

19. századi magyar regényeken végzett kísérletek a magyar nyelvű szerzőazonosítás leghatékonyabb távolságméréseinek megtalálására

Timári Mária

ELTE DH

mariazimanyi@btk.elte.hu

Bajzát Tímea Borbála

ELTE DH

timimimi@student.elte.hu

Palkó Gábor

ELTE DH

palko.gabor@btk.elte.hu

In the field of computational stylistics it is a widespread assumption that there exists a unique pattern of a person's language use and this is a so-called authorial fingerprint (Baayen et al. 2002). Identifying the authorial fingerprint can become a base of using quantitative text similarity studies for authorship attribution. Although the metaphor of the "fingerprint" may give the false impression that this pattern can be read from the author's texts in an objective way. The modelling of this "fingerprint" is in fact a creative digital humanities task which constructs a pattern that is based on a selection and combination of dozens of linguistic features that can be interpreted statistically and that can only be interpreted in comparison with other authorial texts.

Considering the size of the corpora and the complexity of text characteristics and similarity calculations, the method cannot be done without the use of computer algorithms. There are already some softwares (e.g. R-Stylo, JGAAP, Websty) that apparently offer an accessible way to researchers for analyzing texts, (however these methods limit the process of searching for patterns), while on the other hand it is possible to implement the calculations using custom program codes.

In our research we endeavour to find the most efficient measures in authorship attribution for Hungarian texts.

1. Bevezetés

A számítógépes stilsztika területén széles körben elfogadott nézet, hogy léteznek az egyes ember nyelvhasználatát jellemző egyedi mintázatok, szerzői „ujjlenyomatok”¹, amelyek feltérképezése megalapozhatja a kvantitatív szöveghasonlósági vizsgálatok felhasználását szerzőazonosítás céljaira.

Az „ujjlenyomat” metaforája ugyanakkor azt a téves képzetet keltheti, hogy a szerzőre jellemző ésszámszerűsíthető mintákaszövegekből objektív módon volnának kiolvashatók.

1 Harald et al., 2002.



Az „ujjlenyomat” megalkotása valójában egy olyan kreatív digitális bölcsészeti feladat, amely a szövegstatistikailag értelmezhető több tucatnyi nyelvi jellemző, majd az ezekre épülő különféle hasonlósági számítások szelekciója és kombinációja alapján konstruál egy, mindig csak más szerzői szövegekkel való összehasonlítás horizontján értelmezhető „mintázatot”.

Mindez – tekintettel a szövegtörzsek méretére, a szövegjellemzők és a hasonlósági számítások komplexitására – ma már elképzelhetetlen számítógépes algoritmusok használata nélkül. Léteznek (látszólag) kész megoldásokat nyújtó, a kutatók számára könnyebben elérhető, de a mintázatkeresés folyamatát leszűkítő szoftveres megoldások (pl. R-Stylo, Websty, JGAAP), másrésztől lehetséges a számítások implementálása egyedi programkódok segítségével is.

Kutatásunkban arra vállalkozunk, hogy 19. századi magyar regényeken vizsgálva – előtérbe helyezve Jókai Mór szövegterének vizsgálatát – megtaláljuk azokat a beállításokat, illetve távolságmértékeket, melyek a magyar nyelvű szerzőazonosítási és klaszterezési eljárások során a legpontosabb eredményekkel szolgálnak.

2. A vizsgálatokhoz felhasznált (al)korpuszok

A legelső és a legfontosabb lépés annak irányába, hogy felvázolhassuk azon nyelvhasználati és stilisztikai relevanciákat, amelyek feltehetően jellemzik Jókai Mór írói ujjlenyomatát, a vizsgálati korpuszok létrehozása. Ezen feladat nem nélkülözheti a széleskörű és irányított mintavételt Jókai Mór életművéből, illetőleg kiemelt fontosságú a referenciaanyag összeállítása is, amely segítségével komparatív méréseket tudunk végezni, s ezek tükrében döntést hozni abban, hogy a 19. századi magyar prózairodalomra vonatkozóan mely paraméterbeállítások lesznek azok, amelyek valószínűsíthetően képesek eredményt felmutatni a szerzőazonosítás céljából elvégzett vizsgálatokban.

Az első kialakított vizsgálati korpuszunk a Jókai korpusz volt, amely összeállítása során törekedtünk arra, hogy Jókai Mór életművéből minél teljesebb gyűjtemény kerüljön bele a korpuszba. Az elsődleges forrása a Jókai Mór által írt regényeknek az Unikornis Kiadó Jókai sorozata volt.² Mivel ezen kiadássorozat nem tartalmazta Jókai Mór teljes életművét, ezért bővítenünk kellett a forrásainkat, a korpuszba kerülő többi regény forrása így a Révai Testvérek kiadó kiadványai lettek.³ Végül a két kiadássorozat alapján összeállított korpusz anyagát összevetettük a Magyar Országos Közös Katalógus (MOKKA)⁴ listájával (a katalógusban Jókai Mór névvel indítottunk keresést). Az így kapott lista alapján a korpuszunkat bővítettük azokkal a prózai munkákkal, amelyek a két kiadássorozatban nem szerepeltek, viszont szabadon hozzáférhetőek voltak interneten, tehát a Magyar Elektronikus Könyvtár gyűjteményéből vételeztük ezen szövegeket.⁵ Összesítve a Jókai korpuszunk 66 darab regényt, 6 darab novellát és 14 darab más prózakötetet tartalmazott.

2 *Jókai Mór munkái - gyűjteményes díszkiadás.* Unikornis Kiadó. Budapest. 1992–2004.

3 *Jókai Mór összes művei.* Révai Testvérek Irodalmi Intézet. Budapest. 1894–1898.

4 <http://www.mokka.hu/> (hozzáférés: 2021.05.24.)

5 <https://mek.oszk.hu/> (hozzáférés: 2021.05.24.)

A referencia korpusz összesen 132 regényt tartalmaz, a szövegek forrása a Magyar Elektronikus Könyvtár, ennek a korpusznak egy része az ELTE DH Regénykorpusz gyűjteményében is megtalálható.⁶ A referencia korpusz összeállításánál az volt a kiindulópontunk, hogy olyan szabadon elérhető regényeket keressünk, amelyek első kiadása az 1846 és 1907 közötti időszakra esik, mert ezen időszak párhuzamos a Jókai Mór életmű keletkezési idejével. A 132 regény összesen 56 szerzőtől származik. Mind a Jókai korpusz mind pedig a referencia korpusz szövegeinek tárolási- és munkaformátuma a TXT formátum.

Azonban önmagában ez a két korpusz nem bizonyult alkalmasnak a mérések beállítására azért, mert a referencia korpusz teljes anyagának jelentős része olyan szövegkiadásokból áll, amelyek a 19. századi helyesírási és szövegkiadási hagyományokat tükrözik, míg a Jókai korpuszba kerülő művek zöme (Unikornis Kiadás sorozata például) a helyesírást tekintve a mai sztenderdet követi, ez pedig a nyelvstatisztikai mérések eredményeit s azok hitelességét alapvetően megkérdőjelezheti. Ezen hipotézis tesztelésére további alkorpuszokat hoztunk létre. Az első volt egy szűkített regénykorpusz, amely Jókai Mór három regényét tartalmazta, illetve négy szerzőtől további 12 regényt. Olyan szerzőket választottunk ebbe az alkorpuszba, akik életműve kanonikusként tartható számon a magyar irodalom- és szövegkiadás történetben, így feltételezhető volt az, hogy a MEK-en olyan szövegkiadásai is elérhetők, amelyek már a XX. század második felére datálhatók, így a helyesírásukat tekintve nem különböznek jelentős mértékben a Jókai Mór szövegek modernebb kiadványaitól. Az így kiválogatott szerzők és műveik a következők: Mikszáth Kálmán (Beszterce ostroma;⁷ A beszélő köntös;⁸ Új Zrínyiász⁹), Kemény Zsigmond (Férj és nő;¹⁰ Gyulai Pál;¹¹ Az özvegy és leánya¹²), Gárdonyi Géza (Az isten rabjai;¹³ A világyáró angol;¹⁴ A lámpás¹⁵), Eötvös József (Magyarország 1514-ben;¹⁶ A karthauzi,¹⁷ A falu jegyzője¹⁸). Jókai Mórtól pedig Az arany ember,¹⁹ A lőcsei fehér asszony²⁰ és a Szomorú napok²¹ című regény került a szűkített korpuszba.

A nagyobb korpusz, amelyben a vételezett szövegek kiadási évétől eltekintettünk, 69 szöveget tartalmazott. 3 Jókai Mór által írt művet és 66 másik regényt 32 szerzőtől a

6 <https://regenykorpusz.elte-dh.hu/> (hozzáférés: 2021.05.24.)

7 Mikszáth Kálmán, *Beszterce ostroma*, Arcanum, Budapest, 1998.

8 Mikszáth Kálmán, *A beszélő köntös*, Arcanum, Budapest, 1998.

9 Mikszáth Kálmán, *Új Zrínyiász*, Arcanum, Budapest, 1998.

10 Kemény Zsigmond, *Férj és nő*, Unikornis Kiadó, Budapest, 1996.

11 Kemény Zsigmond, *Gyulai Pál*, Szépirodalmi Könyvkiadó, Budapest, 1967.

12 Kemény Zsigmond, *Az özvegy és leánya*, Európa, Budapest, 2000.

13 Gárdonyi Géza, *Az Isten rabjai*, Szépirodalmi Könyvkiadó, Budapest, 1974.

14 Gárdonyi Géza, *A világyáró angol*, Rozsnyai Kiadó, Budapest, 1905.

15 Gárdonyi Géza, *A lámpás*, Irod. és Műv. Kiadó, Bukarest, 1957.

16 Eötvös József, *Magyarország 1514-ben*, Szépirodalmi Könyvkiadó, Budapest, 1972.

17 Eötvös József, *A karthauzi*, Franklin társulat, Budapest, 1911.

18 Eötvös József, *A falu jegyzője*, Szépirodalmi Könyvkiadó, Budapest, 1971.

19 Jókai Mór, *Az arany ember*. Unikornis Kiadó, Budapest, 2001.

20 Jókai Mór, *A lőcsei fehér asszony*, Unikornis Kiadó, 2001.

21 Jókai Mór, *Szomorú napok*, Unikornis Kiadó, 2001.



referencia korpuszunkból. Az így létrehozott két alkorpusz lehetőséget biztosított arra, hogy teszteljük azt, hogy mi történik akkor a szerzőazonosításra alkalmazott kódok és paraméterbeállítások alkalmazásánál, amikor a vizsgált szövegek kiadásának évét (és ezen keresztül a helyesírási hagyományt) nem vesszük figyelembe.

A korpuszok létrehozásánál kiemelkedően fontos szemponttá vált a duplumok szűrése, hiszen például a Jókai Mór kötetek korpuszba való illesztésénél előfordult az, hogy ismétlődő szövegeket azonosítottunk. A célunk az volt a duplumszűrésnél tehát, hogy minden szöveg lehetőleg csak egyszer forduljon elő a vizsgálatra kialakított szöveggyűjteményeinkben.

Az ismétlődő szövegek felbukkanásának alapvetően két oka volt. Az első, hogy különböző novellák, elbeszélések nem egy, hanem akár több kötetnek is részét képezték, vagyis ilyenkor az egész novella, illetve elbeszélés egynél többször szerepelt a korpuszban. Ezekben az esetekben a duplumszűrés segítségével megkerestük az ismétlődő szövegeket, s csupán egyszer hagytuk meg őket a korpuszban. A másik lehetőség a szövegek ismétlődésére az intratextualitás, azaz a szerző szövegei közti kapcsolat volt, ugyanis jellemzően az útleírások és életrajzi írások tartalmaztak olyan szövegrészt, melyek egy-egy regényben, elbeszélésben is felbukkantak. Ezek ugyanakkor jórészt csupán egy-egy bekezdésnyi szöveget jelentettek, s olykor nem is teljesen pontosan voltak átvéve, némi módosítás történt rajtuk. Ez utóbbi esetekben tehát a korpuszban hagytuk az ismétlődő részeket, egyrészt, mert az intratextualitást nem hagyhatjuk figyelmen kívül teljes mértékben a szerzőségi vizsgálatban, másrészt, mivel a szűkített korpuszból kihagytuk az útleírásokat, valamint az önéletrajzi írásokat, tekintve, hogy a referenciakorpusz is jobbára csak fikciós műveket tartalmaz.

3. A kutatás módszere

A korpuszok összeállítása után a vizsgálatot a Stylo-tól ellépve, pythonban végeztük. A Stylo csomag egy R-ben - mely egy programozási nyelv és szoftverkörnyezet statisztikai számításokhoz és ábrázoláshoz - kialakított függvénykönyvtár, mely a stilometriai elemzések során gyakran használt funkciókat tartalmazza, beleértve a szerzői hozzárendelést és a műfaji felismerést.

A python egy általános célú, programozási nyelv, melyet az 1980-as évek végén, eredetileg oktatási célra alakítottak ki. Szövegelemzési célokra is kiválóan alkalmazható, így ma a digitális bölcsészetben gyakran használt programnyelv.

A Mike Kestemont által készített, ám 5 éve nem frissített, python programnyelvre átírt Stylo implementációt, a Pystyl-t, valamint egy teljes dokumentációjú kódot, melyben a szöveg és a végrehajtható kód váltja egymást, a Pydeltát felhasználva lefuttattunk a korpuszokon több száz beállítást, hogy lássuk, melyek a leghatékonyabbak a 19. századi magyar szövegek klaszterezésére, vagyis adattömbök homogén csoportokba sorolására, klasszifikálására.

A Stylo-tól való ellépés egy igen fontos lépés volt, méghozzá több szempontból. Először is azért, mert ezt megelőzően csupán manuálisan, a különböző paramétereket egyesével beállítva végeztünk hasonló vizsgálatokat a Stylo-ban.

Továbbá, bár a Stylo kódjaihoz is hozzá lehet férni, úgy véltük, az ellenőrizhetőség, a megismételhetőség és az átláthatóság kedvéért célszerű ezeket a kódokat pythonba implementálni, s ott végezni ezeket a vizsgálatokat. A nyílt tudomány elveinek mentén a futtatott kódokat, valamint a korpuszokat természetesen közzé tesszük.

Harmadrészt az indokolta a Stylo-tól való ellépést, hogy olyan beállításokat is kívántunk alkalmazni, melyeket a stylo nem tud. A Stylo-ban ugyanis nincs lehetőség az általunk vizsgált egyik távolságmérték, a Burrows delta használatára, továbbá a Feature-ek keverése sem megoldható.

Feature alatt azokat a megszámlálható jellemzőket értjük, melyeket automatikusan kivonunk a korpuszból, majd ezeket változóként használjuk a további elemzéshez. A stylo alapbeállításában ezek a korpuszokban leggyakrabban előforduló szövegelemek, azaz az MFW lista. Ezek lehetnek akár szó vagy karakter n-gramok. Az n-gram a számítógépes nyelvészetben olyan összefüggő/folyamatos szekvencia, amely n elemből áll (tehát például szó-bigram esetében ez azt jelenti, hogy 2 egymást követő lexéma).

Bizonyos esetekben azonban van értelme személyre szabott funkciók halmazát használni, pl. azokat a szavakat, amelyek érzelmekkel társulnak, vagy mondjuk a funkciószavak meghatározott rész-halmazát, erre ugyanakkor a Styloban szintén nincs lehetőség.

Az előbbieken említett MFW rövidítés a „Most Frequent Words” szóhármast takarja, habár pontosabb lenne szavak helyett szövegelemről beszélni. Az MFW mint paraméter azért bír kiemelt jelentőséggel, mert nagymértékben hatással lehet az elemzés kimenetelére, és mindemellett a megfelelő érték megtalálása is összetett feladat, például függ a szövegek hosszától.

Szintén fontos paraméter a **Culling**. Ez egy százalékos érték. Az itt megadott szám jelentése, hogy az adott szónak/szövegelemnek a korpuszban található szövegek hány százalékában kell előfordulnia ahhoz, hogy felhasználásra kerüljön az elemzés során. Ez a paraméter, az MFW-hez hasonlóan nehezen megjósolható, habár a szövegek hossza és a szövegek szóhalmazai közti átfedés elég jól meghatározza az optimális mértéket.

Az előzetes, manuálisan végzett vizsgálatok eredményei alapján úgy tűnt, a 19. századi magyar szövegek esetén a legpontosabb eredményeket a Cosine (Koszinus) Távolság adja, szó-bigrammokon futtatva, 1000-es MFW listával, Culling és Sampling nélkül. Habár n-dimenziós koordináták közti távolságok kiszámítására rendkívül sokféle módszer létezik, az előzetes eredményekből, továbbá Christof Schöch egy korábbi vizsgálatából kiindulva három távolságmértéket választottunk ki a vizsgálatához.²²

22 Baayen et al., 2017

Lásd még a kutatás során felhasznált kódokhoz <https://github.com/christofs/stylometry-bibliography> (hozzáférés: 2021.05.27.)



Burrows Deltája nagy előrelépést jelentett az eddig használatos távolságmértékekhez képest, és mind a mai napig az egyik leggyakrabban használt távolságmérték. Gyakran használják más módszerekkel való összehasonlítás céljára.

$$\Delta_{(A,B)} = \frac{1}{n} \sum_{i=1}^n \left| \frac{A_i - \mu_i}{\sigma_i} - \frac{B_i - \mu_i}{\sigma_i} \right|$$

1. ábra

Eder Deltája egy módosított verziója a Burrows Deltájának. A gyakoribb szavak súlyozását megnöveli valamelyest, a ritkébbakét pedig lecsökkenti.

$$\Delta_{(A,B)} = \frac{1}{n} \sum_{i=1}^n \left(\left| \frac{A_i - B_i}{\sigma_i} \right| \times \frac{n - n_i + 1}{n} \right)$$

2. ábra

A Koszinusz Távoóság is egy gyakran használt, jól ismert mértékegység. Két N dimenziójú vektor közötti hasonlóság mértékét jelenti a vektorok közötti szög koszinuszának megállapításával.

$$1 - \frac{\sum_{i=1}^n A_i B_i}{\sqrt{\sum_{i=1}^n A_i^2} \cdot \sqrt{\sum_{i=1}^n B_i^2}}$$

3. ábra

Ezekkel a távolságmértékekkel dolgoztunk tehát, és a lehető legtöbb beállítást futtattuk le, még hozzá egy kisebb és egy nagyobb korpuszon is, a már említett 15 valamint 69 szöveget tartalmazó korpuszokon. A következő fejezetben (4.) az ábrán különböző beállításokkal való futtatás eredményei láthatók.

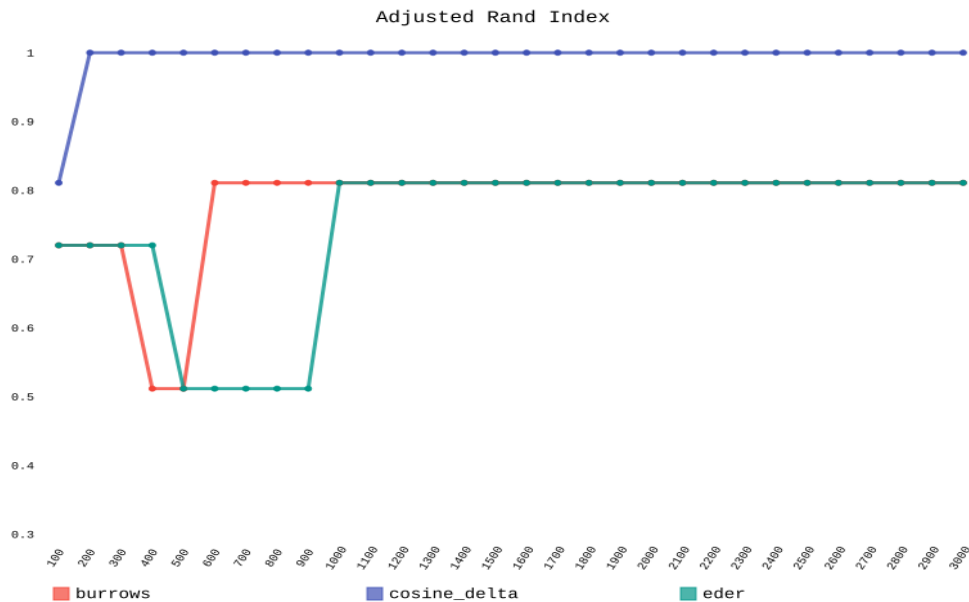
4. A vizsgálatban tesztelt paraméterbeállítások és függvények eredményei²³

Az ábrákon a függőleges tengely 0-tól 1-ig van beosztva, s az 1 jelenti a maximum értéket, vagyis azt, hogy a függvény maximális pontossággal tudta meghatározni a szövegek szerzőjét, és az összes művet hibátlanul csoportosította a szerzőség alapján.

A vízszintes tengelyen az MFW (Most Frequent Word) lista értékének növekedése látható, 100-as kezdőértéssel, egészen 3000-ig, 100-as lépésközzel. Ez azt jelenti, hogy minden egyes beállításról 30 különböző MFW listán való lefutás eredményét tudjuk megmondani. Az adott szám azt jelenti, hogy hány leggyakoribb szövegelemét tartalmazza a lista a korpusznak.

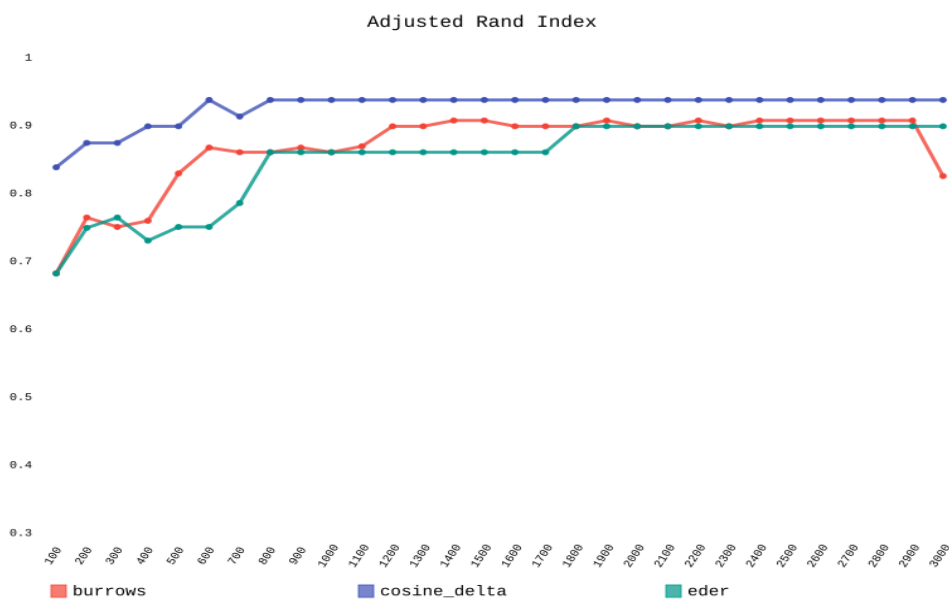
²³ A kutatáshoz alkalmazott kódok pythonba történő implementációját Nemeskey Dávid végezte el.

Ezen az ábrán egy olyan lefutás eredménye látható, mely a 15 szöveget tartalmazó korpuszon ment végbe, karakter-trigrammokon, 10%-os cullinggal.



4. ábra

Szembevetendő, hogy a Koszinusz Távolság már 200-as MFW listánál eléri maximumértékét (4. ábra), s ez után a vizsgált MFW intervallumon belül képes maximális pontossággal a szerzők csoportosítására, míg a másik két függvénynek 300, illetve 400 MFW-nél a hatékonysága csökken, s csak 600, illetve 1000 MFW esetén ível fel újra, azonban még ekkor sem érik el azt a teljesítményt, amit a Koszinusz.

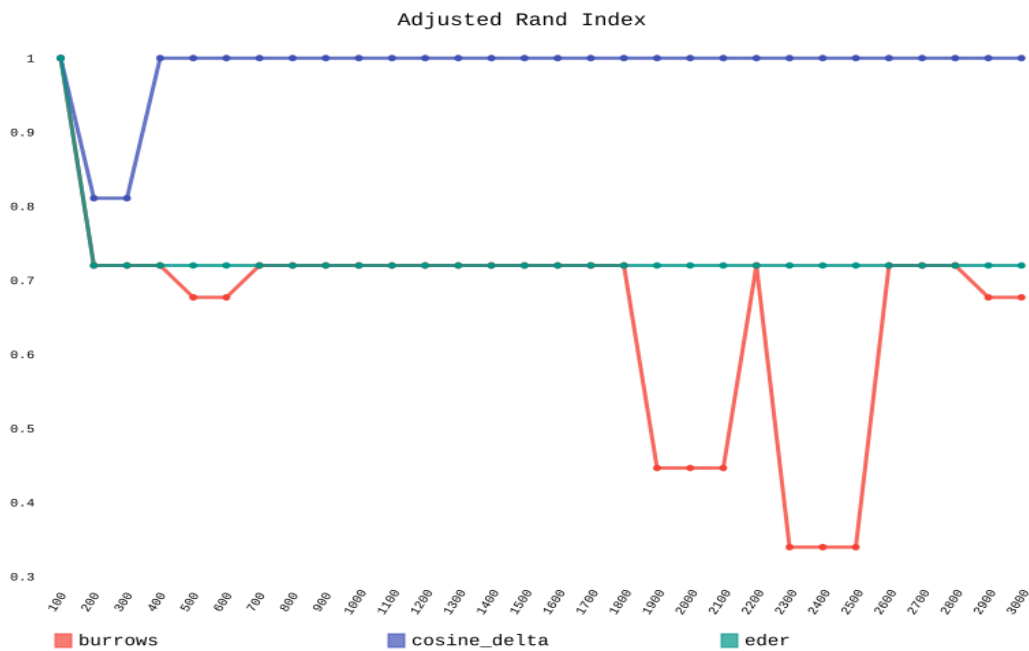


5. ábra

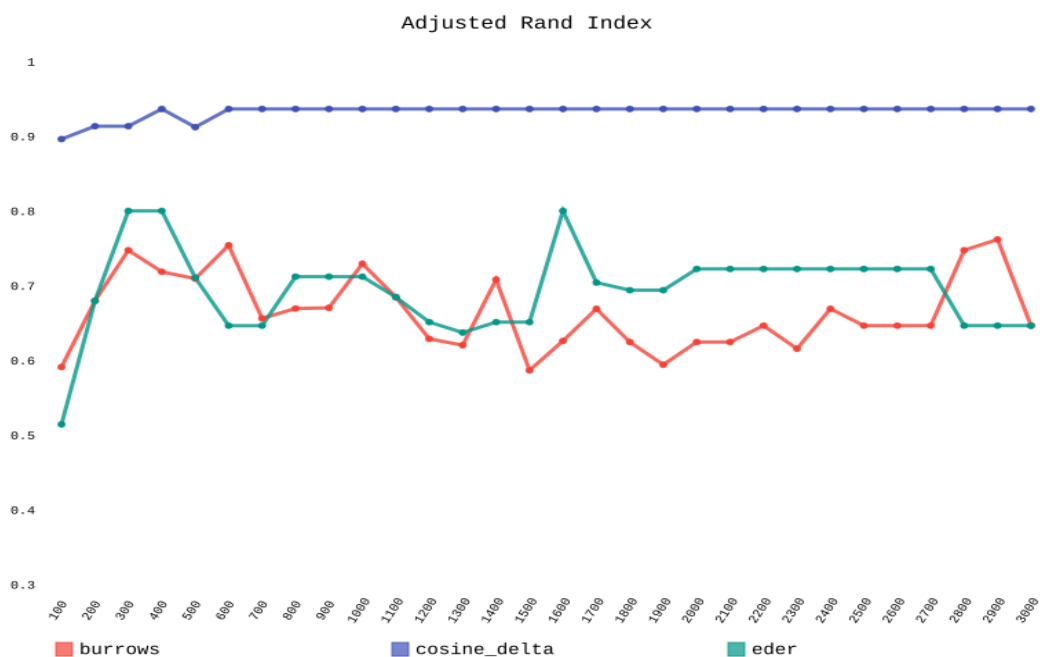


Az 5. ábrán ugyanezen paraméterek eredményei láthatóak a 69 szövegből álló korpuszon futtatva.

Megjegyzendő, hogy míg a 15 szövegen végzett vizsgálat során igen, addig a 69 szövegen végzett vizsgálat során még a Koszinusz Távolság sem érte el a maximum hatékonyságot, azaz nem tudta hibátlanul klaszterezni a szövegeket, azonban ahogy a többi ábrán is látszik, a másik két függvényhez képest így is kiemelkedő teljesítményt nyújtott.



6. ábra



7. ábra

A 6.-7. ábra a szó-bigrammokon, szintén 10%-os Cullinggal futtatott eredményeket mutatja mind a 15 szöveges, mind a 69 szöveget tartalmazó korpusz esetén. Látható az, hogy a szó-bigrammok esetén szintén a Koszínusz Távolság függvény az, amely az általunk alkalmazott paraméterbeállítások mentén a legmagasabb pontossággal képes klasszifikálni a vizsgált szöveganyagot.

5. Összefoglalás

Vizsgálatunk eredményeiből az látszik, hogy a Koszínusz Távolság bizonyult a leghatékonyabb távolságmérésnek a 19. századi magyar nyelvű szövegek szerzőazonosítási, illetve klaszterezési eljárásai során. A különböző beállítások közül a 3-as és 4-es karakter n-gramok, valamint az 1-es és 2-es szó n-gramok adták a legpontosabb eredményeket. Ez utóbbiakat azonban fenntartással kezeljük abból a megfontolásból, hogy a magyar nyelv esetén a szó n-gramok használata torzíthatja a vizsgálatot, mivel a regény tematikája helyeződik előtérbe. Ennek kiküszöbölésére szolgál majd az a további vizsgálatunk, mely során az e-magyar²⁴ emtsv verziójának segítségével a szövegek szemantikai tartalmú szavait (a főneveket, melléneveket, valamint az igéket) poszttag-ekre (morfológiai elemzést tartalmazó címkékre) cseréljük, s az így kapott szövegeken futtatjuk újra a leghatékonyabbnak bizonyult beállításokat. Így egy tematikafüggetlen módszerrel egészítenénk ki ezen kutatásunkat.

Szintén további kutatási terveink közé tartozik, hogy Jókai Mór szövegeinek nyelvezetéről egy szélesebb körű stilometriai elemzést készítsünk, mely a Jókaira jellemző nyelvstatisztikai alapú mintázatok feltérképezését jelenti, vagyis megkíséreljük megkonstruálni Jókai Mór szerzői „ujjlenyomatát”. Ehhez kapcsolódik az a szerzőséggel foglalkozó vizsgálat, amelyet a Jókai Kritikai Kiadást készítő kutatócsoporttal együttműködve végzünk. A vizsgálat során stilometriai módszerek alkalmazásával teszünk kísérletet arra, hogy a kutatók által kétes hitelűnek tekintett szövegekről eldöntsük, vajon a konstruált „ujjlenyomat” alapján a biztos szerzőségű Jókai szövegekhez mennyire állnak közel a nyelvstatisztika perspektívájából.

6. Források és bibliográfia

6.1. Források

A tanulmányban idézett képletekhez lásd:

<http://tdk.bme.hu/VIK/DownloadPaper/Szovegelemzesi-modszerek-automatizacioja>
(utolsó elérés: 2021.05.27.)

A kutatás során felhasznált kódok: <https://github.com/christofs/stylometry-bibliography>

Jókai Mór összes művei. Révai Testvérek Irodalmi Intézet. Budapest. 1894–1898.

Jókai Mór munkái – gyűjteményes díszkiadás. Unikornis Kiadó. Budapest. 1992–2004.

²⁴ Indig et al., 2019.



6.2. Bibliográfia

Harald Baayen, Hans van Halteren, Anneke Neijt, Fiona Tweddle, “An experiment in authorship attribution” Conference Paper (JADT 2002: 6es Journées internationales d’Analyse statistique des Données Textuelles) 2002.

Indig Balázs, Sass Bálint, Simon Eszter, Mittelholcz Iván, Kundráth Péter, Vadász Noémi, „emtsv – egy formátum mind felett”, In *XV. Magyar Számítógépes Nyelvészeti Konferencia*, szerkesztette Berend Gábor, Gosztolya Gábor, Vincze Veronika, 235–247. Szeged: Szegedi Tudományegyetem TTIK, Informatikai Intézet. 2019.

Stefan Evert, Fotis Jannidis, Thomas Proisl, Isabella Reger, Steffen Pielström, Christof Schöch, Thorsten Vitt, “Understanding and explaining Delta measures for authorship attribution”, *Digital Scholarship in the Humanities*, 2017(2): ii4–ii16. hozzáférés: 2021.05.27. doi: [10.1093/llc/fqx023](https://doi.org/10.1093/llc/fqx023)