Song Lyrics Adaptations: Computational Interpretation of the Pentathlon Principle

Barbora Štěpánková and **Rudolf Rosa**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
barbora.stepankova320@student.cuni.cz, rosa@ufal.mff.cuni.cz

Abstract

Songs are an integral part of human culture, and they often resonate the most when we can sing them in our native language. However, translating song lyrics presents a unique challenge: maintaining singability, naturalness, and semantic fidelity. In this work, we computationally interpret Low's Pentathlon Principle of singable translations to be able to properly measure the quality of adapted lyrics, breaking it down into five measurable metrics that reflect the key aspects of singable translations. Building on this foundation, we introduce a text-to-text song lyrics translation system based on generative large language models, designed to meet the Pentathlon Principle's criteria, without relying on melodies or bilingual training data.

We experiment on the English-Czech language pair: we collect a dataset of English-to-Czech bilingual song lyrics and identify the desirable values of the five Pentathlon Principle metrics based on the values achieved by human translators. Through detailed human assessment of automatically generated lyric translations, we confirm the appropriateness of the proposed metrics as well as the general validity of the Pentathlon Principle, with some insights into the variation in people's individual preferences. All code and data are available at https://github.com/stepankovab/Computational-Interpretation-of-the-Pentathlon-Principle.

1 Introduction

Songs are a prominent part of human culture, everywhere in the world. Since the old days, people have been singing folk songs, and adapting them to different situations. One of these adaptations is translation. Rewriting a song's lyrics into another language while keeping the song singable, naturally sounding and semantically close to the original is a very complex task without a straightforward definition. Franzon (2008) defined five levels of song adaptations, ranging from leaving the song as it is, to making completely new lyrics with zero connection to the original meaning. In this paper, we are going to focus on song lyrics adaptations, while keeping both the singable aspect, as well as the semantic aspect.

There have been many attempts to formalise what makes a good song translation. Low (2003, 2005) proposed a set of rules, called the Pentathlon Principle of Singable Translations. These guidelines are still accepted by song translators today (Sardiña, 2021; Pidhrushna, 2021; Saragih and Natsir, 2023). Kim et al. (2023) proposed metrics for computationally evaluating song translation quality for Japanese and Korean. However, to the best of our knowledge, we are the first to try to computationally interpret and verify the Pentathlon Principle as a whole instead of using it as a given thing.

In this work, we computationally interpret Low (2003) in terms of collecting and proposing metrics for measuring the song translation quality. We experiment on the Czech-English language pair: we collect a dataset of bilingual song lyrics and evaluate the official human-translated songs by these metrics, finding the desirable values of the metrics.

As mentioned above, song lyrics translation is a difficult and complex task even for human translators. In recent years, many works try to simplify and automatize this process by using computational methods, to make translated songs more accessible. The first step of creating a singable adaptation is generating text to a given melody. Many studies in generating song lyrics used datasets of melody-lyrics pairs (Watanabe et al., 2018; Sheng et al., 2021; Zhang et al., 2024b). Recently, Chen and Teufel (2024) used scansion as an intermediate step between melody and lyrics and generated Chinese texts. Tian et al. (2023) generated lyrics to a melody without needing melody-lyrics aligned data for training. Studies on automatic song translations were done mainly on Chinese: Guo et al. (2022) fo-

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 117–128

cused on translating lyrics for tonal languages, and Ou et al. (2023) used prompted machine translation with melody-based word boundaries for Chinese lyrics translation.

In this work, we propose an approach which explores text-to-text song lyric translation without the need for melody-aligned or bilingual training data, using generative large language models (LLMs). We evaluate various setups of our system using the Pentathlon Principle metrics, comparing the setups to the human-translated song lyrics, and through a thorough human evaluation conclude the importance of individual aspects of the Pentathlon Principle, and their balance.

2 Pentathlon Principle Metrics

The Pentathlon Principle, as defined by Low (2003), consists of five aspects of lyrics. It states that all these aspects should be balanced, the same as an athlete competing in a pentathlon has to have balanced skills in all five activities to be successful. The five aspects of singable translations are Singability, Sense, Naturalness, Rhyme and Rhythm. In this Section, we discuss each aspect of the Pentathlon Principle from the computational point of view. We present five metrics, each measuring one aspect of the Pentathlon Principle.[1]

First, let us introduce the notation used for the metric descriptions. All proposed metrics are section-wise, giving scores for each section (e.g. verse or chorus) separately. The Pentathlon Principle was proposed in the context of singable translations, so most of the metrics have the source-language lyrics and the targe-language lyrics as inputs. We denote the source-language lyrics section consisting of n lines as $X = \{x_1, \ldots, x_n\}$ and the translated target-language lyrics section as $\tilde{X} = \{\tilde{x}_1, \ldots, \tilde{x}_n\}$.

2.1 Singability

Low (2003) describes singability as the comfort of the lyrics being sung to a certain melody. While singability is closely tied with stress patterns and line lengths, Low addresses these under the term *Rhythm*, as do we. Singability encompasses the adequateness of certain syllables being placed at certain parts of the song. Low (2003) emphasises consonant clusters and vowel openness as key parts of singability, explaining that large consonant clusters and tight syllables are awkaward to sing if the melody is not adapted to it.

Pressure that'll tip										**CCVO distance = 0.06**
N	VO	N	VO	N	OO	N	VO	N	VV	N
pɹ	ɛ	ʃ	ə	ɹ ð	æ	t	ə	l t	ɪ	p
Dál to na mě syp *(Keep throwing it at me)*										
N	OO	N	VO	N	OO	N	VO	N	VV	N
d	aː	l t	ɔ	n	a	mɲ	ɛ	s	ɪ	p
Change the fates' design										**CCVO distance = 0.56**
N	VO	C	VO	N	VO	C	VV	N	OO	N
ʧ	eɪ	nʤð	ə	f	eɪ	ts d	ɪ	z	aɪ	n
Osud převracej *(Overturn fate)*										
N	VO	V	VV	C	VO	N	OO	N	VO	N
-	ɔ	s	u	t pɾ̥	ɛ	vɾ	a	ts	ɛ	j

Table 1: Example of two lines with $CCVO_{dist} = 0.06$ signifying high mutual singability and of two lines with $CCVO_{dist} = 0.56$ signifying low mutual singability, even though the number of syllables of the compared lines is the same.

Proposed method We propose the *Consonant Cluster and Vowel Openness Distance (CCVO Dist)* metric. We define a consonant cluster as three or more consecutive consonants in the phonetic transcription of the line. We determine vowel openness from the IPA chart[2]. For each pair of lyrics x_i and $\tilde{x}_i$, we extract the *CCVO* (see Table 1): a string marking whether there is a consonant cluster between vowels of adjacent syllables (*C* for a cluster, *N* for no cluster) and the openness of the most open vowel from the syllable (*OO* for open, *VO* for mid and *VV* for a closed vowel). The Levenshtein distance is then computed between these two *CCVO*s and divided by the length of $CCVO(x_i)$, representing the original line.

$$CCVO_{Dist}(X, \tilde{X}) = \frac{1}{n} \sum_{i=1}^{n} \frac{Lev_{Dist}(CCVO(x_i), CCVO(\tilde{x}_i))}{len(CCVO(x_i))} \quad (1)$$

2.2 Sense

Sense is defined as the similarity in meaning, but as Franzon (2008) emphasizes, there are different levels of song translations, and one should not prioritise meaning over other aspects of the pentathlon if the final adaptation should be singable.

Preliminary experiments Preliminary experiments with BLEU score (Papineni et al., 2002)

[1] Our implementation of the Pentathlon Principle metrics is at https://github.com/stepankovab/Computational-Interpretation-of-the-Pentathlon-Principle

[2] https://en.wikipedia.org/wiki/IPA_vowel_chart_with_audio [Accessed 2025-02-14]

showed that even human-translated lyrics reach a near zero BLEU-2 score[3]. This might be because the translator usually can not choose the most straightforward way of translating the lyrics due to the melody constraints. Even though BLEU is used in measuring song translation quality (Ou et al., 2023; Guo et al., 2022), oftentimes BLEU decreases while meaning-unrelated metrics improve.

Proposed method To have more freedom in reformulating the same thought in different words, we adopted the *Semantic Similarity* metric from Kim et al. (2023). The metric measures the similarity of individual song sections X and $\tilde{X}$ based on the cosine similarity of text embedding vectors obtained using a pre-trained Sentence BERT model (Reimers and Gurevych, 2019).

$$\mathit{SemantSim}(X, \tilde{X}) = \frac{\text{SBERT}(X) \cdot \text{SBERT}(\tilde{X})}{||\text{SBERT}(X)||||\text{SBERT}(\tilde{X})||} \quad (2)$$

2.3 Naturalness

According to Low (2003), naturalness *'involves considerations of features such as register and word order'*. To quantify this, we propose using the perplexity of a language model pre-trained on the target language, measured on the $\tilde{X}$ section.

Perplexity reflects how well a sequence aligns with common linguistic patterns, with lower values indicating more natural phrasing. Since a well-trained model captures typical syntax and idiomatic usage, perplexity serves as a reasonable proxy for naturalness: high perplexity suggests unnatural word order or phrasing, while low perplexity indicates fluency.

$$\mathit{Naturalness} = \text{PPL}_{\text{LM}}(\tilde{X}) \quad (3)$$

2.4 Rhyme

Low (2003) notes that fewer or differently placed rhymes are often better than forcing a rhyme scheme at the expense of other Pentathlon Principle aspects.

Preliminary experiments We experimented with metrics based on recall of rhymes rather than accuracy. When considering accuracy, the translation is penalized more for changing the rhyme scheme than for not rhyming at all. It is also penalized for introducing new rhymes, thus making the translation more artistic. The flip side shows that recall oriented metrics prefer song sections with n same lines: when all the lines rhyme, the recall is perfect, which is not what we desired. In the end, we settled on using the Jaccard Index, as an average song section has an imbalance between rhyming pairs of lines and non-rhyming pairs of lines.

Proposed method Let the original rhyme scheme be a graph R and the new scheme a graph $\tilde{R}$, both with vertices $\{1, \ldots, n\}$, representing the indices of lines in the song sections X and $\tilde{X}$ respectively. An edge between nodes i and j in R means lines x_i and x_j rhyme. Function $Edges(R)$ returns the set of (i, j) tuples where the i and j correspond to the indices of rhyming lines in X.

The Rhyme Scheme Jaccard Index is computed as follows, effectively computing the number of common edges divided by the number of all edges.

$$RS_{JI}(R, \tilde{R}) = \frac{|Edges(R) \cap Edges(\tilde{R})|}{|Edges(R) \cup Edges(\tilde{R})|} \quad (4)$$

2.5 Rhythm

The main aspect of rhythm is whether the lyrics can fit the melody. The key focus when measuring rhythm computationally usually lies in syllable counts (Guo et al., 2022; Ou et al., 2023), and almost never in stress patterns.

Preliminary experiments We conducted preliminary experiments measuring stress pattern distance, similarly to how we measure CCVO distance in Section 2.1. The results were partially promising, but we have not managed to devise a metric that would capture all of the important rhythmic aspects. We leave a better stress pattern distance metric as a future work, and focus on the more wide-spread syllable count based metrics. We experimented with syllable accuracy as used by Guo et al. (2022); Ou et al. (2023), however we found it too strict. When a 3-syllable line translates to a 10 syllable line, it is much worse than when an 11-syllable line is missing one syllable.

Proposed method We use the *Syllable distance* from Kim et al. (2023). With *syl* as a syllable counter function, syllable distance can be computed as:

$$SylDist(X, \tilde{X}) = \frac{1}{2n} \sum_{i=1}^{n} \left(\frac{|syl(x_i) - syl(\tilde{x}_i)|}{syl(x_i)} + \frac{|syl(x_i) - syl(\tilde{x}_i)|}{syl(\tilde{x}_i)} \right) \quad (5)$$

[3]Measured on En-Cs parallel data introduced in Section 4.1.1

3 Lyrics Generation System

While previous approaches to song lyrics adaptation were through machine translation, song lyric adaptation using generative LLMs is underexplored. In this Section, we propose a text-to-text song lyrics generation system based on the Pentathlon Principle. This system can be trained using only the target language data, and when provided with lyrics in a source language, it produces a singable adaptation of the lyrics in a target language.

Our pipeline (see Figure 1) has several steps, each described in more detail in the following Subsections. The pipeline input is the lyrics of a song section in the source language, divided into individual lines. The output is lyrics in the target language that are singable to the same melody as the input lyrics, while also retaining similar meaning, naturalness, rhythm and rhyme.

First, defining features of the source lyrics are extracted (Section 3.1). Then, a prompt for an LLM is built based on the extracted features of the source lyrics (Section 3.2). Based on this prompt, a fine-tuned LLM generates the lyrics in the target language. The training process of the model is described in Section 3.3 and the inference process is described in Section 3.4. Finally, the generated lyrics are post-processed (Section 3.5).

3.1 Feature Extraction

The first step of our pipeline is the extraction of relevant features from the input song section. During inference, the input section is in the source language and during the training phase, this section is in the target language. Therefore, we need to be able to do feature extraction in both the source and the target languages.

We are extracting three things: syllable counts for rhythm and singability, rhyme scheme for rhyme, and the maximum of five keywords for sense.

3.2 Prompt Format

In this Subsection, we describe the various formats of the LLM prompt created from the extracted features. We tried two main approaches: first, generating the whole lyrics section at once and second, generating each line separately. We also experimented with which of the extracted features to include in the prompt. For examples of prompts, see Table 2.

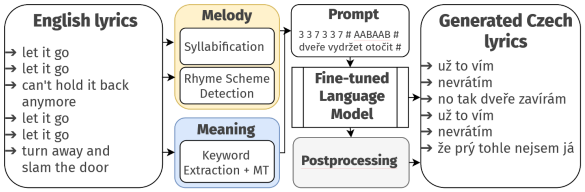


Figure 1: Inference pipeline visualisation. The generated Czech lyrics as translated by DeepL: *I already know, I'm not coming back, so I'm closing the door. I already know, I'm not coming back, they say this isn't me.*

syllables, rhyme scheme and keywords (Sections)
3 3 7 # AAA # najednou, hrou, skončit # *3 # A #* Najednou *3 # A #* Najednou *7 # A #* Chci skončit s tou hloupou hrou
syllables, endings and keywords (Lines)
7 # ou # hrou, skončit # Chci skončit s tou hloupou hrou

Table 2: Training examples for fine-tuning LLMs to generate song lyrics. The first example is for generating whole sections at once, the second one for generating an individual line with the E model (for the S model, the ending parameter is missing in the format).

3.2.1 Prompt for Generating Sections

The prompt has two parts: the first line containing all relevant information, and the annotated lines of the song section lyrics.

The first line of the prompt contains syllable counts for each line, the rhyme scheme and keywords of the section, all separated by the # separator. This first line is the prompt during inference. To enforce dependencies of lines on syllable counts and the rhyme scheme during training, the corresponding syllable count and letter of the rhyme scheme are added at the beginning of each line of the song section as an annotation. The prompt format is inspired by Chudoba and Rosa (2024).

3.2.2 Prompt for Generating Lines

When generating each line individually, the line is generated as a continuation of the prompt without a new line. There is the syllable count the same as when generating a line in a section. Instead of a letter of the rhyme scheme, there is the desired line ending. Then there are the line keywords.

3.3 Model Finetuning

As mentioned in Section 3.1, during training the prompts are built from features extracted from tar-

get song sections. These target song sections are then showed as the 'correct answer', teaching the model to predict the target sections based on the features extracted from them.

When generating full sections at once, we fine-tuned one LLM using the prompt for sections followed by the annotated lyrics.

When generating lines individually, we fine-tuned two models, S and E, that take turns during inference. The S model generates a line without a pre-specified ending, while the E model generates a line to rhyme with an already generated line and thus has the desired ending specified in its prompt.

3.4 Inference

During inference, the information needed for the prompt creation is extracted from the source language lyrics. The prompt is created and then based on that, target lyrics are generated. When generating whole sections, the prompt consists only of the first line, relying on the model to know which annotations belong to which line.

Multiple outputs are sampled and ranked according to each of the Pentathlon Principle metrics; we choose the one with the lowest sum of the ranks.

3.5 Lyrics Post-Processing

As postprocessing, we correct the lengths of the section lines where needed by removing or adding stopwords in appropriate places. For removing, we remove only words from a 'stopwords list' of the target language, and for adding, we suggest making a list of neutral phrases in the target language from one to three syllables, such as 'Then', 'So' or 'And', which can be easily inserted into the line. For both postprocessing techniques, we are minimising the syllable and CCVO distance while keeping the rhyme intact and the naturalness score of the section either the same or better, which ensures that no unnatural insertion or deletion is made.

4 Experimental Setup for EN→CS

We tested everything on an English-Czech language pair, in the direction of English to Czech. We describe the EN→CS data in Section 4.1, the implementation details of the Pentathlon Principle metrics and the Feature Extraction function specific for Czech and English in Section 4.2, and in Section 4.3 we discuss the LLM selection, training and inference.

Musical name	# Songs	# Sections
Frozen	8	65
Frozen 2	8	62
Moana	8	64
Encanto	6	110
Tangled	7	53
The Jungle Book	3	24
The Lion King	6	44
The Little Mermaid	5	45
Grease	1	6
Les Miserables	17	176
	69	649

Table 3: English-Czech aligned dataset distribution. The first part shows Disney songs and the second part shows songs from other musicals.

4.1 Data

In this Section, we will describe the data used for both evaluating the metrics and training the lyric-generating model.

4.1.1 Parallel Data

We collected 69 official English song lyrics and their Czech translations made for commercial musical films translated by professionals. The final dataset consists of 649 parallel song sections, where a song section is usually a single verse or a chorus, or, for example, a four-liner of a rap part. After splitting the songs into song sections, we cleaned them of metadata and meticulously mapped them onto each other by hand line by line to ensure correctness. In Table 3, we present a closer analysis of the dataset.

4.1.2 Monolingual Data

Our training dataset consists of 77478 Czech song sections obtained from the *Velký zpěvník* (translates to *The Great Songbook*) webpage[4]. The web contains 17599 mainly Czech songs from 1381 interprets, both recent and from the previous century.

We split the scraped data into sections and filtered out those not in Czech. A comparison with the parallel data can be seen in Table 4.

4.2 Pentathlon Principle Metrics and Feature Extraction Implementation

There are multiple language-specific functions throughout the Pentathlon Principle metrics and the Feature Extraction function in the lyric-generating system. In this Section, we describe which tools we used for Czech and English.

[4] www.velkyzpevnik.cz [Online Accessed 2024-02-02]

	Parallel Data	Czech Data
# Sections	649	77478
Avg lines per section	4.7	5.2
Avg line length	6.88 syll.	7.58 syll.
Most common themes	life	love
	sea	night
	day	sleep
	night	morning
	world	life
	dream	singing
	love	wind
	wind	sun
	time	world

Table 4: Comparison of the Parallel and the monolingual Czech dataset. The most common themes are obtained by counting the most frequent keywords.

For singability and rhythm, syllabification of the text is needed. First, we transcribe the text into IPA[5]. Then we use rule-based syllabification of the IPA inspired by a Czech syllabification script[6]. The complete list of our syllabification rules can be found in our GitHub repository. We made our syllabification function instead of using a premade one to have control over the output, as well as have the output in IPA directly.

For sense, we first translate the Czech sections into English, then obtain the sentence embeddings by all-MiniLM-L6-v2 (Wang et al., 2020). For naturalness, we chose to measure the perplexity by CsMPT7B (Fajčík et al., 2024), a Czech version of MPT7b (MosaicML, 2023), rather than a multilingual model, as our concern is the naturalness of the text in the target language, not the overall commonness of the text. For a rhyme scheme extraction, RhymeTagger (Plecháč, 2018) is used for both Czech and English. On top of that, we also accept identical rhymes, as song lyrics often use repetition to emphasise both meaning and rhythm. For keyword extraction, we used KeyBERT (Khan et al., 2022).

4.3 LLM Selection, Fine-Tuning and Inference Parameters

We chose TinyLlama pre-trained on large amounts of Czech text, CSTinyLlama-1.2B (Fajčík et al., 2024), as a base model. We also experiment with TinyLlama (Zhang et al., 2024a), which has 1.1 billion parameters and is not Czech-specific, and with a GPT2-small pre-trained on Czech (Chaloupský, 2022) which has only 137M parameters. For evaluation of these models, see Appendix A.

For fine-tuning the models, we used a batch size of 64, a learning rate of 5×10^{-4} and trained the model for one epoch, as there was no change of the loss function when continuing training.

For inference, we generate using sampling, with the top_p of 0.9, temperature of 0.8 and repetition penalty of 1 as lyrics often repeat. We tried randomly sampling 1 to 50 outputs, ranking them in each aspect of the pentathlon principle as described in Section 3.4 and outputting the one with the lowest sum of ranks. There was an improvement in both the metrics and the subjective quality of the output lyrics with more returned samples to choose from. As a compromise between quality and speed, we proceeded with 10 samples. A small experiment on 30 inputs showed that the ranking selects the 1st or 2nd best output according to human evaluation.

5 Evaluation and Discussion

In this Section, we evaluate our experimental setup on the test part of the parallel dataset introduced in Section 4.1.1. We use the English song lyrics as the source for all the following evaluations. As the target language song sections, we are using the official Czech translations from the parallel data in Section 5.2, machine translations (MT)[7] of the English lyrics into Czech in Section 5.3 and data generated by the Lyrics Generating System from Section 3 in Section 5.4. We also evaluate a random baseline in Section 5.1. All of the above-mentioned evaluations are automatic, using the Pentathlon Principle metrics. The results can be seen in Table 5. A visualization of the metric values distribution for individual setups can be seen in Figure 2.

In Section 5.5, we present a manual evaluation of the various Czech song lyrics adaptations, paired with statistics about human preference of individual Pentathlon Principle metrics and the dependencies of these preferences on choices in the evaluation.

5.1 Automatic Evaluation of Random Baseline

We create a baseline by randomly pairing up the English sections and the Czech official translations, truncating the longer of the pair, and evaluating these by the Pentathlon Principle metrics. We can see that the only well-performing metric is naturalness, as naturalness is measured independently of the source lyrics.

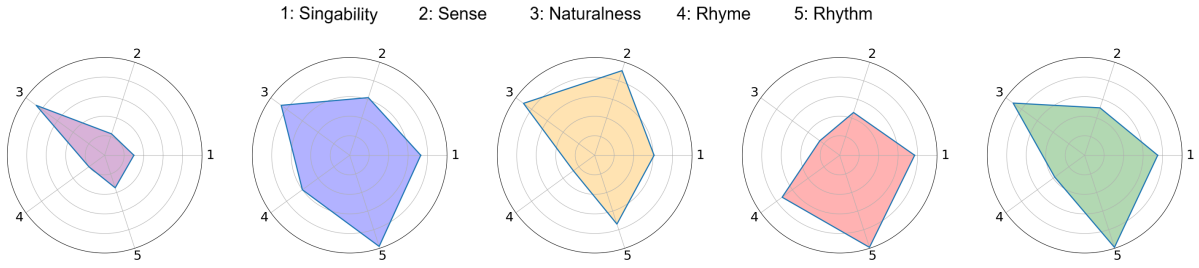
[5]For English https://pypi.org/project/eng-to-ipa/, for Czech https://github.com/lukyjanek/phonetic-transcription

[6]https://github.com/Gldkslfmsd/sekacek

[7]Translated using Lindat translator (Popel et al., 2020)

			Baseline	Official	MT	Lines	Sections
Singability	CCVO Distance	↘	0.70	0.27	0.39	**0.23**	0.25
Sense	Semantic Similarity	↗	0.23	0.62	**0.91**	0.46	0.51
Naturalness	Perplexity CsMPT	↘	131	131	97	748	**92**
Rhyme	Rhyme Scheme JI	↗	0.20	0.60	0.27	**0.73**	0.38
Rhythm	Syllable Distance	↘	0.65	0.02	0.26	**0.01**	**0.01**

Table 5: Random baseline, official translations of musical songs, MT of English part, and our proposed system generating by lines and sections, evaluated by the Pentathlon Principle metrics. For each metric, we show the direction depending on whether we are aiming for higher or lower values in that metric.



(a) Random Baseline (b) Official Translations (c) Machine Translation (d) Generated - Lines (e) Generated - Sections

Figure 2: Visualisation of the balance between individual normalised aspects of the Pentathlon Principle on all setups. *CCVO Distance* (Singability) and *Syllable Distance* (Rhythm) are normalised as '$1 - metricValue$', *Perplexity* (Naturalness) is normalised as $\frac{(1000 - metricValue)}{1000}$

		EN → CS	EN → KO	EN → JP
Semantic Similarity	↗	0.62	0.55	0.54
Syllable Distance	↘	0.02	0.11	0.17

Table 6: Comparison of EN→CS singable human translations (our) with EN→KO and EN→JP singable human translations (Kim et al., 2023).

5.2 Automatic Evaluation of Parallel Data

In this Subsection, we discuss the values of the Pentathlon Principle metrics reached by professional song lyrics translators. We hypothesise that these values are the optimal balanced distribution of the individual metrics. In table 5, we can see that all of the metrics except naturalness[8] increased significantly compared to the random baseline. The rhyme scheme Jaccard Index is quite low at 0.6, which shows that translators do not strictly stick with the original rhyme scheme. Also, sense is mediocre with only 0.62 semantic similarity, suggesting that translators change the meaning a bit to accommodate the text to the melody.

5.2.1 Comparison with Japanese and Korean

Two of the Pentathlon Principle metrics are adapted from Kim et al. (2023) who evaluated EN→JP and EN→KO human-translated singable lyrics. We compare our results measured on the EN→CS human translated singable dataset with theirs in Table 6. We can see that Czech reaches both better syllable distance and semantic similarity. This suggests that translating English lyrics into Japanese and Korean might be a more difficult task than translating into Czech.

5.3 Automatic Evaluation of MT

Next, we evaluate the machine translations. The MT outperforms both the random baseline and official translations in naturalness and sense. It is not surprising, as MT systems are crafted with these two goals in mind, while a human translator has to sacrifice both to abide by the constraints of the song. On the other hand, the system performed mediocrely in singability and rhythm and failed to retain the correct rhyme scheme.

5.4 Automatic Evaluation of Generated Data

Lastly, we evaluate the quality of the generated target-language adaptations. When generating each line separately, the outputs perform very poorly in the naturalness metric and mediocrely in sense. This might be because we generate the section a few words at a time. On the other hand, the generated outputs beat all other setups in singability,

[8]The sections of random baseline and official translations are the same, just shuffled.

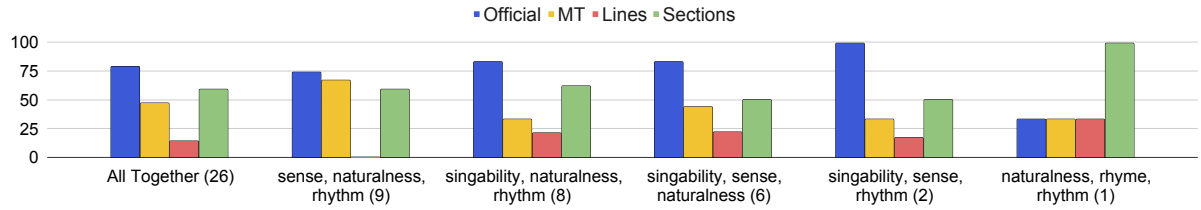


Figure 3: Percentages of times people chose a specific setup during manual evaluation. The first graph shows all participants. The following graphs show the preferences of groups divided based on what they consider the three most important aspects of the Pentathlon Principle. Number of people in each group is in brackets.

rhyme and rhythm.

The outputs generated as whole sections score very well in the rhythm and singability metrics. The naturalness of this setup is the best of all setups. Both the base model and the model used for measuring perplexity are pre-trained on Czech texts, so there is a possible training data overlap, which could make the perplexity (naturalness) biased. Both rhyme and sense are mediocre: rhyme outperforms the baseline and MT, and sense outperforms the baseline and 'Lines' model, however neither reaches the level of the official translations.

We can see that while the 'Lines' model focused a lot on the structure and ignored the language side, the 'Sections' model tried to retain balance in all metrics, coming out the weakest on rhyme.

5.5 Human Evaluation

We asked 26 people to participate in an A/B testing survey, providing them with a melody, the source English lyrics and two versions of the Czech target lyrics (see Appendix B). The conditions for participating in the survey were to speak both Czech and English and to be able to listen to the melody. No musical background was required, as we wanted to measure the preference of the general audience, not of music performers. We randomly sampled 10% of song sections out of the test set, recorded piano recordings of melodies of these sections and further randomly sampled sections for each survey separately, resulting in each survey being different.

The participants were to imagine that they were to sing the song adaptation as a part of a musical performance based on the original and choose the 'better' of the two. After all of the comparisons, they were asked to rank the 5 aspects of the Pentathlon Principle based on perceived importance. Results of the ranking are in Table 7. We can see that the most important aspect is naturalness, and the least important aspect by far is rhyme.

When looking at the percentages of the people who chose a given model when they had a choice, we get that 79% of the time people chose the official translation when given a choice. They chose the lyrics generated by sections 59% of the time, the MT 47% of the time and the lyrics generated by lines only 14% of the time.

	Ranked as #1	Avg ranking
Singability	5 x	2.85
Sense	6 x	2.85
Naturalness	8 x	2.12
Rhyme	0 x	4.54
Rhythm	7 x	2.65

Table 7: Pentathlon Principle aspects ranked from the most important (1) to the least important (5) by 26 survey participants.

Next, we divided the people into groups based on their choice of the Pentathlon Principle's top three most important aspects. The distribution of group preferences based on their first and second priority choices is provided in Appendix C. In Figure 3 we can see that 9 people who prefer sense, naturalness and rhythm favour the MT, which has high sense and naturalness scores, almost the same as the official translations. They prefer it more than the generated sections and do not give the generated lines a single vote. On the other hand, the group of 8 people favouring singability, naturalness and rhythm gave the most votes to the official translations, followed by the generated sections, where both of these setups excel in these three aspects. The generated lines which lack naturalness were chosen almost as many times as the MT which is mediocre in singability and rhythm. The other groups yield similar distributions except for one single person who prefers generated sections.

The human evaluation confirms that people generally prefer song translations with balanced aspects of the Pentathlon Principle, as well as that our metrics capture individual aspects well. It also

suggests that people's preferences differ, highlighting the necessity of producing balanced adaptations to be liked by the majority.

6 Conclusion

In this work, we propose an automatic metric system based on the pentathlon principle: metrics measuring the singability, sense, naturalness, rhyme and rhythm of translated song lyrics, and measure the ideal values of the metrics on human-translated official song lyrics. We propose a lyric translation system based on the pentathlon principle and implement it for the English-Czech language pair. We use the proposed metrics and human evaluation to compare the official translations, our generated translations and machine translations. The evaluation shows the validity of both our metrics and our lyric translation approach, as well as some insight into human preference when it comes to song translations, confirming Low's Pentathlon Principle.

Limitations

Limitations of this work are verifying the pentathlon principle for just one language, as well as the training-inference mismatch, which is necessary for training without bilingual data. Due to copyright reasons, the data are released under the Research Licence only. Lastly, due to our limited resources, we were able to verify the validity of our proposed lyrics generation system using only the smaller models from the LLM family.

Ethics Statement

We believe that our research does not inflict any harm on any group of people. We state that our goal is not to replace human translators with automated translators but rather to ultimately provide tools that could aid both professional and non-professional translators of human lyrics, and/or to allow automatically translating lyrics which would otherwise stay untranslated.

We believe that the way in which we use copyrighted materials (Czech and English song lyrics) does not violate any rules, as it falls under the copyright exception for scientific research (as defined by the European DSM Directive,[9] in Czechia implemented by §39d of Act 121/2000 Coll.). Our research is non-commercial and we do not further distribute the copyrighted materials except for further non-commercial research.

[9] https://www.europarl.europa.eu/doceo/document/A-8-2018-0245-AM-271-271_EN.pdf

Acknowledgements

The work has been partially supported by the EduPo grant (TQ01000153 Generating Czech poetry in an educative and multimedia environment), which is co-financed from the state budget by the Technology agency of the Czech Republic under the SIGMA DC3 Programme. Computational resources were provided by the e-INFRA CZ project (ID:90254), supported by the Ministry of Education, Youth and Sports of the Czech Republic. The work described herein has also been using data, tools and services provided by the LINDAT/CLARIAH-CZ Research Infrastructure (https://lindat.cz), supported by the Ministry of Education, Youth and Sports of the Czech Republic (Project No. LM2023062).

References

Lukáš Chaloupský. 2022. Automatic generation of medical reports from chest X-rays in Czech.

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

Michal Chudoba and Rudolf Rosa. 2024. GPT Czech Poet: Generation of Czech Poetic Strophes with Language Models. *arXiv preprint arXiv:2407.12790*, pages 1–9.

Martin Fajčík, Martin Dočekal, Jan Doležal, Karel Beneš, and Michal Hradiš. 2024. BenCzechMark: Machine language understanding benchmark for Czech language.

Johan Franzon. 2008. Choices in song translation. *The Translator*, 14(2):373–399.

Fenfei Guo, Chen Zhang, Zhirui Zhang, Qixin He, Kejun Zhang, Jun Xie, and Jordan Boyd-Graber. 2022. Automatic song translation for tonal languages. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 729–743, Dublin, Ireland. Association for Computational Linguistics.

Muhammad Qasim Khan, Abdul Shahid, M Irfan Uddin, Muhammad Roman, Abdullah Alharbi, Wael Alosaimi, Jameel Almalki, and Saeed M Alshahrani. 2022. Impact analysis of keyword extraction using contextual word embedding. *PeerJ Computer Science*, 8:e967.

Haven Kim, Kento Watanabe, Masataka Goto, and Juhan Nam. 2023. A computational evaluation framework for singable lyric translation. *Ismir 2023 Hybrid Conference*.

Peter Low. 2003. Singable translations of songs. *Perspectives*, 11(2):87–103.

Peter Low. 2005. The pentathlon approach to translating songs. In *Song and significance*, pages 185–212. Brill.

NLP team MosaicML. 2023. Introducing MPT-7B: A New Standard for Open-Source, Commercially Usable LLMs. Accessed 2024-08-09.

Longshen Ou, Xichu Ma, Min-Yen Kan, and Ye Wang. 2023. Songs across borders: Singable and controllable neural lyric translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 447–467, Toronto, Canada. Association for Computational Linguistics.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Olena Pidhrushna. 2021. Functional approach to songs in film translation: Challenges and compromises. In *SHS Web of Conferences*, volume 105, page 04003. EDP Sciences.

Petr Plecháč. 2018. A collocation-driven method of discovering rhymes (in Czech, English, and French poetry). *Taming the Corpus: From Inflection and Lexis to Interpretation*, pages 79–95.

Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtský. 2020. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Bahagia Saragih and Muhammad Natsir. 2023. The singable techniques are used in Emma Heesters's translated lyrics. *Randwick International of Education and Linguistics Science Journal*, 4:766–773.

Lucía Camardiel Sardiña. 2021. The translation of disney songs into spanish: Differences between the peninsular spanish and the latin american spanish versions. Master's thesis, University of Hawai'i at Manoa.

Zhonghao Sheng, Kaitao Song, Xu Tan, Yi Ren, Wei Ye, Shikun Zhang, and Tao Qin. 2021. Songmass: Automatic song writing with pre-training and alignment constraint. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 35, pages 13798–13805.

Yufei Tian, Anjali Narayan-Chen, Shereen Oraby, Alessandra Cervone, Gunnar Sigurdsson, Chenyang Tao, Wenbo Zhao, Yiwen Chen, Tagyoung Chung, Jing Huang, and Nanyun Peng. 2023. Unsupervised melody-to-lyrics generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9235–9254, Toronto, Canada. Association for Computational Linguistics.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. 2020. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788.

Kento Watanabe, Yuichiroh Matsubayashi, Satoru Fukayama, Masataka Goto, Kentaro Inui, and Tomoyasu Nakano. 2018. A melody-conditioned lyrics language model. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 163–172, New Orleans, Louisiana. Association for Computational Linguistics.

Peiyuan Zhang, Guangtao Zeng, Tianduo Wang, and Wei Lu. 2024a. Tinyllama: An open-source small language model. *arXiv preprint arXiv:2401.02385*.

Zhe Zhang, Karol Lasocki, Yi Yu, and Atsuhiro Takasu. 2024b. Syllable-level lyrics generation from melody exploiting character-level language model. In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 1336–1346, St. Julian's, Malta. Association for Computational Linguistics.

A Additional Experiments results

Table 8 shows the results of other base models used in the same way as described in the main body of the paper, measured on the Pentathlon Principle metrics. Using the TinyLlama as a base model yields results that do not respect the rules of the Czech language, it creates new words to comply with the length and rhyme requirements. It is interesting to see that using a small 137M parameter LLM yields just slightly worse results than using a 1.2B parameter model.

Table 9 shows results of the metrics discussed as preliminary experiments in Section 2 measured on the setups described in the main body of the paper. We can see that the BLEU score yields low results, verifying that song translations are not simple translations. The perplexity of a multilingual Mistral

		TinyLlama Lines	CsTinyLlama lines	TinyLlama sections	CsTinyLlama sections	CsGPT2-small sections
CCVO Distance	↘	**0.22**	0.23	0.26	0.25	0.29
Semantic Similarity	↗	0.44	0.46	0.45	**0.51**	0.48
Perplexity CsMPT	↘	938	748	212	**92**	99
Rhyme Scheme JI	↗	**0.81**	0.73	0.44	0.38	0.33
Syllable Distance	↘	**0.01**	**0.01**	0.02	**0.01**	0.06

Table 8: Additional results of TinyLlama (Zhang et al., 2024a) and Czech TinyLlama (Fajčík et al., 2024) fine-tuned to generate each line of the song section individually, and of the TinyLlama, Czech TinyLlama and Czech GPT2-small (Chaloupský, 2022) models fine-tuned to generate a whole section at once.

			Baseline	Official	MT	Lines	Sections
Singability	**CCVO Distance**	↘	0.70	0.27	0.39	**0.23**	0.25
Sense	**Semantic Similarity**	↗	0.23	0.62	**0.91**	0.46	0.51
	BLEU2	↗	0.00	0.04	-	0.01	0.01
Naturalness	**Perplexity CsMPT**	↘	131	131	97	748	**92**
	Perplexity Mistral	↘	41	41	**25**	67	34
Rhyme	**Rhyme Scheme JI**	↗	0.20	0.60	0.27	**0.73**	0.38
	Recall-based rhyme	↗	0.55	**0.77**	0.32	0.74	0.64
Rhythm	**Syllable Distance**	↘	0.65	0.02	0.26	**0.01**	**0.01**
	Syllable Accuracy	↗	0.10	0.83	0.20	**0.99**	0.98
	Stress Distance	↘	0.20	**0.72**	0.56	0.68	0.67

Table 9: Baseline, official translations of musical songs, MT and the lyrics generated by lines and by sections evaluated by a portion of metrics we experimented with. For each metric, we show the direction depending on whether we are aiming for higher or lower numbers in that metric.

favours MT and can not see the unnaturalness of the lyrics generated by lines, as it can not generate Czech well. The recall-based rhyme scheme metric shows that even human translators do not strictly keep the rhyme scheme.

B Human Evaluation Questionnaire

An example of one question from the human evaluation questionnaire can be seen in Figure 4.

C Human Evaluation Results

In this Section, we present additional graphs showing the results of the human evaluation. Preliminary experiments revealed that identifying the single most important aspect of the Pentathlon Principle is very difficult. For this reason, in the main body of the paper, we show the graph dividing people into groups by their top three aspects of the Pentathlon Principle. Nevertheless, as shown in Figure 5, individuals who prioritized rhythm favoured rhythmic models, and similar patterns emerged for other preferences. Figure 6 shows that 10 participants prioritized naturalness and sense, while 6 favoured rhythm and singability. The remaining participants

```
original:
flower gleam and glow
let your power shine
make the clock reverse
bring back what once was mine

1:
květinový lesk a záře
nech svou sílu zářit
zvrátit čas
vrať mi to co bylo kdysi moje

2:
jak se mi líbíš
svítíš jako květ
že je to tak rok
co jsem tě uviděl
```

Figure 4: One question from the human evaluation questionnaire. The participants were provided with a matching melody together with each question. The first song section is the machine translation of the original, and the second song section is a generated adaptation, that translates to: *How I like you, you shine like a flower, it's been a year, since I saw you.*

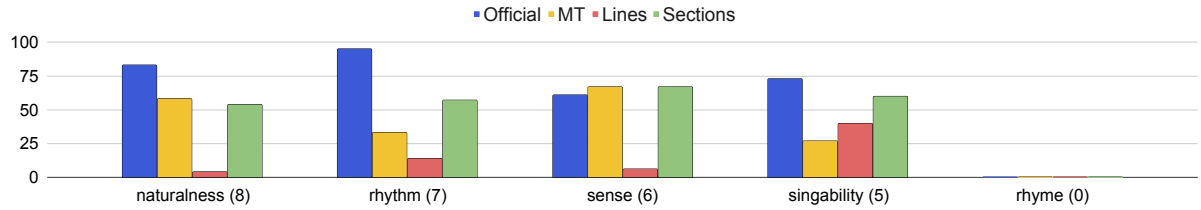


Figure 5: Percentages of times people chose a specific setup during manual evaluation. The graphs show the preferences of groups divided based on what they consider the most important aspect of the Pentathlon Principle. Number of people in each group is in brackets.

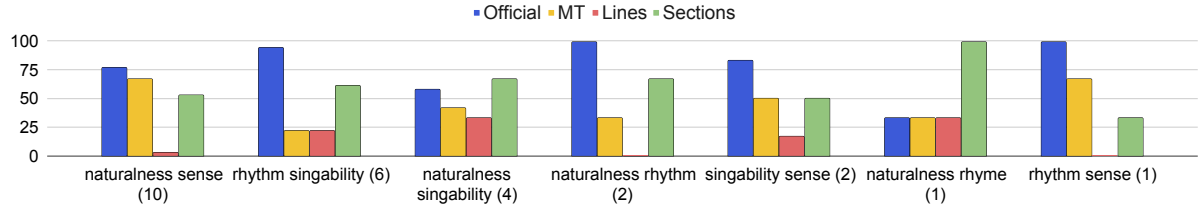


Figure 6: Percentages of times people chose a specific setup during manual evaluation. The graphs show the preferences of groups divided based on what they consider the two most important aspects of the Pentathlon Principle. Number of people in each group is in brackets.

showed more mixed preferences, leading to less clear distinctions.