Simon Gábor

A számolás joga, avagy korpusznyelvészet és irodalomtudomány[1]

A talán sokak által ismert, Ridley Scott rendezte 2012-es science fiction mozifilm, a *Prometheus* egyik jelenetében David, az android a következő kérdéssel fordul dr. Holloway-hez: „Mondja, önök miért csináltak engem?" A rövid, de karakteres válasz így hangzik: „Mert meg tudtuk tenni." Miközben a számítógépes elemzéseket alkalmazó humántudományos kutatások minden korábbinál nagyobb léptékeket öltenek, ironikusnak hathat a legitimáció kérdéseivel foglalkozni, hiszen azok eredményei a digitálisan hozzáférhető és feldolgozott szöveganyagtól kezdve a szerzőazonosításon át a távoli olvasásig nyilvánvalónak tűnnek. Mégis gyakran jut eszembe ez a jelenet, amikor a számítógépes szövegelemzés lehetőségei kerülnek szóba, mert remekül illusztrálja a technikai tudás és a tudományos felfedezés ambivalens viszonyát. És bár Babitsot is idézhetnénk („ez a sok szépség mind mire való?"), mégiscsak stílszerűbb a mesterséges intelligencia humanoid formájának fiktív öndefiniálási kísérletéből kiindulni a dilemma felméréséhez: pusztán azért végzünk el méréseket, mert van ehhez alkalmas digitális eszközünk („meg tudjuk tenni"), vagy ezek valóban elvezetnek új kérdésekhez és válaszokhoz a poétikai kutatás területén?

Adott tehát egyfelől egy olyan informatikai technológia, amely az emberi megismerő kapacitásokat meghaladó megfigyelésekre is alkalmas. És adott másfelől a felfedezés vágya, olyan jelenséget meglátni, amely még nem tárult fel a maga teljességében. Szerencsés esetben ez a két tényező találkozik a tudományos megismerésben, ez lenne az, amit Michael Stubbs a számítógéppel támogatott kvantitatív szövegelemzés (computer-assisted quantitative analysis of literary texts) legerősebb fegyvertényének tekint: ekkor nem csupán olyan megfigyeléseket tehetünk, amelyeket saját megismerő apparátusunk nem tenne lehetővé, de magáról az olvasásról, a befogadói reakciókról is új ismereteket szerzünk. Szerencsétlen esetben pedig a triviális, haszontalan ténymegállapításokat gyarapítjuk: semmi olyat nem tudunk meg a számítógépes elemzésből, ami ne lenne feltárható szoros olvasással, vagy ahogyan Martin Paul Eve 2022-es monográfiája elején fogalmaz, „a *bálna* szó megszámlálása a *Moby-Dick*ben egyetlen tényről árulkodik csupán: milyen gyakran használják a *bálna* szót a *Moby-Dick*ben." Ez utóbbi eset pedig egybecseng Holloway válaszával:

1 A tanulmány elkészítését a Magyar Tudományos Akadémia Bolyai János Kutatási Ösztöndíja, valamint az Innovációs és Technológiai Minisztérium ÚNKP-21-5 Új Nemzeti Tehetségprogramja támogatta a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból.

pusztán azért, mert meg tudunk tenni valamit, nem szükségszerű, hogy az hasznára is válik a tudományos megismerésnek.

E tanulmányban arra teszek kísérletet, hogy a szerencsés és a szerencsétlen eset közötti széles tartományt feltérképezzem a korpusznyelvészet eszköztára felől, azaz rámutassak arra, miként gazdagíthatja a korpusznyelvészet az irodalomtudományi kutatást. Ez a kérdésfelvetés természetesen korántsem újszerű vagy eredeti; elég talán, ha ezen a ponton csak Sass Bálint ismeretterjesztő tanulmányaira utalok az elmúlt évekből, Dodé Réka, Ludányi Zsófia, Falyuna Nóra és Kuna Ágnes tanulmányára a 2018-as *Nyelv, poétika, kogníció* kötetben, a 2021-es *Líra, poétika, diskurzus* egyes tanulmányaira (Laczkó Krisztina és Tátrai Szilárd szerkesztésében), vagy akár a *Helikon Irodalom- és Kultúratudományi Szemle* 2020/1. lapszámára, amelyet a számítógépes irodalomtudománynak szenteltek. Ugyanakkor az efféle kérdések hasznossága éppen vissza-visszatérő aktualitásukban rejlik, mert arra késztetnek, hogy a megújuló technológiai tudás mellett folyamatosan reflektáljunk bölcsészeti kutatásaink céljaira és eredményeire. Maróthy Szilvia szerkesztői bevezetője is erről tanúskodik az említett Helikon-szám elején: „Az irodalomtudományokban régi ismerős a számítógép, ám mégis újra és újra szükséges bemutatkoznia." Úgy is fogalmazhatunk, hogy ez az ismétlődő bemutatkozás, a digitális módszerekkel való állandó számvetés nem mentegetőző fordulat, hanem a tudományos kutatásra irányuló (metatudományos) reflektálás nélkülözhetetlen művelete. A jelen tanulmány nem kínál átfogó képet minden számítógépes elemzési lehetőségről, csupán a korpusznyelvészettel foglalkozik, és annak is leginkább az alapfogalmaival, az adattípusaival és a szemléletmódjával kívánja megismertetni az érdeklődő olvasót. Korpuszt építeni ugyanis – kis túlzással állítható – ma már bárki tud (ha rendelkezésére állnak a jogtiszta digitális szövegváltozatok), éppen ezért ezen a területen válik különösen aktuálissá a „miért?", „mire való?" kérdése.

Korpusz, mintázat, gyakoriság

Általánosságban szólva a korpusz sok szöveg összegyűjtésével, elektronikus tárolásával és nyelvészeti elemzéshez történő előkészítésével áll elő – legalábbis a nyelvtudomány számára. Ezzel szorosan összefügg a mintavételezés fogalma, tehát a vizsgálni kívánt szövegek kiválogatása. Innen nézve a fő különbség a hagyományos szoros olvasáson alapuló irodalomtudományi megközelítés és a korpusznyelvészet között nem is annyira az „egy szöveg alaposan vagy sok szöveg felületesen" módszertani döntésében rejlik (noha a távoli olvasás látszólag éppen e döntés köré szerveződik), hanem sokkal inkább abban, hogy míg az előbbi számára az irodalmi szöveg önmagában vett esztétikai tárgy, addig az utóbbi a szövegre egy minta részeként tekint. Másként fogalmazva, egy irodalomtudományos elemzésnek nem magát a szöveget mint az elemzés alapegységét kell legitimálnia, hanem azt a kiindulópontot, amelyből az adott szöveget szóra bírja, és amelyből a műalkotás új kérdésekhez vezetheti el az elemzőt. Ezzel szemben a korpusznyelvészeti vizsgálat során a kutatás alapkérdése határozza meg, mely szövegek és miért válhatnak az elemzés alapját képező minta részévé. Jól láthatóan különböző a megismerés logikája: az első esetben a szöveg adott, s a cél az autentikus értelmezés; a második esetben a kutatási kérdés adott, s a cél olyan szövegkorpusz összeállítása, amelynek az elemzésével választ találhatunk rá. Ebből következően a korpusznyelvészeti kutatás következtetéseit, eredményeinek általánosíthatóságát is meghatározza, milyen vizsgálati korpuszon történt az elemzés.

Egy példával élve, érdekelheti az elemzőt az a kérdés, miként ábrázolja a fizikai környezetet József Attila költészete. Az elemzés egyik lehetséges útja olyan versek kiválogatása az életműből (legyenek azok ismert vagy kevésbé kanonizált darabok), amelyekben hangsúlyossá válik a versvilág fizikai környezetének leírása, majd egy elemzői kiindulópontból (legyen az a látás fenomenológiája, az észlelés nyelvi megformálása vagy éppen a biopoétika) sajátos, a szerzőre jellemző poétikai megoldások megfigyelése. Sőt, akár egyetlen alkotás részletező feltárása is újabb észrevételekkel gazdagíthatja a József Attila költészetéről zajló tudományos diskurzust. A korpusznyelvész számára ugyanakkor két egészen más kérdés adódik mindjárt a vizsgálat kezdetén: mi kerüljön be a vizsgálat korpuszába, és mit tudunk a szövegekben kereshetővé tenni. Az első kérdés komoly módszertani dilemma, amennyiben bármely József Attila-versben előfordulhat a fiktív versvilág valamilyen részletezettségű bemutatása, tehát vagy a teljes lírai életművet vesszük alapul (s még ekkor is kérdéses, mit kezdjünk például a rögtönzésekkel, be nem fejezett szövegekkel), vagy válogatunk abból valamilyen szempont szerint (például a cím utaljon a környezetre, szerepeljenek bizonyos kifejezések a szövegben és így tovább).

Ha a mintavételezés megtörtént, további problémát jelent a kutatni kívánt jelenség azonosíthatósága. A korpusznyelvészet számára ugyanis alapvetően kétféle tényező informatív: a gyakoriság (frekvencia) és az eloszlás (disztribúció). Következésképpen egy poétikai jelenséget megragadhatóvá kell tenni nyelvi szerkezetként, hogy az a korpuszban azonosíthatóvá, lekérdezhetővé, mérhetővé váljon. Az iménti példánál maradva, dönthetünk úgy, hogy a tájábrázolást bizonyos szavak megjelenésével próbáljuk megragadni (ilyenek lehetnek a természeti vagy városi tájra utaló főnevek), de alkalmazhatunk sokkal elvontabb térjelentésű kifejezéstípusokat is (mint amilyenek a névutós szerkezetek vagy helyhatározói esetraggal ellátott főnevek, sőt, a táj észlelésére utaló igealakok is szóba jöhetnek). Minél több szerkezettípust emelünk be, annál nagyobb az esélyünk arra, hogy azok valóban elvezetnek a vizsgálni kívánt poétikai jelenséghez, ugyanakkor annál differenciáltabb lesz az eredmény. Amennyiben ugyanis sikerül a számítógépes kereső számára is azonosítható típusokat kijelölni (azaz operacionalizálni tudjuk a jelenség vizsgálatát), már nem az egyes előfordulások lesznek lényegesek, hanem a belőlük formálódó mintázat: sűrűsödési és ritkulási pontjaik, illetve együttállásaik, társulásaik és a korpuszon belüli eloszlásuk. A korpusznyelvész a mintázatot elemzi különböző szempontok figyelembevételével, mint amilyen az életmű korszakolása vagy a mintázat egyszerűsége/összetettsége. A mintázat feltérképezése pedig elvezetheti a korpusz határain belül érvényes általános konklúziókhoz, például hogy az életmű bizonyos szakaszaiban gazdagabb vagy éppen szegényesebb a fizikai környezet leírása, illetve hogy egyes kulcsszavakkal vagy címkonstrukciókkal jellemzően együtt fordul elő az összetettebb tájábrázolás.

A gyakoriság tehát nem önmagában lesz fontos, ugyanis a mintázat megfigyelését segíti. Nem véletlen, hogy a korpusznyelvészet az összes előfordulás számát (abszolút gyakoriság, raw frequency) nem is tekinti igazán informatívnak: a gyakoriságot rendre viszonyítani szükséges valamihez (a korpusz egészének terjedelméhez, a kötet szószámához, vagy a versek átlagos hosszúságához), hogy megkaphassuk a relatív gyakoriságot, amely aztán már valóban jellemzi a mintázatot. A számolás így nem öncélú és önértékű: az adja meg a létjogosultságát, ha tudjuk, mit, miért és hogyan számolunk.

Ezen a ponton kitérhetünk egy másik gyakori kritikára is, miszerint a kvantitatív számítógépes elemzések körben forgó érveléshez vezetnek, amennyiben egy jelenséget azonosítunk valamilyen

nyelvi szerkezettel, majd annak korpuszbeli megfigyelését követően magáról a jelenségről fogalmazunk meg állításokat. A cirkularitás abban rejlene, hogy a jelenség korpuszbeli azonosítását az elemző végzi el, így amikor a méréseinek az eredményét interpretálja, nem magáról a jelenségről beszél, hanem a saját döntésének a következményeiről. Ez a feltételezés azonban csak akkor jogos, ha a jelenséget azonosítjuk magukkal a kereshető nyelvi szerkezetekkel, vagyis ha redukcionista módon szemléljük azt. A korpusznyelvészetnek két hagyományos mutatója is van az ilyen hibák elkerülésére: az egyik a pontosság (precision), a másik a fedés (recall). Míg az előbbi arról ad információt, milyen arányban találtunk számunkra megfelelő adatokat egy kereséssel, addig az utóbbi azt mutatja meg, hogy a minket érdeklő adatok közül mennyit sikerült megtalálnunk. Köznapi példával élve, ha gyomlálás közben minden lágy szárú, zöld növényt kihúzunk a földből, nagy eséllyel az összes gyomot eltávolítjuk (tehát jó lesz a keresésünk fedése), de velük együtt az értékes palánták is a gyomokkal együtt végezhetik (alacsony lesz a keresésünk pontossága). Ha viszont csak bizonyos formájú növényeket keresünk, kellően pontos lehet a megfigyelésünk (hiszen a paradicsompalánták mint téves találatok nem keveredtek a gyomok közé), ám könnyen lehet, hogy egyes, a palántákra hasonlító gyomok a földben maradnak (vagyis a fedés alacsony marad). Visszatérve a tájpoétikára, minden keresés után célszerű ellenőriznünk, milyen mértékű a mérésünk pontossága: míg a környezetre vonatkozó főnevek valószínűleg igen nagy arányban valóban a tájat írják le, addig a névutós, esetragos kifejezések sok esetben metaforikusan más jellegű viszonyokra (is) utalnak (például *tájban* és *márciusban*, vagy *a sínek között* és *a barátok között*), ezért az utóbbiak pontossága gyengébb. Az is belátható, hogy ezek kombinációja nagyobb mértékű fedést fog eredményezni (akár alacsonyabb pontosság mellett is), mint külön-külön történő alkalmazásuk. Mindebből pedig ismét az következik, hogy a számolás önmagában nem vezet el érvényes adatokhoz, ám a körültekintő és ellenőrzött számolás, valamint a keresés revíziója már igen. Ám úgy is fogalmazhatunk, hogy a számolás pusztán adatokhoz vezet, nem pedig evidenciákhoz: az elemző méréseket kipróbáló és kiértékelő tevékenysége avatja a puszta adatokat valamilyen kérdés megválaszolásához szükséges tényekké, a korpuszt pedig adatforrásból evidenciaforrássá.

Konkordancia, kulcsszó, kollokáció

Bármely korpuszelemző programmal, felülettel lehetőségünk van a minket érdeklő nyelvi jelenségek korpuszbeli előfordulásait listázni, mégpedig közvetlen szövegkörnyezetükben. (A korpusznyelvészet ezt nevezi KWIC (keyword in context) megjelenítésnek.) Az így előálló, a korpusz szövegeiben a keresett kifejezés(eke)t kiemelő adatsort konkordanciának nevezzük, amely arra alkalmas, hogy az elemző találatról találatra végignézhesse az egyes előfordulásokat, összevethesse azokat, vagy szempontok szerint csoportosítsa. Könnyű belátni, hogy többezres gyakoriságnál ez a kézi elemzés (hand and eye analysis) már nagyon sok időt és energiát igényel, így bár a szoros olvasáshoz ez áll a legközelebb, ebben az esetben támaszkodunk a legkevésbé a digitális korpuszelemzés eszköztárára. Mégis fontos lépése ez a kutatásnak, mert segít a lekérdezés pontosságának meghatározásában, és újabb keresésekhez vezethet. Továbbá olyan szöveghelyekre irányíthatja a figyelmet, amelyek a korpusz szöveganyagának átolvasása során akár rejtve is maradhatnak, esetleg átsiklunk felettük, gondoljunk a teljes lírai életműveket vagy nagyregényeket megcélzó vizsgálatokra. Végül arra is lehetőséget adnak a számítógépes eszközök, hogy bizonyos feltételek mentén

szűrjük a konkordanciát, azaz például csak azokat a találatokat tartsuk benn a keresési mintában, amelyeknél a lekérdezett kifejezés baloldali vagy jobboldali kontextusában szerepel/nem szerepel egy másik kifejezés, vagy valamilyen nyelvi kategória (például szófaj). Az egymásra épülő szűrésekkel egyre pontosabb képet kaphatunk a nyelvi megformálás tendenciáiról, így a konkordanciák mintázatok manuális azonosítását is lehetővé teszik. Ezeknél az elemzéseknél a korpusz voltaképpen előzetes intuícióink, olvasói benyomásaink alátámasztását és/vagy finomítását teszi lehetővé adatok forrásaként, ezért az ilyen vizsgálatokat korpusszal támogatott elemzéseknek (corpus-assisted analysis) nevezhetjük.

A konkordanciák mellett a korpusznyelvészet másik bevett, hagyományos adattípusát a gyakorisági listák adják. Az egyszerű szógyakorisági lista is sok információt adhat: megerősítheti például azt a benyomásunkat, hogy egy kötet/versciklus valamilyen központi téma köré szerveződik, ha a gyakorisági listán előkelő helyen állnak a témához kapcsolódó kifejezések. Vagy ha például Cormac McCarthy regényéről a *Véres délkörök*ről (*Blood Meridian*) olyan olvasói tapasztalatunk alakul ki, hogy a szöveg védjegye az erőszakábrázolás,[2] ezt alátámasztja a regény szógyakorisági listája: az első 150 leggyakoribb szó között szerepelnek a *fire, dead, pistol, blood* alakok, együtt a táj (*desert, sun, dark*) és a vadnyugati világ (*horse, ride, old, men*) jellemző kifejezéseivel.[3] A lista élén természetesen az angol nyelv grammatikai kifejezései (névmások, névelők, prepozíciók, a létige alakjai, tagadószó, segédigék) állnak, valamint nagyon általános igék (mint például a *said*, ami a MacCarthy-próza függő idéző technikáját is adatolhatóvá teszi), de a 100. elem körül sűrűsödnek az erőszakhoz kapcsolódó kifejezések. Következésképpen már a gyakorisági lista önmagában alkalmas arra, hogy benyomásainkat precíz, adatokkal alátámasztott megállapításokká formáljuk.

Vannak azonban a gyakoriságra épülő speciálisabb mérések is, melyek közül a kulcsszók azonosítására térek ki részletesebben. A korpusznyelvészet kulcsszónak tekinti egy korpusz azon kifejezéseit, amelyek nem csupán gyakoriak a korpuszban, hanem gyakoriságuk elsősorban a korpuszra jellemző sajátosság, összevetve egy másik szöveghalmazzal. A kulcsszóknak kulcsértékük (keyness) van, amely matematikai módon (leginkább a log likelihood és a log ratio függvényekkel) mérhető. Ha tehát egy kifejezés kulcsértéke magas (meghalad például egy előre meghatározott szignifikanciaszintet), az azt jelenti, hogy a kifejezés az adott korpuszban jellemzően gyakrabban fordul elő, mint az összevetés alapjául szolgáló úgynevezett referenciakorpuszban. McCarthy imént említett művéről azt állítja a moly.hu bejegyzése, hogy a „legpokolibb regénye". Ha kulcsszóelemzést végzünk a regényen, és referenciakorpuszként egy másik művét, ez esetben a *The Road* (*Az út*) című, posztapokaliptikus regényt választjuk (amely ugyancsak gazdag embertelen erőszakábrázolásban), a mérés egyrészt felmutatja, hogy a két regény igen eltérő történeti-kulturális környezetben játszódik (jellemző kulcsszók lesznek a *horse, desert, riders, mule, hat* kifejezések), ugyanakkor a gyilkosságok elkövetésének jellemző szókészlete, valamint az erőszakos jelenetek szereplői is helyet kapnak a szignifikáns kulcsszók között (*shot, rifle, captain, sergeant, blood, mexican, war, indian*). Ezeknek a szavaknak a korpuszbeli gyakorisága jellemzően magasabb a másik regényhez viszonyítva, és bár ez nem jelenti sem azt, hogy nem fordulhatnak elő a referenciakorpuszban,

2 Meggyőződhetünk erről, ha a moly.hu oldalán megnézzük a regényről bejegyzett hozzászólásokat.

3 A listát a szabadon letölthető LancsBox program 6.0-ás verziójával készítettem, elérhető a http://corpora.lancs.ac.uk/lancsbox/ url címen.

sem pedig azt, hogy a vizsgált regényben az abszolút gyakoriságuk szükségszerűen kiemelkedő lenne, ám a statisztikai elemzés rámutat arra, hogy ezek az adott mű szaliens (feltűnő) kifejezései. A kulcsszóelemzés már olyan kvantitatív technika, amely nem helyettesíthető manuális elemzéssel, miközben felmutatja a vizsgált szöveg(ek) sajátos lexikális mintázatait. Ezt a típusú kutatást, amelyben a korpuszelemzés a kezdeti intuícióinkat, hipotéziseinket alapos adatolással teszi ellenőrizhetővé, korpuszalapú elemzésnek (corpus-based analysis) nevezi a nyelvészeti szakirodalom, utalva a korpusz növekvő jelentőségére az előző metódusokhoz képest.

A gyakorisági mérések a listákon túl további érdekes adattípushoz is vezethetnek, ha nem csupán a szöveg szavainak önmagában vett előfordulásait tekintjük, hanem azt mérjük, milyen gyakran fordul el egy szó egy másik szó társaságában. Így jutunk el az együttes előfordulási mintázatokhoz, azaz a kollokációkhoz. A kollokáció kiemelt kifejezése a csomópont (node), amelynek a környezetében (ez az úgynevezett kollokációs ablak) a szoftver más kifejezések előfordulási gyakoriságát számolja. Ha ez utóbbi bizonyos kifejezések esetében magas érték, akkor a csomópont asszociálódik a korpuszban az adott kifejezésekkel, vagyis azok a kollokáltjai lesznek. Például a modern magyar elégiákból épített korpuszban (amely az 1850-es évektől tartalmazza egyfelől az elégiaként kanonizált alkotásokat, másfelől azokat a költeményeket, amelyek fellelhetők a Digitális Irodalmi Akadémia adatbázisában,[4] és amelyeknek a címében szerepel az *elégia* kifejezés) a *lát* igetővel erősen kollokálódik egyrészt az *én* személyes névmás, a *mit* vonatkozó névmás és a *hogy* kötőszó, másfelől viszont a *nem* tagadószó is.[5] Míg az első mintázat arra utal, hogy az elégiákban a látás folyamata erősen kötődik a lírai szubjektumhoz, és rendre valamilyen látással érzékelt jelenet is bemutatásra kerül, addig a második adat éppen azt húzza alá, hogy a látás hiánya, vagyis a vizuális érzékelés (átmeneti vagy tartós) megszűnése lesz a modern elégiák jellegzetessége. Ez a megfigyelés egybecsenghet azzal az olvasói tapasztalattal, hogy a modern elégiák gyakran tematizálják a világ megismerésének elbizonytalanodását és/vagy ellehetetlenülését, a temporális kontraszt mellé (vagy annak helyébe) az ismeretelméleti kételyt helyezve. (Fontos adalék lehet, hogy a tagadás a *néz* és a *hall* igetövekkel is kollokálódik, noha nem annyira erősen, mint a *lát* esetében.)

A kollokációk elemzésénél az első és legfontosabb döntés a csomópont kiválasztása, amely következhet előzetes benyomásainkból vagy tesztelni kívánt feltevésekből (esetleg a recepció korábbi megállapításaiból), de a szógyakorisági lista vagy a kulcsszók köre is bemenetként szolgálhat. Meg kell továbbá határozni az átvizsgált szövegkörnyezet (a kollokációs ablak) méretét, valamint azt a statisztikai függvényt, amellyel a kollokálódás, másként az együttes előfordulás erősségét a szoftver mérni fogja. (E tekintetben igen széles a korpusznyelvészeti paletta, de a részletektől ebben a tanulmányban megkímélem az olvasót, és egyszerűen a logDice-nak nevezett asszociációs mutatót ajánlom használni.) A kollokációs mérések tagadhatatlan előnye, hogy olyan mintázatokhoz is elvezethetik a kutatót, amelyek semmilyen módon nem jelentek meg az előzetes benyomásokban, mert az emberi szem nem regisztrálta azokat olvasás közben, mégis hatással lehetnek a szövegvilág szerveződésére. Ezek a vizsgálatok tehát akkor is eredményt hozhatnak, ha nem fogalmazunk meg feltevéseket, hipotéziseket, hanem mindjárt a korpuszhoz fordulunk. Ezért az ilyen típusú, korpuszvezérelt (corpus-driven) kutatásokban a legnagyobb a jelentősége a korpusznak, egyben a leginkább meghatározó a módszertani tudatosság és reflexivitás, vagyis az, hogy

4 https://pim.hu/hu/dia

5 A méréseket ezúttal is a LancsBox szoftverrel végeztem.

tudjuk, mit, hogyan és miért éppen úgy vizsgálunk. Értelemszerűen a leginkább induktív módja ez a számítógépes szövegelemzéseknek.

Zárszó

Ebben az áttekintő tanulmányban arra tettem kísérletet, hogy a számítógéppel végzett szövegelemzés legalapvetőbb adattípusait bemutassam röviden. Nem tértem ki sem az önálló módszercsaládként kibontakozó kutatási irányokra (mint amilyen a stilometria), sem a korpusznyelvészet másik lényeges területére, az annotálásra sem, hiszen annak módszertani elvei és megvalósítási lehetőségei önálló tanulmányt érdemelnének. Ez utóbbi témáról bőven tájékozódhat egyébiránt az olvasó a két korábban említett tanulmánykötet (l. *Nyelv, poétika, kogníció* 2018-ból, illetve *Líra, poétika, diskurzus* 2021-ből) tanulmányaiból, melyek többek között egy újonnan kutatásnak, a poétikai személyjelölés korpuszalapú vizsgálatának a részleteiről is beszámolnak.[6] (A vonatkozó tanulmányok mind a kutatás átfogó tervét, mind pedig a személyjelölés annotálásának módszertani részleteit bemutatják.)

A körkép mégoly leegyszerűsítő jellege mellett is azt remélem, sikerült némiképp oldanom a kvantitatív mérésekkel szembeni bizonytalanságot. Egyrészt annak a hangsúlyozásával, hogy miközben a digitális elemzési módszerek folyamatosan terjesztik ki a határaikat, a technológia önmagában nem üdvözít, és a megvalósíthatóság még nem garantálja (és végképp nem legitimálja) a hatékonyságot. Ugyanakkor a korpusznyelvészeti kutatásba számos olyan lépés iktatható be, amely a kapott eredmények reflektált értelmezésén keresztül elkerüli a technológiába vetett puszta bizalom csapdáját. Ráadásul a számolás nem csupán finomítható, de fejleszthető is, amely mindenképpen tágítja az irodalmi szövegekről folytatott diskurzus horizontját: precízebbé teszi azt, alátámaszt vagy árnyal megállapításokat, és új mintázatok felismeréséhez is elvezethet.

Éppen ez az utóbbi az, ami a távolsággal összefügg: a mintázatok felfedezéséhez el kell távolodnunk a konkrét szöveg(ek)től, ahogyan a kamera sem csupán ráközelít egy tárgyra, hanem távolítani is tud. Mindeközben azt is érdemes észrevennünk, hogy a számítógépes elemzések sok esetben igen kicsi egységeit mutatják fel az elemzésnek: szavakat, együttes előfordulásokat, kifejezéseket közvetlen szövegkörnyezetükben, csakhogy a távlat, ahonnan ezekre ráközelítünk, megmarad. Éppen ezért szükséges a szoros és a távoli olvasás dichotómiáján túllépnünk: ezek a folyamatok egymást feltételező és egymásba forduló elemzési műveletek, hiszen a hagyományos olvasás kelthet olyan benyomásokat, amelyeket távolító technikákkal vizsgálhatunk, miközben e technikák maguk vezethetnek el olyan szöveghelyekhez vagy mintázatokhoz, amelyeket szoros olvasással tovább elemezhetünk. Hagyományos és új módszerek egymást kiegészítő viszonyára hívja fel a figyelmet a már idézett Martin Paul Eve is, amikor azzal érvel, hogy a közelségnek és a távolságnak nem kellene ellentétbe kerülnie: a távolság is lehet mélység (azaz a mélyére hatolhat a nyelvi megalkotottságnak, ezzel együtt a poétikusságnak), mégpedig olyan mintázatok felmutatásával, amelyek az olvasó számára közvetlenül nem percipiálhatók, mégis akár a legelemibb

6 A kutatás az Eötvös Loránd Tudományegyetem DiAGram Funkcionális Nyelvészeti Kutatóközpontjában működő Stíluskutató csoportjának aktuális projektje, amely egy annotált magyar lírakorpusz kialakítását és elemzését célozza meg.

módon határozzák meg a befogadás élményét. Az úgynevezett számítógépes szoros olvasás (computational close reading) egyesíti a különböző olvasási módokat, és megvalósítja azt a módszertani sokféleséget, amely napjainkban mérvadó a humántudományos kutatásokban is. És amellyel vélhetően mind David, mind dr. Holloway kiegyezne.

Ajánlott irodalom

Eve, Martin Paul, *The Digital Humanities and Literary Studies*, Oxford, Oxford University Press, 2022.
Helikon Irodalom- és Kultúratudományi Szemle 2020/1. szám. Számítógépes irodalomtudomány.
Líra, poétika, diskurzus, szerk. Laczkó Krisztina, Tátrai Szilárd, Bp., ELTE Eötvös József Collegium, 2021.
Nyelv, poétika, kogníció. Elmélet és módszer a poétikai kutatásban, szerk. Domonkosi Ágnes, Simon Gábor, Eger, Líceum, 2018.
Sass Bálint, *Nyelvészeti szövegkeresők, nemzeti korpuszportál*, Magyar Tudomány, 2016/7, 798–808.
Stubbs, Michael, *Quantitative methods in literary linguistics = The Cambridge Handbook of Stylistics*, eds. Peter Stockwell, Sara Whiteley, Cambridge, Cambridge University Press, 2014, 46–62.