B-4

単語埋め込みに基づくサプライザルのモデル化
Modeling surprisal by word embeddings

人間文化研究機構 国立国語研究所　浅原　正幸
National Intitute for Japanese Language and Linguistics, Japan, Masayuki ASAHARA

1 はじめに

本研究では日本語を対象として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評価・統語解析モデルの構築・言語受容者側の特徴を取り入れた動的な言語処理について検討を行う。

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 et al., 2014) の読み時間データ BCCWJ-EyeTrack (Asahara et al., 2016) 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 (2001) は、頻度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 Surprisal theory として定式化している。この Surprisal 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評価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 et al., 2013b) 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、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., 2013a)。

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 et al., 2014) から fastText (Bojanowski et al., 2017) により構成した NWJC2vec (Asahara, 2018c) を用いた。ベイジアン線形混合モデル (Sorensen et al., 2016) に基づく統計分析の結果、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。前者のノルムが連接する文節の多様性を、後者の隣接文節間のコサイン類似度が隣接確率をモデル化することが分かった。

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。

2 前提

2.1 BCCWJ-EyeTrack

BCCWJ-EyeTrack (Asahara et al., 2016) 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。Table 1 に BCCWJ-EyeTrack のデータ形式について示す。データには、BCCWJ に対する文節係り受けアノテーション BCCWJ-DepPara (Asahara and Matsumoto, 2016) に基づいた、読み時間評価対象文節の係り受けの数 (`dependent`) が付与されている。

Table 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
sample_screen	factor	画面識別子
length	int	文字数
space	factor	文節境界空白の有無
setorder	int	文節境界空白の呈示順
subj	factor	実験協力者 ID
dependent	int	係り受け関係

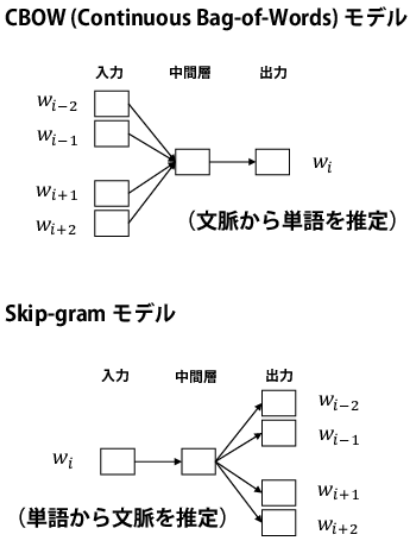

Figure 1: CBOW モデルと Skip-gram モデル

2.2 Surprisal

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 surprisal と呼んだ：

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

Surprisal は x の（文脈による条件付き）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、確率的言語モデルに基づくもの [1]・N-gram Surprisal [2]・Parser Surprisal [3] 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 [4] を提案した。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単語列を期待しうるか否かの困難さをモデル化する。

2.3 単語埋め込みと NWJC2vec

単語埋め込みは (Mikolov et al., 2013b) 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 (Figure 1 上) と単語から文脈を推定する skip-gram モデル (Figure 1 下) 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、skim-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。

[1] $\texttt{Surprisal}_{k+1} = -\log_2 P(w_{k+1}|w_1 \dots w_k)$
[2] $\texttt{Surprisal}_{w_{k+1}} = -\log_2 P(w_{k+1}|w_{k-2}, w_{k-1}, w_k)$
[3] $\texttt{Surprisal}_n = \log_2 P(T, w_1 \dots w_{n-1}) - \log_2 P(T, w_1 \dots w_n)$
[4] $\texttt{Surprisal}_{k+1} = D(P_{k+1}||P_k) = -\log P(w_{k+1}|w_1 \dots w_k)$

NWJC2vec (Asahara, 2018c) 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) を用いてモデル化したもの [5] 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。

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。浅原他 (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。Asahara et al. (2016) は文節係り受けアノテーション BCCWJ-DepPara (Asahara and Matsumoto, 2016) 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。Asahara (2018a) は節情報アノテーション BCCWJ-ToriClause (Matsumoto et al., 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。Asahara and Kato (2017) は分類語彙表番号アノテーション BCCWJ-WLSP (Kato et al., 2018) 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。Asahara (2017) は情報構造アノテーション BCCWJ-Infostr (Miyauchi et al., 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 (2018b) は述語項構造・共参照情報アノテーション BCCWJ-PAS (植田他, 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

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

Figure 2: 推定に用いた線形式

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。Figure 2 に推定に用いた線形式を示す。

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、$\beta^{\texttt{dependent}}$ は当該文節に係る文節の数に対する係数である。$\beta^{\texttt{sessionN}}, \beta^{\texttt{articleN}}, \beta^{\texttt{screenN}}, \beta^{\texttt{lineN}}, \beta^{\texttt{segmentN}}$ が試行順に対する係数、$\beta^{\texttt{is\_first}}, \beta^{\texttt{is\_last}}, \beta^{\texttt{is\_second\_last}}$ 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、被験者に対するランダム係数として $\gamma^{\texttt{subj}=s(x)}$ を考慮する。

[5] 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`

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。単語埋め込みのノルムは大きければ大きいほど、多くの単語の結びつきやすい傾向にあり、結果として連接曖昧性を増す傾向にある。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 $\mu_{\texttt{wv}}$ 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距離に対する係数であり、これを評価する。モデルとして、CBOW と skip-gram の二つを評価する。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、文節間の連接確率を考慮しない。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。相乗平均を評価する際にゼロ頻度は 1 を乗じた。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 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

Table 2: 分析結果 (概要)

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

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 未満のものである
+ 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

$\beta_{\texttt{wm\_norm}}$: skip-gram において読み時間が長くなる
$\beta_{\texttt{wm\_sim}}$: 読み時間が短くなる
$\beta_{\texttt{freq\_ave}}$: 読み時間が短くなる

Table 2 に各モデルの分析結果を示す。

まず頻度の相乗平均に基づくモデル ($\mu_{\texttt{freq}}$) においては、SPT 以外で高頻度のものが読み時間が短くなる。次に単語埋め込みに基づくモデル ($\mu_{\texttt{wv}}$) 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。最後に単語埋め込みと頻度の双方を考慮したモデル ($\mu_{\texttt{all}}$) においては、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。

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。単語埋め込みのノルムは、他の単語の隣接可能性を表現している。加法構成性に基づき文節単位のベクトルを構成しても、そのノルムは文節の隣接可能性を表現しており、結果として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になったと考える。これは、文脈から単語を推定する CBOW モデルではその傾向が見られなかったが、skip-gram モデルでは確認することができた。文節間の隣接可能性については、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。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。

最後に Total Time の分析結果の詳細を Table 3 に示す。

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

Table 3: 分析結果 (Total Time)

Parameter	Rhat	n_eff	mean	sd	se_mean
α	1.001	2527	5.949	0.096	0.002
$\beta_{\texttt{length}}$	1.000	18000	0.063	0.003	0.000
$\beta_{\texttt{space}}$	1.000	18000	-0.065	0.012	0.000
$\beta_{\texttt{dependent}}$	1.000	18000	-0.064	0.007	0.000
$\beta_{\texttt{sessionN}}$	1.000	18000	-0.075	0.012	0.000
$\beta_{\texttt{articleN}}$	1.000	5337	-0.004	0.013	0.000
$\beta_{\texttt{screenN}}$	1.000	18000	-0.039	0.006	0.000
$\beta_{\texttt{lineN}}$	1.000	18000	-0.022	0.004	0.000
$\beta_{\texttt{segmentN}}$	1.000	18000	-0.021	0.003	0.000
$\beta_{\texttt{is\_first}}$	1.000	12605	0.124	0.019	0.000
$\beta_{\texttt{is\_last}}$	1.001	14859	-0.021	0.020	0.000
$\beta_{\texttt{is\_second\_last}}$	1.000	16749	0.090	0.017	0.000
$\beta_{\texttt{wv\_norm}}$	1.000	18000	0.025	0.002	0.000
$\beta_{\texttt{wv\_sim}}$	1.000	15203	-0.309	0.037	0.000
$\beta_{\texttt{freq\_ave}}$	1.000	18000	-0.009	0.002	0.000
σ	1.000	18000	0.653	0.004	0.000
$\sigma_{\texttt{article}}$	1.000	8997	0.083	0.018	0.000
$\sigma_{\texttt{subj}}$	1.000	12999	0.297	0.047	0.000
log-posterior	1.000	5489	-912.243	5.725	0.077

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
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

$\beta_{\texttt{length}}$: 文字数が多ければ視線停留対象面積が大きくなり、結果読み時間が長くなる
$\beta_{\texttt{space}}$: 文節間に空白を入れたほうが読み時間が短くなる
$\beta_{\texttt{dependent}}$: 係り受けの数が多いほど読み時間が短くなる
$\beta_{*\texttt{N}}$: 実験が進めば進むほど読み時間が短くなる
$\beta_{\texttt{is\_*}}$: レイアウト上の制約。両端で読み時間が変化する
$\beta_{\texttt{wv\_norm}}$: ノルムが大きいほど読み時間が長くなる
$\beta_{\texttt{wv\_sim}}$: 類似度が大きいほど読み時間が短くなる
$\beta_{\texttt{freq\_ave}}$: 頻度が高いほど読み時間が短くなる

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隣接可能性をモデル化し、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、隣接尤度を適切にモデル化できることを確認した。

作例に基づく読み時間評価の被験者実験では、例文中の語選択において、頻度による統制が一般的に行われる。本研究の結果から、単語埋め込みのノルムが読み時間に対して影響を与えることが分かった。作例時に単語埋め込みのノルムについても統制が必要であるだろう。

しかしながら、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の線形式であることから、簡単に読み時間の推定が計算できる。ゆえに、統計分析時に頻度情報や単語埋め込みの計量を用いることで、統制が不要になる可能性がある。

Acknowledgement

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および科研費 JP15K12888, JP17H00917, JP18H05521 によるものです。

References

Asahara, Masayuki (2017) "Between Reading Time and Information Structure," in *Proceedings of the 31st Pacific Asia Conference on Language, Information and Computation (PACLIC 31)*, pp. 15–24.

——— (2018a) "Between Reading Time and Clause Boundaries in Japanese – Wrap-up Effect in a Head-Final Language," in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32)*, p. (To Appear).

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

——— (2018c) "NWJC2Vec: Word embedding dataset from 'NINJAL Web Japanese Corpus'," *Terminology: International Journal of Theoretical and Applied Issues in Specialized Communication*, Vol. 24, No. 2, pp. 7–25, Feb.

Asahara, Masayuki and Sachi Kato (2017) "Between Reading Time and Syntactic/Semantic Categories," in *Proceedings of the eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 404–412.

Asahara, Masayuki and Yuji Matsumoto (2016) "BCCWJ-DepPara: A Syntactic Annotation Treebank on the 'Balanced Corpus of Contemporary Written Japanese'," in *Proceedings of the 12th Workshop on Asian Language Resources (ALR12)*, pp. 49–58.

Asahara, Masayuki, Kikuo Maekawa, Mizuho Imada, Sachi Kato, and Hikari Konishi (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*, Vol. 25, No. 1–2, pp. 129–148.

Asahara, Masayuki, Hajime Ono, and Edson T. Miyamoto (2016) "Reading-Time Annotations for 'Balanced Corpus of Contemporary Written Japanese'," in *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics*, pp. 684–694.

Bojanowski, Piotr, Edouard Grave, Armand Joulin, and Tomas Mikolov (2017) "Enriching Word Vectors with Subword Information," *Transactions of the Association for Computational Linguistics*, Vol. 5, pp. 135–146.

Hale, John (2001) "A Probabilistic Earley Parser as a Psycholinguistic Model," in *Proceedings of the second conference of the North American chapter of the association for computational linguistics*, Vol. 2, pp. 159–166.

Kato, Sachi, Masayuki Asahara, and Makoto Yamazaki (2018) "Annotation of 'Word List by Semantic Principles' Labels for Balanced Corpus of Contemporary Written Japanese," in *Proceedings of the 32nd Pacific Asia Conference on Language, Information and Computation (PACLIC 32)*, p. (To Appear).

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

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*, Vol. 48, pp. 345-371.

Matsumoto, Satomi, Masayuki Asahara, and Setsuko Arita (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, Tomas, Ilya Sutskever, Kai Chen, Greg Corrado, and Jeffrey Dean (2013a) "Distributed Representations of Words and Phrases and their Compositionalyty," in *Advances in Neural Information Processing Systems*, pp. 3111–3119.

Mikolov, Tomas, Kai Chen, Greg Corrado, and Jeffrey Dean (2013b) "Efficient Estimation of Word Representations in Vector Space," in *International Conference on Learning Representations*.

Miyauchi, Takuya, Masayuki Asahara, Natsuko Nakagawa, and Sachi Kato (2018) "Information-structure annotation of the "Balanced Corpus of Contemporary Written Japanese"," in *Communications in Computer and Information Science, 781: International Conference of the Pacific Association for Computational Linguistics*, pp. 155–164.

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*, Vol. 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』, 205–214 頁.

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』, 489–492 頁.

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』, 473–476 頁.