ELTE Verskorpusz – a magyar kanonikus költészet gépileg annotált adatbázisa

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

[1] Eötvös Loránd Tudományegyetem, Bölcsészettudományi Kar
1088 Budapest, Múzeum krt. 4., 4–6.
{horvath.peter,indig.balazs,szlavich.eszter,
lindner.zsofia,vida.bence,karabulut.aslihan,timari.maria,
palko.gabor}@btk.elte.hu
[2] Bölcsészettudományi Kutatóközpont, Irodalomtudományi Intézet
1118 Budapet, Ménesi út 11–13.
{fellegi.zsofia,bajzat.timea}@abtk.hu
peter.kundrath@gmail.com

Kivonat: Az ELTE Verskorpusz a magyar kanonikus költészet vizsgálatára létrejött, a versek szerkezeti egységeinek, a szavak bizonyos grammatikai és fonológiai jellemzőinek, valamint a vershangzás bizonyos jellemzőinek az automatikusan létrehozott annotációit tartalmazó, online elérhető lekérdezőfelülettel rendelkező adatbázis. A tanulmányban bemutatjuk az ELTE Verskorpusz fontosabb mennyiségi jellemzőit, a verskorpusz létrehozásának főbb lépéseit, valamint az egyes annotálási lépésekhez használt eszközöket. Részletesebben ismertetjük a vershangzás annotálásának módját, valamint a verskorpusz formátumát. Emellett bemutatjuk a korpuszhoz készült lekérdezőfelület főbb funkcióit is.

1. Bevezetés és kapcsolódó munkák

Az ELTE Verskorpusz[1] egy olyan online lekérdezőfelülettel[2] rendelkező adatbázis, amely különböző típusú annotációkkal ellátva tartalmazza a kanonikusnak tekintett magyar költészet jelentős részét. Korábbi, magyar nyelvű versek hangzásjellemzőit is tartalmazó, automatikusan annotált korpusz létezéséről nincs tudomásunk. A korpusz létrehozása során ugyanakkor támaszkodhattunk az olyan általánosabb célú, magyar nyelvű szövegeket tartalmazó korpuszokra, mint a Magyar Nemzeti Szövegtár[3] (Oravecz és mtsai, 2014) vagy a Magyar Történeti Szövegtár[4] (Csengery, 2006; Sass, 2017). Bár nem tartalmaz annotált szövegeket, előzményként megemlítendő *A régi magyar vers leltára a kezdetektől 1600-ig/Répertoire de la poésie hongroise ancienne* című

[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

adatbázis is, amely magyar versek különböző adatait, többek között a metrikai jellemzőit tartalmazza kereshető formában.[5] A verskorpusz létrehozása során fontos kiindulópont volt számunkra a Cseh Tudományos Akadémia által fejlesztett Cseh verskorpusz (Korpus českého verše)[6], amely közel 80 000 annotált verset tartalmaz a 19. századból és a 20. század elejéről. A korpusz a szavak lemmája, szófaja, morfológiai és fonológiai jellemzői mellett a ritmusra és a rímekre vonatkozó automatikusan létrehozott annotációkat is tartalmaz (Plecháč és Kolár, 2015; Ibrahim és Plecháč, 2011). Ugyancsak támaszkodtunk az ELTE Verskorpuszhoz hasonlóan TEI XML formátumban készült, 16. és 17. századi spanyol szonetteket tartalmazó Corpus de Sonetos del Siglo de Oro (Corpus of Spanish Golden-Age Sonnets)[7] nevű verskorpusz megoldásaira, amelyben ritmusra vonatkozó, automatikusan létrehozott annotációk is szerepelnek (Navarro-Colorado, 2015; Navarro-Colorado és mtsai, 2016).[8]

Az ELTE Verskorpusz jelenleg 49 költő összes versét tartalmazza.[9] Ez összesen 13063 verset jelent, amelyek együttes szószáma durván 2,7 millió. A korpusz fő formátuma TEI XML, amely a különféle digitális bölcsészeti és nyelvészeti adatbázisok egyik legelterjedtebb formátuma. A versek annotációit gépileg hoztuk létre. A verseket három szinten annotáltuk. Egyrészt annotáltuk a versek szerkezeti egységeit: a címeket, a versszakokat és a verssorokat. Emellett tokenizáltuk a versszövegeket, és annotáltuk a szavak grammatikai jellemzőit, azaz a szavak lemmáját, szófaját és morfoszintaktikai tulajdonságait. Végezetül a versek vershangzáshoz kapcsolódó, formailag egyszerűbben megragadható jellemzőit is címkéztük, ami a rímképletek, a rímpárok, a ritmus, az alliterációk és a szavak fontosabb fonológiai jellemzőinek az annotálását jelentette.

A 2. részben bemutatjuk a korpusz létrehozásának lépéseit és az egyes lépésekhez használt eszközöket. A 3. részben ismertetjük a vershangzás automatikus annotálásának a módját, a 4. részben pedig a korpusz formátumát. Végezetül az 5. részben röviden ismertetjük a korpuszhoz készített, online elérhető lekérdezőfelület főbb funkcióit.

2. A korpusz létrehozásának lépései

A korpusz jelenlegi tartalmának forrását a Magyar Elektronikus Könyvtár[10] adatbázisában megtalálható, a public domain körébe tartozó szerzők összes versét tartalmazó dokumentumfájlok adták. Ezek a dokumentumok jellemzően többféle formátumban is megtalálhatók a MEK oldalán. Elsődlegesen az RTF-formátumú fájlokat használtuk, ha ez nem állt rendelkezésre, akkor a HTML-formátumút. Első lépésben a MEK oldaláról leszedett fájlokból egy szkripttel létrehoztuk a versek szerkezeti egységeinek az annotációit tartalmazó TEI XML fájlokat. Az RTF-formátumú dokumentumok esetében ehhez egy XQuery szkriptet, a HTML-formátumú szövegek esetében pedig egy

[5] https://f-book.com/rpha.

[6] https://versologie.cz/v2/web_content/corpus.php?lang=en; https://github.com/versotym/corpusCzechVerse

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

[8] További verskorpuszokról és azok típusairól lásd Dodé et al. 2018.

[9] A korpusz Radnóti kivételével tartalmazza többek között a középiskolai tananyagban szereplő, hetvenn

[10] http://mek.oszk.hu

Python szkriptet használtunk. A szkriptek futtatásának eredményeképpen mindegyik vers belekerült egy külön TEI XML fájlba, amely XML elemekként tartalmazza a címekre, versszakokra és verssorokra vonatkozó annotációkat. Ugyancsak ebben a fázisban hoztuk létre automatikusan a TEI XML fájlok <teiHeader> elemeit, amely tartalmazza a fontosabb metaadatokat, például a vers szerzőjét és címét, valamint az eredeti MEK-forrás fontosabb adatait.

Ezt követően a szerkezeti egységek annotációit tartalmazó TEI XML fájlokat kézzel ellenőriztük, azaz összevetettük őket a kiinduló RTF- vagy HTML-formátumú dokumentumokkal. Erre azért volt szükség, mert a szkriptek bizonyos speciális eseteket nem tudtak helyesen annotálni, illetve a kiinduló dokumentumfájlokban is voltak olyan inkonzisztens megoldások, amelyek annotációs hibához vezettek. A kézi ellenőrzéshez az Oxygen XML Editor programot[11] használtuk.

Ezt követően annotáltuk, úgyszintén automatikusan, a vers szavainak a grammatikai jellemzőit, azaz a lemmát, a szófajt és a morfoszintaktikai tulajdonságokat. Ehhez az e-magyar elemzőlánc emtsv változatát használtuk (Váradi és mtsai, 2017; Indig és mtsai, 2019; Simon és mtsai, 2020), amelyet egy saját Python szkriptbe ágyazva futtattunk. Erre azért volt szükség, mert az e-magyar bemenete TXT, a kimenete pedig TSV, a verskorpusz formátuma azonban TEI XML. A szkript kiszedi a TEI XML-ből a szöveget, a szövegen lefuttatja az e-magyart, majd az e-magyar TSV-formátumú elemzését visszaalakítja TEI XML-lé. A szófaji és morfoszintaktikai jellemzőket az e-magyar Universal Dependenciesnek (UDv1)[12] megfelelő kimenetével annotáltuk, amely a grammatikai jellemzők annotálásának a legelterjedtebb címkézési rendszere (Vincze és mtsai 2017).

Az annotációs folyamat következő lépése a vershangzáshoz kapcsolódó, formailag egyszerűbben megragadható jellemzőknek, a rímnek, a rímpároknak, a sorok ritmusának, az alliterációknak és a szavak fontosabb fonológiai jellemzőinek az annotálása volt. Ehhez egy saját fejlesztésű, kifejezetten a projekt számára készített, Python nyelvben írt programot használtunk.

A korpusz utolsó annotálási lépéseként egy XSLT stíluslap segítségével elvégeztünk néhány átalakítást az XML fájlokban szereplő annotációk pozícióján, átneveztünk bizonyos XML elemeket és attribútumokat, illetve további, a versek, versszakok és sorok szó- és szótagszámára vonatkozó információkkal bővítettük a már meglévő annotációkat.[13] Az így előálló XML fájlok bár TEI-közeliek, de nem felelnek meg a TEI specifikációjának. Erre az utolsó annotálási fázisra azért volt szükség, mert a TEI által specifikált formátum a részletesebb annotációk tárolására kevésbé alkalmas. E lépés eredményeképpen előállt egy olyan formátuma is a verskorpusznak, amelyben az annotációkat tartalmazó elemek és attribútumok a lehető legegyértelműbb módon utalnak a tartalmukra, illetve amely esetében az annotációk logikusabb elrendezése, valamint a bevezetett további annotációk miatt egyszerűbben meg lehet írni a keresőkifejezéseket, és azokat gyorsabban le lehet futtatni.

Az alábbi felsorolás mutatja be a korpusz létrehozásának a leírt lépéseit.

[11] https://www.oxygenxml.com

[12] https://universaldependencies.org/treebanks/hu_szeged/index.html

[13] A bevezetett, illetve átnevezett elemekről és attribútumokról lásd a https://github.com/ELTE-DH/poetry-corpus oldalon szereplő leírást.

level0: szerkezeti egységek annotálása
Bemenet: RTF, HTML
Kimenet: TEI XML
Eszköz: XQuery szkript, Python szkript

level1: szerkezeti egységeket tartalmazó TEI XML-ek kézi ellenőrzése
Kimenet: TEI XML
Eszköz: manuális, Oxygen XML Editor használatával

level2: tokenizálás, lemmatizálás, szófaji és morfológiai annotálás
Kimenet: TEI XML
Eszköz: Python szkriptbe ágyazott e-magyar

level3: a vershangzás jellemzőinek annotálása
Kimenet: TEI XML
Eszköz: a projekthez fejlesztett Python program (hunpoem_analyzer-TEI)

level4: formátumátalakítás és az annotációk bővítése
Kimenet: XML
Eszköz: XSLT stíluslap

A korpusz annotálására kialakított folyamat lehetővé teszi, hogy a szerkezeti annotációkat tartalmazó, kézileg ellenőrzött TEI XML fájlokon bármikor újrafuttassuk a további annotációs lépéseket, amennyiben azok valamelyikén valamilyen fejlesztés vagy hibajavítás történt.

3. A vershangzás gépi annotálása

Angol nyelvű versek hangzástulajdonságainak a gépi felismertetésére számos programot írtak. Ilyen például a Scandroid (Hartman, 2005) és a ZuScansion (Agirrezabal és mtsai, 2016) nevű eszköz, amelyek angol nyelvű versek ritmusát és metrumát ismerik fel, vagy az AnalysePoem (Plamondon, 2006) nevű eszköz, amely angol nyelvű versek ritmusának, metrumának és rímképletének a felismerésére képes. Több, angol nyelvű versek hangzástulajdonságainak az automatikus elemzésére épülő kutatás is született az utóbbi években. Kao és Jurafsky (2012) kutatása például professzionális és amatőr amerikai versek különbségeit vizsgálta, amelynek során a szókincs mellett olyan vershangzáshoz kapcsolódó, automatikusan elemzett jellemzőket is felhasználtak, mint a versekben szereplő alliterációk vagy rímpárok. De megemlíthető Tanasescu és mtsai (2016) kutatása is, amely angol nyelvű versek ritmus és rím alapján történő automatikus osztályozására irányult.

Magyar nyelvű versek hangzástulajdonságainak az automatizált, valamilyen számítógépes programmal végzett elemzésére csak kevés példát találhatunk, ugyanakkor ezek között meglepően koraiak is vannak. Voigt Vilmos 1972-es tanulmánya mutatja

be az első kísérletet magyar nyelvű versek számítógépes ritmuselemzésére. A létrehozott programmal három Szabó Lőrinc-szonettnek ismertették fel a megvalósuló időmértékes ritmusát (Voigt, 1972). Saját korát megelőzte Jékel és Papp (1974) könyve, amely Ady összes versének az algoritmikus úton előállított fonémastatisztikai adatait tartalmazza. Ugyancsak a korai példák között tartható számon Jékel és Szuromi (1980) műve, amely Petőfi 300 verse esetében tartalmazza a szótagok részben gépi úton meghatározott, különböző típusú nyomatékértékeit, valamint a nyomatékértékek automatikusan előállított összegzését és különféle statisztikáit. Szorosan kapcsolódik a tanulmány témájához Lesi (2006, 2008) kutatása is, aki tudomásunk szerint elsőként hozott létre olyan többfunkciós programot, amely magyar nyelvű versek rímképletének, alliterációinak és metrumának a gépi elemzésére is alkalmas. Érdemes utalni Labádi (2018) Berzsenyi verseiről írt tanulmányára is, amelyben automatizált módszerekkel végzett, a versek szókészletére és fonémajellemzőire (szóhosszúság, magánhangzók és mássalhangzók eloszlása) vonatkozó vizsgálatok eredményei szerepelnek.

Az ELTE Verskorpuszban szereplő versek hangzásának gépi annotálásához a hunpoem_analyzer-TEI elnevezésű, Python nyelvben írt, saját fejlesztésű programot használtuk (Horváth, 2020a, 2020b). A programmal a sorok időmértékes ritmusát, a versszakok rímképletét, a rímpárokat, az alliterációkat és a szavak fontosabb fonológiai jellemzőit annotáltuk.

A sorok időmértékes ritmusának, azaz a hosszú és rövid szótagoknak az annotálása néhány egyszerű, a magyar verstanban jól ismert szabály alapján elvégezhető, így nem volt szükséges kiejtésszótárakat beépíteni az algoritmusba. Ezek a szabályok a következők: (1) a program rövid szótagként elemzi azokat a szótagokat, amelyekben rövid magánhangzó van, és közvetlenül a rövid magánhangzó után nem áll mássalhangzó, vagy csak egy rövid mássalhangzó áll; (2) a program hosszú szótagként elemzi azokat a szótagokat, amelyekben hosszú magánhangzó áll, valamint azokat a rövid magánhangzós szótagokat, amelyekben hosszú vagy egynél több mássalhangzó követi a magánhangzót. E szabályok alkalmazása során a program a bevett verstani hagyománynak megfelelően nem veszi figyelembe az esetleges szóhatárokat. Az annotáló programba be lett építve az a verstani szabály is, miszerint a szó eleji mássalhangzó-torlódások (pl. *krákog, trottyos, strigula*) nem nyújtják meg az előtte lévő rövid magánhangzóra végződő szótagot, vagyis azok nem hosszúnak, hanem rövidnek számítanak. Az elemzés kimenete minden sor esetében egy 0 és 1 karakterekből álló karaktersor, amelyben a 0 a rövid, az 1 pedig a hosszú szótagokat reprezentálja (pl. *Húnyt szemmel bérceken futunk* – 11110101). Mivel a szótagok hosszúságának a megállapításában fontos információ az, hogy a magánhangzót egy vagy több mássalhangzó követi-e, szükséges volt a programba beépíteni annak vizsgálatát is, hogy egy két- vagy háromjegyű mássalhangzónak tűnő karaktersorozat valóban két- vagy háromjegyű mássalhangzónak, azaz egy fonémának tekintendő. Ehhez az e-magyar program morfológiai elemzőjét használtuk. Amennyiben a két- vagy háromjegyű mássalhangzónak tűnő karaktersorozat közé morfémahatár esik, az nem tekinthető egy fonémának.[14]

[14] Például a *gazság* szóra az e-magyar morfológiai elemzője többek között megadja a következő részletes morfológiai elemzési lehetőséget: gaz[/Adj]=gaz+ság[_Nz_Abstr/N]=ság+[Nom]= . Az elemzésből kiderül, hogy a *gaz* és a *ság* karaktersorok között morfémahatár van, azaz a *z* és az *s* nem tekinthető egy zs hangot alkotó rövid mássalhangzónak.

A program a versszakok rímképleteinek az annotálását a hagyományos módon végzi el: minden versszak esetében egy olyan karaktersor jelöli a rímképletet, amelyben a rímelő sorok az ábécé azonos betűjével reprezentálódnak (pl. aabbcb). A program azokat a sorvégeket tekinti egymással rímelőnek, amelyek megegyeznek egymással a sorvégi záró mássalhangzó megléte vagy meg nem léte tekintetében, amelyekben az utolsó szótag magánhangzója a hosszúságot nem számítva megegyezik, illetve amelyekben megegyezik az utolsó előtti szótagok hosszúsága. A rímelés e szabályának alkalmazásában a fő szempont az volt, hogy a szabály ne legyen túl szűkös, de ne is generáljon túl. Mindkét eset ugyanis ahhoz vezet, hogy a konzisztensen, azaz azonos rímképletű versszakokkal is leelemezhető verseket a program nagyobb eséllyel kezelné inkonzisztens módon, vagyis a túl specifikus vagy túl általános szabály alkalmazása miatt bizonyos versszakokat a többihez képest eltérő rímképlettel annotálna.

A rímképlet mellett a program annotálja az egymással rímpárt alkotó, azaz hívó- és felelőrím viszonyban lévő szavakat. A program jelenlegi verziója csak versszakon belül elemez rímpárokat. Egy rímpár szavai között maximum négy sor lehet, például egy hat soros abbcca rímképletű versszakban az első és az utolsó sor rímhelyzetben lévő szavait rímpárként azonosítja a program, de például egy abbccca rímképletű hétsoros versszak első és utolsó sorának rímhelyzetben lévő szavait már nem elemzi rímpárként. Egy rímhelyzetben lévő szó két rímpárnak is a része lehet, az első rímpárban felelő-, a másodikban hívórímként. Például egy négysoros bokorrím, azaz egy aaaa rímképletű versszak második sorának a rímhelyzetben lévő szava felelőrímként az első sor, hívórímként pedig a harmadik sor sorvégi szavával is rímpárt alkot, ugyanakkor a program jelenlegi elemzésében nem alkot rímpárt a negyedik sor sorvégi szavával, vagyis egy rímelő szó hívórímként mindig csak a hozzá legközelebbi rímelő szóval alkothat rímpárt.

Az alliterációk annotálása során a program nem csupán azokat a szerkezeteket elemzi alliterációként, amelyekben egymást követő szavak ugyanazzal a hanggal kezdődnek, hanem azokat is, amelyekben két ugyanolyan hanggal kezdődő szó közé beékelődik egy másik hanggal kezdődő szó. Minden alliterációként elemzett szerkezet kap egy "a" és "n" betűből álló karaktersort annotációként, amelyben az "a" betű az egymással alliteráló szavakat, az "n" betű pedig az alliteráló szavak közé beékelődő nem alliteráló szavakat jelöli (pl. „Bus donna barna balkonon” – anaa). Az alliterációk elemzése során a program az *a, az, és, s* szavakat stopwordökként kezeli, azaz ezek a szavak egy kételemű alliterációnak nem lehetnek részei, ugyanakkor részei lehetnek olyan kettőnél több elemű alliteráló szerkezeteknek, amelyekben legalább kettő nem stopword előfordul alliteráló szóként.[15]

A szavak fonológiai tulajdonságainak annotálása a szótagszám, a hangrend (magas, mély vagy vegyes), valamint a szó egyszerűsített fonológiai szerkezetének a megadására terjedt ki. A fonológiai szerkezet megadása során a Magyar Nemzeti Szövegtárban alkalmazott megoldást követtük némi módosítással (Oravecz és mtsai, 2014): minden szó kap egy karaktersort, amely c, b, f, B és F karakterekből állhat. Az egyes karakterek a szó hangjainak néhány fontosabb fonológiai tulajdonságát jelölik. Ezek a következők:

[15] Csak versszakon belül elemeztünk alliterációkat, ami azt jelenti, hogy egy versszak utolsó és a következő versszak első, azonos hanggal kezdődő szava nem annotálódik alliterációként.

c – mássalhangzó, b – hátul képzett rövid magánhangzó, f – elöl képzett rövid magánhangzó, B – hátul képzett hosszú magánhangzó, F – elöl képzett hosszú magánhangzó (pl. *szerszámaival* – cfccBcbfcbc).

4. A korpusz formátuma

A korpusz formátuma – leszámítva az utolsó annotációs fázissal létrejövő XML-eket – TEI XML. A TEI XML a digitális bölcsészeti és nyelvészeti korpuszok egyik széles körben elterjedt sztenderdje, amely számos szövegtípus, többek között versek annotálására is felkínál elemkészletet (TEI Consortium, 2021). Az alábbiakban bemutatandó level1, level2, level3 és level4 formátumok a korpusz egyre több annotációs réteget tartalmazó verzióinak a formátumai. Ezek megfelelnek a korpusz gitHub oldalán található könyvtáraknak.

4.1. A level1 formátuma

A versek szerkezeti egységeinek az annotálása, illetve az annotációk kézi ellenőrzése révén létrejövő level1-es TEI XML-ek az alábbi módon tartalmazzák a verseket.

```
<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>
```

A versek címei a <head> elembe, a versszakok az <lg> elembe, a sorok pedig az <l> elembe kerülnek. A mottók, a különböző szeparátorelemek, illetve a versek keletkezésének helyére és idejére vonatkozó megjegyzések pedig a <p> elemben vannak.

4.2. A level2 formátuma

Az e-magyar lefuttatásával létrejövő, a szavak grammatikai annotációit tartalmazó level2-es TEI XML fájlok esetében minden szó belekerül egy külön <w> elembe, amelyek attribútumokként tartalmazzák a szavak e-magyarral felismertetett jellemzőit. A @lemma attribútumba kerül a szó szótári alakja, a @pos attribútumba a szó szófajának a címkéje, az @msd attribútumba pedig a szó morfoszintaktikai jellemzői az Universal Dependencies rendszerének megfelelő tulajdonság-érték párokként.[16] A központozások <pc> elembe kerülnek, amelynek a @join attribútuma jelzi a tapadás irányát. Az <lg>, <l>, <w> és <pc> elemek az @xml:id attribútum értékeként egy egyedi azonosítót is kapnak ebben az annotációs fázisban. Az alábbi XML-részlet az idézett Babits-vers harmadik sora alapján mutatja be az említett annotációk korpuszbeli szerepeltetésének módját.

```
<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>
```

4.3. A level3 formátuma

A hunpoem_analyzer-TEI program lefuttatásával létrejövő level3-as TEI XML fájlokba további, a vershangzás annotációit tartalmazó elemek és attribútumok kerülnek bele.

```
<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>
    […]
```

[16] Az attribútumok megnevezéseit a TEI specifikáció írja elő.

Az <lg> elemek @rhyme attribútumában szerepel a versszak rímképlete, az <l> elem @n attribútumában a sor szótagszáma, a @real attribútumban pedig a sor időmértékes ritmusa. A TEI alapsémája nem teszi lehetővé, hogy a szavak fonológiai jellemzőit a <w> elem attribútumaiként annotáljuk, így azokat standoff módon, azaz a versszövegtől leválasztva, az XML fájlnak egy későbbi részén annotáltuk az alábbi módon.

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

A <spanGrp type="phonStructures"> elemben szereplő elemek egy-egy szó főbb fonológiai jellemzőinek az annotációit tartalmazzák. A @target attribútum értéke annak a szónak az xml:id-je, amire az annotáció vonatkozik. A @subtype attribútum értéke a szó szótagszáma, a @type attribútum értéke a szó hangrendje, magának a elemnek a tartalma pedig a szó egyszerűsített fonológiai reprezentációja.

Ugyanígy, standoff módon annotáltuk a rímpárokat és az alliterációkat is. A rímpárokat a <linkGrp type="rhymePairs"> elemben lévő <link> elemek annotálják. Minden <link> elem egy adott rímpárt annotál oly módon, hogy a @target attribútum értékeként szereplő két xml:id utal a rímpárt alkotó két szóra.

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

Az alliterációk annotációi hasonló módon szerepelnek, a <spanGrp type="alliterations"> elem egyes elemei egy-egy alliterációt annotálnak oly módon, hogy a @target attribútum értékei utalnak az alliteráló szerkezetet alkotó szavak xml:id-ire. A @type attribútumban egy "a" és "n" betűből álló karaktersor áll. Az "a" karakterek az alliteráló, azaz ugyanolyan hanggal kezdődő szavakat reprezentálják, az "n" karakterek pedig az alliteráló szavak közé beékelődő nem alliteráló, más hanggal kezdődő szavakat.

```
<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>
```

4.4. A level4 formátuma

Az annotációs folyamat utolsó lépésével létrehozott level4-es fájlokban több szempontból is eltértünk a TEI XML sémától annak érdekében, hogy a versek olyan formátumban is meglegyenek, amelyen egyszerűbben és gyorsabban lehet lekérdezéseket végrehajtani. Egyrészt az érthetőség kedvéért több elemnek és attribútumnak is megváltoztattuk a nevét úgy, hogy az egyértelműen utaljon az általa tartalmazott annotáció típusára. Másrészt a level3 TEI XML fájljaiban standoff módon annotált fonológiai jellemzőket áthelyeztük a szavak <w> elemeiben szereplő attribútumokba. Harmadrészt bővítettük a versek szerkezeti egységeinek az annotációit versszakszámra, sorszámra, szószámra és szótagszámra vonatkozó annotációkkal.

```
<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>
```

A fenti level4-es XML példából látható, hogy a elemben szereplő @div_numStanza attribútum értékeként tüntettük fel a versszakok számát, a @div_numLine attribútum értékeként a vers sorainak a számát, a @div_numWord attribútum értékeként a vers szavainak a számát, a div_numSyll attribútum értékeként pedig a vers szótagjainak a számát. A @div_numShortSyll és a @div_numLongSyll attribútumokban szerepel a rövid és a hosszú szótagok száma. Az egész versnek a rímképletét a @div_rhyme attribútum tartalmazza. A @div_rhyme attribútum értékeként megadott karaktersorokban virgula választja el az egyes versszakokra vonatkozó rímképleteket. Az azonos betűk csak egy versszakon belül jelölnek egymással rímelő sorokat. A @div_syllPattern attribútumban szerepel a vers szótagmintája, amely egy kötőjelekkel és virgulákkal elválasztott, számokból álló karaktersor, amelyben a számok a vers egyes sorainak a szótagszámát jelölik. A versszakokat tartalmazó <lg> elemeket és a sorokat tartalmazó <l> elemeket hasonló módon bővítettük további, sorszámra, szószámra és szótagszámra vonatkozó attribútumokkal.

A szavaknak a level3-as TEI XML verzióban standoff módon annotált fonológiai jellemzői a level4-es verzióban – ahogyan a példában is látható – a <w> elemek attribútumai közé kerülnek. A @w_numSyll attribútumban szerepel a szó szótagszáma, a

@phonType attribútumban a szó hangrendje, a @phonStruct attribútumban pedig a szó fonológiai reprezentációja.

A rímpárok standoff annotációiban az egyértelműség kedvéért megváltoztattuk az elemek neveit, illetve a lekérdezések megkönnyítése érdekében a rímpárok tagjaira vonatkozó elemek tartalmaként feltüntettük a szóalakokat, attribútumokként pedig a szavak annotált jellemzőit. Az alliterációk esetében is megváltoztattuk az alliterációkat tartalmazó elemek neveit, a lekérdezések megkönnyítése érdekében pedig az elemek tartalmaiként itt is feltüntettük az alliterációt alkotó szavakat, attribútumokként pedig az alliterációban szereplő szavak szófaját, morfoszintaktikai jellemzőit és lemmáit.

5. A lekérdezőfelület

A verseket és azok annotációit tartalmazó level4-es XML fájlokból létrehoztunk egy MariaDB-alapú SQL-adatbázist, ebben keres a verskorpusznak a https://verskorpusz.elte-dh.hu oldalon elérhető, bárki által szabadon használható online lekérdezőeszköze. A lekérdezőeszköz amellett, hogy az egyes versek annotált jellemzőit megjeleníti, számos keresőfunkcióval is rendelkezik. A keresőfunkciók megtervezése során nagymértékben támaszkodhattunk a már létező magyar nyelvű korpuszok lekérdezőfelületeire, különösen a Magyar Nemzeti Szövegtár lekérdezőfelületére (Oravecz és mtsai, 2014). Kereshetünk szóalakokra, lemmákra, szófajokra, morfoszintaktikai jellemzőkre, illetve ezek tetszőleges kombinációjából alkotott szókapcsolatokra is. A fonológiai jellemzők és a ritmus annotálásának köszönhetően szavak szótagszáma, hangrendje, egyszerűsített fonológiai reprezentációja és a szavak szótagjainak hosszúsága alapján is végezhetünk lekérdezéseket, illetve ezeket kombinálhatjuk a szóalakra, lemmára és morfoszintaktikai jellemzőkre vonatkozó keresőkifejezésekkel. A szerzők mellett rímképletek alapján is szűrhetjük a verseket.

A lekérdezőfelülettel a megadott keresési feltételek alapján szóalakokra és lemmákra vonatkozó gyakorisági listákat is generálhatunk. Amennyiben a keresőmezőben több szóból álló szerkezetre vonatkozó keresési kifejezést adunk meg, akkor e szerkezetre vonatkozóan is létrehozhatunk gyakorisági listát. Az 1. táblázat három, a lekérdezőfelülettel generált gyakorisági listát mutat be: a verskorpuszban szereplő leggyakoribb öt rímhelyzetben lévő főnevet, a leggyakoribb öt *bús* + főnév szókapcsolatot, valamint a leggyakoribb öt alliteráló melléknév + főnév szókapcsolatot. A megadott szófaji kategóriákba tartozó szavak lekérdezése lemmák alapján történt.

	rímhelyzetben lévő főnevek		*bús* + főnév szókapcsolatok		alliteráló melléknév + főnév szókapcsolatok	
1	élet	2585	bús szív	63	szép szem	221
2	szem	2362	bús szem	56	szép szó	129
3	világ	2350	bús lélek	46	nagy név	93
4	ég	2099	bús fej	45	kis kéz	50
5	kéz	2002	bús dal	36	szép szerelem	30

1. táblázat. A verskorpusz leggyakoribb rímhelyzetben lévő főnevei, *bús* + főnév szókapcsolatai, valamint alliteráló melléknév + főnév szókapcsolatai.

A keresési találatokat, az azokhoz kapcsolódó fontosabb kvantitatív adatokat, a gyakorisági listákat, illetve a vizsgálat számára kijelölt alkorpusz legfontosabb kvantitatív jellemzőit TSV-formátumban letölthetjük, és bármelyik táblázatkezelő programban megnyithatjuk. A lekérdezőfelülethez tartozik egy részletes használati útmutató is, amely bemutatja az egyes keresési funkciókat.

6. Összegzés

Az ELTE Verskorpusz építésével a célunk egy olyan, bárki számára elérhető annotált korpusz létrehozása volt, amely reményeink szerint mind az irodalomtudományos, mind pedig a nyelvészeti kutatásokat segítheti. A korpuszhoz készült lekérdezőfelület lehetővé teszi, hogy különösebb informatikai tudás nélkül is olyan információkhoz juthassunk a magyar kanonikus költészetről, amelyhez a szoros olvasás eljárásai révén nem juthatnánk. Bízunk abban, hogy a lekérdezőfelület révén a korpusz nemcsak a kutatásban, hanem egyéb színtereken, például a közoktatásban is hasznosítható lesz. Az annotált XML fájlok a projekt gitHub oldaláról (https://github.com/ELTE-DH/poetry-corpus) letölthetőek, és kutatás számára szabadon felhasználhatóak. Az XML fájlok közzétételének köszönhetően a programozási tudással rendelkező kutatónak lehetősége van arra, hogy olyan összetettebb lekérdezéseket is elvégezhessen a korpuszon, amelyeket az online lekérdezőfelület nem tesz lehetővé. Az ELTE Verskorpusz nem egy lezárt projekt, a jövőben szeretnénk további szerzőkkel és annotációs rétegekkel bővíteni a korpuszt, illetve további funkciókkal kiegészíteni a lekérdezőfelületet.

Köszönetnyilvánítás

Az ELTE Verskorpusz elkészítését a Felsőoktatási Intézményi Kiválósági Program és a Digitális Örökség Nemzeti Laboratórium támogatta.

Hivatkozások

Agirrezabal, M., Astigarraga, A., Arrieta, B., Hulden, M.: ZeuScansion: A Tool for Scansion of English Poetry. Journal of Language Modelling 4(1), 3–28 (2016)

Csengery, K.: Az elektronikus korpusz. In: Ittzés, N. (szerk.) A magyar nyelv nagyszótára 1. Segédletek. pp. 18–19. MTA Nyelvtudományi Intézet, Budapest (2006)

Dodé, R., Ludányi Zs., Falyuna, N., Kuna, Á.: Poétika és korpusz. Hogyan nyújthat segítséget a korpusznyelvészet poétikus szövegek vizsgálatához? In: Domonkosi, Á., Simon, G. (szerk.) Nyelv, poétika, kogníció. Elmélet és módszer a poétikai kutatásban. pp. 175–196. Líceum Kiadó, Eger (2018)

Hartman, C., O.: The Scandroid. Version 1.1. [User guide] http://charlesohartman.com/verse/scandroid/ScandroidManual.pdf (2005)

Horváth, P.: A vershangzás jellemzőinek automatikus feltárása József Attila verseiben. Digitális Bölcsészet 3, M:3–M:27 (2020a)

Horváth, P.: Az ELTE Verskorpusz automatikus annotációs eljárásai révén nyerhető kvantitatív adattípusok. In: Simon, G., Tolcsvai Nagy, G. (szerk.) Nyelvtan, diskurzus, megismerés. pp. 313–332. Eötvös Kiadó, Budapest (2020b)

Ibrahim, R., Plecháč, P.: Toward Automatic Analysis of Czech Verse. In: Scherr, B. P., Baily, J., Kazartsev, E. V. (szerk.) Formal Methods in Poetics. pp. 295–305. RAM, Lüdenscheid (2011)

Indig, B., Sass B., Simon, E., Mittelholcz, I., Kundráth, P., Vadász, N.: emtsv – egy formátum mind felett. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 235–247. Szegedi Tudományegyetem TTIK, Informatikai Intézet, Szeged (2019)

Jékel, P., Papp, F.: Ady Endre összes költői műveinek fonémastatisztikája. Akadémiai Kiadó, Budapest (1974)

Jékel, P., Szuromi, L.: Petőfi metrumai. Kossuth Lajos Tudományegyetem, Debrecen (1980)

Kao, J., Jurafsky, D.: A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry. In: Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature. pp. 8–17. Association for Computational Linguistics, Montréal (2012)

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

Lesi, Z.: Automatikus verselemzés tanuló algoritmusok alkalmazásával. In: Alexin, Z., Csendes, D. (szerk.) IV. Magyar Számítógépes Nyelvészeti Konferencia. pp. 402–407. Szegedi Tudományegyetem Informatikai Tanszékcsoport, Szeged (2006)

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

Navarro-Colorado, B.: A Computational Linguistic Approach to Spanish Golden Age Sonnets: Metrical and Semantic Aspects. In: Proceedings of the Fourth Workshop on Computational Linguistics for Literature. pp. 105–113. Association for Computational Linguistics (ACL), Denver (2015)

Navarro-Colorado, B., Ribes Lafoz, M., Sánchez, N.: Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation. In: Calzolari, N., Choukri, K., Declerck, T. et al. (szerk.) Proceedings of the 10th edition of the Language Resources and Evaluation Conference (LREC 2016). pp. 4360–4364. European Languages Resources Association (ELRA), Portorož (2016)

Oravecz, Cs., Váradi, T., Sass, B.: The Hungarian Gigaword Corpus. In: Calzolari, N., Choukri, K., Declerck, T., Loftsson, H., Maegaard, B., Mariani, J., Moreno, A., Odijk, J., Piperidis, S. (szerk.) Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC-2014). pp. 1719–1723. European Languages Resources Association (ELRA), Reykjavik (2014)

Plamondon, M. R.: Virtual Verse Analysis: Analysing Patterns in Poetry. Literary and Linguistic Computing 21(1), 127–141 (2006)

Plecháč, P., Kolár, R.: The Corpus of Czech Verse. Studia Metrica et Poetica 2(1), 107–118 (2015)

Sass, B.: Keresés korpuszban: a kibővített Magyar történeti szövegtár új keresőfelülete. In: A nyelvtörténeti kutatások újabb eredményei. pp. 267–277. Szegedi Tudományegyetem Magyar Nyelvészeti Tanszék, Szeged (2017)

Simon, E., Indig, B., Kalivoda, Á., Mittelholcz, I., Sass, B., Vadász, N.: Újabb fejlemények az e-magyar háza táján. In: Berend, G., Gosztolya, G., Vincze, V. (szerk.) XVI. Magyar Számítógépes Nyelvészeti Konferencia. pp. 29–42. SZTE Informatikai Intézet, Szeged (2020)

Tanasescu, C., Paget, B., Inkpen, D.: Automatic Classification of Poetry by Meter and Rhyme. In: Proceedings of the Twenty-Ninth International Florida Artificial Intelligence Research Society Conference. Florida Artificial Intelligence Research Society. https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS16/paper/view/12923/12883 (2016)

TEI Consortium: TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.5.0. https://tei-c.org/release/doc/tei-p5-doc/en/Guidelines.pdf (2019)

Váradi, T., Simon, E., Sass, B., Gerőcs, M., Mittelholtz, I., Novák, A., Indig, B., Prószéky, G., Vincze, V.: Az e-magyar digitális nyelvfeldolgozó rendszer. In: Vincze, V. (szerk.) XIII. Magyar Számítógépes Nyelvészeti Konferencia. pp. 49–60. Szegedi Tudományegyetem, Informatikai Intézet, Szeged (2017)

Vincze, V., Simkó, K., Szántó, Zs., Farkas, R.: Universal Dependencies and Morphology for Hungarian – and on the Price of Universality. In: Mirella, L., Phil, B., Alexander, K. (szerk.) Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics (EACL 2017): Volume 1. Long papers. pp. 356–365. Assosiation for Computational Linguistics (ACL), Valencia (2017)

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