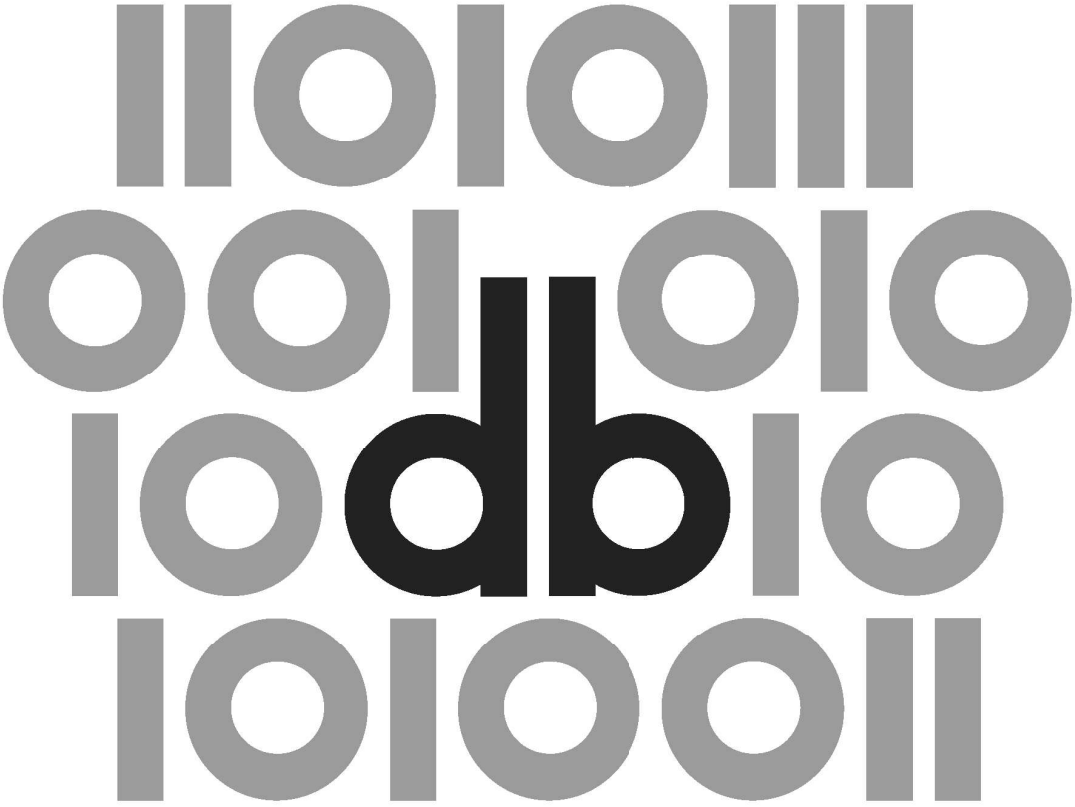
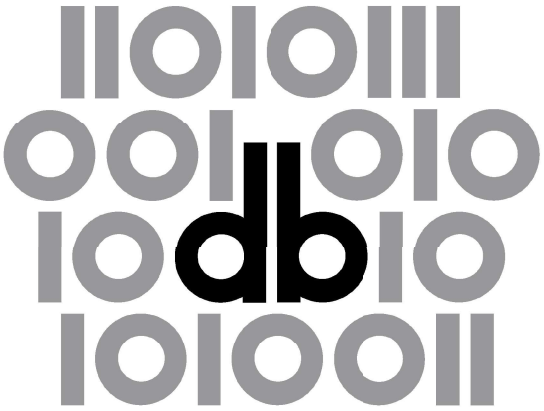
4 (2021) <DIGITÁLIS BÖLCSÉSZET>



4 (2021) </DIGITÁLIS BÖLCSÉSZET>

Digitális Bölcsészet
2021., negyedik szám

<DIGITÁLIS BÖLCSÉSZET>



4 (2021)

Felelős szerkesztő:
Maróthy Szilvia

Szerkesztőség:
Kokas Károly, Parádi Andrea

Rovatvezetők:
Tanulmányok: Kiss Margit
Műhely: Péter Róbert
Kritika: Almási Zsolt
Labor: Maróthy Szilvia

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.2021.4

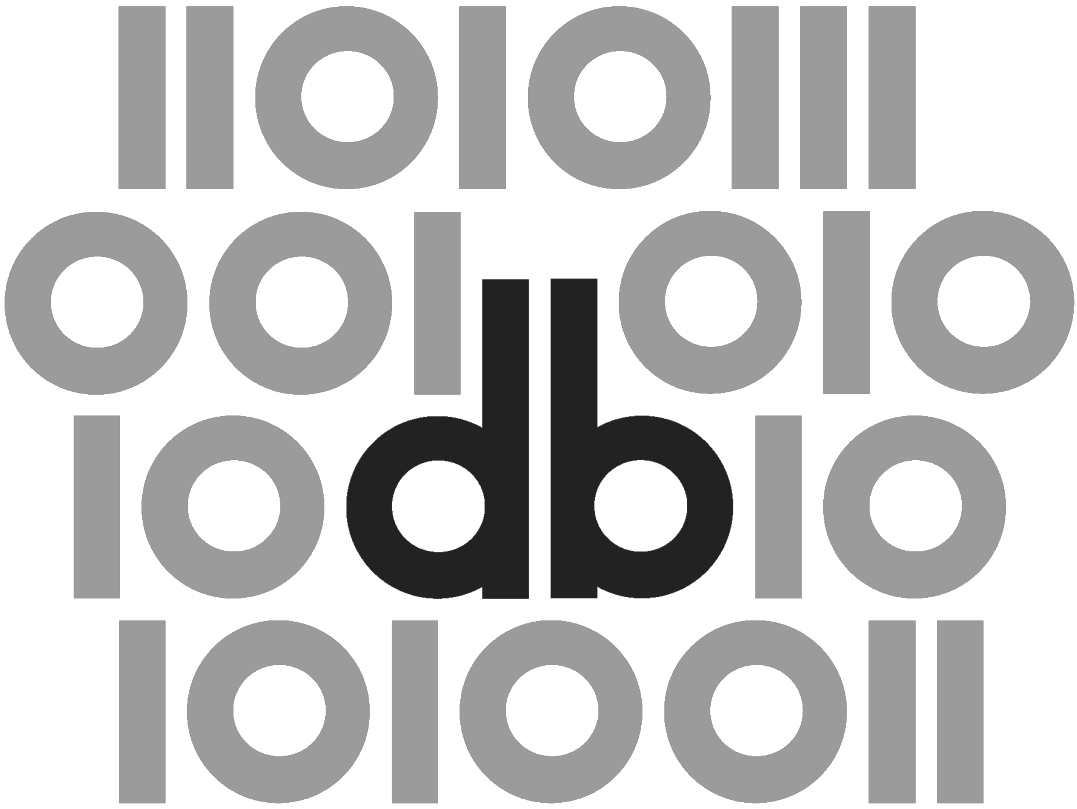
Kiadja a Bakonyi Géza Alapítvány és az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A).
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

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

Ez a mű a Creative Commons *Nevezd meg! – Ne add el! – Így add tovább! 2.5 Magyarország Licenc* (`http://creativecommons.org/licenses/by-nc-sa/2.5/hu/`) feltételeinek megfelelően felhasználható.

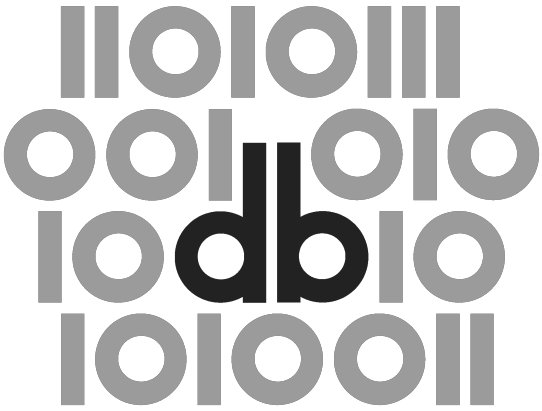
Honlap: `http://ojs.elte.hu/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

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



Digitális Bölcsészet
2021., negyedik szám

<DIGITÁLIS BÖLCSÉSZET>



4 (2021)

Felelős szerkesztő:
Maróthy Szilvia

Szerkesztőség:
Kokas Károly, Parádi Andrea

Rovatvezetők:
Tanulmányok: Kiss Margit
Műhely: Péter Róbert
Kritika: Almási Zsolt
Labor: Maróthy Szilvia

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.2021.4

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.

(cc) BY-NC-SA

Ez a mű a Creative Commons *Nevezd meg! – Ne add el! – Így add tovább! 2.5 Magyarország Licenc* (`http://creativecommons.org/licenses/by-nc-sa/2.5/hu/`) feltételeinek megfelelően felhasználható.

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

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

<TANULMÁNYOK>

Ted Underwood

University of Illinois Urbana-Champaign,

College of Liberal Arts & Sciences,

Department of English

tunder@illinois.edu

A műfajok életciklusai*

A kutatók még a „műfaj" jellegét illetően sem értenek egyet. Kérdéses például, hogy szöveges mivoltuk vagy társadalmi recepciójuk határozza meg inkább a műfajokat? Állandó kategóriáknak, évszázadokon átívelő történelmi konstrukcióknak vagy éppen rövid távú generációs jelenségeknek számítanak? A tanulmány e véget nem érő vitákhoz kapcsolódik újszerű kvantitatív módszerekkel. Kimutatja, hogy a műfajok társadalmi és szövegalapú definíciói jól összehangolhatók: az olvasó ítélete alapján meghatározott műfajok a szöveg bizonyos jegyei alapján is felismerhetők. Erre az eredményre építve a tanulmány a műfajok szokásos élettartamával kapcsolatos ellentmondásokat vizsgálja. Kiderül, hogy a különféle műfaji modellek gyakran még akkor is összeegyeztethetők egymással, ha definíciójuk különböző korok olvasói ítélete alapján született. Így például azok a modellek, amelyek felismerik a tizenkilencedik századi „detektívtörténeteket" és „tudományos románcokat", a huszadik századi „krimit" és „sci-fit" is felismerik. A bizonyítékok tehát arra utalnak, hogy – bár a műfajok bizonytalan történelmi konstrukciók – egy részük jóval maradandóbb, mint ahogy azt az irodalomtörténészek feltételezik. Ám ez nem minden esetben igaz. A gótikus irodalom például éppen olyan laza és változékony, mint amilyennek a kritikusok eddig is tekintették.

Kulcsszavak:
gépi tanulás, műfaj, irodalom, irodalomtörténet

A műfaj fogalma szinte egyidős az irodalomelmélettel, de az évszázadok óta húzódó viták sem tudtak túl sok egyetértést szülni a témában.[1] Ennek részben az az oka, hogy a műfaj egy adott szöveg életének különböző pontjain más és más dolognak látszik.

* William E. Underwood, „The Life Cycles of Genres," *Journal of Cultural Analytics* 1, 1. sz. (2017), https://www.doi.org/10.22148/16.005.

[1] A szerző köszönetét fejezi ki a kanadai Social Sciences and Humanities Research Council által finanszírozott NovelTM projekt támogatásáért. A projektet szintén támogatta a University of Chicago Knowledge Lab és a University of Illinois, Urbana-Champaign Center for Advanced Study. A HathiTrust Digital Library feltárásában a HathiTrust Research Center nyújtott segítséget, és hasznos tanácsokat kaptam a Novel™ csapat minden tagjától és Jim Englishtől.

A retorika tudósai általában a kommunikatív cselekvés olyan mintázataira összpontosítanak, amelyekből a feljegyzések vagy a tragédiák születnek.[2] A szociológusokat gyakran inkább a befogadást szervező intézmények érdeklik.[3] Az irodalomtudósok pedig hagyományosan maguknak a szövegeknek a mintaalkotási jellemzőivel vannak elfoglalva. A műfaj mindezen szempontjai természetesen kapcsolódnak egymáshoz. De a kapcsolatokat nem könnyű leírni.

A távoli olvasás látszólag tükrözi az irodalomtudósok mániáját, mely szerint a műfaj a szöveges műalkotások egyik alapvető jellemzője. De a kvantitatív módszerek valódi értéke abban rejlik, hogy segítségükkel a tudósok összehangolhatják a műfaj szöveges és társadalmi megközelítéseit. A jelen dolgozat egy ilyen típusú kísérleti kapcsolat felvázolására vállalkozik. Az alábbiakban a műfajt először a befogadás történetével kapcsolatos kérdésként közelítjük meg – összegyűjtve az egyes olvasók vagy intézmények által bizonyos történelmi pillanatokban csoportosított címek listáját. E címek mögé tekintve azonban magukat a szövegeket is megvizsgáljuk. A statisztikai modellezés mai gyakorlatának köszönhetően párbeszédet kezdeményezhetünk különböző szövegcsoportok között, így felmérhetjük például, hogy a gótikus* irodalom egymással versengő meghatározásai (amelyek különböző időpontokban születtek, és merőben eltérő könyvlistákat eredményeztek) valóban annyira kompatibilisek voltake egymással, mint azt egyes kritikusok állítják.

A történelmi összehasonlítás problémája azért is különösen sürgető, mert az irodalomtudósok mindeddig nem tudtak konszenzusra jutni a regényszerű műfajok életciklusával kapcsolatban. A gótikus irodalom kategóriája például egyes vélemények szerint 25, más vélemények szerint 250 évet ölel fel. *Graphs, Maps, Trees* című könyvében Franco Moretti átfogó képet nyújtott a műfajt érintő akadémiai stúdiumokról, és arra a következtetésre jutott, hogy a műfajoknál „igen gyakori az őrségváltás […] ilyenkor fél tucat műfaj hirtelen eltűnik a színről, sok új műfaj bejön, majd ezek nagyjából huszonöt évig a helyükön maradnak."[4] A késő tizennyolcadik századból származó gótikus regény 1825 körül átadja a stafétát (például) a Newgate-regénynek, ezt váltja az 1850-es évek végén a szenzációs regény, végül a tizenkilencedik század végére megérkezik a „birodalmi gótika" (pl. *Drakula*), amelyet Moretti diagramja a régebbi szerzeteses-banditás gótikától már minden tekintetben különböző jelenségként kezel. Moretti összefüggést feltételez a sorozat huszonöt éves ritmusa és a generációk váltakozása között.

Ám bizonyos jól megalapozott recepciós hagyományok alapján a műfajok jóval hosszabb időtávon is képesek fenntartani egységes identitásukat, mint ahogy azt a

[2] Amy Devitt, *Writing Genres* (Carbondale: SIU Press, 2004); Carolyn Miller, „Genre as Social Action," *Quarterly Journal of Speech* 70 (1984): 151–167, https://doi.org/10.1080/00335638409383686.

[3] Pierre Bourdieu, *Distinction: A Social Critique of the Judgment of Taste* (Cambridge: Harvard University Press, 1984); Paul DiMaggio, „Classification in Art," *American Sociological Review* 52 (1987): 440–455, https://doi.org/10.2307/2095290.

* Horace Walpole *Castle of Otranto [Otrantoi kastély]* című regénye óta a „gótikus" olyan regényműfajt jelöl, amelyben a rémület és a fantasztikus játszik nagy szerepet, és amely regények cselekménye szellemjárta, romos és félelmetes helyeken játszódik. Vö. David Mikics. *A New Handbook of Literary Terms* (New Haven–London: Yale University Press, 2007), 137. (A szerk.)

[4] Franco Moretti, *Graphs, Maps, Trees* (London:Verso, 2005), 18.

 DOI 10.31400/dh-hun.2021.4.2492

generációnkénti ritmus megengedné. A kortárs krimi kedvelői gyakran egyforma örömmel olvassák Agatha Christie-t és Wilkie Collinst.[5] És legalábbis a kritikusok szívesen láttatják úgy Stephen Kinget, mint aki a H. P. Lovecraft és Bram Stoker művein keresztül egészen *Az otrantói kastély*ig *(The Castle of Otranto)* visszanyúló folytonos gótikus hagyomány örököse.[6]

Tisztában vagyunk vele, hogy a fentiek mind a recepciót illető érvényes megállapítások. Morettinek igaza van, hogy a műfaj akadémiai vizsgálatai gyakran egy generációnyi időszakra terjednek ki. De az is igaz, hogy a „detektívtörténethez" hasonló kategóriák több, mint egy évszázada fontosak az olvasóknak. A szöveges elemzés egyik állítást sem fogja cáfolni, de arra talán rávilágíthat, hogy hogyan lehetnek összeegyeztethetők. Az áttekintések ellentmondásosságát feloldhatjuk például azzal a kézenfekvő magyarázattal, hogy Morettinek a műfajok ritmusával kapcsolatban tett megállapításai az általa vizsgált századra helytállóak – de a huszadik századra már nem, mert a műfajok eddigre tartós marketinges intézményekké szilárdulnak. Moretti azonban utalt arra, hogy[7] még egy olyan hosszú életű huszadik századi műfaj, mint a detektívtörténet is alapjában véve hasonló generációnyi szakaszok sorozatából állhat (a holmesi „ügy", a gyanúsítottak zárt köre a vidéki házban, a bűnügyi thriller), melyeket csak e viszonylag gyenge leszármazási szál fűz össze.[8] Előfordulhat, hogy egy ilyen sorozat eleje egyáltalán nem hasonlít a végére. Maurizio Ascari például amellett érvel, hogy „a *detektívtörténet* [detective fiction] kifejezést egyre inkább felváltja a *krimi* [crime fiction]", és szkeptikusan kezeli azt a régi narratívát, mely „Poe-t alapító atyának tekintette, és az ő »okoskodó meséiben« látta a műfaj modelljének a kibontakozását."[9]

Ha dönteni szeretnénk e megközelítések között, akkor lényegében arra a kérdésre keressük a választ, hogy a rövidebb életű „generációs" műfajok vajon koherensebbek-e, mint a hosszú életű műfajok. Ezen a ponton lehetnek segítségünkre a szöveges bizonyítékok. Matthew Jockers kimutatta, hogy a huszonöt-harminc éves skálán mozgó műfajok a tizenkilencedik században nyelvileg koherens jelenségek. Az úgynevezett ezüstvillás [silver-fork]* vagy a szenzációs regény példáin keresztül betanított statisztikai modell elfogadható pontossággal tud azonosítani további példákat ugyanarra a műfajra.[10]

[5] P. D. James, *Talking About Detective Fiction* (New York: Vintage, 2011).

[6] John Sears, *Stephen King's Gothic* (Chicago: University of Chicago Press, 2011), 51, 174.

[7] Moretti, *Graphs, Maps, Trees,* 31.

[8] A science fiction hasonló elméletéről lásd Adam Roberts, „A Brief Note on Moretti and Science Fiction," in Jonathan Goodwin and John Holbo, eds., *Reading Graphs, Maps, Trees: Critical Responses to Franco Moretti,* 49–55 (Anderson: Parlor Press, 2011).

[9] Maurizio Ascari, „The Dangers of Distant Reading: Reassessing Moretti's Approach to Literary Genres," *Genre* 47, 1. sz. (2014): 1–19, 15, https://doi.org/10.1215/00166928–2392348.

* Az „ezüst villás" [silver-fork] regények 1820–1830 között voltak népszerűek, a társadalom felső rétegének életéről szóltak (luxus, elegáns bálok). A regények az úgynevezett régens korban (1811–1820) játszódtak, azaz amikor Őrült III. György helyett a fia, IV. György mint régensherceg uralkodott. Vö. David Mikics, *A New Handbook of Literary Terms* (New Haven–London: Yale University Press, 2007), 278. (A szerk.)

[10] Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Urbana: University of Illinois Press, 2013), 67–81, https://doi.org/10.5406/illinois/9780252037528.001.0001.

Érdekes lenne kideríteni, hogy az olyan hosszabb kiterjedésű jelenségek, mint a detektívtörténet vagy a science fiction kisebb vagy nagyobb nyelvi koherenciát mutatnak-e, mint ezek a harmincéves kategóriák. Ösztönösen azt várnánk, hogy a hosszabb életű kategóriák határai elmosódnak; nehéz elhinni, hogy *A Holdgyémánt* (*The Moonstone,* 1868) és *A hosszú álom* (*The Big Sleep,* 1939) között szövegszinten túl sok hasonlóság lenne. Ha kiderül, hogy a „szenzációs regényeket" könnyebb felismerni, mint a „detektívtörténeteket", akkor lesz némi bizonyítékunk arra, hogy Morettinek igaza volt a műfajok mögöttes generációnkénti logikájával kapcsolatban. Ilyesfajta bizonyítékokkal nem zárnánk ki a hosszabb távú folytonosság lehetőségét: végül is nem tudjuk, hogy a könyveknek kell-e szövegesen is hasonlítaniuk egymásra ahhoz, hogy ugyanahhoz a műfajhoz tartozzanak. De ha azt látjuk, hogy a szöveges koherencia rövidebb időtávokon volt a legerősebb, abból annyi következtetést legalábbis levonhatunk, hogy az egy generációnyi műfajokban megtalálható egy bizonyos *fajtájú* koherencia, amely a hosszabb életű műfajokból hiányzik.

Másrészt legalább ugyanilyen ésszerű mértékben számíthatunk ezektől eltérő történelmi pályákra is. Például kifejezetten sok energiát szántak arra a kutatók, hogy leírják a műfaji konvenciók fokozatos szabványosulását a huszadik század elején, és a műfaj megszilárdulásának sarkalatos pontjaként emelték ki a műfajspecifikus ponyva megjelenését vagy az olyan kritikai kinyilatkoztatásokat, mint Ronald Knox a detektívtörténetek szabályait bemutató *Decalogue*-ja (*Tízparancsolat,* 1929).[11] Bizonyos műfajok esetében szerintük a folyamat még ennél is tovább tarthatott. Gary K. Wolfe azt állítja, hogy „a sci-fi regény" az 1940-es évek előtt „nem tudott egy műfajjá összeállni".[12] Ha ez az irodalomtörténeti rekonstrukció helyes, akkor nem egymástól eltérő, egymást követő generációs fázisokra számíthatunk, hanem a határok folyamatos keményedésére, amelyek a huszadik század közepén már sokkal világosabban elkülönülnek, mint a század elején. Erre a történetre számítottam, amikor belevágtam a jelen projektbe.

E kérdések megválaszolásához a recepció tizennyolc különböző helyszínén gyűjtöttem össze a műfajhoz rendelt címek listáját. A listák egy része a közelmúlt kritikai álláspontjára épít, más listákat írók és szerkesztők jelöltek ki a huszadik század elején, míg egyes listák nagy számú könyvtári katalóguskészítő gyakorlatát tükrözik (lásd az *A függelék*et). Bár minden egyes lista némileg eltérően határozza meg a tárgyát, nagyjából három kategóriába sorolhatók: detektívtörténet [detective fiction] (vagy „rejtélyes történet" [mystery] vagy „krimi" [crime fiction]), (a szintén többféleképpen definiált) sci-fi [science fiction] és a gótikus [Gothic]. (Vitatható, hogy a „gótikus" műfajnak számít-e egyáltalán – de éppen emiatt érdekes eset.) A Chicago Text Lab és a HathiTrust Digital Library gyűjteményére támaszkodva én magam is kigyűjtöttem az ezekbe a kategóriákba tartozó szövegeket.[13] Ha összehasonlítjuk a különböző recepciós helyszínekkel és az idővonal különböző szegmenseivel társított szövegcsoporto-

[11] Ronald Knox, „'Introduction' to *The Best Detective Stories of 1928–29* (1929)", reprinted in Howard Haycraft, ed., *Murder for Pleasure: The Life and Times of the Detective Story,* Revised Edition (New York: Biblio and Tannen, 1976).

[12] Gary K. Wolfe, *Evaporating Genres: Essays on Fantastic Literature* (Middletown: Wesleyan University Press, 2011), 21.

[13] Ha óvatlanok vagyunk, a szövegek normalizálási folyamata elrejtheti előlünk a buktatókat. Megpróbáltam például eltávolítani a címnegyedeket, az előszókat és az élőfejeket, amelyek egyébként tartalmazhatnak olyan explicit műfaji címkéket, amelyekre a modell kíváncsi lenne. Az is aggo-

 DOI 10.31400/dh-hun.2021.4.2492

kat, meg tudjuk állapítani, hogy a különböző kategóriák pontosan mennyire voltak stabilak.

Az ebből a kísérletből körvonalazódó történet az imént vázolt alternatív elképzelések – a generációk váltakozása vagy a műfaji konvenciók fokozatos megszilárdulása – egyikéhez sem tudott szépen igazodni. Kevés bizonyítékot látok a Moretti elmélete által előre jelzett generációnkénti hullámokra. Sőt, még csak arról sincs szó, hogy az időrendileg egy koncentrált műfajba (mint például a „szenzációs regény, 1860–1880") tartozó könyvek szükségszerűen jobban hasonlítanának egymásra, mint a hosszú idővonalat lefedő könyvek. A detektívtörténet és a sci-fi legalább akkora szöveges koherenciát mutat, mint a Moretti-féle rövidebb életű műfajok, és ezt a koherenciát nagyon hosszú ideig (160 vagy akár 200 évig is) fenn tudják tartani. Tehát, azt hiszem, félretehetjük azt a (termékeny) feltételezést, mely szerint a huszonöt éves generációs ciklusok különösebb jelentőséggel bírnának a műfajok tanulmányozása szempontjából.

De várakozásaim ellenére nem találtam sok bizonyítékot a fokozatos megszilárdulást feltételező elméletre sem. Bár az kétségtelenül igaz, hogy a műfajt irányító kiadói intézmények fokozatosan fejlődtek, úgy tűnik, tévedtem, amikor arra számítottam, hogy a műfajok közötti szöveges különbségek is fokozatosan alakulnak ki. A detektívtörténetek esetében például azok a szöveges eltérések, melyek a huszadik századi leleplezős történeteket más műfajoktól megkülönböztetik, nagyon világosan visszavezethetők *A Morgue utcai kettős gyilkosság*ig – de sokkal tovább nem. A detektívtörténet abban az értelemben fokozatosan terjedt, hogy Poe és Vidocq kezdetben még elszigetelt, másodrangú, utánzók nélküli alakok voltak, műfajspecifikus magazinokról és könyvklubokról pedig végképp nem lehetett még beszélni. De a szöveges mintázatoknak nem kell olyan fokozatosan fejlődniük, mint az intézményeknek. Poe történeteiben már számos olyan jellemző megtalálható, amelyek a huszadik századi krimit elkülönítik a többi műfajtól.

Prediktív modellezés

A számítógép tulajdonképpen azért lép színre ebben a dolgozatban, hogy a segítségével próbáljuk meg kibogozni a kortárs műfajelméletek szülte kusza problémákat. Ha azonosítani tudnánk egyetlen olyan formai alapelvet, amely alapján egyszer s mindenkorra meg lehetne határozni a műfajok mibenlétét, azzal jelentősen megkönnyíthetnénk a kritikusi feladatunkat. Elég lenne annyit mondanunk, hogy sci-fi az, amit Darko Suvin „a kognitív elidegenítés irodalmaként"[14] definiál. Az olvasók között azonban ritka az egyetértés egy adott műfaj meghatározó jegyeit illetően, míg a különböző közösségek ugyanazon művekkel kapcsolatban gyakran más-más dolgot értékelnek. A teoretikusok gyanúja szerint egyre valószínűbb, hogy a műfaj valójában „családi hasonlóságon" alapul, amelyet egy sor egymással átfedésben álló

dalomra adhat okot, hogy a szövegek két különböző könyvtárból (HathiTrust és Chicago Text Lab) származnak.

[14] Darko Suvin, „On the Poetics of the Science Fiction Genre," *College English* 34, 3. sz. (1972): 372–382, 372, `https://doi.org/10.2307/375141`.

funkció teremt meg.[15] Ráadásul a műfajok történelmi alkotások: fontos jellemzőik változhatnak.[16]

Összefoglalva: egyre inkább úgy tűnik, hogy a műfaj nem egyetlen, megfigyelhető és leírható tárgy. Talán inkább különböző módon kapcsolódó művek közötti összefüggések változékony rendszereként kell elképzelnünk, amelyek különböző mértékben hasonlítanak egymásra. Az ehhez hasonló problémák olyan módszertant igényelnek, amely óvatos az ontológiai feltételezésekkel, a részleteket illetően pedig türelmes. A prediktív modellezés éppen ilyen módszer. Leo Breiman szerint a prediktív modellek leginkább abban térnek el a szokványos statisztikai módszerektől (és, hozzátenném, a hagyományos kritikai eljárásoktól), hogy előzetes elvárások nélkül próbálják azonosítani a vizsgált jelenséget ténylegesen okozó és magyarázó mögöttes tényezőket.[17] A műfaj esetében ez azt jelenti, hogy a műfaj meghatározása helyett egy olyan modell azonosítására törekszünk, amely képes reprodukálni az egyes valós megfigyelők ítéleteit. A méretre vonatkozó jelzők (például „hatalmas" [huge], „gigantikus" [gigantic], de az „apró" [tiny] is) például a legmegbízhatóbb olyan szöveges kulcsok közé tartoznak, amelyek jelzik, hogy egy könyvet sci-finek fognak nevezni. Kevesen *definiálnák* a scifit a méreteken való merengésként, de kiderült, hogy a (bizonyos forrásokban) sci-fiként minősített művek valójában feltűnően sokat tárgyalják ezt a témát. Adjunk hozzá még néhány száz szót, és talán kapunk egy olyan statisztikai modellt, amely azonosítani tudja az ugyanezekben a forrásokban „science fiction" névvel illetett műveket, akkor is, ha a műfaj alapját képező definíció továbbra is nehezen fogalmazható meg (vagy talán soha nem is létezett).

Újabban Hoyt Long és Richard Jean So hasonlóképpen prediktív modellekkel igyekezte felderíteni a haiku stílus „látens, nem explicit nyomait" az angol költészetben.[18] Az ilyen projektekben a gépi tanulás célja nem is annyira az, hogy megnöveljük az általunk figyelembe vett könyvek számát, hanem hogy észlelhessük és összehasonlíthassuk az elmosódott családi hasonlóságokat, amelyek redukció nélkül nehezen lennének verbálisan definiálhatók. Sarkosabban fogalmazva: a számítógépes módszerek hasznossá teszik a kortárs műfajelméletet. Megszabadulhatunk a berögzült meghatározásoktól, és a műfaj tanulmányozását ehelyett csak bizonyos történelmi szereplők változó gyakorlatára alapozhatjuk – de még mindig olyan műfaji modelleket kapunk, amelyek értelmes összehasonlításokat és szembeállításokat tesznek lehetővé.

Mivel a prediktív modellben az egyes változókhoz nincs ok-okozati erő rendelve, a tulajdonságok kiválasztása nem elsődleges fontosságú. Az alábbi kifejtésben ugyan használok szavakat kulcsfogalomként, de nem szeretném azt sugallni, hogy a műfaj nyelvi jelenség. Mert nem az. A műfaj egy tág értelemben vett társadalmi jelenség, és a szavak történetesen kényelmes prediktív kulcsként működnek; segítségükkel nyo-

[15] Paul Kincaid, „On the Origins of Genre," *Extrapolation* 44, 4 sz. (2003): 409–419, 413–414, https://doi.org/10.3828/extr.2003.44.4.04.

[16] John Rieder, „On Defining SF, or Not: Genre Theory, SF, and History," *Science Fiction Studies* 37, 2. sz. (July 2010): 190–209, 193.

[17] Leo Breiman, „Statistical Modeling: The Two Cultures," *Statistical Science* 16, 3. sz. (2001): 199–231, https://doi.org/10.1214/ss/1009213726.

[18] Hoyt Long and Richard Jean So, „Literary Pattern Recognition: Modernism between Close Reading and Machine Learning," *Critical Inquiry* 42, 2. sz. (2016): 235–267, 266, https://doi.org/10.1086/684353.

 DOI 10.31400/dh-hun.2021.4.2492

mon követhetjük a különféle kiválasztási gyakorlatok közötti implicit hasonlóságokat és különbségeket. Tetszés szerint akár a szöveg más jellemzőit is használhatnánk. Egyes kutatók írásjeleket vagy karakterhálózatokat használtak a műfaj előrejelzésére; amikor egy 850000 kötetből álló gyűjteményben próbáltam műfajokat azonosítani, az oldalformátumhoz kapcsolódó jellemzőkben gondolkodtam.[19] De annak a projektnek az volt a célja, hogy maximalizáljuk a modell teljes prediktív pontosságát, mivel ismeretlen vizekre vetettük ki a hálónkat, és egyszerűen csak a lehető legnagyobb fogást akartuk elérni.

A jelen projekt célja más. Címkézett példákkal dolgozom, nem pedig címkézetleneket próbálok elkapni. Ha az itt leírt modellek mindegyikét javítani lehetne 1%-kal, az semmit sem változtatna az érvelésen. Ami igazán számít, az a különböző szövegcsoportok közötti határok *relatív* erőssége. Ezért fektettem kevés energiát a pontosság optimalizálására; inkább a minél jobb olvashatóságra és a következetességre törekedtem. Az itt leírt modellek mind ugyanazt a tulajdonsághalmazt használják, amelyet úgy kaptunk meg, hogy egyszerűen vettük a teljes gyűjtemény 10000 leggyakoribb szavát (dokumentumgyakoriság szerint), minden műfajban és a meghatározott műfaj nélküli művekben is. (Vehettük volna a leggyakoribb 5000 vagy 3000 szót is, a pontosság így körülbelül 1%-kal változna.) A tanulási algoritmushoz L2-szabályos logisztikus regressziót használok, egy jól ismert algoritmust, mely viszonylag egyszerű becsléseket ad a jellemzők fontosságáról.[20]

A puszta pontosság iránt tanúsított gavalléros hozzáállásunknak persze megvannak a maga határai. Ha a statisztikai modellek egyáltalán nem tudnák megjósolni a műfajt, nyilvánvalóan nem szolgálnának hasznos bizonyítékokkal. Ebbe a problémába azonban itt nem fogunk beleütközni. Az itt tárgyalt modellek 70–93%-os pontossággal tudnak előrejelzéseket tenni, inkább e tartomány felső vége felé csoportosulva. És bár most egy olyan műfajt fogunk jellemezni, amelyet mindössze 76%-os pontossággal, „viszonylag laza" csoportosulásként jósolnak meg a modellek, és ezt állítjuk szembe egy 91%-os biztonsággal felismerhető műfajjal, az igazság az, hogy ezek a számok mind nagyon jelentős hasonlóságokra utalnak egy szövegcsoporton belül: jóval a szokásos társadalmi-tudományos hatásnagysági küszöbértékek felett vagyunk.[21]

Bár minden műfajnál ugyanazt a módszert alkalmaztam, nem tudom előre garantálni, hogy a módszer minden műfajnál egyformán működőképes. Ha egy műfajt különösen nehéz a szókincs alapján „megfogni", akkor nehézkes lehet a szózsák [bag-of-words] modell segítségével osztályozni. A sci-fi például különleges problémákat vet fel, mivel a tengeralattjárók egy idő után már nem számítanak futurisztikus elképzeléseknek, és a holnap újabb csodái váltják fel őket. A gyakorlatban ritkán találkozunk ilyen problémákkal. A lexikális modelleknek nem okoz nehézséget megtalálni a tematikusan eltérő műveket összekapcsoló közös formális elemeket. Általában a

[19] [üres lábjegyzet az eredetiben]

[20] Fabian Pedregosa et al., „Scikit-learn: Machine Learning in Python," *Journal of Machine Learning Research* 12 (2011): 2825–2830; Az előállított képek alapja: Hadley Wickham, *ggplot2: Elegant Graphics for Data Analysis* (New York: Springer, 2009), `https://doi.org/10.1007/978-0-387-98141-3`.

[21] Ha a prediktív pontosságot Cohen-féle „d" értékre alakítjuk, 65% felett már minden a „nagy hatás" kategóriába tartozik.

szövegek közötti olyan hasonlóságokat mutatnak ki, amelyek szorosan követik a kritikai intuíciónkat. De el is térhetnek a kritikusok elvárásaitól (amelyek viszont nem egységesek). A pozitív eltérések könnyen értelmezhetők: a lexikális modell által felfedezett folytonosságok mindjárt kizárják azt a tételt, hogy két recepciós helyszínnek nincsenek közös vonásai. A negatív eltérésekkel kapcsolatban lassabban fog kialakulni bizonyosság, mivel nem lehet *eleve* kizárni a modellt elkerülő hasonlóságok létezését. Abban, hogy a lazábbnak tekintett csoportosulások szöveges szinten valóban kevésbé koherensek, csak akkor kezdhetünk el bízni, ha módszereink már számos különböző műfaj koherenciáját felismerték.

Ha egy tanulási algoritmusnak elegendő változót adunk, akkor az lényegében „memorizálni" tud egy adathalmazt, és így valószerűtlenül pontos előrejelzéseket készít a korábban már látott példákról. Tehát egy számos változóval operáló modellt valójában csak visszatartott példákkal lehet tesztelni. A jelen dolgozatban használt modelleket folyamatosan visszatartott szerzők keresztellenőrzésével értékeltük ki.[22] Vagyis megmutatjuk a modellnek az egy halmazban szereplő összes szerzőt (egy kivételével), majd a modellt a nem látott szerző művein teszteljük. A folyamatot addig ismételjük, amíg az összes szerzőt lefedtük, és ki tudjuk számítani az egész halmazra vonatkozó pontosságot.

A műfajok háromnál több dimenziós térben találhatók

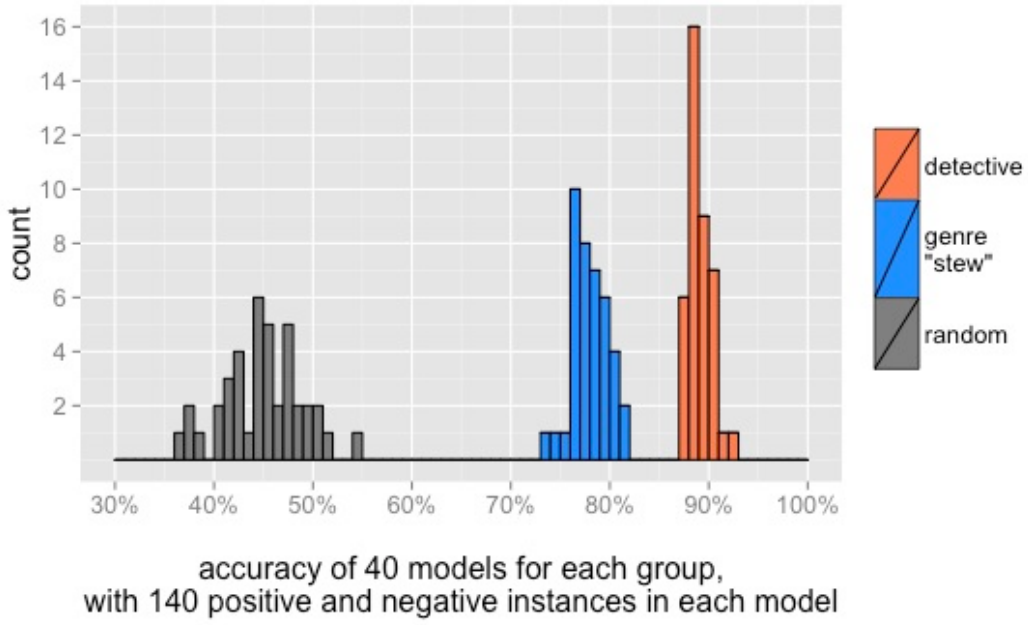
A „műfaj" szó alapvetően egy térképszerű mentális képet idéz fel, ahol a táj kisebb parcellákra van felosztva, és minden irodalmi mű szigorúan csak egy parcellában található meg. Ám a fiktív szövegek kategorizálásának a gyakorlata egyáltalán nem ilyen térképet eredményez, és én sem törekszem egy ilyen térkép előállítására. *A fehér ruhás nő*höz (*The Woman in White,* 1859) hasonló regényeket egyes megfigyelők a gótikus irodalomhoz, mások a szenzációs regényhez sorolják. Az én metaadataim között mindkét besoroláshoz kapcsolódó címkék szerepelnek. Más regények egyetlen határozott műfajhoz sem kapcsolódnak: valójában a tizenkilencedik századi művek többsége soha nem kapott konkrét besorolást. A fikciós művek tehát tartozhatnak sok műfajhoz vagy egyetlen egyhez sem. Ahelyett, hogy megpróbálnék felfedezni egyetlen olyan sémát, amely szerint ez az egész tér szerveződik, külön összehasonlítások sorát futtatom, és mindig olyan műveket csoportosítok, amelyek egy bizonyos műfaji besorolást tükröző címkét kaptak, és ezt a halmazt hasonlítom össze egy azonos méretű kontraszthalmazzal. A kontraszthalmaz általában véletlenszerűen lett kiválasztva valamely digitális könyvtárból (azt leszámítva, hogy milyen mértékben zár ki a pozitív halmazban szereplő címkéket), és olyan időbeli eloszlás szerint, mely a lehető legszorosabban egyezik a pozitív halmaz eloszlásával.

Ha nincs érdemi különbség a két összehasonlított kategória között, akkor a modell várakozásaink szerint a véletlenszerű találgatásnál (50%-os pontosság) nem sokkal pontosabb előrejelzéseket készít. A hitelesség ellenőrzése végett célszerűnek látszik e teszt lefuttatásával kezdeni. Véletlenszerűen két „csapatba" soroltam az ebben a cikk-

[22] Ennek kifejtését lásd: D. Sculley and Bradely M. Pasanek, „Meaning and Mining: The Impact of Implicit Assumptions in Data-Mining for the Humanities," *Literary and Linguistic Computing* 23, 4. sz. (2008): 409–424, https://doi.org/10.1093/llc/fqn019.

 DOI 10.31400/dh-hun.2021.4.2492

ben felhasznált összes szerzőt, mindkét „csapatból” véletlenszerűen kiválasztottam 140 kötetet, majd megpróbáltam betanítani a modellt a két csoport megkülönböztetésére. Az *1. ábra* negyven kísérlet eredményeit ábrázolja (szürkével). Ahogy látható, az átlagos pontosság csak egy kicsivel alacsonyabb (45%), mintha véletlenszerűen tippeltünk volna. Amikor egy osztályozási algoritmus próbálja megtalálni a különbségeket két véletlenszerűen kiválasztott csoport között, képes ugyan még felfedezni egy halvány mintát, de az esetleges minta nem fog értelmes kapcsolatban állni a visszatartott példákkal, és egy ilyen haszontalan szabály könnyen olyan elfogultságot adhat az előrejelzéseknek, amely a találgatásnál is rosszabb. Így tehát elenyésző a veszélye annak, hogy az algoritmus látszólagos különbséget azonosít két olyan csoport között, amelyeknél ilyen különbségek nem állnak fenn.



1. ábra. 3 vélelmezett „műfaj” 40 modelljének pontosságát ábrázoló hisztogram. Minden modellhez 140 pozitív eset lett véletlenszerűen kiválasztva egy hosszabb listáról. A „detektívtörténet” lista a következő szakaszban kifejtett módszerekkel lett kialakítva.

De van egy másik fajta alapteszt is, amelyet le kell futtatnunk. Mi történik, ha az általunk tanulmányozott műfajok *bármelyikével* felcímkézett valamennyi műből egyetlen szörnyű ragut készítünk, és ezt a szuperhalmazt hasonlítjuk össze az összes véletlenszerűen kiválasztott, semmilyen műfaji címkével nem társított művel? Ahogy az *1. ábrá*n látható, a modell gyakran felismeri a „műfaji raguból” érkező köteteket, jóllehet a gótikus, a Newgate, a szenzációs, a detektív és a sci-fi ezen kombinációja valószínűleg egyáltalán nem képez egy szokványosan koherens műfajnak nevezett csoportot. A modell előrejelzései átlagosan 78%-ban helyesek.[23] Ha rápillantunk né-

[23] Mivel a jelen dolgozatban mindig két egyenlő méretű kategóriát fogunk összehasonlítani, nem fogom megkülönböztetni a precizitást a felidézéstől.

hány olyan szóra, melyek az ebben a műfaj-szuperhalmazban megfigyelhető tagságra utalnak, könnyen megérthető, mi is történik. Az ide tartozó egyes műfajok nem teljesen különböznek egymástól. Közös bennük a szenzációs tárgy és néhány olyan kellék vagy cselekményszövési eszköz, amelyek nagyobb valószínűséggel fordulnak elő bennük, mint egy véletlenszerűen kiválasztott műben. A prediktív szavak listájának tetején áll például a „gyilkosság" [murder], a „szörnyű" [ghastly], a „zár" [lock], a „kulcs [key]", az „elmélet" [theory] és a „laboratórium" [laboratory]. Nehezebben azonosíthatók a véletlenszerű kontraszthalmaz jellegzetes szavai, de (összehasonlításképpen legalábbis) általában a hétköznapi családi élet különböző vonatkozásait idézik fel („házas" [married], „hibáztat" [blame], „reggelek" [mornings], „büszke" [proud], „barátok" [friends], „délutánok" [afternoons]). A kép talán megváltozna, ha olyan műfajokat tanulmányoznánk, mint a fejlődésregény. De az ehhez a cikkhez összeállított adatkészletben nagy különbségek figyelhetők meg a műfaji fikció, mint olyan, illetve a véletlenszerűen kiválasztott, viszonylag köznapi háttér között.

Ha jobban meg akarnánk érteni ezt a különbséget, nem lenne elég egy tízezernél több jellemzőt tartalmazó lista alját és tetejét megnézni. Kétségtelenül ki lehetne alakítani egy olyan szemantikai keretet, amely támogatná és pontosítaná a „szenzációs" és a „köznapi" témát érintő, rendszertelen következtetéseimet. Ám ez helyet és időt venne igénybe, és ennek a dolgozatnak nem az az elsődleges célja, hogy a tárgyalt műfajok mindegyikének a tartalmáról új leírást kínáljon. Így tehát, bár a jellemzők teljes listája elérhető egy online kód- és adatmellékletben,[24] a leírásaim rövidek lesznek, és mindegyik modellnél csak néhány olyan prediktív szóra térek ki, amelyek általános benyomást nyújtanak a tárgyalt szembeállításról.

A műfajok újradefiniálása helyett ebben az esszében alapvetően a műfajok változó élettartamával és szöveges koherenciájuk fokozataival kapcsolatban hozok fel érveket. E célból nem az a fontos, hogyan jellemezzük a nyelvezetet, inkább az, hogy hogyan értelmezzük a szövegcsoportok közötti hasonlóság fokát. Különösen fontos megérteni, hogy a műfajok hasonlóságának a tere háromnál több dimenzióra terjedhet ki. Az egyik tengelyen egymáshoz közel álló elemek más szempontból nagyon messze lehetnek egymástól. Láttuk például, hogy ha összekeverjük a detektívtörténetet, a sci-fit és más műfajokat, akkor az esetek 78%-ában megkülönböztethetők a véletlenszerűen kiválasztott művektől. Csábító lenne ebből arra következtetni, hogy a detektívtörténet és a sci-fi „alapvetően folyamatos" vagy „78%-ban hasonló"; akár el is kezdhetnénk felépíteni egy narratívát, amely Sherlock Holmes kémiai iskolázottságát a két műfaj közötti döntő, eddig figyelmen kívül hagyott kapcsolatnak tekintené. Valójában azonban a különbségek e műfajok *között* még erősebbek, mint a véletlenszerűen kiválasztott háttérrel szemben fennálló közös különbözőségük. Ha betanítunk egy modellt, hogy megkülönböztesse a detektívtörténetet a sci-fitől vagy a gótikustól, akkor az esetek több mint 93%-ában igaza lesz (mindkét esetben).

A prediktív modellek kicsit olyanok, mint az emberek: mindig találnak olyan módszereket, amelyek szerint a két műcsoport hasonló, de olyanokat is, amelyek szerint különböznek. Ha tudni akarjuk, hogy a detektívtörténet és a sci-fi értelmesen egy kalap alá vehető-e, az egyszerű kétoldali összehasonlítás soha nem fog válaszolni a

[24] Underwood, „Code and Data to Support The Life Cycles of Genres," 12 May 2016, `https://github.com/tedunderwood/fiction`, https://doi.org/10.22148/16.005.

 DOI 10.31400/dh-hun.2021.4.2492

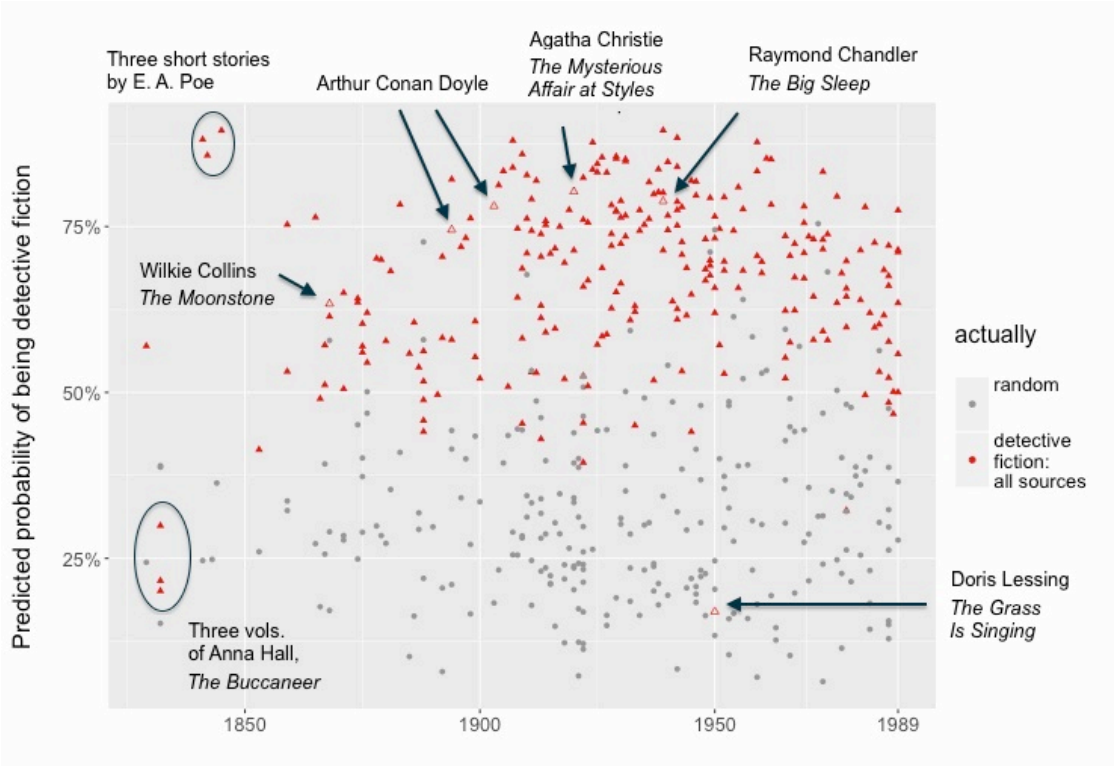
kérdésre. Érdemesebb lehet több összehasonlítás relatív erejére figyelnünk. Az *1. ábrá*n például azt látjuk, hogy a detektívtörténetet lényegesen könnyebb megkülönböztetni a véletlenszerű kontraszthalmaztól, mint a „műfaji ragunktól", ami már arra utalhat, hogy ez egy szorosabb kötésű kategória. Alternatív megoldásként feltehetnénk egy háromoldalú kérdést, amelynek segítségével két műfajt ugyanabban a koordinátarendszerben vizsgálhatunk. Például *ugyanúgy* különböznek-e a detektívtörténetek a többi fikciós műtől, mint a sci-fi? A prediktív modellek jól tudják egy bizonyítékhalmaz alapján a másikat kikövetkeztetni. Így betaníthatjuk a modellünket a detektívtörténet és egy véletlenszerűen kiválasztott háttér közötti megkülönböztetésre, majd kérhetjük ugyanazt a modellt a sci-fi műfajú művek és ugyanazon háttér megkülönböztetésére. Ahogy az várható, a modell teljesen megbukik: az esetek kevesebb, mint a felében lesz igaza. Bár a két műfajban néhány dolog közös (például az elméletek és a laboratóriumok), a *legtöbb* olyan jellemzőjük, amely megkülönbözteti őket a háttértől, eltérő. Amikor el kell döntenem, hogy két műfaji modell hasonló-e, az alábbiakban ezt a tesztet fogom a legmegbízhatóbbnak tekinteni. Annyit biztosan meg tud mondani, hogy a szörnyű „műfaji ragu" elkülöníthető detektívtörténetre és sci-fire. A következő kérdés az, hogy maga a „detektívtörténet" feloszlik-e további alműfajokra vagy bizonyos recepciós helyszínek köré csoportosuló művekre, amelyek jobban különböznek egymástól, mint amennyire hasonlítanak.

A detektívtörténet

A huszadik század második feléből származó műfaji listák összeállítása során elsősorban a Kongresszusi Könyvtár (Library of Congress, Washington) műfaj/forma címsoraira támaszkodtam. A köteteket különböző könyvtárosok látták el ezekkel a címkékkel, és sok különböző ember műfajokkal kapcsolatos hallgatólagos feltevéseit tükrözik. A detektívtörténet esetében több különböző címsort egy kategóriába soroltam, többek között a „Rejtélyes történet" [Mystery fiction], a „Detektívtörténet és rejtélyes történet" [Detective and mystery fiction] és a „Nyomozók" [Detectives] tárgyi címsort (ha többnyire fikciós művekre alkalmazták őket). Ahogy megyünk vissza 1940 előttre, a címkék egyre szórványosabbá válnak; olyan művekkel találkozunk, amelyeket eredetileg akkor katalogizáltak, amikor a Kongresszusi Könyvtár még nem a jelenlegi formájában használta a rendszerét. E művek közül csak néhányat katalogizáltak újra a modern, bőbeszédű módszerrel. Így tehát a korábbi időszak tekintetében leginkább bibliográfiákra és kritikai tanulmányokra kell támaszkodnunk.

Több óriási detektívtörténeti bibliográfia is létezik, igazi kihívás olyat találni, amely elég kicsi az átíráshoz. A háború előtti detektívtörténettel kapcsolatban leginkább az Indiana University Library által 1973-ban szervezett „A krimi első száz éve 1841–1941" kiállítás katalógusára támaszkodtam. A kiállítás regények mellett novellásköteteket (és néhány önálló novellát) is katalogizál. A lefordított művekre is kiterjed. Én is hasonlóan nyitott leszek a jelen esszé során. A Jules Verne kaliberű írók Franciaországon kívül is elemi mértékben alakították a műfajt, így a fordítások kizárásával sokat veszítenénk. Sőt, kétlem, hogy az itt vizsgált mintázatok bármilyen lefordíthatatlan problémát támasztanának: mint látni fogjuk, Verne még fordításban is a sci-fi kivételesen tipikus képviselője marad.

Előfordulhat persze, hogy a detektívtörténet egyetlen (1941 előtti kötetekre korlátozódó) kiállítási katalógusa olyan képet alakít ki a műfajról, amely lényegileg különbözik a sokféle ember által katalogizált, 1829–1989 közötti kötetek által tükrözöttektől. De a statisztikai modellezés éppen az ilyen típusú kérdéseket teszi vizsgálhatóvá. Ha csak az indianai kiállítás 88 kötetét (amelyeket meg tudtam szerezni digitálisan) modellezzük, akkor igen magas, 90,9%-os pontosságot érhetünk el. A Kongresszusi Könyvtár műfaji címkéivel ellátott 177 kötet már vegyesebb képet mutat, és csak az esetek 87,3%-ában ismerhetők fel. Ha egyesítjük a két halmazt, 249 kötetet kapunk (mivel 16 mindkét halmazban szerepelt), amelyeket 91,0%-os biztonsággal tudunk felismerni. Így a különböző módokon kiválasztott csoportok keverése nem csökkenti a pontosságot; ez egy „szintlépést" biztosító kompromisszum.



2. ábra. A „detektívtörténet" halmazból való származás várható valószínűsége; 91,0%-os átfogó pontosság.

De, mint már említettem, az algoritmikus modellek hatékonyan találják meg a közös elemeket a művek egy-egy csoportjában. A két kategória (A és B) közötti hasonlóság valódi próbája az, ha megkérünk egy A és C közötti szembeállításra betanított modellt arra, hogy B-t is különböztesse meg C-től. Ha például azt kérjük a Kongresszusi Könyvtár detektívtörténet kategóriáján betanított modelltől, hogy különböztesse meg az indianai kiállítás anyagát egy hasonló véletlenszerű háttértől, azt még mindig 89,3%-os pontossággal tudná megtenni. Ez valóban igazolja, hogy a detektívtörténet esetében nagyrészt egybevágó definíciókkal van dolgunk.

Az általunk használt modell valószínűségi jellege miatt láthatóvá válik, hogy a detektívtörténet képviselői közül kik a különösen jellegzetesek vagy a különösen

 DOI 10.31400/dh-hun.2021.4.2492

nehezen osztályozhatók. Kivetíthetjük a köteteket az y tengelyre, amely azt jelzi, hogy a modell milyen bizonyossággal tudja a „detektívtörténet" halmazba sorolni őket. A *2. ábrá*n ezt készítettem el mind a 249, vagy egyes könyvtárosok által felcímkézett, vagy az indianiai kiállításon szerepeltetett kötettel.

Az egyik nagyon szembetűnő dolog Edgar Allan Poe három, az 1840-es évek elejéről származó történetének a pozíciója. E művek nemcsak a saját idejükben, hanem a teljes 1829–1989 közötti idővonalat meghatározó normák szerint is a műfaj példaértékű modelljeinek látszanak. Bár az, hogy Poe tartósan sablon tud maradni, összhangban áll a detektívtörténetek egyik elterjedt származástörténetével,[25] a dologról a kritikusok között nincs egyetértés. Moretti például azt állítja, hogy a detektívtörténet csak 1890 körül érte el ma ismert formáját,[26] Ascari pedig nyíltan tagadja, hogy Poe még mindig fontos lenne a krimi szempontjából.[27] Ilyesfajta folytonosságra egy statisztikai modellben egyáltalán nem számítottam. Sőt, inkább azt vártam, hogy a detektívtörténet határai egyre jobban elmosódnak, ahogy visszamegyünk Conan Doyle elé, abba az időszakba, amikor a leleplező történetek még gyakran összeolvadtak más műfajokkal, például a szenzációs regénnyel. Talán valamivel több elmosódás figyelhető meg az 1860–1870-es években, mint a huszadik század közepén. (És az 1830-as években néhány kirívó hiba is megfigyelhető: azok a katalóguskészítők, akik megpróbálták odáig nyújtani a „detektívtörténet" kategóriát, hogy egy 1832-es regény is beleférjen, végső soron tönkreteszik a koncepciót.) De a kritikusok által általában prototípusként azonosított korai művek *(A Morgue utcai kettős gyilkosság, A Holdgyémánt)* továbbra is példaértékűek ebben a modellben. Ez a bizonyíték nem feltétlenül jelöl ki egy „eredetet", és azt sem bizonyítja, hogy bizonyos írók határozták volna meg a műfajt. Csupán annyit bizonyít, hogy a huszadik század végén a detektívtörténet prototípusait azonosító kritikusok helyesen jártak el, amikor a saját koruk gyakorlatát visszavetítették a múltra.

De a bizonyíték azt is mutathatja, hogy a detektívtörténet folytonossága több, mint egy sor genealógiai kapcsolat a különböző formák között. Tegyük fel például, hogy összevonjuk az indianai kiállítást és a Kongresszusi Könyvtár címkéit egyetlen, 249 szövegből álló csoportba, de az 1930-as évnél időrend szerint kettéosztjuk a csoportot. Milyen mértékben változik a detektívtörténet definíciója a két időszak között? Ha az 1930 utáni 130 kötetet modellezzük, 88,5%-os pontosságot kapunk. De ha a modellt az 1930-ig megjelent 119 kötet alapján tanítjuk be, majd e modell alapján készítünk az 1930 utáni művekre vonatkozó előrejelzéseket, akkor a modell pontossága még mindig 86,9%-os. Az 1930-ig a detektívtörténetet jelölő verbális különbségek nagyrészt a későbbiekben is jellemzőek maradnak. Ez nem jelenti azt, hogy a műfaj nem változott. A kemény [hardboiled] detektív fő divatja például 1930-ban még sehol sincs, ez kétségtelenül fontos változás. De a *detektívtörténet* és az irodalmi mező többi része közötti különbség természete már nem változott drámai mértékben. A műfaj határai stabilak. (Érdemes megjegyezni, hogy ez annak ellenére igaz, hogy az 1930 előtti

[25] Stephen Rachman, „Poe and the Origins of Detective Fiction," in Catherine Ross Nickerson, ed., *The Cambridge Companion to American Crime Fiction* New York: Cambridge University Press, 2010), 17–28, https://doi.org/10.1017/CCOL9780521199377.003.

[26] Moretti, *Graphs, Maps, Trees,* 31.

[27] Ascari, „The Dangers of Distant Reading," 15.

véletlenszerű kontrasztanyag javarészt a HathiTrust könyvtárból, az 1930 utáni anyag viszont többnyire a Chicago Text Lab könyvtárából származik. A gyűjteményeket eltérő módon választották ki, de a különbségek nem elég nagyok ahhoz, hogy belezavarjanak a műfaji jelbe.)

A modellünk bizonyítékokat hoz arra, hogy a folytonosság elég erős ahhoz, hogy valódi problémát jelentsen a jelenleg uralkodó műfajelméleti nominalizmus számára. Azt az (érvényes) feltevést, hogy a műfajok nem feltétlen foghatók össze egy világos definíció vagy már létező esszencia mentén, gyakran egy-két lépéssel még tovább viszik, és azt sugallják, hogy a műfajokat csak egy genealógiai szál kapcsolja össze – hogy a múlt és a jövő pusztán a „különböző gyakorlatokat folytató közösségek" közötti folyamatos manőverezésként kapcsolódik össze.[28] Maurizio Ascari bírálja Morettit, amiért „pozitivista" módon közelíti meg a detektívtörténetet, és emlékeztet rá, hogy „a huszadik század folyamán a detektívtörténet alapjaiban megváltozott", sőt, „minden irodalmi műfaj szüntelenül változik".[29] Mark Bould és Sherryl Vint (2009) szerint „bár gyakran úgy érzékeljük őket, a műfajok soha nem a világban már eleve meglévő dolgok, amelyeket később a műfaj szakértői tanulmányoznak, hanem különböző állítások és gyakorlatok szövetéből álló, képlékeny, testetlen szerkezetek" – vagy még merészebben fogalmazva, „nincs olyan dolog, hogy sci-fi."[30] Amikor elkezdtem ezt a tanulmányt, a hasonló állítások egy részével talán magam is azonosulni tudtam; a műfaj stabilitásával kapcsolatos szkepticizmus összességében előnyösebbnek tűnt, mint a definíciókkal kapcsolatos parttalan vita. A prediktív modellek azonban lehetővé teszik a középutat. Kezdhetünk óvatos lépésekkel, meghatározott történelmi szereplők által kijelölt vélelmezett határokkal, majd empirikusan rákérdezhetünk, hogy hallgatólagos kiválasztási kritériumaik mennyire egyeznek vagy térnek el egymástól. A detektívtörténetek esetében a különböző emberek által különböző időpontokban gondozott szöveglisták igen nagy mértékben összeegyeztethetők. Sőt, a műfaj múltjának modellje kiválóan előre tudja jelezni a jövőjét.

De pontosan mi *volt* az a „detektívtörténet"-definíció, amely a háttérben működött? A válasz javarészt egyáltalán nem sokkoló. A „rendőrség" [police], a „gyilkosság" [murder], a „nyomozás" [investigation] és a „bűntény" [crime] szabja meg a műfaj tematikai alapvetését. A „gyanú" [suspicion], a „bizonyíték" [evidence], a „bizonyítani" [prove], az „elmélet" [theory], a „véletlen" [coincidence] (és, hogy egy finomabb példával is szolgáljunk, a „bárki" [whoever]) szavak alapozzák meg a cselekményt hajtó, kétségre és bizonyításra építő szerkezetet. Ha egy kicsit mélyebbre ásunk a modellben, kevésbé nyilvánvaló részletekre bukkanhatunk. Így például fontos kulcsfogalmak jönnek az építészet és a lakberendezés területéről: az „ajtó" [door], a „szoba" [room], az „ablak" [window], az „íróasztal" [desk] mind nagy prediktív erővel bíró szavak. A skála másik végén pedig a gyermekkort és az oktatást leíró szavak szerepelnek („született" [born], „nőtt" [grew], „tanított" [taught], „gyermekek" [children], „tanárok" [teacher]), melyek nagy bizonyossággal jelzik, hogy az adott kötet nem detektívtörténet. Talán

[28] Rieder, „On Defining SF, or Not," 15.

[29] Ascari, „The Dangers of Distant Reading," 7, 15.

[30] Mark Bould and Sherryl Vint, „There is No Such Thing as Science Fiction," in James Gunn, Marleen Barr, and Matthew Candelaria, eds., *Reading Science Fiction* (New York: Palgrave, 2009), 43–51, 48, 43.

 DOI 10.31400/dh-hun.2021.4.2492

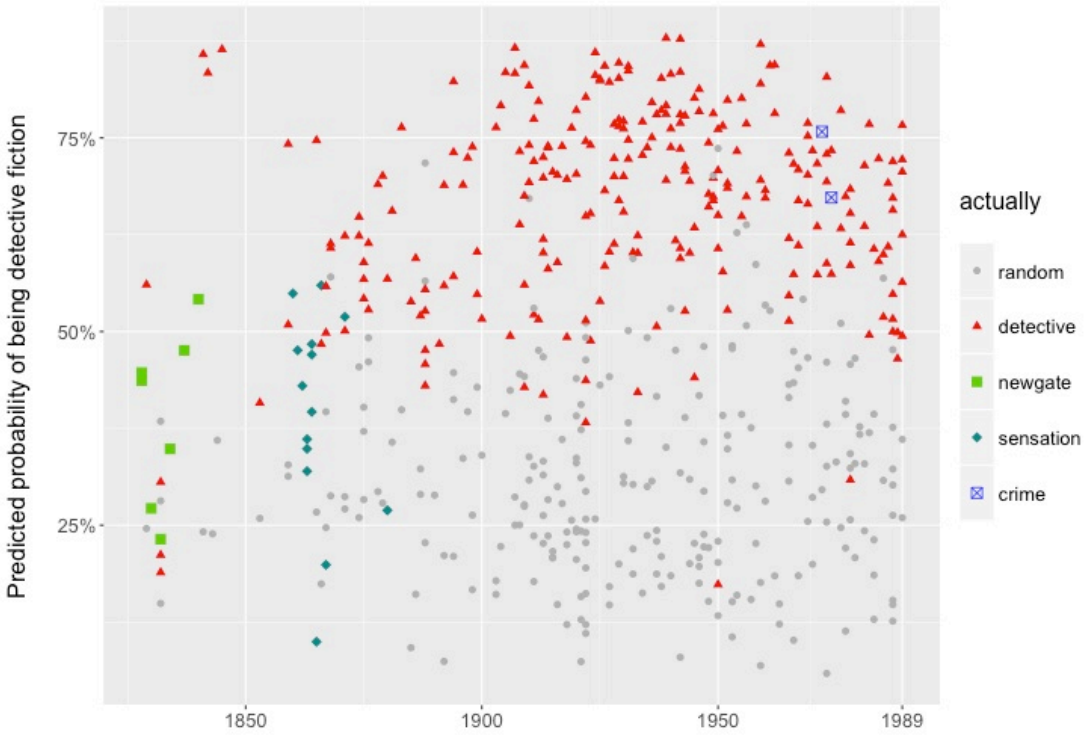
amiatt, hogy a műfaj jobbára egy bizonyos titokzatos eseményre koncentrál (vagy mert hajlamos novella formájában megjelenni), az életrajzi keretek viszonylag szűkösek.

A detektívtörténetről tehát megállapítottuk, hogy 160 éven keresztül (1829–1989) szövegesen koherens maradt. De addig nem teszteltünk le alaposan egy modellt, amíg nem találjuk meg, hol omlik össze. Mi történik például, ha a nyomozó alakja helyett a bűnözői miliőt állítjuk előtérbe? A „detektívtörténet" és a „krimi" közötti határ nagyon képlékeny, és Patricia Highsmith regényei, illetve az úriember tolvaj Arthur J. Raffles alakja szinte meg is kérdőjelezi e határok létjogosultságát. Mi történik, ha felveszünk néhány „krimi" címkével igen, de „detektívtörténet" címkével el nem látott huszadik századi regényt, valamint egy csoport 1820–1840 közötti Newgate-regényt? Elvégre a detektívtörténetek egyik klasszikus feldolgozása a *Twist Olivér*rel kezdődik.[31] A szenzációsregényeket is azonosították már a detektívtörténetek előfutáraként;[32] *A Holdgyémánt* az indianai kiállításon is szerepelt, de megpróbálhatnánk további szenzációs regényeket hozzáadni, hogy lássuk, azok is beleférnek-e ebbe a kategóriába.

Bizonyos helyeken a modellünk könnyen magába fogad más kategóriákat, más helyeken viszont egyáltalán nem rugalmas. A *3. ábrá*n egy olyan modell látható, amelyet a *2. ábrá*nál felhasznált 249 detektívtörténettel tanítottam be, de azt is engedélyeztem neki, hogy a betanításhoz használt halmazon kívüli más művekről is készítsen előrejelzéseket. A „krimi" címkével ellátott újabb regényekről (például Patricia Highsmith, *A Dog's Ransom [Silány váltságdíj]*) kiderült, hogy kifejezetten jól összeegyeztethetőek a detektívtörténeti modellünkkel. A tizenkilencedik századi szenzációs regények és a Newgate-regények azonban nem férnek bele ugyanabba a szövegdobozba. Mindez nem azt bizonyítja, hogy D. A. Miller tévesen tárgyalta együtt a Newgate-regényeket és a detektívtörténeteket a *The Novel and the Police* című könyvében. Végső soron semmilyen törvény nem írja elő, hogy az irodalomtörténeti fogalmaknak a nyelvezet szintjén is felismerhetőnek kell lenniük. Ha úgy szeretnénk definiálni a krimi nevű műfajt, mint amely magába foglalja a Newgate-regényt, bátran megtehetjük, és a fogalom megvilágító erejű lehet. De így már egy kissé más típusú műfajról fogunk beszélni, és ebből a műfajból hiányzik a nyelvi homogenitásnak az a foka, amely Edgar Allan Poe-t Agatha Christie-vel és Patricia Highsmith-szel egyaránt összeköti. A vizsgálat lényege más szóval nem az, hogy eldöntse, mit lehet vagy mit nem lehet „műfajnak" nevezni, hanem hogy segítsen megkülönböztetni azokat a különböző mintákat, amelyek leírására az irodalomtörténészek használni szokták a szót.

[31] D. A. Miller, *The Novel and the Police* (Berkeley: University of California Press, 1988).

[32] Christopher Pittard, „From Sensation to the Strand," in Charles J. Rzepka and Lee Horsley, eds., *A Companion to Crime Fiction* (Oxford: Wiley-Blackwell, 2010), 105–116, https://doi.org/10.1002/9781444317916.ch7.



3. ábra. Egy csak a detektívtörténetekkel és véletlenszerű példákkal betanított modell előrejelzései három másik kategóriában.

A gótikus irodalom

A gótikus fikció eredete, stílszerűen, kései leszármazottaikat már pusztán halovány emlékek formájában kísértő ősalakok rejtélyes történeteként körvonalazódik. A kritikusok látszólag igencsak biztosak benne, hogy a gótikus regény egy 1760-tól talán 1830-ig tartó, kifejezetten koherens jelenség volt. De ahogy haladunk tovább a tizenkilencedik századba, egyre kevésbé egyértelmű, hogy a gótikus továbbra is folyamatos hagyomány marad-e. Mint már említettem, Franco Moretti a tizenkilencedik századi gótikus irodalmat a század elején és végén uralkodó két műfajra osztja fel. A huszadik századi Amerikában a „déli gótikus" irodalmat gyakran különálló irodalmi jelenségként kezelik. Erős érvek szólnak ezen túl egy kifejezetten női gótikus hagyomány létezése mellett, amely DuMaurier *Rebecca* és Brontë *Jane Eyre* című művéig vezethető vissza.[33] Egyes kritikai hagyományok viszont ragaszkodnak a felsorolt dolgok közötti folytonossághoz, és a gótikus irodalom kategóriáját addig tágítják, hogy az már a kortárs „horror" kiadói kategóriát is magába foglalja.[34] Kaphatók olyan, a gótikus irodalmat gyűjtő antológiák, melyek az *Otranto* és Anne Rice közötti teljes távolságot lefedik, így a recepció gyakorlati oldalát illetően kell lennie *valamilyen* folytonosság-

[33] Juliann Fleenor, ed., *The Female Gothic* (Montreal: Eden, 1983).
[34] Mark Edmundson, *Nightmare on Main Street: Angels, Sadomasochism, and the Culture of Gothic* (Cambridge: Harvard University Press, 1999).

 DOI 10.31400/dh-hun.2021.4.2492

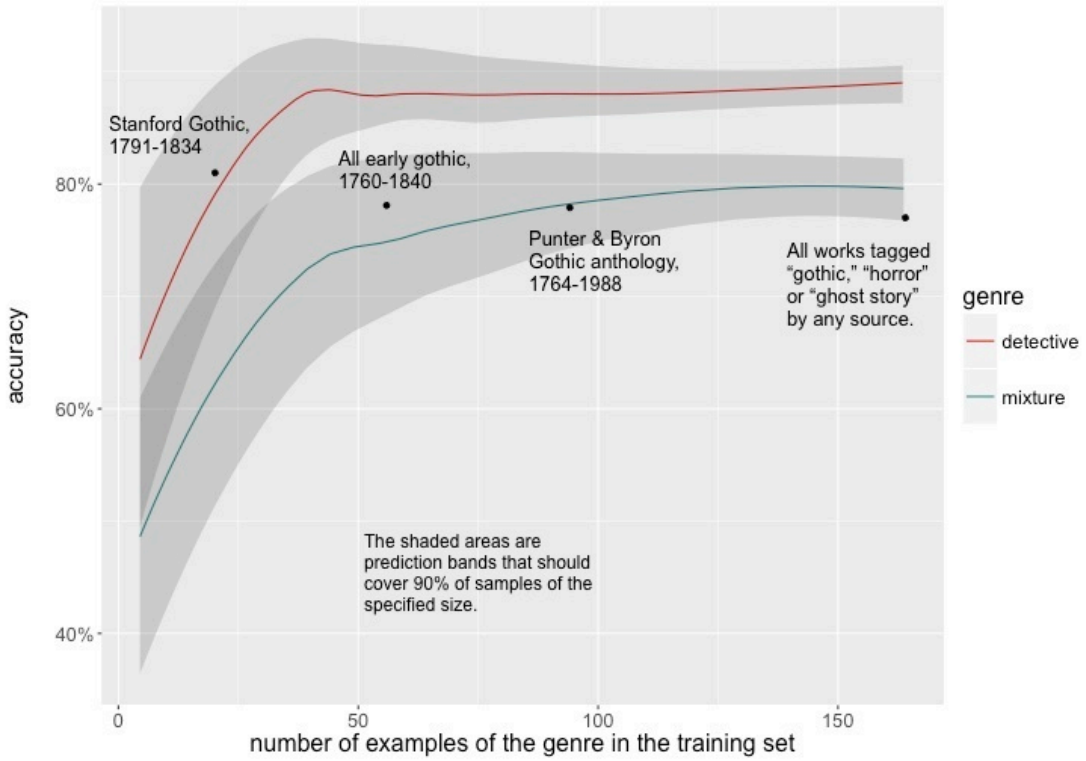
nak, nevezzük azt műfajnak, módnak, rajongói (szub)kultúrának (fandomnak) vagy témák laza sorozatának.

Mivel megalapozottan kételkedhetünk abban, hogy a „gótikus" irodalom valóban egyetlen, erősen egységet mutató hagyománynak tekinthető-e, ebben az esetben különösen fontos volt, hogy össze tudjuk hasonlítani a különböző forrásokból érkező bizonyítékokat. Az 1840 előtti időszak tekintetében nagy mértékben támaszkodtam a Stanford Literary Lab gótikus irodalomról készített listájára;[35] az 1940 utáni időszak kapcsán pedig növekvő mértékben tudtam a Kongresszusi Könyvtár „horrorhoz" vagy „kísértethistóriához" kapcsolódó műfaji címkéire építeni. Magam is számos művet kigyűjtöttem a *The Gothic* című Blackwell-kézikönyvben (szerkesztette David Punter és Glennis Byron, 2004) szereplő említések alapján, amely Walpole-tól egészen Brett Easton Ellisig próbálja meg feltérképezni a gótikus hagyományt, és őket egy sor közvetítő, köztük Charlotte Brontë, Henry James és H. P. Lovecraft alakján keresztül köti össze.

Egyik lista sem mutat a detektívtörténet esetében észlelt mértékű koherenciát. A legkisebb méretű mintát sikerült a legpontosabban előre jelezni: a Stanford Literary Lab által gótikusként azonosított 21 művet (1791 és 1834 között) 81,0%-os pontossággal sikerült felismerni. Az összes listát magába foglaló szuperhalmazt volt a legnehezebb modellezni: a 165 kötetet csak 77,0%-os pontossággal tudtuk felismertetni. A 81% és 77% közötti számszerű eltérés itt kicsit drámaibb, mint első hallásra gondolnánk, mivel itt nem almát hasonlítunk össze almával. A pontosság várakozásaink szerint a modellezett készlet méretével párhuzamosan nő. E növekedést érzékeltetendő különböző görbéken ábrázolom a más műfajok esetében különböző mintaméreteknél megfigyelt átlagos pontosságot.

A gótikus regények stanfordi részhalmaza szövegesen éppolyan koherens, mint a detektívtörténetek hasonló méretű mintája. A gótikus fikció évszázadokon átívelő mintái ellenben méretükhöz képest kirívóan gyengén teljesítenek; nem könnyebb modellezni őket, mint a projektben vizsgált minden műfaj közös halmazát. Ez arra utal, hogy a romantika korának gótikus regényei nyelvileg nem sok közös vonást mutatnak a huszadik századi horror és természetfeletti fikció hagyományával. Ezt úgy erősíthetjük meg, ha az egyik ilyen időszak alapján tanítunk be egy modellt, majd a másik időszakra alkalmazzuk azt; ez alig teljesít jobban, mint a véletlenszerű halmaz.

[35] A „stanfordi gótika" esetében használt metaadatokat a Stanford Literary Lab fejlesztette ki; és alighanem sok ember, köztük Ryan Heuser és Matthew L. Jockers munkájáról van szó.



4. ábra. A gótikus fikció különböző mintái más műfajok különböző mintaméretek esetében mutatott pontossági tartományához viszonyítva.

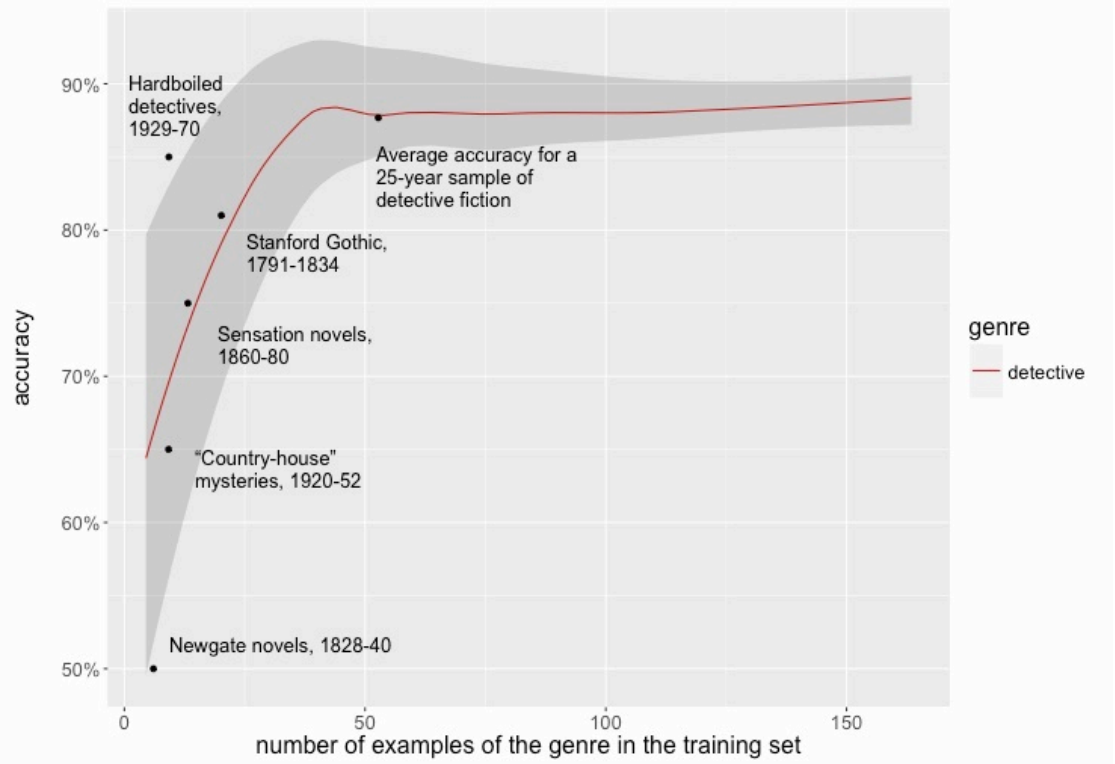
Nem hinném, hogy ez túlzottan meglepi az olvasókat. Nagyon kevés kritikus állítja azt, hogy a gótikus olyan szorosan kötött műfaj, mint a szenzációsregény vagy a detektívtörténet. Punter és Byron (2004) még egy két évszázados antológia építése során is elismeri annak a lehetőségét, hogy „nagyon kevés tényleges irodalmi szöveg van, amelyek »gótikusak«; vagyis a gótikus talán inkább bizonyos elszórtan észlelhető pillanatokon, trópusokon, ismétlődő motívumokon keresztül jelentkezik a nyugati irodalmi hagyományon belül".[36] A gótikus fikció két évszázados skálán történő modellezésének a nehézségei részben megerősítik ezt a gyanút, de talán módszerünk hitelességi ellenőrzésének is tekinthetők. (Ha egy módszer nem ismeri fel a nehéz eseteket, megbízhatósága más esetekben is megkérdőjelezhető.) Ez arra is emlékeztethet bennünket, hogy bizonyos folytonosságokat egy nyelvezeten alapuló modell nem tud észlelni. A kritikusok között egyetértés figyelhető meg a gótikus által jelölt dolgok spektrumát illetően, és azt, ha nem is éppen műfajnak, de legalábbis egy módnak vagy laza tematikus hasonlóságnak tekintik. Modellünk azonban ezt a spektrumot nem látja egységesebbnek, mint a népszerű műfajú művek véletlenszerűen kiválasztott halmazát.

Az egyik nyilvánvaló magyarázat erre az lehetne, hogy a gótikus fogalma az idővonal túl nagy részére terjed ki, és a nyelv egyszerűen túl sokat változik két évszázad alatt ahhoz, hogy a műfajt ilyen távlatban nyelvileg modellezni lehessen. De van olyan példánk is, ahol ez mintha nem jelentene problémát – a *4. ábra* például arra emlékeztet

[36] David Punter and Glennis Byron, *The Gothic* (Malden: Blackwell, 2004), xviii.

 DOI 10.31400/dh-hun.2021.4.2492

bennünket, hogy a detektívtörténet 1829-től 1989-ig egészen jól egyben marad. És valóban, ahogy azt az *5. ábra* mutatja, nem igazán van bizonyítékunk arra, hogy a kronológiailag koncentrált műfajok nyelvi értelemben általában koherensebbek, mint az a mintánk, amely a detektívtörténetek 160 évét fedi le.



5. ábra. Több, körülbelül generációnyi méretű műfaj helye, a detektívtörténetek 1829 és 1989 közötti véletlenszerű mintáján belül megfigyelt pontossági tartományt ábrázoló görbéhez képest. A sötétebb sáv egy, a detektívtörténetek véletlenszerű mintája alapján készített modell 90%-át lefedő előrejelzési sáv.

Ha a detektívtörténetek mintáját egy, a gótikussal, a Newgate-regénnyel vagy a szenzációstörténetekkel megegyező méretű betanítási halmazra csökkentjük, akkor a detektívtörténet éppolyan koherensnek tűnik, mint e kronológiailag koncentrált műfajok, annak ellenére, hogy a köteteket több, mint 150 éves időszakból válogattuk ki. Ez a döntő bizonyíték Franco Moretti azon feltételezése ellen, hogy a műfajok élettartama egy generációra terjedne ki. Úgy tűnik, hogy az egy generációnál jóval több ideig élő műfajokat olyan szöveges hasonlóságok tartják egybe, melyek éppen olyan erősek, mint a rövidebb életű műfajokat egybetartó hasonlóságok. Mivel nehéz bizonyítani, hogy ezek a modellek a szövegek közötti összes lehetséges hasonlóságot rögzítik, fogalmazzunk óvatosan: ha van is valamilyen generációs ritmus a műfajok történetében, ez a módszer nem észlelte azt – pedig úgy tűnik, az összes olyan mintát fel tudja ismerni, melyet a tudósok műfajnak szoktak nevezni.

Vajon még koherensebbé tehetnénk a detektívtörténetet, ha egy szűkebb generációs időtávra koncentrálnánk? Nem igazán. Ha a detektívtörténet hosszú ívét huszonöt éves szakaszokra bontjuk, majd azokat külön-külön modellezzük, olyan átlagos pon-

tosságot kapunk, amely nagyon közel áll az egész idővonalról vett hasonló méretű minta pontosságához. Legalább egy, a kritikusok által is elismert alműfaj esetében valamivel magasabb pontosság érhető el: ez nem más, mint a „kemény krimi". Még egy viszonylag kisebb, tíz darab 1929–1970 közötti kemény krimiből álló mintát is 85%-os pontossággal lehet kiemelni a halmazból. E regények stilisztikai homogeneitása talán társadalmi homogeneitásukkal áll összefüggésben. Dorothy Hughes az egyetlen nő a kemény krimiket tartalmazó századközepi mintánkban. A kemény krimi és a vidékiház-hagyomány* közötti különbség azonban egyáltalán nem mutat generációs megosztottságot; inkább a nemek és az Atlanti-óceán áll a hátterében. Ám ez sem egy leküzdhetetlen szakadék. A vidékiház-hagyományon betanított modell éppolyan pontosan (85%) azonosítja a kemény nyomozókat, mint a magukkal a kemény krimikkel betanított modell. Kevés okunk van tehát arra következtetni, hogy ezek az alműfajok függetlenedtek volna a „detektívtörténet" átfogó fogalmától.

Sci-fi

Eddig amellett érveltem, hogy az egy évszázadnál tovább fennálló műfaji fogalmak nyelvileg éppolyan koherensek lehetnek, mint a néhány évtizedig fennálló műfajok. Eddig azonban a detektívtörténet az egyetlen példám, és okkal gyaníthatjuk, hogy a detektívtörténet/rejtélyes történet/krimi műfaj szokatlanul stabil premisszákra épül. Mindig adott a bűntény, mindig adott a nyomozó, mindig adott a nyomozás. A sci-fi viszont mintha nagyobb kihívást jelentene, hiszen e műfaj premisszái eredendően változékonyak. A műfaj Verne-féle prototípusai gyakran írnak le léggömböket, tengeralattjárókat és további olyan szállítóeszközöket, amelyek ma már nem számítanak tudományos-fantasztikusnak. A műfaj legújabb képviselői már nagyon másfajta technológiákra építenek. Elsőre nem tűnik nyilvánvalónak, hogy Gibson *Neuromanc*ában, Wells *Az időgép*ében és Mary Shelley *Frankenstein*jében túl sok lenne a közös szókincs. Számos szkeptikus műfaji elmélet éppen a sci-fi változékonyságából indult ki – ahogy arra *A sci-fi nem létezik [There Is No Such Thing As Science Fiction]* cím emlékeztet minket.[37] Összefoglalva: ösztönösen nem világos, hogy a sci-fi esetében azt várhatjuk-e, hogy hosszú időtávon át is egységes marad, mint a detektívtörténet, vagy éppen hogy szétesik, mint a gótikus irodalom.

Mivel a sci-fi korai történetéről többféle narratíva létezik, ezt az időszakot különböző források alapján dolgoztam fel. A *The Anatomy of Wonder* című közismert bibliográfia a sci-fi korai történetéről szóló fejezeteket is tartalmaz, melyek szerzője a sci-fi író Brian Stableford. Stableford műfajtörténete erősen hangsúlyozza H. G. Wells és a jövőbeli háború hagyományának szerepét, ugyanakkor kissé óvatosabb az olyan elődökkel kapcsolatban, mint Mary Shelley és Jane Loudon. (A sci-fi számos tör-

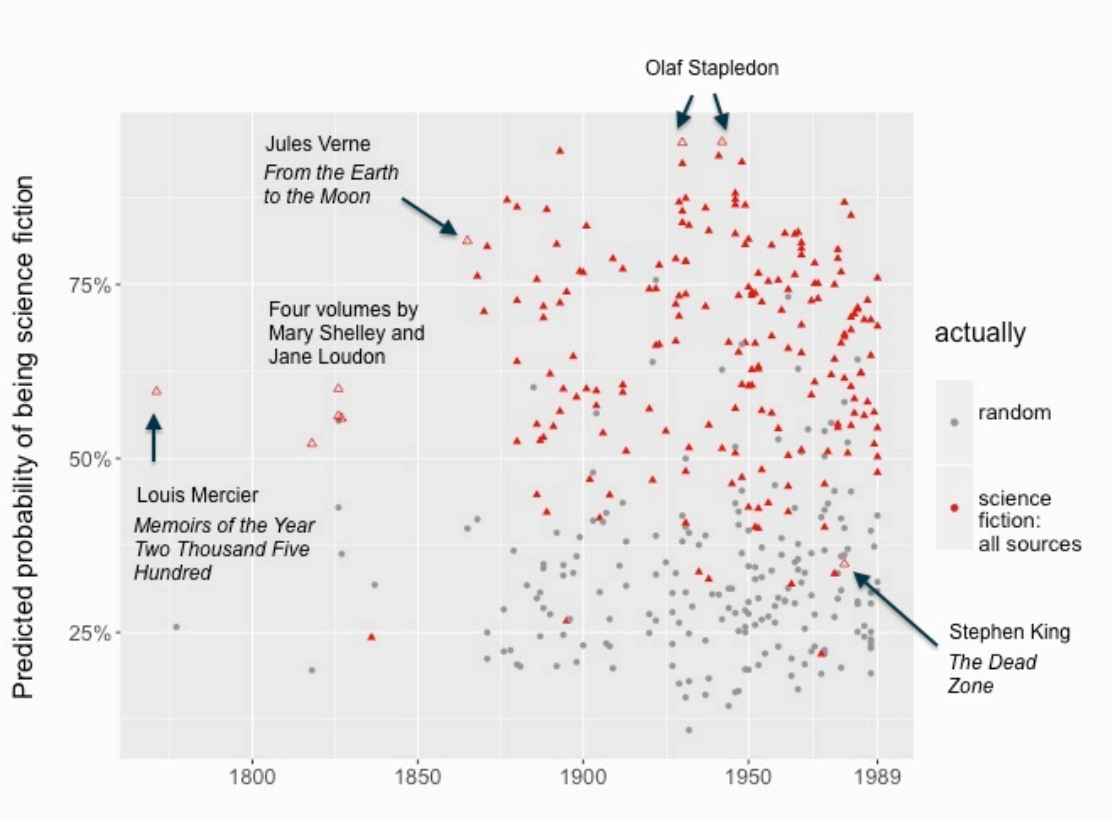
* A vidékiház-hagyomány a detektívtörténetek egyik alműfaja. A deketívtörténetek egy része olyan helyen játszódik, hogy a lehetséges gyanúsítottak egy zárt halmazból kerülhetnek ki, nincs lehetőség külsős elkövetőre, mert a gyilkosság egy vonaton, hajón, repülőn stb. játszódik. Ilyen a vidéki ház hagyomány is, ahol a ház lakói a gyanúsítottak. Általában vidéki arisztokrata, dzsentri házakról van szó. Agatha Christi is írt ilyen regényt. (A szerk.)

37 Bould és Vint: „No Such Thing as Science Fiction"; lásd még: Kincaid, „On the Origins of Genre"; Rieder, „On Defining SF, or Not."

 DOI 10.31400/dh-hun.2021.4.2492

ténészéhez hasonlóan Stableford is hajlamos a tudományos tartalom révén meghatározni a műfajt, és ő is szkeptikus azokkal a művekkel kapcsolatban, melyekből az ilyen tartalom hiányzik.) Mivel a nők műfajon belüli megjelenéséről is teljesebb képet szerettem volna kapni, egy, a nők korai sci-fiben betöltött szerepéről szóló bibliográfiára is építettem, amelyet Mary Mark Ockerbloom (a University of Pennsylvania könyvtárának munkatársa) állított össze. Eltérő fogalmi hangsúlyaik ellenére e források olyan szöveglistákat kínálnak, amelyek nyelvi értelemben nagyon hasonló módon modellezhetők. Ha az egyik lista modelljét összekapcsoljuk a könyvtárosok által felcímkézett huszadik századi szövegekkel, akkor a modell 90%-os vagy nagyobb pontossággal tudja előre jelezni a másik listát.

Ha valamennyi bibliográfiai forrásunkat egyesítjük, akkor egy 196 kötetből álló, az 1771 és 1989 közötti időszakot lefedő listát kapunk, amely 88,3%-os pontossággal modellezhető. A műfaj határai egy kicsit kevésbé világosak, mint a detektívtörténet esetében, de koherenciája még így is egyértelműen közelebb áll ehhez a műfajhoz, mint a gótikushoz. (A pontosság még akkor is körülbelül háromszor közelebb áll a detektívtörténethez, ha az összes műfaji mintát azonos számú kötetre csökkentjük.)



6. ábra. Tudományos-fantasztikus művek 1771–1989 között, 88,3%-os pontossággal osztályozva.

Hogyan lehet egy ilyen hosszú és ezerarcú történettel bíró műfajt pusztán a szavak megszámlálása alapján modellezni? A *Frankenstein* a romantikus korszak szövege, amely még magára a „tudományra" sem hivatkozik túl gyakran. Kiderült azonban,

hogy vannak olyan szóbeli nyomok, amelyek 170 éven keresztül is fennmaradtak. A méretre való utalások („hatalmas", „messze", „nagyobb") nagyon jellemzőek a scifire, csakúgy, mint a nagy számok („ezrek"). A „földre" és az „emberi" dolgokra mutató öntudatos hivatkozásokat általában kiegészíti a „teremtmények" társasága, melyektől az emberiség különbözik, és gyakori a semleges birtokos névmás [its], mivel gyakran olyan szereplőkkel találkozunk, akik nem rendelkeznek könnyen felismerhető emberi nemmel. Ezt egyáltalán nem szánom a műfaj kimerítő leírásának – csak ízelítő a modellből.

A *6. ábra* egyik vizuálisan kiugró jellemzője, hogy a piros háromszögek 1950 után csekély mértékben csökkenni kezdenek. Korántsem biztos, hogy ez statisztikailag szignifikáns tendencia, de érdekes módon hasonló tendencia figyelhető meg a detektívtörténet esetében is *(3. ábra)*. Mindkét esetben úgy tűnik, hogy a háború utáni kötetek fokozatosan elveszítik az 1930-as és 1940-es évekre jellemző markáns műfaji elkülönülésüket. Van okunk óvatosságra: ez egy finom trend, amit akár a kiválasztás viszontagságai is okozhatnak, hiszen a század közepén megjelenő kötetek, ha rendelkeznek műfajspecifikus rajongói közönséggel, esetenként nagyobb valószínűséggel állnak rendelkezésre digitálisan. Az is elképzelhető, hogy egy idővonal közepén szereplő művek általában jobban illeszkednek a modellhez, mint az idővonal két végén lévő művek (bár ez a minta nem volt nyilvánvaló a korábbi, ezzel a módszerrel lebonyolított kutatásoknál). Ám ha ez egy valós tendencia, akkor elképzelhető, hogy Gary Wolfe hipotézise köszön itt vissza, mely szerint a fantasztikus műfajok határai az utóbbi időben instabillá váltak. „A fantasy elpárolog […] egyre csak sugároz, átjárja a körülötte lévő levegőt, és furcsa szagot áraszt az irodalmi légkörben."[38] Vannak azonban kisebb különbségek Wolfe téziséhez képest, és a „mágikus realizmus" kapcsán megfogalmazott, idevágó megállapításokhoz képest is. Úgy tűnik, az ezekben a modellekben megfigyelhető váltás közelebb vitte a műfajspecifikus fikciót az irodalmi mező további részeihez. Az ezzel ellentétes tendencia – hogy a műfaji trópusokat a fősodorba tartozó irodalmi szerzők játékosan felhasználják – még nem különösebben látható. Talán 1990 után válna láthatóvá.

A másik, ami esetleges várakozásaink ellenére sem jelenik meg ebben a modellben, az a műfaji konvenciók fokozatos megszilárdulása, pedig a sci-fi szakértői nem kevés időt töltenek e folyamat vizsgálatával. E műfaj történetének kutatói ritkán hajlandóak akkora jelentőséget tulajdonítani Verne és Shelley műveinek, mint amennyit a detektívtörténetek kutatói tulajdonítanak Poe-nak. A történeti áttekintések jelentős részének az a narratív alapvetése, hogy a sci-fi eredetileg csökevényes jelenség volt (amely elszórtan jelent meg az utópia, a planetáris románc stb. műfajokban), míg bizonyos 1925 és 1950 közötti ponyvamagazinok és antológiák új köntösbe nem öltöztették és új irányokat adtak neki. Hugo Gernsback *Amazing Stories* (*Csodálatos történetek,* 1926) című műve gyakran központi szerepet játszik. Wolfe például azt mondja, hogy a „sci-fi, annak ellenére, hogy az egész tizenkilencedik században élő erőteljes örökségre épít, 1926 után lényegében *megtervezett* műfaj lett".[39] De még e pont után is „a sci-fi regény kitartóan elkerülte a rejtélyes történetekhez és a westernekhez hasonló műfaji koherenciát", míg 1943-ban meg nem jelent a *Pocket Book of Science*

[38] Gary K. Wolfe, *Evaporating Genres,* viii.
[39] Uo., 34.

 DOI 10.31400/dh-hun.2021.4.2492

Fiction (Sci-fi zsebkönyv).[40] Egyik ilyen kritikus megszilárdulási pillanat sem látható a modellben. A nyelv tekintetében a Verne és Gernsback közötti fél évszázad (1875–1925) ugyanolyan koherensnek és a fikció más formáitól ugyanannyira eltérőnek tűnik, mint az 1926 utáni időszak. Természetesen lehetséges, hogy a modell rossz. Talán a sci-fi pusztán nyelvi alapú megkülönböztethetősége nem olyan fontos, mint a megszilárdulás egyéb formái. De az is lehetségesnek tűnik, hogy a tudományos-fantasztikus irodalomtörténet-írást indokolatlanul lenyűgözte a Gernsback által bevezetett „tudományos-fantasztikus" kifejezés vagy a ponyvakorszak romantikája vagy bizonyos technológiák (például az űrrepülés) szimbolikus jelentősége, és ez összességében háttérbe szorította azt a tényt, hogy a műfaj alapvetően összhangban áll a tudományos románc korábbi hagyományaival.

Mi a tanulság?

Az ebben a cikkben összegyűjtött bizonyítékok három meglévő műfajelméletet vonnak kétségbe. E célpontok közül talán Franco Moretti feltevése a legkevésbé fontos, mely szerint a műfaj egy generációs ciklus: Moretti ezt eleve szabadkozva, afféle spekulációként kínálta fel, és csak néhány más kutató fogadta el.[41] Az a feltevés, hogy a műfaji határok fokozatosan „megszilárdulnak" a huszadik század elején, már komolyabb kérdés. Az, hogy a ponyva adott formát a korábban „koherenssé válni képtelen" sokszínű hagyománynak, nagyon bevett a sci-fi kutatásban.[42] Nem gondolom, hogy ezt az állítást már megcáfoltam volna, de a nyelvi modellek mindenesetre a bizonyítékok feltűnő hiányát mutatják: a sci-fi jellegzetes nyelvezete, úgy tűnik, már a megszilárduló intézmények előtt kialakult. A harmadik műfajelmélet, amelyet megkérdőjeleztem, az az újabban népszerű elképzelés, mely szerint a műfajok története nem több, mint egy egymástól eltérő kulturális formákat összekötő genealógiai szál. A prediktív modellek nyíltan megkérdőjelezik ezt az állítást. Ha az 1930 előtti detektívtörténeteken betanított modell később is fel tudja ismerni a detektívtörténeteket (és a krimiket), akkor a műfajt az irodalmi mező többi részétől elválasztó különbségek viszonylag stabilak.

De újra hangsúlyoznom kell, hogy a stabil műfaji *határ* nem ugyanaz, mint az adott határon belül található tartalmak stabil *definíciója*. Bár a műfajokhoz kapcsolódó konkrét szavak gyakran lenyűgözőek, az erre vonatkozó részleteket szándékosan mellékesen kezeltem itt, hogy elkerüljem a definíciós állításokat. Sőt, ez a tanulmány éppen azért támaszkodik prediktív modellekre, mert így a definíció problematikáját elkerülve inkább a műfajjal kapcsolatos kortárs szkepticizmus alapját képező, óvatos, nominalista alapvetésekből indulhattam ki. A prediktív modellek például könnyen magukba építik azt az elképzelést, hogy a műfajok egyetlen meghatározó jellemző helyett számos tulajdonságból álló „családi hasonlóságokat" foglalnak magukba.[43] A prediktív modellek továbbá azzal a feltételezéssel is kifejezetten jól összeegyeztethetők, mely szerint a műfajok nem feltárásra váró formai mélystruktúrákból, hanem egymással

[40] Uo., 21.
[41] Roberts, „A Brief Note on Moretti."
[42] Wolfe, *Evaporating Genres,* 21.
[43] Kincaid, "On the Origins of Genre."

versengő, szubjektív „címkézési" cselekményekből állnak.[44] A jelen cikk azért épít a számítógépes módszerekre, mert a segítségükkel tudunk ilyen típusú pluralisztikus és távlati alapokra támaszkodni. Máskülönben (számomra legalábbis) nehéz lett volna jellemezni és összehasonlítani a sok különböző megfigyelő által azonosított összetett családi hasonlóságok erejét.

De még egy távlati alapvetésekből kiinduló vizsgálat is képes lehet végül azt felfedni, hogy az egymással versengő címkézési cselekmények valójában bizonyos esetekben hallgatólagosan összeegyeztethetők. És az is előfordulhat, hogy még az időbeli változékonyság alaptételéből kiinduló vizsgálatok is azt mutatják ki végül, hogy egyes műfajok határai akár másfél évszázadon keresztül is stabilak maradtak. Véleményem szerint ezt láttuk a detektívtörténetek és a sci-fi esetében (bár talán mindkét műfaj esetében igaz, hogy a huszadik század végére kezdenek „elpárologni").

Nem várhatunk el minden esetben ugyanolyan stabilitást. A gótika nevű jelenségcsoport például kevésbé hajlamos az összeolvadásra. Számos tizenkilencedik századi műfaj (például a Newgate-regény és a szenzációstörténetek) pedig, úgy tűnik, éppen olyan rövid életű, mint ahogy azt Moretti állította. Még a sci-fi is valamivel változékonyabb, mint a detektívtörténet. Ha lehet valamilyen fő következtetést levonni ebből a tanulmányból, az az lenne, hogy a „műfajnak" nevezett entitásoknak különböző típusai vannak, életciklusaik különbözőek lehetnek, és különböző fokú szöveges koherenciát mutathatnak. Az irodalomtudósok intuitíve már elismerték ezt például a „műfajok" és a „módok" megkülönböztetése révén. E bináris megkülönböztetés azonban nem tudja kielégítő módon leírni az itt feltárt történelmi folytonosságot. Bár ez a cikk elutasítja Franco Moretti azon feltételezését, hogy a műfajok generációkhoz hasonló ritmust követnek, azt hiszem, egyúttal igazolja azt az átfogóbb állítását, mely szerint a kvantitatív módszerek rugalmasabb és az általunk vizsgált anyag összetettségére jobban reagáló leíró forrásokat adhatnak az irodalomtudósok kezébe.

Fordította: Maczelka Csaba

[44] Rieder, „On Defining SF, or Not," 192–193.

 DOI 10.31400/dh-hun.2021.4.2492

The Life Cycles of Genres

Critics disagree even about the kind of thing a "genre" is. For instance, are genres defined mostly by textual form or by social reception? Are they permanent categories, century-sized historical constructs, or short-lived generational phenomena? This essay tries to address some of these persistent debates with recent quantitative methods. It shows, first, that social and textual definitions of genre can align fairly well: genres defined by the judgments of specific readers can also be recognized by textual clues. Building on that success, it investigates controversies about the typical lifespan of a genre. It turns out that different models of a genre are often compatible, even if defined by the judgments of readers in different periods. For instance, the models that can recognize nineteenth-century "detective stories" and "scientific romance" also recognize twentieth century "crime fiction" and "science fiction." This evidence suggests that—while genres are contingent historical constructs—many of them are more durable than literary historians have been willing to claim. But this is not universally true. The Gothic, for instance, is just as loose and mutable as critics have typically believed.

Keywords:
machine learning, genre, literature, literary history

A) függelék: metaadatok

A modellezéshez használt metaadatok a GitHub repozitórium finalmeta.csv fájljában találhatók.

Minden egyes mű korlátlan számú „műfaji címkét" viselhet, amelyek különböző, az adott művel összekapcsolt csoportokat jelölnek.

Az alábbi táblázat a címkék jelentését mutatja be: összesen 962 szövegre vonatkoznak. A szövegek nagy része kötetnyi méretű, de előfordul néhány novella is.

címke	#szövegek	dátum	leírás vagy forrás
det100	89	1829–1941	*The First Hundred Years of Detective Fiction, 1841–1941*. 1973. Lilly Library, Bloomington, IN. `http://www.indiana.edu/~liblilly/etexts/detective/`
chimyst	146	1923–1989	A könyvtárosok által a „detektív" [detective] vagy a „rejtélyes történet" [mystery fiction] kategóriába sorolt művek a Chicago Text Lab gyűjteményében.
locdetmyst	45	1832–1922	A könyvtárosok által „detektív- és rejtélyes történet" [detective and mystery fiction] kategóriába sorolt művek a HathiTrust gyűjteményében.

locdetective	16	1865–1912	A könyvtárosok által a „detektívek” [Detectives] kategóriába sorolt művek. Gyakran esetnapló-jellegű történetek.
crime	2	1972–1974	A könyvtárosok által a „krimi” [crime fiction] kategóriába felvett, de a „detektívtörténetek” [detective fiction] kategóriából kihagyott művek.
cozy	10	1920–1952	Olyan művek, amelyek szerzői a *The Mystery Readers' Advisory: The Librarian's Clues to Murder and Mayhem* című kötetben (John Charles, Joanna Morrison és Candace Clark, Chicago: ALA, 2002) vidéki házban játszódó rejtélyes történetek íróiként szerepelnek.
hardboiled	10	1929–1970	Geoffrey O'Brien, *Hardboiled America* (New York: Van Nostrand Reinhold, 1981) című kötetének melléklete.
newgate	7	1828–1840	Keith Hollingsworth, *The Newgate Novel*, Detroit 1963.
sensation	14	1860–1880	„The Sensation Novel,” Winifred Hughes, in Patrick Brantlinger and William B. Thesing, eds., *A Companion to the Victorian Novel* (Blackwell, 2002).
lockandkey	10	1800–1903	A *The Lock and Key Library: Classic Mystery and Detective Stories* című, Julian Hawthorne által szerkesztett (New York, 1909) antológiába beválogatott szövegek. Olyan írók is szerepelnek itt, mint Dosztojevszkij, akit valószínűleg még 1909-ben sem tekintettek krimiírónak, a cikkben használt „detektívtörténet” modellbe nem vettük fel.
pbgothic	96	1764–1988	David Punter and Glennis Byron, *The Gothic* (Malden: Blackwell, 2004), „Kronológia”.
stangothic	21	1791–1834	A Stanford Literary Lab metaadataiban a „gótikus irodalom” [Gothic] címkével ellátott művek kisebb halmaza.
lochorror	4	1818	A könyvtárosok által a „horror” [horror] kategóriába sorolt művek a HathiTrust gyűjteményében.
chihorror	23	1933–1989	A könyvtárosok által a „horror” [horror] kategóriába sorolt művek a Chicago Text Lab gyűjteményében.
locghost	28	1826–1922	A könyvtárosok által a „kísértethistóriák” [ghost stories] kategóriába sorolt művek.

 DOI 10.31400/dh-hun.2021.4.2492

locscifi	21	1836–1909	A könyvtárosok által a „sci-fi" [science fiction] kategóriába sorolt művek a HathiTrust gyűjteményében.
chiscifi	144	1901–1989	A könyvtárosok által a „sci-fi" [science fiction] kategóriába sorolt művek a Chicago Text Lab gyűjteményében.
femscifi	9	1818–1922	Ockerbloom, Mary Mark. 2015. „Pre-1950 Utopias and Science Fiction by Women."
anatscifi	36	1771–1922	Stableford, Brian. 2004. „The Emergence of Science Fiction" és „Science Fiction Between the Wars." In *Anatomy of Wonder,* ed. Neil Barron, 5th edition, 3–44.
chiutopia	13	1920–1976	A könyvtárosok által az „utópiák" [Utopias] kategóriába sorolt művek a Chicago Text Lab gyűjteményében, a jelen tanulmányban nincs beépítve a „sci-fi" kategóriába.
chifantasy	53	1901–1989	A könyvtárosok által a „fantasy" [fantasy] vagy „fantasztikus irodalom" [fantasy fiction] kategóriába sorolt művek, a jelen tanulmány célkitűzéseinek megfelelően nincs beépítve a „sci-fi" kategóriába.
–	–	–	–
juvenile	23	1904–1922	Ifjúsági közönségnek szánt művek; kigyűjtve, de a jelen tanulmány nem használja.
drop	33	1838–1922	Olyan művek, amelyeket végül nem használtam fel, de az átláthatóság érdekében meghagytam a metaadatokat. Leggyakrabban az ifjúsági jelleg miatt maradtak ki.
random	169	1769–1922	A HathiTrust Digital Library anyagából véletlenszerűen, a NEH [National Endowment for the Humanities] által finanszírozott „Understanding Genre in a Collection of a Million Volumes" című projekt keretében kidolgozott, fikcióval kapcsolatos metaadatok alapján kiválasztott művek. A „véletlenszerű kiválasztás" ebben az esetben azt jelenti, hogy a kötetek véletlenszerűen lettek kiválasztva, ám a szerző jóváhagyta vagy elutasította a kiválasztott szövegeket, hogy elkerülje a fikciós szövegeket, klasszikus költészetet, ifjúsági műveket tartalmazó kóbor köteteket.

chirandom	202	1920–1989	A Chicago Text Lab anyagából véletlenszerűen kiválasztott művek. A kiválasztás ebben az esetben közelebb állt a valódi véletlenszerű kiválasztáshoz.[45]
teamred	484	1760–1989	A hitelesség ellenőrzésére véletlenszerűen kiválasztott szerzők.
teamblack	500	1764–1989	A hitelesség ellenőrzésére véletlenszerűen kiválasztott szerzők.
stew	224	1764–1989	Véletlenszerűen kiválasztott, a gótikus irodalom [Gothic], a sci-fi [science fiction] és a krimi/detektívtörténet [crime/detective] hagyományát kiegyensúlyozottan reprezentáló kötetek halmaza, célja egy szörnyű műfaji ragu elkészítése.

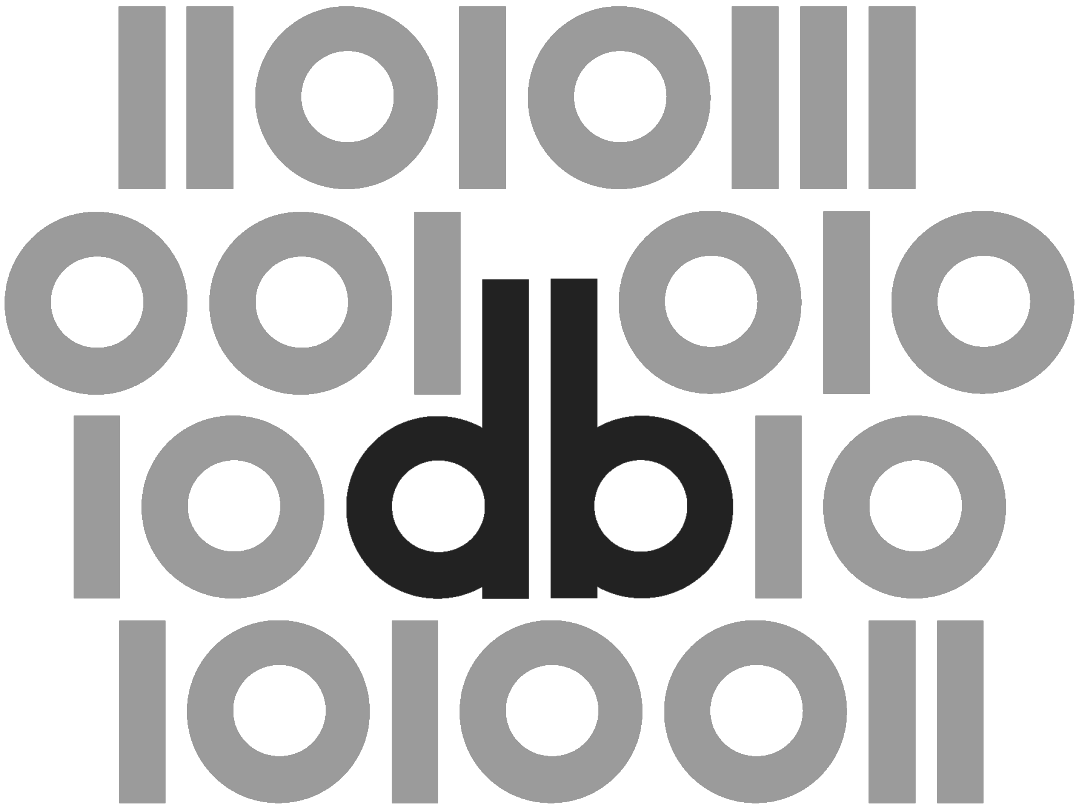
B függelék

A B függelék az a GitHub repozitórium, amely a projekthez tartozó kódokat, adatokat és metaadatokat tartalmazza, és az alábbi címen érhető el:

`https://github.com/tedunderwood/fiction`

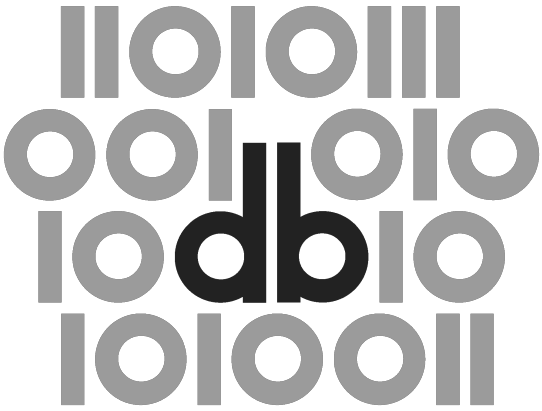
[45] Megjegyzendő, hogy a két „véletlenszerű” [random] címke más műfaji címkékkel együtt is előfordulhat. A véletlenszerűen kiválasztott kötetek például a „chimyst” kategóriának is tagjai lehetnek – ilyen esetekben csak akkor hagytam ki az adott művet a negatív (kontroll) halmazból, ha a „chimyst” szerepelt a pozitív halmazban.

 DOI 10.31400/dh-hun.2021.4.2492



Digitális Bölcsészet
2021., negyedik szám

<DIGITÁLIS BÖLCSÉSZET>



4 (2021)

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.2021.4

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.

Ez a mű a Creative Commons *Nevezd meg! – Ne add el! – Így add tovább! 2.5 Magyarország Licenc* (`http://creativecommons.org/licenses/by-nc-sa/2.5/hu/`) feltételeinek megfelelően felhasználható.

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

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

<TANULMÁNYOK>

Szemes Botond

ELTE BTK Irodalomtudományi Doktori Iskola

boboszemes@gmail.com

Kidolgozott viszonyok

A tagmondatkapcsolatok automatikus azonosításának hasznosíthatósága a stilisztikában és az irodalomtörténet-írásban*

A tanulmány egy olyan módszer kidolgozására tesz kísérletet, amely lehetővé teszi a magyar írott nyelvben a tagmondatkapcsolatok különböző típusainak automatikus azonosítását a kötőszavak és a relatív névmások mondatban elfoglalt helyzetének elemzésén keresztül. Ez a módszer új távlatokat nyithat meg a stilometriai kutatások számára: a kötőszavak és a névmások egyfelől mint funkciószavak statisztikailag megfelelő mennyiségű adatot szolgáltatnak a szövegek mélystruktúráinak feltérképezéséhez, másfelől tartalmas – grammatikai – jelentést is hordoznak (a tagmondatok között létesülő kapcsolatokról), ezáltal az eloszlásukra vonatkozó eredmények könnyen értelmezhetők. Az egyes típusok relatív gyakoriságának vizsgálatával ugyanis lehetővé válik, hogy meghatározzuk az egy szövegre vagy szövegcsoportra legjellemzőbb/legkevésbé jellemző kapcsolattípusokat. Ennek köszönhetően a regények stílusa és poétikája a mondat szintjén válik megragadhatóvá, miközben a szövegeknek olyan topológiai-logikai karaktere is feltárulhat, amely az olvasás során általában reflektálatlan marad. A tanulmány 100 magyar regény vizsgálatán keresztül tesz kísérletet az egyes szövegek mondatstilisztikai elemzésére, valamint stílustörténeti tendenciák és poétikai hagyományok azonosítására a magyar regényirodalomban.

Kulcsszavak:
tagmondatkapcsolat, kötőszó, stilometria, irodalomtörténet, magyar regény

Bevezetés

A tanulmány 100 kanonikus, 1832 és 2005 között megjelent magyar regény mondatstilisztikai vizsgálatára vállalkozik. Ennek során kvantitatív és algoritmikus eljárásokat

* A kutatásban nyújtott segítségükért köszönettel tartozom a krakkói Computational Stylistic Group és az ELTE BTK Digitális Bölcsészet Tanszék munkatársainak (elsősorban Indig Balázsnak, Szlávich Eszternek és Bajzáth Tímeának); valamint Melhardt Gergőnek. Köszönöm továbbá a folyóirat bírálóinak a korábbi verziókhoz fűzött hasznos megjegyzéseit is!

részesít előnyben, hiszen ezektől a szövegek olyan jellemzőinek a feltárása remélhető, amelyek nem vagy csak kevéssé reflektáltak az olvasás folyamatában. Az alábbiakban ennek megfelelően egy olyan módszert ismeretetek, amely az összetett mondatok különböző típusainak gyakorisági vizsgálatát teszi lehetővé, kiindulópontként szolgálva az irodalomtörténeti, illetve az egyes regények nyelvi szerveződésére vonatkozó megállapítások számára.

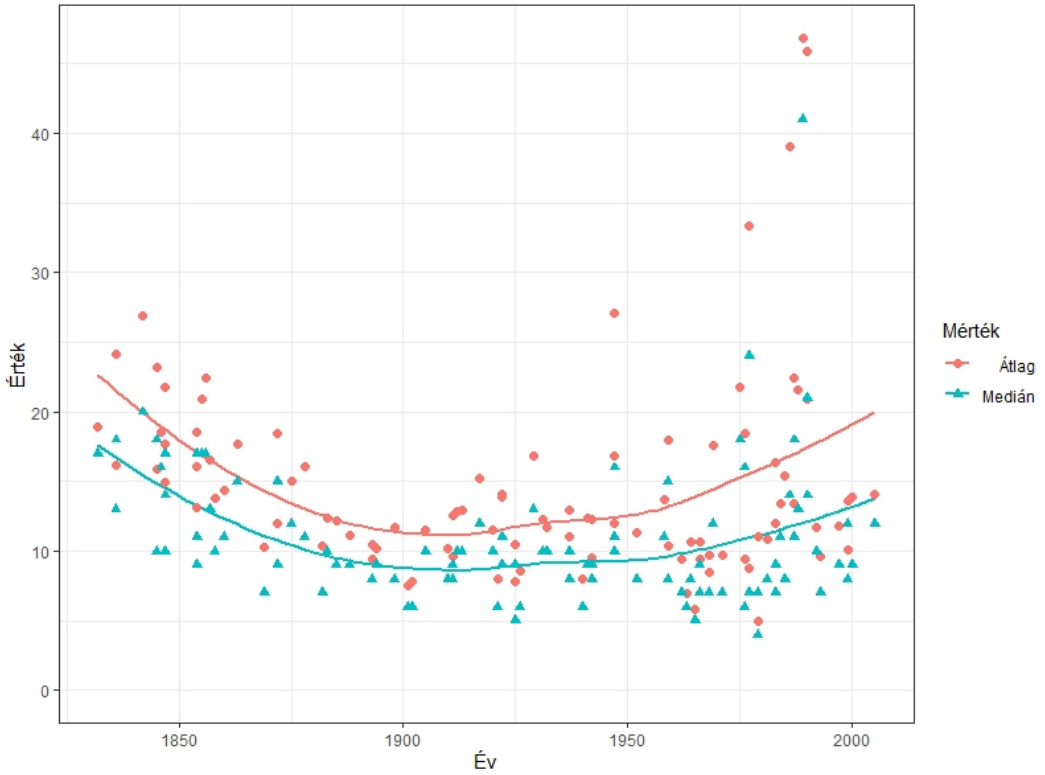
Az ezt megalapozó előzetes kutatás szintén hasonló megközelítést érvényesített, amikor a szövegeket mondatokra, majd a mondatokat szavakra bontva számította ki 100 kanonikus, 1832 és 2005 között keletkezett magyar regény[1] mondathosszúságának átlagát és mediánját. Az eredményeket bemutató ábrákon egyértelmű tendenciák rajzolódnak ki: már ezekben a tendenciákban is tetten érhetők a magyar regény stílustörténetének egyes csomópontjai és főbb változásai. Az *1.1. ábrá*n a korpusz egészét felölelő, az *1.2. ábrá*n pedig Jókai Mór 23 regényét tartalmazó, 1846 és 1894 közötti időszakban láthatjuk a szavak számában meghatározott mondathosszúságok átlagát és mediánját, valamint az adatpontokra illesztett regressziós görbét.[2] A 19. században kirajzolódó csökkenő tendencia – amely Jókai Mór életművében is megfigyelhető – a magyar prózahagyomány stiláris átalakulásán túl a sajtó szerepének megerősödésével, az új íróeszközök elterjedésével, valamint az írás és az olvasás oktatására vonatkozó iskolai reformokkal hozható összefüggésbe.[3] A huszadik század utolsó harmadában kevésbé egységes a megfigyelhető tendencia, inkább a különböző prózapoétikák együttes jelenlétéről beszélhetünk: az extrém értékeket mutató hosszúmondatos szerzők regényei (pl. Krasznahorkai László, Nádas Péter, Kertész Imre, Konrád György egyes művei) mellett kifejezetten rövidmondatos prózák (pl. Mándy Iván, Mészöly Miklós egyes művei) is megtalálhatók a korszakban.

[1] A korpuszok részletes leírását lásd az alábbi, *A korpusz* című részben.

[2] Ez a görbe olyan becsült értékek alapján jön létre, amelyek y valószínűsített helyzetét adják meg x változóhoz (jelen esetben az évszámokhoz) képest úgy, hogy a becsült értékek és a valódi adatpontok közti különbség minimális legyen. Ennek a függvényszerű viszonynak a kiszámítását statisztikai egyenletek segítségével végezhetjük el, amely műveletek az R nyelvhez írt *ggplot* csomagba implementálva a vizualizáció során futnak le. Jelen esetben – a később bemutatott ábrákhoz hasonlóan – a használt kód alapbeállítása a *polinomiális regressziós illesztést* (a kódban: „loess") alkalmazta. Vö. Hadley Wickham, Danielle Navarro and Thomas Lin Pedersen, *ggplot2: Elegant Graphics for Data Analysis* (New York: Springer, 2016), `https://doi.org/10.1007/978-3-319-24277-4`.

[3] A mondathosszúságok vizsgálatát és a tendenciák elemzését lásd: Szemes Botond, „Mondathosszúság és irodalomtörténet: 100 magyar regény szövegstatisztikai elemzése," *Literatura* 46, 3. sz. (2020): 335–367.

 DOI: 10.31400/dh-hun.2021.4.2415



1.1. ábra. Mondathosszúság: átlag és medián, 1832–2005.

1.2. ábra. Mondathosszúság – Jókai Mór. Átlag és medián, 1846–1895.

Már tehát egy ilyen banális szempont (a mondatok átlagos hosszúsága) mentén létrehozott vizualizációk is képesek hírt adni irodalom- és stílustörténeti mozgásokról, és segíthetnek a rövid- és hosszúmondatos prózapoétikák elkülönítésében. Differenciáltabb kérdés azonban csak akkor merülhet fel, ha ezeknek a poétikáknak a hasonlóságaira és különbségeire, valamint a mondatok belső szerveződésére is rákérdezünk, ugyanis nem feltétlenül jelenti ugyanazt a „hosszúmondatosság" különböző alkotók vagy korszakok esetében. Annak érdekében, hogy részletesebb képet kapjunk e 100 regény stílusáról, vizsgálhatjuk többek között a szövegek szókincsgazdagságát vagy a különböző szófajok eloszlását is (pl. nominális és verbális stílusok megkülönböztetésekor),[4] ám jelen dolgozat elsősorban a mondatszerkezetek, pontosabban az összetett mondatok kötőszavakkal kidolgozott, így nyelvileg hozzáférhetővé tett fajtáinak statisztikai elemzésén keresztül kívánja felvázolni a magyar regény stílustörténetének egy szeletét. Azért a mondatstilisztikai szempontot részesítem előnyben, mert kiinduló hipotézisem szerint, amennyiben meg tudjuk adni egy szövegben vagy szövegcsoportban a tagmondatkapcsolatok relatív gyakoriságát, akkor azt is meg tudjuk határozni, hogy mely kapcsolattípusok jellemzők leginkább egy adott szövegre/szövegcsoportra, amin keresztül annak rejtett topológiai-logikai[5] karakterére is következtetni tudunk. Könnyen belátható ugyanis, hogy más annak a szövegnek a karaktere és a világhoz fűződő viszonya, amelyik például arányosan több feltételes tagmondatkapcsolatot dolgoz ki, mint amelyikben inkább az ellentétes kapcsolatok nagy száma figyelhető meg – ahogyan egy kapcsolattípus hiánya is sokat elárulhat a vizsgált műről.

Az ilyen következtetéseknek fontos előképei a magyar stilisztika kvantitatív módszereken nyugvó korábbi (elsősorban 70-es, 80-as évekbeli) munkái, amelyekben szintén erőteljesen érvényesül a mondatokra irányuló elemzési szempont:

> a legjellemzőbb képet a szövegről mondatainak minősítése adja, hiszen a mondatok közlésegységek, meglétük elengedhetetlen szövegsajátság, a szöveg alatt közvetlenül elhelyezkedő szinten állnak. Kvantitatív feldolgozásuk nem jelenti azt, hogy a szövegből kiragadott, attól elszigetelve szemlélt egyedi mondatok egyedi »stilisztikai szerepét« néznénk, hanem éppen egészében tekintjük át a szöveget, különféle mondatainak, mondatszerkezeti formáinak előfordulási arányaiban.[6]

Ezek a munkák bizonyították, hogy az összetett mondatok a szövegek legalapvetőbb típusai, és még a „rövidmondatosként" meghatározott prózastílusban is túlsúlyban

[4] Ilyen irányba tett kísérletekhez vö. Uo., 360–365.

[5] „Topológia" alatt egy hálózat elemeinek elrendezését értem. A „logika" kifejezést nem szigorú értelemben használom, amely így elsősorban nem a mellérendelő kapcsolatok formális logikai megfelelőire vonatkozik, csupán a tagmondatok között kidolgozott viszonyt jelöli. Éppen ezért pontosabb megfogalmazásnak tűnik a mondatok „diagrammatikus" karakteréről beszélni. Vö. Frederik Stjernfelt, „The Extension of Peircean Diagram Category: Charting the Implications of a Diagrammatical Revolution in Semiotic," in Olga Pombo and Alexander Gerner, eds., *Studies in Diagrammatology and Diagram Praxis,* 57–73 (London: College, 2010).

[6] V. Raisz Rózsa, „A kvantitatív módszer mondatstilisztikai alkalmazhatóságáról," *Magyar Nyelvőr* 110, 2. sz. (1986): 161–169, 164. Lásd még Deme László, *Mondatszerkezeti sajátságok gyakorisági vizsgálata* (Budapest: Akadémiai Kiadó, 1971); Nagy Ferenc, *Kvantitatív nyelvészet* (Budapest: Tankönyvkiadó, 1972); Zsilka Tibor, *Statisztika és stilisztika* (Budapest: Akadémiai Kiadó, 1974).

 DOI: 10.31400/dh-hun.2021.4.2415

szerepelnek.[7] Ugyanakkor ezeket a kvantitatív vizsgálatokat főként a stílustörténeti besorolás motiválta (például az impresszionista stílusjegyek felismerése), amely szempont ma kevésbé tűnik termékenynek az irodalomtörténet-írás horizontjából. Innen nézve a tanulmány következtetéseire a legnagyobb hatást a magyar stíluskutatás jelentős, de napjainkban alig idézett alakjának, Herczeg Gyulának a munkássága gyakorolta, aki szintén a mondatszerkezetek változásain keresztül tett kísérletet a magyar próza történetének megírására, és a nyelvi jellegzetességekből érvényesen következtetett egy szerző, vagy korszak világképére is, érdemben hozzájárulva a szövegek értelmezéséhez[8] – ám még digitális eszközök hiányában sokszor kis méretű korpusz rövid szemelvényeire és az olvasói intuícióra hagyatkozva. A mostani eljárás a Herczeg által kidolgozott mondatstilisztikai megközelítést nagy mennyiségű adat statisztikai feldolgozásával és vizualizációjával egészíti ki.

Ugyanakkor meg kell említeni a digitális kereséseken nyugvó kvantitatív elemzések egy sajátosságát is. Egy ilyen elemzés ugyanis más logikát kíván meg, mint a hagyományos mondatstilisztikai megközelítés. Jan Rybicki, napjaink számítógépes stíluskutatásának kiemelkedő művelője például azt vetette fel a kutatásról tartott egyik előadásom vitájában, hogy minek fáradozni annyit a tagmondathatárok (*A módszer* című részben ismertetett) azonosításával és a kidolgozott tagmondatkapcsolatok más szerkezetektől való elkülönítésével, amikor egyszerűbben lenne mérhető pusztán a kötőszavak gyakorisága. Ezáltal több adat állna rendelkezésre, amelyek segítségével statisztikailag is megalapozottabb következtetések tehetők, illetve a kutatás amúgy sem előre meghatározott grammatikai kategóriák azonosítását, hanem a szövegek említett topológiai-diagrammatikus karakterének feltárását tűzte ki céljaként, ez a kérdésfeltevés pedig nem kizárólag a tagmondatkapcsolatok gyakoriságára vonatkozik. A hozzászólás sokat alakított a kutatás előfeltevésein; mindazonáltal úgy gondolom, hogy egyszerűen a kötőszavak összeszámlálása a poliszém jelentések és a szemantikailag is jól elkülöníthető grammatikai szerkezetek miatt valójában nem vezetne értelmezhető eredményre. Az előre definiált nyelvészeti kategóriáktól esetenként elrugaszkodó digitális bölcsészeti logika ugyanakkor megjelent a keresések módszertanának véglegesítéseben, és a tagmondatkapcsolatok azonosításakor megengedő módon jártam el, azaz a tagmondatkapcsolatokhoz hasonló grammatikai jelentéseket kidolgozó mondatrészek közötti kapcsolatok egy részét nem zártam ki a keresési találatok közül, vállalva ezzel azt, hogy a kialakított módszer csak korlátozásokkal alkalmazható kizárólag a tagmondatkapcsolatok azonosítására.

[7] „Kiderül a vizsgálatból, hogy az úgynevezett »rövidmondatos« stílust sem lehet egyoldalúan egyszerű mondatosnak, esetleg egyoldalúan mellérendelőnek elképzelni. Még Gárdonyi elvszerűen megalkotott rövid mondatos szövegeiben is csaknem 50%-nyi az összetett mondat, és kevés olyan összefüggő részt lehet találni a művekben, amelyben »nagyobb távon« csak egyszerű mondatot ismer fel a kereső." V. Raisz Rózsa, „Mondatgyakorisági vizsgálatok néhány századforduló kori szépprózai szövegben," *Magyar Nyelv* 72, 2. sz. (1976): 202–209, 208.

[8] Főbb művei e tekintetben: Herczeg Gyula, „A mondatstilisztikai kutatás mint módszer," in Telegdi Zsigmond, szerk., *Általános nyelvészeti tanulmányok XI: A szöveg megközelítései,* 127–142 (Budapest: Akadémiai Kiadó, 1976); Uő., *A régi magyar próza stílusformái* (Budapest: Tankönyvkiadó, 1985); Uő., *A XIX. századi magyar próza stílusformái* (Budapest: Tankönyvkiadó, 1981); Uő., *A modern magyar próza stílusformái* (Budapest: Tankönyvkiadó, 1975).

A korpusz

Labádi Gergely a korai magyar regény adatbázisának létrehozásakor György Lajos korszakolását követve az 1730-tól 1830/40-ig tartó időszakban jelöli ki a magyar regény kialakulását vagy előtörténetét.[9] Vaderna Gábor és Szilágyi Márton irodalomtörténeti leírása szerint „az ekkor megalapozódó új [prózai] műfaj azonban eltér attól, amit később, az 1830-as évektől *regényként* határoz meg a korabeli kritika."[10] Éppen ezért a kiforrott magyar regényirodalom kezdetét, az 1830-as éveket tettem meg a korpusz kiindulópontjaként. A végpontot Nádas Péter *Párhuzamos történetek* című 2005-ös regénye jelenti, kizárva a kutatásból a kortárs magyar irodalom elmúlt évtizedeit, mivel nehéz lenne olyan szempontot találni, amelynek segítségével csupán pár kanonikus mű kiválasztható a jelenlegi magyar irodalomból. Kanonikus művekre azért van szükség, mert a 19. század első feléből jóformán csak ezek érhetők el digitalizált formában, ami így meghatározta a korpusz egészének szerkezetét. Ugyanígy az ebből a korszakból elérhető alkotók száma szabja meg azt is, hogy későbbi időszakokból hány szöveget vettem fel a kutatásba, mivel arányos időbeli eloszlásra törekedtem. Így minden évtizedből legkevesebb 3, legfeljebb 8 szöveg került a korpuszba. Ez összesen 100 regényt[11] és 173 évet, azaz átlagosan 1,7 évenként egy regényt jelent. Szintén az arányosság miatt egy írótól maximum 4, de inkább kevesebb regényt válogattam; ez összesen 58 szerzőt eredményezett, mindegyiknek átlagosan 1,7 regényét. Egy szerzőtől abban az esetben válogattam be 4 írást, ha azok között jelentős az időbeli távolság, hogy ezáltal vizsgálható legyen az is, hogy az egyes tendenciákat követi-e az adott alkotó életműve, vagy inkább időszaktól független szerzői „ujjlenyomatokról" beszélhetünk. Ugyanezért hoztam létre a Jókai Mór 23 regényét magagába foglaló alkorpuszt is. A kutatás egy későbbi szakaszában érdemesnek tűnik a vizsgált szempontokat más műfajú (kisprózai, nem fikciós, akár tudományos-értekező) szövegeket tartalmazó korpusz esetében is érvényesíteni.

A szövegek összetételét tehát az irodalomtörténeti konszenzuson túl a Magyar Elektronikus Könyvtár és a Digitális Irodalmi Akadémia adatbázisai határozták meg. Mivel egyik sem érvényesít határozott irodalomtörténeti koncepciót a digitalizáció során, ezért fordultam az ELTE BTK Digitális Bölcsészet Tanszék által egy nemzetközi projekt keretében összeállított regénykorpuszhoz, amely az 1850 és 1920 között megjelent magyar nyelvű prózairodalomból előre meghatározott szempontok szerint válogat,[12] valamint az Akadémia Kiadó *Magyar irodalom* című kézikönyvéhez és a magyar próza stílustörténetét tárgyaló korábbi monográfiákhoz (például Herczeg Gyula köteteihez). A korpuszt ezenkívül egy-két népszerű szerzővel egészítettem még

[9] Labádi Gergely, „A magyar regény adatbázisa," *Acta Historiae Litterarum Hungaricarum: Acta Universitatis Szegediensis* 32, 1. sz. (2016): 11–30. Lásd még: György Lajos, *A magyar regény előzményei* (Budapest: Akadémiai Kiadó, 1941).

[10] Szilágyi Márton és Vaderna Gábor, „A klasszikus magyar irodalom (kb. 1750-től kb. 1900-ig)," in *Magyar irodalom,* főszerk. Gintli Tibor, 313–641 (Budapest: Akadémiai Kiadó, 2010), 370.

[11] A regények listáját lásd a *Függelék* 1.1. pontjában.

[12] A projekthez hozzáférés: 2021.04.18, `https://www.distant-reading.net/eltec/`, amelynek magyar nyelvű adatbázisa elérhető az alábbi linken, hozzáférés: 2021.04.18, `https://github.com/COST-ELTeC/ELTeC-hun`. Az ebből az adatbázisból kinövő *Regénykorpusz* az összetett kereséseket lehetővé tevő keresőfelülettel elérhető az alábbi linken, hozzáférés: 2021.04.18, `https://regenykorpusz.elte-dh.hu/`.

 DOI: 10.31400/dh-hun.2021.4.2415

ki (például Rejtő Jenő, Gárdonyi Géza). Újabb szövegek digitalizációja és nagyobb, megbízható adatbázis kiépítése elsődleges feladata a hazai irodalomtudománynak; meglehet, a most bemutatott eredmények is árnyalhatók lennének egy szélesebb körű kutatás során.

A statisztika alapvető belátása, hogy minden eredmény a feldolgozott adatok közti viszonyt mutatja és csak bizonyos számú adat felett tekinthető ez a viszony reprezentatívnak. Az itt feldolgozott 100 regény semmiképpen sem elegendő ahhoz, hogy a magyar prózairodalom egészére nézve biztos következtetéseket vonhassunk le, inkább egy ma élő kánonra és a kanonikus szövegek történeti alakulására láthatunk rá általuk.

A módszer

A magyar helyesírásban a tagmondathatárok mondatközi írásjelekkel (vessző, pontosvessző, kettőspont, gondolatjel) jelöltek, így azonosításuk is könnyebben elvégezhető, mint más nyelvek esetében, amelyek nélkülözik az ilyen típusú jelölést. A tagmondatkapcsolatok azonosítása így a kutatásban a mondatközi írásjelek és kötőszavak vagy vonatkozó névmások kombinációjának azonosításán alapul: ezek a kombinációk reguláris kifejezésekkel könnyen detektálhatók. Felmerülhetnek más, összetettebb grammatikai jellemzők szerinti keresések is, amelyek szerkesztetlen szövegek (pl. lejegyzett beszéltnyelv), vagy más nyelvek esetében is használhatók lehetnek: ezek közül a legkifinomultabbnak a véges igék között elhelyezkedő kötőszavak azonosítása tűnik, hiszen egy tagmondat prototipikusan egy véges (lehorgonyzott)[13] igéből mint állítmányból és annak kidolgozásából áll. Ez a módszer ugyan általánosabb és kifinomultabb szempontot biztosít a tagmondatkapcsolatok felismeréséhez, ugyanakkor három nehézségbe ütközik. Egyrészt a véges igék felismerését a kutatásban használt *e-magyar* szóelemző[14] nem tudja hibátlanul elvégezni – ez különösen igaz a 19. század első felében keletkezett szövegekre –, másrészt több kötőszó gyakran nem a véges igék között helyezkedik el, hanem például a mondat elején, harmadrészt ez a szempont nem veszi figyelembe azokat az eseteket, amikor zéró kopulával alkotott névszói-igei állítmány építi fel a tagmondatot („Az a ház szép, amelyik magas"). Ez utóbbira, a zéró kopulás szerkezetek automatikus azonosítására ígéretes eljárások kifejlesztése kezdődött meg, amelyek a későbbiekben mind az *e-magyar,* mind a tagmondathatárok elemzésének hatékonyságát növelhetik.[15] A nehézségek miatt azonban a jelenlegi

[13] Episztemikus lehorgonyzás alatt a kognitív nyelvészetben azt a szemantikai műveletet értjük, amely működése révén a típusból példány jön létre (pl. a szótári igealakból teljes igei alak), tehát hozzáférhetővé válnak a referenciális jelenetre való vonatkozások az alaphoz viszonyítva a megnyilatkozó és a befogadó figyelemirányulásának és mentális feldolgozásának, valamint tudásának megfelelően. Vö. Ronald W. Langacker, *Foundations of Cognitive Grammar, Vol. I* (Stanford: Stanford UP, 1987). Az igék episztemikus lehorgonyzásához lásd Tolcsvai Nagy Gábor, *Nyelvtan: A magyar nyelv kézikönyvtára* (Budapest: Osiris Kiadó, 2017), 349–352.

[14] Váradi Tamás, Simon Eszter, Sass Bálint, Mittelholcz Iván, Novák Attila és Indig Balázs, „E-magyar – A Digital Language Processing System," in Nicoletta Calzolari et. al., eds., *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018),* 1307–1312 (Miyazaki: European Language Resources Association, 2018).

[15] Dömötör Andrea, Yang Zijian Győző és Novák Attila, *„Nesze semmi, fogd meg jól! Zéró kopulák automatikus felismerése neurális gépi fordítással,"* in *XVI. Magyar Számítógépes Nyelvészeti Konferencia,*

kutatás megmaradt a mondatközi írásjeleket követő, vagy a mondat elején álló kötőszavak/vonatkozó névmások reguláris kifejezésekkel történő azonosításánál, amely egy megbízhatóbb és pontosabb módszert jelent szerkesztett szövegeket elemezve.[16]

A reguláris kifejezéseknél az alábbi jellemzők meghatározása vált kiemelten fontossá: (1) a kötőszavak/névmások pozíciója, azaz az írásjeltől való lehetséges távolsága; (2) a poliszém kötőszók funkcióeloszlása és prototipikus jelentése, valamint (3) az adott kötőszó prototipikusan tagmondatok vagy mondatrészek között létesít-e inkább kapcsolatot. Ezeket a kérdéseket a korpuszból vett 100 mondatos minták kézi elemzésével igyekeztük megválaszolni Bajzát Tímea és Szlávich Eszter nyelvészek, a Digitális Bölcsészet Tanszék munkatársai segítségével.[17] Amennyiben egy kötőszó/vonatkozói névmás a 100 esetből legalább 60-szor egy típusú tagmondatkapcsolatot dolgozott ki, akkor az adott fajta alá soroltuk be, ellenkező esetben kizártuk a kutatásból. Például a *hogy* kötőszót vizsgálva a tagmondatkapcsolat típusa nem határozható meg egyértelműen, emiatt nem vettük fel a kutatásba. Egyedül abban az esetben kezeltük rugalmasan a kategorizációt, ha egy kötőszó egyértelműen beazonosítható grammatikai jelentést dolgoz ki, ugyanakkor formailag nem elkülöníthető, hogy mondarészek, vagy tagmondatok között szerepel mint kötőelem („[…] csak Lengyelországban lesz az emberből kiszolgáltatott, légüres térben kapkodó irodalmár, azaz semmi." [Spiró György, *Az Ikszek*]). Ekkor, ha az esetek több, mint 50%-ában (tehát 100-ból legalább 50-szer) tagmondatok közötti kapcsolatot dolgoz ki az adott kötőszó, akkor felvettük a kutatásba: ilyen volt például az *azaz* kötőszó a korpuszból vett 100-as mintát vizsgálva. A meglehetősen alacsony hibahatár kapcsán fontos itt is hangsúlyozni, hogy ugyan a kutatás a tagmondatkapcsolatok azonosítására irányul, ám valójában nem az előre meghatározott grammatikai kategóriák gyakoriságának, hanem a szövegek diagrammatikus szerveződésének elemzését tűzi ki célul, amelyhez elsősorban minél több adat előállítása szükséges. Így a hasonló grammatikai jelentést hordozó, de mondatrészek között létesülő viszonyokat megengedő módon kezeltem és nem minden esetben zártam ki a keresési eredmények közül, abból a hipotézisből kiindulva, hogy így a regények mondatainak domináns kapcsolattípusai jobban kidomborodhatnak a

MSZNY: Szeged, 2020. január 23–24, 385–398 (Szeged: Szegedi Tudományegyetem TTIK Informatikai Intézet, 2020).

[16] Az így kapott eredményeket ugyanakkor összevetettem a véges igék azonosításán alapuló módszer eredményeivel: ez utóbbi pontatlanabb, de arányait tekintve hasonló értékeket adott meg az egyes szövegekre vonatkozóan. Azaz egy-egy regény jellemzői ezen keresztül is kimutathatónak tűnnek. A véges igék azonosításán alapuló módszer eredményeit lásd a *Függelék* 1.2. pontjában.

[17] Az elemzéskor a tagmondatokat funkcionális keretben vizsgálva olyan egységként határoztuk meg, amelyek egy lehorgonyzott folyamattípus előhívásáért felelős, úgynevezett „magmondattal" rendelkeznek. Ehhez lásd Imrényi András és Kugler Nóra, „Mondattan," in *Nyelvtan,* szerk. Tolcsvai Nagy Gábor et. al., 663–889, (Budapest: Osiris Kiadó, 2018), 703, `https://doi.org/10.14232/JENY.2018.1.8`, valamint Imrényi András, *A magyar mondat viszonyhálózati modellje* (Budapest: Akadémiai Kiadó, 2013), `https://doi.org/10.1556/9789634544258`. A minták elemzésének eredményeit lásd a *Függelék* 2.1. pontjában. Az összetett mondat meghatározásához vö. „A több jelenetből, több elemi mondatból álló szerkezeteket összetett mondatoknak nevezi a szakirodalom. […] A tagmondatkapcsolat két (vagy több) jelenetnek nagyobb integráltságban megvalósuló közlését képviseli ahhoz képest, mintha a jelenetek elkülönülő mondatokban lennének megjelenítve. […] A tagmondatkapcsolatokban mindig bizonyos viszonytípusok valósulnak meg, a tagmondatkapcsolódási típusok grammatikalizálódott mintázatok." Imrényi és Kugler, „Mondattan," 808.

 DOI: 10.31400/dh-hun.2021.4.2415

keresések során. Fontos azonban, hogy a másik véglet, a megkötések nélküli, pusztán a kötőszókra vonatkozó és nagyobb adatbőséget ígérő keresések sem tűnnek célravezetőnek, hiszen a poliszém jelentések (az *így* mutatónévmási, a *bár* főnévi jelentése stb.) és a különböző szerkezetek közötti markánsabb eltérések miatt az így elért találatok már nem a szövegek egy jól körülhatárolható tulajdonságára vonatkoznának. A magyar helyesírás segítségünkre lehet ennek a megengedő, de korlátozásokat is érvényesítő szempontnak a gyakorlati alkalmazásában, amennyiben a mondatrészek között kidolgozott viszonyok csak abban az esetben jelöltek mondatközi írásjellel (az előbeszédben pedig szünettel), ha azok a tagmondatkapcsolatokhoz hasonló grammatikai jelentéssel bírnak („Hogy volt az embernek történelem előtti, azaz olyan korszaka, amelyről semmi, nemcsak írásban, de még szájról-szájra adott mondákban sem maradt fenn: az kétségen felül áll." [Móricz Zsigmond: *Légy jó mindhalálig*]). Kizárólag a tagmondatkapcsolatok azonosítása az itt bemutatott keresések szigorúbb korlátozásával hozható létre.

Fontos továbbá, hogy bizonyos kötőszavak csak közvetlenül az írásjelet követően dolgoznak ki tagmondatkapcsolatot *(és)*, míg mások a szórendi rugalmasság miatt nagyobb távolságban szerepelhetnek (pl. *ugyanis, azonban*). Sass Bálint az ilyen eseteknél a vesszőtől való 0–3 szó távolságot határozza meg a kötőszavak lehetséges előfordulásaként,[18] a minták elemzése azonban azt mutatta, hogy ennél is nagyobb távolságban, tulajdonképpen a következő tagmondatokat elválasztó vagy mondatvégi írásjelig terjedő szakaszban bárhol előfordulhatnak ezek a kötőszavak. Az egyes szavakra vonatkozó megállapításainkat a típusok jellemzésekor ismertetem. A kutatás tárgya a teljes mondat, amely a benne kidolgozott kapcsolat szerint sorolódik valamely típus alá. A többszörösen összetett mondatok minden abban előforduló típusba besorolhatók, például a „Válaszában nem találok ellentmondást, s így ha jól meggondolom, igaza volt" mondat egyszerre tartozik a kapcsolatos, a következtető és a feltételes típusokhoz. Azokat az tagmondatkapcsolatokat, amelyek nem kidolgozottak kötőszavak/vonatkozói névmások által, a vizsgálat figyelmen kívül hagyta, például a következőt: „A mondatban nem találtunk kötőszót, [tehát] kihagytuk a kutatásból." Ez a mozzanat az eljárás alapvető jellemzőjére világít rá: a kutatás valójában nem általánosan az összetett mondatokra vonatkozik, hanem az olyan összetett mondatokra, ahol a tagmondatkapcsolatok a kötőszavak/vonatkozó névmások által kidolgozottak. Egy ilyen kidolgozott kapcsolat jobban hozzáférhetővé teszi a tagmondatok közötti viszonyt, azaz nem létrehozza, „hanem profilálja azt, hozzáférhetővé teszi az összefüggés jellegét, és ezzel irányítja az értelmezői folyamatot".[19] A kidolgozatlan kapcsolatoknál ezzel szemben többféle értelmezés is elképzelhető, illetve a tagmondatok közötti viszony kevésbé kerül a résztvevők figyelmének előterébe. Ezen túl a kötőszavak azért is lehetnek különösen fontosak egy kvantitatív elemzés számára, mert mint funkciószavak[20] nagyon gyakoriak egy szövegben, és így statisztikailag megfelelő

[18] Sass Bálint, *Igei szerkezetek gyakorisági szótára: Egy automatikus lexikai kinyerő eljárás és alkalmazása,* doktori disszertáció (Budapest: Pázmány Péter Katolikus Egyetem, 2011), 36.

[19] Imrényi és Kugler, „Mondattan," 811.

[20] A számítógépes stíluskutatás erősen hagyatkozik a szavak olyan felosztására, amely szerint léteznek nyitott nyelvosztályba tartozó, tartalmas jelentéssel bíró szavak (prototipikusan főnevek, melléknevek, igék) és zárt nyelvosztályba tartozó (azaz alig bővülő, meghatározott készletet alkotó), pontos

mennyiségű adatot szolgáltatnak a szövegek mélystruktúráinak feltérképezéséhez; ugyanakkor – több más funkcionális szóval ellentétben – tartalmas (grammatikai) jelentést is hordoznak (az általuk kidolgozott kapcsolatról),[21] ezáltal az eloszlásukra és gyakoriságukra vonatkozó eredmények könnyen értelmezhetők. A magyar regény kvantitatív módszerek segítségével felvázolt stílustörténete a kötőszavaknak ezen a kettős tulajdonságán alapul.

A minták kiértékelését követően a következő típusokat és szavakat különítettük el, amely típusok csak módszertani szempontok miatt alkotnak élesen elhatárolt kategóriákat, valójában inkább a családi hasonlóság mintájára szerveződnek. Ugyanígy az egyszerű és összetett mondatok elkülönítése sem magától értetődő, hiszen egy kötőszós szerkezet mind a kettőben szerepelhet. A kutatás ezt a nehézséget az alább felsorolt kritériumokkal igyekezte áthidalni, miközben a hibaszázalékok kapcsán említett megengedő szempontot is szem előtt tartotta.

1. Kapcsolatos mellérendelés. A mondatközi írásjelet közvetlenül követő *és, sőt, s, valamint, majd, aztán, azután* szavak valamelyike – az írásjelhez való közelség oka, hogy ezáltal kiszűrhető az *aztán* partikulaként való szereplésének nagy része, illetve a mondatközi írásjelet nem közvetlenül követő esetek is, amelyek nem a tagmondatok, hanem mondatrészek viszonyára vonatkoznak. Ezenkívül a tagmondatok közti írásjelet 0–2 szó után követő *ráadásul* és az *egyrészt … másrészt*, *hol … hol* összetételek valamelyike.
   A *majd* határozószó ugyan nem tekinthető kötőszónak (hiszen tartalmas jelentése van, és valódi kötőszavak mellett is szerepelhet, megjelenése voltaképp nem befolyásolja az összetett mondat típusát), ám 100 olyan esetből, amikor más kapcsolatos kötőszó nincs a tagmondatban, 63-szor kapcsolatos kronologikus jelentésű viszony kidolgozásában vett részt.[22] Ehhez hasonlóan az *aztán* is 80-szor szerepel kronologikus kapcsolatos mellérendelésben, és csak 14-szer diskurzusjelölőként a mondatközi írásjelet követően.
   A *nemcsak… hanem* többtagú kötőszó 100 esetből csak 18-szor jelölt tagmondatok közötti kapcsolatot, a többi esetben mondatrészek viszonyát dolgozta ki, így nem képezi a vizsgálat tárgyát.
2. Ellentétes mellérendelés. Egy mondatközi írásjel és egy másik írásjel között álló *azonban, ellenben, viszont, ámde, csakhogy, ellenkezőleg, mindazonáltal, ugyan-*

jelentést nélkülöző, erősen grammatikalizálódott szavak, azaz funkciószavak (pl. névelők, kötőszók, prepozíciók stb.) Vö. Daniel G. Morrow, „Grammatical morphemes and conceptual structure in discourse processing," *Cognitive Science* 10, 4. sz. (1986): 423–455, https://doi.org/10.1207/s15516709cog1004_2.

[21] A *hogy* kötőszó ebből a szempontból is kivételt jelent, hiszen „a tartalomkifejtő mellékmondat élén álló *hogy* kötőszó csak a főmondattól való függést jelöli, míg más kapcsolóelemek fogalmi szerkezete ennél általában kidolgozottabb (a kidolgozottság azonban mindig fokozat kérdése). Például a *mert,* a *mint* és egyéb, szófajuk szerint is kötőszói kifejezések úgy töltenek be kontextualizáló szerepet a mellékmondatban, hogy megnyitnak valamilyen fogalmi tartományt (mentális teret) az ábrázolt jelenet értelmezése számára: a *mert* az ok/magyarázat, a *mint* a hasonlítás tartományát jelöli ki ehhez." Imrényi és Kugler, „Mondattan," 832.

[22] A kronologikus egymásra következés a kapcsolatos mellérendelések prototipikus esetének tekinthető az MNSz adatai alapján: „Az és kötőszós tagmondatkapcsolatban az a leggyakoribb (30%), hogy J1 és J2 időben egymást követő jelenetek, J2 a cselekvésláncban feltételezi J1-et." Uo., 860.

 DOI: 10.31400/dh-hun.2021.4.2415

akkor, márpedig, mégse(m), mégis kötőszavak valamelyike, illetve a *csak hát* szerkezet – az írásjelet követő nagy lehetséges terjedelem a szórendi rugalmasság miatt fontos –, és az írásjelet közvetlenül követő *de, ám, pedig* kötőszavak. Az *ellenkezőleg* és a *csak hát* ugyan nem kötőszavak, de prototipikusan kapcsolóelemként működnek tagmondatok között; miközben az *ellenkezőleg* pontosítást és fokozást is kifejez. A reguláris kifejezés nem tesz különbséget a *nemcsak … hanem/de* típusú kapcsolatok második tagjában szereplő ellentétes jelentésű kötőszavak között.

3. Választó mellérendelés. A tagmondatok közti írásjelet közvetlenül követő *vagy, avagy* kötőszavak valamelyike. Az írásjelhez való közelséggel az olyan esetek nagy része kiszűrhető, amikor a *vagy* mondatrészek közötti kapcsolatot létesít, vagy egyes szám második személyű létigeként szerepel. A választó viszony esetenként nem tagmondatok, hanem mondatrészek között dolgoz ki kapcsolatot („Nem minden mutatványnak halál a vége, ahogy a mutatványoknak egyébként sok köze van a halálhoz és a pusztuláshoz, vagy legalábbis ahhoz a rettenethez, amely a halállal van összefüggésben." Darvasi László, *A könnymutatványosok legendája*), amit a kutatás nem különített el és tagmondatkapcsolatként azonosít.
4. Magyarázó mellérendelés. Egy mondatközi írásjel és egy másik írásjel között álló *azaz, azazhogy, ugyanis, hiszen, vagyishogy, tudniillik, miszerint, mármint, mégpedig, egyszóval, elvégre, utóvégre, illetve* kötőszavak valamelyike – az írásjelet követő lehetséges nagy terjedelem a szórendi rugalmasság miatt fontos (a korpuszt vizsgálva 5–6, vagy annál több szó távolságra egyébként elenyészőnek mondható a kötőszavak felbukkanása).
   Az *illetve* sajátosan viselkedő és poliszém kötőszó: magyarázó, választó és kapcsolatos jelentést is felvehet. A 100 összetett mondatból álló mintában azonban 70-szer magyarázó tagmondatkapcsolatban, legtöbbször a *pontosabban* szinonimájaként szerepel, sokszor kapcsolatos jelentésárnyalattal. (29 esetben is hasonló jelentést vesz fel, ám mondatrészek közötti kapcsolat kidolgozását végzi el.)
   Az *azaz* kötőszó 100 összetett mondatból 53-szor szerepelt tagmondatok közötti kapcsolóelemként, de a többi esetben is hasonló grammatikai jelentést hordoz, így a kutatás részét képezi, ugyanakkor a hasonló jelentésű és viselkedésű *vagyis* csupán 43-szor dolgozott ki tagmondatok közötti kapcsolatot, így nem vettük fel a kutatásba (szemben a *vagyishogy* kötőszóval, amely ugyan elenyésző számban szerepel a korpusz mondatai között, akkor viszont jellemzően tagmondatkapcsoló elemként).
   Ebbe a kategóriába tartoznak továbbá a mondatközi írásjelet közvetlenül, vagy a *mert, de, és, s* kötőszavakat követő *hisz, pontosabban.*
5. Következtető mellérendelés. Egy mondatközi írásjel és egy másik írásjel között álló *tehát, ennélfogva, eszerint, úgyhogy, szóval* kötőszavak valamelyike.
6. Megengedő alárendelés. Az alábbi kötőszavak, ha a mondat legalább egy mondatközi írásjelet tartalmaz: *bár, ha… is, habár, igaz, hogy, jóllehet, még ha, mégha, noha, ugyan, ámbár.* A reguláris kifejezés ekkor a leghatékonyabb, mivel figyelembe veszi a mondat elején álló kötőszavakat is, amelyek ebben az esetben prototipikusan fordított, mellékmondat-főmondat tagmondatsorrendet többször vezetnek be, és nem az előző mondatra vonatkoznak, például: „Bár szerelem

miatt kimondhatatlanul szenvedett, e reggelen se mulasztotta el mindennapi tisztelgését a Ferenciek terén…" (Krúdy Gyula, *Hét bagoly*). Ha előírjuk, hogy a mondat legalább egy mondatközi írásjelet tartalmazzon, akkor a vizsgált 100 mondatból 66-szor fordított tagmondatsorrendű megengedő kapcsolatot dolgozott ki a mondat elején álló kötőszó.

Ebbe a kategóriába tartozik továbbá a mondatközti írásjelet közvetlenül követő *holott* kötőszó, mivel a *holott* az előző mondatra vonatkozik prototipikusan a mondat elején, és nem fordított tagmondatsorrendű összetett mondatot vezet be.

7. Hasonlító alárendelés. A mondatközi írásjelet közvetlenül, vagy 1 szó után követő, valamint a mondat elején, vagy az első szó után álló *mint, mintha, akár, akárha, akárcsak, minthogyha* kötőszavak valamelyike. Kivételt képeznek azok az esetek, amikor a *nem* vagy az *a* előzi meg a *mint* kötőszót, mert az előbbi esetben a *nem mint + főnév* szerkezetek, utóbbiban pedig a régiesen használt *amint (a mint)* megvalósulása a jellemző.

   Az *akár* hasonlító jelentése csak akkor van túlsúlyban, ha nem *akár… akár…* sémájú konstrukcióban szerepel, így ennél a kötőszónál előzetesen biztosítani kell, hogy egy *akár* legyen a mondatban – ekkor 82-szer hasonlító és 18-szor partikula szerepben volt azonosítható a 100-as mintában. Ezt a kritériumot a reguláris kifejezésben is meghatározhatjuk, ám a teljesítmény optimalizálása miatt célszerűbb a reguláris kifejezést megelőzően a korpusz mondatait egy új változóhoz rendelni, és a *nem mint* és az *a mint* szószerkezetekkel együtt azokat a mondatokat is eltávolítani ebből a korpuszból, amelyekben több *akár* szerepel – majd ezen a szűkített korpuszon végezni a keresést.

   Amennyiben egy szó áll a mondatközi írásjel és a kötőszó között, 60-szor hasonlító kapcsolat, 30-szor nem tagmondatkapcsolat állt fenn a vizsgált 90 esetből.

   A hasonlító összetételnél figyelembe kell venni, hogy a főmondatrész gyakran csak odaértett módon jelenik meg a mondatban: „[Olyan vagy,] Mintha megváltoztál volna." A mondat elején álló kötőszó ebben az esetben 100-ból 31-szer jelentett tipikus hasonlatot, 16-szor nem tagmondatok közötti viszonyt írt le, és 53-szor, leginkább a *mintha* kötőszó esetében odaértett főmondatrészről volt szó. (A *mint* kötőszó a mondat elején 50-es mintában 37-szer vonatkozott kidolgozott hasonlítói összetételre, 7-szer nem tagmondatok közötti viszonyra, 7-szer pedig odaértett főmondatra). A jelenlegi kutatás nem tett különbséget a valóban kidolgozott és az odaértett hasonlatok között, ám a későbbiekben érdemes lehet a hasonlító összetett mondatok differenciáltabb tipológiáját megadni.

   A reguláris kifejezés továbbá tévesen ismer fel hasonlító összetett mondatként mondatrészek közötti kapcsolatot is, például: „Találkozott Jánossal, kivel mint szakemberrel beszélhetett." Az ilyen hibák azonban statisztikailag elenyészők a hasonlító összetételek számához képest. Nem jelent hibát továbbá, hogy a minősítő minőségjelzői kapcsolatok közül csak a kötőszóval kidolgozott tagmondatkapcsolatokat sorolja be a keresés a hasonlító összetett mondatok alá („Egy olyan nyakkendőre lenne szükségem, mint [amilyen] az öné"), szemben azokkal, amelyeknél a viszonyt nem dolgozza ki kötőszó („Olyan vagyok, [mint]

 DOI: 10.31400/dh-hun.2021.4.2415

amilyen te"), hiszen a kutatás éppen a formai oldalon is kidolgozott és a figyelem előterében álló kapcsolatokra vonatkozik.

8. Ok-, célhatározói alárendelés. A tagmondatok közti írásjelet közvetlenül, vagy 1 szó után követő *mert, merthogy, minthogy, mivel, nehogy* kötőszavak valamelyike – az írásjeltől való nagyobb távolság többnyire téves eredményekre vezet, elsősorban a *mert* és a *mivel* homonim jelentései miatt. Ebbe a kategóriába tartozik továbbá a mondatközi írásjelet közvetlenül követő *különben* kötőszó, illetve az *amiatt, avégre, avégett, azért …, hogy* szerkezetek.
   Ezenkívül szintén ebbe a típusba tartoznak a mondat eleji, vagy az első szót követő *minthogy, mivel* kötőszavak valamelyike a legalább egy mondatközi írásjelet tartalmazó mondatokban, mivel ezek prototipikusan fordított tagmondatsorrendű összetett mondatot vezetnek be („Minthogy többségben vannak az ilyen esetek, felvettük őket is a kutatásba"). A korpuszból vett 100 mondatban 90-szer fordult elő fordított tagmondatsorrendű ok-/célhatározói összetett mondat, amennyiben a reguláris kifejezés előírta, hogy legalább egy mondatközi írásjelet tartalmazzon a mondat.
9. Időhatározói alárendelés. A mondatközi írásjelet közvetlenül követő *mire, közben,* valamint az alábbi kötőszavak/névmások, ha a mondat legalább egy mondatközi írásjelet tartalmaz: *alighogy, ameddig, amíg, amikor, amióta, attól fogva/kezdve …, hogy, amint, mígnem, amidőn, midőn, mielőtt, míg, mialatt, mihelyt, mikor, mióta, miután.* Ezek a kötőszavak ugyanis a mondat elején prototipikusan (100 esetből 94-szer) fordított tagmondatsorrendű időhatározói kapcsolatot vezetnek be, amennyiben a reguláris kifejezés előírja, hogy legalább egy mondatközi írásjelet tartalmazzon a mondat (kizárva az egyszerű kérdő mondatokat, például „Mikor jössz?"). A kifejezés – tévesen – olyan eseteket is felismer, mint a „Mikor jössz, he?", ám ezek száma statisztikailag elenyésző a valódi összetett mondatokhoz képest. A *mire* kötőszó 100 esetből 64-szer dolgozott ki időhatározói kapcsolatot, és csupán 14-szer vezet be prototipikus vonatkozói alárendelést.
   Ezenkívül kezelni kell a mentális állapotról beszámoló *hogy* kötőszós mellékmondatokat, amelyben időhatározói vonatkozó névmás is szerepel, mert ez a konstrukció nem időhatározói alárendelésre utal („Azt kérdezte, hogy mikor jön el a boldogság és a nyugalom ideje"). Ugyanígy ki kell szűrni azoknak a mellékmondatoknak a nagy részét is, amelyekből hiányzik a *hogy* kötőszó („Nem tudom, mikor oldják fel a korlátozásokat"). Ezek közül azokat lehet kis hibaszázalékkal azonosítani, amelyek esetében a mentális folyamatokat kidolgozó igék és ragozott formáik közvetlenül[23] előzik meg a mondatközi írásjelet (lásd az előző példát). Ennek érdekében a Magyar Nemzeti Szövegárból (*MNSz* v2.0.5) vételezett mintákon vizsgáltuk, hogy prototipikusan milyen igék szerepelnek a *hogy* kötőszós mellékmondatok előtti főmondatokban,[24] majd ezeknek az igék-

[23] Azokban az esetekben, amikor nem közvetlenül előzik meg az írásjelek a felsorolt igéket, az eredmények megbízhatatlanok voltak és nem csupán a *hogy* kötőszós mellékmondati konstrukciókra vonatkoznak.

[24] Oravecz Csaba, Váradi Tamás, and Sass Bálint, „The Hungarian Gigaword Corpus," in Nicoletta Calzolari et. al., eds., *Proceedings of LREC 2014,* 1719–1723 (Reykjavík: European Languages Resour-

nek a viselkedését a korpuszból vett mintákon elemeztük. A minták kiértékelését követően az alábbi töveket tudtuk meghatározni, amelyek mentális folyamatot kidolgozó igeként prototipikusan ilyen konstrukciókban szerepelnek: *főnévi igenév + akar, elárul, eldönt, elképzel, érdekel, tud, kérdez, kérd, képzel, kitalál, megállapít, megért, megmutat, sejt.*

Ezenkívül fontos a *mint* kötőszó után álló időhatározói vonatkozói névmások kiszűrése is („Olyan volt, mint amikor elfelejted a locsolóverset"). Ezeknek figyelmen kívül hagyását a *hogy* kötőszós mellékmondatokkal együtt a reguláris kifejezésben is definiálhatjuk, ám a teljesítmény optimalizálása miatt célszerűbb a reguláris kifejezést megelőzően a korpusz mondatait egy új változóhoz rendelni, és ezeket a konstrukciótípusokat eltávolítani ebből a korpuszból – majd ezen a szűkített korpuszon végezni a keresést.

10. Helyhatározói alárendelés. Az alábbi névmások, ha a mondat legalább egy mondatközi írásjelet tartalmaz: *ahol, ahonnan, ahova, ahová, ameddig, amerre* (a mondat elején álló kötőszó ekkor a 100-as mintából 77-szer helyhatározói alárendelést vezetett be), valamint a mondatközi írásjelet közvetlenül – vagy a régies írásmód miatt az *a* mutató névmás után – *hol, honnan, hová, hova, meddig, merről, merre* kötőszavak valamelyike – kiszűrve a nem a tagmondatok közti viszonyt jelölő kérdő névmások egy résztét („Megtántorodott, nem volt hová hátralépnie").

A reguláris kifejezés nem tesz különbséget a *hol …, hol* típusú kapcsolatos mellérendeléssel, amely azonban hordoz helyre vonatkozó jelentésárnyalatot.

Ezenkívül ennél a típusnál is kezelni kell a *hogy* kötőszós mellékmondatokat („Nem tudom, [hogy] hol vagyok."), amelyek egy részét a fent említett módszerrel szűrtük ki a korpuszból.

11. Prototipikus (nominatívuszi) vonatkozói névmással bevezetett mellékmondat. Az alábbi névmások és ragozott alakjaik (azaz azok a vonatkozó névmások, amelyeknek van személyjelölő párja), ha a mondat legalább egy mondatközi írásjelet tartalmaz: *ami, aki, amely, mely,* kivéve, ha a *mint* kötőszó áll előttük („Úgy viselkedtem, mint aki villanyszerelő"). A mondat elején az említett vonatkozói névmások közül az *aki* 100-ból 91-szer, az *ami/amely* 100-ból 70-szer vezetett be fordított tagmondatsorrendű vonatkozói alárendelést összetett mondatokban. Ahhoz, hogy csak a névmások ragozott alakjaira vonatkozzon a kifejezés, a szövegből (a fent ismertetett módon) előzetesen ki kell szűrni az *amint, amikor, amiként, amiképp, amiatt, amidőn, amióta, amialatt, amielőtt, amiután, amíg, melyik* (amennyiben nem az *a melyik* régies írásmód érvényesül) szavakat.

Ebbe a kategóriába tartoznak továbbá az *és, s, de* kötőszavak, (a régies írásmód miatt) az *a* mutató névmás, továbbá a mondatközi írásjelet közvetlenül követő *ki, a mit, kit, mik, kik, miben, kiben, mibe, kibe, a mire, kire, a miről, kiről, mitől, kitől, miből, kiből, kinél, kivel, a mivel, minek, kinek, min, kin, mihez, kihez, miket, kiket, mikben, kikben, mikbe, kikbe, mikre, kikre, mikről, kikről, miktől, kiktől, mikből, kikből, miknél, kiknél, mikkel, kikkel, miknek, kiknek, miken, kiken,*

ces Association, 2014), `https://doi.org/10.13140/2.1.3957.1840`. A konkordanciákat lekérő eljárás és az eredményekből létrehozott gyakorisági listák a *Függelék* 2.2. pontjában találhatók.

 DOI: 10.31400/dh-hun.2021.4.2415

mikhez, kikhez névmások (kiszűrve a nem a tagmondatkapcsolatot kidolgozó kérdő névmások egy részét [„Egyedül harcolok, hiszen kiben is bízhatnék."]), illetve a mondatközi írásjelet közvetlenül követő *mi* névmás (kiszűrve a többes szám első személyű személyes névmások java részét, bár a kifejezés így is tartalmaz olyan eseteket, mint például: „ők dolgoznak, mi beszélünk róla.")
A kifejezés nem tartalmazza a *mivel, mire, mit, miről* névmásokat (csak a régies írásmóddal írt *a mivel, a mire, a mit, a miről* alakokat), mert az első prototipikusan ok-/célhatározói alárendelésben, a második időhatározói kapcsolatban, utóbbi kettő pedig kérdő névmásként, vagy *hogy* kötőszós mellékmondatokban fordul elő; valamint nem különíti el a *ki* szó igekötői használatát, ám a mondatközi írásjel utáni távolság korlátozásával ezek nagy része kiszűrhető.
Ezenkívül ennél a típusnál is kezelni kell a *hogy* kötőszós mellékmondatokat („Nem értem, [hogy] mit mondasz."), amelyek egy részét a fent említett módszerrel szűrtük ki a korpuszból, azzal a kiegészítéssel, hogy ebben az esetben a mentális igékre vonatkozó tövek közé felvettük az *ért* tövet is; a *mi* vonatkozó névmás és ragozott alakjai előtt pedig nem előírás, hogy a mentális folyamatot kidolgozó ige közvetlenül a vessző előtt álljon.

12. Feltételes alárendelés. Az alábbi kötőszavak, ha a mondat legalább egy mondatközi írásjelet tartamaz: *ha, hacsak, amennyiben, hogyha*. Ezek a kötőszavak a mondat elején prototipikusan (100-as mintából 94-szer) fordított tagmondatsorrendű összetett mondatot vezetnek be összetett mondatokban.
A reguláris kifejezés nem tesz különbséget a *ha … is* típusú megengedő mellérendeléssel, ám az is tartalmaz feltételes jelentésárnyalatot.

Az R programnyelvben lefuttatott keresés ezeket a szempontokat érvényesítette, amikor a mondatokra felbontott regényekben (vagy azok szűkített változatában) számlálta meg az egyes típusokhoz tartozó kötőszavak számát. Ezt a számot az adott regény szavainak számával elosztva (és a kezelhetőség érdekében 1000-rel megszorozva) kaphatjuk meg a típusok relatív gyakoriságát. A mondatok számával való osztás nem lenne célravezető, hiszen abban az esetben a hosszúmondatos prózák magasabb értéket vennének fel a rövidmondatos regényekkel szemben. A véges igék száma viszont szintén jó osztó lehetne, de ehhez ismét szükség lenne egy olyan nyelvi elemzőre, amely a 19. századi szövegeken is megbízhatóan működik (és tudja kezelni például a múlt idők különböző fajtáit). A következő részben ismertetett számok és ábrák a regények egészén lefuttatott keresések eredményei – az összes érték (12 típus 100 regényre vonatkozóan) megtalálható a *Függelék* 1.1. pontjában. Célszerű ugyanakkor ezeket az értékeket olyan számításokkal is kiegészíteni, amelyek nem a regények egészén, hanem az azokból vett, lehetőleg dialógusmentes[25] és egyenlő hosszúságú mintákon (jelen esetben: 250 mondat, amely alacsony szám oka Esterházy Péter *Fuharosok* című

[25] A leíró és párbeszédes részek automatikus elkülönítése a MEK és a DIA felületein elérhető különböző kiadások miatt nehézkes (a DIA a könyvbeli forráskiadást imitálja, így a különbségek ott a szerzők és kiadók szokásaira vezethetők vissza); nem adható meg ehhez csupán egyetlen tipográfiai szempont. Az itt használt kódok a szólamváltást jelölő írásjelek mellett a mondást kifejező igéket vették figyelembe a dialógusmentes részletek elkülönítésekor. A kutatás során meghatározott mondást kifejező igék tövei: *mond, kérd, felel, szól, folytat, így, kiált, int, hökk, ord, üvölt, sik, súg, suttog, morm, vél, ismét, nevet, bólog, tett, tev, hall.*

művének rövidsége) alapulnak: míg az előbbi módszer a művek egészéről mond el többet, addig az utóbbi statisztikailag jobban összehasonlítható értékeket szolgáltat (bár ezek az értékek kevesebb adatra vonatkoznak, így esetenként nem is található egy adott kapcsolattípus a vizsgált 250 mondatban). Az eredmények megtalálhatók a *Függelék* 1.3. pontjában.

Ugyanilyen kereséseket végeztem a regények időben egymást követő 10-es csoportjain is a történeti tendenciák vizsgálatához; valamint külön kódok számították ki, hogy az egyes kapcsolattípusok egymáshoz képest milyen arányban találhatók meg a regényekben (lásd *Eredmények 3.*) – ekkor az egy regényben azonosított összes találat számával osztottam el az egyes típusokhoz tartozó értékeket és szoroztam meg 100-zal (esetenként két típust összevontam egy tágabb kategóriába a jobb átláthatóság kedvéért). Az ezeket bemutató értékek a *Függelék* 1.4. (évcsoportok) és 1.5. pontjában (belső arány) találhatók.

Eredmények 1.

A kiszámított értékeket három szempont szerint ábrázolhatjuk: (1) az idővonal mentén a történeti változásokra összpontosítva; (2) összehasonlító módon vizsgálva, hogy mely regényekre mennyire jellemzők az egyes kapcsolattípusok; (3) valamint aszerint, hogy egy regényen belül milyen arányban oszlik el a típusok gyakorisága.

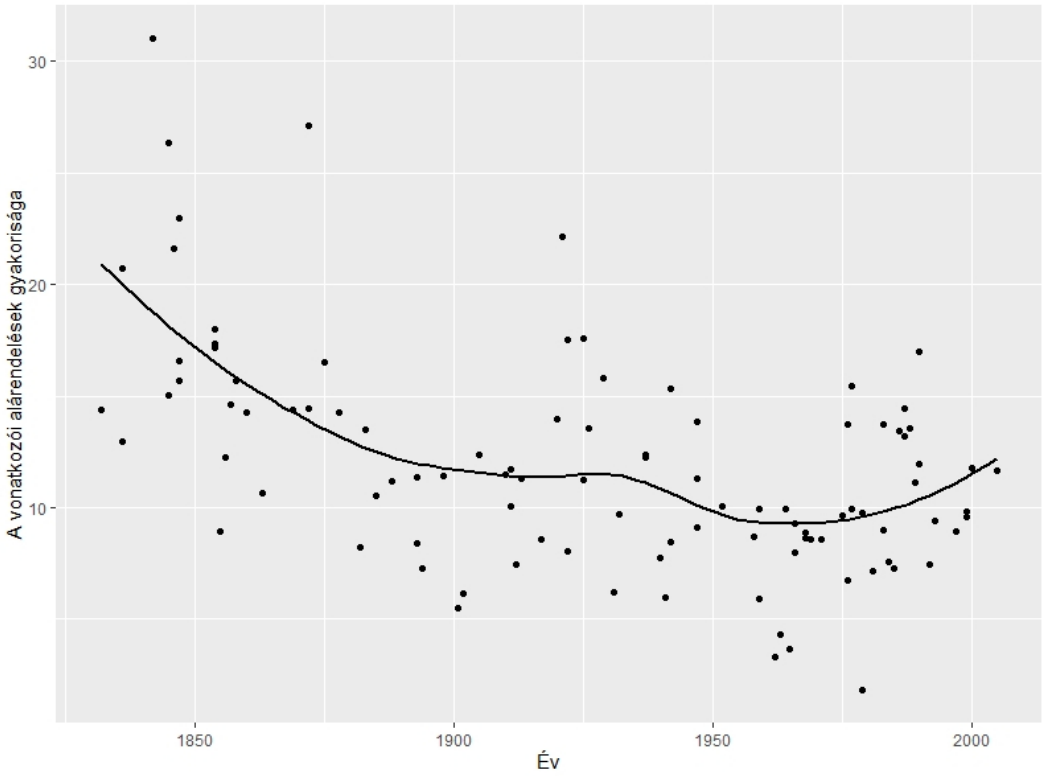
Az első szempont szerint kétféle grafikont hozhatunk létre. Egyfelől ábrázolhatjuk a regényekhez rendelt értékeket a megjelenési évek viszonyában, valamint az így felrajzolt adatpontokra illeszthető regressziós görbét, másfelől a 10 regényenkénti évcsoport értékeit oszlopdiagramok formájában is összehasonlíthatjuk.

A prototipikus vonatkozói alárendelések (időbeli változás)

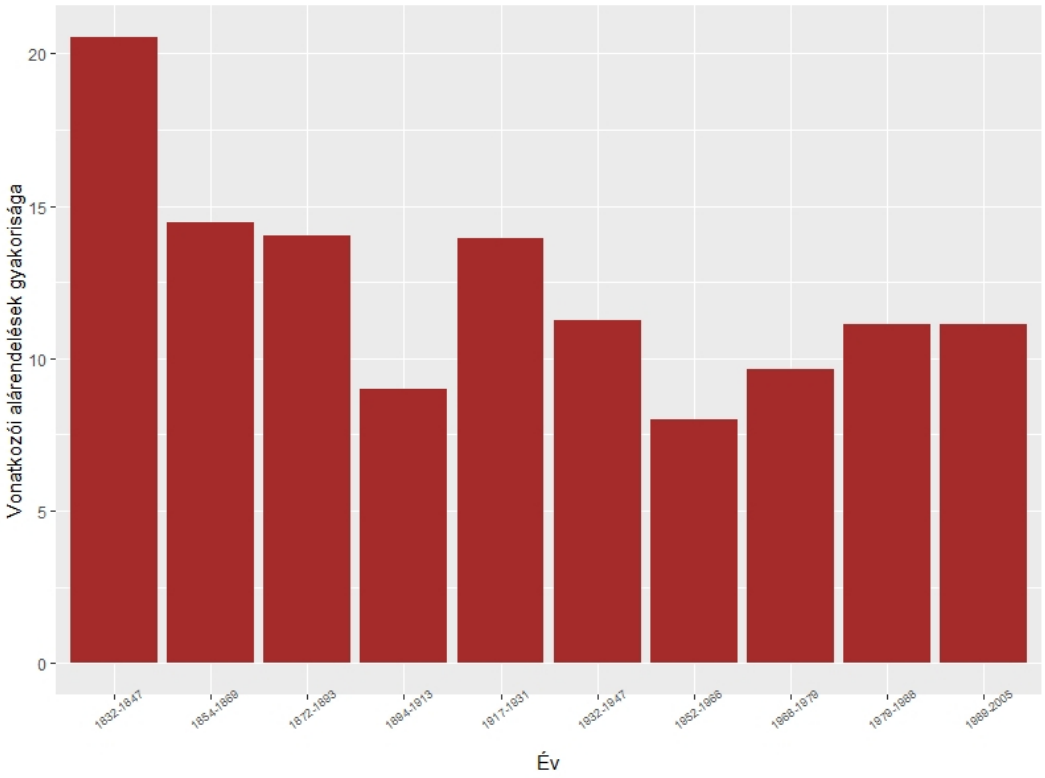
A *2.1.* és *2.2. ábrá*n a prototipikus (nominatívuszi) vonatkozói alárendelések változása figyelhető meg. A *2.1. ábrá*n megfigyelhető csökkenő tendenciát[26] egyszerre okozzák az 1840–50-es évek kiugró értékei, valamint az, hogy az 1870-es évekig alig található szöveg a grafikon alsó régióiban. Az ábra ezáltal arra a poétikai-történeti összefüggésre hívja fel a figyelmet, miszerint a korszak előszeretettel dolgozza ki a mondatok szereplőit részletekbe menően, külön tagmondatok segítségével.

[26] Az egyszempontos varianciaanalízis (ANOVA) alapján erős korreláció mutatható ki a vonatkozói kapcsolattípus gyakorisága és a megjelenési évek között (F-érték: 35.44), azaz a kirajzolódó tendencia által jelölt különbségek statisztikailag szignifikánsnak tarthatók.

 DOI: 10.31400/dh-hun.2021.4.2415



2.1. ábra. A prototipikus vonatkozói alárendelések változása, 1832–2005.



2.2. ábra. A prototipikus vonatkozói alárendelések változása, évcsoportok.

A vonatkozói alárendelés a 19. század első felében nagy hatást kifejtő ritmikus-retorikus prózastílus esetében a leggyakoribb. Ez a stílus olyan többszörösen összetett mondatokból (az úgynevezett „romantikus körmondatokból") épül fel, amelyekben a mondatrészek és a tagmondatok azonos pozícióban és azonos szinten, szimmetrikusan helyezkednek el: ez jelentheti egyrészt azt, amikor ugyanolyan vagy hasonló jelentésű mellékmondatok dolgozzák ki a tagmondatokban ugyanazt a helyet elfoglaló (például a tagmondat elején álló), azonos funkciót betöltő mondatrészeket (pl. alanyokat), másrészt azt, amikor egyetlen mondatrészhez azonos fokú, egymással mellérendelő viszonyban álló tagmondatok felsorolásszerűen kapcsolódnak. Ennek az anaforisztikus stílusnak legfontosabb hazai képviselője Eötvös József,[27] akinek a korpuszba felvett mindhárom regénye az első négy legnagyobb értéket mutató szöveg között található – a 100 regény közül pedig *A karthauzi*ban a leggyakoribb a vonatkozói alárendelések használata. A szimmetrikusan elhelyezkedő vonatkozói alárendelésekre az alábbi példákat hozhatjuk életművéből:

> Ő, ki e házban született s növekedett, ki e falak közt szorítá apját utólszor kebeléhez, ki, mióta az meghalt s őt hivatalában követé – itt tölté életét, kit a faluban minden ház, a temetőn minden sír olyanokra inte, kiknek ő volt legjobb vigasztalója, s ki a pálya vége felé közelgve érzé, hogy emlékei közt más keserűek nincsenek, mint melyeket a halál hagyott: - miként álmodhatott más szerencséről […] (Eötvös József, *A falu jegyzője*)[28]

> […] ti bércek, melyeknek látásánál a gyermekszív egykor vágyakkal telt; te dörögve lezúgó hegypatak, melynek hatalmas szava túl nem fogja hangozni az öröm érzetét, mi e szívben felzajog; te virágzó mező, mely soha nem hazudtál még, s minden tavasszal meghozád virágidat! (Eötvös József, *A karthauzi*)

Az első példában öt, ugyanahhoz a főmondathoz tartozó, a *ki* névmás által bevezetett alanyi mellékmondat található, amelyek egy szinten, egymáshoz képest mellérendelői viszonyban találhatók. A második példában különböző alanyokat dolgoznak ki a *mely* névmással bevezetett mellékmondatok, ám ezek az alanyok és a tagmondatok is a mondat azonos szintjén, azonos pozícióban helyezkednek el. Ezáltal jól követhető, átlátható szerkezet jön létre, amely a mondatok hosszúságának ellenére is könnyen értelmezhető. Ez a szerkesztésmód egészen a 13. századi magánoklevelek és kancelláriai iratok stílusára vezethető vissza, amelyekben a párhuzamos részek az iratok áttekinthetőségét szavatolták: a jogi és retorikai műveltség összefonódása pedig megteremtette a ritmikus-retorikus próza alapjait, amely a kora reneszánsztól mint irodalmi nyelv is kifejtette hatását.[29] Ez az irodalmi nyelv a 19. század Magyarországán Herczeg

[27] Vö. Herczeg Gyula, „Eötvös körmondatai (Folytatás)," *Magyar Nyelvőr* 77, 1–2. sz. (1953): 165–179.

[28] A szépirodalmi idézetek hivatkozott kiadását a *Függelék* 4. pontjában ismertetem és nem jelölöm külön az egyes idézeteket követően.

[29] Herczeg, „A mondatstilisztikai kutatás mint módszer," 133–135. Érdemes elgondolkodni azon az összefüggésen is, amely az államszervezés, a retorika és a perspektivikus képábrázolás között áll fenn: ekkor a párhuzamos mondatszerkesztés nemcsak a táblázatos és egyéb formanyomtatványokba történő elrendezés nyelvi előképeként jelenik meg, hanem a látvány leképezésének mint a hely kiosztásának szintén a reneszánszban kifejlődött technikájával is összevethetővé válhat. Ennek

 DOI: 10.31400/dh-hun.2021.4.2415

Gyula megfogalmazásában „a politikai igények stilisztikai vetületeként" értelmezhető, amennyiben a felsorolásszerűen egymásra következő tagmondatok halmozásával „buzdít, mozgósít, patetikus hangvétellel bírál vagy magasztal".[30]

A ritmikus-retorikus próza a korpusz szövegei közül elsősorban Eötvös regényeiben fedezhető fel, ám találhatunk rá példát Kemény Zsigmond, Jósika Miklós vagy Toldy István műveiben is.[31] Esetükben azonban – ami a korszak prózairodalmára általánosan is igaz – gyakoribbak a beékelésszerű alárendelések, amelyek nem kerülnek a fent bemutatott módon párhuzamba más mellékmondatokkal; illetve a kevésbé szabályosan elrendezett, különböző szinteken elhelyezkedő vonatkozói névmásokkal bevezetett mellékmondatokból építkező összetettebb szerkezetek. Ez utóbbira az alábbi példákat említhetjük:

> Ali azon különös tekintetek egyikét veté most az alig tizenhat éves leányra, melyek szívtelen embereknél, mintegy átküzdvén kedélyök lemezén, inkább valami eszményi igazság érzetét a természetnek látszanak bebizonyítani, mint érzést, mely a szívből kelne – azon igazság érzetét, mely oly harsányul hangzik, a legelfásultabb kedélyek dacára, s akaratlanul kivíjja a méltánylást önkénytelen kifejezésében a vonásoknak, melyek néha többet tolmácsolnak, mint lelkünk érez vagy akar érezni. (Jósika Mikós, *Diamante*)

> […] és alig van társadalom, mely egy bukott nőt ritkábban kényszerítene pirulni, mint volt a párisi a császárság alatt, melyben a demi-monde úgy szólván bizonyos erkölcsi jogosultsággal bírt, melyet el nem ismerni annyi volt, mint „elmaradni a világtól." (Toldy István, *Anatole*)

az összevetésnek adhat alapot a tény, hogy Alberti is a retorikai mondatelemzés mintájára és – éppen a térbeli viszonyokból eredő – terminológiájával *(compositio, locus, circumscriptio)* írja le a képelemzés folyamatát. Vö. Bernhard Siegert, *Cultural Techniques: Grids, Filters, Doors, and Other Articulations of the Real,* trans. Geoffrey Winthrop-Young (Fordham: Fordham UP, 2015), 132, `https://doi.org/10.1515/9780823263783`.

[30] Herczeg, *A XIX. századi magyar próza stílusformái,* 84–85. Innen nézve talán nem véletlen, hogy Eötvös politikus is volt; stílusán érezhető országgyűlési beszédeinek és egyéb publicisztikáinak hatása is.

[31] Kemény Zsigmondnak inkább korai prózájában találhatunk ilyen szerkezeteket, de idézhető például *A rajongók* (1858–59) alábbi mondata is, ahol az *aki* mint általános alany szerepel és négy mellékmondatban vezeti fel a főmondatot: „*Aki* figyelemmel kísérte az ellentétet e férfiú szavai és kedélyhangulata közt, *aki* elleste a fájó benyomást, mely idegein átrezgett, midőn magasztalni hallá magát, *aki* látta, hogy máskor meleg, részvékeny és áradozó szíve mennyire elfeledte a hőbb dobogással együtt a lelkesülést és életörömet, *aki* fanyar mosolyát, sötét merengéseit, nyugtalan álmait és álmatlan éjeit aggódva követé szemeivel s gondjaival: az kénytelen volt valami mélyen ható, valami rejtélyes és bősz történetben keresni e rögtöni átalakulás okát." (Kemény Zsigmond, *A rajongók*) Jósika prózájából a következő idézet szolgálhat példaként, melynek bevezetése az azonos szinten és pozícióban lévő alanyok és a hozzájuk tartozó alárendelések felsorolásából áll: „A fez, *mely* fejét borítá; az ujjas, *mely* derekára szorult; a törökös csálvár, *mely* szárain folyt alá: minden fekete fustanella, *mely* derekán alul térdéig ért, volt fehér és tiszta, mint a tavaszi hó." (Jósika Miklós, *Diamante*) Toldy István 1872-es *Anatole* című regényéből pedig az alábbi példát említhetjük, amelyben egyetlen alanyhoz (a kereskedőhöz) kapcsolódnak a vonatkozói alárendelések: „A Flavigneux család minden ismerősei és rokonai közt alig volt ház, mely több tiszteletet érdemelt volna, mint e derék kereskedőé, *ki* ötven éves korában egy húsz éves ifjú gyöngédségével szerette nejét, *kinek* vágyait a negyven éves nő odaadó szerelme teljesen kielégíté s *ki* mindent elkövetett, hogy három gyermekét a lehető legjobb nevelésben részesítse." (Toldy István, *Anatole*)

Ezekben az esetekben a szimmetrikus szerkesztés hiányában és az egymásból kinövő alárendelések következtében a mondatok szerkezetének az átláthatósága, érthetősége is jelentősen csökken. Ami viszont közös az eddig említett szerzőkben – és a *2.1. ábra* alapján a korszak többi regényírójában is –, az a vonatkozói névmással bevezetett alárendelések nagy száma. Ezek az alárendelések a főmondat egy szereplőjét teszik megfigyelhetővé egy más jelenetben is, ezáltal többletinformációt szolgáltatnak róla, képesek annak meghatározására. A ritmikus-retorikus prózastílusban a mellékmondatok halmozása egyben az olyan jelenetek halmozását is jelenti, amelyben a korábbi tagmondat szereplőjét figyelhetjük meg, és ez a halmozás, az egymást erősítő (mert a korábbi szereplő hasonló tulajdonságait bemutató), és sokszor csattanószerű, rövid zárlatot előkészítő mondatrészek adják a szövegek retorikusságát.[32] Ez a retorikus elem csak helyenként található meg a század közepének prózájában, ám a tagmondatok egyes elemeinek vonatkozói névmással bevezetett mellékmondatokkal történő kidolgozása, azaz egy új jelenetben a tulajdonságok és viselkedésmódok meghatározása alapvetően jellemző a korszakra. Míg Eötvös regényeiben ez az eljárás az éles kritika és az ironikus hangvétel sokszínű eszközeként jelenik meg, addig például Toldy István 1872-es, amúgy nagyon gondosan megformált, *Anatole* című regényében néhol idejétmúltnak hat a társadalom helyes működése és a társadalmi szerepek meghatározásának ez az igénye, amelyet a vonatkozói alárendelések gyakori alkalmazásán keresztül kíván elérni.[33] Ennek megfelelően a század végére nemcsak a ritmikus-retorikus stílus szorul vissza, hanem a vonatkozói alárendelések gyakorisága is csökken (Eötvös három regényét tekintve is megfigyelhető ez a csökkenő tendencia), és inkább rövidmondatos próza veszi át a korábbi stílusok helyét, amelynek szerkezetei nem a szereplők új jelenetben történő megfigyeltetését részesítik előnyben.

A következtető és a magyarázó kapcsolatok (időbeli változás)

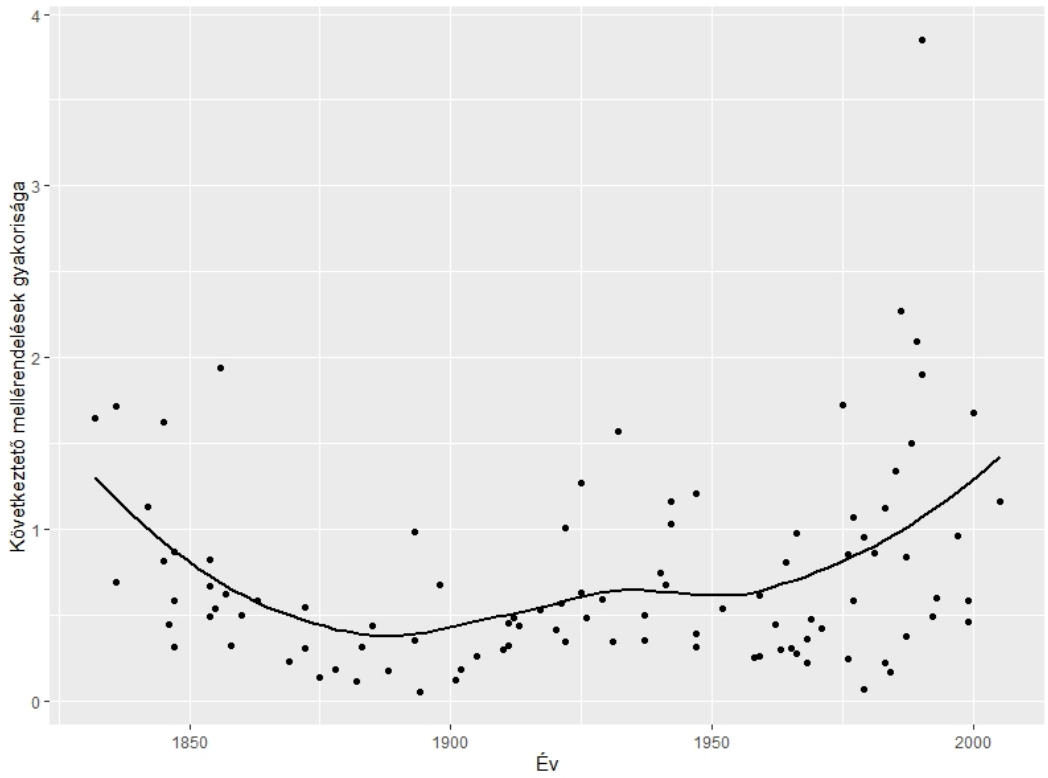
A mérési adatok alapján megfigyelhető egy másfajta prózastílus is a 19. század közepén, amely szintén többszörösen összetett mondatokat hoz létre, viszont kevésbé alárendelésekkel, hanem inkább mellérendelő kapcsolatokkal dolgozza ki a tagmon-

[32] A stílus fennköltsége ugyanakkor gyakran parodisztikus felhangokkal is társul, leginkább abban az esetben, amikor a honi viszonyok tarthatatlanságának leírásával párosul a fennkölt stílus, vagy amikor a vármegyei közigazgatást saját nyelvével gúnyolja ki a szerző. Vö. Herczeg, *Eötvös körmondatai (Folytatás)*, 171.

[33] Szemléletes példa lehet ez utóbbira az alábbi részlet, amely ugyan azt állítja, hogy vannak annyira szép nők, akiket nem lehet szavakkal leírni, ám éppen hogy egy nagy ívű és szemléletes leírását adja ennek a szépségnek: „Vannak nők, *kikről* hívebb fogalmat adhatunk az olvasónak, ha nem is kisértjük meg leírásukat; ha fantáziájokra bízzuk elképzelni ama női eszményképet, *melyet* lelke egyszer megálmodott […] Mert e nő szépsége olyan volt, *melyet* nemcsak szemünkkel, elménkkel, hanem lelkünk, kedélyünk minden ízével élvezünk. (…) ha bámuljuk Vénust, *melynek* szépsége a vonások szabályos tökélyében rejlik, vajjon nem kell-e vele egy polczra helyeznünk az oly nőt, *kinek* szépsége nem a vonások anyagában, hanem az arcz szellemében, kifejezésében, a tekintet bájában s mosolya tiszta költészetében vakítólag lép elénkbe? […] E mosoly, *mely* szemében, ajkán s arczának minden izében honolt, ragyogott, mint a nap, s káprázatos glóriával vette körül az egész alakot, *melynek* fényében a körülötte sürgő fekete frakkok úgy tüntek föl előttem, mint egy csapat oktalan moly, *mely* meggondolatlan ösztönét követve, körülrepkedi a tüzet, hogy lángjában megperzselve szárnyait, lehulljon."

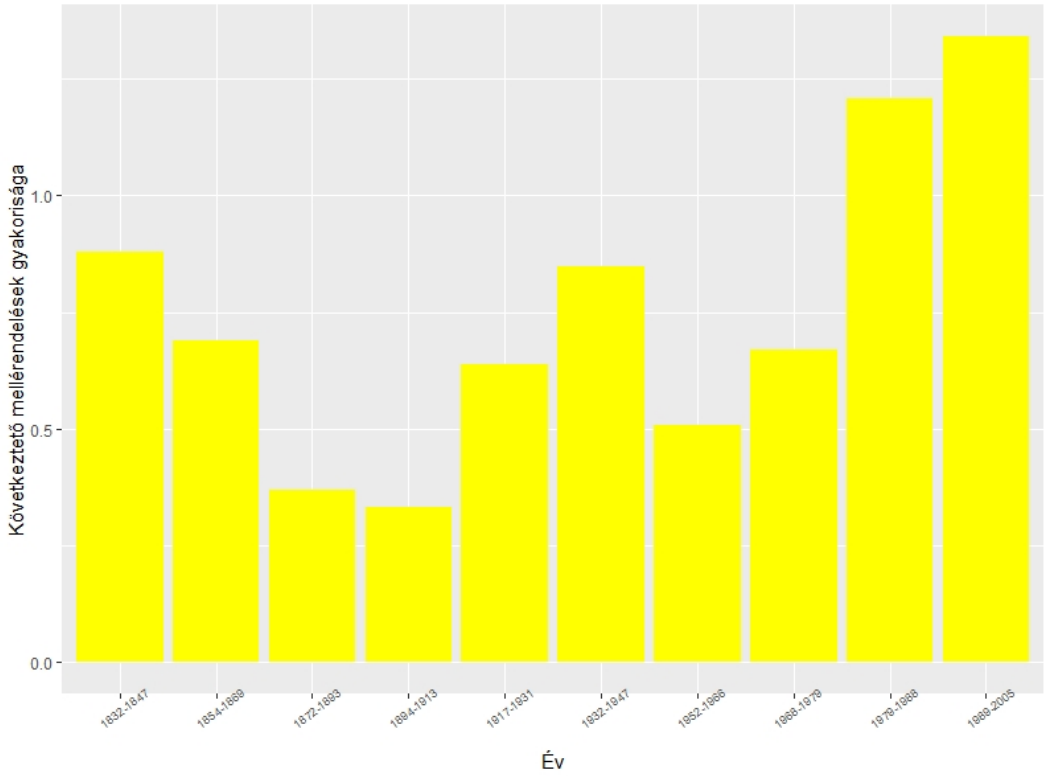
 DOI: 10.31400/dh-hun.2021.4.2415

datok közötti viszonyt. Ezek a tagmondatok akár önálló mondatokként is szerepelhetnének a szövegben, a kötőszavakkal történő összedolgozásuk a közöttük lévő logikai és/vagy ok-okozati összefüggést erősítik. Gaál József *Szirmay Ilona* című regényében az ellentétes, a következtető és megengedő, Nagy Ignác *Magyar titok* című regényében a következtető, a választó és a megengedő, Fáy András két regényében és Kuthy Lajos *Hazai rejtelmek* című regényében a választó, Vas Gereben *Nagy idők, nagy emberek* című regényében pedig a következtető mellérendelések nagy száma figyelhető meg a korpusz egészéhez képest. A tagmondatok ilyen összefűzése kevésbé szerkesztett és szabályos prózanyelvet hoz létre, ám lehetővé teszi az események dinamikus egymásra következését, valamint a motivációk és élethelyzetek részletes kifejtését. Olyan hosszúmondatos poétika ez, amely az 1970-es évektől jelenik meg ismét a magyar prózairodalomban – sőt még jellemzőbb lesz szerzők egy csoportjára, ahogy az a következtető mellérendelések gyakoriságának alakulását bemutató grafikonokon is látható. *(2.3.*[34] és *2.4. ábra)*



2.3. ábra. A következtető mellérendelés változása, 1832–2005.

[34] Az egyszempontos varianciaanalízis (ANOVA) alapján csak a három nagyobb időszak (1832–1860, 1861–1874 és 1975–2005) között mutatható ki korreláció a következtető kapcsolattípus gyakorisága és a megjelenési évek között (F-érték: 21.05), azaz a kirajzolódó tendencia által jelölt különbségek csak ezek között az időszakok között tarthatók statisztikailag szignifikánsnak.



2.4. ábra. A következtető mellérendelések gyakoriságának változása, évcsoportok.

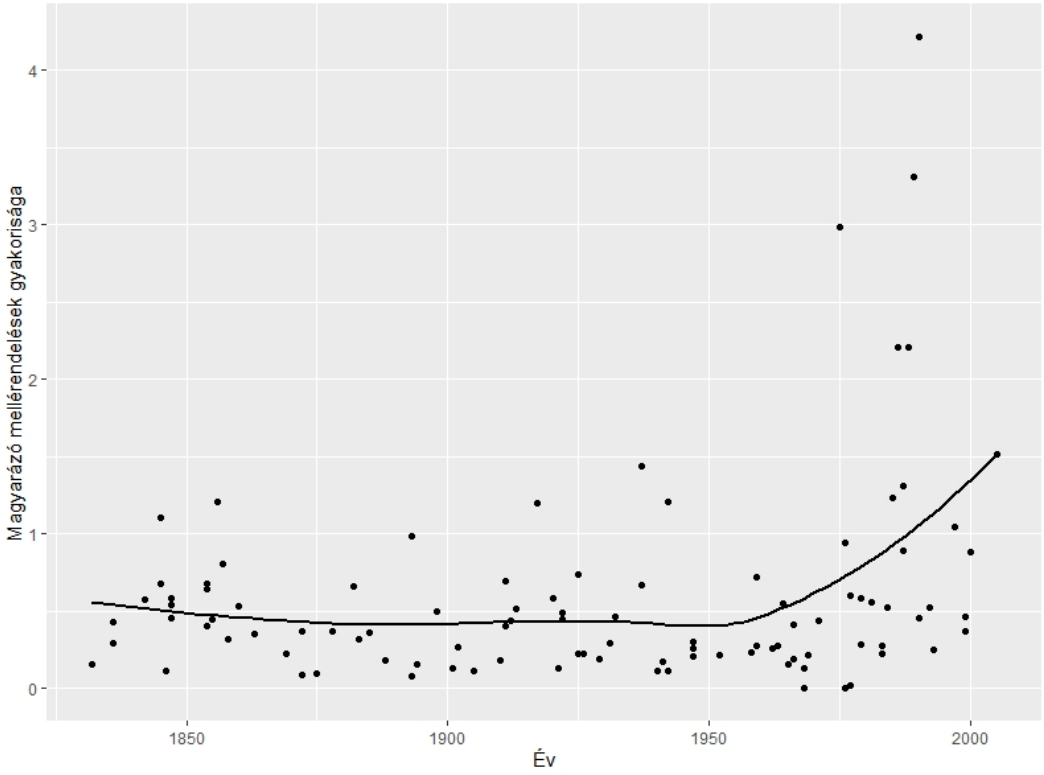
Ezek az ábrák és értékek kapcsolatot teremthetnek a két korszak regényirodalmának egyes képviselői között (és akár a stílustörténet ciklikusságára is felhívhatják a figyelmet), ám érdemes rávilágítani a köztük lévő különbségekre is. A következtető kötőszavak a 19. században legtöbbször az előzményekből következő eseménysort vezetik fel és egy cselekvő motivációját dolgozzák ki. Az egymásból következő események jelölése azért is szükséges, mert sokszor többszörösen összetett, bonyolult szerkezetek közötti viszony válik ezáltal könnyebben átláthatóvá.[35] Ugyanakkor már a 19. században megjelenik a következtetések ironikus-parodisztikus használata is, amelyek inkább meghökkentő, egymásból logikailag nem feltétlenül következő kapcsolatokat dolgoznak ki következtető mellérendelések vagy ok- és célhatározói alárendelések formájában.[36]

[35] Például: „A következő éjszaka, az után, melyet most leírtunk, fényesen volt a teli holdtól világítva, s már éjfél felé annyira haladtak a haramiák a várvíváshoz megkivántató szerek és eszközök elkészítésével, hogy ahhoz magához fogni lehetett; megparancsolá *tehát* Pintye Gregor, hogy az ágyúkat még az éji homályban a kitűzött helyekre vonják, oda hordván egyszersmind a már tetemes számmal öntött golyókat és puskaport, melyben a rablók hiányt nem szenvedtek." (Gaál József, *Szirmay Ilona*)

[36] Például: „Az árkokban rothadó víz és szemét igen ocsmány bűzt terjeszt és ez rendkívül hasznos, *mert* a faluról bejövő uraságokat arra figyelmezteti, hogy mielőbb siessenek ismét vissza egészséges falusi levegőjökre, hol a csődületi betegség nem érheti őket el oly könnyen, mint Pesten." (Nagy Ignác, *Magyar titok*)

 DOI: 10.31400/dh-hun.2021.4.2415

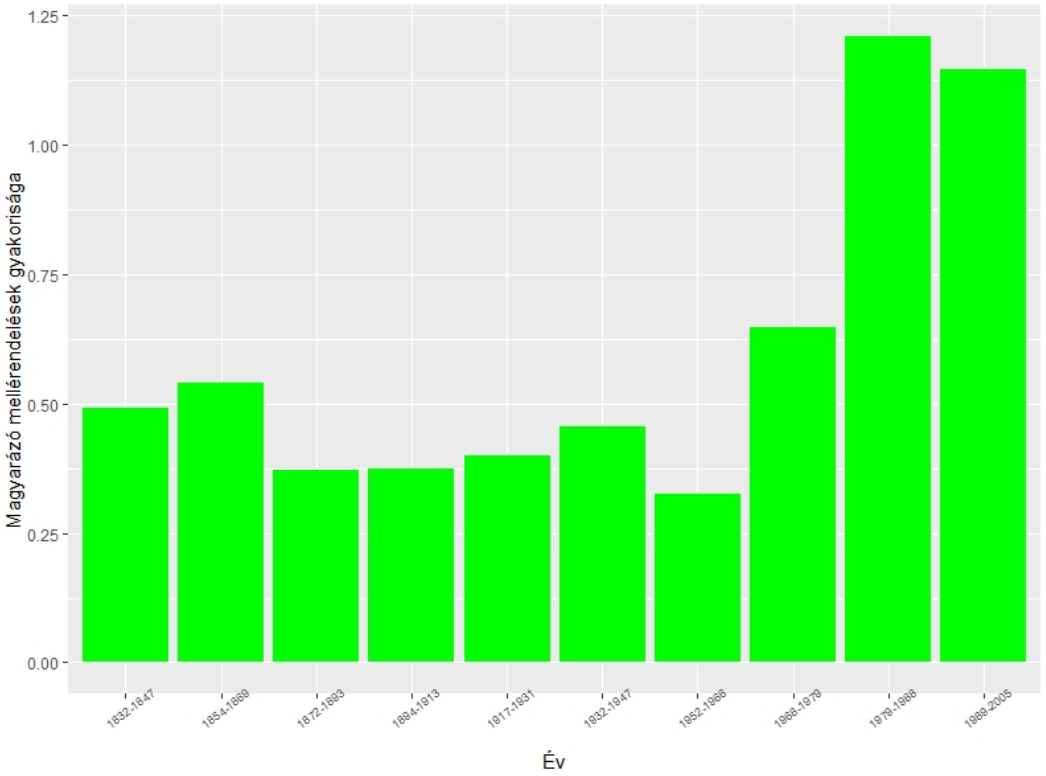
A vizsgálat fókuszát az időtengely másik végére helyezve megállapíthatjuk, hogy a következtetésekben rejlő ilyen potenciált aknázzák ki a 20. század utolsó harmadában jelentkező hosszúmondatos prózák is, amelyek, miközben részletes összefüggéseket képesek felvázolni a többszörösen összetett szerkezetekben, ezeknek az összefüggéseknek a megalkotottságára, esetleg abszurditására is felhívják a figyelmet. Ez a megalkotottság a 20. század második felének mondatszerkezeteiben összefügghet a kor irodalmának azzal a felismerésével, miszerint az eseményeket bemutató prózanyelv is egy beszédmód a sok közül, amely egy eleve valamiként értelmezett világot képes csak ábrázolni.[37] Ezt nem csak a következtető, hanem a magyarázó mellérendelések nagy számával érik el – olyan jellemző ez, amely a 19. századi regényirodalomból hiányzik. *(2.5.*[38] és *2.6 ábra)*



2.5. ábra. A magyarázó mellérendelés változása, 1832–2005.

[37] Vö. Kulcsár Szabó Ernő, *A magyar irodalom története 1945–1991* (Budapest: Argumentum Kiadó, 1994), 479–496.

[38] Az egyszempontos varianciaanalízis (ANOVA) alapján egy gyenge, de egyértelmű korreláció mutatható ki a vonatkozói kapcsolattípus gyakorisága és a megjelenési évek között (F-érték: 6.864), azaz a kirajzolódó tendencia által jelölt különbségek statisztikailag szignifikánsnak tarthatók.



2.6. ábra. A magyarázó mellérendelések gyakoriságának változása, évcsoportok.

A magyarázatok és következtetések halmozása összefüggést létesít a tagmondatok között, leginkább pontosítva a kapcsolatot megelőző tagmondat jelentését, ugyanakkor – éppen a következtetések és magyarázatok nagy száma miatt – ezeknek az összefüggéseknek a konstruáltsága, természetellenessége vagy pontatlansága (hiszen bármeddig pontosíthatók) is a figyelem előterébe kerül; „ez pedig nem más, mint magyarázat, a magyarázatok, mi több, a tényeket agyonmagyarázó, *vagyis* végső soron e tényeket megsemmisítő, de legalábbis elködösítő magyarázatok halmaza" – ahogyan Kertész Imre *Kaddis*ának elbeszélője fogalmaz. Kertész Imre mindhárom, a korpuszban szereplő regénye *(Sorstalanság, A kudarc, Kaddis a meg nem született gyermekért)* az első öt hely valamelyikén található a magyarázó mellérendelések gyakoriságát tekintve, és szintén magas értékeket vesznek fel a következtető kapcsolattípusok esetében. Ez a regények tematikus összefüggésén túl a köztük lévő stiláris hasonlóságra hívja fel a figyelmet, amelyet eddig kevésbé tárgyalt a szakirodalom.[39] A stiláris és tartalmi szintek azonban nem elválaszthatók egymástól: az említett Kertész-művek az egyes eseményeket a társadalmi rend működésének szükségszerű következményeként írják le, céljuk ennek a működésnek a minél pontosabb magyarázata – legyen szó a holokauszt történetéről, az 50-es évek vagy a Kádár-korszak Magyarországáról. A *Sorstalanság*ban Köves felismerése éppen az, hogy története nem érthető meg

[39] Hasonlóan magas értékek jellemzik a *Jegyzőkönyv, Az angol lobogó* című kisregényeket, sőt a *Gályanapló* című naplóregényt is, amely újabb – ezidáig feltáratlan – összefüggésekre világíthat rá az életművön belül.

 DOI: 10.31400/dh-hun.2021.4.2415

utólagos nézőpontból, hanem az események egymásból következő, logikus rendjét kell bemutatnia, ha saját élményeiről kíván beszélni. Az utólagosságot elutasító utólagos tudás paradox egysége szervezi az egyszerre naplószerű és múlt idejű-visszatekintő narrációt,[40] és ez érhető tetten a mondatszerkezetekben is, hiszen a magyarázatok nemcsak a dolgok természetes egymásra következését mutatják be, hanem azok hátborzongató abszurditását is: „Már egészen más dolog azonban – tették rögtön hozzá – az Arbeitslager, *vagyis* munkatábor: ott az élet könnyű, a viszonyok és az élelmezés, járt híre, hasonlíthatatlan, s ez természetes is, *hisz* ott a cél is más elvégre" (Kertész Imre, *Sorstalanság*). Ritkán hangoztatott jellemzője a Kertész-prózának, ám az ok-okozati kapcsolatok ilyen természetellenessége egyben – ha mégoly sötét és kínzó módon, de – humoros hatást is kiválthat. Ugyanígy a *Kaddis* elbeszélője is a világrend alapvető működéséről számol be, amely működés a legszörnyűbb pillanatokban is magyarázható és érthető (hiszen „amire tényleg nincs magyarázat, az nem a rossz, ellenkezőleg: a jó"); a hosszú magyarázatok egyszerre gyűrik maguk alá az olvasót és győzik meg igazukról, valamint egyszerre hatnak parodisztikusnak, helyenként humorosnak, sőt akár patologikusnak,[41] mániákusnak, vagy logikailag inkonzisztensnek,[42] de mindenképp magukban hordozzák a megértés, a magyarázat és az (ön)felszámolás ellentmondásos kapcsolatát:

> […] ez a kérdés te vagy, pontosabban én vagyok, de általad kérdésessé téve, még pontosabban (és ezzel nagyjából doktor Obláth is egyetértett): az én létezésem a te léted lehetőségeként szemlélve, vagyis én mint gyilkos, ha a pontosságot a végletekig, a képtelenig akarjuk fokozni, és némi önkínzással ez meg is engedhető, hiszen, hál' isten, késő, mindig is késő lesz már… (Kertész Imre, *Kaddis a meg nem született gyermekért*)

Hasonló figyelhető meg a 20. század utolsó harmadában íródott más regények esetében is, amelyek nagy számban dolgoznak ki magyarázó és következtető kapcsolatot, például Nádas Péter *Emlékiratok könyve* című munkája és Krasznahorkai László két regénye, elsősorban *Az ellenállás melankóliája* (de említhető a korpuszból Esterházy Péter *Harmonia cælestis* vagy Spiró György *A jövevény* című műve is). *Az ellenállás melankóliájá*ban a korpuszon belül a második leggyakoribbak a magyarázó és harmadik leggyakoribbak a következtető típusú kapcsolatok (és magas az ok- és célhatározói alárendelések száma is). Az ezekkel kidolgozott ok-okozati viszonyok leginkább szereplői tudattartalmakhoz kötődnek, azaz – a *Sorstalanság*hoz hasonlóan – az egyes szereplők alkalmazkodásáról értesítenek az egyre inkább fenyegető mindennapokhoz, ahogyan érthetővé és magyarázhatóvá kívánják tenni környezetük működését:[43] „Mindez már

[40] „A szöveg megkísérli csinálni is, amit mond, kétféleképpen beszéli el egyszerre a történetet, visszatekintő és szinkrón perspektívából." Vári György, *Kertész Imre:* Buchenwald fölött az ég (Budapest: Kijárat Kiadó, 2003) 57.

[41] Vö. Radnóti Sándor, „Auschwitz mint szellemi életforma," *Holmi* 3, 3. sz. (1991): 372–378, 377.

[42] Vö. Lengyel Imre Zsolt, „»egy lekerekített, üvegkemény tárgy.« Kísérlet a *Kaddis a meg nem született gyermekért* határainak olvasására," in Gintli Tibor, Kelemen Pál, Kulcsár-Szabó Zoltán és Szemes Botond, szerk., *Kertész másként* (Budapest: Kijárat Kiadó, 2021), megjelenés előtt.

[43] A Kertész esetében megállapított parodisztikus-humoros hatás Krasznahorkai stílusára is érvényes – ugyan bírálólag, de erre hívja fel a figyelmet például Horkay Hörcher Ferenc is, amikor megjegyzi,

senkit sem lepett meg igazán, *hiszen* az uralkodó állapotok természetesen ugyanúgy éreztették hatásukat a vasúti közlekedésben, mint bármi másban." Krasznahorkai korai prózájában „tudatkövető" technika érvényesül, amely a szereplők belső vagy kimondott beszédét dolgozza össze,[44] amely során a szereplők következtetései és motivációi felől ismerjük meg az ábrázolt világot – például: „[…] az »alapos szemügyre vételt« korántsem könnyítette meg, a tömeg nagyságára vonatkozó kérdéssel *ugyanis* (*hisz* mit figyeljen meg rajta, ha minden változatlan?) sehogy sem boldogult…" Ezáltal az ábrázolt világ összefüggései „belülről", a szereplői tudatokon keresztül bomlanak ki, aminek köszönhetően átélhetővé válnak az amúgy abszurdnak tűnő eseménysorok is.

Nádas Péter *Emlékiratok könyve* című regényében (amelyben arányosan az ötödik legtöbb magyarázó és második legtöbb következtető összetétel található a korpuszban) a következtetés és magyarázat halmozásának e kettős tulajdonsága, azaz az ok-okozati összefüggések feltárása és egyben természetellenességük, megalkotottságuk színre vitele főként abban érhető tetten, ahogyan az elbeszélők részletes pontosságra törekvő leírásai a helyzetek bármeddig pontosítható, így lezárhatatlan magyarázatának érzetét keltik, és látványosan bonyolult viszonyrendszereket hoznak létre:

> […] egyáltalán nem nagy baj, ha ezt gondolom róla, mert ő meg tudja rólam, amit szeretnék titokban tartani, hiszen Thea előtt Melchior igazán nem fegyelmezte a tekintetét, amit tehát titoknak szántunk, mármint, hogy nem egyszerűen jó barátok, hanem szerelmesek vagyunk, ez bizonyára előtte sem titok. (Nádas Péter, *Emlékiratok könyve*)

A hasonlító kapcsolatok (időbeli változás)

Az eddigi tendenciák főként a hosszúmondatos prózákat érintették, és az idővonal két végén mutattak kiugró értékeket. A hasonlító kapcsolatok relatív gyakorisága ezzel szemben a 20. század első felében éri el a csúcspontját. *(2.7.*[45] és *2.8. ábra)* A korpusz alapján elmondható, hogy a korszak prózája ritkábban dolgoz ki kapcsolatot a tagmondatok között, és a viszonyok létesítése helyett inkább az önálló jelenetek vagy képek egymás mellé helyezésével kívánja a helyzetek leírását adni. Ez alól azonban kivétel a hasonlatok alkalmazása, amelyek éppenhogy távoli területek összekapcsolására képesek – például: „[…] a gyeplőszár olyan egyhangúan rángatózik, *mint* a falusi temetőben reszket a nyárfa" (Krúdy Gyula, *A vörös postakocsi*). A korpuszból Márai Sándor *A gyertyák csonkig égnek,* Gelléri Andor Endre *A nagymosoda,* Szomory Dezső *A párizsi regény* című műveire és Krúdy Gyula regényeire jellemző leginkább ez a kapcsolattípus.

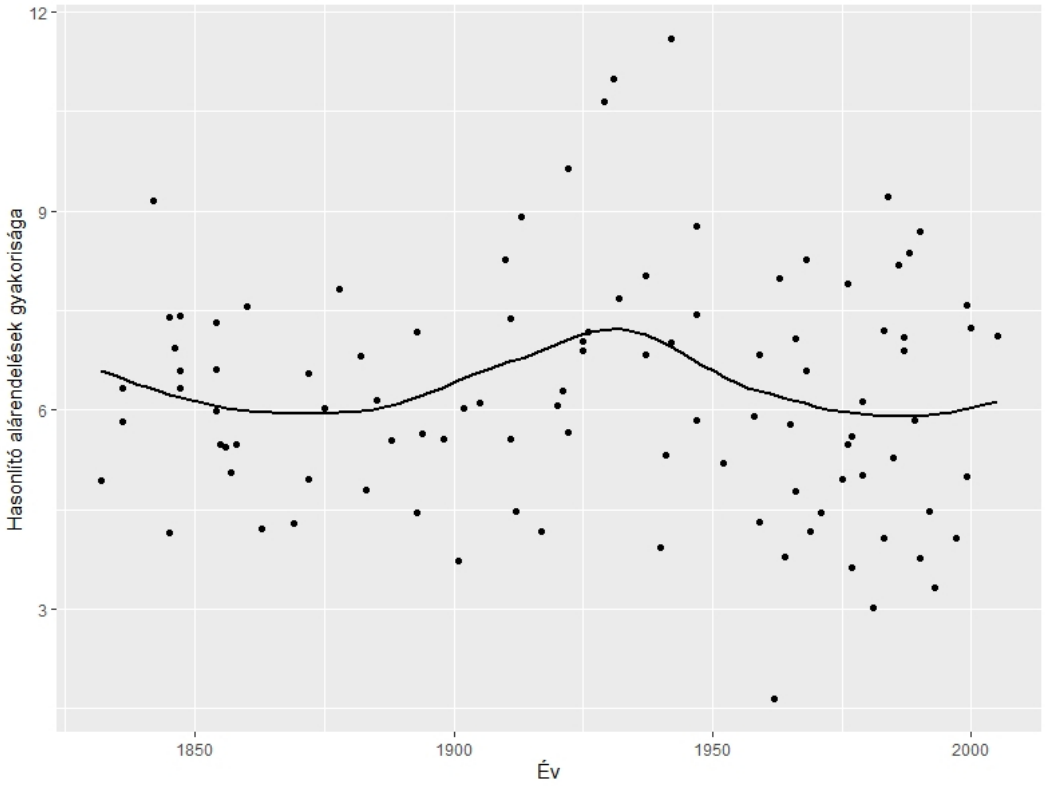
---

hogy Krasznahorkai körmondatai finoman szerkesztettek és szépek, „azonban ez az erősen artisztikus nyelv alkalmanként a paródia határára sodorja az írói elbeszélésmódot." Horkay Hörcher Ferenc, „A pusztulás anatómiája," *Holmi* 2, 5. sz. (1990): 576–580, 579.

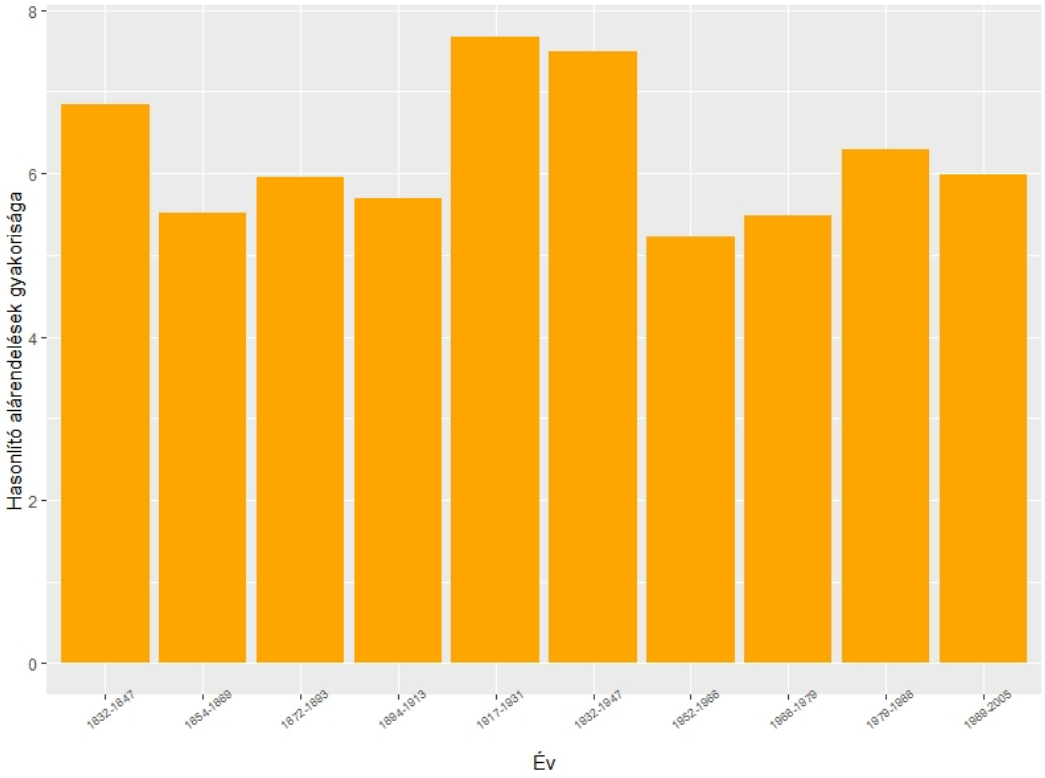
[44] Szirák Péter, „Látja, az egész nincsen: Krasznahorkai László prózájáról," *Alföld* 39, 7. sz. (1993): 41–51, 47.

[45] Az egyszempontos varianciaanalízis (ANOVA) alapján csak a három nagyobb időszak (1832–1909, 1910–1949 és 1950–2005) között mutatható ki korreláció a következtető kapcsolattípus gyakorisága és a megjelenési évek között (F-érték: 12.67), azaz a kirajzolódó tendencia által jelölt különbségek csak ezek között az időszakok között tarthatók statisztikailag szignifikánsnak.

 DOI: 10.31400/dh-hun.2021.4.2415



2.7. ábra. A hasonlító alárendelés változása, 1832–2005.



2.8. ábra. A hasonlító alárendelés gyakoriságának változása, évcsoportok.

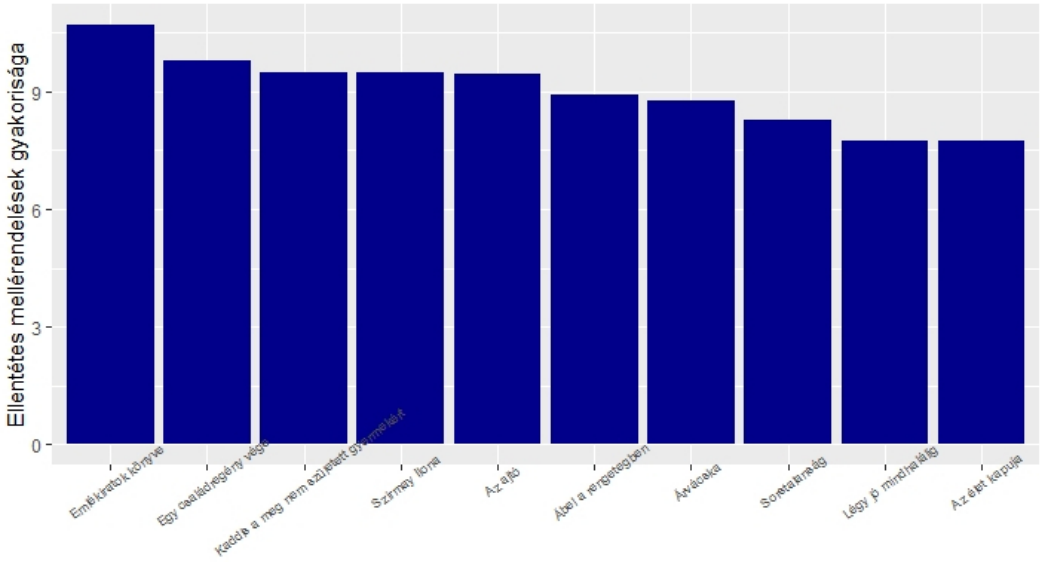
Ezek a grafikonok újabb, időben is elkülöníthető paradigmát jelölnek ki. Míg a vonatkozói alárendelések nagy száma a 19. század első felében, a magyarázó-következtető kapcsolattípusoké pedig a 20. század végén (illetve részben a 19. század első felében) figyelhető meg, addig a 20. század első felében a hasonlatok alkalmazása a jellemző.

Ezek a stiláris, mondatszerkesztési paradigmák egyúttal a világhoz fűződő, episztemológiai viszony történeti különbözőségeire is utalnak. A vonatkozói alárendelések a mondatok egy szereplőjének adják részletesebb meghatározását azáltal, hogy egy új jelenetben szerepeltetik, ami egyúttal a mondat jelentésének rögzítését is szolgálja. Ezzel szemben a folyamatos pontosítás és kifejtés állandóan új értelmezések felé nyitja meg a mondatot, így a világhoz kevésbé stabil hozzáférést tételez fel, ugyanakkor inkább képes azt belső perspektívából megfigyelhetővé tenni. A harmadik paradigmát jelentő hasonlatok két terület összehasonlítása révén tudnak új információt szolgáltatni – ebben az esetben nem a jellemzés vagy a kifejtés, hanem egy analógia létrehozása („Szép vagy, mint a rózsaszál”), vagy a különbségek artikulálása („Erősebb, mint a te apukád”) játszik központi szerepet.

Eredmények 2.

A regényekben az egyes kapcsolattípusok gyakoriságát nemcsak idővonal mentén, hanem összehasonlító módon, oszlopdiagramok formájában is ábrázolhatjuk (ahogyan eddig a regényenkénti évcsoportok esetében). Ekkor nem a történeti tendenciák, hanem az egyes művek sajátosságai válnak megfigyelhetővé. A csökkenő sorrendbe rendezett ábrák kezdeti és záró szakaszai bírnak a legnagyobb információértékkel, hiszen a legnagyobb és legalacsonyabb értékeket mutató regényekről mondható el, hogy az adott típus jobban/kevésbé jellemző rájuk a korpusz egészéhez képest.

Ellentétes kapcsolatok (legjellemzőbb regények)



3.1. ábra. Ellentétes mellérendelés – Az első 10 regény.

 DOI: 10.31400/dh-hun.2021.4.2415

A *3.1. ábrá*n az ellentétes mellérendeléseket leggyakrabban kidolgozó regények láthatók. Feltűnő, hogy Nádas Péter két regénye található az első két helyen, azaz ezekben a szövegekben a legnagyobb az ellentétes tagmondatkapcsolatok relatív gyakorisága. Az *Egy családregény vége* – amelyben gyakoriak a feltételes és ok-/célhatározói kapcsolatok is, lásd *Függelék* 1.1. – a gyerek elbeszélő perspektívájából mutatja be az 1950-es évek Magyarországát, amely perspektívát elsősorban a felnőtt–gyerek, a képzelet–valóság, az egyén–családi/politikai közösség feszültsége jellemez a legjobban. Ezek az ellentétek szervezik az elbeszélő világát, aki azáltal mutat rá környezete ellentmondásaira, hogy megpróbál alkalmazkodni hozzájuk (ez a kettősség nagyon hatásosan képes kiaknázni a gyerekperspektívában rejlő poétikai lehetőségeket), mígnem a regény végére a teljes elutasításhoz jut el – ezt a szembehelyezkedést érzékelteti a szöveg híres zárómondata is: „Nem." Az ellentétek az eltérő horizontból megszólalók szólamainak különbségeiből is adódnak, ugyanis a gyerek elbeszélőével a bibliai hagyományokig visszanyúló nagypapa és a pártfunkcionárius apuka szólama is keveredik a szövegben. Mindennek megfelelően több értelmező is rámutatott a regényt szervező oppozíciók működésére;[46] amelyet Bazsányi Sándor a János evangéliumából vett mottóban megfogalmazott feszültségre vezet vissza, („És a világosság a sötétségben fénylik, / *de* a sötétség nem fogadta be azt."), miközben a nagypapa által elmesélt történetrészlet és a mottó szoros kapcsolatára is felhívja a figyelmet:[47]

> Egy pillanatra, míg valahová taszítják, lökik, látja a középső meztelen szemét; nem a szemet, csak a pillantás fényét rögzíti; és amikor sötét, forró szobájában imádkozik, s várja a hírt, hogyan vélekedik a történtekről a Szanhedrin és rabbi Abjatar, és imádkozik, azt kéri imáiban az Úrtól, hogy világosítsa meg elméjét a történtek felől, akkor a nagy sötétségben, amiben fuldokol, néha felsejlik ennek a pillantásnak a fénye, de nem érti a fényt; Simon nem érti, sötétsége a fényt nem tudja befogadni, pedig a fény a sötétségben világít; abba kéne kapaszkodnia, abba a fénybe, de ő nem érti és világosságért imádkozik az Úrhoz, de közben nem veszi észre a fényt, annak a tekintetnek a fényét, amit az Úr elküldött.

Az idézetben (ahogy a mottóban is) a fény és a sötétség ellentéte mellett a megértés/befogadás versus nem-értés/nem-befogadás ellentéte, illetve a cselekvések eredményeinek különbsége (például az imádkozás, amely nem talál meghallgatásra) válik hangsúlyossá, amely ellentéteket elsősorban ellentétes tagmondatkapcsolatok dolgoznak ki – ez a stiláris-tematikai összefüggés a szöveg alapvető jellemzőjének tekinthető.

Balassa Péter monográfiájában arról ír, hogy a *Családregény* igazi tétje „az egység és egyben látás megőrzésében létrejövő *vagy* elpusztuló *személyiség* (ezt vette észre a korabeli kritikák többsége, amikor többször az „igen" és a „nem" oppozíciójával írta le a könyvet),"[48] és maga is oppozíciópárok fontosságára hívja fel a figyelmet – például a regényben többször felbukkanó és a zárlatra motívummá váló fekete-fehér kövezetet is

[46] Kálmán C. György szerint például a különböző szólamok összeütközésének köszönhetően a „szerkezet szétfeszülése" figyelhető meg a Bazsányi Sándor által „tagadáselvű regény"-ként jellemzett szövegben. Kálmán C. György, „Csalárd regény: Újraolvasás", *Jelenkor* 44, 10. sz. (2002): 1051–1056, 1052; Bazsányi Sándor, *Nádas Péter* (Pécs: Jelenkor Kiadó, 2018), 62.

[47] Uo., 60.

[48] Balassa Péter, *Nádas Péter* (Pozsony: Kalligram Kiadó, 1997), 97, kiemelés az eredetiben.

mint a művön végigvonuló ellentétek metaforáját azonosítja: „részben a fekete-fehér egyértelműség és a véletlen kombinációjának játszmája ez, részben a két szín poláris ellentétével való játszadozás, mely a mű egész kontextusában az igen/nem, élet/halál bináris oppozíciójának, nyers kérdezz-felelek-jének a metaforája."[49] Elemzése az ezeket összefogó, végső, és az egész Nádas-prózát meghatározó ellentétre fut ki, azaz magában az ellentétben rejlő kettősségre, miszerint az *Családregény* mindenekelőtt a szembenállást felszámoló *dialógus* és a valódi kapcsolódást nélkülöző *dualitás* oppozícióját, dilemmáját viszi színre.[50] Másképpen megfogalmazva: Nádas első regényének alapkérdése a tőlünk elválasztott Másikkal való azonosulás lehetősége és a felszámolhatatlan különválasztottságunk bizonyossága közti feszültségre vonatkozik. Ez a dilemma az ellentétes tagmondatkapcsolatok gyakori alkalmazásával kerül kifejtésre, azaz ebben az esetben egyértelmű kapcsolat tételezhető a szöveg tematikus és stiláris szintjei, a történetet szervező ellentétek és az ellentétes tagmondatkapcsolatok nagy száma között. Sőt, ahogy már a kifejezésben rejlő oximoron (ellentétes kapcsolat) is tanúsítja, maga a grammatikai szerkezet is hordozni képes a Balassa által megfogalmazott dilemmát összekapcsolódás és különbözőség kettőségéről: az összetett mondat e fajtája ugyanis egyszerre állítja szembe és köti össze a tagmondatokat.

Az ellentétes kapcsolatok nagy száma az *Emlékiratok könyve* hasonló szerveződéséről értesít, topológiai-logikai karakterének szintén fontos összetevője az ellentétezés. A nagyregény hosszú, többszörösen összetett mondatait eddig inkább azok – tagadhatatlanul erős – hatása felől elemezték[51] vagy a világ leírásának egy sajátos módozataként azonosították,[52] ám ritkán vizsgálták tüzetesebben a mondatok szerkezetét. Ez alól kivétel Jolanta Jastrzębska nyelvészeti megközelítésből írt, fontos megfigyeléseket tartalmazó tanulmánya, amelynek általános megállapításai azonban nélkülözik a kvantitatív alapozást.[53]

Ezidáig a regényben nagyszámú következtető és magyarázó mellérendelések kerültek tárgyalásra, amely a részletes pontosságra és az összefüggések feltárására törekvő elbeszélő[54] stílusának fontos jellemzőjére világított rá. Mindezt az ellentétes kapcsolatok gyakoriságával egészíthetjük kis. Fontos látni, hogy az ellentétek és a következtetések/magyarázatok különböző funkciót látnak el a mondatokon belül. A mondatok szerkezete ugyanis olyan alapsémára vezethető vissza, amely nagyobb gondolati egységek részletes kifejtéseiből és ezeknek az egységeknek a szembeállításából jön létre. A részletes kifejtés a következtetések és magyarázatok mellett alárendelő kapcsolatokkal és mondatrészek halmozásával[55] történik. A halmozás az érzékek (nem értelmező

[49] Uo., 122.

[50] „Dialógus és dualitás ellentéte az igazi, általános oppozíció az egész életműben." Uo., 139.

[51] Például: Boros Gábor, „Nagy fúga: Adalék az *Emlékiratok könyve* értelmezéséhez," in Balassa Péter, szerk., *Diptychon: Elemzések Esterházy Péter és Nádas Péter műveiről,* 169–179 (Budapest: Magvető Kiadó, 1988).

[52] Bagi Zsolt, *A körülírás* (Pécs: Jelenkor Kiadó, 2005).

[53] Jolanta Jastrzębska, „Körmondatos stílus a modern magyar prózában. Nádas Péter: *Emlékiratok könyve* példáján," *Magyar Nyelv* 89, 1. sz. (1993): 26–40.

[54] Jelen elemzés nem tér ki az elbeszélők közötti különbségre és nem foglalkozik külön Krisztián mondatainak szerkezetével, amelyek bizonyosan torzítják a többi elbeszélő stílusáról mondottakat.

[55] A legtöbbször szinonim szavak halmozására többen felhívták már a figyelmet, vö. Jastrzębska, „Körondatos stílus," 29–30; Bernáth Árpád, „Mi háromban", in *Diptychon,* 137–157. A kutatás külön

 DOI: 10.31400/dh-hun.2021.4.2415

jellegű) működésével hozható kapcsolatba és inkább a szöveg érzéki dimenzióit erősíti, míg a magyarázatok és alárendelések az értelmi összefüggések feltárásáért felelnek.[56] Az ellentétes kapcsolatok többek között ezeknek a részletes kifejtést tartalmazó nagyobb egységeknek a szembeállításáért felelősek – az egységek elhatárolását gyakran pontosvesszők is erősítik. Terjedelmi okok miatt most csak egyetlen hosszú mondatot idézek, amelyben jelölöm a szembeállított gondolati egységeket (és amelyben a második egység éppen a „részletek részleteiben elvesző analízissel" szemben határozza meg magát).

> [1] Ám akkor beszélnem kell arról is, az igazság nevében és a teljesség kedvéért igyekezve eltüntetni azon látszatot, ami ezidáig keletkezhetett, miszerint egész akkori életem csupán nyúlós keservekből, szégyenteljes kegyetlenkedésekből, siralmas vereségekből, elviselhetetlen, igen, elviselhetetlen szenvedésekből állott volna, nem, beszámolóm kétségtelen egyoldalúságát ellensúlyozandó hangsúlyoznom, hogy nem és nem! mert az öröm és az élvezet természetesen legalább oly jellemző része volt, a szenvedés pedig talán csak azért hagy mélyebb nyomot, mert igénybe véve a tudat képességét a gondolkodásra, töprengéssel és rágódással megnyújtja az időt, míg az igazi öröm, mely minden tudatost megkerül, az érzékire szorítkozik, csupán annyi időt ad magának és nekünk, amíg létezik, miáltal végzetesen véletlenszerűnek és esetlegesnek tűnik, különállónak és elszakítottnak a szenvedéstől, s így a szenvedés hosszú és bonyolult történeteket hagy az emlékezésnek, míg az öröm percnyi felvillanásokat; [2] de félre a részletek részleteiben elvesző analízissel s a részletek értelmét kereső filozófiával is! ami szükséges ugyan, ha szemügyre akarjuk venni, hogy lelkiekben milyen gazdagok vagyunk, s miért ne vennénk szemügyre élvezettel? ám éppen mert e gazdagság oly végtelen, s a végtelen a leginkább megfoghatatlan dolgok közé tartozik, analitikus sietségünkben hajlamosak vagyunk oknak, sérüléseink, torzulásaink, szenvedéseink, lelki betegségeink, mondjuk ki, hogy nyomorékságunk kiváltó okának tartani egészen egyszerű és végső soron természetesnek mondható történéseket, mivel éppen e történések egészét veszítettük el bizonyos önkényesen kiválasztott részletek javára a szemünk elől, a részletek végtelen gazdagságától való rettenetünkben megálljt parancsoltunk magunknak ott, ahol még jócskán tovább lehetne menni, rettenetünk bűnbakot csinál, kis áldozati oltárokat emel, szertartásosan szúr levegőbe késeket, s így nagyobb zavart okoztunk, mintha egyáltalán nem is gondolkodtunk volna magunkról, bizony boldogok a lelki szegények! ne gondolkodjunk hát! adjuk át magunkat szabadon és minden elfogultságtól mentesen annak a kellemes

vizsgálta a 100 regényben a halmozások gyakoriságát, amely eljárás az *e-magyar* által elemzett szövegekben az egymást követő, vagy vesszővel elválasztott melléknevek (elsősorban jelzős szerkezetek) és főnevek hármas vagy négyes sorozatának azonosításán alapult. Tehát a bonyolultabb szerkezeteket (pl. jelzős határozók, mint az alábbi példamondat elején) ezzel a módszerrel nem lehet azonosítani, így az eredmények is csak az egyszerűbb felsorolásokra, halmozásokra vonatkoznak. A jelzős szerkezetek halmozásokban kiugróan magas értéket mutatnak Kaffka Margit regényei, de az *Emlékiratok könyvé*ben is gyakoriak az ilyen szerkezetek, lásd a *Függelék* 3. pontjában.

[56] Vö. Szemes Botond, „Borítójáról a könyvet: Szellemi, anyagi és történeti meghatározottságok az *Emlékiratok könyvé*ben," *Kalligram* 27, 5. sz. (2018): 70–71.

> érzetnek, hogy anyánk betegágya mellett ülünk a földön, fejünket úgy hajtva testét takaró hűvös selyempaplanára, hogy ajkunk a karjának csupasz bőrén pihenjen, hajunkban érezzük finom ujjait, a hajtövek kellemes borzongását, bizsergését, mert a másik kezével beletúrt zavarában, mintegy vigasztalón, kipróbálva, hogy szavainak élét tudja-e venni a mozdulattal, s bár e borzongás lassan testünk egész felületét birtokába veszi, [3] az elhangzottakat mégsem lehet ilyen egyszerűen visszavonni: bennünk is fölrémlett már a sejtés, hogy az apánk talán nem is apánk, s ha ama bizonyos két férfi között ő nem tudott dönteni, akkor e sejtés most csaknem bizonyossággá erősödhetik, [4] erről viszont, megértjük, nem lehet többet mondani, legyünk csendben hát, s akkor pontosan érezni fogjuk, hogy mindaz, mi szavaink hatására emlékként fölvillant és immár elült bennünk, bármennyire fontos, bármennyire meghatározó legyen, csupán érzelmeink és valódi érdeklődésünk hátterében áll, mert abban a térben, melyben mindezt megpróbáljuk megragadni, sajátunkként mérlegelni, melyben valódinak nevezhető történéseink zajlanak, teljesen egyedül vagyunk, ahhoz nekik nincsen és nem is lehet semmi közük.

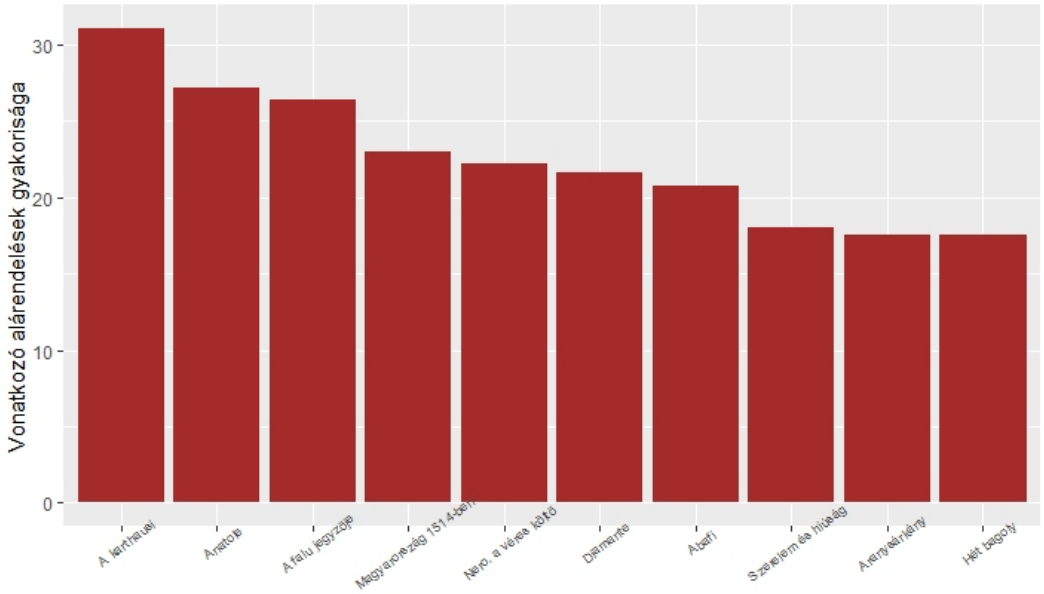
Az ehhez hasonló szembeállítások ezúttal is a regényt szervező oppozíciókkal hozhatók összefüggésbe. A látszat–valóság, a testi–akaratlagos/szellemi, az (el)képzelt–megvalósult ellentéteit dolgozzák ugyanis ki az ilyen ellentétes kapcsolatok, amelyek általában az elvárthoz, az érzethez, az elképzelthez vagy a korábbi emlékekhez képesti történéseket, gondolatokat vezetik fel. A regénynek ezt a szerkezetét erősítik a kisebb hatókörű ellentétes kapcsolatok is, amelyek a nagyobb gondolati egységeken belül létesítenek ellentétes viszonyt a tagmondatok között. Fontos látni, hogy az ellentétezések, ahogyan a *Családregény* esetében, itt sem a teljes elválasztásért felelősek, hanem a különbözőség és azonosság (dualizmus és dialógus) kettősségét hozzák játékba – ezáltal a regény központi problémáját[57] hordozzák magukban. Bagi Zsolt egy másik szöveghely elemzésekor hasonló következtetésre jut, amikor megjegyzi, hogy „egy »viszont« ellentételezése vezeti be a mondatba az új eseményt, az teremti meg benne helyét. Ez a »viszont«, amelynek rokonait oly gyakran fellelhetjük Nádas bekezdésnyi mondataiban, a mondatba mint test-írás alapegységébe egy olyan törést ír, amely a mondat felütésében adott szituációt anélkül borítja fel, hogy azzal való közösségét felmondaná. Ha egy mondat állna ellentétben az előzővel, akkor a törés elválasztaná a kettőt, és izolálná őket egymással szemben, így azonban bensőséges kapcsolatuk fennmarad; egymásba folynak, de megjelenik bennük egy vetemedés vagy gyűrődés, amely a folytonosságban a megszakítottságot képviseli."[58]

[57] Ez a probléma a személyes identitás és az interszubjektív világban átélhető közösségiség kapcsolatára vonatkozik, ahogyan azt Bagi Zsolt is megjegyzi a regény egy, a másikkal való azonosulás lehetőségét és lehetetlenségét ellentétes szerkezetek halmozásával körüljáró mondata kapcsán: Bagi, *A körülírás,* 63–65.

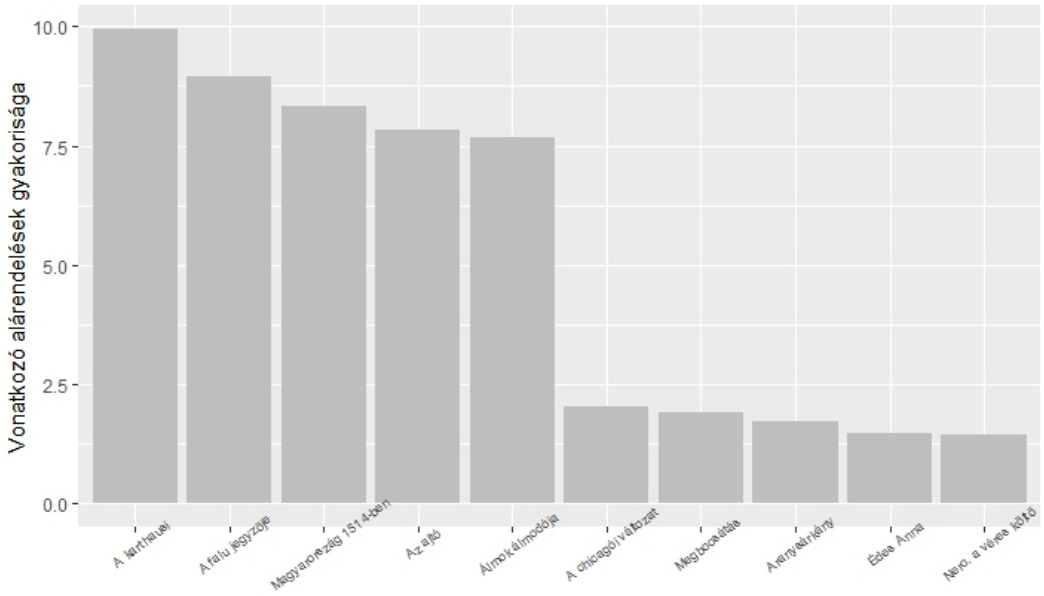
[58] Uo., 112.

 DOI: 10.31400/dh-hun.2021.4.2415

Prototipikus vonatkozói alárendelések és feltételes kapcsolatok (legjellemzőbb regények)



3.2. ábra. Prototipikus vonatkozói alárendelés – Az első 10 regény.



3.3. ábra. Feltételes alárendelés – Az első és az utolsó 5 regény.

A *3.2. ábra* a prototipikus vonatkozói, a *3.3. ábra* a feltételes alárendeléseket leggyakrabban, illetve legkevésbé gyakran alkalmazó regényeket mutatják. Eötvös József mindhárom regényében kifejezetten gyakorinak mondhatók ezek a kapcsolattípusok, ami a már tárgyalt anaforisztikus, ritmikus-retorikus prózastílusra vezethető vissza, hiszen az azonos pozícióban lévő mellékmondatok rendszeres halmozása nemcsak

a vonatkozói, hanem a feltételes mellékmondatokra is érvényes.[59] Ám korántsem szükségszerű a két típus ilyen összefüggése – ahogyan azt például Kosztolányi Dezső műveinél láthatjuk. Míg a *Nero, a véres költő* az ötödik, az *Aranysárkány* a kilencedik legnagyobb értéket veszi fel, és az *Édes Anna* is magas értéket mutat a vonatkozói alárendeléseket tekintve (*Függelék* 1.1.), addig e három szövegben a legkevésbé gyakoriak a feltételes tagmondatkapcsolatok. Fontos példa lehet ez arra, hogy nemcsak egy típus gyakorisága, hanem annak hiánya is beszédes lehet egy író stílusával kapcsolatban.

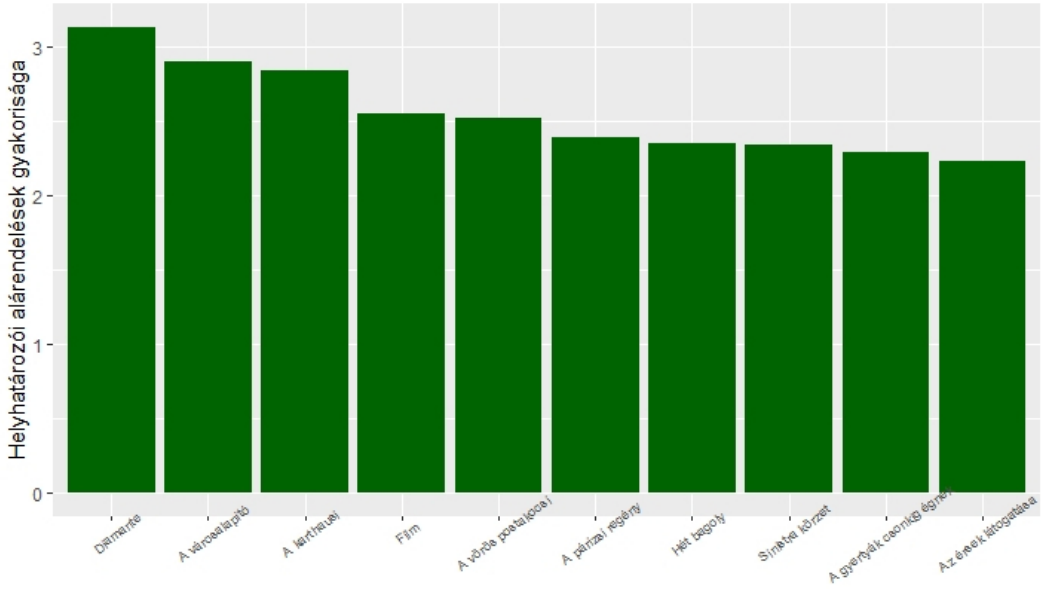
Kosztolányi rövidmondatos prózája mentes a tagmondatok halmozásától és a másod- vagy harmadfokú alárendelésektől, ahogyan a hiányos szerkezetek se jellemzők rá: éppen az egyszerű, hagyományos (névszói alanyok és igei állítmányok viszonyát kidolgozó) mondatok adják ennek a prózapoétikának az egyediségét.[60] A vonatkozói névmással bevezetett mellékmondatok, ahogy fentebb látható volt, a leírások pontosságát erősítik, hiszen azáltal, hogy egy szereplőt új jelenetben teszik megfigyelhetővé, annak jellemzéséhez és meghatározásához járulnak hozzá. Erre lehet példa a Senecától származó idézet a *Nero, a véres költő*ből, amelyben a szereplőket jellemző alárendelések egyben a regény gyújtópontjában lévő két, egymással nem összebékíthető szerepet (költő és császár) jelölik ki: „Hibát követtem el, tudom. Akkor a legnagyobbat, mikor odaadtam magamat Neronak, én, *aki* költő vagyok, őneki, *aki* csak császár." A vonatkozói névmásokkal kidolgozott mellékmondatok továbbá az elbeszélő külső, megfigyelő pozíciójához is hozzájárul; az átélt, szabad függő beszéd más kapcsolattípusokat is előnyben részesít, ám ezek nem jellemzők Kosztolányi vizsgált regényeire. Külön érdekes ebből a szempontból a feltételes alárendelések alacsony száma, amely a tárgyalt művek logikai karakterének alapvető jellemzőjének tekinthető: a regényekben a lehetőségek és feltételek számbavétele helyett a konkrét világ tárgyilagos leírására korlátozódik az elbeszélés. Ebben a konkrétságban jelölhetjük ki Kosztolányi prózastílusának alapját, amely stílus gazdagságát nem a mondatszerkezeti bonyolultság, hanem a pontos, érzékletes szóhasználat adja.

[59] Például: „S ha végre a vész kitört, ha minden vágyait s érzelmeit egy tárgyba pontosítva, ifjú lelkének egész hatalmával egy kedves lényt fogott körül; ha egy asszonyt, nem – ha egy angyalt talált, kit határtalan képzetének kincseivel fölruházott, s leborulva bálványa előtt, minden gyönyört s bánatot, mindent, mi szörnyű s fenséges életünkben, egyszerre, egy pillanatban érezte; ha istennek s elátkozottnak gondolva magát, egy pillanatban bánat- s örömének terhe alatt lesüllyedett, s ami keblét tölti, több, mint hogy magába rejteni, nagyobb, mint hogy kimondani tudná; ki találna szót ezen érzelme leírására?!" (Eötvös József, *A falu jegyzője*) Idézi Herczeg, „Eötvös körmondatai (Folytatás)," 175.

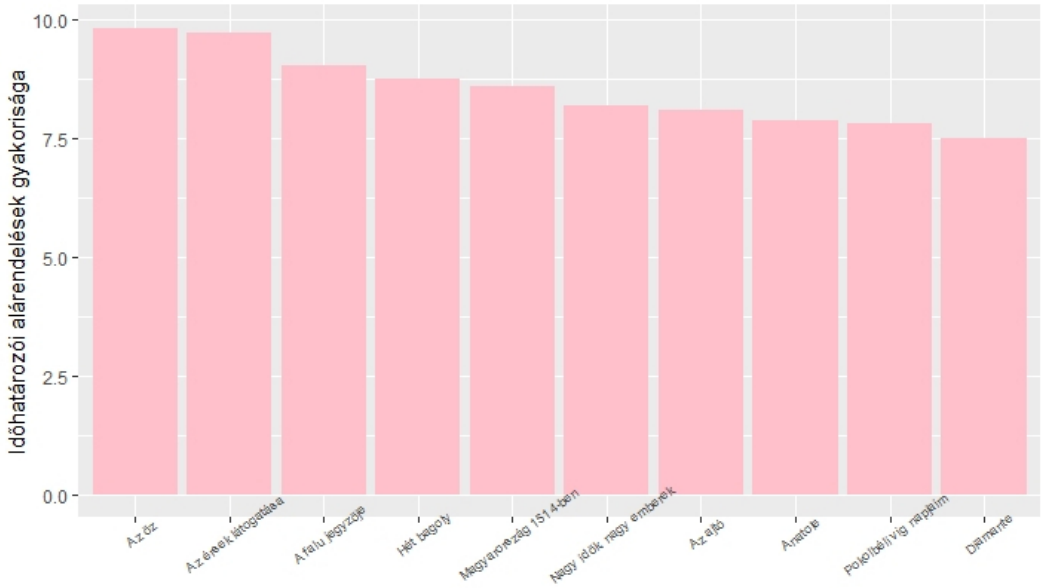
[60] Vö. Herczeg Gyula, „A prózaíró Kosztolányi," *Magyar Nyelvőr* 110, 4. sz. (1986): 432–449.

 DOI: 10.31400/dh-hun.2021.4.2415

Hely- és időhatározói kapcsolatok (legjellemzőbb regények)



3.4. ábra. Helyhatározói alárendelés – Az első 10 regény.



3.5. ábra. Időhatározói alárendelés – Az első 10 regény.

Röviden említhetjük még a *3.4.* és *3.5. ábrá*n bemutatott hely-, illetve időhatározói kapcsolatok gyakoriságát is. Ezek a típusok a vonatkozói alárendelések aleseteinek tekinthetők, amelyek helybeli vagy időbeli viszonyok kidolgozásáért felelősek. A helyhatározói kapcsolatok nem meglepő módon azokban a regényekben a leggyakoribbak, amelyekben amúgy is magas a vonatkozói névmásokkal bevezetett mellékmondatok

száma, vagy amelyeknek középpontjában egy terület leírása áll: ilyen Konrád György *A városalapító,* Mészöly Miklós *Film* és Bodor Ádám *Sinistra körzet,* valamint *Az érsek látogatása* című regénye. Ez utóbbiakban az időbeliséget kidolgozó kapcsolatok is gyakorinak tekinthetők, hasonlóan Krúdy Gyula vizsgált regényeihez *(A vörös postakocsi, Hét bagoly),* amelyek szintén magas értékeket mutatnak mind a két kategóriában. Ez az események idő- és térbeli viszonyainak részletes kidolgozottságát jelöli, ami fontos szempontot kínálhat a regények stílusának leírásához.

Krúdy prózájában például a helyszínek és időpontok részletes jellemzése a különböző síkok egymás mellé rendezését és összehasonlítását szolgálják. Hely és idő gyakran nem is elválasztható regényeiben egymástól, amennyiben egy helyszín is a múltra vonatkozó emlékeket hordoz és az idősíkok összehasonlításának teremt alapot, például:

> A régi Arany Sas fogadóban szállott meg, ahol már egyetemi hallgató korában is lakott: akkor egy szobában, most kettőben, mert legényének is ott kellett aludnia, és éjjel levetkőztetni, lábát, kezét megdörzsölni, betakarni, és jó éjszakát mondani, mint otthon Ungban, az ősi házban, ahol a föld, a mező és a patak az első foglalásból maradott meg az Alvinciak kezén. (Krúdy Gyula, *A vörös postakocsi*).

Ahogyan az idézetben is látható, a pontos idő- és térbeli kidolgozottságot erősíti a hasonlatok gyakori alkalmazása is, amely egyfelől a különböző síkok összehasonlítását teszi lehetővé (a példában az Ungon töltött gyerekkort és az Arany Sas fogadóban telt egyetemista éveket), másfelől kevésbé az időpontok és helyszínek, hanem inkább a mentális tartományok között teremt kapcsolatot. A kötőszavak nélküli, felsorolás jellegű mellérendeléseken túl ezek a kapcsolattípusok tekinthetők gyakorinak Krúdy korpuszba felvett két regényében, amelyek sajátosságát így a különböző (valós, vagy mentális) területek és idősíkok részletes leírásában és a közöttük való mozgásban ragadhatjuk meg.

Eredmények 3.

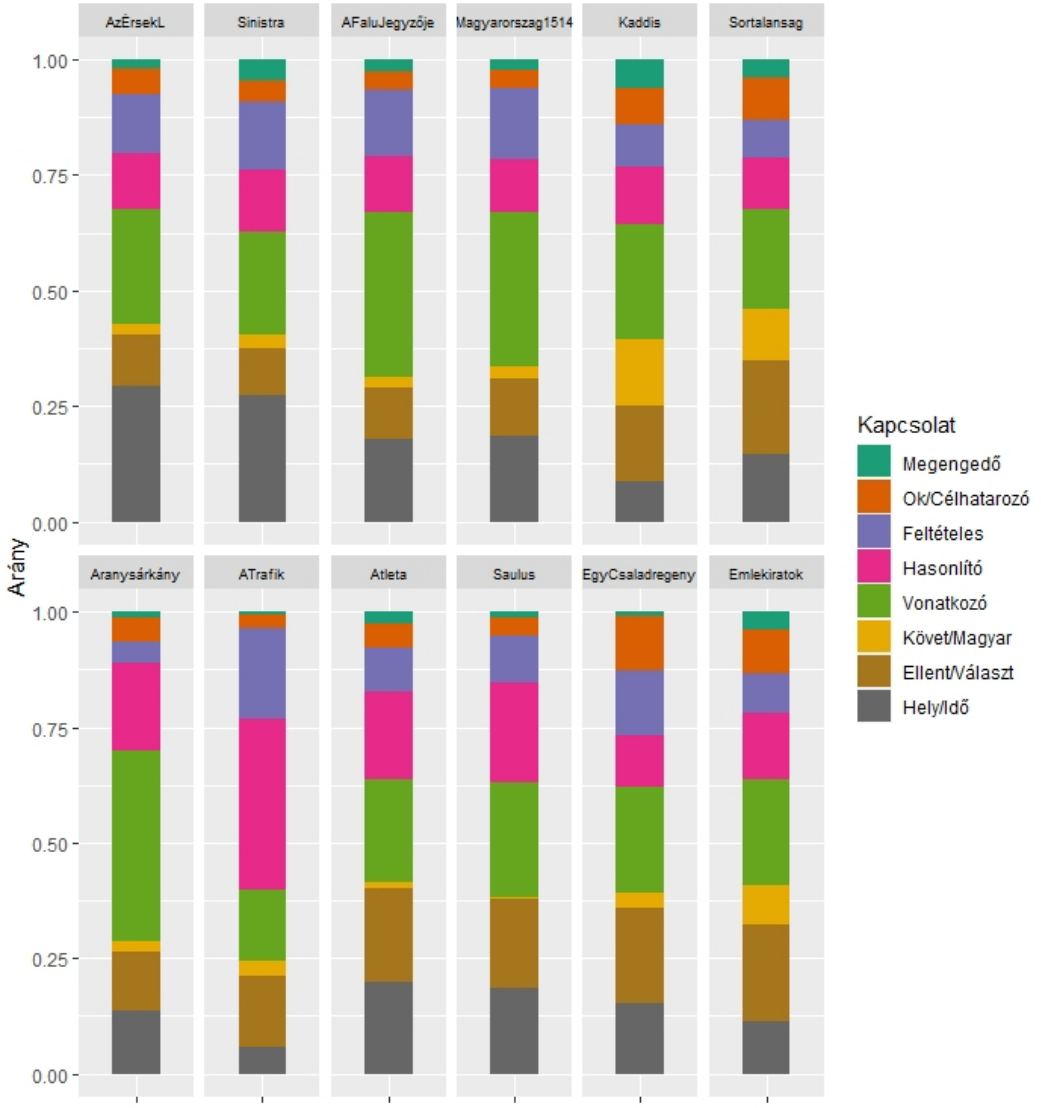
A harmadik lehetőség az eredmények ábrázolásában, amikor a regények belső szerveződésére összpontosítunk. Ebben az esetben nem a relatív gyakoriságokat, hanem csupán azt számítjuk ki, hogy összesen hány kidolgozott kapcsolat található egy szövegben, és ennek hány százalékát teszik ki az egyes típusok. Mindez természetesen csak a vizsgált típusokra vonatkozik és azoknak sem mindegyikére: azon túl, hogy az ábrák nem mutatják a kötőszóval nem kidolgozott, valamint a *hogy* kötőszós mellékmondatokat, nem látható a kapcsolatos mellérendelések aránya sem, mivel ezek bírnak a legkisebb magyarázó erővel egy szöveg stílusáról, illetve ezek esetében nagyon gyakori, hogy a kapcsolatos viszony nincs külön kötőszóval jelölve a szövegben, így a mért eredmények is csak megközelítőleg érvényesek. Mindezek miatt tehát nem mondható el, hogy a diagramok a regények egészének belső szerkezetét tükröznék – csupán a tagmondatkapcsolatok egy részének eloszlását ábrázolják. Továbbá a könnyebb átláthatóság kedvéért a *4.1.* és *4.2. ábrá*n a választó és ellentétes mellérendeléseket, a magyarázó és következtető mellérendeléseket, valamint az idő-

 DOI: 10.31400/dh-hun.2021.4.2415

és helyhatározói alárendeléseket egy összevont típusba rendeztem, mivel az ezek által kidolgozott grammatikai jelentések viszonylag közel esnek egymáshoz. Könnyen létrehozható ugyanakkor egy részletesebb ábra is, ám csak az átláthatóság rovására – illetve ábrázolható az is, hogy csak az összevont kategóriákba tartozó típusok (pl. hely- és időhatározói alárendelések) hogyan aránylanak egymáshoz.

A relatív gyakoriságot mutató ábrák mellett több szempont miatt is érdemes ezeket a grafikonokat is szemügyre venni. Egyrészt ezáltal láthatóvá válik az eddig külön kezelt típusok gyakoriságának aránya a szövegeken belül. Mivel az egyes típusok (pl. a megengedő, vagy a magyarázó és következtető kapcsolatok) szinte mindig kevesebb számban fordulnak elő, mint mások (pl. vonatkozói alárendelések), az ábrák segítségével megválaszolható kérdés nem feltétlenül az, hogy melyik típus a leggyakoribb egy szövegben, hanem hogy milyen mértékűek a különbségek a típusok között. Másrészt így azoknak a regényeknek a sajátosságai is jobban kirajzolódhatnak, amelyek több korábbi kategóriában is kiugróan magas értéket mutattak (pl. Kertész Imre, *Kaddis*), vagy amelyek semelyik kategóriában nem mutattak kimagasló értéket, de belső szerveződésüket tekintve érdekes jellegzetességek figyelhetők meg (pl. Mány Iván, *A trafik*). Harmadrészt e diagramok segítségével a szövegek másképpen hasonlíthatók össze, mint eddig: mivel az ábrák egyszerre mutatják az összes kapcsolattípust, így mindegyiket figyelembe tudjuk venni az összehasonlítás során; a típusok arányai közti különbség pedig szinte azonnal érzékelhető.

A 12 regény belső szerveződését bemutató *4.1. ábrá*n például Mándy Iván *A trafik* című regényének jellegzetessége szembetűnő, hiszen esetében a hasonlatok egyértelmű dominanciája figyelhető meg. Azaz a korábbi (a relatív gyakoriságokat kiszámító) mérések alapján inkább az mondható el, hogy a regényre a tagmondatkapcsolatok hiánya és az egymástól elkülönült rövid benyomásszerű mondatok a jellemzők – a *4.1. ábra* mindezt azzal egészíti ki, hogy ha a szöveg mégis viszonyt dolgoz ki két tagmondat között, akkor azt leginkább hasonlítás útján teszi. Érdemes még megfigyelni, hogy a szövegben milyen kis számban fordulnak elő vonatkozói névmással bevezetett mellékmondatok a többi regényhez képest: gyakoribbnak tekinthetők ezeknél a szerkezeteknél a feltételes kapcsolatok, valamint a választó és ellentétes típusok összevont kategóriája is. Ezzel szemben az *Emlékiratok könyvé*ben vagy a *Kaddis*ban inkább a kapcsolatok arányos eloszlása figyelhető meg: más szövegekhez képes ezekben a regényekben például a megengedő és következtető/magyarázó kapcsolatok száma is jelentősnek tekinthető. Ez elsősorban az idő- és helybeli viszonyokat kidolgozó kapcsolatok rovására történik, legalábbis más regényekhez képest kisebb súllyal szerepelnek ezek a kapcsolattípusok az említett szövegekben. Az *Emlékiratok könyvé*ben például, habár nagyon fontosak az események idő és térbeli koordinátái és a különböző síkok viszonya, ez a viszony nem mindig kidolgozott nyelvileg, és inkább a történetszálak puszta egymás mellé rendeléséből jön létre. Gyakran csak több mondat után lehet megkülönböztetni például, hogy melyik fejezet hol és mikor játszódik, és csak egy-egy szereplő vagy történés kapcsán következtethetünk az elmesélt történet idejére/helyére – hiszen, amint az eddigiek során is láttuk, a kapcsolatok inkább egy konkrét jelenet összefüggéseit és nem a tágabb idő- és helybeli kontextust dolgozzák ki. Bodor Ádám regényeiben *(Sinistra körzet, Az érsek látogatása)* éppen ellenkezőleg, az ábrázolt kapcsolatoknak több mint a negyede dolgoz ki hely- és időbeli viszonyokat.



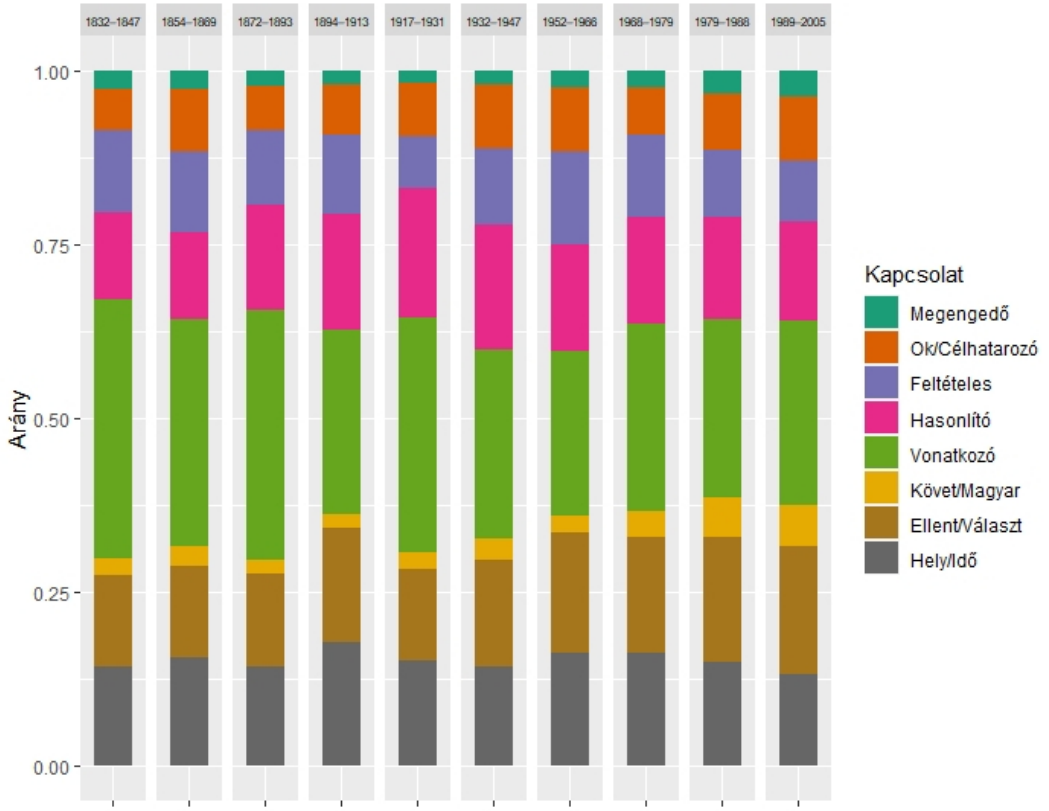
4.1. ábra. A kapcsolattípusok egymáshoz viszonyított aránya – 12 regény.

Szembetűnő az ábrán továbbá Mészöly Miklós *Saulus*ában és *Az atléta halálá*ban a magyarázó és következtető kapcsolatok elenyésző és az ok-/célhatározói kapcsolatok kis szerepe, ami a szövegek olyan szerveződésére hívja fel a figyelmet, amely a logikai kapcsolatok korlátozásán alapul.

Mindezek mellett lehetőség nyílik a 10 regényenkénti évcsoport belső szerveződését is hasonló módon vizualizálni. A *4.2. ábrá*n a korábban ismertetett történeti folyamatokat figyelhetjük meg az egyes oszlopok közti különbségeket vizsgálva (lásd *Eredmények 1.*). Egyedül a vonatkozói alárendelések arányának változásában nem rajzolódik ki egyértelműen a relatív gyakoriság alapján felfedezett tendencia, ami arra vezethető vissza, hogy míg a 19. század elejének hosszúmondatos regényei rendszerint másfajta kapcsolatokat is létesítenek a tagmondatok között, addig a 20. század első

 DOI: 10.31400/dh-hun.2021.4.2415

felében azok a regények, amelyekben gyakoriak a vonatkozói alárendelések, kisebb számban dolgoznak ki más típusokat (elsősorban Kosztolányi Dezső művei).



4.2. ábra. A kapcsolattípusok egymáshoz viszonyított aránya, évcsoportok.

Eredmények 4.

A regények belső szerveződését mutató diagram *(4.1. ábra)* segítségével könnyen felismerhetők az azonos mondatszerkezeteket előnyben részesítő művek, és szembetűnő az egy alkotóhoz (pl. Eötvös József) tartozó regények hasonlósága is. Ennek kapcsán merülhet fel a kérdés, hogy vajon a stílustörténeti tendenciák felrajzolásán és az egyes szövegek értelmezésén túl segíthet-e a mondatkapcsolatok kvantitatív vizsgálata az ismeretlen szerzőséget azonosító kutatások során is. Azaz milyen hatásfokon lehet elkülöníteni az egy szerzőhöz tartozó szövegeket a korpusz többi szerzőjétől a meglévő adatokra hagyatkozva. A szerzőségre irányuló kutatások több ízben bizonyították már, hogy a szövegek tematikus szintje alatt létezik egy úgynevezett „szerzői ujjlenyomat", amely a leggyakrabban – és ezért nem tudatosan – használt szavak, főként konkrét jelentés nélküli funkciószavak eloszlására vonatkozik, és amely eloszlás a szerzők különböző időszakban írt, különböző műfajú szövegeiben is hozzávetőlegesen állandó.[61] A leggyakrabban használt szavak relatív gyakorisága alapján tehát lehetséges

[61] Lásd például Harald Baayen, *Word Frequency Distributions* (Dordrecht: Kluwer, 2001), `https://doi.org/10.1007/978-94-010-0844-0`; John Burrows, „'Delta': A Measure of Stylistic

egy szerzőt (társadalmi-esztétikai kondícióitól függetlenül) más alkotóktól megkülönböztetni; kérdés, hogy ez igaz-e kizárólag a kötőszavakat és a tagmondatkapcsolatot létesítő vonatkozói névmásokat figyelembe véve is.

Az alábbi ábrák létrehozásakor tehát a kutatás során kiszámított értékeket vettem alapul. Ezek az értékek az egyes kapcsolattípusok relatív gyakoriságát vagy az egymáshoz viszonyított arányát mutatják: egy regényhez ezáltal több (12, illetve 8) érték is kapcsolódik, amelyek mentén egy többdimenziós térben/koordinátarendszerben lesznek a szövegek elhelyezhetők. Ez a pozíció a többdimenziós térben ugyanúgy határozható meg, mint ahogyan egy kétdimenziós koordinátarendszerben is jelölhető bármely pont, amelyhez megadjuk az x és az y tengelyre vonatkozó értékeket. Egy ilyen többdimenziós teret ugyanakkor sem ábrázolni, sem elképzelni nem tudunk – azonban létezik két lehetőség, hogy ebben a térben elfoglalt helyük alapján megadjuk a szövegek távolságát és közelségét, azaz (az értékek által leírt jellemzőkre vonatkozó) hasonlóságukat és különbözőségüket. Az egyik lehetőség, ha nem ábrázoljuk a pontokat ebben a térben, csupán a pozíciók közti távolságot számítjuk ki a két dimenzióban is érvényes metrikák szerint, majd a szövegeket az egymástól való közelségük alapján csoportosítjuk, végül ezeket a csoportokat egy dendrogram formájában jelenítjük meg. A kutatás során több távolságmetrika teljesítményét is összevetettem a regények ilyen csoportosításakor, amelyek közül a kétféle adatsor (a kapcsolattípusok relatív gyakorisága és egymáshoz viszonyított aránya – *5.1* és *5.2 ábra*) tekintetében a koszinusztávolság bizonyult a leghatékonyabbnak, míg az összevont adatsor alapján a Manhattan-távolság teljesített a legjobban (lásd a 65. és a 66. lábjegyzetet).[62] Az alábbiakban minden esetben a leghatékonyabb csoportosítást létrehozó módszereken alapuló eredményeket ismertetem, ami természetesen az eredmények kimazsolázását *(cherry-picking)* jelenti: ám ebben az esetben csupán a lehető legjobb csoportosítási eredményeket és azokon keresztül a kalasszifikáció korlátait kívánom bemutatni. A másik lehetőség a többdimenziós tér dimenzióinak a csökkentésére irányul, amelynek segítségével kétdimenziós felületen tudjuk ábrázolni az adatpontok közötti összefüggéseket. Jelen

Difference and a Guide to Likely Authorship," *Literary and Linguistic Computing* 17 (2002): 267–287, https://doi.org/10.1093/llc/17.3.267; Patrick Juola, „Authorship Attribution," *Foundations and Trends in Information Retrieval* 1, 3. sz. (2006): 233–334, http://doi.org/10.1561/1500000005. Stilometriai kutatásokat összefoglaló tanulmány: Maciej Eder, „Style-Markers in Authorship Attribution: A Cross-Language Study of the Authorial Fingerprint," *Studies in Polish Linguistic* 7, 1. sz. (2011), hozzáférés: 2021.04.28, http://www.ejournals.eu/sj/index.php/SiPL/article/view/2261/0.

[62] Mindazonáltal a szakirodalomban elfogadottnak tűnik az a megállapítás, miszerint a szerzősazonosítás esetében a koszinusztávolság a legmegbízhatóbb metrika, vö. Evert Stefan et. al., „Understanding and Explaining Delta Measures for Authorship Attribution," *Digital Scholarship in the Humanities* 32, 2. sz. (2017): 4–16, https://doi.org/10.1093/llc/fqx023. A Manhattan-távolság a kétdimenziós térben két ponthoz egy harmadikat jelöl, amellyel együtt egy olyan derékszögű háromszöget alkotnak, amelynek átfogója a két pont közötti egyenes; a távolság mértékét pedig a befogók hosszának összege határozza meg. A metrika elnevezését New York jellegzetes építészeti elrendezéséről kapta, hiszen az egymásra merőleges utcákon általában derékszöget bezáró útvonalon lehet eljutni egyik pontból a másikba. A koszinusztávolság ezzel szemben azt számítja ki, hogy ha két pontot egy koordinátarendszerben ábrázolunk, akkor mekkora a pontokat és az origót összekötő egyenesek által bezárt szög koszinusza.

 DOI: 10.31400/dh-hun.2021.4.2415

esetben ehhez az úgynevezett főkomponens-analízis eljárását alkalmaztam, amely az adatpontok varianciáját megtartva redukálja a dimenziók számát.[63]

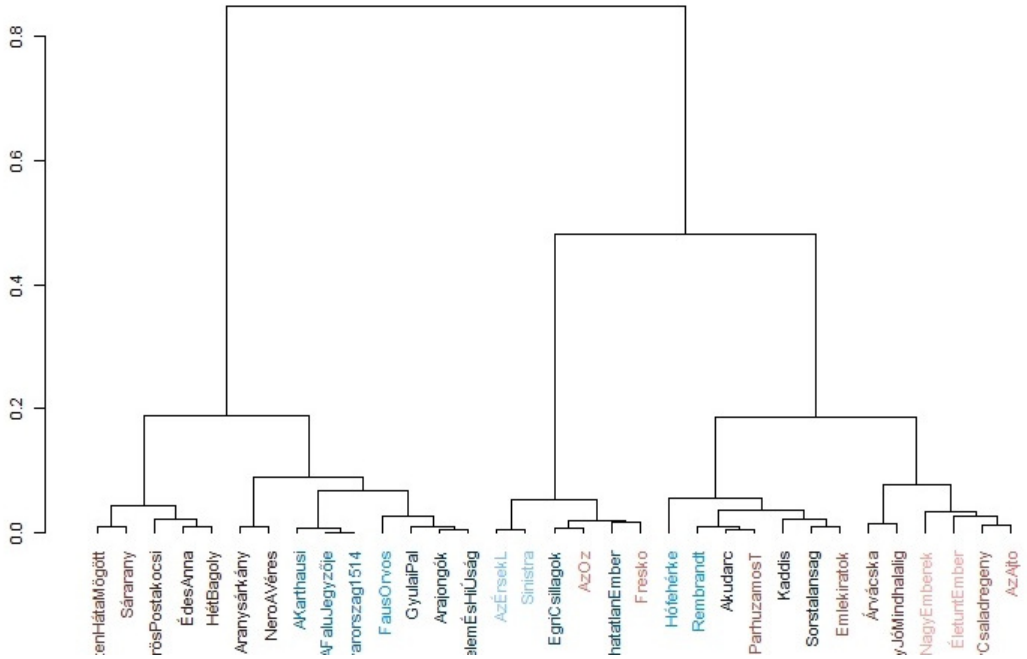
Az *5.1.* és *5.2. ábrá*n a koszinusztávolság szerint és a Ward-módszer[64] alapján csoportosított 33 regény látható a korpuszból, amelyek közül legalább kettőnek azonos szerzője van (az azonos szerzőséget színkód jelöli), hogy ezáltal elemezhető legyen, hogy milyen hatékonysággal kerülnek ugyanahhoz az alkotóhoz tartozó szövegek egymás mellé a dendrogram ágrajzán. Az *5.1. ábra* a kapcsolatok belső arányát, a *5.2. ábra* a kapcsolatok relatív gyakoriságát vette figyelembe (a kapcsolatos mellérendelést kivéve) mint a szövegek helyzetét meghatározó értékeket. Mindkettőn helyesen került egymás mellé Bodor Ádám és Gárdonyi Géza két-két, valamint Eötvös József és Kemény Zsigmond három-három regénye, ahogyan Krúdy Gyula két és Kosztolányi Dezső három műve is közel, egy fürtön helyezkedik el az ábrákon. Ugyanakkor mindkét csoportosításban kisebb-nagyobb hibák is mutatkoznak: Szabó Magda és Bródy Sándor egyes regényei meglehetősen távolra kerültek egymástól, Móricz Zsigmond négy regénye pedig két csoportra oszlik *(Az Isten háta mögött* és *Sárarany,* illetve *Légy jó mindhalálig* és *Árvácska).* A regények szerzőség alapján történő elkülönítése így csupán közepesen sikeresnek mondható a kapcsolattípusok gyakoriságát vizsgálva.[65] Az eredmények kis mértékben javíthatók,[66] ha mind a két szempontot (relatív gyakoriság és belső arány) egyszerre érvényesítjük, ám az azonos szerzőségű művek elkülönítése így sem tökéletes *(5.3. ábra).*

[63] A főkomponens-analízis eljárása az alábbi módon képzelhető el: a módszer első lépésben az értékek átlaga által kijelölt pontot teszi meg a koordinátarendszer középpontjául, majd egy olyan egyenest húz ezen az origón keresztül, amelyhez a lehető legközelebb esnek az egyes adatpontok. Ezt követően erre az egyenesre további merőlegeseket állít a változók számának megfelelően. A végső lépésben ezek közül a merőlegesek közül azt a kettőt teszi meg az eredményül megadott kétdimenziós ábra két egyenesének (PC1 és PC2), amelyekre vetítve az adatpontok origótól való távolságának négyzetösszege a legnagyobb. A létrehozott diagram mutatja, hogy PC1 és PC2 hány százalékban játszik szerepet az adatpontok közötti különbség meghatározásában: minél nagyobb ez a szám, annál jelentősebb az egységnyi távolság e tengely mentén. Vö. Zakaria Jaadi, *A Step-by-Step Explanation of Principal Component Analysis (PCA),* hozzáférés: 2021.04.28, `https://builtin.com/data-science/step-step-explanation-principal-component-analysis`.

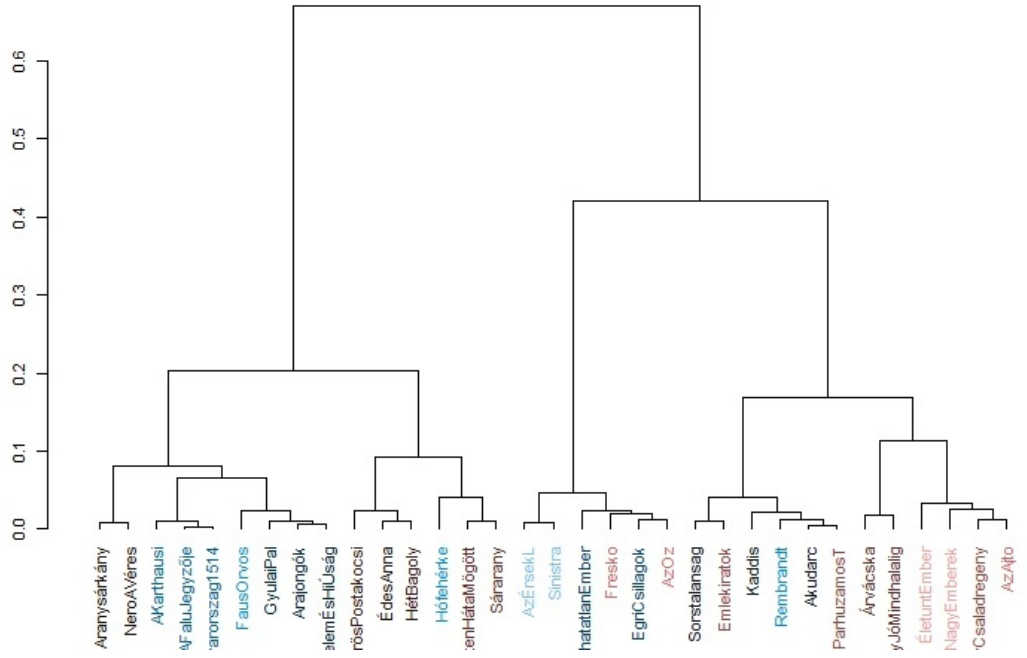
[64] A Ward-módszer arra törekszik, hogy az egy csoportba rendezett elemek varianciája, azaz a csoportátlaghoz képesti eltérések négyzetösszege a lehető legkevesebb legyen.

[65] A kapcsolattípusok egymáshoz viszonyított aránya alapján létrehozott csoportok a 12 szerző helyes elrendezéséhez képest 0,43-as értéket mutat a két csoportosítást összehasonlító korrigált Rand index (ARI) szerint (Manhattan-távolsággal: ARI = 0,43); a relatív gyakoriságok alapján létrehozott csoportosítás esetében ARI = 0,42 (Manhattan-távolsággal: ARI = 0,37). Ez a mérőszám minél inkább 0-hoz közelít, annál inkább egyezik meg az eredmény egy pusztán véletlenszerű csoportosítás eredményével; míg minél inkább 1-hez közelít, annál inkább egyezik meg a szerzőség szerinti helyes csoportosítással. (Ugyanakkor a statisztikailag kevésbé megbízható, mert a véletlenszerű eloszlás hatását nem, csupán az egy fürtre kerülő elempárokat vizsgáló nem-korrigált Rand index alapján mindkét csoportosítás magas, 0,9 feletti értéket mutat a szerzők mentén elrendezett csoportokhoz képest.)

[66] Manhattan-távolsággal: ARI = 0,54. Koszinusztávolsággal: ARI = 0,45.

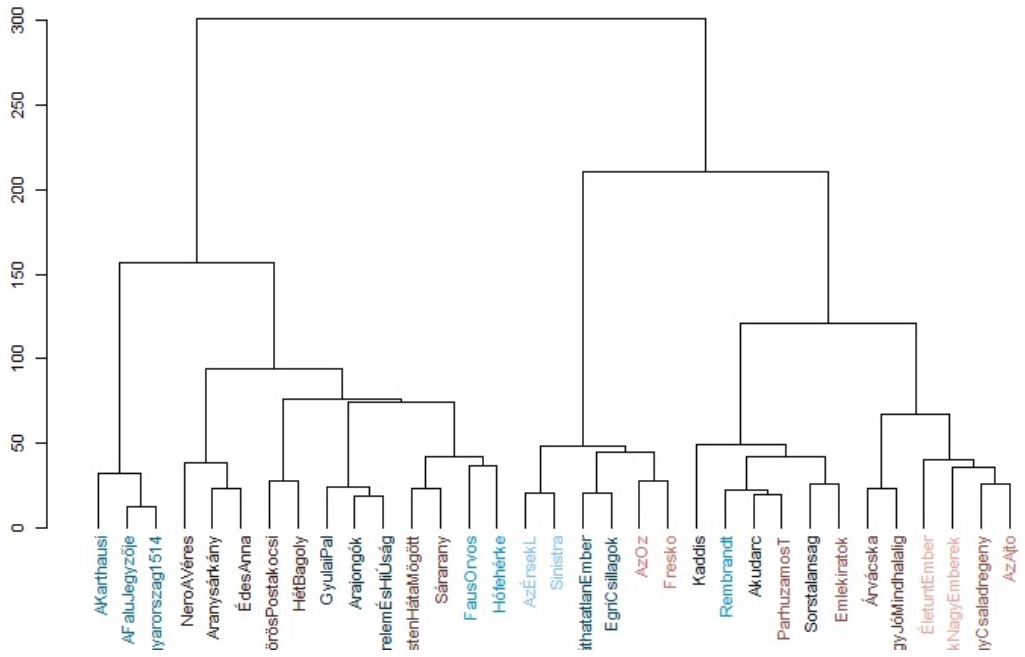


5.1. ábra. Koszinusztávolság. 33 regény csoportosítása a kapcsolattípusok egymáshoz viszonyított aránya alapján.



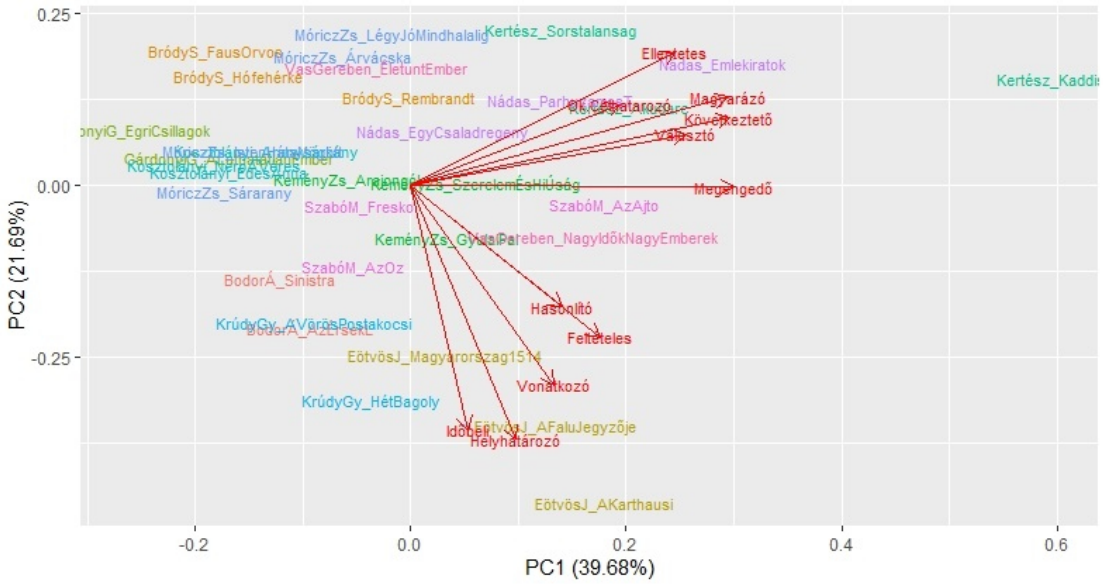
5.2. ábra. Koszinusztávolság. 33 regény csoportosítása a kapcsolattípusok relatív gyakorisága alapján.

 DOI: 10.31400/dh-hun.2021.4.2415



5.3. ábra. Manhattan-távolság. 33 regény csoportosítása a kapcsolattípusok egymáshoz viszonyított aránya és relatív gyakorisága alapján.

A relatív gyakoriságokra vonatkozó főkomponens-analízis eredménye látható az *5.4. ábrán*. Itt a szövegek közti különbség és hasonlóság a térben elfoglalt pozíció, valamint a tengelyekhez (PC1 és PC2) rendelt százalék szerint alakul – ez a szám mutatja, hogy milyen mértékben játszik szerepet a tengelyeken megfigyelhető távolság az adatpontok elkülönítésében. Tehát a kiválasztott szempont szerint hasonló szövegek egymáshoz közel helyezkednek el; míg az azonos szerzőségű művek azonos színnel szerepelnek az ábrán. Ezenkívül mint vektorok jelöltek az egyes kapcsolattípusok is az ábrán, amelyeknek iránya mutatja, hogy ezek a típusok hogyan befolyásolják a szövegek elhelyezkedését.



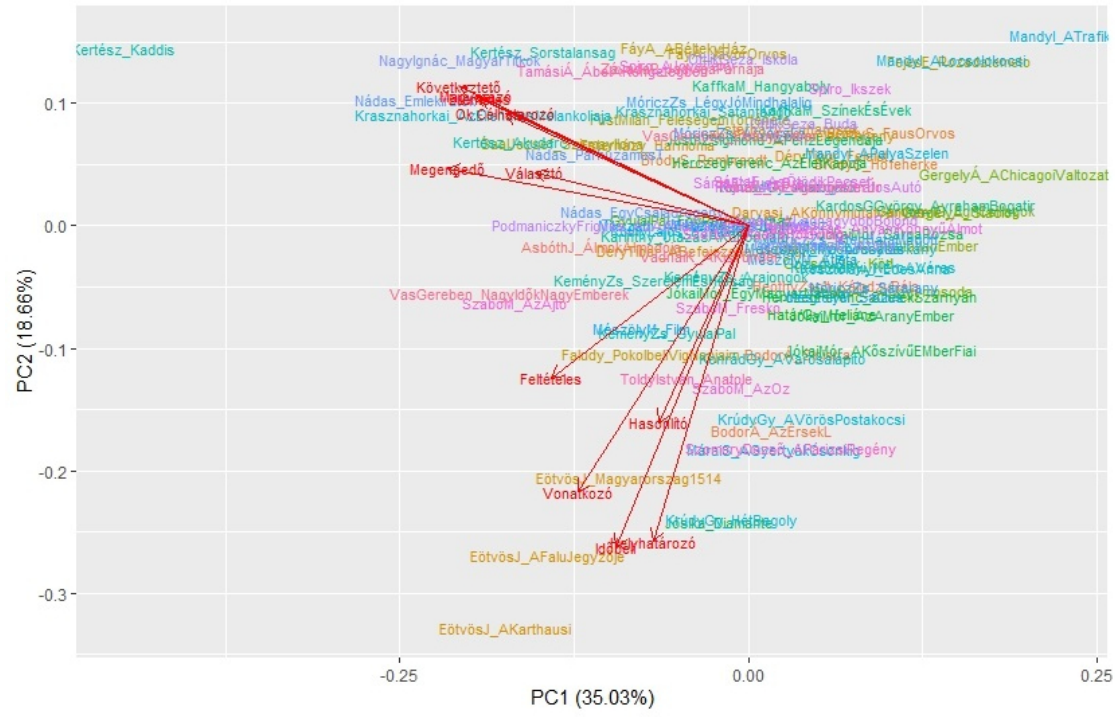
5.4. ábra. Főkomponens-analízis a kapcsolattípusok relatív gyakorisága szerint, 33 regény.

Az azonos szerzőségű regények elkülönítése ezúttal is egyszerre nevezhető sikeresnek és sikertelennek: a legtöbb esetben hasonló helyzetűek, de nem különíthetők el egyértelműen más szövegektől vagy szövegcsoportoktól. Mindez magukra a vizsgált jellemzőkre vezethető vissza. A kötőszavak és a tagmondatkapcsolatot létesítő vonatkozó névmások használata ugyanis a szöveg *félig tudatos* szintjére vonatkozik; hiszen míg más funkciószavak (pl. névelők) eloszlása teljesen független az alkotói szándéktól – gyakoriságuk ezért is képes a szerző stílusának „ujjlenyomatát" hordozni –, addig a tagmondatkapcsolatok kidolgozása nem teljesen független az alkotó esztétikai és tartalmi megfontolásaitól. A tudatos és a nem tudatos szintek között fejtik ki a hatásukat, amennyiben használatuk (a tartalmas jelentéssel bíró szavakhoz képest) nem kontrollált az alkotói folyamat során, ugyanakkor a tudatos írói döntések befolyásolhatják gyakoriságukat, illetve az általuk kidolgozott kapcsolatok gyakoriságát. Ezért tekinthető a kísérlet közepesen sikeresnek, azaz emiatt lehet több azonos szerzőhöz tartozó szöveget egy csoportba rendezni a vázolt módszerekkel, és ugyanígy emiatt nem felismerhető a más mondatszerkezeteket előnyben részesítő regények közös szerzősége. A kutatás eredményei azt mutatják, hogy nem is a szerzőazonosítás kérdésének megválaszolására alkalmas a tagmondatkapcsolatok gyakoriságának elemzése, hanem a különböző hagyományokhoz tartozó prózapoétikák különíthetők el általa.

Az *5.4. ábrá*n világosan látható, hogy a relatív gyakoriság szerint a kapcsolattípusok két csoportja egymástól élesen elválik, azaz két irány különíthető el a piros vektorokat figyelembe véve: egyfelől a mellérendelő (választó, ellentétes, következtető, magyarázó és megengedő) kapcsolatok és az ok-/célhatározói alárendelések, másfelől a hasonlító, feltételes, vonatkozói, valamint hely- és időhatározói alárendelések fejtik ki hatásukat megegyező irányban. Mind a 100 regényt vizsgálva hasonló eredményre juthatunk *(5.5. ábra)*. Ami alapján arra a következtethetünk, hogy a mondatszerkezetek kvantitatív vizsgálatán keresztül a kötőszó- és vonatkozónévmás-használat szem-

 DOI: 10.31400/dh-hun.2021.4.2415

pontjából három prózastílus határozható meg a magyar regényirodalom kanonikus szövegeit tekintve. Egyrészt elkülönül egy, a tagmondatok közötti, elsősorban logikai viszonyt kidolgozó stílus – abban az értelemben, ahogyan a formális logika is főként a mellérendelő kapcsolatokat ruházza fel logikai értékkel; illetve az ok-/célhatározói kapcsolatok is ilyen összefüggések kidolgozásában játszanak szerepet, nagyon hasonló módon a következtető és magyarázó mellérendelésekhez, amelyektől való elhatárolása esetenként mesterségesnek is tűnhet. Másrészt egy csoportba rendeződnek az inkább a mondatok szereplőit vagy az ábrázolt jelenetek körülményeit részletesebben kidolgozó stílushoz tartozó regények; ezekben a vonatkozói, feltételes és hasonlító mellékmondatok száma tekinthető gyakorinak. A harmadik stílust képviselik azok a szövegek, amelyekre egyik típus sem jellemző, hanem inkább rövidmondatos,[67] tagmondatkapcsolatokat ritkábban kidolgozó prózanyelvet érvényesítenek. Természetesen ezek a stílusok inkább mint ideáltípusok írhatók le, és ritkán valósulnak meg tiszta formában. Mindazonáltal – ahogyan az *5.5. ábra* is mutatja – az első hagyományra Kertész Imre *Kaddis a meg nem született gyermekért,* a második hagyományra Eötvös József *A karthauzi,* míg a harmadikra Mándy Iván *A trafik* című regénye tekinthető a korpuszból a legjellemzőbbnek: a többi szöveg az ezek által határolt térben oszlik el.



5.5. ábra. Főkomponens-analízis a kapcsolattípusok relatív gyakorisága szerint, 100 regény

[67] A mondathosszúságokhoz lásd Szemes, *Mondathosszúság és irodalomtörténet* idézett tanulmányát és annak függelékét.

Összegzés

A tagmondatkapcsolatok eloszlásának vizsgálata nemcsak a szövegek nyelvi szerveződésének jobb megismeréséhez vezethet, hanem diagrammatikus, topológiai-logikai karakterüket is képes megvilágítani. Automatikus azonosításuk és gyakoriságuk mérése így a stilisztikai leíráson túl egyaránt lehetőséget teremt irodalomtörténeti tendenciák felvázolására, az egyes regények értelmezésére és különböző prózapoétikai hagyományok elkülönítésére. A dolgozat három hagyományt különböztetett meg egymástól: egy inkább alárendeléseket alkalmazó, a mondatok egy-egy szereplőjét részletesebben kidolgozó, ezáltal a jelentést jobban stabilizáló stílust; egy, a tagmondatok közötti logikai kapcsolatot létesítő stílust, amely állandóan új értelmezések felé nyitja meg a mondatot; valamint egy rövidmondatos stílust, amely inkább egyszerű mondatokból, valamint nem kidolgozott tagmondatkapcsolatokból (illetve a kidolgozott kapcsolatokat tekintve elsősorban hasonlatokból) épül fel. Az ilyen következtetésekhez járulnak hozzá az eredmények különböző vizualizációi, amelyek különböző összefüggéseket mutató elrendezésekben képesek ábrázolni a mérési eredményeket. A továbbiakban érdemesnek tűnik az egyes kapcsolatok részletesebb, alkategóriákra bontott elemzését is elvégezni, valamint a mondatok között kidolgozott kapcsolatok vizsgálatával kiegészíteni az eddigi kutatást.

Elaborated Relations. On the usability of automatic identification of clause relationships in stylistics and literary history

The paper aims to develop a method that allows an automatic identification of the different types of complex and compound sentences in the Hungarian written language through the analysis of conjunction words, relative pronouns and their positions in the sentence. This method opens up new perspectives in the stylometric research: on the one hand, conjunctions and pronouns as function words provide a large amount of data for statistical analyses, and on the other hand they also carry meaning – about the relationships between the clauses (eg. opposition, conditionality). By examining the relative frequency of each type, it becomes possible to identify the most characteristic relationship of the clauses in a given text or corpora. Thus, the style and poetics of the novels can be grasped at the scale of the sentence, while also revealing the topological-logical structure of the texts, which is not usually reflected during the reading process. The paper analyzes the frequency of each type of relation in a corpus of 100 Hungarian novels. This provides an opportunity to (1) register stylistic tendencies of the Hungarian novel in the period between 1832–2005); (2) examine which texts are the most typical of each type compared to other texts; (3) to visualize the 'internal structure' of a text, based on the relative proportions of types to each other; and (4) to distinguish different stylistic traditions.

Keywords:
clause relation, conjunctions, stylometry, literary history, Hungarian novel

 DOI: 10.31400/dh-hun.2021.4.2415

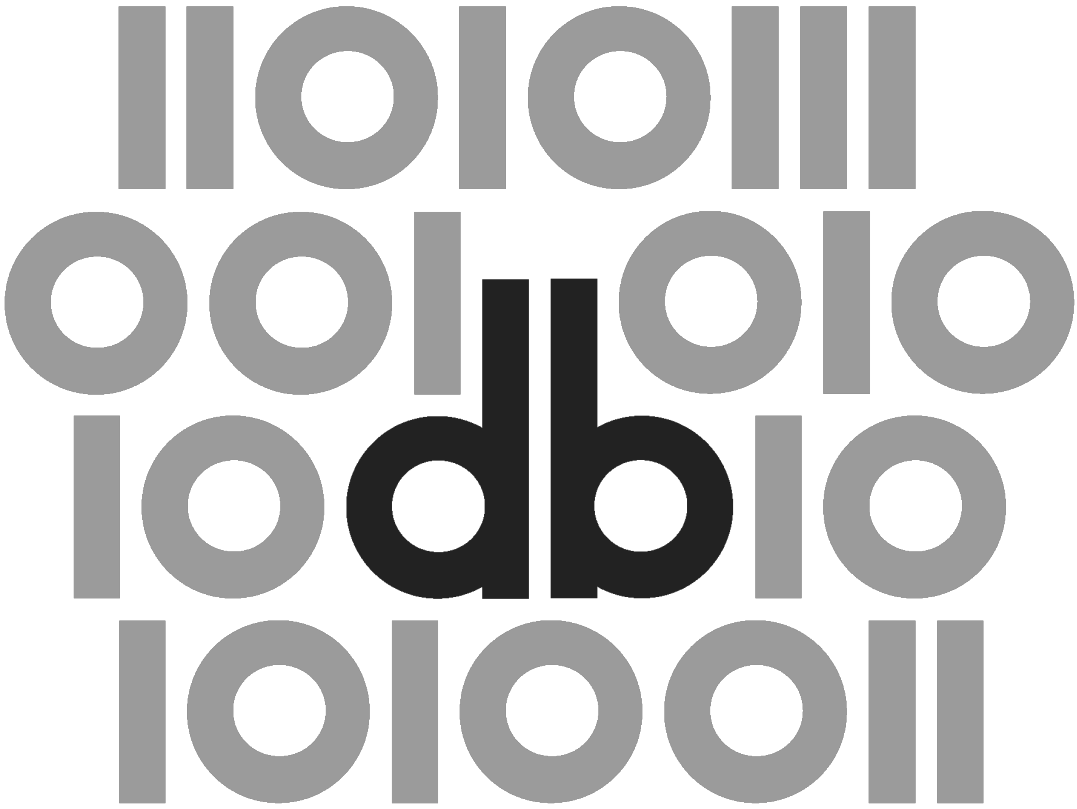
Függelék

A kutatás során használt kódok, valamint az egyes regényekre vonatkozó eredmények elérhetők:

`https://github.com/SzemesBotond/tagmondatkapcsolat`

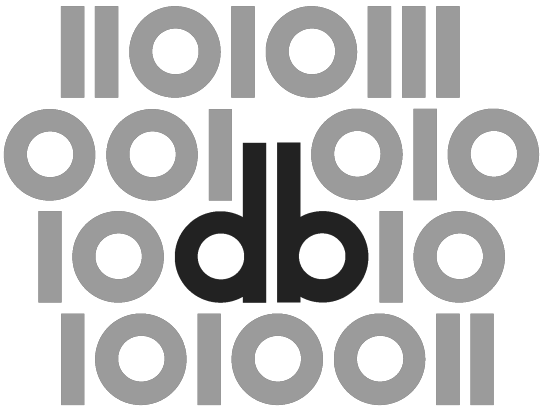
A tanulmány függelékei és ábrái a cikk mellékleteként is elérhetők annak adatlapjáról:

`https://doi.org/10.31400/dh-hun.2021.4.2415`



Digitális Bölcsészet
2021., negyedik szám

<DIGITÁLIS BÖLCSÉSZET>



4 (2021)

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.2021.4

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.

Ez a mű a Creative Commons *Nevezd meg! – Ne add el! – Így add tovább! 2.5 Magyarország Licenc* (`http://creativecommons.org/licenses/by-nc-sa/2.5/hu/`) feltételeinek megfelelően felhasználható.

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

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

<TANULMÁNYOK>

Horváth Péter 0000-0002-3517-5623

ELTE BTK TI Digitális Bölcsészet Tanszék

horvath.peeteer@gmail.com

Két eljárás magyar nyelvű versek metrumának gépi felismertetéséhez*

A tanulmány a magyar időmértékes és ütemhangsúlyos versmetrum gépi felismertetésének egy-egy eljárását mutatja be. A magyar időmértékes metrum elemzését végző algoritmus a verset besorolja a daktilikus, anapesztikus, jambikus vagy trochaikus kategóriák valamelyikébe, és megadja a megvalósult versritmus szabályosságának mértékét. A magyar ütemhangsúlyos metrum elemzését végző algoritmus pedig megadja az ütemek szótagszámát. A tanulmány a két algoritmus alkalmazási lehetőségeit egy tesztkorpusz segítségével mutatja be, amely tizenegy, a 19. század közepén és második felében, valamint a 20. század első felében alkotó szerző verseit tartalmazza.

Kulcsszavak:
időmértékes metrum, ütemhangsúlyos metrum, gépi elemzés, korpuszvizsgálat, ELTE Verskorpusz

1. Bevezetés

A tanulmány a magyar időmértékes és ütemhangsúlyos metrum gépi felismertetésének egy-egy eljárását mutatja be. Az időmértékes metrum gépi elemzését elvégző algoritmus a négy nagy időmértékes metrumkategória, a daktilikus, az anapesztikus, a jambikus és a trochaikus felismerésére, valamint a megvalósuló versritmus szabályosságának a mérésére képes. Természetesen a négy metrumkategórián belül további alkategóriák is vannak, mint amilyen például a hexameter, illetve olyan kötött képletű metrumok is léteznek, amelyek nem sorolhatók be ezekbe az általános kategóriákba. Ezen specifikusabb metrumok automatikus felismerésére az algoritmus nem képes. Az ütemhangsúlyos verselés felismerését végző algoritmus is az egyszerűbb, aaaa vagy abab szerkezetű versek felismerésére képes, azaz azon versek esetében állapít meg ütemhangsúlyos metrumot, ahol minden sorra ugyanaz az ütemosztás jellemző, vagy pedig kétfajta, a versszakok páratlan és páros soraira jellemző ütemosztás váltogatja egymást. Megjegyzendő ugyanakkor, hogy az ütemhangsúlyos verseknek a legnagyobb részét lefedi ez a két típus.

* A tanulmány az NKFIH K_21 137659 számú, „A személyjelölés konstrukcióinak korpuszalapú, kognitív poétikai vizsgálata” pályázat, valamint a Digitális Örökség Nemzeti Laboratórium keretében készült.

A 2. részben röviden bemutatok több hazai példát a vershangzás gépi elemzésére, valamint a metrum gépi felismertetésének néhány olyan nemzetközi példáját is ismertetem, amelyek bizonyos elgondolásait felhasználtam a két algoritmus megtervezése során. A 3. részben ismertetem az időmértékes metrum felismerésére, illetve a megvalósuló versritmus szabályosságának a mérésére létrehozott algoritmus főbb lépéseit. Az algoritmust Python nyelvben implementáltam, és az ELTE Verskorpusz részét képező tizenegy, a 19. század közepén és második felében, valamint a 20. század első felében alkotó költő versein futtattam le. Az így nyert, időmértékes metrumra vonatkozó kvantitatív adatokat a 4. rész mutatja be. Az 5. rész az ütemhangsúlyos verselés gépi felismerését elvégző algoritmus főbb lépéseit veszi sorra. Az úgyszintén Python nyelvben implementált algoritmusnak az említett tizenegy szerző versein való lefuttatásával kapott előfordulási adatokat a 6. részben ismertetem. Végezetül a 7. részben bemutatom azt is, hogy az időmértékes és az ütemhangsúlyos verselést elemző két algoritmus kimenetének együttes figyelembevételével hogyan vizsgálgató a szimultán versmetrum.

2. A vershangzás és a metrum gépi elemzésének néhány példája

Magyar nyelvű versek hangzástulajdonságainak az automatizált, valamilyen számítógépes programmal végzett elemzésére viszonylag korai példákat is találhatunk. Voigt Vilmos 1972-es tanulmánya mutatja be az első kísérletet magyar nyelvű versek számítógépes ritmuselemzésére. A létrehozott programmal három Szabó Lőrinc-szonettnek ismertették fel a megvalósuló időmértékes ritmusát, azaz a program meghatározta az egyes szótagok hosszúságát egy négyfokozatú skálán, és összegezte, illetve átlagolta az egyes sorokra és szótaghelyekre kapott értékeket.[1] Saját korát megelőzte Jékel Pál és Papp Ferenc 1974-ben kiadott műve, amely Ady összes versének az algoritmikus úton előállított fonémastatisztikai adatait tartalmazza, valamint az adatokat elemző bevezető tanulmányt.[2] Ugyancsak a korai példák között tartható számon Jékel Pál és Szuromi Lajos 1980-ban megjelent műve, amely Petőfi 300 verse esetében tartalmazza a szótagok részben gépi úton előállított különböző típusú nyomatékértékeit, valamint a nyomatékértékek automatikusan előállított összegzését és különféle statisztikáit.[3] Szorosan kapcsolódik a tanulmány témájához Lesi Zoltán kutatása is, aki tudomásom szerint elsőként hozott létre olyan többfunkciós programot, amely magyar nyelvű

[1] Voigt Vilmos, „Számítógépes ritmuselemzési kísérlet," *Irodalomtörténeti Közlemények* 76, 2. sz. (1972): 203–211.

[2] Jékel Pál és Papp Ferenc, *Ady Endre összes költői műveinek fonémastatisztikája* (Budapest: Akadémiai Kiadó, 1974). Jékel és Papp kutatása nemcsak abban volt újszerű, hogy a korhoz képest viszonylag nagy mennyiségű szöveget dolgoztak fel automatikusan, hanem abban is, hogy a szöveghasonlóság mérésére egy olyan matematikai modellt alkalmaztak, amely csak harminc évvel később vált általánosan elterjedtté a számítógépes nyelvészetben. Ez a szövegek hasonlóságának vektortérben történő, koszinusz távolságon alapuló mérése.

[3] Jékel Pál és Szuromi Lajos, *Petőfi metrumai* (Debrecen: Kossuth Lajos Tudományegyetem, 1980). A kutatás módszertani problémáira hívja fel a figyelmet: Vadai István, „Számítógép a verstan szolgálatában: Módszertani megjegyzések egy számítógépes ritmuselemzési kísérlet kapcsán," *Irodalomtörténeti Közlemények* 88, 1. sz. (1984): 74–86.

 DOI: 10.31400/dh-hun.2021.4.2361

versek rímképletének, alliterációinak és metrumának a gépi elemzésére is alkalmas.[4] Érdemes utalni Labádi Gergely Berzsenyi verseiről írott tanulmányára is, amelyben automatizált módszerekkel végzett, a versek szókészletére és fonémajellemzőire (szóhosszúság, magánhangzók és mássalhangzók eloszlása) vonatkozó vizsgálatok eredményei szerepelnek.[5] Bár nem tartalmaz automatikusan előállított információkat, de megemlítendő még a Horváth Iván főszerkesztésével létrehozott, *A régi magyar vers leltára a kezdetektől 1600-ig (Répertoire de la poésie hongroise ancienne)* nevű adatbázis, amelyben többek között a versek rímképletére és metrumára vonatkozó adatok is szerepelnek kereshető formában.[6]

A versritmus és versmetrum gépi elemzésére számos nemzetközi példát is találhatunk. Például a 16. és 17. századi spanyol szonettekből álló *Corpus de Sonetos del Siglo de Oro (Corpus of Spanish Golden-Age Sonnets)* minden verssor esetében tartalmazza a hangsúlyos és hangsúlytalan szótagok gépileg létrehozott – és a kétséges esetekben manuálisan ellenőrzött – annotációit.[7] A *Cseh Verskorpusz (Korpus českého verše)* közel 80000 verset tartalmaz számos automatikusan létrehozott annotációs réteggel.[8] Többek között a spanyol korpuszhoz hasonlóan szerepelnek benne a hangsúlyos és hangsúlytalan szótagok jelölései, de ezek mellett tartalmazza a hangsúlyos és hangsúlytalan szótagok annotációiból absztrahált metrumkategóriákat is. Az itt bemutatandó két algoritmusnak is az a célja, hogy ne csupán a versritmust ismerje fel, hanem a ritmusból, azaz a hosszú és rövid szótagok, illetve a hangsúlyos és hangsúlytalan szótagok sorából valamilyen metrumra is képes legyen absztrahálni.

A metrum gépi elemzésének tehát jellemzően két fő lépése van. Az első a ritmus elemzése, azaz az adott metrikai rendszertől függően a hangsúlyos és hangsúlytalan vagy a hosszú és rövid szótagok megállapítása a verssorokban. A második lépés pedig a megállapított ritmusból a metrumra történő absztrahálás. Az angol (és más indoeurópai nyelvek, például a már említett spanyol és cseh) esetében a versmetrum

[4] Lesi Zoltán, „Automatikus verselemzés tanuló algoritmusok alkalmazásával," in Alexin Zoltán és Csendes Dóra, szerk., *IV. Magyar Számítógépes Nyelvészeti Konferencia,* 402–407 (Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2006); Lesi Zoltán, „Automatikus formai verselemzés," *Alkalmazott Nyelvtudomány* 8, 1–2. sz. (2008): 197–208. A nyilvánosan nem elérhető programot bemutató két tanulmányban sajnos csak néhány rövid megjegyzés és egy képernyőfotó utal a metrum felismertetésére. Uo., 205.

[5] Labádi Gergely, „Az olvasó gép: Berzsenyi Dániel versei távolról," *Digitális Bölcsészet* 1. sz. (2018): 17–34, https://doi.org/10.31400/dh-hun.2018.1.126.

[6] *Répertoire de la poésie hongroise ancienne,* hozzáférés: 2021.04.08, https://f-book.com/rpha.

[7] Borja Navarro-Colorado, „A Computational Linguistic Approach to Spanish Golden Age Sonnets: Metrical and Semantic Aspects," in Anna Feldman, Anna Kazantseva, Stan Szpakowicz et al., eds., *Proceedings of the Fourth Workshop on Computational Linguistics for Literature,* 105–113 (Denver: Association for Computational Linguistics, 2015), https://doi.org/10.3115/v1/W15-0712; Borja Navarro-Colorado, Marí Ribes Lafoz and Noelia Sánchez, „Metrical Annotation of a Large Corpus of Spanish Sonnets: Representation, Scansion and Evaluation," in Nicoletta Calzolari, Khalid Choukri, Thierry Declerck et al., eds. *Proceedings of the 10th Edition of the Language Resources and Evaluation Conference (LREC 2016),* 4360–4364 (Portorož: ELRA, 2016).

[8] Petr Plecháč and Robert Kolár, „The Corpus of Czech Verse," *Studia Metrica et Poetica* 2, 1. sz. (2015): 107–118, https://doi.org/10.12697/smp.2015.2.1.05; Robert Ibrahim and Petr Plecháč, „Toward Automatic Analysis of Czech Verse," in Barry P Scherr, James Baily and Evgeny V. Kazartsev, eds., *Formal Methods in Poetics: A Collection of Scholarly Works Dedicated to the Memory of Professor M.A. Krasnoperova,* 295–305 (Lüdenscheid: RAM, 2011).

nem a görög, a latin és a magyar időmértékes metrumra is jellemző hosszú és rövid szótagok, hanem a hangsúlyos és hangsúlytalan szótagok váltakozására épül. Például az angol nyelvű versekben a jambus egy hangsúlytalan és egy hangsúlyos szótagból áll. Mivel az angolban nincs általános szabály arra vonatkozólag, hogy egy szó melyik szótagja hangsúlyos, gyakori megoldás, hogy az angol nyelvű versek metrumát elemző algoritmusokba a szavak szótagjainak a hangsúlyértékét jelölő lexikális adatbázist építenek be. Ezzel a megoldással él például a *Scandroid* nevű program, amely a verseket a jambikus és anapesztikus kategóriák valamelyikébe sorolja be.[9] A *Scandroid*nál több metrumtípus felismerésére is képes *AnalysePoem* és *ZeuScansion* eszközök is egy programba beépített adatbázist használnak a hangsúlyos és hangsúlytalan szótagok megállapításához.[10] Az előbbi esetben ez egy felhasználó által bővíthető lexikai adatbázis, az utóbbiban pedig két kiejtési szótár, amelyek segítségével a program a szótárakban nem szereplő szavak hangsúlyviszonyaira is kínál elemzést, oly módon, hogy az elemzendő szó írásmódjára legjobban hasonlító szótári szó elemzését választja ki.

Szemben az angollal, magyar nyelvű versek esetében a rövid és hosszú szótagok váltakozására épülő időmértékes metrum elemzéséhez nem szükséges beépíteni az algoritmusba hangzásjellemzőkre vonatkozó lexikai adatbázist, hiszen néhány általános, könnyen algoritmizálható szabály alapján meghatározható, hogy mely szótagok tekintendők rövidnek, és melyek hosszúnak (ezeket a szabályokat lásd a következő részben). A magyar nyelvű versek ütemhangsúlyos metrumának a gépi felismertetésekor is szerencséje van a kutatónak, hiszen, szemben az angollal, a magyar kötött hangsúlyú nyelv, azaz a szavak hangsúlyos szótagjai mindig az első szótagok. Ez tehát azt jelenti, hogy az ütemhangsúlyos versek ütemeinek az első, hangsúlyos szótagjainak a megállapításához sem szükséges szóhangsúlyokat jelölő lexikai adatbázist beépíteni a programba.

A metrum szabályalapú elemzésének második fő lépése általában egy számos allépésből álló absztrakciós eljárás, amelynek során a program a megállapított versritmus alapján meghatározza a metrumot (pl. hexameter, jambikus, felező nyolcas). Ez az absztrakciós eljárás sok esetben valamilyen pontozásos módszert használ, amelynek során a legtöbb pontot kapó metrumalternatívát választja ki a program. Például az említett *AnalysePoem* nevű program hat metrumot képes megállapítani. A program a domináns metrum megállapításához számos tesztet végez el egymást követően, amelynek kimeneteként megbízhatósági értéket *(confidence number)* rendel a megállapított metrumhoz. Minél nagyobb a megbízhatósági érték, annál szabályosabban valósítja meg a versritmus a metrumot, egy bizonyos megbízhatósági érték alatt pedig a vers nem tekinthető az adott metrumhoz tartozónak.[11] Ehhez hasonló pontozásos módszert alkalmaztam én is a magyar időmértékes metrum felismerésére létrehozott

[9] Charles O. Hartman, *The Scandroid. Version 1.1.* [User guide] (2005), hozzáférés: 2020.01.19, `http://charlesohartman.com/verse/scandroid/ScandroidManual.pdf`.

[10] Marc R. Plamondon, „Virtual Verse Analysis: Analysing Patterns in Poetry," *Literary and Linguistic Computing* 21, 1. sz. (2006): 127–141, `https://doi.org/10.1093/llc/fql011`; Manex Agirrezabal, Aitzol Astigarraga, Bertol Arrieta et al., „*ZeuScansion:* A Tool for Scansion of English Poetry," *Journal of Language Modelling* 4, 1. sz. (2016): 3–28, `https://doi.org/10.15398/jlm.v4i1.102`.

[11] Plamondon, „Virtual Verse Analysis: Analysing Patterns in Poetry".

 DOI: 10.31400/dh-hun.2021.4.2361

algoritmusban. Az úgyszintén említett *ZeuScansion* nevű eszköz a versre általánosan jellemző metrumot oly módon állapítja meg, hogy kiszámítja az egyes szótaghelyekre eső szótagok átlagos hangsúlyértékét, majd az ebből absztrahált ritmusképletnek több alternatív, verslábakra tagolt elemzését is megadja, amelyekből egy pontozási rendszer alapján választja ki a legmegfelelőbbet (az algoritmus azonos szótagszámú sorokon működtethető).[12] A hangsúlyos szótagok egyes szótaghelyekre eső arányainak a számítását én is alkalmaztam a magyar ütemhangsúlyos metrum felismerésére létrehozott algoritmusban.

Bár a fenti, angol nyelvű versek metrumát elemző algoritmusokból bizonyos ötleteket, megközelítési módokat át tudtam venni, hangsúlyozandó, hogy a magyar időmértékes és ütemhangsúlyos versmetrum különbözik az angol és a többi ma is beszélt indoeurópai nyelv nagy részének a versmetrumától, vagyis ezen algoritmusokat nem lehet alkalmazni a magyar nyelvű versekre. A magyar versmetrum szabályalapú gépi felismertetéséhez a magyar nyelv, illetve a magyar időmértékes és ütemhangsúlyos metrum speciális jellemzőit figyelembe vevő algoritmusok kidolgozása szükséges. Az alábbiakban bemutatandó két algoritmus létrehozása során elsődleges verstani kézikönyvként a Szepes Erika és Szerdahelyi István által írt Verstan című összefoglaló monográfiát használtam.[13]

3. Az időmértékes metrum gépi felismertetésének lépései

A cél egy olyan algoritmus kialakítása volt, amely első lépésben a verssorokat, második lépésben pedig a verseket teszteli a négy időmértékes metrumra, a daktilikus, az anapesztikus, a jambikus és a trochaikus kategóriákra, mindegyik esetében egy 0 és 1 közötti szabályossági pontszámot rendelve először a verssorokhoz, majd pedig a versekhez. Minél nagyobb a tesztelt metrumra kapott szabályossági pontszám, annál inkább megvalósítja a vers az adott metrumot. A kapott szabályossági pontszámok alapján megtudhatjuk, hogy a négyből melyik metrumot valósítja meg leginkább az adott vers, illetve megadhatunk egy küszöbértéket is, amelyet ha nem halad meg a vers szabályossági pontszáma egyik tesztelt metrum esetében sem, akkor a vers nem sorolódik be egyik metrumkategóriába sem. Ez az eljárás tehát az adott metrumhoz tartozást nem igen/nem kérdésként, hanem fokozati jellemzőként ragadja meg, amely fokozati jellemző persze a küszöbértékek használatával átalakítható igen/nem bináris tulajdonsággá.[14] Ennek az eljárásnak több előnye is van. Egyrészt az algoritmus futtatása során különböző küszöbértékekkel kísérletezhetünk, adott esetben a kutatási kérdés függvényében változtatva azt. Másrészt az időmértékes metrum vizsgálata nemcsak arra a kérdésre terjedhet ki, hogy hány darab jambikus, trochaikus stb. vers van egy adott korpuszban, hanem arra is, hogy egy adott metrumkategória megvalósulásaiként felismert versek mennyire szabályosak, azaz milyen mértékben valósítják meg az adott metrum elvont ritmusképletét. Például a szabályossági pontszám alapján

[12] Agirrezabal et al., „*ZeuScansion:* A Tool for Scansion of English Poetry".

[13] Szepes Erika és Szerdahelyi István, *Verstan* (Budapest: Gondolat Kiadó, 1981).

[14] Mindez persze nem jelenti azt, hogy ne lennének minimumfeltételei annak, hogy egy verssor egy adott metrumkategóriába besorolható-e.

vizsgálhatjuk azt, hogy a jambikus versek kötöttebb ritmusa hogyan lazult fel bizonyos szerzők vagy időszakok esetében.

A versek metrumának elemzése a verssorok ritmusának az elemzésére épül rá. A ritmus elemzésére a *hunpoem_analyzer-TEI* elnevezésű programot használtam, amelyet az ELTE Verskorpusz[15] különböző, formailag egyszerűbben megragadható hangzásjellemzőinek az automatikus annotálásához írtam.[16] Ennek a programnak az egyik funkciója a verssorok időmértékes ritmusának elemzése, azaz annak megállapítása, hogy a sorokban az egyes szótagok hosszúak vagy rövidek-e. A sorok hosszú és rövid szótagjainak az elemzése néhány egyszerű, a magyar verstanban jól ismert szabály alapján elvégezhető, így az előző részben bemutatott példáktól eltérően nem szükséges kiejtési szótárakat beépíteni az algoritmusba. Ezek a szabályok a következők:

(1) a program rövid szótagként elemzi azokat a szótagokat, amelyekben rövid magánhangzó van, és közvetlenül a rövid magánhangzó után nem áll mássalhangzó, vagy csak egy rövid mássalhangzó áll;
(2) a program hosszú szótagként elemzi azokat a szótagokat, amelyekben hosszú magánhangzó áll, valamint azokat a rövid magánhangzós szótagokat, amelyekben hosszú vagy egynél több mássalhangzó követi a magánhangzót;
(3) a program figyelembe veszi azt a verstani szabályt is, miszerint a szó eleji mássalhangzó-torlódások (pl. *krákog, trottyos, strigula*) nem nyújtják meg az előtte lévő rövid magánhangzóra végződő szótagot, vagyis azok nem hosszúnak, hanem rövidnek számítanak.

Mivel a szótagok hosszúságának a megállapításában fontos információ az, hogy a magánhangzót egy vagy több mássalhangzó követi-e, szükséges volt a programba beépíteni annak vizsgálatát is, hogy egy két- vagy háromjegyű mássalhangzónak tűnő karaktersorozat valóban két- vagy háromjegyű mássalhangzónak, azaz egy fonémának tekintendő. Ehhez az e-magyar program morfológiai elemzőjét használtam.[17] Amennyiben a két- vagy háromjegyű mássalhangzónak tűnő karaktersorozat közé morfémahatár esik, akkor az nem tekinthető egy fonémának. Az időmértékes metrum automatikus meghatározását elvégző algoritmus bemenetét a ritmuselemzés kimenete adja, amely minden verssor esetében egy 1 és 0 karakterekből álló karaktersor, amelyben a 0 karakter a rövid, azaz egymorás, az 1 pedig a hosszú, azaz kétmorás szótagokat reprezentálja (pl. „Nincsen apám, se anyám” – 1001001).

[15] ELTE Verskorpusz, hozzáférés: 2021.09.01, `https://verskorpusz.elte-dh.hu/`.

[16] Horváth Péter, „A vershangzás jellemzőinek automatikus feltárása József Attila verseiben,” *Digitális Bölcsészet* 3. sz. (2020): 3–27, `https://doi.org/10.31400/dh-hun.2020.3.422`; Horváth Péter, „Az ELTE Verskorpusz automatikus annotációs eljárásai révén nyerhető kvantitatív adattípusok,” in Simon Gábor és Tolcsvai Nagy Gábor, szerk., *Nyelvtan, diskurzus, megismerés,* 313–332 (Budapest: Eötvös Kiadó, 2020).

[17] Váradi Tamás, Simon Eszter, Sass Bálint et al., „Az e-magyar digitális nyelvfeldolgozó rendszer,” in Vincze Veronika, szerk., *XIII. Magyar Számítógépes Nyelvészeti Konferencia,* 49–60 (Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2017); Novák Attila, Rebrus Péter és Ludányi Zsófia, „Az emMorph morfológiai elemző annotációs formalizmusa,” in Vincze, *XIII. Magyar Számítógépes Nyelvészeti Konferencia,* 70–78; Indig Balázs, Sass Bálint, Simon Eszter et al., „emtsv – Egy formátum mind felett,” in Berend Gábor, Gosztolya Gábor és Vincze Veronika, szerk., *XV. Magyar Számítógépes Nyelvészeti Konferencia,* 235–247 (Szeged: Szegedi Tudományegyetem TTIK, Informatikai Intézet, 2019).

 DOI: 10.31400/dh-hun.2021.4.2361

Mivel az időmértékes ritmus elemzését elvégző *hunpoem_analyzer-TEI* programnak az ELTE Verskorpusz annotációjában is szereplő kimenete 0 (rövid szótag) és 1 (hosszú szótag) karakterekből áll, az időmértékes ritmusból metrumra absztraháló algoritmus a metrumok tesztelése előtt átkonvertálja a beolvasott karaktersorok 0 és 1 karaktereit a moraszámnak megfelelő karakterekké, azaz a 0 karaktereket 1 karakterekre, az 1 karaktereket pedig 2 karakterekre változtatja (pl. „Nincsen apám, se anyám" – 2112112). Ez az átalakítás megkönnyíti a moraszámokon alapuló verslábakra bontást.

A metrumok tesztelése során az algoritmus első lépésben verslábakra bontja a sorok ritmusát reprezentáló, 1 és 2 karakterekből álló karaktersorokat. Ez az anapesztikus és daktilikus, valamint a jambikus és trochaikus metrumok tesztelése esetében eltérő módon valósul meg. Az anapesztikus és daktilikus metrumok esetében a teljes lábak mindig négymorásak, hiszen nemcsak az alaplábak, a daktilus (– ∪∪), illetve az anapesztus (∪∪ –), hanem az ezeket gyakran helyettesítő spondeus (– –) és elvétve helyettesítő proceleuzmatikus (∪∪∪∪) is négymorás, azaz időértékük négy rövid szótag időértékét teszi ki. Az anapesztikus és daktilikus metrumok tesztelése során tehát az algoritmus első lépésben megpróbálja a verssorok hosszú és rövid szótagjait reprezentáló, 1 és 2 karakterekből álló karaktersorokat négy moránként verslábakra bontani, azzal az engedménnyel, hogy az utolsó láb lehet csonka láb vagy olyan teljes láb, ahol a négymorás érték úgy jön ki, hogy a sorban szereplő utolsó, rövid szótag konvencionálisan hosszú szótagot, vagy a sorban szereplő utolsó, hosszú szótag rövid szótagot helyettesít. Ez alapján tehát a négymorás metrumú verssorok utolsó lába lehet egy hosszú vagy rövid szótagból álló csonka láb, illetve egy négymorás teljes lábként funkcionáló, anapesztusnak (∪∪ –) számítható tribrachisz (∪∪∪), egy úgyszintén teljes láb értékű, spondeusnak (– –) számítható trocheus (– ∪), valamint egy daktilusnak (– ∪∪) számítható krétikus (– ∪ –) és egy proceleuzmatikusnak (∪∪∪∪) számítható negyedik paión (∪∪∪ –). Ha a verssor négymorás lábakra való felbontása nem megvalósítható, akkor az azt jelenti, hogy a verssor nem lehet anapesztikus vagy daktilikus, vagyis a sor szabályossági pontszáma mind az anapesztikus, mind a daktilikus metrum esetében 0 lesz.

Bár az itt bemutatandó algoritmusnak nem célja, hogy felismerje azokat a kötöttebb metrumképleteket, ahol nemcsak a sor végén, hanem a sor belsejében is szerepel csonka láb, a pentameter esetében mégis kivételt tettem, hiszen a pentameter a hexameter mellett a daktilikus metrumok legtipikusabb fajtája, a hexameter és a pentameter együttes előfordulása adja a magyar költészetben is gyakori disztichont. A hexametert az algoritmus a fent bemutatott módszer alapján be tudja sorolni a daktilikus kategóriába, hiszen a hexameterben hat darab négymorás teljes láb szerepel. A pentameterben azonban a harmadik versláb egy fél láb, egy darab hosszú szótag, amely után ráadásul cezúra, azaz szóhatár van.[18] A daktilikus metrum tesztelésébe így külön beépítettem egy funkciót, amely a verssort megpróbálja pentameterként elemezni, és csak ezt követően, ennek sikertelensége esetén kezdi meg az algoritmus a sor ritmusát jelölő karaktersor négymorás verslábakra szabdalását. A pentameter felismeréséhez az algoritmus egyrészt egy egyszerű illesztéses módszerrel megvizsgálja,

[18] Szepes és Szerdahelyi két hémiepesz kólon kapcsolataként írja le a pentametert. Szepes és Szerdahelyi, *Verstan,* 219–220. Mivel az algoritmus elemzésének alapegysége a versláb, megmaradtam a pentameter hagyományos megközelítésénél.

hogy az előzetesen listázott, pentameternek tekinthető ritmusképletek valamelyikének megfelel-e az adott sor ritmusa, majd pedig, amennyiben az illesztés sikeres volt, megvizsgálja, hogy a sorban a harmadik, csonka lábat követően új szó kezdődik-e. Ha mind a két feltételnek megfelel a sor, akkor az algoritmus a pentameternek megfelelő tagolással, azaz egy soron belüli fél lábbal osztja verslábakra a sort.

A jambikus és trochaikus metrumú versek esetében a teljes lábak általában két szótagosak. A lábak moraszámára nincsen megkötés: a két metrum alaplábát adó jambus (∪ –), illetve trocheus (– ∪) hárommorás, az ezeket helyettesítő spondeus (– –) négymorás, de úgyszintén megjelenhet helyettesítő lábként a kétmorás pirrichius (∪∪) is. A jambikus és trochaikus metrumok tesztelésénél az algoritmus első lépésben két szótagonként felosztja a verssorok hosszú és rövid szótagjait reprezentáló, 1 és 2 karakterekből álló karaktersorokat. Az utolsó láb ebben az esetben is lehet egy szótagú csonka láb. Speciális esetekben megjelenhetnek a jambikus és trochaikus sorokban három szótagú verslábak is: jambust helyettesítő anapesztus (ciklikus anapesztus) és trocheust helyettesítő daktilus (ciklikus daktilus). Az ilyen sorokat az algoritmus nem tudja helyesen elemezni.

Az algoritmus második lépése annak ellenőrzése, hogy a verssor megfelel-e bizonyos, a tesztelt metrumhoz kapcsolódó minimumfeltételeknek. Az anapesztikus és daktilikus metrumok tesztelése esetében, amennyiben a sort fel lehetett osztani négymorás verslábakra, az algoritmus megvizsgálja, hogy a verslábak között van-e amphibrachisz (∪ – ∪). Az amphibrachisz helyettesítő lábként való megjelenése az anapesztikus vagy daktilikus metrumoknál nem jellemző, az amphibrachisz jelenléte így általában arra utal, hogy annak ellenére, hogy a verssort négymoránként verslábakra lehet tagolni, a sor anapesztikus vagy daktilikus metrumúként való elemzése rossz irány, és az sokkal inkább lesz jambikus vagy trochaikus (vagy pedig nem jellemző rá semmilyen időmértékes metrum). Amennyiben tehát szerepel amphibrachisz a vizsgált verssorban, a verssor szabályossági pontszáma 0 lesz mind az anapesztikus, mind a daktilikus metrum tesztelése esetén.

A jambikus és trochaikus metrumok tesztelése esetében a verssor két szótagonként történő verslábakra osztását követően az algoritmus megvizsgálja a sor utolsó teljes lábát. A jambikus metrum minimumfeltétele ugyanis az, hogy amennyiben a sor teljes lábra végződik, az utolsó láb jambus (∪ –) vagy pirrichius (∪∪), amennyiben csonka lábra végződik, akkor pedig az utolsó előtti lábnak, azaz az utolsó teljes lábnak jambusnak kell lennie. A trochaikus metrum minimumfeltétele a sor teljes lábra végződése esetén az utolsó láb trocheus (– ∪) vagy spondeus (– –) volta, csonka lábra végződés esetén pedig az utolsó előtti láb, azaz az utolsó teljes láb trocheus volta. A jambikus vagy trochaikus kategóriába való besorolásnak e minimumfeltételei követik a verstanoknak azt az elképzelését, miszerint a jambikusság, illetve trochaikusság elsődleges kritériuma az utolsó teljes láb jambus vagy trocheus volta, ahol a jambust konvencionálisan pirrichius, a trocheust pedig konvencionálisan spondeus helyettesítheti, amennyiben az utolsó teljes lábat nem követi egy további csonka láb. Ha a verssor a tesztelt jambikus vagy trochaikus metrum e minimumfeltételeit nem teljesíti, akkor a szabályossági pontszáma a sornak az adott metrum esetében 0 lesz.

Az alábbiakban felsorolásszerűen összegzem, hogy az algoritmus a négy metrum esetében milyen minimumfeltételeket érvényesít. Amennyiben a verssor nem felel

 DOI: 10.31400/dh-hun.2021.4.2361

meg a tesztelt metrum alább megadott minimumfeltételeinek, akkor a verssornak az adott metrumra vonatkozó szabályossági pontszáma 0.

Anapesztikus és daktilikus metrum minimumfeltételei:

- A verssor teljes lábai négymorásak.
- A verslábak között nincs amphibrachisz.

Jambikus metrum minimumfeltételei:

- Ha a verssor teljes lábra végződik, akkor az utolsó láb jambus vagy pirrichius, ha csonka lábra végződik, akkor az utolsó előtti láb, azaz az utolsó teljes láb jambus.

Trochaikus metrum minimumfeltételei:

- Ha a verssor teljes lábra végződik, akkor az utolsó láb trocheus vagy spondeus, ha csonka lábra végződik, akkor az utolsó előtti láb, azaz az utolsó teljes láb trocheus.

Amennyiben az elemzett verssor megfelel a tesztelt metrum minimumfeltételeinek, akkor az algoritmus kiszámolja a sor szabályossági pontszámát, azaz megad egy 0 és 1 közötti értéket, amely azt jelzi, hogy az adott sor ritmusa milyen mértékben felel meg a tesztelt metrumnak. Minél közelebb van a szabályossági pontszám az 1-hez, a tesztelt metrum szempontjából annál szabályosabb a sor. A szabályossági érték számolásának alapja a verslábak pontozása. Minden teljes láb 0-tól 4-ig terjedő pontszámot kap attól függően, hogy melyik metrumra teszteli az algoritmus a verssort. Négy pontot azok a verslábak kapnak, amelyek a tesztelt metrumkategóriának az alaplábai. A daktilikus metrum esetében a daktilus, az anapesztikus metrum esetében az anapesztus, a jambikus metrum esetében a jambus, a trochaikus metrum esetében pedig a trocheus. Két pontot kap a spondeus, amely mind a négy metrum esetében konvencionálisan használható helyettesítő láb. Egy pontot kapnak a kevésbé konvencionális helyettesítő lábak: a daktilikus és anapesztikus metrum tesztelése esetén a proceleuzmatikus, a jambikus és trochaikus metrum tesztelése esetén pedig a pirrichius. Nulla pontot kapnak a tesztelt metrum alaplábával ellentétes lábak, azaz daktilikus metrum tesztelése esetén az anapesztus, anapesztikus metrum tesztelése esetén a daktilus, jambikus metrum tesztelése esetén a trocheus, trochaikus metrum tesztelése esetén pedig a jambus. A sorvégi fél lábakat, illetve pentameter esetében a sor végi és soron belüli fél lábat az algoritmus nem veszi figyelembe a sorok szabályossági pontszámainak a számítása során. Az algoritmus minden sor esetében összeadja a verslábak pontszámait, majd az összeget elosztja a sorban szereplő teljes lábak számának a négyszeresével. Az így kapott pontszám a sornak egy adott metrumra vonatkozó szabályossági pontszáma. Az algoritmus tehát az alábbi képlet alapján számítja ki a sorok szabályossági pontszámát.

$$s = \frac{4a+2b+c}{4d}$$

s = a verssor szabályossági pontszáma
a = alaplábak száma
b = elsődleges helyettesítő lábak száma
c = másodlagos helyettesítő lábak száma
d = teljes lábak száma (d = a + b + c)

Ahogyan azt fentebb már írtam, a sorvégi rövid szótagok konvencionálisan hosszúnak, a sorvégi hosszú szótagok pedig rövidnek is számíthatnak. Így az algoritmus a daktilikus és anapesztikus metrum tesztelésénél az utolsó lábnál szereplő tribrachiszt (∪∪∪) anapesztusnak (∪∪ –), a trocheust (– ∪) spondeusnak (– –), a krétikust (– ∪ –) daktilusnak (– ∪∪), a negyedik paiónt (∪∪∪ –) pedig proceleuzmatikusnak (∪∪∪∪) veszi, és ennek megfelelően pontozza őket. A jambikus és trochaikus metrum tesztelésekor pedig az utolsó lábnál szereplő pirrichiust (∪∪) jambusnak (∪ –), a spondeust (– –) pedig trocheusnak (– ∪) veszi, és a pontozás is ennek megfelelően történik. A fenti általános képletet az algoritmusban természetesen az egyes metrumokra specifikálva alkalmaztam, ezek a képletek az alábbi módon adhatók meg.

daktilikus:

$$s = \frac{daktilusszám \times 4 + spondeusszám \times 2 + proceleuzmatikusszám}{verslábszám \times 4}$$

anapesztikus:

$$s = \frac{anapesztusszám \times 4 + spondeusszám \times 2 + proceleuzmatikusszám}{verslábszám \times 4}$$

jambikus:

$$s = \frac{jambusszám \times 4 + spondeusszám \times 2 + pirrichiusszám}{verslábszám \times 4}$$

trochaikus:

$$s = \frac{trocheusszám \times 4 + spondeusszám \times 2 + pirrichiusszám}{verslábszám \times 4}$$

A képletekből láthatjuk, hogy amennyiben egy sor összes teljes lába a tesztelt metrum alaplába, például jambikus metrum tesztelése esetén a sor összes teljes lába jambus, akkor a számlálóba és a nevezőbe ugyanaz a szám kerül, és így a sor jambikus metrumra kapott szabályossági pontszáma 1 lesz. Minél több a verssorban a helyettesítő versláb (például jambikus metrum tesztelése esetén spondeus vagy pirrichius), illetve a metrum alaplábával ellentétes láb (például jambikus metrum tesztelése esetén trocheus), annál alacsonyabb szám kerül a számlálóba, és így annál alacsonyabb lesz a sor szabályossági pontszáma. Az alábbi *1. táblázat* néhány négy lábból álló sornak a daktilikus és jambikus metrum tesztelése esetén előálló szabályossági pontszámát mutatja be. A szabályossági pontszámokat két tizedesjegyre kerekítettem.

 DOI: 10.31400/dh-hun.2021.4.2361

1. táblázat. Különböző négy lábú ritmusok szabályossági pontszáma daktilikus és jambikus metrumok tesztelése esetén

Daktilikus		Jambikus	
Ritmusképlet	Szabályossági pontszám	Ritmusképlet	Szabályossági pontszám
– ∪∪ | – ∪∪ | – ∪∪ | – ∪∪	1,00	∪ – | ∪ – | ∪ – | ∪ –	1,00
– ∪∪ | – ∪∪ | – ∪∪ | – –	0,88	∪ – | ∪ – | – – | ∪ –	0,88
– ∪∪ | – ∪∪ | – – | – –	0,75	– – | ∪ – | – – | ∪ –	0,75
– – | – ∪∪ | – ∪∪ | – –	0,75	– – | – – | – – | ∪ –	0,63
– – | – ∪∪ |– – | – –	0,63	∪ – | ∪ – | ∪∪ | ∪ –	0,81
∪∪∪∪ | – ∪∪ | – – | – –	0,56	– – | ∪ – | ∪∪ | ∪ –	0,69
∪∪∪∪ | – ∪∪ | ∪∪∪∪ | – –	0,50	∪∪ | ∪ – | ∪∪ | ∪ –	0,63
– ∪∪ | – ∪∪ | – ∪∪ | ∪∪ –	0,75	∪∪ | – – | ∪∪ | ∪ –	0,50
– ∪∪ | – – | – ∪∪ | ∪∪ –	0,63	– ∪ | – – | ∪∪ | ∪ –	0,44
– ∪∪ | ∪∪∪∪ | – ∪∪ | ∪∪ –	0,56	– ∪ | – ∪ | – ∪ | ∪ –	0,25

A táblázatból látható, hogy minél kevésbé követi egy verssor ritmusa a tesztelt metrum elvont ritmusképletét, annál kisebb a szabályossági pontszám, egy bizonyos pontszám alatt pedig az adott sor már nem igazán tekinthető a daktilikus vagy jambikus metrum megvalósulásának.

Miután az algoritmus kiszámolta egy vers minden sorának a tesztelt metrumhoz kapcsolódó szabályossági pontszámát, megadja az egész versnek a tesztelt metrumra vonatkozó szabályossági pontszámát is. Ez egy egyszerű számítás alapján történik: az algoritmus összeadja az egyes sorok szabályossági pontszámait, majd az összeget elosztja az összes sor számával. A képlet tehát a következő (v a vers szabályossági pontszáma, s a verssorok szabályossági pontszáma, n a sorok száma):

$$v = \frac{\sum_{i=1}^{n} s_i}{n}$$

A számítás eredménye a verssorok szabályossági pontszámához hasonlóan egy 0 és 1 közötti szám, ahol az 1 érték a versnek a tesztelt metrum szempontjából teljesen szabályos voltát jelzi. Az 1 érték akkor jöhet ki, ha minden sor szabályossági pontszáma 1. Minél alacsonyabb a sorok szabályossági pontszáma, annál alacsonyabb lesz a vers egészére kapott szabályossági pontszám is.

Az algoritmus utolsó lépése az elemzett vers négy metrumra kapott szabályossági pontszámainak az összevetése. A verset az algoritmus abba a metrumkategóriába sorolja be, amelyre a legnagyobb szabályossági pontszám jött ki. Az algoritmus futtatása előtt megadható egy küszöb is. Ez egy olyan 0 és 1 közötti érték, amelyet meg kell haladnia a legnagyobb szabályossági pontszámnak ahhoz, hogy a verset az algoritmus valóban besorolja az adott metrumkategóriába. Mindenképpen érdemes küszöbértéket megadni, mivel az időmértékes metrum fokozati megközelítése miatt, még ha alacsony szabályossági pontszámmal is, de gyakorlatilag minden vers besorolódik a négyből

valamelyik metrumkategóriába.[19] A küszöbérték megadásával biztosíthatjuk, hogy az algoritmus kimeneteként csak a valóban időmértékes metrumú versek sorolódjanak be valamelyik időmértékes metrum kategóriájába.

Az alábbi felsorolás pontokba szedve összegzi a fent bemutatott, az időmértékes metrum gépi felismertetésére kidolgozott algoritmus főbb lépéseit:

1. A szótagok hosszúságának (rövid vagy hosszú) megállapítása
2. A vers tesztelése a négy metrumra
    a. A verssorok verslábakra osztása a tesztelt metrumnak megfelelően
    b. A tesztelt metrum minimumfeltételeinek ellenőrzése
    c. A verssorok szabályossági pontszámának kiszámítása a tesztelt metrumnak megfelelően
    d. A vers szabályossági pontszámának kiszámítása a sorok szabályossági pontszáma alapján
3. A vers négy metrumra kapott szabályossági pontszámának és a megadott küszöbértéknek az összevetése

A felsorolásban az (a)–(d) lépéseket beljebb szedtem. Ezzel azt próbáltam jelezni, hogy az (a)–(d) olyan, a (2) nagyobb lépés részét képező ismétlődő lépéssorozat, amelyet az algoritmus egy vers elemzésekor mind a négy metrum tesztelésénél elvégez. Ezen ismétlődő lépéssorozat eredményeképpen áll elő a verset jellemző, négy metrumra kapott négy szabályossági pontszám, amelyeket az algoritmus a (3) lépésben összevet egymással, illetve az előzetesen megadott küszöbértékkel.

4. Az időmértékes metrumot elemző algoritmus alkalmazása tizenegy költő versein

Az előző részben bemutatott algoritmust Python nyelvben implementáltam[20] oly módon, hogy annak bemenetét az ELTE Verskorpusz automatikus annotálására írt, úgyszintén Pythonban fejlesztett *hunpoem_analyzer-TEI* program TEI XML kimenete adja, amely többek között tartalmazza a verssoroknak a hosszú és rövid szótagokat megkülönböztető ritmuselemzését. Az időmértékes metrum felismerésére létrehozott programot tizenegy, a 19. század közepén, illetve második felében, valamint a 20. század első felében alkotó költőnek az ELTE Verskorpuszból[21] vett versein futtattam le.

[19] Megjegyzendő, hogy van elvi lehetősége annak, hogy egy vers mind a négy metrum esetében 0 szabályossági pontszámot kapjon. Ez abban az esetben következhetne be, ha a versnek egy olyan sora sincsen, amely a négy metrumból legalább az egyik minimumfeltételeit teljesítené.

[20] A kód a tanulmány mellékleteként letölthető: `https://doi.org/10.31400/dh-hun.2021.4.2361`.

[21] Horváth Péter, Kundráth Péter, Indig Balázs et al., „ELTE Verskorpusz – a magyar kanonikus költészet gépileg annotált adatbázisa," in Berend Gábor, Gosztolya Gábor és Vincze Veronika szerk., *XVIII. Magyar Számítógépes Nyelvészeti Konferencia,* 375–88 (Szeged: Szegedi Tudományegyetem TTIK, Informatikai Intézet, 2022); *ELTE Poetry Corpus,* hozzáférés: 2021.09.01, `https://github.com/ELTE-DH/poetry-corpus`. A level3 mappában szerepel a vershangzásra vonatkozó annotációkat is tartalmazó TEI XML kimenet.

 DOI: 10.31400/dh-hun.2021.4.2361

A tizenegy költő születésük időrendjében a következő: Tompa Mihály, Arany János, Petőfi Sándor, Gyulai Pál, Vajda János, Reviczky Gyula, Komjáthy Jenő, Ady Endre, Babits Mihály, Kosztolányi Dezső, József Attila.

A *2. táblázat* a vizsgált tesztkorpusz időmértékes metrumú verseinek az előfordulási adatait mutatja be. A metrumok szabályossági pontszámához kapcsolódó küszöböt 0,5 értékben adtam meg.

2. táblázat. Időmértékes metrumú versek gyakorisága (küszöbérték: 0,5)

Szerző	Összes	Időm.	Arány	Jambikus		Trochaikus	
				Előford.	Arány	Előford.	Arány
Tompa Mihály	491	429	0,87	307	0,72	116	0,27
Arany János	417	285	0,68	115	0,40	156	0,55
Petőfi Sándor	839	600	0,72	352	0,59	240	0,40
Gyulai Pál	156	112	0,72	43	0,38	69	0,62
Vajda János	199	156	0,78	85	0,54	69	0,44
Reviczky Gyula	335	331	0,99	271	0,82	57	0,17
Komjáthy Jenő	246	213	0,87	188	0,88	21	0,10
Ady Endre	1116	369	0,33	260	0,70	109	0,30
Babits Mihály	514	314	0,61	247	0,79	59	0,19
Kosztolányi Dezső	630	534	0,85	503	0,94	26	0,05
József Attila	599	309	0,52	249	0,81	50	0,16

A táblázat második oszlopában szerepel a szerzőkhöz tartozó versek száma, a harmadik oszlopban pedig azt tüntettem fel, hogy ebből a program mennyit elemzett időmértékes metrumúnak, azaz hány vers szabályossági pontszáma haladta meg a 0,5 értéket a négyből legalább az egyik metrum esetében.[22] A negyedik oszlop az időmértékesként elemzett versek arányát mutatja be az összes vershez képest. Az utolsó négy oszlop a jambikusként és trochaikusként felismert versek számát és az időmértékesként felismert versekhez viszonyított arányát mutatja be. A táblázatból látható, hogy az elemzőprogram jól meg tudta ragadni azt az ismert tendenciát, hogy míg a 19. század első harmadában született szerzők (Tompa, Arany, Petőfi, Gyulai, Vajda) nagy arányban írtak trochaikus metrumú verseket – Aranynál és Gyulainál az így elemzett versek meghaladják az időmértékesként elemzett versek 50%-át –, addig a 19. század utolsó harmadától a trochaikus metrum visszaszorul, a jambikus metrumú versek aránya pedig még nagyobb lesz. A jambikus versek Kosztolányinál fordulnak elő a legnagyobb arányban: az időmértékesként azonosított versek 94%-át ismerte fel jambikusként a program. A táblázat harmadik oszlopából azt is láthatjuk, hogy a program Reviczkynek szinte az összes versét időmértékesként elemezte, ezzel szemben Ady esetében mindössze a versek 33%-át ismerte fel időmértékesként.

[22] Ha egy vers több metrumra is pontosan ugyanazt a szabályossági pontszámot kapja, a program nem sorolja be a verset egyik időmértékes metrumkategóriába sem, és azt az általános „időmértékes" kategóriába sem sorolja be. Ez az eljárás a kapott előfordulási adatokat csak minimálisan módosítja. 0 küszöbérték esetén 10 vers kapott egynél több metrumra azonos szabályossági pontszámot, 0,3 küszöbérték esetén 5 vers, 0,5 küszöbérték esetén pedig nem volt ilyen vers.

A programot lefuttattam 0,3 küszöbértékkel is. Az így nyert előfordulási adatokat a *3. táblázat* mutatja be. Látható, hogy ezzel a küszöbértékkel már Ady verseinek a jelentős részét is időmértékesként ismeri fel a program. Mindez azt mutatja, hogy Ady versritmusa a tesztkorpuszban szereplő többi költő versritmusához képest kevésbé szabálykövető, vagy legalábbis nem azokat a szabályokat követi, mint a többi szerző versritmusa, és amelyek a programban az időmértékes metrum tipikus jellemzőiként implementálva lettek.[23]

3. táblázat. Időmértékes metrumú versek gyakorisága (küszöbérték: 0,3)

Szerző	Összes	Időm.	Arány	Jambikus		Trochaikus	
				Előford.	Arány	Előford.	Arány
Tompa Mihály	491	483	0,98	326	0,67	150	0,31
Arany János	417	396	0,95	136	0,34	226	0,57
Petőfi Sándor	839	791	0,94	416	0,53	351	0,44
Gyulai Pál	156	148	0,95	49	0,33	96	0,65
Vajda János	199	196	0,98	99	0,51	95	0,48
Reviczky Gyula	335	334	1,00	271	0,81	60	0,18
Komjáthy Jenő	246	243	0,99	198	0,81	36	0,15
Ady Endre	1116	910	0,82	694	0,76	209	0,23
Babits Mihály	514	429	0,83	295	0,69	117	0,27
Kosztolányi Dezső	630	606	0,96	535	0,88	61	0,10
József Attila	599	504	0,84	366	0,73	126	0,25

A *4. táblázat* az anapesztikusként, illetve daktilikusként elemzett versek számát mutatja be 0,5 és 0,3 küszöbértékkel. Babitsnál 0,5 küszöbértékkel a következő verseket ismerte fel a program daktilikusként: *Extasis, A sziget nem elég magas, Ekloga, Klasszikus álmok, Új Leoninusok, Vágy ered, Május huszonhárom Rákospalotán, Márciusi reggelen.* Ezek között a versek között disztichonok is vannak *(Új Leoninusok, Május huszonhárom Rákospalotán).* A disztichon egy hexameter és egy pentameter váltakozására épül. Mivel a programba külön beépítettem a pentameter felismerésének funkcióját, a disztichonban írt verseket a program fel tudja ismerni daktilikusként.

4. táblázat. Anapesztikus és daktilikus versek előfordulása (küszöbérték: 0,5 és 0,3)

Szerző	Anapesztikus		Daktilikus	
	> 0,5	> 0,3	> 0,5	> 0,3
Tompa Mihály	3	3	3	4
Arany János	1	8	13	26
Petőfi Sándor	5	12	3	12
Gyulai Pál	0	3	0	0

[23] Ady speciális versritmusára több szerző is felhívta már a figyelmet, például: Horváth János, *Rendszeres magyar verstan* (Budapest: Akadémiai Kiadó, 1969), 164–165; Gáldi László, *Ismerjük meg a versformákat!* (Budapest: Móra Ferenc Ifjúsági Könyvkiadó, 1961), 103–106; Vargyas Lajos, *Magyar vers – magyar nyelv: Verstani tanulmány* (Budapest: Szépirodalmi Könyvkiadó, 1966), 158–165.

 DOI: 10.31400/dh-hun.2021.4.2361

Vajda János	1	1	1	1
Reviczky Gyula	2	2	1	1
Komjáthy Jenő	0	1	4	8
Ady Endre	0	3	0	4
Babits Mihály	0	7	8	10
Kosztolányi Dezső	3	6	2	4
József Attila	1	2	9	10

Mivel a program nem csupán besorolja egy metrumkategóriába az időmértékesként felismert verset, hanem minden vers esetében egy szabályossági pontszámot is megad, vizsgálhatjuk azt is, hogy az egyes szerzők esetében milyen eltéréseket mutatnak a szabályossági pontszámok középértékei. Az *5. táblázat* a 0,5 küszöbértékkel jambikusként elemzett versek szabályossági pontszámainak az átlagát és mediánját mutatja be. Minél nagyobb egy szerző esetében az átlag, illetve a medián, annál szabálykövetőbben alkalmazta az adott szerző a jambikus metrumot. Látható, hogy a tesztkorpusz szerzői közül Reviczky és Komjáthy esetében a legmagasabbak ezek a középértékek, vagyis az ő jambikus versritmusuk követte leginkább a jambikus metrum elvont ritmusképletét. A legalacsonyabb középértékeket Adynál találjuk, ami megint csak megerősíti azt, hogy Ady versritmusa elüt a tesztkorpuszban szereplő többi szerző versritmusától, kevésbé követte azokat a szabályokat, amelyeket a többiek.

5. táblázat. Jambikus versek szabályossági pontszámainak középértékei (küszöbérték: 0,5)

Szerző	Átlag	Medián
Tompa Mihály	0,72	0,73
Arany János	0,71	0,72
Petőfi Sándor	0,68	0,69
Gyulai Pál	0,67	0,66
Vajda János	0,65	0,65
Reviczky Gyula	0,77	0,78
Komjáthy Jenő	0,77	0,77
Ady Endre	0,58	0,57
Babits Mihály	0,67	0,68
Kosztolányi Dezső	0,74	0,75
József Attila	0,63	0,61

5. Az ütemhangsúlyos metrum gépi felismertetésének lépései

Az alábbiakban bemutatandó második eljárás célja az egyszerűbb, aaaa, illetve abab szerkezetű ütemhangsúlyos metrumú versek felismerése, tehát azon ütemhangsúlyos versek felismerése, amelyeknek minden sorára ugyanaz az ütemosztás jellemző, vagy pedig két típusú, a versszakok páratlan és páros soraira jellemző ütemosztás váltogatja

egymást. Az ütemhangsúlyos verselés alapja az ütemeleji hangsúlyos szótagok szabályszerű visszatérése. A magyar kötött hangsúlyú nyelv, azaz a szavaknak – speciális helyzetektől eltekintve – csupán az első szótagja lehet hangsúlyos, a többi szótag pedig hangsúlytalan. Bár a különféle egy szótagos funkciószavak (névelők, kötőszavak stb.) általában hangsúlytalanok, ezek a magyar költészeti hagyományban adhatják az ütem elején álló hangsúlyos szótagokat, így az algoritmusnak nem szükséges megkülönböztetnie a funkciószavakat a többi szótól. Az ütemhangsúlyos verselés automatikus elemzését elvégző algoritmus első lépése a verssorok szó eleji hangsúlyos szótagjainak megállapítása. Ennek a lépésnek a kimenete az időmértékes ritmus elemzéséhez hasonlóan minden sor esetében egy 0 és 1 számokból álló karaktersor. Ezúttal azonban az 1 a hangsúlyos, azaz a szó eleji szótagokat, a 0 pedig a hangsúlytalan, azaz a nem szó eleji szótagokat reprezentálja.

Az algoritmus a kapott hangsúlyértékek összesítése alapján állapítja meg, hogy az adott versre jellemző-e valamilyen ütemhangsúlyos metrum. Ennek alapja az, hogy egy vers sorainak a hangsúlyértékeit reprezentáló, 0 és 1 számokból álló karaktersorok olyan mátrixot adnak ki, amelyben minden oszlop egy adott sorszámú szótaghelyre vonatkozik. Az oszlopokban szereplő, szó eleji szótagokat jelölő egyesek összeadásával megállapítható, hogy a vers adott sorszámú szótaghelyére esik-e ütemkezdet vagy nem.[24] Természetesen egy ütemhangsúlyos vers esetében is vannak olyan verssorok, amelyben valamelyik ütem elejére nem esik szó eleji szótag. Emiatt az algoritmust úgy alakítottam ki, hogy a program futtatásának kezdetén a felhasználó megadhatja, hogy a verssorok minimum hány százalékában legyen egy adott szótaghelyen szó eleji szótag (azaz a mátrixban 1-es karakter) ahhoz, hogy a program ütemkezdetet állapítson meg. Én a programot 0,75 értékkel futtattam, vagyis a program egy adott szótaghelyen akkor állapított meg ütemkezdetet, ha legalább a sorok 75%-ában szó eleji szótag esett az adott szótaghelyre.

Az algoritmusba nincsen beépítve a hosszú szótagokra vonatkozó ütemalkotási licencia, miszerint az ütem eleji hangsúlyos szótagokat helyenként helyettesíthetik hosszú szótagok.[25] Ugyanígy nem építettem be az összetett szavakra vonatkozó ütemalkotási licenciát sem, amely szerint az ütem eleji hangsúlyos szótag az összetett szavak második vagy további összetételi tagjainak az első szótagja, illetve nem elváló igekötős igék esetében az igekötő utáni első szótag is lehet.[26] Az ütemalkotási licenciák

[24] Megjegyzendő, hogy már a Voigt-féle 1972-es első magyar gépi ritmuselemzés is ilyen mátrixok alapján számolta az egyes szótaghelyekre eső átlagos értékeket, azzal a különbséggel, hogy a szótagok hosszúságát jelölték a mátrixban egy négyfokozatú skálán. Voigt, „Számítógépes ritmuselemzési kísérlet”. Jékel és Szuromi ugyancsak hasonló módszert használt az egyes szótaghelyekre eső szótagnyomatékok átlagolására. Jékel és Szuromi, *Petőfi metrumai.*

[25] Szepes és Szerdahelyi, *Verstan,* 375–376.

[26] Uo., 373–375. Az algoritmusnak létrehoztam egy olyan változatát, amely az *e-magyar* morfológiai elemzése alapján figyelembe veszi az összetételi tagok első szótagjait is, de ezt a verziót végül elvetettem, mivel az *e-magyar* (és minden más szabadon felhasználható, magyar nyelvre alkalmazható elemző) a szavak egyszerűsített morfológiai címkéi alapján végzi el az egyértelműsítést, azaz a több lehetséges morfoszintaktikai elemzésből az adott mondatkontextusban a legvalószínűbb kiválasztását. Ezeknek az egyszerű címkéknek azonban abban az esetben több összetett, a morfológiai felbontást is tartalmazó elemzés is megfeleltethető, ha a szótő felbontható összetételi tagokra. Ökölszabálynak jónak tűnt, hogy ha az egyértelműsített egyszerű címkének több morfológiai felbontás is megfeleltethető, akkor azt a morfológiai felbontást hasznosítsa az algoritmus, amelyben

 DOI: 10.31400/dh-hun.2021.4.2361

figyelmen kívül hagyását kompenzáltam azzal, hogy viszonylag alacsonyan, 75%-ban határoztam meg az ütem elejére eső szó eleji szótagok minimális arányát.

Az ütemek megállapításánál két megszorítást is érvényesítettem: (1) a második szótaghely nem lehet ütemkezdet; (2) egy ütem nem lehet hosszabb egy előre megadott szótagszámnál. Az (1) megszorítást azért volt szükséges bevezetni, mert a verssorok első szava sok esetben egy szótagú szó, ugyanakkor verstani szempontból nem lenne észszerű egy szótagos ütemeket feltételezni a sorok elején, főleg, hogy az itt szereplő egy szótagú szavak sok esetben hangsúlytalan kötőszavak vagy egyéb típusú funkciószavak.[27] A (2) megszorítás összhangban áll a magyar verstani kutatások azon gyakori megállapításával, miszerint egy bizonyos szótagszám felett nem beszélhetünk ütemről. Ugyanakkor ezt a maximális szótagszámot a felhasználó maga adhatja meg a program futtatásának kezdetén. Én a programot hat szótagos maximális ütemhosszúsággal futtattam. Az ütemek hosszúságára vonatkozó megszorításból az is következik, hogy az ütemekre megadott maximális szótagszámmal rendelkező vagy annál rövidebb verssorok a program elemzésében lehetnek ütemosztást nem tartalmazó önálló ütemek.[28]

---

a legtöbb morfémára lett felbontva a szó, hiszen az összetételi tagok ebben az esetben állapíthatók meg. Ez a megoldás azonban jelentősen túlgenerálta az összetett szavakat, és például olyan szavak is összetett szavakként értelmeződtek, mint a *szerelem,* a *tanács* vagy a *babér (szer|elem, tan|ács, bab|ér).* Emellett ha az algoritmus minden igekötős ige esetében hangsúlyosnak venné az igekötő utáni első szótagot, akkor valószínűleg jelentősebben megnőne annak az esélye, hogy a program olyan helyen is ütemhatárt állapítson meg, ahol valójában nem kezdődik új ütem. Például, ha az említett Weöres-vers esetében az igekötős igék igekötő utáni szótagját is automatikusan hangsúlyosnak venné az algoritmus – azaz 1-es karakterrel reprezentálná –, akkor a vers harmadik szótagja esetében is ütemkezdetet állapítana meg, ami félrevinné a vers metrumának elemzését.

[27] Felvethető, hogy a sorok elején álló egy szótagos szavak felütéseknek tekintendők. Ugyanakkor az általam áttekintett verstani szakirodalomban csak megjegyzésszerű, meglehetősen bizonytalan, az algoritmizálhatóság szempontjából nem eléggé egyértelmű megállapításokat találtam az ütemhangsúlyos metrumú versekben megjelenő esetleges felütésekről, ráadásul a szerzők többnyire a jelenség kifejezetten ritka voltát hangsúlyozzák. Gáldi László az ütemelőzőnek a magyar verstani elemzésekben „a magyar versrendszertől *teljesen idegen* és zeneileg is aránylag modern fogalmához csak kivételesen" folyamodik (Gáldi, *Ismerjük meg a versformákat!,* 19). Szepes és Szerdahelyi úgyszintén kuriózumként tartja számon a jelenséget, amely csak „bizonyos – kivételes – esetekben" jelenik meg ütemhangsúlyos verselésünkben (Szepes és Szerdahelyi, *Verstan,* 371). Megjegyzendő az is, hogy felütések leginkább csak hangsúlytalan funkciószavak lehetnének, és szinte kizárólag olyan pozícióban, ahol közvetlenül utánuk nem egy hozzájuk hasonló hangsúlytalan funkciószó, hanem egy hangsúlyos első szótaggal rendelkező szó áll. Az itt bemutatott algoritmus ugyanakkor nem vizsgálja, hogy a versben szereplő szavak funkciószavak vagy nem. A Weöres-vers is jól példázza, hogy hiába van a sor elején sok egy szótagú szó, ellenkezne a ritmusérzékünkkel, ha azokat megpróbálnánk felütésekként ritmizálni.

[28] A verstani szakirodalom alapján az korántsem egyértelmű, hogy mi lenne az ütemek maximális szótagszáma. Vargyas Lajos megközelítésében például az ütem maximálisan négy szótagos lehet, bár elemzéseiben kivételes esetekben szerepelnek négynél több szótagból álló ütemek is (Vargyas, *Magyar vers – magyar nyelv,* 9–14). Gáldi László úgyszintén alapvetően négy szótagban állapítja meg a maximális szótagszámot, de megjegyzi, hogy az kivételesen lehet öt is (Gáldi, *Ismerjük meg a versformákat!,* 27). Horváth János a szokványos és a gyors ütemtípust különíti el, az előbbi maximális szótagszáma négy, az utóbbié hat (Horváth, *Rendszeres magyar verstan,* 18–51). Arany János leírásában is az ütemek maximum négy szótagosak, ugyanakkor az ütemek nagyobb, sormetszetekkel elválasztott tagokat alkothatnak, amelyek maximum hét szótagosak lehetnek (Arany János, „A magyar nemzeti vers-idomról," in Arany János, *Tanulmányok és kritikák I.* [második,

Az alábbi *6. táblázat* Weöres Sándor *A medve töprengése* című, tizenkét hét szótagos sorból álló versének a szó eleji és nem szó eleji szótagjait mutatja be. Az 1 jelöli a szó eleji, a 0 pedig a nem szó eleji szótagokat.

6. táblázat. Weöres Sándor *A medve töprengése* című versének szó eleji szótagjai

	1. szótag	2. szótag	3. szótag	4. szótag	5. szótag	6. szótag	7. szótag
1. sor	1	1	1	0	1	1	1
2. sor	1	0	1	0	1	0	0
3. sor	1	0	0	1	1	0	0
4. sor	1	1	1	0	1	0	0
5. sor	1	1	0	0	1	0	0
6. sor	1	1	0	0	1	0	0
7. sor	1	1	0	0	1	1	0
8. sor	1	1	0	0	1	0	0
9. sor	1	0	0	1	0	0	1
10. sor	1	1	1	0	0	0	1
11. sor	1	1	1	0	1	0	0
12. sor	1	1	1	0	1	1	0
Összesen	12	9	6	2	10	3	3
Arány	100%	75%	50%	17%	83%	25%	25%

Ahogyan azt a 6. táblázattal igyekeztem szemléltetni, az algoritmus által előállított, a vers egyes soraira vonatkozó, 1 és 0 karaktereket tartalmazó karaktersorok olyan mátrixként értelmezhetők, amelyben a karaktersorok adott sorszámú karakterei a verssorok adott sorszámú szótaghelyére vonatkoznak. Egy meghatározott szótaghelyen, azaz a mátrix egy adott oszlopában szereplő számok összeadásával, majd az így kapott összegnek a vers sorainak a számával való elosztásával megkaphatjuk, hogy az adott szótaghelyre milyen arányban esik szó eleji szótag. Amennyiben ez az arány eléri a megadott küszöbértéket – az általam használt beállítással 75%-ot –, akkor a szótaghelyen az algoritmus ütemkezdetet állapít meg. A 2. táblázatból például látható, hogy ezen eljárás alapján a program az első szótaghely mellett az ötödik szótaghelynél állapítana meg ütemkezdetet, hiszen ezen a szótaghelyen a sorok 83%-ában szó eleji szótag áll, vagyis a vers metruma kétütemű hetes, ötödik szótagra eső ütemkezdettel. A második szótaghelyen is a szótagok 75%-a szó eleji, az algoritmusba beépített megszorítás miatt azonban itt nem lehet ütemkezdet.

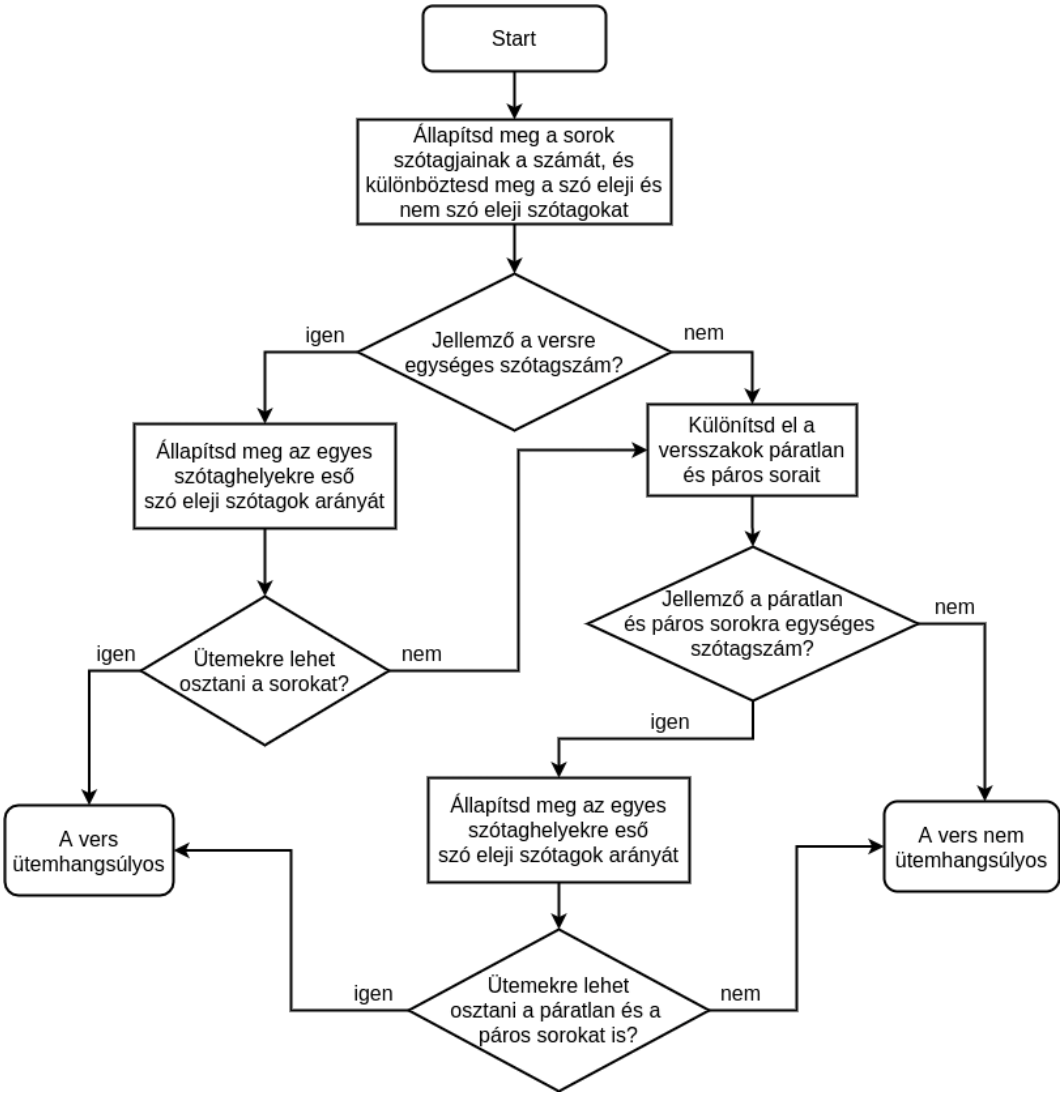
javított kiadás], 288–320 [Debrecen: Debreceni Egyetemi Kiadó, 2012]). Szepes és Szerdahelyi a négy szótagos ütemek mellett az öt és hat szótagos ütemeket a magyar ütemhangsúlyos metrumú versek leggyakoribb ütemtípusainak tartja, de hét, nyolc, kilenc, sőt tíz szótagon felüli ütemek létezése mellett is érvelnek, és nem állapítanak meg felső határt az ütemek maximális szótagszámára vonatkozóan (Szepes és Szerdahelyi, *Verstan,* 365–371). A program futtatásánál alkalmazott hat szótagos ütemhatár mögött elsősorban praktikus szempontok állnak. Bár én sem tartom kizártnak a hat szótagnál hosszabb ütemek létezését, az ütemhangsúlyos metrumú versek azonosítását minden bizonnyal félrevinné, ha például a program hét vagy nyolc szótagos ütemekből álló ütemhangsúlyos versekként elemezné a hét vagy nyolc szótagú sorokból álló, sormetszetet nem tartalmazó összes verset.

 DOI: 10.31400/dh-hun.2021.4.2361

Ütemhangsúlyos verselést csak azokban a versekben lehet megállapítani, amelyekben a sorok szótagszáma valamilyen szabályosságot mutat, hiszen azonos ütembeosztású sorokról csak azonos szótagszámú sorok esetében beszélhetünk. Ennek folytán az ütemhangsúlyos metrumot elemző algoritmus a szó eleji és nem szó eleji szótagok meghatározása mellett azt is megvizsgálja, hogy a vers esetében megadható-e olyan szótagszám, amely általánosan jellemzi a vers sorait. Az algoritmust úgy alakítottam ki, hogy a program futtatásának kezdetén a felhasználó megadhatja, hogy a sorok minimum hány százalékának kell azonos szótagszámúnak lennie ahhoz, hogy ezt a szótagszámot a program a vers általános szótagszámának tekintse. Az algoritmus jelenlegi implementációjában ez mindig ugyanaz a 0 és 1 közötti érték, mint az ütemkezdetek megállapításához megadandó, szó eleji szótagokra vonatkozó minimum arány. Ahogyan arra már többször is utaltam, 0,75 értékkel futtattam a programot, vagyis a verssorok minimum 75%-ának kell azonos szótagszámúnak lennie ahhoz, hogy a program az ezen sorokra jellemző szótagszámot a vers alapszótagszámának tekintse, és e szótagszám alapján elvégezze az ütemhangsúlyos metrum elemzését a fent jelzett módon. Az algoritmus a megállapított alapszótagszámnál hosszabb vagy rövidebb sorok szó eleji szótagjait nem számolja bele az egyes szótaghelyekre eső szó eleji szótagok összegébe, ugyanakkor ezek a sorok is beleszámítanak a vers összes sorába az egyes szótaghelyekre eső szó eleji szótagok arányainak a számítása során. Ez tehát azt jelenti, hogy az alapszótagszámnál hosszabb vagy rövidebb sorokat az algoritmus úgy tekinti, mintha a mátrixban lenne egy csupa nulla karakterből álló sor.

Amennyiben nem jellemző a versre valamilyen általános szótagszám, vagy bár megállapítható általános szótagszám, de ütemhangsúlyosságot nem sikerült megállapítania az algoritmusnak, akkor az algoritmus különválasztja a versszakok páratlan és páros sorait, és megvizsgálja, hogy a szétválasztott sorokra külön-külön jellemző-e valamilyen általános, a sorok előzetesen megadott arányára – az általam használt beállításban a sorok minimum 75%-ára – jellemző szótagszám, illetve valamilyen ütemosztás. Ezzel az eljárással az abab szerkezetű ütemhangsúlyos metrumokat lehet felismertetni, azaz azokat az eseteket, amelyekben a versszakok páratlan és páros sorainak ütemosztása eltérő (pl. abab–abab–abab, aba–aba–aba, aba–abab–ab[29]). Ha a program a szétválasztott sorok esetében sem tudott általános szótagszámot, majd ütemhangsúlyosságot megállapítani, akkor a verset nem ütemhangsúlyosként kategorizálja. Az ütemhangsúlyos verselést elemző algoritmus főbb lépéseit az *1. ábra* mutatja be.

[29] A versszakoknak nem muszáj ugyanannyi sorból állniuk ahhoz, hogy az algoritmus soronként váltakozó ütemosztást állapítson meg.



1. ábra. Az ütemhangsúlyos verselést felismerő algoritmus főbb lépései

Az, hogy a program egy vers esetében nem tud megállapítani ütemhangsúlyos metrumot, nem feltétlenül jelenti azt, hogy a versre valóban nem jellemző valamilyen ütemhangsúlyos metrum, hiszen léteznek a fentieknél bonyolultabb szerkezetű ütemhangsúlyos versek is, amelyek felismerésére a bemutatott algoritmus nem képes.

6. Az ütemhangsúlyos metrumot elemző algoritmus alkalmazása tizenegy költő versein

A fent bemutatott, az ütemhangsúlyos metrum felismertetésére kitalált algoritmust úgyszintén implementáltam Python programnyelvben,[30] és lefuttattam ugyanazon a tesztkorpuszon, mint amelyen az időmértékes metrum felismerését elvégző programot. A *7. táblázat* mutatja be a program futtatásával kapott előfordulási adatokat.

[30] A kód a tanulmány mellékleteként letölthető: `https://doi.org/10.31400/dh-hun.2021.4.2361`.

 DOI: 10.31400/dh-hun.2021.4.2361

A második oszlopban szerepel a tesztkorpusz egyes szerzőihez tartozó összes vers száma, a harmadik oszlopban pedig a program által ütemhangsúlyosként felismert versek száma. Az utolsó oszlop mutatja be az ütemhangsúlyos versek arányát az adott szerző összes verséhez viszonyítva.

7. **táblázat**. Ütemhangsúlyos versek gyakorisága (megegyező szótagszámú sorok és ütemkezdetre eső szó eleji szótagok minimális aránya: 0,75; ütemek maximális szótagszáma: 6)

Szerző	Összes vers	Ütemhangsúlyos	Arány
Tompa Mihály	491	203	0,41
Arany János	417	217	0,52
Petőfi Sándor	839	405	0,48
Gyulai Pál	156	73	0,47
Vajda János	199	58	0,29
Reviczky Gyula	335	89	0,27
Komjáthy Jenő	246	45	0,18
Ady Endre	1116	238	0,21
Babits Mihály	514	114	0,22
Kosztolányi Dezső	630	82	0,13
József Attila	599	148	0,25

A program futtatásával kapott előfordulási adatok megfelelnek az irodalomtörténeti tudásunknak, miszerint a 19. század közepének népies, „magyaros” megszólalásmódjaival is kísérletező költői (Tompa, Arany, Petőfi, Gyulai) jóval nagyobb mértékben használtak ütemhangsúlyos metrumokat, mint az őket követő költők.

Az ütemhangsúlyosságot elemző program nem csupán azt mondja meg egy versről, hogy az ütemhangsúlyos-e vagy sem, hanem azt is megállapítja, hogy pontosan milyen ütemhangsúlyos metrumot követ az adott vers (azzal a már említett megszorítással, hogy a program csak aaaa, illetve abab típusú ütemhangsúlyos metrumokat képes felismerni). Ennek köszönhetően megvizsgálható az is, hogy egy adott szerző milyen típusú ütemhangsúlyos metrumokat használt a leggyakrabban. A *8. táblázat*ban Petőfi leggyakoribb tíz ütemhangsúlyos metrumát tüntettem fel a program elemzése alapján. A táblázatban látható két leggyakoribb metrum a kétütemű tízes és a felező nyolcas. Az előbbiben az első ütem négy, a második pedig hat szótagból áll. A táblázat negyedik helyén szereplő metrumnál csupán egy szám szerepel: 6. Ezek olyan versek, amelyekben a vers sorai hat szótagosak, a sorokon belül pedig nincsen ütemosztás. Mivel a programot hat szótagos maximális ütemhosszúsággal futtattam, a program a megegyező szótagszámú, hat szótagnál nem hosszabb verssorokból álló verseket akkor is ütemhangsúlyosként elemzi, ha a sorokban nincsen ütemosztás. Megemlítendő még az ötödik helyen álló 4 | 6 és 4 | 5 típusú metrum, ahol a kétféle kimenet arra utal, hogy a versszakok páratlan és páros sorai eltérő ütemosztást követnek. A páratlan sorok kétütemű tízesek, négy és hat szótagból álló ütemekkel, a páros sorok pedig kétütemű kilencesek, négy és öt szótagos ütemekkel.

8. táblázat. Petőfi leggyakoribb ütemhangsúlyos metrumai (megegyező szótagszámú sorok és ütemkezdetre eső szó eleji szótagok minimális aránya: 0,75; ütemek maximális szótagszáma: 6)

	Metrum	Előfordulás
1.	4 | 6	39
2.	4 | 4	38
3.	6 | 6	36
4.	6	35
5.	4 | 6 és 4 | 5	24
6.	5 | 5	20
7.	4 | 4 | 3	18
8.	4 | 4 és 4 | 3	13
9.	4 | 4 és 3	8
10.	4 | 4 és 6	6

7. A szimultán metrumú versek gépi felismertetése

A két algoritmus együttes alkalmazásával lehetőségünk van a szimultán metrumú versek gépi felismertetésére is. A szimultán versmetrum megközelítésében Szepes és Szerdahelyi értelmezését követem, amely szerint azokban az esetekben beszélhetünk szimultaneitásról, „amikor ugyanazok a költemények egyszerre felelnek meg valamilyen időmértékes és ütemhangsúlyos képletnek, tehát kétféle ritmusúak”.[31] A szerzők hangsúlyozzák, hogy „a szimultán versek szövegeiben nem egy – a hangsúlyos és időmértékes metrumok összegzéséből adódó – sajátos metrum érvényesül, hanem e szövegekben a hangsúlyos és időmértékes nyomatékrend egymás mellett áll, s így nem összegezhető.”[32] Az időmértékes és az ütemhangsúlyos metrumot elemző két algoritmus együttes futtatásával lehetőség van azon versek gépi felismertetésére, amelyekben egyszerre érvényesül az időmértékes metrum és az ütemhangsúlyos metrum, vagyis amelyek Szepes és Szerdahelyi definíciója alapján szimultán metrumú verseknek tekinthetők.

A *9. táblázat* a tesztkorpusz szerzői esetében bemutatja az ütemhangsúlyosként felismert versek számát, a szimultánként (ütemhangsúlyosként és időmértékesként) felismert versek számát, valamint ez utóbbiak arányát az adott szerző összes verséhez képest. A táblázat utolsó két oszlopa azt mutatja be, hogy a szimultán versek közül a program elemzése alapján mennyi a jambikus és a trochaikus időmértékes metrumot érvényesítő vers. A program futtatása során az időmértékes metrumokra vonatkozó küszöböt 0,5 értékben adtam meg.

[31] Szepes és Szerdahelyi, *Verstan,* 510.

[32] Uo., 513. Szepes és Szerdahelyi ugyanakkor a szimultán ritmusú versek mellett megkülönbözteti a kevert ritmusú és az ötvözött ritmusú verseket is. Az előbbi esetében a versben ütemhangsúlyos és időmértékes részek váltogatják egymást soronként vagy néhány soros szakaszonként, az utóbbi esetben pedig a versnek csak egy metruma van, ez azonban az időmértékes és ütemhangsúlyos rendszer elveinek az összegyúrásából jön létre. Uo., 513–517. Ezek gépi azonosítása nem volt célom.

 DOI: 10.31400/dh-hun.2021.4.2361

9. táblázat. Szimultán metrumú versek gyakorisága (megegyező szótagszámú sorok és ütemkezdetre eső szó eleji szótagok minimális aránya: 0,75; ütemek maximális szótagszáma: 6; időmértékes metrumok küszöbértéke: 0,5)

Szerző	Összes vers	Ütemhangsúlyos	Szimultán	Arány	Szimultán jambikus	Szimultán trochaikus
Tompa Mihály	491	203	165	0,34	80	84
Arany János	417	217	148	0,35	31	116
Petőfi Sándor	839	405	261	0,31	102	158
Gyulai Pál	156	73	51	0,33	10	41
Vajda János	199	58	42	0,21	14	28
Reviczky Gyula	335	89	86	0,26	48	38
Komjáthy Jenő	246	45	32	0,13	19	12
Ady Endre	1116	238	122	0,11	49	73
Babits Mihály	514	114	73	0,14	47	25
Kosztolányi Dezső	630	82	69	0,11	62	7
József Attila	599	148	90	0.15	61	28

A táblázatból látható, hogy a szimultán metrumú versek a 19. század első harmadában született, az ütemhangsúlyos metrumot amúgy is nagyobb mértékben használó költők esetében jelennek meg a legnagyobb arányban (Tompa, Arany, Petőfi, Gyulai). Ugyancsak feltűnő, hogy annak ellenére, hogy Aranyon és Gyulain kívül az összes többi vizsgált szerző nagyobb arányban használta a jambikus metrumot, mint a trochaikus metrumot (lásd a 2. táblázatban), a szimultán versek esetében az említett két szerző mellett Tompánál, Petőfinél, Vajdánál és Adynál is nagyobb számban jelennek meg a trochaikus metrumú versek, mint a jambikusak. Ez összefügghet azzal az elképzeléssel, miszerint az ütemhangsúlyos metrum természeténél fogva vonzza a trochaikus lüktetést.[33]

8. Összegzés

A tanulmány magyar nyelvű versek metrumának gépi felismertetésére mutatott be két szabályalapú algoritmust. Az első algoritmus célja a versek besorolása a négy nagy időmértékes metrumkategória, az anapesztikus, a daktilikus, a jambikus vagy a trochaikus metrumok valamelyikébe. Az algoritmus először a verssorok, majd a versek szabályossági pontszámát számítja ki mind a négy metrumra vonatkozóan. A szabályossági pontszám mind a sorok, mind a versek esetében egy 0 és 1 közötti szám, amely azt jelzi, hogy a verssor, illetve a vers milyen mértékben valósítja meg az adott metrum elvont ritmusképletét. A versekhez rendelt szabályossági pontszám lehetővé teszi, hogy a kutatói elemzésből kizárjuk azokat verseket, amelyek ritmusa a felismert metrum tekintetében nem ér el egy meghatározott szabályossági szintet. A bemutatott másik algoritmus célja az aaaa, illetve abab szerkezetű ütemhangsúlyos

[33] Szepes és Szerdahelyi, *Verstan,* 267–269; Horváth, *Rendszeres magyar verstan,* 128–133; Vargyas, *Magyar vers – magyar nyelv,* 148–150.

metrumú versek felismerése, azaz az algoritmus azon verseknél állapít meg ütemhangsúlyos metrumot, amelyekben minden sorra ugyanaz az ütemosztás jellemző, vagy pedig két típusú, a versszakok páratlan és páros soraira jellemző ütemosztás váltogatja egymást. Ennek az algoritmusnak az alapja az egyes szótaghelyekre eső szó eleji szótagok arányának a kiszámítása.

A két algoritmust Python nyelvben implementáltam, és az ELTE Verskorpuszból kiszedett tesztkorpuszon futtattam le, amely tizenegy, a 19. század közepén és második felében, valamint a 20. század első felében alkotó költő verseit tartalmazta. A korpuszvizsgálatokban többek között bemutattam, hogy az időmértékes metrum gépi elemzése során a küszöbérték módosítása miképpen befolyásolja a kapott előfordulási adatokat, hogy a kimenetként kapott szabályossági pontszámok középértékei hogyan használhatóak fel a versritmusra irányuló kutatásokban, illetve hogy az ütemhangsúlyos metrumot elemző algoritmus hogyan használható az ütemhangsúlyos metrumtípusok gyakorisági listájának az előállítására. Azt is bemutattam, hogy a két algoritmus együttes használata miképpen járulhat hozzá a szimultán metrumú versek vizsgálatához.

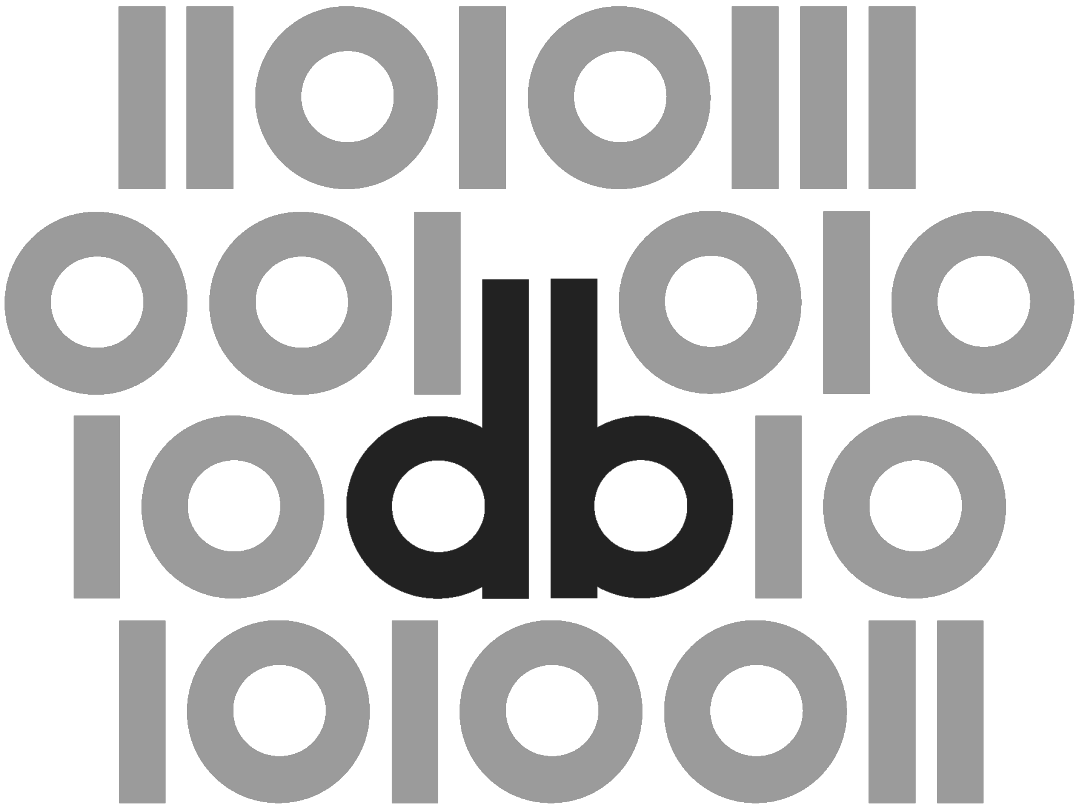
Bár a két algoritmus csupán az egyszerűbb metrumok felismerésére képes, és az elemzés nem közelíti meg az emberi elemzés pontosságát és árnyaltságát, arra alkalmasnak tűnnek, hogy kvantitatív vizsgálatokhoz használható adatokat állítsanak elő. A metrumra vonatkozó, automatizáltan előállított kvantitatív adatok persze nem feltétlenül az olyan nagymértékben kanonikus szerzők esetében hozhatnak új eredményeket, mint akiket a tanulmány tesztkorpuszában is szerepeltettem, hiszen e szerzők nagy részének a versritmusával sokat foglalkozott az irodalomtörténet. A kanonikus szerzők szerepeltetése a tanulmányban elsősorban azt a célt szolgálta, hogy teszteljem, illetve bemutassam a kialakított két algoritmus használhatóságát: mivel a 19. és 20. század költészetét reprezentáló, a versmetrum kézi annotációit tartalmazó magyar verskorpusz nem létezik, az algoritmusok használhatóságát úgy tudtam valamelyest feltérképezni, hogy megnéztem, a kanonikus szerzőkön lefuttatott program kimenete összhangban van-e azzal, amit ezen költők versritmusa kapcsán lehetett tudni. Ugyanakkor a nem vagy kevésbé kanonikus költészet vizsgálatához a metrum gépi elemzése már jóval produktívabb módon járulhat hozzá, hiszen ez olyan, a költészeti kánonnál sokszorosan nagyobb szövegkorpusz, amelynek az elolvasására, illetve a szoros olvasás technikáin alapuló elemzésére nem igazán van lehetőség. A bemutatott két algoritmus természetesen több ponton továbbfejleszthető vagy akár teljesen újragondolható. A magyar időmértékes és ütemhangsúlyos metrum gépi felismertetésére minden bizonnyal számos további, más elveket és számításokat alkalmazó eljárás is kitalálható.

 DOI: 10.31400/dh-hun.2021.4.2361

Two Computational Methods for Detecting Meter in Hungarian Poetry

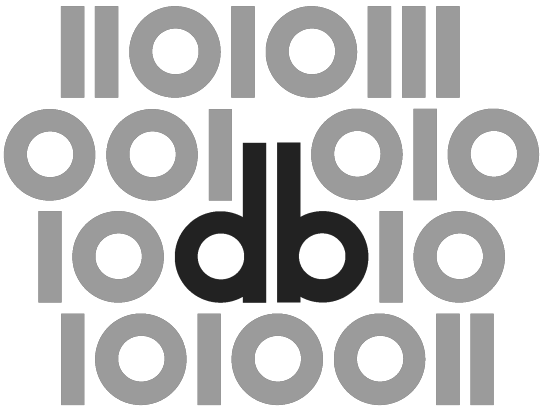
This paper presents two methods for the machine recognition of meter in Hungarian poetry, one for quantitative and another for qualitative meter. The algorithm analysing the former classifies the poems as dactylic, anapestic, iambic or trochaic, and defines the degree of regularity of the rhythm in the poems. The algorithm detecting the Hungarian qualitative meter returns the number of syllables of each beat. The paper also demonstrates the applicability of the two algorithms by using a test corpus containing the poems of eleven authors written between 1838 and 1941.

Keywords:
quantitative meter, qualitative meter, machine analysis, corpus research, ELTE Poetry Corpus



Digitális Bölcsészet
2021., negyedik szám

<DIGITÁLIS BÖLCSÉSZET>



4 (2021)

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.2021.4

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.

Ez a mű a Creative Commons *Nevezd meg! – Ne add el! – Így add tovább! 2.5 Magyarország Licenc* (`http://creativecommons.org/licenses/by-nc-sa/2.5/hu/`) feltételeinek megfelelően felhasználható.

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

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

<KRITIKA>

Seláf Levente 0000-0002-6052-9841

ELTE BTK Régi Magyar Irodalom Tanszék

selaf.levente@btk.elte.hu

Petr Plecháč. *Versification and Authorship Attribution*. Prague: Institute of Czech Literature–Karolinum Press, 2021. ISBN 9788024648712. 98 oldal. https://doi.org/10.14712/9788024648903*

E rövidke, vékony, kevesebb mint 100 oldalas kötet témája egy új kutatási módszer, a verselési jellemzők felhasználása versben írt szövegek szerzői attribúciós problémáinak megoldásához.

Az elmúlt bő száz évben a szerzőség megállapításának, ellenőrzésének módszere nagyon sokat változott. Alapvetően stilometriai, a szerző szókincsét, nyelvhasználatát vizsgáló eljárásokat használnak a kétes hitelű, kérdéses szerzőségű, esetleg hamisított szövegek kiszűrésére és pontos szerzőjük azonosítására. Plecháč könyve azt mutatja be, hogy a költői szövegek ritmikai és verselési jellemzői ugyanolyan jól használhatók az ilyen jellegű feladatoknál, mint a szókincs elemei.

A négy fejezet közül az első a kvantitatív alapú szerzőségi vizsgálatok történetét vázolja röviden. Általában a stilometriai elemzések módszertanának változásait tekinti át, a matematikai és statisztikai alapú elemzések első alkalmazásaitól egészen a 21. századi módszerekig: bemutatja, hogyan finomodott a vizsgálat a szókincs általános elemzésétől a funkciószavak variációinak összevetéséig, majd azt, hogy a különféle statisztikai módszerek hogyan tudták egyre pontosabban meghatározni egy mű szerzőségét. Egészen a legfrissebb, jelenkori módszerekig jut el, Burrows deltáján (2002, 2003) és annak módosított változatain és a különféle vektorgeometriai technikákon keresztül egészen a szupport vektor gépekig (SVM), melyek két szöveg hasonlóságát (vagy több szöveg hasonlóságán keresztül egy adott szerző stílusát) ábrázolni tudják. A fejezet utolsó része a verselési tulajdonságokon alapuló attribúciós kísérletek történetét vázolja röviden, (melyek egyébként, a Shakespeare-kutatást leszámítva, egészen későn, csak az elmúlt évtizedekben születtek meg) bemutatva, hogy míg a szógyakoriság statisztikai vizsgálata rengeteget fejlődött, finomodott az elmúlt évtizedekben, ez a módszertani áttörés a verselési tulajdonságok vizsgálatában nem következett be.

A második, rövid fejezet egy verstani áttekintés, amely rögzíti azokat az alapfogalmakat, paramétereket, melyeknek segítségével a verselés kvantitatív jellemzői vizsgálhatóvá válnak. Három kategóriát vesz fel, ezek a ritmus, a rím és az eufónia. A ritmus vizsgálatának egyik módja a ritmikai profil létrehozása: ez a hangsúlyos verselésű szövegekben az egyes metrikai pozíciókra eső hangsúlyok gyakoriságát vizsgálja. Hátránya, hogy nem tud mit kezdeni azzal, ha egy pozícióba két szótag is jut, ami nem ritka például az angolban, vagy a ritmusképleten felüli, esetleg a hiányzó szótagokkal.

* A publikáció az Innovációs és Technológiai Minisztérium Nemzeti Kutatási Fejlesztési és Innovációs Alapból nyújtott támogatásával, a K 135631 számú, *A régi magyar költészet számítógépes metrikai és stilometriai vizsgálata* című pályázati program keretében készült.

A ritmustípus *(rhythmic type)* ezzel ellentétben lekottázza az adott sor ritmusát, nem az absztrakt képlettel veti össze, ennek pedig az a hátránya, hogy az adatok nehezen lesznek mintába rendezhetők. Ezekhez képest Plecháč a ritmikus n-gramok vizsgálatát javasolja: ez a sort kisebb egységekre osztja, és ezek gyakoriságát vizsgálja, a két és három hangsúlytalan szótagból álló elemtől a csak egy hangsúlyosat tartalmazóig. A cseh és a spanyol költészetben ez tűnik a legjobbnak, míg a német, hangsúlyos verselésben a második módszert tartja célravezetőnek. A rímek vizsgálatában hét különböző jegyet vesz figyelembe két egymásra rímelő szó viszonyának meghatározásához: a morfológiai jegyeket, a rímszavak szótagszámát, az utolsó hangsúlyos szótag utáni szótagok számát, az utolsó szótag mássalhangzózárlatát, az utolsó magánhangzót, és gyenge rímeknél az utolsó magánhangzó előtti mássalhangzót és az utolsó előtti (hangsúlyos) magánhangzót. Az eufónia pedig bizonyos hangok gyakoriságát jelenti – ez gyakorlatilag az alliteráció vizsgálatának kiterjesztése teljes művekre.

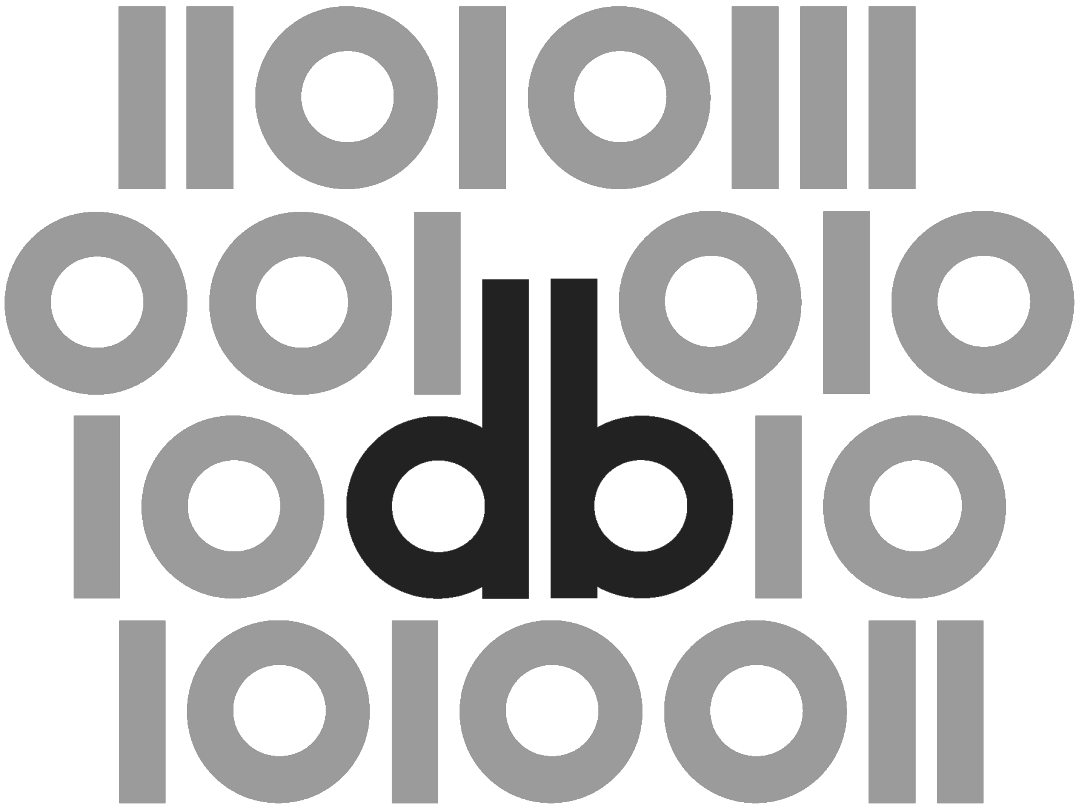
A következő fejezet a sztichometriai módszerek kísérleti vizsgálatát tartalmazza három nagy korpuszon, a cseh, német és spanyol költészetből, ezek a következők: *The Corpus of Czech Verse, Metricalizer – the Corpus of German Verse, Corpus de sonetos del Siglo de Oro.* A kísérlet célja annak megállapítása, hogy e verseknél a gépi tanulással mennyire felismertethetők egyes szerzők szövegei stilometriai, illetve sztichometriai adatok alapján. Ez a vizsgálat sok érdekes, bár megjósolható részeredménnyel is járt, így például a nagyon termékeny szerzők esetében, akik változatos stílusban írtak, a szövegeik felismerése kevésbé hatékonyan sikerült az algoritmusnak. Izgalmas, hogy melyik tulajdonságok voltak a legjellemzőbbek egy-egy adott alkorpuszban. A cseh verseknél például a rímek fonetikai összetétele, a német versekben a morfológiai jegyek, míg a spanyol szövegek valahol a kettő között foglaltak helyet. A ritmikus n-gramok közül a négyeleműek jellemezték a legjobban a versek szerzőit az ebből a szempontból releváns cseh és spanyol anyagban. A fejezet konklúziója megállapítja, hogy a verselési jegyek alapján nagyon jó arányban felismerhetők a szerzők, sőt, néha ezek felülmúlják a lexikai elemekből kinyerhető jellemzők pontosságát. A két szempont kombinációja pedig nagyon biztos eredményt ad az szerzőiség szempontjából.

A kötet utolsó része két esettanulmányt tartalmaz. Az egyik egy klasszikus attribúciós probléma tisztázása, William Shakespeare és John Fletcher *Két nemes rokon* című drámájában állapítja meg a szerzőpáros tagjainak hozzájárulását a műhöz. Már 1812-ben született elemzés, amely megpróbálta tisztázni, pontosan melyikük melyik jelenetet, felvonást írta, és azóta is sok, eltérő metodikájú és szempontú elemzés született. Plecháč tanulmánya kombinálja a szókincs és a metrikai jellemzők gyakoriságának vizsgálatát, és így nagyon pontos eredményt kap, főleg a csúszó vagy gördülő attribúció *(rolling attribution)* módszerével (itt a szöveget százsoros részekben vizsgálja, de ötsoronként előrehaladva (1–100, 5–105, 10–110 stb.), ami nagyon jó százalékban kiszűri a hibákat. Végeredményben a tanulmány megerősíti az elmúlt évtizedek filológiai álláspontját a két szerző által írt részekről. Nagyon meggyőző bizonyítása ez annak, hogy a verselés ugyanolyan egyedi, szerzőspecifikus jellemző, mint a szóhasználat. A verselés szempontjából egyébként az adott esetben a legnagyobb különbség az volt, hogy Shakespeare sokkal több hangsúlyos végződésű sort írt, míg Fletchernél többségben voltak a hangsúlytalan sorvégek.

 DOI 10.31400/dh-hun.2021.4.3503

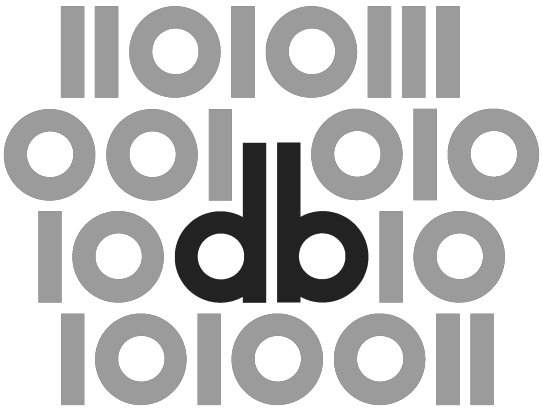
A kötet utolsó tanulmányának társszerzője Artjoms Šeļa, a tartui egyetem kutatója. Ebben a vizsgálatban egy hamisított szövegkorpusz beazonosítása a cél. 1978-ban A. A. Iljusin publikálta G. S. Batyenkov költő, katonatiszt és dekabrista lázadó összes költeményét, köztük számos olyan verssel, amelyeket állítása szerint Batyenkov hosszú fogsága és szibériai száműzetése során írt. E versek hitelességét sokan megkérdőjelezték, a legalaposabb vizsgálatnak M. I. Sapir vetette őket alá, de számítógép használata nélkül. Sapir konklúziója a vizsgálatból az volt, hogy lehetetlen stilisztikai és verselési tulajdonságok alapján meghatározni egy szerzőt. Ezzel a túlzó általánosítással száll szembe a tanulmány, mely meggyőzően mutatja ki Batyenkov és kortársai verseinek elemzésével, és ezeket összevetve a Pszeudo-Batyenkov-féle korpusszal, sőt, Iljusin saját verseivel, hogy igenis, nagy pontossággal tisztázni lehet, hogy melyek az Iljusin által írt, hamisított versek.

A rövid, de bravúros könyv tehát több szempontból is bemutatja vizsgálati módszere érvényességét. A magyar filológiában nincs sok olyan kérdéses szerzőségű verses szöveg, amelyek ilyen komplex, stilo- és sztichometriai elemzése komoly meglepetéseket tartogathat. De azért vannak esetek, amikor érdemes ezzel megpróbálkozni, például az *Eurialus és Lucretia* című széphistória esetében. Emellett a módszer pontosságát magyar versanyagon olyan, már bizonyított hamisítások korpuszának elemzésével lehetne ellenőrizni, mint Thaly Kálmán kuruc versei. Izgalmas lehet a többféle költői személyiséget és stílust kidolgozó Kovács András Ferenc verseinek összevetése is, hogy azonosítsuk, mely jegyek jellemzik az alteregóit, s melyek, megcáfolhatatlanul, a közös szerzőt.



Digitális Bölcsészet
2021., negyedik szám

<DIGITÁLIS BÖLCSÉSZET>



4 (2021)

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.2021.4
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.

(cc) BY-NC-SA

Ez a mű a Creative Commons *Nevezd meg! – Ne add el! – Így add tovább! 2.5 Magyarország Licenc* (`http://creativecommons.org/licenses/by-nc-sa/2.5/hu/`) feltételeinek megfelelően felhasználható.

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

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

<KRITIKA>

Bátorfy Attila 0000-0003-2447-1010

ELTE BTK Művészetelméleti és Médiakutatási Intézet,

Média és Kommunikáció Tanszék

batorfy.attila@btk.elte.hu

Kocsis Károly (főszerk.). *Magyarország Nemzeti Atlasza – Társadalom (Interaktív verzió)*. Budapest: CSFK Földrajztudományi Intézet, 2021. `https://emna.hu/`

2021. szeptember 16-án megjelent a nagy múltú *Magyarország Nemzeti Atlasza* harmadik, társadalomról szóló kötete Kocsis Károly főszerkesztésében. Ugyanezen a napon publikálta a Csillagászati és Földtudományi Kutatóközpont Földrajztudományi Intézete a kötet interaktív verzióját. Ebben a recenzióban kizárólag a hivatkozott interaktív verzióval foglalkozom.

Leírás

A honlap nyitóoldala szürke, monokróm térképet mutat Magyarországot középre helyezve. A bal oldali, összecsukható menüsorból választhatjuk ki, milyen térképrétegeket szeretnénk megjeleníteni az alaptérképen. A térképrétegek főcsoportokra vannak osztva. Összesen tizenkilenc ilyen kategóriát találunk, amelyek további alkategóriákra bomlanak, közel négyszázra. Kattintásra jobbról beúszik az adott térképréteg címe, a térkép szín- és jelmagyarázata, néhány hasznos, a réteg adataiból számított statisztika, illetve némi beállítási lehetőség. Ha akarjuk, mi magunk is keverhetjük a színeket, beállíthatjuk, hogy milyen lépték vagy variáció szerint színezze az algoritmus a térképeket. Ez hasznos, mivel a különféle léptékek és színek bizonyos mintázatokat jobban láthatóvá tesznek. Ezen kívül a jobboldali menüsorból megkapjuk a térkép teljes leírását, az adatokat szűrhetjük, a földrajzi nevek között kereshetünk, hangsúlyossá tehetjük az adminisztratív határokat egészen települési szintig. Emellett a térképeket elmenthetjük, rajzolhatunk rájuk. További funkció, hogy a monokróm szürke alaptérképet kicserélhetjük műholdképre, egy előre generált úgynevezett „színes" térképre, de választhatjuk azt is, hogy ne jelenjen meg a réteg alatt semmi. Ugyan fel lehetett volna kínálni egy sötét tónusú alaptérképet, már csak azért is, mert vannak színek, amelyek jobban érvényesülnek a sötét háttéren, és adatábrázolási szempontból éppen az alapként kínált szürke tónus az egyik legproblémásabb, mert kevés színt bír el. Hasznos, hogy emellett a domborzati modellt *(hillshade)* is meg tudjuk jeleníteni, de társadalmi adatoknál fontos lenne látni a ténylegesen lakott területeket. A számos adatréteg emellett nem pusztán a jelenről szól, hanem történeti távlatot is nyújt a Kárpát-medence, a történelmi Magyarország és az Osztrák-Magyar Monarchia viszonyrendszerében. Mindez önmagában jelentős előrelépés ahhoz képest, hogy társadalmi-gazdasági térinformatikai adat, statisztika ekkora mennyiségben korábban nem volt elérhető interaktív formában a nagyközönség számára.

Hiányok

A hazai és nemzetközi akadémiai-kutatói szektor hasonló megoldásai arra az elgondolásra épülnek, hogy a felhasználó szeret játszani, felfedezni, rétegeket váltani, színeket keverni, és a súlyos adatmennyiségtől önmagában úgy érezheti, hogy olyan szolgáltatást kap, amely minden igényét kielégíti. Bár számosan lehetnek olyanok, akik hajlandók nagyon sok időt eltölteni a böngészéssel, az ilyen vizuális adatfelfedező szolgáltatások (*data explorer*ek) könnyen elveszíthetik a potenciális felhasználók egy jelentős részét. Itt megmutatkozik az adatelemzés és -ábrázolás két nagy csoportja, az exploratív és az explanatív közötti eltérés. A készítők az előbbi megoldás javára lemondtak a magyarázatokról, az információ és a tudás átadásáról. Az oldalon a felhasználó vizuális reprezentációkat lát, amelyeknek van szakmailag pontos leírása, mégsem derül ki, hogy az adott térkép miért fontos, mit kellene belőle különösen látni, mi az, amiért érdemes rákattintani. A vizuális reprezentáció ritkán állja meg a helyét egymagában, főleg akkor, ha információábrázolásról beszélünk. A legtöbb esetben nem elég megmutatni az adatokat, hanem el is kell mondani, hogy mit látunk, továbbá érdemes vizuális és szöveges fogódzókat adni a felhasználónak. Ezek nélkül az MNA interaktív verziója nem más, mint a nyomtatott kiadvány felszerelése digitális kapcsolókkal, miközben a felhasználók többsége továbbra is sötétben tapogatózik.

A projekt kissé régivágású profiljára szintén jó példa, hogy demográfiai adatok esetében is kizárólag konvencionális felületszínezéses *(choropleth)* térképeket látunk, miközben ez a módszer félrevezető. Talán ezért választották azt a megoldást, hogy az adminisztratív határokon belüli adatokat legtöbb esetben a lakosságra normalizált, százalékos arányok adják. Ez azért problematikus, mert épp azt fedi el, hogy hány főnek a valahány százalékáról van szó. Szerencsésebb lett volna, ha ki lehetne választani, hogy abszolút számot vagy százalékos arányt mutasson a térkép. Megoldás lehetne továbbá az is, ha az alaptérképen fel van tüntetve a települések lakott/beépített területe, akkor csak arra vonatkozóan mutassa az adatokat.

Megfontolandó lett volna az értékarányos jelmódszer.[1] Ez ugyan a települési, járási szinteken nagyon sűrűvé tette volna a térképet, ám ha nem a hagyományos négyzet-háromszög jelekben gondolkodnak, hanem *spike-map*ben (nevezzük ezt tűtérképnek),[2] akkor a vizuális jelek sűrűségéből/fedettségéből adódó értelmezhetőségi problémákat jelentősen csökkenteni lehetett volna. Használhattak volna továbbá alakzaton belüli pontelosztásos módszert,[3] illetve *gridezés*t is.[4] Mindezeknek nem csupán

[1] Az értékarányos jelmódszer általában a geometrikus alakzatok, leggyakrabban a kör, négyzet és háromszög adatalapú méretezése.

[2] A *spike-map* az értékarányos jelmódszer egyik hátrányára, a vizuális sűrűség csökkentésére és az alakzatok fedésének kiküszöbölésére kitalált módszer. A *spike-map* háromszög alakzatot használ, ám az értékek szerint csak a háromszög magassága változik.

[3] Az alakzaton (poligonon) belüli pontelosztás a felületszínezéses térképekben rejlő manipulációra válasz. Az adott alakzaton, általában az adminisztratív határon belül annyi pontot osztunk ilyenkor el, ahány ember tartozik ahhoz az adminisztratív egységhez. Így az egy pont egy ember ábrázolásnál a pontok sűrűsége lesz a vizuális információ és nem a felület színezése. Ez a módszer akkor működik, ha embereket akarunk megmutatni.

[4] A *gridezés* szintén a felületszínezéses térképek egyik, manapság népszerű alternatívája. A *gridezés* vagy négyzetelés, más esetekben kaptár- vagy méhsejtalakzat valamekkora négyzetkilométerre aggregálja az adatokat.

 DOI 10.31400/dh-hun.2021.4.3132

az a funkciója, hogy megoldást kínáljanak a felületszínezésből eredő méret/lakosság, illetve településsűrűségi/domborzati problémára, hanem az is, hogy a felhasználó többféleképpen lássa az adatokat. Sok esetben ugyanis éppen egy másik módszer az, amely segít megérteni olyan összefüggéseket, kiugrásokat, mintázatokat egy térképen, amelyek egyébként nem lennének láthatók. Követendő példák erre a sokszínűségre többek között a svájci vagy a spanyol nemzeti atlaszok digitális verziói.[5]

Felhasználói szempontból a kapcsolókkal, parancsgombokkal teli kezelőfelület sem tekinthető korszerűnek. Még olyan egyszerű helyzetekben is, mint a hasznos adatszűrési funkció, ki kell adnunk az „Alkalmaz" parancsot ahhoz, hogy a *backend* végrehajtsa a szűrést. Ez a helyzet adódhat egy szintén elavult, szerveroldali adatkiszolgálási rezsimből is, miközben *on-the-fly* megoldások[6] egy évtizede léteznek az online térképeknél is. Jó példa erre a *Morphocode data explorer*e, amely szintén különféle paraméterezési lehetőségeket kínál, és a *scrollytelling*[7] megoldással az adatkiszolgálás és a vizualizáció, továbbá az ábrázolás módszere is dinamikusan változik.[8]

Végül nem mehetünk el szó nélkül amellett, hogy a projekt és annak résztvevői a térképeket kiszolgáló térinformatikai adatállományra magántulajdonként tekintenek. A térinformatikai fájlokhoz és azokhoz kapcsolódó vagy az azokkal már eleve összekötött adatokhoz nem lehet hozzáférni WMS/WMTS szerveren keresztül, hogy a felhasználók behúzhassák azokat más szoftverekbe. Az adatokat és a térképállományokat sem lehet letölteni: se adatbázisként, se térinformatikai nyers adatállományként, se az adatokkal eleve összekötött GIS-fájlként.[9] Így a rendelkezésre álló adatokat a projektben részt vevő állami intézményeken és munkatársakon kívül senki nem használhatja, nem ellenőrizheti őket, nem készíthet az adatokból saját feldolgozásokat, vagy médiaképes tartalmakat. Ezek az adatok közkincsek, közpénzből állították őket elő, és a projektben részt vevő egyetemek, kutatóközpontok és állami szervek birtoklási vágyán kívül semmilyen érv nem szól amellett, hogy azokat elzárják a nyilvánosság elől. A láthatóvá tétel nem jelent egyet a hozzáférhetőséggel és a nyilvánossággal. A korábban már említett *Svájc Nemzeti Atlasza* minden egyes térképnél megadja, hogy az azon található rétegek (domborzati raszter, *digital elevation model,* adminisztratív határok, továbbá a konkrét társadalmi vagy egyéb adat) térinformatikai adatai honnan tölthetők le hivatalos és nyilvános svájci vagy globális forrásból, adatbázisból. A spanyolok atlasza a megjelenített adat CSV és XLSX exportálását teszi lehetővé. Bogotá nyíltadatoldala pedig nem csupán a *data explorer* jelleget köti össze a kiválasztott

[5] Institut für Kartografie und Geoinformation Zürich: *Atlas der Schweiz,* hozzáférés: 2022.03.07, `https://www.atlasderschweiz.ch/`; Instituto Geográfico Nacional, Centro Nacional de Información Geográfica: *ANE – Atlas Nacional Interactivo de España,* hozzáférés: 2022.03.07, `https://interactivo-atlasnacional.ign.es/`.

[6] *On-the-fly* megoldásoknak nevezzük azokat az interakciókat, amikor nem kell külön parancsot adni a művelet futtatásához, például az alkalmazás érzékeli, hogy görgetés, zoomolás történt, vagy éppen valamilyen beállítás megváltozott.

[7] *Scrollytelling* megoldásoknak nevezzük azokat az online interakciókat, amikor egy vizualizáció/animáció az oldalon való lefelé görgetéssel együtt változik.

[8] *Morphocode Explorer,* hozzáférés: 2022.03.07, `https://explorer.morphocode.com/`.

[9] GIS: geographic information system. A térképészeti digitális fájlokat nevezzük összefoglaló néven GIS-fájloknak. Ezek lehetnek vektorosak (leggyakrabban SHP, KML, GEOJSON kiterjesztésűek), vagy raszteres képfájlok (pl. GEOTIFF). A GIS-fájlokat össze lehet kötni adatbázisokkal, és azokat új GIS-fájlként el is lehet menteni.

területi egység GIS-adatainak letölthetőségével, hanem úgy ahogy van, letölthetővé teszi a teljes, fél- vagy egyévente frissülő referenciatérkép hatalmas térinformatikai állományát mindenki számára.[10] Léteznek hazai, szintén az akadémiai közegből érkező gyakorlatok is, például a *GISta Hungarorum* projekt,[11] Ignácz Károly *Budapest választ* online is elérhető monográfiája[12] vagy a Magyar Bányászati és Földtani Szolgálat WMTS-szerver csatlakozási lehetősége.[13]

Konklúzió

Az előzményeket ismerve érthető, ha a projektben résztvevő szakemberek, kutatók és akadémikusok úgy érzik, hogy az MNA interaktívvá tétele mérföldkő, hiszen korábban nem volt ilyen szolgáltatás. Ez valóban fontos lépés, de nem elég jelentős. Az oldal azonban technológiailag, felhasználói szempontból, az adatkiszolgálás terén, az interakció fogalmi felfogásában és az ábrázolás-módszertanban is mintegy tíz évvel le van maradva az aktuális (ám nem feltétlenül a nemzeti atlaszgyártási) trendektől. A projektre az Eötvös Loránd Kutatási Hálózatnak a Nemzeti Összetartozás Évéről szóló, 2020-as évi februári kormányhatározatban megítélt 280,7 millió forintnak egy részét lehetett volna előtanulmányokra fordítani, hiszen egészen biztosan találtak volna jó és korszerű gyakorlatokat. Gondolhatunk itt a nagyrészt nyílt forráskódú Kepler.gl-re, vagy nagytestvérére, az Unfolded.ai-ra, amelyek a felhasználói felület szempontjából is kínáltak volna ötleteket, és tökéletesen alkalmasak nemcsak *mashup*ok készítésére, külső adatintegrációra, tárolásra, külső dizájnok behúzására, de akár még együttműködésre is.

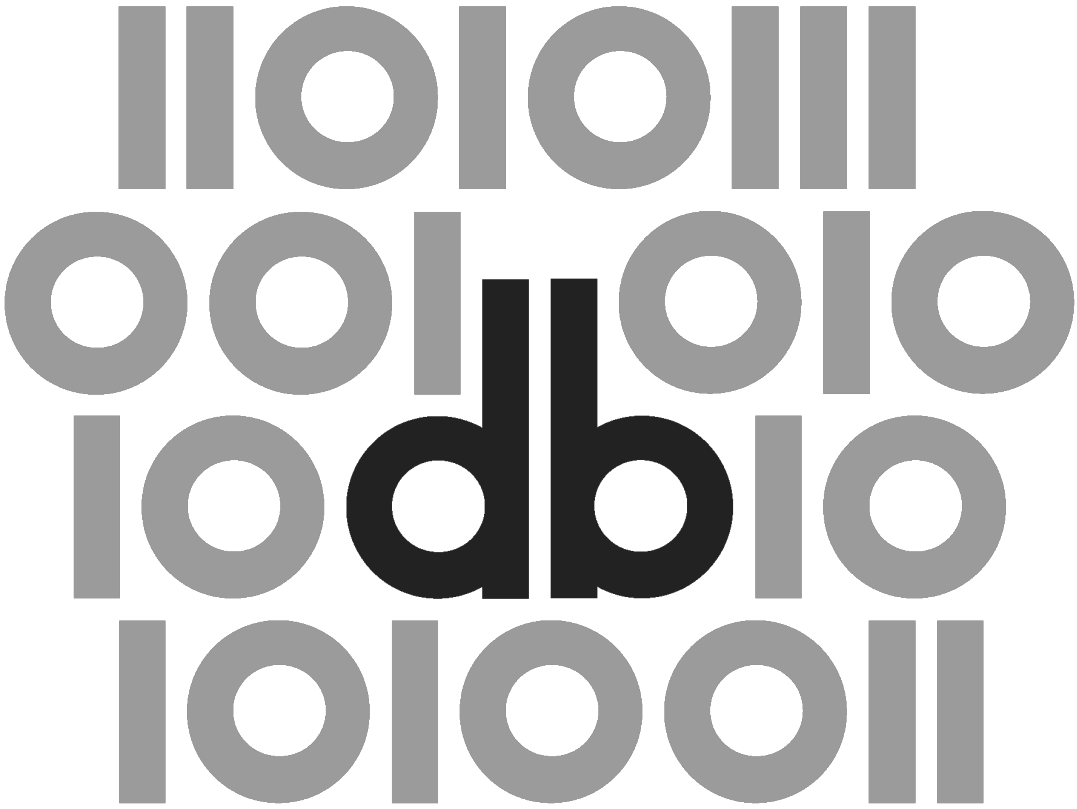
[10] Alcaldía Mayor de Bogotá D.C. – Municipio de Bogotá: *Datos Abiertos de Bogotá – Mapa de Referencia,* hozzáférés: 2022.03.07, `https://datosabiertos.bogota.gov.co/dataset/mapa-de--referencia`.

[11] Demeter Gábor, Szulovszky János et al., *GISta Hungarorum – több, mint térkép: Történeti térinformatikai rendszer*, hozzáférés: 2022.03.14, `http://gistory.hu/g/hu/gistory/otka`.

[12] Ignácz Károly, *Budapest választ: Parlamenti és törvényhatósági választások a fővárosban 1920–1945* (Budapest: Napvilág Kiadó, 2013). Online verzió a letölthető térinformatikai adatokkal, hozzáférés: 2022.03.07, `https://bpvalaszt.hu/`.

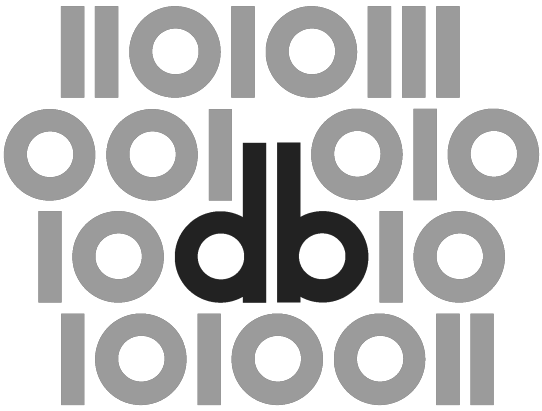
[13] Magyar Bányászati és Földtani Szolgálat, hozzáférés: 2022.03.07, `https://map.mbfsz.gov.hu/`.

 DOI 10.31400/dh-hun.2021.4.3132



Digitális Bölcsészet
2021., negyedik szám

<DIGITÁLIS BÖLCSÉSZET>



4 (2021)

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.2021.4
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.

(cc) BY-NC-SA

Ez a mű a Creative Commons *Nevezd meg! – Ne add el! – Így add tovább! 2.5 Magyarország Licenc* (`http://creativecommons.org/licenses/by-nc-sa/2.5/hu/`) feltételeinek megfelelően felhasználható.

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

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

<KRITIKA>

Király Péter 0000–0002–8749–4597

Göttingen eResearch AllianceGesellschaft für wissenschaftliche Datenverarbeitung mbH, Göttingen

peter.kiraly@gwdg.de

Kulturális adatelemzés – bevezetés és önéletrajz

Lev Manovich. *Cultural analytics*. Cambridge, Massachussetts–London, England: The MIT Press, 2020.
https://doi.org/10.7551/mitpress/11214.001.0001. 318 oldal.
ISBN 9780262037105, e-ISBN 9780262360647

Lev Manovich a digitális kultúra és média egyik vezető kortárs elméletalkotója, a City University of New York professzora, világszerte számos konferencia vendégelőadója. Néhány írása magyarul is olvasható,[1] sőt figyelemre méltó magyar recepciója is van.[2] 2007-ben alkotta meg a könyv címét is adó *cultural analytics* (a továbbiakban: kulturá-

[1] Lev Manovich, „A kulturális interfészek elmélete," ford. Kolumbán Melinda, *Magyar műhely* 37–38, 108–109. sz. (1998): 67–87; Uő, „A képernyő mögött," ford. Teutsch Áron, *Korunk* 3, 4. sz. (2000): 93–95; Uő, „Az adatbázis logikája," ford. Kiss Pál Szabolcs, *Magyar Műhely* 39, 115. sz. (2000): 48–69; Uő, „A film mint kulturális interface," ford. Vajdovich Györgyi, *Metropolis* 5, 2. sz. (2001): 24–43; Uő, „Posztmédia esztétika: Krízisben a médium," ford. Kiss Pál Szabolcs, *Erdélyi Művészet* 5, 2. (17.) sz. (2004): 11–20 és *Exindex,* 2004. márc. 10, http://exindex.hu/index.php?page=3&id==227; Uő, „Remixelhetőség," ford. Fükköh Borbála és Roszík Linda, in Halácsy Péter, Vályi Gábor és Barry Wellman, szerk., *Hatalom a mobiltömegek kezében,* 79–91 (Budapest: Typotex Kiadó, 2007), http://mediaremix.hu/remix1/letolt/manovich.pdf; Uő, „Az újmédia nyelve: Mi az újmédia," ford. Gerencsér Péter, in Gerencsér Péter, szerk., *Új, média, művészet,* 12–45 (Szeged: Universitas Kiadó, 2008), 13; Uő, „Flash generáció," ford. Gerencsér Péter, in Gerencsér, *Új, média, művészet,* 145–158, 145; Uő, „Beutazható tér – A Doom és a Myst," ford. Tóth Tamás Boldizsár, *Metropolis* 12, 1. sz. (2008), 72–98; Uő, „Mi a film?," ford. Gollowitzer Dóra Diána, *Apertúra,* 2009, ősz, http://apertura.hu/2009/osz/manovich–3; Uő, „Az adatbázis mint szimbolikus forma," ford. Kiss Julianna, *Apertúra,* 2009, ősz, http://apertura.hu/2009/osz/manovich; Uő, „Mi a digitális mozi?," ford. Kiss Gábor Zoltán, in Kiss Gábor Zoltán, szerk., *Narratívák 10: A narrációtól az attrakcióig,* 164–167 (Budapest: Kijárat Kiadó, 2011); Uő, „A mindennapi (média) élet gyakorlata," ford. Mátyus Imre, *Apertúra,* 2011, tavasz, http://apertura.hu/2011/tavasz/manovich.

[2] Mindenekelőtt Gerencsér Péter doktori disszertációja *A web 2.0 mint a net art neoavantgárdja. Folytonosságok és törésvonalak az internetes művészet diskurzusában.* Doktori értekezés. Témavezető: Dr. Dragon Zoltán. Szegedi Tudományegyetem Bölcsészettudományi Kar Irodalomtudományi Doktori Iskola, 2017. (https://doi.org/10.14232/phd.4006) Ezen felül Andok Mónika, „Manovich újmédia-elmélete: a Software Studies", in: Balázs Géza és H, Varga Gyula, szerk., *Sajtónyelv-médianyelv: Kutatás, elemzés, dokumentumok,* 11–22 (Budapest, Hungarovox Kiadó, 2017); Ármeán Otília, „Interfész és esztétikum Lev Manovich elméletében," *ME.DOK: Média-Történet-Kommunikáció,* 4. sz. (2012): 27–40, https://www.medok.ro/sites/medok/files/publications/pdfs/ME.dok–2012–4.pdf; Sághy Miklós, „Az adatbázis-logika és a film: Reflexiók

lis adatelemzés) kifejezést, amely az utóbbi években mint egyfajta új interdiszciplináris tudományág terjedt el: számos intézmény elnevezésében helyet kapott, létrejött a saját folyóirata[3] és felsőoktatási képzési rendszere. A könyv egyrészt a kulturális adatelemzés meghatározása, történetének, céljainak és módszereinek ismertetése, másrészt viszont bevallottan szerzői önéletrajz és visszatekintés az elmúlt másfél évtized munkásságára.

A kulturális adatelemzés kiindulópontja Manovich 2005-ben feltett kérdéssorozata: hogyan érthetjük meg a naponta milliárdnyi képpel gyarapodó kortárs fotográfiát? vagy a húszmillió felhasználó által megosztott százmilliónyi dallal reprezentálható kortárs zenét? vagy a négymilliárd Pinterest-oldalt? Manovich válasza az ehhez hasonló kérdésekre az, hogy a mintázatok, trendek és a kortárs kultúra dinamikájának tanulmányozásához adattudományi módszereket kell alkalmazni. A könyv azonban szándékosan nem érinti az adattudomány gyakorlati lépéseit. Arra fókuszál inkább, hogy milyen tudományos (és művészeti) kérdéseket lehet feltenni és megválaszolni, így a könyvet olyan kutatóknak is ajánlja, akik az adattudományban járatlanok. A szerző az adatvezérelt kutatások számára a következő munkafolyamatot ajánlja: 1) a kutatás tárgyának kvantitatív módszerekkel való elemezésén vagy reprezentálásán gondolkodás; 2) a megfelelő adatok elérhetőségének, illetve előállíthatóságának vizsgálata; 3) adatgyűjtés; 4) az adatok áttekintése vizuális módszerekkel; 5) statisztikai és adattudományi adatelemzés; végül lehetőség szerint 6) interaktív adatvizualizáció, hogy mások is áttekinthessék az adatokat (20).

Manovich 2007-ben alapította meg a ma San Diegóban és New Yorkban Cultural Analytics Lab néven működő kutatási csoportját –, amelynek eredményei a könyv gerincét adják. A laboratóriumnak két célja volt: egyrészt használatba venni a számítástudomány, az adatvizualizáció és a médiaművészet meglévő módszereit különféle jellegű kortárs médiumok áttekintése és elemzése érdekében, másrészt megvizsgálni, hogy mindezek a módszerek és a kulturális média nagy adattárai milyen módon kérdőjelezik meg kultúráról alkotott mai elképzelésinket és elemzési módszereinket. Számos helyen kifejti, hogy a digitális bölcsészet elsősorban a múlt szövegkorpuszaival foglalkozik, részben ennek ellensúlyozására a labor elsősorban a kortárs vizuális kultúra tanulmányozására fekteti a hangsúlyt. A kulturális adatelemzés azonban több ennél, a kifejezést (és változatait) ma tucatnyi kutatási-oktatási intézmény használja, szakfolyóiratai jöttek létre, és ezek jóval szélesebb körben értelmezik mind az adatelemzés, mind a kultúra fogalmát, mint amiről a jelen könyv számot ad. A szerző az öt legfontosabb, őt „legszenvedélyesebben" érdeklő témával határolja le a könyv mondanivalóját. 1) A kulturális adatelemzés a kortárs kultúra informatikai és formatervezési módszerek alkalmazásával történő áttekintését és elemzését jelenti. A kultúra fogalmába beletartoznak az átlagember, a művészek és a kreatív ipar által megvalósított kulturális tevékenységek. Célja lehetővé tenni annak áttekintését, hogy világszerte mit is alkotnak, mit képzelnek el, és mit értékelnek az emberek. A magas művészet mellett a kulturális élet „hosszú farka" *(long tail)* is érdekes és elemzendő, amelynek

Lev Manovich és Dragon Zoltán írásaira," *Apertúra,* 2011, tavasz, `http://apertura.hu/2011/tavasz/saghy`.

[3] *Journal of Cultural Analytics,* 2016– (ISSN 2371–4549). Kiadja a montreali McGill Egyetem Nyelvek, Irodalmak és Kultúrák Tanszéke, hozzáférés: 2022.04.07, `https://culturalanalytics.org/`

 10.31400/dh-hun.2021.4.3510

segítségével olyan városok, országok, csoportok és személyek is felkerülnek a kulturális térképre, akiket – így Manovich – mind a történeti, mind a kortárs narratívák, de sokszor a nemzeti digitalizációs programok is figyelmen kívül hagynak. 2) A nyelv nem biztosít elég kifejező erőt a nagy mennyiségű adatokban fellelhető számos árnyalat és eltérés leírására, ezért a számok és a vizuális kifejezés eszközéhez kell folyamodnunk. 3) A könyv, a digitális bölcsészet szövegközpontúságát ellensúlyozandó, elsősorban a vizuális médiát elemzi. 4) Kikerülhető-e az a nyelv által támasztott kényszer, hogy kategóriarendszerekben gondolkodjunk a kulturális tevékenységekről, kikerülhető-e a statisztikai megközelítésből következő kvantifikáció, egyáltalán: lehet-e számok nélkül nagy mennyiségű kulturális adatot elemezni? 5) A kulturális adatelemzés nemcsak felhasznál bizonyos számítási módszereket, de ezeket, illetve a módszerek céljait és rejtett előfeltevéseit kritikai vizsgálat alá is veszi.

Már a kutatás megkezdésekor számba vett néhány kihívást, amellyel a tudományterületnek számolnia kell. Ilyen – többek között – az, hogy milyen alapvetően új értelmezéseket tesznek lehetővé a számítási módszerek és a nagy adattárak? Hogyan térképezhetünk fel milliárdnyi képet és videót? Hogyan kombináljuk a kvalitatív és kvantitatív módszereket? Hogyan tudjuk elemezni az interaktív médiát? Milyen elméleti fogalmakkal és modellekkel kell elemezni a felhasználók által generált, hatalmas mennyiségű és sebességű tartalmat és interakciót? Milyen lesz a „kultúra természettudománya" (*science of culture,* 50) és hol lesznek a határai? Elemek, témák és stratégiák statisztikai megoszlásaival és kombinációival le lehet-e írni a kultúrát egyszerre objektívan és pontosan? Végül – és Manovich számára éppen ez a legérdekfeszítőbb kérdés: mi a redukció megfelelő szintje és mi veszik el, amikor sok milliárd kulturális objektumot kevés számú témává szűkítünk le?

A szerző számára a könyv a médiaelmélet részét képezi, benyomásom szerint leginkább az erről a területről érkező olvasóknak lesznek ismerősek Manovich hivatkozásai, de jellegéből adódóan a könyv számos tudományos és gyakorlati területtel is érintkezik. Külön felhívja a figyelmet a szoftverek társadalmi-gazdasági hatásait elemző *software studies*ra, melynek korábban egy önálló kötetet szentelt.[4] Épp ezen hatások miatt úgy véli, „minden kulturális és médiakutatónak és -hallgatónak alaposan meg kell értenie az adattudományt és a mesterséges intelligenciát" (19).

Manovich amellett érvel, hogy bár számos előzményt találunk a 19. század közepétől kezdve a matematika, statisztika, adatvizualizáció bölcsészeti alkalmazására (ilyen történeti előzményeket a *Digitális Bölcsészet* több cikkében is találunk, de a jelen kötetben a kevésbé ismert orosz vonatkozások a leginkább érdekesek). A fordulat kulcsa a kulturális termelés, terjesztés és részvétel új nagyságrendje, ami 2005 táján következett be és immár megközelíti a fizikusok és biológusok által vizsgált jelenségeket, bizonyos tulajdonságai (például a hatványfüggvényszerű eloszlás, a skálafüggetlen hálózatok jelenléte) pedig azonosak e jelenségekével. A tudományos fordulatot egy 2007-es konferencia jelentette, amely elindította a közösségi média kvantitatív elemzésének virágkorát. Egy másik fontos, korábban kevésbé figyelembe vett változás: a 90-es évektől nemcsak a professzionális kulturális ipar növekedett világszerte ugrásszerűen, de a kreatív területek oktatása is. Számos helyen felhívja a figyelmet a földrajzi vonat-

[4] Lev Manovich, *Software Takes Command* (London: Bloomsbury Academic, 2007, jav. kiad.: 2013), https://doi.org/10.5040/9781472544988.

kozásokra, állítása szerint a korábbi centrum–periféria felosztás értelmét vesztette, ennek helyébe a néhány kiugró centrum mellé sok-sok, korábban figyelembe nem vett fejlődő és exkommunista ország városa zárkózott fel a kulturális tevékenységek aktivitási térképén. Kérdéses persze, hogy pusztán az aktivitás növekedett-e meg, vagy pedig azzal párhuzamosan annak globális elérhetősége is.

A szerző gondolkodásmódjára nagy hatással voltak az interaktív, *real-time* vizualizációs felületek, például a *Google Analytics* (ami a *cultural analytics* kifejezés közvetlen ötletadója is lett), a közelmúltban elhunyt svéd népegészségügyi professzor Hans Rosling *Gapminder* nevű szoftvere, vagy a marketingcégek kialakította vásárlói aktivitást mérő felületek. Manovich víziója az volt, hogy weboldalak vagy áruházak forgalma helyett a kultúra folyamát, a „kulturális földrengéseket" lehetne hasonló módon sok szempont alapján megjeleníteni. A labor egyik első eredménye egy különféle kulturális adatokat és folyamatokat megjelenítő monitorfal, melyen a felhasználó tetszés szerint összevethetett adatokat, nagyíthatott, új szempontokból kombinálhatott. Egyúttal azonban, figyelmeztet Manovich: bizonyos adatelemzések már lehetségesek (pl. nevekhez tartozó egyedek, vagy képeken szereplő tárgyak azonosítása), mások azonban (például az esztétikai vagy szemantikus tulajdonságok figyelése) továbbra is nagyon távoli célnak tűnnek.

A kulturális adatelemzés nem a semmiből lett hirtelen, a *Kultúra természettudománya?* című fejezet számos előzményt és párhuzamos kutatást (többek között a társadalomtudományi célú számítástechnika, a szociális média elemzése, a médiatörténet vagy a digitális bölcsészet köréből) és egyéb kulturális produktumot (pl. *Ngram Viewer,* könyvtári adatvizualizációk) említ, melyeknek hasonló a célja és a módszertana mint a laborban folyó munkának. Ahhoz, hogy a kulturális adatelemzés önálló diszciplína legyen, meg kell vizsgálni, mi köti össze és választja szét a három legfontosabb összetevőt, vagyis a bölcsészetet és kvalitatív társadalomtudományokat, a kvantitatív társadalomtudományokat, valamint a számítástudományt. A bölcsészet – Manovich értékelése szerint – az egyedire fókuszál, a dolgok jelentését keresi és a múlt felé tájolt. Ezzel szemben a természettudományok az általánosra koncentrálnak, tudományos módszertant és matematikát használnak, és valamiképpen szeretnék megjósolni a jövőt. A szerző érdekes módon nem nyújt annál bővebb kvantitatív bizonyítékot arra az állítására, hogy a digitális bölcsészet elsősorban hivatásos alkotók által létrehozott történeti tárgyakat vizsgál, minthogy a kedves olvasónak a legfontosabb szakfolyóirat és konferencia tartalomjegyzékeinek átböngészését ajánlja (holott például a régészet, a néprajz, a kulturális antropológia, de még a történelem vagy az irodalomtudomány is számos, nem az „elit" által létrehozott tárgyat és egyéb jelenséget vizsgál). A közösségi média iránt érdeklődő informatika pedig éppen ellenkezőleg: a tömegek által létrehozott kortárs „népi" produktumokra koncentrál *(nonprofessional vernacular culture, popular culture).* Összességében három ellentétpárt lát Manovich, amiket az új területnek egyszerre kellene figyelembe vennie: történelmi perspektíva a jelenre való fókuszáltsággal szemben; hivatásos alkotók az amatőrökkel szemben; általános az egyedivel szemben. A természettudományok 18–19. századbeli előtörésével többekben megszületett a gondolat, hogy a természeti törvények mintájára, sok munkával, hasonló tételeket lehetne alkalmazni a társadalomra is; ezek a várakozások azonban nem igazolódtak be. Manovich a determinisztikus törvények megalkotása

 10.31400/dh-hun.2021.4.3510

helyett a természettudományok két másik alapvető megközelítési módját venné át: a valószínűségen alapuló statisztikai modelleket – melyek nem az egyedre vonatkozóan határoznak meg kikerülhetetlenül érvényesülő törvényszerűségeket, hanem a vizsgált népesség teljességére valószínűségeket –, illetve a tudományos szimulációt.

Manovich hosszan ír a kulturális iparban folyó médiaelemzésről *(media analytics)*. Bár módszereik hasonlóak (és így a kulturális adatelemzés sokat tanulhat belőle), a médiaelemzés alapvető célja a Horkheimer és Adorno 1947-es *A felvilágosodás dialektikája* alapján „kulturális iparnak" nevezett terület valamilyen gyakorlati-üzleti folyamatának támogatása (például annak eldöntése, hogy melyik hirdetés kerüljön a felhasználó elé). Ez a fajta médiaelemzés két nagy tevékenységi kört fed le: a felhasználói interakciók elemzését és az ennél kevésbé ismert tartalomelemzést, amely a tartalom különféle jellegzetességeit igyekszik megragadni (például a *Spotify* olyan elvont fogalmakat, mint a „táncolhatóság", „hangszeresség", „beszédarány"). A média egyénre szabott, automatizált akciói (például a keresési eredmények összeállítása, a hirdetések) mindkét tevékenységi kör eredményeit felhasználják. Ezek lehetnek részben felhasználói tevékenységek (például szűrők és más beállítások segítségével) kontrollálhatók, vagy teljes mértékben automatizáltak. Továbbá az akciók lehetnek determinisztikusak (vagyis ugyanazon feltételek mellett ugyanazt eredményezőek), illetve nem determinisztikusak. A médiatörténet azt mutatja, hogy a 21. században elmozdulás történik a kulturális ipar 20. századi szorosan determinisztikus technológiái és gyakorlata felől a nem determinisztikus technológiák felé, mint amilyen a neurális hálókat használó felügyelt gépi tanulás, amely ugyan jó teljesítményt nyújt, de működése szinte interpretálhatatlan. Ezért a korábban elterjedt algoritmikus jelzőt Manovich nem javasolja használni, hiszen klasszikus értelemben az algoritmus determinisztikus, helyette inkább az elemzést *(analytics)* vagy a szoftverest javasolja, melyek lefedik mindkét módszert (63–64).

Fontos probléma, hogy az újfajta tartalomiparhoz a kutatók gyakorlatilag alig, vagy egyáltalán nem férnek hozzá; néhány cég a kutatás számára megosztja ugyan tartalmuk egy szeletét, de lehetetlen például a közösségi médumok adott napi teljes terméséhez hozzáférni, a cégek pedig igen ritkán publikálják saját kutatásaik eredményeit. A szerző Horkheimer és Adorno állítását, miszerint „ma a kultúra mindent az egyformasággal fertőz meg" kutatási kérdéssé módosítja: mi a médiaelemzés hatása a kulturális ipar által előállított és a fogyasztók számára megmutatott, általuk választott termékekre? Vannak ugyan ilyen jellegű kutatások (69–71), de egyelőre kevés eredménnyel szolgálnak.

A könyv középső szakasza a kultúra adatként való reprezentációjával foglalkozik. A kutatás forrásaként szolgáló adatokat Manovich a következő kategóriákba sorolja: média (a professzionalisták vagy amatőrök által előállított tényleges tartalmak), viselkedés (digitális nyomot hagyó online vagy természetes aktivitás), ember-gép interakció (az előbbi kategória kiemelésre érdemes alárendeltje, amely alatt az interaktív, algoritmusvezérelt média, a virtuális és kiterjesztett realitás alkalmazások és a felhasználóik közötti interakciót érti) és az esemény (időtartammal rendelkező és embereket érintő kulturális események, melyeknek összetevői, alkategóriái a helyek és a szervezetek).

Az adatoknak más felosztási módjai is elképzelhetők: kulturális adatok (kulturális tárgyak és rendszerek), kulturális információk (az előbbiek leírására szolgáló metaadatok) és kulturális diskurzus (az első kategóriára vonatkozó reflexiók) vagy digitálisan született tárgyak, más mediális formában keletkezett digitalizált tárgyak és kulturális élmények vagy megint csak másik felosztás szerint szerzők, szövegek (üzenetek) és hallgatóság. Számba veszi az egyes kategóriák fontosabb adatforrásait, a korábban már említett hozzáféréssel kapcsolatos nehézségeket, illetve megemlít néhány példaként szolgáló kutatást. figyelmeztet néhány elhanyagolt szempontra, például a nagy, sokak által használt és ezért sok kutatás által elemzett szolgáltatások mellett elsikkad az egyedi honlapokon található tartalom vizsgálata és az ebből fakadó kutatási kérdések feltétele. A viselkedési adatok esetében felhívja a figyelmet a személyes adatok kezelésére és az ebből fakadó következményekre, például, hogy az EU-s Általános Adatvédelmi Szabályozás (GDPR) következményeként a közösségi médiumok API-jai 2018-tól fogva jóval kevesebb, a felhasználóra vonatkozó adathoz engednek hozzáférést. Rámutat arra is, hogy az ezen adatok elemzéséhez is keretet nyújtó, a társadalomtudományok által még a digitális korszak előtt kifejlesztett kvalitatív kutatásmódszertani eszköztár (résztvevők megfigyelése, különféle interjútípusok, a sűrű leírás stb.), továbbá és az ehhez kapcsolódó elméleti viták a bölcsészetben ismeretlenek (Manovich itt is láthatólag valamilyen szűkebb értelemben beszél a bölcsészetről, hiszen létezik erre reflexió[5]). Az ember-gép interakcióval kapcsolatos adatok jelentik a legtöbb kutatási problémát: hogyan lehet egyáltalán rögzíteni és feldolgozni ezeket? Itt – a többi adattípustól eltérően – nem is hoz kutatási példát, de a könyv egy másik pontján beszámol saját kísérletéről.

A kulturális mintavétel önálló fejezetet kapott, melynek fő kérdése, hogy a gyűjtemények mennyire reprezentálják a teljességet, illetve kell-e és lehet-e a gazdaság- vagy társadalomtudományok mintájára olyan reprezentatív, vagyis a vizsgált populáció főbb tulajdonságait tükröző adatcsomagokat előállítani (rétegzett mintavételezéssel), amelyek alapul szolgálhatnak az elemzéseknek. Jelenleg nincsenek ilyen adatcsomagok. Elkészítésükhöz két fontos szempontot kellene figyelembe venni: 1) mintavétel előtt az adott kulturális univerzumot szükséges tanulmányozni és megérteni 2) nagyon sokféle elemzési szempont létezhet (a művek vagy a befogadás ismérvei vagy akár véletlen választások alapján), de a minta ezek közül csak néhány alapján lehet reprezentatív, nem lehetséges egyszerre minden szempontot figyelembe venni. A különféle szempontok alapján létrehozott minták épp ezért kiegészítik egymást. Manovich számos példát felsorol, amikor hagyományos és digitális állományok nem reprezentálják a saját maguk által kitűzött kulturális világ teljességét, az általa elemezni kívánt teljességet pedig végképp nem. Kitér a kánonok kritikájára, és azt javasolja, hogy az új bölcsészeti kánonok legyenek kiegyensúlyozottak a korábban nem reprezentált csoportok irányába. (Olvasóként felmerül bennem, ez lehetséges-e, hiszen a kánon alapvetően nem szociológiai, hanem esztétikai elvű kategória. Ha a munkásszobrász

[5] Például a sűrű leírás és a digitalizálás összefüggéseivel kapcsolatos elgondolásokra lásd Tóth-Czifra Erzsébet nagyszabású áttekintését: „The Risk of Losing the *Thick Description:* Data Management Challenges Faced by the Arts and Humanities in the Evolving FAIR Data Ecosystem", in Jennifer Edmond, ed., *Digital Technology and the Practices of Humanities Research,* 235–266 (Cambridge, UK: Open Book Publishers, 2020), https://doi.org/10.11647/obp.0192.10.

 10.31400/dh-hun.2021.4.3510

nem alkot remekművet, akkor is kerüljön be, mivel egy adott társadalmi réteget reprezentál? A könyv egy másik pontján említi Barabási kutatásait, amely alapján a művészeti siker a tehetségen felül a jó hálózati kapcsolódások következménye; érdemes lenne azon gondolkozni, hogy ennek a nem esztétikai kategóriának a reprezentativitást elnyomó hatásait hogyan lehetne a kánonalkotásban ellensúlyozni, de Manovichnál ez a szempont nem merül fel.) Bár nincsenek ilyen adatbázisok, voltak hasonló kísérletek, elsősorban Bourdieu *Különbség: Az ízlésítélet társadalomkritikája* című 1979-es, az emberek ízlése és gazdasági-társadalmi státusza összefüggésének elméletét tárgyaló műve nyomán, amelyben a szerző az akkor új statisztikai módszert, a korreszpondenciaelemzést alkalmazza. Fontos, és máshol is előkerül a könyvben, hogy a statisztikakészítés mindig redukció: míg a tipikusra rámutat, az egyedit elfedi, tehát a könyv elején kitűzött célok közül nem mindet lehet teljesíteni általa. Ezért ideális esetben a vizsgálni kívánt kulturális folyamat során létrejövő minden adatot meg kell próbálnunk beszerezni és elemezni.

A metaadatok és tulajdonságok fejezete egy definícióval kezdődik: „Egy kulturális jelenség adatreprezentációja számos adatobjektumot és ezen objektumok valamilyen rendszer szerint kódolt jellemzőit tartalmazza." (122.) A reprezentáció készítése során három kulcskérdésre kell választ adni: hol vannak a jelenség határai, melyek a reprezentálandó tárgyak és mik a tulajdonságai. Ezen felül el kell döntenünk, hogy két klasszikus nézőpont melyikét tesszük magunkévá: a jelenség vajon tőlünk függetlenül is létező dolog, vagy éppen mi állítjuk elő a reprezentálással és a tárgyalással? Foucault nyomán a fejezet megkísérli az adattudomány archeológiai elemzését. A tulajdonságokra különféle területek számos eltérő, részben azonos jelentésű terminust használnak, Manovich elsősorban azt emeli ki, hogy a tulajdonságok egy része emberi tevékenység (például könyvtári katalogizálás vagy közösségi médiás aktivitás) nyomán született, más részük pedig számítógépes elemzés következtében. Nagyon fontos, hogy az adatot mindig előállítják, nem csak úgy organikusan terem, vagyis puszta léte és jellemzői mindig emberi döntések eredménye. A digitális objektum befoglalt, rögzített tulajdonságai mellett az alapul szolgáló objektumnak lehetnek olyan tulajdonságai is, amelyek kívül maradtak az adatrögzítés során. Egy gyűjteményben az objektumok tulajdonságait úgy kellene kódolni, hogy azokkal műveleteket tudjunk végezni, ennek megfelelően azonos tulajdonságok formátuma ideális esetben mindig azonos (például nagyon megnehezíti az elemzést, ha a dátum egyszer szám, másszor számtartomány vagy szöveges adat). Az adatreprezentáció tehát adattudományi technológiák felől nézve számításra alkalmas, kezelhető, megismerhető és megosztható. Az emberek által az adatokból kinyerhető információ és az adat számítógép számára látható képe között jelentős szemantikus hézag van, amit az adatelemzés során mindig figyelembe kell vennünk.

A *Nyelv, kategóriák és érzékelés* fejezet további szempontokkal bővíti az adatok általános elemzését. Ismerteti a négy alapvető mérési skálát (névleges, sorrenden alapuló, intervallum- és arányskálák) és ezek alkalmazását a kulturális adatelemzésben. Bő teret szentel a könyv több pontján már említett gondolatnak, miszerint a nyelv korlátozott és az érzékszervi észlelés spektrumához közelebb áll a számszerű ábrázolásmód, amely lényegében új nyelvet kínál a kultúra leírásához és tárgyalásához, ugyanakkor a nyelv és az érzékelés komplementer rendszert alkot. Az érzékeléssel

kapcsolatos tapasztalatok művészeti kitágítására az egyik példa Andy Warhol 1964-es *Empire* című filmje. Warhol letett egy kamerát az Empire State Buildinggel szemben egy irodában és nyolc órán keresztül filmezte az épületet, majd némileg gyorsítva, hat és fél órán keresztül vetítette (megtagadva, hogy a filmet vágva vagy eltérő sebességgel játsszák le).[6] Warhol munkája nemcsak Manovichnak adott támpontot, hanem Mészöly Miklósnak is, aki *A tettenérés dialektikája: A cinéma direct útmutatásai* című esszéjében[7] egy új esztétika lehetséges alapjaként tekintett a műre. *Film* című regényének szerkezetére is nagy hatást gyakorolt, annak ellenére, hogy Mészöly, esszéje tanúsága szerint nem látta a felvételt, csak egy francia filmes szaklapban olvasott róla. Az érzékelés mérésére léteznek módszerek, de eddig igen kevés olyan tanulmány született, mely a művészeti hatás mértékét vizsgálta volna, holott el lehetne képzelni akár olyan filmkritikát is, amely a cselekmény vagy színészi játék helyett a nézők reakcióit írja le. A nyelvhez kötött kategóriák tökéletlenül írják le a művészi produktumot, és ez néha a művészettörténeti elemzésre is hatással lehet. Példaként egy saját kutatást hoz, amelyben Van Gogh képeiben a világosság és a színtelítettség időbeni változását elemezve jutott arra a művészettörténeti konszenzussal némileg szembemenő következtetésre, hogy a festő újításai – amelyeket egy-egy lakóhelyváltoztatás után bevezetett és amelyek alapján életművét hagyományosan korszakolják – a megállapított korszakain belül sem kizárólagosak, és e korszakokon átívelőek (vagyis Van Gogh egy-egy újítás után is festett „régi" stílusú képeket).

A könyv utolsó része, a *Kulturális adatok feltárása* elsősorban gyakorlati problémákkal és ebből fakadó elméleti javaslatokkal foglalkozik, sorra véve az információvizualizációt, a feltáró médiaelemzést és a médiavizualizáció módszereit. Az információvizualizáció az adat és a vizuális reprezentáció közti leképezést jelenti. Célja, hogy egy (tipikusan nagy) adathalmaz szerkezetét megismerjük, ellentétben az információdizájnnal, amely ismert, világos struktúrából indul ki és célja ennek a struktúrának a képi megmutatása. Két fontos alapelvre épül: az adatok redukciójára és a térbeli változók (elhelyezkedés, méret, forma, és újabban: görbület, illetve mozgás) alkalmazására. (Ha már fentebb Mészölyt szóba hoztuk, meg kell említsünk egy másik párhuzamot: Manovichnak a redukcionizmusról adott leírása párhuzamba állítható Nádas Péter több helyen elmondott hasonló leírásával, tudniillik saját írásművészetét némileg redukcionista mesterei – pl. Mészöly – ellenében alakította ki.) A legtöbb ilyen térbeli leképezés- vagy grafikontípust 1800 és 1850 között találták ki, azóta nagyon kevés új elem jelent meg az eszközkészletben (a számítógépek alkalmazása elsősorban a hatékonyságot

[6] A cikk írása idején YouTube-on megtalálható volt az eredeti tempójú nyolcórás változat (hozzáférés: 2022.02.15, `https://www.youtube.com/watch?v=py_WBhCxoKc`), amelyet időközben az Andy Warhol Museum szerzői jogi problémák miatt eltávolíttatott. Jelenleg egy hat és fél órás (hozzáférés: 2022.04.05, `https://www.youtube.com/watch?v=UM0QrH0c268`) változat érhető el, valamint a legendás avantgárd filmes Jonas Mekas visszaemlékezése, aki jelen volt a forgatáson (hozzáférés: 2022.04.05, `https://www.youtube.com/watch?v=XnN1NqXr1Qs`).

[7] Mészöly Miklós, „A tettenérés dialektikája: A cinéma direct útmutatásai," *Filmkultúra* 2. sz. (1969): 49–54, hozzáférés: 2022.04.05, `https://mandadb.hu/dokumentum/6512/hatvankilenc2.pdf`). Átdolgozott formában „Warhol kamerája – a tettenérés tanulságai" címen megjelent *A tágasság iskolája* kötetében (Budapest, 1977). Elektronikus verziója a DIA-n olvasható, hozzáférés: 2022.04.07. `https://reader.dia.hu/epub–reader.html?token=2LLIp2su9kPCG976%2FnJJV2uBO%2ByELDmofxbPEg93tNyD9143f%2BKse2SVKyHOkBWL&lng=hu`.

 10.31400/dh-hun.2021.4.3510

növelte). A vizuális szimbólumok hatékonysága azonban nem egyforma, az érzékelés biológiai tulajdonságaiból következően a szín, tónus, átlátszóság és forma általában kevésbé hatékony információközvetítők, így a vizualizációban az adat másodlagos aspektusait tükrözik. Manovich számára ezzel szemben az az igazán érdekes, hogyan lehet redukció *nélkül* ábrázolni az adatokat. Ezt a megközelítést médiavizualizációnak *(media visualization)* vagy közvetlen vizualizációnak *(direct visualization)* nevezi (197–198), amely esetben az eredeti vizuális médiaobjektumokból vagy részeikből születik új képi ábrázolás. Ennek egyik, általánosan ismert változata a szófelhő, amelyben maguk a szavak szerepelnek, mindenféle redukció nélkül. Az ábrázolásmódnak a közelmúlt művészetében voltak előzményei, ezek közül a *Digitális Bölcsészet* olvasói számára talán legérdekesebb egy, a kritikai szövegkiadáshoz kapcsolódó alkotás, Ben Fry *A kedvelt nyomvonalak megőrzése*[8] című alkotása, amely Darwin *A fajok eredete* hat kiadásának sorban következő szövegváltozatait mutatja be mozgóképszerűen. Ezeknek az alkotásoknak közös nevezője, hogy időbeli változásokat transzponálnak térbeli megjelenítésbe, de valamilyen módon minden pillanatban a teljes kiinduló objektumot megjelenítve. Manovich úgy gondolja, hogy a korábbi, információt grafikus jelekké konvertáló ábrázolásmód csak a technológiai korlátokból fakadó gyakorlati kényszer, tehát történeti jelenség volt. A feltáró adatelemzés *(exploratory data analysis)* mintájára megalkotja a feltáró médiaelemzés *(exploratory media analysis)* fogalmát (207). Az eredeti fogalmat (csakúgy mint az adattudomány kifejezést) John Tukey alkotta meg és az elemzés első, többnyire vizualizációval együtt járó lépésének tekintette, melynek célja megismerni az adatok alapstruktúráját. A médiaelemzés ennek mintájára olyan vizualizációt jelent, amelyben a teljes gyűjteményt ábrázoljuk, valamilyen tulajdonságai szerint rendezve. A közgyűjteményi vagy a médiaszolgáltatások keresőfelületei, de a médiakezelő alkalmazások is igen korlátozottak abban, hogyan mutatják meg az állományt, milyen keresési és rendezési szempontokhoz férhetnek hozzá a felhasználók, és a médiaelemeket hogyan képezik le a felhasználói felületre (üdítő kivétel Benoit Seguin *Replica* nevű alkalmazása, amely reneszánsz és barokk festmények elemzését segíti képi motívumok keresésének és a képek rendezésének lehetővé tételével).[9] Manovich célkitűzése a médiavizualizációs megközelítési mód „szabványosítása" és ennek érdekében szabad forráskódú, és különféle vizuális média esetében alkalmazható szoftverek elkészítése lett.

A médiavizualizáció mélyen kvalitatív és sokkal kevésbé kvantitatív módszer. A képeket valamilyen metaadatuk (például készítési idejük vagy helyük, kategóriájuk) alapján szokás elrendezni, Manovich ezzel szemben izgalmasabbnak tartja, ha valamilyen belső tulajdonság alapján „újratérképezi" *(remap)* az elrendezést (223). A másik fontos döntés a művelet tervezésekor arról kell szóljon, hogy az összes objektumot, vagy annak valamilyen mintavételezett részét jelenítsük meg. Ez a mintavétel lehet

[8] Hozzáférés: 2022.04.05, `https://fathom.info/traces/`.

[9] Benoit Seguin, „Replica Project: Status, Issues and Directions," `https://profile.benoitseguin.net/2016/12/19/replica-project-status-and-roadmap.html`. Magyar szempontból érdekesség, hogy az alkalmazás annak a Foundation Giorgo Cini-nek a 300 000 darabos digitális képgyűjteményére épült, ami a Klaniczay Tibor-féle reneszánsz és barokk kutatásoknak első, és talán legfontosabb külföldi partnere volt, ezért igen jól ismert a korszak magyar kutatói számára. Lásd. Ács Pál és Székely Júlia, *A reneszánsz reneszánsza* (Budapest: Bölcsészettudományi Kutatóközpont Történettudományi Intézet, 2021).

idő- (egy mozgófilmből snittenként egy képkocka) vagy téralapú (minden kép egy bizonyos részlete). A metaadat-alapú elrendezésben a képek például időrendi sorrendben és ugyan kicsinyítve jelennek meg egyetlen oldalon, de az emberi szem számára a legfeltűnőbb vizuális jegyek változása (például egy múzeumi gyűjtemény esetében bizonyos évek termésének kiugróan nagy aránya) itt is feltűnhet. A kép fizikai tulajdonságainak (világosság, színtelítettség, kontraszt stb.) számszerűsítésével és erre alapuló információvizualizálással együtt pedig világossá válhatnak rejtett tendenciák is. Az automatikus videóösszefoglaló készítése, médiaipari fontossága miatt aktív kutatási téma, amelynek kialakult a kritériumrendszere, ez azonban elsősorban az egyedi, illetve nem esztétikai (pl. marketing) szempontokból fontos pillanatok megragadására összpontosít, vagyis a kulturális adatelemzés szempontjából nem kielégítő. A téralapú mintavétel egyik példájában az amerikai *Time* magazin címlapjainak közepéből vett egy pixel széles csíkokat illesztettek össze, amelyből a keret és a kép arányának, a keret színárnyalatainak, illetve a kép jellegének változásai olvashatók le. Tudományos szempontból leginkább talán az utolsó, az újratérképezést illusztráló példa érdekes. Manovich Dziga Vertov szovjet rendező és filmteoretikus, a montázstechnika és a későbbi *cinema verité (kinopravda)* irányzat atyja 1928-as *A tizenegyedik év* című filmjéről készített egy szemléletes vizualizációt. A képernyőn két hosszú filmcsík jelent meg: a felsőn a film snittjeinek kezdő, az alsón a befejező képkockáit helyezte el úgy, hogy az egymáshoz tartozó kiinduló- és végpontok mindig összevethető módon, egymás alatt helyezkedtek el. Vertovot a dinamikus kompozíciók és a szokatlan nézőpontok alkotójaként ismerik, de ez a médiavizualizáció egy másik oldalára világít rá: bár a snittek egymásutánja tényleg igazolja mindezt, ha az egyes snitteket nézzük, azok kimondottan statikusak, mozgást, látószögváltást nem tartalmaznak, az első és az utolsó kocka csaknem mindig egyforma.

A könyv utolsó, rövid fejezete a konklúzió, benne megírásának célja („hogy bemutassam a kulturális adatelemzésben tett saját utamat, és hogy mit tanultam annak 2007-es kezdete óta", illetve „hogy megvizsgáljam az egyes lépések során felmerülő fogalmakat, megkérdőjelezzem a hagyományos módszereket és rámutassak fel nem fedezett lehetőségekre", 245) és a szerző egyfajta ars poeticája is elhangzik. Újra megismétli, hogy célja mind a normális, mind a kivételes vizsgálata, egyikről sem akar lemondani. Minden egyes kulturális kifejeződésben és interakcióban van valami egyedi. Ez az egyediség bizonyos elemzéseknél fontos, másoknál nem. „A kulturális adatelemzés végső célja a hivatásosok és amatőrök által a teljes földkerekségen létrehozott kortárs alkotások változatosságának feltérképezése és megértése lehet, vagyis arra is rá kell világítania, hogy a számos alkotásban mi az eltérés és nem csak arra, hogy mi bennük a közös." (251.) E tekintetben a kultúra öt legfontosabb mérendő jellegzetessége a változatosság, az egyediség, a dinamika (időbeni változás), a szerkezet és a változtathatóság.

Manovich munkája összességében igen fontos, inspiratív és iránymutató mű – nem kevés szubjektivitással, valamint néhány hiányossággal és tévedéssel. Legfőbb hiányának azt tartom, hogy noha az olvasó már az első oldalak elolvasása után is kellő lökést érezhet magában ahhoz, hogy elkezdje e csodálatos elemzések alkalmazását saját adatain, ehhez mégsem kap iránymutatást. A szerző sokszor utal arra, hogy a könyv nem gyakorlati útmutató, de jó lenne legalább jegyzetben ajánlani néhány olyan

 10.31400/dh-hun.2021.4.3510

könyvet vagy online tananyagot, amin az olvasók elindulhatnak, amit Manovich a saját kurzusain minden bizonnyal meg is tesz. Annál inkább fontos lenne, mivel az adattudományi könyvek sok esetben egyetlen adott tudományterületre koncentrálnak, onnan veszik az adataikat és annak a tudományos kérdéseit próbálják számítógépes módszerekkel megválaszolni.

A másik probléma, véleményem szerint, a történeti aspektus korlátozott volta. Vegyük például a következő felvezetést: „A kulturális globalizáció, amely divathetek és művészeti biennálék százaihoz vezetett világszerte, […]" (27). Biennáléket természetesen már jóval korábban rendeztek – hazánkban a '60-as, '70-es évektől vált szokássá, például a Miskolci Országos Grafikai biennálét 1961 óta, a soproni Országos Érembiennálét 1977 óta rendezik –, igaz, a '90-es évektől számuk megugrott. Általában elmondható, hogy Manovich kérdései érdekesek, elemzései, válaszai azonban néha pontatlanok. Nem, vagy csak keveset szól néhány figyelmen kívül nem hagyható kontextusról, nevezetesen, hogy milyen számítási és emberi erőforrások szükségesek az elemzésekhez, mik a jogi vonatkozások. Bár maga is szóvá teszi, hogy sok kutatás nem publikálja a kutatási adatokat és szoftvereket, Manovich sem jeleskedik abban, hogy az olvasó számára ezeket elérhetővé tegye (holott néhány szoftverprojektje elérhető a *GitHub*on).[10] Végezetül – és tudományterületünk szempontjából ez a legérzékenyebb pont – úgy tűnik, félreismeri a bölcsészet teljesebb spektrumát, és emiatt sokszor félreinterpretálja a bölcsészet általános célkitűzéseit, módszereit. Például: „A digitális bölcsészet nagyrészt ignorálta a népi digitális kultúra tanulmányozásának lehetőségeit, mivel, mint korábban kifejtettem, a hivatásos és magaskultúra tanulmányozásának hagyományos bölcsész paradigmáját követi." (55.) Csak néhány hazai példa ennek cáfolatára: a népdalgyűjtés keretében közel száz éve két elkülönülő szemlélettel folyik a gyűjtés: az egyikben esztétikai kategóriák érvényesülnek („csak tiszta forrásból"), a másikban viszont kulturális antropológiaiak („mit énekel a nép"). A Régi Magyar Költők Tára[11] (az '50-es évektől) vagy a RPHA[12] (a '70-es évektől) esztétikai értékre való tekintet nélkül minden magyar verset számba vesz és elemez, az előbbi egyik leágazása pedig, a közköltészet[13] hívószavával, kifejezetten az anonim alkotásokkal teszi ezt. Az 1994-ben elhunyt Kunt Ernő egyik kutatási területe a paraszti használatú fényképekre irányult, de a vizuális antropológia tudományága[14] (amelynek egyetemi tanszékét ő

[10] Néhány ilyen, Manovich vagy munkatársai által készített szoftver: *Software Studies* tool (kód: `https://github.com/softdetours/softwarestudies`, weboldal: `http://lab.softwarestudies.com/`). *ivpy: Iconographic Visualization Inside Computational Notebooks* (`https://github.com/damoncrockett/ivpy`). *ImagePlot* visualization software: explore patterns in large image collections (kód: `https://github.com/culturevis/imageplot`, weboldal: `http://lab.softwarestudies.com/p/imageplot.html`). ImageMontage plugin for *ImageJ* (kód: `https://github.com/culturevis/imagemontage`, leírás: `https://imagej.nih.gov/ij/plugins/image-montage/index.html`). Ezek a szoftverek a 2010-es évek első felében készültek.

[11] Hozzáférés: 2022.04.05, `https://hu.wikipedia.org/wiki/R%C3%A9gi_Magyar_K%C3%B6lt%C5%91k_T%C3%A1ra`.

[12] Horváth Iván és munkatársai: *Répertoire de la poésie hongroise ancienne (RPHA) A régi magyar vers leltára a kezdetektől 1600-ig*. 7.3. kiadás, `https://f-book.com/rpha/v7/index.php`.

[13] Hozzáférés: 2022.04.05, `https://szovegtar.iti.mta.hu/hu/sorozatok/rmkt-xviii-szazad/`.

[14] Hozzáférés: 2022.04.05, `https://hu.wikipedia.org/wiki/Vizu%C3%A1lis_antropol%C3%B3gia`.

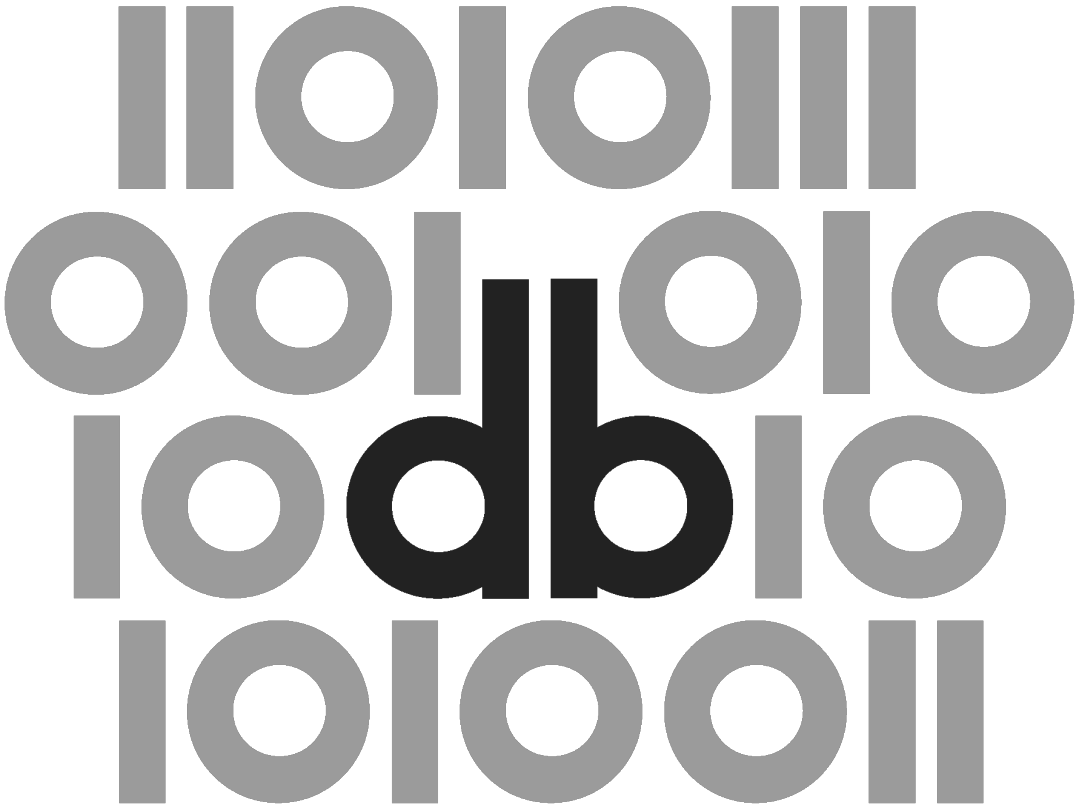
szervezte meg) szintén a mindennapi vizuális médiumok vizsgálatával foglalkozik. A Néprajzi Múzeum 2003-ban indult MaDok programja a múzeumi jelenkutatást célul tűzve a „kortárs tárgyi világ megőrzésérére és a jelenkor múzeumi dokumentációjára helyezi a hangsúlyt”,[15] amely már megfogalmazását tekintve is közel áll Manovich elképzeléséhez.

Mindezen felsorolt problémák nem kisebbítik a könyv érdemeit, hiszen az egy vállaltan szubjektív útinapló, amelyből mindenki sok ötletet meríthet saját adatkutatásában. El tudnék képzelni olyan együttműködést, amelynek keretében közgyűjtemények valamilyen, az adatvédelmet figyelembe vevő módon, a jelenlegi kísérleteknél[16] hatékonyabban őriznék meg a web Manovich által elemzett szegmenseit, illetve remélem, hogy a könyv hatására születik majd egy kevésbé szubjektív, a kulturális adatelemzés szélesebb körét átfogó, kézikönyv jellegű kiadvány.

[15] Hozzáférés: 2022.04.05, https://neprajz.hu/madok/muzeumi-jelenkutatas/madok-program.html.

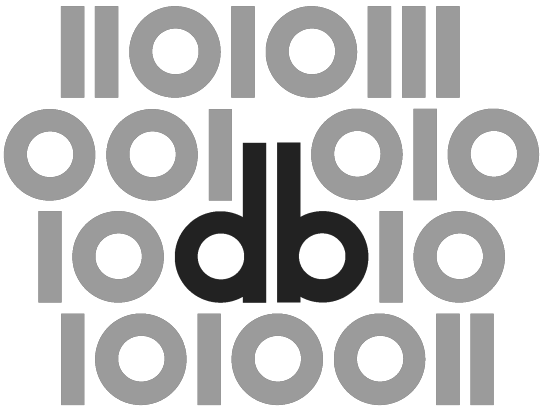
[16] Például a Kongresszusi Könyvtár Twitter-archívumáról lásd Axel Bruns, „The Library of Congress Twitter Archive: A Failure of Historic Proportions,” in *DMRC at large*, 2018. jan. 2, https://medium.com/dmrc-at-large/the-library-of-congress-twitter-archive-a-failure--of-historic-proportions-6dc1c3bc9e2c.

 10.31400/dh-hun.2021.4.3510



Digitális Bölcsészet
2021., negyedik szám

<DIGITÁLIS BÖLCSÉSZET>



4 (2021)

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.2021.4

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.

Ez a mű a Creative Commons *Nevezd meg! – Ne add el! – Így add tovább! 2.5 Magyarország Licenc* (`http://creativecommons.org/licenses/by-nc-sa/2.5/hu/`) feltételeinek megfelelően felhasználható.

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

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

<MŰHELY>

Péter Róbert 0000-0002-7972-4751

Szegedi Tudományegyetem

robert.peter@ieas-szeged.hu

Szántó Zsolt 0000-0002-8924-206X

Szegedi Tudományegyetem

szantozs@inf.u-szeged.hu

Bilicki Vilmos 0000-0002-7793-2661

Szegedi Tudományegyetem

bilickiv@inf.u-szeged.hu

Berend Gábor 0000-0002-3845-4978

Szegedi Tudományegyetem

berendg@inf.u-szeged.hu

Az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) többnyelvű kutatási eszköz bemutatása

E dolgozat célja, hogy bemutassa az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) többnyelvű kutatási eszköz működéséhez kapcsolódó munkafolyamatot és a különböző elemzőfunkciókat. A webes alkalmazás segítségével nagy mennyiségű metaadatot és szöveget lehet feldolgozni és kritikusan elemezni adatvezérelt, mesterséges intelligenciával és természetesnyelv-feldolgozásos technológiákkal támogatott módszerekkel és eszközökkel. Az AVOBMAT szöveg- és adatbányászati eszköz újdonságai a következők: (i) számos nyelven képes előfeldolgozni, (szemantikusan) gazdagítani és elemezni metaadatokat és szövegeket; (ii) a beépített funkciók lehetőséget biztosítanak a szoros és távoli olvasásra egyaránt; (iii) egy felhasználóbarát, interaktív grafikus felületen integrál metaadat és szövegelemzéssel kapcsolatos kutatási eszközöket. A platformfüggetlen alkalmazás elsősorban olyan felhasználók számára lett kifejlesztve, akik nem rendelkeznek programozási ismeretekkel. Az egyszerűen használható felület interaktív paraméterbeállítást és vezérlést biztosít a normalizálást is támogató előfeldolgozástól az analitikai szakaszokig.

A felhasználók interaktív módon kísérletezhetnek az elemzések különböző beállításaival a munkafolyamat során. Ezáltal az AVOBMAT segít felismerni a számítógépes szöveg- és adatelemzés episztemológiai kihívásait, korlátait és erősségeit, valamint kritikus módon értelmezni az alkalmazott módszereket és eredményeket.

Kulcsszavak:
szöveg- és adatbányászat, természetesnyelv-feldolgozás, többnyelvű kutatási eszköz, metaadat, szemantikus adatgazdagítás

1. Bevezetés

Az elmúlt két évtizedben hatalmas mennyiségű nyomtatott forrást digitalizáltak és kódoltak eltérő minőségben és módokon. Ezek a digitális anyagok és hozzájuk tartozó bibliográfiai adatok nyílt hozzáférésű vagy előfizetést igénylő adatbázisokban korlátozott módon elérhetőek és kereshetőek. Temérdek – szövegek, metaadatok tárolására és keresésére használható – adatbázis jött létre (és tűnt el) az elmúlt időszakban,[1] de ezek az alkalmazások ritkán megfelelőek kutatási kérdések megválaszolására.[2] Emellett a főként angol, német és francia nyelvű dokumentumok vizsgálatára finomhangolt szoftverek is problémákat, nehézségeket okoznak a nem világnyelveken írott szövegekkel foglalkozó (digitális) bölcsészeti kutatások során. Ezeket a kihívásokat felismerve számos európai országban – jelentős részben a Digital Research Infrastructure for the Arts and Humanities (DARIAH) támogatásával – az elmúlt években kezdték kiépíteni azokat a digitális bölcsészeti kutatási infrastruktúrákat, amelyek lehetőséget biztosítanak a digitális tartalmak feltáró – kutatási problémák megoldását elősegítő – elemzéséhez, megosztásához és megőrzéséhez.[3] Többek között ilyen elképzelések

[1] Mivel a digitális bölcsészeti projektek jelentős része pályázati forrásból valósul meg, a pályázatok lezárását követően számos esetben nincs lehetőség a létrejött adatbázisok és szoftverek frissítésére, valamint ezek szervereken való működtetésére és tárolására. Ez az egyik fő oka annak, hogy számos digitális bölcsészeti eszköz és adatbázis csak korlátozott ideig elérhető. A problémát felismerve az elmúlt néhány évben kiírt európai uniós és hazai pályázatok különös hangsúlyt fektetnek a fenntarthatóságra és a kutatási infrastruktúrák kiépítésére.

[2] John Bradley, „Digital Tools in the Humanities: Some Fundamental Provocations?," *Digital Scholarship in the Humanities* 34, 1. sz. (2019): 13–20, https://doi.org/10.1093/llc/fqy033; Marijn Koolen, Jasmijn van Gorp and Jacco van Ossenbruggen, „Toward a Model for Digital Tool Criticism: Reflection as Integrative Practice," *Digital Scholarship in the Humanities* 34, 2. sz. (2019): 368–385, https://doi.org/10.1093/llc/fqy048; John Unsworth, „Scholarly Primitives: What Methods Do Humanities Researchers Have in Common, and How Might Our Tools Reflect This?" hozzáférés: 2021.12.15, https://johnunsworth.name/Kings.5–00/primitives.html.

[3] Maria Ågren, Claudine Moulin, Marko Tadic, Julianne Nyhan, Arianna Ciula, Margaret Kelleher, Elmar Mittler, Andrea Bozzi and Kristin Kuutma, *Science Policy Briefing: Research Infrastructures in the Digital Humanities* (Strasbourg: European Science Foundation, 2011), https://www.esf.org/fileadmin/user_upload/esf/RI_DigitalHumanities_B42_2011.pdf. Németországban, Finnországban és Lengyelországban 2021-ben indultak olyan projektek, melyek célja digitális bölcsészeti

 DOI 10.31400/dh-hun.2021.4.3530

motiválták az AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) kutatási eszköz – önmagában nem infrastruktúra – fejlesztését, amely vállalkozás 2017-ben kezdődött a Szegedi Tudományegyetemen.

A dolgozat célja, hogy bemutassa az AVOBMAT többnyelvű kutatási eszköz működéséhez kapcsolódó munkafolyamatot és a különböző elemző funkciókat.[4] A programozási ismereteket nem igénylő webes alkalmazás segítségével nagy mennyiségű metaadatot és szöveget lehet feldolgozni és kritikusan elemezni korszerű, adatvezérelt és természetesnyelv-feldolgozásos technológiákkal támogatott módszerekkel és eszközökkel.

Első lépésben ismertetjük és összehasonlítjuk az AVOBMAT-hoz funkcionalitásában hasonló alkalmazásokat és bemutatjuk az AVOBMAT újdonságait. A következő fejezet az elemezni kívánt dokumentumok előfeldolgozásáról és a különböző feltöltési opciókról ad áttekintést. Ezután szemléltetjük, milyen módokon kereshetünk a feltöltött dokumentumok metaadatai és szövegei között. A keresési funkciók segítségével szűkíthetjük a korpuszunkat és meghatározhatunk egy alkorpuszt, amelyen a különböző metaadat- és szövegelemzéseket elvégezhetjük. A dolgozat következő részében a metaadat-elemzési lehetőségek és ezekhez kapcsolódó vizualizációk kerülnek bemutatásra konkrét példákon keresztül. Majd a tanulmány leghosszabb fejezete betekintést nyújt a szövegek tartalmi vizsgálatára vonatkozó paraméterezhető elemzők (pl. témamodellezés, szóstatisztikai vizsgálatok, névelem-felismerés) működésébe. Az

kutatási infrastruktúrák létrehozása. Vannak olyan futó projektek is, melyek nemzetközi kutatási infrastruktúrát szeretnének létrehozni a digitális bölcsészet egy adott részterületén. Lásd például a Computational Literary Studies Infrastructure-t, hozzáférés: 2021.12.15, `https://clsinfra.io/`. Hazánkban jelenleg nincs digitális bölcsészeti kutatási infrastruktúra, bár 2021-ben nálunk is megfogalmazódtak ilyen irányú tervek (pl. Digital Humanities Platform: dHUpla) a Petőfi Irodalmi Múzeum Digitális Bölcsészeti Központjában, valamint a Kulturális Örökség Nemzeti Laboratóriumban, de konkrét, kutatók által is használható infrastruktúrák még nem állnak rendelkezésre. dHUpla, hozzáférés: 202112.15, `https://dhupla.hu/` és DH-LAB, hozzáférés: 2021.12.15, `https://dh--lab.hu/`.

[4] Az AVOBMAT korábbi verzióját az alábbi publikációban és poszteren mutattuk be: Róbert Péter, Zsolt Szántó, József Seres, Vilmos Bilicki and Gábor Berend, „AVOBMAT: A Digital Toolkit for Analysing and Visualizing Bibliographic Metadata and Texts," in Berend Gábor, Gosztolya Gábor és Vincze Veronika, szerk., *XVI. Magyar Számítógépes Nyelvészeti Konferencia,* 43–55 (Szeged: Szegedi Tudományegyetem, Informatikai Intézet, 2020), `http://acta.bibl.u-szeged.hu/67682/`; Zsolt Szántó, József Seres, Vilmos Bilicki, Bendegúz M. Bendicsek, Gábor Berend and Róbert Péter, „Introducing the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) Multilingual Research Tool," in *DARIAH: Virtual Annual Event 2020. Poster Exhibition,* hozzáférés: 2021.12.15, `https://www.virtualdariah2020.dariah.eu/posters/#lightbox--gallery-1/7/`. Az AVOBMAT fejlesztését részben az EFOP-3.6.1-16-2016-00008 és az EFOP-3.6.3-VEKOP-16-2017-0002 azonosítószámú pályázatok támogatták. Az előbbi pályázat keretében jött létre a TANIT *(Text ANalysIs Tools)* morfológiai elemző (`http://dighum.bibl.u-szeged.hu/tanit/`). Labádi Gergely, Farkas Richárd, Nagy Roland és Péter Róbert, „TANIT: Magyar nyelvű szövegeket elemző eszköz összehasonlító digitális bölcsészeti feladatokhoz," in Vincze Veronika, szerk., *XIV. Magyar Számítógépes Nyelvészeti Konferencia*, 450–455 (Szeged: JATEPress, 2018), `http://real.mtak.hu/86149/1/teljesB5-460-465.pdf`. Köszönettel tartozunk Ficand Tamás, Simon Gábor, Dér Gergely, Seres József és Bendicsek M. Bendegúz hallgatóknak, akik részt vettek az AVOBMAT fejlesztésében, valamint Kokas Károly, Sándor Ákos, Nagy Gyula és Erdődi Zoltán (SZTE Klebelsberg Könyvtár) informatikus könyvtárosoknak, akik biztosították a technikai hátteret és számos adatbázist az említett szoftverek teszteléséhez.

összegzés és a fejlesztési tervek felvázolása előtt röviden szólunk a jogosultságkezelésről, valamint a felhasználók által nem látható adminisztrátori felület által kínált lehetőségekről.

2. Hasonló alkalmazások és újdonságok

Az olyan digitális bölcsészethez köthető kereskedelmi termékekben, mint a *Gale Digital Scholar Lab*[5] (az ára miatt csak a leggazdagabb egyetemeken és kutatóintézetekben érhető el), az egyéni vagy könyvtári adatbázisok, a(z előre betanított) nyelvi modellek és szótárak nem tölthetők fel. A szintén magas előfizetési áron megrendelhető *ProQuest Text* and *Data Mining Studio*[6] alkalmazás használatához Python és R programozási ismeretek szükségesek. A legtöbb jelenleg elérhető webes szövegelemző eszköz, mint például az angol nyelvű dokumentumok feldolgozására fókuszáló *Voyant Tools*[7] vagy a *Topics Explorer,*[8] nem tud megbirkózni nagy szövegkorpuszokkal. A szöveg- és adatelemzésre alkalmazható Python- és R-könyvtárak konfigurációs beállításaihoz képest a viszonylag kevés böngészőalapú digitális bölcsészettudományi alkalmazásban csak korlátozott számú konfigurációs paramétert lehet beállítani az egyes elemzők esetében. A *Paper Machine*[9] (a *Zotero* hivatkozáskezelő szoftver bővítménye) egykor ötvözte az alapvető bibliográfiai metaadatokat (dátum, cím és kiadási hely) és témamodellezési elemzőket, de ez már nem kompatibilis a *Zotero* jelenlegi, 5-ös verziójával. Az *Interactive Text Mining Suite*[10] webes alkalmazás előfeldolgozza a TXT és PDF formátumú szövegeket, amelyeken klaszter-, téma- és gyakoriságelemzéseket végez, de nagyon kevés metaadatmezőt (szerző, év, cím és kategória) tud kezelni, valamint a metaadatokat nem képes önállóan elemezni. A böngészőalapú alkalmazások közül a *Lexos*[11] kínálja a legtöbb eszközt a szövegek (TXT, HTML és XML) előfeldolgozására és szegmentálására. A *Lexos* tokenizálja a szövegeket, azonosítja az n-gramokat, statisztikai összefoglalókat készít, vizualizálja az eredményeket különböző típusú szófelhők, dendrogramok és konszenzusfák segítségével. A szövegek összehasonlítása mellett, a MALLET[12] által generált adatokon alapuló „témafelhőt" is lehet a *Lexos*szal készíteni,

[5] *Gale Digital Scholar Lab,* hozzáférés: 2021.12.15, https://www.gale.com/primary-sources/digital-scholar-lab.

[6] *ProQuest Text and Data Mining Studio,* hozzáférés: 2021.12.15, https://about.proquest.com/en/products-services/TDM-Studio/.

[7] Stéfan Sinclair and Geoffrey Rockwell, *Voyant Tools*, hozzáférés: 2021.12.15, http://voyant-tools.org/.

[8] *TopicsExplorer*, hozzáférés: 2021.12.15, https://dariah-de.github.io/TopicsExplorer/.

[9] *Paper Machines*, hozzáférés: 2021.12.15, http://papermachines.org/.

[10] *Interactive Text Mining Suite,* hozzáférés: 2021.12.15, https://languagevariationsuite.wordpress.com/2016/03/18/interactive-text-mining-suite-itms/; Olga Scrivner and Jefferson Davis, „Interactive Text Mining Suite: Data Visualization for Literary Studies," in Thierry Declerck and Sandra Kübler, eds., *Proceedings of the Workshop on Corpora in the Digital Humanities*, 29–38 (Bloomington, 2017), http://ceur-ws.org/Vol-1786/scrivner.pdf.

[11] *Lexos*, hozzáférés: 2021.12.15, http://lexos.wheatoncollege.edu/upload; Scott Kleinman, Mark D. LeBlanc, Michael D. C. Drout, and Weiqi Feng, *Lexos. v4.0*, hozzáférés: 2021.12.15, https://github.com/WheatonCS/Lexos/.

[12] Andrew Kachites McCallum, „MALLET: A Machine Learning for Language Toolkit," hozzáférés: 2021.12.15, http://mallet.cs.umass.edu.

 DOI 10.31400/dh-hun.2021.4.3530

de hagyományos (pl. Latent Dirichlet Allocation alapú) témamodellezést nem lehet ezzel végezni. Az eddig említett ingyenesen hozzáférhető eszközök egyike sem képes bibliográfiai adatok és szövegek (szemantikus) gazdagítására és elemzésére, valamint a szövegelemzéshez feltöltött digitális gyűjtemények szűrésére metaadatok vagy (teljes szövegű) kulcsszavas keresések segítségével, beleértve a közelítő *(fuzzy),* szószomszédsági *(proximity)* és parancssori lekérdezéseket. Az eleddig felsorolt szövegelemzési funkciók jelentős része (és számos egyéb, szövegek összehasonlítására alkalmas elemző) megtalálható az *impresso: Media Monitoring of the Past,*[13] újságok elemzésére specializálódott kutatási eszközben. Ennek a korszerű szövegbányászati eszköznek az a hátránya, hogy csak fix, főleg svájci újságkorpuszokon működik. Az ismertetett alkalmazásokból merítettünk ötleteket és számos elemző funkciót beépítettünk az AVOBMAT-ba is.

Az AVOBMAT kutatási eszköz újdonságai a következők: (i) számos nyelven képes előfeldolgozni, (szemantikusan) gazdagítani és elemezni metaadatokat és szövegeket; (ii) a beépített funkciók lehetőséget biztosítanak a szoros *(close)* és távoli *(distant)* olvasásra egyaránt; (iii) egy felhasználóbarát, programozási tudást nem igénylő grafikus felületen integrál metaadat- és szövegelemzéssel kapcsolatos kutatási eszközöket. Az interaktív felületen a legtöbb esetben ki-be kapcsolhatóak a megjelenített metaadatmezők, valamint módosíthatóak az elemzési paraméterek és ezután újrafuttathatók az elemzések. A távoli és szoros olvasási megközelítések elemzési keretrendszerünkben történő kombinálásával a felhasználók új perspektívákat azonosíthatnak a bibliográfiai adatok és a szövegelemzés kapcsán, valamint eleddig ismeretlen összefüggéseket fedezhetnek fel a digitális gyűjteményekben. Az AVOBMAT lehetővé teszi, hogy a felhasználó által tetszőlegesen konfigurálható előfeldolgozás után feltöltött adatbázisokat különböző típusú metaadatok és teljes szöveges keresések alapján szűrjék, és a szűrt alkorpuszon bibliográfiai, hálózati, valamint természetesnyelv-feldolgozással kapcsolatos elemzéseket végezzenek. Az eddigi digitális módszereket használó kutatások nem igazán aknázták ki a bibliográfiai (meta)adatok elemzésében rejlő lehetőségeket. A legújabb kutatások igazolják, hogy a szövegbányászathoz hasonlóan a bibliográfiai adatok (úgyis mint *big data*) kritikus vizsgálata is számos új felismerést nyújthat, eddig figyelmen kívül hagyott mintákat és trendeket tárhat fel, új típusú bizonyítékokkal és eredményekkel szolgálhat, valamint megkérdőjelezhet, finomíthat régi hipotéziseket a bölcsészettudományok területén.[14] Például a 18. századi tanulmá-

[13] impresso. Media Monitoring of the Past, hozzáférés: 2021.12.15, `https://impresso-project.ch`; Matteo Romanello, Maud Ehrmann, Simon Clematide and Daniele Guido, „The Impresso System Architecture in a Nutshell," *Technical Report, EuropeanaTech Insights*, 16 (2020), `https://pro.europeana.eu/page/issue-16-newspapers#the-impresso-system-architecture-in-a--nutshell`.

[14] Iraklis Varlamis and George Tsatsaronis, „Visualizing Bibliographic Databases as Graphs and Mining Potential Research Synergies," in Randall Bilof, ed., *2011 International Conference on Advances in Social Networks Analysis and Mining*, 53–60 (Piscatawy, NJ: The Institute of Electrical and Electronics Engineers, 2011), `https://doi.org/10.1109/ASONAM.2011.52`; Róbert Péter, „Researching (British Digital) Press Archives with New Quantitative Methods," *Hungarian Journal for English and American Studies* 17, 2. sz. (2011): 283–300, https://www.jstor.org/stable/43487818; Katrina Fenlon, Miles Efron and Peter Organisciak, „Tooling the Aggregator's Workbench: Metadata Visualization through Statistical Text Analysis," *Proceedings of the American Society for Information Science and Technology* 49, 1. sz. (2012): 1–10, `https://doi.org/10.1002/meet.14504901161`; Franco Mo-

nyok kapcsán ezekre kiváló példákat találunk többek között Mikko Tolonen, Simon Burrows, Dan Edelstein, Mark Towsey és Alicia Montoya vezette kutatócsoportok publikációiban.[15]

3. A konfigurálható előfeldolgozás és feltöltés

Az előfeldolgozási fázisban a felhasználó konfigurálhatja az egyes elemzőket, valamint a metaadat- és szemantikus gazdagítást végző eszközöket. A munkafolyamat első lépése az a feltöltendő szövegekre épülő automatikus nyelvdetekció, melynek eredményét az elemzések során figyelembe veszi a program. Ehhez a művelethez a *lang-*

retti, *Distant Reading* (London; New York: Verso, 2013); Andrew Prescott, „Bibliographic Records as Humanities Big Data," in Xiaohua Tony Hu et al., eds., *2013 IEEE International Conference on Big Data,* 55–58 (Piscatawy, NJ, 2013), https://doi.org/10.1109/BigData.2013.6691670; Matthew L. Jockers, *Macroanalysis: Digital Methods and Literary History* (Champaign, IL: University of Illinois Press, 2013); Andrew Prescott, *Big Data in the Arts and Humanities: Some Arts and Humanities Research Council Projects* ([Glasgow]: University of Glasgow, 2015); Shawn Graham, Ian Milligan and Scott Weingart, *Exploring Big Historical Data: the Historian's Macroscope* (London: Imperial College Press, 2016), https://doi.org/10.1142/p981; Jean-Philippe Moreux, „Innovative Approaches of Historical Newspapers: Data Mining, Data Visualization, Semantic Enrichment: Facilitating Access for various Profiles of Users," in *IFLA News Media Section, Lexington, August 2016, At Lexington, USA,* 1–16 (Lexington: IFLA, IFLA, 2016), https://hal-bnf.archives-ouvertes.fr/hal-01389455/document; Giovanni Schiuma and Daniela Carlucci, *Big Data in the Arts and Humanities: Theory and Practice* (Boca Raton, FL: Taylor and Francis, 2018); DARIAH Bibliographical Data Working Group, „An Analysis of the Current Bibliographical Data Landscape in the Humanities. The Joint Bibliodata Agendas of Public Stakeholders", (2022), https://doi.org/10.5281/zenodo.6559857.

[15] Mikko Tolonen, Leo Lahti and Niko Ilomäki, „A Quantitative Study of History in the English Short-Title Catalogue (ESTC), 1470-1800," *Liber Quarterly* 25, 2. sz. (2015): 87–116, https://doi.org/10.18352/lq.10112; Péter Róbert, „Digitális és módszertani fordulat a sajtókutatásban: A 17–18. századi magyar vonatkozású angol újságcikkek »távolságtartó olvasása«," *Aetas* 29, 1. sz. (2015), 5–30, http://acta.bibl.u-szeged.hu/35222/; Dan Edelstein, Paula Findlen, Giovanna Ceserani, Caroline Winterer and Nicole Coleman, „Historical Research in a Digital Age: Reflections from the Mapping the Republic of Letters Project," *American Historical Review* 122, 2. sz. (2017): 401–424, https://academic.oup.com/ahr/article/122/2/400/3096208, https://doi.org/10.1093/ahr/122.2.400; Simon Burrows, *The French Book Trade in Enlightenment Europe II: Enlightenment Bestsellers* (London: Bloomsbury Academic, 2018); Mark Towsey, „Book Use and Sociability in Lost Libraries of the Eighteenth Century: Towards a Union Catalogue," in Flavis Bruni and Andrew Pettegree, eds., *Lost Books: Reconstructing the Print World of Pre-Industrial Europe*, 414–438 (Leiden: Brill, 2016), https://doi.org/10.1163/9789004311824_021; Leo Lahti, Jani Marjanen, Hege Roivainen and Mikko Tolonen, „Bibliographic Data Science and the History of the Book (c. 1500–1800)," *Cataloging & Classification Quarterly* 57, 1. sz. (2019): 5–23, https://doi.org/10.1080/01639374.2018.1543747; Mark J. Hill, Ville Vaara, Tanja Säily, Leo Lahti and Mikko Tolonen, „Reconstructing Intellectual Networks: From the ESTC's Bibliographic Metadata to Historical Material," in *Proceedings of the Digital Humanities in the Nordic Countries,* 201–219 (Copenhagen: CEUR-WS.org, 2019), http://ceur-ws.org/Vol-2364/19_paper.pdf; Alicia C. Montoya, „Enlightenment? What Enlightenment? Reflections on Half a Million Books (British, French and Dutch Private Libraries, 1665 – 1830)," *Eighteenth-Century Studies* 54, 2. sz. (2021): 909–934, https://doi.org/10.1353/ecs.2021.0097; Simon Burrows and Terhi Nurmikko-Fuller, „Charting Cultural History Through Historical Bibliometric Research: Methods, Concepts, Challenges, Results," in Kristen Schuster and Stuart Dunn, eds., *Routledge International Handbook of Research Methods in Digital Humanities,* 109–124 (New York: Routledge, 2020), https://doi.org/10.4324/9780429777028-9.

 DOI 10.31400/dh-hun.2021.4.3530

detect nyelvfelismerő programot használjuk.[16] Ha egynyelvű korpuszt elemzünk, az automatikus nyelvfelismerés helyett megadhatjuk az elemezni kívánt szövegek nyelvét: 61 nyelv közül tudunk választani egy gördülősáv segítségével. Az automatikus nyelvdetekció gazdagítja és pontosíthatja az analóg módon megadott, dokumentumok nyelvére vonatkozó metaadatokat.[17]

Ha nem szeretnénk a teljes szövegállományt elemezni, akkor a kontextusszűrő *(Context filter)* funkció segítségével megadhatunk kulcsszavakat, valamint a kulcsszavaktól balra és jobbra (külön-külön) található szavak számát. A későbbi elemzések során az AVOBMAT csak az így definiált szövegdobozokban található szavakat elemzi, a többi szöveget eltávolítja a dokumentumokból. Ez a funkció hasznos lehet például kisebb szövegrészek elkülönítésére, így cikkek szerint nem szegmentált újságkorpuszok kezdeti feldolgozására is.

A helyettesítés *(replace)* funkció segítségével az optikai szövegfelismerésből (Optical Character Recognition: OCR) adódó hibákat javíthatunk, összevonhatunk szinonimákat, vagy modernizálhatjuk a régi, nem standardizált helyesírást használó szövegeinket. A cserepárok megadása során reguláris kifejezéseket is használhatunk. Így például lehetőségünk van különleges karakterek törlésére, rövidítések feloldására, az elválasztott szavak összevonására, melyek fontos lépések a szövegtisztítás és normalizálás során.

A metaadat-gazdagítás magában foglalja a szövegek nyelvének detektálását, valamint a szerzők nemének automatikus azonosítását. Az utóbbi célra a *gender-guesser* nevű Python-csomag általunk továbbfejlesztett verzióját használjuk. Az androgün keresztnevek létezése miatt a nemek azonosítása nem mindig kivitelezhető egyértelműen. A dokumentumok szerzőit férfi, női vagy ismeretlen (például ha csak a keresztnév rövidítése adott) kategóriákba soroljuk. Külön metainformációként kezeljük, ha egy dokumentumnak egyáltalán nincs szerzője. Annak érdekében, hogy csökkentsük azon esetek számát, amikor nem tudjuk megállapítani a szerző nemét, a dokumentum nyelvét is bevonjuk a döntéshozatali folyamatba. Azt az egyszerűsítő feltételezést vesszük alapul, hogy a szerzői nevek ugyanazon a nyelven szerepelnek, mint maguk a dokumentumok. E feltételezés alapján tudjuk kezelni az apofóniát, például Kovács Imréné magyar szerző esetén következtethetünk arra, hogy az Imréné név az Imre névből származik, tehát női szerzőséget rendelünk hozzá. A nyelvi információk felhasználásával csökkenthetjük azt a bizonytalanságot is, amely abból ered, hogy ugyanaz a keresztnév különböző nyelvekben különböző nemű személyekre utalhat. A programba beépített női és férfi keresztnév-adatbázisok (*gender-guesser* csomag) tartalmát a felhasználó bővítheti saját női és férfi névlistáival is. A magyar nyelv esetén

[16] *Langdetect*, hozzáférés: 2021.12.15, `https://pypi.org/project/langdetect/`.

[17] A Szegedi Tudományegyetem *Egyetemi Kiadványok* nevű repozitóriumában található Kurdy Fehér János *Oleskeluharjoituksia* című finn versét magyar nyelvű versként tünteti fel a katalógus. Kurdy Fehér János, „Oleskeluharjoituksia," *Gondolat-jel* (1993), 1–2. sz., 22, `http://acta.bibl.u-szeged.hu/11488/`. Az AVOBMAT számos – a katalógusban a dokumentumok nyelvére vonatkozó metaadatmezőben nem szereplő – nyelvű dokumentumot azonosított. Itt meg kell jegyezni, hogy az automatikus nyelvdetekció tévedhet rövid (pl. képaláírások) vagy rosszul OCR-ezett dokumentumok esetében. Az utóbbira jó példa a *Délmagyarország* 1945. február 21-i száma, melynek OCR-ezett első oldala nem összefüggő szavakat, hanem ezek szóközökkel elválasztott betűsorát tartalmazza. *Délmagyarország*, 1945. febr. 21., 1, `http://dmarchiv.bibl.u-szeged.hu/10369/`.

például az ELKH Nyelvtudományi Kutatóközpont által anyakönyvi bejegyzésre alkalmasnak minősített utónevek jegyzékét[18] is használhatjuk erre a célra, de tetszőleges – a szoftveres elemzést pontosító – névlistákat is megadhatunk. A kiegészítésképpen megadott férfi és női nevek felülírják a program által valószínűsített nemi kategóriákat.[19]

A szövegelemző funkciókhoz köthető előfeldolgozási műveletek minden elemző esetén egyedileg konfigurálhatók. Opcionálisan beállítható a következő hat paraméter: (i) lemmatizálás (24 nyelven);[20] (ii) kisbetűsítés; (iii) számok; (iv) nem alfanumerikus karakterek; (v) írásjelek és (vi) stopszavak eltávolítása. A stopszavas és a punktuációs előfeldolgozás során a felhasználó is kiegészítheti az AVOBMAT-ba épített spaCy nyelvmodulokban található listákat. A stopszavak kiszűrésénél és a szótövesítés során a program figyelembe veszi az adott dokumentum automatikusan azonosított vagy manuálisan megadott nyelvét. Továbbá az n-gram elemző esetében a felhasználó megadhatja az azonosítandó szógramok hosszát (1–5). A lexikaidiverzitás-elemző segítségével nyolc metrika szerint tudjuk kiszámoltatni az egyes szövegek lexikai gazdagságát: Type-token ratio (TTR), Guiraud (Root TTR), Herdan (Log TTR), Mass TTR, Mean Segmental TTR (MSTTR), Moving Average TTR (MATTR), Measure of Textual Lexical Diversity (MTLD) és Hypergeometric Distribution Diversity (HDD). Az MSTTR és az MATTR esetében – a korábban említett előfeldolgozási paraméterek mellett – beállíthatjuk a „szövegablak" méretét (szószám) is. Itt fontos megjegyezni, hogy a különböző metrikák eltérő módon érzékenyek a szöveghosszra. Az utóbbi szempontból az MSTTR-, HDD- és MTLD-kalkulációk a legstabilabbak.[21]

[18] ELKH Nyelvtudományi Kutatóközpont, „ELKH Nyelvtudományi Kutatóközpont által anyakönyvi bejegyzésre alkalmasnak minősített utónevek jegyzéke," hozzáférés: 2021.12.15, http://www.nytud.mta.hu/oszt/nyelvmuvelo/utonevek/index.html.

[19] Ha az adatbázisunkban vannak hiányos szerzői nevek (pl. csak a vezetéknév adott), de a felhasználó tudja a szerző nemét, akkor ily módon is pontosíthatjuk a szerzői nevek felismerését. Például ha csak a Shakespeare név van megadva, akkor ezt az algoritmus ismeretlen nemű kategóriába sorolja. Ezt mi korrigálhatjuk, ha Shakespeare-t hozzáadjuk a férfi nevek listájához.

[20] A lemmatizáláshoz a spaCy nyelvmodelljeit és a *lemmagen* Python-csomagot használjuk. spaCy, hozzáférés: 2021.12.15, https://spacy.io/usage/models; Matjaž Juršic, et al., „Lemmagen: Multilingual Lemmatisation with Induced Ripple-down Rules," *Journal of Universal Computer Science* 16. 9 sz. (2010): 1190–1214; *Lemmagen*, hozzáférés: 2021.12.15, https://pypi.org/project/Lemmagen/.

[21] George Udny Yule, *The Statistical Study of Literary Vocabulary* (Cambridge: Cambridge University Press, 1944); Edward H. Simpson, „Measurement of Diversity," *Nature* 163 (1949): 688, https://doi.org/10.1038/163688a0; Gustav Herdan, „A New Derivation and Interpretation of Yule's 'Characteristic' *K*," *Zeitschrift für angewandte Mathematik und Physik* 6, 4. sz. (1955): 332–334, https://doi.org/10.1007/BF01587632; Heinz Dieter Maas, „Über den Zusammenhang zwischen Wortschatzumfang und Länge eines Textes," *Zeitschrift für Literaturwissenschaft und Linguistik* 2, 8 sz. (1972): 73–96; 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; Philip M. McCarthy and Scott Jarvis, „vocd: A Theoretical and Empirical Evaluation," *Language Testing* 24, 4. sz. (2007): 459–488, https://doi.org/10.1177/0265532207080767; Michael A. Covington and Joe D. McFall, „Cutting the Gordian Knot: the Moving-Average Type-Token Ratio (MATTR)," *Journal of Quantitative Linguistics* 17, 2. sz. (2010): 94–100, https://doi.org/10.1080/09296171003643098; Philip M. McCarthy and Scott. Jarvis, „MTLD, vocd-D, and HD-D: A Validation Study of Sophisticated Approaches to Lexical Diversity Assessment," *Behaviour Research Methods* 42, 2. sz. (2010): 381–392, https:

 DOI 10.31400/dh-hun.2021.4.3530

Minden elemző esetében vannak javasolt alapbeállítások, amelyeket a felhasználó tetszőlegesen módosíthat. Az egyes szövegelemzők ki és bekapcsolhatóak. Ha nincs szükségünk valamelyik elemzésre, akkor ily módon gyorsíthatjuk az előfeldolgozást. Az összes előfeldolgozás során megadott paramétert és adatot exportálhatjuk egy JSON-fájlba, amelyeket importálhatunk is a későbbi elemzések, kísérletezések során.

Az előfeldolgozási paraméterek megadása után feltölthetjük az adatbázisainkat többféle formátumban. Az AVOBMAT automatikusan importálja a *Zotero*-gyűjteményeket CSV- és RDF-formátumban (teljes szöveggel), valamint a számos könyvtárban használt EPrintses-adatbázisokat (EP3 XML a teljes szövegek URL-jeivel). A *Zotero* 20 különböző típusú bibliográfiai formátumot (pl. MARC, BibTex) tud importálni, amelyeket a felhasználók gyűjteménybe rendezhetnek. Ezeket a gyűjteményeket manuálisan vagy automatikusan tisztíthatják, bővíthetik új metaadatokat és szövegeket tartalmazó tételekkel.[22] A 87 bibliográfiai metaadatmezőt (pl. szerző, kiadó, publikáció cím) tartalmazó *Zotero* alapú CSV-struktúrát számos egyéb metaadatmezővel kiegészítettük (pl. könyvkereskedő, publikációk gyakorisága). A különböző bibliográfiai metaadatstandardok közötti különbségeket egyeztettük. Például az EP3 XML „Publication" mezője megegyezik a Zotero „Publication Title" mezőjével, így mindkettő egy közös „Elasticsearch" mezőbe kerül „publicationTitle" néven.[23]

Az adatbázisok feltölthetők egy egyszerű CSV-fájl segítségével is. A teljes szövegek hozzáadása többféleképpen történhet: (i) a szövegek rögzíthetők egy külön erre a célra létesített CSV-s mezőben; a szövegekre mutató (ii) relatív vagy (iii) internetes útvonalat is megadhatjuk egy másik mezőben. A második opció esetében a teljes szövegeket tartalmazó mappát és a metaadatokat tartalmazó CSV-fájlt tömörítve kell feltöltenünk. Az AVOBMAT minden olyan szövegformátumot tud importálni, amelyet az *Apache Tika*[24] program képes kezelni, mivel ez alakítja át egyszerű szöveggé a bemeneti fájlokat.

4. A keresés és kiválasztás

Az *Elasticsearch* motort használó alkalmazásban a kutatók kereshetnek a gazdagított bibliográfiai adatokban és az előfeldolgozott szövegekben fazettás, összetett és parancssori keresések segítségével. A fazettás keresésnél minden metaadatmező esetében megjelenik az értékkel nem rendelkező tételek (pl. a szerző nincs mindenhol

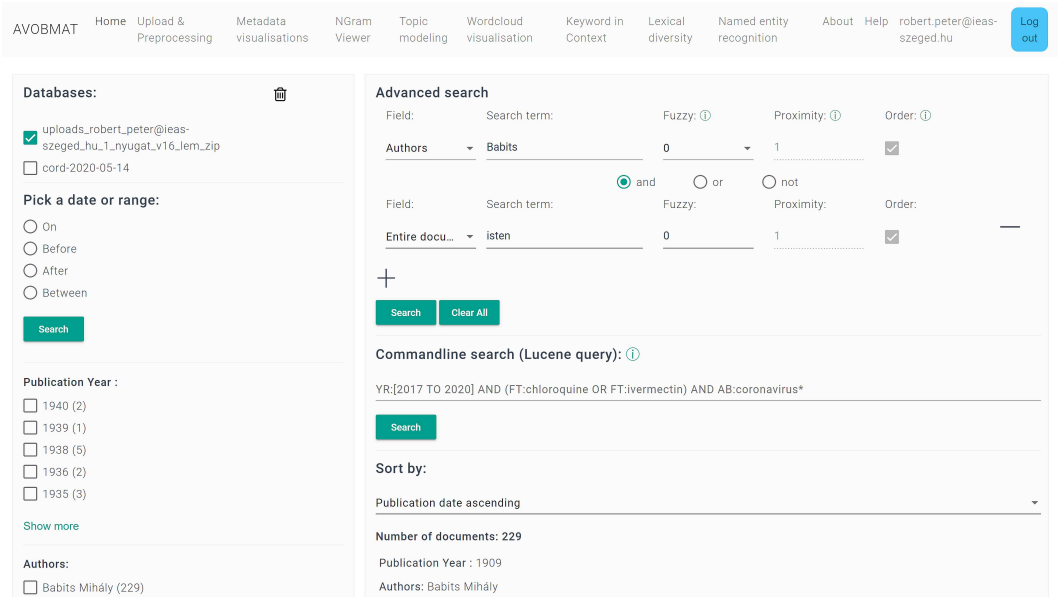
://doi.org/10.3758/BRM.42.2.381; Joan Torruella and Ramon Capsada, „Lexical Statistics and Tipological Structures: A Measure of Lexical Richness," *Procedia: Social and Behavioral Sciences* 95 (2013): 447–454, https://doi.org/10.1016/j.sbspro.2013.10.668; Kristopher Kyle, *Lexical diversity*, hozzáférés: 2021.12.15, https://github.com/kristopherkyle/lexical_diversity. Az AVOBMAT a fenti metrikákhoz tartozó értékek mellett jelzi az egyes szövegekhez tartozó szavak számát (token) valamint a különböző szóalakok számát (típus) is táblázatos formában.

[22] A metaadatok minőségellenőrzésével kapcsolatban lásd Király Péter publikációit és programkódjait. Például Péter Király und Rudolf Ungváry, „Bemerkungen zu der Qualitätsbewertung von MARC-21-Datensätzen," in Michael Franke-Maier, Anna Kasprzik, Andreas Ledl und Hans Schürmann, Hg., *Qualität in der Inhaltserschließung*, 177–228 (Berlin: De Gruyter Saur, 2021), https://doi.org/10.1515/9783110691597−011 és https://github.com/pkiraly.

[23] *Elasticsearch*, hozzáférés: 2021.12.15, https://www.elastic.co/.

[24] *Apache Tika*, hozzáférés: 2021.12.15, https://tika.apache.org/.

megadva) száma is egy külön metaadatmezőben *(missing_value)*, ami segíti a felhasználót az adatok és eredmények kritikus értelmezésében. Az AVOBMAT támogatja a paraméterezhető közelítő *(fuzzy)*, szószomszédsági *(proximity)* kereséseket, valamint a Lucene-szintaxist használó parancssori lekérdezéseket is. Így lehet használni például helyettesítő karaktereket (pl. ? vagy *) és reguláris kifejezéseket (melyeket *slash* közé kell tenni) is a kereséseknél.[25] Az utóbbi segítségével például egy szó összes ragozott alakjára is rákereshetünk. A paraméterezhető közelítő keresés különösen hasznos OCR-ezett dokumentumok esetén vagy régi szövegek elemzése során a nem standardizált helyesírás miatt. A szótávolság-keresés figyelembe veszi a megadott szótávolságban lévő szavak sorrendjét. Az összetett keresőben kombinálhatjuk a szótávolság- és a közelítő keresési funkciókat a Boole-operátorokkal (AND, OR, NOT) összekapcsolt keresésekkel. A metaadatmezők esetében csak azok jelennek meg a grafikus felületen, amelyeket egy adott adatbázis tartalmaz. A feltöltött adatbázisokat a keresés során egyszerű kijelöléssel lehet egyesíteni. A különböző keresési funkciók segítségével leszűkített korpuszon végezhetők el a metaadat- és szövegelemzések.



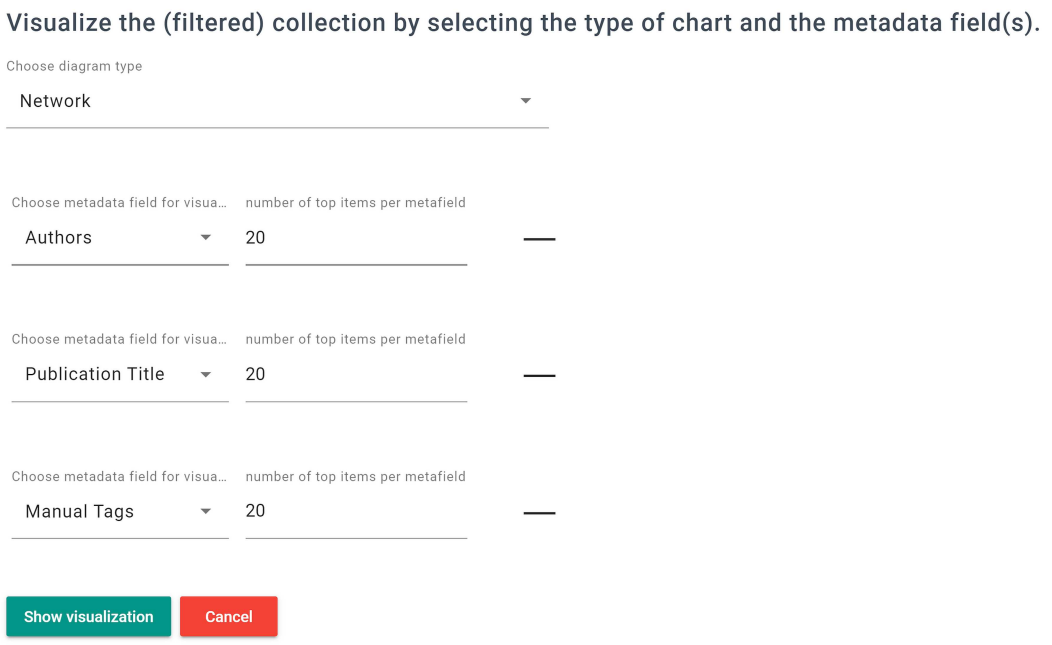
1. ábra. Az AVOBMAT grafikus felülete

[25] A keresési szintaxissal kapcsolatos dokumentáció az alábbi linken elérhető: `https://www.elastic.co/guide/en/elasticsearch/reference/7.17/query-dsl-query-string-query.html#query-string-syntax`.

 DOI 10.31400/dh-hun.2021.4.3530

5. Metaadat-elemzés és vizualizáció

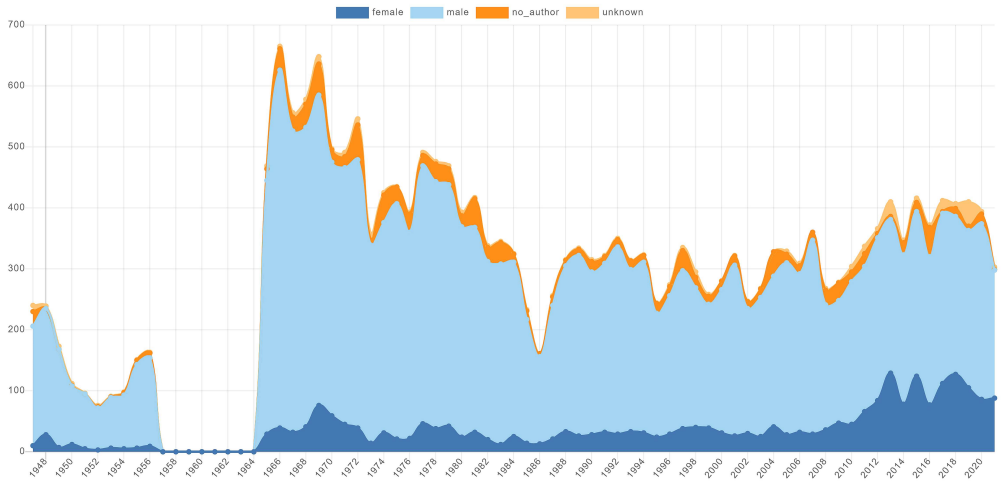
A felhasználók elemezhetik és vizualizálhatják a bibliográfiai adatokat (i) kronologikusan, vonal- és területdiagramokon, normalizált és aggregált formátumban;[26] (ii) interaktív hálózati elemzést készíthetnek legfeljebb három metaadatmező segítségével; (iii) a megadott paraméterek alapján tetszés szerinti kör-, sáv- és oszlopdiagramokat készíthetnek a bibliográfiai adatok felhasználásával. Az elemzők esetében a kutató határozza meg, mely metaadatmezőket szeretne elemezni és az adott mezőhöz tartozó adatsoron belül az első hány leggyakrabban előforduló tételt szeretne megjeleníttetni. Az adatpontok azért vannak külön ábrázolva, mert ezekre kattintva megjelennek a hozzájuk tartozó értékek. Az ábrákon található színmagyarázatok egyben szűrőként is funkcionálnak: minden vizualizáció esetében az egyes megjelenített metaadatok interaktív módon ki-be kapcsolhatóak, és ugyanez vonatkozik a hiányzó értékek *(missing_values)* és egyéb értékek *(other_values)* mezőkre is. Az utóbb azokra a „kimaradt" értékekre utal, melyeket a felhasználó dob el a paraméterezés során, amikor kiválasztja, hány leggyakrabban előforduló tételt szeretne ábrázolni egy metaadatmezőn belül. A diagramok egyes pontjaira kattintva a program megjeleníti az adott ponthoz tartozó értékeket (pl. név, szám, százalék).



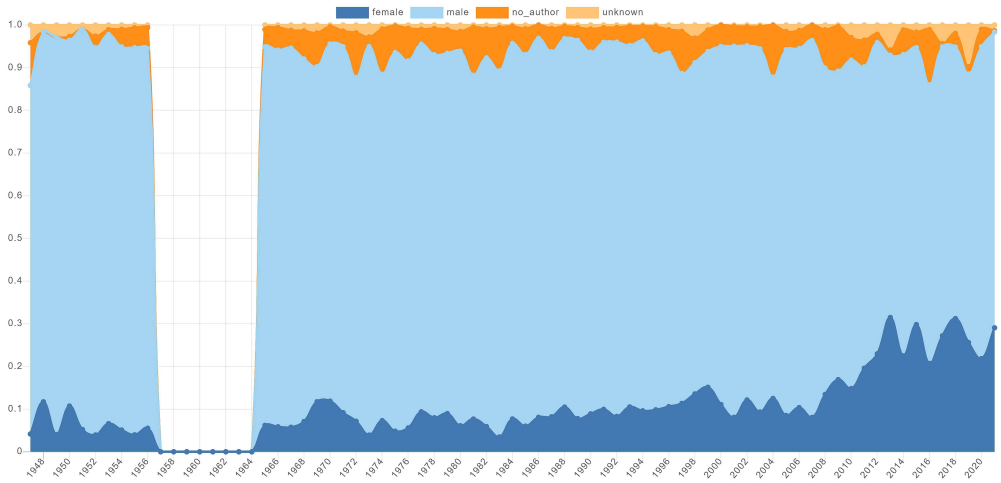
2. ábra. A metaadat-vizualizációs beállítási panel

[26] Az idősoros ábrázolás esetében ez például azt jelenti, hogy az aggregált módban az adott évhez tartozó nyers adatok száma jelenik meg a grafikonon, normalizált ábrázolás esetén pedig a relatív gyakoriságot láthatjuk százalékokban kifejezve. Ha például egy napilap cikktípusait (hír, hirdetés stb.) jelenítjük meg, akkor az aggregált ábrázolás során az adott évhez tartozó összes hirdetés száma jelenik meg a függvényen, míg a normalizált verzióban azt láthatjuk, hogy az adott évben megjelent összes cikk hány százaléka volt hirdetés.

Nézzünk néhány konkrét példát a különböző típusú metaadat-vizualizációkra:



3. ábra. Női, férfi, szerző nélküli és azonosíthatatlan nemű szerzők a *Tiszatáj* folyóiratban (aggregált) 1948 és 2021 között



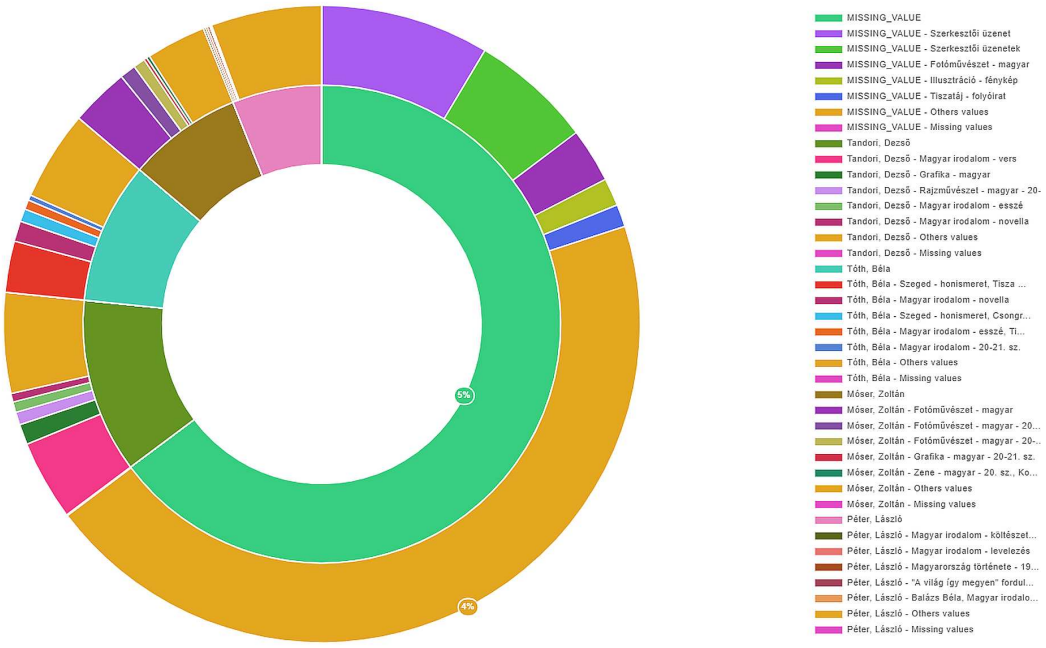
4. ábra. Női, férfi, szerző nélküli és azonosíthatatlan nemű szerzők a *Tiszatáj* folyóiratban (normalizált) 1948 és 2021 között

A *Tiszatáj*ban publikáló női és férfi szerzők eloszlása mellett a fenti ábrákon az is látszik, hogy 1957 és 1964 között nincsenek értékek a diagramokon. Ennek az az oka, hogy ebben az időszakban formátumot váltott a lap, a korábbi, oldaltól oldalig terjedő forma helyett ebben a néhány évben hasábos formában jelent meg, így a szegedi könyvtárosok nem darabolták szét, nem bontották cikkekre ezeket a számokat.

Az alábbi többszintű gyűrűdiagrammon a *Tiszatáj*ban publikáló 5 leggyakoribb szerző és a műveikhez analóg módon rendelt kulcsszavak eloszlását láthatjuk. A felsorolásban megfigyelhető, hogy az egyes kulcsszavak nem minden esetben vannak

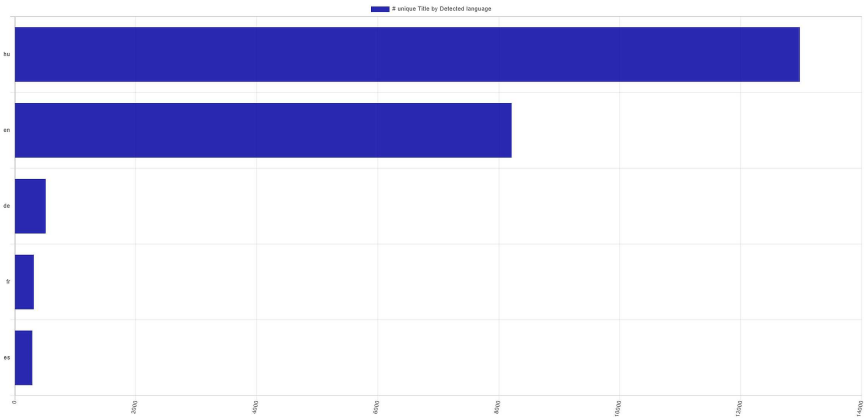
 DOI 10.31400/dh-hun.2021.4.3530

egységesítve (pl. a „Szerkesztői üzenet” és a „Szerkesztői üzenetek” külön kategóriákat alkotnak). Az ilyen pontatlanságok beazonosítását és kijavítását követően, a metaadatok tekintetében megtisztított, normalizált adatbázist újra feltölthetjük az AVOBMAT-ba.



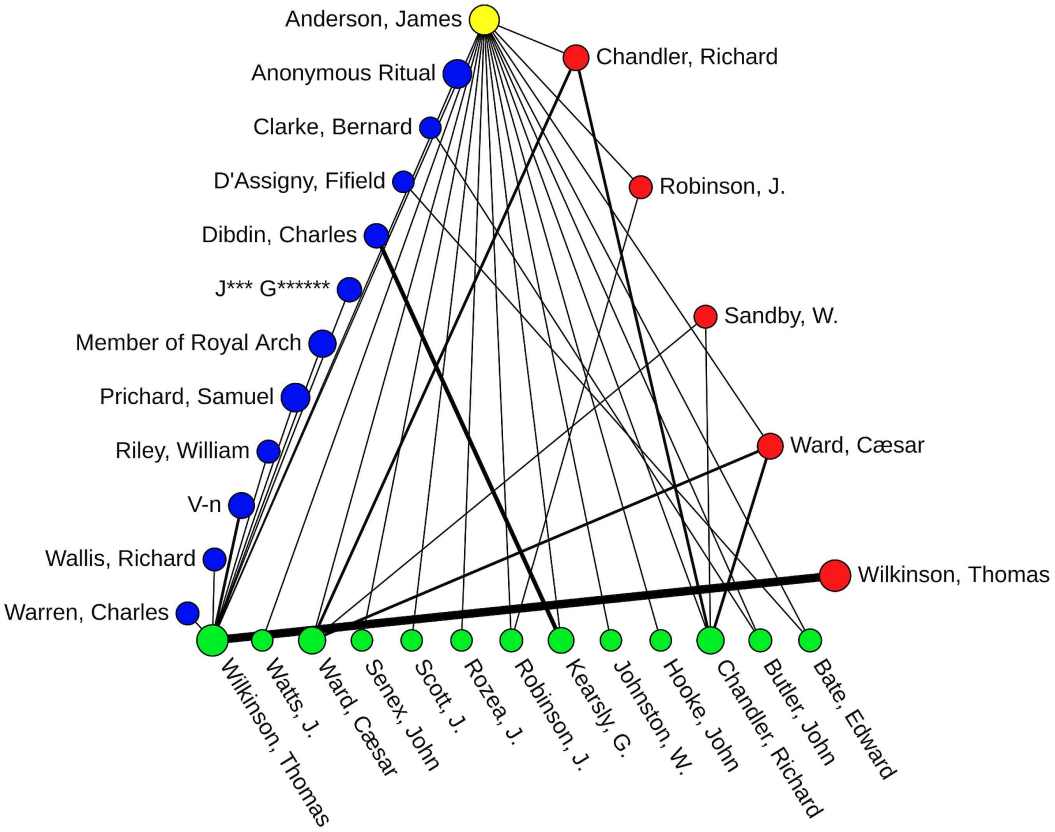
5. ábra. A *Tiszatáj*ban publikáló 5 leggyakoribb szerző és a műveikhez rendelt kulcsszavak (egyéb értékek kikapcsolva)

Tandori Dezső írásait sötétzölddel láthatjuk a belső gyűrűben. A gyűrűsáv méretéből látható, hogy ő publikálta a legtöbb írást a folyóiratban. Az adott gyűrűrészre kattintva a pontos százalék is megjelenik. A Tandori-gyűrűsáv külső gyűrűben található folytatásában a szerző által írt művek kulcsszavazott kategóriái találhatóak: pl. a sötétebb rózsaszín jelöli a „magyar irodalom – vers” kategóriát, ahogy az a jobb oldali színskála színeihez rendelt jelöléseknél is megfigyelhető.



6. ábra. Az SZTE *Egyetemi Kiadványok* repozitóriumában az 5 leggyakoribb nyelven 2000 után írt cikkek száma (automatikus nyelvfelismerés)

A metaadatokhoz tartozó tételek hálózati kapcsolatát is megjeleníthetjük, s ezekhez tartozó alhálózatokat is vizualizálhatunk. A hálózati csúcsoknál található kör mutatja a csúcshoz tartozó kapcsolatok számát, amely a kör méretével arányos. A csúcsokat összekötő vonalak (élek) vastagsága a kapcsolatok számát jelzi, melyek száma megjelenik, ha egy adott élre kattintunk.



7. **ábra**. A 18. századi brit és ír szabadkőműves könyvek szerző–nyomdász–könyvkereskedő alhálózata, James Anderson szerző kapcsolati hálója

A bibliográfiai adatok elemzése amellett, hogy feltárja a különböző metaadatmezőkben tárolt rekordok közötti, korábban ismeretlen összefüggéseket, és rávilágít az eleddig figyelmen kívül hagyott trendekre, fényt deríthet az adatbázisok (például bibliográfiai adatokat érintő) hiányosságaira, korlátaira, beleértve az adatbázis-készítésnél alkalmazott előítéleteket (például a gyűjteménybe kerülő szövegek kiválogatásával vagy osztályozásával kapcsolatban).[27] A legtöbb adatszolgáltató (könyvtárak és profitorientált cégek egyaránt) vagy nem ismerte fel, vagy ha igen, nem szívesen hozza nyilvánosságra ezeket az információkat. Ilyen típusú problémák előzetes számítógépes feltérképezése segíti a kutatókat abban, hogy megalapozott szakmai döntéseket hozzanak projektjeikről, és kritikusan elemezzék a digitális gyűjtemények tartalmát. Továbbá az adatgazdáknak is lehetőséget nyújt a bibliográfiai adatok minőségének javításához.

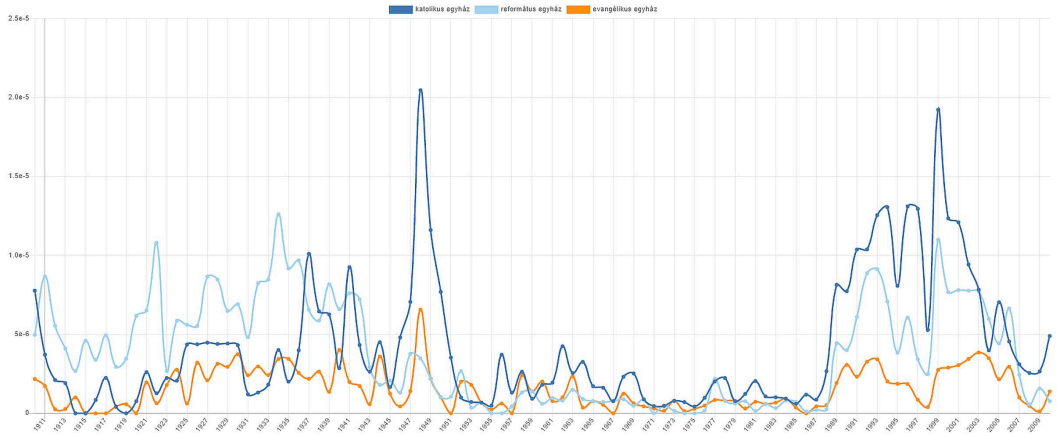
[27] Katherine Bode, „Why You Can't Model Away Bias?" *Modern Language Quarterly* 81, 1 sz. (2020): 95–124, https://doi.org/10.1215/00267929–7933102.

 DOI 10.31400/dh-hun.2021.4.3530

6. Szövegelemzés és vizualizáció

6.1. N-gram elemzés

A szövegek diakronikus elemzését az AVOBMAT n-gram elemzője támogatja. Idősoron megjeleníti a felhasználó által megadott – teljes szövegben található – n-gramok (itt egymás után következő *n* darab szó) éves eloszlását aggregált és normalizált módon. A legfeljebb öt szó hosszúságú n-gramokat az előfeldolgozási szakaszban azonosítja a program. A normalizált nézet esetében a százalékos gyakoriságot úgy kapjuk meg, hogy az adott évben fellelhető, felhasználó által keresett n-gramok számát elosztjuk az ugyanahhoz az évhez tartozó szövegekben található szavak számával.



8. ábra. A *katolikus egyház, református egyház* és *evangélikus egyház* bigramok normalizált eloszlása a *Délmagyarország* napilapban, 1911 és 2009 között[28]

6.2. Témamodellezés

A témamodellezés segítségével rejtett és absztrakt témákat, szemantikai információkat fedezhetünk fel szövegekben. Az algoritmus statisztikai módszereket használ a szövegekbe ágyazott témák feltárására, valamint e témák kapcsolatainak és időbeli változásainak feltárására.[29] Az AVOBMAT rendelkezik egy böngészőbe épített Latent Dirichlet Allocation (LDA) funkcióval, amely a *jsLDA*-könyvtárra[30] támaszkodik a témamodellek kiszámításánál és grafikus ábrázolásánál. Az LDA a felhasználó által megadott számú látens témát azonosít, ahol minden dokumentum e témák keverékének tekinthető. A módszer az együtt előforduló szavakat csoportosítja témákba, a dokumentumokhoz pedig valószínűségekkel hozzárendeli az egyes témákat. A témaelemzés mellett a modellezés eredményeit is különböző módon tudja megjeleníteni az

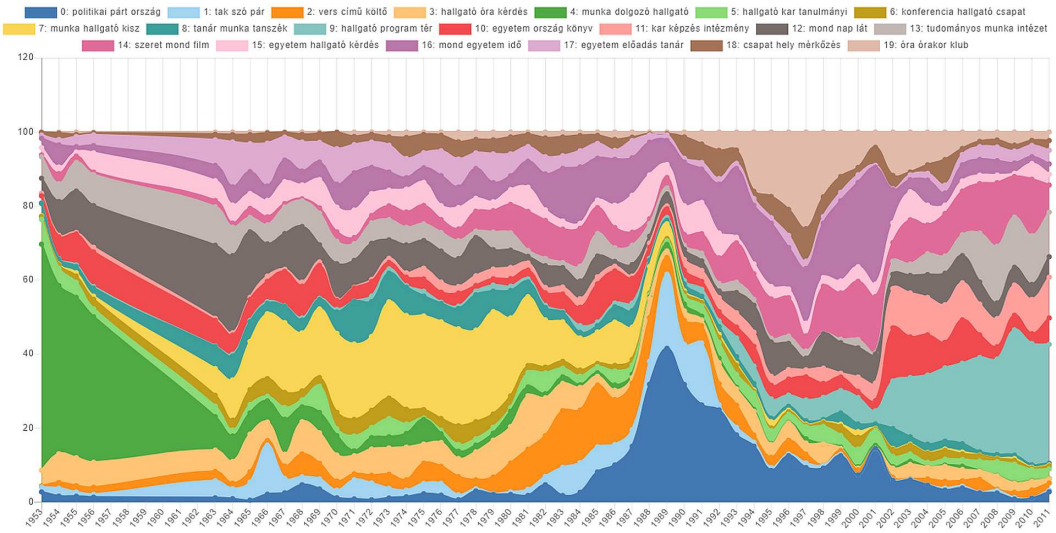
[28] 1920 és 1925 között csak részben vagy egyáltalán nem jelent meg a *Délmagyarország*, ekkor *Szeged* néven volt elérhető napilap. 1956. november 20. és 1957. április 30. között pedig a *Szegedi Néplap* váltotta fel a *Délmagyarország*ot. Az n-gram elemzés ezen újságok cikkeit is tartalmazza.

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

[30] *jsLDA*, hozzáférés: 2021.12.15, `https://mimno.infosci.cornell.edu/jsLDA/`.

AVOBMAT. Megmutatja az egyes témákhoz kapcsolódó legrelevánsabb szavakat és dokumentumokat, megjeleníti e témák eloszlását idősoron, vizualizálja a különböző témák közötti korrelációkat, és különböző formátumokban exportálja az eredményeket. A bibliográfiai adatok felhasználása lehetővé teszi, hogy diakronikus témamodellezéseket végezzünk, amelyek általánosabb szemantikai mintákat tárnak fel a nyelvhasználatban, mint amilyeneket a gyakorta nagy méretű digitális gyűjtemények szoros olvasása nyújtana.

Az eredeti *jsLDA*-implementáció paraméterként a témák számát és az iterációkat igényli. Ezt három új paraméterrel bővítettük. A felhasználó beállíthatja az elemezni kívánt korpuszban a szavak minimális előfordulási számát. Ha ez a minimum nagy, az algoritmus gyorsabb lesz a szerver és a böngésző közötti csökkentett adatközlés miatt, de hátránya, hogy elveszíthetjük a dokumentumokra vonatkozó információk egy részét. A leggyakrabban előforduló (stop)szavakat interaktív módon távolíthatjuk el a „Vocabulary” ikonra kattintva. Az ilyen szűrés után mindig újra kell futtatni az elemzőt. A felhasználók beállíthatják az alfa és béta LDA-hiperparamétereket is: az alfa a dokumentum–téma sűrűséget, a béta pedig a téma–szó sűrűséget jelöli.[31] A *jsLDA* programot még kiegészítettük egyrészt azzal, hogy a témák időbeli eloszlását aggregált és normalizált módokon is ábrázolhatjuk, másrészt az egyes témákhoz kapcsolható dokumentumok alapvető bibliográfiai adatait is megjeleníthetjük a témákra vonatkozó dokumentumokhoz tartozó valószínűségi értékek mellett.



9. ábra. A *Szegedi Egyetem* folyóirat egy témamodellezése, 1953–2011 (témák száma: 20, alfa = 0,1; béta = 0,01)

A fenti témamodellezés esetén így is értelmezhetjük az alábbi témákhoz tartozó szavakat: [0] *politikai, párt, ország, kérdés, tart, lát, helyzet* – pártpolitikai hírek; [2] *vers,*

[31] Szimmetrikus Dirichlet-eloszlást feltételezve, az alacsony alfa érték nagyobb súlyt helyez arra, hogy minden dokumentum csak néhány domináns témából álljon, míg a magas érték sokkal több viszonylag domináns témát ad vissza. Hasonlóképpen, egy alacsony béta érték nagyobb súlyt helyez arra, hogy az egyes témák csak néhány domináns szóból álljanak. Ha magasabb a béta érték, a témák nagy számú – korpuszban található – szóból állnak.

 DOI 10.31400/dh-hun.2021.4.3530

című, költő, kötet, szerző – a folyóiratban megjelent versek, költeményes kötetek bemutatása; [5] *hallgató, kar, tanulmányi, ösztöndíj, félév, szociális, támogatás, szak* – hallgatói támogatásokkal, ösztöndíjakkal kapcsolatos hírek; [7] *munka, hallgató, KISZ, kollégium, bizottság, főiskola, tag, éves, feladat, tevékenység* – KISZ-es eseményekhez, tagsághoz köthető témák; [18] *csapat, hely, mérkőzés, pont, bajnokság, második, játékos, együttes, verseny* – egyetemi sportbajnokságokkal kapcsolatos híradás.

6.3. Szóstatisztikai elemzések

A szófelhők hatékony eszközök lehetnek egy korpuszban valamilyen szempontból prominens szavak kiemelésére. Háromféle szóstatisztikai elemzőt integráltunk az AVOBMAT alkalmazásba. A legegyszerűbb vizualizáció a szógyakoriság alapján készíti el a szófelhőt és mutatja az egyes szavakhoz tartozó gyakorisági adatokat. Minden egyes vizualizáció esetében megadhatjuk, hogy hány darab szó jelenjen meg a szófelhőben. A második elemző *(Significant text)* azt mutatja, hogy milyen, az átlagostól jelentősen eltérő gyakoriságú szavak különböztetik meg egy digitális gyűjtemény általunk szűréssel kiválasztott részhalmazát a korpuszban található összes szövegtől.[32] A harmadik elemző *(TagSpheres)* lehetővé teszi a felhasználók számára, hogy egy szó kontextusát vizsgálják.[33] A különböző szófelhők mellett a szóstatisztikai adatokat oszlopdiagrammokban is láthatjuk, és az itt szereplő adatsorokat exportálhatjuk.

A *Significant text* elemző egy lekérdezés által definiált alkorpuszra leginkább jellemző (jelentősen eltérő gyakoriságú) szavakat azonosítja. Például ha a felhasználó az AVOBMAT keresési lehetőségeit használva kiválaszt egy szerzőt a korpuszból, akkor ez az eszköz megmutatja azokat a szavakat, amelyek e szerző műveihez legszignifikánsabban kapcsolódnak (jelentősen eltérő a gyakoriságuk) a teljes korpuszban található szövegekhez képest. Az előbbi részhalmazt előtérhalmaznak *(foreground set),* a dokumentumok teljes halmazát pedig háttérhalmaznak *(background set)* nevezzük.[34] Az *Elasticsearch* ezen halmazok statisztikai összehasonlításával rangsorolja az egyes szavakat. A következő képlet mutatja az úgynevezett JLH-érték kiszámítását, amelyet a szavak rangsorolásához alkalmazunk:

$$JLH = \left(p_{előtérhalmaz} - p_{háttérhalmaz}\right) \frac{p_{előtérhalmaz}}{p_{háttérhalmaz}}$$

ahol a $p_{előtérhalmaz}$ a relatív gyakorisága az előtérhalmazban található kifejezésnek, míg a $p_{háttérhalmaz}$ a relatív gyakorisága ugyanennek a kifejezésnek a háttérhalmazban. Az

[32] A *significant text* elemző dokumentációját lásd, hozzáférés: 2021.12.15, `https://www.elastic.co/guide/en/elasticsearch/reference/8.0/search-aggregations-bucket-significanttext-aggregation.html`.

[33] Stefan Jänicke and Gerik Scheuermann, „On the Visualization of Hierarchical Relations and Tree Structures with TagSpheres," in José Braz et al., eds., *Computer Vision, Imaging and Computer Graphics Theory and Applications,* 199–219 (Cham Springer International Publishing, 2017), `https://doi.org/10.1007/978-3-319-64870-5_10`.

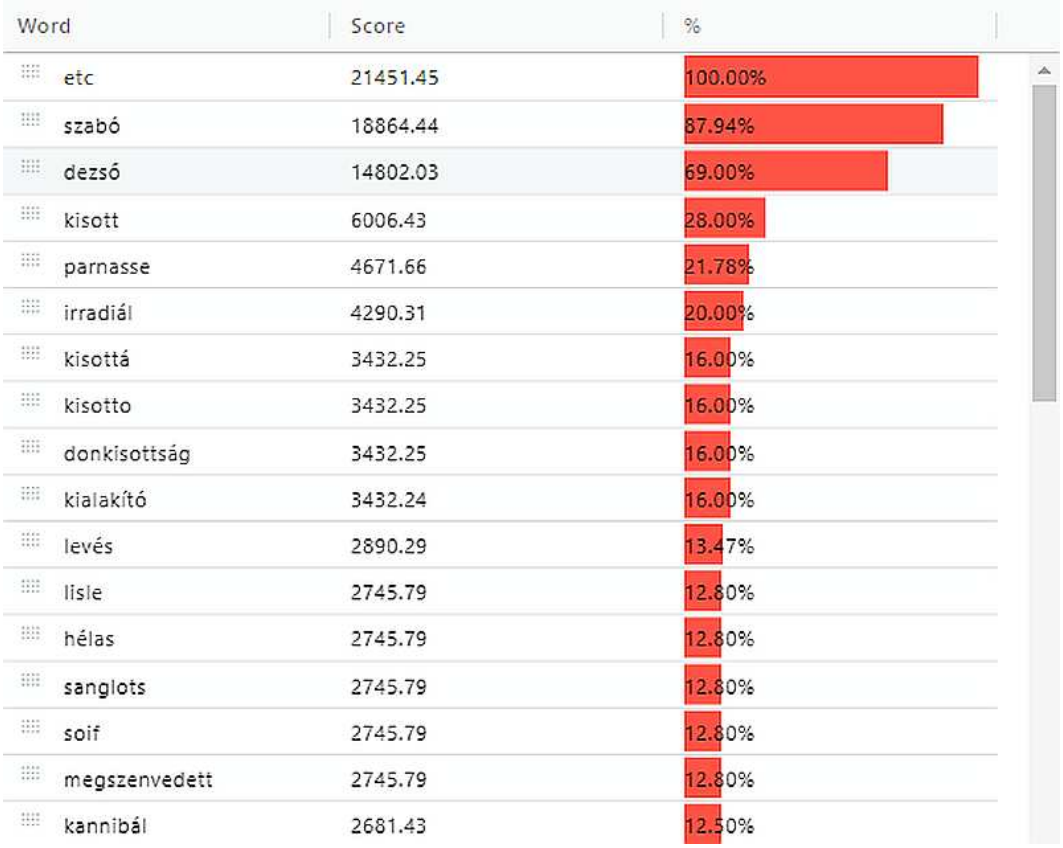
[34] Ezt az *Elasticsearch*ben használt alapbeállítást az eredmények értelmezésénél figyelembe kell venni. A háttérhalmazt úgy is megadhatjuk az *Elasticsearch* konfigurációjában (a *background_is_superset* paramétert hamisra állítva), hogy ez diszjunkt halmazt képezzen az előtérhalmazzal, így csak azokat a szövegeket tartalmazza, amelyeket nem választott ki a felhasználó. Ezt a választási opciót szeretnénk a grafikus felületre is kivezetni a jövőben.

AVOBMAT ezt az összehasonlító elemzést még három közismertebb metrika *(mutual information, chi square, Google normalized distance)* segítségével is kiszámolja. Az elemzés elvégzése előtt beállíthatunk két paramétert: (i) hány darab szó jelenjen meg a szófelhőben; (ii) hány darab dokumentumot válasszon ki mintaként *(sample)* az előtérhalmazból és a háttérhalmazból.[35]

10. ábra. Szabó Dezső *Nyugat* folyóiratban megjelent 230 írására legjellemzőbb szavak (JLH-metrika, 100 szó). A szófelhő bizonyos szavai jól tükrözik a szerző francia műveltségét.

[35] Christopher D. Manning, Prabhakar Raghavan and Hinrich Schütze, *An Introduction to Information Retrieval* (Cambridge: Cambridge University Press, 2009), 272–275 *(mutual information),* 275–277 *(chi square),* https://nlp.stanford.edu/IR-book/pdf/irbookonlinereading.pdf; Rudi L. Cilibrasi and Paul M. B. Vitányi, „The Google Similarity Distance," *IEEE Transactions on Knowledge and Data Engineering* 19, 3. sz. (2007): 370–383, https://arxiv.org/pdf/cs/0412098v3.pdf, https://doi.org/10.1109/TKDE.2007.48.

 DOI 10.31400/dh-hun.2021.4.3530



Word	Score	%
etc	21451.45	100.00%
szabó	18864.44	87.94%
dezső	14802.03	69.00%
kisott	6006.43	28.00%
parnasse	4671.66	21.78%
irradiál	4290.31	20.00%
kisottá	3432.25	16.00%
kisotto	3432.25	16.00%
donkisottság	3432.25	16.00%
kialakító	3432.24	16.00%
levés	2890.29	13.47%
lisle	2745.79	12.80%
hélas	2745.79	12.80%
sanglots	2745.79	12.80%
soif	2745.79	12.80%
megszenvedett	2745.79	12.80%
kannibál	2681.43	12.50%

11. ábra. Szabó Dezső *Nyugat* folyóiratban megjelent 230 írására legjellemzőbb szavak (JLH-metrika) és az ezekhez tartozó statisztikai adatok

Melyik metrikát válasszuk? A *mutual information* a magas gyakoriságú kifejezéseket részesíti előnyben, még akkor is, ha azok a háttérhalmazban is gyakran előfordulnak. Így ez a stopszavak kiválasztásához is vezethet. A *mutual information* nem valószínű, hogy nagyon ritka kifejezéseket, például helytelen helyesírással írott szavakat emel ki. A *Google normalized distance (gnd)* a magas együttes előfordulási gyakoriságú kifejezéseket részesíti előnyben, és elkerüli a stopszavak kiválasztását; talán jobban alkalmas a szinonimák felismerésére. A *gnd* azonban hajlamos a nagyon ritka kifejezések kiválasztására, amelyek például helyesírási hibákból származnak. A *chi square* és a JLH hozzávetőlegesen a kettő között helyezkedik el.[36]

A hagyományos szófelhők a szavakat egymástól függetlenül kezelik, és elveszítik a szavak közötti kontextuális információt. A szavak szövegkörnyezetének grafikus ábrázolásához a *TagSpheres* programot integráltuk. Ez olyan szófelhőt hoz létre, amely egy megadott keresőszó környezetében együttesen előforduló szavakat mutatja. A különböző szótávolságra található szavakat eltérő színekkel jelöli. A keresőkifejezés mellett a felhasználó megadhatja (i) az együtt előforduló szavak minimális gyakoriságát; (ii) az együtt előforduló szavak maximális szótávolságát a megadott szótól; (iii) a szavak a

[36] „Significant text aggregation," hozzáférés: 2021.12.15, `https://www.elastic.co/guide/en/elasticsearch/reference/8.0/search-aggregations-bucket-significanttext-aggregation.html`.

keresőkifejezéstől csak balra, csak jobbra vagy mindkét környezetben való előfordulását. Ennél az elemzőnél különös jelentősége van annak, hogy az előfeldolgozás során kiszűrtük-e a stopszavakat.

12. ábra. Babits Mihály „Istenképe” a *Nyugat* folyóiratban megjelent művei alapján (3 szótávolság stopszavak nélkül, minimum szógyakoriság: 2). Ilyen típusú elemzést más szerzők esetén is elvégezhetünk és összehasonlíthatjuk az eredményeket.

 DOI 10.31400/dh-hun.2021.4.3530

Word	Word distance	Count	%
isten	0	577	100.00%
tud	1	21	3.64%
ad	1	20	3.47%
ember	1	19	3.29%
ó	1	14	2.43%
isten	1	12	2.08%
szó	1	8	1.39%
hisz	1	8	1.39%
régi	1	7	1.21%
kér	1	7	1.21%
mond	1	7	1.21%
ég	1	7	1.21%
harcol	1	7	1.21%
lát	1	7	1.21%
süket	1	7	1.21%
anya	1	6	1.04%
világ	1	6	1.04%
szeret	1	6	1.04%

13. ábra. Az *Isten* szó környezte Babits Mihály *Nyugat* folyóiratban megjelent műveiben

6.4. Konkordancia

A konkordancia eszköz segíti az elemezni kívánt szövegek szoros vagy lassú olvasását. Megadhatjuk, hogy az adott keresési kifejezés (akár több szó) környezetében hány betűt jelenítsen meg a program, valamint azt is, hogy maximum hány találatot mutasson. A kulcsszavak kontextusát kétféle nézetben jeleníthetjük meg: az egyikben („View occurances line by line") soronként jelennek meg a találatok (így a szövegkontextus kisebb), a másikban („View occurances in context") pedig a szövegdobozban annyi karakter jelenik meg a keresési kifejezés körül, ahányat a felhasználó beállít. Mindkét esetben a találatokat rendezhetjük szerző, megjelenési év és szöveg szerint.

Authors	Title	Publication year	Text
Schöpflin Aladár	A két Vörösmarty	1908	tudatja vele rosszallását s röviden, pregnánsan, rá nézve jellemzően fejti ki a függetlenség értékét az életben, szemben a szolga-szerep adta viszonylagos jóléttel. Egész világfelfogását bizonyos józan egyensúly jellemzi. Vallja a francia forradalom alapelveit, de forradalmi színezet nélkül. A ***magyar nemzet*** függetlenségéért lelkesül, de még a 48 mámora sem ragadja magával, óvatos higgadtsággal áll elébe. Sarkallja a nemzetet haladásra, de ostor nélkül. A jövőbe néz, de ugyanekkor a múltba is bele tud merülni. Szellemi érdeklődésében megfér Horatius és Hugo Viktor; az egyikhez temperamentuma húzza
Junius	Gyulai Pál	1909	amelyhez tartozott egyébként minden gondolkozó agy, amely mérsékletet hirdetett, amely a tisztes kiegyezést többre tartotta volna a kétségbeesés heroikus erőfeszítéseinél, az élet-halál küzdelem kockázatánál. Világos után, teóriában, jogilag legalább nem volt többé Magyarország és nem volt többé ***magyar nemzet***. Földünkön mindenfelé csak rom és pusztulás; testünk-lelkünk merében vérző seb. Nem csoda, ha nemzetünk a becsülettel megállott heroikus erőfeszítés után zsibbadt, tompa aléltságba esett; ha biztatva magát jövendővel, megadta magát sorsának. A magyarnak egy időre, akkor azt hihettük, hogy örökre
Schöpflin Aladár	A fiatal Gyulai	1909	kell?" S nem ezerszer gúny-e, hogy ez a vitatkozás még ma, ötven év múlva sem vesztette el teljesen aktualitását? Legélesebbé akkor válik a hangja, mikor az ötvenes évek lírikusainak kulturálatlanságát tépdesi. Már akkor ki van benne alakulva az az ideál, amelyet a magyar irodalom elé tűzött: a ***magyar nemzet*** egyéniségét európai színvonalon művészileg kialakító irodalom. A mai napig is ez maradt legfőbb kritériuma minden magyar irodalmi műnek az ő szemében. Általában egész kritikai pályájának mozgató elvei már ekkor megállapodtak benne. Felfogása a későbbi évek nyugodt alkotó munkájában bővült és

14. ábra. Konkordancianézet. A *magyar nemzet* kifejezés a *Nyugat* folyóirat cikkeiben

6.5. Névelem-felismerés

Az AVOBMAT-ba integráltuk a spaCy neurális hálókra épülő nyelvmodelljeit, melyek segítségével szövegekből automatikus módszerekkel kinyerhetünk névelemeket (Named Entity Recogniton: NER), többek között közneveket, tulajdonneveket (pl. személynevek, helyek, szervezetek nevei) és dátumokat. Ez a funkció 16 nyelven működik, a magyar nyelvet is beleszámítva, bár az utóbbinak jelenleg még nincs hivatalos spaCy nyelvmodellje.[37] Az alábbi táblázat mutatja, milyen nyelveken milyen névelemeket azonosít az AVOBMAT. A névelem-felismerés eredményeit többféle módon lehet megjeleníteni. A szemantikus gazdagítás során létrejött névelemek egyes típusai külön metaadatmezőkben tárolódnak és jelennek meg a fazettás és összetett keresőben, valamint a metaadat-vizualizációs beállítási panelben. A szövegben felismert névelemek a teljes szövegben is megtekinthetők: ehhez a találati listában ki kell választanunk egy szöveget és a megjelenési módot a „Named Entity Recognition"-re kell állítani. Ekkor a névelemeket és ezek típusait eltérő színekkel látjuk majd a szövegben. Az AVOBMAT a névelem-felismerés eredményeiről egyszerű statisztikákat is készít. Az „Entities in all documents" funkció az adatbázisunkban vagy annak általunk szűkített részhalmazában mutatja a felismert névelemeket, számukat és azt, hogy hány dokumentumban fordulnak elő. Az „Entities by documents" pedig a névelemek számát mutatja dokumentumonként. A nyelvi modellek frissítése lehetséges. A névelem-felismerés pontossága nyelvenként, ezeken belül elérhető (általában kis, közepes és nagy) modellenként és szövegtípusonként változik.[38]

	Person	Organization	Location	Miscellaneous	Language	Work of art	Geopolitical	National or religious group	Date	Ordinal	Product	Quantity	Time	Money	Infrastructure	Cardinal	Event	Law	Percent	Period	Movement	Phone	Pet name	Title affix
Chinese	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X					
Danish	X	X	X	X																				
Dutch	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X					
English	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X					
French	X	X	X	X																				
German	X	X	X	X																				
Greek	X	X	X				X				X						X							
Italian	X	X	X		X	X	X	X	X	X	X	X	X	X	X	X	X	X	X		X	X	X	X
Japanese	X	X	X	X																				
Lithuanian	X	X	X				X				X		X											
Hungarian	X	X	X	X																				
Norwegian Bokmål	X	X	X				X		X				X											
Polish	X	X	X	X																				
Portuguese	X	X	X		X	X	X	X	X	X	X			X	X	X	X			X				
Romanian	X	X	X	X																				

15. ábra. Névelem-felismerés különböző nyelveken az AVOBMAT-ban

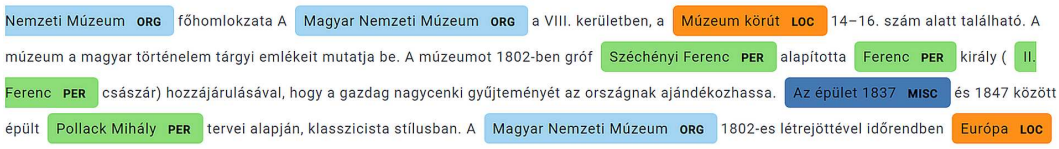
[37] György Orosz, Zsolt Szántó, Péter Berkecz, Gergő Szabó and Richárd Farkas, „HuSpaCy: An Industrial-Strength Hungarian Natural Language Processing Toolkit," in Berend Gábor, Gosztolya Gábor és Vincze Veronika, szerk., *XVIII. Magyar Számítógépes Nyelvészeti Konferencia,* 59–73 (Szeged: JATEPress, 2022), `https://rgai.inf.u-szeged.hu/file/427`.

[38] „SpaCy Models and Languages," hozzáférés: 2021.12.15, `https://spacy.io/usage/models`.

 DOI 10.31400/dh-hun.2021.4.3530

Entity	Count ↓	Type	Documents
Szeged	346	LOCATION	63
Juhász Gyula	311	PERSON	52
József Attila	216	PERSON	34
Szegeden	175	LOCATION	52
Illyés	159	PERSON	9
Bálint Sándor	142	PERSON	14

16. ábra. Névelem-statisztika Péter László *Tiszatáj* folyóiratban publikált cikkeire vonatkozóan



17. ábra. A teljes szövegben megtekinthető névelemek a *Wikipédia* „Budapest” szócikkében

7. Jogosultságkezelés és adminisztrátori felület

Az AVOBMAT jelenleg egy egyszerű jogosultságkezelési funkcióval rendelkezik. A regisztráció után az ideiglenes vagy állandó felhasználók feltölthetnek egyéni adatbázisokat. A kutatási projektek keretében létrejött adatbázisokat igény szerint akár publikussá is lehet tenni egy külön adminisztrátori felületen.[39] Az ily módon bárki számára hozzáférhető metaadatok és szövegek innovatív módon, korszerű szövegbányászati eszközökkel elemezhetők egy egyszerű grafikus felületen. Mivel jelenleg a feltölthető adatbázisok méretét nem lehet az adminisztrátori felületen szabályozni és a rendelkezésre álló szerver kapacitása korlátozott, ezért az AVOBMAT még nem tud tömeges felhasználói igényeket kielégíteni. Egyéni egyeztetést követően kutatóknak

[39] Jelenleg egy COVID-19-cel kapcsolatos tudományos cikkeket (63571 darab) tartalmazó adatbázis (cord-2020-05-14) érhető el nyilvánosan. Ezt és az adatbázis korábbi verzióit 44 országban használták a pandémia kezdetén. Fontos megjegyezni, hogy ezen a felületen nem tesztelhető az összes cikkben említett elemző (pl. a névelem-felismerés). Lásd hozzáférés: 2021.12.15, `https://avobmat.hu/covid-19/` és hozzáférés: 2021.12.15, `http://dighum.bibl.u-szeged.hu/avobmat-covid/home`. Lucy Wang, Kyle Lo and Róbert Péter, „COVID-19 Open Research Dataset and AVOBMAT Text Mining Tool,” New York NLP Meet-up, 2020. ápr. 27., hozzáférés: 2021.12.15, `https://www.youtube.com/watch?v=GivUfb8KhZY`; Christopher Nunn, „Research COVID-19 with AVOBMAT,” *OpenMethods: Highlighting Digital Humanities Methods and Tools,* 2020. jún. 8, `https://openmethods.dariah.eu/2020/06/08/research-covid-19-with-avobmat/`.

és kutatócsoportoknak viszont tudunk hozzáférést biztosítani az eszközhöz, az erőforrások függvényében. Az adminisztrátori felületen beállíthatjuk még a metaadatmezők neveit, rövidítéseit és megjelenéseit a különböző elemzőkben és a keresési felületen, valamint új metaadatmezők felvételére is van lehetőség.

8. Összegzés és tervek

E dolgozat a platformfüggetlen AVOBMAT szöveg- és adatbányászati kutatási eszközt mutatta be, amely elsősorban olyan felhasználók számára lett kifejlesztve, akik nem rendelkeznek programozási ismeretekkel. Láttuk, hogy ez a felfedezést segítő alkalmazás számos dinamikus szöveg- és adatbányászati elemzést tesz lehetővé. Az egyszerű, felhasználóbarát grafikus felület interaktív paraméterbeállítást és vezérlést biztosít a normalizálást is támogató előfeldolgozástól az analitikai szakaszokig. A szoros és távoli olvasás kombinálása mellett, az AVOBMAT-eszköztár egyedülálló tulajdonsága, hogy egyesíti a bibliográfiai (meta)adat- és a szövegelemzést számos nyelven. Lehetővé teszi a felhasználók számára, hogy a feltöltött és nyilvánosan elérhető adatbázisokat metaadatok és teljes szövegű kulcsszavas keresések segítségével szűrjék, és a szűrt adatkészleteken elvégezzék az összes metaadat-vizualizációs, hálózati és nyelvtechnológiai elemzést. A felhasználók így könnyen és interaktív módon kísérletezhetnek az elemzések különböző beállításaival a munkafolyamat során. Ezáltal a program segít felismerni a számítógépes szöveg- és adatelemzés episztemológiai kihívásait, korlátait és erősségeit, valamint kritikus módon értelmezni az alkalmazott módszereket és eredményeket. Bár adatvezérelt technológiákat használunk az elemzéseknél, fontos nyomatékosítani, hogy a sokszor hiányos, zajos, esetleges adatokat nem tekinthetjük adottnak vagy objektívnek, az adatok elméletekkel terheltek, a kapcsolatuk a kontextustól függően folyamatosan változik.[40] Ezért olyan funkciókat építettünk be az alkalmazásba, melyek segítik az adatbázisok korlátainak és az adatbázis-építés mögött húzódó előítéletek feltérképezését. A mennyiség nem garantálja a minőséget, az adatok természetéből adódó pontatlanságát számos esetben, bizonyos mértékig kompenzálja az adatok volumene és nagy száma *(big data)*. Az eredmények értelmezésénél nem feltétlenül a pontos számokon van a hangsúly, hanem azokon az esetenként új típusú bizonyítékként szolgáló trendeken, mintákon és modelleken, amelyeket ezek feltárnak. Az AVOBMAT csupán egy eszköz, módszertani apparátus, a digitális forrás- és kódkritika *(digital source criticism, critical code studies)* alkalmazása nélkülözhetetlen az eredmények interpretációja során. Az AVOBMAT export és import funkciói (konfigurációs beállítások, adatsorok, vizualizációk) az eredmények reprodukálhatóságát, valamint az előfeldolgozás és a szövegelemzés átláthatóságát hivatottak megkönnyíteni.

[40] Tobias Blanke and Andrew Prescott, „Dealing With Big Data," in Gabriele Griffin and Matt Hayler, eds., *Research Methods for Reading Digital Data in the Digital Humanities,* 184–205 (Edinburgh: Edinburgh University Press: Edinburgh, 2016), https://doi.org/10.1515/9781474409629-012; Giovanni Schiuma and Daniela Carlucci, eds., *Big Data in the Arts and Humanities* (Boca Raton, FL: Auerbach Publications, 2018); Jennifer Edmond, Nicola Horsley, Jörg Lehmann and Mike Priddy, eds., *The Trouble With Big Data: How Datafication Displaces Cultural Practices* (London: Bloomsbury Academic, 2021), https://doi.org/10.5040/9781350239654.

 DOI 10.31400/dh-hun.2021.4.3530

A további fejlesztési terveket illetően, az ismert apró kódhibák javítását követően a feltöltési lehetőségeket szeretnénk kiegészíteni többek között a TEI XML formátummal, valamint olyan API-k integrálásával, melyek metaadatokat és teljes szöveget tudnak kinyerni különböző, nemzetközi standardokat használó tudományos adatforrásokból. A felhős környezetben működő és skálázható szerverkapacitás függvényében a felhasználóknak lehetőséget biztosítanánk saját, egyénileg tanított spaCys nyelvi modellek feltöltésére. Az AVOBMAT-ba feltöltött adatbázisok szemantikus gazdagítását a szófaji egyértelműsítéssel (POS tagging) szeretnénk bővíteni, melynek a grafikus felületen is kereshető eredményeiről szintén készítenénk statisztikai kimutatásokat. Az egyértelműsített metaadatok, a névelem-felismeréssel és névelem-összekapcsolással (Named Entity Linking: NEL), valamint magyar nyelvű szövegekre finomhangolt *wikifier*rel azonosított és a *Wikidata* szemantikusan összekapcsolt névelemeihez rendelt adatok felhasználásával *Neo4j*-re épülő tudásgráfokat szeretnénk létrehozni.[41] A komoly nyelvtechnológiai kihívásokat okozó névelem-azonosítás és -összekapcsolás hatékonyságának növelését segítik az AVOBMAT többnyelvű automatikus morfológiai elemzői (pl. lemmatizálás). A magyar *DBpedia*[42] és *Nemzeti Névtér*[43] elemeivel is gazdagított tudásgráf-adatbázisok és szemantikus tudáshálók segítségével további, eleddig ismeretlen kapcsolatokat fedezhetnénk fel, többek között szövegek, metaadatok, személyek, helyszínek, események és fogalmak között.[44]

Introducing the AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) Multilingual Research Tool

The objective of this paper is to demonstrate the workflow, different analytical functions and features of the multilingual AVOBMAT (Analysis and Visualization of Bibliographic Metadata and Texts) digital tool. This web application enables researchers to critically analyse bibliographic data and texts at scale with the help of data-driven methods and tools supported by artificial intelligence and natural language processing techniques. The unique features of the AVOBMAT toolkit are that (i) it can preprocess, analyse and (semantically) enrich a huge number of texts and metadata in several languages; (ii) the implemented analytical and visualization tools provide interactive close and distant reading of texts and bibliographic data; (iii) it combines bibliographic data and natural language processing research methods in one integrated, interactive and user-friendly web application. In the preprocessing phase, the user can set nine optional parameters such as lemmatization and stopword filtering. Users can create different configurations for the different analyses and

[41] *Neo4j*, hozzáférés: 2021.12.15, `https://neo4j.com/`.

[42] *DBpedia*, hozzáférés: 2021.12.15, `https://hu.dbpedia.org/`.

[43] *Nemzeti Névtér*, hozzáférés: 2021.12.15, `https://abcd.hu/`.

[44] Lásd pl. Biography Shampoo: A Network of Finnish Biographies on the Semantic Web, hozzáférés: 2021.12.15, `http://biografiasampo.fi/`; David Lindemann and Christiane Klaes, „Zotero to Wikidata Through Wikibase: A Workflow for Publication Metadata LOD-ification Using Free Software," megjelenés alatt.

visualizations. The metadata enrichment includes the automatic identification of the gender of the authors and automatic language detection. Users can search and filter the uploaded and enriched bibliographic data and preprocessed texts in faceted, advanced and command line modes. Having filtered the uploaded databases and selected the metadata field(s), users can (i) analyze and visualize the bibliographic data chronologically in line and area charts in normalized and aggregated formats; (ii) create an interactive network analysis; (iii) make pie, horizontal and vertical bar charts of the bibliographic data. As for the content analysis, the diachronic analysis of texts is supported by the N-gram viewer. Two types of frequency analyses are implemented: the significant text function shows what differentiates a subset of documents from other texts in the corpus, and the TagSpheres enables users to investigate the context of a word. The close reading is also fostered by the Keyword in Context tool. AVOBMAT has an in-browser Latent Dirichlet Allocation function to calculate and visualize topic models. It semantically enriches the texts and metadata by the use of named entity recognition in 16 languages. The export functions of AVOBMAT facilitate the reproducibility of the results and transparency of the preprocessing and text analysis. It helps users realize the epistemological challenges, limitations and strengths of computational text analysis and visual representation of digital texts and datasets.

Keywords:
text and data mining, multilingual digital tool, natural language processing, metadata, semantic enrichment

 DOI 10.31400/dh-hun.2021.4.3530