# Shtylo: stilometriai elemzések webes támogatása

Dobi Jan Sándor[1], Mészáros Tamás[1], Kiss Margit[2]

[1] BME Méréstechnika és Információs Rendszerek Tanszék
H-1117 Budapest, XI. Magyar tudósok körútja 2. I ép. E437.
meszaros@mit.bme.hu
[2] MTA BTK Irodalomtudományi Intézet
1118 Budapest, Ménesi út 11–13.
kiss.margit@btk.mta.hu

**Kivonat:** A stilometria a számítógépes nyelvészet dinamikusan fejlődő területe. Széles körű felhasználását azonban gátolja az a tény, hogy alkalmazóinak többsége nincs a szükséges informatikai tudás birtokában. Cikkünkben egy olyan rendszert mutatunk be, amely az R nyelven írt stylo programcsomaghoz nyújt egy teljes értékű webes felhasználói felületet, valamint segítséget nyújt a stilometriai kísérletekhez szükséges korpuszok összeállításában és tárolásában is. Az elkészített szoftver működését egyrészt történeti szövegek elemzésén, másrészt plágiumkeresési feladat végrehajtásán mutatjuk be.

## 1 Bevezetés

A digitalizált szövegkorpuszok létrehozása a hagyományos módszerektől eltérő, új elemzési eljárások kialakítását teszi lehetővé, illetve kívánja meg. Az egyik, napjainkban egyre inkább elterjedő ilyen korpusznyelvészeti módszer a stilometria, amelynek egyik széles körben használt eszköze az R nyelven írt Stylo programcsomag [7, 9]. Az ilyen, alapvetően statisztikai vizsgálatokhoz az R nyelv alkalmazása kézenfekvő megoldás, ám használata olyan speciális szakértelmet kíván, amellyel éppen azok a kutatók kevéssé rendelkeznek, akik a számítógéppel támogatott nyelvi elemzésektől komoly eredményeket várnak. Célunk egy olyan támogató eszköz létrehozása volt, amellyel a felhasználók az R nyelv mélyreható ismerete nélkül is elvégezhetik a manapság leggyakrabban alkalmazott stilometriai elemzéseket. Ehhez kifejlesztettünk egy webalkalmazást, amely egyrészt egyszerűen használható felülettel látja el a Stylo programcsomagot, másrészt kiegészíti korpuszkezelési funkciókkal. Tanulmányunkban az eszköz alkalmazási lehetőségeit két, merőben eltérő adatszerkezetű korpuszon mutatjuk be: egyrészt 18. századi történeti szövegek téma- és stíluskategorizálására alkalmaztuk, másrészt mesterséges nyelvű szöveges dokumentumokon végzett plágiumkeresésre.

### 1.1 A számítógépes stilometria és problémái

A stilometria, a stílus elemzésének számszerűsített változata, a számítógépes nyelvészet egyik sokat kutatott és széles körben alkalmazott területe. Módszerei alapvetően a statisztikai szövegelemzés területéről kerülnek ki, tipikusan olyan jellemzőket és azokra épülő klaszterezési eljárásokat alkalmaz, amelyek a szövegeket hasonlóságát különféle jellemzőik és azokon értelmezett távolságmértékek szerint ítélik meg. A mértéket úgy szabja meg, hogy a bemeneti szövegeknek a feladat szempontjából releváns jellemzőit ragadja meg.

Egy stilometriai vizsgálat több lépésből áll: a korpusz összeállítása, a kívánt jellemzők, valamint a meghatározásukhoz legmegfelelőbb módszerek kiválasztása, az előfeldolgozás, a jellemzőkinyerés és -feldolgozás, majd az eredmény előállítása. Bár a lépések többsége jól támogatható informatikai eszközökkel, és ily módon lehetőség nyílik nagy méretű korpuszok egyszerű és gyors elemzésére, a gépi stilometriai elemzésnek megvannak a maga akadályai, amelyek megnehezítik az effajta törekvéseket.

Ezek az alábbi csoportokba sorolhatók:

- **Technológiai nehézségek:** Noha a számítógépek kapacitása rohamosan nő, de ugyanez elmondható az elemző módszerek komplexitásáról és a korpuszok méretéről is. A számítógéppel szemben támasztott igényeink általában felemésztik a rendelkezésre álló kapacitást. Az egyre növekvő méretű adathalmazokon futtatott elemzőmódszerek jelentős számítási igénnyel rendelkeznek, ezért van egy felső korlátja annak, hogy egy átlagos kutató mekkora szövegkorpuszokon dolgozhat értelmes időkeretek betartásával.
- **Tudás hiánya:** A korpuszmodell megalkotása és a kívánatos jellemzők meghatározása igen tudásintenzív feladat. A kutatónak tudnia kell, hogy milyen jellegeket kíván kinyerni a szövegekből. Erre vannak ajánlások, de minden kísérlet egyedi. Mi több, manapság elemzőalgoritmusok hada áll rendelkezésre: melyiket válasszuk egy adott szövegkorpusz esetén? Hagyományos statisztikai vagy gépi tanulás alapú módszert válasszunk? Ezekre a kérdésekre nincsen egyértelmű és könnyű válasz.
- **Megfelelő módszerek hiánya:** Noha számos publikált és implementált módszer áll egy kutató rendelkezésére, de a gépi stilometria mégiscsak egy fiatal terület. Egyelőre nem alakultak ki általánosan optimális megoldások, bizonyítottan „jó” eljárások a feladatok megoldására. Néha már a probléma formalizálása is akadályokba ütközik. Ide sorolhatók a nyelvfeldolgozással kapcsolatos problémák is. Egy korpusz részletes nyelvi elemzése minden bizonnyal sokkal gazdagabb eredményeket szolgáltatna, de a természetes nyelvű (különösen a történeti) szövegek informált nyelvi elemzését számos tényező gátolja (nyelvtani változatosság, elírások, nyelvi hibák). Ehelyett a statisztikai és a gépi tanulásos black-box módszerek alkalmazhatók egy-egy kutatási kérdés megválaszolásához.

### 1.4 Tipikus feladatok

A stilometrián belül vannak tipikusnak mondható kérdések, amelyeket megfogalmaznak egy szövegkorpusz elemzése kapcsán, és alapvetően a szerző stílusára, írásmódjára vonatkoznak. Az alkalmazások alapvetően a szöveg „megmért" stílusjegyeinek felhasználásban különböznek. Az alábbiakban ezekre mutatunk be néhány példát.

**Törvényszéki stilometria.** A stilometria egy jelentős alkalmazási területe a jogászi munka módszertanába is bekerült, jelentős mértékben a bűnüldözés területén van jelen [2]. Amikor nem állnak rendelkezésre fizikai szövegek, csak digitálisak (ami egyre gyakrabban előfordul) a szöveg íróját nem lehet többé az íráskép alapján beazonosítani. A digitális szövegeknek nincsen írásképük, így a szöveg más tulajdonságai alapján kell beazonosítani, hogy ki írta őket. Tovább nehezíti a feladatot az, hogy esetenként számolni kell a megtévesztés lehetőségével [3]. Ide sorolható a plágiumkeresés is, ugyanis ilyenkor is egy megtévesztési kísérletet kell kiszűrnünk. A mintapéldákról szóló rész bemutat egy, a rendszerünkkel végzett plágiumkeresési tesztet is.

**Történelmi szövegek szerzőségi, autentikusságának vizsgálata.** A régebbi korok szövegei kapcsán különböző okokra visszavezetve gyakran merülhet fel a valós szerzőség azonosításának a kérdése. Az ilyen kérdések megválaszolása a nyelv- és irodalomtudomány egyik fontos kutatási területe, amely a digitális módszertannak köszönhetően a korábbiaktól hatékonyabb módon képes megválaszolni a tisztázatlan kérdéseket. A stilometriának ez az egyik legtermékenyebb területe [18], amely a modern a korban jelentős eredményeket hozott kezdetben különösen a Shakespeare-kutatásban [4, *19]*; de azóta számtalan más szerzővel kapcsolatban is, ami következésképpen a módszertan folyamatos változását eredményezi [20]. A cikkünkben egy magyar 18. századi szövegeken végzett kísérletet mutatunk be.

**Időbeliséganalízis.** A harmadik jelentős terület, ahol eredményeket értek el stilometriai eszközökkel: a bizonytalan korú szövegkorpuszok időbeliségének vizsgálata. Lorenzo Valla kísérlete részben szintén ide tartozik, aki elemzését a szöveg latin stílusának korára alapozta. Wincenty Lutosławski elemzése is egy effajta kísérlet: Platón párbeszédeit rendezte időrendi sorba stilisztikai jellemzőik alapján [6]. A platóni dialógusok kérdése azóta számos alkalommal megragadta egy-egy kutató figyelmét, hisz az időbeliség kulcsfontosságú Platón filozófiájának megértéséhez. Napjainkban a szerzőségi vizsgálatokat alkalmazzák nyelvtörténeti periodizációban, de szerzői életmű korszakolásában is [21]. Esettanulmányként mi is végeztünk egy kísérletet, amelyben Mikes Kelemen műveinek kronológiai vizsgálata volt a cél.

**Csoportbéli hovatartozás vizsgálata.** Egy adott szöveg esetén nem csak a konkrét szerzőre lehetünk kíváncsiak. Számos esetben elegendő, ha el tudjuk dönteni, hogy valamely csoporthoz való tartozása fennáll-e [22]. Bizonyos elemzésekben már ez is értékes információkkal szolgálhat. Kérdés lehet, hogy a szerző férfi-e vagy nő, milyen nemzetiségű, idősebb vagy fiatalabb, milyen iskolázottságú, milyen személyiségtípusba tartozik. Az efféle elemzések egy speciális esete a szerző anyanyelvének meghatározása az idegen nyelvű írása alapján.

### 1.5 Stilometriai eszközök

A stilometria egyre népszerűbb tudományterület, számos módszerének létezik szoftver-megvalósítása.

A Signature [7] ingyenes szoftver, melyet Peter Millican fejleszt az Oxfordi Egyetemen. A weboldal szerint főleg szerzőségi vizsgálatra alkalmas. Számos esettanulmány található a weboldalon, melyet azonban úgy tűnik nem frissítettek egy ideje. A Java Graphical Authorship Attribution Program, vagy JGAAP [8] egy nyílt forrású, szabad szoftver, melyet a Githubon lehet elérni. Fő fejlesztője Patrick Juola. Például a Robert Galbraith álnéven publikált Kakukkszó című regény valódi szerzőjének (J. K. Rowling) azonosításához a JGAAP szoftvert alkalmazták, az elemzésben Juolanak jelentős szerepe volt. A Stylene az Antwerpeni Egyetem CLiPS kutatócsoportja által működtet stilometriai webszolgáltatás holland nyelvre.

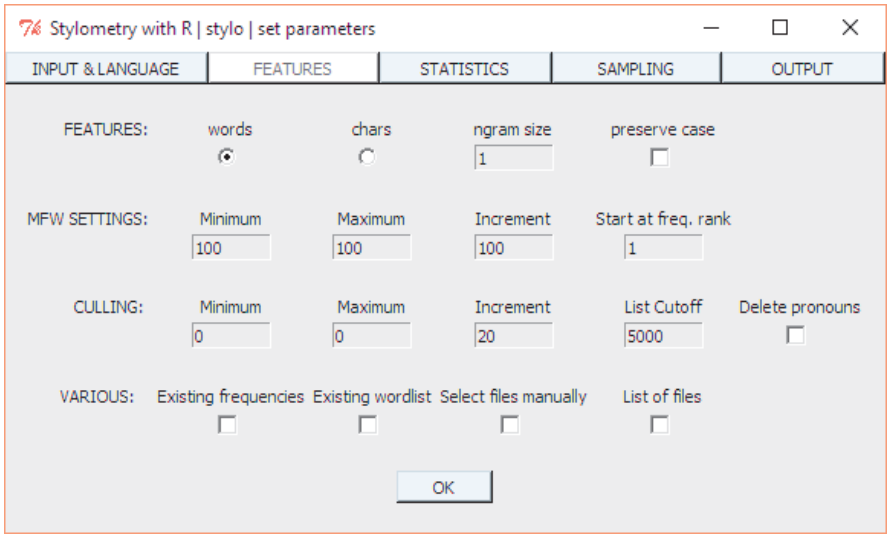
Az általunk is használt stylo, mely a Shtylo alapját képezi, egy R-csomag, melyet Maciej Eder, Jan Rybicki és Mike Kestemont fejlesztett, szabad nyílt forrású szoftver. A következő rész ezt a csomagot mutatja be.

## 2 A stylo csomag

A csomagot Maciej Eder, Jan Rybicki és Mike Kestemont írta [9]. A csomag szabad szoftver, letölthető a Githubról [10]. 2013-ban készült, megjelenése előtt [11] a teljes stilometriai workflow – feltehetőleg a JGAAP kivételével – nem állt rendelkezésre egy szoftver részeként. A stylo magába foglal számtalan releváns elemzési módszert, továbbá megkönnyíti a teljes stilometriai folyamatot. A csomagba foglalt magas szintű csomagolófüggvények a workflow majdnem minden lépését elvégzik helyettünk, csak a konfigurációjukat kell megadnunk.

### 2.1 A Stylo főbb funkciói

A csomag öt fő magas szintű csomagolófüggvénnyel rendelkezik, melyek magukba foglalják az egyes lépéseket és egy egész elemzést végigvisznek az elejétől a végéig. Hívható grafikus felülettel, ekkor a paraméterek grafikus vezérlőelemek segítségével állíthatók be, vagy grafikus felület nélkül, ekkor a paraméterek R kódból állíthatók. A felhasználói felületre látható példa az 1. ábrán, amelyen a stylo függvény paraméterek nélküli meghívásának eredményét látjuk. A többi csomagolófüggvény felülete is nagyon hasonló.




**1. ábra**. *A stylo csomag stylo() függvényének grafikus felülete.*

Az öt függvény:

- **stylo**: A függvény feladata, hogy a stilometria fő irányzatának számító modellalapú, statisztikai módszereken alapuló elemzést hajtsa végre. Számos elemzési eljárás közül választhatunk, mint a klaszteranalízis, a Multidimensional Scaling (MDS), a főkomponens-analízis kovarianciamátrixszal (PCV), a főkomponens-analízis korrelációs mátrixszal (PCR) és a konszenzusfa (BCT). A módszerekben különféle távolságmértékeket alkalmazhatunk (euklideszi távolság, Manhattan-távolság, Burrows-delta [12], Argamon-delta [13], Eder-delta, egyszerű Eder-delta, Canberra-távolság, koszinusztávolság).
- **classify**: A függvény gépitanulás-alapú osztályozó algoritmusok futtatására alkalmas. Az eljárás kétlépcsős: először betanítunk egy osztályozót egy megfelelően strukturált tanítóminta-készlettel, majd a számunkra érdekes szövegeket megkíséreljük az osztályozóval besorolni. A függvény számos eljárást ismer (Burrows távolságmérték, k-Nearest Neighbours, Szupportvektor-gépek, Naiv Bayes, Nearest Shrunken Centroid).
- **rolling.delta, rolling.classify**: A már említett elemzések görgetett változata: céljuk, hogy egy korpusz szövegein belül stilisztikai váltások felismerését tegyék lehetővé. A szövegeket adott méretű átlapolható ablakokra osztjuk, és minden ablakon egy elemzést hajtunk végre. Az így kapott értékeket egy grafikonon ábrázolva hirtelen változásokat lehet felfedezni, melyek számos dolgot jelenthetnek, pl. egy szerző átvette a művet egy másiktól.
- **oppose**: A függvény két szövegkorpusz összehasonlítására való. A függvény eredményeként két szólistát kapunk: az első a tesztelt szerző által kifejezetten

preferált szavakat tartalmazza, míg a második ezzel szemben a szerző által kifejezetten került szavakat.

## 3 A Shtylo webalkalmazás

A stylo potenciális alkalmazóinak jelentős részétől nem lehet elvárni, hogy ismerje az R programozási nyelvet, és tudjon használni egy R-csomagot. Ez más jellegű szakértelmet igényel. Manapság webrendszerekkel mindenki kapcsolatban áll, így egy megfelelő webes felhasználói felület alkalmazásával kezelhetjük ezt a problémát. Noha a stylo csomagnak létezik egy grafikus és egy kezdetleges webes felülete is (a websty), ezek nehezen használhatók. Számos érv szól amellett, hogy a stylohoz készítsünk egy színvonalas webalkalmazást.

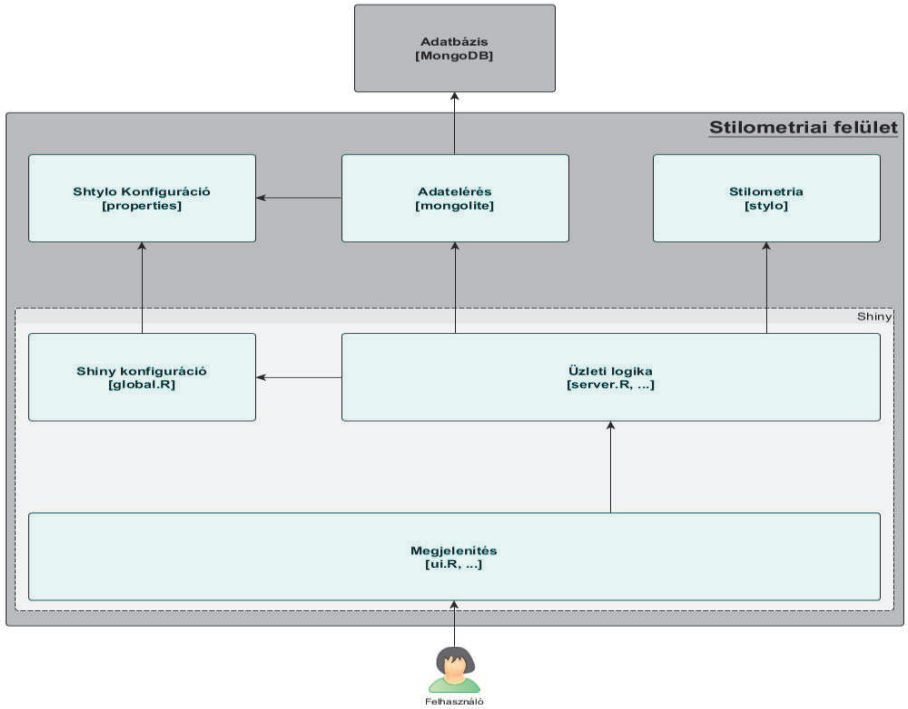
A felhasználói felület azonban nem az egyetlen megoldandó probléma. Érdemes még egy terhet levenni a kutató válláról: a futtatókörnyezet kialakítását (szoftverek telepítését és beállítását). Egy webalkalmazás esetén csupán egy böngészőre van szükség, az R program egy szervergépen fut, ott egyszer kell a szoftvert telepíteni és beállítani.

Egy harmadik szempont is említést érdemel. A stylo elemzései igen számításigényesek is tudnak lenni: nagy korpuszokon sok memóriát és processzoridőt tudnak fogyasztani. Ez jelentősen szűkítheti azon eszközök halmazát, ahol az eszköz használható. Ha egy központi, erős szervergépen fut az alkalmazás, akkor a szolgáltatásokat egy gyenge kliensről is el lehet érni, bárhova magunkkal lehet vinni, és igény szerint lehet elemzéseket futtatni. Ez a megoldás jól illeszkedik az informatikai szolgáltatások felhőalapú (azaz jól skálázható) megvalósításához is.

Az általunk készített Shtylo rendszer az eddigiek mellett egy további jelentős hozzáadott értékkel is bír, ez pedig a korpusz adatbázis-alapú tárolása. A stylo csomag a korpuszt, az elemzési eredményeket mind fájlok alakjában tárolja. Ez rosszul skálázódik, és nem nagyon alkalmas arra, hogy a segítségével korpuszokat építsünk és tartsunk nyilván. A Shtylo ezzel szemben a korpuszépítésben is segítséget nyújt. A felületen szövegeket lehet adott korpuszokban feltölteni, és a korpuszokat egy adatbázisban tárolja el. Ez jól skálázódik, tehát jelentős mennyiségű szöveg és sok felhasználó esetén a replikáció és terheléselosztás is megoldható. Ugyanez a fájlrendszerben tárolt fájlokkal jóval nehezebb lenne.

### 3.1 A Shtylo felépítése és technológiái

A Shtylo alapvetően egy kliens-szerver alkalmazás, amely kliens oldalon standard webböngészőt alkalmaz, szerver oldalon pedig az R környezet mellett egy webkiszolgálóra és egy korpusztárra (adatbázisra) épül.



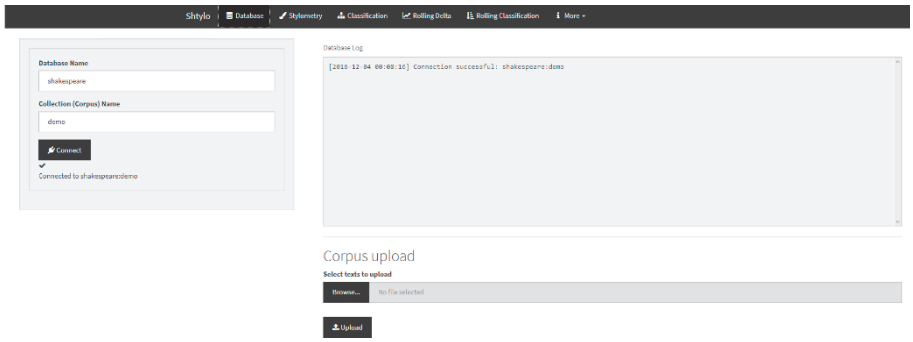

**2. ábra**. *A Shtylo főbb részei*

Technológiai oldalról nézve a Shtylo a webes felület összeállítására a Shiny [14] keretrendszert használja, mely egy R-nyelvű webalkalmazások fejlesztésére alkalmas reaktív szoftverkönyvtár. Innen ered alkalmazásunk neve is: (Sh)iny + s(tylo).

Adatbázis-technológiának a MongoDB [15] dokumentumalapú adatbáziskezelőt választottuk, ugyanis adattárolási modellje jól illeszkedik a feladathoz és minden teljesítménybeli vagy funkcionális igényt kielégít.

### 3.2 A korpuszokról

A Shtylo rendszerben az adatréteg egy MongoDB adatbázis, amelyben a szövegek ún. gyűjteményekbe vannak szervezve és több gyűjtemény alkot egy adatbázist. Amikor egy szöveget feltöltünk, az a szöveg mindig egy MongoDB gyűjtemény, vagyis a mi értelmezésünkben egy korpusz tagja lesz. Amikor a Shtylo meghívja a stylo csomagot, akkor mindig az aktuálisan csatlakoztatott MongoDB gyűjteményt adja át neki elemzésre, azaz egy korpuszt. A gyűjtemények, vagyis korpuszok mindig egy adatbázisnak a tagjai, amelyeket felhasználókhoz, illetve projektekhez rendelhetünk.

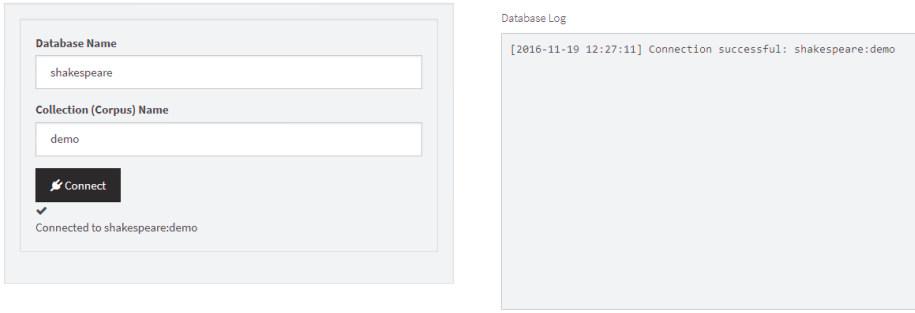



**3. ábra.** *A Shtylo adatbázis-felülete. Balra az adatbázis és a gyűjtemény kiválasztása, jobbra lent a feltöltésre szánt fájlok megadása, felette az adatbázisnapló látható.*

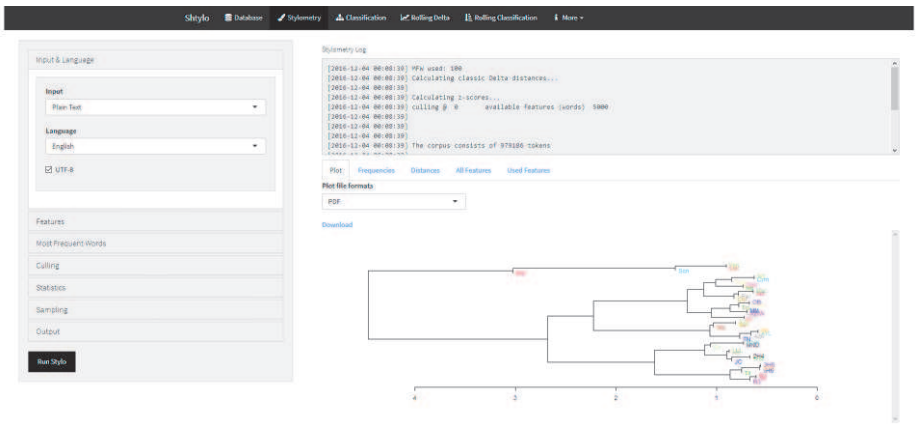
### 3.3 A rendszer a felhasználó szemével

Felhasználóként a rendszer igen egyszerű. Az alábbi cselekvésekkel lépünk kapcsolatba a rendszerrel:

1. Kiválasztjuk, mely adatbázisnak mely gyűjteményén kívánunk dolgozni.
2. Ha ez még üres, akkor feltölthetjük a szövegeinket.
3. Elemzéseket futtatunk: kiválasztjuk, hogy jelenleg melyikre vagyunk kíváncsiak, és beállítjuk a paramétereit. A paraméterek a stylo beállítási lehetőségeitől függnek.
4. Az eredményeket a felületen látjuk, az adatokat, képeket változatos formátumokban elmenthetjük a lokális gépünkre.




**4. ábra.** *A csatlakozási űrlap és a napló egy sikeres csatlakozás esetén.*




**5. ábra.** *A stylo() függvény felülete a Shtylo webalkalmazásban egy elemzés lefuttatása után.*

A rendszer jelen pillanatban még csak a stylo() függvény számára biztosít felületet, mivel azonban ez a leggyakrabban használt funkció, ezért már így is hasznos lehet a stylo felhasználói számára. A továbbiakban három elemzési példát mutatunk be, melyek elvégzése során a Shtylo segített a stylo gördülékenyebb használatában.

# 4 Példák

A három példa közül kettő a Mikes-életművel foglalkozik. A harmadik egy kísérlet arra, hogy a stylo csomagot plágiumkeresésre használjuk egy egyetemi tárgy házi feladatára beküldött megoldásokon keresztül.

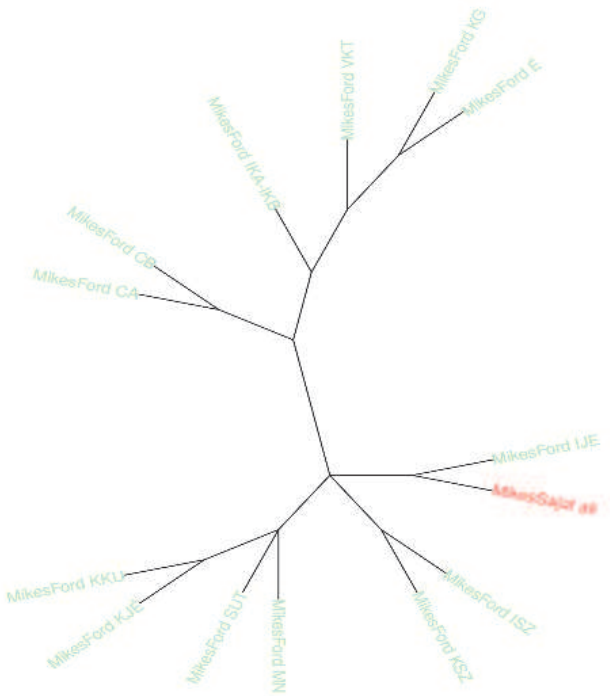
A Mikes-műveket a stilumetriai elemzés előtt egy szótárra épülő módszerrel normalizáltuk annak érdekében, hogy a szavak jellegzetesen sokféle írásmódja minél kevésbé torzítsa az eredményeket. A normalizálás részleteiről a [23] irodalomban számoltunk be.

## 4.1 Első kísérlet: saját művek és fordítások

Ebben a kísérletben arra a kérdésre kerestük a választ, hogy eldöntsük a stylo segítségével, a mikesi életműben hogyan viszonyulnak a saját művek a szerző által készített fordításokhoz. A kísérlet megtervezésekor az alábbi dolgokat tartottuk szem előtt:

- Mikes Kelemen saját művei egy fájlban voltak megadva, míg a fordítások külön-külön. Így az egyik osztályban több volt a mű, mint a másikban. Ilyenkor a Klasszikus Delta torzíthatja az eredményeket, így Eder deltájára került a választás, mely a nem izoláló nyelveken jobban teljesít.

- A konszenzusfa építése során magasra állítottuk a konszenzusküszöböt, mivel célszerű maximalizálni azt a hasonlóságot, ami mellett egy fordítás összekötésre kerül az eredeti művekkel.
- A műhosszak jelentős szórása miatt a mintavételezés mellett döntöttünk, és a szövegeket 10000-jelleg—hosszú darabokra osztottuk fel, ezzel csökkentve a statisztikai torzulást.



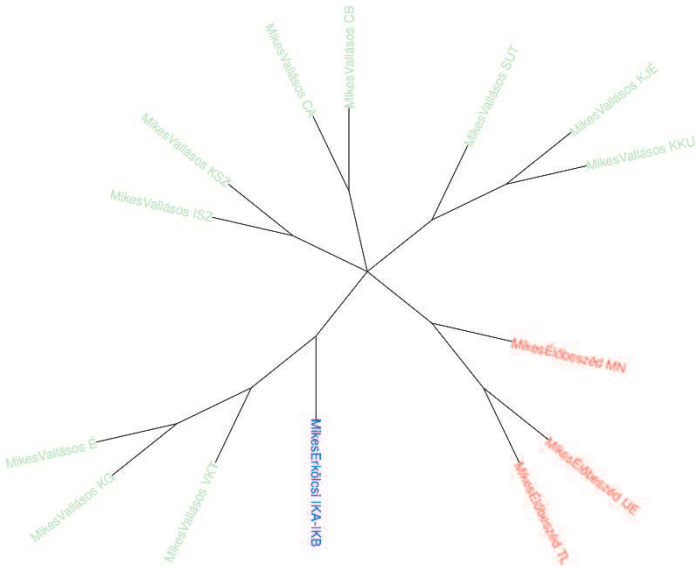

**6. ábra.** *Az első kísérlet konszenzusfája.*

Ennek az elemzésnek az eredményeképp a kísérlet során a művek jól láthatóan négy csoportba sorolódtak, amely árnyalja az életmű eddigi felosztásáról alkotott képet is. A beállítások finomításakor azt tapasztaltuk, hogy Az idő jól eltöltésének módja című fordítása mindig nagyon közel esik a Törökországi levelekhez (saját szerzőségű mű), bármilyen beállítást is használunk. Ezen kívül látni, hogy a többi fordítást külön konszenzuságakban találjuk meg, tehát csak ez az egyetlen fordítás volt olyan közel az eredetihez, hogy a futások 90%-ában volt közöttük közvetlen kapcsolat. Meglepetés az is, hogy az eddigi hagyományos kutatásokban a Törökországi levelekhez sok szempontból kötődő Mulatságos napok című fordítás nem kapcsolódik olyan szorosan a levelekhez.

### 4.2 Második kísérlet: művek tematikája

Ebben a kísérletben az volt a cél, hogy tematikai vizsgálatot végezzünk az életmű egészében. Arra voltunk kíváncsiak, hogy a stylo képes-e a különféle tematikájú szövegeket, így az egyházi, erkölcsi és élőbeszéd kategóriájába sorolható műveket különválasztani, illetve hogy ezek a művek stilisztikai szempontból megkülönböztethetőek-e. Három élőbeszédes, egy erkölcsi, és tíz vallási témájú szöveg alkotta a korpuszt. Ennél a kísérletnél:

- A Canberra távolság mellett döntöttünk, mert az érzékeny a ritka szókincsbeli különbségekre.
- A művek hosszeltérését itt is mintavételezéssel ellensúlyoztuk.
- A konszenzusküszöböt itt alacsony, mert itt az a kérdés, hogy a művek kapcsolatban lesznek-e egymással Canberra távolság mellett, vagy távol maradnak.

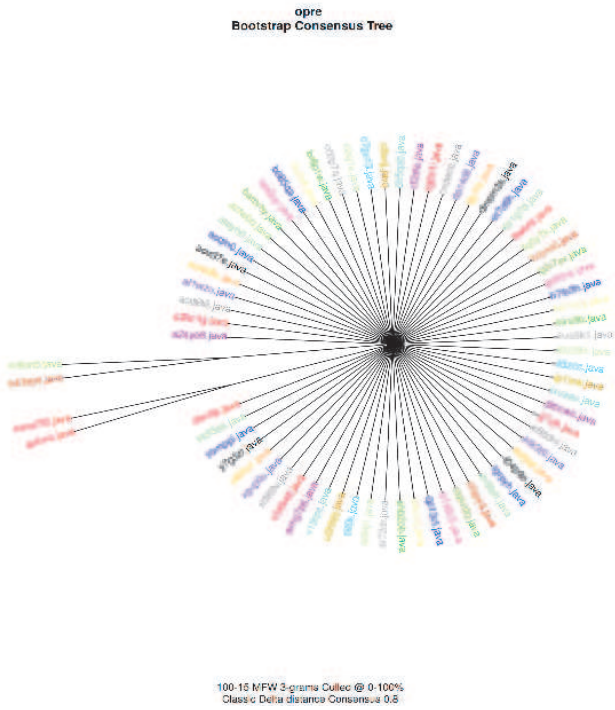



**7. ábra.** *A második kísérlet konszenzusfája.*

A kísérlet eredményeképpen két fő dolog látszódik világosan kirajzolódni. Először is az élőbeszéd jellegű szövegeket jól láthatóan külön lehet választani a többitől. Másodszor pedig az erkölcsi témájú szöveg besorolódik egy fába a vallási témájú szövegek egy részével. Tehát itt a stylo főleg arra volt képes, hogy az élőbeszédes szövegeket és a nem élőbeszédes szövegeket különválassza – amely a stilisztikai, szövegnyelvészeti kutatások számára nagy előnnyel bír –, s ez a legerősebb stilisztikai marker ebben a korpuszban.

### 4.3 Harmadik kísérlet: plágiumkeresés

Ez a kísérlet eltér az eddigiektől annyiban, hogy a stylo efféle használata nem dokumentált. Érdekes kérdés, hogy használható-e plágiumkeresésre is, ráadásul nem természetes nyelvű szövegeken, hanem forráskódokon, jelen esetben Java nyelven írt házi feladatokon. Az eredmények meglepően jók lettek, a referenciaként használt Sherlock [16] is azt a kettő párt adta valószínű plágiumként, mint a stylo. Emögött az ok valószínűleg az, hogy a plágiumok nem voltak túl szofisztikáltak. Az egyik eleve teljesen azonos másolat volt, ez nem tekinthető sikernek, a másik pedig relatíve egyszerűen azonosítható gépies módosításokat eszközölt a forráskódon. Ennek ellenére is biztató eredmény, hogy a stylo eredményesen működött.




**8. ábra.** *A harmadik kísérlet konszenzusfája.*

## 5 Összefoglalás

Célkitűzésünk egy olyan rendszer megvalósítása volt, amely az informatikában nem jártas kutatók számára is elérhetővé teszi az R nyelven írt stylo programcsomag alkalmazását. Használatához nincs szükség egyéni szoftvertelepítésre és -konfigurációra, mivel a program központi szerveren és ahhoz kapcsolódó böngésző

alkalmazás segítségével működik. A megvalósított programunk azonban többet nyújt egy egyszerű webes felhasználói felületnél.

Egyrészt olyan elemzési módszereket állítottunk össze, amelyek a stylo legfontosabb eszközeihez nyújtanak egyszerűen használható felületet azok bonyolult numerikus paraméterezését mellőzve. Ezek lehetővé teszik a szövegek stilometriai elemzését az R programozási nyelvet és a stylo programcsomagot nem ismerő felhasználók számára is.

Másrészt a rendszert kiegészítettük egy MongoDB-re épülő korpusztárral, amely a stylo fájlalapú korpuszbeviteli megoldásához képes lényegesen rugalmasabb tárolási és menedzselési lehetőségeket kínál. A felhasználók a webes felületen keresztül állíthatják össze a korpuszt, és adhatják meg annak kiválasztott részeit az elemzések számára.

A rendszer működését több kísérletben ellenőriztük történeti szövegek klaszterezése és programszövegek plágiumkeresése céljából.

A megvalósított rendszer nyílt forráskódú (https://github.com/dobijan/shtylo), a DHmine keretrendszer részét képezi (http://dhmine.mit.bme.hu). Közzétételével szeretnénk az egyre nagyobb népszerűségnek örvendő stylo eszköztárra felhívni a hazai kutatók figyelmét. Reményeink szerint az egyszerűen használható webes felülettel hozzájárulunk a stilometriai módszerek hazai elterjedéséhez.

## Köszönetnyilvánítás

A projekt az Európai Unió támogatásával, az Európai Regionális Fejlesztési Alap társfinanszírozásával valósul meg (EFOP-3.6.2-16-2017-00013).

## Bibliográfia

1. Lutosławski, W.: Principes de stylométrie. (1898)
2. Totty, R.N., Hardcastle, R.A., Pearson, J.: Forensic linguistics: The determination of authorship from habits of style. Journal of the Forensic Science Society, 27(1):13–28. (1987)
3. Brennan, M., Afroz, S., Greenstadt, R.: Adversarial stylometry: Circumventing authorship recognition to preserve privacy and anonymity. ACM Trans. Inf. Syst. Secur., 15(3):12:1–12:22. (2012.)
4. Elliott, W.E.Y., Valenza, R.J.: And then there were none: Winnowing the Shakespeare claimants. Computers and the Humanities, 30(3):191–245. (1996)
5. Mosteller, F., Wallace, D.: Inference and Disputed Authorship: The Federalist. Addison-Wesley. (1964)
6. Lutosławski, W.: Sur une nouvelle méthode pour déterminer la chronologie des dialogues de Platon. Paris: H. Welter. (1896)
7. Millican, P.: The Signature Stylometric System http://www.philocomp.net/humanities/signature.htm (2017. november 24.)
8. Juola, P.: Java Graphical Authorship Attribution Program. https://github.com/evllabs/JGAAP (2017. november 24.)

9. Eder, M., Rybicki, J., Kestemont, M.: stylo: Functions for a Variety of Stylometric Analyses, https://cran.r-project.org/package=stylo. (2017. november 24.)
10. Eder, M., Rybicki, J., Kestemont, M.: computationalstylistics/stylo: R package for stylometric analyses https://github.com/computationalstylistics/stylo (2017. november 24.)
11. Eder, M., Rybicki, J., Kestemont, M.: Stylometry with R: a suite of tools. In Digital Humanities 2013: Conference Abstracts, pages 487–489. (2013)
12. Burrows, J.: 'Delta': A measure of stylistic difference and a guide to likely authorship. (2002)
13. Argamon, S.: Interpreting Burrows's delta: geometric and probabilistic foundations. Literary and linguistic computing, 17(3):267–287. (2002)
14. Rstudio: Shiny by RStudio http://shiny.rstudio.com/ (2017. november 24.)
15. MongoDB, Inc.: Introduction to MongoDB https://docs.mongodb.com/manual/introduction/ (2017. november 24.).
16. Department of Computer Science, University of Warwick: Sherlock – Plagiarism Detection Software https://www2.warwick.ac.uk/fac/sci/dcs/research/ias/software/sherlock/ (2017. november 24.)
17. Dobi, J. S.: Shtylo Wiki. https://github.com/dobijan/shtylo/wiki (2017. november 24.)
18. Jack Grieve, Quantitative Authorship Attribution: An Evaluation of Techniques, Literary and Linguistic Computing, Vol. 22, No. 3, 2007. 251–270.
19. *Craig DH, Kinney AF, Shakespeare, Computers and the Mystery of Authorship, Cambridge University Press, Cambridge, UK, 234 (2009)*
20. Kestemont, M., Stover, J., Koppel, M., Karsdorp, F., Daelemans, W. (2016). Authorship Verification with the Ruzicka Metric. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 246-249.
21. van Hulle, D., Kestemont, M. Stylochronometry and the Periodization of Samuel Beckett's Prose. In *Digital Humanities 2016: Conference Abstracts*. Jagiellonian University & Pedagogical University, Kraków, pp. 393-395. (2016)
22. Deeper Delta Across Genres and Languages: Do We Really Need the Most Frequent Words? Literary and Linguistic Computing 1-7, (2011)
23. Margit Kiss, Tamás Mészáros, Creating an extended author's dictionary to support digital literary research In: DH Benelux 2016. Luxembourg (2016)