ニューラル言語モデルによるかき混ぜ構文のモデリング

竹本健悟 * 武田更紗 * 井上景太 * 大関洋平
東京大学
{takemoto-kengo2262, sarasa-t, inoue-keita819, oseki}@g.ecc.u-tokyo.ac.jp

概要

フィラー・ギャップ依存は，wh 句とその解釈位置との間に成立する長距離依存関係である．近年，言語モデルがフィラー・ギャップ依存に対応してどのような予測を形成するのかをサプライザルを用いて評価する研究が進められているが，その多くは英語の wh 移動を対象としている．本研究では，複数の言語モデルについて日本語のかき混ぜ構文を対象にサプライザルの分析を行う．実験の結果，言語モデルは部分的にかき混ぜ構文におけるフィラー・ギャップ依存の交互作用を一般化し，人間の処理傾向とも整合的な予測傾向を示した．

1 はじめに

フィラーギャップ依存（Filler-Gap Dependency; FGD) とは，wh 句などの先行要素（filler）とそれが解釈されるべき表層的に現れない引数位置（gap）との間に成立する長距離依存関係のことである．FGD は，長距離にわたり成立しうる一方で，その成立可否が統語構造に依存して体系的に変化する [1, 2, 3]. filler と gap の関係が単純な線形順序では規定しきれないという事実は，言語の構造が階層的であることの裏付けの一つとされてきた．

近年，この FGD をニューラルネットワークを基盤とする言語モデル（Language Model; LM）において定量的に評価する研究が注目されている [4, 5, 6]. Wilcox らは，英語の wh 移動文において filler と gap の有無を独立に操作した 4 条件について，自己回帰型 LM における次単語予測にもとづいて与える各語の確率（サプライザル）を比較することで， LM が FGD の交互作用を一般化できることを示した [4, 5]. これらの結果は，言語学における階層的な移動現象の学習可能性をめぐる議論に定量的証拠を与えるという点で重要である．

しかし，FGD のモデリングに関する既存研究の多くは英語の wh 移動を中心に設計されており，FGD の現れ方が異なる言語や構文を横断的に扱う試みは限定的である [4, 5, 6].

そこで本研究では，語順が変化する言語現象である日本語のかき混ぜ（scrambling）構文に焦点を当てる．かき混ぜは基底構造での位置から要素を抜き出すことで FGD を形成する．Aoshima らは wh 認可子（wh-licensor）を伴う節を跨いだかき混ぜによる移動について自己ペース読み（Self-paced Reading; SPR) 課題を行うことで，人間のかき混ぜ構文の処理傾向を明らかにした [7, 8].

本研究では，FGD を伴うかき混ぜ構文における LM のサプライザルを，1) FGD の交互作用を正しく捉えることができているか，2) SPR 課題によって示された人間の処理傾向と整合するか，という 2 つの観点から分析を行う．

2 関連研究

2.1 LM における FGD 処理の分析 [4]

まずは，サプライザルによって言語モデルが FGD を一般化できているかを確かめた試みとして，Wilcox らによる [4] の研究を取り上げる．この研究では，サプライザルを自己回帰型 LM が文を順に処理していく中での各トークンにおける次単語確率から求める．そして，これを心理言語学におけるオンライン処理負荷のに対応する指標として用い，LM が FGD を一般化しているかを検証した．

サプライザルの計算は，gap の有無と wh 認可子の有無の交絡による次の (1) ような例文のパラダイムを通じて行われる (例文は [4], p.212-213, 一部改変)．

(1)

a. **no wh-licensor, no-gap**
I know that the lion devoured a gazelle at sunrise.

b. **wh-licensor, no-gap**
*I know what the lion devoured a gazelle at sunrise.

* 共同第一著者

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

表 1: Aoshima らによる日本語のかき混ぜと補文標識の交絡の例（[8]，p.30, 一部改変.）

	r_1	r_2	r_3	r_4	r_5	r_6	r_7	r_8
a. Scrambled, Decl.Comp.	どの生徒に	担任は	校長が	本を	読んだと	図書室で	司書に	言いましたか
b. In-situ, Decl.Comp.	*担任は	校長が	どの生徒に	本を	読んだと	図書室で	司書に	言いましたか
c. Scrambled, Q-particle	*どの生徒に	担任は	校長が	本を	読んだか	図書室で	司書に	言いました
d. In-situ, Q-particle	担任は	校長が	どの生徒に	本を	読んだか	図書室で	司書に	言いました

c. **no wh-licensor, gap**
*I know that the lion devoured __ at sunrise.
d. **wh-licensor, gap**
I know what the lion devoured __ at sunrise.

Wilcox らは，filler の解釈が決定する位置である gap の直後に続く語（at）におけるサプライザルに注目し，上記の 4 条件におけるその値から wh 認可相互作用（wh-licensing interaction; WLI）を以下の様に定義した．ここで，S_{cond} は条件 *cond* におけるサプライザルの値を表す．

$$\text{WLI} = \left(S_{\text{wh,no-gap}} - S_{\text{no-wh,no-gap}}\right) - \left(S_{\text{wh,gap}} - S_{\text{no-wh,gap}}\right)$$

WLI が正であれば，wh 認可子がある場合に gap がない条件でサプライザルが増え，gap を含む条件では wh 認可子の存在がサプライザルを相対的に下げるという FGD に特徴的なパターンを LM が一般化していることを意味する．

結果として，少なくとも一部の条件で WLI が有意に正となり，LM が FGD に対応する予測を形成していることを Wilcox らは報告している．

ただ，FGD は英語の wh 疑問文でのもののみならず，長距離の依存関係一般のことを指す．そのため，ここまで概観した [4] の試みは，英語以外の多様な FGD にも拡大して検討される余地がある．

2.2 かき混ぜ構文における FGD

日本語に目を向けてみると，語順入れ替えの分析において仮定されるかき混ぜの現象が好例である．かき混ぜの操作を経て派生された文には，抜き出される要素（filler）と，その基底構造での位置（gap）の間に長距離依存関係があるもの考えることができる[1)]．ただ，英語と異なる点として wh 認可子の位置があり，これは Aoshima らによる SPR 課題を通じた研究 [7], [8] によって報告されている．

彼女らはかき混ぜにおける FGD と，日本語の疑問標識との交絡を利用し，表 1 のような刺激のパラダイムを用いて実験を行っている．r とは，SPR 課題で読み手に一度に表示される領域（region）を r で示す．日本語の「か」は補文内の動詞にも主文の動詞にも付加されうるが，それによって間接疑問かそうでないかのスコープが決まる．FGD は同じスコープ内で規定されるので，主文疑問であるにもかかわらず wh 要素が補文内に止まっていると解釈ができず非文となる（表 1b）．反対に，間接疑問文で wh 要素が補文の外へ抜き出されている場合も，非文となる（表 1c).

この実験で報告されているところによれば，表 1 a，b，d のような条件では補文標識を含む領域（r_5）で読み時間の増大が観察された一方，c に関しては，その前の領域から読み時間の有意な上昇が確認されなかった．このことから，Aoshima らは前置された wh 句が本来解釈されるべき主文ではなく，補文の形式を処理する際に優先的に解釈されるものと結論づけている．

さらに，r_7 から r_8 にかけては全ての傾向において読み時間が長くなっており，これについては文末に到達したときに wh 要素との依存関係を計算しているからであるとの考察を行っている．これらの心理言語学的知見は，本研究でサプライザルによって評価する FGD のオンライン形成に対して行動指標に基づく比較基準を与える．

1) 本稿ではかき混ぜの理論的な分析の詳細は割愛する．代表的な論考としては，[9] や [10] を参照されたい．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

3 実験設定

本研究では，LM が日本語かき混ぜ構文における FGD に対応する予測を形成しているかをサプライザルに基づいて検証する．具体的には，[8] における SPR 課題で使用された 24 組の例文に対して，後述する複数のモデルでサプライザルを計測する．例文は全て表 1 と同一の構文パターンである．なお以下では，便宜上，表 1 の a, c のように r_8 で filler が解釈される条件を「かき混ぜ条件」，反対に，表 1 の b, d のように r_5 で解釈される条件を「元位置条件」と呼ぶ．

3.1 サプライザルの計算

自己回帰型 LM においては，各位置で次単語確率 $P(w_t \mid w_{<t})$ を与える．本研究では [4] と同様に，各トークン w_t のサプライザルを

$$S(w_t) = -\log P(w_t \mid w_{<t})$$

で定義する．

刺激文における領域 r_i がサブワード列 $\{w_t\}_{t \in r_i}$ に対応するとき，領域サプライザルを

$$S(r_i) = \sum_{t \in r_i} S(w_t)$$

で与える．

3.2 サプライザルの計算領域

サプライザルを計算する領域は，かき混ぜの有無によって異なる．これは，[4] と同様に，wh 句の解釈が定まる最初の位置における処理負荷を捉えることを目的としているためである．

本稿では，元位置条件とかき混ぜ条件それぞれにおける fi の解釈位置に合わせ，元位置条件では r_5 を，かき混ぜ条件では r_8 をサプライザルの比較対象とした．

もし LM がかき混ぜ構文における FGD の交互作用を一般化している場合，次の 2 点が成り立つことが予測される．(i) 元位置条件（In – situ）では，補文での wh 句の登場に対して補文内の動詞に平叙補文標識 *-to* ではなく疑問標識 *-ka* を期待するため，そのサプライザルの差 $S_{\text{In-situ,to}}(r_5) - S_{\text{In-situ,ka}}(r_5)$ は正の値を取る．反対に，(ii) かき混ぜ条件（Scr）は，$S_{\text{Scr,to}}(r_8) - S_{\text{Scr,ka}}(r_8)$ は負の値を取る．したがって，本研究では次のように WLI を定義する．

表 2: 本研究で用いる事前学習済み言語モデル

モデル	パラメータ数	学習トークン数
LSTM [11]	54.7 M	146 M
Transformer [11]	55.8 M	146 M
GPT2 [12]	337 M	-

$$\begin{aligned}\text{WLI} = &\left(S_{\text{In-situ,to}}(r_5) - S_{\text{In-situ,ka}}(r_5)\right) \\ &- \left(S_{\text{Scr,to}}(r_8) - S_{\text{Scr,ka}}(r_8)\right)\end{aligned}$$

この式に基づき，本研究では，WLI が正であれば，LM がかき混ぜ構文における FGD の交互作用を一般化しているものと考える．

3.3 言語モデル

実験に使用したモデルの詳細を表 2 に示す．使用したモデルはいずれも自己回帰型の事前学習済み LM であり，日本語コーパスで学習されている．また，本研究では領域内サプライザルを算出するため，形態素分割を先行させた後にサブワード分割を行うモデルのみを用い，領域境界を跨ぐ分割を可能な限り回避した．

4 結果・考察

4.1 FGD 交互作用の分析

図 1 に，各モデルにおける条件ごとのサプライザル差分と，それらから算出された WLI を示す．

WLI は LSTM で負，Transformer と GPT2 で正となった．このことは，LSTM が FGD を一般化していない一方で，Transformer および GPT2 が FGD に対応する予測傾向を少なくとも部分的に捉えていることを示唆する．この差は，LSTM では節を跨ぐ手がかりの保持・更新が難しいのに対し，自己注意型モデルは離れた位置の情報を参照しやすいというアーキテクチャの性質と整合的である．

Transformer と GPT2 において，元位置条件では

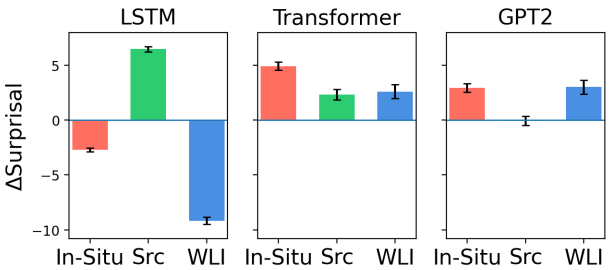

図 1: 各モデルにおけるサプライザル差分と WLI

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

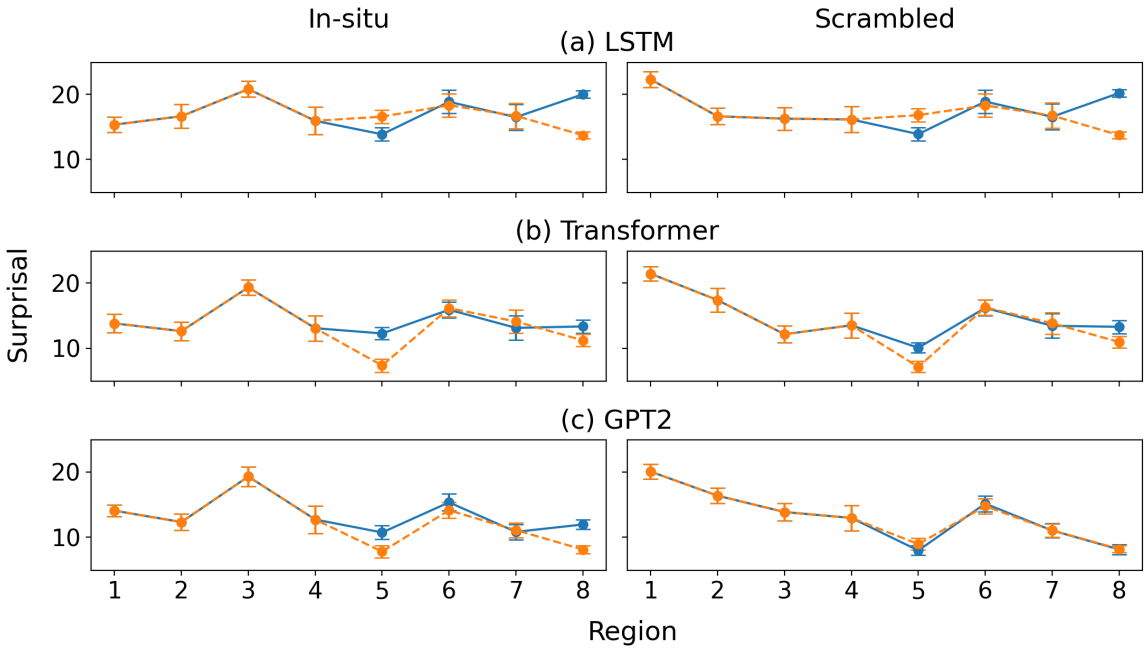

図 2: 各モデルにおけるサプライザルの推移（実線：平叙補文 *-to*，破線：間接疑問 *-ka*）

FGD に対応する予測傾向 ($S_{\text{In-situ,to}}(r_5) - S_{\text{In-situ,ka}}(r_5) > 0$）が見られたことから，同一節内の局所的な依存に対しては安定した予測を形成できることを意味する．

一方で，かき混ぜ条件においては 本来期待される符号反転（($S_{\text{Scr,to}}(r_8) - S_{\text{Scr,ka}}(r_8) < 0$）が明確には現れず，Transformer では正の傾向が残った．これは，かき混ぜ条件では依存関係の保持期間が長く，wh 句に由来する期待が干渉や減衰の影響を受けやすいため，スコープ確定の遅延や節を跨ぐ長距離依存を十分に反映できていない可能性を示す．

以上より，一部のモデルは FGD を一定程度捉えるものの，その多くは「wh 句の出現に伴う疑問標識への期待の形成」といった表層的な予測にとどまり，階層構造に基づく長距離依存への一般化は限定的である可能性が高い．

4.2 人間の処理傾向との対照

各モデルのサプライザルの推移を図 2 に示す．本節では，2.2 節で紹介した [8] における人間の処理に関する報告と対応する傾向が，LM のサプライザルにおいても観察されるかを検討する．

補文標識の現れる r_5 において，平叙補文標識 *-to* と疑問標識 *-ka* のサプライザルを比較すると，LSTM では r_5 において元位置・かき混ぜの両条件で wh の出現に伴う疑問標識への期待が形成されなかった．

一方で，Transformer では常に *-to* のサプライザルが *-ka* より高い．これは，先行する wh 要素により補文標識における *-ka* の出現を期待していることを示唆し，[8] における人間の処理傾向と整合している．

GPT2 でも元位置条件では同様に *-to* のサプライザルが *-ka* より高い一方，かき混ぜ条件では両者の間に差は見られなかった．これは，GPT2 は Transformer よりもかき混ぜ条件における長距離依存を捉えており，その結果として補文標識位置で局所的な *-ka* 期待に早期収束せず，*-to*/*-ka* の差が現れにくくなったと解釈できる．

5 おわりに

本研究では，日本語のかき混ぜ構文を対象に，LM が FGD に対応する予測を形成しているかを，サプライザルに基づいて検証した．実験の結果，LM は部分的に FGD の交互作用を一般化し，人間の処理傾向と整合的な予測傾向を示した一方で，階層的な長距離依存の追跡は限定的であることが明らかになった．今後の課題として，wh 句と解釈位置の距離を段階的に操作して依存距離が FGD 交互作用に与える影響を検証するとともに，wh 句が異なる領域に現れる構文にも対象を広げ，観測された予測傾向が構文横断的に一般化するかどうかを明らかにする必要がある．

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).

謝辞

本研究は，JSPS 科研費 JP24H00087，JST さきがけ JPMJPR21C2，JST CREST JPMJCR2565，JST BOOST JPMJBY24B2 の支援を受けたものです．

参考文献

[1] John Robert Ross. **Constraints on Variables in Syntax**. PhD dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1967.

[2] Noam Chomsky. On Wh-movement. In Peter W. Culicover, Thomas Wasow, and Adrian Akmajian, editors, **Formal Syntax**, pp. 71–132. Academic Press, New York, 1977.

[3] C.-T. James Huang. **Logical Relations in Chinese and the Theory of Grammar**. PhD dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1982.

[4] 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, 2018. Association for Computational Linguistics.

[5] Ethan Gotlieb Wilcox, Richard Futrell, and Roger Levy. Using computational models to test syntactic learnability. **Linguistic Inquiry**, Vol. 55, No. 4, pp. 805–848, 2024.

[6] Nur Lan, Emmanuel Chemla, and Roni Katzir. Large language models and the argument from the poverty of the stimulus. **Linguistic Inquiry**, pp. 1–28, 2024.

[7] Sachiko Aoshima, Colin Phillips, and Amy Weinberg. Processing of Japanese Wh-scrambling constructions. In William McClure, editor, **Japanese/Korean Linguistics 12**, pp. 179–191. CSLI Publications, Stanford, CA, 2003.

[8] Sachiko Aoshima, Colin Phillips, and Amy Weinberg. Processing filler-gap dependencies in a head-final language. **Journal of Memory and Language**, Vol. 51, No. 1, pp. 23–54, 2004.

[9] Mamoru Saito. **Some Asymmetries in Japanese and Their Theoretical Implications**. PhD dissertation, Massachusetts Institute of Technology, Cambridge, MA, 1985.

[10] Hajime Hoji. **Logical form constraints and configurational structures in Japanese**. PhD thesis, University of Washington, Seattle, 1985.

[11] Tatsuki Kuribayashi, Yohei Oseki, Takumi Ito, Yoshida Ryo, Masayuki Asahara, and Kentaro Inui. Lower perplexity is not always human-like. In **Proceedings of the Joint Conference of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing**, pp. 5203–5217, Online, August 2021. Association for Computational Linguistics.

[12] Okazaki Lab. japanese-gpt2-medium-unidic. Hugging Face model card. https://huggingface.co/okazaki-lab/japanese-gpt2-medium-unidic (accessed 2026-01-09). License: CC-BY-SA 4.0.

This work is published without peer review and is licensed by the author(s) under CC BY 4.0 (https://creativecommons.org/licenses/by/4.0/).