

KOVÁCS LÁSZLÓ

Nyugat-magyarországi Egyetem

klaszlo@btk.nyme.hu**Nyelvi hálózatok a mentális lexikonban – Agykapocs-kutatás**

The investigation of network structures has come to the centre of attention in several fields of scientific research. Although linguists have been showing a growing interest in this subject as well, only few projects come up with empirical data which prove that network properties can be equally found in language. This article introduces a project which aims at collecting word associations in several languages from Internet-users. The database taking shape in this way can be used for mental lexicon research, furthermore some features of the software developed for this purpose can be used to prove the existence of, and to examine the network structure of the mental lexicon. The paper sets out from the description of the project, elaborates on the problems encountered in the first phase of its development, and finally demonstrates that it is possible to collect reliable linguistic data (word associations) via the Internet.

A mentális lexikon hálózatos szerveződésével kapcsolatban eddig kevés kutatás ismert (Fóris, 2007; Kovács, 2007; Kovács, 2008; Steyvers & Tenenbaum, 2005; Vitevitch, 2008; Lengyel, 2007b), a nyelvi hálózatok kutatásának és elemzésének módszertana kidolgozásra vár. Az általános hálózatelmélettel foglalkozó kutatások eredményeiből, módszereiből kiindulva alkothatunk meg olyan eszközöket és eljárásokat, amelyek segítséget nyújtanak a nyelven belüli hálózatok minél pontosabb feltérképezésére. A különböző diszciplínák hálózatokkal kapcsolatos kutatásai (Barabási, 2003; Csermely, 2005; Strogatz, 2003; Watts, 2004) egyértelműen rámutatnak, hogy hálózatos jelenségek (pl. kis-világ jelenség) csak nagy mennyiségű adat segítségével mutathatók ki. A nagy mennyiségű adat feldolgozása számítógépek felhasználásával felgyorsítható. Az informatikai eszközök nyelvészeti célú felhasználása természetesen ma már széles körben elterjedt, az adatok gyűjtése ugyanakkor az esetek többségében ma is többnyire hagyományos úton történik.

Az informatika és nyelvészet viszonylag új kapcsolódási pontja a fogalmi rendszerek automatikus feltérképezésére kifejlesztett szoftverek megjelenése (Auger & Barrière, 2008; Halskov & Barrière, 2008; Aussenac-Gilles & Jacques, 2008; Sierra, *et al.*, 2008; Soler & Alcina, 2008; Marshman, 2008).

Internet-felhasználók segítségével végzett kutatások

A gazdasági versenyszférában területükön vezető szerepet betöltő (főként kozmetikai illetve gyógyszeripari, genetikai kutatásokkal foglalkozó) cégek ma versenyképességük megőrzéséhez a hagyományos kutatás mellett felhasználják az internet-felhasználók tudását is: aktuális K+F projektjüket, megoldásra váró problémáikat az internet segítségével (akár erre szakosodott portálokon, pl. InnoCentive) közzéteszik. A kutatásban és a probléma megoldásában bárki részt vehet; sikeres együttműködés esetén a cég előre meghatározott összeget fizet. Az ilyen jellegű, az utóbbi 5-10 évben elterjedt megoldások a megbízást adó cég számára előnyösek, ugyanis nem csak saját alkalmazásban lévő kutatóik dolgoznak egy adott problémán (rendkívül magas bér és költségvonzattal), hanem a világ bármely részén élő (akár „amatőr”, pl. még egyetemi hallgató) kutatók potenciális megoldásai is a cég rendelkezésére állnak. A most kialakulóban lévő megoldások hosszú távon a K+F minden területét megváltoztathatják (Szabó és Hámori, 2006; Tapscott & Williams, 2007).

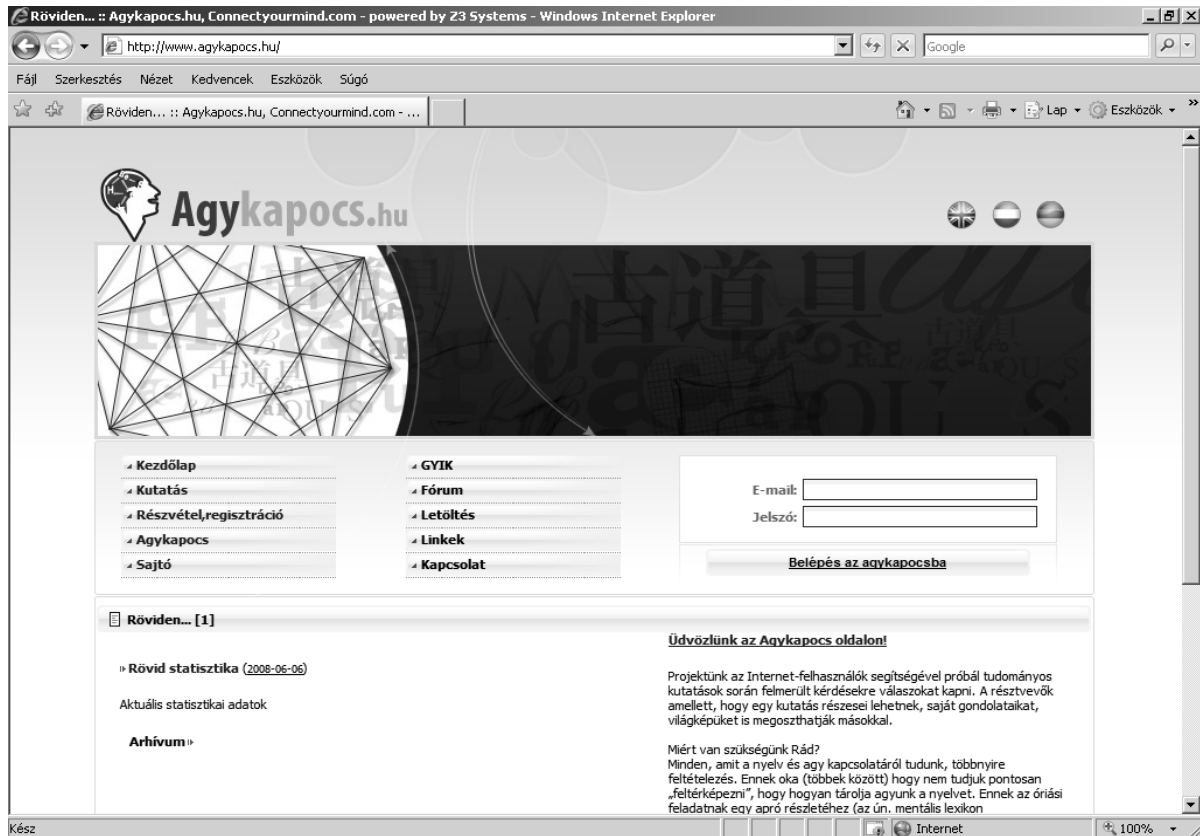
Egyre gyakoribb, hogy kutatók is az internet-felhasználók segítségét veszik igénybe kutatási projektjükhöz. Több tudományos (főleg szociológiai jellegű) kutatás illetve piacorientált marketingkutatás használ fel interneten keresztül kitöltött kérdőíveket, on-line felméréseket. Az ilyen jellegű felmérések előnye, hogy a kutatók nagy mennyiségű adathoz jutnak viszonylag kis anyagi (és idő) ráfordítással, illetve olyan mennyiségű adatot képesek összegyűjteni, amely hagyományos úton nem lenne lehetséges (Schmidt, 2005).

A kifejezetten kutatási céllal kifejlesztett számítógépes programok egyik fajtája a felhasználók szabad gépidejét (a számítógép be van kapcsolva, de senki nem dolgozik rajta) használja fel sikeresen, pl. a Berkeley által üzemeltetett Seti@Home, BOINC stb. programok. A projektek nagyságrendjére jellemző, hogy pl. a BOINC kb. 350 ezer önkéntes „kutatót”, és közel 600 ezer számítógépet „foglalkoztat” (2008. június vége). A kutatásba bekapcsolódott felhasználó otthoni számítógépére egy programot letöltve, majd azt képernyőkímélőként „futtatva” nagy mennyiségű adatot képes feldolgozni. Amint a feldolgozás sikerült, a felhasználó elküldi a már feldolgozott adatokat, és új, „nyers” adatcsomagot tölthet le – az egész folyamathoz semmilyen szaktudásra nincs szükség, csak minimális informatikai ismeretekre (a program telepítése). A kutatók által írt (megrendelt) program biztosítja, hogy az adatok pontosan az elvárt formában, mélységben kerülnek elemzésre.

Az internet-felhasználók segítségét más módon veszi igénybe a NASA Stardust@home kutatása, valamint a GalaxyZoo, illetve többek között az eBird projekt. Ezen esetekben nem automatikus a számítógép által elvégzendő feladatokról van szó, ellenkezőleg: a kutatók az internet-felhasználók aktív közreműködését kérik. (Az angol nyelvben az ilyen jellegű projektekre megalkották a „citizen science”, valamint az outsourcing mintájára a „crowdsourcing” fogalmakat.) A projektek többségében a részvételhez semmilyen szakmai előképzett-

ségre nincs szükség. A GalaxyZoo projekt esetében (Lintott, *et al.*, 2008) angliai és egyesült államokbeli egyetemek kutatói a Sloan Digital Sky Survey rádióteleszkóppal távoli galaxisokról készített képek osztályozását tűzték ki célul. A projektben bárki részt vehet. Rövid tréning után (a weboldalak mindenki számára érthetően elmagyarázzák az elvégzendő feladatot, közben interaktív módon tesztelik is a leendő projektrésztvevőt) a résztvevő elkezdheti a galaxisok osztályozását a megadott szempontok szerint. A kutatók által kitűzött cél (minden galaxist legalább 20 alkalommal osztályozzanak, így biztosítva a viszonylag megbízható eredményeket) hetek alatt teljesült, így jelenleg a program továbbfejlesztésén dolgoznak GalaxyZoo 2, illetve 3 néven. A laikusoktól kapott eredmények (galaxis-besorolások) megegyeznek a szakértő csillagászok által végzett besorolásokkal (Lintott, *et al.*, 2008; Land, *et al.*, 2008; Bamford, *et al.*, 2008). A projekt első részében közel 1 millió galaxist osztályozott 100 000 önkéntes. A kapott adatok (osztályozások) száma kb. 40 millió (Lintott, *et al.*, 2008).

Hasonló jellegű, nyelvészeti adatokat gyűjtő és feldolgozó projekt számunkra jelenleg nem ismert. A nyelvi hálózatok kutatása, hipotézisek felállítása, illetve (legalább részleges) igazolása ugyanakkor elképzelhetetlen (viszonylag) nagy mennyiségű adat nélkül. Fenti projektekből, valamint kutatásaink egy kitűzött céljából (mentális lexikon hálózatos jelenségeinek megismerése, Kovács, 2007; Kovács, 2008) kiindulva jött létre a magyar Agykapocs (1. ábra) illetve angol nyelven a ConnectYourMind projekt.



1. ábra

Az Agykapocs projekt kezdőlapja

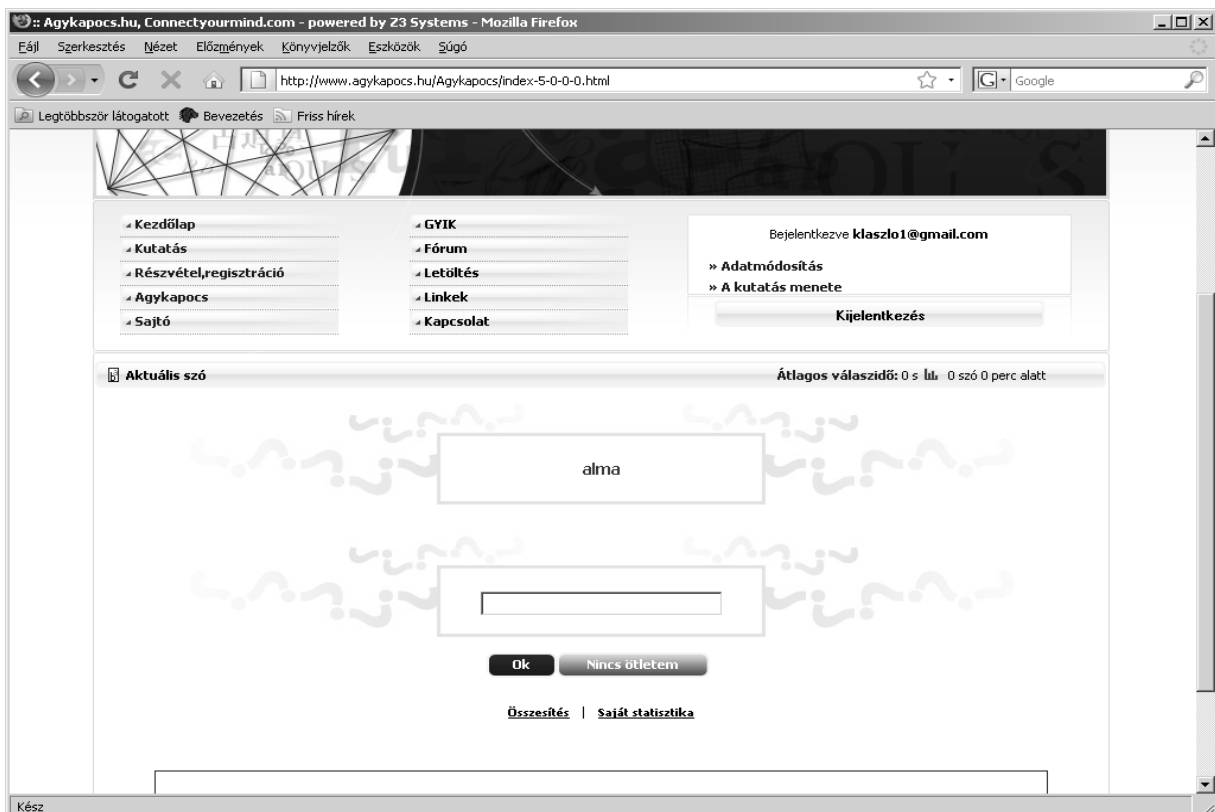
Az Agykapocs-kutatás

Az Agykapocs-kutatás célja minél több szót tartalmazó, többnyelvű szóasszociációs adatbázis kialakítása, amely adatbázisban a szavak hálózatot alkotnak. A szavak a hálózat pontjai, közöttük a kapcsolatot irányított, súlyozott gráfok jelzik (a kapcsolat irányának és erősségének megfelelően).

A projekt jelenleg az interneten elérhető, bárki részt vehet benne. Az adatbázis elektronikus formában kerül tárolásra (MySQL). A felhasználó regisztrálásánál felhasználónevet, jelszót, valamint statisztikai adatokat (kor, nem, nyelvismeret, végzettség) ad meg, amelyek segítségével szűkíthetők később a szavakra adott asszociációk (pl. milyen választ adtak a gazdaság szóra a 25-35 év közötti gazdasági egyetemest vagy főiskolát végzett nők). Regisztrálás után a felhasználó a megjelenő szavakra gépeli be az első eszébe jutó szót, illetve ha nem jut eszébe semmi, a megfelelő gomb megnyomásával a következő szóra lép. A program az egyes szavakra adott válaszok idejét is rögzíti.

A felhasználó szabad asszociációs tesztet végez (Gósy, 2005), ahol a hívószóra kimondja az első eszébe jutó szót, azzal a különbséggel, hogy az

asszociációs tesztek többségétől eltérően, itt olvasott szavakra írásban ad választ (2. ábra). Az asszociációs tesztek gyakoribb formája – amikor hallott szóra szóban válaszol a kísérletben résztvevő személy – számos technikai nehézséget okozna, és pontosan a nagy mennyiségű adat automatikus gyűjtését tenné lehetetlenné (pl. minden egyes, csak egy kicsit is másképp kiejtett szót a számítógép új szóként értelmezne – ez a probléma gépelt válaszok esetében csak sokkal kisebb mértékben jelentkezik). Ugyanakkor a szóasszociációs teszt írásbeli formája is ismert: ezzel a módszerrel nagyméretű (5000 szó, 6000 válaszadó, 72 000 válasz, 1973-1998 között) angol nyelvű szóasszociációs adatbázist hoztak létre (Nelson, *et al.*, 1998).



2. ábra

Az Agykapocs projekt hívószót adó és válaszszót váró oldala

A programban szabályozható, hogy a felhasználó mely hívószavakra, milyen sorrendben adjon választ. Ezek a szavak (fix szólista) után a felhasználónak a rendszerben lévő további szavak véletlenszerűen jelennek meg. Válaszként a rendszer (szintén technikai és módszertani megfontolásból) nem csak egy szót fogad el, hanem több szóból álló szócsoporthoz, kifejezéshez, mondatot, 255 karakter hosszúságig. (A programban szereplő adatokkal kapcsolatban kizárólag a

„szó” megnevezést használom, itt utalva rá, hogy a „szó” jelen esetben szócsoportot is takarhat).

A program a szavakat, illetve a köztük lévő kapcsolatokat relációs adatbázisban tárolja. Rögzíti az egyes szavak közötti kapcsolatot, a kapcsolat irányát, valamint rögzíti a kapcsolatok erősségét.

Statisztikai adatokat automatikusan számol a program (3. és 4. ábrák):

- felhasználók száma a rendszerben, nem, kor stb. szerinti megoszlása,
- szavak száma a rendszerben (nyelvenként és összesítve),
- kapcsolatok száma (**alma**->**körte** kapcsolat: egy kapcsolat, ha van **körte**->**alma** kapcsolat is, akkor kettő kapcsolatot),
- adatok száma (**alma**->**körte** válasz 15, **körte**->**alma** 10, akkor a két kapcsolattól az adatok száma 25),
- 100 legerősebb kapcsolat a rendszerben,
- 100 legtöbb bejövő kapcsolattal rendelkező szó,
- 100 legtöbb kimenő kapcsolattal rendelkező szó.

Név	#
Felhasználók száma a rendszerben	574 db
Összes szó nyelvenként	
hu	11703 db
en	167 db
de	154 db
it	107 db
fr	102 db
lv	99 db
ru	99 db
pl	97 db
Szavak száma a rendszerben	12528 db
Összes adat nyelvenként	
hu	64745 db
de	180 db
it	48 db
en	25 db
Összes adat a rendszerben	64998 db
Összes kapcsolat nyelvenként	
hu	29829 db
.	1 db

3. ábra

Az Agykapocs adminisztrációs felületén látható statisztikai adatok: felhasználók száma, szavak száma, adatok száma

Regisztrált felhasználók ▾ Szólista ▾ Riport ▾ Alap adatok ▾ Szövegek ▾ Beállítások ▾ Fórum ▾	
Összes kapcsolat nyelvenként	
hu	29829 db
de	140 db
it	48 db
en	25 db
Összes kapcsolat a rendszerben	
30042 db	
Első 10 legerősebb kapcsolat	
mobil » telefon (hu)	217 db
bank » pénz (hu)	195 db
tőzsde » pénz (hu)	190 db
kevés » sok (hu)	184 db
sok » kevés (hu)	179 db
euró » pénz (hu)	171 db
hitel » pénz (hu)	166 db
állat » kutya (hu)	162 db
bankkártya » pénz (hu)	155 db
szék » asztal (hu)	149 db
Áfa » adó (hu)	143 db
enni » inni (hu)	139 db
új » régi (hu)	138 db
teve » sivatag (hu)	137 db
Magyarország » haza (hu)	137 db

Kész

4. ábra

Az Agykapocs adminisztrációs felületén látható statisztikai adatok: kapcsolatok száma nyelvenként, legerősebb kapcsolatok

Külön lekérdezéssel három fő funkció érhető el:

a) Adott szó kapcsolatai egyszerű lekérdezéssel:

(Hívószóra adott válaszok száma, valamint egyidejűleg azon szavak, amire a hívószó válaszként jelent meg – pl. jelzi, hogy az **alma** szóra hányszor volt *körte* a válasz, valamint azt is, ha a **dinnye** hívószóra a válaszcso volt az *alma*. l. 5. ábra.)

A lekérdezés első része sima asszociáció, amelyet asszociációs kutatások során megkapunk (Postman & Keppel, 1970; Lengyel, 2005; Lengyel, 2007a; Lengyel, 2007b; Lengyel, 2008a; Lengyel, 2008b; Lengyel, 2008c; Balló és Jaugisztinné, 1983; Navracsics, 2007). A lekérdezés másik része (mely szavakra volt válasz az „alma” szó) lehetővé teszi az asszociációk mélyebb kutatását. Tudomásunk szerint olyan adatbázis, ami a szóból kimenő, valamint a bemenő kapcsolatokat is tanulmányozza, még nem készült.

#	M	T	Név	Válasz szó	Nyelv	Típus
<input type="checkbox"/>			ember => állat (109/495) [22%]	állat => ember (33/473) [7%]	hu	-
<input type="checkbox"/>			ember => élőlény (35/495) [7.1%]	élőlény	hu	-
<input type="checkbox"/>			ember => élet (29/495) [5.9%]	élet => ember (1/7) [14.3%]	hu	-
<input type="checkbox"/>			ember => férfi (21/495) [4.2%]	férfi	hu	-
<input type="checkbox"/>			ember => majom (14/495) [2.8%]	majom => ember (1/1) [100%]	hu	-
<input type="checkbox"/>			ember => ember (11/495) [2.2%]	ember => ember (11/495) [2.2%]	hu	-
<input type="checkbox"/>			ember => no_idea (10/495) [2%]	no_idea	hu	-
<input type="checkbox"/>			ember => barát (9/495) [1.8%]	barát => ember (1/13) [7.7%]	hu	-
<input type="checkbox"/>			ember => fehér (7/495) [1.4%]	fehér	hu	-

5. ábra

Asszociációk mutatása a rendszerben

b) Adott szó kapcsolatai részletes lekérdezéssel:

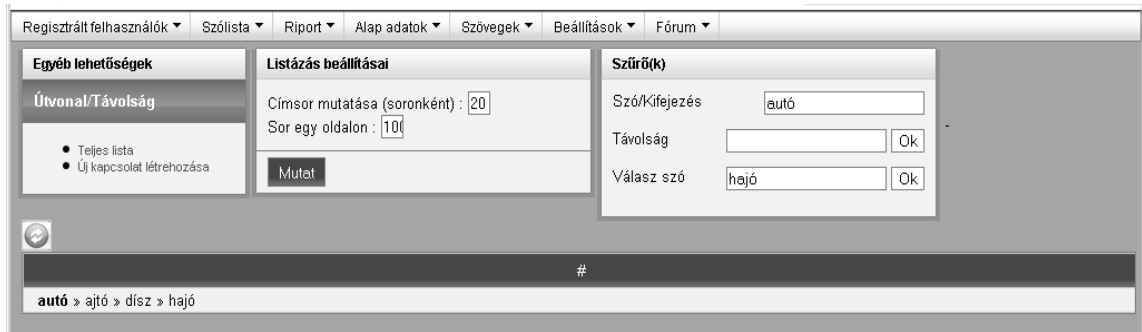
A rendszerben adott szó kimenő kapcsolatai különböző, a regisztrációnál megadott statisztikai adatok figyelembe vételével/szűkítésével is lekérdezhetők: pl. milyen válaszokat adtak nem, korosztály, végzettség stb. szerint. Tudomásunk szerint ilyen mélységig „differenciál” és ilyen szempontok szerint visszakereshető asszociációs adatbázis nem készült.

c) Legrövidebb útvonal lekérdezése:

A programban megadható bármely, a rendszerben szereplő szó kiindulási pontként, valamint bármely, a rendszerben szereplő szó célszóként. A rendszer automatikusan kiszámolja két szó között a legrövidebb útvonalat. (Pl. *alma* kiindulás, *kutya* célszó egy lehetséges útja: *alma*->*gyümölcs*->*növény*->*állat*->*kutya*. l. 6. ábra.) Mindez két formában történik, egyrészt figyelembe véve a kapcsolatok irányát, másrészt a kapcsolatok irányát figyelmen kívül hagyva (Kovács, megjelenés alatt).

Ezen funkció segítségével igazolható a szókincs kis-világ szerveződése.

Tudomásunk szerint ehhez hasonló, ilyen méretű magyar nyelvű kutatás (szavak „távolsága” a mentális lexikonban) eddig nem zajlott. A mentális lexikon hasonló (csak az átlagos legrövidebb utat számoló) feltérképezésére angol nyelven Steyvers & Tenenbaum (2005) kutatása ismert, a Dél-floridai Egyetem szó-asszociációs adatbázisának (The University of South Florida Word Association, Rhyme and Word Fragment Norms) (Nelson, *et al.*, 1998) segítségével.



6. ábra

Legrövidebb kapcsolat mutatása a rendszerben

A projekt 2008 áprilisában indult összesen 100 szóból álló szólistával (1. Melléklet). Tesztelés és javítások után a felhasználók, szavak, kapcsolatok száma a projekt első fázisában a következőképpen változott (1. táblázat):

1. táblázat

Felhasználók, szavak, kapcsolatok száma az Agykapocs rendszerben

Felhasználók száma	Szavak száma (magyar)	Kapcsolatok száma (magyar)
69	3228	6105
108	4006	8086
142	5066	10242
207	5697	13950
268	7219	18491
491	10449	25754
539	11307	28389
576	11857	30342

576 fő esetében:

- Kor szerinti eloszlás: 0-10 (3db); 10-14 (23db); 15-18 (80db); 19-24 (242db); 25-35 (125db); 36-50 (61db); 51-65 (36db); 66< (6db)
- Nem szerinti eloszlás: Férfi (222db); Nő (354db)

A projekt vége, illetve nagysága nyitott, az adatok (számítógépes illetve anyagi erőforrások függvényében) szinte bármilyen számú felhasználótól, bármilyen nyelven gyűjthetők. Jelen publikációban kiemelt részek tehát egy későbbi, sokkal nagyobb adatbázis megvalósításának, és egy szélesebb körű (reményeink szerint nemzetközi) kutatásnak első lépései.

Elvárások a projekttel kapcsolatban

A projekt célkitűzései a következők:

- szóasszociációs adatbázis készítése, amely a jelenlegi adatbázisokkal szemben:
 - folyamatosan bővül,
 - gyors, differenciált lekérdezéseket tesz lehetővé,
- segítségével az asszociációk a jelenlegi kutatásoknál mélyebben elemezhetőek,
- a jelenlegi kutatásokhoz képest lényegesen (nagyságrendekkel) nagyobb adatmennyiséggel, több nyelvvvel dolgozik,
- segítségével a mentális lexikonban kimutatható a hálózatos jelenségek bizonyos megjelenése; többek között, hogy a mentális lexikon kis világgént viselkedik,
- segítségével igazolható, hogy „[...] a mentális lexikon tulajdonképpen a szavak közötti dinamikus hálózat, melynek egyes elemei változó erősségű potenciállal rendelkeznek az aktiválást illetően” (Lengyel, 2007b: 55).

További várható eredmények

- bizonyítani, hogy nagy mennyiségű anonim nyelvészeti adatfelvétel viszonylag kis anyagi és időráfordítással lehetséges a korábbi, hasonló jellegű anonim, pl. Labov adatgyűjtés kritikáinak (Schlobinski, 1996) kikerülésével,
- új nyelvészeti kutatómódszertani (adatgyűjtési) eljárás alapjainak letétele.

A projekt során felmerülő problémák, kérdések

A mentális lexikon asszociációs vizsgálata tudomásunk szerint a mai napig nem tűzte ki célul egy egyén mentális lexikonjának teljes feltérképezést. A mentális lexikon nagyságát így csak hozzávetőlegesen tudjuk megbecsülni (az Egyesült Államokban egy egyetemi hallgató esetében kb. 150 000 szó), mivel nem tisztázott, mi számít egy egységnek a mentális lexikonban (Gósy, 2005):

- Ha minden szám külön fogalom, akkor a mentális lexikon nagysága végtelen.
- Külön számoljuk-e az összetett szavakat?
- Külön számoljuk-e a képzett szavakat?
- A mentális lexikon vizsgálatakor szavakat vagy fogalmakat vizsgálunk? (A „zöld” szó is, fogalom is lehet egyben. Ugyanakkor, ha egy színmegnevezési tesztben valaki „katonai zöld” vagy „lilas-barnás” (Navracsics, 2007) kifejezéssel nevez meg egy színt, az azt jelzi, hogy a színt felfogta, érzékelte, csak a megfelelő szó hiányzik a szókészletéből – jelen esetben tehát a „katonai zöld” egy egység (mivel egy színre utal, ami egy fogalomként van a mentális lexikonban), vagy két egység (szó forma, katonai + zöld)?)

– **nem egy ember asszociációi, hanem „közösségé”**

Jelen projekt nem egy személy mentális lexikonjának feltérképezését tűzte ki célul, hanem (az asszociációs vizsgálatoknak megfelelően) a kutatásban részt vevő személyek szavakra adott asszociációiból próbál következtetéseket levonni egy adott nyelv mentális lexikonjának általános felépítésére vonatkozóan.

– **megbízhatóság**

A kutatás során gyűjtött adatok megbízhatósága döntő fontosságú. Az adatok gyűjtésével kapcsolatban a megbízhatósági kérdések egy része magából a szó-asszociációs vizsgálatból mint módszerből származik. A módszer a pszicholingvisztikai kutatásokban 120 éve ismert (Francis Galton 1883-ban alkalmazta először: l. Gósy, 2005), így a szóasszociációs vizsgálatot mint módszert a mentális lexikon-kutatásoknál elfogadottnak tekintjük.

Kérdéses, hogy milyen mértékben változtatja meg az eredményeket a tény, hogy a válaszadás olvasott szavakra történik, és a válaszokat írásban adják az alanyok. Időkorlát beiktatása a válaszoknál megoldható lett volna, ugyanakkor az adatok egy része (pl. lassú internet-kapcsolat, lassabban gépelő felhasználó, rövidebb/hosszabb szavak gépelése más időt igényel stb.) elveszett volna. Ezért megoldásként a válaszoknál az idő rögzítését választottuk, így részletes lekérdezésnél – amennyiben szükséges – csak azokat az adatokat jelenítjük meg, amelyek pl. 5 mp-en belül érkeztek.

100 szóval, 10 fővel végzett kontrollvizsgálattal a projekt tesztelési fázisában eredményként nagyságrendileg hasonló válaszokat kaptunk, mint a gép előtt végzett tesztekkel. A nagyságrendileg hasonló válaszok jelen esetben azt jelentik, hogy amennyiben adott szóra gyakori volt egy adott válasz írásban (**alma->körte**, 20%), körülbelül hasonló nagyságrendben érkezett ugyanaz a válasz szóban is (**alma->körte** jelen esetben 18,9%). Ki kell viszont hangsúlyoznunk, hogy egyértelműen nem igazolhatók (tudomásunk szerint nem igazolták) sem eltérések, sem teljes azonosság szóban, illetve írásban végrehajtott asszociáció tesztek esetében. Ennek oka egyrészt, hogy az adott asszociációs válaszokat az egyén legutóbbi tapasztalatai is befolyásolhatják, (tehát a hívószóra adott válasz szó ugyanazon személynél más és más lehet, amennyiben többször vizsgáljuk) másrészt, hogy szóban adott válaszok esetében sem tudjuk igazolni, hogy a kísérleti alany az első eszébe jutó választ adta. Jelen kutatásban hangsúlyozzuk, hogy olvasott szóra írásban kértünk választ, ugyanakkor a kontrollkísérlet alapján, fent említett okokból, valamint feltételezésünket megcáfoló kutatások hiányában feltételezzük, hogy az eredmények ugyanúgy elemezhetők, mint a hallott szóra szóban adott válaszok.

Problémát jelenthet, hogy az asszociációs tesztet számítógép előtt töltik ki a kísérleti alanyok. Kérdéses, mennyire befolyásolja a nyelvhasználó személyiségét az internetes kommunikáció. Az internet-felhasználók nyelvhasználatát általánosságban Balázs (2005) és Bódi (2004, 2005), a chat nyelvezetét Balaskó (2005) elemezte. Domonkosi (2005: 156) az internet nyelvhasználatának empirikus megismerésére három lehetőséget lát: „1. a nyelvhasználót, a nyelvhasz-

nálat forrását problematizáló kérdőívezés; 2. a nyelvhasználatot mint produkciót, dinamikát vizsgáló megfigyelés, 3. illetve a nyelvhasználatot mint produktumot vizsgáló nyelvi elemzés.”

Jelen dolgozatban feltételezzük, hogy a módszer (interneten gyűjtött adatok) az asszociációs kutatás minőségét nem befolyásolja, mivel a kapott adatok eloszlása csak minimális eltéréseket mutat a hagyományos asszociációs tesztek eredményeivel szemben. Ugyanakkor asszociációs kutatásoknál az elektronikus kitöltés papír-alapútól eltérő hatása további vizsgálatok tárgya kell, hogy legyen. A jelenleg létrejövő adatbázis az ilyen jellegű kutatásoknak is forrása lehet.

A szólistába bekerült, a papír-alapútól eltérő nyelvhasználatra utaló, további kutatások szükségességét prognosztizáló asszociációs példák:

PIN-kód -> ****

mosoly -> :) :-D

nevetés -> :)

freemail -> @

kukac -> @@

csapatszellem -> .hu

e-mail -> @freemail.hu; @citromail.hu

kiakasztó -> ez a kérdőív:)

e-mail -> kukac

Módszertani szempontból – mivel tudomásunk szerint hasonló jellegű internetes nyelvészeti kutatás még nem zajlott – teljesen új területet térképezünk fel. Segítséget nyújthat az adatok megbízhatóságának megállapításában, hogy a korábban említett internetes kutatások esetében a kutatók az internet-felhasználóktól kapott adatokat megbízhatónak tekintik (Lintott, *et al.*, 2008; Land, *et al.*, 2008; Bamford, *et al.*, 2008). A részvétel a kutatásban önkéntes, így valószínűsíthető, hogy egy résztvevőnek sem célja az adatok meghamisítása (illetve nem nagyobb mértékben kell a problémával számolnunk, mint hasonló jellegű hagyományos kérdőíves stb. felmérések esetében). Hibás adatok keletkezhetnek egyrészt elgépelésből, másrészt amennyiben a kutatásban részt vevő tudatosan hamis adatot (válaszszót) ad meg. Ezen lehetőségek minimalizálása érdekében az új szavak csak ellenőrzés után kerülnek a rendszerbe.

Telefonos kérdőívek és web alapú kérdőívek összevetésének eredményeként kimutatták, hogy a web alapú felmérésekkel kapott adatok megbízhatóbbak, a felmérések olcsóbbak (telefonos felmérések költségének 10-30%-a) és gyorsabbak (Roster, *et al.*, 2004; Braunsberger, *et al.*, 2007).

Jelen kutatás során viszonylag gyorsan, nagy mennyiségű adatot kívánunk felvenni és feldolgozni. Az adatfelvétellel kapcsolatban felmerülhetnek hasonló jellegű módszertani kérdések, mint Labov r kiejtési kísérletével kapcsolatban (Schlobinski, 1996; Wardaugh, 2002). A kutatás során megpróbáltuk a módszertani hibák lehetőségét minimálisra csökkenteni. Ugyanakkor mivel új módszerről van szó, illetve a kísérletben résztvevőkkel nincs személyes kapcsola-

tunk, nem zárható ki, hogy az asszociációk egyes esetekben „hibásak” (vagyis nem az első válaszcót adták meg stb.). Ezek kiszűrése nem megoldható, viszont a kutatás tényleges eredményeit nem befolyásolják. Amennyiben egy szóra 200 válasz érkezett, az esetleges néhány „hamis” válasz nem befolyásolhatja az eredményeket. (A probléma abból a szempontból is további kutatást igényel, hogy egyáltalán lehet-e hibásnak tekinteni bármilyen asszociációt válaszként – természetesen eltekintve az olyan egyértelmű esetektől, ahol minden hívószóra következetesen „xxx” vagy „kutya” a válasz.)

Az esetleges hamis válaszokat figyelembe véve a „legrövidebb út” számolása esetében automatikusan a legtöbb kapcsolatú utat keresi a program. Vagyis amennyiben alma->pite->étel->etet->macska út létezik 1-1-1-1 erősségű kapcsolatokkal, és létezik az alma->gyümölcs->növény->állat->macska azonos hosszúságú útvonal 3-5-5-3 erősségű kapcsolatokkal, a rendszer utóbbit fogja megjeleníteni. Így a valamilyen okból hibásan bekerült adatok nem (vagy csak kisebb mértékben) befolyásolják az eredményeket.

– szavak ellenőrzése

A felhasználó által a hívószóra adott válasz gépelt formában nagy- és kisbetűket, szóközöket, írásjeleket figyelembe véve kerül rögzítésre.

- Amennyiben a válaszcó adott formában szerepel már az adatbázisban, a kapcsolatok hozzáadása a hívó- és válaszcóhoz automatikusan történik.
- Új szó akkor kerülhet be az adatbázisba, ha (adminisztrátori jogosultsággal bíró) kutató manuálisan ellenőrizte a szót, valamint azt, hogy a szó milyen nyelvű. A szó ellenőrzés előtt az egyszerű és részletes lekérdezésben már szerepel, de nem kerül még felvételre a szólistába, vagyis a szó nem jelenhet meg hívószóként más felhasználónak. Amennyiben a szó helyes (elgépelés stb.), az ellenőrző személy a szót egyszerűen elfogadja, így a szó bekerül az adatbázisba. Az adatbázisba bekerült szó (és a létrejött kapcsolat) szerepel minden statisztikában, egyszerű és részletes lekérdezésben, legrövidebb útvonal keresésében, továbbá az elfogadott szó része a szólistának, vagyis a felhasználóknak megjelenhet hívószóként.

A szó ellenőrzésénél javításra, törlésre az alábbi esetekben kerülhet sor:

- hibásan írt (elgépelt) szó, amennyiben az elírás egyértelmű (*kutya* helyett *kuyta*),
- helyesírási egységesítés (pl. minden tulajdonnév nagy kezdőbetűvel szerepeljen),
- hibás a szóhoz rendelt nyelv (a magyar **asztal** szóra a német *Tisch* szó érkezik válaszként. A rendszer alapértelmezésben magyar szóra magyar szót vár, vagyis feltételezi, hogy a *Tisch* szó magyar. Itt az ellenőrzésnél, a nyelv megadásánál a legördülő menüből a magyar helyett kiválasztjuk a „német” besorolást),
- egyértelműen hibás válaszok törlése (hívószóra xxxyyy a válasz),
- nem eldönthető, hogy a válaszadó milyen választ kívánt megadni. (Pl. **kutya** hívószóra *akcsa* válasz. Ebben az esetben az asszociáció lehet

kacsa de macska is, ami az elgévelt válaszból nem eldönthető. Itt az adat elvesztését tudomásul véve a válaszszó törlésre kerül.)

Lehetőségként szerepel a szó „parkolópályára” állítása, vagyis a szó ellenőrizetlenül marad. Így később ellenőrizhető, a lekérdezésekben szerepel, ugyanakkor a szó hívószóként nem kerül be a szólistába.

A válaszsavak egy részénél (a projekt jelenlegi fázisában) ezt a megoldást alkalmaztuk. Az oldalnak nem célja, hogy korhatárhoz kösse a felhasználókat. Ugyanakkor válaszként több esetben megjelentek káromkodások, nemiségre utaló (durva) kifejezések. Ezeket a szavakat (jogi és adatvédelmi előírásokat szem előtt tartva) nem tettük be a szólistába, azokat ellenőrizetlenül, „parkolópályán” hagytuk, mivel nem tudjuk garantálni, hogy ezen szavak ne kiskorúaknak jelenjenek meg hívószóként. Az adatbázisban tehát ezek a szavak rögzítésre kerülnek (tehát a statisztikák, lekérdezések részét képezik), ugyanakkor nem jelennek meg hívószóként.

– nyelv/kódlap

A projekt során probléma a nyelvek betűinek (karaktereinek) számítógépes kezelése. Sajnos jelenleg nincs olyan informatikai megoldás, ami minden esetben problémamentesen kezelné a különböző nyelvek karaktereit. Kódolásnál az UTF-8 kódolást választottuk. Az egyes nyelvek kódolása és helyes karaktermegjelenítése így (az UTF határain belül) egységes. Ugyanakkor nem találtunk megoldást annak a problémának a kezelésére, ha valaki nem adott nyelvű billentyűzetről viszi be a szavakat, és így bizonyos karaktereket más karakterekkel helyettesít (pl. orosz nyelvű szavakat magyar billentyűzetről, latin betűkkel gépel). Ilyen esetekben megoldásként egyelőre csak a megfelelő szavak manuális átírását találtuk.

A projekt tervei között szerepel az adatbázis bővítése további nyelvekkel (kínai, japán, arab stb.), ahol a kódolás problémája fokozottan jelentkezik.

Nyelvi, illetve nyelvi ellenőrzési problémát jelent, hogy a különböző nyelvű válaszsók esetében a szót ellenőrző személy ideális esetben adott nyelv anyanyelvi beszélője, lehetőleg nyelvész (alkalmazott nyelvész), pszicholingvisztikai ismeretekkel. Jelenleg ez anyagi okokból (a projekt teljes egészében magánfinanszírozású, bármilyen intézményi vagy pályázati támogatás nélkül) nem megoldható. Ezért jelenleg (habár az adatbázisban 10 nyelv szerepel) három nyelv – magyar, német, angol –, prioritást élvez. A kutatás jelenlegi fázisában (előbb említett okból) csak a magyar nyelvű asszociációk során elért eredmények kerülnek elemzésre (Kovács, megjelenés alatt). Ugyanakkor az adatbázis használhatósága és értéke megsokszorozódik, amennyiben sikerül több nyelven nagy mennyiségű, megbízható adatot összegyűjteni.

– a kutatásban résztvevők

A kutatás résztvevői önkéntesek. Jelen kutatási módszer nem alkalmas kvótás vagy többlépcsős csoportos mintavétel lefolytatására (Freedman, *et al.*, 2005). Mivel a kutatás az internetet használók asszociációit rögzíti, a résztvevők többsége feltételezhetően a fiatalok, illetve fiatal felnőttek köréből kerül ki: így az ő

asszociációik lényegesen nagyobb számban lesznek reprezentálva az adatbázisban. Ezt a problémát a lekérdezés korosztályonkénti szűkítésének lehetősége megoldja. Az adatbázis természetesen a „manuális úton” internetet nem használók adataival is bővíthető. Ez az asszociációs teszt (pl. első 100 fix szavának) írásban, papír alapon való kitöltésével majd az adatok manuális (adott adatokkal felhasználó regisztrálása) úton történő rendszerbe vitelével történhet. Jelenleg viszont semmilyen módszert nem ismerünk annak biztosítására, hogy az adatbázis – akár csak egy nyelv esetében is –, reprezentatív mintaként (minden korosztályból, végzettség stb. szerint a népesség eloszlásnak megfelelő arány) szolgáljon, ugyanakkor hangsúlyozzuk, hogy ez nem is célunk.

– a munka nagysága (automatikusan is)

A létrejövő adatbázis nagysága és bizonyos (előre nem tervezhetett), a projekt nyitottságából eredő sajátosságok számos gyakorlati megfontolást, illetve néhány esetben az eredeti elképzelések felülbírálását tették szükségessé.

Felülbírálásra került többek között a csak helyesen leírt szavak rögzítése. A kezdeti szakaszban is több esetben találoztunk olyan adatokkal, melyek következetesen, nagy számban, helytelen helyesírással érkeztek. Egyszerű idő- és munka-megfontolásból néhány esetben engedélyeztük, hogy a szó (elvileg) hibásan kerüljön a rendszerbe (EU helyett eu). Helyesírási szempontból ez nem indokolható, ugyanakkor módszertani szempontból igen – a hibásan írt asszociációkból is levonhatók tanulságok.

Több esetben adott szóról nem volt eldönthető, hogy mi a helyes magyar írásmód, mivel a gyakorlatban több alak honosodott meg (menedzser, manager), így mindkét alak a rendszerbe került. Más esetben a magyar helyesírás nem egyértelmű, illetve az elvi helyesírási szabály tudatosan kerül megszegésre: e-mail, ímél – mely utóbbi alak nem helyes, ugyanakkor nem tekinthető egyértelműen hibásnak sem (tudatosan alkotott, „kreatív”, fonetikus alak).

– az adatbázis lezárása, a projekt vége

Kérdésként felmerül, hogy a projekt mondható-e bármely szakaszában lezártnak, illetve célszerű-e lezárni az adatbázist egy adott adatmennyiségnél és időpontban. Amennyiben a cél egy adott helyzetkép rögzítése (pl. asszociációk 2008-ban), az adatbázis lezárása látszik célszerűnek. (Asszociációs adatbázisok létrehozásakor többnyire kitűzött cél elérése után a projekt befejezése jellemző.) Ugyanakkor jelen módszer alkalmas az asszociációk „folyamatos” vizsgálatára, vagyis egy folyamatosan bővülő adatbázis megalkotása a cél. Problémaként mérülhet fel, hogy amennyiben a projekt hosszabb ideig fut, a statisztikai adatok (végzettség, kor) csak részben lesznek igazak: a 2008-ban 14 éves és 2018-ban az akkori 14 éves alany hívószavakra adott válaszai egymás mellett fognak szerepelni, vagyis 2018-ban a projekt nem fogja tudni kimutatni az akkori (2018) leggyakoribb asszociációkat, hiszen az adatbázis 2008-tól folyamatosan bővül, vagyis az előző 10 év adatai is szerepelni fognak benne. Ez a probléma megoldható pl. az állapot évenkénti rögzítésével, azaz az adatbázis részleges lezárásával (és külön tárolásával). Az adatok ekkor évente elemezhetők, a változások jobban

követhetők. Így egyszerre válik lehetővé az asszociációk statikusabb (egy adott időszakban) és dinamikusabb, változó (pl. 10 évig folyamatos) vizsgálata.

– **további lehetőségek**

A rendszer a fent említett, jelenleg működő elemeken túl további funkciókkal bővíthető. Quillain (Collins & Quillian, 1969), illetve a WordNet (Miller, *et al.*, 1990; Miller, 1998; Fellbaum, 1998) elképzeléseit követve a rendszerben a kapcsolatok milyensége jelen esetben is definiálható (pl. madár – szárny kapcsolatnál rész-egész viszony, apa – anya esetében egyenrangú viszony stb.). A munka ezen része (ismét erőforrás hiányában) nem kezdődött meg.

A kapcsolatok milyenségének definiálása több módon képzelhető el:

- valamely korábbi, a szavak közötti kapcsolatokat osztályozó szerző alapján (pl. Quillain ISA, HASA stb. kapcsolatok) (Collins & Quillian, 1969),
- a szemantikus webek számára kifejlesztett RDF nyelv parancsait és logikáját követve (Gottfink, 2006; Szeredy, *et al.*, 2005),
- teljesen új, lehetőleg számítógéppel értelmezhető, de nem az RDF elveit követő kapcsolati kategorizálás.

Az első megoldás mellett a nyelvészeti/pszicholingvisztikai hagyományok szólnak, ugyanakkor az ehhez hasonló kapcsolati rendszerek nem kerültek teljes mélységben kidolgozásra, adott szerzők inkább (a teljesség igénye nélkül) példákat hoznak a lehetséges osztályozásra.

Az RDF-ben alkalmazott osztályozás mellett szól, hogy a kapcsolatok számítógéppel értelmezhetők. Az RDF informatikai megközelítésben kezeli a különböző fogalmi (és adat-) viszonyokat. Az RDF előnye (számítógéppel értelmezhető) egyben hátrány is: csak olyan kapcsolatokat (viszonyokat) enged meg és definiál, amelyek számítógépes értelmezése a nyelv kifejlesztésénél megoldható volt. Az RDF elleni érv lehet továbbá, hogy jelenlegi szemantikai analízisek nem az RDF nyelven alapulnak.

A legcélszerűbb megoldás egy az RDF-en, valamint logikai és/vagy nyelvészeti kutatásokon alapuló új kapcsolatrendszer kialakítása és a számítógép számára értelmezhető módon történő megvalósítása. A dolgozatnak nem célja ezen kapcsolatrendszer alapjainak letétele és elgondolások megvalósítása, csak utalni kívánunk a rendszerben rejlő lehetőségekre.

A rendszerben szereplő egyes szavak további jellemzőkkel bővíthetők (definíció, nyelvtani információk stb.), így a rendszer ebben is hasonlítható a több országban is elindított WordNet projekthez.

Összegzés

Jelen publikáció célja egy a nyelvészetben eddig nem használt kutatási forma/módszer (módszertani megfontolások, a kutatás első szakaszában felmerült problémák) bemutatása. A dolgozat tudatosan nem törekszik és törekedhet (terjedelmi korlátok miatt) az eredmények bemutatására és elemzésre. (Az eddigi eredmények bemutatása további publikációk tárgya lesz, pl. Kovács, megjelenés alatt.) Sokkal fontosabbnak tartottuk a módszer ismertetését, mivel eredményeink szerint jelen módszer az alkalmazott nyelvészet, valamint a nyelvtudomány kutatómódszertani repertoárját egy olyan eszközzel/módszerrel bővítheti, amelynek előnyei:

- nagy mennyiségű adat gyors felvételét teszi lehetővé viszonylag kis emberi és anyagi ráfordítással,
- az adatok egyszerre elektronikusan kerülnek rögzítésre, így gyors, differenciált feldolgozás/elemzés válik lehetővé,
- olyan emberek vonhatók be nyelvészeti kutatásba, akik más módon nem vagy csak nehezen lennének elérhetőek,
- a rendszer moduláris jellegéből adódóan további funkciókkal bővíthető.

Elektronikus források (honlapok)

Agykapocs

<http://www.agykapocs.hu>

BOINC

<http://boinc.berkeley.edu/> (2008.06.30)

ConnectYourMind

<http://www.connectyourmind.com>

eBird

<http://ebird.org/content/ebird/about> (2008.06.30)

GalaxyZoo

<http://www.galaxyzoo.org/> (2008.06.30)

InnoCentive

<http://www.innocentive.com> (2008.07.11)

SETI@Home

<http://setiathome.berkeley.edu/> (2008.06.30)

Stardust@home

<http://stardustathome.ssl.berkeley.edu/> (2008.06.30)

WordNet

<http://wordnet.princeton.edu> (2005.06.20.)

Irodalom

- Auger, A. and Barrière, C.** (2008) Pattern-based approaches to semantic relation extraction: A state-of-the-art. *Terminology*. 14/1 2008. Special Issue: Pattern-based Approaches to Semantic Relation Extraction. Amsterdam: John Benjamins. pp. 1-19.
- Aussenac-Gilles, N. and Jacques, M. P.** (2008) Designing and evaluating patterns for relation acquisition from texts with CAMÉLÉON. *Terminology*. 14/1 2008. Special Issue: Pattern-based Approaches to Semantic Relation Extraction. Amsterdam: John Benjamins. pp. 45-73.
- Balaskó M.** (2005) Virtuális közösségek kommunikációja a cybertérben. In: Balázs G. és Bódi Z. (szerk.) *Az internetkorszak kommunikációja*. Budapest: Gondolat-Infonia. 58-88.
- Balázs G.** (2005) Az internetkorszak kommunikációja. In: Balázs G. és Bódi Z. (szerk.) *Az internetkorszak kommunikációja*. Budapest: Gondolat-Infonia. 25-57.
- Balló L. és Jagusztinné U. K.** (1983) *Magyar verbális asszociációk*. Szeged: Juhász Gyula Tanárképző Főiskola.
- Bamford, S. P., Nichol, R. C., Baldry, I. K., Land, K., Lintott, C. J., Schawinski, K., Slosar, A., Szalay, A. S., Thomas, D., Torki, M., Andreescu, D., Edmondson, E. M., Miller, C. J., Murray, P., Raddick, M. J. and Vandenberg, J.** (2009) Galaxy Zoo: the independence of morphology and colour. *Monthly Notices of the Royal Astronomical Society*. 393:4. pp. 1324-1352.
- Barabási A. L.** (2003) *Behálózva*. Budapest: Magyar Könyvklub.
- Bódi Z.** (2004) *A világháló nyelve*. Budapest: Gondolat.
- Bódi Z.** (2005) Szimbolikus írásbeliség az internetes interakcióban. In: Balázs G. és Bódi Z. (szerk.) *Az internetkorszak kommunikációja*. Budapest: Gondolat-Infonia. 195-212.
- Braunsberger, K., Wybenga, H. and Gates, R.** (2007) A comparison of reliability between telephone and web-based surveys. *Journal of Business Research*. 60/7. pp. 758-764.
- Collins, A. M. and Quillian, M. R.** (1969) Retrieval time from semantic memory. *Journal of verbal learning and verbal behavior* 8 (2) pp. 240-248, változatlan formában In: Balota, D. A. and Marsh, E. J. (eds., 2004) *Cognitive Psychology: Key Readings*. New York: Psychology Press. 395-402.
- Csermely P.** (2005) *A rejtett hálózatok ereje*. Budapest: Vince.
- Domonkosi Á.** (2005) Az internet nyelvhasználatának empirikus kutatási lehetőségei. In: Balázs G. és Bódi Z. (szerk.) *Az internetkorszak kommunikációja*. Budapest: Gondolat-Infonia. 143-158.
- Fellbaum, Ch.** (ed., 1998) *WordNet*. Cambridge: MIT Press.
- Fóris Á.** (2007) A skálafüggetlen hálók nyelvészeti vonatkozásai. *Alkalmazott Nyelvtudomány*, VII. / 1-2. 105-125.
- Freedmann, D., Pisani, R. and Purves, R.** (2005) *Statisztika*. Budapest: Typotex.
- Gósy M.** (2005) *Pszicholingvisztika*. Budapest: Osiris.
- Gotttdank T.** (2006) *Szemantikus Web*. Budapest: ComputerBooks.
- Halskov, J. and Barrière, C.** (2008) Web-based extraction of semantic relation instances for terminology work. *Terminology*. 14/1 2008. Special Issue: Pattern-based Approaches to Semantic Relation Extraction. Amsterdam: John Benjamins. pp. 20-44.
- Kovács L.** (2007) Mentális lexikon és kis világok. *Alkalmazott Nyelvtudomány*, VII. / 1-2. 140-150.
- Kovács L.** (2008) Kis világok egy pszicholingvisztikai kutatás tükrében. In: Balaskó M. és Balázs G. (szerk.) *Konvergenciák 2003-2006*. Szombathely: NYME SEK. 305-314.
- Kovács L.** (megjelenés alatt) Irányított kapcsolatok a mentális lexikonban. (*Modern Nyelvoktatás*, 2009).
- Land K., Lintott C. J., Schawinski, K., Slosar, A., Bamford, S., Thomas, D., Raddick, J.M., Nichol, R.C., Szalay, A., Andreescu, D., Murray, P. and Vandenberg, J.** (2008) Galaxy Zoo: The large-scale spin statistics of spiral galaxies in the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*. 388/4. pp. 1686-1692.
- Lengyel Zs.** (2005) *Asszociáció vizsgálatok*. (előadás) VIII. Pszicholingvisztikai Nyári Egyetem, 2005. május 29.-június 3.
- Lengyel Zs.** (2007a) Asszociációs normák enciklopédiája. *Alkalmazott Nyelvtudomány* 2007/1-2. 125-141.

- Lengyel Zs.** (2007b) Rövidzárlat. In: Benő A., Fazekas E. és Szilágyi N. S. (szerk.) *Nyelvek és nyelv-változatok. Köszöntő kötet Péntek János tiszteletére*. Kolozsvár: Anyanyelvápolók Erdélyi Szövetsége. 2. kötet. 46-55.
- Lengyel Zs.** (2008a) Férfi – nő: 10-14 évesek körében végzett verbális asszociációs vizsgálat alapján. In: Bodnár I., Kegyesné Szekeres E. és Simigné Fenyő S. (szerk.) *Sokszínű nyelvészet. "Női szóval - női szemmel" Gender kutatás a nyelvészetben és az irodalomban*. Miskolc: Miskolci Egyetem. 55-63.
- Lengyel Zs.** (2008b) *Magyar Asszociációs Normák Enciklopédiája: új perspektívák*. Elhangzott MANYE 2008. április 3-5.
- Lengyel Zs.** (2008c) *Magyar asszociációs normák enciklopédiája I*. Budapest: Tinta.
- Lintott, C. J., Schawinski, K., Slosar, A., Land, K., Bamford, S., Thomas, D., Raddick, J.M., Nichol, R.C., Szalay, A., Andreescu, D., Murray, P. and Vandenberg, J.** (2008) Galaxy Zoo: Morphologies derived from visual inspection of galaxies from the Sloan Digital Sky Survey. *Monthly Notices of the Royal Astronomical Society*. 389/3. pp. 1179-1189. http://arxiv.org/PS_cache/arxiv/pdf/0804/0804.4483v1.pdf (2008.09.17.)
- Marshman, E.** (2008) Expressions of uncertainty in candidate knowledge-rich contexts: A comparison in English and French specialized texts. *Terminology*. 14/1 2008. Special Issue: Pattern-based Approaches to Semantic Relation Extraction. Amsterdam: John Benjamins. pp. 124-151.
- Miller, G. A., Beckwith, R., Fellbaum, C., Gross, D. and Miller, K.** (1990) Introduction to WordNet: an on-line lexical database. *International Journal of Lexicography*. pp. 235-244. <http://wordnet.princeton.edu/5papers.pdf> (Javított, bővített változat, 1993)
- Miller, G. A.** (1998) Nouns in WordNet. In: Fellbaum, Ch. (ed.) *WordNet*. Cambridge: MIT Press.
- Navracsics J.** (2007) *A kétnyelvű mentális lexikon*. Budapest: Balassi.
- Nelson, D. L., McEvoy, C. L. and Schreiber, T. A.** (1998) *The University of South Florida word association, rhyme, and word fragment norms*. <http://www.usf.edu/FreeAssociation/>.
- Postman, L. and Keppel, G.** (eds., 1970) *Norms of Word Association*. New York: Academic Press.
- Roster, C. A., Rogers, R. D., Albaum, G. and Klein, D.** (2004) A comparison of response characteristics from web and telephone surveys. *International Journal of Market Research*. 46/3. pp. 359-373.
- Schlobinski, P.** (1996) *Empirische Sprachwissenschaft*. Opladen: Westdeutscher Verlag.
- Schmidt, J.** (2005) Online-Forschung. In: Lehmann, K. und Schetsche, M. (Hg.) *Die Google-Gesellschaft*. Bielefeld: Transcript.
- Sierra, G., Alarcón, R., Aguilar, C. and Bach, C.** (2008) Definitional verbal patterns for semantic relation extraction. *Terminology*. 14/1 2008. Special Issue: Pattern-based Approaches to Semantic Relation Extraction. Amsterdam: John Benjamins. pp. 74-98.
- Soler, V. i Alcina, A.** (2008) Patrones lexicos para la extracción de conceptos vinculados por la relación parte-todo en español. *Terminology*. 14/1 2008. Special Issue: Pattern-based Approaches to Semantic Relation Extraction. Amsterdam: John Benjamins. pp. 99-123.
- Steyvers, M. and Tenenbaum, J. B.** (2005) The Large-Scale Structure of Semantic Networks: Statistical Analyses and a Model of Semantic Growth. *Cognitive Science*, 29. pp. 41-78.
- Strogatz, S.** (2003) *Sync*. London: Penguin.
- Szabó K. és Hámori B.** (2006) *Információgazdaság*. Budapest: Akadémiai.
- Szeredi P., Lukácsy G. és Benkő T.** (2005) *A szemantikus világháló elmélete és gyakorlata*. Budapest: Typotex.
- Tapscott, D. and Williams, A.** (2007) *Wikinómia*. Budapest: HVG.
- Vitevitch, M. S.** (2008) What can graph theory tell us about word learning and lexical retrieval? *Journal of Speech, Language, and Hearing Research*, 51. pp. 408-422.
- Wardaugh, R.** (2002) *Szociolingvisztika*. Budapest: Osiris.
- Watts, D.** (2004) *Six Degrees*. London: Random House.

Melléklet

Megjegyzés: a táblázatban minden szó nyelvi megjelölése: hu, azaz magyar, ezért minden egyes szónál ezt külön nem jeleztük.

Szó / Nyelv (hu)	Szó / Nyelv (hu)	Szó / Nyelv (hu)	Szó / Nyelv (hu)
alma	motor	információ	kevés
ember	nyelv	politikus	kutya
barátság	tudás	mosoly	Magyarország
hitel	reklám	vám	Áfa
szép	foci	bolt	hobbi
piros	inni	állat	gazdaság
könyv	diploma	tiszta	szabadidő
ünnepelni	sport	számítógép	fordítás
autó	pénz	divat	határ
biztonság	kenyér	marketing	vizsga
tőzsde	bevásárolni	kórház	öröm
szépség	manager	bankkártya	múlt
politika	wellness	Európa	orvos
Tv	adó	film	euró
olvasni	olcsó	gyűlölni	kedves
drága	szabad	egészség	piszkos
e-mail	teve	étel	világ
vidám	hír	internet	új
szerelem	enni	sokkolni	remény
bank	gazdag	kereskedelem	
segíteni	tolmácsolás	hangulat	
mobil	környezetvédelem	választás	
zöld	fizetés	ár	
nyaralás	szín	környezet	
ajándék	újság	aludni	
vállalkozás	sok	nevetni	
szék	venni	nyár	