**Horváth Péter**

*ELTE Digitális Bölcsészet Központ*

horvath.peeteer@gmail.com

# A vershangzás jellemzőinek automatikus feltárása József Attila verseiben*

A tanulmány József Attila versei alapján mutatja be a vershangzáshoz kapcsolódó tulajdonságok automatikus feltárásának egy módszerét. A tanulmány első fele ismerteti a vershangzás jellemzőinek automatikus annotálására létrehozott *hunpoem_analyzer-TEI* program funkcióit, valamint az annotációs folyamat főbb lépéseit. A tanulmány második fele különböző, az elemzett korpuszból kinyert, a versek lexikai tulajdonságaihoz és a vershangzás jellemzőihez kapcsolódó gyakorisági listákat mutat be.

Kulcsszavak:
automatikus verselemzés, korpusznyelvészet, vershangzás, rím, ritmus, József Attila, *hunpoem_analyzer-TEI, e-magyar*

## 1. Bevezetés

Számítógépes, automatikus módszerek használata lehetővé teszi nagy mennyiségű vers elemzését, ami manuálisan nem, vagy csak nagyon hosszú idő alatt lenne elvégezhető. Az elemezhető versek mennyiségének ez az ugrásszerű növekedése újfajta tudományos kérdések megválaszolását teszi lehetővé. Teljes életművekkel, korszakokkal vagy akár költészeti hagyományokkal kapcsolatban fogalmazhatunk meg kvantitatív alapokon nyugvó állításokat. Ennek a tanulmánynak ugyanakkor nem célja József Attila költészetével kapcsolatban irodalomtudományos megállapításokra jutni. Az alábbiakban pusztán azt mutatom be, hogy milyen típusú jellemzőket nyerhetünk ki egy verseket tartalmazó korpuszból számítógépes eszközök segítségével. A József Attila-versek hangzásjellemzőinek az elemzésére használt *hunpoem_analyzer-TEI* elnevezésű programot a készülő ELTE Verskorpusz számára írtam. A program jelenleg is fejlesztés alatt áll, vagyis a tanulmányban bemutatott funkciói nem tekinthetők véglegesnek.

A tanulmány második részében bemutatok néhány nemzetközi példát versek automatikus elemzésére, a harmadik részben pedig ismertetem a vershangzás jellemzőinek

* Köszönöm a tanulmány két névtelen lektorának a részletes javaslatokat. Javaslataik nem csupán ennek a tanulmánynak a megírásában segítettek sokat, hanem a vershangzást elemző program továbbfejlesztésében is nagyon hasznosak lesznek.

az elemzésére használt *hunpoem_analyzer-TEI* program funkcióit. A negyedik részben röviden áttekintem a József Attila-korpusz létrehozásának és automatikus elemzésének főbb lépéseit. A tanulmány ötödik részében néhány, a versek szókészletére, illetve grammatikai tulajdonságaira vonatkozó jellemzőt mutatok be, melyek az MTA Nyelvtudományi Intézetében fejlesztett *e-magyar* programnak a verseken való lefuttatásával váltak kinyerhetővé. A hatodik részben a korpusznak a *hunpoem_analyzer-TEI* program futtatása révén kinyert, vershangzáshoz kapcsolódó jellemzőit ismertetem, az utolsó, hetedik részben pedig röviden összegzem a tanulmányban bemutatott módszert.

## 2. Néhány példa a vershangzás jellemzőinek automatikus elemzésére

Grammatikai tulajdonságok automatikus elemzése Magyarországon is jól ismert eljárás. Például a magyar nyelvészeti kutatásokban széles körben használt Magyar Nemzeti Szövegtár kereshető formában tartalmazza a szavak automatikus elemzés útján létrehozott grammatikai annotációit.[1] A vershangzáshoz kapcsolódó jellemzők automatikus elemzése azonban kevésbé gyakori eljárás, így az itt ismertetendő módszer előtt érdemes néhány, a vershangzás automatikus elemzését megcélzó nemzetközi példát is röviden bemutatni. Az alábbiakban szereplő eszközök, korpuszok és kutatások az utóbbi tizenöt évhez kapcsolódnak. Hangsúlyozandó, hogy az áttekintés ezen időszakra vonatkozóan sem kimerítő jellegű.

A vershangzáshoz kapcsolódó különböző jellemzők közül kifejezetten intenzíven foglalkoznak a ritmus automatikus elemzésével. A Charles O. Hartman által fejlesztett *Scandroid* nevű program például angol nyelvű versek ritmusát elemzi.[2] A program az elemzés során megkülönbözteti a vers szavainak hangsúlyos és hangsúlytalan szótagjait, majd ez alapján eldönti, hogy a vers jambikusnak vagy anapesztikusnak tekinthető-e inkább, és megállapítja az egyes verslábakat.[3] A hangsúlyos és hangsúlytalan szótagok megkülönböztetése részben általános szabályok alapján, részben pedig a programba beépített lexikon alapján történik. A *Scandroid* program módosított, nagy adatmennyiségen lefuttatható változatát használja Chris Tanasescu, Bryan Paget és Diana Inkpen, akiknek a kutatása angol nyelvű versek ritmus és rím alapján történő automatikus osztályozására irányul.[4]

Az úgyszintén angol nyelvű versek elemzésére fejlesztett *ZuScansion* nevű eszköz a *Scandroid*tól eltérően a jambikus és anapesztikus metrumok mellett egyéb metrumok

[1] Csaba Oravecz, Tamás Váradi and Bálint Sass, „The Hungarian Gigaword Corpus," in *Proceedings of the Ninth International Conference on Language Resources and Evaluation* (LREC-2014), eds., Nicoletta Calzolari, Khalid Choukri, Thierry Declerck et al. (Reykjavik: European Language Resources Association [ELRA], 2014), 1719–1723.

[2] Charles O. Hartman, *The Scandroid. Version 1.1.* [User guide] (2005), hozzáférés: 2020.01.19, `http://charlesohartman.com/verse/scandroid/ScandroidManual.pdf`.

[3] Az angol (és más indoeurópai nyelvek) versritmusa nem a rövid és hosszú szótagok, hanem a hangsúlyos és hangsúlytalan szótagok váltakozására épül.

[4] Chris Tanasescu, Bryan Paget and Diana Inkpen, „Automatic Classification of Poetry by Meter and Rhyme," in *AAAI Publications: The Twenty-Ninth International Flairs Conference (2016),* eds., Zdravko Markov and Ingrid Russell, hozzáférés: 2020.01.19, `https://www.aaai.org/ocs/index.php/FLAIRS/FLAIRS16/paper/view/12923/12883`.

 DOI 10.31400/dh-hun.2020.3.422

elkülönítésére is képes.[5] A metrum megállapításának alapját az elsődleges hangsúlyú, a másodlagos hangsúlyú és a hangsúlytalan szótagok automatikus elkülönítése adja. A program a szótagok hangsúlyait egyrészt két kiejtésszótár alapján, másrészt a lexikai hangsúlyt adott esetben felülíró, a szavak szófajához kapcsolódó prozódiai szabályok alapján állapítja meg. Amennyiben egy szó nem szerepel a program által használt szótárakban, akkor az algoritmus az írásmódja alapján az elemzendő szóra leginkább hasonlító szótári szó hangsúlyviszonyai alapján elemzi azt. A program a versre általánosan jellemző metrumot oly módon állapítja meg, hogy kiszámítja az egyes szótaghelyekre eső szótagok átlagos hangsúlyértékét, majd az ebből absztrahált ritmusképletnek több alternatív, verslábakra tagolt elemzését is megadja, amelyekből egy pontozási rendszer alapján választja ki a legmegfelelőbbet (az algoritmus azonos szótagszámú sorokon működtethető).

Az *AnalysePoem* program angol nyelvű versek ritmusának és rímképletének az elemzésére lett fejlesztve.[6] A ritmus elemzésének első lépéseként a program a beépített és a felhasználó által bővíthető lexikonja alapján, a *ZeuScansion*höz hasonlóan megkülönböztet erősen és gyengén hangsúlyos, valamint hangsúlytalan szótagokat. Amennyiben egy szó nem szerepel a lexikonban, az algoritmus általában meg tud adni a szóra egy valószínűsíthető elemzést. A szótagok hangsúlyai alapján a program hat domináns metrumot képes megállapítani. A domináns metrum megállapításához a program számos tesztet végez el egymást követően, amelynek kimeneteként egy megbízhatósági értéket *(confidence number)* rendel a megállapított metrumhoz. Minél nagyobb a megbízhatósági érték, annál szabályosabb a ritmus, egy bizonyos megbízhatósági érték alatt a vers nem tekinthető az adott metrumhoz tartozónak. Az *AnalysePoem* a versek rímképletét is képes megadni, aminek előfeltétele a ritmus elemzése. A program a rímképletek megfelelő elemzéséhez egy folyamatosan bővülő adatbázist használ, amely a korábban már elemzett, egymással rímelő szavakat tartalmazza.

Justine Kao és Dan Jurafsky kutatása professzionális és amatőr amerikai versek különbségeit vizsgálja egy amatőr és egy professzionális verseket tartalmazó, a kutatás számára létrehozott korpusz különböző nyelvi tulajdonságainak kvantitatív összevetésével.[7] A szókincs elemzése mellett a szerzők a David Kaplan által fejlesztett *PoetryAnalyzer*t használva automatikusan elemzik a versekben szereplő alliterációkat (egymást követő, azonos mássalhangzóval kezdődő szavak), asszonáncokat (ugyanazon magánhangzók ismétlődése), konszonáncokat (ugyanazon mássalhangzók ismétlődése) és a sorvégi rímpárokat. A rímpárok elkülönítése során a program tiszta rímeket *(perfect rhyme)* és nem tiszta rímeket *(slant rhyme)* különböztet meg. Tanasescu, Paget és Inkpen már idézett, a ritmus mellett a sorvégi rímek automatikus elemzésére is kiterjedő kutatása a rímek két csoportja, a tiszta rím *(perfect rhyme)* és a Kao és Jurafsky kutatásánál tágabban értelmezett nem tiszta rím *(slant rhyme)* mellett egy

[5] Manex Agirrezabal, Aitzol Astigarraga, Bertol Arrieta and Mans Hulden, „ZeuScansion: A Tool for Scansion of English Poetry," *Journal of Language Modelling* 4, 1. sz. (2016): 3–28.

[6] Marc R. Plamondon, „Virtual Verse Analysis: Analysing Patterns in Poetry," *Literary and Linguistic Computing* 21, 1. sz. (2006): 127–141, `https://doi.org/10.1093/llc/fql011`.

[7] Justine Kao and Dan Jurafsky, „A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry," in *Proceedings of the NAACL-HLT 2012 Workshop on Computational Linguistics for Literature,* eds., David Elson, Anna Kazantseva, Rada Mihalcea et al. (Montréal: Association for Computational Linguistics, 2012), 8–17.

harmadik csoportot, az *eye rhyme* kategóriát is bevezeti. Ide azok az esetek tartoznak, amikor a szavak az írásmódjuk miatt tiszta rímnek tűnnek, de valójában nem rímelnek (pl. *slaughter-laughter*).

Nem angol nyelvű versek automatikus elemzésére is találhatunk példákat. A versek szerkezeti elemei mellett a ritmus annotálására törekszik például a 16. és 17. századi spanyol szonetteket tartalmazó Corpus de Sonetos del Siglo de Oro (Corpus of Spanish Golden-Age Sonnets).[8] A korpuszban egy erre fejlesztett programmal automatikusan annotálják a verssorokban a hangsúlyos és hangsúlytalan szótagokat, a nem egyértelmű ritmusú, többféleképpen is annotálható verssorokat pedig manuálisan ellenőrzik. Érdemes megemlíteni a Cseh Tudományos Akadémia által fejlesztett Cseh verskorpuszt (Korpus českého verše) is, amely közel 80 000 annotált verset tartalmaz a 19. századból és a 20. század elejéről.[9] A korpusz a szavak lemmája, szófaja, morfológiai és fonológiai jellemzői mellett a ritmusra és a rímekre vonatkozó automatikusan létrehozott annotációkat is tartalmaz.[10] A korpusz honlapján számos lekérdező eszközből választhatunk.[11]

## 3. A *hunpoem_analyzer-TEI* program funkciói

A József Attila-versek hangzásjellemzőinek automatikus elemzésére az általam írt *hunpoem_analyzer-TEI* programot használtam, amely a készülő ELTE Verskorpusz hangzásjellemzőinek az annotálására lett fejlesztve.[12] A Python nyelvben írt program bemenetét a versek szerkezeti egységeinek, azaz a címnek, a versszakoknak és a soroknak a szövegközi *(inline)* annotációit tartalmazó TEI XML-fájlok adják.[13] A program két fő modulból tevődik össze. Az első modul az MTA Nyelvtudományi Intézetében fejlesztett *e-magyar* program *emtsv* nevű változatát futtatja, és a szavaknak az *e-magyar* TSV-kimenetében szereplő grammatikai elemzésével, azaz a szavak lemmá-

[8] Borja Navarro-Colorado, „A Computational Linguistic Approach to Spanish Golden Age Sonnets: Metrical and Semantic Aspects," in *Proceedings of the Fourth Workshop on Computational Linguistics for Literature,* eds., Anna Feldman, Anna Kazantseva, Stan Szpakowicz et al. (Denver: Association for Computational Linguistics, 2015), 105–113, https://doi.org/10.3115/v1/w15–0712; Borja Navarro-Colorado, Marí Ribes Lafoz and Noelia Sánchez, „Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation," in *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference* (LREC 2016), eds., Nicoletta Calzolari, Khalid Choukri, Thierry Declerck et al. (Portorož: ELRA, 2016), 4360–4364.

[9] Petr Plecháč and Robert Kolár, „The Corpus of Czech Verse," *Studia Metrica et Poetica* 2, 1. sz. (2015): 107–118.

[10] Robert Ibrahim and Petr Plecháč, „Toward Automatic Analysis of Czech Verse," in *Formal Methods in Poetics,* eds., Barry P Scherr, James Baily and Evgeny V. Kazartsev (Lüdenscheid: RAM, 2011), 295–305.

[11] Hozzáférés: 2020.01.19, http://versologie.cz/v2/web_content/tools.php?lang=en.

[12] Köszönöm Indig Balázsnak a kód átnézését, és a kevésbé szerencsés kódolási megoldások javítását. Emellett köszönöm neki az *e-magyar* futtatásában nyújtott segítséget.

[13] TEI Consortium, *TEI P5: Guidelines for Electronic Text Encoding and Interchange. Version 3.5.0.* (2019), hozzáférés: 2020.01.19. https://tei–c.org/release/doc/tei–p5–doc/en/Guidelines.pdf.

 DOI 10.31400/dh-hun.2020.3.422

jával (szótári alakjával), szófajával és morfoszintaktikai tulajdonságaival kiegészíti a TEI XML-fájlokat.[14]

A második modul végzi el a vershangzás jellemzőinek elemzését, amelynek során az elemzett jellemzők is bekerülnek annotációként a TEI XML-fájlokba. A program automatikusan megadja a szavak főbb fonológiai tulajdonságait, a versek versszakainak rímképletét, a rímpárokat alkotó szavakat, a verssorok ritmusát, valamint az alliterációkat alkotó szavakat. A szavak fonológiai tulajdonságainak elemzése a szótagszám, a hangrend (magas, mély vagy vegyes),[15] valamint a szó fonológiai szerkezetének a megadását jelenti. A fonológiai szerkezet elemzése során minden szó kap egy C, V, B, F, 1 és 2 karakterekből álló karaktersort, amelyben a karakterek a szó hangjainak néhány fontosabb fonológiai tulajdonságát jelölik. Ezek a következők: C – mássalhangzó, V – magánhangzó, B – hátul képzett magánhangzó, F – elöl képzett magánhangzó, 1 – rövid magánhangzó, 2 – hosszú magánhangzó (pl. „szerszámaival" – C, VF1, C, C, VB2, C, VB1, VF1, C, VB1, C).

A rímképlet elemzése során a program minden versszak esetében megadja annak rímképletét a megszokott módon, az ábécé betűiből álló karaktersorral (pl. aaaa, aabbcc, abcb).[16] A program azokat a sorvégeket tekinti egymással rímelőnek, amelyekben az utolsó szótag magánhangzója a hosszúságot nem számítva megegyezik, valamint 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.[17] 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. Természetesen a későbbiekben a rímelés automatikus elemzése finomítható. Valószínűleg ki lehetne dolgozni olyan algoritmust, amely a konzisztens elemzés elvének a megtartása mellett az előző részben bemutatott példákhoz hasonlóan elkülöníti egymástól a tiszta

[14] Váradi Tamás, Simon Eszter, Sass Bálint, Gerőcs Mátyás, Mittelholtz Iván, Novák Attila, Indig Balázs, Prószéky Gábor és Vincze Veronika, „Az e-magyar digitális nyelvfeldolgozó rendszer," in *XIII. Magyar Számítógépes Nyelvészeti Konferencia,* szerk., Vincze Veronika (Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2017), 49–60; Mittelholcz Iván, „emToken: Unicode-képes tokenizáló magyar nyelvre," in *XIII. Magyar Számítógépes Nyelvészeti Konferencia,* szerk. Vincze, 61–69; Novák Attila, Rebrus Péter, Ludányi Zsófia, „Az emMorph morfológiai elemző annotációs formalizmusa," in *XIII. Magyar Számítógépes Nyelvészeti Konferencia,* szerk. Vincze, 70–78; Indig Balázs, Sass Bálint, Simon Eszter, Mittelholcz Iván, Kundráth Péter és Vadász Noémi, „emtsv – Egy formátum mind felett," in *XV. Magyar Számítógépes Nyelvészeti Konferencia,* szerk., Berend Gábor, Gosztolya Gábor és Vincze Veronika (Szeged: Szegedi Tudományegyetem TTIK, Informatikai Intézet, 2019), 235–247.

[15] A program a hangrendnek a hagyományos, közoktatásban alkalmazott megközelítését használja azzal a módosítással, hogy az összetett szavak esetében a hangrend megállapítása nem az utolsó tag, hanem a teljes szó alapján történik, hiszen a hangrend annotálásának a célja esetleges poétikai következtetések elősegítése, nem pedig a toldalékok magánhangzóinak az előrejelzése (amire egyébként a közoktatásban alkalmazott hagyományos megközelítés nem is igazán alkalmas).

[16] Az elemzőprogram egy lehetséges továbbfejlesztésének iránya, hogy ne csak a versszakon belüli, hanem az interstrofikus, versszakon túlnyúló rímelést is elemezni tudja. Például a József Attila költészetében is megjelenő szonettformára jellemző ez a rímelési mód.

[17] Egy, az alkalmazottnál még általánosabb szabály lenne például, ha a program nem venné figyelembe a sor utolsó előtti szótagjának a hosszúságát, vagyis minden azonos magánhangzóra végződő sorvéget egymással rímelőnek tekintene.

és nem tiszta rímeket (azaz az asszonáncokat), adott esetben a nem tiszta rímeket is több csoportba sorolva.[18]

A rímképlet mellett a program megadja az egymással rímpárt alkotó, azaz hívó- és felelőrím viszonyban lévő szavak listáját, rímpáronként elkülönítve azokat. 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. Ezzel a fajta elemzéssel az vizsgálható, hogy egy adott típusú rímszó milyen típusú, a rímszerkezetben azt közvetlenül követő (vagy megelőző) rímszót hív elő a legvalószínűbben. Ennek például a rímelés pszicholingvisztikai irányultságú vizsgálataiban lehet jelentősége. A rímek automatikus elemzését a későbbiekben ugyanakkor érdemes lenne úgy továbbfejleszteni, hogy a bokorrímek és az egyéb, kettőnél több elemű rímek esetében ne csak az egymást követő tagok által alkotott rímpárokat lehessen vizsgálni, illetve hogy a kettőnél több tagú szerkezeteket egy egységként is le lehessen kérdezni. Ez utóbbi megoldás segíthetné a rímelés összetettebb szerkezeteinek a leírására irányuló vizsgálatokat. A rímek automatikus annotálása egy megfelelő méretű korpuszban – a tervek szerint az ELTE Verskorpusz ilyen lenne – lehetőséget adhat egy rímszótár elkészítésére is.[19]

A *hunpoem_analyzer-TEI* program úgyszintén elemzi a sorok ritmusát. Szemben az angol és egyéb indoeurópai nyelvű versek szóhangsúlyra épülő versritmusával, a magyar időmértékes versritmus a rövid és hosszú szótagok különbségére épül. A sorok hosszú és rövid szótagjainak az annotálása néhány egyszerű, a magyar verstanban jól ismert szabály alapján elvégezhető, így az előző részben bemutatott példáktól eltérően nem 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

[18] Az asszonáncoknak ilyen fokozati leírását adja: Arany János, „Valami az asszonáncról," in *Arany János összes művei X. kötet. Prózai művek 1.*, szerk. Keresztury Dezső (Budapest: Akadémiai Kiadó, 1962), 213–217. Arany János leírásának megkülönböztető jegyekre épülő, fonológiai keretben történő újraértelmezését adja: Szépe György, „Nyelvészeti jegyzetek Arany Jánosnak »Valami az asszonánczról« című tanulmányáról," *Magyar Nyelvőr* 93, 1. sz. (1969): 1–32. Funkcionális kognitív nyelvészeti keretben a magyar rím úgyszintén Arany leírásából kiinduló, prototípusalapú szerveződését vázolja fel: Simon Gábor, „A magyar rím fonológiai leírása funkcionális-kognitív megközelítésben," *Magyar Nyelvőr* 136, 2. sz. (2012): 65–82, valamint Simon Gábor, *Egy kognitív poétikai rímelmélet megalapozása* (Budapest: Tinta Könyvkiadó, 2014).

[19] Számítógépes nyelvészeti eljárásokat használó rímszótár elkészítésének módszeréről lásd Mártonfi Attila, „Egy magyar rímszótár terve," in *„Mielz valt mesure / que ne fait estultie": A hatvanéves Horváth Iván tiszteletére,* szerk., Bartók István et al. (Budapest: Krónika Nova Kiadó, 2008), 198–204.

 DOI 10.31400/dh-hun.2020.3.422

azokat a szótagokat, amelyekben hosszú magánhangzó áll, valamint azokat a rövid magánhangzós szótagokat, amelyekben hosszú vagy egynél több mássalhangzó követi a magánhangzót. 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. *„Nincsen apám, se anyám"* – 1001001). Ezt az elemzést a program automatikusan elvégzi minden versen, függetlenül attól, hogy jellemző-e rá valamilyen szabályos időmértékes versritmus vagy nem. Az így megadott ritmusból a program jelen állapotában nem kísérel meg valamilyen, a versre általánosan jellemző metrumot megállapítani, és verslábakat sem határoz meg. A távlati tervek között szerepel, hogy az előző részben bemutatott példákhoz hasonlóan a program az időmértékes ritmusból absztrahálni tudjon metrumot, illetve hogy az időmértékes verselés mellett az ütemhangsúlyos (és a szimultán) verselést is felismerje.[20]

A *hunpoem_analyzer-TEI* program utolsó funkciója az alliterációk automatikus felismerése. Az alliteráció tágabb értelmezése alapján nem csupán az azonos mássalhangzóval, hanem az azonos magánhangzóval kezdődő szavak is alliterációként annotálódnak. A program jelenlegi változata csak versszakon belül elemez alliterációkat. Ez tehát azt jelenti, hogy egy versszak utolsó és a következő versszak első, azonos hanggal kezdődő szava nem annotálódik alliterációként. 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. Így például Babitsnak a *„Bus donna barna balkonon"* sorát egy egységként elemezné mint alliterációt. Minden alliterációként elemzett szerkezet kap egy *a* és *n* betűből álló karaktersort, 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). A programba egyelőre nincs beépítve *stopword*-lista, aminek következtében az *az asztal* típusú szerkezetek is alliterációként annotálódnak.

A *hunpoem_analyzer-TEI* programnak nemcsak a bemenete, hanem az elemzett tulajdonságokat tartalmazó kimenete is megfelel a TEI XML-szabványnak.

## 4. A József Attila-korpusz létrehozásának és automatikus elemzésének főbb lépései

A József Attila verseit tartalmazó korpuszt a Magyar Elektronikus Könyvtár (MEK)[21] oldaláról letölthető *József Attila összes költeménye* című, MEK-00708 azonosítószámú digitális dokumentumból hoztam létre.[22] A töredékeket, az idegen nyelvű verseket,

[20] Az időmértékes metrum felismerése első megközelítésben a jambikus, trochaikus, anapesztikus és daktilikus általános kategóriákba történő automatikus besorolást jelenthetné. Az ütemhangsúlyos verselés automatikus felismerésének előfeltétele a szóhangsúlyok automatikus annotálása. A versritmusról például lásd: Szepes Erika és Szerdahelyi István, *Verstan* (Budapest: Gondolat Kiadó, 1981).

[21] Magyar Elektronikus Könyvtár, hozzáférés: 2019.01.19, `https://mek.oszk.hu/`.

[22] Az ELTE Verskorpusz építése során alapvetően a MEK adatbázisában található versgyűjteményeket használjuk, amelyek többnyire egységes formátumban vannak, így nem kell minden egyes szerző esetében külön szkriptet írni a versek szerkezeti egységeinek az annotációit tartalmazó TEI XML-fájlok létrehozására, amelyek a további annotálási lépések bemenetét adják. Bár a `http://jozsefattila.elte.hu/` oldalon jobb minőségben szerepelnek a szövegek, a MEK-szövegkiadás haszná-

valamint a *Változatok* és a *Rögtönzések* címszó alatt szereplő szövegeket mellőzve 532 vers került a korpuszba. Egy XQuery szkripttel minden versből létrehoztam egy, a versek szerkezeti egységeinek (cím, versszakok, sorok) az *inline* annotációit tartalmazó TEI XML-fájlt. Ezt követően a versek versszakbeosztását ellenőriztem a kritikai kiadás alapján, és ahol eltérés volt, ott javítottam.[23] Ez a rímképlet és a rímpárok pontosabb megállapítása miatt volt fontos. Egyéb szempontból nem javítottam a szövegeket. Következő lépésben a TEI XML-fájlokon lefuttattam a *hunpoem_analyzer-TEI* programot, illetve az ebbe beépített *e-magyar* program *emtsv* változatát. Ennek eredményeként a TEI XML-fájlokba bekerültek a szavak grammatikai tulajdonságai, azaz a szavak lemmája, szófaja és morfoszintaktikai jellemzői, valamint az előző pontban részletezett, vershangzáshoz kapcsolódó annotációk, vagyis a szavak főbb fonológiai jellemzői, a versszakok rímképlete, a rímpárok, a sorok ritmusa és az alliterációk elemzései. Ezt követően a TEI XML-fájlok formátumán egy XSLT stíluslappal kisebb módosításokat hajtottam végre, mivel a TEI XML-formátum alapvetően szövegek szabványos tárolására és megosztására lett kitalálva, nem pedig arra, hogy bonyolultabb lekérdezéseket is egyszerűbben meg lehessen írni, illetve gyorsan le lehessen futtatni. A kimenetül kapott XML-fájlokat az *eXist-db* nevű XML-adatbázis-kezelő programba töltöttem be, ahol XQuery nyelven meg lehet írni a lekérdezéseket. Hangsúlyozandó tehát, hogy az alábbiakban szereplő gyakorisági listákat nem az *e-magyar,* és nem is a *hunpoem_analyzer-TEI* program hozza létre, hiszen ezek funkciója pusztán egy szó vagy egy több szóból álló szerkezeti egység (sor, versszak stb.) elemzése, illetve az elemzésnek az XML-fájlba való beírása. A különböző gyakorisági listák létrehozása az XQuery-lekérdezésekbe lett beépítve.

A József Attila-korpusz alapját adó MEK-dokumentum a címe ellenére természetesen nem tartalmazza az összes József Attila-verset, és a szövegek is több esetben eltérnek a kritikai kiadás alapszövegétől. Ennek a tanulmánynak az elsődleges célja azonban nem az, hogy minél pontosabb kvantitatív adatokat szolgáltasson József Attila költészetéről, hanem az, hogy József Attila verseinek a példáján keresztül bemutassa a vershangzáshoz kapcsolódó jellemzők automatikus feltárásának egy módszerét.[24] Mindazonáltal feltételezhető (vagy legalábbis remélhető), hogy a korpuszban nem szereplő, illetve rontott szöveggel szereplő versek nem befolyásolják lényegesen az alábbiakban bemutatandó gyakorisági listák elemeinek a sorrendjét.

## 5. A József Attila-versek szókészletének néhány jellemzője

A versek automatikusan feltárt hangzásjellemzőinek az ismertetése előtt érdemes a korpusz főbb lexikai tulajdonságaiból is bemutatni párat. A lexikai tulajdonságok

latával részben az volt a célom, hogy teszteljem a készülő ELTE Verskorpusz számára kialakított, a MEK adatbázisában található szövegekre épülő annotációs folyamatot.

[23] József Attila, *József Attila összes versei*, kiad. Stoll Béla (Budapest : Balassi Kiadó, 2005).

[24] A József Attila költészetéről pontosabb adatokat szolgáltató, a különböző szövegváltozatokat is figyelembe vevő írói szótár tervezetéről lásd : Mártonfi Attila, „Számítógép és írói szótár – különös tekintettel a készülő József Attila szótárra," *Magyar Nyelv* 110, 1. sz. (2014) : 30–46.

 DOI 10.31400/dh-hun.2020.3.422

automatikus kinyerését az *e-magyar*nak a verseken való lefuttatása tette lehetővé. Az alábbi *1. táblázat* a korpusz legfontosabb mennyiségi jellemzőit mutatja be.[25]

| | |
|---|---|
| versek száma | 532 |
| szavak száma | 58144 |
| lemmák száma | 10038 |

**1. táblázat**. A korpusz legfontosabb mennyiségi jellemzői

Egy szó lemmája a szó szótári alakja. Például a *futott, futnánk, fut, fuss* szavak lemmája a *fut* szó. A lemmák száma tehát a korpusz szókincsének a méretét mutatja meg. A *2. táblázat* a korpuszban szereplő húsz leggyakoribb lemmát mutatja be. A lemmák utáni oszlopban adtam meg az előfordulási számot, a később szereplő gyakorisági listáknál is ezt a módszert követem. A táblázatban egy adott lemmához különböző szófajú, adott esetben homonim szóalakok is tartozhatnak, például a *ki* lemma a névmási és az igekötői szóalakokat is magában foglalja.

| | | |
|---|---|---|
| 1 | a | 4174 |
| 2 | az | 1432 |
| 3 | s | 1260 |
| 4 | és | 877 |
| 5 | nem | 844 |
| 6 | én | 768 |
| 7 | van | 742 |
| 8 | is | 585 |
| 9 | hogy | 553 |
| 10 | mint | 525 |
| 11 | ha | 483 |
| 12 | csak | 384 |
| 13 | de | 375 |
| 14 | ki | 325 |
| 15 | egy | 323 |
| 16 | már | 319 |
| 17 | meg | 303 |
| 18 | te | 298 |
| 19 | ő | 291 |
| 20 | mi | 289 |

**2. táblázat**. A leggyakoribb lemmák

Látható, hogy a húsz leggyakoribb lemma között többnyire névelők, kötőszavak, névmások szerepelnek, azaz olyan fogalmilag kevésbé tartalmas szavak, amelyek nem túl sokat árulnak el a szöveg jellegéről. Érdemes tehát fogalmilag tartalmasabb jelentésű

[25] A verscímeken nem lett lefuttatva az *e-magyar* és a vershangzást elemző program, így ezek nem alkotják részét a tanulmányban bemutatott adatoknak.

szófajokhoz kapcsolódó szólistákat lekérdezni. A *3. táblázat* a korpuszban szereplő tíz leggyakoribb igei, főnévi és melléknévi lemmát mutatja be.

**Ige**

| | | |
|---|---|---|
| 1 | van | 742 |
| 2 | lesz | 273 |
| 3 | tud | 192 |
| 4 | szeret | 172 |
| 5 | lát | 143 |
| 6 | nincs | 130 |
| 7 | jön | 130 |
| 8 | kell | 127 |
| 9 | vár | 113 |
| 10 | él | 112 |

**Főnév**

| | | |
|---|---|---|
| 1 | szem | 179 |
| 2 | szív | 172 |
| 3 | ég | 170 |
| 4 | föld | 162 |
| 5 | ember | 160 |
| 6 | lélek | 147 |
| 7 | világ | 129 |
| 8 | isten | 111 |
| 9 | szó | 102 |
| 10 | kéz | 98 |

**Melléknév**

| | | |
|---|---|---|
| 1 | szép | 211 |
| 2 | nagy | 208 |
| 3 | jó | 155 |
| 4 | kis | 97 |
| 5 | bús | 74 |
| 6 | erős | 65 |
| 7 | tiszta | 58 |
| 8 | friss | 52 |
| 9 | lágy | 49 |
| 10 | lassú | 48 |

**3. táblázat**. A leggyakoribb igei, főnévi és melléknévi lemmák.

Bár ennek a tanulmánynak nem célja a különböző lekérdezett listák elemzése, talán érdemes megjegyezni azt a számomra meglepő eredményt, hogy a tíz leggyakoribb melléknév többsége pozitív konnotációjú. Ez persze összefügghet azzal a nyelvi univerzáléval, hogy a minőségre vonatkozó kategóriák megnevezésére jellemzően pozitív konnotációjú mellékneveket használunk.

 DOI 10.31400/dh-hun.2020.3.422

Természetesen specifikusabb nyelvtani kategóriák szólistáit is le lehet kérdezni, az alábbi, *4. táblázat*ban például a tíz leggyakoribb E/1, E/2 és E/3 alakban előforduló igei lemma szerepel. Félkövérrel emeltem ki azokat az igéket, amelyek a háromból csak az egyik kategóriában szerepelnek a leggyakoribb tíz igei lemma között.

| | E/1 | | E/2 | | E/3 | |
|---|---|---|---|---|---|---|
| 1 | van | 186 | van | 48 | van | 401 |
| 2 | tud | 96 | lesz | 44 | lesz | 134 |
| 3 | szeret | 70 | jön | 39 | **nincs** | 128 |
| 4 | lesz | 52 | **néz** | 28 | **kell** | 122 |
| 5 | lát | 50 | szeret | 27 | **száll** | 66 |
| 6 | vár | 29 | **mond** | 25 | jön | 56 |
| 7 | **hisz** | 28 | tud | 23 | **hull** | 47 |
| 8 | **érez** | 26 | **ad** | 22 | vár | 45 |
| 9 | **megy** | 24 | lát | 22 | **volna** | 45 |
| 10 | **él** | 24 | **alszik** | 22 | tud | 44 |

**4. táblázat**. Az E/1, E/2 és E/3 alakban leggyakrabban előforduló igei lemmák.

Nemcsak szólisták, hanem a szavaknál absztraktabb, grammatikai kategóriák is lekérdezhetők. Az *5. táblázat* a korpusz leggyakoribb tíz szófaját mutatja be.[26]

| | | |
|---|---|---|
| 1 | főnév | 15774 |
| 2 | ige | 11613 |
| 3 | melléknév | 6984 |
| 4 | határozószó | 6184 |
| 5 | determináns | 5588 |
| 6 | névmás | 4896 |
| 7 | mellérendelő kötőszó | 3131 |
| 8 | alárendelő kötőszó | 1716 |
| 9 | névutó | 685 |
| 10 | tulajdonnév | 523 |

**5. táblázat**. A leggyakoribb szófajok.

## 6. A József Attila-versek automatikusan feltárt hangzásjellemzői

Az előző részben bemutatott, szavakat és grammatikai kategóriákat tartalmazó gyakorisági listák minden bizonnyal nem jelentenek nagy újdonságot azok számára,

[26] Az *e-magyar*nak az UD (Universal Dependencies) típusú szófaji és morfoszintaktikai címkéket kiadó kimenetét használtam. Az UD-szabványt követő *e-magyar* nem minden esetben úgy osztja ki a szófaji címkéket, ahogyan azt a magyar leíró nyelvtanok teszik. Az e-magyar UD kimenetében szereplő eredeti szófaji címkék és azoknak az általam használt fordításai a következők: NOUN – főnév, VERB – ige, ADJ – melléknév, ADV – határozószó, DET – determináns, PRON – névmás, CONJ – mellérendelő kötőszó, SCONJ – alárendelő kötőszó, ADP – névutó, PROPN – tulajdonnév. NUM – számnév. Az UD szabvány magyar nyelvre való alkalmazását lásd az alábbi címen, hozzáférés: 2020.01.19, `https://universaldependencies.org/treebanks/hu_szeged/index.html`.

akik foglalkoznak korpusznyelvészettel, hiszen ilyen gyakorisági listákat a nyelvileg elemzett korpuszok lekérdezőfelületei általában tudnak generálni (erre jó példa a már említett Magyar Nemzeti Szövegtár lekérdezőfelülete[27]). Egy szövegcsoportot kvantitatív módon azonban nem csak a szókészletre vonatkozó tulajdonságok mentén jellemezhetünk. Egy lírai szövegeket tartalmazó korpuszt például az alapján is leírhatunk, hogy melyek a legtipikusabb, leggyakoribb rímképletek, rímpárok, ritmusok, vagy éppen alliterációtípusok. A továbbiakban olyan gyakorisági listákat mutatok be, amelyek lekérdezését a vershangzáshoz kapcsolódó tulajdonságok előzetes, automatikus annotálása tette lehetővé.

Egy verskorpusz kapcsán feltehetjük azt az egyszerű kérdést, hogy melyek a leggyakoribb versszakszámok. A *6. táblázat* a tíz leggyakoribb versszakszámot mutatja be.

| **Versszakszám** | | |
|---|---|---|
| 1 | 4 | 152 |
| 2 | 3 | 82 |
| 3 | 1 | 66 |
| 4 | 5 | 54 |
| 5 | 6 | 46 |
| 6 | 2 | 45 |
| 7 | 7 | 21 |
| 8 | 8 | 17 |
| 9 | 9 | 14 |
| 10 | 10 | 8 |

**6. táblázat**. A leggyakoribb versszakszámok.

A táblázatból láthatjuk, hogy a négy versszakból álló versek a leggyakoribbak, ezekből 152 darab szerepel a korpuszban. A lista második helyén pedig a három versszakos versek állnak.

Egy verskorpusz egyik fontos jellemzője az is, hogy a versszakok esetében melyek a legtipikusabb sorszámok és rímképletek. A *7. táblázat* a versszakokra jellemző leggyakoribb tíz sorszámot és rímképletet mutatja be.

| **Sorszám** | | |
|---|---|---|
| 1 | 4 | 1301 |
| 2 | 3 | 348 |
| 3 | 2 | 251 |
| 4 | 5 | 142 |
| 5 | 8 | 139 |
| 6 | 6 | 88 |
| 7 | 1 | 69 |
| 8 | 7 | 64 |
| 9 | 10 | 32 |

[27] *Magyar Nemzeti Szövegtár,* hozzáférés: 2020.01.19, `http://clara.nytud.hu/mnsz2-dev/`.

 DOI 10.31400/dh-hun.2020.3.422

| | | |
|---|---|---|
| 10 | 11 | 16 |
| 11 | 9 | 16 |
| 12 | 12 | 16 |

**Rímképlet**

| | | |
|---|---|---|
| 1 | abab | 445 |
| 2 | aabb | 242 |
| 3 | abcb | 158 |
| 4 | abba | 143 |
| 5 | aab | 142 |
| 6 | aa | 138 |
| 7 | ab | 113 |
| 8 | aba | 111 |
| 9 | aaaa | 82 |
| 10 | # | 69 |

7. **táblázat**. A versszakok leggyakoribb sorszámai és rímképletei.

A táblázatból látható, hogy a négysoros versszakok a legtipikusabbak, ezekből 1301 darab szerepel a korpuszban, a második leggyakoribb típus pedig a háromsoros versszak. A leggyakoribb rímképlet az abab típus, azaz a négysoros keresztrímek, amelyből 445 szerepel a korpuszban. A lista második helyén pedig az aabb típus áll, vagyis a négysoros párosrímek. A rímképletek oszlopban a 10. helyen szereplő # jel az egysoros versszakokra utal.

Nemcsak versszakok, hanem verssorok kapcsán is lekérdezhetünk különböző gyakorisági listákat. A *8. táblázat* a korpusz soraira jellemző tíz leggyakoribb szószámot, szótagszámot és ritmusképletet mutatja be. A ritmusképletekben a 0 a rövid szótag, az 1 pedig a hosszú szótag jele.[28]

**Szószám**

| | | |
|---|---|---|
| 1 | 5 | 3072 |
| 2 | 4 | 2965 |
| 3 | 6 | 1891 |
| 4 | 3 | 1672 |
| 5 | 7 | 914 |
| 6 | 2 | 574 |
| 7 | 8 | 365 |
| 8 | 9 | 164 |
| 9 | 1 | 136 |
| 10 | 10 | 72 |

[28] A sorok utolsó szótagjának az elemzésében a program nem veszi figyelembe a következő sor elejét.

**Szótagszám**

| | | |
|---|---|---|
| 1 | 10 | 2509 |
| 2 | 8 | 2296 |
| 3 | 9 | 2229 |
| 4 | 11 | 1852 |
| 5 | 7 | 708 |
| 6 | 6 | 509 |
| 7 | 12 | 383 |
| 8 | 4 | 243 |
| 9 | 13 | 205 |
| 10 | 5 | 166 |

**Ritmus**

| | | |
|---|---|---|
| 1 | 110101010 | 57 |
| 2 | 011101010 | 55 |
| 3 | 010111010 | 48 |
| 4 | 010101010 | 46 |
| 5 | 11010101010 | 46 |
| 6 | 11011100 | 46 |
| 7 | 111101010 | 44 |
| 8 | 01111100 | 43 |
| 9 | 110111010 | 42 |
| 10 | 01011101 | 41 |
| 11 | 01111101 | 41 |

**8. táblázat**. A sorok leggyakoribb szószámai, szótagszámai és ritmusképletei.

A vizsgált korpuszban a leggyakoribb tehát az öt szóból álló sor, szótagszám tekintetében pedig a tíz szótagos sor. A leggyakoribb ritmusképlet viszont kilenc szótagból áll, hagyományos jelöléssel: – – U – U – U – U. Megjegyzendő, hogy az első négy ritmusképlet valójában mind ötödfeles jambus. A különbség pusztán annyi, hogy a jambusokat helyettesítő spondeusok máshol jelennek meg. A többi ritmusképlet is mind jambikus lejtésű. Ahogyan arra már utaltam, az annotáló program továbbfejlesztésének a távlati céljai között szerepel a ritmusból metrumra történő absztrahálás funkciójának a beépítése. Más szerzők verseivel való összehasonlításban érdekes lehet a hosszú és rövid szótagok korpuszbeli előfordulásának a száma is, ezt az alábbi, 9. táblázat mutatja be.

| | |
|---|---|
| hosszú szótagok száma | 62731 |
| rövid szótagok száma | 51908 |

**9. táblázat**. A hosszú és rövid szótagok száma.

Ahogyan arról a tanulmány elején szó volt, a versek automatikus elemzéséhez fejlesztett *hunpoem_analyzer-TEI* program a szavak szótagszámát, hangrendjét és fonológiai szerkezetét is elemzi. A *10. táblázat* a magas, mély és vegyes hangrendű

 DOI 10.31400/dh-hun.2020.3.422

szavak gyakorisági sorrendjét, valamint a szavak szótagszámának a gyakorisági sorrendjét mutatja be. A szótagszám oszlopban a 0 érték ötödik helyen való szereplése a *s* kötőszó gyakori használatával magyarázható.

**Hangrend**

| | | |
|---|---|---|
| 1 | magas | 25452 |
| 2 | mély | 23908 |
| 3 | vegyes | 7461 |

**Szótagszám**

| | | |
|---|---|---|
| 1 | 1 | 21176 |
| 2 | 2 | 19608 |
| 3 | 3 | 10985 |
| 4 | 4 | 4134 |
| 5 | 0 | 1272 |
| 6 | 5 | 773 |
| 7 | 6 | 125 |
| 8 | 7 | 19 |
| 9 | 8 | 1 |

**10. táblázat**. A szavak hangrendjének és szótagszámainak a gyakorisági listája.

A *11. táblázat* a tíz leggyakoribb fonológiai szerkezetet mutatja be. A fonológiai szerkezetek reprezentációiban a C a mássalhangzókat, a V pedig a magánhangzókat jelöli, a V után álló F az elöl képzett, a B pedig a hátul képzett magánhangzókra utal. Az ezt követő 1 szám arra utal, hogy a magánhangzó rövid, a 2 pedig, hogy a magánhangzó hosszú. A VB1 szerkezetnek az első helyen szereplése tehát az *a* névelőnek köszönhető.

**Fonológiai szerkezet**

| | | |
|---|---|---|
| 1 | VB1 | 4174 |
| 2 | C, VB1, C | 2000 |
| 3 | C, VF1, C | 1911 |
| 4 | VF2, C | 1668 |
| 5 | C, VF1, C, C | 1634 |
| 6 | VB1, C | 1393 |
| 7 | C, VF1 | 1362 |
| 8 | C, VF1, C, VF1, C | 1348 |
| 9 | C, VB1, C, VB1, C | 1300 |
| 10 | VF1, C | 1290 |

**11. táblázat**. A szavak leggyakoribb fonológiai szerkezetei.

A korpuszban ugyancsak annotálva lettek az alliterációk, így ezekre vonatkozó gyakorisági listákat is le lehetett kérdezni. A *12. táblázat* a tíz leggyakoribb alliterációként elemzett szerkezetet mutatja be. Az *a* betű jelöli a szerkezetben egymással alliteráló szavakat, az *n* pedig az alliteráló szavak közé esetlegesen beékelődő nem

alliteráló szavakat. A *hunpoem_analyzer-TEI* program csak azokat a szerkezeteket elemzi alliterációként, amelyekben két alliteráló szó közé maximum egy nem alliteráló szó ékelődik be. A programba egyelőre nincs beépítve stopword-lista, így az *az asztal* típusú szerkezetek is részei a listának.

| | **Szerkezet** | |
|---|---|---|
| 1 | ana | 2237 |
| 2 | aa | 2155 |
| 3 | anana | 157 |
| 4 | aaa | 147 |
| 5 | aana | 133 |
| 6 | anaa | 120 |
| 7 | ananana | 16 |
| 8 | anaaa | 10 |
| 9 | ananaa | 10 |
| 10 | aaaa | 10 |

**12. táblázat**. Az alliterációk leggyakoribb típusai.

A *13. táblázat* az alliterációk leggyakoribb típusait mutatja be szófajok alapján. Az első oszlopban a beékelődő, nem alliteráló szavakat nem tartalmazó szerkezetek leggyakoribb tíz típusa szerepel az alliterációt alkotó szavak szófaja alapján. Ebből a listából – kompenzálva a *stopword*ök hiányát – kihagytam azokat a kételemű alliterációtípusokat, amelyek determinánsokat tartalmaznak. A másik oszlop a leggyakoribb tíz, beékelődő szót tartalmazó szerkezetet mutatja be a szavak szófaja alapján. Az alliteráló szavak közé beékelődő szó szófaját félkövérrel emeltem ki. Ebből a listából is kihagytam azokat a – beékelődő szóval együtt – háromelemű alliterációtípusokat, amelyekben a determináns nem beékelődő, hanem alliteráló szóként jelenik meg.

| | **Szófaj – nincs benne nem alliteráló szó** | |
|---|---|---|
| 1 | melléknév, főnév | 193 |
| 2 | főnév, főnév | 170 |
| 3 | főnév, ige | 166 |
| 4 | ige, ige | 97 |
| 5 | ige, névmás | 70 |
| 6 | ige, főnév | 64 |
| 7 | határozószó, ige | 63 |
| 8 | ige, határozószó | 58 |
| 9 | főnév, melléknév | 56 |
| 10 | melléknév, melléknév | 55 |

| | **Szófaj – beékelődő nem alliteráló szóval** | |
|---|---|---|
| 1 | főnév, **determináns**, főnév | 76 |
| 2 | ige, **determináns**, főnév | 74 |
| 3 | melléknév, **főnév**, ige | 49 |

 DOI 10.31400/dh-hun.2020.3.422

| | | |
|---|---|---|
| 4 | főnév, **melléknév**, főnév | 43 |
| 5 | főnév, **főnév**, ige | 34 |
| 6 | határozószó, **ige**, határozószó | 34 |
| 7 | ige, **határozószó**, ige | 32 |
| 8 | főnév, **főnév**, főnév | 30 |
| 9 | ige, **melléknév**, főnév | 29 |
| 10 | melléknév, **főnév**, főnév | 29 |

**13. táblázat**. Az alliterációk leggyakoribb típusai szófajok alapján.

A táblázatból látható, hogy a beékelődő szavakat nem tartalmazó leggyakoribb tíz szerkezet között nincsen olyan szófaj-kombináció, amely kettőnél több elemű lenne, és a beékelődő szavakat tartalmazó leggyakoribb tíz szerkezet között sincs olyan, amelyben kettőnél több alliteráló szó szerepelne.

A következő típusa az itt bemutatandó, vershangzásra vonatkozó kvantitatív jellemzőknek a rímhelyzetben lévő szavakhoz kapcsolódó gyakorisági listák. A rímnek nem csupán fonológiai, hanem szemantikai funkciója is van. A rímhelyzet a versvilág felépítése szempontjából jelentéstanilag kitüntetett pozíció, így egy verskorpusz esetében érdekes lehet, hogy milyen típusú szavak kerülnek leggyakrabban rímhelyzetbe.[29] A *14. táblázat* a József Attila-korpuszban szereplő tíz leggyakoribb rímhelyzetben lévő tokent (szóalak), lemmát (szótári alak) és szófajt mutatja be. Ezen listák esetében tehát nincsen annak jelentősége, hogy a szó hívó- vagy felelőrím pozíciót tölt be, vagy esetleg mindkettőt.

**Token**

| | | |
|---|---|---|
| 1 | is | 53 |
| 2 | el | 44 |
| 3 | meg | 39 |
| 4 | már | 35 |
| 5 | van | 34 |
| 6 | vagyok | 32 |
| 7 | én | 30 |
| 8 | alatt | 25 |
| 9 | engem | 25 |
| 10 | ég | 22 |

**Lemma**

| | | |
|---|---|---|
| 1 | van | 168 |
| 2 | én | 133 |
| 3 | ég | 74 |
| 4 | maga | 61 |

[29] A rím jelentéstani szerepét mutatja be Simon, *Egy kognitív poétikai rímelmélet megalapozása.* A monográfiában szerepel egy József Attila-verseket is tartalmazó verskorpusz rímhelyzetben lévő szavainak a manuálisan létrehozott szófaji annotációira épülő vizsgálat.

| | | |
|---|---|---|
| 5 | lesz | 57 |
| 6 | világ | 56 |
| 7 | is | 53 |
| 8 | te | 50 |
| 9 | szeret | 50 |
| 10 | ő | 46 |

**Szófaj**

| | | |
|---|---|---|
| 1 | főnév | 5002 |
| 2 | ige | 3018 |
| 3 | melléknév | 810 |
| 4 | határozószó | 728 |
| 5 | névmás | 510 |
| 6 | névutó | 223 |
| 7 | tulajdonnév | 77 |
| 8 | számnév | 44 |
| 9 | mellérendelő kötőszó | 37 |
| 10 | PART[30] | 37 |

**14. táblázat**. A leggyakoribb rímhelyzetben lévő tokenek, lemmák és szófajok.

Érdemes megjegyezni, hogy míg a korpusz leggyakoribb húsz lemmája között egyáltalán nem szerepelt főnév, és igeként is csak a *van* lemma szerepelt, addig a leggyakoribb tíz rímhelyzetben lévő lemma között a *van* mellett szerepel a *lesz* és a *szeret* ige, valamint a *világ* főnév és az *ég* lemma, amely főnév és ige is lehet. Ha a szófajok gyakorisági listáját nézzük, akkor megállapíthatjuk, hogy a rímhelyzetben lévő első négy leggyakoribb szófaj (főnév, ige, melléknév, határozószó) ugyanaz, mint a korpusz összes szava esetében. Ezt követően változik a sorrend, például nincs a leggyakoribb tíz rímhelyzetben lévő szófaj között determináns és alárendelő kötőszó, a mellérendelő kötőszó pedig hátrébb került. Ezek a változások persze nem meglepőek, hiszen a versek sorai a legtöbb esetben tagmondatok, amelyek végén nincsen névelő vagy kötőszó.

A rímhelyzetben lévő szavak esetében is lekérdezhetjük azok fonológiai tulajdonságait. A *15. táblázat* a rímhelyzetben lévő szavak hangrendjének és szótagszámának a gyakorisági listáját mutatja be.

**Hangrend**

| | | |
|---|---|---|
| 1 | magas | 4804 |
| 2 | mély | 3939 |
| 3 | vegyes | 1749 |

[30] Az *e-magyar* UD típusú kimenetében a PART (particle) címkét kizárólag a *meg* igekötő kapja. Ez a címke tehát nem egyezik meg a magyar leíró nyelvészetben használt partikula szófajának a kategóriájával. A többi igekötő az ADV (határozószó) címkét kapja.

 DOI 10.31400/dh-hun.2020.3.422

| | Szótagszám | |
|---|---|---|
| 1 | 2 | 4439 |
| 2 | 3 | 3076 |
| 3 | 4 | 1497 |
| 4 | 1 | 1140 |
| 5 | 5 | 292 |
| 6 | 6 | 44 |
| 7 | 7 | 4 |

**15. táblázat**. A rímhelyzetben lévő szavak hangrendjének és szótagszámának gyakorisági listája.

A *15. táblázat*ból kiderül, hogy a mély hangrendű rímhelyzetben lévő szavak aránya csökken, míg a vegyes hangrendű szavak aránya nő az összes szó mély és vegyes hangrendű szavának az arányához képest. Az összes szónak 42,1%-a mély hangrendű és 13,1%-a vegyes hangrendű. Ezzel szemben a rímhelyzetben lévő szavaknak 37,5%-a mély hangrendű és 16,7%-a vegyes hangrendű. A jövőben érdemes lenne megvizsgálni, hogy a hangrendi arányoknak ez a változása a magyar rímelésre jellemző általános, azaz más szerzőknél is megjelenő mintázatnak tekinthető-e. A rímhelyzetben lévő szavak szótagszámának a gyakorisági listája egy dologban különbözik a korpusz összes szavára vonatkozó gyakorisági listától: az egy szótagos szavak hátrébb kerültek a listában. Ez minden bizonnyal azzal a szófajok kapcsán megállapított jellemzővel magyarázható, hogy az egy szótagú névelők nem állnak a verssorok végén, és a gyakran egy szótagú kötőszavak sem jellemzőek ebben a pozícióban.

A *16. táblázat* a rímhelyzetben lévő szavak leggyakoribb tíz fonológiai szerkezetét mutatja be.

| | Fonológiai szerkezet | |
|---|---|---|
| 1 | C, VB1, C, VB1, C | 394 |
| 2 | C, VF1, C, VF1, C | 350 |
| 3 | C, VF1, C, C, VF1, C | 324 |
| 4 | C, VB1, C, C, VB1, C | 189 |
| 5 | C, VB1, C, C, VB1 | 160 |
| 6 | C, VB2, C | 158 |
| 7 | C, VF1, C, VF1, C, VF1, C | 153 |
| 8 | C, VF1, C, VF1, C, C, VF1, C | 152 |
| 9 | C, VF1, C, VF2, C | 145 |
| 10 | C, VB1, C, VB2, C | 142 |

**16. táblázat**. A rímhelyzetben lévő szavak leggyakoribb fonológiai szerkezetei.

A szótagszámnak a fentebb említett változása magyarázza a leggyakoribb tíz fonológiai szerkezet listájában tapasztalható jelentősebb eltéréseket a korpusz összes szavára vonatkozó listához képest. A rímhelyzetben lévő szavak esetében az első öt leggyakoribb fonológiai szerkezet két magánhangzót tartalmaz, és a leggyakoribb tíz szerkezet között is csak egy darab egy magánhangzót tartalmazó szerkezet szerepel,

ugyanakkor két darab három magánhangzót tartalmazó szerkezet is megjelenik. Ezzel szemben az összes szóra vonatkozó listában csupán két darab két magánhangzót tartalmazó szerkezet szerepel a lista nyolcadik és kilencedik helyén, az összes többi szerkezet egy szótagú.

Végezetül érdekes lehet az is, hogy melyek a leggyakoribb rímpárok, illetve rímpártípusok. A *17. táblázat* első oszlopa a korpuszban legalább háromszor szereplő rímpárokat mutatja be gyakoriságuk sorrendjében, a második oszlop pedig a rímpárok lemmatizált alakjaiból tünteti fel azokat, amelyekből legalább három szerepel a korpuszban. Hangsúlyozandó, hogy a táblázatban szereplő gyakorisági listák figyelembe veszik a rímpárt alkotó szavak sorrendjét. Például a *tenger-ember* rímpárba nincsenek beleszámolva az *ember-tenger* sorrendű rímpárok, ezek a lista külön tételét alkotják.

**Token**

| | | |
|---|---|---|
| 1 | tenger – ember | 4 |
| 2 | szemegödre – mindörökre | 4 |
| 3 | vagyok – csillagok | 3 |
| 4 | az – igaz | 3 |
| 5 | szellem – ellen | 3 |
| 6 | hamis – is | 3 |
| 7 | pajtás – hajtás | 3 |
| 8 | magam – van | 3 |
| 9 | lelkem – telken | 3 |
| 10 | össze – fürössze | 3 |
| 11 | szerelem – velem | 3 |
| 12 | kell – el | 3 |
| 13 | apámat – Amerikának | 3 |
| 14 | agyunk – vagyunk | 3 |

**Lemma**

| | | |
|---|---|---|
| 1 | maga – van | 5 |
| 2 | lélek – telek | 5 |
| 3 | tenger – ember | 5 |
| 4 | szemegöd – mindörökre | 4 |
| 5 | van – csillag | 3 |
| 6 | hisz – visz | 3 |
| 7 | között – köd | 3 |
| 8 | az – igaz | 3 |
| 9 | tör – meggyötör | 3 |
| 10 | szellem – ellen | 3 |
| 11 | elme – szerelem | 3 |
| 12 | maga – szó | 3 |
| 13 | vágy – ágy | 3 |
| 14 | hamis – is | 3 |
| 15 | pajtás – hajtás | 3 |
| 16 | össze – fürészt | 3 |

 DOI 10.31400/dh-hun.2020.3.422

| | | |
|---|---|---|
| 17 | szerelem – én | 3 |
| 18 | kell – el | 3 |
| 19 | test – este | 3 |
| 20 | apa – Amerika | 3 |
| 21 | van – maga | 3 |
| 22 | agy – van | 3 |

**17. táblázat**. Rímpárok gyakorisági listája tokenek és lemmák alapján – a rímpárok tagjainak a sorrendje számít.

A táblázat kapcsán megjegyzendő, hogy sok esetben bizonyos rímpároknak a többszöri előfordulása kizárólag egy verssel magyarázható. Például a *szemegödre–mindörökre* rímpárnak mind a négy előfordulása a *Bús magyar éneke* című versben található, az *apámat-Amerikának* rímpárnak pedig mind a három előfordulása a „Csak most…" kezdetű versben jelenik meg.

A *18. táblázatban* a leggyakoribb tíz, rímpárt alkotó szófaj-kombináció, valamint a leggyakoribb tizenöt, rímpárt alkotó szótagszám–kombináció szerepel. Ezekben az esetekben is számít a rímpárok tagjainak a sorrendje.

**Szófaj**

| | | |
|---|---|---|
| 1 | főnév – főnév | 1312 |
| 2 | főnév – ige | 624 |
| 3 | ige – ige | 608 |
| 4 | ige – főnév | 601 |
| 5 | melléknév – főnév | 203 |
| 6 | határozószó – főnév | 175 |
| 7 | főnév – melléknév | 169 |
| 8 | főnév – határozószó | 156 |
| 9 | névmás – főnév | 123 |
| 10 | főnév – névmás | 113 |

**Szótagszám**

| | | |
|---|---|---|
| 1 | 2 – 2 | 1088 |
| 2 | 2 – 3 | 624 |
| 3 | 3 – 3 | 590 |
| 4 | 3 – 2 | 520 |
| 5 | 2 – 4 | 335 |
| 6 | 4 – 2 | 266 |
| 7 | 3 – 4 | 211 |
| 8 | 1 – 2 | 201 |
| 9 | 4 – 4 | 192 |
| 10 | 1 – 1 | 177 |
| 11 | 2 – 1 | 176 |
| 12 | 1 – 3 | 173 |

| | | |
|---|---|---|
| 13 | 4 – 3 | 153 |
| 14 | 3 – 1 | 119 |
| 15 | 2 – 5 | 77 |

**18. táblázat**. Rímpárok gyakorisági listája szófajok és szótagszámok alapján – a rímpárok tagjainak a sorrendje számít.

A *18. táblázatban* szereplő szótagszámok kapcsán egy érdekességre érdemes felhívni a figyelmet. Egy különböző szótagszámokat tartalmazó kombináció esetében mindig több előfordulással jelenik meg az a változat, amelyben az első tag, vagyis a hívórím a kevesebb szótagszámú. Vagyis a gyakorisági listában a 2-3 kombináció előrébb szerepel, mint a 3-2 kombináció, a 2-4 kombináció előrébb szerepel, mint a 4-2 kombináció, a 3-4, az 1-2, az 1-3 és a 2-5 pedig előrébb szerepel, mint a 4-3, a 2-1, a 3-1 és az 5-2 kombinációk. Ebből persze még nem következik, hogy egy József Attilára specifikusan jellemző mintázatot találtunk, hiszen könnyen lehet, hogy egy, a rímhez kapcsolódó általános, más szerzőkre is jellemző sémáról van szó, amely valamilyen kognitív, pszicholingvisztikai tényezővel magyarázható. A kérdés eldöntéséhez a jövőben érdemes lenne más szerzők verseit is megvizsgálni ebből a szempontból.

Az előző, *17. és 18. táblázatok* gyakorisági listái figyelembe vették a rímpárok elemeinek a sorrendjét. Bizonyos kutatási kérdések megválaszolásához azonban szükséges lehet olyan listák előállítása is, amelyek nem veszik figyelembe a sorrendet. Ezt mutatja be a 19. és a 20. táblázat. A *19. táblázat* két oszlopában a korpuszban legalább négyszer előforduló rímpárokat, illetve lemmatizált rímpárokat tüntettem fel, a rímpárok tagjainak a sorrendjét figyelmen kívül hagyva (vagyis az *ember-tenger* vagy a *maga-van* rímpár a *tenger-ember* és a *van-maga* kombinációkat is magában foglalja).

| | **Token** | |
|---|---|---|
| 1 | ember – tenger | 6 |
| 2 | hamis – is | 5 |
| 3 | magam – van | 5 |
| 4 | engem – szivemben[31] | 4 |
| 5 | is – mégis | 4 |
| 6 | lelkem – telken | 4 |
| 7 | végtelenbe – egyre | 4 |
| 8 | agyunk – vagyunk | 4 |
| 9 | szerelem – velem | 4 |
| 10 | költemény – én | 4 |
| 11 | mindörökre – szemegödre | 4 |

[31] A különböző írásváltozatú alakok természetesen különböző tokenekként kerülnek be a listába. Ugyanakkor az *engem-szivemben* rímpár esetében rákerestem az *engem-szívemben* rímpárra, amelyből egyet találtam, ezzel együtt tehát összesen öt előfordulásról beszélhetünk.

 DOI 10.31400/dh-hun.2020.3.422

**Lemma**

| | | |
|---|---|---|
| 1 | maga – van | 8 |
| 2 | ember – tenger | 7 |
| 3 | lélek – telek | 6 |
| 4 | maga – szó | 5 |
| 5 | agy – van | 5 |
| 6 | hamis – is | 5 |
| 7 | hisz – visz | 4 |
| 8 | elme – szerelem | 4 |
| 9 | is – mégis | 4 |
| 10 | nyom – ottan | 4 |
| 11 | ellen – szellem | 4 |
| 12 | egyre – végtelen | 4 |
| 13 | ragyog – van | 4 |
| 14 | vágy – ágy | 4 |
| 15 | szerelem – én | 4 |
| 16 | költemény – én | 4 |

**19. táblázat**. Rímpárok gyakorisági listája tokenek és lemmák alapján – a rímpárok tagjainak a sorrendje nem számít.

A *20. táblázat*ban a rímpárok tíz leggyakoribb szófaj- és szótagszám-kombinációját tüntettem fel a rímpárok tagjainak sorrendjét figyelmen kívül hagyva (vagyis a főnév-ige vagy a szótagszámra vonatkozó 2-3 listaelemek az ige-főnév és a 3-2 kombinációkra is vonatkoznak).

**Szófaj**

| | | |
|---|---|---|
| 1 | főnév – főnév | 1312 |
| 2 | főnév – ige | 1225 |
| 3 | ige – ige | 608 |
| 4 | melléknév – főnév | 372 |
| 5 | határozószó – főnév | 331 |
| 6 | főnév – névmás | 236 |
| 7 | melléknév – ige | 189 |
| 8 | határozószó – ige | 153 |
| 9 | névmás – ige | 126 |
| 10 | névutó – főnév | 109 |

**Szótagszám**

| | | |
|---|---|---|
| 1 | 2 – 3 | 1144 |
| 2 | 2 – 2 | 1088 |
| 3 | 2 – 4 | 601 |
| 4 | 3 – 3 | 590 |
| 5 | 1 – 2 | 377 |
| 6 | 3 – 4 | 364 |

| | | |
|---|---|---|
| 7 | 1 – 3 | 292 |
| 8 | 4 – 4 | 192 |
| 9 | 1 – 1 | 177 |
| 10 | 2 – 5 | 116 |

**20. táblázat**. Rímpárok gyakorisági listája szófajok és szótagszámok alapján – a rímpárok tagjainak a sorrendje nem számít.

## 7. Összegzés

A tanulmány József Attila verseinek a példáján keresztül mutatta be a vershangzáshoz kapcsolódó jellemzők automatikus feltárásának egy módját, amely két fő lépésből állt. Az első lépés a korpusz automatikus annotálása volt. Ennek során egy XQuery szkripttel annotáltam a versek szerkezeti egységeit, azaz a versszakokat és a sorokat, valamint az *e-magyar* és a *hunpoem_analyzer-TEI* program lefuttatásával a korpusz szavainak a grammatikai tulajdonságait és a vershangzáshoz kapcsolódó jellemzőket. A grammatikai tulajdonságok annotálása a szavak lemmájának, szófajának és morfoszintaktikai tulajdonságainak az automatikus elemzését jelentette. A vershangzás tulajdonságainak az annotálása pedig a szavak főbb fonológiai tulajdonságainak, a versszakok rímképletének, a rímpároknak, a sorok ritmusának, valamint az alliterációknak az automatikus elemzésére terjedt ki. A korpusz hangzásjellemzőinek feltárásában a második lépés a gyakorisági listák generálása volt, amelyhez az XML-adatbázisok lekérdezésére szolgáló XQuery nyelvet használtam. A vershangzás automatikus elemzése számos ponton továbbfejleszthető. A távlati tervek között szerepel az időmértékes ritmusból metrumra absztrahálás funkciójának, az ütemhangsúlyos verselés megállapításának, valamint a rímek fokozati alapon történő elkülönítésének a beépítése az annotáló programba.

Nem volt célom, hogy a gyakorisági listák alapján következtetéseket vonjak le József Attila költészetéről. Csupán annak bemutatására törekedtem, hogy hogyan lehet a vershangzás jellemzőire vonatkozó olyan kvantitatív adatokhoz jutni, amelyek alapját adhatják különböző irodalomtudományos vizsgálatoknak. Az ilyen vizsgálatok nem pusztán egy adott költő verseire terjedhetnek ki, hiszen különböző alkotók verseiből nyert kvantitatív jellemzők összevetéséből derülhet ki, hogy mi az, amiben sajátos, és mi az, amiben általános mintázatokat valósít meg egy életmű vagy annak egy része. A lexikai tulajdonságok és a vershangzás jellemzőinek az automatikus feltárása arra is lehetőséget ad, hogy egy-egy időszakra, vagy éppen egy adott lírai műfajra vonatkozóan állapítsunk meg kvantitatív jellemzőket. Irodalmi szövegek kvantitatív jellemzőinek a feltárása, illetve az ilyen jellemzők alapján történő vizsgálatok természetesen nem helyettesíthetik a szövegek szoros olvasásával megvalósuló kvalitatív elemzéseket. A két megközelítést egymást támogató módszerként érdemes kezelni, amelyek másmás perspektívából tekintenek az irodalmi szövegekre, és így eltérő típusú kérdésekre adhatnak válaszokat.

 DOI 10.31400/dh-hun.2020.3.422

## Automatic Analysis of Sound Devices in Attila József's Poems

The paper presents a method of automatic analysis of sound devices by using Attila József's poems as a case study. The first half of the paper discusses the functions of the program *hunpoem_analyzer-TEI,* which was developed for the automatic annotation of sound devices and addresses the main steps of the annotation process. The second half of the paper presents different frequency lists of lexical features and sound devices extracted from the annotated corpus.

Keywords:
automatic poetry analysis, corpus linguistics, sound devices, rhyme, rhythm, Attila József, *hunpoem_analyzer-TEI, e-magyar*