F-3 再帰的ニューラルネットワーク文法によるヒト文処理のモデリング

吉田　遼[1]，能地　宏[2]，大関　洋平[3]
[1]東京大学，[2]産業技術総合研究所，[3]東京大学大学院総合文化研究科

要旨

近年、自然言語処理において、人工ニューラルネットワークを用いた高性能なニューラル言語モデルが多く提案されている。しかし、一般的なニューラル言語モデルは文を階層構造を持たない単語列として処理しており、ヒト文処理のモデルとしての妥当性は定かでない。一方、自然言語の階層構造を考慮したニューラル言語モデルも提案されており、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 Grammars, RNNG; Dyer et al., 2016）は、構文解析及び言語モデリングの二つのタスクで高い精度を達成している。これらの精度は工学的な評価指標によるものだが、本研究では、サプライザル理論（Hale, 2001）に基づき、言語モデルが推定した単語・文節の情報量（サプライザル）が、ヒトの視線計測データをどの程度説明できるか、という認知科学的な評価指標を用いる。実験の結果、再帰的ニューラルネットワーク文法のサプライザルが、階層構造を考慮しないニューラル言語モデルのサプライザルよりも読み時間のモデル化に有用であることが確認できた。ヒト文処理のモデルとして、階層構造を考慮しないモデルよりも、階層構造を明示的に考慮するモデルの方が妥当であることが示されたと言える。

1. はじめに

近年の自然言語処理では、人工ニューラルネットワークを用いたニューラル言語モデルによる文処理が主流になっている。Long Short Term Memory (LSTM) (Hochreiter and Schmidhuber, 1997) は、再帰的ニューラルネットワークの一種であり、その時系列データ処理への有用性から、言語モデルとして高い精度を達成している（Melis et al., 2018）。LSTM は、言語学で伝統的に扱われてきた自然言語の階層構造を考慮せず、単に文脈から次の単語の確率分布を算出することで、単語列を生成する。しかしながら、LSTM は単語間の長距離依存関係をある程度捉えられることができる (Linzen et al., 2016; Wilcox et al., 2018）など、言語モデルに階層構造を考慮するアーキテクチャを組み込む必要性に疑問を呈するような結果を出している。それに対し、再帰的ニューラルネットワーク文法（Recurrent Neural Network Grammars, RNNG; Dyer et al., 2016）は、ニューラルネットワークを用いて逐次的に階層構造と単語列を生成する。RNNG は、言語モデリング、構文解析という 2 つの自然言語処理タスクで高い精度を達成しており、また、LSTM よりも正確に長距離依存関係が捉えられることが示されている（Kuncoro et al., 2018; Wilcox et al., 2019）。このように、先行研究では、工学的精度や文法能力は、自然言語の階層構造を考慮しない言語モデルでもある程度高い精度を達成することができる一方で、自然言語の階層構造を考慮することでさらに向上することが示されてきた。

一方で、階層構造を考慮しない言語モデルと、階層構造を考慮する言語モデルの、ヒト文処理のモデルとしての妥当性を比較しようとする先行研究も存在する。近年、サプライザル理論（Hale, 2001）に基づき、言語モデルが推定した単語・文節の情報量（サプライザル）が、ヒトの視線計測データや脳波の回帰に有効であることが示されてきた（e.g. Goodkind and Bicknell, 2018）。この枠組みに則り、その回帰の精度を言語モデルの心理言語学的精度とみなすことで、言語モデルのヒト文処理のモデルとしての妥当性が比較されている。工学的指標や文法能力と異なり、言語モデルが自然言語の階層構造を考慮することの心理言語学的精度における有効性については、統一的な結論が得られていない。例えば、Frank and Bod (2011）、Frank et al. (2015）では、それぞれ、読み時間、脳波の予測において階層構造を考慮しないモデルの優位性を示す結果、Fossum and Levy (2012)、Hale et al. (2018）では階層構造を考慮するモデルの優位性を示す結果となっている。

これらの、言語モデルの心理言語学的精度に関する研究の問題点の一つは、ほとんどが英語に焦点を当てていることである。英語は SVO 語順であり、動詞をもとにして項構造を作ることができ

る。一方で、日本語は SOV 語順で、かき混ぜが可能な言語であり、読者は項構造を逐次的に予測しながら文処理を行う必要がある。このように、日本語は文処理に際しより多くの構造を予測する必要があると考えられ、階層構造を考慮しない言語モデルと、階層構造を考慮する言語モデルの、心理言語学的精度の比較に有用な言語であるといえる。よって、本研究では、日本語を用いて LSTM と RNNG の心理言語学的精度の比較を行う。さらには、異なるパージングストラテジーを持つ RNNG の心理言語学的精度を比較することで、ヒト文処理の妥当なパージングモデルについて検証する。

2. 実験

2.1. 言語モデル

2.1.1 Long Short Term Memory 言語モデル

単語列の生成モデルであり、階層構造を帰納バイアスとして持たない。本研究では、隠れ層のユニット数 256、単語埋め込み層のユニット数 256 の 2 層 LSTM を用いた。

2.1.2 再帰的ニューラルネットワーク文法（Recurrent Neural Network Grammars, RNNG）

RNNG (Dyer et al., 2016) は、再帰的ニューラルネットワーク（RNN）を用いたトップダウンな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 は、Stack 内に逐次的に句構造を生成していく。Stack LSTM は、それまでに生成された句構造全体を表現するベクトルを得る LSTM である。生成の各ステップでは、そのベクトル表現に基づき、以下の 3 つのアクションに対する確率分布を算出する。一つ目は、現在の句のノードの直下に新しい句を生成する NT、二つ目は、現在の句のノードの直下に新しい語を生成する SHIFT、三つ目は、現在の句を閉じる REDUCE である。最初の二つのアクションが選択された際には同じベクトル表現に基づき、それぞれ生成する句のラベル、生成する語の確率分布を算出する。REDUCE が選択された際には、その句のラベルとノードの直下に属する句や語が双方向 LSTM によって単一の部分木ベクトルとして集約される（図 2)。本研究では、Dyer et al. (2016) を踏襲し、隠れ層のユニット数 256 の 2 層 LSTM を Stack LSTM として用いた。

2.1.3 レフトコーナーRNNG

2.1.2 のように、Dyer et al. (2016) の RNNG は、トップダウンに句構造の生成を行う。本研究

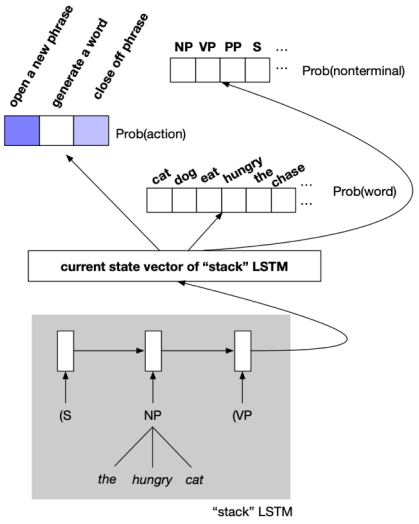

図 1: RNNG のアーキテクチャ。図は（Hale et al. 2018）より。

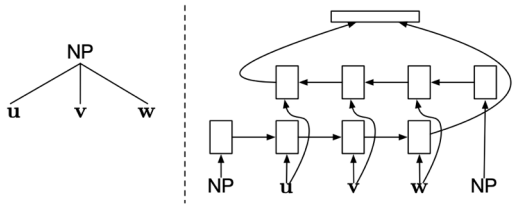

図 2: NP の娘ベクトルである u、v、w を、部分木ベクトルに集約する。図は(Dyer et al. 2016)より。

では、パージングストラテジーによる心理言語的精度の比較を行うために、新たにレフトコーナー

に句構造の生成を行う RNNG を導入する。レフトコーナーRNNG では、母ノードの句ラベルの生成が、その娘ノードのうちの最左の要素が生成、または部分木ベクトルに集約された後に行われる。その後、生成された母ノードの句ラベルと娘ノードのうちの最左の要素が Stack 内で入れ替わる。その他はトップダウン RNNG と同様にして、句構造の生成が行われる。本研究では、隠れ層のユニット数 256 の 2 層 LSTM を Stack LSTM として用いた。

2.2. 訓練データ、検証データ

Kaede treebank (Tanaka and Nagata 2013) を用いた。京都大学テキストコーパスの 1 万文に対して、国語研長単位を基準として句構造のアノテーションが付与されている。エラーを含む文を除いた 9014 文を 4:1 の割合で訓練データと検証データに分割した。テストデータにおける未知語を減らすため、国語研長単位品詞一覧（小椋秀樹ほか, 2011）に基づき、固有名詞、記号、補助記号をそれぞれ 6、2、7 種類のカテゴリに抽象化した。RNNG の教師データは、単語列の他に構成素境界と、機能を示す拡張タグを除いた文法範疇を含むが、LSTM の教師データは単語列のみである。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った時点で訓練を終了した。

2.3. 視線計測コーパス

BCCWJ-EyeTrack（浅原, 小野, 宮本 2019）を用いた。『現代日本語書き言葉均衡コーパス』(Maekawa et al. 2014)のコアデータの新聞データの一部に対して 24 人の日本語母語話者の視線走査法と自己ペース読文法を用いた読み時間付与がなされている。

Demberg and Keller (2008) を踏襲し、データポイントのうち、視線が停留していないもの、行頭または行末に提示されたもの、句読点を含むもの、二つ以上の大文字英字を含むもの、文字以外を含むものについては、読み時間が文処理負荷を純粋に反映しているとはいえないため、分析から除外した。当該処理によりデータポイントの大部分が除外された（約 69%が除外され、6140 のデータポイントが分析対象となった)。

2.4. サプライザル

サプライザルは、以下の式で与えられる文脈（context）における当該単語や文節（x）の負の対数確率であり、当該単語・文節の情報量を表す。サプライザルが大きいほど、文脈から当該単語・文節を予測することが難しく、当該単語・文節の処理負荷が大きくなると考えられている。

$$surprisal(x) = -\log p(x|\text{context})$$

本研究では、言語モデルの算出したサプライザルを用いて、ヒト文処理の負荷を反映していると考えられる視線計測データを回帰することで、言語モデル間の心理言語学的精度を比較する。

2.4.1 単語サプライザル

LSTM の単語サプライザルは、言語モデルが算出した当該単語の文脈条件付き確率$p(x|\text{context})$から直接求められる。

RNNG の単語サプライザルは、Hale et al. (2018) を踏襲し、ビームサーチ（Stern et al., 2017）を用いて求める。単語列から予想される階層構造のうち確率の高いもの複数保持しつつ、それらの確率を階層構造について周辺化することで、$p(x|\text{context})$を求める。本研究では、Hale et al. (2018) を踏襲し、生成の各アクションの上位いくつまでを保持するか、を表すアクションビーム幅として$k = \{100, 200, 400, 600, 800, 1000\}$を採用した。

文節サプライザル

言語モデルは訓練データの単語分割単位である国語研長単位に対して、単語単位で確率を付与するが、日本語では読み時間は文節単位に対して付与される。このように、日本語では生起確率と読

み時間の集計単位に齟齬が生じているが、本研究では、単語サプライザルの和を、文節サプライザルとして用いる。

2.5. 評価指標

Frank and Bod (2011) に則り、言語学的精度（言語モデルの言語モデリング精度）と、心理言語学的精度（言語モデルの読み時間に対する説明力）を評価する。

2.5.1 言語学的精度

モデルのテストデータ全体に対するサプライザルの平均値の負で定義される。値が大きいほど、テストデータを高い精度で予測できる、より良い言語モデルであるといえる。

2.5.2 心理言語学的精度

読み時間に関係するとされている文節長などの説明変数で読み時間を回帰した線形混合モデルに、それぞれの言語モデルのサプライザルを説明変数として加えた際の、deviance の減少分で定義される。線形混合モデルには、R (Baayen et al., 2008) の、`lmer` パッケージを用いた。ベースラインの回帰モデルは、Frank and Bod (2011) を踏襲し、`sentpos`（文中の文節位置）、`nrchar`（文節中の文字数）、`prevfix`（直前の文節に視線停留があったか否か）、`nextfix`（直後の文節に視線停留があったか否か）、`freq`（文節頻度）を説明変数に持つ。浅原 (2019) に倣い、文節頻度としては、『国語研日本語ウェブコーパス』(NWJC) (Asahara et al., 2014) によって算出された文節中の単語頻度の相乗平均を用いた。Frank and Bod (2011) では、ベースラインの回帰モデルの説明変数に、この他に単語のバイグラム頻度を持つが、日本語では文節単位のバイグラム頻度の算出が難しいため、本実験では除いている。回帰の際にはこれらすべての説明変数、説明変数の交互作用のうち有意なもの、被験者・文節のランダム切片、を用いて読み時間のうちの Total Reading Time をモデリングした。Frank and Bod (2011) では、被験者のランダムスロープのうち最も効果が有意なものを加えているが、本研究ではこれ（`nrchar` の被験者のランダムスロープ）を加えたベースラインモデルに LSTM のサプライザルを加えた際に、回帰モデルが収束しなかったため、除いた。

二つの言語モデル A、B の心理言語学的精度間（ここでは A の心理言語学的精度>B の心理言語学的精度、とする）に有意差があるかどうかは、両方の言語モデル A、B のサプライザルを含む回帰モデルが、一つの言語モデル B のサプライザルのみを含む回帰モデルよりも、λ^2テスト($p \leq 0.05$)下で有意に精度が高いか、で評価した。有意に精度が高い場合、言語モデル A が、言語モデル B よりも有意に心理言語学的精度が高いといえる。

3. 結果

Total Reading Time に対する結果を図 3 に示した。x軸に言語学的精度、y軸に心理言語学的精度をプロットした。l が LSTM、r[n]がトップダウン RNNG、c[n]がレフトコーナーRNNG を表し、n は(アクションビームの幅/100) を表す。すべてのサプライザルについて、説明変数として加えた際にベースラインの回帰モデルよりも有意に精度が向上した ($\lambda^2 > 32.3, p < 0.0001$)。すべての RNNG が、LSTM よりも有意に心理言語学的精度が高かった ($\lambda^2 > 11.405, p < 0.001$)。また、すべてのレフトコーナーRNNG が、すべてのトップダウン RNNG よりも有意に心理言語学的精度が高かった($\lambda^2 > 33.516, p < 0.0001$)。トップダウン RNNG 内では、いくつかの組み合わせでのみ心理言語学的精度に有意差がみられた。r2 はすべてのその他のトップダウン RNNG よりも心理言語学的精度が低く($\lambda^2 > 6.6157, p \leq 0.01011$)、また、r1 は r10 よりも有意に心理言語学的精度が低かった($\lambda^2 = 3.8606, p = 0.04943$)。また、レフトコーナーRNNG 内で心理言語学的精度に有意差がみられたのは、c6>c8 だけであった($\lambda^2 = 4.0396, p = 0.04444$)。

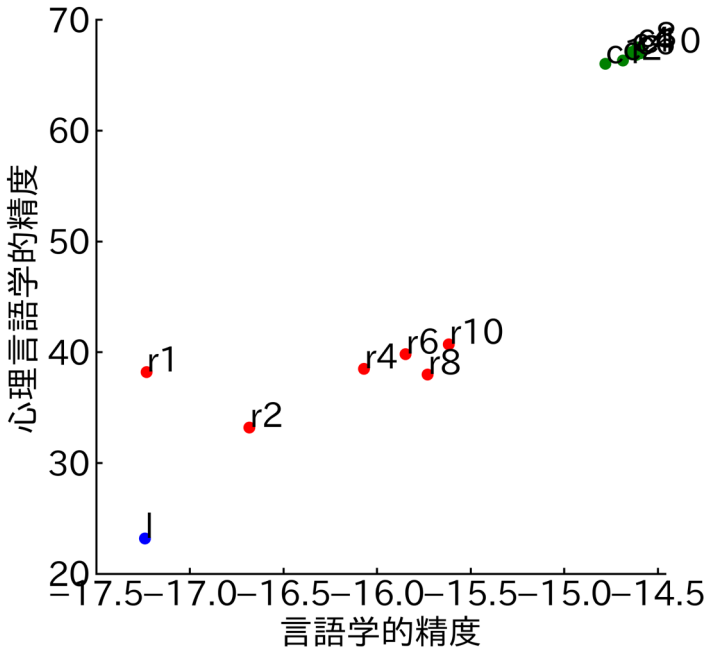

図 3: Total Reading Time に対する結果。x軸に言語学的精度、y軸に心理言語学的精度をプロットした。言語学的精度はサプライザルの平均値の負、心理言語学的精度は言語モデルのサプライザルによる回帰モデルの deviance の減少分を表す。l（青）が LSTM、r[n]（赤）がトップダウン RNNG、c[n]（緑）がレフトコーナ RNNG を表し、n は(アクションビームの幅/100)を表す。

4. 考察

階層構造を考慮する言語モデルである RNNG が、階層構造を考慮しない言語モデルである LSTM よりも、心理言語学的精度が高いという結果が得られた。これは、自然言語の階層構造を考慮する言語モデルが、自然言語の階層構造を考慮しない言語モデルよりも、ヒト文処理のモデルとして妥当であるということを示唆する。LSTM のような、階層構造を考慮せず語順のみを扱う言語モデルは、出現した単語列にのみ基づいて、単に次の単語の確率を決定するため、チョムスキー階層（Chomsky, 1956）における有限状態文法に相当する生成力を持つ。一方で、RNNG のような、階層構造を考慮するモデルは、単語を句にまとめ、さらにその句同士を上位の句にまとめるという操作が可能であり、チョムスキー階層における文脈自由文法に相当する生成力を持つ。自然言語には有限状態文法では記述できない、再帰や長距離依存といった文が存在する（Chomsky, 1957）。よって有限状態文法までしか扱えない LSTM ではヒト文処理の近似に限界があり、よりヒト文処理に近いモデルには、文脈自由文法を扱える、RNNG のように階層構造を考慮するモデルが妥当なのだと考えられる。

また、日本語におけるヒト文処理の際のパージングストラテジーとしては、レフトコーナーが、トップダウンよりも妥当であることを示唆する結果が得られた。レフトコーナーRNNG 内では、ビーム幅による言語学的精度・心理言語学的精度の差は見られなかったが、トップダウン RNNG 内では、ビーム幅が大きいほど言語学的精度が高くなる傾向が見られた。心理言語学的精度については、アクションビーム幅が一定以上（>400）になると有意な差は存在しないが、ビーム幅が小さい時（<200）には、ビーム幅が大きいものよりも低くなる傾向が見られた。ヒト文処理では、可能な階層構造のうち確率の高い複数のものについて並列処理が行われているという仮説がある（cf.

Jurafsly, 1996）。少なくとも、トップダウン RNNG の結果からは、保持する構造の幅が少なく、直列処理に近くなると、ヒト文処理のモデルとしての妥当性が低くなることが示唆される。

最後に、本研究の課題点を挙げる。まず、LSTM、トップダウン RNNG、レフトコーナーRNNG、の間の心理言語学的精度の差が生じている原因について、詳細な分析が行われていない点である。これらの差が生じるデータポイントについて特定し、ヒト文処理について得られている知見が、言語モデルでも観察されるかを調べる必要がある。

次に、訓練データの数が少なく、単語単位のサプライザルが国語研長単位に対して付与されるため、テストデータ内に未知語を含むデータポイントが残されている点である。文の前処理によってデータポイントの多くが失われていることもあり、さらに未知語を含む文節をデータポイントから除外してしまうと、回帰モデルが収束しなくなってしまう。そのため、未知語を含むデータポイントを本研究では除外しなかった。今後は、単語分割単位が国語研短単位の訓練データや、より大規模なデータである NINJAL Parsed Corpus of Modern Japanese (NPCMJ, http://npcmj.ninjal.ac.jp) での言語モデルの訓練などによって、テストデータ内の未知語を減少させ、より正確な分析をする必要があるといえる。

5. 結論

本研究では、日本語を用いて LSTM と RNNG の心理言語学的精度の比較を行った。さらには、異なるパージングストラテジーを持つ RNNG の心理言語学的精度を比較することで、ヒト文処理の妥当なパージングモデルについて検証した。その結果、RNNG のように階層構造を明示的に考慮する、レフトコーナーなパージングモデルの、ヒト文処理のモデルとして優位性が示唆された。

参照文献

Msayuki Asahara, Kikuo Maezono, Mizuho Ikeda, Sachi Kato, and Hiraki 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*, 25 (1–2): 129–148.

R. H. Baayen, D. J. Davidson, and D. M. Bates. (2008) Mixed-effects modeling with crossed random effects for subjects and items. In *Journal of Memory and Language,* 59: 390-412.

Noam Chomsky. (1956) Three models for the description of language. *IRE Transactions on Information Theory* (2): 113-124

Noam Chomsky. (1957) *Syntactic Structures*. Mouton, The Hague/Paris.

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

Chris Dyer, Adhiguna Kuncoro, Miguel Ballesteros, and Noah A. Smith. (2016) Recurrent neural network grammars. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*: 199–209, San Diego, California.

Victoria Fossum and Roger Levy. (2012) *Sequential vs. hierarchical syntactic models of human incremental sentence processing.* In *Proceedings of the 3rd work- shop on cognitive modeling and computational linguistics*: 61–69.

Stefan L. Frank, Leun J. Otten, Giulia Galli, and Gabriella Vigliocco. (2015) The ERP response to the amount of information conveyed by words in sentences. *Brain and Language*, 140:1–11.

Stefan L. Frank and Rens Bod. (2011) Insensitivity of the human sentence-processing system to hierarchical structure. *Psychological Science*, 22(6): 829–834.

Adam Goodkind and Klinton Bicknell. (2018) 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): 10–18.
John Hale. (2001) A probabilistic Earley parser as a psycholinguistic model. In *Proceedings of the second meeting of the north american chapter of the association for computational linguistics on language technologies*: 199–209.
John Hale, Chris Dyer, Adhiguna Kuncoro, and Jonathan R. Brennan. (2018) Finding Syntax in Human Encephalography with Beam Search. In *Proceedings of ACL*.
Sepp Hochreiter and Jürgen Schmidhuber. (1997) Long short-term memory. *Neural Computation*, 9(8):1735-1780.
Daniel Jurafsky. (1996) A probabilistic model of lexical and syntactic access and disambiguation. Cognitive Science 20 (2): 137-194.
Adhiguna Kuncoro, Chris Dyer, John Hale, Dani Yogatama, Stephen Clark, and Phil Blunsom. (2018) 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)*, 1: 1426–1436.
Tal Linzen, Emmanuel Dupoux, and Yoav Goldberg. (2016) Assessing the ability of lstms to learn syntax-sensitive dependencies. *Transactions of the Associ- ation for Computational Linguistics*, 4:521–535.
Gabor Melis, Chris Dyer, and Phil Blunsom. (2018) On the state of the art of evaluation in neural language models. In *Proc. of ICLR*.
Mitchell Stern, Daniel Fried, and Dan Klein. (2017) Effective inference for generative neural parsing. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*: 1695–1700, Copenhagen, Denmark.
Takaaki Tanaka, and Masaaki Nagata. (2013) Constructing a Practical Constituent Parser from a Japanese Treebank with Function Labels. *Proceedings of 4th Workshop on Statistical Parsing of Morphologically-Rich Languages (SPMRL'2013)*: 108–118.
Ethan Wilcox, Roger Levy, Takashi Morita, and Richard Futrell. (2018) What do rnn language models learn about filler-gap dependencies? *arXiv preprint arXiv:1809.00042*.
Ethan Wilcox, Peng Qian, Richard Futrell, Miguel Ballesteros, and Roger Levy. (2019) 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*: 3302–3312.
浅原正幸（2019）「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・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（*Gengo Kenkyu*) 156: 67–96
小椋秀樹ほか(2011）「国立国語研究所内部報告書『現代日本語書き言葉均衡コーパス』形態論情報規程集 第 4 版」