Jan Rybicki 0000-0003-2504-9372

Uniwersytet Jagielloński w Krakowie

Instytutu Filologii Angielskiej

jkrybicki@gmail.com

Vive la différence! Írók nemének azonosítása többváltozós szógyakorisági elemzések során*

A kutatás első fázisa a szógyakoriságok többváltozós elemzését a szerzők nemének azonosítására használta egy 18. századi és 19. század eleji szentimentális és gótikus regényeket tartalmazó korpuszban. Ennek érdekében kerültek összehasonlításra a különböző nemekre vonatkozó leggyakoribb és a Burrows Zeta-módszerével létrehozott közepes gyakoriságú szavak listája. A kutatás második fázisában a már két korszakból származó (18–19. és 19–20. század) kifejezések összehasonlító elemzése szintén arra kereste a választ, hogy mennyire használhatók ezek a szerzők nemének meghatározásakor.

Kulcsszavak:
szerzőazonosítás, szerzői nem, Bootstrap Consensus Network, Zeta-módszer

Egek, megint a hűtőbe tetted a mogyoróvajat!
Basszus, már megint a hűtőbe tetted a mogyoróvajat!
(George Lakoff: *Language and Woman's Place*)

1. Bevezető

Nemrég megbíztak – ha ez a jó szó rá –, hogy vizsgáljak meg egy 18. századi, zömében angol női szerzők által írt, de pár ismeretlen szerzőségű regényt is tartalmazó korpuszt, hátha a névtelen munkák között felfedezhető férfi írók kézjegye is; illetve nézzem meg, hogy ugyanebben a korpuszban az alig ismert nők által írt regények miben különböznek a korszak sokkal híresebb női szerzőinek szövegeitől. A megbízás a „Női írók a történelemben: az európai irodalmi kultúra új megértése felé" (2009–13) című COST Action projekttől érkezett, amelyet Prof. Suzan van Dijk vezetett, és akinek hálával ajánlom ezt a tanulmányt. A bemutatott kutatás egy része korábbi munkáimból származik, amelyeket két COST Action konferencián mutattam be a következő

* Eredeti megjelenés: Jan Rybicki, „*Vive la différence:* Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies," *Digital Scholarship in the Humanities* 31, 4. sz. (2016): 746–761, https://doi.org/10.1093/llc/fqv023.

címeken: *Transzkulturális, transznacionális és transzdiszciplináris perspektívák a női irodalomtörténetben* (Poznan, 2012. november 26–28.), valamint *Európai női szerzőség: hálózatok és akadályok* (Hága, 2013. június 19–21.).

Bármilyen diskurzus a férfi és nő közötti különbség természetéről – és egyáltalán, a létezéséről is – rizikós vállalkozás; sőt jelen esetben, ugyan különböző okokból, de már a „férfi és nő" kifejezéssel szemben is ellenérzéseim vannak, amelyet a cikk eredeti nyelve kényszerít rám. Az angolban, ha ez nem egy bevett kifejezés volna (de az), ellenkezőleg kellene lennie, ahogy a lengyelben szinte mindig: „kobiety i mezczyz´ni", azaz 'nő és férfi'. Történetesen ez a lengyel címe Françoise Giroud és Bernard-Henri Lévy híres könyvének is: *A férfi és a nő (Les Hommes et les femmes),*[1] amelynek lengyel fordítója, Kalina Szymanowsky, hasonlóan érezhetett, mint én most. Persze ezen túl is számos oka van, hogy igyekszem elkerülni minden ideológiai előfeltevést ebben a tanulmányban – ehelyett inkább a férfias empirizmusra hagyatkozom: először a kísérletezés, aztán az eredmény (ha van egyáltalán), és harmadjára annak megvitatása.

Természetesen ez is értelmezhető ideológiai döntésként, ugyanakkor logikusnak tűnik, hogy ilyen távolságtartással mutassuk be, a többváltozós szógyakorisági vizsgálat mennyiben képes a nemiség meghatározására, amely a műfajjal, a kronológiával vagy a témával együtt oly sokszor befolyásolják a stilometriai kutatások eredményét a szerzőazonosítás során. Úgy gondolom, hogy a nemiségre vonatkozó nyelvi jegyek megkülönböztetése és azonosítása valóban komoly kihívás lesz a számítógépes stilisztikának a közeljövőben, és sok munka van még e téren, annak ellenére, hogy Matt Jockers már egy egész fejezetet szentelt a problémának népszerű *Macroanalysis* című könyvében.[2]

Mivel nem igazán beszélhetünk olyan elméleti kiindulópontról, amely megmagyarázná, hogy miért vezet a szerzőazonosítás során sokkal jobb eredményre a leggyakoribb szavak vizsgálata bármely más jellemzőhöz képest, azzal kell dolgoznunk, amink van. Ha a társadalmi nemre vonatkozó nyelvi jellemzők a kérdésesek, akkor James Pennebaker *The Secret Life of Pronouns (A névmások titkos élete)* című munkájára érdemes hagyatkoznunk.[3] Innen nézve nem meglepő, hogy a stilometriával foglalkozó közösség nagy lelkesedéssel fogadta ezt a szöveget, ahogy arról az alábbi recenzió is tanúskodik:

> A könyv mindenképpen megérdemel egy recenziót a Literary and Linguistic Computingban, egyrészt mert nyelvészeti és irodalmi kérdésekre egyaránt figyelmet fordít, másrészt és mindenekelőtt azért, mert interpretatív dimenziókkal gazdagítja a stilometriát (a stílus technikai vizsgálatát), amely dimenziók még a mai napig sem eléggé kidolgozottak […] Ahogy azt e folyóirat olvasói is tudják, a stilometria egyre inkább a szerzőazonosításra fókuszál, amelyben saját tevékenységének objektív hitelesítését látja. Szintén közismert, hogy a funkciószavak eloszlása a legjobb indikátora a szerzőségnek. […] Egy kicsit

[1] Françoise Giroud i Bernard Henri Lévy, *Kobiety i mężczyźni* (Warszawa: Puls, 1994).

[2] Matthew Jockers, *Macroanalysis: Digital Methods and Literary History* (Champaign: University of Illinois Press, 2013), https://doi.org/10.5406/illinois/9780252037528.001.0001.

[3] James Pennebaker, *The Secret Life of Pronouns: What Our Words Say about Us* (New York: Bloomsbury Press, 2011), https://doi.org/10.1016/S0262-4079(11)62167-2.

 DOI: 10.31400/dh-hun.2021.5.3143

> később bírálni fogom Pennebakert, amiért ignorálja a stilometrikus irodalmat, de inkább arra helyezném a hangsúlyt, hogy mivel gazdagította, és ez nem kevés.[4]

Pennebaker e cikk központi témáját könyvének harmadik, *The Words of Sex, Age and Power (A nem, a kor és a hatalom szavai)* című fejezetében tárgyalja: „A nők gyakrabban használják az egyes szám első személyt, kognitív és társadalmi vonatkozású szavakat; a férfiak gyakrabban használnak névelőt, de nincs jelentős különbség férfiak és nők között a többes szám első személy vagy a pozitív érzelmi töltetű szavak használatában."[5] Majd kiegészíti ezt a felsorolást:

> A férfiak gyakrabban használnak nagy szavakat, főneveket [ez egy másik megfogalmazása annak, hogy több névelőt használnak – J. R.], prepozíciókat, számokat és káromkodásokat. A nők több személyes névmást, igét (beleértve a segédigéket is), negatív érzelmeket (különösen a szorongás/aggódás viszonyában), tagadásokat (ne, nem, soha), bizonyosságot kifejező szavakat (mindig, teljesen), óvatosságot és valószínűséget kifejező szavak („Úgy gondolom", „Azt hiszem") használnak.[6]

Ami még fontosabb ebben a munkában, hogy a diskurzust a való életből áthelyezi az irodalomba (pontosabban drámákba és filmforgatókönyvekbe) és kijelöl egy „kilenc fokozatú férfi-női nyelvi skálát".[7] Ez alapján

> Shakespeare és Tarantino férfiak, és úgy is írnak, mint a férfiak. Azaz a férfi és női karaktereik egyaránt úgy használják a funkciószavakat, ahogyan azt férfiak szokták. A két szerző bár szóhasználatában hasonlít, írásuk tartalmában és terjedelmében egyértelműen különböznek. Shakespeare azért érdekes, mert briliánsan közvetíti a való élet témáit és a női problémákat. A funkciószavak használatából ítélve viszont úgy tűnik, hogy Tarantinóhoz hasonlóan ő sem tud a női elmébe belehelyezkedni.[8]

Itt Pennebaker egy nagyon fontos ponthoz ér. Az irodalomban ugyanis, ellentétben a való élettel, megeshet, hogy a szerző nemre jellemző nyelvezete megváltozik attól függően, hogy férfi vagy női narrátort vagy karaktert beszéltet; azaz hogy a szerző képes elrejteni saját nemének nyelvezetét. Ennek sikerességét köthetjük értékítéletekhez is, és világos, hogy Pennebaker sem habozik ezt megtenni: számára a *Periklész* és a *Ponyvaregény* alkotói elbuktak a teszten. Persze kérdés, hogy valaha valaki átment-e már rajta.

[4] John Nerbonne, „*The Secret Life of Pronouns: What our Words Say about Us.* James Pennebaker (review)," *Literary and Linguistic Computing* 29, 1. sz. (2014): 140, https://doi.org/10.1093/llc/fqt006.

[5] Pennebaker, *The Secret Life of Pronouns,* 40.

[6] Uo., 43.

[7] Uo., 49.

[8] Uo., 56.

A stilometrián belül a nemhez kötődő nyelvi jegyeket sikerült nyomon követni a politikai beszédektől[9] a beszélt[10] és formális írott nyelven[11] át egészen a blogokig[12] és a hírességek Twitter-bejegyzéséig.[13] A szépirodalomra vonatkozóan a leginkább említésre méltó munka Koppel és munkatársai tanulmánya, akik 80%-os sikert értek el az írók nemének azonosításában.[14] Érdekes módon az eredmények a fikciós szövegek esetében nem igazán különböznek a nem irodalmiaktól, amiből arra következtethetünk, hogy Pennebakernek igaza lehetett, és a legtöbb szerző nem képes meghamisítani nyelvezetét, a férfiak nem képesek „nőit írni" és fordítva. Magam is szomorúan szembesültem ezzel: abból a harminc regényből, amelyet korábban angolról lengyelre fordítottam, csupán három származott nőtől és mindhárom (különösen Nadine Gordimertől a *None to Accompany Me*) esetében végigkísérte a szerencsétlen fordítót a félelem, hogy férfiként ír inkább, mint nőként. A stilometriai kutatás ezt csak erősíti – annál is inkább, mivel leggyakrabban teljes regényekre alkalmazzák a módszereket, miközben a *cherchez la femme* ('keresd a nőt') szempontját ésszerűbb lenne a karakterek sajátos nyelvhasználatának tekintetében érvényesíteni. Köztudott például, hogy a szerzőazonosításon túli modern stilometria Burrows *Computation into Criticism* című munkájával kezdődött, amely éppen a karakterek különböző nyelvhasználatának elemzését végezte el. De ez csak még több problémát szül: egy átlagos regényben nagyon kevés karakter beszél 10000 vagy 5000 szónál többel,[15] holott ezek a leggyakrabb

[9] Mats Dahllöf, „Automatic Prediction of Gender, Political Affiliation, and Age in Swedish Politicians from the Wording of Their Speeches—a Comparative Study of Classifiability," *Literary and Linguistic Computing* 27, 2. sz. (2012): 139–153, https://doi.org/10.1093/llc/fqs010; Bei Yu, „Language and Gender in Congressional Speech," *Literary and Linguistic Computing* 29, 1. sz. (2014): 118–132, https://doi.org/10.1093/llc/fqs073.

[10] Sameer Singh, „A Pilot Study on Gender Differences in Conversational Speech on Lexical Richness Measures," *Literary and Linguistic Computing* 16, 3. sz. (2001): 251–264, https://doi.org/10.1093/llc/16.3.251; Yoko Iyeiri, Michiko Yaguchi and Yasumasa Baba, „Principal Component Analysis of Turn-initial Words in Spoken Interactions," *Literary and Linguistic Computing* 26, 2. sz. (2011): 139–152, https://doi.org/10.1093/llc/fqr005.

[11] Shlomo Argamon, Moshe Koppel, Jonathan Fine and Anat Rachel Shimoni, „Gender, Genre, and Writing Style in Formal Written Texts," *Text* 23, 3. sz. (2003): 321–346, https://doi.org/10.1515/text.2003.014; George K. Mikros, „Systematic Stylometric Differences in Men and Women Authors: A Corpus-based Study," in Reinhard Köhler and Gabriel Altmann, eds., *Issues in Quantitative Linguistics 3: Dedicated to Karl-Heinz Best on the Occasion of His 70th Birthday,* 206–223 (Lüdenscheid: RAM-Verlag, 2013).

[12] Jonathan Schler, Moshe Koppel, Shlomo Argamon and James Pennebaker, „Effects of Age and Gender on Blogging," *AAAI Spring Symposium: Computational Approaches to Analyzing Weblogs* 6 (2006): 199–205; George K. Mikros, „Authorship Attribution and Gender Identification in Greek Blogs," in Ivan Obradović, Emmerich Kelih and Reinhard Köhler, eds., *Selected Papers of the VIIIth International Conference on Quantitative Linguistics,* 21–32 (Belgrade: Academic Mind, 2013).

[13] George K. Mikros and Konstantinos Perifanos, „Authorship Attribution in Greek Tweets Using Multilevel Author's Ngram profiles," in E. Hovy, V. Markman, C. H. Martell, and D. Uthus, eds., *Papers from the 2013 AAAI Spring Symposium "Analyzing Microtext", Stanford, CA, 25–27 March 2013,* 17–23 (Palo Alto, CA: AAAI Press, 2013).

[14] Moshe Koppel, Shlomo Argamon and Anat Rachel Shimoni, „Automatically Categorizing Written Texts by Author Gender," *Literary and Linguistic Computing* 17, 4. sz. (2002): 401–412, https://doi.org/10.1093/llc/17.4.401.

[15] Jan Rybicki, „Does Size Matter? A Re-Examination of a Time-Proven Method," in *Digital Humanities 2008: Book of Abstract,* 184 (Finland: University of Oulu, 2008).

 DOI: 10.31400/dh-hun.2021.5.3143

szavakra irányuló kutatás valóban-biztonságos és majdnem-biztonságos határai,[16] és általában is nagyon nehéz olyan hősnőt találni, aki megközelítené ezeket az értékeket például a történelmi regényekben – ha a könyvet férfi írta –, továbbá ugyanez igaz Shakespeare női karaktereire is.[17] Tulajdonképpen a több elbeszélős regényekben lehet a legjobban megítélni a szerzőt abból a szempontból, hogy mennyire sikerült túllépnie a nemét jellemző nyelvhasználaton. Az *Üvöltő szelek,* az *A puszta ház* vagy az *Ulysses*[18] például elég anyagot adhat a stilometrikusnak, hogy az idiolektusok alapján végezzen nemek közti összehasonlítást.

De ez egy kissé más történet. Ha visszatérünk saját feladatunkhoz, azt látjuk, hogy minden tanulmány, amely hasonló problémákkal foglalkozik, Mark Olsen munkájára hivatkozik, aki a korai női irodalmat és „a hiányzó női hang" tudatos megteremtését vizsgálta az Écriture féminine: *Searching for an Indefinable Practice?* című könyvében.[19] Olsen különböző műfajokban végzett összehasonlító elemzést a szógyakoriságok alapján mindkét nemre vonatkozóan, hogy azonosítsa a „férfi" és „női" nyelvhasználat kulcsszavait – sőt az időbeliséggel is számolva mintegy öt évszázadon át vizsgálta e jelenséget. Munkája példaadó jelen kutatás számára is.

És itt van még Matt Jockers óvatosságra intő története is: amikor a különböző jellemzők relatív hatását kutatta a stilometriában, azt találta, hogy „a klasszifikációs és a lineáris regressziós tesztek során csak kis szerepet játszik az alkotó neme", hiszen az eredményeknek csak 8 százalékáért felelős ez a szempont.[20] Miközben azt állítja, hogy „a 19. századi regények esetében nem különösebben nehéz elkülöníteni a férfiakat a nőktől" (ez egyáltalán nem ellentmondásos, mivel a nemeket, az irodalmit is, ekkor még elég szigorú apartheid rendszer jellemezte). Jockers felállított egy listát is azokról a 19. századi írókról, akiket a leginkább nehéz a társadalmi nem szempontjából besorolni – a lista tartalmazza ennek a tanulmánynak is néhány főhősét: William Beckford, Maria Edgeworth, Matthew Lewis és William Godwin.[21]

Látszólag tehát a feladat, amit a COST Action keretében kaptam, sokkal egyszerűbb volt a korábbiakhoz képest: *cherchez l'homme!,* azaz 'keresd a férfit!' egy csak női írótól származó, a 18. századtól a korai 19. századig tartó korszak szövegkorpuszában, majd megvizsgálni, hogy van-e bármilyen különbség azok között a nők között, akik egy bizonyos ponton bekerültek az angol irodalmi kánonba (mielőtt azt elfújta a szél) és azok között, akik nem. Az „egy bizonyos ponton" itt nem csak egy klisé: a kanonizált nők között például ott van Austen és Burney is, akiknek az irodalmi karrierjük nagyon

[16] Maciej Eder, „Does Size Matter? Authorship Attribution, Small Samples, Big Problem," *Literary and Linguistic Computing* 30, 2. sz. (2015): 167–182, https://doi.org/10.1093/llc/fqt066.

[17] Jan Rybicki, „Twelve *Hamlet*s: A Stylometric Analysis of Major Characters' Idiolects in Three English Versions and Nine Translations," in *Digital Humanities 2007: Conference Abstracts,* 191–192 (Urbana-Champaign: University of Illinois, 2007).

[18] Ami azt illeti, legalább egy szerző, úgy tűnik, kiválóan teljesít ezen a teszten. Az alább ismertetett módszerekkel végzett előzetes vizsgálatok azt mutatják, hogy Joyce „női" epizódjai az *Ulysses*ben, mint például a *Nauszikaá* és a *Penelopé,* az angol női modernisták szövegei köré csoportosulnak; míg Joyce remekművének más részei megmaradnak a férfi környezetben.

[19] Mark Olsen, „*Écriture Féminine:* Searching for an Indefinable Practice?" *Literary and Linguistic Computing* 20, Issue Suppl. (2005): 147–164, https://doi.org/10.1093/llc/fqi020.

[20] Jockers, *Microanalysis,* 92.

[21] Uo., 94–95.

különböző utakat járt be végül. A *Büszkeség és balítélet* szerzője ma egy valódi szent az angol irodalomban és a filmadaptációkban, valamint meghatározó szereplője az akadémiai kutatásoknak is, ezzel ellentétben a *Cecilia* írója elveszítette egykori előnyét riválisával szemben. Ezt illusztrálja az alábbi folyamat: egy népszerű angol irodalmi kézikönyv 1874-es kiadása egy rövid fejezetet szentel Burneynek, és egy szóval sem említi Austent; 1891-ben ugyanaz a rövid bekezdés jelent meg az előbbiről, az utóbbiról viszont már legalább háromszor olyan hosszú ismertetés.[22] (Persze még így is mindkét szerző sokkal ismertebb, mint a hamarosan bemutatott Chawton House bármely másik írója.) Nem meglepő módon a tanulmányban is szereplő híres férfi szerzők ugyanabban a tiszteletre méltó korabeli kiadványban jóval részletesebben kerültek bemutatásra. A kánonok persze változnak térben és időben; a lengyel nézőpontból például a 18. századi angol irodalmi kánonnak tartalmaznia kellene Jane Portert, összesen két rövid regény szerzőjét (eltekintve az egyetlen színdarabjától és novellisztikájától), amelyek közül az egyik, a *Thaddeus of Warsaw* Lengyelországban játszódik és valószínűleg az első történelmi regény, ami foglalkozik a lengyel történelem drámai eseményeivel – időben megelőzve a hasonló témájú, de lengyel nyelvű műveket.

2. A kutatási anyag és a módszer

A tanulmányozandó korpusz a korai női irodalom megújult kutatóközpontja, a Chawton House könyvtári anyagából jött létre. Már a központ elhelyezkedése is nagyon találó, hiszen Edward Austen Knight, Jane Austen testvérének ingatlanában működik, de maga Jane is a környéken lakott. A korpusz a Chawton House digitalizációs projektjének eredménye: a tanulmány megírásának idejében 46, nők által írt regényt tartalmazott, amelyek 1723 és 1830 között keletkeztek – ebből 34-nek van nevesített szerzője, 5 szerző két regénnyel is szerepel, 12 pedig névtelen. A korpusz készítői szerint a szövegek „jól jelzik az 1600 és 1830 között létrejött női irodalom gazdag szövegvilágát és innovatív jellegét", és abban bíznak, hogy „azáltal, hogy ezeket az alig ismert regényeket elérhetővé teszik a szélesebb közönség számára [...], és érdeklődést váltanak ki az olvasók új generációjának körében, egyben felélénkítik a kevésbé ismert szerzőkről szóló tudományos diskurzust is."[23] Ezeket a szövegeket két referenciakorpusszal hasonlítottam össze: az egyik a híresebb női szerzőket (Austen, Radcliffe, Burney, Edgeworth, Shelley, összesen 22 regénnyel), a másik a kor híres férfi íróit tartalmazza (Swift, Johnson, Richardson, Fielding, Sterne, Smollett, Goldsmith, Beckford, Peacock, összesen 21 regénnyel). Hogy a két fő kérdésünket megválaszoljam, ezeknek a korpuszoknak a különböző kombinációját alkalmaztam.

A kutatás során különböző beállításokat használtam az R nevű, nyílt forráskódú, elsősorban statisztikai feladatokra kialakított programozási környezet *stylo* névre hallgató, külön stilometriai kutatások számára létrehozott bővítményében,[24] az így kapott

[22] Truman Jay Backus, *Shaw's New History of English Literature* (New York and Chicago: Sheldon & Co., 1874, 1891).

[23] Hozzáférés: 2021.11.28, `https://chawtonhouse.org/the-library/womens-writing-in-english-2/novels-online/`.

[24] Maciej Eder, Mike Kestemont and Jan Rybicki, „Stylometry with R: A Suite of Tools," in *Digital Humanities 2013: Conference Abstracts,* 487–489 (Nebraska: University of Nebraska, 2013).

 DOI: 10.31400/dh-hun.2021.5.3143

eredményeket pedig a *Gephi* vizualizációs platform segítségével ábrázoltam hálózatok formájában. A munkafolyamat a szövegcsoportok fáinak és konszenzushálózatoknak (Bootstrap Consensus Network, BCN) a létrehozásából állt, amelyeket a leggyakrabban használt szavak klasszikus Delta-távolsága,[25] valamint a közepesen gyakori szavak Burrows Zetájának Craig által módosított eljárásával kapott értékei alapján hoztam létre (ez utóbbi a *stylo* „oppose" függvényébe került implementálásra).[26] Mindkét eljárás a klaszteranalízis vizualizációjának bemeneti értékeit képezte. Magukat a hálózatokat a *Gephi* „Force Atlas 2" algoritmusa hozta létre, amely különösen alkalmas a különbségek – például irodalmi szövegek közti különbségek – ábrázolására. A bemenet – a klaszterek kapcsolatainak kereszthitelesített erőssége – kétféleképpen reprezentálható: a szövegek közötti élek nagyságával, valamint az őket reprezentáló csomópontok közötti távolsággal.[27] A program készítőit idézve:

> A „ForceAtlas2" egy erő-irányított *(force-directed)* elrendezés: szimulál egy fizikai rendszert annak érdekében, hogy térbelivé tegyen egy hálózatot. A csomópontok taszítják egymást, mint a töltött részecskék, míg az élek magukhoz vonzzák a saját csomópontjaikat, mint a rugók. Ezek az erők olyan mozgást képeznek, ami konvergál a kiegyensúlyozott állapothoz. A végső konfiguráció így képes segíteni az adatok értelmezését.[28]

Itt kell megjegyeznem, hogy szeretném elkerülni a különböző statisztikai módszerek hívei között jelenleg is futó csatározást, hogy mely eljárások a legoptimálisabbak, a legkorszerűbbek, vagy egyszerűen csak divatosak a stilometriai kutatásokban. A szomorú igazság az, hogy nincs egyetemes egyetértés, és az összehasonlító tanulmányok továbbra is csupán egy százalékpontnyi javulásokról adnak hírt; és bár bizonyos módszerek (mint például a Support Vector Machines) némi előnyt jelenthetnek kevésbé intenzív eljárásokkal szemben (például a jelen tanulmányban is használt Delta-alapú klaszterelemzés), a módszerek közötti választás jelentősége egy irodalmi tanulmányban (szemben egy módszertani-logikaival) valószínűleg nulla. Őszinte véleményem, hogy sokkal fontosabb a stabil módszertan használata, még akkor is, ha ez azt jelenti,

[25] John Burrows, „'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship," *Literary and Linguistic Computing* 17, 3. sz. (2002): 267–287, `https://doi.org/10.1093/llc/17.3.267`.

[26] Maciej Eder, „Metody ścisłe w literaturoznawstwie i pułapki pozornego obiektywizmu – przykład stylometrii," *Teksty Drugie* 2. sz. (2014): 90–105; Maciej Eder, Visualization in Stylometry: Some Problems and Solutions," *Literary and Linguistic Computing* 32, 1. sz. (2017): 50–64; Jan Rybicki, „Visualizing Literature: Artistic Statistics," in Magdalena Bleinert, Isabela Curyłło-Klag and Bożena Kucała, eds., *Art of Literature, Literature in Art,* 135–146 (Krakow: Jagellonian University Press, 2014). Ez igazán kínos. Egy gyors felmérés stilometristák körében világszerte megerősítette a gyanúmat, hogy David Hoover volt az első, aki azt hangoztatta, hogy a Zeta-szavak egy Delta-szerű eljárásban hasznosíthatók lennének: tehát ebben egyetértés van – ugyanakkor nem tudtunk megegyezni (még David sem), hogy mikor és hol történt ez pontosan.

[27] Mathieu Bastian, Sebastien Heymann and Mathieu Jacomy, „Gephi: An Open Source Software for Exploring and Manipulating Networks," in *Proceedings of the Third International Conference on Weblogs and Social Media, ICWSM. San Jose, California, May 17–20, 2009,* 361–362 (Menlo Park, CA: The AAAI Press, 2009), http://doi.org/ 10.13140/2.1.1341.1520.

[28] Mathieu Jacomy, Tommaso Venturini, Sebastien Heymann and Mathieu Bastian, „ForceAtlas2, a Continuous Graph Layout Algorithm for Handy Network Visualization Designed for the Gephi Software," *PLoS ONE* 9, 6. sz. (2014), `https://doi.org/10.1371/journal.pone.0098679`.

hogy itt-ott néhány százalékkal csökken a szerzőazonosítás sikere, azzal szemben, amikor ismeretlen változókkal végzünk műveleteket mindig új algoritmusok segítségével, miközben egy irodalmi (nyelvészeti) kérdést próbálunk megoldani.

3. Eredmények

Először a legegyszerűbb módon, azaz a leggyakoribb szavak elemzésével kellett ellenőriznem, hogy egyáltalán elkülöníthető-e a nemek nyelvhasználata. E tekintetben reménykeltők Pennebaker eredményei, miközben szem előtt kell tartani, hogy bármely szöveggyűjtemény leggyakoribb szavainak listája ritkán azonos egy kizárólag a funkciószavakból előzetesen összeállított listával. De hiába minden remény, a leggyakoribb szavak nem tártak fel semmilyen eredményt a nemek tekintetében. A híres férfiak és híres nők szógyűjteményeinek klaszteranalízise során létrehozott konszenzusfák csupán a szerzőség felismerésében teljesítettek jól, és csak a gótikus szerzők esetében nem működtek megfelelően.[29] A Chawton House regényeinek (a névteleneket is beleértve) hozzáadása keveset változtatott a dolgon.

A helyzet akkor válik érdekesebbé, amikor a klaszteranalízis eredményét a hálózati vizualizáció bemeneteként használjuk a *Gephi*ben. Mivel ez utóbbi nem csak egy jellemzőt és nem csak a legerősebb klasztereket emeli ki, a szerzőségi kapcsolatok mindent uraló erejét ekkor más jellemzők gyengítik. Az *1. ábrá*n két további, a csoportosítást befolyásoló szempontot is felfedezhetünk: a műfaj/téma, ami miatt Shelley Radcliffe és a *Vathek* mellé kerül (jobbra), és a nemi hovatartozás, amiért a férfiak egy csomóba rendeződnek középen, a többi nő pedig egy másikba (balra). Az egyetlen nemi szempontból rosszul azonosított szerző, Richardson, részben felmenthető: a *Pamela* és a *Clarissa* talán valóban jól sikerült esetei a másik nem nyelvi sajátosságainak átvételére, amelyet a levélregény műfaja is ösztönözhetett. A *Grandison* (szintén Richardson műve) viszont talán a legkevésbé remélt érték ezeknek közelében, ami azt sugallhatja, hogy a mű elején Harriet Byron jobban dominál, mint később a névadó hős – vagy talán

[29] Valójában Lewis és Godwin olyan megrögzött bajkeverők voltak a kutatás kezdeti szakaszában, hogy teljesen ki kellett őket zárni a referenciakorpuszból. Lewis hajlamos volt csatlakozni az összes többi gótikus íróhoz (Walpole *Otrantói kastély*ának ugyanezen okból kellett távoznia), férfiakhoz és nőkhöz egyaránt, akik pedig a legtöbb kezdeti elemzésben folyamatosan külön csoportot alkottak. A gótikus regényben – legalábbis az én szöveghalmazomban – számszerűleg erősen domináltak a nők, viszont a műfaji/tematikus jegyek sikeresen elfedték a nemek közti különbséget. Beckford azért maradt a korpuszban, hogy ezt a jelenséget a gótikus *Vathek* című regényével demonstrálja, valamint az *Azemia* érdekes viselkedése miatt, amely a Chawton House-tól származó regények egy részének műfaji paródiája, és amely a vizsgálat során végig ragaszkodott is a parodizált művekhez. Ez nem újdonság, hiszen a paródiák stilometrikus viselkedését, mint a legtöbb izgalmas jelenséget a területen, már Burrows is leírta. (John Burrows, „Who Wrote Shamela? Verifying the Authorship of a Parodic Text," *Literary and Linguistic Computing* 20, 4. sz. [2005]: 437–450, `https://doi.org/10.1093/llc/fqi049`.) Godwin viselkedése még furcsább volt egy olyan korpuszban, amely lánya műveit is tartalmazta (elvégre a Frankenstein állítólag Mary Shelley gyermekkori élményeinek egy részét is magába foglalja), olyannyira, hogy ezzel egy külön dolgozatban foglalkozom majd; a szülői beavatkozás legkisebb gyanúja miatt azonban neki is mennie kellett. Ennek kapcsán meg kell említenem, hogy más, potenciálisan érdekes gyermek-szülő kérdések is kapcsolódnak a korpuszhoz. Érdemes lenne nyomon követni, hogy milyen hatással volt Mary Shelleyre édesanyja, Mary Wollstonecraft, és édesapja; ahogy Maria Edgeworth írásai is hajlamosak nagyon próteuszivá válni abban az időben, amikor apja önéletrajzán dolgozott.

 DOI: 10.31400/dh-hun.2021.5.3143

még logikusabb, hogy míg a *Pamela* és a *Clarissa* a női nyelvhasználat miatt került a női írók körébe, addig a szerzői kézjegy rendelte hozzájuk Richardson harmadik regényét.

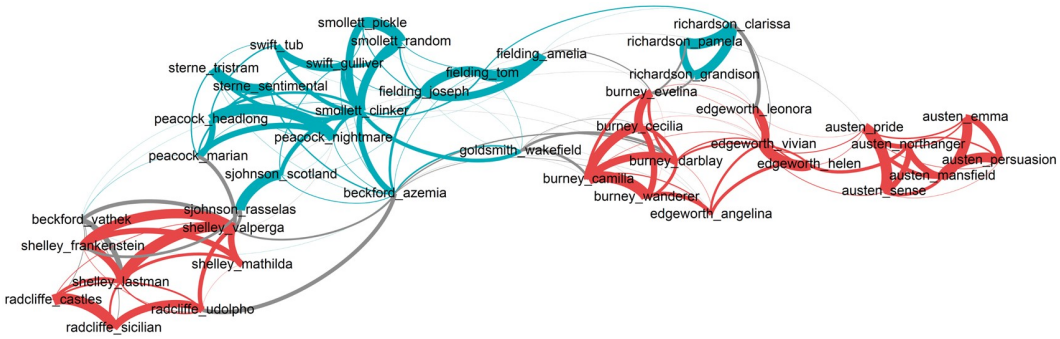
A Chawton House gyűjteményének kiegészítése a „híres férfiak" és „híres nők" referenciakorpuszával biztató eredményekre vezettek a tizenkét ismeretlen szerzőségű szöveg nemi identifikációjakor *(2. ábra)*. A férfiak csoportja változatlan, Richardson továbbra is kívülálló, de az anonim szövegek némelyike most egészen izgalmas pozíciót foglal el. Ez különösen igaz a *The Imposters Detected: or, the Life of a Portuguese*-re *(Az azonosított imposztorok, avagy egy portugál élete)* (1760), és egy nagyon richardsoninak tűnő szövegre, a *The Reward of Virtue: or, the History of Miss Polly Graham*re (1769). Ha a leggyakoribb szavak eredményét megbízhatónak tartjuk, akkor ezek volnának az első gyanúsítottak arra nézvést, hogy férfiak kerültek a nők közé.

Ezek az eredmények azonban csak akkor lennének igazán megbízhatók, ha létezne egy elméleti modell a férfi–női különbségekre a lexikai választások tekintetében. Ami először eszembe jut az Pennebaker már idézett listája (vagy pontosabban fogalmazva lexikai kategóriái). Ezek a kategóriák egy olyan háromszáz szavas listává alakíthatók, amely megfelel Pennebaker leírásának, de ezek alapján egyáltalán nem észlelhető elkülönülés férfiak és nők között. Ez tehát ismét hiú reménynek bizonyult, ami nem feltétlenül meglepő: jelen kutatás korpusza ugyanis 18. századi és 19. század eleji szövegeket tartalmaz, Pennebakeré viszont sokkal általánosabb volt. Jockers már említett kutatásában viszont szerepel néhány szerző az én referenciakorpuszomból is, ezért még egy kísérletet tettem az ő „Jellemzők, amik legjobban megkülönböztetik a férfiakat és nőket" című szólistájával.[30] Volt némi átfedés ebben a listában, a Pennebaker-félében és az én kutatásom leggyakoribb szavai között, de a nemek elkülönülő nyelvhasználatának kérdésében nem tapasztaltam előremutató eredményt ez esetben sem.

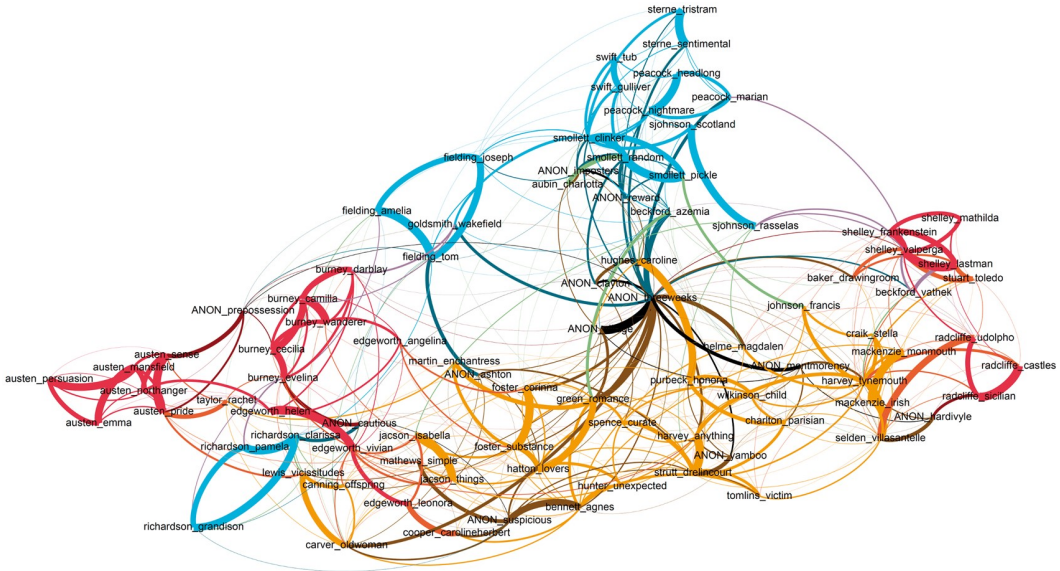
A megfelelő szólista keresése Burrows egy másik módszeréhez vezetett, a Zetához, amely azonos méretű részekre darabolja a szövegeket, majd olyan szavakat keres, amelyek következetesen felbukkannak egy szövegben, vagy szövegcsoportban, miközben jellemzően hiányoznak egy másikban.[31] Ezzel a módszerrel az adott szövegre jellemző, közepes gyakoriságú szavak halmazát alkotjuk meg (durván mondva a klasszikus *log-likelihood* eljárással kinyerhető kulcsszavakat a funkciószavak kivételével). E megközelítés legvonzóbb tulajdonságai közé tartozik, hogy az ilyen szavak sokkal tartalmasabb jelentésűek, mint a magas frekvenciájú funkciószavak és emiatt hagyományos irodalmi szempontból is jobban értelmezhetők. A „híres férfiak" és „híres nők" következetesen előnyben részesített szavainak, és azon keresztül a férfi és női nyelvhasználat közti különbségnek a megtalálásához a fenti módszer implementációját, a *stylo* „oppose" függvényét alkalmaztam. Az eredményként kapott körülbelül hatszáz szóból a Chawton House és a „kanonikus szerzők" egységes korpuszán tesztelve került kiválasztásra az az optimális mennyiség, amellyel az elemzés sikeresnek tekinthető. Majd az eredmények becsületes és alapos kimazsolázását *(cherry-picking)* követően csak azokat a grafikonokat tekintettem jelentősnek, amelyek helyesen szétválasztották a „híres férfiak" és „híres nők" alkorpuszát.

[30] Jockers, *Macroanalysis,* 94.

[31] John Burrows, „All the Way Through: Testing for Authorship in Different Frequency Strata," *Literary and Linguistic Computing* 22, 1. sz. (2006): 27–47, https://doi.org/10.1093/llc/fqi067.



1. ábra. A „híres férfiak” és „híres nők” szövegeihez készült hálózat, a 100–1000 leggyakoribb szóra vonatkozóan.



2. ábra. A „híres férfiak” és „híres nők”, valamint a Chawton House regényeinek ismert és ismeretlen (ANON előtaggal ellátott) szerzők által írt szövegeihez készült hálózat, a 100–1000 leggyakoribb szóra vonatkozóan.

A *3. ábra* egy ilyen gráfot mutat be 248 közepes frekvenciájú férfi és női szó alapján (az eredmények a hosszabb szólistáknál, 490 darabig nagyon hasonlók). Mindenekelőtt szembetűnő, hogy a szövegek, amelyekből a szólisták létrejöttek, szinte tökéletesen elkülönülnek az ábrán a nemek szerint. Csak egy kivétel volt: Beckford *Azemiá*ja a klaszterfa női részére került. Bár ez sem olyan meglepő, hiszen a regény a korszak „női” írásainak paródiájaként jött létre. Beckford világosan kijelöli céljait a regény alcímében: „Kortárs szerzők stílusának imitációja versben és prózában”, sőt egy női személy, Jacquette Agenta Mariana Jenks szerepeltetésével tovább fokozza az illúziót, amelyet már parodisztikus ajánlásában is felvezet:

> Kétségbeesve, hogy e lapoknak vajon átadhatom-e mindazt az éleselméjűséget, ragyogást, következetességet, finomságot, emelkedettséget, fantáziát, zsenialitást, humort, ítélőképességet, lényeglátást, tudást, fényűzést, vidámságot, nai-

 DOI: 10.31400/dh-hun.2021.5.3143

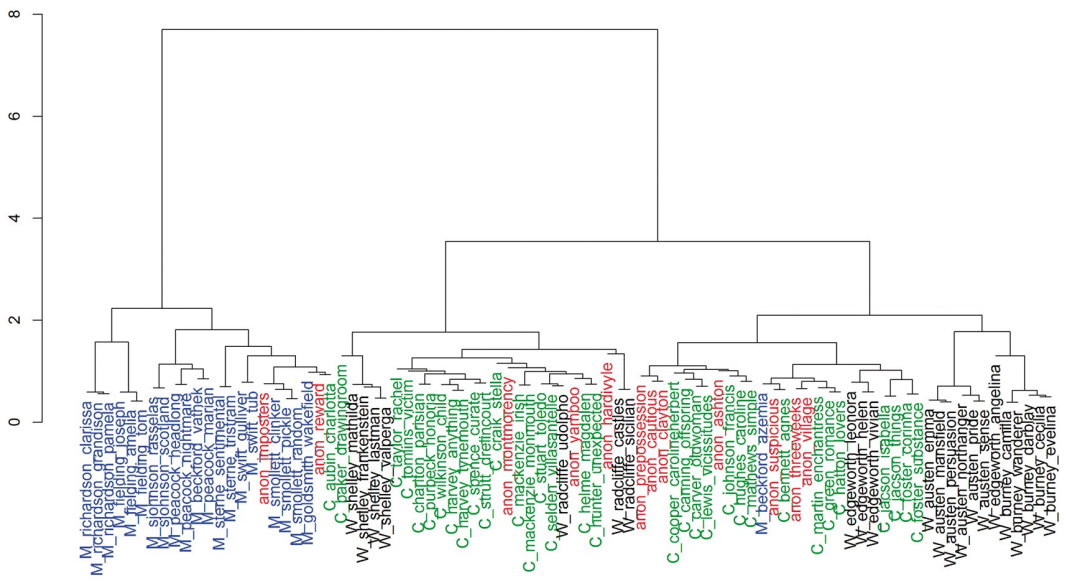
> vitást, mindentudást, pátoszt, gyorsaságot, fanyarságot, szelídséget, gyengédséget, városiasságot, lendületességet, szellemességet, kiválóságot, fiatalosságot és élénkséget, amely Kegyed tollából árad, mégis megkockáztatom, hogy ez az irodalmi szárnypróbálgatás, amelyet megtiszteltetés számomra az Ön oltalmazó jóindulatának ajánlanom, inkább az Ön mosolygó támogatásának (amely oly kedves az irodalmi lelkeknek), mint bármilyen egyéni érdemnek köszönhetően, arra szolgálhat, hogy nem kevésbé elfogadhatatlanul felkeltse a kifinomult barátok kedves érdeklődését a brit szellem azon magasabbrendű régiójában, ahol Kegyed jóságos és sugárzó csillagként ragyog.

Míg a „híres nők" esetében sehol nem történt téves azonosítás, a Chawton House egy ismert szerzőségű szövege a férfiak csoportjában jelent meg: Penelope Aubin *The Life of Charlotta Du Pont, an English Lady; Taken from her own MEMOIRS* (1723) című regénye. A szerző vitathatatlanul létezett, közismert, hogy volt férje és három gyermeke, számos regényt és fordítást jegyzett, így tehát nem lehet egyszerűen félresöpörni a hibát azzal, hogy ő is csak egy kitalált Jenks. Hozzáteszem, bár nem túl nagy meggyőződéssel, hogy a félrepozicionálása talán a regény rendkívül kalandos cselekményével magyarázható (ide értve néhány madagaszkári kalózt is, akik úgy tűnik, rossz óceánon tevékenykedtek), amely nagyrészt a szerző bátyjának a kolóniákon szerzett élményein alapul:

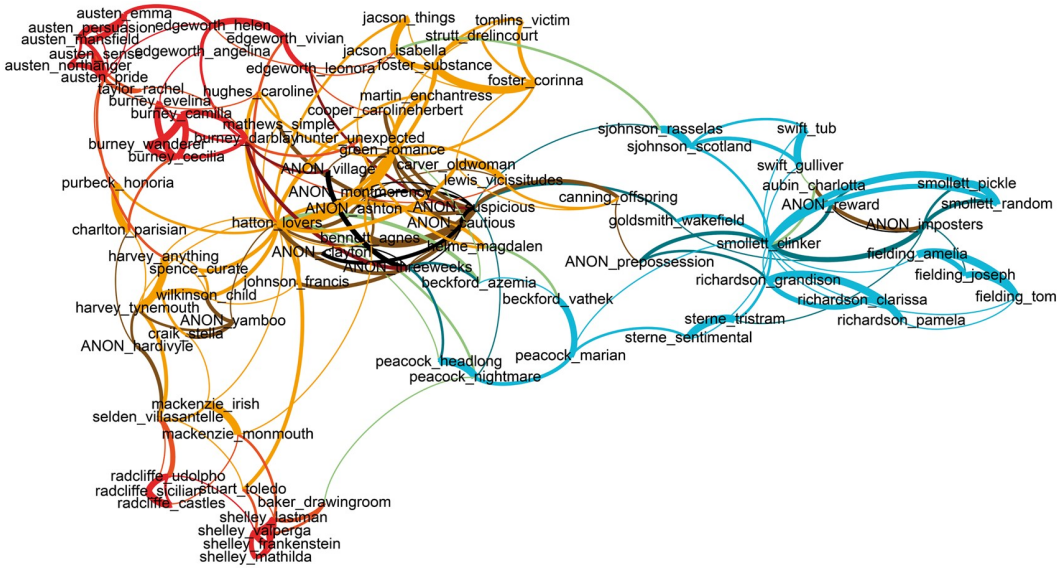
> Beszámolva arról, hogy a mostohaanyja hogyan száműzte őt Virginiába, hogyan rabolták el a hajót madagaszkári kalózok, és hogyan foglalta azt vissza egy spanyol hadfi. A spanyol Nyugat-Indiában kötött házasságáról és az ott átélt kalandjairól, valamint az Angliába való visszatéréséről. És több úriember és hölgy történetéről, akikkel utazásai során találkozott; némelyikük rabszolga volt Berberföldön, mások pedig hajótörés szenvedtek a barbár partokon a nagy Oroonoko folyónál: onnan menekültek el és tértek haza végül biztonságban Franciaországba és Spanyolországba.[32]

További eredménye a kutatásomnak, hogy két névtelen mű a Chawton House korpuszából – és következetesen ez a kettő – nem került be a női szerzők halmazába a klaszterelemzés során, ahogyan a funkciószavak alapján történő csoportosításkor sem: a *The Imposter Detected* és a *The Reward of Virtue*. A *4. ábra* hálózatán, amely ugyanezekkel a mérésekkel jött létre, a férfiak és nők jól elkülönülnek egymástól az említett két regény kivételével. Ezen a gráfon egy másik Chawton House-regény is megközelíti a férfi szerzők csomópontját: a *Prepossession,* és csak találgatni lehet, hogy vajon a névtelen írónő írói tehetsége miatt került ide a könyv, amelynek alcíme: *Memoirs of Count Touloussin. Written by Himself (Count Touloussin saját kezűleg írt emlékei)* vagy, ami kevésbé valószínű, hogy valóban létezett egy Count Touloussin nevű személy. Azt is érdemes megjegyeznünk ugyanakkor, hogy a többi névtelen szöveg gondtalanul beilleszkedik a Chawton-gyűjtemény klaszterébe.

[32] Van egy másik szerzői csontváz is ebben a szekrényben, bár ez ebben az esetben nem sokat segít: majdnem egy évszázaddal később, 1770-ben valaki (nyilvánvalóan egy könyvkereskedő) megváltoztatott néhány nevet Aubin művében (aki már több évtizede halott volt), és névtelenül kiadta *The Inhuman Stepmother; or the History of Miss Harriot Montague* címmel.



3. ábra. A klaszterelemzés során létrejött fa a „híres férfiak” (M előtaggal), a „híres nők” (W előtaggal), illetve a Chawton House ismert (C előtaggal) és ismeretlen (ANON előtaggal) szerzőinek szövegeihez, a 284 „híres férfi”/„híres nő” Zeta-kulcsszó alapján.



4. ábra. A hálózatelemzés során létrejött gráf a „híres férfiak”, a „híres nők”, illetve a Chawton House ismert és ismeretlen (ANON előtaggal) szerzőinek szövegeihez, a 284 „híres férfi”/„híres nő” Zeta-kulcsszó alapján.

Ami még érdekesebb, az annak a szólistának az összetétele, amelynek a fenti eredmény köszönhető. A listában található szavak nagy része értelmezhető Jockers és Pennebaker eredményei alapján is. A női rész különösen feltűnő, mivel – eltekintve az igék nagyobb számú arányától – olyan elemeket tartalmaz, amelyek direkt kapcsolatban állnak a regények témájával és általános hangulatával, és megerősítik a legtipikusabb

 DOI: 10.31400/dh-hun.2021.5.3143

sztereotípiákat a kor női irodalmával kapcsolatban. Ezek a szavak legalább három kategóriába oszthatóak:

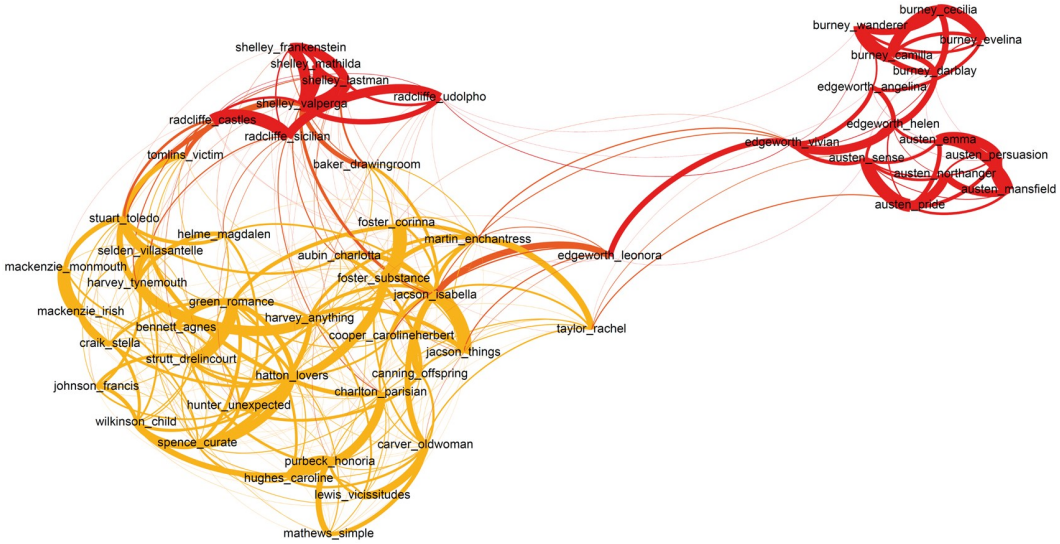
> 1. érzelmi (és azok kifejezésére szolgáló) szavak: *érzések, érezte, érez, szorongó, érezni, egyedül, fájdalmas, aggodalom, figyelem, gyönyörű, meglepetés, kedvesség, megbán, mosoly, kötődés, zaklatott, izgatott, boldogság, érzelmek, óhajtott, izgatott, barátságos, kíván, csodálat, szeretet, mohó, félt, hiú, szomorkodik, érzés, agónia, riasztás, extrém, szeretetre méltó/kedves, elfogult, érdeklődő, meglepett, illendő, bátorság, mohón, érzelmek, magány, arckifejezés, érzékenység, élénk, stabil, bánat, kifejezés, megerőltető, szenvedés, pillantás, elveszett, félelmetes, csalódottság, zárkózott, nyugodt, melankólia;*
>
> 2. felkiáltások és közbeszólások, közbevetések: *így kiáltott fel, ó, igen, ah, kiáltotta;*
>
> 3. társadalom és család szavai: *társadalom, szokások, anya, elegáns, forma/alak, rezidencia, körülmények, elkísér/kíséret.*

E kutatás szerint, a férfiak kulcsszavai kevésbé kötődnek a történet tartalmához, de annyi elmondható, hogy e szerzők kedvelik az erényeket kifejező szavakat, mint például: *segítség, őszinte, becsület, szívesség, megbocsájtás, megérdemel, érdem, rend, reputáció, minőség;* udvariasak más férfiakkal: *úriember, földesúr, társ;* bőven van mondanivalójuk a másik nemről: *csinos, ruhák, jó hír/hírnév;* kedvelik az archaizmusokat (ez részben Swift szövegeivel magyarázható); előszeretettel használnak rövidítéseket; gyakran szólítják meg az olvasót, és bár emlegetik Istent és az ördögöt is, sosem úgy káromkodnak a karakterek, hogy túlzottan szókimondóak volnának: *káromkodott, esküdözött;* beszélnek a testről és testrészekről: *száj, orr;* és egész biztos, hogy érdekli őket a pénz és a számok: *darab, kiadás, drága, hat, húsz, három.*

De van itt még több is. A *4. ábra* azt mutatja, hogy némileg megosztott a hálózat bal oldala. E rész centruma tartalmazza a Chawton House regényeit, az ismert szerzőségű és az anonim munkákat egyaránt; míg a híres riválisaik a perifériára szorultak, amely elkülönülésen belül saját részhalmazt képeznek a gótikus műfaj képviselői (baloldalt, alul). Van egy módja annak, hogy még precízebben értékeljük ezt a jelenséget: fel kell mérni a kanonizált női szerzők sajátosságait a férfi szerzők és a Chawton House kánonból kimaradt szerzőinek vonatkozásában. A két csoport hálózatanalízisét a vonatkozó Zeta-szavak Delta-távolságára alapozni egyszerű tautológiának tűnhet (de nem teljesen az, mivel a Delta és a Zeta két független módszer, nagyon különböző szólistákkal, lásd *5. ábra*), sőt a valódi különbség csak úgy érthető meg, ha összehasonlítjuk a két csoportra jellemző Zeta-szavakat. Ennek alapján azt mondhatjuk, hogy a Chawton House szerzői egyértelműen azokat a szavakat preferálják, amelyek a legsztereotipikusabban köthetőek a szentimentális női irodalomhoz. A sztereotípiát pedig talán pont az teszi kevésbé sztereotíppá (azaz egy egyszerű előítéletnél többé), hogy tetten érhető a létrehozott szólistákban is. Még egyszer tehát, ez a hosszú lista értelemszerűen és könnyedén osztható különböző szemantikai kategóriákra: ezek közé tartoznak a érzékenyen árnyalt anatómiai részletek (úgy mint *kebel, orca, karok, mellek, ajkak*); családi és társadalmi kifejezések (például: *férj, gyermek, szülő, úrhölgy,*

lány, szerető, feleség, apa, újszülött, Anglia, rezidencia, szolga, fiú, uraság, özvegy, kapitány, házi), absztrakt, elvont fogalmak *(hősiesség, kívánságok, ragaszkodás, gondviselés, halál, barátság, elvek, vallás, lélek),* közbeszólások és dicsérő kifejezések *(elbűvölő, gyöngéd, erényes, ártatlan, elegáns, ó!, angyal, jóképű, divatos, értelmes, elbűvölő, őszinte),* valamint a negatív érzelmek *(szerencsétlen, végzetes, nélkülöző, bűnös, felkavart, sérült, kín, bánat, nyomorult).*

A kanonizált női írók által kedvelt szavak ezzel szemben sokkal hétköznapibbnak, földhöz ragadtabbnak tűnnek. Az első húsz ezekből (a sorrendet a csökkenő Zeta-pontszámok alapján állítottam fel) a következők: *alig(ha), bármi, egyéb, fajta/féle, senki, sírt, néz, beszél, jelenleg, minden, jön, kezdődik, van, beszélt, között, beszél, előre, sétált, mindenki, nem fog.* Igazából úgy tűnik, mintha egy sima szógyakorisági lista elemei lennének, illetve osztoznak pár kategóriában (rövidítések, pozitív érzelmek, jellemzők) az előző vizsgálat „férfi" szólistájával is. Ez elég sokat elárul a kanonizációs folyamatokról: egy női szerző akkor tör be az általánosan elfogadott kánonba, minél inkább úgy ír, mint egy férfi.



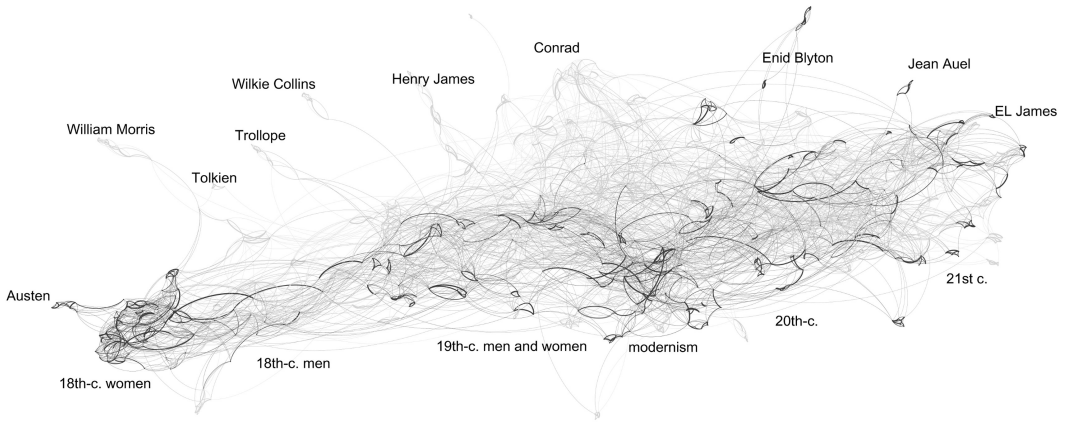
5. ábra. A hálózatelemzés során létrejött gráf a „híres nők", illetve a Chawton House ismert szerzőinek szövegeihez, a megfelelő Zeta-kulcsszavak alapján.

Mivel a nemi alapú megoszlás egyre tisztábban rajzolódik ki a vizsgált anyagban – abban az anyagban, amely a regény műfajának 18. századi felemelkedéséből jött létre –, csábító, hogy egy lépéssel továbbhaladva megvizsgáljuk, vajon ezek (vagy más) „férfi" és „női" szavak továbbra is tetten érhetők-e a 19., a 20., vagy akár a 21. században. Pennebaker megállapításai ellenére elképzelhető, hogy nem léteznek a későbbiekben ezek a különbségek, mivel a 18. századi, szigorúan behatárolt nemi szerepek oldódása jelentős az elkövetkező évszázadokban, és ennek nagy része az irodalomnak köszönhető (persze nem hagyható figyelmen kívül az irodalom lelkes szerepvállalása a sztereotípiák megerősítésében sem). A problémát súlyosbítja az a jelenség, amivel már ebben a cikkben is találkoztunk – azaz a tény, hogy a nemek nyelvhasználatát előzetesen leíró szólisták kudarcot vallanak, ha nem az adott vizsgálati anyagból származnak. Tehát problémás lehet, hogy egy 18. századi regényekből

 DOI: 10.31400/dh-hun.2021.5.3143

kinyert szólistának lehet-e bármi jelentősége később keletkezett művek szerzőinek nemi azonosításakor a megváltozott társadalmi és történelmi feltételek között egy kvantitatív vizsgálat során.

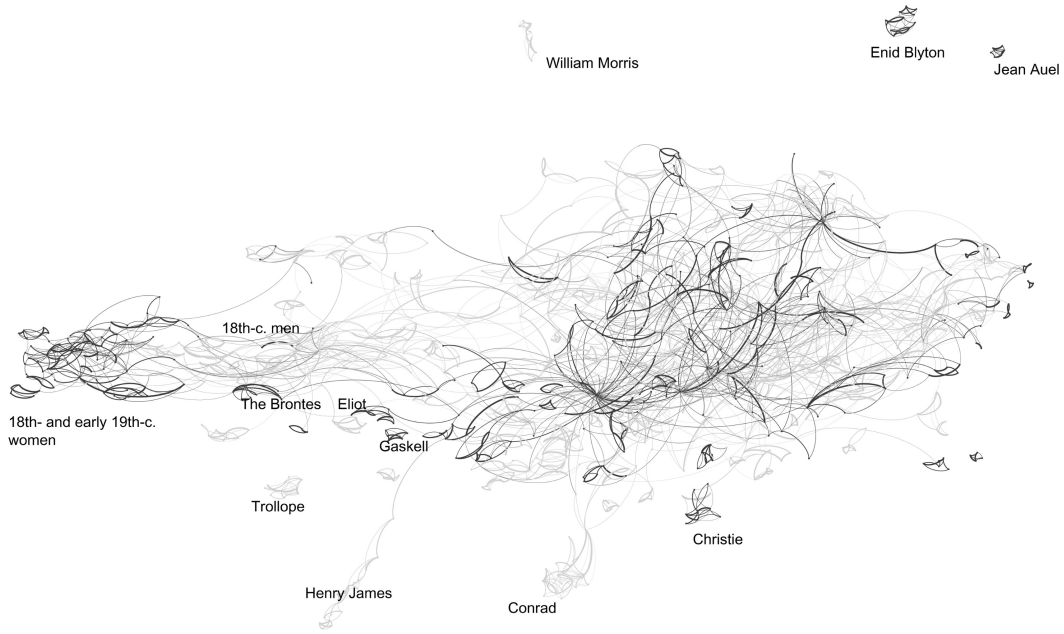
Hogy eloszlassuk vagy megerősítsük ezeket a félelmeket, egy 1000 kötetes korpuszon végeztem egy sor hálózatelemzést. Összesen valamivel több mint 111 millió tokenből áll a 635 férfi és 365 női szövegét felölelő korpusz, amely magában foglalja az összes fent vizsgált művet, amelyek jól reprezentálják a 18. és 19. századi regényt, továbbá még nagyobb arányban tartalmaz regényeket a 20. és 21. századból. Az első hálózat arra keresi a választ, hogy egy ilyen kiterjesztett korpuszban a leggyakoribb szavak önmagukban biztosítják-e a nemek azonosíthatóságát. A kvázi-egydimenziós diagramban makroszinten a kronologikus jegyek (a legkorábbi szövegektől a legfrissebbekig, azaz ebben az esetben: balról jobbra), mikroszinten pedig a szerzőség szempontjai (a szövegcsoportok szerzők szerint) dominálnak. Az előbbi valóban igen erős tényező: az első csoport, balról, magába foglalja Jane Austent és a többi 18. századi női szerzőt (kanonizáltakat és nem kanonizáltakat egyaránt), néhány átfedéssel a szomszédos férfi írók csoportjával, ugyanabból a századból. A nemi megoszlás azonban már nem észlelhető a későbbi századokban, ahogy azt a férfiakat mutató világosszürke és a nőket mutató sötétszürke időtengelyen végigvonuló mintázatai mutatják. A női szerzőknek csak egy kisebb csoportosulása figyelhető meg, ami megfelel a Woolf, Hall, West és Mansfield által fémjelzett modernitásnak, de több férfi modernista szerző is fellelhető a környékükön. Néhány kiugró érték szintén megemlíthető, amelyek kapcsán egy eltökélt kommentátor megkockáztathatná a megjegyzést, miszerint, ami elkezdődött az angolszász női irodalomban Jane Austennal, az most E. L. Jamesszel végződik…



6. ábra. A hálózatelemzés során létrejött gráf 1000, férfiak (világosszürke) és nők (sötétszürke) által írt regényhez, a 100–1000 leggyakoribb szóra vonatkozóan.

Hogy megbizonyosodjunk a 18. században a nemhez kötődő, kulcsszavakon alapuló nyelvhasználati különbségek érvényességéről, ellenőriznünk kell ezt a hipotézist az 1000 regény esetében is *(7. ábra)*. Vészjósló, hogy nem sok minden történik, a grafikon csak vertikálisan növekszik. Az alaposabb vizsgálat a 18. századi férfi szerzőknek valamivel nagyobb távolságát mutatja Austentól, Burneytől és a Chawton House regényeitől. Látható talán egy jobban elkülönülő evolúciós ív is a Brontëktól kezd-

ve George Elioton és Gaskellen keresztül, de aztán a női írók közötti sötétszürke kapcsolatok elkezdenek beleolvadni a férfi írók világosszürke tengerébe, és tartósan együtt haladnak tovább a 21. század felé. Ez azt jelzi, hogy a 18. századi, nemiséghez kötődő szavak egyszerűen elavulttá váltak, és kevésbé hasznosak, mint ha pusztán a leggyakoribb szavakat vesszük alapul a szerzők nemének feltérképezéséhez. Továbbá érdekes módon az előző hálózat több kiugró értéke megjelenik ezen az ábrán is. Tisztán látszik, hogy ehhez a fajta analízishez lényegtelen, hogy mely szavakból indulunk ki, amíg elég sok van belőlük – Maciej Eder és jómagam néhány korábbi megállapításának igazolására.[33]



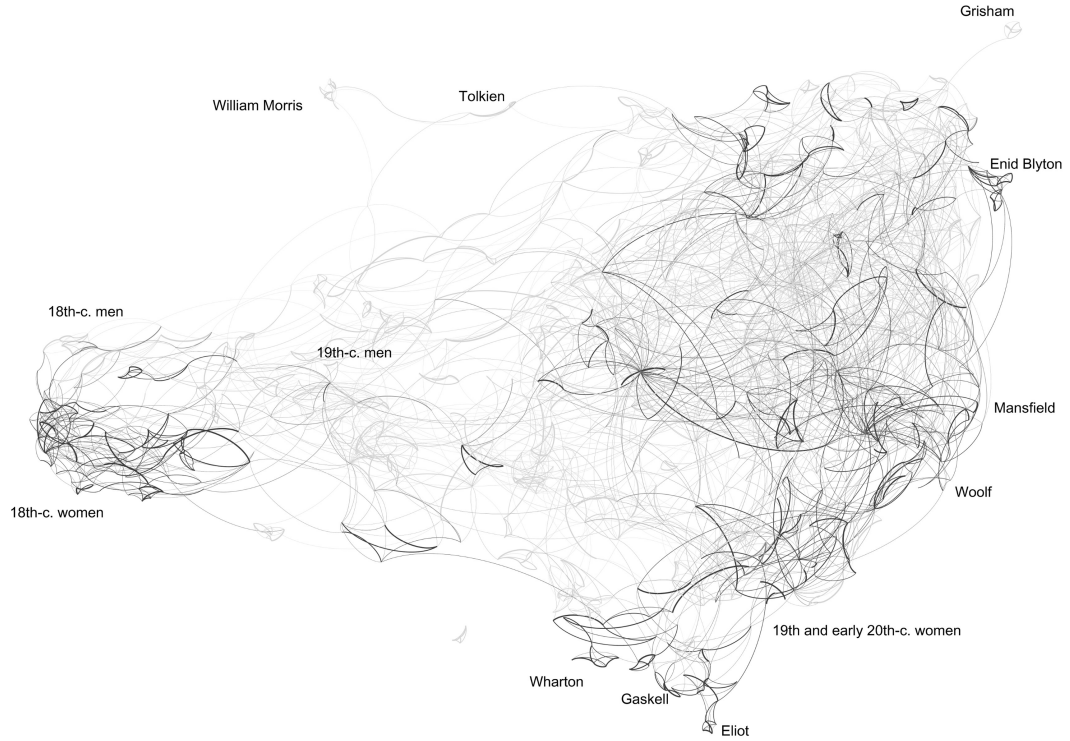
7. ábra. A hálózatelemzés során létrejött gráf 1000, férfiak (világosszürke) és nők (sötétszürke) által írt regényhez, 18. századi „híres férfiak”/„híres nők” Zeta-kulcsszavai alapján.

A társadalmi nemek szempontjából némileg frissebb szólista összeállításához egy olyan százszöveges alkorpuszt alkalmaztam, amely 67 férfi és 33 női szerzőtől származó, 1839 és 1939 között létrejött szöveget tartalmazott. Ebben az esetben is mindkét nemre előállíthatók a Zeta-szavak. A női szerzőknél kimutatott szavak az elődeikhez képest sokkal komplexebb készletet alkotnak. Ezek közé tartoznak az érzékeléssel kapcsolatos és kognitív kifejezések: *figyelni, nézte, nézni, tűnődött, megállt, magába szívta, tudatosság, tudatos, kifejezés;* a negatív és pozitív állapotok és érzelmek: *érzések, kellemes, gyengéd, bonyolult, széles, csöndesen, szenvedve, összezár, szenvedélyes, merészelt, kedvelte, napfény, szeretőn, felemelt, elfoglalt, felriadt, ragyogott, hajlított, erőfeszítés, sápadt, találó;* az olvasás szavai: *olvasni, könyvek;* a „feminin” vagy háztartási tárgyak: *selyem, virágok, rózsák, fürtök, szék;* és a színek: *karmazsin, barna.* Ezzel ellentétben

[33] Jan Rybicki and Maciej Eder, „Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?” *Literary and Linguistic Computing* 26, 3. sz. (2011): 315–321, https://doi.org/10.1093/llc/fqr031.

 DOI: 10.31400/dh-hun.2021.5.3143

a férfiak listája még sztereotipabbá vált; egyre inkább tükrözte a korszak zűrzavaros politikai hangulatát, dominálnak benne a főnevek: *ellenség, becsület, elnézést, csata, kapitány, tiszt, kard, lő, lövés, harc, hadsereg, tisztek, elesett, fegyveres, füst, harag;* előfordulnak számok és/vagy pénz: *ezer, tucat, ötven;* hirtelen az alsó tagok kezdik uralni az emberi testet: *lábak, sarkak;* végre a nőket *nőneműként* emlegetik; a trágár beszéd egyre szókimondóbb: *káromkodás, káromkodott, ördög, átkozott;* és gyakoriak az ivással kapcsolatos kifejezések: *palack, részeg*. Az így létrejövő hálózat *(8. ábra)* még a korábbinál is jobban eltér a leggyakoribb szavakon alapuló lineáris sorrendtől, és megtöri a kronológiai sorrend hegemóniáját azzal, hogy a 19. és a 20. század elejei írónőket perifériára helyezi (az ábra közepén, alul). Fontos megjegyezni, hogy az ábra más szövegeket is felhasznál, nem csak a szólisták létrehozásához alkalmazottakat.



8. ábra. A hálózatelemzés során létrejött gráf 1000, férfiak (világosszürke) és nők (sötétszürke) által írt regényhez, az 1839–1939 között alkotó férfiak és nők Zeta-kulcsszavai alapján.

Van még egy fontos különbség a férfi és női irodalmi nyelv között, azonban ez csak az angolnál ragozóbb nyelvekben válik nyilvánvalóvá. Volt ugyanis egy hasonló kísérlet lengyel regényeken egy ugyanígy 100 szövegből álló korpuszon, amely, bár hasonló szemantikai kategóriákat eredményezett a nemeket illetően, egy másik szóosztály ugyanakkor tovább erősítette a különbségeket: a múlt idejű igealakok hím- és nőnemű személyraggal. Hiszen a 300 Zeta-szó listájára csak azok az igék kerültek rá, amelyek a megfelelő nemű toldalékkal fordultak elő a 19. és 20. század eleji szövegekben. Egy ilyen jelenség önmagában nem meglepő, hiszen a férfi perspektíva dominanciája a férfi írók regényeiben és a női perspektíva dominanciája a nők írásaiban nemcsak észszerű

elvárás, hanem egy jól megalapozott irodalmi tény is. A 19. századi történelmi románcok hősnői nagyon keveset beszélnek a férfi hősökhöz képest; míg Jane Austennál soha nem fordul elő egy jelenetben két férfi nő nélkül. Ami *tényleg* meglepő, az a jelenség nagysága: a második nemnek mindig nagyon kevés dolga vagy mondanivalója van az egyes szerzők műveiben.

4. Konklúzió

Úgy tűnhet, hogy a tanulmány elérte elsődleges céljait. A nemek szógyakoriság szerinti azonosításának két különböző módja meglehetősen következetesen rámutatott két lehetséges gyanúsítottra az esetlegesen előforduló férfiak keresésében az anonim szerzők között a 18. és a 19. század eleji angol női szövegek Chawton House korpuszában. Főként a *The Imposters Detected: or, the Life of a Portuguese* című szöveg tűnik gyanúsnak, hiszen a részben pikareszk, részben katolikusellenes szatíra eltér a gyűjteményt uraló szentimentalista művektől. Ha a *The Imposters* tényleg egy férfi imposztor műve, akkor bizony elkövetett egy apró, de leleplező baklövést, nem is annyira a női szövegekre jellemző funkciószó- vagy a kulcsszóhasználat utánzásában, mintsem a történet előszavában elejtett, igencsak árulkodó mondatban: „Csakis a *nőies (womanish)* és gyenge lelkek sértődnének meg az ebben a könyvecskében szereplő történeteken". Más szövegekben is előfordul ez a durva dőlt betűs szó az 1000-es korpuszban, ám csak a *The Imposters* volt annyira szemtelen, hogy a szerzői előszóban közvetlenül ezzel illesse az olvasót. A többváltozós elemzés, a távoli olvasás, az előszó egyetlen szava, illetve a szoros olvasás és ezek kombinációi is mind gyanúba keverik a névtelen szerzőt. Ezzel a könyvvel egyébként más kétes dolog is akad. A szerkesztő ugyanis úgy tesz, mintha a történet a Padovában talált kézirat francia fordítása lenne. Nem mintha ez ritkaság volna azokban az időkben: az angol gótikus regények fele fordításnak adja ki magát. Az viszont már érdekesebb, hogy maguk a franciák nem vállalják a felelősséget, amikor az *Annales typographiques ou notice du progrès des connoissances humaines* című, Párizsban kiadott 1760-as munka második része a könyv eredeti angol címét adja meg, és egy kíméletlenül őszinte kommentárt fűz hozzá: „Az összes kiadvány közül, amelyek a legújabb portugál ügyek alapján születtek, nem volt rosszabb, mint az, amelynek a címét most olvasták."[34]

A másik rendszeres gyanúsított, a *The Reward of Virtue; or, the History of Miss Polly Graham* némileg titokzatosabb, mert esetében nincs sokatmondó előszó és más egyértelmű nyom. Legalább a világképe valamivel konstruktívabb, hiszen a mű utolsó fejezete egy elég hasznos intézményt mutat be: a Bounty Hall olyan hely, ahol „egy hölgyekből álló társaság, miután figyelembe vette azokat a kellemetlenségeket, amelyek sok erényes háziasszonyt az elkerülhetetlen szerencsétlenségek következtében a szegénységbe taszítottak, nagylelkűen úgy döntött, hogy menedéket nyújt ezeknek a boldogtalan személyeknek". Sajnos ezt a nemes írást a *The Monthly Review, or*

[34] „De toutes les brochures auxquelles les dernieres affaires du Portugal ont donné lieu, il n'y en a pas eu de plus mauvaise que celle dont on vient de lire le titre," *Annales typographiques ou notice du progrès des connoissances humaines* (Paris: Vincent, 1760), vol. 2., 261.

 DOI: 10.31400/dh-hun.2021.5.3143

Literary Journal névtelen kritikusa mint „valószínűtlen és összefüggéstelen mesék zűrzavarát”[35] utasította el.

A cikk második kérdésére – azaz hogy mi a különbség a Chawton House női írói és szerencsésebb riválisaik, például Austen vagy Shelley között – olyan választ érdemes adni, amely értelmezhető az elmúlt fél évszázad kánonháborúival összefüggésben is. Érdekes ugyanis, hogy az ebből a stilometriai elemzésből is kirajzolódó nézet (miszerint a nők akkor válhatnak a kánon részévé, ha minél inkább úgy írnak, mint a férfiak) hogyan kapcsolódhat a Harold Bloom által védett nyugati kánon és az általa „neheztelés iskolájának” nevezett elmélet közötti huzavonához. A hagyományos irodalomtudomány talán még soha nem hangoztatta olyan egyértelműen, mint a kvantitatív kutatás, hogy a „férfiként író” és „nőként író” fogalmak olyan állandó változásnak vannak kitéve, ami alapján nem használhatjuk azokat problémamentesen. Ezt sugallja az is, hogy a tanulmány képtelen volt olyan stabil „kánont” létrehozni a férfi és női kulcsszavakból, amelyek átívelnének a korpusz változásain vagy az irodalmi evolúció mozgásain; és éppolyan kevés sikerrel tudott statisztikai elemzéssel ilyen szavakat találni, mint előre meghatározott listákkal és kategóriákkal próbálkozva.

Ugyanakkor a kulcsszavak kiszűrése hasznosnak bizonyul a hagyományos irodalomtörténet szempontjából, és felhasználható a gyakran igen gyanús szózsákmodell igazolására is: úgy tűnik, ezzel megbízható eredmények tárhatók fel, amíg megalapozott a statisztika és a megfelelőek a módszerek. Hasznos lehet továbbá, hogy nyomon kövessük a kulcsszavak időbeli változását – ezt bizonyítja a genderszenzitív Zeta-szavak eltolódása a szentimentalizmus szókincsétől – ami a Chawton House korpuszt olyan egységes szöveggyűjteménnyé teszi – a női írásokat egy évszázaddal később meghatározó, sokkal kevésbé egyoldalú szavak gyűjteménye felé. Mindeközben az egyoldalúságtól a komplexitás felé zajló folyamat a visszájára is fordulhat, ahogyan az a férfiak kulcsszavaiból is kitűnik: a *Twist Olivér,* a 100 regényt tartalmazó korpusz legkorábbi művének megjelenése, és e gyűjtemény határa (1939) között a kulcsszavak különböző szemantikai kategóriáit kétségtelenül eluralta az akkor dúló háborúk (a búr háborúk, az első és a már érezhető második világháború) szókincse.[36]

Az teljesen világosan látszik ebből a vizsgálatból, hogy a szerző neménél nagyobb hatalmak befolyásolják az irodalmi szókincs alakulását. A legfontosabb ezek között az idő, ami John Burrows másik klasszikus tanulmányának, a *Tiptoeing into the Infinite*nek is a hőse: az időbeliség mint megkülönböztető jegy még akkor is felbukkan, hívatlanul, ha az elemzés középpontjában a társadalmi nem (vagy igazából bármi más)

[35] *The Monthly Review, or Literary Journal* 41 (1769): 479.

[36] Ezt össze kellene hasonlítani Jane Austen regényeiben a külvilágban zajló események látszólagos elkerülésével (szemben levelezéseivel). Austen műveinek kulcsszavai nem hozhatók összefüggésbe a napóleoni háborúkkal – kivéve a *Meggyőző érvek* című regényt. Ezen az egyetlen regényen kívül a háború soha nem kerül előtérbe, nem úgy, mint rengeteg más műben, amelyet a konfliktus ideje alatt írtak a kontinensen. Ahogy Austen császári (vagy gyarmati) összefonódásait is csak a közelmúltban fedezték fel (vö. Edward Said, *Culture and Imperialism* [London: Chatto and Windus, 1993], 80–96.), a francia hírszerzésnek csak egy nagyon hozzáértő tisztje ismerné fel, hogy Wickham ezredének Longbourne-ből Brightonba történő eltávolítása, amely a *Büszkeség és balítélet*ben zajlik, azt sugallja, hogy a *Galád Albion* egészen másfajta pusztítást tervezhet a La Manche kontinentális oldalán.

áll.[37] Jockers úgy becsüli, hogy az „évtized" kategóriájának hatása 14%-nyi a szövegek eltérésében.[38] Vajon az idő, talán általánosabban kezelve, nem sokkal-sokkal fontosabb ennél? Azt azonban meg kell hagyni, hogy van egy érdekes kettősség az irodalmi nyelv esetében az időbeliség kapcsán, hiszen az éppúgy kifejtheti a hatását az egyetlen szerző életművét tartalmazó, mint a hosszabb időt és több szerzőt felölelő korpuszokban; márpedig mindkét jelenséget nem lehet a nyelvi változásnak ugyanazzal a mechanizmusával magyarázni. De ez már egy egészen más történet – amihez vissza kellene fordítani Kipling *A dzsungel könyve* utolsó sorának bowdlerizált lengyel fordítását az eredetire.

Fordította: Szlávich Eszter és Matis-Zöllner Anna

Vive la différence! Tracing the (Authorial) Gender Signal by Multivariate Analysis of Word Frequencies

Multivariate analysis of word frequencies is used to identify the gender of authors in a corpus of 18th and early 19th century English sentimentalist and Gothic fiction. Results obtained with most frequent words are compared to those produced with medium-frequency Burrows's Zeta words characteristic for both genders. Gender-sensitive words from two periods (18th/19th century and 19th/20th century) are compared in terms of their usefulness for gender identification in literary texts.

Keywords:
authorship attribution, gender of author, Bootstrap Consensus Network, Burrows's Zeta

[37] John Burrows, „Tiptoeing into the Infinite: Testing for Evidence of National Differences in the Language of English Narrative," in Susan Hockey and Nancy Ide, eds., *Research in Humanities Computing 4: Selected Papers from the 1992 ACH/ALLC Conference,* 1–33 (Oxford: Clarendon Press, 1996).

[38] Jockers, *Macroanalysis,* 96.

 DOI: 10.31400/dh-hun.2021.5.3143