

生成 AI を用いた鹿児島方言生成

—日琉諸語の低資源言語・方言の生成に向けた試み—

坂井美日

鹿児島大学 共通教育センター

mika.sakai@km.kagoshima-u.ac.jp

概要

本稿では GPT-4 をベースに、低資源方言のひとつである鹿児島方言の生成を試みる。一般に低資源言語・方言の生成は困難であり、鹿児島方言についても zero-shot では精度 0% である。本稿では、方言のテキストと辞書を標準日本語と対訳の形で整備し、更なるその体系の言語知識をプロンプトに入力することで、精度を向上させられるということを示す。また、当方言については、言語知識を仮名ベースの分析を示す方が精度を向上させた。本稿の提案手法による精度は、現時点で 85% 以上である。

1 はじめに

いわゆる生成 AI のベースは大規模言語モデル (以下 LLM) であり、精度はデータ量に依存する。そのため、データ量が少ない低資源言語・方言の生成は一般に困難である。しかし、低資源言語・方言こそ生成 AI による地域課題解決が期待される場所であり (消滅危機方言の継承支援教材や、方言の壁解消ツールの開発等)、これを実現化するための研究が必要である。本稿の試みは、鹿児島方言の生成を実現化することである。後述のように鹿児島方言の生成は、現段階の GPT-4 で zero-shot の精度がほぼ 0% である。このような方言を対象とするモデルの開発には、ファインチューニング等の前段階作業として、機械の方言処理の可否や傾向を観察する必要がある。本稿はその確認作業として、プロンプティングによる実験をおこなった結果を示すものである。

1.1 鹿児島方言について

本稿の対象は、日琉諸語に属する鹿児島方言である。本稿では鹿児島市のデータに絞る。(1) に例を示す。(1) へがわっぜふっでも一びんたがしれ、はよなけにぐっど。

語順や形態素の承接順序等は、標準日本語や他の日琉諸語と共通するが、一方で方言固有語彙や、音

韻、文法面で標準日本語等と異なる特徴を持つことから (木部 2001、平塚 2017 等)、標準語話者が学習無しに理解するのは難しいとされる方言の一つである。(1) は、次のような内容である。

(2) he//hai//=ga wazze fur-Q=de
灰=nom 非常に 降る -NPST=CSL
moo binta=ga sire//siro-i//
感嘆 頭=NOM 白い-ADJS
hajo//haja-u// nake//naka=i// nigu-Q=do
はやくい-ADV 中=DAT 逃げる -NPST=SFP
(灰がひどく降るので、もう頭が白い。はやく中に逃げるよ。)

この例だけでも、方言固有語彙 (びんた=あたま等) や、標準語と異なる音韻特徴 (母音融合: hai→he 等、促音化: fur-Q 等) や、文法特徴 (二段活用: にぐっ=逃げる等) があることが窺える。

当方言は、少数方言かつ低資源方言である。安定した話者は 60 代以上で、若年層への継承が途絶えつつある消滅危機方言である。方言辞書は複数あるものの (橋口 2004 等)、電子公開された方言データは限られており、機会学習可能な言語資源は乏しい。

1.2 低資源言語の学習と生成

低資源言語・方言の生成精度向上に有効だと考えられる一つは、他の高資源言語で補う手法である。

たとえば機械翻訳の先行研究では、低資源言語のモデルの性能を上げるため、転移学習にて他言語のモデルを活用している (Zoph et al. 2016 等)。これは言語生成にも応用できる可能性がある。また、他言語で補う際には、言語間の類似度が性能の向上に影響することが知られている (Martínez-García et al. 2021 等)。これらを踏まえ今回、鹿児島方言と語順などの共通性を持つ日琉諸語の中で最も LLM の学習が進んでいる標準日本語の活用を試みる。

また、2.2 節に示すように、現段階の GPT-4 は、言語学的分析が不得意な傾向にある。本稿では試みとして、言語知識を与えることで精度が向上するか

を検証する。言語学的観点からプロンプティングを検討することは有効だと期待される (Torrent et al. 2023)。本稿の問いは、次のとおりである。

- (3) 問い：日琉諸語の低資源方言である鹿児島方言の生成は、標準日本語と対訳形式で整備されたデータと、その体系の言語知識の説明とをプロンプトで入力することで、精度が上がるか。

以下まず、言語知識の入力で精度が上がるかを検証するため、母音融合現象の処理で実験をおこなう。

2 言語知識の入力は精度を上げるか

2.1 母音融合の実験—枠組みと方法—

当該方言には、母音が連続した際に融合するという現象がある（以下、母音融合現象）。ここではその現象が発生する環境の一つである「イ形容詞」を対象にタスクを設定し、実験をおこなう。

当該方言では形容詞に二系統、伝統方言形の「カ形容詞」と流入形の「イ形容詞」がある。前者は形容詞語幹に形容詞化語尾-ka が後接したものである。後者は形容詞化語尾が-i であるため、語幹末の母音と-i 母音が連続することから、当該方言の体系のなかでは融合（かつ短縮）した結果が表層形となる。

- (4) 例：意味「甘い」

カ形容詞：あまか

イ形容詞：あめ (//amai//→//ame://→//ame/))

今回の実験では、母音融合したイ形容詞を生成できるか検証するため、タスクとして(5) の 20 問を設定する。正解は(6) のとおりである。

- (5) タスク

次の意味の形容詞を鹿児島方言でどのように言うか答えて下さい。

1. ひどい、2. ながい、3. 古い、4. 美しい、5. 高い、6. 安い、7. 速い、8. 遅い、9. 苦しい、10. 熱い、11. 冷たい、12. 重い、13. 軽い、14. まぶしい、15. 明るい、16. 寂しい、17. 強い、18. 弱い、19. 悲しい、20. 白い

【条件】語尾が「か」で終わる形式以外で答えて下さい。つまり「ひどか」「ながか」等、語尾が「か」で終わる形式以外の言い方を答えて下さい

- (6) 正解 1. ひで、2. なげ、3. ふり、4. うつくし、5. たけ、6. やし、7. はえ、8. おせ、9. くるし、10. あち、11. つめて、12. おめ、13. かり、14. まぶし、15. あかり、16. さびし、17. つえ、18. よえ、19. かなし、20. しれ

なお、これらは 4 つの融合規則に沿った例である。

- (7) 融合規則

I. //ai//→/e/ (2, 5, 7, 11, 18)

II. //ui//→/i/ (3, 6, 10, 13, 15)

III. //oi//→/e/ (1, 8, 12, 17, 20)

IV. //i://→/i/ (4, 9, 14, 16, 19)

実験は、Zero-shot prompting と Few-shot prompting、そして言語知識を入力したプロンプトでおこなう。

Zero-shot prompting では、(5) のみ入力する。

Few-shot prompting では(5) の前に(8) を入力する。a~l は、a, c, k が規則 I、b, g, i が規則 II、e, j, l が規則 III、d, f, h が規則 IV にあたる。

- (8) Few-shot prompting

鹿児島方言の形容詞の例を挙げます。

標準日本語＝鹿児島方言

- a. 甘い＝あめ、b. 寒い＝さみ、c. 固い＝かて、d. 嬉しい＝うれし、e. 太い＝ふて、f. 清々しい＝清々し、g. 悪い＝わり、h. 美味しい＝おいし、i. えぐい＝えぎ、j. 黒い＝くれ、k. 痛い＝いて、l. 細い＝ほせ

それでは、上記の例を参考に、鹿児島方言の法則を分析しつつ、（一以下(2) を入力一）

言語知識の入力は、手法として 2 種類考えられる。一つは前述/ame/のように、仮名文字にとらわれず音韻論上の最小単位である「音素」をベースに説明する手法（以下便宜的に「音素ベース」）、もう一つは、仮名文字以上の分析をおこなわずに説明する手法（以下「仮名ベース」）の 2 種類である。それぞれ(9) (10) のプロンプトを用いることとする。実験方法としては、10 回ずつ毎回新規に GPT-4 を立ち上げて生成をおこなう。評価方法としては、正答 1 点／誤答 0 点の 20 点満点で得点化する。

- (9) 音素ベース

鹿児島方言では、母音が連続すると母音が融合します。

（法則）

1. ai→e、2. ii→i、3. ui→i、4. ei→e、5. oi→e

この法則は、形容詞の語末にも適用され、発音に反映されます。

（例）あまい (amai) →あめ

嬉しい (ureshii) →うれし

寒い (samui) →さみ

黒い (kuroi) →くれ

（以下、(5) のタスクを入力）

(10) 仮名ベース

鹿児島方言の形容詞は、次の手順で作ります。

1. 語末の「い」を削除する。
2. 語末から2番目の仮名の発音を変更する。
 - ア段ならばエ段に変更。
 - イ段ならば変更なし。
 - ウ段ならばイ段に変更。
 - オ段ならばエ段に変更。

(例)

甘い (あまい: 「ま」はア段) →あめ
 嬉しい (うれし: 「し」はイ段) →うれし
 寒い (さむい: 「む」はウ段) →さみ
 黒い (くろい: 「ろ」はオ段) →くれ
 (以下(5)のタスクを入力)

2.2 結果

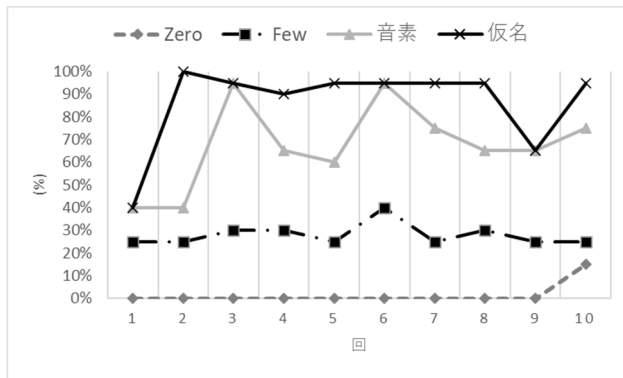


図 1 形容詞の母音融合 (得点率)

表 1 正答数および得点率

母音融合	Zero-shot		Few-shot		仮名ベース		音素ベース	
	正答数	得点率	正答数	得点率	正答数	得点率	正答数	得点率
1回目	0	0%	5	25%	8	40%	8	40%
2回目	0	0%	5	25%	20	100%	8	40%
3回目	0	0%	6	30%	19	95%	19	95%
4回目	0	0%	6	30%	18	90%	13	65%
5回目	0	0%	5	25%	19	95%	12	60%
6回目	0	0%	8	40%	19	95%	19	95%
7回目	0	0%	5	25%	19	95%	15	75%
8回目	0	0%	6	30%	19	95%	13	65%
9回目	0	0%	5	25%	13	65%	13	65%
10回目	3	15%	5	25%	19	95%	15	75%

Zero-shot prompting の得点率は、ほぼ 0%である。回答の多くは標準日本語のまま出現し、複数例で他方言 (ひどい→えらい: 関西方言等、美しい→ちゅら: 沖縄等)、別の表現 (悲しい→しょんぼり等) と、実在しない語彙 (長い→ながーと等) が含まれた。10 回目で 3 問正解が出ているが、これは全ての項目で語尾 (い) を避けるという処理がなされた結果、たまたま 14, 16, 19 が正答と同形になったという

だけである。この結果からも、現時点の GPT-4 では、鹿児島方言をデフォルトのまま正確に生成することは不可能であるということが分かる。

Few-shot prompting の平均正答数は 5.6、平均得点率は 28%である。誤答の多くは、語尾 (い) の削除 (冷たい→つめた等) や、漢字部分の読みを残す処理 (冷たい→つめ等) 等であった。字面上の単純な処理に留まっていることが窺える。「法則を分析」することを指示しても、現段階の GPT-4 は言語学的分析にまでは踏み込まない(もしくは踏み込めない)。

それでは言語学的分析を与えるとどうであろうか。

まず、音素ベースの説明を入力した場合、平均正答数は 13.5、平均得点率は 67.5%であった。Few-shot prompting よりも精度が高くなっており、言語学的分析を与えた方が、精度が向上するということが示される ($t(18)=-6.405$, p 値は約 $4.97e-06 < 0.05$)。

そして仮名ベースの説明を与えた場合、平均正答数は 17.3、平均得点率は 86.5%であった。8/10 回で得点率が 90%以上となり、全問正解も含まれている。音素ベースと比較すると、仮名ベースの説明のほうが、精度が高い ($t(18)=2.242$, $p=0.0378 < 0.05$)。

なお、仮名ベースのほうが精度を上げる傾向は、動詞活用の処理にも見られたので、以下参考に示す。

参考：動詞活用の処理

この実験では、正格活用 (一段・二段・多段) を対象に、8 形式 (否定・丁寧・複合・過去・文中止・基本・假定・命令) を問う全 32 問のタスクを設定したものである。母音融合の実験と同様 10 回ずつ生成。正答 1 点、誤答 0 点で、記号のみずれるケース (言語面は正しく生成されるが、[] の位置がずれる等) は部分点 0.75 点とし、32 点満点で得点化した。

(11) タスク例

問題: はやく【逃げる】ぞ! = はよ [] ど!
 正解: にぐっ

(12) 音素ベースの入力例: 二段

二段系は、語幹末の母音が交代します。

基本形と假定形では語幹 2 を使い、それ以外の場合は語幹 1 を使います。

(例) 起きる・・・語幹 1: oki- / 語幹 2: oku-

- 基本形 語幹 2 oku-φ [おくっ] 【起きる】

- 假定形 語幹 2 oku-reba [おくれば] 【起きれば】

- 命令形 語幹 1 oki-re [おきれ] 【起きろ】

(以下、各形式の音素ベースの分析例)

(13) 仮名ベースの説明例：二段

二段系は、2種類の語幹が交代します。
 基本形と仮定形では語幹2使い、それ以外の場合
 は語幹1を使います。
 (例) 起きる・・・語幹1：おき-／語幹2：おく
 - 基本形 語幹2 おく-っ【起きる】
 - 仮定形 語幹2 おく-れば【起きれば】
 - 命令形 語幹1 おき-れ【起きろ】
 (以下、各形式の仮名ベースの分析例)

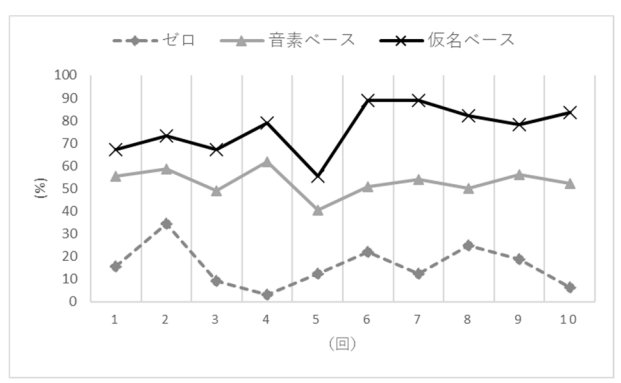


図 2 動詞活用 (得点率)

音素ベースの平均得点率は 52.9%、仮名ベースは 76.4%である (小数点第二位四捨五入)。t 検定の結果、t 統計量は約 6.10、p 値は約 0.00000914 < 0.05 であり、仮名ベースのほうが、精度が高いようである。

3 鹿児島方言の生成

本節では(3)の問いを検証するため、少量の方言資料と方言辞書とを標準語対訳に整理したものと、前節の検証をふまえ仮名ベースの簡易文法概説を作成し、併せてプロンプティングすることで鹿児島方言を生成できるか実験する。

(14) タスク：「鹿児島方言を使って、東京スカイツリーについて説明してもらえますか？」

(15) 使用したデータ

- a. 会話文 (約 20 分)：鹿児島市方言話者 (1959 年代生男性、1960 年代生女性) の会話データ (東京スカイツリーの話はない) を標準日本語と対訳セットにしたもの。
- b. 語彙集 (120 語)：a に出る方言固有語彙を標準語対訳セットにし、動詞は活用の種類と語幹の情報をつけたもの。
- c. 簡易文法：項目は格、形容詞、動詞、文末、接続。

生成結果は(16) の通りである。評価は、鹿児島方言話者らに確認した上で、表 2 の基準で点数化した。

(16) 生成結果と評価

東京スカイツリーは、東京のスキダクにあるどー。この^Bタワーは、ふっとかで、世界ん中で一番高かタワーじゃった。スカイツリーんてっぺんから見っと、東京ん景色が一望できっと、よか眺めじゃらいね。特に夜ん景色は、光がきらきらして^B、きれいかどー。スカイツリーん中には^Bおみせもずんばいあつて、おきゃっさーが買い物^Bしたり、食事したりできっと。また、そこには^B展望台もあつて^B、わっぜ高かところから^B東京を見がでくっど。天気よか日にゃ、富士山も見えっど。スカイツリーへの行き方は、電車で行くのが^B便利じゃっど。東京スカイツリー駅や押上駅からちかいから^B、そっから歩いていけっど。

表 2 基準と点数

評価	基準	点数
A	問題無し	1点
B	不十分 (方言文脈の中にあっても良いが、より伝統的な言い方がある。)	0.3点
C	不自然 (意味はわかるが、当該方言として違和感がある)	0.1点
D	非文 (文法的、意味的に不可)	0点

生成文章には 59 文節あり、その中に D (非文) が含まれていないことが注目される。C (不自然) は 1 か所あり、促音化すべきところがこれに該当した (×あるど／○あつど)。B (不十分) は、主に発音面で標準語寄りとされる所が 11 か所であった。例えば、2 文目「このタワー」については、属格 no の母音脱落形のほうが良いとされる (△この／○こん)。

結果の点数としては、50.4 / 59 点。得点率は 85.4% (小数点第二四捨五入) である。

この結果は、(15) の入力によって方言生成の精度が高くなることを示し、(3) 証明の根拠となりうる。

4 まとめ

本稿では、日琉諸語の低資源方言である鹿児島方言の生成において、データを標準日本語と対訳形式で整備し、その体系の言語知識をプロンプティングすることで、精度が上がることを示した。

今後の課題は、更に精度を上げるため、より機械と相性の良い言語知識提示法を検討することである。

謝辞

本研究は令和5年度鹿児島大学地域活性化研究支援事業の助成を受け、かつ科研費19K13213、22H00007、23H00007Jの成果の一部を反映したものである。

参考文献

木部暢子. 2001. 鹿児島方言に見られる音変化について. 音声研究. 第5巻第3号. pp. 42-48.

橋口満. 2004. 鹿児島方言大辞典(上)(下). 高城書房

平塚雄亮. 2017. 鹿児島県鹿児島市方言. 方言文法研究会編. 全国方言文法辞典資料集(3)活用体系(2). pp. 107-116.

Martínez-García Antonio, Badia Toni, and Barnes Jeremy. 2021. Evaluating morphological typology in zero-shot cross-lingual transfer. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3136–3153, Online. Association for Computational Linguistics.

Torrent Timponi Tiago, Hoffmann Thomas, Almeida Lorenzi Arthur and Mark Turner. 2023. *Copilots for Linguists AI, Constructions, and Frames*. Cambridge : Cambridge University Press.

Zoph Barret, Yuret Deniz, May Jonathan, and Knight Kevin. 2016. Transfer Learning for Low-Resource Neural Machine Translation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1568–1575, Austin, Texas. Association for Computational Linguistics.