# Kulcsszókinyerés magyar nyelvű tudományos publikációkból

Berend Gábor[1], Farkas Richárd[2]

[1] Szegedi Tudományegyetem, Informatikai Tanszékcsoport,
6720 Szeged, Árpád tér 2.
berendg@inf.u-szeged.hu
[2] SZTE-MTA Mesterséges Intelligencia Kutatócsoport,
6720 Szeged, Tisza Lajos körút 103.
rfarkas@inf.u-szeged.hu

**Kivonat:** A szöveges dokumentumokból történő kulcsszókinyerés számos alkalmazási területen hasznosítható, a katalogizáló és kivonatoló rendszerektől kezdve egészen az információ-visszakereső módszerekig. Különösen igaz mindez a tudásalapú társadalom sajátosságaiból adódóan a tudományos publikációkra: [1] szerint a 2006-ban megjelenő publikációk száma meghaladta az 1,3 milliót; [8] szerint pedig az elkészült tudományos munkák közel 50%-a olvasatlanul marad. A publikációk számának folyamatos és markáns növekedése miatt feldolgozásuk mára csupán automatikus eszközök segítségével képzelhető el. Jelen cikk az első magyar nyelvű tudományos publikációkra kidolgozott kulcsszókinyerő rendszert mutatja be.

## 1 Bevezetés

Kulcsszavak alatt az egyes dokumentumok tartalmát tömören és jól reprezentáló fogalmakat értjük, az automatikus kulcsszókinyerés alapfeladata pedig ezen fogalmak egy halmazának számítógépes meghatározása a dokumentum tartalma alapján. Habár a tudományos publikációk hatékony feldolgozhatóságához ezek szerzőik által történő megadása hasznos lenne, legtöbb esetben mégsem találkozhatunk a cikkekhez rendelt kulcsszavakkal. Jelen cikkben bemutatjuk az első automatikus kulcsszókinyerő rendszert, amelyet magyar nyelvű tudományos publikációkra dolgoztunk ki. Tanított modellünk és rendszerünk kipróbálható és elérhető a Creative Commons licenc alatt a www.inf.u-szeged.hu/rgai/kpe url-ről.

A szöveges dokumentumokból történő kulcsszókinyerés felhasználhatósága széles körökre terjed ki – a katalogizáló és kivonatoló rendszerektől kezdve egészen az információ-visszakereső alkalmazásokig. Különösen igaz mindez a tudás alapú társadalom sajátosságaiból adódóan a tudományos publikációkra: Björk és társai [1] szerint a csupán 2006-ban megjelenő publikációk száma meghaladta az 1,3 milliót; Taleb [8] szerint pedig az elkészült tudományos munkák közel 50%-a olvasatlanul marad. A publikációk számának folyamatos és markáns növekedésének fényében feldolgozásuk mára csupán automatikus eszközök segítségével képzelhető el.

Kulcsszókinyerő rendszerünk eredményességének teszteléséhez két eltérő témájú (politika- és neveléstudományi), szerzői kulcsszavakkal ellátott cikkeket tartalmazó újságra támaszkodva építettünk adatbázist.

Az így előálló adatbázis segítségével – a korábbi kulcsszókinyeréssel foglalkozó munkák többségéhez hasonlóan [4, 7, 9] – mi is felügyelt tanulási feladatot fogalmaztunk meg, ahol a cél a szerzői kulcsszavak szövegből történő kinyerése volt. Munkánk során az eddigi – a feladat angol nyelven történő megvalósítására kialakított – jellemzőkészletet bővítettük ki, illetve adaptáltuk a magyar nyelv sajátosságaihoz alkalmazkodva. Az egyes cikkekből kinyert kulcsszavak meghatározását ezen kibővített jellemzőkészlet segítségével, a belőlük lehetségesnek tartott frázisok kulcsszóként való viselkedésének poszteriori valószínűségének (különböző modellek szerint történő) kiszámítását alapul véve hajtottuk végre.

A keresztvalidáció során tapasztalt eredmények alapján kijelenthető, hogy az angol nyelvre szánt megközelítések – a megfelelő átalakítások után – magyar nyelvre is átültethetők, a jellemzőkészlet kiterjesztésével pedig további javulások érhetők el az eredményekben.

## 2 Kapcsolódó munkák

A kulcsszógenerálást végző tanuló algoritmusokat a bennük használt felügyelet mértékén túl működési alapelvük alapján különböztethetjük meg. A két nagy irányzat a kulcsszóajánlás és a kulcsszókinyerés.

A *kulcsszóajánló* rendszerek működési elve, hogy egy szóban forgó dokumentum címkéinek meghatározásához az adott dokumentumhoz bizonyos szempontok alapján hasonló dokumentum kulcsszavai közül választ. Az ilyen eljárások előnye, hogy a hasonló dokumentumból vett kulcsszó nem feltétlen van jelen a vizsgált dokumentumban, vagyis támogatja az absztrakt (a vizsgált dokumentum szövegében ténylegesen elő nem forduló) kulcsszavak generálását. Ugyanakkor az efféle eljárások hátránya is épp abból ered, hogy az ajánlott kulcsszavak a hasonló dokumentumok kulcsszavai közül kerülnek ki, vagyis csak a dokumentumhalmaz szintjén legalább egy alkalommal kulcsszóként definiált kifejezések kinyerésére képes, a kulcsszavak dinamikájához nem tud alkalmazkodni.

A *kulcsszókinyerő* módszerek az előbbiekkel ellentétben az aktuálisan vizsgált dokumentum szövegéből nyerik ki a kulcszavakat, olyan módon, hogy egy alkalmasan választott stratégia mellett legenerálják a dokumentum összes potenciális kulcsszójelöltjét, majd ezeket egy gépi tanulási modell alapján rangsorolják, a rangsor első elemeit pedig kulcsszóként kezelik. Ebben az esetben már nem áll fenn az a megszorítás, hogy egy tanítóhalmazbeli dokumentum tényleges kulcsszavai között szerepelnie kelljen a kulcsszójelölteknek, tehát az ebbe a csoportba tartozó algoritmusok képesek alkalmazkodni a kulcsszavak időbeli változásához. Ugyanakkor az is elmondható, hogy mivel a kulcsszójelöltek egytől egyig a dokumentum szövege alapján lettek generálva, így olyan kulcsszavak kinyerésére, amelyek a dokumentum szövegében nem voltak leírva, az ilyen eljárások önmagukban nem képesek, ráadásul ha a cél egy teljes dokumentumhalmaz felkulcsszavazása, a szinonimák és szemantikailag hasonló kulcsszavak egységes kezelésének is külön figyelmet kell szentelnünk.

A kulcsszavak meghatározásának egyik leggyakoribb célja a tudományos publikációk kulcsszavakkal való ellátása, az idei évben a SemEval konferencia egy versenyt is hirdetett a témában [5]. Ezen felül számos publikáció jelent már meg a kulcsszavazás témakörén belül speciálisan az angol nyelvű tudományos publikációk kulcsszavazásával foglalkozva. [4] webes keresésekkel igyekezett pontosabb eredményekre jutni, míg [7] a tudományos cikkek strukturáltságát és a bennük szereplő rövidítések fontosságát hangsúlyozta.

# 3 Módszertan

Az automatikus címkézés magyar nyelvre való adaptálásához felügyelt tanítás mellett végrehajtott *kulcsszókinyerési* módszertant alkalmaztunk. Tanuló algoritmusnak a generatív logisztikus regressziót választottuk a kulcsszójelöltek poszteriori valószínűségének meghatározására. Ebben a fejezetben részletesen bemutatjuk a jellemzőtér építését megelőző előfeldolgozó lépéseket, valamint magát a jellemzőteret.

## 3.1 Előfeldolgozás

Az előfeldolgozás magában foglalta a „nem hasznos” dokumentumrészek elhagyását, a fejezethatárok meghatározását, valamint a kulcsszójelöltek és velük kapcsolatos statisztikák kigyűjtését.

A PDF formában fellelhető publikációk feldolgozásának első lépése a szöveges tartalmuk kinyerésére irányult, melyhez a szabadon elérhető PDFBox[1] konvertert használtunk. Következő lépésként a szöveg megtisztítását hajtottuk végre: az egyes dokumentumon belül túl gyakran előforduló szövegrészeket – mint amilyenek a fejlécben található szövegek – egyszerűen eltávolítottuk a feldolgozandó szövegek közül, ily módon megtisztítva szövegeinket a fölöslegesen mondatközbe beékelődő szövegektől. Más jellegű megközelítést igényelt a táblázatok tartalmának kezelése, melyek szintén képesek mondatok belsejébe beékelődni a PDF konvertálása során. Ezeket a tartalmakat a sorok (token- és karakter-) hosszára számított statisztikákat és reguláris kifejezéseket alkalmazva lehetett eredményesen eltávolítani a folyó szövegből. A következő feladat az előálló nyers szövegek nyelvi elemzése volt. A nyelvi elemzést a magyarlanccal [10] végeztük el, amely egyúttal a lemmatizálásért is felelős volt.

A szövegre irányuló előfeldolgozó lépések végeztével tehát előálltak a dokumentumok tisztított, szöveges részeinek szófaji kódokkal ellátott verziói. Ezt követően lettek meghatározva a kulcsszójelöltek az egyes dokumentumokra. Kulcsszójelöltként kezeltük azokat az 1 és 4 tokenhossz közötti kifejezéseket, melyek melléknevekből és főnevekből álltak csupán, és se nem kezdődtek stopszóval, se nem végződtek stopszóval vagy melléknévvel.

[1] http://pdfbox.apache.org/

### 3.2 Jellemzőtér

A jellemzőtér kidolgozása során a hagyományos jellemzők beépítésén túl új jellemzők hozzáadott értékét is megvizsgáltuk. [9] az egyes kulcsszójelölteket azok tf-idf mértékével, valamint a dokumentumbeli első előfordulásuk relatív pozíciójával jellemezte. Cikkük megjelenése óta ezekre tekint a kulcsszókinyeréssel foglalkozó irodalom standard jellemzőkként. Munkájuk egy kiterjesztésében [6] már felhasználták a kulcsszójelöltek tokenszámát is a kifejezések kulcsszóként való előfordulásának leírására. A fenti jellemzők továbbiak mellett a mi rendszerünkben is szerepet kaptak.

[7] angol nyelvű számítástudományi publikációkon megmutatta, hogy a rövidítések feltérképezése segítségünkre lehet a kulcsszavak meghatározására. Ezért mi úgy jártunk el, hogy kigyűjtöttük a publikációk szövegeiből az összes olyan tokent, amely több nagybetűs karaktert tartalmazott, mint kisbetűst, egy-egy több token hosszú kulcsszójelölt esetében pedig döntést hoztunk, hogy az kiterjesztése lehet-e az őt tartalmazó dokumentum valamelyik rövidítésének, oly módon, hogy ugyanazzal a betűvel kezdődnek, majd pedig ugyanabban a sorrendben fordulnak benne elő egyazon rövidítés betűi.

A kulcsszójelöltek dokumentumbeli elhelyezkedését az első előfordulás relatív pozícióján túl a pozíciókban mutatkozó szórás értéke alapján is vizsgáltuk. Miután vettük egy kulcsszójelölt összes dokumentumbeli előfordulását, egyszerűen kiszámítottuk azok értékeiben jelentkező szórás nagyságát, és ezt az értéket adtuk meg a kérdéses kulcsszóaspiránsnak az adott jellemzőre nézve. Egy másik statisztikai módszerrel számított jellemző a PMI (Pointwise Mutual Information) [2] volt, amely a több tokenből álló kifejezések alkotóelemeinek együttes előfordulási gyakoriságának mértékét vizsgálni. A mérőszám csak akkor adott 1 értéket, ha a kifejezés minden egyes tokenje kizárólag egymást követve fordult elő a dokumentumon belül, valószínűvé téve ezáltal, hogy egy többszavas kifejezéssel van dolgunk.

A pozíciókkal kapcsolatos tulajdonságokat a tokenpozíciókon kívül szekcióbeli elhelyezkedésre is alkalmaztuk. Hasonlóan az első előfordulás tokenhossz függvényében történő relatív meghatározásához, kiszámítottuk a szekciók arányában egy kulcsszójelölt relatív pozícióját. Ezen túl, hasonlóan a tf-idf mértékhez, kiszámításra került minden kulcsszóaspiránshoz egy sf-isf (szekciófrekvencia – invertált szekvenciafrekvencia) érték is,

$$sf\text{-}isf(t_i, d_j) = sf(t_i, d_j) * isf(t_i) \quad \textbf{(1)}$$

alapján, amelyből $sf(t_i, d_j)$ azt mutatja meg, hogy a $j$-edik dokumentum szekcióinak milyen arányában van jelen a $t_i$ kifejezés, $isf(t_i)$ pedig azt határozza meg, hogy a korpuszban lévő összes szekció mekkora részében szerepel $t_i$ kifejezés.

A jellemzők egy másik fontos részhalmaza a kulcsszójelöltek nyelvi elemzéséből származó információkat használta föl. A szófaji kódokért felelős nominális jellemző az adott kulcsszójelölt szófaji kódjának sorozatát tartalmazta, a *magyar pártfejlődés* kifejezés esetében értéke *melléknév+főnév* volt. A szófajok egyszerű nyilvántartásán túl egy másik jellemző a kulcsszójelölt összes előfordulási formáján belül annak az arányát írta le, hogy mekkora valószínűséggel szerepelt a kifejezésben főnév. Ezeken túlmenően jellemzőként tekintettünk arra is, hogy az adott kulcsszóaspiráns hány elemet tartalmazott egy előre definiált magyar nyelvű stopszólistáról. A motiváció

ezen jellemző használata mögött az volt, hogy magas értéke egy kifejezés kulcsszóként való viselkedés ellen szólhat.

Fontosnak éreztük azt a tényt is egy kulcsszójelölttel kapcsolatban, hogy az azt tartalmazó mondatok közül mennyi esetében fordult elő hivatkozás. Az ez alapján a statisztika alapján számított bináris jellemző igaz értéket vett fel, ha a kulcsszóaspiráns dokumentumában legalább egy olyan mondat szerepelt, amely mind a kulcsszójelöltet, mind pedig legalább egy hivatkozást tartalmazott. Egy további, a dokumentum szerkesztéséből nyerhető információ volt számunkra, hogy egy-egy kulcsszójelölt szerepelt-e a publikáció címében, avagy legalább egy fejezet főcímében. A kulcsszójelöltekkel kapcsolatban fölhasználtuk azon információkat is, hogy más tanító adatbázisbeli dokumentumon az adott kifejezés szerepelt-e kulcsszóként, illetve, hogy a magyar nyelvű Wikipédián található-e azonos névvel szócikk.

Végül egy kifejezés kulcsszóként való szereplését egy dokumentumban nagymértékben meghatározza, hogy milyen igék társaságában szerepel együtt, azaz hogy milyen kontextusban szerepel. Éppen ezért a tanítás folyamán legyűjtöttük a tanítóhalmazban szereplő igéket és azok gyakoriságát, majd a tanítódokumentumok méretének függvényében meghatározott küszöbszámnál többször előforduló igékre szorítkozva folytattuk vizsgálódásainkat. Egy kulcsszójelölt esetében megvizsgáltuk, hogy melyek azok a tanítóhalmazon legalább százszor szereplő igék, amelyekkel a saját dokumentumában együtt előfordultak, majd azon igéknek megfeleltethető jellemzőket, ahol ez a mennyiség 0-tól különböző volt, egyenlővé tettük az együttes előfordulásuk számával.

## 4 Adatbázisok

A felügyelt tanulás végrehajtásához, valamint a kiértékelés könnyebbé tételéhez kulcsszavakkal ellátott magyar nyelvű online folyóiratokat gyűjtöttünk össze. A feladat nehezebbnek bizonyult, mint azt elsőre gondoltuk, hiszen sok publikációs forrás nem tartalmaz kulcsszavakat, mi több, a kulcsszavakat tartalmazók közül több sem volt végül alkalmas a feldolgozásra, azok többnyelvűsége (angol-magyar) vagy a PDF-dokumentumok minőségbeli hiányosságai miatt. Végül méréseinket a Politikatudományi Szemle[2] archívumán, valamint a Magyar Neveléstudományi Konferencia 2009-es konferenciakiadványa[3] alapján hajtottuk végre.

A Politikatudományi Szemle szerkesztősége a 2007-es évfolyammal kezdődően vezette be részlegesen a megjelent cikkek kulcsszavazását, így adathalmazunkat az ez idő alatt megjelent, kulcsszavakkal bíró cikkeik képezték. A PDF-dokumentumok tartalmát sima szöveges fájlba konvertáló eszköz a 86 kulcsszavazott publikáció esetében 6 alkalommal nem járt sikerrel szerkesztési hibáknak köszönhetően, így legvégül 80 dokumentum képezte vizsgálódásunk tárgyát.

A teljes dokumentumhalmazhoz 351 kulcsszó lett hozzárendelve összesen 412 alkalommal, ami átlagosan 5,15 kulcsszót jelent dokumentumonként. A kulcsszavak között 79 absztrakt kulcsszó volt, ami azt jelenti, hogy ennyi esetben volt egy kifeje-

[2] http://www.poltudszemle.hu/archivum,7.html
[3] http://www.nevelestudomany.hu/onk2009/

zés úgy megadva kulcsszónak, hogy a publikáció szövegében nincsen megemlítve maga a kifejezés. Az ilyen esetek az összes kulcsszó 19,17% tették ki tehát, méghozzá úgy, hogy az összes absztrakt kulcsszó egyedi volt abban az értelemben, hogy nem volt olyan kulcsszó, amely egynél többször fordult volna elő absztrakt minőségében (nem absztraktként ettől még többször is előfordulhatott). A 80 dokumentumból összesen 146508 kulcsszójelöltet generált rendszerünk, ami dokumentum szinten átlagosan 1831,35 kulcsszóaspiránst jelent. Generatív modellünk feladata minden dokumentumra ezeknek a jelölteknek a rangsorolása volt.

1. táblázat: A leggyakoribb politikatudományi kulcsszavak listája.

| Kulcsszó | Előfordulási gyakoriság | Absztrakt előfordulás |
|---|---|---|
| pártok | 5 | 0 |
| politikai kommunikáció | 5 | 0 |
| politikatudomány | 5 | 0 |
| kormányzás | 4 | 0 |
| média | 4 | 0 |

2. táblázat: Példa absztrakt kulcsszavakra.

| Kulcsszó | Előfordulási gyakoriság | Absztrakt előfordulás |
|---|---|---|
| nacionalizmus | 2 | 1 |
| választói magatartás | 2 | 1 |
| kormányzat minősége | 1 | 1 |
| komparativisztika | 1 | 1 |
| témakisajátítás | 1 | 1 |

A neveléstudománnyal foglalkozó konferenciakiadvány a 2009-es Magyar Neveléstudományi Konferenciára elfogadott publikációk absztraktjait tartalmazta, de a szerzők által meghatározott szabad szavas kulcsszavakat nem minden esetben (a témakörök megjelölése általános volt, de azok használata nem lett volna megfelelő számunkra). A kiadványban végül 19 darab kulcsszavazott tartalmi összefoglaló volt található, melyekhez összesen 63 egyedi kulcsszó volt rendelve, ami átlagosan 3,32 szerzői kulcsszót jelent dokumentumonként. A cikkek szerzői által meghatározott kulcsszavak közül 19 volt jelen absztrakt minőségben, a politikatudományosénál magasabb, 25,4%-os absztraktkulcsszó-arányt eredményezve mindezzel. A neveléstudománnyal foglalkozó dokumentumok esetében a rendszerünk által generált kulcsszójelöltek száma 3169 volt, ami átlagosan 166,79 kifejezést jelent dokumentumokként.

## 5 Kísérletek

A kiértékelés során keresztvalidációt alkalmaztunk mindkét korpusz esetében. A kiértékelések között egyaránt végrehajtottunk szigorú egyezésen alapulót, valamint a kulcsszavak speciális természetére való tekintettel emberi ellenőrzést is. Mindezen

felül az emberi ellenőrzés megbízhatóságának tesztelésére megvizsgáltuk az annotátorok közötti egyetértés szintjét is.

### 5.1 Annotátorok közötti egyezés

A kézi kiértékelést két nyelvész egymástól függetlenül végezte. Feladatuk az volt, hogy egy dokumentum automatikus kulcsszavazásának bírálatakor döntést hozzanak egyrészről az eredeti szerzői kulcsszavak automatikus kulcsszavakkal való lefedettségéről, valamint az automatikus kulcsszavaknak az adott dokumentum témájába vágóságáról, azaz pontosságukról.

Az annotációk közötti egyezés mértékét a politikatudományi témájú publikációk esetében egy baseline, illetve a végső módszerre is teszteltük, a neveléstudományi cikkek esetében pedig a kiterjesztett jellemzőteret használó modell kiértékelését mértük. Fontos kiemelnünk, hogy ezen mérések során a két független annotátor az automatikus rendszer által kinyert kulcsszavakról hozott döntéseket, így ezen döntések egymáshoz való hasonlóságának vizsgálatán keresztül a számukra kitűzött feladat jól definiáltságát állt módunkban megvizsgálni.

Az annotátorok közötti egyetértés jellemzését az azonos módon megítélt kulcsszavak arányán túl κ-mértékkel [3] is elvégeztük, mely egy olyan statisztikai mutató, amely a tapasztalt egyezési szintet megpróbálja korrigálni a véletlennek köszönhető egyezéssel. A κ-mérték értéke -1 és 1 között mozog, -1-et akkor veszi föl, ha az annotátorok jelölései teljes ellentétben állnak egymással, 1 értéke pedig akkor lesz, ha tökéletes összhang tapasztalható az annotátorok között. Általánosságban, a 0,4 és 0,6 közötti értékű egyezéseket közepes szintűknek, a 0,6 és 0,8 értékek között mozgókat kielégítőnek, a 0,8 és 1,0 közöttieket pedig közel tökéletes egyezésűként szokás jellemezni.

3. táblázat: Annotátorok közötti egyetértés.

| | **Egyezés** | **κ-mérték** |
|---|---|---|
| Baseline fedés (politikatudomány) | 90,42% | 0,7689 |
| Baseline precízió (politikatudomány) | 81,68% | 0,55411 |
| Végső fedés (politikatudomány) | 85,92% | 0,7223 |
| Végső precízió (politikatudomány) | 81,92% | 0,6256 |
| Végső fedés (neveléstudomány) | 92,06% | 0,8382 |
| Végső precízió (neveléstudomány) | 91,10% | 0,7206 |

### 5.2 Rendszereredmények

A szigorú egyezés alatt az automatikusan kinyert kulcsszavak és az eredeti szerzői kulcsszavak szótöveinek egyezését követeltük meg az elfogadáshoz. Ebben az esetben csak a szerzői kulcsszavakhoz mért pontos egyezés (fedés) mérésére volt lehetőség, a kizárólag az automatikus kulcsszavak között szereplő, a cikk témáját különben jól összegző kifejezések nem kerültek elfogadásra.

Éppen ezért az emberi kiértékelést is szükségesnek éreztük, hiszen könnyen előfordulhatott, hogy az eredeti kulcsszavak jelentéstartalmát lefedő, de attól eltérő formában álló kifejezést nyert ki rendszerünk, ám ilyen esetekben a szigorú egyezésen alapuló kiértékelés hibás kifejezésként kezelte az egyébként szemantikája alapján elfogadható kifejezéseket is. A szigorú, valamint a megengedőbb, emberi kiértékelést figyelembe vevő rendszereredmények a 4., illetve az 5. táblázatban olvashatók.

Baseline módszerünk jellemzőtere a standard módszert követte, azaz az egyes kifejezéseket az azt tartalmazó dokumentum alapján számított tf-idf mértékkel és dokumentumon belüli legkorábbi előfordulásának relatív pozíciójával jellemezte.

4. táblázat: A szigorú kiértékelés eredményei.

| Rendszer | Fedés |
|---|---|
| Baseline (politikatudomány) | 13,02% |
| Teljes jellemzőtér (politikatudomány) | 31,07% |
| Teljes jellemzőtér (neveléstudomány) | 30,16% |

5. táblázat: Kézi kiértékelés eredményei.

| Rendszer | Pontosság | Fedés | F-mérték |
|---|---|---|---|
| Baseline (politikatudomány) | 36,79% | 33,91% | 0,3529 |
| Teljes jellemzőtér (politikatudomány) | 49,76% | 54,12% | 0,5185 |
| Teljes jellemzőtér (neveléstudomány) | 24,15% | 46,03% | 0,3168 |

## 6 Diszkusszió

Dolgozatunkban magyar nyelvű tudományos publikációk eltérő területein (politika- és neveléstudomány) működőképes kulcsszókinyerő rendszert mutattunk be, amely eredményei a baseline rendszert jelentősen meghaladták, valamint az idei évben angol nyelvű tudományos publikációkra meghirdetett hasonló témájú verseny végeredményeit összegző cikk [5] eredményeivel összehasonlítható eredményeket produkált a szigorú kiértékelés alkalmazása mellett.

A 4. táblázat eredményeiből kitűnik az is, hogy a szigorú kiértékelés esetében (fedés) hasonló eredmények születtek mind a politikatudományi, mind pedig a neveléstudományi cikkek esetében. A 30% körül mozgó eredmények kapcsán fontos megjegyezni, hogy a két korpuszon belüli absztraktcímke-eloszlásoknak köszönhetően a szigorú egyezéssel elérhető eredmények maximális értékei 80,83%, illetve 74,6% voltak a politikatudományi, illetve a neveléstudományi cikkekre nézve.

Ha az eredményeket az 5. táblázatban látható emberi kiértékelés szintjén vetjük össze, akkor már eltérés tapasztalható – elsősorban pontosság tekintetében – a két korpuszon elért eredményességet illetően. A politikatudományi publikációk esetében tapasztalható jobb eredmények annak tudhatók be, hogy azok esetében a teljes cikk szövege rendelkezésre állt a kulcsszavak meghatározása során, míg a neveléstudományi dokumentumok között kizárólag a teljes publikációk absztraktjait tudtuk használni, jellemzőterünknek pedig a hosszabb dokumentumok (több információ) kedveznek.

Az annotátorok kiértékelései közötti egyezés elemzése is érdekességeket mutat. A 3. táblázatból kiolvasható, hogy a fedéssel kapcsolatos döntéseket minden esetben jóval nagyobb összhang mellett voltak képesek meghatározni, mint az automatikus kulcsszavak helyességére irányuló döntéseket. Az eltérő típusú jelölésekben mutatkozó különbségeken túl megfigyelhető volt még az is, hogy a jobb minőségű automatikus címkézést nagyobb annotátorok közötti egyetértés is kísérte, hiszen ekkor a több jó kulcsszó miatt vélhetően kevesebb kérdéses szituációban kellett dönteniük az annotáció során.

Jövőbeli munkaként fontosnak érezzük kulcsszókinyerő rendszerünk kiterjesztését abban a tekintetben, hogy a dokumentum szövegében le nem írt (absztrakt) kulcsszavak meghatározására is képes legyen, valamint a továbbiakban a korpuszszintű kohézió (kulcsszavak normalizálása) szem előtt tartásával is szeretnénk foglalkozni.

## Köszönetnyilvánítás

A cikk szerzői köszönettel tartoznak a kulcsszavak annotációját végző nyelvészeknek, Vincze Veronikának és Almási Attilának. A kutatást – részben – a TEXTREND kódnevű projekt keretében az NKTH támogatta.

## Bibliográfia

1. Björk, B-C., Roos, A., Lauri, M.: Scientific journal publishing: yearly volume and open access availability. Information Research, ISSN 1368-1613, Vol. 14, No. 1 (2009)
2. Church, K.W., Hanks, P.: Word association norms, mutual information and lexicography. In: Proceedings of the 27th Annual Conference of the ACL. ACL, New Brunswick, NJ. (1989) 76–83
3. Cohen, J. A.: A coefficient of agreement for nominal scales. Educational and Psychological Measurement Vol. 20 No. 1 (1960) 37–46
4. Frank, E., Paynter, G. W., Witten, I. H., Gutwin, C., Nevill-Manning, C, G.: Domain-Specific Keyphrase Extraction. In: Proceedings of the 16th International Joint Conference on Artificial Intelligence (IJCAI99) (1999) 668–673
5. Kim, S. N., Medelyan, O., Kan, M-Y., Baldwin, T.: SemEval-2010 task 5: Automatic keyphrase extraction from scientific articles. In: Proceedings of the Fifth International Workshop on Semantic Evaluations (SemEval-2010) (2010)
6. Medelyan, O., Witten, I H.: Thesaurus based automatic keyphrase indexing. In: Proceedings of the 6th ACM/IEEE-CS joint conference on Digital Libraries (2006) 296–297
7. Nguyen, T. D., Kan, M-Y.: Keyphrase extraction in scientific publications. In: Proceedings of International Conference on Asian Digital Libraries (ICADL07) (2007) 317–326
8. Taleb, N. N.: The Black Swan, Chapter 7. ISBN: 1400063515 (2007)
9. Witten, I. H., Paynter, G. W., Frank, E., Gutwin, C., Nevill-Manning, C, G.: Kea: Practical Automatic Keyphrase Extraction. In: ACM DL (1999) 254–255
10. Zsibrita J., Nagy I., Farkas R.: Magyar nyelvi elemző modulok az UIMA keretrendszerhez. In: VI. Magyar Számítógépes Nyelvészeti Konferencia (2009) 394–395