

PSEUDO-AURELIUS
VICTOR *ORIGO GENTIS*
ROMANAE CÍMŰ MŰVÉNEK
STILOMETRIAI ELEMZÉSE

TARTALOMJEGYZÉK

1	Bevezetés	1
2	A stilometriáról	2
2.1	A stilometria története	2
2.1.1	Kezdetek	3
2.1.2	A számítógép elterjedése	4
2.1.3	Újabb módszerek.....	5
3	Az <i>Origo gentis Romanae</i>	7
4	Alkalmazott eszközök és eljárások	9
4.1	Szövegek.....	9
4.2	Szoftverek	9
4.2.1	JGAAP.....	10
4.2.2	Stylometry with R	10
4.3	Statisztikai eljárások.....	11
4.3.1	Főkomponens analízis	12
4.3.2	Klaszteranalízis	12
5	Vizsgálatok	13
5.1	Első vizsgálat	13
5.2	Második vizsgálat.....	16
6	Összegzés.....	22
7	Irodalomjegyzék:	23

1 BEVEZETÉS

Dolgozatom témája a Sextus Aurelius Victor neve alatt fennmaradt *Origo gentis Romanae* című, Róma őstörténetével foglalkozó mű stilometriai elemzése. A vizsgálat több célkitűzéssel rendelkezik:

1. Megállapítani, vajon az *Origo gentis Romanae* nyelvhasználata beleillik-e az Aurelius Victor neve alatt fennmaradt gyűjteménybe, vagyis lehet-e ugyanaz a szerzője, mint a gyűjteményben hagyományozott többi szövegnek.
2. Ha az előző kérdésre negatív a válasz, akkor kísérletet tenni az *Origo gentis Romanae* szerzőjének megállapítására.

A dolgozat elején röviden ismertetem a stilometriát mint módszert: kialakulását és fejlődésének történetét, különös tekintettel a klasszika-filológiában való alkalmazásának előzményeire. Ezután rátérek az *Origo gentis Romanae* ismertetésére, majd a használt stilometriai módszerek, eljárások és eszközök leírása következik. Végül az ismertetett eljárások gyakorlati alkalmazásának leírása következik: a vizsgálatok menetének leírása és az eredmények elemzése.

2 A STILOMETRIÁRÓL

A stilometria (a. m. stílusmérés) az irodalmi stílus elemzése statisztikai módszerekkel, leggyakrabban egy mű szerzőjének meghatározása céljából. Bár eredetileg szépirodalmi művekre alkalmazták, ma már olyan egyéb írásos anyagok vizsgálatát is végzik stilometriai eszközökkel, mint például interneten keletkezett szövegek: blogok¹, e-mailek², Twitter-bejegyzések³ és egyéb hasonló típusú szövegek, valamint a szoftverek forráskódjai⁴. Sőt, manapság már nemcsak szövegek, hanem zeneművek⁵ és festmények⁶ elemzésére is használják a stilometriát. A módszer megbízhatóságát bizonyítja, hogy egyes országokban az igazságügy is használja szöveges bizonyítékok elemzésére.⁷ Eredményessége másik fokmérője, hogy egyes szakemberek az internet anonimitását féltik a statisztikai szövegelemzéstől, és már „ellenszereket” is kidolgoztak ellene.⁸ Bár a stilometria az angolszász szakirodalomban ismert és elismert módszer, Magyarországon mégis alig írnak róla és alig használják, ugyanakkor ez jellemző az úgynevezett digitális bölcsészettudományok egyéb területeire is. Többek között a klasszika-filológiában⁹, illetve a közép-¹⁰ és neolatin¹¹ irodalom kutatásában is használt és bevált eszköz.

2.1 A stilometria története

A matematika használatára az irodalomban már az ókorban is találhatunk példákat. Saunaka, szanszkrit grammatikus Kr. e. 300 körül megszámlolta a *Rigvéda* szótagjait, szavait és versszakait, hogy a szöveg helyes megőrzését elősegítse. Kr. e. 180 körül az alexandriai könyvtárban tevékenykedő Samothrakéi Aristarchos egyebek mellett az *Ilias*ban és *Odysseiá*ban megjelenő kifejezéseket csoportosította gyakoriság szerint.¹² A maszoréták, a

¹ GOSWAMI-SARKAR-RUSTAGI (2009)

² CALIX-CONNORS-LEVY-MANZAR-MCCABE-WESTCOTT (1998)

³ BHARGAVA-MEHNDIRATTA-ASAWA (2013)

⁴ CALISKAN-ISLAM (2013)

⁵ BACKER-KRANENBURG (2005)

⁶ QI-TAEB-HUGHES (2013)

⁷ WALL (2009, 65)

⁸ BRENNAN-AFROZ-GREENSTADT (2012); BRENNAN-GREENSTADT (2009); MCDONALD-AFROZ-CALISKAN-ISLAM-STOLERMAN-GREENSTADT (2012)

⁹ FORSYTH-HOLMES-TSE (1999); SMITH-KELLY (2002); MICHAELSON-MORTON (1972); GURNEY-GURNEY (1998); TSE-TWEEDIE-FRISCHER (1998); WEISBERG-BINGHAM-SELLEW-ADERINKOMI (2011)

¹⁰ KESTEMONT-MOENS-DEPLOIGE (2013); KESTEMONT (2012)

¹¹ TWEEDIE-HOLMES-CORNS (1998)

¹² RUDMAN (2006), 611

zsidó szövegkritikusok, főleg Aaron Ben-Asher, a héber Biblia szövegének pontos hagyományozása érdekében szintén használtak matematikai eszközöket.¹³

2.1.1 Kezdetek

A gondolat, hogy egy irodalmi kérdés megválaszolására egy matematikai módszert, a statisztikát alkalmazzák, először a XIX. század második felében bukkant fel. 1851-ben Augustus de Morgan, angol logika-szakértő egy barátjának írt levelében felvetette, hogy Szent Pál apostol leveleinek kérdéses szerzősége eldönthető, ha megnézzük, hogy az egyik szöveg nem tartalmaz-e nagyobb arányban hosszabb szavakat, mint a másik.¹⁴ Általánosan elfogadott, hogy ezt tartjuk a nem hagyományos szerzőség-meghatározás kezdetének.

Lewis Campbell 1867-ben megjelent szövegkiadásának előszavában Platón *Sophista* és *Politicus* című dialógusait a ritka szavak előfordulásának aránya alapján a görög filozófus kései művei közé sorolja.¹⁵ De Morgan elméletét Thomas C. Mendenhall vizsgálta és fejlesztette tovább az 1880-as években, és eredményeit 1887-ben publikálta.¹⁶ Ő végezte el az első nem hagyományos kísérletet, melyben Bacon, Marlowe és Shakespeare műveiben vizsgálta a szavak átlagos hosszúságát, hogy eldöntse néhány megkérdőjelezett szerzőségű színdarab hovatartozását.¹⁷ Talált is olyan hasonlóságokat Shakespeare és Marlowe között, amelyek még ma is (sokkal kifinomultabb) vizsgálatok tárgyát képezik, bár a szavak átlagos hosszán alapuló összehasonlításról már rég bebizonyosodott, hogy nem csak, hogy szerzők közötti vizsgálatokban nem állja meg a helyét, de egyetlen szerző életművén belül sem.¹⁸

Mendenhall tanulmányával majdnem egy időben kisebb fellendülés figyelhető meg a stilometriai kutatásban. 1888-ban jelent meg a Conrad Mascol álnéven publikáló William Benjamin Smith két tanulmánya szintén a szent páli levelek kapcsán.¹⁹ A korszakban Lucius Sherman a mondatok átlagos hosszúságát és a központosítást, mint a szerző meghatározására alkalmas jellemzőket vizsgálta.²⁰ Szintén a mondathosszúságot állította vizsgálati középpontjába G. W. Gerwig.²¹ Wincenty Lutosławski ezzel szemben elsősorban kronológiai

¹³ DELCOURT (2002), 980

¹⁴ HOLMES (1998), 112; DE MORGAN (2010), 215-216

¹⁵ BURNET (1928), 9; CAMPBELL (1867); GRZYBEK (2014), 70

¹⁶ MENDENHALL (1887)

¹⁷ RUDMAN (2006), 611

¹⁸ JUOLA (2008), 8

¹⁹ MASCOL (1888a); (1888b); RUDMAN (1997) 354

²⁰ SHERMAN (1888); (1893)

²¹ GERWIG (1894); RUDMAN (2006) 611

kérdések megválaszolására keresett megoldást. A platóni dialógusok sorrendjének meghatározása során használta a stilometriát, amely kifejezést is ő használta először tudományos értelemben.²² Kiindulópontnak azokat a dialógusokat használta, amelyek keletkezési idejét biztosan tudjuk, és az ezektől való stílusbeli eltérést mérte, így állapítva meg a bizonytalan datálású művek időrendjét. A stilometriát a grafológiához hasonlította, mivel a stílus is olyan egyedi (ha nem egyedibb), mint a kézírás, és amellet érvelt, hogy ha a grafológia általánosan és hivatalosan is elfogadott eszköz, akkor a stilometria is legyen az.²³

A XX. század elején Robert E. Moritz kritizálta Sherman mondathosszúságon alapuló stilometriáját, és többféle stílusjegyet is megvizsgált.²⁴ Andrej Andrejevics Markov a betűsorokat tanulmányozta.²⁵ George Kingsley Zipf amerikai nyelvész 1932-es híres tanulmányában közzéteszi megfigyeléseit, melyekre a természetes nyelvek statisztikai elemzése közben jutott. A később róla elnevezett törvény szerint egy szó gyakorisági rangja és az előfordulási gyakorisága között fordított arányosság áll fenn. Tehát a gyakorisági sorban első helyen álló szó körülbelül kétszer gyakoribb, mint a második, és háromszor gyakoribb, mint a harmadik stb.²⁶ Később azt is megállapították, hogy ez a Zipf-eloszlás az élet más területein is megjelenik,²⁷ többek között a városok népességeloszlásában²⁸, a weboldalak látogatottságában²⁹ és a cégek méretében³⁰.

2.1.2 A számítógép elterjedése

A XX. század második felében a számítógépek elterjedésével átalakult a statisztika és ezért a stilometria is. Több adatot, gyorsabban és pontosabban tudtak feldolgozni és ez igaz volt a szöveges adatokra is. Nem csoda, hogy a stilometria talán legnagyobb hatású alapl műve is ebben a korszakban született. Ez Charles Frederick Mosteller és David Wallace 1964-ben megjelent könyve³¹, melyben az ún. *Föderalista értekezéseket*³² elemzik. A könyv sikerét jelzi, hogy az

²² PAWŁOWSKI (2004), 84

²³ LUTOSŁAWSKI (1897a) (1897b) (1898) PAWŁOWSKI (2004)

²⁴ MORITZ (1903); (1909); RUDMAN (1997); (2006), 611

²⁵ DELCOURT (2002); PETRUSZEWCZ (1981); RUDMAN (2006), 611

²⁶ HOLMES (1994), 93; ZIPF (1932)

²⁷ LI (2002)

²⁸ CZALLER (2012)

²⁹ ADAMIC–HUBERMAN (2002)

³⁰ AXTELL (2001)

³¹ MOSTELLER–WALLACE (2011); MOSTELLER–WALLACE (1964)

³² *The Federalist Papers*. Magyar fordításban: Hamilton, Alexander – Madison, James – Jay, John: *A föderalista : értekezések az amerikai alkotmányról*. (Ford. Balabán Péter). Budapest. 1998. A könyv 85 esszét tartalmaz, melyeket Alexander Hamilton, James Madison és John Jay nevéhez kötnek, és amelyekben az amerikai alkotmány

új stilometriai módszereket ezeken a szövegeken próbálják ki először. Mosteller és Wallace a funkciósavakra fókuszált a kutatásban. A funkciósavak a beszéd olyan összetevői, melyek önmagukban nem hordoznak jelentést, szerepük a mondatok összetartása (pl. kötőszavak, névelők, névutók, előjárósók) és ezért függetlenek a szöveg témájától. A szerzőpáros 30, a szövegekből kinyert funkciószó eloszlását elemezte, és azt az eredményt kapták, hogy a funkciósavak kiválóan alkalmasak az egyéni stílusok elkülönítésére.³³ Az általuk alkalmazott módszer annyira sikeresnek mutatkozott, hogy az azóta eltelt ötven évben a stilometriai elemzések körében domináns a funkciósavak használata.³⁴

1979-ben Ian Marriott a mondathosszúságot vette alapul a *Historia Augusta* szerzőségének megállapításához.³⁵ Eredményül azt kapta, hogy a kézirati hagyományban hat szerző neve alatt fennmaradt császárelőletrajzokat valójában egy szerző írta. 1990-ben David Sansone kritizálta³⁶ Marriott módszerét, és felhívta a figyelmet arra, hogy Tore Janson már 1964-ben megállapította, hogy antik szövegek esetében a mondathosszúság és a központosítás kiadásonként változó³⁷, ezért a mondatok és tagmondatok hosszúsága nem alkalmazható ókori latin szövegek stilometriai elemzésének alapjaként. Egy 1996-os tanulmány, amely szintén a *Historia Augusta* szerzőségével foglalkozik, több latin biográfus és történetíró prózai műveit elemezve szintén arra a következtetésre jutott, hogy a mondathosszúság nem alkalmazható megkülönböztető jegyként, mivel a *Historia Augusta* és a többi vizsgált szöveg átlaga nem mutat szignifikáns különbséget.³⁸

2.1.3 Újabb módszerek

Az 1980-as évek végén és az 1990-es évek elején John Burrows publikálta azt a cikksorozatot,³⁹ amelyben bevezeti a stilometriába a többváltozós statisztikai eljárásokat. Ezzel a kutatás új lendületet kapott, és Burrows tudománytörténeti jelentősége Mosteller és Wallace

elfogadására igyekeztek rábírní a new yorkiakat. Az összes levél *Publius* álnév alatt jelent meg 1787 és 1788 között new yorki újságokban, és bár később mindhárom író közzétette az által írt levelek sorszámainak listáját, van 12 olyan levél, amiket két író (Madison és Hamilton) is a magáénak tulajdonít. A hagyományos kutatás mind a tizenkettőt madisoninak ítélte. JUOLA (2008), 10.

³³ JUOLA (2008), 10; WEISBERG–BINGHAM–SELLEW–ADERINKOMI (2011), 5

³⁴ JUOLA (2008), 11

³⁵ MARRIOTT (1979)

³⁶ SANSONE (1990)

³⁷ JANSON (1964) A kiadások között olyan érdekes eltéréseket is kimutatott, mint például azt, hogy a pontosvesszőt gyakrabban használják a német szerkesztők, mint az angolok illetve a régebbiek, mint a modernek. (p. 29)

³⁸ FRISCHER–GUTHRIE–TSE–TWEEDIE (1996); HOCKEY (2000), 110

³⁹ BURROWS (1987); (1989); (1992a); (1992b)

mellé emelkedett. Burrows kiválasztott 75 nagy gyakoriságú funkciószt, kiszámolta ezer szavankénti előfordulási arányukat és az így kapott többváltozós adattömbön többváltozós statisztikai eljárásokat: főkomponens analízist, klaszteranalízist, diszkriminancia-analízist és korrespondencia-elemzést hajtott végre. Ezekkel vizualizálni is tudta a szövegek egymáshoz való viszonyát.⁴⁰ A többváltozós statisztikai eljárások olyan jól működtek, hogy azóta szinte minden stilometriai vizsgálat illet alkalmaz a legtöbbször a funkciósavak gyakoriságával együtt.

A számítástechnika fejlődésével nem csak, hogy egyre bonyolultabb és számításigényesebb statisztikai eljárások alkalmazása előtt nyílt meg az út, hanem egy teljesen új irányban is megindult kutatás: ez a gépi tanulás, ezen belül pedig a mesterséges neurális háló. Ez, hasonlóan a természeteshez, csomópontokból (ún. neuronok) és kapcsolatokból épül fel. Első lépésben „be kell tanítani” a hálót, ekkor a szövegben mintázatok figyel meg, és megjegyzi azokat, hogy aztán más szövegekben felismerje a megtanult mintázatok.⁴¹ A mesterséges intelligencia és a genetikai algoritmusok kutatásának újabb eredményei szintén kamatoztathatók a stilometriában.⁴²

⁴⁰ HOLMES (1998), 113; WEISBERG–BINGHAM–SELLEW–ADERINKOMI (2011), 5

⁴¹ HOLMES (1998) 115; JUOLA (2008) 53; TWEEDIE–SINGH–HOLMES (1996a) (1996b)

⁴² HOLMES–FORSYTH (1995); PAVLYSHENKO (2012); STAMATATOS (2009), 11

3 AZ ORIGO GENTIS ROMANAE

Az *Origo gentis Romanae* Itália őstörténetét meséli el Saturnus uralkodásától Róma alapításáig. A mű két XV. századi kódexben maradt fent: a *Codex Oxoniensis* (Bodleianus Canon. Class. Lat. 131, fol. 85–96) és a *Codex Bruxellensis* (Bibliotheca Regia 9755–9763, fol. 52^r–56^v).⁴³ A szöveg mindkettőben egy három műből álló korpusz részeként szerepel, amelynek tagjai az *Origo*-n kívül a *Liber de viris illustribus urbis Romae* és az *Aurelii Victoris Historiae abbreviatae* (közismertebb nevén a *Caesares*) címen szerepelnek.⁴⁴ Ez a *Historia Tripertita* néven is emlegetett összeállítás tévesen egy Kr. u. IV. századi római szerző, Sextus Aurelius Victor neve alatt maradt fent, holott a három mű közül csak az utolsót, a *Caesares*-t írta biztosan ő.⁴⁵ A három művet valószínűleg a IV. században szerkesztették össze. Az így kapott gyűjtemény lefedi a római történelmet a mondai őstörténelemtől egészen Kr. u. 360-ig.⁴⁶ Valószínűleg mindhárom mű szerzője más volt, és azt is valószínűnek tartják, hogy az összeszerkesztő nem a szerzők egyike, hanem egy negyedik személy lehetett. A kereszténység említésének hiánya arra utal, hogy mind a hárman (vagy négyen) pogányok voltak.⁴⁷

A hármas gyűjtemény első darabja az *Origo gentis Romanae*. Az *Origo* előtt rövid bevezető található:

„*Origo gentis Romanae a Iano et Saturno conditoribus, per succedentes sibimet reges, usque ad consulatum decimum Constantii, digesta ex auctoribus Verrio Flacco, Antiate (ut quidem idem Verrius maluit dicere, quam Antia), tum ex annalibus pontificum, dein Cincio, Egnatio, Veratio, Fabio Pictore, Licinio Macro, Varrone, Caesare, Tuberone, atque ex omni priscorum historia; proinde ut quisque neotericorum asseveravit, hoc est et Livius et Victor Afer.*”

Ezt az előszót a gyűjtemény összeállítója írta, és az egész korpusz élére szánta, mivel a benne említett időtartam átfogja a három mű által tárgyalt történelmet: *a Iano et Saturno ... usque ad consulatum decimum Constantii*.⁴⁸ Ebből következik, hogy az *Origo gentis Romanae* valójában nem az első mű címe, sokkal inkább az egész gyűjteményé. Momigliano több példát

⁴³ GRUENDEL–PILCHMAYR (1970), XIII–XIV

⁴⁴ MOMIGLIANO (1958), 56

⁴⁵ ADAMIK (2009), 720

⁴⁶ ALBRECHT (2003), 1104

⁴⁷ CAMERON (2010), 624; MOMIGLIANO (1963), 96

⁴⁸ MOMIGLIANO (1958), 57

is felhoz, hogy az *origo* kifejezés általános értelemben vett *történelem* jelentésben is állt.⁴⁹ Mindezek ellenére a szakirodalomban, jobb elnevezés híján, a gyűjtemény első darabját *Origo gentis Romanae* néven emlegetik, ezért a továbbiakban is ezt az elnevezést használom.

Az *Origo*, rövidségéhez képest rendkívül sok forrást idéz. Összesen 18 név szerint említett szerzőtől idéz sokszor az idézett művet is megnevezve, de csak háromtól idéz szó szerint (Vergilius, Ennius, Sallustius, illetve Plautustól egy kifejezést).⁵⁰ Az *Origo* idézeteinek és forrásainak eredetisége kérdéses volt, a kutatásban megjelentek mindkét végletet képviselő vélemények, például Hermann Peter teljesen kihagyta a római történetírók töredékeit összegyűjtő művéből⁵¹, míg Nicholas Horsfall szerint igen értékes anyagokat őrzött meg.⁵² Momigliano is arra az eredményre jutott, hogy az *Origo gentis Romanae* idézetei eredetiek.⁵³

A hagyományos stilisztikán alapuló kutatás annyit biztosan megállapított, hogy a három művet három különböző szerző alkotta, azonban nem tudott konkrét jelölteket állítani arra a kérdésre, hogy ki lehetett az *Origo gentis Romanae* szerzője. Többek között ezen kérdés megválaszolására is alkalmazható a stilometria.

⁴⁹ MOMIGLIANO (1958), 58

⁵⁰ CORNELL (2013) 96-97

⁵¹ PETER (1883)

⁵² BREMMER–HORSFALL (1987); CAMERON (2004), 11

⁵³ MOMIGLIANO (1958), 66-70

4 ALKALMAZOTT ESZKÖZÖK ÉS ELJÁRÁSOK

A következőkben ismertetem a vizsgálat során használt szövegeket, programokat és a módszerek elméleti alapjait. A szoftverek kiválasztásánál elsődleges szempont volt az, hogy rendelkezzenek grafikus felhasználói felülettel, illetve jól dokumentáltak legyenek, mivel mindkettő tulajdonság nagyban megkönnyíti a használatot. Ezenkívül fontos volt, hogy a program multifunkcionális legyen, sok beállítási lehetőséggel rendelkezzen, és többféle eljárás elvégzésére is alkalmas legyen. Végül, hogy az interneten ingyenesen elérhetőek legyenek, mert ez nagyban megkönnyíti a mérések megismétlését, ellenőrzését.

4.1 Szövegek

Az elemzett szövegeket az internetről szereztem be a *The Latin Library*⁵⁴, a *Perseus Digital Library*⁵⁵ és a *Packard Humanities Institute*⁵⁶ weboldaláról. A szövegek nyilvános elérhetősége szintén megkönnyíti a vizsgálatok megismétlését, ellenőrzését.

4.2 Szoftverek

A számítógépes stilometriai vizsgálatok hajnalán még a kutatóknak maguknak kellett programot írni, sőt még a szövegeket is ők gépelték. Később, a számítógépek elterjedésével együtt megjelentek a statisztikai, illetve korpusz- és szövegkezelő programok, amelyek nagyban megkönnyítették a munkát. Az alább ismertetett programok ezt a két programtípust egyesítik magukban, ugyanis egyaránt képesek a szövegek kezelésére (pl. szavak, betűk, mondatok megszámlálása, sorrend felállítása, előfordulási gyakoriság kiszámolása stb.), és az így kapott adatokkal képesek bonyolult matematikai számításokat elvégezni. Így a stilometriai vizsgálat ideje lerövidül, és az eljárás leegyszerűsödik. Ezeknek a programoknak köszönhetően a matematikában nem jártas, a humán tudományok felől közelítő kutatók is könnyen végezhetnek stilometriai elemzéseket.

⁵⁴ <http://www.thelatinlibrary.com/>

⁵⁵ <http://www.perseus.tufts.edu/hopper/>

⁵⁶ <http://latin.packhum.org/>

4.2.1 JGAAP⁵⁷

A JGAAP (*Java Graphical Authorship Attribution Program*) egy, Patrick Juola vezetésével fejlesztett, nyílt forráskódú, Java-alapú, ezért többféle operációs rendszeren futtatható, ingyenes program. Juola a stilometriai elemzést három fázisra osztotta:⁵⁸

1. Egységesítés: A dokumentumokat egységesen egyszerű szöveggé alakítani, és törölni minden zavaró, vagy az adott vizsgálat szempontjából fölösleges jelenséget (pl. mondatvégi írásjelek törlése, nagybetűk átalakítása kisbetűkké, számok eltávolítása stb.)
2. A szöveg felbontása: beállítani, hogy a szövegnek mely paraméterét, paramétereit használjuk fel a statisztikai elemzésre (betűk, szavak, mondatok, n-gramok, funkciószavak stb.)
3. Analízis: a kiválasztott paraméterek elemzése valamilyen osztályozási módszerrel (pl. főkomponens-analízis, klaszteranalízis, legközelebbi szomszéd-analízis, tartóvektor-gép stb.)

A fentiek beállításai széles skálán variálhatóak, ennek eredményeként a programmal több, mint 20 000 féle stilometriai elemzést lehet elvégezni. Ezek persze nem mind alkalmasak a stílusok elkülönítésére.⁵⁹ A szoftver weboldalán nem túl részletes, képekkel illusztrált használati utasítás található, a programban pedig minden opcióhoz tartozik egy rövid magyarázat.

4.2.2 Stylometry with R⁶⁰

Az R egy számítógépes programozási nyelv és szoftverkörnyezet statisztikai számítások céljára. Széles tárházát nyújtja a különféle statisztikai és ábrázolási technikáknak és könnyen bővíthető. Nyílt forráskódú, ingyenes, többféle operációs rendszeren futtatható.⁶¹ A felhasználók által írt és közzétett programokat a CRAN-ből⁶² lehet letölteni, ahol már több, mint 5000 csomag található. Aktív fejlesztői és felhasználói közösség áll mögötte, köztük

⁵⁷ JUOLA (2009a)

⁵⁸ JUOLA–SOFKO–BRENNAN (2006); JUOLA (2009b), 2

⁵⁹ JUOLA (2009b), 3

⁶⁰ R CORE TEAM (2013)

⁶¹ <http://r-projekt.hu/mi-az-r/>

⁶² „The Comprehensive R Archive Network” <http://cran.r-project.org/>

magyar is.⁶³ Az R kiválóan alkalmas szöveges állományok statisztikai elemzésére is (pl. szövegbányászat⁶⁴).

Ezt a tulajdonságot használja ki a *Computational Stylistics Group* (Maciej Eder, Jan Rybicki és Mike Kestemont)⁶⁵ kutatócsoport által fejlesztett *Stylometry with R* nevű programcsomag, amely négy különálló R szkriptből áll:⁶⁶

- *Stylo*: Ez a fő eszköz a csomagban. Főkomponens analízist (PCA), klaszteranalízist (CA), sokdimenziós skálázást (MDS) és bootstrap konszenzusfa-elemzést (BCT) képes végezni.
- *Classify*: különféle osztályozásokat tud elvégezni.
- *Rolling Delta*: többszerzős műveket képes elemezni és elkülöníteni a szakaszokat szerző szerint.
- *Oppose Test*: két szövegtörzs közötti különbséget elemzi.

A program grafikus felhasználói felülettel rendelkezik, amelyen minden beállítás elérhető. Alapból támogatja a latin nyelvet. Az eredményeket többféle képformátumban képes menteni. A csomag weboldalán képekkel illusztrált, részletes használati utasítás található, illetve a programban magyarázatok tartoznak az opciókhoz. A vizsgálatokban a *stylo* szkriptet használtam.

4.3 Statisztikai eljárások

Ebben a fejezetben a vizsgálatok során használt statisztikai eljárásokat ismertetem röviden és felületesen. A kifejezetten stilometriai elemzések végrehajtásának céljából készül szoftverek megjelenése lehetővé tette, hogy a matematikában és a statisztikában nem járatos kutatók is végezzenek stilometriai méréseket, a bonyolult matematikai képletek és egyenletek ismerete nélkül. Ezért jelen dolgozat nem tartalmazza az alábbi eljárások részletes matematikai leírásait.

⁶³ <http://r-projekt.hu/>

⁶⁴ FEINERER (2008); MEYER–HORNIK–FEINERER (2008)

⁶⁵ <https://sites.google.com/site/computationalstylistics/>

⁶⁶ EDER–KESTEMONT–RYBICKI (2013)

4.3.1 Főkomponens analízis

A főkomponens analízis (*Principal Component Analysis* – PCA) a legrégebbi és legismertebb többváltozós statisztikai eljárás. 1901-ben találta fel Karl Pearson,⁶⁷ majd 1933-ban fejlesztette tovább Harold Hotelling.⁶⁸ Az eljárás a dimenziószám csökkentésére használható. Hogy ezt elérje a főkomponens analízis az eredeti változókat átalakítja át új, nem korreláló változókká, az úgynevezett főkomponensekké. A folyamat során minimális az adatvesztés, az első néhány főkomponens egyesíti magában az eredeti változók információit és segítségével eredetileg rejtett hatások figyelhetőek meg. Az első két főkomponens grafikusán megjeleníthető koordináta-rendszerként, egyfajta árnyékaként a sokdimenziós térnek. Alkalmazási területe igen széleskörű: pszichológia, meteorológia, biológia, mérnöki tudományok, stb.⁶⁹

4.3.2 Klaszteranalízis

A klaszterelemzés (*Cluster Analysis* – CA), a főkomponens analízishez hasonlóan, egy dimenziócsökkentő eljárás, melynek során az adatokat csoportokba (klaszterekbe) soroljuk a köztük lévő hasonlóság és különbség alapján.⁷⁰ Az elemzés eredménye egy dendrogram (mivel hasonlít egy fához), ami vizuálisan is megjeleníti a klasztereket és a távolságokat. A klaszteranalízis először minden elemzett tagot külön klaszternek tekint, majd a legközelebbiket (tehát a leghasonlóbbakat) egy-egy klaszterbe csoportosítja, aztán az így létrejött klasztereket is csoportosítja az egymás közti távolságaik alapján. Ezt addig ismétli, amíg minden tag egy nagy klaszterbe kerül, és ez alapján megrajzolja a dendrogramot.⁷¹

⁶⁷ PEARSON (1901)

⁶⁸ BINONGO-SMITH (1999), 445; HOTELLING (1933)

⁶⁹ BINONGO-SMITH (1999) 445

⁷⁰ SZÉKELY-BARNA (2002), 109

⁷¹ JOHNSON-WICHERN (2007), 680

5 VIZSGÁLATOK

A vizsgálatok elvégzése előtt az összes kiválasztott szövegből töröltem minden szószerinti idézetet. A szövegek formázását egységesítettem: eltávolítottam minden központosítást, bekezdésjelet és számot. A szavakat kisbetűsre formáztam és az összes v betűt u betűre cseréltem. A végeredményben nem maradt más csak szavak és szóközök. A szövegek .txt kiterjesztésű szövegfájlokban voltak tárolva, UTF-8 kódolással.

A szövegekkel nem lehet számításokat végezni, ezért valamilyen tulajdonságaikat számszerűsíteni kell. Jelen esetben ez a szavak gyakorisága. A program megszámlálja az egyes szavak előfordulását a vizsgálatban részt vevő szövegekben és az így kapott adatokkal végzi el a statisztikai műveletet.

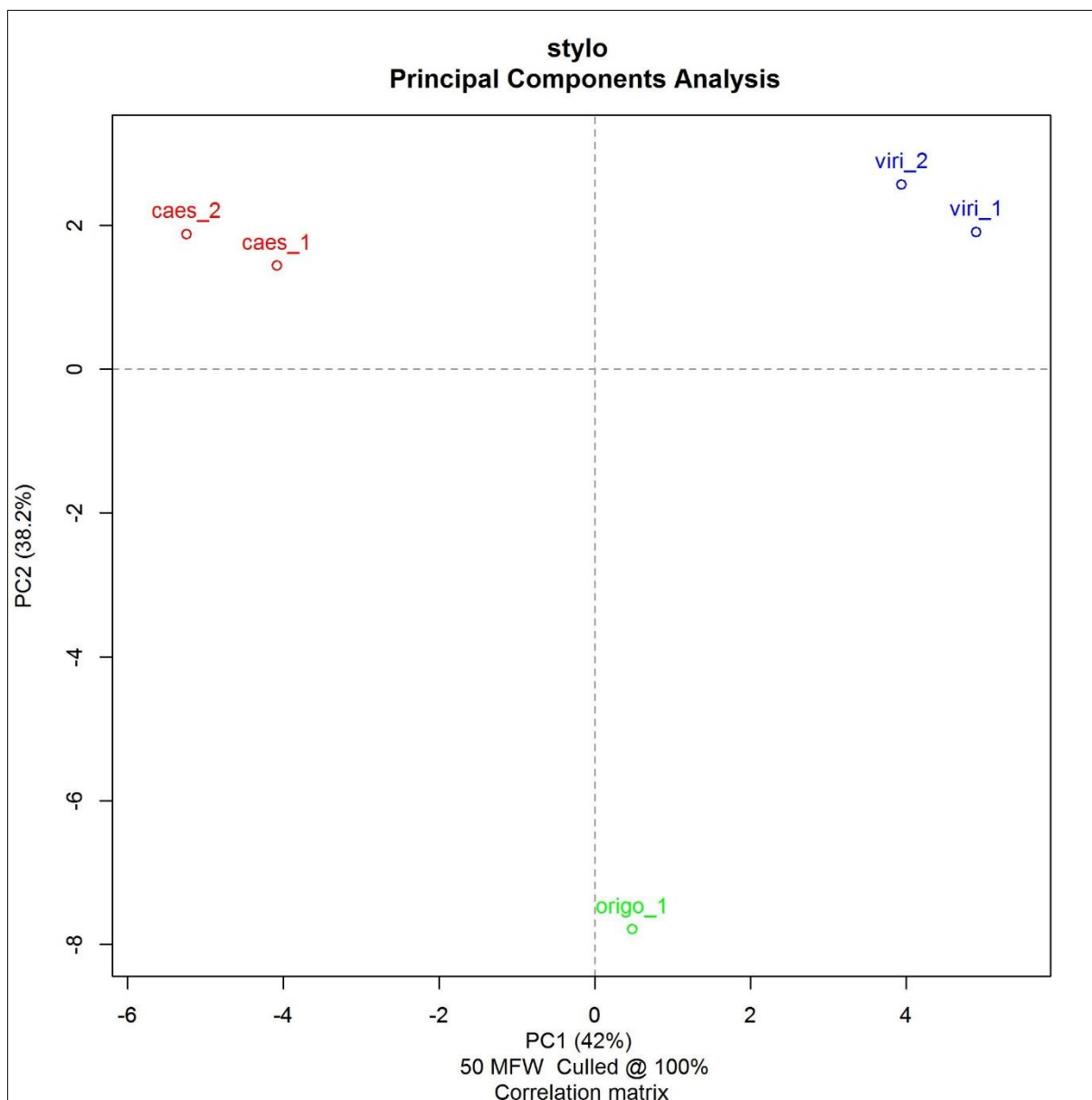
5.1 Első vizsgálat

Az első vizsgálat célja eldönteni, hogy a Sextus Aurelius Victor neve alatt fennmaradt korpusz egységes-e, azaz hogy ugyanazon szerző tollából származik-e a három mű. A vizsgálatot főkomponens-analízis segítségével végeztem a *Stylometry with R* programcsomag *stylo* szkriptjével. Az elemzésben a következő művek szerepeltek:

Név	Fájlnév	Szavak száma
<i>Origo gentis Romanae</i>	origo.txt	4229
<i>De Viris Illustribus</i>	viri.txt	9277
<i>De Caesaribus</i>	caes.txt	10555

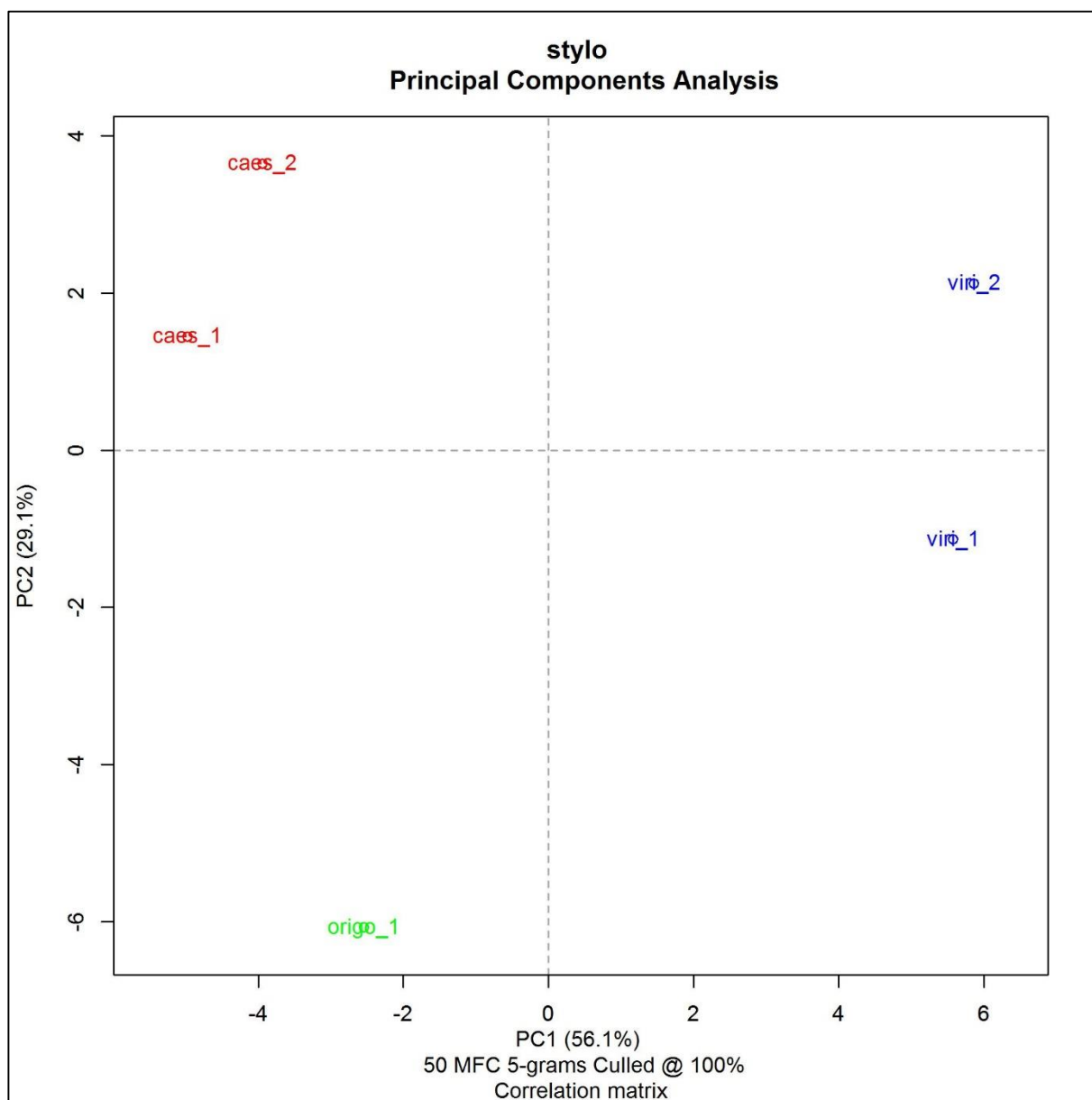
1. táblázat: az első vizsgálatban szereplő művek

Az *Origo* és a két másik szöveg hosszúsága közti eltérést mintavételezéssel küszöböltem ki, ezért minden szöveg 4200 szavas darabokra (mintákra) lett osztva. A leggyakoribb 50 szó előfordulási gyakorisága képezte a számítások alapját. A főkomponens analízis eredménye az 1. ábrán látható.



1. ábra

Az első (PC1) és a második főkomponens (PC2) együtt az összes variancia 80,2%-át fedile és a koordináta-rendszerében tisztán látszik a három mű elkülönülése. Ez az eredmény (a hagyományos vizsgálatokkal összhangban) alátámasztja azt az elméletet, miszerint a három művet három különböző szerző írta.



2. ábra

Ezután a szöveg 5-gramjait vettem górcső alá főkomponens elemzéssel. Az eredmény a 2. ábrán látható. Az ötven leggyakoribb 5-gram alapján a három mű elkülönülése tökéletesen megfigyelhető. Az eredmények magukért beszélnek, ezért nem tartom szükségesnek további elemzések elvégzését ezzel a kérdéssel kapcsolatban (ti. a három mű közös szerzőtől való származása).

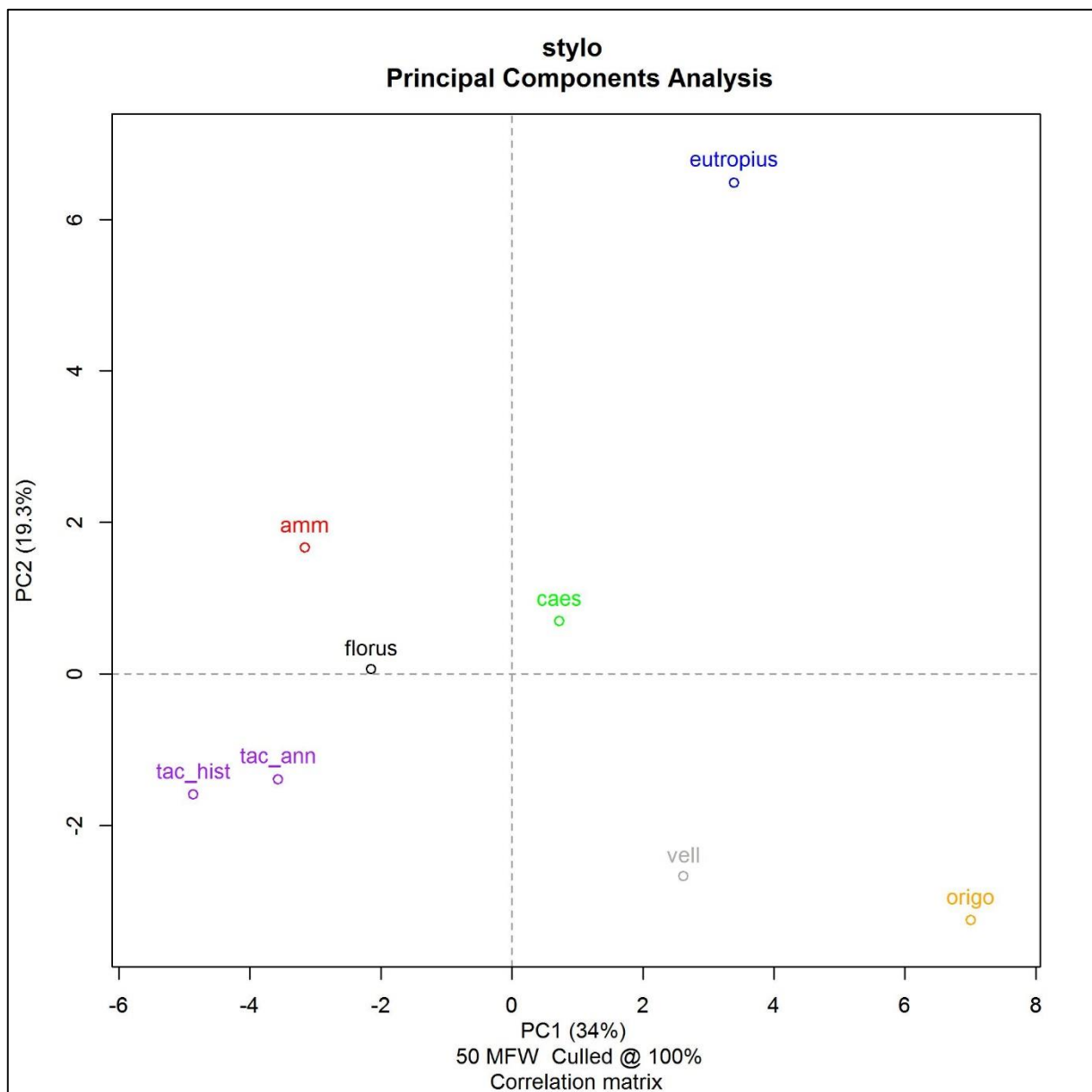
5.2 Második vizsgálat

A második vizsgálat célja az *Origo gentis Romanae* szerzőjének meghatározása stilometria segítségével. Mivel nincsenek konkrét jelöltek a szerzői szerepre, ezért a császárkori pogány történetírókat vontam vizsgálat alá. Kizártam a keresztény szerzőket, mivel a mű egy szóval sem említi a kereszténységet, ami pogány eredetre utal.⁷² Azoknál a szerzőknél, akiknek több könyvnyi írásuk is fennmaradt, válogatást készítettem, lehetőleg úgy, hogy a műveik elejéből, közepéből és végéből is legyen egy-egy könyvnyi szöveg. A vizsgálatot főkomponens analízis és klaszteranalízis segítségével végeztem a *Stylometry with R* programcsomag *stylo* szkriptjével. A vizsgálatban a következő művek vettek részt:

Név	Fájlnév	Szavak száma
Ammianus Marcellinus: <i>Historiae</i> 14., 20., 25., 31. könyv	amm.txt	29340
<i>Epitome de Caesaribus</i>	caes.txt	8576
Eutropius: <i>Breviarium</i>	eutropius.txt	18544
Florus: <i>Epitome</i>	florus.txt	26386
<i>Origo gentis Romanae</i>	origo.txt	4229
Tacitus: <i>Annales</i> 1., 12., 15. könyv	tac.ann.txt	27300
Tacitus: <i>Historiae</i> 1., 3., 5. könyv	tac.hist.txt	26438
Velleius Paterculus: <i>Historiae Romanae</i> 1., 2. könyv	vell.txt	26167

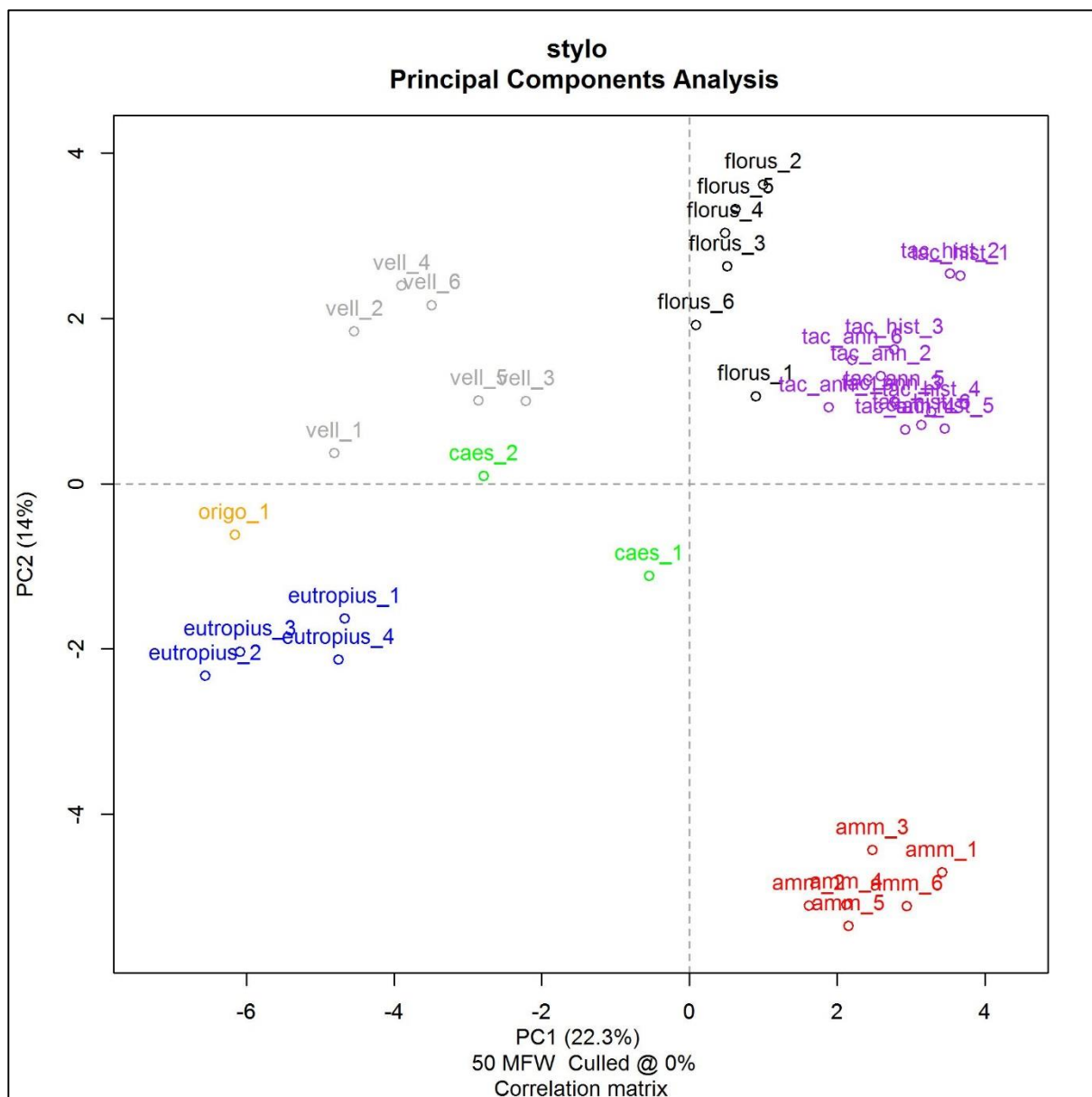
2. táblázat: a második vizsgálatban szereplő művek

⁷² CAMERON (2010), 624; MOMIGLIANO (1963), 96



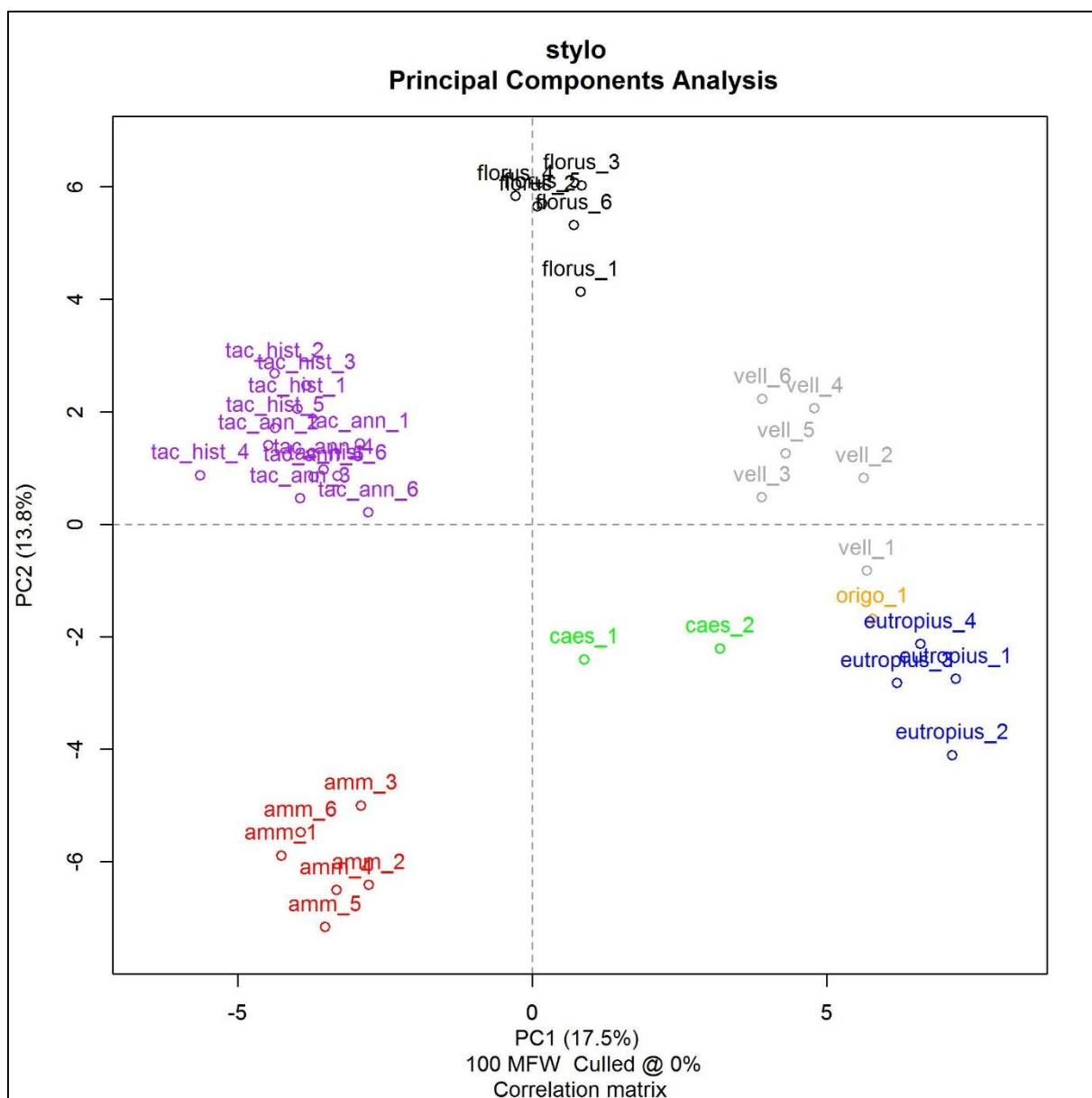
3. ábra

Először főkomponens-analízist végeztem a szövegeken mintavételezés nélkül. A leggyakoribb 50 szó képezte az elemzés alapját. Az eredmény a 2. ábrán látható. Az első két főkomponens (PC1 és PC2) együtt az összes variancia 53,3%-át adja vissza. A koordináta-rendszerben egyértelműen látszik a szerzők közti elkülönülés. Tacitus művei egymáshoz közel helyezkednek el, hozzájuk legközelebb pedig (kicsit meglepő módon) Florus és (nem meglepő módon) Ammianus Marcellinus. Az *Origo gentis Romanae* egyik szerzőhöz sincs közel szignifikáns módon, az elemzés szerint a vizsgáltak közül Velleius Paterculushoz áll legközelebb a stílusa.



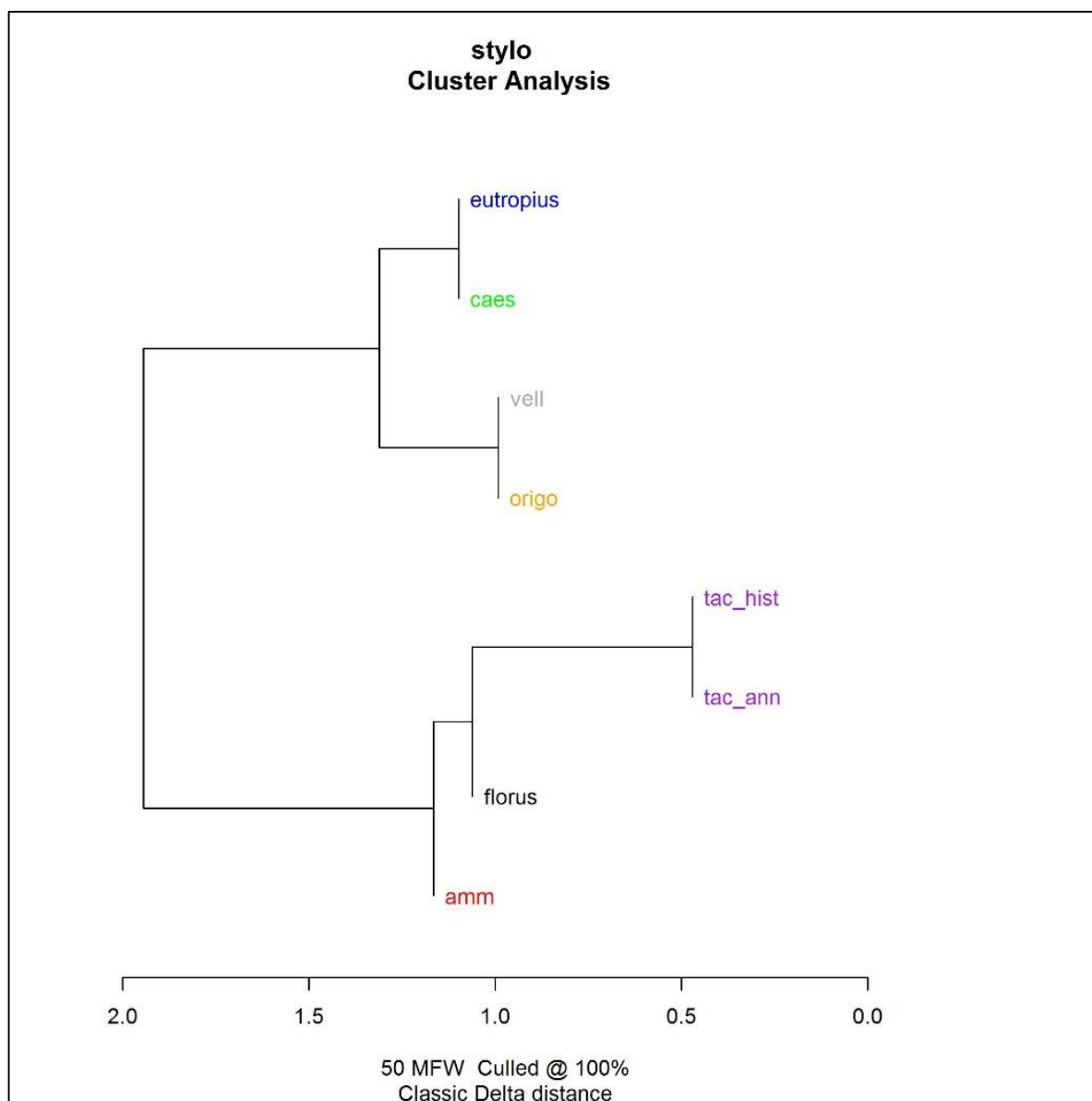
4. ábra

Ezután megismételtem a kísérletet, de 4200 szavas mintavételezéssel. Az eredmény a 4. ábrán látható. 4200 szavas mintavételezéssel az első két főkomponens már csak az összes variancia 34,4%-át fedi le. Ennek ellenére szépen kirajzolódnak a határok, elkülönülnek a szerzők egymástól. Tacitus és Florus feltűnően közeli helyeket foglalnak el, de a második főkomponens jobban elkülöníti tőlük Ammianus Marcellinust. Az *Origo gentis Romanae* most is Velleius Paterculushoz áll a legközelebb, de nem eléggé, ahhoz, hogy szerzőségről beszéljünk. Szintén feltűnő az *Epitome de Caesaribus* két részének egymáshoz viszonyított távolságának túlzott nagysága: ez akár további vizsgálatok tárgyát képezheti.



5. ábra

Ezután ötvenről százra emeltem a leggyakrabban használt szavak számát. Ez a változtatás a csoportok élesebb elkülönülését eredményezte, bár az első két főkomponens veszített a teljes varianciából, amint az az 5. ábrán is látható. Megfigyelhető, hogy Florus teljesen elkülönül Tacitustól. Az *Origo gentis Romanae* még mindig Velleius Paterculus közelében található, főleg annak első mintájával mutat hasonlóságot, ami lényegében a *Historiae Romanae* teljes első könyvét takarja. Szintén a közelében található Eutropius, de az *Epitome de Caesaribus* távolabb került.



A klaszteranalízis eredménye hasonló csoportosítást mutat, mint a főkomponens analízisé (6. ábra). A tacitusi szövegek itt is egy csoportba kerültek, és az *Origo* most is Velleius Paterculushoz áll a legközelebb, egy kicsit távolabb áll tőle Eutropius és az *Epitome de Caesaribus*.

Ezekkel az eredményekkel összhangban van a JGAAP által adott távolságlista. A 3. táblázat az *Origo* és a többi mű 50 leggyakoribb szaván végzett számítás eredményét mutatja. A sorrend eleje hasonló a főkomponens analízis eredményéhez. Az *Origo gentis Romanae* stílusához legközelebb álló szerző a vizsgáltak közül Velleius Paterculus. A második és harmadik helyen az Epitome és Eutropius áll, csakúgy, mint az előző méréseknél.

1.	Velleius Paterculus	60.75873770506712
2.	Epitome de Caesaribus	68.54931777914739
3.	Eutropius	74.74875864242489
4.	Florus	83.75697710004718
5.	Tacitus: Annales	85.05675737911345
6.	Ammianus Marcellinus	86.66205311851587
7.	Tacitus: Historiae	93.38006609003132

3. táblázat

6 ÖSSZEGZÉS

Az elvégzett vizsgálatok közül egyik sem hozott pozitív eredményt. Az első vizsgálat eredménye alátámasztja az eddigi ismereteinket a Sextus Aurelius Victor neve alatt fennmaradt korpusz egységességével (vagy inkább az egységesség hiányával) kapcsolatban, miszerint a három művet három különböző szerző írta. A második vizsgálat érdekes eredményeket hozott: többféle beállítással is elvégezve a méréseket kiderült, hogy az *Origo gentis Romanae* stílusához Velleius Paterculus stílusa áll a legközelebb a vizsgálatban elemzett művek közül. Az, hogy az *Origo* szerzője Velleius Paterculus lenne, nem zárható ki teljesen, legalábbis kronológiailag lehetséges, mivel az *Origoban* található idézetek legkésőbbi forrása Verrius Flaccus, aki kortársa volt Velleius Paterculusnak. Az eredmény szerint a második esélyes jelölt, Eutropius időben közelebb élt az *Origo* feltételezett keletkezési idejéhez. Az *Origo* és a két szerző kapcsolata további vizsgálatok tárgyát képezheti, egyrészt további stilometriai eljárások bevonásával, másrészt hagyományos stilisztikai elemzéssel.

7 IRODALOMJEGYZÉK:

- ADAMIC, Lada A. – HUBERMAN, Bernardo A.: *Zipf's law and the Internet*. Glottometrics, 3 (1). 2002, 143–150
- ADAMIK Tamás: *Római irodalom: a kezdetektől a Nyugatrómai Birodalom bukásáig*. Pozsony. 2009
- ALBRECHT, Michael von: *A római irodalom története*. Budapest. 2003
- AXTELL, Robert L.: *Zipf Distribution of U.S. Firm Sizes*. Science, 293 (5536). 2001, 1818–1820
- BACKER, Eric – KRANENBURG, Peter van: *On musical stylometry—a pattern recognition approach*. Pattern Recognition Letters, 26, 2005, 299–309
- BERGER, Helmut – MERKL, Dieter: *A comparison of text-categorization methods applied to n-gram frequency statistics*. In *AI 2004: Advances in Artificial Intelligence* (pp. 998–1003). Springer. 2005
- BHARGAVA, Mudit – MEHNDIRATTA, Pulkit – ASAWA, Krishna: *Stylometric Analysis for Authorship Attribution on Twitter*. in: *Big Data Analytics (Lecture Notes in Computer Science)*, 2013, 37-47
- BINONGO, J. N. G. – SMITH, M. W. A.: *The application of principal component analysis to stylometry*. Literary and Linguistic Computing, 14 (4). 1999, 445–466
- BREMMER, Jan – HORSFALL, Nicholas: *Roman Myth and Mythography*. University of London Institute of Classical Studies. 1987
- BRENNAN, Michael – AFROZ, Sadia – GREENSTADT, Rachel: *Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity*. ACM Transactions on Information and System Security, 15 (3), 2012,
- BRENNAN, Michael – GREENSTADT, Rachel: *Practical Attacks Against Authorship Recognition Techniques*. Proceedings of the Twenty-First Innovative Applications of Artificial Intelligence Conference, 2009, 60-65
- BURNET, John: *Platonism*. University of California Press. 1928

- BURROWS, John F.: *“An ocean where each kind...”: Statistical analysis and some major determinants of literary style*. *Computers and the Humanities*, 23 (4-5). 1989, 309–321
- BURROWS, John F.: *Computers and the study of literature*. *Computers and Written Texts*. 1992a, 167–204
- BURROWS, John F.: *Not unless you ask nicely: The interpretative nexus between analysis and information*. *Literary and Linguistic Computing*, 7 (2). 1992b, 91–109
- BURROWS, John F.: *Word-patterns and story-shapes: The statistical analysis of narrative style*. *Literary and Linguistic Computing*, 2 (2). 1987, 61–70
- CALISKAN-ISLAM Aylin: *Supervised Source Code Authorship Attribution*. IEEE Symposium on Security & Privacy, 2013.
- CALIX, K. – CONNORS, M. – LEVY, D. – MANZAR, H. – MCCABE, G. – WESTCOTT, S.: *Stylometry for E-mail Author Identification and Authentication*. Proceedings of CSIS Research Day, Pace University, 2008.
- CAMERON, Alan: *Greek Mythography in the Roman World*. Oxford University Press. 2004
- CAMERON, Alan: *The last pagans of Rome*. Oxford University Press. 2010
- CAMPBELL, Lewis: *The Sophistes and Politicus of Plato*. Clarendon Press. 1867
- CAVNAR, William B. – TRENKLE, John M.: *N-grambased text categorization*. In *In Proc. of SDAIR-94, 3rd Annual Symposium on Document Analysis and Information Retrieval* (pp. 161–175). 1994
- CORNELL, T. J.: *The Fragments of the Roman Historians*. Oxford University Press. 2013
- CZALLER, László: *A Zipf-törvény érvényesülése a világ országaiban*. *Területi Statisztika*, (05). 2012, 461–478
- DE MORGAN, Sophia Elizabeth – DE MORGAN, Augustus: *Memoir of Augustus De Morgan: With Selections from His Letters*. Cambridge University Press. 2010
- DELCOURT, Christian: *Stylometry*. *Revue Belge de Philologie et D’histoire*, 80 (3). 2002, 979–1002

- EDER, Maciej – KESTEMONT, Mike – RYBICKI, Jan: *Stylometry with R: a suite of tools*. In *Digital Humanities 2013: Conference Abstracts* (pp. 487–89). Lincoln, NE. 2013
- EDER, Maciej: *Style-Markers in Authorship Attribution A Cross-Language Study of the Authorial Fingerprint*. *Studies in Polish Linguistics*, 2011 (Volume 6, Issue 1). 2011, 99–114
- FEINERER, Ingo: *A text mining framework in R and its applications*. WU Vienna University of Economics and Business. 2008
- FORSYTH, Richard S – HOLMES, David E. – TSE, Emily K.: *Cicero, Sigonio and Burrows: Investigating the Authenticity of the "Consolatio"*. *Literary and Linguistic Computing*, 14 (3), 1999, 1-26.
- FRISCHER, Bernard – GUTHRIE, Donald – TSE, Emily – TWEEDIE, FJ: *"Sentence" Length and Word-Type at "Sentence" Beginning and End. Reliable Authorship Discriminators for Latin Prose? New Studies on the Authorship of the Historia Augusta*. *Research in Humanities Computing*, 5. 1996, 110–142
- GERWIG, George William: *On the Decrease of Predication and of Sentence Weight in English Prose*. *University Studies*, 2 (1). 1894, 17
- GOSWAMI, Sumit – SARKAR, Sudeshna – RUSTAGI, Mayur: *Stylometric Analysis of Bloggers' Age and Gender*. *Proceedings of the Third International ICWSM Conference*, 2009, 214-217
- GRUENDEL, Roland – PILCHMAYR, Franz (Eds.): *Sexti Aurelii Victoris liber de Caesaribus : praecedunt Origo gentis Romanae et Liber de viris illustribus urbis Romae, subsequitur Epitome de Caesaribus*. Leipzig. 1970
- GRZYBEK, Peter: *The Emergence of Stylometry: Prolegomena to the History of Term and Concept*. In Katalin Kroó & Peeter Torop (Eds.), *Text within Text - Culture within Culture* (pp. 58–75). L'Harmattan. 2014
- GURNEY, Penelope J. – GURNEY, Lyman W.: *Authorship Attribution of the Scriptores Historiae Augustae*. *Literary and Linguistic Computing*, 13 (3), 1998, 119-131

- HOCKEY, Susan: *Electronic Texts in the Humanities*. Oxford University Press. 2000
- HOLMES, David I. – FORSYTH, Richard S.: *The Federalist Revisited: New Directions in Authorship Attribution*. *Literary and Linguistic Computing*, 10 (2). 1995, 111–127
- HOLMES, David I.: *Authorship Attribution*. *Computers and the Humanities*, 28 (2). 1994, 87–106
- HOLMES, David I.: *The Evolution of Stylometry in Humanities Scholarship*. *Literary and Linguistic Computing*, 13 (3). 1998, 111–117
- HOTELLING, H.: *Analysis of a complex of statistical variables into principal components*. *Journal of Educational Psychology*, 24 (6). 1933, 417–441
- HOUVARDAS, John – STAMATATOS, Efstathios: *N-gram feature selection for authorship identification*. In *Artificial Intelligence: Methodology, Systems, and Applications* (pp. 77–86). Springer. 2006
- JANSON, Tore: *The Problems of Measuring Sentence-Length in Classical Texts*. *Studia Linguistica*, 18 (1). 1964, 26–36
- JOHNSON, Richard A. – WICHERN, Dean W.: *Applied Multivariate Statistical Analysis* (6 edition.). Upper Saddle River, N.J. 2007
- JUOLA, Patrick – SOFKO, John – BRENNAN, Patrick: *A Prototype for Authorship Attribution Studies*. *Literary and Linguistic Computing*, 21 (2). 2006, 169–178
- JUOLA, Patrick: *Authorship Attribution*. *Foundations and Trends® in Information Retrieval*, 1 (3). 2008, 233–334
- JUOLA, Patrick: *JGAAP, a Java-based, modular, program for textual analysis, text categorization, and authorship attribution*. 2009a
- JUOLA, Patrick: *JGAAP: A System for Comparative Evaluation of Authorship Attribution*. *Journal of the Chicago Colloquium on Digital Humanities and Computer Science*, 1 (1). 2009b

- KESTEMONT, M. – MOENS, S. – DEPLOIGE, J., (2013). *Stylometry and the complex authorship in Hildegard of Bingen's oeuvre*. Digital Humanities 2013. Conference Abstracts, 2013, 255–258
- KESTEMONT, Mike: *Stylometry for Medieval Authorship Studies: An Application to Rhyme Words*. Digital Philology: A Journal of Medieval Cultures, 1 (1), 2012, 42-72
- LI, Wentian: *Zipf's law everywhere*. Glottometrics, 5. 2002, 14–21
- LÓPEZ-ESCOBEDO, Fernanda – MÉNDEZ-CRUZ, Carlos-Francisco – SIERRA, Gerardo – SOLÓRZANO-SOTO, Julián: *Analysis of Stylometric Variables in Long and Short Texts*. Procedia - Social and Behavioral Sciences, 95. 2013, 604–611
- LUTOSŁAWSKI, Wincenty: *On Stylometry. Abstract of a paper read at the Oxford Philological Society on May 21st by Dr. W. Lutosławski, of Drozdowo, near Lomza, Poland*. Classical Review. 1897a, 11–284
- LUTOSŁAWSKI, Wincenty: *Principes de stylométrie*. Revue Des Études Grecques. 1898, 41–61
- LUTOSŁAWSKI, Wincenty: *The origin and growth of Plato's logic*. Georg Olms Verlag. 1897b
- MARRIOTT, Ian: *The Authorship of the Historia Augusta: Two Computer Studies*. The Journal of Roman Studies, 69. 1979, 65–77
- MASCOL, Conrad: *Curves of pauline and pseudo-pauline style I*. Unitarian Review, 30. 1888a, 453–460
- MASCOL, Conrad: *Curves of pauline and pseudo-pauline style II*. Unitarian Review, 30. 1888b, 539–546
- MCDONALD, Andrew – AFROZ, Sadia – CALISKAN-ISLAM, Aylin – STOLERMAN, Ariel – GREENSTADT Rachel: *Use Fewer Instances of the Letter "i": Toward Writing Style Anonymization*. The 12th Privacy Enhancing Technologies Symposium, 2012.
- MENDENHALL, T. C.: *The characteristic curves of composition*. Science, ns-9 (214S). 1887, 237–246

- MEYER, David – HORNIK, Kurt – FEINERER, Ingo: *Text mining infrastructure in R*. Journal of Statistical Software, 25 (5). 2008, 1–54
- MICHAELSON, S. – MORTON, A. Q.: *The New Stylometry: A One-Word Test of Authorship for Greek Writers*. The Classical Quarterly (New Series), 22 (01). 1972, 89–102
- MOMIGLIANO, Arnaldo: *Some Observations on the “Origo Gentis Romanae.”* The Journal of Roman Studies, 48 (1/2). 1958, 56–73
- MOMIGLIANO, Arnaldo: *The conflict between paganism and christianity in the fourth century: essays*. Oxford, Clarendon P. 1963
- MORITZ, Robert E.: *On a Quantitative Relation Governing Certain Linguistic Phenomena*. Modern Language Notes, 24 (8). 1909, 234–241
- MORITZ, Robert E.: *On the Variation and Functional Relation of Certain Sentence-Constants in Standard Literature*. University Studies, 8 (11). 1903, 229–253
- MOSTELLER, Frederick – WALLACE, David: *Applied Bayesian and Classical Inference: The Case of The Federalist Papers* (2nd ed. 1984. Softcover reprint of the original 2nd ed. 1984 edition.). New York, NY. 2011
- MOSTELLER, Frederick – WALLACE, David: *Inference and disputed authorship: The Federalist*. Addison-Wesley. 1964
- PAVLYSHENKO, Bohdan: *Genetic Optimization of Keywords Subset in the Classification Analysis of Texts Authorship*. arXiv:1211.3402 [cs]. 2012
- PAWŁOWSKI, Adam: *Wincenty Lutosławski-a forgotten father of stylometry*. Glottometrics, 8. 2004, 83–89
- PEARSON, Karl: *On lines and planes of closest fit to systems of points in space*. Philosophical Magazine Series 6, 2 (11). 1901, 559–572
- PETER, H. W. G.: *Historicorum Romanorum fragmenta*. In aedibvs B.G. Teubneri. 1883
- PETRUSZEWCZ, Micheline: *Les chaînes de Markov dans le domaine linguistique* (Vol. 19). Genova. 1981

- QI, Hanchao –TAEB, Armeen –HUGHES, Shannon M.: *Visual stylometry using background selection and wavelet-HMT-based Fisher information distances for attribution and dating of impressionist paintings*. *Signal Processing* 93, 2013, 541–553
- R CORE TEAM: *R: A Language and Environment for Statistical Computing*. Vienna, Austria. 2013
- RUDMAN, Joseph: *Authorship Attribution: Statistical and Computational Methods*. In Keith Brown (Ed.), *Encyclopedia of Language and Linguistics*. Elsevier. 2006
- RUDMAN, Joseph: *The State of Authorship Attribution Studies: Some Problems and Solutions*. *Computers and the Humanities*, 31 (4). 1997, 351–365
- SANSONE, David: *The Computer and the Historia Augusta: A Note on Marriott*. *The Journal of Roman Studies*, 80. 1990, 174–177
- SHERMAN, Lucius Adelno: *Analytics of literature: A manual for the objective study of English prose and poetry*. Ginn. 1893
- SHERMAN, Lucius Adelno: *Some observations upon the sentence-length in English prose*. 1888
- SMITH, J. A. – KELLY, C.: *Stylistic Constancy and Change across Literary Corpora: Using Measures of Lexical Richness to Date Works*. *Computers and the Humanities*, 36 (4). 2002, 411–430
- SMITH, M. Wilfrid A.: *Attribution by statistics: a critique of four recent studies*. *Revue, Informatique et Statistique Dans Les Sciences Humaines*, 26. 1990, 233–251
- STAMATATOS, Efstathios: *A Survey of Modern Authorship Attribution Methods*. *J. Am. Soc. Inf. Sci. Technol.*, 60 (3). 2009, 538–556
- STAMATATOS, Efstathios: *On the Robustness of Authorship Attribution Based on Character N-Gram Features*. *Journal of Law & Policy*, 21 (2). 2013, 421–439

- SUEN, Ching Y.: *n-Gram Statistics for Natural Language Understanding and Text Processing*.
IEEE Transactions on Pattern Analysis and Machine Intelligence, PAMI-1 (2). 1979,
164–172
- SZÉKELY, Mária – BARNA, Ildikó: *Túlélőkészlet az SPSS-hez*. Budapest. 2002
- TSE, Emily – TWEEDIE, Fiona J. – FRISCHER, Bernard D.: *Unravelling the Purple Thread. Function Word Variability and the Scriptorum Historiae Augustae*. Literary and Linguistic Computing, 13 (3), 1998, 142-149
- TWEEDIE, Fiona J. – HOLMES, David I. – CORNS, Thomas N.: *The Provenance of De Doctrina Christiana, attributed to John Milton. A Statistical Investigation*. Literary and Linguistic Computing, 13 (2), 1998, 77-87
- TWEEDIE, Fiona J. – SINGH, Sameer – HOLMES, David I.: *An introduction to neural networks in stylometry*. Research in Humanities Computing, 5. 1996a, 249–263
- TWEEDIE, Fiona J. – SINGH, Sameer – HOLMES, David I.: *Neural network applications in stylometry: The Federalist Papers*. Computers and the Humanities, 30 (1). 1996b, 1–10
- WALL, Wilson J.: *Forensic Science in Court*. 2009, Hoboken, NJ 2009
- WEISBERG, S. – BINGHAM, C. – SELLEW, P. – ADERINKOMI, T.: *Stylometric Analysis of Ancient Greek Texts*. 2011, (<http://www.thompsonaderinkomi.com/wp-content/uploads/2011/08/A-Stylometric-Analysis-of-Ancient-Greek-Texts.pdf>) (2014. 08. 10.)
- ZIPF, George Kingsley: *Selected studies of the principle of relative frequency in language*. 1932