Evaluation of Generated Poetry

David Mareček, Kateřina Motalík Hodková, Tomáš Musil, Rudolf Rosa
Insitute of Formal and Applied Linguistics
Faculty of Mathematics and Physics, Charles University
Prague, Czechia
{marecek, hodkova, musil, rosa}@ufal.mff.cuni.cz

Abstract

We propose a range of automated metrics for evaluation of generated poetry. The metrics measure various aspects of poetry: rhyming, metre, syntax, semantics, and amount of unknown words. In a case study, we implement the metrics for Czech language, apply them to poetry generated by several automated systems as well as human-written, and correlate them with human judgment. We find that most of the proposed metrics correlate well with corresponding human evaluation, but semantically oriented metrics are much better predictors of the overall impression than metrics evaluating formal properties.

1 Introduction

With current Large Language Models (LLMs), automated *generation* of creative texts is becoming easier than ever, including tasks that have always been considered difficult to achieve, such as automated generation of poetry (Shahriar, 2022; Belouadi and Eger, 2023; Agirrezabal and Oliveira, 2024; Valença and Calegario, 2025). While there is probably little reason in trying to automate poetry generation in the sense of simulating the human artistic practice per se, it may be useful e.g. for educating students of literature. An interactive poetry generator can bring dusted poetry to life, allowing students to generate new variants of existing poems by differing some of their aspects (e.g. style, language, rhyme, metre, themes), provide them with full interpretative freedom when working with completely newly generated poems, as well as support starting writers by helping them to express their ideas and improve their style.

In order to train models for any task, it is crucial to be able to reliably perform automated evaluations of the model outputs, as this guides model development, allows comparison of quality to human performance, and enables automated output selection/reranking at inference. However, automated *evaluation* of generated creative texts remains a challenge for multiple reasons, such as:

- The task is considerably open-ended, making it impossible to list a relevant set of optimal outputs to compare to.
- The output cannot be easily treated as fulfilling a set of clear subtasks, completion of which could be easily measured.
- The task is not completely well defined, as even human evaluators struggle to reach a consensus in evaluating poetry, generated or written by human poets.
- While LLMs can be successfully used to evaluate various aspects of texts, there is a significant threat of skewed results when using an LLM to evaluate outputs generated by the identical LLM (or a similar one).

In our work, we specifically focus on automated ways of evaluating the quality of automatically generated poetry, which makes the task even more difficult in some aspects. It may be argued that poetry—like other art forms—cannot be fully understood or evaluated by machines alone, since aesthetic judgment presupposes human experience and self-reflection. Moreover, Porter and Machery (2024) show that evaluating the quality of poetry is not straightforward even for humans, as their study revealed that humans may actually prefer generated poetry to human-written poetry under some circumstances. Without disputing these claims, we counter that *some* aspects of a poem, relevant to the quality of the poem, can presumably be rather objectively evaluated and measured. These include formal properties such as rhyming and metre, which are irrelevant in general prosaic texts.

In this paper, we propose a range of automated metrics related to various aspects of poetry. The metrics are reference-free, requiring only the text of the poem on input. As a case study, we implement and test the metrics in the context of poetry written

Proceedings of the 5th Workshop on Evaluation and Comparison of NLP Systems, pages 109–118

in the Czech language. We analyze the relevant metrics on poems generated by LLMs and a large corpus of human-written poetry. In addition to automated evaluation, the texts are evaluated by human annotators.

Our work follows similar directions as Erato (Agirrezabal et al., 2023), which also evaluates poetry quality along multiple dimensions using statistical metrics. More recently, Sahu and Vechtomova (2025) also employ LLM prompting to evaluate poetry quality. There are also works on evaluating poetry e.g. in Russian (Koziev, 2025) or Chinese (Zhao and Lee, 2022). Most of the evaluators are language-dependent, and we are not aware of any previous work evaluating quality of Czech poetry.

2 Metrics

We now describe our proposed metrics; a case study implementing the metrics for the Czech language poetry follows in Section 3. Our metrics come in three variants, based on how the poem is processed:

STAT Quality assessed by computing a statistic.

LLM Quality assessed by prompting an LLM.

HUMAN Quality assessed by a human evaluator.

Our STAT metrics are based on structured analyses of the poems. LLM and HUMAN metrics amount to asking the LLM or the human annotator a question about the quality of the poem, such as "Rate the rhyming of the following poem on a scale 0-10." We propose identical prompts/instructions for humans and LLMs (detailed in Table 1).

2.1 RHYMING

In many poetic traditions, a poem is organized around a rhyme scheme, which specifies which lines should rhyme with each other. The exact definition of what constitutes "ending in a similar way" sufficiently to be considered rhyming is language-specific. However, the general principle is that we find the rhyming part (*reduplicant*) in each verse, take its phonetic transcription, and check whether it is identical or sufficiently similar to the reduplicant of the corresponding verse.

In STAT-RHYMING, we propose to compute the ratio of verses v_i in poem P rhyming with at least one other verse v_j within a context window of K verses before:

$$s_r = \frac{\sum_{i=1}^{|P|} \mathbf{1}_{i-K \le j < i} \, \text{rhymes}(v_i, v_j)}{|P|} \qquad (1)$$

A potential future improvement of the metric might also take into account rhyme scheme consistency across stanzas, as all stanzas of a poem typically pertain to the same rhyme scheme.[1]

2.2 METRE

METRE is a metric that examines how regular the rhythmic structure of a poem is. The rhythmic structure is achieved by the alternation of stressed and unstressed syllables, according to an intended metre (e.g. iamb, trochee, or dactyl). As our proposed evaluation setting has no information about the intended metre on the input, the first step is to determine the most likely metre of the poem. The next step is to assess how perfectly the poem pertains to the metre.

As for STAT-METRE, we propose to compute consistency of each verse v in poem P with the apparent metre M,[2] averaged over all verses:[3]

$$s_m = \frac{\sum_{v \in P} \text{consistency}(v, M)}{|P|} \qquad (2)$$

We found that properly implementing the consistency measure may be difficult. Our initial approach was to automatically mark syllable stresses and to measure the ratio of syllables stressed consistently with the metre, but we found that blindly following the formal metre rules in this way is an oversimplification and does not correlate well with human-perceived metric quality. Therefore, our proposed approach, which we use in our case study, is to estimate the consistency of the stress pattern with the metre using a model trained on metre annotations in a poetry corpus, if available.

2.3 KNOWN-WORDS

While neologisms are a productive part of language development, in general text, we usually consider the appearance of non-existent words to be an error. Poetry is considerably more free in this aspect, with poets frequently introducing new words, e.g. by deriving, compounding or blending existing words. However, in generated poetry, we have observed a considerable amount of non-existent words that

[1] However, care should be taken when designing such a metric, as many poems systematically use multiple rhyme schemes, including the prime example of sonnets.

[2] In a polymetric poem, the metre may differ across verses.

[3] As we do not presuppose the knowledge of the intended metre, the apparent metre first needs to be detected. Alternatively, one may compute this metric for all possible metres, and then take the maximum value.

Metric	Quality	Gloss
SEMANTICS	smysluplnost	meaningfulness of
SYNTAX	syntaktickou konzistenci	syntactic well-formedness of
RHYMING	rýmování	rhyming of
METRE	metrickou konzistenci	metrical consistency of
KNOWN-WORDS	nesmyslná slova	nonsense words of
OVERALL IMPRESSION	celkový dojem z	overall impression from

Table 1: The prompts/instructions used for evaluating the poems given to the LLM/to the human annotators. For all metrics, the complete prompt/instruction followed the following template:
Na škále 0 až 10 ohodnoť <quality> následující básně. Napiš pouze to číslo.\n\n <poem>
(*On a scale from 0 to 10, rate the <quality> the following poem. Write only the number.\n\n <poem>*)
All the prompts/instructions were given in the language of the poems (English glosses provided here for reference).

even proficient users of the language cannot meaningfully interpret in the context of the poem. This seems to most frequently happen at the end of the verse, apparently with the model trying to fulfill the formal requirements of the poem (rhyming, and/or metre).[4]

As judging the transparency of a neologism is hard even for humans, let alone automated tools, we propose this metric as a ratio of words that are part of the lexicon of the language.

In STAT-KNOWN-WORDS, this is a matter of a simple check in a sufficiently large morphologically inflected lexicon of the language. We define STAT-KNOWN-WORDS as the ratio of tokens of the poem P present in the lexicon L:

$$s_{kw} = \frac{\sum_{i=1}^{|P|} \mathbf{1}\{ P_i \in L \}}{|P|} \quad (3)$$

In HUMAN-KNOWN-WORDS, we suggest to rely on the introspection of native speakers of the language (who can always consult a lexicon if unsure).

2.4 SYNTAX

Syntactic properties of poetic text are complex and do not directly fully map to syntactic properties of prosaic text, yet there are numerous rules and strong tendencies that are mostly or fully observed even in poetry (Cinková et al., 2024; Karimovna and Saurikova, 2025).[5] We thus believe that a structured statistical approach evaluating some of the syntactic aspects of the poem could be implemented, and their observation or violation may be a useful indicator of the poem quality.

Unfortunately, we are not aware of any practically usable tools for automated syntactic analysis of poetry, as syntactic parsers are typically trained on prosaic texts (Straka and Straková, 2017) and syntactically annotated corpora of poetry are extremely scarce and tiny. Therefore, we only implement the HUMAN and LLM variant of the SYNTAX metric, leaving the investigation of a potential STAT-SYNTAX for future work.

2.5 SEMANTICS

Meaningfulness or semantics in poems (or generally in art) can be difficult to define and to apply strict rules to, as everyone may interpret it differently, finding or ignoring connections between its elements, chosen lexical units, stylistic devices, etc. We are not aware of any usable automated tools applicable to poetry that would provide us with useful semantic analyses; therefore, we propose this metric only in the HUMAN and LLM variants.

Inspired by the work of Rastier (2009) on Interpretative Semantics and isotopy, and by practical feedback provided to us by our evaluators, we believe that a viable future path for a more structured measure of meaningfulness may focus on the coherence, continuity and recurrence of various themes or motives introduced in the poems. Unfortunately, the research on automated motive analysis of Czech poetry has been unsuccessful so far (Kořínková et al., 2024). There is some promising work in progress on our side, but at this point, we need to leave a potential STAT variant of this metric for future work.

2.6 OVERALL IMPRESSION

The HUMAN-OVERALL IMPRESSION is our main target metric that we are typically ultimately trying to maximize. While we may assume that the hu-

[4] This is of course made possible by the use of subwords in most current LLMs.

[5] In Czech, the already considerably flexible word order is even more free in poetry, whereas morphological agreement is strictly observed, and the verb-complement structure is generally observed but occasionally violated (*anacoluthon*).

man evaluator presumably takes all the previously mentioned qualities of the poem into account when assessing the OVERALL IMPRESSION, the metric is not necessarily an aggregate of the other metrics. The final scores are influenced by the subjective impression of each poem. Although not an objective method, we believe that individuals may respond to the same work of art with diverse emotions and judgments, perceiving it positively or negatively in different ways. We thus think that this metric simulates how potential users of our poems-generating models may perceive the models' output, as users without deeper knowledge of the domain and without the access to a set of evaluation metrics or tools are unlikely to analyse various aspects of poem in detail before formulation a conclusion about the poem's quality.

3 Experimental Settings

In our case study, we focus on evaluating generated poetry in Czech language. We implement the proposed metrics for Czech poetry, gather several datasets of Czech poems for evaluation, hire annotators, and compare results of the automated metrics to human evaluations. This section describes the experimental settings; the results are presented and discussed in the next section. All our codes, data and results are available in our public repository.[6]

3.1 Poetry Data

We compiled an evaluation corpus of 100 poems originating from the following five sources, 20 poems from each source. As we partially focus on the formal aspects of rhyme and meter, we did not include free verse and/or non-metrical poems.

CCV Real poems written by existing Czech poets, randomly sampled from the Corpus of Czech Verse (Plecháč and Kolár, 2015).[7]

LLM Poems generated by ChatGPT.[8]

our16-40000 Poems generated by our model[9] (trained for 40,000 epochs, 16-bit precision).

our16-7500 Poems generated by our model (7,500 epochs, 16-bit precision).

our4 Poems generated by our model (7,500 epochs, 4-bit precision for inference).

All the poems were converted into a simple unified plaintext format, featuring only the title and text of the poem,[10] and their order was randomized.

3.2 Metric Implementation

We decided to implement all the metrics in the $[0, 1]$ range (higher is better). For HUMAN and LLM metrics, we ask the annotator/model to produce a score in the more natural $[0, 10]$ range, and then normalize it into the target range.

We used the same simple prompts/instructions for both LLM and HUMAN, detailed in Table 1. We also experimented with more detailed instructions for LLM, based on the few-shot and chain-of-thought approaches, but did not find them to lead to a notable improvement of the results.[11]

For all LLM metrics, we used a `gpt-4o-mini` with `temperature=0` (deterministic generation).

For STAT-RHYMING, we use the automatic rhyme detection tool RhymeTagger[12] (Plecháč, 2018). This tool examines each pair of verses in a given context window, estimates the probability that the verses' reduplicants rhyme with each other,[13] and identifies the rhyming verses as those that exceed a given threshold.

For STAT-METRE, we use the tool Květa (Plecháč, 2016), which analyzes the poem by detecting syllables and stresses, and for each verse, it computes the probabilities of four metres (iamb, trochee, dactyl, amphibrach),[14] which we use as measures of consistency of the verse with the metres. The resulting STAT-METRE score is the probability of the globally highest-scoring metre averaged over all verses.[15]

[6] https://github.com/ufal/edupo

[7] We skipped poems that were too old (written before 1850) or too long (more than 32 lines).

[8] We generated poems with `gpt-4o-mini`, using the prompt *"Vygeneruj českou rýmovanou báseň"* ("Generate a rhymed poem in Czech"). To achieve some diversity, we iteratively specified more parameters, such as a specific theme, metre, and/or rhyme scheme.

[9] Specifically, our poetry-generation model is a Llama 3.1 model, fine-tuned on CCV using LoRA (Hu et al., 2021) with Unsloth (Han et al., 2023) [anonymized citation].

[10] We omit the author name, even though almost all of the sources provide one, as we do not focus on stylometry and do not find the author to be important to assess the poem quality (potentially even biasing the annotators).

[11] Some further discussion in Section 4.5.

[12] https://github.com/versotym/rhymetagger

[13] The rhyming probabilities are simple statistical estimates on the CCV, i.e. a statistic of how often such a pair of reduplicants was marked as rhyming by the annotators of the corpus.

[14] The tool does not detect other possible metres.

[15] The probability of a metre for a verse is not a direct rule-based computation of the average stress consistency, as the tool also takes other aspects into account, and then trains a metre identification model on the CCV corpus; the consistency score is thus an estimation of how consistent the particular stress pattern is with the given metre based on corpus observations.

For STAT-KNOWN-WORDS, we use the large inflected MorfFlex lexicon (Hajič et al., 2024) indirectly through analyzing the text with the UDPipe morphological tagger (Straka and Straková, 2017); when its guesser is turned off, it does not produce analyses for words not present in MorfFlex.

3.3 Human Evaluation

We employed three human experts: a *linguist*, a *versologist*, and a *literary expert*.[16] A first round of evaluation was done by the *linguist*, annotating all six HUMAN metrics. As HUMAN-SEMANTICS was clearly identified as the most useful metric in the first round, the other two experts were then only asked to provide annotations for HUMAN-SEMANTICS. Also, we found the *linguist* to be incapable of providing high-quality annotations for HUMAN-METRE; therefore, the HUMAN-METRE was scratched and redone by the *versologist*.[17] Thus, in the reported results, HUMAN-SEMANTICS is an average of 3 human experts, and all other HUMAN metrics are by one expert only.[18]

4 Results

Figure 1 shows the evaluation of the poetry datasets using all of the proposed metrics, and Table 2 measures how each of the metrics correlates with the human-reported overall impression, using Pearson coefficient. Table 4 and Figures 2 and 3 evaluate some further inter-correlations among the metrics.

4.1 What Are Optimal Values of the Metrics?

All the proposed metrics are in the $[0.0, 1.0]$ range, thus the apparent optimal value for each metric is 1.0. However, Figure 1 clearly shows that even for the professional human-written poems in the CCV corpus, none of the metrics typically reach this value, as human-written poetry often deviates from the theoretical ideals in various ways. Therefore, to simulate human-written poetry, one may wish not to maximize the metrics but rather to reach values similar to those observed on human-written poetry.

Metric	Corr. HOI
HUMAN-SEMANTICS	0.90
HUMAN-SYNTAX	0.87
HUMAN-KNOWN-WORDS	0.67
HUMAN-METRE	0.16
HUMAN-RHYMING	0.28
LLM-SEMANTICS	0.59
LLM-SYNTAX	0.56
LLM-METRE	0.64
LLM-RHYMING	0.61
STAT-KNOWN-WORDS	0.54
STAT-METRE	0.10
STAT-RHYMING	-0.11

Table 2: Correlation of all the metrics with HUMAN-OVERALL IMPRESSION.

4.2 SEMANTICS

Already when examining the human evaluations, we can clearly see that the human annotators find semantics to be crucial for the overall impression (correlation 0.9 in Table 2; the pairwise inter-annotator correlations are $\{0.53; 0.71; 0.75\}$). The corresponding automated LLM-SEMANTICS metric seems to be highly useful, as it is rather reliable (correlation 0.65 with HUMAN-SEMANTICS in Figure 2, which is competitive with the inter-annotator correlations) and has a high impact on the overall impression (correlation 0.59 in Table 2).

However, Figure 1 shows the well-known self-favoring bias of LLMs, as `gpt-4o-mini` favors its own results over all other systems (including human-written poems) in all LLM-based metrics, which is not warranted by the human evaluation. Therefore, LLM-SEMANTICS can be used to compare the quality of multiple individual poems generated by one system, but cannot reliably compare the quality of poems generated by the judging LLM to poems generated by other systems (although it presumably can rank multiple systems that are similarly different from the judging system).

4.3 SYNTAX

Table 2 shows HUMAN-SYNTAX highly correlated with the overall impression (0.87). On the other hand, Table 2 reveals that SYNTAX is highly correlated with SEMANTICS in both HUMAN and LLM variants (0.75 and 0.71), much higher than any other HUMAN metric. Figure 2 shows that the

[16] All of the experts are members of our paid research team (distinct from the designers of the metrics and the generator models) and are thus fully compensated for their work.

[17] We did not observe such issues with the other metrics; it seems the metre is not sufficiently well known and requires prior training for non-experts.

[18] Identically to the LLM evaluator, we did not provide the human annotators with specific instructions as of what do specific values of the metrics correspond to, as long as poems perceived as better get a higher value. The same value of the metric thus does not necessarily mean the same thing across different annotators. This is not an issue when simply correlating the results, but we note that absolute values of the metrics should not be compared across annotators and/or LLMs without prior adjustment.

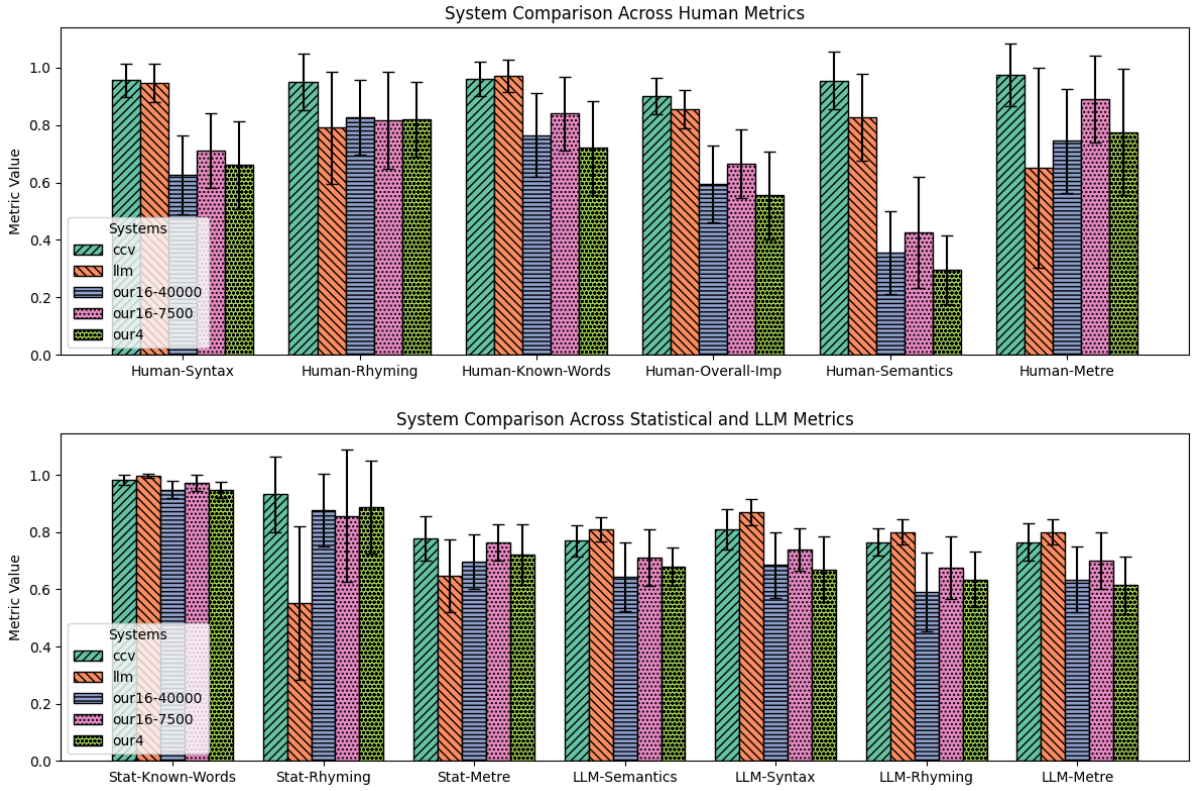


Figure 1: Values of the proposed metrics on Czech poetry generated by various systems as well as human-written.

correlation between HUMAN-SYNTAX and LLM-SYNTAX is 0.52, which is respectable, but can alternatively be explained through both of these metrics being highly correlated with SEMANTICS. It is thus unclear to what extent SYNTAX measures something useful in addition to SEMANTICS. On the other hand, a potential future STAT-SYNTAX measure, based on classical syntactic parsers (Straka and Straková, 2017) and syntactic properties of poetry (Cinková et al., 2024), might be a cheaper proxy to LLM-SEMANTICS.

4.4 KNOWN-WORDS

We have found that it is very useful to look at the ratio of out-of-vocabulary words in the poems (HUMAN-KNOWN-WORDS has 0.67 correlation with OVERALL IMPRESSION in Table 2). Our systems often generated too many non-existent and mostly nonsensical words, which the annotators found to severely hurt the semantics of the poems, even if this was apparently done due to the effort of the system to fulfill the formal rules of metre and rhyming.[19]

STAT-KNOWN-WORDS is quite reliable (correlation 0.73 with HUMAN-KNOWN-WORDS in Figure 3), fast and easy to compute, and useful for predicting OVERALL IMPRESSION (correlation 0.54 in Table 2).

To investigate to what extent the success of this measure is an artifact of the generators producing too many unknown words, we also measured its correlation with OVERALL IMPRESSION only on CCV human-written poems. The correlation stays moderate (0.50), suggesting that STAT-KNOWN-WORDS may be rather useful in general. However, it is worth noting that most CCV poems have no or very few unknown words and they all received very high OVERALL IMPRESSION scores, and thus no strong conclusions can be drawn here.

4.5 RHYMING and METRE

Although LLM-RHYMING and LLM-METRE correlate well with the overall impression (0.61 and 0.64 in Table 2), we have found that, in fact, all LLM based metrics highly correlate with each other (see Table 4) while showing only low correlations with the corresponding HUMAN evaluations (0.17 for METRE and 0.21 for RHYMING, see Figure 2). I.e., it seems that `gpt-4o-mini` is rather good at judging the meaningfulness of the

[19] Our systems are clearly overtuned for formal quality of the generated poems, at the cost of their meaningfulness.

Metric A	Metric B	CCV poems	generated poems	all poems
LLM-SEMANTICS	HUMAN-SEMANTICS	0.26	0.69	0.65
STAT-KNOWN-WORDS	HUMAN-KNOWN-WORDS	0.37	0.74	0.73
STAT-RHYMING	HUMAN-RHYMING	-0.04	0.48	0.48

Table 3: Correlations of human and automated variants of several metrics, measured separately on human-written (CCV) and generated subsets of the evaluation dataset.

Metric A	Metric B	LLM	HUMAN
SEMANTICS	SYNTAX	0.71	0.75
SEMANTICS	RHYMING	0.76	0.14
SEMANTICS	METRE	0.78	0.30
RHYMING	SYNTAX	0.80	0.25
METRE	SYNTAX	0.76	0.10
RHYMING	METRE	0.79	0.21

Table 4: Correlation between various pairs of metrics (metric A and metric B), either in LLM variant or HUMAN variant (i.e. not a correlation of LLM metrics with HUMAN metrics).

poems, but is mostly unable to judge other qualities and resorts to judging meaningfulness even when prompted to judge metre or rhyming.[20] Using LLMs to asses formal properties of poetry thus does not seem very promising and STAT metrics seem to be superior; this is in line with findings of Agirrezabal and Oliveira (2025).

STAT-METRE and STAT-RHYMING are rather reliable (0.77 and 0.48 correlations with HUMAN-METRE and HUMAN-RHYMING in Figure 3). However, the results in Table 2 clearly show that our annotators strongly favor meaningfulness over these formal aspects, with low correlations with OVERALL IMPRESSION already for HUMAN-METRE and HUMAN-RHYMING (0.16 and 0.28), and subsequently with no meaningful relation between the overall impression and STAT-METRE or STAT-RHYMING (correlations 0.10 and -0.11, respectively). This is thus partially a negative result: Even professional human evaluators do not care much about the metre and rhyming in generated poetry, and thus measuring these aspects, even if with a high accuracy, is not a good predictor of the human-perceived quality of the generated poems. In general, it seems to be much more fruitful to focus on the semantic quality rather than formal qualities in poetry generation; this is in line with findings of Porter and Machery (2024).

Our annotators also noted that they were reluctant to rate a poem poorly if it was not formally perfect in rhyming and/or metre, since historically, the adherence to the rules in human-written poetry varied, and many authors violated some of the rules on purpose for various reasons. Thus, it is not straightforward to decide for some of the violations if these should be treated as intentional deviations or unintentional errors. On the other hand, they also noted that our proposed automated metrics do not capture various other relevant formal aspects, such as syllable count regularity, tautological rhymes,[21] or ingenuity of the rhyme scheme.[22] This constitutes potential future improvements, although of questionable importance given the low correlation with OVERALL IMPRESSION.

4.6 Metric Combination

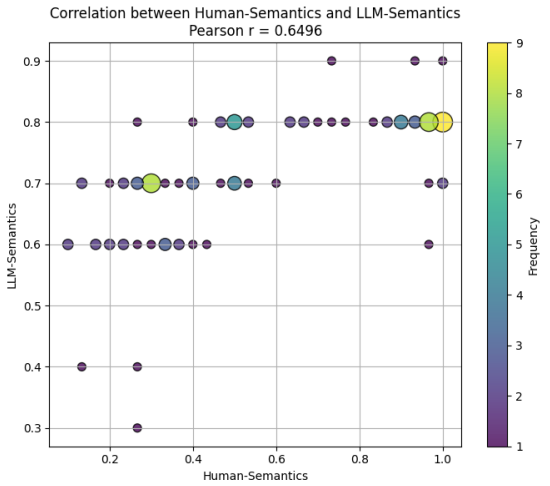
The two best-performing automated metrics are LLM-SEMANTICS and STAT-KNOWN-WORDS, and they are only moderately correlated (0.65), which suggests options for a combined metric. However, the small amount of human-rated poems currently available to us does not allow for any extensive tuning and testing of the metric combination parameters. Therefore, we only evaluate a single straightforward combination metric, computed as a multiplication of LLM-SEMANTICS and STAT-KNOWN-WORDS.

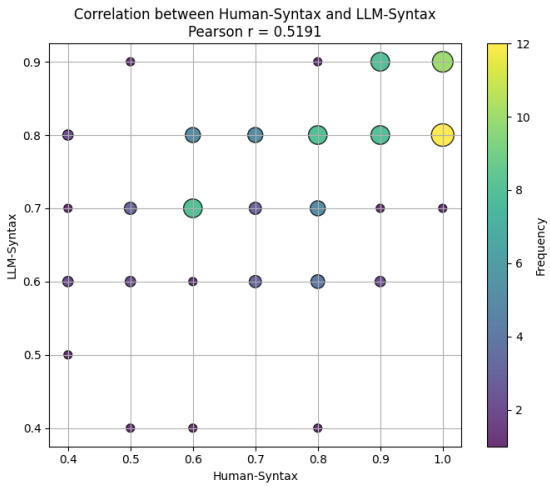
The correlation of the combined metric with OVERALL IMPRESSION is 0.62, which is a slight improvement over the individual metrics (0.59 and 0.54 respectively).

[20] Conversely, we found that `gpt-4o-mini` is rather apt at generating poems reasonably well pertaining to the specified metre (and to some extent also to the rhyme scheme), i.e. these are generative but not analytical capabilities of the model. We have confirmed this with further experiments based on the chain-of-thought approach, where we prompted the model to analyze the rhyming and metre of various poems verse by verse and stanza by stanza. The model produced correct theoretical knowledge and correctly identified many key features of the poems, but then nevertheless produced mostly incorrect metre and rhyme scheme labels.

[21] Rhyming a word with itself.

[22] In Czech poetry, e.g. couplet-based rhyme schemes (AABBCCDD...) are typically considered low style, typical for folk poetry and children poetry, while high style uses more intricate rhyme schemes.





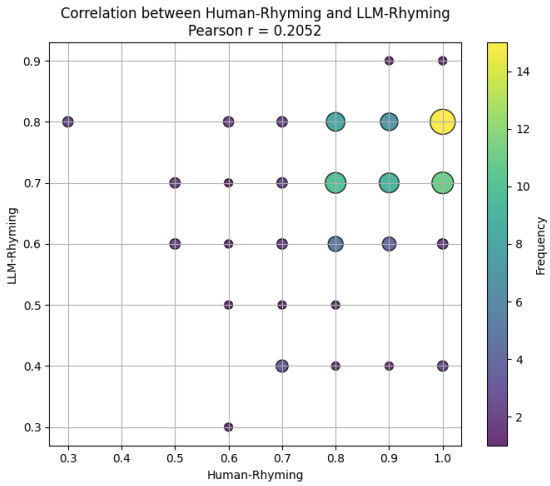
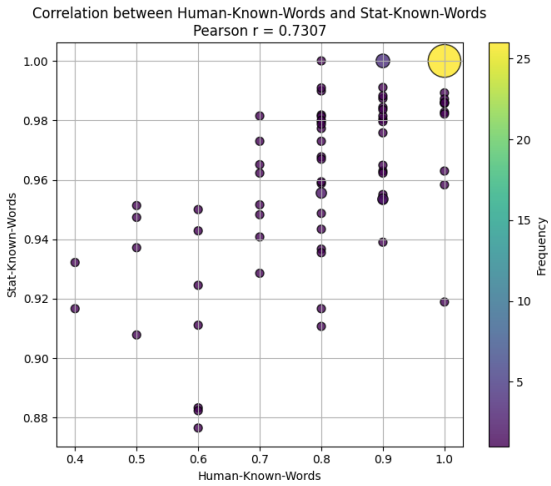
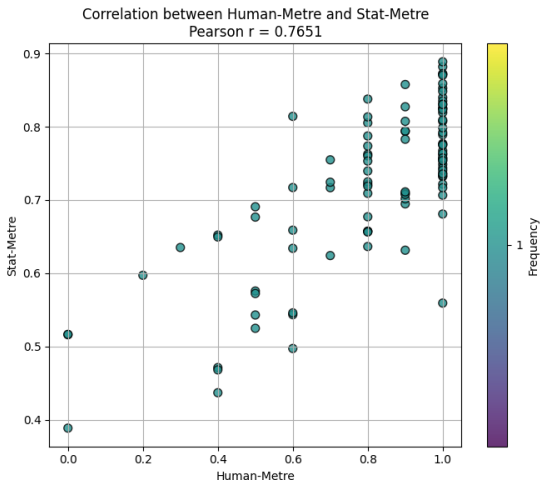


Figure 2: Correlation between HUMAN and LLM metrics (`gpt-4o-mini`) for SEMANTICS, SYNTAX and RHYMING for individual poems.





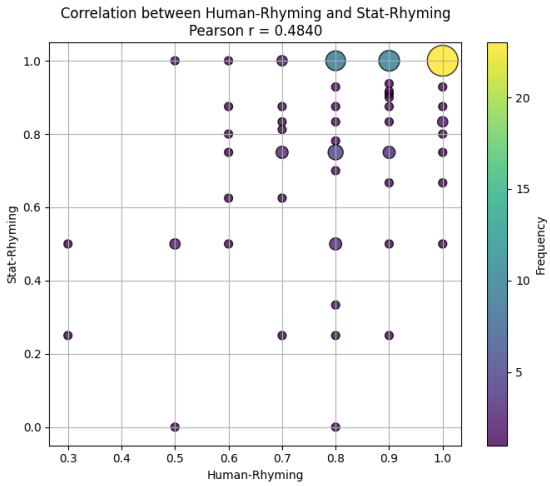


Figure 3: Correlation between HUMAN and STAT metrics for KNOWN-WORDS, METRE and RHYMING for individual poems.

4.7 Reliability of Metrics on Human-written vs. Generated Poems

While our metrics are primarily designed to be used on generated poetry, all the results reported so far have been measured on a mix of generated and human-written poetry. In Table 3, we investigate the reliability of several metrics separately on human-written (CCV) and on generated poems, by correlating the automated metrics with the human annotations.[23]

The results clearly show that the metrics perform rather poorly on human-written poems, and thus should only be used on generated poetry.

5 Conclusion

In this paper, we proposed a range of automated metrics that measure various aspects of poem quality, both statistics-based and LLM-based. The metrics are designed to evaluate automatically generated poetry, both for comparing multiple poetry generation systems or variants of one system, as well as to allow for automated selection/reranking of generated poems based on their quality.

In our case study on Czech poetry, we identified the metrics LLM-SEMANTICS (prompting `gpt-4o-mini` to assess how meaningful the poem is) and STAT-KNOWN-WORDS (computing the ratio of out-of-vocabulary words based on a morphological dictionary) as the most useful. Both of these metrics are rather reliable, correlating well both with their human variants as well as with the human-perceived overall poem quality; the combination of these two metrics (by multiplication) performs even slightly better than each of the metrics alone. However, both metrics also have clear limitations. STAT-KNOWN-WORDS is fast and cheap to compute, although its success in our case study may be due to the fact that many of the evaluated poetry generating models simply generated too many nonsensical words (in order to fulfill the formal poetry rules), and its usefulness might thus diminish with better generator models. As for LLM-SEMANTICS, it is only useful for ranking multiple poems generated by one system, and for ranking multiple systems sufficiently different from the judging LLM, as we have reconfirmed the pre-existing observation that LLMs tend to judge their own outputs more favorably.

[23]Note that the CCV subset only constitutes 20% of the evaluation dataset, and thus the performance on generated poems has much stronger influence on the evaluation of the metrics on the whole dataset.

We were also able to reliably implement versologically motivated metrics evaluating metre and rhyming, but we did not find them useful for evaluating the overall quality of the generated poems, as the human annotators favored content over form.

Despite being confined to the setting of our case study, our findings seem to reaffirm conclusions drawn in several related studies.

Acknowledgments

The work has been 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. The first three authors are supported by the project "Human-centred AI for a Sustainable and Adaptive Society" (reg. no.: CZ.02.01.01/00/23_025/0008691), co-funded by the European Union. 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). The research was also supported by Czech Science Foundation Grant No. 25-14501L.

References

Manex Agirrezabal and Hugo Gonçalo Oliveira. 2024. Zero-shot metrical poetry generation with open language models: a quantitative analysis. In *Proceedings of ICCC24*.

Manex Agirrezabal and Hugo Gonçalo Oliveira. 2025. Refining metrical constraints in LLM-generated poetry with feedback. In *Proceedings of ICCC25*.

Manex Agirrezabal, Hugo Gonçalo Oliveira, and Aitor Ormazabal. 2023. Erato: Automatizing poetry evaluation. In *EPIA Conference on Artificial Intelligence*, pages 3–14. Springer.

Jonas Belouadi and Steffen Eger. 2023. ByGPT5: End-to-end style-conditioned poetry generation with token-free language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7364–7381.

Silvie Cinková, Petr Plecháč, and Martin Popel. 2024. Rhymes and syntax: A morpho-syntactic analysis of Czech poetry. *Primerjalna književnost*, 47(2).

Jan Hajič, Jaroslava Hlaváčová, Marie Mikulová, Milan Straka, and Barbora Štěpánková. 2024. MorfFlex CZ

2.1. LINDAT/CLARIN digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Daniel Han, Michael Han, and Unsloth team. 2023. Unsloth.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. LoRA: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Karimova Shaxnoza Karimovna and Gulbagira Saurikova. 2025. Syntactic devices used in poetic language. *Modern American Journal of Linguistics, Education, and Pedagogy*, 1(2):291–295.

Lucie Kořínková, Tereza Nováková, Michal Kosák, Jiří Flaišman, and Karel Klouda. 2024. Motivické a tematické klastry v básnických textech české poezie 19. a počátku 20. století: k novým možnostem využití databáze česká elektronická knihovna. *Ceska Literatura*, 72(2):204–217.

Ilya Koziev. 2025. Automated evaluation of meter and rhyme in Russian generative and human-authored poetry. *arXiv preprint*.

Petr Plecháč. 2018. *A Collocation-Driven Method of Discovering Rhymes (in Czech, English, and French Poetry)*, pages 79–95. Springer International Publishing, Cham.

Petr Plecháč. 2016. Czech verse processing system KVĚTA – phonetic and metrical components. *Glottotheory*, 7(2):159–174.

Petr Plecháč and Robert Kolár. 2015. The corpus of Czech verse. *Studia Metrica et Poetica*, 2(1):107–118.

Brian Porter and Edouard Machery. 2024. AI-generated poetry is indistinguishable from human-written poetry and is rated more favorably. *Scientific Reports*, 14(1):26133.

François Rastier. 2009. *Sémantique interprétative*, 3rd edition edition. Presses Universitaires de France.

Gaurav Sahu and Olga Vechtomova. 2025. Computational modeling of artistic inspiration: A framework for predicting aesthetic preferences in poetic lines using linguistic and stylistic features. In *Proceedings of ICCC25*.

Sakib Shahriar. 2022. GAN computers generate arts? a survey on visual arts, music, and literary text generation using generative adversarial network. *Displays*, 73:102237.

Milan Straka and Jana Straková. 2017. Tokenizing, POS tagging, lemmatizing and parsing UD 2.0 with UDPipe. In *Proceedings of the CoNLL 2017 Shared Task: Multilingual Parsing from Raw Text to Universal Dependencies*, pages 88–99, Vancouver, Canada. Association for Computational Linguistics.

André Valença and Filipe Calegario. 2025. Experimenting with large language models for poetic scansion in Portuguese: A case study on metric and rhythmic structuring. In *Proceedings of ICCC25*.

Jianli Zhao and Hyo Jong Lee. 2022. Automatic generation and evaluation of Chinese classical poetry with attention-based deep neural network. *Applied Sciences*, 12(13).