

Generating Interactive Japanese Puns Based on Phoneme Similarity

Yilin Wang¹[0009–0007–3516–993X], Takehiro Yamamoto¹[0000–0003–0601–3139],
and Hiroaki Ohshima¹[0000–0002–9492–2246]

University of Hyogo, Kobe, Japan
ad24r004@guh.u-hyogo.ac.jp
t.yamamoto@sis.u-hyogo.ac.jp
ohshima@ai.u-hyogo.ac.jp

Abstract. *Dajare*, a form of Japanese pun, utilizes phonetically identical or similar words and phrases with different meanings to create a funny effect. This paper presents a method for generating *interactive Dajare*, a Japanese pun formed by a response that includes a segment phonetically similar to a part of the original utterance. Our approach involves retrieving candidate words and phrases based on phoneme similarity. We then leverage large language models (LLMs) to generate response sentences for these candidates, which are subsequently ranked according to conversational naturalness to select the most suitable interactive *Dajare*. Our experiment revealed that the proposed approach, by explicitly providing a phonetically similar segment, makes it easier for annotators to identify interactive *Dajare* compared to the baseline, which generates them without providing such a segment. This suggests that our method more effectively produces recognizable interactive *Dajare*. Additionally, the results showed that word selection positively influences the perceived cleverness (the ingenuity of transforming words with similar pronunciations) of interactive *Dajare* than on its funniness. Furthermore, current LLMs still fall significantly short of human capabilities in generating genuinely funny text.

Keywords: Japanese puns · interactive *Dajare* · funniness · cleverness

1 Introduction

The widespread adoption of smartphones and AI has led to dialogue systems like Siri and ChatGPT, which efficiently deliver information. However, these task-oriented systems prioritize accuracy and speed, often neglecting the inherent fun in interactions. Concurrently, the increasing number of single individuals now seeks more engaging, friendship-like conversational experiences, desiring enjoyment that goes beyond mere accurate responses. These are known as *Non-Task-Oriented Dialogue Systems* [3], which prioritize fun over functionality in open-ended dialogues. The primary challenge, however, is for these systems to genuinely emulate the naturalness and intrigue of human conversation.

To address the identified issue, this study aims to enhance the funniness of dialogue by employing wordplay, with a particular emphasis on puns. Wordplay, in its broadest sense, can be defined as the act of playing with words (or the result thereof), which is based on variations of linguistic units without involving formal similarities. It can be categorized based on its functions within discourse, such as amusing people, achieving funny effects, or providing aesthetic pleasure [15]. Consequently, this research intends to utilize a specific form of wordplay known as puns to enhance the funniness in dialogue. The Oxford English Dictionary (OED), a highly regarded reference for the English language, defines “a pun as employment of a word in a manner that suggests two or more meanings or associations, or the use of two or more words that sound alike but have different meanings, thereby producing a funny effect”. This study will concentrate on the Japanese pun known as *Dajare*, which has a definition closely aligned with that of a pun in English [14]. A classic illustration of traditional *Dajare* is provided below, showcasing two words that share identical phonetic qualities (e.g., *kōchō*) yet convey distinct meanings within a single sentence.

Japanese:	校長	先生	絶好調	
Pronunciation:	<u><i>Kōchō</i></u>	<i>sensei</i>	<u><i>zekkōchō</i></u>	1
English:	<u>principle</u>	teacher	<u>favorable</u>	
Gloss:	The <u>principle</u> is in a <u>favorable</u> condition.			

Additionally, another instance of *Dajare* can be observed in dialogues, as demonstrated below. In this context, two parts that are phonetically identical or similar yet semantically distinct appear in two separate sentences, with one part located in the utterance and the other in the response.

Utterance: 先生 は 校長 に 昇進 した よう です!

Utterance: *Sensei wa kōchō ni shōshin shita yō desu!*

Gloss: It seems that the teacher has been promoted to principal !

Response: おお, それ は 好調 な ニュース だ ね!

Response: *Ō, sore wa kōchō na nyūsu da ne!*

Gloss: Oh, that's favorable news !

Utterance: 荒川 の 女の子 は どう です か?

Utterance: *Arakawa no onnanoko wa dō desu ka?*

Gloss: How are the girls at Arakawa ?

Response: あら, 可愛い です よ.

Response: *Ara, kawaii desu yo.*

Gloss: Oh, they are cute...

In this study, we refer to this particular type of *Dajare* as **Interactive Dajare**, and our primary aim is to both generate and comprehensively evaluate them. Our investigation focused on the efficacy of large language models (LLMs)

¹ Adapted from <https://ja.wikipedia.org/wiki/dajare>

² Based the example on the one published in <https://dajare.jp/works/89096/>

in producing interactive Dajare, emphasizing several critical dimensions. The experiment results indicate that the naturalness of the phonetic relationship in interactive Dajare does not substantially affect the perceived funniness. Although surprise alone does not serve as a reliable predictor of funniness, our analysis suggests that semantic similarity can contribute to the generation of more coherent and engaging responses. Crucially, the results of the experiment demonstrate that the interactive Dajare generated by our proposed method are more easily recognized by annotators compared to those from the baseline method, which generates them without providing a phonetically similar segment. Ultimately, despite notable advancements, our findings consistently illustrate that contemporary LLMs remain significantly inferior to human capabilities in the production of genuinely funny text. These observations provide valuable insights into both the potential and the existing limitations of artificial intelligence in the realm of nuanced linguistic funniness.

2 Related Work

2.1 Pun Generation

Hong et al. [7] introduced T-PEG, a system that utilized phonetic and semantic linguistic resources to automatically extract word relationships in puns and store this knowledge in a template format for generating computer-generated punning riddles. Yu et al. [18] introduced an approach for generating homographic puns by utilizing a conditional neural language model, without depending on pre-existing pun data. In addition, He et al. [6] proposed an unsupervised method for generating pun sentences by leveraging a “local-global surprisal” principle seeks a strong association between the pun word and its distant context.

There is some research on the generation of Japanese puns, known as Dajare. For instance, Binsted et al. [2] proposed a method for generating Dajare riddles using templates that substitute homonyms for words with long strings. Yokogawa [17] proposed a method for generating Japanese puns by leveraging the phonetic similarities between words, with the aim of creating funny wordplay. Minami et al. [11] proposed a method for learning Japanese pronunciation. A trained language model is then employed to generate Dajare, which is fine-tuned with pairs of Dajare and their paraphrases.

This study differs from previous studies in that it focuses on the generation of interactive Japanese puns, which differ from non-interactive puns in that they require careful consideration of word choice and the overall naturalness of the dialogue. This presents a significant technical challenge when exploring the creation of interactive puns.

2.2 Pun Evaluation

Kao et al. [8] presented a computational model of fun in puns based on the Bayesian framework of surprise and incongruity. Their research demonstrated that the funniness of a pun correlates with the listener’s probabilistic surprise.

In addition, Binsted et al. [2] conducted a subjective evaluation experiment via email on Dajare riddles created by humans. This study suggests that dictionaries have a significant influence on the funniness of Dajare. Furthermore, Motoki et al. [16] evaluated the funniness of Dajare as a binary classification task and proposed a method for detecting “funny” Dajare using a classifier with two categories: “funny” Dajare and “not funny” Dajare.

While previous research has often focused on traditional Dajare, our study specifically targets interactive Dajare. We employ human evaluation to investigate how dialogue naturalness, topic unexpectedness, and the cleverness of word choice influence their perceived funniness. Crucially, we extend this analysis by exploring the funniness and cleverness derived from the phonetic variations of words within these interactive Dajare.

3 Generation of Interactive Dajare

3.1 Definition of Interactive Dajare

In this study, an interactive Dajare consists of two phonologically identical or similar words/phrases that are semantically distinct, with one word placed in the *utterance* and the other in the *response* to the utterance. The two examples provided in Section 1 both illustrate instances of interactive Dajare.

The **utterance** refers to the sentence uttered by the user in a dialogue and serves as the input for this study. The **response** refers to the reply to the utterance and is the output of this study. In addition, words within the utterance that serve as the basis for phonologically identical or similar elements in the response, such as “校長” (*kōchō* - principal) or “荒川” (*arakawa* - Arakawa) from the examples in Section 1, are referred to as **Dajare source words**. The words or phrases in the response that are phonologically similar to the Dajare source words are referred to as **Dajare target words**. For example, Dajare target words for the Dajare source word “荒川” (*arakawa* - Arakawa) include “新川” (*arakawa* - river), “唐川” (*karakawa* - surname), and “あら,可愛い” (*ara,kawaii* - Oh, cute).

3.2 Overview of the Proposed Method

In this study, we aim to create an interactive Dajare. Figure 1 illustrates the overview of the proposed method, which consists of three phases: retrieval, generation, and ranking.

In the retrieval phase, we extract Dajare source words from a given utterance. Also, the Dajare target words that are phonologically identical or similar to the Dajare source words are then retrieved from a Japanese dictionary. The candidates for Dajare target words are filtered either based on their frequency of occurrence in daily life or their semantic similarity to all Dajare source words.

In the generation phase, a large language model (LLM) is employed to generate candidate responses based on the obtained Dajare target words candidates. At this stage, the output meets certain constraints of interactive Dajare in the

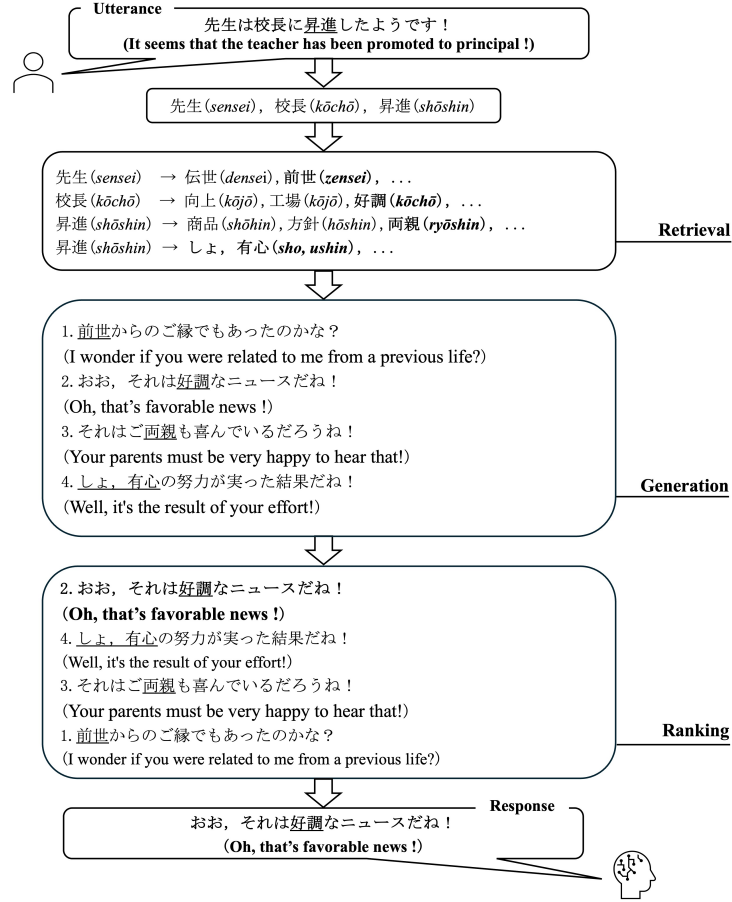


Fig. 1. Overview of the proposed method.

prompt, which include the requirement that the Dajare target words must be incorporated into the responses.

In the ranking phase, we assess the naturalness of the generated candidate responses by considering the interactive Dajare, which includes both the utterance and the response. The most natural response is then selected and returned.

3.3 Retrieval

Mora and Phoneme In order to identify words that are phonologically identical or similar to the Dajare source words, we need to calculate the phoneme similarity between the words, as phonemes are the smallest units of sound in a language. Thus, we introduced the mora, the smallest unit of timing, which is equal to or shorter than a syllable and is used to measure syllable weight [9]. In

Table 1. Phonemes elements.

Consonants	k, g, s, z, c, t, d, n, h, b, p, m, r, v
Semivowels	j, w
Vowels	a, i, u, e, o
Special Phonemes	N(ん), T(っ), R(ー)

Japanese, a mora is composed of the following phonemes: consonants, semivowels, and vowels [1], as illustrated in Table 1.

Therefore, we summarize all possible occurrences of moras and represent them separately as phoneme sequences. For example, the word “校長[コウチヨウ]” (*kōchō* - principal) will first be segmented into individual mora “コ,ウ,チヨ,ウ”³, and then be converted into a phoneme sequence “*k_o,--u,cj_o,--u*”.

Extraction of Dajare Source Words To search for Dajare target words, it is essential to extract Dajare source words from the utterance. In this research, all nouns, verbs, adjectives, and adverbs in the utterance that contain three or more moras are extracted as Dajare source words. In this context, MeCab [10], a morphological analysis tool, is used to process this part.

Then, the obtained Dajare source words are converted into phoneme sequences. For example, from the utterance “先生は校長に昇進したようです!” (It seems the teacher has been promoted to principal!), the Dajare source words “先生” (*sensei* - teacher), “校長” (*kōchō* - principal), and “昇進” (*shōshin* - promotion) are extracted, and their corresponding phoneme sequences (“*s_e,N--,s_e,--i*”, “*k_o,--u,cj_o,--u*”, “*sj_o,--u,s_i,N--*”) are generated.

Candidates of Dajare Target Words We found that homonyms (phonologically identical words) and paronyms (phonologically similar words) commonly appear in Dajare, based on an analysis of existing Dajare data. To further investigate how various linguistic expressions impact the surprise, funniness, and cleverness of interactive Dajare, we additionally include Dajare-Contained Words and Dajare-Contained Phrases. This allows us to explore a broader range of phonetic relationships beyond simple word-for-word identity in interactive Dajare.

In this study, we search for Dajare target words in existing Japanese dictionaries. We utilized the UniDic dictionary [4] and the Balanced Corpus of Contemporary Written Japanese (BCCWJ) Short Word Corpus [12]. The UniDic dictionary contains 878,229 Japanese words, along with a selection of English words and emojis. The BCCWJ Short Word Corpus specifically includes words that occur with a frequency greater than one in various media, such as newspapers and television. It contains 185,137 Japanese words, including their pronunciation and frequency of occurrence.

1. Homonyms

³ <https://qiita.com/shimajiroxyz/items/a133d990df2bc3affc12>

Homonyms are two words that are phonologically identical but have different meanings. We only use the phoneme sequences of Dajare source words to identify homonyms in the dictionary that match these sequences exactly. For example, using the Dajare source word “校長” ($k_o, _ _ u, cjo, _ _ u$ - principal), the Dajare target word “好調” ($k_o, _ _ u, cjo, _ _ u$ - favorable) can be derived as a homonym.

2. Paronyms

Paronyms are two words that are phonologically similar but not identical. Specifically, for this study, paronyms are words where the edit distance between their vowel sequences is zero, and the edit distance between their consonant sequences is one. In this study, semivowels and special phonemes are treated as consonants due to their minimal phonological impact in Japanese.

For example, when comparing the Dajare source word “歯医者” ($h_a, _ _ i, sja$ - dentist) and “会社” ($k_a, _ _ i, sja$ - company) in the dictionary, their vowel sequences (a, i, a) have an edit distance of zero. The consonant sequences ($h, _ _ , sj$) and ($k, _ _ , sj$) have an edit distance of one. This matches our definition, identifying “会社” as a paronym for “歯医者”.

3. Dajare-Contained Words

Dajare-Contained Words are words that include the phoneme sequences of the original Dajare source words. For example, “戦い” ($t_a, t_a, k_a, _ _ i$ - fight) is a Dajare-contained word that contains the Dajare source word “田高” (t_a, t_a, k_a - surname). Based on the position of the phoneme sequence of Dajare source word, we categorize Dajare-contained words into three types: those where the sequence appears at the beginning, such as “戦い” ($t_a, t_a, k_a, _ _ i$ - fight); those where it appears at the end, like “強か” (s_i, t_a, t_a, k_a - determined); and those where it is found in the middle, as seen in “暖かい” ($_ _ a, t_a, t_a, k_a, _ _ i$ - warm).

4. Dajare-Contained Phrases

All the methods described so far have been based on seeking words as Dajare target words. Dajare-Contained Phrases are defined as phrases that include the phoneme sequences of the Dajare source words. These phrases typically consist of two comma-separated words, like “はい, 社長” ($h_a, _ _ i, sja, cjo, _ _ u$ - Yes, Boss).

Our analysis of existing interactive Dajare revealed that combinations like “interjection, noun/adjective/adverb” are particularly common. By acquiring these phrases, we aim to generate more diverse responses of interactive Dajare. We proposed a novel method to obtain these phrases. First, the Dajare source word’s mora sequence is divided into two segments, each with at least one mora. To comprehensively identify interjections, we consider common Japanese interjection endings (“ッ”(T₋) or “ー”(R₋)) when forming the first segment. If an interjection matches the first segment, we then search the dictionary for words corresponding to the second segment (nouns, adjectives, or adverbs). These are then combined with the interjection, separated by a comma. If no match is found, the segmentation is adjusted, and the search is repeated. For instance, if a Dajare source word has four moras, initial segmentation might be 1+3 moras. If no match occurs, this would be adjusted to 2+2 moras, and subsequently to 3+1 moras, exploring all possible two-segment divisions. This approach effectively retrieves Dajare-contained phrases for broader Dajare generation.

With this method, it is possible to generate phrases such as “あら, 可愛い”($_a, r_a, k_a, w_a, _i, _i$ - Oh, cute) and “あら, 代わり”($_a, r_a, k_a, w_a, r_i$ - Oh, substitute) using the Dajare source word “荒川”($_a, r_a, k_a, w_a$ - Arakawa).

Filtering The Dajare target words candidates obtained through the method described above are sourced from the UniDic dictionary; however, some of these words are not commonly used in everyday language. If such uncommon words are included in the response, it is considered an interactive Dajare that may not be understood by many people.

Therefore, we proposed two filtering methods for getting Dajare target words that are easily understood by a wide audience.

(1) Frequency Filtering

This method utilizes the BCCWJ short word corpus to identify the ten Dajare target words with the highest frequency of occurrence.

(2) Similarity Filtering

We manually analyzed the interactive Dajare data and found that, in some cases—such as the example in Section 1—the Dajare target word “あら, 可愛い”($_a, r_a, k_a, w_a, _i, _i$ - Oh, cute) in the response sentence was semantically similar to the Dajare source word “女の子”($_o, N_a, n_a, n_o, k_o$ - girl) in the utterance sentence. Therefore, the similarity filtering proposed in this study ranks Dajare target words based on the semantic similarity between the Dajare source words and the Dajare target words.

First, let $W = \{w_1, w_2, \dots, w_n\}$ be the set of Dajare source words in the utterance, where n represents the number of Dajare source words. Let $P = \{p_1, p_2, \dots, p_m\}$ be the set of Dajare target words obtained as candidates using the method described above. Let m represent the number of Dajare target words selected as candidates by choosing the top 50 Dajare target words based on frequency, which in this study is defined as $m = 50 \times n$. The semantic similarity between the Dajare source words and the Dajare target words is calculated using cosine similarity based on the embedded representations of each word. The similarity between a Dajare target word $p_j \in P$ and a set of Dajare source words W is calculated according to the following formula.

$$\text{sim}(p_j, W) = \max_{w_i \in W} \text{sim}(\mathbf{p}_j, \mathbf{w}_i), \quad j = 1, 2, \dots, m \quad (1)$$

$\mathbf{p}_j$ and $\mathbf{w}_i$ denote the embedded representations of p_j and w_i , respectively. In this study, we utilize the *intfloat/multilingual-e5-large* embedding model⁴. The top 10 Dajare target words, listed in descending order based on the value of the formula 1, are utilized for the generation described in the following section.

Otherwise, in order to verify whether different types of Dajare target words affect the quality of the final generated responses, we filter four types of Dajare target words separately. However, in the Dajare-contained phrases, due to the presence of interjections, the process of filtering these phrases over P applies only to a subset of the second part of the Dajare-contained phrases, specifically the portion that excludes the interjections.

⁴ <https://huggingface.co/intfloat/multilingual-e5-large>

 Generate a response that satisfies the following conditions.

Conditions:

1. Given an utterance S, a Dajare source word W and a Dajare target word P.
2. The generated response must include the Dajare target word P.
3. Words included in the utterance S should be avoided to be included in the response as much as possible.
4. Dajare source word W must not appear in the generated response.
5. The length of the generated response should be short and concise.

Example 1:
 Utterance S: 「先生は校長に昇進したようです！」
 (It seems that the teacher has been promoted to principal !)
 Dajare source word W: 「校長(kōchō - principal)」
 Dajare target word P: 「好調(kōchō - favorable)」

Output 1:
 「おお、それは好調なニュースだね！」
 (Oh, that's favorable news !)

Example 2:
 Utterance S: 「この毛布、すごく温かいんだよ！冬にぴったり！」
 (This blanket is very warm! Perfect for winter!)
 Dajare source word W: 「毛布(mōfu - woolen cloth)」
 Dajare target word P: 「もう、不思議(Mō, fushigi - mysterious)」

Output 2:
 「もう、不思議なくらい温かくて、手放せなくなりそうだね！」
 (It's so mysteriously warm that I'm not sure I'll ever be able to part with it!)

Input:
 Utterance S: 「{Utterance S}」
 Dajare source word W: 「{Dajare source word W}」
 Dajare target word P: 「{Dajare target word P}」

Fig. 2. Prompt for proposed method.

 Generate a response that satisfies the following conditions.

Conditions:

1. Given an utterance S.
2. Generate a response sentence using words or phrases that are phonologically similar to the words in the utterance S.
3. Words included in the utterance S should be avoided to be included in the response as much as possible.
4. The length of the generated response should be short and concise.

Example 1:
 Utterance S: 「先生は校長に昇進したようです！」
 (It seems that the teacher has been promoted to principal !)

Output 1:
 「おお、それは好調なニュースだね！」
 (Oh, that's favorable news !)

Example 2:
 Utterance S: 「この毛布、すごく温かいんだよ！冬にぴったり！」
 (This blanket is very warm! Perfect for winter!)

Output 2:
 「もう、不思議なくらい温かくて、手放せなくなりそうだね！」
 (It's so mysteriously warm that I'm not sure I'll ever be able to part with it!)

Input:
 Utterance S: 「{Utterance S}」

Fig. 3. Prompt for baseline.

3.4 Generation

The method described in Section 3.3 produces a set of Dajare target words that should be incorporated into the responses. In this study, we utilize large language models (LLMs) to generate responses. This is achieved by providing prompts that meet the constraints necessary for creating interactive Dajare.

The prompt provided in this study is shown in Figure 2. The constraints of the prompt are as follows.

- **Explicit Inclusion of a Dajare Target Word:** The prompt explicitly includes a desired Dajare target word, guiding the LLMs to incorporate it into the response sentence.
- **No Words Repetition:** Words present in the utterance are excluded from the response sentence to prevent misunderstanding and ensure clear interactive Dajare.
- **Concise Output:** We instruct the LLMs to generate short and concise responses, reflecting the typical length of existing Dajare.

Using this prompt, we generated candidate responses which included the Dajare target words filtering in Section 3.3.

3.5 Ranking

Finally, we select the appropriate response as the output by ranking the candidate responses obtained. We believe that a suitable response should be a natural response to the utterance, easily understood by a wide audience.

In this study, we use perplexity as a measure of text naturalness. Perplexity is a metric that evaluates how effectively a language model predicts a given text. If the model’s parameters are denoted as θ , the probability of the model generating the text $T = t_1 t_2 \dots t_N$ can be expressed as follows.

$$p(T; \theta) = \prod_{i=1}^N p(t_i | t_1, t_2, \dots, t_{i-1}; \theta) \quad (2)$$

The perplexity of this text, denoted as T , is calculated as follows.

$$\text{Perplexity}(T; \theta) = \exp \left(-\frac{1}{N} \log p(T; \theta) \right) \quad (3)$$

The smaller the perplexity, the higher the probability that the model will generate the text, making it more likely to be perceived as general and natural. In this study, the model utilizes *rinna/japanese-gpt2-small*⁵. Using this perplexity, we obtain the naturalness of the combined dialogues between the utterance and the generated response sentences. The response with the lowest perplexity is selected as the final response.

4 Experiment

4.1 Data

To assess the effectiveness of the proposed method, we construct data for evaluating interactive Dajare. Specifically, we develop our own interactive Dajare data, in addition to those inspired by existing Dajare websites. For our references, we consult DajareStation⁶ and Dajarenavi⁷.

In this study, a total of 50 interactive Dajare (i.e., utterance-response pairs) are created and utilized for the experiments.

4.2 Comparative Methods

In this study, we compare a total of 19 methods, including an Ideal Example, to evaluate manually crafted interactive Dajare, as described in Section 4.1.

The remaining methods utilize either *GPT-4o*⁸ or *Llama-3.1-Swallow-8B-Instruct-v0.3*⁹ [5,13](referred to as Swallow) for response generation. For each model, we explore various configurations:

- Baseline: Generating a response without providing a Dajare target word, based on the prompt in Figure 3.

⁵ <https://huggingface.co/rinna/japanese-gpt2-small>

⁶ <https://dajare.jp/>

⁷ <https://dajarenavi.net/>

⁸ <https://openai.com/index/hello-gpt-4o/>

⁹ <https://swallow-llm.github.io/llama3.1-swallow.ja.html>

- **Dajare Target Words Inclusion:** Generating responses by incorporating specific Dajare target words into the prompt. These words are categorized and filtered using two filtering methods:
 - **Frequency Filtering** (from Section 3.3): Using homonyms, paronyms, Dajare-contained words, and Dajare-contained phrases.
 - **Similarity Filtering** (from Section 3.3): Using homonyms, paronyms, Dajare-contained words, and Dajare-contained phrases.

To ensure a diverse range of responses, we set the temperature parameter of the GPT-4o model and the Swallow model to 1. In the Swallow model, the top-p parameter is set to 0.9, and the max_tokens parameter is set to 128.

4.3 Evaluation Criteria

In order to assess the quality of the generated responses, we seek three annotators to evaluate the responses produced by each method based on five criteria.

- **Recognition(Rec.):** Evaluate whether the utterance and response constitute an interactive Dajare. If the annotators recognize that the response contained a word or phrase phonetically similar to that in the utterance sentence, they assigned a score of 1; otherwise, they assigned a score of 0.
- **Naturalness(Nat.):** Evaluate whether there are grammatical errors or discrepancies in the coherence of the dialogue’s meaning. Even if it is not an interactive Dajare, a response sentence that feels natural in the context of a dialogue is scored as 1; otherwise, it receives a score of 0.
- **Surprise(Sur.):** Evaluate whether the response takes an unexpected turn in relation to the topic of the utterance. If the response feels surprising, it is rated as 1; if it does not feel surprising, it is scored as 0.
- **Funniness(Fun.):** Evaluate the funniness of the response. If the response feels funny immediately upon reading, it is rated as 1; if it feels funny after some reflection, it is rated as 0.5; and if it does not feel funny at all, it is rated as 0.
- **Cleverness(Cle.):** Evaluate whether the word changes in response are clever, considering that the dialogue is recognized as an interactive Dajare. If the word changes are particularly clever, the score is rated as 1; if the word changes are predictable and lacked creativity, the score is rated as 0.5; and if the word changes are entirely unoriginal, the score receives a rating of 0.

Otherwise, we define **Fun(Bin).** and **Cle(Bin).**, where values evaluated as 0.5 or 1 are considered to be 1, while values evaluated as 0 remain 0. Specifically, we develop a new evaluation criteria called “Funniness as an interactive Dajare(**Daj.**)” using the following calculation formula: if the annotators scored “Recognition” as 0, the “Funniness as an interactive Dajare” score is set to 0; if the annotators scored “Recognition” as 1, the “Funniness as an interactive Dajare” score is the same as “Funniness” score. This criteria is to judge the funniness of the interactive Dajare.

Furthermore, we use Fleiss’ kappa to assess the reliability of agreement among three annotators, and the arithmetic mean of the scores provided by the three annotators to evaluate the performance of each criteria.

Table 2. Results of Fleiss’ κ .

Rec.	Nat.	Sur.	Fun.	Fun(Bin).	Cle.	Cle(Bin).
0.69	0.67	0.57	0.40	0.44	0.21	0.24

Table 3. Results of comparative methods.

Model	Filtering	Dajare Target Words	Rec.	Nat.	Sur.	Fun.	Daj.	Cle.
Ideal Example	N/A	N/A	0.89	0.85	0.33	0.35	0.34	0.67
GPT-4o	Baseline	N/A	0.23	0.43	0.36	0.10	0.05	0.09
	Frequency	Homonyms	0.74	0.36	0.37	0.05	0.04	0.18
		Paronyms	0.76	0.48	0.32	0.06	0.06	0.16
		Dajare-Contained Words	0.74	0.26	0.40	0.05	0.05	0.19
		Dajare-Contained Phrases	0.81	0.23	0.46	0.04	0.04	0.18
	Similarity	Homonyms	0.73	0.41	0.34	0.05	0.05	0.21
		Paronyms	0.59	0.48	0.25	0.05	0.04	0.15
		Dajare-Contained Words	0.77	0.27	0.39	0.04	0.03	0.18
		Dajare-Contained Phrases	0.82	0.27	0.47	0.07	0.06	0.22
Swallow	Baseline	N/A	0.00	0.69	0.11	0.05	0.00	0.00
	Frequency	Homonyms	0.63	0.42	0.30	0.06	0.06	0.17
		Paronyms	0.67	0.32	0.31	0.01	0.01	0.15
		Dajare-Contained Words	0.67	0.29	0.43	0.03	0.02	0.14
		Dajare-Contained Phrases	0.82	0.18	0.53	0.07	0.05	0.20
	Similarity	Homonyms	0.62	0.39	0.29	0.02	0.02	0.13
		Paronyms	0.55	0.43	0.23	0.05	0.03	0.15
		Dajare-Contained Words	0.81	0.21	0.38	0.06	0.05	0.17
		Dajare-Contained Phrases	0.76	0.31	0.40	0.07	0.07	0.23

5 Result

The results of Fleiss’ kappa are presented in Table 2, which indicate that moderate agreement was observed for Recognition (0.69), Naturalness (0.67), and Surprise (0.57), suggesting that annotators had a fairly consistent judgment regarding grammaticality, contextual appropriateness, and overall quality. In contrast, agreement on Funniness (0.40 and 0.44) and especially Cleverness (0.21 and 0.24) was notably lower, indicating that perceptions of funniness and creativity in wordplay were more subjective and varied significantly among annotators.

Meanwhile, the results of comparative methods are presented in Table 3. The Ideal Example, manually constructed, established a clear benchmark by achieving the highest scores across nearly all dimensions (e.g., Recognition: 0.89, Naturalness: 0.85, Funniness: 0.35, Cleverness: 0.67). In contrast, both the GPT-4o and Swallow models demonstrated significantly lower overall performance, with baseline methods yielding the worst results (e.g., Swallow Baseline: Recognition = 0.00, Cleverness = 0.00). Notably, our proposed methods significantly improved Dajare Recognition (Rec.) for both LLMs (GPT-4o: 0.59-0.82; Swallow: 0.55-0.82) compared to their baselines. This substantial gap underscores that LLMs inherently struggle to identify phonetically similar segments within an utterance and seamlessly integrate them into a coherent response.

For both models, Dajare-contained phrases under the similarity-based filtering tend to yield the highest scores among automatic generation methods across several dimensions. Notably, “GPT-4o + Similarity + Dajare-Contained Phrases” achieves the highest scores in Recognition (0.82), Surprise (0.47) and

Table 4. Examples of generated responses.

Details	Interactive Dajare
GPT-4o	
Baseline	今日の海、波浪($h_a, r_o, _u$)がすごいね! (The ocean today, the waves are amazing!) 波が騒いでいて、まるでハロー($h_a, r_o, _u$)と言ってるみたいだね! (The waves are making so much noise, it's like they're saying hello!)
Similarity, Homonym	最近、体調($t_a, _i, c_j o, _u$)はどうだ? (How are you feeling these days?) ああ、体長($t_a, _i, c_j o, _u$)が伸びちゃったかな? (Oh, have I grown too long?)
Frequency, Paronym	破竹(h_a, c_i, k_u)の勢いで走ってきたんだって? (I hear you've been running at breakneck speed.) それはまるで社畜($s_j a, c_i, k_u$)のように全力疾走だね! (That's like running as fast as you can, like a company animal!)
Swallow	
Similarity	今日是用事($j o, _u, z_i$)があつて忙しかったんだ。 (I've been busy with things to do.)
Dajare-Contained Word	今日是要人($j o, _u, z_i, N_u$)がいっぱいいいたみたいだね! (There seemed to be a lot of important people today!)
Frequency	秋田($_a, k_i, t_a$)の名物料理、きりたんぽ鍋ばかりで...
Dajare-Contained Phrase	(All I could eat was kiritampo hot pot, a specialty of Akita...) あ、北方($_a, k_i, t_a, g_a, t_a$)の味が恋しいなあ。 (Oh, I miss the taste of the north.)

Cleverness (0.22) within model-generated responses. Despite advancements in recognition and naturalness, funniness scores remain low across all settings, typically below 0.07. The derived metric “Funniness as an interactive Dajare” shows similarly low values.

Meanwhile, we presented several examples of the generated interactive Dajare in Table 4. Aside from the dialogue (Swallow + Similarity + Dajare-Contained Word), which was primarily evaluated for its cleverness, the other dialogues were generally regarded as funny, showcasing clever interactive Dajare. We note that most of the highly-rated funny interactive Dajare exhibit a natural conversational flow and incorporate word transformations that stimulate associations. However, the dialogue (GPT-4o + Similarity + Homonym) highlights an intriguing phenomenon: when words have extremely high phonetic similarity, they can lead to misunderstandings that unexpectedly enhance the funniness. This also suggests that in interactive Dajare, phonetic ambiguity occasionally adds an extra layer of unexpected amusement.

6 Discussion

This study examined the performance of large language models (LLMs) in generating interactive Dajare, focusing on several key aspects: the impact of word selection on funniness and cleverness, the correlation between surprise and funniness, differences in filtering strategies, and variations among models.

To begin with, across various models and filtering strategies, the funniness scores (Fun. and Daj.) consistently remain low, regardless of the method used

for selecting Dajare target words. This indicates that the type of phonetic relationship alone does not significantly influence perceived funniness.

In contrast, cleverness scores exhibit greater variation depending on the word selection method. Specifically, Dajare-contained phrases consistently attain higher scores across both models and filtering methods (e.g., GPT-4o + Similarity: 0.22; Swallow + Similarity: 0.23). This suggests that phrase-level substitutions are more likely to be recognized as phonetic matches and are often perceived as more clever.

Next, while certain settings yield higher surprise scores (e.g., GPT-4o + Frequency + Dajare-Contained Phrases: 0.46), the funniness scores remain low (0.04 and 0.04). This lack of correlation indicates that surprise alone does not predict funniness. In other words, although unexpected twists in dialogue may enhance engagement, they do not inherently elicit laughter.

Additionally, comparing two filtering methods within the same word selection strategy reveals that similarity-based method generally achieved slightly higher scores across multiple dimensions, including naturalness, funniness, and cleverness. This trend indicates that filtering by semantic similarity may enhance the generation of more coherent and engaging responses.

Ultimately, GPT-4o and Swallow performed comparably in generating interactive Dajare, suggesting a similar level of linguistic sophistication for this specific task. However, our findings confirm that current LLMs still fall significantly short of human capabilities in generating genuinely funny text.

Otherwise, the lack of a large-scale interactive Dajare dataset makes it difficult to conduct quantitative analyses on the factors that contribute to the perceived funniness in such dialogues. And the absence of multi-turn dialogue contexts made it challenging for annotators to assess funniness effectively.

7 Conclusion

In this study, We proposed a method for generating interactive Dajare, which includes retrieval, generation, and ranking. The experimental results revealed that our approach is more effective in generating recognizable interactive Dajare compared to the baseline, which lacks providing a phonetically similar segment. The results of the experiment also showed that word selection positively influences the perceived cleverness of interactive Dajare than on its funniness. Additionally, we confirmed that current large language models (LLMs) still fall notably short of human capabilities in generating genuinely funny text. In the future, we will concentrate on constructing a specialized dataset of interactive Dajare. And, focus on Reinforcement Learning from Human Feedback (RLHF) to guide models in generating more funny responses.

Acknowledgments

This work was supported by JSPS KAKENHI Grant Numbers JP25K03229, JP24K03228, and JP25K03228.

References

1. Akinaga, K.: How many syllables (mora) are there in Japanese? Lectures on Japanese Language Education **5**, 11–21 (1969), (in Japanese)
2. Binsted, K., Takizawa, O.: Computer generation of puns in Japanese. Sony Computer Science Lab, Communications Research Lab (1997)
3. Chen, H., Liu, X., Yin, D., Tang, J.: A Survey on Dialogue Systems: Recent Advances and New Frontiers. *Association for Computing Machinery* **19**(2), 25–35 (2017)
4. Den, Y., Ogiso, T., Ogura, H., Yamada, A., Minematsu, N., Uchimoto, K., Koiso, H.: The development of an electronic dictionary for morphological analysis and its application to Japanese corpus linguistics. *Japanese Linguistics* **22**, 101–123 (2007)
5. Fujii, K., Nakamura, T., Loem, M., Iida, H., Ohi, M., Hattori, K., Shota, H., Mizuki, S., Yokota, R., Okazaki, N.: Continual Pre-Training for Cross-Lingual LLM Adaptation: Enhancing Japanese Language Capabilities. In: *Proceedings of the 1st Conference on Language Modeling* (2024)
6. He, H., Peng, N., Liang, P.: Pun Generation with Surprise. In: *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1*. pp. 1734–1744 (2019)
7. Hong, B.A., Ong, E.: Automatically Extracting Word Relationships as Templates for Pun Generation. In: *Proceedings of the Workshop on Computational Approaches to Linguistic Creativity*. pp. 24–31 (2009)
8. Kao, J.T., Levy, R., Goodman, N.D.: A computational model of linguistic humor in puns. *Cognitive science* **40**(5), 1270–1285 (2016)
9. Kubozono, H.: On the Universality of Mora and Syllable (Features on Theories of Syllable and Mora). *Journal of the Phonetic Society of Japan* **2**(1), 5–15 (1998)
10. Kudo, T.: *McCab : Yet Another Part-of-Speech and Morphological Analyzer* (2005)
11. Minami, T., Kiyoo, Y., Tahara, Y., Osuga, A.: Generation of Puns by Paraphrasing Sentences Using a Language Model Trained on Japanese Kana Notation. In: *The 16th JSAI International Symposia on AI, 2G5–GS–6–03*. pp. 1–4 (2024), (in Japanese)
12. National Institute for Japanese Language and Linguistics, Center for Corpus Development: Balanced Corpus of Contemporary Written Japanese (BCCWJ) Word-unit Frequency List (Version 1.0). <https://repository.ninjal.ac.jp/records/3234>
13. Okazaki, N., Hattori, K., Shota, H., Iida, H., Ohi, M., Fujii, K., Nakamura, T., Loem, M., Yokota, R., Mizuki, S.: Building a Large Japanese Web Corpus for Large Language Models. In: *Proceedings of the 1st Conference on Language Modeling* (2024)
14. Takashi, O.: Dajare is more Flexible than Puns: Evidence from Word Play in Japanese (<Feature Articles> Current Trends in Experimental OT and Laboratory Phonology). *Journal of the Phonetic Society of Japan* **14**(1), 76–85 (2010)
15. Thaler, V.: *Varieties of Wordplay*, pp. 47–62. De Gruyter (2016)
16. Yatsu, M., Araki, K.: An Analysis of Factors of Funniness of Japanese Puns. *Proceedings of the Fuzzy System Symposium* **32**, 237–242 (2016), (in Japanese)
17. Yokogawa, T.: Generation of Japanese puns based on similarity of articulation. In: *Proceedings of the Joint 9th IFSA World Congress and 20th NAFIPS International Conference (Cat. No. 01TH8569)*. vol. 4, pp. 2259–2264 (2001)
18. Yu, Z., Tan, J., Wan, X.: A Neural Approach to Pun Generation. In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. pp. 1650–1660 (2018)