DOI: 10.31915/NWS.2021.6

Szöveghasonlósági vizsgálatok automatizálása

Kiss Margit
BTK Irodalomtudományi Intézet, XIX. századi osztály
kiss.margit@abtk.hu

Palkó Gábor
ELTE BTK TI Digitális Bölcsészeti Tanszék, Digitális Örökség Nemzeti Laboratórium, BTK Irodalomtudományi Intézet
palko.gabor@abtk.hu

Szakács Béla Benedek
BME VIK, MA hallgató
skyraiderfighter@gmail.com

A stilometriai elemzések a digitális bölcsészeti kutatásokban központi szerepet töltenek be: a szövegalkotás egyéni jellemzőinek a feltárásában segítenek úgy, hogy a meghatározó jegyek kiszűrése révén tulajdonságok tízezrei alapján hasonlíthatók össze a kérdéses szövegek. E számítógéppel végzett összetett statisztikai elemzőeljárások elvégzése közel sem egyszerű művelet. Különféle elemzőszoftverek (Websty, JGAAP, Stylene stb.) érhetők ma már el, azonban azt a feladatot, amely a megfelelő elemzési eredmények eléréséhez szükséges, a felhasználónak komoly nyelvészeti-statisztikai háttértudással és empirikus kísérletek révén kell elvégeznie.

Tanulmányunkban felvázoljuk a stilometriai elemzés különféle alkalmazási területeit és eddigi eredményeit. Mint hazai innováció, bemutatunk egy a BME Méréstechnika és Információs Rendszerek Tanszék fejlesztésében, az ELTE Digitális Bölcsészet Tanszék, illetve a BTK Irodalomtudományi Intézet együttműködésében kialakított webszolgáltatást, a *Shtylo*-t,[1] amelyet a Digitális örökség nemzeti laboratórium üzemeltet. Az alkalmazás korszerű újításai révén több szempontból is megkönnyíti a stilometriai elemzést végzők munkáját. Újdonsága, hogy webes környezetbe ágyazva a felhasználó mentesül a szoftver telepítése és a számítási feladatok elvégzéséhez szükséges, igen jelentős gépi erőforrás biztosítása alól. Egyedülálló módon, a felhasználótól elvárt tudásigényes feladatot automatikus paraméterezéssel támogatja: a heurisztikus paraméterbeállítás elemzi a vizsgálandó korpuszokat, és javaslatot tesz az adekvát beállításokra; ezen túlmenően a már meglévő paraméterek több elemzés automatikus futtatásával még optimálisabbá tehetők a pontosabb eredmények elérése érdekében.

Ma a stilometria megnevezés a stílus statisztikai alapú vizsgálatát jelenti, amely különféle tudományterületeket integrál: helye van benne a nyelvészetnek, irodalomtudománynak, filológiának, stilisztikának, statisztikának és az informatikának. A szó és a diszciplína megalkotója Wincenty Lutosławsky, aki a módszert Platón dialógusainak a

1 https://github.com/szakacsb/shtylo

kronologizálásához alkalmazta.[2] Feladata az adott szerzőre jellemző megkülönböztető jegyeknek, az ún. szerzői ujjlenyomatnak a feltárása. Tipikus alkalmazási területe a szerzői idiolektus vizsgálata, az anonim vagy vitatott szerzőségű szövegek azonosítása,[3] az egyéni nyelvezet alakulása, formálódása, korszakolás,[4] csoporthoz tartozás vizsgálata,[5] műfaji jelleg, de akár a nyelvi szempontból megmutatkozó hatás elemzése is.[6] Az irodalom- és nyelvtudományi kutatásokon kívül törvényszéki eljárások során alkalmazzák,[7] az utóbbi időben azonban más tudományterületeken is hozott eredményeket.[8]

Már a 19. században is felfigyeltek a szerzői jelleg statisztikai alapú megközelítésére,[9] azonban a jelentős áttörést mégis a 20. század hozta el az informatika térhódításával, amelynek révén lehetővé vált a szövegek stilisztikai jegyeinek a mérése, eredmények összevethetősége, értékelése.[10]

A mai modern stilometria esetében, a számítógép azokon a területeken nyújt segítséget, amelyeken több szempontú, átfogó összehasonlítások szükségeltetnek: a nyelvi modellek vizsgálatakor a szövegalkotás, kifejezésmód egyéni jellemzőinek a feltárásában úgy, hogy a szerzőt meghatározó jegyek kiszűrésére törekszik. Mivel a kézi összehasonlítás csak rövidebb terjedelmű szövegeken egy-egy jellemző esetében lehet sikeres, digitális bölcsészeti eszközökkel jellemzően nagy terjedelmű szövegkorpuszok vizsgálatára nyílik már lehetőség. Ehhez a gépi feldolgozásra alkalmas jegyek kigyűjtését és összehasonlítását kell elvégeznünk a számítógép segítségével. Mindaz, amit a számítógépes nyelvészet egy adott szövegről meg tud határozni, az egy-egy tulajdonság lehet az összehasonlítás során: például teljes szóstatisztikák, szavak gyakorisága, szófaja, karakter n-gram száma, szóhossz, különféle gyakorisági jellemzők a teljes szövegre stb. A korszerű vizsgálatokban a legszélesebb körben elterjedt a minimum 100 leggyakoribb szó elemzése, ezt követi a mondathossz, a szóhosszúság, a hangsúlyos és hangsúlytalan szótagok váltakozása, a szókészlet gazdagsága, a leggyakoribb funkciószavak, a

2 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 .

3 Patrick Juola, „The Rowling Case: A Proposed Standard Analytic Protocol for Authorship Questions," *Digital Scholarship in the Humanities*, 30, 1. sz, 30, (2015), 100–113.

4 Dirk Van Hulle and Mike Kestemont, „Periodizing Samuel Beckett's Works: A Stylochronometric Approach," *Style* 6, 2. sz. (2016), 172–202.

5 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.

6 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.

7 Pelin Canbay, Ebru Akcapinar Sezer and Hayri Sever, „Deep Combination of Stylometry Features in Forensic Authorship Analysis", *International Journal of Information Security Science* 9, 3. sz. , (2020) 154–163.

8 Haixia Liu, Raymond H. Chan and Yuan Yao, „Geometric tight frame based stylometry for art authentication of van Gogh paintings," *Applied and Computational Harmonic Analysis* 41, 2. sz., (2016), 590–602.

9 Thomas Corwin Mendenhall, „The Characteristic Curves of Composition," *Science* 9, 214. sz. (1887): 237–249.

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

központozás, a kollokációk, bizonyos betűsorozatok gyakoriságának a vizsgálata.[11] Ennek megfelelően ezeket a gépi feldolgozásra alkalmas jegyeket kell kigyűjteni, és a különböző szövegeket ez alapján összehasonlítani. A számítógép tulajdonságok tízezrei alapján képes erre, s a nagy mennyiségű jellemzőknek a vizsgálatára statisztikai módszereket alkalmazunk. Ezek a statisztikai elemző módszerek rendkívül bonyolult és összetett eljárások, amelyek megfelelő kiválasztása függ attól, hogy konkrétan mire vagyunk kíváncsiak az adott szöveg esetében. Hiszen ami az egyik esetben sikeres attribúciós eljárás lehet, az nem feltétlen eredményes a másikban: meg kell határozni, hogy adott feladatra, adott szövegekre milyen tulajdonságokat kell vizsgálni, s ehhez mi az adekvát módszer. Az elemzéshez reprezentatív korpusz összeállítása rendkívül fontos feladat, hiszen tudnunk kell, hogy az elemzés során miből adódik a szövegek közti különbség. Az elemzendő szövegek mérete ugyancsak fontos paraméter: bizonyos távolságmértékek erre nagyon érzékenyek, és emiatt torzíthatnak. A sikeres stilometriai elemzés tehát közel sem triviális feladat, az elméleti háttér alapos ismerete mellett az empirikus módon szerzett tudás szintúgy alapvető fontosságú.

A hazai innovációként fejlesztett *Shtylo*[12] webalkalmazás abban nyújt segítséget, hogy azok számára is megkönnyítse a stilometriai elemzések elvégzését, akiknek komolyabb nehézséget okoz ennek az elemzőmódszernek az alapos megismerése és számítógépes futtatása.

A *Shtylo* webalkalmazás alapját a *stylo* jelenti, egy R nyelven írt könyvtár, amelyet Maciej Eder és társai készítettek és publikáltak 2016-ban[13]. A stilometriai elemzés során végzett műveletek mind az ő implementációjukra támaszkodnak. Ez egy gazdag könyvtár, sok funkcióval, de a használata az R nyelv ismeretét és az ilyen típusú programok futtatásához szükséges környezetet igényli. Dobi Jan Sándor erre a problémára adott választ, amikor létrehozta a *Shiny* elnevezésű R nyelvű framework használatával ennek a szoftvernek az első változatát, amely már webes környezetben adott hozzáférést a *stylo* funkcióihoz.[14] Mi ezt az alkalmazást dolgoztuk át és bővítettük ki a *Wizard* és *Analyzer* funkciókkal.

A *Shtylo* elsődleges előnye, hogy bárhonnan elérhető egy böngészővel, és képes url-ekről korpuszokat betölteni, hogy aztán azokon különböző elemzéseket lehessen elvégezni. Az elemzésekhez használt paraméterezés egy egyszerű *JSON* formátumú szövegként kimásolható és újra felhasználható más korpuszoknál. Ezen felül pedig hozzáférést biztosít a *stylo* fő funkciójához, egy részletesen paraméterezhető elemzéshez, annak eredményének elmentéséhez, valamint a *Wizard* és *Analyzer* felületekhez, amelyek között a paraméterezés szabadon mozgatható. Mivel az elemzések meglehetősen számításigényes műveletek, ezért ezek a háttérben futnak, hogy a felhasználó közben használni tudja a felületet egyéb célokra.

11 Maciej Eder, „Style-Markers in Authorship Attribution: A Cross-Language Study of the Authorial Fingerprint," *Studies of Polish Linguistics* 6, 1.sz. (2011), 99–114.

12 https://github.com/szakacsb/shtylo

13 Maciej Eder, Jan Rybicki, Mike Kestemont "Stylometry with R: a package for computational text analysis." *R Journal*, 8(1) (2016), 107–21.

14 Dobi Jan Sándor, Mészáros Tamás és Kiss Margit „Shtylo : stilometriai elemzések webes támogatása", in Magyar Számítógépes Nyelvészeti Konferencia, (14)., 2018., 423–436.

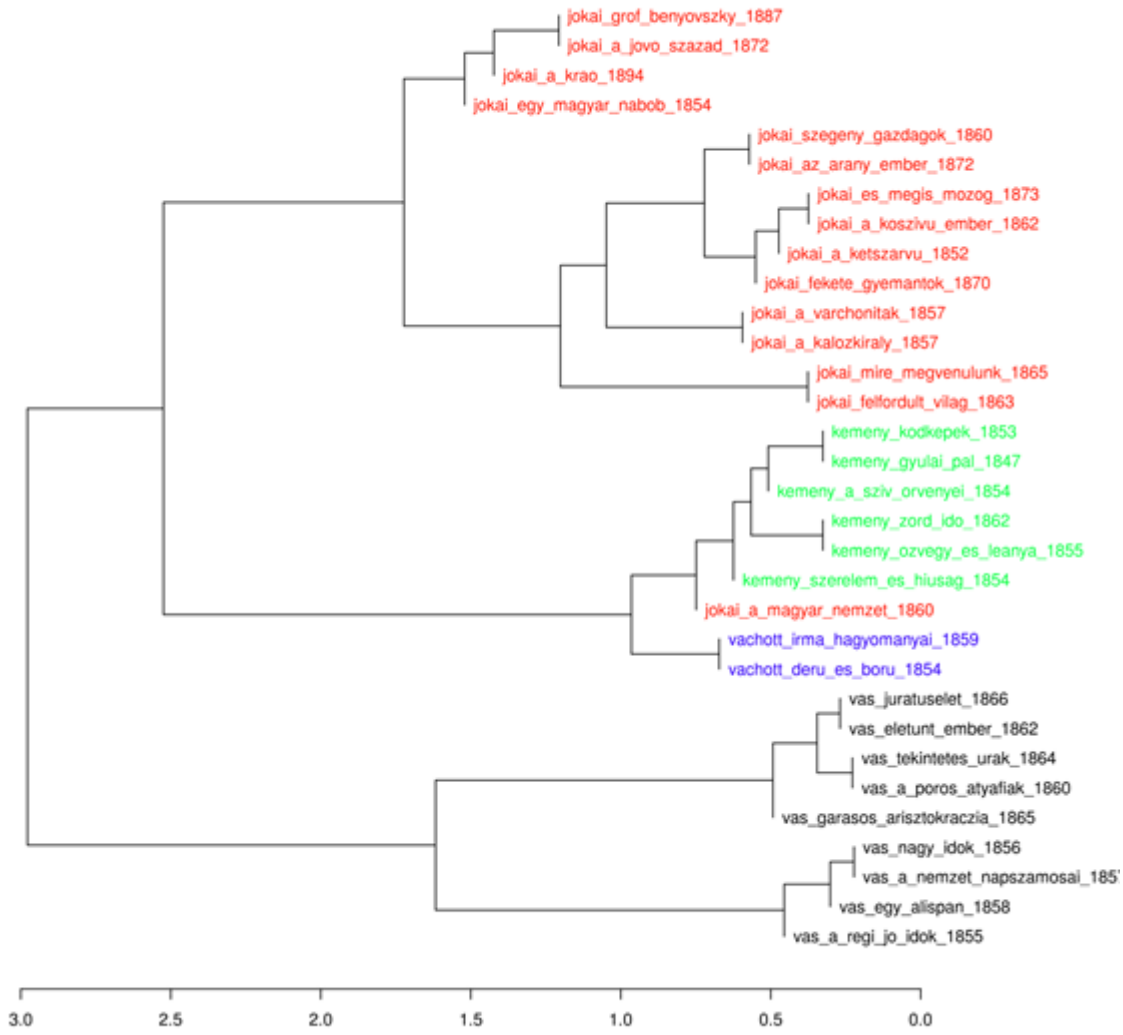
Maga a *Wizard*, hogy minél gyorsabb legyen, heurisztikus módon elemzi a szövegeket, szabályok segítségével, nem végez rajtuk tényleges stilometriai elemzést, csak alapvető statisztikai adatokat használ fel: a szó- vagy karakter-n-gramok eloszlását, a szövegek számát, a szövegek hosszát, és a szövegek szerzők közötti megoszlását. Pár fontosabb paramétert állít be vagy szűkít le: azt, hogy szó- vagy karakter-n-gram legyen használva, hogy mekkora legyen az n, a Most Frequent Words (a feature-ök hány százaléka legyen felhasználva) és a Culling (a szövegek hány százalékában kell jelen lennie a feature-nek, hogy használva legyen), és egyéb paraméterek. Ezek után a felhasználóra van bízva, hogy végighaladjon rajtuk, és módosítsa őket igényei szerint. A webes felületen láthatja a hozzájuk fűzött magyarázatokat is, amelyek segítenek a paraméterek elemzésében.

Az *Analyzer* esetében a cél az volt, hogy minél könnyebben csoportosíthatóvá tegye a megadott szövegeket. A működésének alapelve, hogy a szimulált lehűtés nevezetű lokális keresési algoritmus segítségével próbál egy olyan paraméter-kombinációt találni, amivel a megadott, különböző szerzőjű szövegek egymástól a lehető legtávolabb helyezkednek el. Ehhez minden lépésben megváltoztatja valamelyik paramétert egy egységgel (ez paramétertől függ, hogy mennyi). Mivel sok paraméter van, ezért csak a legfontosabbakat vizsgálja: ezek az elemzéshez használt n-gram típusa, hossza, a Most Frequent Words, a Culling és a távolságmérték típusa. A keresés az ezen paraméterek lehetséges értékeiből álló térben történik. Használhat előre megadott kezdőpontot ebben a térben, ami lehet akár a *Wizard* kimenete is, vagy van lehetőség arra, hogy több véletlenszerűen kiválasztott közül a legjobbat használja. Az algoritmus egy előre beállított lépésszám után áll meg. Jóval lassabb a *Wizard*-nál, mivel minden lépéskor lefut egy teljes stilometriai elemzés, viszont jelentősen pontosabb is. A használatával előállított paraméterezést aztán fel lehet használni olyan korpuszokon, ahol ismeretlen szerzőjű szövegek is vannak, és azok helyzetét lehet viszonyítani az ismert szerzőjű szövegekhez.

Az alábbi kísérlettel mutatjuk be az itt említett eljárások *(Wizard, Analyzer)* hatékonyságát. Az ELTE Digitális Bölcsészet Tanszéken létrehozott regénykorpuszból[15] összesen 32 regényt elemeztünk négy XIX. századi szerzőtől. Úgy válogattuk ki a műveket, hogy a keletkezési idejük közel essen egymáshoz – tehát a szerzőség és ne az adott időszak legyen a megkülönböztető elem. Az első elemzés a *Shtylo* alapbeállításaival futott.[16]

15 https://regenykorpusz.elte-dh.hu/

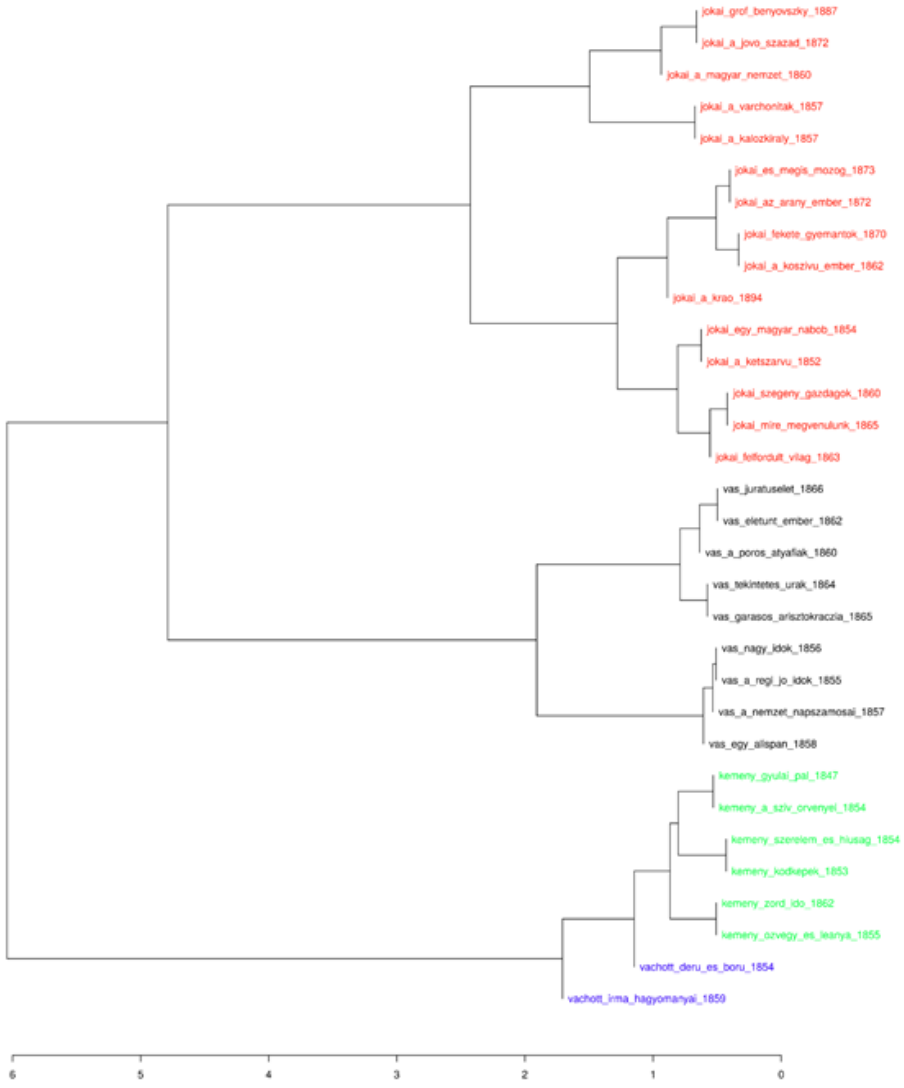
16 filetype: plain, language: English, encoding: yes, features: c, NgramInput: 1, CaseCheckbox: no, MfwMinimumInput: 100, MfwMaximumInput: 100, MfwIncrementInput: 100, MfwFreqRankInput: 1, CullingMinimumInput: 0, CullingMaximumInput: 0, CullingIncrementInput: 20, CullingListCutoffInput: 5000, CullingPronoun: no, Statistics: CA, StatisticsConsensus: 0.5, Scatterplot: labels, ScatterplotMargin: 2, ScatterplotOffset: 3, PcaFlavour: classic, ClusteringHorizontal: yes, Distances: dist.delta, SamplingMethod: no.sampling, SamplingNumber: 10000, OutputPlotHeight: 10, OutputPlotWidth: 10, OutputPlotFont: 10, OutputPlotLine: 1, OutputPlotColour: colors, OutputPlotDefault: no, OutputPlotTitles: no



Itt azt láthatjuk, hogy nagyrészt elkülönülnek az egyes szerzők művei, ám Jókai egyik regényét az elemző más klasztercsoporthoz sorolja, és az azonos szerzők művei is meglehetősen szétszóródnak egymástól.

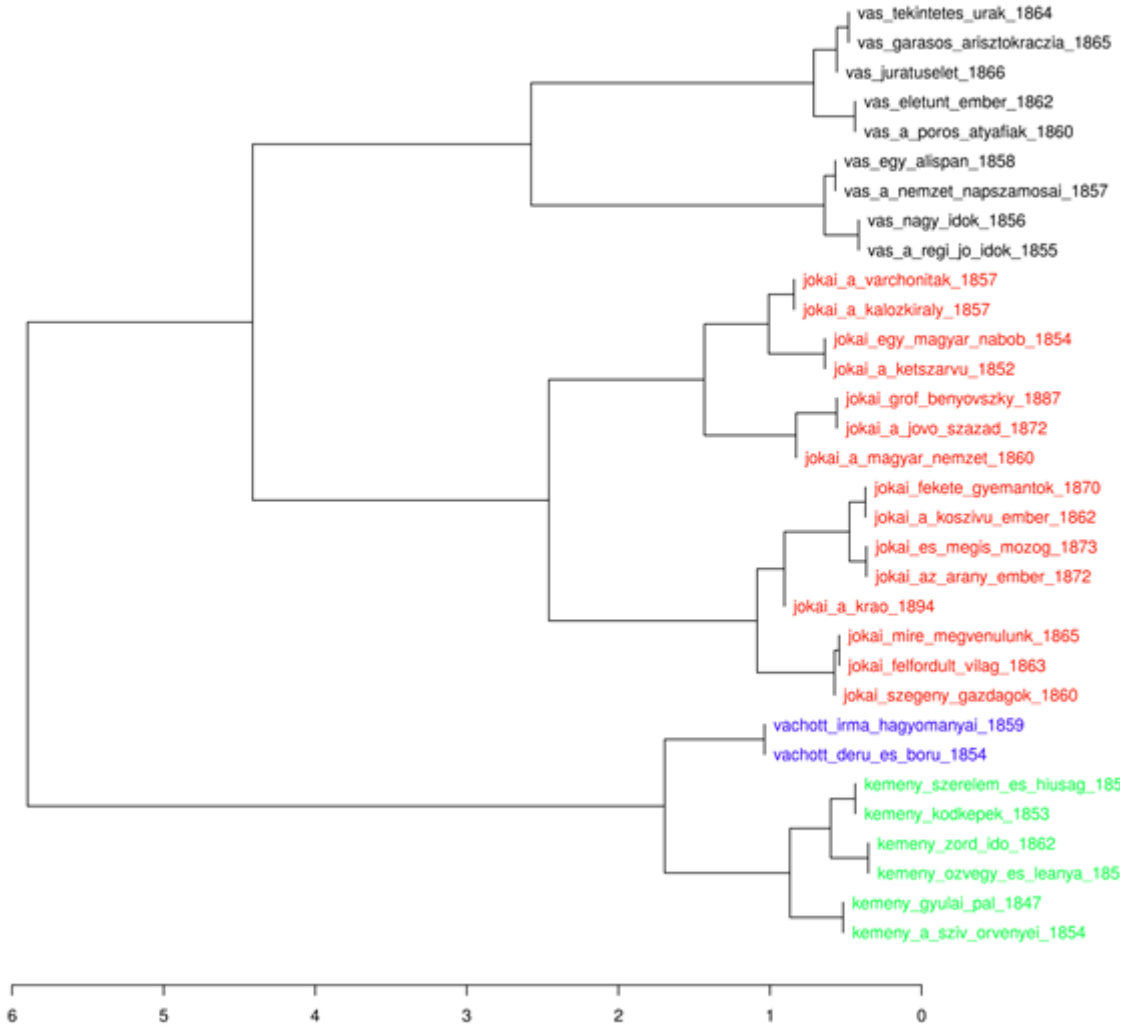
A következő kísérletben a *Wizard* futtatása után az alábbi eredményt kaptuk:[17]

17 filetype: plain, language: hungarian, encoding: yes, features: c, NgramInput: 3, CaseCheckbox: no, MfwMinimumInput: 67, MfwMaximumInput: 67, MfwIncrementInput: 100, MfwFreqRankInput: 1, CullingMinimumInput: 96, CullingMaximumInput: 96, CullingIncrementInput: 20, CullingListCutoffInput: 5000, CullingPronoun: yes, Statistics: CA, StatisticsConsensus: 0.5, Scatterplot: labels, ScatterplotMargin: 2, ScatterplotOffset: 3, PcaFlavour: classic, ClusteringHorizontal: yes, Distances: dist.delta, SamplingMethod: normal.sampling, SamplingNumber: .na, OutputPlotHeight: 16, OutputPlotWidth: 13, OutputPlotFont: 10, OutputPlotLine: 1, OutputPlotColour: colors, OutputPlotDefault: no, OutputPlotTitles: no



Ebben az esetben már egyértelműbben kirajzolódik a négy elkülönülő szerző csoportja, ugyanakkor az is látszik, hogy Vachott egyik regénye az elemzés alapján közelebbi rokonságban van Kemény Zsigmond szövegeivel, mint a másik általa írt művel.
Ezt követően az *Analyzer*t 10 iterációban futtattuk.[18]

18 filetype: plain, language: hungarian, encoding: yes, features: c, NgramInput: 3, CaseCheckbox: no, MfwMinimumInput: 67, MfwMaximumInput: 67, MfwIncrementInput: 100, MfwFreqRankInput: 1, CullingMinimumInput: 96, CullingMaximumInput: 96, CullingIncrementInput: 20, CullingListCutoffInput: 5000, CullingPronoun: yes, Statistics: CA, StatisticsConsensus: 0.5, Scatterplot: labels, ScatterplotMargin: 2, ScatterplotOffset: 3, PcaFlavour: classic, ClusteringHorizontal: yes, Distances: dist.delta, SamplingMethod: normal.sampling, SamplingNumber: .na, OutputPlotHeight: 16, OutputPlotWidth: 13, OutputPlotFont: 10, OutputPlotLine: 1, OutputPlotColour: colors, OutputPlotDefault: no, OutputPlotTitles: no



A négy szerző regényei ebben az elemzési eljárásban határozottan elkülönülnek egymástól, ugyanakkor az azonos szerzőhöz tartozó művek nem szóródnak külön ágakba, hanem sokkal szorosabb tömböket alkotnak, mint a korábbi kísérletekben. Az is tisztán látszik, hogy a négy szerző olyan különálló tömböket alkot, amelyek nem keverednek más szerzők klasztereivel.

A fenti kutatások és fejlesztések szakmai hátterét a Bölcsészettudományi Kutatóközpont Irodalomtudományi Intézet, valamint az ELTE BTK TI Digitális Bölcsészet Tanszék együttműködésével létrejött Stilometriai Kutatócsoport biztosítja, az infrastruktúrát pedig a Digitális örökség nemzeti laboratórium nyújtja. A *Shtylo* eszközre épülő szolgáltatás megvalósításához a technológiai know-how-t a BME VIK Méréstechnika és Információs Rendszerek Tanszék oktatója és hallgatói biztosították, nélkülük nem jöhetett volna létre a szolgáltatás.

A szolgáltatás nyitott minden érdeklődő számára, de célcsoportként a stilometriai kutatások iránt érdeklődő kutatókat és egyetemi hallgatókat határoztuk meg, azokat méghozzá, akik bár használni szeretnék a számítógépes stilisztika digitális eszközeit, de - híján a programozói kompetenciának - eddig ki voltak zárva ezen eszközök használati köréből. A limitált számítógépes infrastruktúrával, illetve tudással rendelkező kutatók számára különösen a fent bemutatott, a hiperparaméterek beállításokat automatizáló eszközök jelentenek nagy segítséget, ezek képezik a *Shtylo* újdonságát az eredeti R csomag, a *stylo*-hoz képest. Azonban még az automatizált paraméterezés sem teszi magától értetődővé, még kevésbé tudományos szempontból körültekintően argumentálttá a stilometriai kalkulációkat. Ehhez olyan képzésekre van szükség, ahol szerzőazonosításban vagy más stilometriai műveletekben technológiailag járatos, gyakorlott, de a szakirodalmi kontextust is ismerő kutatók bemutatják az eszközök helyes használatát, konkrét elemzéseken keresztül. Az elkövetkező években a Stilometriai Kutatócsoport ilyen képzések elindítását tervezi, személyes jelenlétre épülő kurzusok mellett eLearning anyagok formájában is.

A kvantitatív stilometriai kalkuláció nyelvfüggő, a magyar nyelv nyelvstatisztikai vizsgálata egészen más technológiákat követel meg, mint a stilometriai elemzések fősodrába tartozó világnyelvekre épülő elemzések. Ezért nagyon fontos, hogy a kalkulációk eredménye mögötti hiperparamétereket, és a paraméterezés számszerű eredményeit publikáljuk, hogy a jövőbeli, feltételezhetően egyre nagyobb számú egyedi kutatási eredmény, szerzőazonosító algoritmusra támaszkodó érvelés egy már létező és jól dokumentált kutatási térbe illeszkedjen, növelve az eredmények ellenőrizhetőségét. Eddig nagyon kevés ilyen eredményt publikáltak magyar nyelvű anyagok vizsgálatára építve. Ezért is fontos, hogy miközben mérésekkel alátámasztott érvelések segítségével erősödjön meg a magyar stilometria szakirodalma, folyamatosan a nálunk előbbre járó diskurzusok összefüggésében helyezzük el a kutatásainkat.

A Stilometriai Kutatócsoport ezen dolgozik, részt veszünk az International Journal of Digital Humanities szerzőazonosítással foglalkozó különszámának szerkesztésében, illetve a COST Action „Distant Reading for European Literary History" nemzetközi projektben, amelynek keretei között olyan korpuszokat hozunk létre és publikálunk, amelyek eleve a stilomteriai vizsgálatokra vannak optimalizálva.

A szolgáltatás további fejlesztésének rövid távú céljai között szerepel a felhasználói fiókok kialakításának lehetősége, lehetőség szerint EduID alapú azonosítás használatával, valamint a korpuszok kezeléséhez felhő alapú tárhely szolgáltatás integrálása. Hosszabb távon a továbbra is az R programnyelv környezetében működő *stylo* funkcionalitását a digitális bölcsészetben előre törő, hovatovább egyeduralkodóvá váló, így a kutatók szélesebb köre számára könnyebben elérhető *Python* nyelvi környezetbe fogjuk átültetni. Ezen munka első lépéseit a jelen kötetben megjelenő Jókai stilometriai elemzés írja le. Célunk továbbá, hogy a szerzőazonosítás módszertanát az irodalmi szövegeken túli szövegműfajokra is kiterjesszük, kézenfekvő az eszközök felhasználása webaratással létrejövő korpuszokra.

Távlati célunk olyan új, a *stylo* algoritmusainak frissebb technológiák integrálása a stilometriai kutatásba, mint a vektortér technológia, illetve a mélytanulás eszközeinek alkalmazása. Ezzel a munkával párhuzamosan haladunk a Maciej Eder vezette, és Európa egyik legfontosabb számítógépes stilisztikai műhelynek számító krakkói „Computational Stylistics Group" tevékenységével.