言語モデルの文法知識評価における間接肯定証拠の分析

大羽未悠 [1,3]　大関洋平 [2]　深津聡世 [2]　芳賀あかり [1]　大内啓樹 [1]　渡辺太郎 [1]　菅原朔 [3]
[1] 奈良先端科学技術大学院大学　[2] 東京大学　[3] 国立情報学研究所
{oba.miyu.ol2, haga.akari.ha0, hiroki.ouchi, taro}@is.naist.jp
{oseki, akiyofukatsu}@g.ecc.u-tokyo.ac.jp　saku@nii.ac.jp

概要

なぜ大規模言語モデルは人間の数千倍以上の量の訓練データを必要とするのだろうか？本研究では，人間が未知の言語現象を含む文の容認性を関連する観測事例（間接肯定証拠）から類推可能であることに着目する．言語モデルも人間と同様に間接肯定証拠を利用できるかを分析することで，間接肯定証拠の利用可否が人間と言語モデルのデータ効率の差の要因となる帰納バイアスの候補になりうるかを調査する．実験の結果，少なくとも本研究の範囲では，言語モデルは学習時に間接肯定証拠を利用していないことが明らかとなり，このことがデータ効率の差を引き起こしている可能性が示唆された．

1　はじめに

昨今様々なタスクで進展を遂げた言語モデルは，学習時に大量のデータを使用する．例えば，大規模言語モデルは，子どもが大人と同じレベルの文法を獲得するまでに触れるデータ量の数千倍以上のデータで学習しており [1]，人間よりも非効率な学習をしていると考えられる．言語モデルと人間のデータ効率の差にどんな帰納バイアスが起因するかについての検証を見据え，言語モデルが人間のような言語知識を獲得できるか，何が効率的な言語獲得に寄与するのかについて，モデルアーキテクチャや訓練データなどの観点から関心が寄せられている [2, 3]．言語モデルは，ラベルのない自然なテキストデータから階層的な汎化を獲得できることが知られており [4, 5]，そのデータ量や種類は言語獲得に異なる影響を与える．どのようなドメインののデータが寄与するかについて，例えば子供向けに発話されたテキストからなるコーパスは同量の Wikipedia よりも人間らしい言語知識の獲得に良い影響を与えることが知られている [6, 3]．一方で，より詳細な訓練事例単位での分析は限られており，データのどのような要素が言語知識の獲得に影響を与えているのかは明らかになっていない．

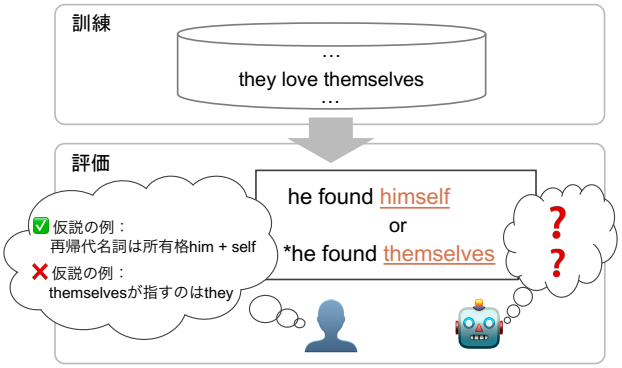

図 1　間接肯定証拠の例．人間は，観測された事例（図上部）から何らかの仮説を立て，与えられた文（図下部）が文法的に正しいかを推論できる．本研究では，言語モデルが間接肯定証拠を使用しているか調査する．

本研究では，訓練事例レベルでどのようなデータが言語獲得に寄与するか，人間の言語獲得における「証拠」の観点から分析する．子どもが親の発話などから観測した文に含まれる文法現象の情報をそのまま学習するとき，その文は直接肯定証拠と呼ばれる．一方で，Pearl ら [7] によると観測した事例に何らかの関連する言語現象が含まれる文の容認性を，何らかの仮説を立てて推論するとき，その情報を間接肯定証拠と呼ぶ．図 1 に例を示す．人間は観測事例（図 1 上部）から，例えば活用や再帰代名詞の接尾辞のような何らかの規則に関する仮説を立て，与えられた文（図 1 下部）が文法的に正しいかを推論できるということが明らかになってきている [8]．しかし，言語モデルが同様に間接的な情報を用いて学習しているかは明らかではなく，学習のための情報に何を選ぶかという帰納バイアスの解明が効率的な言語獲得の実現には有用だと考えられる．

言語モデルがデータのどのような情報を証拠として用いているか調査するため，文法知識を問う評価データを解く際の各訓練事例の寄与度を計算す

る．まず，寄与度の計算の代表的な手法である影響関数 [9] が言語モデルの訓練でも利用可能かを検証する．既存の影響関数は，GLUE [10] のようなラベル付きの評価事例に対してラベル付きの訓練事例の寄与度の測定を対象にしている．本研究は，自然な言語獲得のシナリオとして一般的な言語モデルの学習を行い，文法的な文と非文法的な文の尤度による識別タスクで評価を行う点で既存の影響関数の用途と異なる．そこで，本研究の目的に沿う形で影響関数を修正し，計算された寄与度が高い事例が実際に文法知識の獲得に寄与しているかを調査し，実際に文法知識の性能を大きく変化させる傾向が観察された．

次に，言語モデルが間接肯定証拠から言語知識を学習しているかについて，間接肯定証拠の対象となる可能性のある訓練例とその寄与度の観点から分析した．その結果，少なくとも本実験の範囲では，現行の言語モデルは間接肯定証拠からの学習はしていないことが明らかになり，間接肯定証拠を使うための何らかの仮説が構築できるようなモデルを実装することで，人間のようなデータ効率の良い言語獲得ができる可能性があることが示唆された．

2 言語モデルの文法知識の評価における影響関数の検証

2.1 文法知識評価タスク

言語モデルの文法知識の評価には，与えられた文の文法性を判断させる BLiMP などのベンチマークが用いられる [11]．BLiMP は 12 種類の英語の言語現象から構成される．各言語現象に文法的に正しい文（1a）とそうでない文（1b）の対が数千ずつ含まれており，言語モデルがどちらを好むかを調査する．

（1a） Many teenagers were helping **themselves**.

（1b）* Many teenagers were helping **herself**.

穴埋め言語モデルを評価する場合，振る舞いを調査するための標準的な指標である疑似パープレキシティ (PPPL) [12] を利用できる．Salazar ら [12] に従い，文 $\boldsymbol{s} = [w_1, w_2, \cdots, w_n]$ の PPPL は，穴埋め言語モデルを使用して式 (1) で計算される．

$$\mathrm{PPPL}(\boldsymbol{s}, \theta) = \prod_{t=1}^{n} p_\theta(w_t | \boldsymbol{s}_{\setminus w_t})^{-\frac{1}{|\boldsymbol{s}|}} \tag{1}$$

w_t は文中の t 番目のトークン，$\boldsymbol{s}_{\setminus w_t}$ は文中の w_t を除く全てのトークンの系列を表す．$p_\theta(w_t | \boldsymbol{s}_{\setminus w_t})$ は，文脈 $\boldsymbol{s}_{\setminus w_t}$ が与えられたときに言語モデル θ が w_t を予測する確率である．

BLiMP の文対のうち，文法的に正しい文に低い PPPL が割り当てられた文対の割合をスコアとして，各言語現象における性能を求めることで，言語モデルの文法知識を測ることができる．

2.2 影響関数

影響関数とは，各訓練事例がモデルの推論結果にどの程度影響するかを近似的に計算する手法である [9, 13]．Koh ら [9] は，ある評価事例 z_{test} に対するある訓練事例 z_{train} の影響関数 $\mathcal{I}$ を式 (2) の通り定義している．$\mathcal{I}(z_{\text{train}}, z_{\text{test}})$ の値が小さい訓練事例ほど有用な事例であり，その事例を除いたデータセットで訓練したモデルは評価データに対する損失が大きくなる．値が大きくなるほど，有害な事例となり，その事例を除いたデータセットは対象の評価データに対する損失を小さくさせる．

$$\mathcal{I}(z_{\text{train}}, z_{\text{test}}) := -\nabla_\theta L(z_{\text{test}}, \hat{\theta})^T H_{\hat{\theta}}^{-1} \nabla_\theta L(z_{\text{train}}, \hat{\theta}) \tag{2}$$

ここで，$L(z, \hat{\theta})$ は事例 z に対するモデル $\hat{\theta}$ の損失関数であり，$H_{\hat{\theta}}^{-1}$ はモデルパラメータ $\hat{\theta}$ のヘッセ行列である．既存手法は，訓練・評価事例とも GLUE のようなラベル付きのファインチューニングのためのデータを用いている．

2.3 言語モデルの評価のための影響関数

本研究では，ラベルなしテキストでの訓練によりモデルの言語知識の獲得を調査するため，式 (2) の損失関数の代わりに PPPL を使用する．また，文法的な文に低い PPPL，非文法的な文に高い PPPL を割り当てるモデルは目的の言語知識を獲得していると考える．評価事例のうち文法的な文を $z_{\text{test}}^{\text{good}}$，非文法的な文を $z_{\text{test}}^{\text{bad}}$ とする．$z_{\text{test}} = (z_{\text{test}}^{\text{good}}, z_{\text{test}}^{\text{bad}})$ とし，式 (3) をある評価事例 z_{test} に対するある訓練事例 z_{train} の寄与度と新たに定義する．$\mathcal{I}(z_{\text{train}}, z_{\text{test}}^{\text{good}})$ が小さいほど z_{train} が文法的な文 $z_{\text{test}}^{\text{good}}$ に低い PPPL を与えるのに寄与し，$\mathcal{I}(z_{\text{train}}, z_{\text{test}}^{\text{bad}})$ が大きいほど z_{train} が非文法的な文に高い PPPL を与えるのに寄与するため，式 (3) のように差を取ることで，ある訓練事例がどれほど寄与するかを測れる．寄与度が大きいほど，その訓練事例が評価事例を正しく判断するのに役立つことを示唆する．

$$\Delta\mathcal{I}(z_{\text{train}}, z_{\text{test}}) = \mathcal{I}(z_{\text{train}}, z_{\text{test}}^{\text{bad}}) - \mathcal{I}(z_{\text{train}}, z_{\text{test}}^{\text{good}}) \tag{3}$$

BLiMP は各言語現象に対して複数の評価事例が

表 1 Ana. agr の評価事例に寄与した上位 0, 1, 5, 20%の訓練事例をランダムな文に置き換えて訓練したモデルを BLiMP の各言語現象で評価したスコア．

言語現象	寄与度上位 n%を置き換え			
	$n = 0$	$n = 1$	$n = 5$	$n = 20$
Ana. agr	**83.3**	**75.1**	**73.8**	**70.5**
D-n agr	64.4	63.2	62.4	63.6
Irregular	69.6	69.7	68.1	68.5
S-v agr	60.1	59.3	59.2	60.0
Arg. str	89.3	87.9	88.8	87.4
Ellipsis	78.5	78.0	78.8	77.1
Filler-gap	61.3	63.0	63.1	61.3
Island	92.8	91.0	89.4	89.6
NPI	38.5	40.6	42.3	40.0
Quantifiers	58.1	54.4	54.0	49.6
Binding	59.8	64.0	62.6	61.6
Ctrl. Rais.	76.0	76.2	75.7	75.3

含まれる．ある言語現象の評価事例の集合を $\mathcal{X}_{\text{test}}$ としたとき，その言語現象に対する訓練事例 z_{train} の寄与度 $\mathcal{I}_{\text{phen}}$ を式 (4) のように定義する．

$$\mathcal{I}_{\text{phen}}(z_{\text{train}}, \mathcal{X}_{\text{test}}) = \frac{1}{|\mathcal{X}_{\text{test}}|} \sum_{z_{\text{test}} \in \mathcal{X}_{\text{test}}} \Delta\mathcal{I}(z_{\text{train}}, z_{\text{test}}) \quad (4)$$

2.4 寄与度の大きい訓練事例の効果

定義した影響関数の評価として，ある言語現象の評価データに対して計算された寄与度 $\mathcal{I}_{\text{phen}}$ が大きい訓練事例が，実際にその現象に関する知識獲得に寄与しているかを調査する．訓練データには Wikipedia を 2,000 万単語分，約 670,000 文を用いた．寄与度を測る評価事例は，BLiMP の各言語現象の評価データからランダムに 100 文対ずつ取得する．モデルは，RoBERTa [14] のパラメータや訓練データを小規模にした BabyBERTa [6] を用いる．

寄与度の高いデータは実際の文法獲得に寄与するか 当該言語現象に対して寄与度の高い訓練データをソートし，上位 1, 5, 20%の文を除去し，トークン数を維持したまま Wikipedia で訓練データに用いていない文と置き換える．置き換える前と後の訓練データで訓練したモデルをそれぞれ評価し，寄与度が上位の文が実際に寄与していたかを調べた．

表 3 は，Anaphor agreement（3.1 節参照）の評価事例に寄与した上位 0, 1, 5, 20%の文をランダムな文に置き換えたデータで訓練したモデルを BLiMP の各言語現象で評価した結果である．Ana.agr は寄与度の高い例を置き換える割合が大きくなるほどに下がっている傾向が見られる．わずか上位 1%のデータ，約 6,700 文を置き換えただけでも，スコアが十分に下がっている．また，Ana.agr が飛び抜けて大きく下がり他の言語現象は少し下がるかほとんど変わらない結果となり，ターゲットの言語現象に寄与する訓練事例は実際にその現象に関する知識獲得に寄与することが示唆された．NPI（否定極性項目）は他と比較するとスコアが下がることが確認された．Ana.agr と NPI はどちらも c-command が関連する言語現象であり，言語現象間の言語学的な類似度と寄与度の高い訓練事例の重複に関する解釈は今後の課題としたい．一方で，他の現象に対する寄与度が上位の訓練事例を置き換えて学習した場合は，Ana.agr よりも顕著でない結果となったことから，寄与度の分散の大きさと結果の関連性の分析などが今後必要となる．

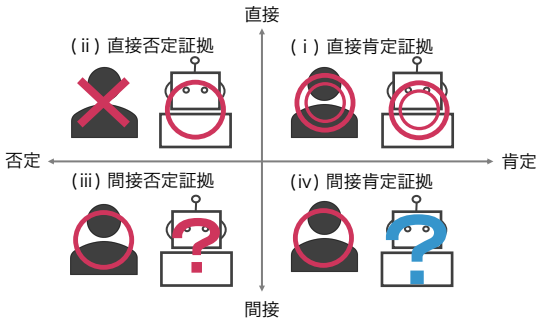

図 2 言語獲得における証拠の分類．肯定か否定か，直接か間接の 2 軸に分けられる．本研究では言語モデルの間接肯定証拠を分析対象とする．○・×は当該証拠を使用するか否かを示し，？は未検証であることを示す．

3 間接肯定証拠の分析

言語獲得の分野では，言語知識の学習に用いられる情報を証拠と呼ぶ．証拠はその性質に応じて 2 軸に分けられる（図 2）．ある表現がその言語で現れるかどうかによって肯定・否定証拠に分けられ，その文の文法性が明らかか，それとも学習者の推論が必要かによって直接・間接証拠に分けられる [7]．

直接肯定証拠（図 2 (i)）は，学習者が観測するデータの中に出現する情報そのものであり，話者が使用すること自体が文法的であるという前提で学習に用いられる．従来は人間が利用できる唯一のデータと考えられてきた [15]．言語モデルも同じく入力テキストから出現確率を学習するため，明らかに利用可能な証拠である．直接否定証拠（図 2 (ii)）は，あるデータが非文法的であることを明示的に知らされることである．人間の言語獲得時は，特に構文知識に関する明示的な訂正の利用ができないまたは無視すると考えられてきた [16]．一方で，言語モデル

は直接否定証拠により統語能力が向上するという知見がある [17]．間接否定証拠（図 2 (iii)）は，ある情報をもつ文がデータに存在しないことから，そのような文は非文法的だと推論することを指す．間接否定証拠に関しては，人間の言語獲得における有効性が唱えられている [18]．間接肯定証拠（図 2 (iv)）は，出現したデータから初見のデータに対して何らかの仮説を基に推論することである．人間の言語獲得に有益なデータとして認識されるようになったのは，比較的最近のことである [8]．これら間接証拠から言語モデルが学習可能かは明らかになっていない．間接否定証拠からの学習可能性を調べるには，「非文が出現しないこと」をモデルが認識している状態を観測する必要があり難しいため，本研究では事例の出現について寄与度による観測が可能な間接肯定証拠の検証を優先する．

3.1 実験設定

言語現象 Ana. agr をケーススタディとし，言語モデルが間接肯定証拠を利用しているかについて分析を行う．Ana. agr は，例えば文法的に正しい文（2a）と正しくない文（2b）から構成される [1]．

（2a） He found **himself**.

（2b）* He found **themselves**.

このような文対の文法性を正しく判断するために，言語モデルが間接肯定証拠を用いているのか，直接肯定証拠のみかを明らかにしたい．もし直接肯定証拠を使用していれば，文法的な評価事例と同じ再帰代名詞を含む訓練事例（3a）の寄与度が高くなることが期待できる．間接肯定証拠を使用していたら，非文法的な評価事例と同じ再帰代名詞を含む訓練事例（3b）や，評価事例に含まれていない再帰代名詞を含む訓練事例（3c）の寄与度も高くなることが予想される．

（3a） He hurt **himself**.

（3b） They love **themselves**.

（3c） She saw **herself**.

実験では，訓練データのうち Ana. agr が含まれる事例を対象とする．評価事例は 2.4 節と同様に Ana. agr から 100 例使用する．

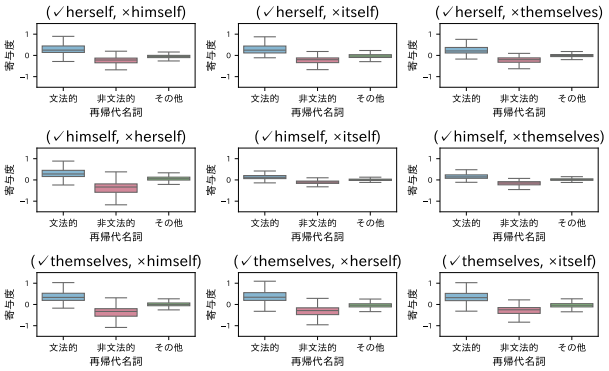

図 3 Ana. agr の訓練事例を，ある評価事例のうち文法的な文（✓）と同じ再帰代名詞を含む事例，非文法的な文（✗）と同じ再帰代名詞を含む事例，その評価事例に含まれていない再帰代名詞を含む事例に分け，その評価事例に対して計算された各訓練事例の寄与度の分布．

3.2 結果

図 3 は，訓練事例を，ある評価事例のうち文法的な文（✓）と同じ再帰代名詞を含む事例，非文法的な文（✗）と同じ再帰代名詞を含む事例，その評価事例に含まれない再帰代名詞を含む事例に分け，その評価事例に対する各訓練事例の寄与度の分布を表している [2]．どの評価事例の対でも直接肯定証拠となる再帰代名詞が含まれた訓練例の寄与度が高く，間接肯定証拠を表す再帰代名詞が含まれた訓練例の寄与度は低い結果となった．現行の言語モデルは，少なくとも本実験の範囲では，間接肯定証拠からの学習は見られなかった．間接肯定証拠を使うための何らかの仮説を構築できるようなベイジアンモデルなどを利用することで，人間のようなデータ効率の良い言語獲得に近づく可能性があり，今後の課題としたい．

4 おわりに

本研究では，訓練事例単位でどのようなデータが言語知識の獲得に寄与しているか，主に Ana. agr を対象に間接肯定証拠の観点から調査を行った．その結果，少なくとも本実験設定の範囲では，言語モデルは間接肯定証拠から学習しておらず，人間の学習時の帰納バイアスとは異なることが示唆された．本実験で得られた人間との帰納バイアスのギャップを埋めるような機構を言語モデルに実装してデータ効率の向上を試みることが今後の課題となる．

1） BLiMP の Ana. agr（照応の一致）は全て目的語に再帰代名詞が位置している文から構成されている．言語モデルの文法能力を測るために代表的な言語現象の種類である一致の中では，文法性が一つの単語の出現で判定可能で分析しやすいことから選択した．

2） 評価事例の文法的な文と非文法的な文の再帰代名詞の対が 2 例以上評価事例に存在する再帰代名詞のみ記載した．

謝辞

本研究は JSPS 科研費 JP21H05054, JST さきがけ JPMJPR21C2, JST さきがけ JPMJPR20C4 の助成を受けたものです.

参考文献

[1] Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell. Findings of the BabyLM challenge: Sample-efficient pretraining on developmentally plausible corpora. In Alex Warstadt, Aaron Mueller, Leshem Choshen, Ethan Wilcox, Chengxu Zhuang, Juan Ciro, Rafael Mosquera, Bhargavi Paranjabe, Adina Williams, Tal Linzen, and Ryan Cotterell, editors, **Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning**, pp. 1–34, Singapore, December 2023. Association for Computational Linguistics.

[2] 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.

[3] Aaron Mueller and Tal Linzen. How to plant trees in language models: Data and architectural effects on the emergence of syntactic inductive biases. In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, **Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)**, pp. 11237–11252, Toronto, Canada, July 2023. Association for Computational Linguistics.

[4] Alex Warstadt and Samuel R. Bowman. Can neural networks acquire a structural bias from raw linguistic data?, 2020.

[5] Aaron Mueller, Robert Frank, Tal Linzen, Luheng Wang, and Sebastian Schuster. Coloring the blank slate: Pre-training imparts a hierarchical inductive bias to sequence-to-sequence models. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, **Findings of the Association for Computational Linguistics: ACL 2022**, pp. 1352–1368, Dublin, Ireland, May 2022. Association for Computational Linguistics.

[6] Philip A. Huebner, Elior Sulem, Fisher Cynthia, and Dan Roth. BabyBERTa: Learning more grammar with small-scale child-directed language. In Arianna Bisazza and Omri Abend, editors, **Proceedings of the 25th Conference on Computational Natural Language Learning**, pp. 624–646, Online, November 2021. Association for Computational Linguistics.

[7] Lisa S. Pearl and Benjamin Mis. The role of indirect positive evidence in syntactic acquisition: A look at anaphoric ”one”. **Language**, Vol. 92, No. 1, pp. 1–30, 2016.

[8] Florencia Reali and Morten H. Christiansen. Uncovering the richness of the stimulus: Structure dependence and indirect statistical evidence. **Cognitive Science**, Vol. 29, No. 6, pp. 1007–1028, 2005.

[9] Pang Wei Koh and Percy Liang. Understanding black-box predictions via influence functions. In Doina Precup and Yee Whye Teh, editors, **Proceedings of the 34th International Conference on Machine Learning**, Vol. 70 of **Proceedings of Machine Learning Research**, pp. 1885–1894. PMLR, 06–11 Aug 2017.

[10] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In Tal Linzen, Grzegorz Chrupa la, and Afra Alishahi, editors, **Proceedings of the 2018 EMNLP Workshop BlackboxNLP: Analyzing and Interpreting Neural Networks for NLP**, pp. 353–355, Brussels, Belgium, November 2018. Association for Computational Linguistics.

[11] Alex Warstadt, Alicia Parrish, Haokun Liu, Anhad Mohananey, Wei Peng, Sheng-Fu Wang, and Samuel R. Bowman. BLiMP: The benchmark of linguistic minimal pairs for English. **Transactions of the Association for Computational Linguistics**, Vol. 8, pp. 377–392, 2020.

[12] Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. Masked language model scoring. In **Proceedings of ACL**, pp. 2699–2712, Online, July 2020. Association for Computational Linguistics.

[13] Han Guo, Nazneen Rajani, Peter Hase, Mohit Bansal, and Caiming Xiong. FastIF: Scalable influence functions for efficient model interpretation and debugging. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, **Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing**, pp. 10333–10350, Online and Punta Cana, Dominican Republic, November 2021. Association for Computational Linguistics.

[14] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized bert pretraining approach, 2019.

[15] Noam Chomsky. Rules and representations. **Behavioral and Brain Sciences**, Vol. 3, No. 1, p. 1–15, 1980.

[16] Jane Grimshaw and Steven Pinker. Positive and negative evidence in language acquistion. **Behavioral and Brain Sciences**, Vol. 12, No. 2, p. 341–342, 1989.

[17] Hiroshi Noji and Hiroya Takamura. An analysis of the utility of explicit negative examples to improve the syntactic abilities of neural language models. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel Tetreault, editors, **Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics**, pp. 3375–3385, Online, July 2020. Association for Computational Linguistics.

[18] Barbara C. Lust. **Child Language: Acquisition and Growth**. Cambridge Textbooks in Linguistics. Cambridge University Press, 2006.