Makkai T. Csilla – Varga Emese

Csomópontok és metszetek

Móricz Zsigmond műveinek hálózatos megjelenítése – Esettanulmány

Módszertani kérdések

A klasszikus irodalomértelmezés jellemzően a szavak, mondatok, történetek szemantikai regisztere felől közelít a szövegekhez. Az elemeire bontható, egyedi struktúrákba rendeződő hálózatként elgondolt műalkotás vizsgálata új perspektívát jelenthet egy olyan szakterületen, ahol az általános tapasztalat mégiscsak az, hogy a hangsúlyok időről időre eltolódnak, a recepció az éppen aktuális tendenciáknak megfelelően változik.

Móricz életművének recepciójára is hatással voltak a kanonizációs eljárások, viszont ennek részletes felvázolása egy másik tanulmány témája lehetne. A korábbi tendenciákat ismerve és a 2000-es évektől kezdődő Móricz-újraértelmezéshez kapcsolódva felvetjük a kérdést, hogy mit állíthatunk teljes bizonyossággal, mennyire ismerhető meg kutatásunk tárgya, Móricz Zsigmond terjedelmes életműve? Illetve az eddigi irányokat kiegészítve, mit tudunk meg az írói életműről, ha egy új módszert alkalmazunk? Megszámoljuk azokat a nyelvi jegyeket, amelyek megkülönböztető funkcióval bírnak a szerző stílusára nézve: pl. a szókészleti gazdagság, a szintaktikai jellemzők, a szókörnyezet és ezek eloszlását vizsgálva egymással összevetjük a szövegeket.

A pálya alakulástörténetének vizsgálatához gépi elemzési módszert használtunk, a statisztikai alapokon működő stilometriát. Ez egy automatizált szövegelemzési módszer, amely képes nagyméretű szövegkorpuszok elemeinek összehasonlítására. Az elemzés egy részben kontrollálható folyamat, hiszen a kutató választja ki a vizsgálatok paramétereit, hogy a szoftver a teljes szólista hány százalékát használja fel az elemzéshez, és hány százalékát vonja ki az elemzésből, mennyi és mekkora mintát vegyen ki a szövegből, melyik távolságmértéket használja. Ezáltal az algoritmus számszerűsíti a szövegek egyes nyelvi jellemzőit és azok eloszlását a textusban, valamint a távolságmértékek szerint veti össze a korpusz elemeit. Ilyen módon a kimenetként kapott ábrákról a szövegek egymáshoz való viszonya olvasható le.

A stilometriát elsősorban bizonytalan szerzőségű szövegek alkotójának azonosítására alkalmazzák, hisz ez a statisztikai módszer megkülönbözteti a szerzői stílusok közötti különbségeket a szövegekben.[1] Így az úgynevezett „szerzői ujjlenyomat" kerül detektálásra, mivel a szoftver a szövegek tematikus szintje alatt meghúzódó nyelvi jellemző-

[1] Lásd Patrick Juola: The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions, Digital Scholarship in the Humanities, 2015. 1. sz. 100–113.; Maciej Eder: Elena Ferrante: Egy „virtuális" szerző, Digitális Bölcsészet, 2021. 5. sz. 3–17.

ket térképezi fel. Ez főként a nem tudatosan használt szavak/szókapcsolatok, vagyis a jelentés nélküli funkciószavak eloszlásán és gyakoriságának vizsgálatán keresztül történik. Kutatások azt bizonyítják,[2] hogy a funkciószavak eloszlása fontosabb stílusmarker, mint gondolnánk, ráadásul emberi olvasással nem mérhető vetülete a szövegnek.

A stilometria n emcsak szerzőazonosításra lehet alkalmas, hanem a szerzői életmű szóhasználatváltozásának feltérképezésére, korszakolására is.[3] Az írói ujjlenyomat változásai a jelentés nélküli funkciószavak eloszlásának és gyakoriságának vizsgálatával szintén fellelhetőek. Ezek a jelenségek emberi szemmel, szoros olvasás révén nehezen érzékelhetőek, így ebben a számítógépes nyelvészetnek lényeges szerepe van.

Tanulmányunkban többféle mérést és vizualizációt mutatunk be, arra keresve a választ, hogy a stilometriailag kimutatható-e a szövegek publikálásának időrendje, megcélozható-e valamiféle korszakolás, láthatóvá válnak-e eltérések vagy új összefüggések Móricz regényeinek hálózatában?

Szövegműveletek, mérések

A vizsgált szövegkorpusz 30 regényből áll, amelyeket a Magyar Elektronikus Könyvtár szolgáltatásából értünk el. Az életműből kizárólag a regényeket vizsgáltuk, mivel bizonyos alkalmazott távolságmetrikák érzékenyek a szövegek hosszúságára. A kiválasztott 30 mű Móricz összes regényének kb. 80 százalékát jelenti; ez reprezentatív korpusznak tekinthető. A regények között előzetesen három kategóriát állítottunk fel: az elsőbe az 1910-es évek regényei, a másodikba az 1920-as években, a harmadikba pedig a kései, 1930–1940-es években kiadott művek tartoznak.

A Móricz-korpuszon végzett első stilometriai kísérletünk egy korábbi klaszteranalízis volt.[4] Ennél az elemzésnél a 3000 leggyakoribb szó-bigramot (tehát két egymás után következő lexémát) vizsgáltuk a szövegekben, illetve nem használtunk cullingot, ami azt jelenti, hogy nem szűrtünk ki egyetlen szó-bigramot sem a szövegekből. A használt távolságmérték pedig a szövegvizsgálatokra fejlesztett Eder deltája volt.[5]

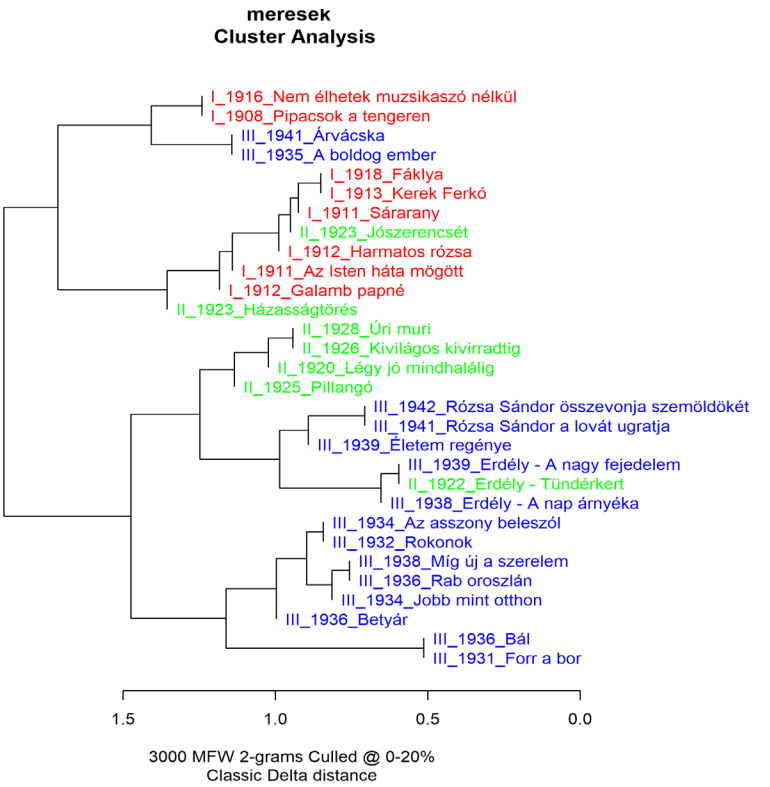
A vizsgálat eredményét ábrázoló dendrogramon a Móricz-szövegek két nagy ágra oszlanak, ahol a határvonal az 1920-as évek regényeinél húzódik meg. A két klaszteren belül pedig nagyrészt időrendi sorrend állt fel a művek között. Megfigyelhető azonban, hogy az 1923-as *Jószerencsét* és *Házasságtörés* is a korai regényekkel került egy csoportba. Ezek a művek tematikailag nagyon eltérnek egymástól, mint ahogy általában az 1910-es években írt regények; ezért ebből arra következtethetünk, hogy a

[2] Lásd Harald BAAYEN: Word Frequency Distributions, Dodrecht, 2001, Kluwer; John Burrows: „Delta": A Measure of Stylistic Difference and a Guide to Likely Authorship, Literary and Linguistic Computing, 2002. 17. sz. 267–287.

[3] Lásd Lutosławski, Wincenty: The Origin and Growth of Plato's Logic: With an Account of Plato's Style and of the Chronology of his Writings, London, 1897, Longmans; Van Hulle Dirk – Kestemont Mike: Periodizing Samuel Beckett's Works: A Stylochronometric Approach, Style, 2016. 2. sz. 172–202.

[4] Makkai T. Csilla – Varga Emese – Vétek Bence: Egy életmű metszete. Stilometriai vizsgálat Móricz Zsigmond regényein, Szépirodalmi Figyelő, 2022. 1. sz. 71–85.

[5] Maciej Eder: Taking stylometry to the limits: Benchmark study on 5,281 texts from Patrologia Latina, In. Digital Humanities 2015: Book of Abstracts, Szerk. Sydney, 2015, University of Western Sydney.



művek hasonlósága nem tematikai összefüggésekben áll, hanem a szóhasználat szintjén húzódó stílusbeli kapcsolatokban.

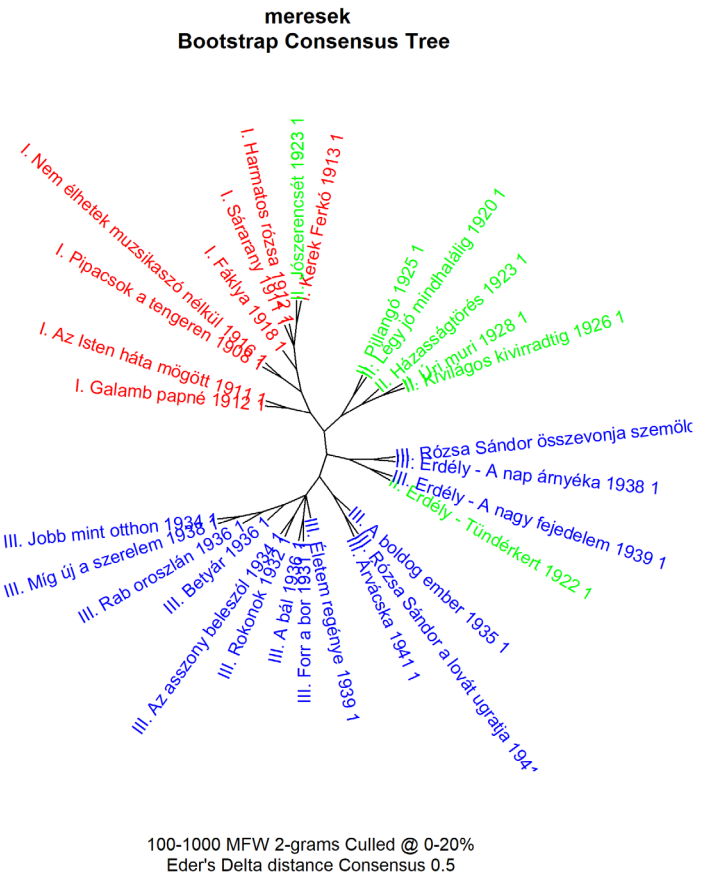
Az *Árvácska* (1941) és a *Boldog ember* (1935) is több rokonságot mutat a korai művekkel, mint a későbbiekkel. Mindkét szöveget az elhangzott történetek, párbeszédek – tehát az élőbeszéd – irodalmi nyelven történő elbeszélhetősége jellemzi. Ez arra enged rákérdezni, hogy a korai művek érőteljesebben közvetítik-e a hangzó nyelvet, az élőbeszédet.

Mivel a különféle paraméterek részben eltérő eredményeket hoznak, egy kutatás során több mérést ajánlott elvégezni, és összehasonlításuk alapján értékelhetjük az eredmény megbízhatóságát. A megbízhatóság maximalizálása és a „véletlenek" kiszűrése létező probléma a stilometriában, ezért egy olyan ábrázolási mód került kifejlesztésre, amely több mérést képes egy ábrán vizualizálni. Ezen voltaképpen az egyváltozós, többszöri mérések együtteséből vont konszenzus kerül megjelenítésre.

A konszenzusfa gyakorlatilag egy sorozatnyi klaszterezés eredményeit összegzi egy ábrába,[6] jelen esetben a beállított paramétereink szerint, a leggyakoribb 100–1000 szavas tartományban vizsgált a szoftver. A mérést mintavétellel készítettük, ami annyit

[6] MACIEJ EDER: Visualization in Stylometry: Cluster Analysis Using Networks, Digital Scholarship in the Humanities, 2017. 1. sz. 50–64.

jelent, hogy a szoftver a textusokból random 15000 szó hosszúságú részleteket vetett össze. Mivel a szöveghosszak közötti eltérések torzíthatják az eredményt, szükséges volt az elemzéshez a mintavételezésre. Az iterált mérés során a szógyakorisági lista hosszúsága változott: az algoritmus először a 100 leggyakoribb szó-bigramig (tehát két egymás után következő szóig), majd a 200 leggyakrabban előfordulóig, egészen az 1000 leggyakoribb egymást követő szóig vizsgálta a részleteket. Ezen kívül meghatároztuk a konszenzusküszöböt is, tehát azt, hogy az elvégzett mérések hány száza-



lékában jelenjenek meg a hasonlóságok ahhoz, hogy az algoritmus figyelembe vegye azt. Mi egy közepes erősségű, 0,5-ös konszenzusküszöböt állítottunk be. A vizsgálat eredménye az alábbi ábrán látható.

A klaszteranalízissel ellentétben a konszenzusfán a szövegek egymáshoz való távolsága kevésbé olvasható le, viszont jól megfigyelhetőek az egyes szövegcsoportok, melyek szorosabb kapcsolatot mutatnak egymással, illetve láthatóvá válnak azon szövegek, amelyek nagyobb stílusbeli eltérést mutatnak a többitől. „A konszenzusfa másik fontos különlegessége, hogy egy plusz paramétert is magával hoz: a konszenzus erősségét.

Ezzel a paraméterrel lehet megadni számára, hogy a klaszterezési eljárások hány százalékában kellett közvetlen kapcsolatot találnia a két szöveg között ahhoz, hogy itt is közvetlen kapcsolatban levő szövegként jelenjenek meg."[7]

Ezt a paramétert a mérési beállításoknál adjuk meg, a mi vizsgálatunk esetében ez egy közepesen erős mérték volt. A Móricz-regények a konszenzusfán három nagy szövegcsoportra oszlanak: a korai, a '20-as évekbeli és a kései regényekre. A sorozatnyi klaszterezés tehát a dendrogramhoz hasonlóan egyfajta kronológiát eredményezett a konszenzusfán. Hasonlóképpen, a *Jószerencsét* itt is a korai regényekhez sorolódik, a *Házasságtörés* viszont már a '20-as évek csoportjába kerül. Mivel a *Jószerencsét* megírásának körülményei ismertek,[8] tehát az a tény, hogy Móricz 1914-ben folytatásokban kezdte el közölni a regényt a *Nyugat*ban, majd az első világháború kitörése megszakította a szövegfolyamot, egyáltalán nem meglepő, hogy stilometriai értelemben a szöveg a '10-es években kiadott szövegek közé sorolódik. A regény végül csonka változatban, 1923-ban jelent meg, a könyv végén pedig Móricz is értékeli a félbehagyott szöveget: „Ez a regény a Világháború által torzóvá szakadt. Írtam, mert az Élet diktálta; nagynak indult, bőségesnek, fulladtnak, sokrétűnek, mint volt Akkor a Sorsunk. // S jött a Háború: és felrobbantotta a regényt, mint minden életet. Azóta kilenc év múlt el: sem változtatni, sem folytatni, sem kiadni nem tudtam; ma itt a könyv, mint történelmi dokumentum."[9]

A Móricz-kutatók egyetértenek abban, hogy a világháború változást hoz az írói kifejezésmódban;[10] ezt az időszakot inkább az új beszédmódok keresése jellemzi: „Az életműben azonban változást tapasztalhatunk. Aktívan folytatta ugyan az írást, mégis a *megjelentetett* művek száma elmarad a korábbi tempóhoz képest. Nagyobb terjedelmű művel, regénnyel csak 1917-ben jelentkezik. A háború alatt főként rövidebb műfajokban közöl írásokat, ez az időszak leginkább az új forma, az új téma keresésének idejeként jellemezhető."[11] A stilometriai kísérlet kimutatta, hogy a *Jószerencsét* nyelvhasználati szempontból a korai írói korszakhoz sorolható, emellett pedig azt, hogy a világháború lezárását követően, hozzávetőlegesen a '20-as évek környékére már egy másik írói korszak kezdete tehető.

Mivel azt feltételezzük, hogy ez a változás sokkal rétegzettebb, komplexebb folyamat, folytatni szerettük volna a regények vizsgálatát más beállítási paraméterekkel és vizualizációkkal. Célunk a művek mélyebb, nyelvhasználati összefüggésrendszerének, hálózatosságának megmutatása volt.

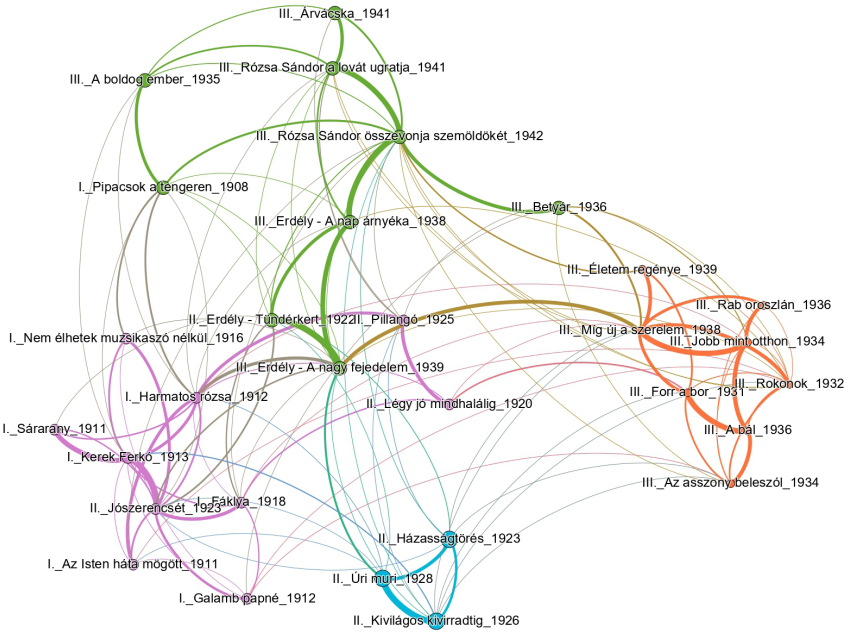
[7] Szakács Béla Benedek: Szövegelemzési módszerek automatizációja, BME VIK, TDK-dolgozat, 2018. Lásd: http://tdk.bme.hu/VIK/SW5/Szovegelemzesi-modszerek-automatizacioja.

[8] Szilágyi Zsófia: Móricz Zsigmond. Monográfia. Pozsony, 2013, Kalligram, 175 [Szilágyi 2013]; Cséve Anna: A Tükör mint hallucinációs fal. Alföld, 2005. 9. sz. 74–81.

[9] Móricz Zsigmond: Jószerencsét, Regények I., Budapest, 1975, Szépirodalmi Kiadó, 834.

[10] Szilágyi 2013, 218.; Tardy Anna: Tükör-élet-mű. Móricz Zsigmond Tükör-kötetei, Első Század, 2012, 2. sz. 315–354. 321.; Tardy Anna: Levelek a frontról. A Móricz-hagyaték emlékei az első világháború kezdeti időszakáról, In: Emlékezés egy nyár-éjszakára. Interdiszciplináris tanulmányok 1914 mikrotörténelméről, Szerk. Kappanyos András, 2015, MTA Bölcsészettudományi Kutatóközpont, 105–115. [Tardy 2015.]

[11] Tardy 2015, 109.



A konszenzusfa előállítása mellett a mérés [2. ábra] olyan adatbázist eredményez, amelyen a megadott számú jellemzők alapján a szövegek egymáshoz való viszonya jelenik meg. A szövegek közötti konszenzus erőssége számokkal van súlyozva. Ez képezi a konszenzusháló alapját, amelyet a Gephi szoftverbe[12] importálva egy hálózatos ábrán tudtunk megjeleníteni. Ez a vizualizációs módszer a kapcsolatrendszer egyszerűsített ábrázolását teszi lehetővé, mely az aktuális szempontoknak megfelelően kiemeli az összefüggéseket, így az eredmények könnyebben értelmezhetővé válnak. A hálózatban már nemcsak az egymáshoz közeli szövegcsoportok láthatóak, hanem az egyes szövegek összes kapcsolata és azok erőssége is megjelenik.

A kapcsolat erősségét az élek vastagsága, valamint a csomópontok közötti távolság jelöli. Ez által láthatóvá válik, hogy valamely szöveg melyik másik regénnyel rokonítható stilometriai értelemben. A kapcsolatok számának megfelelően pedig növekszik a regényt jelző csomópont mérete is. A színek az egyes klasztereket jelölik, vagyis azon regények csoportját, amelyek szorosabb stílusbeli hasonlóságot mutatnak egymással. Eszerint láthatóvá válik, hogy a regények nagyrészt korszakok szerint különülnek el egymástól. A lila klaszterbe a korai regények kerültek: főként a *Sárarany*, *Kerek Ferkó*, *Jószerencsét*, *Fáklya*, *Isten háta mögött* mutatnak szoros kapcsolatot egymással. A kék klaszterbe a huszas évek regényei közül a *Házasságtörés*, az *Úri muri*, a *Kivilágos kivirradtig* került, viszont az is megfigyelhető, hogy egyes művek inkább későbbi regényekkel mutatnak szorosabb rokonságot, ilyen a *Pillangó* (1925) és a *Légy jó mindhalálig*

[12] Gephi, hozzáférés: 2022. 09. 20., https://gephi.org/.

(1920). A harmincas években megjelent művek, az *Életem regénye*, *Míg új a szerelem*, *Rab oroszlán*, *Forr a bor*, *Rokonok*, *A bál*, *Az asszony beleszól* a piros színű klaszterbe tartoznak. A szintén harmincas években kiadott *Betyár* (1936) és *Boldog ember* (1935) szorosabb kapcsolatot mutatnak a későbbi, Rózsa Sándor-regényekkel.

Ha azt szeretnénk megvizsgálni, hogy melyik az a szöveg, amelyik a legtöbb regénynyel áll kapcsolatban, tehát melyik az a Móricz-regény, mely az életmű „origójának" tekinthető, akkor azt látjuk, hogy kiemelt helyet foglal el az Erdély-trilógia, illetve a Rózsa Sándor-regények csoportja. Gyanakodhatunk arra, hogy a szövegek hosszúságából adódó nagy szókincs meghatározó tényező ebben, viszont mivel a mérés mintavételezéssel készült, tehát random kiragadott egyforma hosszúságú szövegrészeket vetettünk össze, a torzítás lehetőségét kizárhatjuk. Mindezek tudatában, érdemes megvizsgálni, hogy a Rózsa Sándor- és az Erdély-regények, ezen belül pedig a *Rózsa Sándor összevonja szemöldökét* (1942) miért milyen képez a legtöbb regénnyel stilometriai rokonságot.

Móricz nyelvjárás-használatának fejlődésvonalát nézve a kutatások arra mutatnak rá, hogy a kezdeti nagyarányú tájnyelviség a '20-as esztendőktől letisztultabb, egyszerűbb lesz: „A húszas évektől kezdve már a nagyvonalú ecsetkezelés jellemzi: nem a hangzó külsőségek visszaadására, hanem csak „színek" megvillantására törekszik a dialógusokban, majd a gondolati szövegekben."[13] Ez a stíluseszköz jellemzi a Rózsa Sándor-regényeket is: az elbeszélésmód meghatározó eleme a szegedi nyelvjárás megjelenítése a párbeszédekben és a belső monológszerű leírásokban. A nyelvi vonatkozású elemzések gyakran utalnak a nyelvjárás következetlen használatára. Péter László olyan jelenségekre hívja fel a figyelmet, mint az ö-zés helytelen használata a regényekben, a szegedi nyelvjárásokon kívül másvidéki tájszavak megjelenése és a fonetikus írásmód alkalmazása, valamint egyazon nyelvi jelenségek többféle jelölése.[14] A jelenségek együttese olyan heterogenitást kölcsönözhetett a regény stílusának, amelyhez stilometriailag sok korábbi szöveg tud kapcsolódni. Úgy tűnik, ezek a regények magukba olvasztják a móriczi írásmód különböző karakterjegyeit, amelyek elemeikben az egyes szövegekben is felfedezhetőek.

A hálózatos megjelenítés segít ábrázolni a szövegek közötti stilometriai kapcsolatot, de az összefüggésrendszerek megfejtése további filológiai kutatómunkát és szoros olvasást igényel. Vizsgálatunk rávilágít arra, hogy a gépi elemző módszerek új irányokat nyithatnak meg az életmű értelmezéséhez, s ez minden irodalomtörténész számára nyitott feladat lehet…

[13] Sebestyén Árpád: A tájnyelv, mint stíluseszköz Móricz Zsigmond műveiben In: Magyar nyelvjárások. Szerk. Kálmán Béla – Papp István – Sebestyén Árpád, Debrecen, 1972, Kossuth Lajos Tudományegyetem, 9–41., 19.

[14] Péter László: A „Rózsa Sándor" nyelve: a nyelvjárás írói alkalmazásának kérdéséhez: Móricz Zsigmond halála 10. évfordulójára. Tiszatáj, 1952. 3. sz. 152–181.