Eszterházy Károly Főiskola

Kivonatoló program kontra emberi kivonatolás

Lengyelné dr. Molnár Tünde

Eszterházy Károly Főiskola

Eszterházy Károly Főiskola

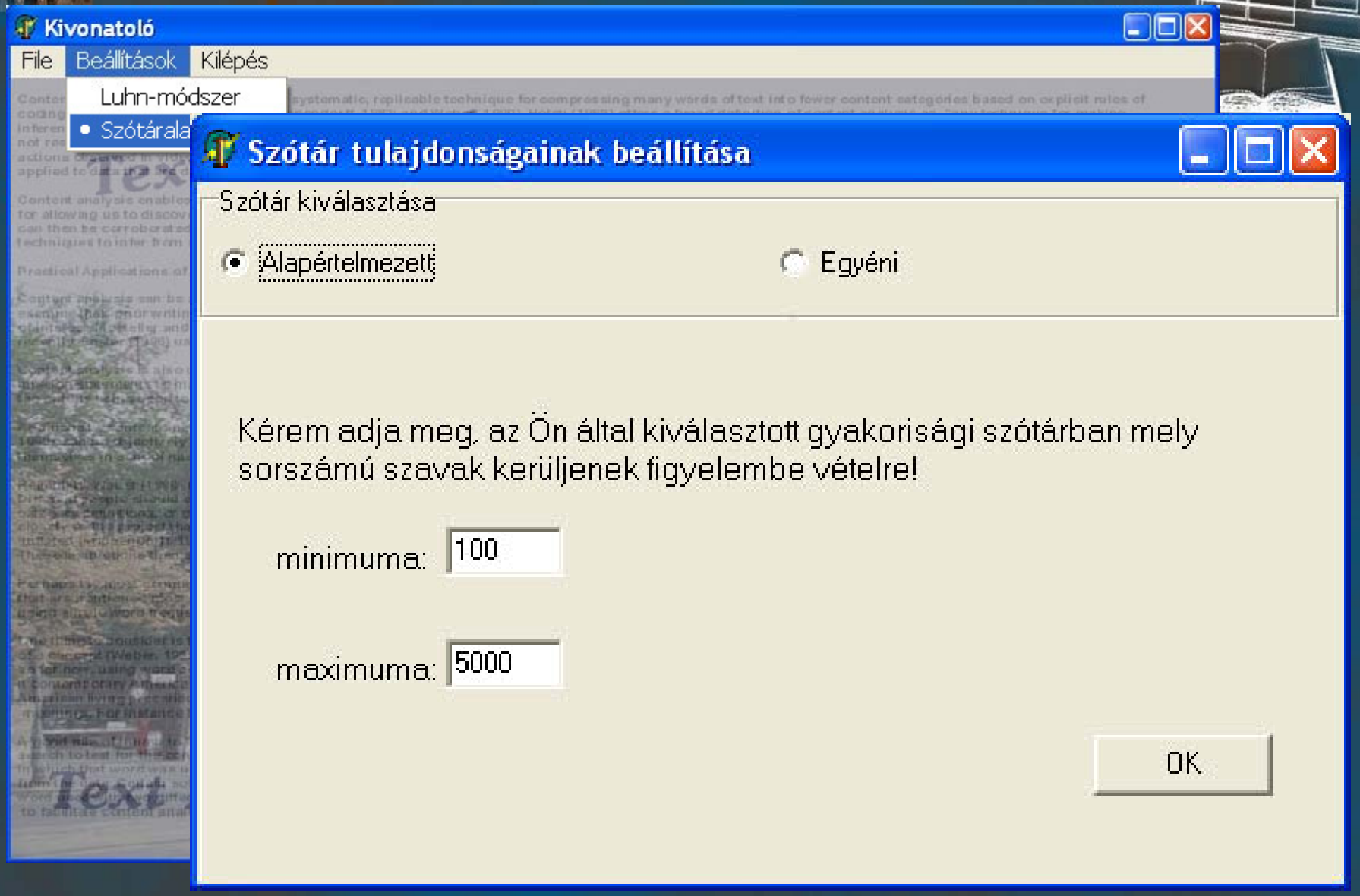
Kivonatoló program

- Magyar nyelvű offline kivonatoló program
  - kvantitatív tartalomelemzés
  - egységeit a szöveg szavai képezik,
  - az output: a szöveg mondatai
  - Jelenlegi állapot: elkészült a program első verziója
  - Tesztelés:
    - emberi kivonatokkal való összevetéssel


Kivonatoló
File Beállítások Kilépés
Referátumkészítés
Text Mining
Kivonatolás

Eszterházy Károly Főiskola

Szótár használat



Eszterházy Károly Főiskola

Output

Kivonatoló

File Beállítások Kilépés

Mentés Nyomtatás

A következő meghatározás is mutatja, hogy nincs éles határ a kivonat és a referátum között.
Egy szöveg szavaiból képzett előfordulási lista négy részre osztható.
A jövőben véleményem szerint egyre több gyakorisági szótárral fogunk találkozni, melyek közül néhánynak az alapját talán a Magyar Nemzeti Szövegtár szolgáltatja.
Berelson a szisztematikus és mennyiségi leírására alkalmas kutatási technikákat tekinti tartalomelemzésnek.
Berelson a tartalomelemzést a következőképpen definiálja: a kommunikáció valóságos tartalmának objektív, szisztematikus és mennyiségi leírására alkalmas kutatási technika.
Akkor használjuk, ha az egységek a vizsgálat szempontjából heterogének.
A nem véletlen mintavételi eljárások alapján levont következtetésekből nem általánosíthatunk.
Zipf-törvénye alapján a szignifikáns szavak listájának meghatározása során figyelembe kell venni az adott terület kifejezéseiből készített gyakorisági szótárakat.
Viszont ezt a kategóriát nehéz figyelembe venni a későbbi elemzés során.
Mint ahogy említettem és az eddig ismertetett szótárak is mutatják a szépirodalomtól eltérő tudományághoz kapcsolódó átfogó gyakorisági szótár egyenlőre nem létezik magyar nyelven.
Nem az explicit tartalom vizsgálata a cél, hanem a szövegkörnyezet kontextusából szeretnénk következtetéseket levonni.
A függelékben található gyakorisági szótár a 10-nél többször előforduló szavakat tartalmazza.
A gyakorisági vizsgálatok nem állnak meg a szöveg szavaiból képzett előfordulási csoportok felállításának szintjén.
Míg a korábbi vizsgálatok jobbára kvantitatív elemzések voltak, a II. világháború idején a propaganda-elemzés vált meghatározóvá.
Az elmúlt években ugrásszerűen megnőtt a gyakorisági szótárak készítése, kiadása, és egyre több szótárt építő vállalkozással találkozhatunk, azonban szakterülethez kapcsolódó átfogó gyakorisági szótár továbbra sem létezik magyar nyelven.
A szöveg szavainak szótöveiből képzett szógyakorisági lista felállítását a szöveget jellemző szignifikáns szavak meghatározása követi.
A szóhasználat gyakorisági értékeinek meghatározásához a 150 millió szót tartalmazó Magyar Nemzeti Szövegtárat vették alapul.
Ha a szöveget manuálisan előkészítjük és a következő módon történik a gépelésük, akkor a problémák nagy része kiküszöbölhető: Rövidített szavak eredeti formában történő begépelése
A dokumentációs nyelvészeti csoport korszakát (1967-1971) az 1966-ban, az MTA Számítóközpontjában, a gépi nyelvészeti munkacsoportból átalakult Dokumentációs Nyelvészeti Csoport munkája jellemzi.
Az eredeti dokumentum ismerete nélkül is érthető.
A mondat meghatározásakor is több szempontot kell figyelembe venni.

Eszterházy Károly Főiskola

Felmérés

Eszterházy Károly Főiskola

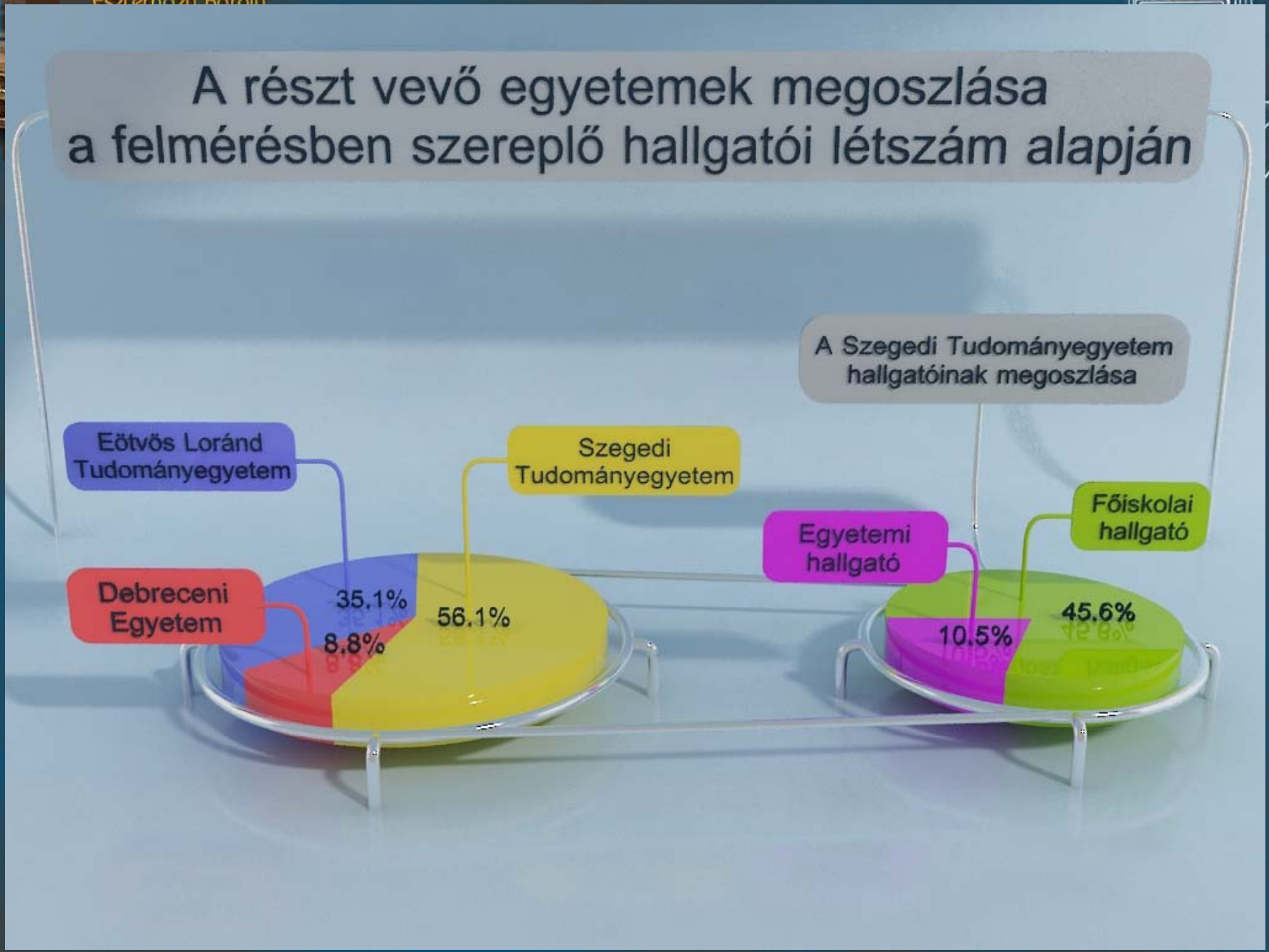
A felmérés alapja

- KOLTAY Tibor:
  Szöveg, információ, relevancia: néhány adalék a témakörhöz
- PROKNÉ Palik Mária:
  A tartalmi feltárás problémái online könyvtári katalógusokban

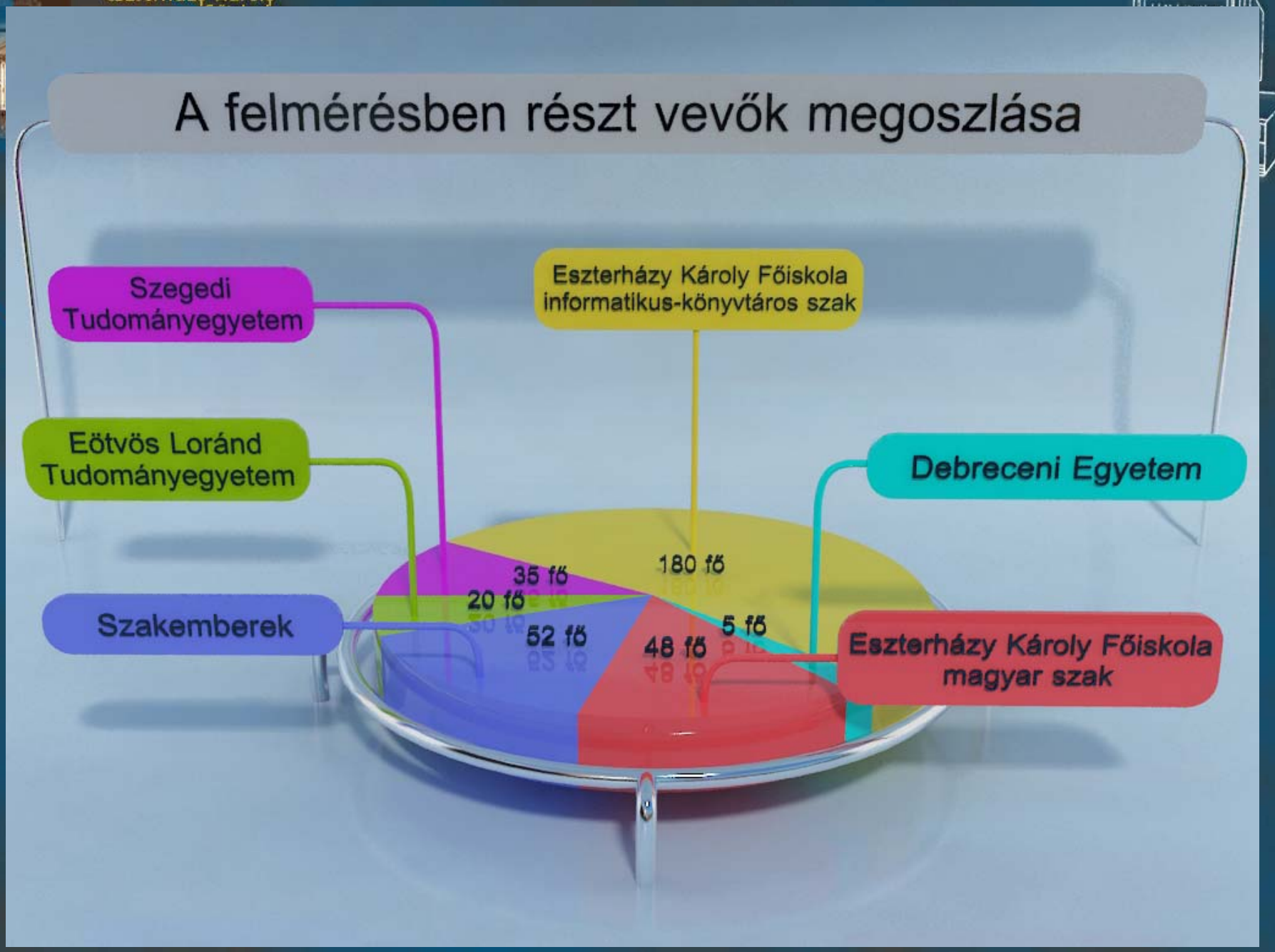
- Kitöltés módja:
  - online
  - papír alapú

- Feladat:
  -20%-s tömörítés → 17-17 releváns mondat kiválasztása és rangsorolása
  -Kulcsszó megjelölés (korlát nélkül)

A részt vevő egyetemek megoszlása a felmérésben szereplő hallgatói létszám alapján



A felmérésben részt vevők megoszlása



Súlyozás

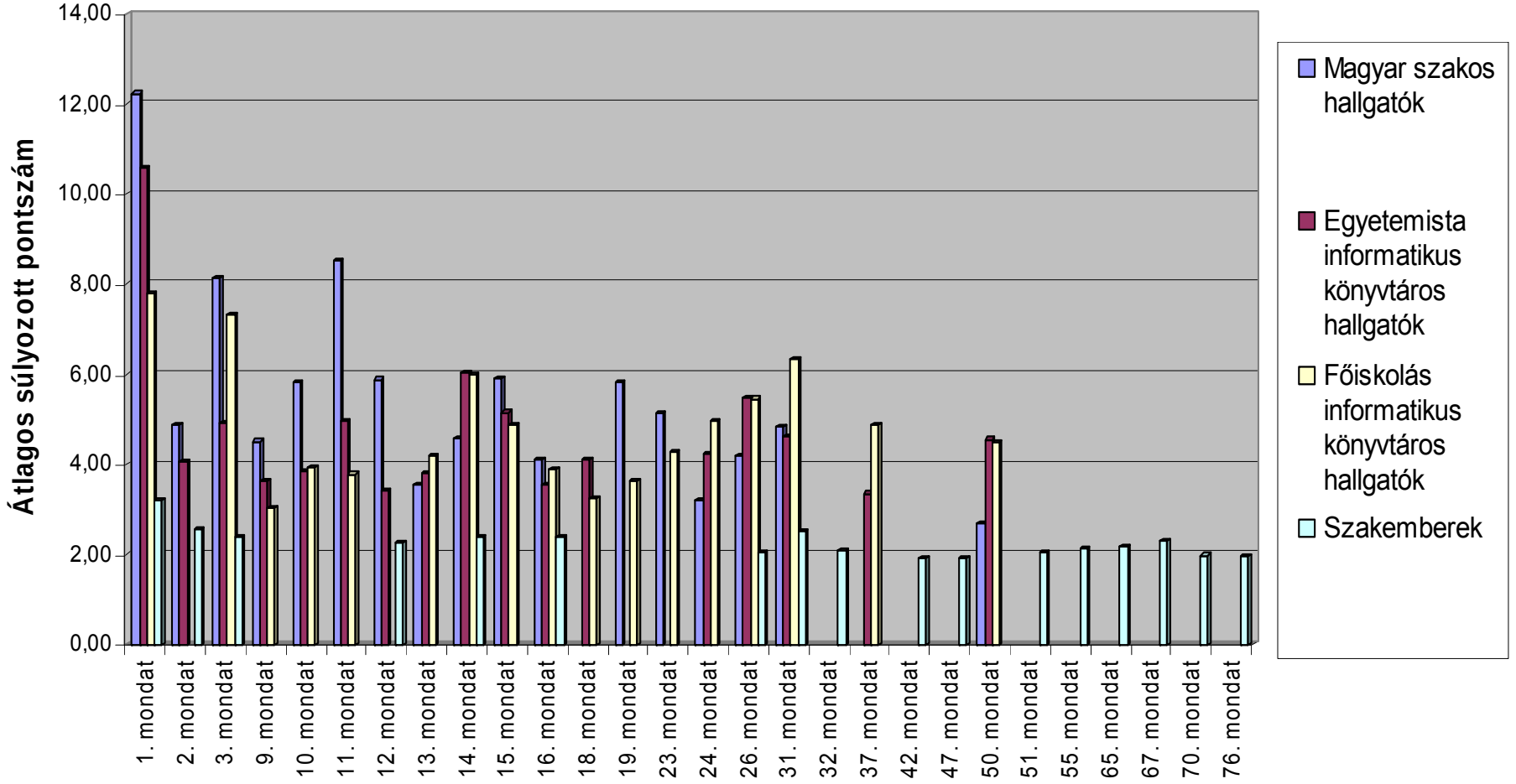
- **Borda-módszer alapján**
  - A rangsorbeli első helyezetteket **n** súllyal vettem figyelembe,
  - a másodikat **n-1**-s értékkel,
  - ….
  - az utolsó helyezett **1**-s súllyal fog szerepelni az értékelésben.

Eszterházy Károly Főiskola

KOLTAY Tibor: *Szöveg, információ, relevancia: néhány adalék a témakörhöz*

Sorrend	A főiskolások legrelevánsabb mondatai	Az egyetemisták legrelevánsabb mondatai
1.	1. mondat	1. mondat
2.	3. mondat	14. mondat
3.	31. mondat	26. mondat
4.	14. mondat	15. mondat
5.	26. mondat	11. mondat
6.	24. mondat	3. mondat
7.	37. mondat	31. mondat
8.	15. mondat	50. mondat
9.	50. mondat	24. mondat
10.	23. mondat	18. mondat
11.	13. mondat	2. mondat
12.	10. mondat	10. mondat
13.	16. mondat	13. mondat
14.	11. mondat	9. mondat
15.	19. mondat	16. mondat
16.	18. mondat	12. mondat
17.	9. mondat	37. mondat

A kivonat mondatai a négy mintacsoportnál



Eszterházy Károly Főiskola

Súlyozott pontszámokból képzett korrelációs mátrix

	Informatikus–könyvtáros egyetemista hallgatók	Informatikus–könyvtáros főiskolás hallgatók	Szakemberek	Magyar szakos hallgatók
Informatikus könyvtáros egyetemista hallgatók	1			
Informatikus könyvtáros főiskolás hallgatók	0,917	1		
Szakemberek	0,486	0,445	1	
Magyar szakos hallgatók	0,872	0,856	0,441	1

Eszterházy Károly Főiskola

Kivonatok mondatainak eloszlása

A szöveg	A kivonat eloszlása Koltay Tibor cikke esetén				
	Egyetemista informatikus könyvtáros hallgatók	Főiskolás informatikus könyvtáros hallgatók	Szak-emberek	Magyar szakos hallgatók	Szerző
I. negyede	70,59%	64,71%	35,29%	70,59%	46,67%
II. negyede	23,53%	29,41%	23,53%	23,53%	40,00%
III. negyede	5,88%	5,88%	23,53%	5,88%	6,67%
IV. negyede	-	-	17,65%	-	6,67%

Eszterházy Károly Főiskola

Egyezés a szerző kivonatával

- A minta 340 tagja közül 7 fő jelölte meg ugyanazon mondatokat mint a szerző

- A mintacsoportok súlyozott kivonatait alapul véve az egyezés:
  - Egyetemista informatikus könyvtáros hallgatók esetén **8 mondat**
  - Főiskolás informatikus könyvtáros hallgatók esetén **8 mondat**
  - Szakemberek esetén **5 mondat**
  - Magyar szakos hallgatók esetén **7 mondat**

Eszterházy Károly Főiskola

Számítógépes output

- A két módszer esetén 10 mondat egyezik meg.
- A mintacsoportokkal az egyezés:

Mintacsoportok	Luhn módszere alapján		Szószablya szótár alapján	
	Egyező monda-tok száma	Egyezés aránya	Egyező monda-tok száma	Egyezés aránya
Informatikus könyvtáros egyetemista hallgatók	3	17,65%	3	17,65%
Informatikus–könyvtáros főiskolás hallgatók	4	23,53%	3	17,65%
Szakemberek	4	23,53%	6	35,29%
Magyar szakos hallgatók	3	17,65%	3	17,65%
Szerző kivonata	5	29,41%	3	17,65%

Eszterházy Károly Főiskola

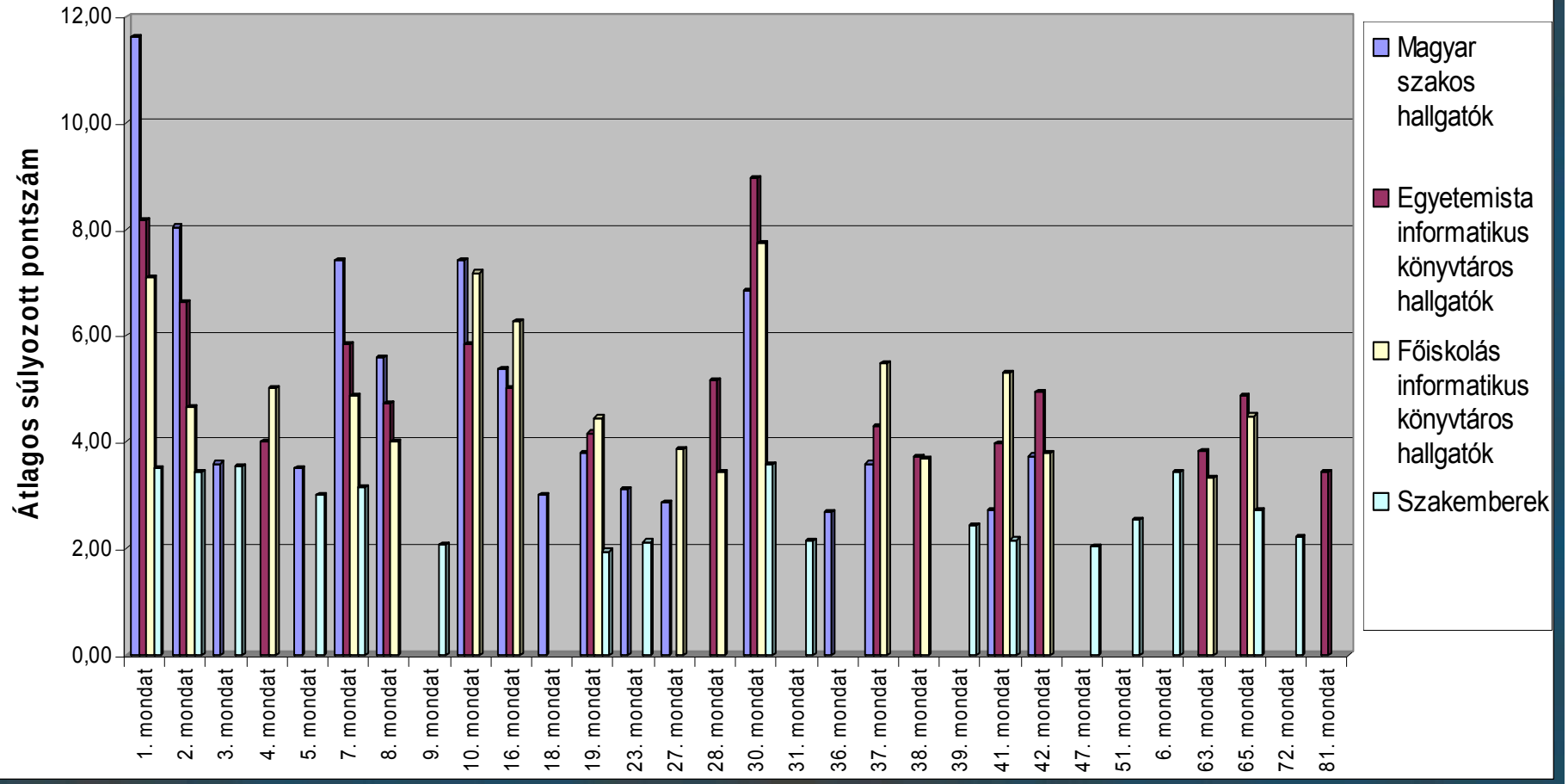
PROKNÉ Palik Mária: *A tartalmi feltárás problémái online könyvtári katalógusokban*

Eszterházy Károly Főiskola

Sajátosságok

- Az egyetemista és főiskolás informatikus-könyvtáros hallgatók kivonata 94%-ban megegyezik
- Az első bekezdés 4-4 illetve 3-3 mondata megtalálható a mintacsoportok kivonatában
- A szakembereknél eltérőek a vélemények:
  - 52 szakember 37 különböző mondatot tett első helyre
  - egyetlen mondat kapott 25% feletti jelölést (a hallgatóknál 80, illetve 70%-70% fölötti)

A kivonat mondatai a négy mintacsoportnál



Eszterházy Károly Főiskola

Korrelációs mátrix a megjelölt mondatok alapján

	Informatikus könyvtáros egyetemista hallgatók	Informatikus könyvtáros főiskolás hallgatók	Szakemberek	Magyar szakos hallgatók
Informatikus könyvtáros egyetemista hallgatók	1			
Informatikus könyvtáros főiskolás hallgatók	0,936	1		
Szakemberek	0,526	0,471	1	
Magyar szakos hallgatók	0,878	0,849	0,576	1

Eszterházy Károly Főiskola

Kivonatok mondatainak eloszlása

A szöveg	A kivonat eloszlása Prokné Palik Mária cikke esetén				
	Egyetemista informatikus könyvtáros hallgatók	Főiskolás informatikus könyvtáros hallgatók	Szak-emberek	Magyar szakos hallgatók	Szerző
I. negyede	47,06%	47,06%	47,06%	58,82%	5,88%
II. negyede	23,53%	29,41%	23,53%	29,41%	29,41%
III. negyede	11,76%	11,76%	17,65%	11,76%	23,53%
IV. negyede	17,65%	11,76%	11,76%	0,00%	41,18%

A szerző többször nem 1-1 mondatot jelöl meg, hanem 1-1 bekezdésnyi részt.

Eszterházy Károly Főiskola

Egyezés a szerző kivonatával

- Összesen **7** olyan fő van a 340 mintatag közül, akiknek a kivonata 9 mondatban megegyezik a szerző kivonatának mondataival (sorrend elhagyásával)

- A mintacsoportok súlyozott kivonatait alapul véve az egyezés:
  - Egyetemista informatikus könyvtáros hallgatók kivonatával **6 mondat**
  - Főiskolás informatikus könyvtáros hallgatók kivonatával **6 mondat**
  - Szakemberek kivonatával **2 mondat**
  - Magyar szakos hallgatók kivonatával **4 mondat**

Eszterházy Károly Főiskola

Számítógépes output

- A két módszer esetén 11 mondat egyezik meg.
- A mintacsoportokkal az egyezés:

Mintacsoportok	Luhn módszere alapján		A Szószablya szótár alapján	
	Egyező mondatok száma	Egyezés aránya	Egyező mondatok száma	Egyezés aránya
Informatikus–könyvtáros egyetemista hallgatók	6	35,29%	2	11,76%
Informatikus–könyvtáros főiskolás hallgatók	6	35,29%	3	17,75%
Szakemberek	8	47,06%	4	23,53%
Magyar szakos hallgatók	8	47,06%	5	29,41%
A szerző kivonata	4	23,53%	2	11,76%

Eszterházy Károly Főiskola

Végkövetkeztetés

- Nincs két ember, aki ugyanazt a referátumot állítaná elő.
  - Bár a 340 felmérésben részt vevő személy közül 7 illetve 6 személy van, akinek mondatai megegyeznek, de eltérő sorrendben
  - kijelenthető, hogy nagyon kevés személy ítéli meg egyformán a cikkek lényeges mondatait.
- Nincs nagy betűs KIVONAT

Kivonatoló program kontra emberi kivonatolás: DÖNTETLEN

Eszterházy Károly Főiskola

Köszönöm a figyelmet!

mtunde@ektf.hu