**Labádi Gergely †**

*Szegedi Tudományegyetem, Magyar Irodalmi Tanszék*

# Az olvasó gép: Berzsenyi Dániel versei távolról*

Tanulmányom helyzetfelmérés és kísérlet. Az érdekel, a rendelkezésre álló magyar nyelvi számítógépes elemzőkkel és statisztikai programokkal lehetséges-e a Berzsenyi-versek csoportosításával kapcsolatos, a szoros olvasás révén felmerült hipotézist igazolni, illetve meg lehet-e válaszolni olyan, Berzsenyi hangzóelméleti és -használati gyakorlatát illető kérdéseket, amelyeket szoros olvasással nyilvánvalóan nem lehetne, emellett bemutatok egy lehetséges elemzési módszert, amely egyes fogalmak felbukkanását, eloszlását és összefüggéseit kutatja. Olyan belső kapcsolatokat, struktúrákat tárhatunk fel, illetve olyan modelleket, törvényszerűségeket állapíthatunk meg, amelyek a szoros olvasás révén egyszerűen nem állnának rendelkezésünkre. Egy olyan szemléletmódról és módszertanról van tehát szó, amely korábban elképzelhetetlen mértékű kontextualizációt tesz lehetővé, amely az egyedi szövegek környezetének, az egyes szövegeket meghatározó irodalmi, társadalmi, kulturális trendeknek az azonosítását, elemzését ígéri: ez a történeti és stilisztikai kérdéseknél egyértelmű előnnyel rendelkezik a szoros olvasással szemben.

Kulcsszavak:
Berzsenyi Dániel, R, Magyarlanc, tartalomelemzés, távoli olvasás

## 1. Géppel mérni

Tanulmányom helyzetfelmérés és kísérlet. Az elmúlt néhány évben a klasszikus irodalomtörténeti, irodalomtudományos kérdések mellett – mint oly sokunkat – a digitális bölcsészet elméleti és gyakorlati kérdései, problémái is foglalkoztattak.[1] A számítógép használata immár végérvényesen és egyértelműen nem pusztán eszközprobléma, mint

* A kutatást az EFOP-3.6.1–16–2016–00008 azonosítójú, EU társfinanszírozású projekt támogatta (Intelligens élettudományi technológiák, módszertanok, alkalmazások fejlesztése és innovatív folyamatok, szolgáltatások kialakítása a szegedi tudásbázisra építve).

1 Labádi Gergely, „A filológiai tudás formái," in *Textológia – filológia – értelmezés: Klasszikus magyar irodalom„* szerk. Czifra Mariann és Szilágyi Márton (Debrecen: Debreceni Egyetemi Kiadó, 2014), 173–190; Labádi Gergely, „A magyar regény adatbázisa," *Acta Historiae Litterarum Hungaricarum* 32 (2016): 11–30. Szintén ennek az érdeklődésnek jegyében szerveztük meg Kokas Károllyal és Péter Róberttel közösen 2015 októberében a *Digitális bölcsészet Szegeden* című workshopot (hozzáférés: 2018.05.23, `http://digibolcsesz.ek.szte.hu/`).

aminek az 1970-es években esetleg látszott.[2] A '90-es évek közepének felkiáltása („The Information Society is now upon us!") ugyan jelezheti egy buzzword aktuális tündöklését, de mint tudjuk, ennél éppenséggel többről van szó. Az információs társadalom a modernitás uralkodó médiumának, a könyv pozícióinak megroppantásával kihívást jelent a humántudományok számára mind a gyakorlat, mind az elmélet vonatkozásában. Hivatkozhatunk akár a mindennapokra is, arra, hogy – legalábbis a nyugati tudományosságban – a *Digital Humanities* a 2000-es évek második felére szervezetileg, intézményileg végleg és egyértelműen áttört,[3] azaz vannak képzések, folyóiratok, konferenciák vagy minimum konferenciaszekciók, a bölcsészállásoknál rendszerint elvárás valamiféle *DH*-képesség vagy -gyakorlat. A számítógépek használata azonban sokáig csak a filológia klasszikus gyakorlatainak könnyítését célozta, úgy tettünk, mintha valójában semmi sem történt volna.[4] Mellőzve most a humántudományok gyakorlatának korábban már tárgyalt kérdéseit,[5] pusztán az a módszertani megközelítés, amelyet ki *macroanalysis*ként, ki *distant,* ki *machine reading*ként, ki pedig *algorithmic criticism*ként nevez meg, sürget bennünket, hogy újragondoljuk eddigi, *close reading*ként emlegetett gyakorlatunkat és annak eredményeit.

Az irodalomtudományban eddig a „bizonyítékok" gyűjtése alapvetően szubjektív megfigyeléseken alapult, a belőlük levont általánosítás érvénye pedig a minta reprezentativitásán múlt. Matthew Jockers példája Ian Wattnak a regény felemelkedéséről szóló klasszikus munkája, amely mindössze Daniel Defoe, Samuel Richardson és Henry Fielding munkái alapján készült, jóllehet a szóba jöhető szerzők és szövegek köre nagyságrendekkel nagyobb. Adódik a kérdés, ha ezt a több ezer további szöveget el tudnánk „olvasni," más történetet mondanánk? Eddig ezt a kérdést esélyünk sem volt megválaszolni, de most már megvannak az eszközeink.

A számítógép-használat adatok gyűjtésére valójában mégsem eszközkérdés. Hiszen – durva leegyszerűsítéssel – a statisztikák irodalmi használatának elméleti és módszertani alapjai több évtizedesek, a szerzőattribúció elméletei éppenséggel még régebbiek. Tehát nem pusztán eszközkérdés: mostanra „egy olyan fordulóponthoz, eseményhorizonthoz értünk, ahol már elegendő szöveget és szakirodalmat kódoltunk ahhoz, hogy lehetővé tegye számunkra, sőt hogy kényszerítsen bennünket, új kérdéseket tegyünk fel az irodalomról és az irodalmi adatról."[6] Matthew Jockers elbeszélésében 2008 a fordulópont, és bár nem reflektál rá, de megállapítása természetesen csak az angol nyelvű szöveghagyományra vonatkozik. A német kultúra kapcsán Fotis

[2] Lásd a Jockers-kritikámban (Labádi Gergely, „Matthew Jockers, Macroanalysis: Digital Methods and Literary History," *Irodalomtörténet* 97, 4. sz. (2016): 496–500) idézett tanulmányokat: Martin L. West, *Szövegkritika és szövegkiadás,* ford. Bolonyai Gábor (Budapest: Typotex Kiadó, 1999 [1973]); Voigt Vilmos, „Számítógépes ritmuselemzési kísérlet," *Irodalomtörténeti Közlemények* 76, 2. sz. (1972): 203–211.

[3] Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana: University of Illionis Press, 2013), 11–23.

[4] „Der Einsatz von EDV in der editorischen Arbeit grundsätzlich eine Arbeitserleichterung und eine weitaus grössere Präzision in den zu erzielenden Ergebnissen mit sich bringt." Bodo Plachta, *Editionswissenschaft: Eine Einführung in Methode und Praxis der Edition neuerer Texte* (Stuttgart: Reclam, 1997), 131.

[5] Labádi, „A filológiai tudás," 173–190.

[6] Jockers, *Macroanalysis,* 4.

 DOI 10.31400/dh-hun.2018.1.126

Jannidis és Gerhard Lauer szerint 2011, a TextGrid repozitóriumának megindulása óta áll fenn a kvantitatív vizsgálatok alapfeltétele, a megfelelő mennyiségű és minőségű digitalizált szöveg.[7] Mi a helyzet a magyar nyelven írott hagyománnyal? Az Arcanum szolgáltatásában az elmúlt 200 év magyar tudományának több mint 11 millió oldalnyi anyaga van digitalizálva, a Magyar Elektronikus Könyvtár (MEK) 16000 dokumentum fölött jár, a Digitális Irodalmi Akadémiában (DIA) pedig több mint 80 kortárs életmű férhető hozzá. Az Országos Széchényi Könyvtár ELDORADO szolgáltatása a még nem digitalizált források kérdését próbálja, kívánja rendezni. Érdemes tehát nekünk is elgondolkodni, új kérdéseket feltenni, még ha jelenlegi formájában az előbb felsoroltak közül nem is minden alkalmas kvantitatív vizsgálatok lefolytatására – a kétrétegű PDF például korántsem a digitális mint olyan kvintesszenciája; a DIA anyaga nem könnyen hozzáférhető, a MEK szövegeinek megbízhatósága pedig problematikus, azaz a magyarországi digitalizált szövegek nem egyenletes minőségűek, és nem egységes elvek szerint készültek.

Az új megközelítés kiindulópontja, bárhogy is nevezi magát, hogy az egyedi szövegtől eltávolodva olyan redukciós és absztrakciós eszközöket nyerünk, amelyek a tudás speciális, a konkrét szöveg realitása felől nem észlelhető formáját kínálják. Olyan belső kapcsolatokat, struktúrákat tárhatunk fel, illetve olyan modelleket, törvényszerűségeket állapíthatunk meg, amelyek a szoros olvasás révén egyszerűen nem állnának rendelkezésünkre. Egy olyan szemléletmódról és módszertanról van tehát szó, amely korábban elképzelhetetlen mértékű kontextualizációt tesz lehetővé, amely az egyedi szövegek környezetének, az egyes szövegeket meghatározó irodalmi, társadalmi, kulturális trendeknek az azonosítását, elemzését ígéri: ez a történeti és stilisztikai kérdéseknél egyértelmű előnnyel rendelkezik a szoros olvasással szemben. Ugyanakkor nagy mennyiségű adat számítógépes elemzése egyszerűen csak egy alternatív módszer adatokat gyűjteni, hipotéziseket felállítani, ám ezeket ugyanúgy kritikával kell kezelni, és értelmezni szükséges.

A „nem-olvasás" – ahogy Franco Moretti saját gyakorlatát provokatívan nevezi[8] – tehát használható korábbi hipotézisek ellenőrzésére, új hipotézisek felállítására, korábban fel nem ismert, lehetséges kutatási területek körülhatárolására. Mivel Moretti és Jockers könyvéről írott recenzióimban már utaltam néhány konkrét példára, most inkább olyan területeket neveznék meg, amelyek megoldására eddig nemigen volt lehetőségünk, ugyanakkor a számítógépek használata révén lehetséges, vagy legalábbis erős érveket kaphat egyik vagy másik hipotézis, tehát lépéseket tehetünk a megoldás érdekében. Ha tétel bíró vizsgálatot keresünk, akkor rögtön adódnak az eldöntetlen szerzőségű szövegek. Ilyen akad bőséggel a magyar irodalomban, és a kanonikus szövegek között például a *Fanni hagyományai* szerzősége az egyik leg-

[7] Fotis Jannidis and Gerhard Lauer, „Burrows's Delta and Its Use in German Literary History" in *Distant Readings: Topologies of German Culture in the Long Nineteenth Century,* eds. Matt Erlin and Lynne Tatlock (Rochester: Camden House, 2014), 30.

[8] „What does it mean, studying world literature? How do we do it? I work on West European narrative between 1790 and 1930, and already feel like a charlatan outside of Britain or France. World literature? Many people have read more and better than I have, of course, but still, we are talking of hundreds of languages and literatures here. Reading more seems hardly to be the solution. […] Reading more is always a good thing, but not the solution." Franco Moretti, „Conjectures on World Literature," *New Left Review* 1 (2000. jan.-febr.): 55.

nagyobb filológiai rejtélyek egyike. Az értelmezéstörténet sokkal inkább az aktuális irodalompolitikai kívánalmak szerint alakult, semmint valamiféle ellenőrizhető filológiai, nyelvi módszertan alapján. Pedig Szinnyei Ferenc egyszer már megpróbálta nyelvészeti módszerekkel megoldani a kérdést, de a szakirodalom nem fogadja el érveit – még ha egybeesik is a mostani konszenzussal.[9] Ennek azonban nemcsak a kiindulópont bizonytalansága az oka, maga a követett módszertan sem meggyőző. A szerzőattribúció során ugyanis tipikusan nem egy-egy tartalommal bíró kifejezést kell vizsgálni, hanem a leggyakoribb szavakat *(most frequent words),* amelyek között a funkciószavak (például névmások, névelők, viszonyszók) dominálnak – Jockers szerint még az írásjelhasználat is árulkodó –, illetve e szavaknak a szöveg egészéhez viszonyított arányát kell nézni.[10] Szinnyei viszont már egy-egy tartalommal bíró szó felbukkanását döntőnek tartja. Időnként persze maga is elismeri, hogy némelyik kifejezés elég népszerű a kor érzékeny prózájában, tehát szerzőattribúcióra kevéssé alkalmas – mint például a *csendes*, az *édes* vagy a *kedves* –, vagy hogy valójában a szöveg „numerozitása," a felkiáltások szintén az érzékenység stílusjellemzői. És akkor még nem beszéltünk a szerkesztői négykezesek lehetőségéről sem – ami a korszakból az *Egyéni és eszményi* attribúciójának is kérdése egyébként –, amelyet viszont a számítógépes módszerek már kezelni tudnak.[11]

Szintén szerzőattribúciós kérdés, ugyanakkor jóval hevesebb indulatokat váltott ki még szakmai körökben is Kosztolányi Dezső név nélküli cikkeinek kérdése, különösen az *Új Nemzedék*ben vállalt szerepének története.[12] A vitához új adatokkal járulnának hozzá a körültekintően használt, előtte alaposan, semleges szövegeken kipróbálva „belőtt" szerzőattribúciós vizsgálatok. A századelő újságírói gyakorlatának filológiai kérdéseit az irodalomtörténészi szakma szintén nyelvi, stiláris eszközökkel kívánta megoldani, ugyanakkor a követett módszertan, ha talán közelebb is áll a szerzőattribúciós kutatások mai állásához, mint Szinnyei javaslata, nem gépi eszközöket használ,[13] ezért az eredményei kétségesek, szubjektívabbak. Sok bennük a megérzés: Péter László nem véletlenül említi fent hivatkozott tanulmánya utolsó bekezdésében, hogy számítógéppel kellene folytatni a vizsgálatokat, többet és többfélét lehetne elemezni.

Az irodalmi szövegekkel kapcsolatban vannak ugyanakkor más jellegű, az írói nyelvhasználathoz (is) kötődő, de sosem igazolt kijelentések. Mikszáth elbeszélőmű-

[9] Szinnyei Ferenc, *Kármán József és az Uránia névtelenjei* (Budapest: MTA, 1924). Vö. Szilágyi Márton, *Kármán József és Pajor Gáspár Urániája* (Debrecen: Kossuth Egyetemi Kiadó, 1998), 80.

[10] Jockers is foglalkozik a kérdéssel, de nagyhatású volt Burrows tanulmánya: John Burrows, „Questions of Authorship: Attribution and Beyond," *Computers and the Humanities* 37 (2003): 5–32, `https://doi.org/10.1023/A:1021814530952`. Informatív történeti összefoglalásként pedig: Harold Love, *Attributing Authorship: An Introduction* (Cambridge: Cambridge University Press, 2002), `https://doi.org/10.1017/CBO9780511483165.001`.

[11] Brian Vickers, *Shakespeare, Co-Author: A Historical Study of Five Collaborative Plays* (Oxford: Oxford University Press, 2002).

[12] Számos cikk foglalkozik a kérdéssel, csak egy összefoglalót idézek, amely utal a vita több résztvevőjének írására: Lengyel András, „Egy s más az Új Nemzedék Pardon rovatáról," *Kalligram* 19, 6. sz. (2010): 89–99. Köszönöm Bíró-Balogh Tamásnak a kérdés történetéről adott felvilágosítását.

[13] Péter László, „Juhász Gyula névtelen cikkeinek felismerése stílussajátságai alapján" in *Jelentéstan és stilisztika: A Magyar Nyelvészek 2. Nemzetközi Kongresszusának előadásai,* szerk. Imre Samu, Szathmári István és Szüts László (Budapest: Akadémiai Kiadó, 1974), 454–457; Lengyel András, „Kosztolányi-dubiózák," *Forrás* 38, 11. sz. (2006): 91–113.

 DOI 10.31400/dh-hun.2018.1.126

vészete, különösen a *Tót atyafiak* és *A jó palócok* kapcsán már az első megjelenés óta gyakran elhangzik az élőbeszédszerűség kategóriája. Az értelmezésekben azonban ennek kifejtése általában meglehetősen homályos marad. A Tahin Szabolcs által összegyűjtött kijelentések jól jelzik, mennyire nehéz megfogni e sajátosságát a novelláknak: „Mintha csakugyan eleven beszéd csengene fülembe olvasásakor…," „Mintha nem is volna köztünk az a nagy távolság, mely írót és olvasót elválasztja…," „Mintha minden szó egyenesen az ajkáról lebbent volna a papírosra…"[14] Ugyanakkor bármilyen homályosak is az idézetek, az egyértelmű, hogy az élőbeszéd az írásbeliséggel van szembeállítva. Ebben az esetben pedig lennie kell mérhető különbségeknek – vannak is.[15] Az élőbeszédszerűség narratológiai elemzései, amelyek a novellák elbeszélőjéhez, elbeszélői tudatához kapcsolva hozzák fel, természetesen meggyőzők, és ettől a vizsgálattól függetlenek, bár a továbblépés útját mutatják a különböző „hangok" stiláris sajátosságainak vizsgálatára figyelmeztetve. Herczeg Ferenc egyes regényeinek szókincsével kapcsolatban is elhangzott a szegényes minősítés – ez is mérhető adat.[16]

Heves reakciókat vált ki, amikor egy-egy klasszikust modernizálnak. Az átdolgozók amellett érvelnek, hogy így közelebb hozzák a szöveget a mai diákokhoz, hogy majd így kedvet kapnak az eredetihez, az ellenzők viszont – már-már a morális pánik jeleit mutatva – azt állítják, hogy csak elcsökevényesítjük a diákokat, ha nem az eredetivel szembesülnek. A vitát megoldani persze nem lehet, de a számítógépes szövegvizsgálatok abban segíthetnek, hogy megválaszoljuk, valójában mely tulajdonságai tűntek el a régi szövegnek, megőrzött-e valamit az eredeti nyelvi struktúrákból, mit domborít ki az új verzió, mennyiben veszi át az átdolgozó szövegeinek nyelvi sajátosságait.[17] A mérhető sajátságok tárháza széles – a mondatok összetettségétől kezdve a szófaji arányokon át az átlagos szóhosszúságig –, és valójában hozzásegíthet bennünket az eredeti néhány, korábban talán nem is észlelt sajátosságának fölismeréséhez. Valamint a sikeres/nem sikeres írók közti különbségek is mérhetők – még ha a mérések nem is magyaráznak meg mindent.[18]

## 2. Verseket mérni

Az eddig hivatkozott tanulmányok, a Shakespeare-szakirodalom kivételével, mind prózai szövegekkel foglalkoznak, de természetesen vannak olyan tanulmányok is, amelyek verseket vizsgálnak. Ezek azonban Berzsenyi esetében kevésbé segítenek. Az angol nyelvű költészet sikeres, antológiákban gyakran publikált darabjait vizsgáló tanulmány eredményei – az első sor jellemzően kevesebb szótagból áll, mint „ismeretlen" kortársának verse, és jellemzően monoszillabák alkotják, továbbá a sikeres költeményeknek egyszerűbb a szótára, rövidebbek a szavai – nem túl érdekesek, illetve

[14] Tahin Szabolcs, „»Élőbeszédszerűség« Mikszáth prózájában," *Tiszatáj* 57, 11. sz. (2003): 53–71.

[15] Labádi Gergely, *Géppel mért irodalom: a mikszáthi élőbeszédszerűség,* kézirat, 2017.

[16] Tóth Mihály, *Herczeg Ferenc írói szókincsének vizsgálata,* kézirat, 2017.

[17] A több klasszikust, például az *Egri csillagok*at, *A kőszívű ember fiai*t átdolgozó Nógrádi Gergely maga is író.

[18] Jodie Archer and Matthew L. Jockers, *The Bestseller Code: Anatomy of the Blockbuster Novel* (New York: St. Martins Press, 2016).

egy-egy költői korpusz vizsgálatára kevéssé alkalmasak.[19] Van ugyan izgalmasabb eredményeket felmutató számítógépes kutatás is, amely az angol nyelvű amatőr és professzionális költők versei közötti különbségeket keresve arra jutott, hogy elsősorban nem a versírás megtanulható és könnyen mérhető sajátosságai jelzik a különbséget (mint a rímek vagy a prozódia), hanem a használt kifejezések konkrétsága.[20] Ez azonban a megfelelő háttéranyag nélkül, ti. a magyar szavak konkrétsági fokát jelző szótár nélkül nálunk nem alkalmazható – a szerzők és Simon Eszter jóvoltából hozzám eljutott kéziratos cikk sajnos nem túl nagy (296 főnév) korpuszt dolgoz fel[21] –, jóllehet a megállapítás egybeesik a 19. század második felének Hász-Fehér Katalin által feltárt kritikai diskurzusával.[22] A lengyel költészet grammatikai rímeinek, a grammatikairím-használat történetének statisztikai vizsgálata ugyan a magyar anyag kapcsán is ígéretes, ám megfelelő háttéranyag nélkül egy-egy költői korpusz elemzésére nem lehet használni.[23]

A kísérletet ugyanakkor támogatja, hogy versek statisztikai jellegű vizsgálatának magyar szakirodalma, kutatástörténete is van, amely pontosan megnevezi, milyen adatokat milyen intencióval mér. Zsilka Tibor egyik tanulmánya például kifejezetten a magyar költői nyelv statisztikai vizsgálatának, a vizsgálat szabályainak kialakítását tűzte ki célul azzal a szándékkal, hogy „egy oly korban, amikor garmadával jelennek meg a különböző színvonalú s tartalmú verskötegek és egyéb írásos művek," legyenek olyan „egzakt módszerek," amelyekkel – túl „az intuíció"-n – megítélhetők.[24] Bár számos izgalmas megállapítást tartalmaznak Zsilka vizsgálatai, azzal nem tudok egyetérteni, hogy egy szerző azért volna jobb költő, mint a másik, mert a jeltípusok (szótári szavak, a képletben: V) és a tényleges szóalakok (N) aránya (Type Token Ratio) nála a magasabb, azaz választékosabb, „gazdagabb" a szókincse, mint költőtársáé, vagy esetleg kisebbek egyes szófajok ismétlődési arányai, vagy mert több igét használ, mint a másik, tehát „stílusa dinamikus."[25] A cikk – mint a példák is mutatják – egyedül a szó- és szófaji statisztikákkal foglalkozik Cselényi László és Juhász Ferenc versei kapcsán. De nemcsak az eredmények értékelése lehet kérdéses. Zsilka a szókincsgazdaság mérésére Pierre Guiraud[26] nemzetközileg elfogadott képletét alkalmazza:

[19] Richard S. Forsyth, „Pops & Flops: Some Properties of Famous English Poems," *Empirical Studies of the Arts* 18, 1. sz. (2000): 49–67.

[20] Justine Kao and Dan Jurafsky, „A Computational Analysis of Style, Affect, and Imagery in Contemporary Poetry," *Linguistic Issues in Language Technology* 12, 3. sz. (2015), `https://nlp.stanford.edu/pubs/kaojurafsky12.pdf`. Lásd még Michael Dalvean, „Ranking Contemporary American Poems," *Digital Scholarship in the Humanities* 30, 1. sz. (2015): 6–19.

[21] Fekete István és Babarczy Anna, *Főnévi fogalmak konkrétsági, elképzelhetőségi és definiálhatósági értékeinek összefüggései,* kézirat, 2008.

[22] Hász-Fehér Katalin, „A dilettantizmus kérdése a 19. század közepének kritikáiban," *Acta Historiae Litterarum Hungaricarum* 32 (2016): 77–116.

[23] Karol R. Opara, „Grammatical Rhymes in Polish Poetry: A Quantitative Analysis," *Digital Scholarship in the Humanities* 30, 4. sz. (2015): 589–598.

[24] Zsilka Tibor, *Stilisztika és statisztika* (Budapest: Akadémiai Kiadó, 1974), 19.

[25] Zsilka, *Stilisztika és statisztika,* 16.

[26] Guiraud munkájáról részletes ismertető olvasható magyarul: J. Soltész Katalin, „Guiraud statisztikai módszere a szókincs vizsgálatában" in *Általános nyelvészeti tanulmányok I.,* szerk. Telegdi Zsigmond (Budapest: Akadémiai Kiadó, 1963), 263–272.

 DOI 10.31400/dh-hun.2018.1.126

$$R = \frac{V}{\sqrt{N}}$$

Egy másik tanulmányában ugyan jelzi, hogy Guiraud e képlettel végzett szókincsgazdagsági vizsgálatai a francia nyelven írt szövegek esetében eredményeznek 20,5-ös átlagot,[27] de ezt az értéket mégis felhasználja Cselényi és Juhász költészetének összevetésére, mégis erre hivatkozva állítja, hogy a képlet nem használható 2000 szó alatti szövegek esetében, mert „a kapott eredmény nem egyezik az empirikus elvárásszinttel," azaz egy 742 szavas ciklus esetében a kapott 17,60 egy Kosztolányi-korpusztól kevés, mert „nem tükrözi híven a költő szókincsgazdagságát."[28] Ami éppenséggel lehet, hogy igaz, de mivel nem végzett kontrollvizsgálatot, nem nézte meg, 2000 szó fölött miként alakul ez az érték Kosztolányinál, valójában nem tudjuk, mekkora Kosztolányi verseinek TTR-értéke, a 17,60 valóban túl alacsony érték-e.

A kötet azonban még így is rendkívül inspiráló. A *Mérések a szöveg fonetikai, ritmikai és morfológiai szintjén* című tanulmánya 20. század elejének magyar költészetből vett példákat elemez.[29] A különböző képzési helyű magán- és mássalhangzók egymáshoz viszonyított arányainak, a szavak hosszúságának, szerkezeti fölépítésének vizsgálata ugyan öncélúnak tűnhet, de egyrészt valóban sikerül megragadni a stílusbeli különbségeket – másképpen: a különbségek magyarázhatók a stílusbeli eltérésekkel –, másrészt a háttérben Fónagy Iván nagy anyagot mozgató izgalmas vizsgálatai állnak, például a 19. századi francia költészet mássalhangzó-használata statisztikai és stilisztikai eredményeinek Petőfi költészetével való összevetése.[30] Ráadásul ez a szempont egyértelműen illeszkedik Berzsenyi versszemléletéhez is. A *Poétai harmonistika* egyik passzusa ugyanis így hangzik:

> Harmóniába hozza a' poéta tárgyaival a' beszédet akkor, midőn stylusát tárgyai' természetéhez alkalmaztatja. […] Kiterjed ezen harmóniázat még a' beszéd' külhangjaira is, úgy hogy a' poéta, valamennyire csak a' nyelv' természete engedi, a' gyengébb érzelmekhez és szebb tárgyakhoz lágyabb és szebb hangu szavakat válogat, a' zordonabb tárgyakhoz pedig keményebb hanguakat.[31]

Azt sajnos nem lehet tudni, hogy Berzsenyi egészen pontosan mely hangokat gondolta lágyabbnak és szebbnek, melyeket pedig keményebbeknek. Fónagy Iván ugyan részletesen adatolja, hogy ez az elképzelés az ókortól kezdve a retorikai-poétikai gondolkodás része, de mivel nem teljesen egybevágók az elképzelések, hiba volna bármelyiket is kiemelni. Egy levélrészlet ugyanakkor segít valamelyest közelebb lépni a kérdéshez. Berzsenyi egy Kazinczynak 1811 nyarán írt levélben helyesírási kérdésekről elmélkedve írja: „Annyi, mennyi, honnyi, olly, melly, st. mert ezeknek hangjok

[27] Zsilka, *Stilisztika és statisztika,*16.
[28] Zsilka, *Stilisztika és statisztika,* 22.
[29] Zsilka, *Stilisztika és statisztika,* 46–75.
[30] Fónagy Iván, *A költői nyelv hangtanából* (Budapest: Akadémiai Kiadó, 1989), 36–42. Zsilka értelemszerűen az első kiadást használta [1959].
[31] Berzsenyi Dániel, *Prózai munkái,* kiad. Fórizs Gergely (Budapest: Editio Princeps, 2011), 388.

bizonytalan, vagy legalább a' különbség nem igen metsző a' kemény l, n, és lágy ly, ny között, és mivel a lágy ly többnyire még csak végső betűinkbe csuszott bé."[32]

Bármily változatos is a hangok jelentésével kapcsolatos hagyomány az ókortól a 19. századig (és tovább), az *l* és az *n* kemény volta nem szerepel benne – legalábbis Fónagy anyaggyűjtése alapján. Nehéz értelmezni Berzsenyi kijelentését. Egyedüli nyom, amely talán magyarázhatja az állítást, Révai stilisztikájának egyik szöveghelye. A „szép hanggal" kapcsolatos hibák („keménység," „egyenlő hangozat," „egyhangúság") kapcsán olvashatjuk:

> A' magános szavakban a sértő keménység minden nyelvekben a' mássalhangzóktól vagyon. A' mássalhangzók öszveállató valóságos részeik ugyan a' szavaknak, mert tulajdonúl azokban áll minden jelentésök; a' magánhangzók pedig tsak nyilásai a' szájnak, mellyekel a' mássalhangzók kimondatnak.[33]

Révai tehát azt állítja, hogy alapvetően minden mássalhangzó kemény, torlódásukat érdemes hát elkerülni, kivált, ha az adott mássalhangzók „már magokban is keményebbek" – teszi hozzá ugyanitt. Bár stilisztikájából nem derül ki, melyek ezek a „magokban" is keményebb mássalhangzók, Révai felfogása mégis magyarázhatja Berzsenyi kijelentését, és ami fontosabb, számunkra lehetőséget nyújt arra, hogy költészete sajátszerűségét megpróbáljuk megmérni – legalábbis egy vonatkozásban, a jólhangzás Révai ugyanezen passzusokban megfogalmazta követelményének és feltételeinek ismeretében. A tanulmány második részében erre teszek kísérletet, még ha ennek a hozadéka önmagában nem is túl sok, mivel egyelőre még hiányzik a háttér, azaz a korszak nyelvi jellemzőinek szélesebb, több szerzőt, több szövegtípust feldolgozó felmérése. Emellett a következőkben bemutatok még egy lehetséges elemzési módszert, amely egyes fogalmak felbukkanását, eloszlását és összefüggéseit kutatja. Ennek eredményei persze vita tárgyát képezhetik – egyáltalán, megfelelő fogalmakat vizsgáltam? –, de egyrészt jelzik a számítógépes vizsgálatok potenciálját, másrészt az eredmények elrendezésekor, vizualizálásakor felmerülő problémák is érdekes módszertani kérdéseket vetnek fel.

### 2. 1. A korpusz előkészítése

A vizsgálatokhoz a Magyar Elektronikus Könyvtárban található Berzsenyi-verseket választottam. Az Arcanum *Verstár* CD-je alapján készült verzió textológiai szempontból ugyan nem megfelelő, de mivel ez az interneten ma hozzáférhető magyar költészeti korpusz alapja, modernizálása egységes szempontok alapján és egyenletes minőségben készült, valamint az Országos Széchényi Könyvtár mint befogadó intézmény mégiscsak hitelesíti, végül úgy döntöttem, kiindulásként e kísérletben érdemes elfogadni.[34] Magam még annyit módosítottam a morfológiai elemzés pon-

[32] Berzsenyi Dániel Kazinczy Ferencnek, Nikla, 1811. június 5. in Berzsenyi Dániel, *Levelezése,* kiad. Fórizs Gergely (Budapest: Editio Princeps, 2014), 222.

[33] Révai Miklós, *A magyar szép toll,* kiad. Éder Zoltán (Budapest: Akadémiai Kiadó, 1973), 52.

[34] A szövegek átírása mindig is a számítógépes vizsgálatok neuralgikus pontja volt. Lásd Vadai István kritikáját az első magyar számítógépes ritmuselemzésről: Vadai István, „Számítógép a verstan szolgálatában: Megjegyzések egy számítógépes ritmuselemzési kísérlet kapcsán," *Irodalomtörténeti Közlemények* 88, 1. sz. (1984): 77.

 DOI 10.31400/dh-hun.2018.1.126

tossága érdekében, hogy a kis- és nagybetűket a mondatstruktúrához igazítottam – kivéve értelemszerűen a tulajdonneveket –, valamint a versek tördelését is prózaivá alakítottam, mert a szövegek MEK-es kódolása sajnos üres sorokat tett be a versszakok közé, ami a morfológiai elemző számára megtévesztő volt. A legsúlyosabb, feltehetően a karakterfelismerés során elkövetett tévesztéseket javítottam (például *m* helyett *rn* – és fordítva). A verseket az 1816-os kötet sorrendjében mentettem le és neveztem el, azaz a vizsgálatok során a kötetrend dominált, az időrend ugyanis kevés kivételtől eltekintve hipotetikus, nagy intervallumokkal dolgozik, így informatív értéke meglehetősen csekély. Érdekes kérdés persze, hogy a számítógépes elemzések eredményét fel lehet-e használni ennek pontosítására, de ez most nem volt célom. A fogalmak korrelációinak vizsgálata a kötetrend fényében ugyanakkor hozzájárul a kötetkompozíció vizsgálatának kérdéséhez, új érveket szolgáltathat az egyes elképzelésekhez, illetve segíthet megtalálni magukat a kulcsfogalmakat is.

A fogalomkeresés előtt a korpuszt a *Magyarlanc* elnevezésű nyelvi elemzővel preparáltam.[35] Az első oszlop az eredeti szöveget tartalmazza, a második a szóalakok lemmatizált változatát, a harmadik szófaját, a negyedik pedig a szóalak morfológiai elemzését (részletesebb, mondatszintű elemzéseket is képes a program készíteni). Mint a mellékelt képből látszik, a program alapvetően elboldogul a 19. század eleji versszöveggel is, még ha nem is sikerül mindig mindent pontosan megállapítania – erre később lesz egy konkrét példa is –, hiszen például a *legnemesebb* lemmája voltaképp a *nemes* kellene, hogy legyen, de azon a szinten, amire most nekünk kell, megfelelő.

**2. 2. Fogalmak és összefüggéseik**

A fogalom- vagy inkább kifejezéskeresést azért gondoltam Berzsenyi esetében értelmes lekérdezésnek, mivel – mint jeleztem – a kötet verseinek elrendezésével kapcsolatos értelmezésekhez szolgálhat adatokkal. Berzsenyi kötetkompozíciójával kapcsolatban Onder Csaba elképzelését vettem alapul.[36] Eszerint a kötet versei a „privatum" és a „publicum" szférájáról szólnak. A privatum meghatározó élménye a szerelem és a költészet, a publicumé pedig egy virágzó közösséggel szembeállított romló, hanyatló közösség. Ennek megfelelően három kulcsfogalom köré szerveztem a lekérdezéseket: az első a magyar/hon/haza szócsoport (publicum), a második a múzsa/cüpris/camoena/szerelem (privatum), a harmadik pedig a tagadást kifejező nem/sem/ne/se/nincs. Ezek a fogalmak ugyan megfelelőnek tűntek, de huszonegy vers kimaradt, ti. sem az első, sem a második fogalomcsoport szavai nem fordultak elő bennük – feltételezve természetesen, hogy a versek vagy a privatum, vagy a publicum körébe tartoznak, harmadik csoport nincs. A következő versekről van szó: *A Melancholia, Jámborság és középszer, Gróf Török Sophiehez, A közelítő tél, Osztályrészem, Egy hívtelenhez, Chloe, Egy szilaj leánykához, Amathus, A csermelyhez, Egy leánykához, A megelégedés, Keszthely, Fohászkodás, Az est, Szerelmes bánkódás, Lilihez, Cencimhez,*

[35] János Zsibrita, Veronika Vincze and Richárd Farkas, „magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian," *Proceedings of RANLP 2013* (2013): 763–771. *Magyarlanc*, hozzáférés: 2017.02.03, `http://rgai.inf.u-szeged.hu/index.php?lang=en&page=magyarlanc`.

[36] Berzsenyi Dániel, *Versek: teljes gondozott szövegek: 1816,* szerk. Onder Csaba, Matúra Klasszikusok (Budapest: RaabeKlett, 1998), 21–37.

```
Boldog  boldog  ADJ Case=Nom|Degree=Pos|Number=Sing
vagy    van VERB    Definite=Ind|Mood=Ind|Number=Sing|Person=2|Tense=Pres|VerbForm=Fin|Voice=Act
,   ,   PUNCT   _
Áon Áon PROPN   Case=Nom|Number=Sing
szűzei  szűz    NOUN    Case=Nom|Number=Plur|Number[psor]=Sing|Person[psor]=3
kedvese kedves  NOUN    Case=Nom|Number=Sing|Number[psor]=Sing|Person[psor]=3
,   ,   PUNCT   _
S   s   CONJ    _
nagy    nagy    ADJ Case=Nom|Degree=Pos|Number=Sing
,   ,   PUNCT   _
mint    mint    SCONJ   _
hazádnak    haza    NOUN    Case=Dat|Number=Sing|Number[psor]=Sing|Person[psor]=2
legnemesebb nemesebb    ADJ Case=Nom|Degree=Sup|Number=Sing
fia fia NOUN    Case=Nom|Number=Sing|Number[psor]=Sing|Person[psor]=3
!   !   PUNCT   _

E   e   PRON    Case=Nom|Number=Sing|Person=3|PronType=Dem
két két NUM Case=Nom|NumType=Card|Number=Sing
remek   remek   ADJ Case=Nom|Degree=Pos|Number=Sing
dísz    dísz    NOUN    Case=Nom|Number=Sing
kéri    kér VERB    Definite=Def|Mood=Ind|Number=Sing|Person=3|Tense=Pres|VerbForm=Fin|Voice=Act
méltán  méltán  ADV _
A   a   DET Definite=Def|PronType=Art
Ganymédes Ganymédes PROPN   Case=Nom|Number=Sing
poharát pohár   NOUN    Case=Acc|Number=Sing|Number[psor]=Sing|Person[psor]=3
az  az  DET Definite=Def|PronType=Art
égben   ég  NOUN    Case=Ine|Number=Sing
.   .   PUNCT   _
```

**1. ábra**. A korpusz elemzése a *Magyarlanc* alkalmazásával

*Esztihez, Báró Wesselényi Miklós képe, A temető.* – Itt valójában indulhattam volna másik irányba is – hiszen az ilyen fogalomkeresés nem más, mint számítógéppel megsegített *close reading* –, ti. ráhagyatkozom a számítógépre, és megnézem, miként csoportosítja maga a verseket, majd ebből megpróbálok következtetéseket levonni. A dolgozat egy későbbi pontján ezt röviden bemutatom.

A hiba egyik oka, hogy a *Magyarlanc* nem tudta helyesen elemezni a *múzsát, múzsához* szóalakokat és egy *múzs* alapszót tételezett. A mai magyar nyelvre épülő, meglehetősen nagy, 1200000 szót és 250000 írásjelet tartalmazó, hat területről (szépirodalom, jog, hír- és hetilapok cikkei, gazdasági hírek, kamaszoktól származó szövegek, informatikai irodalom) összegyűjtött, nyelvészek által morfológiailag elemzett tanító korpuszában nyilván nem szerepelt ez a kifejezés.[37] Tehát ezt módosítani kellett – persze ha alanyesetben szerepel a versben, akkor az elemző is *múzsa*-ként azonosítja. Ezt a csoportot kiegészítettem még az *ifjú* és *kebel* kifejezéssel, illetve a *cüpris*, *camoena* nagybetűs alakját is beemeltem a biztonság kedvéért, mert a *Magyarlanc* nem minden szót kisbetűsített. Az *isten* viszont például nem volna elég specifikus, mivel nagyon gyakran fordul elő az első csoport verseiben is (például „honnunk isteni"). Hat vers viszont még mindig kihullott. Ezért felvettem ide még a *mirtus*-t, az első csoportba pedig a *báró*-t és a *Rómá*-t, de négy vers, *A közelítő tél, Egy hívtelenhez, A csermelyhez, Fohászkodás* ezek után is fennmaradt. A második csoportba tettem ezek után a *fohász-*

[37] Csak a Szeged Dependency Treebank adatait ismertettem, jóllehet a *Magyarlanc* tanító készletében ezenkívül még a szintén kézzel annotált Szeged Corpus is szerepel. Zsibrita, Vincze and Farkas, „magyarlanc," 763–771.

 DOI 10.31400/dh-hun.2018.1.126

*kodás*-t, valamint a *labirintus*-t is, ezzel sikerült minden verset valamelyik csoportba osztani.

Mint jeleztem, a fogalmak megfelelősége kérdés tárgya, de kiinduló elemzésnek, mintának, megfelelnek. Az érdekelt, van-e valami kimutatható, számszerűsíthető összefüggés közöttük, illetve mutat-e különbséget az eredmény ahhoz képest, mintha véletlenszerűen rendeznénk el a szövegeket, azaz a versekhez véletlenszerűen rendelnénk, hogy a privatum vagy a publicum szavai fordulnak elő bennük. A vizsgálatot értelemszerűen a lemmatizált alakokon végeztem a statisztikai elemzések, illetve a természetesnyelv-feldolgozás során gyakran használt R programozási nyelven, és mindig az R beépített függvényeit, parancsait használtam az értékek kiszámításához.[38]

Mint a mellékletben látható, ha a nyers szöveg versenkénti bontásban rendelkezésünkre áll, elég kevés, mindössze huszonegy parancssorral megoldható a vizsgálat – persze a lényeg, az értelmezés, ezután következik. Az első táblázat egy részlet, amely azt mutatja, a kötet első öt szövegében a keresett kifejezések előfordulnak-e vagy sem (az első esetet „1”, a másodikat „0” jelöli). A táblázat természetesen mind a nyolcvanhét versre vonatkozó adatokat tartalmazza: az első szócsoport összesen 31, a második 73, a harmadik 66 versben van jelen, tehát lesznek olyanok, amelyekben mind a privatum, mind a publicum szavai előfordulnak. Érdekesebb azonban a második táblázat, amely azt vizsgálja, van-e összefüggés a fogalmak felbukkanása között (az értékeket az átláthatóbb eredmények érdekében négy tizedesjegyig adom meg, az R tíz tizedesig számítja ki).

| | magyarok | istennők | nemek |
|---|---|---|---|
| *Ajánlás* | 1 | 1 | 0 |
| *Küpriszhez* | 0 | 1 | 0 |
| *A melancholia* | 0 | 1 | 0 |
| *A szerelemhez* | 0 | 1 | 1 |
| *A jámborság és középszer* | 0 | 1 | 1 |

| | magyarok | istennők | nemek |
|---|---|---|---|
| magyarok | 1,0000 | -0,5886 | 0,1393 |
| istennők | -0,5886 | 1,0000 | -0,0277 |
| nemek | 0,1393 | -0,0277 | 1,0000 |

**1. táblázat**. A vizsgált fogalmak előfordulása

De mit jelent mindez? A korreláció értéke -1 és +1 között lehet. Ha az előbbi, akkor teljes a korreláció, de fordított, tehát ha az egyik érték nő, akkor a másik ugyanolyan mértékben csökken, ha az utóbbi, akkor mindig ugyanabba az irányba, ugyanolyan mértékben változnak. Minél közelebb van egy érték a nullához, annál kisebb a korreláció, nulla esetében egyáltalán nem lehet kimutatni a két fogalom előfordulása

[38] A vizsgálat megismételhetősége/ellenőrizhetősége érdekében a használt kódok kommentárokkal kísért listája az interneten is olvasható a http://labadigergely.github.io/szovegek/2017/02/17/R5/ címen. Az értékek kiszámításakor az R által használt képletek a program súgójából megismerhetők. R Core Team, *R: A Language and Environment for Statistical Computing. R Foundation for Statistical Computing* (Vienna, 2016), hozzáférés: 2018.05.23, `https://www.R-project.org/`.

között bármiféle összefüggést. A +/-0.5 és a +/-1 közötti érték esetében kifejezetten erős korrelációról szoktunk beszélni, a +/-0.1 és +/-0.3 közöttieknél van ugyan, de gyenge a korreláció – a Pearson-féle korrelációs koefficiens értelmezésében Jockers könyvét követem.[39]

A vizsgált fogalmak esetében a „magyarok" címszóval összefoglaltak és az „istennők" között ténylegesen kimutatható, és kifejezetten erős negatív kapcsolat van (-0,5886), tehát azokban a versekben, amelyekben az első csoport szavai/témái előfordulnak, a másodiké nem – és fordítva: a privatum és a publicum világa nem keveredik. Szintén kimutatható a kapcsolat a „magyarok" és a „nemek" között (0,1393), azaz a publicum világában a tagadás, a hiány jelen van, de ez az érték alacsony, a korreláció gyenge. A „nemek" és az „istennők" között viszont gyakorlatilag nincs semmiféle kapcsolat (-0,0277), azaz a privatumban a hiány nem meghatározó élmény – persze a keresett kifejezésektől, ismétlem, sok függ.[40]

Mindez így önmagában nem több, mint Onder kötetkoncepciójának számszerű igazolása, ami éppenséggel nem kevés, de tény, nyolcvanhét vers nem akkora adatmennyiség, amekkorát egy ember ne tudna feldolgozni. Innen több irányba is tovább lehet lépni: más fogalmakra keresve talán más koncepciót erősebben lehet igazolni/cáfolni. De talán izgalmasabb, látványosabb, ha az eredményt randomizáltan ellenőrizzük, azaz megnézzük, hogyha az egyes versekben előforduló kifejezések értékeit összekeverjük – tehát az első táblázat nulláit és egyeseit véletlenszerűen osztjuk ki a versek között –, miként alakulnak a korrelációs értékek, mert ekkor derül ki, hogy a tényleges eredmény mennyire egyedi, váratlan, vagy ha úgy tetszik, a szerző mennyire volt „tudatos." Az R által kínált lehetőségekkel élve természetesen nem nekünk kell újra és újra kiosztanunk a számokat, ezt elvégzi a gép. Miután a program tízezerszer véletlenszerűen kiosztotta, hogy melyik vers mely csoportba tartozzon, a következő eredményeket kapjuk az összefüggésekről – mivel korrelációt csak két érték között tudunk számolni ezért a következő diagram csak a publicum és a privatum lehetséges korrelációit mutatja be, de mint az utána következő táblázatból kiderül, természetesen mindegyik párosra kiszámoltattam az értékeket.

Bár a lényeg talán ebből is látszik, az alapvető statisztikai adatok kiírása egyértelműbbé, és könnyebben elemezhetővé teszi az eredményeket, ráadásul mindet tartalmazza:

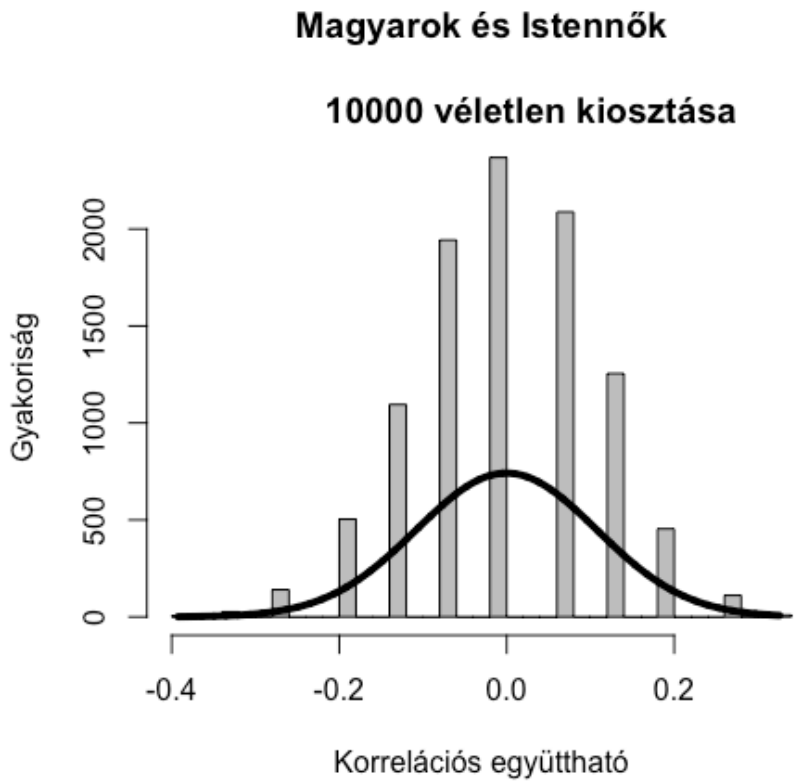
| | min. | max. | átlag | szórás | tényleges érték |
|---|---|---|---|---|---|
| magyarok–istennők | -0,3926 | 0,3258 | -0,0002 | 0,1076 | -0,5886 |
| magyarok–nemek | -0,3655 | 0,3636 | -0,0002 | 0,1077 | 0,1393 |
| nemek–istennők | -0,2470 | 0,4840 | 0,0003 | 0,1083 | -0,0277 |

**2. táblázat**. A korrelációs értékek számszerűsítve páronkénti bontásban

[39] Matthew L. Jockers, *Text Analysis with R for Students of Literature* (New York: Springer, 2014), https://doi.org/10.1007/978-3-319-03164-4, 49. A korrelációanalízis ötletét innen vettem, viszont a kódok jelentős részét én írtam, lévén Jockers más jellegű korpuszon és más módszertannal dolgozott: egy regény, a *Moby Dick* fejezeteit és két kifejezésének (*whale, ahab*) darabszám szerinti és százalékos előfordulását és összefüggését vizsgálta.

[40] Lehetséges, hogy ebben a kérdésben a számítógépes nyelvészetben használatos szentiment- és emócióelemzés részletesebb válaszokat tudna adni.

 DOI 10.31400/dh-hun.2018.1.126




**2. ábra**. A publicum és a privatum korrelációi

A keresett szavak előfordulásának tízezer véletlenszerű kiosztása esetén az átlagos korreláció minden esetben lényegében nulla, azaz nincs semmiféle összefüggés a szavak versekbeli előfordulása között. A szórás is nagyon alacsony, éppen eléri a 0,1-es határt, amelytől valamiféle gyenge kapcsolatról már lehet beszélni, mindez azt jelenti, hogy a tízezer eset túlnyomó részében az adatok nulla körül maradnak. Ehhez kapcsolódóan van egy statisztikai szabály is, a 68–95–99,7, amely azt jelzi, hogy a mért adatok 68%-a az átlagtól legfeljebb egy szórásnyira tér el, a 95%-uk az átlagtól kétszórásnyi távolságon belül helyezkedik el, 99,7%-uk pedig legfeljebb háromszórásnyi távolságban található az átlaghoz képest, így a szórásérték a tényleges adat váratlanságát, valószínűségét mutatja. A szélsőértékek persze azt is megmutatják, hogy ilyen alacsony esetszám (31, 73, 66) és kevés érték (0, 1) esetén nehéz olyan elosztást találni, amelynél igazán erős korrelációról beszélhetnénk. Igaz, a fenti táblázat szélsőértékei majd mindegyik esetben kívül esnek a szélsőértékek háromszorosán (egyedül a nemek–istennők minimumértéke van belül a háromszorosán), azaz tízezer eset háromtized százaléka (30), ha tényleges eredményként jelenne meg, váratlannak volna minősíthető. Éppen ezért fontos, hogy a „magyarok” és az „istennők” közötti tényleges adat még a tízezer randomizált kiosztás minimumértékén is jóval túl található, tehát a statisztikai előrejelzés szerint teljességgel váratlan és valószínűtlen a ténylegesen mért adat. Ami az egyébként már önmagában is erősként interpretálható korrelációt (-0,5886) még erősebbé teszi, azaz a privatum és a publicum – az általam használt fogalmakkal mért – koncepciója erős támogatást kap, hiszen a tényleges értéket

tízezer próbálkozásból egyszer sem sikerült előállítania a programnak. A „magyarok" és „nemek" között mért kapcsolat távol van a lehetséges szélsőértékektől, de belül a szórás kétszeresén, ami a körülmények ismeretében (alacsony érték- és esetszám) a mutatottnál (0,1393) talán valamivel erősebb összefüggést valószínűsít. A „nemek" és az „istennők" közötti -0,0277-es érték már eleve a kapcsolat teljes hiányát mutatta, amit a randomizálás is megerősített: bár a mért adat magasabb, mint az átlag, ám jócskán belül van az egyszeres szóráson. Figyelembe véve azonban, hogy a szélsőértékek itt is ugyanakkora intervallumot fognak át, mint a másik két esetben, tehát akár erősebb összefüggés is elképzelhető volna, a kapcsolat teljes hiányát igazoló tényleges adat megint a koncepciót erősíti – vagy a csoportosítás során jól megválasztott fogalmakat dicséri.

Természetesen további vizsgálatokat is el lehetne végezni. Ha magukon a fogalmakon nem is változtatunk, még akkor is értelmesnek tűnik az egyes könyveken belül kiszámolni az értékeket vagy visszalépni az időben és Berzsenyi 1808-as kötetkompozíciójával vetni össze az eredményeket – az induló kódkészlet most már mindenki rendelkezésére áll. A legkézenfekvőbb azonban, ha megnézzük, hogy a számítógépes algoritmusok milyen tematikus csoportokat találnak a nyolcvanhét versben, ezek igazolják-e a koncepciót. Egy módszertani megjegyzést kell azonban előre bocsátanom: a *topic modelling* hosszabb szövegek esetén megbízhatóbb. A kísérlet ugyanakkor érdekes eredményt adott még így is.

A tartalomelemzés során meg kell adni, hány témát keressen az R algoritmusa – én a *tm* csomagot használtam[41] –, értelemszerűen kettőt adtam meg, és nem a versekben ténylegesen előforduló szavak, hanem ezek szótári formái alapján keresett a program. Az első csoport első ötven szava: *szent, kéz, lát, égi, föld, néz, bölcs, dicső, isten, világ, nap, magyar, nép, ember, gyenge, vér, haza, lélek, örök, vad, boldog, virág, érez, harc, tesz, barát, víg, halál, ész, fény, róma, század, fej, fényes, kor, tenger, elme, kény, kevély, kincs, por, száll, tűz, zár, bér, gyönyörű, kebel, múzsa, pálya, szabad.* A sorrendet a gép állította fel, s minél előrébb van egy szó, annál jellemzőbb az adott témára. A második csoport első ötven szava: *szív, öröm, lélek, szem, bús, szerelem, édes, élet, kar, kép, megy, könny, szelíd, hív, láng, arany, idő, tündér, kedves, szeret, szűz, csendes, homály, kegyes, liget, mély, forró, jer, szerető, kebel, mosolyogva, tud, vár, zöld, magas, völgy, kér, mennyei, erő, gond, hall, int, lel, mosolyog, rózsa, tér, vesz, vidám, fed.*

Ezzel a módszerrel le kellett mondanom arról, hogy a publicum világában (első csoport) jelen lévő hiányt vizsgáljam, mivel a tagadószavakat mint tipikus funkciószavakat a *topic modelling* nem veszi figyelembe. A két listát összevetve látszik, hogy az eredeti koncepció publicum–privatum megoszlása érvényes, de a tartalmuk – legalábbis az általam fentebb adott összefoglalóhoz képest – megváltozott. A múzsaverseket, a tudomány dicséretét tartalmazó darabokat a publicum körébe osztja az algoritmus – ami abból a szempontból aligha meglepő, hogy a megénekelt kortárs személyiségek érdemei közt az ősi hadi erények mellett a művelődésben betöltött szerepüket is kiemelik az ódák –, a privatumban pedig a szerelmi költészet mellett hangsúlyosan jelennek meg a sztoikus életfilozófiát, életvilágot bemutató versek. Ha mindezt az egyes darabokra lebontva kívánjuk vizsgálni, akkor jóval árnyaltabb képet kapunk,

[41] Ingo Feinerer, Kurt Hornik and David Meyer, „Text Mining Infrastructure in R," *Journal of Statistical Software* 25, 5. sz. (2008): 1–54, `http://www.jstatsoft.org/v25/i05/`.

 DOI 10.31400/dh-hun.2018.1.126

mint a fogalomkeresős módszer esetében, hiszen a *topic modelling* kiindulópontja, hogy minden szöveg több témából áll. Azaz valamilyen mértékben minden versben megtalálható mindkét téma. Nézzük meg, hogy alakulnak az előző táblázat értékei a tartalomelemzés során:

| | publicum | privatum |
|---|---|---|
| *Ajánlás* | 59,67% | 40,32% |
| *Küpriszhez* | 42,67% | 57,33% |
| *A melancholia* | 37,78% | 62,22% |
| *A szerelemhez* | 38,46% | 61,54% |
| *A jámborság és középszer* | 49,65% | 50,35% |

**3. táblázat**. Az értékek alakulás a tartalomelemzés során

Az *Ajánlás*ban tehát az első téma dominál közel 60%-kal, a többi négyben viszont a második. Maguk az adatok tehát alátámasztják az előző táblázat 0-it és 1-eit, ha 50%-nál húzzuk meg a határt. *A jámborság és középszer* értékei azonban a gépi olvasás sajátos mechanizmusára is felhívják a figyelmet. Egy értő olvasó aligha sorolná (kerekítve) 50%-ban a publicum világába tartozónak a szöveget, a program számára azonban a kulcsszavak megléte a kérdés – a *„szent,"* a *„templom,"* az *„áldozópap,"* a *„rabiga"* valóban értelmesek a közösségi lét diskurzusának keretében. A vers értékei mindemellett talán Berzsenyi alkotómódszerére nézve is érdekesek, mennyire könnyen vándorolnak egyes képzetek, képek a különböző időpontban, műfajban és céllal született versek között. E vers mellett a *Küpriszhez* értékei talán még meglepőbbek: nehéz értő szemmel 43%-ban a publicum világába sorolni – persze például a *„bér,"* a *„tanít"* megint csak részesei a másik diskurzusnak, tehát nincs szó tévedésről. Inkább az a kérdés ezek után, hol húzzuk meg a határt, azaz hány százaléknál minősítsünk egy szöveget egyértelműen az egyik vagy a másik csoportba tartozónak. Ha 50%-nál, akkor az első témába 28, a másodikba 59 szöveg tartozik; 60% esetében 13, illetve 74. A publicum körébe tartozó szövegek közül legnagyobb értékkel *A pesti magyar társasághoz* rendelkezik, a számítógépes algoritmus szerint 75,47%-ban tartozik ide, s csak 24,53%-ban a privatumba. Ebben legnagyobb értékkel a *Lilihez* („Jer, Lili! nézd…") rendelkezik (78,16%).

### 2. 3. Magán- és mássalhangzók, szóhosszúság

A tanulmány befejező részében a „szép hanggal" kapcsolatos stilisztikai kérdésekre próbálok meg válaszokat adni. A korábban idézett Berzsenyi-passzus és a vele összekapcsolt Révai-részlet nyomán többféle szempont alapján is lehet a verseket vizsgálni. Az egyik a magán- és mássalhangzók aránya, mivel ezek „illő öszvemérsékeltetése szerzi főképen minden nyelvnek az ő kedves hangját."[42] A másik szintén könnyen mérhető tulajdonság a szavak hosszúsága, amely túlzott monotonitás esetén szintén „keménység"-et okoz Révai szerint.

A szövegeket értelemszerűen most a tényleges szóalakok alapján kell vizsgálni, ami még élesebben veti fel a szöveg-előkészítés nehezen megválaszolható módszertani kérdését, a helyesírás normalizálását, és óhatatlanul lesznek hibák. A program

[42] Révai, *A magyar szép*, 52.

alapvetően csak a latin betűket ismeri fel, be lehet ugyan állítani, hogy például a *th*-t, *ph*-t, vagy éppen az *ae*-t egy hangzóként értse, de az ilyen megoldások nem mindig elegendők. A *vaszár* ebben a korpuszban szerencsére nem fordul elő (csak acélból készült zárak, závárok), de egy szimpla gépi számolásnál a helyes mássalhangzószám a *vaszár* esetében négy volna, viszont ha beállítjuk az *sz*-et, akkor a program hármat fog találni. A korban ugyan az *s*-nek volt egy másik karaktere is *(ſ)*, amelyet Verseghy szerint arra kell használni, hogy az összetett mássalhangzókban *(sz, zs)* az *s*-et jelölve a szóösszetételekből adódó félreértéseket el lehessen kerülni, illetve különböző szabályokat lehessen tanítani a diákoknak kiejtésről, elválasztásról,[43] ám ezt az átiratok nem őrzik meg – és az egykorúak közül sem mindenki, nem minden szerző, kiadó, nyomda alkalmazta ennyire következetesen. A hosszú összetett mássalhangzókat viszont nem kódoltam külön, mert az értékük mindenképpen kettő lesz: a gép persze azért számolja annak az *lly* esetében, mert az *l*-et és az *ly*-et észleli, az ember pedig azért, mert tudja, hogy az *ly* duplázódik. Ezzel természetesen le kell mondani az egyes mássalhangzók ismétlődésének vizsgálatáról – legalábbis nekem. És akkor még nem beszéltünk arról, hogy az *y*-t hová számoljuk: bizonyos mássalhangzók mellett *ü*-ként vagy *i*-ként kell számolni, mások mellett pedig egy összetett mássalhangzó része. Ezt viszont legalább meg lehet oldani egy erre is ügyelő átírással. Mindezt szükséges még kiegészíteni azzal Jékel és Papp munkája nyomán,[44] hogy bár a „szép hanggal" kapcsolatos vizsgálódást emlegettem korábban, valójában fonémákat fogok vizsgálni, hiszen az írásban nem jelölt összeolvadásokat, részleges és teljes hasonulásokat nem vettem figyelembe – a különbség Jékelék ellenpróbái szerint elhanyagolható, mindössze 4 ezrelék[45] – a címeket azonban velük ellentétben igen.

A számolást ismét hamar elvégzi a gép, eredményét az alábbi táblázatban foglaltam össze. Az érdekesebb rész, mint mindig, utána következik.

| | Ajánlás | Első könyv | Második könyv | Harmadik könyv | Negyedik könyv |
|---|---|---|---|---|---|
| msh | 354 | 8296 | 12588 | 12707 | 9520 |
| mgh | 250 | 5643 | 8548 | 8686 | 6341 |
| átlag szóhossz | 5,2983 | 5,1722 | 5,1189 | 5,1711 | 4,8713 |
| msh/mgh | 1,4160 | 1,4701 | 1,4726 | 1,4629 | 1,5013 |
| min. msh | 0 | 0 | 0 | 0 | 0 |
| max. msh | 8 | 9 | 10 | 10 | 10 |
| átlag msh | 3,1053 | 3,0783 | 3,0487 | 3,0716 | 2,9238 |
| szórás msh | 1,1708 | 1,6408 | 1,6805 | 1,7120 | 1,7132 |
| min mgh | 0 | 0 | 0 | 0 | 0 |
| max mgh | 5 | 6 | 6 | 7 | 6 |
| átlag mgh | 2,1930 | 2,0939 | 2,0702 | 2,0996 | 1,9475 |
| szórás mgh | 1,1279 | 1,1046 | 1,1021 | 1,1259 | 1,1041 |

**4. táblázat**. A fonémavizsgálat eredménye

[43] Verseghy Ferenc, *Magyar grammatika avvagy Nyelvtudomány* (Buda: Egyetemi Nyomda, 1818), 56.
[44] Jékel Pál és Papp Ferenc, *Ady Endre összes költői műveinek fonémastatisztikája* (Budapest: Akadémiai Kiadó, 1974).
[45] Jékel és Papp, *Ady Endre összes,* x.

 DOI 10.31400/dh-hun.2018.1.126

A táblázat sorai közül talán egyedül az átlagos szóhossz szorul magyarázatra: összeadja az adott szakasz mással- és magánhangzóinak számát, majd elosztja a szavak számával. A szavak átlagos hosszai nagyon közel esnek egymáshoz, ami a mással- és magánhangzók aránya kiegyenlítettségének fényében persze nem meglepő. Kérdés, ez mennyire az egyéni nyelvhasználat jellemzője, vagy mennyire befolyásolhatja maga a mérés. Az utóbbit kizárhatjuk, mert a keresés során harminc mássalhangzót adtam meg és csak 16 magánhangzót,[46] amelyek aránya 1,875, ami jóval magasabb, mint a mért 1,47-es, 1,5-ös érték. Mielőtt kontrollanyaggal vetném össze, a táblázat értelmezéséhez érdemes még hozzátenni, hogy az egyes hangzócsoportoknak közelebb esik az átlaga a legalacsonyabb szélsőértékhez, és az ezzel párosuló magas szórási érték azt jelzi, hogy kevés a kifejezetten rövid szó – amit Révai expliciten tilt is –, viszont ha megnézzük, a szórás háromszoros értéke közel van a maximumhoz, de sosem éri el. Azaz a szavak meglehetős változatosságot mutatnak a szóhosszúságot tekintve, és az extrémhosszú, tíz mássalhangzót vagy hét magánhangzót tartalmazó szavak, legfeljebb a szöveg 0,3%-át teszik ki – a túlzottan hosszú szavaktól is óv Révai. Tehát a versek a Révai-féle jólhangzás követelményeinek (változatos szóhossz, magán- és mássalhangzók arányos eloszlása) megfelelnek.

Kontrollként megvizsgálva egy Berzsenyi-levél, valamint -értekezésrészlet, illetve néhány Virág-vers és -értekezésrészlet szövegét,[47] a következő táblázatot kapjuk:

| | Berzsenyi-értekezés | Berzsenyi-levél | Virág-értekezés | Virág-versek |
|---|---|---|---|---|
| msh | 1435 | 1229 | 2904 | 994 |
| mgh | 961 | 2025 | 2025 | 671 |
| átlag szóhossz | 5,3363 | 5,6590 | 5,6590 | 4,9554 |
| msh/mgh | 1,4932 | 1,4341 | 1,4341 | 1,4814 |
| min. msh | 0 | 0 | 0 | 0 |
| max. msh | 9 | 10 | 10 | 9 |
| átlag msh | 3,1960 | 3,1035 | 3,3341 | 2,9583 |
| szórás msh | 1,8915 | 1,8751 | 2,0633 | 1,7151 |
| min mgh | 0 | 0 | 0 | 0 |
| max mgh | 6 | 6 | 6 | 5 |
| átlag mgh | 2,1403 | 2,1035 | 2,3250 | 1,9970 |
| szórás mgh | 1,1497 | 1,2657 | 1,2789 | 1,0968 |

**5. táblázat**. Berzsenyi- és Virág-szövegek fonémavizsgálatának eredménye

Szembetűnő, hogy a vizsgált Virág-versek adatai egy kivétellel teljesen összhangban vannak a Berzsenyi-kötet adataival. A magánhangzók szórásának háromszoros értéke

[46] A hangzók száma a *th-*, *ae*-típusú kiegészítések miatt nőtt meg. A függelékben olvasható kódban valamennyi látszik.

[47] A levél kivételével MEK-en elérhető szöveget vettem alapul. Az értekezésrészletek: Berzsenyi Dániel: *A magyarországi mezei szorgalom némely akadályairul;* Virág Benedek: *Magyar századok.* A Virág-versek a következők: *Gróf Forgács Miklós emlékezete, Bosszúállás, Egy jeles poétára.* A levél pedig Berzsenyi fent idézett levele Kazinczynak 1811 nyaráról a kritikai kiadás alapján. A feldolgozás a versekkel megegyező módon történt.

az átlaghoz adva eléri a maximumértéket, míg a Berzsenyi-verseknél ez egyik esetben sem történt meg. Azaz a Berzsenyi-versek szóhosszai valamivel változatosabbak. Virág prózai szövegeinek magánhangzószórása szintén ugyanezt mutatja (a háromszoros értéke az átlaghoz hozzáadva meghaladja a magánhangzók szélsőértékét), de Berzsenyié már nem. Érdekes továbbá, hogy a prózai szövegek esetében az átlagos szóhossz valamivel nagyobb, mint a versek esetében, Virág szövegénél ez elég markáns különbség, a másik két esetben legfeljebb tendenciáról beszélhetünk – elképzelhető, hogy a témaválasztás volt hatással erre az értékre, ezt a későbbiekben érdemes megvizsgálni. A mással- és magánhangzók arányában azonban nem látszik különbség. Egy nagyobb kontrollanyag alapján a konklúzió is erőteljesebb lehetne, de mindezek után úgy tűnik, Berzsenyi a jólhangzás követelményének jegyében mind prózában, mind versben odafigyelt a szóhosszúság változatosságára, a mással- és magánhangzók egyenletes eloszlása azonban sokkal inkább nyelvi sajátságnak tűnik, semmint a szerzői nyelvhasználat jellemzőjének.

## 3. Következtetések

A tanulmányból kiderült, hogy a gépi olvasás mechanikus voltának megvannak a maga gyenge pontjai – emlékezzünk például az *lly* értékének kérdésére –, de még egy ilyen kis korpusz esetén is látszik a potenciálja. A versek gépi, tematikus csoportosítása után felmerült, hogy a vizsgált koncepció tartalmát esetleg érdemes módosítani, hogy a művelődés problémái, illetve a sztoikus világlátás kérdései hangsúlyosabban jelenjenek meg a Berzsenyi-versek értelmezésében. De mindkét vizsgálat egyértelművé tette: amíg a magyar írott örökség kapcsán nem kezdődik meg egy szisztematikus, egységes átírási, modernizálási elvek alapján történő, nyílt hozzáférésű digitalizálás, addig az ellenőrizhetőség, megismételhetőség alapvető tudományos kritériumainak sok vizsgálat nem fog tudni megfelelni, a *distant reading* módszerei közül sokat nem lehet biztonsággal használni.

## The Reading Machine
## A Distant Reading of Dániel Berzsenyi's Poems

This experimental study is an attempt to explore and test whether, by using current computational linguistics tools for the Hungarian language, it is possible to confirm hypotheses concerning the classification of Dániel Berzsenyi's (1776–1836) poems drawing on their close reading. The paper also investigates if we can answer questions related to the phoneme theory and practice in Berzsenyi's poems, which obviously cannot be examined by traditional close reading methods.

Keywords:
Dániel Berzsenyi, R, Magyarlanc, topic modelling, distant reading

 DOI 10.31400/dh-hun.2018.1.126