Husain, Samar, Shravan Vasishth, and Narayanan Srinivasan (2015) Integration and prediction difficulty in Hindi sentence comprehension: Evidence from an eye-tracking corpus. *Journal of Eye Movement Research* 8(2): 1–12.

Just, Marcel A., Patricia A. Carpenter, and Jacqueline D. Woolley (1982) Paradigms and processes in reading comprehension. *Journal of Experimental Psychology: General* 3: 228–238.

加藤祥・浅原正幸・山崎誠（2019）「分類語彙表番号を付与した『現代日本語書き言葉均衡コーパス』の書籍・新聞・雑誌データ」『日本語の研究』15: 134–141.

Kennedy, Alan, Robin Hill, and Joël Pynte (2003) The Dundee Corpus. *Proceedings of the 12th European Conference on Eye Movement*.

Kliegl, Reinhold, Antje Nuthmann, and Ralf Engbert (2006) Tracking the mind during reading: The influence of past, present, and future words on fixation durations. *Journal of Experimental Psychology: General* 135(1): 12–35.

Kohsom, Chananda and Fernand Gobet (1997) Adding spaces to Thai and English: Effects on reading. *Proceedings of the Cognitive Science Society 19*: 383–393.

Konieczny, Lars (2000) Locality and parsing complexity. *Journal of Psycholinguistic Research* 29(6): 627–645.

Konieczny, Lars and Phillipp Döring (2003) Anticipation of clause-final heads. Evidence from eye-tracking and SRNs. *Proceedings of the 4th International Conference on Cognitive Science*. 330–335.

Levy, Roger (2008) Expectation-based syntactic comprehension *Cognition* 106: 1126–1177.

Levy, Roger and Frank Keller (2013) Expectation and locality effects in German verb-final structures. *Journal of Memory and Language* 68: 199–222.

Luong, Minh-Thang, Timothy J. O'Donnell, and Noah D. Goodman (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, Kikuo, 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: 345–371.

松田真幸（2001）「日本語文の読みに及ぼす文節間空白の影響」『基礎心理学研究』19(2): 83–92.

Matsumoto, Satomi, Masayuki Asahara, and Setsuko Arita (2018) Japanese clause classification annotation on the 'Balanced Corpus of Contemporary Written Japanese'. *Proceedings of Workshop on Asian Language Resources 13 (ALR13)*: 1–8.

松吉俊（2014）「否定の焦点情報アノテーション」『自然言語処理』21(2): 249–270.

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』: 699–702.

宮内拓也・浅原正幸・中川奈津子・加藤祥（2018）「『現代日本語書き言葉均衡コーパス』への情報構造アノテーションとその分析」『国立国語研究所論集』16: 19–33.

Nakatani, Kentaro and Edward Gibson (2010) An On-Line of Japanese Nesting Complexity. *Cognitive Science* 34(1): 94–112.

苧坂満里子（編）（2002）『ワーキングメモリー脳のメモ帳』新曜社.

Patterson, Clare and Janna Drummer (2016) EyeTracking – Focus: eyetracking during reading. *Linguistischer Methodenworkshop* (HU Berlin).

Pynte, Joël and Alan Kennedy (2007) The influence of punctuation and word class on distributed processing in normal reading. *Vision Research* 47(9): 1215–1227.
Pynte, Joël, Boris New, and Alan Kennedy (2008) On-line contextual influences during reading normal text: A multiple-regression analysis. *Vision Research* 48(21): 2172–2183.
Rayner, Keith, Martin Fischer, and Alexander Pollatsek (1998) Unspaced text interferes with both word identification and eye movement control. *Vision Research* 38(8): 1129–1144.
Roland, Douglas, Gail Mauner, Carolyn O'Meara, and Hongoak Yun (2012) Discourse expectations and relative clause processing. *Journal of Memory and Language* 66(3): 479–508.
Rouder, Jeffery N. (2005) Are unshifted distributional models appropriate for response time. *Psychometrika* 70: 377–381.
Sainio, Miia, Jukka Hyönä, Kazuo Bingushi Raymond Bertram (2007) The role of inter spacing in reading Japanese: An eye movement study. *Vision Research* 47: 2575–2584.
van Schijndel, Marten and William Schuler (2013) An analysis of frequency- and memory-based processing costs. *Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*: 95–105.
Seminck, Olga and Pascal Amilli (2017) A computational model of human preferences for pronoun resolution. *Proceedings of the Student Research Workshop at the 15th Conference of the European Chapter of the Association for Computational Linguistics*: 53–63.
Seminck, Olga and Pascal Amilli (2018) A gold anaphora annotation layer on an eye movement corpus. *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*: 3518–3522.
Smith, Nathaniel J. and Roger Levy (2008) Optimal processing times in reading: A formal model and empirical investigation. *Proceedings of the 30th Annual Conference of the Cognitive Science Society*: 595–600.
Smith, Nathaniel J. and Roger Levy (2013) The effect of word predictability on reading time is logarithmic. *Cognition* 128(3): 302–319.
Sorensen, Tanner, Sven Hohenstein, and Shravan Vasishth (2016) Bayesian linear mixed models using Stan: A tutorial for psychologists, linguists, and cognitive scientists. *Quantitative Methods for Psychology* 12: 175–200.
Uchida, Shodai, Edson T. Miyamoto, Yuki Hirose, Yuki Kobayashi, and Takane Ito (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*.
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』: 205–214.
Universal Dependencies (2014) Universal Dependencies. https://universaldependencies.github.io/docs/.
Vasishth, Shravan and Richard L. Lewis (2006) Argument-head distance and processing complexity: Explaining both locality and antilocality effects. *Language* 82(4): 767–794.
Vasishth, Shravan and Heiner Drenhaus (2011) Locality in German. *Dialogue and Discourse* 2(1): 59–82.
Winskel, Heather, Ralph Radach, and Sudaporn Luksaneeyanawin (2009) Eye movements when reading spaced and unspaced Thai and English: A comparison of Thai-English bilinguals and English monolinguals. *Eye movements when reading spaced and unspaced Thai and English: A comparison of Thai-English bilinguals and English monolinguals* 61: 339–351.
Yan, Ming, Reinhold Kliegl, Eike Martin Richter, Antje Nuthmann, and Hua Shu (2010) Flexible saccade target selection in Chinese reading. *The Quarterly Journal of Experimental Psychology* 63(4): 705–725.

執筆者連絡先：
浅原正幸
190-8561 東京都立川市緑町 10-2
国立国語研究所　コーパス開発センター
TEL: 042-540-4300　e-mail: masayu-a@ninjal.ac.jp

［受領日　2017 年 8 月 11 日
最終原稿受理日　2019 年 3 月 7 日］

Abstract

BCCWJ-EyeTrack: Reading Time Annotation on the 'Balanced Corpus of Contemporary Written Japanese'

Masayuki Asahara
National Institute for Japanese Language and Linguistics

Hajime Ono
Tsuda University

Edson T. Miyamoto
Future University Hakodate

We report on a new Japanese corpus of eye-tracking data. The corpus design is partly modeled on the Dundee Eye-Tracking Corpus for English and French texts, but it addresses language-specific issues such as the lack of segmentation spaces in Japanese texts. Twenty-four native Japanese speakers read excerpts from the Balanced Corpus of Contemporary Written Japanese, presented with or without a space between segments. Segments were based on *bunsetsu* units (a content word plus functional material). Two types of methodologies were used for data collection: eye-tracking and self-paced reading. We report two analyses to illustrate the advantages of having such a large reading-time data set for texts that have annotations such as syntactic-dependency relations. First, contrary to previous eye-tracking reports based on relatively small sets of sentences, texts segmented with spaces were read more quickly than the same texts presented without spaces. Second, across the various types of sentences in the corpus, reading a *bunsetu* was faster the more it was preceded by dependent phrases. This evidence for anti-locality effects, is more general than what was thus far available in the literature.