# SZÖVEG AZ OLVASÓNAK: KVANTITATÍV MÓDSZEREK ÉS DIGITÁLIS SHAKESPEARE-FILOLÓGIA

Almási Zsolt

PhD, tanszékvezető egyetemi docens,
Pázmány Péter Katolikus Egyetem Angol Nyelvű Irodalmak és Kultúrák Tanszéke
almasi.zsolt@btk.ppke.hu

A számítógépes technológia és módszertan az irodalomtudományban várakozásainkkal ellentétben nem oldja meg a filológiai problémákat, hanem inkább radikalizálja, és még világosabban felszínre hozza őket. Shakespeare-kutatóként a Shakespeare-filológia problémáira reflektálok, de reményem az, hogy ezeket az észrevételeket más filológiai kontextusban is adaptálni lehet. A Shakespeare-filológiában a számítógépes módszerek alkalmazásáról, a kvantitatív módszereken alapuló elemzésekről, a nagyobb szövegkorpuszokon alapuló ún. *távolvasásról (distant reading)* jelentősebb mennyiségű publikáció a 21. században látott napvilágot. Ezen cikkek és monográfiák meglehetősen nagy vitákat kavartak a szövegek irodalmisága, esztétikuma tekintetében, másfelől pedig módszertani, irodalomelméleti, matematikai, statisztikaelméleti kérdéseket is felvetettek. Ebbe a vitába kapcsolódik ez a tanulmány is – de más megközelítésben. A kvantitatív elemzéseket közlő cikkek eredményeit nagyon nehéz ellenőrizni, hiszen számos esetben nem világos, pontosan milyen szövegkorpuszra alapul a kutatás, az alkalmazott számítógépes módszerek, szoftverek miként működnek, milyen nyelven íródott a szoftver, és ennek a nyelvnek milyen előfeltételezéseit fogadta el a program írója. Mindezen megfontolások a Shakespeare-filológia számára fontosak, hiszen eleve egy meglehetősen képlékeny, bizonytalan szövegkorpusszal dolgozik a filológus, és ezt a bizonytalanságot tovább fokozza a számítógépes módszertan. Mivel a hagyományos publikációk keretei nem adnak lehetőséget mindezen bizonytalanságok kiküszöbölésére, ezért javaslatot szeretnék tenni egy újfajta publikációs módozatra, amely lehetőséget teremt a filológusközösségnek, hogy a publikációk eredményei ellenőrizhetőek, ismételhetőek legyenek.

A Shakespeare-kutatásban a kvantitatív módszereket stilisztikai elemzésekhez, szerzőségi kérdések *(attribution studies)* felvetéséhez, valamint műfajok vizsgálatához *(genre studies)* használják leginkább. Az utóbbira talán legjobb példa Michael Witmore és Jonathan Hope: ők „iteratív, azaz ismétlődésen alapuló irodalomtudománynak" *(„iterative criticism")* nevezik tudományos kutatásukat (Hope – Witmore, 2010). Ennek az a célja, hogy létrehozzanak „az angol szavak címkézéséhez egy olyan fenomenológiai alapú architektúrát – lényegileg szóhalmazok vagy szótárak gyűjteményét, mint például a *DocuScope* program –, amely a műfaji jellegzetességeket a mondatok szintjén mutatja meg" (Hope – Witmore, 2010, 360–361). Bizonyos nyelvi jellemzők, a szavak címkéinek és azok gyakoriságának alapján a *DocuScope* program segítségével csoportokba sorolják Shakespeare műveit, és ezek a csoportok szinte meglepő módon megegyeznek a művek hagyományos műfaji kategorizálásával. Ezek az eredmények azért jelentősek, mert így kiderül, hogy a műfaji besorolásnak nemcsak a cselekmény, a karakterek viselkedése szolgál alapul, hanem hasonló kézzelfogható módon a textualitás, a szavak csoportosítása és gyakorisága is.

A szerzőségi kérdések tekintetében az egyik legutóbbi könyv Hugh Craig és Arthur F. Kinney tollából és szerkesztésében jelent meg *Shakespeare, Computers, and the Mystery of Authorship* címmel (Craig – Kinney, 2009). Az egyértelműen Shakespeare-nek tulajdonított művekből létrehoztak egy korpuszt, és kétezer szavas szekvenciákra bontott szegmenseket vizsgálva kvantitatív módszerrel megállapították, hogy milyen lexikális jellemzők írják le ezt az anyagot. Ezzel a módszerrel a lexikális szavak gyakoriságának alapján létrehozták Shakespeare ún. *markerszavainak* ötszázas listáját. Hasonló módon elkészítettek egy olyan korpuszt is, amely bizonyosan más kortárs szerzők műveit tartalmazza, és itt is felállítottak egy ötszáz szavas listát kétezer szavas szekvenciák alapján. A szerzőségükben kétséges műveket vagy azok egyes kérdéses részeit elemezték ugyanezzel a módszerrel, majd megvizsgálták a proximitásukat a két korpuszhoz képest (Craig – Kinney, 2009). Az így kapott eredményeket aztán összevetették egy olyan elemzéssel, ahol ugyanezt a módszert követték, ám ebben az esetben nem a lexikális, hanem a nyelvtani szavakból állították fel a két markerlistát.

Önmagában véve nem forradalmi módszerekről és eredményekről van szó, hanem inkább arról, hogy ezek az eljárások egy szövegkorpusz-elemző hagyományba, a Shakespeare-kutatásban meglévő hagyományba illeszkednek. Ami nóvum, és ami előrelépést jelent, az a módszertan finomhangolásában rejlik, kihasználva a számítógép erejét, azaz a számoláson alapuló műveleteket, amelyek a számítógépet fáradhatatlansága és pontossága révén verhetetlenné teszik az emberi olvasóval történő összehasonlításban. Az említett szerzőpárosok – eredményeik ismertetése során – soha nem azt állítják, hogy a régi problémák megoldása a hagyomány és a korábbi kutatások figyelembe vétele nélkül történne. Úgy vélik, hogy bár a kapott vizsgálati eredmények, a statisztikai módszertan egyáltalán nem vezet meglepő és forradalmi eredményekhez, de a matematizálható tényekkel bizonyos értelmezési irányokat meg tudnak jelölni a műfaji sajátságok vagy a szerzőségi kérdések tekintetében. Mindezen megfontolások mellett azonban mindkét szerzőpárost meglehetősen sok kritika érte. Hugh és Craig módszertanát a transzparencia hiányával vádolta Peter Kirwan (Kirwan, 2010), Brian Vickers pedig a szavakon és nem a szókapcsolatokon alapuló metodikát kritizálta (Vickers, 2011). Alan Galey (2010) Hope-ot és Witmore-t bírálta az elemzett szövegek filológiai minősége miatt. A kritikák ellenére a két szerzőpáros meghatározó a matematizálható tényeken alapuló Shakespeare-kutatások terén.

A matematizálható tények kérdését, azaz a szógyakoriság problémáját tárom fel a Shakespeare-filológia tükrében. A szógyakoriság vizsgálatakor nagyon fontos szempont,

hogy milyen szövegkiadást használ a kutató, hiszen a szövegben található szavakat számoltatjuk a számítógéppel. A Shakespeare-kutatásban régóta tudott, hogy bármennyire szeretnénk, sajnos nincs olyan színmű, amely esetében rendelkeznénk egy végső kézirattal, vagy egy, a szerző által jóváhagyott, nyomtatásban megjelent verzióval. Kézirat nem maradt fenn, a korai nyomtatványok még Shakespeare életében és közvetlenül utána egyes színművek esetében eltéréseket, néha nagyon jelentős különbségeket mutatnak. Ennek oka a korabeli kulturális hangulatban, a tulajdonjogi problémákban keresendő, és így a szöveget, a nyomdába szánt művet sem előtte, sem a kiadáskor nem vették féltő gonddal körül, aminek egyik kényelmetlen következménye, hogy nem hagyományozódott ránk egy végső változat. Amikor a 18. századtól a Shakespeare-i szövegeket elkezdték gondozni, azt is egyéni ízlésbeli megfontolásokkal fűszerezték a kor elvárásainak megfelelően. Azaz mára a színműveket tekintve olyan sok szövegvariáns áll a rendelkezésünkre, hogy egy kritikai kiadás készítője filológus legyen a talpán, ha azt a célt tűzi ki, hogy egy olyan szöveget állít elő a hagyomány alapján, amely minden olvasói elvárásnak megfelel.

A Shakespeare-szövegek hosszú története és ebből következő instabilitása annak ellenére, hogy közismert tény volt, a 20. század végéig valójában nem vált kulcskérdéssé, hiszen mindig készültek kritikai kiadások, amelyek valamiféle végső szövegváltozattal kecsegtettek – lehetett *Hamletre* utalni, maximum azt tettük hozzá, hogy most Harold Jenkins *Hamlet*-kiadását használjuk. A 21. századi digitális technológiának köszönhetően ma nagyon könnyen beláthatjuk az instabilitás meglétét, hiszen egy-két kattintással megtekinthetjük a korai nyomtatványokat s különféle adatbázisok segítségével a későbbi szerkesztett szövegek majdnem teljes történetét. Éppen ezért a szövegek instabilitása nem megoldásra váró probléma, nemcsak tudományosan elfogadott tény, hanem, ahogy Jowett állítja, „gondolkodásmód" (Jowett, 2009). Az, hogy milyen szövegváltozatot elemeztetünk a számítógéppel, azonban mindenképpen hatással lesz arra, hogy milyen eredményeket ad a statisztikai analízis.

Térjünk azonban vissza a matematizálható tények problémájához, és vegyük szemügyre a szavak számlálását egy példa segítségével. Érdemes kísérletet tenni a legegyszerűbb számolási adatokkal a *Sok hűhó semmiért* című komédiát alapul véve. Ha a nagyon kifinomult és alapos *WolframAlpha* (URL1) keresőfelületet használva kérdezünk le statisztikai adatokat a komédiáról, akkor sokat tanulhatunk a darabról, annak világáról a számok tükrében is. Megtudhatjuk például, hogy Dogberry jóval többet beszél, mint Hero, és hogy Beatrice jóval kevesebbet, mint Benedek, valamint, hogy hány szóból állnak az egyes jelenetek. Kiderül, melyek a leggyakrabban használt szavak, melyik a leghosszabb szó a műben, továbbá arra is fény derül, hogy az egész mű 21 183 szóból áll.

Ha azonban a *WordHoard* (URL2) nevű alkalmazást hívjuk segítségül, némileg eltérő adatokat kapunk. Az alkalmazás letölthető, az eredményeket könnyedén el lehet menteni a saját számítógépen. Ezt az eszközt irodalmi korpuszok elemzésére tervezték, és online hozzáfér a felcímkézett szövegkorpuszokhoz. Bár magukat a szövegeket nem lehet megtekinteni, a dokumentációból kiderül, hogy a híres Moby Shakespeare-szövegváltozatnak egy szerkesztett változatáról van szó, amely az egyik legjobb 19. századi kritikai kiadáson alapul. A *WordHoard* az alábbi szofisztikált





szempontrendszer alapján elemzi a szöveget: szógyakoriság, kollokációk, szófajok, a beszélő neme, beszélő halandósága, vers, metrikus alakzat. Ha ebben az alkalmazásban tekintjük meg a szavak számát, akkor az eredmény 20 910 lesz.

Az eltéréseket látva egy saját készítésű egyszerű szövegelemző programmal (URL3) is megszámoltattam a szavakat. Az elemzés karakterszámlálást, szószámot, a leggyakrabban használt tíz szót és a legritkábban, azaz a műben egyetlen egyszer előforduló szavakat, illetve a kötőjellel írt összetett szavakat listázza. A szövegelemző szkript Python nyelven készült, és bizonyos jellemzőket adottságoknak vesz. A szó valójában sztringet, üres karakterek között elhelyezkedő karaktersort jelent, ahol a karakterek szigorú bináris oppozícióban tételeződnek. A bináris oppozíció igen-nem szigorúságában ugyanannak a betűnek a nagy és kisbetűs változata külön karakternek, a sorvégi törés, *whitespace* ugyancsak karakternek számít.

A *Sok hűhó semmiért* első kvartókiadása alapján készítettem egy elemezhető szövegváltozatot. A szövegben a korabeli standardizálatlan helyesírás miatt ugyanaz a szó több betűsorként is megjelenik, de ez nem számít a szószámolásnál, csak a gyakoriságot torzítja, ám a jelen elemzés szempontjából ez a torzítás nem releváns. További probléma, hogy a sortörést a koramodern nyomdai szedő sokszor nem tudta a szóvégre pozicionálni, hanem kénytelen volt elválasztani a hosszabb szavakat. Az elválasztásnak azonban az lett a következménye, hogy az elválasztott szavak külön szavaknak, sztringeknek látszanak a gép számára. Természetesen az elválasztójelet és a sortörést is egyszerűen el lehetne távolítani a szövegből mechanikusan, ez azonban a sorszámok felborulásához vezetne. Itt sokkal egyszerűbbnek látszott kézzel eltávolítani őket a szövegből, és az elválasztás, illetve a sorhossz alapján vagy az adott, vagy a következő sorhoz csatolni az egyesített szavakat. Azért sem lehetett volna mechanikusan kitörölni a kötőjeleket, mivel bizonyos esetekben a szedő a túl hosszúnak ítélt szavakat sorközi helyzetben is kötőjellel választotta el. Ezeket a kötőjeleket nem lett volna érdemes eltávolítani. A standardizálatlanság és a sorközi kötőjelek kitörlése a szöveg történetiségének bizonyos rétegeit fedné el, tehát ez a fajta egységesítés és modernizálás nem célszerű minden esetben.

Ideális helyzetben a program megírása és a számítógéppel elemeztetendő szöveg megfelelő formátummá alakítása egyetlen ember feladata, vagy pedig egy programozó és egy filológus együttműködésén alapul. A szövegelemző alkalmazás írójának ismernie kell az elemzendő szöveg egyedi sajátosságait, hiszen ami egy szoftverfejlesztő számára adottságnak tűnik, az a filológus számára nem: minden korszak, szerző, szöveg más és más problémák elé állítja a filológust. Ugyanakkor a filológusnak is ismernie kell az adott alkalmazás jellemzőit, hiszen ennek az alkalmazásnak, a szkriptnyelv előfeltételezéseinek ismeretében lehet csak előkészíteni az elemzendő szöveget, hogy érvényes eredményeket adhasson az elemzés.

Az általam írt program szerint 22 171 szó található a *Sok hűhó semmiért*-ben, ami új eredmény az előzőekhez képest. A számbeli különbségeket az is indokolhatja, hogy az előzőektől eltérő szövegeket elemeztettem. A *WolframAlpha* esetében semmilyen információnk sincs az elemzett szövegről, bár a kvartókiadás bibliográfiai adatai jelennek meg a korabeli kiadás címlapjával együtt, ám a statisztikai adatok egy része nem ennek a szöveg-

változatnak az irányába mutat. Ilyen például a felvonások és jelenetek szószámát illető adatok felsorolása, hiszen a kvartókiadás a felvonások és jelenetek felosztását nem tartalmazta. A WordHoard esetében azt tudjuk, hogy egy sokáig közismert, de mégiscsak egy 19. századi szövegkiadás némileg szerkesztett, modernizált változatával van dolgunk. Én a számítógépes olvasatra alkalmazott kvartókiadás szövegével dolgoztam.

A különböző szövegek elemzéséből adódó eltéréseket azzal kerülhetjük el, ha azonos szöveget olvastatunk a számítógéppel: tehát feltölthetjük az általam alkalmazott szövegvariánst egy független szövegelemző alkalmazásba, a *Voyant Tools*-ba (URL4). A *Voyant Tools* olyan nyílt hozzáférésű online szövegelemző eszköz, amelynek segítségével szófelhő készíthető a megfelelő formátumban feltöltött szövegről a szógyakoriság alapján, valamint statisztikai adatokat tudhatunk meg a szó- és kollokációgyakoriságról, a keresett szavak előfordulásáról a szövegben elfoglalt helyük szerint. Ebbe az alkalmazásba betöltve a szövegverziómat ismét újabb eredményt kapunk: 22 162 szót. Ez utóbbinál tehát valószínűleg nem a szövegből adódó, hanem a kódban rejlő különbségek számítanak, például, hogy mit tart az alkalmazás szónak és mit nem, hogyan dolgozza fel a kis- és nagybetűket, a kötőjellel írt szavakat, a számokat.

Amellett is érvelhetünk, hogy ezek a minimum kilenc-, maximum ezerszavas különbségek az egyes alkalmazások végeredményei között nem relevánsak. Statisztikailag az ilyen mértékű eltérések nem számottevőek, hiszen hozzávetőlegesen 21 ezer szavas szövegről van szó. A Shakespeare-kritika 2014 óta használja a *good enough text* (Rowe, 2014) fogalmát, ami arra utal, hogy egy hozzávetőlegesen megbízható szöveg a kutatás szempontjából, ha nem is tökéletes, de elfogadható – különösen nagy korpuszok vizsgálata esetén. A good enough text analógiájára az adott alkalmazást *good enough application*-nek nevezhetjük. A szöveg is és az alkalmazás is megfelelő egyfajta kutatás szempontjából, különösen, ha nagy korpuszt elemeztetünk a számítógéppel. A *good enough text* és a *good enough application,* azt gondolom, nem megoldandó probléma, mert megoldhatatlan, mint a szövegek pluralitása – ezért gondolkodásmóddá kell válnia.

Mi következik tehát mindebből? A számítógép kérlelhetetlen alapossága felszínre hoz olyan problémákat, amelyekkel a filológusnak eddig nem kellett feltétlenül számolnia. Amint számokat látunk, az egzaktság hite keríthet minket hatalmába. Ám amint a színfalak mögé tekintünk, kiderül, hogy meglehetősen sok bizonytalansági tényező alakítja az eredményeket. Shakespeare esetében először is a szövegváltozatok sokaságával kell számolni. Másodszor az alkalmazás is bizonytalansági tényező: ami érvényesen és jól működik az egyik szövegnél, az téves eredményekre vezethet egy másiknál – hiszen minden szöveg egyedi. Továbbá kiderült az is, hogy a szkriptnyelv is nyelv, amely bizonyos előfeltételezésekkel él, és amit számításba kell venni, ha egzakt eredményekre törekszünk. A szöveget nemcsak az emberi olvasónak kell elkészíteni, fogyaszthatóvá tenni, hanem a számítógép számára is. Sőt az is világossá vált, hogy amikor a gépi olvasásra előkészítünk egy szöveget, akkor bizonyos célra, bizonyos felhasználásra készítjük a szöveget, és ami egyfajta célnak megfelel, esetleg egy másiknak nem. Ez a néhány bizonytalansági tényező kiiktathatatlan, ezekkel együtt kell élni, ennek gondolkodásmóddá kell válnia. Ha ezt elfogadjuk, akkor az a fajta hagyományos publikáció, amelyhez hozzászoktunk, nem lesz





alkalmas az információ közlésére. Nem elég az eredményeket összefoglalni egy cikkben, hiszen a számítógépes szövegelemzés, adatkinyerés meglehetősen bizonytalan lábakon áll a tudós közösség szempontjából, mivel egy hagyományos publikációban nem ellenőrizhetőek az eredményhez vezető módszerek és adatok.

Ebben a helyzetben az tűnik megfelelő megoldásnak, hogy hozzáférhetővé, sőt nyílt hozzáférésűvé kell tenni az eredményekhez vezető módszereket és adatokat is. A hagyományos publikáció mellett az elemzett szöveget és a kódot a megfelelő licencekkel mindenki számára elérhetővé és ellenőrizhetővé kell tenni. Ideális esetben olyan kiadásra gondolok, mint a multimédiás kiadvány vagy az értéknövelt kiadás (Karen van Godtsenhoven, 2009). Ám ezek egyelőre tervek. Ennek a hozzáférhetőségnek egy másik lehetséges módja a repozitórium használata, ahova fel lehet tölteni a szöveget a megfelelő metaadatokkal, a szerző nevével, a készítés dátumával, a nyílt forráskódú programmal, a programozó nevével. Mivel erre a kiadók nem készültek fel, sem az egyetemek nem biztosítanak ilyen repozitóriumokat, más megoldást kell találni.

Egyelőre a *GitHub* (URL5) nevű ingyenes szolgáltatás tűnik erre a legmegfelelőbbnek, hiszen ide munkaanyagok és metaadatok is egyaránt feltölthetők, és bárki számára hozzáférhetők. Alkalmat biztosít a közös kutatásra, megjelenítve, hogy melyik felhasználó mivel járult hozzá a munkához; továbbá a feltöltött fájlokat lemásolhatja, továbbfejlesztheti, sőt az eredeti tulajdonosa követheti is feltöltéseinek útját. Ez természetesen átmeneti megoldás, hiszen a repozitóriumot valójában a kiadóknak, esetleg a felsőoktatási intézményeknek kellene biztosítaniuk, ám amíg ez nem történik meg, a *GitHub* jó barátja lehet a számítógépet az irodalomtudomány gazdagítására használó kutatóknak.

Kulcsszavak: *Shakespeare, digitális filológia, kvantitatív módszerek, digitális publikáció, Sok hűhó semmiért, digitális szövegelemzés, digitális repozitórium*

**IRODALOM**


Craig, Hugh – Kinney, Arthur F. (eds.) (2009): *Shakespeare, Computers, and the Mystery of Authorship*. Cambridge University Press, New York

Galey, Alan (2010): Comments. alangaley 4 May 2010 at 12.05 am. • http://tinyurl.com/jxqquh5

Godtsenhoven, Karen van et al. (2009): *Emerging Standards for Enhanced Publications and Repository Technology. Survey on Technology*. Amsterdam University Press, Amsterdam 15–18. • http://dare.uva.nl/cgi/arno/show.cgi?fid=150752

Hope, Jonathan – Witmore, Michael (2010) „The Hundredth Psalm to the Tune of "Green Sleeves": Digital Approaches to Shakespeare's Language of Genre". *Shakespeare Quarterly.* 61, 3, 357–390. DOI: 10.1353/shq.2010.0002

Jowett, John (2007): *Shakespeare and Text*. Oxford University Press, Oxford-New York

Kirwan, Peter (2010): Review of Shakespeare, Computers, and the Mystery of Authorship, ed. Hugh Craig and Arthur F. Kinney. *Early Theatre.* 13, 1, DOI: 10.12745/et.13.1.824 • https://earlytheatre.org/earlytheatre/article/view/824/887

Rowe, Katherine (2014): Living with Digital Incunables, or a 'good enough' Shakespeare Text. In: Carson, Christie – Kirwan, Peter (eds.): *Shakespeare and the Digital World*. Cambridge University Press, UK

Vickers, Brian (2011): Shakespeare and Authorsip Studies in the Twenty-first Century. *Shakespeare Quarterly.* 62, 1, 106–142. DOI: 10.1353/shq.2011.0004

URL1: WolframAlpha • http://tinyurl.com/hqkwuvv

URL2: WordHoard • http://tinyurl.com/gycnd

URL3: http://tinyurl.com/zmxbkjy

URL4: Voyant Tools • http://voyant-tools.org/

URL5: GitHub • https://github.com/