Main properties of the corpus:

• https://github.com/ELTE-DH/poetry-corpus
• Source: Hungarian Electronic Library
• Number of poems: 49
• Time period: 16th century - 20th century
• Number of poems: 13,063
• Number of words: 2.7 million

Annotation layers:

• Structural units: title, stanzas, lines, separators
• Grammatical features: lemma, part of speech, morphosyntactic features
• Sound devices: quantitative rhythm of lines, rhyme patterns of stanzas, rhyme pairs, alliterations, phonological features of words

Main text:

The level0, level1, level2, level3 and level4 formats presented below are the formats of the different versions of the corpus produced by each annotation stage. These versions contain an increasing number of annotation layers. The levels correspond to the libraries on the gitHub page of the corpus.

The TEI XML format of level1:

```
<xml:id="11">...
</xml:id>
```

The TEI XML format of level2:

```
<w lemma="Húnyt" msd="Case=Nom|Degree=Pos|Number=Sing" pos="ADJ" xml:id="w1">Húnyt</w>
```

The TEI XML format of level3:

```
<lg rhyme="abab" xml:id="lg1">...
</lg>
```

The XML format of level4:

```
<div type="poem">
<head>Húnyt szemmel...
</head>
<l rhyme="abab">...
</l>
</div>
```

Manual evaluation of the automatic annotation of rhythm:

Rules of syllable length: (1) Syllables with a short vowel and no consonant or only one consonant immediately after the vowel are short syllables; (2) syllables with a long vowel and syllables with a short vowel followed by a long consonant or more than one consonant are long syllables; (3) more than one consonant at the beginning of a word (e.g. krákog, trottys, strigula) do not lengthen the syllable ending in a short vowel in the preceding word.

Method: To measure the accuracy of the rhythm annotation, we divided the corpus into three sub-corpora on the basis of the poets' year of birth, after which 200 lines with their rhythm annotation were randomly selected from each sub-corpus. We then manually checked the rhythm annotation of lines and marked the incorrect annotations in spreadsheets. In the manual evaluation, only the three rules listed above were taken into account: the special metrical rules of Hungarian poetry before the mid-19th century were not applied.

Results:

<table>
<thead>
<tr>
<th>Time period</th>
<th>Error rate</th>
</tr>
</thead>
<tbody>
<tr>
<td>1505 - 1771</td>
<td>3.5%</td>
</tr>
<tr>
<td>1772 - 1854</td>
<td>1.5%</td>
</tr>
<tr>
<td>1855 - 1909</td>
<td>2%</td>
</tr>
<tr>
<td>1905 - 1909</td>
<td>2.33%</td>
</tr>
</tbody>
</table>

Automatic evaluation of three rule sets for rhyming: The rules of rhyming should not be too restrictive, but they should not over-generate. Both cases lead to more inconsistent annotations, where the rhyme patterns of certain stanzas in a poem are annotated differently than the others.

Method: We implemented three sets of rules to test which is the most efficient. The rule set considered most effective was the one that resulted in the largest number of poems annotated consistently, where all stanzas were annotated with the same rhyme pattern.

Results:

<table>
<thead>
<tr>
<th>Rule set</th>
<th>Consistent poems</th>
</tr>
</thead>
<tbody>
<tr>
<td>4593</td>
<td>same vowel in the last syllables without counting vowel length AND same length of the second to last syllables</td>
</tr>
<tr>
<td>4974</td>
<td>same vowel in the last syllables without counting vowel length AND same length of the second to last syllables AND last phonemes are vowels OR last phonemes are consonants</td>
</tr>
<tr>
<td>4740</td>
<td>same vowel in the last syllables with counting vowel length AND same length of the second to last syllables AND last phonemes are vowels OR last phonemes are consonants</td>
</tr>
</tbody>
</table>