BOLONYAI FLÓRA – SEBŐK MIKLÓS

KVANTITATÍV SZÖVEGELEMZÉS ÉS SZÖVEGBÁNYÁSZAT

BEVEZETÉS

A fejezet áttekintést nyújt a kvantitatív szövegelemzés és szövegbányászat elméleti alapjairól, legfontosabb feladatairól és módszereiről, valamint néhány konkrét példával bemutatja, hogyan alkalmazhatók ezek a jogi szövegek vizsgálatára. A társadalom- és jogtudományban a szövegek tartalmának elemzése hagyományosan kvalitatív módszerekkel történt. Az 1990-es évektől kezdődően ugyanakkor előtérbe kerültek a szövegre mint adatra tekintő kutatások. A szövegek adattá konvertálása leggyakrabban egy dokumentum-kifejezés-mátrix létrehozásával történik, melyben minden dokumentum egy megfigyelés (sor), s a dokumentumban szereplő minden kifejezés egy oszlop. Az így kvantitatívvá alakított kvalitatív adatokkal már számos leíró és haladó statisztikai művelet elvégezhető. A leggyakoribb feladatok között megtalálható a csoportosítás (klaszterezés), az osztályozás (például a szövegek pozitív vagy negatív konnotáció szerinti besorolása), a tulajdonnevek felismerése, valamint a szöveg-összehasonlítás. Ezeket a feladatokat sokszor (a mesterséges intelligencia egy részterületének számító) gépi tanulással oldják meg. A fejezet röviden bemutat két, a jogi szövegekre vonatkozó alkalmazást is. Az elsőben az Alkotmánybíróság egyes döntéseinek tartalmát elemezzük nemzetközi jogi orientációjuk szempontjából. A másodikban az Országgyűlésnek benyújtott törvényjavaslatok és az elfogadott törvények szövegét hasonlítjuk össze.

A jogi szövegek elemzésének sok évszázados gyakorlata jellemzően a jogdogmatika területének sajátos módszertanára épül. Ezt a paradigmát olyan kulcsfogalmak jellemezik, mint a fogalmi rendszerépítés, a logika, a retorika, az argumentáció, az interpretáció és az analógia (Szabó, 1996). A módszertanukként a jogdogmatikát használó művekben ugyanakkor közös két elem. Egyrészt kiindulópontjukat jellemzően az egyes szövegek

jelentik (akkor is, ha a jogeseteket más esetek vagy a jogrendszer egészének összefüggésében tárgyalják). Másfelől tárgyukhoz elsősorban nem a társadalomtudományokban megszokott kvantitatív módszertannal közelítenek.

Miközben a nemzetközi és a hazai szakirodalomban is megjelentek már a jogi szövegeket nyelvközpontúan (lásd a szövegempirizmust, Blutman, 2014), ezek aggregátumaként, illetve kvantitatív nézőpontból vizsgáló munkák (lásd Jakab–Dyevre–Itzcovich, 2017; Gárdos-Orosz–Lőrincz–Ződi, 2017; Ződi, 2014b), inkább a kivételt, mint a szabályt jelentetve a jogtudományi fősodorban (e tudományfejlődési narratíva alternatív értelmezései kapcsán lásd Bódig–Ződi, 2016; Jakab–Menyhárd, 2015; Pokol, 2015). Még az olyan, egyébként a dogmatikai paradigmával kompatibilis kvalitatív társadalomtudományi tartalomelemzési technikák sem váltak a gyakorló jogtudós eszköztárának részévé, mint például a diskurzuselemzés (lásd a *Kvalitatív esettanulmány és diskurzuselemzés* című fejezetet). Másfelől a jogrendszerre mint hálózatra tekintő kvantitatív munkák jellemzően nem használták ki a jogi szövegek kvantitatív elemzésében rejlő lehetőségeket (ez alól kivételként lásd például Hamp–Syi–Markovich, 2016).[1] Mindezen kérdésekről e kötetben bővebben is szól *A jogi szövegek mint big data* és a *Hálózatkutatás* című fejezet, így itt csak annyit jegyzünk meg, hogy a *szöveggel mint adattal* foglalkozó kutatások átfedésben vannak a *jog mint adat* (Livermore–Rockmore, 2019) egyéb területeivel (mint amilyenek például a bírói döntések hivatkozási hálózatai).

A fejezet legfontosabb célja az, hogy bevezetést nyújtson a jogi szövegek kvantitatív társadalomtudományi elemzésének módszertanába (e tekintetben jelentős mértékben épít a Ződi Zsolt által jegyzett big data fejezetben bevezetett fogalmakra és jogi alkalmazásokra). A szövegek kvantitatív elemzése (*quantitative text analysis*, QTA; Mehl, 2006) annyiban különbözik a hagyományos, kvalitatív elemzési technikáktól, hogy a szövegre mint adatra *(text as data)* tekint (Gentzkow–Kelly–Taddy, 2019; Grimmer–Stewart, 2013).

A kvantitatív szövegelemzés tágan értelmezett tudományterületét szokták (a szűkebb tudásanyagot lefedő) szövegbányászat (Tikk, 2007), szöveganalitika, „szavak mint adat" (Slapin–Proksch, 2014; Laver et al., 2003) néven is említeni, illetve nagy átfedést mutat a természetesnyelv-feldolgozás (*natural language processing*, NLP), a számítógépes nyelvészet és a gépi tanulás *(machine learning)* területével is. Bármelyik elnevezést is használják az adott kutatások, annyi közös bennük, hogy valamilyen – jellemzően, bár nem kizárólag – szabadon hozzáférhető szoftveres környezetben végzik a tényleges elemzést, legyen szó az R programnyelvről (Silge–Robinson, 2017) vagy a másik gyakran használt opcióról, a pythonról (Bengfort–Bilbro–Ojeda, 2018).

[1] Köszönjük Ződi Zsolt jogtudományi-korpusznyelvészeti hátterű megjegyzéseit itt és alább.

■ 1. A SZÖVEG MINT ADAT

A szövegek – hasonlóan más kvalitatív adatforrásokhoz (filmek, képek) – közvetlenül nem elemezhetők kvantitatív módon, így az ilyen célú vizsgálatoknál be kell iktatni az adattá (itt: kvantitatív információvá) alakítás lépéseit. E lépések számos sajátos módszertani problémát vetnek fel, amelyekről bőséges áttekintést nyújt Sebők, 2016, valamint a Kúria joggyakorlat-elemzése kapcsán illusztrálja Mészáros–Sebők, 2018. Így e helyt csak egy rövid összefoglalót adunk a módszertani problémákról, és a fejezet nagy részét két konkrét kutatás bemutatásának szenteljük.

A társadalom- és jogtudományban a szövegek tartalmának kvantitatív elemzése az 1990-es évektől kezdődően került előtérbe, de az informatikai irodalomban már a hatvanas évektől lerakták a téma fogalmi alapjait. A szövegek adattá konvertálása már a korai tanulmányokban is (lásd például Borko, 1962) gyakran egy dokumentumkifejezés-mátrix (*document-term matrix*, DTM) létrehozásával történik, melyben minden dokumentum egy megfigyelés (sor), és a dokumentumban szereplő minden kifejezés egy oszlop. Az *1. táblázat* az 1999. évi magyar törvények 125 elemű (sorból álló) korpuszán (dokumentumgyűjteményén) szemlélteti a DTM logikáját.

1. TÁBLÁZAT ■ Az 1999. évi törvények szöveg-előkészítés nélküli DTM-e (részlet)

	1999	évi	c	törvény	az	európai	szociális
1999 C.txt	5	820	5	21	5	235	10
1999 CI.txt	3	416	3	17	5	108	0
1999 CII.txt	4	176	1	3	2	73	0

Mint az az ábrára kiemelt három törvény kapcsán megfigyelhető, ezek többször, akár több ezer alkalommal is tartalmazhatnak egy-egy kifejezést (lásd az első törvény esetében az „évi” terminust). Másfelől tekintettel arra, hogy az ilyen mátrixok mérete attól függ, hogy mekkora az adott korpusz szókincse, az oszlopok száma akár a több tízezret is elérheti (míg a sorok száma esetünkben 125-re korlátozott). A teljes, előzetes beavatkozás nélküli mátrix oszlopainak száma 60 586, ami rendkívül nagy számítási igényű feladatot ró a későbbi elemzéseket végző program mögött álló hardverre.

De nem csak a számítás kivitelezhetőségének biztosítása, illetve a futási idő csökkentése miatt célszerű redukálni e mátrix méretét egyes oszlopok eltávolításával. Az ábrán látható, hogy önálló elemként (szakirodalmi nyelven: tokenként) jelenik meg a „c” betű, mely nyilván a százas szám római megfelelőjére, valamint a felsorolások „c” pontjára utal. Ugyanakkor az elemzések többségéhez (például a törvények tartalmának megállapításához) ez nem járul hozzá hasznos információval, így az ilyen és ehhez hasonló tokenekre vonatkozó adatok elhagyhatók.

Hasonlóképpen elhagyható az „az” névelő, amely szintén nem utal a törvény tartalmára (szemben az „európai” vagy a „szociális” kifejezéssel). Gyakran el szokták távolítani a számokat, írásjeleket, a szóközöket (hiszen a szavak már önálló cellákba kerülnek) és az

ún. tiltólistás szavakat (mint a példában a névelők és a „c"). A nagybetűs szavakat sokszor kisbetűssé alakítják.

Összességében az ilyen lépéseket a szövegek elemzésre való előkészítéseként *(preprocessing)* használjuk, ami a legtöbb szövegbányászati projekt fontos eleme. Az 1999-ben kihirdetett 125 törvény korpuszának ilyen „tisztítása" egy R-ben írt program segítségével 25 368 oszlopot tartott meg, melynek a *2. táblázat* mutatja egy részletét.

2. TÁBLÁZAT ■ Az 1999. évi törvények szöveg-előkészítés utáni DTM-e (részlet)

	európ	szociális	kart	torinó	kel	függele
1999C.txt	21	50	44	47	2	3
1999CI.txt	17	0	0	0	0	1
1999CII.txt	3	0	0	0	0	0

Mint az illusztrációból látható, az „európai"-ból „európ" lett, ami a szótövesítés *(stemming)* nyelvtechnológiai megoldására vezethető vissza. Az „európ" ragozott alakjai megint csak nem tesznek hozzá az elemzésünkhöz, ha arra vagyunk kíváncsiak, hogy milyen gyakoriak az *európai* vagy az *Európában* zajló eseményekre való hivatkozások. Másfelől a „szociális" kifejezés esetében a magyar nyelven elérhető szótövesítő program nem alkotott új alakot, amit az is indokol, hogy például a „szociális" és a „ szocialista" kifejezés eltérő tartalmú (az ilyen kemény diók már az ún. lemmatizálással kezelhetők, mely nemcsak nyelvtani, hanem lexikális információkat is felhasznál a tokenizáláshoz, és mindig értelmes szavakat eredményez).

■ 2. A SZÖVEGBÁNYÁSZATI TECHNIKÁK ALKALMAZÁSA A GYAKORLATBAN

Az ily módon kvantitatívvá alakított kvalitatív adatokkal már számos szokásos leíró és haladó statisztikai művelet elvégezhető. A leggyakoribb feladatok között megtalálható az információ-visszakeresés *(information retrieval)*, a kulcsszavas keresés, a csoportosítás (klaszterezés), az osztályozás (például a szövegek pozitív vagy negatív konnotáció szerinti besorolása), a névelemek (mint például a tulajdonnevek) felismerése, valamint a szöveg-összehasonlítás. Ezeket a feladatokat sokszor a mesterséges intelligencia egy részterületének tekinthető gépi tanulással oldják meg, azáltal, hogy egy adatbázis statisztikai információi alapján határoznak meg csoportokat a szövegekben vagy tesznek előrejelzést az eredetileg nem vizsgált elemek kapcsán. Az egyszerűbb feladatokat ugyanakkor szótáralapú technikákkal is kezelni lehet (az egyes módszertani megoldásokról bővebben lásd Grimmer–Stewart, 2013; Sebők, 2016; Welbers – Van Atteveldt – Benoit, 2017).

Az alábbiakban két gépi szövegelemzési példát mutatunk be az empirikus jogtudomány területéről. Az első projektben kulcsszavak segítségével vizsgáljuk, hogy az alkotmánybírósági határozatok indokolásaiban milyen témák köszönnek vissza. A má-

sodikban automatizált szöveg-összehasonlítást végzünk: azt vizsgáljuk, hogy mennyire változik a megjelent törvények szövege az eredetileg benyújtott törvényjavaslatok szövegéhez képest.

2.1. Első példa: alkotmánybírósági határozatok kulcsszavas elemzése

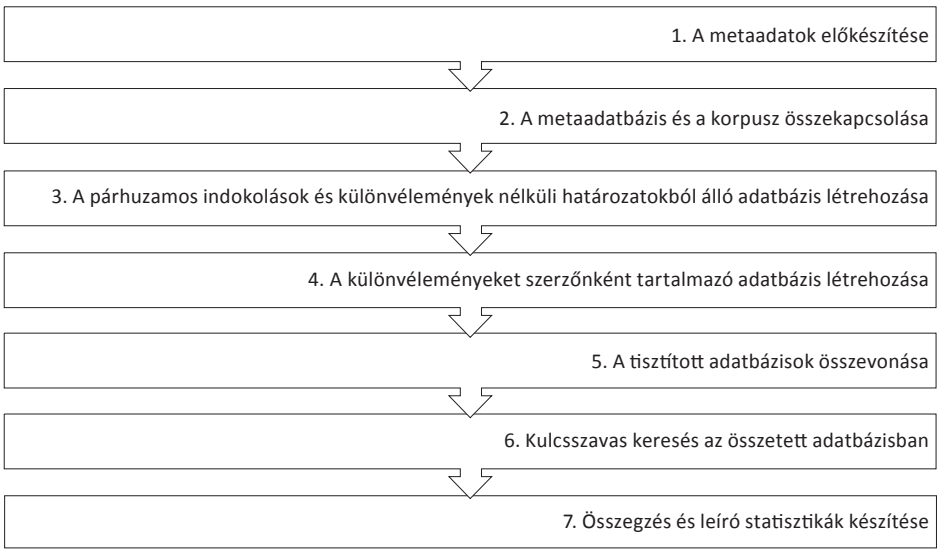
A magyar Alkotmánybíróság (AB) határozatait számos módon lehet elemezni a szöveg mint adat paradigma segítségével. Ezek közül talán a legegyszerűbbek egyike az automatizált kulcsszavas keresés, mely ugyanakkor már túlmutat például egy szövegszerkesztőben kényelmesen kivitelezhető elemzésen.

Minikutatásunkban arra voltunk kíváncsiak, hogy a határozatok indokolásában, illetve az esetleges különvéleményekben megjelennek-e olyan kulcsszavak, amelyek az indokolást vagy különvéleményt három kiemelt terület, az európai uniós, a „németes" vagy a nemzeti szuverenitásra épülő jogi hagyomány egyikéhez kötik. Természetesen a kulcsszavas keresés egyszerűsítő „lefordítása" annak a problémának, hogy az Alkotmánybíróság az egyes korszakaiban melyik hagyományt követi, ugyanakkor a vizsgálat számszerűsíthető eredményei jó kiegészítését jelenthetik a hagyományos, szövegértelmezésen alapuló dogmatikai elemzésnek. Így például a kulcsszavas keresést felhasználva arra a kérdésre is választ kerestünk, hogy megfigyelhető-e bármilyen idősoros mintázat, azaz az indokolások tartalma hogyan változik a testület egyes elnökeinek regnálása alatt.

Az AB határozatok kulcsszavas elemzésének kiindulópontja két adatbázis, amelyeket a Társadalomtudományi Kutatóközpont munkatársai állítottak össze:[2] egyrészt az 1990 és 2018 között született AB határozatok döntő többségét tartalmazó korpusz, másrészt pedig az ugyanebben a korszakban megjelent AB döntések metaadatait tartalmazó adatbázis. Az adatforrás méretéből adódóan kevésbé alkalmas a kézi-kvalitatív elemzésre: a korpusz több mint 3500 AB határozat szövegét tartalmazza, így a gépi elemzés hatékony megoldást jelenthet.

Az adatbázis metaadatokat is tartalmaz, ami hasznos kiegészítési lehetőségét adja a pusztán szövegekre épülő elemzésnek. A metaadat táblák az adott dokumentumok valamilyen lényeges tulajdonságát tartalmazzák, így esetünkben a kapcsolódó döntés online elérési útvonalát, azonosítóját, típusát, címét, a megjelenés évét, a dokumentumban idézett AB határozatok és egyéb jogi dokumentumok felsorolását, a döntéshozásban részt vevő alkotmánybírók nevét, az esetlegesen különvéleményt megfogalmazó bírák nevét, a dokumentum fogalmazóját, valamint az ügy típusát, amelynek kapcsán a döntés született. Az *1. ábra* röviden összefoglalja az elemzés folyamatát.

[2] Az itt használt adatbázis létrehozásában a szerzőkön kívül közreműködött Jakab András, Kacsuk Zoltán és Ződi Zsolt. Az adatbázisból kihagytuk a kutatási kérdés szempontjából nem releváns döntéseket.



1. ÁBRA ■ Az elemzés folyamata

Ahogy az elemzés folyamatát bemutató *1. ábrán* is látszik, a kulcsszavas keresést hosszas adattisztítás és adatbázis-kezelési feladatok előzik meg, ami általánosságban is jól illusztrálja a szövegbányászati tevékenység komplexitását. A nyers szöveges adatok önmagukban gyakran nem alkalmasak elemzésre, a metaadatokkal történő összekapcsolás így gyakori feladat a gépi szövegelemzés során. Másfelől a strukturálatlan adatforrásokból (például egy e-mailben kapott tömörített .txt-csomag) elemzésre alkalmas, strukturált adatbázis létrehozása sok kutatási projektben nagyobb kihívás, mint maga a kvantitatív szövegelemzés.

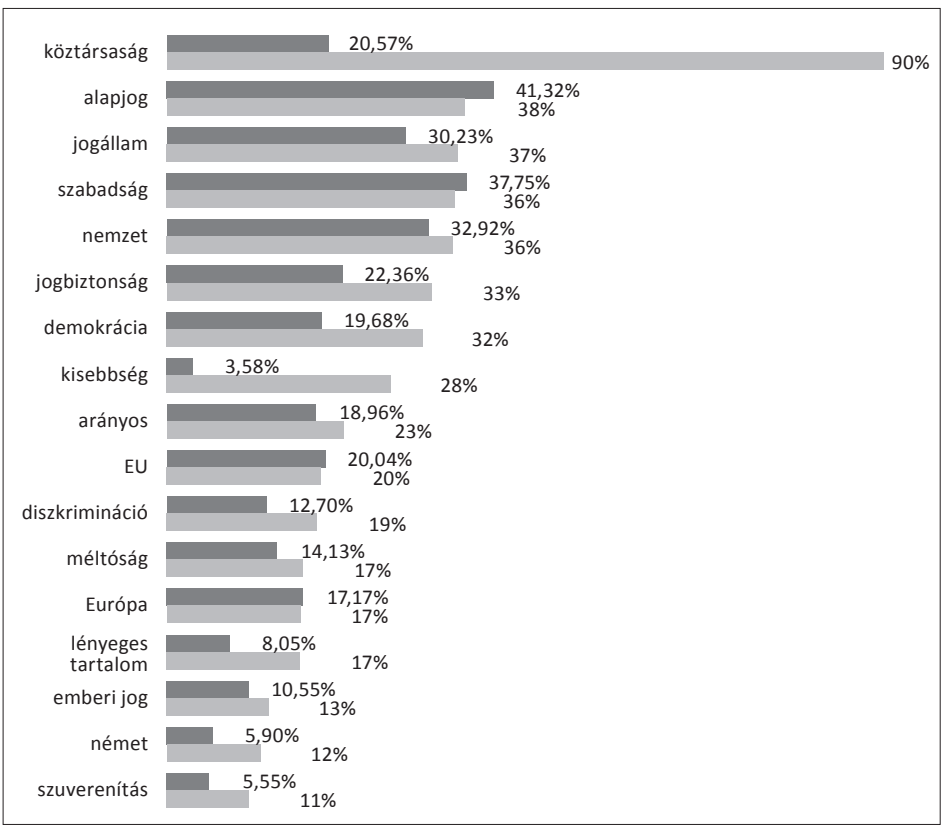
A *3. táblázatban* a kutatási kérdésünknek megfelelő három nagy csoportra osztott kulcsszavak láthatók. A kulcsszavak egy részére a kifejezések eredeti formájában, változtatás nélkül kerestünk rá (például „Emberi Jogok Európai Bírósága"), míg más esetekben szótöveztük a szótárban szereplő szót, hogy a keresett kulcsszó toldalékolt formái is megjelenjenek a találatok között (például „demokr", „diszkrim" – az egyes kulcsszavakat a gyakori kontextusuk alapján sorolták jogász kollégáink az egyes kategóriákba).

Immár az elemzés eredményeire rátérve a *2. ábra* azt mutatja be, hogy az egyes nagyobb területekhez kapcsolódó, leggyakrabban előforduló kulcsszavak az egyes alkotmánybírósági elnökök alatt kiadott állásfoglalások és különvélemények mekkora hányadában fordulnak elő. Az ábrán látszik, hogy az 1990-es évek elején még gyakran előfordult a német joghoz kapcsolható indokolás az alkotmányjogi döntésekben és az ezekhez fűzött különvéleményekben is.

A 2000-es évek közepétől kezdve ezek gyakorlatilag teljesen eltűntek. Ez nem specifikusan magyar jelenség, és leginkább azzal magyarázható, hogy a német alkotmánybíróság mintaadó szerepét részben nyelvi (ti. az angol nyelv dominanciájának erősödése a

3. TÁBLÁZAT ■ A kulcsszavas keresés során használt szavak

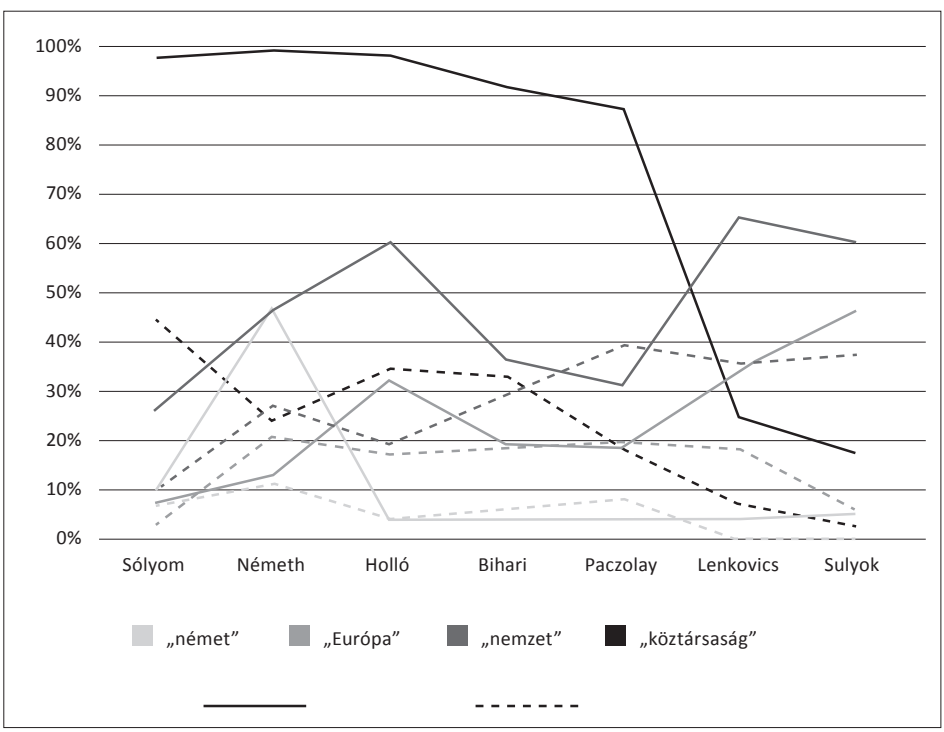
Európai jog	Német jog	Nemzeti szuverenitás
Európa	német	méltóság
EU	BVerfG	jogállam
Strasbourg	BVerfGE	jogbiztonság
strasbourgi	Drittwirkung	demokr
Emberi Jogok Európai Bírósága	Wesensgehalt	köztársaság
Európai Emberi Jogi Bíróság	lényeges tartal	parlamentáris
Európai Unió Bírósága		alapjog
		egyház
		nemzet
		kisebbség
		arányos
		diszkrim
		szabadság
		ezeréves



2. ÁBRA ■ A gyakori kulcsszavak százalékos előfordulása a két dokumentumtípusban

jogtudományban a német kárára, különösen a fiatalabb jogtudósi generációkban), részben jogi okokból (nemzetközi jogi kötelezettségekből fakadóan) az Emberi Jogok Európai Bírósága vette át (Groppi–Ponthoreau, 2013, 429). Az „Európa" szó gyakoriságnövekedése is nagyrészt ennek tudható be.

További érdekesség, hogy a „köztársaság" kifejezés szinte minden döntés indokolásában szerepelt Paczolay Péter alkotmánybírósági elnökségéig (2008–2015), azóta viszont jóval ritkább a használata. Ez rávilágít egy gyakori módszertani problémára: a formai változások hatására. Az ország hivatalos neve 2012. január 1. óta nem „Magyar Köztársaság" (addig a határozatokat is „A Magyar Köztársaság nevében!" kezdettel hirdették ki), hanem „Magyarország". A korábbi Alkotmány számos helyen tartalmazta a „Magyar Köztársaság" kifejezést, amelyet aztán az Alkotmánybíróság rendre beidézett az indokolásában, és ez jelent meg az adatokban; a 2012. január 1. óta hatályos Alaptörvény esetében azonban a „köztársaság" szó már leginkább csak a „köztársasági elnök" szóösszetételben jelenik meg, ez azonban jóval kevesebb határozatnál releváns. Ugyanakkor pusztán a formális elnevezés megváltozásával nem magyarázható teljesen a kifejezés gyakorisága a vizsgált időszak során. Összességében ez a példa is jól mutatja, hogy az adatok helyes értelmezéséhez minden szövegbányászati elemzésnél szükséges az adott szövegek szubsztatív (jelen esetben jogászi) ismerete.



3. ÁBRA ■ A gyakori kulcsszavak százalékos előfordulása az egyes AB elnökök időszakában

2.2. Második példa: szövegek gépi összehasonlítása

A második példa már némileg komplexebb és a szövegbányászat több különböző technikájára is épít.[3] A gépi szöveg-összehasonlítás – más automatizált szövegelemzési módszerekhez hasonlóan – lehetőséget teremt arra, hogy nagy mennyiségű szöveget tudjunk viszonylag rövid idő alatt feldolgozni. Gépi eszközökkel a hasonlóságot kvantitatív módon mérjük, így az összehasonlítás első lépéseként mindig meghatározzuk azt a mérőszámot, amely alapján a szövegeket összehasonlítjuk.

A szövegek, akár gépi, akár kézi úton, több szempontból is összehasonlíthatók. Gondolhatunk egészen egyszerű mutatószámokra, például a szöveg hosszára vagy a szövegben szereplő szavak átlagos hosszára, de ezeknél bonyolultabb változókra is. Ilyen, a klasszikus korpusznyelvészetben gyakran használt összetettebb mutató a szövegben szereplő mondatok összetétele (például hogy a szöveg mekkora része áll határozószavakból, igékből stb.) vagy a szöveg lexikális diverzitása (szókincse).

Az egyik legegyszerűbb lexikális diverzitást mérő mérőszám a *type-token ratio* (TTR). Kiszámításához a vizsgált szövegben található összes *különböző* szó számát kell elosztani a szövegben található szavak számával (ideértve a duplikációkat, azonos szótöveket is). A TTR gyakori kritikája, hogy nagyon érzékeny a szövegek terjedelmére. Egy rövid szövegnek ugyanis – minden mást adottnak véve – nagyobb eséllyel lesz magasabb a TTR-mutatója, hiszen ilyenkor kisebb a szóismétlés valószínűsége. A TTR e hátrányának kiküszöbölésére több megoldás is létezik, például a mozgó átlagokra építő MATTR (*moving average TTR*) mutató. Ez egy előre definiált hosszúságú szövegrészleten számolja ki a szövegre jellemző TTR-t, s a kijelölt fix hosszúságú részlet folyamatosan csúszik előre a szövegben.

A lexikális diverzitás leginkább akkor lehet a szövegek összehasonlításának hasznos eszköze, ha a szerzők vagy a potenciális célközönség nyelvhasználatának választékosságát akarjuk felmérni. Ugyanakkor kevésbé használható, ha két szöveg tartalmát akarjuk összehasonlítani, hiszen a szókincs mérete önmagában nem mond semmit a szöveg tartalmáról. Ez utóbbi feladatnál a legkézzelfoghatóbb megoldás az, ha a szövegeket alkotó egységeket (például a szavakat vagy egymást követő szavak nagyobb blokkjait, az ún. n-gramokat, lásd alább, illetve *A jogi szövegek mint big data* című fejezet példáit) kezeljük adatként, és a két szöveget az alapján vetjük össze, hogy az alkotóelemeik mennyire hasonlítanak egymásra.

Fontos hangsúlyozni, hogy ez a módszer sem képes tökéletesen felismerni a jelentésbeli hasonlóságokat vagy azonosságokat (például hogy a „Fehér Ház" és az „amerikai elnök" bigram ugyanarra a politikai döntéshozóra utal), és bizonyos esetekben túlhangsúlyozza a formális különbségeket (például ugyanannak a szónak az eltérő írásmódjait). Ennek ellenére összességében az automatizált módszerekkel mért kvantitatív hasonlóság és a hagyományos módszerekkel mért (például emberek által olvasás során megállapított) kvalitatív hasonlóság között nincs drasztikus eltérés, miközben a kvantitatív meg-

[3] A vizsgálat módszertanát és eredményeit összefoglaló tanulmány: Sebők et al., 2020 (kézirat).

oldással megbízhatóan lehetünk képesek nagy mennyiségű szöveget is elemezni rövid idő alatt. Ez a képesség rendkívül hasznos emberi olvasással feldolgozhatatlan nagyságú korpuszoknál, mint amilyen példáinkban a rendszerváltás utáni AB határozatok vagy a törvények és törvényjavaslatok adatbázisa.

A szövegek tartalmi összehasonlítására alkalmas mérőszámokra jellemző, hogy kiszámolásuk során a szövegek tartalmát valamilyen matematikai eszközzel reprezentáljuk, és a szövegeket reprezentáló egységek között definiáljuk a hasonlóságot. A szövegek reprezentálásához általában a szöveg valamilyen kisebb egységét használjuk fel, azaz a szöveget *tokenizáljuk*. A leggyakrabban használt egységek közé tartoznak a szövegben szereplő karakterek vagy egymás után következő karakterek csoportja *(karakter n-gram)*, illetve a szövegben szereplő szavak és az egymás után következő szavakból álló *n-gramok*.[4] A következőkben két hasonlósági mérőszámot mutatunk be röviden.

Jaccard-távolság. A Jaccard-távolság kiszámításához először a Jaccard-együttható kiszámítására van szükség (erről lásd a Miskolc Jogi Korpuszt, Szabó–Vinnai, 2018; Vincze, 2018). A Jaccard-együttható kalkulálása során azt vizsgáljuk meg, hogy a két összehasonlítandó szövegben a megegyező n-gramok száma hogyan aránylik a két szövegben szereplő összes n-gram számához. A mutató értékét befolyásolja, hogy mekkora n-re esik a választás, valamint az, hogy milyen hosszú a két összehasonlítandó szöveg. A Jaccard-együttható halmazműveleti jelekkel a következőképpen fejezhető ki:

$$J(doc_1, doc_2) = \frac{|doc_1 \cap doc_2|}{|doc_1 \cup doc_2|}$$

A fenti képletben a $|doc_1 \cap doc_2|$ a mindkét dokumentumban szereplő n-gramok számát (vagyis a két dokumentumot reprezentáló n-gramokból álló halmazok metszetének a méretét), a $|doc_1 \cup doc_2|$ pedig a két dokumentumban szereplő összes n-gram számát (vagyis a két dokumentumot reprezentáló n-gramokból álló halmazok uniójának a méretét) jelöli. Mivel a közös n-gramok száma mindig kisebb, mint az összes n-gram száma, ezért a Jaccard-együttható értéke mindig 0 és 1 közé esik. A szintén 0 és 1 közötti értékkészlettel rendelkező Jaccard-távolság képlete a következő:

$$d_J(doc_1, doc_2) = 1 - J(doc_1, doc_2) = 1 - \frac{|doc_1 \cap doc_2|}{|doc_1 \cup doc_2|}$$

A 0 értékű Jaccard-távolság azt jelenti, hogy a két szöveg teljesen megegyezik a vizsgált n-gramok szintjén, az 1 értékű Jaccard-távolság pedig azt, hogy a két vizsgált szövegben egyáltalán nincs egyező n-gram. A Jaccard-távolság felfogható százalékos eltérésként is: ha két szöveg között a Jaccard-távolság 0,5, az azt jelenti, hogy a két szöveg (a vizsgált egységek szintjén) 50%-os eltérést mutat.

[4] Ha a szövegelemzés egysége egy karakter vagy egy szó, akkor a (karakter) n-gramok egy speciális esetével dolgozunk: az unigramokkal.

Koszinusztávolság. A koszinusztávolság két vektor közötti távolságot mér, s kiszámításához először a koszinuszhasonlóság kiszámítására van szükség. A koszinuszhasonlóság nem más, mint a két nem nullvektor által bezárt szög koszinusza, ami kifejezhető a vektorok skaláris szorzatának,[5] valamint a két vektor hosszának[6] a hányadosával. Szövegek összehasonlítása esetén a szövegeket egy-egy szógyakorisági *(term frequency)* vektor reprezentálja, melynek alapja – akárcsak a Jaccard-hasonlóság esetén – tetszőleges hosszúságú n-gram lehet.

Az n=1 esetben (azaz ha csak különálló szavakat nézünk, függetlenül a sorrendjüktől) ezt a vektort úgy kell elképzelni, hogy ha van két „dokumentumunk", amelyik négy szóból áll („polgári törvénykönyv polgári törvénykönyv", illetve „polgári polgári polgári törvénykönyv"), akkor a vonatkozó szóvektor értékek a két dokumentumból álló korpuszunkra nézve (ahol a vektor elemei a következők: polgári; törvénykönyv) rendre (2; 2), illetve (3; 1). E koordináta-rendszerben a két szó adja a két tengelyt és a szavak gyakorisága az adott tengelyen elfoglalt pozíciót. Az origóból a koordináták által meghatározott két pontba húzott vonalak által bezárt szög koszinusza lesz így a két dokumentum eltérését meghatározó mutató. Természetesen a komplexebb, akár több ezer szót is tartalmazó korpuszok esetében a vizuális ábrázolás már nem kivitelezhető, így a távolság kiszámítására a megfelelő képlet használható (valamint egy kellő kapacitású számítógép).

A koszinuszhasonlóság képlete a következő:

$$\cos(\mathrm{doc}_1, \mathrm{doc}_2) = \cos(\theta) = \frac{\mathrm{doc}_1 \cdot \mathrm{doc}_2}{|\mathrm{doc}_1||\mathrm{doc}_2|}$$

A fenti képletben a $\mathrm{doc}_1 \cdot \mathrm{doc}_2$ kifejezés a két szöveget reprezentáló vektorok skaláris szorzata, míg a $|\mathrm{doc}_1||\mathrm{doc}_2|$ kifejezés a két szöveget reprezentáló vektorok hosszának a szorzata. Hasonlóan a Jaccard-hasonlósághoz, a két szöveg által bezárt szög koszinusza is 0 és 1 közötti értéket vehet fel.[7] A koszinusztávolságot úgy kapjuk meg, ha ezt a 0 és 1 közötti értéket kivonjuk 1-ből, képlete így a következő:

$$d_{cos}(\mathrm{doc}_1, \mathrm{doc}_2) = 1 - \cos(\mathrm{doc}_1, \mathrm{doc}_2) = 1 - \frac{\mathrm{doc}_1 \cdot \mathrm{doc}_2}{|\mathrm{doc}_1||\mathrm{doc}_2|}$$

A 0 értékű koszinusztávolság azt jelenti, hogy a két szöveg által bezárt szög koszinusza 1, vagyis a két szöveget reprezentáló vektorok egy irányba mutatnak, ami a gyakorlatban azt jelenti, hogy a szövegek egyformák, vagy legalábbis egyforma n-gramokból épülnek fel. Ezzel szemben az 1 értékű koszinusztávolság azt jelenti, hogy a két szöveget reprezentáló

[5] Az n hosszúságú *a* és *b* vektorok skaláris szorzataként az alábbi számot (skalárt) értjük: $\mathrm{a} \cdot \mathrm{b} = \sum_{i=1}^{n} a_i b_i$, ahol a_i és b_i rendre *a* és *b* vektor i-edik eleme.

[6] Az n hosszúságú *a* vektor hosszát az alábbi képlettel számolhatjuk ki: $|\mathrm{a}| = \sqrt{\sum_{i=1}^{n} {a_i}^2}$

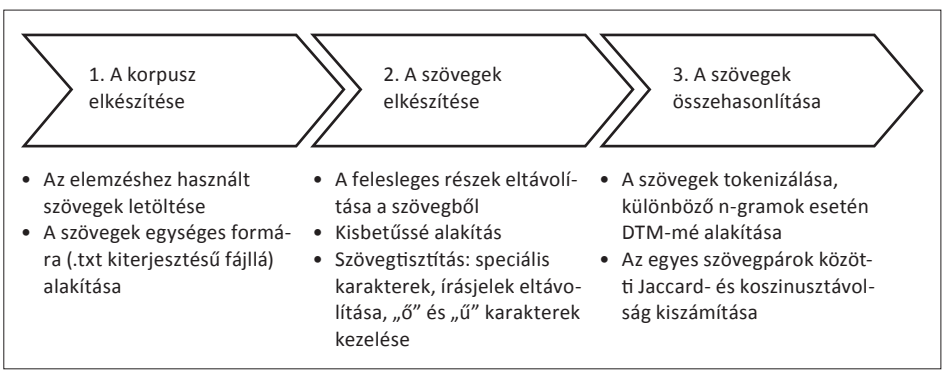
[7] Habár a koszinuszfüggvény értékkészlete a [-1,1] zárt intervallum, a szógyakorisági vektor sajátossága miatt (a vektor minden eleme nem negatív) két szöveg által bezárt szög esetén a koszinuszfüggvény 0 és 1 közötti értékeket vesz fel.

vektorok merőlegesek egymásra (és így az általuk bezárt szög koszinusza 0). Ez a gyakorlatban akkor fordul elő, ha a két szövegben nincs egyezés a vizsgált n-gramok szintjén.

A legelterjedtebb szövegelemző szoftverek a bemeneti szövegek megfelelő előkészítése után beépített függvények segítségével számolják ki a szövegek hasonlóságát mérő ilyen mérőszámokat, így amennyiben ezek elméleti hátterével tisztában vagyunk, maga a számítás minimális programozási ismeretet igényel.

2.3. A jogszabályszövegek gépi szöveg-összehasonlítása

A következőkben egy példán mutatjuk be az automatikus szöveg-összehasonlítás egy lehetséges alkalmazását. Az elemzés tágabb tudományos célja annak vizsgálata volt, hogy a rendszerváltás óta eltelt időszakban az Országgyűlésnek milyen törvényhozási hatalma volt, azaz mennyiben tudta befolyásolni a (jobbára a végrehajtó hatalom által benyújtott) törvényjavaslatok szövegét. Az elemzés fontosabb lépéseit a *4. ábra* mutatja be.



4. ÁBRA ■ Az elemzés folyamata

1. lépés: a korpusz előkészítése

Az elemzéshez használt korpusz az Országgyűlés által tárgyalt és elfogadott, 1990 és 2018 között benyújtott törvényjavaslatokból és elfogadott törvényekből állt. A törvényjavaslatok többsége elérhető a parlament honlapjáról, az elfogadott törvények pedig szintén letölthetők.[8] Mivel néhány törvényjavaslat nem állt rendelkezésre az elemzéshez szükséges formában (például mert csak gyenge minőségű szkennelt dokumentumok voltak elérhetők az 1990-es évekből), a végső adatbázisban nincs benne a vizsgált időszak alatt elfogadott összes törvény. Az adatbázis azonban így is szinte teljes volt, több mint 4100 törvényjavaslat-törvény párból állt.

[8] *Bit.ly/329gSdZ; bit.ly/3iHdjSK.*

A híres *Supreme Court Database*[9] és sok más további, az 1980–90-es évek empirikus szövegvizsgálatainak megközelítésével dolgozó kutatás mintájára természetesen lehetett volna kézzel is kódolni a szöveget, és minden dokumentumpár megkaphatta volna az eltérésükre vonatkozó értéket (a kézi kódolás kapcsán lásd *A kutatási terv* című fejezetet is). Ugyanakkor az adatbázis mérete és az, hogy egy dokumentumpáron belül akár több száz egyező vagy különböző blokk is azonosítható, különösen indokolta az automatikus szöveg-összehasonlítás alkalmazását. Ekkora szövegmennyiség egyenletes színvonalú, kézzel történő kódolása még tapasztalt kutatók számára is nagy kihívás, ráadásul – szemben előző példánkkal – érdemi hozzáadott értékkel sem járna ebben az esetben a kutatói kódolás.

E lépés kapcsán még megjegyzendő, hogy habár az összehasonlítandó szövegek többsége online elérhető, ezek formátuma nem egységes, ezért a korpusz elkészítéséhez a szövegeket egységes formátumúvá konvertáltuk, egyszerű szöveggé (.txt kiterjesztésű fájlokká) alakítottuk.

2. lépés: a szövegek előkészítése

A szöveg-előkészítés során eltávolítottuk a törvényjavaslatok és törvények azon részeit, amelyek nem tartoznak a szövegek lényegi, s így elemzendő részéhez. Elhagytuk a címeket, fejléceket, a lábjegyzeteket, valamint mellékleteket (például általános és részletes indokolás, amelyek nem részei az elfogadott jogszabálynak, s így nem is lehet ezek kapcsán a hasonlóságokat vizsgálni). A végső szöveg tehát a cím után kezdődött, és az első mellékletig vagy az indokolás kezdetéig tartott.

A szöveg-előkészítés lépéséhez tartozik a szövegek tisztítása is. Mivel a törvényjavaslatok és törvények teljes szövegét akartuk összehasonlítani, nem használtunk sem szótövezést vagy lemmatizálást, és nem távolítottunk el korábban meghatározott tiltólista alapján sem szavakat a szövegből. Megtartottuk a szövegben szereplő számokat is.

Ugyanakkor a szövegeket kisbetűssé alakítottuk, és a „§" karakter kivételével eltávolítottuk az írásjeleket és a speciális karaktereket. Mivel a hosszú ékezetes magánhangzók karakterkódolása bonyodalmakat okozhat, az általános szövegtisztításon felül külön kellett foglalkozni az „ő" és „ű" karakterekkel. Ezeket először „@" és „#" karakterekkel helyettesítettük, majd visszacseréltük az „ö" és „ü" karakterekre (azért a „@" és „#" karaktereket használtuk, mert ezekről feltételezhető volt, hogy nem szerepelnek máshol a korpuszban). Az *5. ábra* egy konkrét törvényjavaslat segítségével mutatja be, hogy a benyújtott szöveg hogyan változott meg az előkészítés eredményeképpen.

[9] Ez az ún. Spaeth-adatbázis, *bit.ly/3lfwlBS*.

Ilyen volt: a 2005-ös T/15784. számú törvényjavaslat eredeti szövege. A szövegtisztítás során törölt karakterek aláhúzással, az átalakított „ő” és „ű” karakterek pedig **félkövér** kiemeléssel vannak jelölve

Magyar Szocialista Párt
Országgyűlési képviselőcsoport
2005. évi…. törvény a gyermekek védelméről és a gyámügyi igazgatásról szóló 1997. évi XXXI. törvény módosításáról

1.§ A gyermekek védelmér**ő**l és a gyámügyi igazgatásról szóló 1997. évi XXXI. törvény (a továbbiakban: Gyvt.) 162. §-a az alábbi (3) bekezdéssel egészül ki:
„(3) Felhatalmazást kap az ifjúsági, családügyi, szociális és esélyegyenl**ő**ségi miniszter, a belügyminiszter, továbbá a pénzügyminiszter, hogy együttes rendeletben határozzák meg a települési önkormányzat részére szociális gyermekétkeztetés céljából nyújtott támogatás igénylésének, folyósításának és elszámolásának részletes szabályait.”
2.§ Ez a törvény a kihirdetését követ**ő** harmadik napon lép hatályba.

Indoklás
A Kormány a közelmúltban döntést hozott arról (…) A Javaslat az e rendelet megalkotásához szükséges felhatalmazó rendelkezéssel egészíti ki a Gyvt. szabályait.
Budapest, 2005. április 19.

Ilyen lett: a törvényjavaslat szövege a szövegelőkészítés után

l § a gyermekek védelméröl és a gyámügyi igazgatásról szóló 1997 évi xxxi törvény a továbbiakban gyvt 162 §a az alábbi 3 bekezdéssel egészül ki 3 felhatalmazást kap az ifjúsági családügyi szociális és esélyegyenlöségi miniszter a belügyminiszter továbbá a pénzügyminiszter hogy együttes rendeletben határozzák meg a települési önkormányzat részére szociális gyermekétkeztetés céljából nyújtott támogatás igénylésének folyósításának és elszámolásának részletes szabályait 2 § ez a törvény a kihirdetését követö harmadik napon lép hatályba

5. ÁBRA ■ A szövegelőkészítés eredményei

3. lépés: a szövegek összehasonlítása

A tisztított és rövidített szövegek már alkalmasak voltak az összehasonlításra, ezek már ténylegesen csak a törvényjavaslatok és törvények lényegi részét tartalmazták. Első lépésben a már tisztított szövegek alapján elkészítettük az adott szövegpár szószedetét, amely az általunk meghatározott n-nek megfelelő n-gramokból állt. Ehhez először tokenizáltuk a szövegeket, majd definiáltuk az n-gramok esetében szükséges hosszúságot, mely esetünkben kettő (azaz két egymást követő szó volt az elemzés egysége). Ezt követően mindkét szöveget DTM-formára alakítottuk, amivel végbement a kvalitatív információforrásnak tekinthető szövegek kvantitatív adattá alakítása (lásd a *4. táblázatot,* amely a fentebb még az unigramok kapcsán bevezetett *1.* és *2. táblázat* formátumának megfelelően mutatja, hogy milyen gyakoriak voltak a vizsgált bigramok).

4. TÁBLÁZAT ■ A 2013. évi T/10516. számú törvényjavaslat és a 2013. évi LXXIII. törvény DTM-e (részlet)

	a_kizárólag	kizárólag_energiahasznosítás	energiahasznosítás_céljából	céljából_kitermelt	kitermelt_termálvíz	termálvíz_visszatáplálására	visszatáplálására_kérelemre	kérelemre_a	a_vízügyi	vízügyi_hatósági	hatósági_feladatokat	feladatokat_ellátó	ellátó_szerv	szerv_ad	ad_engedélyt
Törvényjavaslat	1	1	1	1	1	1	1	1	1	1	1	1	1	1	1
Törvény	0	1	1	0	0	0	0	0	0	0	0	0	0	0	1

Az így kapott objektum egy olyan mátrix, amelynek „hossza" (az oszlopok száma) az előre meghatározott szószedetben szereplő kifejezések számával egyezett meg. A mátrix cellái pedig azt mutatják meg, hogy az adott kifejezés hányszor fordult elő a vizsgált szövegben. Ha egy kifejezés nem szerepelt a szövegben, a kifejezéshez tartozó cellába nulla került.

A szövegek mátrixszá alakítása után már kiszámolható volt a Jaccard-hasonlóság és a két szöveg vektora által bezárt szög koszinusza, ezeket 1-ből kivonva pedig megkaptuk a Jaccard- és a koszinusztávolságot. A *6. ábra* szemlélteti a konkrétan vizsgált szövegpáros kapcsán az összehasonlítás bemeneti adatait: a **félkövér** kiemeléssel jelölt részek a szövegek egyező részeit jelölik, az aláhúzottak a különbségeket.

A két szöveg között mért bigram alapú Jaccard-távolság értéke 0,51, a koszinusztávolságé 0,36, azaz az eltérés közepesnek mondható.

Az *5. táblázat* jelzi a kiemelt szövegpárunk átfedésben lévő részeit. A **félkövérrel** jelölt bigramok mindkét szövegben szerepelnek, a *dőlt betűsek* csak a törvényjavaslatban, a normál betűsek pedig csak a kihirdetett törvényben.

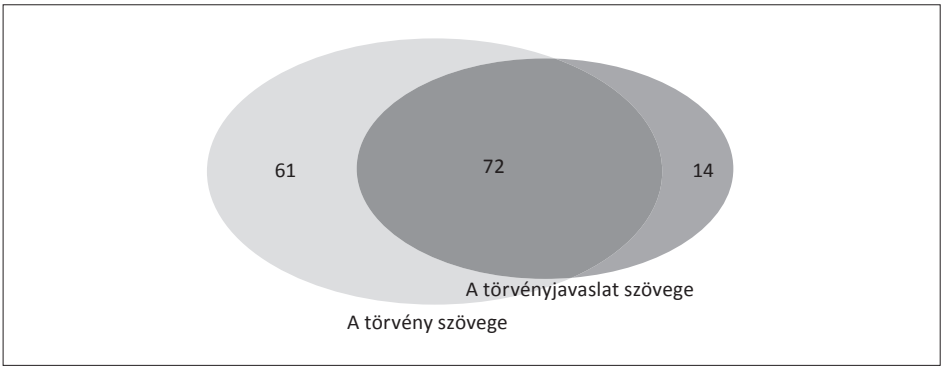
A 2013-as T/10516. számú törvényjavaslat	A 2013. évi LXXIII. törvény szövege
1 § a vízgazdálkodásról szóló 1995 évi lvii törvény a továbbiakban vgtv 15 § 3 bekezdése helyébe a következö rendelkezés lép 3 az ásvány gyógy és termálvizek felhasználásánál elönyben kell részesíteni a gyógyászati illetve a gyógyüdülési használatot a kizárólag energiahasznosítás céljából kitermelt termálvíz visszatáplálására kérelemre a vízügyi hatósági feladatokat ellátó szerv **ad engedélyt 2 § hatályát veszti a a vgtv 15 § 3a bekezdése valamint 1 számú mellékletének 35 és 36 pontja b a vízgazdálkodásról szóló 1995 évi lvii törvény módosításáról szóló 2012 évi cv törvény 3§ ez a törvény a kihirdetését követö 3 napon lép hatályba**	**1 § a vízgazdálkodásról szóló 1995 évi lvii törvény a továbbiakban vgtv 15 § 3 bekezdése helyébe a következö rendelkezés lép 3 az ásvány gyógy és termálvizek felhasználásánál elönyben kell részesíteni a gyógyászati illetve a gyógyüdülési használatot** 2 § a vgtv 28 §a a következö 5 bekezdéssel egészül ki 5 a szénhidrogén kitermelési céllal mélyített de arra alkalmatlan kutak termálvíz kitermelési célú hasznosítására a vízügyi hatóság **ad engedélyt** 3 § a vgtv 29 §a a következö 1a bekezdéssel egészül ki 1a termálvíz **kizárólag energiahasznosítás céljából** történö kitermelésére vonatkozó létesítési engedélyben az e törvény végrehajtására kiadott rendeletben meghatározott feltételek szerint rendelkezni kell a kitermelt víz elhelyezésének módjáról 4 § h**atályát veszti a a vgtv 15 § 3a bekezdése valamint 1 számú mellékletének 35 és 36 pontja b a vízgazdálkodásról szóló 1995 évi lvii törvény módosításáról szóló 2012 évi cv törvény** 5 § **ez a törvény a kihirdetését követö 3 napon lép hatályba**

6. ÁBRA ■ A szöveg-összehasonlítás eredménye

5. TÁBLÁZAT ■ A két dokumentum szövegének bigram alapú szószedete (részlet)

céljából_történö	kutak_termálvíz	**és_36**	**pontja_b**
§_ez	hasznosítására_a	**törvény_módosításáról**	**a_törvény**
létesítési_engedélyben	kiadott_rendeletben	**energiahasznosítás_céljából**	**a_gyógyászati**
kérelemre_a	vízügyi_hatóság	**helyébe_a**	**következö_rendelkezés**
alkalmatlan_kutak	*ellátó_szerv*	**2012_évi**	**§_3a**
kitermelési_célú	ki_1a	**a_továbbiakban**	**3_az**

Az *5. táblázatban* csak a teljes anyag egy részét látjuk: abban összesen 72 félkövér betűs cella van, tehát ennyi bigram szerepel mindkét szövegben. A két szövegben összesen 147 bigram található, így a két szöveg Jaccard-hasonlósága $\frac{72}{147} = 0{,}49$. Ezek alapján a két szöveg Jaccard-távolsága 0,51. A *8. ábrán* látható Venn-diagram a konkrét törvényjavaslat-törvény pár szövege közötti viszonyt ábrázolja.



8. ÁBRA ■ A két dokumentum Jaccard-hasonlósága Venn-diagramon

A két szövegrészlet között a koszinusztávolságot az alábbi két vektor összehasonlításával számoljuk ki (s a végén az eredményt kivonjuk egyből):

$$\boldsymbol{v_1} = (1,1,1,1,1,1,1,1,1,1,1,1,1,1,1); \quad \boldsymbol{v_2} = (0,1,1,0,0,0,0,0,0,0,0,0,0,0,1)$$

$$|\boldsymbol{v_1}| = \sqrt{15}; \quad |\boldsymbol{v_2}| = \sqrt{3}; \quad \boldsymbol{v_1} \cdot \boldsymbol{v_2} = \sqrt{3} \rightarrow$$

$$d_{cos}(\boldsymbol{v_1}, \boldsymbol{v_2}) = 1 - \frac{\sqrt{3}}{\sqrt{15} * \sqrt{3}} = 1 - \frac{1}{\sqrt{15}} = 0{,}74$$

A két hasonlósági mérce érdemi (bár messze nem drasztikus) eltérése egyben jelzi azt is, miért fontos, hogy körültekintően használjuk ezeket az eszközöket, s az eredmények validálása érdekében hasznos lehet a különböző megoldások összehasonlító használata is.

■ 3. TOVÁBBI KUTATÁSI IRÁNYOK

A rendelkezésre álló terjedelmi keretek között nem vállalkozhattunk arra, hogy valamennyi fontosabb kvantitatív szövegelemzési technika kapcsán részletesen ismertessük a kutatási folyamatot. A fejezet zárásaként ugyanakkor érdemes kicsit kitekinteni a fokozatosan bővülő magyar irodalomra annak érdekében, hogy további ötleteket adjunk a módszertan iránt érdeklődő olvasónak.

Legalább három különböző szempontot érdemes mérlegelni a szövegbányászati projektek tervezésekor. Az első s gyakran legfontosabb korlát a korpusz kérdése. A hazai empirikus jogtudományi kutatások számára egyelőre általában nem állnak rendelkezésre olyan jól strukturált, metaadatokkal bőségesen ellátott, nyilvános repozitóriumokban elhelyezett korpuszok, amelyek számos más fejlett demokrácia kapcsán elérhetők (ez alól kivételként lásd például a Társadalomtudományi Kutatóközpont Comparative Agendas Projectjének [CAP] törvényhozási és rendeleti adatbázisát: *cap.tk.mta.hu,* bővebb leírá-

sa Boda–Sebők, 2018). Az induló kutatások esetében így kulcskérdés annak felmérése, hogy a megfelelő korpusz beszerzése milyen költséggel és munkaigénnyel jár.

A második mérlegelendő szempont a releváns kutatási kérdés megfogalmazása. E tekintetben a társadalomtudományi hátterű szövegbányászati irodalom számos hasznos példát kínál. Lehet vizsgálni, hogy ki a szerzője egy adott szövegrészletnek (Kapronczay–Plangár, 2017), milyen családnevek, földrajzi nevek vagy szervezetek szerepelnek gyakran egy korpuszban (Sebők et al., 2015; Sebők–Mészáros–Kis, 2018) vagy hogy milyen témák jellemeznek egy korpuszt, s ezek hangsúlyai hogyan változnak idővel. Az, hogy az ilyen és ezekhez hasonló kutatási kérdéseket milyen módszertannal érdemes vizsgálni, már átvezet a kutatási terv összeállításának – harmadik fontos – eleméhez.

Az itt bemutatott egyszerűbb elemzésekhez képest a kvantitatív szövegelemzés főárama a gépi tanulásra épít. Ennek két fő ága van: a felügyelet nélküli és a felügyelt tanulási megoldások. Ezek közül a CAP adatai alapján az előbbit megidézve a *9.* és a *10. ábra* azt mutatja be, hogy ha tíz témát szeretnénk kijelölni két magyar parlamenti ciklus törvényhozási termésében, akkor milyen kulcsszavak határozzák meg ezeket. A felügyelet nélküli eljárás itt arra utal, hogy az algoritmus fontosabb jellemzőinek meghatározásán (témák száma, mintavételi eljárás stb.) túl érdemi kutatói közreműködés nélkül, pusztán statisztikai alapon milyen csoportok vagy klaszterek képezhetők a teljes szövegállományból.

9. ÁBRA ■ Két téma az 1990 és 1994 közötti törvények topikmodelljéből

A két pár ábrán (melyek a jogtudományi szempontból legfontosabb témacsoportokat emelik ki) látszik, hogy több mint húsz év távlatában nemcsak a gyakran szereplő, s ezért meghatározónak bizonyuló kulcsszavak változnak, de a többé-kevésbé koherensnek tekinthető témák határai és az ezeket létrehozó kisebb szakterületek is (mint amilyen a közbeszerzések egyszerre jogi és gazdasági témaköre).

Az ábrákon látható szófelhők egyben arra is utalnak, hogy egyre nagyobb figyelmet kap a kvantitatív szövegelemzési folyamat önálló elemeként az adatvizualizáció (egy példa erre Balogh–Fülöp–Szabó, 2017). A nagy méretű szöveges adatbázisok tartalmának

10. ÁBRA ■ Két téma a 2014 és 2018 közötti törvények topikmodelljéből

áttekintésére ezek a képek vagy akár a videók sokszor hasznosabbak, mint a csak a gyakorlati szövegbányászati tapasztalatokkal rendelkező kutatók számára könnyen értelmezhető dokumentumkifejezés-mátrixok vagy a korpuszokat leíró kvantitatív mérőszámok. A vizualizáció így egy újabb eszköz a kvantitatív szövegelemzés eredendően interdiszciplináris eszköztárában.

■ AJÁNLOTT IRODALOM

Grimmer, Justin – Stewart, Brandon M. (2013): Text as Data. The Promise and Pitfalls of Automatic Content Analysis Methods for Political texts. *Political Analysis*, vol. 21, no. 3, 267–297.

Jakab, András – Dyevre, Arthur – Itzcovich, Giulio (szerk.) (2017): *Comparative Constitutional Reasoning*. Cambridge, Cambridge University Press.

Sebők Miklós (szerk.) (2016): *Kvantitatív szövegelemzés és szövegbányászat a politikatudományban*. Budapest, L'Harmattan.

Tikk Domonkos (szerk.) (2007): *Szövegbányászat*. Budapest, Typotex.

Ződi, Zsolt (2014b): Analysis of Citation Patterns of Hungarian Judicial Decisions. Is Hungarian Legal System Really Converging to Case Laws? *SSRN Electronic Journal, bit.ly/37PIozu*.