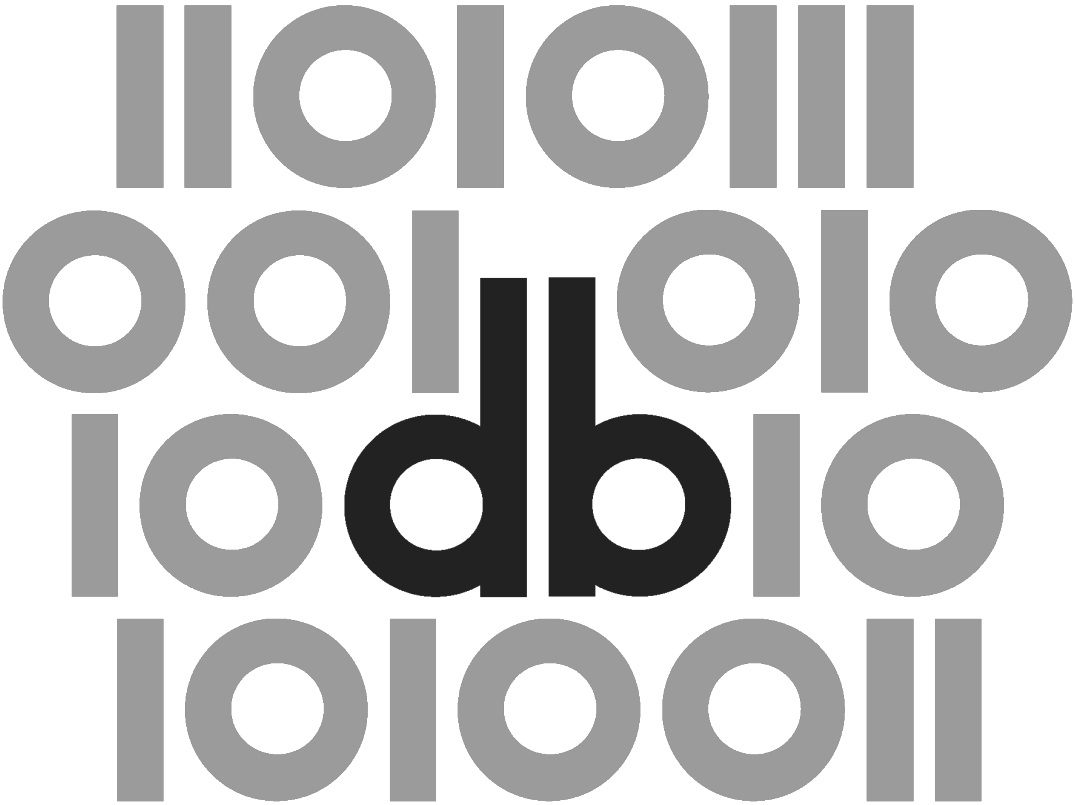
5 (2021) <DIGITÁLIS BÖLCSÉSZET>

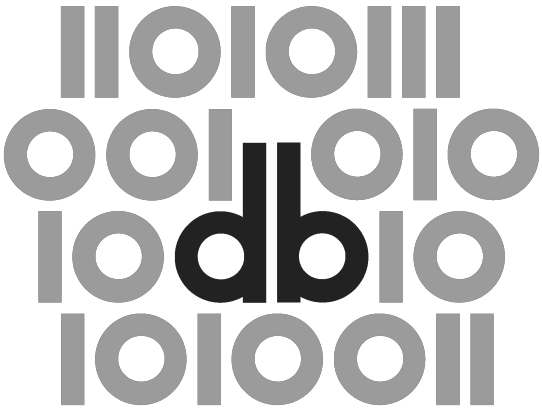
A krakkói Computational Stylistics Group (Különszám)



5 (2021) </DIGITÁLIS BÖLCSÉSZET>

Digitális Bölcsészet
2021., ötödik szám
A krakkói Computational Stylistics Group (Különszám)

<DIGITÁLIS BÖLCSÉSZET>



5 (2021)

A krakkói
Computational Stylistics Group

(Különszám)

A különszámot Szemes Botond szerkesztette.

Felelős szerkesztő:
Maróthy Szilvia

Szerkesztőség:
Kokas Károly, Parádi Andrea

Rovatvezetők:
Tanulmányok: Kiss Margit
Műhely: Péter Róbert
Kritika: Almási Zsolt
Labor: Mártonfi Attila

Tanácsadó testület:
Bartók István, Fazekas István, Golden Dániel, Horváth Iván, Palkó Gábor, Pap Balázs, Sass Bálint, Seláf Levente

Korábbi munkatársaink:
Bartók Zsófia Ágnes (szerkesztő, rovatvezető), Fodor János (szerkesztő), †Labádi Gergely (szerkesztő, rovatvezető), †Orlovszky Géza (tanácsadó testület)

ISSN 2630-9696
DOI 10.31400/dh-hun.2021.5

Kiadja a Bakonyi Géza Alapítvány és az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A).
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

Megjelenik az Open Journal Systems (OJS) v. 3. platformon, melynek működtetését az ELTE Egyetemi Könyvtár- és Levéltár biztosítja.

A különszám megjelenését a Wacław Felczak Alapítvány támogatta.

Honlap: `http://ojs.elte.hu/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Olvasószerkesztő: Bucsics Katalin
Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

Tartalom

<ELŐSZÓ>

Joanna Byszuk 0000-0003-2850-2996

Instytutu Języka Polskiego PAN

joanna.byszuk@ijp.pan.pl

Szemes Botond 0000-0002-0637-6776

ELTE BTK Irodalomtudományi Doktori Iskola; ELTE BTK Digitális Bölcsészet Tanszék

boboszemes@gmail.com

A krakkói Computational Stylistics Group bemutatkozása

Előszó a *Digitális Bölcsészet* folyóirat tematikus lapszámához

A „stilometria” fogalmát a világhírű lengyel filológus-filozófus-nyelvész, Wincenty Lutosławski alkotta meg a 19–20. század fordulóján, amikor Platón dialógusainak időrendjét a bennük előforduló nyelvi elemek gyakoriságvizsgálatán keresztül kívánta meghatározni.[1] Módszerének sok követője akadt, ám a valódi előrelépést a számítástechnika elterjedése hozta el a területen: a nyelvi elemek statisztikai vizsgálatát innentől kezdve már nagy korpuszokra kiterjesztve és ellenőrzött módon lehet megvalósítani. A számítógépes stíluskutatás legjelentősebb képviselői közül sokan – Lutosławskihoz hasonlóan – szintén Lengyelországból érkeznek, akik elsősorban a krakkói székhelyű Computational Stylistics Group kutatócsoportjába tömörülnek.[2] Az alábbi folyóiratszám e kutatócsoport munkájába nyújt bepillantást a tagok által írt tanulmányok fordításain keresztül.

A stilometria – a Magyarországon is nagy hagyományokkal rendelkező kvantitatív stilisztikához hasonlóan – egy meglehetősen egyszerű elképzelésből indul ki: egy szerző, egy mű vagy egy szövegcsoport stílusa meghatározható a rá jellemző vagy nem jellemző összetevői (pl. szavak, nyelvtani szerkezetek) alapján; így, ha ezeket megfelelő módon azonosítani tudjuk, akkor az előfordulásokat összeszámolva a szövegek stílusa kvantifikálható és összehasonlítható lesz, valamint eddig nem reflektált

[1] Wincenty Lutołsawski, *The Origin and Growth of Plato's Logic: With an Account of Plato's Style and of the Chronology of His Writings* (London: Forgotten Books, 2018 [1890]).

[2] A kutatócsoport egy intézményközi szervezet, amelynek tagjai leginkább a Lengyel Nyelvtudományi Intézetben (Lengyel Tudományos Akadémia) a Jagelló Egyetemen és az Antwerpeni Egyetemen működnek. A kutatócsoport honlapja, hozzáférés: 2021.12.13, `https://computationalstylistics.github.io/`.

tulajdonságaik is láthatóvá válhatnak.[3] Hiszen egy ilyen gyakoriságvizsgálat a nyelvi működés más szintjére vonatkozik, mint az olvasás művelete – térbeli metaforával élve a szövegek „mélystruktúráira" irányul. A stilometriai kutatások eredményei ennek megfelelően a szerzőazonosítás[4] területén érték el a leginkább szembetűnő sikereket; a számítógépes kapacitást kihasználva ugyanis bizonyíthatóvá vált, hogy minden szerzőre (de akár korszakra, műfajra, vagy más szövegcsoportra) egyénileg jellemző az általa használt szavak eloszlása: azaz a leggyakoribb[5] szavak előfordulását mérve elkülöníthetők az egyes alkotók szövegei egymástól – aminek segítségével az ismeretlen szerzőségű művek írója is meghatározható lehet.[6] Ebből is látható, hogy mit jelent a „mélystruktúra" megjelölés, hiszen egy szövegben a leggyakrabban a konkrét jelentést nélkülöző úgynevezett funkciószavak (pl. névelők, kötőszavak) fordulnak elő, így ezek eloszlása nem a művek szemantikai karakteréről értesít bennünket, hanem az adott íróra jellemző nyelvhasználatról, vagyis – újabb beszédes metaforát alkalmazva – a „szerzői ujjlenyomat" jelenlétéről.[7]

Ez az egyszerű elképzelés akkor válik összetetté, ha átültetjük a gyakorlatba, és rákérdezünk arra, hogy milyen nyelvi elemeket, milyen korpuszokon és hogyan érdemes azonosítani, illetve összeszámolni a statisztikai alapú stíluskutatás során. A Computational Stylistics Group munkássága innen nézve válik megkerülhetetlenné, hiszen ennek előterében a módszertani kísérletezés, az optimális működés megtalálása és más kutatók számára erre vonatkozó ajánlások megfogalmazása áll. Ennyiben pedig a digitális bölcsészet alapvetően kísérletező jellegét erősíti meg és teszi látványossá, amely talán a diszciplína egyik legfontosabb sajátosságaként jelölhető meg.[8] Ennek

[3] Ugyan a stilometria a szövegelemzés területén alakult ki, meg kell jegyezni, hogy manapság más médiumok és művészeti ágak (pl. zene, színház, film) esetében is alkalmazzák a fent kifejtett módszert. A lapszám ugyanakkor olyan tanulmányokat közöl, amelyek az irodalom- és nyelvtudomány viszonyában hasznosítják a statisztikai alapú stíluskutatás belátásait.

[4] A korábbi magyar nyelvű tanulmányokat követve ezt a kifejezést használjuk a lapszámban. Ez azonban annyiban különbözik a nemzetközi szakirodalomban használatos angol *authorship attribution* megjelöléstől, hogy annak 'szerzőség hozzárendelés' jelentése magában foglalja, hogy az eredmények nem az 'azonosítás' bizonyosságára, hanem a valószínűségek között mozgó kutatás konstrukcióira vonatkoznak.

[5] Általában azért a leggyakoribb szavak vizsgálata kerül az előtérbe, mivel ezek biztosítanak statisztikailag elegendő mennyiségű adatot a következtetések számára.

[6] Ennek a módszernek az alapítószövege: 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.

[7] Ehhez lásd Maciej Eder, „Style-Markers in Authorship Attribution: A Cross-Language Study of the Authorial Fingerprint," *Studies in Polish Linguistics* 6 (2011): 99–114.

[8] A digitális bölcsészet ilyen irányú meghatározását lásd Ted Underwood, „A Genealogy of Distant Reading," *DHQ* 11, 2. sz. (2017), http://www.digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

 DOI DOI 10.31400/dh-hun.2021.5.3147

megfelelően kísérleteket végeztek, hogy milyen,[9] illetve hány darab[10] nyelvi elemen keresztül és legkevesebb hány szóból álló szövegeket vizsgálva[11] érhető el legjobban a szövegek/szerzők megkülönböztető stílusa, milyen metrikák segítségével ragadható meg ez a különbség,[12] hogyan érdemes a kiinduló tanítókorpuszt létrehozni,[13] és milyen vizuális megjelenítés tudja a szövegek közötti kapcsolatokról a legtöbb információt láthatóvá tenni.[14] Ezek a kísérletek két szempontból is alapvető fontosságúak. Egyrészt mivel a kortárs stilometria a digitális bölcsészet részeként nem csupán a kvantitatív stilisztika hagyományához kapcsolódik, hanem erősen támaszkodik az adattudomány és a statisztika eljárásaira is – a hivatkozott tanulmányok pedig megkerülhetetlen szerepet játszanak abban, hogy ezeket az eljárásokat ne pusztán átvegyék a bölcsészeti érdekeltségű projektek, hanem azokat megértve saját céljaikhoz legyenek képesek igazítani. Másrészt az említett vizsgálatokat nemcsak a módszereknek, hanem magának a nyelvi működésnek a jobb megértése is motiválja, azaz hogy milyen olyan eloszlások, törvényszerűségek, történeti alakulások figyelhetők meg az irodalmi és hétköznapi nyelvhasználatban a kvantitatív megközelítések segítségével, amelyek korábban nem voltak láthatók – és ezek hogyan köthetők a kvalitatív megközelítések eredményeihez.

Fontos megemlíteni, hogy a kutatócsoport nem csupán tanulmányok formájában kommunikálja eredményeit, hanem nagy hangsúlyt fektet eljárásaik és szemléletük tanítására is. Számtalan workshop, nyári egyetem,[15] blogbejegyzés[16] és a módszerek elsajátítását segítő anyag kifejlesztése mellett gyakran fogadnak vendégeket szemé-

[9] Például Mike Kestemont, „Function Words in Authorship Attribution: From Black Magic to Theory?" in Anna Feldman, Anna Kazantseva and Stan Szpakowicz, eds., *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL),* 59–66 (Gothenburg: Association for Computational Linguistics, 2014), http://doi.org/10.3115/v1/W14-0908. 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.

[10] Jan Rybicki, „Reading Novels with Statistics: What Numbers of Words Tell Us about Authorship, Genre, or Chronology," in John August Dobelman, ed., *Models and Reality: Festschrift For James Robert Thompson,* 207–224 (Chicago: T&NO Company, 2017).

[11] Maciej Eder, „Does Size Matter? Authorship Attribution, Small Samples, Big Problem," *Digital Scholarship in the Humanities* 30, 2. sz. (2015): 167–182, https://doi.org/10.1093/llc/fqt066. Maciej Eder, „Short Samples in Authorship Attribution: A New Approach," in Rhian Lewis, Cecily Raynor, Dominic Forest, Michael Sinatra and Stéfan Sinclair, eds., *dh2017: Digital Humanities 2017, Conference Abstracts, McGill University & Université de Montréal, Montréal, Canada, August 8–11, 2017,* 223 (Montréal: Alliance of Digital Humanities Organizations [ADHO], 2017).

[12] Ennek kapcsán fontos megemlíteni Burrows *Delta*jának a flektáló nyelvekre kialakított változatát, Eder *Delta*ját. Ezeknek elemzéséhez lásd Fotis Jannidis et. al., „Improving Burrows' Delta – An Empirical Evaluation of Text Distance Measures," in *Book of Abstracts of the Digital Humanities Conference 2015, ADHO* (Sydney, UWS, 2015).

[13] Maciej Eder and Jan Rybicki, „Do Birds of a Feather Really Flock Together, Or How to Choose Training Samples for Authorship Attribution," *Literary and Linguistic Computing* 28, 2. sz. (2012): 229–236, https://doi.org/10.1093/llc/fqs036.

[14] Maciej Eder, „Visualization in Stylometry: Cluster Analysis Using Networks," *Digital Scholarship in the Humanities* 32, 1. sz. (2017): 50–64, https://doi.org/10.1093/llc/fqv061.

[15] A legnagyobbakat említve: Digital Humanities Summer Institute (hozzáférés: 2021.12.13, https://dhsi.org) és European Summer University in Digital Humanities (hozzáférés: 2021.12.13, https://esu.fdhl.info/).

[16] Hozzáférés: 2021.12.13, https://computationalstylistics.github.io/blog/.

lyesen is, hogy ismereteiket első kézből tudják átadni az érdeklődő kutatók számára.[17] Ezen kívül a csoport „DH Lunch" címmel saját előadás-sorozatot szervez, amelynek keretében a világ különböző pontjairól érkeznek előadók, hogy a stilometria eszköztárának egyéni felhasználásairól számoljanak be.[18] Tevékenységeik közül azonban minden bizonnyal az R programozási környezetre kifejlesztett, *Stylo* programcsomag[19] gyakorolta a legnagyobb hatást a digitális bölcsészeti közösségre. A csomag magába foglalja a stilometriai kutatás minden lépését a korpuszok előkészítésétől a gyakoriságvizsgálat paramétereinek beállításán és az eredményeken végzett statisztikai eljárásokon át a szövegek hasonlóságának/különbségének kiszámításáig és a viszonyok különféle vizuális megjelenítéséig. Ennek kifejlesztésekor szintén elsősorban a felhasználóbarát, széles körű használat szempontjai érvényesültek, amelynek eredményeképp a *Stylo* kifejezetten elterjedté vált a szerzőazonosításra és más, szövegek stílusára irányuló digitális bölcsészeti projektekben.[20] Ezt erősíti tovább az oktatóanyagok fejlesztése és a stilometria területén dolgozó kutatóközösséggel tartott szoros kapcsolat: az említett fórumokon kívül érdemes még megemlíteni a programcsomag működése kapcsán felmerült kérdések számára kialakított felületet,[21] valamint az új felhasználóknak szóló oktatóvideókat is.[22]

A folyóiratszámban közölt tanulmányok ugyanakkor túlmutatnak az említett alapkutatásokon, amennyiben konkrét – elsősorban irodalom- és nyelvtudományi – kérdések megválaszolását tűzik ki célul. Azonban ezek a kutatások is kísérletként valósulnak meg, hiszen rendszerint különböző módszerek teljesítményét ütköztetik egymással, sőt az eredményeket is inkább mint valószínűségeket prezentálják, jelezve az azokból levonható következtetések határait. Miközben ez a tudatosság természetesen nem gátolja meg a szerzőket abban, hogy óvatos, de alapvető irodalom- és kultúrtörténeti összefüggéseket vázoljanak fel a kísérletek alapján.

Az első három tanulmány szorosan összetartozik és egy külön blokkot alkot a lapszámon belül. Mindegyik szöveg a szerzőazonosítás témaköréhez kapcsolódik, ugyanakkor annál tágabb horizontot fognak át, mivel kérdéseik nem csupán az ismeretlen szerzőségű művek alkotójának azonosítására irányulnak. Maciej Eder tanulmánya az Elena Ferrante álnéven megjelentetett, nemzetközi sikert aratott regényeket[23] vizsgál-

[17] Így került a kutatócsoporttal kapcsolatba jelen folyóiratszám vendégszerkesztője és az előszó társszerzője, Szemes Botond is.

[18] A rögzített előadások szintén szabadon megtekinthetők, hozzáférés: 2021.12.13, `https://www.youtube.com/channel/UCQfYhxastnHg6jiZU-H3PLA/videos`.

[19] Maciej Eder, Jan Rybicki and Mike Kestemont, „Stylometry with R: A Package for Computational Text Analysis," *R Journal* 8, 1. sz. (2016): 107–121, `https://doi.org/10.32614/RJ-2016-007`.

[20] A csomag hazai fejlesztésű, webes alkalmazása *Shtylo* néven érhető el: Dobi Jan Sándor, Mészáros Tamás és Kiss Margit, „*Shtylo:* Stilometriai elemzések webes támogatása," in Vincze Veronika, szerk., *XIV. Magyar Számítógépes Nyelvészeti Konferencia,* 423–436 (Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2018). Ez utóbbi fejlesztést alkalmazta Kiss Margit, a folyóirat szerkesztője is tanulmányában: Kiss Margit, „Stilometriai elemzés lehetőségei magyar történeti szövegkorpuszon," *Digitális Bölcsészet* 2 (2019): 15–33.

[21] Hozzáférés: 2021.12.13, `https://groups.google.com/g/computationalstylistics`.

[22] Hozzáférés: 2021.12.13, `https://www.youtube.com/watch?v=Rv7u4UNZJrA`.

[23] Amelyek közül már nagy költségvetésű filmadaptáció is készült: éppen idén szeptemberben a Velencei Filmfesztiválon került bemutatásra *Az elveszett gyerek története* című regényen alapuló játékfilm.

 DOI DOI 10.31400/dh-hun.2021.5.3147

ja, és habár módszereivel komoly érveket szolgáltat az írói név mögött rejlő valódi szerző kilétére vonatkozóan, a kutatás fókuszában mégsem egyszerűen a szerzőség kérdése áll. Jobban érdekli Edert, hogy miután azonosíthatóvá vált a művek valódi szerzője, észlelhető-e a két néven megjelentetett szövegek között stiláris különbség. A leggyakrabb szavakon alapuló vizsgálatai azt bizonyítják, hogy ugyan az álnéven írt szövegek is tartalmazzák valódi írójuk „ujjlenyomatát", mégis elkülöníthető egymástól a kétféle írói identitás. Külön kiemelendő a dolgozat módszertani sokszínűsége: míg a szerzőazonosításhoz makroperspektívát érvényesít, és 150 olasz regény hálózatát rajzolja fel, addig az egy személyhez tartozó, de különböző írói identitások vizsgálatakor a szövegek apró részleteit elemzi az úgynevezett Rolling Classify (szintén a *Stylo*-csomagba implementált) módszerével.

Jan Rybicki tanulmánya ugyanígy a szerzőazonosítás kérdéséből indul ki: 18–19. századi angol írónők korpuszában az ismeretlen szerzőségű szövegek alkotójának meghatározására tesz kísérletet, valamint annak ellenőrzésére, hogy nem „keveredett-e" férfi alkotó az írónők közé. Ez a kérdés vezeti el a „férfi" és a „női" írásmód megkülönböztethetőségének igazán izgalmas és nem kevésbé problematikus kérdéséhez – eredményei alapján úgy tűnik, hogy míg a 18. században kulcsszavaik alapján elkülöníthetők egymástól a férfiak és a nők által írt regények, addig a 19. és 20. században ez a szétválasztás már nem lehetséges. A tanulmány, miközben a társadalmi nem kérdésköréhez járul hozzá kvantitatív szempontokkal, érdekes részleteket közöl az angol irodalomtörténet kapcsán is.

A blokk harmadik szövege egyszerre kapcsolódik a Computational Stylistics Group alapkutatásaihoz és mutat fel önálló eredményeket. Az ebben részletezett, nemzetközi kollaboráció során megvalósult projekt leginkább azt a kérdést teszi fel, hogy mennyiben befolyásolja a szerzőazonosítás eredményeit, ha különböző módon és különböző minőségben digitalizált szövegekből áll az elemezni kívánt korpusz. Ennek kapcsán részletesen bemutatják, hogyan működik a nyomtatott és a kézzel írt szövegek digitalizálása és a betűkarakterek automatikus felismerése – előbbi az OCR (optikai karakterfelismerés), utóbbi a HTR (kézzel írott szövegfelismerés) technikáján alapul. Mindkét eljárás automatikusan ismeri fel a digitalizált karaktereket (a HTR esetében ehhez először létre kell hozni az adott szerző kézírásának modelljét), ám gyakran sok hibával dolgoznak. A digitális bölcsészet kutatói sokszor kényszerülnek így létrejött, különböző minőségű és forrású, „zajos" szövegekre támaszkodni – a tanulmány arra keresi a választ, hogy ezek mennyiben nehezítik meg a szerzőazonosítás feladatát, azaz mennyire „mosódik el" az egyes szerzők „ujjlenyomata" a digitalizálás folyamatában.

Artjoms Šeļa, Boris Orekhov és Roman Leibov tanulmánya mind módszertanilag, mind a kutatási kérdést tekintve talán a legambiciózusabb vállalkozás a lapszámban. Az orosz költészet kiterjedt korpuszán vizsgálják, hogy a versek metrikai képlete vajon azok tematikus szerveződését is meghatározza-e. A „témák" kijelölésekor jól érzékelhető a digitális bölcsészet hagyományos stíluskutatástól eltérő logikája: a témamodell *(topic modelling)* algoritmusa mindössze az egymás kontextusában előforduló szavak csoportjait határozza meg; a szerzők pedig nem értelmezik – sőt a főszövegben nem is idézik – ezeket a csoportokat, csupán az egyes versmértékekre jellemző eloszlásukról adnak számot. (Arról, hogy a módszer jól értelmezhető eredményeket is biztosít, a mellékletben közölt eredmények értesítenek: például az ötös trocheushoz korábban

rendelt „éjszaka", „táj", „halál", „szerelem", „út" címkékhez hasonló szócsoportokat azonosít az algoritmus is – többek között „tudni, élni, lenni, meghalni, semmi"; „kert, zöld, levél, ág, hárs"; „menni, ösvény, út, keresztezni, láb".) A modellek létrehozásakor roppant körültekintő és invenciózus módon járnak el a szerzők; külön érdemes megemlíteni ötletüket, amely szerint, ha a költemények kevésbé gyakori szavait szóbeágyazási módszerrel *(word-embedding)* a gyakrabban előforduló szinonimákra cseréljük, még hatékonyabb eredményekre vezethet a témamodellezés folyamata. Tesztjeik kimutatják, hogy erős kapcsolat áll fenn az orosz költészeti hagyományban a versmérték és a tartalom között – azaz már a vers formájából következtetni lehet annak szemantikai karakterére. Mindezt a kulturális evolúció *(cultural evolution)* fiatal tudományterület kereteiben értelmezik, amely azt a folyamatot vizsgálja, ahogyan a társas tanulás során megszerzett információk változnak, fennmaradnak és differenciálódnak az idő során. A szerzők véleménye szerint az általuk azonosított összefüggések és tendenciák általánosan, nyelvektől és irodalmi hagyományoktól függetlenül igazak lehetnek, ami által a versmérték–jelentés kapcsolatban az irodalmi termelés egy alapvető dinamikáját érthetjük meg.

A lapszám utolsó két írása a nyelvészet területéről közelít a digitális bölcsészet és a stilometria felé. Albert Leśniak és Zbigniew Pasek munkájának kiindulópontja, hogy a korpusznyelvészet tulajdonképpen a foucault-i értelemben vett diskurzuselemzés. Ebből az alapállásból kiindulva vetik össze két vallási közösség, a neoprotestáns és a katolikus hívők tanúságtételeinek szövegeit. Ezeknek kulcsszavait és kulcsszavak kollokációit elemezve alapvető különbségekre tudnak rámutatni az egyes közösségek működésében, leginkább a hívőknek a bűnhöz való viszonyát és a megtérés folyamatának időszerkezetét tekintve. Míg a katolikus tanúságtételek elsősorban a szexualitás témája körül forognak, addig a neoprotestáns változatokban inkább a különböző függőségek (alkohol, drog, cigaretta) kerülnek előtérbe a bűnök felsorolásakor; illetve míg a katolikus közösségben éles váltás figyelhető meg a megtérés előtti és utáni időszak között, addig a neoprotestáns elbeszélések a lassabb átmenet sémáját részesítik előnyben.

Az utolsó tanulmány az eddigiekkel szemben kevésbé a saját kutatási eredmények ismertetésében érdekelt. Egy olyan folyamatról számol be, amely minden esetben az eddig elmondottak feltételeként szolgál: magának az elemezni kívánt digitális korpusznak a létrehozásáról. Mivel ezt a kérdést a dialektológia területén járják körül, a lengyel Szepesség digitális és kereshető adatbázisának megalkotásáról szóló írásuk külön érdekes lehet a nyelvjárások és a szociolektusok kutatói számára.

A krakkói Computational Stylistics Group egy nemzetközileg jelentősen beágyazott műhely. Az említett előadás-sorozatokon és workshopokon túl szoros kapcsolatot ápolnak az Antwerpeni Egyetemmel (amellyel külön projektben vizsgálják a *deep learning* módszerek stilisztikai hasznosíthatóságát[24]), az ELTE BTK Digitális Bölcsészet Tanszékéhez hasonlóan tagjai a „Distant Reading for European Literary History" COST Action programnak[25] és a „CLS Infra" elnevezésű Horizon 2020 projektnek is.[26] Jelen lapszámmal reméljük, hogy munkásságuk a magyar digitális bölcsészeti közösséghez is közelebb kerül.

[24] Hozzáférés: 2021.12.13, `https://computationalstylistics.github.io/projects/deep_learning/`.

[25] Hozzáférés: 2021.12.13, `https://www.distant-reading.net/`.

[26] Hozzáférés: 2021.12.13, `https://clsinfra.io/`.

<TANULMÁNYOK>

Maciej Eder 0000-0002-1429-5036

Instytutu Języka Polskiego PAN

Uniwersytetu Pedagogicznego w Krakowie, Instytucie Filologii Polskiej

maciej.eder@ijp.pan.pl

Elena Ferrante: Egy „virtuális" szerző*

A jelen tanulmány az Elena Ferrante írói álnév alatt publikált regényeket vizsgálja azért, hogy kísérletet tegyen az álnév mögött rejtőző szerzőség azonosítására. A tanulmány ugyanakkor többre vállalkozik annál, minthogy egyszerűen újranyissa a szerzőazonosítás kérdését; a kutatás valódi jelentősége az, hogy magának a szerzői ujjlenyomatnak az állandóságát teszteli Ferrante műveinek vonatkozásában, rejtőzzön akárki is az írói álnév mögött. A kutatási kérdés megválaszolásához egy 150 regényből álló korpusz állt rendelkezésre, a stilisztikai hasonlóság méréséhez az úgynevezett Bootstrap Consensus Network eljárást használta a kutatás. Azon szerzők listája, akiknek a szerzői ujjlenyomata a leginkább egyezőnek bizonyult „Ferrante" lenyomataival – köztük Domenico Starnonéval a lista élén – az úgynevezett Rolling Classify ('mozgó osztályozás') metódus alkalmazásával vált hozzáférhetővé. Ez a módszer – amit Ferrante és Starnone regényein egymástól függetlenül alkalmaztunk – az irodalmi szövegek lokális stilisztikai sajátosságainak feltárására alkalmas, ami ezáltal részletesebb megfigyeléseket is megfogalmazhatóvá tesz. Az összkép, néhány kivételtől eltekintve, megerősíti, hogy Starnone és Ferrante stílusa megkülönböztethető egymástól, ami viszont erős érvnek mutatkozik a „virtuális szerzőség" hipotézise mellett. Úgy tűnik, hogy Domenico Starnone – különösen kései műveiben – képes megkülönböztetni egymástól saját és alteregója hangját.

Kulcsszavak:
stilometria, szerzőazonosítás, virtuális szerzőség, Bootstrap Consensus Network, Rolling Classify

Bevezetés

Elena Ferrante szerzői nevét csaknem három évtizede tartják számon a nemzetközi irodalmi színtéren. A hét regény és a *The Guardian* heti rovatában rendszeresen publikált művek szerzőjének valódi kiléte azonban sajátságos módon anonimitásba burkolózik. Az olvasóközönség ugyanis tisztában van azzal, hogy az „Elena Ferrante" egy, a szerző(k) által választott álnév, amely elfedi a valós szerző(k) kilétét. Azaz

* Eredeti megjelenés: Maciej Eder, „Elena Ferrante: A Virtual Author," in Arjuna Tuzzi and Michele A. Cortelazzo, eds., *Drawing Elena Ferrante's Profile,* 31–47 (Padova: Padova University Press, 2018).

talán csak a közelmúltig volt ez így: egyre több cikkben olvasható ugyanis a feltevés, miszerint az írói álnév mögött Dominico Starnone szerzősége rejtőzik.[1] Az azóta megjelent tudományos munkákban, amelyek a legkorszerűbb statisztikai elemzéseket alkalmazzák vizsgálatuk során, szintén beigazolódott ez a feltevés.[2] Starnone 1943-ban született író és újságíró, aki számos, kritikusok által elismert regény és kisebb prózai művek alkotója. Érdekes tény, hogy a másik szerző, akinek a közvélekedés szintén az írói álnév alatt publikált regényeket tulajdonítja, nem más, mint Anita Raja – Domenico Starnone felesége, akit a német irodalom egyik fordítójaként ismerhetünk.[3]

A jelen tanulmány nem a fentiekben részletezett szerzői kérdést kívánja újra megnyitni – bár a Starnone-hipotézis most is a vizsgálat egyik előterében álló jelenség –, ehelyett a vizsgálat a szerzői „ujjlenyomat" stabilitásának tesztelésére tesz kísérletet, legyen akárki is Ferrante műveinek tényleges szerzője. Azontúl, hogy a vizsgálat törekszik azonosítani az anonimitásba burkolózó szerző nevét, mérhetővé teszi azt is, hogy az azonosított szerző neve alatt megjelent művek, valamint a Ferrante név alatt publikált prózai munkák milyen mértékű egyezést és eltérést mutatnak egymáshoz képest a stilisztika szempontjából. Tehát a kutatási kérdés a következő: vajon két elhatárolható, de egymással mégis összefüggő „szerzői" profillal állunk-e szemben, amelyek egyfelől az íróhoz, másfelől az író által konstruált szerzői identitáshoz tartoznak? A szerzőazonosításra alkalmazott módszerek az előbb vázolt kérdések kontextusában az alábbi hipotézist teszik lehetővé: ha a gépi tanulással működő klasszifikációs eljárás képes elválasztani egymástól a két különböző, de egymáshoz erősen kapcsolódó szerzői ujjlenyomatot – tehát a „Ferrante"-hez és a tényleges olasz íróhoz tartozót – az visszaigazolja, hogy egyetlen személy képes lehet egyszerre több stilisztikai perszónát kialakítani.

A potenciális írok listájának szűkítése

A szerzőazonosításhoz elvégzendő vizsgálatok első lépése az, hogy olyan szerzők műveit gyűjtsük össze, akik az ismeretlen szöveg potenciális írói lehetnek. „Ferrante" esetében több, mint kockázatos azt feltételeznünk, hogy a lehetséges szerzők pontosan listázhatók, még akkor is, ha az előzetes vizsgálatok alapján úgy tűnik, Domenico

[1] Luigi Galella, „Ferrante-Starnone: Un amore molesto in via Gemito," *La Stampa*, 2005. jan. 16., 27; Simone Gatto, „Una biografia, due autofiction. Ferrante-Starnone: cancellare le tracce," *Lo Specchio di carta: Osservatorio sul romanzo italiano contemporaneo,* 2016. okt. 22, `https://lospecchiodicarta.it/2016/10/22/una-biografia-due-autofiction-ferrante-starnone-cancellare-le-tracce/`.

[2] Michele A. Cortelazzo e Arjuna Tuzzi, „Sulle tracce di Elena Ferrante: Questioni di metodo e primi risultati," in Giuseppe Palumbo, ed., *Testi, corpora, confronti interlinguistici: Approcci qualitativi e quantitative,* 11–25 (Trieste: Edizioni Università di Triest, 2017), http://doi.org/10.13137/978-88-8303-913-3/18478; Michele A. Cortelazzo, George K. Mikros and Arjuna Tuzzi, „Profiling Elena Ferrante: A Look Beyond Novels," in Domenica Fioredistella Iezzi, Livia Celardo and Michelangelo Misuraca, eds., *Jadt '18. Proceedings of the 14th International Conference on Statistical Analysis of Textual Data,* 165–173 (Roma: UniversItalia, 2018); Anjura Tuzzi and Michele A. Cortelazzo, „What is Elena Ferrante? A Comparative Analysis of a Secretive Bestselling Italian Writer," *Digital Scholarship in the Humanities* 33, 3. sz. (2018): 685–702, `https://doi.org/10.1093/llc/fqx066`.

[3] Claudio Gatti, „Elena Ferrante: An answer?" *The New York Review of Books,* 2016. okt. 2., `http://www.nybooks.com/daily/2016/10/02/elena-ferrante-an-answer/`.

 DOI 10.31400/dh-hun.2021.5.3142

Starnone lapul az írói álnév mögött. Ilyen esetekben érdemes a jelöltek nyílt halmazát, tágabb körét vizsgálnunk, amely során megbízhatóan magas számú művet viszünk be a korpuszba – jelen esetben kortárs olasz szerzőktől –, azért, hogy a stilometriai vizsgálatok minél szélesebb spektrumon legyenek elvégezhetők, és ezáltal kiszűrhetővé váljanak azok az alkotók, akik szerzői ujjlenyomata potenciálisan megfeleltethető a kérdéses szerzőségű szöveg írójának lenyomatával. A jelen tanulmányban bemutatott vizsgálat tesztkorpusza összesen 150 olasz 20. századi regényt tartalmaz, 40 különböző szerzőtől, köztük hét olyat, amely a Ferrante álnév alatt került publikálásra. A tesztkorpusz megegyezik azzal a gyűjteménnyel, amelyet Michele Cortelazzo és Arjuna Tuzzi használt egy korábbi, Ferrante írói kilétének kérdéséhez kapcsolódó vizsgálatban.[4]

Mivel a cél a potenciális jelöltek listájának szűkítése, nem pedig egyetlen szerző hozzárendelése a kérdéses szerzőségű szövegekhez – ez olyan eset, ahol a nagy merítés, azaz a felidézés *(recall)* fontosabb, mint a precízió *(precision)* –, ezért egy meglehetősen egyszerű eljárást érdemes alkalmaznunk. Annak ellenére, hogy a kifinomultabb gépi tanuláson alapuló módszerek (pl. Support Vectors Machines) rendkívül hatékonynak bizonyulnak a hasonló többdimenziós problémák feltárásában,[5] a szerzőazonosításban a viszonylag egyszerű távolságalapú eljárások is kellően jól teljesítenek,[6] különösen akkor, amikor magas a vizsgált szerzők száma.[7] A Ferrante-regények lehetséges szerzőinek kiválasztásához a Bootstrap Consensus Network (BCN) módszert[8] használta a kutatás. Ez a metódus a klaszteranalízisnek olyan továbbfejlesztett változata, melynek célja egy adott korpusz többszörös kiértékelése különböző jellemzők mentén (jelen esetben a leggyakoribb szavak értékei 100–1000 szavas tartományban). Az iterációkban a stilometriában használatos távolságalapú metrikák mérik a szövegek közelségét; a cél ekkor az egymáshoz stilárisan kapcsolódó szövegek azonosítása.

A stiláris kapcsolat e koncepciója lényegében abból indul ki, hogy a korpusz minden szövege jellemezhető aszerint, hogy többé-kevésbé hasonlít (stilometriailag) az összes többi szöveghez, ezáltal mindegyiknek létezik egy szomszédja, amely a legmagasabb egyezést mutatja vele. E feltételezés geometriai interpretációja azt jelenti, hogy a szövegek egy többdimenziós térben (a dimenziók számát a vizsgált jellemzők száma adja meg) pontokként reprezentálódnak; a pontok közötti térbeli távolság pedig azt

[4] Cortelazzo e Tuzzi, „Sulle tracce di Elena Ferrante"; Cortelazzo, Mikros and Tuzzi, „Profiling Elena Ferrante"; Tuzzi and Cortelazzo, „What is Elena Ferrante?"; Michele A. Cortelazzo, Paolo Nadalutti, Stefano Ondelli and Arjuna Tuzzi, „Authorship Attribution and Text Clustering in Contemporary Italian Novels: Does Elena Ferrante's and Domenico Starnone's Regional Origin Play a Role?" in Lu Wang, Reinhard Köhler and Arjuna Tuzzi, eds., *Scructures, Properties, and Interrelations: Selected Papers from Qualico 2016,* 1–14 (Lüdenscheidt: RAM Verlag, 2018).

[5] Gareth James, Daniela Witten, Trevor Hastie and Robert Tibshirani, *An Introduction to Statistical Learning with Applications in R* (New York: Springer, 2013), https://doi.org/10.1080/24754269.2021.1980261.

[6] Matthew Jockers and Daniela Witten, „A Comparative Study of Machine Learning Methods for Authorship Attribution," *Literary and Linguistic Computing* 25, 2. sz. (2010): 215–223, https://doi.org/10.1093/llc/fqq001.

[7] Koen Luyckx and Walter Daelemans, „The Effect of Author Set Size and Data Size in Authorship Attribution," *Literary and Linguistic Computing* 26, 1. sz. (2011): 35–55, https://doi.org/10.1093/llc/fqq013.

[8] Maciej Eder, „Visualization in Stylometry: Cluster Analysis Using Networks," *Digital Scholarship in the Humanities* 31, 1. sz. (2017): 50–64, https://doi.org/10.1093/llc/fqv061.

szimbolizálja, hogy a szövegek egymáshoz képest milyen stilometriai hasonlóságot mutatnak. Ezért a szövegek közelsége meghatározható geometrikusan, távolságalapú metrikák alapján. A jelen tanulmányban a Manhattan-távolságot alkalmaztuk a standardizált szógyakoriságokra *(z-score),* amelyeket a dimenziók számával (vagyis az összes vizsgált szó számával) osztottunk el. Ez a mérési procedúra Burrows Deltaként vált ismertté a stilometriában.[9]

A BCN-eljárás minden egyes iterációja egy listát eredményez, amelyek a megadott számú jellemzők alapján a szövegekre vonatkozó szomszédsági viszonyokat tartalmazzák. Az első ilyen „pillanatkép" a 100 leggyakoribb, míg a második iteráció a 200 leggyakoribb szóra irányult, majd a 100-as lépésköz további alkalmazásával eljutottunk a 300-as, 400-as, és egészen az 1000 leggyakoribb szót figyelembe vevő beállításig. A következő lépésként ezeknek a pillanatképeknek az eredményeit szinkronizáljuk egymással, és egy hálózatalapú leképezést hozunk létre belőlük. A hálózat élei a szövegek (csomópontok) közötti legközelebbi szomszédos kapcsolatokat jelölik. Pontosabban, minden egyes csomóponthoz három él rendelődik – a hasonlóság alapján a szöveg három legközelebbi szomszédja. Ebből kifolyólag a BCN-eljárás több szempontból megegyezik a k-NN klasszifikációval,[10] amely egy szöveg *k* legközelebbi szomszédja alapján hozza meg klasszifikációs döntését.

A fenti módon generált hálózat meglehetősen informatív számunkra, ugyanakkor az így kirajzolódó ábra szabad szemmel nehezen olvasható, így érdemes a hálózatot erőirányított gráfként *(force-directed graph)* is vizualizálni. Ebben az esetben az eljárás részét képezi a csoportokba rendezett csomópontok (azaz a szövegek), illetve a még nagyobb csoportokba rendezett csoportok kézi ellenőrzése is. A kialakuló klaszterek emberi értelmezése, több más magyarázó módszerhez hasonlóan, meglehetősen egyszerűvé teszi ennek használatát.

A tanulmányban közzétett összes számítási elemzés elvégzéséhez – beleértve a stilometriai hálózatok megrajzolását is – az R program *Stylo* bővítményét alkalmaztuk.[11] A végleges hálózatok rendezéséhez pedig a *Gephi* szoftvert és annak a beépített ForceAtlas2 algoritmusát használtuk.[12]

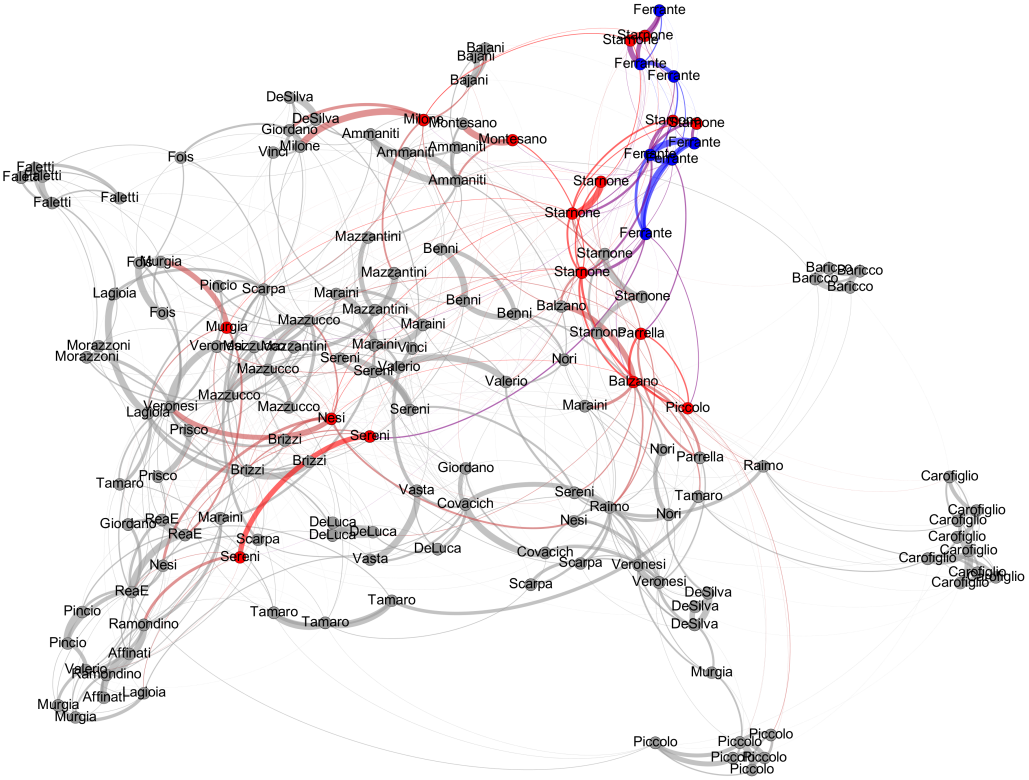
[9] John Burrows, „'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship," *Literary and Linguistic Computing* 17, 3. sz. (2002): 267–287, https://doi.org/10.1093/llc/17.3.267.

[10] Gareth, Witten, Hastie and Tibshirani, *An Introduction to Statistical Learning,* 104.

[11] Maciej Eder, Jan Rybicki and Mike Kestemont, „Stylometry with R: A Package for Computational Text Analysis," *R Journal* 8, 1. sz. (2016): 107–121, https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

[12] Mathieu Bastian, Sébastien 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.

 DOI 10.31400/dh-hun.2021.5.3142



1. ábra. Bootstrap Consensus Network: 150 kortárs olasz regény. A legközelebbi szomszédsági hasonlóságok hálózati kapcsolatokként ábrázolva. A Ferrantéhez leginkább hasonló szövegek kékkel kiemelve.

A vizsgálatba bevont 150 kortárs olasz regény konszenzushálózata az *1. ábrán* látható. Attól eltekintve, hogy az így kirajzolódó regényhálózat egészének elemzése további izgalmas eredményeket szolgáltathat, a jelen vizsgálat középpontjában azon hálózatrészek állnak, amelyek Ferrante regényei környezetében fordulnak elő. A hálózatot szemlélve feltűnő lehet számunkra Ferrante és Starnone közelsége. Ebből következtethetünk arra, hogy az ábrán látható írók közül Starnone az, akinek a regényei stilometriailag leginkább egyezést mutatnak Ferrante műveivel. Megállapítható még, hogy lényegében mind a hét Ferrante-regény szoros kapcsolatban áll Starnone egyik művével. Azonban vannak más olyan szerzők is a hálózatban, akik időnként kapcsolódnak Ferrante regényeihez, tehát nem zárhatók ki a további stilometriai mérésekből. E regényírók a következők: Rosella Milone, Giuseppe Montesano, Valeria Parrella, Francesco Piccolo, Clara Sereni, Michela Murgia és Edoardo Nesi. A felsorolás vegyesen tartalmaz női és férfi szerzőket, ennek ellenére mindegyiket potenciális jelöltnek tekinthetjük az Elena Ferrante írói álnév mögött megbúvó szerzőségre.

Ferrante, darabokban

Az előző részben ismertetett vizsgálat eredményeit áttekintve megerősíthetjük azt a hipotézist, miszerint Elena Ferrante valóban Domenico Starnone írói alteregója lehet, továbbá azonosíthatóvá váltak olyan szerzők, akik művei ugyan kevésbé mutatnak

egyezést stilometriai szemszögből Ferrante regényeivel, azonban a térbeli közelség miatt mégsem zárhatók ki a potenciális szerzőjelöltek listájáról. Az így listázott művek közötti lehetséges kapcsolatok feltárásához ezután az előzőektől alapjaiban eltérő, újszerűnek számító kísérleti módszert alkalmazunk. A Rolling Stylometry eljárásban a fókuszpontban lévő (anonim) szövegeket kisebb részekre szegmentáljuk, majd ezeket szekvenciális elemzéssel dolgozzuk fel.[13] Ez a módszer azon feltételezésre épül, hogy a stilometriai lenyomatoknak nem szükségszerűen kell egyenletes mértékűnek lenniük egy adott szöveg különböző részeiben. Gondoljunk csak arra, hogy egy regényben vélhetően a narratív szövegrészek a dialógusoktól eltérő stiláris profillal bírnak. Hasonlóképpen gondolkodhatunk azokról a szövegegységekről, amelyek kiterjedtebb intertextuális kapcsolatokat hordoznak; az alkalmazott módszeren keresztül ezen szekvenciákat is detektálni tudjuk, hiszen mindegyik regény szegmenseit külön, ugyanakkor egymás után rendezve teszteljük.

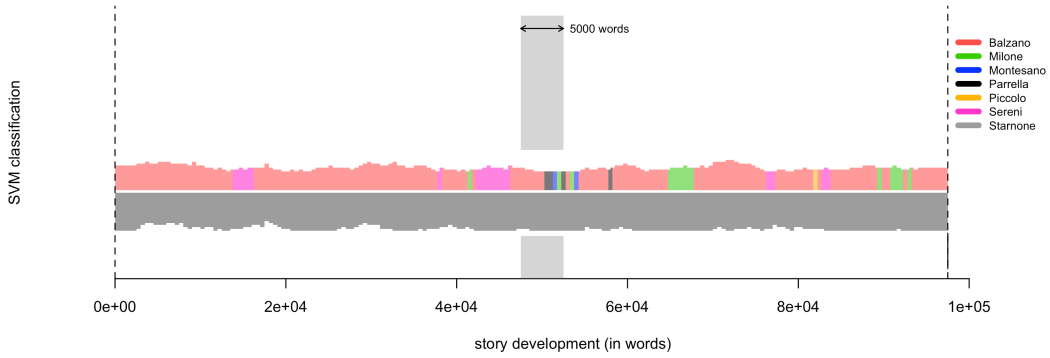
Két oka van, hogy a Rolling Classify eljárást alkalmazzuk a regények vizsgálatára. Az első, hogy könnyen lehet, hogy a Ferrante álnév mögött nem egy, hanem több szerző áll. A második pedig, hogy az előzőekben bemutatott vizsgálat mintegy madártávlatból enged betekintést a szövegek hálózatrendszerébe, míg az imént vázolt eljárás egyfajta mikroszkópként is működhet, hiszen itt a kulcs az, hogy a szövegek minden szekvenciáját egyenként vetjük analízis alá.

Az alkalmazott eljárási mód ellenőrzött gépi tanuláson alapuló klasszifikációs módszer, amely használatához szükségünk van egy tanítókorpuszra *(training set)*, azaz egy manuálisan válogatott, ismert szerzőségű szövegekből álló gyűjteményre, amely reprezentatívnak bizonyul az általa tartalmazott szerzői csoportokra vonatkozóan; és természetesen szükségünk van magára a tesztkorpuszra *(test set)*, mely tartalmazza a vizsgálni kívánt szövegtesteket. A különbség abban rejlik a szokásos ellenőrzött beállítások és a Rolling Classify között, hogy az utóbbi a bemeneti (ismeretlen szerzőségű) szövegeket egyenlő méretű szekvenciákra darabolja fel, majd a tesztkorpuszt ezekkel a diszkrét szövegdarabokkal tölti fel, megtartva azok eredeti sorrendjét. Az eljárás kiegészül egy komplex vizualizációval is, amely az eredeti szöveget, vagyis a darabjainak sorrendjét színes csíkok formájában mutatja. A csík *alsó* része a klasszifikációs metódus döntését ábrázolja: minél szélesebb a csík, annál megbízhatóbb az osztályzás.

A jelen vizsgálatban az összes szekvenciális kísérletet ugyanazon hiperparaméterbeállításokkal hajtottuk végre. A tanítókorpusz 30 regényt tartalmazott az előző mérések által azonosított hét különböző szerzőtől: Marco Balzano, Rossella Milone, Giuseppe Montesano, Valeria Parrella, Francesco Piccolo, Clara Sereni és Domenico Starnone. A tesztek során Ferrante minden regényét 5000 szavas blokkokra daraboltuk fel, úgy, hogy a blokkok közötti lépésköz 500 szó legyen; tehát a blokkokat meghatározó vizsgálati „ablak" mindig 500 szóval haladt tovább az eredeti szövegen.

[13] Maciej Eder, „Rolling Stylometry," *Digital Scholarship in the Humanities* 31, 3. sz. (2016): 457–469, https://doi.org/10.1093/llc/fqv010.

 DOI 10.31400/dh-hun.2021.5.3142



2. ábra. Ferrante *L'amica geniale* című regénye, részletenként a szűkített korpusz 7 szerzőjével szembeállítva. Jellemzők: Rolling Classify módszer, SVM-osztályozás, a 100 leggyakoribb szóra.

A *L'amica geniale* (2011) című regényre vonatkozó eredményeket a *2. ábra* szemlélteti. A vonal alatt látható egységes csík nem mutat stiláris különbséget, tehát az azonosító eljárás a regény minden szegmensét Domenico Starnone szerzőségéhez rendeli. A vizsgálatban minden kapott vizualizáció azonos képet vázol fel: a klasszifikációs metódus kivétel nélkül Starnonénak tulajdonítja a vizsgált szövegtesteket, amely erős bizonyítékul szolgál az előzetes, szerzőségre irányuló hipotézis megerősítéséhez.

Egy virtuális szerző?

Noha a szerzőazonosításra irányuló vizsgálatok sohasem zárhatók le biztos eredményekkel, a vizsgálat statisztikai alapokon működő eljárásai stabil empirikus bizonyítékokat nyújtottak annak megerősítésére, hogy az anonim szerzőség mögött Domenico Starnone szerzői ujjlenyomatát sejthessük. Bármennyire is érdekes ez, azzal tisztában kell lennünk, hogy önmagában a szerzőazonosítási eljárás nem biztosít rálátást a vizsgált művek keletkezésének körülményeire. Ugyan izgalmas feltárni azt, ki áll az ismeretlen szerzőségű szövegek mögött, még izgalmasabb kérdés, hogy egy adott szerző mennyire képes szándékosan megváltoztatni a saját írói stílusát. Tételezzük fel, hogy Ferrante azonos Starnonéval. Vajon milyen kapcsolat áll fenn a „két" szerzői életmű között, és milyen stiláris, stilometriai különbséget mutathatunk ki közöttük?

Természetesen a kettős identitás problematikája nem új keletű. Az egyik legismertebb példa a világirodalomban Romain Gary (1914–1980) francia regényíró életműve, aki szerzői pályafutásának egy bizonyos időszakában úgy döntött, hogy személyét „elmaszkolva", Émile Ajar álnéven publikálja műveit. Ő az egyetlen olyan szerző, akinek kétszer ítélték oda a Goncourt-díjat (a díj csak egyszer nyerhető el). Ez úgy történhetett meg, hogy egyszerűen nem derült ki, hogy valójában egyetlen szerző áll a két életmű mögött. A ritmikai mintázatokon alapuló stilometriai vizsgálatok ráadásul azt sugallták, hogy a két szerzői profil valóban mutat különbséget, még akkor is, ha az nem minden esetben egyértelmű.[14]

[14] Adam Pawłowski, *Séries temporelles en linguistique: Avec application à l'attribution de textes Romain Gary et Emile Ajar* (Lausanne: Slatkine, 1996).

A saját szerzői profil megváltozásának azonban nem feltétlen kell szándékosnak lennie. Érdemes megemlíteni azokat az eseteket is, amikor ezek a stiláris elmozdulások organikusan jönnek létre az idő múlásával.[15] Ezt a jelenséget figyelhetjük meg például Henry James életművében is.[16] Érdekesség továbbá, hogy a stilometriai módszer egyik első alkalmazása hasonló módon Platón szerzői ujjlenyomata időbeli alakulásának feltárására irányult.[17] A szerzői lenyomat megváltozása mögött olykor betegségek is állhatnak, mint például a demencia Agatha Christie esetében,[18] de a változás akkor is kimutatható lehet, amikor egy szerző segítséget kér a művei lejegyzéséhez, majd a lejegyző stílusa dominánsabbá válik az eredeti szerzőéhez képest – példaként említhető Francis Bacon pályájának egy szakasza.[19]

A jelen tanulmány esetében releváns hivatkozási pont a Mike Kestemont és szerzőtársai által végezett kutatás. Ennek során két középkori latin nyelven lejegyzett látomást vetettek analízis alá, amelyeket a szakirodalom hagyományosan Bingeni Hildegardnak tulajdonít.[20] A kutatók megfigyelései alapján azonban kijelenthető, hogy Hildegard és lejegyzője, Gibloux-i Guibert együttműködése elkülönülő szerzői profilt alkot, amely nem feleltethető meg egyszerűen Hildegard és Guibert stílusának kombinációjaként. Ebben a példában a kooperációt nevezhetjük inkább egy virtuális „harmadik" szerző profiljának is. A virtuális szerzőség létezésének felvetése a szinergiahipotézis elméleti keretéből származtatható,[21] miszerint egy közösen írt mű szerzői lenyomata mutathat hasonlóságot a domináns szerzői kéz ujjlenyomatával, azonban akár el is térhet mindegyik alkotó önálló szerzői profiljától.

A következő vizsgálat arra irányult, hogy feltárja, vajon Starnone és Ferrante esetében beszélhetünk-e ilyen virtuális szerzőségről, valamint hogy Starnone saját neve alatt megjelent írásai mennyire különböznek el a Ferrante név alatt publikált szövegektől. Ez utóbbi kérdés valójában igen összetett, amely mentén további kutatói kérdések fogalmazhatók meg. Például az, hogy a Ferrante név alatt publikált szövegek több női szerzőre jellemző tulajdonságot mutatnak-e fel, mint azok, amelyeket az író a saját neve alatt jelentetett meg. Vajon a tényleges szerzőnek sikerült kilépnie a saját nyelvi

[15] Constantina Stamou, „Stylochronometry: Stylistic Development, Sequence of Composition, and Relative Dating," *Literary and Linguistic Computing* 23, 2. sz. (2008): 181–199, https://doi.org/10.1093/llc/fqm029.

[16] David L. Hoover, „Corpus Stylistics, Stylometry, and the Styles of Henry James," *Style* 41, 2. sz. (2008): 174–203.

[17] Wincenty Lutosławski, *The Origin and Growth of Plato's Logic: With an Account of Plato's Style and of the Chronology of his Writings* (London: Longmans, Green & Co, 1897).

[18] Xuan Le, Ian Lancashire, Graeme Hirst and Regina Jokel, „Longitudinal Detection of Dementia through Lexical and Syntactic Changes in Writing: A Case Study of three British Novelists," *Literary and Linguistic Computing* 26, 4. sz. (2011): 435–461, https://doi.org/10.1093/llc/fqr013.

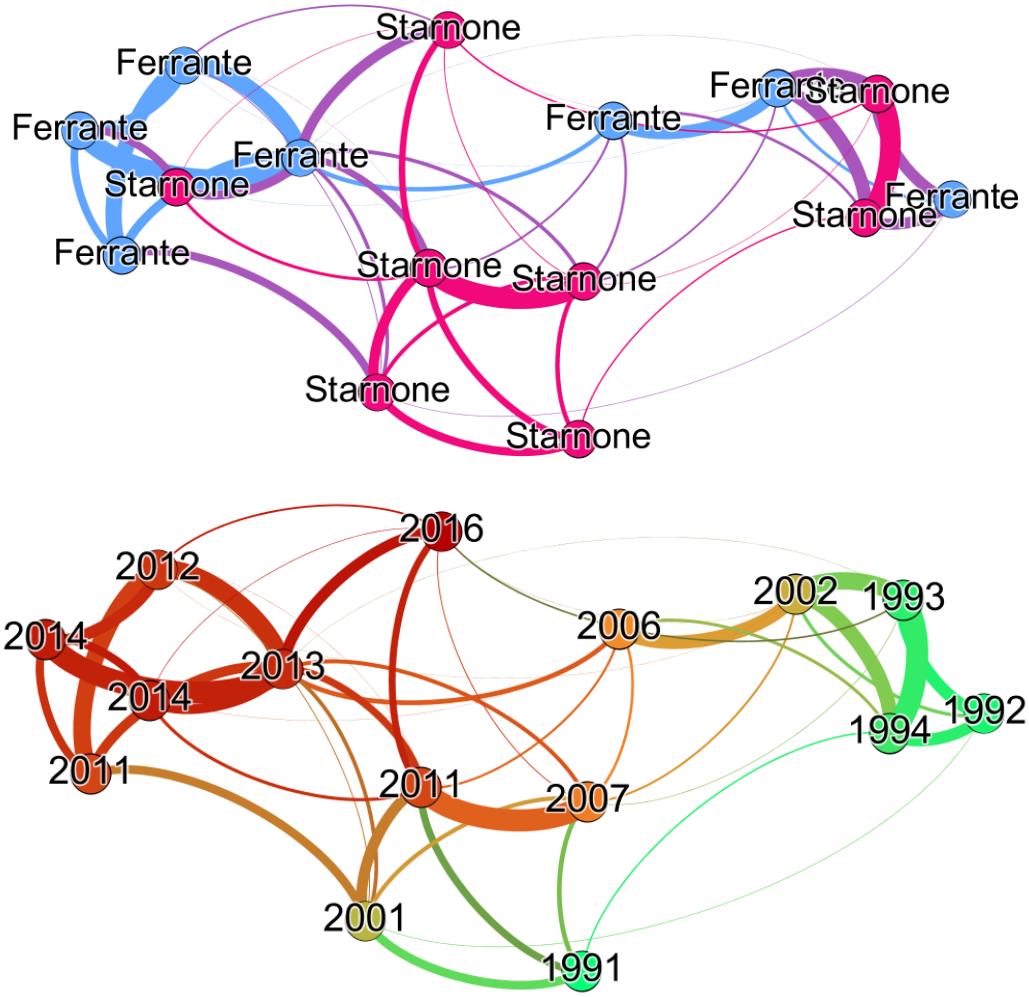
[19] Noel B. Reynolds, Bruce G. Schaalje and John L. Hilton, „Who Wrote Bacon? Assessing the Respective Roles of Francis Bacon and his Secretaries in the Production of his English Works," *Literary and Linguistic Computing* 27, 4. sz. (2011): 409–425, https://doi.org/10.1093/llc/fqs020.

[20] Mike Kestemont, Sara Moens and Jeroen Deploige, „Collaborative Authorship in the Twelfth Century: A Stylometric Study of Hildegard of Bingen and Guibert of Gembloux," *Literary and Linguistic Computing* 30, 2. sz. (2015): 199–224, https://doi.org/10.1093/llc/fqt063.

[21] James W. Pennebaker, *The Secret Life of Pronouns: What Our Words Say About Us* (New York: Bloomsbury Press, 2011).

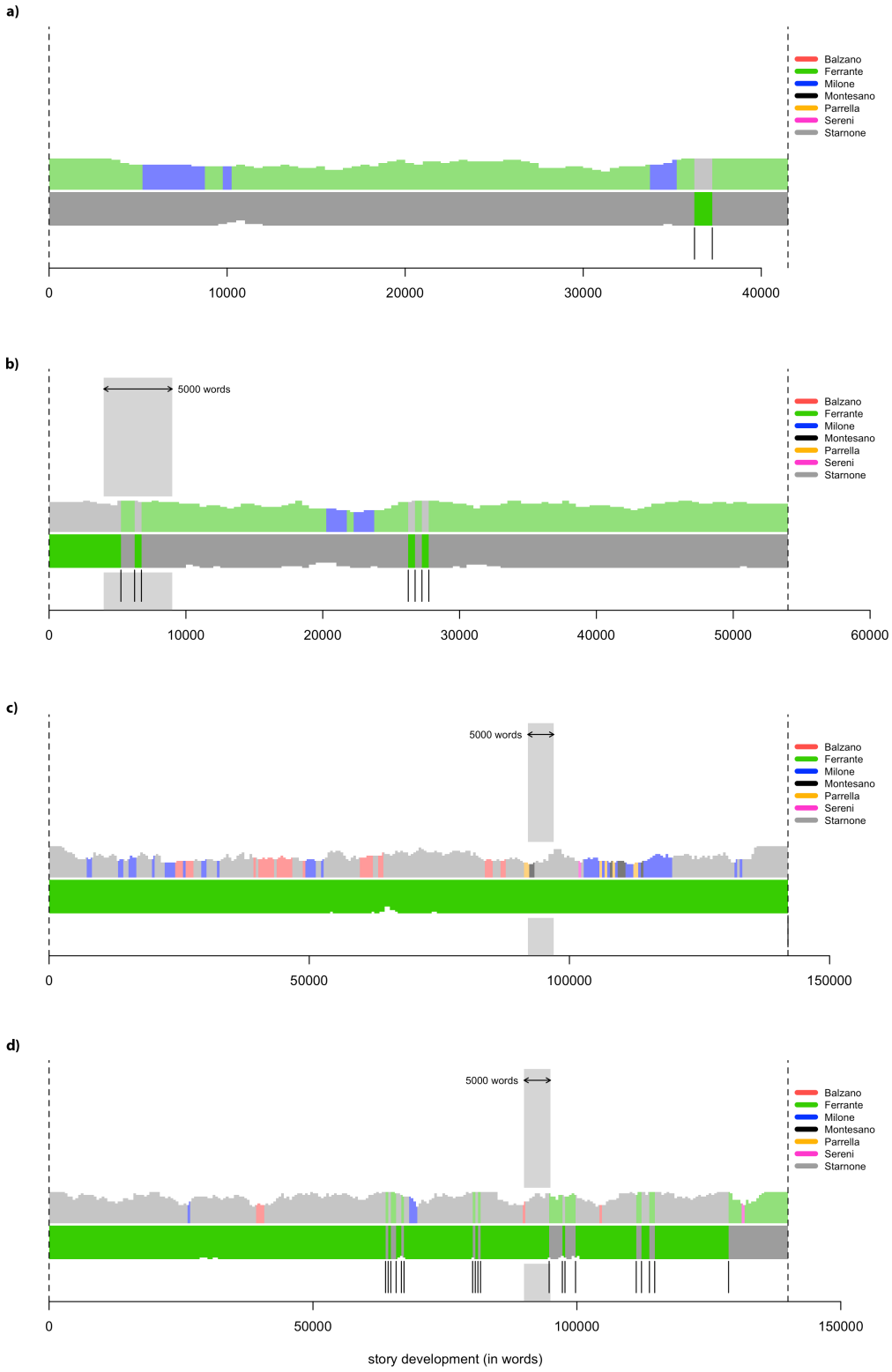
 DOI 10.31400/dh-hun.2021.5.3142

beágyazottságából? Azonban ezen kérdéseknek csupán töredéke az, amelyre a jelen tanulmányban választ tudunk adni.



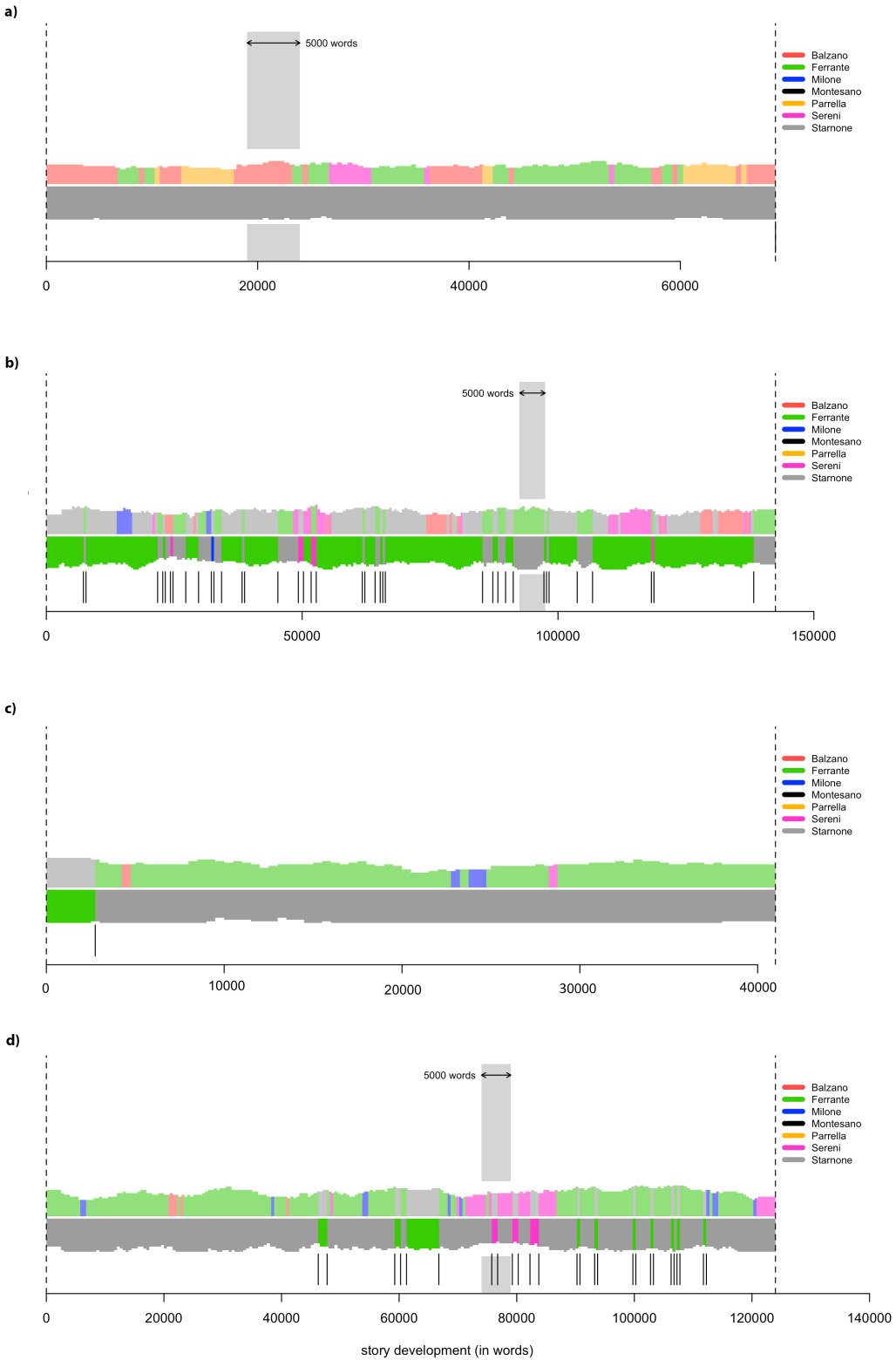
3. ábra. Bootstrap Consensus Network a Ferrante és Starnone által írt regények közötti hasonlóságokról: a) a két szerző közötti különbség: nem rajzolódik ki egyértelmű minta, b) egy időbeli jel megléte, amely erősebbnek tűnik, mint a két szerzői hang.

Az elemzés egy kis méretű konszenzushálózatból indul ki, amely kizárólag a Starnone és Ferrante regényei közötti kapcsolatok feltárására koncentrálódik. Ahogyan az a *3a. ábrán* látható, a két szerzői profil különböző klasztereket alkot, de ezek a különbségek messze nem egyértelműek. Inkább az válik feltűnővé, hogy a regények időrendi mintába rendeződnek (lásd *3b. ábra*). Még ha előzetesek is, ezek az eredmények egy további – és viszonylag erős – karakterjegyet fednek fel, amire érdemes odafigyelni: Starnone és alteregója feltételezett különálló hangját beárnyékolhatja a szerző általános stilisztikai változása az idők során.



4. ábra. Elena Ferrante regényeinek szekvenciális elemzése a Rolling Classify módszerrel. a) *L'amore molesto* (1992) b) *I Giorni dell'abbandono* (2002) c) *Storia del nuovo cognome* (2012) d) *Storia della bambina perduta* (2014). A tanítókorpusz 7 szerző 30 regényét tartalmazta, köztük Ferrante és Starnone műveit is.

 DOI 10.31400/dh-hun.2021.5.3142



5. **ábra**. Domenico Starnone regényeinek szekvenciális elemzése a Rolling Classify módszerrel. a) *Il salto con le aste* (1989) b) *Via Gemito* (2000) c) *Prima esecuzione* (2007) d) *Lacci* (2014). A tanítókorpusz 8 szerző 36 regényét tartalmazta, köztük Ferrante és Starnone műveit is.

A fent vázolt probléma feloldásához további vizsgálatsorozatokat végeztünk. Ezúttal a tanítókorpusz a feljebb említett hét szerző munkáit és Ferrante regényeit tartalmazta. Az így elvégzett vizsgálatok Ferrante és Starnone műveit külön csoportba sorolták. A kutatási kérdés a következő tehát: a klasszifikáció vajon azonosítja-e a „virtuális" Ferrantét? A hiperparaméter-beállítások ugyanazok maradtak, mint az előző vizsgálat esetében is: Support Vectors Machine módszer (SVM), 100 leggyakoribb szó, 5000 szóból álló vizsgálati „ablak", 500 szavas csúsztatással. A kísérletek eredményeit a *4a–d. ábra* szemlélteti; Ferrante elemzett regényei kronologikus sorrendben vannak feltűntetve. A mintázat nem rajzol ki teljesen egyértelmű képet: míg a korai Ferrante-regények viszonylag kevés hasonlóságot mutatnak a feltételezett „virtuális" Ferrante szerzői ujjlenyomatával, addig a kései művei esetében már nagyobb egyezést vélhetünk felfedezni a klasszifikációs eljárás alapján. A *L'amore molesto* (1992) című Ferrante-regény szinte minden szegmensét a klasszifikáció Starnonénak tulajdonítja, kivéve egy viszonylag rövid szövegrészletet a mű végén. A „virtuális" Ferrante-hang az *I Giorni dell'abbandono* (2002) című regény esetében élesebben körvonalazódik, ezúttal a szöveg elején. A *La figlia oscurá*ban (2006) a ferrantei szegmensek aránya nagyjából megegyezik a Starnonénak tulajdonított szekvenciák arányával. A *L'amica geniale. Infanzia adolescenza* (2011) esetében a szerzői profil túlnyomórészt Ferrantére utal, ezen tendencia még erősebben kimutatható a *Storia del nuovo cognome* (2012) című regényben (lásd *4c. ábra*). Ez utóbbi regényben már egy markáns virtuális szerzői hang azonosítható, hiszen a klasszifikációs eljárás az összes szegmenst Ferrantéhez kapcsolta. A két későbbi művében – *Storia di chi fugge e di chi resta* (2013) és *Storia della bambina perduta* (2014) – a szerzői ujjlenyomat már homályosabb foltokat is mutat (lásd *4d. ábra*), azonban a virtuális szerzői hang jelenléte tagadhatatlan ezen szövegek esetében is.

Bármennyire is egyértelműnek tűnnek a kapott eredmények, akár némi kételyt is felvethetnek. A vizsgálat során kinyert eredmények elég erős bizonyítékként támasztják alá azt a hipotézist, hogy Ferrante „szerzői" profilja fokozatosan alakult ki, és vált uralkodóvá a kései regényeiben. Mielőtt azonban kétségek nélkül elfogadnánk ezen állítást, meg kell vizsgálnunk, hogy Starnone szerzői hangjának stabilitása fennáll-e a vizsgált periódusban. Hiszen felmerülhet egy olyan lehetőség is, amely szerint Starnone szerzői hangja egyszerűen „ferranteivé" változott (jelen esetben lényegtelen a publikáló név). Ez a felvetés az eddig bemutatott vizsgálatok alapján nem kizárható. Sőt az előzőekben bemutatott kísérlet *(3. ábra)* már rámutatott az időbeliség központi szerepére az adathalmaz vizsgálatakor.

A Ferrante regényeivel összefüggő eredmények kiértékeléséhez így további kísérletekre volt szükség, ezúttal Starnone regényeire helyezve a fókuszt. Természetesen a paraméterbeállításokon nem módosítottunk az előző vizsgálathoz képest. Az eredményeket az *5a–d. ábra* mutatja (a tíz regényből négyet). Ezen ábrák egy egyedi, a virtuális ferrantei hangtól különböző starnonei hang létezéséről értesítenek, még ha a felvázolt kép nem is annyira egyértelmű, mint ahogyan azt előzetesen vártuk. Különösen akkor szembetűnő néhány stilisztikai sajátosság, ha időrendi sorrendbe rendezzük az elemzett regényeket. Nézzük először a legelső regényeket: az *Ex cattedra* (1987) szegmentált szövegteste egyértelműen Starnone szerzői profilját tükrözi, akárcsak az 1989-ben megjelent *Il salto con le aste (5. ábra).* A *Fuori registro* (1991) esetében nagyjából hasonló

 DOI 10.31400/dh-hun.2021.5.3142

a kép, eltekintve attól, hogy a regény elején, egy néhány ezer szavas szegmens ferrantei besorolást kapott a klasszifikációs eljárás során. Ezután nézzünk rá az *Eccesso di zelo* (1993) című műre, amelyben ismételten Starnone szerzői ujjlenyomata érvényesül, noha két rövidebb rész Ferrante szerzőségét mutatja fel. Változó tendencia mutatkozik a *Denti* (1994) regény klasszifikációjánál, mivel az eljárás fele-fele arányban kapcsolta a szövegtestet Ferrantéhoz és Starnonéhoz, viszont a *Via Gemito* (2000) című műben a módszer már egyáltalán nem azonosította Starnonét. Meglepő, hogy a klasszifikáció az utóbb említett regényt még nem is csupán Ferrantéhoz rendelte, hanem néhány kisebb szövegszegmensét Clara Sereninek (!) tulajdonította. Ezen eredmény ellentétben áll azzal a hipotézissel, miszerint Starnone képes elhatárolni egymástól a szerzői hangokat. Másrészről a következő regény, nevezetesen a *Prima esecuzione* (2007), semlegesíti az *5c. ábrá*n felmerülő kételyt, hiszen Starnone szerzői hangja ebben újra tisztává válik, alig tartalmaz olyan részeket, amelyek esetében felmerül a „ferrantei" hang. Az *Autobiografia erotica di Aristide Gambìa* című regényben a szerzőnek továbbra is adatolhatóvá válik a stabil írói profilja, még annak ellenére is, hogy néhány részlet véletlenszerűen Elena Ferrante vagy Clara Sereni szerzői hangjával kapcsolódik össze.

A két perszóna közötti váltás azonban nem egyértelmű a világhírű író esetében, erre bizonyíték a következő, *Lacci* (2014) című mű, amely egyértelműen tartalmaz olyan, viszonylagosan hosszú szegmenseket, amelyek már újra a ferrantei hangon szólalnak meg (lásd *5d. ábra*). Azonban az öt éve megjelent (2016) *Scherzetto* már újra evidenciaként szolgál arra, hogy a szerző képes sikeresen elnémítani a ferrantei szólamokat: néhány marginális szövegdarabon kívül a mű túlnyomórésze egyértelműen Starnonénak tulajdonítható. Annak ellenére is sejthető a szerzői hangok mögötti tudatosság, hogy a fenti eredmények némelyike nem teszi lehetővé a határozott konklúzió megfogalmazását. A Ferrante- és a Starnone-regények jelentős többsége képes tiszta szerzői hangot megszólaltatni. Kiderült számunkra, hogy a nápolyi író irodalmi színjátéka több mint sikeresnek tekinthető, legalábbis a feltárt stilometriai eljárások felől közelítve.

Konklúziók

A tanulmány arra vállalkozott, hogy olyan regények stilometriai alapú szerzőazonosítását mutassa be, amelyek Elena Ferrante írói álnév alatt jelentek meg. Ahogy azt vártuk, a már létező hipotézist, miszerint Domenico Starnone szerzősége lapul az álnéven publikált regények mögött, majdnem cáfolhatatlanul bizonyította a bemutatott vizsgálat. Azonban a tanulmány fő célja azt volt – a szerzőazonosítás viszonylag egyszerű kérdésén túl –, hogy választ találjon arra is, a szerzői álnév alatt megbújó író milyen mértékben, vagy egyáltalán képes-e stilisztikailag szétválasztani a két szerzői identitását.

A Rolling Classify tesztsorozat – függetlenül attól, hogy Starnone vagy Ferrante regényein alkalmaztuk – lehetővé tette az általános megfigyelések rögzítését. Az összkép, néhány kivételtől eltekintve, megerősíti, hogy Starnone és Ferrante hangja megkülönböztethető egymástól, és ez elég erős evidenciának tűnik ahhoz, hogy a virtuális szerzőség hipotézise fenntartható legyen. Domenico Starnone egyértelműen

bizonyítja – különösen a kései műveiben – hogy saját stílusprofilját képes elhatárolni szerzői alteregójától.

Azonban további vizsgálatokat igényel az a jelenség, hogy Starnone műveinek néhány részletét a klasszifikáció tévesen Clara Sereninek *(Via Gemito),* vagy Marco Balzanonak *(Lacci)* tulajdonította. A téves hozzárendeléseket magyarázhatja, hogy a klasszifikáció nem volt megfelelően betanítva a feladatra. Az ilyen példákat „hamis pozitívnak" nevezzük a gépi tanulásban. Másrészt itt foglalkozhatnánk a regények helyi stílusbeli sajátosságaival is. Hasonló módon például Harper Lee *Ne bántsátok a feketerigót!* című regényének stilometriai vizsgálatakor is lokálisan „beárnyékolt" szerzői ujjlenyomatokat figyeltünk meg,[22] Truman Capote esetenként felismert stílusa azonban korántsem utal arra, hogy a vizsgált szöveg vegyes szerzőségű lenne. Valószínűbb, hogy inkább az intertexuális hangok *(intertextual voices)* jelenségével kell számolnunk, amelyek bizonyos szöveghelyeken átsugározhatnak az – általában átlátszatlan – eredeti szerzői ujjlenyomaton. Vélhetően a *Via Gemito* és a *Lacci* esetében is hasonló jelenséggel állunk szemben.

A vizsgálat során feltárt eredmények azért is szembetűnőek, mert egy fokozatos – ugyanakkor sikeres – virtuális szerzői identitáskonstruálás művelete rajzolódik ki előttünk, amely a korai művekben alig fellelhető, míg a késeiekben már egyértelműen domináns jeggyé válik. Másrészt pedig az eredmények jelentős mértékben hozzájárulnak a fent említett szinergiahipotézis igazolásához.[23] Ez nemcsak két különböző szerzői hang stiláris kombinációjára vonatkozhat, hanem arra is, hogy két különböző stiláris személyiség rejtőzhet egyetlen tényleges szerző mögött.

Fordította: Bajzát Tímea

Elena Ferrante: A Virtual Author

The present study scrutinizes the novels by Elena Ferrante, in order to discover the actual writer hidden behind the pseudonym. Rather than simply reopen the authorship question, however, the paper attempts at testing the stability of the authorial signal in the works by "Ferrante", whoever the actual author might be. To address the research question, a network of 150 novels and their stylistic similarities has been computed using the Bootstrap Consensus Network method. A list of authors most similar to "Ferrante", including Domenico Starnone at the first place, was then analyzed using the technique Rolling Classify, which was designed to detect local stylistic idiosyncrasies in literary texts. The series of Rolling Classify tests – performed independently for the novels by both Ferrante and Starnone – allows for formulating general observations. The overall picture confirms, with a few exceptions, that Starnone and Ferrante can

[22] Maciej Eder and Jan Rybicki, „Go Set a Watchman while We Kill the Mockingbird in Cold Blood, with Cats and Other People," in *Digital Humanities: Conference Abstracts,* 184–186 (Kraków: Jagiellonian University & Pedagogical University, 2016).

[23] Pennebaker, *The Secret Life of Pronouns.*

 DOI 10.31400/dh-hun.2021.5.3142

be told apart, which, in turn, seems to be a strong argument in favor of the virtual author hypothesis. Apparently, Domenico Starnone demonstrates, particularly in his late works, the ability to differentiate his own stylistic profile and the voice of his *alter ego*.

Keywords:
stylometry, authorship attribution, virtual author, Bootstrap Consensus Network, Rolling Classify

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.

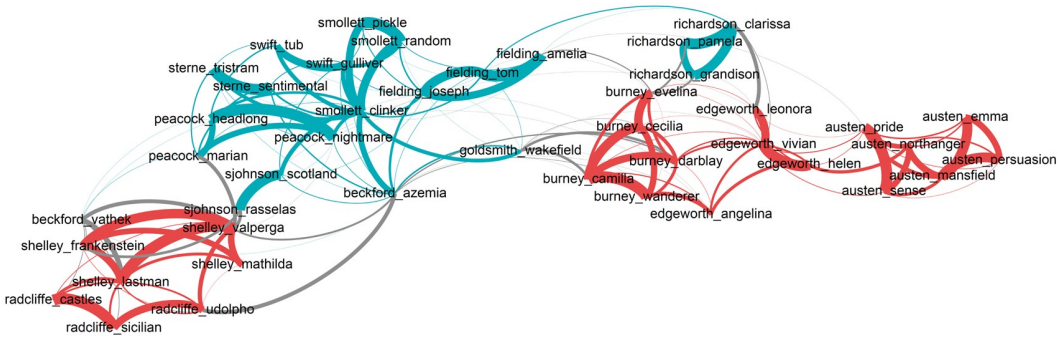
A Chawton House gyűjteményének kiegészítése a „híres férfiak" és „híres nők" referenciakorpuszával biztató eredményekre vezettek a tizenkét ismeretlen szerzőségű szöveg nemi identifikációjakor *(2. ábra)*. A férfiak csoportja változatlan, Richardson továbbra is kívülálló, de az anonim szövegek némelyike 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.

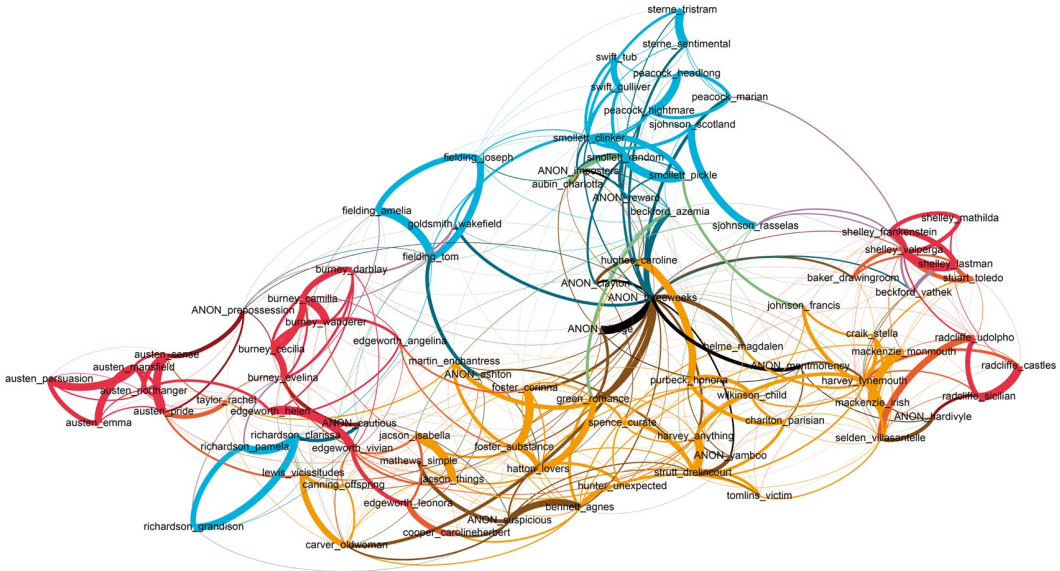
A megfelelő szólista keresése Burrows egy másik módszeréhez vezetett, a Zetához, amely azonos méretű részekre darabolja a szövegeket, majd olyan szavakat keres, amelyek következetesen felbukkannak egy szövegben, vagy szövegcsoportban, miközben jellemzően hiányoznak egy másikban.[31] Ezzel a módszerrel 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

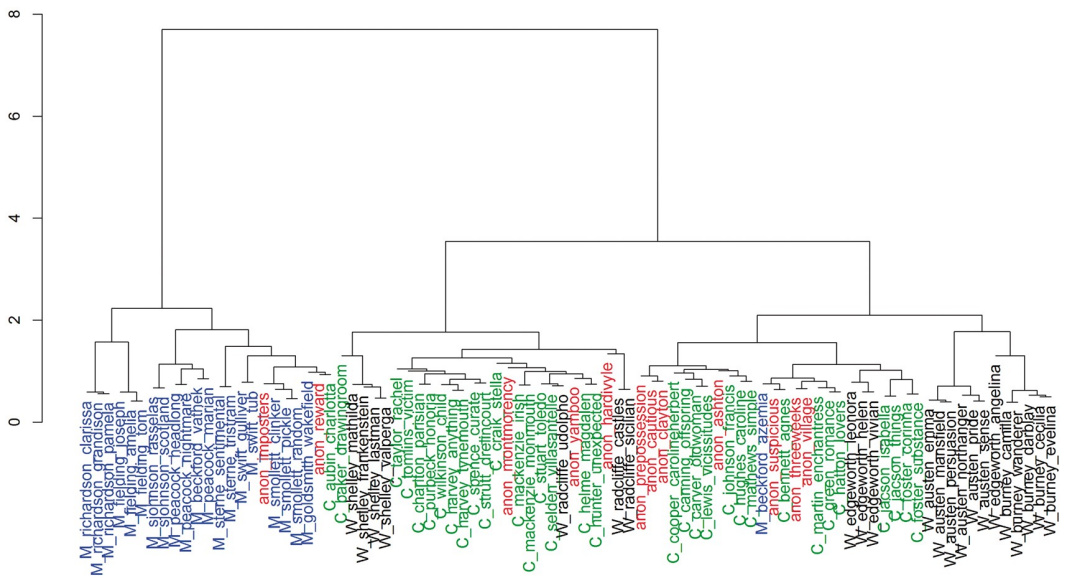
> vitást, mindentudást, pátoszt, gyorsaságot, fanyarságot, szelídséget, gyengédséget, városiasságot, lendületességet, szellemességet, kiválóságot, fiatalosságot és élénkséget, amely Kegyed tollából árad, mégis megkockáztatom, hogy ez az irodalmi szárnypróbálgatás, amelyet megtiszteltetés számomra az 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:

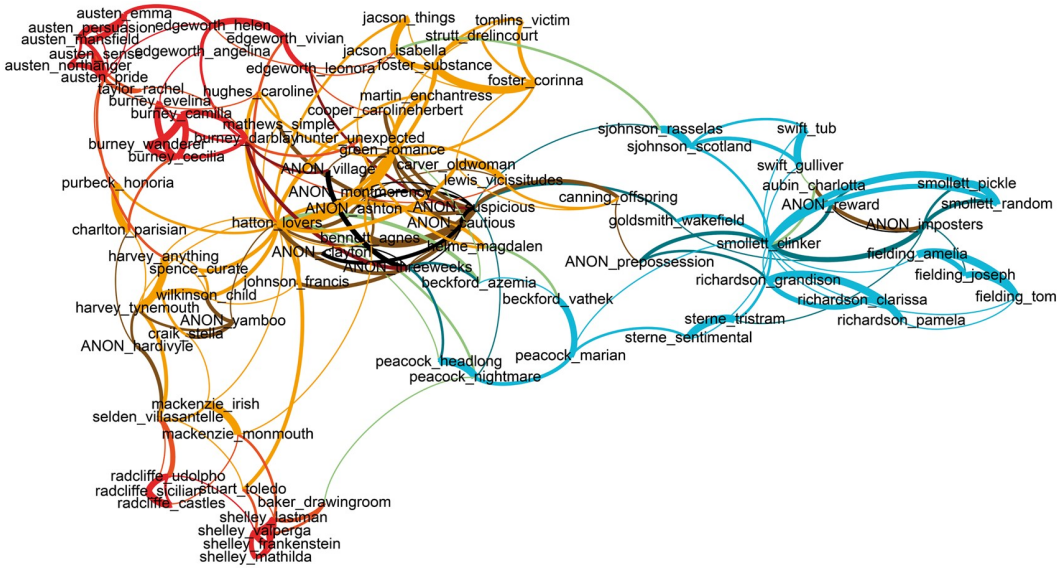
> Beszámolva arról, hogy a mostohaanyja hogyan száműzte őt Virginiába, hogyan rabolták el a hajót madagaszkári kalózok, és hogyan foglalta azt vissza egy spanyol hadfi. A spanyol Nyugat-Indiában kötött házasságáról és az ott átélt kalandjairól, valamint az Angliába való visszatéréséről. És több úriember 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:

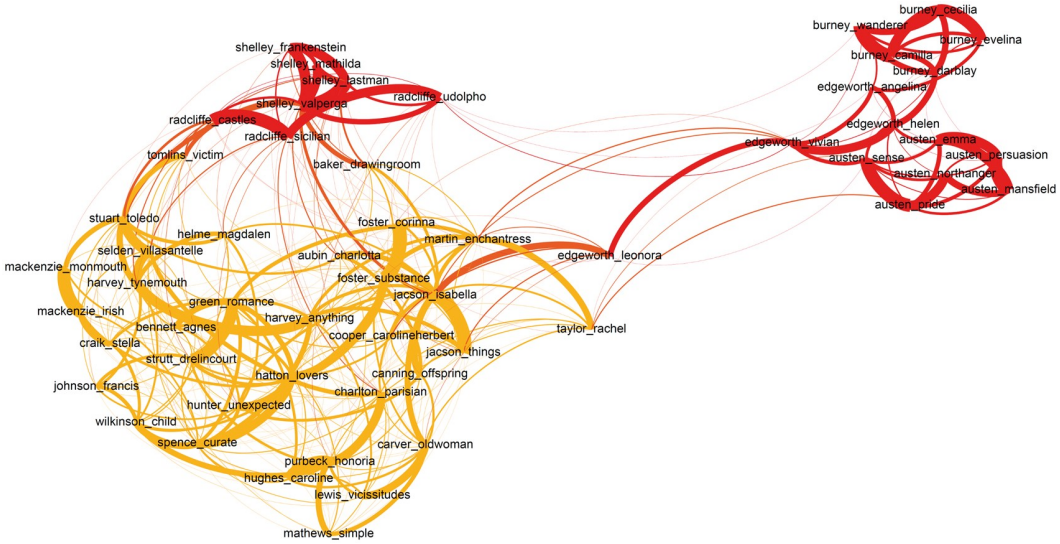
> 1. érzelmi (és azok kifejezésére szolgáló) szavak: *érzések, érezte, érez, szorongó, érezni, egyedül, fájdalmas, aggodalom, figyelem, gyönyörű, meglepetés, kedvesség, megbán, mosoly, kötődés, zaklatott, izgatott, boldogság, érzelmek, óhajtott, izgatott, barátságos, kíván, csodálat, szeretet, mohó, 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.



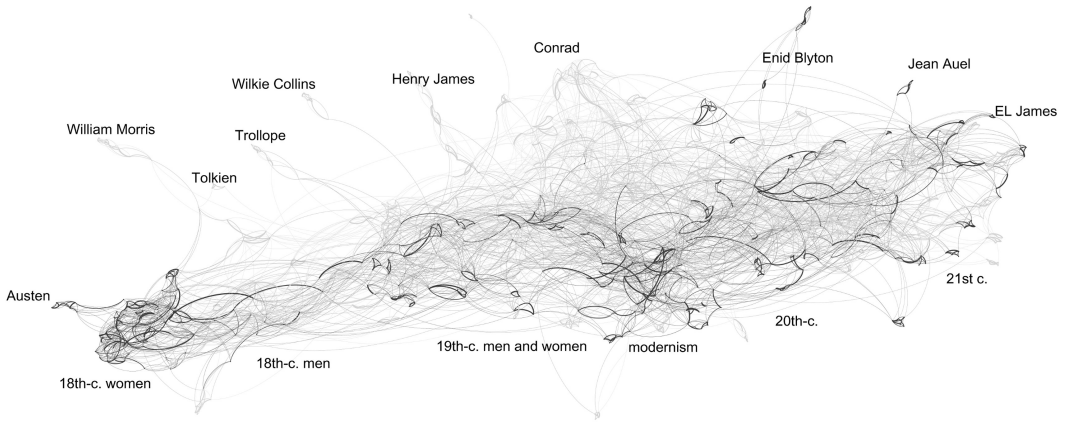
5. ábra. A hálózatelemzés során létrejött gráf a „híres nők", illetve a Chawton House ismert szerzőinek szövegeihez, a megfelelő Zeta-kulcsszavak alapján.

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.

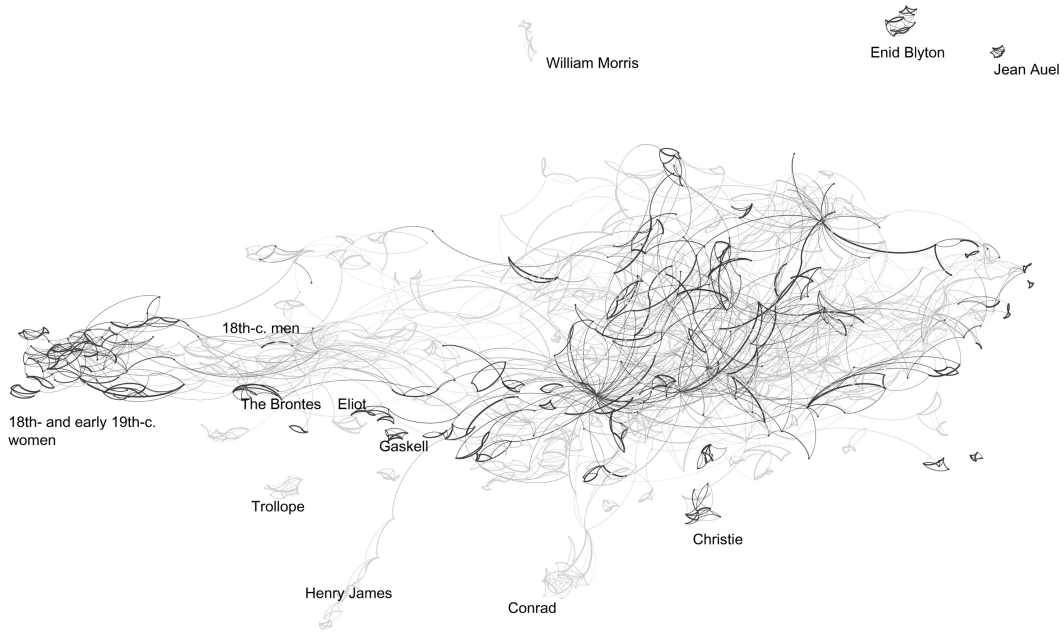
Hogy eloszlassuk vagy megerősítsük ezeket a félelmeket, egy 1000 kötetes korpuszon végeztem egy sor hálózatelemzést. Összesen valamivel több mint 111 millió tokenből áll a 635 férfi és 365 női szövegét felölelő korpusz, amely magában foglalja az összes fent vizsgált művet, amelyek jól reprezentálják 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]



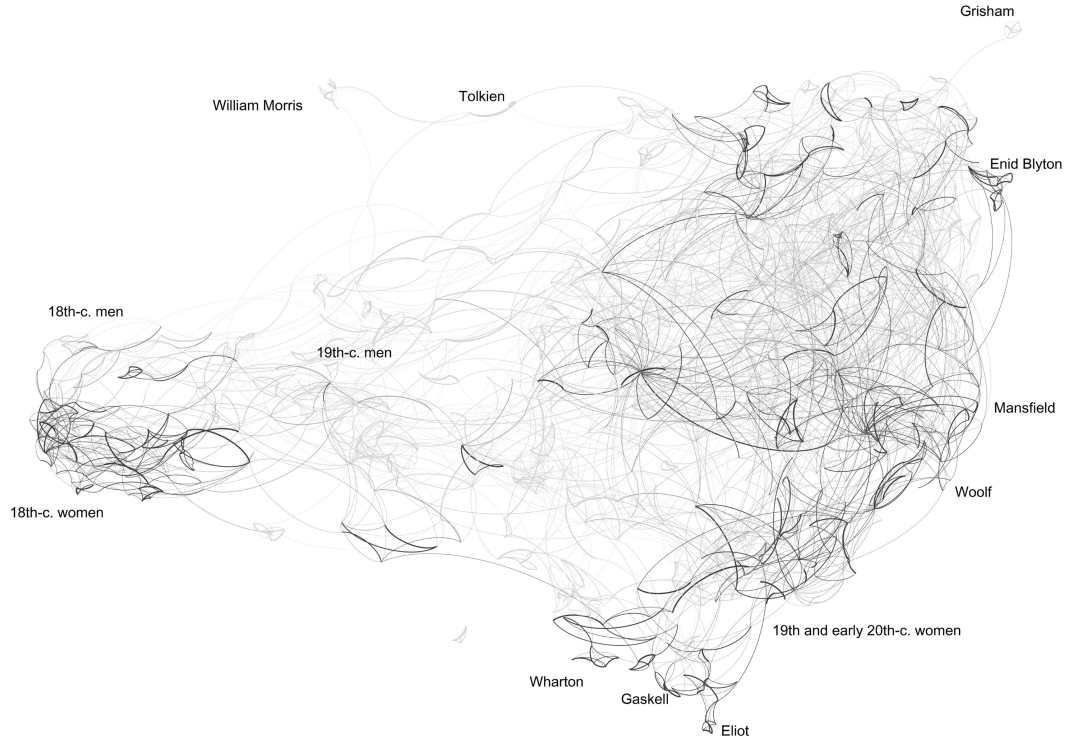
7. ábra. A hálózatelemzés során létrejött gráf 1000, férfiak (világosszürke) és nők (sötétszürke) által írt regényhez, 18. századi „híres férfiak”/„híres nők” Zeta-kulcsszavai alapján.

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

Greta Franzini 0000-0003-1159-5575

Georg-August-Universität Göttingen

greta.franzini@eurac.edu

Mike Kestemont 0000-0003-3590-693X

Universiteit Antwerpen

mike.kestemont@uantwerpen.be

Gabriela Rotari

Georg-August-Universität Göttingen

gabriela.rotari@gmail.com

Melina Jander 0000-0003-1646-6836

Georg-August-Universität Göttingen

jander@sub.uni-goettingen.de

Jeremi K. Ochab 0000-0002-7281-1852

Uniwersytet Jagielloński w Krakowie

jeremi.ochab@uj.edu.pl

Emily Franzini

Georg-August-Universität Göttingen; Decoded Ltd., London

Joanna Byszuk 0000-0003-2850-2996

Instytut Języka Polskiego PAN

joanna.byszuk@ijp.pan.pl

Jan Rybicki 0000-0003-2504-9372

Uniwersytet Jagielloński w Krakowie

jkrybicki@gmail.com

Szerzőazonosítás Jacob és Wilhelm Grimm zajos, digitalizált levelezésében*

Az alábbi cikk egy multidiszciplináris projekt eredményeit mutatja be, amely a különböző digitalizációs stratégiák számítógépes szöveganalízisben való használ-

* Eredeti megjelenés: Greta Franzini, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi K. Ochab, Emily Franzini, Joanna Byszuk and Jan Rybicki, „Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm," *Frontiers in Digital Humanities* 5 (2018), https://doi.org/10.3389/fdigh.2018.00004.

hatóságát járja körül. Pontosabban Jacob és Wilhelm Grimm szerzőségének automatizált megkülönböztetésére tettünk kísérletet, melyet egy HTR (Handwritten Text Recognition – kézzel írott szöveg felismerése) és OCR (Optical Character Recognition – optikai karakterfelismerés) által feldolgozott levelezéskorpuszban hajtottunk végre, korrekció nélkül – felmérve, hogy az így keletkezett zaj milyen hatással van a fivérek különböző írásmódjának azonosítására. Összegezve, úgy tűnik, hogy az OCR megbízható helyettesítője lehet a manuális átírásnak, legalábbis a szerzőazonosítás kérdéskörét illetően. Eredményeink továbbá abba az irányba mutatnak, miszerint még a különböző digitalizációs eljárásokból származó tanító- és tesztkorpuszok *(training and test set)* is használhatók a szerzőazonosítás során. A HTR-t tekintve a kutatás azt demonstrálja, hogy ez az automatizált átírás ugyan az OCR-hez képest szignifikánsan növeli a szövegek félrecsoportosításának veszélyét, ám körülbelül 20% feletti tisztaság már önmagában elegendő ahhoz, hogy a véletlennél nagyobb esélye legyen a helyes bináris megfeleltetésnek.

Kulcsszavak:
stilometria, szerzőazonosítás, német irodalom, Grimm, digitalizáció, OCR, HTR

1. Bevezetés

Több tanulmány is beszámol arról, hogy adatelemzéssel foglalkozó kutatók a kutatási idejüknek akár 80%-át is az adatok előkészítésével tölthetik, míg csak 20%-ot szentelnek magukra a kutatási kérdésekre.[1] Ez az egyensúlyhiány azt sugallja, hogy a kutatók abban a hitben dolgoznak, hogy az előkészítésbe fektetett idő egyenesen arányos az eredmények minőségével – más szavakkal: a magas minőségű munka összeegyeztethetetlen az alacsony minőségű, zajos adatokkal.

Jelen tanulmány erre a jelenségre adott válasznak is tekinthető: olyan kutatásokra épül, melyek a számítógépes szerzőazonosításban elfogadható mértékű digitalizációs zajokat járják körül,[2] és arra törekszik, hogy megbízható modellt adjon Jacob és Wilhelm Grimm[3] kézírásának átírására, illetve hogy képes legyen meghatározni szerzőségüket levelezésük alapján.[4] Ennek érdekében elemzéseket futtatunk le a szövegek nyomtatott kiadásának OCR-rel feldolgozott és nem korrigált változatain, valamint a

[1] Például Hadley Wickham, „Tidy Data," *Journal of Statistical Software* 59, 10. sz. (2014), https://doi.org/10.18637/jss.v059.i10.

[2] A téma 2015 óta az EMNLP éves workshopjának is tárgya. Hozzáférés: 2021.07.30, https://noisy-text.github.io/2017/.

[3] Jacob (1785–1863) és Wilhelm Grimm (1786–1859) német írók, tudósok és akadémikusok, akik a 19. század folyamán népmesék gyűjtésével és publikálásával váltak ismertté.

[4] A cikk egy, a németországi Göttingeni Egyetem hat hónapos kísérleti projektjének keretében végzett kutatásokat ismerteti. A TrAIN (Tracing Authorship In Noise) néven ismert projekt a zajos OCR- és a HTR-adatok számítógépes szövegelemzésére gyakorolt hatását kívánta vizsgálni. A projekt 2016 júliusa és 2017 januárja között zajlott. A projekt weboldala, hozzáférés: 2021.07.30 http://www.etrap.eu/research/tracing-authorship-in-noise-train/.

 DOI 10.31400/dh-hun.2021.5.3144

kézzel írt dokumentumok digitalizált képeiből készített, szintén korrigálatlan HTR-verziókon egyaránt. A két verzió alapvetően eltérő feldolgozási zajokat emel a diskurzusba, számunkra hasznos összehasonlítási szempontokat biztosítva. E tanulmányban ugyanis a következő kérdést tesszük fel: az OCR és a HTR során keletkezett zaj milyen mértékben befolyásolja a Grimm testvérek egyéni stílusának *(stylome)*[5] azonosítását? A munkával a jelen technológia lehetőségeit térképezzük fel, miközben nagyobb rálátást kívánunk biztosítani Jacob és Wilhelm Grimm stílusára is. Akadályként a következő tényezőkkel kell számolnunk: az eredeti szövegek digitális átalakítása során keletkezett textuális zajok torzító hatása, valamint a rendelkezésünkre álló adatok sokszínűsége és mennyisége. A tanulmány szerkezete az alábbiak szerint alakul: az 1. és 2. rész a projekt motivációját tárgyalja, a 3. részben a kutatás alapját képező anyagról lesz szó, míg a 4. az anyag digitalizációját és a nem korrigált szövegeken végzett szerzőazonosítás lépéseit mutatja be, továbbá itt tárgyaljuk a kutatás eredményeit is. Végül az 5. rész az összegzés és további kitekintések megtétele számára biztosít teret.

2. Kapcsolódó kutatások

Az online elérhető szövegekkel dolgozó kutatóknak gyakran zajos vagy strukturálatlan adatokkal kell megküzdeniük. Egészen pontosan kétféle zaj nehezíti munkájukat: 1. amely a szöveg előállításakor keletkezik (úgy mint: helyesírási hiba, a standardtól eltérő szóalak, speciális karakterek, szándékolt rövidítések, nyelvtani hibák stb.); 2. amely a szöveg más formátumba történő konvertálása során (úgy mint: digitalizáció vagy digitális transzformáció) jön létre.[6] Az utóbbi típus rendszerint az interneten fellelhető irodalmi szövegeknél, az OCR-rel vagy HTR-rel feldolgozott nyomtatott vagy kézírásos művek esetében gyakori. Ezek kapcsán az alábbiak mondhatók el: a történeti szövegeknél az OCR pontossága a karakterek szintjén a 95%-ot is meghaladhatja,[7] bár a forrás típusától függően ez a szám lehet alacsonyabb is (klasszikus szövegek kritikai kiadása versus ősnyomtatványok), a HTR esetében pedig már 80–90% közé esik, a kézírás tisztaságától függően.[8] Lopresti az OCR-hibáknak az információvisszakeresésre (Information Retrieval – IR) és a természetes nyelvfeldolgozásra (Natural Language Processing – NLP) gyakorolt hatásával is foglalkozik,[9] és habár léteznek a zaj

[5] Hans van Halteren, Harald Baayen, Fiona Tweedie, Marco Haverkort and Anneke Neijt, „New Machine Learning Methods Demonstrate the Existence of a Human Stylome," *Journal of Quantitative Linguistics* 12, 1. sz. (2005): 65–77, https://doi.org/10.1080/09296170500055350.

[6] L. Venkata Subramaniam, Shourya Roy, Tanveer A. Faruquie and Sumit Negi, „A Survey of Types of Text Noise and Techniques to Handle Noisy Text," in *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data – AND '09,* 115–122 (Barcelona: ACM Press, 2009), 115, https://doi.org/10.1145/1568296.1568315.

[7] Florian Fink, Klaus U. Schulz and Uwe Springmann. „Profiling of OCR'ed Historical Texts Revisited," in *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage,* 61–66 (Göttingen Germany: ACM, 2017), https://doi.org/10.1145/3078081.3078096.

[8] Sumeet Agarwal, Shantanu Godbole, Diwakar Punjani and Shourya Roy, „How Much Noise is Too Much: A Study in Automatic Text Classification," in *Seventh IEEE International Conference on Data Mining (ICDM 2007),* 3–12 (Omaha: 2007), 5, https://doi.org/10.1109/ICDM.2007.21.

[9] Daniel Lopresti, „Optical Character Recognition Errors and Their Effects on Natural Language Processing," *International Journal on Document Analysis and Recognition (IJDAR)* 12, 3. sz. (2009): 141–151, https://doi.org/10.1007/s10032-009-0094-8.

mértékét és a felismerés pontosságát meghatározó módszerek,[10] sőt félig automatizált eljárások is segíthetik a tévesztések javítását és a korabeli helyesírási normának a gépi hibától való megkülönböztetését,[11] mégsem véletlen, hogy egyes tanulmányok arra következtetnek, hogy az adatelemzéssel dolgozó kutatók a kutatási idő akár 80%-át is az adatok előkészítésével tölthetik el.[12]

Az előkészítés idejének megrövidítése érdekében érdemes az algoritmusok zajtoleranciájára vonatkozó teszteket végezni. Agarwal és munkatársai például egy egész sor kísérletről számolnak be, amelyek a digitalizáció során keletkezett hibáknak a szövegcsoportosító algoritmusokra gyakorolt hatását tesztelték, és amelyek alapján megállapították, hogy az osztályozás pontossága akár 40%-ig is tolerálja a bevezetett zajokat.[13] A stilometriában Eder angol, német, lengyel, ógörög és latin prózaszövegeken mutatta be több szerzőazonosításhoz használt módszerének stabilitását: kutatásának eredménye, hogy a zajtolerancia ugyan eltérő a különböző nyelveknél, de még a 20%-os torzulás sem befolyásolja szignifikánsan a módszerek teljesítményét.[14] A számítógépes szerzőazonosítás jelenlegi megközelítései ugyanis olyan jellemzőkre irányulnak, mint a szavak unigrammáinak vagy karakter n-gramoknak az eloszlása;[15] ezek pedig nagyon gyakori és átfogó elemek egy szövegben (és sokkal kevésbé ritkák, mint például a jelentéses szavak), ami magyarázhatja, hogy miért ellenállók a jelentős mértékű, látszólag sztochasztikus zajjal szemben. Továbbá bizonyos szabályozási technikák (például a Support Vector Machine osztályozó eljárása) segítségünkre lehetnek abban, hogy ne essünk a túlillesztés *(overfitting)* csapdájába a zajos környezetben. Mindeddig azonban még egyetlen szisztematikus tanulmány sem modellezett le ilyesfajta zajt, miközben a kutatók az NPL-szoftverek használatával rutinszerűen normalizálják eredményeiket.[16]

[10] Subramaniam et. al, „A Survey of Types of Text Noise," 117–118.

[11] Thorsten Vobl, Annette Gotscharek, Uli Reffle, Christoph Ringlstetter and Klaus U. Schulz, „PoCoTo – an Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage – DATeCH '14,* 57–61 (Madrid: ACM Press, 2014), https://doi.org/10.1145/2595188.2595197. Példának lásd a CIS-LMU *Post Correction Tool*ját (PoCoTo). Hozzáférés: 2021.07.30, https://www.digitisation.eu/tools-resources/tools-for-text-digitisation/cis-lmu-post-correction-tool-pocoto/.

[12] Wickham, „Tidy Data."

[13] Agarwal et. al., *How Much Noise.*

[14] Maciej Eder, „Mind Your Corpus: Systematic Errors in Authorship Attribution," *Literary and Linguistic Computing* 28, 4. sz. (2013): 603–614, 612, https://doi.org/10.1093/llc/fqt039.

[15] Bradley Kjell, W. Addison Woods and Ophir Frieder, „Discrimination of Authorship Using Visualization," *Information Processing & Management* 30, 1. sz. (1994): 141–150, https://doi.org/10.1016/0306-4573(94)90029-9; Efstathios Stamatatos, „On the Robustness of Authorship Attribution Based on Character N-Gram Features," *Journal of Law and Policy* 21, 2. sz. (2013): 421–439; Mike Kestemont, „Function Words in Authorship Attribution: From Black Magic to Theory?" in *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL),* 59–66 (Gothenburg: Association for Computational Linguistics, 2014), https://doi.org/10.3115/v1/W14-0908.

[16] Például az alábbi esetekben: Patrick Juola, „Authorship Attribution," *Foundations and Trends in Information Retrieval* 1, 3. sz. (2007): 233–334, http://doi.org/10.1561/1500000005; Stamatatos, „On the Robustness;" Moshe Koppel, Jonathan Schler and Shlomo Argamon, „Computational Methods in Authorship Attribution," *Journal of the American Society for Information Science and Technology* 60, 1. sz. (2009): 9–26, https://doi.org/10.1002/asi.20961.

 DOI 10.31400/dh-hun.2021.5.3144

A szerzőazonosítás és a stilometriai analízis eredményeinek megbízhatóságát a szövegminőségen túl a minta mérete (vagyis az elemzett szavak száma) is befolyásolja. Eder egy, a témában írt cikkében előző megállapításait[17] cáfolva a minimális elfogadható mintaméret küszöbét 2000 szóban határozza meg – amennyiben a szerzői ujjlenyomat a szövegben markánsnak mondható.[18] Ebből az következik, hogy noha a nagy méretű minták általánosan előnyösebbek, a kis méretűek (2000 szóig) is képesek lehetnek pontos eredményeket biztosítani.

3. Anyagok

3.1. A Grimm-levelezés

Az OCR és HTR során keletkezett zajnak a szerzőazonosításra gyakorolt hatását a Grimm család levelezésének egy részén teszteltük. Ez a diakrón szövegegyüttes kifejezetten alkalmas a feladatra, mivel a kézírásos dokumentumok és a nyomtatott kiadás egyaránt rendelkezésre állnak.

3.1.1. Kézírásos levelek ♦ 2015 októberében jutottunk hozzá a Grimm család körülbelül 36000 levelének digitalizált korpuszához a marburgi Állami Levéltárnak köszönhetően.[19] Ezek közt több olyan levél található, melyeket Jacob és Wilhelm Grimm egymással, illetve ismerőseikkel váltottak több mint 70 év leforgása alatt: a szerteágazó (a betegségektől az utazásokig terjedő) témájú levelek a testvérek életének és stilisztikai fejlődésének tanújaként is szolgálhatnak. A Marburg-gyűjtemény azonban nem teljes: további 1000 hivatalos levelet őriz a berlini Humboldt Egyetem is,[20] ám az ezeknek a megszerzésére irányuló tárgyalások még nem zárultak le a Grimm Levelezés Központjával.

Mivel a kutatás tárgya Jacob és Wilhelm Grimm kézírásának vizsgálata, a teljes Marburg-gyűjteményből csupán a fivérek által írt leveleket választottuk ki.

3.1.2. A levelek nyomtatásban megjelent kritikai kiadása ♦ A testvérek levelezésének egyetlen kritikai kiadása Heinz Rölleke 2001-ben megjelent *Jacob és Wilhelm Grimm levelezése* című, két kötetet számláló könyve.[21] A kiadvány előszavában az olvasható, hogy a kritikai kiadás a szövegek közlésekor követi az eredeti szövegekhez hű szerkesztői konvenciót.[22] Rölleke azonban apró változtatásokat eszközöl: mind a gon-

[17] Maciej Eder, „Does Size Matter? Authorship Attribution, Small Samples, Big Problem," *Digital Scholarship in the Humanities* 30, 2. sz. (2015): 167–182, https://doi.org/10.1093/llc/fqt066.

[18] Maciej Eder, „Short Samples in Authorship Attribution: A New Approach," *Digital Humanities 2017: Conference Abstracts,* 221–224 (Montreal: McGill University, 2017), 223.

[19] A TIFF-fájlok és a publikáció jogait a marburgi Hesseni Állami Levéltártól vásároltuk meg. A gyűjtemény neve: *340 Grimm.* Bővebben, hozzáférés: 2021.07.30, https://landesarchiv.hessen.de/.

[20] Lásd bővebben, hozzáférés: 2021.07.30, http://www.grimmbriefwechsel.de/arbeitsstelle/arbeitsstelle.html.

[21] *Briefwechsel der Brüder Jacob und Wilhelm Grimm: Kritische Ausgabe in Einzelbänden,* Hg., Heinz Rölleke Bd. 1, in 3 Teil (Stuttgart: Hirzel Verlag, 2001).

[22] A szerkesztői jegyzet teljes szövegét német nyelven lásd: Uwe Meves und Jens Haustein, „Vorwort," in Rölleke, *Briefwechsel der Brüder,* 1.1: 5–8, http://www.grimmnetz.de/bwfiles/!grimm-bw1-1_kopie.pdf.

dolatjeleket, mind a nagykötőjeleket gondolatjellé egységesíti a szókapcsolatokban; pótolja a hiányzó központozást; a szokatlan rövidítéseket dőlttel szedve közli szögletes zárójelben; kiteszi a hiányzó umlautokat (de nem jelzi a mulasztások tényét); hiányzó karaktereket pótol anélkül, hogy meghatározná a helyüket a kéziratban; nem jelöli sem a pecsétes helyeket, sem pedig a kihúzott szövegrészeket.

3.2. A kutatásba felvett levelek

A Marburg-korpuszból 85 levél került kiválasztásra a kutatáshoz – ebből 50 Jacob Grimmnek, a fennmaradó 35 pedig Wilhelm Grimmnek tulajdonítható. Ezek többnyire egymásnak vagy egy rokonuknak, Karl Weigandnak címzett levelek. Weigand a kor szerzője és filológusa, aki Grimmékkel együtt részt vett a *Deutsches Wörterbuch (Német Nagyszótár)* elkészítésének munkálataiban. Az *1.* és *2. táblázat* időbeli sorrendben mutatja a Jacobtól, illetve Wilhelmtől származó leveleket.

 DOI 10.31400/dh-hun.2021.5.3144

1. táblázat. Jacob Grimm 50 levele. Az olvashatóságra vonatkozó szakértői értékelés szögletes zárójelben szerepel. A második korszakot (1800) a második HTR-modell során vettük fel a korpuszba.

LETTERS WRITTEN BY JACOB GRIMM: 50			
Epoch	**Letter ID**	**Year**	**Readability**
1. 1793	Br 5995	1793	low [to v. low]
2. 1800	Ms 237	1800	low
3. 1805-1806	Br 2164	1805	low
	Br 2165	1805	low
	Br 2169	1805	low
	Br 2163	1805	low [to v. low]
	Br 2166	1805	low
	Br 2167	1805	low
	Br 2168	1805	low [to v. low]
	Br 2170	1805	low
	Br 2176	1805	very low
	Br 2174	1806	low
4. 1814-1863	Br 2175	1833	low
	Br 2171	1833	medium [to high]
	Br 2172	1838	medium [to high]
	Br 5996	1838	medium
	Br 2237	1840	medium
	Br 2238	1840	low
	Br 2239	1840	low [to medium]
	Br 2240	1841	high
	Br 2241	1844	very low [to medium]
	Br 2242	1846	very low [to medium]
	Br 2243	1847	medium
	Br 2173	1848	low
	Br 2269	1848	low
	Br 2244	1849	low [to medium]
	Br 2245	1849	low
	Br 2268	1850	low
	Ms 131	1850	high
	Br 2246	1852	low
	Br 2247	1853	low
	Br 2248	1854	low
	Br 2249	1855	low
	Br 2250	1856	low
	Br 2266	1857	low
	Br 2251	1858	low
	Br 2252	1858	low
	Br 2253	1859	low
	Br 2254	1859	low
	Br 2255	1859	low
	Br 2267	1859	low
	Br 2256	1860	low
	Br 2257	1860	low
	Br 2258	1861	medium [to low]
	Br 2259	1861	medium [to low]
	Br 2260	1861	low
	Br 2261	1862	very low
	Br 2262	1862	low
	Br 2263	1862	low
	Br 2264	1863	very low
	Br 2265	1863	very low

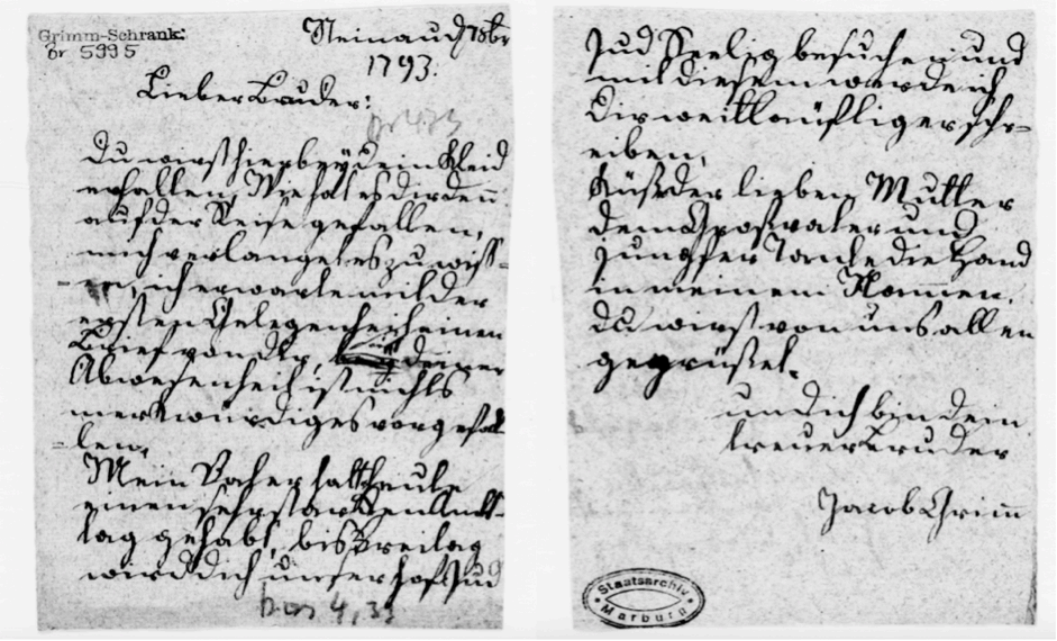
2. táblázat. Wilhelm Grimm 35 levele. Az olvashatóságra vonatkozó szakértői értékelés szögletes zárójelben szerepel.

LETTERS WRITTEN BY WILHELM GRIMM: 35			
Epoch	**Letter ID**	**Year**	**Readability**
1. 1793	Br 5993	1793	low
	Br 5994	1793	low
	Br 2678	1793	low
	Br 2679	1793	low
2. 1802-1805	Br 2677	1805	low
3. 1831-1843	Br 2680	1831	low
	Ms 426 Bl 7	1833	very low [to low]
	Ms 428 Bl 7b	1833	very low [to low]
	Ms 426 Bl 10	1833	very low
	Ms 426 Bl 11	1833	very low
	Ms 426 Bl 13	1833	very low
	Ms 426 Bl 15	1833	very low
	Br 1687	1843	low
	Br 2681	1843	very low [to medium]
	Br 1688	1843	low
4. 1846-1859	Br 2734	1846	medium [to high]
	Br 2682	1847	low
	Br 2683	1848	medium
	Ms 161	1850	low
	Br 2735	1851	high [low to medium or high]
	Br 2736	1855	high [low to medium or high]
	Br 2684	1856	high [low to medium or high]
	Br 2685	1856	medium
	Br 2687	1856	medium
	Br 2686	1856	high [low to medium or high]
	Br 2688	1856	medium
	Br 2689	1856	medium
	Br 2737	1857	medium
	Br 2690	1858	low
	Br 2738	1858	medium
	Br 2739	1858	low
	Br 2740	1859	low
	Br 2741	1859	low
	Br 2742	1859	low
	Br 2743	1859	medium

 DOI 10.31400/dh-hun.2021.5.3144

3.2.1. A levelek kategorizálása: korszakok és olvashatóság ♦ A 85 levelet korszakok és olvashatóság mentén manuálisan is kategorizáltuk.

3.2.1.1. Korszakok * Az idő előrehaladtával a testvérek írásképe változott. Ezek a változások legjobban akkor észlelhetők, ha a leveleket egymás mellett vizsgáljuk. Így például Jacob kézírása az 1805–1806 közti időszakban (20–21 éves korában) látványosan különbözik az élete végére kialakult írásképtől. Az *1. ábrá*n a gyűjtemény legkorábbi, Jacob 8 éves korában írt levele látható 1793-ból.



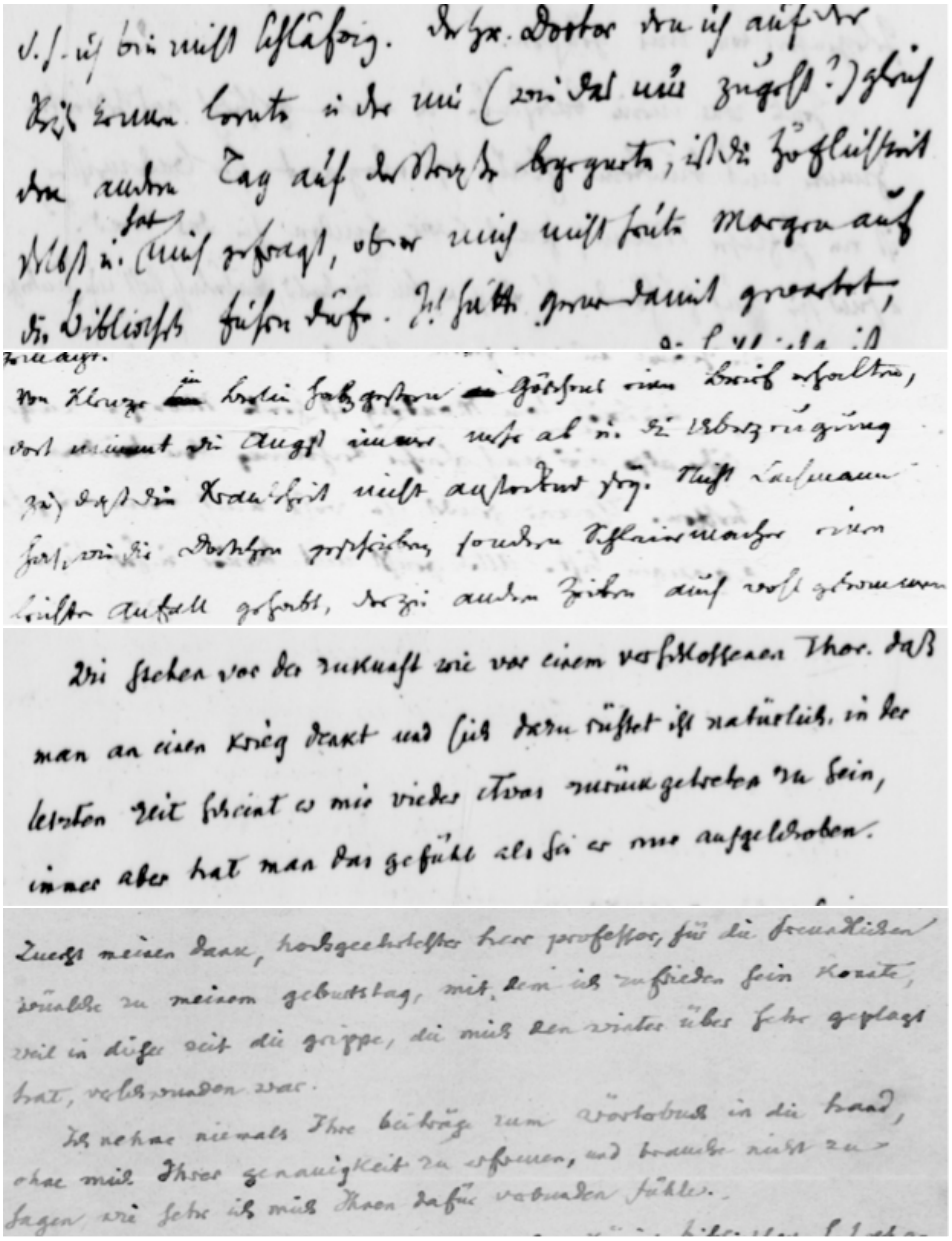
1. ábra. Jacob Grimm levele 1793-ból. A levél teljes átirata a *Függelék*ben olvasható.

A változások mentén a leveleket kézírásos periódusok, korszakok szerint csoportosítottuk: eszerint fejenként négy csoportra oszthatók a fivérek levelei. Jacob korszakai: 1793, 1800, 1805–1806, 1814–1863,[23] míg Wilhelméi: 1793, 1802–1805, 1831–1843 és 1846–1859.

3.2.1.2. Olvashatóság * A leveleket négy csoportra osztottuk olvashatóságuk alapján is. Ezek a csoportok: nagyon alacsony *(very low),* alacsony *(low),* közepes *(medium)* és magas *(high)* olvashatóság, mint ahogy az a *2. táblázat*ban is látható. Az olvashatóságot *(readability)* a papír minősége (a rossz minőségű papíron kiütköznek a tintafoltok) és a kézírás kiolvashatósága *(legibility)* befolyásolja. A Grimm-kutatókkal való egyeztetés alapján az állapítható meg, hogy az olvashatóság kritériumai a testvérek kézírásának szabályosságában, az egyértelműen megkülönböztethető karakterhosszúságban, valamint a szóvégi betűk önkényes elhagyásában ragadható

[23] Jacob Grimm itt jelölt utolsó korszakában megfigyelhető ugyan némi változás a kalligráfiát illetően, de ez nem akkora mértékű, hogy ezért indokolt volna a negyedik csoportot továbbosztani.

meg.[24] Ebből származtatható az a megfigyelés is, hogy általánosságban mind Jacob, mind Wilhelm korai írásai nehezebben olvashatók. A levélcsoportok ilyen kronológiai elkülönítése szükséges egy megbízható HTR-modell betanításához, amely a korpusz egészét tekintve is a lehető legalacsonyabb hibaarányt produkálja.



2. ábra. Négy levél Wilhelm Grimmtől. Fentről lefelé haladva: nagyon alacsony olvashatóság (Br 5993, 7 évesen); alacsony olvashatóság (Br 2680, 45 évesen); közepes olvashatóság (Br 2743, 73 évesen); magas olvashatóság (Br 2736, 69 évesen).

[24] Bernhard Lauer és Rotraut Fischer a Brüder Grimm-Gesellschafttól ['Grimm testvérek Egyesület'], Kassel, Németország.

 DOI 10.31400/dh-hun.2021.5.3144

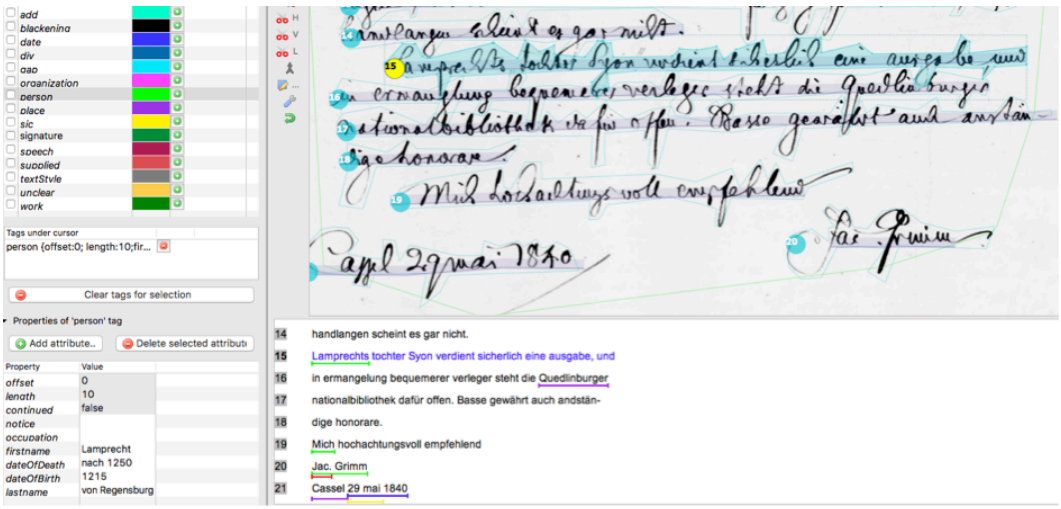
4. Módszerek és eredmények

4.1 Digitalizáció: manuális átírás (MAN), HTR és OCR

Az alábbi rész a Grimm-levelezés háromféle digitalizálási módszerét részletezi, és a címben jelzett utóbbi kettőt (HTR és OCR) össze is veti egymással.

4.1.1. Manuális átírás ♦ A levelek automatikus és manuális átírására a legújabb technológián alapuló *Transcribus* szoftvert alkalmaztuk.[25] Azonban egy olyan HTR-modell létrehozásához, amely lehetővé teszi a kézírásos dokumentumok automatizált átírását, a *Transcribus* programnak minimum száz oldal terjedelmű, manuálisan átírt szövegre van szüksége.[26] A 85 darab így átírt levél hossza változó – néhány csupán egyoldalas, néhány több oldalt is felölel. Jacob 50 levele 90 oldalnyi átírt szövegnek felel meg, míg Wilhelm 35 levele 64 oldalt tesz ki: összesen tehát 154 manuálisan átírt oldalról beszélhetünk a vizsgált szövegek esetében.

Amellett, hogy diplomatikus átírást készítettünk,[27] a leveleket metaadatokkal is elláttuk a fivérek írásképét és a tartalmat illetően – miként az a *3. ábrán* is látható.[28]



3. ábra. A *Transcribus* felülete. A képernyő három területre oszlik: a bal oldalsáv a metaadatokat és a jegyzetelési funkciókat; a felső terület az eredeti dokumentumot (Br 2238, Jacob Grimm); az alsó terület pedig a jegyzetekkel ellátott átiratot tartalmazza.

[25] A *Transcribus*ról több információért lásd, hozzáférés: 2021.07.30, `https://transkribus.eu/Transkribus/`.

[26] További információért lásd a *Transkribus* wiki-oldalát: `https://transkribus.eu/wiki/index.php/Questions_and_Answers/What_is_needed_for_the_HTR_to_work.3F`.

[27] A „diplomatikus átírás" definíciójához lásd: *Lexicon of Scholarly Editing,* hozzáférés: 2021.11.18, `https://lexiconse.uantwerpen.be/lexicon/transcriptionDiplomatic.html`.

[28] Melinda Jander, „Handwritten Text Recognition – Transkribus: A User Report," in *Göttingen, Germany: eTRAP Research Group* (Göttingen: University of Göttingen, 2016), hozzáférés: 2021.11.18, `http://www.etrap.eu/wp-content/uploads/2016/11/TrAIN-Transkribus_User_Report-2016.pdf`.

4.1.2. A kézírásos levelek HTR-modellje ♦ A manuális átírásokat a HTR-modell kidolgozására használtuk, hogy az képes legyen felismerni és automatizáltan átírni további, a fivéreknek tulajdonítható leveleket és dokumentumokat (mint például a már említett 1000 darab levelet a berlini gyűjteményből).

Mint említettük, egy megbízható HTR-modellnek ideális esetben egy legalább 100 oldalnyi kézzel átírt szöveget tartalmazó tanítókorpuszra van szüksége.[29] A tény, hogy Jacob és Wilhelm Grimm 85 levele összesen 154 oldalnyi szöveget tesz ki, a következő választás elé állított minket: inkább a HTR 100 oldalnyi szövegszükségletét elégítsük ki úgy, hogy a testvérek leveleit egyesítjük, vagy két különálló modellt képezzünk ki, egyet-egyet a fivérek számára, viszont kevesebb terjedelemben (90 oldal Jacobnak és 64 oldal Wilhelmnek). Végül úgy döntöttünk, hogy mindkét lehetőséget teszteljük, az eredményeket pedig összevetjük egymással.[30]

4.1.2.1. Az első HTR-modell * Az első HTR-próbakör során Jacob és Wilhelm mind a 85 manuálisan átírt levelét (154 oldal és 26983 szó) felhasználtuk. Amellett, hogy így eleget tettünk a minimum oldalszám követelményének, a két kézírás egyesítése és egy ilyen HTR-modell megalkotása mögött meghúzódó feltételezésünk az volt, hogy a kevert modell nagyobb ellenállóságot mutat majd a testvérek kézírásának diakrón változásaival szemben. A karakterhiba-arány (CER – Character Error Rate) a kevert modell eredményeként 18,83% volt – azaz a szöveg minden ötödik karakterét helytelenül ismerte fel a program. Ennek javítása érdekében további 2000 szónyi (17 oldalnak megfelelő) Grimm-kézírást írtunk át manuálisan. A várakozással ellentétben az új hibaarány 40%-ra nőtt – azaz minden kettő és feledik karakter hibásan szerepelt az átiratban. Alaposabb vizsgálat után rájöttünk azonban, hogy 13 nagyon alacsony olvashatóságú levél felelős ezért a magas karakterhiba-arányért. Ennek csökkentése céljából egy olyan második HTR-kört futtattunk le, amelyben a 13 problematikus levelet 11 másik, a fivérek által írt dokumentumra cseréltük (35 oldal, 5788 szó).[31] A *3.* és *4. táblázat* az elvett és hozzáadott dokumentumokat listázza.

[29] További információért a HTR-hez szükséges adathalmazképzéssel kapcsolatban lásd: `http://read.transkribus.eu/wp-content/uploads/2017/01/READ_D7.7_HTRbasedonNN.pdf`.

[30] A HTR-modellt a *Transcribus* használatával dr. Günther Mühlberger hozta létre.

[31] A dokumentumok forrása a marburgi Hesseni Állami Levéltár weboldala, lásd, hozzáférés: 2021.07.30, `https://arcinsys.hessen.de/arcinsys/detailAction.action?detailid=g195109&icomefrom=search`.

 DOI 10.31400/dh-hun.2021.5.3144

3. táblázat. A HTR-tesztkorpuszból alacsony olvashatóságuk miatt kizárt levelek, amelyek negatívan hatottak az első HTR-kör működésére.

LETTERS DISCARDED FROM HTR CORPUS			
Author	**Letter ID**	**Year**	**Readability**
Jacob	Br 2176	1805	very low
	Br 2241	1844	very low
	Br 2242	1846	very low
	Br 2261	1862	very low
	Br 2264	1863	very low
	Br 2265	1863	very low
Wilhelm	Ms 426 Bl 7	1833	very low
	Ms 426 Bl 7b	1833	very low
	Ms 426 Bl 10	1833	very low
	Ms 426 Bl 11	1833	very low
	Ms 426 Bl 13	1833	very low
	Ms 426 Bl 15	1833	very low
	Br 2681	1843	very low

4. táblázat. A HTR-tesztkorpuszba újonnan felvett levelek magas és közepes olvashatósággal, amelyek az előző kör után kizárt 13 levél helyettesítésére szolgálnak. A dokumentumok a levelek mellett verseket és dalokat is tartalmaznak.

LETTERS ADDED TO THE HTR CORPUS					
Author	**Epoch**	**Document ID**	**Year**	**Readability**	**HTR Word-count**
Jacob	2. 1800	Ms 237 (Song)	1800	low	343
	4. 1814-1863	Ms 239 (Diary entry)	1815	high	1218
		Br 2231	1829	high	699
		Br 2230	1839	high	245
		Br 2232	1841	high	246
		Br 2235	1850	medium	316
		Br 2233	1860	high	107
		Ms 242 (Dictionary entry draft)	n.d.	high	485
Wilhelm	2. 1802-1805	Ms 245 (poem)	1802	medium	177
	3. 1831-1843	Br 2579	1833	medium	726
	4. 1846-1859	Br 2580	1854	medium	1226

4.1.2.2. A második HTR-modell * A második HTR-körben így 83 dokumentumot vizsgáltunk, amelyek összesen 28963 szót (ebből 10250 Wilhelmé, 18686 Jacobé) és 128 oldalt (ebből 44 Wilhelmé és 84 Jacobé) tartalmaztak. Ebben a körben Jacob és Wilhelm írásait, az „oszd meg és uralkodj” elvével, egymástól függetlenül tápláltuk be a különböző tanító- és tesztkorpuszokba, ami így 8 különálló esetet eredményezett (fejenként és korszakonként egyet) – a szándékunk ezzel annak a felderítése volt, hogy vajon kevesebb, de jobban kontrollált adattal stabilabb modell hozható-e létre. Biztató eredmények születtek; az átlagos karakterhiba-arány az egyes esetekben kevesebb volt, mint 10%.

Alább Jacob Br 2238-as jegyzékszámú, 1840-re datálható levelének HTR-rel átírt részlete látható, melyet Jacob kézírásának 1814–1863 közti periódusán tanított modellel hoztunk létre.

A levél eredeti szövege:

> [...] handlangen scheint es gar nicht. Lamprechts tochter Syon verdient sicherlich eine Ausgabe, und in ermangelung bequemerer verleger steht die quedlinburger nationalbibliothek dafür offen. Rasse gewährt auch andständigehonorare. Mich hochachtungsvoll empfehlend Jac. Grimm

A HTR-átírás [a hibák aláhúzással jelölve]:

> [...] handlangen scheint es gar nicht. Lamprechts tochter von verdient sicherlich eine ausgabe und in ermangelung bgineneber verleger steht die quedlinburger natüoalbiblittchke der für offen. Rasse gewährt auch wurdendigehonorare mich hochachtungsvoll empfehlend Ihr. Grimm

4.1.3. A kritikai kiadás OCR-verziója ♦ Rölleke hétkötetes kritikai kiadásából a digitalizációt és az OCR-t követően 7 fájl készült.[32] Az alábbi szövegrészlet a zajos OCR-eredményre hoz példát (Jacob Grimm fent idézett levele, Br 2238-as jegyzékszámmal):

> Handlangen scheint es gar nicht.
>
> Lamprechts tochter Syon verdient sicherlich eine ausgabe, und in er-manglung bequemerer Verleger steht die Quedlinburger nationalbibliothek dafür offen. Basse gewaährt auch anständige honorare. Mich hochachtungsvoll empfehlend
>
> Jac. Grimm.

Ahogy látható, az OCR az *er-manglung* szóban megőrizte a kiadás által használt kötőjelet, a *Verleger* szót pedig nagy kezdőbetűvel írta, míg a nyomtatott kiadásban kis kezdőbetűvel szerepel. Ezektől a hibáktól eltekintve az *5.* és *6. táblázat* az OCR nagy fokú pontosságát mutatja: a levelek esetében a helyesen felismert szavak mediánja 91% fölötti (a helyes karakterfelismerés pedig 98%-os).

5. táblázat. A gyűjtemény 72 levelének átlagos tisztasága. A félkövérrel szedett számok a levelek eloszlásának mediánját mutatják. Az átlagok standard hibái a gyűjteményen belül elhanyagolhatók.

MEAN COLLECTION CLEANLINESS		
	Clean words in %	**Clean characters in %**
OCR	88.25	97.79
HTR	80.85	94.41
LETTER CLEANLINESS (THREE QUARTILES)		
OCR	86.80 **91.69** 94.06	97.95 **98.70** 99.18
HTR	79.28 **84.29** 88.39	94.09 **95.89** 97.44

[32] A digitalizációt és az OCR-t a Göttinger Digitisation Centre *Abbyy Fine Reader*rel végezte. Lásd bővebben, hozzáférés: 2021.07.30, `https://www.sub.uni-goettingen.de/en/copying-digitising/goettingen-digitisation-centre/`.

 DOI 10.31400/dh-hun.2021.5.3144

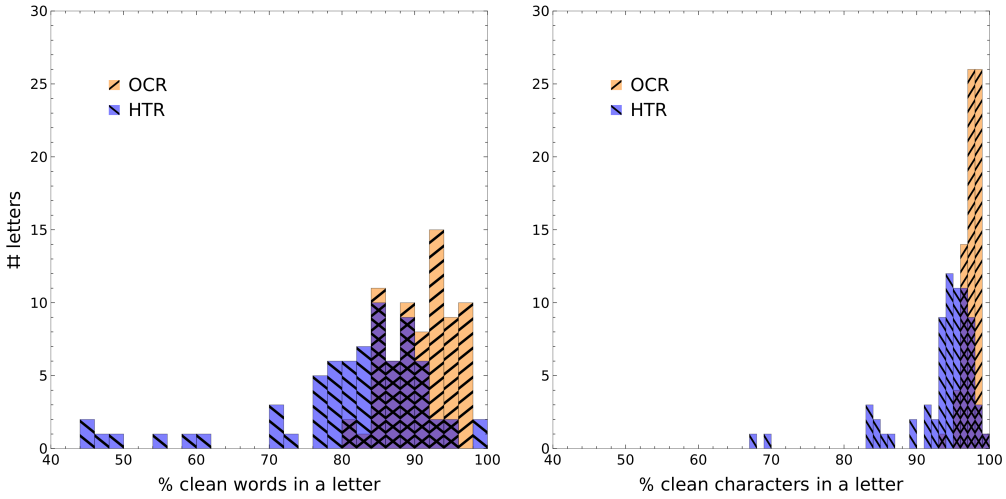
6. táblázat. A gyűjtemény 72 levelének átlagos tisztasága szerzők szerint. Wilhelm levelei következetesen magasabb értéket vesznek fel, habár kevesebb van belőlük. A félkövérrel szedett számok a levelek eloszlásának mediánját mutatják. A standard hibák: <0.0026% (szavak) és <0.00016% (karakterek).

MEAN COLLECTION CLEANLINESS				
	Clean words in %		**Clean characters in %**	
	Jacob	Wilhelm	Jacob	Wilhelm
OCR	87.10	91.12	97.60	98.26
HTR	79.44	84.21	94.24	94.81
LETTER CLEANLINESS (THREE QUARTILES)				
OCR	86.65 **91.69** 93.87	87.51 **91.98** 94.24	98.29 **98.86** 99.17	97.49 **98.43** 99.19
HTR	76.93 **81.93** 85.68	83.61 **87.30** 90.41	94.00 **95.50** 96.96	95.22 **96.77** 98.39

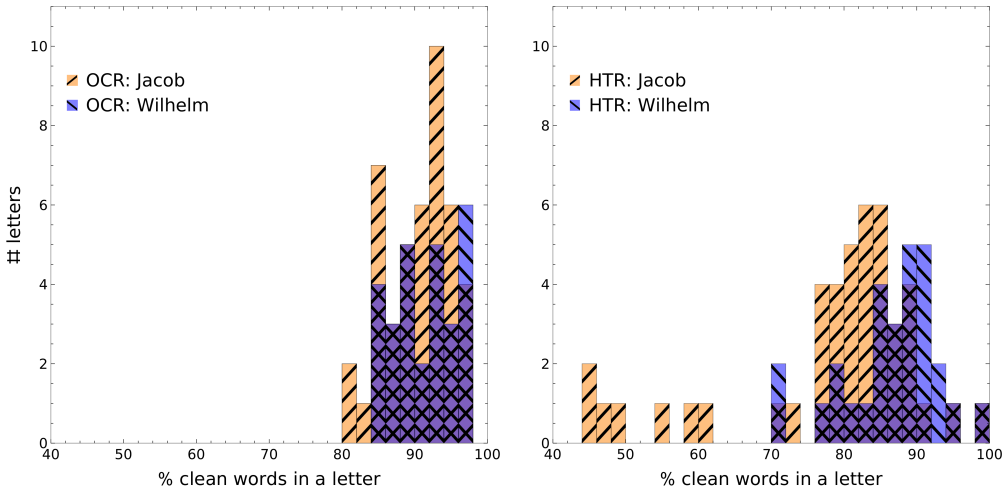
4.1.4. A HTR és OCR által feldolgozott adatok tisztaságának kiértékelése ♦ A következő lépésben a HTR és az OCR segítségével, valamint manuális módon feldolgozott (MAN) levelek tisztaságának összevetését végeztük el. E formátumok mindegyikében 72 darab levél érhető el. A manuális átírást vettük etalonkorpusznak és a többi verzió tisztaságát ehhez mértük (vö. *5. táblázat*). Fontos még megjegyezni, hogy míg a HTR esetében a különbségek csak a felismerési hibákból adódnak, addig az OCR esetében a felismerési hibák és Rölleke esetleges szerkesztői beavatkozásai egyaránt számításba jöhetnek.

Attól függően, hogy mely stilometriai vizsgálatot végezzük, a tisztaság a rosszul felismert szavak (ha az osztályozás szavakon, szó n-gramokon, vagy lemmákon alapul) vagy a rosszul felismert karakterek (amennyiben karakter n-gramokat használunk) százalékos arányán áll, hiszen minden ilyen hiba módosítja a szó/karakter/n-gram gyakoriságát, következésképpen megváltoztathatja a szövegek között mért távolságokat (lásd *4. ábra*).[33] Ezen túl a szerzők eltérő hibaaránya szintén nehézséget jelent a módszerek értékelésekor (lásd *6. táblázat* és *5. ábra*).

[33] Vö. 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.



4. ábra. A hisztogramok azt mutatják, hogy hány levélhez tartozik egy adott százalékban helyesen felismert szó/karakter. Feltűnő, hogy a HTR meglehetősen instabil eredményeket produkál (a bal oldali kiugró értékek). A láthatóság érdekében két hisztogram közötti átfedést a színek keverésével és keresztmintázat hozzáadásával jelezzük.



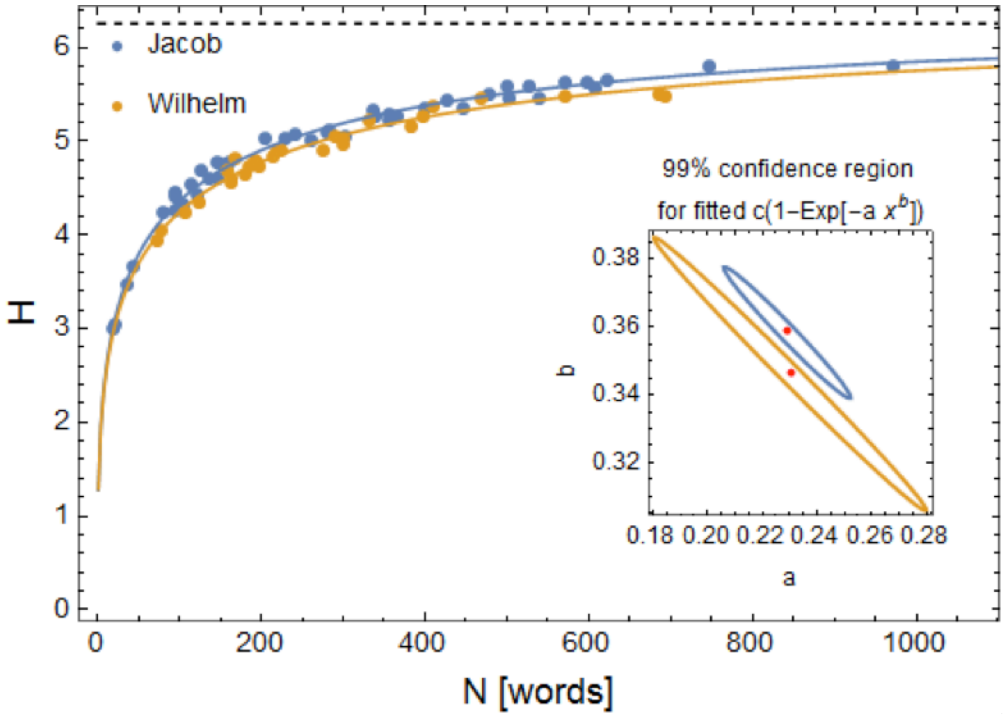
5. ábra. A Jacob és Wilhelm leveleinek hisztogramjai. A jobb oldali panel mutatja, hogy a HTR számára Jacob kézírása okozott nagyobb problémát.

Kezdésként talán nem volna szerencsés a digitalizációs eljárások hatását a szerzőazonosítás kapcsán felmerülő összes jellemzőn (szó és karakter n-gramok) egyszerre vizsgálnunk. Éppen ezért elsőként inkább a felismerési hibáknak a lexikális gazdagságra gyakorolt hatásáról számolunk be, és csak később fordulunk a szerzőazonosítás felé (a 4.2. részben). A lexikális gazdagság nem feltétlen biztosítja a jó szerzőazonosítást,[34] habár határozott stilisztikai jelentéssel bír és a zajos adatokhoz kapcsolódó problémákat is jól illusztrálhatja. A típusok számát megadó képletek közül *(richness*

[34] David L. Hoover, „Another Perspective on Vocabulary Richness," *Computers and the Humanities* 37, 2. sz. (2003): 151–178, https://doi.org/10.1023/A:1022673822140.

 DOI 10.31400/dh-hun.2021.5.3144

score)[35] kettőt alkalmaztunk: Shannon entrópiáját $H = -\sum_{t=1}^{T} p_t \log p_t$ és a Simpson-indexet $D = \sum_{t=1}^{T} p_t^2$ (más néven fordított részvételi arány [Inverse Participation Ratio – IPR]). Mindkettő a diverzitásindexek egy alesete, ahol T a típusok számát jelöli, míg p_t a t típus megjelenésének valószínűségét (azaz t összes előfordulása osztva a szöveg szavainak számával). Néhány kritika,[36] valamint a képletek nagyon erős (bár nem lineáris) korrelációja ellenére ezek a legegyszerűbbek, a legkevésbé önkényesek és elméleti szempontból is a legjobban megérthetők. Ha N a szövegben lévő tokenek száma, az IPR 1/N és 1 között mozog (maximális gazdagság és zéró gazdagság), és főként a legjellemzőbb értékekre (azaz a leggyakoribb szavakra) összpontosít, és így gyorsan stabilizálódik az N szöveghosszal. Az entrópia 0-tól logN-ig terjed (zéró és maximális gazdagság), és a szóeloszlás görbéjének farkára összpontosít (azaz a legritkább szavakra, mint a *hapax legomenon*), amelyek lassabban stabilizálódnak, és az apró változásokra is érzékenyebbek a szöveg előrehaladtával *(6. ábra)*.



6. ábra. Jacob és Wilhelm átírt leveleinek entrópiája (pontok) azt mutatja, hogy Jacob nagyobb lexikai változatossággal rendelkezik. Az eredmény statisztikailag szignifikáns (0,99-es megbízhatósági szinten). A fekete szaggatott vonal a két görbére illesztett közös c aszimptotát jelöli.

[35] Fiona J. Tweedie and R. Harald Baayen, „How Variable May a Constant be? Measures of Lexical Richness in Perspective," *Computers and the Humanities* 32, 5. sz. (1998): 323–352, https://doi.org/10.1023/A:1001749303137; Gejza Wimmer and Gabriel Altmann, „Review Article: On Vocabulary Richness," *Journal of Quantitative Linguistics* 6, 1. sz. (1999): 1–9, https://doi.org/10.1076/jqul.6.1.1.4148.

[36] D. I. Holmes, „The Analysis of Literary Style – A Review," *Journal of the Royal Statistical Society. Series A (General)* 148, 4. sz. (1985): 328–341, 328, https://doi.org/10.2307/2981893; Philippe Thoiron, „Diversity Index and Entropy as Measures of Lexical Richness," *Computers and the Humanities* 20, 3. sz. (1986): 197–202, https://doi.org/10.1007/BF02404461.

Ennek kapcsán megállapítható, hogy a HTR hibáinak köszönhetően a levelenkénti szógazdagság statisztikailag jelentősen csökken az átiratokban (a T-tesztek alapján: $p = 1.04 \times 10^{-7}$ [entrópia] és $p = 8.03 \times 10^{-6}$ [IPR]). A rövidebb levelekben ennek az lehet az oka, hogy a HTR szavakat hagy ki, vagy olvaszt egybe, ami alacsonyabb N-értéket, és – entrópia esetében – alacsonyabb logN-értéket eredményez. Egyébiránt nincs statisztikai korreláció a HTR- és OCR-feldolgozások szöveggazdagsága és -tisztasága közt. Mindazonáltal az eredményeket mérlegelve az OCR járhatóbb megoldásnak tűnik a stilometriai vizsgálatok esetében.

Annak érdekében, hogy Jacob és Wilhelm leveleinek lexikai gazdagsága közti különbséget részletesebben feltárjuk, vissza kellene térni ahhoz a problémakörhöz is, miszerint az IRP és az entrópia a szöveg hosszától is függ. Ezt egy exponenciális függvénnyel tudjuk modellezni, mely alulról közelít egy állandóhoz (az adatokon alapuló vizualizációt mutatja a *6. ábra*). Ezután lehetne megvizsgálni, hogy az illesztett görbék paraméterei között van-e számottevő különbség (ez szintén a *6. ábrá*n látható).

4.2. Szerzőazonosítás

4.2.1. Alapok és beállítások ♦ Ebben a részben a Grimm testvérek egyéni stílusának azonosításról írunk a levelezés már fent tárgyalt zajos digitalizálásának tükrében. A szövegek szerzőségét gépi tanulással végzett kategorizációs és osztályozási műveletekkel igyekeztük meghatározni.[37] Ennek során egy standard bináris osztályozási gyakorlathoz, a Support Vector Machine-hoz (SVM) folyamodtunk, lineáris kernellel és a jól ismert *scikit-learn* könyvtár alapbeállításaival.[38] Egyes tanulmányok szerint az SVM erős alapot biztosít a szerzőazonosításhoz, még kiemelkedően szegényes bemeneti értékek mellett is.[39] Tekintve, hogy adathalmazunk kicsi volt, a visszatartott szerzőkön alapuló keresztellenőrzési eljárást alkalmaztunk (*leave-one-out,* LOO), amelyet minden levéllel elvégeztünk. Azaz minden esetben egyetlen levelet tartottunk vissza tesztpéldányként, míg az osztályozó algoritmust a fennmaradó elemekkel tanítottuk be. Ezt követően rögzítettük a betanított modell előrejelzését a visszatartott minta szerzőségére vonatkozóan. Az egyes modellek teljesítményét a pontosság *(accuracy),* valamint az F1-érték bevett mérőszámai segítségével írjuk le. A kísérlettel kapcsolatban érdemes megemlíteni, hogy ez a felállás egy viszonylag kevés kihívást rejtő szerzőségi problémának tekinthető, hiszen a szerzők száma nagyon korlátozott,[40] és az

[37] Fabrizio Sebastiani, „Machine Learning in Automated Text Categorization," *ACM Computing Surveys* 34, 1. sz. (2002): 1–47, https://doi.org/10.1145/505282.505283; Moshe Koppel, Jonathan Schler and Shlomo Argamon, „Computational Methods in Authorship Attribution," *Journal of the American Society for Information Science and Technology* 60, 1. sz. (2009): 9–26, https://doi.org/10.1002/asi.20961; Efstathios Stamatatos, „A Survey of Modern Authorship Attribution Methods," *Journal of the American Society for Information Science and Technology* 60, 3. sz. (2009): 538–556, https://doi.org/10.1002/asi.20961.

[38] Fabian Pedregosa et al., „Scikit-learn: Machine Learning in Python," *arXiv:1201.0490 [cs]*, 2018. június 5., http://arxiv.org/abs/1201.0490.

[39] Stamatatos, „A Survey."

[40] Vö. Kim Luyckx and Walter Daelemans, „The Effect of Author Set Size and Data Size in Authorship Attribution," *Literary and Linguistic Computing* 26, 1. sz. (2011): 35–55, https://doi.org/10.1093/llc/fqq013. Eder, „Does Size Matter?"

 DOI 10.31400/dh-hun.2021.5.3144

érintett szövegek műfaja is viszonylag állandó.[41] Habár az adathalmaz mérete a gépi tanulás szempontjából kicsinek tekinthető, jól reprezentálja a szerzőségre vonatkozó tanulmányok nagy részét, ahol gyakoriak a rövid szövegekből álló korpuszok, és amelyek ezért szintén a LOO-eljárást részesítik előnyben. Az egyes levelek hossza már önmagában kihívásnak tekinthető, mivel a legtöbb levél lényegesen kevesebb szót tartalmaz annál, mint amit a korábban tárgyalt minimális terjedelmi küszöbök megkövetelnének.[42]

Mi azonban főként nem erre, hanem a különböző digitalizálási módok (a manuális átírás – MAN, és automatikus HTR- és OCR-eljárások) szerzőazonosításban nyújtott teljesítményére koncentráltunk. Ezért a kísérletek során nem csupán egymástól függetlenül teszteltük az egyes módszerek szerzőazonosításra gyakorolt hatását, hanem a különböző eljárások között is. Célkitűzésünk ugyanis, hogy közelebb kerüljünk az irányítottságból adódó zavarok *(directionality artifacts)* megértéséhez; hiszen, ha az egyik digitalizációs eljáráson alapuló modellek jól alkalmazhatók más eljárásokkal átírt szövegek esetében is, akkor azok vonzóbbá válhatnak a jövőbeli projektek során. A szerzőség kérdését tárgyaló tanulmányokban használt szövegek jellemzően meglehetősen eltérő eredetűek, és nagyon különböző forrásokból és kiadásokból érkeznek, és/vagy a különböző módon (OCR és HTR) feldolgozott anyagok gyakran keverednek egymással. Az eltérő anyagok közötti irányultsági hatások megismerése lehetővé tenné, hogy a szövegosztályozás kontextusában hasznos ajánlásokat fogalmazzunk meg a jövőbeli adatgyűjtések számára.

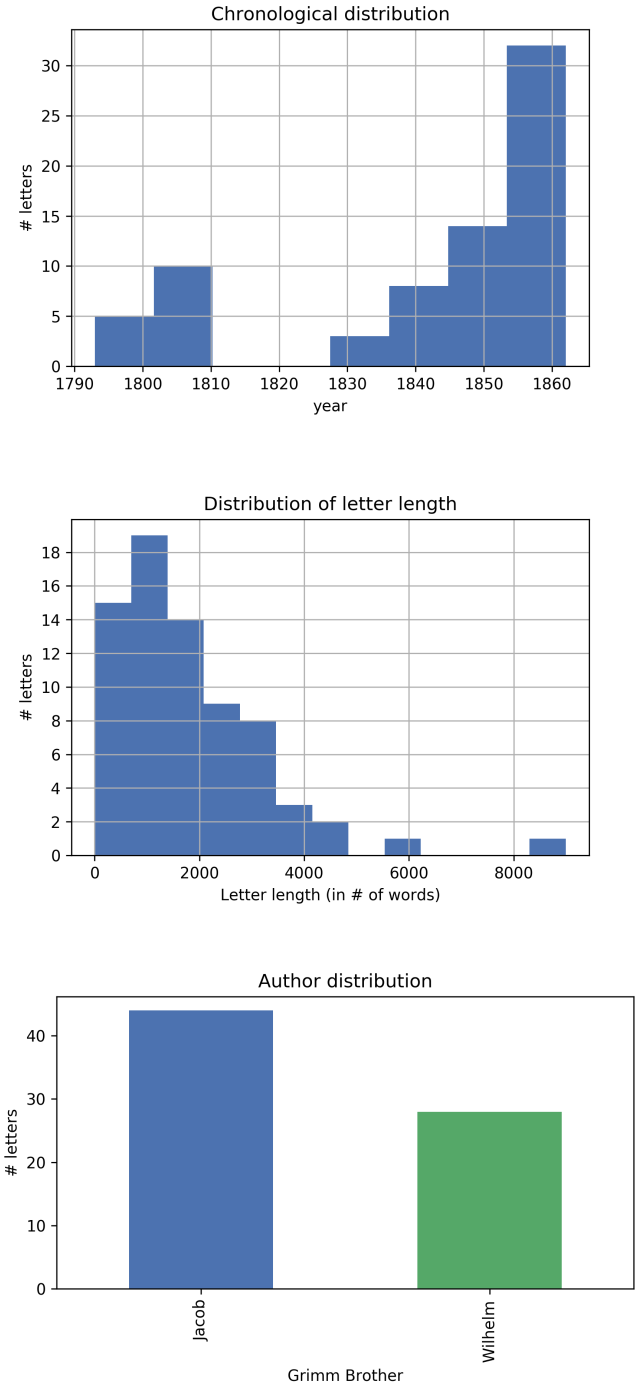
Az előkészítés során a nagybetűket kisbetűvé alakítottuk minden dokumentumban, azért, hogy csökkentsük a szórványjelenségek számát, továbbá eltávolítottuk az üdvözlő formulákból a testvérek neveit, amelyek megzavarhatják a szerzőazonosítást (például kivettük a „W.-t" egy olyan kifejezésből, mint a „Lieber W."). Mindegyik feldolgozás pontosan ugyanazt a 72 levelet tartalmazta, az ezekből készített statisztika a *7. ábrá*n látható. A legtöbb levél a lefedett időszak második feléből származik, bár számos ifjúkorban írt levelet is tartalmaz. A levelek átlagos hossza (szóhosszban) ~1,832, de a hosszúságok jelentős szórást mutatnak (SD = ~1,464). Fontos megjegyezni, hogy Wilhelmet mennyiségileg felülmúlja (n = 28 érték) amúgy is termékenyebbnek mondható testvére (n = 44). A módszert tekintve: a karakter n-gramok vizsgálata nemcsak az egyik legkorszerűbb szövegelemző stratégia a szerzőséggel foglalkozó tanulmányokban, hanem lehetővé teszi a modellek számára a finom, a szavak szintje alatti információ rögzítését is.[43] Az adatokat a TF–IDF (kifejezésgyakoriság–fordított dokumentumgyakoriság) szerint súlyozott vektortérben modelláltuk, az 5000 leggyakoribb, és egy dokumentumban legalább kétszer előforduló karakter n-gramok (bigram, trigram és tetragram) alapján. Végezetül az így kapott mátrixra a soronkén-

[41] Vö. Stamatatos, „A Survey."

[42] Eder, „Does Size Matter?"

[43] Kestemont, „Function Words;" Upendra Sapkota, Steven Bethard, Manuel Montes and Thamar Solorio, „Not All Character N-Grams Are Created Equal: A Study in Authorship Attribution," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 93–102 (Denver, Colorado: Association for Computational Linguistics, 2015), https://doi.org/10.3115/v1/N15-1010.

ti és oszloponkénti normalizálást a stilometriában bevett módon alkalmaztuk (L1-normalizálás, illetve *feature scaling*).[44]

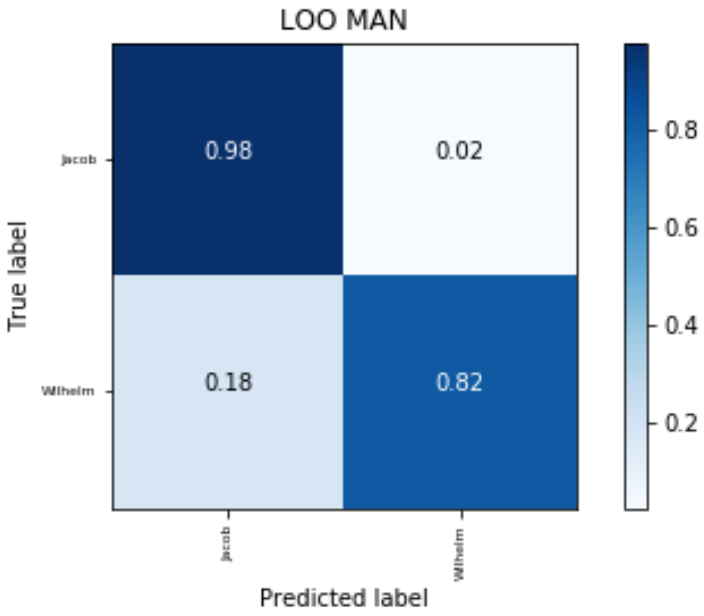


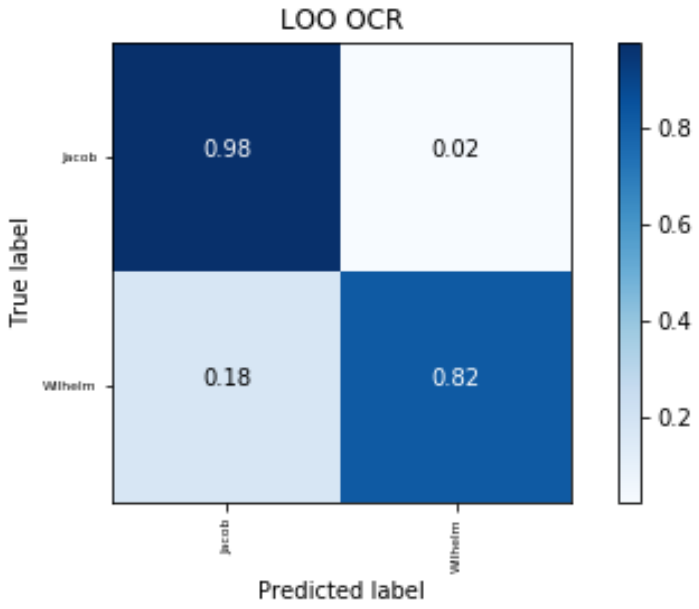
7. **ábra**. Általános információk az ebben a részben tárgyalt korpuszról: a levelek kronológiai, hosszbeli és szerzői címke szerinti megoszlása.

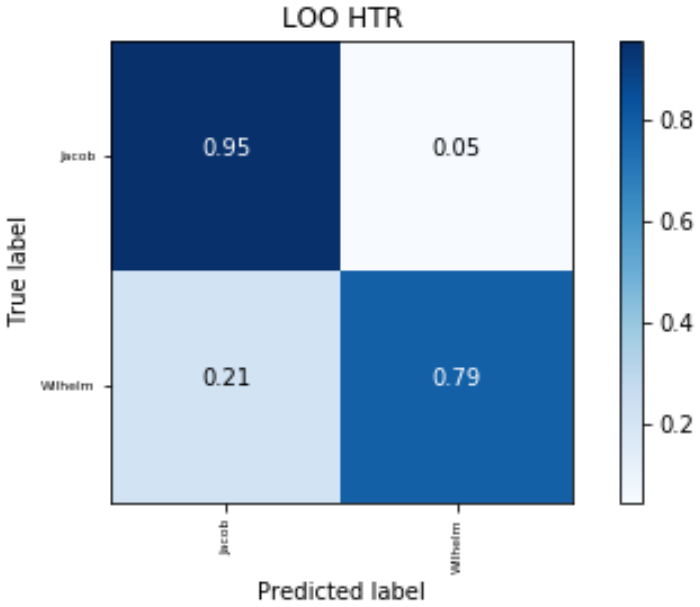
[44] Azaz lényegében Burrows Deltájának normalizációs eljárását követtük. Vö. Burrows, „‘Delta’.”

 DOI 10.31400/dh-hun.2021.5.3144

4.2.2. Azonosítás az egyes eljárásokon belül *(Intra-modality attribution)* ♦ Modellünk általános teljesítményének felméréséhez elsőként a fentebb tárgyalt beállításokkal létrehozott visszatartáson alapuló módszer (LOO) eredményeit személyenként közöljük az adatcsoportokra vonatkozóan (MAN, OCR, HTR). A hibamátrixokat a *8. ábra* és a *7. táblázat* mutatják: ezek részletezik az egyes adatkészletek pontosságát és F1-pontszámát. Általánosságban elmondható, hogy az eredmények viszonylag jók mindkét szempont tekintetében, de semmiképp sem tökéletesek – az SVM egyértelműen Jacob javára hajtja végre az azonosítást, valószínűleg a tanulókorpuszban való relatív túlsúlya miatt. A manuális feldolgozás (MAN) és az OCR eredményei megegyeznek, míg a HTR mindkét értékelési mutatóban rosszabbul teljesít.







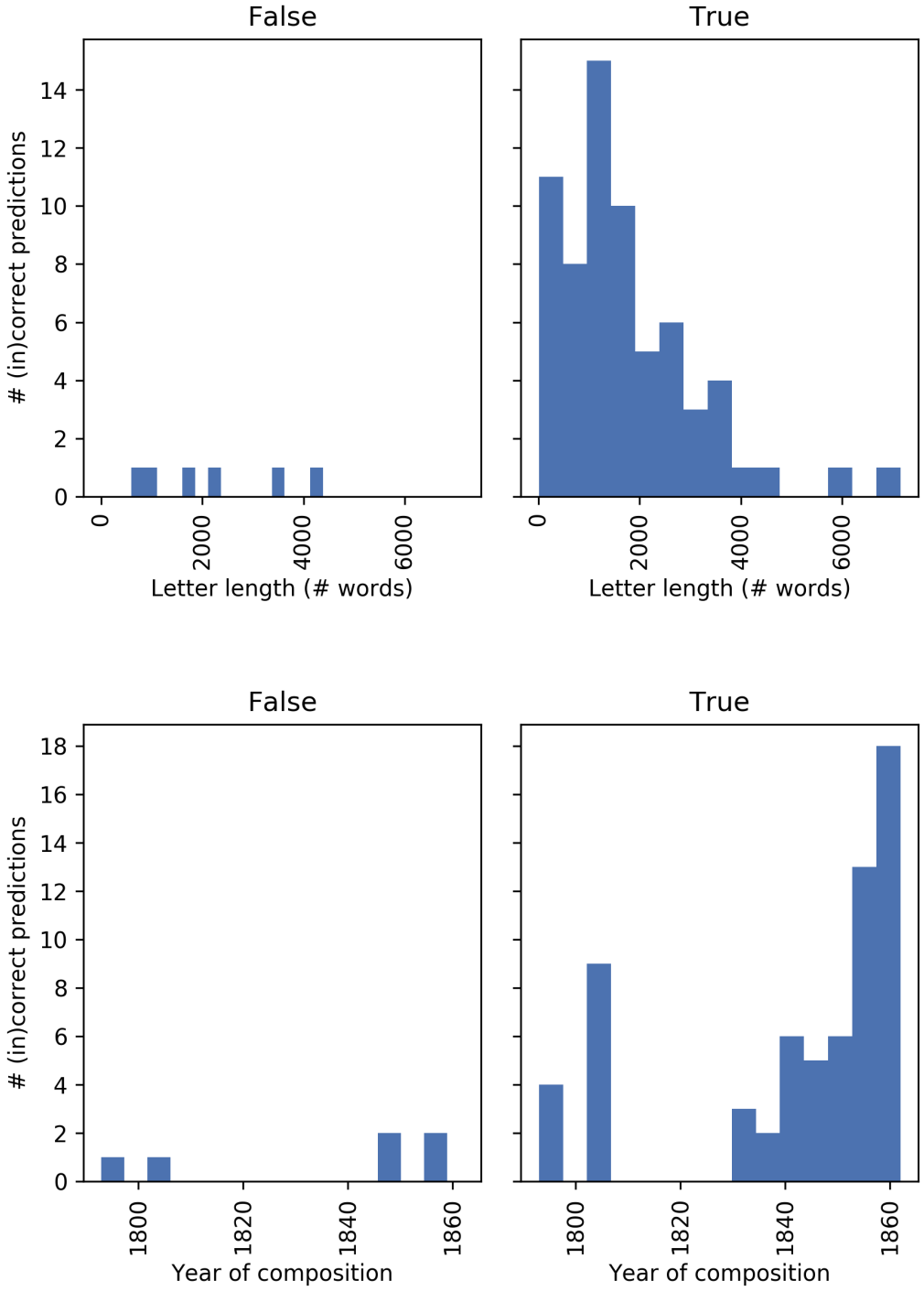
8. ábra. A visszatartáson alapuló módszer *(leave-one-out)* hibamátrixai a három adathalmazban (MAN, OCR, HTR).

7. táblázat. Az egyes eljárásokon belül végzett visszatartásos módszer eredménye a pontosság és az F1-pontszám tekintetében.

	MAN	OCR	HTR
Accuracy	91.66	91.66	88.88
F1-score	88.46	88.46	84.61

A MAN-ra vonatkozó előrejelzéseket – amelyeknek elvileg a legmegbízhatóbbnak kell lenniük – a levelek hosszának és keletkezési idejének függvényében közelebbről megvizsgálva azt látjuk, hogy a téves hozzárendelések – némileg meglepő módon – nem korlátozódnak a rövidebb levelekre (lásd *9. ábra*). Figyelembe véve a levelek általános időrendi megoszlását, azt is láthatjuk, hogy a hibás hozzárendelések az anyag teljes terjedelmében előfordulnak.

 DOI 10.31400/dh-hun.2021.5.3144



9. ábra. A helyes (igaz) és helytelen (hamis) azonosítások eloszlása a kézileg feldolgozott adatsorban: fent a karakterek száma, lent dátum szerint.

Miközben a dátum vagy a hossz nem tűnik döntő tényezőnek, az S1 táblázat *(Függelék, Kiegészítő anyagok)* alapján elmondható, hogy valójában ugyanazoknál a leveleknél rendszeres a rossz szerzőhöz való hozzárendelés. Míg az OCR és a manuális feldolgozás esetében teljesen konzisztensek ezek a hibák, a HTR néhány esetben kiszámíthatatla-

nul viselkedik. Érdemes megjegyezni, hogy az adathalmazok számos rendkívül rövid levelet is tartalmaznak (akár csupán nyolc szóból állót), amelyeken azonban a modellek helyesen végzik el a szerzőazonosítást – valószínűleg a Jacob-levelek irányába való torzítás miatt.

4.2.3. Eljárásokon átívelő hozzárendelés *(Cross-modality attribution)* ♦ A digitális bölcsészet kutatóinak sokszor különböző eredetű, és nem feltétlenül összeegyeztethető módon digitalizált szöveges anyagokat kell egymással kombinálniuk. Bár ez a gyakorlat egyértelműen nem optimális, gyakran elkerülhetetlen. Annak érdekében tehát, hogy felmérhessük a digitalizálási módoknak a szerzőazonosításra gyakorolt együttes hatását, egy eljárásokon átívelő kísérlethez fordultunk. Az adatok összehangolásához mindhárom adatkészletet egyszerre vektorizáltuk a korábbi módon, az 5000 leggyakoribb karakter n-gramot figyelembe véve. A LOO-módszert pedig a következő módon alkalmaztuk:

1. Az adathalmaz minden olyan levelének, amely mind a három adatkészletben megtalálható, meghatároztuk a három verzióját (MAN, OCR, HTR): ezek adták az elemzésekben a visszatartott tételeket.

2. Három különböző osztályozó eljárást tanítottunk be az így fennmaradó egyenként 71 levelére minden alkorpusz esetében.

3. Végül a három osztályozó alapján rendeltünk egy-egy szerzőt mindhárom visszatartott tételhez (összesen 9 szerzőattribúció).

Mindez lehetővé tett egy modelleken átívelő elemzést: hogy megvizsgálhassuk, az egyik digitalizációs eljárásra betanított modell miként teljesít, ha más eljárásokkal digitalizált szövegeken is tesztelik. Ez összesen kilenc betanításteszt kombinációját és összevetését eredményezi (HTR → HTR, HTR → MAN, HTR → OCR, MAN → HTR, MAN → MAN, MAN → OCR, OCR → HTR, OCR → MAN, OCR → OCR), beleértve azt is, amikor ugyanazt a modellt alkalmaztuk a betanításra és a tesztelésre is (azt azonban érdemes észben tartani, hogy az utóbbi eredmények eltérhetnek az előző részben bemutatottaktól a különböző vektorizációs megközelítések miatt). Ezen túlmenően fontos, hogy nyomon kövessük az irányultságot is: az „A" eljárással létrehozott korpuszon betanított modell teljesítménye a „B" eljárással digitalizált szövegre nem feltétlenül egyezik meg annak fordítottjával. Döntő fontosságú, hogy ez lehetővé teszi számunkra, hogy kezdeti ajánlásokat adhassunk arra vonatkozóan, mely modellt részesítsük előnyben a betanítás és a tesztelés során (ezek nem feltétlenül egyeznek meg).

A *10. ábra* négyzetes mátrixában a téves osztályozás kimeneteit mutatjuk be. Három dolgot figyelhetünk meg:

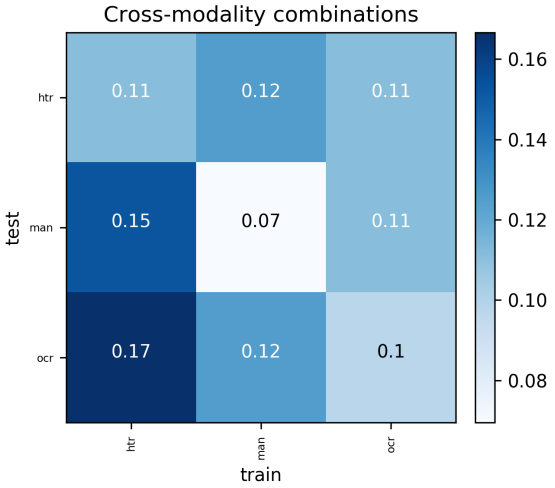
1. Az optimális kombinációt akkor érhetjük el, ha egy rendszert manuálisan átírt adatokon (MAN) tanítunk be és tesztelünk (ennek a hibaaránya: 0,07).

2. Szembeötlő, hogy a HTR-rel létrehozott szövegeken tanított modellek nem teljesítenek jól a háromfajta teszt egyikén sem (különösképpen az OCR-rel kombinálva) (hibaarány: 0,17).

 DOI 10.31400/dh-hun.2021.5.3144

> 3. Úgy tűnik, hogy az OCR viszonylag jól teljesít tanító modellként; sőt elvárásainkkal ellentétben az OCR-rel létrehozott szövegeken tanított és HTR-el létrehozott szövegen tesztelt modell még a manuális átíráson tanított modelleknél is jobb értékeket eredményezett.

Ezeket a megfigyeléseket szemlélteti az egyes levelekhez készített S2 táblázat *(Függelék: Kiegészítő anyagok)*. Ismét azt látjuk, hogy ugyanazoknak a leveleknek a szerzőségét azonosítják tévesen a modelleken átívelő különböző beállítások. Mindazonáltal ez a táblázat azt is mutatja, hogy ezen a szinten jellemzően a HTR-t magába foglaló modellek vezetnek hibás hozzárendeléshez.



10. ábra. Az eljárásokon átívelő hozzárendelés eredményei: a hibás klasszifikáció mátrixa.

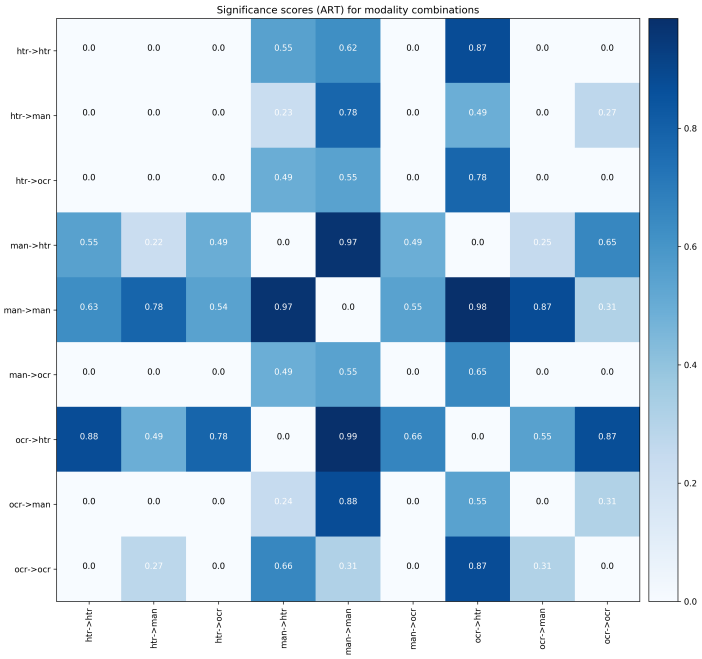
A kombinációk közötti eltérések számos értékes mintát mutattak, de bizonyos esetekben a különbségek felettébb aprók. Ezért fordultunk szignifikanciavizsgálatokhoz, hogy felmérhessük, a modellkombinációknál megfigyelt eltérések relevánsnak tekinthetők-e statisztikai szempontból. A különböző osztályozók eredményeinek szignifikanciavizsgálata vitatott téma a gépi tanulásban, különösen az olyan kis adathalmazok esetében, mint az itt vizsgáltak, ahol az osztálycímkék valódi eloszlása nem ismert vagy nem becsülhető meg megfelelően. Ezért a „közelítő randomizációs tesztelés" *(approximate randomization testing)* néven ismert megközelítést választottuk.[45] Ez a nem paraméteres teszt gyakori a számítógépes szerzőazonosításban[46] – például két olyan azonosítás eredményének összehasonlításakor, ahol nem lehet előzetesen felbecsülni a (potenciálisan rendkívül összetett) eloszlásokat. A teszt olyan pontszámot ad, amelynek segítségével meghatározhatjuk, hogy két bináris osztályozás eredménye

[45] Eric W. Noreen, *Computer-Intensive Methods for Testing Hypotheses: An Introduction* (New York: Wiley, 1989).

[46] Efstathios Stamatatos et al., „Overview of the Author Identification Task at PAN 2014," in Linda Cappellato, Nicola Ferro, Martin Halvey and Wessel Kraaij, *Working Notes for CLEF 2014 Conference,* 877–897 (Sheffield, 2014).

statisztikailag szignifikáns-e az F1-pontszám tekintetében. Ha ezek az értékek nem teszik lehetővé a nullhipotézis (H0) elutasítását, akkor az osztályozók *nem* szolgálnak szignifikánsan eltérő eredményekkel.

A pontszámokat táblázatos formában közöljük *(11. ábra)*. Az ábra nagyrészt megerősíti a korábbi megfigyeléseket: a HTR-rel feldolgozott szövegeken tanított modellek rosszabbul teljesítenek, és kimeneteik sem mutatnak jelentős eltérést egymástól. Amint azt a sötétebb színnel szedett sorok és oszlopok mutatják, a HTR-t tartalmazó kombinációk általában határozottan rosszabb eredményeket hoznak, mint azok melyek (kizárólagosan csak) manuális átíráson vagy OCR-en alapulnak. Emellett a világosabb cellák igazolják azt a feltevést is, hogy sok esetben az OCR-t tartalmazó kombinációk nem térnek el jelentősen a manuális feldolgozástól. Ez arra enged következtetni, hogy az OCR-digitalizálás megbízható helyettesítője lehet a jóval fáradságosabb manuális átírásnak – legalábbis ami a szerzőazonosítást illeti.

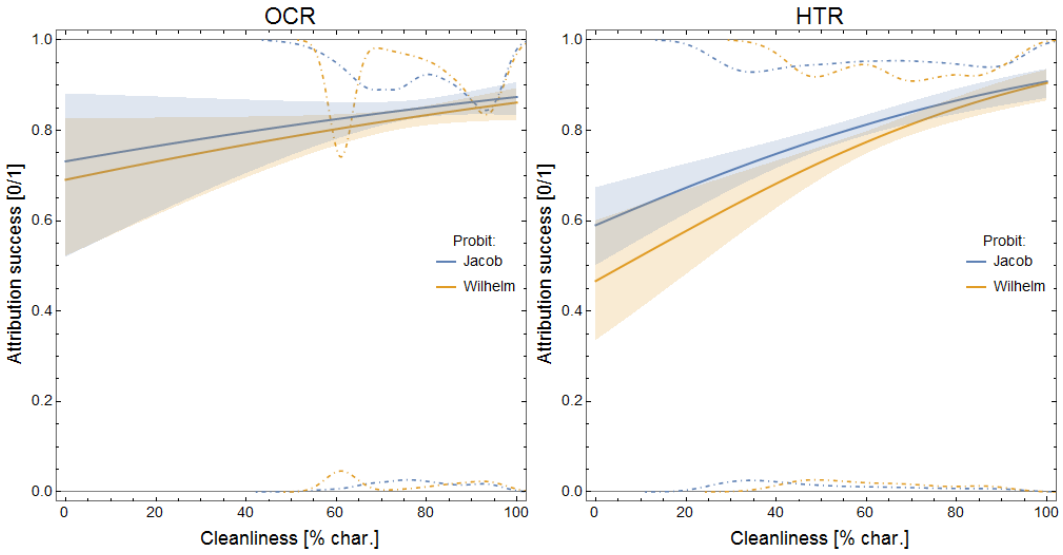


11. ábra. A különböző digitalizációs eljárások kombinációinak a szerzőazonosítás hatékonyságára kapott pontszámok táblázatos ábrázolása. A magasabb pontszámok azt jelzik, hogy a két rendszer jelentősen eltérő eredményeket produkál a többi kombinációhoz képest.

4.2.4. Bináris azonosítás kontra szövegtisztaság ♦ Az itt közölt kísérlet arra irányul, hogy feltárjuk, létezik-e összefüggés a Grimm-levelezés bináris szerzőazonosításának (tehát a szerzőség Jacobnak vagy Wilhelmnek tulajdonítása) sikere és az automatizált szövegfelismerés (OCR vagy HTR) között. Az alábbiakban kitérünk arra is, hogy mi történik, ha az azonosítás folyamata eltérő modelleket használ: tanítókorpuszként manuális átírást, tesztként pedig OCR-t vagy a HTR-t.

 DOI 10.31400/dh-hun.2021.5.3144

Ez esetben ahelyett, hogy a meglévő levelek tisztasága (lásd 4.1.4. rész) és az osztályozási eredmények közti összefüggést keresnénk, finomabb módon jártunk el, és véletlenszerű mintavételezéssel fejlesztettük a korpusz statisztikai megbízhatóságát. A tesztkorpuszt tehát a manuálisan átírt levelekből, véletlenszerű mintavételezéssel állítottuk össze, amely így 72 szöveget tartalmazott (eloszlásuk: 44 levél Jacobtól, 28 levél Wilhelmtől). Mindegyik legalább 1500 karakterből (nagyjából 250 szóból) áll. Ez lehetővé tette számunkra, hogy normalizáljuk a mintákat, miközben megőriztük az eredeti adatkészlet néhány reális tulajdonságát is. Ezt követően egy újabb 1500 karakter hosszú véletlen mintát hoztunk létre az automatikusan átírt (OCR, HTR) szövegek soraiból úgy, hogy nyomon tudtuk követni azok tisztaságát (azaz a helyesen átírt karakterek arányát) és osztályozni is tudtuk azokat (Burrows-féle Delta-távolság alkalmazásával a száz leggyakoribb szó alapján). Az eljárást (tanító és tesztkorpusz kijelölése, az osztályozás lefuttatása) szerzőkként és átírási módszerenként is addig ismételtük, míg összesen 3800 véletlenszerű mintát kaptunk, amelyeket tisztaságuk (0-tól 100%-ig terjedő skálán) és osztályozási eredményük (0 vagy 1 – helyes vagy helytelen szerzőazonosítás) jellemez. Ezeket az adatokat aztán a probitregresszió segítségével lehet elemezni, amint azt a *12. ábra* szemlélteti.



12. ábra. Véletlenszerű mintavételezésű szövegekre illesztett probitmodellek (folytonos vonalak); az árnyékolt területek a 0,95-ös megbízhatósági intervallumot jelölik. A pontozott vonalak grafikonjai a minták helyes (fent) és téves (lent) azonosítására vonatkoznak. Fontos, hogy a felső grafikon fejjel lefelé értelmezendő. Például az OCR esetében a Wilhelm-minták többsége 60%-os vagy 95%-os tisztaságú (a felső grafikon csúcsértékei), és 60%-os tisztaságnál ötször annyi helyes, mint helytelen azonosítás van (a felső és az alsó grafikon csúcsértékeinek aránya).

Az eredmény mindig a tanítókorpusz méretétől függ (valamint a testvérek részhalmazának relatív gyakoriságától). Ezért ellenőriztük, hogy a *12. ábrán* látható modellek megfelelnek-e az azonos méretű részhalmazok esetében is. Néhány (száz) minta még megbízhatatlan eredményre vezetett, ugyanakkor több ezer elegendő volt a konver-

gens, stabil eredmény eléréséhez. A mérések továbbá azt mutatják, hogy az OCR-eljárás tisztasága és a szerzőazonosítás sikere között csekély mértékű (0,95 – de nem 0,99 megbízhatósági szinten), míg a HTR-rel feldogozott írások tekintetében jelentős az összefüggés (0,99 megbízhatósági szint): ez esetben minél tisztábbak a szövegek, annál valószínűbb a helyes felismerés. Figyelemre méltó, hogy a körülbelül 20% feletti tisztaság már elegendő a HTR-rel feldolgozott szövegek szerzőjének a véletlennél nagyobb valószínűségű helyes azonosítására (OCR esetén a helyes felismerés valószínűsége mindig nagyobb a véletlennél).

5. Következtetések

A tanulmány a Jacob és Wilhelm Grimm leveleinek digitalizációja során keletkező zaj automatikus szerzőazonosításában mért hatásáról ír. Az összes lehetséges digitalizálási „forgatókönyv" teszteléséhez három különböző kimenetet hasonlítottunk össze: 1. az eredeti levelek manuális átírását; 2. a Grimm-levelek 2001-es nyomtatott, kritikai kiadásának OCR-feldolgozását; és 3. az eredeti levelek automatikus átírásához készült HTR-modellt. A manuális átírást etalonkorpuszként használtuk az OCR- és a HTR-feldolgozás tisztaságának értékeléséhez. A várakozásoknak megfelelően a HTR hibaaránya magasabb volt az OCR-nél a kézírás esetlegessége miatt (szemben a nyomtatás egységességével). Mindezeken túl a kísérletek azt mutatták, hogy a HTR-feldolgozásra messze nem tökéletes adatkészlet ellenére is átlagosan 6%-nál kevesebb karakterhibaaránnyal dolgoztak a Grimm testvérek számára létrehozott modellek (azaz minden tizenhetedik karaktert olvasták be hibásan).

Az ilyen hibaarány már önmagában elég magas ahhoz, hogy jelentősen csökkentse a HTR-rel feldogozott levelek szókincsbeli gazdagságát. Mivel ez a testvérek szerzőségét tekintve megkülönböztető tényező, megvizsgáltuk a három különböző digitalizálási kimenet (manuális átírás [MAN], zajos OCR és zajos HTR) hatását a szerzőazonosításra. Ennek eredményeként megállapítottuk, hogy (legalább a szerzőazonosítás során) az OCR-rel végzett digitalizálás megbízható alternatívaként szolgál a gondosabb manuális átiratok mellett is.

Érdekes, hogy a hozzárendelés akkor is működőképesnek tűnik, ha a tanító és a tesztkorpuszokat különböző módon digitalizált szövegekből építik fel. Ami a HTR-t illeti, a kutatásaink arra vezettek, hogy ugyan az automatikus átírás jelentősen növeli a szöveg téves osztályozásának kockázatát az OCR-hez képest, már körülbelül 20% feletti szövegtisztaság is elegendő ahhoz, hogy a véletlennél nagyobb valószínűséggel legyen sikeres a bináris azonosítás (OCR esetén ez a valószínűség mindig nagyobb).

Habár eredményeink még kezdetlegesek, érvként szolgálhatnak abban a diskurzusban, ami a szöveg abszolút tisztaságát nem a szerzőazonosítás elsődleges feltételének teszi meg[47] – mindenesetre a Grimm testvérek által írt levelek kapcsán. Az általunk létrehozott HTR-modell az első olyan modell, amely a Grimm fivérek kézírásának felismerésére jött létre – ami még tovább finomítható több kézzel írott dokumentum (például a jelenleg Berlinben található szakmai levelezés) betáplálásával. Kitekintésként: egy következő kutatási téma lehetne a közös szerzőség vizsgálata a *Gyermek- és*

[47] Vö. Eder, „Does Size Matter?"

 DOI 10.31400/dh-hun.2021.5.3144

családi mesék (Grimm's Kinder und Hausmärchen) című mű esetében, azt ellenőrzendő (már ha egyáltalán lehetséges), hogy mely mesékben érvényesül markánsabban Jacob, illetve Wilhelm szerzői ujjlenyomata.

Fordította: Kustos Júlia

Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm

This article presents the results of a multidisciplinary project aimed at better understanding the impact of different digitization strategies in computational text analysis. More specifically, it describes an effort to automatically discern the authorship of Jacob and Wilhelm Grimm in a body of uncorrected correspondence processed by HTR (Handwritten Text Recognition) and OCR (Optical Character Recognition), reporting on the effect this noise has on the analyses necessary to computationally identify the different writing style of the two brothers. In summary, our findings show that OCR digitization serves as a reliable proxy for the more painstaking process of manual digitization, at least when it comes to authorship attribution. Our results suggest that attribution is viable even when using training and test sets from different digitization pipelines. With regard to HTR, this research demonstrates that even though automated transcription significantly increases risk of text misclassification when compared to OCR, a cleanliness above 20% is already sufficient to achieve a higher-than-chance probability of correct binary attribution.

Keywords:
stylometry, authorship attribution, german literature, Grimm, digitization, OCR, HTR

Köszönetnyilvánítás

A szerzők köszönettel tartoznak kollégáiknak: Maria Moritznak, Kirill Bulertnek, Marco Büchlernek és volt kollégájuknak Linda Brandtnak a projekthez nyújtott értékes hozzájárulásukért. Ezúton is szeretnék megköszönni a németországi Kasselben a Brüder Grimm-Gesellschaft e.V munkatársainak: Bernhard Lauernek és Rotraut Fischernek a szakértői tanácsokat és a Grimm testvérek kalligráfiájával kapcsolatos tudásuk megosztását. Köszönik dr. Stephan Tulkensnek a 4.2.3. alfejezet kutatásában nyújtott támogatását, valamint dr. Günther Mühlbergernek a Grimm-kézírás HTR-modellezésében nyújtott nélkülözhetetlen hozzájárulását. Végül külön köszönet illeti Gerhard Lauer professzort állandó javaslataiért és támogatásáért.

Függelékek

A cikkhez készült kiegészítő anyag (*S1 táblázat:* Az egyes digitalizációs eljárásokon belül létrehozott modellek eredményeinek az áttekintése; *S2 táblázat:* A eljárásokon átívelő kísérletek eredménye) hozzáférhető az alábbi címen:

> `https://www.frontiersin.org/articles/10.3389/fdigh.2018.00004/full#supplementary-material`, valamint a jelen cikk mellékleteként annak adatlapján: `http://doi.org/10.31400/dh-hun.2021.5.3144`.

Jacob Grimm 1793-as levelének manuális leirata (vö. 3.2.1.2. alfejezet):

> Montag
>
> Steinau den 7 8br 1793.
>
> Lieber Bruder!
>
> Du wirst hierbey dein Kleid erhalten, Wie hatt es dir denn auf der Reise gefallen, mich verlanget es zu wissen, ich erwarte mit der ersten Gelegenheit einen Brief von dir, seit deiner Abwesenheit ist nichts merkwürdiges vorgefallen.
>
> Mein Vater hatte heute einen sehr starken Amtstag gehabt, bis Freitag wird dich unser hofjud Jud Seelig besuchen und mit diesem werde ich dir weitläufliger schreiben, Küβe der lieben Mutter dem Groβvater und jungfer Tante die Hand in meinem Nammen. Du wirst von uns allen gegrüβet, und ich bin dein treuer
>
> Bruder
>
> Jacob Grimm

 DOI 10.31400/dh-hun.2021.5.3144

Artjoms Šeļa 0000-0002-2272-2077

Instytutu Języka Polskiego PAN

artjoms.sela@ijp.pan.pl

Boris Orekhov 0000-0002-9099-0436

Higher School of Economics, Moscow

borekhov@hse.ru

Roman Leibov 0000-0002-5521-2954

Tartu Ülikool

hv.dekanaat@ut.ee

Gyenge műfajok
A költői versmérték és a jelentés közötti kapcsolat modellálása az orosz költészetben*

A dolgozat egy már meglévő, „a versmérték jelentésmezőjeként" ismert költészetelmélet formalizálását kísérli meg, amely elmélet azt állítja, hogy a modern líra különböző metrikai formái bizonyos jelentésbeli asszociációkat halmoznak fel és őriznek meg. Az LDA témamodellező *(topic modelling)* algoritmussal vizsgáltuk az orosz költészet tág korpuszát (1750–1950), hogy ezáltal minden egyes verset egy tématérben, a versmértékeket pedig a témák valószínűségének eloszlása szerint reprezentáljunk. Nem felügyelt osztályozást és kiterjedt mintavételt alkalmazva megmutatjuk, hogy a verselési formákon belül és között erős a forma és a jelentés kapcsolata: ugyanahhoz a versmértékhez tartozó két minta sokszor nagyon is hasonlóként tűnik fel, és ugyanannak a családnak két verselési formája legtöbbször szintén egy klaszterbe kerül. Ez a kapcsolat akkor is kimutatható, ha a korpusz kronológiai szempontból ellenőrzött, és nem következménye a populáció méretének. Amellett érvelünk, hogy hasonló megközelítést nyelvek és költészeti hagyományok szemantikai mezőinek összehasonlításakor is alkalmazni lehet, amelynek révén az irodalomtörténet legalapvetőbb kérdéseire adhatók releváns válaszok.

Kulcsszavak:
költészet, szemantika, versmértékek, témamodellezés, klaszterezés

* Eredeti megjelenés: Artjoms Šeļa, Boris Orekhov and Roman Leibov, „Weak Genres: Modeling Association Between Poetic Meter and Meaning in Russian Poetry," in *CHR 2020: Workshop on Computational Humanities Research*, 2020, `http://ceur-ws.org/Vol-2723/long35.pdf`.

1. Bevezetés

A költői forma és annak jelentése közötti kapcsolat talán triviálisnak tűnhet. Történetileg a metrikai megkülönböztetés vezetett a műfajok és a költői beszéd típusainak elkülönítéséhez, egészen az indoeurópai epika „hosszú" és a lírai vers „rövid" soraiig.[1] Általában nem várunk önelemző meditációt egy limericktől, míg egy szonettől talán számítunk rá. A daktilikus hexameter európai imitációi vagy az elégikus disztichon a modern verselési rendszerekben tematikusan kötődnek klasszikus kori forrásaikhoz. Vajon a versforma és annak szemantikája közötti kapcsolat érvényes a versmértékek „általános használatára" a modern költészeti hagyományban is, ahol a műfaj és a forma közötti normatív kapcsolatok gyorsabban elenyésznek? A válasz, amellyel mindenki egyetért: igen.

A versmértékek képessége, hogy az idő múlásával felhalmozzanak és megőrizzenek különböző szemantikai jellemzőket, „a versmérték szemantikai mezőjeként" is ismert a kvantitatív metrikai tudományok orosz iskolájában.[2] A kezdeti megfigyelések azonban csupán egy-egy költő metrumhasználatán[3] vagy anekdotikus bizonyítékokon alapultak (nevezetesen néhány időben elszórt, trochaikus pentameterben megalkotott költeményre). A korai kutatók mindazonáltal a versmérték-jelentés kapcsolatot organikusnak tekintették, vagyis úgy gondolták, hogy a ritmus néhány belső tulajdonsága formálja a vers jelentését.[4] Mikhail Gasparov több ezer 19. századi költemény szoros olvasására alapozva demonstrálta, hogy ezek a kapcsolatok történeti jellegűek, amelyeket a versmérték helyi hagyományai és a későbbi alkalmazásai határoznak meg, ami egy szétszóródott, de mégis megkülönböztethető szemantikai profil létrejöttét eredményezi.[5]

[1] Mikhail Leonovich Gasparov, *A History of European Versification,* trans. Gerald Stanton Smith and Leofranc Holford-Strevens (Oxford–New York: Clarendon Press–Oxford University Press, 1996), https://doi.org/10.1093/acprof:oso/9780198158790.001.0001; Antoine Meillet, *Les origines indo-européennes des mètres grecs* (Paris: Les Presses universitaires de France, 1923), hozzáférés: 2021.12.07, https://archive.org/details/lesoriginesindoe00meiluoft.

[2] Maxim Iljics Shapir, „'Semanticheskii oreol metra': termin i poniatie," in Maxim Iljics Shapir, *Universum versus: iazyk, stikh, smysl v russkoi poezii XVIII–XX vekov,* Vol. 2., 395–404 (Moszkva: Iazyki slavianskoi kul tury, 2015); Marina Tarlinskaja and Naira Oganesova, „Meter and Meaning: The Semantic Halo of Verse Form in English Romantic Lyrical Poems (Iambic and Trochaic Tetrameter)," *The American Journal of Semiotics* 4, 3–4. sz. (1986): 85–106, https://doi.org/10.5840/ajs198643/422; Mikhail Trunin, „Towards the Concept of Semantic Halo," *Studia Metrica et Poetica* 4, 2. sz. (2017): 41–66, https://doi.org/10.12697/smp.2017.4.2.03.

[3] Grigoriy Vinokur, „Vol'nye iamby Pushkina," in *Pushkin i ego sovremenniki: Materialy i issledovania,* Vol. 38–39, 23–26 (Leningrad: 1930).

[4] Roman Jakobson, „Toward a Description of Mácha's Verse," in Roman Jakobson, *Selected Writings, Vol. 5: On Verse, Its Masters and Explorers,* eds., Stephen Rudy and Martha Taylor, 433–485 (The Hague–Paris–New York: Mouton Publishers, 1979), https://doi.org/10.1515/9783110803068.433; Kiril Taranovskii, „O vzaimootnosheniah stikhotvornogo metra i tematiki," *American Contributions to the Fifth International Congress of Slavists, Sofia, September 1963: Vol. 1: Literary contributions,* 287–332 (The Hague: Mouton and Co., 1963).

[5] Mikhail Gasparov, *Metr i smysl: ob odnom iz mekhanizmov kulturnoi pamiati* (Moscow: Izdatelskii tsentr RGGU, 1999).

 DOI 10.31400/dh-hun.2021.5.3145

Az ilyen megállapítások vonzereje ellenére a szemantikai mező koncepciója a formalizálás hiányában könnyen kritizálható és nehezen védhető elképzelés. Még ha egyes konkrét „mezők" nem is egyszerű mintavételi hiba termékei, magára a mechanizmusra és a metrikus formák közötti kapcsolatok szerkezetére vonatkozó általánosítások továbbra is megfoghatatlanok maradnak. Ugyanakkor néhány korábbi empirikus kísérlet, amely az orosz[6] és a baskír[7] költészetben a versmérték-jelentés viszony megközelítésére irányult, szólisták összehasonlítására támaszkodva, egészen jól körül tudta írni a metrikus formák közötti lexikális különbségeket, amely eredmények számunkra is fontos kiindulási pontot biztosítanak az alábbiakban.

Ez a dolgozat ugyanis a szemantikus mező jelenlétét igyekszik megvizsgálni az orosz költészetben, absztrakt szemantikus jellemzők (témák) alapján, amelyek minden egyedi verset egységes módon képesek leírni. Azáltal, hogy a szövegeket egyetlen modellen belül helyezzük el, rugalmasan tudunk teszteket végezni és osztályozási algoritmusokat használni a tudományos feltételezések kifejtésére és ellenőrzésére. Ennek során a hierarchikus klaszterezés módszerére támaszkodunk, hogy a versmértéken belüli szemantikus hasonlóságok szintjét (hasonlítanak-e a versmértékek önmagukra) és a versmértékek közötti kapcsolatokat (kapcsolódnak-e egymáshoz az egy családhoz tartozó, különböző versmértékek) felmérjük. Az elemzést követően tárgyaljuk, hogy a versmérték szemantikus mezőjének a formalizálása miként járulhat hozzá a metrumnak mint kulturális átadásnak *(cultural transmission)* a megértéséhez, és hogy miként lehetne hasonló megközelítést alkalmazni a témamezők különböző nyelveken és hagyományokon átívelő vizsgálatára.

2. Korpusz

A kutatásban felhasznált adatok az Orosz Nemzeti Korpusz Költészeti algyűjteményéből[8] származnak, amely a 18. századtól a 20. századig terjedő korszakban született szövegeket tartalmaz. Nagyjából tehát lefedi a modern orosz versmérték egész történetét, amely a német időmértékes verselés 1730-as megjelenésével kezdődött. A korpusz már elgondolásában is egyértelműen elfogult a kánon iránt: csak azok a 18–19. századi szövegek kerültek a gyűjteménybe, amelyek elérhetők a 20. századi kritikai kiadásokban.[9] Ezért nem vesz figyelembe sok korábbi, akadémiai kánonon kívüli költészeti teljesítményt, és egyenlőtlenséget teremt a költemények kronológiai eloszlásában is: a szövegeknek több mint 75%-a a 20. századból való. Ráadásul egyáltalán nem egységes a merítés, mivel 1917-tel kezdődően az orosz költészet három, általában véve elszigetelt hagyománnyá válik szét: szovjet, emigrációs és nem hivatalos underground. Mivel nem áll módunkban automatikusan elkülöníteni őket, a korpusz határát 1950-ben állapítjuk meg, ami kizárja a legtöbb underground művet, és megállítja az órát, mielőtt

[6] Alexander Piperski, „Semantic Halo of a Meter: A Keyword-Based Approach," in *Kompiuternaia lingvistka i intellectualnyie tekhnoloigii, Vol. 2: Kompiuternaia lingvistika: lingvisticheskiie issledovaniia,* 342–354 (Moscow: RGGU, 2017).

[7] Boris Orekhov, *Bashkirskii stikh XX veka: Korpusnoe issledovanie* (St. Petersburg: Aletejia, 2019).

[8] Orosz Nemzeti Korpusz (2003), hozzáférés: 2021.12.07, `https://ruscorpora.ru/new/en/index.html`.

[9] Kirill Korchagin, „Poezija XX veka v poeticheskom podkorpuse Natsional'nogo korpusa russkogo iazyka: problema reprezentativnosti," *Trudy instituta im. V. V. Vinogradova* 6 (2015): 235–256.

megkezdődne az észrevehető sodródás a nem klasszikus versmértékek felé. Miután minden felosztási művelet és előzetes feldolgozási lépés (lásd alább, a 3. részben) megtörtént, 47804 szöveg (2275233 szó) maradt a korpuszban.

Jelen dolgozat legfőképpen a szótagszámláló hangsúlyos verselésű *(accentual-syllabic, AS)*[10] – és általában rímes – költészetre összpontosít, ami sokkal tovább fennmaradt az orosz lírában, mint a szabadvers felé forduló nyugati tradíciókban.[11] Az orosz szótagszámláló hangsúlyos verselési rendszerek a hangsúlyok és a soron belüli szótagok számának szigorúbb behatárolásán alapszanak a pusztán hangsúlyos (ahol csak a hangsúlyok száma fontos) vagy a pusztán szótagoló (ahol csak a szótagszám számít) verseléshez képest. Ebben a verselési módban a versmértékek a ritmus visszatérő kisebb egységeire épülnek – verslábakra, amelyek mintákba rendezik a hangsúlyos és nem hangsúlyos szótagokat, általában kettőt vagy hármat (bináris vagy hármas láb). Mivel a metrikai séma a versritmus absztrakciója, és folyamatosan módosul (a várt hangsúlyos pozíciók hangsúlytalanok maradnak vagy fordítva), általában erős versus gyenge pozíciókról beszélünk „hangsúlyos" vagy „hangsúlytalan" helyett. A *B.3. táblázat* összegzést nyújt a klasszikus időmértékes formákról, amelyeket ebben a dolgozatban használunk. Kivételt képeznek az úgynevezett „dolnikok", amelyek a szótagszámra vonatkozó szabályok meglazításával eltávolodnak a szótagszámláló hangsúlyos verseléstől, ám mégsem lehet őket figyelmen kívül hagyni, olyan nagy számban vannak jelen a 20. században.

Annak érdekében, hogy – ha lehetséges – minden verset egyetlen, egyértelmű versképlettel írhassunk le, a korpusz metaadatait használtuk fel, amelyek a versformára vonatkozó annotációkat tartalmazzák. A korpuszannotációt intézményes keretek közt végezték a nyelvészet és a prozódia területén jártas szakemberek felügyeletével, ugyanakkor nem jegyezték a hibaszázalékot, vagy hogy mekkora volt az egyetértés az annotációt végzők körében.[12] Mindazonáltal nagyon magasra becsüljük a munka pontosságát, különösen a klasszikus formák tekintetében, amelyek még minimális képzettséggel is könnyen megkülönböztethetők. Megkértünk három irodalomtudóst, hogy igazoljanak 100 eredeti, a versformára vonatkozó korpuszannotációt: átlagosan 97,7% címkét jelöltek meg „igazként", a köztük lévő egyetértés mértéke pedig 96,6% volt (az alacsony mértékű egyetértés azokban az esetekben volt gyakori, amikor a címkét „hamisnak" tekintették).

[10] Az angol terminológia az angol és az ebből a szempontból hasonló orosz verselést tükrözi, amely a hangsúlyos és hangsúlytalan szótagok szabályos váltakozásán alapul (és ahol a hangsúlyos szótagok pozíciója nem kötött egy szón belül, mint például a magyarban). A klasszikus görög-latin időmértékes verselés rendszere adaptálható a hangsúlyos verselés viszonyaira, azonban ebben az esetben a verslábakat nem a rövid és hosszú, hanem a hangsúlytalan és hangsúlyos szótagok hozzák létre. Ez egyben egy kötött szótagszámú („szótagszámláló") verselést eredményez, hiszen a verslábak a szótagok számát is meghatározzák. A verstani kérdésekben nyújtott segítségért hálával tartozunk Ferencz Győzőnek – *a szerk.*

[11] Mikhail Gronas, *Cognitive Poetics and Cultural Memory: Russian Literary Mnemonics* (New York: Routledge, 2010), https://doi.org/10.4324/9780203842430.

[12] E. Grishina et al, „Poeticheskii korpus v ramkah NKRIA: obschaia struktura i perspektivy ispolzovania," in *Natsionalnii korpus russkogo iazyka: 2006–2008. Novye rezultaty i perspektivy,* 71–113 (St. Petersburg: Nestor-Istoria, 2009).

 DOI 10.31400/dh-hun.2021.5.3145

A metaadatokkal való címkézéskor meglehetősen konzervatívok voltunk, előnyben részesítettük a homogén metrikai lejegyzéseket, és kizártuk a polimetria vagy egyéb heterogén formák összetett eseteinek nagy részét. Szintén egyszerűsítéseket végeztünk a versszak tekintetében, és csak az általánosan elterjedt rímképletekre támaszkodtunk. E dolgozatban az orosz versmértékrendszerből származó metrikai lejegyzést használjuk, vagyis Jambus-4-fm a négyes jambust jelenti, amelyben rendszeresen váltakoznak a nőrímet és a hímrímet tartalmazó (akatalektikus vagy katalektikus) sorok.

A metrikus kifejezésnek tehát három szintje különböztethető meg egyetlen metrikai formulából:

1. A metrikai mintázat általános *család*ja (pl. a trocheus olyan versmérték, amely bináris verslábon alapul, erős pozícióval az első szótagon);

2. A *versmérték* a lábak számán alapul (pl. az ötös trocheus [Trocheus-5], 5 trochaikus verslábból álló trochaikus pentameter);

3. A versmérték katalektikus *variáns*a, amely az utolsó hangsúlyos szótag után álló nem hangsúlyos szótagok mintázatát írja le (pl. Trocheus-5-fm; f – nőrím (Xu), m – hímrím (X), d – daktilikus (Xuu) sorvég).

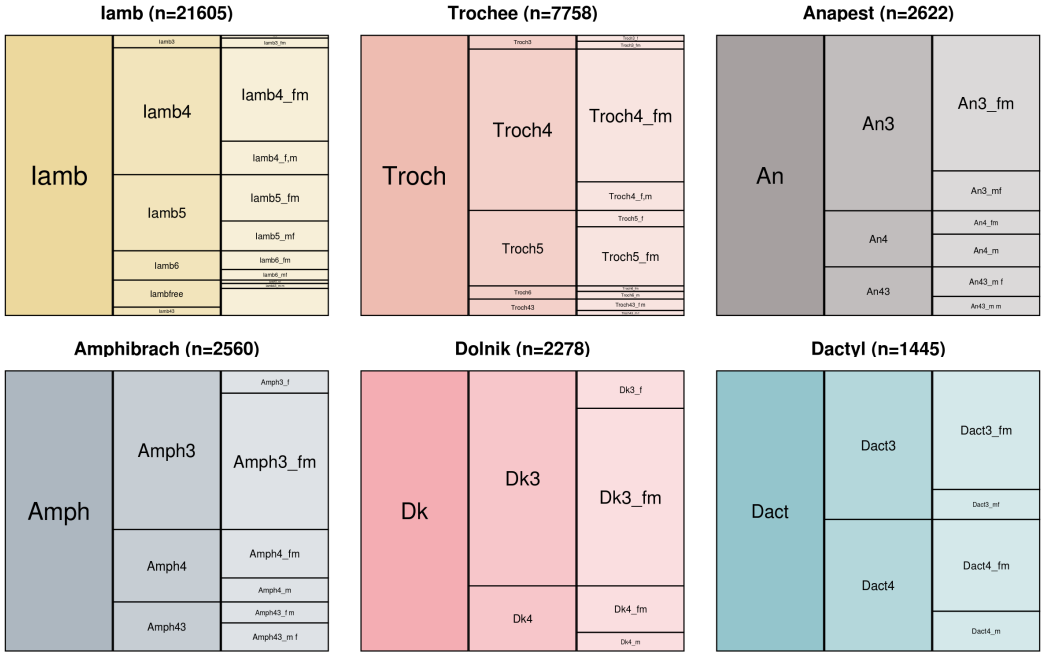
Az *1. ábra* a verseknek ezt a három szintű elosztását mutatja be a korpusz hat leggyakoribb metrikai családjához viszonyítva (versmértékenként csak a két leggyakoribb variáns látható), és megadja a versek családon belüli abszolút számát is. A jambikus versmértéknek, különösen pedig a négyes jambusnak mint az orosz időmértékes verselés „normatív versmértékének" túlsúlya egyértelmű. Azért, hogy kezeljük a versformák ennyire szélsőséges egyenlőtlenségét, a továbbiakban erősen támaszkodunk a véletlenszerű mintavételezésre és az iteratív kísérletekre.

3. A szemantika modellezése

Az egyedi versek szemantikai jellemzői révén igyekszünk modellezni a versmérték és a jelentés közötti kapcsolatot. Ennek érdekében az LDA (Latent Dirichlet Allocation) témamodellező algoritmusát[13] alkalmaztuk a teljes korpuszon a versek csoportosítása nélkül, a metrikai címkéket és az egyéb metaadatokat a dokumentumok nevében feltüntetve.

A témamodell az információt kinyerő algoritmusok egy nagy családjának együttes elnevezése, amelyek az egymás közelében előforduló elemek csoportjait keresik egy dokumentumgyűjteményben. Ezeket a csoportokat témának nevezzük (hiszen az eredeti cél a szövegbányászat volt), de a modellek alkalmazhatók például molekulák-

[13] D. M. Blei, Andrew Y. Ng and Michael I. Jordan, „Latent Dirichlet Allocation," *Journal of Machine Learning Research* 3 (2003): 993–1022.



1. ábra. A metrikai formák aránya a családokon belül. Minden egyes, legalább 200 vershez tartozó versmérték esetében a két leggyakoribb variánst ábrázoljuk. Az abszolút versszámok azonban minden verset magukba foglalnak.

ra,[14] zenére[15] és génekre[16] is, vagy bármilyen feladatra, amely megkívánja a hasonló viselkedésű csoportok kinyerését nagyszámú tulajdonság alapján (szavak, akkordok, gének, kémiai elemek stb.). A témamodellezést ma már széles körben alkalmazzák a bölcsészet- és társadalomtudományokban szövegbányászat és -osztályozás céljából;[17] és már az is többször bebizonyosodott, hogy az LDA kisebb költői szövegek korpuszára is alkalmazható.[18]

[14] Michał Woźniak et al., „Linguistic Measures of Chemical Diversity and the 'Keywords' of Molecular Collections," *Scientific Reports* 8 (2018), https://doi.org/10.1038/s41598-018-25440-6.

[15] Matthias Mauch et al., „The Evolution of Popular Music: USA 1960–2010," *Royal Society Open Science* 2, 5. sz. (2015): https://doi.org/10.1098/rsos.150081.

[16] Manuele Bicego et al., „Investigating Topic Models' Capabilities in Expression Microarray Data Classification," *IEEE/ACM Transactions on Computational Biology and Bioinformatics* 9, 6. sz. (2012): 1831–1836, http://doi.org/10.1109/TCBB.2012.121.

[17] David Hall, Daniel Jurafsky and Christopher D. Manning, „Studying the History of Ideas Using Topic Models," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing. (EMNLP '08)*, 363–371 (Stroudsburg, PA: Association for Computational Linguistics, 2008), https://doi.org/10.3115/1613715.1613763; Paul DiMaggio, Manish Nag and David M. Blei, „Exploiting Affinities Between Topic Modeling and the Sociological Perspective on Culture: Application to Newspaper Coverage of U.S. Government Arts Funding," *Poetics* 41, 6. sz. (2013): 570–606, https://doi.org/10.1016/j.poetic.2013.08.004; Christof Schöch, „Topic Modeling Genre: An Exploration of French Classical and Enlightenment Drama," *Digital Humanities Quarterly* 11, 2. sz. (2017), https://doi.org/10.5281/zenodo.166356.

[18] Ehsaneddin Asgari, Marzyeh Ghassemi and Mark Alan Finlayson, „Confirming the Themes and Interpretive Unity of Ghazal Poetry Using Topic Models," in *Proceedings of the Neural Information Processing Systems (NIPS) Workshop on Topic Models* (Lake Tahoe, NV, December 2013); Borja Navarro-Colorado, „On Poetic Topic Modeling: Extracting Themes and Motifs From a Corpus of

 DOI 10.31400/dh-hun.2021.5.3145

Ígéretesnek tűnik továbbá a témamodellek alkalmazása a kultúrtörténet általános kérdéseinek modellezésére is: megragadhatóvá vált például a popzenében a változás mértéke,[19] az információ tudományos kutatásának módjai[20] vagy az innováció és a korábbi témák napirenden tartása közti különbség a történeti-politikai diskurzusban.[21] Ezekben az esetekben az entitások témareprezentációja pusztán a „tartalom" megragadására irányul. A költői nyelv hasonlóan absztrakt reprezentációját célozzuk meg mi is, tétován utánozva az irodalmárokat, akik olyan magas rendű szemantikai címkéket használtak a versmérték-specifikus jelentések leírásához, mint az este, az út vagy a halál (olyan témák, amelyek Gasparov szerint együttesen fejezik ki a trochaikus pentameter legfőbb szemantikai irányait az orosz költészetben).[22]

Az LDA egy generatív valószínűségmodell, amely néhány nagyon fontos feltételezésen alapul: 1) a gyűjtemény minden szövege *k* számú témából áll; 2) minden egyes téma leírható az összes rendelkezésre álló jellemző (jelen esetben: szavak) valószínűségi eloszlásával (ahol a legtöbb jellemző nagyon valószínűtlen az adott témára). Az LDA a szövegeket *k* téma eloszlása mentén határozza meg, így minden dokumentumot lényegében egyenlő méretű vektorként lehet leírni egyetlen „tématérben". Más szavakkal, az LDA megpróbál az együttesen előforduló szavak bizonyos számú csoportjára automatikusan következtetni; ennek köszönhetően minden egyes dokumentum ezeknek a csoportoknak a kombinációjaként lesz értelmezhető. A témamodellek használatát döntő fontosságúnak tekintjük, mert 1) az LDA lehetővé teszi az egyes versek szintjén az egységes szemantikai absztrakciót; 2) a dokumentumokat potenciálisan kis számú és jól értelmezhető dimenzióval fejezi ki; 3) a versekben a témák valószínűségei lehetővé teszik a lényegre törő utólagos elemzést; 4) a témamodellek függetlenítik a megközelítésünket a nyelv- és egyéb specifikus szakterületektől.

A modell betanítása előtt a korpusz előzetes feldolgozását az alábbi lépésekben végeztük el:

Spanish Poetry," *Frontiers in Digital Humanities* 5 (2018), https://doi.org/10.3389/fdigh.2018.00015; Thomas N. Haider, „Diachronic Topics in New High German Poetry," in *Proceedings of the International Digital Humantities Conference. Utrecht, 8–12 July* 2019, https://dev.clariah.nl/files/dh2019/boa/1031.html; Petr Plechac and Thomas N. Haider, „Mapping Topic Evolution Across Poetic Traditions," in *arXiv:2006.15732* [cs. stat], August 2020, hozzáférés: 2021.12.07, https://arxiv.org/abs/2006.15732.

[19] Mauch et al., „The Evolution of Popular Music."

[20] Jaimie Murdock, Colin Allen and Simon DeDeo, „Exploration and Exploitation of Victorian Science in Darwin's Reading Notebooks," *Cognition* 159, február (2017): 117–126, https://doi.org/10.1016/j.cognition.2016.11.012.

[21] Alexander T. J. Barron et al., „Individuals, Institutions, and Innovation in the Debates of the French Revolution," *Proceedings of the National Academy of Sciences of the United States of America* 115, 18. sz. (2018), 4607–4612, https://doi.org/10.1073/pnas.1717729115.

[22] Gasparov, *Metr i smysl.*

1. Minden szöveget lemmatizáltunk a *mystem* 3.1-et használva.[23]

2. A korpuszra egy általános stopszólistát alkalmaztunk (eltávolítottuk a kötőszókat, a partikulákat, az elöljárószókat, a névmásokat és a számneveket).

3. Csökkenteni akartuk a korpusz lexikális változatosságát, ezért csak az 5000 leggyakoribb szóra képeztük ki a modellt. Az LDA eredménye általában annál jobb, minél kevesebb a szórványosan előforduló adat, ezért szokás eltávolítani a ritka szavakat az előkészítés során. Ugyanakkor a költői nyelv szemantikai leegyszerűsítése érdekében különböző, ugyanezen a korpuszon betanított szóbeágyazási *(word-embedding)* modelleket is használtunk, hogy az 5000 szavas „magon" kívül eső szavakat helyettesíteni tudjuk (a word2vec implementációja a gensim nevű *Python* könyvtáron keresztül, a vektor mérete=300). Egy adott szót akkor helyettesítettünk a leggyakoribb 1000 szó valamelyikével, ha annak volt szemantikai szomszédja a hozzá kontextuálisan leginkább hasonlító 10 szó között (a megfelelő vektorok koszinuszhasonlóságában mérve). Ez az eljárás lehetővé teszi számunkra, hogy a szavakat hiponímiájukkal, gyakoribb szinonimáikkal vagy grammatikai variánsaikkal helyettesítsük (pl. kicsinyítő alakok kicserélése), és néhány esetben azt, hogy a költői nyelv tradicionális metonímiáit megmagyarázzuk (pl. a „Pontus" „óceán"-ra cserélése). Az eljárás nem volt tökéletes, némi zajjal járt, ami ugyanakkor nem volt észrevehető hatással a modellre. Azt is meg kell jegyeznünk, hogy az eredményeink akkor sem változtak szélsőségesen, ha nem hajtottuk végre a kontextuális cserét, vagy ha más felső határt szabtunk a leggyakoribb szavak kijelölésénél. A jelentéktelen hatásoktól függetlenül továbbra is a kontextuális cserével létrejött adatokra vonatkozó eredményeket közöljük, mivel hiszünk abban, hogy ez a szemantikus absztrakció felé vett irány fontos és a jövőben fejlesztésre érdemes. Az eredeti korpuszra vonatkozó főbb eredmények a mellékletben találhatók *(B.5. táblázat)*.

4. A korpuszt a szövegméret alapján is korlátoztuk, hogy az LDA-t legalább összehasonlítható szóeloszlású dokumentumokon képezzük ki. Eltávolítottuk a nagyon kicsi (kevesebb mint 4) és a nagyon nagy (több mint 100 soros) verseket, ami az összes szöveg körülbelül 95%-át meghagyta nekünk. Ezután tovább szűrtük a korpuszt a szavak száma alapján, csak a 10 és 90 százalék közötti szövegeket meghagyva (20 és 102 szavas versek, ami nagyjából megfelel a 12 és 50 soros verseknek, ha beleszámoljuk a stopszavak eltávolítását). Ezek a korlátozások mutatják, hogy a modellünk alapvetően a rövid lírai költészetre összpontosít (ami az uralkodó forma az orosz hagyományban, amelyben a vershossz összezsugorodása tapasz-

[23] Ilya Segalovich, „A Fast Morphological Algorithm with Unknown Word Guessing Induced by a Dictionary for a Web Search Engine," in Hamid R. Arabnia and Elena B. Kozerenko, eds., *Proceedings of the International Conference on Machine Learning; Models, Technologies and Applications. MLMTA'03, June 23–26, 2003, Las Vegas, Nevada,* 273–280 (CSREA Press, 2003).

 DOI 10.31400/dh-hun.2021.5.3145

> talható).[24] Úgy véljük azonban, bármilyen eredményünk is van, annak a hosszú elbeszélő költészetre is érvényesnek kell lennie, ahol a versritmus szemantikai hagyományai sokkal hangsúlyosabbnak tűnnek.[25] Az összes művelet után 47804 szöveg maradt a korpuszban (amiből 39220 versnek egyedi, a korpusz annotációjából származó, formára vonatkozó címkéje van).

Nincs általánosan elismert mód a modell számára optimális témaszám meghatározására:[26] ebben a dolgozatban a 80 témával tanított LDA eredményeiről számolunk be, ami a témakoherencia (log-valószínűség) és a témazavar (a modell „meglepetése", amikor nem látott adatot jósol meg) közötti kompromisszum középponti modellje. Más témaszámokkal (10 és 200 között) is teszteltük a főbb eljárásokat, amelyek szintén nagy teljesítményt mutattak (lásd *B.5. táblázat*ot a mellékletben). Az LDA-ra az alpha=0.1 (nem akartuk, hogy sok téma generáljon egyetlen szöveget, s hogy kezelhetetlenek legyenek az eloszlások) és a beta=0.3 (nem akartuk, hogy túl sok szó járuljon hozzá egy témához, inkább csak néhány) beállításokat alkalmaztuk.

A végső modell gyors ellenőrzéséhez összevethetjük a korábban kvalitatív módon meghatározott témákat a versmértékekhez rendelt szavak csoportjával (lásd *B.4. táblázat*ot a mellékletben). Míg néhány téma összeegyeztethetőnek tűnik a versmértékek feltételezett jelentésmezőjével, természetesen nincs közvetlen kapcsolat közöttük. A témák, ha nem is egyeznek meg az absztrahált metrikai témákkal (Gasparov sem rendszerszerűen használta őket a különböző versmértékek leírásakor), még így is jól értelmezhetők és felhasználhatók a céljainkra a versmértékek tartalmának eloszláson alapuló reprezentációjakor.

4. A mező feltérképezése

4.1. A versmértéken belüli hasonlóságok

A „versmérték jelentésmezőjének elmélete" feltételezi, hogy a jelentés nem véletlenszerűen oszlik el a metrikai formák között, azaz hogy minden egyes versmérték történeti módon egyedi szemantikai profilt épít ki. Az elmélet továbbá kumulatívnak tekinti a mezőhatást (legalábbis implicit módon): nem volnánk képesek rekonstruálni a versmérték szemantikáját egy elszigetelt versben, de sajátos mintázat tűnik fel, ha a versmérték használatának sokkal nagyobb körét vizsgáljuk egy tradíción belül. Ezeket az elveket újrafogalmazva azt mondhatjuk, hogy a jelentés-versmérték kapcsolat bizonyos önhasonlóságot feltételez egy versformán belül. Ha a mezőhatás létezik, akkor az ugyanolyan versmértékű versek két független csoportjának egymáshoz közelebb kellene lenniük a jelentés tekintetében, mint a különböző versmértékűekhez.

[24] Artjoms Šeļa and Oleg Sobchuk, „The Shortest Species: How the Length of Russian Poetry Changed (1750–1921)," *Studia Metrica et Poetica* 4, 1. sz. (2017): 66–84, https://doi.org/10.12697/smp.2017.4.1.03.

[25] Vö. Gasparov, *Metr i smysl.*

[26] Stefano Sbalchiero and Maciej Eder, „Topic Modeling, Long Texts and the Best Number of Topics: Some Problems and Solutions," *Quality & Quantity* 54, 4. sz. (2020): 1095–1108, https://doi.org/10.1007/s11135-020-00976-w.

Tegyük fel, hogy a hagyomány egészének a megfigyelői vagyunk, és a metrikai mezőkre az 1950-es évekből tekintünk (ez a korpuszunk felső időbeli határa). Ahhoz, hogy teszteljük, vajon a jelentés-versmérték kapcsolat észlelhető-e általános szinten, nem felügyelt osztályozást hajtunk végre minden, legalább 500 versre jellemző versmérték 200 példányának két véletlenszerű mintáján (visszatevés nélkül). Minden mintára kiszámítjuk a témavalószínűségeket, hogy az összesített témaeloszlást reprezentálni tudjuk az adott versmértéken belül. Mivel valószínűségek eloszlásával van dolgunk, következő lépésként a Jensen–Shannon divergenciát (amely szimmetrikus a Kullback–Leibler divergenciával[27]) számítjuk ki a minták között, és az így kapott távolságok alapján alkotjuk meg a hierarchikus klasztereket (dendrogramokat). Ezután folytatódik a mintavétel és az újraszámítás még 100-szor. A 100 klaszterelemzés információiból egy összesített, a „többségi szabály" elvével létrehozott konszenzusfa rajzolható fel:[28] az ábra akkor kapcsol össze elemeket, ha az összes dendrogramon 50%-os egyezés figyelhető meg, vagyis két ág nem kapcsolódik, ha nem tartoznak egy klaszterbe legalább a fák felében *(2a. ábra)*.

Ugyanezt az eljárást lehet alkalmazni a metrikai variánsok szintjén is. A metrikai annotációban levő szórványos adatok és zaj miatt csak a négyes jambust (Iamb-4) és azokat a variánsait használjuk, amelyek legalább 200 versre jellemzők, miközben eltávolítjuk a leggyakoribb variánst diffúz szemantikája miatt (Iamb-4-fm). Ez a négyes jambusnak mindössze négy formáját hagyja meg nekünk *(2b. ábra)*.

Anélkül, hogy ennek a megközelítésnek további komplikációit szóba hoznánk, világos, hogy a korpuszban vannak versmértéken belüli szemantikai hasonlóságok (ugyanazon mértékhez tartozó variánsok egymás mellé rendeződnek az ábrán). Természetesen a metrikai variánsok között felfedezhető szemantikai különbség is, bár az ilyen szintű részletességhez sokkal jobb annotációkra és strófainformációkra volna szükség. Mindenesetre a témainformáció önmagában elég arra, hogy két, ugyanabból a versmértékből származó, vitathatóan nagy mintát következetesen egy csoportba rendezzünk (ha a korpuszunkban egy vers méretének mediánja 50 szó, akkor a versmérték-jelentés kapcsolat 10000 szónyi mintában már meglehetősen hangsúlyos). A metrikai mező „kumulatív" hatását ellenőrizhetjük azáltal, hogy megnézzük, hogyan változik a hierarchikus csoportosítás teljesítménye a minta méretével.

A nem felügyelt osztályozás értékelésére két módszert használtunk: az egyszerű klasztertisztaságot (Cluster Purity, CP: a klaszteregyezések összege osztva az egyedi minták számával[29]) és a korrigált Rand-indexet (Adjusted Rand Index, ARI),[30] amelyeket arra terveztek, hogy összehasonlítsanak két osztályozást; az utóbbi pedig számot ad a véletlenszerű osztályozásról is (visszatérési értéke 0 körüli). Ebben az esetben is a versmértékenként 500 elérhető vers küszöbét használjuk, és az eddigi eljárásokat alkalmazzuk az egyre nagyobb mintákon (250 versig), kiszámolva a CP-t és az ARI-t

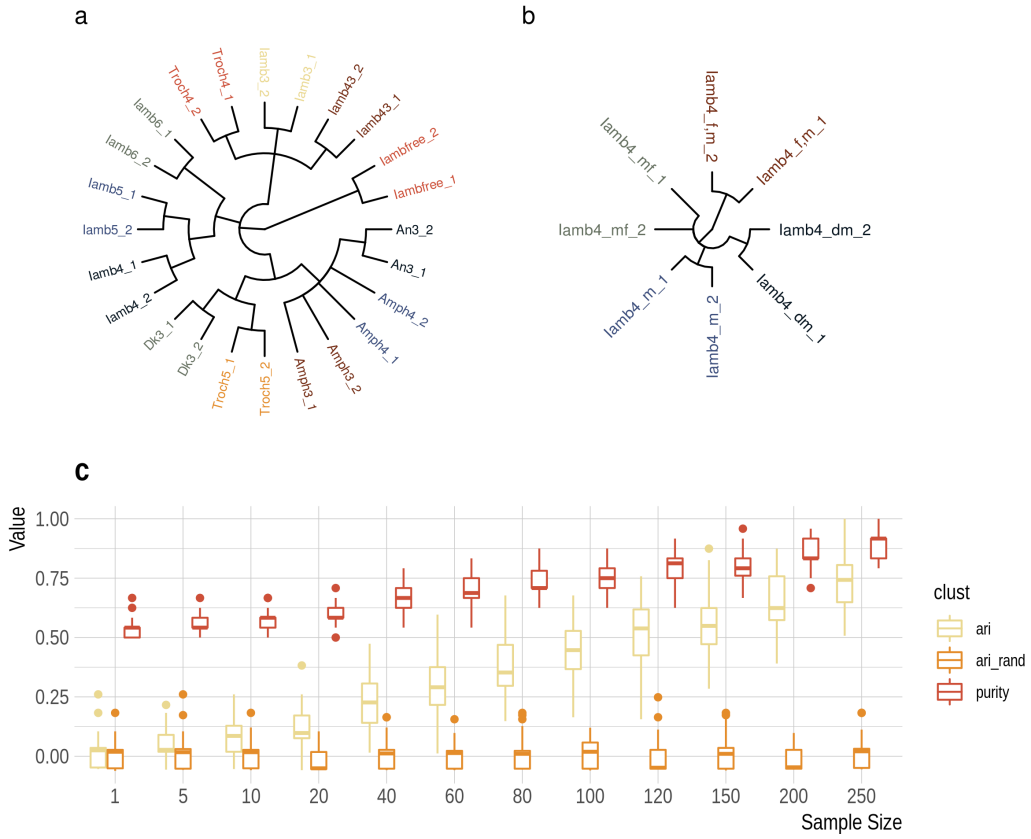
[27] Solomon Kullback and Richard A. Leibler, „On Information and Sufficiency," *The Annals of Mathematical Statistics* 22, 1. sz. (1951): 79–86, http://doi.org/10.1214/aoms/1177729694.

[28] Joseph Felsenstein, „Confidence Limits on Phylogenies: An Approach Using the Bootstrap," *Evolution* 39, 4. sz. (1985): 783–791, http://doi.org/10.1111/j.1558-5646.1985.tb00420.x.

[29] Florian Cafiero and Jean-Baptiste Camps, „Why Molière Most Likely Did Write His Plays," *Science Advances* 5, 11. sz. (2019): 2375–2548, http://doi.org/10.1126/sciadv.aax5489.

[30] Lawrence J. Hubert and Phipps Arabie, „Comparing Partitions," *Journal of Classification* 2, 1. sz. (1985): 193–218, http://doi.org/10.1007/BF01908075.

 DOI 10.31400/dh-hun.2021.5.3145



2. ábra. a) A klaszterezésbeli megegyezéseket mutató, a „többségi szabályt” alkalmazó konszenzusfa (Jensen–Shannon divergencia, teljes kapcsolat) – 100 iteráció, 9 versmérték 2 véletlenszerű mintája, mintánként 200 vers. b) A jambikus variánsok klaszterezése közötti megegyezést mutató konszenzusfa, mintánként 100 vers. c) A hierarchikus klaszterezés teljesítménye CP-ben és ARI-ban az „alapigazsággal” (metrikai címkék) szemben, a véletlenszerűen hozzárendelt klaszterekhez képest. Ugyanazokon a versmértékeken futtatva (mindegyik esetben legalább 500 verssel), mintaméreteként 100 iteráció.

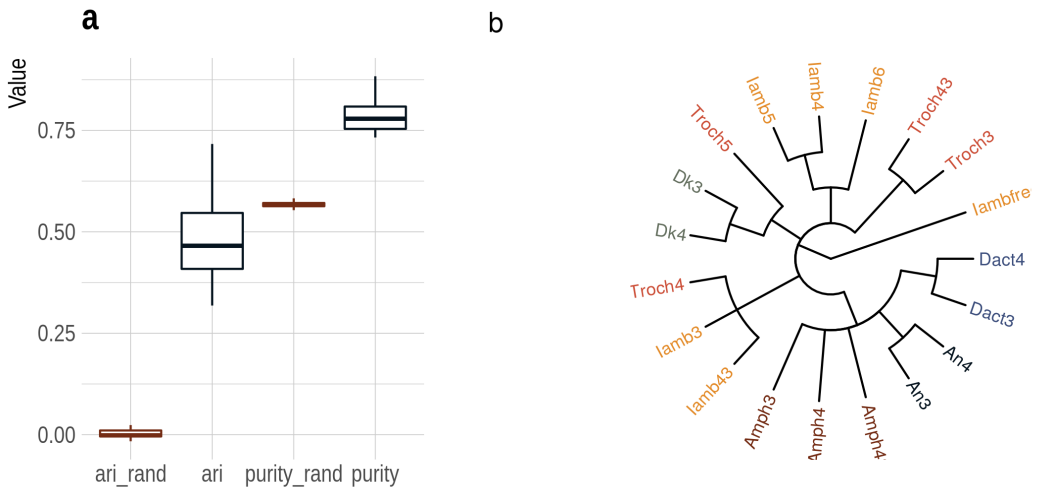
a klaszterezés minden egyes esetére. Mint várható volt, a klaszterezés pontossága a mintában lévő versek számával együtt nő, egészen az ARI=0.73 és a CP=0.90 mediánig *(2c. ábra)*. Ugyanakkor fontos, hogy a nem véletlenszerű klaszterezés hamar észrevehető, és a versmértékekben néhány szemantikus minta felismerhető már mintánként 20–40 versnél.

4.2. A versmértékek közötti hasonlóságok

A *2c. ábra* konszenzusfája rámutat az egyes családokhoz tartozó metrikai formák jelentésbeli kapcsolataira is. Több jambikus méter hajlamos egy klaszterbe kerülni, és ugyanez történik azokkal a hármas formákkal, amelyek világosan egy csoportot alkotnak (az amphibrachus formák közötti minimális eltéréssel). Nem igazán van „alapigazság” arra nézve, hogy hogyan kellene a költői formáknak egymáshoz viszo-

nyulniuk jelentéstani szempontból, kivéve a történeti alkalmazásukra és a hasonló eredetükre vonatkozó néhány megfigyelést. Ugyanakkor számíthatunk rá, hogy a jelentés legalább részben a metrikai családokhoz kötődik (pl. jambikus vagy trochaikus versek), mivel ezek nagyon hasonló ritmikus és grammatikai határokat szabnak a nyelvnek.[31] Néhány esetben az egyes családba tartozó versmértékek történeti kötődéseikben is hasonlóak. Például a legtöbb jambikus mértéket kezdetben magas presztízsű műfajokban használták: a négyes jambust [Iamb-4] az ódákban, az ötös jambust [Iamb-5] a drámában, a hatost [Iamb-6] az elégiában. Ugyanakkor a trochaikus formákat gyakran az „elitet" alkotó jambikus vers ellenpontjaként fogták fel; sőt bizonyos ritmikai vonásuk átfedésbe került a szóbeli hagyománnyal, ami folklórimitációkként határozta meg használatukat, és ennek megfelelő asszociációs mező alakult ki körülöttük.

A feltételezett „családi hatás" tesztelésére egy igen konzervatív kísérletet terveztünk: mivel az elérhető versmértékek száma nem egyezik meg családonként, csak azokat a családokat vettük figyelembe, amelyekben legalább két gyakori versmérték van (> 400 vers). Ezután családonként 20-szor véletlenszerűen veszünk két versmértéket; egy versmértékkészletben 300 véletlenszerűen kiválasztott vers szerepel; ugyanúgy számítjuk ki a klasztereket, mint ahogyan azt leírtuk a 4.1. pontban, ám ezúttal a formák klaszterezését családjuk „alapigazságával" szemben igazoljuk (jambus, trocheus stb.). A folyamatot 100-szor megismételjük a 20 versmértékkészlet mindegyikén. Ennek során jegyezzük az átlagos ARI- és CP-értékek eloszlását minden mintavételezett versmértéknél a véletlenszerű csoportosításhoz mérve.



3. ábra. a) A versmértékek kapcsolatának erőssége a nekik megfelelő családdal. A klasztereket családonként egyenlő számú versmértékkel számolva (k=6, n=12, mintaméret = 300). 100 iteráció minden 20 metrikai készletben. b) Versmértékek közötti stabil szemantikai kapcsolatokat bemutató konszenzusfa (k=6, n=19, mintaméret = 300, fák = 100).

[31] Mikhail Leonovich Gasparov and Marina Tarlinskaja, „The Linguistics of Verse," *The Slavic and East European Journal* 52, 2. sz. (2008): 198–207.

 DOI 10.31400/dh-hun.2021.5.3145

Habár az így létrejött klaszterezés teljesítménye talán nem tűnik magasnak (a medián ARI 0.44 körüli, a CP – 0.76), az értékek mégis elegek ahhoz, hogy megerősítsék, legalábbis bizonyos fokig, hogy a versmértéken belüli kapcsolatokat [vagyis az azonos mértékhez tartozó minták kapcsolata, vö. 4.1. alrész – *a szerk.*] a metrikai családok motiválják. Hogy jobban szemléltessük ezt a hatást, 100 klaszterezés eredményéből kiszámítottunk egy konszenzusfát a családonkénti versmértékek számának bármilyen korlátozása nélkül (*3b. ábra:* a jambusok, a daktilusok, az anapesztusok és néhány trocheus rendszerint következetesen egy klaszterbe kerül; a hármas formák szemantikája valamiképpen diffúz marad, de még így is egy klasztert formálnak egymással).

A „rossz attribúciók" esetei szintén informatívok lehetnek, és összhangban vannak a tárgyhoz kapcsolódó tudományos munkákkal. A hármas jambus [Iamb-3] és a négyes trocheus [Trochee-4] hasonlósága jól ismert: mindkettő a 18. századi anakreóni versből ered, és sok variánsban megegyezik a „dal" ["song"] szemantikájával.[32] A négyes és hármas jambus [Iamb-43] a különböző verslábú sorok rendszeres módosulásával szintén a lírai dalból és a balladából fejlődött ki, és a lírikus-epikus költészethez kapcsolódik.

4.3. A jelentésmező időbelisége

Ideje elhagynunk annak a megfigyelőnek a pozícióját, aki „visszatekint" az egész hagyományra. A *2.* és a *3. ábra* világossá teszi, hogy a klaszterezést bizonyos mértékig erősíti az eltérő időben feltűnő versmértékek közötti különbség. Mivel az LDA algoritmus a dokumentumokon belül a szavak együttes előfordulását használja fel, természetszerűen eredményez olyan szócsoportokat, amelyek különböző időből származnak (például a „nép" témája, a „szovjet" téma vagy a naturalisztikus háború témája). Ez határozza meg a divergenciaszámításokat is: a szovjet témának közel nulla a valószínűsége a 18–19. századi szövegekben. Továbbá ezt láthatjuk abból is, ahogyan a dolnik és az ötös trocheus [Trochee-5] következetesen egy klaszterben tűnik fel – két nagyon népszerű 20. századi forma, amelyek korábban ritkán fordultak elő.

A „szabad" jambus jó példa lehet erre. Ezt a formát a változó verslábhosszok jambikus sorainak a szabálytalan módosulásai alkották, és majdnem kizárólag sajátos műfajokban alkalmazták: költői episztolákban, fabulákban és epigrammákban, valamint teljes mértékben elhagyták a használatát az 1850-es évek után. A szabad jambusban írt 1200 versben csak két téma fordul elő, amelyek együttesen 20% valószínűséggel bírnak (állatok és kommunikáció). Az elnevezése ellenére ez a versmérték megfagyott az időben, műfajok kombinációja nyomja rá a bélyegét, ezáltal két innen származó mintát nem nehéz egy klaszterbe rendezni. Röviden, a szemantikai absztrakciónk nem eléggé absztrakt ahhoz, hogy figyelmen kívül hagyjuk a kronológiai különbségeket.

Az idő ugyanakkor nem érvényteleníti a metrikai mező általános jelenlétét; végső soron a metrikai formák aszinkron fejlődése és divatbeli változásai alakítják az észlelt különbségeket, és határozott korokhoz kötik őket (a négyes jambus [Iamb-4-fm] az 1820-as, 1830-as évek „aranykorához", a hármas daktilus [Dactyl-3] az 1850–1880-as évek polgári és politikai érzületéhez, a dolnik a modernista költészethez). A korpusz

[32] Gasparov, *Metr i smysl.*

ellenőrzése kronológiai szempontból nemcsak a mezőhatás kisebb léptékű tesztelése szempontjából hasznos, hanem lehetőségeket teremt a versmérték-jelentés kapcsolat működésének időbeli vizsgálatához is. Mindezt azonban csak röviden érintjük, mivel ez külön probléma, amely célzott kísérleteket és adatokat érdemel.

Először is azt szeretnénk látni, hogy a versmértéken belüli szemantikai hasonlóságok jelen vannak-e, ha minden versminta ugyanabból az időből származik. Ennek érdekében a korpuszunkat 30 éves szakaszokra osztjuk fel (kizárva a 18. századot, mert ott a népszerű versmértékek változatossága nem nagy). Minden egyes időkeretből vesszük annak hat leggyakoribb versmértékét, és jelezzük az átlagos ARI-értékeket *(1. táblázat)*. Ezek az értékek nem hasonlíthatók közvetlenül össze, mivel különböző versmértékekre és mintaméretekre vonatkoznak (az alsó határ időszakonként a legritkább versmértékű szövegek számának a fele), de elég ahhoz, hogy rámutassunk: a mezőhatás észlelhető korlátozott időkereteken belül is, sőt bizonyos időszakokban kisebb mintákban is, mint várható (vö. *2. ábra*).

Másodszor arra is fel lehet használni a kronológiai információt, hogy kérdéseket fogalmazzunk meg a mezőviselkedéssel és a szemantikai felhalmozással kapcsolatban. Ha egy versmérték jelentésmezőjét történeti és nem organikus jelenségnek tekintjük, akkor arra számíthatunk, hogy a versmérték és a jelentés közötti kapcsolat gyengül az idő múlásával. Egészen pontosan arra számítunk, hogy eltérést találunk a 19. század eleji és végi költészet klaszterezésének értékelésében. Ez megerősítené a versmérték szemantikájának történetiségét, és feloldaná a forma és a műfaj közötti merev kapcsolatot, amelyet a normatív esztétika erőltetett a költészetre. Mivel az osztályozásunk függ a minta méretétől, egyszerűen elfelezzük a 19. századi adatokat, és megfigyeljük az ARI-értékek eloszlását: a klaszterezést szigorúan csak ugyanazon a versmértékkészleten és ugyanakkora mintával végezzük el mindkét időbeli csoport esetében.

Kiderült, hogy a két periódus közötti különbség jelentős *(2. táblázat)*. A 40 versből álló mintákban – a két csoport esetében a lehető legnagyobb mintaméret – jobban elkülöníthetők a versmértékek a 19. század első felében, amely tudatosabb versmértékhasználatot jelez. Ha a 19. század második felében növeljük a minta méretét, akkor az átlagos klaszterpontosság megjósolhatóan emelkedik (mintaméret=100, ARI=0.43), ami a versmértékekben a szemantikai felhalmozódás folyamatára mutat – azaz szemantikailag diffúzabbak lesznek, de nem válnak felismerhetetlenné.

1. táblázat. Versmérték-jelentés kapcsolat különböző időkeretekben, 100 iteráció periódusonként

Period	Median ARI	Poems per sample	Meters
1800-1829	0.51	30	I4 If T4 I6 I5 I3
1830-1859	0.47	50	I4 T4 I6 I5 If Amph3
1860-1889	0.23	30	I4 T4 I6 I5 An3 An43
1890-1919	0.77	270	I4 I5 T4 I6 An3 T5
1920-1949	0.58	250	I4 I5 T4 T5 Dk3 An3

 DOI 10.31400/dh-hun.2021.5.3145

2. táblázat. A 19. század első és második fele, összehangolt klaszterezés. 1000 mintavételező iteráció minden periódusban. Jelentős a különbség az ARI-k eloszlásában (t-test, t=19,6, $p < 0{,}0001$)

Period	Median ARI	Poems per sample	Meters
1800-1849	0.48	40	I4 T4 If I6 I5 Amph4
1850-1899	0.33		

4.4. A témakifejezés és a versmérték gyakorisága

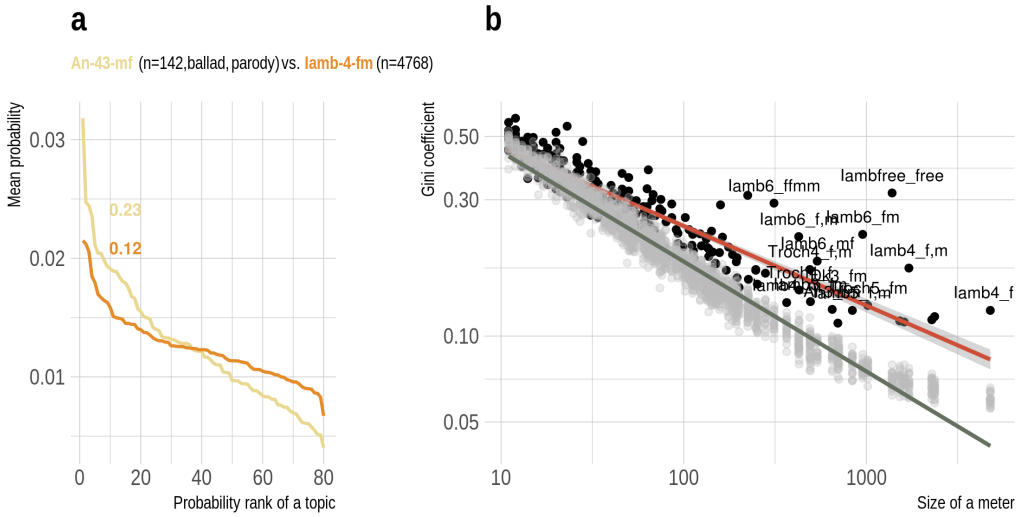
A dolgozat elejétől fogva ijesztő kérdés árnyékolja be az eredményeket: mi van, ha az egész mezőhatás abból az egyszerű tényből ered, hogy a metrikai formák népszerűsége változó? Könnyű bármilyen témakonfigurációra következtetni a négyes jambus, és nehezebb a hármas jambus esetében, míg a hármas trocheus [Trochee-3-dm] esetében szinte nem is lehet. A jelentésmező így talán a mintavételi hibáknak, az időbeli különbségeknek és az azonos metrumú verseket hasonlóan olvasó irodalmárok torzító visszaigazolásainak a kombinációjaként született.

Sőt ennél is bonyolultabb a helyzet, mivel az irodalmárok szerint számítanunk kell arra is, hogy egy mező „jellegzetessége" természeténél fogva csökken, ha nő egy versmérték gyakorisága. Egyszerűen a ritka formákban nincs helye a szemantikai variációnak. Ennélfogva feltételezzük, hogy 1) egy versmérték kifejezésének erőssége és gyakorisága között lineáris kapcsolatnak kell lennie; 2) amennyiben a jelentéstulajdonítás a versmértékek esetében véletlenszerű, másmilyen tendenciát kellene látnunk.

Ahhoz, hogy felmérjük, mennyire „jellegzetes" egyetlen versmértéken belül a jelentés, felhasználhatjuk a témavalószínűségek eloszlásának grafikonjait, és megnézhetjük őket az „egyenlőtlenség"[33] perspektívájából. A kevésbé népszerű versmértékek esetében arra számítunk, hogy kevesebb jellemző témát találunk, mint a nagyon gyakori formákban. Erre mutat példát a *4a. ábra:* a témavalószínűségeket a két metrikai variáns (a négyes jambus [Iamb-4-fm] és a négyes–hármas anapesztus [An-43-mf]) összes verséből hoztuk létre, és aszerint rendeztük el, hogy összességében hogyan járulnak hozzá a metrum szemantikájához. Minden egyes így létrejött grafikonra ki lehet számítani az úgynevezett Gini-együtthatót – amelyet eredetileg arra terveztek, hogy mérje a nemzeti egyenlőtlenséget a vagyoneloszlásban. A Gini 1-es értéket vesz fel, amikor egy disztribúció ördögi módon egyenlőtlen (egyetlen témának 100% a valószínűsége), és 0-t, amikor tökéletesen egyenlő (mind a 80 téma valószínűsége 1,25%). Nyilvánvaló, hogy ez az együttható képes megragadni, hogy mennyire koncentrált a versmérték szemantikája, legalábbis viszonylagosan; a Gini abszolút értékeit azonban befolyásolnák az LDA alapbeállításai (a 0,1 alpha magasabb fokú egyenlőtlenséget feltételez egy vers témavalószínűségeiben, mint például 0,5 alpha).

Hogy igazoljuk a mező kiterjedtségét és (nem-)véletlen természetét, először kiszámoljuk a Gini-együtthatókat minden metrikai variánsra, amely legalább 10-szer előfordul a korpuszban. Majd elvégezzük ugyanezt a számítást a szövegek újraelosztását

[33] Azaz, hogy a témák mennyire egyenlően oszlanak meg a verstípusokban – *a szerk.*



4. ábra. a) A témaegyenlőtlenségbeli különbség (Gini-együttható) az általánosan elterjedt négyes jambus (Iamb-4-fm) (0,12) és a négyes és hármas anapesztus (An-43-fm) ritka formája között. b) A szemantikai egyenlőtlenség csökkenése: a versmérték gyakorisága (fekete) versus a véletlenszerűen újraelosztott versek (szürke) alapján, 20 független újraelosztás. A két lineáris modell lejtői különbözőek (-0,28, -0,39), és a modell nagyobb variációt ír le az újraelosztott (R^2 = 0,96), mint az empirikus (R^2 = 0,81) adatban.

követően: minden egyes *n* gyakoriságú metrikai forma esetén azonos számú, *n* versnyi véletlenszerű mintát veszünk (visszatevés nélkül) a korpuszból. Végül minden verset véletlenszerű módon áthelyezünk üres „versmértékkosarakba" – ezt az újraelosztást 20 különböző alkalommal végezzük el. Ha a mezőből nem lehet véletlenszerűen mintát venni, akkor a két pontcsoport között észrevehetőnek kell lennie az eltérésnek, ahogyan az egyenlőtlenség a mintanagysággal korrelál.

A *4b. ábra* az egyenlőtlenségben felfedezhető különbségeket mutatja a véletlenszerűen összesített versek és a valódi metrikai formák közötti logaritmikus skálán. A várakozásoknak megfelelően a versmérték gyakorisága mentén csökken a szemantikai egyenlőtlenség (azaz több téma is jellemző az adott csoportra), ám árulkodók a kivételek, amelyek a gyakori metrikai formákban (mindenekelőtt a szabad jambusban) is koncentrált szemantikai mezőt jeleznek. Másfelől az egyenlőtlenség a véletlenszerűen újraosztott adatban gyorsabban csökken, és sok valódi versmértéket hagy a vonal felett. Ez arra utal, hogy míg a nagyon ritka formákban véletlenszerűen *talán* előfordulhat hasonló szintű egyenlőtlenség, nincs okunk erre számítani a jelentésmező egészében. Nagyon is valószínűtlen, hogy még a mindig semlegesként és általánosan elterjedtként számontartott négyes jambus szemantikai görbéjét is képesek legyünk véletlenszerű mintavétellel létrehozni.

 DOI 10.31400/dh-hun.2021.5.3145

5. Diszkusszió

A dolgozatban megmutattuk, hogy önmagában a témára vonatkozó információ alapján felismerhető egy versforma. Az egyazon versmértékben írt költemények szemantikailag hasonlók maradnak egymáshoz; a különböző versmértékek pedig gyakran akkor mutattak stabil viszonyt, ha egy családból származnak. Az osztályozás pontosságának történeti különbsége szintén azt sugallja, hogy a metrikai formákban szemantikai felhalmozódás történik, valamint a versmérték „jelentése" szétszóródik az idő folyamán anélkül, hogy felismerhetetlenné válna. Ezek a felfedezések, úgy hisszük, megerősítik a jelentésmező elméletét és legfőbb feltételezéseit, legalábbis egy általános szinten.

A jövőben a metrikai mező hatását jobban megértjük majd a kulturális evolúció keretében,[34] ami lehetőséget biztosít arra, hogy jobban kifejtsük azt is, miként gondolkodunk a történeti folyamatokról és a kulturális átadásról *(cultural transmission)*. A „kulturális evolúció" egy kialakulóban lévő tudományterület, amely a kulturális információ (amit általában úgy határoznak meg, hogy minden olyan információ, amelyet szociális tanulással szerzünk meg) változását, fennmaradását és szétszóródását tanulmányozza. Ez a keret sokféle diskurzuson és területen átívelő kutatást fog át: alkalmazták a régészeti leletek,[35] népmesék[36] és középkori kéziratok kulturális törzsfejlődésének rekonstruálására;[37] az emberi tanulás és a kultúra felgyülemlő erőihez való hozzájárulás megértésére;[38] a populáris zenében történő újítások mértékének vizsgálatára;[39] a nyelvfejlődés makromintáinak tanulmányozására[40] vagy a kulturális információ szétszóródásában és fennmaradásában a népesség méretének szerepét vizsgálva .[41]

[34] Alex Mesoudi, *Cultural Evolution: How Darwinian Theory Can Explain Human Culture and Synthesize the Social Sciences* (Chicago: University of Chicago Press, 2011), https://doi.org/10.7208/chicago/9780226520452.001.0001; Oleg Sobchuk, *Charting Artistic Evolution: An Essay in Theory*. PhD-thesis (Tartu: University of Tartu Press, 2018).

[35] Michael J. O'Brien and R. Lee Lyman, „Evolutionary Archeology: Current Status and Future Prospects," *Evolutionary Anthropology: Issues, News, and Reviews* 11, 1. sz. (2002): 26–36, http://doi.org/10.1002/evan.10007.

[36] Sara Graça da Silva and Jamshid J. Tehrani, „Comparative Phylogenetic Analyses Uncover the Ancient Roots of Indo-European Folktales," *Royal Society Open Science* 3, 1. sz. (2016), http://doi.org/10.1098/rsos.150645.

[37] Adrian C. Barbrook et al., „The Phylogeny of *The Canterbury Tales,*" *Nature* 394, 839. sz. (1998), http://doi.org/10.1038/29667; Joris van Zundert. „Computational Methods and Tools," in Philipp Roelli, ed., *Handbook of Stemmatology: History, Methodology, Digital Approaches,* De Gruyter Reference, 292–356 (De Gruyter, 2020), https://doi.org/10.1515/9783110684384-006.

[38] Claudio Tennie, Josep Call and Michael Tomasello, „Ratcheting Up the Ratchet: On the Evolution of Cumulative Culture," *Philosophical Transactions of the Royal Society B: Biological Sciences* 364, 1528. sz. (2009): 2405–2415, http://doi.org/10.1098/rstb.2009.0052.

[39] Mauch et al., „The Evolution of Popular Music."

[40] Russell D. Gray and Fiona M. Jordan, „Language Trees Support the Express-Train Sequence of Austronesian Expansion," *Nature* 405, 6790. sz. (2000): 1052–1055, http://doi.org/10.1038/35016575; Remco Bouckaert et al., „Mapping the Origins and Expansion of the Indo-European Language Family," *Science* 337, 6097. sz. (2012): 957–960, http://doi.org/10.1126/science.1219669.

[41] Joseph Henrich, „Demography and Cultural Evolution: How Adaptive Cultural Processes can Produce Maladaptive Losses: The Tasmanian Case," *American Antiquity* 69, 2. sz. (2004): 197–214, http://doi.org/10.2307/4128416; Adam Powell, Stephen Shennan and Mark G. Thomas, „Late

Nem túlzás azt mondani, hogy minden új vers korábbi versekből ered. Sőt azok legtöbbször az imitáció termékei: legalábbis rendkívül ritka, amikor egy költő teljesen meg tud szabadulni a hagyománytól, vagy egyedül képes megalkotni egy teljes verselési rendszert. Ha mégis megtörténik, nagy rá az esély, hogy ezek az egyéni erőfeszítések nem lesznek hosszú életűek, egyszerűen azért, mert nem lesz elég követőjük. A költői formák kitartók és konzervatívok: az olyan dolgok, mint a jambikus pentameter, a rímelés vagy a szonett mintázata, századokon át képesek túlélni. Ez azt jelenti, hogy az új verseknek elődeikkel nagyon sok közös formai jellemzőjük van – mint például a versmérték –, hatékonyan ismétlik a korábban alkalmazott formát. Érvelhetnénk amellett, hogy semmi sem állíthatja meg a lírai önkifejezés individualizált modern hagyományának költőjét abban, hogy teljesen szabadon használjon egy metrikai formát, függetlenül annak szemantikai vonatkozásaitól; de láthatjuk, hogy nem ez a helyzet.

A versmértékekre és a versformákra tekinthetünk úgy, mint amelyek a kultúra „TRIM"-jeihez (Transmission Isolating Mechanism – 'átviteli elszigetelő mechanizmus')[42] hasonlóan viselkednek. Ezek a mechanizmusok olyan kondíciók (házassági hagyományok, háztartási szerveződés stb.), amelyek fenntartják az információátvitel „vertikális" szintjét (a szülőktől az utódokig) a kultúrában, amit általában a „horizontális" kapcsolatok (kortársaktól kortársakig) kiterjedt területének tekintenek. Hasonló módon korlátozzák a versmértékek a versek szemantikai lehetőségeit, és a jelentéselőállítást homályos, mégis határozott utakra terelik. Ez biztosítja, hogy a modern költészettörténetekben a versmértékek „gyenge műfajokként" viselkednek, és jellemzők bővülő készletét termelik újra azokban a versekben, amelyek szintén hasonló formai eredettel bírnak.

Miért kell a költészetben ennek a formai elszigetelődésnek egyáltalán megtörténnie? Az egyik kézenfekvő válasz, hogy a versmérték képes hatékony mnemotechnikai rendszerként működni, amely a nyelvet egy magasabb szintű mintázatba helyezi és növeli a megjegyezhetőséget.[43] A költői formák a szóbeli hagyományból erednek, ami nagyon sok formális működésen alapult (versmérték, rím, nyelvi formulák, történetek stb.), ezek behatárolták, hogyan lehet létrehozni és újramondani egy szöveget úgy, hogy elősegítse a memorizálását és az átvitelét.[44] A versmérték emlékeztető ereje nyilvánvalóan számít az írásos tradícióban is. Nem csupán egy forma, amelyre emlékeznek és amelyet újraalkotnak; már azáltal, hogy használatban van, a versmérték továbbcipeli örökké növekvő batyuját. Az orosz költészet sok fordulatot és tudatosan irányított forradalmat foglal magába az egyes versmértékekkel kapcsolatos elvárások-

Pleistocene Demography and the Appearance of Modern Human Behavior," *Science* 324, 5932. sz. (2009): 1298–1301, http://doi.org/10.1126/science.1170165.

[42] William H. Durham, „Advances in Evolutionary Culture Theory," *Annual Review of Anthropology* 19, 1. sz. (1990): 187–210, http://doi.org/10.1146/annurev.an.19.100190.001155; Jamshid Tehrani and Mark Collard, „Do Transmission Isolating Mechanisms (TRIMS) Influence Cultural Evolution? Evidence from Patterns of Textile Diversity Within and Between Iranian Tribal Groups," in Roy Ellen, Stephen J. Lycett and Sarah E. Johns, eds., *Understanding Cultural Transmission in Anthropology: A Critical Synthesis,* 148–164 (Oxford: Berghahn Books, 2013).

[43] Gronas, *Cognitive Poetics and Cultural Memory.*

[44] David C. Rubin, *Memory in Oral Traditions: The Cognitive Psychology of Epic, Ballads, and Counting-Out Rhymes* (New York–Oxford: Oxford University Press, 1995).

 DOI 10.31400/dh-hun.2021.5.3145

kal szemben (mert ezek az elvárások léteztek), ugyanakkor úgy tűnik, hogy senki nem menekülhet meg igazán az időmértékes vers mnemonikus zsarnokságától.

Úgy gondoljuk, a szemantikai mező jelen lesz (nagyobb vagy kisebb mértékben) bármely költészeti hagyományban, bármilyen verselési rendszeren is alapuljon, amely lehetővé teszi a sajátos és stabil versformák létrejöttét az idők során. A témamodellek absztrakt, származtatott eszközök készletét nyújtják nekünk (osztályozási pontosság, egyenlőtlenség stb.), amelyek mentén különböző nyelvek és hagyományok válnak összehasonlíthatóvá. Ez egyúttal hozzáférést biztosít az irodalomtörténet általános kérdéseihez is: hogyan „buknak el" a költői műfajok az idők során, milyen mértékig maradnak felismerhetők a versmértékek (ha egyáltalán), hogyan történik az új formák feltalálása, vagy hogy mi a szerepe az egyéni költőknek és egyedi verseknek a szemantikai mező alakításában.

Fordította: Vásári Melinda

Weak Genres: Modeling Association Between Poetic Meter and Meaning in Russian Poetry

This paper aims to formalize an established theory in versification studies known as "semantic halo of a meter" which states that different metrical forms in modern poetry accumulate and retain distinct semantic associations. We use LDA topic modeling on a large-scale corpus of Russian poetry (1750-1950) to represent each poem in one topic space and then proceed to represent each meter as a distribution of aggregated topic probabilities. Using unsupervised classification and extensive sampling we show that robust form-meaning associations are present both within and between metrical forms: two samples of the same meter tend to appear most similar, while two metrical forms of the same family tend to group together. This effect is present if corpus is controlled for chronology and is not an artifact of population size. We argue that similar approach could be used to align and compare semantic halos across languages and traditions to give meaningful general-level answers to questions of literary history.

Keywords:
poetry, semantics, meters, topic modeling, clustering

Köszönetnyilvánítás

Artjoms Šeļát a „Large-Scale Text Analysis and Methodological Foundations of Computational Stylistics" (NCN 2017/26/E/ HS2/01019) elnevezésű projekt keretében a Polish National Science Centre támogatta. Szeretnénk megköszönni a két névtelen bírálónak, hogy figyelmesen olvasták és javították a dolgozatunkat. Köszönjük Joanna Byszuknak, Maciej Edernek, Antonina Martynenkónak, Vera Polilovának és Oleg Sobchuknak a hozzájárulásukat, segítségüket és támogatásukat.

Függelék A

A kódok és az adatok hozzáférhetősége

A Document-Term Matrix, a feldolgozási lépések, a végső modellek és a teljes elemzés szabadon hozzáférhető: `https://github.com/perechen/semantic_halo_rus`. Az elemzés során az *R* 4.0.2. verzióját használtuk, az LDA-implementációt a *topicmodels,*[45] a modell kimenetének kezelését a *tidytext,*[46] a számításokat a *phylentropy*[47] és a *ineq,*[48] a fák megrajzolását a *ggtree,*[49] az ábrákat pedig a *patchwork*[50] csomaggal hoztuk létre. Az ábrákhoz a MonikeMedium színpalettát a *ghibli* csomag szolgáltatta.[51]

[45] Bettina Grün and Kurt Hornik, „Topicmodels: An R Package for Fitting Topic Models," *Journal of Statistical Software* 40, 13. sz. (2011): 1–30, `http://doi.org/10.18637/jss.v040.i13`.

[46] Julia Silge and David Robinson, *Text Mining with R: A Tidy Approach* (O'Reilly Media Inc., 2017), hozzáférés: 2021.12.07, `https://www.tidytextmining.com/`.

[47] Hajk-Georg Drost, „Philentropy: Information Theory and Distance Quantification with R," *Journal of Open Source Software* 26, 3. sz. (2018), `https://doi.org/10.21105/joss.00765`.

[48] Tze-I Yang, Andrew Torget and Rada Mihalcea, „Topic Modeling on Historical Newspapers," in Kalliopi Zervanou and Piroska Lendvai, eds., *Proceedings of the 5th ACL-HLT Workshop on Language Technology for Cultural Heritage, Social Sciences, and Humanities,* 96–104 (Portland OR: Association for Computational Linguistics, 2011).

[49] Guangchuang Yu et al., „Two Methods for Mapping and Visualizing Associated Data on Phylogeny Using Ggtree," *Molecular Biology and Evolution* 35, 12. sz. (2018): 3041–3043, `http://doi.org/10.1093/molbev/msy194`.

[50] Thomas Lin Pedersen, *patchwork: The Composer of Plots* (2020), hozzáférés: 2021.12.07, `https://CRAN.R-project.org/package=patchwork`.

[51] Ewen Henderson, Danielle Desrosiers and Michael Chirico, *ghibli: Studio Ghibli Colour Palettes* (2020), hozzáférés: 2021.07.12, `https://CRAN.R-project.org/package=ghibli`.

 DOI 10.31400/dh-hun.2021.5.3145

Függelék B

B.3. táblázat. A dolgozatban használt legfőbb versmértékek. 1 – a verslábon belül erős pozíciót jelöli (valószínű a hangsúly), 0 – gyenge. A zárójeles szótagok vagy jelen vannak, vagy nem (a sor végén ; dolniknál sor közben is). A klasszikus versmértékekre angol példákat adunk.

Meter	Type	Foot	Example	Comment
Iamb	binary	01	01|01|01|01|01(00) Thus was|I, slee|ping, by | a bro|ther's hand Of life,| of crown,| of queen,| at once| dispatch'd	Iambic Pentameter
Trochee	binary	10	10|10|10|1(00) Tell me| not in| mournful | numbers, Life is | but an | empty | dream	Trochaic Tetrameter
Dactyl	ternary	100	100|100|100|1(00) Brightest and| best of the | sons of the| morning	Dactylic Tetrameter
Amphibrach	ternary	010	010|010|010|01(00) Oh, hush thee,| my baby|, thy sire was | a knight Thy mother | a lady| both lovely | and bright	Amphibrachic Tetrameter
Anapest	ternary	001	001|001(00) He is gone | on the moun|tain He is lost | to the for|est	Anapestic Dimeter
Dolnik	/	/	(00)1(0)01(0)01(00)	3-ictus Dolnik based on number of stressed positions (3) but unstressed syllable interval is limited to 1-2 syllables

B.4. táblázat. A megkülönböztető témák (a szavakat lefordítottuk) három versmértékben Gasparov leírásaihoz hasonlítva. Az átlagtól leginkább eltérő top 10 témát listáztuk. A korábban meghatározott jelentésmező szempontjából releváns témákat kiemeltük.

Meter	Halo (Gasparov)	Topic	Top words
Trochee-5-fm	Night, Landscape, Love, Death, Road	69	**to know, to live, to be, to die, nothing**
		41	war, to go, soldier, battle, bullet
		61	**goodbye, last, to go (away), hand, parting**
		25	**wind, steppe, sand, grass, desert**
		66	**garden, green, leaf, branch, linden**
		45	**train, wheel, smoke, to fly, wind**
		38	window, house, wall, room, table
		39	**water, river, shore, to swim, lake**
		31	**to go, path, road, to cross, leg**
Trochee-3-fm	Song, Road, Nature, Yearning, Love, Death	76	**to sing, song, nightingale, voice**
		77	matter, take, give, comrade, most
		23	**red, to go, oi, white, "ka" (folksong love topic)**
		43	**snow, white, ice, winter, snowy**
		51	door, house, enter, window, wait
		51	**woods, pine, green, tree**
		10	**wind, leaf, autumn, rain, autumn**
		21	**dream, to dream, night, to wake, morning**
		22	night, darkness, murk, dark
		31	**to go, path, road, to cross, leg**
Iamb-4-dm	dangerous movement through space / love	62	**city, tower, wall, stone**
		59	**horror, death, evil, blood**
		1	**star, world, sky, earth, abyss**
		6	**shade, dream, ghost, pale**
		4	**soul, dream, beauty, world, power**
		68	**hour, wait, to come, soon, or**
		12	god, temple, tsar, before, world
		61	**goodbye, last, to go (away), hand, parting**
		80	**city, road, house, light, to go**
		50	poem, write, poet, book, word

B.5. táblázat. Az Ajusted Rand Index mediánértékei a versmértéken belüli klaszterezéshez *(2c. ábra)* különböző mintában (ARI@*n* vers mintánként), LDA-modelleket használva változó *k* számú témával. Az „ARI family" oszlop ARI-értékeket tartalmaz a versmértékek közötti hasonlóságtesztekhez *(3a. ábra)*. A klaszterezés erős teljesítménye különböző modellekben azt mutatja, hogy a témák száma kevéssé van hatással a kísérleti eredményekre, és így meglepő módon nem igazán befolyásoló tényező. A 20 témával ellátott LDA az egyik legmagasabb teljesítményt mutatja a mintánkénti 250 versnél, ami még inkább kiemeli a szemantikai információ redukciójának hatékonyságát. Ugyanakkor van egy éppenhogy észrevehető kompromisszum a lokális (versmértéken belüli) és a globális (versmértékek közötti) felismerés között (úgy tűnik, a 80 és 100 témát magukba foglaló modellek szolgáltatják a legkiegyensúlyozottabb teljesítményt). További teszteket hajtottunk végre az eredeti Document-Term Matrixon tanult LDA-modelleken (a kevésbé gyakori szavaknak a gyakoribb szemantikai szomszédjaikra való cserélése nélkül – lásd a „w/o replacement" részt).

DTM	k	ARI@10	ARI@60	ARI@100	ARI@150	ARI@250	ARI family
w/ replacement	10	0.07	0.28	0.47	0.55	0.62	0.33
	20	0.09	0.28	0.45	0.57	0.76	0.40
	40	0.08	0.31	0.44	0.60	0.71	0.36
	60	0.09	0.30	0.41	0.55	0.70	0.43
	80	0.09	0.29	0.44	0.56	0.73	0.43
	100	0.09	0.31	0.45	0.60	0.76	0.39
	120	0.04	0.30	0.45	0.54	0.70	0.48
	150	0.08	0.28	0.42	0.57	0.70	0.41
	200	0.04	0.28	0.42	0.54	0.70	0.42
w/o replacement	20	0.08	0.34	0.46	0.55	0.70	0.32
	80	0.09	0.31	0.41	0.56	0.76	0.32
	150	0.08	0.27	0.45	0.60	0.73	0.45

 DOI 10.31400/dh-hun.2021.5.3145

Albert Leśniak 0000-0002-7141-576X

Instytutu Języka Polskiego PAN

albert.lesniak@ijp.pan.pl

Zbigniew Pasek 0000-0003-2580-4366

Akademia Górniczo-Hutnicza w Krakowie

pasek@agh.edu.pl

Neoprotestáns és katolikus tanúságtételek a korpuszalapú diskurzuselemzés perspektívájából*

Jelen tanulmány tárgya evangelizáló tanúságtételek komparatív elemzése római katolikus (a Deon.pl internetes portál) és protestáns – pünkösdi-karizmatikus (a *Chrześcijanin* folyóirat stb.) források alapján. A szerzők a korpusznyelvészet eszközeinek felhasználásával rávilágítanak a két korpusz közötti eltérésekre. Érdekesnek bizonyult a „bűn területeinek" rekonstruálása – a rossznak, amit az Isten segítségével a tanúságot tevő emberek legyőztek. Ez két eltérő terület: a katolikus tanúságtételek esetén a szerelem–szex–aszkézis fogalmak köré épül a szemantikai mező. A pünkösdi tanúságtételek az élvezeti cikkek–függőség–patológia fogalmak köré összpontosulnak. A két részből álló kutatás (a korpuszon végzett gyakoriság- és a kulcsszóelemzés, valamint a leglényegesebb kollokációk vizuális bemutatása) eredményeképpen következtetések vonhatók le azokról a domináns jelentéstartományokról, amik meghatározzák a keresztény egyház e két ágához tartozó hívek vallásos életének mentális térképét.

Kulcsszavak:
korpusznyelvészet, keresztény tanúságtételek, lengyel vallási kultúra, korpuszalapú diskurzuselemzés

A modern vallási nyelvben egyre fontosabb helyet foglal el az úgynevezett tanúságtétel, amely Małgorzata D. Nowak definíciója szerint „egy személyes vallási élményről

* Eredeti megjelenés: Albert Leśniak i Zbigniew Pasek, „Świadectwa ewangelikalne i katolickie w perspektywie korpusowej analizy dyskursu," *Socjolingwistyka* 34 (2020), 57–75, https://doi.org/10.17651/SOCJOLING.34.4. Hálásan köszönjük a szerzőknek a magyar nyelvű ábrák létrehozását! – *a szerk.)*

(az Istennel való »találkozásról«) szóló beszámoló".[1] A szerző hangsúlyozza, hogy a tanúságtételek általában az „azelőtt (az élmény előtt) – akkor (az élmény) – most (az élmény után)" séma alapján jönnek létre.[2] Grzegorz Pełczyński megállapítja, hogy a tanúságtétel „az orális (kvázifolklorisztikus) és irodalmi nyelvhasználatban funkcionáló önálló epikai műfajjá" vált.[3] Az amúgy igen változatos keresztény vallásos nyelvben a 20. század második felétől nő meg a tanúságtétel mint vallási kifejezésmód szerepe. A protestáns világból származó terminust (valamint a tanúságtevés gyakorlatát) számos kortárs katolikus megújulási mozgalom is átvette, kicsit másképp hangsúlyozva a tanúságtétel szerepét saját lelkiségükben. A neoprotestantizmus számára a tanúságtétel továbbra is az evangelizáció egyik fő formája, míg a katolikus vallásosságban ez inkább a számos gyakorlat egyike. A pünkösdi tanúságtétel fő célja az, hogy a hallgatókat ösztönözze életük radikális megváltoztatására vagy megerősítse hitükben a már hívőket. Erőteljesebben megnyilvánul benne a megtérés előtti és utáni helyzet ellentétes bemutatásán alapuló séma. A katolikus gyakorlatban a tanúságtétel nemcsak egy gyökeres lelki fordulatról számol be, hanem arról is, hogyan avatkozik be Isten hétköznapi, kevésbé látványos módon az ember életébe. A keresztény tanúságtételeket vizsgálva érdemes megjegyezni továbbá, hogy domináns nyelvi funkciójuk a meggyőzés.

Jelen tanulmány célja a katolikus és protestáns – pünkösdi – tanúságtételek komparatív elemzése, oly módon, hogy a tanúságtételekre az egyes felekezetek – azaz katolikus és pünkösdi – diskurzusának részeként tekint. Ennek során arra vállalkozunk, hogy körülírjuk a vallási diskurzus e két típusának bizonyos sajátosságait. Aleksandra Pawlikowska nyomán a vallási diskurzust következőképpen értjük:

> […] egy adott felekezeti közösség szövegeinek és megnyilvánulásainak összessége, amely tükrözi azt, hogyan tesz állításokat a világról ez a közösség vallási szempontból. A diskurzus a valóság ábrázolásának meghatározott rendszerét tartalmazza, valamint sajátos ontológiát, axiológiát és kommunikálási stratégiák, illetve szabályok rendszerét.[4]

Cikkünkben a következtetéseket a kulcsszavak és a kollokációk hasonlóságainak, illetve eltéréseinek elemzésére alapoztuk. Szándékunk az volt, hogy a korpusznyelvészet eszközeinek felhasználásával bemutassuk, miként jelennek meg vallási eszmék a lengyel római katolikusok és a pünkösdi-karizmatikusok (*protestanckich-zielonoświątkowych,* a továbbiakban: pünkösdi) vallási nyelvezetében. Figyelmünket különösen is a bűn és a bűnösség forrásainak kérdésére fordítottuk, mivel a kutatás folyamán nyilvánvalóvá vált, hogy a két korpusz másképpen alakítja ki a „bűn területeinek" profilját.

[1] Małgorzata Danuta Nowak, *Świadectwo religijne: Gatunek – język – styl* (Lublin: Towarzystwo Naukowe KUL. „Nowenna pompejańska", 2005), hozzáférés: 2021.07.26, `http://mysliborska28.pl/nowenna-pompejanska`.

[2] Uo., 22.

[3] Grzegorz Pełczyński, „«Świadectwa». Opowieści o własnym nawróceniu," *Lud* 90 (2006): 37–52, 43.

[4] Aleksandra Pawlikowska, *Zróżnicowanie leksykalnosemantyczne dyskursów wyznaniowych na materiale polskich tekstów ewangelickich, katolickich i prawosławnych* (Wrocław: Quaestio, 2015), 20.

 DOI 10.31400/dh-hun.2021.5.3146

Kutatási módszertan

Kutatásunk során egy, a római katolikus és a pünkösdi egyházak híveinek tanúságtételeit tartalmazó szövegkorpuszt vizsgáltunk, az ismétlődő kulcsszavakra és azok kollokációira összpontosítva. A vizsgálat alapjául szolgáló adatokat programozási eszközökkel, illetve a korpusznyelvészet és a kvantitatív nyelvészet eredményeire támaszkodva kidolgozott statisztikai mérőszámok segítségével nyertük ki és különítettük el. Az ilyen típusú nyelvi korpuszok, azaz reprezentatív, digitális formátumú, véges méretű szöveges adatok[5] ugyanis kiindulópontként szolgálhatnak egy adott nyelvváltozatról (pl. a vallásos nyelvről) szóló állításoknak. Emellett a visszatérő nyelvi minták (nyelvi szokások) feltárására is alkalmasak, így segítségükkel következtethetünk arra, hogy a nyelvhasználók hogyan használják az adott nyelvet a diszkurzív világ konstruálására.[6] Mellőzve a részletes okfejtéseket magával a diskurzus fogalmával kapcsolatban, szeretnénk kiemelni annak dinamikus jellegét. Ez bizonyos értelemben Michel Foucault meghatározásából fakad: számára a diskurzus nem más, mint „gyakorlatok, amik rendszeresen létrehozzák tárgyuk értelmezését",[7] vagyis esetében a „nyelv a cselekvésben" szemlélet érvényesül. Ez a megközelítés pedig – Paul Baker és Tony McEnery megfigyelése szerint – tökéletesen illeszkedik a korpusznyelvészethez, hiszen ez a tudományterület egy adott nyelv nagy szövegtárainak az elemzésére vállalkozik, sőt „legmélyebb értelmében az egész korpusznyelvészet tulajdonképpen diskurzuselemzés".[8]

A tanulmányban használt két elemzési módszer a lexémák előfordulásának és együttes előfordulásának gyakoriságelemzésén alapul: ez a kulcsszavak és a kollokációk kinyerésének a művelete. A korpuszvizsgálatokban a kulcsszavakra két fő feltevés vonatkozik: először is a jelentés nem oszlik szét egyenletesen a szövegben, azaz vannak szavak, amelyek szemantikai potenciálja nagyobb, „jelentésteljesebbek". Másodszor, a tágabb jelentéssel bíró szavak megkülönböztetése mérhető, ami alapjául szolgál annak, hogy azokat a szavakat tekintsük kulcsszavaknak, amelyek gyakorisága a korpuszban a referenciakorpuszhoz képest statisztikailag jelentős. Amíg a szavak gyakorisági listája azt írja le, hogy a szavak milyen (relatív vagy kumulatív) *gyakoriság*gal fordulnak elő a korpuszban, a kulcsszavak listája azt mutatja meg, hogy mennyire *lényeges*ek ezek a szavak a vizsgált korpusz szempontjából.[9] Mike Scott szerint a kulcsszavak olyan „szavak, amik egy adott szövegben váratlanul gyakran fordulnak elő",[10] és Christopher Tribble-lel együtt hozzáteszi:

[5] Vö. Tony McEnery and Andew Wilson, *Corpus Linguistics: An Introduction* (Edinburgh: Edinburgh University Press, 2001).

[6] Paul Baker, *Using Corpora in Discourse Analysis* (London, New York: Continuum, 2006), 1, https://doi.org/10.5040/9781350933996.

[7] Michael Foucault, *A tudás archeológiája,* ford. Perczel István (Budapest: Atlantisz Könyvkiadó, 1972), 49.

[8] Paul Baker and Tony McEnery, eds., *Corpora and Discourse Studies: Integrating Discourse and Corpora* (London: Palgrave Macmillan, 2015), 4, https://doi.org/10.1057/9781137431738.

[9] Baker, *Using Corpora,* 125.

[10] Mike Scott, „PC Analysis of Key Words – And Key Key Words," *System: An International Journal of Educational Technology and Applied Linguistics* 25, 2. sz. (1997): 233–245, 236, https://doi.org/10.1016/S0346-251X(97)00011-0.

> […] a kulcsfontosság olyan tulajdonság, amivel szavak rendelkezhetnek egy adott szöveg vagy szöveggyűjtemény keretei között és aminek megléte azt sugallja, hogy az érintett szavak fontosak és visszatükrözik, hogy miről szól a szöveg valójában, kihagyva az apróságokat és a lényegtelen részleteket.[11]

Nyelvstatisztikai szempontból ezek tehát olyan lexémák, amelyek gyakorisága egy adott korpuszban váratlanul magas (pozitív kulcsszavak) vagy alacsony (negatív kulcsszavak), emellett pedig olyan szemantikai potenciállal rendelkeznek, ami fontos információkat hordozhat magáról a szövegről vagy a korpuszról.

A diskurzuselemzés számára a kulcsszavak sajátos nyelvi esetek, amelyekre alapozva történik a társadalmi és kulturális valóság konstrukciója. Ennek a kultúraalkotó potenciált figyelembe vevő megközelítésnek előfutára William Raymond volt, aki a mára már híressé vált *Keywords* című munkájában felhívta a figyelmet arra, hogy a kulturális kulcsszavak nemcsak koruk kultúráját írják le, hanem azok is használják őket, akik erről a kultúráról folyó diskurzusban részt vesznek. Williams szerint a kulturális kulcsszavak „szavak és jelentések közös gyűjteménye a legáltalánosabb angol nyelvű diskurzusban azokról a gyakorlatokról és intézményekről, amiket együttesen kultúrának és társadalomnak nevezünk”.[12] Williams nem használt statisztikai módszereket a kultúra szempontjából jelentős szemantikai potenciállal rendelkező szavak kinyerésére, főleg azért nem, mert még nem állt rendelkezésére olyan módszer, ami lehetővé tette volna megfelelő nagyságrendben a szövegek vizsgálatát. Alan Duran úgy vélte:

> A nagy nyelvi korpuszokhoz használt digitális keresési lehetőségek fejlődése (a módszerek, amiket együttesen korpusznyelvészetnek hívunk) arra ösztönöz, hogy figyelmünket újra a kultúrával kapcsolatos kulcsszavak felé irányítsuk.[13]

Hangsúlyozzuk, hogy egy kulcsszólista létrehozása a kultúrával foglalkozó tudományok esetében csak egy bevezető szakasz lehet, amely a vizsgált szavak kontextusának mélyebb elemzéséhez vezet el. Ahogy Victoria Kamasa írja: „a kulcsszavak fontos útmutatók, amik a kutatók figyelmét ráirányítják azokra a jelenségekre, amik az általuk vizsgált diskurzusra jellemzőek. Egyfajta kiindulási pontot jelenthetnek a gyűjtött adatokhoz.”[14]

A korpuszvizsgálatok legtöbbször emergens és disztributív jellegűek.[15] Anna Bączkowska szerint az emergencia, azaz amikor a következtetések levonása előzetes feltevések nélkül, kizárólag a korpuszelemzésből fakadó megfigyelések alapján történik,

[11] Mike Scott and Christopher Tribble, *Textual Patterns: Keyword and Corpus Analysis in Language Education* (Amsterdam: John Benjamins Publishing, 2006), 55–56, https://doi.org/10.1075/scl.22.

[12] Raymond Williams, *Keywords: A Vocabulary of Culture and Society* (London: Fontana, 1983), 15.

[13] Alan Durant, „Raymond Williams's *Keywords:* Investigating Meanings ‘Offered, Felt for, Tested, Confirmed, Asserted, Qualified, Changed’,” *Critical Quarterly* 48, 4. sz. (2006): 1–26, 16, https://doi.org/10.1111/j.1467-8705.2006.00743.x.

[14] Victoria Kamasa, „Techniki językoznawstwa korpusowego wykorzystywane w krytycznej analizie dyskursu. Przegląd,” *Przegląd Socjologii Jakościowej* 10, 2. sz. (2014): 100–117, 107.

[15] Barbara Lewandowska-Tomaszczyk, *Podstawy językoznawstwa korpusowego* (Łódź: Wydawnictwo Uniwersytetu Łódzkiego, 2005).

 DOI 10.31400/dh-hun.2021.5.3146

lényegesen felértékeli a kontextus szerepét, mert feltételezi, hogy a jelentés nemcsak egyes szavak statikus definíciójában rejlik, hanem azok kontextusában is, amit a vizsgált szót körbevevő szavak csoportja kódol.[16] Bączkowska szerint a kontextus részt vesz a kulcsszavak jelentésének megalkotásában, profilálja és pontosítja őket és rámutat a rájuk jellemző szemantikai preferenciákra. Érdemes itt megfigyelni, hogy a kontextus fogalmát szélesebben is lehet érteni, nemcsak mint az adott kulcsszót körbevevő lexikai egységek csoportját, hanem arra a kommunikációs helyzetre is vonatkoztatva, amiben az adott megnyilatkozás elhangzott.[17] Ehhez kapcsolódóan a kontextust a következő értelemben is használni fogjuk: „egy struktúra, ami tartalmazza a társadalmi helyzet összes olyan tulajdonságát, ami lényeges a diskurzus létrehozása és befogadása szempontjából".[18]

Szorosabb értelemben a kontextus azáltal is rekonstruálható, hogy azonosítjuk az úgynevezett kollokációkat, vagyis azokat a szavakat, amelyek hajlamosak az együttes előfordulásra, gyakrabban, mint ahogy ez a valószínűség-elmélet szerint következne – feltéve, hogy a szavak eloszlása a korpuszban véletlenszerű. John R. Firth szintén úgy vélte, hogy a szó jelentéseinek egy része meghatározható kollokációi alapján, például az *éjszaka* (angolul: *night*) szó egyik jelentése megfogalmazható a *sötét* (angolul: *dark*) szón keresztül, ami gyakran kapcsolódik hozzá. Az állítását ebben a híres mondatban foglalta össze: „egy szót megismerhetsz a környezetéből".[19] Baker számára azonban a kollokációk elemzése egy „módszer, amin keresztül megérthetjük a szavak jelentéseit és a szavak közti kapcsolatokat, amiket kisebb skálán, egyetlen szöveg elemzése alapján nehéz lenne meghatározni".[20] A diskurzuselemzésben az effajta kollokációs profilokat összetettebb diszkurzív struktúrák azonosítására használjuk:

> […] legtöbbször azért használjuk őket, hogy részletesebb információkat szerezzünk az adott szavak működéséről a vizsgált szövegekben. Ilyen információ alapján a kutató azután azonosíthatja a diszkurzív struktúrákat.[21]

A statisztikai jelentőség mérése a két szó együttes előfordulása esetén sok szempontból hasonló lehet a korpuszok lexikális eltéréseinek kulcsszavak alapján történő méréséhez, sőt gyakran használják ebből a célból mindkét esetben ugyanazt az asszociációs mértéket. A lexikális egységeket (mind a kulcsszavakat, mind a kollokációkat), amelyek magas gyakorisága a korpuszban statisztikailag jelentős, azaz a véletlenszerűnél jóval magasabb, a jelen vizsgálatban a *log-likelihood* módszer segítségével azonosítottuk, amit először Ted Dunning javasolt a Vaclav Brezina kézikönyvében leírt eljárás szerint.[22] Ez az egyik leggyakrabban használt módszer, amivel meg lehet határozni, hogy az esemény előfordulása (azaz a szó előfordulása az A korpuszban a B korpuszhoz

[16] Anna Bączkowska, „Korpusowa analiza dyskursu związanego z tematyką imigracji w brytyjskiej prasie opiniotwórczej," *Conversatoria Linguistica* 10 (2016): 8–9.

[17] Vö. Barbara Boniecka, „Tekst w kontekście (problemy metodologiczne)," *Polonica* 16 (1994): 43–67.

[18] Katarzyna Skowronek, *Między sacrum a profanum: Studium językoznawcze listów pasterskich Konferencji Episkopatu Polski (1945–2005)* (Kraków: Wydawnictwo Lexis, 2006), 26–27.

[19] John Rupert Firth, *Papers in Linguistics 1934–1951* (London: Oxford University Press, 1957), 179.

[20] Bake, *Using Corpora,* 196.

[21] Kamasa, *Techniki językoznawstwa,* 108.

[22] Lásd Ted Dunning, „Accurate Methods for the Statistics of Surprise and Coincidence," *Computational Linguistics* 19, 1. sz. (1993): 61–74, illetve: Vaclav Brezina, *Statistics in Corpus Linguistics: A Practical*

képest, vagy két szó együttes előfordulása egy korpuszban) statisztikailag jelentős-e, vagy véletlenszerű. Az összes számítást lemmatizált formátumok alapján végeztük, ingyenes licenc alapján használható *Python* 3.7 programozási környezetben. A számítások (a generált adatokkal együtt) elérhetők a `http://github.com/alblesniak/swiadectwa` oldalon.

A kollokációs eredmények egyik lehetséges bemutatási formája a gráf és a kollokációs hálózat.[23] Ez a vizuális módszer lehetővé teszi bonyolult összefüggések elemzését a nyelvi adatkészletben, ami hagyományos táblázatos formában jóval nehezebben volna vizsgálható. Míg a kollokációs gráfokban a csomópontok a leglényegesebb kollokációkat mutatják (egy definiált távolság alapján, amivel meghatározzuk a vizsgált szó előtt és után álló szavaknak azt a lehetséges számát, amit a számítás során figyelembe veszünk, illetve az asszociációs mérték egy szintén előre meghatározott minimális értéke esetén), a kollokációs hálózatok még egy lépést tesznek előre, és a hálózat csomópontjaiban több (esetünkben 50) kulcsszó kapcsolódásait is mutatják, illetve figyelembe veszik a másodrendű kollokációkat (a kollokációk kollokációit) is. A két csomópontot összekötő élek szélessége a kollokációs erőnek felel meg, a csomópontok nagysága pedig a szavaknak a referenciakorpuszhoz képest mutatott kulcsfontosságát ábrázolja.

Munkánk során azt is be akartuk mutatni, hogyan nyitnak utat a korpusznyelvészeti eszközök a digitális bölcsészet irányába, és hogy nemcsak hatékonyabbá teszik a hatalmas méretű szövegkorpuszokkal (adatgyűjteményekkel) folytatott munkát, hanem annak köszönhetően, hogy alkalmasak tendenciák (vagy azok hiányának) bizonyítására, lehetővé teszik a korábbi megfigyelések megerősítését vagy elutasítását, ebből fakadóan pedig megbízhatóbb következtetéseket tesznek megfogalmazhatóvá egy adott csoport modern vallási nyelve és vallási kultúrája terén is.

Korpuszok

A vizsgált korpusz összesen 281746 tokent tartalmaz (373 dokumentumot). Két kisebb alkorpuszból áll: a Pünkösdi Egyház hívei által leírt tanúságtételek (73554 token, 78 dokumentum) és a Római Katolikus Egyház híveinek tanúságtételei (208192 token, 295 dokumentum). A protestáns tanúságtételek korpuszában található szövegek két forrásból származnak: a *Chrześcijanin* folyóiratból – a Pünkösdi Egyház lapjából – és az egyház gyülekezeteinek hivatalos weboldalairól. A referenciakorpusz alapját képező katolikus tanúságtételeket a Deon.pl katolikus internetes portálról gyűjtöttük. Az adatok gyűjtésére automatizált módszert alkalmaztunk, a szövegeket a Scrapy keresőrobot-rendszer felhasználásával, úgynevezett *web scraping* módszerrel nyertük

Guide (New York: Cambridge University Press, 2018), `https://doi.org/10.1017/9781316410899`; Wit Piotr Chlondowski, „Pentekostalizacja w ujęciu kardynała J. Ratzingera oraz dokumentów episkopatu – dar czy zagrożenie?", hozzáférés: 2021.07.30, `https://deon.pl/magazyn/pentekostalizacja-w-ujeciu-kard-j-ratzingera-oraz-dokumentow-episkopatow-dar-czy-zagrozenie,481475`.

[23] Vaclav Brezina, „Collocation Graphs and Networks: Selected Applications," in Pascual Cantos-Gómez and Moisés Almela-Sánchez, eds., *Lexical Collocation Analysis: Quantitative Methods in the Humanities and Social Sciences,* 59–83 (Cham: Springer, 2018), `https://doi.org/10.1007/978-3-319-92582-0_4`.

 DOI 10.31400/dh-hun.2021.5.3146

ki.[24] A protestáns alkorpuszhoz tartozó szövegek túlnyomó többsége a 20–21. század fordulójáról származik, a katolikus tanúságtételek pedig a 2010–2019 közötti évtizedben keletkeztek.

Annak ellenére, hogy a forrásszövegeket különféle médiumokból gyűjtöttük, kommunikációs csatornájuk homogén, ugyanis mindegyik írott szöveg. A szövegekhez való szerkesztői hozzájárulás mértéke nehezen megállapítható, némi óvatossággal mégis kijelenthetjük, hogy szerzői jellegűek, keletkezésük ideje pedig hasonló, így területi, kronológiai és szociológiai-kulturális szempontból is egységesek.[25] A vizsgált korpuszok közötti méretbeli különbség abból fakad, hogy a katolikus forrásszövegek könnyebben elérhetőek digitális formátumban. Érdemes emellett szem előtt tartani azt is, hogy a lengyelországi pünkösdi egyház híveinek száma (kb. 25 ezer fő) jóval kisebb a katolikus egyházénál, ami nem teszi lehetővé egymásnak teljesen megfelelő korpuszok létrehozását. Bár az alkorpuszok a dokumentumok számában eltérnek, a szövegek átlaghosszúsága hasonló, a méretbeli különbségekből eredő esetleges problémákat pedig olyan módon minimalizáltuk, hogy az analízisben a tokenek relatív előfordulási számát (a korpusz méretével normalizált számot), valamint megfelelően megválasztott statisztikai mérőszámot használunk. Ennek fényében az a véleményünk, hogy a korpuszok reprezentatív és összehasonlítható jellegűek.

1. táblázat. A vizsgált korpuszok statisztikai jellemzői.

Tanúságtételek	PROTESTÁNS	KATOLIKUS
Szövegek száma	78	295
Tokenek száma	73 554	208 192
Lemmák száma	6209	11 448
Átlagos tokenszám egy dokumentumban	943	705,8
Minimális tokenszám egy dokumentumban	193	28
Maximális tokenszám egy dokumentumban	5201	3299

A korpuszban szereplő összes dokumentumot tokenekre bontottuk, majd ezeket (a központozás törlése után) lemmatizáltuk, azaz nyelvtani alapalakjukra, úgynevezett lemmákba csoportosítottuk. Ez a folyamat lehetővé teszi minden lexéma számszerűsítését, függetlenül attól, hogy milyen ragozott alakban szerepeltek a szövegben. Az így előkészített korpuszon különféle statisztikai számítási sorozatok végezhetők el az úgynevezett asszociációs mértékek felhasználásával; ezeket két elem közti korreláció vizsgálatában alkalmazzuk, mint például egy lexéma gyakorisága a két különböző korpuszban vagy két lexéma együttes előfordulása ugyanabban a korpuszban. Az első

[24] Lásd Ryan Mitchell, *Ekstrakcja danych z językiem* Python: *Pozyskiwanie danych z Internetu,* tłum Krzysztof Sawka (Gliwice: Helion, 2019).

[25] Jadwiga Sambor, *Słowa i liczby: Zagadnienia językoznawstwa statystycznego* (Wrocław: Zakład Narodowy im. Ossolińskich, 1972), 24.

említett eset lehetővé teszi egy adott korpuszra jellemző lexémák, azaz a kulcsszavak meghatározását, a második segítségével pedig kinyerhetjük azokat a szókapcsolatokat, amelyek hajlamosak az együttes előfordulásra – ezek az úgynevezett kollokációk.

A kutatás eredményei 1.

Kulcsszavak

Az itt bemutatott *2.* és *3. táblázat*ban összesítve mutatjuk a vizsgálat eredményeit. A jelen dolgozatban csak az 50 leggyakoribb kulcsszót vettük figyelembe a két korpuszból. Következtetéseink levonásához felhasználtuk a táblázatokban található további adatokat is; ezeket elérhetővé tettük a Github portálon.

2. táblázat. Az 50 leggyakoribb kulcsszó a katolikus tanúságtételek korpuszában, a *log-likelihood* érték alapján számolva.

	Kulcsszó	log-likelihood érték	Előfordulások száma 1000 szóra
1	rekolekcja – 'lelkigyakorlat'	101,5917	1,124
2	spowiedź –'gyónás'	80,5576	0,8598
3	świadek –'tanú'	71,0707	1,4266
4	chłopak –'srác'	70,6767	1,0231
5	związek –'kapcsolat'	65,8248	0,9126
6	czystość – 1. 'tisztaság', 2. 'szüzesség'	56,9192	0,6148
7	różaniec –'rózsafüzér'	53,0052	0,538
8	ciąża –'terhesség'	50,2416	0,5572
9	odmawiać – 1.'[imát] mondani', 2. 'visszautasítani'	46,1537	0,5956
10	Maryja –'Mária'	43,5417	0,6052
11	modlitwa –'ima'	36,388	2,6802
12	ślub – 1. 'esküvő', 2. 'fogadalom'	36,2593	0,8358
13	badanie –'vizsgálat'	35,5301	0,3891
14	sakrament – 'szentség'	34,9741	0,3843
15	intencja –'intenció, szándék'	34,4187	0,3795
16	tydzień –'hét'	34,1442	1,1096
17	matka –'anya'	33,8066	0,7829
18	małżeństwo –'házasság'	29,4044	0,8021
19	ksiądz –'pap'	27,0809	0,8406
20	ciężki –'nehéz'	24,4665	0,3602
21	miłosierdzie – 'irgalom'	24,3482	0,3266
22	cierpienie –'szenvedés'	23,3219	0,3795

 DOI 10.31400/dh-hun.2021.5.3146

23	kolejny –'következő'	23,2697	1,3305
24	siebie –'magát'	23,2233	5,2404
25	miesiąc –'hónap'	21,8457	1,0951
26	operacja –'műtét'	20,9935	0,3266
27	poród –'szülés'	19,719	0,2113
28	przypadek – 1. 'eset', 2. 'véletlenség'	19,6632	0,4227
29	dług –'adósság'	19,1594	0,2065
30	dziewczyna – 'lány'	18,7765	0,5091
31	uzdrowienie – 1. '(meg)gyógyítás', 2. 'gyógyulás'	18,271	0,3554
32	walczyć –'küzdeni'	18,1164	0,3266
33	pojawiać –'megjelenni'	17,1941	0,317
34	dobra – 'jó'	16,7309	0,2546
35	konkretny –'konkrét'	16,3781	0,1825
36	choroba – 'betegség'	16,3686	0,5236
37	zaufanie – 'bizalom'	16,0594	0,2161
38	przyczyna – 'ok'	15,8257	0,1777
39	pierwsza – 'első'	15,681	1,1192
40	walka – 'küzdelem'	15,6145	0,3266
41	seks – 'szex'	15,6145	0,3266
42	warto – 'érdemes'	15,2374	0,269
43	lekarz –'orvos'	14,979	0,61
44	oglądać –'néz'	14,7715	0,2642
46	obój –'oboa'	14,7251	0,1681
45	owoc –'gyümölcs'	14,7681	0,2354
47	wspólny –'közös'	14,5784	0,4995
48	kryzys –'válság'	14,1772	0,1633
49	znak –'jel'	13,8904	0,3074
50	doświadczenie –'tapasztalat'	13,5202	0,6677

3. táblázat. Az 50 leggyakoribb kulcsszó a protestáns tanúságtételek korpuszában, a *log-likelihood* érték alapján számolva.

	Kulcsszó	log-likelihood érték	Előfordulások száma 1000 szóra
1	bóg –'isten'	189,9654	12,5894
2	Jezus –'Jézus'	173,5969	6,3627
3	Biblia –'Biblia'	164,6221	1,278
4	życie –'élet'	161,3172	9,8295
5	zbór 'gyülekezet'	154,2113	0,8293
6	alkohol –'alkohol'	133,1858	1,3595
7	Chrystus –'Krisztus'	112,4584	1,7402
8	czytać –'olvasni'	96,0239	1,645

9	chrzest –'keresztelés, keresztelő'	86,2822	0,6798
10	społeczność –'közösség'	85,4751	0,4758
11	pastor –'lelkész'	77,085	0,4622
12	kościół – 1. 'templom', 2. 'egyház'	73,506	2,5831
13	nabożeństwo –'istentisztelet'	72,8846	0,7613
14	papieros –'cigaretta'	67,4022	0,5166
15	ośrodek –'központ'	64,5074	0,6526
16	więzienie –'börtön'	57,3104	0,3807
17	przyjąć –'elfogadni'	56,7842	1,0876
18	pić – 'inni'	54,6864	0,6254
19	zbawienie 'megváltás'	54,1362	0,5302
20	słowo –'szó'	52,7294	2,8414
21	chrześci- jański –'keresztény'		
		52,0894	0,503
22	zbawić –'megváltani'	50,7596	0,4351
23	testament –'testamentum'	49,8966	0,3399
24	wieczny –'örök'	48,5263	0,4215
25	człowiek –'ember'	46,4807	4,133
26	Basia (női becenév)	43,8891	0,4079
27	uczęszczać –'jár, eljár [pl. közösségbe, iskolába]'	43,1411	0,3399
28	pustka –'üresség'	42,3944	0,503
29	pieśń –'ének'	39,5986	0,3807
30	wodny –'vízi'	38,5425	0,2311
31	dom – 'ház'	38,3874	2,488
32	pismo – 'írás'	37,2055	0,5846
33	wieczność – 'örökkévalóság'	35,3232	0,2583
34	narkotyk – 'kábítószer'	33,65	0,4622
35	zbawiciel – 'megváltó'	33,3054	0,4486
36	żyć – 'él'	32,722	1,7266
37	religia – 'vallás'	31,4581	0,3535
38	boży – 'isteni'	30,1777	2,2432
39	grzesznik –'bűnös'	29,2594	0,2719
40	serce – 'szív'	28,6687	2,8686
41	swój – 'saját'	28,4267	2,2432
42	Wojtek (férfi becenév)	28,3178	0,1767
43	werset – '(szentírási) vers'	28,3178	0,1767
44	wołać – 'hívni, kiáltani'	27,3432	0,3127
45	duch ' – lélek'	27,2701	1,4547

 DOI 10.31400/dh-hun.2021.5.3146

46	palić – 1. 'éget', 2. 'cigarettázik'	27,084	0,2855
47	śpiewać – 'énekelni'	23,7291	0,3943
48	syn – 'fiú(gyermek)'	23,7145	1,1284
49	kolega – 'kolléga'	23,6056	0,5982
50	przyjść – 'jönni'	23,6048	0,9789

A *2.* és *3. táblázat*ban foglalt adatok elemzése alapján az alábbi következtetések fogalmazhatók meg. Először is fontos megjegyezni, hogy a katolikus korpuszban erőteljesen jelen van az általunk ideiglenesen szexuális-aszketikusnak elnevezett szemantikai mező (szerelem–szex–aszkézis). A listán szereplő első 1000 kulcsszó elemzése kimutatta, hogy ez a szemantikai szerkezet olyan magas gyakoriságú lexémákból áll, mint: *chłopak* 'fiú, barát', *związek* 'kapcsolat', *czystość* 'tisztaság, szüzesség', *ciąża* 'terhesség', *małżeństwo* 'házasság', *poród* 'szülés', *dziewczyna* 'lány, barátnő', *seks* 'szex', *małżeński* 'házassági', *ukochana* 'kedves[em][nőnemű]', *łóżko* 'ágy', *przytulić* 'ölel', *zakochać* 'beleszeret', *rodzinny* 'családi', *kobieta* 'nő'. A protestáns (pünkösdi) korpuszban azonban egy másik szemantikai mező figyelhető meg, amit a következőképpen lehet megfogalmazni: élvezeti cikkek–függőség–patológia. Olyan szavakból áll, mint: *alkohol* 'alkohol', *papieros* 'cigaretta', *pić* 'inni', *więzienie* 'börtön' és ehhez hasonlók. Szeretnénk hangsúlyozni, hogy tudatosan kihagytunk olyan nyilvánvaló jelenségeket, mint például hogy a protestáns korpuszban olyan szavak dominálnak, mint: *czytać* 'olvasni', *Biblia* 'Biblia', *pismo* 'írás', *testament* 'testamentum', *pastor* 'lelkész', a katolikus korpuszban pedig: *różaniec* 'rózsafüzér', *Maryja* 'Mária', *sakrament* 'szentség', *ksiądz* 'pap', *eucharystia* 'eucharisztia'.

Az elemzésben azokra a szavakra összpontosítottunk, amelyek többé-kevésbe kapcsolatban állnak a kiemelt fogalmi tartományokkal. A két tartományt alkotó lexémák relatív és kumulatív gyakoriságára vonatkozó adatok a *4. táblázat*ban találhatók. A komparatív elemzés eredményeként képet kaptunk az általunk kulcsszóként meghatározott szavak gyakoriságában előforduló aránytalanságokról.

4. táblázat. A szerelem–szex–aszkézis szemantikai szerkezetre jellemző lexémák.

Kulcsszó	Előfordulások száma a katolikus korpuszban	Előfordulások száma a katolikus korpuszban 1000 szóra	Előfordulások száma a protestáns korpuszban	Előfordulások száma a protestáns korpuszban 1000 szóra	Különbség
chłopak – 'fiú barát'	213	1,02	11	0,15	0,87
związek – 'kapcsolat'	190	0,91	9	0,12	0,79
czystość – 1. 'tisztaság', 2. 'szüzesség'	128	0,61	3	0,04	0,57
małżeństwo – 'házasság'	167	0,8	19	0,26	0,54

ciąża – ’terhesség’	116	0,56	3	0,04	0,52
dziewczyna – ’lány’	106	0,51	12	0,16	0,35
seks ’szex’	68	0,33	6	0,08	0,25
poród – ’szülés’	44	0,21	1	0,01	0,2
małżeński – ’házassági, házas’	44	0,21	5	0,07	0,14
łóżko – ’ágy’	54	0,26	9	0,12	0,14
kobieta – ’nő’	102	0,49	26	0,35	0,14
rodzinny – ’családi’	47	0,23	9	0,12	0,1
ukochana – ’kedves[em] [nőnemű]’	21	0,1	1	0,01	0,09
zakochać – ’beleszeret’	20	0,1	2	0,03	0,07
przytulić – ’átölelni’	15	0,07	1	0,01	0,06

5. táblázat. Az élvezeti cikkek–függőség–patológia szemantikai szerkezetre jellemző lexémák.

Kulcsszó	Előfordulások száma a katolikus korpuszban	Előfordulások száma a katolikus korpuszban 1000 szóra	Előfordulások száma a protestáns korpuszban	Előfordulások száma a protestáns korpuszban 1000 szóra	Különbség
alkohol – ’alkohol’	36	0,17	100	1,36	1,19
pić – ’inni’	20	0,1	46	0,63	0,53
papieros – ’cigaretta’	7	0,03	38	0,52	0,48
narkotyk – ’kábítószer’	19	0,09	34	0,46	0,37
więzienie – ’börtön’	3	0,01	28	0,38	0,37
palenie – 1. ’égetés’, 2. ’cigarettázás’	4	0,02	15	0,2	0,18
picie –’ivás’	4	0,02	15	0,2	0,18
wyrok – ’ítélet’	4	0,02	15	0,2	0,18

 DOI 10.31400/dh-hun.2021.5.3146

alkoholik – 'alkoholista'	8	0,04	14	0,19	0,15
karny – 'büntető'	1	0	8	0,11	0,1
kac – 'másnaposság'	1	0	6	0,08	0,08

Az előfordulások számát 1000 szóra vetítve megalapozottnak bizonyult kiemelni a katolikus korpuszban a férfi és nő, fiú és lány közötti kapcsolatra vonatkozó fogalmi tartományt, melynek hangsúlya erősen szexuális és aszketikus (e tekintetben lásd a *czystość* 'szüzesség' szó magas pozícióját, ami a szexuális önmegtartóztatásra vonatkozik). A gondolatmenetet folytatva feltehetjük a kérdést: lehet-e ezeket a statisztikai jellegű tendenciákat úgy értelmezni, mint annak a tézisnek a megerősítését, hogy a Lengyelországban ma élő katolikusok alapvető problémaként tapasztalják meg a (szerelmi) kapcsolatok erotikus dimenziójában a két partner között fellépő feszültségeket. A tanúságtételek elemzése „Isten működésének erejét" ezeknek a kapcsolatoknak a rendezésében mutatja ki. A kapcsolatok ezen szférája kiemelkedik mint a vallásos nyugtalanság, a problémák, vagy – keresztény nyelvezettel – a bűn fő forrása. Az elemzés kimutatja, milyen intenzitással és milyen mértékben van jelen a szexualitás komplexuma ezekben a tanúságtételekben, ami alátámasztja azt az tézist, hogy ez a szféra egyike azoknak, amik a katolikusok mindennapi lelki küzdelmeit dominálják. A probléma mérlegelése során érdemes felidézni azt a vallásszociológusok által már számtalanszor kiemelt ellentmondást, ami az Egyház erre vonatkozó tanítása és a lengyel katolikusoknak ezzel szembemenő hozzáállása és gyakorlata között feszül (például a házasság előtti szex vagy a fogamzásgátló szerek alkalmazásának tiltása). Vallásszociológiai kutatások szerint a katolikus egyház által hirdetett etikai szabályokat minden ötödik lengyel fogadja el.[26]

A Deon.pl portál szerzői (a *Chrześcijanin* szerkesztőihez hasonlóan) a publikáció előtt válogatnak a beérkezett tanúságtételek között. Feltételezhetjük, hogy a szövegek közzététele során figyelembe veszik azok aktualitását, vagyis hogy kapcsolódnak-e a hívők mindennapi problémáihoz vagy az evangelizáció szempontjából potenciálisan hatékonyak-e. A közzétett tanúságtételek azokra a problémákra reflektálnak, amikkel a hívők együttélnek, ezért úgy tűnik, hogy az ezek alapján felállított diagnózis érvényes. A szerelem szexuális dimenziójával kapcsolatos fogalmi komplexum jelenlétét csak részben magyarázza az elsősorban a fiatalokhoz szóló Deon.pl portál sajátos célcsoportja (bár a korpuszban számos idősebbektől származó tanúságtétel található, a szerzők életkorának megállapítása sajnos nem mindig lehetséges). Így hát, némileg szűkítve korábbi megállapításunkat, felállíthatjuk azt a tézist, hogy a fiatal lengyel katolikusok számára jelentős, sőt kulcsfontosságú problémát jelent az intim élet területén összeegyeztetni az egyház tanítását és személyes szükségleteiket.

Érdemes azt is megjegyezni, hogy a katolikus korpuszban a tanúságtételek 65%-a női szerzőtől származik *(6. táblázat)*. Ez a besorolás automatikusan történt, az egyes vagy

[26] Többek között: Janusz Mariański, „Religia i moralność w świadomości Polaków: zależność czy autonomia?" *Konteksty Społeczne* 3, 1. sz. (2015): 12–13.

többes számú, múlt idejű igék ragozott alakjainak gyakorisága alapján – ezek a lengyel nyelvben rendelkeznek olyan tulajdonságokkal, amelyekből megállapítható a beszélő neme. Ez a tény részben megmagyarázza például azt a jelenséget, hogy a *chłopak* 'fiú, barát' szó kétszer gyakrabban fordul elő, mint a *dziewczyna* 'lány, barátnő'.

6. táblázat. A tanúságtételek megoszlása a szerző neme szerint.

Tanúságtételek	férfiként besorolt szerző	nőként besorolt szerző	nem megállapítható
katolikus	89 (30,07%)	194 (65,54%)	13 (4,39%)
protestáns	37 (47,44%)	41 (52,56%)	0 (0,0%)

Az erotikára vonatkozó fogalmi mező jelenlétét kimutató adatok olyan általánosabb trendeket is megvilágítanak, mint például azt, miért növekedett meg az utóbbi tíz évben azoknak a lelki útmutatóknak (könyveknek, közösségi médiában látható filmeknek, szemináriumoknak, workshopoknak és lelkigyakorlatoknak) a népszerűsége, amelyek témájukban egyre gyakrabban érintik az intim természetű kérdéseket és problémákat keresztény perspektívából. Elegendő itt megemlíteni például a kapucinus szerzetes, Ksawery Knotz katolikus kiadói tevékenységét[27] vagy a rendszeres médiaszereplő domonkos szerzetest, Adam Szustakot, akinek YouTube-csatornáját több mint 500000 felhasználó követi.

A pünkösdi tanúságtételek szerzői vegyes társadalmi háttérrel rendelkeznek, az általuk leírtakat pedig a lehető legszélesebb közönségnek szánják. A „bűn és bűnösség" hasonló tartományát keresve ebben a korpuszban érdemes megfigyelni, milyen mértékben felülreprezentáltak (statisztikailag gyakoribbak) azok a szavak, amelyekből kirajzolódik az élvezeti cikkek – függőség – patológia tartományra vonatkozó szemantikai mező. A pünkösdista tanúságtételek magas szintű sematizációja (a szavak/témák változatosságának alacsonyabb foka) alapján ez a fogalmi mező elsősorban a megtérés, a neoprotestáns hívő (ebbe az irányzatba tartozik a pünkösdi mozgalom) életének fordulópontja előtti időszakra vonatkozik. Annak érdekében, hogy a tanúságtétel megmutassa az egyén bensőjét megváltoztató Isten mindenhatóságát, szuggesztív módon írja le a korábbi életszakaszokat. A vizsgált tanúságtételek által érintett területek lefedik a modern lengyel patológia térképét (főként alkoholizmus és kábítószerhasználat).

A kutatás eredményei 2.

Kollokációk és kollokációs hálók

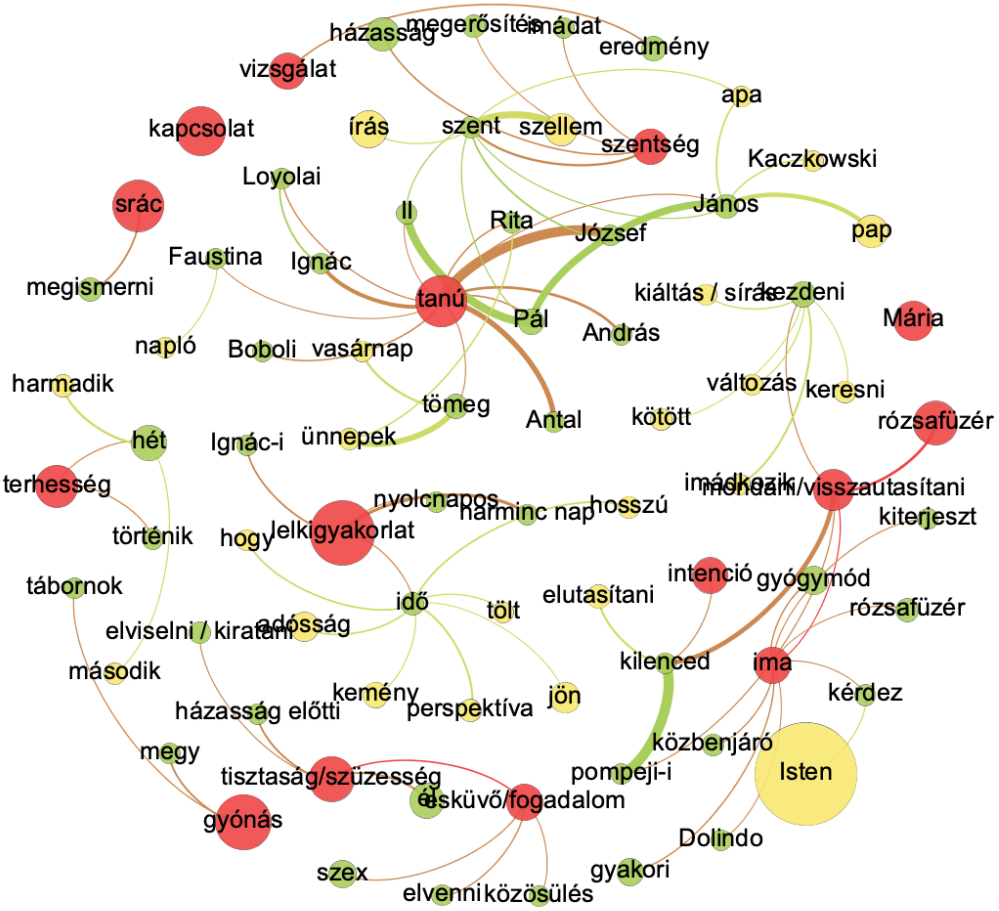
Az analízis következő szakaszai lehetővé tették új adatok begyűjtését. Az eddig megfogalmazott tételeket követve megvizsgáltuk a kulcsszavak kollokációját mindkét korpuszban – amint a módszertani fejezetben már kifejtettük, a nyelvi nyilatkozatok

[27] Ksawery Knotz i Krystyna Strączek, *Seks jest boski, czyli erotyka katolika* (Kraków: Znak, 2010).

 DOI 10.31400/dh-hun.2021.5.3146

kontextusa meghatározza a szavak jelentését, így az elemzésnek ez a fázisa elengedhetetlenül szükséges a kezdeti hipotézisek alátámasztásához.

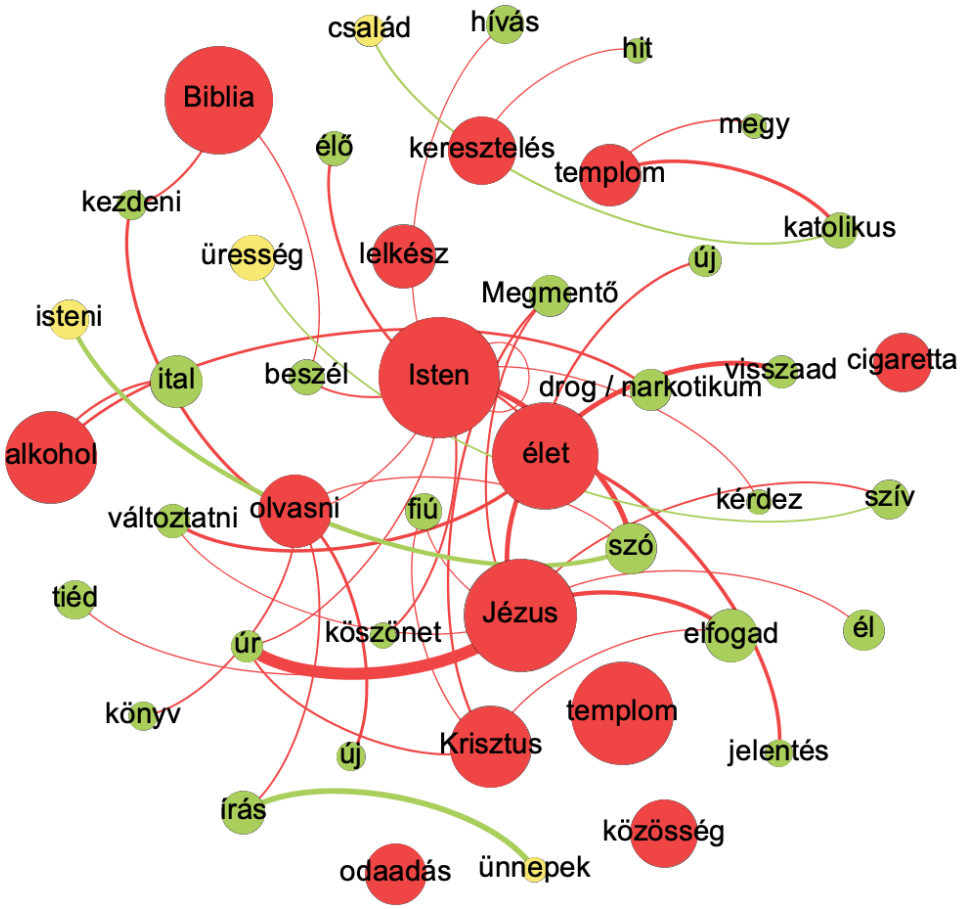
A digitális bölcsészet a vizualizációt mind a vizsgálat alkotóelemeként, mind az eredmények bemutatásának módszereként felhasználja. A továbbiakban bemutatott két háló és két gráf segítségével vizuálisan jelenítjük meg a korpuszok kulcsszavainak kollokációs rendszerét. A pontok mérete megfelel a kulcsszó diskurzív mértékének, amelyet az előfordulásoknak a korpusz méretéhez viszonyított száma alapján állapítottunk meg. Piros színnel a korpuszból kinyert kulcsszavakat jelöltük, zöld színnel ezek kollokációit, sárgával pedig a másodrendű kollokációs kapcsolatokat.



1. ábra. A katolikus korpusz 15 leggyakoribb kulcsszavának vizualizált kollokációs hálózata.

A vizualizált kollokációs háló láthatóvá teszi a protestáns tanúságtételek már említett szemantikai homogenitását *(2. ábra)*. A protestáns korpuszt ábrázoló gráf feltárja, milyen domináns szerepet töltenek be a szemantikai kapcsolatok felépítésében az olyan fogalmak, mint *Bóg* 'Isten', *Jezus*, *Chrystus* 'Jézus Krisztus', *Biblia* 'Biblia', és közvetlenül utánuk: *społeczność* 'közösség' (a neoprotestáns mozgalmak egyik kulcsszava), *alkohol* 'alkohol', *czytać* 'olvasni', *zbór* 'gyülekezet'.

A katolikus kollokációs hálóból *(1. ábra)* a tanúságtételek nagyobb tematikai változatossága olvasható ki: ezek nem mutatnak olyan erős lexikális és strukturális egységességet, mint a protestáns szövegek. A gráf elemzése során látható, hogy olyan szavak foglalnak el központi helyet, mint: *świadek* 'tanú' (ez a *św.* – *święty* 'szent' – rövidítés hibás lemmatizálásának következménye), *rekolekcja* 'lelkigyakorlat', *różaniec* 'rózsafüzér', *Maryja* 'Mária', *sakrament* 'szentség', *spowiedź* 'gyónás', *modlitwa* 'ima'. Nem bizonyultak azonban kevésbé fontosnak a *chłopak* 'fiú barát', *związek* 'kapcsolat', *czystość* 'szüzesség', *ciąża* 'terhesség', *ślub* 'esküvő' szavak sem. Figyeljük meg, hogy a *ślub* 'esküvő' szó az egyértelmű *wziąć* 'megkötni' kifejezésen kívül kapcsolatot alkot a *współżycie* 'együttélés' és *seks* 'szex' szavakkal is.



2. ábra. A protestáns korpusz 15 leggyakoribb kulcsszavának vizualizált kollokációs hálózata.

A katolikus tanúságok „hagyományos" olvasása, amelyet a „digitális" elemzéssel együtt végeztünk, még nagyobb változatosságot mutat a hívők életébe történő „isteni beavatkozás" leírásában. Az Istennel való kapcsolat elérésében hatékonynak bizonyult Galgani Szent Gemma (40 – a zárójelekben itt és a továbbiakban a tanúságtételek sorszáma szerepel) vagy Szent József (38) imája, illetve a zarándoklat Jasna Górára

 DOI 10.31400/dh-hun.2021.5.3146

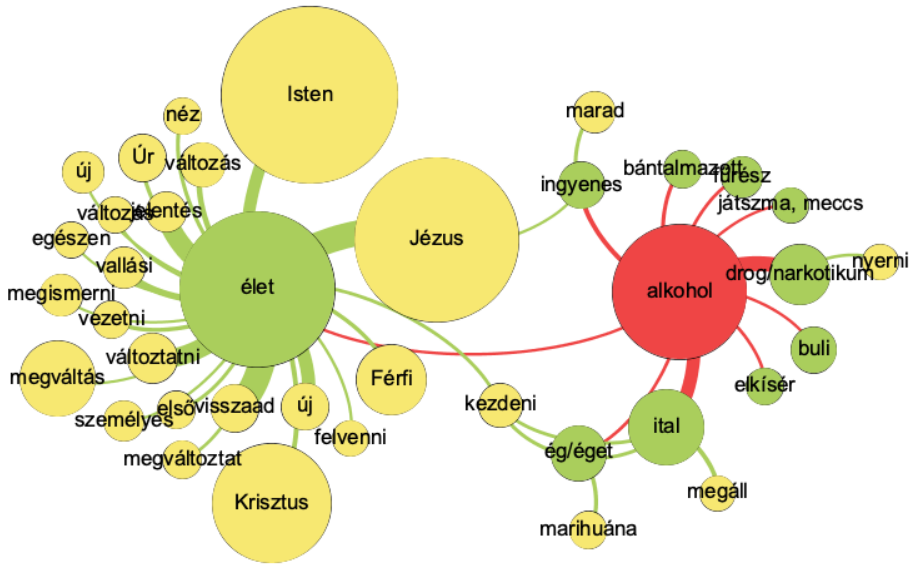
(54). A katolikus korpuszban találhatók olyan tanúságok is, amelyek struktúrájukban hasonlítanak a pünkösdista tanúságtételekre – ezek az emberi életben történő radikális változásokról számolnak be és a protestánshoz hasonló szókincset használnak, mint például *msza o uwolnienie* 'a szabadulásért [mondott] mise' (90), *osobisty egzorcyzm* 'személyes ördögűzés' (100). Ezek mellett feltűnnek olyan tanúságtételek is, mint például az, amelyikben a szerző visszaemlékezik gyermeke csodás megmenekülésére Bobola Szent András képének segítségével a varsói felkelés idején (118), vagy olyan tanúságtételek, amelyekben Isten segítségével megkerül az elveszett kulcscsomó.

Ahogy korábban is említettük, statisztikai szempontból kiemelkedő csoportot alkotnak a tanúságtételek között a házasság előtti tisztaságról szólók, vagy azok, ahol az isteni beavatkozás segít az életre szóló társ megtalálásában és megtartásában (301 és mások). Figyeljük meg, hogy például a *ciąża* 'terhesség' szó nagyon ritkán fordul elő a protestáns tanúságtételekben (*log-likelihood* értéke 10 vagy kevesebb), míg a katolikus korpuszban számos kollokációt alkot, a *być w ciąży* 'terhesnek lenni' pedig a korpusz egyik kulcskifejezése.

3. ábra. A *czystość* 'tisztaság, szüzesség' kulcsszó vizualizált kollokációs hálózata a katolikus korpusz alapján.

A katolikus korpuszban figyelemre méltó a két szóból álló *nowenna pompejańska* 'pompeji kilenced' kollokáció magas gyakorisága is. Ez az ima a Fortunata Agrelli által megtapasztalt Mária-jelenések hatására született 1884-ben, Olaszországban. A kilencedet Bartolo Longo domonkos harmadrendi szerzetes tette népszerűvé. Az ezzel kapcsolatos tanúságok elemzése kimutatja, hogy azok, akik ezt az imát mondják, mélyen hisznek annak hatékonyságában és azt „ellenállhatatlan"-ként írják le.

A következő szakaszban megvizsgáltuk a kulcsszavakat és felállítottuk a kollokációs hálózatokat, hogy verifikáljuk az eddigi következtetéseket. A jelen tanulmány céljaira kiválasztottunk két kifejezést: a *czystość* 'tisztaság, szüzesség' (katolikus korpusz) – *3. ábra* – és az alkohol (protestáns korpusz) – *4. ábra* – szavakat. Ezek vizualizációja azzal a céllal készült, hogy a szövegek mélyebb rétegeibe nyújtsanak bepillantást. A *3. ábrán* például megfigyelhetjük a *czystość* 'tisztaság' erős kapcsolatát a *żyć z Bogiem* 'Istennel élni' frazémával, de ugyanez a szó erős kapcsolatot alkot a következőkkel is: *przedmałżeński* 'házasság előtti', *współżycie* 'együttélés', *seks* 'szex', valamint *dochować* 'megtartani', *walka* 'küzdelem', *trwanie/wytrwać* 'kitartás/kitartani', *zachować/zachowanie* 'fenntartani'. A katolikus diskurzusban lexikalizálódott *dochować* vagy *zachować czystość* 'megtartani a tisztaságot' fordulat hídként tűnik fel két erősen központi komponens, az 'Istennel való élet' és a házasság előtti szexuális együttélés között.



4. ábra. Az alkohol 'alkohol' kulcsszó vizualizált kollokációs hálózata a protestáns korpusz alapján.

A kollokációk elemzése rendkívül érdekes fogalmi metaforákat is feltár, amelyek funkciója a katolikusok házasság előtti szexuális önmegtartóztatásával kapcsolatos nehézségek belső tapasztalatának konceptualizálása. A korábban már említett (részben

 DOI 10.31400/dh-hun.2021.5.3146

várható) szóhasználat mellett, mint amilyen a *zachowanie czystości* 'a tisztaság megtartása', megjelennek kevésbé egyértelmű, de a megtett erőfeszítések mértékéről sokat eláruló, a KÜZDELEM forrástartományból merített metaforikus kifejezések is. A korpuszból olyan, ehhez a fogalmi tartományhoz tartozó mondatokat tudunk példaként kiemelni, mint: „Ne féljetek harcolni a tisztaságért, mert a várakozás ugyan keserű, de gyümölcsei annál édesebbek" (114), „[…] a fiatal nők átlagos életét élem. Mindennap küzdök a tisztaságért" (159), „Máig gondban vagyok ezzel, de mégis harcolok – sikeresen. Már régóta tartom magam a tisztaságban" (183), „Mi magunk küzdünk saját magunkkal, hogy minden ellenére ne lépjük át a határt" (123), „Több olyan emberre van szükség, aki akar harcolni a tisztaságért" (286) stb. A szexuális önmegtartóztatás szünet nélküli *küzdelemként* történő konceptualizációja egy folyamatos feszültséggel teli alak nyelvi képét rajzolja meg, akinek célja az, hogy uralja szexuális ösztönét az általa vallott értékekhez való hűség nevében, ennek érdekében pedig aktív tevékenységet fejt ki.

Az *alkohol* szónak a protestáns korpusz alapján vizualizált hálózata egyfelől bemutatja a kifejezés erős kapcsolódásait az élvezeti cikkek és patológia fogalmaival: *pić/piła* 'ital/ivott', *palić* 'dohányozni', *narkotyk* 'drog', *nadużywać* 'túlzott mértékben használni'. Másfelől megfigyelhető az erős kapcsolódás a *życie* 'élet' szóval és a *zmienić* és *odmienić* 'megváltoztatni', *zmiana* 'változás', *nowe* 'új' szemantikai szerkezettel, amely olyan főkulcsszavak köré épül (lásd a gráfon a pontok méretét), mint például *Bóg* 'Isten', *Jezus* 'Jézus', *Chrystus* 'Krisztus'. A diagram egyrészt ábrázolja a pünkösdi hívő Krisztus-központúságát (az élet változásának okozója Jézus), másrészt láthatóvá válnak rajta a gonosz és a fenyegetés fő területei is, amelyekkel a pünkösdi-karizmatikusok harcban állnak.

Összegzés

Vizsgálatunk lehetővé tette, hogy bepillantást nyerjünk a katolikus és pünkösdi vallási diskurzus egy szeletébe, a vallási tanúságtételekbe. A két korpusz közötti különbséget csak részben magyarázza, hogy azokat különböző médiumokból (internetes portálról és újságból) gyűjtöttük. A megnyilatkozók a két diskurzusban különböző módon közelítik meg magát a keresztény tanúságtételt. A pünkösdiek számára a tanúság beszámoló az egész életet megváltoztató megtérésről, míg a katolikusok számára Istennek a mindennapi életben megnyilvánuló folyamatos segítségéről és jelenlétéről szóló elbeszélés. Az evangelikál életmodell drasztikus megtérést feltételez, a katolikus minta viszont a tökéletesedés irányába kifejtett állandó és fokozatos erőfeszítés modelljéhez áll közelebb. A tanúságtételek az ezeken a területeken kapott isteni segítség különféle formáiról számolnak be.

A különbség azt eredményezi, hogy a pünkösdi tanúságtételek sematikusabbak, kisebb szókincset használnak, így az egyes kulcsszavak hangsúlyosabbak ebben az alkorpuszban. Az élethelyzetek leírása erősebben standardizált, ugyanis a megtérés előtti sötét, gonosz élet (mely leggyakrabban valamilyen függőséghez kapcsolódik) és a lelki fordulat hatására bekövetkező pozitív, gyökeres változás közötti kontrasztra összpontosítanak. A megtérés – a protestáns teológiai nézőpontnak megfelelően – kegyelmi ajándék, Isten adománya. A katolikus tanúságok lexikálisan jóval sokszí-

nűbbek, mert az isteni beavatkozás és segítség jóval változatosabb következményeiről számolnak be. Bár találhatunk kivételeket, fő szabályként mégis hiányzik belőlük az áttörésszerű lelki megtérés motívuma. A katolikus diskurzusban a hitet annak a folyamatos támogatásnak és segítségnek tulajdonítják, amit Isten nyújt az ember mindennapjaiban. A protestáns diskurzusban Isten úgyszólván „pontszerűen" működik, és teljes mélységében megváltoztatja az egyén életét.

Elemzésünk feltárta, hogy a korpuszokat erősen megkülönböztető elem a gonosz problematikája. A katolikus szövegekben a gonosz szemantikai jelentésszerkezete a barát-barátnő, férfi-nő kapcsolatok kontextusában bukkan fel és nagyon gyakran érinti az erotika témáját. A szavak gyakoriságára vonatkozó statisztikai elemzés kimutatja a terület állandó jelenlétét a partneri kapcsolataikat építő fiatalok tudatában. Felállítható az a hipotézis, hogy a kapcsolatok szexuális (testi) dimenziójának hangsúlyosságát a fiatal katolikusokat feszítő dilemmák okozzák. Ezek a fiatalok szeretnének hűek maradni az egyház támasztotta követelményekhez (a házasság előtti szex tilalmához és a közösülést illető szabályozáshoz), egyúttal viszont a kortárs nyugati kultúra liberális szabályaival hasonlítják össze őket. A katolikus tanúságtételek a két értékrend konfliktusából származó feszültségről is tanúskodnak.

Háromfokozatú (a lexikálistól a kollokációkig terjedő) és több száz tanúságtételt figyelembe vevő analízisünkből látható, hogy a lengyel katolikusok életében az isteni jelenlét fő területe az erotika és az érzések szférájának „rendezése". A protestánsok számára az isteni beavatkozás mindenekelőtt a függő ember hívő kereszténnyé változtatásában nyilvánul meg (a függőség palettája rendkívül széles, az alkoholtól a szexen át a kábítószerekig terjed).

A fiatal katolikusokat kínzó lelkiismereti válságok elemzése rámutat, milyen dilemmákkal szembesülnek, miközben vallási identitásukat építik. A kutatás alátámasztja azt a feltevést, hogy az elemzett tanúságtételek nyelvi világa a helyes szexualitás problematikájára koncentrál, mintha katolikusnak lenni egyet jelentene bizonyos konkrét szabályokkal, amelyekben a moralitás elválaszthatatlanul összekapcsolódik a helyes módon folytatott szexuális élettel. A pünkösdi diskurzusban a gonosz valójában a múlt része. A megtérés „megszabadít a gonosztól" és azt a múltba száműzi.

A jelen dolgozatban bemutatott kísérleti kutatás szemlélteti, milyen lehetőségeket kínál a bölcsészettudományok számára a kvantitatív és kvalitatív módszerek összekapcsolása, példánkban a korpuszalapú diskurzuselemzés. Megjeleníti, hogy a korpusznyelvészeti eszközök (amelyeket a digitális bölcsészet egyik elemeként kezelünk) hogyan lehetnek segítségünkre nagy terjedelmű szöveges adatok interpretálásában. Szándékunk szerint demonstráltuk, miképpen egészíthetik ki ezek az eszközök a modern vallásos életről és a lengyel vallásos kultúráról szóló ismereteinket.

Fordította: Zöldy Anna

 DOI 10.31400/dh-hun.2021.5.3146

Evangelical and Catholic Testimonies in the Perspective of Corpus-based Discourse Analysis

The article presents the results of a comparative analysis of Roman Catholic (from the portal Deon.pl) and Protestant, Pentecostal (from the magazine *Chrześcijanin* et al.) Christian testimonies. Using the tools of corpus linguistics, the authors show the differences between both collections of texts. Especially interesting was the reconstruction of the "areas of sin", the evils the testimony authors overcame thanks to God's help. These are different spheres for both types of texts, in the case of Catholic testimonies it is formed by a semantic complex built around the terms of "love – sex – asceticism", and Pentecostal testimonies are focused on the "stimulants – addictions – antisocial behaviour" complex. The analysis consisting of two stages (frequency, describing the role of keywords in the corpus, and visual recognition of the most important collocations) allowed us to formulate conclusions on the dominant areas of meaning which define the mental map of the religious life of the followers of both the Christian denominations.

Keywords:
corpus linguistics, Christian testimonies, Polish religious culture, corpus-based discourse analysis

Helena Grochola-Szczepanek 0000-0002-1511-0486

Instytutu Języka Polskiego PAN

helena.grochola@ijp.pan.pl

Ruprecht Von Waldenfels 0000-0001-5822-5040

Friedrich-Schiller-Universität, Jena

ruprecht.waldenfels@uni-jena.de

Rafał L. Górski 0000-0003-4727-2639

Instytutu Języka Polskiego PAN

rafal.gorski@ijp.pan.pl

Michał Woźniak 0000-0001-9018-2204

Instytutu Języka Polskiego PAN

michal.wozniak@ijp.pan.pl

A szepességi lengyel nyelvjárás korpusznyelvészeti elemzése*

A tanulmány a lengyel Szepesség nyelvjárásának korpuszát létrehozó projektet ismerteti. A lengyelországi dialektológiai kutatások többségétől eltérően korpuszunk a fiatal és a középgeneráció beszédét is tartalmazza, mivel célja a régió nyelvjárásának, szociolingvisztikai helyzetének dokumentálása is. A felvételeket nem fonetikusan, hanem a sztenderd lengyel ortográfiával írtuk át, ami nemcsak a korpuszban való egyszerű keresést teszi lehetővé, hanem azt is, hogy a meglévő eszközökkel lemmatizáljuk és morfoszintaktikai annotációval egészítsük ki a szövegeket. A fonetika iránt érdeklődő felhasználók a felvételeket mondatonként érhetik el. A cikk ismerteti a korpusz összeállításának lépéseit, és tárgyalja a lehetséges alkalmazásokat. A szerzők amellett kívánnak érvelni, hogy egy nagy korpusz, amely egy kis, homogén területet fed le, sokkal értékesebb forrás a dialektológusok számára, mint egy sor kisebb korpusz, amely egy nagyobb régiót dokumentál.

Kulcsszavak:
korpusz, beszélt nyelv, dialektológia, szepességi nyelvjárás

* Eredeti megjelenés: Helena Grochola-Szczepanek, Ruprecht Von Waldenfels, Rafał L. Górski i Michał Woźniak, „Korpus języka mówionego mieszkańców Spisza," *LingVaria* 14, 1. sz. (2019): 165–180.

1. Bevezetés

A nyelvjárási szövegek lejegyzése során mindig nehéz meghozni a döntést, hogy vajon arra kell-e törekednünk, hogy a lehető legnagyobb területen történjen a dokumentáció, beérve ezáltal egy meglehetősen felületes eredménnyel, vagy arra, hogy csak kiválasztott helyszínek elmélyült felmérését végezzük el. Képletesen szólva: tárjuk-e fel a felszín egészét, vagy végezzünk inkább helyszíni mélyfúrásokat? E dilemma nem csupán a dialektológiát vagy az areális nyelvészetet érinti, de tágabb értelemben a művészettörténetet, a botanikát, a zoológiát, a geológiát stb. is. Erre a kérdésre persze nem lehet kizárólagos választ adni, az elkövetkezőkben mégis annak bizonyítására teszünk kísérletet, hogy egy kisebb terület nyelvjárásának részletes feltérképezése tudományos szempontból kiemelten értékes lehet; annak ellenére, hogy ezáltal számos más területet hagyunk felfedezetlenül.

Az alábbi tanulmány a szepességi lengyel nyelvjárás nagy méretű elektronikus korpuszának létrehozását és annak szempontjait hivatott bemutatni.[1] E tudományos munka forradalmian újnak számít a lengyel dialektológiában; habár nyelvjárási szövegek több mint száz éve kerülnek kiadásra nyomtatott formában, ahogyan bizonyos hangfelvételek is már fél évszázada elérhetők a kutatók számára.[2] A szóban forgó munka meglehetősen nagy léptékű, digitális szövegkorpuszát hangfelvételek és azok lejegyzései alkotják. A létrehozott korpusz keresője továbbá lehetővé teszi, hogy egy adott szó előfordulásait, valamint ragozott alakjait másodpercek alatt kikereshessük. Elérhetők ugyan lengyel köznyelvi (például *A lengyel nyelv nemzeti korpusza,* a továbbiakban: NKJP[3]) és lengyel nyelvtörténeti korpuszok is,[4] a *Szepességi korpusz* az első

[1] A „Język mieszkańców Spisza. Korpus tekstów i nagrań gwarowych" ['A lengyel Szepesség lakóinak nyelve. Nyelvjárási szövegek és hangfelvételek korpusza'] című tudományos kutatás 2015–2019 között a Narodowy Program Rozwoju Humanistyki [Nemzeti Bölcsészettudományi Fejlesztési Program] finanszírozásával valósult meg (1bH 15 0166 83).

[2] Ma már léteznek teljesen digitalizált nyelvjárási gyűjtemények is, mint például *A mazóviai nyelvjárások akusztikus adatbázisa* vagy a *Lengyel dialektusok és nyelvjárások. Internetes kézikönyv.* (*Akustyczna baza danych gwar mazowieckich. Wokalizm,* 2013–2017, hozzáférés: 2021.09.22, `http://www.bazamazak.uw.edu.pl`; *Dialekty i gwary polskie. Kompendium internetowe,* pod. red. Haliny Karaś, hozzáférés: 2021.09.22, `http://www.dialektologia.uw.edu.pl/index.php`.) Míg az előbbi kizárólagosan a hanganyagon alapszik, addig az utóbbi, bár szöveget is rendel a hangfelvételekhez, csekély mérete miatt nem képezheti elmélyült kutatás alapját, csupán a dialektológiai ismeretterjesztést célozhatja. A szakirodalomban szó esik Maćkowce falu nyelvjárási korpuszának összeállításáról, e gyűjtést azonban a mai napig nem publikálták: Aleksandra Krawczyk-Wieczorek, „Automatyczna lematyzacja tekstu w zapisie fonetycznym: Korpus polskiej gwary południowokresowej," *Język Polski* 92, 1. sz. (2012): 11–19. Ugyanígy A *Lengyel Nyelvjárások Korpuszának* összeállítása szintén a tervezés fázisában van. Halina Karaś, Monika Kresa i Aleksandra Krawczyk-Wieczorek, „Towards a Corpus of Polish Dialect Texts," *Prace Filologiczne* 63 (2012): 129–145.

[3] *NKJP: Narodowy Korpus Języka Polskiego* (red. A. Przepiórkowski, M. Bańko, R. L. Górski, B. Lewandowska-Tomaszczyk, Varsó, 2012), hozzáférés: 2021.09.22, `http://nkjp.pl./`.

[4] *KorBa: Elektroniczny korpus tekstów polskich z XVII I XVIII w. (do 1772 r.),* hozzáférés: 2021.09.22, `http://korba.edu.pl/`. Lásd még Włodzimierz Gruszczyński, Dorota Adamiec i Maciej Ogrodniczuk, „Elektroniczny korpus tekstów polskich z XVII i XVIII w. (do 1772 r.) – prezentacja projektu

 DOI 10.31400/dh-hun.2021.5.3148

elektronikus lengyel nyelvjárási gyűjtemény, amely a mai kutatási eszközök minden kihívásának megfelel. A projekt az alábbi kutatásokkal áll szoros összefüggésben:

- az Usztja-folyóvidék lakóinak nyelvi korpusza (Arhangelszki terület, Oroszország) `http://parasolcorpus.org/Pushkino/login.php`;
- ruszin nyelvi korpusz `https://www.russinisch.uni-freiburg.de/`;
- Litvánia, Fehéroroszország és Oroszország határvidéki nyelvjárásainak korpusza `http://www.trimcocorpus.de/spoco/`.

Ezek a munkák szakmódszertani megközelítésükben, technikai megoldásaikban és (bizonyos mértékig) információs infrastruktúrájukban is megegyeznek egymással.

2. Földrajzi kiterjedés

A korpusznyelvészeti kutatás a Szepesség lengyelországi részét érinti (15 falut), és nem vonatkozik a Szepesség szlovákiai részére, amelynek kiterjedése jelentősen nagyobb a lengyelországinál. A teljes Szepesség feltérképezése szükségszerű volna ugyan, e vállalkozás a munka jelen fázisában azonban erős pénzügyi és munkaerőbeli korlátokba ütközne.

3. A korpusz összetétele

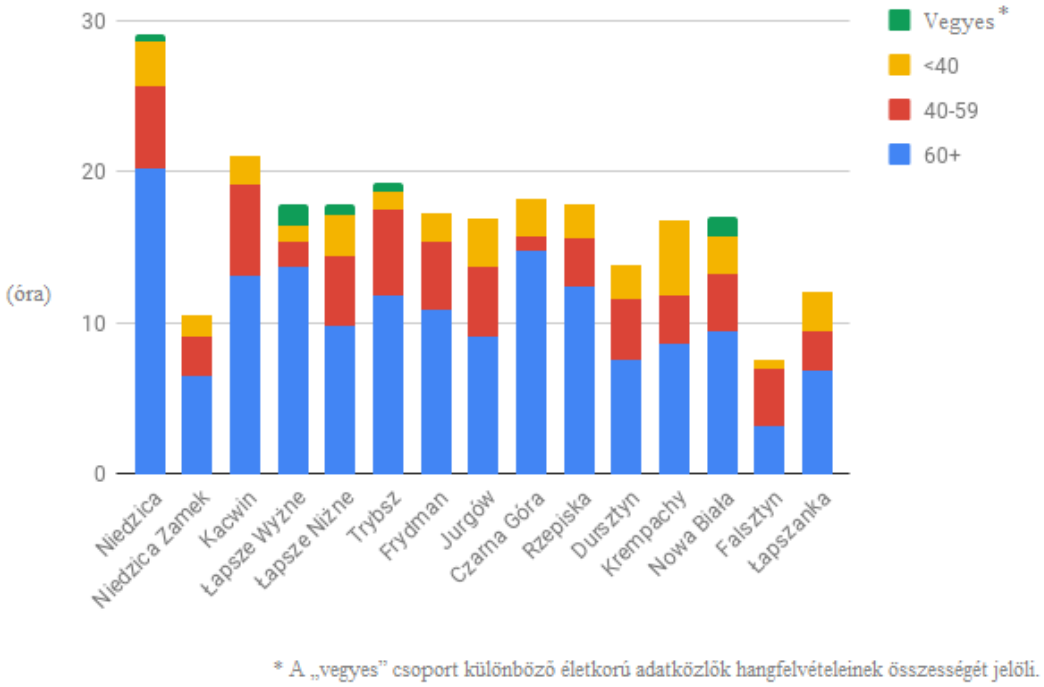
A *Szepességi korpusz* hangfájlok és lejegyzéseik egymáshoz rendelt rendszereként írható le: minden hangfájlhoz egy XML-formátumú szöveges fájl tartozik. A gyűjtemény 340 adatközlőtől származó, több mint 320 felvételből áll, ezek összesen nagyjából 250 órányi anyagot tesznek ki.[5]

A korpusz szövegszinten nem kevesebb mint kétmillió szóból, pontosabban tokenből áll.[6] Token lehet egy szó vagy egy írásjel, bár bizonyos szavak három különálló részre osztva kerültek rögzítésre, például a *chciałbyś* [~'szeretnél'] ige *chciał* [~'szeret, akar'], *by* [a feltételes mód morfémája] és *ś* [E/2 személyrag] egységekre tagolva szerepel a korpuszban, amely elemek a *chcieć* ['akarni'], *by* és *być* [létige] szótári alakokra (lemmákra) vezethetők vissza. Ez utóbbi – az NKJP terminológiája szerint – az úgynevezett agglutináns: olyan mozgó morféma, amely nemcsak az igéhez, de más szófajú szavakhoz is képes kapcsolódni. A korpusz keresési eredményei a szövegek hosszabb részleteibe ágyazva jelennek meg a keresőfelületen, ezeket a lejegyzés *szegmenseinek* nevezzük, és megközelítőleg mondatértékűnek tekinthetők. A korpusz közel kilencvenezer ilyen szegmensből áll.

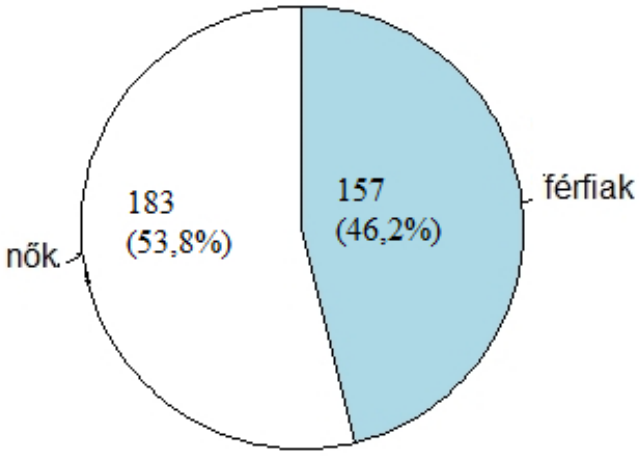
badawczego," *Polonica* 33 (2013): 311–318; Magdalena Derwojedowa, Witold Kieraś, Dorota Skowrońska i Robert Wołosz, „Współczesne narzędzia leksykograficzne a analiza tekstów dawniejszych," *Polonica* 34 (2014): 21–27.

5 A felvételek és adatközlők száma közti különbség azzal magyarázható, hogy bizonyos hangfelvételek két vagy több adatközlő közreműködésével készültek – lásd „vegyes" csoport *(1. ábra)*.

6 Adam Przepiórkowski, *Korpus IPI PAN. Wersja wstępna* (Varsawa: IPI PAN, 2004).



1. ábra. A hangfelvételek hosszúsága lakhely és korcsoport függvényében.



2. ábra. A férfiak és nők aránya a hangfelvételeken.

3.1. A szemléletesség és a hitelesség értékei közötti kompromisszum megtalálása

Egy nyelvi korpusznak reprezentatívan kell tükröznie egy adott nyelvi közösség beszédét. Ennek során figyelmen kívül hagyjuk az írott nyelvi szövegek kérdését (e feladat jóval bonyolultabb volna), és kizárólag beszélt nyelvi adatokra szorítkozunk.[7] Ebben

[7] Vö. Rafał L. Górski i Marek Łaziński, „Reprezentatywność i zrównoważenie korpusu," in Adam Przepiórkowski, Mirosław Bańko, Rafał L. Górski i Barbara Lewandowska-Tomaszczyk, red., *Narodowy Korpus Języka Polskiego,* 25–36 (Warsawa: Wydawnictwo Naukowe PWN, 2012).

 DOI 10.31400/dh-hun.2021.5.3148

az esetben a reprezentativitást hasonlóképpen értelmezhetjük, mint egy közvéleménykutatásnál, amennyiben egy korpusztól is azt várjuk, hogy abban a férfiak és a nők (vö. *2. ábra*), a fiatalok és az idősek, a felsőfokú képesítéssel rendelkezők, illetve nem rendelkezők aránya megegyezzen a magában a közösségben megfigyelhető arányokkal. Ilyen megközelítés adhat „átlagolt" képet az adott nyelvről. A korpusz létrehozásakor kismértékben el kellett térnünk ettől a módszertől, méghozzá úgy, hogy az alanyok kiválasztásánál arra törekedtünk, hogy az adatközlők között nagyobb számban legyenek az idősebbek, akik körében a nyelvjárás korábbi állapotában őrződött meg; beszédükön kevésbé figyelhető meg a lengyel köznyelv hatása. E csoport mérsékelt túlsúlya olyan adathalmazt eredményezhet, amely a nyelvjárási sajátosságokat a lehető legtisztább formájában teszi elérhetővé. Ez azonban nem lehet az egyedüli szempont. A középkorú és fiatal generációk felvételeinek segítségével ugyanis más, hasonlóan fontos jellegzetességeket is megfigyelhetünk, például a nyelvjárási elemek eltűnésének és a nyelv sztenderdizációjának folyamatát. Továbbá fontos az is, hogy a nyelvjárást hitelesen, aktuális állapotában őrizzük meg egy ilyen kutatásban, beleértve ebbe a fiatalabb korosztályok beszédét is.

Mindezeket figyelembe véve e korpusznyelvészeti kutatás legfontosabb célkitűzése, hogy a szepességi falvak lakóinak beszédét kortól, végzettségtől és egyéb tényezőktől függetlenül archiváljuk – és csak ezt követi az autentikus nyelvhasználókat előnyben részesítő szempont. Ez a megközelítés eltér a hagyományos dialektológiai kutatásoktól, amelyek a nyelvjárás legidősebb rétegének rögzítésére törekszenek, figyelmen kívül hagyva a fiatalabb, illetve a magasabb iskolai végzettséggel rendelkező adatközlőket.[8] A szepességi nyelvjárás rögzítésének vonatkozásában ezen kívül lényeges szempont, hogy az adatközlők kötetlenül beszéljenek a felvétel ideje alatt azon a nyelven, amelyet a mindennapjaikban használnak. Ezen elsősorban a nyelvjárást, illetve a regionális köznyelvet értjük.

Az alanyok kiválasztását ily módon két szempont határozta meg: egyszerre kellett biztosítani a felmérés reprezentatív voltát, valamint dokumentálni a szepességi nyelvjárás jellegzetességét. Tisztában vagyunk vele, hogy e két kritérium bizonyos mértékben kölcsönösen kizárja egymást, mindazonáltal igyekeztünk olyan kompromisszumot találni, amely segítségével a korpusz mindkét szempontból kutathatóvá válik. Így tehát, bár a felvételeken minden generáció képviseltette magát, mégsem tekinthető arányosnak a korcsoportok eloszlása. A falusi lakosság nyelvének archiválása esetében csupán törekedhetünk e kompromisszum megtalálására, amelyet valóban elérni csaknem lehetetlen (vö. *1. táblázat*).

1. táblázat. A korcsoportok megoszlása a tokenek számában.

Korcsoport	40 év alatt	40–59 év	59 év felett
Tokenek száma	281 632 (14,4%)	510 337 (26%)	1 168 900 (59,6%)

Érdemes kiemelni, hogy jelen korpusz a 2015 és 2018 közötti nyelvállapotot dokumentálja. Nem tartalmaz korábbi szövegeket, még ha léteznek is ilyen felvételek, és az is

[8] AJPP: Mieczysław Małecki i Kazimierz Nitsch, *Atlas językowy polskiego Podkarpacia. Cz. I: Mapy. Cz. II: Wstęp, objaśnienia, wykazy wyrazów* (Kraków: Polskiej Akademji Umiejętności, 1934), 18.

valószínű, hogy ez a kiterjedt dokumentációs munka később sem fog megismétlődni. E korpusz tehát szigorúan véve szinkrón természetű, a nyelvjárásnak nem múltbeli állapotát, hanem fejlődésének aktuális fázisát tárgyalja.

4. A kutatómunka állomásai

A nem sztenderd beszélt nyelvi korpusz feldolgozásának folyamata a következő állomásokra osztható: adatgyűjtés (az adatközlőkkel folytatott beszélgetések rögzítése és archiválása), adatfeldolgozás (a lejegyzés elkészítése, a morfoszintaktikai annotáció hozzáadása) és végül az eredmények rögzítése egy korpuszkeresővel ellátott adatbázis formájában.

4.1. Terepmunka

A nyelvjárási anyag rögzítését kutatók egy csoportja végezte el a terepmunka során, amely a lengyelországi Szepesség mind a 15 települését érintette. Ez a feladat semmiben nem különbözik a nyelvjárásgyűjtő kutatók hagyományos munkájától. A felvételeket az adatközlők beleegyezésével rögzítették; magától értetődik tehát, hogy a felvett párbeszédek mellőzik a spontaneitást; épp ellenkezőleg, a felvett szövegek jobbára narratív jellegű megszólalások, az alanyok közötti legcsekélyebb interakcióval. Az adatközlőktől írásbeli hozzájárulást kértünk a kutatásban való részvételhez, a felvételek felhasználásához és a korpuszban való elhelyezésükhöz. A hangfelvételeket Olympus LS-12 és Olympus LS-14 típusú diktafonok segítségével WAV-formátumban rögzítették. Tudatosan mellőzzük az MP3-formátum használatát, amely lényegesen kisebb fájlméretben menthető, éppen ezért nem használható fonetikai kutatás céljaira.

A 2015 és 2018 közötti terepmunka folyamán mintegy 600 szepességi lakossal rögzítettünk felvételt, ez összesen nagyjából 400 órányi hanganyagot tesz ki. Ebből 250 órányit szántunk lejegyzésre. A legtöbb beszélgetést az 1940-es években született személyekkel vettük fel (137 adatközlő). A legidősebb felvételi alanyunk egy 1915-ben, Frydman faluban született nő, a legfiatalabb egy 2008-ban született tanuló volt, Nowa Biała községből. Az adatközlők átlagéletkora 58 év (medián: 61, szórás: 22,3).

Az adatfelhalmozás során a társadalom nyelvre gyakorolt hatását vizsgáló szocilingvisztikai módszer került a kutatás homlokterébe.[9] Ezt a megközelítést a falusiak nyelvhasználatában megfigyelhető nagy fokú eltérések alapozzák meg,[10] amelyek jelen esetben olyan tényezők mentén alakulnak, mint például az életkor, a nem, a végzettség, a származás vagy a hosszabb életvitelszerű tartózkodás a falu határain kívül. Ezért

[9] Wladyslaw Lubaś, *Społeczne uwarunkowania współczesnej polszczyzny*: *Szkice socjolingwistyczne* (Kraków: Wydawn. Literackie, 1979). Illetve: Bogusław Dunaj, „Dialektologia a socjolingwistyka," *Acta Universitatis Lodziensis. Folia Linguistica* 12 (1986): 15–23.

[10] Helena Grochola-Szczepanek, „Badanie języka mieszkańców wsi w kontekście przemian społecznych," *Socjolingwistyka* 27 (2013): 43–53; Bogusław Wyderka, „Problemy teoretyczne współczesnej dialektologii," in Maciej Rak i Kazimierz Sikora red., *Badania dialektologiczne: Stan, perspektywy, metodologia, Biblioteka LingVariów 17,* 13–21 (Kraków: Księgarnia Akademicka, 2014).

 DOI 10.31400/dh-hun.2021.5.3148

minden adatközlő a felvétellel egy időben kitöltött egy szociológiai kérdőívet is, amely a nyelvhasználatot befolyásoló lehetséges tényezőket hivatott feltárni.[11]

4.2. A lejegyzés

Egy nem sztenderd nyelvváltozat annotációja során a köznyelvtől eltérő nyelvi rendszerrel kell szembenézni, azaz a sztenderd nyelvváltozatban ismeretlen lexémák előfordulásával, valamint fonetikai vagy morfológiai változatokkal. A lejegyzés módját az a tényező is meghatározta, hogy e szövegek egy nyelvi korpusz létrehozása céljából készülnek. Tehát a hangfelvételek annotációját a korpusz összeállíthatóságának igényeihez kell szabni, különböző nyelvi és műszaki szabályoknak megfelelően.[12]

A lejegyzést meghatározó elvek kidolgozása kulcsfontosságú már a korpuszkészítés tervezési folyamatában. A készítőkben felmerülő problémákra a következő három lejegyzési mód adható válaszként:

1. fonetikus lejegyzés (a szlavisztikának megfelelő IPA-variáns);
2. félig fonetikus lejegyzés;
3. a sztenderd lengyel nyelvváltozat helyesírásának követése.

Esetleges negyedik megoldásként egy önálló szepességi helyesírás kialakítása is felmerülhet, ez azonban a korpusz esetében nem jöhetett szóba.

Az első megoldás, érthető módon, több akadályba ütközik. Először is, ez a teljesen fonetikus lejegyzés meglehetősen munkaigényesnek bizonyulna, mivel a lejegyzőknek minden esetben el kellene dönteniük, hogy az adott szóelőfordulás pontosan milyen kiejtésben realizálódott. Ez a döntés ráadásul gyakran vitatható. Ugyanakkor mivel a lejegyzés mellett a hangfelvétel is rendelkezésre áll, a fonetika iránt érdeklődők vagy akár a fonetikus ábécét nem ismerő laikusok is tanulmányozhatják az adatközlők kiejtésének sajátosságait.

Miközben a harmadik pontban említett normalizált lejegyzés hátránya a fonetikus írásmóddal szemben az, hogy közel sem azonosítható a tényleges kiejtéssel vagy annak idealizált változatával. Így például a selypes fogréshangokat <s> vagy <sz> alakban is lejegyezhetjük, annak függvényében, hogy az adott hang hogyan szerepel a köznyelvi szóban – bár az adatközlő kiejtésében ugyanarról a hangról beszélhetünk.

Ezen kívül viszont két tényező is a normalizált lejegyzési mód mellett szól. Először is ily módon megkönnyíthetjük a korpuszban történő keresést, leginkább abban az esetben, ha a hangérték a meghatározó (fonetika, fonológia, morfológia, esetleg ezeknek a szociolingvisztikához kapcsolódó határterületei), hozzásegítve a kutatót ahhoz, hogy a

[11] A terepmunka módszertani leírása egy külön szaktanulmányban kapott helyet. Helena Grochola-Szczepanek, „Nowe badania języka mieszkańców wsi regionu polskiego Spisza," in Błażej Osowski, Paulina Michalska-Górecka, Justyna Kobus i Agnieszka Piotrowska-Wojaczyk, red., *Język w regionie, region w języku 2,* 103–119 (Poznań: Wydawnictwo „Poznańskie Studia Polonistyczne", 2017).

[12] A lejegyzés irányelveinek kialakításáról külön szaktanulmányban esik szó: Helena Grochola-Szczepanek i Michal Woźniak, „Transkrypcja języka mieszkańców wsi w aplikacji ELAN w Korpusie Spiskim," Renata Przybylska, Maciej Rak i Agata Kwaśnicka-Janowicz, red., *Historia języka, dialektologia i onomastyka w nowych kontekstach interpretacyjnych,* 267–278 (Kraków: Wydawnictwo Uniwersytetu Jagiellonskiego, 2018).

normalizált fonetikus lejegyzés segítségével találja meg a megfelelő hangfájlt. Másodszor, a normalizált lejegyzés eszközéül szolgálhatnak a lengyel köznyelv morfológiai jelölései is (vö. 4.3. alfejezet).

A fenti okokból döntöttünk a harmadik, és bizonyos esetekben az első megoldás mellett (ekkor mindkettőt alkalmaztuk). Azért nem használtunk félig fonetikus lejegyzést, mert az a módszer egyesíti magában a két bemutatott megoldás hátrányait, ellenben nem jár együtt azok előnyeivel. Egyrészt e megközelítés nem teszi lehetővé a már meglévő nyelvelemzési eszközök használatát, és nem könnyíti meg a keresést, hiszen tükrözi a kiejtésbeli következetlenségeket, másrészt – ahogy már a neve is sugallja – csupán megközelíti, sőt meglehetősen gyengén adja vissza a tényleges kiejtést.

A morfológia esetében hasonló helyzettel állunk szemben. Azok a morfémák, amelyek csak alakjukban különböznek a sztenderd nyelvváltozatban használatostól, a nyelvi sztenderd készletével jelölhetők. Más szóval, ha a nyelvjárási szó morfémája visszavezethető a sztenderd nyelvváltozatban szereplő morfémára, akkor a sztenderd változat jelenik meg a lejegyzésben. Ha azonban a szepességi morféma a sztenderdben használthoz képest más morfémára vezethető vissza, akkor a szepességi változatot visszaadó írásmódhoz folyamodunk. Például a *chodzić* 'jár' ige egyes szám első személyű múlt idejű alakját *chodziyłef* alakban, nem pedig *chidziłem* alakban tüntetjük fel, hiszen az első az *-ech* régies múlt idejű igevégződésre vezethető vissza, míg a második, az *-em,* a ma használatos múlt idejű személyrag továbbélése. A *biała* 'fehér' melléknév ugyanakkor a sztenderd alakban szerepel, mivel a köznyelvi és a szepességi változat is ugyanarra a morfémára vezethető vissza; az eltérésre rendszerszerű hangváltozás ad magyarázatot.

Tehát négy különböző esettel állunk szemben, amelyek mindegyike a lengyel köznyelvhez való közeledés különböző fokozatának tekintendő, ennélfogva különbözőképp írhatók le:

1. A köznyelvnek megfelelő vagy szabályszerűen módosult alakváltozatok (például a zártabb magánhangzók) lejegyzése a sztenderd szerint történt.
2. Morfológiailag eltérő szavak: a lengyel köznyelvhez közvetlenül kapcsolódó, de egy-egy morfémában vagy ragozott alakban eltérő nyelvi egységeket a lejegyzésben mindkét – sztenderd és nyelvjárási – változatban feltüntetjük // jellel elválasztva.
3. Azok az alakváltozatok, amelyek megfelelői a lengyel köznyelvben megtalálhatók, de szemantikájukban eltérnek attól, a sztenderd szerint szerepelnek a lejegyzésben ^ szimbólummal jelölve.
4. Azok a lexémák, amelyek a lengyel köznyelvben nem, kizárólag a nyelvjárásban jelennek meg, hangzás utáni fonetikus alakban szerepelnek # szimbólummal ellátva. A kutatás későbbi munkafázisaiban ezek egységesítését is elvégezzük.

 DOI 10.31400/dh-hun.2021.5.3148

2. táblázat. A lejegyzés és a szóosztályok lejegyzésben szereplő annotációjának példái, valamint ezek előfordulásának száma a korpuszban.

Szóosztályok száma	Példák (hangzás utáni alakban)	A lejegyzésben szereplő adat	Annotáció	Tokenek száma
1	*mlyko* 'tej'	mleko	nincs	1 844 353
2	*dałak* 'adtam'	dałam//dałak	//	116 516
3	*ślafrok* 'köpeny, köntös'	szlafrok	^	35 086
4	*odziywacka* 'szabó'	*odziywacka* (végleges formában: *odziewaczka*// *odziywacka*)	#	70 812

A 2-es és 4-es számú osztályokba tartozó szavakat tehát egyszerre két alakban, a köznyelvi és a nyelvjárási változatban is feltüntettük, az érvényes helyesírási szabályoknak megfelelően. A köznyelvi változat annotációja mesterséges, míg a nyelvjárási változaté megközelítőleg tükrözi a szepességi nyelvjárási kiejtés sajátosságait. Mindkét változat feltüntetése lehetővé teszi, hogy a sztenderd és a nyelvjárási alakváltozatot egyszerre lássuk. A lejegyzésben jelöltünk egyéb nyelvjárási változatokat is: elkülöníthetők több szóból álló kifejezések (*młodzi panowie* a *państwo młodzi* vagy a *nowożeńcy* alakok helyett ['friss házasok, ifjú pár']), a sztenderdtől eltérő szószerkezetek (*ku moście* ['a hídhoz'] a *do mostu* vagy a *w stronę mostu* alakok helyett) és jövevényszavak<*pridi*> (a *przyjdź* ['jön'] alak helyett – szlovák hatás), <*dawaj sało*> (a *słonina* szó ['szalonna'] helyett – orosz hatás).

Az összes hangfelvétel visszajátszása és a lejegyzések elkészítése a korpuszkészítés folyamatának legkimerítőbb része. Egyórányi hanganyag átdolgozása, dokumentációja és a lejegyzés nagyjából 40 munkaórányi feladat. A hangfelvétel és a lejegyzés összevetését néhány munkatárs több ízben elvégezte. Kiemelendő továbbá, hogy a tisztázat elkészítése nem esik egybe a lejegyzés befejezésével: a hallás utáni leírás során számos hiba merülhet fel, ezért lehetőséget kell biztosítani azok későbbi módosítására.

4.3. A morfoszintaktikai annotáció

Az elkészült lejegyzéseket XML-formátumba kódolva mentjük el a felvételekhez külön fájlként hozzárendelve. Ekkor zárul le a fájlok kézi szerkesztése, innen automatikus rendszerek veszik át a munkát. Az első ilyen munkafolyamat a morfoszintaktikai annotáció elvégzése, tehát a tokenek szótári alakjának (lemma) és nyelvtani sajátosságainak azonosítása. Irányadóként az NKJP egységes jelölőelem-készletét (a nyelvtani kategóriák és a hozzájuk tartozó jelölések címkézőrendszerét) határoztuk meg. Például az *akordeonie* ('harmonika') token a következő annotációt kapja subst:sg:loc:m3, ahol a kettősponttal tagolt egységek megfeleltethetők a szófajnak, a számnak, a nyelvtani esetnek és a nemnek.

Az automatikus nyelvi elemzést két lépésben végeztük. A köznyelvi szavak elemzésére a *Pantera* programot használtuk. Ezek közé tartoznak a sztenderddel megegyező

szófajú és homonim jelentésű nyelvjárási szavak is (lásd a 4.2. alfejezetben tárgyalt osztályozás hármas számú osztályát). Azoknak a szavaknak az elemzései pedig, amelyeket a köznyelvre adaptált szoftver nem képes kezelni, a *Kuźnia* nevű program kiegészítő adatbázisán alapulnak.

4.4. A korpusz elkészítése

A következő lépés a szövegek és a hangfájlok feldolgozása és átalakítása egy konkordanciaprogram segítségével. Ez a fázis ugyancsak teljesen automatikus, és a következő szakaszra tagolható:

1. a hangfájlok feldarabolása a lejegyzésének megfelelően;
2. a szöveges fájlok átalakítása a CWB[13] programnak megfelelő formátumra;
3. a metaadatok hozzárendelése a fájlokhoz;
4. a személyes adatok anonimizálása;
5. magyarázatok hozzárendelése a sztenderdből hiányzó szavakhoz;
6. az adatok mentése a CWB-adatbázis formátumában;
7. az adatbázis csatlakoztatása a webes felülethez.

5. Szabványos eszközök

Egy beszélt nyelvi korpusz többlépcsős munkával állítható össze, minden egyes lépés során jelentős erőfeszítésre és gondosságra van szükség, hogy megőrződjön az adatok konzisztenciája. A rendelkezésre álló eszközök nagymértékben felgyorsították és megkönnyítették a munkafolyamatot. A *Szepességi korpusz* az alábbi eszközök felhasználásával valósulhatott meg:

> 1. ELAN – a Max Planck Pszicholingvisztikai Intézet multimédiás források annotációjára használható programja.[14] A programot széles körben alkalmazzák beszéd rögzítésére, adatok feldolgozására és archiválására. Az ELAN a hangfelvételeket feldolgozók és lejegyzők számára létrehozott speciális munkakörnyezet. Lehetővé teszi az adatközlések szegmentálását és többszintű címkézését. Az ELAN a lejegyzést XML-formátumban rögzíti, amely a legnépszerűbb és legszélesebb körben használt formátum az efféle kutatásokban, ezzel egy időben a hangfelvételeket WAV-formátumba írja át, amely a hangfelvételek rögzítésének egyik legnépszerűbb módja.
>
> 2. *Pantera* nyelvi elemző. A morfoszintaktikai annotációra alkalmazott szoftver igényei döntő szerepet játszottak abban, hogy az adatok sztenderd nyelvváltozatban történő lejegyzése mellett döntsünk. Ez az alkalmazás

[13] The IMS Open Corpus Workbench, hozzáférés: 2021.09.22, `http://cwb.sourceforge.net`.

[14] Hennie Brugman and Albert Russel, „Annotating Multimedia/Multi-modal Resources with ELAN," in Maria Teresa Lino, Maria Francisca Xavier, Fátima Ferreira, Rute Costa and Raquel Silva, eds., *Proceedings of Fourth International Conference on Language Resources and Evaluation (LREC'4),* 2065–2068 (Lisbon: European Language Resources Association [ELRA], 2004), `http://www.lrec-conf.org/proceedings/lrec2004/pdf/480.pdf`.

 DOI 10.31400/dh-hun.2021.5.3148

lehetővé teszi, hogy egy adott szó összes előfordulására rákereshessünk, alakváltozattól függetlenül. Ugyanígy elérhető bármely szó a ragozott alakja alapján is. A lengyel fejlesztésű elemzők kizárólag a lengyel köznyelvre adaptálva működnek, ezért nem használhatók olyan szavak esetében, amelyek nem felelnek meg egyetlen lengyel köznyelvi szó szótári alakjának sem. A szepességi kutatási projektben azért használtuk a *Panterá*t, mert kimagaslóan pontos annotációt tesz lehetővé.

3. A Lengyel Tudományos Akadémia Számítástechnikai Kutatóintézet (IPI PAN) által lengyel nyelvű ragozási szótárak összeállítására kifejlesztett *Kuźnia* program. Mivel a meglévő elemzők nem boldogulnak a nyelvjárás lexikai elemeivel, azok morfoszintaktikai annotációját kézzel, a *Kuźnia* szoftver segítségével végeztük el. Ez lehetővé teszi ragozási paradigmák, szótári alakok, illetve (a sztenderdből hiányzó szavak szótárához használt) magyarázatok hozzárendelését a lengyel köznyelvben ismeretlen lexémákhoz.

A rendelkezésre álló digitális eszközök használata számos jól látható előnnyel bír: felgyorsítja a munkát, és lehetővé teszi a feldolgozott adatok következetes kezelését, valamint olyan normatív szabályozások eszközlését (például XML-formátum, az NKJP annotáló rendszerének alkalmazása), amelyek növelik az adatok egységességét, ezáltal könnyítve azok felhasználhatóságát más hasonló jellegű kutatások során, illetve biztosítva azok nagyobb stabilitását a hosszú távú adatmegőrzés számára.[15] Mindazonáltal meg kell jegyeznünk, hogy a szabványos eszközök használata nem sztenderd nyelvváltozatok feldolgozása során szükségszerű módosításokat követel meg. Ezen módosítások a következő két csoportra bonthatók:

a) Ami az adatok sokféleségét illeti: ahogy fent említettük, ahhoz, hogy a nyelvi elemző szoftver használatát lehetővé tegyük, a lejegyzés során normalizálnunk kellett a szavak helyesírását.

b) Az eszközöket illetően: a *Kuźnia* a nyelvi sztenderd ragozási rendszerének kezelésére készült. Mivel a nyelvjárás több helyen eltér ettől, elengedhetetlenné vált bizonyos módosítások bevezetése, amelyek lehetővé tették a kizárólag nyelvjárási lexémák paradigmáinak kezelését. Minthogy ezek a szavak két (normalizált és nyelvjárási) alakváltozatban is szerepelnek a korpuszban, indokolttá vált a *Kuźnia* szoftver oly jellegű bővítése, amely megengedi, hogy egy adott lexéma két változatát is tárolja az adatbázisban. Ezeket a módosításokat az tette lehetővé, hogy a *Kuźnia* nyílt forráskódú alkalmazás, amelynek forráskódja BSD-licenc alatt érhető el.

6. Hatás

A kutatási projekt elsődleges eredményének a beszélt nyelvi szövegekből álló korpusz létrejötte tekinthető, amely a `https://spisz.ijp.pan.pl` címen érhető el az

[15] Ruprecht von Waldenfels and Michał Woźniak, „SpoCo – A Simple and Adaptable Web Interface for Dialect Corpora," *Journal for Language Technology and Computational Linguistics* 31 (2016): 155–170.

interneten. Ehhez kapcsolódik a korpusz igényeit kiszolgáló online felület is, amely a korábbi Spoco-projekt módosított verziójaként jött létre.

A korpusz lehetővé teszi a szövegegységek szerinti keresést, amelyek a hangfelvétel egy részletéből és a lejegyzésükből állnak. Egy CWB-alapú keresőről van szó, amelyet széles körben használnak korpuszkészítésre, mivel ez a keretrendszer engedi az összetett keresést, valamint lehetőséget teremt a kvantitatív és a kvalitatív elemzés számára is. A korpuszban a CQL formális lekérdezőnyelv segítségével lehet keresni, ez nagy fokú szabadságot biztosít a felhasználónak, ugyanakkor a lekérdezőnyelv ismeretét is feltételezi. A felhasználóbarát kialakítás érdekében ugyanakkor a lekérdezőnyelv használata helyett egyszerű szerkezeti egységekből álló kereséseket indíthatunk a felületen – elegendő bevinni a kívánt szót valamelyik keresőmezőbe. A felhasználói felület négyféle keresőmezőt kínál: szóalak *(token)* – az adott szó normalizált alakjának szövegszintű keresését teszi lehetővé; lemma – az adott szótári alakú szó minden alakváltozata kikereshető; nyelvjárási alak – kikeresi a kívánt szó nyelvjárási alakváltozatát; és grammatikai tulajdonság – nyelvtani sajátosságok szerinti szókeresést tesz lehetővé. A lekérdezési eredményeket szűrők segítségével tudjuk korlátozni (szűkíteni lehet: nem, nemzetiség, végzettség, lakhely, születési év és adatközlő szerint).

Az imént tárgyalt modul az egyszerű lekérdezési feltételek szerinti keresést teszi lehetővé. Az összetett keresés (például az összes olyan szó megkeresése, amelynek eltér a köznyelvi és a nyelvjárási alakja) bár jóval nagyobb szabadságot biztosít a felhasználónak, megkívánja a CQL ismeretét. A keresési eredmények a lekérdezési feltételek által meghatározott szövegegységekbe ágyazva jelennek meg. Az összes szegmenshez tartozó hangfájl meghallgatható, ezzel egy időben a vonatkozó normalizált és nyelvjárási szövegrészletet is meg lehet tekinteni. A keresési eredmények kétféleképpen jeleníthetők meg: a szegmensek egyszerűsített megjelenítése, valamint KWIC-listaként (Key Word in Context – kontextusos kulcsszó), amely a szövegegységeket három oszlopra tagolja: a keresett elemet megelőző kontextus, a keresett elem és az azt követő kontextus. Mindkét megjelenítési mód lehetővé teszi a keresési eredmények csoportosítását, a szegmensekhez rendelt metaadatok, valamint a szélesebb (hét szövegegységből álló) kontextus megtekintését. Ezenfelül a keresési eredmények (a lejegyzések és a hangértékek egyaránt) letölthetők.

A szövegkorpuszhoz és a hangfelvételek gyűjteményéhez továbbá egy szótár is kapcsolódik. E szótár olyan szavakból és kifejezésekből áll, amelyek a hangfelvételen szerepelnek, viszont a lengyel köznyelv számára ismeretlenek, illetve amelyek a nyelvjárásban más jelentéshez köthetők, mint a sztenderd nyelvváltozatban. A valódi tájszók a szócikkekben köznyelviesített vagy nyelvjárási alakváltozatban szerepelnek. A jelentésbeli tájszók szócikkei a sztenderd és a nyelvjárási alakváltozatot is mutatják. A szótár továbbá olyan kifejezéseket is tartalmaz, amelyek szerepelnek lengyel nyelvű szótárakban, ám régies vagy nyelvjárási szónak számítanak.

7. Alkalmazás

A korpusznyelvészet mindenekelőtt olyan módszertan, amely számos lehetőséget kínál, ugyanakkor korlátokat is állít a kutatók elé. Az egyik ilyen korlátot a hiányzó adatok jelentik: az a tény, hogy valamely kifejezés nem szerepel a korpuszban, nem

 DOI 10.31400/dh-hun.2021.5.3148

jelenti egyértelműen azt, hogy ez a szó vagy alak ne képezné a nyelvhasználat részét. A korpusz összeállítói gyakorta találkoztak ezzel a problémával a ragozott alakok kapcsán, hiszen néhány szó egészen sajátos paradigmával rendelkezik, amelynek nem minden eleme jelenik meg az adatközlők szövegeiben. Az ilyen alakok korpuszba való felvételéhez a kutató nyelvi kompetenciájára kell hagyatkoznunk.

7.1. Szociolingvisztika és nyelvföldrajz

A lekérdezések eredményeinek szociológiai adatokkal történő szűrhetősége lehetővé teszi az adatközlők bármely csoportjának részletes elemzését. Lekérdezhetjük például a *dom* 'ház' szó összes előfordulását azokban a közlésekben, amelyeket 1950 előtt született, Niedzica (Nedec) községben élő nőkkel vettek fel. Az ilyen jellegű szűrések lehetővé teszik az olyan demográfiai és társadalmi tényezők nyelvre gyakorolt hatásának korpuszalapú vizsgálatát, mint az életkor vagy a nem kategóriája.[16]

7.2. Nyelvtan : ragozás és szintaxis

Ahogy azt már fentebb említettük, a korpusz elemeit morfológiai annotációval láttuk el. Ez a jelölőrendszer egy sor nyelvtani kutatás elvégzését biztosítja ragozás, szóalkotás és szintaxis viszonylatában egyaránt. Meg kell jegyeznünk azonban, hogy az annotáció az adott szóhoz tartozó nyelvtani kategóriákra vonatkozik, és nem a külön morfémákra ; morfémák keresésére karakterszekvenciák[17] megadásával van lehetőség, amely csupán megközelítőleges keresési eredményt biztosít. Hasonló helyzet áll fenn a megadott ragozási jellemzők szerinti szerkezetekre irányuló keresés esetében is.[18] Ekkora méretű korpusz esetén ugyanakkor az adott nyelvtani jelenségek előfordulásának száma már elegendő ahhoz, hogy mennyiségi mérést végezhessünk, és elkülöníthessük egymástól a marginális és a tipikus eseteket.

7.3. Pragmatika

Meg kell jegyeznünk, hogy a *Szepességi korpusz* beszélt nyelvi szövegek olyan gyűjteménye, amely méretében megközelíti az NKJP társalgási alkorpuszát. Az adatbázisban történő keresés lehetővé teszi a szélesebb kontextus vizsgálatát is, ami a pragmatika területén végzett kutatómunka alapjául is szolgálhat. A szélesebb szövegkörnyezet vizsgálatának lehetősége elengedhetetlen a téma-réma szerkezetek tanulmányozása esetén, illetve akkor, amikor e szerkezetek szintaxisra gyakorolt hatását kívánjuk elemezni. (Miközben ez a jelenség még a sztenderd nyelvváltozat esetében sem gyakran kerül az elemzések középpontjába.) További feladat, amelynek megoldására a *Szepességi korpusz* alkalmas lehet, a diskurzusjelölők korpuszalapú kutatása.

[16] Külön megjelenő cikkben tárgyaljuk azt, miként befolyásolják a metaadatokba felvett jellemzők az adatközlők kódját.

[17] A kicsinyítő képzős szavakat például *-eczek, -eczka, -eczko* végződéssel lehet lekérdezni [lemma=".+*ecz*(*ek*|*ka*|*ko*)"].

[18] A folyamatos jövő idő a *być* (lét)ige és a főnévi igenév vagy az ige múlt idejű alakjának egymást követő sorrendjével fejezhető ki. Ez a séma azonban nem alkalmazható az olyan szekvenciák esetén, mint például *będzie szybko szedł* [a *być* ige ragozott alakja és az *iść* – 'megy' – ige múlt idejű alakja között határozószó – *szybko* – áll, jelentése : 'ő (hímnem) gyorsan fog menni' – *a ford.*].

7.4. Fonetika

A hanganyag kiváló alapul szolgálhat fonetikai és prozódiai kutatásokhoz. Éppen ezért már a kutatási projekt kezdeti szakaszában meghoztuk a döntést: a hangfelvételeket veszteségmentes formátumban fogjuk rögzíteni, nem pedig a legnépszerűbb MP3-formátumban, amelynek vitathatatlan előnye, hogy kisebb fájlméretben menthető – ugyanakkor a veszteségmentes formátum jóval alkalmasabb a kutatásokra.

Az adatközlések részleteit (általánosan 15 másodperc) le lehet tölteni és fel lehet dolgozni fonetikai elemzőprogramok (például *Praat*) segítségével. Mindazonáltal nem hallgathatjuk el, hogy nem minden felvétel minősége felel meg fonetikai kutatás igényeinek. Elsődleges feladatunknak azt tekintettük, hogy minél nagyobb mennyiségű adatot gyűjtsünk; a munkatársak nem töröltek egy felvételt sem a gyenge hangminőség miatt.

Megemlítendő még, hogy a keresőrendszer lehetővé teszi például a két zárhang között álló magánhangzók kigyűjtését is. A metaadatok segítségével pedig tovább szűkíthető a lekérdezett eredmények listája nem, életkor vagy lakóhely szerint. A hangfelvételek ösztönözhetik a prozódia területén végzett kutatásokat is, amelyek a szintaxis vizsgálatának fontos elemét képezik.

7.5. Szókészlet

Bár egy ilyen méretű korpusz nem meríti ki az alapos lexikai kutatás igényeit, a mintegy 10 000 különböző szót[19] számláló gyűjtemény mégis tekintélyesnek nevezhető. Igaz, a szókészlet gazdagsága függ a felvételeken hallható beszédtémáktól is, amelyek gyakran érintik a hagyományok, szokások, gyermekjátékok, mondókák és a mezőgazdasági munka stb. területeit. Ezen felül a korpusz lehetőséget biztosít a szókapcsolatok, többszavas kifejezések megfigyelésére, továbbá konkordanciaszótár elkészítésére.

Mindehhez hozzátehetjük, hogy a korpusz alkalmas lehet a szepességiek kultúráját és szokásait illető ismereteink bővítésére is.

8. Összegzés

A fent leírt problémák egyike sem oldható meg nagy méretű, digitalizált korpusz létrehozása nélkül. Nem helyettesítheti sem szótár, sem szöveglejegyzések, sem hangfelvételek puszta gyűjteménye.

Hangsúlyoznunk kell, hogy a leírt kutatási projekt érdeme a feldolgozott anyag mennyiségében keresendő. Egy kis méretű korpusz még viszonylag gyakori jelenségekre is csak kevés előfordulást tartalmaz, így csupán meglehetősen korlátozott kutatásokat tenne lehetővé. A lengyel dialektológiának a *Szepességi korpusz* révén olyan eszköze jött létre, amely bár pontszerű, mégis lehetővé teszi a Kis-Lengyelország déli részén található nyelvjárások egyikének alapos és sokrétű tanulmányozását.

Fordította: Bali Farkas Péter

[19] Összehasonlításként az *Árva mente nyelvjárásának szótára* (KąśSGO) megközelítőleg 28 000 szócikkből áll, beleértve a sztenderd nyelvvel közös szavakat is.

 DOI 10.31400/dh-hun.2021.5.3148

A Spoken Corpus of Inhabitants of Polish Spisz

The article describes a dialect corpus project that documents the dialect of Polish Spisz. In contrast to the majority of dialectological research in Poland, our corpus also includes the speech of the youngest and middle generations, as its aim is also to document the sociolinguistic situation of the dialect of the region. Recordings have been transcribed into standard Polish orthography, not phonetically, which makes it possible not only to easily search the corpus but also to use existing tools to lemmatize and add morphosyntactic annotation to the texts. Users interested in the phonetic layer can access the recordings on a per-utterance basis. The article describes the stages of compiling the corpus and discusses its potential applications. The authors argue that a large corpus which covers a small, homogeneous area is a more valuable resource for dialectologists than a series of small corpora documenting a larger region.

Keywords:
corpus, spoken language, dialectology, Spisz dialect