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>