Kozák Dániel

Digitális szövegkorpuszok a klasszika-filológiában[1]

1. Bevezetés

Gregory Crane, a digitális bölcsészet professzora a lipcsei egyetemen[2] és a klasszika-filológia digitális fordulatának jelenleg talán legismertebb proponense egy 2004-ben megjelent tanulmányában azt írja, hogy a klasszika-filológia és a számítógép történetéről egy ideális világban nem kellene beszélni, mivel a diszciplína művelését segítő digitális eszközök nem eléggé speciálisak ahhoz, hogy indokolt legyen egy „digitális klasszika-filológiát" elkülöníteni az általánosabb értelemben vett „digitális filológiától", illetve „digitális bölcsészettől".[3] A holt nyelveken írott szövegek tanulmányozása felvet ugyan speciális problémákat is, de valójában minden egyes nyelvnek, élőknek és holtaknak egyaránt megvannak a maga sajátosságai, melyek kezelésére digitális eszközeinket alkalmassá kell tenni. Végső soron pedig a feladat minden esetben nagyon hasonló: a túlnyomórészt nem digitális közegben született szövegek digitális kódolása a lehető legtöbb

[1] A tanulmány az Emberi Erőforrások Minisztériuma ÚNKP-17-4 kódszámú Új Nemzeti Kiválósági Programjának támogatásával készült.

[2] A Lipcsei Egyetem digitális klasszika-filológiai (és más filológiai) projektjeiről lásd az egyetem Digitális Bölcsészet Központjának honlapját: www.dh.uni-leipzig.de/wo (a jegyzetekben és a bibliográfiában feltüntetett online tartalmak elérési dátuma egységesen 2018.02.25).

[3] Crane 2004: 47. Crane tanulmánya Theodore Brunnernek, az alábbiakban tárgyalt TLG-adatbázis egyik alapító szerkesztőjének a klasszika-filológia és számítógép történetét összefoglaló korábbi írására reagál (Brunner 1993).

információ megőrzésével, s az így létrejött digitális szövegek kutathatóvá tétele egyes szövegenként és lehetőleg egybefüggő korpuszként is, minél több nyelvi és tartalmi paraméter alapján.

Két okból mégis indokoltnak látszik külön tanulmányt szentelni a klasszika-filológia digitális eszközeinek. Egyrészt: amint azt Crane maga is hangsúlyozza, a számítógép bölcsésztudományi használatának hajnalán – a kilencvenes évek végéig, vagy akár a kétezres évek közepéig – a klasszika-filológia a rokon diszciplínákétól sok tekintetben eltérő utakon és más ritmusban közeledett a mai értelemben vett digitális bölcsészet felé. Ha tehát elméleti szempontból nem indokolt is, egy történeti áttekintés keretei között mégis hasznos lehet a digitális klasszika-filológia elkülönített tárgyalása. Másrészt: a klasszika-filológia elméletét és módszertanát tárgyaló szakirodalomban mindeddig meglepően csekély mértékben jelent meg az alábbiakban tárgyalt digitális eszközök vizsgálata. Még sok tekintetben feltárásra vár és önreflexió tárgyává teendő, hogy használatuk miként és milyen mértékben változtatta és változtatja meg a diszciplína művelését.

Nem vállalkozom arra, hogy a digitális eszközök teljes spektrumát áttekintsem: kifejezetten a digitális szövegkorpuszokat tárgyalom majd, minthogy ezek – nem pedig, példának okáért, a digitális kritikai szövegkiadások – megjelenése gyakorolta a legjelentősebb hatást a klasszika-filológusok munkájára.[4] Jelen tanulmány első részében a szövegkorpuszok kifejlesztésének történe-

[4] A digitális klasszika-filológiai (illetve tágabb értelemben: ókortudományi) projektek és eszközök legteljesebb – ám sajnos nem minden tekintetben naprakész – listáját lásd wiki.digitalclassicist.org. Néhány magyar vonatkozású projektet érdemes továbbá kiemelni e helyütt. Ilyen Catullus költeményeinek online kritikai kiadása Kiss Dániel gondozásában (Catullus Online, www.catullusonline.org), Adamik Béla kutatócsoportjának nyelvészeti-epigráfiai adatbázisa (Császárkori latin feliratok számítógépes nyelvtörténeti adatbázisa, lldb.elte.hu), valamint a Szépművészeti Múzeum Antik Gyűjteményének Hyperión projektje, a kutatási eredmények széles, nem kizárólag szakmai közönséghez való eljuttatása szempontjából kiemelten fontos digitális enciklopédiák egyik példája (www2.szepmuveszeti.hu/hyperion).

tét s ezen eszközök mai állapotát tekintem át, második részében pedig néhány szempont szerint azt igyekszem körüljárni, hogy a digitális korpuszok használata miként illeszkedik egyrészt a klasszika-filológia hagyományos, „analóg" módszertanába, s mennyiben alakítja át azt másrészt, különös tekintettel az intertextuális értelmezésre, vagyis annak vizsgálatára, hogy különböző szövegek mennyiben és milyen módon alakítják egymás jelentését és értelmezését olyan esetekben is, amikor a szövegek közti hasonlóságot minden bizonnyal nem valamely szerzői szándék hozza létre.

2. A digitális szövegkorpuszok kifejlesztése

A tanulmány alapját képező konferencia-előadásra készülve áttekintettem a Digital Classicist címen, párhuzamosan Londonban és Berlinben megrendezett szemináriumsorozat utolsó három évadában elhangzott 51 előadás címét és absztraktját.[5] Tanulságosnak tűnik az előadások tematikus megoszlása:

epigráfia, papirológia, kodikológia	9 előadás
3D szkennelés, megjelenítés, nyomtatás; régészet	7
adatok vizualizációja, hálózatelemzés	7
szemantikus web, *linked open data* technológiák	7
digitális eszközök a (felső)oktatásban; hallgatók bevonása a kutatásba	5
természetes nyelvfeldolgozás (NLP)	5
digitális földrajz, topográfia	5
digitális szakirodalom: hivatkozások kezelése, katalogizálás, publikálás	3
a digitális bölcsészet általános módszertani és kulturális kérdései	2
digitális szövegkorpuszok	1

[5] www.digitalclassicist.org/wip/index.html. Nem vettem figyelembe a 2015–2017-es évadokban elhangzott néhány, az ókortudomány valamely más ágát (például egyiptológia) képviselő előadást.

Nyilvánvalóan lehetséges lett volna más kategorizálás, és figyelembe vehettem volna más előadásokat, publikációkat is; reményeim szerint azonban a lista ebben a formában is alkalmas arra, hogy alátámassza fenti kijelentésemet a digitális szövegkorpuszokra irányuló tudományos önreflexió viszonylagos hiányáról. Mindössze két előadó – egyikük éppen a már említett Gregory Crane – tárgyalt a digitális bölcsészettel és annak keretei között a klasszika-filológiával kapcsolatos általános módszertani és kulturális kérdéseket. Ez mégis csupán a második legnépszerűtlenebb témának bizonyult: az utolsó helyre kerültek éppen a tanulmányom tárgyát képező digitális szövegkorpuszok, ráadásul a velük foglalkozó egyetlen előadó sem általában tárgyalta ezeket, hanem egy speciális adatbázist: a mükénéi görög nyelvemlékeket összegyűjtő korpuszt mutatott be.

Meglepőnek tűnhet e téma alulreprezentáltsága. Azt sugallja, hogy a klasszika-filológusok az antik szövegeket kifejezetten azok materiális valójában: mint feliratot, papiruszt, kódexet vizsgálják digitális eszközökkel, s újabban a természetes nyelvfeldolgozás (NLP) eszközeit is adaptálni igyekeznek az ógörög és a latin nyelvhez. Mi a helyzet azonban magukkal az antik irodalmi szövegek lehető legnagyobb részét elérhetővé és kereshetővé tévő adatbázisokkal? Nem arról van szó, hogy a klasszika-filológia ne foglalkozna kifejlesztésükkel. Valójában ezek a korpuszok immár mintegy harminc éve léteznek, bár használatuk természetesen csak fokozatosan, a személyi számítógépek és különösen az internet terjedésével párhuzamosan vált a kutatók hétköznapjainak részévé. Az alulreprezentáltságnak az említett előadás-sorozaton tehát inkább az lehet az oka, hogy bizonyos tekintetben a téma mára „lerágott csontnak" tűnhet a klasszika-filológusok szemében: a klasszika-filológusok (és informatikus kollégáik) ma jellemzően más jellegű, informatikai szempontból sok esetben bonyolultabb digitális eszközök kifejlesztésén munkálkodnak, és értelemszerűen ezekről a munkálatokról számolnak be előadásaikban és írott (jellemzően persze online) publikációkban.[6] Ez

[6] Lásd például a Heidelbergi Egyetem *Digital Classics Online* open-access folyóiratát (journals.ub.uni-heidelberg.de/index.php/dco).

már önmagában is ezen adatbázisok korai kifejlesztésének negatív mellékhatása lehet – két további mellékhatásra később még visszatérek.

A digitális filológia történetének összefoglalásaiban megkerülhetetlen két korai, a latin nyelvű irodalom kutatását segítő eszköz: Roberto Busa *Index Thomisticus*a, Aquinói Szent Tamás műveinek az IBM közreműködésével már az 1940-es években fejleszteni kezdett számítógépes konkordanciája,[7] melynek segítségével a teológus terjedelmes munkásságának egészére vonatkozóan lehetett lekérdezni egy-egy szó előfordulásait és azok közvetlen szövegkörnyezetét, valamint David Packard 1960-as években létrehozott digitális konkordanciája a római történetíró, Titus Livius *Ab urbe condita* című részlegesen fennmaradt, de még így is meglehetősen terjedelmes szöveget alkotó munkájához. A több szerző műveit felölelő, általánosabb célú ógörög és klasszikus latin szövegkorpuszok evolúciójának rövid összefoglalását azonban valamivel később érdemes kezdenünk.[8] 1972-ben a Kaliforniai Egyetemen egy ógörög lexikográfiai adatbázis kifejlesztése ürügyén összegyűlt klasszika-filológusok végül azt a döntést hozták, hogy *Thesaurus Linguae Graecae* (a továbbiakban: TLG) címmel mégsem lexikográfiai adatbázist, hanem ennél sokrétűbb felhasználást lehetővé tévő digitális szövegkorpuszt hoznak létre, melynek tartalmaznia kell az összes fennmaradt ógörög irodalmi szöveget a homéroszi eposzoktól kezdve a Kr. u. 2. század végéig (ez mintegy 27 millió szó terjedelmű korpuszt jelent). Az „irodalom" fogalma ez esetben igen tágan értelmezendő. Vonatkozik lényegében minden terjesztésre szánt és jellemzően kéziratos formában ránk hagyományozott szövegre: a költészeten és a mai értelemben vett „szépirodalmon" túl például történetírói munkákra, szónoki beszédekre, filozófiai és tudomá-

[7] A továbbfejlesztett *Index Thomisticus* ma már szabadon elérhető az interneten: www.corpusthomisticum.org.

[8] Bővebb áttekintés olvasható a TLG honlapján: stephanus.tlg.uci.edu/history.php, lásd továbbá Brunner 1993 és Crane 2004 már említett tanulmányait.

nyos szövegekre is. Ami tehát kimaradt a TLG-ből, az a késő antik görög irodalom, a nem irodalmi papiruszok (szerződések, magánlevelek stb.), illetve a feliratok.[9]

A nagy mennyiségű szöveg rögzítésének megkezdése előtt a rendszer infrastruktúráját is meg kellett tervezni és teremteni. Szükség volt az ógörög szövegekben, illetve azok kritikai kiadásaiban nagy számban előforduló diakritikus és egyéb textuális jeleket is rögzítő kódolási rendszerre (ez a Beta Code nevet kapta),[10] valamint az adatok tárolását és keresését lehetővé tévő speciális szoftverre és hardverre (ezek együttesét az archaikus görög költők egyike után Ibycusnak nevezték el). A digitális bölcsészet jelene szempontjából sem mellékes, hogy a bölcsészek már akkor sem pusztán megrendelőként fordultak az informatikusok felé a digitális eszközök kifejlesztésekor: a két csoport között személyi átfedések is voltak. Az Ibycus rendszer kifejlesztője, az első digitális konkordanciák kapcsán már említett David Packard mérnök és ókortudós volt egy személyben, s nem mellesleg a számítástechnika történetében fontos szerepet játszó Hewlett–Packard cég társalapítójának fia; a TLG általános szövegkorpusszá fejlesztésének ötlete (s a szükséges anyagi források egy része) pedig az elektronikai eszközöket fejlesztő Zenith Corporation alapítójának lányától, Marianne McDonaldtól származott, aki éppen a boldogság kifejezéseit vizsgálta Euripidészről szóló doktori disszertációjában. David Packard nem sokkal később, az 1980-as években tervezett egy speciális személyi számítógépet is a TLG használatához, ez azonban az általános célú PC-k terjedésével hamar idejétmúlttá vált. Maga az adatbázis idővel mágnesszalagról CD-ROM-ra költözött; 1985-ben éppen a TLG lett az első publikált, nem zenei adatot tartalmazó CD. A David Packard alapította Packard Humanities Institute eközben elkészítette a

[9] A görög nyelvű feliratok viszonylag bő, de korántsem teljes adatbázisa: epigraphy.packhum.org; a papiruszgyűjtemények tekintetében lásd papyri.info.

[10] A Beta Code rövid áttekintése és teljeskörű dokumentációja egyaránt elérhető: stephanus.tlg.uci.edu/encoding.php.

TLG latin megfelelőjét, a PHI adatbázist, mely szintén Kr. u. 200-ig tartalmazta a római irodalmi szövegek lényegében teljes korpuszát. E két, technikailag is azonos felépítésű és ugyanazon kliensprogramokkal kezelhető adatbázis tehát a 90-es években egy-egy CD-n a klasszikus görög és latin irodalom túlnyomó részét hozzáférhetővé tette – méghozzá megbízható és sok esetben meglepően friss kritikai kiadások szövegét átvéve –, és az egész korpuszra kiterjedő, gyors keresési lehetőségeket biztosított.

Ezt követően viszont a két adatbázis útjai elváltak egymástól. A TLG 2001-ben webes felületet kapott, és előfizetéses üzleti modellre váltott (bár egy igen korlátozott szövegválogatás továbbra is szabadon elérhető), a kronológiai lefedettség pedig jelentősen bővült: ma már a bizánci irodalom is elérhető egészen 1453-ig, így a TLG összesen mintegy százmillió szó terjedelmű szöveget tartalmaz.[11] Integráltak több, a szövegek értelmezését segítő eszközt: a sztenderd ógörög–angol nagyszótár legfrissebb kiadását, valamint egyes speciális szerzői szótárakat. Megtörtént a szövegek lemmatizálása (a ragozott szóalakok megfelelő lexémához rendelése), így könnyen megtalálható egy-egy szó bármely előfordulása, függetlenül attól, hogy milyen alakban jelenik meg az egyes szövegekben – ez a görög nyelv rendkívüli morfológiai és dialektikus változatosságát tekintve különösen fontos. E fejlesztéseknek köszönhetően tehát egyetlen lekérdezéssel kideríthető, hogy egy adott jelentésmezőhez tartozó görög szavaknak, vagy egy adott jelzős kifejezésnek bármely ragozott alakja (szórendtől függetlenül) hol fordul elő a teljes korpuszban, vagy azon belül egy szerzői életműben, irodalomtörténeti korszakban, esetleg műfajban. Újabban a TLG ún. n-gram keresője arra is képessé vált, hogy két kiválasztott szöveget összehasonlítva automatikusan felismerjen egyes nem szó szerinti idézeteket, illetve mindkét szövegben előforduló idiómákat is.

Miközben a TLG így fejlődött, a latin PHI adatbázis hosszú időre tetszhalott állapotba került, míg végül 2015-ben készítői

[11] stephanus.tlg.uci.edu.

ingyenes online felületen tették elérhetővé a legutolsó, 1998-ban megjelent verziót.[12] Az adatbázis tehát nem bővült, s azóta sem bővül új szövegekkel,[13] integrált szótárakkal, s a felhasználói felület is – noha sok szempontból praktikus – igen korlátozott keresési lehetőségeket nyújt. A TLG-nek ma a korpusz kiterjedtségét és a keresési lehetőségeket tekintve inkább a Brepols kiadó 2009 óta elérhető *Library of Latin Texts* (LLT) nevű, ugyancsak előfizetéses adatbázisa feleltethető meg, mely az antik latin nyelvű szövegek lényegében teljes (tehát nem csak a Kr. u. 200 előtti) fennmaradt korpuszán túl hozzáférést biztosít rengeteg középkori, újkori és kortárs szöveghez egészen a II. vatikáni zsinatig bezárólag, s jól paraméterezhető, lemma-alapú keresővel rendelkezik.[14]

3. A digitális korpuszok és a nagyközönség

A fentiekben tárgyalt három szövegkorpusz kifejlesztéséhez a klasszika-filológia szaktudományos igényei vezettek, ez pedig az adatbázisok és keresőfelületük jellegét is alapvetően befolyásolja. Célközönségüket maguk a klasszika-filológusok alkotják: azok a latin és ógörög nyelvben, valamint az antik kultúrában és történelemben jártas szakemberek, akiknek a szövegek szó szoros értelmében vett, elsődleges megértéséhez jellemzően nincs szükségük fordításokra, az egyes szóalakok nyelvtani magyarázatára, és szótárhoz sem minden egyes mondat értelmezésekor

[12] latin.packhum.org (lásd továbbá az online változatról írt recenziókat: Loar 2017; Kozák 2018).

[13] A PHI-t (is) kiegészítő speciális korpusznak tekinthető a késő antik, elsősorban nem keresztény szövegeket egybegyűjtő digilibLT: digiliblt.lett.unipmn.it. A latin nyelvű feliratok legteljesebb, teljes szövegű keresést biztosító online metakeresője az Epigraphik-Datenbank Clauss/Slaby (db.edcs.eu; Adamik Béla szíves szóbeli közlése).

[14] clt.brepolis.net/llta. Az LLT egy keresztény latin szövegeket tartalmazó CD-ROM adatbázis (CETEDOC) alapjaira épült.

kell nyúlniuk. Érdemes röviden két olyan szöveges adatbázisról is szót ejteni, mely (részben legalábbis) más közönséget céloz meg. Az egyik a Latin Library, mely amatőr vállalkozásként már a kilencvenes évek végén – amikor még a PHI adatbázis is csak CD-ROM-on volt elérhető – szabadon hozzáférhetővé tett az interneten rengeteg antik és posztantik latin szöveget.[15] A mai szemmel nézve rendkívül egyszerű honlapnak nincs saját keresője; a szerzői jogi korlátok miatt régi és sok esetben eleve megbízhatatlan szövegkiadásokon alapul, ráadásul a digitalizálás során keletkezett szöveghibák egy része sincs javítva. A Latin Libraryt tehát nem tekinthetjük megbízható tudományos eszköznek (erre a latin szövegeket maguk számára gyakran innen nyomtató egyetemi hallgatókat is figyelmeztetnünk kell). A másik adatbázis az 1995 óta elérhető és folyamatosan fejlesztett Perseus Digital Library,[16] mely ógörög és latin szövegeket egyaránt tartalmaz, vagyis ahelyett, hogy az antik irodalom két nagy korpuszát egymástól nyelvi alapon elválasztaná, sokkal inkább azt hangsúlyozza, hogy azok sok tekintetben kulturális egységet alkotnak, ezért vizsgálatuk elválaszthatatlan egymástól. Legalább ennyire fontos, hogy a Perseus hozzáférhetővé tesz fordításokat, szótárakat, kommentárokat, valamint egy-egy szaklexikont is, melyek vonatkozó szakaszai a tárgyalt antik szövegek mellett is megjeleníthetők, az antik szövegek megértését pedig az egyes szavak nyelvtani analízise is segíti. A Perseus tehát nem csupán szövegkorpusz, hanem „digitális filológiai munkakörnyezet" (vagy annak prototípusa) is egyben, mely sok esetben a tudományos munkát is megbízhatóan támogatja. Kifejezetten szövegkorpuszként viszont hagy kívánnivalót maga után: a már tárgyalt korpuszokkal (TLG, PHI, LLT) ellentétben az ógörög és római (vagy latin) irodalom egyetlen korszakára vonatkozólag sem tekinthető teljesnek, vagyis a segít-

[15] www.thelatinlibrary.com; lásd még Gellar-Goad 2016 recenzióját, továbbá Tarrant 2016: 149–151 kritikus megjegyzéseit.

[16] www.perseus.tufts.edu; recenzió erről: Lang 2018. A Perseus legújabb, 5.0-ás verziója megjelenés előtt áll; fenti megjegyzéseim és Lang recenziója még a 4.0-ás verzión alapulnak.

ségével végrehajtott keresések sem lehetnek minden tekintetben reprezentatívak.

E helyütt mégsem a Latin Library és a Perseus hiányosságait szeretném hangsúlyozni, hanem azt, hogy a digitális szövegkorpuszok műfaján belül mindkettőnek meghatározó szerepe van a klasszika-filológia társadalmi beágyazottságának, a szakmán kívüliekkel való kapcsolatának erősítésében. Az antik irodalmi szövegeket nem elzárt, a kutatók számára hozzáférhető és vizsgálandó korpuszként mutatják be, hanem mindannyiunk közös és éppen ezért szabadon hozzáférhetővé teendő kulturális örökségeként. Ennek az örökségnek a megértése nehéz ugyan és klasszika-filológus szakértőket igényel, de utóbbiak feladata az is, hogy a megértés legalább egyes útjainak bejárásában az érdeklődők lehető legszélesebb köre számára segítséget nyújtsanak.

4. A korpuszok korai kifejlesztésének néhány mellékhatása

A klasszika-filológusok joggal büszkélkedhetnek azzal, hogy rendkívül korán hozzákezdtek a digitális szövegkorpuszok kifejlesztéséhez; ritkábban esik azonban szó ennek néhány káros mellékhatásáról. E mellékhatások részben technikai jellegűek. Feltehetőleg már az adatbázisok (nem minden esetben jól dokumentált) felépítése sem ideális a mai elvárások felől nézve – különösen igaz lehet ez az 1998 óta lényegében nem fejlesztett PHI-re. Még feltűnőbb azonban az elmaradás az egyes szövegek digitális reprezentációját tekintve. A TLG és a PHI tervezésekor még nem léteztek a szövegek digitális, filológiai igényű kódolásának olyan, ma lényegében megkerülhetetlen szabványai, mint a Unicode, a TEI és az XML. A Beta Code (lásd fönt) karakterkódolását viszonylag egyszerű Unicode-dá konvertálni, de más elemeinek TEI/XML-re való átalakítása nem automatizálható, és sok esetben a Beta Code nem is rögzít minden olyan információt, mely nélkül a TEI/XML kódolás

nem valósítható meg.[17] A meglévő korpuszok technikai szempontból vett korszerűsítése tehát hatalmas emberi és anyagi erőforrásokat igényelne.

Kérdés ugyanakkor – amint azt Franz Fischer nemrég megjelent tanulmányában vizsgálja –, hogy reális-e ugyanolyan elvárásokat támasztani a digitális szövegkorpuszokkal, mint egyes szövegek digitális (értsd: eleve digitális közegben született, *born digital*) kritikai kiadásaival szemben.[18] Értelemszerűen több időt és energiát szánhatunk egységnyi terjedelmű szöveg digitális kódolására, ha egyetlen mű kiadását készítjük el, mint ha több száz vagy ezer szövegből építünk korpuszt; az utóbbi esetben az egységes reprezentáció igénye felülírhatja az egyes szövegek jellegzetességeihez legjobban idomuló és legaprólékosabb kódolás iránti igényt. A Fischer által javasolt kompromisszum szerint a digitális szövegkorpuszban helyet kapó szövegek esetében a „digitális" kritériumok rugalmasan kezelendők, a filológiai kritériumokból viszont nem engedhetnek a szerkesztők.

A klasszika-filológia meglévő korpuszainak minősége e kompromisszumos javaslat szerint sem kielégítő. Korai kifejlesztésük káros mellékhatásai között ugyanis filológiai jellegűek is akadnak, s ezek közül minden bizonnyal éppen a kritikai apparátus hiánya a legsúlyosabb. A TLG és a PHI kifejlesztésekor még nem zajlott le az az „új filológiához" köthető szemléletbeli változás, melynek eredményeképpen a szövegkiadók által elvetett szövegvariánsok is felértékelődtek, és az interpretáció szempontjából érdekesebbé váltak.[19] Szerzői jogi és technikai problémák mellett

[17] Kifejezetten technikai szempontból a Perseus fejlettebb, mint a TLG és a PHI (például az újabban feltöltött szövegek már TEI/XML szabvány szerint kódoltak), ám a korpusz nagyságát tekintve nem veheti fel a versenyt az utóbbiakkal, s ezért nem is léphet a helyükre.

[18] Fischer 2017: 275–287; a szerkesztői szerepről a digitális szövegkorpuszok vonatkozásában lásd még Crane–Rydberg-Cox 2000.

[19] Az „új filológiáról" magyarul lásd például a *Helikon* 2000/4-es tematikus számát; a szöveg variabilitásáról elméleti összefüggésben Déri–Kelemen–Krupp–Tamás 2011 vonatkozó tanulmányait (219–432).

valószínűleg ebből is fakad, hogy a TLG és a PHI megbízható nyomtatott kritikai kiadásokon alapul ugyan, de kizárólag a szövegkiadó által megállapított főszöveget veszi át; a szövegkiadások digitalizálása a bevezetőkre és a kritikai apparátusra nem terjedt ki. A TLG történetében többször is előfordult már, hogy egy szövegnek új szövegkiadása került az adatbázisba, ezek azonban a régi helyére léptek, a két kiadás nem vált párhuzamosan megjeleníthetővé, illetve kereshetővé.[20] A korpuszok jelenlegi változatában tehát nemcsak az esetleges szövegvariánsokra nem lehet rákeresni, hanem minden szövegtörténeti információ, továbbá a hagyományozott szöveg inherens variabilitásának ténye is elfedésre kerül – ennek pedig különösen a klasszika-filológusok legújabb, immár digitális szövegeken és szövegkorpuszokon nevelkedő, s azokat nem annyira a nyomtatott kiadások mellett, hanem inkább helyettük használó generációjára gondolva lehetnek módszertani kockázatai.[21]

[20] Vö. Crane–Bamman–Jones 2013: 35.

[21] A Lipcsei Egyetemen működő *Open Greek and Latin* projekt (www.dh.uni-leipzig.de/wo/projects/open-greek-and-latin-project) egyik célja éppen az, hogy egy-egy szövegnek több szövegkiadását is digitalizálja, és azok főszövegét összehasonlítva legalább részleges rálátást biztosítson a szövegtörténetre is. Az antik és posztantik latin nyelvű költészet speciális korpusza, a *Musisque Deoque* (www.mqdq.it) célja, hogy a nyomtatott kiadások kritikai apparátusát is legalább részben digitalizálja: azokra az olvasatokra kiterjedően, melyek ugyancsak értelmes, az irodalmi interpretáció szempontjából figyelembe vehető szöveget eredményeznek. A szövegkritikai információk jelenleg csak egyes szövegek (például Vergilius *Aeneis*e) esetében érhetők el. E helyütt említendő továbbá, hogy az egyesült államokbeli Society for Classical Studies új digitális kritikai kiadássorozatot tervez indítani (digitallatin.org), melynek technikai és filológiai alapelvei jelenleg még kidolgozás alatt állnak (digitallatin.github.io/guidelines/LDLT-Guidelines.html).

5. A klasszika-filológia „hagyományos” gyakorlata és a digitális szövegkorpuszok

Klasszika-filológusként csábító lehetőség természetesen, hogy azt mondjam: kollégáim valamiféle jóstehetségtől indíttatva kezdtek hozzá már a hetvenes években az említett digitális korpuszok létrehozásához. Az okok azonban minden bizonnyal ennél praktikusabbak voltak, illetve részben a klasszika-filológia hagyományos módszertanából fakadtak. Lássuk először a praktikus okokat. Az antik (és azon belül is különösen a Kr. u. 200-nál korábbi, a TLG- és a PHI-korpuszban eredetileg összegyűjtött) latin és ógörög irodalmi szövegek közepes méretű és statikus korpuszt alkotnak. Közepes méreten e helyütt azt értem, hogy a korpusz egyrészt van annyira terjedelmes, hogy a filológus egyéni, természetes memóriája már ne tudja teljes egészében befogni, sőt, a nyomtatott szótárak, konkordanciák és egyéb kézikönyvek is csak korlátozottan legyenek képesek reprezentálni, egy-egy jelenség összes előfordulása helyett csupán példákat hozva. A korpusz ugyanakkor – egy-egy élő nyelvi irodalomével összehasonlítva – eléggé korlátozott terjedelmű is ahhoz, hogy a szövegek kézi felvitele (az ellenőrzést is beleértve) ugyan hatalmas, de mégsem beláthatatlan mennyiségű munkát jelentsen. A szóban forgó korpusz továbbá majdnem teljesen statikus. Új antik szövegek értelemszerűen nem keletkeznek, s bár időről időre még mindig előkerülnek eddig ismeretlen irodalmi szövegek (gondoljunk például a kétezres évek közepén publikált Szapphó-papirusztöredékekre), ezek a korpusz összterjedelmét érdemben nem módosítják. A korpusz statikus jellegéből fakadóan nem kellett és ma sem kell attól tartani, hogy a digitalizálás nem tud majd lépést tartani a korpusz bővülésével. Az említett adatbázisok kifejlesztése tehát meglévő tudományos igények kielégítését ígérte, ugyanakkor reális célnak is bizonyult. Összehasonlításképpen: a hetvenes évek technikai infrastruktúrájával az angol vagy német nyelvű irodalmi szövegek teljes körű digitalizálása, akár csak a századfordulóig bezárólag, vagy akár csak egy-egy országra korlátozva minden bizonnyal elképzelhetetlen lett volna.

A digitális szövegkorpuszok korai kifejlesztésének másik oka, mint említettem, vélhetőleg a diszciplína hagyományos, digitális kort megelőző „analóg" hagyományaiban keresendő. A ma ismert antik görög és latin szövegek korpusza nemcsak közepes méretű (a fenti értelemben), hanem jelentős mértékben fragmentált is. Nemcsak azokra a művekre gondolok, melyek töredékesen maradtak ránk, vagy akár kizárólag címüket ismerjük, hanem azokra is, melyek többé-kevésbé hiánytalanul átvészelték az évezredeket – s ez utóbbiakra sem kizárólag azért utalok, mert maguk is többé-kevésbé bizonytalan kézirati hagyomány alapján rekonstruálhatók csupán. Mindazok az antik szövegek, fragmentumok és címek, melyeket ma ismerünk, csupán töredékek egy szöveges rommezőn: emlékeztetnek minket egy valaha sokkal terjedelmesebb korpuszra. Minden kultúrára igaz természetesen, hogy egyes szövegei elveszhetnek a jövő filológusai számára; de a helyreállíthatatlan szövegromlás és pusztulás mértéke az ókori kéziratos kultúrák esetében értelemszerűen különösen nagy.

Ezt a fragmentált korpuszt ugyanakkor a klasszika-filológusok – bár töredékességével természetesen ők maguk vannak a leginkább tisztában – hagyományosan nagyon is egységes és egybefüggő szöveganyagként kénytelenek kezelni. A tőlünk nemcsak időben, hanem kulturális és nyelvi tekintetben is távoli szövegek értelmezéséhez – sőt gyakran elsődleges, grammatikai megértéséhez is – minden rendelkezésre álló összehasonlító adatra szükségünk van. Két, az irodalomtörténeti narratívákban egymástól igen távoli szöveget, illetve azok egy-egy részletét is közvetlenül összevetjük egymással (például egy Kr. e. 5. századi görög nyelvű orvosi szöveg néhány szavas töredékét Vergilius *Aeneis*ének egy sorával), ha úgy véljük, hogy ez segít megoldani egy szövegkritikai problémát, megválaszol valamilyen nyelvi vagy tartalmi kérdést, vagy bármilyen tekintetben előmozdítja az interpretációt. Ezzel természetesen nem feltétlenül állítjuk, hogy a két szöveg között irodalmi értelemben vett intertextuális kapcsolat is fennállna (azaz idézetről vagy allúzióról volna szó), s még kevésbé azt, hogy a későbbi szöveg szerzője ismerte a korábbi szöveget. Mégis, filológiai tevékenységünk a szó legáltalánosabb – ugyan-

akkor nagyon is pragmatikus – értelmében intertextuális, szövegközi kapcsolatokat hoz létre a korpusz különböző szövegei, pontosabban azok egy-egy jellemzően apró részlete között. Természetesen nem azt állítom, hogy kizárólag klasszika-filológiai gyakorlat volna ez, csupán azt, hogy az antik szöveg megértése és értelmezése esetében más filológiákkal összehasonlítva különösen jellemző praxisról beszélhetünk.

Nem véletlen, hogy az értelmezett szöveget mintegy atomjaira bontva – Roland Barthes szavaival: „csillagokká repesztve"[22] – vizsgáló kommentár mindig is ennyire nélkülözhetetlen és jellemző szakirodalmi műfaj volt és maradt mindmáig a klasszika-filológiában.[23] Egy adott szövegrészhez – egyszerre gyakran csak egy-két szóhoz – írt tipikus kommentár-bejegyzés kisebb-nagyobb részben éppen az értelmezést segítő „párhuzamos szöveghelyek" *(loci paralleli/similes)* listájából épül fel. E listát gyakran a „vö. pl." hírhedt formulája vezeti be, helyettesítve és a kommentár olvasójára bízva a párhuzamok interpretációját.[24] Ez a formula és a párhuzamos szöveghelyeket jelölő többé-kevésbé szabványos, minden kiadás által követett rövidítések (például Cic. *Leg*. 2.30 = Cicero, *De legibus* 2. könyv, 30. *caput* [fejezet]) mintegy pre-digitális linkek módjára[25] biztatják a kommentár olvasóját, hogy az egyik szövegről a másikra „ugorjon", majd onnan vissza az elsőre, vagy éppen (a másik szöveghez írt kommentárbe-

[22] Barthes 1997: 25–26.

[23] A kommentár műfaja kiemelt figyelmet kapott az utóbbi két évtized filológiatörténeti és -elméleti kutatásaiban. Lásd például a következő tanulmánykötéteket: Most 1999, különös tekintettel Don Fowler tanulmányára; Gibson–Kraus 2002; Kraus–Stray 2016.

[24] Erről bővebben lásd Gibson 2002. Gibson a költői szövegek értelmezési hagyományára koncentrálva a párhuzamos szöveghelyek egy lehetséges tipológiáját is kidolgozza (333–346).

[25] A Classical Works Knowledge Base (cwkb.org) a *linked open data* technikájával lehetővé teszi e hagyományos szöveghely-jelölések digitális linkekké alakítását, a felhasználó pedig választhat, hogy melyik digitális korpuszban szeretné az adott szöveghelyet megnyitni. Lásd még a Canonical Text Services szolgáltatást (cite-architecture.github.io/cts/).

jegyzést is megnézve) tovább egy harmadikra, negyedikre és így tovább a végtelenségig. Az antik irodalom kutatása során tehát a klasszika-filológusok az ilyen értelemben vett szövegközi kapcsolatok végtelenül sűrű és komplex hálózatát is felépítették; és éppen ezek a kapcsolatok teszik az egyébként töredékes korpuszt – hangsúlyozottan mint a tudományos vizsgálódás tárgyát – a fent tárgyalt értelemben mégis egységessé és egybefüggővé. Ez a hálózat természetesen csak virtuális, hiszen egészében senki sem rajzolta még fel – a vállalkozás minden bizonnyal lehetetlen is volna –, de az antik szövegek tudományos recepciójáról minden bizonnyal nagyon sokat elárulna, ha látnánk, hogy mely szövegek (illetve szövegrészek) játszanak középponti szerepet benne: melyekre hivatkoznak különösen sokszor egy-egy másik szöveg magyarázataképpen, vagy fordítva, melyek esetében szokás átlagon felüli számú párhuzamot felvonultatni a magyarázathoz.[26]

A digitális szövegkorpuszok korai kifejlesztése valószínűleg részben annak is köszönhető (ha ez nem is feltétlenül tudatosult, s jellemzően ma sem tárgyalt szempont), hogy azok jutnak a legközelebb az antik szövegek imént vázolt, hálózatként elképzelt modelljének gyakorlati megvalósításához. Ha a vizsgálatba bevonható szövegeket nemcsak közös fizikai térben (könyvtárban, könyvespolcon, de különálló kötetekben) helyezzük egymás mellé, hanem egyetlen, konzisztensen kódolt digitális szövegtérben (adatbázisban) helyezzük el őket, akkor új lehetőségek nyílhatnak meg a kutatók előtt. A legkézenfekvőbb haszon nyilvánvalóan a szövegekhez való hozzáférés megkönnyítése és felgyorsítása: a filológus a számítógépe előtt ülve néhány másodperc alatt az őt érdeklő szövegek bármelyikének bármely

[26] Az antik szövegeket a tudományos publikációkban összekapcsoló „hivatkozási hálózatok" automatikus digitalizálásának lehetőségeiről lásd Romanello 2016: 21–39. Vergilius *Aeneis*ére – és csupán a szakirodalom egy részletére – vonatkozólag lásd az „*Aeneid* in Jstor" (aeneid.citedloci.org/) szolgáltatást, mely az eposz „hőtérképét" is felrajzolva mutatja, hogy egy-egy *Aeneis*-sorra mely Jstorban elérhető tanulmányok hivatkoznak (idézettel vagy a nélkül).

részletét meg tudja jeleníteni (és készülő tanulmányába átmásolni). A szövegről szövegre ugrás meggyorsításán túl azonban bizonyos típusú szövegközi párhuzamok keresésének is új útjai nyílnak meg. A legkönnyebben természetesen az azonos lexikai bázison alapuló szövegközi párhuzamok találhatók meg, illetve egy-egy szó vagy kifejezés nyelv- és irodalomtörténeti „életútja" vázolható fel;[27] de a lehetőségeknek csak az szab határt, hogy digitális szövegkorpuszunk keresője milyen paraméterek szerint „finomhangolható", s hogy emberi közreműködéssel vagy automatizált, gépi eszközökkel mennyi és milyen (nyelvi, irodalmi, kulturális) információt kapcsolunk a szöveg egy-egy szegmenséhez (betűhöz, szóhoz, mondathoz stb.). A TLG kifejlesztésekor éppen a teljes korpuszra kiterjedő keresés, nem pedig a gyors hozzáférés biztosítása volt az elsődleges cél. Ez annak tükrében tűnik különösen említésre méltónak, hogy – egy nemrég megjelent tanulmány szerint – az angol filológia egyes hasonló digitális korpuszainak létrehozásakor a kilencvenes és kétezres években a szövegeken átívelő kereshetőség másodlagos szerepet játszott, s e funkció majdnem ki is maradt a fejlesztésből.[28]

6. Az olvasás módozatai

A digitális szövegkorpuszok által nyújtott szolgáltatások nem egyszerűen felgyorsíthatják a korábban is végzett műveleteket, hanem – ezzel összefüggésben – át is alakíthatják a létező praxisokat, munkamódszereket. Egyszerű példával élve: egy engem érdeklő tanulmányban említett, de nem idézett szöveghelyek kö-

[27] Erre jó példa lehet John Richardson monográfiája az eredetileg „parancsnoki jogkör", s csak később (a kora császárkortól) „birodalom" jelentéssel bíró *imperium* szó jelentésfejlődéséről írott, a korábbi munkáknál lényegesen több forráshelyet vizsgáló monográfiája (Richardson 2008); a szerző digitális munkamódszeréről bővebben lásd Richardson 2005: 139–140.

[28] Bilansky 2017: 513–515, az Early English Books Online és a Women Writers Online adatbázisokról.

zül minden bizonnyal többet fogok ténylegesen is megnézni, ha nem kell mindegyikért könyvtárba mennem, és ott adott esetben a raktárból kikérnem. Nem feltétlenül arról van tehát szó, hogy ugyanazt a munkát gyorsabban végzem el, hanem arról, hogy a munka elvégzésére szánt, szükségképpen véges időben reprezentatívabb áttekintést szerezhetek az adott téma szempontjából releváns forrásokról.

Az imént tudatosan választottam az „olvasni" helyett a „megnézni" és az „áttekintést szerezni" kifejezéseket. A digitális médiumok fejlődésével párhuzamosan növekvő figyelmet kaptak a szöveges tartalmak befogadásának, tág értelemben vett olvasásának különböző módozatai és stratégiái. Az (irodalmi) szöveg aprólékos, a lehető legtöbb részletre kiterjedő és ezeket a részleteket az esztétikailag egységesnek tekintett mű egészének interpretációjában hasznosító lineáris „szoros olvasás" *(close reading)* hosszú időn át megkérdőjelezhetetlen stratégiája mellett – vagy éppen helyett – ma olyan módszerek és jellemzően digitális eszközeik képezik kiemelt figyelem tárgyát, mint a makroanalízis, a „távoli olvasás" *(distant reading)* és az „algoritmikus olvasás".[29] A különböző kutatók által javasolt – részben egymást kiegészítő, részben egymással rivalizáló – terminusok, illetve az általuk jelölt értelmezői stratégiák közös jellemzője, hogy az „olvasás" tárgyaként az emberi elme által terjedelmük miatt nehezen vagy egyáltalán nem belátható szövegeket, illetve szövegkorpuszokat határozzák meg, az „olvasás" digitális eszközeire koncentrálnak, és sok esetben a szövegek „szabad szemmel", a szoros olvasás stratégiáját követve nem is tudatosuló, inkább grafikonokon és térképeken ábrázolható tulajdonságainak jellemzően kvantitatív elemzésére

[29] Makroanalízis: Jockers 2013; távoli olvasás: Moretti 2005 és 2013; algoritmikus olvasás: Ramsay 2011. Az olvasási praxisok átalakulásáról általában lásd Hayles 2012: 55–79; a távoli olvasásban rejlő lehetőségekről magyarul lásd például Labádi 2014; Péter 2016; vö. továbbá a *Publications of the Modern Language Association of America* 2017 májusában megjelent 132 (3)-as számának tematikus – és Moretti megközelítésével szemben több esetben kritikus tanulmányokat tartalmazó – összeállítását.

építik az interpretációt. Franco Moretti provokatív hipotézise szerint az irodalmi kánont alkotó viszonylag kevés mű szoros olvasása helyett a teljes irodalmi produkció „távoli olvasása" révén tudunk egy-egy műfajról vagy irodalomtörténeti korszakról releváns értelmezéseket adni – Moretti például az angol regény és alműfajai 18–19. századi történetét elemzi több ezer regény címét és narratív alapsémáját (de hangsúlyozottan *nem* a szó szoros értelmében vett szövegét) vizsgálva.[30]

Az ilyen értelemben vett „távoli olvasás" lehetőségei a klasszika-filológiában korlátozottnak tűnnek, nem utolsósorban éppen az antik irodalmi szövegek korpuszának fentiekben már tárgyalt közepes mérete és fragmentáltsága következtében. A könyvnyomtatás és az irodalmi tömegtermelés kezdete előtt keletkezett szövegekről van szó, melyek közül sok mára végérvényesen elveszett. A szövegek összességeként definiált korpusz még eléggé terjedelmes ahhoz, hogy kvantitatív, statisztikai megközelítésben vizsgáljuk például egy-egy szó, kifejezés vagy akár grammatikai szerkezet előfordulásait; ám ha például az egyes műfajokba sorolható és legalább címként fennmaradt művek számát vesszük alapul, akkor már nem kapunk megfelelően nagy elemszámú adathalmazt ahhoz, hogy olyan „távoli olvasásnak" vessük alá azt, mint Moretti tette az angol regénnyel. A klasszika-filológia szempontjából ezért azok a digitális olvasás módozataival összefüggő kutatások tűnnek különösen relevánsnak, melyek a *full text* adatbázisokat mint szöveges „műfajt", s ezzel összefüggésben az adatbázis „olvasását", kezelését mint filológiai praxist vizsgálják. E kutatások közül több is hangsúlyozza, hogy az egyedi szöveg szoros olvasása és az adatbázison lefuttatott keresések eredménylistáinak a távoli olvasás gyakorlatával rokonítható áttekintése és elemzése inkább egymást kiegészítő, mintsem szembeállítható és önállóan alkalmazandó stratégiák, s hogy éppen az adatbázis mint műfaj az, ami elősegíti a két értelmezői stratégia kooperatív alkalmazását.[31] A digitális szövegkorpuszok kiváló eszközök arra,

[30] Moretti 2005: 3–33.
[31] Hayles 2012: 175–247; Bilansky 2017: 518–519.

hogy egyszerre kapjunk nem negatív értelemben véve felületes áttekintést[32] egy-egy nyelvi jelenség vagy téma előfordulásairól az adott korpuszban (s ezt mint kvantitatív információt akár listákban, grafikonokon stb. ábrázoljuk is), valamint e tevékenység közben egyúttal ki is választjuk a találatok közül azokat, melyeket a szoros olvasás módszerével részletesebb (kvalitatív) vizsgálatra is érdemesnek tartunk. Hogy a *distant reading* és a *close reading* egymást kiegészítve alkalmazható értelmezői stratégiák, arra jó példát szolgáltathat a klasszika-filológia fent vázolt jellemző – de természetesen nem kizárólagos – értelmezési gyakorlata is. A szövegek közötti hasonlóságok kimutatása és ezáltal a szövegek komplex hálózatának építése során az adott esetben digitális szövegkorpuszok és azok keresői segítségével megtalált szöveges párhuzamokat mint „adatokat" jellemzően egy adott szöveg értelmezésében, annak szoros olvasásához kapcsolódóan használjuk fel.

7. Szövegpárhuzamok keresése digitális eszközökkel

Amint arról fent már esett szó, a szövegpárhuzamok keresése a digitális kort megelőzően is a klasszika-filológia jellemző gyakorlata volt. Kérdés tehát, hogy milyen tekintetben nyújtanak új lehetőséget erre a digitális szövegkorpuszok és keresőik. Először is természetesen abban, hogy több párhuzamot találhatunk a segítségükkel; csakhogy a párhuzamos helyek sokszor már így is kénye lmetlenül hosszú listájának[33] további bővítése egy ponton túl

[32] A „hiperolvasásnak" többek a digitális keresők által összeállított találati listák feldolgozásában hasznos technikáit tárgyaló tanulmányában James Sosnoski nyolcat különböztet meg; ezek közül a digitális filológia szempontjából a szűrés *(filtering)*, a felületes *(skimming)* és válogató olvasás *(pecking),* a szövegek közti határok átlépése *(trespassing)* és a fragmentálás *(fragmenting)* tűnnek különösen fontosnak (Sosnoski 1999: 163–172).

[33] A „párhuzamvadászat" negatív aspektusairól lásd Gibson 2002: 347–353.

minden bizonnyal nem járna együtt a filológiai megértés elmélyítésével, s a digitális kutatás során ellen is kell állnunk a puszta halmozás kísértésének. A hangsúlyt inkább a változatosságra helyezném. Olyan párhuzamokat is könnyedén megtalálhatunk a digitális keresők – és persze helyesen megválasztott keresőkifejezések – segítségével, melyekre korábban saját olvasottságunk, emlékezetünk szükségszerű korlátai, a szótárak és más kézikönyvek korlátozott példaanyaga miatt nem bukkantunk volna rá, vagy éppenséggel azért nem, mert akár tudatosított, akár tudat alatti előzetes értelmezési gesztussal bizonyos szövegeket eleve nem találtunk relevánsnak a komparatív vizsgálat szempontjából. Ha Vergilius *Aeneis*ének egy részletét vizsgálva a tíz, már korábban is felismert homéroszi párhuzamot további öttel egészítjük ki, azzal viszonylag keveset nyertünk; de ha találunk akár csak egy korábban nem dokumentált releváns párhuzamot például a jogi szövegek között, az lényegesen módosíthatja a vizsgált szövegrészről alkotható értelmezéseket. A digitális szövegkorpuszokon végrehajtott keresések többek között éppen arra bizonyulhatnak különösen alkalmasnak, hogy a korábbi kutatásban rögzült, esetenként egyoldalú értelmezéseket újraértékeljük és kiegészítsük egy-egy szöveghely intertextuális hátterének teljesebb feltárásával.[34]

A változatosság mellett a reprezentativitás is a digitális korpuszok használata mellett szólhat. A nyomtatott szótárak (a rendkívül ritka szavaktól eltekintve) nem sorolhatják fel egy-egy szó vagy akár szókapcsolat összes előfordulását. A filológusok gyakran relatív és pontatlan kijelentéseket kénytelenek tenni arról egy-egy szöveghely kapcsán, hogy az adott kifejezés „gyakran/ritkán fordul elő a klasszikus latinban" vagy „a költői szövegekben". De vajon mi volna az átlagos: nem is túl gyakori és nem is túl ritka előfordulás? És tételesen mely szövegek képezik a „klasszikus

[34] Ennek lehetőségeit egy készülő esettanulmányban mutatom be részletesen, egy kulcsfontosságú *Aeneis*-helyhez (*imperium sine fine dedi,* „határtalan/végtelen hatalmat adtam [a rómaiaknak], *Aen*. 1.279) kapcsolódóan: Kozák (megjelenés előtt).

latin" vagy a „költői szövegek" korpuszát? A digitális eszközök lehetővé teszik, hogy pontosítsuk ezeket a kijelentéseket: „az adott kifejezés X alkalommal található meg az Y digitális korpuszban" vagy annak „Z alkorpuszában".[35] A hasonló kijelentések nem is annyira önmagukban, mint inkább egymással összehasonlítva sugallhatnak a korábbiaknál megalapozottabb értelmezéseket egyes kifejezések relatív gyakoriságáról.

A fentiekkel összefüggésben fontosnak tűnik a digitális korpuszokban lefuttatott keresések reprodukálhatósága, illetve ennek a publikálás során való biztosítása is. A filológusnak legalább a kulcsfontosságúnak ítélt esetekben érdemes pontosan dokumentálnia, hogy milyen keresőkifejezést és milyen keresési paramétereket használt. Az ideális eset az volna, ha a digitális szövegkorpuszok lehetőséget adnának arra, hogy a keresésekhez egyedi azonosítókat rendeljünk, s ezeket mint hivatkozásokat tüntessük fel publikációnkban, így megadva a lehetőséget az olvasók számára, hogy megvizsgálhassák ugyanazokat az adatsorokat, és esetleg más következtetésekre jussanak velük kapcsolatban.[36]

Nem állíthatjuk ugyanakkor, hogy a digitális korpuszok segítségével talált párhuzamok – még ha számosabbak és reprezentatívabbak is – szükségképpen objektívebb képet is rajzolnának egy-egy adott szövegrész tág értelemben vett intertextuális kontextusáról. Fontos tudatosítani, hogy digitális keresésünk első lépése, a keresés tárgyát képező szó vagy kifejezés kiválasztása és ennek tükrében a keresőkifejezés összeállítása már eleve interpretatív aktus, és interpretatív jellegű a keresés eredményeképpen kapott találati lista elemzése is. A teljes találati lista csupán nyers adatokat tartalmaz, melyeket egyenként kell ellenőriznünk, értékelnünk és értelmeznünk. A lista tartalmazhat egyértelműen

[35] Vö. Crane–Bamman–Jones 2013: 53.

[36] E tekintetben a PHI-korpusz felhasználói felületének minimalizmusa kifejezetten hasznosnak mondható: minden egyes keresés, sőt minden megjeleníthető tartalmi egység saját URL-lel rendelkezik. Az LLT-adatbázisból egyedi formátumú fájlként letölthető egy-egy keresés összes paramétere.

téves találatokat: ilyenek lehetnek a homonímák, vagy a roszszul (esetleg a keresőrendszer korlátai miatt kényszerűen) választott keresőkifejezés következményei. A latin és az ógörög nyelv morfológiai jellemzői miatt ezeket a téves találatokat a digitális korpuszok jelenlegi verziói nem tudják automatikusan kiszűrni. A lista részét képezhetik továbbá a szó szoros értelmében véve nem téves, a kérdésfeltevésünk szempontjából mégis irrelevánsnak ítélhető találatok is. Kereshetjük például két szó együttes előfordulásait, csakhogy a latin és az ógörög viszonylag szabad szórendje miatt ez nem feltétlenül jelenti azt, hogy a keresett szavak közvetlenül egymás mellett fordulnak elő egy mondatban. Ki tudom gyűjteni azokat a mondatokat, melyekben mindkét szó szerepel, ám a találatokat egyenként megvizsgálva tudom csak eldönteni a kontextus, a két keresett szó egymástól való távolsága és szintaktikai viszonya tükrében, hogy csupán esetleges vagy pedig jelentésképző erővel bíró jelenségről van-e szó, vagyis az értelmezés szempontjából is beszélhetünk-e „együttes előfordulásukról".

A téves és az irrelevánsnak ítélt találatok kiszűrése után a lista gyakran még mindig viszonylag hosszú és heterogén marad: minél gyakoribb egyik vagy mindkét szó előfordulása a teljes korpuszban, annál inkább. Ilyen esetekben a találatok kategorizálása válik kulcsfontosságú értelmezői műveletté. Listánk tartalmazhat egyrészt olyan párhuzamokat is, melyeket a két szöveg között közvetlen kapcsolatot létrehozó idézetként vagy allúzióként azonosítunk: ezek az irodalmi intertextualitás minősített esetei. Egy-egy keresés lefuttatásakor a filológus valószínűleg abban reménykedik, hogy ilyeneket is felfedez majd – de persze éppen ezek esetében van viszonylag nagyobb esély arra is, hogy a korábbi szakirodalomban már dokumentált és interpretált párhuzamokról van szó, különösen, ha irodalomtörténeti szempontból kézenfekvő a két szöveg közti kapcsolat vizsgálata. E helyütt nem is a találatok ezen kategóriáját emelném ki, hanem azt a „szürke zónát", amely az egyértelműen téves/irreleváns találatok és az egyértelműen releváns, az irodalmi interpretációt döntő mértékben alakító idézetek/allúziók között helyezkedik el. Eb-

ben a szürke zónában a „valamilyen szempontból potenciálisan releváns" találatok helyezkednek el. A digitális eszközökkel végzett kutatás módszertanát illetően sok szempontból ezek lehetnek a legérdekesebbek, éspedig éppen azért, mert nehéz a skála végpontjait képező másik két kategória egyikében elhelyezni őket. Segítségükkel jól tesztelhetjük – és kényszerítenek is arra, hogy teszteljük – az intertextualitás természetéről alkotott implicit vagy explicit elképzeléseinket, elméleteinket.[37] Hol vannak az értelmezhető intertextualitás határai? Az irodalmi allúzióktól az irreleváns találatok felé elindulva vajon hol lesz az a határ az említett szürke zónában, ahol még éppen megfogalmazhatónak, illetve már éppen megfogalmazhatatlannak tartjuk, hogy a kereső által adott találatok külön-külön vagy együtt milyen módon befolyásolhatják a digitális eszközökkel vizsgált szöveg értelmezését irodalmi, nyelvi, kulturális szempontból?

A digitális korpuszoknak és keresőiknek az a fentiekben még másodlagosnak tekintett sajátossága, hogy a nem digitális módszerekhez képest több párhuzamos helyre hívhatják fel a kutató figyelmét, ebben az összefüggésben bizonyulhat mégis kifejezetten fontosnak. Vannak ugyanis olyan szövegpárhuzamok, melyek egyenként vizsgálva nem tűnnek különösebben relevánsnak – nem volna meggyőző állítás tehát, hogy a két szóban forgó szöveg között közvetlen és lényegi kapcsolat van –, ha viszont felismerjük, hogy a párhuzamként kínálkozó kifejezés ismétlődően fordul elő a korpuszban, akkor ezek a szöveghelyek együttesen mégis elárulhatják az elsődlegesen vizsgált szövegről, illetve annak kérdéses részletéről azt, hogy milyen műfaji, nyel-

[37] Ebből a szempontból különösen feltűnő és sajnálatos, hogy az ezredforduló környékén (tehát már a digitális szövegkorpuszok létrejöttét követően, de még jellemzően azok online elérhetővé válását megelőzően) megjelent, az antik irodalmi szövegek intertextualitását, illetve az intertextuális értelmezés klasszika-filológiai gyakorlatait általában vizsgáló monográfiák vagy egyáltalán nem tárgyalják a digitális szövegkorpuszokat (mint Hinds 1998), vagy csupán egy-egy mondatban tesznek említést róluk (így például Edmunds 2001: 22).

vi, kulturális regisztert szólaltat meg, milyen diskurzus(ok)hoz kapcsolható.

Az intertextualitás elméleteinek és gyakorlatának legradikálisabb, az egyes párhuzamok értelmezésétől legerőteljesebben elrugaszkodó tesztéléséhez a digitális klasszika-filológiában jelenleg a Tesserae projekt szolgáltat eszközt.[38] Speciális, a plágiumkereső rendszerek elvén működő keresőjével nem egyes szavak, kifejezések előfordulásaira kereshetünk rá, hanem bizonyos paraméterek beállítása után a keresőrendszer két kiválasztott szövegből gyűjti ki a nyelvi hasonlóságot mutató részleteket, s ezeket bonyolult algoritmus szerint (elsősorban a hasonlóságot megalapozó szavak gyakorisága alapján) súlyozza is.[39] A találati listákon szereplő egyezések száma zavarba ejtő: Ovidius *Metamorphoses*ének 779 soros első éneke és Vergilius *Aeneis*e (összesen mintegy tízezer sor) között az alapbeállításokat változatlanul hagyva nem kevesebb mint 6714 tételből álló találati listát kapunk, s e „párhuzamok" nagy részét minden bizonnyal nehéz lenne valamilyen irodalmi interpretáció során hasznosítani – gyakran az algoritmus által magas pontszámmal kiemelt egyezések is igencsak esetlegesnek tűnnek. A Tesserae célja végső soron nem is az, hogy egyes tételekkel gazdagítsa a dokumentált párhuzamok számát (bár sok idő és türelem birtokában éppenséggel erre is használható), hanem az, hogy a távoli olvasás eszközeivel mérhetővé és összehasonlíthatóvá tegye különböző szövegpárok „intertextuális távolságát",[40] illetve ábrázolhatóvá tegye adott esetben e távolság relatív változásait – tehát a fenti példánál maradva például azt, hogy a *Metamorphoses* első énekében az *Aeneis*szel mint az epikus hagyomány meghatározó szövegével hasonlóságot mutató sorok

[38] tesserae.caset.buffalo.edu. A Tesserae korpusza nagyrészt a Perseus (lásd fent) latin szövegein alapul, ám annál lényegesen kisebb. A rendszer részletesebb bemutatását lásd: Coffee et al. 2013.

[39] A Tesserae súlyozási algoritmusáról lásd Forstall et al. 2015.

[40] Erről részletesen lásd Gawley–Diddams 2017.

gyakorisága többé-kevésbé állandó-e, vagy pedig valamilyen értelmezhető tendencia szerint változik.[41]

8. Összegzés

A tanulmányomban részletesen tárgyalt „hagyományos", első generációs digitális szövegkorpuszok és a Tesserae intertextuális keresési lehetőségeit összehasonlítva kimondható, úgy vélem, hogy a klasszika-filológia immár mintegy fél évszázados „digitális történelme" jelenleg korszakhatárhoz ért. A TLG, a PHI és a hozzájuk hasonló korpuszok új eszközöket adtak a kutatók kezébe, s nem csupán könnyebbé, gyorsabbá tették munkájukat, hanem a kutatás módszertanára is hatást gyakoroltak – például azáltal, ahogyan a találati listák feldolgozásában a kvantitatív elemzési módszerek a korábbiaknál hangsúlyosabban egészítik ki a kvalitatív interpretációs stratégiákat. Az első generációs korpuszok és az általuk lehetővé tett keresések ugyanakkor, amint azt a fentiekben tárgyaltam, jól illeszkednek a klasszika-filológia hagyományos gyakorlatába: inkább továbbfejlesztik, illetve kiegészítik a korábbi praxisokat, mintsem radikális paradigmaváltást valósítanak meg. A digitális korpuszok megjelenése tehát a klasszika-filológia történetében döntő változás volt ugyan, mégis sok tekintetben evolúciós lépésként, az adott diszciplína saját történeti kontextusában a szerves fejlődés példájaként tekinthetünk rá.

Ezzel szemben a Tesserae és a hozzá hasonló, fejlesztés alatt álló második generációs digitális eszközök az intertextuális kutatás terén inkább forradalmi változásokat ígérnek, illetve ilyenekkel fenyegetnek – nézőpont kérdése. Ma még egyáltalán nem látszik, hogy az általuk lehetővé tett és javasolt, a távoli olvasás

[41] A projektben részt vevő filológusok legrészletesebb – a rendszer „kalibrálása" során készült – esettanulmánya (Coffee et al. 2012) a kora császárkori Lucanus *Pharsalia* című eposza első énekének, valamint Vergilius *Aeneis*ének intertextuális kapcsolatát vizsgálja.

stratégiáival rokonítható értelmezési módszerek hosszabb távon beilleszthetők lesznek-e a klasszika-filológia sztenderd, a kutatók által széles körben legitimként elismert praxisai közé. Abban azonban már most segíthetnek, hogy történeti kontextusba helyezzük a közelmúlt első generációs digitális eszközeit. Már látjuk nemcsak azok „analóg" előzményeit, hanem digitális utódjait is. A ma rendelkezésünkre álló informatikai infrastruktúra felől nézve a klasszika-filológia első generációs digitális korpuszai feltűnően egyszerű eszközök: egyrészt minél több (egy-egy korszakra kiterjedően lehetőleg az összes fennmaradt) latin és ógörög szöveget kell magukba foglalniuk, másrészt viszont informatikai szempontból jellemzően nagyon egyszerű, karakteralapú keresések végrehajtását teszik lehetővé az így létrejött korpuszon. Ha új eszközökként új praxisok kialakulásához vezettek is, a klasszika-filológusoknak a számítógép kezelésén túl nem kellett alapvetően új, szakmájuk hagyományaitól idegen készségeket elsajátítaniuk e korpuszok eredményes használatához. A Tesserae keresési paramétereinek megfelelő beállításához, illetve az eredmények értelmezéséhez már legalább alapjaiban meg kell érteniük azokat az algoritmusokat, melyek szerint a kereső a találatokat súlyozza, a rendszer működését bemutató tanulmányok olvasója pedig matematikai képletekkel is találkozik. A második generációs digitális eszközök tehát elődeiknél sokkal erőteljesebben kényszerítik ki az interdiszciplináris gondolkodást, más tudományok (ez esetben értelemszerűen a matematika és az informatika) szemléletmódjának alkalmazását. Akárhogyan alakuljon is hosszú távon a klasszika-filológia története a digitális korban, ez már önmagában is fontos hozzájárulás lehet ahhoz, hogy e diszciplína termékenyen vehessen részt a digitális bölcsészet formálásában, illetve fenntartsa párbeszédképességét más, korunk tudományképét erőteljesebben meghatározó tudományokkal.

Irodalom

Barthes, Roland 1997. *S/Z*. Budapest: Osiris Kiadó.

Bilansky, Alan 2017. Search, Reading, and the Rise of Database. *Digital Scholarship in the Humanities* 32 (3): 511–527.

Brunner, Theodore F. 1993. Classics and the Computer: The History of a Relationship. In Solomon, J. (ed.): *Accessing Antiquity: The Computerization of Classical Databases.* Tucson: University of Arizona Press. 10–33.

Coffee, Neil – Koenig, Jean-Pierre – Poornima, Shakthi – Ossewaarde, Roelant – Forstall, Christopher – Jacobson, Sarah 2012. Intertextuality in the Digital Age. *Transactions of the American Philological Association* 142 (2): 383–422.

Coffee, Neil – Koenig, Jean-Pierre – Poornima, Shakthi – Forstall, Christopher W. – Ossewaarde, Roelant – Jacobson, Sarah L. 2013. The Tesserae Project: Intertextual Analysis of Latin Poetry. *Literary and Linguistic Computing* 28 (2): 221–228.

Crane, Gregory – Rydberg-Cox, Jeffrey A. 2000. New Technology and New Roles: The Need for "Corpus Editors". *Proceedings of the Fifth ACM Conference on Digital Libraries*. New York: ACM Press. 252–253.

Crane, Gregory 2004. Classics and the Computer: An End of the History. In Schreibman, Susan – Siemens, Ray – Unsworth, John (edd.): *A Companion to Digital Humanities*. Malden: Wiley–Blackwell. 46–55. [elérhető online: http://www.digitalhumanities.org/companion]

Crane, Gregory – Bamman, David – Jones, Alison 2013. ePhilology: When the Books Talk to Their Readers. In Siemens, Ray – Schreibman, Susan (ed.): *A Companion to Digital Literary Studies*. Malden: Wiley, Blackwell. 29–64.

Déri Balázs – Kelemen Pál – Krupp József – Tamás Ábel (szerk.) 2011. *Metafilológia 1: Szöveg – variáns – kommentár*. Budapest: Ráció.

Edmunds, Lowell 2001. *Intertextuality and the Reading of Roman Poetry*. Baltimore, London: Johns Hopkins University Press.

Fischer, Franz 2017. Digital Corpora and Scholarly Editions of Latin Texts: Features and Requirements of Textual Criticism. *Speculum* 92 (S1): 265–287.

Forstall, Christopher – Coffee, Neil – Buck, Thomas – Roache, Katherine – Jacobson, Sarah 2015. Modeling the Scholars: Detecting Intertextuality Through Enhanced Word-level N-gram Matching. *Digital Scholarship in the Humanities* 30 (4): 503–515.

Gawley, James O. – Diddams, A. Caitlin 2017. Comparing the Intertextuality of Multiple Authors Using Tesserae: A New Technique for Normalization. *Digital Scholarship in the Humanities* 32 (2): 53–59.

Gellar-Goad, Ted H. M. 2016. Review: The Latin Library. *Society for Classical Studies Blog:* classicalstudies.org/scs-blog/ted-gellar-goad/review-latin-library.

Gibson, Roy K. – Kraus, Christina S. (eds.) 2002. *The Classical Commentary: Histories, Practices, Theory.* Leiden, Boston: Brill.

Gibson, Roy K. 2002. 'Cf. e. g.': A Typology of 'Parallels' and the Role of Commentaries on Latin Poetry. In Gibson – Kraus (eds.) 2002. 331–357.

Hayles, N. Katherine 2012. *How We Think: Digital Media and Contemporary Technogenesis*. Chicago: Chicago University Press.

Hinds, Stephen 1998. *Allusion and Intertext. Dynamics of Appropriation in Roman Poetry*. Cambridge: Cambridge University Press.

Jockers, Matthew L. 2013. *Macroanalysis: Digital Methods and Literary History*. Urbana: University of Illinois Press.

Kozák Dániel 2018. PHI Latin Texts (review). *RIDE. A Review Journal for Digital Editions and Resources* 8: ride.i-d-e.de/issues/issue-8/phi.

Kozák Dániel (megjelenés előtt). The Intertextual Frontiers of Vergil's „Empire Without Limit": Digital Comments on *Aeneid* 1. 278–279.

Kraus, Christina S. – Stray, Christopher (eds.) 2016. *Classical Commentaries: Explorations in a Scholarly Genre*. Oxford: Oxford University Press.

Labádi Gergely 2014. Franco Moretti: Distant Reading (recenzió). *Irodalomtörténet* 95 (4): 561–564.

Lang, Sarah 2018. Review of Perseus Digital Library. *RIDE. A Review Journal for Digital Editions and Resources* 8: ride.i-d-e.de/issues/issue-8/perseus.

Loar, Matthew 2017. Review: The Packard Humanities Institute (PHI)—Classical Latin Texts. *Society for Classical Studies Blog:* clas-

sicalstudies.org/scs-blog/matthew-loar/review-packard-humanities-institute-phi%E2%80%94classical-latin-texts.
Moretti, Franco 2005. *Graphs, Maps, Trees: Abstract Models for a Literary History*. London, New York: Verso.
Moretti, Franco 2013. *Distant Reading*. London, New York: Verso.
Most, Glenn W. (ed.) 1999. *Commentaries – Kommentare*. Göttingen: Vandenhoeck & Ruprecht.
Péter Róbert 2016. A Big Data kihívás és lehetőség a bölcsészettudományokban: digitális szövegek és metaadatok távoli olvasása. *Magyar Tudomány* 177 (11): 1323–1330.
Ramsay, Stephen 2011. *Reading Machines: Toward an Algorithmic Criticism*. Urbana: University of Illinois Press.
Richardson, John S. 2005. Indexing Roman Imperialism. *The Indexer* 24 (3): 138–140.
Richardson, John S. 2008. *The Language of Empire: Rome and the Idea of Empire from the Third Century BC to the Second Century AD*. Cambridge: Cambridge University Press.
Romanello, Matteo 2016. Exploring Citation Networks to Study Intertextuality in Classics. *Digital Humanities Quarterly* 10 (2). www.digitalhumanities.org/dhq/vol/10/2/000255/000255.html.
Sosnoski, James J. 1999. Hyper-Readers and Their Reading-Engines. In Hawisher, Gail E. – Selfe, Cynthia L. (eds.): *Passions, Politics, and 21st Century Technologies*. Urbana: Utah State University Press. 161–177.
Tarrant, Richard 2016. *Texts, Editors, and Readers: Methods and Problems in Latin Textual Criticism*. Cambridge: Cambridge University Press.