Kiss Margit

Bölcsészettudományi Kutatóközpont, Irodalomtudományi Intézet

kiss.margit@btk.mta.hu

Stilometriai elemzés lehetőségei magyar történeti szövegkorpuszon

Tanulmányomban magyar nyelvű történeti szövegek számítógépes elemzésének egy olyan lehetőségével foglalkozom, amely ötvözi a nyelv- és irodalomtudomány, az informatika és a statisztika eredményeit. A szerzőségi, illetve a stilometriai vizsgálat bár nem új keletű az irodalmi szövegelemzések esetében, módszertanát tekintve folyamatosan formálódik, megújul. Munkámban áttekintem e szövegelemzési módszer jellemzőit és alkalmazási lehetőségeit, majd esettanulmányként különböző elemzéseket mutatok be Mikes Kelemen művei alapján. Stilometriai módszerekkel vizsgálom az életműben a saját szerzőségű művek és a fordítások kapcsolatát, valamint a művek tematikai elkülönülését. Végezetül bemutatom, hogy a digitális írói szótár alkalmazása – mint történeti szöveget normalizáló eszköz – hogyan javíthatja ezeknek az elemzőmódszereknek a hatékonyságát.

Kulcsszavak:
szerzőségi vizsgálat, stilometria, digitális írói szótár, Mikes Kelemen

> A nyelv csodálatossága abban rejlik, hogy bár közös forrásból táplálkozik, mi mind valami egyedit hozunk létre belőle. A számítógépes elemzés lehetővé teszi, hogy sokkal pontosabban nyomozzunk a lexikai elemek után, mintha egyszerűen csak a puszta intuíciónkra hagyatkoznánk.[1]

1. Bevezetés

A nagyméretű szövegkorpuszokat számítógép támogatásával elemző kutatók manapság egyre több módszer közül választhatnak, s olyan kérdésekre adhatnak választ, amelyekre korábban manuális módszerek felhasználásával még nem vagy csak jelentős időráfordítással volt lehetőség. Az egyre nagyobb méreteket öltő digitalizálás

[1] Hugh Craig, a University of Newcastle professzor emeritusának egyetemi weboldalán megjelent összegzés. Hugh Craig, „Figures of Speech," hozzáférés: 2019.02.20, `https://www.newcastle.edu.au/profile/hugh-craig`. (Ford. tőlem.)

mellett arra is érdemes hangsúlyt fektetni, hogy akik korpusz alapú szövegvizsgálatokat végeznek, megismerjék, alkalmazzák, majd továbbfejlesszék az elemzőeljárásokat.[2] A tanulmány célja kettős: a modern nemzetközi kutatások tükrében a nyelv-, az irodalomtudomány, az informatika, valamint a statisztika eredményeit ötvöző, a magyar nyelvű szövegek vizsgálatában kevéssé elterjedt szerzőségi, illetve stilometriai elemzőmódszert mutatom be. Másfelől arra a kérdésre keresem a választ, hogy vajon a magyar történeti szövegek elemzésében hogyan alkalmazhatjuk ezeket a statisztikai alapú módszereket, és hogy miként növelhetjük az elemzések hatékonyságát. Ehhez konkrét esettanulmányokat mutatok be Mikes Kelemen életművének különböző szempontok alapján történő stilometriai elemzésével.

2. Előzmények

2.1. Szerzőségi vizsgálatok

A vitatott vagy bizonytalan szerzőség megállapításával kapcsolatos vizsgálatok éppoly régre nyúlnak vissza, mint amióta az írás létezik. A szerzőségi vizsgálatokat és alakulásukat kutató Hugh Craig megjegyzi, hogy a *Biblia,* a homéroszi művek vagy Shakespeare munkái még egy olyan időszakban születtek, amikor a szerzői homogeneitás nem volt különösebben fontos szempont.[3] Későbbi generációk mégis jelentőséget tulajdonítottak annak, hogy szerzőségi szempontból is megvizsgálják ezeket a szövegeket. A bekövetkezett szemléletváltás a reneszánszra tehető a szövegek komparatív vizsgálatának lehetőségével, a nyelvi és textológiai diszciplínák alkalmazásával. Egyik leghíresebb példa erre Lorenzo Valla 15. századbeli humanista munkája, aki filológiai módszerekkel bizonyította be, hogy a *Donatio Constantini* adománylevél hamisítvány.[4] Azóta számos kétes vagy bizonytalan szerzőségű művet tulajdonítottak valamely szerzőnek, vagy zártak ki egy adott szerzőség alól, de nem kevés szöveg maradt anonim vagy legalábbis vitatott szerzőségű.

A szerzőségi vizsgálatok hagyományos megközelítésben a filológia, nyelv- és irodalomtörténet, paleográfia, kodikológia, történettudomány és az igazságügy egyes területeit érintik, de idővel a diszciplína nem hagyományos eljárásokkal is kiegészült, úgy mint a statisztikai elemzőmódszerek.[5] A szerzőségi vizsgálatok esetében a statisztika alkalmazása ugyan nem nevezhető tradicionálisnak, ugyanakkor a statisztikára támaszkodó szövegelemzés nem új fejlemény.[6] 1851-ben Augustus de Morgan egy

[2] Mészáros Tamás, „Mit nyújthat a modern informatika az irodalomtudomány számára?" *Magyar Tudomány,* 11. sz. (2016): 1310–1315, http://www.matud.iif.hu/2016/11/06.htm.

[3] Hugh Craig, „Stylistic Analysis and Authorship Studies" in *A Companion to Digital Humanities,* eds. Susan Schreibman, Ray Siemens and John Unsworth (Oxford: Blackwell Publishing, 2007), 282, https://doi.org/10.1002/9780470999875.

[4] Harold Love, *Attributing Authorship* (Cambridge: Cambridge University Press, 2002), 18–19, https://doi.org/10.1017/cbo9780511483165.003; Christopher B.Coleman, ed, trans., *The Treatise of Lorenzo Valla on the Donation of Constantine. Text and Translation into English* (New Haven: Yale University Press, 1922), 131–133.

[5] Maciej Eder, „Style-Markers in Authorship Attribution: A Cross-Language Study of the Authorial Fingerprint," *Studies of Polish Linguistics* 1 (2011): 100.

[6] A jelen tanulmány szempontjából releváns mérföldkövekre vonatkozóan felhasznált és áttekintő összegzést adó irodalom: David Holmes and Judit Kardos, „Who Was the Author? An Introduction to

 DOI 10.31400/dh-hun.2019.2.336

barátjának írott levelében arról a megfigyeléséről számol be, hogy a szavak hosszúságának meghatározó szerepe van a szerzőség megállapításával kapcsolatban.[7] Thomas Mendenhall amerikai fizikus az 1880-as években az írói stílus kvantitatív elemzésével foglalkozott, elsősorban angol szerzők munkái alapján.[8] Évtizedekkel később George Udny Yule és George Zipf meghatározó eredményeket ért el az elemzésben alkalmazható szövegjegyek felkutatásában.[9] Az 1960-as években Frederick Mosteller és David Wallace analízise már megnyitotta az utat a modern, digitális kor stilometriája felé: munkáik úttörő jelentőségűvé váltak az irodalmi szövegek szerzőségi vizsgálatainak tekintetében.[10] A *Federalist Papers* 1787 és 1788 között 85 politikai esszét publikált, amelyben a szavazókat az Egyesült Államok számára készített alkotmány jóváhagyásáról igyekeztek meggyőzni. Az esszéket mind „Publius" névvel jegyezték, de azt azért lehetett tudni, hogy Alexander Hamilton, James Madison, illetve John Jay írhatta őket. Több nyelvi megkülönböztető jegyet, valamint valószínűségi modelleket is felhasználtak a különösen nehéznek mutatkozó szerzőségi probléma miatt, amelyet a stílus- és a politikai tartalombeli hasonlóság is nehezített. Frederick Mosteller és David Wallace mind a tizenkét vitatott írást Madisonnak tulajdonította, így a kapott eredmény lényegében összhangban állt a történészkutatók eredményeivel. Az 1980–1990-as években John Burrows jelentős eredményeket ért el új elemzési eljárások kialakításával, amelynek során a megkülönböztető jegyek közül a funkciószavak elemzésére támaszkodott. Burrows több szerzőt és eltérő műfajú munkákat elemzett, például Austent, a Brontë testvéreket, Scottot és Byront.[11] Lexikális szintű elemzések végzése során a Burrows-módszer ma is elterjedt.[12] A számítógép térhódítása a bölcsészettudományokban új szakaszt nyitott a szerzőségi vizsgálatok területén is a szövegek stilisztikai jegyeinek a mérésével, valamint az eredmények összevethetősége és értékelése terén, ugyanakkor

Stylometry," *Chance* 16, 2. sz. (2003): 5–8, http://doi.org/10.1080/09332480.2003.10554842; David Holmes, „The Evolution of Stylometry in Humanities Scholarship," *Literary and Linguistic Computing* 13, 3. sz. (1998): 111–117, https://doi.org/10.1093/llc/13.3.111; Harold Love, *Attributing Authorship* (Cambridge: Cambridge University Press, 2002), https://doi.org/10.1017/cbo9780511483165.003.

7 R. D. Lord, „Studies in the History of Probability and Statistics. VIII. de Morgan and the Statistical Study of Literary Style," *Biometrika* 45, 1–2. sz. (1958): 282, https://doi.org/10.1093/biomet/45.1–2.282.

8 Thomas Corwin Mendenhall, „The Characteristic Curves of Composition," *Science* 9, 214. sz. (1887): 237–249, https://doi.org/10.1126/science.ns–9.214s.237.

9 Udny Yule, *The Statistical Study of Literary Vocabulary* (Cambridge, Cambridge University Press, 1944); George Kingsley Zipf, *Selected Studies of the Principle of Relative Frequency in Language* (Cambridge: Harvard University Press, 1932), https://doi.org/10.4159/harvard.9780674434929.

10 Frederick Mosteller and David Wallace, *Inference and Disputed Authorship: The Federalist.* Reprinted With a New Introduction by John Nerbonne (Stanford: CSLI Publications, 2007 [1964]).

11 John Burrows, *Computation into Criticism: A Study of Jane Austen's Novels and an Experiment in Method* (Oxford: Clarendon Press, 1987).

12 John Burrows, „'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship," *Literary and Linguistic Computing* 17, 3. sz. (2002): 267–287, https://doi.org/10.1093/llc/17.3.267; David Hoover, „Testing Burrows's Delta," *Literary and Linguistic Computing* 19, 4. sz. (2004): 453–475, https://doi.org/10.1093/llc/19.4.453.

azt is el kell ismerni, hogy a stilometria nemritkán heves szakmai viták kereszttüzébe kerül.[13]

2.2. Stilometria

A stilometria szó és a diszciplína megalkotója Wincenty Lutosławsky.[14] Ő az új módszert Platón dialógusainak a kronologizálásához alkalmazta, amellyel a filozófus eszmerendszerének az értelmezéséhez nyújtott újfajta segédletet. Ma a stilometria megnevezés a stílus statisztikai alapú vizsgálatát jelenti, a szerzőség statisztikai szempontú, nyelvészeti és statisztikai feltevéseken alapuló megközelítését *nem hagyományos szerzőségi vizsgálat*nak (*non-traditional authorship attribution,* ford. tőlem) nevezik.[15] Mindkét esetben az a kérdés áll a középpontban, hogy melyek azok a nyelvi tényezők, amelyek meghatározók a szerzői művekkel kapcsolatban. Az irodalmi nyelv és stílus statisztikai alapú elemzésének nem az a célja, hogy felforgassa a hagyományos elemzések során alkalmazott eszközkészletet, hanem hogy kiegészítse, komplexebbé tegye a hagyományos módszerekkel végzett vizsgálatokat a kétséges jelenségeket illetően. Minden szerzőnek van egy olyan sajátos stílusa, ami állandó, és olyan jegyeket tartalmaz, amely mennyiségileg is meghatározható, ezáltal megkülönböztető funkcióval rendelkezik, így bizonyos nyelvi jellemzők (szókészlettani gazdagság, kollokációk, sajátos szintaktikai jellemzők, szókörnyezet) statisztikai eszközökkel mérhetők. Az a cél, hogy fel lehessen tárni ezeket a szerzői megkülönböztető jegyeket, különösen azokat, amelyek a szoros olvasás során észrevétlenek maradnak. Kísérletek azt bizonyítják, hogy ezek az emberi olvasás során rejtve maradó rögzült minták a stílusparódiák vagy az álnéven írt munkák szerzőit is leleplezhetik azzal, hogy a saját nyelvezetük ujjlenyomatait hordozzák magukon.[16] A szövegelemzés során a számítógépes stilisztika tendenciákkal dolgozik. A tendenciák jobban megfigyelhetők az olyan összetett jelenségek mögött, amelyekhez az emberi feldolgozó- és felfogóképesség már nem elegendő. Azokon a területeken nyújt segítséget, amelyeken több szempontú, átfogó összehasonlítások szükségeltetnek: a nyelvi modellek vizsgálatakor a szövegalkotás, kifejezésmód egyéni jellemzőinek a feltárásában úgy, hogy az egyént meghatározó jegyek kiszűrésére törekszik.

David Holmes és Judit Kardos rámutat arra, hogy a modern stilometria a kezdetek óta sokat változott a számítógép nyújtotta lehetőségek és a mesterséges intelligencia hatására, amely a meghatározó stílusjegyek felismerésében is szerepet játszhat.[17] Rámutatnak továbbá, hogy a gépi tanulás sikeresen alkalmazható e területen. A *neurális*

[13] M. W. A. Smith, „Shakespeare, Stylometry and »Sir Thomas More«," *Studies in Philology* 89, 4. sz. (1992): 434–444.

[14] 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: Longmans, 1897), `https://archive.org/details/originandgrowth00lutogoog/page/n44`.

[15] Eder, „Style-Markers in Authorship," 100–101.

[16] Hugh Craig, „Stylistic Analysis," 285; John Burrows, „I Lisp'd in Numbers: Fielding, Richardson and the Appraisal of Statistical Evidence," *The Scriblerian,* 33 (1991): 234–241. J. K. Rowling Cuckoo's Calling című regényének szerzőazonosítása Patrick Juola, „The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions," *Digital Scholarship in the Humanities* 30, 1. sz. (2015): 100–113, `https://doi.org/10.1093/llc/fqv040`.

[17] Holmes and Kardos, „Who Was the Author," 5–8.

 DOI 10.31400/dh-hun.2019.2.336

hálók segítségével tökéletesíthetjük az elemzést, amely azáltal javítja a módszer működését, hogy maga próbál olyan tulajdonságot felfedezni, amely az általunk megadottat tökéletesíti. A tanítófolyamat számos előnye mellett a hátránya az, hogy nagy mennyiségű adat szükséges hozzá. A *genetikus algoritmus* a tanítókorpuszon kalibrálódik evolúciós jelleggel, és a stilometriai vizsgálatokban a meglévő szabályok közül a legadekvátabb megkülönböztető funkció megtalálását segíti. Úgy vélik, vitatott szerzőség esetében nagyon jó eredménnyel alkalmazható, ha elegendő adat áll rendelkezésre a mintatanuláshoz. 1993–1994-ben Robert Matthews és Tom Merriam alkalmazta a módszert sikerrel: Shakespeare és Fletcher műveiből tanítókorpuszt hoztak létre, majd a *The Two Noble Kinsmen* című műben vizsgálták a két szerző kollaborációját.[18]

2.2.1. Mire alkalmazható a stilometria?

Az, hogy szövegeket a lexikális jegyeik alapján mérünk és hasonlítunk össze, lehetővé teszi a vizsgált szövegek közti azonosságok és különbségek értékelését. A vizsgálati szempontok vonatkozhatnak anonim vagy vitatott szerzőségű szövegek azonosításának a támogatására,[19] de akár egy szerzői munkásságon belül a nyelvezet, a szövegformálás változásának a feltárására is, amely segítséget nyújthat az életmű korszakolásában.[20] Elemezhetünk csoporthoz való tartozást: férfi és női szerzők munkái közti különbséget,[21] műfaji jelleget,[22] nyelvi szempontból is megmutatkozó hatást, előzményt, inspirációt[23] stb.

Ezek az elemzési metódusok jellemzően nem önmagukban állnak, sőt magukban alkalmazva félre is vezethetik az elemzőt. Például e vizsgálatok eredményeképpen ma úgy vélik, hogy Shakespeare-nek nem volt átlagon felüli gazdagságú szókészlete, az ő kivételessége sokkal inkább abban rejlik, hogy egyedülálló módon használta az átlagos, hétköznapi szavakat. A stilometriai módszerek bevonásának köszönhetően ma már valószínűsíthető, hogy a *VI. Henrik* című dráma egyes részeiben Marlowe is közreműködött.

[18] Uo., 5–8. Robert Matthews and Tom Merriam, "Neural Computation in Stylometry I: An Application to the Works of Shakespeare and Fletcher," *Literary and Linguistic Computing* 8, 4. sz. (1993): 203–209, https://doi.org/10.1093/llc/8.4.203; Tom Merriam and Robert Matthews, "Neural Computation in Stylometry II: An Application to the Works of Shakespeare and Marlowe," *Literary and Linguistic Computing* 9, 1. sz. (1994): 1–6, https://doi.org/10.1093/llc/9.1.1.

[19] Ward E. Y. Elliott and Robert J. Valenza, „Two Tough Nuts to Crack: Did Shakespeare Write the 'Shakespeare' Portions of Sir Thomas More and Edward III? Part I," *Literary and Linguistic Computing* 25, 1. sz. (2010): 67–83, https://doi.org/10.1093/llc/fqp029; Ward E. Y. Elliott and Robert J. Valenza, „Two Tough Nuts to Crack: Did Shakespeare Write the 'Shakespeare' Portions of Sir Thomas More and Edward III? Part II: Conclusion," *Literary and Linguistic Computing* 25, 2. sz. (2010): 165–177, https://doi.org/10.1093/llc/fqp029.

[20] Dirk Van Hulle and Mike Kestemont, „Periodizing Samuel Beckett's Works: A Stylochronometric Approach," Style 6, 2. sz. (2016), 172–202, https://doi.org/10.5325/style.50.2.0172.

[21] Sean G. Weidman and James O'Sullivan, „The Limits of Distinctive Words: Re-evaluating Literature's Gender Marker Debate," *Digital Scholarship in the Humanities* 33, 2. sz. (2018): 374–390, https://doi.org/10.1093/llc/fqx017.

[22] Alexandre Sotov, „Lexical Diversity in a Literary Genre: A Corpus Study of the *Ṛgveda*," *Literary and Linguistic Computing* 24, 4. sz. (2009), 435–447, https://doi.org/10.1093/llc/fqn044.

[23] Regula Hohl Trillini and Sixta Quassdorf, „A 'Key to all Quotations'? A Corpus-Based Parameter Model of Intertextuality," *Literary and Linguistic Computing* 25, 3. sz. (2010), 269–286, https://doi.org/10.1093/llc/fqq003.

ködött.[24] Bár alapvetően az angol nyelvű munkákra és a klasszikus művek vizsgálatára koncentrálódnak a számítógépes szerzőségi elemzővizsgálatok,[25] az utóbbi időszakban más nyelvekre és szövegtípusokra is alkalmazzák őket.[26]

3. Szerzői életmű stilometriai vizsgálata

Ha a szövegek közti eltéréseket kutatjuk, akkor nemcsak az egyes szerzők közti különbözőségeket vizsgálhatjuk, hanem a szerzői életművön belüli váltásokat is nyomon követhetjük. Ez esetben fontos látnunk, hogy a szerzők egymás közti kifejezésmódbeli különbözősége és a szerzői életmű alakulása – bár e két típus közel sem egyforma mértékben – szövegstatisztikai szempontból eltéréseket rejt. Mérhető különbségek nemcsak szerzők között lehetnek, hanem egy életpálya különböző szakaszai között is vizsgálható a nyelvezet változása, amelynek módszeres vizsgálata különféle értelmezői keretek kialakításában is segítséget nyújthat, így például a szerzői életművek szakaszolásában.[27] Az írói-költői nyelvezet alakulásának a vizsgálata során a *Does "Late Style" Exist? New Stylometric Approaches to Variation in Single-Author Corpora* című tanulmány[28] szerzője arra az eredményre jutott, hogy nem az egyes szerzők kései korszakának a beszédmódjai térnek el jelentősen a megelőzőekétől, hanem éppen a korai írói életpálya különül el markánsan a későbbi alkotói szakaszoktól. Az önálló szerzői korpusz vizsgálata az alkotásmód alakulása tekintetében egy lehetséges út a szépirodalmi szövegek stilometriai elemzésében. Különösen azokban az esetekben hatékony eszköz, amelyekben jelentős mennyiségű szöveg áll rendelkezésre az életmű adekvátabb megértése érdekében.

Magyar nyelvű szépirodalmi szövegek vizsgálatában nem általánosan elterjedt gyakorlat a stilometriai módszertan. A magyar történeti szövegek számítógépes elemzése különösen nehézségekkel terhelt feladat a nyelv standardizálatlansága és a gépi elemzés szabályelvűsége miatt. Arra voltunk kíváncsiak, hogy a stilometriai módszerek vajon ezzel együtt is támogatást nyújtanak-e a szövegvizsgálatokban, s a magyar nyelvű történeti szövegek vizsgálatához alkalmazható-e megbízható eredménnyel ez a

[24] Hugh Craig, „Ignore the Doubters: Here's Why Christopher Marlowe Co-wrote Shakespeare's Henry VI," *The Conversation,* 2016. nov. 9, `https://theconversation.com/ignore-the-doubters-heres-why-christopher-marlowe-co-wrote-shakespeares-henry-vi-68229`; Hugh Craig and Arthur F. Kinney, eds., *Shakespeare, Computers and the Mystery of Authorship* (Cambridge: Cambridge University Press, 2009), `https://doi.org/10.1017/cbo9780511605437`.

[25] Vö. 6. jegyzet, különösen Harold Love, *Attributing Authorship* (Cambridge: Cambridge University Press, 2002), `https://doi.org/10.1017/cbo9780511483165.003`.

[26] Érdekes kísérlet a lengyel nyelv történeti korszakolásának vizsgálatára: Maciej Eder and Rafal L. Górski, „Historical Linguistics' New Toys, or Stylometry Applied to the Study of Language Change" in *Digital Humanities 2016: Conference Abstracts,* eds. Maciej Eder and Jan Rybicki (Krakow: Jagellonian University & Pedagogical University, 2016), 182–184.

[27] Van Hulle and Kestemont, „Periodizing Samuel Beckett's Works," 172–202.

[28] Jonathan Pearce Reeve, „Does »Late Style« Exist? New Stylometric Approaches to Variation in Single-Author Corpora" in *Digital Humanities 2018, DH 2018, Book of Abstracts,* eds. Jonathan Girón Palau and Isabel Galina Russell (Mexico City: El Colegio de México, UNAM, and RedHD, June 26–29, 2018), 478–480.

 DOI 10.31400/dh-hun.2019.2.336

módszer. Kísérletképpen ismert szerzőségű szövegeken, a Mikes-korpuszon[29] végeztünk elemzéseket.[30] Előtte azonban érdemes áttekinteni, hogy milyen elemekből áll egy stilometriai elemzés.

3.1. Mit mérjünk?

Lutoslawsky 1897-ben azt írta, hogyha a kézírás meghatározza az írója személyét, akkor az egyéni stílus még ennél is személyesebb és jellemzőbb.[31] Maciej Eder szerint a mai stilometriai módszereket alkalmazók messze állnak ettől a határozottságtól, mégis úgy vélik, hogy az írás folyamatára hatással van a tudattalan.[32] A legfontosabb feladat kinyomozni az erről árulkodó jegyet, a „szerzői ujjlenyomatot" a különféle nyelvi (lexikális, morfológiai, szintaktikai) jellemzők közül. Arra a kérdésre kell választ adni, hogy melyik az a nyelvi jelenség, amely mérhető az egyes szerzői szövegekben a szerzői ujjlenyomat meghatározása érdekében. Úgy véli, hogy a sikeres vizsgálathoz minél több egyedi stíluselem, ún. stílusmarker[33] meghatározása a cél. Hogy a stilisztikai egyéni jellemzők kialakítása során azonban mi a közös és mi az egyedi a nyelvben, nem teljesen magától értetődő. Véleménye szerint a legjobb stílusmarkerek azok, amelyek szabad szemmel felfedezhetetlenek, így a szerzői kontrollon túlmutatnak, és az utánzás sem fog rajtuk. Bár idővel egy szerző stílusa, kifejezésmódja változhat, nem különböznek olyan meghatározó mértékben egymástól a saját szövegeik, mintha más szerzőkhöz hasonlítanánk őket. Hugh Craig például rámutatott, hogy a korai Henry James és a kései Henry James is különböző, de nem annyira eltérő, mint Henry James és Thomas Hardy.[34] Burrows pedig egy vizsgálatában rávilágított arra, hogy Henry Fielding Samuel Richardson ellenében álnéven írt stílusparódiája sokkal közelebb maradt a saját stílusára jellemző nyelvi elemekhez, mint a kifigurázandó szerzőéhez.[35]

A stílusmarkerek változatossága legalább olyan gazdag, mint a története. David Holmes *Authorship Attribution* című írása[36] kiválóan összefoglalja a különféle stílusmarkerek alkalmazásával elért eredményeket, és egyben feltárja gyenge pontjaikat is. A tudomány jelen állása szerint a stilometriában a nyelvi változásnak nem a kevésbé, hanem éppen a jobban ellenálló szókészleti elemek vizsgálata ígérkezik eredményesebbnek, mert ezek mutatnak rá az egyéni kifejezésmód rögzült formáira. A lexikai szint mára meghatározóvá vált, ennek mérhetővé tételére számos különféle statisztikai

[29] Mikes Kelemen, *Összes művei,* s. a. r. Hopp Lajos (Budapest: Akadémiai Kiadó, 1966–1988). Elektronikus verzió: Magyar Elektronikus Könyvtár, 2011, `http://mek.oszk.hu/09000/09000/`. A kísérlethez felhasznált szövegkorpusz a cikk online mellékletében megtalálható (vö. 53. jegyzet): `https://doi.org/10.31400/dh–hun.2019.2.336`.

[30] A stilometriai kísérletek futtatásában Dobi Jan Sándor hallgató (BME VIK) és Mészáros Tamás egyetemi oktató (BME VIK) volt segítségemre.

[31] Eder, „Style-Markers in Authorship," 103 alapján Lutosławski, "The Origin and Growth of Plato's," 66.

[32] Eder, „Style-Markers in Authorship," 103.

[33] A kifejezés az angol terminológia alapján történő tükörfordítás tőlem. A szót 'stílust jelölő' értelemben használom.

[34] Craig, „Stylistic Analysis," 285.

[35] Burrows, „I Lisp'd in Numbers," 234–241.

[36] David Holmes, „Authorship Attribution," *Computers and the Humanities* 28, 2. sz. (1994): 87–106, `https://doi.org/10.1007/bf01830689`.

módszer született.[37] Eder kutatása[38] rámutat, hogy a korszerű vizsgálatokban a legszélesebb körben elterjedt a minimum 100 leggyakoribb szó elemzése (MFW, min. 100), ezt követi a mondathossz, a szóhosszúság, a hangsúlyos és hangsúlytalan szótagok váltakozása, a szókészlet gazdagsága, a leggyakoribb funkciószavak, a központozás, a kollokációk, bizonyos betűsorozatok gyakorisága és a szóbigramok vizsgálata.[39]

Jelen kutatások arra is felhívják a figyelmet, hogy a stílusmarkerek nem tekinthetők teljesen nyelvfüggetlennek.[40] Különbözőségük a nyelvtípusok közti különbségből is adódik. Egyre több vizsgálat irányul e nyelvspecifikus jegyek feltárására.[41] Hogy végül ténylegesen egy adott szöveg elemzéséhez melyik marker válik megkülönböztetővé, az erősen függ magától a korpusztól. Grieve a *Quantitative Authorship Attribution: An Evaluation of Techniques* című tanulmányában[42] harminckilenc szerzőségi módszert hasonlít össze, hogy választ kapjon arra a kérdésre, melyik lehet a leghasznosabb a szerzőség megállapításához. Ismert szerzőségű szövegeken hasonlította össze a különféle lehetőségeket, és arra az eredményre jutott, hogy az algoritmusok kombinációja meggyőző eredményt nyújt a megfelelő stílusmarker megtaláláshoz, de még a valószínűség megfogalmazásához is több módszer együttes alkalmazását tartja indokoltnak.

3.2. Hogyan mérjünk?

A nyelv vizsgálatára alkalmazott statisztikai technikáknak egyik csoportja a szövegkorpuszból körültekintően kiválasztott egyetlen jelenségre fókuszál, mint például a szókészlet gazdagsága, különféle indexek stb. A másik csoport nagy mennyiségű jellemzőt vizsgál, ezek a multidimenzionális statisztikai módszerek, amelyek finomabb különbségek feltárására is alkalmasak.[43] Ezek lényege, hogy a tulajdonságok sokdimenziós terében helyezik el a vizsgált szövegeket. Ilyen például a *klaszteranalízis,* amely egy olyan csoportosító eljárás, amellyel elemeket homogén csoportokba ren-

[37] Holmes, „Authorship Attribution," 87–98.

[38] Eder, „Style-Markers in Authorship," 103.

[39] David Holmes, „The Evolution of Stylometry," 111–117; David L. Hoover, „Frequent Word Sequences and Statistical Stylistics," *Literary and Linguistic Computing* 17, 2. sz. (2002): 157–180, https://doi.org/10.1093/llc/17.2.157; Juan-Pablo Posadas-Duran, Grigori Sidorov and Ildar Batyrshin, „Complete Syntactic N-grams as Style Markers for Authorship Attribution," in *Human-Inspired Computing and Its Applications,* MICAI 2014, Lecture Notes in Computer Science, vol. 8856, eds. A. Gelbukh, F. C. Espinoza and S. N. Galicia-Haro (New York: Springer, 2015), 9–17, https://doi.org/10.1007/978−3−319−13647−9_2.

[40] Maciej Eder, Jan Rybicki and Mike Kestemont, „Stylometry with R: A Package for Computational Text Analysis," *The R Journal* 8, 1. sz. (2016): 107–121, https://doi.org/10.32614/rj−2016−−007; Maciej Eder, „Style-Markers in Authorship," 103.

[41] E témában részint a magyar nyelvre vonatkozóan is találunk megállapításokat: Jan Rybicki and Maciej Eder, „Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words?" *Literary and Linguistic Computing* 26, 3. sz. (2011): 315–321, https://doi.org/10.1093/llc/fqr031.

[42] Jack Grieve. „Quantitative Authorship Attribution: An Evaluation of Techniques," *Literary and Linguistic Computing* 22, 3. sz. (2007): 251–270, https://doi.10.1093/llc/fqm020.

[43] A leggyakrabban alkalmazott elemzőmódszereket és távolságmértékeket Eder ide tartozó kutatási eredményei alapján mutatom be: Maciej Eder, „Style-Markers in Authorship," 99–114; Maciej Eder, Jan Rybicki and Mike Kestemont, „Stylometry with R: a package for computational text analysis," *R Journal* 8, 1. sz. (2016): 107–121, https://doi.org/10.32614/rj−2016−007.

 DOI 10.31400/dh-hun.2019.2.336

dezünk. Az egyes klasztereken belüli adatok valamely jellemzők mentén hasonlítanak és különböznek a többi klaszter elemeitől. Hasonló multidimenzionális eljárások a *főkomponens-analízis,* a *faktoranalízis,* a *többdimenziós skálázás,* a *diszkriminanciaanalízis,* a *Support Vector Machine (SVM),* a *Nearest Shrunken Centroids (NSC),* és Burrows attribúciós tesztjei: a *Delta, Zeta* és *Iota.*

A stilometriai elemzés során statisztikai módszerrel különféle stílusmarkerek előfordulási gyakoriságát vizsgáljuk a szövegekben, azaz a szövegekhez a stílusmarkerek terében egy-egy vektort rendelünk. Az így kapott, adott szövegre jellemző értékeket távolságmértékkel elemezzük, hogy meghatározzuk a szövegek egymáshoz való viszonyát. Két szöveg hasonlósága a tulajdonságok terében az őket reprezentáló vektorok között lévő távolsággal határozható meg.

A multidimenzionális eljárás során a szövegkorpusz gyakorisági tényezői közti távolság mérésére alkalmazott *távolságmérték* kiválasztását nagyban meghatározza az, mit akarunk elemezni. Az *euklideszi távolság* csak azokban az esetekben megfelelő, ha a markerek eloszlása a szövegekben hasonló, amely sokféle markernél nem áll fenn, pl. a szavak gyakorisága jellemzően nem ilyen. Alkalmas lehet azonban a ritka, a témát megjelölő szavakra, hiszen azok jellemzően egyenlő mértékben szerepelnek a korpuszokban. A *Manhattan távolság* már a normalizált távolságot méri. A *Classic Delta* normalizált szógyakoriságot mér, de függ az elemzett szövegek arányától és a szerzők szövegarányától. Argamon *Lineáris Deltá*ja Burrows *Deltá*jának és az *euklideszi távolság*nak a keveréke: a normalizált jelleggyakoriságokra alkalmazott *euklideszi távolság,* amely érzékeny a szövegek számára. Eder *Deltá*ja a flektáló nyelvekre jól alkalmazható, a *Classic Delta* módosítása. A *Canberra távolság* nagyon szenzitív a szerzők közti ritka szóhasználatra, és érzékeny az elemzett szavak számára.

A sikeres stilometriai elemzés közel sem triviális feladat. Nemcsak az összetett elméleti háttér alapos ismeretére van szükség, de az empirikus módon szerzett tudás is meghatározó szerepű. A megfelelő reprezentatív korpusz összeállításának a fontossága alapvető, hiszen fel kell ismernünk, hogy miből adódik a szövegek közti különbség. Ha például a vizsgált szövegek tematikailag nagyon eltérőek, akkor könnyen meglehet, hogy a témák közti különbség domborodik ki, és mégsem a szerzők közti eltérést elemeztük, ahogy terveztük. Ugyanez igaz a különféle műfajok és az időszakok közti különbözőségekre is. A vizsgálandó szövegek méretének a figyelembe vétele szintén lényeges szempont, hiszen bizonyos *távolságmérték*ek erre nagyon érzékenyek, és emiatt torzíthatnak. Az eredményességhez hozzájárul az adott vizsgálathoz legadekvátabb módszerek megtalálása és ezek kombinációja, együttes alkalmazása. Ami az egyik esetben sikeres attribúciós eljárás, az nem feltétlenül működik a másikban. Mára már megdőlni látszik az a feltételezés, hogy a stilometriai elemzés során azt a technikát kellene megtalálni, amely sikeres lehet minden műfajra, nyelvre és korszakra.[44] Helyette inkább az adott feladathoz és elemzéshez érdemes a legadekvátabb módszert kialakítani. Az ily módon elvégzett vizsgálat esetén is inkább valószínűségről, mint teljes bizonyosságról beszélhetünk, és nem nélkülözhető a kritikai szellemű nyelvi-filológiai kontroll sem.

[44] Holmes and Kardos „Who Was the Author," 5.

3.3. Mivel mérjünk?

A szerzőségi és stilometriai elemzésekhez ma már különféle informatikai eszközcsomagok állnak rendelkezésre. A bölcsészkutatók számára egyszerűen alkalmazható a magyar nyelv statisztikai alapú szövegelemzésére is alkalmas, nyílt hozzáférésű, Maciej Eder, Jan Rybicki és Mike Kestemont által R-ben kialakított *Stylo* programcsomag,[45] amelynek hazai fejlesztésben már webes alkalmazása, a *Shtylo* is elérhető.[46] Ez utóbbi előnye, hogy a futtatókörnyezet kialakításának a terhét leveszi a kutató válláról. A webes alkalmazáshoz egy böngészőre van szükség, a sok memóriát és processzoridőt igénybe vevő feladatok egy központi szervergépen futnak. További előnye, hogy a korpuszokat adatbázisban tárolja el. Ugyan az alkalmazás a munkafolyamatok különböző lépéseit elvégzi helyettünk, de a konfiguráció és a paraméterezés ezzel együtt is komoly hozzáértést és tapasztalatot igénylő feladat, amely érinti a bemenettel és a nyelvvel, a választott vizsgálandó és a leggyakoribb vizsgálandó elemekkel, a selejtezéssel, a választott statisztikai elemzőmódszerrel, a mintavételezéssel és a kimenet formátumával kapcsolatos beállításokat. A továbbiakban bemutatott kísérleteket ezzel az alkalmazással végeztük el.

4. Kísérletek a Mikes-korpuszon

A vizsgálat alapkérdése az, hogy vajon magyar történeti szövegen eredménnyel tudjuk-e alkalmazni a statisztikai elemzésnek ezt a típusát. Ehhez azt vizsgáltuk, hogy a Mikes-művek[47] különböző jellegű csoportosítása a nyelvi megformáltság alapján stilometriai eszközökkel megvalósítható-e, illetve az életművel kapcsolatban meglévő ismereteink alapján igazolható-e a módszer alkalmazhatósága.[48] A tanulmány három kísérletet tárgyal: az első a saját szerzőségű szövegek és a fordítások kapcsolatát, a második a műfaji-tematikai besoroláson alapuló beszédmód szerinti elkülönülést mutatja be. A harmadik kísérletben a stilometriai elemzések hatékonyságának a növelését vizsgáljuk, amelynek egyik lehetőségét egy olyan fázis beiktatásával képzeljük el, amelyben a digitális szótár mint történeti szöveget normalizáló eszköz jut szerephez.

A teljes Mikes-életmű betűhű kritikai kiadása mintegy 6000 oldalnyi terjedelmű és kb. 1,5 millió szót tartalmaz. A saját szerzőségű *Törökországi levelek* mellett Mikes munkásságának jelentősebb része franciából való fordításokból áll, amelyeket Hopp Lajos a következő kategóriákba sorolt: erkölcsnevelő értekező próza önálló átültetése;

[45] Eder, Rybicki and Kestemont, „Stylometry with R," 107–121, `https://doi.org/10.32614/rj--2016-007`.

[46] Az alkalmazásról részletesen: *Shtylo,* hozzáférés: 2019.02.20, `https://github.com/dobijan/shtylo/wiki`. Dobi Jan Sándor, Mészáros Tamás és Kiss Margit, „Shtylo: stilometriai elemzések webes támogatása," in *XIV. Magyar Számítógépes Nyelvészeti Konferencia,* szerk. Vincze Veronika (Szeged: Szegedi Tudományegyetem Informatikai Tanszékcsoport, 2018), 423–436.

[47] A vizsgálat a kritikai kiadás szöveganyaga alapján történt: Mikes Kelemen, *Összes művei,* s. a. r. Hopp Lajos (Budapest: Akadémiai Kiadó, 1966–1988).

[48] Az elemzések informatikai hátterét Dobi Jan Sándor „Shtylo: egy webalkalmazás az R-beli stilometriacsomag, a Stylo számára" című önálló laboratóriumi dolgozat (BME Villamosmérnöki és Informatikai Kar Méréstechnika és Információs Rendszerek Tanszék, konzulens Mészáros Tamás, 2016) taglalja.

 DOI 10.31400/dh-hun.2019.2.336

szépprózai átdolgozások; elmélkedő, didaktikus, kegyességi próza; klasszikus történeti értekező próza.[49] E műfaji-tematikai besorolást alapul véve a teljes életmű beszédmód szerinti elkülönítésben három főbb kategóriába sorolható: élőbeszéd, vallásos, erkölcsi (1. táblázat).

Műcím	Rövidítés	Tokenek száma	Saját vagy fordítás	Beszédmód
Törökországi levelek, Misszilis levelek	TL	105860	saját	élőbeszéd
Épistolák	É	268611	fordítás	vallásos
Keresztényi Gondolatok	KG	29694	fordítás	vallásos
A Kristus Jésus Életének Historiája	KJÉ	64146	fordítás	vallásos
A Keresztnek királyi uttya	KKU	160581	fordítás	vallásos
Mulattságos napok	MN	80386	fordítás	élőbeszéd
A Valoságos Keresztényeknek Tüköre	VKT	39291	fordítás	vallásos
Az Ifjak Kalauza (A, B)	IKA, IKB	182515	fordítás	erkölcsi
Catechismus Formájára valo kőzőnséges Oktatasok (A)	CA	200489	fordítás	vallásos
Catechismus Formájára valo közönséges. Oktatások (B)	CB	193533	fordítás	vallásos
Az idö Jóll el Töltésének Módgya Minden féle rendben	IJE	40872	fordítás	élőbeszéd
Az Izraéliták Szokásárol	ISZ	30333	fordítás	vallásos
A Keresztényeknek Szokásirol	KSZ	51695	fordítás	vallásos
A Sidok és az Ujj Testámentumnak Historiája	SUT	98295	fordítás	vallásos

1. táblázat. *Mikes műveinek áttekintő táblázata a korpusz mérete, a szerzőség és a beszédmód szerinti besorolás alapján*

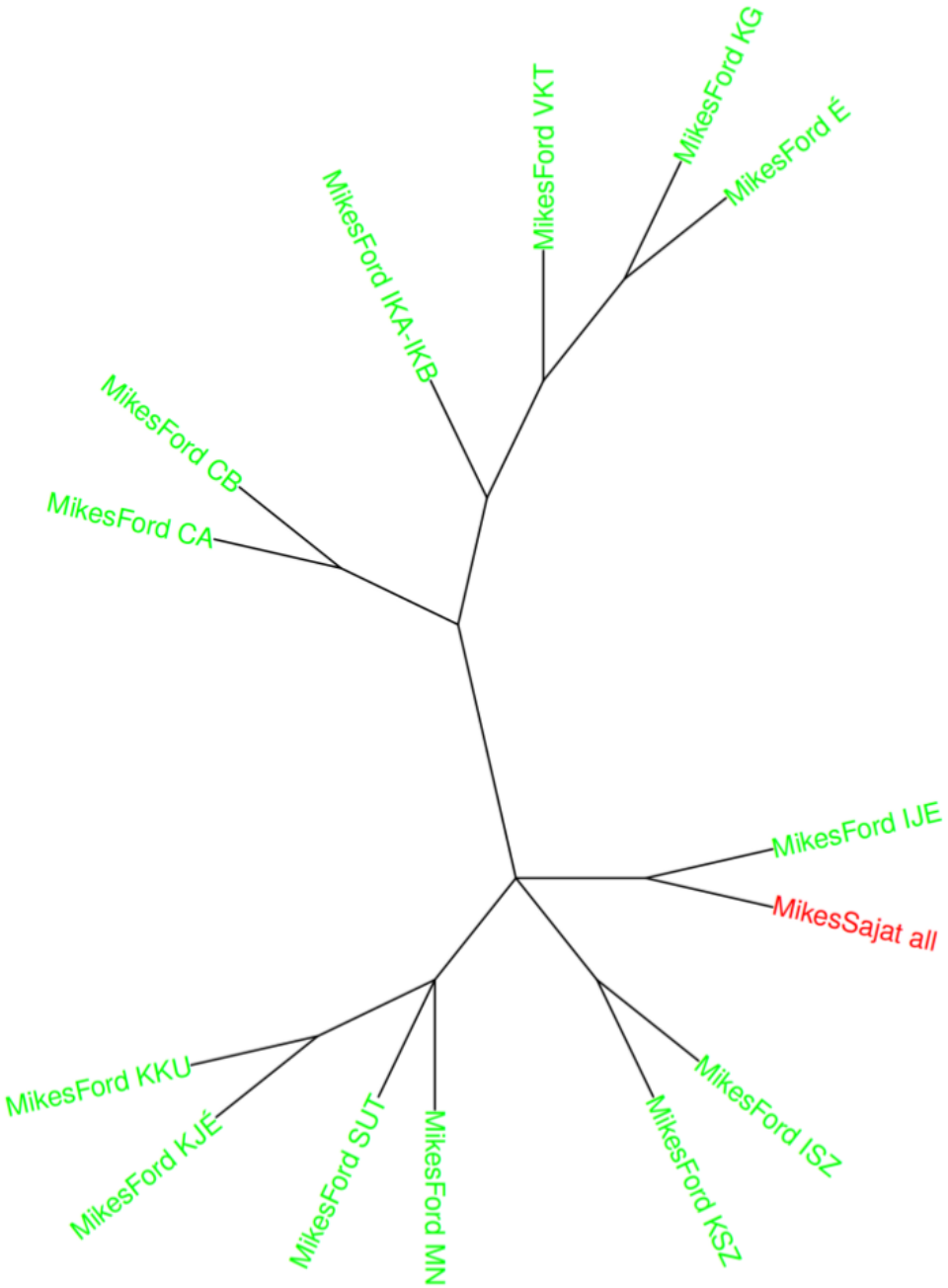
Az első kísérletben a saját szerzőségű levelek és a fordítások kapcsolatát vizsgáltuk. Mikes Kelemen fordítói munkásságát kevésbé tartják számon, holott maga az író sem határolta el egymástól alkotói tevékenysége e két területét, a levélírás és a fordítás szorosan érintkezik.[50] A szerzői életmű ugyanakkor túl nagy terjedelmű ahhoz, hogy manuális eszközökkel átfogó nyelvi vizsgálatot lehessen rajta végezni. Áttekintő elemzés elvégzéséhez számítógépes elemzőmódszerek nyújthatnak segítséget. Ennek egy korai példája az a részint számítógéppel, részint manuálisan, a szókészlet reprezentatív mennyiségén végzett lexikológiai elemzés, amely rávilágított, hogy az író saját szerzőségű munkái, a levelek és a fordítások között különbség tapasztalható a szókészlet markáns elkülönülése, az előremutató szóalkotási technikák és a szövegformálás tekintetében.[51] Jelen kutatásban arra voltunk kíváncsiak, hogy a stilometriai elemzés

[49] Hopp Lajos, *A fordító Mikes Kelemen* (Budapest: Universitas Kiadó, 2002), 133–385.

[50] Hopp, *A fordító Mikes,* 133–385.

[51] Kiss Margit, „»más értelmet adni ezeknek a szóknak«: Mikes Kelemen szóhasználatához," in *Nunquam autores, semper interpretes: A magyarországi fordításirodalom a 18. században,* szerk. Lengyel Réka (Budapest: MTA BTK Irodalomtudományi Intézet, 2016), 58–68.

hogyan tudja elkülöníteni a saját szerzőségű művet a fordításoktól, vagyis a Mikes-szókészlet teljes egészét érintő majdani vizsgálatba a stilometriai elemzés bevonható-e, s alkalmazható-e magyar nyelvű 18. századi szövegekre. Az első kísérletben a saját művek (piros) és a fordítások (zöld) (l. ábra) szétválasztására tettünk kísérletet a *Shtylo*val. Általánosságban elmondható, hogy a paraméterek beállítása egy iteratív folyamat, a beállítás helyességét az a priori tudással próbáljuk ellenőrizni.



1. ábra. *A fordítások és a saját művek elrendeződése konszenzusfán. Paraméterezés a Shtyloban: 100-800 MFW 2-grams Culled @ 0-100 %, Eder's Delta distance Consensus 0,9*

 DOI 10.31400/dh-hun.2019.2.336

Két csoportra osztottuk a műveket (1. táblázat, 1. ábra).[52] A fordításokat tartalmazó csoport jóval több művet és hosszabb szövegeket tartalmazott, mint a másik. Mivel a *Classic Delta* érzékeny a korpuszok méretére, ezért *Eder Deltá*-ját alkalmaztuk. Emellett szólt még az az érv is, hogy ez a *távolságmérték* a nem izoláló jellegű nyelvek esetében jobb eredményeket ad. Az elemzési eljárások közül a *konszenzusfá*t választottuk, amely széles körben elterjedt mód a stilometriai elemzésekben, és alkalmas arra, hogy a különböző művek közti hasonlóságot és eltérést jól láttassa. Ebben az eljárásban több egymás utáni *klaszteranalízis* fut, amelynek során több különböző beállítás mellett történik az összehasonlítás. A beállítások többségében egy adott hasonlóság kimutatható, a *konszenzusfa* ezeket ábrázolja. Ebben az elemzésben csak azokat a hasonlóságokat tartjuk meg, amelyek a beállítások többségénél megjelennek. Maga az ábrázolás nem a szövegek közti távolság nagyságát ábrázolja, hanem a hasonlóság gyakoriságát mutatja. A különböző beállításokkal végzett kísérletek minél többször mutatnak hasonlóságot, annál szorosabb kapcsolatot mutatnak, és annál közelebb helyezkednek el egymáshoz a fán.[53] Az elemzés eredményeképpen (1. ábra) a Mikes-művek négy fő csoportba különültek el. A beállítások módosításait követően is ugyanazt láttuk, hogy egyedül egy fordítás (IJE) esik közel a saját szerzőségű műhöz (TL), a futtatások 90%-a azt mutatta, hogy van köztük kapcsolat. Ez az eredmény nem hozott váratlan meglepetést abban a tekintetben, hogy a Mikes-korpusz feldolgozásával készülő *Mikes-szótár*[54] szócikkeinek írása során szoros olvasással is valószínűsíthetőnek tűnt e két mű szókészlettani közelsége, de a hasonlóság gyanújába egy másik mű is keveredett, amelyet majd az elemzéshatékonyság növelésével végzett kísérlet fog igazolni. A többi fordítás külön konszenzuságban található, ugyanakkor az látszik, hogy e művek között is fennáll a kapcsolat, amely a konszenzus erősségének beállítása során végig megmaradt, így nem rendeztük a korpuszt egymástól független művekre. Az is kiolvasható ugyanakkor, hogy a fordításvariánsok (CA, CB) – amelyek között minimális mértékű nyelvi eltérés található – egy közös ágon találhatók, ezen túlmenően a saját szerzőségű mű és a fordítások jól elkülönülnek egymástól. Az elemzés során a konszenzusküszöböt magasra állítottuk, hogy a fordítások és a saját művek közti hasonlóság a legjobban látszódjék. A konszenzusküszöb megadásánál azt határozzuk meg, hogy az elvégzett kísérletek hány százalékában jelenjen meg a hasonlóság.

Az eltérő hosszúságú szövegek (lásd az adatokat az 1. táblázatban) torzíthatják a statisztikai elemzéseket, éppen ezért az elemzés során lehetőség van a szövegek mintavételezésére, amelynek során a szöveghosszakat hasonló méretűre állítjuk be. Hogy a Mikes-szövegek eltérő hossza közti különbség ne torzítsa a statisztikai elemzést, a

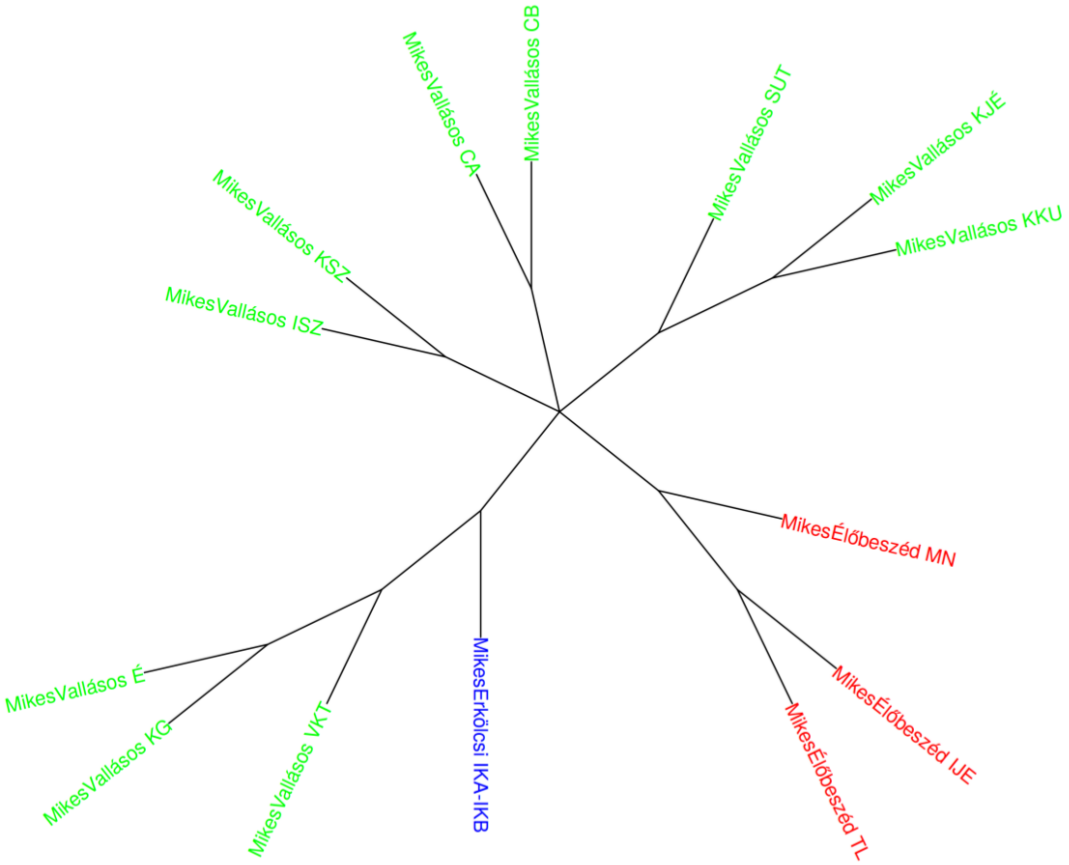
[52] A kísérletekben szereplő szövegkorpusz a mellékletben található: a digitalizált kritikai kiadás betűhű Mikes-szövegeit tartalmazza a sajtó alá rendező bejegyzései, kommentárjai nélkül.

[53] Eder, Rybicki and Kestemont, „Stylometry with R," 107–121, `https://doi.org/10.32614/rj--2016-007`; Maciej Eder, „Visualization in Stylometry: Cluster Analysis Using Networks," Digital Scholarship in the Humanities 32, 1. sz. (2017): 50–64, `https://doi.org/10.1093/llc/fqv061`.

[54] A digitális Mikes-szótár a teljes írói korpuszt feldolgozó szótár, amely 2010 óta folyamatosan készül. Jelenlegi fázisában alaki rendszerezést ad, ami azt jelenti, hogy minden mikesi szóelőfordulás mai alakú címszóhoz tartozik. Az állomány folyamatosan bővül, az eddig elkészült anyag itt érhető el: Kiss Margit szerk., *Mikes-szótár: elektronikus adatbázis* (Budapest: MTA BTK Irodalomtudományi Intézet), hozzáférés: 2019.02.20, `http://www.mikesszotar.iti.mta.hu`.

mintavételezés segítségével normalizáltuk a szöveghosszúságot (Sampling 1000). A stílusmarkerek közül a leggyakoribb bigramok (MFW 2-grams) beállítás bizonyult megfelelőnek. Biztató eredmény, hogy a stilometriai elemzés alátámasztotta a saját művek és fordítások viszonyáról meglévő eddigi ismereteinket az életművel kapcsolatban, s ez egyben azt is jelenti, hogy ezzel a módszerrel az egyes művek közti lexikai alapú hasonlóságok, különbözőségek feltérképezése további, részletes kutatás tárgyát tudja képezni a jövőben.

A következő kísérletben egy olyan vizsgálatot végeztünk, amelyben a mikesi életmű darabjain a beszédmód szerinti elkülönülést kívántuk láttatni (1. táblázat, 2. ábra). Arra voltunk kíváncsiak, hogy a *Shtylo* segítségével lehetőségünk van-e az író életművében jól elkülöníthető egyházi, erkölcsi tematikájú és élőbeszédszerű műveket a szókészlet elkülönülése alapján statisztikai szempontból is igazolhatóan megkülönböztetni.



2. ábra. *A művek tematikus elrendeződése konszenzusfán. Paraméterezés a Shtyloban: 100-1000 MFW 2-grams, Culled @ 0-80%, Canberra distance, Consensus 0,5*

A Mikes-korpuszt három élőbeszédszerű (piros), egy erkölcsi (kék) és tíz vallásos mű (zöld) alkotja. Mivel a tematikai meghatározottság ebben az esetben erősen a tartalmas szavak vizsgálatára helyezi a hangsúlyt, így a *Canberra távolság* tűnt a legadekvátabbnak. A három csoport ez esetben ugyancsak eltérő hosszúságú műveket tartalmazott (1. táblázat), így ezt mintavételezéssel kompenzáltuk (lásd az előző kísérletben leírtakat), hogy a statisztikai elemzés ne torzuljon. Az eredmény vizualizálására itt is a *konszenzusfa* tűnt megfelelőnek. A konszenzusküszöb értéke ebben a kísérletben

 DOI 10.31400/dh-hun.2019.2.336

alacsony (0,5), mert az volt a kérdés , hogy a *Canberra távolság* alkalmazásával a művek kapcsolatban lesznek-e egymással, vagy távol kerülnek. Az látszik, hogy a *Shtylo* segítségével az élőbeszédszerű szövegeket (MN, IJE, TL) jól külön tudtuk választani a többitől, továbbá az erkölcsi témájú szöveg (IKA, IKB) egy ágba sorolódik a vallási témájú szövegek egy részével (É, KG, VKT), ami igazolhatja azt is, hogy ez a fajta tematikai megkülönböztetés nem jár feltétlenül a szókészlet jelentős elkülönülésével. Ebben a kísérletben a stilometriai elemzés arra volt képes, hogy az élőbeszédszerű szövegeket markánsan elkülönítse a többitől, s ez tekinthető a legerősebb stilisztikai markernek ebben a kísérleti korpuszban. Ez esetben az élőbeszédre jellemző csoportban a saját szerzőségű művek mellett ott találunk két fordítást is.

A stilometria történeti fejlődésében fontos szerep jutott a mennyiségileg meghatározható jelenségek, a szerzői megkülönböztető jegyek meghatározásának – állítja Holmes –, s ebben a tekintetben a lexikális jegyek túlsúlyba kerültek, ám az utóbbi időszakban a szintaktikai, szemantikai, grammatikai, szófajtani, morfológiai megkülönböztető jegyek is megjelentek, amelyek elemzéséhez egyre több informatikai támogatás kínálkozik, és amelynek eredményeképpen az összetett elemzések pontosabb, megbízhatóbb eredmények elérését teszik lehetővé.[55] Minthogy a gépi feldolgozásra alkalmas szövegek mennyisége folyamatosan nő, fejlődik, és egyre hatékonyabbá válik a stilometriai módszerek eszköztára,[56] így lehetővé válik a szerzőségi vizsgálatok elvégzése nagyméretű szövegkorpuszokon is.[57] A lexikai alapú elemzés javításának egyik lehetséges, további módját a harmadik kísérlet mutatja be. Ha a statisztikai szövegelemzés szógyakoriság-alapú vizsgálatai során az adott szövegkorpuszban megjelenő szóelőfordulásokkal számolunk, akkor történeti szövegek esetében különösen nagy alaki változatossággal találkozunk, Mikes estében például *ekepen, eképen, e képen, ekeppen, eképpen, e képpen, ekkepen, ekképen, ekképpen*. Ha ezt a sokféleséget a történeti alakok normalizálásával csökkenteni lehetne, akkor az elemzés hatékonyságát növelhetnénk, mivel a szóalakok változatainak a redukálásával csak az alapalakban álló szavakat (pl. *ekképpen*) hasonlítanánk össze és nem az alakváltozataikat, illetve paradigmatikus alakjaikat is (pl. *ekepen, eképen, e képen, ekeppen, eképpen, e képpen, ekkepen, ekképen*), mintha külön szótári alakok lennének. Ezt az előfeldolgozást támogathatják a gépi morfológiai elemzők is, ám magyar történeti szövegek esetében az ilyen jellegű automatizált elemzés közel sem egyszerű megoldás.[58] A történeti szövegek gépi automatikus morfológiai elemzését más, megbízhatóbb megoldással is pótolhatjuk, például ha a normalizálást szótár segítségével végezzük el. A Mikes-korpusz normalizálásához a készülő digitális *Mikes-szótár* segítséget ad, hiszen alaki rendszerezés révén minden egyes szövegben szereplő régies alakú szót mai címszóhoz rendel, ezáltal a stilometriai elemzés során nem a régies, paradigmatikus alakban

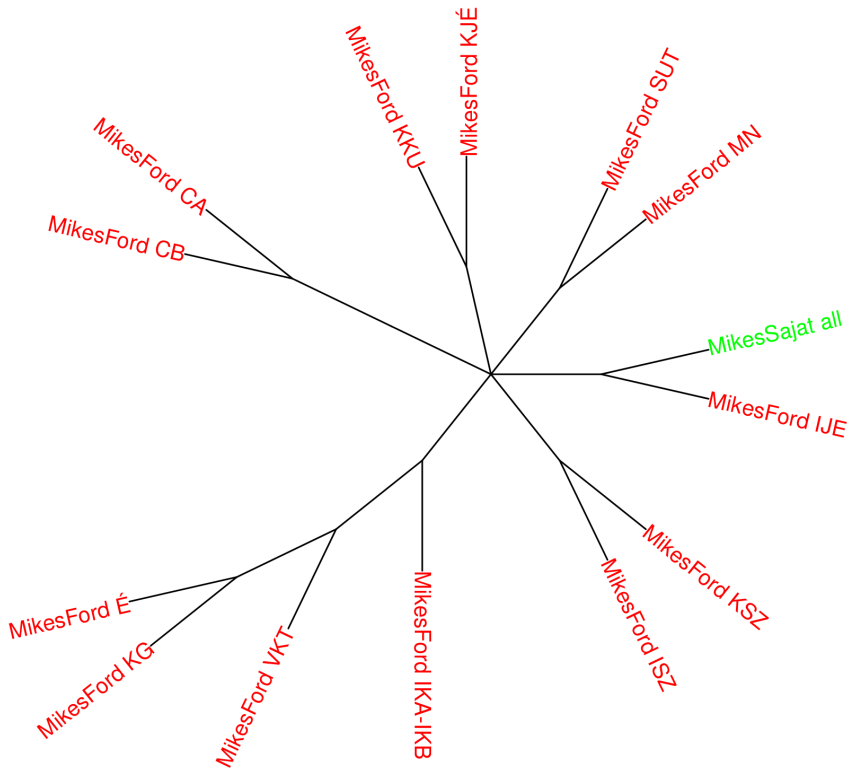
[55] David Holmes, „Authorship Attribution," *Computers and the Humanities* 28, 2. sz. (1994): 87–106, https://doi.org/10.1007/bf01830689.

[56] Évről évre újabbak látnak napvilágot, pl. Justin Stover and Mike Kestemont, „The Authorship of the Historia Augusta: Two New Computational Studies," *Bulletin of the Institute of Classical Studies* 59, 2. sz. (2016): 140–157, http://dx.doi.org/10.1111/j.2041–5370.2016.12043.x.

[57] Craig, „Stylistic Analysis," 280; MacDonald Pairman Jackson, „Determining Authorship: A New Technique," *Research Opportunities in Renaissance Drama* 41 (2002): 1–14.

[58] Kiss Gabriella, Kiss Margit és Pajzs Júlia, „A Nagyszótár történeti korpuszának elemzéséről," *Magyar Nyelv* 100, 2. sz. (2004): 185–191.

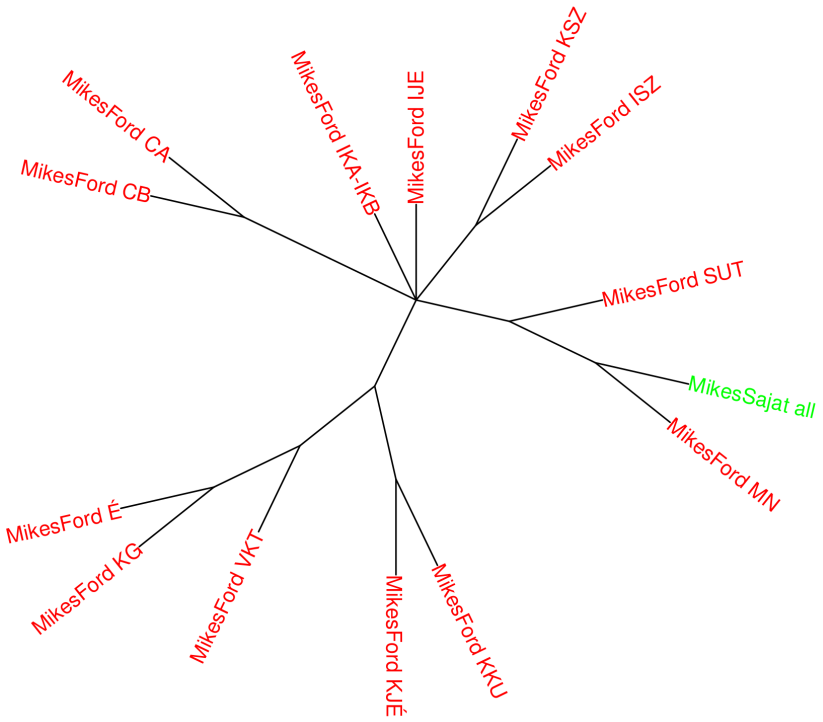
lévő szavakat hasonlíthatjuk össze egymással, hanem a mai szótári alapalakokat. Így megbízhatóbb eredményt kaphatunk az életmű szókészletére alapuló vizsgálattal kapcsolatban. Az utolsó kísérletben tehát a stilometriai elemzést kiegészítettük egy olyan előzetes munkafázissal, amelyben a *Mikes-szótár* segítségével végeztük el a szövegek normalizálását.[59] A szavak szövegbeli előfordulási alakjait helyettesítettük a standardizált, mai szótári alakban (nem toldalékolt!) álló megfelelőikkel annak érdekében, hogy növeljük a statisztikai elemzés hatékonyságát a szavak közti különbségek redukálásával, amely a toldalékolás és a történeti szöveg egyenetlensége miatt van jelen. Ebben a kísérletben két elemzést végeztünk ugyanazokkal a beállításokkal, hogy összehasonlíthatóvá váljék a különbség a két futtatás között.



3. ábra. *Mikes-művek elemzése az eredeti szövegek felhasználásával. Paraméterezés a Shtyloban: 100-2000 MFW 2-grams, Culled @ 0%, Classic Delta distance, Consensus 0,5*

[59] Margit Kiss and Tamás Mészáros, „Creating an Extended Author's Dictionary to Support Digital Literary Research," *Abstracts of DH Benelux Conference, June 6–8, 2016,* hozzáférés: 2019.02.20, `http://2016.dhbenelux.org/wp-content/uploads/sites/4/2016/05/89_Kiss-Meszaros_FinalAbstract_DHBenelux_2016_long.pdf`.

 DOI 10.31400/dh-hun.2019.2.336



4. ábra. *Mikes-művek elemzése a normalizált szóalakokkal. Paraméterezés a Shtyloban: 100-2000 MFW 2-grams, Culled @ 0%, Classic Delta distance, Consensus 0,5*

Az első esetben az eredeti Mikes-szövegeket elemeztük (3. ábra), a másodikban a szótári szavakra lecserélt normalizált változatot (4. ábra). A leggyakoribb vizsgálandó elemeknél szintén a leggyakoribb bigramok (MFW 2-grams) beállítást választottuk. A *Classic Delta* távolságmértéket alkalmaztuk, amelyet a normalizálás indokolttá tesz. A szövegkorpusz egyenetlenségét mintavételezéssel kompenzáltuk. Az elemzés eredményének a vizualizálásához a konszenzusfát alkalmaztuk. A konszenzusküszöb kialakításánál arra törekedtünk, hogy az életmű egyes darabjai kapcsolatban maradjanak egymással. Az elemzés eredményéből látható, hogy mindkét futtatásnál a CA, CB fordításvariánsok értelemszerűen nagyon közel maradt egymáshoz, amely az elemzés relevanciáját erősíti, hiszen nagyon minimális eltérés van a két szöveg között. Az eredeti szövegek elemzésénél az É, a KG, VKT ágához az IKA, IKB variánsai esnek közelebb. Míg a normalizált korpuszon az IKA, IKB variánsai helyet cserélnek a KJÉ és a KKU írásokkal. További látványos különbség, hogy a saját szerzőségű művek (TL) az eredeti szövegeket tartalmazó korpuszvizsgálat esetében IJE írással vannak legközelebbi kapcsolatban (ahogy a saját művek és a fordítások esetén is láthattuk), ugyanakkor a normalizált korpuszon végzett elemzés során az MN esik hozzá legközelebb, továbbá ugyanazon az ágon található még kicsit távolabb a SUT. Az eredeti szövegvizsgálat során a SUT és az MN került szoros kapcsolatba egymással.

A szótár szerkesztése során empirikus megfigyeléssel is érzékelhető volt, hogy a *Törökországi levelek* (TL) lexikális anyaga szorosabb kapcsolatban van azokkal a fordí-

tásokkal, amelyeket az itt bemutatott stilometriai elemzések eredményeztek. Mindennek további, mélyreható és átfogó feltárásához megvan a kiindulási eszköz, amely a terjedelmes életmű módszeres feldolgozásának egyik lehetséges módja. Az itt bemutatott példák a stilometriai módszerek Mikes-korpuszra történő alkalmazhatóságát támasztják alá. Jelen keretek között nem cél a bemutatott eredmények további, mikesi életművel kapcsolatos mikrofilológiai elemzése, ugyanakkor egy nyelvi-alkotói folyamatokat feltáró későbbi, további adatelemzésen alapuló munka kezdeti lépéseként értelmezendő. Az írói munkásságot terjedelme miatt nyelvi szempontból részleteiben, egyes aspektusaiból vizsgálták ez idáig.[60] E munkának további kutatási iránya lehet a stílus fogalmának, értelmezési kereteinek a továbbgondolása, amely az új módszertannak köszönhetően is formálódik.[61] A digitális korpusz és a szótár segítségével, valamint az informatikai támogatású elemzőmódszerek alkalmazásával lehetőség nyílik nagyobb léptékű vizsgálatok elvégzésére a jövőben. Egyúttal ez azt is jelenti, hogy a digitális szótárakkal szembeni elvárásokat, feladatokat is revideálnunk kell. A digitális szótár ellátja a hagyományos szótári funkciókat, ezen túlmenően strukturált szövegkorpuszként az informatikai alapú szöveg- és korpuszelemzést, annak hatékonyságát növelő eszközként is képes támogatni.[62]

5. Összegzés

A szépirodalmi szövegeket feldolgozó stilometriai kutatások gyakran heves viták kereszttüzébe kerülnek, sokan a létjogosultságukat is kétségbe vonják, holott ha segédeszközként tekintünk rájuk a filológiai vizsgálatokban, és nem egyedüli módszerként, akkor árnyaltabb képet kaphatunk e területről.[63] A dolgozatnak ezeket az anomáliákat nem volt célja bemutatni, helyette inkább azokra az eredményekre koncentrált, amelyek azt támasztják alá, hogy a statisztikai alapú szerzőségi, stilometriai elemzés olyan eljárások közé tartozik, amely támogatni képes a szoros olvasás során vizsgálandó problémák megoldását. Ennek érdekében e tudományterület jelenlegi eredményeinek és a kísérletek hátterének módszertani feltárásához nyújtott áttekintésén túl konkrét stilometriai elemzéseket is bemutatott a dolgozat, amelyek alátámasztották az elvégzett kísérletekben e módszer relevanciáját. A továbblépés egyik lehetséges iránya az, hogy még pontosabbá tegyük a stilometriai elemzést, amelynek például

[60] Például Szabó T. Attila, „A székely nyelvjárások a magyar irodalomban," *Új Látóhatár* 4 (1989): 549–557.

[61] Nemzetközi diskurzusban pl. Berenike Herrmann, Karina van Dalen-Oskam and Christof Schöch, „Revisiting Style, a Key Concept in Literary Studies," *Journal of Literary Theory* 9, 1. sz. (2015): 25–52, https://doi.org/10.1515/jlt−2015−0003.

[62] Margit Kiss and Tamás Mészáros, „Rethinking the Role of Digital Author's Dictionaries in Humanities Research" in *Proceedings of the XVIII EURALEX International Congress,* Simon Krek, Jaka Čibej, Vojko Gorjanc and Iztok Kosem eds. (Ljubjana: Ljubjana University Press, 2019), 871–880; Mark Andrew Algee-Hewitt, „The Hidden Dictionary: Text Mining Eighteenth-Century Knowledge Networks" in *Digital Humanities 2018, DH 2018, Book of Abstracts,* eds. Jonathan Girón Palau and Isabel Galina Russell (Mexico City: El Colegio de México, UNAM, and RedHD, 2018), 146–147.

[63] A többféle elemzés kombinációján alapuló vizsgálatot és a vizsgálati eredmények valószínűségéről ír Patrick Juola, „The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions," *Digital Scholarship in the Humanities* 30, 1. sz. (2015): 100–113, https://doi.org/10.1093/llc/fqv040; Grieve, „Quantitative Authorship Attribution," 251–270, https://doi.10.1093/llc/fqm020.

 DOI 10.31400/dh-hun.2019.2.336

egyik módja lehet a digitális szótár bevonása a történeti szöveg normalizálásába. Másik lehetséges irány a kapott eredmények felhasználásával további nyelvi-filológiai elemzések elvégzése, amellyel a hatalmas életművel kapcsolatos tudásunk tovább gazdagodik, ez azonban már e dolgozat keretein túlmutató feladat.

> Sokat vitatott dolog még mindig, hogy azok, akik az irodalommal foglalkoznak, miért nem szeretik a számok világát: nem bíznak bennük, és nem értik, hogy az ember hogy tud valami olyan banalitással foglalkozni, mint hogy irodalmi szövegekben megszámoljon valamit.[64]

The Potentials of Stylometry Analysis of Hungarian Historical Text Corpora

This paper discusses the potentials of the computer assisted analysis of Hungarian historical texts, which relies on the methods of linguistic, literary studies, information technology and statistics. It reviews the characteristics and applicability of different stylometric methods and demonstrates their use by case studies based on Kelemen Mikes' (1690-1761) works. It examines the relationships between Mikes' own works and his translations as well as the thematic separation of his works. The case studies highlight that the effectiveness of the stylometric analyses of Mikes' writings can be improved by applying the digital Mikes' dictionary.

Keywords:
authorship attribution, stylometry, digital author's dictionary, Kelemen Mikes

[64] Hugh Craig, a University of Newcastle professzor emeritusa, egyetemi weboldalán *Figures of speech* címmel megjelent összegzés, hozzáférés: 2019.02.20, `https://www.newcastle.edu.au/profile/hugh-craig`, (ford. tőlem).