Katona Eszter Rita[1]

Természetesnyelv-feldolgozás a korrupciókutatásban

Új adatforrások, új módszerek, új tartalmi kérdések[2]
A doktori értekezés tézisei

https://doi.org/10.18030/socio.hu.2023.3.76

1. Bevezetés

Témaválasztásom hátterében két indok áll, az egyik a tartalomra, a másik a módszertanra vonatkozik. Olyan problémát szerettem volna választani, aminek van társadalomtudományi tétje és hatása. Ami nemcsak a tudományos szférának szól, hanem lehetőséget ad arra, hogy a civil szféra és a hétköznapok is összekapcsolódjanak vele. Korábban is volt már lehetőségem a K-Monitorral közösen dolgozni, valamint az adatbázisaikat már korábban is elemeztem, és mindig is kimondottan fontosnak tartottam a munkájukat, a korrupció elleni küzdelmet. Másrészről, témaválasztásom idején a természetesnyelv-feldolgozás *(Natural Language Processing, NLP)* új, kevesek által alkalmazott módszer volt, a magyar szociológus közösség akkoriban ismerkedett vele.

Disszertációmban módszertani jellegű, a korrupciót a nyilvánosság különböző szintjein, és ennek megfelelően különböző adatforrásokon, a természetesnyelv-feldolgozás különböző módszereivel elemzem. Egyfelől a „média szintjén" vizsgálom a korrupció reprezentációját az online sajtó elemzésén keresztül, másfelől a „hivatalos szinten" a kormányzati kiadásokban megjelenő korrupciós kockázatot elemzem a közbeszerzési szerződéseken keresztül. A disszertációban bemutatott empirikus tanulmányok több tudományterülethez kapcsolódnak, melyeket a módszertan köt össze.

Disszertációm négy részből áll: először az alkalmazott módszereket mutattam be, majd a(z automatizált) szövegelemzés perspektíváit a korrupciókutatásban. Ezt követően két empirikus kutatáson keresztül hoz példát az értekezés a természetesnyelv-feldolgozás két típusának: a nem felügyelt és felügyelt módszerek alkalmazására.

A disszertációm jelentősége többek között abban áll, hogy nagy adatbázisok alapján tesz állítást arról, amit korábban csak feltételeztünk. Az értekezés általánosíthatósága, replikálhatósága és szkópja is más, mint az interjús vagy kvalitatív tartalomelemzéssel végzett kutatásoké, hiszen pontos, kvantitatív mérésekkel rendelkezünk, melyek jó kiegészítései, sőt akár helyettesítői a kvalitatív technikáknak.

Disszertációmat annak reményben írtam, hogy példát mutat a szöveg mint adat korrupciókutatásban való felhasználására. Vannak olyan problémák, kérdések, melyekre kvantitatív mérés hiányában korábban nem tudtunk (megfelelő) választ adni. Interjúkkal, kvalitatív és kérdőíves kutatásokkal vagy nem tudnánk ezeket a kérdéseket megválaszolni, vagy csak kis elemszámú megfigyelésből származhatna benyomásunk. Bár eddig

1 ELTE Társadalomtudományi Kar; ELTE RC2S2.

2 ELTE Szociológiai Doktori Iskola, 2023. Témavezetők: Németh Renáta és Fazekas Mihály.

is voltak szakértőkkel felvett interjúk, de nem voltak szisztematikus eredmények nagy adatbázisokon. Olyan módszereket alkalmaztam, melyek segítségével ezekhez a kérdésekhez közelebb kerülünk.

2. A módszertan háttere: természetesnyelv-feldolgozás

A fejezet célja megteremteni az empirikus elemzések hátterét, bevezetni az olvasót a „szöveg mint adat" megközelítésbe.[3] A disszertáció felépítése törekszik arra, hogy minden olvasó megtalálja azt, ami leginkább érdekli. A fejezet „ideális olvasója" az a kvantitatív szociológus, aki érdeklődik a természetesnyelv-feldolgozás iránt, de maga nem űzi azt. A fejezet azt tárgyalja, hogyan lesz a szövegből adat, és kitér az értekezésben használt adatforrások módszertani jellemzőire. Mivel az NLP elemzések oroszlánrészét a szövegek előfeldolgozása teszi ki, és mivel e lépések ismeretlenek lehetnek az „ideális olvasó" számára, ezeket is e fejezet ismerteti.

2.1. Az NLP módszerek rövid áttekintése

Az NLP módszerek alapja legtöbbször a gépi tanulás. A modellalkotás során megkülönböztetünk felügyelt *(supervised),* és nem-felügyelt *(unsupervised)* módszereket.

A felügyelt gépi tanulás esetében rendelkezünk olyan kategóriákkal, amiket előzetesen már ismerünk, és a cél az, hogy új elemeket tudjunk beilleszteni a már rendelkezésre álló kategóriarendszerbe. Például célunk lehet közbeszerzési pályázatok felhívásait „korrupciógyanús" és „nem korrupciógyanús" kategóriába sorolni. De az is elképzelhető, hogy a modellünk diszkriminációját, klasszifikációs „logikáját" szeretnénk mélyebben megérteni, tehát azt, hogy például milyen szövegrészek, megfogalmazások mentén kerül a dokumentum egyik vagy másik kategóriába. Ilyenkor tehát arra vagyunk kíváncsiak, hogy van-e a korrupciós csalásnak nyelvi leképeződése, milyen kifejezések, szóösszetételek valószínűsítik a „korrupciógyanús" kategóriába kerülést.

Felügyelt esetben rendelkezésünkre áll a modellillesztés során egy felcímkézett, tanuló adathalmaz. Ilyenkor a kutató célja az, hogy új (felcímkézetlen) elemeket is be tudjon illeszteni a kategóriarendszerbe. Tudjuk, hogy milyen struktúrába szeretnénk az adatainkat kódolni, és ezt tanítjuk meg az algoritmusnak. A felügyelt módszerekkel ellentétben a nem-felügyelt modellek esetén nem rendelkezünk semmilyen előzetes kategóriarendszerrel, nincsenek előzetes feltevéseink, ismereteink a szöveg külső jellemzőiről. Ilyenkor tanítóhalmazzal sem rendelkezünk, hanem a modellünkre bízzuk, hogy a szövegekben különböző statisztikai feltevések alapján valamilyen látens struktúrát (szinonimákat, névelemeket vagy éppen látens témákat) találjanak.

2.2.A hagyományos társadalomtudományi szövegelemzéstől az NLP-ig

A természetesnyelv-feldolgozás megjelenése nem hozott új társadalomtudományi paradigmát, hiszen a kvalitatív és a kvantitatív szövegelemzés korábban is jelen volt. Azonban az NLP egy egészen új elemzési eszköztárat igényel és új, komplexebb elemzések elvégzésére nyújt lehetőséget.

A szövegből kinyert, azonosított tartalom két csoportba sorolható: a manifeszt és a látens tartalom típusaiba. A kvalitatív tartalomelemzés egy, a szöveges adatok szubjektív értelmezésének és az adatokban rejlő, elsősorban látens tartalmak kutatására alkalmas módszer, mely empirikusan és módszertanilag ellenőrzött eszközöket alkalmaz. Segítségével a szövegeket kommunikációs kontextusukban elemezhetjük, bennük mintázatokat kereshetünk, és az adatokat klasszifikálhatjuk – anélkül, hogy szövegeinket kvantifikálnánk. A ko-

3 Ez a rész több, a kutatócsoporttól független szerző tanulmánya mellett részben a témavezetőmmel, Németh Renátával és Kmetty Zoltánnal közösen írt cikkre épít, ami a *Szociológiai Szemlében* jelent meg (Németh–Katona–Kmetty 2020).

rai kvantitatív tartalomelemzés ezzel szemben a szövegekben előforduló szavak vagy témák gyakoriságának vizsgálatára alkalmas. A szavak együttes előfordulása alapján elsősorban manifeszt – de ezen túlmenően bizonyos esetekben látens – tartalmakat kívánnak felderíteni előre meghatározott kategóriák szerint. A korai kvantitatív elemzések alapvető módszere, hogy „megszámolják" az egyes manifeszt szövegelemek (kvalitatív elemzés során azonosított „kódok" vagy konkrét kifejezések, esetleg metaadatok) előfordulási gyakoriságát, majd ezeket az információkat használják a mintázatok feltárására.

A két módszer alkalmazásának eredményeképpen előálló produktumok is különböznek. A kvantitatív eljárások statisztikai módszerekkel elemezhető, számszerűsített adatok létrehozására alkalmasak, a kvalitatív tartalomelemzés során leírások, tipológiák születnek, valamint szükségszerűen szubjektív, részletes beszámolók az adott szövegről, amelyek bemutatják a vizsgált jelenség jelentésének árnyalatait (Zhang–Wildemuth 2005). A disszertációmban bemutatott két esettanulmány is azt mutatja, hogy az NLP a két módszert igen gyakran ötvözi.

A kvantitatív szövegelemzés széleskörű elterjedéséhez, az automatizált szöveganalitika és az NLP megjelenéséhez két tényező, a számítógépek számítási kapacitásának jelentős növekedésével párhuzamosan a digitalizált szövegek elérhetővé válása járult hozzá (Miner et al. 2012). A legelső statisztikai eljárások, melyek lehetőséget teremtettek a nagy mennyiségű szöveges információ feldolgozására, a 80-as, 90-es évek fordulóján jelentek meg. Az áttörést azonban a 2000-es évek hozták, amikor az automatizált szöveganalitika piaci alkalmazása megkezdődött. E módszerek alkalmazása először az üzleti szférában jelent meg, de gyorsan elérte a humántudományokat is. Az irodalomtudományban például már 2000-ben megjelent a „távoli olvasás" *(distant reading)* fogalma (Moretti 2000). A társadalomtudományokra alkalmazva Moretti (2000, 2013) terminológiájával azt mondhatnánk, hogy a hagyományos módszerek (sokszor még a kvantitatív szövegelemzés is) a „szoros olvasásra", *(close reading)* épít: egy dokumentumgyűjteményből (korpuszból) származó mintát vizsgál. A természetesnyelv-feldolgozás pedig a távoli olvasást alkalmazza egy digitális, teljes és nagy méretű korpuszt vizsgálva. A távoli olvasás során nem a konkrét, egyedi szövegekre fókuszálunk, hanem ezektől eltávolodva mintázatokat és trendeket figyelünk meg.

2.3. A természetesnyelv-feldolgozás relevanciája a társadalomkutatásokban

Egy fontos érv az NLP társadalomtudományi relevanciája mellett az, hogy legtöbbször beavatkozásmentes vizsgálatokat végezhetünk olyan jellegű információk elemzésére, melyekre máskülönben csak kérdőívek, interjúk révén kaptunk volna választ. Utóbbi technikák nem beavatkozásmentes vizsgálatok, tehát a vizsgálódásunkkal hatással lehetünk eredményeinkre is. Például, ha a korrupció elterjedtségének megítélésével kapcsolatban végzünk kutatást, hagyományos empirikus eszközöket használva megkérdezhetjük az állampolgárok véleményét, attitűdjét (ám ilyenkor maga a tény, hogy kutatást végzünk, kérdéseket teszünk fel, hatással lehet azokra a válaszokra, amit alanyaink megfogalmaznak) vagy az újfajta adatelemzési módszerek segítségével elemezhetünk természetes megnyilvánulásokat, mint például kommenteket a közösségi médiában (ilyenkor „talált" adatokkal dolgozunk, melyek nem kutatási céllal jöttek létre, így maga a kutatás nem befolyásolja a véleményeket). Módszertani szempontból érdemes lehet hangsúlyozni, hogy a két helyzetben azonban eltér az elemzési egység. A kérdőívekkel felvett kutatásban az elemzési egység az egyén, a kommentek elemzésénél viszont maga a komment jelenti az elemzési egységet, nem pedig az egyén. Ez azért lehet problémás, mert egy egyén több kommentet is írhat, így azoknak, akik gyakrabban kommentelnek, felerősödik a hangjuk, felülreprezentálttá válik a véleményük.

Az NLP alkalmazhatósága mellett szól, hogy a társadalomtudományi kutatások során sokszor dolgozunk nehezen operacionalizálható fogalmakkal vagy olyan szenzitív témákkal, melyek nehezen hozzáférhetők – ezek vizsgálatát is megkönnyíti a természetesnyelv-feldolgozás. Nem beszélve arról, hogy egyre nehezebb klasszikus surveykutatások során adatot gyűjteni az egyre növekvő válaszmegtagadás következtében (Messing–Szeitl–Ságvári 2022).

3. Az NLP helye a korrupciókutatásban[4]

3.1. A kutatás előzményei, problémafelvetés

A korrupció mérése több megközelítést követhet. A *Médiakutatóban* megjelent tanulmányban, melyet Németh Renáta és Kmetty Zoltán társszerzőimmel készítettünk (Katona–Németh–Kmetty 2021), bemutattunk egy lehetséges megoldást, valamint a *Socio.hu Társadalomtudományi Szemlében* megjelent, Németh Renátával közösen írt tanulmányunkban (Katona–Németh 2021) pedig bemutattunk egy másik lehetséges megoldást arra, hogy a korrupcióval kapcsolatos vizsgálatokat milyen szempontok szerint különíthetjük el egymástól.

Egyrészt:

1. vizsgálhatjuk a korrupciós észlelések jellegét (tapasztalati vagy perceptuális),
2. a szintjeit (egyéni vagy társadalmi),
3. vagy a távolságát az egyéntől (hétköznapi vagy politikai).

Ezek a kutatások leggyakrabban kérdőíves megkérdezésen alapulnak, tehát nem beavatkozásmentes vizsgálatok. Bár az anonimitás garantált, de maga a kutatás ténye is befolyásolhatja az eredményeket. A téma igen szenzitív, és ennek köszönhetően magas a nemválaszolási arány, emellett kétségbe vonható a válaszok megbízhatósága is (Axelsson–Dahlberg 2018, Fazekas–Tóth–King 2016). Emellett problematikus, hogy a lakosság igen kis része rendelkezik (főként a politikai) korrupció kapcsán közvetlen tapasztalattal (Fazekas–Tóth–King 2016).

Ám a korrupció mérése során használhatók beavatkozásmentes vizsgálatok is, melyek meglévő adatokon alapulnak, jellemzően a korrupciós kockázat elemzésére, a korrupciógyanús esetek azonosítására alkalmazhatók. A beavatkozásmentes vizsgálatokban a hagyományos kvantitatív és kvalitatív módszerek mellett használhatóak az újabb típusú módszerek is, mint a hálózatelemzés vagy akár a természetesnyelv-feldolgozás. Míg a nem beavatkozásmentes módszerek alapvetően a korrupció volumenét és elterjedtségét mérik, de nem adnak képet arról, hogy milyen korrupciós tematikák mentén alakul a közbeszéd, a kvantitatív sajtóelemzések éppen ilyen kérdésekre tudnak választ adni. Hajdu és munkatársai (2018a, 2018b) tartalomelemzéssel vizsgálták a korrupció médiareprezentációját a 2004 és 2013 közötti időszakban, több ország összevetésében. Azért is különösen fontos a téma médiareprezentációjának vizsgálata, mert a korrupció percepcióját nagyban befolyásolja a média helyzete, a sajtó szabadsága (Suphachalasai 2005). A korrupcióról szóló híradások hatására ugyanis csökkenhet a politikai korrupció mértéke (Németh–Körmendi–Kiss 2011), de ezzel együtt növekedhet a korrupció társadalmi percepciója.

A fenti kutatásokban közös, hogy főként szógyakoriság-elemzéseket használnak, esetleg kézi kódolást, szakértői besorolást. Az első esettanulmányomban magam is a korrupció médiareprezentációjával foglalkoz-

4 A fejezetben többek között a Németh Renáta társszerzőségével, a *Socio.hu Társadalomtudományi Szemlében* publikált elemzést is felhasználom (Katona–Németh 2021).

tam, azonban az automatikus szöveganalitika segítségével. Bár az NLP-t sokszor „csupán" információkinyerésre *(information retrieval)* használják, például nehezen kezelhető pdf-fájlok elemezhetővé tételére vagy olyan adatok gyűjtéséhez, amik szükségesek az indikátorok elkészítéséhez, a szövegbányászati megoldások társadalomtudományi elterjedésével párhuzamosan megjelentek az első olyan nemzetközi cikkek is, amelyekben a természetesnyelv-feldolgozás nem kizárólag adatgyűjtési, hanem elemzési eszköz is.

A fejezet célja a már létező szakirodalom terjedelmének meghatározása, a meglévő irodalom felkutatása, ezáltal a téma átfogó áttekintése, ami lehetőséget ad arra, hogy rámutasson a meglévő szakirodalom hiányosságaira is, valamint hogy azonosítsa a még nem kutatott területeket.

3.2. Adatok és módszerek

Az elérhető irodalom áttekintése során célunk a 2010 után született angol nyelvű tanulmányok összegyűjtése volt, melyek szövegelemzési módszereket alkalmazvaa korrupciót vizsgálják. Az áttekintés módszeréül a *scoping review*-t választottuk, hiszen a módszer lényege, hogy pontosan nem definiálható témában keres irodalmat (Arksey–O'Malley 2005). Az adatgyűjtés során a *Google Scholaron* kerestünk olyan cikkeket, melyek címében szerepel a *corruption* kifejezés, az absztraktjában valamilyen szövegelemzésre utaló kifejezés. Így összesen 131 cikket gyűjtöttünk, melyekből 65 cikk volt releváns.[5]

A cikkek kisebb része (28) kvalitatív módon nyúl a szövegekhez. Ezek a cikkek főképp diskurzuselemzést használnak, melyek esetében az automatizálhatóság nem feltétlen van jelen. A kvantitatív módszereket alkalmazó cikkek közül négy olyan tanulmányt találtam, ahol az algoritmizálhatóság egyértelműen megjelenik, ezekről a disszertációban írtam részletesen, a többi nagyon specifikus, egy-egy esetre fókuszáló esettanulmány.

3.3. Eredmények

A releváns cikkek elemzésével lényeges eltéréseket találtunk a felhasznált szöveges adatforrást, a korrupciómérési módot és az elemzési megközelítést tekintve, ugyanakkor kevés (adatforrását, módszerét vagy a korrupciómérés módját tekintve) kevert típusú tanulmányt találtunk. Legyen a módszer automatikus vagy sem, a legtöbb szövegelemzés médiareprezentáció-elemzés. A klasszikus, a korrupció volumenét vagy a vele kapcsolatos attitűdöt, percepciót leíró, illetve észlelésének következményeit vizsgáló (Muço 2019) munkákon kívül találtunk a korrupció megelőzésére felhasználható eredményeket (Fazekas–Tóth 2016; Pan–Chen 2018), sőt intervencióra közvetlenül alkalmasakat is (lásd Li és munkatársai (2019) surveillance-rendszerét).

Az IMF *„news-flow index"*-éhez összeállított szöveges adatbázis értékes gyűjtemény, mely számos elemzési lehetőséget rejt magában. Mivel hosszú időt ölel fel (1995–2017), így a korrupció kifejezés változásának, a korrupció tematizáltságának dinamikája is vizsgálható lenne rajta. Mivel 30 ország hírforrásait tartalmazza, így nemzetközi összehasonlítás alapjául is szolgálhat: feltárhatók lennének a korrupció szó használatának, jelentésének kulturális aspektusai és az országok közötti különbségek is. Egy másik észrevételünk, hogy bár a tanulmány mintát használ, nem egy teljes populációt, ennek ellenére a szerzők gyakran nem fogalmazzák meg expliciten, hogy a cikkek összegyűjtéséhez kiválasztott hírportálok mit kívánnak reprezentálni, nem tudjuk, hogy mire általánosíthatók az eredmények.

5 Nem tekintettük relevánsnak azokat a cikkeket, amik nem voltak hozzáférhetőek (4), ahol ismétlődtek (VAGY: többször szerepeltek?) a gyűjtésben (3), vagy amik témájukban nem feleltek meg a kritériumainknak (59).

Kézenfekvő NLP-alapú kiegészítés kínálkozik még Fazekas és Tóth (2016) kutatásaira is, ezt a disszertáció második esettanulmányában mutattam be.

Másik kritikai észrevételünk a nagy társadalomtudományos potenciállal bíró felügyelt klasszifikációs algoritmusokat alkalmazó kutatások (például: Noerlina et al. 2017) annotálására vonatkozott. Nem írnak róla részletesen, hogy kik (szakértők? laikusok?), milyen korrupciódefiníciós instrukciókat követve és milyen szervezésben (egyetlen annotátor? kettős független annotálás?) annotáltak. Pedig a társadalomkutatási alkalmazásoknál éppen ezeknek a döntéseknek van kiemelt szerepe, hiszen a besorolás itt az üzleti és műszaki alkalmazásoknál sokkal komplexebb interpretációs feladatot jelent.

A *scoping review* megmutatta, hogy egy adott tudományterületen hogyan alkalmaznak különböző megközelítéseket (kvalitatív, kvantitatív és automatizált szövegelemzést) az elemzők. Felvillantotta, hogy a különböző módszerek mennyiben eltérő kutatási kérdésekre tudnak választ adni, valamint, hogy milyen fehér foltok azonosíthatók az automatizált szöveganalitika felhasználása terén.

4. A korrupció reprezentációja a médiában[6]

4.1. A kutatás előzményei, problémafelvetés

Lambsdorff (2007) több tanulmány bemutatásán keresztül érvelt amellett, hogy a sajtószabadság segít csökkenteni a korrupciót: ha a sajtó nagy része állami tulajdonban van, a korrupció is nagyobb valószínűséggel van jelen.

Persze, összességében az, ha a médiában sokat írnak a korrupcióról nem elég. Park (2012) megállapította, hogy hiába jelenik meg a korrupciós botrányokról sok cikk a sajtóban, amennyiben az érdemi vita elmarad, az a korrupció elbagatellizálódásán át, politikai tétlenséghez vezet: *„A botrányok nem elsősorban felháborodást szülnek, hanem csömört”* (Martin 2019:17). Ez a folyamat érhető tetten Magyarország kapcsán is. Martin (2019) megállapítása szerint a „korrupcióvakság” összefügg a magyar média polarizáltságával: a Fidesz-kormány által kiépített propagandamédiumokkal szemben álló portálok szerinte egyértelműen kormánykritikus üzeneteket közvetítenek.

A fentiek is jól mutatják, hogy fontos lenne lépéseket tenni az információhoz való hozzáférés széleskörű biztosítása és a média függetlensége felé.

A fejezetben a magyarországi online sajtó médiareprezentációja kerül górcső alá, amit többek között a kormánypárti és nem kormánypárti média tükrében is vizsgáltunk. A 2007–2018 közötti időszakra vonatkozóan azonosítottuk a cikkek főbb témáit, valamint hogy hogyan kapcsolódnak egymáshoz ezek a témák (K1, K2) és a tematikus változás dinamikáját: az egyes korrupciós témacsoportok előtérbe kerülését és háttérbe szorulását, illetve az egyes témák tartalmi változását (K3, K4). A dinamikai változások kapcsán külön vizsgáltuk, hogy a 2014-es választási kampány időszakában a kampányon kívüli időszakhoz képest hogyan változott a hazai sajtó korrupciós tematikája (K5). Az időbeli dinamika vizsgálata mellett jól megragadható dimenzió

6 A negyedik fejezet alapját pedig a *Médiakutató* folyóiratban olvasható, Kmetty Zoltán, valamint Németh Renáta társszerzőségében megjelent tanulmány adja (Katona–Kmetty–Németh 2021). A disszertációmhoz kapcsolódó online mellékletben (https://eszterkatona.github.io/dtm_viz/index.html) elérhető a fejezetben bemutatott vizualizációk interaktív verziója, valamint esetenként kiegészítő ábrák is. Az interpretációt sokszor nagyban támogatja az interaktív adatvizualizáció, így ezek nem mellékes illusztrációként jelennek meg, hanem a mélyebb megértéshez szükséges fontos eszközként szolgálnak, amellyel az olvasó is végezhet saját mélyfúrásokat.

az online újságok politikaipárt-közelsége is. Külön elemeztük, hogy a kormánypárti és a nem kormánypárti médiában mennyire térnek el egymástól a vizsgált korrupciós tematikák (K6). Az adatbázis lehetőséget nyújt egy érdekes természetes kísérlet kiértékelésére is (K7), amely azt vizsgálja, hogyan változtatja meg egy lap korrupcióval kapcsolatos diskurzusát a tulajdonosváltás. Elemzésünkben az *Origo* korrupcióval kapcsolatos cikkeit vizsgáltuk ebből a szempontból.

4.2. Adatok és módszerek

Az elemzésben a K-Monitor által azonosított, korrupcióval kapcsolatos újságcikkeket használtuk. A K-Monitor az „Akták" nevű adatbázisában[7] gyűjt össze minden olyan linket, ami Magyarországhoz köthető korrupciós ügyekkel kapcsolatban megjelenő cikkre mutat a magyar sajtóban. A K-Monitortól kapott adatbázis 33 557 URL-t tartalmazott, az adatok legyűjtését, a duplikátumok eltávolítását és a tisztítást követően 26 262 cikk maradt meg. A Python Newspaper[8] csomagját használtuk a cikkek letöltéséhez. Az elemzés során felhasznált cikkek forrása nagyon változatos, összesen 152 portálról származnak azonosított tartalmak. A portálok döntő többsége híroldal, de bulvároldalak és blogok is megjelennek közöttük.

Az adatokra úgy tekintünk, mint populációra, az esettanulmányban megfogalmazott következtetéseimben a korrupció médiareprezentációját a K-Monitor cikkgyűjteményén tárgyaltam. Az elemzett adatbázis 2007-től, az adatbázis építésének kezdetétől indul, és 2018 augusztusában zárul, tehát a 2018-as cikkek gyűjteménye nem teljeskörű. A cikkek évenkénti megoszlását az 1. ábra mutatja. Ezen látható, hogy 2011-ig folyamatosan nőtt az adatbázisban a cikkek száma, míg 2012-ben, 2013-ban és 2016-ban egy nagyobb visszaesés tapasztalható. Az ábrán szintén jól látható, hogy a korábbi évekből arányaiban kevesebb cikket tudtunk elérni és letölteni. Ebben az esetben a hiányzó adatok úgy értelmezendőek, mintha egy teljes populációt vizsgáló kérdőíves kutatás esetén lennének válaszmegtagadók, vagy olyanok, akiket nem sikerült elérni.

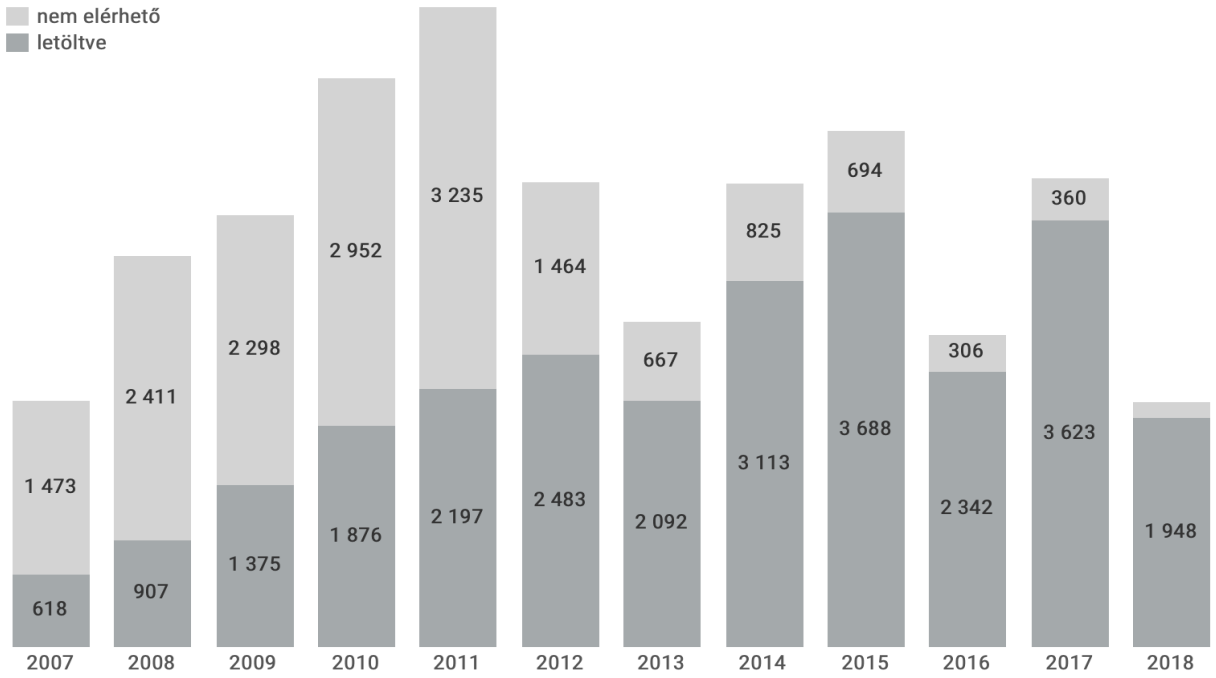
Minden NLP-elemzés első lépése a nyers szövegkorpuszból egy elemzésre alkalmas numerikus adatbázis előállítása. A korrupciós cikkek adatbázisán mi is elvégeztük az előfeldolgozást: azonosítottuk a mondatokat és szavakat *(tokenization),* szótövesítettünk *(lemmatization),* azonosítottuk a szófajokat és más nyelvészeti kategóriákat *(part of speech tagging),* a tartalmatlan szavakat, például névelőket eltávolítottuk a szövegből *(stop word removal)* és összevontuk a szignifikáns bigramokat, valamint a tulajdonneveket és más névelemeket *(named entity recognition).*

A korrupciós diskurzus témáinak automatizált feltárását topikmodellezés segítségével végeztük. A topikmodell azonosítja azokat a látens témákat, melyek a korpusz főbb témáit képezik. Ha például rendelkezésünkre állnak az online sajtó összegyűjtött korrupcióval foglalkozó anyagai, akkor topikmodellezéssel azt vizsgálhatjuk, hogy a szövegek milyen témák köré csoportosulnak. A modell azt feltételezi, hogy a szavak erős szemantikai információkkal rendelkeznek, és a hasonló témákkal foglalkozó dokumentumok hasonló szavak csoportjait használják. A látens témákat tehát a korpusz dokumentumaiban gyakran együtt szereplő szavak csoportjainak azonosításával fedezi fel. A témák (vagy másképp topikok) az elemzés során tárulnak fel, előre nem ismertek, azonban a témák számát (ahogy a klaszterelemzésnél is) előre definiálnunk kell. A topikmodellezés eredményeképp visszakapjuk a dokumentumok topikjainak megoszlását, valamint az egyes topikok szóeloszlását. Az optimális topikszám megválasztása a többféle topikszám mellett illesztett modellek közül a

7 http://k-monitor.hu/adatbazis/aktak

8 https://github.com/codelucas/newspaper/

1. ábra. Az elemzett és a nem elérhető cikkek száma évenként, 2007–2018



Forrás: saját szerkesztés.

„legjobb" modell kiválasztásán alapul. A dinamikus topikmodell az általános topikmodellt (a látens Dirichlet allokációt, LDA-t) az idő dimenziójának megjelenítésével általánosítja, vagyis a folyamatot úgy módosítja, hogy a topikok változhatnak az idő során. Elemzésünkben a látens Dirichlet-allokáció dinamikus változatát használtuk (Blei–Lafferty 2006), a Python Gensim csomagjának LdaSeqModel[9] osztályát alkalmazva.

Az időt években szegmentáltuk, azaz 2007 és 2018 között tizenegy lépésben változhattak a topikok, mégpedig két aspektusukban. Egyrészt évről-évre változhat a topikok valószínűségeloszlása (például az adott téma népszerűbbé válhat). Másrészt a topikok tartalmi változása is megengedett, azaz topikokon belül változhat a kifejezések eloszlása, így például azonosítani lehet a gyorsan és lassan változó témákat.

4.3. Eredmények

A topikszám a topikmodellek bemenő paramétere, vagyis annak megadása az elemző feladata. Több előelemzést futtatva, az interpretálhatóság érdekében és a topikkoherencia *(topic coherence)* mutatóra támaszkodva (Stevens et al. 2012) hét topik megkülönböztetése mellett döntöttünk. A 7-es topikszám esetén jól interpretálható csoportokat kaptunk, melyek tartalmilag egyértelműen elválnak egymástól. Első lépésben azt vizsgáltuk, hogy mi a jelentése az egyes topikoknak, hogyan azonosíthatók tartalmilag (K1). A topikonként legjellemzőbb cikkek „szoros olvasásával" mélyebben is vizsgáltuk, hogy milyen cikkek jelennek meg az egyes topikokban, így az alábbi hét témát azonosítottuk:

9 https://radimrehurek.com/gensim/models/ldaseqmodel.html

1. igazságszolgáltatás,
2. kormányzati és nem kormányzati szervek kapcsolata,
3. közbeszerzések,
4. nemzetközi ügyek,
5. önkormányzati szintű ügyek,
6. pártok és politikusok,
7. vállalkozások, vállalatok.

A kapott hét topik a külső információkkal összhangban van, érvényességük jól alátámasztható. Adódik a kérdés, hogy hogyan kapcsolódnak egymáshoz a topikok, melyek azok, amelyek tartalmilag határozottan elkülönülnek a többitől (K2)? És mely témák kerülnek egymástól távol, melyek közelednek egymáshoz (K4)?

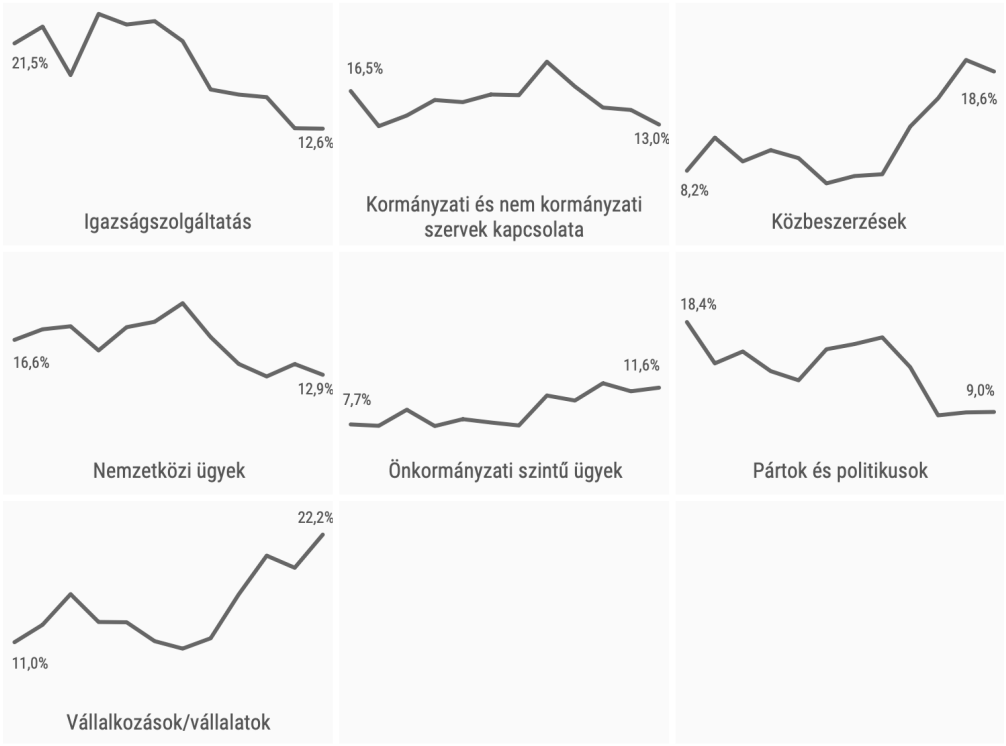
A pártokhoz, politikusokhoz köthető ügyek és az igazságszolgáltatás topikja jól elkülönül a többitől, és ez az elkülönülés stabilan, a vizsgált időszak egészében megmarad. Az elemzett cikkek alapján ez indokolható is: mindkét topik általánosságban hasonló ügyeket dolgoz fel, amelyek a többi topikban kevésbé jelennek meg, ám egymástól nagyon eltérő nézőponttal közelítenek ugyanazokhoz az ügyekhez. Két másik csoportot alkot – a tartalmi elemzés alapján szintén jól indokolhatóan – a nemzetközi ügyek, illetve a kormányzati és nem kormányzati szervek kapcsolatát vizsgáló topik, valamint a közbeszerzések, önkormányzati szintű ügyek és vállalkozások, vállalatok topikcsoportja, de az évek során e két csoport elkülönülése megszűnik, illetve a korábban ugyanazon csoporthoz tartozó topikok (például közbeszerzések és vállalkozások, vállalatok) eltávolodnak egymástól.

A változási dinamika (K3) az adathiány miatt a vártnál nehezebben elemezhetőnek bizonyult, mert a teljes időszakra csupán öt portálról *(Index, Origo, HVG, vilaggazdasag.hu, hetivalasz.hu)* áll rendelkezésre elemezhető tartalom. Három topik válik időben egyre dominánsabbá: a közbeszerzéseket, az önkormányzati ügyeket és a vállalkozások vállalatok ügyeit tematizáló cikkek jelentősége az idő előrehaladtával nő.

Ezt követően a kampányidőszak és a kormányzathoz való viszony hatását vizsgáltuk (K5, K6) 2016 és 2018 között.

Azt vizsgáltuk, mely témákról ír a nem kormánypárti média *(atlatszo.hu, fn.hu, 24.hu, hvg.hu, Index, magyarnarancs.hu, mno.hu, nepszava.hu, nol.hu, 168ora.hu, 444.hu)*, illetve melyekről a kormánypárti média *(figyelo.hu, magyarhirlap.hu, magyaridok.hu, Pesti Srácok, Origo)*, és van-e kapcsolat a tematika és a kormánypárti/nem kormánypárti pozíció között. Ennek kapcsán fontos újra megjegyezni, hogy az egyes cikkek nincsenek pontosan megfeleltetve a topikoknak, csak topikvalószínűséget tudunk rendelni minden cikkhez. Ezek a cikkenkénti topikvalószínűségek voltak a modelljeink függő változói, amelyek kapcsolatát a háttérváltozókkal lineáris regresszió segítségével vizsgáltuk. Külön modell készült az összes topikra. A független változók a következők voltak: a cikk megjelenésének helye (kormánypárti vagy nem kormánypárti média), illetve a cikk megjelenésének időpontja (kampányidőszakban vagy azon kívül jelent-e meg a cikk). A kampányidőszakot a választás előtti négy hónapként definiáltuk. A két magyarázó változó közötti interakciót is bevontuk az elemzésbe, hogy lássuk, módosítja-e például a kampányidőszak a kormánypárti médiumok tematizáltságát, de az interakció egyik esetben sem volt szignifikáns, ezért a bemutatott modellekből ezeket az interakciókat már elhagytuk (1. táblázat).

2. ábra. Topikok eloszlása a 2007–2018 időszakra letöltött összes cikkben



Forrás: saját szerkesztés.

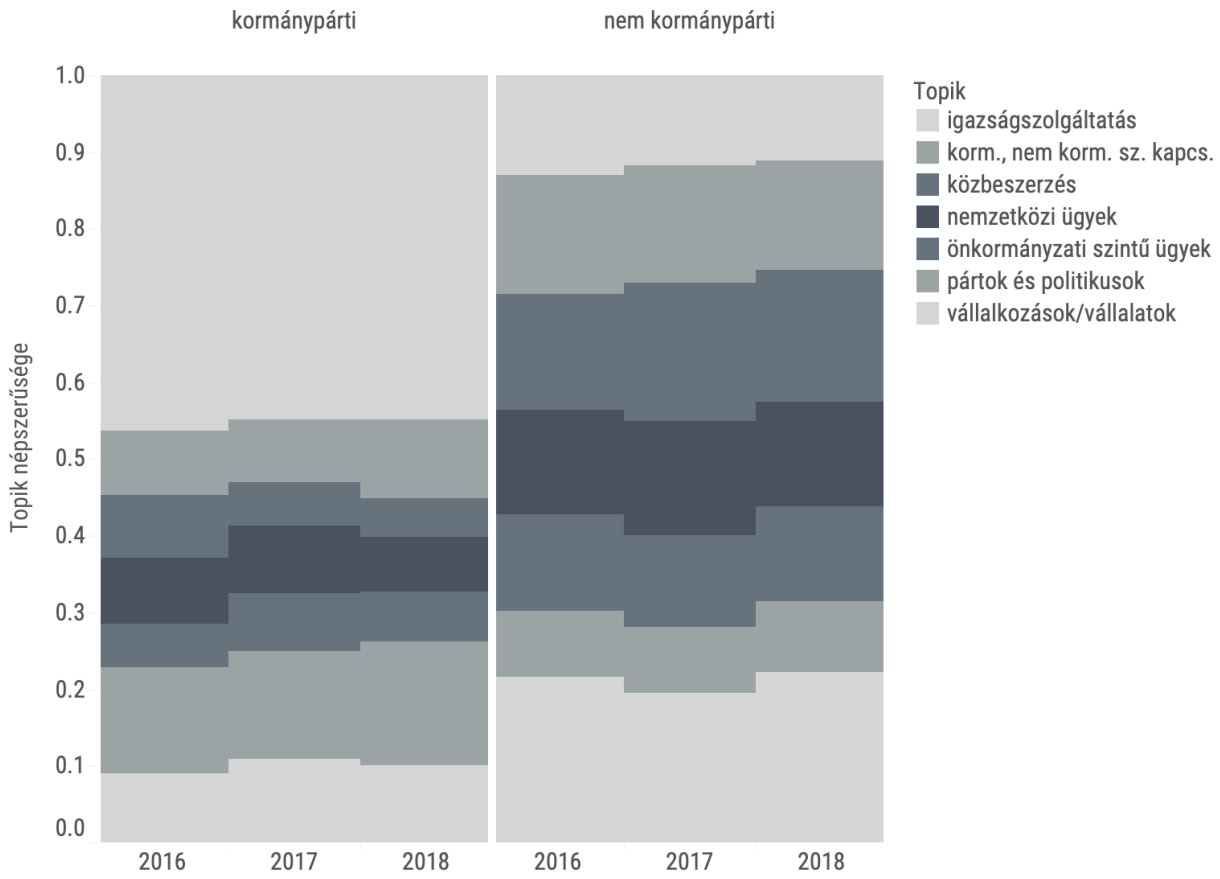
1. táblázat. A lineáris regressziós modellek eredményei

Függő változó	Magyarázó változók	Adj. R2	Koefficiens	t-próba (p)
Igazságszolgáltatás	korm_ell	0,123	-0,335	0,000
	kampány		-0,0093	0,205
Kormányzati és nem korm. szervek kapcs.	korm_ell	0,007	0,0643	0,000
	kampány		-0,0127	0,052
Közbeszerzések	korm_ell	0,012	0,1022	0,000
	kampány		0,0076	0,310
Nemzetközi ügyek	korm_ell	0,007	0,0579	0,000
	kampány		0,0175	0,004
Pártok és politikusok	korm_ell	0,010	-0,0571	0,000
	kampány		0,0134	0,005
Vállalatok, vállalkozások	korm_ell	0,013	0,1101	0,000
	kampány		-0,0189	0,018
Önkormányzati szintű ügyek	korm_ell	0,006	0,0584	0,041
	kampány		0,0024	0,690

A magyarázó változók kódolása: korm_ell esetében kormánypárti = 0, nem kormánypárti = 1, illetve kampány esetében nem kampányidőszak = 0, kampányidőszak = 1

Az összes modell szignifikáns az F-próba alapján, ám a modellek magyarázóereje 7-ből 6 esetben nagyon alacsony, 1% körüli, tehát a tematikát (kézenfekvő módon) más fontos tényezők is magyarázzák a két bevont változó mellett. Egyetlen esetben, az igazságszolgáltatás topik esetében beszélhetünk 12%-os ma-

3. ábra. A topikok eloszlása a 2016–2018 között készült összes cikkben a nem kormánypárti, illetve a kormánypárti sajtóban



Forrás: saját szerkesztés.

gyarázóerőről – eszerint a már igazságszolgáltatási szakaszba került korrupciós esetek megjelenítését erősen befolyásolja, hogy kormánypárti-e az adott médium, és hogy kampányidőszak van-e éppen. Az 1. táblázat alapján azt mondhatjuk el, hogy míg a kormánypárti/nem kormánypárti besorolás hatása mindenhol szignifikáns, addig a kampányidőszak hatása csak a nemzetközi ügyek, a pártokhoz köthető ügyek és a vállalkozások, vállalatok topik esetében szignifikáns. Kampányidőszakban az előbbi két topik nagyobb, míg utóbbi topik kisebb médianyilvánosságot kapott.

A kormánypárti és nem kormánypárti bontást vizsgálva azt mondhatjuk, hogy a kormánypárti médiában a pártokhoz, politikusokhoz köthető ügyek aránya, valamint az igazságszolgáltatás aránya szignifikánsan magasabb (az utóbbi közel 50%-os), a többi topik viszont a nem kormánypárti médiában kapott jelentősebb szerepet. Ezt a 3. ábra szemlélteti.

Utolsó kutatási kérdésünk arra irányult, hogy hogyan hat a tulajdonosváltás egy lap korrupciós tematikájára. Ezt a kérdést az *Origo* kapcsán elemezzük. Az *Origo* körüli változások 2014 júniusában kezdődtek, amikor Sáling Gergő főszerkesztő távozott, de a változás (az *Index* esetével szemben) nem egy adott napon, hanem szépen lassan történik. A kérdésre, miszerint a tulajdonosváltás hatása kimutatható-e egy portál esetében, a 2. táblázat alapján részben már válaszolhatunk is.

2. táblázat. Az Aktákban megtalálható ill. sikeresen letöltött cikkek száma (origo.hu)

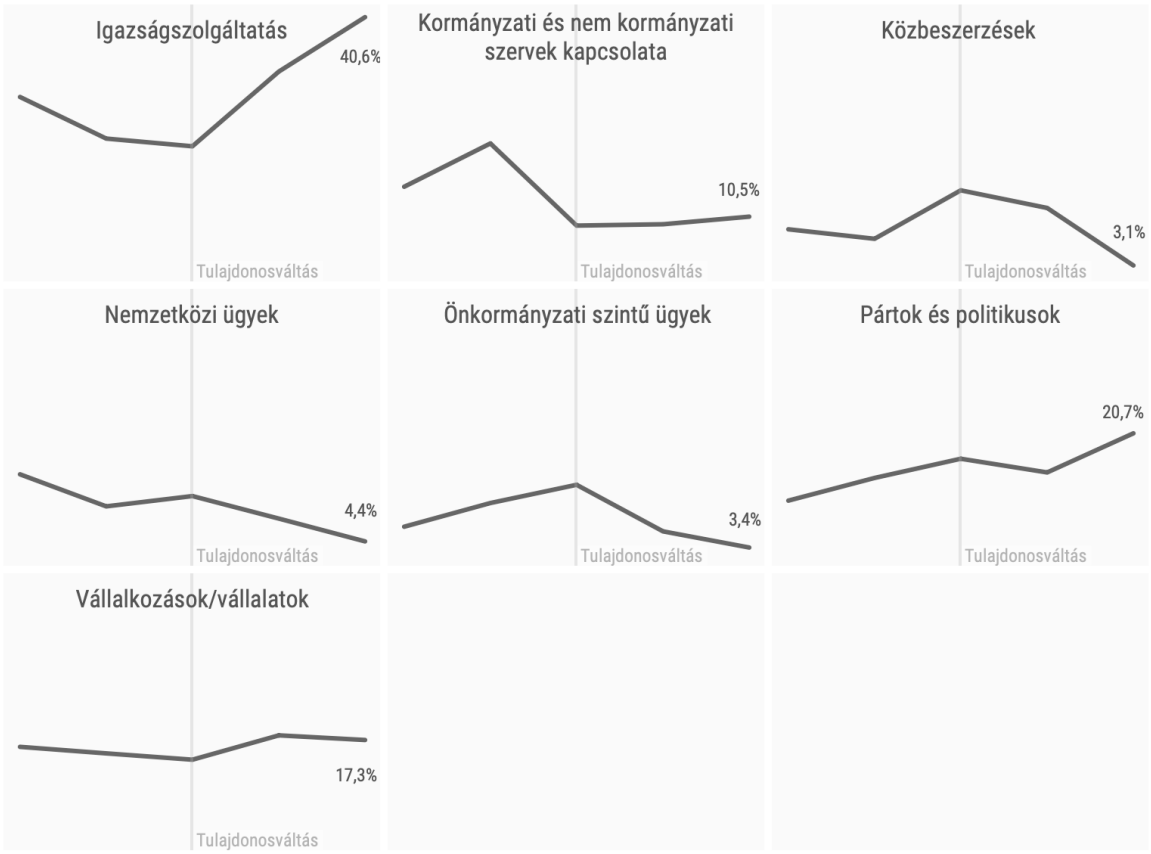
origo.hu	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Összesen	106	140	244	573	594	643	723	241	215	31	72	16
Letöltve	100	132	231	551	545	588	670	229	212	29	70	16

A K-Monitor cikkgyűjteményébe 2016-tól kezdődően radikálisan kevesebb cikk került be, mint a tulajdonosváltás előtt (annak ellenére, hogy az elérési arány nem romlott). Ha a portálon megjelenő topikok arányát vizsgáljuk, a 4. ábra is alátámasztja a tulajdonosváltás befolyásoló erejét. Szignifikáns eltérést három topikban látunk: az igazságszolgáltatás, a kormányzati és nem kormányzati szervek kapcsolata, valamint a nemzetközi ügyek topikban.

4.4. Összegzés

A fejezet azt bizonyítja, hogy emberi olvasás nélkül is jól kimutathatóak a magyar médiapiac azon sajátosságai, amiket más kutatók nem automatizált elemzésekkel állapítottak meg. Például, hogy a sajtóban közölt hírek függenek a tulajdonosi szerkezettől, hiszen a tulajdonosváltás hatását kimutattuk, függenek továbbá a politikai érdekektől, hiszen a politikai ciklus (elemzésünkben a választási időszak) hatása is kimutatható, sőt azt is láttuk, hogy a különböző ideológiai hátterű médiumok különböző módon keretezik a korrupciót. Ez alapján megállapítható, hogy a társadalomtudományi tudástermelésben is jelentős szerepe lehet a természetesnyelv-feldolgozásnak, hiszen segítségével nem csak prevalenciák írhatók le.

4. ábra. A topikok eloszlása az adott évből letöltött összes cikkben 2014 és 2018 között az origo.hu-n



Forrás: saját szerkesztés.

5. A korlátozott verseny előrejelzése a közbeszerzési pályázatok szövegezése alapján[10]

5.1. A kutatás előzményei, problémafelvetés

A közbeszerzések Magyarországon a GDP 15%-át, valamint a kormányzati kiadások mintegy egyharmadát teszik ki. Európa-szerte, a kontinens keleti és nyugati részein egyaránt elterjedtek a korrupcióval kapcsolatos vádak, ennek ellenére keveset tudunk a korrupcióról, és arról, hogy mi mozgatja azt. Az utóbbi években robbanásszerűen megnőtt a hivatalos, kormányzati nyilvántartáson alapuló adatok elérhetősége (Adam–Dávid Barrett–Fazekas 2020). Ez új lehetőséget nyitott a korrupció és a korlátozott verseny tanulmányozására szövegbányászati módszerek alkalmazásával. A fejezetben a korrupciót a közforrásokhoz való korlátozott hozzáférésként értelmeztük, ami a közbeszerzési pályázatok esetében a verseny korlátozásaként jelenik meg (Fazekas–Tóth–King 2016). Az elemzés célja éppen ezért a nyílt verseny korlátozásának előrejelzése a közbeszerzési pályázatok tender szintű szöveges információinak felhasználásával. Az elemzés során online elérhető, hivatalos kormányzati adatokat vizsgáltunk: kb. 120 ezer magyar közbeszerzési szerződést a 2011 és 2020 közötti időszakból. A magyar adatállomány a *Government Transparency Institute* globális szerződés-adatbázisának része.[11]

Akkor beszélhetünk korrupciómentes, jól működő közbeszerzési piacról, ha (ahogyan azt a jogszabályok is rögzítik és a tudományos elméletek is alátámasztják) nyílt és tisztességes a verseny. Tehát, azoknak a vállalatoknak, amelyek képesek a kért feladatokat, szolgáltatásokat ellátni, lehetőséget kell kapniuk az ajánlattételre, és tisztességes elbírálásban kell részesülniük (Yukins 2007). Ezeket az elveket sérti, ha bizonyos ajánlattevők egyenlőtlen bánásmódban részesülnek, például kizárják őket a versenyből, noha optimális esetben részt vehetnének a pályázaton. Ez tulajdonképpen nem más, mint a pártatlanság kritériumának megszegése (Rothstein–Teorell 2008). A közbeszerzési korrupciós kockázatot elemző szakirodalom alapján ennek megfelelően a korrupciót a közforrásokhoz való nyílt hozzáférés korlátozásaként értelmezhetjük (North–Wallis–Weingast 2009, Fazekas–Tóth–King 2016).

Számtalan mérőeszközt, korrupciókockázati mutatót és proxy indikátort fejlesztettek ki, e mutatók az adminisztratív nyilvántartásokban és hivatalos adatbázisokban viszonylag könnyen hozzáférhető (vagy legalábbis könnyen feldolgozható) információkra támaszkodnak (Fazekas–Saussier 2018). Azonban éppen ezért, ezek a mutatók szinte kizárólag strukturált információt, kemény adatot használnak. Közbeszerzési eljárások elemzése esetén ilyen lehet az eljárás típusa vagy a szerződés értéke.

A szövegek és konkrét kifejezések, kulcsszavak fontosságát a közbeszerzési korrupció azonosítására kevés, de értékes szakirodalom mutatja be. Ez a szakirodalom inspirációként és háttérként szolgált elemzésünkhöz. A legtöbb tanulmány szótáralapú módszert használ, azaz a korrupcióhoz és a korrupcióval összefüggő jelenségekhez kapcsolódó kulcsszavak előre meghatározott listájával dolgozik, és a kulcsszavak előfordulását számolja össze.

Rabuzin és Modrusan (2019), akik a horvátországi korrupciógyanús pályázatok azonosításával foglalkoztak, a közbeszerzési pályázatok konkrét szövegezéséből indultak ki. Arra voltak kíváncsiak, hogy a technikai alkalmasság kritériuma segít-e előrejelezni, hogy egy vagy több ajánlattevő nyújt be pályázatot. Ehhez a technikai alkalmasság kifejezést keresték a pályázatokban, és a kulcskifejezést követő 1000 szót használva

10 A fejezet Fazekas Mihállyal közös, jelenleg is folyó kutatásra épít, melyből publikáció még nem született, de folyamatban van.

11 Az adatbázis itt elérhető: opentender.eu/hu/download

építettek modelleket (regressziót, *naive bayes* és SVM modelleket használtak). A legjobb modelljük, egy logisztikus regresszió 69%-os pontossággal működött.

Gorgun és szerzőtársai (2020) szintén azt vizsgálták, hogy a tartalmi információk hogyan befolyásolják a közbeszerzésben részt vevő ajánlattevők számát. Ők azonban nem kétértékű, hanem folytonos változóként tekintettek a kimeneti változóra. Érdekesség a kutatásukban, hogy ők az összes EU tagország adatait elemezték. Mivel ez összesen 24 nyelvet fed le, így az összes szöveget lefordították a Google-fordító segítségével angolra. A cikkükből kiderül (az előző publikáció erre nem tett utalást), hogy a szövegelőkészítés során stopszavazást és szótövezést végeztek, valamint összevontak gyakran együtt járó kifejezéseket is. A legpontosabb előrejelzést náluk a K-legközelebbi szomszéd modell adta, n-gramok bevonásával.

5.2. Adatok és módszerek

Az OECD (2007) által azonosított korrupciós technikák széles palettájáról a legelterjedtebb a pályázati feltételek „testre szabása", azonban a pályázati dokumentáció nem minden szakasza használható egyformán a verseny korlátozására. A pályázati dokumentációnak három olyan fő területe van, amelyet fel lehet használni egy favorizált ajánlattevő előnyben részesítésére: a *részvételi (alkalmassági) feltételek,* a termékleírás és az értékelési szempontok. Jelen kutatás ezekre fókuszált.

A strukturált, formális korrupciós indikátorokkal Fazekas és kutatócsoportja (2016) modelljeit igyekeztünk reprodukálni (3. táblázat), de hosszabb időszakot és kevesebb *„red flag"* változót használtunk. A változószelekcióval a modellünk általánosan érvényesebb, könnyebben alkalmazható különböző országokra (lásd például Fazekas–Kocsis 2020), így széles körben, nemzetközi szinten is alkalmazható lehet a későbbiekben.

3. táblázat. A CRI komponensei

Jelentkezési, beadási időszak	Egy ajánlattevős-e a szerződés
	Nyilvánosan nem meghirdetett pályázat
	Eljárás típusa
	Pályázat benyújtásának feltételeinek számossága
	Túlságosan rövid benyújtási határidő
	Dokumentáció relatív ára
	A pályázati feltételek utólagos módosítása
Értékelési időszak	Egy pályázón kívül mindenki más kizárása
	Más kritériumok figyelembevétele az ár helyett
	Értékelési folyamat megszakítása és újrapályáztatás
	A döntési időszak hossza
Megvalósítási időszak	Szerződésmódosítás
	Szerződés időtartamának/értékének módosítása
	Megnyert pályázatok aránya

Forrás: Fazekas és szerzőtársai (2016) alapján.

A 13 változóból négyet használtunk az elemzésben: a benyújtási időszak hosszát, a döntési időszak hosszát, az eljárás típusát, valamint hogy közzétettek-e ajánlati felhívást. A célváltozónk az, hogy egy vagy több pályázat érkezett az adott közbeszerzésre.

Az elemzésben olyan felügyelt tanulási modelleket alkalmaztunk, melyek a strukturált, formális változók és a strukturálatlan szöveges információk segítségével jelzik előre a verseny korlátozását. A célváltozó bináris

(1 = egy benyújtott ajánlat; 0 = több mint egy benyújtott ajánlat). A használni kívánt modellek kiválasztásakor egyrészt törekedtünk arra, hogy egyensúlyt teremtsünk az értelmezhetőség és a predikciós erő között, másrészt szerettük volna megtartani a kapcsolódást a hagyományos regressziós módszereket alkalmazó korábbi irodalomhoz. Így döntöttünk a bináris logisztikus regresszió és a Random Forest modellek illesztése és összehasonlítása mellett.

5.3. Eredmények

Az alapmodellek csak strukturált adatokat használnak: a kontrollváltozókat, valamint a „*red flag*" változókat. Az alapmodellek pontossága 76–80% között mozog (4. táblázat). Elemzésünkben összességében a Random Forest modellek jobban teljesítettek, mint a logisztikus regressziós modellek. A pályázati dokumentáció különböző részeit először külön-külön adtuk hozzá a modellhez. A modellek mindegyike (bár nem sokkal, de) felülmúlja az alapmodelleket.

4. táblázat. A modellek eredményei

Alapmodellek	precision	recall	f1-score	accuracy
Log. regresszió: kontroll változók	0.57	0.76	0.65	0.76
Random Forest: kontroll változók	0.77	0.79	0.77	0.79
Log. regresszió: kontroll változók + „red flag"-ek	**0.72**	**0.76**	**0.67**	**0.76**
Random Forest: kontroll változók + „red flag"-ek	0.79	0.80	0.79	0.80

Megjegyzés:
a kiemelt modell felépítése hasonlít leginkább a Fazekas kutatócsoportja (2016) által szerkesztett modellre

Szöveges modellek a részvételi feltételekre vonatkozó szövegeken	precision	recall	f1-score	accuracy
LR: szövegek + „red flag" változók	0.78	0.77	0.69	0.77
RF: szövegek + „red flag" változók	0.79	0.77	0.69	0.77
LR: szövegek + „red flag" + kontroll változók	0.74	0.78	0.72	0.78
RF: szövegek + „red flag" + kontroll változók	**0.79**	**0.81**	**0.79**	**0.81**

Megjegyzés: a kiemelt modell teljesít a legjobban

Szöveges modellek a termékleírásra vonatkozó szövegeken	precision	recall	f1-score	accuracy
LR: szövegek + „red flag" változók	0.84	0.85	0.83	0.85
RF: szövegek + „red flag" változók	0.84	0.84	0.82	0.84
LR: szövegek + „red flag" + kontroll változók	**0.84**	**0.85**	**0.84**	**0.85**
RF: szövegek + „red flag" + kontroll változók	0.83	0.84	0.83	0.84

Megjegyzés: a kiemelt modell teljesít a legjobban

Szöveges modellek az értékelési szempontokra vonatkozó szövegeken	precision	recall	f1-score	accuracy

LR: szövegek + „red flag" változók	0.81	0.79	0.73	0.79
RF: szövegek + „red flag" változók	0.81	0.80	0.75	0.80
LR: szövegek + „red flag" + kontroll változók	0.79	0.80	0.76	0.80
RF: szövegek + „red flag" + kontroll változók	**0.81**	**0.83**	**0.81**	**0.83**

Megjegyzés: a kiemelt modell teljesít a legjobban

Szöveges modellek az összes szöveg bevonásával	precision	recall	f1-score	accuracy
LR: szövegek + „red flag" változók	0.85	0.85	0.83	0.85
RF: szövegek + „red flag" változók	0.84	0.85	0.83	0.85
LR: szövegek + „red flag" + kontroll változók	**0.85**	**0.85**	**0.84**	**0.85**
RF: szövegek + „red flag" + kontroll változók	**0.83**	**0.85**	**0.83**	**0.85**

Megjegyzés: a kiemelt modell teljesít a legjobban

A szöveges információkat használó legjobb Random Forest modellek 81% és 85% közötti pontosságot érnek el, míg a logisztikus regressziók jellemzően valamivel kevésbé pontosak, de szintén jobban teljesítenek, mint az alapmodell. A legjobban teljesítő modellek azok, melyeket a termékleírások szövegeire illesztettünk. A modellek eredményei alapján arra következtethetünk, hogy a szöveges információnak a kontroll és a „*red flag*" változókon túl is további magyarázóereje van. Ez arra utal, hogy a pályázati szövegekben szereplő feltételek és kritériumok lehetővé teszik a verseny korlátozását, azaz az egyetlen ajánlattétel valószínűségének növelését, ha a formális, strukturált torzítás nincs is jelen.

Bár a különböző modellek általános teljesítménye alátámasztja hipotéziseinket, a modellek működése mindeddig „fekete doboz" maradt. Az eredmények értelmezése és elméletünkhöz való visszacsatolása érdekében megvizsgáltuk, hogy mely szótöbbesek a legfontosabbak az ajánlattevők számának előrejelzésében, és hogy a legfontosabb n-gramoknak milyen irányú a hatása: csökkentik-e vagy növelik-e az egyetlen ajánlat beérkezésének valószínűségét.

A részvételi feltételek szövegeit vizsgálva azt találtuk, hogy a személyes követelményekből származó legfontosabb n-gramok nagyobb valószínűséggel jelzik előre az egyetlen ajánlat beérkezését, ha a szövegük jogi feltételre utal. Ezzel szemben kisebb az egyetlen ajánlattétel valószínűsége, ha a szöveg azt állítja, hogy az ajánlat benyújtásának nincs jogi feltétele. Emellett a „közös ajánlattevő" kifejezés jelenléte valószínűsíti a több ajánlat benyújtását. Ugyanez látható a gazdasági követelmények legfontosabb kifejezései alapján. Ha egy pályázat lehetővé teszi a közös ajánlattételt, akkor valószínűleg több ajánlattevő jelentkezik, ami csökkenti a korrupció kockázatát. Ugyanakkor az is látható, hogy ágazatspecifikus kifejezések (személygépjármű/autó) is szerepelnek, amit a jövőben érdemes lesz megvizsgálni. A technikai követelményekkel kapcsolatban azt találtuk, hogy ha speciális tapasztalatra van szükség, akkor valószínűbb egyetlen ajánlat beérkezése – ugyanakkor ennek az ellenkezője is igaz: ha általános ismereteket várnak el egy adott témában, akkor valószínűleg több vállalat jelentkezik a feladatra.

A termékleírásokat vizsgálva, azt találtuk, hogy a nagyon specifikus termékek köre inkább valószínűsíti az egyetlen ajánlat beérkezését. Mivel azonban a termékek nagyon sajátos jellegűek, további következtetések levonásához specifikus, ágazati elemzésre van szükség. Az értékelési szempontok alapján a garanciális időszakokat és a jótállási záradékokat értékelő kritériumok általában csökkentik az egyetlen ajánlattétel kockázatát, ami arra utal, hogy az állami beruházások hosszabb távú perspektívája általában csökkenti a korrupciós kockázatokat. Ha az értékelési szempontok a teljes ár helyett az egységköltségeket tartalmazzák,

akkor az egyetlen ajánlattevő pályázásának valószínűsége növekszik. Ez a szerződés végrehajtása során a mennyiségek kijátszására ad lehetőséget: azáltal, hogy az ajánlat az egységárak alapján kezdetben olcsónak tűnik, de a mennyiség kijátszásával az ajánlat később felárazódik (Fazekas–Tóth–King 2013). Az 5. táblázatban látható néhány példa a legfontosabb prediktorokról. A szövegek a lemmatizált és stopszavazott formájukban kerültek be a modellbe, így itt is ebben a formában jelennek meg. A negatív előjelű hatások csökkentik, a pozitív előjelűek pedig növelik a korrupciós kockázatot, amit itt az ajánlattevők számán keresztül a verseny korlátozásával mértünk.

5. táblázat. A legfontosabb prediktorok és hatásuk a részvételi feltételek szövegei alapján

Korrupciós kockázat	Hatásnagyság	n-gram (3, 4 vagy 5 szóból álló kifejezések)
↓	-2,17	szereplő együttes megfelelhet amennyiben
↓	-1,21	kezű aláírás tartalmazó önéletrajz végzettség
↓	-1,01	szervezet kapacitás támaszkodhat gazdasági
↓	-0,75	közös ajánlattevő ajánlat
↓	-0,74	nem tartozik kbt bekezdés
↑	0,75	gazdasági szereplő kbt bekezdés pont
↑	0,82	kbt bekezdés szerinti kizáró
↑	1,02	személygépjármű szállítás származó általános forgalmi
↑	1,18	térkő kivitelezés vonatkozó referencia

5.4. Összegzés

Az elemzés kvalitatív kutatások és szakértői interjúk alapján megfogalmazott hipotéziseket tesztelt és támasztott alá. Bizonyítottuk, hogy a verseny korlátozása gyakran a közbeszerzési pályázatok szövegezésébe bújtatva ér célt. Sőt, elemzésünkkel sikerült olyan szövegrészeket azonosítanunk, melyek akár segíthetnének a korrupciógyanús pályázatok kiszűrésében is. Emellett olyan modellt hoztunk létre, ami pontosabb előrejelzést ad az egyszereplős pályázatok azonosításában, mint a korábbi modellek. Tehát nagy mintán, kvantitatív eszközökkel is sikerült bizonyítanunk a látens tartalom jelentős szerepét a kutatási kérdésünk kapcsán.

Hosszú távú céljaink között szerepel a modellünk nemzetközi szinten is alkalmazható implementációja, valamint, hogy a kutatók és a szakpolitikai szereplők a „szöveg mint adat" felhasználásával bővítsék a korrupciókockázati mutatókat.

6. A kutatás új eredményei, az értekezés főbb megállapításai

Disszertációm négy részből áll: az első szakasz az alkalmazott módszereket helyezi kontextusba: bevezeti a legfontosabb kifejezéseket, módszereket, és adatforrásokat. Ezt követően az előfeldolgozás lépéseinek segítségével szemlélteti, hogy hogyan lesz a szövegből tisztított, statisztikai eszközökkel elemezhető numerikus adatbázis. Végül bemutatja a különbséget a klasszifikációs, predikciós és magyarázó modellek között, valamint a látens témák feltárását szolgáló topikmodellezést.

Az ezt követő fejezet az elmúlt kb. 20 évben korrupciókutatás kapcsán megjelent tanulmányokat vizsgálja, melyben (automatizált) szövegelemzést használtak. Az elemzés bemutatja azt, hogy a kutatók milyen szöveges adatforrást, milyen korrupciómérési módot és milyen elemzési megközelítést használtak. Két tanulmány esetében alternatívát mutat arra, hogy hogyan lehetne hatékonyabbá és szélesebb körre is általánosíthatóvá,

vagy adott esetben objektívebbé, tenni az elemzést az automatizált szöveganalitika segítségével.

A harmadik szakasz, azaz az első empirikus elemzés a korrupció hazai online médiareprezentációjának tematikus elemzését mutatja be, egy nem felügyelt szövegbányászati megközelítést, azon belül is dinamikus topikmodellezést alkalmazva. A szövegkorpuszt a K-Monitor cikkgyűjteménye adta, amely korrupciógyanús ügyeket feldolgozó, online sajtóban megjelent cikkeket tartalmaz. Az esettanulmány egyfelől explorativ jellegű: a 2007–2018 közötti időszakra vonatkozóan azonosítja a cikkek főbb témáit és a tematikus változás dinamikáját. A kutatás magyarázatokat is keresett: vizsgálta, hogy van-e kapcsolat a tematika és a médium kormánypárti/nem kormánypárti pozíciója között, illetve, hogy a kampányidőszak befolyásolja-e a korrupció reprezentációjának tematikáját. Jelentős különbségek mutatkoznak abban, hogy a különböző portálok milyen témákról milyen gyakran számoltak be. Annak köszönhetően, hogy az elemzett időszakban megváltozott az *Origo* hírportál tulajdonjoga, alá lehet támasztani a tulajdonosváltás szerepét a korrupciós témák (és azok mennyiségének) változásában. A természetesnyelv-feldolgozás segítségével egy 11 évet átfogó, sok portált szemléző adatbázis összes cikkét, mintavétel nélkül vizsgálhatjuk, a „távoli olvasás" segítségével a tematikus mintázatok feltárására nyílt lehetőség. Olyan mennyiségű szöveget elemezhettünk, ami „szorosan olvasva" nem lett volna lehetséges.

A negyedik szakasz hivatalos, adminisztratív adatokat vizsgál, felügyelt módszer segítségével. Az elemzés a nyílt verseny korlátozásának előrejelzésével foglalkozik, a közbeszerzési pályázatok tenderszintű szöveges információinak felhasználásával. Több mint 170 ezer magyar közbeszerzési szerződést vizsgál a 2011 és 2020 közötti időszakból. A bemutatott szövegbányászati megközelítés meghatározó lehet azok mellett a mutatók mellett, melyeket jelenleg a korrupciós kockázat indikátoraiként alkalmaznak, hiszen a létező indikátorokra épít, de túlmutat azokon azáltal, hogy a pályázati felhívások szövegeiből indul ki. A korábbi kvalitatív esettanulmányok is megmutatták, valamint jelen elemzés is alátámasztja, hogy a pályázati feltételeket sokszor egyetlen, előnyben részesített ajánlattevőre szabják, és ezáltal kizárják a versenytársakat a pályázatból. Először a korábbi kutatások megismétlésével, majd logisztikus regresszió és Random Forest modellek illesztésével az egyetlen ajánlattevő mint korrupciós kockázat megjóslása a cél. Az eredmények azt mutatják, hogy a szöveges adatokat használó modellek felülmúlják a korábbi előrejelzéseket. A modell jóságának ellenőrzésénél nem áll meg az elemzés: vizsgálja azt is, hogy a különböző szótöbbeseknek milyen hatása van az eredményre. Ez azért különösen fontos, mert sajnos sok tanulmány megáll a modell teljesítményének ismertetésénél, figyelmen kívül hagyva azt, hogy a társadalomtudományos gondolkodásban az interpretációnak jelentős szerepe van.

A két esettanulmány tehát megmutatta, hogy a természetesnyelv-feldolgozás hogyan tud a társadalomtudományi tudástermeléshez hozzájárulni; kiválthatja és megerősítheti a kvalitatív megközelítést területtől függően. Nagy segítség lehet, hogy nem kell elolvasni több tízezer vagy százezer cikket, de a korpusz előállítása és az elemezhető adatbázis kialakítása nagy munka. Emellett meg kell érteni és választ kell adni az adathiány forrására, ami sok időt, erőforrást és nagy szakértelmet igényel. Szeretném hangsúlyozni azt is, hogy az értelmezéshez és az eredmények validálásához elengedhetetlen a tartalmi tudás. A módszer alkalmazása tehát elősegítheti a kutatók együttműködését, erősítheti az interdiszciplinaritást.

Hivatkozások

Adam, I. – Dávid Barrett, E. – Fazekas, M. (2020) *Modelling Reform Strategies for Open Contracting in Low and Middle Income Countries.* London, UK: Transparency International.

Arksey, H. – O'Malley, L. (2005) Scoping studies: Towards a methodological framework. *International Journal of Social Research Methodology: Theory and Practice,* 8(1), 19–32. https://doi.org/10.1080/1364557032000119616

Axelsson, S. – Dahlberg, S. (2018) Corruption Talk: Mapping the Word Corruption in Online Text Data Across the World. General Conference of the European Consortium for Political Research. *Manuscript.*

Blei, D. M. – Lafferty, J. D. (2006) Dynamic topic models. *Proceedings of the 23rd international conference on Machine learning (ICML ,06).* New York, USA: Association for Computing Machinery, 113–120. https://doi.org/10.1145/1143844.1143859

Fazekas, M. – Tóth, I. J. (2016) From Corruption to State Capture: A New Analytical Framework with Empirical Applications from Hungary. *Political Research Quarterly,* 69(2), 320–334. https://doi.org/10.1177/1065912916639137

Fazekas, M. – Saussier, S. (2018) Big Data in Public Procurement. Colloquium. In Piga, G. –Tátrai, T. (szerk.) *Law and Economics of Public Procurement Reforms* (Chapter 3). London: Routledge, 131–146.

Fazekas, M. – Kocsis, G. (2020) Uncovering High-Level Corruption: Cross-National Objective Corruption Risk Indicators Using Public Procurement Data. *British Journal of Political Science,* 50(1), 155–164. https://doi.org/10.1017/S0007123417000461

Fazekas, M. – Cingolani, L. – Tóth, B. (2018) Innovations in Objectively Measuring Corruption in Public Procurement. In Anheier, H. K. – Haber, M. – Kayser, M. A. (szerk.) *Governance Indicators. Approaches, Progress, Promise* (Chapter 7). Oxford: Oxford University Press. https://doi.org/10.1093/oso/9780198817062.003.0007

Fazekas, M. – Tóth, I. J. – King, P. L. (2016) An Objective Corruption Risk Index Using Public Procurement Data. *European Journal of Criminal Policy and Research,* 22(3), 369–397. https://doi.org/10.1007/s10610-016-9308-z

Fazekas, M. – Tóth, I. J. – King, P. L. (2013) Corruption Manual for Beginners: 'Corruption techniques' in Public Procurement with Examples from Hungary. *Corruption Research Center Budapest Working Paper No. CRCB-WP/2013:01.* https://doi.org/10.2139/ssrn.2333354

Gorgun, M. K. – Kutlu, M. – Taş, B. K. O. (2020) Predicting The Number of Bidders in Public Procurement. *2020 5th International Conference on Computer Science and Engineering (UBMK).* Diyarbakir, TR: IEEE, 360–365. https://doi.org/10.1109/UBMK50275.2020.9219404

Hajdu M. – Pápay B. – Szántó Z. – Tóth J. I. (2018a) A korrupció sajtómegjelenése nemzetközi összehasonlításban. *Magyar Tudomány,* 179(4), 496–506.

Hajdu M. – Pápay B. – Szántó Z. – Tóth J. I. (2018b) Content analysis of corruption coverage: Cross-national differences and commonalities. *European Journal of Communication*, 33(1), 7–21. https://doi.org/10.1177%2F0267323117750673

Hajdu M. – Pápay B. – Szántó Z. – Tóth I. J. (2016) *Human Assisted Content Analysis of the print press coverage of corruption in Hungary.* Projekt jelentés. ANTICORRP. Elérhető: http://unipub.lib.uni-corvinus.hu/2689/1/D6_1_16.pdf [Letöltve: 2022-05-30].

Hlatshwayo, S. – Oeking, A. – Ghazanchyan, M. – Corvino, D. – Shukla, A. – Leigh, L. (2018) The Measurement and Macro-Relevance of Corruption: A Big Data Approach. *IMF Working Papers,* 18(195), 1. https://doi.org/10.5089/9781484373095.001

Katona E. – Kmetty Z. – Németh R. (2021) A korrupció hazai online média-reprezentációjának vizsgálata természetes nyelvfeldolgozással. *Médiakutató,* 22(2), 69–88.

Katona E. – Németh R. (2021) Automatizált szöveganalitika a korrupció kutatásában. *Socio.hu Társadalomtudományi Szemle,* 11(1), 108–124. https://doi.org/10.18030/socio.hu.2021.1.108

Lambsdorff, J. (2007) *The institutional economics of corruption and reform: theory, evidence, and policy.* Cambridge: Cambridge University Press.

Li, J. – Chen, W. – Xu, Q. – Shah, N. – Mackey, T. (2019) Leveraging Big Data to Identify Corruption as an SDG Goal 16 Humanitarian Technology. Seattle, WA, USA: *IEEE Global Humanitarian Technology Conference (GHTC)*, 1–4. https://doi.org/10.1109/GHTC46095.2019.9033129

Martin, J. P. (2019) Kéz a kézben a lejtőn. *Médiakutató: Médiaelméleti Folyóirat,* (20)3, 7–21.

Messing V. – Szeitl B. – Ságvári B. (2022) Webre terelés a személyes lekérdezés alternatívája? – Egy „push-to-web" hibrid survey tapasztalatai. *Statisztikai Szemle,* 100(3), 213–233. https://doi.org/10.20311/stat2022.3.hu0213

Miner, G. – Elder, J. – Hill, T. – Nisbet, R. – Delen, D. – Fast, A. (2012) *Practical Text Mining and Statistical Analysis for Non-structured Text Data Applications*. Waltham, MA: Academic Press of Elsevier.

Moretti, F. (2013) *Distant Reading*. London: Verso Books.

Moretti, F. (2000) Conjectures on World Literature. *New Left Review*, (1), 54–68.

Muço, A. (2019) *Learn from thy Neighbor: Do Voters Associate Corruption with Political Parties?* Publikálatlan kézirat.

Németh E. – Körmendi G. – Kiss B. (2011) Korrupció és nyilvánosság. A média hatása a korrupcióra és annak társadalmi megítélésére. *Pénzügyi Szemle,* 56(1), 57–65.

Németh R. – Katona E. – Kmetty Z. (2020) Az automatizált szövegelemzés perspektívája a társadalomtudományokban. *Szociológiai Szemle,* 30(1), 44–62. https://doi.org/10.51624/SzocSzemle.2020.1.3

Noerlina – Wulandhari, L. – Sasmoko, S. – Muqsith, A. – Alamsyah, M. (2017) Corruption Cases Mapping Based on Indonesia's Corruption Perception Index. *Journal of Physics: Conference Series*, 801 012019. https://doi.org/10.1088/1742-6596/801/1/012019

North, D. C. – Wallis, J. J. – Weingast, B. R. (2009) *Violence and Social Orders: A Conceptual Framework for Interpreting Recorded Human History.* Cambridge: Cambridge University Press.

OECD (2007) *Integrity in Public Procurement. Good Practice from A to Z.* Paris: OECD.

Ogunmuyiwa, H. O. (2015) A Critical Discourse Analysis of Corruption in Presidential Speeches. *International Journal for Innovation Education and Research*, 3(12), 31–50. https://doi.org/10.31686/ijier.vol3.iss12.484

Pan, J. – Chen, K. (2018) Concealing Corruption: How Chinese Officials Distort Upward Reporting of Online Grievances. *American Political Science Review,* 112(3), 602–620. https://doi.org/10.1017/S0003055418000205

Park, C. S. (2012) How the media frame political corruption: Episodic and thematic frame stories found in Illinois newspapers. *Paper Originally Prepared for the Ethics and Reform Symposium on Illinois Government (September 27-28, 2012)*. Elérhető: http://paulsimoninstitute.siu.edu/_common/documents/whats-in-the-water/water-illinois/park.pdf [Letöltve: 2022-05-30].

Rabuzin, K. – Modrusan, N. (2019) Prediction of Public Procurement Corruption Indices using Machine Learning Methods. *Proceedings of the 11th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management, KMIS.* Vienna, AT: KMIS, 333–340. https://doi.org/10.5220/0008353603330340

Rothstein, B. – Teorell, J. (2008) What is Quality of Government? A Theory of Impartial Government Institutions. *Governance,* 21(2), 165–190. https://doi.org/10.1111/j.1468-0491.2008.00391.x

Stevens, K. – Kegelmeyer, P. – Andrzejewski, D. – Buttler, D. (2012) Exploring Topic Coherence over Many Models and Many Topics. *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12).* Jeju Island, Korea: Association for Computational Linguistics, 952–961. Elérhető: https://www.aclweb.org/anthology/D12-1087.pdf [Letöltve: 2022-01-23].

Suphachalasai, S. (2005) Bureaucratic Corruption and Mass Media. *Environmental Economy and Policy Research Discussion Paper No. 05.2005*. Cambridge: University of Cambridge, Department of Land Economics. https://doi.org/10.2139/ssrn.722403

Yukins, C. (2007) Integrating Integrity And Procurement: The United Nations Convention Against Corruption and the Uncitral Model Procurement Law. *Public Contract Law Journal,* 36(3), 307–329.

Zhang, Y. – Wildemuth, B. M. (2005) *Qualitative Analysis of Content*. 1(2), 1–12.