言語モデルの事前学習におけるバリエーションセットの効果

芳賀あかり [1]　深津聡世 [2]　大羽未悠 [1]　Arianna Bisazza[3]　大関洋平 [2]
[1] 奈良先端科学技術大学院大学　[2] 東京大学　[3]University of Groningen
{haga.akari.ha0,oba.miyu.ol2}@is.naist.jp
{akiyofukatsu,oseki}@g.ecc.u-tokyo.ac.jp
{a.bisazza}@rug.nl

概要

近年，子供向けの発話（CDS）が言語モデル（LM）の学習効率を向上させる可能性が示唆されている．しかし，CDS のどの特性が LM の学習に有効であるかは明らかになっていない．本研究では，CDS の一つの特徴であるバリエーションセット（VS）に着目し，VS が LM の学習に与える影響を調査する．具体的には，人工的に作成した VS を含む学習データを用いて GPT-2 の事前学習を行った．その結果，文法の学習と自然言語理解において，学習データ中の VS の存在がモデルの性能向上に貢献する可能性を示した．これらの結果から，VS が LM の学習効率向上に有益であることが示唆される一方，効果の詳細を解明するためのさらなる調査が必要である．

1　はじめに

現在の言語モデル（LM）は，多くのタスクにおいて高い性能を示しているが，学習効率の面では改善の余地が大きい．例えば，現在広く使用されている LM の一つである Llama 2 は，事前学習に約 2 兆語のデータを使用する一方で，人間は 13 歳までに 1 億語未満の言語入力で母語を習得するとされている [1]．また，現在の LM が学習するような大量かつ多様なインプットは与えられないのにも関わらず，子供は 6 歳頃までに基本的な文法を習得する [2, 3]．このことから，子供の言語習得の過程は，LM の学習効率を向上させる上で重要な示唆を与える可能性がある．さらに，第一言語獲得の分野では，子供の効率的な学習に，子供の認知能力の制限や子供向けの発話（Child-Directed Speech; CDS）が寄与することが示されている [4, 5, 6, 7, 8]．特に，CDS は，LM の学習においても言語知識の獲得を促進する可能性が示唆されている．例えば，先行研究では，文法知識の学習効率の向上 [9]，意味理解への貢献 [10]，さらに階層構造の学習の促進 [11] が示されている．これらの背景から，CDS が小規模な学習データを用いた学習を助けることは明らかになりつつある．しかし，CDS のどの特性が LM の学習に貢献するかは，未解明の部分が多い．

第一言語および第二言語獲得に関する研究では，このような CDS の特徴の一つとして，バリエーションセット（Variation Set; VS）が注目されている．VS は，連続した発話において，わずかに異なる語彙や構文を用いて類似の意図が表現される発話である [12]．このような発話は CDS に広く見られるが，他の発話にはほとんど存在しない．先行研究では，VS が子供の注意をあるトピックに集中させ，新しい情報を導入しながら理解を促進することで，統語構造の学習を支援することが示されている [13, 14, 15, 16]．したがって，これらの知見は，VS が言語学習全般に有効であり，LM における学習効率の向上にも寄与する可能性を示唆している．

以上の背景から，本研究では，VS が LM の学習効率に与える影響を検証する．具体的には，Küntay らの定義 [12] に基づいて人工的に VS を生成し，異なる割合で非 VS と混合した学習データを作成する．その後，作成した学習データを用いて，2 つのベンチマーク BLiMP [17]，GLUE [18] におけるモデルの性能を比較する．

2　関連研究

子どもの言語発達に関する研究では，CDS が，子どもの言語習得において重要な役割を果たしていることが示唆されている．例えば，Fernald は，生後 4 か月の乳児 48 名の CDS の選好を観察し，乳児が大人向けの発話 (Adult-Directed Speech; ADS) よりも CDS を好むことを示した [5]．また，Jusczyk は，乳児が ADS よりも CDS を聞いた場合の方が，発話を区切る能力が向上することを報告している [6]．さ

らに，Rowe は，50 組の親子を対象とした長期にわたる観察を通じて，親が高度な語彙や文脈から切り離された会話をすることで，子どもの語彙発達が促進されることを実証した [7]. [1)]

自然言語処理分野の研究では，CDS が人間の言語習得だけでなく，モデルの文法知識の獲得にも有益であるかをさらに検証している．Huebner らは，5 百万語の CDS を用いて学習した小規模な RoBERTa [21] モデル (BabyBERTa) が，30 億語で学習した RoBERTa と同等の言語能力を達成できることを示した [9]．また，You らは，CDS が統語構造を持たない場合でも因果関係を把握するための豊かな意味情報を有しており，意味情報の抽出を学習する上で効果的であることを明らかにした [10]．さらに，Mueller らは，CDS を用いて学習した LM が，Wikipedia のような典型的なデータセットで学習した場合よりも階層的な構造をより理解できると主張している [11]．これらの研究結果は，CDS が言語学習に及ぼす効果を示しているが，本研究では，この効果に寄与する CDS の具体的な特性に着目する．

第一言語および第二言語習得の研究において，VS は，言語発達の成功における重要な要因として注目を集めている [12]．Küntay らは，VS の特徴を以下のように説明している．連続する発話において，1) 同じ内容が繰り返されるか言い換えられる，2) 意図が一貫している，3) 語の置き換え，フレーズの追加や削除，フレーズの並び替えなどが行われる [12]．また，Wiré らは，英語における典型的な VS の例を以下のように示している [22].

(1) You can put the animals there.
You can take the pig and the cat and put them there.
Can you put them there?
Good.
Can you put the pig there too?

さらに，既存研究は VS が実際に人間の言語学習を促進することを示唆している．例えば，Hoff-Ginsberg は，同一の発話を繰り返すことが子どもの統語発達を促進し，わずかに変化を加えた連続する発話が文構造に関する手がかりを提供し，統語発達を助けると主張した [13]．また，Brodsky らは，VS のような部分的な反復を含む発話が学習者にとって情報密度が高くなる可能性を示している [14]．さらに，Onnis らは成人に人工言語を教える実験を通じて，VS が文を解析する助けとなることを示した [15]．これにより，連続する類似の文を比較することが統語構造を学習する手がかりとなることが示唆された．これらの背景から，文脈的に一貫した発話でありながら表現にわずかな違いがある場合，その構造的な違いがより顕著になり，LM における統語構造の予測向上が期待できる．

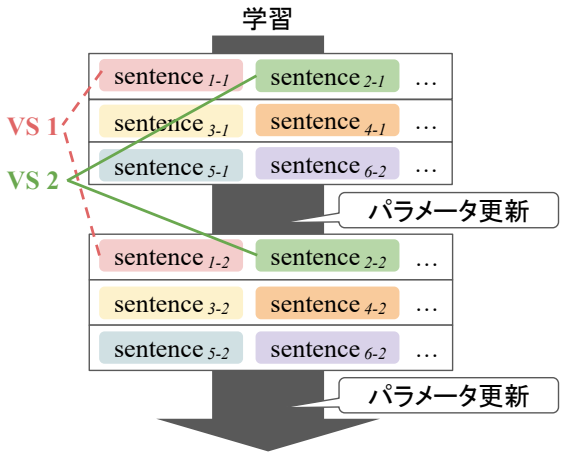

図 1 モデルの学習時に VS を与えるための学習データの作成手法．図は batch size=3 とした例を示す．VS 内の各文を隣接するバッチに配置することで，モデルは VS 内の各文を観測した後にパラメータを更新し，同じ VS 内の次の文を観測する．図中の sentence *i*-*j* は，*i* 番目の VS の *j* 番目の文を示す．

3 実験方法

人間の言語獲得の研究から着想を得て，本研究では，VS が LM においても言語習得の助けとなるかを検証する．我々の知る限り，この効果を検証した研究は片野らの予備実験 [23] に限られている．片野らの実験では，CDS データから自然に発生する VS を複数の自動検出手法を用いて抽出し，学習データ中の VS の量を変化させることによって LM の学習への VS の影響を調査している．しかし，片野らの実験では、VS の影響が有意に示されなかった．これは，高精度な VS の自動抽出手法が未確立であり，VS の正確な数を制御するのが難しいためと考えられる [16]．この課題に対処するため，本研究では人工的に VS を作成し，学習データ内の VS の割合を完全に制御する．具体的には，`gpt4o-mini`[2)] を用いて VS を生成し，異なる割合で学習データに含める．[3)]

1) ただし，一部の文化圏では，CDS がほとんど使用されないことがある [19, 20].

2) https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/

3) 学習データ内の VS の割合を完全に制御するため，人工 VS 以外の文は全てシャッフルし，自然発生した VS が含まれないようにした．

子供の単語学習における次単語予測の効果を視線計測を用いて調査した研究 [24] によれば，訓練時に文を聞きながら指示対象を予測し，その予測を修正する視線パターンを示した子供は，テストで高い成績を示す傾向があることが明らかになった．この結果は，人間の言語習得において，次単語予測を行った後にその予測を修正することが学習に寄与する可能性を示唆している．本研究では，人間の予測を修正する行為が LM のパラメータ更新に相当すると仮定し，LM が VS 内の各文を読むたびにパラメータを更新するよう学習データを構築する．具体的には，図 1 に示すように，VS 内の各文を隣接するバッチに配置する．これにより，モデルは VS 内の 1 文を入力してパラメータ更新を行った後，次に同じ VS 内の次の文を入力する．

3.1 モデル

子どもは文を処理する際，統語情報と意味情報を踏まえて次に来る語句のカテゴリを予測しているとされている [25]．また，最近の研究では，予測的な行動が子どもの言語習得を助けることが示唆されている [26]．これらの知見は，文の続きの予測が文構造の学習に貢献することを示している．そのため，本研究では，事前学習で次単語予測行う LM である GPT-2 [27] を使用する．ハイパーパラメータの詳細は付録 C に示す．

3.2 学習データセットの構築

まず，学習データを作成するために，子供と親の会話を記録したデータセットである CHILDES [28] から英語の CDS を 1,000 万語抽出した．次に，`gpt4o-mini` を使用して人工 VS を作成した．モデルには，CDS からサンプリングした 1 文を与え，これを言い換えることによって VS を作成するよう指示した．また，VS を作成する際には Küntay らの VS の特徴についての記述および典型的な VS の例 [12] を参考にするよう指示した（プロンプトは付録 A を参照）．生成された VS の約 48%は疑問文となった．実際に生成された VS の例を付録 B に示す．

実際の CDS には 10%から 50%の VS が含まれるため [29, 14, 15]，VS と非 VS を混合して学習データを作成する．非 VS には，シャッフルすることにより自然発生の VS が含まれないようにした CDS を用いた．

3.3 評価

モデルの評価には，BLiMP [17] および GLUE [18] を採用する．BLiMP は，モデルの文法知識を評価するための二値分類タスクを提供するベンチマークであり，GLUE は，自然言語理解を評価するためのベンチマークである．GLUE はファインチューニングを行った後に評価を行う．本研究では，これらの指標の評価に BabyLM [30] が提供するパイプライン [30, 31] を使用する．

また，VS をどの比率で学習データに含めるべきかを調査するため，VS の比率を変化させた際のモデルの性能の比較を行う．さらに，同一文の言い換えがデータ内に存在することの効果と，それらが連続して提示されることの効果を分離するため，学習データをシャッフルした結果との比較も行う．

4 実験結果

本実験では，人工 VS を含む学習データを用いて，GPT-2 をスクラッチから学習した．以下では 3 epoch 終了時点（モデルの学習収束後）の結果を報告する．

図 2 に，学習データ中の VS の割合を変化させた際のモデルのパフォーマンスを示す．

まず，VS の割合の影響に着目すると，シャッフルなし条件において，すべての指標で VS を学習データに含めた場合にスコアが向上した．VS の最適な割合はベンチマークによって異なり，BLiMP スコアは VS の割合が増加するにつれて向上した．ただし，付録 D に示すように，人工的な VS が実際の CDS に含まれる VS と比較してノイズが少ないことが結果に影響した可能性が残る．一方で，BLiMP Supplement は VS の割合が 20%のとき，GLUE は VS の割合が 40%のときに最高スコアを達成しており，これは実際の CDS における VS の割合に近い．

シャッフルあり条件下では，BLiMP および BLiMP Supplement のスコアが，学習データに VS を含めた場合に向上した．特に，BLiMP Supplement のスコアは，シャッフルあり条件下で VS の割合が実際の CDS に近い 40%に近づくにつれて向上した．次に，シャッフルなし条件とシャッフルあり条件を比較すると，各ベンチマークにとって最適な VS の割合の場合，シャッフルなし条件がシャッフルあり条件を上回った．

以上の結果より，学習データに VS を含めた場合，シャッフルなし条件は全ての指標でシャッフルあり

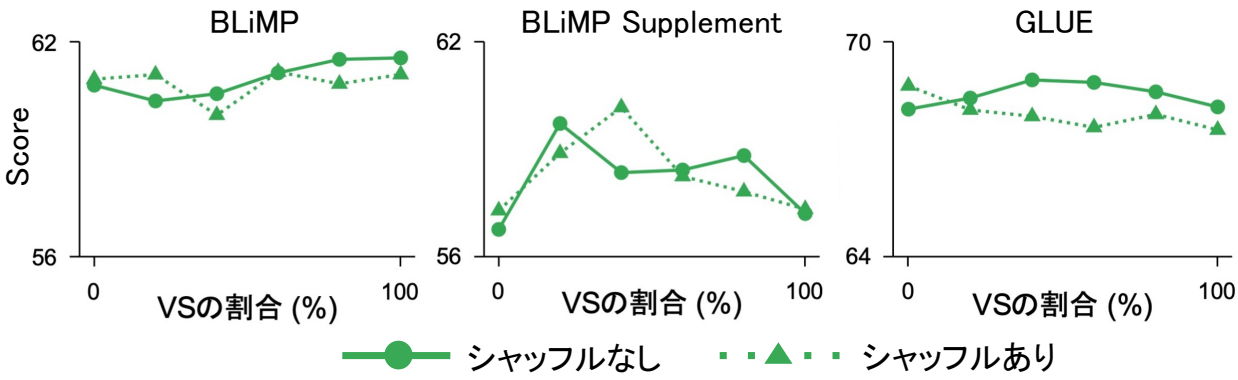

図 2　各学習条件におけるモデルの性能

条件より高いスコアを示し，連続した言い換えが学習に貢献する可能性が示された．しかし，一部の実験設定ではシャッフルあり条件がシャッフルなし条件を上回った．

4.1　考察

本研究の結果は，VS 内の文を隣接したバッチに配置したとき，連続した言い換えの存在が多くの場合でモデルの学習に有益であることを示している．しかし，直感に反して，CDS のように言い換えを連続して提示するよりも，シャッフルした順序で提示する方が良い結果をもたらす場合があることが明らかになった．また，学習データ内の VS の最適な割合は実験条件や評価基準によって異なり，最適な値を見つけることはできなかった．これは，現在の実験設計において，学習データ内の VS の増加が語彙の多様性の減少を引き起こしていることが原因である可能性がある．

5　おわりに

本研究では，CDS に含まれる VS が LM の学習に与える影響について調査を行った．人間を対象とした既存研究では，VS が文構造の理解に効果をもたらすとされていたのに対し，本研究では，LM の学習において，VS が文構造の理解に加え自然言語理解にも有益な影響を与える可能性が示唆された．CDS を用いた学習により学習効率向上を達成している BabyBERTa [9] などの小規模な LM も，CDS 中の VS の恩恵を受けているかもしれない．今後の研究では，学習データ内の人工 VS の増加と語彙の多様性の減少のトレードオフを考慮した，より詳細な分析を行う予定である．

謝辞

本研究は JSPS 科研費 24H00087，JST PRESTO JPMJPR21C2, JST ACT-X JPMJAX24CM, および Dutch Research Council (NWO) within the Talent Programme (VI.Vidi.221C.009) 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, et al. Findings of the babylm challenge: Sample-efficient pretraining on developmentally plausible corpora. In **Proceedings of the BabyLM Challenge at the 27th Conference on Computational Natural Language Learning**, 2023.

[2] Rhea Paul. **Analyzing Complex Sentence Development**, pp. 36–71. University Park Press, 1981.

[3] Nenagh Kemp, Elena Lieven, and Michael Tomasello. Young children's knowledge of the "determiner" and "adjective" categories. **J Speech Lang Hear Res**, Vol. 48, No. 3, pp. 592–609, June 2005.

[4] Elissa L. Newport. Maturational constraints on language learning. **Cognitive Science**, Vol. 14, No. 1, pp. 11–28, 1990.

[5] Anne Fernald. Four-month-old infants prefer to listen to motherese. **Infant Behavior and Development**, Vol. 8, No. 2, pp. 181–195, 1985.

[6] Peter W. Jusczyk. **The discovery of spoken language**. MIT Press, 1997.

[7] Meredith L. Rowe. A longitudinal investigation of the role of quantity and quality of child-directed speech in vocabulary development. **Child Development**, Vol. 83, No. 5, pp. 1762–1774, 2012.

[8] Vera Kempe, Mitsuhiko Ota, and Sonja Schaeffler. Does child-directed speech facilitate language development in all domains? a study space analysis of the existing evidence. **Developmental Review**, Vol. 72, p. 101121, 2024.

[9] 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.

[10] Guanghao You, Balthasar Bickel, Moritz M. Daum, and Sabine Stoll. Child-directed speech is optimized for syntax-free semantic inference. **Scientific Reports**, Vol. 11, No. 1, p. 16527, Aug 2021.

[11] 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.

[12] Aylin C. Küntay and Dan I. Slobin. **Listening to a Turkish mother: Some puzzles for acquisition**, chapter 17. Lawrence Erlbaum, 1996.

[13] Erika Hoff-Ginsberg. Function and structure in maternal speech: Their relation to the child' s development of syntax. **Developmental Pyschology**, Vol. 22, No. 2, pp. 155–163, 1986.

[14] Peter Brodsky and Heidi Waterfall. Characterizing motherese: On the computational structure of child-directed language. In **Proceedings of the Annual Meeting of the Cognitive Science Society**, Nashville, Tennessee, 2007. Cognitive Sciency Society.

[15] Luca Onnis, Heidi R. Waterfall, and Shimon Edelman. Learn locally, act globally: Learning language from variation set cues. **Cognition**, Vol. 109, No. 3, pp. 423–430, 2008.

[16] Nicholas A. Lester, Steven Moran, Aylin C. Küntay, Shanley E.M. Allen, Barbara Pfeiler, and Sabine Stoll. Detecting structured repetition in child-surrounding speech: Evidence from maximally diverse languages. **Cognition**, Vol. 221, p. 104986, 2022.

[17] 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.

[18] 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 **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.

[19] Alejandrina Cristia, Emmanuel Dupoux, Michael Gurven, and Jonathan Stieglitz. Child-directed speech is infrequent in a forager-farmer population: A time allocation study. **Child development**, Vol. 90, No. 3, pp. 759–773, 2019.

[20] Ann Weber, Anne Fernald, and Yatma Diop. When cultural norms discourage talking to babies: Effectiveness of a parenting program in rural senegal. **Child Development**, Vol. 88, No. 5, pp. 1513–1526, 2017.

[21] 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. **ArXiv**, Vol. abs/1907.11692, , 2019.

[22] Mats Wirén, Kristina Nilsson Björkenstam, Gintarė Grigonytė, and Elisabet Eir Cortes. Longitudinal studies of variation sets in child-directed speech. In Anna Korhonen, Alessandro Lenci, Brian Murphy, Thierry Poibeau, and Aline Villavicencio, editors, **Proceedings of the 7th Workshop on Cognitive Aspects of Computational Language Learning**, pp. 44–52, Berlin, August 2016. Association for Computational Linguistics.

[23] Harufumi Katano. Syntax acquisition in language models with small datasets: Data efficiency using variation sets [in Japanese], 2024. Bachelor's thesis.

[24] Tracy Reuter, Arielle Borovsky, and Casey Lew-Williams. Predict and redirect: Prediction errors support children' s word learning. **Developmental Psychology**, Vol. 55, No. 8, p. 1656, 2019.

[25] Arielle Borovsky, Jeffrey L Elman, and Anne Fernald. Knowing a lot for one's age: Vocabulary skill and not age is associated with anticipatory incremental sentence interpretation in children and adults. **J Exp Child Psychol**, Vol. 112, No. 4, pp. 417–436, May 2012.

[26] Tracy Reuter, Arielle Borovsky, and Casey Lew-Williams. Predict and redirect: Prediction errors support children's word learning. **Dev Psychol**, Vol. 55, No. 8, pp. 1656–1665, May 2019.

[27] Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. Improving language understanding by generative pre-training, 2018.

[28] Brian MacWhinney. **The CHILDES Project: Tools for analyzing talk. Third Edition**. Lawrence Erlbaum Associates, 2000.

[29] Heidi R. Waterfall. **A little change is a good thing: Feature theory, language acquisition and variation sets**. PhD thesis, University of Chicago, 2006.

[30] Leshem Choshen, Ryan Cotterell, Michael Y. Hu, Tal Linzen, Aaron Mueller, Candace Ross, Alex Warstadt, Ethan Wilcox, Adina Williams, and Chengxu Zhuang. [call for papers] the 2nd BabyLM Challenge: Sample-efficient pretraining on a developmentally plausible corpus. **Computing Research Repository**, Vol. arXiv:2404.06214, , 2024.

[31] Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. A framework for few-shot language model evaluation, 12 2023.

[32] Thea Cameron-Faulkner, Elena Lieven, and Michael Tomasello. A construction based analysis of child directed speech. **Cognitive Science**, Vol. 27, No. 6, pp. 843–873, 2003.

A　VS 生成のためのプロンプト

VS の生成には以下のプロンプトを用いた．

Rephrase a given sentence based on the characteristics of variation sets. A variation set is a set of utterances that have characteristics as follows:
In successive utterances,
- the same content is repeated or rephrased.
- there is a consistent intent.
- there are operations such as word substitution, addition/deletion of phrases, and reordering of phrases.
Here is an example:
You can put the animals there.
You can take the pig and the cat and put them there.
Can you put them there?
Good.
Can you put the pig there too?
Please use only the vocabulary that 10 year-old children understand.

B　生成された VS の例

`gpt4o-mini` によって生成された VS の例を表 1 に示す．

表 1　`gpt4o-mini` によって生成された VS の例

元の CDS	生成された VS
What do you want?	What do you need? What do you want to have? Can you tell me what you want? What is it that you want? What do you feel like getting?
What did Laura do last night?	What did Laura do yesterday evening? What was Laura doing last night? Can you tell me what Laura did last night? What activity did Laura have last night? What was Laura up to last night?
Hold on for a moment.	Please wait for a little bit. Just give me a second. Can you wait for a minute? Wait here for a moment, please. Just hang on for a minute.
We'll tidy it up.	Let's clean it together. Can we clean it? Great idea! Should we clean it now?
Okay, let's help you down.	Let's get you down now. Can we help you down? Alright, let's get you down safely. Can I help you get down?
How are you doing?	How have you been? How's it going? How are things with you? How do you feel today?

C　実験設定の詳細

モデルの学習時に用いたハイパーパラメータを表 2 に示す．

表 2　モデルの学習に用いたハイパーパラメータ

Model	architecture	GPT-2
	parameters	124M
	vocab size	50,257
	hidden size	768
	heads	12
	layers	12
	dropout	0.1
	layer norm eps	1e-05
	initializer range	0.02
Optimizer	algorithm	AdamW
	learning rates	5e-05
	betas	(0.9, 0.999)
	weight decay	0.0
Scheduler	type	linear
Training	gradient accumulation	1
	epoch	3
	batch size	64
	line by line	true
	NGPU	1

D　実際の CDS と人工 VS の差異

実際の CDS には，以下のような不完全な文を用いた発話が多く含まれている．

(2) To who?
You don't.
To you or to Laura?
To me.
Oh how come?

このような不完全な文からなる発話は CDS のおよそ 30% を占める [32] が，これに対して，人工的な VS には以下のように完全な文が多く含まれている．

(3) It's a blanket that we all share.
We all have a blanket together.
This blanket belongs to everyone.
It's a blanket for all of us to use.
Everyone can use this blanket.

したがって，学習データ中の人工 VS を増加させたとき，学習データ中のノイズが減少し，実験結果に影響する可能性がある．