英語母語話者と生成 AI の文法性判断の差異調査

吉村理一 [1]　陳曦 [2]　伊藤薫 [1]　森部想水 [3]
[1] 九州大学 言語文化研究院 [2] 九州大学 人間環境学府
[3] 九州大学 芸術工学府
{chen.xi.831,moribe.sosui.695}@s.kyushu-u.ac.jp
{r-yoshimura,ito}@flc.kyushu-u.ac.jp

概要

本研究では，英語母語話者の文法性判断と生成 AI のその判断の差異を調査した. 英語の主要 56 構文（合計 4,483 例）を用い，AI の文法性判断と母語話者のそれを比較した結果, 母語話者が文法的と判断した例では約 82 %，非文法的文と判断した例では約 62 %の一致率を確認した．特に話題化，外置，寄生空所など移動を伴う構文において AI と母語話者の判断が乖離する傾向にあることが示された．本研究により，文法性判断に関して生成 AI の苦手分野を特定できたことから，AI の文法性判断能力向上を目指した適切な学習データの構築とベンチマークの開発が今後の研究課題として浮き彫りになった.

1　はじめに

近年の生成 AI の優れた性能に注目した研究が盛んに行われている. Han et al. [1], Toriida [2] に見られるように，英語教育においては, 4 技能 (読む, 書く, 聞く, 話す) に加え, それらの基礎となる文法や語彙力の強化を目的として, ChatGPT のような生成 AI を教師役として活用する取り組みが進んでいる. しかし, ChatGPT が示す模範解答と母語話者の判断が異なるケースが報告されており, その信頼性には不透明な部分がある. 例えば, (1) の例文は that-trace 効果に関する文法的な対比を示しているが, Haider [3] によれば ChatGPT はこれらの文法性の対比を正確に認識できない.

さらに筆者の予備調査では, (2a) に示す (非) 制限関係節の区別や (2b, c) に示す前置詞残留の文法性判断においても母語話者と ChatGPT の判断には溝があることを確認している.[1)]

学習者により正確な文法知識と適切な解説が産出できるようになることを将来的な目標に，本研究では英語母語話者の言語知識と生成 AI のメタ言語知識の根幹である文法性判断の差を測定し，報告する.

1) 2024 年 11 月 1 日参照.

表 1　that 痕跡効果の文法性判断比較

例文	母語話者	ChatGPT-4o
(1a) Who do you think that they invited?	ok	ok
(1b) Who do you think that invited them?	*	ok

[3]

表 2　(非) 制限関係性・前置詞残留の文法性比較

例文	母語話者	ChatGPT-4o
(2a) As a result of working at the newspaper company, I found my future plant, that was placed there.	*	ok
(2b) What aspect of her thesis did you talk to Mary in detail about? (Takami [4])	ok	*
(2c) Which park did you find the rabbit in? ([4])	ok	?*

2　先行研究

本稿と関連が深い先行研究として，Warstadt et al. [5] および [3] を概観する.

2.1　言語モデルの文法性判断力の比較

[5] は言語モデルが持つ英語に関しての文法性判断が人の持つ言語直観とどの程度一致しているかを調査している．彼らは 67 の個別のデータセット

それぞれに 1,000 の最小対の例文を含む BLiMP[2] と称するベンチマークを開発し，n-gram, LSTM, Transformer (GPT-2・Transformer-XL) LMs の精度を比較している．包含されている代表的な文法現象に，再帰代名詞，項構造，束縛，コントロール/繰上げ，決定詞と名詞句の一致，削除，空所との依存関係，動詞の不規則変化，島の効果，否定極性表現，数量詞，主語-動詞の一致の問題があり，これらの最小対の例文に対して文法性の判断を問う形式である．彼らの調査結果を図 1 に示す．

Model	Overall	Ana. agr	Arg. str	Binding	Ctrl. rais.	D-n agr	Ellipsis	Filler-gap	Irregular	Island	NPI	Quantifiers	S-v agr
5-gram	60.5	47.9	71.9	64.4	68.5	70.0	36.9	58.1	79.5	53.7	45.5	53.5	60.3
LSTM	68.9	91.7	73.2	73.5	67.0	85.4	67.6	72.5	89.1	42.9	51.7	64.5	80.1
TXL	68.7	94.1	69.5	74.7	71.5	83.0	77.2	64.9	78.2	45.8	55.2	69.3	76.0
GPT-2	80.1	99.6	78.3	80.1	80.5	93.3	86.6	79.0	84.1	63.1	78.9	71.3	89.0
Human	88.6	97.5	90.0	87.3	83.9	92.2	85.0	86.9	97.0	84.9	88.1	86.6	90.9

図 1　文法性判断の結果比較 [5]

[5] の比較実験 (図 1) から，GPT-2 の性能が一番良いことが分かり，再帰代名詞の認可や主語-動詞の一致などの形態論に関する現象では，GPT-2 は人間の文判断に近いことが示された．一方で，島効果，否定極性や数量詞の認可，項構造の違いの把握においては著しく低い精度であることも報告されている．

2.2　英語構文に対する ChatGPT の文法性判断

[3] は，英語における that 痕跡効果，wh 句の優位性効果，'why' と'how' の非 in-situ 特性，二重目的語構文における間接目的語 wh の移動，主語条件，主要部前置型の語彙句へ左側付加を行う言語は主要部後置型の句への左側付加が制限される Left-Left-Constraint，believe タイプと expect タイプの区別，副詞句や代名詞 strict/sloppy 読みなどの解釈曖昧性についての ChatGPT の文法性判断を報告している（表 3）．

[3] によれば，母語話者が非文法的とした例文を ChatGPT が誤って文法的と判断することが目立った例に that 痕跡効果，'why' と'how' の非 in-situ 特性，Left-Left-Constraint が挙げられている．それとは逆に，母語話者が文法的とした例文を誤って非文法的と判断した例に解釈曖昧性を示す例文群が挙げられている．この解釈曖昧性を示す例文群については，そもそもその曖昧性自体を検知することができなかったとも述べられている．とはいえ，この結果は英語母語話者の文法性判断の平均と比較しても遜色ないと [3] は結論づけている．

表 3　ChatGPT の応答まとめ [3]

	False positive	False negative
That-trace	partially	no
Wh-subject	no	no
'Why' or 'how'	yes	no
Wh-double-objects	no	no
Subject condition	no	no
Left-left constraint	partially	no
Believe vs. expect	no	no
Ambiguities	no	partially[14]

False positive:　Bot judges a sentence as grammatical that is in fact ungrammatical.
False negative:　Bot judges a sentence as ungrammatical that is in fact grammatical.

2.3　先行研究のまとめと本研究の試み

2.1 節の [5] および 2.2 節の [3] の研究を概観して分かるように，GPT 系のモデルが他と比較して高い精度を示している．両者とも生成 AI の精度の高さを謳っているが，その調査方法とカバーされている言語現象の種類には課題が残る．このことを鑑み，本研究では中島 [6] により編纂された英語の主要構文を参考に，その例文群と文法性判断を英語母語話者の判断として ChatGPT の応答と比較する．なお，調査過程の中で ChatGPT が意図しないパターン学習をし，その結果を反映させる可能性を抑制するため，最小対の例文群だけを問うことを控え，（56 種の言語現象に分類される）広範な種類の例文（合計 4,483 例）を取り入れることにした[3]．また，例文群の種類ごとにチャットを改めることも徹底した．詳細は 3 節に譲る．

3　検証実験

本研究では，構文横断的に最新の生成 AI と英語母語話者の文法性判断の違いを明らかにし，その背景にある間違いを引き起こす共通の原因の考察を目的とし検証実験を行った．以下に具体的な実験環境と実施方法を示す．

2）　Benchmark of Linguistic Minimal Pairs

3）　[3] によると，ChatGPT は，敢えて非文法的な例文のテストをしているにもかかわらず，質疑応答の過程でそれを誤植と判断し自ら修正した例文についての文法性判断を行うことがある．例えば，What does she think that cured her disease?という例文のテストを行なっているにも関わらず，What does she think cured her disease?に修正した形での判断を返答することがある．最小対のデータセットで実験が継続されていることやチャットログがリセットされないまま調査が継続されることなどが，その要因として考えられる．

3.1 実験環境

2.3 節でも触れたように，先行研究では網羅されていない言語現象も調査対象に含める意図から，[6] を参考に 56 種 (付録参照) の英語の主要構文を取り上げることにした．

調査は構文の種類ごとにチャットを設定して行ったが，調査過程で意図しない学習が行われ，その結果が反映されることがないように最小対のデータに限定しての実験は行わなかった．調査対象の例文は合計で 4,483 例となった．また，文法性判断を問う対象の言語モデルとしては，執筆時点で無料ユーザーにも開放されており，学習者が質問する際に最も良く用いられると考えられる GPT-4o のブラウザ版 ChatGPT を採用した．

3.2 実験手法

本研究では，言語モデルに与えるペルソナによる性能の変化も加味するために，通常の利用法と考えられる，文法性判断をするよう単純に指示する方法とは別に，教師役として判断をする場合を想定し，それぞれの場合のプロンプト設計を行った．本研究では，同じ言語現象について扱う場合でも教師役・通常の利用法の役割ごとにセッションを分け，生成 AI が与えられている役割が混同されないようにした．以下に教師役と通常の利用法のプロンプト例と，判断をスキップするなどの文法性判断が正常に行われない場合に処理を行わせるように誘導するプロンプト例を示す．基本的には，これらのプロンプトに続けて対象の文章を入力することで生成 AI による文法性判断が得られる．

表 4 教師役・一般的利用と出力調整用のプロンプト例

役割	プロンプト例
教師役	You are an English teacher. Are the following English sentences grammatically correct? Could you please explain why you think that is ungrammatical? Could you please explain why you thought that was unnatural?
一般的利用	Are the following English sentences grammatically correct? Could you please explain why you think that is ungrammatical? Could you please explain why you thought that was unnatural? Give your judgment and reason, sentence by sentence.
出力調整	Please do it in order and give me your judgment for all sentences.

4 結果

以下に，教師役のプロンプトを与えた場合と特に役割を与えずに文法性判断を行うよう指示するプロンプトを与えた場合に関して，再現率 (Recall) と適合率 (Precision) と F 値を求めた[4]．本研究では，文法的な文を文法的だと判断できる能力と，非文法的な文を非文法的であると判断する能力のそれぞれに関して全構文における判定精度を求めた．

表 5 正例（文法的な文）に対する評価結果

	再現率 (正)	適合率 (正)	F 値 (正)
教師役	0.818	0.838	0.828
一般的利用	0.812	0.836	0.824

表 6 負例（非文法的な文）に対する評価結果

	再現率 (負)	適合率 (負)	F 値 (負)
教師役	0.637	0.605	0.621
一般的利用	0.643	0.603	0.622

現状の生成 AI は，文法的な文を文法的だと判断するタスクでは人間の母語話者の判断と 8 割程度一致するようだが，非文法的な文を非文法的だと判断するタスクについては 6 割程度しか人間と一致しないことがわかった．また，上記指標の観点では教師の役割を与えた場合と特にそのような役割を与えず判断を問うプロンプトを与えた場合の間で特に顕著な差が見られなかった．以下に特に判断精度が悪かった構文種を示す[5]．

4.1 文法的な文を AI が非文法的と判断

表 7 文法的な文の判別がうまくいっていない構文種

	教師役			一般的利用		
構文種	再現率	適合率	F 値	再現率	適合率	F 値
19: 話題化	0.828	0.706	0.762	0.310	0.750	0.439
31: 外置	0.560	0.875	0.683	0.600	0.769	0.674
34: 寄生空所	0.606	0.690	0.645	0.758	0.625	0.685

4.1.1 話題化構文

(3a) Beans, I don't like.

(3b) This book, I asked Bill to get his students to read.

目的語の名詞句を前置させる話題化 (3a, b) について，生成 AI はいずれも非文法的と判断した．紙

4) 本実験では生成 AI に文法性判断の理由（生成 AI のメタ言語知識）についても答えさせた．紙幅の都合により，ここでは判断の一致率のみ報告する．

5) なお，本研究で得られた全ての構文種の結果は GitHub に公開している．https://github.com/ryoshimura23/grammaticality-judgment

幅の都合で割愛したが，目的語位置を代名詞で表示（例えば (3a) では them）した転位構文についても，生成 AI は非文法的と誤って判断した．この種の構文について容認することが難しいようである．

4.1.2 外置構文

(4a) A review of a new book about French cooking came out yesterday.

(4b) A review came out yesterday of a new book about French cooking.

A review とそれを修飾する前置詞句が隣接する (4a) は生成 AI も文法的であると判断したが，前置詞句を右方に移動させた (4b) のような例は誤って非文法的であると判断した．話題化構文や外置構文など移動を伴う構文については，生成 AI の判断は母語話者の判断とずれる傾向にある．

4.1.3 寄生空所構文

(5a) Which books about himself did John file before Mary read?

(5b) The report which I filed without reading.

(5c) a person that people that talk to usually end up fascinated with.

(5d) Which report did you file without reading it?

同一の filler が 2 箇所の gap を埋める寄生空所構文（(5a)-(5c)）は比較的低い F 値であった．gap の 1 つを代名詞 it に置き換えた (5d) も調査したが，こちらも生成 AI は誤って非文法的であると判断した．

4.2 非文法的な文を AI が文法的と判断

表 8 非文法的な文の判別がうまくいっていない構文種

	教師役			一般的利用		
構文種	再現率	適合率	F 値	再現率	適合率	F 値
8: 否定倒置	0.455	0.294	0.357	0.636	0.350	0.451
15: 否定文	0.563	0.243	0.340	0.563	0.273	0.367
35: 省略文	0.289	0.786	0.423	0.368	0.700	0.483

4.2.1 否定倒置の有無（全文/構成素否定）

(6a) With no job, John would be happy, wouldn't he?

(6b)* With no job, John would be happy, would he?

(6a, b) は構成素否定を伴う付加疑問文である．主節は肯定環境であるため，否定の question tag を有する (6a) は文法的である一方，肯定の tag を有する (6b) は非文法的である．生成 AI は (6a, b) いずれもを文法的であると判断した．

4.2.2 否定文と否定極性表現 (NPI)

(7a) No student who had ever read anything about phrenology attended the lecture.

(7b) * Some student who had ever read anything about phrenology attended the lecture.

(7c) * No one was a bit happy about these facts.

NPI の ever や anything が含まれる (7a, b) では，否定要素を含む (7a) を生成 AI は母語話者と同じく文法的と判断したが，(7b) も文法的と判断した．また，antimorphic の強い否定環境でのみ認可される NPI の a bit (7c) についても，生成 AI は誤って文法的であると判断した．

4.2.3 省略文

(8a) After Bill did, John tried LSD.

(8b)* John did, after Bill tried LSD.

(8c)* Even though she hoped that, Mary doubted that the bus would be on time.

(8d)* John believed Mary to know French but Peter believed Jane to.

(8a,b) は主節と従属節領域での動詞句削除の可否を表している．生成 AI はどちらも例も文法的と判断した．また，(8c) は補文標識の that 以下の時制句削除，(8d) は不定詞 to 以下の動詞句削除が認可されないことを表しているが，生成 AI はこれらの例も文法的であると判断した．

5 まとめと考察

本研究により，文法的な英文に対する文法性判断について AI と母語話者は約 82%程度一致し，非文法的な文に対する評価は約 62%程度一致することが分かった．また，教師の役割を与えるか否かによって文法性判断の精度に顕著な差は見られなかった．特に精度が低い話題化，外置，寄生空所などに共通する特徴は要素の移動を伴うことである．単語の最終的な配置を学習する Attention (Vaswani et. al [7]) が，その前段階に違う配置を想定する統語構造をカバーできていない場合に起きるエラーの可能性がある．判定可否の条件を今後調査する．また，今回取り上げていないが精度が低い文法現象として使役や属格に関する構文が挙げられる．これらは意味的制約も絡み非常に複雑であることから，様々なコーパスに影響を受ける LLM にとって判断が難しいように思われる．これら LLM が苦手とする文法現象に関するベンチマーク開発を今後予定している．

謝辞

本稿の執筆にあたり，稲田俊明氏，西原俊明氏，古賀恵介氏から示唆に富むご助言を賜りました．また，調査過程で伊瀬知ひとみ氏と高橋大智氏からもご協力をいただきました．特記して，これらの方々に感謝申し上げます．本研究は，九州大学人社系学際融合プログラムならびに JST 国家戦略分野の若手研究者・博士後期課程学生育成事業（博士後期課程学生支援 JPMJBS2406) の助成を受けたものです．

参考文献

[1] Jieun Han, Haneul Yoo, Yoonsu Kim, Junho Myung, Minsun Kim, Hyunseung Lim, Juho Kim, Tak Yeon Lee, Hwajung Hong, So-Yeon Ahn, and Alice Oh. Recipe: How to integrate chatgpt into efl writing education. Proceedings of the Tenth ACM Conference on Learning @ Scale (L@S ' 23), pp. 1–8, Copenhagen, Denmark, 2023. Association for Computing Machinery.

[2] Marie-Claude Toriida. Chatgpt: A review of potential uses in education and reported uses in language education. 花園大学文学部研究紀要, Vol. 56, pp. 27–41, 2024.

[3] Hubert Haider. Is chat-gpt a grammatically competent informant? Manuscript, Salzburg University, 2023. https://lingbuzz.net/lingbuzz/007285.

[4] Ken ichi Takami. **Preposition Stranding: From Syntactic to Functional Analyses**. De Gruyter Mouton, 1992.

[5] 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.

[6] 中島平三. 最新 英語構文事典. 大修館書店, 2001.

[7] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In **Advances in Neural Information Processing Systems**, Vol. 30, pp. 5998–6008, 2017.

A 付録

本研究で用いた英語の主要構文 56 種を以下に示す.

1. 受動態
2. 中間態
3. tough 構文
4. 二重目的語構文
5. 同族目的語
6. 存在構文
7. 場所句倒置
8. 否定倒置（全文/構成素否定）
9. 主語・補語倒置
10. 命令文
11. 感嘆文
12. 疑問文
13. 多重 wh 疑問文
14. 付加疑問文
15. 否定文と否定極性表現
16. 強調構文
17. 擬似分裂文
18. 繰上げ構文
19. 話題化構文
20. 転位構文
21. 小節
22. that 節
23. 間接疑問文
24. 関係節
25. 比較構文
26. 副詞節
27. 譲歩節
28. 挿入節
29. 仮定法
30. 等位構造
31. 外置構文
32. 重名詞句転位
33. 二次述語
34. 寄生空所
35. 省略文
36. 不定詞節
37. 動名詞節
38. 分詞構文
39. 属格化
40. (非) 能格・非対格動詞
41. 知覚動詞
42. 使役動詞
43. 叙実動詞
44. 心理動詞
45. 場所格交替動詞
46. 句動詞
47. 成句
48. (法) 助動詞
49. 準助動詞
50. 相
51. 副詞
52. (再帰) 代名詞
53. 数量詞作用域
54. 数量詞遊離
55. 句構造
56. 優位性効果・複合名詞句等の島制約