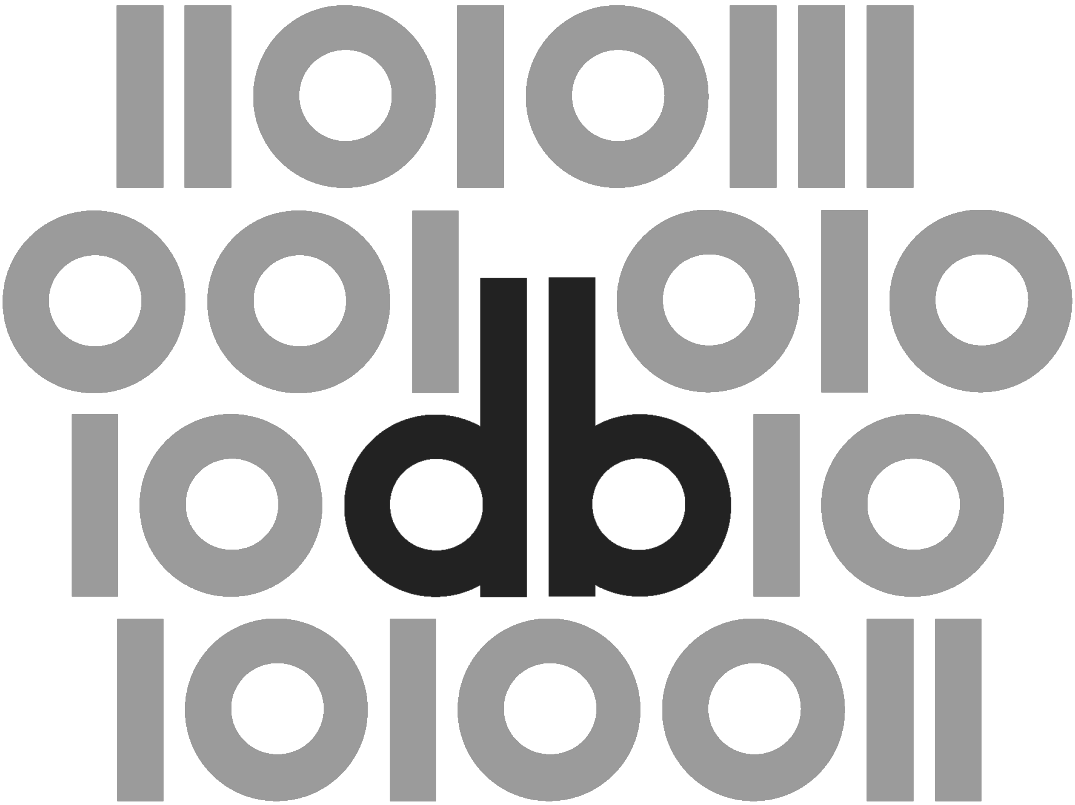
5 (2021) <DIGITÁLIS BÖLCSÉSZET>

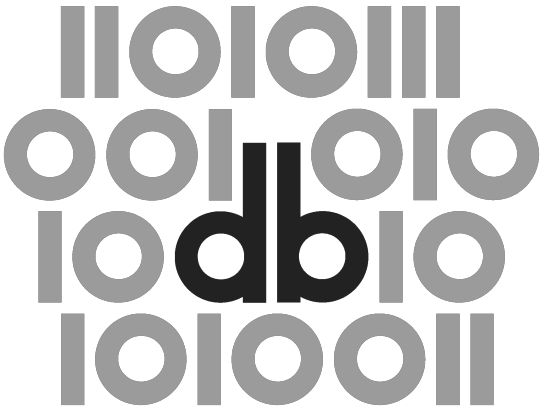
# A krakkói Computational Stylistics Group (Különszám)




5 (2021) </DIGITÁLIS BÖLCSÉSZET>

# Digitális Bölcsészet
# 2021., ötödik szám
## A krakkói Computational Stylistics Group (Különszám)

# <DIGITÁLIS BÖLCSÉSZET>




5 (2021)

A krakkói
Computational Stylistics Group

(Különszám)

**A különszámot Szemes Botond szerkesztette.**

**Felelős szerkesztő:**
Maróthy Szilvia

**Szerkesztőség:**
Kokas Károly, Parádi Andrea

**Rovatvezetők:**
*Tanulmányok:* Kiss Margit
*Műhely:* Péter Róbert
*Kritika:* Almási Zsolt
*Labor:* Mártonfi Attila

**Tanácsadó testület:**
Bartók István, Fazekas István, Golden Dániel, Horváth Iván, Palkó Gábor, Pap Balázs, Sass Bálint, Seláf Levente

**Korábbi munkatársaink:**
Bartók Zsófia Ágnes (szerkesztő, rovatvezető), Fodor János (szerkesztő), †Labádi Gergely (szerkesztő, rovatvezető), †Orlovszky Géza (tanácsadó testület)

## ISSN 2630-9696
## DOI 10.31400/dh-hun.2021.5

Kiadja a Bakonyi Géza Alapítvány és az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A).
Felelős kiadó az ELTE BTK Régi Magyar Irodalom Tanszék vezetője.

Megjelenik az Open Journal Systems (OJS) v. 3. platformon, melynek működtetését az ELTE Egyetemi Könyvtár- és Levéltár biztosítja.

A különszám megjelenését a Wacław Felczak Alapítvány támogatta.



Honlap: `http://ojs.elte.hu/digitalisbolcseszet`
Email cím: dbfolyoirat@gmail.com

Olvasószerkesztő: Bucsics Katalin
Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

# Tartalom

**Greta Franzini** 0000-0003-1159-5575

*Georg-August-Universität Göttingen*

greta.franzini@eurac.edu

**Mike Kestemont** 0000-0003-3590-693X

*Universiteit Antwerpen*

mike.kestemont@uantwerpen.be

**Gabriela Rotari**

*Georg-August-Universität Göttingen*

gabriela.rotari@gmail.com

**Melina Jander** 0000-0003-1646-6836

*Georg-August-Universität Göttingen*

jander@sub.uni-goettingen.de

**Jeremi K. Ochab** 0000-0002-7281-1852

*Uniwersytet Jagielloński w Krakowie*

jeremi.ochab@uj.edu.pl

**Emily Franzini**

*Georg-August-Universität Göttingen; Decoded Ltd., London*

**Joanna Byszuk** 0000-0003-2850-2996

*Instytut Języka Polskiego PAN*

joanna.byszuk@ijp.pan.pl

**Jan Rybicki** 0000-0003-2504-9372

*Uniwersytet Jagielloński w Krakowie*

jkrybicki@gmail.com

# Szerzőazonosítás Jacob és Wilhelm Grimm zajos, digitalizált levelezésében*

Az alábbi cikk egy multidiszciplináris projekt eredményeit mutatja be, amely a különböző digitalizációs stratégiák számítógépes szöveganalízisben való használ-

* Eredeti megjelenés: Greta Franzini, Mike Kestemont, Gabriela Rotari, Melina Jander, Jeremi K. Ochab, Emily Franzini, Joanna Byszuk and Jan Rybicki, „Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm," *Frontiers in Digital Humanities* 5 (2018), https://doi.org/10.3389/fdigh.2018.00004.

hatóságát járja körül. Pontosabban Jacob és Wilhelm Grimm szerzőségének automatizált megkülönböztetésére tettünk kísérletet, melyet egy HTR (Handwritten Text Recognition – kézzel írott szöveg felismerése) és OCR (Optical Character Recognition – optikai karakterfelismerés) által feldolgozott levelezéskorpuszban hajtottunk végre, korrekció nélkül – felmérve, hogy az így keletkezett zaj milyen hatással van a fivérek különböző írásmódjának azonosítására. Összegezve, úgy tűnik, hogy az OCR megbízható helyettesítője lehet a manuális átírásnak, legalábbis a szerzőazonosítás kérdéskörét illetően. Eredményeink továbbá abba az irányba mutatnak, miszerint még a különböző digitalizációs eljárásokból származó tanító- és tesztkorpuszok *(training and test set)* is használhatók a szerzőazonosítás során. A HTR-t tekintve a kutatás azt demonstrálja, hogy ez az automatizált átírás ugyan az OCR-hez képest szignifikánsan növeli a szövegek félrecsoportosításának veszélyét, ám körülbelül 20% feletti tisztaság már önmagában elegendő ahhoz, hogy a véletlennél nagyobb esélye legyen a helyes bináris megfeleltetésnek.

Kulcsszavak:
stilometria, szerzőazonosítás, német irodalom, Grimm, digitalizáció, OCR, HTR

## 1. Bevezetés

Több tanulmány is beszámol arról, hogy adatelemzéssel foglalkozó kutatók a kutatási idejüknek akár 80%-át is az adatok előkészítésével tölthetik, míg csak 20%-ot szentelnek magukra a kutatási kérdésekre.[1] Ez az egyensúlyhiány azt sugallja, hogy a kutatók abban a hitben dolgoznak, hogy az előkészítésbe fektetett idő egyenesen arányos az eredmények minőségével – más szavakkal: a magas minőségű munka összeegyeztethetetlen az alacsony minőségű, zajos adatokkal.

Jelen tanulmány erre a jelenségre adott válasznak is tekinthető: olyan kutatásokra épül, melyek a számítógépes szerzőazonosításban elfogadható mértékű digitalizációs zajokat járják körül,[2] és arra törekszik, hogy megbízható modellt adjon Jacob és Wilhelm Grimm[3] kézírásának átírására, illetve hogy képes legyen meghatározni szerzőségüket levelezésük alapján.[4] Ennek érdekében elemzéseket futtatunk le a szövegek nyomtatott kiadásának OCR-rel feldolgozott és nem korrigált változatain, valamint a

[1] Például Hadley Wickham, „Tidy Data," *Journal of Statistical Software* 59, 10. sz. (2014), https://doi.org/10.18637/jss.v059.i10.

[2] A téma 2015 óta az EMNLP éves workshopjának is tárgya. Hozzáférés: 2021.07.30, https://noisy-text.github.io/2017/.

[3] Jacob (1785–1863) és Wilhelm Grimm (1786–1859) német írók, tudósok és akadémikusok, akik a 19. század folyamán népmesék gyűjtésével és publikálásával váltak ismertté.

[4] A cikk egy, a németországi Göttingeni Egyetem hat hónapos kísérleti projektjének keretében végzett kutatásokat ismerteti. A TrAIN (Tracing Authorship In Noise) néven ismert projekt a zajos OCR- és a HTR-adatok számítógépes szövegelemzésére gyakorolt hatását kívánta vizsgálni. A projekt 2016 júliusa és 2017 januárja között zajlott. A projekt weboldala, hozzáférés: 2021.07.30 http://www.etrap.eu/research/tracing-authorship-in-noise-train/.

 DOI 10.31400/dh-hun.2021.5.3144

kézzel írt dokumentumok digitalizált képeiből készített, szintén korrigálatlan HTR-verziókon egyaránt. A két verzió alapvetően eltérő feldolgozási zajokat emel a diskurzusba, számunkra hasznos összehasonlítási szempontokat biztosítva. E tanulmányban ugyanis a következő kérdést tesszük fel: az OCR és a HTR során keletkezett zaj milyen mértékben befolyásolja a Grimm testvérek egyéni stílusának *(stylome)*[5] azonosítását? A munkával a jelen technológia lehetőségeit térképezzük fel, miközben nagyobb rálátást kívánunk biztosítani Jacob és Wilhelm Grimm stílusára is. Akadályként a következő tényezőkkel kell számolnunk: az eredeti szövegek digitális átalakítása során keletkezett textuális zajok torzító hatása, valamint a rendelkezésünkre álló adatok sokszínűsége és mennyisége. A tanulmány szerkezete az alábbiak szerint alakul: az 1. és 2. rész a projekt motivációját tárgyalja, a 3. részben a kutatás alapját képező anyagról lesz szó, míg a 4. az anyag digitalizációját és a nem korrigált szövegeken végzett szerzőazonosítás lépéseit mutatja be, továbbá itt tárgyaljuk a kutatás eredményeit is. Végül az 5. rész az összegzés és további kitekintések megtétele számára biztosít teret.

## 2. Kapcsolódó kutatások

Az online elérhető szövegekkel dolgozó kutatóknak gyakran zajos vagy strukturálatlan adatokkal kell megküzdeniük. Egészen pontosan kétféle zaj nehezíti munkájukat: 1. amely a szöveg előállításakor keletkezik (úgy mint: helyesírási hiba, a standardtól eltérő szóalak, speciális karakterek, szándékolt rövidítések, nyelvtani hibák stb.); 2. amely a szöveg más formátumba történő konvertálása során (úgy mint: digitalizáció vagy digitális transzformáció) jön létre.[6] Az utóbbi típus rendszerint az interneten fellelhető irodalmi szövegeknél, az OCR-rel vagy HTR-rel feldolgozott nyomtatott vagy kézírásos művek esetében gyakori. Ezek kapcsán az alábbiak mondhatók el: a történeti szövegeknél az OCR pontossága a karakterek szintjén a 95%-ot is meghaladhatja,[7] bár a forrás típusától függően ez a szám lehet alacsonyabb is (klasszikus szövegek kritikai kiadása versus ősnyomtatványok), a HTR esetében pedig már 80–90% közé esik, a kézírás tisztaságától függően.[8] Lopresti az OCR-hibáknak az információvisszakeresésre (Information Retrieval – IR) és a természetes nyelvfeldolgozásra (Natural Language Processing – NLP) gyakorolt hatásával is foglalkozik,[9] és habár léteznek a zaj

[5] Hans van Halteren, Harald Baayen, Fiona Tweedie, Marco Haverkort and Anneke Neijt, „New Machine Learning Methods Demonstrate the Existence of a Human Stylome," *Journal of Quantitative Linguistics* 12, 1. sz. (2005): 65–77, https://doi.org/10.1080/09296170500055350.

[6] L. Venkata Subramaniam, Shourya Roy, Tanveer A. Faruquie and Sumit Negi, „A Survey of Types of Text Noise and Techniques to Handle Noisy Text," in *Proceedings of The Third Workshop on Analytics for Noisy Unstructured Text Data – AND '09,* 115–122 (Barcelona: ACM Press, 2009), 115, https://doi.org/10.1145/1568296.1568315.

[7] Florian Fink, Klaus U. Schulz and Uwe Springmann. „Profiling of OCR'ed Historical Texts Revisited," in *Proceedings of the 2nd International Conference on Digital Access to Textual Cultural Heritage,* 61–66 (Göttingen Germany: ACM, 2017), https://doi.org/10.1145/3078081.3078096.

[8] Sumeet Agarwal, Shantanu Godbole, Diwakar Punjani and Shourya Roy, „How Much Noise is Too Much: A Study in Automatic Text Classification," in *Seventh IEEE International Conference on Data Mining (ICDM 2007),* 3–12 (Omaha: 2007), 5, https://doi.org/10.1109/ICDM.2007.21.

[9] Daniel Lopresti, „Optical Character Recognition Errors and Their Effects on Natural Language Processing," *International Journal on Document Analysis and Recognition (IJDAR)* 12, 3. sz. (2009): 141–151, https://doi.org/10.1007/s10032-009-0094-8.

mértékét és a felismerés pontosságát meghatározó módszerek,[10] sőt félig automatizált eljárások is segíthetik a tévesztések javítását és a korabeli helyesírási normának a gépi hibától való megkülönböztetését,[11] mégsem véletlen, hogy egyes tanulmányok arra következtetnek, hogy az adatelemzéssel dolgozó kutatók a kutatási idő akár 80%-át is az adatok előkészítésével tölthetik el.[12]

Az előkészítés idejének megrövidítése érdekében érdemes az algoritmusok zajtoleranciájára vonatkozó teszteket végezni. Agarwal és munkatársai például egy egész sor kísérletről számolnak be, amelyek a digitalizáció során keletkezett hibáknak a szövegcsoportosító algoritmusokra gyakorolt hatását tesztelték, és amelyek alapján megállapították, hogy az osztályozás pontossága akár 40%-ig is tolerálja a bevezetett zajokat.[13] A stilometriában Eder angol, német, lengyel, ógörög és latin prózaszövegeken mutatta be több szerzőazonosításhoz használt módszerének stabilitását: kutatásának eredménye, hogy a zajtolerancia ugyan eltérő a különböző nyelveknél, de még a 20%-os torzulás sem befolyásolja szignifikánsan a módszerek teljesítményét.[14] A számítógépes szerzőazonosítás jelenlegi megközelítései ugyanis olyan jellemzőkre irányulnak, mint a szavak unigrammáinak vagy karakter n-gramoknak az eloszlása;[15] ezek pedig nagyon gyakori és átfogó elemek egy szövegben (és sokkal kevésbé ritkák, mint például a jelentéses szavak), ami magyarázhatja, hogy miért ellenállók a jelentős mértékű, látszólag sztochasztikus zajjal szemben. Továbbá bizonyos szabályozási technikák (például a Support Vector Machine osztályozó eljárása) segítségünkre lehetnek abban, hogy ne essünk a túlillesztés *(overfitting)* csapdájába a zajos környezetben. Mindeddig azonban még egyetlen szisztematikus tanulmány sem modellezett le ilyesfajta zajt, miközben a kutatók az NPL-szoftverek használatával rutinszerűen normalizálják eredményeiket.[16]

[10] Subramaniam et. al, „A Survey of Types of Text Noise," 117–118.

[11] Thorsten Vobl, Annette Gotscharek, Uli Reffle, Christoph Ringlstetter and Klaus U. Schulz, „PoCoTo – an Open Source System for Efficient Interactive Postcorrection of OCRed Historical Texts," in *Proceedings of the First International Conference on Digital Access to Textual Cultural Heritage – DATeCH '14,* 57–61 (Madrid: ACM Press, 2014), https://doi.org/10.1145/2595188.2595197. Példának lásd a CIS-LMU *Post Correction Tool*ját (PoCoTo). Hozzáférés: 2021.07.30, https://www.digitisation.eu/tools-resources/tools-for-text-digitisation/cis-lmu-post-correction-tool-pocoto/.

[12] Wickham, „Tidy Data."

[13] Agarwal et. al., *How Much Noise.*

[14] Maciej Eder, „Mind Your Corpus: Systematic Errors in Authorship Attribution," *Literary and Linguistic Computing* 28, 4. sz. (2013): 603–614, 612, https://doi.org/10.1093/llc/fqt039.

[15] Bradley Kjell, W. Addison Woods and Ophir Frieder, „Discrimination of Authorship Using Visualization," *Information Processing & Management* 30, 1. sz. (1994): 141–150, https://doi.org/10.1016/0306-4573(94)90029-9; Efstathios Stamatatos, „On the Robustness of Authorship Attribution Based on Character N-Gram Features," *Journal of Law and Policy* 21, 2. sz. (2013): 421–439; Mike Kestemont, „Function Words in Authorship Attribution: From Black Magic to Theory?" in *Proceedings of the 3rd Workshop on Computational Linguistics for Literature (CLFL),* 59–66 (Gothenburg: Association for Computational Linguistics, 2014), https://doi.org/10.3115/v1/W14-0908.

[16] Például az alábbi esetekben: Patrick Juola, „Authorship Attribution," *Foundations and Trends in Information Retrieval* 1, 3. sz. (2007): 233–334, http://doi.org/10.1561/1500000005; Stamatatos, „On the Robustness;" Moshe Koppel, Jonathan Schler and Shlomo Argamon, „Computational Methods in Authorship Attribution," *Journal of the American Society for Information Science and Technology* 60, 1. sz. (2009): 9–26, https://doi.org/10.1002/asi.20961.

 DOI 10.31400/dh-hun.2021.5.3144

A szerzőazonosítás és a stilometriai analízis eredményeinek megbízhatóságát a szövegminőségen túl a minta mérete (vagyis az elemzett szavak száma) is befolyásolja. Eder egy, a témában írt cikkében előző megállapításait[17] cáfolva a minimális elfogadható mintaméret küszöbét 2000 szóban határozza meg – amennyiben a szerzői ujjlenyomat a szövegben markánsnak mondható.[18] Ebből az következik, hogy noha a nagy méretű minták általánosan előnyösebbek, a kis méretűek (2000 szóig) is képesek lehetnek pontos eredményeket biztosítani.

## 3. Anyagok

### 3.1. A Grimm-levelezés

Az OCR és HTR során keletkezett zajnak a szerzőazonosításra gyakorolt hatását a Grimm család levelezésének egy részén teszteltük. Ez a diakrón szövegegyüttes kifejezetten alkalmas a feladatra, mivel a kézírásos dokumentumok és a nyomtatott kiadás egyaránt rendelkezésre állnak.

**3.1.1. Kézírásos levelek** ♦ 2015 októberében jutottunk hozzá a Grimm család körülbelül 36000 levelének digitalizált korpuszához a marburgi Állami Levéltárnak köszönhetően.[19] Ezek közt több olyan levél található, melyeket Jacob és Wilhelm Grimm egymással, illetve ismerőseikkel váltottak több mint 70 év leforgása alatt: a szerteágazó (a betegségektől az utazásokig terjedő) témájú levelek a testvérek életének és stilisztikai fejlődésének tanújaként is szolgálhatnak. A Marburg-gyűjtemény azonban nem teljes: további 1000 hivatalos levelet őriz a berlini Humboldt Egyetem is,[20] ám az ezeknek a megszerzésére irányuló tárgyalások még nem zárultak le a Grimm Levelezés Központjával.

Mivel a kutatás tárgya Jacob és Wilhelm Grimm kézírásának vizsgálata, a teljes Marburg-gyűjteményből csupán a fivérek által írt leveleket választottuk ki.

**3.1.2. A levelek nyomtatásban megjelent kritikai kiadása** ♦ A testvérek levelezésének egyetlen kritikai kiadása Heinz Rölleke 2001-ben megjelent *Jacob és Wilhelm Grimm levelezése* című, két kötetet számláló könyve.[21] A kiadvány előszavában az olvasható, hogy a kritikai kiadás a szövegek közlésekor követi az eredeti szövegekhez hű szerkesztői konvenciót.[22] Rölleke azonban apró változtatásokat eszközöl: mind a gon-

[17] Maciej Eder, „Does Size Matter? Authorship Attribution, Small Samples, Big Problem," *Digital Scholarship in the Humanities* 30, 2. sz. (2015): 167–182, https://doi.org/10.1093/llc/fqt066.

[18] Maciej Eder, „Short Samples in Authorship Attribution: A New Approach," *Digital Humanities 2017: Conference Abstracts,* 221–224 (Montreal: McGill University, 2017), 223.

[19] A TIFF-fájlok és a publikáció jogait a marburgi Hesseni Állami Levéltártól vásároltuk meg. A gyűjtemény neve: *340 Grimm.* Bővebben, hozzáférés: 2021.07.30, https://landesarchiv.hessen.de/.

[20] Lásd bővebben, hozzáférés: 2021.07.30, http://www.grimmbriefwechsel.de/arbeitsstelle/arbeitsstelle.html.

[21] *Briefwechsel der Brüder Jacob und Wilhelm Grimm: Kritische Ausgabe in Einzelbänden,* Hg., Heinz Rölleke Bd. 1, in 3 Teil (Stuttgart: Hirzel Verlag, 2001).

[22] A szerkesztői jegyzet teljes szövegét német nyelven lásd: Uwe Meves und Jens Haustein, „Vorwort," in Rölleke, *Briefwechsel der Brüder,* 1.1: 5–8, http://www.grimmnetz.de/bwfiles/!grimm-bw1-1_kopie.pdf.

dolatjeleket, mind a nagykötőjeleket gondolatjellé egységesíti a szókapcsolatokban; pótolja a hiányzó központozást; a szokatlan rövidítéseket dőlttel szedve közli szögletes zárójelben; kiteszi a hiányzó umlautokat (de nem jelzi a mulasztások tényét); hiányzó karaktereket pótol anélkül, hogy meghatározná a helyüket a kéziratban; nem jelöli sem a pecsétes helyeket, sem pedig a kihúzott szövegrészeket.

### 3.2. A kutatásba felvett levelek

A Marburg-korpuszból 85 levél került kiválasztásra a kutatáshoz – ebből 50 Jacob Grimmnek, a fennmaradó 35 pedig Wilhelm Grimmnek tulajdonítható. Ezek többnyire egymásnak vagy egy rokonuknak, Karl Weigandnak címzett levelek. Weigand a kor szerzője és filológusa, aki Grimmékkel együtt részt vett a *Deutsches Wörterbuch (Német Nagyszótár)* elkészítésének munkálataiban. Az *1.* és *2. táblázat* időbeli sorrendben mutatja a Jacobtól, illetve Wilhelmtől származó leveleket.

 DOI 10.31400/dh-hun.2021.5.3144

**1. táblázat**. Jacob Grimm 50 levele. Az olvashatóságra vonatkozó szakértői értékelés szögletes zárójelben szerepel. A második korszakot (1800) a második HTR-modell során vettük fel a korpuszba.

| LETTERS WRITTEN BY JACOB GRIMM: 50 | | | |
|---|---|---|---|
| **Epoch** | **Letter ID** | **Year** | **Readability** |
| 1. 1793 | Br 5995 | 1793 | low [to v. low] |
| 2. 1800 | Ms 237 | 1800 | low |
| 3. 1805-1806 | Br 2164 | 1805 | low |
| | Br 2165 | 1805 | low |
| | Br 2169 | 1805 | low |
| | Br 2163 | 1805 | low [to v. low] |
| | Br 2166 | 1805 | low |
| | Br 2167 | 1805 | low |
| | Br 2168 | 1805 | low [to v. low] |
| | Br 2170 | 1805 | low |
| | Br 2176 | 1805 | very low |
| | Br 2174 | 1806 | low |
| 4. 1814-1863 | Br 2175 | 1833 | low |
| | Br 2171 | 1833 | medium [to high] |
| | Br 2172 | 1838 | medium [to high] |
| | Br 5996 | 1838 | medium |
| | Br 2237 | 1840 | medium |
| | Br 2238 | 1840 | low |
| | Br 2239 | 1840 | low [to medium] |
| | Br 2240 | 1841 | high |
| | Br 2241 | 1844 | very low [to medium] |
| | Br 2242 | 1846 | very low [to medium] |
| | Br 2243 | 1847 | medium |
| | Br 2173 | 1848 | low |
| | Br 2269 | 1848 | low |
| | Br 2244 | 1849 | low [to medium] |
| | Br 2245 | 1849 | low |
| | Br 2268 | 1850 | low |
| | Ms 131 | 1850 | high |
| | Br 2246 | 1852 | low |
| | Br 2247 | 1853 | low |
| | Br 2248 | 1854 | low |
| | Br 2249 | 1855 | low |
| | Br 2250 | 1856 | low |
| | Br 2266 | 1857 | low |
| | Br 2251 | 1858 | low |
| | Br 2252 | 1858 | low |
| | Br 2253 | 1859 | low |
| | Br 2254 | 1859 | low |
| | Br 2255 | 1859 | low |
| | Br 2267 | 1859 | low |
| | Br 2256 | 1860 | low |
| | Br 2257 | 1860 | low |
| | Br 2258 | 1861 | medium [to low] |
| | Br 2259 | 1861 | medium [to low] |
| | Br 2260 | 1861 | low |
| | Br 2261 | 1862 | very low |
| | Br 2262 | 1862 | low |
| | Br 2263 | 1862 | low |
| | Br 2264 | 1863 | very low |
| | Br 2265 | 1863 | very low |

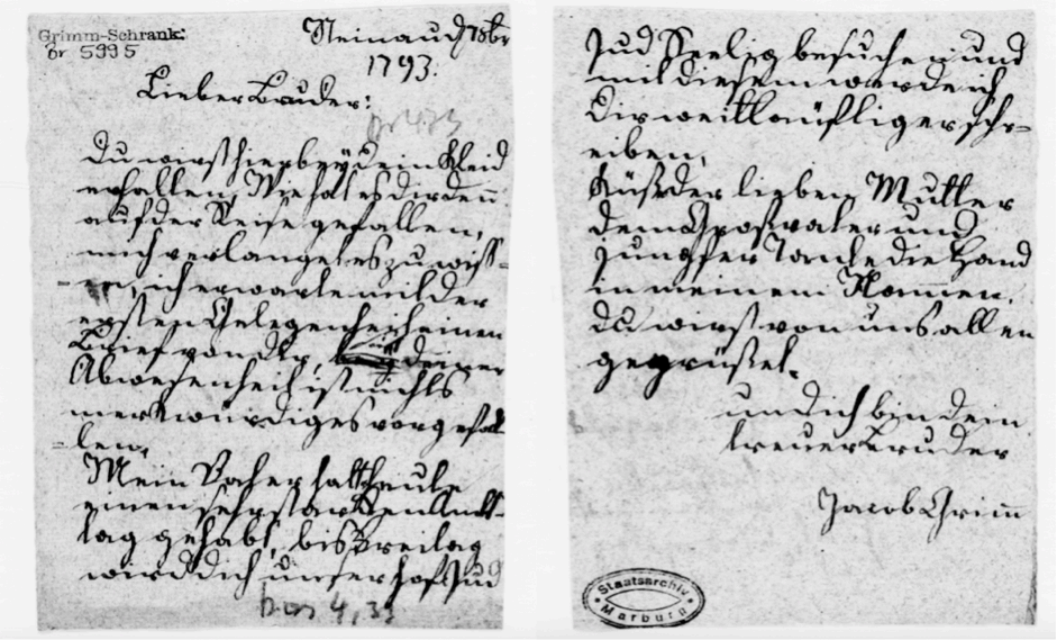
**2. táblázat**. Wilhelm Grimm 35 levele. Az olvashatóságra vonatkozó szakértői értékelés szögletes zárójelben szerepel.

| **LETTERS WRITTEN BY WILHELM GRIMM: 35** | | | |
|---|---|---|---|
| **Epoch** | **Letter ID** | **Year** | **Readability** |
| 1. 1793 | Br 5993 | 1793 | low |
| | Br 5994 | 1793 | low |
| | Br 2678 | 1793 | low |
| | Br 2679 | 1793 | low |
| 2. 1802-1805 | Br 2677 | 1805 | low |
| 3. 1831-1843 | Br 2680 | 1831 | low |
| | Ms 426 Bl 7 | 1833 | very low [to low] |
| | Ms 428 Bl 7b | 1833 | very low [to low] |
| | Ms 426 Bl 10 | 1833 | very low |
| | Ms 426 Bl 11 | 1833 | very low |
| | Ms 426 Bl 13 | 1833 | very low |
| | Ms 426 Bl 15 | 1833 | very low |
| | Br 1687 | 1843 | low |
| | Br 2681 | 1843 | very low [to medium] |
| | Br 1688 | 1843 | low |
| 4. 1846-1859 | Br 2734 | 1846 | medium [to high] |
| | Br 2682 | 1847 | low |
| | Br 2683 | 1848 | medium |
| | Ms 161 | 1850 | low |
| | Br 2735 | 1851 | high [low to medium or high] |
| | Br 2736 | 1855 | high [low to medium or high] |
| | Br 2684 | 1856 | high [low to medium or high] |
| | Br 2685 | 1856 | medium |
| | Br 2687 | 1856 | medium |
| | Br 2686 | 1856 | high [low to medium or high] |
| | Br 2688 | 1856 | medium |
| | Br 2689 | 1856 | medium |
| | Br 2737 | 1857 | medium |
| | Br 2690 | 1858 | low |
| | Br 2738 | 1858 | medium |
| | Br 2739 | 1858 | low |
| | Br 2740 | 1859 | low |
| | Br 2741 | 1859 | low |
| | Br 2742 | 1859 | low |
| | Br 2743 | 1859 | medium |

 DOI 10.31400/dh-hun.2021.5.3144

**3.2.1. A levelek kategorizálása: korszakok és olvashatóság** ♦ A 85 levelet korszakok és olvashatóság mentén manuálisan is kategorizáltuk.

**3.2.1.1. Korszakok** * Az idő előrehaladtával a testvérek írásképe változott. Ezek a változások legjobban akkor észlelhetők, ha a leveleket egymás mellett vizsgáljuk. Így például Jacob kézírása az 1805–1806 közti időszakban (20–21 éves korában) látványosan különbözik az élete végére kialakult írásképtől. Az *1. ábrá*n a gyűjtemény legkorábbi, Jacob 8 éves korában írt levele látható 1793-ból.



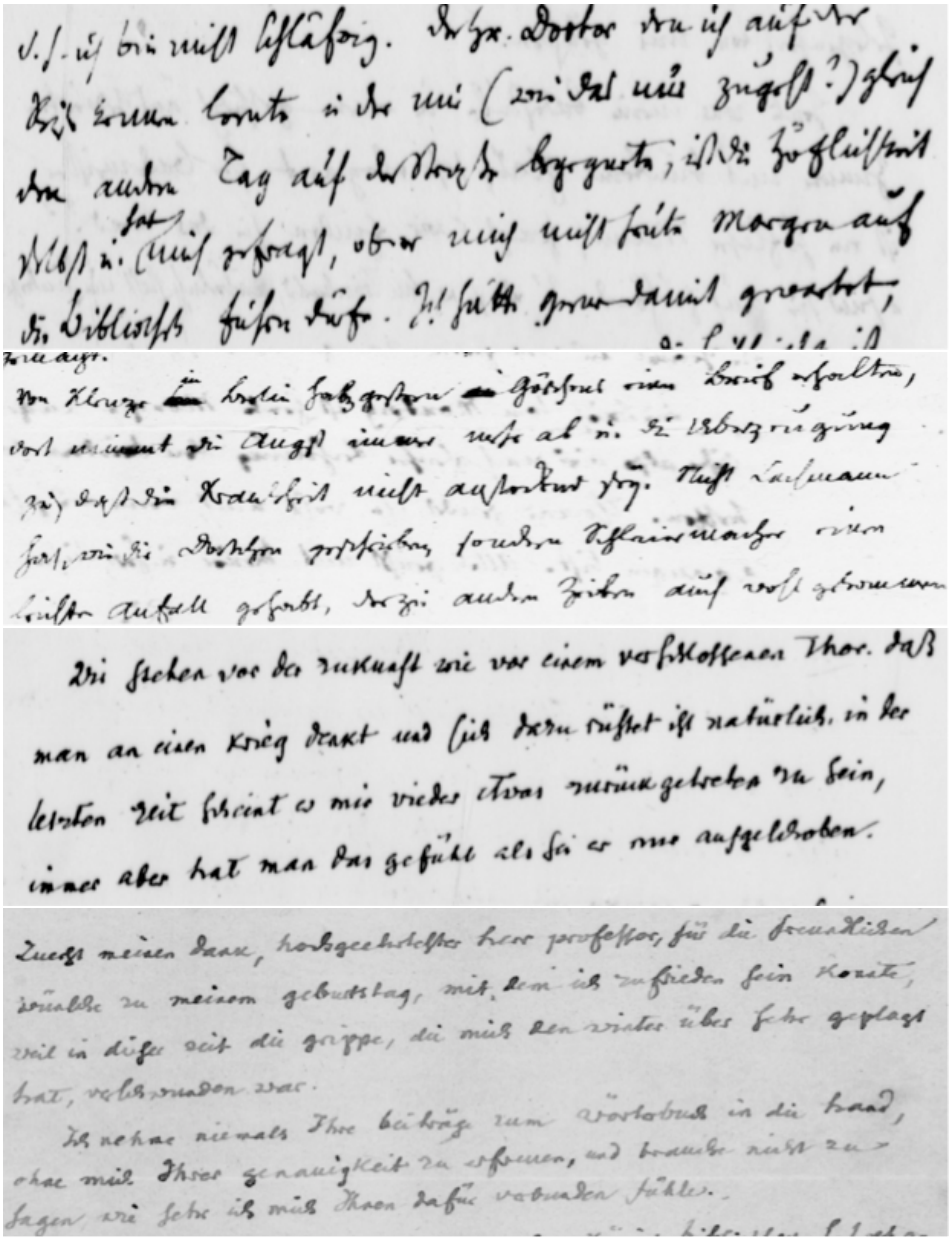

**1. ábra**. Jacob Grimm levele 1793-ból. A levél teljes átirata a *Függelék*ben olvasható.

A változások mentén a leveleket kézírásos periódusok, korszakok szerint csoportosítottuk: eszerint fejenként négy csoportra oszthatók a fivérek levelei. Jacob korszakai: 1793, 1800, 1805–1806, 1814–1863,[23] míg Wilhelméi: 1793, 1802–1805, 1831–1843 és 1846–1859.

**3.2.1.2. Olvashatóság** * A leveleket négy csoportra osztottuk olvashatóságuk alapján is. Ezek a csoportok: nagyon alacsony *(very low),* alacsony *(low),* közepes *(medium)* és magas *(high)* olvashatóság, mint ahogy az a *2. táblázat*ban is látható. Az olvashatóságot *(readability)* a papír minősége (a rossz minőségű papíron kiütköznek a tintafoltok) és a kézírás kiolvashatósága *(legibility)* befolyásolja. A Grimm-kutatókkal való egyeztetés alapján az állapítható meg, hogy az olvashatóság kritériumai a testvérek kézírásának szabályosságában, az egyértelműen megkülönböztethető karakterhosszúságban, valamint a szóvégi betűk önkényes elhagyásában ragadható

[23] Jacob Grimm itt jelölt utolsó korszakában megfigyelhető ugyan némi változás a kalligráfiát illetően, de ez nem akkora mértékű, hogy ezért indokolt volna a negyedik csoportot továbbosztani.

meg.[24] Ebből származtatható az a megfigyelés is, hogy általánosságban mind Jacob, mind Wilhelm korai írásai nehezebben olvashatók. A levélcsoportok ilyen kronológiai elkülönítése szükséges egy megbízható HTR-modell betanításához, amely a korpusz egészét tekintve is a lehető legalacsonyabb hibaarányt produkálja.



**2. ábra**. Négy levél Wilhelm Grimmtől. Fentről lefelé haladva: nagyon alacsony olvashatóság (Br 5993, 7 évesen); alacsony olvashatóság (Br 2680, 45 évesen); közepes olvashatóság (Br 2743, 73 évesen); magas olvashatóság (Br 2736, 69 évesen).

[24] Bernhard Lauer és Rotraut Fischer a Brüder Grimm-Gesellschafttól ['Grimm testvérek Egyesület'], Kassel, Németország.

 DOI 10.31400/dh-hun.2021.5.3144

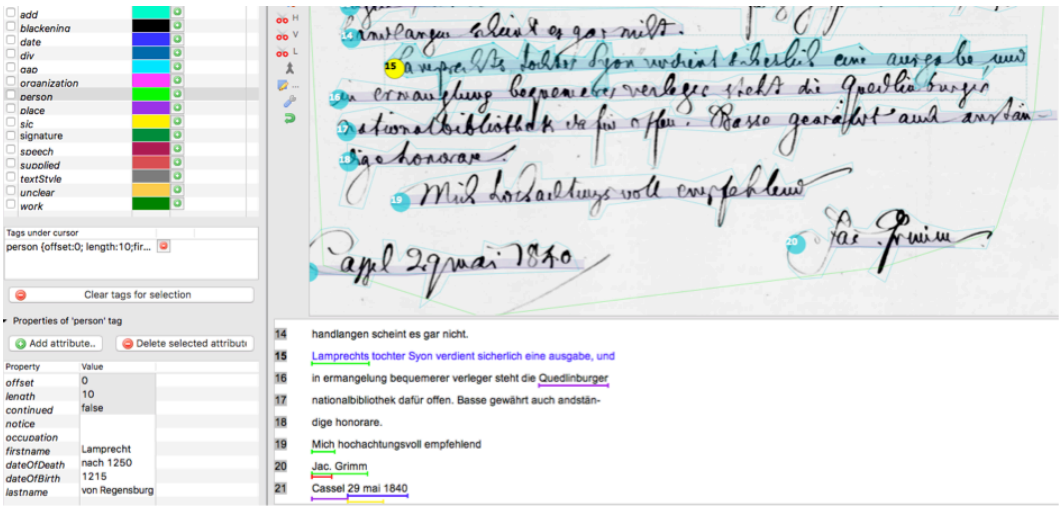
## 4. Módszerek és eredmények

### 4.1 Digitalizáció: manuális átírás (MAN), HTR és OCR

Az alábbi rész a Grimm-levelezés háromféle digitalizálási módszerét részletezi, és a címben jelzett utóbbi kettőt (HTR és OCR) össze is veti egymással.

**4.1.1. Manuális átírás** ♦ A levelek automatikus és manuális átírására a legújabb technológián alapuló *Transcribus* szoftvert alkalmaztuk.[25] Azonban egy olyan HTR-modell létrehozásához, amely lehetővé teszi a kézírásos dokumentumok automatizált átírását, a *Transcribus* programnak minimum száz oldal terjedelmű, manuálisan átírt szövegre van szüksége.[26] A 85 darab így átírt levél hossza változó – néhány csupán egyoldalas, néhány több oldalt is felölel. Jacob 50 levele 90 oldalnyi átírt szövegnek felel meg, míg Wilhelm 35 levele 64 oldalt tesz ki: összesen tehát 154 manuálisan átírt oldalról beszélhetünk a vizsgált szövegek esetében.

Amellett, hogy diplomatikus átírást készítettünk,[27] a leveleket metaadatokkal is elláttuk a fivérek írásképét és a tartalmat illetően – miként az a *3. ábrán* is látható.[28]




**3. ábra.** A *Transcribus* felülete. A képernyő három területre oszlik: a bal oldalsáv a metaadatokat és a jegyzetelési funkciókat; a felső terület az eredeti dokumentumot (Br 2238, Jacob Grimm); az alsó terület pedig a jegyzetekkel ellátott átiratot tartalmazza.

[25] A *Transcribus*ról több információért lásd, hozzáférés: 2021.07.30, `https://transkribus.eu/Transkribus/`.

[26] További információért lásd a *Transkribus* wiki-oldalát: `https://transkribus.eu/wiki/index.php/Questions_and_Answers/What_is_needed_for_the_HTR_to_work.3F`.

[27] A „diplomatikus átírás" definíciójához lásd: *Lexicon of Scholarly Editing,* hozzáférés: 2021.11.18, `https://lexiconse.uantwerpen.be/lexicon/transcriptionDiplomatic.html`.

[28] Melinda Jander, „Handwritten Text Recognition – Transkribus: A User Report," in *Göttingen, Germany: eTRAP Research Group* (Göttingen: University of Göttingen, 2016), hozzáférés: 2021.11.18, `http://www.etrap.eu/wp-content/uploads/2016/11/TrAIN-Transkribus_User_Report-2016.pdf`.

**4.1.2. A kézírásos levelek HTR-modellje** ♦ A manuális átírásokat a HTR-modell kidolgozására használtuk, hogy az képes legyen felismerni és automatizáltan átírni további, a fivéreknek tulajdonítható leveleket és dokumentumokat (mint például a már említett 1000 darab levelet a berlini gyűjteményből).

Mint említettük, egy megbízható HTR-modellnek ideális esetben egy legalább 100 oldalnyi kézzel átírt szöveget tartalmazó tanítókorpuszra van szüksége.[29] A tény, hogy Jacob és Wilhelm Grimm 85 levele összesen 154 oldalnyi szöveget tesz ki, a következő választás elé állított minket: inkább a HTR 100 oldalnyi szövegszükségletét elégítsük ki úgy, hogy a testvérek leveleit egyesítjük, vagy két különálló modellt képezzünk ki, egyet-egyet a fivérek számára, viszont kevesebb terjedelemben (90 oldal Jacobnak és 64 oldal Wilhelmnek). Végül úgy döntöttünk, hogy mindkét lehetőséget teszteljük, az eredményeket pedig összevetjük egymással.[30]

**4.1.2.1. Az első HTR-modell** * Az első HTR-próbakör során Jacob és Wilhelm mind a 85 manuálisan átírt levelét (154 oldal és 26983 szó) felhasználtuk. Amellett, hogy így eleget tettünk a minimum oldalszám követelményének, a két kézírás egyesítése és egy ilyen HTR-modell megalkotása mögött meghúzódó feltételezésünk az volt, hogy a kevert modell nagyobb ellenállóságot mutat majd a testvérek kézírásának diakrón változásaival szemben. A karakterhiba-arány (CER – Character Error Rate) a kevert modell eredményeként 18,83% volt – azaz a szöveg minden ötödik karakterét helytelenül ismerte fel a program. Ennek javítása érdekében további 2000 szónyi (17 oldalnak megfelelő) Grimm-kézírást írtunk át manuálisan. A várakozással ellentétben az új hibaarány 40%-ra nőtt – azaz minden kettő és feledik karakter hibásan szerepelt az átiratban. Alaposabb vizsgálat után rájöttünk azonban, hogy 13 nagyon alacsony olvashatóságú levél felelős ezért a magas karakterhiba-arányért. Ennek csökkentése céljából egy olyan második HTR-kört futtattunk le, amelyben a 13 problematikus levelet 11 másik, a fivérek által írt dokumentumra cseréltük (35 oldal, 5788 szó).[31] A *3.* és *4. táblázat* az elvett és hozzáadott dokumentumokat listázza.

[29] További információért a HTR-hez szükséges adathalmazképzéssel kapcsolatban lásd: `http://read.transkribus.eu/wp-content/uploads/2017/01/READ_D7.7_HTRbasedonNN.pdf`.

[30] A HTR-modellt a *Transcribus* használatával dr. Günther Mühlberger hozta létre.

[31] A dokumentumok forrása a marburgi Hesseni Állami Levéltár weboldala, lásd, hozzáférés: 2021.07.30, `https://arcinsys.hessen.de/arcinsys/detailAction.action?detailid=g195109&icomefrom=search`.

 DOI 10.31400/dh-hun.2021.5.3144

**3. táblázat**. A HTR-tesztkorpuszból alacsony olvashatóságuk miatt kizárt levelek, amelyek negatívan hatottak az első HTR-kör működésére.

| LETTERS DISCARDED FROM HTR CORPUS | | | |
|---|---|---|---|
| **Author** | **Letter ID** | **Year** | **Readability** |
| Jacob | Br 2176 | 1805 | very low |
| | Br 2241 | 1844 | very low |
| | Br 2242 | 1846 | very low |
| | Br 2261 | 1862 | very low |
| | Br 2264 | 1863 | very low |
| | Br 2265 | 1863 | very low |
| Wilhelm | Ms 426 Bl 7 | 1833 | very low |
| | Ms 426 Bl 7b | 1833 | very low |
| | Ms 426 Bl 10 | 1833 | very low |
| | Ms 426 Bl 11 | 1833 | very low |
| | Ms 426 Bl 13 | 1833 | very low |
| | Ms 426 Bl 15 | 1833 | very low |
| | Br 2681 | 1843 | very low |

**4. táblázat**. A HTR-tesztkorpuszba újonnan felvett levelek magas és közepes olvashatósággal, amelyek az előző kör után kizárt 13 levél helyettesítésére szolgálnak. A dokumentumok a levelek mellett verseket és dalokat is tartalmaznak.

| LETTERS ADDED TO THE HTR CORPUS | | | | | |
|---|---|---|---|---|---|
| **Author** | **Epoch** | **Document ID** | **Year** | **Readability** | **HTR Word-count** |
| Jacob | 2. 1800 | Ms 237 (Song) | 1800 | low | 343 |
| | 4. 1814-1863 | Ms 239 (Diary entry) | 1815 | high | 1218 |
| | | Br 2231 | 1829 | high | 699 |
| | | Br 2230 | 1839 | high | 245 |
| | | Br 2232 | 1841 | high | 246 |
| | | Br 2235 | 1850 | medium | 316 |
| | | Br 2233 | 1860 | high | 107 |
| | | Ms 242 (Dictionary entry draft) | n.d. | high | 485 |
| Wilhelm | 2. 1802-1805 | Ms 245 (poem) | 1802 | medium | 177 |
| | 3. 1831-1843 | Br 2579 | 1833 | medium | 726 |
| | 4. 1846-1859 | Br 2580 | 1854 | medium | 1226 |

**4.1.2.2. A második HTR-modell** * A második HTR-körben így 83 dokumentumot vizsgáltunk, amelyek összesen 28963 szót (ebből 10250 Wilhelmé, 18686 Jacobé) és 128 oldalt (ebből 44 Wilhelmé és 84 Jacobé) tartalmaztak. Ebben a körben Jacob és Wilhelm írásait, az „oszd meg és uralkodj” elvével, egymástól függetlenül tápláltuk be a különböző tanító- és tesztkorpuszokba, ami így 8 különálló esetet eredményezett (fejenként és korszakonként egyet) – a szándékunk ezzel annak a felderítése volt, hogy vajon kevesebb, de jobban kontrollált adattal stabilabb modell hozható-e létre. Biztató eredmények születtek; az átlagos karakterhiba-arány az egyes esetekben kevesebb volt, mint 10%.

Alább Jacob Br 2238-as jegyzékszámú, 1840-re datálható levelének HTR-rel átírt részlete látható, melyet Jacob kézírásának 1814–1863 közti periódusán tanított modellel hoztunk létre.

A levél eredeti szövege:

> [...] handlangen scheint es gar nicht. Lamprechts tochter Syon verdient sicherlich eine Ausgabe, und in ermangelung bequemerer verleger steht die quedlinburger nationalbibliothek dafür offen. Rasse gewährt auch andständigehonorare. Mich hochachtungsvoll empfehlend Jac. Grimm

A HTR-átírás [a hibák aláhúzással jelölve]:

> [...] handlangen scheint es gar nicht. Lamprechts tochter von verdient sicherlich eine ausgabe und in ermangelung bgineneber verleger steht die quedlinburger natüoalbiblittchke der für offen. Rasse gewährt auch wurdendigehonorare mich hochachtungsvoll empfehlend Ihr. Grimm

**4.1.3. A kritikai kiadás OCR-verziója** ♦ Rölleke hétkötetes kritikai kiadásából a digitalizációt és az OCR-t követően 7 fájl készült.[32] Az alábbi szövegrészlet a zajos OCR-eredményre hoz példát (Jacob Grimm fent idézett levele, Br 2238-as jegyzékszámmal):

> Handlangen scheint es gar nicht.
>
> Lamprechts tochter Syon verdient sicherlich eine ausgabe, und in er-manglung bequemerer Verleger steht die Quedlinburger nationalbibliothek dafür offen. Basse gewaährt auch anständige honorare. Mich hochachtungsvoll empfehlend
>
> Jac. Grimm.

Ahogy látható, az OCR az *er-manglung* szóban megőrizte a kiadás által használt kötőjelet, a *Verleger* szót pedig nagy kezdőbetűvel írta, míg a nyomtatott kiadásban kis kezdőbetűvel szerepel. Ezektől a hibáktól eltekintve az *5.* és *6. táblázat* az OCR nagy fokú pontosságát mutatja: a levelek esetében a helyesen felismert szavak mediánja 91% fölötti (a helyes karakterfelismerés pedig 98%-os).

**5. táblázat**. A gyűjtemény 72 levelének átlagos tisztasága. A félkövérrel szedett számok a levelek eloszlásának mediánját mutatják. Az átlagok standard hibái a gyűjteményen belül elhanyagolhatók.

| **MEAN COLLECTION CLEANLINESS** | | |
|---|---|---|
| | **Clean words in %** | **Clean characters in %** |
| OCR | 88.25 | 97.79 |
| HTR | 80.85 | 94.41 |
| **LETTER CLEANLINESS (THREE QUARTILES)** | | |
| OCR | 86.80 **91.69** 94.06 | 97.95 **98.70** 99.18 |
| HTR | 79.28 **84.29** 88.39 | 94.09 **95.89** 97.44 |

[32] A digitalizációt és az OCR-t a Göttinger Digitisation Centre *Abbyy Fine Reader*rel végezte. Lásd bővebben, hozzáférés: 2021.07.30, `https://www.sub.uni-goettingen.de/en/copying-digitising/goettingen-digitisation-centre/`.

 DOI 10.31400/dh-hun.2021.5.3144

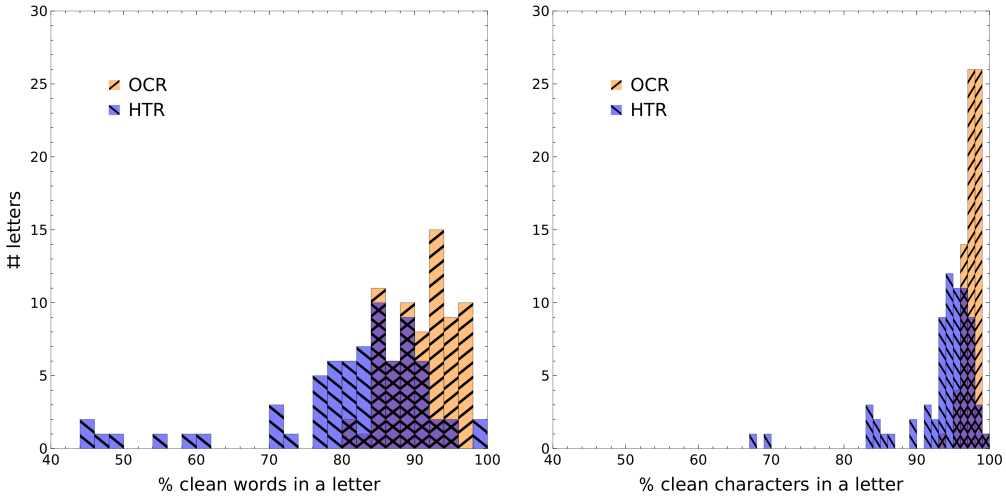
**6. táblázat**. A gyűjtemény 72 levelének átlagos tisztasága szerzők szerint. Wilhelm levelei következetesen magasabb értéket vesznek fel, habár kevesebb van belőlük. A félkövérrel szedett számok a levelek eloszlásának mediánját mutatják. A standard hibák: <0.0026% (szavak) és <0.00016% (karakterek).

| **MEAN COLLECTION CLEANLINESS** | | | | |
|---|---|---|---|---|
| | **Clean words in %** | | **Clean characters in %** | |
| | Jacob | Wilhelm | Jacob | Wilhelm |
| OCR | 87.10 | 91.12 | 97.60 | 98.26 |
| HTR | 79.44 | 84.21 | 94.24 | 94.81 |
| **LETTER CLEANLINESS (THREE QUARTILES)** | | | | |
| OCR | 86.65 **91.69** 93.87 | 87.51 **91.98** 94.24 | 98.29 **98.86** 99.17 | 97.49 **98.43** 99.19 |
| HTR | 76.93 **81.93** 85.68 | 83.61 **87.30** 90.41 | 94.00 **95.50** 96.96 | 95.22 **96.77** 98.39 |

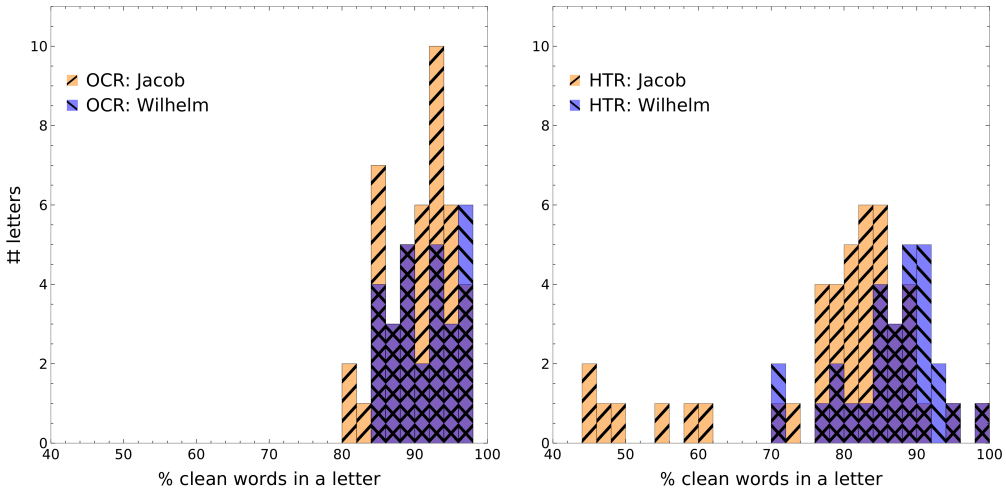
**4.1.4. A HTR és OCR által feldolgozott adatok tisztaságának kiértékelése** ♦ A következő lépésben a HTR és az OCR segítségével, valamint manuális módon feldolgozott (MAN) levelek tisztaságának összevetését végeztük el. E formátumok mindegyikében 72 darab levél érhető el. A manuális átírást vettük etalonkorpusznak és a többi verzió tisztaságát ehhez mértük (vö. *5. táblázat*). Fontos még megjegyezni, hogy míg a HTR esetében a különbségek csak a felismerési hibákból adódnak, addig az OCR esetében a felismerési hibák és Rölleke esetleges szerkesztői beavatkozásai egyaránt számításba jöhetnek.

Attól függően, hogy mely stilometriai vizsgálatot végezzük, a tisztaság a rosszul felismert szavak (ha az osztályozás szavakon, szó n-gramokon, vagy lemmákon alapul) vagy a rosszul felismert karakterek (amennyiben karakter n-gramokat használunk) százalékos arányán áll, hiszen minden ilyen hiba módosítja a szó/karakter/n-gram gyakoriságát, következésképpen megváltoztathatja a szövegek között mért távolságokat (lásd *4. ábra*).[33] Ezen túl a szerzők eltérő hibaaránya szintén nehézséget jelent a módszerek értékelésekor (lásd *6. táblázat* és *5. ábra*).

[33] Vö. John Burrows, „'Delta': A Measure of Stylistic Difference and a Guide to Likely Authorship," *Literary and Linguistic Computing* 17, 3. sz. (2002): 267–287, https://doi.org/10.1093/llc/17.3.267.




**4. ábra**. A hisztogramok azt mutatják, hogy hány levélhez tartozik egy adott százalékban helyesen felismert szó/karakter. Feltűnő, hogy a HTR meglehetősen instabil eredményeket produkál (a bal oldali kiugró értékek). A láthatóság érdekében két hisztogram közötti átfedést a színek keverésével és keresztmintázat hozzáadásával jelezzük.



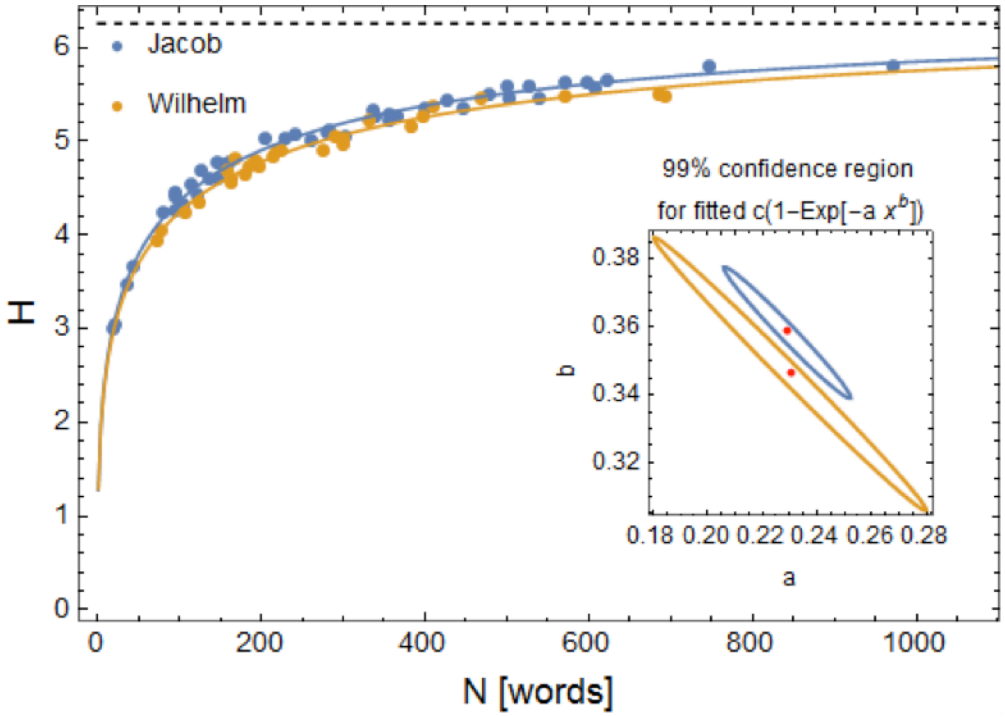

**5. ábra**. A Jacob és Wilhelm leveleinek hisztogramjai. A jobb oldali panel mutatja, hogy a HTR számára Jacob kézírása okozott nagyobb problémát.

Kezdésként talán nem volna szerencsés a digitalizációs eljárások hatását a szerzőazonosítás kapcsán felmerülő összes jellemzőn (szó és karakter n-gramok) egyszerre vizsgálnunk. Éppen ezért elsőként inkább a felismerési hibáknak a lexikális gazdagságra gyakorolt hatásáról számolunk be, és csak később fordulunk a szerzőazonosítás felé (a 4.2. részben). A lexikális gazdagság nem feltétlen biztosítja a jó szerzőazonosítást,[34] habár határozott stilisztikai jelentéssel bír és a zajos adatokhoz kapcsolódó problémákat is jól illusztrálhatja. A típusok számát megadó képletek közül *(richness*

[34] David L. Hoover, „Another Perspective on Vocabulary Richness," *Computers and the Humanities* 37, 2. sz. (2003): 151–178, https://doi.org/10.1023/A:1022673822140.

 DOI 10.31400/dh-hun.2021.5.3144

*score)*[35] kettőt alkalmaztunk: Shannon entrópiáját $H = -\sum_{t=1}^{T} p_t \log p_t$ és a Simpson-indexet $D = \sum_{t=1}^{T} p_t^2$ (más néven fordított részvételi arány [Inverse Participation Ratio – IPR]). Mindkettő a diverzitásindexek egy alesete, ahol $T$ a típusok számát jelöli, míg $p_t$ a $t$ típus megjelenésének valószínűségét (azaz $t$ összes előfordulása osztva a szöveg szavainak számával). Néhány kritika,[36] valamint a képletek nagyon erős (bár nem lineáris) korrelációja ellenére ezek a legegyszerűbbek, a legkevésbé önkényesek és elméleti szempontból is a legjobban megérthetők. Ha $N$ a szövegben lévő tokenek száma, az IPR 1/N és 1 között mozog (maximális gazdagság és zéró gazdagság), és főként a legjellemzőbb értékekre (azaz a leggyakoribb szavakra) összpontosít, és így gyorsan stabilizálódik az $N$ szöveghosszal. Az entrópia 0-tól logN-ig terjed (zéró és maximális gazdagság), és a szóeloszlás görbéjének farkára összpontosít (azaz a legritkább szavakra, mint a *hapax legomenon*), amelyek lassabban stabilizálódnak, és az apró változásokra is érzékenyebbek a szöveg előrehaladtával *(6. ábra)*.




**6. ábra**. Jacob és Wilhelm átírt leveleinek entrópiája (pontok) azt mutatja, hogy Jacob nagyobb lexikai változatossággal rendelkezik. Az eredmény statisztikailag szignifikáns (0,99-es megbízhatósági szinten). A fekete szaggatott vonal a két görbére illesztett közös $c$ aszimptotát jelöli.

[35] Fiona J. Tweedie and R. Harald Baayen, „How Variable May a Constant be? Measures of Lexical Richness in Perspective," *Computers and the Humanities* 32, 5. sz. (1998): 323–352, https://doi.org/10.1023/A:1001749303137; Gejza Wimmer and Gabriel Altmann, „Review Article: On Vocabulary Richness," *Journal of Quantitative Linguistics* 6, 1. sz. (1999): 1–9, https://doi.org/10.1076/jqul.6.1.1.4148.

[36] D. I. Holmes, „The Analysis of Literary Style – A Review," *Journal of the Royal Statistical Society. Series A (General)* 148, 4. sz. (1985): 328–341, 328, https://doi.org/10.2307/2981893; Philippe Thoiron, „Diversity Index and Entropy as Measures of Lexical Richness," *Computers and the Humanities* 20, 3. sz. (1986): 197–202, https://doi.org/10.1007/BF02404461.

Ennek kapcsán megállapítható, hogy a HTR hibáinak köszönhetően a levelenkénti szógazdagság statisztikailag jelentősen csökken az átiratokban (a T-tesztek alapján: $p = 1.04 \times 10^{-7}$ [entrópia] és $p = 8.03 \times 10^{-6}$ [IPR]). A rövidebb levelekben ennek az lehet az oka, hogy a HTR szavakat hagy ki, vagy olvaszt egybe, ami alacsonyabb N-értéket, és – entrópia esetében – alacsonyabb logN-értéket eredményez. Egyébiránt nincs statisztikai korreláció a HTR- és OCR-feldolgozások szöveggazdagsága és -tisztasága közt. Mindazonáltal az eredményeket mérlegelve az OCR járhatóbb megoldásnak tűnik a stilometriai vizsgálatok esetében.

Annak érdekében, hogy Jacob és Wilhelm leveleinek lexikai gazdagsága közti különbséget részletesebben feltárjuk, vissza kellene térni ahhoz a problémakörhöz is, miszerint az IRP és az entrópia a szöveg hosszától is függ. Ezt egy exponenciális függvénnyel tudjuk modellezni, mely alulról közelít egy állandóhoz (az adatokon alapuló vizualizációt mutatja a *6. ábra*). Ezután lehetne megvizsgálni, hogy az illesztett görbék paraméterei között van-e számottevő különbség (ez szintén a *6. ábrá*n látható).

### 4.2. Szerzőazonosítás

**4.2.1. Alapok és beállítások** ♦ Ebben a részben a Grimm testvérek egyéni stílusának azonosításról írunk a levelezés már fent tárgyalt zajos digitalizálásának tükrében. A szövegek szerzőségét gépi tanulással végzett kategorizációs és osztályozási műveletekkel igyekeztük meghatározni.[37] Ennek során egy standard bináris osztályozási gyakorlathoz, a Support Vector Machine-hoz (SVM) folyamodtunk, lineáris kernellel és a jól ismert *scikit-learn* könyvtár alapbeállításaival.[38] Egyes tanulmányok szerint az SVM erős alapot biztosít a szerzőazonosításhoz, még kiemelkedően szegényes bemeneti értékek mellett is.[39] Tekintve, hogy adathalmazunk kicsi volt, a visszatartott szerzőkön alapuló keresztellenőrzési eljárást alkalmaztunk (*leave-one-out,* LOO), amelyet minden levéllel elvégeztünk. Azaz minden esetben egyetlen levelet tartottunk vissza tesztpéldányként, míg az osztályozó algoritmust a fennmaradó elemekkel tanítottuk be. Ezt követően rögzítettük a betanított modell előrejelzését a visszatartott minta szerzőségére vonatkozóan. Az egyes modellek teljesítményét a pontosság *(accuracy),* valamint az F1-érték bevett mérőszámai segítségével írjuk le. A kísérlettel kapcsolatban érdemes megemlíteni, hogy ez a felállás egy viszonylag kevés kihívást rejtő szerzőségi problémának tekinthető, hiszen a szerzők száma nagyon korlátozott,[40] és az

[37] Fabrizio Sebastiani, „Machine Learning in Automated Text Categorization," *ACM Computing Surveys* 34, 1. sz. (2002): 1–47, https://doi.org/10.1145/505282.505283; Moshe Koppel, Jonathan Schler and Shlomo Argamon, „Computational Methods in Authorship Attribution," *Journal of the American Society for Information Science and Technology* 60, 1. sz. (2009): 9–26, https://doi.org/10.1002/asi.20961; Efstathios Stamatatos, „A Survey of Modern Authorship Attribution Methods," *Journal of the American Society for Information Science and Technology* 60, 3. sz. (2009): 538–556, https://doi.org/10.1002/asi.20961.

[38] Fabian Pedregosa et al., „Scikit-learn: Machine Learning in Python," *arXiv:1201.0490 [cs]*, 2018. június 5., http://arxiv.org/abs/1201.0490.

[39] Stamatatos, „A Survey."

[40] Vö. Kim Luyckx and Walter Daelemans, „The Effect of Author Set Size and Data Size in Authorship Attribution," *Literary and Linguistic Computing* 26, 1. sz. (2011): 35–55, https://doi.org/10.1093/llc/fqq013. Eder, „Does Size Matter?"

 DOI 10.31400/dh-hun.2021.5.3144

érintett szövegek műfaja is viszonylag állandó.[41] Habár az adathalmaz mérete a gépi tanulás szempontjából kicsinek tekinthető, jól reprezentálja a szerzőségre vonatkozó tanulmányok nagy részét, ahol gyakoriak a rövid szövegekből álló korpuszok, és amelyek ezért szintén a LOO-eljárást részesítik előnyben. Az egyes levelek hossza már önmagában kihívásnak tekinthető, mivel a legtöbb levél lényegesen kevesebb szót tartalmaz annál, mint amit a korábban tárgyalt minimális terjedelmi küszöbök megkövetelnének.[42]

Mi azonban főként nem erre, hanem a különböző digitalizálási módok (a manuális átírás – MAN, és automatikus HTR- és OCR-eljárások) szerzőazonosításban nyújtott teljesítményére koncentráltunk. Ezért a kísérletek során nem csupán egymástól függetlenül teszteltük az egyes módszerek szerzőazonosításra gyakorolt hatását, hanem a különböző eljárások között is. Célkitűzésünk ugyanis, hogy közelebb kerüljünk az irányítottságból adódó zavarok *(directionality artifacts)* megértéséhez; hiszen, ha az egyik digitalizációs eljáráson alapuló modellek jól alkalmazhatók más eljárásokkal átírt szövegek esetében is, akkor azok vonzóbbá válhatnak a jövőbeli projektek során. A szerzőség kérdését tárgyaló tanulmányokban használt szövegek jellemzően meglehetősen eltérő eredetűek, és nagyon különböző forrásokból és kiadásokból érkeznek, és/vagy a különböző módon (OCR és HTR) feldolgozott anyagok gyakran keverednek egymással. Az eltérő anyagok közötti irányultsági hatások megismerése lehetővé tenné, hogy a szövegosztályozás kontextusában hasznos ajánlásokat fogalmazzunk meg a jövőbeli adatgyűjtések számára.

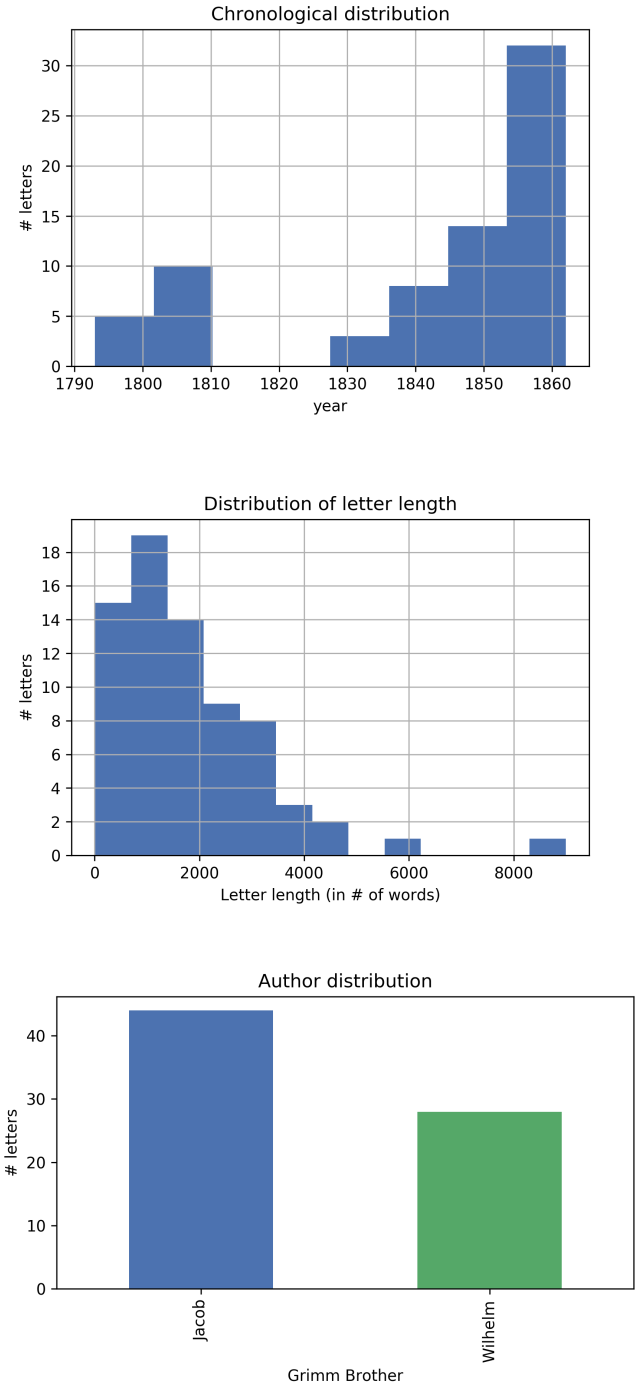
Az előkészítés során a nagybetűket kisbetűvé alakítottuk minden dokumentumban, azért, hogy csökkentsük a szórványjelenségek számát, továbbá eltávolítottuk az üdvözlő formulákból a testvérek neveit, amelyek megzavarhatják a szerzőazonosítást (például kivettük a „W.-t" egy olyan kifejezésből, mint a „Lieber W."). Mindegyik feldolgozás pontosan ugyanazt a 72 levelet tartalmazta, az ezekből készített statisztika a *7. ábrá*n látható. A legtöbb levél a lefedett időszak második feléből származik, bár számos ifjúkorban írt levelet is tartalmaz. A levelek átlagos hossza (szóhosszban) ~1,832, de a hosszúságok jelentős szórást mutatnak (SD = ~1,464). Fontos megjegyezni, hogy Wilhelmet mennyiségileg felülmúlja (n = 28 érték) amúgy is termékenyebbnek mondható testvére (n = 44). A módszert tekintve: a karakter n-gramok vizsgálata nemcsak az egyik legkorszerűbb szövegelemző stratégia a szerzőséggel foglalkozó tanulmányokban, hanem lehetővé teszi a modellek számára a finom, a szavak szintje alatti információ rögzítését is.[43] Az adatokat a TF–IDF (kifejezésgyakoriság–fordított dokumentumgyakoriság) szerint súlyozott vektortérben modelláltuk, az 5000 leggyakoribb, és egy dokumentumban legalább kétszer előforduló karakter n-gramok (bigram, trigram és tetragram) alapján. Végezetül az így kapott mátrixra a soronkén-

[41] Vö. Stamatatos, „A Survey."

[42] Eder, „Does Size Matter?"

[43] Kestemont, „Function Words;" Upendra Sapkota, Steven Bethard, Manuel Montes and Thamar Solorio, „Not All Character N-Grams Are Created Equal: A Study in Authorship Attribution," in *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies,* 93–102 (Denver, Colorado: Association for Computational Linguistics, 2015), https://doi.org/10.3115/v1/N15-1010.

ti és oszloponkénti normalizálást a stilometriában bevett módon alkalmaztuk (L1-normalizálás, illetve *feature scaling*).[44]

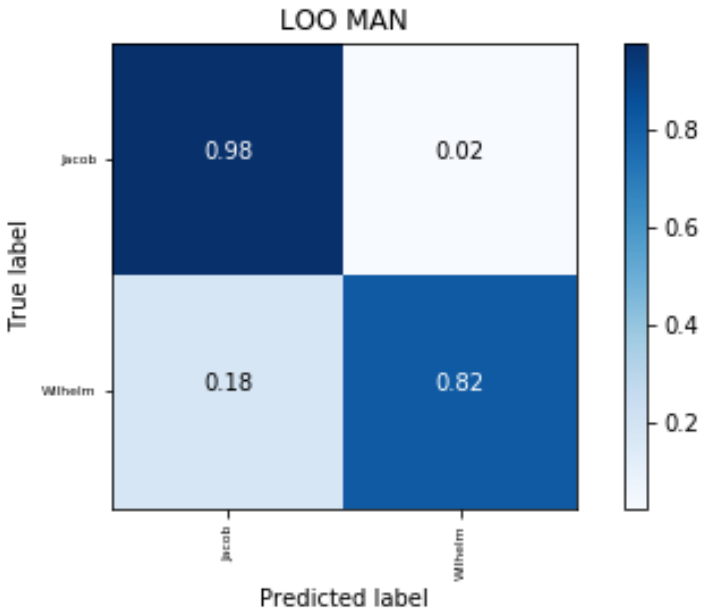


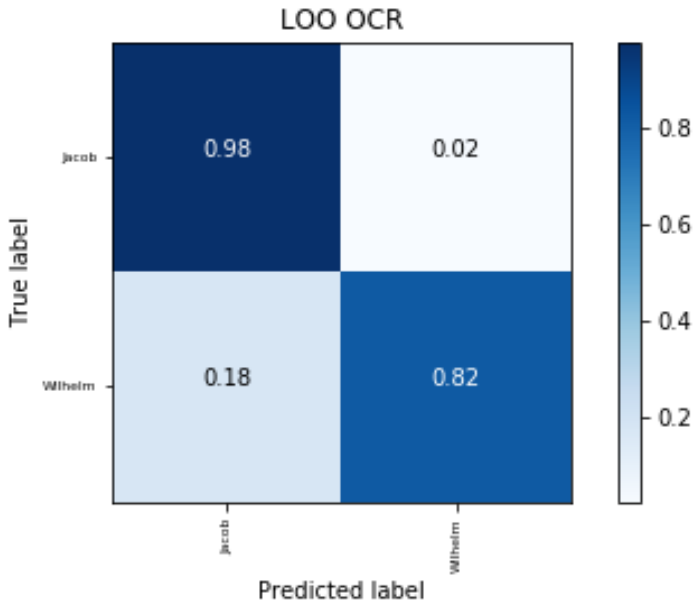

7. **ábra**. Általános információk az ebben a részben tárgyalt korpuszról: a levelek kronológiai, hosszbeli és szerzői címke szerinti megoszlása.

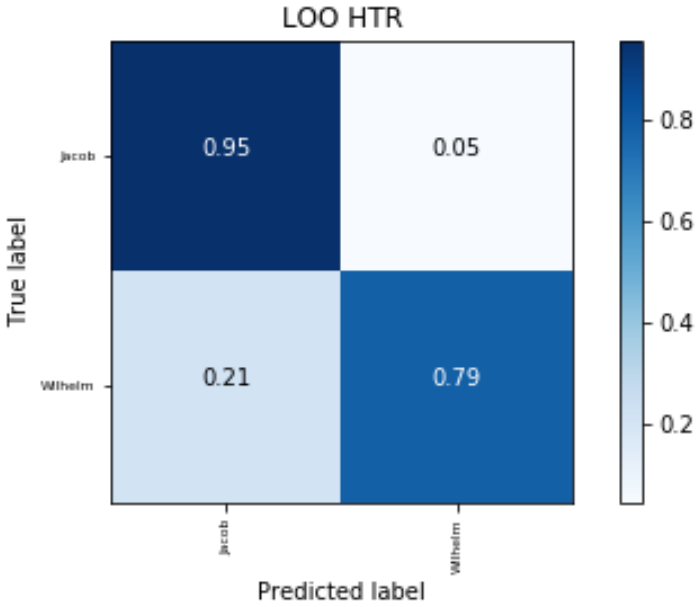
[44] Azaz lényegében Burrows Deltájának normalizációs eljárását követtük. Vö. Burrows, „‘Delta’.”

 DOI 10.31400/dh-hun.2021.5.3144

**4.2.2. Azonosítás az egyes eljárásokon belül *(Intra-modality attribution)*** ♦ Modellünk általános teljesítményének felméréséhez elsőként a fentebb tárgyalt beállításokkal létrehozott visszatartáson alapuló módszer (LOO) eredményeit személyenként közöljük az adatcsoportokra vonatkozóan (MAN, OCR, HTR). A hibamátrixokat a *8. ábra* és a *7. táblázat* mutatják: ezek részletezik az egyes adatkészletek pontosságát és F1-pontszámát. Általánosságban elmondható, hogy az eredmények viszonylag jók mindkét szempont tekintetében, de semmiképp sem tökéletesek – az SVM egyértelműen Jacob javára hajtja végre az azonosítást, valószínűleg a tanulókorpuszban való relatív túlsúlya miatt. A manuális feldolgozás (MAN) és az OCR eredményei megegyeznek, míg a HTR mindkét értékelési mutatóban rosszabbul teljesít.





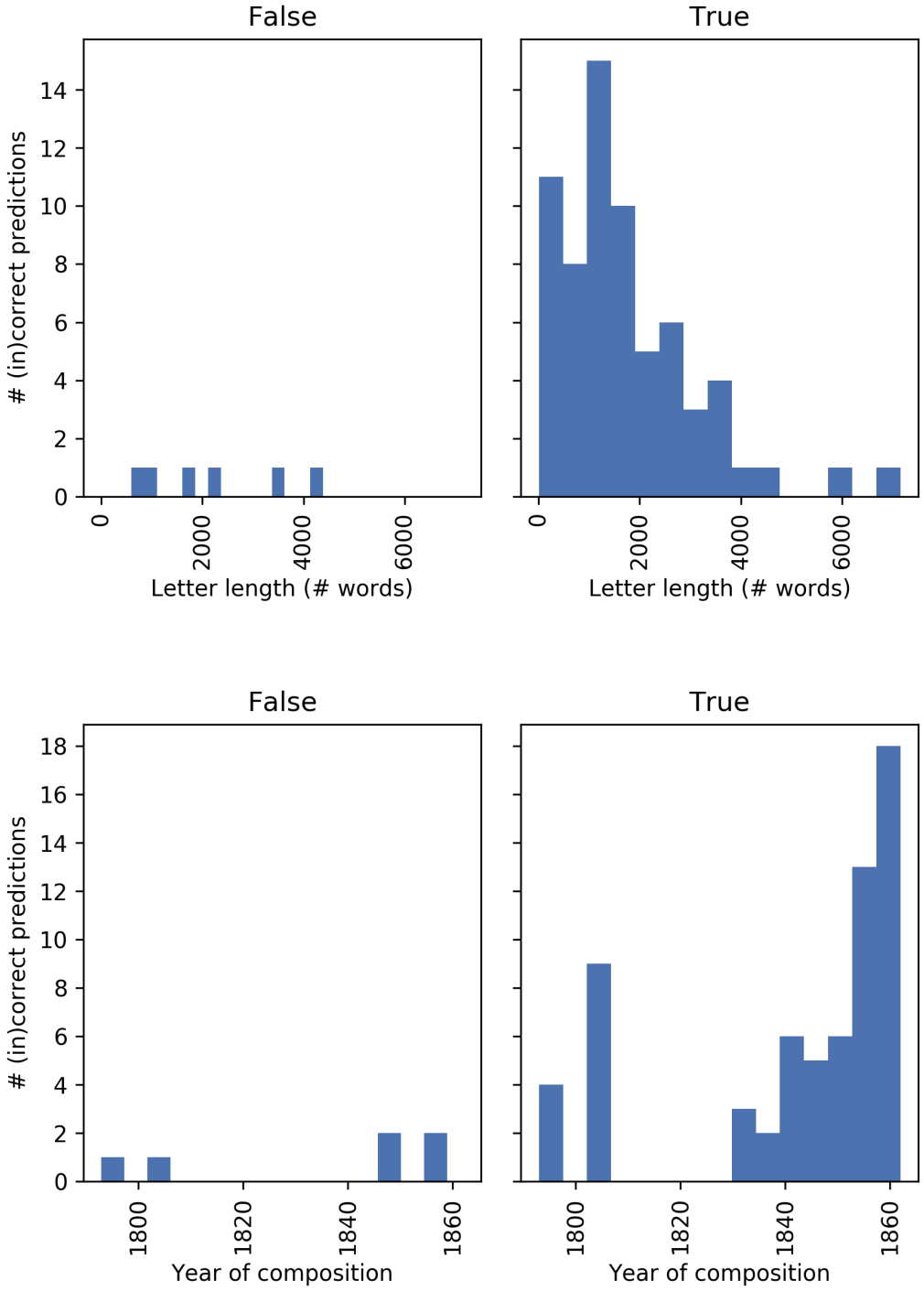

**8. ábra**. A visszatartáson alapuló módszer *(leave-one-out)* hibamátrixai a három adathalmazban (MAN, OCR, HTR).

**7. táblázat**. Az egyes eljárásokon belül végzett visszatartásos módszer eredménye a pontosság és az F1-pontszám tekintetében.

| | MAN | OCR | HTR |
|---|---|---|---|
| Accuracy | 91.66 | 91.66 | 88.88 |
| F1-score | 88.46 | 88.46 | 84.61 |

A MAN-ra vonatkozó előrejelzéseket – amelyeknek elvileg a legmegbízhatóbbnak kell lenniük – a levelek hosszának és keletkezési idejének függvényében közelebbről megvizsgálva azt látjuk, hogy a téves hozzárendelések – némileg meglepő módon – nem korlátozódnak a rövidebb levelekre (lásd *9. ábra*). Figyelembe véve a levelek általános időrendi megoszlását, azt is láthatjuk, hogy a hibás hozzárendelések az anyag teljes terjedelmében előfordulnak.

 DOI 10.31400/dh-hun.2021.5.3144




**9. ábra**. A helyes (igaz) és helytelen (hamis) azonosítások eloszlása a kézileg feldolgozott adatsorban: fent a karakterek száma, lent dátum szerint.

Miközben a dátum vagy a hossz nem tűnik döntő tényezőnek, az S1 táblázat *(Függelék, Kiegészítő anyagok)* alapján elmondható, hogy valójában ugyanazoknál a leveleknél rendszeres a rossz szerzőhöz való hozzárendelés. Míg az OCR és a manuális feldolgozás esetében teljesen konzisztensek ezek a hibák, a HTR néhány esetben kiszámíthatatla-

nul viselkedik. Érdemes megjegyezni, hogy az adathalmazok számos rendkívül rövid levelet is tartalmaznak (akár csupán nyolc szóból állót), amelyeken azonban a modellek helyesen végzik el a szerzőazonosítást – valószínűleg a Jacob-levelek irányába való torzítás miatt.

**4.2.3. Eljárásokon átívelő hozzárendelés *(Cross-modality attribution)*** ♦ A digitális bölcsészet kutatóinak sokszor különböző eredetű, és nem feltétlenül összeegyeztethető módon digitalizált szöveges anyagokat kell egymással kombinálniuk. Bár ez a gyakorlat egyértelműen nem optimális, gyakran elkerülhetetlen. Annak érdekében tehát, hogy felmérhessük a digitalizálási módoknak a szerzőazonosításra gyakorolt együttes hatását, egy eljárásokon átívelő kísérlethez fordultunk. Az adatok összehangolásához mindhárom adatkészletet egyszerre vektorizáltuk a korábbi módon, az 5000 leggyakoribb karakter n-gramot figyelembe véve. A LOO-módszert pedig a következő módon alkalmaztuk:

1. Az adathalmaz minden olyan levelének, amely mind a három adatkészletben megtalálható, meghatároztuk a három verzióját (MAN, OCR, HTR): ezek adták az elemzésekben a visszatartott tételeket.

2. Három különböző osztályozó eljárást tanítottunk be az így fennmaradó egyenként 71 levelére minden alkorpusz esetében.

3. Végül a három osztályozó alapján rendeltünk egy-egy szerzőt mindhárom visszatartott tételhez (összesen 9 szerzőattribúció).

Mindez lehetővé tett egy modelleken átívelő elemzést: hogy megvizsgálhassuk, az egyik digitalizációs eljárásra betanított modell miként teljesít, ha más eljárásokkal digitalizált szövegeken is tesztelik. Ez összesen kilenc betanításteszt kombinációját és összevetését eredményezi (HTR → HTR, HTR → MAN, HTR → OCR, MAN → HTR, MAN → MAN, MAN → OCR, OCR → HTR, OCR → MAN, OCR → OCR), beleértve azt is, amikor ugyanazt a modellt alkalmaztuk a betanításra és a tesztelésre is (azt azonban érdemes észben tartani, hogy az utóbbi eredmények eltérhetnek az előző részben bemutatottaktól a különböző vektorizációs megközelítések miatt). Ezen túlmenően fontos, hogy nyomon kövessük az irányultságot is: az „A" eljárással létrehozott korpuszon betanított modell teljesítménye a „B" eljárással digitalizált szövegre nem feltétlenül egyezik meg annak fordítottjával. Döntő fontosságú, hogy ez lehetővé teszi számunkra, hogy kezdeti ajánlásokat adhassunk arra vonatkozóan, mely modellt részesítsük előnyben a betanítás és a tesztelés során (ezek nem feltétlenül egyeznek meg).

A *10. ábra* négyzetes mátrixában a téves osztályozás kimeneteit mutatjuk be. Három dolgot figyelhetünk meg:

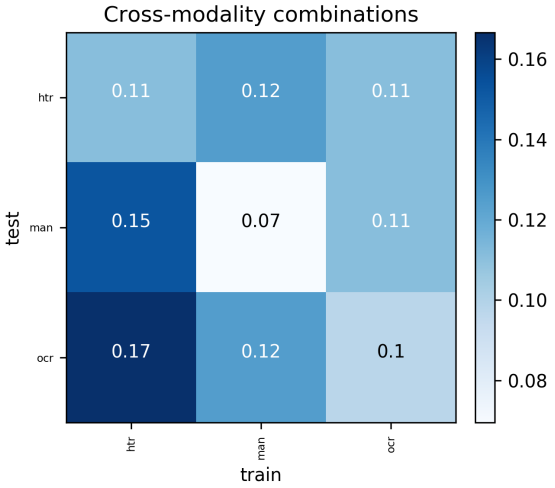
1. Az optimális kombinációt akkor érhetjük el, ha egy rendszert manuálisan átírt adatokon (MAN) tanítunk be és tesztelünk (ennek a hibaaránya: 0,07).

2. Szembeötlő, hogy a HTR-rel létrehozott szövegeken tanított modellek nem teljesítenek jól a háromfajta teszt egyikén sem (különösképpen az OCR-rel kombinálva) (hibaarány: 0,17).

 DOI 10.31400/dh-hun.2021.5.3144

> 3. Úgy tűnik, hogy az OCR viszonylag jól teljesít tanító modellként; sőt elvárásainkkal ellentétben az OCR-rel létrehozott szövegeken tanított és HTR-el létrehozott szövegen tesztelt modell még a manuális átíráson tanított modelleknél is jobb értékeket eredményezett.

Ezeket a megfigyeléseket szemlélteti az egyes levelekhez készített S2 táblázat *(Függelék: Kiegészítő anyagok)*. Ismét azt látjuk, hogy ugyanazoknak a leveleknek a szerzőségét azonosítják tévesen a modelleken átívelő különböző beállítások. Mindazonáltal ez a táblázat azt is mutatja, hogy ezen a szinten jellemzően a HTR-t magába foglaló modellek vezetnek hibás hozzárendeléshez.




**10. ábra**. Az eljárásokon átívelő hozzárendelés eredményei: a hibás klasszifikáció mátrixa.

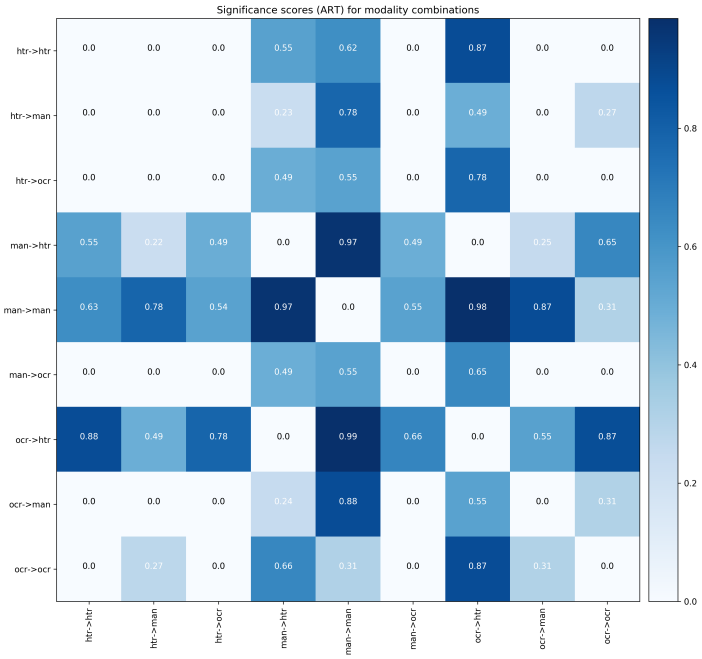
A kombinációk közötti eltérések számos értékes mintát mutattak, de bizonyos esetekben a különbségek felettébb aprók. Ezért fordultunk szignifikanciavizsgálatokhoz, hogy felmérhessük, a modellkombinációknál megfigyelt eltérések relevánsnak tekinthetők-e statisztikai szempontból. A különböző osztályozók eredményeinek szignifikanciavizsgálata vitatott téma a gépi tanulásban, különösen az olyan kis adathalmazok esetében, mint az itt vizsgáltak, ahol az osztálycímkék valódi eloszlása nem ismert vagy nem becsülhető meg megfelelően. Ezért a „közelítő randomizációs tesztelés" *(approximate randomization testing)* néven ismert megközelítést választottuk.[45] Ez a nem paraméteres teszt gyakori a számítógépes szerzőazonosításban[46] – például két olyan azonosítás eredményének összehasonlításakor, ahol nem lehet előzetesen felbecsülni a (potenciálisan rendkívül összetett) eloszlásokat. A teszt olyan pontszámot ad, amelynek segítségével meghatározhatjuk, hogy két bináris osztályozás eredménye

[45] Eric W. Noreen, *Computer-Intensive Methods for Testing Hypotheses: An Introduction* (New York: Wiley, 1989).

[46] Efstathios Stamatatos et al., „Overview of the Author Identification Task at PAN 2014," in Linda Cappellato, Nicola Ferro, Martin Halvey and Wessel Kraaij, *Working Notes for CLEF 2014 Conference,* 877–897 (Sheffield, 2014).

statisztikailag szignifikáns-e az F1-pontszám tekintetében. Ha ezek az értékek nem teszik lehetővé a nullhipotézis (H0) elutasítását, akkor az osztályozók *nem* szolgálnak szignifikánsan eltérő eredményekkel.

A pontszámokat táblázatos formában közöljük *(11. ábra)*. Az ábra nagyrészt megerősíti a korábbi megfigyeléseket: a HTR-rel feldolgozott szövegeken tanított modellek rosszabbul teljesítenek, és kimeneteik sem mutatnak jelentős eltérést egymástól. Amint azt a sötétebb színnel szedett sorok és oszlopok mutatják, a HTR-t tartalmazó kombinációk általában határozottan rosszabb eredményeket hoznak, mint azok melyek (kizárólagosan csak) manuális átíráson vagy OCR-en alapulnak. Emellett a világosabb cellák igazolják azt a feltevést is, hogy sok esetben az OCR-t tartalmazó kombinációk nem térnek el jelentősen a manuális feldolgozástól. Ez arra enged következtetni, hogy az OCR-digitalizálás megbízható helyettesítője lehet a jóval fáradságosabb manuális átírásnak – legalábbis ami a szerzőazonosítást illeti.

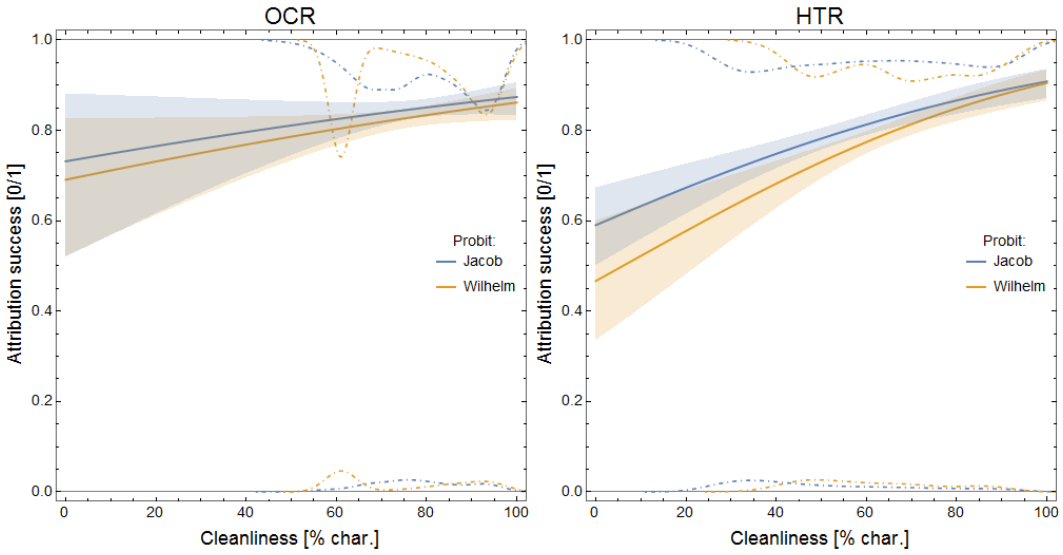



**11. ábra**. A különböző digitalizációs eljárások kombinációinak a szerzőazonosítás hatékonyságára kapott pontszámok táblázatos ábrázolása. A magasabb pontszámok azt jelzik, hogy a két rendszer jelentősen eltérő eredményeket produkál a többi kombinációhoz képest.

**4.2.4. Bináris azonosítás kontra szövegtisztaság** ♦ Az itt közölt kísérlet arra irányul, hogy feltárjuk, létezik-e összefüggés a Grimm-levelezés bináris szerzőazonosításának (tehát a szerzőség Jacobnak vagy Wilhelmnek tulajdonítása) sikere és az automatizált szövegfelismerés (OCR vagy HTR) között. Az alábbiakban kitérünk arra is, hogy mi történik, ha az azonosítás folyamata eltérő modelleket használ: tanítókorpuszként manuális átírást, tesztként pedig OCR-t vagy a HTR-t.

 DOI 10.31400/dh-hun.2021.5.3144

Ez esetben ahelyett, hogy a meglévő levelek tisztasága (lásd 4.1.4. rész) és az osztályozási eredmények közti összefüggést keresnénk, finomabb módon jártunk el, és véletlenszerű mintavételezéssel fejlesztettük a korpusz statisztikai megbízhatóságát. A tesztkorpuszt tehát a manuálisan átírt levelekből, véletlenszerű mintavételezéssel állítottuk össze, amely így 72 szöveget tartalmazott (eloszlásuk: 44 levél Jacobtól, 28 levél Wilhelmtől). Mindegyik legalább 1500 karakterből (nagyjából 250 szóból) áll. Ez lehetővé tette számunkra, hogy normalizáljuk a mintákat, miközben megőriztük az eredeti adatkészlet néhány reális tulajdonságát is. Ezt követően egy újabb 1500 karakter hosszú véletlen mintát hoztunk létre az automatikusan átírt (OCR, HTR) szövegek soraiból úgy, hogy nyomon tudtuk követni azok tisztaságát (azaz a helyesen átírt karakterek arányát) és osztályozni is tudtuk azokat (Burrows-féle Delta-távolság alkalmazásával a száz leggyakoribb szó alapján). Az eljárást (tanító és tesztkorpusz kijelölése, az osztályozás lefuttatása) szerzőkként és átírási módszerenként is addig ismételtük, míg összesen 3800 véletlenszerű mintát kaptunk, amelyeket tisztaságuk (0-tól 100%-ig terjedő skálán) és osztályozási eredményük (0 vagy 1 – helyes vagy helytelen szerzőazonosítás) jellemez. Ezeket az adatokat aztán a probitregresszió segítségével lehet elemezni, amint azt a *12. ábra* szemlélteti.




**12. ábra.** Véletlenszerű mintavételezésű szövegekre illesztett probitmodellek (folytonos vonalak); az árnyékolt területek a 0,95-ös megbízhatósági intervallumot jelölik. A pontozott vonalak grafikonjai a minták helyes (fent) és téves (lent) azonosítására vonatkoznak. Fontos, hogy a felső grafikon fejjel lefelé értelmezendő. Például az OCR esetében a Wilhelm-minták többsége 60%-os vagy 95%-os tisztaságú (a felső grafikon csúcsértékei), és 60%-os tisztaságnál ötször annyi helyes, mint helytelen azonosítás van (a felső és az alsó grafikon csúcsértékeinek aránya).

Az eredmény mindig a tanítókorpusz méretétől függ (valamint a testvérek részhalmazának relatív gyakoriságától). Ezért ellenőriztük, hogy a *12. ábrán* látható modellek megfelelnek-e az azonos méretű részhalmazok esetében is. Néhány (száz) minta még megbízhatatlan eredményre vezetett, ugyanakkor több ezer elegendő volt a konver-

gens, stabil eredmény eléréséhez. A mérések továbbá azt mutatják, hogy az OCR-eljárás tisztasága és a szerzőazonosítás sikere között csekély mértékű (0,95 – de nem 0,99 megbízhatósági szinten), míg a HTR-rel feldogozott írások tekintetében jelentős az összefüggés (0,99 megbízhatósági szint): ez esetben minél tisztábbak a szövegek, annál valószínűbb a helyes felismerés. Figyelemre méltó, hogy a körülbelül 20% feletti tisztaság már elegendő a HTR-rel feldolgozott szövegek szerzőjének a véletlennél nagyobb valószínűségű helyes azonosítására (OCR esetén a helyes felismerés valószínűsége mindig nagyobb a véletlennél).

## 5. Következtetések

A tanulmány a Jacob és Wilhelm Grimm leveleinek digitalizációja során keletkező zaj automatikus szerzőazonosításában mért hatásáról ír. Az összes lehetséges digitalizálási „forgatókönyv" teszteléséhez három különböző kimenetet hasonlítottunk össze: 1. az eredeti levelek manuális átírását; 2. a Grimm-levelek 2001-es nyomtatott, kritikai kiadásának OCR-feldolgozását; és 3. az eredeti levelek automatikus átírásához készült HTR-modellt. A manuális átírást etalonkorpuszként használtuk az OCR- és a HTR-feldolgozás tisztaságának értékeléséhez. A várakozásoknak megfelelően a HTR hibaaránya magasabb volt az OCR-nél a kézírás esetlegessége miatt (szemben a nyomtatás egységességével). Mindezeken túl a kísérletek azt mutatták, hogy a HTR-feldolgozásra messze nem tökéletes adatkészlet ellenére is átlagosan 6%-nál kevesebb karakterhibaaránnyal dolgoztak a Grimm testvérek számára létrehozott modellek (azaz minden tizenhetedik karaktert olvasták be hibásan).

Az ilyen hibaarány már önmagában elég magas ahhoz, hogy jelentősen csökkentse a HTR-rel feldogozott levelek szókincsbeli gazdagságát. Mivel ez a testvérek szerzőségét tekintve megkülönböztető tényező, megvizsgáltuk a három különböző digitalizálási kimenet (manuális átírás [MAN], zajos OCR és zajos HTR) hatását a szerzőazonosításra. Ennek eredményeként megállapítottuk, hogy (legalább a szerzőazonosítás során) az OCR-rel végzett digitalizálás megbízható alternatívaként szolgál a gondosabb manuális átiratok mellett is.

Érdekes, hogy a hozzárendelés akkor is működőképesnek tűnik, ha a tanító és a tesztkorpuszokat különböző módon digitalizált szövegekből építik fel. Ami a HTR-t illeti, a kutatásaink arra vezettek, hogy ugyan az automatikus átírás jelentősen növeli a szöveg téves osztályozásának kockázatát az OCR-hez képest, már körülbelül 20% feletti szövegtisztaság is elegendő ahhoz, hogy a véletlennél nagyobb valószínűséggel legyen sikeres a bináris azonosítás (OCR esetén ez a valószínűség mindig nagyobb).

Habár eredményeink még kezdetlegesek, érvként szolgálhatnak abban a diskurzusban, ami a szöveg abszolút tisztaságát nem a szerzőazonosítás elsődleges feltételének teszi meg[47] – mindenesetre a Grimm testvérek által írt levelek kapcsán. Az általunk létrehozott HTR-modell az első olyan modell, amely a Grimm fivérek kézírásának felismerésére jött létre – ami még tovább finomítható több kézzel írott dokumentum (például a jelenleg Berlinben található szakmai levelezés) betáplálásával. Kitekintésként: egy következő kutatási téma lehetne a közös szerzőség vizsgálata a *Gyermek- és*

[47] Vö. Eder, „Does Size Matter?"

 DOI 10.31400/dh-hun.2021.5.3144

*családi mesék (Grimm's Kinder und Hausmärchen)* című mű esetében, azt ellenőrzendő (már ha egyáltalán lehetséges), hogy mely mesékben érvényesül markánsabban Jacob, illetve Wilhelm szerzői ujjlenyomata.

Fordította: Kustos Júlia

## Attributing Authorship in the Noisy Digitized Correspondence of Jacob and Wilhelm Grimm

This article presents the results of a multidisciplinary project aimed at better understanding the impact of different digitization strategies in computational text analysis. More specifically, it describes an effort to automatically discern the authorship of Jacob and Wilhelm Grimm in a body of uncorrected correspondence processed by HTR (Handwritten Text Recognition) and OCR (Optical Character Recognition), reporting on the effect this noise has on the analyses necessary to computationally identify the different writing style of the two brothers. In summary, our findings show that OCR digitization serves as a reliable proxy for the more painstaking process of manual digitization, at least when it comes to authorship attribution. Our results suggest that attribution is viable even when using training and test sets from different digitization pipelines. With regard to HTR, this research demonstrates that even though automated transcription significantly increases risk of text misclassification when compared to OCR, a cleanliness above 20% is already sufficient to achieve a higher-than-chance probability of correct binary attribution.

Keywords:
stylometry, authorship attribution, german literature, Grimm, digitization, OCR, HTR

## Köszönetnyilvánítás

A szerzők köszönettel tartoznak kollégáiknak: Maria Moritznak, Kirill Bulertnek, Marco Büchlernek és volt kollégájuknak Linda Brandtnak a projekthez nyújtott értékes hozzájárulásukért. Ezúton is szeretnék megköszönni a németországi Kasselben a Brüder Grimm-Gesellschaft e.V munkatársainak: Bernhard Lauernek és Rotraut Fischernek a szakértői tanácsokat és a Grimm testvérek kalligráfiájával kapcsolatos tudásuk megosztását. Köszönik dr. Stephan Tulkensnek a 4.2.3. alfejezet kutatásában nyújtott támogatását, valamint dr. Günther Mühlbergernek a Grimm-kézírás HTR-modellezésében nyújtott nélkülözhetetlen hozzájárulását. Végül külön köszönet illeti Gerhard Lauer professzort állandó javaslataiért és támogatásáért.

## Függelékek

A cikkhez készült kiegészítő anyag (*S1 táblázat:* Az egyes digitalizációs eljárásokon belül létrehozott modellek eredményeinek az áttekintése; *S2 táblázat:* A eljárásokon átívelő kísérletek eredménye) hozzáférhető az alábbi címen:

> `https://www.frontiersin.org/articles/10.3389/fdigh.2018.00004/full#supplementary-material`, valamint a jelen cikk mellékleteként annak adatlapján: `http://doi.org/10.31400/dh-hun.2021.5.3144`.

Jacob Grimm 1793-as levelének manuális leirata (vö. 3.2.1.2. alfejezet):

> Montag
>
> Steinau den 7 8br 1793.
>
> Lieber Bruder!
>
> Du wirst hierbey dein Kleid erhalten, Wie hatt es dir denn auf der Reise gefallen, mich verlanget es zu wissen, ich erwarte mit der ersten Gelegenheit einen Brief von dir, seit deiner Abwesenheit ist nichts merkwürdiges vorgefallen.
>
> Mein Vater hatte heute einen sehr starken Amtstag gehabt, bis Freitag wird dich unser hofjud Jud Seelig besuchen und mit diesem werde ich dir weitläufliger schreiben, Küβe der lieben Mutter dem Groβvater und jungfer Tante die Hand in meinem Nammen. Du wirst von uns allen gegrüβet, und ich bin dein treuer
>
> Bruder
>
> Jacob Grimm

 DOI 10.31400/dh-hun.2021.5.3144