

Magyar Tudomány

SZÁMÍTÓGÉP
AZ IRODALOMTUDOMÁNYBAN
vendégszerkesztő: Kiss Margit

TOURS-I SZENT MÁRTON

Digitális munkásság a 2020-as években
Miért nincs királyi út?

2016•II

A MAGYAR TUDOMÁNYOS AKADÉMIA FOLYÓIRATA. ALAPÍTÁS ÉVE: 1840
177. ÉVFOLYAM – 2016/II. SZÁM

Főszerkesztő:

CSÁNYI VILMOS

Felelős szerkesztő:

ELEK LÁSZLÓ

Olvasszerkesztő:

MAJOROS KLÁRA, SELEANU MAGDALÉNA

Lapterv, tipográfia:

MAKOVECZ BENJAMIN

Szerkesztőbizottság:

BENCZE GYULA, BOZÓ LÁSZLÓ, CSÁSZÁR ÁKOS, HAMZA GÁBOR,
LUDASSY MÁRIA, SOLYMOSI FRIGYES, SPÄT ANDRÁS, VAMOS TIBOR

A lapot készítették:

GIMES JÚLIA, HALMOS TAMÁS, HOLLÓ VIRÁG, MATSKÁSI ISTVÁN, PERECZ LÁSZLÓ,
SIPOS JÚLIA, SZABADOS LÁSZLÓ, F. TÓTH TIBOR, ZIMMERMANN JUDIT

Szerkesztőség:

1051 Budapest, Nádor utca 7. • Telefon/fax: (+36-1)3179-524, telefon: (+36-1)4116-253
matud@helka.iif.hu • www.matud.iif.hu

Előfizetésben terjeszti a Magyar Posta Zrt. Postacím: 1900 Budapest.

Előfizetésben megrendelhető az ország bármely postáján, a hírlapot kézbesítőknél,

www.posta.hu WEBSHOP-ban (<https://eshop.posta.hu/storefront/>),

e-mailen a hirlapelofizetes@posta.hu címen, telefonon 06-1-767-8262 számon,

levélben a MP Zrt. 1900 Budapest címen.

Külföldre és külföldön előfizethető a Magyar Posta Zrt.-nél: www.posta.hu WEBSHOP-ban

(<https://eshop.posta.hu/storefront/>), 1900 Budapest, 06-1-767-8262, hirlapelofizetes@posta.hu

Belföldi előfizetési díj egy évre: 11 040 Ft.

Nyomdai munkák: Inferno Reklám Kft.

Felelős vezető: Farkas Dóra

Megjelenik: 11,4 (A/5) ív terjedelemben

HU ISSN 0025 0325

TARTALOM

Számitógép az irodalomtudományban

Vendég szerkesztő: Kiss Margit

Kiss Margit – Mészáros Tamás: Bevezető.....	1282
Almási Zsolt: Szöveg az olvasónak: kvantitatív módszerek és digitális Shakespeare-filológia	1286
Bodrogi Ferenc Máté: Elektronikus kritikai kiadások a klasszikus magyar irodalomban ...	1292
Dömötör Adrienne – Novák Attila: Főurak levelei, boszorkányok pereit és egy újonnan épített adatbázis: A Történeti magánéleti korpusz	1297
Kiss Margit: Egy XVIII. századi írói korpusz modern feldolgozása	1305
Mészáros Tamás: Mit nyújthat a modern informatika az irodalomtudomány számára? ...	1310
Palkó Gábor: Mit jelent a digitális filológia a szemantikus web korában? A Digiphil projektről	1316
Péter Róbert: A Big Data kihívás és lehetőség a bölcsészettudományokban: digitális szövegek és metaadatok távoli olvasása	1323
Seláf Levente: Digitális versrepertóriumok fejlesztése és összekapcsolása: kutatástörténet és kilátások	1331

Tours-i Szent Márton (316–397)

Pálffy Géza: A pozsonyi Szent Márton-koronázótemplom: a Magyar Királyság egyik legfontosabb temetkezőhelye	1337
Tóth Ferenc: Sabaria patria Sancti Martini, Utazók, katonák, diplomaták a szombathelyi Szent Márton-kultuszról.....	1349
Vánca István: Lúdhúst enni	1359

Tanulmány

Tamás Pál: Digitális munkásság a 2020-as években	1365
Dobos Imre – Michalkó Gábor – Nováky Erzsébet: Miért nincs királyi út a habilitáció publikációs követelményeinek meghatározására?	1379

Vélemény, vita

Hadas Miklós: Az MTA és a férfialalom	1391
---	------

Az MTA új levelező tagjainak bemutatása

Dunai László	1395
Frank András	1397

Kitekintés (Gimes Júlia).....

Könyvszemle (Sipos Júlia)

Két partiumi monográfia (Balázs Géza)	1403
Hűlt nyomokon (Ablonczy Balázs)	1406

IRODALOM

- Bíbor Máté et al. (2005): *A magyar irodalom filológiája*. Gépeskönyv • <http://www.tankonyvtar.hu/en/tartalom/ta/magyar-irodalom/index.html>
- Bíró Szabolcs (2005): *Szövegfeldolgozás XML alapokon*. Neumann Kht., Budapest • <http://www.tankonyvtar.hu/informatika/szovegfeldolgozas-xml-080906-159>
- Gabler, Hans Walter (1989): A kiadói szöveg születése: számítógép bába-szerepben. (ford. Farkas Ildikó) *Helikon*. 3–4, 421–428.
- Huitfeldt, Claus (1994/95): Multi-Dimensional Texts in a One-Dimensional Medium. *Computers and the Humanities*. 28, 235–41. DOI: 10.1007/BF01830270
- Kalsó Gyula (2011): *A TEI-XML felhasználása magyar nyelvű korpuszok építésében*. In: Boda István Károly – Mónos Katalin (szerk.): *Az alkalmazott nyelvészet ma: innováció, technológia, tradíció. XX. Magyar Alkalmazott Nyelvészeti Kongresszus*. Debrecen, Mány–Debreceni Egyetem, Budapest. 65–71. • http://www.inf.unideb.hu/~boda/pub/MANYEXX_elso_resz-B5.pdf
- Kelemen Pál – Kulcsár Szabó E. – Tamás Á. – Váderna G. (szerk.) (2014): *Metafilológia 2. Szerző – könyv – jelenetek*. Ráció, Budapest
- McGann, Jerome J. (1989): *Az Ulysses mint posztmodern szöveg: a Gabler-féle kiadás*. (ford. Friedrich Judit) *Helikon*. 3–4, 429–452.
- Oliver, Andrew (1989): Mikroinformatika és textológia. (ford. Farkas Ildikó) *Helikon*. 3–4, 412–420.
- Palkó Gábor (2015): Digitális filológia: számítógép anyaszerepben. *Filológiai Közöny*. 2, 187–199. • http://www.balassikiado.hu/BB/NET/Filologia/Filopdfek/Filo_2015_2.pdf
- Pichler, Alois (ed.) (2015): *Wittgenstein Source Bergen Nachlass Edition*. Edited by the Wittgenstein Archives at the University of Bergen under the direction of Alois Pichler. In: Wittgenstein Source (2009–). WAB, Bergen • www.wittgensteinsource.org
- Renear, Allen H. (2004): Text Encoding. In: Schreibman, Susan et al. (eds.): *A Companion to Digital Humanities*. Blackwell, Oxford • <http://www.digitalhumanities.org/companion/>
- Renear, Allen, Mylonas, E. – Durand, D. (1996): Refining our Notion of What Text Really Is: The Problem of Overlapping Hierarchies. In: Ide, Nancy – Hockey, Susan (eds.): *Research in Humanities Computing*. Oxford University Press, Oxford. 263–280. • <http://cds.library.brown.edu/resources/stg/monographs/ohco.html>
- Tóth Máté (2010): Könyvtárak a szemantikus web világában. *Könyvtári Közöny*. 3, 413–438. • <http://ki.oszk.hu/ktf/2010/10/konyvtarak-a-szemantikus-web-vilagaban/>
- URL1: digiphil.hu
- URL2: data.digiphil.hu/search/



A BIG DATA KIHÍVÁS ÉS LEHETŐSÉG A BÖLCSÉSZETTUDOMÁNYOKBAN: DIGITÁLIS SZÖVEGEK ÉS METAADATOK TÁVOLI OLVASÁSA

Péter Róbert

egyetemi adjunktus,
Szegei Tudományegyetem Angol Tanszék
rpeter@lit.u-szeged.hu

Mindannyian tudjuk és tapasztaljuk: az információs társadalom korszaka jelentős kihívás a humán tudományok számára: a változás egyaránt érinti a tudás társadalmi intézményeit, legitimációját, ám ugyanúgy kihat mindennapjainkra, tudományos gyakorlatunkra. Mindez alkalmat nyújt – valójában sürget – a humántudományi kutatások tárgyának, módszerének és közegének újrarendelésére.¹

A bölcsészettudományokban zajló digitális fordulatot jelzi, hogy az elmúlt évtizedben hatalmas mennyiségű ismert és ismeretlen forrásanyag vált elérhetővé és kereshetővé szabad felhasználású vagy előfizetést igénylő digitális gyűjteményekben. Nyilvánvaló, hogy a digitális forradalom jóval gyorsabbá, könnyebbé tette a kutatást azzal, hogy több millió szöveget tartalmazó digitális archívumokban kereshetünk, korábban nem, vagy nehezen

hozzáférhető anyagok váltak elérhetővé. Az angol nyelvű írott kultúrkincsre vonatkozóan Matthew L. Jockers 2008-at, a németre vonatkozóan Fotis Jannidis és Gerhard Lauer pedig 2011-et jelöli meg az áttörés éveként (Jockers, 2013; Janidis – Lauer, 2014). Több millió szöveg azonban nemcsak könnyebbé jelent: a digitális fordulat valójában új kihívások elé állította a kutatókat. A gyakorlatban végbe ment digitális forradalmat azonban még nem igazán követte módszertani forradalom. A lehetőségeket és kihívásokat is rejtő digitális kutatás jelenlegi helyzetét tökéletesen példázza a verseny, melyet a British Library hirdet meg 2013 óta: a British Library Labs pályázatot ír ki tudósoknak, kísérletező szakembereknek és szoftverfejlesztőknek, melynek célja a neves brit intézmény digitális gyűjteményeinek felhasználására épülő innovatív és úttörő projektek megvalósítása (URL2). Olyan újszerű kutatási elképzeléseket várnak, melyek hiánypótló ismeretek, felfedezések forrásként támaszkodnak a könyvtár hatalmas digitális gyűjteményeire. A digitális tudományos kutatás valóban formabontó és úttörő kutatási lehetőségekkel kecsegtet a bölcsészettudomá-

¹ Kokas Károly, Labádi Gergely, Péter Róbert, *Digitális bölcsészet Szegeden* – konferenciafelhívás (2015. október 12.) URL1. Ezúton mondok köszönetet Labádi Gergely kollégámnak a tanulmány korábbi verziójához fűzött építő és értékes megjegyzéséért, valamint azért, hogy utalhatam *A magyar regény adatbázisa* című projekt eddigi publikálatlan eredményeire.

nyokban, mivel a szövegek és az ezekhez rendelt metaadatok elemzéséhez számtalan új lehetőséget kínálnak, melyeket e tudományterületen korábban nem alkalmaztak.

Az új digitális módszerek és eszközök használatát a hazánkban a tavaly megszüntetett, de világszerte gomba módra szaporodó digitális bölcsészet tanszékeken oktatják. A hallgatók megtanulják például, miként kell adatbázisokat létrehozni, szövegeket felcímkézni. Született már egyetemi tankönyv, amely kifejezetten a nyelv és irodalom szakosokat célozza meg a hatalmas szövegtörzsek és adatbázisok statisztikai elemzéséhez és ábrázolásához használt R programnyelv megismertetésével, amelynek van kifejezetten nyelvészeti és irodalmi elemzésekhez használható csomagja is (Jockers, 2014; Eder et al., 2013).

Felejtünk el olvasni vagy olvassunk távolról?

Nyilvánvaló, hogy az új adatbázisokban található szövegek mindegyikét képtelenség egy emberöltő alatt elolvasni és hagyományos eljárásokkal feldolgozni. 2010-ig csak a Google 15 millió könyvet digitalizált, a világon valaha megjelent összes könyv körülbelül 12%-át. Ha valaki megpróbálná csak a 2000 után megjelent angol nyelvű könyvek felolvasását – 200 szó per perc tempóval számolva –, akkor ez nyolcvan évig tartana egyhuzamban (Michel et al., 2011). A *big data* / *big text* feldolgozásához új módszerek szükségesek. Többek között ez indította Franco Moretti irodalomtörténész a *távolsági olvasás* (*distant reading*) kifejezés bevezetésére a szakirodalomban.² Részben az *Annales* iskola hagyományaira építve, Moretti a kvantitatív és statisztikai módszerek használatát szorgalmazza egy racionálisabb és

átfogóbb irodalomtörténet megalkotása érdekében. A klasszikus irodalomtudomány módszertana ugyanis az ún. *close reading*-re, a szoros olvasás metodikájára épült, azaz alapvetően néhány szerző, néhány szövegének a vizsgálatára: a kánonformálás ideológiai okai és az emberi feldolgozhatóság korlátai miatt az elemzés minimális mennyiségű kanonizált szöveget vesz csak figyelembe. A távoli olvasás célja nagy mennyiségű szövegek közötti kapcsolatok, párhuzamok, ismétlődő minták, ciklusok feltárása és elemzése, melyeket a limitált szoros olvasás nem képes feltárni (Moretti, 2005; Labádi, 2014). Bár a modellezés során elveszítjük magát a szöveget, az absztrakciós eljárás új típusú ismereteket, összefüggéseket, folyamatokat és struktúrákat világít meg, melyeket az irodalomtörténész grafikonok, térképek, ágrajzok és törzsfejlődési fák segítségével illusztrál. Számos kollégájával együttben Moretti komolyan veszi Sherlock Holmes figyelmeztetését, miszerint kapitális hiba adatok nélkül elméleteket alkotni. Tanulmányai korszakokon, nemzeteken és kultúrákon átívelő tendenciákat tárnak fel. Például a szakirodalomban elklónított negyvenégy különböző regénytípus 1740 és 1900 közötti nagy-britanniai elterjedésével kapcsolatban megállapítja, hogy az egyes regényfajták általában huszonöt-harminc évig maradnak népszerűek, és az egyes típusok esetében körülbelül ugyanannyi jelenik meg évente. Új regénytípus általában akkor születik, ha a korábbi már veszt a népszerűségéből. A brit, japán, nigériai, spanyol és olasz regények 1750 és 1900 közötti elterjedésének grafikonon történő megjelenítése pedig annak a törvényszerűségnek a megfogalmazásához vezetett, hogy a regények meghonosodása mindenütt két fázisban valósult meg, az első fellendülést mindig egy rövidebb visszaesési fázis követte

² Stephen Ramsay *algorithmic criticism*, Matthew Jockers *macroanalysis* fogalmai és módszerei számos hasonló-ságot mutatnak a Moretti-féle *distant reading*-gel.

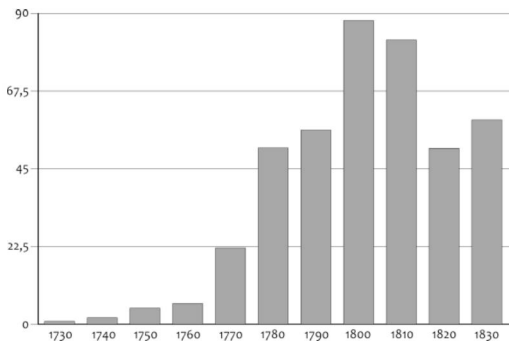
(Moretti, 2005). A Labádi Gergely és Bátori Anna által készített magyar regényadatbázis (1730–1840) bibliográfiai adatainak modellezése is a Moretti-hipotézist látszik igazolni (Labádi – Bátori, 2015).

A Moretti és más digitális bölcsészek által az adatok ábrázolására használt technikák és eljárások az egyes természettudományi diszciplínák esetében már a 19. században megjelentek. Gondoljunk például a Darwin-féle evolúciós életfára vagy a Florence Nightingale által használt rózsadiagramra, amely demonstrálta, hogy a krími háborúban (1853–1856) jóval több katona halt meg megelőzhető betegségek következtében, mint a csatáren szerzett sérülések vagy egyéb okok miatt. Bár a 19. században elterjedt – főként leíró jellegű – ábrázolási technikákat még mindig használjuk, a mai számítógépes-matematikai modellezés és időalapú vizualizáció segítségével jóval szofisztikáltabb módon tudunk hatalmas mennyiségű adatok vagy információs folyamatok közötti korrelációkat, asszociáció-

kat és trendeket megjeleníteni és azonosítani. Erre jó példa a Google *N-Gram Viewer* alkalmazása.

A culturomics mint új kutatási módszer

A közelmúltban elérhetővé vált óriási volumenű és kulcsszavakkal kereshető digitális archívumanyag az emberiség történetében először teszi bárki számára lehetővé több millió könyv áttekintését és szisztematikusan elemzését (*1. ábra*). A Google Labs a Harvard Universityvel karöltve 2010 decemberében útnak indította az *N-Gram Viewer*-t, melynek segítségével lehetővé vált a szóhasználati gyakoriság időbeli, grafikus megjelenítése, pontosabban legfeljebb 5 gramig, 5,2 millió digitalizált könyv (az 1600 és 2000 között kiadott összes könyv körülbelül 4%-a) elemzésével. Az *n-gram* *n* darab egymást követő szóra utal: a „Magyar Királyság” például egy bigram, ha erre vagyunk kíváncsiak, akkor a kereső olyan találatokat ad vissza, ahol e két szó ebben a sorrendben áll egymás mellett. A fejlesztésben



1. ábra • A Magyarországon megjelent regénycímek száma tízéves bontásban, az utánnyomásokkal együtt

részt vevő kutatók analitikai módszerüket *culturomics*-nek nevezték el, és az emberi kultúra tanulmányozására alkalmazott nagy teljesítményű adatgyűjtésként és elemzésként határozták meg (Michel, 2011). Az *N-Gram Viewer*-t többek között nyelvészek, közgazdászok, pszichológusok és (sport)történészek használják kutatási eszközként (Roivainen, 2013; Roth, 2014; Philips et al., 2015).

Ugyanakkor ennek az új, hatékony, bár még kezdeti stádiumban lévő kutatási eszköznek néhány gyenge pontja is van. Először is a Google nem mindig tudományos szempontok és megfontolások alapján válogatja ki a digitalizálandó könyveket, ami szükségszerűen tükröződik a hatalmasra duzzadt *N-Gram* korpuszon is. Másrészt az elemzést gyakran torzíja a korai korszak könyvek gyenge szövegfelismerése. Az *N-Gram Viewer* létrehozói sem tagadják ezeket a hibákat, sőt hangsúlyozzák, hogy az *N-Gram Viewer* leginkább az 1800 után íródott könyvek vizsgálatára alkalmas. Harmadsorban, a Google által digitalizált könyvek metaadatai nemegyszer pontatlanok, ami szintén torzíja az elemzést. Az *N-Gram Viewer* készítői például azt állítják, hogy a folyóiratokat nem foglalták bele az adatbázisba. Az általuk létrehozott korpusz azonban valójában több magazint is tartalmaz. Ráadásul az egymás mellett álló szavak keresésénél finomabb számítógépes szövegfeltárási technikák is léteznek: a kulcsszavak és kontextusaik együttes vizsgálata (KWIC), főkomponensanalízis (PCA), kulcsszavak és kollokációk, valamint a tartalmi elemzés (*topic modelling*) (Jockers, 2013).

Tim Schwartz további kritikával illette az *N-Gram Viewer*-t, mely igen lényeges annak a digitális kutatási módszernek az utólagos igazolásához, amelyet e tanulmány következő részében mutatok be. Schwartz szerint a

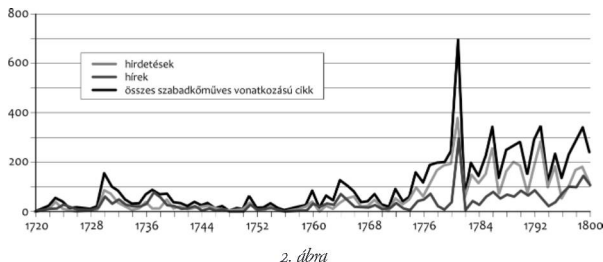
„könyvek természetükből adódóan távolabb vannak a kultúra liktetésétől, mint a periodikák, különösen az újságok. A könyvkiadás során a nyomtatás jelentős időeltolódást eredményez. A sajtótermékek közelebb vannak a valós időhöz” (Schwartz, 2011, 36.).

Újságok távoli olvasása

2010-ben szegedi (programozó) matematikusokkal egy olyan digitális módszer fejlesztésébe kezdtünk, amely többek között képes nagy mennyiségű sajtócikk bibliográfiai- és metaadatainak grafikus megjelenítésére, eloszlásának és gyakoriságának vizsgálatára. Az előbbi segítségével többek között – eddig ismeretlen – hosszú távú történeti, kulturális, nyelvi trendeket és folyamatokat jeleníthetünk meg, valamint tesztelhetünk régi hipotéziseket (Péter, 2011; Péter, 2015).

Egy 2009-ben indult kutatási projekt keretében elkészítettünk egy adatbázist, amely a brit és ír szabadkőművességgel kapcsolatos digitális, ill. levéltári gyűjteményekben elérhető újságcikkek (11 987 tétel) bibliográfiai és metaadatait tartalmazza 1709 és 1813 között. Az említett digitális módszer ezen adatbázis kvantitatív és statisztikai elemzéséből fejlődött ki a projekt során használt elemzési eljárások általánosításán keresztül. Az alap kutatás során azonosított szabadkőművességgel kapcsolatos újság- és folyóiratcikkek adatainak modellezése segítséget nyújtott egy kritikai forrásgyűjtemény egyik kötetében közölt cikkek tematikai és időbeli kiválasztásához is (Péter, 2016). Terjedelmi korlátok miatt itt csak két funkciót tudunk vázlatosan bemutatni.

A 2. ábra a szabadkőművességgel kapcsolatos cikkek évenkénti számának változását mutatja 1720 és 1800 között. Az eltérő árnyalatok a hírek, hirdetések és az egy adott évben megjelenő összes cikk függvényeit ábrázolják.



Az újságok és a bennük megjelenő cikkek számának időbeli változása miatt, ha nem abszolút, hanem relatív, százalékos módon ábrázoljuk a szabadkőművességgel kapcsolatos cikkek számát, akkor is 1781 kiemelkedő jelentőségű év. Az utóbbi két modellezési eljárás (abszolút és relatív) nem veszi figyelembe azt a tény, hogy gyakran szó szerint ugyanazt a cikket több korabeli újság is közölte (plagizálás). Emiatt a maximumhelyeket teszteltük úgy is, hogy a szabadkőműves vonatkozású cikkek megjelenéseit olyan különálló, hosszú futamidejű és minél teljesebb példányszámmal rendelkező újságok alapján vizsgáltuk, mint például a *Public Advertiser* vagy *Gazetteer and New Daily Advertiser*. Ezeknél a függvényeknél szintén 1781-nél találjuk az abszolút maximum értéket. De mi történt ekkor a brit szabadkőművesség történetében? Az 1780. december 29-én először bemutatott *Harlequin Free-mason* című pantomimjátékot a következő évben hatvanhárom alkalommal játszották a londoni Theatre Royalban. A nagysikerű színdarabbal kapcsolatban nagy számban jelentek meg hirdetések és beszámolók az angol sajtóban. Eleddig sem a színháztörténet, sem pedig a szabadkőműves történetírás nem foglalkozott e darab elemzésével és kontextua-

lizálásával, pedig az előadások és az újságcikkek nagy száma alapján minden bizonnyal jelentős hatással volt a szabadkőművességről kialakított korabeli társadalmi képre.³ A színdarab egy eddig javarészt feltérképezetlen kutatási területre irányította figyelmünket. Az erre épülő alapkutatás számos új aspektusból világította meg a 18. századi férfitársaságok, köztük a szabadkőművesség, és a brit színházi élet szoros és élelénk kapcsolatát (Péter, 2016).

Az újságszerkesztőknek és -tulajdonosoknak kulcsfontosságú szerepük volt és van abban, hogy miről és hogyan tudósít egy adott sajtóorgánus. Ebből a szempontból sem mellékes megvizsgálni azt, hogy a szabadkőművességet említő cikkek mely újságokban milyen számban és milyen gyakorisággal jelentek meg. Mindkét listán a *Morning Chronicle* áll a második helyen 1791 és 1813 között. Nem véletlenül, hiszen James Perry, aki az újság tulajdonosa és szerkesztője ebben az időszakban, az Ösiek Nagypáholyának helyettes nagymestere volt 1787 és 1790 között. Személyét a szabadkőműves történetírás mind-egyedül ignorálta. Annyira nem tartották korábban fontosnak őt, hogy a 121 történetész köz-

³ A színdarabot még huszonegyszer játszották 1789-ben és 1793-ban.

remüködésével létrejött háromkötetes, közel háromezer oldalas, a jeles 18. századi szabadtőművészek életrajzát tartalmazó *Le monde maçonnique des Lumières: Europe-Amériques & colonies: dictionnaire prosopographique* (Porset – Révaugé, 2013) című kiadványban Perry nevét még csak meg sem említik, jóllehet ő volt az egyik aláírója annak az *Articles of Union* (1813) címet viselő dokumentumnak is, amely az előtte évtizedekig rivalizáló Ősiek és Modernek Nagypáholyainak egyesítését deklarálta, létrehozván az azóta is működő Egyesült Angol Nagypáholyt.

A távoli olvasás korlátai és veszélyei

A *distant reading* módszere ugyan minden korpuszra és (meta)adatbázisra alkalmazható, ám eredményessége, a használatával szerezhető ismeretek relevanciája adatbázisonként változó. Általánosságban elmondható, hogy minél teljesebb, átfogóbb és metaadatokkal minél gazdagabban és pontosabban ellátott egy adatbázis, annál precízebb és megbízhatóbb eredményeket kapunk lekérdezéseink során, feltéve ha szakmailag releváns és az adatbázis lehetőségeit figyelembe vevő kérdést teszünk fel. Ha például kulcsszavak előfordulását szeretnénk modellezni címekben vagy cikkekben egy olyan adatbázisban, ahol a személynevek nincsenek külön megjelölve vagy felcímkézve, akkor torz eredményekhez vezet, ha a *Christian* (keresztény) szó előfordulást ábrázoljuk, mivel a keresési eredmények között nem csak a keresztény vallásra vonatkozó találatok szerepelnek, hanem azok is, amelyekben szerepel a *Christian* (Krisztián) név. Fontos kritérium az is, hogy a keresési kifejezés jelentése az idő múlásával csak minimális változáson menjen keresztül (például: helynevek, tárgyak pontos megnevezései) – vagy legalábbis az adatok értékelé-

sénél számoljunk ezzel is. Minél kevésbé kontextusfüggő egy szó jelentése, annál értékelhetőbb eredményeket kapunk. A távoli olvasás sokszor jól ismert folyamatokat és eseményeket igazol vissza: miután elkészítettük a 17. századi angol sajtóban a magyar vonatkozású cikkek adatait tartalmazó adatbázist, nem lepődtünk meg, hogy hazánkról a legtöbb cikk Buda 1686-os visszafoglalásakor jelent meg (Péter, 2015).

Összegzés és kitekintés

A humán tudományokban zajló digitális fordulat új lehetőségeket és irányokat kínál a kutatásban. A digitális adatbázisokban korszerű eljárásokkal rögzített és metaadatokkal ellátott szövegekkel kapcsolatban eddig nem alkalmazott módszerekkel új kérdéseket tehetünk fel. Ezek megválaszolása feltáratlan bizonyítékokhoz és ismeretekhez juttathat el, melyek a szoros olvasáson alapuló módszerekkel sokszor nem kimutathatók. Így régóta elfogadott téziseket tudunk tesztelni és újraértékelni, valamint új kutatási csomópontokat tudunk megrajzolni. Bár legtöbbször egyszerű kérdéseket teszünk fel az új digitalizált anyaggal és az ehhez rendelt adatokkal kapcsolatban, ezek megválaszolása bonyolult szakproblémákhoz vezet. Néhány évvel ezelőtt még e tanulmányban említett kérdések feltevése sem volt lehetséges. Az emberiség történetében most van először lehetőség ilyen típusú elemzések elvégzésére, mivel a 21. század előtt nem álltak rendelkezésünkre ilyen volumenű digitális archívumok.

A digitális fordulat nem ássa alá a hagyományos kutatási formákat és módszereket, inkább erősíti azokat. Moretti néhány provokatív kijelentésével vitázva úgy gondolom, hogy a távoli olvasás nem helyettesíti a szövegközi olvasást, melynek használatát az

említett szerző elavult „teológiai gyakorlatnak” nevezi. A szövegközeli olvasás módszertanának megszületésekor a módszer képviselői sokat profitáltak a statisztikai elemzések kínálta lehetőségekből (Igarashi, 2015). A kvantifikációt megvető kritikusokkal ellentétben úgy véljük, hogy a kvantitatív és kvalitatív módszerek a bölcsészettudományok területén is kiegészítik egymást. Nem a pozitívista szemlélet újbóli térhódítását tükrözi a statisztikai módszerek használata a digitális bölcsészettudományban. A kutató – és nem a számítógép – dönti el, milyen változókat mérünk és hogyan. A digitális és módszertani forradalom új kutatási eszközöket ad a tudósok kezébe, de ezek eszközök, és nem célok. Az új eredmények értelmezése és elemzése mindig is a szakavatott bölcsész feladata marad.

Mind a módszerek, mind pedig az eddig digitalizált szöveggállomány mennyisége és minősége tekintetében a digitális bölcsészet még gyermekcipőben jár. Olyan kifinomult kutatási eszközökre és szoftverekre van szükségünk, amelyek úgy képesek megbirkózni hatalmas mennyiségű szöveggel és metaadattal, hogy az elemzésnél a bölcsészettudományi kutatásban hangsúlyos részletek és nüanszok nem vesznek el. Az ilyen innovatív inter-, multi- és transzdiszciplináris projektek külön-

böző tudományterületek képviselői közötti együttműködést kívánnak. Kivitelezésük ideális feltétele az lenne, hogy az elemezni kívánt adatbázis teljes egészében nyílt hozzáféréstű legyen. Manapság az utóbbi részleges vagy teljes hiánya világszerte súlyos feszültségeket és problémákat okoz a bölcsészettudományi kutatások területén. A nemzeti és egyetemi könyvtárak és nem profitorientált magáncégek (például: ProQuest, Cengage Learning) feladata lenne a nemzeti kultúrkincs digitális feldolgozása és tárolása.⁴ A hatalmas digitális és pénzügyi szakadék miatt alapvető tudományos adatbázisok ráadásul csak vagyonos egyetemeken oktatók számára érhetők el, így a kevésbé tehető egyetemeken kutatók fel sem tehetik azokat a kérdéseket, mint a jó módú egyetemeken tevékenykedő kollégáik.

⁴ Óriási veszélyeket hordoz magában, hogy a digitalizált kultúrkincseket magáncégek tárolják. Gondoljunk itt a több milliárd dolláros adósságot felhalmozó Cengage Learning által 2013 júliusában bejelentett csődeljárásra. Az akkori több mint száz adatbázist birtokló céget végül az átszervezés megmentette.

Kulcsszavak: *digitális bölcsészet, statisztika, adatbázis, big data, távoli olvasás, szabadkőműveliség*

IRODALOM

- Ascarì, Maurizio (2014): The Dangers of Distant Reading: Reassessing Moretti's Approach to Literary Genres. *Genre*, 47, 1, 1–19. DOI: 10.1215/00166928-2392348
- Bátóri Anna – Labádi Gergely (2015): Egy regényadatbázis felépítése – kérdések és lehetőségek. Előadás *A számítógép az irodalomtudományban* c. workshopon (Budapest, 2015. november 24.). • <http://tinyurl.com/gvgn9n9>
- Eder, Maciej – M. Kestemont – J. Rybicki (2013): Stylometry with R: A Suite of Tools. In: *Digital Humanities 2013: Conference Abstracts*. Lincoln:

- University of Nebraska (Lincoln.), 487–489. • <http://tinyurl.com/hmgylhy>
- Igarashi, Yohei (2015): Statistical Analysis at the Birth of Close Reading. *New Literary History*, 46, 3, 485–504. DOI: 10.1353/nlh.2015.0023
- Jannidis, Fotis – Lauer, Gerhard (2014): Burrows's Delta and Its Use in German Literary History. In: Erdlin, Matt – Tatlock, Lynne (eds.): *Distant Reading: Topologies of German Culture in the Long Nineteenth Century*. Camden House, Rochester, NY, 29–54. • <http://tinyurl.com/groyymx>
- Jockers, M. L. (2013): *Macroanalysis: Digital Methods and Literary History*. Univ. of Illinois Press, Urbana

- Jockers, Matthew L. (2014): *Text Analysis with R for Students of Literature (Quantitative Methods in the Humanities and Social Sciences)*. Springer-Verlag, Cham
- Juola, Patrick (2013): Using the Google N-Gram Corpus to Measure Cultural Complexity. *Literary and Linguistic Computing*, 28, 4, 668–675. DOI: 10.1093/llc/lqt017
- Labádi Gergely (2014): Franco Moretti, Distant Reading. [könyvismeret] *Irodalomtörténet*, 95, 4, 561–564. • <http://tinyurl.com/zxyu33s>
- Labádi Gergely – Bátori Anna (2015): *A magyar regény adatbázisa. kézirát*
- Michel, Jean-Baptiste et al. (2011): Quantitative Analysis of Culture Using Millions of Digitized Books. *Science*, 331, 6014, 176–182. DOI: 10.1126/science.1199644 • <http://tinyurl.com/7yu5649>
- Moretti, Franco (2005): *Graphs, Maps, Trees: Abstract Models for a Literary History*. Verso, London
- Péter Róbert (2011): Researching (British Digital) Press Archives with New Quantitative Methods. *Hungarian Journal for English and American Studies*, 17, 2, 283–300. • <http://www.jstor.org/stable/43487818>
- Péter Róbert (2015): Digitális és módszertani fordulat a sajtókutatásban: A 17–18. századi magyar vonatkozású angol újságcikkek „távolságtartó olvasása”. *AETAS* 30.1, 5–30. • <http://www.aetas.hu/2015-01.pdf>
- Péter Róbert (ed.) (2016): *British Freemasonry 1717–1813*. I–V. Routledge, New York, (Vol. 5.)
- Phillips, Murray G. – Osmond, G. – Townsend, S. (2015): A Bird's-eye View of the Past: Digital History, Distant Reading and Sport History. *International Journal of the History of Sport*. Published online 28 October 2015. DOI: 10.1080/09523367.2015.1090976 • https://espace.library.uq.edu.au/view/UQ:373008/UQ373008_Post_print.pdf
- Porset, Charles – Révauger, Marie-Cécile (2013): *Le monde maçonnique des Lumières: Europe-Amériques & colonies: dictionnaire prosopographique*. H. Champion, Paris
- Roivainen, Eka (2014): Changes in Word Usage Frequency May Hamper Intergenerational Comparisons of Vocabulary Skills: An Ngram Analysis of Wordsum, WAIS, and WISC Test Items. *Journal of Psychoeducational Assessment*, 32, 1, 83–87. DOI: 10.1177/0734282913485542
- Roth, Steffen (2013): The Fairly Good Economy: Testing the Economization of Society Hypothesis Against a Google Ngram view of Trends in Functional Differentiation (1800–2000). *Journal of Applied Business Research*, 29, 5, 1495–1500. • http://papers.ssrn.com/sol3/papers.cfm?abstract_id=2318228
- Schwartz, Tim (2011): Culturalomics: Periodicals Gauge Culture's Pulse. *Science*, 332, 6025, 36. DOI: 10.1126/science.332.6025.35-c • <http://science.sciencemag.org/content/332/6025/35.long>
- URL1: <http://digibolcsesz.ek.szte.hu/>
- URL2: <http://labs.bl.uk/>



DIGITÁLIS VERSREPERTÓRIUMOK FEJLESZTÉSE ÉS ÖSSZEKAPCSOLÁSA: KUTATÁSTÖRTÉNET ÉS KILÁTÁSOK

Seláf Levente

PhD, egyetemi adjunktus,

Eötvös Loránd Tudományegyetem Bölcsészettudományi Kar

Magyar Irodalom- és Kultúratudományi Intézet

levente.selaf@gmail.com

A bölcsészinformatika s azon belül a digitális filológia világvizonylatban is korán, már az 1990-es években fontos szerepet játszott a magyar tudományban (hálózati szövegkiadások, adatbázisok létrehozásával). Az e tudománynak szentelt első központ az ELTE-n 1997-ben Horváth Iván vezetésével megalakított BIÖP (Bölcsészettudományi Informatika Önálló Program) műhelye volt, amely több hálózati kritikai kiadást, digitális formában elérhető tananyagot és tudományos cikket készített, valamint konferenciákat szervezett a digitális átalakulás következtében a filológiára váró kihívásokról és az ennek következtében elkerülhetetlenné váló szemlélet- és paradigmaváltásról az irodalmi szövegek értelmezésében. Ebből a műhelyből indult el az itt ismertetendő versrepertórium-építő kezdeményezés is, mely túlélte a BIÖP-öt.

Az első papíralapú versleltárak, főleg a nemzeti irodalomtörténetek segédleteiként, a filológia hőskorában, a 19. században készültek: céljuk egy-egy költészeti hagyomány számbavétele, rendszerezése volt több-kevesebb kritérium (szereztetés adatai, műfajiság, metrikai és ritikábban retorikai-poétikai jel-

lemzők) leírásával. Az alaposabb kidolgozott-ságú metrikai repertóriumok készítése az 1960-as években indult – elsőként a provanszál trubadúrok műveinek feldolgozásával (Frank, 1962), s különös módon még a 2000-es években is jelent meg nyomtatott formában ilyen kézikönyv. Természetesen az újlatin és a germán nyelvek gazdag irodalmi kíváncsák meg leginkább az ilyesfajta számvetést; a csekélyebb mennyiségű fennmaradt szöveg alapján ismert költői hagyományok nem szorultak rá feltétlenül hasonló segédeszközre. Az első digitális versrepertórium mégsem az említett irodalmakhoz köthető: Magyarországon az 1601 előtti magyar versek anyagából készítette el a szegedi egyetem kutatócsoportja 1976 és 1992 között Horváth Iván vezetésével az RPHA-t (*Répertoire de la poésie hongroise ancienne*). Ez már a nyomtatott repertóriumok tanulságainak a figyelembevételével készült, nagyon gazdag szempontrendszer alapján dolgozta fel a verseket, és komplex keresésekre is alkalmas volt.

Egy versrepertórium legnagyobb jelentősége valójában ebben rejlik. Ha jó, akkor nemcsak leltároz, hanem a versek sok para-