再帰的ニューラルネットワーク文法による人間の文処理のモデリング

吉田遼[1]　能地宏[2]　大関洋平[1]
[1]東京大学　[2]産業技術総合研究所
{yoshiryo0617, oseki}@g.ecc.u-tokyo.ac.jp, hiroshi.noji@aist.go.jp

1 はじめに

理論言語学では, 自然言語は階層構造を持つと言われている [1]. しかし, RNN 言語モデルの一つである LSTM [2] 言語モデルは, 階層構造を考慮しない線形モデルであるにもかかわらず, 言語モデリング精度が高い [3] だけでなく, 単語間の長距離依存関係を把握できる文法能力を有している [4, 5]. また, 近年では, RNN 言語モデルが算出した確率的な予測が, 階層構造モデルである PCFG よりも高い精度で脳波をモデル化できることから, 線形モデルの認知的妥当性までもが主張されている [6].

一方で, 自然言語の階層構造を明示的に考慮するニューラル言語モデルも提案されている. その一つが, 階層構造と単語列の生成モデルである再帰的ニューラルネットワーク文法 (recurrent neural network grammar, RNNG) [7] である. 先行研究では, RNNG は LSTM 言語モデルよりも高い言語モデリング精度や文法能力を有している [8, 9] など, 言語処理において線形モデルよりも階層構造モデルが優位であることが示されてきた. また, RNNG は LSTM 言語モデルよりも高い精度で脳波をモデル化できることから, 線形モデルは認知的に妥当であるとはいえず, 人間は階層構造を認識しながら文処理をしている可能性が示唆されている [10].

Hale ら [10] は, RNNG の認知的妥当性を主張しているが, その parsing strategy には着目していない. RNNG は, top-down parsing strategy を持つモデルであるが, 中央埋め込み文に対するワーキングメモリ負荷の観点から, 人間の持つ parsing strategy は left-corner であると言われている [11]. また, 先行研究で用いられている英語のように, 右枝分かれ構造を持つ言語とは異なり, 日本語のように左枝分かれ構造を持つ言語は, top-down parsing strategy では解析が難しいと言われており, 人間の持つ parsing strategy は top-down ではない可能性がある.

そこで本研究では, 左枝分かれ構造を持つ日本語を用いて, top-down/left-corner parsing strategy を持つ RNNG の認知的妥当性を, LSTM 言語モデルと比較することにより検証した. 実験により, left-corner parsing モデルが最も認知的に妥当であるが, 線形モデルや top-down parsing モデルも部分的には認知的に妥当であることを示唆する結果が得られた. また, 先行研究では言語モデリング精度の高い言語モデルほど認知的妥当性が高いことが確認されている [12] が, 本研究の追加実験では, 工学的精度の高い RNNG ほど認知的妥当性が高い一方で, LSTM 言語モデルの言語モデリング精度の高さは必ずしも認知的な妥当性を意味しないことが示唆された.

2 Linking hypothesis：surprisal 理論

人間は, 文脈から次に来る単語や文節を予測しながら文を処理しており, 予測しやすい単語や文節は処理負荷が低く, 読み時間が短くなると言われている. これを定式化したのが, surprisal 理論 [13, 14] である. Surprisal 理論では, 文脈に出現する単語や文節の情報量は surprisal（$-\log p$(単語や文節 | 文脈)）により測ることができ, さらに p(単語や文節 | 文脈) が小さいほど大きい値をとる surprisal は, 単語や文節の処理負荷に比例するとされている. 近年では, 言語モデルの算出した surprisal が読み時間や脳波のモデル化に有効であることが示されており [6, 12], 階層構造モデルと線形モデルの読み時間・脳波のモデル化精度を言語モデルの認知的妥当性とみなし比較することで, 人間の文処理に階層構造の認識が伴うか, という問いが検証されてきた [15, 16, 6, 10]. しかし, これらの先行研究では, 統一的な結論は得られておらず, また階層構造モデルの parsing strategy は着目されていない. 本研究では, surprisal 理論に則り, 日本語で LSTM 言語モデルと top-down/left-corner

RNNG の認知的妥当性を比較する.

3 実験

3.1 言語モデル

3.1.1 LSTM 言語モデル

LSTM 言語モデルは, 階層構造を考慮しない単語列の生成モデルである. 本研究では, 単語埋め込みの次元数 256, 隠れ層の次元数 256 の 2 層 LSTM を持つ LSTM 言語モデルを用いた.

3.1.2 Top-down/left-corner RNNG

RNNG は, RNN を用いた階層構造と単語列の生成モデルである. RNNG のアーキテクチャを, 図 1 に示す. RNNG では, 単語（e.g., “*hungry*”）, 開いた句のラベル（e.g., “(NP”）がベクトルで表現される. それらのベクトルから, 閉じた句（e.g., “(NP *the hungry cat*)”）も一つのベクトル表現として算出される. それらのベクトルは, スタックと呼ばれるデータ構造に保持される.

Top-down RNNG では, 各時刻においてスタック LSTM により算出されたスタックの状態を表すベクトルに基づき, 以下の 3 つのアクションに対する確率分布が算出される：

- 開いた句のラベルの生成：開いた句のラベルを表すベクトルが, スタックの先頭に追加される.
- 単語の生成：単語を表すベクトルが, スタックの先頭に追加される.
- 句を閉じる：スタック内の開いた句のラベル及びその句の構成素を表すベクトルが, 閉じた句を表す一つのベクトルに合成される.

前 2 つのアクションが選択された際には, スタックの状態を表すベクトルに基づきどの開いた句のラベル・単語が生成されるかが選択される. 「句を閉じる」アクションの際の, 閉じた句を表すベクトル算出の際には双方向 LSTM が用いられる（図 2）.

Left-corner RNNG [8] では, top-down RNNG の「開いた句のラベルの生成」にあたるアクションとして,「開いた句ラベルの生成及びスタック内での入れ替え」アクションがある. Left-corner parsing では, 開いた句のラベルがその句の構成素の最左要素が生成または句として閉じられた後に生成される. よって, 開いた句のラベルが生成された際に, 句の構成素の最

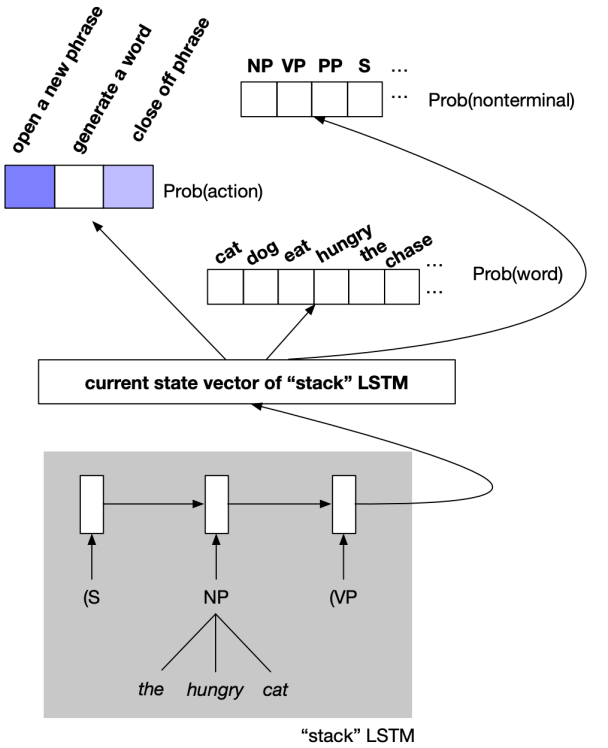

図 1 RNNG のアーキテクチャ. 図は Hale ら [10] より.

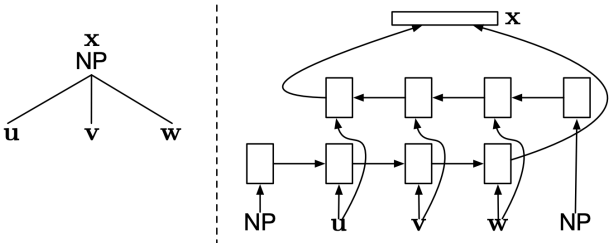

図 2 「句を閉じる」アクションの際の閉じた句ベクトルの算出. 図は Dyer ら [7] より.

左要素がスタックの先頭に存在するため, それらを入れ替えることで, スタック内にその時刻までに構築された階層構造を実現する.

本研究では, 単語埋め込みの次元数 256, 隠れ層の次元数 256 の 2 層スタック LSTM を持つ top-down/left-corner RNNG を用いた.

3.2 コーパス

3.2.1 学習用コーパス

NINJAL Parsed Corpus of Modern Japanese（NPCMJ, http://npcmj.ninjal.ac.jp）を用いた. 現代日本語の書き言葉と話し言葉のテクストに対し文の統語・意味解析情報が付与されている. 総文数 40,831 文,

総語数 560,098 語である．短単位[1)]を入力単位として，文単位に分割されたコーパスで言語モデルを学習した．LSTM 言語モデルの学習には，単語列のみ，RNNG の学習には，階層構造[2)]・単語列が用いられた．

3.2.2 視線計測コーパス

BCCWJ-EyeTrack [20] を用いた．『現代日本語書き言葉均衡コーパス』 [21] の新聞記事サンプルに対して，日本語母語話者 24 人分の読み時間が付与されている．Smith ら [22] に則り，読み時間として first pass time を用いた．浅原ら [20] に則り、読み時間がゼロミリ秒である文節は視線が停留していないとして分析データから除外し，また，本文に出現する文節のみを対象とした．さらに，Fossum ら [16] を踏襲し，言語モデルの訓練データに出現しない単語を含む文節についても除外した．

3.3 Surprisal

LSTM 言語モデルの単語 surprisal は言語モデルが算出した当該単語の文脈条件付き確率（p(単語 | 文脈)）から直接求められる．

RNNG の単語 surprisal は，Hale ら [10] を踏襲し、ビームサーチ [23] を用いて求める．単語列の背後に想定される階層構造のうち確率の高いものを複数保持しつつ，それらの確率を階層構造について周辺化することで p(単語 | 文脈) を求める．本研究では，Wilcox ら [9] を踏襲し，周辺化対象となる構造の数を表す単語ビーム幅として 10 を採用した[3)]．

言語モデルは単語単位の確率を算出するが，日本語では読み時間は文節単位に付与される．本研究では，文節内の単語 surprisal の和を文節 surprisal として用いた．

3.4 評価指標：Deviance accuracy

先行研究にならい，読み時間に関係するとされる説明変数で対数読み時間をモデル化したベースラインの回帰モデルに，言語モデルの surprisal を説明変数として加えた際の deviance の減少分（deviance accuracy）を比較する．Deviance accuracy に統計的な有意差があるかどうかは，nested model comparision で評価した．言語モデル A, B の surprisal を共に説明変数に含む回帰モデルの deviance が，言語モデル A の surprisal のみを含む回帰モデルの deviance よりも χ^2 検定（$p \leq 0.05$）下で有意に小さければ，言語モデル B の surprisal は，言語モデル A の surprisal が説明できていない variance を説明できており，deviance accuracy の差は有意である．逆もまた同様である．

ベースライン回帰モデルには，以下の線形混合モデルを用いる：

$$\begin{aligned}\log(\mathrm{RT}) \sim\ & \mathtt{length} + \mathtt{freq} \\ & + \mathtt{is\_first} + \mathtt{is\_last} + \mathtt{is\_second\_last} \\ & + \mathtt{screenN} + \mathtt{lineN} + \mathtt{segmentN} \\ & + (1|\mathtt{article}) + (1|\mathtt{subj}).\end{aligned} \tag{1}$$

各説明変数の詳細については，付録に示す．数値型の説明変数は，全て中心化を行った．最初に一度モデル化した上で，3 標準偏差を超える文節を除外した．11,504 の文節が最終的なモデル化対象となった．

4 結果

結果を図 3 に示した．横軸が言語モデルを表し，縦軸が deviance accuracy を表す．Deviance accuracy の nested model comparision の結果を，表 1 に示した．図 3 より，ベースライン回帰モデルに surprisal を説明変数として加えた際に，全ての言語モデルで読み時間のモデル化精度が向上した．言語モデル間では，left-corner RNNG が最も deviance accuracy が高く，LSTM 言語モデルと top-down RNNG は left-corner RNNG より低い同程度の deviance accuracy であった．

表 1 より，left-corner RNNG と LSTM 言語モデル，top-down RNNG 間の deviance accuracy の差が有意であった．また，全ての言語モデル間の組み合わせで nested model comparision の結果が有意であり，deviance accracy が低い LSTM 言語モデルや top-down RNNG であっても，left-corner RNNG が説明できない varirnce を説明できていた．これは，left-corner parsing モデルが最も認知的に妥当であるが，線形モデルや top-down parsing モデルも部分的には認知的に妥当であることを示唆する．

5 追加実験

Hale ら [10] では，top-down RNNG が LSTM 言語モデルよりも認知的妥当性が高いことが報告されていた．一方，本研究では，top-down RNNG と LSTM 言語モデルの deviance accuracy は同程度であった．

1） MeCab [17] を用いて，Unidic [18] 辞書により短単位に分割した．
2） Frank ら [15] に則り，standard manner [19] により空要素や機能を示す拡張タグを除去した．
3） アクションビーム幅として 100，ファストトラック幅として 1 を採用した．

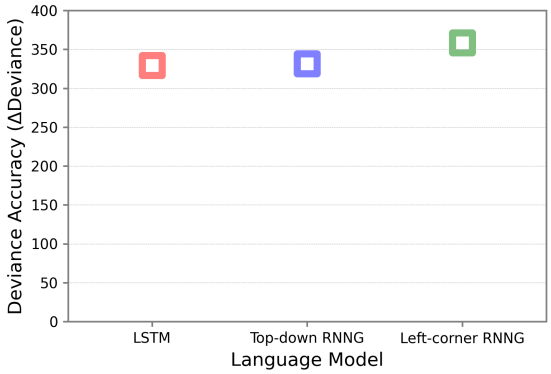

図 3 言語モデルの deviance accuracy. 横軸は言語モデルを, 縦軸は deviance accuracy（ベースライン回帰モデルからの deviance の減少分）を表す. 各色が表す言語モデルは以下：赤＝ LSTM 言語モデル、青＝ top-down RNNG、緑＝ left-corner RNNG.

表 1 Nested model comparison の結果. Bonferroni 法により有意水準 $\alpha = 0.0056$ で検定した.

	χ^2	df	p
Baseline < LSTM	343.45	1	**<0.0001**
Baseline < Top-down RNNG	345.85	1	**<0.0001**
Baseline < Left-corner RNNG	372.67	1	**<0.0001**
LSTM < Top-down RNNG	39.202	1	**<0.0001**
LSTM < Left-corner RNNG	37.242	1	**<0.0001**
Top-down RNNG < LSTM	36.796	1	**<0.0001**
Top-down RNNG < Left-corner RNNG	40.762	1	**<0.0001**
Left-corner RNNG < LSTM	8.0157	1	**0.004637**
Left-corner RNNG < Top-down RNNG	13.941	1	**0.0001886**

Goodkind ら [12] では言語モデリング精度の高い言語モデルほど認知的妥当性が高いことが確認されている. また, Hale ら [10] では, 単語ビーム幅が大きくなると, RNNG の構文解析精度が高くなることが確認されている. これらのことから, より大きな単語ビーム幅を持つ RNNG であれば, 工学的精度が向上し, top-down RNNG でも LSTM 言語モデルよりも高い認知的妥当性が得られる可能性がある.

この仮説を検証するために, 複数の単語ビーム幅 $k = \{10, 20, 40, 60, 80, 100\}$ を持つ RNNG を用いて, 工学的精度と認知的妥当性の関係を調べる[4].

結果

結果を図 4 に示す. 横軸が-perplexity を表し, 縦軸が deviance accuracy を表す. 1（赤）が LSTM 言語モデル, td[n]（青）が top-down RNNG, lc[n]（緑）が left-corner RNNG を表し, n は単語ビーム幅を表す. Top-down/left-corner RNNG 共に, 単語ビーム幅が

4) アクションビーム幅として $10k$, ファストトラック幅として $k/10$ を採用した.

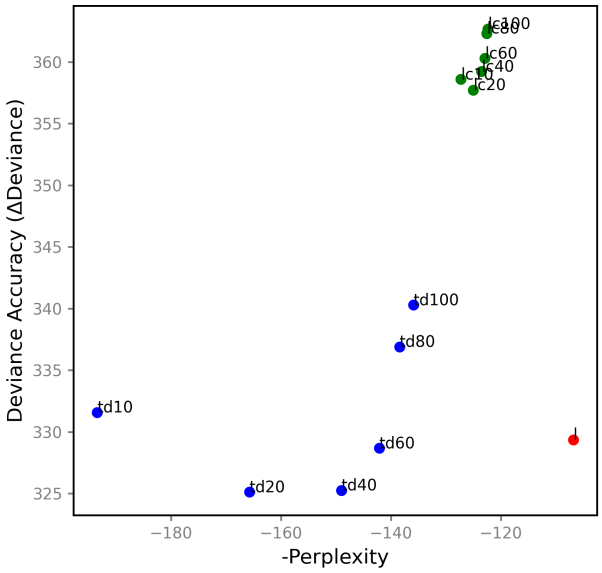

図 4 Perplexity と deviance Accuracy の関係. 横軸が-perplexity を表し, 縦軸が deviance acuracy を表す. 1（赤）が LSTM 言語モデル, td[n]（青）が top-down RNNG, lc[n]（緑）が left-corner RNNG, n は単語ビーム幅を表す.

大きくなるほど言語モデリング精度が向上した. Top-down/left-corner RNNG では, 言語モデリング精度が向上すると読み時間のモデル化精度も向上する傾向が観察された. 一方で, LSTM 言語モデルは, 様々なビーム幅の RNNG と比べ, 言語モデリング精度が高いが読み時間のモデル化精度が低かった. また, 仮説のように, top-down RNNG であっても, 言語モデリング精度が向上すると LSTM 言語モデルの deviance accuracy を上回った. RNNG の構文解析精度と deviance accuracy の関係については, 付録に示すが, 言語モデリング精度と同様の傾向が観察された.

6 おわりに

本研究では, left-corner parsing モデルが最も認知的に妥当であるが, 線形モデルや top-down parsing モデルも部分的には認知的に妥当であることが示唆された. また, 工学的精度の高い RNNG ほど, 認知的妥当性が高い一方で, LSTM 言語モデルの言語モデリング精度の高さは必ずしも認知的妥当性を意味しないことが示唆された.

謝辞

本研究は JSPS 科研費 JP19H04990, および国立国語研究所共同研究プロジェクト「大規模コーパスを利用した言語処理の計算心理言語学的研究」の助成を受けたものです.

参考文献

[1] Noam Chomsky. *Syntactic structures*. Mouton, The Hague, 1957. OCLC: 308125.

[2] Sepp Hochreiter and Jürgen Schmidhuber. Long Short-term Memory. *Neural computation*, Vol. 9, pp. 1735–80, December 1997.

[3] Martin Sundermeyer, R. Schlüter, and H. Ney. LSTM Neural Networks for Language Modeling. In *INTERSPEECH*, 2012.

[4] Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. Assessing the Ability of LSTMs to Learn Syntax-Sensitive Dependencies. *Transactions of the Association for Computational Linguistics*, Vol. 4, pp. 521–535, 2016.

[5] Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. What do RNN Language Models Learn about Filler–Gap Dependencies? In *Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP*, pp. 211–221, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[6] Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, Vol. 140, pp. 1–11, January 2015.

[7] Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. Recurrent Neural Network Grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pp. 199–209, San Diego, California, June 2016. Association for Computational Linguistics.

[8] Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. LSTMs Can Learn Syntax-Sensitive Dependencies Well, But Modeling Structure Makes Them Better. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1426–1436, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[9] Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. Structural Supervision Improves Learning of Non-Local Grammatical Dependencies. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 3302–3312, Minneapolis, Minnesota, June 2019. Association for Computational Linguistics.

[10] John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan Brennan. Finding syntax in human encephalography with beam search. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2727–2736, Melbourne, Australia, July 2018. Association for Computational Linguistics.

[11] Steven P. Abney and Mark Johnson. Memory requirements and local ambiguities of parsing strategies. *Journal of Psycholinguistic Research*, Vol. 20, No. 3, pp. 233–250, May 1991.

[12] Adam Goodkind and Klinton Bicknell. Predictive power of word surprisal for reading times is a linear function of language model quality. In *Proceedings of the 8th Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2018)*, pp. 10–18, Salt Lake City, Utah, January 2018. Association for Computational Linguistics.

[13] John Hale. A Probabilistic Earley Parser as a Psycholinguistic Model. In *Second Meeting of the North American Chapter of the Association for Computational Linguistics*, 2001.

[14] Roger Levy. Expectation-based syntactic comprehension. *Cognition*, Vol. 106, No. 3, pp. 1126–1177, March 2008.

[15] Stefan Frank and Rens Bod. Insensitivity of the Human Sentence-Processing System to Hierarchical Structure. *Psychological science*, Vol. 22, pp. 829–34, June 2011.

[16] Victoria Fossum and Roger Levy. Sequential vs. Hierarchical Syntactic Models of Human Incremental Sentence Processing. In *Proceedings of the 3rd Workshop on Cognitive Modeling and Computational Linguistics (CMCL 2012)*, pp. 61–69, Montréal, Canada, June 2012. Association for Computational Linguistics.

[17] Taku Kudo, Kaoru Yamamoto, and Yuji Matsumoto. Applying Conditional Random Fields to Japanese Morphological Analysis. In *Proceedings of the 2004 Conference on Empirical Methods in Natural Language Processing*, pp. 230–237, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[18] Yasuharu Den, Junpei Nakamura, Toshinobu Ogiso, and Hideki Ogura. A Proper Approach to Japanese Morphological Analysis: Dictionary, Model, and Evaluation. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco, May 2008. European Language Resources Association (ELRA).

[19] Christopher Manning and Hinrich Schutze. *Foundations of Statistical Natural Language Processing*. MIT Press, May 1999. Google-Books-ID: 3qnuDwAAQBAJ.

[20] 正幸浅原, 創小野, エジソン正宮本. BCCWJ-EyeTrack. 言語研究, Vol. 156, pp. 67–96, 2019.

[21] Kikuo Maekawa, Makoto Yamazaki, Toshinobu Ogiso, Takehiko Maruyama, Hideki Ogura, Wakako Kashino, Hanae Koiso, Masaya Yamaguchi, Makiro Tanaka, and Yasuharu Den. Balanced corpus of contemporary written Japanese. *Language Resources and Evaluation*, Vol. 48, No. 2, pp. 345–371, June 2014.

[22] Nathaniel J. Smith and Roger Levy. The effect of word predictability on reading time is logarithmic. *Cognition*, Vol. 128, No. 3, pp. 302–319, September 2013.

[23] Mitchell Stern, Daniel Fried, and Dan Klein. Effective Inference for Generative Neural Parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pp. 1695–1700, Copenhagen, Denmark, September 2017. Association for Computational Linguistics.

表 2　本研究で用いた説明変数.

変数名	型	概要
length	int	文字数
freq	int	文節内単語頻度の幾何平均
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
screenN	int	画面提示順
lineN	int	行提示順
segmentN	int	文節提示順
article	factor	記事情報
subj	factor	実験協力者 ID

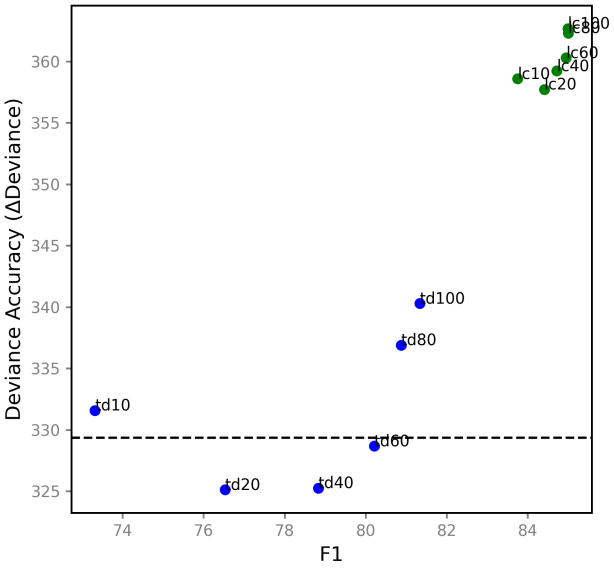

図 5　F1 bracket accuracy と deviance accuracy の関係. 横軸が F1 bracketing accuracy を表し, 縦軸が deviance accuracy を表す. td[n]（青）が top-down RNNG, lc[n]（緑）が left-corner RNNG を表し, n は単語ビーム幅を表す. 水平に引かれた点線は LSTM 言語モデルの deviance accuracy を表す.

A　説明変数

本実験で用いた説明変数を表 2 に示す. これらの説明変数は, 浅原ら [20] で用いられているものを踏襲しつつ, 本研究の分析対象外である空白情報に関する説明変数と, 係数が有意ではない記事提示順に関する説明変数を除いた. その上で, 先行研究 [15, 16, 10] で用いられている頻度情報を加えた.

B　構文解析精度

RNNG は, 単語列と階層構造の生成モデルであるため, RNNG の工学的精度には, 5 節で用いた言語モデリング精度の他に, 構文解析精度がある. RNNG の構文解析精度として, RNNG が単語ビーム幅内で最も高い確率を算出した階層構造の F1 bracket accuracy を用い, 認知的妥当性との関係を調べた. RNNG の構文解析精度と読み時間のモデル化精度の関係を図 5 に示した. 横軸が F1 bracket accuracy を表し, 縦軸が deviance accuracy を表す. td[n]（青）が top-down RNNG, lc[n]（緑）が left-corner RNNG を表し, n は単語ビーム幅を表す. 水平に引かれた点線は, LSTM の deviance accuracy を表す. top-down/left-corner RNNG 共に, 単語ビーム幅が大きくなるほど構文解析精度が向上した. 図 5 のように, top-down /left-corner RNNG では, 構文解析精度が向上すると読み時間のモデル化精度も向上する傾向が観察された. Top-down RNNG でも, 構文解析精度が高い場合には, LSTM 言語モデルの deviance accuracy を数値として上回ることも確認された.

C　学習用コーパスの分割

学習用コーパスである NPCMJ は, 14 の出典のテクストからなる. この各出典におけるテクストの, 90%を訓練データ, 5%を検証データとして用いた. 言語モデルの学習は, 訓練データにより行われ, 検証データにおける損失が 3 エポック連続で減少しなくなるまで行われた. 残り 5%はテストデータとして, 言語モデルの perplexity と RNNG の構文解析精度を測定する際に用いた.