# Poems to Lyrics: Automated Rephrasing with Beat Alignment

**Mohamad Elzohbi, Richard Zhao**
Department of Computer Science, University of Calgary
Calgary, Alberta, Canada T2N 1N4
{melzohbi, richard.zhao1}@ucalgary.ca

## Abstract

This paper explores the generation of English lyrical lines that align with the rhythm perceived from language. We use the ByT5 transformer model, which processes text at the byte level, to rephrase poetry lines according to a given beat pattern. Our approach builds on earlier studies on beat patterns perceived from English conversations by integrating a guided paraphrasing task with rhythmic constraints. Additionally, we introduce PARAPOETRY, a large-scale parallel dataset of automatically generated poetry line rephrases. Our results demonstrate that the proposed model can effectively rephrase lyrics to align with specific beat patterns using only textual data. A human evaluation study further confirms that English-speaking participants largely agree with our model's beat alignment. We further assess other qualities, including fluency, meaningfulness, and poeticness.

## Introduction

Recent advances in natural language generation (NLG) have renewed scholarly interest in creative data generation, particularly for creative language such as poetry (Elzohbi and Zhao 2023). Transformer-based models (Vaswani et al. 2017) have significantly improved text quality through transfer learning, which in turn has opened avenues for integrating and exploring additional creative elements into the generation process. While both poetry and song lyrics are expected to display poetic qualities in a sense, they differ in their structural and rhythmic requirements. Unlike poetry, song lyrics must adhere to a rhythmic structure that aligns with musical accompaniment. This distinction highlights the need for specialized models that incorporate sub-syllabic aspects, ensuring that generated lyrics are not only linguistically and poetically coherent but also musically coherent.

While traditional approaches have primarily focused on rhyme and poetic coherence, comparatively little attention has been given to the rhythmic structure of both poetry and lyrics. In particular, song lyrics require a dual focus: they must display creative language while also conforming to the rhythmic patterns of musical accompaniment. This challenge calls for methods that integrate linguistic creativity with the control over rhythmic patterns.

In this work, we focus on the task of rephrasing poetry lines to achieve alignment with desired beat patterns. Our objective is to preserve the poetic quality of the original lines while ensuring that the rephrased versions conform to the rhythmic requirements essential for musical integration. As a result, our focus will be on *lyrical poetry*. To achieve this, our approach leverages the ByT5 transformer model, renowned for its byte-level processing and fine-grained control over sub-syllabic features critical to beat alignment. We introduce PARAPOETRY, a large-scale dataset of poetry line rephrases automatically generated via GPT-3.5, which we believe it can serve as a valuable resource for various creative language generation tasks.

Our work contributes to the intersection of natural language processing and music, offering insights into beat-aligned language generation and laying groundwork for future research in this domain. Our contributions can be summarized as follows:

- We present PARAPOETRY, a large-scale dataset that pairs human-written poetry lines with automatically generated rephrases. We believe that this dataset can serve as a valuable resource for future research in creative text generation.
- We demonstrate the utility of the proposed dataset for our approach through a guided rephrasing task that integrates beat pattern alignment into the generation process. Our model is trained to generate lyrical poetry that align with desired beat percussion.
- We conduct human evaluations confirming that verses generated by our models adhere to the specified beat patterns, highlighting the potential of our method to support creative lyric writing through a co-creative process.

## Related Work

Many studies focus on the rhyme of lyrics but often overlook their essential rhythmic aspect (Xue et al. 2021). The task of generating lyrics given a melody has been explored using various techniques and in different languages. Some approaches use text-based or symbolic melody inputs, while others use audio-based representations. Songmass (Sheng et al. 2021) introduced transformer-based encoder-decoder models for modelling lyrics and symbolic melodies. The authors used supervised training to align text in lyrics with notes in melodies, generating textual lyrics from MIDI-based melodies and vice versa. Xue et al. (2021) proposed

| Statistic | Value | Description |
|---|---|---|
| Exact Match Percentage | 0.061% | Percentage of cases where original lines and rephrases are exactly identical, indicating that rephrases are not mere copies. |
| Rhythm Match Percentage | 0.285% | Percentage of instances where the beat pattern of original lines and rephrased lines match, demonstrating rhythmic diversity. |
| Total Poetry Lines | 1,038,743 | Total number of poetry lines in the original dataset. |
| Total Rephrased lines | 934,054 | Total number of lines rephrased by GPT-3.5. |
| Unique lines | 907,450 | Number of unique original poetry lines in PARAPOETRY. |
| Line-Rephrase Pairs | 4,674,190 | Total count of line-rephrase pairs present in PARAPOETRY. |
| Unique Rephrases | 4,659,608 | Number of unique rephrases in PARAPOETRY. |
| Maximum Rephrase Count | 271 | Maximum number of rephrases attributed to a single line. |
| Minimum Rephrase Count | 1 | Minimum number of rephrases a line has received. |
| Average Rephrase Count | 5.00419 | Average number of rephrases per line. |

Table 1: Summary Statistics of the PARAPOETRY Dataset.

an autoregressive transformer-based model to produce beat-aligned lyrics by collecting and processing a rap lyric dataset to identify beats and word timestamps, inserting beat positions into the lyrics as special tokens.

Other approaches extract phonetic features from lyrics and compare them with music. For example, Oliveira, Cardoso, and Pereira (2007) proposed heuristics for generating lyrics that align with a MIDI-based melody by matching stressed syllables with strong beats. More recently, Chen and Teufel (2024) explored similarities between tonal contours in melody and lyrics using phonetics and musicology theories, extracting pseudo-melodies from lyrics to fine-tune a transformer-based model (mBART) for lyrics generation given a topic and a melody.

In our previous work (Elzohbi and Zhao 2024), we developed a model that extracts beat patterns from English words by training ByT5 (Xue et al. 2022) to replace or insert words so that they align with a desired beat pattern. We refer to this task as the *substitution task*. This approach builds upon earlier research into perceptual beat locations in spontaneous English conversations (Allen 1972; Rathcke et al. 2021), where participants tapped their fingers in synchrony with perceived beats in spoken words, demonstrating that rhythmic beats often align with vowel onsets.

However, the *substitution task* is limited to replacing words within a lyrical line to match a beat, and lacks the capacity to generate complete lines that align with a given beat pattern. Moreover, while prior evaluation relied on automated metrics, it did not incorporate human judgment to assess rhythmic alignment or the overall quality of the modified lines. In this work, we address these limitations by developing a model capable of rewriting entire lines of lyrical poetry to match a target beat pattern, while preserving both semantic content and poetic expression. We will refer to this task as the *rephrase task*.

A parallel dataset with beat information is ideal for the beat-aligned rephrase task. Unfortunately, specialized datasets for poetry and lyrics are generally scarce, with most studies creating their own datasets using automatic methods (Greer et al. 2019; Sulun, Oliveira, and Viana 2023; Martinez-Sevilla et al. 2023; Chen and Teufel 2024). For this reason, we automatically generated a large-scale dataset of poetry rephrases, that we will call PARAPOETRY, by prompting OpenAI's GPT-3.5. In this work, we demonstrate the utility of this dataset for beat-aligned lyrical poetry rephrasing.

# Methodology

## Task Formalization

Given a line of lyrics $L$ with an undesirable beat pattern $G2B(L)$, the task is to generate a rephrased line $L'$ with a desired beat pattern $G2B(L')$. The function $G2B(\cdot)$, first introduced in our previous work (Elzohbi and Zhao 2024), is a grapheme-to-beat transformation that maps text to a sequence of beat/rest annotations. A beat unit is marked with a "1" at a vowel onset within a consonant-vowel sequence, as well as at the initial vowel of null-onset syllables where either a glottal stop serves as its onset or the onset is borrowed from the preceding syllable. All other positions, including the second elements of long vowels and diphthongs, consonants within clusters, or consonants in codas, are marked with a "0", unless they are repositioned as onsets for the following null-onset syllables.

## Dataset Creation, Augmentation and Processing

Due to the lack of human-generated poetry-rephrase parallel datasets, we created PARAPOETRY, which comprises of human-written poetry lines with five automatically generated rephrases each. We started with the preprocessed subset of the Gutenberg poetry corpus we used for the substitution task, focusing specifically on English quatrains. Using GPT-3.5,[1] we generated five rephrases for each human-written poetry line, with an emphasis on maintaining poetic language.

Through iterative prompt engineering, we manually refined and evaluated sample responses of around 15 prompts to achieve the best results in guiding GPT-3.5 in retaining poetic devices such as metaphors and similes of the originals

[1]We used version: `gpt-3.5-turbo-0125`

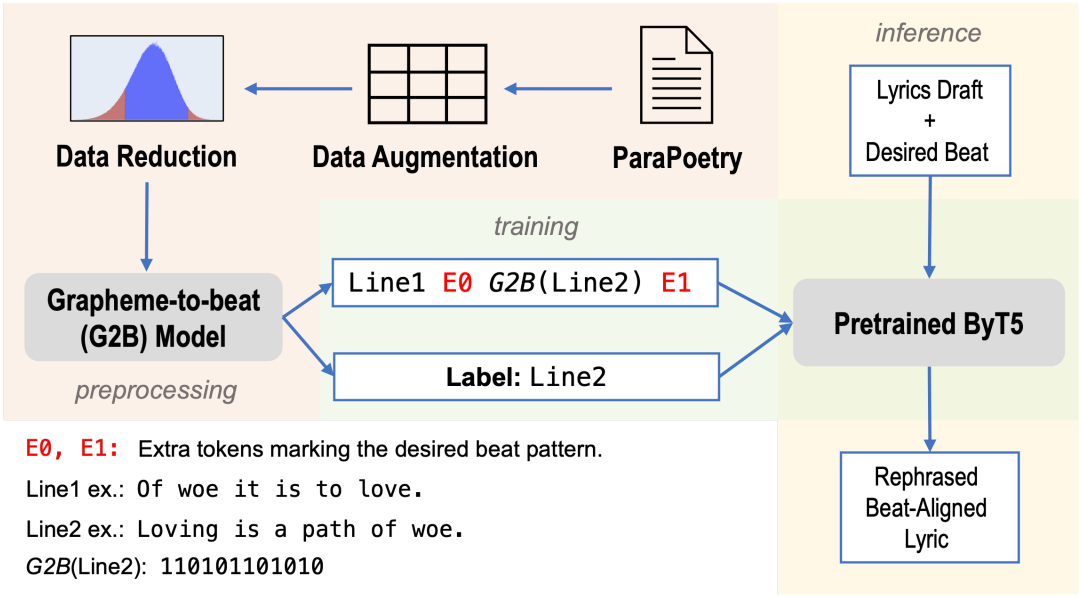


Figure 1: The preprocessing, training and inference pipelines.

in the rephrases while following formatting requirements. The prompt included providing two illustrative examples and demonstrating how to handle anomalies as a few-shot. We also experimented with temperature settings, which control the randomness and creativity of the output. We found that a temperature of $0.7$ struck a balance between following formatting instructions and producing poetic paraphrases. In the generation process, there were instances where GPT-3.5 ignored rephrasing certain lines or generated more or less than 5 rephrases resulting in more than $4.6$ million line-rephrase pairs (see statistics in Table 1).

In PARAPOETRY, the original lines are written by real poets, unlike the generated rephrases. To help our model be biased toward generating human-like rephrases, we augmented the dataset by swapping the rephrases with the original lines, avoiding a bias toward automatically generated output. We then filtered out very long and very short lines, retaining only those with more than $8$ and fewer than $128$, after examining the length distribution of the dataset we found that the majority of the lines fall within this range.

We analyzed the dataset to test lexical dissimilarity between the original lines and the rephrases. We calculate the PINC score (Paraphrase In N-gram Change) (Chen and Dolan 2011). The PINC$(o, p)$ score measures the degree of lexical novelty of the automatically generated rephrases by comparing the generated rephrase $p$ with the original line $o$, analyzing differences in n-grams. It measures the proportion of n-grams in the paraphrase that are absent in the original, averaging this across n-gram lengths. A score of "1" means the original line and the rephrase are very different, and a score of "0" means they are very similar lexically.

$$\text{PINC}(o,p) = \frac{1}{N}\sum_{n=1}^{N} 1 - \frac{|n\text{-gram}_o \cap n\text{-gram}_p|}{|n\text{-gram}_p|}$$

After calculating the PINC scores with $N = 4$, we observed a normal distribution centered around $0.521$, which indicates a moderate level of lexical dissimilarity between the original lines and the rephrases in our dataset. Since lexical dissimilarity can impact the rhythmic diversity of generated lines by altering the positions of vowel onsets, and consequently the beat pattern, a moderate average allows for flexible rephrasing while accommodating varying levels of beat variation relative to the original lines.

We trim the tails of the distribution keeping data with PINC scores between $0.2$ and $0.8$. This excludes data with minimal substitutions and potential garbling.

## Experimental Setup

### Dataset Split

To accommodate our available resources, we sample a subset of 3 million pairs from the processed dataset to train and evaluate our models with an evaluation set containing $15,000$ examples.

### Model Training

For this task, we fine-tuned pre-trained ByT5 models on the processed subsets using a guided rephrasing objective. For each lyric line $L$ and its rephrased version $L'$, we computed the beat pattern $G2B(L')$. To mark the boundaries of the guiding beat pattern, special tokens ($E_0$ and $E_1$) were added and concatenated to the end of the lyric line $L$. The input is structured as $I = (L, E_0, G2B(L'), E_1)$. The model was trained to predict $L'$ (see Figure 1).

We fine-tuned the pre-trained ByT5-base version (Xue et al. 2022) on the task described using the training subset to get **ByT5-R**. As the rephrase task is arguably more complex, yet similar, than the substitution task, we also investigated whether the knowledge gained from the model fine-

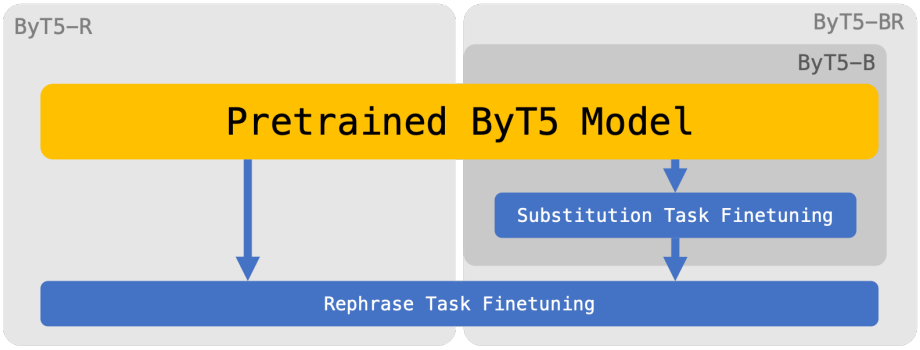


Figure 2: Curriculum learning pipeline.

tuned on the substitution task (**ByT5-B**) could be transferred to enhance rephrasing through a curriculum learning strategy. Curriculum learning is a strategy that gradually raises complexity to help the model learn more effectively. Forms of increased complexity may include increasing the data (or model's experience) complexity during the training, expanding the model's capacity by adding or activating more layers or units, or training on simpler tasks before moving to more complex ones (Soviany et al. 2022).

Here, we apply curriculum learning by first training on the simpler substitution task, then on the more complex rephrasing task (see Figure 2). This yields the model **ByT5-BR** using **ByT5-B** as a starting point.

As a baseline, we train another ByT5-base with attention masks set to zero in the beat pattern positions, named **ByT5-R-Zero**, preventing the model from attending to these patterns during training.

All models were trained for four epochs on an NVIDIA A100 GPU. The learning rate was set to $3e-4$ with a cosine scheduler and a weight decay of $0.1$. The training batch size was set to $128$, and the evaluation batch size set to $16$.

### Automated Evaluation Metrics

The main objective of our task is to replace a verse so that it aligns with a specific beat pattern while ensuring the new verse maintains semantic similarity to the original. We use automated evaluation metrics to measure the sentence semantic similarity, lexical dissimilarity and beat alignment.

**Sentence similarity:** We used Sentence-BERT[2] (Reimers and Gurevych 2019) to encode the original and the rephrased lines. Cosine similarity was utilized to measure the similarity between their embeddings.

**Lexical dissimilarity:** To assess the lexical diversity of the generated rephrases, we use the PINC score with $N = 4$. This score is used as a descriptive metric to show the level of lexical novelty in the rephrased lines in comparison with similarity level.

**Alignment Scores:** assess how well the generated lines align with the required beat pattern. We utilized two evaluation metrics:

- *Exact Alignment Accuracy:* This metric determines if the generated line aligns precisely with the expected beat rhythm, resulting in "0" for non-alignment, and "1" for exact alignment.
- *Levenshtein similarity:* This metric quantifies alignment by calculating the complement of the Levenshtein distance $d_{lev}$ between the beat of the generated line $B_o$ and the expected beat $B_e$, normalized. It allows for some flexibility, recognizing cases where alignment may not be exact but is still reasonably close. The Levenshtein similarity is defined as follows:

$$\text{Sim}(B_o, B_e) = 1 - \frac{d_{lev}(B_o, B_e)}{\max(\text{len}(B_o), \text{len}(B_e))}$$

## Experimental Results

This section examines the performance of the fine-tuned models in meeting the task objectives. We evaluate their performance based on sentence similarity, lexical dissimilarity, and beat alignment.

| Model | Sim. (%) | PINC (%) | Acc. (%) | Lev. (%) |
|---|---|---|---|---|
| ByT5-R-Zero | 74.42 | **45.60** | 7.00 | 84.71 |
| ByT5-R | 81.29 | 30.78 | 91.01 | 99.46 |
| ByT5-BR | **82.04** | 29.71 | **94.34** | **99.67** |

*Sim.* = Sentence Similarity, *Acc.* = Exact Alignment Accuracy, *Lev.* = Levenshtein Similarity Score.

Table 2: Performance Comparison Between Models.

[2] https://huggingface.co/sentence-transformers/paraphrase-multilingual-MiniLM-L12-v2

### Sentence Similarity

The results in Table 2 show that our models achieved higher similarity to the original lines compared to the **ByT5-R-Zero** model. However, the **ByT5-R-Zero** model, which does not account for beat patterns, attained a higher PINC score of 45.6. In contrast, our models generated lines with lower but still moderate PINC scores, indicating higher lexical similarity to the original lines. It seems that our models tend to use words similar to those in the original lines, as compared to the base model with higher PINC score and lower similarity.

Upon reviewing the distribution of Levenshtein similarity scores in PARAPOETRY, we observed a roughly normal distribution centered around 0.8. This suggests that, although there is moderate lexical dissimilarity, the model likely attends to target beat patterns that are very similar to those of the original lines, resulting in rephrased lines with relatively higher similarity scores and lower PINC scores compared to the baseline model.

### Alignment Scores

Our models demonstrated a high level of beat alignment accuracy. The **ByT5-BR** model achieved the highest exact alignment accuracy, exceeding 94%. The **ByT5-R** model also performed well, reaching approximately 91%, though slightly lower than **ByT5-BR**, highlighting the effectiveness of curriculum learning in this context. The Levenshtein similarity scores with both models achieving over 99%, indicates strong rhythmic alignment with only minor deviations. In contrast, the **ByT5-R-Zero** model, which did not attend to the beat patterns, exhibited extremely low exact alignment accuracy of around 7% and a significantly lower Levenshtein similarity score of 84.71%. These results emphasize that proper training is essential for achieving effective beat alignment. These findings confirm that our fine-tuned models can generate verses that align with beat patterns while maintaining similarity to the original lines.

## Human-Centric Evaluation

Although automated natural language processing metrics can provide useful insights, they fall short of capturing the subjective human experience of poetic language and creative qualities. In particular, creative aspects are difficult to quantify using standard metrics. To address this limitation, we designed a survey to collect human judgments on rhythmic alignment and creative qualities. Our evaluation study was organized into three main components:

- **Word-Beat Alignment Experiment:** Participants were presented with a pairwise comparison task, where they were asked to fill in the blank with the word or phrase that best matches a given beat pattern as well as indicating their preference based on fluency and sense in completing the lyrical line.
- **Line-Beat Alignment Experiment:** In a similar pairwise comparison setup, participants were asked to select the line that best matches a given beat pattern as well as indicating their preference based on fluency and sense.
- **Qualitative Evaluation:** Since evaluating creative qualities of a poetic text may require more context than only individual words or lines, participants were asked to rate complete verses generated by our model on multiple aspects, including fluency, sense, and overall poetic quality.

The first two experiments evaluate the ability of the **ByT5-B** and **ByT5-BR** models to align words and rephrase lines according to specific audio-based beat patterns. It also serves as evidence supporting the concept of perceptual beat locations in English text. The third experiment evaluates the **ByT5-BR** model's capability to generate complete quality verses while adhering to the beat alignment task.

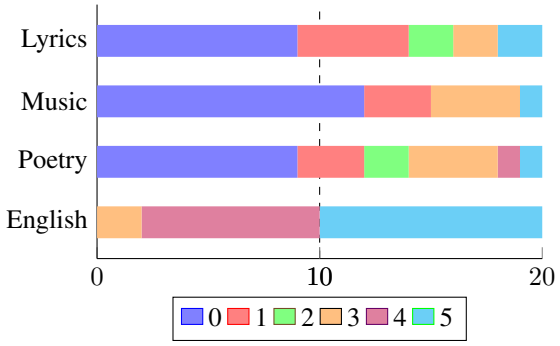


Figure 3: Demographics of the Participants.

### Participants Demographics

Twenty participants were recruited for this study via convenience sampling, drawing on graduate students at the University of Calgary from the Department of Computer Science and acquaintances with a background in poetry or lyrics composition. Each participant self-assessed their English proficiency on a 0 (no understanding) to 5 (native speaker) scale and their familiarity with writing poetry, lyrics, and music composition on a 0 (no experience) to 5 (published-level expertise) scale. Although respondents who selected "0" for English proficiency would have been removed, no exclusions were necessary. Sessions were conducted online through Qualtrics[3] (survey links were sent by email) and took approximately 45-60 minutes to complete. Figure 3 illustrates our sample, which comprised an approximately equal mix of native and non-native English speakers as well as individuals with and without prior experience in poetry, lyrics writing, or music composition.

The participants were informed that the study aimed to investigate the impact of large language models on learning beat patterns and maintaining textual coherence. However, they were not told that some of the lines were human-written while others were generated by a model. To address participant fatigue effect, the survey was structured with simpler questions presented first, progressing to the more complex ones. Additionally, a $20 CAD gift card was promised as an incentive for participation. While our sample of twenty is

[3] https://www.qualtrics.com/

| Experiment | Category | Alignment Experiments | | Fluency & Sense Experiments | |
|---|---|---|---|---|---|
| | | # Chosen | % | # Chosen | % |
| Word-Beat | ByT5-B | 129 | 64.5 | 54 | 27 |
| | Original | 71 | 35.5 | 106 | 53 |
| | Both | - | - | 40 | 20 |
| Line-Beat | ByT5-BR | 131 | 65.50 | 42 | 21 |
| | Original | 69 | 34.50 | 98 | 49 |
| | Both | - | - | 60 | 30 |

Table 3: Experimental results for word-beat and line-beat alignment evaluations.

relatively small, and there is no universally agreed-upon optimal size for this type of exploratory study, we nevertheless evaluated our results for statistical significance.

Note that two participants did not complete the qualitative evaluation which came last in the survey, resulting in 18 responses for that portion of the study.

## Word-Beat Alignment Experiment

In this experiment, we randomly selected 20 lines from the evaluation set. We first performed a manual grapheme-to-beat transformation, identifying the rhythmic beat pattern present in each of the selected lines placing a beat at each vowel onset and a rest elsewhere. For each line, we masked between 3 and 6 syllables from various positions in a line (i.e., beginning, middle, and end) to ensure a balanced sample and sufficient contextual information for the model. The syllables can be one or more complete words. The **ByT5-B** model was then prompted to generate candidate replacements. The required beat pattern for the replacement was deliberately chosen to differ from the original words' beat pattern while matching their syllable count. In this way, we ensure that participants are not simply matching syllable-beat counts.

Each participant was presented with a pairwise fill-in-the-blank task in which the original words and the generated words are both presented. Participants can listen to the provided beat pattern in audio format (a clave hit for each 1 in the beat sequence and a pause for each 0) and were asked to choose the words that best align with it. In addition to assessing rhythmic alignment, participants were asked to select the words that best completes the lyrical line in terms of overall fluency and sense or if they think both words fit the context equally. For this study, we defined *fluency* for the participants being the smoothness, flow, and ease with which the verses can be read or spoken and *sense* being the logical coherence and clarity of the lines where lines that make more sense are those that convey a clear and understandable meaning or narrative. The definitions were presented to the participants every time they were asked to evaluate the lines. To manage the survey duration, each participant evaluates a random subset of 10 out of the 20 examples.

The results of the word-beat alignment evaluation (see Table 3) indicate that the generated words were chosen in 64.5% of cases, compared to 35.5% for the original version. This indicates that human participants found the generated words to align better than the original words. Conversely, the fluency and sense evaluation showed a slight preference for the original version at about 53% over the other choices. Although the generated words solely were preferred only 27% of the cases, in 47% they were preferred at least as much as the original words with no significant statistical difference on a paired two-tailed t-test with a $p$-value = 0.63.

## Line-Beat Alignment Experiment

Following the word-beat alignment task, participants were introduced to the line-beat alignment task. For this experiment, we randomly selected 20 lines from the evaluation set and rephrased each one of them using the proposed **ByT5-BR** model trained on the rephrasing objective. For every line, we manually identified the beat patterns and then assigned a new different beat pattern that preserves the original number of beat count. This ensured that both the original and rephrased lines contained the same number of syllables, thereby preventing syllable count from influencing the evaluation.

Each participant was presented with a random subset of 10 out of the 20 examples. The participants were asked to perform a pairwise comparison between the original and the rephrased lines, selecting the one that best aligned with the beat pattern presented in audio format. Additionally, participants were asked to rate which line has more fluency and sense, or if they believed both lines were equally acceptable.

The line-beat alignment results indicated that the generated lines were chosen in 65.5% of the cases, while the original lines were chosen in 34.5% of the cases. This suggests that participants generally perceived the generated lines as better aligning with the given beat pattern. In contrast, the evaluation of fluency and sense showed a preference for the original lines at approximately 49%, with the generated lines being preferred solely in 21% of cases. However, in 51% of the cases, the generated lines were considered at least as acceptable as the originals, with no significant statistical difference with a $p$-value = 0.88.

We also investigated whether the order of tasks influenced participant responses by examining the practice effect. Comparing responses from the easier word-beat alignment task and the more challenging line-beat alignment task revealed no significant statistical difference with one-tailed t-test with $p$-value = 0.44.

Moreover, we manually assessed the alignment accuracy

| Aspect | Description |
|---|---|
| Fluency | The smoothness, flow, and ease with which the verses can be read or spoken. |
| Sense | The logical coherence and clarity of the lyrics. Lines that make more sense are those that convey a clear and understandable meaning or narrative. |
| Meaningfulness | Being rich in detail, evoking strong mental pictures. |
| Unpredictability | How unexpected or surprising the combination of words in the lyrics or lines are. |
| Originality | In the use of words to you from the perspective of the participant. |
| Humanness | Verses that are more likely to be written by a Human have more hummanness. |
| Poeticness | The qualities that make the lines artistic, expressive, and emotionally resonant. Poetic lines often use poetic devices like metaphor and similes, etc. |

Table 4: Definitions of the qualitative evaluation metrics.

in the first and second experiments. In the word-beat experiment, one out of 20 phrases generated by **ByT5-B** did not align perfectly with the beat pattern. In the line-beat experiment, two of the generated examples by **ByT5-BR** did not align perfectly with the beat pattern. These three instances were retained in the evaluation and received 24 evaluations, the generated semi-matching patterns were chosen 8 times out of 24, suggesting that counting syllables was probably preferred over partially matching sub-syllabic alignment.

## Qualitative Evaluation

In this part of the study, we assessed the linguistic and poetic qualities of our model's output. Our model is trained to rephrase single lines. Single lines by themselves may not contain enough information to be evaluated on a creativity scale. So in this experiment we wanted to test the utility of our rephrasing model to generate complete verses as compared to real poems written by human poets.

We began by randomly selecting 10 complete poems from amateur poets (normal users of the platform) and well-known famous poets scraped from the *allpoetry* platform.[4] To avoid any influence of rhyme schemes which our model does not handle, we removed poems that followed a distinct rhyming pattern. From the remaining poems, we extracted 10 verses from the famous poems and 10 from the amateur poems. These verses were then stripped of punctuation and line breaks before being processed by the **ByT5-BR** model, which is then used to generate rephrased versions of each verse. Punctuation and line breaks were then restored manually to the generated lines to ensure they were presented in a similar format as the original lines. Although this experiment does not focus on beat patterns, our model requires a beat pattern to generate the rephrased lines. For this purpose, we assigned a beat pattern using the Bjorklund algorithm (Bjorklund 2003) to distribute the sequence of 1's and 0's in the original beat as evenly as possible. This demonstrates a potential real-world use case for our model.

Participants were presented with both the original and the rephrased verses, and they were asked to rate several qualities of the generated verses using a slider scale from 0 (worst possible) to 100 (best possible). For clarity, the definitions for each evaluated quality were provided alongside every example (see Table 4 for the definitions we used).

Table 5 summarizes the differences in quality ratings between original and generated verses as evaluated by the participants. The differences are computed as the mean rating for the original verses minus the mean rating for the generated verses ($\overline{X}_o - \overline{X}_r$) on a scale from 0 to 100. The table further breaks down the results for all participants, native English speakers, and participants with poetry, lyrics, or music composition experience, along with the corresponding $p$-values to assess statistical significance.

**Fluency and Sense** *Fluency* was rated significantly higher for the original verses across all groups. All participants reported a mean difference of ($+9.41$, $p = 0.043$), while native speakers and participants with experience showed even larger differences ($+18.22$, $p = 0.010$ and $+20.62$, $p = 0.019$, respectively). Similarly, *Sense* yielded even more noticeable differences, with the original verses outperforming the generated ones by ($+15.98$, $p = 0.002$) for all participants, ($+28.18$, $p = 0.0004$) among native speakers, and ($+29.89$, $p = 0.002$) among those with experience. The results suggest that human-written verses are perceived as more fluent and coherent, especially by native participants who have a deeper familiarity with language and participants who were more exposed to poetic language and song lyrics.

**Meaningfulness, Unpredictability and Originality** *Meaningfulness* also favored the original verses, with a significant overall difference of ($+10.09$, $p = 0.027$). Although the differences for native speakers ($+12.66$, $p = 0.050$) and poetry-experienced participants ($+15.23$, $p = 0.057$) were slightly higher, they only approached conventional significance levels. In contrast, the *Unpredictability* exhibited a slight advantage for the generated verses, as evidenced by negative differences for all participants, though none of these differences reached statistical significance ($p$-values ranging from $0.159$ to $0.404$). This suggests that while our model's output may offer a degree of surprise, this effect is subtle and not strongly differ-

[4] https://allpoetry.com/

| | All Participants | | Native Speakers Only | | Experienced Only | |
|---|---|---|---|---|---|---|
| **Metric** | $(\overline{X}_o - \overline{X}_r)$ | $p$-value | $(\overline{X}_o - \overline{X}_r)$ | $p$-value | $(\overline{X}_o - \overline{X}_r)$ | $p$-value |
| Fluency | +9.41 | 0.043 | +18.22 | 0.010 | +20.62 | 0.019 |
| Sense | +15.98 | 0.002 | +28.18 | 0.0004 | +29.89 | 0.002 |
| Meaningful | +10.09 | 0.027 | +12.66 | 0.050 | +15.23 | 0.057 |
| Unpredictability | -3.87 | 0.159 | -2.71 | 0.299 | -1.57 | 0.404 |
| Originality | -9.47 | 0.013 | -7.58 | 0.078 | -3.77 | 0.142 |
| Humanness | +9.03 | 0.072 | +20.02 | 0.027 | +26.94 | 0.001 |
| Poeticness | +6.67 | 0.156 | +12.00 | 0.118 | +21.83 | 0.009 |

Table 5: Statistical comparison across participant groups.

entiated by the evaluators. *Originality* showed that the generated verses were rated as more original overall, with a statistically significant difference of ($-9.47$, $p = 0.013$). However, when the analysis was restricted to native speakers and those with poetry experience, the differences were reduced ($-7.58$ with $p = 0.078$ and $-3.77$ with $p = 0.142$, respectively). The non-significant difference in meaningfullness, unpredictability and originality for native speakers and experienced participants suggest that the preservation of surprising mental imagery and meaningful novelty is debatable among these groups.

**Humanness and Poeticness** In the case of *Humanness*, original verses again scored higher. This difference reached statistical significance for native speakers ($+20.02$, $p = 0.027$) and poetry-experienced participants ($+26.94$, $p = 0.001$), while the overall group difference ($+9.03$, $p = 0.072$) was marginal. The evaluation of *Poeticness* yielded mixed results. For the overall participant group, the difference of ($+6.67$, $p = 0.156$) suggests a slight advantage for the original verses, although this was not statistically significant. Among native speakers, the difference increased to ($+12.00$, $p = 0.118$), and for those with poetry experience, the original verses were rated significantly more poetic by ($+21.83$, $p = 0.009$). This shows that individuals with a background in poetry and lyrics were more sensitive to the these qualities and that the model's output was perceived as less poetic by these participants.

**Other Implications** The average quality scores of the generated verses ranged from approximately $45\%$ to around $66\%$, while those for the original verses ranged from about $53\%$ to around $76\%$ across the evaluated qualities. This suggest that while the generated verses scored significantly lower than the original verses on some qualities, they were still rated at least as acceptable by the participants on average and that the model is capable of generating verses that are at least comparable to human-written verses in some aspects while maintaining a high level of beat alignment. Finally, when comparing the rephrasing of poems written by amateur poets and those written by famous poets, the difference in quality between the generated and original poems was more noticeable by the participants for amateur poets, even among native speakers and experienced participants. It is possible that famous poems were more challenging for participants to evaluate.

## Conclusion

In this paper, we presented an approach to rephrase English lyric lines so that they align with specific beat patterns, utilizing the ByT5 transformer model. By introducing PARAPOETRY, a large-scale parallel dataset of poetry-rephrase pairs, we provided a valuable resource for researchers interested in poetry and lyrics analysis and generation. We demonstrated the utility of PARAPOETRY by fine-tuning pre-trained ByT5 models on a guided rephrasing task to reconstruct lyric lines with predetermined beat patterns. Our experimental results demonstrate that the ByT5 model can successfully generate beat-aligned lyrical poetry lines while maintaining high semantic similarity with moderate lexical diversity while demonstrating high rhythmic alignment.

The results of our human-centric evaluation provide several noteworthy insights regarding the rhythmic and creative qualities of the model-generated poetry. Overall, the experiments indicate that the proposed generative models trained on the substitution task (**ByT5-B**) and the rephrasing task (**ByT5-BR**) are capable of aligning words and lines with specific beat patterns. In both the word-beat and line-beat alignment experiments, participants selected the model-generated outputs approximately two-thirds of the time when asked to judge rhythmic alignment, which underscores the models' effectiveness in adhering to imposed rhythmic constraints. However, evaluations based on fluency and sense consistently demonstrated a slight preference for the original text. This finding is especially noticeable in the qualitative evaluation of the rephrasing of complete verses, where native speakers and participants with relevant experience rated the original verses significantly higher in terms of fluency, coherence, and overall poetic quality. These results suggest that, although the generated outputs achieve a high degree of rhythmic alignment, they still lack certain creative qualities that are found in human-written poetry.

## Limitations and Future Work

Despite the progress made in aligning lyrical lines with rhythmic patterns, several limitations remain for further discussion. Although PARAPOETRY proved useful, the dataset

| Original Line | Beat Pattern | Rephrased Line |
|---|---|---|
| Feral landscapes bathed in twilight | 100010001011010 | Swamp lands in the sunset |
| Sure we might fish from out the mothers sons | 10100101010110100101100010101l0 | Genuinely from their mothers sons we'd managed to fish away |
| The depth and peak of a grief of this nature | 1100100100101001010010 | Precise and height of such a grief as this |
| Was not thy melody touching and choice | 10010010111100100100011100 | Was not your melody poignant and preferred |
| Then in the heart itself that knoweth | 1011011000100010100 | Dwell within the knowing heart itself |
| Come pity us all ye who see | 10101110010101010100 | Show mercy upon us all who observe |
| And scatter salt on the unmissed sepulcher | 1001010100110110100100 | And sow with salt the unremembered grave |
| Or what of tidings you abroad doo heare | 101010101010100100010 | Pray what whispers of tidings reach you |
| By Fan and Sword and Office box | 10101001001101000 | By Fan an Sword and Office box |
| Stay till the storms are o'er | 1000100101011000100 | Remain close until the gales cease |

Table 6: Examples of original lines and their rephrased versions according to specified beat patterns, showcasing the English rephrase task. The examples are selected from the evaluation set.

was generated using automatic methods with GPT-3.5 and may contain inconsistencies or biases inherent to the automated process. Our current approach determines beat placement primarily based on vowel onsets, thereby neglecting the impact of syllabic stress on beat perception. Additionally, the human evaluation component, while insightful, was conducted with a relatively small participant pool who are largely not experts in the poetry, lyrics and music fields and with limited set of examples. Further, the model was trained to generate single lines even though the qualitative evaluation involved full verses.

The qualitative evaluation highlights a mixed performance in terms of creativity. While human-written verses scored higher in aspects such as fluency, sense, and poeticness, the generated verses were rated as more original and unpredictable, indicating that the models may introduce original word combinations and unexpected expressions. However, this by itself does not guarantee that these combinations are creative as striking a balance between adherence to other creative aspects while generating content with meaningful novelty is necessary.

Future research should consider integrating features that consider the effects of syllabic stress. Incorporating a larger proportion of human-curated data could also enhance the quality of the poetic expressions generated. Expanding the evaluation to include a more diverse participant pool especially experts in the field. Finally, extending the approach to generate full verses or entire songs may introduce additional challenges in maintaining coherence and poeticness as well as overall musical alignment over longer texts.

## Supplementary Material

The source code and the prompts used are available at: `https://github.com/melzohbi/poem-rhythm-para`. PARAPOETRY is available at: `https://doi.org/10.5683/SP3/WMU1BC`. Table 6 provides examples of original lines and their rephrased versions by the proposed rephrasing model.

## Ethical Considerations

All experiments involving human participants in this paper received ethics approval from the Conjoint Faculties Research Ethics Board (CFREB) at the University of Calgary under protocol number REB24-0877. Participation was voluntary, and all participants were provided with a consent form informing them of their right to withdraw from the study at any time or decline participation. Data collection was conducted confidentially and anonymized by removing any identifiable information that could reveal the participants' identities.

## References


Allen, G. D. 1972. The location of rhythmic stress beats in english: an experimental study i. *Language and Speech* 15(1):72–100. PMID: 5073939.

Bjorklund, E. 2003. The theory of rep-rate pattern generation in the sns timing system. *SNS ASD Tech Note, SNS-NOTE-CNTRL-99, Los Alamos National Laboratory, Los Alamos, U.S.A.*

Chen, D., and Dolan, W. B. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, 190–200.

Chen, Y., and Teufel, S. 2024. Scansion-based lyrics generation. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 14370–14381.

Elzohbi, M., and Zhao, R. 2023. Creative data generation: A review focusing on text and poetry. In *Proceedings of the 14th International Conference on Computational Creativity (ICCC'23)*, 29–38.

Elzohbi, M., and Zhao, R. 2024. Let the poem hit the rhythm: Using a byte-based transformer for beat-aligned

poetry generation. In *Proceedings of the 15th International Conference on Computational Creativity, (ICCC'24)*, 407–411.

Greer, T.; Singla, K.; Ma, B.; and Narayanan, S. 2019. Learning shared vector representations of lyrics and chords in music. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 3951–3955. IEEE.

Martinez-Sevilla, J. C.; Rios-Vila, A.; Castellanos, F. J.; and Calvo-Zaragoza, J. 2023. A holistic approach for aligned music and lyrics transcription. In *International Conference on Document Analysis and Recognition*, 185–201. Springer.

Oliveira, H. R. G.; Cardoso, F. A.; and Pereira, F. C. 2007. Tra-la-lyrics: An approach to generate text based on rhythm. In *Proceedings of the 4th. International Joint Workshop on Computational Creativity*, 47–55. London, UK: A. Cardoso and G. Wiggins.

Rathcke, T.; Lin, C.-y.; Falk, S.; and Dalla Bella, S. 2021. Tapping into linguistic rhythm. *Laboratory Phonology* 12(1):1–32.

Reimers, N., and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, 3982–3992. Association for Computational Linguistics.

Sheng, Z.; Song, K.; Tan, X.; Ren, Y.; Ye, W.; Zhang, S.; and Qin, T. 2021. Songmass: Automatic song writing with pre-training and alignment constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 13798–13805.

Soviany, P.; Ionescu, R. T.; Rota, P.; and Sebe, N. 2022. Curriculum learning: A survey. *International Journal of Computer Vision* 130(6):1526–1565.

Sulun, S.; Oliveira, P.; and Viana, P. 2023. Emotion4midi: A lyrics-based emotion-labeled symbolic music dataset. In *EPIA Conference on Artificial Intelligence*, 77–89. Springer.

Vaswani, A.; Shazeer, N.; Parmar, N.; Uszkoreit, J.; Jones, L.; Gomez, A. N.; Kaiser, Ł.; and Polosukhin, I. 2017. Attention is all you need. *Advances in neural information processing systems* 30.

Xue, L.; Song, K.; Wu, D.; Tan, X.; Zhang, N. L.; Qin, T.; Zhang, W.-Q.; and Liu, T.-Y. 2021. Deeprapper: Neural rap generation with rhyme and rhythm modeling. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, 69–81.

Xue, L.; Barua, A.; Constant, N.; Al-Rfou, R.; Narang, S.; Kale, M.; Roberts, A.; and Raffel, C. 2022. Byt5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics* 10:291–306.