Lengyelné Molnár Tünde
Eszterházy Károly Főiskola, Médiainformatika Intézet
mtunde@ektf.hu

AZ EMBERI ÉS GÉPI REFERÁTUMKÉSZÍTÉS HATÉKONYSÁGÁNAK ELEMZÉSE

Kutatásom lényege egy automatikus kivonat előállítására képes program megírása. Ezt a feladatot nem lehet megoldani, az emberi kivonatolás mozzanatainak elemzése nélkül. Ennek elősegítésére készítettem egy empirikus vizsgálatot, melynek célja, megvizsgálni az emberi referátum-készítés sajátosságait.

Empirikus mérésem során különböző témájú szakmai cikkek kivonatának elkészítésére kértem fel több felsőoktatási intézmény könyvtár-informatika szakos hallgatóit.

Az alapul szolgáló cikkek kiválasztásakor a két legfőbb szaklap aktuális számaiból választottam egy-egy cikket, így a Könyvtári Figyelő[1], illetve a Tudományos és Műszaki tájékoztatás[2] folyóiratokból kerültek a cikkek kiválasztásra.

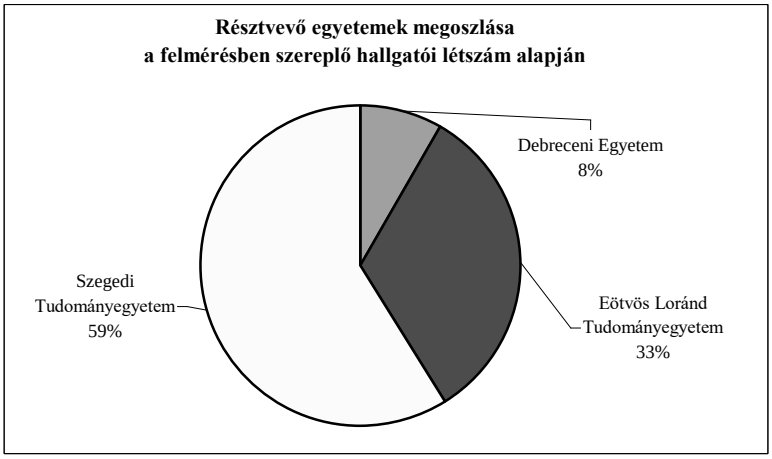
A minta meghatározásakor próbáltam egyetemi-, illetve főiskolai hallgatókat is bevonni a felmérésbe:

- Ennek eredményeként a főiskolai hallgatók saját intézményünkből (azaz az Eszterházy Károly Főiskoláról) kerültek kiválasztásra. A főiskolai hallgatók a teljesség elvével lettek felmérve, így minden[3] nappali és távoktatási évfolyam hallgatója elkészítette a felmérés alapjául szolgáló két cikk referátumát.
- Az egyetemi hallgatók bevonása során sajnos nem volt lehetőségem teljes mintavételt alkalmazni, így három fő képviselő egyetem: a Debreceni Egyetem, az Eötvös Loránd Tudományegyetem, valamint a Szegedi Tudományegyetem könyvtár-informatika szakos hallgatói képezték az alapul szolgáló populációt. A mintába az alábbi hallgatói létszámmal kerültek be:

[1] KOLTAY Tibor: Szöveg, információ, relevancia: néhány adalék a témakörhöz. – In: Könyvtári Figyelő. 2005. (51. évf.) 3. sz. pp. 514–518.

[2] PROKNÉ Palik Mária: A tartalmi feltárás problémái online könyvtári katalógusokban. – In: Tudományos és műszaki tájékoztatás. 2005. (52. évf.) 11–12. sz. pp. 525–527.

[3] Aki a felmérés időpontjában megjelent a konzultáción.



Az egyetemek hallgatóit[4] a referátum elkészítésén túl felkértem az általuk leghasznosabbnak ítélt szavak megjelölésére is. Ennek elemzését szeretném publikációmban részletesen ismertetni.

A leghasznosabbnak ítélt mondatok megjelölésekor a hallgatókat kötötte egy korlát: a szövegnek 20%-os kivonatát kellett elkészíteni, ezért csak adott számú mondatot jelölhettek meg. Ezzel szemben a hasznos szavak megjelölésekor semmiféle kikötés nem szerepelt, melynek következtében a megjelölt szavak száma széles spektrumon mozgott: volt, aki egyetlen szót sem jelölt meg, illetve a legtöbb megjelölt szó mind a két cikk esetén a szöveg 23%-át tette ki.[5] A kitöltők 16,39%-a nem jelölt meg hasznos szavakat (csak a kivonatot készítette el). A további elemzések ezen üres kitöltők figyelmen kívül hagyásával történik. Így, átlagosan a szöveg 4,6%-át jelölték meg hasznos szóként. Ez már a második érték volt, melynél elegendő egyetlen számadatot feltüntetnem, annak ellenére, hogy két cikkről beszélünk, ezért nézzük meg ennek az okát részletesebben!

Az egyik cikk 1274 szóból, a másik pedig 1160 szót tartalmazott (elhagyva a névelőket és kötőszavakat). A két szöveg szavainak száma közti különbség 9,82%.

Az első eredménynek az tekinthető, hogy a maximálisan-; illetve az átlagosan megjelölt szavak száma mindkét cikk esetében azonos érték (–0,5% és +0,05% eltéréssel). A részletek elemzése során már valamivel árnyaltabb képpel találkozunk.

Előbb nézzük meg a megjelölt szavak eloszlását Falus Iván–Oléh János kategória alkotási szabályrendszerével, mely szerint 50 körüli elemszámú minta esetén 8-9 csoportot célszerű alkotni, ahol is a csoport intervallumok nagyságának megválasztásakor a minta értéktartományából kell kiindulni, azaz „a minta legnagyobb és

[4] A Szegedi Egyetem hallgatóinak 19%-a vesz részt egyetemi képzésben, 81% főiskolai szintű képzés hallgatója.

[5] A Könyvtári Figyelőből származó cikk esetén 23,86%, míg a TMT cikkénél a szöveg 23,36% volt a legtöbb megjelölt szó.

legkisebb eleme által behatárolt zárt intervallumból”.[6] A konkrét csoportok nagyságát az alsó és felső érték közötti csoportok számának megválasztása határozta meg, melyet célszerű 1, 2, 3, 5, 10 vagy ennek többszörösére választani.

A két cikk esetén a következőképpen alakul a megjelölt szavak számának abszolút és relatív gyakorisági megoszlása a kategóriák között:

Alsó határ	Felső határ	Könyvtári Figyelő folyóirat cikke	
		Abszolút gyakoriság	Relatív gyakoriság
0 -	30	34	65,38%
31 -	60	7	13,46%
61 -	90	2	3,85%
91 -	120	4	7,69%
121 -	150	3	5,77%
151 -	180	0	0,00%
181 -	210	1	1,92%
211 -	240	0	0,00%
241 -	270	0	0,00%
271 -	300	0	0,00%
301 -	330	1	1,92%
Összesen:		52	100,00%

Alsó határ	Felső határ	Tudományos és Műszaki Tájékoztatás folyóirat cikke	
		Abszolút gyakoriság	Relatív gyakoriság
0 -	0	16	31,00%
1 -	30	20	38,46%
31 -	60	4	7,69%
61 -	90	7	13,46%
91 -	120	2	3,85%
121 -	150	0	0,00%
151 -	180	1	1,92%
181 -	210	0	0,00%
211 -	240	1	1,92%
241 -	270	0	0,00%
271 -	300	1	1,92%
Összesen:		52	100,00%

[6] Falus Iván–Oléh János: Statisztikai módszerek pedagógusok számára. – Budapest: OKKER, 2000. – p. 57.

Mivel mindkét esetben az adatok több, mint 90%-a tartozik az első 5 kategóriába, ezért elemezzük tovább egy másik csoportosításban a gyakorisági eloszlások alakulását!

Csoport határok	KF		TMT	
	Abszolút gyakoriság	Relatív gyakoriság	Abszolút gyakoriság	Relatív gyakoriság
Nincs megjelölt szó	16	30,77%	16	30,77%
1-10 db	2	3,85%	5	9,62%
11-20 db	9	17,31%	7	13,46%
21-30 db	7	13,46%	8	15,38%
31-40 db	4	7,69%	2	3,85%
41-50 db	2	3,85%	2	3,85%
51-60 db	1	1,92%	0	0,00%
61-70 db	0	0,00%	1	1,92%
71-80 db	0	0,00%	2	3,85%
81-90 db	2	3,85%	4	7,69%
91-100 db	2	3,85%	0	0,00%
Több mint 100 megjelölt szó	7	13,46%	5	9,62%
Összes:	52	100,00%	52	100,00%

Megtévesztő lehet, de a „Nincs megjelölt szó” kategóriába tartozó 16–16 fő nem azt jelenti, hogy 16-an nem töltötték ki a felmérést. Azon személyeket, akik egyik cikk esetén sem jelöltek meg szavakat, már az elemzés elején kizártam, így a fenti táblázatokban az ő adataik nem szerepelnek. 16 fő van mind két cikk esetén, akik csak az egyik cikk lényeges szavait jelölték meg, a másik cikkhez érve pedig vagy elvesztették érdeklődésüket, vagy időhiány miatt nem jelöltek meg egyetlen szót sem[7].

Vizsgáljuk meg, a szavakat megjelölő személyek között melyik kategória a leggyakoribb!

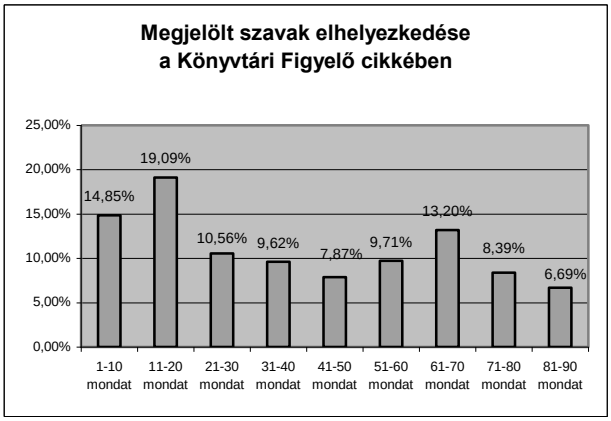
A táblázatból leolvasható, hogy a Könyvtári Figyelőből származó cikk esetén a 11 és 40 közötti szót jelölt meg a résztvevők 38%. Ha csak azon személyeket vesszük figyelembe, akik jelöltek is meg szavakat ezen cikk esetén, akkor a kitöltők több, mint 55% tartozik a 11 és 40 közötti szót lényegesnek tartó személyek közé.

A Tudományos és Műszaki Tájékoztatásból származó cikk esetén 10 szóval alacsonyabban alakulnak az értékek, így az 1 és 30 szó közötti megjelöléshez tartozik a kitöltők 38%-a, míg a cikket fel nem dolgozó kitöltőket figyelmen kívül hagyva ez az érték szintén a minta 55%-a.

[7] A kitöltésre nem volt szabva időkorlát, azonban a gyakorlatban tanórákra vitték be a kollégák a felmérés anyagát, így a tanóra vége jelenthetett időkorlátot a kitöltő hallgatóknak.

Szakmailag nagyon fontos megvizsgálni a megjelölt szavak szövegen belüli elhelyezkedését!

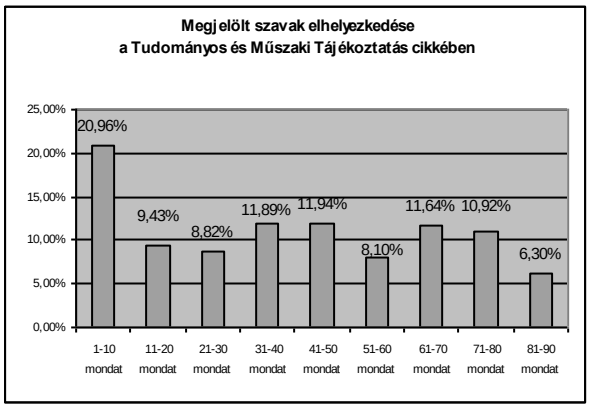
Már a feldolgozás során érezhető volt, hogy előszeretettel jelölik meg a szöveg elején lévő szavakat, majd ez a lelkesedés a szöveg vége felé csökken. Nézzük meg számadatok formájában, hogyan alakul a szavak megjelölésének elhelyezkedése a mondatokat alapul véve!



A Könyvtári Figyelőből származó cikk esetén a szavak 20%-a az első 12 mondatból származik, és a 20. mondat végére már megjelölték a szavak 34%-át.

Ez a cikk 86 mondatot tartalmazott, tehát a szavak 20%-a a mondatok első 14%-ban található, a mondatok első harmadában pedig megtalálható a szavak 46%-a. Ezt követően a szómegjelölés egyenletesebbé válik. Összegezve elmondható, hogy ezen cikknél a lényegesnek tartott szavak majdnem 40%-a a szöveg első negyedéből kerül megjelölésre.

Nézzük meg, hogyan alakulnak ezek az értékek a TMT-ből származó cikk esetén!



Ezen cikk esetén az első három mondatból kerül megjelölésre az összes szó 10%, valamint a megjelölt hasznosnak tartott szavak több, mint 20%-a az első 10 (pontosan az első 9) mondatból került kiválasztásra. A cikk összesen 89 mondatából ez a mondatok 10%-át jelenti. Ezt követően egyenletesnek tekinthető a szókiválasztás elhelyezkedése.

A fent tapasztalt értékek nem meglepőek. A tartalomelemző eljárások közül több is részletesen kitér az első bekezdés fontosságára, mivel itt a szerző bevezeti mondanivalóját, amit rendszerint olyan formában tesz meg, hogy ad egy összefoglaló gondolatsort a cikk tartalmáról, sokszor felsorolva a benne található leglényegesebb témaköröket.

Az első bekezdéseket súlyozó elméletek ugyanilyen fontosnak tartják az utolsó bekezdést is, mivel ott a szerző összegzi a cikkben foglaltak, felsorolja, majd lezárja az elért eredményeket. Nézzük meg a felmérés alapjául szolgáló személyek is fontosnak tartották-e az utolsó mondatokat!

A diagramokról leolvasható, hogy mind a két cikk esetén a legalacsonyabb számú hasznos szó kiválasztás a szöveg utolsó mondataiban történik. Ennek magyarázatát a két cikk alapos vizsgálata után sem könnyű megadni, véleményem szerint ugyanis mind a két cikk szerzője tartalmas gondolatokkal zárja cikkét. Azonban az utolsó mondatokban kevés új kifejezés található, inkább a cikkben már előforduló szavak kerülnek felhasználásra az összegzés során. Ez talán egy magyarázat lehet az alacsonyabb szó megjelölésre.

A felmérés alapjául szolgáló cikkek szakmai cikkek. Fontos kérdésnek tartom, hogy a referátum-készítésben a szaktudás játszik-e fontosabb szerepet, vagy a szövegek tömörítésében, lényegkiemelésben való jártasság. Ennek a kérdésnek az elemzésére néhány magyar szakos hallgatót is felkértem a referátum elkészítésére. Így a felmérésbe nem csak könyvtár-informatika szakos hallgatókat vontam be, hanem az Eszterházy Károly Főiskola magyar szakos hallgatóival is készítettem kivonatot.

A hallgatók száma az egyetemekről származó hallgatói létszám 44%-a, azaz 27 fő vett részt a felmérésben (22%-uk egyetemi képzésen vesz részt). Az általuk készített hasznos szavak megjelölése során azonban csak egy hallgató volt, aki egyetlen cikket sem jelölt meg, így ki kellett zárni ezen elemzésből, illetve a két cikk esetén is sokkal kevesebben éltek azzal a lehetőséggel, hogy nem teljesítik a kért feladatot és nem jelölnek meg hasznos szavakat. Ennek okát a nagyobb rutinban látom, véleményem szerint a magyar szakos hallgatók jobban hozzá vannak szokva a hasonló jellegű feladatokhoz, így nem idegenkednek tőle. Az első cikk esetén az összes hallgató jelölt meg hasznos szavakat, míg a második cikk esetén 3 tanuló hagyta ki ezt a feladatot, de ennek oka az időhiány volt. Mivel a legtöbb felmérést személyesen folytattam le, így volt lehetőségem annak a tapasztalatnak a levonására, hogy a magyar szakosok sokkal lassabban (alaposabban?) olvassák el a szöveget, ezért több személynek is kevés volt a rendelkezésre álló 60 perc (mely nem kötelező korlát volt, csak a tanóra végéig hátra lévő idő), míg a könyvtár-informatika szakosok átlagosan 40 perc alatt teljesítették a kért feladatot.

A magyar szakos hallgatók a Könyvtári Figyelő cikke esetén átlagosan a szavak 5,17%-át jelölték meg hasznos szónak. Míg a könyvtár-informatika szakos hallgatók esetén a két cikknél az átlagosan megjelölt szavak száma szinte azonos érték volt

(~3,2%), addig a magyar szakosoknál jelentős különbséget tapasztalhatunk. A második cikk esetén átlagosan csupán az összes szó 2,02%-át jelölték meg hasznos szónak.

A valósághoz közelebb álló képet kapunk abban az esetben, ha az első cikknél elhagyjuk legmagasabb értéket (mely kiugróan magas: átlagosan kicsivel 50 fölötti szószámot jelöltek meg a hallgatók, de egy személy 384 szót választott ki hasznosnak). Ezen szélsőértéket figyelmen kívül hagyva átlagosan a szavak 4,18%-a került kiválasztásra. Azonban még ez is dupla annyi megjelölést jelent, mint a második cikk esetében. Ennek magyarázatát talán a cikkek témájában találhatjuk meg. *KOLTAY Tibor: Szöveg, információ, relevancia: néhány adalék* a témakörhöz című cikke nem áll annyira távol a magyar szakos hallgatók szakterületétől. Ezen cikk esetén találkozunk a jóval magasabb számú szó kiválasztással. A másik cikk *PROKNÉ Palik Mária: A tartalmi feltárás problémái online könyvtári katalógusokban* című munkája már szinte teljes mértékben a könyvtáros szakma szakterületéhez tartozik. Itt a magyar szakos hallgatók nagyon alacsony számú hasznos szót választottak ki. Talán egy magyarázat lehet a távolabb álló témakör. Több információ birtokába jutunk, ha megvizsgáljuk, hogy hogyan alakultak a konkrétan kiválasztott szavak a két minta esetén. Előtte azonban vizsgáljuk meg, hogy a hallgatók által megjelölt szavak száma hogyan oszlik meg az alábbi kategóriák között a két cikk esetén!

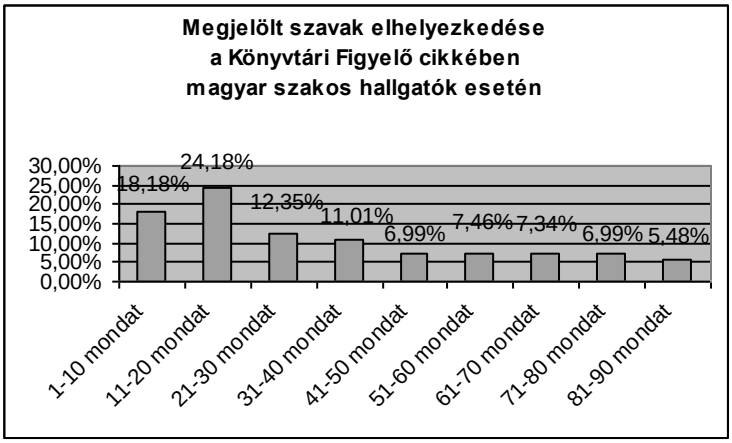
A gyakorisági elemzés során az első cikknél volt néhány kiugróan magas elem, de a legtöbben (a kitöltők majdnem 80%-a) 100-nál kevesebb szót jelölt meg. Ezért az alábbi kategóriákkal létrehozott gyakoriság táblázatot elemezzük!

Csoport határok	KF		TMT	
	Abszolút gyakoriság	Relatív gyakoriság	Abszolút gyakoriság	Abszolút gyakoriság
Nincs megjelölt szó	0	0,00%	3	11,54%
1-10 db	5	19,23%	7	26,92%
11-20 db	4	15,38%	7	26,92%
21-30 db	3	11,54%	3	11,54%
31-40 db	3	11,54%	2	7,69%
41-50 db	2	7,69%	1	3,85%
51-60 db	0	0,00%	1	3,85%
61-70 db	3	11,54%	0	0,00%
71-80 db	0	0,00%	0	0,00%
81-90 db	0	0,00%	0	0,00%
91-100 db	0	0,00%	0	0,00%
Több mint 100 db	6	23,08%	2	7,69%
Összes:	26	100,00%	26	100,00%

A Könyvtári Figyelő cikke esetén a legtöbben 11–50 szót jelöltek meg (a kitöltők több mint 65%-a), míg a TMT-ből származó cikk esetén a 11 és 40 szó közötti

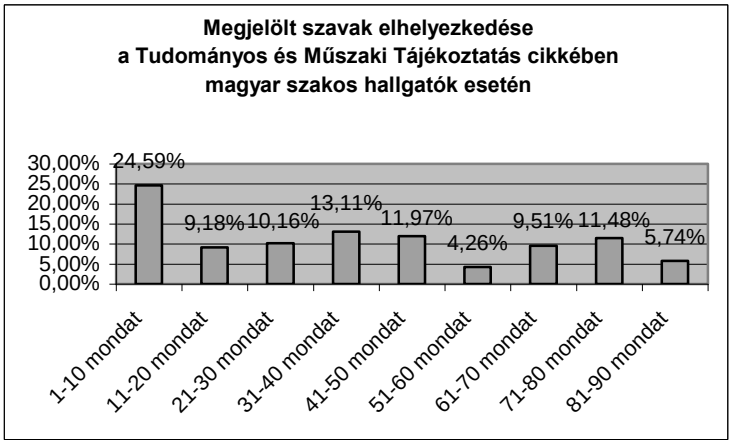
intervallumba tartozik a kitöltők 76%-a. A magyar szakosok átlagos releváns szómegjelölése eltér a könyvtár-informatika szakos hallgatók eredményétől, de mindkét csoportnál a legjellemzőbb kategóriák a második cikk esetén egy intervallummal lentebb találhatóak, mint az első cikk esetén, azaz 10 szóval alacsonyabb a megjelölt szavak száma.

Ha megvizsgáljuk a megjelölt szavak elhelyezkedését a szöveg mondatain belül, a könyvtár-informatika szakos hallgatók választásához nagyon hasonló képet kapunk.



Már a diagramon is látszik, hogy a magyar szakos hallgatók is az első 20 mondatból választották ki a legtöbb hasznos szót, mégpedig az összesnek több mint 42%-át itt jelölték meg, a szavak fele pedig az első 25 mondatból kerül kiválasztása. A hasonlóság nemcsak grafikusan látszik. Ha megnézzük az informatikus-könyvtáros szakos hallgatók abszolút (vagy relatív) gyakorisági adatait, mely megmutatja, hogy mondatonként hány hasznos szó került megjelölésre, és ezt összevetjük a magyar szakos hallgatók gyakorisági adataival, akkor az abszolút gyakoriság vizsgálata esetén erős pozitív korrelációs kapcsolatot kapunk (0,7767), míg a kommulált gyakorisági adatok esetén még szorosabb összefüggést mutat a korrelációs együttható a 0,9878-as értékével.

A TMT cikkének vizsgálatakor teljesen hasonló kép fogad minket.



Szinte 1–2%-os eltéréseket tapasztalunk a szavak mondatokon belüli elhelyezkedésének vizsgálatakor. A megjelölt szavak 20%-a az első 8 mondatban található. Azaz a magyar szakos hallgatók is nagyon lényegesnek tartották az első mondatok kifejezéseit. A kezdeti magas számú releváns szó megjelölés átmegy egy egyenletes eloszlásba, melyet az is mutat, hogy a szavak felét az első 34 mondat tartalmazza (mely az összes 89 mondat majdnem 40%-a), míg a Könyvtári Figyelő cikke esetén a legelső mondatok szavai ugyan nem kapnak akkora szerepet, de a releváns szavak felét az első 25 mondat tartalmazza (mely a cikk mondatainak 29%). Ez az összefüggés azonban mindkét alapul szolgáló minta esetén fenn áll, melyet a korrelációs értékek is alátámasztanak:

Az informatikus-könyvtáros és a magyar szakos hallgatók releváns szavainak mondatonkénti eloszlásához tartozó abszolút (és relatív) gyakorisági értékeinek korrelációs együtthatója: 0,6991, mely pozitív korrelációt mutat, a halmozott, azaz kommulált gyakorisághoz tartozó korrelációs érték pedig 0,9975, mely nagyon erős pozitív kapcsolatra utal.

Már a pozitív korrelációból is adódik, de a diagramokat megtekintve is láthatjuk, hogy mindkét cikk esetén az utolsó bekezdés szavai a magyar szakosoknál sem kaptak nagyon fontosságot, mint a könyvtárosoknál.

Összegezve elmondható, hogy a releváns szavak kiválasztása során nem mutatható ki különbség abban, hogy a referátumot szakemberek (informatikus-könyvtáros hallgatók), vagy a témához nem annyira értő, de kivonatkészítésben nagy rutinnal rendelkező személyek (magyarszakos hallgatók) készítik. Bár a két csoport esetén a releváns szavak számának megválasztása teljesen eltérő, de azok eloszlása, szövegen belüli súlyozása, azaz tartalmi hatása megegyezik a két minta esetén.

Végezetül nézzük meg a számítógépes kivonatkészítő program eredményeinek összevetését a hallgatói minta által kapott adatokkal!

Az automatikus kivonatkészítés első lépései közé tartozik a szavak szótövének meghatározása, majd az előfordulásaik összesítése. Ennek eredményeként előáll egy szógyakorisági lista. Az elemzésünk alapjául szolgáló szavak már nem tartalmazzák a tiltott szólista tagjait, így pl. a leggyakrabban előforduló névelők, kötőszavak és

hasonló szavakkal nem találkozunk az alábbi listába. Az összesítés után a leggyakrabban előforduló 10 szót szeretném bemutatni, illetve megvizsgálni, hogy az informatikus-könyvtáros-, illetve magyar szakos hallgatók relevánsnak tartott szavai között hányadik helyen szerepel.

Szavak	Gépi elemzés alapján	Egyetemi hallgatók	Magyar szakos hallgatók
szöveg	1	1	1
relevancia	2	13	7
jel	3	6	3
tartalom	4	11	11
információ	5	8	2
jelentés	5	11	8
nyelv	5	2	4
adott	6	-	-
kapcsolat	6	9	11
könyvtártudomány	6	5	5
jelölő	7	15	14
szó	8	>30	>30
kognitív	8	12	10
elv	9	4	12
feladó	9	>30	21
informatív	9	20	12
címzett	9	22	15
két	9	18	22
objektum	9	12	19
paradigma	9	>30	18

A leggyakrabban előforduló nem tiltott szó megegyezik mind a két minta esetén a leggyakrabban megjelölt helyre kerül kifejezéssel. A „szöveg” szót jelölték meg a legtöbben. A hasonló rangsorolás további kifejezések esetén is fenn áll. Az első 10 helyre kerül szavak 35%-a található meg az első 10 helyen az informatikus könyvtárosok esetén és 40%-a kapott szintén a legtöbb jelölést a magyar szakos hallgatók esetén.

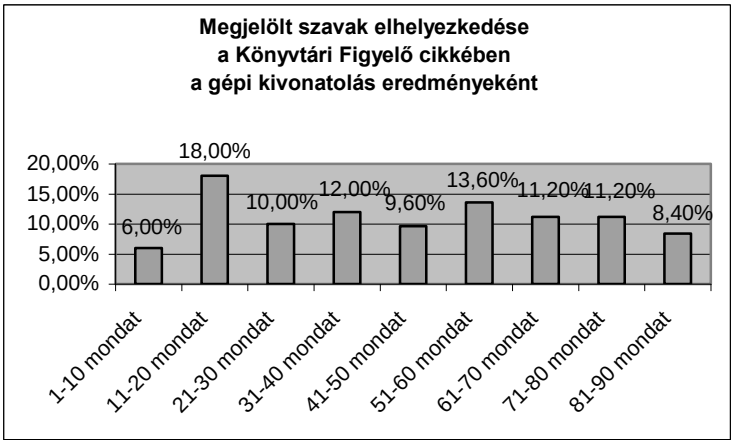
A leggyakrabban előforduló első 20 szónak 70%-a szerepel a legtöbb jelölést kapott első 20 szó között az informatikus könyvtáros hallgatók esetén, a magyar szakos hallgatók pedig ezen szavak 80%-a szerepel a legtöbb jelölést kapott első 20 szó között.

A Tudományos és Műszaki Tájékoztatás cikkének szavait megvizsgálva a számítógépes összesítés alapján, a következő kifejezések szerepeltek a szövegben a leggyakrabban (elhagyva a tiltott szavakat):

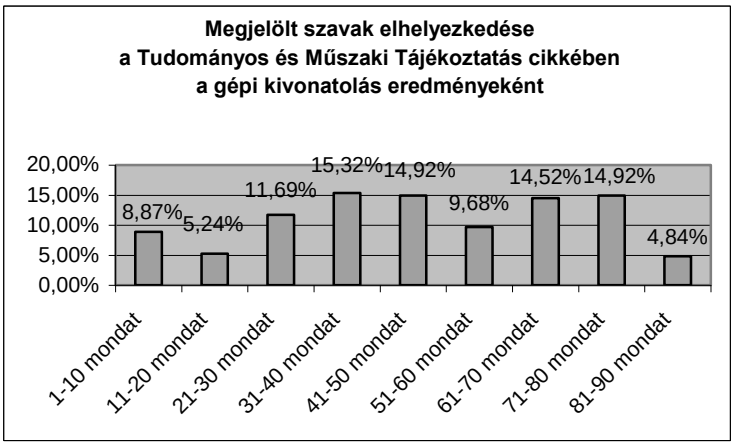
	Gépi elemzés alapján	Egyetemi hallgatók	Magyar szakos hallgatók
jelzet	1	1	6
adatbázis	2	4	2
új	3	11	15
régi	4	>30	14
utaló	5	6	15
különböző	6	>30	16
könyvtár	7	9	16
ETO	8	14	2
egymás	9	-	-
ETO-jelzet	9	2	3
könyvtári	9	14	7
osztályozás	9	5	13
retrospektív	9	12	4
táblázat	9	16	17
feldolgozás	10	11	10
információkereső	10	17	9
katalógus	10	12	16
nyelv	10	9	12
online	10	21	10
számítógépes	10	16	14

Ezen cikknél még az előző esetben tapasztalt hasonlóságnál is nagyobb egyezéssel találkozhatunk. A program által első 10 helyre került gyakorisággal rendelkező 14 szónak a fele szintén az első 10 helyen található mind a két minta esetén. (Az eltérés csak annyi, hogy nem ugyanazon kifejezések). Tovább vizsgálva a hallgatók által legtöbb jelölést kapott szavakat, a gépi leggyakoribb 14 szó közül 13 mindkét minta esetén az első 15 helyezett között szerepel. Ez több mint 90%-os egyezést jelent.

Ezen egyezés feltárása után némi ellentmondásba ütközünk, ha megvizsgáljuk a gépi szógyakoriság leggyakoribb kifejezéseinek mondatokbeli elhelyezkedését. Míg – az első cikknél – a leggyakrabban előforduló nem tiltott szavak 70%, ill. 80%-a található a két minta esetén a leggyakrabban megjelölt szavak között, addig megvizsgálva a szavak mondatokon belüli elhelyezkedését látható, hogy az első 10 mondat a leggyakoribb szavak súlyozása után sem kap előkelő szerepet. Sokkal egyenletesebb a szavak elhelyezkedése a szövegen belül, mint ahogy azt a hallgatók megjelölték.



A másik cikk elemzése során is hasonló a helyzet. Hiába teljesül, hogy a leggyakoribb 14 szó közül 13 a mintáknál is a leggyakrabban megjelölt szavak közé tartozik, a mondatokon belüli elhelyezkedése gyökeresen eltér a hallgatók által megjelölt helyektől. Az első mondatok nemhogy nem kapnak kiugró szerepet, szinte a legalacsonyabb mértékben tartalmazzák ezen szavakat.



A fenti eredményekből két következtetés vonható le:

1. A felmérés alapjául szolgáló minta releváns szókiválasztását nagymértékben befolyásolja a szavak szövegen belüli gyakorisága. A szerző által gyakrabban használt kifejezések kiváltják a megjelölés kényszerét a kivonatot készítő személyekben.
2. A mintába tartozó személyek eredményének elemzése alapján levonható az a következtetés, hogy a lényegesnek tartott szavak elhelyezkedése során a személyek nem következetesek. Míg a szöveg elején megjelölnek bizonyos

szavakat, ha ugyanazon szó, kifejezés a szöveg közepén, illetve vége felé is előfordul, már nem kerül megjelölésre. Ez az oka annak, hogy ha számítógéppel súlyozzuk,[8] hogy mely mondatok tartalmazzák a leggyakoribb szavakat, akkor a szöveg közepén szereplő mondatok magasabb értékeket kapnak, mint az elején lévők.

Összegezve elmondható, hogy ha az emberi kivonatoláshoz hasonló eredményt adó gépi kivonatot szeretnénk kapni, akkor a szöveg elején lévő mondatok szavait nagyobb súllyal kell figyelembe venni, mint a többi kifejezést.

[8] Pontozva a mondatban előforduló azon szavakat, melyek szerepelnek a gyakorisági listán, majd az eltérő mondathosszúságokat kiküszöbölésére egy átlagos pontszámot adva minden mondatnak.