Szociológiai Szemle 32(2): 91–100.

Távolról is olvasható? A szövegbányászat perspektívája a szociológiai kutatásban[1]

Is it readable from a distance? The perspective of text mining in sociological research

Szigeti Ákos[2]

https://doi.org/10.51624/SzocSzemle.2022.2.5

Beérkezés: 2021. 12. 16.
Átdolgozott változat beérkezése: 2022. 07. 13.
Elfogadás: 2022. 07. 22.

„Neither the complications nor the opportunities necessitate outright rejection or unreflective acceptance" (Dobson 2019: 10)
„A komplikációk és a lehetőségek sem vezethetnek teljes elutasításhoz vagy reflexió nélküli elfogadáshoz." (Dobson 2019: 10)

A *Módszeresen* szociológiai metodológia előadás- és vitasorozat kilencedik alkalma a nagy mennyiségű szöveges adat gyűjtésére és automatizált elemzésére szolgáló szövegbányászatról szólt. A Társadalomtudományi Kutatóközpont és az Eötvös Loránd Tudományegyetem Társadalomtudományi Kar közös szervezésében megvalósuló online esemény bevezető előadását Németh Renáta (ELTE TáTK, RC2S2) tartotta. A bevezetőt követő kerekasztal-beszélgetés résztvevőjeként felszólalt Barna Ildikó (ELTE TáTK, RC2S2), Bodor Péter (ELTE TáTK), Géring Zsuzsanna (BGE), Kmetty Zoltán (TK CSS-Recens és ELTE TáTK), Ring Orsolya (TK PTI és TK CSS-RECENS), Sik Domonkos (ELTE TáTK) és Szigeti Ákos (NKE RDI). A beszélgetés moderátora Gárdos Judit (TK SZI) volt.

A szövegbányászat mint társadalomkutatási módszer

Bevezető előadásában Németh Renáta elkülönítette a szociológiában több évtizede használt tartalomelemzést a társadalomkutatás által nemrégiben felfedezett szö-

1 A beszámoló az Innovációs és Technológiai Minisztérium Kooperatív Doktori Program Doktori Hallgatói Ösztöndíj Programjának a Nemzeti Kutatási, Fejlesztési és Innovációs Alapból finanszírozott szakmai támogatásával készült.

2 Nemzeti Közszolgálati Egyetem, email: szigeti.akos@uni-nke.hu

vegbányászattól, mely az adattudomány, a mesterségesintelligencia-kutatás és a számítógépes nyelvészet területéről származik. A szövegbányászat, a természetes nyelvfeldolgozás (angolul natural language processing, NLP) és a szöveganalitika kifejezéseket bár sokszor szinonimaként használják, jelentésük mégis eltérő: a szövegbányászat során a strukturálatlan szöveges adatokat a természetesnyelv-feldolgozási algoritmusok teszik alkalmassá a különböző automatizált szöveganalitikai módszerek általi elemzésre (Németh et al. 2020). Az előadó a szövegbányászat mint társadalomtudományban alkalmazott módszer három jellemzőjét emelte ki: (1) az adattudományi jelleg miatti sajátos logikáját, melyre példa a meglévő elméletek új adatokon való vizsgálatát lehetővé tevő, a dokumentumok előzetes, kézi felcímkézésén alapuló felügyelt tanulás; (2) sajátos kutatási kérdéseit, így azt, hogy az üzleti szférából átvett, a véleményeket és érzelmeket detektáló szentiment- és emócióelemzés képes lehet például a gyűlöletbeszéd felismerésére; (3) sajátos módszereit, amelyek közül kiemelte a topikmodellezést, mely egy dokumentumgyűjtemény tipikus témáinak azonosítását tűzi ki célul, illetve a szóbeágyazást, mely a szavak jelentésének eltérését vagy hasonlóságát képes kimutatni a szavak vektortérbeli reprezentációjának segítségével.

Járatlan terepen a szövegbányászat segítségével

A kerekasztal-beszélgetés résztvevőinek többsége maga is használ szövegbányászati módszereket kutatásaiban, melyre elsősorban a nagy mennyiségű szöveges adat rendelkezésre állása ad okot és lehetőséget. Barna Ildikó antiszemitizmuskutatásaiban a téma legújabb kutatási terepét adó internetes platformokon elérhető hozzászólások, történeti kutatásaiban pedig a holokauszttúlélőkkel készült számos elemezhető interjú feldolgozása céljából adaptálta a szövegbányászatot módszertani eszköztárába. Géring Zsuzsanna az amúgy jellemzően kismintás kvalitatív módszerrel dolgozó diskurzuselemzés során, főként más módszerrel tovább vizsgálható diskurzusrészletek detektálására, kevert módszerű szövegelemzés egyik eszközeként használja a szöveganalitikát. Ring Orsolya szintén diskurzusok elemzésében, illetve a levéltárakban elérhető hatalmas mennyiségű dokumentum mintázatainak feltárásában alkalmazza a módszert. Sik Domonkos szerint a szövegbányászati módszerek segítségével olyan kutatási terep, „terra incognita" tárul fel, melyet másképpen nem tudnánk elérni és vizsgálni – így az ő egyik kutatási területén, a depresszió vizsgálatában is új lehetőségek nyílnak. Kmetty Zoltán szintén a kérdőíves módszerrel nehezen kutatható, internetes fórumokon annál inkább megmutatkozó depresszió, öngyilkosság és önsértés vizsgálata során alkalmaz szöveganalitikát. Szigeti Ákos doktori munkájában a darknetet vizsgálja, melynek felhasználói szintén nehezen érhetőek el közvetlenül, azonban az általuk generált nagy mennyiségű szöveges adat lehetőséget ad szövegbányászat segítségével való vizsgálatukra. Bodor Péter saját bevallása szerint kívülállóként figyeli a szövegbányászat társadalomtudományi al-

kalmazásának jelenlegi „kísérleti fázisát", mikor még számos területen kipróbálják a módszert. Várakozása szerint a szöveganalitika vizsgálati terepe a jövőben leszűkül, hiszen valójában a társas interakcióknak csupán egy szeglete reprezentálódik elemezhető beszéd vagy szöveg formátumban.

Exploratív vagy konfirmatív a szöveganalitika?

A szöveganalitikai módszerek bár a vállalati szférában jellemzően exploratív jellegű vizsgálatokban kerülnek alkalmazásra, a társadalomtudományban mind exploratív, mind konfirmatív kutatási célra alkalmazhatók. Sik Domonkos munkatársaival a depressziós fórumok hozzászólásainak vizsgálatában először exploratív céllal használta a szöveganalitikát, és egy viszonylag egyszerű elkülönítésre, a depresszió biológiai, pszichológiai, társadalmi keretezésének megkülönböztetésére próbálta megtanítani az algoritmust (Németh et al. 2022). Miután ezen a területen kevésbé jártak sikerrel, és az algoritmus nem tudta elég jól bejósolni az előre megadott kategóriákat, egy másik, szintén exploratív irányt próbáltak ki: tematikus csomópontokat kerestek topikmodellezéssel, melyet a kutató sikeresebbnek értékelt (Sik et al. 2021). Ezt követően, az utóbbi időben kezdtek el szóbeágyazással dolgozni, melynek segítségével azt a hipotézist tesztelik, hogy a depresszióról szóló diskurzusban a biomedikális megközelítés leuralja a többi megközelítést. Tehát utóbbi esetben konfirmatív céllal alkalmaznak szöveganalitikát. Ezzel szemben a szövegbányászat „bölcsője", az ipari alkalmazás jellemzően nem konfirmatív, sokkal inkább exploratív modellekben gondolkozik – tette hozzá Kmetty Zoltán. A társadalomtudomány szerepe, hogy konfirmatív kutatásokat is végezzen, hiszen a megfelelő kérdéseket a társadalomtudósok képesek feltenni. Ugyanakkor a szöveganalitika esetében nem mindig különül el egyértelműen az exploratív és a konfirmatív megközelítés.

A kutatói döntés szerepe a szövegbányászatban

A beszélgetés résztvevői a szövegbányászat számos olyan lépését identifikálták, melyek (akár szubjektív) kutatói döntést igényelnek, ugyanakkor ezek előfordulását természetesnek tartják, és nem hibaként tekintenek rájuk. A diskurzuselemzés működésének megértését segítendő, Géring Zsuzsanna a társadalomtudomány nyelvi fordulatára utalt vissza, melyet követően a szociológia a szöveget nem csupán lenyomatként, hanem egyfajta cselekvési térként értelmezi. Ebben a megközelítésben nagy jelentőséggel bír, hogy hogyan jelöljük ki a szövegkorpusz határait, amellyel dolgozunk. A kijelölés elsősorban a webes tartalmak esetében okoz nehézséget és igényel kutatói döntést, hiszen e szövegeknek nincs előre megadott eleje és vége, mint ahogyan a nyomtatásban megjelenő újságcikkeknek vagy könyveknek. Kmetty Zoltán hozzátette, hogy mivel a szövegbányászattal vizsgált digitális adatok – szemben például a kérdőíves adatokkal – nem kutatási célból készülnek, így tágabb teret

adnak az operacionalizációra – például a kérdőívek meglévő moduljaihoz, skáláihoz képest. Megerősítette, hogy fontos lépés a korpusz határainak definiálása az irreleváns elemek kiszűrése céljából, de az is kutatói döntésen múlik, hogy hogyan fogjuk meg a vizsgálandó fogalmakat. Konfirmatív megközelítés esetén az elképzelés operacionalizációja még kritikusabb. További példa kutatói döntésre a futtatások számának meghatározása topikmodellezés esetében, melyre ugyan vannak elfogadott megoldások, kánonszerű megoldás nincs. Hozzátette, hogy a digitális adatokkal való munkában a kutatói döntésnek nagyobb szerepe van, mint más adatok esetében. Sik Domonkos úgy látja, hogy a szövegek elemzése során egy olyan weberi megközelítésmóddal dolgozunk, ahol a szövegeket körkörösen vizsgáljuk az elmélet és az azonosítható ideáltípusok felől, egy olyan illesztési folyamatban, melynek abszolút értelemben nincs végpontja, a megértés idővel szaturálódik és áll össze koherens értelmezéssé. A folyamatot befolyásolja továbbá, hogy ezek a kutatások jellemzően csapatmunkára épülnek, ami ugyanakkor jó is, hiszen minél többféle perspektíva tud bekapcsolódni a kutatásba, annál izgalmasabbak lesznek az eredmények is.

A kutatások szubjektivitásának kérdése valójában a társadalomkutatás egészét érinti. Szigeti Ákos szerint nincs szubjektivitásmentes kutatás, folyamatosan kutatói döntéseket kell hoznunk a résztvevői megfigyeléstől a statisztikai elemzésig, beleértve a szövegbányászatot is. A kérdés inkább az, hogy kutatóként képesek és hajlandóak vagyunk-e ezeket a döntéseket transzparensen bemutatni. Parti Katalinnal közös tanulmányának eredményei szerint ehhez olyan közeget lenne szükséges megteremteni a tudományon belül, mely elfogadja a hibákat, bukásokat, és képes tanulni belőlük (Parti–Szigeti 2021). Interjúalanyaik szerint ehhez elsősorban az akadémiai kapuőröknek és a folyóiratok szerkesztőségében elhelyezkedőknek kellene másként viszonyulniuk a publikációkhoz, és nem csak az ún. „sikeres kutatásokat", tanulmányokat kellene publikálniuk. Rengeteget segíthet továbbá a nyílt tudomány gyakorlása, melyben történt előrelépés a járvány idején: több folyóirat ingyenesen elérhetővé tette tanulmányait, és az online konferenciák is szélesebb körben váltak elérhetővé. Mindemellett a tanulmányokban is jóval nagyobb hangsúlyt kellene fektetni a módszertan bemutatására, bár a szövegbányászatot alkalmazó kutatások esetében – vélhetően a módszer új jellege miatt – viszonylag részletes módszertani fejezetekkel találkozhatunk. Mindez szükséges feltétele az interoperabilitásnak, melynek köszönhetően az adott módszereket más kutatási területeken is adaptálni tudják. A szubjektivitás nem feltétlenül hibaforrás, hiszen nemcsak a kutatók, hanem maguk a kutatási alanyok is szubjektumok – egészítette ki az előbbieket Bodor Péter. A szubjektív szó ebben a tekintetben lecserélhető a perspektivikus kifejezésre: ha nem vesszük figyelembe a kutatás során a társadalmi cselekvők perspektíváját, akkor mechanizmusokká redukáljuk a társadalmi cselekvőket, akiknek nincsen saját perspektívájuk, szándékuk, nem aktív résztvevői a diskurzusnak, így rajtuk kívül álló erők hatásaként nyilvánulnak meg. Amennyiben a szubjektivitást nem azonosítjuk a hibával, hibázással, és nemcsak a kutatóét, hanem a kutatott szubjektivitását is figyelembe vesszük, akkor felül tudunk kerekedni az objektivitás-szubjektivitás problémakörön.

Társadalmi jelenségek magyarázata szövegbányászattal

A hozzászólók egyetértettek abban, hogy a szövegbányászat segítségével lehetséges bizonyos társadalmi jelenségek magyarázata, sőt előállhatnak olyan helyzetek, például természetes kísérletet lehetővé tevő szituációk, melyeket más módon nem is tudnánk vizsgálni. A szövegbányászat alkalmas társadalmi jelenségek magyarázatára, azonban ehhez Sik Domonkos szerint el kell rugaszkodnunk attól a klasszikus szociológiai paradigmától, mely szerint a társadalmi pozícióból magyarázzuk a sajátosságokat, az attitűdöket, a viselkedésmintázatokat és egyéb jellemzőket. A szociológia már említett, 20. században lezajlott nyelvi fordulata és egyrészt a Foucault nyomán kinőtt diskurzus-elemzési hagyomány, másrészt Luhmann perspektívája, aki amellett érvel, hogy a szociológiának nem az egyéni cselekvőkre kell fókuszálnia, hanem a kommunikációs rendszerekként elképzelt entitásokra, más jellegű magyarázó modelleket valószínűsít. Innentől kezdve nem egyéni sajátosságok társadalmi pozícióhoz való hozzárendelése a magyarázat modellje, hanem a diskurzusok és a kommunikációs rendszerek, folyamatok egymáshoz kapcsolódása, egymással való érintkezése, és így a szemantikai terek és határok létrejötte. Szigeti Ákos a darknet kutatásából hozott példát Sik Domonkos felvetésére: a darkneten jellemzően nem közvetlenül a személyeket vizsgálják, hanem például a hozzászólásaikat és az értékeléseiket a kriptomarketeken, melyeket különböző felhasználónevekkel írnak, azonban nem feltétlenül különböző személyek. A társadalmi jelenségek magyarázatára példát szolgáltatnak a kriptomarketekről szóló kutatások. Az illegális tevékenységeket is folytató darknetes piacokat a rendészeti szervek idővel leállítják, a bezárásokat követően pedig a kutatások szerint a felhasználók jelentős része átmigrál egy másik kriptomarketre. Szövegbányászat segítségével vizsgálható, hogy a kriptomarketekhez kapcsolódó darknetes fórumokon milyen hangulat uralkodik egy ilyen bezárás után, így segítve annak a megértését, hogy a leállás miért nem rettenti el a felhasználókat. Az eredmények szerint főleg az adatvesztéssel, anyagi kárral járó bezárás után nagyobb a felzúdulás és így idővel néhányan felhagynak tevékenységükkel, azonban összességében alig változik a forgalom. Mindezt más módszerrel, konkrétan bitcointranzakciók elemzésével is megerősítették (Szigeti 2022). Az ilyen típusú validáció, illetve a más módszerekkel való kombináció a szövegbányászatot alkalmazó kutatások esetében már csak a módszer relatív új jellege miatt is célszerű lehet. Sik Domonkos megerősítette, hogy igazán akkor fognak nagy magyarázó erővel bírni a szövegbányászatot alkalmazó kutatások, amikor mellé tudunk tenni egy párhuzamos, például történeti elemzést vagy más kutatási elemet. Ehhez kapcsolódva Barna Ildikó elmondta, hogy az NLP-t alkalmazó módszereket gyakran éri az a vád, hogy az eredményeik csupán korábbi tudást validálnak, azonban véleménye szerint ez egyrészt szükséges a módszer fejlődéséhez, másrészt a validáció értéke sem lebecsülendő. Kmetty Zoltán tisztázta, hogy a Szigeti Ákos által bemutatott példában az ún. természetes kísérleti dizájn valósult meg, melyre a szövegbányászat több esetben kiváló lehetőséget ad. A társadalomtudományi magyarázat eseté-

ben jellemzően valamilyen beavatkozás, visszacsatolás lehetőségére van szükség, a kísérlet is egyfajta beavatkozás, ahol ténylegesen kauzális kapcsolatok feltárására van lehetőség – tette hozzá Bodor Péter. Akár luhmanni posztstrukturalista, akár habermasi kritikai, akár más egyéb magyarázatról van szó, ezeknek mindig van valamilyen kimenete, beavatkozási síkja, mely lehet az önismeret növelése, a leleplezés vagy valamilyen társadalommérnöki ambíció, ahogy a példában az, hogy hol avatkozzunk közbe a darknetnek, hol fáj neki a legjobban. Ahogy egy másik kutatás esetében az, hogy hogyan mentsünk meg az öngyilkossággal kacérkodó emberek közül minél többet, a kockázatok azonosításának segítségével. Tehát a magyarázatok csupán egy darabig tartoznak a társadalomtudományi tanszékekre, azután kilépnek onnan, így valamilyen reális folyamatokká kívánatos őket fordítani.

Az értő emberi olvasás szerepe a szövegbányászatban

A felszólalók számos példával támasztották alá, hogy az értő olvasást az alapvetően automatizált módszerekre támaszkodó szövegbányászat során sem nélkülözhetjük, illetve azt, hogy miért van kiemelt szerepe a szociológusoknak a kutatási eredmények interpretációjában. Barna Ildikó szerint a szövegbányászatot alkalmazó kutatások több lépése is igényli az értő olvasást, ez a módszer sem úgy működik, hogy bedobjuk a szöveget egy dobozba, és magától kijön valami, ismernünk kell a szöveget, amelyet vizsgálunk. Barna Ildikót részben az inspirálta a szövegbányászat adaptálásában, hogy a kvantitatív és a kvalitatív módszerek kombinációjának nagyobb a magyarázó ereje, a szövegbányászatnak pedig van egy kvantitatív része, mely statisztikára épül, illetve van egy kvalitatív része is, mely pedig épp az értő olvasást nem nélkülözheti. Napjainkban a szociológia keresi a helyét, a társadalomtudósok ráugrottak a mesterségesintelligencia-alapú, így az NLP-t alkalmazó módszerekre is, ami azzal jár, hogy lehet „furcsa dolgokat" olvasni a területen. Lehetséges, hogy bizonyos társadalomtudományon kívüli kutatók ügyesebben használják ezeket a módszereket, azonban az interpretációhoz szükség van a társadalomtudományra, szükség van ránk, szociológusokra. Nagyon jó különböző diszciplínákkal együtt dolgozni, és az NLP-nek az egyik fantasztikus eleme, hogy kikerülhetetlenné teszi az interdiszciplinaritást, ugyanakkor az interpretációhoz szükség van a szociológiai tudásunkra.

Géring Zsuzsanna három pontban emelte ki az értő olvasás szerepét. Egyrészt azokban a kutatásokban, ahol nem kifejezetten egy konkrét elméletet vizsgálunk, a mi szerepünk, hogy detektáljuk a különböző versengő elméletek szövegekben való megjelenését. Az ilyen vizsgálatok során jellemzően nincsenek előre adott szótáraink, így vagy megalkotjuk őket, vagy a megtalált jelenségeket rendezzük kategóriákba. Példaként hozta kutatását, melyben a fenntarthatóság témáját vizsgálták üzleti iskolák képzési szövegeiben: az elmélet hét különböző fenntarthatósági dimenziót különböztet meg, azt vizsgálva, hogy a fenntarthatóság kifejezés milyen kontextusban jelenik meg, ebből négyet tudtak azonosítani (Király et al. 2021). Azonban

ehhez nem maguk építették ki előzetesen a létező összes angol szóból a szótárakat, hanem a találatokat olvasva értelmezték, hogy pontosan milyen dimenzióról van szó. Másrészt elengedhetetlen az értő olvasás a kevert módszerű alkalmazás esetén. Például amennyiben tovább elemzendő szövegeket detektálunk a szöveganalitika segítségével, ezeket valamilyen szempontrendszer szerint kódoljuk, a kódokat pedig később visszaültetjük az adatbázisba további felhasználás céljából és hozzákapcsoljuk a beszélőt (a példa esetében tehát az intézményi sajátosságokhoz). Egy másik kutatásban a jövővel kapcsolatos mondatokat olvasták végig, és arra jutottak, hogy ezek hat különböző cselekvési keretbe rendeződnek el – ennek megállapításához is szükség volt a mondatok értelemzésére és kutatói kódolására (Géring et al. 2022). Harmadrészt, amennyiben az elemzés során kiugró, nem odaillő elemek jelennek meg, az értő olvasást hívhatjuk segítségül, akár társadalomtudományi magyarázatra szorul, akár nyelvészeti okra (például metafora, irónia) vezethető vissza az adatokban felmerülő „furcsaság".

Az értő olvasásra jelentős szerep hárul a topikmodellezés eredményeinek interpretálásában, ahol az azonosított témákat a kutatónak kell elneveznie és értelmeznie – mondta el Ring Orsolya. De nem csupán a topikmodellezés esetében van jelentősége: egy konkrét, politikai szövegeket elemző kutatásukban, szóbeágyazáson alapuló saját szótárral végzett szentimentelemzésük során több ponton az értő olvasásra támaszkodtak (Üveges et al. 2022): (1) kiválasztottak néhány száz pozitív és negatív töltetű szót, melyeket szóbeágyazással vizsgáltak politikai korpuszon, majd az eredményként megjelenő szavakat (akár több tízezret is) kézzel válogatták át; (2) ezt szinonimaszótárral dúsították, melyet újra kézzel kellett átválogatniuk; (3) a kész szótárakkal elemezték egy újság korpuszát, melyből kirajzolódott egy trend, azonban ennek értékeléséhez szükség volt a konkrét cikkek elolvasására, hiszen így tudták ellenőrizni, hogy a szótár alapján negatívként vagy pozitívként értékelt szöveg valójában az-e, ezáltal azt, hogy a szótár megfelelően működik. Szintén a topikmodellezés példáját használva illusztrálta a reflektivitás megjelenését az ilyen – ahogy már korábban is mondta – „terra incognita" megismerését segítő kutatásokban Sik Domonkos. A topikmodellezés eredményeként kapott szócsoportok értékeléséhez szükséges vizsgálni a szavak előfordulásának kontextusát, az így megjelenő elméleti, történeti vagy egyéb hátteret. Ezt a feladatot a gép nem tudja elvégezni, emberi értelmezést igényel. Ugyanakkor a folyamatot kifejezetten izgalmassá teszi az előre adott sémák hiánya, egy saját, kutatói értelmezési keret felállításának sokféle lehetősége, mely a nagymintás kérdőíves kutatási adatok elemzésében rejlő lehetőségek sokféleségéhez hasonlítható. Bodor Péter tisztázta, hogy mind a szövegen kívüli kontextus (például a szövegek forrása, szekvenciája), mind a szövegen belüli kotextus (szóbeágyazás esetén például az adott szót keretező mondat, melyből adott esetben kiderülhet, hogy csupán viccről, iróniáról van szó) értékelésében fontos szerep hárul az értő olvasásra. Az elemzésbe bevont szavak (tokenek) kiválasztása kiemelten fontos lépés, így például a személyes névmások meghagyása vagy kiszűrése

az előfeldolgozás során, hiszen ezek adott esetben szükségesek lehetnek a beszédszituáció megértéséhez. Mikro-diskurzuselemzéssel kiderül, hogy van egy sor olyan nyelvi elem (akár szó, morféma, igeidő), mely az adott térbeli, időbeli, személyes és szövegen belül szituációba rögzíti a szöveget, melyről szó van.

A gépi olvasás sohasem lesz olyan szinten „értő", mint az emberi olvasás, így az eredmények értelmezésénél mindig egyet vissza kell lépnünk, és el kell olvasnunk a releváns szövegeket – ismertette álláspontját Kmetty Zoltán. Saját kutatási tapasztalata alapján az emberi kódolást alkalmazó projektekben a kódolóknak el kell mondani, hogy próbáljanak meg „a gép fejével gondolkodni", tehát próbáljanak úgy kódolni, ahogyan a gép kódolna. Hiszen kódolóként hiába látja valaki például azt, hogy egy adott cikkben egy hosszú mondat mit jelent az előző és a következő mondatok alapján, vagy hogy egy komment milyen másik kommentekre reagál, amennyiben a gép ezt nem ismerné fel, úgy nem szabad eszerint kódolni, mert nem fog működni a módszer. A jelenleg felfutó kontextuális szóbeágyazási modellek (például a Bidirectional Encoder Representations from Transformers, BERT) esetében már lehetőség van beépíteni a szöveg forrását a modellbe. Így például Kmetty Zoltán és munkatársai aktuális, oltástémájú projektjében képes címkézni a kommenteket aszerint, hogy azok oltásellenes vagy nem oltásellenes oldalon jelennek meg. Tehát lehetséges a gépi olvasást „okosítani" és még inkább „értővé" tenni, ugyanakkor mindig meglesznek a limitációi.

A kutató(csoport) szerepe az értő olvasásban

Az értő olvasás során befolyásoló tényező lehet, hogy az adott kutató mely tudományágból érkezik – tette hozzá a korábbiakhoz Géring Zsuzsanna, egy hallgatói kérdésre válaszolva. Az a multidiszciplináris projektek tapasztalata, hogy a különböző tudományágból érkező kutatók mást vesznek észre az eredményekben a tudáshátterük, illetve a tudományterületük szemléletmódja (például emberközpontú nézőpontjuk) alapján. Éppen ezért fontos és jó kutatói csoportban dolgozni az ilyen típusú projekteken, hiszen így tudnak ezek a különböző tudások érvényesülni. Azonban ilyen esetben jellemzően nem mindenki ért ugyanannyira a módszertanhoz, így szükség van olyan szereplőre, aki képes fordítani a különböző aktorok és így a különböző tudások között. Bodor Péter hozzáfűzte, hogy nemcsak az olvasások különböznek, hanem a szövegek is. Egy bevásárlócédula például egészen más olvasást igényel, mint egy mondat Krasznahorkaitól, tehát számít, hogy egy adott szöveg milyen közegben jön létre, és milyen célra szánják. Így szükség lehet differenciálni az elemzett szövegeket; ahogy Kmetty Zoltán példájában említette, lehetséges például a kommentek forrásának címkézése.

Fogalom- és jelentésváltozás vizsgálata szövegbányászattal

A fogalom- illetve jelentésváltozás története kitapintható-e szövegbányászat segítségével? E hallgatói kérdésre válaszolva Barna Ildikó elmondta, hogy vizsgálni tervezi

például a zsidó szó jelentésének változását, melyet egyelőre akadályoz az, hogy az erre alkalmas újságkorpuszok nem állnak rendelkezésünkre történeti távlatban. További nehezítő körülmény a szövegek leiratozásakor bekövetkező sztenderdizálás, melyre példa lehet a már korábban említett, holokausztról szóló interjúk leiratainak (jegyzőkönyveinek) esete, ahol az interjúalanyok és a leiratozók nyelvi jellemzői nem feltétlenül különíthetők el. Sik Domonkos megerősítette, hogy egy szövegbányászattal dolgozó kutatás során rengeteg kompromisszumot kell kötni annak mentén, hogy milyen szövegkorpuszok érhetőek el digitalizált formában. Azonban amennyiben elérhetőek hosszabb távra visszamenőleg a szöveges adatok, úgy a szövegbányászat segítségével lehetséges eszmetörténeti kutatásokat végezni, illetve a szavak jelentésének változását vizsgálni. Németh Renáta szerint a szavak jelentése kiválóan megadható szóbeágyazási modell segítségével, például az egér szó két jelentése azonosítható a szavakat reprezentáló vektortérben hozzájuk közel álló szavak alapján. E képességét kihasználva a szóbeágyazás tehát alkalmas a jelentéstörténeti vizsgálódásokra is. Kmetty Zoltán hozzátette, hogy nemzetközi szinten vannak izgalmas projektek ebben a témában, ilyen például a Google Ngram,[3] mely rengeteg könyv szövegét tartalmazza digitalizált formában, és így lehetővé teszi a különböző fogalmak időbeli változásának elemzését. Vizsgálták például az angol „gay" (meleg) szó jelentését az 1900-as évektől, amiből kiderült, hogy az hogyan ment át a „daft" (bolond), majd „cheerful" (vidám) jelentéséből a jelenlegi, „homosexual" (homoszexuális) jelentésébe (Hamilton et al. 2016). Az ilyen jellegű vizsgálatokhoz tehát nagyon jó korpuszok kellenek, ugyanakkor a módszer kiválóan alkalmas erre a célra is.

Irodalom

Dobson, J. E. (2019): *Critical Digital Humanities. The Search for a Methodology*. Urbana, Illinois: University of Illinois Press, 175.

Géring, Z. – Tamássy, R. – Király, G. – Rakovics, M. (2022): The portrayal of the future as legitimacy construction: discursive strategies in highly ranked business schools' external communication. *Higher Education,* 1–19. https://doi.org/10.1007/s10734-022-00865-1

Hamilton, W. L. – Leskovec, J. – Jurafsky, D. (2016): *Diachronic Word Embeddings Reveal Statistical Laws of Semantic Change*. https://arxiv.org/abs/1605.09096

Király, G. – Géring, Zs. – Csillag, S. – Rakovics, M. (2021): *Sustainability at business schools: fantasy or reality?* 6th Central European Higher Education Cooperation (CEHEC) Conference, online, 2021. április 22–23.

Németh, R. – Katona, E. – Kmetty, Z. (2020): Az automatizált szövegelemzés perspektívája a társadalomtudományokban. *Szociológiai Szemle,* 30(1): 44–62. https://szociologia.hu/dynamic/44_62_oldal.pdf

3 Google Books Ngram Viewer. https://books.google.com/ngrams Letöltve: 2021. december 15.

Németh, R. – Máté, F. – Katona, E. – Rakovics, M. – Sik, D. (2022): Bio, psycho, or social: supervised machine learning to classify discursive framing of depression in online health communities. *Quality & Quantity,* 1–23. https://doi.org/10.1007/s11135-021-01299-0

Parti, K. – Szigeti, Á. (2021): Innováció a szociológiában. A társadalomtudomány és az adattudomány metszetében elhelyezkedő, innovatív kutatási módszerekre irányuló kutatói attitűdök vizsgálata. *Socio.hu,* 11(1): 147–171. https://doi.org/10.18030/socio.hu.2021.1.147

Sik, D. – Németh, R. – Katona, E. (2021): Topic modelling online depression forums: beyond narratives of self-objectification and self-blaming. *Journal of Mental Health*. https://doi.org/10.1080/09638237.2021.1979493

Szigeti, Á. (2022): Szövegbányászat a dark neten: rendészettudományi alkalmazások. *Belügyi Szemle,* 70(4), 757–767. https://doi.org/10.38146/bsz.2022.4.7

Üveges, I. – Csányi, G. – Ring, O. – Orosz, T. (2022): Szövegaugmentálási módszerek összehasonlítása politikai szövegek szentimentanalízise során. In *Magyar Számítógépes Nyelvészeti Konferencia,* 18, 521–534.