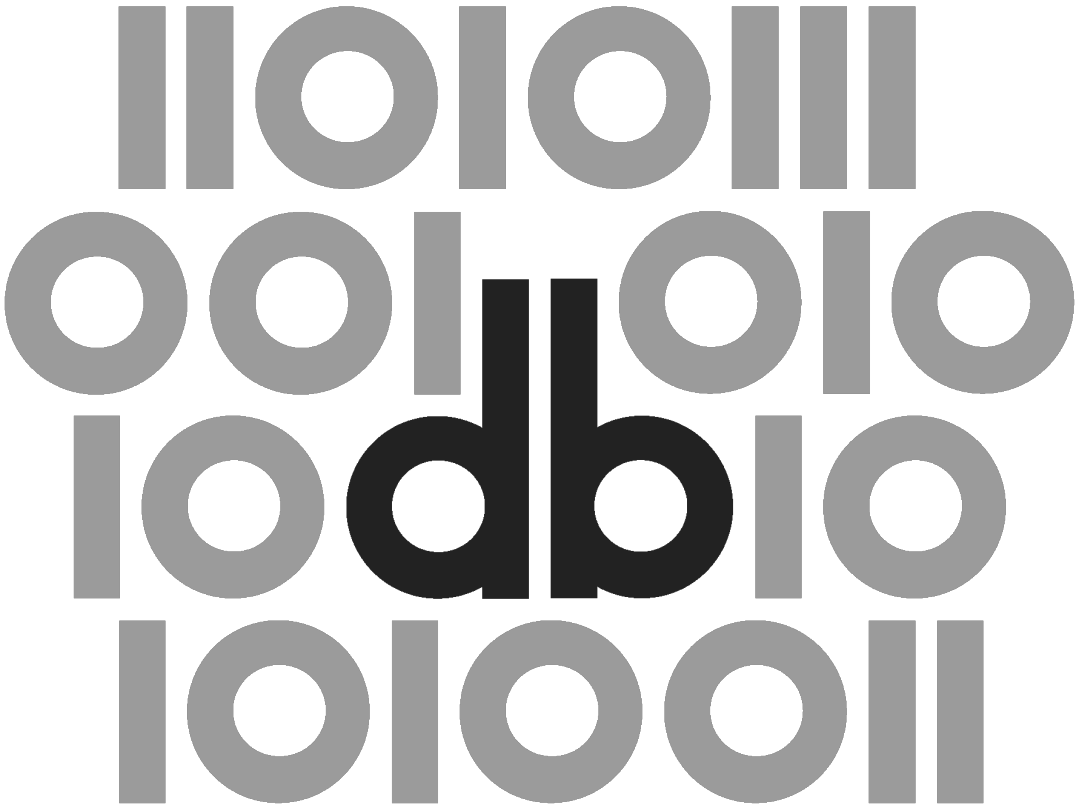
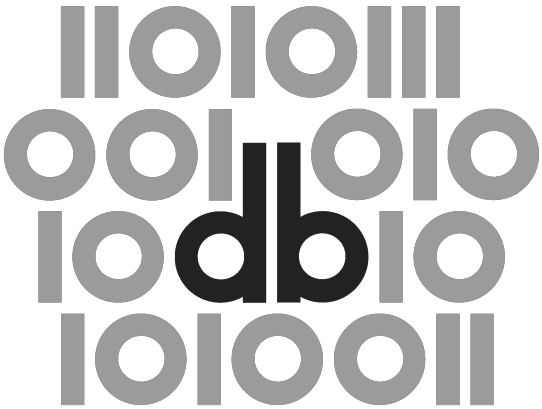
2 (2019) <DIGITÁLIS BÖLCSÉSZET>



2 (2019) </DIGITÁLIS BÖLCSÉSZET>

Digitális Bölcsészet
2019., második szám

<DIGITÁLIS BÖLCSÉSZET>



2 (2019)

ISSN 2630-9696
DOI 10.31400/dh-hun.2019.2
Kiadja az ELTE BTK Régi Magyar Irodalom Tanszéke (1088 Budapest, Múzeum krt. 4/A) és a Bakonyi Géza Alapítvány.
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 Leváltár biztosítja.

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

Tördelés: Hegedüs Béla
Grafika: Hegyi Gábor

<MŰHELY>

Labádi Gergely †

Szegedi Tudományegyetem, Magyar Irodalmi Tanszék

Géppel mért irodalom: a mikszáthi élőbeszédszerűség*

Bár az élőbeszédszerűség terminus a Mikszáth-szakirodalomban *A tót atyafiak* és *A jó palócok* sikere óta folyamatosan jelen van, sőt az életmű egészének fő jellegzetességévé emelkedett, a fogalmat használók nem próbálták meg részletesebben kifejteni, mit értenek rajta. A tanulmány kiindulópontja az, hogy az élőbeszédszerűségnek vannak számszerűsíthető nyelvi jellemzői, hiszen a szakirodalom elég egyértelmű és mérhető különbséget tételez a lejegyzett „irodalmi" és „beszélt" nyelv között. A dolgozat Mikszáth Kálmán és Jókai Mór egyes szövegein végzett morfológiai vizsgálatok eredményeit mutatja be.

Kulcsszavak:
regény, Mikszáth Kálmán, morfológiai elemzés, élőbeszédszerűség, szókincsgazdagság

Tahin Szabolcs 2003-ban megjelent tanulmányának címében – „»Élőbeszédszerűség« Mikszáth prózájában"[1] – nem véletlenül tette idézőjelbe az élőbeszédszerűség kifejezést. Bár a terminus a Mikszáth-szakirodalomban *A tót atyafiak* és *A jó palócok*[2] sikere óta folyamatosan jelen van, sőt az életmű egészének fő jellegzetességévé emelkedett, a fogalmat használók nem próbálták meg részletesebben kifejteni, mit értenek rajta, mindvégig megmaradt a mintha-élmény: „[…] szinte készek volnánk esküt tenni rá, hogy a szép regéket ő maga beszéli el élő szóval. […] A falu vén regélőjére gondolunk, aki téli estéken a kukoricafosztásra összegyűlt népet mulattatja mondásaival."[3]; „Mintha csakugyan eleven beszéd csengene fülembe olvasásakor…"; „Mintha nem is volna köztünk az a nagy távolság, mely írót és olvasót elválasztja…"; „Mintha minden szó egyenesen az ajkáról lebbent volna a papírosra…"; „A mese is mintha önkéntelenül buggyanna ki lelkéből…"[4] Tahin ugyanakkor nem marad a szakirodalom foglya, és nemcsak azért, mert egyértelművé teszi, Schöpflinnek és Bartának az élőbeszédszerűségből levezetett értelmezéseinek helyi értéke ma már meglehetősen

* A kutatást az EFOP-3.6.1–16–2016–00008 azonosítójú, EU társfinanszírozású projekt támogatta 2017-ben.

1 Tahin Szabolcs. „»Élőbeszédszerűség« Mikszáth prózájában," *Tiszatáj* 57, 11. sz. (2003): 53–71.

2 Mikszáth Kálmán, *A tót atyafiak* (Budapest: Grimm, 1881). Mikszáth Kálmán, *A jó palócok* (Budapest: Légrády,1882).

3 Rudnyánszky Gyula, „Mikszáth Kálmán új könyvéről," in *Mikszáth Kálmán Összes Művei* 32. köt. szerk. Bisztray Gyula, Király István (Budapest: Akadémiai Kiadó, 1968), 366–367; Tahin, „»Élőbeszédszerűség«," 54.

4 Schöpflin Aladár, *Magyar Írók* (Budapest: Nyugat, 1917), 42–43; Tahin, „»Élőbeszédszerűség«," 54.

kétes (pl. „operett íz"), hanem azért, mert feltételezi, ez az írói stratégia része lehetett. Tahin újítása, hogy a Mikszáth-szövegek elbeszélői szerepeihez kapcsolva vizsgálja az élőbeszédszerűséget. Márpedig Mikszáth szövegeiben rendszerint többféle elbeszélői hang vegyül: vannak olyanok, amelyek kifejezetten az élőbeszédhez kapcsolódnak, de vannak olyanok is, amelyek egy írásos kultúrát feltételeznek. Az elbeszélői hangok közötti különbségek pedig, mint a például felhozott *Szent Péter esernyője*[5] kapcsán Tahin igazolja, elég látványosak és meggyőzők.

Ha megnézzük a fent idézett részleteket, akkor egyértelmű, hogy a szakirodalomban oly sokszor ismételt élőbeszédszerűség az írásbeliséggel, az írásbeli kultúrával áll ellentétben. Barta János ennek irodalomtörténeti jelentőségét abban látta, hogy Mikszáth, Jókait követve lebontja a magyar regényírói hagyomány retorikusságát: „A magyar prózai epikának erős retorikus hagyományai vannak; Eötvös, Kemény írásait éppen avult retorikus jellegük teszi ma nehezen olvashatóvá. Ezt a hagyományt Jókai rendíti meg, és Mikszáth számolja fel teljesen."[6] Ez a tétel ugyanakkor a Mikszáth-értés egyik nagy keresztje – mutat rá Milbacher Róbert –, mivel egyrészt a „nagy mesélő" Jókaihoz köti Mikszáthot, másrészt pedig egy korszerűtlennek ítélt, naiv, reflexiótlan Mikszáth-próza tétele következik belőle.[7]

Kétségtelen, hogy egyes elbeszélői hangokhoz kötni az élőbeszédszerűséget termékeny stratégia, maguk a szakirodalmi idézetek is utalnak rá (pl. „a falu vén regélője"), de amint arra Tahin is felhívja a figyelmet, *A Noszty fiú…*[8] utószavában maga Mikszáth is a kötetlenül társalgó asztaltársaság fiktív befogadói szituációját ajánlja olvasóinak.

Mindazonáltal az élőbeszédszerűséget nemcsak így lehet vizsgálni. Az írásbeliség és az élőbeszéd nyelvi sajátosságait a nyelvészek elég pontosan leírták. Érsok Nikoletta összefoglalásában[9] a beszélt nyelvre a következők jellemzők: az igék magas arányban szerepelnek, de a melléknevek, jelzők aránya alacsony; a módosítószavak, kötőszavak aránya magasabb, mint írásban, gyakoriak a névmások, határozószók. Rövidebb szintaktikai szerkezetek, félbemaradt mondatok jellemzik a beszédet, gyakoriak a megszólítások, az egyszerű, egytagú mondatok, kevés a többszörösen összetett mondat, de (számomra) meglepő módon az alá- és mellérendelő összetett mondatok számarányában nem figyelhető meg lényeges különbség. Ahogy Érsok összefoglalja: „A fentiekből kifolyólag az írott nyelvi szövegek hosszabb, teljes mondatokból állnak, amelyek egymástól egyértelműen elkülöníthetők. A grammatikalitás, jólformáltság, korrektség és exaktság [sic!], továbbá a választékosabb, a normát követő szóhasználat jellemzi az írott nyelvet."[10]

A számítógépes nyelvészet ma rendelkezésre álló eszközeivel e sajátságok nagy része könnyen mérhetővé, azaz ellenőrizhetővé vált. Mindez természetesen nem azt jelenti, hogy az elbeszélői hangokhoz kötötten vizsgált élőbeszédszerűség tételét el kellene vetni, vagy, hogy így „objektívabb" elemzéseket készíthetünk. Egyszerűen an-

[5] Mikszáth Kálmán, *Szent Péter esernyője* (Budapest: Légrády, [1895]).
[6] Barta János, „Mikszáth-problémák (Első közlemény)," *Irodalomtörténeti Közlemények* 65, 2. sz. (1961): 140–161, 142; Tahin, „»Élőbeszédszerűség«," 58–59.
[7] Milbacher Róbert, „A Mikszáth-befogadás főbb irányairól," *Tiszatáj* 65, 11. sz. (2011): 80–81.
[8] Mikszáth Kálmán, *A Noszty fiú esete Tóth Marival* (Budapest: Franklin, 1908).
[9] Érsok Nikoletta Ágnes, „Szóbeliség és/vagy írásbeliség," *Magyar Nyelvőr* 130, 2. sz. (2006): 165–176.
[10] Érsok, „Szóbeliség és/vagy írásbeliség," 166.

 DOI 10.31400/dh-hun.2019.2.390

nak a lehetőségét kínálja fel a szövegek számítógépes vizsgálata, hogy a korábbiakhoz képest több és más jellegű adatra építve értelmezhessünk irodalmi jelenségeket. Jelen esetben engem az érdekel, az első két sikeres Mikszáth-kötet, *A jó palócok* és *A tót atyafiak* mérhető nyelvi sajátságai miként viszonyulnak Jókai novelláihoz, illetve saját, 1874-es elbeszéléskötetéhez, valamint, hogy egy kései példát is hozzak, az *Öreg szekér, fakó hám* novelláihoz.[11]

1. A korpusz előkészítése

A vizsgált novellakorpusz a következő: Jókaitól az 1856-os *Árnyképek* című kötet (8 novella), a *Dekameron* 1860-ban megjelent tizedik kötete (15), valamint az 1894-es *Athenaeum*-olvasótár szövegei (5) – kíváncsi voltam ugyanis egy olyan válogatásra, amely az élőbeszéd „diadalát" hozó Mikszáth-kötetek után készült.[12] Mikszáthtól az 1874-es *Elbeszélések* (8 novella), az 1881-es *A tót atyafiak* (4), az 1882-es *A jó palócok* (15) és az 1901-ben kiadott *Öreg szekér, fakó hám* (15).[13] Összesen huszonnyolc elbeszélés Jókaitól, negyvenkettő Mikszáthtól. A vizsgálatokhoz a Magyar Elektronikus Könyvtárban elérhető szövegeket választottam. A szövegek egy részét az Arcanum digitalizálta, más részét a Project Gutenberg magyar csapata. Bár textológiai szempontból jogos kritikák érhetik, de mivel egyrészt az Arcanum munkái képezik az interneten ma hozzáférhető magyar irodalmi korpusz alapját, valamint az átírás egységes szempontok alapján és egyenletes minőségben készült, valamint az Országos Széchényi Könyvtár mint befogadó intézmény mégiscsak hitelesíti, végül úgy döntöttem, kiindulásként e kísérletben érdemes elfogadni ezeket. Magam még annyit módosítottam a szükségesnek ítélt elemzés pontossága érdekében, hogy a legsúlyosabb, feltehetően a karakterfelismerés során elkövetett tévesztéseket javítottam (pl. *m* helyett *rn* – és fordítva), illetve a címeket, hogy biztosan külön elemezze őket a program, írásjellel láttam el – ha kellett. A szövegek egykorú helyesírását, ha az az értelmezést befolyásolhatta, modernizáltam, tehát például az *aszszonynyal* alakot *asszonnyal*-ra, az *a mit*, ha nyelvtanilag indokolt volt, *amit*-re javítottam.

Az elemzéshez a korpuszt a *Magyarlanc* elnevezésű nyelvi elemzővel preparáltam.[14] Az alábbi képen *A jó palócok* első novellájának *Magyarlanc*cal elemzett első néhány

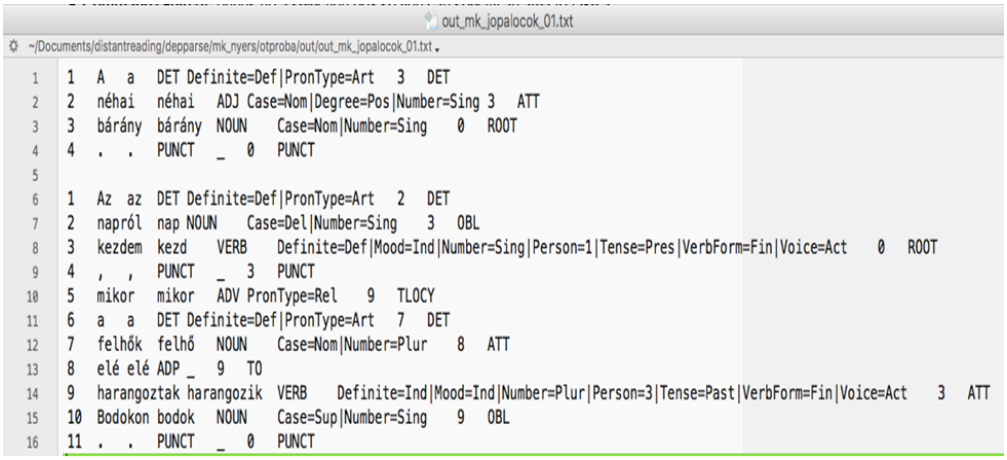
[11] Mikszáth Kálmán, *Öreg szekér, fakó hám* (Budapest: Légrády, 1901).

[12] Jókai Mór, *Árnyképek* (Pest: Emich Gusztáv, 1856), `https://mek.oszk.hu/07300/07324`; Jókai Mór, *Dekameron,* `https://mek.oszk.hu/00800/00845`, eredeti példány: *Jókai összes művei*, CD-ROM (Budapest: Arcanum, 2001); Jókai Mór, *Az egyhuszasos leány, Száz leány egy rakáson és egyéb elbeszélések* (Budapest: Athenaeum, 1894), `https://mek.oszk.hu/16300/16372`.

[13] Mikszáth Kálmán, *Elbeszélések* (Budapest: Vodiáner, 1874), `https://mek.oszk.hu/15400/15499`. Mikszáth Kálmán, *Tót atyafiak,* `https://mek.oszk.hu/00800/00895`, eredeti példány: Mikszáth Kálmán, *Gavallérok; Tót atyafiak: elbeszélések* ([Szentendre]: Interpopulart, 1995); Mikszáth Kálmán, *A jó palócok,* `https://mek.oszk.hu/00900/00950`, eredeti példány: Mikszáth Kálmán, *Tót atyafiak, A jó palócok* (Budapest: Móra, 1978); Mikszáth Kálmán, *Öreg szekér, fakó hám* (Budapest: Légrády, 1901), `https://mek.oszk.hu/11500/11563`.

[14] János Zsibrita, Veronika Vincze and Richárd Farkas, „magyarlanc: A Toolkit for Morphological and Dependency Parsing of Hungarian," in *International Conference on Recent Advances in Natural Language Processing,* eds. G. Angelova, K. Bontcheva and R. Mitkov (Shumen: Incoma Ltd., 2013), 763–771. *Magyarlanc,* hozzáférés: 2017.08.10, `http://www.inf.u-szeged.hu/rgai/magyarlanc`

mondata látható. A többféle elemzési lehetőség közül én a *depparse*-ot választottam, mivel ez nemcsak morfológiai elemzést végez, hanem mondattanit is.



1. ábra. *A korpusz elemzése a Magyarlanc alkalmazásával*

Az első oszlop az adott sztring mondatbeli helyét számolja, a második az eredeti szöveget tartalmazza, a harmadik a szóalakok lemmatizált, szótári alakját, a negyedik szófaját, az ötödik a szóalak morfológiai elemzését, a hatodik azt, hogy az adott szó a mondatban melyik másik szó alá van rendelve, hol van a csomópontja, a hetedik pedig a mondatbeli funkcióját. Mint a mellékelt képből látszik, a program ugyan nem mindent elemez pontosan, hiszen „bodok" valójában tulajdonnév (PROPN), nem pusztán főnév (NOUN), de azon a szinten, amire szükségem van, megfelel. Nagyon kis százalékban, de néhány, ma már nem használatos kifejezés vagy szó-, illetve ragozási alak esetében a *Magyarlanc* nem tudja értelmezni a szófajt, de ez szintén a 19. századi szövegek sajátosságából fakad, így az eredményen érdemben nem változtat.

2. Megmérni az élőbeszédszerűséget

Érsok fent idézett tanulmánya alapján elég jól meghatározhatók az élőbeszédszerűség mérhető elemei. A választékosságot a TTR, azaz a típus–jel-arány mérésével lehet számba venni: hány egyedi szóból (a képletben: V) és hány szóalakból (a képletben: N) áll a szöveg. Minél választékosabb egy szöveg, annál magasabb az egyedi szavak száma. De itt vannak még a szófaji arányok is, a mondatok hosszúságának mérése, szóval meglepően sok mindent meg lehet mérni.

3. Választékosság

Egy másik tanulmányban már foglalkoztam a szókincsgazdagság mérésének néhány aspektusával.[15] A legnagyobb probléma, hogy a szöveg hosszúságával megnő az ismétlés valószínűsége, tehát különböző hosszúságú szövegek összemérése egy egyszerű osztással, hamis eredményekhez vezet. A magyar szakirodalomban Zsilka Tibor a

[15] Labádi Gergely, „Az olvasó gép: Berzsenyi Dániel versei távolról," *Digitális Bölcsészet* 1, 1. sz. (2018): 17–34. http://doi.org/10.31400/dh-hun.2018.1.126.

 DOI 10.31400/dh-hun.2019.2.390

szókincsgazdaság mérésére Pierre Guiraud képletét alkalmazza – a képletet mások is elfogadják[16] –, amely, ahogy a szerző ígéri, kiküszöböli a szöveghosszból fakadó eltéréseket:

$$R = \frac{V}{\sqrt{N}}$$

Forrásai, de főleg saját vizsgálata nyomán Zsilka úgy gondolja, hogy 2000 szó alatti szövegek vizsgálatára még ez a képlet sem alkalmas. Említett cikkemben foglalkoztam e kijelentés mögött álló súlyos módszertani hibával. Meglepő módon a felvett szövegek között vannak kevesebb, mint 2000 szóalakot tartalmazó novellák. Bár teljesen véletlenszerűen választottam őket, Jókai szövegei között tizenhét, kevesebb mint 2000 szóalakból megalkotott novella található, a jóval nagyobb Mikszáth-korpuszban pedig tizenöt. Guiraud képlete így tulajdonképpen vizsgázni fog, más képletekkel is kiszámolom a szövegek „választékosságát", még ha az újabb elképzelésekkel,[17] illetve azzal, hogy érdemes-e egyáltalán efféle zárt korpuszként felfogni a szókincset, nem is foglalkozom.[18]

Ha megnézzük az alábbi grafikont, láthatjuk, hogy Jókai 1854-es kötetének folyamatosan magas szóalakszámával szemben az 1860-as kötet szinte minden írása 2000 szóalak alatti, az 1894-es viszont meglehetősen vegyes képet mutat. Utána helyezve a Mikszáth-szövegek szóhosszúságát mutató ábrát, érdekes következtetésre juthatunk. Mikszáth tulajdonképpen *A jó palócok*kal jutott el a rövidebb novellákig (a kötet alcíme: *Tizenöt apró történet*), előtte maga is közel olyan hosszúakat írt, mint a korai, korábbi Jókai-szövegek. Legalábbis ezen korpusz alapján úgy tűnik, Jókai kezdeményezte a rövidebb rövidtörténeteket a 19. század második felének magyar irodalmában. Erre a következtetésre juthatunk, ha a két korpusz egy-egy adatát szintén figyelembe vesszük. Jókai 28 novellája összesen 100302 szóalakot tartalmaz, azaz novellánként átlag 3582-t. Mikszáth jóval nagyobb korpusza 148879 szóalakot tartalmaz, ami novellánként alig valamivel kevesebb: 3545. Meglehetősen nagy a novelláskötetek által átfogott időszak, ezért a következtetés levonására a számok csak óvatosan alkalmasak: mindazonáltal utólag Mikszáth és Jókai írásművészete „egybemosódhatott" – más szóval, a Barta felvetette fejlődéstörténeti ív,[19] számszerű igazolást kaphat.

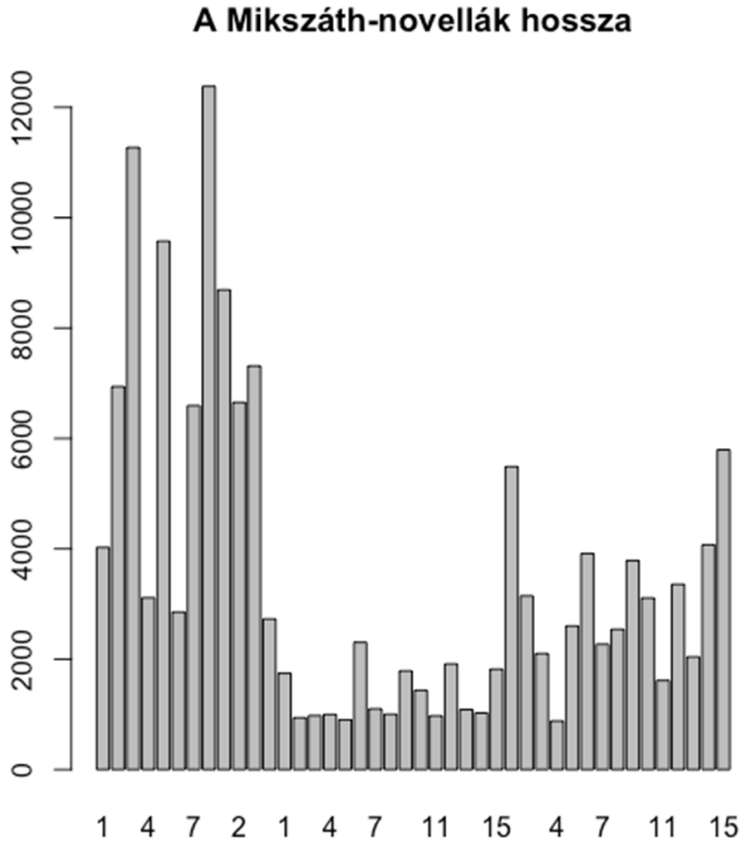
[16] Zsilka Tibor, *Stilisztika és statisztika* (Budapest: Akadémiai Kiadó, 1974). Guiraud munkájáról részletes ismertető olvasható magyarul: J. Soltész Katalin, „Guiraud statisztikai módszere a szókincs vizsgálatában," in *Általános nyelvészeti tanulmányok I.,* szerk. Telegdi Zsigmond (Budapest: Akadémiai Kiadó, 1963), 263–272.

[17] David Mitchell, „Type-token Models: a Comparative Study," *Journal of Quantitative Linguistics* 22, 1. sz. (2014), 1–21, http://doi.org/10.1080/09296174.2014.974456.

[18] András Kornai, „How many words are there?" *Glottometrics* 2, 4. sz. (2002): 61–86.

[19] Jókai és Mikszáth együtt bontják le a magyar próza retorikus hagyományát. Barta, „Mikszáth-problémák," 142.

2. ábra. *A szóalakszám változása Jókai vizsgált köteteiben*



3. ábra. *A szóalakszám változása Mikszáth vizsgált köteteiben*

 DOI 10.31400/dh-hun.2019.2.390

A szókincsgazdaság esetében persze kérdés, hány egyedi lemmából valósítja meg az író a novellát. Lejjebb látható, hogy a Guiraud-féle képlet, bár valamelyest igen, de érdemben nem csökkentette a novellák hosszúságából fakadó különbséget – Zsilkának tehát, módszertani hibája ellenére, igaza van. A Herdanhoz köthető logaritmusos képlet

$$R = \frac{logV}{logN}$$

viszont igen, amennyiben feltételezzük, hogy egy szerző stílusára a TTR többé-kevésbé jellemző – és állandó: a szókincs a szöveg hosszával kétségtelenül nő, de a függvény alapján egyre lassuló mértékben.[20]

1856	1	2	3	4	5	6	7	8
lemma	3742	1757	2397	2237	1871	3146	2034	2243
szóalak	14302	4955	7663	7900	5642	10994	6408	7098

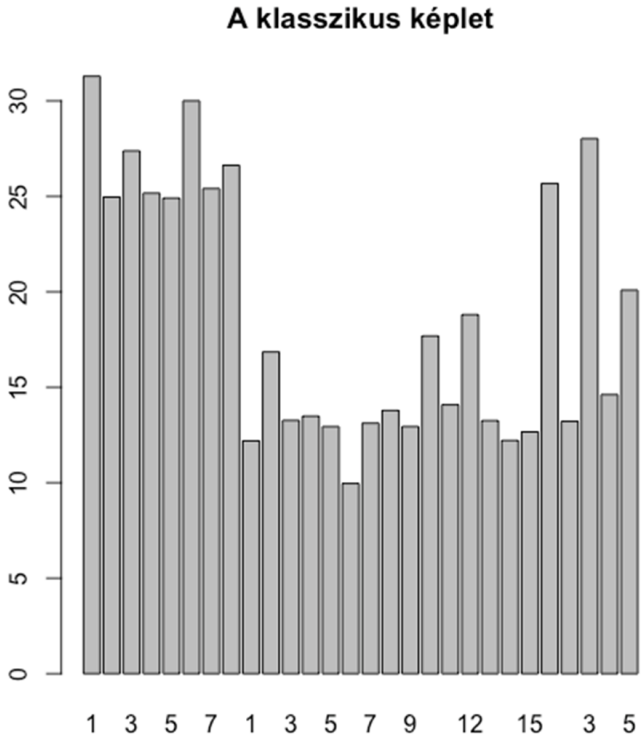
1860	1	2	3	4	5	6	7	8	9
lemma	321	564	400	352	340	179	361	406	300
szóalak	693	1120	909	681	690	323	757	867	537

1860	10	11	12	13	14	15
lemma	720	450	821	334	319	378
szóalak	1657	1021	1906	635	682	891

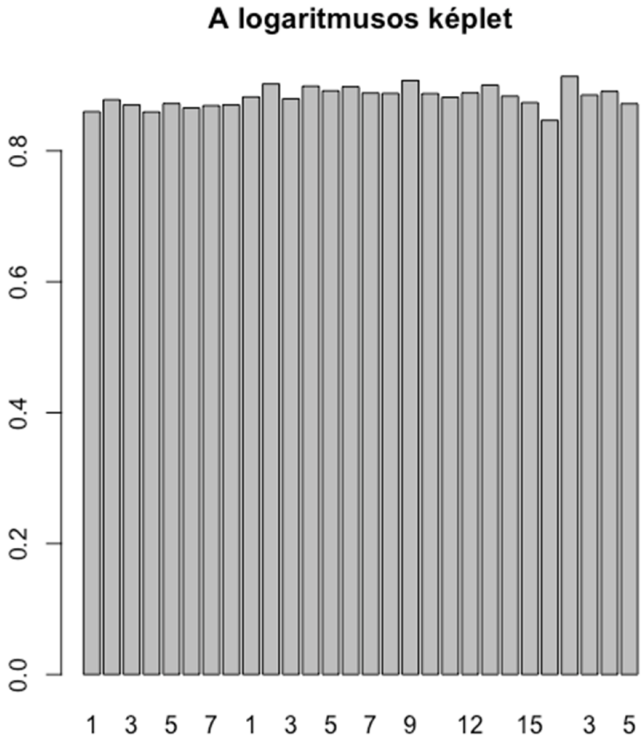
1894	1	2	3	4	5
lemma	2767	299	2116	452	1133
szóalak	11620	512	5702	956	3181

1. táblázat. *Szóalakok és egyedi lemmák száma a három Jókai-kötet novelláira lebontva*

[20] Gustav Herdan, *The Advanced Theory of Language as Choice and Chance* (Berlin: Springer-Verlag, 1966). http://doi.org/10.1007/978-3-642-88388-0.



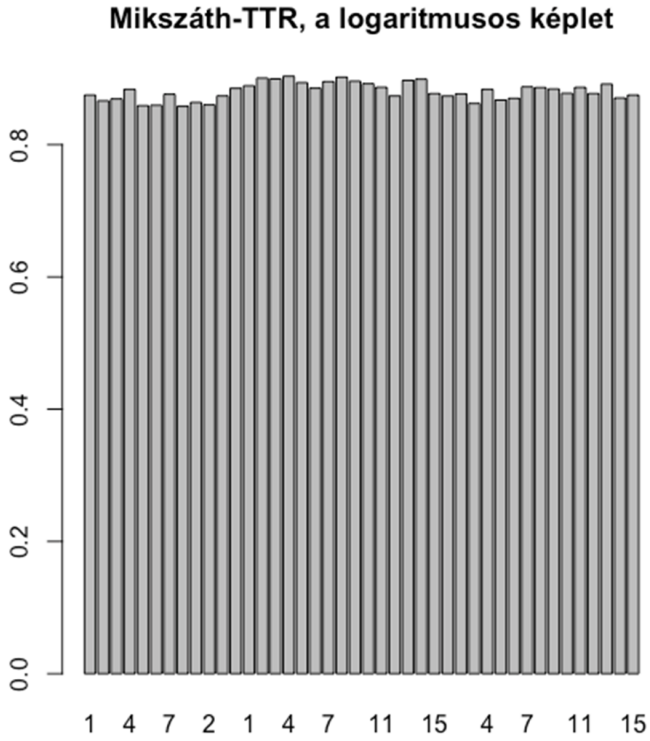
4. ábra. *Szókincsgazdagság Jókai novelláiban a Guiraud-féle képlet szerint*



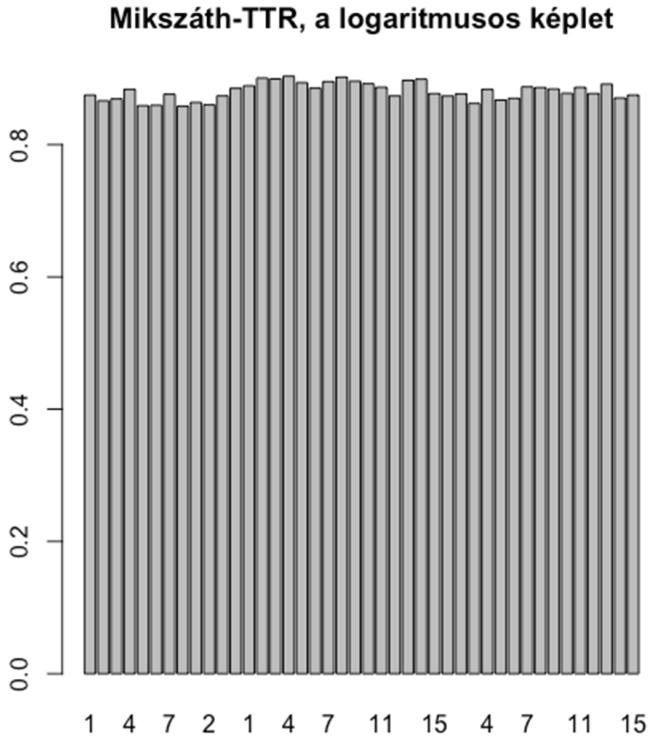
5. ábra. *Szókincsgazdagság Jókai novelláiban a Herdan-féle képlet szerint*

 DOI 10.31400/dh-hun.2019.2.390

Mikszáth esetében nem mutatom be ilyen részletesen az adatokat – a függelékben közölt *R-parancsok* alapján bárki ellenőrizheti –, de a grafikonok hasonló eredményeket mutatnak.



6. ábra. *Szókincsgazdagság Mikszáth novelláiban a Guiraud-féle képlet szerint*



7. ábra. *Szókincsgazdagság Jókai novelláiban a Herdan-féle képlet szerint*

Izgalmasabb azonban kötetekre lebontva vizsgálni az adatokat. Így – a távlati nézőpontnak köszönhetően – túlságosan homogén eredményeket kapunk. Jókai novellás-kötetre lebontott TTR-átlagai a kétféle képlet szerint, majd Mikszáthéi:

Jókai	1856	1860	1894
klasszikus	27	14	20
logaritmusos	0,87	0,89	0,88

Mikszáth	1874	1881	1882	1901
klasszikus	25	25	17	20
logaritmusos	0,87	0,87	0,89	0,88

2. táblázat. *TTR-átlagok a kétféle képlet szerint*

A logaritmusos adatok elég homogénnek mutatkoznak – már csak a kerekítések miatt is –, de ha novellákra lebontjuk a számokat, akkor azért látszanak a különbségek. Jókai szövegeinek értékei a 0,85 és 0,91 között váltakoznak, Mikszáthéi pedig a 0,86 és 0,90 között, azaz Jókaiéi valamivel változatosabbak. Ugyanakkor, ha a kiinduló kérdésre keressük a választ, azaz arra, hogy a novellák „választékossága" kifejezi-e az élőbeszédszerűséget, akkor a válasz inkább a *nem*. A mért értékek szempontjából, bár van némi különbség, ez nem tűnik jelentősnek.

4. Mérhető sajátságok

Érsok korábban idézett összefoglalása alapján azonban más nyelvi sajátosságok esetleg jobban kifejezik az élőbeszédszerűséget. Az általa említett kritériumok közül az egyik legegyszerűbben mérhető a mondathosszúság kérdése. Az alábbi táblázat Jókai novelláinak mondatszámát, illetve egy-egy novella mondatainak átlagos szószámát mutatja, a következő pedig a Mikszáth szövegeivel kapcsolatos adatokat mutatja, kötetenként átlagolva és a szélsőértékeket is feltüntetve.

mondat-/szószám	1856	1860	1894
átlag	523/17	65/13,9	341/13,8
legalacsonyabb	281/11,8	27/8,6	31/10,5
legmagasabb	930/20,1	119/20,2	998/16,5

3. táblázat. *Mondatok és szavak száma a Jókai-novellákban*

mondat-/szószám	1874	1881	1882	1901
átlag	618/11,8	506/12,9	124/10,8	273/11,6
legalacsonyabb	196/10	199/10,8	78/7,6	67/9,9
legmagasabb	989/14,6	729/15,3	196/13,6	507/13,1

4. táblázat. *Mondatok és szavak száma a Mikszáth-novellákban*

Ezek az értékek már jóval informatívabbak. Látszik, hogy Mikszáth, ha nem is feltétlenül írt rövidebb, kevesebb mondatból álló novellákat, mondatainak átlagos szóhosszúsága (értelemszerűen a tényleges szóalakokkal számoltam) már eleve jóval kisebb

 DOI 10.31400/dh-hun.2019.2.390

volt Jókaiénál. Hiába mutatnak Jókai szövegei is (illetve a belőlük készült válogatás) csökkenő tendenciát, Mikszáth maximumértékei sehol nem érik el Jókaiéit, ahogy minimumértékei is alacsonyabbak. Azaz Mikszáth szövegeire inkább igaz, hogy az élőbeszéd rövidebb, egyszerűbb mondatai dominálnak bennük.

Érsok szerint a mondathosszúság mellett a szófajok arányai is sajátosan alakulnak, a módosító- és kötőszavak gyakoribbak, ahogyan a névmások, határozószók és igék is. Szemben például a melléknevekkel. A *Magyarlanc* ún. *depparse* elemzési módja ezeket is jól kezeli. Mint fentebb a bodoki példán láttuk, az elemzés régi szövegek esetén nem mindig pontos, de mivel a nyelvi állapot miatt ugyanakkora hátránnyal indul mindkét szerző, így az eredmények értelmezhetők.

A következőkben két nagy táblázatot fogunk látni, amelyek a két szerző novellásköteteire lebontott szófaji arányait tartalmazzák. A szokásos módon először az átlag, majd a szélsőértékek is szerepelnek. A függelékben közölt *script*ek alapján természetesen részletesebb elemzések is készíthetők.

	1856	szélsőérték	1860	szélsőérték	1894	szélsőérték
ige	16,6	15,6–17,9	17,7	12,1–23,7	15,5	11,9–18,7
főnév	27,4	24,3–29,6	24,8	18,7–36,2	25,5	22,5–29,2
melléknév	10,4	8,3–11,9	8,2	5,3–11,8	10,4	7,8–13,6
kötőszó	7,4	5,9–8,7	8,5	6,5–10,6	8,1	7,4–9,1
névmás	10,6	9,1–13	12,2	9,5–18,5	10,6	8,7–13
névelő	10,1	8,4–11,4	9,5	7,3–12,7	12,3	10,8–13,2
határozó	12,7	11,4–14,7	13,7	6,8–19,3	13,2	10,2–14,9
számnév	1,1	0,8–1,5	1,5	0,6–3,7	1,6	0,8–2,9
névutó	1,8	1,4–2,3	1,5	0,6–3,4	1,3	0,5–2,1
szervetlen	0,2	0,07–0,4	0,7	0–1,9	0,3	0–0,6
igekötő	0,5	0,4–0,7	0,6	0–1,2	0,6	0,3–1

5. táblázat. *Kötetekre lebontott szófaji arányok Jókai novelláiban*

	1874	szélsőérték	1881	szélsőérték
ige	16,6	15,6–17,3	16,2	15,4–17,4
főnév	25,3	24,2–27	25,8	25,3–26,2
melléknév	10,9	9,7–12,6	10,6	9,6–11,3
kötőszó	8,3	7,4–8,9	8,7	7,7–9,1
névmás	9,5	8,5–10,5	9,2	8,6–10,1
névelő	10,7	10,2–11,2	10,4	8,8–11,4
határozó	14,4	12,9–15,4	13,9	12,9–15,3
számnév	1,2	0,8–1,7	1,3	0,9–2,2
névutó	1,2	0,9–1,4	1,2	0,9–1,6
szervetlen	0,7	0,3–0,9	0,6	0,3–1
igekötő	0,6	0,4–0,9	0,7	0,6–0,8

	1882	szélsőérték	1901	szélsőérték
ige	17,4	15,3–20,1	17,5	15,4–19,7
főnév	24,9	22,2–28,4	24,3	21,3–26,5
melléknév	9,7	7,7–12,2	9,5	7,1–10,8
kötőszó	8,1	5,3–9,8	9,2	7,8–11,4
névmás	7,5	6,2–9,8	8,3	6,4–10
névelő	12,5	10,9–14,3	13	10,7–16,1
határozó	15,4	12,9–18,1	13,7	11,6–15,7
számnév	0,9	0,5–1,8	1,1	0,3–2,1
névutó	1	0,5–1,4	1,1	0,7–1,6
szervetlen	0,9	0,4–1,3	0,8	0,2–1,2
igekötő	0,7	0,3–1,1	0,5	0,3–0,8

6. táblázat. *Kötetekre lebontott szófaji arányok Mikszáth novelláiban*

5. Következtetések

A tanulmány kiindulópontja az volt, hogy az élőbeszédszerűségnek vannak számszerűsíthető nyelvi jellemzői, a szakirodalom elég egyértelmű és mérhető különbséget tételez a lejegyzett „irodalmi” és „beszélt” nyelv között.

A felkínált sajátosságok közül, mint láttuk, a választékosság kapcsán nem sikerült olyan jellemzőt találni, amely Jókai és Mikszáth szövegei között döntő különbséget mutattak volna. A mondatok szóhosszúsága kapcsán ugyanakkor világos különbség igazolódott, még ha ennek értelmezése ennyi adat fényében nem is teljesen egyértelmű. A további mért sajátságok esetén megfigyelhettük, hogy Mikszáthnál az igék aránya lassan, de biztosan nő, amit nemcsak a kötetek átlagai, de a szélsőértékek fokozatos növekedése is mutat. Ráadásul ezzel párhuzamosan a főnevek aránya csökken. Jókai adataival összehasonlítva a különbség még feltűnőbb. Az igék aránya, ha csak kevéssel is, de Mikszáthéi alatt maradnak – ez leginkább a szélsőértékeknél mutatkozik meg –, ráadásul a főnevek, ha valamelyest csökkenő tendenciát mutatnak is, magasabbak Mikszáthéinál. Ugyanígy a kötőszavak aránya Mikszáth köteteiben magasabb, ahogyan a szervetlen közbevetéseké is. A névmások ugyan nem, de a határozószók aránya ismét igazolja a felvetést. Ha tehát az Érsok által felsorolt szófaji sajátosságokat megszámoljuk, akkor Mikszáth szövegei kétségtelenül közelebb állnak az élőbeszédhez. Ebből a szempontból éppenséggel az 1874-es kötet sikertelensége elgondolkodtató, hiszen értékei *A tót atyafiak*hoz képest nem térnek el, sőt az élőbeszédszerűség szempontjából sok esetben még „jobbak” is – persze nem állítom, hogy a siker pusztán a szövegek nyelvi sajátosságaiban rejlene.

Ugyanakkor Jókai utólag összeállított, 1894-es kötete is érdekes, mivel a szerkesztők a Jókai saját kötetei esetében már vagy három évtizeddel korábban meginduló folyamatot, – az élőbeszédhez közelítést (növekvő igearány mellett csökkenő főnévarány) – „visszafordítják”. Azaz egy Jókai-válogatást nem tudnak/akarnak – vagy éppenséggel nem lehet – úgy elvégezni, hogy a korabeli tendenciákhoz igazítsák a novellákat.

Néhány rövid történet vizsgálata önmagában persze nem elég. A későbbiekben érdemes volna egy szerzőkre, műfajokra lebontott átfogó vizsgálatot is elvégezni,

 DOI 10.31400/dh-hun.2019.2.390

mivel részletesebb, irodalomtörténetileg releváns következtetések levonásához erre volna szükség – akár a novellákat közlő lapok szintjéig eljutva, mivel a közeg sok esetben befolyásolta magát a szöveg átdolgozását is.[21]

Machine-readable Literature: "Spoken Language" in Mikszáth's Short Stories

Literary scholars have deployed the concept of "spoken language" to describe Kálmán Mikszáth's fiction since the success of his short story collections entitled *A tót atyafiak* (*Slovak Kinsmen,* 1881) and *A jó palócok* (*The Good Palots,* 1882). Although this stylistic concept has become a key characteristic feature of Mikszáth's *oeuvre,* no attempt has been made to elaborate on its definition. As scholarship assumes a clear-cut and measurable distinction between the written "literary" and "spoken" language, this paper claims that this spoken language has quantifiable linguistic markers. This is demonstrated by the morphological analysis of Kálmán Mikszáth and Mór Jókai's fictional writings.

Keywords:
fiction, Kálmán Mikszáth, spoken language, morphological analysis, lexical richness

Függelék

A függelékben a tanulmányban használt *R-script*eket adom meg, hogy bárki újrafuttathassa és elemezhesse az eredményeket, vagy saját céljaira használhassa. Én az *RStudio* nevű programot használtam. A *script*ek előtt álló számok nem a kód részei, hanem a könnyebb követhetőséget szolgálják.[22]

Fájlok behívása

Értelemszerűen a Jókai- és Mikszáth-novellákat külön dolgoztam fel, de a képletek ugyanazok, ezért a nevezéktanban nem teszek különbséget.

```
1. filenames <- list.files(path="~eleresiut", pattern="*.txt")
2. filelist <- lapply(filenames, function(x){read.csv2(x, header
   = FALSE, sep = "\t", stringsAsFactors = FALSE)})
```

[21] Török Erzsébet Zsuzsanna, „A konyhaszolgálótól Szűz Máriáig (Az irodalom hétköz- és ünnepnapi közegei a 19. század végén)" in *A Látható könyv*, szerk. Hász-Fehér Katalin (Szeged: Tiszatáj, 2006), 179–226.

[22] A kódsor a tanulmány mellékleteként TXT formátumban letölthető a cikk weboldaláról. `https://doi.org/10.31400/dh-hun.2019.2.390`

Mondatszám

Mivel a *depparse*-szal végzett elemzés esetében minden mondat minden elemet (szavak, írásjelek) megszámoz, ezért egész egyszerűen össze kell számolni az 1-eseket.

```
3. mondatszam.i <- sapply(filelist, function(df){dim(subset(df,
   V1==1))[1]})
```

Egyedi lemmák száma

Depparse esetén a harmadik oszlop a lemmaoszlop. Írásjeltelenítjük, listátlanítjuk. Kivesszük a ”c”-t, ami a listaformátum miatt kerül bele mindenhová az első helyre. Kiszámoltatjuk, mennyi lemma van az egyes szövegekben. Az integeres verzióval egyszerűbb számolni.

```
4. lemmak <- lapply(filelist, function(df){df["V3"]})
5. csaklemmak <- lapply(lemmak,
   function(x){strsplit(as.character(x), "(\\W+)")})
6. csaklemmak <- lapply(csaklemmak, function(x){unlist(x)})
7. csaklemmak <- lapply(1:length(csaklemmak),
   function(x)csaklemmak[[x]][csaklemmak[[x]]!="c"])
8a. egyedi.szoszam <- sapply(1:length(csaklemmak), function(x)
   unique(unlist(csaklemmak[[x]])))
8b. egyedi.szoszam.i <- sapply(1:length(egyedi.szoszam),
   function(x) length(egyedi.szoszam[[x]]))
```

Betűszám

Ezt értelemszerűen a tényleges szavakból, szóalakokból kell számolni.

```
9. szavak = lapply(filelist, function(df){df["V2"]})
10. szoalakok <- lapply(szavak,
   function(x){strsplit(as.character(x), "(\\W+)")})
11. szoalakok <- lapply(szoalakok, function(x){unlist(x)})
12. szoalakok <- lapply(1:length(szoalakok),
   function(x)szoalakok[[x]][szoalakok[[x]]!= "c"])
13. szoalakok.i <- sapply(1:length(szoalakok),
   function(x)length(szoalakok[[x]]))
14. betuszam.i <- sapply(1:length(szoalakok),
   function(x)nchar(szoalakok[[x]]))
```

Átlagos szóhossz

Hány betűből áll egy szó átlagban.

```
15. szohossz <- sapply(1:length(csakszavak), function(x)
   mean(nchar(szoalakok[[x]])))
```

 DOI 10.31400/dh-hun.2019.2.390

Átlagos mondathossz

Hány szóból áll, hány betűből áll egy mondat.

```
16. mondathossz <- egyedi.szoszam.i/mondatszam.i
17. mondathossz2 <- mondathossz*szohossz
```

Igék száma

```
18. igeszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="VERB"))[1]})
```

Főnevek száma

A tulajdonneveket is ideszámoltam.

```
19. fonevszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="NOUN"))[1]})
20. propnszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="PROPN"))[1]})
21. fonevszam.i <- fonevszam.i + propnszam.i
```

Melléknevek száma

```
22. melleknevszam.i <- sapply(filelist,
    function(df){dim(subset(df, V4=="ADJ"))[1]})
```

Kötőszószám

```
23. kotoszoszam1.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="CONJ"))[1]})
24. kotoszoszam2.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="SCONJ"))[1]})
25. kotoszoszam.i <- kotoszoszam1.i + kotoszoszam2.i
```

Névmások

```
26. nevmasszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="PRON"))[1]})
```

Névelők

```
27. neveloszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="DET"))[1]})
```

Határozószók

```
28. hatarozoszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="ADV"))[1]})
```

Számnév

```
29. szamnevszam.i <- sapply(filelist, function(df){dim(subset(df,
    V4=="NUM"))[1]})
```

Névutók

```
30. nevutoszam <- sapply(filelist, function(df){dim(subset(df,
    V4=="ADP"))[1]})
```

Szervetlen közbevetések

```
31. szervetlenszavak <- sapply(filelist,
    function(df){dim(subset(df, V4=="INTJ"))[1]})
```

Igekötők

```
32. igekotok <- sapply(filelist, function(df){dim(subset(df,
    V4=="PART"))[1]})
```

Szófajok

Egy közös táblázatban összegezzük az eddigi eredményeket. Amint a tanulmányban jeleztem, a szófajok esetében tízes nagyságrendben nem tud megbirkózni a *Magyarlanc* egyes szóalakkal, ami a 100000 fölötti szóalakszámot tekintve elhanyagolható különbség.

```
33. szofajok <- data.frame(igeszam.i, fonevszam.i,
    melleknevszam.i, kotoszoszam.i, nevmasszam.i, neveloszam.i,
    hatarozoszam.i, szamnevszam.i, nevutoszam, szervetlenszavak,
    igekotok)
```

Szófaji arányok

```
34. igearany <- (igeszam.i * 100)/szoalakok.i
35. fonevarany <- (fonevszam.i * 100)/szoalakok.i
36. melleknevarany <- (melleknevszam.i * 100)/szoalakok.i
37. nevmasarany <- (nevmasszam.i * 100)/szoalakok.i
38. neveloarany <- (neveloszam.i * 100)/szoalakok.i
39. hatarozoarany <- (hatarozoszam.i * 100)/szoalakok.i
40. szamnevarany <- (szamnevszam.i * 100)/szoalakok.i
41. nevutoarany <- (nevutoszam * 100)/szoalakok.i
42. szervetlenarany <- (szervetlenszavak * 100)/szoalakok.i
43. igekotoarany <- (igekotok * 100)/szoalakok.i
44. szofajarany <- data.frame(igearany,fonevarany,melleknevarany,
    kotoszoarany, nevmasarany, neveloarany, hatarozoarany,
    szamnevarany, nevutoarany, szervetlenarany, igekotoarany)
```

 DOI 10.31400/dh-hun.2019.2.390

Választékosság

Két képletet használ a dolgozat. Hány egyedi lemma és hány szóalak alkotja a novellákat.

```
45. ttr <- egyedi.szoszam.i/sqrt(szoalakok.i)
46. ttr2 <- log(egyedi.szoszam.i)/log(szoalakok.i)
```

Szófaji arányok átlaga, szélsőértéke

A szófaji arányokat már kiszámoltuk és egy táblázatban elmentettük (szofajaranyok). Ha az egyes novellák értékeire kíváncsiak vagyunk, akkor úgy kell lekérdezni. Mivel az oszlopok az egyes szófajokat tartalmazzák, a sorok pedig az egyes novellák, csak tudni kell, melyik kötetben hány novella van. Jókainál 8, 15, 5, Mikszáthnál 8, 4, 15, 15. A szögletes zárójelben először a sorokat adjuk meg, aztán az oszlopot. Átírva értelemszerűen folyamatosan megy a sorok számozása. Tehát a 47. kód Jókai 1856-ös kötetének novelláira, azok igearányára kérdez rá. Először az átlagra, aztán a szélsőértékekre. Ezek olvashatók a dolgozatban. A 49–50. az 1860-as kötet főnévarányaira kérdez, az 51–52. pedig az 1894-es kötet mellékneveire. A Mikszáth-kötetek esetében hasonló logika alapján kell a sorokat „kiosztani”.

```
47. mean(szofajarany[1:8,1])
48. range(szofajarany[1:8,1])
49. mean(szofajarany[9:23,2])
50. range(szofajarany[9:23,2])
51. mean(szofajarany[24:28,3])
52. range(szofajarany[24:28,3])
```