Makkai T. Csilla – Varga Emese – Vétek Bence

EGY ÉLETMŰ METSZETE*

Stilometriai vizsgálat Móricz Zsigmond regényein

„Ugye furcsa, hogy az ember számmá lett."
(Móricz Zsigmond: Tükör, X.)[1]

A 2000-es évekbeli kibontakozásától egyre nagyobb tudományos érdeklődés övezi a Moretti-féle *distant reading* elméletét és metodológiáját. Kétségtelen, hogy az igazán nagy méretű korpuszok vizsgálata meghaladja az emberi kapacitást, és a statisztikai alapú műveletek elvégzése gépi támogatás nélkül nemcsak rettentő időigényes vállalkozás, hanem a hibázás lehetősége is magas fokú.

Az olyan irodalomértelmezési kísérletek, amelyek kvantitatív módon, adatközpontúan közelítenek a szövegkorpuszokhoz, elsősorban nem a bölcsészetben zajló digitális fordulathoz köthetők,[2] hiszen a 19. század végén is végeztek hasonló kutatásokat,[3] viszont a rendelkezésre álló digitális technológiákban rejlő lehetőségek előirányozták a humán tudományok odafordulását a számokhoz. A statisztikai alapú számítógépes szövegelemzés alternatív vizsgálati módszerként feltárhat olyan megkülönböztető jegyeket, amelyek meghatározása hagyományos módszerekkel nem, vagy csak jelentős időbefektetéssel volna lehetséges, így a filológiai vizsgálatokban elsősorban egyfajta segédeszköznek tekinthető, amely új szempontokkal egészíti ki, de nem helyettesíti a hagyományos elemzési eszközöket.[4]

Az elektronikus kiadások mellett egyre több digitális szövegkiadás születik Magyarországon, azok gépi elemzésére is találunk példát,

* A kutatás a Digitális Bölcsészeti Központ támogatásával valósult meg. Köszönjük dr. Cséve Anna, dr. Kalcsó Gyula és Mihály Eszter kollégáink szakmai segítségét.

1 Móricz Zsigmond, *Tükör*, X. = Móricz Zsigmond, *Tükör (1914–1917)*, szerk. Cséve Anna, Petőfi Irodalmi Múzeum, Budapest, 2021. Lásd: dhupla.hu, https://dhupla.hu/text/o:mzs-tukor.tei.10.

2 Ted Underwood, *A Genealogy of Distant Reading*, Digital Humanities Quarterly 11. (2017/2.), lásd: digitalhumanities.org/dhq/vol/11/2/000317/000317.html.

3 Lucius Adelno Sherman, *Analytics of literature. A manual for the objective study of English prose and poetry*, Ginn and Company, Boston, 1893.; Thomas Corwin Mendenhall, *The Characteristic Curves of Composition*, Science 1887/9., 237–246. Lásd: https://www.science.org/doi/10.1126/science.ns-9.214S.237.

4 David Holmes – Judit Kardos, *Who Was the Author? An Introduction to Stylometry*, Chance 16. (2003/2.), 5., lásd: https://id.art1lib.org/ireader/42465305.

viszont olyan vizsgálat, amely egyetlen szerző életművének feldolgozásával foglalkozik,[5] csak elvétve olvasható.[6] Ebből kiindulva, tanulmányunkban arra teszünk kísérletet, hogy Móricz Zsigmond regényeit statisztikai alapon elemezzük, az R szoftver stilometriai vizsgálatokra alkalmas *stylo* csomagjával.[7] A szoftver különböző távolságmetrikák alapján helyezi el egymáshoz képest a szövegeket a szavak előfordulásának gyakoriságát figyelembe véve. A szerzőattribúciós kísérletek bevett módszere a stilometriai elemzés, de mennyire tud a szókészleti, illetve egyazon életművön belüli tematikai hasonlóságon túlmutatni ez a módszer? Vizsgálatunkban arra voltunk kíváncsiak, hogy a különböző mérések összefüggésbe hozhatók-e a Móricz-korpusszal kapcsolatos recepcióesztétikai vizsgálatok állításaival, illetve találunk-e érvényes metszéspontokat a statisztikai megközelítés és a szövegek szoros olvasata között. Kérdésünk továbbá, hogy a stilometriában rejlő lehetőségek alapján kimutathatók-e a „szerzői ujjlenyomatként" emlegetett rejtőzködő jellegzetességek, melyek meghatározhatják az életmű változását, az esetleges írói korszakokat.

Statisztikai módszerek irodalmi szövegeken

A számítógépes nyelvészet egyik területe a stilometria, amely képes nagy méretű szövegkorpuszok stílusának statisztikai alapú elemzésére. A stilometriai vizsgálatokban a szövegek összehasonlítása a számszerűsített nyelvi jellemzők és különféle távolságmértékek alapján történik. A tudományág megalapozása Wincenty Lutosławsky lengyel filozófus nevéhez fűződik, aki az új vizsgálati módszert 1897-ben Platón műveinek kronologizálásához alkalmazta.[8] A stilometria területéhez tartoznak a szerzőségi vizsgálatok, a vitatott vagy ismeretlen szerzőség azonosítása, az írói nyelvezet formálódásának, korszakolásának meg-

5 Emanuel Modoc – Daiana Gârdan, *Style at the Scale of the Canon. A Stylometric Analysis of 100 Romanian Novels Published between 1920 and 1940,* Metacritic Journal for Comparative Studies and Theory 2020/2., 48–63. Lásd: https://doi.org/10.24193/mjcst.

6 Kiss Margit, *Stilometriai Elemzés Lehetőségei Magyar Történeti Szövegkorpuszon*, Digitális Bölcsészet 2019/2., 15–33., lásd: http://hdl.handle.net/10831/50416.

7 Maciej Eder, Jan Rybicki és Mike Kestemont által fejlesztett programcsomag. Bővebben lásd: Maciej Eder – Jan Rybicki – Mike Kestemont, *Stylometry with R: a package for computational text analysis,* The R Journal 2016/1., 107–121., https://journal.r-project.org/archive/2016/RJ-2016-007/index.html.

8 Wincenty Lutosławski, *The Origin and Growth of Plato's Logic: With an Account of Plato's Style and of the Chronology of his Writings,* Longmans, Green and co, London – New York, 1897. Lásd: https://archive.org/details/originandgrowth00lutogoog/page/n8/mode/2up.

határozása, illetve különböző szerzők stílusának nyelvi szempontú összehasonlítása. Statisztikai eszközökkel mérhető például egy író szókészletének gazdagsága, szókapcsolati és szókörnyezeti jellemzői, sajátos szintaktikai jegyei.

A szerzőségi kérdések statisztikai alapú vizsgálatával már 1851-ben foglalkoztak, Augustus de Morgan angol tudós és Thomas Mendenhall amerikai fizikus kutatásai a stilometriai vizsgálatok első kísérletei közé tartoznak, melyben a szavak hosszúságát és gyakoriságukat vizsgálták. Az 1960-as években Frederick Mosteller és David Wallace analízise újabb előrelépést jelentett a modern kor stilometriája felé. A *Federalist Papers* 1787 és 1788 között publikált esszék álnéven írt munkáit azonosították nyelvi megkülönböztető jegyek és valószínűségi modellek vizsgálatával.[9] Az 1980–1990-es években John Burrows új elemzési eljárásai, melynek a funkciószavak elemzése állt a középpontjában, jelentős eredményeket hozott a lexikális szintű elemzések végzéséhez.[10] A stilometriai vizsgálatok lehetőségei jelentősen bővültek a számítógépek és a mesterséges intelligencia térhódításával.[11]

Bár magyar nyelvterületen a stilometriai vizsgálat kevésbé elterjedt módszer, az utóbbi években eredményes kísérletek születtek magyar történeti szövegeken is. A Mikes-korpuszon végzett kísérletben először saját szerzőségű művek és fordítások, majd a műfaji-tematikai besoroláson alapuló művek beszédmód szerint kerültek elkülönítésre. Végül egy digitális szótár létrehozásával, mely a történeti szöveget normalizáló eszközként funkcionált, a vizsgálat hatékonysága is tesztelésre került.[12] Egy másik, a 19. századi magyar regényeken végzett kísérletben pedig a szerzőazonosítás leghatékonyabb távolságmértékeit kutatták.[13] Egy újabb kísérlet, mely Móricz Zsigmond levelezésének kritikai kiadásán alapszik,[14] az író feleségéhez és másokhoz írt leveleinek textuális és stilometriai sajátosságait vizsgálta. A vizsgálat stilometriailag mérhető

9 Holmes–Kardos, *I. m.*

10 John Burrows, *'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship*, Literary and Linguistic Computing 17. (2002/3.), 267–287.

11 1993–1994-ben Robert Matthews és Tom Merriam gépi tanulás segítségével létrehozott egy tanítókorpuszt Shakespeare és Fletcher műveiből, majd a *The Two Noble Kinsmen* című műben szerzői vizsgálatokat végeztek. Lásd: Holmes–Kardos, *I. m.*

12 Kiss, *I. m.*

13 Timári Mária – Bajzát Tímea Borbála – Palkó Gábor, *19. századi magyar regényeken végzett kísérletek a magyar nyelvű szerzőazonosítás leghatékonyabb távolságméréseinek megtalálására*, Eötvös Loránd Tudományegyetem – Hungarnet Egyesület, Budapest, 2021, 81–90. https://doi.org/10.31915/NWS.2021.9.

14 Móricz Zsigmond *Levelezése* (1892–1913), szerk. Cséve Anna, Petőfi Irodalmi Múzeum, Budapest, 2021. Lásd: https://dhupla.hu/collection/moricz-zsigmond-levelezes.

különbségeket mutatott ki Móricz Zsigmond feleségének, Holics Jankának és a másoknak szóló levelek között.[15]

„Ki a fenének kell collstok"[16]

A stilometriai vizsgálatok különböző szoftveres megoldásokkal (például R-Stylo, Websty, JGAAP) vagy egyedi programkódok segítségével végezhetőek. A Móricz-korpuszon végzett kísérletünkhöz az R nyelven írt *stylo* programcsomagot használtuk. A *stylo* csomag számos elemzési módszert tartalmaz, amelyet írott szövegek számítógépes elemzésére fejlesztettek, pontosabban stiláris értékük kvantitatív jellemzőinek analízisére.[17] A *stylo* minden nyelvhez tartalmaz egy integrált stop-szólistát, ami azért fontos, hogy a leggyakrabban előforduló, különösebb szemantikai vagy egyéb nyelvi, stilometriai szempontból funkcióval nem rendelkező szavak ne kerüljenek bele a vizsgálatba. A sikeres stilometriai elemzéshez az elméleti háttér alapos ismerete szükséges, hogy a szövegkorpuszon végzett, a vizsgálati szempontnak megfelelő mérési beállításokat adjuk meg, majd ezeket empirikus kísérletek révén helyesen tudjuk értelmezni.

A *stylo*ban használt statisztikai módszer lényege, hogy különböző tulajdonságok multidimenzionális terében[18] helyezi el az adott szövegeket. A vizsgálat során a beállításaink alapján releváns stílusjegyek, nyelvi vonások gyakoriságértékeit kapjuk meg, melyeket távolságmértékekkel elemzünk. A szövegek egymáshoz való viszonyát a köztük lévő távolság határozza meg. Mindegyik távolságmetrikai opció mögött egy lineáris algebrai képlet áll. A választható módszerek közül némelyik általánosabban használt matematikai képletek alapján működik, mások kifejezetten stilometriai elemzésre lettek kifejlesztve.

A vizsgálataink során három opciót is kipróbáltunk: az első kettőt, a klasszikus (vagy Burrows) deltát, illetve Eder deltáját kifejezetten stilometriai elemzésre fejlesztették. A normalizált szógyakoriságot mérő[19]

15 Cséve Anna – Kalcsó Gyula – Mihály Eszter, *Stylometric analysis of the correspondence of Zsigmond Móricz*, ACTA Universitatis, Sectio Linguistica XLVII., 2021, 137–146.
16 Móricz Zsigmond, *Tükör,* IV. = Móricz Zsigmond, *Tükör (1914–1917)*, szerk. Cséve Anna, Petőfi Irodalmi Múzeum, Budapest, 2021. Lásd: dhupla.hu, https://dhupla.hu/text/o:mzs-tukor.tei.4.
17 Eder–Rybicki–Kestemont, *I. m.*, 107–108.
18 Kiss, *I. m.*, 22–23.
19 *Uo.*, 23.

klasszikus delta az első olyan kifejezetten stilometriai, specifikusabban szerzőazonosítási célra létrehozott[20] távolságmetrika, amely széles körben ismertté és alkalmazottá vált. Így gyakran használják viszonyítási pontként a mérések során, ezért a mi kutatásunk is ezt a mérést vette alapul. Eder deltája ennek egy módosított változata, amely növeli a gyakrabban használt szavak súlyozását. Ezért második lépésben ezzel a metrikával mértünk a korpuszon, hogy megnézzük, van-e bármilyen markáns vagy akár kevésbé meghatározó eltérés a kapott eredmények között. Végül az általunk kipróbált harmadik metrika a nem kifejezetten stilometriai célokra kifejlesztett, ám ott is előszeretettel használt Canberra-távolság volt, mely beállítás kifejezetten érzékeny a ritkább szavakra.

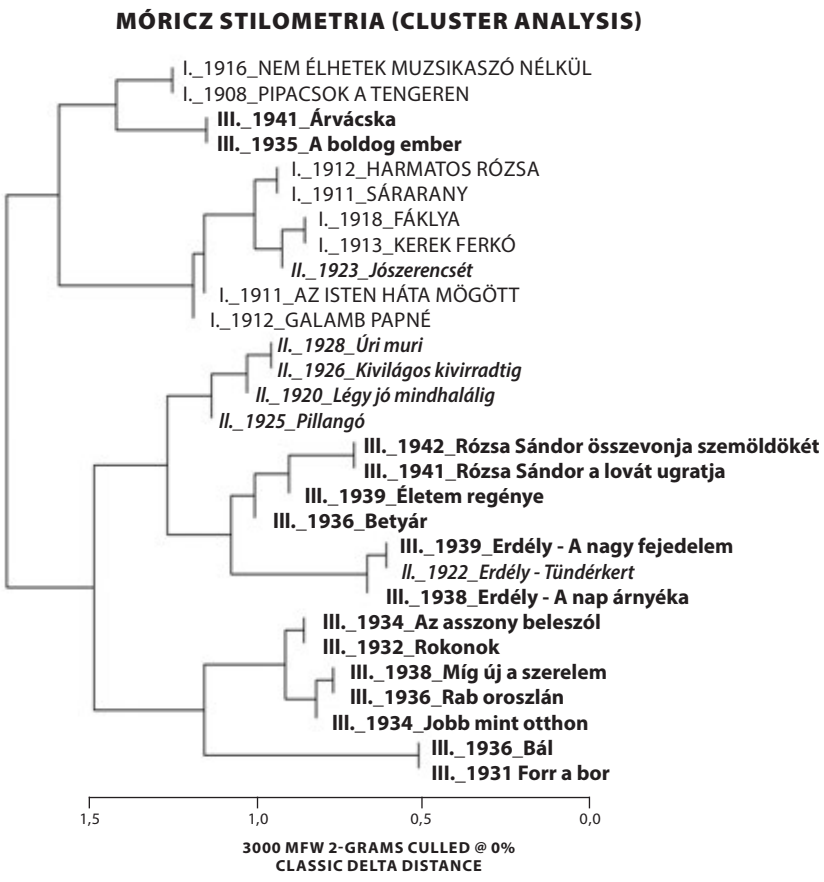
Vektorba zárt regények

Mivel az életmű legjelentősebb részét a regények teszik ki, valamint bizonyos alkalmazott távolságmetrikák érzékenyek a szövegek méretére, vizsgálatunk tárgyát kizárólag Móricz regényei képezik. Ennek eredményeképpen az életmű 29 darabjával foglalkoztunk szorosabban, ez Móricz összes regényének kb. 80 százalékát jelenti. Elsődleges szempont volt, hogy elektronikusan elérhető, letölthető, szerkeszthető szövegekből épüljön fel, ezért elsősorban a Magyar Elektronikus Könyvtárban található szövegekkel dolgoztunk, melyekből kiszűrtük az elbeszéléskötéteket és a meséket. Ha nem is hiánytalan a korpusz, mindenképp reprezentatívnak tekinthető. Ezt követte a fájlok konverziója, hiszen az elemzéseket javasolt plain text formátumú korpuszon végezni.

A méréseinket ábrázoló dendogramokon a korpusz elemei három kategóriára oszlanak, amit a regénycímek sorszámozása és a címek formázása jelöl. Eszerint az első csoportba az 1908-as *Pipacsok a tengeren* és az 1910-es években született regények, a második csoportba az 1920-as évek szövegei, a harmadik csoportba pedig az 1930-40-es évekbeli regények kerültek.

A korpuszon klaszteranalíziseket végeztünk, amely egy multidimenzionális statisztikai módszer, olyan csoportosító eljárás, melyben a korpusz elemeit bizonyos jellemzőik alapján hierarchikus csoportokba rendezzük. A klaszterek egymáshoz való elhelyezkedése is informá-

20 Burrows, *I. m.*



1. ábra. Móricz regényeinek elrendeződése dendogramon Classic Deltával mérve, a 3000 leggyakoribb szó-bigram alapján

cióval bír: minél közelebb kerülnek az alklaszterek elemei egymáshoz, annál nagyobb a közöttük lévő hasonlóság gyakorisága.[21]

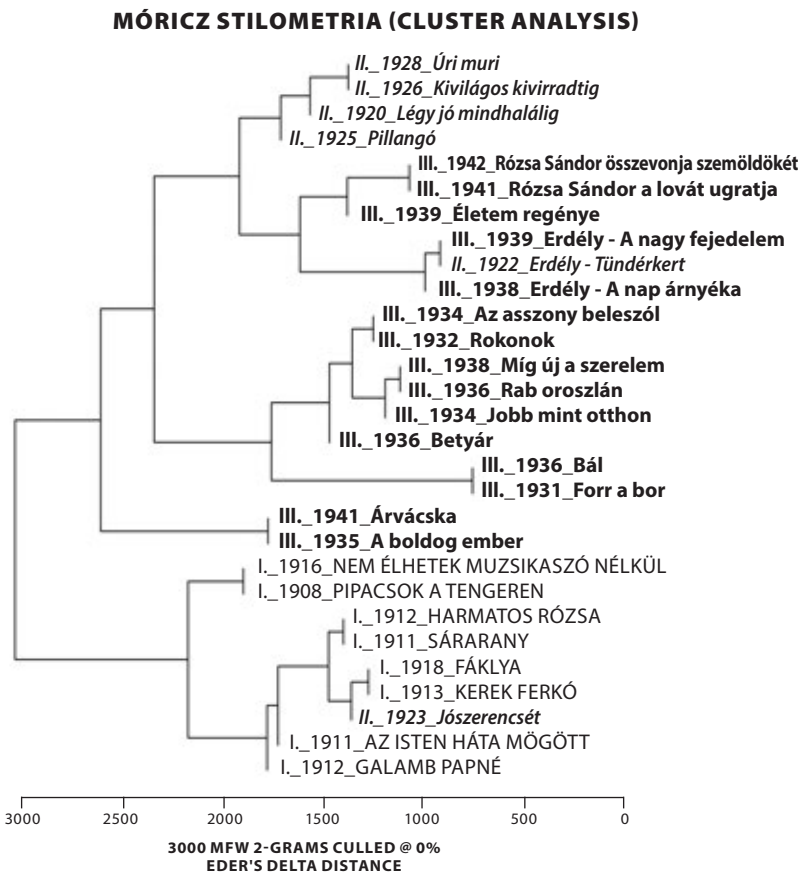
A korpuszon végzett első vizsgálat a 3000 leggyakoribb szó-bigram alapján készült. Mivel irodalmi szövegeken a szószerkezetek vizsgálata hatékonyabb módszer a szerzői stílus meghatározására, szó-bigramokat (két egymást követő lexéma) és nem unigramokat (egy szó) vizsgáltunk. A 25 leggyakoribb szó-bigramot nem vettük figyelembe,

21 Maciej Eder, *Visualization in stylometry: Cluster analysis using networks*, Digital Scholarship in the Humanities 32. (2017/1.), 50. Lásd: https://doi.org/10.1093/llc/fqv061.

annak érdekében, hogy kiszűrjük a stiláris funkcióval nem rendelkező, kevésbé releváns szókapcsolatokat. Mivel megkülönböztető jegyeket és nem hasonlóságokat szeretnénk kiválogatni, nem használtuk a cullingot, ami azt jelenti, hogy nem szűrtünk ki egyetlen szó-bigramot sem a szövegekből.[22] Elsőként a Burrows által fejlesztett Classic Delta távolságmértékkel végeztük el a kísérletet, ami a normalizált szógyakoriságot méri.[23]

Az ábrán megfigyelhető, hogy két nagyobb csoportba rendeződnek a szövegek, ezek pedig kisebb klaszterekre bomlanak. Megfigyelhető továbbá, hogy a korai és kései művek nagyrészt elkülönülnek, a korszakhatár pedig az 1920-as évek közepén húzódik meg. Az 1908 és 1923 között kiadott regények szorosan egymás mellett helyezkednek el, ugyanígy az 1930-as évek szövegei is egy nagyobb tömbbe rendeződnek. *A boldog ember* (1935) és az *Árvácska* (1941) című regények viszont középen helyezkednek el, és több rokonságot mutatnak a korai művekkel, mint a későbbiekkel. Mindkét szöveg összefüggésbe hozható Móricz világháborús feljegyzéseivel, a *Tükör*-kötetekkel. Móricz a megfigyelő pozíciójából jegyzeteli a „valóságot", ahogy az első kötet ajánlójában is olvashatjuk: „Ez a mi életünk nyers tükör képekben megrögzítve."[24] Számos dialógus, impresszió, történettöredék íródik a lapokra, miközben az elbeszélő figurájával és az elbeszélés lehetőségeivel kísérletezik Móricz. Ahogy Cséve Anna,[25] később pedig Szilágyi Zsófia is említi,[26] *A boldog ember*, illetve az *Árvácska* narratopoétikai szempontból összefügg a *Tükör*ben tett kísérletekkel, hiszen az oral history különböző eseteiként értelmezhetők. Mindhárom szöveg narratológiai tétje az elhangzott történetek, párbeszédek – tehát az élőbeszéd – irodalmi nyelven történő elbeszélhetősége. A *Tükör*-kötetekben olvasható, több száz emberrel folytatott beszélgetés sokféle módon kerül lejegyzésre, ezek sokféle módon kerülnek összefüggésbe a párhuzamosan vagy akár később írt szövegekkel. A stilisztikai értelemben megragadható közös nevező, a hangzó szöveg közegváltásával, leírásával áll kapcsolatban. A móriczi írásmódra mindvégig jellemző

22 Cséve–Kalcsó–Mihály, *I. m.*
23 Burrows, *I. m.*
24 Móricz Zsigmond, *Tükör I.* = Móricz Zsigmond, *Tükör (1914–1917)*, szerk. Cséve Anna, Petőfi Irodalmi Múzeum, Budapest, 2021. Lásd: dhupla.hu, https://dhupla.hu/text/o:mzs-tukor.tei.1.
25 Cséve Anna, *A tükör mint hallucinációs fal (Móricz Zsigmond kéziratos könyveiről)*, Alföld 2005/ 9., 74–82., 79.
26 Szilágyi Zsófia, *Móricz Zsigmond*, Kalligram, Pozsony, 2013, 221.



2. ábra. Móricz regényeinek elrendeződése dendogramon Eder Deltájával mérve, a 3000 leggyakoribb szó-bigram alapján

valóságvonatkozás referenciái tehát gyakran szövegelőzményekre mutatnak, elhangzott és lejegyzett dialógusokban, énelbeszélésekben rejlenek, ennek a módszernek eminens példája *A boldog ember* és az *Árvácska*. Az írásmódot és az íráshoz való hozzáállást egyaránt alakító módszer voltaképpen a pálya kezdeteitől jelen van,[27] ez magyarázza azt a tényt, hogy a tárgyalt szövegek közelebb kerültek az életmű első szakaszában írt regényekhez. Bár a kérdés megválaszolása további interdiszciplináris

27 Cséve, *Archívum és terep, írás és élet,* 303–320.

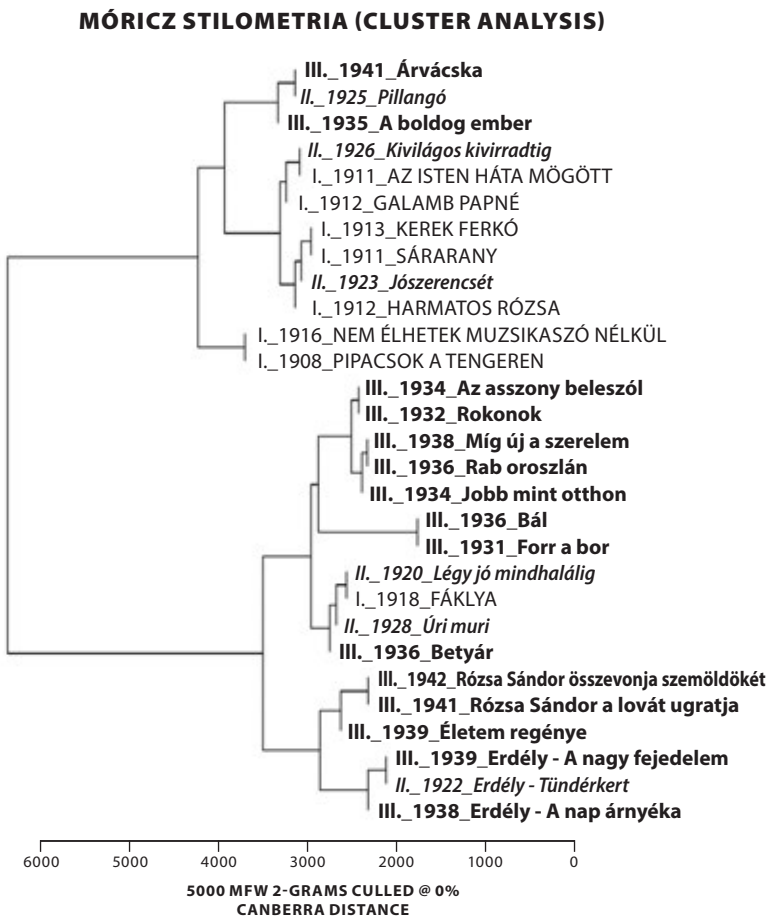
kutatást igényel, mégis, mindezek alapján indokoltnak tűnik rákérdezni, hogy a korai művek erőteljesebben közvetítik-e a hangzó nyelvet, az élőbeszédet. (2. ábra)

A kifejezetten stilometriai szövegvizsgálatokra fejlesztett Eder-féle delták gyakran hoznak pontosabb eredményt.[28] Ebből a megfontolásból készült Eder-deltájával a második vizsgálat. Látható, hogy különösebb eltérés nincs a szövegek klaszterizálása között, kizárólag annyi történt, hogy a két nagy klaszter helyet cserélt, de ennek nincs stilometriai jelentősége. Minél több mérés hoz hasonló eredményt, annál megbízhatóbb a paraméterezés. Tehát a második vizsgálat alátámasztja, hogy a regénykorpusz vizsgálatakor a 3000 leggyakoribb bigrammérés releváns eredményeket hoz.

A következő vizsgálat egy másik távolságmetrikával, a ritkán használt szavakra érzékenyebb Canberra-távolsággal készült. A Canberra-távolsággal elméletileg jobban láthatóvá válnak a szövegek közötti különbségek, például kimutatható egy írói életművön az egyházi, erkölcsi tematikájú és élőbeszédszerű művek elkülönülése.[29] Épp ellenkezőleg, a mi esetünkben nem jött létre több klaszter, sőt, egy nagyobb csoport elemei szóródtak szét. Az első két méréssel összevetve látható, hogy a korszakolás ebben az esetben kevésbé sikeres. Az 1920-as években írt szövegek (*Pillangó*, *Kivilágos kivirradtig*, *Jószerencsét*, *Úri muri*, *Légy jó mindhalálig*) távolabb kerülnek egymástól és elvegyülnek az 1910-es és 30-as évek szövegei között. Feltételezhető, hogy ezek a leginkább a regionális köznyelv regiszterében írt regények rugalmasabban sorolódnak egy-egy klaszter elemei közé, hiszen sokféle elemmel rokonítható tulajdonságokkal rendelkeznek.

A különböző beállításokkal végzett kísérletek kivétel nélkül elkülönítették Móricz pályakezdő műveit a későiektől. A korai művek esetében az egymáshoz közel került regények tematikailag eltérnek egymástól, más-más helyszíneket és társadalmi helyzeteket mutatnak be, a kisvárosi, dzsentri világon át a paraszti sorsokba nyújtanak betekintést. A társadalmi közeget reprezentáló beszédmódok megjelenítésével változnak az elbeszélői eljárások is. Onder Csaba a korai regények, a *Sárarany, Harmatos rózsa, Galamb papné, Az Isten háta mögött, Kerek Ferkó, Jószerencsét* újraolvasásakor arra kereste a választ, hogyan írhatott Móricz alig néhány év alatt regénypoétikai és színvonal tekintetében is egy-

28 Cséve–Kalcsó–Mihály, *I. m.*
29 Kiss, *I. m.*, 28–29.



3. ábra. Móricz regényeinek elrendeződése dendogramon Canberra távolságmetrikával mérve, az 5000 leggyakoribb szó-bigram alapján

mástól ennyire eltérő műveket. Állítása szerint Móricz „hat, közel egyforma terjedelmű, de egymástól eltérő regényben írja szét a *Bovarynét*, a magyar valóságra alkalmazva lekicsinyíti és modernizálja a 19. századi realista nagyepikát és naturalizmust. A regényvilágok hasonlósága, a választott alkotó elbeszélői eljárások különbsége ellenére azonban látható: a hétköznapi, de fordulatos történetek látszólag mind az ún. »magyar valóság« különféle színterein játszódnak, a naturalista iskola

életdarab technikáját követve."[30] Bár ezek olyan aspektusok, melyek statisztikai szövegvizsgálatokkal kevésbé kimutathatóak, a korai művek hasonlóságának stilometriailag releváns gyakorisága kétségtelenül rávilágít arra az írói stílusra, amely Móricz pályakezdő szakaszát átfogóbban jellemzi.

Az ábrákon jól láthatóan Móricz történelmi regényei, az *Erdély-trilógia* és a *Rózsa Sándor*-szövegek az összes mérés alkalmával egymás mellé kerültek, ez jelzi a művek közötti stiláris hasonlóságot, viszont a tematikai hasonlóság következményeként is értelmezhető. Móricz mindkét regényét hatalmas forrásfeltáró munka előzte meg. A felhasznált művekről Csáky Edit készített listát,[31] amelyek között Bethlenről és a korabeli Erdélyről szóló kortárs szaktudományi munkákat, monográfiákat is megtaláljuk. Ezek a szövegek egyaránt segítettek a korszak megismerésében és az archaizáló nyelv megalkotásában. Az a gyakorlat, hogy Móricz szövegreferenciákkal dolgozik és archaizáló nyelveket teremt újra műveiben, túlmutat a tematikai hasonlóságon, és ráirányítja a figyelmet a szövegépítkezésre. Az archaizálás összeköti a történelmi regényeket, de stilometriai kérdésként fontos lehet a forrásanyagok vizsgálata is. A forrásfeltáró munkák alapján felmerül tehát a kérdés, hogy vajon tetten érhető-e stilometriai vizsgálatokkal a felhasznált forrásanyag és az Erdély-regények közötti kapcsolat. Felmerül az is, hogy feltérképezhető-e a Rózsa Sándor-szövegek és a trilógia közös forrásanyaga stilometriai mérésekkel, és ez mennyiben járul hozzá a regények közötti hasonlósághoz. A kérdések megválaszolása újabb irányt adhat a stilometriai kísérleteknek mint forrásfeltáró vizsgálatok eszközének, ugyanakkor új perspektívával egészítené ki a Móricz forráskezelését tárgyaló összevetéseket.

Mivel a szövegek elhelyezkedése egymás viszonylatában történik, felmerült a gyanú, hogy a korpuszban túl nagy hangsúllyal van jelen az *Erdély-trilógia* és a *Rózsa Sándor* szövegegyüttes, ami torzíthatja az eredményt. Ezért egy következő mérés során kivettük a korpuszból ezeket a regényeket. Az előző mérések beállításait megőrizve, a 3000 leggyakoribb szó-bigramokat vizsgáltuk, Eder Delta távolságmértékkel. A második ábrán látható, Eder Deltával végzett vizsgálattal összehasonlítva, a trilógia és a *Rózsa Sándor*-regények nélkül végzett vizsgálatban még mindig elkülönülnek a korai művek a késeiektől, viszont

30 Onder Csaba, *Hangérien bjuti (szex, lektűr, irónia, avagy Móricz Zsigmond indulása)* = *Az újraolvasott Móricz*, szerk. Uő., Nyíregyházi Főiskola, Nyíregyháza, 2005, 123–124.

31 Csáky Edit, *Az* Erdély *történeti forrásai*, Irodalomtörténet 1979/4., 767–785.

a húszas években íródott regények, az *Úri muri* (1928), a *Kivilágos kivirradtig* (1926), a *Légy jó mindhalálig* (1920) és a *Pillangó* (1925), már nem kerülnek egy klaszterbe a harmincas években írt regényekkel, hanem a koraiakhoz sorolódnak. Ez alapján az 1920-as években írt regények egy átmeneti időszakot képviselnek az életműben, melyek ahhoz a pályaszakaszhoz gravitálnak, ami túlsúlyban van.

A korszakolás szempontjából jól láthatóan elkülönülnek a korai és késői művek, ezért ebből kiindulva, a következő kísérletünkhöz a szövegkorpuszt két alkorpuszra osztottuk, a néha vándorló, mindkét szakaszba besorolódó 1920-as évekbeli műveket pedig az első szakaszhoz csatoltuk, mivel legtöbb vizsgálatunk inkább a korai szakaszhoz kapcsolja ezeket. Tehát a korpuszunkat az eddigi eredményeink alapján osztottuk két részre, vagyis a korai és a kései művekre, ahol a korszakhatár az 1920-as évek közepén húzódik meg. Mivel az *Árvácska* és *A boldog ember* a legtöbb esetben a korai művekhez sorolódott be, ezért a korpusz összeállításánál mi is a korai művek csoportjába tettük, az 1922-ben kiadott *Tündérkert* pedig az *Erdély-trilógia* részeként a második alklaszterbe került.

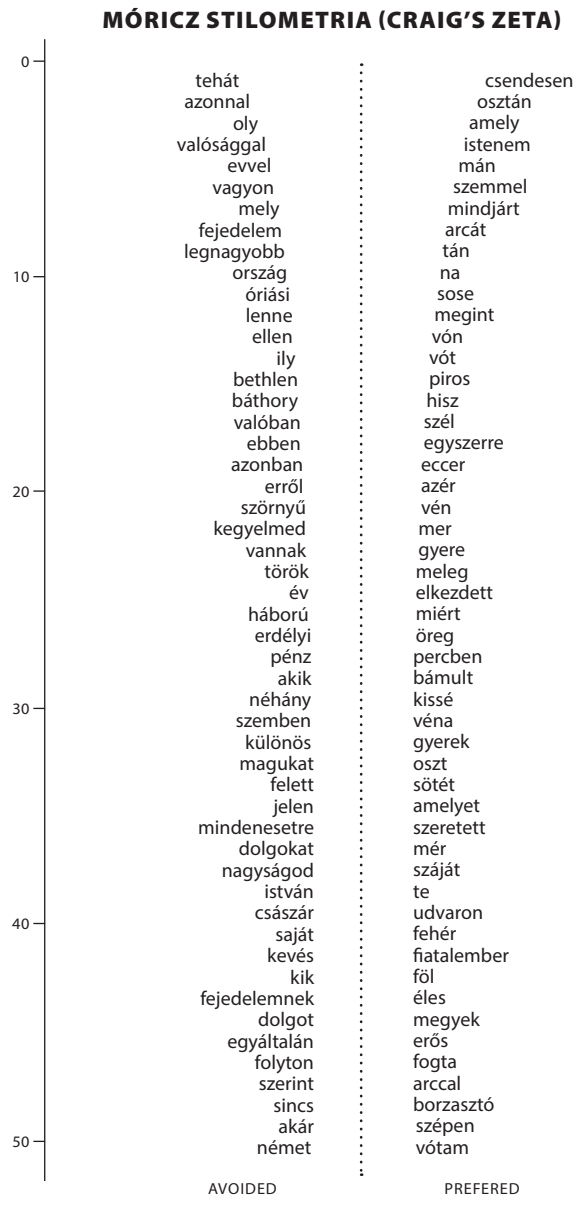
A használt oppose függvény a stylo csomag egyik funkciója, diszkriminanciaanalízisen alapuló, kontrasztív szövegelemző módszer, amelyet két szövegkorpusz összehasonlítására használnak. A függvény a két korpusz szövegét egyenlő méretű szeletekre bontja, ami a mi beállításunkban 3000 szóból áll. Az előfordulási küszöböt (occurence threshold) 2-re állítottuk, így a kevésbé gyakori szavak is bekerülnek az összehasonlításba. A szűrési küszöb (filter threshold) 0.1-es értékű, ami a gyenge diszkriminációs erősségű szavak kiszűrését jelenti.

A vizsgálathoz az eredetileg Burrows által fejlesztett, Craig által kibővített Zeta attribúciós tesztet használtuk,[32] ami nem a leggyakrabban használt szavakra koncentrál, hanem a szófrekvencia-spektrum közepét veszi figyelembe.[33] A függvény két szólistát állít elő egy diagramon. Az egyik, a korai művek kifejezetten preferált szavait tartalmazza a kései regényekhez képest. A másik szólista pedig az első korpusz kifejezetten került szavait tartalmazza a második szövegkorpuszhoz képest.[34]

32 Bővebben lásd az útmutatóban: Maciej Eder – Jan Rybicki – Mike Kestemont – Steffen Pielstroem, *Stylometric Multivariate Analyses,* 2020, 42–44., http://cran-r.c3sl.ufpr.br/web/packages/stylo/stylo.pdf. (Utolsó hozzáférés: 2022. 01. 27.)

33 Tatiana Litvinova – Olga Litvinova, *Analysis and Detection of a Radical Extremist Discourse Using Stylometric Tools = Digital science 2019*, szerk. Tatiana Antipova – Álvaro Rocha, Springer, Berlin, 2020, 37–38.

34 Christof Schöch – Daniel Schlör – Albin Zehe – Henning Gebhard – Martin Becker



4. ábra. Móricz korai és késői regényeinek összehasonlítása az oppose függvénnyel mérve

A mérésből láthatóvá válik, hogy az első csoportba tartozó művek preferált lexémái között gyakoriak a tájnyelvi elemek: 'osztán', 'mán', 'vót', 'vón', 'vóna', 'eccer', 'azér', 'oszt', 'vótam', 'má' stb. Ezek nem kizárólag tájnyelvi jellegük miatt érdekesek, hanem a hangzó nyelv móriczi alkalmazása szempontjából is, mely szövegépítkezésének egyik jellemzője.

A második csoport nem tartalmaz kifejezetten tájnyelvi, vagy a hangzó beszédre jellemző kifejezéseket,[35] ellenben túlsúlyba kerültek az anyagi világhoz, történelemhez, társadalomhoz kapcsolható kifejezések. Az *Erdély-trilógia*ból ismert történelmi személyiségek nevein túl előfordulnak a 'pénz', 'fejedelem', 'ország', 'kegyelmed', 'török', 'háború', 'erdélyi', 'nagyságod', 'császár', 'dolog', 'német', 'magyar' kifejezések, melyek különböző szemantikai mezőkön helyezhetőek el, mégis összekapcsolhatóak.

Mintha az írói tekintet a premier plánban készült portré műfajától eltávolodva, átkapcsolt volna a nagytotálban történő ábrázolásmódra. Viszont ebben az esetben is az eredményt túlságosan befolyásolhatják a kifejezetten történelmi regények, éppen ezért egy következő mérést végeztünk az *Erdély-trilógia* és a *Rózsa Sándor*-szövegek nélkül.

Ebben az esetben az első alkorpusz kifejezetten kedvelt szavai között ismét előfordulnak a már említett nyelvjárási elemek, emellett megfigyelhető, hogy mennyire hangsúlyosan vannak jelen az archoz kapcsolódó lexémák: 'szemmel', 'arcát', 'piros', 'száját', 'öreg', 'bámult', 'szemébe', 'szeme', 'arccal'. Móricz műveinek arc- és testábrázolási szempontú vizsgálata egyre inkább a recepció középpontjába helyeződik.[36] Feltárásra került, hogy a *Tükör*-kötetek bővelkednek a portrékban: „Móricz a Tükörben szereplő arcleírásokban az emberi arcnak a teljes személyiséget képviselő, esszenciális egységét kereste, az arc lényegének, a személyiségnek a megragadhatóságával, leírhatóságával kísérletezett."[37] Vizsgálatunk azt mutatja, hogy ez az érdeklődés, az arc megragadhatóságának, a fej ábrázolhatóságának kérdése nemcsak műhelynaplójában,

– Andreas Hotho, *Burrows'Zeta: Exploring and Evaluating Variants and Parameters*, 2018. Lásd: https://dh2018.adho.org/burrows-zeta-exploring-and-evaluating-variants-and-parameters/.

35 Kivételt képez az 'evvel' szó, mely a mai nyelvhasználatban perifériára került, de a század első felében bár stilisztikailag lehetett jelentősége, nem biztos, hogy kevésbé volt elterjedt forma.

36 Benyovszky Krisztián, *Némi-nemi bizonytalanság. Női és férfi szerepek inverziója az Árvalányokban* = Uő., *Fosztogatás. Móricz-elemzések*, Kalligram, Pozsony, 2010, 40–65.; Baranyai Norbert, *„Az ember még a saját testének sem ura". Test és elbeszélés összefüggései a Sárarany egy fejezetében*, Irodalmi Magazin 2016/1., 61–64.

37 Tardy Anna, *Tükör-élet-mű. Móricz Zsigmond Tükör-kötetei*, Első Század 11., (2012/2.), 327.

hanem a korai művekben is olyan hangsúlyokkal jelentkezik, mely az első írói szakaszt megkülönbözteti a későbbi művektől.

A második kategóriába került szavak részint szintén társadalmi vonatkozással bírnak, bár a tematikai meghatározottság nem annyira erőteljes: 'év', 'pénz', 'dolog', 'dolgokat', 'társadalmi', 'pengő', 'jelen', 'éve', 'csinálni', 'életben', 'időben', 'család', 'évig'. Mégis a világnak kevésbé az érzelmi, mintsem az anyagi részéhez kapcsolódnak a használt kifejezések.

Összegzés

Kutatómunkánk során kiindulópontunkat több irányból is megszabva, összehasonlító célzattal végeztünk méréseket, majd az eredményeket előfeltevéseinkkel, a recepciótörténet egy-egy vonatkozó állításával, illetve egymással is összevetettük. A stilometriai elemzés kvantitatív eredményekre hagyatkozva képes rámutatni bizonyos stílusbeli jellemzőkre, ugyanis kísérletünkben viszonylag pontosan különváltak az e szempontból releváns írói „korszakok". Természetesen, messzemenő következtetések kizárólag a szövegek szoros olvasatával együtt, a Móricz-recepció feldolgozásával vonhatók le.

A korai művek (1908–1920) egy klaszterbe rendeződtek, ahogy az 1930 után írt regények is. Az 1920-as években írt regények, amelyek az életmű középső szakaszát képezik, a mérési beállításoktól függően kapcsolódtak a korai vagy kései szövegekhez, jellemzően köztességet képviselve. A történelmi regények (*Erdély-trilógia* és *Rózsa Sándor*-regények) szorosan illeszkedtek egymáshoz méréseink alkalmával, melynek oka a létrehozott archaizáló nyelvben keresendő.

A Móricz-korpuszon végzett vizsgálataink releváns eredményeket hoztak az életmű belső összefüggése tekintetében, emellett olyan irodalomértelmezési kérdéseket is, amelyek újabb kutatások kiindulópontjai lehetnek. A Móricz-regények forrásainak azonosítása (melyek egyaránt építenek a hangzó beszédre és leírt szövegekre) további történeti és filológiai kutatások tárgya lehet. A gazdag szakirodalom és kéziratanyag gépi támogatású összevetése a művekkel, eddig feltáratlan metszéspontokra közelíthet rá. Emellett megfelelő paraméterezéssel a Móricz-korpusz nyelvi összetétele is vizsgálható, felderítve a regionális köznyelv, tájnyelv és archaizáló nyelv használatának mértékét és módozatait.