Doménspecifikus korpusz építése és validálása

Dodé Réka

ELTE BTK Nyelvtudományi Doktori Iskola
kovacs.reka@nytud.mta.hu

Kivonat: Egy terminuskivonatoló fejlesztésének első lépése az adott domént lefedő korpusz építése, minél relevánsabb és minél több anyagból. Elegendő nagyságú korpusz kézzel történő gyűjtése azonban időigényes. Jelen tanulmány a disszertációmhoz szükséges korpusz építésének módszertani vizsgálata, amelyben az ökoinnováció doménhez tartozó szövegeket gyűjtök és értékelek, oly módon, hogy hasonlóságot mérek a letöltött anyagok között. A szövegek letöltése egy filtereket alkalmazó programkóddal történt. A szkript magyarra történő átültetése után szükség volt egy kezdő URL-lista, illetve egy mintákat tartalmazó fájl összeállítására. A kezdő, manuálisan összeállított lista alapján a program letöltötte a szövegeket, amelyeket az előre megadott minták alapján szűrt. A 2000 szövegből a validáláshoz 20 darab szöveget választottam. Referenciaszövegeknek a legtöbb mintát tartalmazó 4 szöveg lett kijelölve. A szövegek előfeldolgozását követően a JSI Similarity programmal számoltam hasonlóságot (vektorok közötti koszinuszhasonlóság). Nem parametrikus tesztet használtam annak megállapítására, hogy a hasonlósági értékek eltérnek-e az előre megállapított 0,5-es küszöbtől. Egyik szöveg esetében sem találtam szignifikáns eltérést a 0,5-es értéktől. A szövegek szószáma és hasonlósága között enyhe negatív kapcsolatot találtam, amelyet egy kiugró szószámmal rendelkező szöveg okozott.

1 Bevezetés

Napjainkban az interneten fellelhető tartalmak tömegesen elérhetők, és némi ráfordítással milliárdos nagyságrendű korpuszok állíthatók elő. Ezek előnyei, hogy számos nyelvi jelenség megfigyelhető bennük, hátrányuk viszont, hogy sokféle szöveget tartalmaznak, így nem tudunk például ezeken olyan ún. doménspecifikus osztályozókat tanítani és építeni, amelyek csak az adott típusú szövegre jellemzőek.

Jelen tanulmány a doktori disszertációmhoz tartozó előzetes kutatás, mivel értekezésem a magyar szövegeken történő terminuskivonatolással foglalkozik, és a terminuskivonatoló alkalmazás fejlesztésének első lépéseként szükség van egy olyan korpuszra, amely lehetőleg minél relevánsabb és minél több anyagot tartalmaz a vizsgált doménból. (Nagy 2012).

A disszertációm azökoinnováció tárgykörrel foglalkozik, a domén szövegeiből áll majd a használt doménspecifikus korpusz. Az ökoinnováció mint fogalom nem feltétlenül egyértelmű, így nézzük, hogyan definiálja az ökoinnovációt az Európai Bizottság *Az ökoinnovációs cselekvési terv*ben (Eco-AP 2011: 3).

> „Az ökoinnováció az innováció minden olyan formája, amelynek eredménye vagy célja a fenntartható fejlődés irányába történő jelentős és igazolható előrelépés a környezeti hatások csökkentése, a környezetterheléssel szembeni ellenálló képesség növelése, vagy a természeti erőforrások hatékonyabb és felelősségteljesebb felhasználásának megvalósítása révén" (Eco-AP 2011: 3).

A választott domén korszerű technológiákkal (hulladékkezelés, alternatív energiák stb.) kapcsolatos szövegeket, anyagokat foglal magában, így egyrészt folyamatosan változik, másrészt jelenleg még nem áll rendelkezésre nagy mennyiségű magyar nyelvű írásos anyag a témában, ami még inkább megnehezíti a kutatást. Mivel azonban innovatív és kurrens témáról van szó, a későbbiekben rendkívül hasznossá válthatnak a kutatás eredményei.

A tanulmány két fő részből áll: az első felében (2. fejezet) a korpusz építését részletezem, azon belül is a speciális mintákat használó keresőmotor működését, a második részben pedig a korpusz validálásának módszertanát és eredményeit (3. fejezet). A validáláshoz a korpusz egy részét használtam, amely így a 2000 szöveg 1%-ából, azaz 20 véletlenszerűen kiválasztott szövegből, illetve 4 referenciaszövegből áll. Ezek között a szövegek között mértem hasonlóságot a JSI Similarity Service[i] segítségével. Az eredményeket és következtetéseket tartalmazó részben (4. fejezet) részletesen bemutatom a kapott eredményeket, melyeket statisztikailag elemeztem.

A kutatáshoz két hipotézist állítottam fel.

(1) A letöltött szövegek elérik a 0,5-es hasonlósági értéket.
(2) Minél kevesebb szövegszóból áll a szöveg, annál kisebb lesz a vektorok közti hasonlóság értéke az előre kijelölt referenciaszöveghez képest.

2 A korpusz építése

Ha nem az a célunk szövegek letöltésekor, hogy a gyűjtés különösebb tematikus megkötéssel történjen, a letöltés könnyedén kivitelezhető. Egy bizonyos doménhez tartozó korpusz építéséhez azonban már különböző módszerek alkalmazására van szükség. Ezekhez a módszerekhez pedig már szükség van valamilyen korpuszvezérelt eredmény, eszköz, forrás felhasználására, mint például a WordNet, egy előre meghatározott taxonómia vagy egy n-gram statisztika (vö. Remus–Biemann 2016; Safran et al. 2012; Chakrabarti et al. 1999), amely kevésbé támogatott nyelvek esetében nem minden esetben áll rendelkezésre. Az irányított webkeresés (focused crawling, topical crawling, directed crawling) egy olyan webes letöltési folyamat, amely irányított módon specifikus témakörre fókuszál (Remus–Biemann 2016: 3607).

Jelen doménspecifikus korpusz építése egy irányított keresőmotor segítségével történt, amelyet Gregory Greffenstette és Lawrence Muchemi mutattak be 2016-ban egy nemzetközi konferencián (Greffenstette–Muchemi 2016). Greffenstette-ék keresőszkriptje angol nyelvre volt megírva, így azt először némileg módosítani kellett. Ez csupán annak a lépésnek az eltávolítását jelentette, amelyben a program angol szavakat keresett a szövegekben (quickEnglish.awk). A magyarban ezt a lépést amiatt nem tartottuk különösképpen fontosnak, mivel a kulcsszavak listája 48 elemből állt, ezért kisebb volt rá az esély, hogy idegen nyelvű oldalak is bekerülnek a letöltött szövegek közé. A későbbi kutatásokban azonban tervezem alkalmazni.

[i] Elérhető: http://aidemo.ijs.si/xling/wikipedia.html. Letöltve: 2017. július 14.

A szkript shell programozási nyelvben van megírva, és az awk (szövegfeldolgozó és programozási nyelv), illetve fgrep (szövegfeldolgozó) parancsokkal dolgozza fel az előre megadott szöveges mintákat: az URL-eket és a kulcsszavakat. Az awk parancs a UNIX-rendszerek része, így nem szükséges külön installálni. Az awk és az fgrep nagyon hatékony parancsok, amelyeket szöveges fájlok feldolgozására, például minták keresésére használnak. A szkriptben jelen futtatáskor maximum 2000 oldal letöltése volt beállítva paraméterként. A kód egy része azt a célt szolgálta, hogy egy oldal csak egyszer kerülhessen bele a letöltött szövegek közé – amennyiben megtalálható volt benne a mintafájl legalább egy mintája.

A futtatáshoz szükség volt egy 40 elemből álló kezdő URL-listára, amelyet kézzel állítottam össze – a böngészőben az „ökoinnovációra" kapott találatokból kiindulva, illetve egy mintákat tartalmazó fájlra, amelyek gyakorlatilag kulcsszavak, kulcskifejezések és azok variánsai voltak a weboldalakból kiindulva. Összesen 48 kifejezést tartalmaz, amelyek magukba foglalják a kisbetűs, nagybetűs, kötőjeles, kötőjel nélküli variánsokat, a gépelési hibákat (rövid, hosszú magánhangzó) és néhány eltérő szótővariánst. Ezek felvételére a karakterek pontos egyezése miatt volt szükség (pl. *fenntartható energia, Fenntartható energia, ökoinnováció, öko-innováció, napenergia, napenergiá)*. Az eredeti Greffenstette–Muchemi-féle korpusz-összeállításban a mintafájl csupán 2 elemű volt, a domén kis és nagybetűs variánsa. A szkript a letöltött kezdő weboldalakon található szövegek, illetve az ott található URL-ek letöltése után a mintafájlok alapján elkülönítette a jó szövegeket a nem megfelelő szövegektől annak alapján, hogy megtalált-e benne legalább egy kulcskifejezést a felsoroltak közül. A 2000 letöltött szöveg hossza 2 625 164 szövegszó lett.

3 A részkorpusz validálása

A kutatás második felében az összeállított korpuszt validáltam. Azt a tanulmány első felében láthattuk, hogy a letöltés és a szövegek kiválasztása egy mintákat tartalmazó fájl alapján történt, mégis azt feltételeztem, hogy a validálásra szükség lehet ahhoz, hogy ténylegesen úgy kezelhessük az összeállított korpuszunkat, mint az ökoinnovációs domént reprezentatívan lefedő szövegek halmazát, amelyben a szövegek hasonlóak, tehát ugyanahhoz a doménhez tartoznak. A következőkben a validálás módszerét fejtem ki.

3.1 A részkorpusz összeállítása

Mind a 2000 szöveg összehasonlítására az idő hiányában nem volt lehetőségem, így 20 szöveget választottam véletlenszerűen (randomgenerátor), csupán azzal a feltétellel, hogy a szöveg szószáma haladja meg az 500-at. Ezt azért tartottam fontosnak, hogy az összehasonlításnál kellő szó álljon rendelkezésre, illetve mert azt feltételeztem, hogy a kevés szövegszóból álló fájlok olyan kezdőoldalak, amelyeken a szövegek inkább csak menüpontok, illetve egyéb nem releváns pontok, mint Kapcsolat / Elérhetőség, Szervezetről / Rólunk stb.

Az összehasonlítandó részkorpusz összeállítása után kerültek kiválasztásra a referenciaszövegek. A 2000 szövegből összesen négyet választottam referenciaszövegnek annak alapján, hogy egyrészt nem tartoznak a már kiválasztott szövegek közé, illetve ezekben fordult elő a legtöbb kulcskifejezés.

Mind a részkorpusz szövegeit, mind a referenciaszövegeket tokenizáltam, lemmatizáltam, és kiszűrtem belőlük a gyakori nem tartalmas szavakat (stopszószűrés). A tokenizálást és lemmatizálást a 2016-ban elkészült *e-magyar* elemzőláncban lévő eszközökkel, az *emToken* és *emLem* eszközökkel végeztem (Mittelholcz 2017, Váradi et al. 2017). Az így kapott – stopszavak nélküli, szótövesített – szövegeken számoltattam aztán hasonlóságot. A 20 random kiválasztott szöveg átlag hossza stopszavak nélkül 1428 szövegszó, a szövegek hosszának szórása 1054,49. A 4 referenciaszöveg átlag hossza 4936, a szövegek hosszának szórása 3062,63.

3.2 A hasonlóság mérése

A hasonlítás során minden szöveget összevetettem egymással, és a kapott eredményeket elemeztem. A hasonlítás a JSI Similarity Service webszolgáltatással történt, amely a 2012–2014 zajló XLike projekt keretében készült. Az XLike projekt célja egy olyan nyelvtechnológiai infrastruktúra készítése volt, amely hatékonyan nyer ki információkat különböző nyelvű szövegekből, és a nyelveken átívelő adatokkal segíti a különböző eszközök fejlesztését pl. hasonló cikkek ajánlását különböző nyelveken (García-Cuesta et al. 2014: 10). Ennek az infrastruktúrának az egyik eleme a JSI Similarity Service, amely (akár eltérő nyelvű) szövegek hasonlóságát állapítja meg 0 és 1 közé eső mértékben.

3.2.1 A korpusz- és dokumentumhasonlítás elméleti háttere

Nem új keletű igény, hogy korpuszokat, dokumentumokat képesek legyünk objektíven összehasonlítani,. Adam Kilgarriff 2001-es tanulmányának célja olyan különféle módszerek bemutatása, amelyekkel korpuszokat lehet egymással összehasonlítani, minél objektívebb módon (Kilgarriff 2001). Általában elmondható, hogy az összes módszer a szövegek szavainak előfordulásával számol különböző módokon. Kilgarriff először a statisztikai módszereket ismerteti: khi-négyzet-próba, t-teszt, kölcsönös információsó (MI), log-likelihood, Fischer-egzakt-teszt, TF-IDF, melyek közül a Mann–Whitney rangsorolási tesztet emeli ki, majd a valószínűségi eloszlás három módszerét: a Poisson-eloszlást, a binomiális eloszlást és a normál eloszlást. (Ezek a szavak gyakoriságával és azzal a ténnyel számolnak, hogy a tartalmas szavak halmozottan fordulnak elő, szemben a grammatikai szavakkal.) Végül az emberi interpretációs módszerek közül mutat be néhányat (előre meghatározott fogalmak, dimenziók nyelvi megnyilvánulásai). Ezt követően a korpuszhasonlóság, korpuszhomogenitás számolására javasol megoldást. Ehhez elkészíti a „Known-Similarity Corporát", a korpuszhasonlóság mérésére szolgáló gold standardot, majd az 500 leggyakoribb szón Spearman-féle rangkorrelációt, khi-négyzet-próbát és keresztentrópiát számol a hasonlósági értékhez. Bemutatja, hogy a módszerek közül a khi-négyzet-próba teljesített a legjobban.

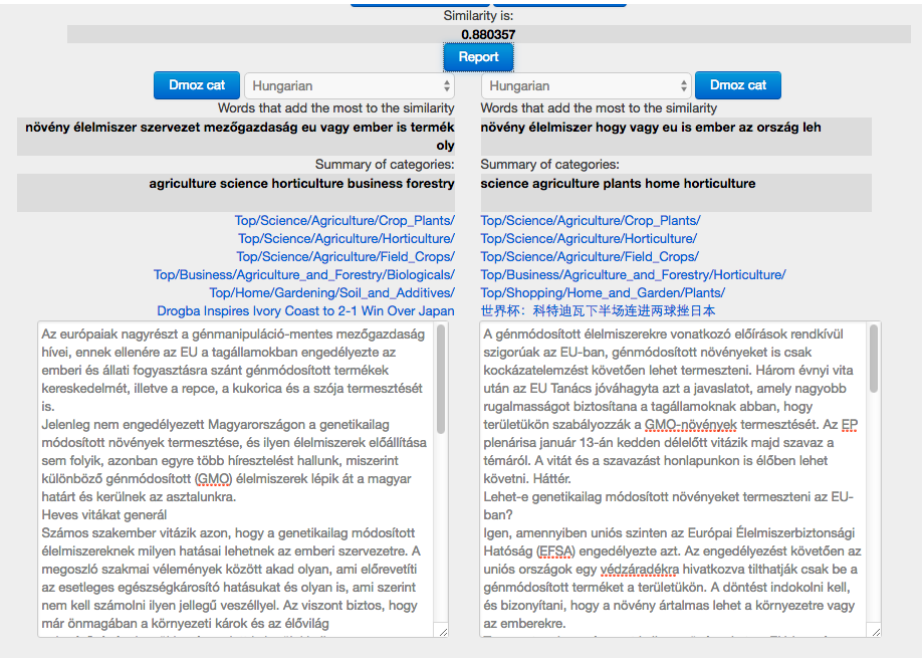
A szövegek, dokumentumok hasonlóságának számolása az általam használt módszerrel egészen a disztribúciós szemantika elméletéig, annak háttere pedig Saussure-ig vezethető vissza, aki megfigyelte, hogy a szavak jelentésének kulcsa a szavak funkciójának különbsége (Saussure 1916/1983). Ennek két típusa van, a szintagmatikus és a paradigmatikus. A szintagmatikus kapcsolat egy szó helyzetére vonatkozik, azon entitások kapcsolatára, amelyek együtt fordulnak elő a szövegben

(szekvenciális kombinációk). Ilyen például egy mondat a szövegben. A paradigmatikus kapcsolat ezzel szemben a szó helyettesítésére vonatkozik, azon entitások kapcsolatára, amelyek ugyanabban a kontextusban fordulnak elő, de nem ugyanakkor (pl. tej, kávé, tea, szörp). A disztribúciós hipotézis Harris nevéhez fűződik. A hipotézis szerint a szavak, amelyek hasonló kontextusban jelennek meg, szemantikailag kapcsolódnak egymáshoz (Harris 1968, 1970). A disztribúciós hipotézis a korpusznyelvészet megjelenésével tudott teret hódítani, mivel modellje az empirikus (korpuszok által szolgáltatott) adatoknak egy geometrikus, vektoralapú interpretációján alapul (Sahlgren 2008). A vektorok és az azokból álló mátrixok képesek dokumentumokat reprezentálni a szavak előfordulási gyakorisága szerint.

3.2.2 A hasonlóságot mérő eszköz

Az XLike dokumentumindexelője, a JSI Similarity Service fogalomalapú módszereket használt a dokumentumok összehasonlítására. E módszer a dokumentumokat egy többdimenziós térbe transzformálja oly módon, hogy a dokumentumok kifejezéseinek súlyai, fontossága (term weight) által reprezentált vektorokat egy ún. illesztett konceptuális vektorrá transzformálja a többnyelvű konceptuális térben, és azokat veti össze egymással (vektortér-modellezés). Ezek a vektorok a fogalmak egyfajta leírásai különböző nyelveken. Ezen belül a fogalmak meghatározásához különböző elméleteket használtak: K-középpontú klaszteranalízist, látens/rejtett szemantikai indexelést, explicit szemantikai analízist és kanonikus korrelációelemzést. A vektorok hasonlóságát a koszinusztávolsággal számolják, amely a két vektor által bezárt szög koszinusza (Rettinger et al. 2012).

Az 1. ábrán látható 0,88-os hasonlósági érték két génmódosítással kapcsolatos szöveghez tartozik. Ezek tartalmilag csak némileg tértek el egymástól. A következő részben foglalkozom a 0,5-es küszöbszint kijelölésével.



1. ábra. A génmódosításról szóló szövegek hasonlósága a JSI programban

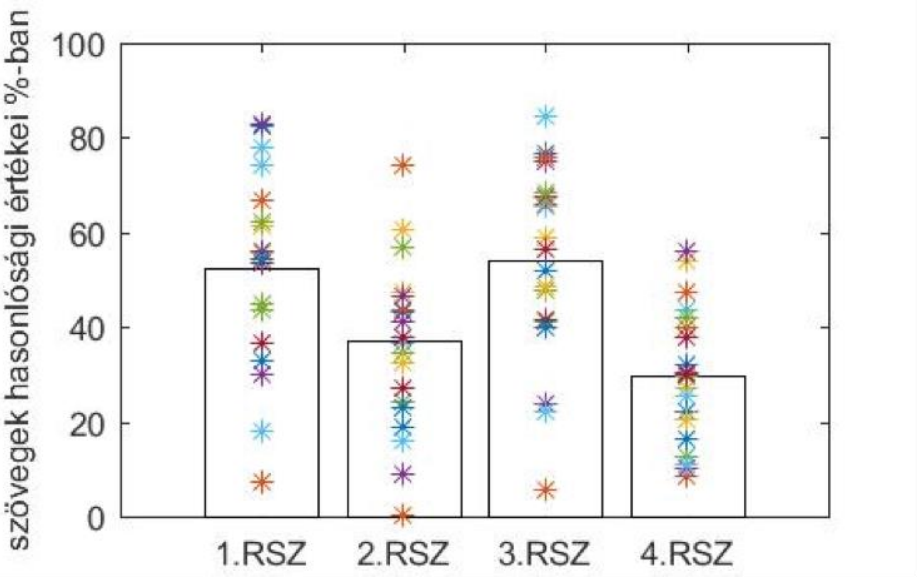
4 Eredmények és következtetések

A következőkben az eredményeket ismertetem a hipotézisek alapján, ezt követően pedig az eredmények alapján levonok néhány következtetést a kutatással kapcsolatban.

4.1 Első hipotézis

A letöltött szövegek elérik a 0,5-es hasonlósági értéket.

A küszöbérték meghatározása előzetes tapasztalatok alapján történt. Míg azok a szövegek, amelyek egymásnak fordításai voltak, 0,92–0,95-os hasonlóági értéket értek el, a tematikusan megegyező szövegek (lásd 1. ábra) 0,85–0,88, a tematikusan eltérő szövegek 0,25–0,3-es értéket. Ezeknek a tapasztalatoknak a birtokában állítottam fel a 0,5-es küszöbértéket, számolva a szövegek eltérő hosszával és az előforduló zajjal (nem releváns szövegek megjelenése). Az eredmények a 2. ábrán és az 1. táblázatban láthatók.



2. ábra. A szövegek hasonlósági értékei referenciaszövegenként százalékban

A 2. ábrán látható az összes szöveghez tartozó hasonlósági érték referenciakorpuszonként. (Az egyszerűség kedvéért a 0 és 1 közötti értékek helyett 0 és 100 közötti értékeket használok.) Az ábrán látható az átlagos érték is, illetve látható a szórás mértéke is. Az 1. táblázat ezeket számszerűsítve mutatja.

Referenciaszöveg	**1**	**2**	**3**	**4**
Átlag hasonlóság	0,524	0,373	0,541	0,297
Szórás	0,198	0,192	0,201	0,14

1. táblázat. Átlag és szórás referenciaszövegenként.

Az átlagokat tekintve látható, hogy az 1-es és a 3-as szövegek esetében átlagosan magasabb a hasonlósági érték, azonban statisztikailag számolva ez nem mondható el. Mivel az értékek nem normál eloszlásúak voltak, ezért non-parametrikus tesztet (az egymintás t-próba nem parametrikus megfelelőjét) használtam annak megállapítására, hogy a hasonlósági értékek eltérnek-e az előre megállapított 0,5-es küszöbtől. Egyik szöveg esetében sem találtam szignifikáns eltérést az 0,5-es értéktől ($p_{1\text{szöveg}} = 0{,}68$; $p_{2\text{szöveg}} = 0{,}36$; $p_{3\text{szöveg}} = 0{,}8$; $p_{4\text{szöveg}} = 0{,}22$).

A referenciaszövegek között is mértem hasonlóságot. A eredményeket a 2. táblázatban láthatók.

Referenciaszövegek	**1–2**	**1–3**	**1–4**	**2–3**	**2–4**	**3–4**
Hasonlóság	0,585	0,844	0,443	0,772	0,568	0,505

2. táblázat. Referenciaszövegek hasonlósága egymáshoz mérve

A 3. táblázat a szövegek hasonlóságainak eltérését mutatja referenciaszövegenként. Ennek megállapítására a páros t-próba nem parametrikus formáját használtam.

Referenciaszövegek		p-érték	Szignifikáns (<0,05*, <0,01**, <0,001***)
1.	2.	0,021	*
1.	3.	0,695	
1.	4.	<0,001 (0,0006)	***
2.	3.	0,011	*
2.	4.	0,223	
3.	4.	<0,001 (0,0004)	***

3. táblázat. Referenciaszövegek hasonlóságának eltérése a szövegek hasonlóságának függvényében

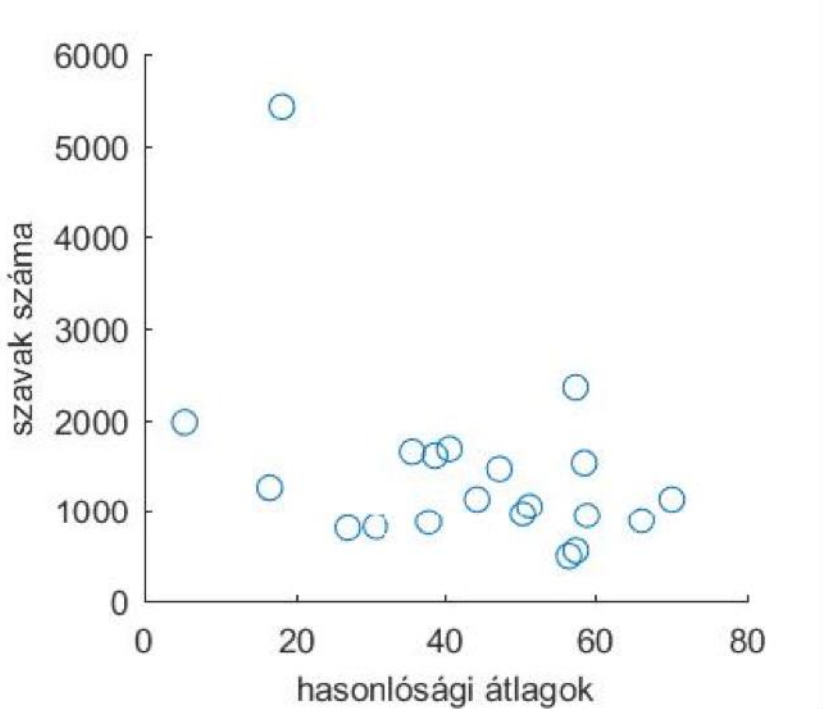
A táblázatból látható, hogy az 1-es, a 4-es, illetve a 3-as és a 4-es referenciaszövegeket között erős szignifikáns, míg az 1-es és 2-es, továbbá a 2-es és a 3-as szövegek között szignifikáns eltérés van.

Az első hipotézis részben teljesült, mivel egyik referenciaszöveghez mért átlag hasonlósági érték sem lett magasabb, mint 0,5.

4.2 Második hipotézis

A második hipotézisem, hogy minél kevesebb szövegszóból áll a szöveg, annál kevesebb lesz a hasonlósági értéke a referenciaszöveghez képest.

A második hipotézis tehát a szövegek szószáma és a hasonlóság mértékének kapcsolatára vonatkozott. Ennek megállapítására korrelációs analízist használtam. Ebből az látszott, hogy egy enyhe tendenciózus ($0,05 < p < 0,1$) negatív kapcsolat van közöttük ($r_{Pearson} = -0,41$, $p = 0,071$), amely ellentétes a várttal, tehát minél több szóból áll a szöveg annál kevésbé hasonló, viszont ahogy a 3. ábrán látható, ezt az eltérést egyetlen kiugró szószámmal rendelkező szöveg okozta. Ennek a szövegnek az ignorálásával az értékek a következőképpen alakulnak: $r_{Pearson} = -0,25$, $p = 0,30$.



3. ábra. A szövegek hasonlósági értékei referenciaszövegenként százalékban

Mindezek alapján elmondható tehát, hogy a második hipotézis nem teljesült, mivel a szövegek mérete és az eredmény között ugyan mutatkozott egy gyenge negatív kapcsolat, azt azonban egyetlen szöveg okozta.

5 Összefoglalás és kitekintés

A kutatás a doktori disszertációmhoz kapcsolódó pilot kutatás volt, melyben egy ökoinnovációhoz tartozó korpuszt állítottam össze egy kezdő URL-lista és egy mintákat tartalmazó fájl alapján. Ezt követően a korpusz egy részkorpuszán végeztem összehasonlítást a szövegek között, melynek során arra voltam kíváncsi, hogy az egyébként irányított módon letöltött szövegek valóban hasonlóak-e, azaz valóban az ökoinnováció tárgykörhöz tartoznak. Az első hipotézis nem igazolódott be, de összességében elmondható, hogy a letöltéshez használt szkript robusztus, egyszerűen átültethető más nyelvekre, és előnye más irányított webkeresővel szemben, hogy nincs szükség a szövegek előzetes elemzésére, illetve más módszerek, eszközök, források felhasználására, csupán egy átgondolt kulcskifejezés-lista összeállítására. A lista összeállításakor számolnunk kell a minták karakterszintű egyezésével, tehát a kulcsszavak variánsait (kisbetű/nagybetű, eltérő helyesírás, eltérő szintaktikai szerkezetek) is érdemes felvenni, illetve átgondolni a túl általános vagy épp túl speciális kulcsszavak megadását.

A kutatás arra is rámutat, hogy mindezek mellett is szükség lehet a korpusz validálásra. A JSI Similarity Service – bár nyelvek közötti hasonlóság számolására készült – egy nyelven belül is használható. A vektortérmodell mint szöveg-összehasonlítási módszer jó irány, mivel bár a szavak előfordulási gyakoriságát használja, a vektorreprezentációk miatt képes egyéb szemantikai információkat is kezelni (pl. szinonimitás). A későbbiekben érdekes volna további szöveg-összehasonlítási módszereket alkalmazni, mint például a khi-négyzet-próba.

A szövegek és a referenciaszövegek is eltéréseket mutattak, ami az ökoinnováció domén ismeretében annak tudható be, hogy rendkívül széles skálán mozognak az aldomének. Láthatjuk az ökoinnováció meghatározásából, hogy a megújuló energia aldomén mellett az újrahasznosítás is az ökoinnováció területe alá tartozik, miközben azok intuitíve és a szókincsben is nagy különbségeket mutatnak. Így tehát a későbbiekben megfontolandó, hogy figyelembe kell venni az ökoinnováció aldoménjeit, és esetleg szűkíteni a vizsgált területeket egy-egy aldoménre. Mindenesetre a további statisztikai vizsgálatok a minta növelését igénylik.

További érdekes kutatási módszer lehetne a szövegek összehasonlítására az emberi annotátorok bevonása, akiknek azt kellene eldönteniük, hogy a szöveg milyen mértékben tartozik a doménhez, illetve miért.

Irodalom

Chakrabarti, S., Van den Berg M., Dom B. 1999. Focused crawling: a new approach to topic-specific Web resource discovery. *Computer Networks*, 31(11): 1623–1640.

De Saussure, F. 1916/1983. *Course in general Linguistics*. London: G. Duckworth.

Európai Bizottság 2011. *Innováció a fenntartható jövőért – Az ökoinnovációs cselekvési terv (Eco-AP)*. Brüsszel, 2011.12.15. COM(2011) 899 végleges. http://kornyezettechnologia.kormany.hu/dow_nload/1/a3/40000/EcoAP.pdf. Letöltve: 2017. február 3.

García-Cuesta, E., Caparros A., Fortuna B., Carreras X., Zhang L., Li Z., Rettinger A. 2014. *Final toolkit architecture specification*. D6.1.2 XLike project. Elérhető: http://www.xlike.org/wp-content/uploads/2012/03/D6.1.2-Final-Toolkit-architecture-specification.pdf. Letöltve: 2017. február 12.

Grefenstette, G., Muchemi, L. 2016. Determining the Characteristic Vocabulary for a Specialized Dictionary using Word2vec and a Directed Crawler. In: Kernerman, I., Kosem I., Krek S., Trap-Jensen L. (szerk.) *Lexicographic Resources for Human Language Technology GLOBALEX 2016 Workshop Proceedings*. Portorož, 24 May 2016. Elérhető: http://ailab.ijs.si/globalex/files/2016/06/LREC2016Workshop-GLOBALEX_Proceedings-v2.pdf. Letöltve: 2016. december. 3.

Harris, Z. 1968. *Mathematical structures of language*. New York: Interscience Publishers.

Harris, Z. 1970. Distributional structure. In: Harris, Z. (szerk.) *Papers in Structural and Transformational Linguistics*. Formal Linguistics Series. Dordrecht: Springer Netherlands. 775–794.

Kilgarriff, A. 2001. Comparing Corpora. *International Journal of Corpus Linguistics*, 6(1): 97–133.

Mittelholcz I. 2017. emToken: Unicode-képes tokenizáló magyar nyelvre. In: Vincze, V. (szerk.) *XIII. Magyar Számítógépes Nyelvészeti Konferencia. MSZNY 2017*. Szeged: JATEPress. 61–70.

Nagy Á. 2012. *Terminológiakivonatolás francia nyelvű szabadalmi leírásokból szabály alapú és statisztikai módszerek segítségével*. PhD-értekezés. Szegedi Tudományegyetem Bölcsészettudományi Kar, Nyelvtudományi Doktori Iskola. Elérhető: http://doktori.bibl.u-szeged.hu/1768/1/disszertacio_NagyAgoston_egyben.pdf. Letöltve: 2016. május. 15.

Remus, S., Biemann, Ch. 2016. Domain-Specific Corpus Expansion with Focused Webcrawling. In: Calzolari, N., Choukri K., Declerck T., Goggi S., Grobelnik M., Maegaard B., Mariani J., Mazo H., Moreno A., Odijk J., Piperidis S. (szerk.) *Proceedings of the Tenth International*

Conference on Language Resources and Evaluation (LREC 2016). Portorož, 23–28 May 2016. Párizs: European Language Resources Association (ELRA). 3607–3611.
Rettinger, A., Zhang L., Rupnik J., Muhič A. 2012. *Cross-lingual document linking prototype.* D4.1.1 XLike project. Elérhető: http://cordis.europa.eu/docs/projects/cnect/2/288342/080/deliverables/001-D411Crosslingualdocumentlinkingprototype.pdf. Letöltve: 2017. február 12.
Safran, M. S., Althagafi A., Che D. 2012. Improving Relevance Prediction for Focused Web Crawlers. In: 2012 IEEE/ACIS 11th International Conference on Computer and Information Science. Los Alamitos, California: IEEE Computer Society. 161–167.
Sahlgren, M. 2008. The Distributional Hypothesis. *Rivista di Linguistica* 20(1): 33–53.
Váradi T., Simon E., Sass B., Gerőcs M., Mittelholcz I., Novák A., Indig B., Prószéky G., Farkas R., Vincze V. 2017. Az e-magyar digitális nyelvfeldolgozó rendszer. In: Vincze, V. (szerk.) *XIII. Magyar Számítógépes Nyelvészeti Konferencia. MSZNY 2017*. Szeged: JATEPress. 49–61.