Katona Eszter – Kmetty Zoltán – Németh Renáta

A korrupció hazai online médiareprezentációjának vizsgálata természetes nyelvfeldolgozással

Cikkünk a korrupció hazai online médiareprezentációjának tematikus elemzését mutatja be szövegbányászati megközelítést, azon belül is dinamikus topikmodellezést alkalmazva. Szövegkorpuszunkat a K-Monitor cikkgyűjteménye adta, amely korrupciógyanús, valamint szabálytalan közpénz-felhasználással kapcsolatos ügyeket feldolgozó, online sajtóban megjelenő cikkeket tartalmaz. Esettanulmányunk egyfelől exploratív jellegű: célunk a 2007–2018 közötti időszakra vonatkozóan azonosítani a cikkek főbb témáit és a tematikus változás dinamikáját, az egyes korrupciós témacsoportok előtérbe kerülését és háttérbe szorulását, illetve az egyes témák tartalmi változását. Kutatásunk másfelől magyarázatokra is kísérletet tesz, annak vizsgálatával, hogy van-e kapcsolat a tematika és a médium ellenzéki/kormánypárti pozíciója között, illetve hogy a kampányidőszak befolyásolja-e a korrupció reprezentációjának tematikáját. Annak köszönhetően, hogy az elemzett időszakban megváltozott az Origo hírportál tulajdonjoga, természetes kísérletként adódik annak vizsgálata, hogy a tulajdonosváltás milyen hatást gyakorolt a portálon megjelenő korrupciós diskurzus tematizáltságára.

Kulcsszavak: automatizált szövegelemzés, dinamikus topikmodell, korrupció, NLP, szövegbányászat

1. Bevezetés

A korrupció kutatása több oldalról kapcsolható szöveges adatokhoz, elsősorban szerkesztett vagy közösségi médián mint forráson alapulva (például Mear 2016, Niklander et al. 2016, Li et al. 2019). Ezeknek a szöveges adatoknak a felhasználásával új szemléletű megközelítésre nyílik lehetőség, ugyanakkor a folyamatosan termelődő digitális szöveges adatforrás kiaknázása részben megoldást is nyújthat meglevő empirikus problémákra, gondoljunk csak a klasszikus survey-hez kapcsolódó egyre jelentősebb válaszmegtagadásra.

Számos kvalitatív jellegű korrupciókutatás készült a szerkesztett médiából kiindulva, jellemzően a kritikai diskurzuselemzést követve (például Touwe & Sultan 2015). De a kvantitatív korrupciókutatások is egyre nagyobb számban elemeznek szerkesztett médiát – olyan komoly szereplők is, mint a Nemzetközi Valutaalap (IMF) (Hlathswayo et al. 2018), amelynek a legújabb, harmadik generációs big data alapú korrupciómutatója, a *news-flow index* kizárólag a korrupció média-reprezentációjára épít.

Tanulmányunkban azt vizsgáljuk, hogy az elmúlt 15 évben a hazai online média milyen tematikai fókuszok mentén mutatta be a korrupciós ügyeket. Ehhez az elemzéshez a K-Monitor (a korrupció visszaszorításáért küzdő civil szervezet) által gyűjtött, korrupcióval kapcsolatos újságcikkeket használjuk fel. Az elemzés során olyan új, kvantitatív szövegbányászati megoldásokra támaszkodunk, amelyek a hazai társadalomkutatásokban eddig kevéssé voltak használatosak. Szövegek gyűjteményének (azaz egy szövegkorpusznak) az elemzése a survey-kutatásokhoz képest egészen más típusú tudást nyújthat számunkra a vizsgált témában. Egyrészről nagyon árnyalt képet lehet arról kapni, hogy a korrupcióval kapcsolatos diskurzus milyen dimenziók mentén alakul; ez a tudás korábban csak kvalitatív, kis mintás kutatásokból tudott kirajzolódni, hátrányként hordozva azok általánosíthatósági korlátját. Másrészről hosszabb időszakokat elemezve képet kaphatunk arról is, hogy a digitális médiában miként alakult a korrupció reprezentációjának dinamikája.

Elemzésünk vállaltan nem technikai központú, tehát elsősorban eredményoldalról mutatjuk be a kutatásunk kimenetét, remélve, hogy inspirációt tudunk nyújtani újabb, hasonló jellegű kutatásoknak. A módszertani részletek iránt érdeklődő olvasónak a természetes nyelvfeldolgozás társadalomtudományi lehetőségeit bemutató összefoglalónkat ajánljuk (Németh et al. 2020).

2. A survey-kutatásoktól a sajtóelemzésekig

A korrupcióval kapcsolatos vizsgálatokat több szempontból is elkülöníthetjük egymástól (Bohn 2012, Kmetty & Kostadinova 2019). Vizsgálhatjuk a korrupciós észlelések jellegét (ami lehet közvetlen tapasztalatra épülő vagy általános, közvetett benyomáson alapuló észlelés), szintjeit (egyéni, társadalmi), vagy az egyéntől való távolságát (hétköznapi, politikai). Ezek a kutatások általában survey-adatfelvételeket használnak fel, amelyek ugyan garantálják az anonimitást, de nem beavatkozásmentes jellegük miatt hatással lehetnek az eredményekre (Axelsson & Dahlberg 2018). A kérdőíves kutatások kapcsán Fazekas Mihály és munkatársai (2016) két aggályt fogalmaznak meg. Egyrészt problematikus a téma szenzitivitása, aminek köszönhetően magas a nemválaszolási arány, és kétségbe vonható a válaszok megbízhatósága is. Másrészt a lakosság igen kis része rendelkezik (főként a politikai) korrupció kapcsán közvetlen tapasztalattal.

A szubjektív becsléseken alapuló mutatók mellett végezhetők beavatkozásmentes vizsgálatok is, amelyek a korrupció volumenét mérik (beavatkozáson értve nemcsak a kísérletet, hanem a kérdezést, az interjút is). Ilyen elemzések során már rendelkezésre álló adatokból készítenek indikátorokat. Ezek a mérések – bár a fentiekkel szemben objektív adatokon alapulnak – jellemzően nem a korrupt tranzakció bekövetkeztét, hanem csupán az annak kedvező feltételek fennállását detektálják, tehát inkább a korrupciós kockázat elemzésére, a korrupciógyanús esetek azonosítására alkalmazhatók. Beavatkozásmentes vizsgálatokban használják a hagyományos kvantitatív (és kvalitatív) módszerek mellett az újabb típusú *(big data)* módszereket is. Az utóbbiak közé tartozik a bevezetőben említett IMF-indikátor is, a *cross-country news flow indices of corruption* (NIC, Hlathswayo et al. 2018). Az NIC-indexet a hivatkozott IMF-riport a korrupció mértékének, dinamikájának és országközi eltéréseinek vizsgálatára használja. A korrupció médiareprezentációjából indul ki, ezzel célja a percepcióalapú indexek hátrányainak kiküszöbölése. A nagy mennyiségű szöveges adatokból kiinduló elemzések közé sorolhatjuk Fazekas és munkatársai (2016) kutatásait is, akik hazai közbeszerzések 150 000 esetet tartalmazó adatbázisán elemezték az eljárás korrupciós kockázatait.

Az eddig tárgyalt módszerek alapvetően a korrupció nagyságát és elterjedtségét mérik, de nem adnak képet arról, hogy milyen korrupciós tematikák mentén alakul a közbeszéd egy adott országban vagy kisebb közösségben. A kvantitatív sajtóelemzések éppen ilyen kérdésekre tudnak választ adni. Ezek az elemzések jelentik a szűkebb fókuszát a mi kutatásunknak is.

E kutatási irány egy másik hazai reprezentánsaként Hajdu Miklós és munkatársai (2018a & 2018b) említhetők, akik cikkükben tartalomelemzéssel vizsgálják a korrupció médiareprezentációját a 2004 és 2013 közötti időszakban, több ország összevetésében. Azért is különösen fontos a téma média felőli vizsgálata, mert a korrupció percepcióját nagyban befolyásolja a média helyzete, a sajtószabadság megléte vagy hiánya (Suphachalasai 2005). A korrupcióról szóló híradások hatására ugyanis csökkenhet a politikai korrupció mértéke (Németh et al. 2011), de ezzel együtt növekedhet a korrupció társadalmi percepciója (ezzel részben ellentétes eredmény Kmetty 2018).* A fenti kutatásokban közös, hogy főként szógyakoriság-elemzéseket használnak az elemzéseikben, illetve kézi kódolást, szakértői besorolást. A szövegbányászati megoldások társadalomtudományi elterjedésével párhuzamosan azonban megjelentek az első olyan nemzetközi cikkek is, amelyekben a természetes nyelvfeldolgozás (*natural language processing*, NLP) nem csupán adatgyűjtési, hanem tartalmi eredményt előállító elemzési eszközként is funkcionál. Erre jó példa Sofia Axelsson és Stefan Dahlberg (2018) kutatása, amelyben disztribúciós szemantika segítségével ragadják meg a korrupciót az online és közösségi médiában. Eredményeik szerint minden régióra jellemző, hogy

* Lásd még Martin József Péter „Kéz a kézben a lejtőn. Korrupció és sajtószabadság a magyar trendek tükrében" című írását lapunk 2019. őszi számában – A szerk.

a korrupcióhoz szemantikailag közel esik a bűnözés (kézenfekvő módon), de posztkommunista sajátosság, hogy a szegénység, a terrorizmus, a migráció is. További példákért és általában a természetes nyelvfeldolgozás korrupciókutatási lehetőségeivel kapcsolatban lásd Katona Eszter és Németh Renáta (2021) irodalmi összefoglaló tanulmányát.

Tanulmányunk szintén a sajtóelemzésen alapuló korrupciókutatások közé tartozik. Kutatásunkban a K-Monitor által összegyűjtött, a hazai online médiában 2007 és 2018 között megjelent, korrupcióval, szabálytalan közpénzfelhasználással kapcsolatos cikkeket elemzünk. Tanulmányunk a szubsztantív fókusz mellett azt is demonstrálja, hogyan lehet a modern szövegbányászati megoldásokat médiaelemzésekben felhasználni. Bár munkánk vállaltan exploratív jellegű, igyekeztünk megfogalmazni olyan kutatási kérdéseket is, amelyek túlmutatnak a jelenségek egyszerű leírásán, és tartalmi szempontból is új tudást generálnak.

3. Kutatási kérdések

Tanulmányunkban tehát a hazai online média korrupcióval, szabálytalan közpénzfelhasználással kapcsolatos diskurzusának tematikus elemzését mutatjuk be. A 2007 és 2018 közötti időszakra vonatkozóan azonosítjuk a cikkek főbb témáit és a tematikus változás dinamikáját: az egyes korrupciós témacsoportok előtérbe kerülését és háttérbe szorulását, illetve az egyes témák tartalmi változását. Ahogy korábban is jeleztük, munkánk exploratív jellegű, ebből következően nem fogalmaztunk meg konkrét hipotéziseket; kutatási kérdéseinket nyitva hagytuk.

A kiinduló kutatási kérdésünk (K1) az, hogy milyen témák kerülnek elő a hazai sajtóban a korrupció kapcsán. Ezzel szorosan összefügg a második kutatási kérdésünk (K2), amely arra vonatkozik, miként hogyan kapcsolódnak egymáshoz ezek a témák, s melyek azok, amelyek tartalmilag határozottan elkülönülnek a többitől.

A tanulmányunkban kiemelt szerepet kap az eredmények dinamikai szempontú elemzése. Mivel több mint tíz évnyi újságcikket elemzünk, lehetőség van arra is, hogy megfigyeljük, miként változik az egyes témák belső szerkezete (K3), és melyek azok a témák, amelyek időben egymáshoz közelebb kerülnek vagy éppen egymástól távolodnak (K4). A dinamikai változások kapcsán külön vizsgáljuk, hogy a 2014-es választási kampány időszakában a kampányon kívüli időszakhoz képest miként változott a hazai sajtó korrupciós tematikája (K5).

Az időbeli dinamika vizsgálata mellett jól megragadható dimenzió az online újságok politikaipárt-közelsége is. Külön elemezzük, hogy az ellenzéki és a kormánypárti médiában mennyire térnek el egymástól a vizsgált korrupciós tematikák (K6). Az utolsó kutatási kérdésünk is részben ezzel kapcsolatos, bár egy speciális fókusszal. Az adatbázis lehetőséget teremt egy érdekes természetes kísérlet kiértékelésére is, amely azt vizsgálja, hogyan változtatja meg a lap korrupcióval kapcsolatos diskurzusát egy tulajdonosiszerkezet-változás. Elemzésünkben az Origó korrupcióval kapcsolatos cikkeit vizsgáljuk ebből a szempontból (K7).

Mivel elemzésünkben az interpretációt nagyban támogatja az interaktív adatvizualizáció, a tanulmányban bemutatott minden ábra, valamint azok interaktív verziói is elérhetőek az alábbi linken.[1] Ez a vizualizáció számunkra nem mellékes illusztrációként, hanem a mélyebb megértéshez szükséges fontos eszközként szolgált, amellyel az olvasó is végezhet saját mélyfúrásokat.

4. Felhasznált adatbázis

Tanulmányunkban a K-Monitor által azonosított, korrupcióval kapcsolatos újságcikkeket használjuk. A K-Monitor az „Akták" nevű adatbázisában[2] gyűjt össze minden olyan cikket, amely korrupcióval kapcsolatban jelent meg a magyar sajtóban. A szervezettől kapott adatbázis az „Akták" cikkeire mutató linkeket tartalmazza. Az „Akták" olyan témákat járnak körül, amelyek gyakrabban előkerülő, több fordulatot vett vagy a jövőben várhatóan folytatódó ügyekről szólnak.

1 Lásd https://rc2s2.github.io/Corruption/mediakutato/.

2 Lásd http://k-monitor.hu/adatbazis/aktak.

Az adatbázisba olyan cikkek linkjei kerülnek be, amelyek az alábbi kritériumok valamelyikének megfelelnek:[3]

1. konkrét korrupciós esetet mutatnak be,
2. szerzője állítja vagy sugallja, hogy valaki a ráruházott hatalmat saját maga vagy egy harmadik fél hasznára fordította,
3. korrupciós ügyben történő jogi eljárásról tájékoztatnak,
4. vagy épp ellenkezőleg: egy korrupciós vádat cáfolnak,
5. a következő témák esetében elkövetett szabálytalanságokról szólnak: közbeszerzés, pártfinanszírozás, pályázatok, kormányzat szerv vagy állami vállalat gazdálkodása, vagyonosodás, juttatások, privatizáció, *whistleblowing*,
6. a következő kulcsszavak valamelyikét tartalmazzák: korrupció, sikkasztás (közszolga által), hűtlen kezelés, vesztegetés, hivatali visszaélés, hatalommal való visszaélés, befolyással üzérkedés, hanyagság, adócsalás, számviteli fegyelem megsértése, protekció, nepotizmus, jogosulatlan gazdasági előny, versenykorlátozás, kartell, whistleblowing/közérdekű bejelentés, közérdekű adatok, átláthatóság.

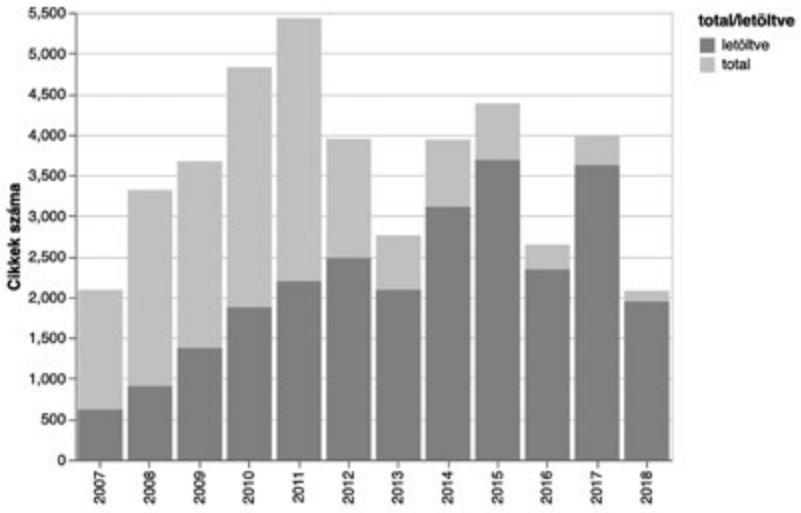
A cikkek gyűjteményébe nem kerülnek be pártközlemények, publicisztikák, illetve hivatkozott vagy más portálról átvett anyagok. Az MTI-cikkek esetében a K-Monitor munkatársai a legrészletesebb forrásból származó cikket viszik be, így törekednek a tartalmi ismétlődések elkerülésére. Az adatbázisba csak online vagy online is megjelenő cikkek kerülnek be, a források minden esetben online hírportálok.

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 csomagját használtuk a cikkek letöltéséhez.[4]

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 (a portálok listáját lásd a Függelékben).

Az elemzett adatbázis 2007-től, az adatbázis építésének kezdetétől indul, és 2018 augusztusában zárul, tehát 2018-ra a cikkek gyűjteménye nem teljes körű. A cikkek évenkénti megoszlását az 1. ábra mutatja. Ezen látható, hogy 2012-ig folyamatosan nőtt az adatbázisban a cikkek száma, míg 2013-ban és 2016-ban nagyobb visszaesés tapasztalható. Az ábrán szintén jól látható, hogy a korábbi évekből arányaiban kevesebb cikket tudtunk elérni és letölteni. Erre még visszatérünk a következőkben.

1. ábra
Elemzett cikkek száma évenként



3 A szervezettől kapott módszertani dokumentáció alapján.

4 Lásd https://github.com/codelucas/newspaper/.

Az 1. táblázatban azt a húsz portált mutatjuk be, amelyről a legtöbb cikk származik. Az adatbázisban előfordulnak olyan hivatkozások is, amelyek esetén a gyűjteményben található link olyan oldalra mutat, ahonnan a cikket már eltávolították. Erre számos példát találtunk; például az mno.hu oldalon csupán 30 százalékos volt az elérési arány. Jól látható a táblákból, hogy a 2010 előtti időszak gyűjtése nem teljes, a kor korrupciós médiatermésének csupán egy szeletét látjuk, s inkább a 2010-es évek vége felé kezelhetjük teljes körűnek kiinduló korpuszunkat.[5]

1. táblázat
A top húsz forrásoldalra mutató linkek, és az azokról letöltött cikkek aránya

	Össz	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	letöltve	%
Index	4882	328	399	428	486	525	395	319	241	291	439	629	240	4720	97%
Origo	3598	100	132	231	551	545	588	670	229	212	29	70	16	3373	94%
NOL	5284	10	56	302	350	441	298	179	383	561	250	0	1	2831	54%
HVG	3350	87	155	184	291	277	496	269	419	145	63	270	128	2784	83%
atlatszo.hu	1853	0	0	0	0	31	89	109	150	249	461	486	232	1807	98%
444	1364	0	0	0	0	0	1	106	144	192	366	374	144	1327	97%
24.hu*	1276	0	0	0	0	25	32	7	236	175	99	342	339	1255	98%
MNO	3824	4	0	2	1	64	189	160	103	52	130	368	72	1145	30%
Népszava	2029	0	1	1	0	2	0	28	92	145	54	296	215	834	41%
ATV	849	0	15	20	4	4	7	6	249	476	3	28	0	812	96%
VS.hu	820	0	0	0	0	0	0	53	420	295	30	5	0	803	98%
Blikk	981	0	0	0	0	58	272	91	98	184	28	7	3	741	76%
napi.hu	712	0	0	0	0	20	17	31	47	269	54	127	19	584	82%
Világ-gazdaság	824	22	52	39	89	61	7	14	15	75	34	13	5	426	52%
Heti Válasz	351	51	53	47	39	38	36	5	11	8	11	33	12	344	98%
Magyar Idők	275	0	0	0	0	0	0	0	0	50	126	36	59	271	99%
mfor.hu	233	1	0	0	0	0	0	0	0	1	3	147	63	215	92%
pestisracok.hu	223	0	0	0	0	0	2	0	39	20	86	50	16	213	96%

** előtte fn.hu, az adatbázisban benne van, de nem került összevonásra*

Minden NLP-elemzés első lépése a nyers szövegkorpuszból egy elemzésre alkalmas numerikus adatbázis előállítása: ezt a lépést előfeldolgozásnak *(pre-processing)* nevezzük. A korrupciós cikkek adatbázisán mi is elvégeztük ezt, a következő lépésekben: mondatok és szavak azonosítása a szövegben *(tokenization),* a tartalmatlan szavak,

5 A teljeskörűség kapcsán fontos megjegyeznünk, hogy azokból a cikkekből indultunk ki, amelyek a K-Monitor adatbázisában benne vannak. Mint minden gyűjtés, természetesen ez az adatbázis is lehet hiányos. Abból a feltételezésből indultunk ki az elemzésben, hogy az összes nagyobb ügy bekerült a K-monitor adatbázisába, és nem volt a készítők részéről semmilyen szándékos torzítás abban, hogy melyik politikai oldalhoz tartozó ügyek kerülhetnek be.

például névelők eltávolítása a szövegből *(stop word removal),* a szótövesítés *(lemmatization),* szófajok és más nyelvészeti kategóriák azonosítása *(part of speech tagging)* és a tulajdonnevek vagy más névelemek felismerése *(named entity recognition).* Az utóbbi lépés célja az volt, hogy az adott tulajdonnév minden előfordulás-változatát ugyanarra a formára hozzuk (például Simicska Lajos és Simicska, Bp és Budapest).

5. Módszerválasztás

A korrupciós diskurzus témáinak automatizált feltárását topikmodellezés segítségével végeztük. A topikmodellezés (Blei & Lafferty 2009) olyan automatizált eljárás, amelynek célja valamely szöveghalmaz látens témáinak azonosítása. A dokumentumok a modell feltevése szerint néhány topik keverékeként azonosíthatók, például egy múzeumfelújításokról szóló cikk 80 százalékban gazdasági, míg 20 százalékban kulturális témát dolgoz fel. Az alkalmazott nyelvi modell az úgynevezett szózsák-modell, azaz a dokumentumoknak csak a szógyakoriságeloszlását vizsgáljuk, a szavak szintaktikai kapcsolatait nem. A topikmodellezés a társadalomtudományokban széles felhasználási spektrummal rendelkezhet: elemezhetjük bármely, a digitális társadalmi térben megjelenő csoport megnyilvánulásainak tematizációját, a témák népszerűségváltozását, a témák tartalmának változását stb. (bővebben a módszerről lásd Németh et al. 2020).

A topikok száma és tartalma *a priori* nem ismert. Akárcsak a (kvantitatív társadalomkutatók számára ismerősebb) k-közép klaszterezés esetén, itt is a modell bemenő paramétere a topikok száma, és az optimális topikszám megválasztása a többféle topikszám mellett illesztett modellek közül a „legjobb" modell kiválasztásán alapul.

A topikmodellezés az utóbbi években gyorsan fejlődött. David M. Blei, Andrew Y. Ng és Michael I. Jordan (2003) írt először a látens Dirichlet-allokációról (*latent Dirichlet allocation,* LDA), amely az egyik legismertebb topikmodellezési eljárás. Az LDA-nak több kiterjesztése létezik, ezek közé tartozik a dinamikus topikmodell *(dynamic topic model),* amely a topikok időbeli változását vizsgálja (Blei & Lafferty 2006). Kutatásunkban ezt a módszert alkalmaztuk.

A dinamikus topikmodell a topikmodellről korábban leírtakat 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 osztályát alkalmazva.

Az időt itt években szegmentáljuk, 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 lehet azonosítható, gyorsan változó és lassan változó téma. Fontos megjegyezni, hogy a modell *chain variance* paramétere korlátozza az adott topik egyik évről következőre történő megváltozását (nem interpretálható és statisztikailag is rosszul definiált lenne a korlátlan változás), amire Blei ajánlását követve az alapértelmezett 0,005-ös paraméterértéket használtuk.[6]

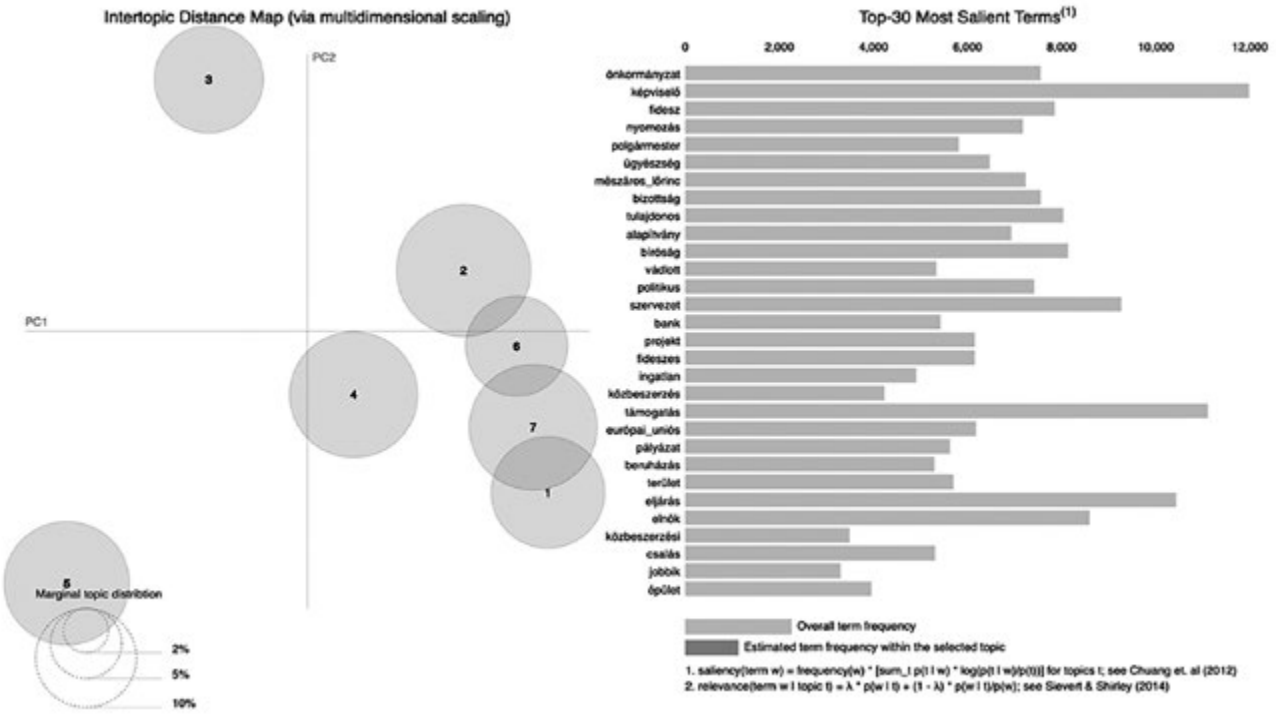
A topikszám a topikmodellek bemenő paramétere, ahogy azt a modellosztály leírása során korábban már említettük, vagyis annak megadása az elemző feladata. Több előelemzést futtatva, a jól interpretálhatóság kívánalmára és a topik-koherencia *(topic coherence)* mutatóra támaszkodva (Stevens et al. 2012) hét topik megkülönböztetése mellett döntöttünk.

6 Blei, a dinamikus topikmodell kifejlesztője a modellt bemutató Google tech talk előadásban tért ki erre, lásd https://www.youtube.com/watch?v=7BMsuyBPx90.

6. Eredmények

A 2. ábrán láthatjuk a hét topik elhelyezkedését egymáshoz képest a 2017-es évben. A grafikon létrehozására a Python pyLDAvis függvényét használtuk (Sievert & Shirley 2014); a korábban hivatkozott online mellékletben elérhető a dinamikus, interaktív verzió is, külön az összes vizsgált évre. Ezt a vizualizáció-sorozatot használtuk arra, hogy megfigyeljük és megértsük a topikok változásának dinamikáját, és interpretáljuk a topikok jelentését (K1, K2, K3).

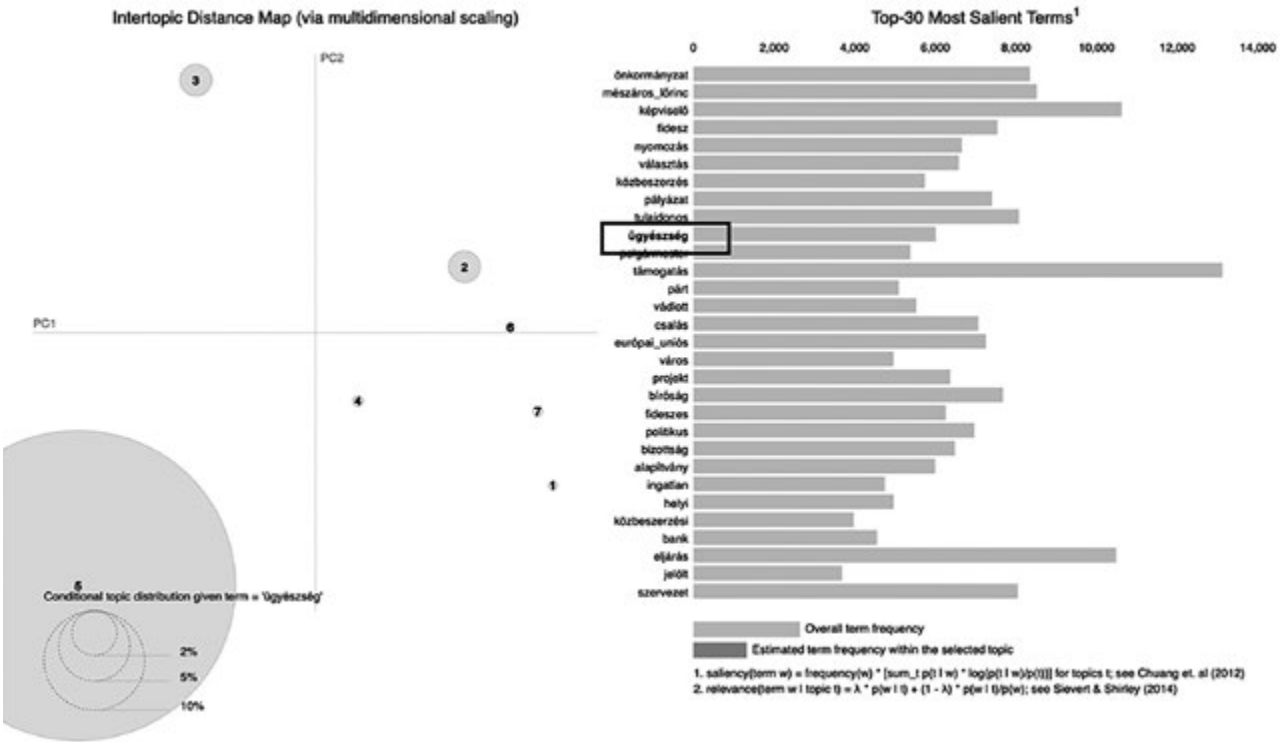
2. ábra
A topikok elhelyezkedése egymáshoz képest a 2017-es évben



A vizualizáción a bal oldalon a topikokat körökkel ábrázoljuk, ahol a körök területe a hozzájuk tartozó szavak számával arányos, a körök középpontjának távolsága pedig a topikok tartalmi távolságának feleltethető meg. A kétdimenziós topik-elhelyezkedés (és ezzel a tartalmi távolság) meghatározása többdimenziós skálázás segítségével történt. A 2. ábra jobb oldalán a 2017-es topikmodell globális jellemzése látható, a 30 leginformatívabb kifejezés felsorolásával. Ez a lista azt mutatja meg, hogy a teljes korpuszon a topikok közötti szóeloszlás-különbséget tekintve mely kifejezések a legfontosabbak. A fontosság mérésére itt használt indikátor nem egyszerűen a szógyakoriság, hanem az informativitás *(saliency),* amelyet a kifejezés relatív gyakoriságának és topik-megkülönböztetési képességének szorzata definiál. Hiába fordul ugyanis elő nagy gyakorisággal egy kifejezés a korpuszban: ha majdnem minden topikban megtalálható (tehát alacsony a megkülönböztető képessége), akkor az adott cikkbeli előfordulása nem segít nekünk abban, hogy eldöntsük, melyik topikhoz tartozik a cikk (Chuang et al. 2012).

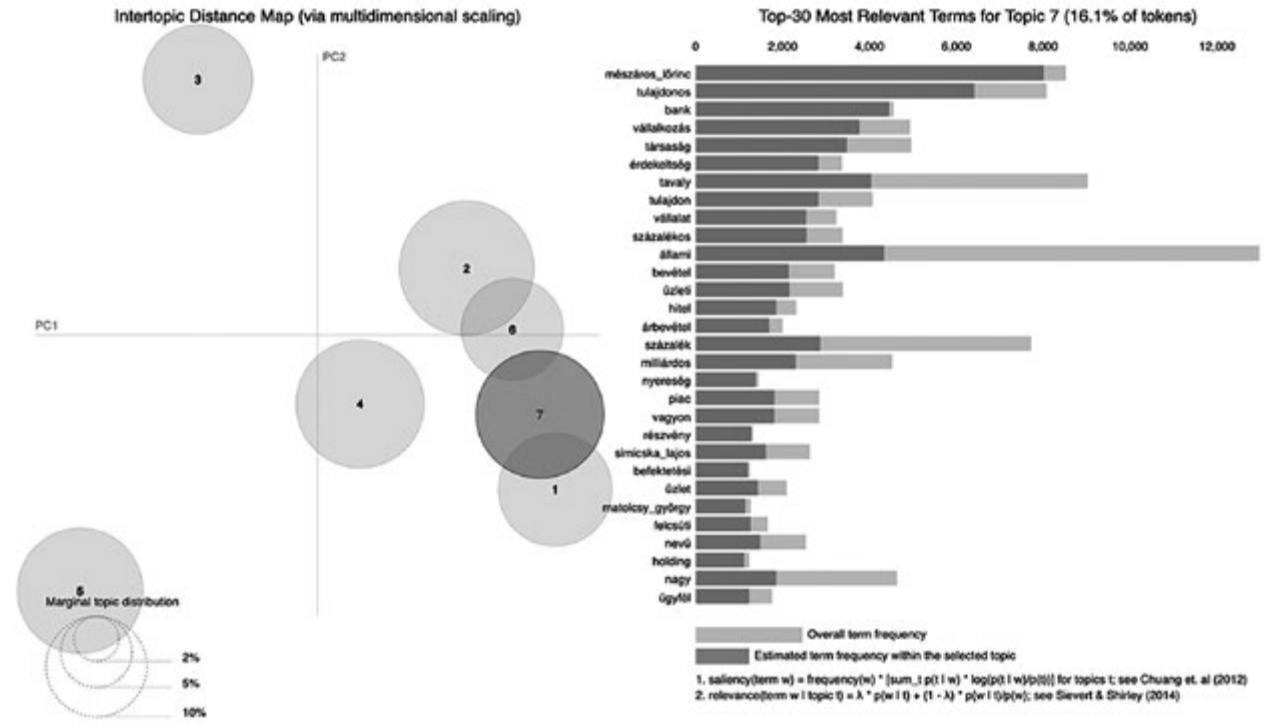
A 2. ábra a korábban hivatkozott linken elérhető interaktív változata egy további, hasznos funkciót tartalmaz, amely nagyban támogatja az értelmezést: a jobb oldalon megjelenített szavak fölé téve a kurzort, a szó előfordulási gyakorisága megjelenik a bal oldalon ábrázolt topikokon (a topikok felett megjelenő új körökkel, ahol a kifejezés topik-specifikus relatív gyakorisága a körök területével arányos, lásd 3. ábrát). Így gyorsan kideríthető például, hogy a 2018-as évet véve az „ügyészség" szinte csak az 5-ös topikban érintett, annak viszont legfontosabb szavai közé tartozik.

3. ábra
Az „ügyészség" szó előfordulási gyakorisága a topikokban 2018-ban



Első lépésben azt vizsgáltuk, hogy mi a jelentése az egyes topikoknak, hogyan azonosíthatók tartalmilag (K1). Ehhez jó kiindulást ad az egyes topikokhoz tartozó legrelevánsabb szavak listája, amelyet a kurzort a topikokat jelölő körök fölé téve kaphatunk meg.

4. ábra
A 7-es topik legrelevánsabb szavai 2018-ban



A 4. ábra jobb oldalán a 2018-as év 7-es topikjához tartozó 30 legrelevánsabb szó topikonkénti eloszlása látható, ahol a relevancia mutató a szó gyakoriságából kapható egy (interaktívan állítható, 0 és 1 közötti) lambda-súlyt használva, az ábrán jelölt képlet segítségével.[7] Azt mondhatjuk, hogy ha közel egyforma hosszú a két horizontális oszlop, akkor az adott topikra exkluzívan jellemző az adott szó. Minél hosszabb a sötétszürke oszlop, annál nagyobb a szó topikspecifikus valószínűsége.

Az ábra elemzése mellett kvalitatív módon is vizsgáltuk a topikok tartalmát, a következő módon. A cikkeket besoroltuk a hét topik egyikébe aszerint, hogy a kialakult témakörök közül melyik a legdominánsabb az adott dokumentumban. Topikonként a legjellemzőbb (legnagyobb dominanciával bíró) cikkek átolvasásával mélyebben is meg tudtuk vizsgálni, hogy milyen cikkek jelennek meg az egyes topikokban.

6.1. Mi a jelentésük az egyes topikoknak, hogyan azonosíthatók tartalmilag? Milyen témák kerülnek elő a hazai sajtóban a korrupció kapcsán (K1)? Hogyan változik az egyes témák belső szerkezete (K3)?

Az 1-es topik egyértelműen a közbeszerzésekkel kapcsolatos ügyeké, a *közbeszerzés* a legrelevánsabb szava, míg e szó más topikban nem is releváns. További jellemző kifejezései a *pályázat, tender, projekt, európai_uniós*. Ide tartozik a *metró* kifejezés is.

A közbeszerzések topikját a budapesti ügyek dominálják, 2013-ig a *főváros* kifejezés meg is jelenik a legrelevánsabb szavak között.

2007-től 2010-ig két nagyon domináns témát fedeztünk fel: a Balatonringgel, valamint a metróépítéssel kapcsolatos cikkeket. A kormányváltással eltűnik a *verseny* kifejezés a releváns szavak közül, és bekerül a *közbeszerzés* szó, amely 2010 után egyre relevánsabb helyet foglal el. 2011 és 2012 között a Közgép, a Margit-híd felújítása, valamint a CET – mára Bálna – beruházása dominálja a cikkeket.

2013. július 1-étől nyitottak meg a dohányboltok, ez visszaköszön a 2013-as év releváns szavai között megjelenő *trafikmutyi*ban. A következő évben pedig *paks* lesz jellemző szó. A sport 2016-tól kezdve minden évben megjelenik a cikkekben: legyen az az Olimpia, a „vizes VB" (A 2017-es úszó-világbajnokság) vagy a stadionépítés témája. 2017-ben *Mészáros Lőrinc* neve is a legrelevánsabb szavak közé kerül.

Az 2-es topik a nemzetközi kapcsolatokat érintő korrupciós ügyeké, jellemző kifejezései a *rendszer, kormány, állam, hatalom,* illetve az *orosz, amerikai, nemzetközi, külföldi, brit, német.* Kezdetben csak a *német, brit, amerikai* kifejezés volt releváns, az előbbi kettő az időszak végére kikopott, míg 2010 táján megjelent az *orosz*, és már 2015-re a topik legfontosabb szavai közé került. Míg 2014-ig jellemzően inkább Oroszországról szóló híreket látunk, egyebek között az ottani korrupciós helyzetről, addig 2014-ben már Oroszország Pakssal kapcsolatban jelenik meg, innentől eltolódik a hangsúly az orosz-magyar kapcsolat irányába.

A topikban jelentős a Transparency Internationalhez (TI) köthető cikkek aránya, akár kutatási jelentések kapcsán, akár Orbán Viktor és a TI viszonyáról legyen szó.

A 3-as topik a pártokhoz és a politikusokhoz köthető esetekkel foglalkozik, jellemző szavai például az *mszp, fidesz, frakció, politikus,* miközben például az *mszp* kifejezés relevanciája más topikok esetén elhanyagolható. Figyelemre méltó a topik dinamikája: míg a vizsgált időszak elején az *MSZP* és *Gyurcsány Ferenc* szerepelt a legrelevánsabb kifejezések között, addig az évek során *Orbán Viktor, Lázár János, Rogán Antal* és a *Jobbik* is megjelent, míg *Gyurcsány Ferenc* eltűnt. Az időszak végére a *Fidesz* került a legrelevánsabb szavak közé, a *Jobbik*

7 Adott szó relevanciája – figyelembe véve azt, hogy a nagyon gyakori szavak kevésbé relevánsak a vizsgált topik szempontjából – egy adott topikon belül a szó topikspecifikus gyakoriságából kapható, úgy, hogy az „büntetve" van a szó általános, nemtopik specifikus gyakoriságával. Minél nagyobb a lambda, annál kevésbé büntetjük a szó általános gyakoriságát. Ha a lambda 1, nem büntetünk, és a nyers topikspecifikus szógyakoriságot használjuk relevancia-mutatóként. Mi a lambda=0,6-os beállítást választottuk, amelyet jól interpretálhatónak láttunk, és amely Carson Sievert és Kenneth Shirley (2014) elemzésében is optimális értékként adódott.

mögötte kissé lemaradva, s az *MSZP* jóval hátrébb sorolódott. Mindez megfelel a 2010-es kormányzópárt-váltásnak (a korrupciós ügyek nyilván az elosztási mechanizmusokhoz jobban hozzáférő kormánypártokhoz köthetők).

A 4-es topik a kormányzati és a nem-kormányzati szervezetek kapcsolatát érintő cikkeket tartalmazza. Legrelevánsabb szavai között van a *minisztérium, tárca, közigazgatási, államtitkár* (ezek a kifejezések más topikokban nem szerepelnek lényegesen), az *adat, nyilvánosság,* illetve az *alapítvány, civil, egyetem, közérdekű.* A tartalmi dinamikára jellemző, hogy a *civil* kifejezés 2012-ben megjelenik, és egyre fontosabbá válik a topikban.

Az 5-ös topik azokat a cikkeket fogja össze, ahol a korrupciógyanús eset az igazságszolgáltatás valamelyik lépcsőfokáig eljutott már. Jellemző szavai a *nyomozás, ügyészség, bíróság, vallomás, vádlott, ítélet*. A cikkek különböző ügyekben hozott jogerős ítéleteket, vádakat, különböző pereket tárgyalnak. Fő témái a letartóztatások, a pénzmosás, a gyilkosság, az adócsalás és a kémkedés. A későbbi elemzés kapcsán még fontos lesz, hogy 2016–2017-ben nagyon sok cikk származik a Pesti Srácokról, míg más topikokban nem volt jellemző az oldalról származó cikkek túlsúlya. Emellett a tulajdonosváltás utáni Origó is fontos forrás a topik szempontjából. A 3-as topikban megjelenő szereplők (Zuschlag, Wieszt, Simon, Hagyó) ügyei ebben a topikban is jelentősek, itt is megjelennek, de ebbe a topikba főként a jogi eljárásokkal kapcsolatos cikkek kerültek.

A 6-os topik az önkormányzati szintű, illetve lokális, helyi kötődésű ügyeké. Összességében tartalmilag ez a legkevésbé változó topik. Jellemző szavai az *önkormányzat, polgármester, város, helyi.* Az érintett ügyeket jól jellemzi a legrelevánsabb szavak között szereplő *ingatlan, négyzetméter, lakás, bérlő, épület, föld*. 2010 után a *kerület* helyét a *föld* veszi át a releváns szavak listájának elején. A *vagyonnyilatkozat* kifejezés a kormányváltás évében jelenik meg, innentől kezdve nagyon sok cikk a polgármesterek és az önkormányzati képviselők vagyonnyilatkozatairól szól.

A 7-es topik a vállalkozásokkal, vállalatokkal kapcsolatos korrupciós diskurzus csoportja. Jellemző kifejezései a *vállalkozás, vállalat, társaság, holding, piac, vagyon, nyereség, részvény, tulajdon, érdekeltség.* A legjellemzőbb és legerősebb fordulat ebben a topikban történik: 2014-ig kizárólag Simicska Lajos szerepel személynévként, 2015-re Mészáros Lőrinc jelenik meg, rögtön a legrelevánsabbak közé ugorva. Ugyanekkor Matolcsy György is megjelenik új kifejezésként (2015-ben még csak a 25., míg 2016-ban már a 3. legrelevánsabb kifejezés), míg Simicska Lajos lecsúszik kettejük mögé. 2018-ra Mészáros Lőrinc neve a legrelevánsabb szó, messze kimagaslóan a többi szó közül.

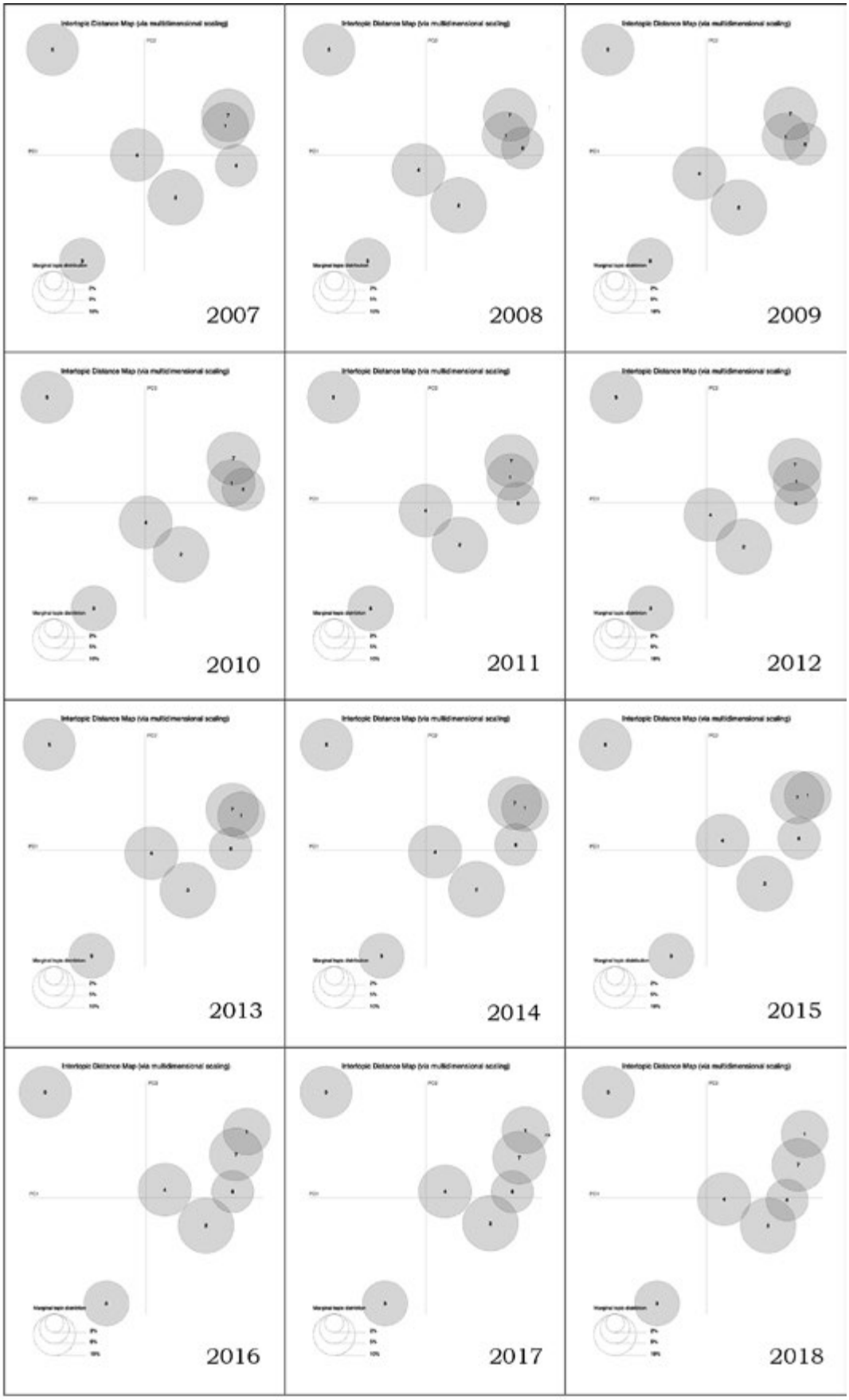
A többivel szemben ez a topik tartalmilag instabil. Ebben a topikban igen jellemző, hogy felbukkannak, majd eltűnnek szavak, szereplők. A kifejezések pozíciója sem állandó, és a kifutásuk sem olyan lassú, mint a többi topik releváns szavainak. Azaz ha adott évben nagy relevanciája is van egy kifejezésnek, a következő évre eltűnhet a legrelevánsabb szavak közül.

6.2. Hogyan kapcsolódnak egymáshoz a topikok, melyek azok, amelyek tartalmilag határozottan elkülönülnek a többitől (K2)? Mely témák kerülnek egymástól távol, melyek közelednek egymáshoz (K4)?

Az 5. ábra a többdimenziós skálázás eredménye alapján a topikok pozícióváltozását mutatja a vizsgált időszakban.

Láthatjuk, hogy a 3-as topik (pártokhoz, politikusokhoz köthető ügyek) és az 5-ös topik (igazságszolgáltatás) 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 2-es és 4-es (a nemzetközi ügyek, illetve a kormányzati és nem kormányzati szervek kapcsolatát vizsgáló topik), valamint az 1-es (közbeszerzések), 6-os (önkormányzati szintű ügyek) és 7-es (vállalkozások, vállalatok) topik csoportja, de az évek során e két csoport elkülönülése megszűnik, illetve a korábban ugyanazon csoporthoz tartozó topikok (például közbeszerzések és vállalkozások, vállalatok) eltávolodnak egymástól.

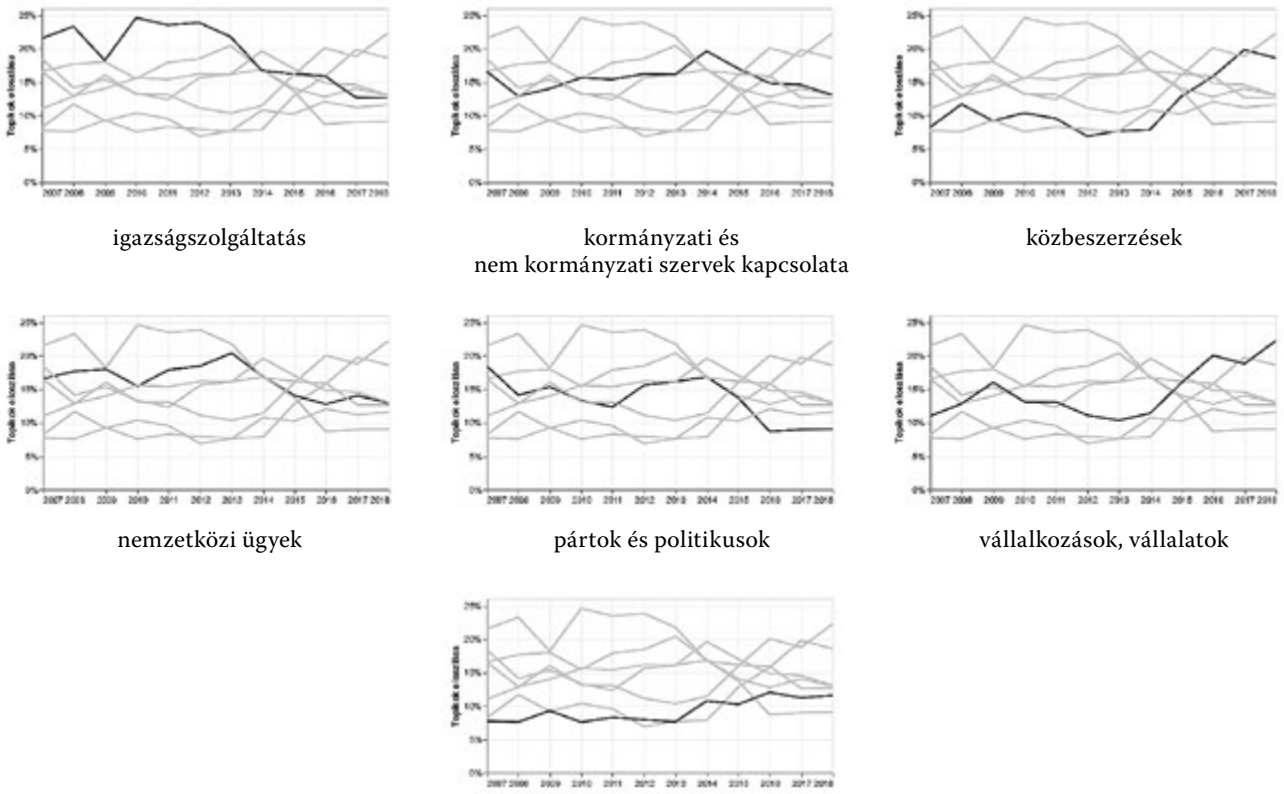
5. ábra
A topikok pozíciójának változása 2007 és 2018 között



6.3. Milyen arányban jelennek meg az egyes topikok?

Az 5. ábrán a körök területével jelölt „topik-nagyság" változási dinamikáját fogja egybe a 6. ábra. Az ábra a teljes cikkgyűjteményt vizsgálja, itt is fontos azonban szem előtt tartanunk, hogy korpuszunkat csupán a 2010-es évek vége felé tekinthetjük teljesnek.

6. ábra
Topikok eloszlása az adott évben letöltött összes cikkben



igazságszolgáltatás | kormányzati és nem kormányzati szervek kapcsolata | közbeszerzések

nemzetközi ügyek | pártok és politikusok | vállalkozások, vállalatok

önkormányzati szintű ügyek

A teljes időszakra csupán öt portálról (Index, Origó, HVG, vilaggazdasag.hu, hetivalasz.hu) áll rendelkezésre elemezhető tartalom. Több jelentős portál, mint az Átlátszó, a 444, a vs.hu és a Pesti Srácok csak 2010 után jött létre, illetve több olyan portál is van, ahonnan a régebbi cikkeket nem (vagy csak alacsony arányban) tudtuk letölteni.

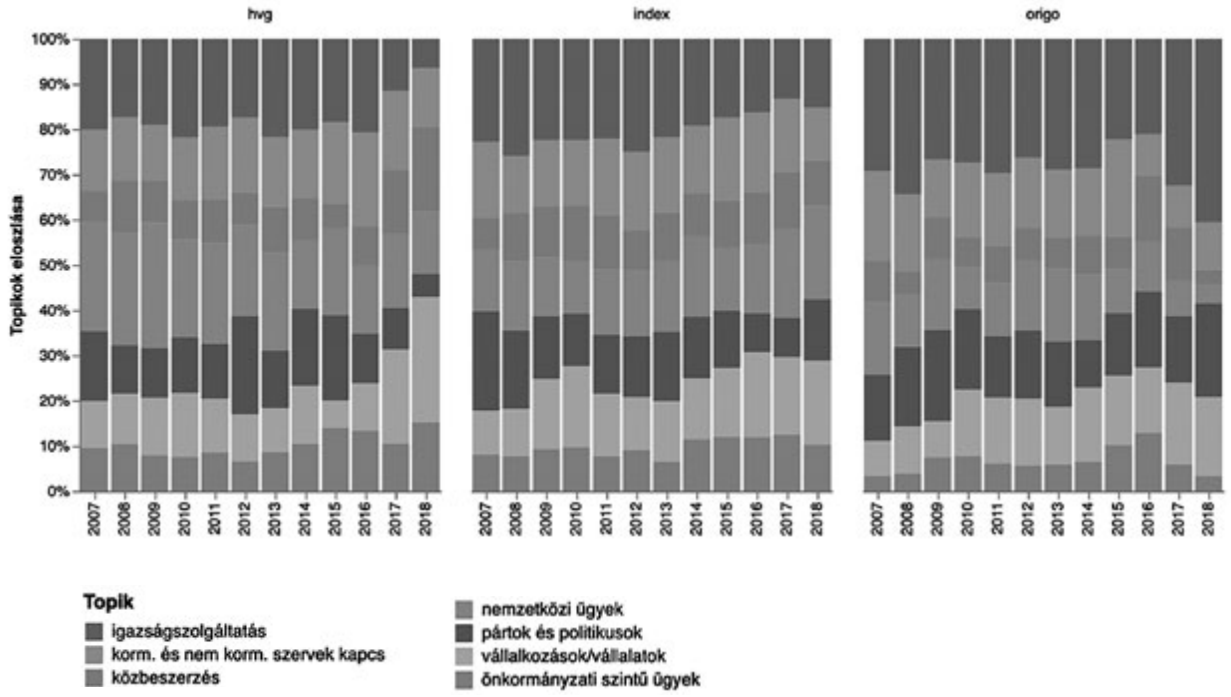
A 6-os ábra alapján látszik, hogy míg az igazságszolgáltatás topik a kezdeti időszakban kiemelkedően fontos volt, a jelentősége az időszak második felében zuhan. Ez vélhetően egyebek között annak köszönhető, hogy a fent bemutatott ügyek (a Zuschlag-, a Wieszt-, a Simon- és a BKV-ügy) jó része lezárul, kisebb jelentőségűvé válik a vizsgált periódus második felében. A kormányváltás után a nemzetközi ügyek aránya megugrik. Két kiugró és egymással szinkron változás még, hogy a közbeszerzések és a vállalkozások, vállalatok topik 2014-ben láthatóan megnő, és ezzel egy időben a 2-es (kormányzati és nem kormányzati szervek kapcsolata) topik aránya (ami elsősorban a civil szervezetekkel kapcsolatos ügyeké), valamint a pártok és politikusok ügyeit feldolgozó topik aránya rohamosan csökken. A közbeszerzések topik aránya 2015-ig 15 százalék alatti, 2018-ra azonban a második legdominánsabb topikká válik. Az önkormányzati szintű ügyek aránya viszonylag konstans.

A következőkben azt a három portált (Index, Origó és hvg.hu) vizsgáljuk meg részletesebben, amelyek esetében a legnagyobb arányú az elérés (az adatbázisban szereplő cikkek közül a legtöbb elérhető volt), a gyűjtés pedig (közel) teljes körű (lásd a 2. táblázatot és a 7. ábrát).

2. táblázat
Elért cikkek száma (Index, Origo, HVG)

	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018	összesen
Index	328	399	428	486	525	395	319	241	291	439	629	240	4720
Origo	100	132	231	551	545	588	670	229	212	29	70	16	3373
HVG	87	155	184	291	277	496	269	419	145	63	270	128	2784

7. ábra
Topikok eloszlása az adott évből letöltött összes cikkben (hvg.hu, Index, Origó)



A 7. ábra alapján látható, hogy a hvg.hu és az Index korrupciós cikkeinek tematizáltsága egymáshoz hasonló, az Origó azonban – főleg a tulajdonosváltás után – eltér tőlük (tanulmányunk későbbi részében külön vizsgáljuk az Origo tulajdonosváltásának hatását). A leglátványosabb eltérés az, hogy az igazságszolgáltatásra fordított hangsúly az Origó esetén elkezd meredeken nőni 2017-ben, míg a másik két portálnál csökken. Ugyanez mondható el a pártok és politikusok topikról.

A hvg.hu és az Index esetében a dinamika hasonló, ám az Index esetében jóval konstansabb képet figyelhetünk meg. A hvg.hu-n 2013-tól meredeken csökken a nemzetközi ügyek aránya, majd 2015-től a pártok és politikusok topik aránya is, ezzel párhuzamosan pedig növekedni kezd a közbeszerzések és a vállalkozások/vállalatok topikok aránya. Az Origón 2016-ban megugrik a közbeszerzések topik aránya, ami a 2015-ös Simicska-botrány hozadékaként értelmezhető. A kiemelt eredmények jól mutatják, hogy jelentős különbséget mutatnak az oldalak annak tekintetében, hogy milyen típusú korrupciós ügyeket mutatnak be.

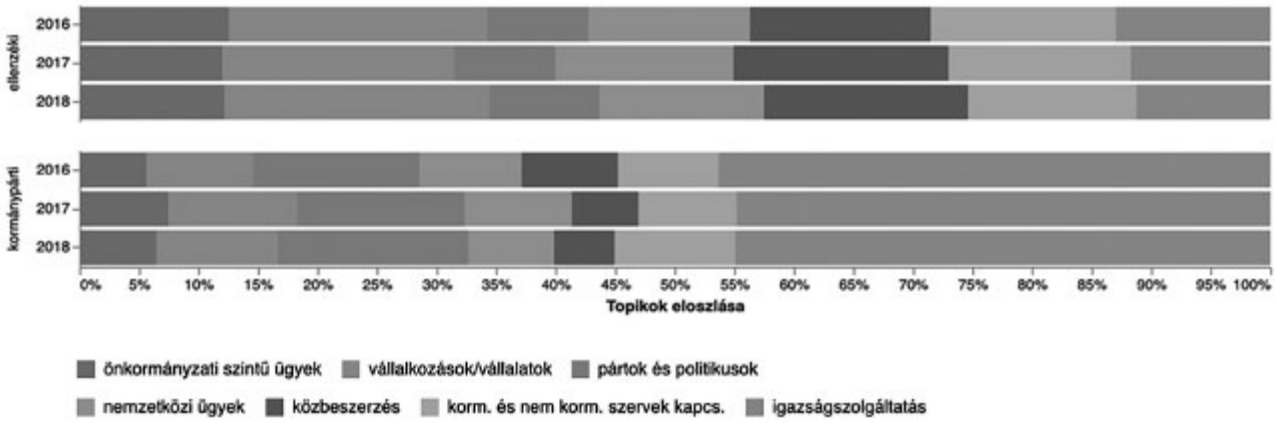
6.4. Kampányidőszak és a kormányzathoz való viszony (K5, K6)

A következő elemzésben csak a 2016 és 2018 között megjelent cikkeket vizsgáljuk. Azért szűkítettük erre az időszakra az elemzést, mert erre az időszakra áll rendelkezésünkre megfelelő mennyiségű szöveg kormánypárti és ellenzéki oldalról is (lásd az 1 táblázatot).

Azt vizsgáltuk, mely témákról ír az ellenzéki média (atlatszo.hu, fn.hu, 24.hu, HVG, Index, magyarnarancs.hu, mno.hu, Nepszava, nol.hu, 168ora.hu, 444.hu), illetve melyekről a kormánypárti média (figyelo.hu, magyarhirlap.hu, magyaridok.hu, Pesti Srácok, Origó), és van-e kapcsolat a tematika és az ellenzéki/kormánypárti pozíció között (lásd a 8. ábrát). 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 ellenzéki média), illetve a cikk megjelenésének időpontja (kampányidőszakban vagy azon kívül jelent-e meg a cikk). A kampányidőszakot a választás előtti négy hónapként definiáltuk.

8. ábra

Topikok eloszlása a 2016–2018 között készült összes cikkben az ellenzéki, illetve a kormánypárti sajtóban



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 (lásd a 3. táblázatot).

Az összes modell szignifikáns az F-próba alapján, ám a modellek magyarázóereje 7-ből 6 esetben nagyon alacsony, 1 százalék körüli, tehát a tematikát (kézenfekvő módon) más fontos tényezők is magyarázzák a két bevont változó mellett. Egyetlen esetben, az igazságszolgáltatás topik esetében beszélhetünk 12 százalékos magyarázóerőről – eszerint a már igazságszolgáltatási szakaszba került korrupciós esetek megjelenítését erősen befolyásolja, hogy kormánypárti-e az adott médium, és hogy kampányidőszak van-e éppen. A tábla alapján azt mondhatjuk el, hogy míg a kormánypárti/ellenzéki 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 kap.

A kormánypárti és ellenzéki bontást vizsgálva azt mondhatjuk, hogy a kormánypárti médiában a pártokhoz, politikusokhoz köthető ügyek aránya, valamint az igazságszolgáltatás aránya szignifikánsan magasabb (az utóbbi közel 50 százalékos), a többi topik viszont az ellenzéki médiában kap jelentősebb szerepet. Ezt a 8. ábra is jól szemlélteti.

3. táblázat
A lineáris regressziós modellek eredményei

Függő változó	Magyarázó változók	Adj. R^2	Koefficiens	t-próba (p)
Igazságszolgáltatás	korm_ell	0,123	-0,335	0,000
	kampány		-0,0093	0,205
Kormányzati és nem kormányzati szervek kapcsolata	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, ellenzéki = 1, illetve kampány esetében nem kampányidőszak = 0, kampányidőszak = 1

6.5. A tulajdonosváltás hatása: a korrupciós tematika változása ugyanazon lapon belül (K7)

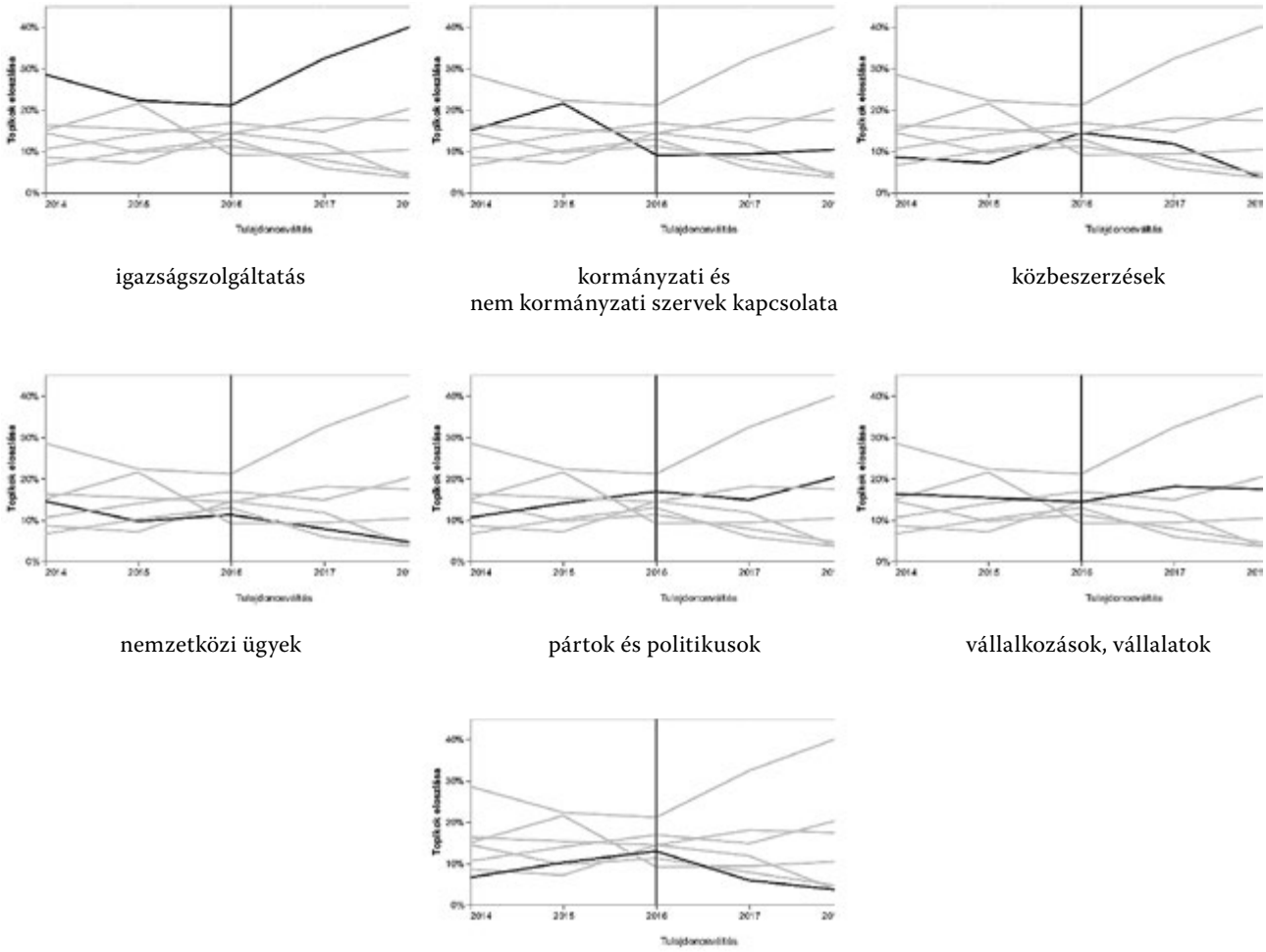
Az utolsó kutatási kérdésünk azt vizsgálja, hogyan hat egy tulajdonosváltás egy lap, jelen esetben az Origó korrupciós tematikájára. Az Origó körüli változások 2014 júniusában kezdődtek, amikor Sáling Gergő főszerkesztő távozott, és helyét Pálmai L. Ákos vette át. A főszerkesztőváltásra az országgyűlési választás után került sor. 2015-ben a Telekom eladásra bocsátotta az Origót, amelyet 2016 februárjában a New Wave Media Kft vett meg, és az új tulajdonos Száraz István lett. 2017 áprilisa is fontos dátum a híroldal életében: többen távoznak az újságtól, és ekkor szűnik meg az oldalon a kommentelési lehetőség is (Facebookon továbbra is lehet kommentelni a cikkeiket).

4. táblázat
Az Aktákban megtalálható, illetve sikeresen letöltött cikkek száma (Origó)

origo.hu												
	2007	2008	2009	2010	2011	2012	2013	2014	2015	2016	2017	2018
Total	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 kérdésre, hogy a tulajdonosváltás hatása kimutatható-e egy portál esetében, a 4. tábla alapján részben már válaszolhatunk. A K-Monitor cikkgyűjteményébe 2016-tól kezdődően jóval kevesebb cikk került be, mint a tulajdonosváltás előtt (lásd a 9. ábrát).

9. ábra
Topikok eloszlása az adott évből letöltött összes cikkben
2014 és 2018 között (Origó)



igazságszolgáltatás | kormányzati és nem kormányzati szervek kapcsolata | közbeszerzések

nemzetközi ügyek | pártok és politikusok | vállalkozások, vállalatok

önkormányzati szintű ügyek

Ha a portálon megjelenő topikok arányát vizsgáljuk, a 9. ábra is alátámasztja a tulajdonosváltás befolyásoló erejét. Bár mindkét időszakban domináns az igazságszolgáltatás topik jelentősége, ez a tulajdonosváltás előtt csökkenő tendenciát mutat, a tulajdonosváltással viszont azonnal megugrik (magasabb szintre, mint ahonnan 2014-ben indult). Figyelemre méltó még a közbeszerzések topik visszaesése is a tulajdonosváltás előtti és utáni időszak összevetésében, ez a változás azonban a varianciaanalízis alapján statisztikailag nem szignifikáns. Szignifikáns eltérést három topikban látunk: az igazságszolgáltatás, a kormányzati és nem kormányzati szervek kapcsolata és a nemzetközi ügyek topikban. Ezek a változások nagyságukat tekintve is jelentősnek ítélhetők.

5. táblázat
Az ANOVA és Welch-próba eredményei

Függő változó	Tulajdonosváltás	Átlag	Próba	p-érték
Igazságszolgáltatás	előtt	0,275	ANOVA	0,000
	után	0,415		
Kormányzati és nem kormányzati szervek kapcsolata	előtt	0,162	Welch	0,000
	után	0,065		
Közbeszerzések	előtt	0,075	ANOVA	0,649
	után	0,067		
Nemzetközi ügyek	előtt	0,134	Welch	0,000
	után	0,069		
Pártok és politikusok	előtt	0,152	ANOVA	0,464
	után	0,168		
Vállalatok, vállalkozások	előtt	0,136	ANOVA	0,487
	után	0,149		
Önkormányzati szintű ügyek	előtt	0,065	ANOVA	0,953
	után	0,066		

Magyarázat: az ANOVA alternatívájaként Welch próbát alkalmaztunk, ha az ANOVA szórás-azonosságra vonatkozó alkalmazási feltétele nem teljesül

Összegzés

Kutatásunkban elsősorban arra a kérdésre kerestük a választ, hogy milyen látens témák azonosíthatók a magyar online sajtó korrupciós diskurzusában. Az azonosított hét topik alapján az igazságszolgáltatás, a kormányzati és a nem kormányzati szervek kapcsolata, a közbeszerzések, a nemzetközi ügyek, a pártok és a politikusok, a vállalatok, vállalkozások és az önkormányzati szintű ügyek kerülnek elő a hazai sajtóban a korrupció kapcsán. A kapott hét topik tartalmilag jól elkülöníthető, könnyen interpretálható. Külső információkkal (mint a 2010-es kormányváltás és Simicska Lajos 2015-ös háttérbe szorulása) érvényességük jól alátámasztható.

A pártokhoz, politikusokhoz köthető ügyek és az igazságszolgáltatás topikja időben stabilan elkülönül a többitől, míg a nemzetközi ügyeket, illetve a kormányzati és nem kormányzati szervek kapcsolatát vizsgáló topik csak kezdetben alkot különálló szigetet. Az idő multával a nemzetközi ügyek topik leválik az utóbbiról, és közel kerül a közbeszerzések, az önkormányzati szintű ügyek és a vállalkozások, vállalatok topik alkotta csoporthoz. Az utóbbi hármas kezdeti közelsége, összefonódása az idő során lazul.

Az általunk vártnál nehezebben volt vizsgálható az a kutatási kérdés, hogy milyen arányban jelennek meg az egyes topikok, mert az adatbázisban a 2013 előtti időszakból származó cikkek lefedettsége nem teljes. Ennek ellenére a topikok tartalmának és egymáshoz való viszonyának elemzése értelmezhetőnek tűnt, és a külső megfigyelésekkel jó egyezést mutat. Fontos adalék, hogy a kormányzathoz való viszony csupán a 2013 utáni időszakra volt értelmezhető. Azt láthattuk, hogy a kampányidőszak hatása a legritkább esetben függ azzal össze, hogy milyen témafókusz van egy lapban, míg a kormánypárti, illetve ellenzéki besorolás a legtöbb esetben hatást mutat. A kormánypárti médiában a pártokhoz, politikusokhoz köthető ügyek topikjának aránya, valamint az igazságszolgáltatás topikjának aránya szignifikánsan magasabb, a másik öt topik viszont az ellenzéki médiában kap jelentősebb szerepet. Vizsgáltuk azt is, hogy milyen változások figyelhetők meg, ha egy lap ellenzéki pozícióból (tulajdonosváltás révén) kormánypárti pozícióba kerül. Azt mondhatjuk, hogy a tulajdonosváltás után az Origón azonnal megváltozott mind a korrupcióval kapcsolatos cikkek mennyisége, mind tartalmi fókuszuk.

Az itt bemutatott szövegbányászati eszköztár azonban nemcsak újságcikkeken alkalmazható, hanem más típusú adatokon is, mint például sajtóhír-kommenteken vagy közösségimédia-posztokon. Az ezekre az adatforrásokra támaszkodó elemzések a lakossági korrupciós észleléssel is jobban rezonálnak. Ahogy az „átlagemberek" is egyre inkább hírkurátorokká válnak az egyszerű hírfogyasztó helyett, annál inkább válik relevánssá a közösségi médiában megjelenő vélemények letapogatása a korrupciós percepciók kapcsán – bár ez a tartalom kétségkívül nem független a szerkesztett médiában megjelenő információktól.

A bevezetőben említettekhez visszatérve fontos megjegyeznünk: az, hogy a médiában mennyi és milyen korrupciós ügy jelenik meg, a korrupció tényleges mértékén kívül számos tényezőtől függ, kezdve a korrupciós ügyek felderítéséhez szükséges erőforrások nagyságától akár addig, hogy milyen egyéb nagy témák tematizálják a közbeszédet (például a COVID). Ugyanakkor az online tér adatai jól használhatóak mind a különböző jellemző korrupciós típusok, mind e típusok időbeli változásának körvonalazására. Az ilyen típusú felhasználást támasztja alá Sofia Axelsson és Stefan Dahlberg (2018) már idézett kutatása is, amelyben a szerzőknek online szerkesztett és közösségimédia-szövegekre támaszkodva specifikus NLP (szemantikus disztribúciós) modellekkel sikerült feltárniuk az adott országok korrupciós mintáit. Médiakutatási szempontból annak is kiemelt jelentősége van, hogy kutatható, a különböző pártállású oldalak milyen korrupciós témákat, típusokat vesznek elő.

Az eredményeink jól mutatják, hogy jelentős különbség van a vizsgált portálok között abban, hogy milyen típusú korrupciós ügyeket mutatnak be. Ez a különbség vélhetően még nagyobb a konkrét ügyek kiválasztásában vagy az egyes ügyek keretezésében. Tanulmányunk korlátozott terjedelme miatt az ilyen típusú mélyebb elemzéseket már nem tudjuk ebben az írásban bemutatni. A szövegkorpusz további – akár kvantitatív, akár kvalitatív – vizsgálata azonban számos új eredményt hozhat a témában.

Összességében az alkalmazott módszer jól használhatónak bizonyult nagy szöveges adatbázisok tematikus struktúrájának automatizált azonosítására, ami előrevetíti azt, hogy potenciálisan más médiakutatási területeken is jól alkalmazható lehet. E megközelítés jelentőségét nehéz volna túlbecsülni: a szövegbányászat elsősorban a növekvő digitizáció miatt válik egyre fontosabb vizsgálati tereppé, hiszen a digitális szövegek fontos lenyomatai az emberek gondolkozásának, attitűdjeinek. A szövegelemzés másik nagy előnye, hogy nemcsak a jelenről tud szólni, hanem a múltról is képes megállapításokat tenni – hiszen a digitális média már Magyarországon is két évtizedes múltra tekint vissza, a korábbi szövegek digitalizációja pedig még szélesebb távlatokat nyithat. Az újságírói közléseken keresztül képesek lehetünk olyan képet rajzolni, amely legalább a professzionális diskurzus szintjén megmutatja a felmerülő korrupciós tematikákat.

Irodalom

Axelsson, Sofia & Stefan Dahlberg (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, David M. & John D. Lafferty (2009): Topic models, http://www.cs.columbia.edu/~blei/papers/BleiLafferty 2009.pdf.

Bohn, Simone R. (2012): Corruption in Latin America: Understanding the Perception–Exposure Gap, *Journal of Politics in Latin America,* vol. 4, no. 3, pp. 67-95.

Chuang, Jason, Christopher D. Manning & Jeffrey Heer (2012): Termite. Visualization Techniques for Assessing Textual Topic Models. *Proceedings of the international working conference on advanced visual interfaces,* pp. 74–77, http://vis.stanford.edu/files/2012-Termite-AVI.pdf.

Fazekas, Mihály, István János Tóth & Lawrence King (2016): *An Objective Corruption Risk Index Using Public Procurement Data.* Springer Sciencet Business Media Dordrecht.

Hajdu Miklós, Pápay Boróka, Szántó Zoltán & Tóth János István (2018a): A korrupció sajtómegjelenése nemzetközi összehasonlításban. *Magyar Tudomány,* 179. évf. 4. sz. 496–506. o.

Hajdu Miklós, Pápay Boróka, Szántó Zoltán & Tóth János István (2018b): Content analysis of corruption coverage: Cross-national differences and commonalities. *European Journal of Communication,* vol. 33, no. 1, pp. 7–21.

Hlatshwayo, Sandile, Anne Oeking, Manuk Ghazanchyan, David Corvino, Ananya Shukla & Lamin Y Leigh (2018): *The Measurement and Macro-Relevance of Corruption: A Big Data Approach.* IMF Working Papers, https://www.imf.org/en/Publications/WP/Issues/2018/08/31/The-Measurement-and-Macro-Relevance-of-Corruption-A-Big-Data-Approach-46157.

Katona Eszter & Németh Renáta (2021, megjelenés alatt): Automatizált szöveganalitika a korrupció kutatásában. *Socio.hu: Társadalomtudományi Szemle.*

Kmetty Zoltán (2018): Incumbent party support and perceptions of corruption. An experimental study. *Szociológiai Szemle,* 28. évf. 4. sz. 152–165. o.

Kostadinova, Tatiana & Kmetty Zoltán (2018): Corruption and Political Participation in Hungary: Testing Models of Civic Engagement. *East European Politics and Societies,* vol. 33, no. 3, pp. 555–578.

Li, Jiawei, Wen-Hao Chen, Quing Xu, Neal Shah & Timothy Mackey (2019): Leveraging Big Data to Identify Corruption as an SDG Goal 16 Humanitarian Technology. *IEEE Global Humanitarian Technology Conference (GHTC),* Seattle, WA, USA, pp. 1–4, https://ieeexplore.ieee.org/document/9033129

Mear, Camilla (2016): An International NGO Startup's Use of Social Media Technology. The Case of Anti Corruption International: A Discursive Analysis on the Organizational Use of the Term 'Corruption' on Twitter. *Master's Thesis,* https://ntnuopen.ntnu.no/ntnu-xmlui/bitstream/handle/11250/2449754/Camilla%20Mear%20-%20MSGLOPOL%20-%202016.pdf?sequence=1&isAllowed=y.

Németh Erzsébet, Körmendi Gábor & Kiss Beatrix (2011): Korrupció és nyilvánosság. A média hatása a korrupcióra és annak társadalmi megítélésére. *Pénzügyi Szemle,* 56. évf. 1. sz. 57–65. o.

Németh Renáta, Katona Eszter & Kmetty Zoltán (2020): Az automatizált szövegelmezés perspektívája a társadalomtudományokban. *Szociológiai Szemle,* 30. évf. 1. sz. 44–62. o.

Niklander, Stefanie, Ricardo Soto, Broderick Crawford, Claudio León de la Barra & Eduardo Olguín (2016): Facilitating Analysis of Audience Reaction on Social Networks Using Content Analysis: A Case Study Based on Political Corruption. In: Stephanidis, Constantine (ed.): HCI Internatinal 2016 – Poster's Extended Abstracts. HCI 2016. *Communications in Computer and Information Science*, vol. 618. Springer, Cham, https://link.springer.com/chapter/10.1007/978-3-319-40542-1_10.

Sievert, Carson & Kenneth E. Shirley (2014): LDAvis: A method for visualizing and interpreting topics. *ACL Workshop on Interactive Language Learning, Visualization, and Interfaces, Baltimore,* https://www.aclweb.org/anthology/W14-3110.pdf.

Stevens, Keith, Philip Kegelmeyer, David Andrzejewski & David Buttler (2012): Exploring topic coherence over many models and many topics. In: *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL '12). Association for Computational Linguistics,* pp. 952–961, https://www.aclweb.org/anthology/D12-1087.pdf.

Suphachalasai, Suphachalasai (2005): Bureaucratic Corruption and Mass Media. *Environmental Economy and Policy Research Working Papers,* University of Cambridge, Department of Land Economics.

Tóth János István & Hajdu Miklós (2018): A korrupció mérési lehetőségei. Lehetséges objektív indikátorok bemutatása egy magyar példa alapján. *Magyar Tudomány,* 179 évf. 4. sz. 507–528. o.

Touwe, Mochtar & M. Iqbal Sultan (2015): The Investigation of Tempo Weekly News Magazine in the Corruption Case of the Megaproject Hambalang Sport Facilities. *International Journal of Sciences: Basic and Applied Research,* vol. 20, no. 2, pp. 233–239.

Függelék

A korpusz forrásoldalainak listája: Index, Origó, NOL, HVG, atlatszo.hu, 444, 24.hu, MNO, Népszava, ATV, VS.hu, Blikk, napi.hu, Világgazdaság, Heti Válasz, Magyar Idők, mfor.hu, pestisracok.hu, fn.hu, HírTv, G7, 168 óra, Klubrádió online, bama.hu, Hírextra, Portfolio, Magyar Hírlap, Vasárnapi Hírek, Átlátszó Oktatás, Bors, délmagyar, Jogi Fórum, cink, abcúg, TASZ, Élet és Irodalom, Kitekintő, orszagszerte.atlatszo.hu, K-Monitor, Direkt36, Kisalföld, Nemzeti Sport, bács-kiskun online, Komment.hu, romnet.hu, Privátbankár, valasz.hu, RTL Klub, HEOL, vasarnapihirek.hu, KEMMA, Mandiner, Pécsi Újság, nyugat.hu, SZOLJON, Sonline, alfahir.hu, Forbes, Hetek, Mérce, TV2, haon.hu, Szabad Pécs, BOON, Magyar Hang, Qubit, Szabad Föld, Szabolcs Online,

Dunántúli Napló, BEOL, Napi Gazdaság, agrárszektor.hu, ÁSZ, BAON, Korrupcióinfó.hu, MIX, PécsMa.hu, Beszélő, ma.hu, Figyelő, Inforádió, Magyar Narancs, dontwasteit.hu, Dunaújváros Online, építészfórum, ezahir.hu, focitipp.hu, HWSW, IHO.hu, IT.news, ittlakunk.hu, Jobbik.hu, Magyar Építők, Médiapiac, mindennapi.hu, mmonline, NRG Report, Nyelv és Tudomány, nyugatmagyar.hu, Privátkopó, Propeller, Revizor, Szoboszlói Színes, Ténytár, totalcar, Transindex, Urbanista (Index), valasztasirendszer.hu, Város Szíve blog, vezess.hu, Zircinfo, Kárpátinfo, Reflektor, Észak Online, Hir6, Gazdasági Rádió, Zalai Hírlap, Pécsi Stop, TEOL, Kreatív, Hírszerző, Bank és Tőzsde, Eszmélet, HavariaPress, rakosmente.fidesz.hu, romaweb.hu, stop.hu, ZOOM, Magyar Rádió, MTV, msn, Manager Magazin, Vas Népe, Pénzcentrum, zóna, ingatlanmagazin.com, globusz.net, DH-online, Fejér Megyei Hírlap, samsungsport.hu, DunaTV, hirado.hu, Új Szó (Szlovákia), 0m2, Deutsche Welle, Mozgó Világ, NPORT, Pwc.com, rakosmente.hu, Somogy TV, Szeged Kurir, turizmusonline.hu

Abstract in English

Applying natural language processing to analyise the representation of corruption in the Hungarian online media

This paper presents a thematic analysis of the representation of corruption in the Hungarian online media, using a text mining tool called dynamic topic modeling. The text corpus was provided by K-Monitor and includes online articles on corruption and issues related to the misuse of public funds. Our study is exploratory in nature: it is aimed at identifying the main topics of the articles and the dynamics of thematic changes in the period 2007–2018, including the meaning, the background and the changes of each corruption topic. The causal links revealed by this research lie in whether the medium is of an oppositional or of a pro-government position, and how election campaign periods affect the thematic structure of the representation of corruption. Owing to the fact that the ownership of the news portal Origó changed during the analysed period, a natural experiment has also been possible in an attempt to reveal the impact of this change on the thematic structure of the corruption discourse on the portal in question.

Keywords: automated text analytics, corruption, dynamic topic model, NLP, text mining

Katona Eszter survey-statisztikus, az ELTE Társadalomtudományi Kar Szociológiai Doktori Iskolájának hallgatója, a Társadalomkutatások Módszertana Tanszék tanársegédje, a Research Center for Computational Social Science kutatócsoport munkatársa. Kutatási témája a természetes nyelvfeldolgozás (NLP) alkalmazása a korrupciókutatásban. Jelenleg Fazekas Mihály témavezetésével UNKP-kutatást végez a közbeszerzési pályázatok korrupciókockázatát szövegbányászati módszerekkel vizsgálva.

Kmetty Zoltán szociológus, a Társadalomtudományi Kutatóintézet CSS-RECENS kutatócsoportjának munkatársa, az ELTE TÁTK adjunktusa és a „Társadalmi viselkedés megértése és modellezése digitális adatokkal" című MILAB-projekt szakmai vezetője. Az elmúlt négy-öt évben elsősorban CSS területen aktív, rendszeresen előad vezető nemzetközi konferenciákon, számos tanulmánya jelent meg Q1-es lapokban.

Németh Renáta az ELTE Társadalomtudományi Kar Empirikus Tanulmányok Intézetének vezetője, a Research Center for Computational Social Science kutatócsoport társvezetője. Alkalmazott matematika és szociológa szakon végzett az ELTE-n, ugyanitt szerezte doktori fokozatát a Szociológia Doktori Iskolában. Szakigazgatója a kar Survey Statisztika és Adatanalitika MSc programjának. Érdeklődése a kvantitatív társadalomkutatás különböző módszereire terjed ki, az utóbbi években az automatizált szöveganalitika szociológiai felhasználása foglalkoztatja. Számos D1-es és Q1-es folyóiratban publikált.