# „tök jó, de nincsenek benne csúnya mondatok”[40]

*egy WAP-alapú szótári rendszer üzemeltetésének tapasztalatai*

Földes András

MorphoLogic Kft; Orbánhegyi út 5. 1126 Budapest
lafoldes@morphologic.hu

**Kivonat:** A MorphoLogic 2003 nyara óta működteti a MoBiWAP szótári rendszert. A MoBiWAP nem más, mint a jól ismert MoBiDic szótári rendszernek a WAP lehetőségei szerint átalakított változata. A program a kezdetektől fogva naplózza a kéréseket. A cikk következtetései ennek a naplófájlnak a vizsgálatán alapulnak. A keresett szavak meglepően nagy hányada obszcén. Az adatok elemzése a jelenség okainak kutatásához próbál támpontokat nyújtani. A szokványos statisztikákon túl, az egyszerre keresett szavak közti kapcsolatokat egy skálafüggetlen gráfban is ábrázolom. Az így kapott koincidenciagráf a szavak szemantikai kapcsolatainak a feltárásában is segíthet. A kapott adatok hozzájárulnak a felhasználói elégedettség növeléséhez.

## Figyelmeztetés

**Az alábbi cikk természetéből fakadóan nagy számban tartalmaz obszcén, pornográf és más offenzív kifejezéseket. A trágár szavak esetenként olyan sűrűn és nagy mennyiségben fordulnak elő, hogy mindennemű körülírás, helyettesítés vagy „kipontozás” a szöveg érhetőségének a rovására ment volna.**
A cikk jórészt egyedi azonosításra alkalmatlan, összesített statisztikai adatokat mutat be. Abban a néhány esetben, ahol a mondandóm illusztrálásra egyedi szövegeket használok, azokból minden, esetlegesen a személyes azonosításra alkalmas adatot eltávolítottam.

## A MoBiWAP rendszer

A MorphoLogic 2003 nyara óta működteti a MoBiWAP szótári rendszert. A MoBiWAP nem más, mint a jól ismert MoBiDic[41] szótárunknak a WAP lehetőségei szerint átalakított változata.

[40] Egy a sok ezer beérkezett vélemény közül

A felhasználók a lekérdezés nyelvi irányának megadása után (az alapértelmezés a bármi-bármi „nyelvpár”, ekkor a keresés az összes lehetséges nyelven történik) beírják a keresendő szót, vagy többszavas kifejezést és elindítják a tényleges keresést. A keresés eredménye vagy a keresett szóhoz tartozó szócikk, vagy „kifejezésben keresés” esetén minden, a szót tartalmazó kifejezés-szócikk. A többi MoBiDic szótárhoz hasonlóan, a keresés előtti szótövesítésnek köszönhetően ragozott alakban is megadható a keresendő szó.
Túl sok találat esetén az eredmények között lapozni lehet, ha pedig nincs találat, akkor a szótár felkínálja a keresett szó ábécésorrend szerinti környezetét.
A **Vélemény** menüpontban a felhasználók értékelhetik a szótár működését, és javaslatokat tehetnek a továbbfejlesztésre.


1. ábra: A MoBiWAP használatának fontosabb lépései

A rendszerben jelenleg egy kéziszótár méretű, mindkét irányból kereshető angol–magyar és német–magyar középszótár, illetve magyar szinonimaszótár működik. A szócikkek méretét a WAP igényeinek megfelelően csökkentettük.
A MoBiWAP része a Pannon GSM WAP-portáljának[42], ennek köszönhetően eddig közel három és fél millió kérést szolgált ki; véleményből eddig több mint hatezer érkezett.

### A keresések naplózása

Az összes szótári keresés adata egy naplófájlba kerül. A fájl (jelen cikk szempontjából érdekes) mezői a következők:

**1. táblázat:** A naplófájl rekordleírása

| **Mező** | **Lehetséges értékek** |
|---|---|
| Időpont | éééé-hh-nn óó:pp:mp |
| Telefontípus + WAP verzió | szöveg |
| Lekérdezési mód | 0 (keresés); 1 (lapozás); 2 (környezet) |
| Keresendő szó, kifejezés | szöveg |
| Keresési mód | 0 (szó); 1 (kifejezés részeként) |
| Forrásnyelv | 0 (bármi); 1038 (magyar); 2057 (angol), 1031 (német) |
| Célnyelv | 0 (bármi); 1038 (magyar); 2057 (angol), 1031 (német) |
| A találatok száma | szám |

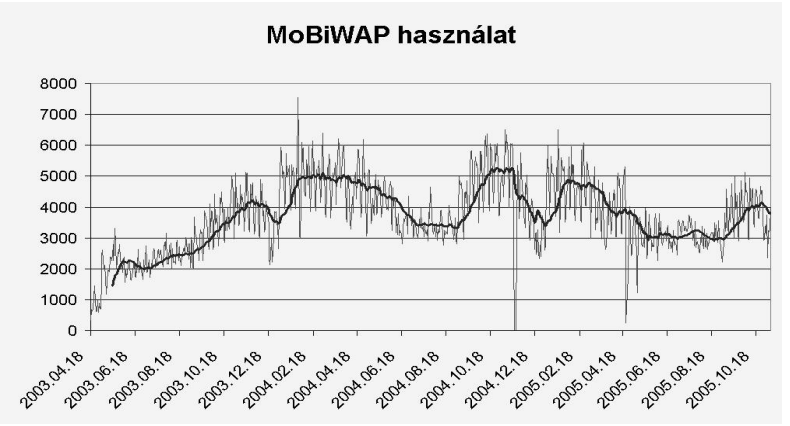
A fenti szerkezetű naplófájlt egy MySQL adatbázisba töltöttem be. A cikk ennek az adatbázisnak az elemzése során nyert adatokra épül.

[41] A MoBiDic rendszer leírása megtalálható a www.morphologic.hu oldalon.
[42] A MoBiWAP a **Hasznos/Szótár** menüpontban érhető el. Más rendszerekből a www.mobidic.hu/scripts/mobiwap.exe. címen lehet hozzáférni.

# Üzemeltetési adatok

## Időbeli eloszlás

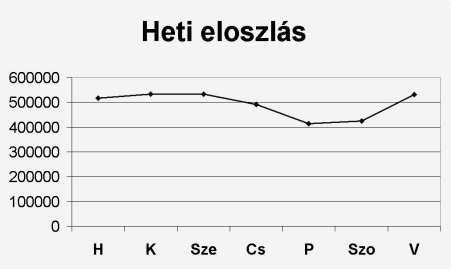
A vizsgált időszakban (2003.04.18. – 2005.11.07.) közel három és félmillió rekord került naplófájlba. Ilyen mennyiségű adat egyedülálló lehetőséget nyújt a szótárhasználat alapos statisztikai elemzéséhez.



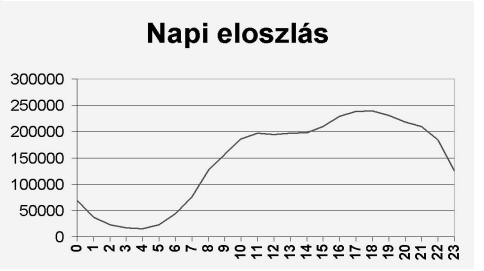

2. ábra

Az ábrán a napi hozzáférések száma látható a teljes időszakban. A kezdeti felfutási időt leszámítva átlagosan napi 3-5 ezer kérés érkezik. A szótárhasználat mérsékeltebb a nyári szünetben, karácsony és újév között**.**

A keresések számának 2005-ös kismértékű általános visszaesése feltehetőleg a Pannon portál átrendezésével magyarázható.
A szótárhasználat heti periodicitása is jól követhető (3. ábra). A forgalom pénteken és szombaton esik vissza, ekkor a potenciális fiatal felhasználók nagy része nyilván nem idegen szavak után kutat a neten.




3. ábra

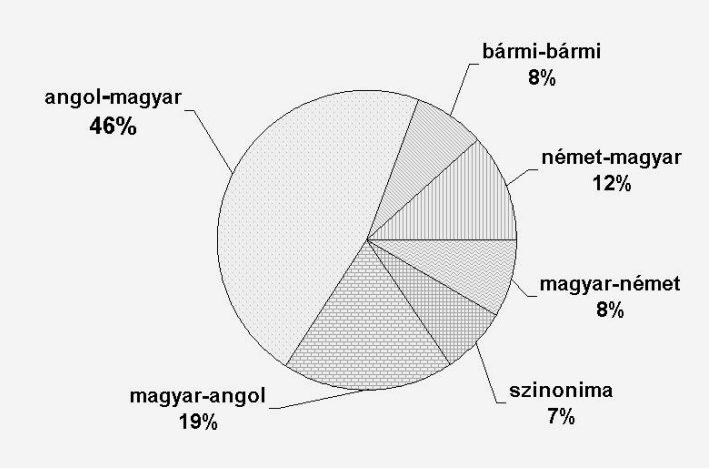



4. ábra

A napi forgalommegoszlás (4. ábra) érdekes módon a felhasználás esti emelkedését mutatja.

**Nyelvek szerinti eloszlás**

A forgalom nyelvenkénti megoszlása (5. ábra) megfelel az idegen nyelvek iránti magyarországi érdeklődésnek.[43] (További nyelveket még kevesebben igényelnének.) A szinonimaszótár 7%-os adata egy hosszú üzemszünet miatt alacsonyabb a valóságos igénynél.

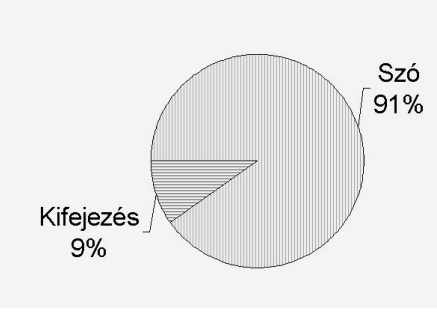



5. ábra: a keresések nyelvi irány szerinti megoszlása

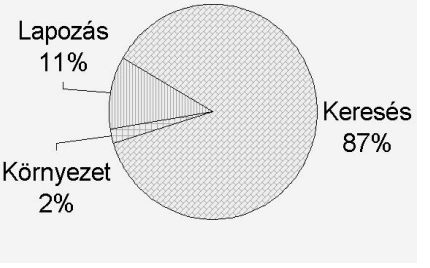
Meglepő, hogy viszonylag kevesen élnek az alapértelmezett nyelvbeállítással (bármi-bármi) ehelyett inkább - egy plusz lépésben - kiválasztják a keresett nyelvi irányt is.

**A felhasználói felület kezelésének adatai**

A felhasználói felület funkcióinak kihasználatlanságát illusztrálja az alábbi két ábra is. A felhasználók túlnyomó része nem él a kifejezésben keresés, illetve a szótár tartalmában való navigálás lehetőségével.




6. ábra




7. ábra

[43] A MorphoLogic és más kiadók szótáreladásai is hasonló eloszlást mutatnak.

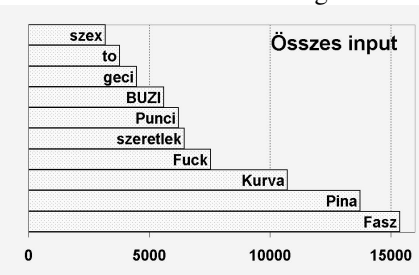
Szolgáltatásunkkal sok olyan embert is elértünk, akinek nem igazán volt még szótár a kezében. Az ő esetükben összemosódik a szótár, a fordítógép, az idegen szavak szótára, a nagylexikon, és a mindent meghallgató beszélgetőtárs fogalma. Mindezt jól illusztrálja a következő néhány „szótári lekérdezés”:

```
Ez a terület,ahol semmi változás nem tapasztalható az elodhöz képest.
Finally out comes a Crow, Comimg quickly to a stop.
Menjen egyenesen elore és a harmadik keresztezodésnél balra.
Bazdmeg de egy köcsög buzi vagy.Te szemét szutyok paraszt csicska.
```

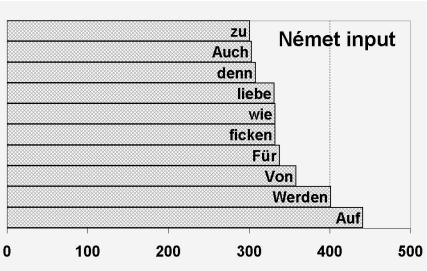
# Gyakorisági adatok

## Gyakoriság általában és nyelvenként

A statisztikai elemzés első kézenfekvő lépése: megkeresni a gyakran kért szavakat. A 8. ábra a teljes adatmennyiség 10 leggyakrabban keresett kifejezését tartalmazza. Megdöbbentő az obszcén szavak elsöprő többsége.
A helyzet a további pozíciókban sem igazán javul, egészen a két-háromszázadik helyig kell elmennünk, hogy „valódi” szótári kereséseket találjunk.
A hihetetlenül sok trágár szó jelenléte döntő részben nyilván nem az idegen nyelvi jelentések iránti érdeklődéssel indokolható.[44] A jelenség mindenképpen pszicholingvisztikai vagy pszichológiai magyarázatra szorul. Ehhez a későbbiekben a további érdekes adatokkal szolgálok.




8. ábra




9. ábra

Ha a szógyakoriságokat forrásnyelv szerinti megbontásban vizsgáljuk, megállapítható hogy az angol nyelv esetében némileg, a német nyelv esetében jelentősen csökken az obszcenitás túlsúlya. Úgy látszik (legalábbis a hazai WAP-felhasználók körében) a magyar és az angol trágárság is része az „általános műveltségnek”, a német nem. (9. ábra)
Elegendő mélységben vizsgálva a szógyakorisági listát a szavak három többé-kevésbé jól elkülöníthető csoportba oszthatók:

[44] Vajon ki tud elképzelni olyan szituációt, amikor valakinek égető szüksége van a “*szutyok paraszt csicska*” angol jelentésére, ezért gyorsan utánanéz WAPon?

**Disznó és más szexuális vonatkozású szavak**: *fasz, pina, kurva, fuck, szeretkezés, szar, segg, bitch, ficken, muschi, szeretlek, csók, hiányzol* és így tovább, oldalakon át.
**Tesztszavak**: Ezek gyakori vagy ritkább szavak, főleg főnevek. A felhasználó feltehetőleg nem igazán szó jelentésére kíváncsi, hanem a szótárat teszteli.[45]: *szép*, *autó, asztal, have, jó, ablak, alma, kutya.*
**Ténylegesen keresett szavak**: ezek között több az ige és a kifejezés, a (magyar) jelentés gyakran nehezen adható meg egy szóval, magyarul a szó több jelentésű. Ide tartozik a kifejezések szavankénti fordításából származó sok prepozíció, segédige is: *issue, cool, imaginative, distress, serendipity* illetve *for, with, have, get, take, could.*

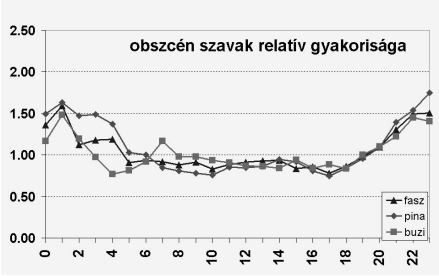
## Gyakoriság időben és egyes szócsoportokra

További érdekes megállapításokat tehetünk, ha a szógyakorisági táblázatokat a teljes anyag különböző szempontok szerint kiválogatott részhalmazaira készítjük el:

**Időbeli eloszlás:**
A táblázat a szógyakoriság-toplistát mutatja különböző időpontokban. Látható, hogy az „élmezőny” gyakorlatilag változatlan:

| idő/helyezés | 0-1 | 6-7 | 12-13 | 18-19 |
|---|---|---|---|---|
| 1 | *fasz* | *fasz* | *fasz* | *fasz* |
| 2 | *pina* | *pina* | *pina* | *pina* |
| 3 | *kurva* | *szeretlek* | *kurva* | *kurva* |
| 4 | *szeretlek* | *kurva* | *fuck* | *fuck* |
| 5 | *fuck* | *fuck* | *punci* | *szeretlek* |

Az alábbi ábrán egyes obszcén szócsoportok használatának relatív gyakorisága látható az idő függvényében (10. ábra).
Látható, hogy az obszcén szavakat napközben az átlagosnál kevésbé, este és késő este viszont jóval gyakrabban keresik.




10. ábra


11. ábra

Néhány beküldött véleményből kiderül[46], hogy sokan szódolgozatok írásakor puskázásra is használják a MoBiWAPot. A jelenség ellenőrzésére megvizsgáltam a közép-

[45] Az én gyakori tesztszavaim: *alma, almafa, mosómedve, apple, raccoon, Tag*

[46] Ilyen vélemények többek közt: `Nagyon zsir ez a xotàr! Tök jòl lehet wele puskàzni!; Kirra a szotar, ezzel puskaztuk a dogankat.`

iskolai tananyagban szereplő néhány szó[47] relatív gyakoriságát (11. ábra). Az ábrán egyértelműen látszik, hogy ezeknek a szavaknak iskolaidőben átlag feletti a gyakorisága. Bár a vizsgált részhalmaz elég kicsi, kis „beleérző képességgel" a görbén talán az ebédszünet is látszik. Ugyanezen az ábrán látható az étlapokon szereplő[48] szavak csoportjának relatív gyakorisága is. A görbe mintha kötődne az étkezési időpontokhoz.

## A találati arány javítása

A sok érdekes és meghökkentő megállapítás mellett a naplófájl elemzésének a legfontosabb haszna a találati hatékonyság és általában a felhasználói elégedettség növelése.

A szótár működtetésének első hónapja után elemeztem a „nincs találat" válaszok lehetséges okait.

Nyolc kategóriát különítettem el:

1. **A nagyszótárakban megvan.** pl.: *középpályás*
2. **Más nyelven a nagyszótárakban megvan.** A felhasználó létező szót írt be, de hibásan adta meg a nyelvi irányt. Pl.: *Spielen*, magyar–német irányban
3. **Gyengén ékezetesítve megvan**. Gyenge ékezetesítésen az í, ó, ő, ú, ű magánhangzók rövid párjukkal való helyettesítését értem. Pl.: *vizköpö*. Nem tartoznak ide a vegyes (néhány ékezet hibás, néhány nem), a fordított (rövid helyett hosszú magánhangzó) és a ragozott (a morfológia nem működik hibás ékezetekkel) alakok.
4. **Erősen ékezetesítve megvan.** Az erős ékezetesítés esetében az összes ékezetes magánhangzót az ékezet nélküli megfelelőjével (a, e, i, o, u) helyettesítettem. Pl.: *gyulolet*
5. **Szavanként megvan**. A beírt többszavas kifejezés kifejezésként nincs meg a szótárban, de az őt alkotó szavak egyenként igen. Pl.: *minél gazdagabb leszel*
6. **Gyengén ékezetesítve szavanként megvan**.
7. **Erősen ékezetesítve szavanként megvan**. A megfelelő esetek kombinációi.
8. **Egyéb**. Minden más eset. Ide tartoznak a fent említett bonyolultabb ékezetesítések (*A testvérem gyereket szűlt*), a helyesírási hibák *(Alles klahr, himveszo*), a kisbetűvel írt német főnevek (*zuschauer*), a nagyon hosszú beírások (*Menj a jó büdös kurva anyádba te kétszínü durva francos nagyseggü majom* és még sokkal hosszabbak is), a nem szöveges input (*1m1m1m1m1*), a tulajdonnevek (*AUCHAN*), más nyelvű kérések *(bune ziua, le roi est mort, vive le roi!*), stb.

[47] *factual, occasion, demand, fiction, suppose, attend, attic, agreement, intention, possible, fame, conclusion, deal, beggar*

[48] *soup, sirloin, tenderloin, beef, pork, mousse*


12. ábra

A hibakategóriák eloszlásának (12. ábra) ismeretében a következő javítások voltak elképzelhetők:

**Technikai jellegű javítások**

Megtörtént az indexek kiegészítése az ékezet nélküli alakokkal, és a kisbetűs német főnevekkel. Tovább javítható a találati arány a helyesírási hibák automatikus javításával. A többszavas inputok esetében szóba jöhet a MetaMorpho fordítószoftver alkalmazása is.

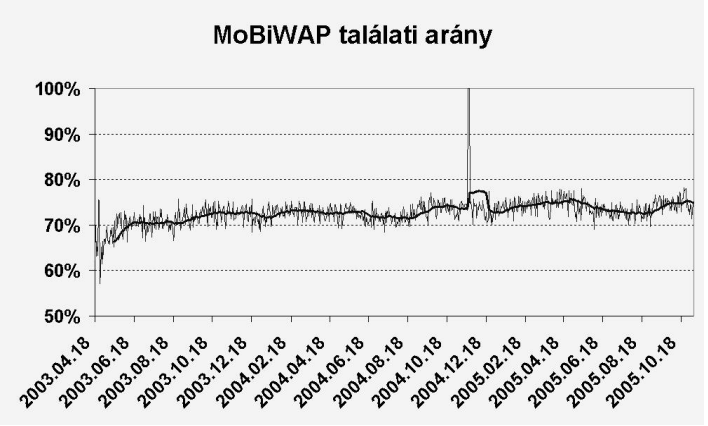
**A lexikont érintő javítások**

Kézenfekvő, hogy ha nagyobb a szótári adatbázis, akkor kevesebb az ismeretlen szó. Ezért megvizsgáltuk mi történne, ha a saját fejlesztésű középszótárunkat (negyvenezer szótári tétel) felcserélnénk az Akadémiai Nagyszótárral. Ez további öt százalékkal növelte sikeres lekérdezések arányát. A naplófájl alaposabb elemzése után azonban itt sem maradt el a meglepetés: nagyjából ugyanekkora javulás volt elérhető a húsz leggyakoribb, a szótárból hiányzó szó felvételével. (Hogy melyik ez a húsz szó? Kérem, lapozzanak vissza a gyakorisági statisztikákhoz...)

**Új adatbázis-modulok hozzáadása**

A „nem talált” listák elemzésével az is megállapítható, hogy nagy szükség lenne, egy, a WAP lehetőségeit kihasználó, magyar nyelvű, „idegenszavakszótára-lexikon-enciklopédia”-szerű adatbázismodulra is. Tudomásom szerint ilyesmi még nincs a piacon, és az adatbázis összeállítása is érdekes feladatnak ígérkezik.
A mérsékelt igény ellenére továbbra is tervezzük újabb nyelvek bevezetését.

A végrehajtott változtatásoknak köszönhetően a találati arány a kezdeti 64 – 66%-ról 73 - 75%-ra nőtt (13. ábra)




13. ábra

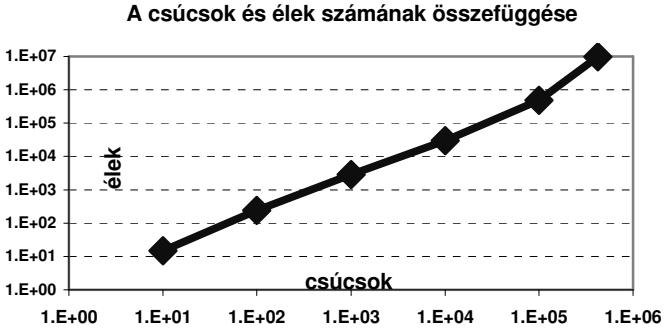
## Az egy „menetben” egyszerre keresett szavak vizsgálata

Az eddigi „magától értetődő” statisztikai feldolgozás mellett az egyes tételek egymáshoz való viszonyát is vizsgálhatjuk. Azaz: akik az x szót keresik, milyen szavakat keresnek még?
A naplófájl tartalmazza a telefon típusát és a WAP-böngésző verziószámát is. Mindez elég specifikus adat ahhoz, hogy azt állíthassuk: az időben szorosan egymás után, azonos telefonról és WAP-browserről érkező kérések nagy valószínűséggel ugyanarról a telefonról, „egy menetben” érkeznek.

### A „koincidenciagráf” előállítása

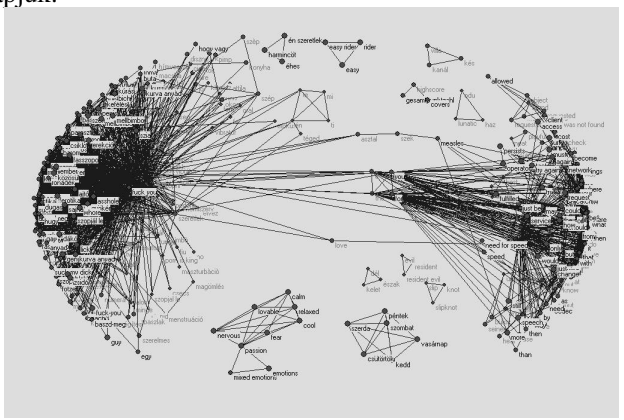
Az adatokat egy gráfban ábrázoltam, melynek csúcsai a keresett kifejezések. Két csúcsot akkor köt össze él, ha a csúcsoknak megfelelő szavak együtt szerepeltek egy lekérdezési menetben. Az élhez hozzárendeltem az együtt szereplés gyakoriságát.

Az így kapott teljes gráf 421 233 csúcsot és 9 776 746 élt tartalmaz. Ekkora adatmennyiséget sajnos egyetlen általam ismert elemző szoftver sem tud kezelni.
Az adatmennyiség csökkenthető, ha csak a nagy értékű éleket és a hozzájuk tartozó csúcsokat tartjuk meg. (Azaz csak azokat a szavakat, amelyek nagyon gyakran szerepelnek együtt egy „menetben”.)




14. ábra

A csökkentett csúcs- és élszámokat loglog skálán ábrázolva (14. ábra), megállapítható, hogy a kapott gráf nem más, mint egy – a napjainkban a legkülönbözőbb tudományterületeken „felfedezett" –skálafüggetlen hálózat[49].
A továbbiakban csak az ezer csúcsot tartalmazó részgráfot vizsgáltam. Ebben a Pajek[50] hálózatkezelő és –megjelenítő szoftver volt a segítségemre.
A gráf csúcsait a Fruchtermann–Reingold[51] algoritmussal átrendezve a következő elrendezést kapjuk:



15. ábra

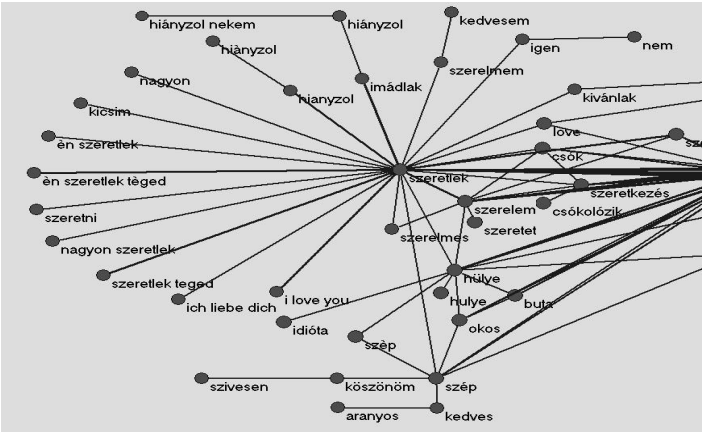
Jól látható, hogy a szavak nagy része két nagy clusterben csoportosul: A baloldali erősen centralizált cluster (a „Pornográf Birodalom") tartalmazza az obszcén szavakat, középpontban a legtöbb kapcsolattal rendelkező *pina* és *fasz* szavakkal. Jobboldalon látható, a sokkal több egyenrangú csúcsot tartalmazó részgráf, alap- és középfokú angol szavakkal (az „Angol Köztársaság"). A két tartomány között csak laza

[49] A téma jó összefoglalása olvasható Barabási Albert-László, vagy Csermely Péter könyvében
[50] http://vlado.fmf.uni-lj.si/pub/networks/pajek/
[51] http://www.boost.org/libs/graph/doc/fruchterman_reingold.html

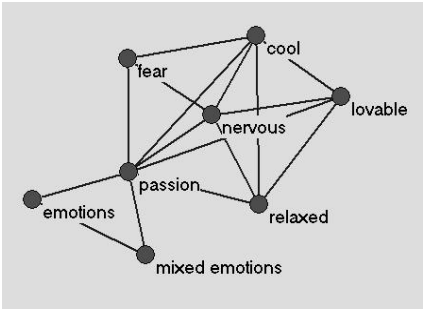
kapcsolat van tesztszavakon (*asztal, szék)* és nagyon gyakori angol szavakon *(a, the)* keresztül.

### *„Felfedezések" a gráfban*

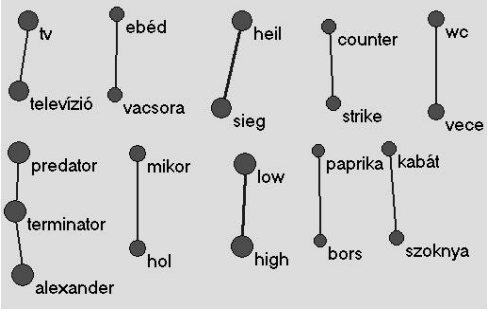
A két nagy tartomány mellett sok kicsi izolált vagy részgráf is feltűnik. Lássunk néhány érdekesebb szemantikai tartalmút:




16. ábra:
A „Pornográf Birodalom" *szerelem* szóhoz tartozó részgráfja, könnyű kitalálni hová is kapcsolódhatnak a jobboldali élek




17. ábra: „Emóciók" egy izolált gráfban




18. ábra Kételemű izolált részgráfok

## Következtetések

A MoBiWAP rendszer üzemeltetése szempontjából a koincidenciagráf vizsgálatának egyértelmű a tanulsága: A sok kapcsolattal rendelkező szavakhoz tartozó szótári tartalmat különös gonddal kell összeállítani: ha a felhasználók elégedettek ezen szavak szócikkeivel, jó esélyünk van arra, hogy többször és más szavak esetében is ezt a szótárat fogják választani.

A koincidenciagráf beható vizsgálata meghaladja jelen cikk lehetőségeit, de (különösen, ha a négyszázszor nagyobb teljes gráfot tekintjük) minden bizonnyal nagyon sok hasznos információt tartalmaz a szemantikai csoportok vizsgálatához, valamint más nyelvszociológiai és nyelvészeti kutatásokhoz.

De erről majd egy másik alkalommal...

## Köszönetnyilvánítás

Köszönettel tartozom Prószéky Gábornak és Kis Balázsnak, akik unszolása és biztatása nélkül soha sem írtam volna meg ezt a cikket; Vöröss Ferencnek, aki gondozta a szótári adatbázis tartalmát.
Köszönettel tartozunk a Pannon GSM-nek a MoBiWAP szolgáltatás megrendeléséért, máskülönben nem jött volna létre a hatalmas vizsgálható szöveganyag.

## Bibliográfia

1. Barabási Albert László: Behálózva – a hálózatok új tudománya (Magyar Könyvklub, 2003)
2. Csermely Péter: A rejtett hálózatok ereje (Vince Kiadó, 2005)