

ELTE Poetry Corpus: A Machine Annotated Database for Canonical Hungarian Poetry

Péter Horváth¹, Péter Kundráth, Balázs Indig¹, Zsófia Fellegi², Eszter Szlávi¹, Tímea Borbála Bajzát², Zsófia Sárközi-Lindner¹, Bence Vida¹, Aslihan Karabulut¹, Mária Timári¹, Gábor Palkó^{1,2}

¹ Eötvös Loránd University, National Laboratory for Digital Heritage (Hungary)

² Research Centre for the Humanities: Institute for Literary Studies, National Laboratory for Digital Heritage (Hungary)
horvath.peter@btk.elte.hu, palko.gabor@btk.elte.hu

Main properties of the corpus:

- <https://github.com/ELTE-DH/poetry-corpus>
- Source: Hungarian Electronic Library
- Number of poets: 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

The stages of the annotation process:

- level0: annotation of structural units
Input: RTF, HTML, Output: TEI XML, Tool: XQuery script, Python script
- level1: manual checking of TEI XML files containing annotations of structural units
Output: TEI XML, Tool: manual, using Oxygen XML Editor
- level2: tokenization, lemmatization, part of speech and morphosyntactic annotation
Output: TEI XML, Tool: e-magyar (Váradi et al. 2018, Indig et al. 2019) embedded in a Python script
- level3: annotation of sound devices
Output: TEI XML, Tool: a Python program developed for the project
- level4: format conversion and the addition of further annotations
Output: XML, Tool: XSLT stylesheet

The query interface:

- <https://verskorporusz.elte-dh.hu>
- MariaDB-based SQL database
- Search functions:
 - Searching for word forms, lemmas, parts of speech, morphosyntactic features, phonological features and any combinations of these
 - Searching for multiple tokens on the basis of the above features
 - Generating frequency lists of word forms and lemmas
 - Generating frequency lists of multi-word structures
 - Filtering poems on the basis of rhyme patterns
 - Displaying quantitative characteristics of the sub-corpora selected
- Search results can be downloaded in TSV

References:

- Indig, B., Sass, B., Simon, E., Mittelholcz, I., Vadász, N., and Makrai, M. 2019. One format to rule them all – the emtsv pipeline for Hungarian. In: Proceedings of the 13th Linguistic Annotation Workshop. Florence. ACL. 155–165.
- Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., Indig, B., Farkas R., Vincze V. 2018. e-magyar – a digital language processing system. In: Proceedings of the Eleventh International Conference on Language Resources and Evaluation. Miyazaki. ELRA. 1307–1312.

Format: The 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:

```
<div type="poem">
  <head>Húnyt szemmel...</head>
  <lg>
    <l>Húnyt szemmel bérceken futunk</l>
    <l>s mindig csodára vágy szivünk:</l>
    <l>a legjobb, amit nem tudunk,</l>
    <l>a legszebb, amit nem hiszünk.</l>
  </lg>
  [...]
</div>
```

The TEI XML format of level2:

```
<l xml:id="l1">
  <w lemma="Húnyt" msd="Case=Nom|
  Degree=Pos|Number=Sing" pos="ADJ"
  xml:id="w1">Húnyt</w>
  <w lemma="szem" msd="Case=Ins|
  Number=Sing" pos="NOUN"
  xml:id="w2">szemmel</w>
```

The TEI XML format of level3:

- Rhyme patterns, rhythm:

```
<div type="poem">
  <head>Húnyt szemmel...</head>
  <lg rhyme="abab" xml:id="lg1">
    <l n="8" real="11110101"
    xml:id="l1">
```
- Standoff annotation of rhyme pairs:

```
<linkGrp type="rhymePairs">
  <link target="#w4 #w14"/>
  <link target="#w9 #w19"/>
  <link target="#w28 #w34"/>
</linkGrp>
```
- Standoff annotation of alliterations:

```
<spanGrp type="alliterations">
  <span target="#w1 #w2 #w3 #w4"
  type="anaa"/>
  <span target="#w20 #w21 #w22"
  type="aaa"/>
  <span target="#w29 #w30" type="aa"/>
</spanGrp>
```

The XML format of level4:

```
<div type="poem" div_numStanza="2"
div_numLine="8" div_numWord="34"
div_numSyll="63" div_numShortSyll="24"
div_numLongSyll="39" div_rhyme="abab|abcb"
div_syllPattern="8-8-8-8|8-8-5-10">
  <head>Húnyt szemmel...</head>
  <lg xml:id="lg1" lg_numLine="4" lg_numWord="19"
  lg_numSyll="32" lg_numShortSyll="11"
  lg_numLongSyll="21" rhyme="abab"
  lg_syllPattern="8-8-8-8">
    <l xml:id="l1" l_numWord="4" l_numSyll="8"
    l_numShortSyll="2" l_numLongSyll="6"
    real="11110101">
      <w xml:id="w1" lemma="Húnyt" msd="Case=
      Nom|Degree=Pos|Number=Sing" pos="ADJ"
      w_numSyll="1" phonType="low" phonStruct=
      "cBcc">Húnyt</w>
```

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*, *trottyos*, *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:	Time period	Error rate
	1505 - 1771	3.5%
	1772 - 1854	1.5%
	1855 - 1909	2%
	1505 - 1909	2.33%

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:	Rule set	Consistent poems
	same vowel in the last syllables without counting vowel length AND same length of the second to last syllables	4593
	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	4974
	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	4740