Proceedings of the 13th Conference on Language Resources and Evaluation (LREC 2022), pages 3471–3478
Marseille, 20-25 June 2022

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

Péter Horváth[1], Péter Kundráth, Balázs Indig[1], Zsófia Fellegi[2], Eszter Szlávich[1], Tímea Borbála Bajzát[2], Zsófia Sárközi-Lindner[1], Bence Vida[1], Aslihan Karabulut[1], Mária Timári[1], Gábor Palkó[12]

[1]Eötvös Loránd University, National Laboratory for Digital Heritage
1088 Múzeum krt. 6-8., Budapest, Hungary
[2]Research Centre for the Humanities: Institute for Literary Studies, National Laboratory for Digital Heritage
1118 Ménesi út 11–13., Budapest, Hungary
{horvath.peter, indig.balazs, szlavich.eszter, lindner.zsofia, vida.bence, karabulut.aslihan, timari.maria, palko.gabor}@btk.elte.hu, {fellegi.zsofia, bajzat.timea}@abtk.hu, peter.kundrath@gmail.com

Abstract

ELTE Poetry Corpus is a database that stores canonical Hungarian poetry with automatically generated annotations of the poems' structural units, grammatical features and sound devices, i.e. rhyme patterns, rhyme pairs, rhythm, alliterations and the main phonological features of words. The corpus has an open access online query tool with several search functions. The paper presents the main stages of the annotation process and the tools used for each stage. The TEI XML format of the different versions of the corpus, each of which contains an increasing number of annotation layers, is presented as well. We have also specified our own XML format for the corpus, slightly different from TEI, in order to make it easier and faster to execute queries on the corpus. We discuss the results of a manual evaluation of the quality of automatic annotation of rhythm, as well as the results of an automatic evaluation of different rule sets used for the automatic annotation of rhyme patterns. Finally, the paper gives an overview of the main functions of the online query tool developed for the corpus.

Keywords: poetry corpus, Hungarian, automatic annotation, sound devices

1. Introduction

ELTE Poetry Corpus[1] is a database with an online query interface[2], which stores a significant part of canonical Hungarian poetry with different types of annotations. Currently, the corpus contains the complete poems of 49 Hungarian authors. The number of poems in the corpus is 13,063 and the number of words is roughly 2,7 million. The main format of the corpus is TEI XML, which is one of the most widely used formats in the fields of digital humanities and linguistics. The annotations of the poems have been created automatically. We annotated the poems on three levels. First, we annotated the structural units of poems: the titles, the stanzas and the lines. Second, we tokenized the texts and annotated the grammatical features of words, i.e. the lemma, the part of speech and the morphosyntactic properties. Third, we annotated several features of sound devices as well: the rhyme patterns, the rhyme pairs, the rhythm, the alliterations and the main phonological properties of words.

In section 2, we give an overview of previous works that served as inspirations for the present project. Section 3 discusses the stages of the corpus building process and the tools used for each stage. Section 4 describes the format of the corpus. Section 5 explains how the accuracy of the automatic annotation of rhythm was evaluated and how the different rule sets applied for the annotation of rhyme patterns was automatically evaluated. Finally, in section 6, we briefly describe the main features of the online query tool developed for the corpus.

2. Related Work

We are not aware of any previous large-scale Hungarian poetry corpus that contains automatically generated annotations of sound devices. However, in the course of corpus building, we could rely on general-purpose Hungarian corpora, such as the Hungarian Gigaword Corpus[3] (Oravecz et al., 2014) and the Hungarian Historical Corpus[4] (Sass, 2017). The database Répertoire de la poésie hongroise ancienne[5] should also be mentioned. While it does not contain annotated texts, it does provide search functions for various data of Hungarian poems written before 1600, including their metrical characteristics (Horváth et al., 1979). The Czech Poetry Corpus (Korpus českého verše)[6] (Plecháč and Kolár, 2015) developed by the Czech Academy of Sciences was also an important inspiration for us. This corpus contains nearly 80,000 poems from the 19th and

[1] https://github.com/ELTE-DH/poetry-corpus
[2] https://verskorpusz.elte-dh.hu
[3] http://clara.nytud.hu/mnsz2-dev
[4] http://clara.nytud.hu/mtsz
[5] https://f-book.com/rpha
[6] https://versologie.cz/v2/web_content/corpus.php?lang=en;https://github.com/versotym/corpusCzechVerse

early 20th centuries with automatically generated annotations of not only the lemmas, the morphosyntactic features and the phonological features of words, but also the rhyme and rhythm of the poems (Ibrahim and Plecháč, 2011). We also relied on the design of the Corpus of Spanish Golden-Age Sonnets (Corpus de Sonetos del Siglo de Oro)[7] created in TEI XML format similar to ELTE Poetry Corpus. This corpus contains automatically generated annotations of rhythm (Navarro-Colorado, 2015; Navarro-Colorado et al., 2016).

Several programs have also been developed for the machine recognition of the sound devices of poems, especially for English-language poems. For instance, the programs Scandroid (Hartman, 2005) and ZuScansion (Agirrezabal et al., 2016) analyze the rhythm and meter of English poems. Another program, AnalysePoem can even recognize the rhyme patterns of English poems in addition to their rhythm and meter (Plamondon, 2006). In recent years, there have also been several research projects based on the automatic analysis of the sound devices of poems. Kao and Jurafsky (2012), for example, studied the differences between professional and amateur American poems, using not only vocabulary but also automatically analyzed features of sound devices, such as alliterations and rhyme pairs. The research of Tanasescu et al. (2016) is also worth mentioning, which aims to automatically classify English poems on the basis of rhythm and rhyme.

There are only a few examples of automatic analysis of sound devices in Hungarian-language poems, but some of them are surprisingly early. Vilmos Voigt's paper represents the first attempt at computer-based rhythm analysis of Hungarian poems: the rhythm of three sonnets was analyzed with a program (Voigt, 1972). Another early attempt is Jékel and Papp's book, which offers computer-generated phoneme statistics of Endre Ady's complete poems (Jékel and Papp, 1974). Also among the early examples is Jékel and Szuromi's book, which offers a multidimensional, partly automatically generated rhythm analysis of 300 poems written by Sándor Petőfi (Jékel and Szuromi, 1980). The first and only multi-functional program for analyzing the rhyme patterns, alliterations and meter of Hungarian poems was developed by Lesi (2008). Currently, this program is not accessible. Labády's research should also be mentioned, since it studies Dániel Berzsenyi's poems on the basis of the lexical and phonemic properties (length of words, distribution of vowels and consonants) analyzed automatically (Labádi, 2018). The most recent research using automatic analysis of sound devices of Hungarian poems was carried out by Maróthy et al. (2021); they automatically analyzed the rhymes of 26 historical songs written in the 16th century.

[7] https://github.com/bncolorado/CorpusSonetosSigloDeOro

3. The Stages of the Annotation Process

The source of the corpus was the document files of the Hungarian Electronic Library[8], a repository that stores the complete poems of many authors who are in the public domain. These documents are typically available in several formats on the website of the Hungarian Electronic Library. RTF files were used as a first choice, and HTML files were used if the RTF files were not available. In the first stage, we used a script on the document files of the Hungarian Electronic Library to create from them the TEI XML files that contain the annotation of structural units. In the case of RTF files, we used an XQuery script; in the case of HTML files, we used a Python script. As a result of running the scripts, every poem is included in a separate TEI XML file, which contains the annotations for titles, stanzas and lines as XML elements. In this phase, we also automatically created the <teiHeader> element of the TEI XML files, which contains metadata about the poems, such as author, title and source document information.
We then manually checked the TEI XML files that contain the annotation of structural units, that is, we compared them with the original RTF or HTML documents. This was necessary because the scripts could not annotate certain special cases correctly, and there were also inconsistencies in the source files that led to annotation errors. For manual checking, Oxygen XML Editor[9] was used.
After the manual checking, we automatically annotated the grammatical properties of the words in the poems, i.e. the lemma, the part of speech and the morphosyntactic features. For this we used the emtsv[10] version of the e-magyar pipeline, which is an NLP tool for the grammatical analysis of Hungarian texts (Váradi et al., 2018; Indig et al., 2019). We used e-magyar with a script that extracts the text from the TEI XML file, runs e-magyar on the text, and then converts the TSV output of e-magyar back to TEI XML. The part of speech and the morphosyntactic features were annotated using the tag set of Universal Dependencies (Vincze et al., 2017), which is one of the output options of e-magyar.
The next stage in the annotation process was the automatic annotation of sound devices: we annotated the rhyme patterns of stanzas, the rhyme pairs within stanzas, the rhythm of lines, the alliterations and the phonological features of words. The latter consists of the annotation of syllable number, vowel type and phonological structure. For the automatic annotation of sound devices, we used the Python program hunpoem_analyzer-TEI developed for this project (Horváth, 2020).
As a final annotation stage, using an XSLT stylesheet, we made some transformations to the position of the annotations in the TEI XML files and renamed some XML elements and attributes.[11] Additional annotations

[8] http://mek.oszk.hu
[9] https://www.oxygenxml.com
[10] https://github.com/nytud/emtsv
[11] On the new and renamed elements and attributes, see the

on word and syllable counts were also added to the XML files in this phase. The resulting XML files are close to TEI, but do not conform to the TEI specification. This last annotation stage was necessary because the format specified by TEI is less suitable for storing more detailed annotations. As a result of this annotation stage, we provided the poems in an XML format in which the arrangement and the naming of elements and attributes are more logical and self-explanatory, allowing easier writing and faster execution of query expressions. Naturally, the poems are also available in the TEI XML format to ensure interoperability.
The following list summarizes the stages of the workflow described above.

- 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 embedded in a Python script)
- level3: annotation of sound devices (Output: TEI XML, Tool: a Python program developed for the project (hunpoem_analyzer-TEI)
- level4: format conversion and the addition of further annotations (Output: XML, Tool: XSLT stylesheet)

The workflow designed for annotating the corpus allows re-running the annotation steps on the manually checked TEI XML files at any time if any of the tools of these steps have been improved.

4. The Format of the Corpus

The format of the corpus, except for the XML files produced by the last annotation stage, is TEI XML. The TEI specification offers tag sets for the annotation of a number of text types, including poems (TEI Consortium, 2021). 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.

4.1. The Format of level1

Figure 1 presents the TEI XML files of the poems produced by the automatic annotation of structural units and the manual checking of the annotations.

description on the gitHub page of the corpus.

```
<text>
  <body>
    <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>
      <lg>
        <l>Az álmok síkos gyöngyeit</l>
        <l>szorítsd, ki únod a valót:</l>
        <l>hímezz belőlük</l>
        <l>fázó lelkedre gyöngyös takarót.</l>
      </lg>
    </div>
  </body>
</text>
```

Figure 1: Annotation of structural units.

The titles of the poems are placed in the <head> element, the stanzas in the <lg> element, and the lines in the <l> element. The subtitles, the epigraphs, the separators and the notes on where and when the poem was written are in the <p> elements.

4.2. The Format of level2

In the case of level2 files produced by running e-magyar, every word is placed in a separate <w> element that contains the grammatical properties of the word as attributes. The dictionary form of the word is in the @lemma attribute, the part of speech is in the @pos attribute, and the morphosyntactic features of the word are in the @msd attribute. The morphosyntactic features are represented as feature-value pairs, based on the tag set of Universal Dependencies. The punctuation marks are in <pc> elements in which the @join attribute indicates the direction of adjacency. In this annotation stage, the elements <lg>, <l>, <w> and <pc> get a unique identifier as the value of the @xml:id attribute. Figure 2 shows the layout of the aforementioned annotations for one line of a poem.

```
<l xml:id="l3">
  <w lemma="a" msd="Definite=Def|PronType=Art"
    pos="DET" xml:id="w10">a</w>
  <w lemma="jó" msd="Case=Nom|Degree=Sup|Number=Sing"
    pos="ADJ" xml:id="w11">legjobb</w>
  <pc join="left" pos="PUNCT" xml:id="pc2">,</pc>
  <w lemma="ami" msd="Case=Acc|Number=Sing|Person=3|
    PronType=Rel" pos="PRON" xml:id="w12">amit</w>
  <w lemma="nem" msd="PronType=Neg" pos="ADV"
    xml:id="w13">nem</w>
  <w lemma="tud" msd="Definite=Ind|Mood=Ind|
    Number=Plur|Person=1|Tense=Pres|VerbForm=Fin|
    Voice=Act" pos="VERB" xml:id="w14">tudunk</w>
  <pc join="left" pos="PUNCT" xml:id="pc3">,</pc>
</l>
```

Figure 2: Annotation of grammatical features.

4.3. The Format of level3

By running the program hunpoem_analyzer-TEI, additional XML elements and attributes containing the annotations of sound devices are added to the TEI XML files (Figure 3).

```
<lg rhyme="abab" xml:id="lg1">
  <l n="8" real="11110101" xml:id="l1">
    <w lemma="húnyt" msd="Case=Nom|Degree=Pos
    Number=Sing" pos="ADJ" xml:id="w1">Húnyt</w>
    […]
```

Figure 3: Annotation of rhyme pattern and rhythm.

The rhyme pattern of stanzas is annotated as the value of @rhyme attributes of the <lg> elements. The program used for the annotation of sound devices outputs the rhyme patterns in the traditional way: for each stanza, the rhyme pattern is denoted by a character string in which the rhyming lines are indicated by matching letters of the alphabet (pl. *aabbcb*). The @n attribute of <l> elements contains the number of syllables in the line, while the @real attribute contains the quantitative rhythm of the line. The rhythm is annotated by a string of *0* and *1* characters, where *0* indicates short syllables and *1* indicates long syllables (e.g. *Húnyt szemmel bérceken futunk – 11110101*).

The TEI specification does not allow the phonological features of words to be annotated as attributes of the <w> elements. Therefor, standoff annotation was used, which means that the annotations of phonological features are separated from the text of the poem and placed in a later part of the XML files, as shown in Figure 4.

```
<spanGrp type="phonStructures">
  <span subtype="1" target="#w1" type="low">cBcc</span>
  <span subtype="2" target="#w2" type="high">cfccfc
  </span>
  […]
</spanGrp>
```

Figure 4: Annotation of phonological features.

The elements enclosed by the <spanGrp> element contain the phonological properties of each word. The value of the @target attribute is the @xml:id of the word annotated. The value of the @subtype attribute is the syllable number; the value of the @type attribute is the vowel type of the word. The content of the element is the word's phonological structure. In the annotation of phonological structure, we followed the representational format of the Hungarian Gigaword Corpus with some minor modifications (Oravecz et al., 2014). Every word gets a string of *c*, *b*, *f*, *B* and *F* characters indicating some important features of the phonemes. The meaning of the characters is the following: *c*: consonant, *b*: short back vowel, *B*: long back vowel, *f*: short front vowel, *F*: long front vowel (e.g. *szerszámaival – cfccBcbfcbc*).

We also annotated the rhyme pairs within stanzas and the alliterations using standoff annotation. The rhyme pairs are annotated by <link> elements contained in the <linkGrp> element (Figure 5). Each link element annotating a rhyme pair has a @target attribute, which contains the two @xml:id identifiers referring to the two rhyming words. There can be up to four lines between the two words annotated as forming a rhyme pair. A word can be a member of two rhyme pairs, as the second member of the first rhyme pair and as the first member of the second rhyme pair. For instance, in the case of a six-line stanza annotated with the rhyme pattern *aabbaa*, the rhyming word of the second line forms a rhyme pair with the rhyming word of the first line and another rhyme pair with the rhyming word of the fifth line. However, the program currently does not treat the rhyming words in the second and sixth lines as a rhyme pair. In other words, a rhyming word as first member can form a rhyme pair only with the rhyming word closest to it.

```
<linkGrp type="rhymePairs">
  <link target="#w4 #w14"/>
  <link target="#w9 #w19"/>
  <link target="#w28 #w34"/>
</linkGrp>
```

Figure 5: Annotation of rhyme pairs.

Alliterations are annotated in a similar way (Figure 6). The elements enclosed by the <spanGrp> element annotate the alliterations in such a way that the value of the @target attribute refers to the @xml:id identifiers of the words that make up the alliterating structure. We annotate as alliterations not only those word structures in which consecutive words begin with the same phoneme, but also those in which a word beginning with another phoneme is inserted between two words beginning with the same phoneme. Therefore, the value of the @type attribute is a string consisting of the characters *a* and *n*, in which character *a* indicates alliterating words, while character *n* indicates non-alliterating words between two alliterating words (e.g. *Bus donna barna balkonon – anaa*). The Hungarian articles *a* and *az*, and the Hungarian conjunctions *s* and *és* are handled as stop words, which means that these words cannot be the members of two-word alliterations. However, they can be members of an alliteration containing at least two other alliterating words.

```
<spanGrp type="alliterations">
  <span target="#w10 #w11 #w12 #w13" type="anaa"/>
  <span target="#w29 #w30 #w31" type="ana"/>
  <span target="#w34 #w35" type="aa"/>
  […]
</spanGrp>
```

Figure 6: Annotation of alliterations.

4.4. The Format of level4

In the case of the XML files produced by the last annotation stage, we have deviated from the TEI specification in order to make the poems accessible in two formats: in TEI and in a format that is easier and faster to

query. Firstly, for the sake of clarity, we have changed the names of several elements and attributes to clearly indicate the type of annotation they contain. Secondly, the phonological features annotated using standoff annotation at level3 have been transferred to the attributes of the <w> elements. Thirdly, the elements of structural units have been expanded with attributes for the number of stanzas, lines, words and syllables.

```
<div type="poem" div_numStanza="2" div_numLine="8"
div_numWord="34" div_numSyll="63" div_numShorSyll="24"
div_numLongSyll="39" div_rhyme="abab|abcb"
div_syllPattern="8-8-8-8|8-8-5-10">
  <head type="title">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="11110
    101">
      <w xml:id="w1" lemma="Húnyt" pos="ADJ"
      msd="Case=Nom|Degree=Pos|Number=Sing"
      w_numSyll="1" phonType="low" phonStruct="cBcc">
      Húnyt</w>
```

Figure 7: Format of level4.

Figure 7 shows the new attributes of the element added to the XML files. The attribute @div_numStanza contains the number of stanzas, @div_numLine the number of lines, @div_numWord the number of words, @div_numSyll the number of syllables, @div_numShortSyll the number of short syllables and @div_numLongSyll the number of long syllables in the poem. The rhyme pattern of the whole poem is contained in the @div_rhyme attribute. In the string given as the value of @div_rhyme, vertical bars separate the rhyme patterns of each stanza. Identical letters indicate rhyming lines only within one stanza. The @div_syllPattern attribute contains the syllable pattern of the poem, which is a string of numbers separated by hyphens and vertical bars, where the numbers indicate the number of syllables in the lines. The <lg> elements of stanzas and the <l> elements of lines have been similarly expanded with additional attributes for line number, word number and syllable number.

As shown in Figure 7, the phonological properties of words annotated in level3 files in standoff format have been transferred to the attributes of the <w> elements. The @w_numSyll attribute contains the syllable number, the @phonType attribute contains the vowel type, and the @phonStruct attribute contains the phonological structure of the word.

In the standoff annotations of rhyme pairs, we have changed the names of the elements for clarity, and to facilitate queries, we have indicated the rhyming word forms as the content of the elements and also indicated the grammatical and phonological properties of the rhyming words as attributes here. We have also changed the name of the elements containing the standoff annotations of alliterations. Furthermore, the word forms of the alliteration are indicated as the content of the elements, while the lemmas, parts of speech and morphosyntactic features of the words forming the alliteration are indicated as the values of attributes.

5. Evaluation of the Annotation of Sound Devices

We manually evaluated the accuracy of the annotation of the quantitative rhythm of lines and used an automatic method to evaluate three rule sets of rhyming to select the most efficient one.

5.1. Manual Evaluation of the Annotation of Rhythm

The automatic annotation of the quantitative rhythm of lines, i.e. the long and short syllables, could be done on the basis of some simple rules well known in Hungarian poetry, which means that it was not necessary to include pronunciation dictionaries in the algorithm. These rules are the following: (1) the program analyzes syllables with a short vowel and no consonant or only one consonant immediately after the vowel as short syllables; (2) the program analyzes syllables with a long vowel and syllables with a short vowel followed by a long consonant or more than one consonant as long syllables; (3) we have also implemented the Hungarian metrical rule that 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.

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. All authors in a sub-corpus participated with the same amount of lines in the random sample. 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. We also did not consider as errors cases where the length of a phoneme was written differently from today's spelling (long instead of short or short instead of long), as we assumed that this difference reflected the pronunciation of an older language variety or a deliberate deviation of the author from the standard pronunciation.

The results of the manual evaluation are presented in Table 1. The first three rows show the percentage of incorrect lines in the random samples from the three different time periods; the last row shows the percentage of incorrect lines in the three samples combined. As the table shows, the error rate in each period is quite low. The algorithm was able to treat the two-digit consonants as a single phoneme, which was the main problem to be solved for quantitative rhythm annotation. Many of the incorrect annotations were caused by foreign words, mostly proper names.

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

Table 1: Evaluation of the annotation of rhythm

5.2. Automatic Evaluation of Three Rule Sets for Rhyming

In the case of automatic annotation of rhymes, the question arises: what rules should be implemented to determine the rhyming lines? The rules of rhyming should not be too restrictive, but they should not overgenerate. Both cases lead to more inconsistent annotations, where the rhyme patterns of certain stanzas in a poem are annotated differently than the others because of the too narrow or too general rules. 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. Table 2 presents the three sets of rules for rhyming and the number of poems annotated with the same rhyme pattern in the case of every stanza, by using the given rule set.

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

Table 2: Evaluation of rule sets for rhyming

As the table shows, the second rule set proved to be the most effective with the largest number of consistently annotated poems, so we eventually annotated the corpus according to this rule set. In the future, we would like to further develop the algorithm so that in the case of an inconsistent annotation, it would analyze the stanzas of the poem using the other two rule sets or even additional ones to achieve a consistent annotation.

It is worth noting that all three rule sets include the condition of equal length of the second to last syllables, which means that the annotation of rhymes is based on the annotation of quantitative rhythm. The better the quality of the rhythm annotation, the more effectively the rule sets of rhyming can be applied.

6. The Query Interface

A MariaDB-based SQL database has been created from the level4 XML files. The query tool developed for the corpus searches this database. The tool is freely accessible for anyone at `https://verskorpusz.elte-dh.hu`. In addition to displaying the annotated properties of the poems, the query tool has a number of search functions. In designing the search functions, we could rely on the query interfaces of existing Hungarian-language corpora, especially the Hungarian Gigaword Corpus (Oravecz et al., 2014). The query tool can be used to search not only for a single token, but also for structures consisting of multiple tokens. By using the different functions of the tool, it is possible to search for word forms, lemmas, parts of speech, morphosyntactic features, or any combination of these. Thanks to the annotation of phonological features and rhythm, the user can also query words by syllable number, vowel type, phonological structure and syllable length, or combine these with search terms for word form, lemma, part of speech and morphosyntactic features. In addition to the authors, the poems can be filtered on the basis of rhyme patterns as well.

Frequency lists of word forms and lemmas can also be generated based on the search terms specified by the query tool. If a search term for a multi-word structure is entered in the search field, it is also possible to generate a frequency list for that structure. Table 3 and Table 4 present two frequency lists generated by the query tool. The first list shows the five most frequent rhyming noun lemmas in the corpus, while the second list presents the five most frequent alliterating adjective + noun collocations by lemma.

Rhyming nouns	Occurrences
élet (life)	2585
szem (eye)	2362
világ (word)	2350
ég (sky)	2099
kéz (hand)	2002

Table 3: The most frequent rhyming nouns

Alliterating adjective + noun	Occurrences
szép szem (beautiful eyes)	221
szép szó (beautiful word)	129
nagy név (great name)	93
kis kéz (small hand)	50
szép szerelem (beautiful love)	30

Table 4: The most frequent alliterating adjective + noun collocations

The tool can also display some important quantitative characteristics of the sub-corpora selected for analysis, such as the number of different word forms and the number of different lemmas of the sub-corpora. The latter indicates the size of the vocabulary used by the poets analyzed. Search results, their associated quantitative data, frequency lists and the quantitative characteristics of the sub-corpora selected can be downloaded in TSV format and opened in any spread-sheet program.

7. Summary

Our aim in building ELTE Poetry Corpus was to create an open access annotated database that can be used for both literary and linguistic research. The query interface of the corpus allows access to quantitative data on canonical Hungarian poetry without any special IT skills. This type of information cannot be obtained by close reading. We believe that the query interface makes the corpus useful not only for researchers but also for teachers. The annotated XML files can be downloaded from the gitHub repository of the project (`https://github.com/ELTE-DH/poetry-corpus`) and may be used for research without restrictions. Publishing the XML files allows researchers with programming skills to perform more complex queries on the corpus that the online query interface does not support. The ELTE Poetry Corpus is not a closed project: in the future, we would like to add more authors and annotation layers to the corpus. We also plan to implement new features in the query tool.

8. Acknowledgements

This project was supported by the National Laboratory for Digital Heritage and the Higher Educational Institutional Excellence Program of the National Research, Development and Innovation Office of Hungary.

9. Bibliographical References

Agirrezabal, M., Astigarraga, A., Arrieta, B., and Hulden, M. (2016). Zeuscansion: A tool for scansion of english poetry. *Journal of Language Modelling*, 4(1):3–28.

Hartman, C. O., (2005). *The Scandroid. Version 1.1. [User guide]*.

Horváth, P. (2020). A vershangzás jellemzőinek automatikus feltárása józsef attila verseiben. *Digitális Bölcsészet*, 3:M:3–M:27.

Ibrahim, R. and Plecháč, P., (2011). *Toward Automatic Analysis of Czech Verse*, pages 295–305. RAM, Lüdenscheid, Germany.

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 Annemarie Friedrich, et al., editors, *Proceedings of the 13th Linguistic Annotation Workshop*, pages 155–165, Florence, Italy, august. Association for Computational Linguistics (ACL).

Jékel, P. and Papp, F. (1974). *Ady Endre összes költői műveinek fonémastatisztikája*. Akadémiai Kiadó, Budapest, Hungary, 1nd edition.

Jékel, P. and Szuromi, L. (1980). *Petőfi metrumai*. Kossuth Lajos Tudományegyetem, Debrecen, Hungary, 1nd edition.

Kao, J. and Jurafsky, D. (2012). A computational analysis of style, affect, and imagery in contemporary poetry. In David Elson, et al., editors, *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature*, pages 8–17, Montréal, Canada, june. Association for Computational Linguistics (ACL).

Labádi, G. (2018). Az olvasó gép: Berzsenyi dániel versei távolról. *Digitális Bölcsészet*, 1:17–34.

Lesi, Z. (2008). Automatikus formai verselemzés. *Alkalmazott Nyelvtudomány*, 8(1-2):197–208.

Maróthy, S., Seláf, L., and Plecháč, P., (2021). *Rhyme in 16th-Century Hungarian Historical Songs: A Pilot Study*, pages 43–58. Institute of Czech Literature of the Czech Academy of Sciences, Prague, Czech Republic.

Navarro-Colorado, B., Lafoz, M. R., and Sánchez, N. (2016). Metrical annotation of a large corpus of spanish sonnets: Representation, scansion and evaluation. In Nicoletta Calzolari, et al., editors, *Proceedings of the Tenth edition of the Language Resources and Evaluation Conference (LREC 2016)*, pages 4360–4364, Portorož, Slovenia, may. European Languages Resources Association (ELRA).

Navarro-Colorado, B. (2015). A computational linguistic approach to spanish golden age sonnets: Metrical and semantic aspects. In Anna Feldman, et al., editors, *Proceedings of the Fourth Workshop on Computational Linguistics for Literature*, pages 105–113, Denver, Colorado, june. Association for Computational Linguistics (ACL).

Oravecz, C., Váradi, T., and Sass, B. (2014). The hungarian gigaword corpus. In Nicoletta Calzolari, et al., editors, *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014)*, pages 1719–1723, Reykjavik, Iceland, may. European Languages Resources Association (ELRA).

Plamondon, M. R. (2006). Virtual verse analysis: Analysing patterns in poetry. *Literary and Linguistic Computing*, 21(1):127–141.

Plecháč, P. and Kolár, R. (2015). The corpus of czech verse. *Studia Metrica et Poetica*, 2(1):107–118.

Sass, B., (2017). *Keresés korpuszban: a kibővített Magyar történeti szövegtár új keresőfelülete*, pages 267–277. Szegedi Tudományegyetem Magyar Nyelvészeti Tanszék, Szeged, Hungary.

Tanasescu, C., Paget, B., and Inkpen, D. (2016). Automatic classification of poetry by meter and rhyme. In Zdravko Markov et al., editors, *Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference*, Key Largo,

Florida, may. Artificial Intelligence Research Society.

Vincze, V., Simkó, K., Szántó, Z., and Farkas, R. (2017). Universal dependencies and morphology for hungarian – and on the price of universality. In Mirella Lapata, et al., editors, *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017): Volume 1. Long papers*, pages 355–364, Valencia, Spain, april. Assosiation for Computational Linguistics (ACL).

Voigt, V. (1972). Számítógépes ritmuselemzési kísérlet. *Irodalomtörténeti Közlemények*, 76(2):203–211.

Váradi, T., Simon, E., Sass, B., Mittelholcz, I., Novák, A., and Indig, B. (2018). e-magyar – a digital language processing system. In Nicoletta Calzolari, et al., editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*, pages 1307–1312, Miyazaki, Japan, may. European Language Resources Association (ELRA).

10. Language Resource Reference

Horváth, I. and Font, Zs. and H. Hubert, G. and Herner, J. and Szőnyi, E. and Vadai, I. (1979). *Répertoire de la poésie hongroise ancienne*.