EduPo: Progress and Challenges of Automated Analysis and Generation of Czech Poetry

Rudolf Rosa and **David Mareček** and **Tomáš Musil**
and **Michal Chudoba** and **Jakub Landsperský**
Charles University, Faculty of Mathematics and Physics
Institute of Formal and Applied Linguistics
Praha, Czechia
rosa@ufal.mff.cuni.cz

Abstract

This paper explores automated analysis and generation of Czech poetry. We review existing tools, datasets, and methodologies while considering the unique characteristics of the Czech language and its poetic tradition. Our approach builds upon available resources wherever possible, yet requires the development of additional components to address existing gaps. We present and evaluate preliminary experiments, highlighting key challenges and potential directions for future research.

1 Introduction and Related Work

In Natural Language Processing (NLP), there is a small but permanent interest in dealing with poetry, as its unique features make it rather different from most other texts and thus more challenging in some aspects than other genres (Gonçalo Oliveira, 2017). In particular, the strong importance of formal properties, intertwined with semantic content of the text, makes it impossible to simply apply standard domain adaptation techniques to fit general-domain systems to poetry; instead, poetry-specific approaches need to be used.

Even in the times where large language models (LLMs) are gradually becoming the solution to most NLP tasks, often with no or little training required, the situation with poetry is different: while standard off-the-shelf LLMs can be used to analyze some properties of poems (typically semantic ones) as well as to possibly generate poetry of reasonable quality in English and a few other major languages, usefulness of vanilla LLMs for poetry in many languages is poor (Shao et al., 2021; Hämäläinen et al., 2022; Sawicki et al., 2023; Porter and Machery, 2024). We believe this is due to the focus of the LLMs on meaning rather than form, as exemplified by their low performance at even simple form-based tasks, such as counting characters in words (Xu and Ma, 2025). This is at least partially due to inadequate tokenization, as the poetry-relevant units (such as syllables) do not correspond well to the LLM subwords (Wöckener et al., 2021), encouraging the use of syllable-based tokenization (Oncevay and Rojas, 2020) or tokenization-free approaches (Belouadi and Eger, 2023). The latter work also reveals another shortcoming of standard LLMs, which is the fact that many poetry-relevant features of the text, such as stress, are not directly apparent to the LLM, and performance can thus be greatly improved by revealing such features via automatically generated annotations of the data.

In our work, we focus on automated generation of Czech poetry, with poetry analysis as an indispensable component for automated data annotation and evaluation.

While there is a range of attempts at generating poetry in several major languages (Piorecký and Husárová, 2024, chapter 5), we are not aware of any substantial work on generating Czech poetry since Neverilová and Pala (2015) and Materna (2016). We thus mostly base our approach on works focusing on other languages, and adapt and extend these approaches for the specifics of the Czech setting.

On the other hand, there has been extensive work on automated *analysis* of Czech poetry, centered around the Květa tool by Plecháč (2016).[1] We thus use Květa as the basis for our analyses, identifying and rectifying some of its shortcomings as well as implementing several missing components. We also take inspiration from the alternative approach to metre detection by Klesnilová et al. (2024).

There has also been some work on automatically identifying themes and motives present in Czech poetry (Bendík, 2023; Kořínková et al., 2024); however, the reported results were mostly negative, concluding that the chosen methods for theme and motive identification do not yield satisfactory results. We thus attempt to solve the problem by using different methods.

[1] And related tools developed by the same team: https://versologie.cz/v2/web_content/tools.php?lang=en

Proceedings of the 5th International Conference on Natural Language Processing for Digital Humanities, pages 524–542
May 3-4, 2025 ©2025 Association for Computational Linguistics

Regarding the theory of Czech verse and Czech poetry, we mainly base our approach on the works of Ibrahim et al. (2013) and Plecháč and Kolár (2017), which had also been the basis for Květa and for the KČV poetry corpus which we use.

Automated evaluation of the quality of generated texts is a long-standing problem which still lacks complete and satisfactory solutions (Schmidtová et al., 2024). In generated poetry, we are interested in some rather standard qualities of text, such as correct grammatical structures and meaningfulness, but already these standard qualities are complicated by the fact that various language constructions, unacceptable in standard writing, can be allowed or even encouraged in poetry (e.g. non-standard word order or creatively deriving new words). Besides that, we would ideally also like to assess some other literary values, such as creativity, beauty, etc., where even human agreement is low; although some automated approaches are appearing, such as the recent work on evaluating novelty of texts by Lu et al. (2025). On the other hand, many of the formal properties of poetry (such as metre or rhyming) are quite rigidly defined and thus rather easy to evaluate automatically (although we need to keep in mind that human authors typically do not follow the rules perfectly).

We discuss specifics of the Czech language and Czech poetry in Section 2, we present our approaches to analyzing Czech poetry in Section 3, and we describe and evaluate our experiments in poetry generation in Section 4. As this paper presents ongoing work, we also discuss a range of plans for future work throughout the paper.

All our source codes and models are publicly available under permissive licences.[2][3] A live beta-version demo of our tools is also available online;[4] screenshots are attached in Appendix G. Some of our experiments have already been described in (Chudoba and Rosa, 2024).

The main practical motivation for our work, within a broader project titled EduPo,[5] is to develop an interactive educational application to be used in teaching about poetry in Czech schools; however, we do not discuss this axis of our work in more detail here, as we find this out of scope for the target reader, and we thus focus solely on the NLP aspects of our work in this paper.

2 Specifics of the Czech Setting

In this section, we discuss several specifics of dealing with Czech poetry.

2.1 Large Corpus of Poetry (KČV)

There exists a very large poetry corpus, the Corpus of Czech Verse[6] (*KČV*, Korpus českého verše) by Plecháč and Kolár (2015), which is freely available and contains 80,229 Czech poems.[7]

The poems in KČV are annotated with various metadata (author, book, publishing year, etc.), versological features (metre and rhythm, rhyming, stanzaicity and stanzas, poetic forms), phonetics (phonetic transcription), and morphology (lemma, part of speech, morphological features).

Most of the features are pre-annotated automatically using Květa and then manually checked and corrected. The annotations can thus be rather reliably used for analyses, model training, and automated evaluation.[8] We use the KČV corpus as our dataset for all experiments.[9]

The KČV only contains poems with expired copyright, thus mostly coming from the 19th century and the beginning of the 20th century. There is an ongoing project of collecting and annotating contemporary poetry (Škrabal and Piorecký, 2022), which we intend to use in our future work.

2.2 Phonetic Transparency

Czech orthography is very regular and rather close to phonetics. Therefore, rule-based approaches can be used to obtain phonetic transcriptions, with only a small amount of harder ambiguous phenomena (such as diphthongs; see Section 2.3). Still, our experiments revealed that foreign words are rather common in Czech poetry (mostly named entities), usually using their original foreign spelling, which means that the results of the rule-based phonetic transcription are unreliable in such cases.

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

[3] http://hdl.handle.net/11234/1-5871

[4] https://quest.ms.mff.cuni.cz/edupo/

[5] https://ufal.mff.cuni.cz/grants/edupo

[6] https://versologie.cz/v2/web_content/corpus.php?lang=en

[7] 2 664 989 lines, 14 592 037 words

[8] At the same time, Plecháč and Kolár (2015) admit (and our experience confirms) that an unknown amount of pre-annotation errors slipped the manual checks and are still part of the corpus, which needs to be taken into account when interpreting any evaluations against these annotations.

[9] We do not re-publish the dataset as it is freely available. We intend to release an enriched version of the dataset in future once we enhance it by adding further automated annotations not present in the original dataset.

2.3 Elusive Syllables

While the concept of syllables is generally accepted for Czech language and syllables are important units for poetry, there is no universal agreement on the syllable definition and *boundaries* (Bičan, 2013; Šturm and Bičan, 2022). The available syllable splitting tools, such as Sekáček (Macháček, 2014), are not very reliable, and we also have not been aware of any datasets with the necessary annotation to train our own splitter. There is a Czech 'PhonCorp' lexicon annotated with phonological features, including syllable boundaries (Bičan, 2015a,b), published already 10 years ago but made publicly available only recently.[10]

The *number of syllables* is easier to get, determined by the number of syllable nuclei – typically vowels (possibly diphthongs) and vocalic consonants. Květa provides an indirect estimate for the number of syllables in a word, but it does not take diphthongs into account,[11] and handles most but not all cases of vocalic consonants.[12]

2.4 Weak Regular Stress

The prosodic stress in Czech language is rather weak and difficult to directly map onto any explicit acoustic qualities of speech (Janota, 1967). However, there is a widely accepted tradition of regular stress placement, which is mostly respected in classical Czech poetry.

In standard Czech (and especially in poetry), stress is traditionally placed on the first syllable of each polysyllabic word (and never on subsequent syllables of the word). Monosyllabic words can be generally regarded as stressed or unstressed as required by the metre of the poem, with a preference of stressing content words and not stressing auxiliary words (but author styles differ in the preferences of stressing of monosyllabic words). Additionally, for words immediately preceded by a monosyllabic preposition, the stress is traditionally moved from the word onto the preposition (and the whole polysyllabic word is left unstressed). Thus, simple rule-based approaches can be used to distinguish stressed and unstressed positions (provided that part-of-speech information is available to distinguish prepositions and ideally also content/auxiliary words).

2.5 Limited Variety of Metres

The properties of Czech prosodic stress implicate that the range of meter types available to Czech poets is rather limited. Traditionally, six (syllabotonic) basic meter types are recognized for Czech poetry (and annotated in KČV), with only three of them being common: iamb (*J*), trochee (*T*), and dactyl (*D*).[13] Trochee (strong-weak)[14] and dactyl (strong-weak-weak)[15] are straightforward to achieve within the Czech stress patterns. Iamb (weak-strong) is realized either by initiating the verse with a monosyllabic word,[16] or by starting the verse with a three-syllable word (dactyl incipit).[17]

2.6 Rhyming and Reduplicants

Verses rhyme with each other if their reduplicants are sufficiently phonetically similar.

Traditionally, the reduplicant (i.e. the rhyming part of the verse) in Czech poetry is defined as the sequence of phones from the penultimate syllable nucleus till the end of the verse. However, if the last word of the verse is monosyllabic, the reduplicant starts either with the last nucleus (in case of a closed verse, ending with a consonant), or with the consonant preceding the last nucleus (in case of an open verse, ending with a nucleus).

Theory of Czech rhyme is rather vague in terms of defining the phonetic similarity of the reduplicants, often listing tendencies rather than hard rules and allowing a lot of freedom to the individual style and preferences of the poet.

3 Automated Analysis of Poetry

We have built a poetry analysis framework that takes a plaintext poem as input (one verse per line, empty lines separating stanzas), performs a sequence of automated analyzes, and produces annotations of the poem text in JSON format. The annotations include phonetic transcriptions, syllabic features, morphological and syntactic features, versological annotation of reduplicants, rhymes, stresses,

[10] https://www.phil.muni.cz/phoncorp/

[11] In Czech, we cannot distinguish diphthongs ('au', 'ou', 'eu') from separate vowels ('a-u', 'o-u', 'e-u') based on orthography. The distinction could be made based on phonetic transcription, but none of the phonetic transcription tools that we tried does distinguish these cases.

[12] Several consonants are potentially vocalic and thus can form the nucleus of a syllable: typically 'r' and 'l' (but only in some words), possibly 'm' and 'n' in some cases, and very rarely also a few other consonants such as 's', 'š' or 'z'.

[13] In KČV, 98% of metric verses pertain to iamb (54%), trochee (41%) and dactyl (3%).

[14] E.g. *__Prav__-da __prav__-da __dál__ by __rá__-di*

[15] E.g. *__ná__-ro-dy __ži__-jí jen __o__-svě-tě*

[16] E.g. *Já __ne__-vím __chvím__ se __od__-va-ha mně __mi__-zí*

[17] E.g. *ne-zná-mou __to__-bě __ci__-zí __spi__-rá-lu*

and metres, motives of the poem, and stylometric analysis. In future, we also plan to try identifying some poetic forms (such as a sonnet or a limerick), some figures of speech (some schemes, such as alliterations or anaphoras, and possibly also some tropes, such as metaphors), and probably also some euphonic qualities.

The framework is built on top of Květa (Plecháč, 2016) as its backbone, with many improvements and complements as needed, and uses UDPipe (Straka and Straková, 2017) to provide morphological and syntactic analyses.

For simplicity, most of the analyses are largely context-independent, which is sufficient in typical cases, but fails to fully correctly cover all situations. Often, multiple ways to analyze the same part of the poem are theoretically possible, and the context of the neighbouring phones, words, verses, or of the whole poem, should be taken into account to correctly select the most adequate variant of the analysis within the given context.[18] For future work, we envision a solution that would keep some analyses ambiguous at certain stages of the processing and disambiguate them through post-processing.

While the analyses may also be useful on their own, we use them to automatically annotate training data and to define evaluation measures.

3.1 Phonetic transcription

The phonetic transcription is rule-based, based on the implementation in Květa, using the Czech Phonetic Transcription (ČFT) formalism.[19] However, the existing method does not properly handle diphthongs and foreign words, and we also complemented it by adding missing vocalic consonants.

Diphthongs Květa does not distinguish between diphthongs and separate vowels (e.g. *ou/o-u* as in *proudit* which could be either *prou-dit* or *pro-u-dit*). As this is a crucial distinction for determining the number of syllables, which in turn is vital for the metre, we designed and implemented a diphthong disambiguation tool. The training data were obtained from the KČV and PhonCorp, which both contain phonetic transcriptions capturing this distinction. We use 'patgen',[20] a tool originally developed for generating TEX hyphenation patterns, to generate efficient patterns for distinguishing diphthongs from separate vowels, and 'hyphenator'[21] to apply the learned patterns to words. The patgen algorithm ensures that all word forms present in the training data are handled correctly, while also generalizing to some word forms not present in the training data. This approach is context-independent and thus cannot distinguish homonyms that differ in the diphthong vs. 2 vowels pronunciation, but these are very rare in Czech.

Foreign words We also complemented the existing method with an automatically built list of foreign characters and words and their phonetic transcriptions, extracted from the KČV corpus. However, we found that our straightforward solution is not completely satisfactory, as there is a sort of intentional ambiguity: for many foreign words, their Czech pronunciation is not completely stable, and poets actively utilize this flexibility to fit the desired rhyming and syllable count.[22] Therefore, a correct phonetic transcription is only achievable with taking the context of the neighbouring verses into account; we leave this for future work.

We use the UDPipe morphological lexicon to define the poem-level measure of **Unknown words** as the proportion of words not present in the lexicon.

3.2 Syllables

Since determining the syllable boundaries is not easily achievable, we only focus on determining the syllable count in each word.[23] We use the phonetic transcription of the word, with diphthongs and vocalic consonants already marked as (single) vowels; thus, the number of syllables is equal to the number of vowels.

A slight but easy-to-solve complication are nonsyllabic prepositions (*k*, *s*, *v*, *z*), which need to be conjoined with the following word in preprocessing (e.g. *k letišti*: *kle-tiš-ti*).

A harder complication, which we have not solved yet, are shortcuts, whose pronunciation is

[18] E.g. the number of syllables intended by the author in ambiguous cases can often be determined from the metrical properties of other verses and/or the regularity of syllable counts within stanzas. Or, whether two verses are to be treated as rhyming or non-rhyming in some edge cases can often be determined by the regularity of the rhyme scheme within the stanza or across stanzas.

[19] https://versologie.cz/v2/web_content/phoebe.php?lang=en

[20] https://ctan.org/pkg/patgen

[21] https://github.com/tensojka/cshyphen

[22] E.g. *Baudelaire* can be easily split into either 2 or 3 syllables – *Bau-de-laire* or *Baude-laire* – or even 4 syllables if needed – *Bau-de-lai-re*; all these variants are attested in KČV.

[23] For future work, we consider the possibility of automatic syllable splitting using Optimality Theory (Prince and Smolensky, 1993).

ambiguous (e.g. 1-syllable *FOK* vs. 3-syllable *F-O-K*) and largely based on conventions which are not recorded by any resource known to us. Additionally, poets take the liberty of bending the rules and conventions as needed within the context of the poem, which must be taken into account.

We use syllable counts to define two evaluation measures: **Syllable count entropy** is the entropy of syllable counts across verses of a poem, and **Syllable accuracy** is the proportion of generated verses that adhere to a pre-specified syllable count.

3.3 Rhyme

The rhyming component in Květa is based on RhymeTagger (Plecháč, 2018). It identifies rhyming verses by estimating the probability of the verse reduplicants rhyming with each other.

We additionally implemented rule-based reduplicant marking from scratch according to the rhyming theory as explained in Section 2.6.

We use rhyming to define two evaluation measures: **Rhyming** (poem-level) is the proportion of verses that rhyme with other verses in a selected window, and **Rhyme accuracy** (corpus-level) is the proportion of generated poems that adhere to the rhyme scheme specified on input.

3.4 Metre

Květa identifies the metre of the poem based on the stressed syllables automatically assigned in a rule-based way, scoring the compatibility of each verse with each of potential metres, averaging the compatibility scores over all verses in the poem, and returning the highest scoring metre.[24]

We use metre to define two evaluation measures: **Metre consistency** (poem-level) is the score of the highest scoring metre, and **Metre accuracy** (corpus-level) is the proportion of generated poems that adhere to the metre specified on input.

3.5 Motives

As the previous approaches on identification of poetry motives in Czech poetry were mostly unsuccessful (Kořínková et al., 2024), we take a different approach, instructing a LLM (`gpt-4o-mini`)[25] to identify up to 5 main themes of the poem (in practice, the LLM seems to always return *exactly* 5 motives); the full prompt is shown in Appendix A.

Label	OK	DEL	EDIT	ADD
Motives abs.	286	66	24	26
Motives rel.	76%	18%	6%	7%
Avg. per poem	3.8	0.9	0.3	0.3

Table 1: Manual evaluation of automatically generated motives for 75 poems (5 motives per poem).

Exploratory experiments evaluated by poetry experts showed that considerably better results are achieved by open-ended motive identification, rather than using a predefined list of allowed motives from Bendík (2023), which leads to less informative results. However, we did not find a strong influence of using a Czech or English prompt, or of machine-translating the poem text into English.

We then performed a manual evaluation of automatically generated motives for 75 poems, split among 3 poetry experts as annotators. They annotated each motive as correct ('OK'), superfluous ('DEL'), or partially correct ('EDIT'), and they could also mark a missing motive ('ADD'). A summary of the evaluation results is shown in Table 1; more details can be found in Appendix A, and examples of automatically identified motives (for generated poems) are included in Appendix E.[26]

The method is rather solid, with most (76%) of the identified motives being correct; additionally, for 32% of the poems, all 5 motives were marked as correct. This confirms that LLMs may struggle with formal aspects of poetry, but are well suited for semantic tasks. The most common error reported by the annotators is a surplus motive, suggesting that it would be useful to design a post-processing step to check and remove (and potentially also edit) some of the motives. On the other hand, 5 motives proved to be a sensible maximum (only for 3 poems, 6 motives were suggested by the annotators).

3.6 Stylometric Analysis

Stylometry is used to attribute authorship of a given text (Plecháč, 2021). In our setting, we use it to estimate author styles, and to measure whether the generated poems successfully imitate a certain author.

We use the sentence embedding architecture SBERT (Reimers and Gurevych, 2019) which re-

[24] Květa does not try to detect polymetric poems, but these are very rare in KČV.

[25] `https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/`

[26] We did not carry out an evaluation of automatically identified motives for generated poems, as manual motive assignment is a laborious task even for high-quality human-written poems, and a hard and frustrating task on the poems generated by our models.

turns an embedding vector for a given text (in our case a poem enriched with number of syllables, metre, and rhyme annotation). Specifically, we use the Robeczech model (Straka et al., 2021), which is pre-trained on Czech texts and further finetune it on examples of poem triplets, where two are always by the same author and one is by a different author.

Once we have a vector for each poem, we can simply measure euclidean distances between any two poems. We use KNN method for the authorship attribution. For a given poem, we can find k nearest poems with known authors (we use $k = 5$) and predict the most frequent author among the k as the poem author.[27]

The efficiency and accuracy of author prediction depends on the number of authors among which we are choosing. For our preliminary experiments, we chose a set of 12 well-known authors with distinct styles.[28] Using a leave-one-out method (train/test split with the ratio of 9:1), we measure the accuracy of the proposed method as 74% on this subset of KČV; it is thus already useful in practice, but still needs further improvements.

We use our current stylometry to define the evaluation measure of **Style accuracy** as the proportion of generated poems where the predicted author is identical to the author specified on input.

4 Automated Generation of Poetry

Our generation approach consists of enriching the plain texts of the poems with relevant annotations and fine-tuning a LLM on the dataset. At inference, desired parameters of the poem to be generated are transformed into a prompt structured in the same way as the annotations in the training data.

So far, we have performed two sets of exploratory experiments in poetry generation (referred to as *first set* and *second set*), experimenting with base model choice, data formatting, and tokenization. The best performing model in each of these sets is further referred to as *first model* and *second model*, respectively.

The first model is released on HuggingFace as *jinymusim/gpt-czech-poet*,[29] together with other models from the first set of experiments which use different tokenizations.[30]

The second model is released on HuggingFace as *tomasmcz/edupo_v0.5*.[31]

Examples of generated poems are attached in the Appendix E.

4.1 Data Deduplication

KČV often contains multiple copies of the same poem, typically with some slight variations of formatting, text, title, and/or segmentation. This creates data imbalance, interferes with our stylometry experiments, and would cause further issues when measuring novelty/plagiarism.

We detect and remove duplicates following Plecháč et al. (2023), computing Levenshtein distances of all poem instances for each author. Additionally, we use Akin[32] to also find duplicates attributed to different authors.

4.2 Fine-tuned Base LLM

For the *first set*, we fine-tuned a Czech GPT-2 model[33] by Chaloupský (2022).[34] Due to the limited context size of the model, we limited these experiments to individual stanzas of 4 or 6 verses.[35] We found the model to generate poems which are mostly good in terms of the formal properties (rhyming, metre, number of syllables), but low-quality in terms of meaning, often forcefully generating completely non-sensical text to fulfill the desired formal properties.

For the *second set*, we switched to Llama-3.1 (Grattafiori et al., 2024), which allows us to train on full poems and yields better results also in terms of meaning. Llama-3.1 is a multilingual model with a very good performance on Czech language in comparison to other freely available models, as attested in BenCzechMark (Fajcik et al., 2024).[36] We use the whole KČV corpus for training the

[27] In case of tie, we take the author with lower average distance from the poem.

[28] Auředníček, Březina, Čelakovský, Dostál-Lutinov, Dyk, Erben, Hálek, Kollár, Mácha, Neruda, Puchmajer, Zeyer

[29] https://huggingface.co/jinymusim/gpt-czech-poet

[30] https://huggingface.co/jinymusim/gpt-czech-poet-base, https://huggingface.co/jinymusim/gpt-czech-poet-our, https://huggingface.co/jinymusim/gpt-czech-poet-syllable, https://huggingface.co/jinymusim/gpt-czech-poet-unicode.

[31] https://huggingface.co/tomasmcz/edupo_v0.5

[32] https://github.com/justinbt1/Akin

[33] Although significantly older and less capable than current LLMs, we still find GPT-2 to be useful for preliminary experiments, as it is quick and cheap to fine-tune.

[34] https://huggingface.co/lchaloupsky/czech-gpt2-oscar

[35] The resulting subset of KČV, which we used to train the first set of models, consists of 374,537 stanzas (composed of 2,310,917 verses); we use 95% of the dataset as training data and the remaining 5% as test data.

[36] https://huggingface.co/blog/benczechmark

second set of models.[37] We use LoRA (Hu et al., 2021) with Unsloth (Han et al., 2023) to fine-tune the models.

More details on the model fine-tuning and hyperparameters are included in Appendix C.

4.3 Data Formatting

It is highly beneficial for poetry modelling to enrich the training data with explicit versological annotations, which helps the model by making the relevant properties overt (Belouadi and Eger, 2023). Moreover, we also need to encode the desired parameters into a prompt for the model generation to follow; at inference, any of the parameters can be specified by the user and inserted into the prompt during the generation process, or left for the model to 'decide'.

We show here two formatting schemes we used. Examples of a poem formatted according to the two formats are enclosed in Attachment B.

In the *first set*, we tried out several formats, eventually settling for:

```
# rhymescheme # year # metre
syllables # reduplicant # verse
syllables # reduplicant # verse
...
```

The rhyme scheme, publication year (as a proxy for style), and metre are included as input parameters for the generation. Explicitly marking the number of syllables and the reduplicant string of each verse proved to be crucial hints for the model; without them, the rhyme accuracy drops tremendously (49.6% compared to 86.9%).

For the *second set*, we slightly modified the format to be more regular, and also to include the name of the author and the title of the poem. We also decided to annotate the metre at each verse independently to support polymetric poems:

```
authorname: poemtitle (year)

# rhymescheme #
# metre # syllables # reduplicant # verse
# metre # syllables # reduplicant # verse
...
```

With unspecified author name, the model often generates texts that do not follow the format. In future experiments, we plan to counter this by introducing a sequence of tokens at the beginning of the prompt to indicate the format of the poem.

Tokenizer	Syll. acc.	Rhyme acc.	Metre acc.
Original	92.3%	86.9%	94.5%
Our BPE	91.0%	80.6%	94.8%
Our syll.	94.4%	88.7%	94.6%
Our char.	97.8%	94.0%	94.0%

Table 2: Effect of tokenization on accuracy of adhering to the specified syllable count, rhyme scheme and metre, evaluated within the first set of generation experiments.

We also plan to experiment with various formats of the reduplicant. In the current format, the reduplicant field contains the ending of the text that follows on that line. It may be better to supply the model with the reduplicant of the previous verse that the current line is supposed to rhyme with, according to the rhyme scheme.

The data annotations, and thus possible input parameters, reflect the analyses which are annotated in KČV and/or which we are already able to automatically produce with sufficient accuracy. For other useful annotations (e.g. poem motives), we first need to develop a sufficiently accurate analysis method and use it to automatically annotate the corpus; then such parameters can be included into the generation process.

4.4 Tokenization

In the first set of experiments, we compared several tokenization strategies:

1. use the original (Czech) tokenizer of the LLM,
2. train a BPE tokenizer on our training data,
3. use a syllable splitter as a tokenizer,[38] inspired by Oncevay and Rojas (2020),
4. tokenize the text into individual characters, inspired by Belouadi and Eger (2023).

Unless the original tokenization was used, we needed to refit the base model to the new tokenization before fine-tuning it on the dataset; we used model recycling of de Vries and Nissim (2021).

Table 2 compares the four tokenization setups in terms of accuracy of adhering to the specified number of syllables, rhyme scheme, and metre, measured on 3,321 poems generated with inputs sampled from KČV.

We did not find any benefit in exchanging a general-domain Czech subword tokenizer for a

[37] For training the second set of models, we do not split off a test set as we do not perform any evaluations of the trained model that require a test set.

[38] We used Sekáček (Macháček, 2014).

BPE tokenizer trained on Czech poetry; we rather observe a deterioration, which may be due to loss of information from pretraining, as the token overlap of the vocabularies is only 33%.

We found that the syllable-based and character-based tokenization leads to higher Syllable and Rhyme accuracies, while having no effect on Metre accuracy, which is already quite high with the original tokenization. However, a small-scale manual evaluation suggested that these improvements are at the expense of meaningfulness, with the sensibility of the generated poems being notably worse for the syllable-based tokenization than for the subword-based tokenization, and still much worse for the character-based tokenization. We thus decided to settle for the original tokenizer in the first model.

We still believe that syllable-based tokenization seems highly suitable for poetry, but we found it is not straightforward to use due to various reasons. The vocabulary token overlap with pretrained models is low (19% in our experiments, with many frequent longer tokens missing in the intersection), which means that a lot of information from the pretraining is lost. There is also the problem of no reliable syllable splitter being available for Czech; we would thus need to devise such a tool. We were also expecting the frequency distribution of syllables to be less balanced than the distribution of standard subwords, which could prevent the model from properly learning the meanings of the tokens (Zouhar et al., 2023); however, at least for Czech, we found this not to be an issue, with the frequency distributions of subwords and syllables being rather similar (see Appendix D for an analysis).

In our second set of experiments, we used the original tokenizer which is part of the (non-Czech) base model. Our future plan is to switch to Czech-specific tokenization; while subword tokenization is still the standard, our results encourage us to also explore syllable-based tokenization, as well as tokenizer-free approaches (Xue et al., 2022; Deiseroth et al., 2024). However, our experiments suggest that after refitting the tokenizer, the loss of information from pretraining is too large and requires fine-tuning the refitted model not only on the (medium-sized) poetry corpus, but also on much larger general-domain Czech data.

4.5 Comparison to KČV Corpus

We evaluated the *second model* by comparing distributions of values of 5 evaluation measures computed on a sample of 94 generated poems versus the poems in KČV. The main results are presented here, with some additional details in Appendix F.

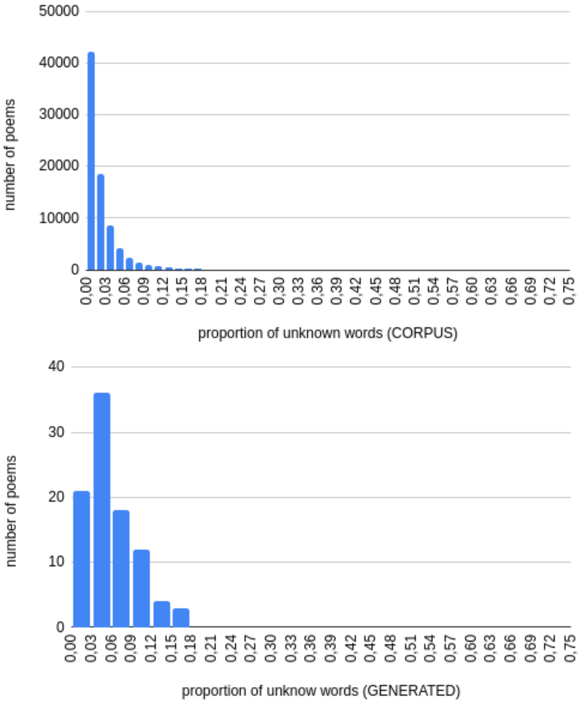


Figure 1: Histogram of unknown words proportions in the corpus and in the generated poems.

The histogram (Figure 1) of values of the Unknown words measure (defined in Section 3.1) shows that the generated poems typically contain slightly more unknown words (around 5%) than typical real poems. We have observed that the model is able to create novel words, which is generally acceptable in poetry; however, human poets tend to create novel words which are understandable to the reader, whereas most of the novel words created by the model are not understandable.

Figure 2 evaluates Rhyming (Section 3.3). In the corpus, we clearly observe fully-rhymed poems (around 1.0), half-rhymed poems (around 0.5, e.g. XAXA[39]), and poems not rhymed at all. On the contrary, the model most often produced poems with 10%-20% non-rhymed verses, as well as a substantial but lower amount of fully-rhymed poems, and no non-rhymed poems. We believe that this either shows that the model primarily 'tries' to generate fully-rhymed poems (which is the most frequent type) but is imperfect at rhyming; or that it did not learn the concept of distinct regular rhyming patterns on the level of poems and thus 'tries' to

[39] A, B, C etc. mark rhyming verses in the rhyme scheme, while X marks non-rhyming verses.

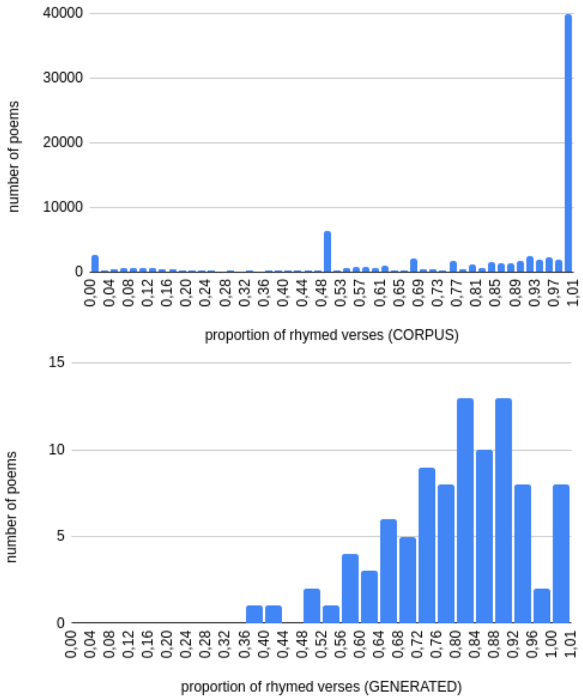


Figure 2: Histogram of proportions of rhymed verses in the corpus and in the generated poems.

produce something between fully-rhymed and half-rhymed poems. This requires further investigation.

Similarly, Syllable count entropy (Section 3.2) of the generated poems is higher, showing that the syllable counts are not as regular as in the corpus poems (Figure 4 in Appendix F). This might have similar reasons to the rhyming irregularities.

On the other hand, Metre consistencies (Section 3.4) are similar both for poems from the corpus and for generated poems (Figure 5 in Appendix F), suggesting that the model managed to learn the aspects of metre.

The measured Style accuracy (Section 3.6) of the generated poems, computed using the selected 12 authors, is 28%. This is much lower than the 74% accuracy on KČV, but still much higher than the random chance at 8%. The model thus already shows some limited success in imitating author styles, but further effort is needed.

We have tried performing exploratory small-scale manual evaluations of qualities such as meaningfullness, poeticity or overall quality, but the evaluation yielded very inconclusive results with stark disagreements among the annotators.[40]

We also plan to measure word/token n-gram overlap of the generated poetry with the training data as a measure of novelty (Lu et al., 2025).[41]

5 Conclusion

In this paper, we presented our ongoing effort of devising a comprehensive framework for automated analysis and generation of Czech poetry. Our approach is largely based on existing tools and datasets and on methods described for other languages, but still faces numerous issues, pertaining to various imperfections and omissions of the available tools and datasets, as well as to specific properties of Czech language and Czech poetry.

We described a range of improvements to existing tools as well as newly designed and implemented components. We also performed various evaluations, shedding light on the tasks and the performance of the proposed methods, as well as at language generation and Czech poetry in general.

The current state of our work leaves many open opportunities for future research and improvements, which we discussed throughout the paper.

Limitations

The paper reports on ongoing research, therefore, many aspects are not yet final and many evaluations are rather indications than hard evidence. Especially, proper manual evaluation of meaningfulness of the generated poems is vital, but so far only has been performed in a preliminary way due to encountered issues with defining the evaluation criteria and explaining them to annotators.

The paper only deals with Czech language and Czech poetry, and we do not claim any language-independence or applicability to other languages. We hope that the proposed methods could be applicable to other languages with similar poetry traditions (such as Slovak), but we have not evaluated that in any way.

The size of models we can train is limited by the computational power available to us. It can be presupposed that by fine-tuning larger base models, better results could be achieved.

[40] Apparently, particular care needs to be taken when designing the manual evaluation, as the desired qualities are not universally understood and somewhat hard to define and explain to annotators. Once we manage to devise a proper manual evaluation scheme, we will also attempt to measure some of these aspects automatically.

[41] This will also be useful once we enrich our training corpus with contemporary poetry, where we will need to ensure that the generated poetry does not infringe upon the copyright of the poem authors by leaking sequences of considerable length directly copied from the poems.

Ethics Statement

We are currently only using poems with expired copyright to train our models. Once we move on to also using copyrighted materials to train our models (which we by itself believe to be acceptable under the research exemptions to copyright law), we will ensure that the generated poems do not constitute inacceptable infringments to the copyright of the poem authors by excessively copying from the copyrighted poems present in the training data.

We also make sure to always explicitly label all the generated poems as automatically generated.

While such concerns have already been raised towards us, we do by no means intend to replace human poets. On the contrary, our broader goal is to develop an interactive educational application, with which we hope to raise the interest in poems and encourage more people to actively interact with poetry.

We have been using GPT and Llama LLMs as base models. It is beyond our control to what extent these models had been created in an ethical way. However, we believe it is more ethical environmentally to fine-tune pretrained models than to train new models from scratch, as this would require a substantially larger amount of computation. In case the consensus becomes that some base models are unethical and it is unethical to use them, we will switch to using different base models.

We are tracking the approximate amounts of computational power used to train our models so that we can estimate the environmental impact of our experiments.

Acknowledgements

The work has been supported by the EduPo grant (TQ01000153 Generating Czech poetry in an educative and multimedia environment), which is cofinanced from the state budget by the Technology agency of the Czech Republic under the SIGMA DC3 Programme. 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

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.

Martin Bendík. 2023. Automatická detekce témat v básnických textech. Master's thesis, České vysoké učení technické v Praze. Vypočetní a informační centrum.

Aleš Bičan. 2013. *Phonotactics of Czech*. PL Academic Research, Imprint der Peter Lang GmbH.

Aleš Bičan. 2015a. Corpus-based analysis of the Czech syllable. *Beiträge der Europäischen Slavistischen Linguistik (POLYSLAV)*, 18.

Aleš Bičan. 2015b. Fonologický lexikální korpus češtiny a slabičná struktura českého slova. *Bohemica Olomucensia*, 7(3-4):45–59.

Lukáš Chaloupský. 2022. Automatic generation of medical reports from chest X-rays in Czech. Master's thesis, Univerzita Karlova, Matematicko-fyzikální fakulta.

Michal Chudoba and Rudolf Rosa. 2024. GPT Czech poet: Generation of Czech poetic strophes with language models.

Wietse de Vries and Malvina Nissim. 2021. As good as new. How to successfully recycle English GPT-2 to make models for other languages. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 836–846, Online. Association for Computational Linguistics.

Björn Deiseroth, Manuel Brack, Patrick Schramowski, Kristian Kersting, and Samuel Weinbach. 2024. T-FREE: Subword tokenizer-free generative LLMs via sparse representations for memory-efficient embeddings. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21829–21851.

Martin Fajcik, Martin Docekal, Jan Dolezal, Karel Ondrej, Karel Benes, Jan Kapsa, Michal Hradis, Zuzana Neverilova, Ales Horak, Michal Stefanik, Adam Jirkovsky, David Adamczyk, Jan Hula, Jan Sedivy, and Hynek Kydlicek. 2024. BenCzechMark: A Czech-centric multitask and multimetric benchmark for language models with duel scoring mechanism.

Hugo Gonçalo Oliveira. 2017. A survey on intelligent poetry generation: Languages, features, techniques, reutilisation and evaluation. In *Proceedings of the 10th International Conference on Natural Language Generation*, pages 11–20, Santiago de Compostela, Spain. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. 2024. The llama 3 herd of models.

Mika Hämäläinen, Khalid Alnajjar, and Thierry Poibeau. 2022. Modern French poetry generation with RoBERTa and GPT-2. In *13th International Conference on Computational Creativity (ICCC) 2022*.

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.

Robert Ibrahim, Petr Plecháč, and Jakub Říha. 2013. *Úvod do teorie verše*. Akropolis.

Přemysl Janota. 1967. An experiment concerning the perception of stress by Czech listeners. *Acta Universitatis Carolinae-Philologica, Phonetica Pragensia I*, pages 45–68.

Kristýna Klesnilová, Karel Klouda, Magda Friedjungová, and Petr Plecháč. 2024. Automatic poetic metre detection for Czech verse. *Studia Metrica et Poetica*, 11(1):44–61.

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.

Ximing Lu, Melanie Sclar, Skyler Hallinan, Niloofar Mireshghallah, Jiacheng Liu, Seungju Han, Allyson Ettinger, Liwei Jiang, Khyathi Chandu, Nouha Dziri, and Yejin Choi. 2025. AI as humanity's Salieri: Quantifying linguistic creativity of language models via systematic attribution of machine text against web text.

Dominik Macháček. 2014. Sekacek. `https://github.com/Gldkslfmsd/sekacek`.

Jiří Materna. 2016. *Poezie umělého světa*. Backstage Books.

Zuzana Neverilová and Karel Pala. 2015. Generating Czech iambic verse. In *RASLAN*, pages 125–132.

Arturo Oncevay and Kervy Rivas Rojas. 2020. Revisiting neural language modelling with syllables. *CoRR*, abs/2010.12881.

Karel Piorecký and Zuzana Husárová. 2024. *The culture of neural networks: Synthetic literature and art in (not only) the Czech and Slovak context*. Nakladatelství Karolinum.

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 and Robert Kolár. 2017. Kapitoly z korpusové versologie. *Slovo (modelové příklady)*, 37(3.1):2.

Petr Plecháč, Robert Kolár, Silvie Cinková, Artjoms Šeļa, Mirella De Sisto, Lara Nugues, Thomas Haider, Benjamin Nagy, Éliane Delente, Richard Renault, et al. 2023. PoeTree. poetry treebanks in Czech, English, French, German, Hungarian, Italian, Portuguese, Russian and Spanish. *Research Data Journal for the Humanities and Social Sciences*.

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

Petr Plecháč. 2021. *Versification and Authorship Attribution*. Institute of Czech Literature, Prague.

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.

Alan Prince and Paul Smolensky. 1993. *Optimality Theory: Constraint Interaction in Generative Grammar*. Rutgers University Center for Cognitive Science, New Brunswick, NJ.

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.

Piotr Sawicki, Marek Grzes, Fabricio Goes, Dan Brown, Max Peeperkorn, and Aisha Khatun. 2023. Bits of grass: Does GPT already know how to write like Whitman? In *Proceedings of the 14th International Conference for Computational Creativity*.

Patrícia Schmidtová, Saad Mahamood, Simone Balloccu, Ondřej Dušek, Albert Gatt, Dimitra Gkatzia, David M Howcroft, Ondřej Plátek, and Adarsa Sivaprasad. 2024. Automatic metrics in natural language generation: A survey of current evaluation practices. In *Proceedings of the 17th International Natural Language Generation Conference*, pages 557–583.

Yizhan Shao, Tong Shao, Minghao Wang, Peng Wang, and Jie Gao. 2021. A sentiment and style controllable approach for Chinese poetry generation. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*, pages 4784–4788.

Milan Straka, Jakub Náplava, Jana Straková, and David Samuel. 2021. *RobeCzech: Czech RoBERTa, a Monolingual Contextualized Language Representation Model*, page 197–209. Springer International Publishing.

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.

Pavel Šturm and Aleš Bičan. 2022. *Slabika a její hranice v češtině*. Charles University in Prague, Karolinum Press.

Jörg Wöckener, Thomas Haider, Tristan Miller, Thanh Tung Linh Nguyen, Minh Vu Pham, Jonas Belouadi, Steffen Eger, et al. 2021. End-to-end style-conditioned poetry generation: what does it take to learn from examples alone? In *Proceedings of the 5th Joint SIGHUM Workshop on Computational Linguistics for Cultural Heritage, Social Sciences, Humanities and Literature*, pages 57–66.

Nan Xu and Xuezhe Ma. 2025. LLM the genius paradox: A linguistic and math expert's struggle with simple word-based counting problems.

Linting Xue, Aditya Barua, Noah Constant, Rami Al-Rfou, Sharan Narang, Mihir Kale, Adam Roberts, and Colin Raffel. 2022. ByT5: Towards a token-free future with pre-trained byte-to-byte models. *Transactions of the Association for Computational Linguistics*, 10:291–306.

Lu Yu, Bartlomiej Twardowski, Xialei Liu, Luis Herranz, Kai Wang, Yongmei Cheng, Shangling Jui, and Joost van de Weijer. 2020. Semantic drift compensation for class-incremental learning. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*.

Vilém Zouhar, Clara Meister, Juan Gastaldi, Li Du, Mrinmaya Sachan, and Ryan Cotterell. 2023. Tokenization and the noiseless channel. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5184–5207, Toronto, Canada. Association for Computational Linguistics.

Michal Škrabal and Karel Piorecký. 2022. The corpus of contemporary Czech poetry: A database for research on contemporary poetic language across media. *Digital Scholarship in the Humanities*, 37(4):1240–1253.

A Motives Identification

Prompt

The system prompt in Czech language for motives identification is as follows:

Jste literární vědec se zaměřením na poezii. Vaším úkolem je určit až 5 hlavních témat básně {poemtitle}. Napište pouze tato témata, nic jiného, každé na samostatný řádek. Takto:\n 1. A\n 2. B\n 3. C

An English translation of the prompt is:

You are a literary scholar with a focus on poetry. Your task is to identify up to 5 main themes of the poem {poemtitle}. Write only these themes, nothing else, each on a separate line. Like this:\n 1. And 2. B\n 3. C

The title of the poem is inserted at the position of the *{poemtitle}* placeholder.

This is then followed by the user prompt, which contains the text of the poem, in plain text, with no annotations.

Full evaluation results

Full results of manual evaluation of motive generation can be found in Table 3.

Examples of automatically identified motives (for generated poems) are included in Appendix E.

B Examples of Poem Formats

We show here the poem 'Jaroslavu Vrchlickému' by Eduard Albert, formatted according to the formats used for the first and second set of experiments. The poem is in iambic metre (J) with the ABAB rhyme scheme and was published in the year 1900.

First format

The format:

```
# rhymescheme # year # metre
syllables # reduplicant # verse
syllables # reduplicant # verse
...
```

The poem:

```
# ABAB # 1900 # J
9 # oři # Tvá loď jde po vysokém moři,
9 # eje # v ně brázdu jako stříbro reje,
9 # oří # svou přídu v modré vlny noří
9 # eje # a bok svůj pěnné do peřeje.
```

Second format

The format:

```
authorname: poemtitle (year)

# rhymescheme #
# metre # syllables # reduplicant # verse
# metre # syllables # reduplicant # verse
...
```

The poem:

```
Eduard Albert: Jaroslavu Vrchlickému (1900)

# A B A B #
# J # 9 # oři # Tvá loď jde po vysokém moři,
# J # 9 # eje # v ně brázdu jako stříbro reje,
```

	Annotator K				Annotator R				Annotator O			
Poem	OK	DEL	EDIT	ADD	OK	DEL	EDIT	ADD	OK	DEL	EDIT	ADD
1	5				5				5			1
2	5				3	2		1	4		1	1
3	2	1	2		4	1			5			
4	5				4	1			5			
5	3	2			4	1			3	1	1	1
6	3	2		1	5				5			
7	3		2		5				5			
8	4		1		3	2			2	3		1
9	3	2		1	2	2	1	1	3	2		1
10	5				5				2	2	1	1
11	5				5				4	1		1
12	4	1			5	1			4	1		1
13	3	1	1		2	2	1	1	5			
14	4		1		2	2	1		5			1
15	3	1	1		3	2			3		2	
16	3	1	1		5				5			
17	3	2			4		1	1	4	1		1
18	4	1			2	2	1	1	4	1		1
19	4		1		3	2			5			
20	2	2	1	1	3	2		1	5			
21	5				4	1			1	2	2	1
22	3	2			5				5			
23	3	2		1	3	1	1		3	2		1
24	5				4	1			2	3		2
25	5				4	1			4	1		1

Table 3: Manual annotation of automatically generated motives for 3x25 poems (each annotator annotated a different set of poems). Each of the 5 generated motives was marked as correct (OK), surplus (DEL), or partially correct (EDIT); additionally, the annotator could mark missing motives (ADD).

```
# J # 9 # oří # svou přídu v modré vlny noří
# J # 9 # eje # a bok svůj pěnné do peřeje.
```

C Model Fine-tuning Details

C.1 Training of the first model

The code used to train the first model is published in a separate Github repository: https://github.com/jinymusim/GPT-Czech-Poet

Learning rate We used Cosine Schedule with warm-up.

Secondary Tasks An additional strategy to enhance learning involves incorporating classification heads to utilize available data. Given that the processed data includes annotations for rhyme schema, meter, year of publishing, and number of syllables, these annotations can be used to compute additional losses, thereby influencing the computed gradient. To implement this, a densely connected layer with softmax activation was introduced over the first token output of the last hidden layer for each named parameter. This configuration essentially makes the first token act as a class token. However, since it can be ensured that the first token is consistently the same, this should have minimal impact. A point of caution arises from the potential dominance of secondary task losses over the main loss, as they outnumber it at a ratio of 4 to 1. This could lead the model to 'focus' more on fine-tuning the secondary tasks rather than the primary task. To maintain control over the model, the weight assigned to secondary tasks was limited to a value of 0.1 for each task.

Drift compensation While finetuning on strophes is expected to be adequate, the temporal scope of the data from 1790 to 1940 raises the possibility that the base model czech-gpt2-oscar might contain inaccurate semantic and grammatical representations of words due to etymological fallacy. To address this concern, a strategy inspired by the article 'Semantic Drift Compensation' (Yu et al., 2020) was implemented. The model was initially trained on raw verses without any parameters, altering the language expressions without changing the structure first. This allowed the model to initially 'focus' on adapting to potential linguistic differences that are present in used dataset.

C.2 Training of the second model

We use LoRA (Hu et al., 2021) with Unsloth (Han et al., 2023) to fine-tune the model with the following parameters:

```
max_seq_length = 1024
warmup_ratio = 0.1
num_train_epochs = 30
lora_r = 64
lora_alpha = 64
```

D Token distribution

In Figure 3, we compare the frequency distributions of syllable versus subword tokens. We tokenized the deduplicated KČV dataset in two ways:

- subword tokenization, using the llama-3.1 tokenizer
- syllable tokenization, using Sekáček (Macháček, 2014)

We can see that the distribution of the token frequencies are quite similar, suggesting that syllable-based tokenization may be a viable alternative to standard subwords.

E Examples of Generated Outputs and Motives

We show one example of generation outputs for each of the models. The examples were selected to illustrate the typical quality of the generated poems, as well as some common error types that we often see in the outputs.

We also show automatically identified motives for the poems.

The input parameters for generation were: a poem of 1 or 2 stanzas of 4 verses (quatrains), using the AABB rhyme scheme, trochee metre, 8 syllables in the first verse, with the title and starting word 'Láska' ('Love').

Output of the first model

The generated poem:

```
Láska, když oni pějí,
jak kdo chce tu nejraději.
jako když se v roucho kryjí,
jako když si cudnou šíji
```

Automated translation by DeepL,[42] manually post-edited to match the original more closely:

```
Love when they sing,
as one likes it best.
As when they cover themselves in robes,
as when they their necks
```

[42] https://www.deepl.com/

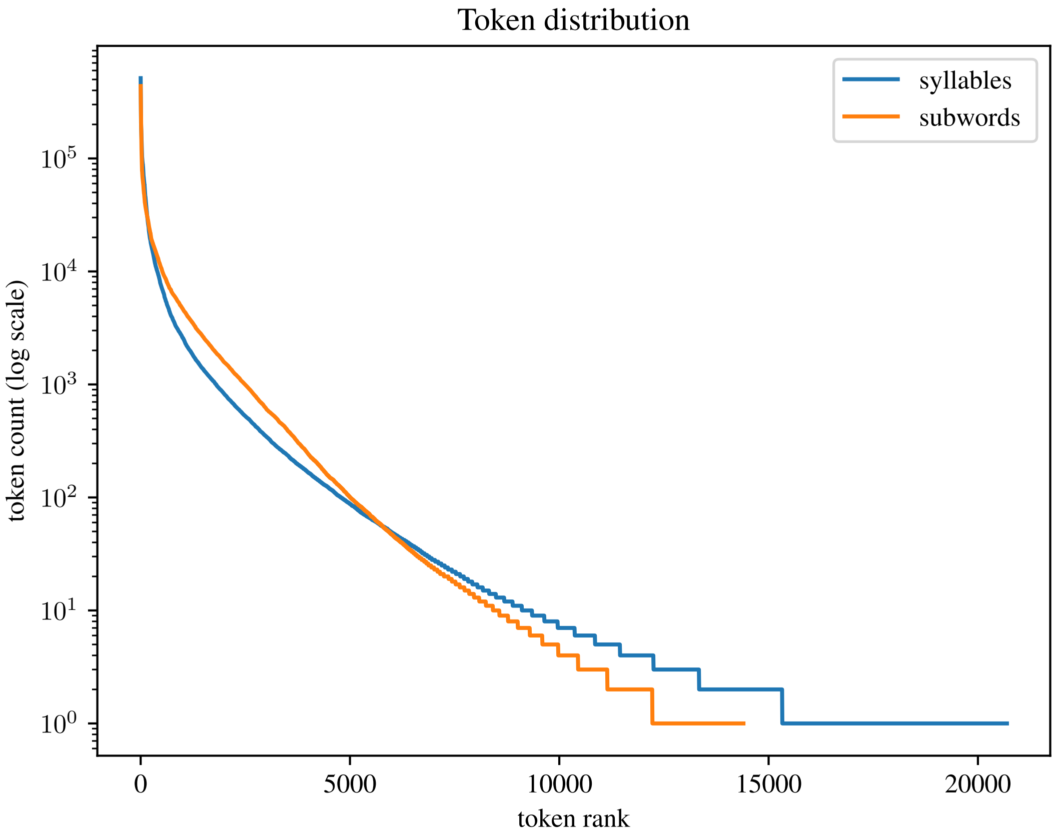


Figure 3: Comparison of frequency distributions of syllables and subwords.

In the first verse, one syllable is missing (7 syllables instead of 8) and the metre is disrupted by that; this could have been easily fixed by using an addition monosyllabon, such as *je* (‘is’) within the verse. In the second verse, the metre is adhered to and the number of syllables (7) matches the first verse, which is good, but the rhyming is imperfect and could have easily been made perfect by using a synonymous 4-syllable word *nejraději* instead of the 3-syllable word *nejraděj*; the model even generated the reduplicant hint *ěji* but then did not follow it. I.e., the first two verses could be very easily post-edited to be perfect in syllable count, metre and rhyming.

The second two verses are perfect in terms of metre, rhyming and number of syllables; although the poem unnecessarily rhymes more than is prescribed by the rhyme scheme.

The text is mostly meaningful, although the last verse seems unfinished.

Automatically identified motives for the poem:

```
1. Láska a její vyjádření
2. Cudnost a intimita
3. Vztah mezi jednotlivci
4. Přírodní metafory
5. Hra a radost v lásce
```

Automated translation of the motives by DeepL:

```
1. Love and its expression
2. Chastity and intimacy
3. Relationship between individuals
4. Natural metaphors
5. Play and joy in love
```

Output of the second model

The generated poem:

```
Láska utekla mi žena,
S kým je zřídka v lásce žena?
Jíž jsem dávno nevyzvád’,
Na nížto již nevádím.

Žena – žena – no, zcestí!
S ní jsem šťasten byl i v štěstí,
Za vinu si křivdou vidí –
Křivdou-li, co lid i sklidí?
```

Automated translation by DeepL, manually post-edited to match the original more closely:

```
Love my wife escaped,
```

With whom is a woman rarely in love?
I've not *invoted her for a long time,
I am not *invoting her any more.

A woman - a woman - well, astray!
I've been happy with her, even in happiness,
She sees herself guilty with injustice.
Is it injustice, what the people reap?

The third and fourth verse use only 7 syllables, but this is correct in the context of the poem. The first verse of the second stanza also uses only 7 syllables, which does not fit the structure well and disrupts the metre. Other verses adhere to the metre well. The rhyme scheme is observed except for the 3rd and 4th verse.

The poem seems mostly meaningful, although somewhat obscure at places. The poem uses a pair of similar non-existent words of unclear meaning (*nevyzvád*, *nevádím*). We often see this in the models when they are 'trying' to follow the rhyme scheme by 'inventing' non-existent rhyming words or word forms; yet, here, the two non-existent words do not adhere to the rhyme scheme. The model generated the reduplicant hint *ádám* for these two verses, but it did not follow this hint in neither of the verses.

Automatically identified motives for the poem:

1. Ztráta lásky
2. Osamělost
3. Reflexe vztahu
4. Vina a křivda
5. Hledání štěstí

Automated translation of the motives by DeepL:

1. Loss of love
2. Loneliness
3. Reflection on the relationship
4. Guilt and injustice
5. The search for happiness

Output of base Llama-3.1

For comparison, we also include an example of a poem generated by the base model used in the second set (llama-3.1) with no fine-tuning. For fine-tuning, we used the base *text completion* model, but for this experiment we used the *instruct variant* of the model to get a poem through instructing the model to generate a rhyming poem. (As the base model is not trained for our formatting of the poems, it obviously does not generate a poem when presented with our prompt.)

The generated poem:

Slunce svítí na obloze jasně,
jejich teplo cítíme v každém místě.
Den je plný života a radosti s vámi,
a noci jsou tiché, plné klidu a štěstí.
Větru vlající vlajky hrdě,
nám připomínají svobodu a pravdu.
Lidi se shromažďují, aby si hráli,
a jejich smích ozývá se do dálky, s radostí.

Automated translation by DeepL:

The sun shines brightly in the sky,
their warmth can be felt in every place.
The day is full of life and joy with you,
and the nights are quiet, full of peace and happiness.
Flags flutter proudly in the wind,
reminding us of freedom and truth.
People gather to play,
and their laughter echoes in the distance, with joy.

The text is semantically quite good, but mostly lacks the formal properties expected from poetry. It does not pertain to any metre; our analysis detects trochee as the most probable, but still below the threshold and thus analyzes the poem as non-metrical. The numbers of syllables also differ on most lines. There is no detectable rhyme scheme, as no two verses rhyme according to our analyses; some verses could be seen as imperfectly rhyming (there are traces of vowel rhyming).

Although we are showing only one example, these properties are quite typical for what we have observed in multiple experiments. The same is true for other free models and older commercial models.

Newest commercial models, such as GPT-4o, are able to generate poems which are formally better, with some rhyming, and often also with partial adherence to a metre, but based on our investigations, formal properties of typical outputs are still below the quality of outputs produced by our models. We intend to carry out a proper evaluation comparing our models to commercial models in future.

Automatically identified motives for the poem:

1. Příroda a její krása
2. Radost a štěstí
3. Svoboda a pravda
4. Společenské soužití
5. Klid a pohoda

Automated translation of the motives by DeepL:

1. Nature and its beauty

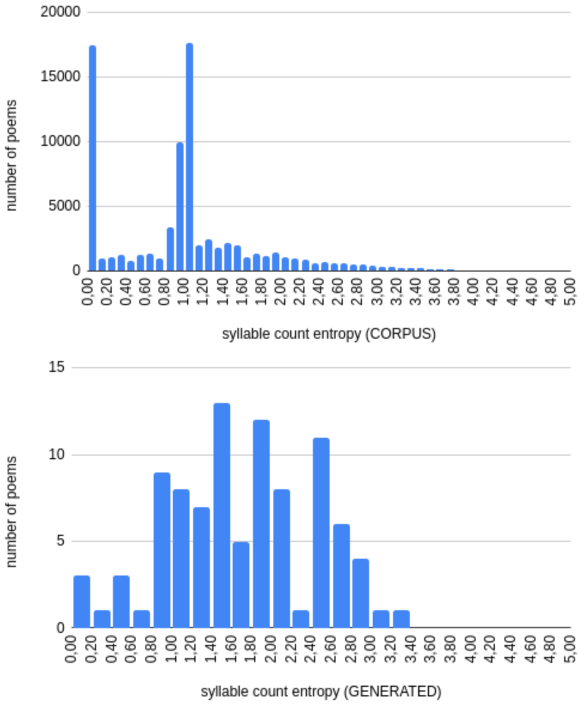


Figure 4: Histogram of the syllable-count entropies in the corpus and in the generated poems.

```
2. Joy and happiness
3. Freedom and truth
4. Social coexistence
5. Peace and well-being
```

F More Automated Evaluation Plots

In addition to the plots presented in the main body of the paper, we present two further histograms of the values of measures computed on poems generated by the second model, compared to values measured on poems in KČV.

Figure 4 shows the values of Syllable count entropy (defined in Section 3.2), and Figure 5 shows the values of Metre consistency (defined in Section 3.4).

G Screenshots of the Tool

We show two screenshots from the preliminary version of the online tool: Figure 6 shows the input screen, specifying the generation parameters, and Figure 7 shows the output screen, displaying the poem and its analyses.

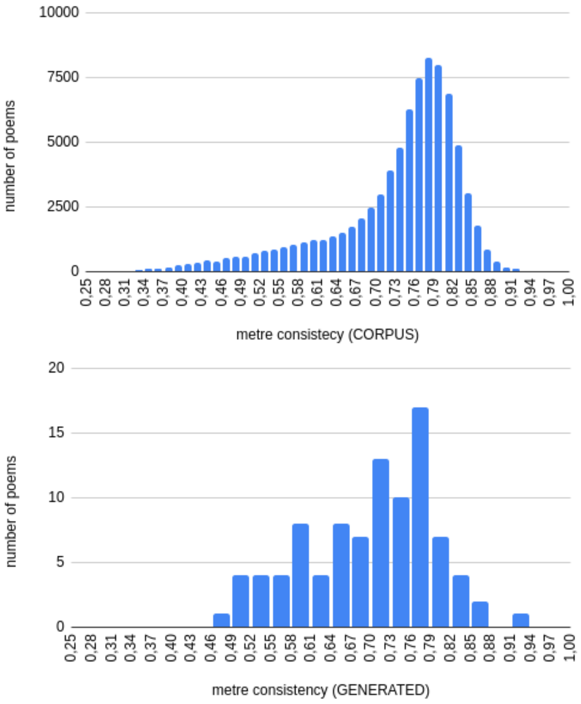
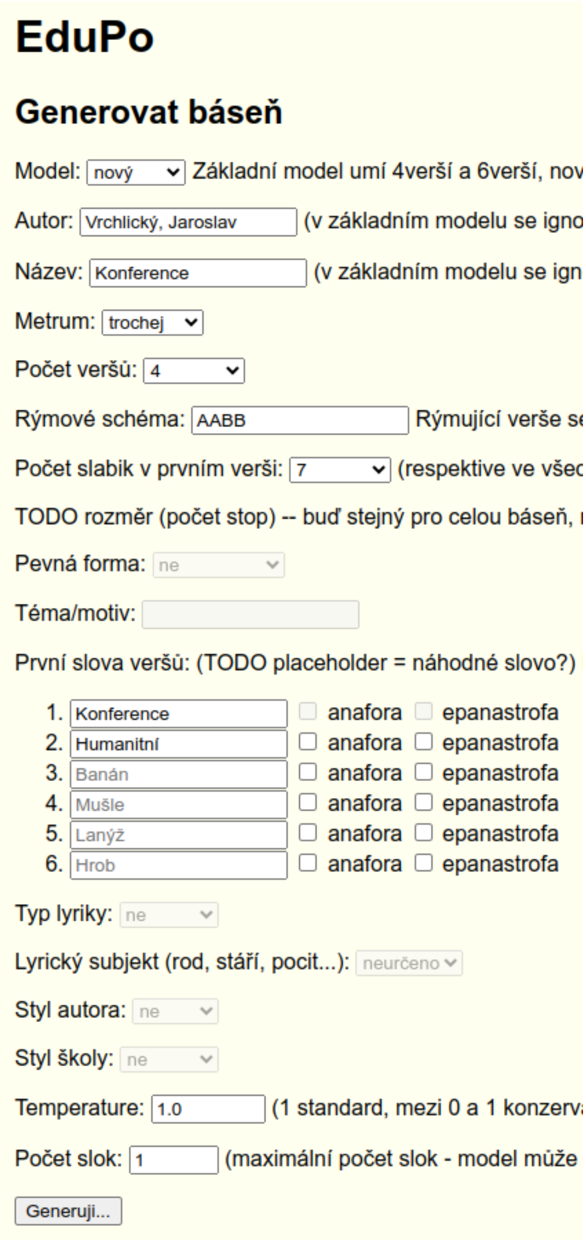


Figure 5: Histogram of metre consistencies of poems in the corpus and generated poems.



EduPo

Generovat báseň

Model: nový Základní model umí 4verší a 6verší, nov

Autor: Vrchlický, Jaroslav (v základním modelu se igno

Název: Konference (v základním modelu se ign

Metrum: trochej

Počet veršů: 4

Rýmové schéma: AABB Rýmující verše se

Počet slabik v prvním verši: 7 (respektive ve všec

TODO rozměr (počet stop) -- buď stejný pro celou báseň,

Pevná forma: ne

Téma/motiv:

První slova veršů: (TODO placeholder = náhodné slovo?)

1. Konference □ anafora □ epanastrofa
2. Humanitní □ anafora □ epanastrofa
3. Banán □ anafora □ epanastrofa
4. Mušle □ anafora □ epanastrofa
5. Lanýž □ anafora □ epanastrofa
6. Hrob □ anafora □ epanastrofa

Typ lyriky: ne

Lyrický subjekt (rod, stáří, pocit...): neurčeno

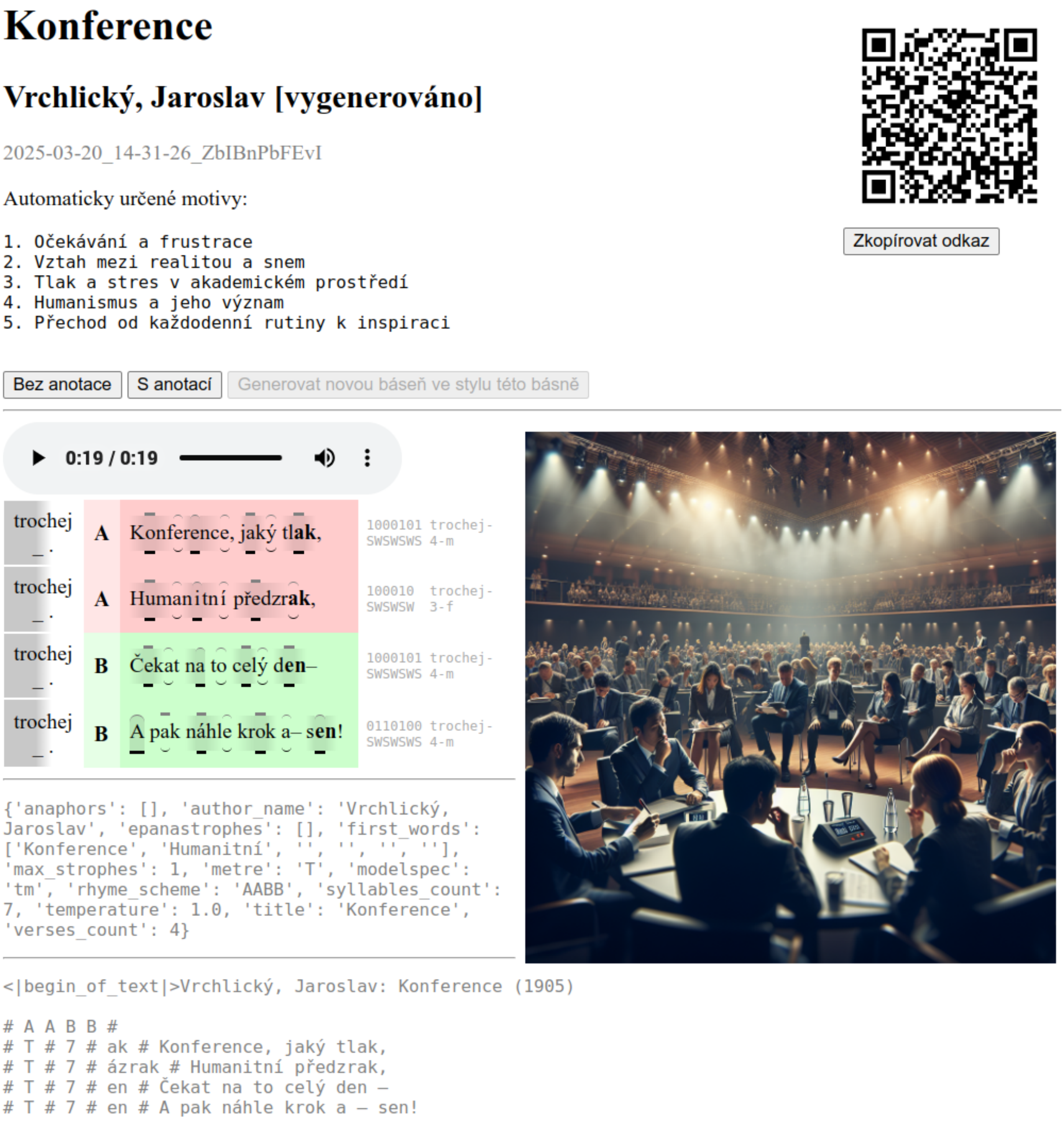
Styl autora: ne

Styl školy: ne

Temperature: 1.0 (1 standard, mezi 0 a 1 konzerv

Počet slok: 1 (maximální počet slok - model může

Figure 6: Screenshot of the input screen of the preliminary version of the tool.
The selected input parameters are: second model, style of the poet *Jaroslav Vrchlický*, title *Konference* ('The Conference'), trochee metre, 4 verses, AABB rhyme scheme, 7 syllables in the first verse, 1st verse starting with the word *Konference* ('conference'), 2nd verse starting with the word *Humanitní* ('humanities'), temperature 1.0, 1 stanza.



Konference

Vrchlický, Jaroslav [vygenerováno]

2025-03-20_14-31-26_ZbIBnPbFEvI

Automaticky určené motivy:

```
1. Očekávání a frustrace
2. Vztah mezi realitou a snem
3. Tlak a stres v akademickém prostředí
4. Humanismus a jeho význam
5. Přechod od každodenní rutiny k inspiraci
```

trochej _ .	A	Konference, jaký tlak,	1000101 trochej- SWSWSWS 4-m
trochej _ .	A	Humanitní předzrak,	100010 trochej- SWSWSW 3-f
trochej _ .	B	Čekat na to celý den–	1000101 trochej- SWSWSWS 4-m
trochej _ .	B	A pak náhle krok a– sen!	0110100 trochej- SWSWSWS 4-m

{'anaphors': [], 'author_name': 'Vrchlický, Jaroslav', 'epanastrophes': [], 'first_words': ['Konference', 'Humanitní', '', '', '', ''], 'max_strophes': 1, 'metre': 'T', 'modelspec': 'tm', 'rhyme_scheme': 'AABB', 'syllables_count': 7, 'temperature': 1.0, 'title': 'Konference', 'verses_count': 4}

```
<|begin_of_text|>Vrchlický, Jaroslav: Konference (1905)

# A A B B #
# T # 7 # ak # Konference, jaký tlak,
# T # 7 # ázrak # Humanitní předzrak,
# T # 7 # en # Čekat na to celý den –
# T # 7 # en # A pak náhle krok a – sen!
```

Figure 7: Screenshot of the output screen of the preliminary version of the tool.
Output generated by the second model according to the parameters set on the input screen, with automated versological analyses (metre, rhythm, stress, foot, reduplicants, rhyme scheme), automatically identified motives, an illustration automatically generated by DALL-E based on the title and text of the poem, and a speech transcription of the poem automatically generated by gTTS library.
The text of the poem, as translated by DeepL, is: 'Conference, what pressure, // Humanities premonition, // Waiting for it all day - // And then suddenly a step and - a dream!'
The automatically identified motives, as translated by DeepL, are: '1. Expectation and frustration; 2. Relationship between reality and dream; 3. Pressure and stress in the academic environment; 4. Humanism and its meaning; 5. Moving from daily routine to inspiration'
The annotation above each verse marks the stress pattern of the line (stressed syllable peaks are marked by lines and unstressed by curves), the annotation below marks the strong/weak positions expected by the metre. Below the poem, the input paremeters and the generated output are shown in raw form.