

Kulturális különbségek a mentális lexikonban

Szóasszociációs vizsgálat és számítógépes klaszterelemzés

Jelige: aletheia

Tartalom

1. Bevezetés.....	3
2. A mentális lexikon vizsgálata szóasszociációs tesztekkel	3
2.1. A mentális lexikon fogalma.....	3
2.2. A szóasszociációs tesztek, mint a mentális lexikon megismerésének eszközei	5
2.3. A South Florida Word Association Norms.....	5
3. A mentális lexikon hálózatalapú elemzése.....	6
3.1. Rövid hálózatelméleti bevezető.....	6
3.2. Hálózatok és a nyelv	7
4. Átfedő klaszterek a mentális lexikonban.....	10
5. A vizsgálat módszerei	15
5.1. A teszt összeállításának szempontjai, folyamata, a teszt végső állapota.....	15
5.2. A megkérdezettek, a teszt kitöltésének körülményei	18
5.3. Adatok felvitele, a formátumok.....	18
6. Az eredmények.....	19
6.1. A CFinder használatával kapott eredmények	23
6.2. Az élklasszterező eljárással kapott eredmények és a két algoritmus teljesítményének összehasonlítása.....	27
7. Összegzés	33
8. Bibliográfia.....	35

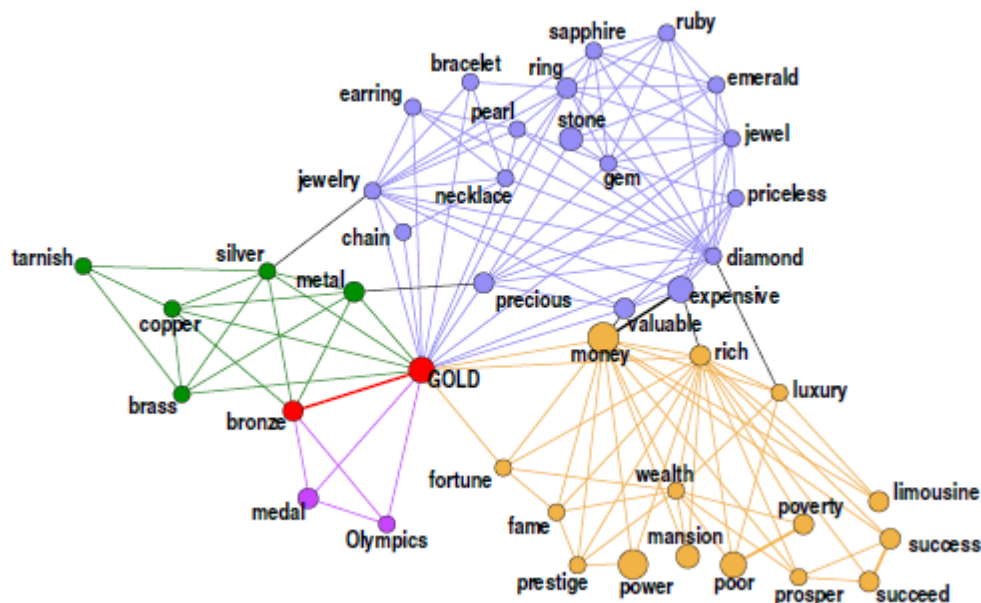
1. Bevezetés

A dolgozat a magyar anyanyelvű beszélők mentális lexikonának egy szeletét, az *arany* szót tartalmazó átfedő klasztereket hasonlítja össze az amerikaiakon végzett tesztek eredményével. Palla és munkatársai (2005a) a hálózatokban jelen lévő, nem diszjunkt csoportosulások kimutatására szolgáló CFinder algoritmusuk működését mutatták be a South Florida Word Association Norms anyagán (Nelson – McEvoy – Schreiber 1998). Az egyik szó, melynek környezetét ábrázolták, a *gold* volt (1. ábra). A módszer bemutatása magyarul is megjelent (Derényi és mtsai 2006). A cikk tartalmaz egy ábrát, melyben a *gold* átfedő klaszterei láthatók, de a csúcsokat (azaz az angol szavakat) magyarra fordították a szerzők (2. ábra). A feltételezésem Kovács (2011) alapján az, hogy a magyar anyanyelvű beszélők mentális lexikonában az *arany* szó körüli jelentésmezők nem ábrázolhatók tökéletesen a *gold* környezetének magyarra fordításával. Ennek ellenőrzéséhez először szükség volt magyar asszociációs normákra, amelyeken az algoritmus futtatható. A 100 emberen elvégzett szóasszociációs teszt eredményeit először Excel-táblába vittem be, majd a CFinder használatához szükséges adatokat egy txt fájlba másoltam. Először a *goldra* és az *aranyra* adott válaszokat hasonlítottam össze. A két asszociációs lista között kulturális és nyelvi különbségek is vannak. Ezt követően csak az eredeti (tehát az angol ábrából fordítással képzett) szavakat figyelembe véve kerestettem klasztereket a CFinderrel. Ha minden beállítást a Palla és munkatársai (2005a) által alkalmazott értékkel hagytam, a program a *gold* körüli négy csoportosulás helyett egyetlen nagy klasztert mutatott csak ki. Ennek oka a magyar szavak közötti sűrű kapcsolatrendszer. Bizonyos beállításokat finomítgatva néhány klaszter kimutatható az adathalmazon a CFinder segítségével, de az olimpia témakörét jelentő csoportosulás egyáltalán nem jelenik meg. Ez magyarázható az *érem* és a *medal* szavak eltérő asszociációival. Végül érdekességképpen megmutatom, milyen klasztereket talál egy átfedések kimutatására nem alkalmas algoritmus a CFinder által egyetlen nagy csoportosulásként ábrázolt hálózatban.

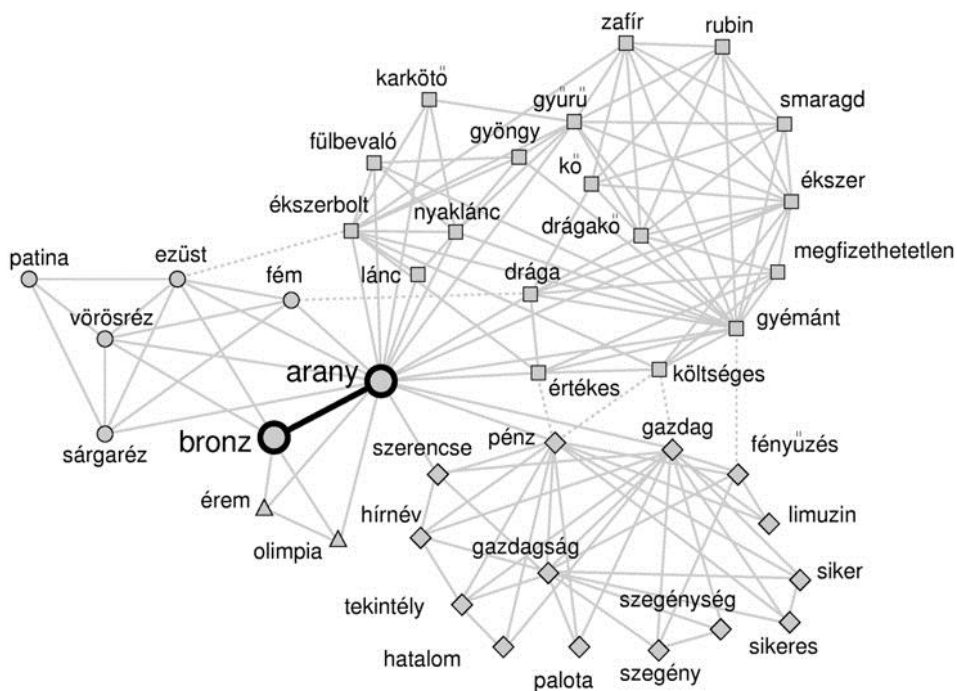
2. A mentális lexikon vizsgálata szóasszociációs tesztekkel

2.1. A mentális lexikon fogalma

A mentális lexikon egyfajta 'agyi szótárát' jelent; „az egyén mindenkori, beszéd- és nyelvi jeleket tároló rendszere képezi”(Gósy 1999: 121) azt. A mentális lexikon mindig az egyénhez



1. ábra: A *gold* körüli átfedő klaszterek a University of South Florida Word Association Norms adatbázisa alapján, ahogyan Palla és munkatársai (2005b)-ben megjelent



2. ábra: A *gold* körüli átfedő klaszterek tagjainak magyarra fordításával kapott ábra, ahogyan Derényi és munkatársai (2006)-ban megjelent

A mentális lexikon felépítését modelláló elméletek közül az úgynevezett háló-elmélet az, ami jelenleg a legvalószínűbb: ez a mentális lexikont hálózatként képzei el, ahol a csúcsok a szavak, egy szó természetesen több szóval is kapcsolatban állhat, és az egymással közvetlenül nem szomszédos pontok is elérik egymást más egységeken keresztül (Gósy 1999: 121, 125).

2.2. A szóasszociációs tesztek, mint a mentális lexikon megismerésének eszközei

A mentális lexikonban lévő szavak közti összefüggések feltárásának régóta használt módja a szóasszociációs teszt (SZAT). Az asszociáció ugyanis feltehetőleg utal arra a kapcsolatra, ami valamilyen módon összeköti a szavakat az egyén elméjében (Wang – Bing – Hou 2010: 48). A SZAT-nak több fajtája van: a szabad asszociációktól a kötött asszociációk abban különböznek, hogy valamilyen kritériumnak meg kell feleljen a válaszsó (például az adott kategória elemeit kell megnevezni, vagy választani kell előre megadott válaszokból). A szabad asszociációk lehetnek egyszavasak vagy meghatározott számú válaszsót várók, illetve lehetnek folyamatosak, amikor annyi szót ír az alany a hívószóra, amennyi csak eszébe jut. Ezen kívül létezik még a megkötés nélküli asszociáció, amikor nincs hívószó, a részt vevő azt ír vagy mond, ami eszébe jut (Wang – Bing – Hou 2010: 48, Kovács 2011: 34–35).

2.3. A South Florida Word Association Norms

A University of South Florida és a University of Kansas közreműködésében jött létre a 36 MB nagyságú, online kezelhető The University of South Florida Word Association, Rhyme and Word Fragment Norms (Nelson – McEvoy – Schreiber 1998, a továbbiakban SFWAN). A készítőik célja az volt, hogy egy nagy, minden kutató számára ingyenesen hozzáférhető szóasszociációs adatbázis váljon elérhetővé. Több mint 6000 résztvevőtől 750 000 választ gyűjtöttek 5019 hívószóra. Az egyszavas, szabad szóasszociáció módszerét alkalmazták. Az anyaggyűjtés 1973-ban kezdődött, az adatbázis 1998-ra készült el. A felületen kereshetünk hívószó vagy válaszsó szerint, és minden kapcsolatnál további információkat is megtudhatunk: az adott szóra válaszolók számát, az adott szóra az adott választ adók számát és arányát, a szavak közti kapcsolat erősségét (a közvetlen kapcsolatukat és esetleg más, közvetett elérési útvonalat a kettő között) és az úgynevezett „közös asszociációkat”.

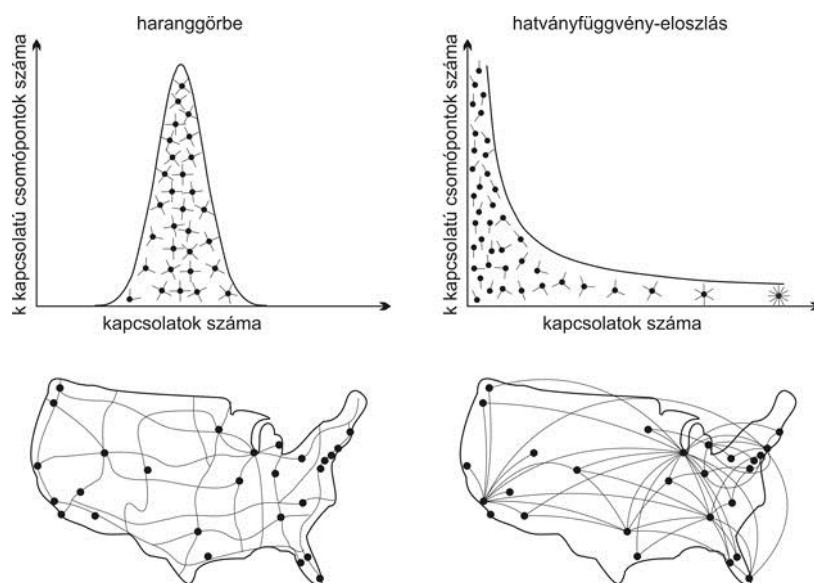
3. A mentális lexikon hálózatalapú elemzése

3.1. Rövid hálózatelméleti bevezető

A nagy mennyiségű és nagy összetettségű adatok hálózatként való reprezentációja lehetővé teszi azt, hogy ezeket hatékonyan elemezzük, megértsük és új tudást hozzunk létre. „Hálózatok mindenhol vannak.” (Barabási 2006: 1298). Hálózatot alkotnak az idegsejtek az agyban, hálózatot alkotnak az emberek egy közösségben, a társadalomban, hálózat az internet is. Az elmúlt 10-20 évben a különböző hálózatok vizsgálata során a kutatók ráébredtek, hogy ezeknek a szerkezetét egyetemes alapelvek irányítják, közös tulajdonságaik révén pedig bizonyos vonásaik megjósolhatóvá válnak (Barabási 2006: 1298). A hálózattudományt általánosan és széles körben lehet alkalmazni különböző jellegű problémákra (terroristák felkutatása és azonosítása, járványok terjedésének előrejelzése stb.), azonban valószínűleg a biológiai jellegű kérdések megválaszolásához használják a leggyakrabban.

A hálózatok gráfok, amelyeknek csúcsaik vannak, azokat pedig élek kötik össze. Az agyban a csúcsok az idegsejtek, melyek között az axonok jelentik a kapcsolatot; a közösségekben az embereket a rokonság, a közös szakma, a barátság köti össze. A legtöbb komplex rendszert jellemző egyik fontosabb tulajdonság a skálafüggetlenség: a fokszámeloszlás, azaz az egyes csúcsokhoz tartozó élek száma hatványfüggvény-eloszlást követ. Ez annyit jelent, hogy nagyon sok olyan pont van, amely kevés kapcsolattal rendelkezik, viszont van néhány olyan pont, amelyből nagyon sok él indul (szemben egy véletlen hálózattal, ahol kevés pontnak lesz nagyon kevés vagy nagyon sok kapcsolata, a legtöbb pontnak ugyanannyi kapcsolata lesz. Ekkor a fokszámeloszlás Poisson-eloszlást követ, lásd 3. ábra.).¹ A skálafüggetlenségnek köszönhetően válnak a hálózatok szilárdná a véletlen hibákkal, viszont sebezhetővé az összehangolt támadásokkal szemben (Barabási 2006: 1298–1307).

¹ A Zipf-törvény lényegében egy hatványfüggvény-eloszlást ír le: ha a szavakat sorbarendezzük gyakoriságuk szerint, akkor a gyakoriság és a szó előfordulási száma (hányadik a sorrendben) közötti kapcsolat hatványfüggvény-eloszlást követ: lesz néhány szó, ami alig-alig fordul elő, és lesz néhány nagyon gyakori (Kovács 2011: 63).



3. ábra: a véletlenszerű és a skálafüggetlen hálózatok fokszám-eloszlásának különbsége. Bal oldalt látható a Poisson-eloszlás, azaz a haranggörbe, amely például az amerikai közúti hálózathoz hasonlít. Jobb oldalt a hatványfüggvény-eloszlás látható, amely viszont a légi útvonalakat jellemzi (Barabási 2006: 1307)

Másik, a nyelvészeti kutatásokban is többször előkerülő tulajdonsága a komplex rendszereknek a kisvilágság. Ez nem más, mint a szociológusok által biztosan jól ismert „hatlépésnyi távolság”: bármely hálózatban, akármilyen nagy is legyen, bármely pont elérhető bármely pontból egy bizonyos (meglepően kis) maximális lépésszámmal. Ennek oka, hogy ezekben a hálózatokban a sűrű helyi kapcsolatrendszerek mellett vannak távolra mutató kapcsolatok is (Barabási 2006: 1301).

Még sok tulajdonságról és érdekességről lehetne írni a hálózatok kapcsán. Én minden további, a jelen kutatáshoz szükséges fogalmat a következő fejezetekben, a hálózatkutatás konkrét nyelvészeti alkalmazásai kapcsán mutatok be.

3.2. Hálózatok és a nyelv

Az eddigi hálózatvezérelt² nyelvészeti kutatásokat egyrészt Kovács (2011) másrészt Ferrer i Cancho (2005) mutatja be részletesen. Ezek nagy részében nem nyelvészek, hanem természettudósok vizsgálták a nyelvet a hálózatok tulajdonságaiból kiindulva. Elsősorban kétféle jellegű hálózatot definiálnak és elemeznek ezek a kutatások: az egyiket a szövegekben egymás mellett vagy egymáshoz közel álló szavak alkotják (szintaktikai hálózatok), a

² A hálózatvezérelt és a hálózatalapú vizsgálat elkülönítésének lényege a kiinduló nézőpont: az utóbbi esetben a nyelvészeti alapról indulunk ki, az előbbiben viszont a hálózat és annak tulajdonságai jelentik a kiindulási alapot (Kovács 2011: 57).

másikban pedig a mentális lexikon elemei jelentik a hálózat csúcsait (szemantikai hálózatok; Kovács 2011: 73–74).

A szintaktikai hálózatok vizsgálata többnyire abból indul ki, hogy két szó között akkor van kapcsolat, ha egy mondaton belül együtt előfordulnak – néhány esetben elegendőnek tekinti a vizsgálat az együttes előfordulást, máskor pedig bizonyos ablakokon belül kell lennie a két szónak. Ez utóbbi „legszigorúbb” formája, amikor a szomszédos szavak között van csak kapcsolat. Szintaktikai hálózatokat vizsgált az angol és más (német, román, cseh) nyelvben több tanulmányban is Ferrer i Cancho és Solé, a magyarban Dominich Sándor és Kiezer Tamás (ezek összefoglalását ld. Kovács 2011: 74–78).

Az előbbieknél tovább ment Ferrer i Cancho, és munkatársaival a függőségi nyelvtan segítségével definiálta a nyelvi hálózatot. Két szó között akkor van kapcsolat, ha függőségi viszony van köztük (Ferrer i Cancho 2005: 63–66).

Az említett vizsgálatok általában kimutatják, hogy az általuk definiált nyelvi hálózat skálafüggetlen fokszámeloszlást mutat, jellemző rá a kisvilágosság. A legtöbb kapcsolattal rendelkező szavak jelenthetik az alapszókincset, amiből kiindulva, és egyéb tényezőket figyelembe véve modellezhető az új szavak elsajátításának menete (Kovács 2011: 74–78).

Áttérve a szemantikai hálózatokra a teauruszok, a WordNet és a szóasszociációs hálózatok vizsgálatáról kell említést tenni. A teauruszok (szinonimaszótárak) esetében két szó között akkor van kapcsolat, ha az egyik a másiknak címszava volt legalább egy bejegyzésben. A legtöbb ilyen jellegű vizsgálat a Roget-féle teauruszra épül (Ferrer i Cancho 2005: 61). A WordNet hálózatalapú vizsgálata többek között Sigman és Cecchi, valamint a Roget-féle teauruszt is elemző Steyvers és Tenenbaum nevéhez kötődik (Sigman – Cecchi 2002, Steyvers – Tenenbaum 2005; a két vizsgálat közti különbség elsősorban az, hogy az utóbbiak mind a négy szófajt (ige, főnév, határozószó és melléknév) vizsgálták, az előbbiek csak a főnevekkel foglalkoztak, ld. Ferrer i Cancho 2005: 62). A WordNet esetében a hálózat adott: a csúcsok a fogalmak (vagy a *synsetek*, vagy akár a szavak is – ebben is eltér a két előbb említett vizsgálat), melyek között a kapcsolat már definiálva van.

A szóasszociációk hálózatalapú vizsgálata kapcsán az előbb említett Steyvers és Tenenbaum (2005) az egyik legfontosabb munka: nem csak a Roget-féle teauruszt és a WordNetet, de az SFWAN egy részét is elemezték (Steyvers – Tenenbaum 2005). A szóasszociációk vizsgálata abban tér el a korábbiaktól, hogy többnyire irányított gráfként definiálják őket: a hívószótól indul az él az asszociáció felé.

Magyar nyelven Kovács László és Lengyel Zsolt foglalkozott szóasszociációk hálózatalapú vizsgálatával (Kovács 2011: 83–85). Kovács (2011)-ben alapos elméleti

bevezető után bemutatja az Agykapocs-kutatást, majd ennek eredményeit elemzi. A mentális lexikonban súlyozott (a gyakoribb asszociációk jelentik az erősebb kapcsolatot, a ritkább asszociációk a gyengét) és irányított kapcsolatokat definiál (157–163). Amennyiben azonban irányított gráfként képzeljük el a mentális lexikont, azt gondolom, nem szabad megállni ennyinél: meg kell vizsgálni, hogy vajon megtaláljuk-e az irányított hálózatokat jellemző kontinenseket³ a mentális lexikonban. Dorogovtsev, Mendes és Samukhin (2001) mutatták ki, hogy mindegy, hogy egy hálózatnak milyen egyéb tulajdonságai vannak, ha irányítottak a kapcsolatai, akkor a négy kontinens meglesz benne.

A szóasszociációs hálózatok alapos, hálózatalapú elemzését végezte el de Deyne és Storms (2008b). Ők a saját, holland nyelvű gyűjtésük adatait vizsgálták. Dolgozatukban kimerítően bemutatják a hálózatok elemzésekor rendre alkalmazott statisztikai módszereket és fogalmakat. Ezeket most röviden ismertetem, hogy később, saját adataim elemzésénél alkalmazhassam őket. A könnyebb érthetőség kedvéért zárójelben mindig feltüntettem az angol terminusokat is. Erre azért is szükség van, mert a magyar terminológia még nem kiforrott, előfordulhat, hogy különböző cikkekben máshogy nevezik meg ugyanazt az angol fogalmat.

A szóasszociációs hálózatban a *csúcsok* (*vertices* vagy *nodes*) a hívószavak, köztük súlyozott és/vagy irányított *élek* (*edges* – *weighted arcs* vagy *directed edges*) vannak, melyek azt jelentik, hogy az adott szó (irányított kapcsolat esetén az él végén álló szó) a másik szóra adott válaszként szerepelt. Fontos kitétel, és az asszociációs hálózatokat vizsgáló tanulmányokra általában igaz, hogy csak a hívószóként is szereplő válaszokat ábrázolják. Ha azokat a válaszokat is felvennénk a hálózatba, amelyek nem voltak hívószavak, egyenlőtlené válna a rendszer, hiszen például ezekből a szavakból biztosan nem lesz kiinduló él, ráadásul helyzetüknél fogva súlyuk is eleve kisebb (de Deyne – Storms 2008b: 216).

A két csúcs közötti legrövidebb távolságot *geodézikus távolságnak* nevezzük (*geodesic path* vagy *geodesic distance*). Ez a csúcsokat összekötő utak (*path*) közül a legrövidebb. Az utak úgy kapcsolják össze az éleket, hogy rajtuk minden csúcs csak egyszer szerepelhet. Két csúcs *szomszédos* (*neighbors*, 'szomszédok'), ha közvetlen kapcsolat van köztük. Irányított gráf esetén minden csúcs rendelkezik egy bejövő és egy kimenő *fokszámmal* (*indegree* és *outdegree*), ami a csúcsba érkező és a csúcsból induló élek számát jelenti. Irányítatlan hálózatban minden csúcsnak egy fokszáma van (*degree*), ami az összes hozzá

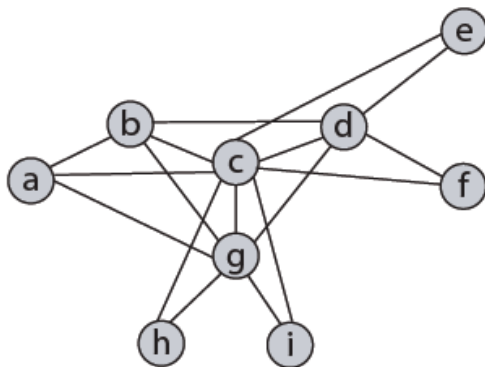
³ Központi mag – minden itteni pont elérhető minden más itteni pontból; BE kontinens – az itteni pontokból csak a központi magba juthatunk, a BE kontinensre visszatérni nem tudunk; KI kontinens – ezek a pontok csak a központi magból érhetőek el, de visszamenni nem tudunk; szigetek – elszigetelt pontok, amelyek nem érhetőek el másik pontokból – ld. például Barabási 2008: 182.

kapcsolódó él száma. Már a fokszám alapján is megállapítható egyfajta központiság (*degree centrality*): a legtöbb kapcsolattal rendelkező pontok a hálózatban a *hubok*. A központiságnak azonban más mérőszámai is vannak. A *közelségi központiság* (*closeness centrality*) kiszámításánál az adott csúcs összes többi ponttól számított legrövidebb távolságát nézzük – ezzel oszjuk el a többi pont számát. A *közöttiségi központiság* (*betweenness centrality*) azt méri, hogy egy pont milyen gyakran szerepel más pontok közötti legrövidebb útvonalakban. Ez egy normalizált alak, 0 és 1 közé esik az értéke. Ha egy pont az összes geodézikus távolságnak része, akkor 1 lesz a közöttiségi központisége, ha egyik ilyen útvonalon sincs rajta, akkor 0. A közöttiség fogalma azért jelentős, mert ha egy nagy közöttiségi központiségű csúcsot eltávolítunk a hálózatból, akkor az darabjaira hullik. A *klaszterezési együttható* (*clustering coefficient*) esetén egy csúcs szomszédjai közötti kapcsolatok számát osztjuk a közöttük lehetséges összes kapcsolat számával. Ez a szám tehát annak a valószínűségét mutatja, hogy két, közös szomszédal rendelkező pont egymásnak is szomszédja. A klaszterezési együtthatók átlaga adja egy hálózat klaszterezési együtthatóját. Szintén nem a csúcsok, hanem a hálózat egészének tulajdonsága a *sűrűség* (*density*). Ez az összes él száma osztva a lehetséges élszámmal, azaz a valós és a lehetséges kapcsolatok aránya. Az *átlagos távolság* (*average distance*) a hálózatban lévő összes csúcspár közötti legrövidebb utak átlaga. Egy hálózat *átmérőjén* a bármely két pont közötti maximális távolság hosszát értjük. Végül a *fokszámeloszlás*, $P(k)$ azt mondja meg, hogy egy véletlenszerűen kiválasztott csúcs milyen valószínűséggel k fokszámú (de Deyne – Storms 2008b: 216–217). A fogalmak könnyebb megértését a 4. ábra segíti.

4. Átfedő klaszterek a mentális lexikonban

A valóvilágbeli hálózatoknak vannak olyan részei, melyekben a csúcsok egymással szorosabb kapcsolatban állnak, mint a hálózat többi részével. Ezeknek a sűrűn kapcsolódó csúcsoknak a halmazát nevezzük klaszternek, csoportosulásnak. Számos módszer létezik a klaszterek kimutatására. Ami ezekben közös, hogy diszjunkt klasztereket definiálnak a hálózatokon, holott a legtöbb esetben ezeket egymással átfedő csoportosulások alkotják (Palla – Derényi – Farkas – Vicsek 2005a: 814).

A



K fokszámok a B hálózatban:

a = 1 b = 2 c = 3
d = 3 e = 1 f = 1
g = 3 h = 1 i = 1

A c csúcs (C hálózat) közelsége (cl):

$$cl(c) = \frac{5-1}{6} = 0.67 \quad /cg+cd+ce+cf=1+1+2+2=6$$

A c csúcs (C hálózat) közöttsége (b):

$$b(c) = \frac{(gd+ge+gf)}{(gd+ge+gf+de+df+ef)} = 0.5$$

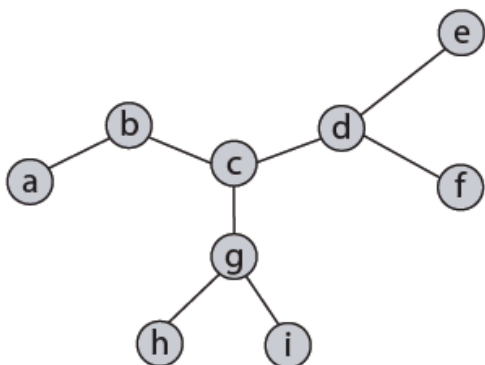
/ gd+ge+gf= azoknak a csomópontokat összekötő
legrövidebb utaknak a hossza, melyek tartalmazzák
c-t
/ gd+ge+gf+de+df+ef = az összes csomópontpárt
összekötő legrövidebb utak hossza

**A B hálózat c csúcsának klaszterezési
együtthatója (CC):**

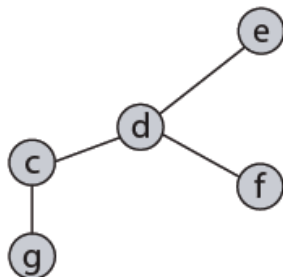
$$CC(c) = \frac{2(0)}{3(3-1)} = 0$$

/ a számlálóban a c szomszédjai közötti kapcsolatok
száma, 2-vel szorozva
/ a nevezőben a c szomszédjai közötti összes
lehetséges kapcsolat száma

B



C



Az A hálózat sűrűsége (d):

$$d(A) = \frac{17}{\frac{9*8}{2}} = 0.47$$

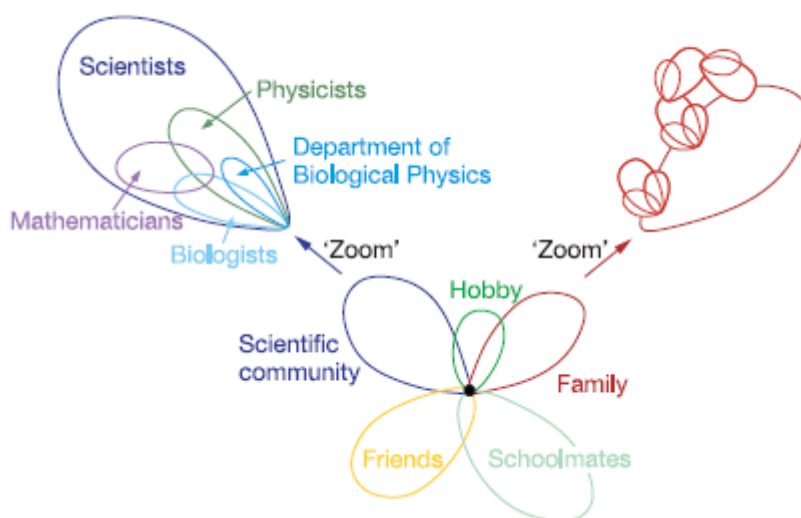
/ a számlálóban az élek száma
/ a nevezőben a lehetséges élek száma

4. ábra: A szóasszociációs hálózatok elemzésére használt néhány fogalom szemléltetése. A három gráf forrása de Deyne - /Storms 2008b: 216

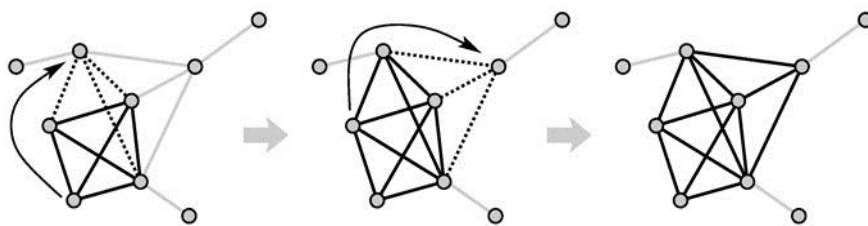
Palla Gergely, Vicsek Tamás, Derényi Imre és Farkas Illés módszerének, a CFindernek az újdonsága épp abban rejlik, hogy az átfedő klasztereket keresi és mutatja ki. Kiinduló gondolatuk, hogy egy tipikus közösség több teljes részgráfból⁴ áll, melyeknek sok csúcsa közös.

⁴ A teljes részgráf olyan gráf, amelynek minden csúcsa össze van kötve minden más csúccsal (Palla – Derényi – Farkas – Vicsek 2005a: 815).

Definíciójuk szerint a közösség, azaz egy k -klikk közösség (*k-clique community*) nem más, mint azon k -klikkek uniója, amelyek elérhetőek egymásból szomszédos k -klikkek⁵ sorozatán keresztül. Ez az alapja a CFinder algoritmusnak: a módszerrel először megkeresik az összes klikket a hálózatban, és ebből kiindulva azonosítják a közösségeket. A 6. ábra szemlélteti az ún. klikk-perkolációs módszert.⁶



5. ábra: Egy szociológiai példa az átfedő klaszterekre: a fekete pont az ábrát tartalmazó cikk akármelyik szerzője lehet. Számos csoportnak tagja, amelyeknek eleve vannak közös tagjai, viszont ha ráközelítünk egy-egy közösségre, azon belül is feltűnnek az átfedő csoportosulások: például a „tudományos csoport” esetében vannak matematikus, fizikus, biológus stb. tagjai a közösségnek, de van, aki egyszerre többhöz is tartozik (Palla – Derényi – Farkas – Vicsek 2005a: 814)



6. ábra: A klikk-perkolációs módszer ($k=4$). A bal szélső ábrán fekete élekkel jelölt klikkből indulunk ki. Minden lépésben egyetlen csúcsot mozdítunk el. A lényeg, hogy a klikk a hálózat csúcsain és élein „gördüljön”. Így teljesül a szomszédosság: két egymást követő lépésben a két klikk egy csúcsban különbözik, $k-1$ csúcsuk közös. A gördülés során bejárt élek és az általuk összekötött csúcsok alkotják a hálózat egy csoportosulását (Derényi – Farkas – Palla – Vicsek 2006).

⁵ Két k -klikk akkor szomszédos, ha $k-1$ közös pontjuk van (Palla – Derényi – Farkas – Vicsek 2005a: 815).

⁶ Ejtteni kell néhány szót a CFinder hiányosságairól, hátrányairól is. Az egyik legfontosabb probléma a saját vizsgálatomban is előkerül majd, tudniillik hogy a közösségekbe nem tartozó csúcsok (márpedig ilyenek vannak, nem csupán a szóasszociációkban, de az emberi közösségekben stb. is) nem jelennek meg az ábrán, a CFinder ezekkel nem tud mit kezdeni. Ezen kívül az sem egyértelmű, hogy a k értékét pontosan mi alapján, hogyan kell megválasztani ahhoz, hogy értelmes struktúrát kapjunk. Fortunato (2010)-ben az összes kritika össze van foglalva, többek között az is, hogy ha túl sok a klaszter, a program egy egyszerű közösség képével tér vissza, azaz az egész hálózatot egy nagy közösségként fogja megjeleníteni (131).

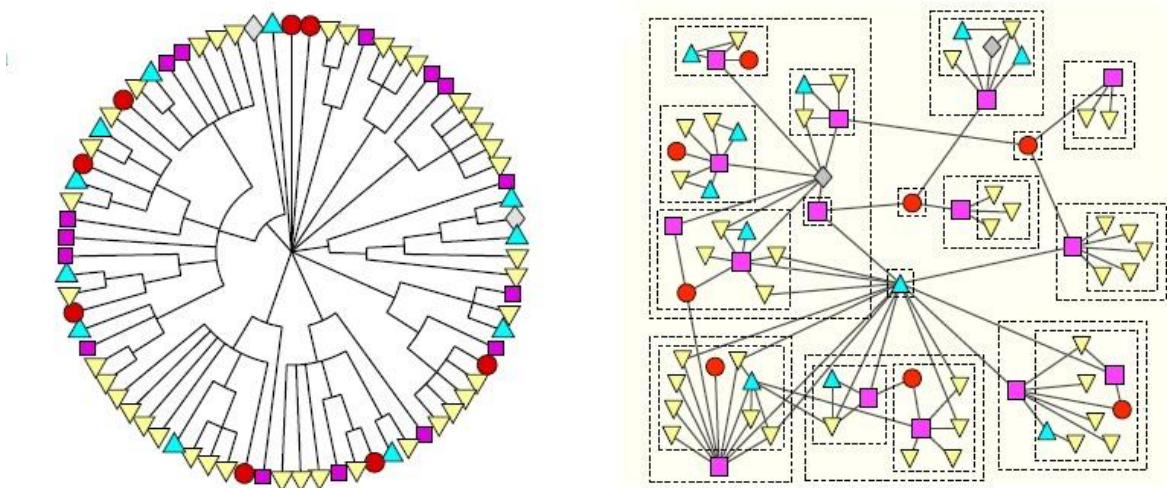
Azóta már számtalan átfedő klaszterek kimutatására szolgáló algoritmust fejlesztettek ki. Mivel dolgozatomban nem a különböző klaszterezési eljárások összehasonlítása a célom, ráadásul ezt már megtette Fortunato (2010), most csak utalok arra, hogy a cikkben minden klaszterezési eljárás bemutatásra kerül, előnyével és hátrányával együtt, beleértve ebbe az átfedő csoportosulások kimutatására alkalmas módszereket (és a CFindert) is. A klaszterek elemzésének kérdése a nyelvészetben megjelenik például a szófelismerés vizsgálatában: A klaszterezési együttható és a szófelismerés kapcsolatát vizsgálta Altieri, Gruenenfelder és Pisoni (2010). Arra az eredményre jutnak, hogy a sűrű szomszédság (azaz a magas klaszterezési együttható) lassítja a hallás utáni szófelismerést és rontja a pontosságot. Emellett például klaszterelemzéssel próbálja megragadni a kváziszinonimák jelentésének különbségét Divjak és Gries (2008).

A vizsgálatomban használt másik klaszterezési eljárás neve hierarchikus élklasztterezés (*hierarchical link clustering*, Ahn – Bagrow – Lehmann 2010a). Ezzel kapcsolatban két fogalmat kell tisztázni, a hálózatok hierarchiájának kérdését és az élklasztterezés jelentését.

Számos hálózatban figyelhető meg, hogy a közösségek rekurzívan hierarchikus struktúrába csoportosulnak, azaz a csúcsok klaszterekbe osztódnak, amelyek tovább oszthatók klaszterek klasztereivé, és így tovább, több lépésen keresztül. Ezekről a különböző szinteken lévő csoportokról több esetben ki lett mutatva, hogy valamilyen funkcionális szerepük van: biokémiai hálózatok (fehérje-interakciós hálózatokban, metabolikus hálózatok) moduljait vagy társadalmi hálózatok közösségeit jelentik (Clauset – Moore – Newman 2008: 98). A hálózati hierarchia (egyik fajta) szemléltetése látható a 7. ábrán.

A gond azonban (legalábbis Ahn és munkatársai szerint) az, hogy a legtöbb való világbeli hálózatban a közösségek átfednek, minden egyes csúcs több csoporthoz is tartozik, ezt pedig, azaz az átfedő csoportosulások közti kapcsolatot a csúcsok alapján megrajzolt globális hierarchia nem tudja szemléltetni. Ezért indulnak ki a szerzők abból, hogy a klaszterek élek, nem pedig csúcsok csoportjai, ami lehetővé teszi az átfedések és a hierarchikus szerveződés egyidejű detektálását (Ahn – Bagrow – Lehmann 2010a: 761).

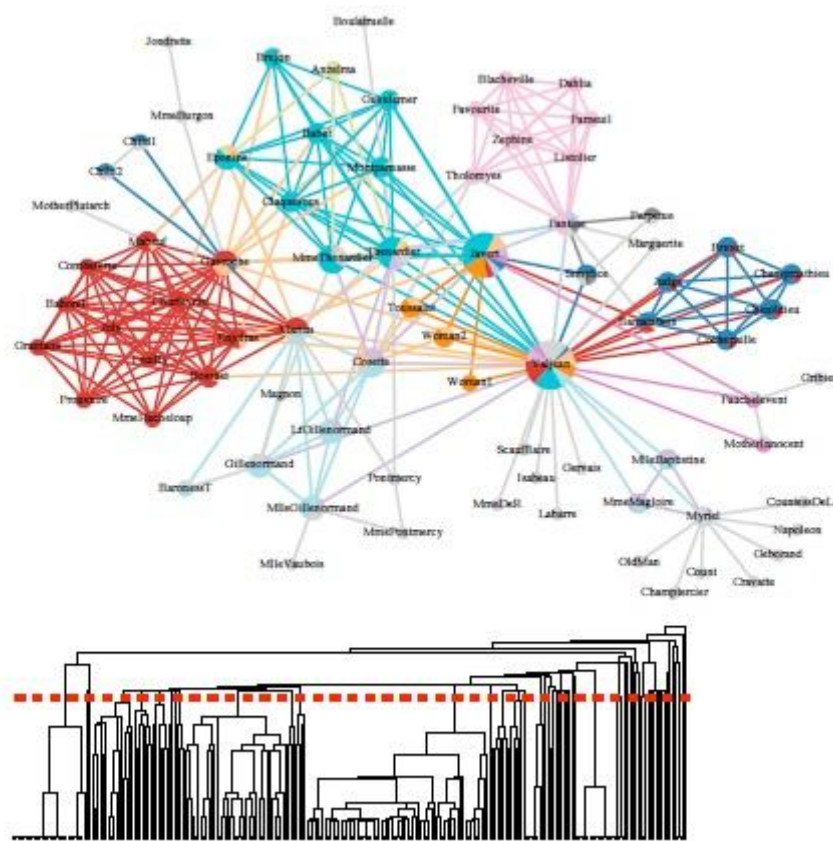
Az élklasztterezés lényege tehát, hogy nem a csúcsok csoportokba tartozását vizsgáljuk elsősorban, hanem az azokat összekötő éleket, és a klasztereket egymással szoros kapcsolatban álló élek halmazának tekintjük.



7. ábra: A hálózati hierarchia szemléltetése (füves területek élővilágának interakciós hálózatán – az 5 különböző jel 5 fajt jelöl). A bal oldali dendrogramon azt látjuk, hogy az egymással szorosabb kapcsolatban álló elemeknek alacsonyabban van közös ősök (alacsonyabban vannak összekapcsolva), mint a távolabb állóknak. A jobb oldali vizualizációban pedig a csúcsok köré rajzolt dobozok jelentik a dendrogram szintjeit (az ábra forrása Clauset – Moore – Newman 2008: 99).

A megközelítés egyik fontos alapja, hogy míg a csúcsok számos közösség tagjai lehetnek, a kapcsolatok jellemzően egy domináns csoportosulásba tartoznak. Tehát például egy személy tagja a családjának, a munkahelyének, a baráti közösségeinek és így tovább. Azonban két ember között a kapcsolatot az jelenti, hogy azonos családnak a tagjai, vagy egy helyen dolgoznak, vagy közös az érdeklődési körük (persze bizonyos esetekben a kapcsolat jelen lehet több területen is, például ha egy család tagjai egy helyen dolgoznak, de a köztük lévő él dominánsan az egyik terület, mondjuk a család klaszterére jellemző). Az élklaszterezésben az élek egyedi pozíciót kapnak, a csúcsok pedig természetesen öröklök az összes, őket tartalmazó él „csoporttagságát” (Ahn – Bagrow – Lehmann 2010a: 761).

Az élklaszterezést szemlélteti a 8. ábra. Ezen jól látszik, hogy bár a dendrogram (alul) jó szemléltető eszköz, és a hálózat gazdag hierarchikus struktúráját segíti feltárni, az alsó része „túl sűrű”. Annak érdekében, hogy ebből a halmazból kiválassza a legrelevánsabb klasztereket, a módszer az eloszlási sűrűség (*partition density*) függvényét alkalmazza (Ahn – Bagrow – Lehmann 2010a: 761). Ez tulajdonképpen egy küszöb, ahol kettévágjuk a dendrogramot, és az ez alatti közösségek belső szerveződésével nem foglalkozunk (lásd a piros szaggatott vonalat a 8. ábrán). Az eloszlási sűrűség a klasztereken belüli élsűrűségen alapszik.



8. ábra: az élklasszterezés szemléltetése (a *Nyomorultak* című regény szereplőinek kapcsolati hálóján). Fent a színes élek a klaszterek elkülönítésére szolgálnak, a csúcsokat, azaz a szereplőket pedig apró diagramok jelképezik, ahol a különböző színek az adott pont közösségekre való tartozását jelentik. Lent az él-alapú dendrogram látható. A dendrogram legalján az egyes pontok éleket, és nem csúcsokat jelölnek (az ábra forrása Ahn – Bagrow – Lehmann 2010b).

5. A vizsgálat módszerei

5.1. A teszt összeállításának szempontjai, folyamata, a teszt végső állapota

A teszt összeállításában, a vizsgálat kivitelezésének módjában igyekeztem minél inkább igazodni az SFWAN készítői által alkalmazott módszerekhez (Palla és mtsai ezen az adatbázison mutatták be algoritmusuk működését, tehát a *gold* körüli klaszterek ennek az adatait tükrözik). Ezért a szabad, egyszavas választ igénylő szóasszociáció formáját választottam: minden résztvevőnek minden egyes szóra egyetlen másikat kell írnia, amely elsőre eszébe jut a hívószóról. Természetesen izgalmasabb lehet, ha három szót is írhat válaszként a résztvevő, azonban annak definiálása, hogy a második és harmadik asszociáció milyen kapcsolatban áll a hívószóval, az első válaszzóval, illetve a harmadik szó a másodikkal, és vajon melyik kapcsolat milyen erős, meghaladja ennek a dolgozatnak a kereteit, és alapos elméleti megalapozást igényel.⁷

⁷ De Deyne és Storms (2008a és 2008b) a már említett, holland nyelvű szóasszociációs vizsgálatában három válaszzót várt minden résztvevőtől minden egyes hívószóra. Pont azzal indokolták választásukat, hogy így a

Mivel kutatásomban egy konkrét szó környezetét akarom felderíteni, ezért ez jelentősen meghatározza a szavak listáját. Az *arany* természetesen minden kérdőívben szerepel, ezen kívül felhasználom az összes szót, ami az angol cikk magyar fordításában megjelent a csoportokban:⁸

- az olimpia köréből: *bronz, érem, olimpia*;
- a nemesfémek: *patina, vörösréz, sárgaréz, ezüst, fém*;
- a jólét: *szerencse, hírnév, pénz, gazdag, gazdagság, fényűzés, limuzin, siker, sikeres, szegénység, szegény, palota, hatalom, tekintély*;
- az ékszerek: *lánc, ékszerbolt, nyaklánc, fülbevaló, karkötő, gyöngy, gyűrű, kő, drágakő, zafír, rubin, smaragd, ékszer, megfizethetetlen, gyémánt, költséges, értékes, drága*.

Ezen kívül szerepel a listán a *János* (esetleges *Arany* - *János* asszociációk tesztelésére), illetve egyéb olyan szavak, amelyek nem részei a *gold* körüli jelentésmezőknek. Ezekre azért van szükség, hogy az asszociációk lehetőleg egymástól függetlenek legyenek, tehát a *gyűrű*, a *karkötő* és az *ékszer* szavak például ne kövessék egymást, mert így az első két hívószó által aktivizált mezők miatt az *ékszer* asszociációja leszűkül.

Ezeket a „töltelékszavakat” is véletlenszerűen választottam: a Magyar Nemzeti Szövegtárban leggyakoribb szónak számító *a*-ra kerestem rá a keresőfelületen, 38 elemű véletlen mintával, 10 szavas kontextusban. Mindegyik sor esetén random számgenerátorral választottam ki, hogy az adott sorból hányadik szót vegyem fel a listára. Ha a kiadott számnak megfelelő helyen nem tartalmas szó állt, addig lépkedtem balra a sorban, míg nem találtam egyet, és azt vettem fel végül (- amennyiben balra haladva nem leltem egyetlen tartalmas szót sem, jobbra indultam). Minden szóalakot jelek és ragok nélkül vettem fel, de a képzőket, és így a szófajt meghagytam, ezért szerepel a listán például főnévi igenév is.

Azért, hogy biztosítsam, hogy a korábban említett befolyásolás ne legyen jelen a kérdőívben, a szavakat véletlenszerűen, számítógépes permutációgenerátorral rendeztem el, minden kérdőívben más-más sorrendben.

gyenge asszociációk is megjelenhetnek. Ők kifejezetten azt a különbséget akarták vizsgálni, ami az első, illetve a második és a harmadik asszociáció generálásában jellemzi a válaszolókat. Éppen ezért dolgozatuk jelentős részében ezeket a különböző sorszámú válaszokat hasonlították össze.

⁸ A magyar fordítás nem felel meg mindenhol az eredeti angol szavaknak: például a *success* és a *succeed* *sikerként* és *sikeresként* van fordítva, a *prosper* fordítását egyáltalán nem is találni a magyar ábrán. Mivel mégis csak ez az ábra a kutatásom kiindulópontja, megtartottam ezeket a szavakat, és nem nyúltam vissza az eredeti angol szavakhoz.

A szavak listája, kezdve a „kötelező” szavakkal:

bronz	limuzin	gyűrű
érem	siker	kő
olimpia	sikeres	drágakő
platina	szegénység	zafír
vörösréz	szegény	rubin
sárgaréz	palota	smaragd
ezüst	hatalom	ékszer
fém	tekintély	megfizethetetlen
szerencse	lánc	gyémánt
hírnév	ékszerbolt	költséges
pénz	nyaklánc	értékes
gazdag	fülbevaló	drága
gazdagság	karkötő	János
fényűzés	gyöngy	

Ezek pedig a random kiválasztott „töltelékszavak”:

megtartani	érez	három
nagykapu	család	embercsempész
köztársaság	közlekedés	pontos
aggasztó	hivatal	sör
kovács	jog	hitel
felmerül	szervező	kártya
filmszakma	megy	szabadságharc
tánc	elfogadni	pozitív
vall	földrengés	mese
jó	sláger	kíván
dolog	paragrafus	település
távozik	beszélget	eső
vezető	kamasz	

5.2. A megkérdezettek, a teszt kitöltésének körülményei

A tesztet pontosan 100 magyar anyanyelvű személy, 22 általános iskolás, 23 végzős gimnazista, 33 egyetemista és 17 középkorú töltötte ki. Mindenkit arra kértem, hogy az első szót, ami az olvasott szóról eszükbe jut, írják le. Sorban haladjanak, ne olvassanak előre, és ne gondolkozzanak a válaszon. Ha nem jut eszükbe semmi, akkor menjenek tovább. Ne írjanak mondatokat, viszont szerepelhet ugyanaz a szó többször is a válaszaikban.

5.3. Adatok felvitele, a formátumok

A kérdőíveket először Excel-táblába írtam be. Egy sor egy asszociációt jelképezett. Négy oszlopba kerültek az adatok: a hívószó, a válasz, a csoport kódja (I= iskolás, E=egyetemista, F=felnőtt) és a kitöltő sorszáma (csoportonként újrakezdve a számozást). Ez utóbbira azért volt szükség, hogy ha később olyan lekérdezésekre lenne szükség, amelyek azzal kapcsolatosak, hogy mi szerepelt ugyanazon a kérdőíven belül, vagy egy adott csoport mit válaszolt stb., akkor ezek elérhetőek legyenek. Példaként íme az Excel-tábla egy szelete:

MEGY	SÉTA	I	23
MEGTARTANI	TÁMASZ	I	23
EZÜST	GYŰRŰ	I	23
ÉREZ	SZERELEM	I	23

7. ábra: Egy példa az asszociációkat tartalmazó Excel-táblából. Az első oszlop a hívószó, a második a válasz, a harmadik és a negyedik pedig a tesztlap egyedi azonosítója (jelen esetben az I, azaz az iskolás csoport 23-as számú résztvevőjének néhány válasza látható)

Ha valaki két szót írt egy helyre, akkor az elsőt vettem figyelembe. Ha írt valamit, áthúzta, és fölé írt egy újabbat (az általános iskolás csoportnál történt néha), akkor az eredeti, áthúzottat vettem figyelembe, feltételezve, hogy az volt a résztvevő első gondolata, csak valamilyen megfelelési igény miatt egy szebbet, jobbat, szalonképesebbet írt oda. Ha valaki egész mondatban, vagy körülírással válaszolt, akkor azt úgy hagytam. Mivel egyszeri előfordulás, ezért a továbbiakban nem lesz már jelen az adathalmazban.⁹

Különféle szűrések után (ezekről részletesebben az eredményeknél) sima szöveges fájlokat hoztam létre a táblázatokból, és ezeken dolgozott a CFinder és az élklaszterező algoritmus.

⁹ Még egy esetet meg kell említenem: az egyik gimnazista papírján a második oldaltól kezdve mindenhol a *pénz* szerepelt válaszként, majd az utolsó oldalon a kitöltő meg is jegyezte, hogy ő ezt nem vette komolyan. Válaszait az első *pénztől* kezdve nem vettem figyelembe.

6. Az eredmények

Az első táblázat a *goldra* illetve az *aranyra* adott válaszok listáját tartalmazza. Az egyes asszociációk melletti szám a válasz relatív gyakorisága.¹⁰ Jelen esetben a válaszsavakat gyakoriságuk szerinti csökkenő sorrendbe rendeztem. A második táblázatban a jobb átláthatóság és a könnyebb összehasonlíthatóság érdekében a mindkét listán szereplő szavakat egymás mellé raktam, a csak az egyik vizsgálatban előforduló asszociációk melletti helyet a másik oszlopban üresen hagytam. A sorrend a SFWAN első táblázatbeli sorrendjét követi. Helytakarékoság céljából az idioszinkretikus, ráadásul a másik adathalmazban párral nem rendelkező adatokat csak felsorolom a táblázat után. A táblázatok közepén üres sor jelzi a *hapax*, egyszer szereplő asszociációk listájának kezdetét. Bár a továbbiakban ezekkel az adatokkal nem foglalkozom, ennél az összehasonlításnál mégis érdekes ezeket is számba venni: van olyan szó, amely az egyik oldalon idioszinkretikus, a másikon nem. Mindkét táblázatban az angol szavak mellett ahol szükséges, zárójelben megadom a magyar fordítást, és ezt a gyakorlatot követem az eredmények további bemutatásánál is. Ahol nem egyértelmű, hogy a szó melyik jelentésére gondolhatott a válaszoló, ott erre külön kitérek.

SILVER (ezüst)	.472	ÉKSZER	.09
MONEY (pénz)	.102	ÉREM	.09
JEWEL (ékszer)	.096	EZÜST	.06
RICH (gazdag)	.046	JÁNOS	.05
DIAMOND (gyémánt)	.030	PÉNZ	.04
METAL (fém)	.025	BÁNYA	.03
VALUE (érték)	.020	GYŰRŰ	.03
CHAIN (lánc)	.015	LÁNC	.03
EXPENSIVE (drága)	.015	ALMA	.02
NECKLACE (nyaklánc)	.015	BULLA	.02
RING (gyűrű)	.015	DRÁGA	.02
SHINE (fénylik) ¹¹	.015	ÓRA	.02
WATCH (óra)	.015	RÚD	.02
WEALTH (gazdagság/vagyon)	.015	SÁRGA	.02
YELLOW (sárga)	.015		

¹⁰ A *goldra* 197 válasz érkezett, tehát az angol szavak esetében a relatív gyakoriság az adott szót válaszként írók száma osztva 197-tel. Én 100 embertől gyűjtöttem választ az *aranyra*, így itt 100-zal osztottam az adott válaszsavot tartalmazó kérdőívek számát. Például 93-an írták a *silver*t asszociációként, így annak a relatív gyakorisága 93/197. Az *ékszer* 9 résztvevő lapján szerepelt az *aranyra* adott válaszként, így annak relatív gyakorisága 9/100.

¹¹ A fordításoknál figyelembe vettem a SFWAN adatbázisban található szófaji jelöléseket. Így például a *shine* fordítása 'fénylik' (V), és nem 'ragyogás' (N), mivel az adatbázisban igeiként van megjelölve.

PRECIOUS (drága/értékes)	.010		
14 K	.005	1. HELYEZETT	.01
BAND (karikagyűrű) ¹²	.005	ARANY	.01
BEAUTY (szépség)	.005	ARÁNY	.01
BLACK (fekete)	.005	ARANYÉREM	.01
BOUILLON	.005	ARANY JÁNOS	.01
BRACELET (karkötő)	.005	ARANYLÁZ	.01
BRIGHT (ragyogó) ¹³	.005	ARANYOS	.01
CARD (kártya) ¹⁴	.005	ARANYRÖG	.01
COIN (érme)	.005	ARANYRÚD	.01
ELEMENT (elem) ¹⁵	.005	ARANYSÁRGA	.01
FORTUNE (szerencse / vagyon)	.005	ARANYTALLÉR	.01
MINERAL (ásvány)	.005	ÁSÓ	.01
RUSH (hajsza) ¹⁶	.005	DRÁGAKŐ	.01
SPARKLE (csillogás)	.005	ELEKTRÓD	.01
		ÉRME	.01
		ÉRTÉK	.01
		EZÜSTÉREM	.01
		FÉNYEK	.01
		FÉNYLIK	.01
		GAZDAGSÁG	.01
		HAJ	.01
		HARAG	.01
		HÖRCSÖG	.01
		JÓKAI MÓR	.01
		KAPU	.01
		KARÁT	.01
		KARKÖTŐ	.01
		KINCS	.01
		KORONA	.01
		KUPA	.01
		NAGYMAMA	.01
		NAP	.01
		PAPRIKA	.01
		SÁRGARÉZ	.01
		SZÉF	.01

¹² Bár a *band*nek számtalan jelentése van, a *gold* miatt én ebben a helyzetben a 'karikagyűrűt' tartom a legvalószínűbbnek.

¹³ A *bright* sok jelentése közül talán a 'fényes', 'ragyogó' lehet a *gold*dal kapcsolatban álló. Mivel ez az asszociáció egyszerű, ezért nem foglalkozom mélyrehatóbban a szó jelentésével.

¹⁴ A *GoldCard* egy játékkonzol, egy memóriakártya, egy hitelkártya-típus sőt egy texasi egészségbiztosítási kártya neve is.

¹⁵ *Chemical element* – 'kémiai elem'.

¹⁶ *Gold rush* – 'aranyláz'.

		TALLÉR	.01
		TEHÉN	.01
		TÖMB	.01
		TUTANKHAMON	.01

1. táblázat: A *goldra* és az *aranyra* adott válaszok listája. Az asszociációk után külön oszlopban a szó relatív gyakorisága

SILVER (ezüst)	.472	EZÜST	.06
MONEY (pénz)	.102	PÉNZ	.04
JEWEL (ékszer)	.096	ÉKSZER	.09
RICH (gazdag)	.046		
DIAMOND (gyémánt)	.030		
METAL (fém)	.025		
VALUE (érték)	.020		
CHAIN (lánc)	.015	LÁNC	.03
EXPENSIVE (drága)	.015	DRÁGA	.02
NECKLACE (nyaklánc)	.015		
RING (gyűrű)	.015	GYŰRŰ	.03
SHINE (fénylik)	.015	FÉNYLIK	.01
WATCH (óra)	.015	ÓRA	.02
WEALTH (gazdagság/vagyon)	.015	GAZDAGSÁG	.01
YELLOW (sárga)	.015	SÁRGA	.02
PRECIOUS (drága/értékes)	.010	DRÁGA ¹⁷	.02
		ÉREM	.09
		JÁNOS	.05
		BÁNYA	.03
		RÚD	.02
		BULLA	.02
		ALMA	.02
BRACELET (karkötő)	.005	KARKÖTŐ	.01
COIN (érme)	.005	ÉRME	.01
RUSH (hajsz)	.005	ARANYLÁZ ¹⁸	.01

2. táblázat: A *goldra* és az *aranyra* adott válaszok listája, megfeleltetésekkel. Az angol hapax, magyar párral nem rendelkező asszociációk a következők: *14 K* ('14 karát'), *band* ('karikagyűrű'), *beauty* ('szépség'), *black* ('fekete'), *Bouillon*, *bright* ('ragyogó'), *card* ('kártya'), *element* ('elem'), *fortune* ('vagyon', 'szerencse'), *mineral* ('ásvány'), *sparkle* ('csillogás'). A csak a magyar adatok között és csak egyszer előforduló szavak: *1. helyezett, arany, arány, aranyérem, Arany János, aranyos, aranyrúd, aranyárga, aranytallér, ásó, drágakő, elektród, érték, ezüstérem, fények, haj, harag, hörcsög, Jókai Mór, kapu, karát, aranyrög, kincs, korona, kupa, nagymama, nap, paprika, sárgaréz, széf, tallér, tehén, tömb, Tutankhamon*

¹⁷ Mivel mind az *expensive*, mind a *precious* jelentése megfelelhet a *drága* jelentésének, ezért mindkét sorban feltüntettem a *drágát*.

¹⁸ Bár nem fordítása a *rush*nak az *aranyláz*, én a jelentéskapcsolat (*gold rush* = 'aranyláz') miatt mégis egy sorba írtam a kettőt. A továbbiakban úgysem lesznek részei az elemzésnek.

A gyakoriságokat megnézve az egyik legszembetűnőbb különbség, hogy míg az amerikaiak majdnem fele ugyanazt a választ adta a *goldra*, a *silver*-re, addig a magyar teszt kitöltőinél nincs ilyen nagy gyakorisággal rendelkező asszociáció. A lista első két eleme, az *ékszer* és az *érem* is csak 9-9%-ban jelent meg a válaszok között. Tehát míg az angol anyanyelvűeknél van egy kiugróan gyakori asszociáció, ezt követi egy szintén elég magas, .102 –es relatív gyakoriságú (*money*), addig a magyaroknál több, nagyjából azonos gyakoriságú szó áll az első helyen, de ezek egyike sem szerepelt a kitöltött tesztek több mint 10%-án (*ékszer*, *érem*, *ezüst*, *János*). Egyszóval a magyar anyanyelvűek asszociációja diverzebb.

Szintén az első táblázatból látszik, hogy – az egyszeri előfordulásokat nem véve figyelembe – nagyjából ugyanannyi válasz érkezett a *goldra* és az *aranyra*. A hapax asszociációk esetén azonban más a helyzet: a magyar nyelvű vizsgálatban két és félszer több egyszeri előfordulású válasz szerepel (39) mint az angolban (14). Ez a szám is a *gold* nagyobb homogenizáló erejére utal.¹⁹

A második táblázatra térve megállapíthatjuk, hogy 10 közös asszociáció van a két vizsgálatban.²⁰ Csak angolul szerepel a *rich*, a *diamond*, a *metal*, a *value*, és a *necklace*. Ezek egyikét sem mondhatjuk az amerikai kultúrára, vagy az angol nyelvre specifikusnak, ráadásul a *value* kivételével, amely nem szerepelt hívószóként a vizsgálatomban, mindegyik magyar fordítása előhívja a magyar anyanyelvűek szóasszociációiban az *aranyat*²¹ - igaz az első kettő esetében mindössze egy-egy adat van erre. Csak a magyarok válaszai közt fordul elő az *érem*, a *János*, a *bánya*, a *rúd*, a *bulla*, és az *alma*. Ez utóbbiak közül néhány egyértelműen kultúraspecifikus: a *János* (Arany János), a *bulla* (aranybulla) és a – nem csupán magyar – népmesék, mondák, mítoszok fontos motívuma, az *alma* (aranyalma). Az *érem*, a *bánya* és a *rúd* nem sajátosan magyar asszociációk, viszont a *gold* mint hívószó nem hívja elő ezeket az amerikaiak vizsgálatában. Ezzel szemben a *medal* és a *mine* hívószóra is érkezett *gold* asszociáció (.211 és .048 relatív gyakorisággal), tehát ha irányított gráfként képzeljük el a mentális lexikont, akkor fordított irányban ez a két kapcsolat az angol adatok szerint is fennáll. Ha irányítatlan gráfként reprezentáljuk a mentális lexikont, akkor pedig egyszerűen mondhatjuk, hogy ez a két asszociáció mind a két nyelv beszélőinek asszociációiban megjelenik. A *rúd* esetében azonban fordított irányú asszociációt sem találunk a SFWAN

¹⁹ A hívószavak homogenizáló erejéről ld. Lengyel 2009.

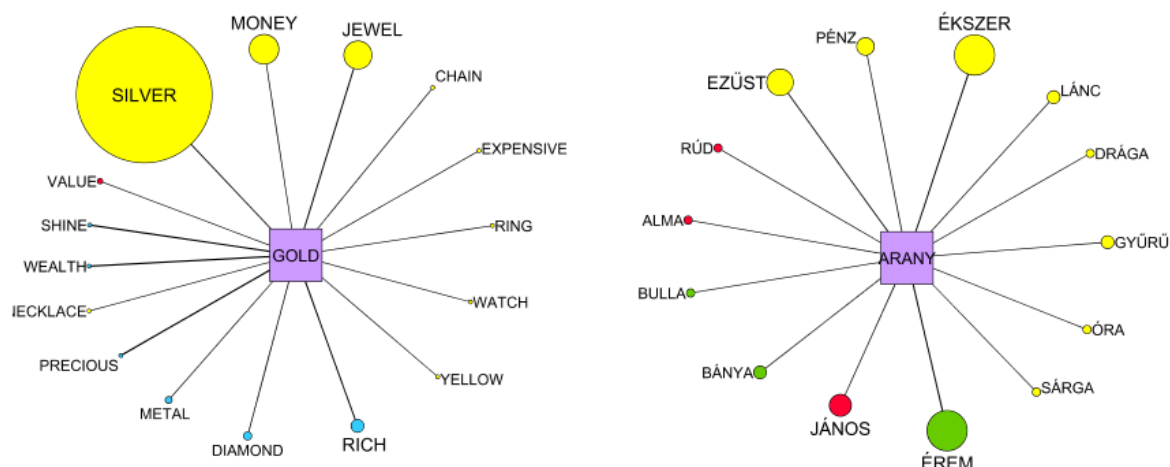
²⁰ A *precious*, *expensive* és a *drága* már említett nem egyértelmű kapcsolata miatt 11 közös asszociációról is lehet beszélni, ami az angol nyelvű adatokban 11, a magyar nyelvűekben 10 szó kapcsolatát jelenti.

²¹ Tehát van *gazdag* - *arany*, *gyémánt* – *arany*, *fém* – *arany* és *nyaklánc* – *arany* asszociációra adatunk, ahol az első szó a hívószó, az *arany* pedig a válasz.

adatbázisában: a *barra* nem adták a *goldot* válaszként. Ehhez hozzátartozik, hogy a *bar* asszociációi egy kivétellel mind a 'bár' jelentéséhez kapcsolódnak: *drink, beer, stool, alcohol, drunk, grill, room, bartender, liquor, club, dance, pub*. A kivétel a *pole*.²²

Három közös hapax asszociációt látunk, bár a *rush* és az *aranyláz* nem teljesen pontos megfelelői egymásnak. Viszont a többi egyszeri előfordulás abszolút egyedi. Van azonban két olyan eset, amelyben a magyar adatok között csupán egyszer előforduló szó angol párja az angol tesztben nem hapax: a *shine* (.015) és a *wealth* (.015).

Az eddig elmondottak szemléltetését szolgálja a 8. ábra.²³ Ezen a csúcsok nagysága a gyakoriságot mutatja: a nagyobb kör súlyozottabb kapcsolatot, azaz az asszociáció nagyobb relatív gyakoriságát jelenti. Sárga színnel kitöltve, a két ábrán nagyjából azonos helyen láthatók a közös asszociációk, és késsel a csak az angol, zölddel a csak a magyar oldalon szereplő szavakat jelképező csúcsok. Kék szín jelzi azokat az angol szavakat is, amelyeknek a magyar megfelelője csak idioszinkretikus asszociáció. Piros színűek azok a pontok, melyek a „valódi egyedi” asszociációk, azaz amelyeknek fordított irányú asszociációjára sincsen példa a másik adathalmazban.



8. ábra: A *goldra* és az *aranyra* adott asszociációk. Sárga színnel a közös, kék, illetve zöld színnel az ebben az irányban (tehát a *goldra* és az *aranyra* adott válaszként) egyedi, piros színnel az abszolút egyedi (a fordított irányú kapcsolatként, tehát hívószóként a *goldot* vagy az *aranyat* asszociálva sem megjelenő) asszociációk. A csúcsok nagysága a kapcsolat súlyával arányos

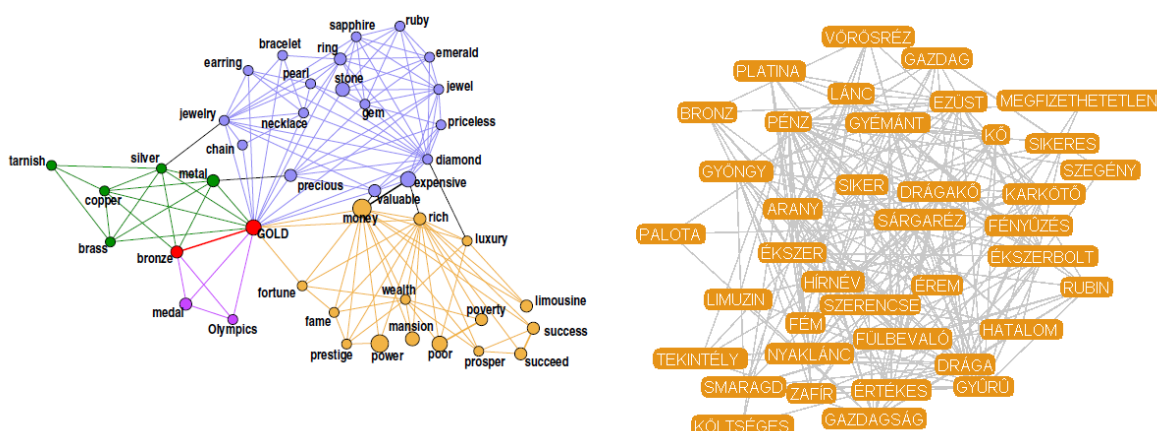
6.1. A CFinder használatával kapott eredmények

Áttérve a CFinder által detektált közösségekre, ha megnézzük a program által a $k=4$ esetén talált klasztereket, érdekes dolgot látunk: míg az angol szóasszociációs teszt adatai alapján a

²² Azt is hozzá kell tenni ehhez, hogy a *John* nem is szerepelt hívószóként a SFWAN-ban, az *apple* igen, de nem érkezett rá *gold* válasz, a *bullt* pedig egyértelműen 'bika' jelentésében értelmezték a tesztet kitöltők, és minden válasz ehhez kapcsolódik (*cow, horn, fight* stb.).

²³ A vizualizációhoz a Cytoscape programot használtam.

gold körül kirajzolódik négy csoportosulás (olimpia, fémek, ékszerek, jólét), addig a magyar adatokban $k=4$ mellett a CFinder egyetlen „óriásklasztert” mutat ki.



9. ábra: A CFinder által a *gold* és az *arany* körül kimutatott klaszter(ek), $k=4$ esetén. Az angol ábrán négy közösség, az olimpiához, a fémekhez, az ékszerekhez és a jóléthez kapcsolódó fogalmak mutathatók ki. A magyar adatokból a k ezen megválasztása esetén nem rajzolódnak ki klaszterek. Ennél az ábránál csupán azokat az asszociációkat vettem figyelembe, ahol minden válasz hívószó is volt egyben, és minden egyes szó szerepelt az angol mintán. A bal oldali ábra forrása Palla – Derényi – Farkas – Vicsek (2005b: 8)

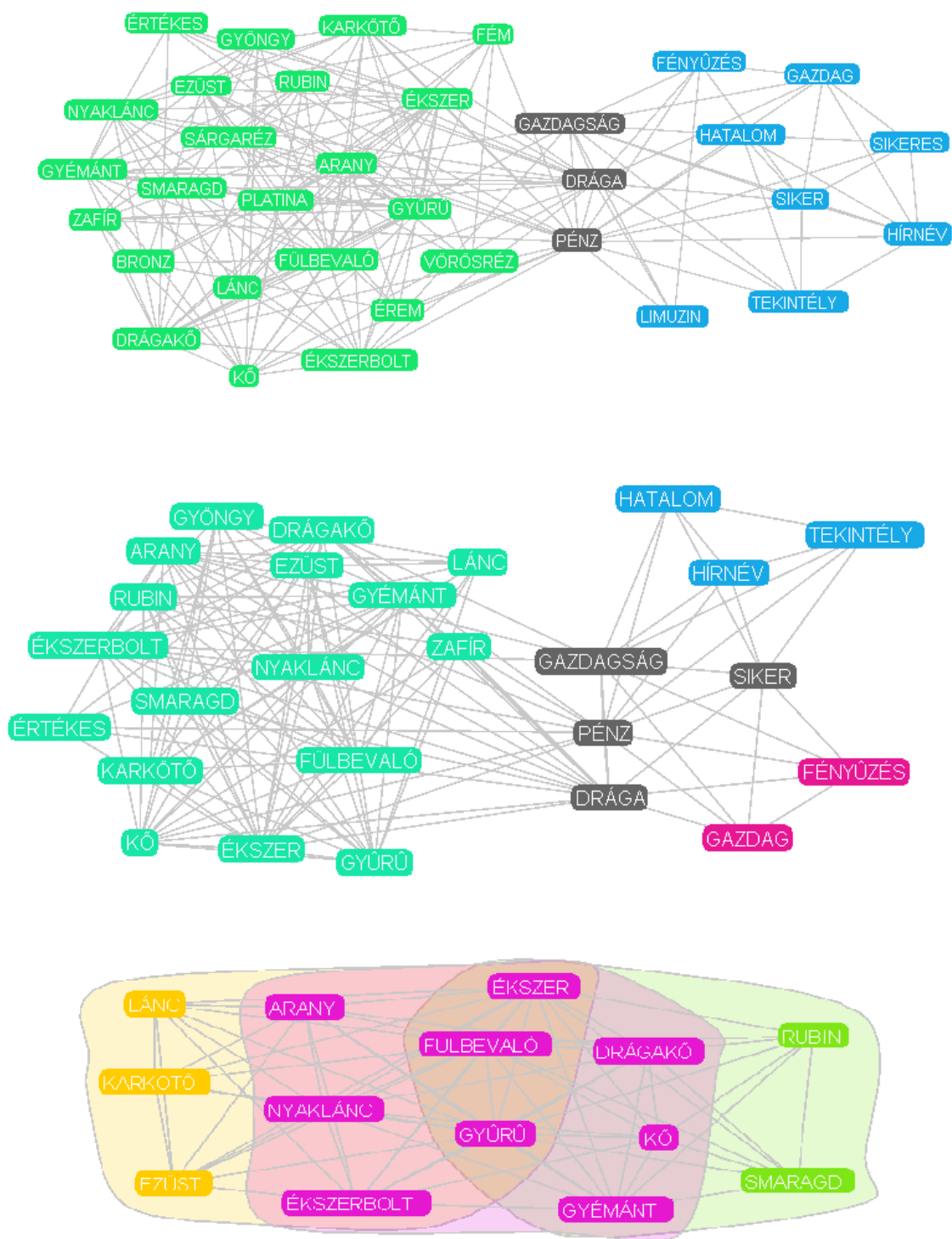
Jól látszik, hogy az *arany* környezete sokkal sűrűbb, mint a *goldé* – „minden mindennel összefügg”. Nézzünk egy példát: a *silver* a fémek-klaszter tagja (*tarnish*, *copper*, *bronze* és *metal* kapcsolatai révén), ezen kívül kapcsolatban áll az ékszerek-klaszter *jewelry* csúcsával is. Az *ezüst* ezzel szemben – a *gold* körüli klaszterek neveinél maradva – kapcsolatai alapján része (lenne) a fémek-klaszternek (*fém*, *platina*), az ékszerek-klaszternek (*fülbevaló*, *gyűrű*, *karkötő*, *lánc*, *nyaklánc*, *ékszer*, *ékszerbolt*) és az olimpia-klaszternek is (*érem*). Vagy kiváló példa még a *drágakő* (*gem*), amelynek az angol megfelelője csak az ékszerek-klaszter tagja, magyarul viszont vannak kapcsolatai a jólét-klaszterben (*gazdagság*, *pénz*) és a fémek-klaszterben is (*fém*). Látható tehát, hogy 4-tagú klikkek „görgetésével” a magyar asszociációkban nem találunk olyan csoportosulásokat, amelyek egymással szorosabb kapcsolatban állnának, mint a hálózat többi részével.

Ha a k -t 5-re, 6-ra vagy 8-ra állítjuk már több, kisebb közösséget talál a CFinder a hálózatban (8. ábra). A „legszebb” az 5-elemű klikkeket tartalmazó közösségeket ábrázoló legfelső rajz: két, jelentésében is jól körülhatárolható csoportot látunk. Zöld színnel az ékszerek, és ehhez kapcsolódóan a fémekhez kötődő fogalomkör elemei, kék színnel a jóléthez köthető szavak láthatók. A szürke szavak, a *gazdagság*, a *drága* és a *pénz* jelentik az átfedést: ezek mindkét klaszternek elemei. $K = 5$ esetén tehát kapunk két klasztert, melyek valóban értelmesek. Ha a k értékét 6-ra állítjuk, akkor az történik, hogy a jóléttel kapcsolatos

fogalmak eddig koherens csoportja kettészakad, és bár a *gazdagság* és a *siker* mindkét klaszternek tagja marad, a kék és a lila szín jelzi a különbségüket. Némi erőfeszítéssel lehet magyarázatot adni a két csoportosulásra – az egyik a gazdagság anyagi (*pénz, fényűzés, drága*) oldalát, a másik a nem annyira kézzelfogható oldalát (*hatalom, hírnév, tekintély*) világítja meg. Azonban a $k = 8$ esetén látható klaszterekre semmilyen értelmes magyarázat sincs: miért része az egyiknek a *karkötő* és a *nyaklánc* (sárga háttérű), ha a másiknak nem (zöld háttérű), holott a *fülbevaló*, a *gyűrű* és az *ékszer* közös? Nyilvánvalóan a k ilyen megválasztása után minimum 8 tagú, egymással nagyon szoros kapcsolatban álló (hiszen 8-8 tagonként minden csúcs minden csúcscsal összeköttetésben kell álljon) csoportokat kapunk, ami nem biztos, hogy helyesen modellezi a mentális lexikont. Az utolsó két példát azért hoztam, hogy a CFinder lehetőségeinek bemutatása mellett kitérjek a korlátaira is. Ami azonban feltűnik, akár a $k = 4$, akár a $k = 5$ esetét nézzük, az az olimpia-klaszter teljes hiánya. Míg a SFWAN anyagán egy picit, de jól behatárolható csoportosulás jelenti az olimpia, elsősorban az olimpiai érmek körét, addig ez a magyar anyagon nem mutatható ki.

Az olimpia fogalomkörbe tartozó szavak csoportosulásának hiányát – ebben az adathalmazban – magyarázhatja a *medal* és az *érem* szavak asszociációinak különbsége. A SFWAN anyagában jól elkülönül egymástól a *medal* és a *coin*: az előbbire adott válaszok alapján (*gold, award, honor, prize, win, winner*) az egyértelműen 'érem mint jutalom, díj', míg az utóbbi (asszociációi: *money, penny, nickel, dollar, dime, toss*) 'érme mint pénz' jelentéssel bír. Ezzel szemben a magyar vizsgálatban az *érem* szóra beérkezett asszociációk a következők: *arany, aranyérem, elismerés, győzelem, helyezés, medál, olimpia, siker, sport, verseny*, illetve *2 oldal* (vagy *két oldal*), *érme, oldal, pénz, pénzérme* és *kérem*. Jól látszik, hogy a magyar *érem* szónak az olimpiához kapcsolódó fogalmakon kívül az 'érme mint pénz' jelentéshez köthető szavakkal is vannak kapcsolatai. Az olimpia-klasztert alkotó többi kapcsolat közül az *olimpia – bronz* és az *olimpia – arany* asszociáció a magyarban hiányzik.

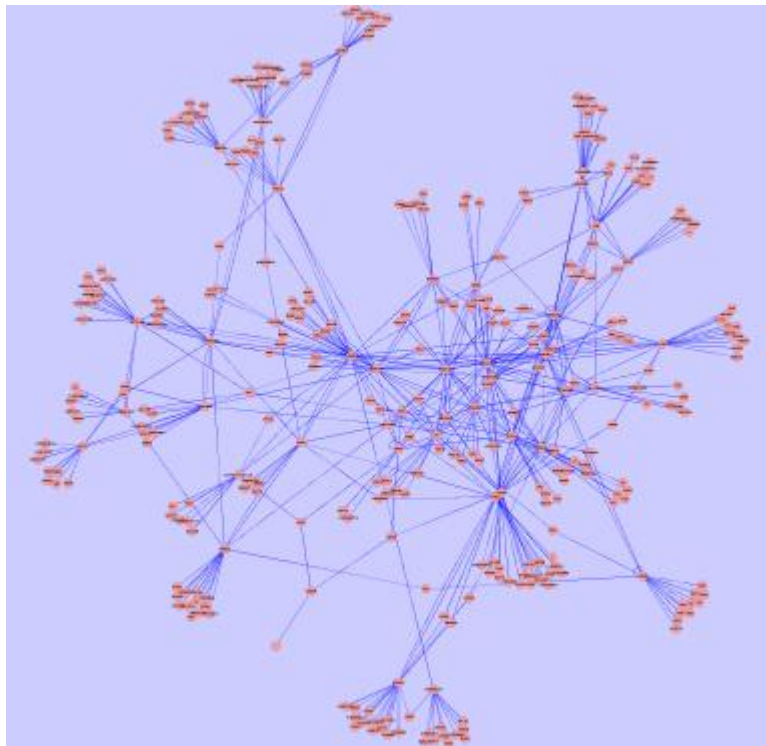
Az eddigieket összefoglalva megállapíthatjuk, hogy a két vizsgálat eredménye közt két nagy különbség van: az egyik a *gold*, illetve az *arany* szóra adott válaszok eltérése, amely egyrészt kulturális okokkal magyarázható (*János, bulla, alma*), másrészt néhány szópár (angol-magyar megfelelők) eltérő jelentéseivel (*bar – rúd*). Ez utóbbi az egyik oka a másik nagy különbségnek: az olimpia-klaszter teljes egészében hiányzik a magyar nyelvű hálózatban (*medal – érem* különbsége), míg a másik három csoportosulást a k értékétől függően vagy egy nagy klaszterként, vagy két klaszterbe összevonva mutatja ki a CFinder.



10. ábra: A CFinder által talált klaszterek $k=5$ (legfelül), $k=6$ (középen) és $k=8$ (legalul) esetén. Az adatfájl csak a hívószóként is szereplő, és az angol minta fordításából nyert asszociációkat tartalmazta

Hogyha nem csak azokat az asszociációkat adom meg, amelyek egyben hívószók is voltak, akkor a CFinder már nem bír el a bemeneti fájjal. A Cytoscape nevű, hálózatok

elemzésére és vizualizálására használható program (Smoot és mtsai 2011) a 11. ábrán látható hálózattal tér vissza. Jól látszik, melyek azok a csúcsok, melyek nem voltak hívószók a vizsgálatban: ezek adják a fürtöket a hálózat peremén.



11. ábra: Minden asszociációt tartalmazó hálózat, a Cytoscape programmal vizualizálva

6.2. Az élklasszterező eljárással kapott eredmények és a két algoritmus teljesítményének összehasonlítása

A következőkben az élklasszterező algoritmus által kapott eredményeket mutatom be. A magyar adathalmazon futtatva a következő 37 „közösséggel” tér vissza:²⁴

- 1: pénz, drága, költséges
- 2: gazdagság, szegénység, fényűzés
- 3: arany, János
- 4: bronz, érem

²⁴ A listát az alapján a kimeneti fájl alapján készítettem, amely a közösségeket a bennük lévő éleket alkotó csúcsok listájaként tünteti fel. Ezt úgy kell értelmezni, hogy például a 4. klasztert a *bronz – érem* él alkotja, tehát ennél a *bronz* és az *érem* a felsorolandó csúcsok. Az 5. közösségnél a *fém – karkötő*, *ezüst – fém*, *fém – gyűrű*, *arany – fém*, *gyűrű – sárgaréz*, *drágakő – fém*, *arany – sárgaréz* és a *karkötő – sárgaréz* élek az építőelemek, így a felsorolt hét csúcs az, amely tagja ennek a klaszternek. Természetesen az élklasszterezés alapvetésére gondolva a csúcsok helyett az élek feltüntetése lenne a megfelelő, de jelen esetben az egyszerű szemléltetés igényét ez a lista is kielégíti.

- 5: gyűrű, arany, sárgaréz, ezüst, karkötő, drágakő, fém
- 6: szegénység, megfizethetetlen
- 7: ékszer, gyűrű, platina, arany, gyémánt, kő, ezüst
- 8: pénz, érem
- 9: platina, fém
- 10: ékszer, arany, bronz, fülbevaló, ezüst
- 11: arany, költséges
- 12: olimpia, érem
- 13: siker, érem
- 14: ékszer, lánc, drága, nyaklánc, kő, gyémánt, zafír, rubin, ékszerbolt, pénz, gyöngy, gyűrű, arany, smaragd, fülbevaló, karkötő, drágakő, ezüst
- 15: bronz, vörösréz
- 16: ékszer, érem, arany, ezüst
- 17: limuzin, szerencse, hírnév, gazdag, siker, tekintély, gazdagság, sikeres, fényűzés, hatalom
- 18: bronz, fém
- 19: ékszer, pénz, értékes, drága, arany, gyémánt, drágakő
- 20: olimpia, siker
- 21: bronz, zafír
- 22: palota, hatalom
- 23: palota, zafír
- 24: ékszer, arany, vörösréz, kő, karkötő
- 25: pénz, sárgaréz, fém
- 26: gazdagság, sárgaréz
- 27: gazdag, palota, fényűzés, gazdagság
- 28: gazdagság, platina
- 29: gazdag, szegény
- 30: pénz, drága, megfizethetetlen
- 31: sárgaréz, vörösréz, fém
- 32: érem, platina
- 33: értékes, megfizethetetlen
- 34: limuzin, szerencse, hírnév, drága, gazdag, siker, tekintély, ékszerbolt, gazdagság, pénz, rubin, arany, sikeres, fényűzés, drágakő, hatalom
- 35: gazdag, gyémánt

36: sárgaréz, fém

37: pénz, szegény, drága

Ha végignézzük a 37 soron, értelmet nyer a lista előtti mondatban a közösség szó körül használt idézőjel. A klaszterek, bármelyik definícióval is szimpatizálunk, olyan részei a hálózatnak, ahol a hálózat valamely építőelemei (a csúcsok vagy az élek) egymással szorosabb kapcsolatban vannak, mint a hálózat többi részével. Ezt szem előtt tartva nehéz elfogadni, hogy van külön egy *gazdag*, *szegény* és egy *pénz*, *szegény*, *drága* klaszter (amelyek tulajdonképpen a *gazdag* – *szegény*, illetve a *pénz* – *szegény* és a *szegény* – *drága* élek klaszterei), illetve van számos olyan közösség, amelynek mindössze egy (él, két csúcs) tagja van. Az élklaszterezés megálmodói hangsúlyozták, hogy fontos a nagyon részletes dendrogramot valahol elvágni annak érdekében, hogy értelmes csoportokat kapjunk (Ahn – Bagrow – Lehmann 2010a: 761). Azonban a felhasználó csak nehezen tud beleszólni az eloszlási sűrűség függvényének működésébe, nehezen tudja befolyásolni azt, hogy hol legyen a határ, ami fölött a klaszterekről beszélünk. Márpedig elnézve ezt a 37 közösséget, egyetlen megoldásként az merül fel, hogy még följebb „vágjuk el” a képzeletbeli fát, és így talán van lehetőség összevonni az apró, 1-2 tagú klasztereket.

Tovább folytatva az élklaszterezéssel kapott eredményeknek a CFinder teljesítményével való összehasonlítását megállapíthatjuk, hogy az olimpia témakörébe tartozó fogalmak eszerint az elemzés szerint sem alkotnak szorosan kapcsolódó közösséget. Van ugyan *bronz*, *érem*; *olimpia*, *érem*; *siker*, *érem* és *olimpia*, *siker* „klaszter”, azonban ezek a fenti elemzés alapján inkább csak egy bizonyos gyakoriságot meghaladóan jelen lévő élek. A 16. sorszámú közösség (*ékszer*, *érem*, *arany*, *ezüst*) már kicsit hasonlít az amerikai vizsgálatban látható „olimpia-klaszterre” (*olimpia*, *érem*, *bronz*, *arany*), azonban az *ékszer* helye nem magyarázható. Kérdés, hogy az elosztási sűrűség függvényének valamilyen módosításával (a határvonal „feljebb húzásával”) el lehetne-e érni, hogy ezek az élek egy klaszterbe kerüljenek. Ezt azonban a már említett korlátok miatt nem tudom tesztelni.

A 14., a 17. és a 34. közösség viszont valóban klaszternek tekinthető, méghozzá olyannak, amit el is tudunk képzelni a mentális lexikonban. A 14.-ben lévő szavak jelentik az ékszerek, drágakövek csoportját, a 17.-ben találjuk a gazdagsághoz, hatalomhoz kapcsolódóakat, a 34. pedig a kettőnek egyfajta egyvelege – az átfedések kérdésére még később visszatérek.

Fontos különbség a két módszer által kapott eredményekben, hogy az élklaszterezés esetén megjelent az *arany* és a *János* mint egy közösség. Igaz ugyan, hogy ez lényegében

csak egy él, amely a jelen vizsgálatban nem is tud más élekkel/csúcsokkal értelmes, intuíciónknak megfelelő klasztert alkotni²⁵, ráadásul kérdés, hogy az eloszlási sűrűség függvényének esetleges megváltoztatása után (ami a korábbi elemzésből fakadóan szükséges lehet) is megmarad-e, vagy pedig egy több élt tartalmazó közösség tagja lesz.

Mivel nincs elérhető eszköz az élklasszterező algoritmus eredményeinek vizualizációjára, ezért nehezebb az átfedéseket vizsgálni, mint a CFinder esetében. A fentihez hasonló listával igyekszem szemléltetni egyes csúcsok több csoportba tartozását. Először lássuk azokat a szavakat, amelyek a CFinder által, $k=5$, illetve $k=6$ esetén detektált klaszterekben az átfedésekért feleltek (a szavak utáni számok az élklasszterezés utáni közösségek közül azoknak a számai, amelyekben az adott szó szerepel):

1. *gazdagság* ($k=5$): 2., 17., 26., 27., 28., 34. – vegyesen vannak ezek közt a jóléttel (2., 17., 27.) és egyéb, fémekkel (26.: *gazdagság*, *sárgaréz*; 28.: *gazdagság*, *platina*) vagy drágakövekkel kapcsolatos klaszterek (a 34., korábban már említett „vegyes” klaszter)
2. *drága* ($k=5$): 1., 14., 19., 30., 34., 37. – jóléttel, anyagi kérdésekkel (1., 30., 37. és a 34.), ékszerekkel, drágakövekkel kapcsolatos klaszterek (14., 19. és 34.)
3. *pénz* ($k=5$): 1., 8., 14., 19., 25., 30., 34., 37. – jóléttel (1., 30., 37. és a 34.), fémekkel (25.), ékszerekkel, drágakövekkel (14., 19.) kapcsolatos klaszterek és a korábbi négy fogalomkör egyikébe se sorolható 8. klaszter (*pénz*, *érem*).
4. *siker* ($k=6$): 13., 17., 20., 34. – olimpiával (13., 20.), jóléttel (17.) kapcsolatos klaszterek és a 34. „vegyes” klaszter

A fentiek azt mutatják, hogy a CFinder segítségével talált, egy-egy fogalomkörnek nagyjából megfeleltethető klaszterek esetén²⁶ az átfedésért felelős szavak az élklasszterezési eljárás során is több közösségnek, több fogalomkörhöz tartozó élcsoportnak a tagjai, tehát ez esetben is átfedést biztosítanak. Az általuk így módon összekapcsolt közösségek is körülbelül ugyanazokat a fogalomköröket fedik le mindkét klaszterezés esetében: egyrészt a jólét és gazdagság szavait, másrészt a drágakövek, ékszerek szavait képviselik. Újdonság az olimpia jelenléte a *siker* esetén, illetve a fémek témakörének egy-két képviselője.

Ha megnézzük a korábban kiemelt 14. és 17. klasztert, mint egységes, egy-egy jól definiálható fogalomkör (az ékszerek és drágakövek, illetve a jólét, gazdagság) szavait összefoglaló csoportokat, azt látjuk, hogy két diszjunkt közösségről van szó: nincsen közös

²⁵ Hiszen az *Arany Jánossal* kapcsolatos fogalomkörnek más szavai nem szerepeltek hívószóként a kérdőíven.

²⁶ Tehát a k -t 5-re vagy 6-ra állítva.

tagjuk.²⁷ Ezt a képet árnyalja egy kicsit a 34., korábban már „vegyesnek” aposztrofált lista: ebben szerepelnek olyan szavak, amelyek a 14.-nek (*ékszerbolt, rubin, arany* stb.), és olyanok, amelyek a 17.-nek (*limuzin, szerencse, hatalom* stb.) is tagjai. Ez a klaszter mintha azt próbálná kifejezni, hogy a másik két közösség tagjai között azért van valamilyen kapcsolat. Persze önmagában a 34. klaszter nehezen elképzelhető egység a mentális lexikonban (miért lenne egy szorosabb közösség tagja a *tekintély* és a *rubin*?), így ebben is az algoritmus korlátait láthatjuk inkább.

Összefoglalva az eddigieket megállapíthatjuk, hogy az élklaszterezés erőssége a CFinderrel szemben egyértelműen az, hogy a CFinder által nem is ábrázolt, ottani klaszterekbe nem tartozó csúcsok itt megjelennek²⁸ mint olyan klaszterek, amelyeknek csupán egy-egy él a tagja (lásd például az *arany* – *János* asszociációt). Ugyanebből az okból kifolyólag bukkannak fel az olimpia témaköréhez tartozó szavak is (*olimpia* – *siker, bronz* – *érem*). Viszont az is megállapítható, hogy a CFinder, elsősorban $k=5$ esetén értelmes, a mentális lexikonban is elképzelhető közösségek képével tér vissza; az élklaszterezéssel, egy-két olyan listától eltekintve, amelynek szavai valóban egy-egy fogalomkör elemei (14., 17. klaszter) számos apró, többnyire több fogalomkör szavait egyszerre tartalmazó csoportot kapunk. Nagyon jellemző az eredményül kapott klaszterek száma: míg a CFinder 1 ($k=4$), 2 ($k=5$) vagy 3 ($k=6$), addig az élklaszterezés 37 közösséget detektál ugyanazon a hálózaton. Természetesen nem szabad elfelejteni, hogy a hierarchikus élklaszterezés egyik fontos szerepe, hogy a hálózatok hierarchiáját és az átfedéseket egyszerre ábrázolja. Így talán az, ami most a CFinderrel szemben hátránya, előnyére válhatna, ha annak a bizonyos küszöbnek a módosításával feljebb tudnánk lépni a hierarchiában: a több kis közösség talán összekapcsolódna egy nagyobbá, értelmet adva ezeknek az apró csoportoknak, felfedve a mentális lexikon hálózatának hierarchiáját. Így azonban, hogy nem tudunk változtatni az eloszlási függvényen, ez csak egy feltételezés marad.

Ez utóbbi különbség, azaz a klaszterek számának jelentős eltérése jelenik meg akkor is, hogyha az eredeti amerikai szóasszociációs vizsgálat eredményein futtatjuk az élklaszterező algoritmust.²⁹ A CFinder ezen a hívószó-válaszszó kapcsolatokat tartalmazó adatbázison detektálta a négy klasztert, amelynek ábrájából vizsgálatom indult, és amely négy klaszter a négy fogalomkörnek (*olimpia, fémek, ékszerek, jólét*) szépen megfeleltethető. Az

²⁷ Emlékeztetőül: a CFinder eredményeiben a két klaszter közös tagja volt a *gazdagság*, a *drága* és a *pénz*.

²⁸ Éppen ez volt az egyik fontos kritika a CFinderrel szemben: ha egy csúcs nem tagja egy klaszternek sem, nem ábrázolja a program (Fortunato 2010: 135).

²⁹ Természetesen egy jóval szűkebb halmazról van szó, mint a teljes South Florida Word Association Norms: azokat a kapcsolatokat szűrtem ki, amelyeknek egyik oldalán (tehát akár hívó- akár válaszszóként) szerepelt az a 41 szó, amely a kiinduló ábrán (Palla és munkatársai (2005b)-ben) megjelent.

élklaszterezéssel 130 közösség listáját kapjuk, bennük számtalan olyan szóval, amelyek a CFinder klasztereiben nem szerepelnek. Ennek oka feltehetően az, ami már a magyar eredmények közti különbségek egy részét is okozta: a CFinder csak azokat a csúcsoakat jeleníti meg, amelyek szorosan kapcsolódnak egy klaszterbe, míg az élklaszterezés nem „hagy le” egy csúcsot sem. Véleményem szerint a 130 klaszter felsorolásának és elemzésének ennek a dolgozatnak a keretein belül nincs jelentősége: egyrészt a két algoritmus működésének különbségéből fakadóan valószínűleg ugyanazokra a következtetésekre jutnánk, mint a magyar adatokról kapott eredmények összehasonlításánál, másrészt 130 közösséget értelmesen kezelni nagyon nehéz vizualizáció nélkül, ahogyan nehéz az átfedéseket, hierarchiát felfedni is.

A bemenet kisebb módosításával (csak azokat a kapcsolatokat (éleket) nézzük, amelyeknek mindkét tagja szerepel Palla és munkatársai (2005b)-ben) az élklaszterezéssel talált csoportosulások száma lecsökken 18-ra. Ezek a következők:

1. valuable, priceless, precious, expensive
2. copper, tarnish, brass, metal, silver, bronze
3. necklace, gold, jewelry, chain
4. copper, jewelry, gold, metal, olympics, brass, medal, silver, bronze
5. prosper, succeed, success
6. bracelet, watch
7. stone, diamond, jewelry, gold, jewel, valuable, gem, priceless, precious, expensive
8. stone, diamond, jewelry, gold, pearl, bracelet, jewel, earring, sapphire, emerald, necklace, ring, ruby, gem
9. money, valuable, expensive, rich
10. poor, wealth, power, money, poverty, luxury, rich, mansion, prestige, limousine, fame
11. precious, metal
12. diamond, luxury
13. limousine, luxury
14. money, gold, rich
15. medal, olympics, bronze
16. prosper, succeed, success, rich, money, wealth
17. money, value
18. expensive, value

Végignézve a listán rögtön felfedezhetünk néhány “ismerős” klasztert: a CFinder-ábra fémek-klaszterét (2. közösség), a jólét-klasztert (10. közösség), az ékszerek klaszterét (8. közösség) és még az olimpia klaszterét is megtalálhatjuk (15. közösség). Ezek nem tökéletes megfelelések, de a párhuzamba állított közösségek csak néhány csúcsban különböznek. Ez azért jelentős, mert így látszik, hogy a legnagyobb különbséget a két algoritmus működésében a már említett, közösségekbe nem szorosan tartozó csúcsok elhagyása vagy nem elhagyása jelenti: ha a „teljes”³⁰ hálózaton vizsgálódik, az élklaszterező eljárás sokkal több mindent fog figyelembe venni, mint a CFinder, ezért nagyobb lesz a különbség. Némi módosítással azonban (csak olyan éleket vizsgálunk, amelyeknek mindkét tagja szerepel a kiinduló ábrán) a két eredmény összehasonlíthatóvá válik, és ez esetben nem olyan nagy az eltérés.

7. Összegzés

Dolgozatomban azt vizsgáltam, hogy vajon az angol anyanyelvű beszélők mentális lexikonában a *gold* körüli klaszterek, melyek egy-egy fogalomkörnek feleltethetők meg, megtalálhatóak-e a magyar anyanyelvű beszélők mentális lexikonában, az *arany* szó körül, és vajon a két hálózat tényleg ugyanolyan-e, ahogyan azt a Derényi et al. (2006)-ban található ábra sugallja. Mivel az angol nyelvű ábra (Palla – Derényi – Farkas – Vicsek 2005b) forrását jelentő adatbázishoz hasonló nem állt rendelkezésemre, először egy saját szóasszociációs vizsgálatot kellett végezni. A 100 emberrel elvégzett teszt alapján egyértelműen kiderült, hogy van különbség a két nyelv beszélőinek mentális lexikonában, a *gold*, illetve az *arany* körüli hálózatban. A különbség egyrészt magára a *goldra* és az *aranyra* adott válaszokban található, másrészt a klaszterek hiányában (pl. a „magyar olimpia-klaszter” teljes hiánya), különbözőségében. Az eltérések oka egyrészt kulturális - ez látható abban, hogy az *arany* - *János*, *arany* - *bulla* és az *arany* - *alma* asszociációknak nincs angol párja. Másrészt különbséget okoz egy-egy szópár eltérő jelentése: erre példa a *bar* és a *rúd* esete, ahol az előbbi feltehetőleg azért nem hívta elő a *goldot* az amerikai vizsgálatban, mert a válaszadók a szó ’bár’ jelentésére asszociáltak, nem pedig a ’rúd’-ra. Ezen kívül két különböző algoritmus mást talál ugyanazon az adathalmazon, ami megmagyarázható a két eljárás eltérő működési mechanizmusával, és világosan mutatja, hogy nem fogadhatjuk el feltétel nélkül a mentális lexikon abszolút képének egy algoritmus eredményeit sem.

³⁰ Azért kell zárójelbe tenni, hogy teljes a hálózat, mert valójában nem az egész adatbázison dolgozott az algoritmus, csupán azokon az éleken, amelyek tartalmazták az eredeti angol ábrán látható szavakat – a teljesség arra vonatkozik, hogy más szavak is részei a vizsgálatnak, mint egy-egy él másik elemei.

Azt gondolom, hogy egy, a South Florida Free Association Normshoz hasonló méretű és felépítésű adatbázisnak és több algoritmus kipróbálásának a segítségével átfogóbb képet kaphatnánk arról, milyen különbségek lehetnek két nyelv beszélőinek a mentális lexikonában.

8. Bibliográfia

- Ahn, Yong-Yeol – Bagrow, James P. – Lehmann, Sune. 2010a. Link communities reveal multi-scale complexity in networks. *Nature* 761–764.
- Ahn, Yong-Yeol – Bagrow, James P. – Lehmann, Sune. 2010a. Link communities reveal multi-scale complexity in networks. Supplementary information.
- Aitchison, Jean 2003. *Words in the Mind: An Introduction to the Mental Lexicon*. Basil Blackwell. Oxford.
- Altieri, Nicholas – Gruenenfelder, Thomas – Pisoni, David B. 2010. Clustering coefficients of lexical neighborhoods: Does neighborhood structure matter in spoken word recognition? *The Mental Lexicon* 1–21.
- Barabási Albert-László 2006. A hálózatok tudománya: a társadalomtól a webig. *Magyar Tudomány* 1298–1308.
- Barabási Albert-László 2008. *Behálózva*. Helikon Kiadó. Budapest.
- Clauset, Aaron – Moore, Christopher – Newman, M. E. J. 2008. Hierarchical structure and the prediction of missing links in networks. *Nature* 98–101.
- De Deyne, Simon – Storms, Gert 2008a. Word associations: Norms for 1,424 Dutch words in a continuous task. *Behavior Research Methods* 198–205.
- De Deyne, Simon – Storms, Gert 2008b. Word associations: Network and semantic properties. *Behavior Research Methods* 213–231.
- Derényi Imre – Farkas Illés – Palla Gergely – Vicsek Tamás 2006. Csoportosulások szociológiai, technológiai és biológiai hálózatokban. *Magyar Tudomány* 1319–1324.
- Divjak, Dagmar S. – Gries, Stefan Th. 2008. Clusters in the mind?: Converging evidence from near synonymy in Russian. *The Mental Lexicon* 188–213.
- Dorogovtsev, Sergey N. – Mendes, José Fernando F. – Samukhin, Alexander N. 2001. Giant strongly connected component of directed networks. *Physical Review E* 64: 1–5.
- Ferrer i Cancho, Ramon 2005. The structure of syntactic dependency networks: insights from recent advances in network theory. *Problems of Quantitative Linguistics* 60–75.
- Fortunato, Santo 2010. Community detection in graphs. *Physics Report* 486: 75–174.
- Gósy Mária 1999. *Pszicholingvisztika*. Corvina. Budapest.
- Kovács László 2011. *Fogalmi rendszerek és lexikai hálózatok a mentális lexikonban*. Tinta Könyvkiadó. Budapest.
- Lengyel Zsolt 2009: Magyar Asszociációs Normák Enciklopédiája: új perspektívák. In: Lengyel Zsolt – Navracsics Judit (főszerk.): *Tanulmányok a mentális lexikonról*.

Nyelvelsajátítás - beszédprodukció – beszédpercepció. Tinta Kiadó. Budapest. 15–20.

Nelson, Douglas L. – McEvoy, Cathy L. – Schreiber, Thomas A. 1998. *The University of South Florida word association, rhyme, and word fragment norms.*
<http://www.usf.edu/FreeAssociation/>.

Palla Gergely – Derényi Imre – Farkas Illés – Vicsek Tamás 2005a. Uncovering the overlapping community structure of complex networks in nature and society. *Nature* 814–818.

Palla Gergely – Derényi Imre – Farkas Illés – Vicsek Tamás 2005b. Uncovering the overlapping community structure of complex networks in nature and society. Supplementary Information.

<http://www.nature.com/nature/journal/v435/n7043/extref/nature03607-s1.pdf>

Pléh Csaba – Lukács Ágnes – Kas Bence 2008. A szótár pszicholingvisztikája. In: Kiefer Ferenc (szerk.): *Strukturális magyar nyelvtan 4.: A szótár szerkezete.* Akadémiai Kiadó. Budapest. 789–852.

Sigman, Mariano – Cecchi, Guillermo A. 2002. Global organization of the Wordnet lexicon. *Proceedings of the National Academy of Sciences of the United States of America* 1742–1747.

Smoot, Michael – Ono, Keiichiro – Ruscheinski, Johannes – Wang, Peng-Liang – Ideker, Trey 2011. Cytoscape 2.8: new features for data integration and network visualization. *Bioinformatics* 431–432.

Steyvers, Mark – Tenenbaum, Joshua B. 2005. The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science* 41–78.

Wang, Hui-li – Beng, Wen-duo – Hou, Yan 2010. A Comparison Study on Word Association Between English Native Speakers and Chinese English Learners. *Canadian Social Science* 6: 45–60.