一般論文

単語埋め込みに基づくサプライザル

浅原　正幸 †

ヒトの文処理のモデル化として Hale によりサプライザルが提案されている．サプライザルは文処理の負荷に対する情報量基準に基づいた指標で，当該単語の文脈中の負の対数確率が文処理の困難さをモデル化するとしている．日本語において眼球運動測定を用いて文処理の負荷をモデル化する際に，統語における基本単位である文節単位の読み時間を集計する．一方，単語の文脈中の生起確率は形態素や単語といった単位で評価し，この齟齬が直接的なサプライザルのモデル化を難しくしていた．本論文では，この問題を解決するために単語埋め込みを用いる．skip-gram の単語埋め込みの加法構成性に基づき，文節構成語のベクトルから文節のベクトルを構成し，隣接文節間のベクトルのコサイン類似度を用いて，文脈中の隣接尤度をモデル化できることを確認した．さらに，skip-gram の単語埋め込みに基づいて構成した文節のベクトルのノルムが，日本語の読み時間のモデル化に寄与することを発見した．

キーワード：リーダビリティ評価，読み時間，単語埋め込み，サプライザル

Surprisal through Word Embeddings

Masayuki Asahara †

The concept of surprisal was proposed by Hale as a psycholinguistic model of sentence processing costs based on the information theory. Surprisal measures a word's negative log probability in context and can be used to model the difficulty in processing a sentence. If this difficulty is estimated using the eye-tracking method, the reading time can be estimated using base phrase units in Japanese. In addition, word probability is estimated from the frequency of morphemes or word units in Japanese. We introduced word embeddings to address the discrepancy in units, which makes it difficult to model surprisal in Japanese. The additive property of skip-gram word embeddings enabled us to compose a base phrase vector from word vectors in the base phrase. We confirmed that the cosine similarity between two adjacent base phrase vectors can be used to model the contextual probability of the bi-gram of the base phrase and found that the norm of the base phrase correlates with reading time in Japanese.

Key Words: *Readability, Reading Time, Word Embeddings, Surprisal*

† 人間文化研究機構国立国語研究所, NINJAL, Japan

1 はじめに

本研究では眼球運動に基づき文の読み時間を推定し，ヒトの文処理機構の解明を目指すとともに，工学的な応用として文の読みやすさのモデル構築を行う．対象言語は日本語とする．

データとして2.1節に示す『現代日本語書き言葉均衡コーパス』(BCCWJ) (Maekawa, Yamazaki, Ogiso, Maruyama, Ogura, Kashino, Koiso, Yamaguchi, Tanaka, and Den 2014) の読み時間データ BCCWJ-EyeTrack (浅原，小野，宮本 2019) を用いる．2.4節に示す通り，過去の研究は統語・意味・談話レベルのアノテーションを重ね合わせることにより，コーパス中に出現する言語現象と読み時間の相関について検討してきた．一方，Hale は，言語構造の頻度 (Structural frequency) が文処理過程に影響を与えると言及し，漸進的な文処理の困難さについて情報量基準に基づいたモデルをサプライザル理論 (Surprisal Theory) として定式化している (Hale 2001)．このサプライザル理論に基づく日本語の読み時間の分析が求められている．

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

この齟齬を吸収するために，単語埋め込み (Mikolov, Chen, Corrado, and Dean 2013a) の利用を提案する．単語埋め込みは前後文脈に基づき構成することにより，単語の置き換え可能性を低次元の実数値ベクトル表現によりモデル化する．このうち skip-gram モデルは加法構成性を持つと言われ[1]，句を構成する単語のベクトルの線形和が，句の置き換え可能性をモデル化できる (Mikolov et al. 2013b)．

日本語の単語埋め込みとして，『国語研日本語ウェブコーパス』(NWJC)(Asahara, Maekawa, Imada, Kato, and Konishi 2014) から fastText (Bojanowski, Grave, Joulin, and Mikolov 2017) により構成した NWJC2vec (Asahara 2018b) を用いた．ベイジアン線形混合モデル (Sorensen, Hohenstein, and Vasishth 2016) に基づく統計分析[2]の結果，skip-gram モデルに基づく単語埋め込みのノルムと隣接文節間のコサイン類似度が，読み時間を予測する因子となりうることが分かった．前者のノルムが読み時間を長くする文節の何らかの特性を，後者の隣接文節間のコサイン類似度が隣接尤度をモデル化すると考える．「隣接尤度」は文節の bigram 隣接尤度のようなものを想定する．

以下，2節に前提となる関連情報について示す．3節に分析手法について示す．4節に結果と考察について示し，5節でまとめと今後の展開を示す．

[1] 原論文 (Mikolov, Sutskever, Chen, Corrado, and Dean 2013b) 5節 Additive Compositionality を参照．
[2] 本研究では，複雑な要因分析の際にモデルの収束が容易なベイズ主義的な統計分析を行う．頻度主義的な統計分析を用いない理由については，『言語研究』論文 (浅原 他 2019) の付録を参照されたい．

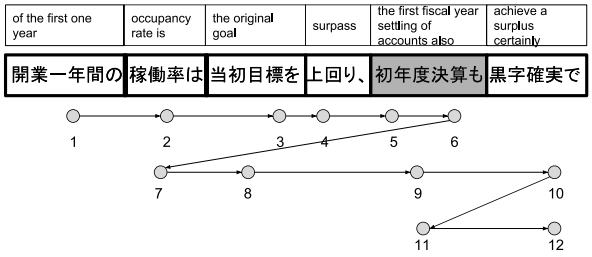
2 前提

2.1 BCCWJ-EyeTrack

BCCWJ-EyeTrack (浅原 他 2019) は，BCCWJ の新聞記事サンプル 20 記事に対して，日本語母語話者 24 人分（女性 19 人, 未回答 1 人, 男性 4 人; 20–55 歳）の読み時間を収集して，データベース化したものである．自己ペース読文法 (SELF: Self-Paced Reading) と視線走査法に基づく文節単位の 5 種類の読み時間 (FFT: First Fixation Time, FPT: First Pass Time, SPT: Second Pass Time, RPT: Regression Path Time, Total: Total Time) が視線停留オフセット値に基づいて算出されている．自己ペース読文法は，実験協力者がスペースキーを押しながら 1 文節ずつ読む方法で，スペースキーを押す時間間隔が当該文節を読む時間として計測される．基本的に後戻りして読むことはできない．視線走査法は，視線走査装置を用いて眼球運動を計測することにより，視線停留時間から読み時間を直接評価する手法である．図 1 に読み時間のタイプの集計例を示す．

表 1 に BCCWJ-EyeTrack のデータ形式について示す．`surface` は単語の表層形である．読み時間 (i.e., `time`) は対数に変換したデータ (i.e., `logtime`) も保持し，一般化線形混合モデル用に用いられる．`measure` は読み時間のタイプ {SELF, FFT, FPT, RPT, SPT, TOTAL} を表す．`sample, article, metadata_orig, metadata` は記事に関連する情報である．`space` は文節境界に半角空白を入れたか否かを示す．`length` は表層形の文字数である．`is_first,is_last,is_second_last` はレイアウトに関する特徴量である．`sessionN, articleN, screenN, lineN, segmentN` は要素の呈示順に関する特徴量である．`subj` は被験者の ID で統計処理においてランダム要因として用いる．`dependent` は当該文節に係る文節の数を人手で付与したもの (浅原，松本 2018) である．

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)



対象領域を「初年度決算も」の文節とする：
FFT は視線停留 5
FPT は視線停留 5 と 6 の総計.
SPT は視線停留 9, 11 の総計
RPT は視線停留 5, 6, 7, 8, 9 の総計
TOTAL は視線停留 5, 6, 9, 11 の総計

図 1 視線走査データの読み時間のタイプの集計例

表 1　BCCWJ-EyeTrack のデータ形式

列名	データ型	摘要
surface	factor	出現書字形
time	int	読み時間
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
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
space	factor	文節境界空白の有無
subj	factor	実験協力者 ID
length	int	文字数
dependent	int	係り受け関係

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)[3].

2.2　サプライザル

Hale (2001) は文脈中に出現する言語的な事象 x（音韻的特徴・単語・発話）が伝達する情報を次式によりはかることができ，これをサプライザル (surprisal)[4] と呼んだ：

$$\texttt{Surprisal}(x) = \log_2 \frac{1}{P(x|\texttt{context})}$$

Surprisal は x の（文脈 context による条件付き）確率が低い場合に大きい値をとり，確率が高い場合に小さい値をとる．さらに単語を処理する認識努力 (cognitive effort) はその surprisal に比例するとしている：

$$\texttt{Effort} \propto \texttt{Surprisal}$$

Surprisal は前方部分単語列に基づいて選好される parse 木を再考するコストとともに後方部分

[3] 自己ペース読文法は，読み戻しができないために正解率が低くなったと考える．
[4] 以下，サプライザル理論一般を表す場合にカタカナ表記で，個々の式を表す場合にアルファベット表記を用いる．

単語列を期待しうるか否かの困難さをモデル化する．Surprisal は，確率的言語モデルに基づくもの[5]・N-gram Surprisal[6]・Parser (PCFG) Surprisal[7]などがあり，Hale (2001) は Earley 法に基づく Parser Surprisal を提案した．Levy (2008) は，前方部分単語列に対する可能な parse 木の確率分布を反映する KL ダイバージェンスに基づく surprisal[8]を提案した．

また Pynte らは Latent Semantic Analysis (LSA) (Landauer and Dumais 1997) による分散表現を用いた semantic surprisal (Pynte, New, and Kennedy 2008) を提案し，英語の視線走査データである Dundee Corpus のモデル化を行っている．具体的には，300 次元の LSA モデルを用いて， wLSA[9]-based surprisal （前隣接単語と注目単語との意味的類似度）と sLSA[10]-based surprisal （従前に出現する前隣接単語以外の単語と注目単語の意味的類似度）について一般化線形混合モデルにより調査した．wLSA-based surprisal は読み時間を予測する効果が確認されたが，sLSA-based surprisal には効果が確認できなかった．Mitchell らは，Latent Direchret Allocation (LDA) (Blei, Ng, and Jordan 2003; Griffiths, Steyvers, and Tanenbaum 2007) による分散表現を用いた LDA-based surprisal を提案した (Mitchell, Lapata, Demberg, and Keller 2010).

2.3 単語埋め込みと NWJC2vec

単語埋め込み (Mikolov et al. 2013a) は単語を数百次元のベクトルで表現する技術である．従来はその単語か否かを表す one-hot 表現が用いられていたため，大規模語彙を表現するために高次元ベクトルになっていた．学習の際のモデルとして，文脈から単語を推定する CBOW モデルと単語から文脈を推定する skip-gram モデルが提案されている．

単語埋め込みにより，単語の入れ替え可能性を低次元のベクトルで表現できるようになったほか，skip-gram モデルには加法構成性と呼ばれる句を構成する語ベクトルの和が，句ベクトルとして利用できるという良い性質を持つ．本研究ではこの性質を，日本語における語を計数する単位と読み時間を評価する単位の齟齬の吸収に用いる．

NWJC2vec (Asahara 2018b) は NWJC 258 億語から訓練した日本語の単語埋め込みデータである．fastText (Bojanowski et al. 2017) を用いて訓練した 300 次元の CBOW および skip-gram モデル[11]を用いる．この学習した単語ベクトルを用いて，視線走査データの集計単位である文節単位のベクトルを合成する．合成には線形和を用いた．

[5] 確率的言語モデル：$\texttt{Surprisal}_{k+1} = -\log_2 P(w_{k+1}|w_1 \dots w_k)$
[6] N-gram：$\texttt{Surprisal}_{w_{k+1}} = -\log_2 P(w_{k+1}|w_{k-2}, w_{k-1}, w_k)$
[7] PCFG：$\texttt{Surprisal}_n = -\log_2 P(Ti, w_i|T_1 \dots T_{i-1}, w_1 \dots w_{i-1})$
[8] KL ダイバージェンス：$\texttt{Surprisal}_{k+1} = D(P_{k+1}||P_k) = -\log P(w_{k+1}|w_1 \dots w_k)$
[9] word level LSA
[10] sentence level LSA
[11] CBOW か skip-gram か以外のオプションは次の通り：`-size 300 -window 8 -negative 25 -hs 0 -sample 1e-4 -iter 15`

2.4 BCCWJ-EyeTrack の過去の分析

BCCWJ-EyeTrack に対して，統語・意味・談話レベルのアノテーションを重ね合わせて，様々な言語現象に対してヒトがどのような反応をするのかについて検討を進めてきた．浅原，小野，宮本 (2017) は被験者属性を対象とし，記憶力がある群は読む速度が速いが全読み時間は記憶力がない群と変わらないこと，語彙力がある群が読み時間が長いことを明らかにした．浅原ら (2019) は文節係り受けアノテーション BCCWJ-DepPara (浅原，松本 2018) と対照比較を行い，係り受けの数が多い文節ほど読み時間が短くなることを明らかにした．浅原 (2019) は節情報アノテーション BCCWJ-ToriClause (Matsumoto, Asahara, and Arita 2018) と対照比較を行い，節末の読み時間が短いことを明らかにした．浅原，加藤 (2019) は分類語彙表番号アノテーション BCCWJ-WLSP (加藤，浅原，山崎 2019) と対照比較を行い，統語分類の「用の類」<「相の類」<「体の類」の順で読み時間が長くなる傾向と，意味分類の「関係」が他の分類（「主体」「活動」「生産物」「自然物」）と比べて読み時間が短くなる傾向を明らかにした．浅原 (2018) は情報構造アノテーション BCCWJ-Infostr (宮内，浅原，中川，加藤 2018) と対照比較を行い，共有性において旧情報 (hearer-old) が新情報 (hearer-new) よりも読み時間が短いことを明らかにした．Asahara (2018a) は述語項構造・共参照情報アノテーション BCCWJ-PAS (植田，飯田，浅原，松本，徳永 2015; 浅原，大村 2016) と対照比較を行い，主語がゼロ代名詞の際に外界照応として二人称を指す場合の述語において，SPT が短くなることを明らかにした．これらの分析には，サンプルと被験者をランダム要因とし，アノテーションを固定要因とした対数時間に対する一般化線形混合モデルかベイジアン線形混合モデル (Sorensen et al. 2016) に基づく方法を用いている．

3 分析手法

分析においては，いくつかの要因に基づく線形式に基づいて，読み時間をベイジアン線形混合モデル (Sorensen et al. 2016) により推定し，その係数を見ることにより進める．図 2 に推定に用いた線形式を示す．分析は，分散表現のノルムと隣接文節類似度に基づくもの ($\mu_{\texttt{wv}}$)，頻度情報に基づく当該文節の出現確率のみに基づくもの ($\mu_{\texttt{freq}}$)，両方を用いたものに基づくもの ($\mu_{\texttt{all}}$) を対比する．分散表現は 2.3 節に示した fastText に基づく NWJC2vec 300 次元のものを用い，CBOW モデルと skip-gram モデルの両方を比較した．対象とする読み時間は，読み戻しが可能な視線走査法の FFT, FPT, RPT, TOTAL とする．SPT については，付録 B 節に述べる．

まず読み時間 `time` を対数正規分布 lognormal によりモデル化し，期待値を μ_*，分散を σ とする．式において $\mu_{\texttt{base}}$ は基本的な要因を表し，α を切片とする．$\beta^{\texttt{length}}$ は文節の文字数に対する係数，$\beta^{\texttt{space}}$ は文節間に半角空白を入れたか否かの係数である．$\beta^{\texttt{sessionN}}$, $\beta^{\texttt{articleN}}$, $\beta^{\texttt{screenN}}$, $\beta^{\texttt{lineN}}$,

$$
\begin{aligned}
\texttt{time} &\sim lognormal(\mu_*, \sigma) \\
\mu_{\texttt{base}} &= \alpha + \beta^{\texttt{length}} \cdot \texttt{length}(x) + \beta^{\texttt{space}} \cdot \chi_{\texttt{space}}(x) + \beta^{\texttt{dependent}} \cdot \texttt{dependent}(x) + \beta^{\texttt{sessionN}} \cdot \texttt{sessionN} \\
&\quad + \beta^{\texttt{articleN}} \cdot \texttt{articleN}(x) + \beta^{\texttt{screenN}} \cdot \texttt{screenN}(x) + \beta^{\texttt{lineN}} \cdot \texttt{lineN}(x) + \beta^{\texttt{segmentN}} \cdot \texttt{segmentN}(x) \\
&\quad + \beta^{\texttt{is\_first}} \cdot \chi_{\texttt{is\_first}}(x) + \beta^{\texttt{is\_last}} \cdot \chi_{\texttt{is\_last}}(x) + \beta^{\texttt{is\_second\_last}} \cdot \chi_{\texttt{is\_second\_last}}(x) \\
&\quad + \sum_{a(x) \in A} \gamma^{\texttt{article}=a(x)} + \sum_{s(x) \in S} \gamma^{\texttt{subj}=s(x)}. \\
\mu_{\texttt{wv\_norm}} &= \mu_{\texttt{base}} + \beta^{\texttt{wm\_norm}} \cdot \texttt{wv\_norm}(x). \\
\mu_{\texttt{wv\_sim}} &= \mu_{\texttt{base}} + \beta^{\texttt{wm\_sim}} \cdot \texttt{wv\_sim}(x). \\
\mu_{\texttt{wv\_all}} &= \mu_{\texttt{base}} + \beta^{\texttt{wm\_norm}} \cdot \texttt{wv\_norm}(x) + \beta^{\texttt{wm\_sim}} \cdot \texttt{wv\_sim}(x). \\
\mu_{\texttt{freq}} &= \mu_{\texttt{base}} + \beta^{\texttt{freq\_ave}} \cdot \texttt{freq\_ave}(x). \\
\mu_{\texttt{all}} &= \mu_{\texttt{base}} + \beta^{\texttt{freq\_ave}} \cdot \texttt{freq\_ave}(x) + \beta^{\texttt{wm\_norm}} \cdot \texttt{wv\_norm}(x) + \beta^{\texttt{wm\_sim}} \cdot \texttt{wv\_sim}(x).
\end{aligned}
$$

図 2 推定に用いた線形式

$\beta^{\texttt{segmentN}}$ が呈示順に対する係数，$\beta^{\texttt{is\_first}}$, $\beta^{\texttt{is\_last}}$, $\beta^{\texttt{is\_second\_last}}$ がレイアウト情報に対する係数である．その他，記事に対するランダム係数として $\gamma^{\texttt{article}=a(x)}$ を，実験協力者に対するランダム係数として $\gamma^{\texttt{subj}=s(x)}$ を考慮する．このランダム係数により記事間の揺れと実験協力者間の揺れを，それぞれ $\sigma_{article}$，σ_{subj} を標準偏差とする正規分布によりモデル化することにより吸収する．

$\beta^{\texttt{dependent}}$ は当該文節に係る文節の数に対する係数である．先行研究においては，PCFG の部分木により統語構造に基づく効果をモデル化していた．日本語においては，比較的語順が自由な言語であるために句構造木ではなく，文節係り受け木により評価する．日本語は主辞後置言語であり，当該文節に係る文節は基本的に全て前置することから，この係数によって実質的に統語構造に基づく効果がモデル化できると考える [12]．

単語ベクトルから構成した文節ベクトルの情報の二つの情報を用いる．一つは当該文節ベクトルのノルム $\texttt{wv\_norm}(x)$ である．もう一つは当該文節ベクトルと左隣接ベクトルのコサイン類似度 $\texttt{wv\_sim}(x)$ である．この上で，単語埋め込みを考慮した期待値として $\mu_{\texttt{wv\_all}}$ を検討する．$\beta^{\texttt{wm\_norm}}$ は単語埋め込みに基づく文節ベクトルのノルムに対する係数，$\beta^{\texttt{wm\_sim}}$ は単語埋め込みに基づく左隣接文節とのコサイン類似度に対する係数であり，これを評価する．分散表現のモデルとして，CBOW と skip-gram の二つを評価する．また，比較のためにノルムのみのもの $\mu_{\texttt{wv\_norm}}$ とコサイン類似度のみのもの $\mu_{\texttt{wv\_sim}}$ も評価する．

比較対照として，単語の頻度を考慮した期待値として $\mu_{\texttt{freq}}$ を検討する．$\beta^{\texttt{freq\_ave}}$ は文節内頻度に対する係数である．単語の頻度に基づく手法については，文節間の連接尤度を考慮しない．文節内頻度は文節内の単語の頻度の相乗平均を評価する [13]．相乗平均を評価する際にゼロ

[12] 詳細については『言語研究』論文 (浅原 他 2019) の付録を参照されたい．
[13] 頻度を確率値とした場合に連接単語数で正規化した対数線形モデルに相当し，Surprisal の式と親和性が良いと考える．

頻度は 1 を乗じた．なお，相加平均でも評価したがモデルが収束しなかった．

最後に，単語埋め込みと単語の頻度の両方を考慮した $\mu_{\texttt{all}}$ を検討する．

分析においては，RStan[14] を用いた．500 iter の warm up のあと，5000 iter を 4 chains 行った．

4　結果と考察

表 2 に各モデルの分析結果を示す．詳細な結果については，付録 A 節に示す．

推定される mean が 0.00 から ±2 SD 以上の差があるものに + もしくは − を付与する．0 は mean が ±2 SD 以内のものである．+ はその値が大きければ，読み時間が長くなることを示す．− はその値が大きければ，読み時間が短くなることを示す．

表 2　分析結果（概要）

			FFT	FPT	RPT	TOTAL
$\mu_{\texttt{wv\_norm}}$	CBOW	$\beta_{\texttt{wm\_norm}}$	0	0	0	0
		$\beta_{\texttt{dependent}}$	−	−	−	−
$\mu_{\texttt{wv\_norm}}$	skip-gram	$\beta_{\texttt{wm\_norm}}$	+	+	+	+
		$\beta_{\texttt{dependent}}$	−	−	−	−
$\mu_{\texttt{wv\_sim}}$	CBOW	$\beta_{\texttt{wm\_sim}}$	0	−	−	−
		$\beta_{\texttt{dependent}}$	−	−	−	−
$\mu_{\texttt{wv\_sim}}$	skip-gram	$\beta_{\texttt{wm\_sim}}$	−	−	−	−
		$\beta_{\texttt{dependent}}$	−	−	−	−
$\mu_{\texttt{wv\_all}}$	CBOW	$\beta_{\texttt{wm\_norm}}$	0	0	0	0
	CBOW	$\beta_{\texttt{wm\_sim}}$	0	−	−	−
		$\beta_{\texttt{dependent}}$	−	−	−	−
$\mu_{\texttt{wv\_all}}$	skip-gram	$\beta_{\texttt{wm\_norm}}$	+	+	+	+
	skip-gram	$\beta_{\texttt{wm\_sim}}$	−	−	−	−
		$\beta_{\texttt{dependent}}$	−	−	−	−
$\mu_{\texttt{freq}}$	相乗平均	$\beta_{\texttt{freq\_ave}}$	−	−	−	−
		$\beta_{\texttt{dependent}}$	−	−	−	−
$\mu_{\texttt{all}}$	CBOW	$\beta_{\texttt{wm\_norm}}$	0	0	0	0
	CBOW	$\beta_{\texttt{wm\_sim}}$	0	−	−	−
	相乗平均	$\beta_{\texttt{freq\_ave}}$	−	−	−	−
		$\beta_{\texttt{dependent}}$	−	−	−	−
$\mu_{\texttt{all}}$	skip-gram	$\beta_{\texttt{wm\_norm}}$	0	+	+	+
	skip-gram	$\beta_{\texttt{wm\_sim}}$	−	−	−	−
	相乗平均	$\beta_{\texttt{freq\_ave}}$	−	−	−	−
		$\beta_{\texttt{dependent}}$	−	−	−	−

[14] https://mc-stan.org/users/interfaces/rstan

まず，いずれの結果も ($\beta_{\texttt{dependent}}$) で係り受けが多ければ多いほど，読み時間が短くなった[15]．つまり，次に述べる結果は係り受けの効果を確認したうえでの付加的な効果である．単語埋め込みに基づくモデル ($\mu_{\texttt{wv\_all}}$) においては，隣接文節間類似度が大きければ大きいほど読み時間が短くなる傾向が見られた ($\beta_{\texttt{wm\_sim}}$)．skip-gram モデルにおいては，ベクトルのノルムが大きければ大きいほど読み時間が長くなる傾向が見られた ($\beta_{\texttt{wm\_norm}}$)．この傾向は CBOW には見られなかった．また，ノルムと類似度を個別にモデル化したもの ($\mu_{\texttt{wv\_norm}}$, $\mu_{\texttt{wv\_sim}}$) でも同じ傾向がみられた．次に，頻度の相乗平均に基づくモデル ($\mu_{\texttt{freq}}$) においては，高頻度のものが読み時間が短くなる傾向がみられた．単語埋め込みと頻度の双方を考慮したモデル ($\mu_{\texttt{all}}$) では，skip-gram の FFT 以外において，両者を個別にモデル化したものを合成したような結果が得られた．

以下，結果について考察する．

まず，文節間の隣接尤度ついては，隣接文節間類似度に関する係数 $\beta_{\texttt{wm\_sim}}$ で確認できる．類似度が大きいほど読み時間が短くなることにより，隣接尤度による予測が効くことが考えられる．つまり，wLSA-based surprisal (Pynte et al. 2008) や LDA-based surprisal (Mitchell et al. 2010) などの既存の semantic surprisal と同様に，fastText による単語埋め込みに基づくモデルでもモデル化できることが確認された．単語の頻度情報からは文節単位の隣接尤度の推定が困難であった．skip-gram の加法構成性に基づき構成した文節単位のベクトルのコサイン類似度が，適切に隣接尤度をモデル化できた．一方，CBOW については加法構成性をもつか否かについては管見の限り報告されていない．本稿の結果，読み時間の推定において CBOW は skip-gram ほどの明確な加法構成性が認められないことが示唆される．

文節内の単語の頻度の相乗平均は，その文節の生起確率を表す．確率が高ければ高いほど読み時間が短くなることが適切にモデル化できている．これは文節単位の unigram surprisal を適切にモデル化できていると考える．

skip-gram の FFT 以外においては，この unigram surprisal と異なる文節単位の特徴として，単語埋め込みのノルムの効果が認められた．単語埋め込みのノルム ($\beta_{\texttt{wm\_norm}}$) は，他の単語の特徴を表現していると考えられる．Schakel らは，単語埋め込みのノルムが，単語の頻度と同様に単語の重要度 (word significance) を表している (Schakel and Wilson 2015) とし，Luhn (1958) らの議論を引用しながら語の共起に関連する何らかの尺度を示していることを議論している．我々はこのノルムが文節間の何らかの特性を表していると考える．加法構成性に基づき文節単位のベクトルを構成することにより，ノルムが大きければ大きいほど予測が難しく，読み時間が長い傾向が確認できた．skip-gram のノルムの大きなものの例を見ると，数値表現[16]・長い付

[15] 先行文脈が統語的な関係を持ち，予測に効くために読み時間が短くなる (Levy 2008).
[16] 数値表現例：「２２６２億ドルで」「９９７５億１３００万円と」

属語連接[17]・長い複合語[18]が見られた．数値表現や付属語連接は，一般的に構成要素の頻度が高くなり，頻度の相乗平均の効果 ($\beta_{\texttt{freq\_ave}}$) のみの場合は読みやすいと判定されるが，この部分の読みにくさがモデル化されているのではないかと考える．文脈から単語を推定する CBOW モデルではこの傾向が見られなかった．skip-gram においてはその加法構成性を持つことが解説されている (Mikolov et al. 2013b) が，CBOW に関してはそのモデル化の方向性のためか加法構成性について議論がされていない．本研究結果でも CBOW の線形和に基づくノルムに対する読み時間の効果は確認できなかった．

いずれの結果も係り受けの効果とともに表れていることから，先行研究で示されている効果と従属性が低い特徴量が発見できたといえる．

5　おわりに

本研究では，日本語の読み時間の推定のために単語埋め込みを用いることを提案した．英語などで進められている surprisal の分析において，単語の頻度に基づく確率が用いられている．しかしながら，日本語においては頻度を計数する単位と読み時間を評価する単位との齟齬があり，この分析を難しくしていた．今回 skip-gram の単語埋め込みを用いて，ベクトルの線形和により文節ベクトルを構成することにより，この問題を解決した．文節ベクトルのノルムが当該文節の頻度とは異なる何らかの特性をモデル化し，ノルムが大きければ大きいほど読み時間が長くなることを確認した．さらにこれらの結果は統語的なモデルとともに導入されるものであり，新しい surprisal を発見したといえる．また，先行研究と同様に，左隣接文節のベクトルと当該文節ベクトルのコサイン類似度が，読み時間を適切にモデル化できることを確認した．

今回，この文節ベクトルのノルムが読み時間に影響を与える何らかの特性を持っていることを経験的に発見したが，数学的に言語学的に何であるのかについての理論的検証については今後の課題としたい．

工学的な応用として重要な点として，これらの単語埋め込みに関する情報は形態素解析器などで単語単位に割り当て可能であり，線形和やコサイン類似度など比較的軽い演算で計算できる．今回用いた統計モデルは線形式であることから，任意の日本語文章に対して簡単に読み時間の推定が計算できる．これまでの BCCWJ-EyeTrack の分析は，高度な統語・意味・談話情報アノテーションに基づくものであり，読み時間のモデル化については，人手によりアノテーションを行う必要があった．本研究で提案する単語埋め込みに基づくモデルは，NWJC2vec に収録されている語で構成される文章であれば，人間が解釈しやすい線形式で読み時間を与える

[17] 長い付属語連接例：「掲載させていただくことがあります。」「持たせてくれるんですよね」
[18] 長い複合語例：「クラーク元北大西洋条約機構（ＮＡＴＯ）欧州連合軍最高司令官が、」「「２１世紀ＣＯＥ（センター・オブ・エクセレンス）プログラム」に」「農畜産業振興事業団（現農畜産業振興機構）から、」

ことができる．本モデルを用いて，読み時間に基づく文章の読みやすさの自動評価ができると考える．

謝　辞

本研究は，国立国語研究所コーパス開発センター共同研究プロジェクト「コーパスアノテーションの拡張・統合・自動化に関する基礎研究」によるものです．本研究の一部は JSPS 科研費 基盤研究 (A) 17H00917，挑戦的研究（萌芽） 18K18519，新学術領域研究 18H05521 の助成を受けたものです．

参考文献

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

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

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), pp. 527–554.

浅原正幸 (2019). 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), pp. 301–328.

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), pp. 219–230.

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

浅原正幸，松本裕治 (2018). 『現代日本語書き言葉均衡コーパス』に対する文節係り受け・並列構造. 自然言語処理, **25** (4), pp. 331–356.

浅原正幸，小野創，宮本エジソン正 (2017). 『現代日本語書き言葉均衡コーパス』の読み時間とその被験者属性. 言語処理学会第 23 回年次大会発表論文集, pp. 473–476.

浅原正幸，小野創，宮本エジソン正 (2019). 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. 言語研究, **156**, p. To Appear.

浅原正幸，大村舞 (2016). BCCWJ-DepParaPAS: 『現代日本語書き言葉均衡コーパス』の係り受け・並列構造と述語項構造・共参照アノテーションの重ね合わせと可視化. 言語処理学

会第 22 回年次大会発表論文集, pp. 489–492.

Blei, D. M., Ng, A. Y., and Jordan, M. I. (2003). "Latent Dirichlet Allocation." *Journal of Machine Learning Research*, **3**, pp. 993–1022.

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

Griffiths, T. L., Steyvers, M., and Tanenbaum, J. B. (2007). "Topics in Semantic Representation." *Psychological Review*, **114** (2), pp. 211–244.

Hale, J. (2001). "A Probabilistic Earley Parser as a Psycholinguistic Model." In *Proceedings of the 2nd Conference of the North American Chapter of the Association for Computational Linguistics*, Vol. 2, pp. 159–166.

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** (2), pp. 134–141.

Landauer, T. K. and Dumais, S. T. (1997). "A Solution to Plato's problem: The Latent Semantic Analysis Theory of Acquisition, Induction and Representation of Knowledge." *Psyochological Review*, **104** (2), pp. 211–240.

Levy, R. (2008). "Expectation-based Syntactic Comprehension." *Cognition*, **106**, pp. 1126–1177.

Luhn, H. P. (1958). "The Automatic Creation of Literature Abstracts." *IBM Journal of Research and Development*, **2** (2), pp. 159–165.

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

Matsumoto, S., Asahara, M., and Arita, S. (2018). "Japanese Clause Classification Annotation on the 'Balanced Corpus of Contemporary Written Japanese'." In *Proceedings of the 13th Workshop on Asian Language Resources (ALR12)*, pp. 1–8.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013a). "Efficient Estimation of Word Representations in Vector Space." In *International Conference on Learning Representations,* **abs/1301.3781**.

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

Mitchell, J., Lapata, M., Demberg, V., and Keller, F. (2010). "Syntactic and Semantic Factors in Processing Difficulty: An Integrated Measure." In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pp. 196–206.

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**, pp. 19–33.

Patterson, C. and Drummer, J. (2016). "EyeTracking–Focus: Eyetracking During Reading." Linguistischer Methodenworkshop (HU Berlin).

Pynte, J., New, B., and Kennedy, A. (2008). "On-line Contextual Influences During Reading Normal Text: A Multiple-regression Analysis." *Vision Research*, **48**, pp. 2172–2183.

Schakel, A. M. J. and Wilson, B. J. (2015). "Measuring Word Significance using Distributed Representations of Words." *CoRR*, **abs/1508.02297**.

Sorensen, T., Hohenstein, S., and Vasishth, S. (2016). "Bayesian Linear Mixed Models Using Stan: A Tutorial for Psychologists, Linguists, and Cognitive Scientists." *Quantitative Methods for Psychology*, **12**, pp. 175–200.

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

Vasishth, S. and Drenhaus, H. (2011). "Locality in German." *Dialogue & Discourse*, **2** (1), pp. 59–82.

付録

A 分析結果（詳細）

本節では，skip-gram の最大モデル (μ_{all}) の結果の詳細について示す．Rhat が収束判定指標で chain 数 3 以上ですべての値が 1.1 以下を収束とみなす．本研究では全ての設定で chain 数 4 とした．n_eff が有効サンプル数，mean がサンプルの期待値（事後平均），sd が MCMC 標準偏差（事後標準偏差），se_mean が標準誤差で，MCMC のサンプルの分散を n_eff で割った値の平方根を表す．なお，表中 σ が対数正規分布の標準偏差，$\sigma_{article}$ が記事に対するランダム係数をモデル化する分布の標準偏差，σ_{subj} が実験協力者に対するランダム係数をモデル化する分布の標準偏差を表す．

A.1 分析結果：FFT skip-gram

表 3 に FFT skip-gram の最大モデル (μ_{all}) の結果を示す．FFT は文節内の最初の停留の注視時間を評価する．1 回の停留のみの評価のために，文節の長さが長い場合（文節内に複数回視線が停留する場合）に，その 2 回目以降の停留は加算されない．このため短い文節でかつ文処理負荷が高い文節で長くなる傾向がある．このため文節長の効果 ($\beta_{\texttt{length}}$) や文節間に空白を入

表 3 分析結果：FFT skip-gram (μ_{all})

Parameter	Rhat	n_eff	mean	sd	se_mean	2.5%	50%	97.5%
α	1.001	1,417	5.487	0.064	0.002	5.361	5.488	5.611
$\beta_{\texttt{length}}$	1.000	25,093	−0.002	0.002	0.000	−0.006	−0.002	0.001
$\beta_{\texttt{space}}$	1.000	17,950	−0.012	0.009	0.000	−0.030	−0.012	0.006
$\beta_{\texttt{dependent}}$	1.000	22,393	**−0.016**	0.005	0.000	−0.027	−0.016	−0.006
$\beta_{\texttt{sessionN}}$	1.000	17,212	0.001	0.009	0.000	−0.016	0.001	0.018
$\beta_{\texttt{articleN}}$	1.000	4,757	−0.007	0.007	0.000	−0.021	−0.007	0.007
$\beta_{\texttt{screenN}}$	1.000	18,221	−0.005	0.004	0.000	−0.013	−0.005	0.004
$\beta_{\texttt{lineN}}$	1.000	27,233	**−0.015**	0.003	0.000	−0.022	−0.015	−0.008
$\beta_{\texttt{segmentN}}$	1.000	14,648	**0.010**	0.002	0.000	0.005	0.010	0.014
$\beta_{\texttt{is\_first}}$	1.000	9,955	0.027	0.014	0.000	−0.002	0.027	0.056
$\beta_{\texttt{is\_last}}$	1.000	11,364	**−0.039**	0.015	0.000	−0.068	−0.039	−0.009
$\beta_{\texttt{is\_second\_last}}$	1.000	10,947	0.006	0.013	0.000	−0.019	0.006	0.031
$\beta_{\texttt{wv\_norm}}$	1.000	21,605	0.003	0.002	0.000	−0.000	0.003	0.006
$\beta_{\texttt{wv\_sim}}$	1.000	11,118	**−0.169**	0.028	0.000	−0.225	−0.169	−0.113
$\beta_{\texttt{freq\_ave}}$	1.000	20,787	**−0.004**	0.001	0.000	−0.007	−0.004	−0.002
σ	1.000	26,753	0.502	0.003	0.000	0.496	0.502	0.508
$\sigma_{\texttt{article}}$	1.001	5,225	0.038	0.010	0.000	0.022	0.037	0.061
$\sigma_{\texttt{subj}}$	1.000	7,762	0.194	0.032	0.000	0.144	0.190	0.267

れるか否かの効果 ($\beta_{\texttt{space}}$) が確認されなかった．呈示順 ($\beta_{\texttt{sessionN}}$，$\beta_{\texttt{articleN}}$，$\beta_{\texttt{screenN}}$，$\beta_{\texttt{lineN}}$，$\beta_{\texttt{segmentN}}$)・レイアウト情報 ($\beta_{\texttt{is\_first}}$，$\beta_{\texttt{is\_last}}$，$\beta_{\texttt{is\_second\_last}}$) に関しては，行呈示順 ($\beta_{\texttt{lineN}}$) で読み時間が短くなる傾向が，文節呈示順 ($\beta_{\texttt{segmentN}}$) で読み時間が長くなる傾向が，行内最右要素 ($\beta_{\texttt{is\_last}}$) で読み時間が短くなる傾向が見られた．

係り受けの数が多いほど読み時間が短くなる傾向 ($\beta_{\texttt{dependent}}$) がみられる．ベクトルのノルム ($\beta_{\texttt{wm\_norm}}$) については，ノルムが大きくなればなるほど読み時間が長くなるという弱い効果がみられる．ベクトルのノルムは，長い文節ほど大きくなる傾向があるために，1 回の停留のみの評価では強い効果が出なかったのではないかと考える．隣接要素との類似度 ($\beta_{\texttt{wm\_sim}}$) や頻度 ($\beta_{\texttt{freq\_ave}}$) については，値が大きいほど読み時間が短くなるという強い効果がみられる．

A.2 FPT skip-gram

表 4 に FPT skip-gram の最大モデル (μ_{all}) の結果を示す．FPT は文節内に初めて視線が停留し，その後文節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．一般に，予想と異なる文節とは異なるものが出てきた場合 (expectation-based)，短期記憶の負荷がかかる文節が出てきた場合 (memory-based) に値が大きくなるとされている．文節長の効果 ($\beta_{\texttt{length}}$) は視線停留対象の面積に相当するために，正方向に効果が出る．また，レジビ

表 4 分析結果：FPT skip-gram (μ_{all})

Parameter	Rhat	n_eff	mean	sd	se_mean
α	1.001	3,547	5.426	0.091	0.002
$\beta_{\texttt{length}}$	1.000	24,188	**0.066**	0.002	0.000
$\beta_{\texttt{space}}$	1.000	19,516	**−0.038**	0.011	0.000
$\beta_{\texttt{dependent}}$	1.000	20,172	**−0.061**	0.006	0.000
$\beta_{\texttt{sessionN}}$	1.000	20,113	**−0.052**	0.011	0.000
$\beta_{\texttt{articleN}}$	1.000	7,037	−0.009	0.012	0.000
$\beta_{\texttt{screenN}}$	1.000	20,552	**−0.026**	0.005	0.000
$\beta_{\texttt{lineN}}$	1.000	25,128	**−0.021**	0.004	0.000
$\beta_{\texttt{segmentN}}$	1.000	19,244	**−0.008**	0.003	0.000
$\beta_{\texttt{is\_first}}$	1.000	15,063	**0.175**	0.018	0.000
$\beta_{\texttt{is\_last}}$	1.000	14,175	0.027	0.019	0.000
$\beta_{\texttt{is\_second\_last}}$	1.000	17,814	**0.086**	0.016	0.000
$\beta_{\texttt{wv\_norm}}$	1.000	23,236	**0.024**	0.002	0.000
$\beta_{\texttt{wv\_sim}}$	1.000	14,256	**−0.296**	0.035	0.000
$\beta_{\texttt{freq\_ave}}$	1.000	20,866	**−0.006**	0.001	0.000
σ	1.000	25,598	0.622	0.004	0.000
$\sigma_{\texttt{article}}$	1.000	9,399	0.071	0.015	0.000
$\sigma_{\texttt{subj}}$	1.000	13,240	0.301	0.047	0.000
log-posterior	1.000	6,394	−279.449	5.671	0.071

リティの観点である文節間に空白を入れるか否かの効果 ($\beta_{\texttt{space}}$) ついては，空白を入れたほうが読み時間が短くなることが確認された．呈示順 ($\beta_{\texttt{sessionN}}$，$\beta_{\texttt{articleN}}$，$\beta_{\texttt{screenN}}$，$\beta_{\texttt{lineN}}$，$\beta_{\texttt{segmentN}}$) に関しては，記事呈示順以外で実験が進むにつれて読み時間が短くなる．記事呈示順は実験計画として 4 パターンのみ準備しており，記事に対するランダム効果 ($\sigma_{\texttt{article}}$) に吸収されたと考える．以上の傾向は，RPT， TOTAL についても共通してみられる．

レイアウト情報 ($\beta_{\texttt{is\_first}}$，$\beta_{\texttt{is\_last}}$，$\beta_{\texttt{is\_second\_last}}$) は，行内最左要素 ($\beta_{\texttt{is\_first}}$) と右から 2 番目の要素 ($\beta_{\texttt{is\_second\_last}}$) で読み時間が長くなる傾向が見られた．これは，注視点の復帰改行の移動の効果だと考える．

係り受けの数が多いほど読み時間が短くなる傾向 ($\beta_{\texttt{dependent}}$) がみられる．ベクトルのノルム ($\beta_{\texttt{wm\_norm}}$) については，ノルムが大きくなればなるほど読み時間が長くなるという強い効果がみられる．隣接要素との類似度 ($\beta_{\texttt{wm\_sim}}$) や頻度 ($\beta_{\texttt{freq\_ave}}$) については，値が大きいほど読み時間が短くなるという強い効果がみられる．

A.3 RPT skip-gram

表 5 に RPT skip-gram の最大モデル (μ_{all}) の結果を示す．RPT は文節内に初めて視線が停留し，その後文節の右側から出るまでの総注視時間である．左側に抜ける場合は継続して合算

表 5 分析結果：RPT skip-gram (μ_{all})

Parameter	Rhat	n_eff	mean	sd	se_mean
α	1.001	2,671	5.656	0.101	0.002
$\beta_{\texttt{length}}$	1.000	26,837	0.066	0.003	0.000
$\beta_{\texttt{space}}$	1.000	20,083	−0.040	0.013	0.000
$\beta_{\texttt{dependent}}$	1.000	22,361	−0.050	0.008	0.000
$\beta_{\texttt{sessionN}}$	1.000	21,139	−0.085	0.013	0.000
$\beta_{\texttt{articleN}}$	1.001	5,418	−0.012	0.013	0.000
$\beta_{\texttt{screenN}}$	1.000	21,178	−0.025	0.006	0.000
$\beta_{\texttt{lineN}}$	1.000	26,959	−0.008	0.005	0.000
$\beta_{\texttt{segmentN}}$	1.000	19,514	−0.025	0.004	0.000
$\beta_{\texttt{is\_first}}$	1.000	14,364	0.035	0.022	0.000
$\beta_{\texttt{is\_last}}$	1.000	14,064	0.178	0.023	0.000
$\beta_{\texttt{is\_second\_last}}$	1.000	16,721	0.117	0.019	0.000
$\beta_{\texttt{wv\_norm}}$	1.000	24,222	0.012	0.002	0.000
$\beta_{\texttt{wv\_sim}}$	1.000	15,533	−0.290	0.043	0.000
$\beta_{\texttt{freq\_ave}}$	1.000	22,056	−0.007	0.002	0.000
σ	1.000	31,074	0.757	0.005	0.000
$\sigma_{\texttt{article}}$	1.000	8,773	0.075	0.017	0.000
$\sigma_{\texttt{subj}}$	1.000	11,137	0.308	0.050	0.000
log-posterior	1.000	6,088	−2,876.653	5.692	0.073

する．当該文節が予想と異なった場合に事前文脈を再確認する時間を評価している．

文節長・空白・呈示順については，FPT と同じ傾向であった．

レイアウト情報 ($\beta_{\texttt{is\_first}}$，$\beta_{\texttt{is\_last}}$，$\beta_{\texttt{is\_second\_last}}$) は，行内最右要素 ($\beta_{\texttt{is\_last}}$) と右から 2 番目の要素 ($\beta_{\texttt{is\_second\_last}}$) で読み時間が長くなる傾向が見られた．これは，RPT の読み時間の定義から，最右要素や右から 2 番目の要素はこれ以上右につきぬけにくいためであろう．

係り受けの数が多いほど読み時間が短くなる傾向 ($\beta_{\texttt{dependent}}$) がみられる．ベクトルのノルム ($\beta_{\texttt{wm\_norm}}$) については，ノルムが大きくなればなるほど読み時間が長くなるという強い効果がみられる．隣接要素との類似度 ($\beta_{\texttt{wm\_sim}}$) や頻度 ($\beta_{\texttt{freq\_ave}}$) については，値が大きいほど読み時間が短くなるという強い効果がみられる．

A.4 TOTAL skip-gram

表 6 に TOTAL skip-gram の最大モデル (μ_{all}) の結果を示す．TOTAL は文節内の総注視時間である．2 回目以降の確認作業も含めた読み時間を評価する．文節長・空白・呈示順については，FPT と同じ傾向であった．

レイアウト情報 ($\beta_{\texttt{is\_first}}$，$\beta_{\texttt{is\_last}}$，$\beta_{\texttt{is\_second\_last}}$) は，FPT と同様に，行内最左要素 ($\beta_{\texttt{is\_last}}$) と右から 2 番目の要素 ($\beta_{\texttt{is\_second\_last}}$) で読み時間が長くなる傾向が見られた．

表 6 分析結果：TOTAL skip-gram (μ_{all})

Parameter	Rhat	n_eff	mean	sd	se_mean
α	1.001	2,527	5.949	0.096	0.002
$\beta_{\texttt{length}}$	1.000	26,389	0.063	0.003	0.000
$\beta_{\texttt{space}}$	1.000	19,587	−0.065	0.012	0.000
$\beta_{\texttt{dependent}}$	1.000	23,809	−0.064	0.007	0.000
$\beta_{\texttt{sessionN}}$	1.000	20,366	−0.075	0.012	0.000
$\beta_{\texttt{articleN}}$	1.000	5,349	−0.004	0.013	0.000
$\beta_{\texttt{screenN}}$	1.000	21,009	−0.039	0.006	0.000
$\beta_{\texttt{lineN}}$	1.000	27,650	−0.022	0.004	0.000
$\beta_{\texttt{segmentN}}$	1.000	18,581	−0.021	0.003	0.000
$\beta_{\texttt{is\_first}}$	1.000	12,557	0.124	0.019	0.000
$\beta_{\texttt{is\_last}}$	1.001	14,770	−0.021	0.020	0.000
$\beta_{\texttt{is\_second\_last}}$	1.000	16,749	0.090	0.017	0.000
$\beta_{\texttt{wv\_norm}}$	1.000	25,069	0.025	0.002	0.000
$\beta_{\texttt{wv\_sim}}$	1.000	14,568	−0.309	0.037	0.000
$\beta_{\texttt{freq\_ave}}$	1.000	21,219	−0.009	0.002	0.000
σ	1.000	27,957	0.653	0.004	0.000
$\sigma_{\texttt{article}}$	1.000	9,042	0.083	0.018	0.000
$\sigma_{\texttt{subj}}$	1.000	12,845	0.297	0.047	0.000
log-posterior	1.000	5,505	−912.243	5.725	0.077

係り受けの数が多いほど読み時間が短くなる傾向 ($\beta_{\texttt{dependent}}$) がみられる．ベクトルのノルム ($\beta_{\texttt{wm\_norm}}$) については，ノルムが大きくなればなるほど読み時間が長くなるという強い効果がみられる．隣接要素との類似度 ($\beta_{\texttt{wm\_sim}}$) や頻度 ($\beta_{\texttt{freq\_ave}}$) については，値が大きいほど読み時間が短くなるという強い効果がみられる．

B Second Pass Time について

Second Pass Time (SPT) は研究者によって，ゼロ読み時間（読み飛ばし）の扱いが異なり，査読などで議論が対立することが多く，本稿では SPT の結果を除外した．

自己ペース読文法においては，実験協力者は必ずすべての文節を見るために読み飛ばしが発生しない．視線走査法の FFT, FPT, RPT, TOTAL においては，ゼロ読み時間を考慮しないことが研究コミュニティにおいて共有されている．SPT はゼロ読み時間を扱う研究とゼロ読み時間を扱わない研究があり，BCCWJ-EyeTrack では後者の扱いをとっている．

著者らが考える理由は三つある．一つ目は TOTAL と FPT の関係である．SPT においてゼロ読み時間をする場合，TOTAL においてゼロ読み時間の場合を除いて TOTAL – FPT の値と

SPT の値が完全従属する．二つ目は対数正規分布 lognormal によりモデル化できる点である．対数正規分布は定義域が $0 < x < \infty$ であり，ゼロ読み時間を評価することができない．しかしながら，正規分布と異なり，サンプリングの際に自然に負の時間を回避できるほか，外れ値の影響が軽減されるというメリットがある．三つ目は，本質的に 2 回目の読み時間がないということは欠損値であると考え，モデル化する対象から外すことで，0 の値を割り当てるという overspecified の問題を回避することができる．

Vasishth らは，ゼロ読み時間を扱うものを rereading time と呼び，ゼロ読み時間を扱わないものを SPT として区別したうえで，SPT を扱うべきとしている (Vasishth and Drenhaus 2011)．さらに rereading time については UMASS の eyedry[19]など rereading time = RPT − FPT としているものもある．Patterson らは 2016 年の時点で “controversy over including 0 when no rereading” (Patterson and Drummer 2016) とし，この扱いについては，まだ議論が収束していない．

なお，SPT 分析結果としては，分散表現のノルムのみ（CBOW, skip-gram とも）が効果として確認され，それ以外の効果（単語頻度の幾何平均・隣接文節との類似度）は確認されなかった．

略歴

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 年より人間文化研究機構国立国語研究所コーパス開発センター特任准教授．2019 年より同教授．博士（工学）．言語処理学会，日本言語学会，日本語学会各会員．

（2019 年 1 月 31 日　受付）
（2019 年 4 月 26 日　再受付）
（2019 年 6 月 22 日　採録）

[19] http://blogs.umass.edu/eyelab/software/