Kísérlet magyar szavak jelentéshasonlóságának meghatározására a *Magyar szókincstár* segítségével

Bárdosi Vilmos[1], Kiss Gábor[2], Kiss Márton[3], Rapcsák Tamás[4]

[1] ELTE BTK Francia Tanszék, tanszékvezető egyetemi tanár;
1088 Budapest, Múzeum krt. 4/c.;
e-mail: vbardosi@ludens.elte.hu
[2] MTA Nyelvtudományi Intézete, Korpusznyelvészeti Osztály; TINTA Könyvkiadó,
http://tintakiadó.hu, igazgató, főszerkesztő; 1117 Budapest, Kondorosi út 17.;
e-mail: kissgabo@tintakiado.hu
[3] SZTE Műszaki Informatika Szak, egyetemi hallgató; TINTA Könyvkiadó,
1117 Budapest, Kondorosi út 17.;
e-mail: Kiss.Marton.1@stud.u-szeged.hu
[4] MTA SZTAKI Operációkutatás és Döntési Rendszerek Laboratórium és Osztály,
tudományos osztályvezető; 1111 Budapest, Kende u 13–17.;
e-mail: rapcsak@oplab.sztaki.hu

Kivonat: A szavak jelentéshasonlóságának meghatározására irányuló kutatások és kísérletek a mintegy fél évszázados asszociációs pszicholingvisztikai kísérletek után az utóbbi évtizedben ugrásszerűen megnőttek. A növekedés okai a természetes nyelvek gépi feldolgozása technológiájának látványos fejlődése és a ma már széles körben elérhető elektronikus nagy nyelvi adatbázisok (egynyelvű szótárak, tezauruszok, korpuszok, WordNet) létrehozása. Előadásunkban bemutatjuk kísérletünket, melyben a *Magyar szókincstárat* [Kiss 1998], pontosabban az abban lévő 25787 címszó alatt található 42976 szinonimasort miként használtuk fel kiindulási nyelvi tudásbázisként szópárok (egyes aljelentések szerint megkülönböztetett) jelentéshasonlóságának meghatározására. Ismertetjük a szópárok jelentéshasonlósági mérőszámaiból felépített – szófajokra szétbontott – jelentéshasonlósági mátrixok létrehozásának menetét. Kísérletet végeztünk, hogy a jelentéshasonlósági mátrixokból kiindulva szinguláris érték dekompozíció (SVD) alkalmazásával miként lehet automatikusan fogalomköröket generálni.

1. A nyelvi intuíció és a szavak jelentésének hasonlósága

Közhelyszerű tény, hogy míg egyes szavak jelentése közel van egymáshoz, más szavaké távolinak tűnik. Az anyanyelvi beszélő a *ballag ~ sétál* szavak jelentését igen hasonlónak érzi egymáshoz, ugyanúgy mint a *fut ~ szalad* szópár tagjainak a jelentését is. Azonban a nyelvhasználó a *ballag ~ fut* szópár tagjainak a jelentését már távolabbinak véli egymástól, még ha a hasonlóság nagyságát, mértékét szavakkal nehezen is tudja megfogalmazni. Minden beszélő nyelvi kompetenciája azt sugallja, hogy ezek a szavak valamilyen módon egy csoportba tartoznak, hiszen mindegyik a

helyváltoztatással, mozgással kapcsolatos. Sőt belső nyelvi intuíciója alapján azt is érzi, hogy a *ballag, sétál* szavak mellé oda kívánkozik a *bandukol,* és ugyanúgy a *fut, szalad* szavak után felsorolható például a *rohan* szó. Mindeközben a nyelvi intuíció azt sugallja, hogy a *bandukol, sétál, ballag* és a *fut, szalad, rohan* két szóhármas valamiféle polaritást fejez ki, valamilyen képzeletbeli skála két végpontján helyezkedik el, és közéjük a *megy, jár, halad* szavak illenek be.

2. A szójelentés-hasonlóság mérésének eddigi útjai

Felvetődik a kérdés, hogy ezt az intuitív érzést, hogy az egyes szavak jelentése hol jobban, hol kevésbé hasonlít egymásra, ki tudjuk-e fejezni valamilyen számértékkel. Valamilyen módon meghatározható-e olyan mutató, amely kifejezi a szavak jelentésbeli hasonlóságát, illetve távolságát. Vagyis megállapítható-e, hogy a *ballag, sétál, bandukol, megy, jár, halad, fut, szalad, rohan* szavakból alkotott jelentéshasonlóság mátrixnak az egyes celláiban milyen mérőszámok helyezkednek el.

	ball	*sétá*	*ban*	*meg*	*iár*	*hal*	*fut*	*szal*	*roh*
ballag	1	?	?	?	?	?	?	?	?
sétál	-	1	?	?	?	?	?	?	?
banduk	-	-	1	?	?	?	?	?	?
megy	-	-	-	1	?	?	?	?	?
jár	-	-	-	-	1	?	?	?	?
halad	-	-	-	-	-	1	?	?	?
fut	-	-	-	-	-	-	1	?	?
szalad	-	-	-	-	-	-	-	1	?
rohan	-	-	-	-	-	-	-	-	1

A feladat tehát N szó (objektum) hasonlóságának kiszámításakor az (N * (N–1)) / 2 darab mérőszám meghatározása, majd továbblépésként a hasonlósági mátrix matematikai módszerekkel való feldolgozása, fogalomkörök automatikus előállítása.

Két szó jelentéshasonlósága meghatározásának 5 igen jól elkülöníthető útját ismerteti a szakirodalom:

1. A pszicholingvisztikai indíttatású ún. asszociációs tesztek és módszerek a legrégebbiek [Osgood 1952], [Osgood 1957], [Miller 1971]. Ezeknek a következő négy főbb típusa van: 1. skálázás, 2. asszociáció, 3. helyettesítés, 4. osztályozás. Az asszociációs fogalmi pszichológiai kísérletek tapasztalatinak ismertetése során helyesen mutat rá Varga Dénes, hogy az eredmény sok esetben különböző szintű, akár 8–10 féle asszociációból tevődik össze: 1. tárgy / tulajdonság, 2. közös / alárendeltség, 3. ellentét, 4. faj / nem, 5. hasonlóság, 6. objektum / cselekvés, 7. közös előfordulás, 8. rész /egész, 9. ok / okozat [Varga 1968]. Mi is úgy véljük, hogy az asszociációs tesztek éppen ezért finom jelentéshasonlóságok meghatározására alkalmatlanok, még ha erre jó néhány kísérlet történt is az utóbbi években.

2. Szójelentés-hasonlóságok meghatározására gyakran felhasználják a szóanyagukat hierarchikusan bemutató fogalomköri szótárakat. Ezek közül is a Roget's thesaurus [Kirkpatrick 1998] a legtöbbet vizsgált és alkalmazott [Morris, Hirst 1991], [Manuba, Takco 1994].

3. A fogalomköri egynyelvű szótárak mellett, az egynyelvű értelmező szótárak is számos vizsgálatnak jelentették a kiinduló pontját. Hideki és Teiki az LDOC (Longman Dictionary of Contemporary English) szótár definícióiból kiindulva végeztek szójelentés-hasonlósági vizsgálatokat. Rámutattak, hogy az egynyelvű értelmező szótárakból azért is célszerű kiindulni, mert azok értelmezéseit lexikográfusok szerkesztették meg értő módon [Hideki, Teiji 1993].

4. Minden kétséget kizárólag a WordNet megjelenése áttörést hozott a természetes nyelvek gépi feldolgozásában és ezen belül is a szópárok jelentéshasonlóságának vizsgálatában, Az új lehetőségekre G. A. Miller hívta fel talán először a figyelmet [Miller 1990], majd számos tudományos és népszerűsítő tanulmány, könyv ismerteti azokat [Fellbaum 1998]. A WordNet alapú szójelentés-hasonlóságok számításának 5 féle módját fejlesztették ki az elmúlt évtizedben [Leacock, Chodorow 1998], [Jiang, Conrath 1997], [Resnik 1995], [Lin 1998], [Hirst, St. Onge 1998], ezeket összefoglalóan ismerteti a Budanitsky, Hirst szerzőpáros [Budanitsky, Hirst 2001].

5. Korpusz alapú szójelentés-hasonlósági vizsgálatok igen sok helyen folynak, ezek alapja a ma már több nyelven is létező, nagyságrendileg 100 millió szövegszót tartalmazó nemzeti nyelvű korpuszok [Jiang, Conrath 1998].

3. Magyar szavak jelentéshasonlóságának megállapításához a magyar nyelvi tudásbázis kiválasztása

Ha magyar szavak jelentését, illetve szópárok jelentéshasonlóságát kívánjuk vizsgálni, először is ésszerűnek tűnik, hogy szótárhoz, esetünkben a legnagyobb magyar köznyelvi szótárhoz, *A Magyar Nyelv Értelmező Szótárához* [ÉrtSz.] forduljunk. A következő jelentésmeghatározásokat, definíciókat találjuk ott, példaképp a *ballag* és a *sétál* szócikkeket adjuk közre:

ballag: tn ige

1. <Ember, állat> lassan kényelmesen lépegetve megy vagy jön.

2. (Isk) <Végzős közép- vagy főiskolás diák> az utolsó tanév legvégén az intézet helyiségeit és környékét csoportosan, hagyományos módon bejárva búcsúzik az iskolától.

3. felé ballag: <személy> az ötvenedik, a hatvanadik, hetvenedik stb. évéhez közeledik.

sétál: tn ige

1. <Személy> testmozgás, levegőzés végett, vagy kedvtelésből lassú, nyugodt léptekkel megy, jár; sétát tesz.

|| **a.** <Rendszerint zárt térben> fel s alá, ide-oda jár, járkál.

|| **b.** (nép) <Ingaóra sétálója> ide-oda leng, <az óra> jár.

2. sétál valahová: sétálva valahova megy.

3. (rosszalló) <Személy> feltűnő ráérő(s)en, illetve a szükségesnél lassabban, kényelmesen lépked.

|| **a.** (rosszalló) Henyén, dologkerülő módon jár-kel.

4. (szoc. e. biz) Állás, munka nélkül van.

Amennyiben többé-kevésbé automatikusan, valamelyest algoritmizáltan szeretnénk az ÉrtSz.-nek a fenti jelentésmeghatározásaiból, azaz szöveges, mondat formájú definícióiból a jelentéshasonlóságokat jelző mérőszámokat meghatározni, nehéz dolgunk lenne. Ezért az ÉrtSz.-et, mint segédeszközt és kiindulási pontot a jelentések hasonlóságának meghatározásához el kell vetnünk.

Az ÉrtSz. tanulmányozása a fentiek ellenére nem volt hiábavaló, hiszen rávilágít arra, hogy egy-egy szónak a jelentése természetszerűleg aljelentésekből tevődik össze. Így, amikor szavak jelentésének hasonlóságát kívánjuk meghatározni, akkor a két vagy több összevetni kívánt szónak az aljelentéseit kell összevetnünk, összehasonlítanunk.

4. A *Magyar szókincstár* mint magyar szavak jelentéshasonlóság mérésének a nyelvi tudásbázisa

Ha figyelmesen szemléljük az ÉrtSz. fenti definícióit, észrevehetjük, hogy a szótár szerkesztői a mondatszerű definíciókat igen gyakran szinonimákkal toldották meg. A gazdag tartalmú, szókészletünk elemeit lineárisan feldolgozó *Magyar szókincstár – Rokon értelmű szavak, szólások és ellentétek szótára* [Kiss 1998] előszavában a főszerkesztő a következőket írja: „A Magyar Szókincstárnak azon kívül, hogy szinonimaszótárként forgatva valamely szó szinonimáit, sőt ellentéteit is kikereshetjük belőle, van egy további haszna is. Ugyanis a szótár rokon értelmű szavai a maguk módján magyarázzák, értelmezik azt a címszót, mely alá be vannak sorolva. Így a *Magyar szókincstár* bizonyos tekintetben értelmező szótári funkciót is betölthet, hiszen az olvasó számára egy kevésbé ismert, homályos jelentésű szónak a jelentését a szinonimák pontosíthatják, megvilágosíthatják." Az idézet elegendő indokot szolgáltat arra, hogy a fenti 9 tagú (*ballag, sétál, bandukol, megy, jár, halad, fut, szalad, rohan)* szósornak megvizsgáljuk a szinonima sorait a *Magyar szókincstárban*, abból a célból, hogy azok alkalmasak-e jelentéshasonlóságok számításának kiinduló pontjaként. Példaképpen megmutatjuk, hogy a *Magyar szókincstárban* a *ballag* és a *sétál* címszavak alatt a következő szinonimasorokat találjuk:

ballag (ige)
1• bandukol, baktat, mendegél, megy, lépked, kullog, cammog, andalog, battyog, slattyog (biz), sétál, poroszkál, kutyog (táj), ballagdál (táj), ballókál (táj), bandikál (táj), bandukál (táj)
2• [iskolától] búcsúzik
sétál (ige)
1• jár, ballag, megy, gyalogol, mozog, kimozdul, levegőzik, kirándul, fordul egyet, kerül egyet
2• sétálgat, sétafikál (pej), sétifikál, járkál, jár-kel, grasszál (pej), korzózik (biz), flangál (biz), flangíroz (rég), spacíroz (rég), promenál (rég), andalog, lötyög (biz), kószál, kódorog (pej), őgyeleg, lődörög, csatangol, cselleng, gévalyog (táj)

A fenti mintából látszik, a *Magyar szókincstár* szerkesztői is (az ÉrtSz.-hez hasonlóan) egy-egy címszó jelentését aljelentésekre bontották, és ezeknek az aljelentéseknek adták meg a szinonimáit. A vizsgált 9 szónak összesen 48 aljelentése van (*ballag* 2, *sétál* 2, *bandukol* 1, *megy* 10, *jár* 13, *halad* 4, *fut* 9, *szalad* 5, *rohan* 2).

Ez azt jelenti, hogy a fenti 9 szó jelentéshasonlóságának meghatározáshoz egy 48 x 48 méretű hasonlósági mátrix kitöltése a feladat.

Míg az ÉrtSz. mondatszerű, szöveges definícióit nem tudtuk eredményesen felhasználni jelentéshasonlóságok meghatározására, úgy tűnik, hogy a *Magyar szókincstárban* közölt szinonimasorok jó kiinduló pontként szolgálhatnak a fenti feladat megoldásához.

A nyelvi tudásbázisként felhasználni kívánt *Magyar szókincstár* a következő jellemzőkkel rendelkezik:

szófaj	szótárban használt rövidítés	címszavak száma	jelentések száma	római sz. homonima	arab sz. homonima
főnév	fn	12400	19118	525	275
ige	ige	7618	15162	23	136
melléknév	mn	4491	6997	523	60
határozószó	hsz	932	1224	98	17
névmás	nm	105	156	18	6
kötőszó	ksz	71	98	37	14
mutatószó	msz	57	66	25	5
névutó	nu	53	85	54	9
számnév	szn	45	53	41	12
indulatszó	isz	13	15	60	9
igenév	ign	2	2	2	2
összesen		**25787**	**42976**	**1406**	**545**

5.1. A szójelentés-hasonlóság mérőszámának számítása

Egy-egy szinonimasort matematikai értelemben vett halmaznak tekinthetünk, amelyben az elemeket az egyes szinonimák adják. Egy szópár (pontosabban a szópár egy-egy aljelentésének) jelentéshasonlóság mérőszámának – melyet H-val jelölünk, H(SZÓa, SZÓb) – a kiszámítása a két halmaz hasonlóságának a meghatározását jelenti.

Például a *fut ~ szalad* szavak **első** aljelentéseinek a jelentéshasonlóságának a meghatározása a következő két halmaz hasonlóságának a kiszámításából áll:

fut 1 = fut, szalad, száguld, vágtat, üget, vágtázik, kocog, limel, rohan, robog, lohol, lót, nyargal, galoppozik, inal, iramlik, iramodik, cikázik, viharzik, kotor, darizik, sprintel, skerál, spurizik, teker, tép, kóstat

szalad 1 = szalad, fut, siet, repül, futkos, fáradozik, spurizik, rohan, lohol, lót-fut, szedi a lábát, iramlik, kotrecel, limel, lófol, inal, őringel, rőföl, trappol, skerál

Két halmaz hasonlósága mérésének kiinduló pontja a két halmazban előforduló közös elemek (P_{11}) és csak az egyes halmazokban meglévő elemek számának (P_{10}, P_{01}) valamiféle összevetése. Párniczky Gábor 1976-ban megjelent művében halmazok hasonlóságának meghatározására öt képletet ismertet (Russel és Rao, Sokal és Michene, Jaccard, Yule, Csuprov) [Párniczky1976]. Számunkra a Jaccard-képlet alkalmazása tűnik célszerűnek. Két szóhalmaz (két szinonimasor) metszete azokból a szavakból áll, amelyek közösek a két halmazban, két halmaz uniója az összes szó

(azaz az összes szinonima) együttese. Így a Jaccard képlettel eredményként kapott metszet – unió arány, azaz a hasonlósági mérték (H), 0 és 1 közé esik. Képletben:

$$H = P_{11} / (P_{11} + P_{10} + P_{01}). \quad \text{(Jaccard)} \qquad (1)$$

A metszet – unió arány használatának hatásos nyelvészeti alkalmazását a szakirodalom is alátámasztja. Példaképpen a *fut ~ szalad* hasonlóságának számítása:

P_{11} (közös elemek) = 9 szó: *fut, inal, iramlik, limel, lohol, rohan, skerál, spurizik, szalad.*

P_{10} (a *fut*-ban meglévő és a *szalad*-ban nem előforduló elemek) =18 *cikázik, darizik, galoppozik, iramodik, kocog, kóstat, kotor, lót, nyargal, robog, sprintel, száguld, teker, tép, üget, vágtat, vágtázik, viharzik.*

P_{01} (a *szalad*-ban meglévő és a *fut*-ban nem előforduló elemek) = 11 szó: *fáradozik, futkos, kotrecel, lófol, lót-fut, őringel, repül, rőföl, siet, szedi a lábát, trappol.*

		fut szinonimasora	
		1	0
szalad szinonimasora	1	9	11
	0	18	

$$H(\textit{fut, szalad}) = P_{11} / (P_{11} + P_{10} + P_{01}) = 9 / (9 + 11 + 18) \approx 0.24. \qquad (2)$$

Számításaink szerint tehát a *fut ~ szalad* szavak **első** aljelentéseinek a hasonlósága a Jaccard-képlet alapján (súlyozás nélkül) számolva: 0.24.

5.2. A szójelentés-hasonlóság mérőszámának tulajdonságai

A szójelentés-hasonlóság mérőszámára (H) a következő tulajdonságok teljesülnek:

– A két szó hasonlóságát mutató mérőszám 0 és 1 közé esik:
$0 \leq H(\text{SZÓa}, \text{SZÓb}) \leq 1$;

– A szópárok jelentéshasonlósága szimmetrikus, azaz a hasonlóság nem függ a szavak sorrendjétől (ezért a hasonlósági mátrix szimmetrikus és elegendő csak az egyik felét, pl. a főátló feletti részét megfelelő módon kitöltenünk):
$H(\text{SZÓa}, \text{SZÓb}) = H(\text{SZÓb}, \text{SZÓa})$;

– Minden szó önmagával vett hasonlósága 1, ezt jelzik a hasonlósági mátrix főátlójában szereplő 1 értékek:
$H(\text{SZÓa}, \text{SZÓa}) = 1$.

Akkor mondjuk, hogy SZÓb jobban hasonlít SZÓc-re mint a SZÓa-ra, ha
$H(\text{SZÓa}, \text{SZÓb}) \leq H(\text{SZÓc}, \text{SZÓb})$.

5.3. Az elemek súlyozásának szükségessége

Párniczky Gábor [Párniczky 1976] könyve felhívja a figyelmünket arra, hogy a gyakorlatban sok esetben a vizsgált halmazok egyes elemei nem egyforma fontosak, ezért célszerű, akár szubjektív módon is – súlyok meghatározása. Mi úgy döntöttünk,

figyelembe véve, hogy a vizsgált szinonimasorokban elhelyezkedő szavak „fontossága" a címszótól távolodva csökken, hogy a szinonimasor egyes tagjait olyan szorzóval látjuk el, amely a fenti szemléletet tükrözi. Ezt indokolja a *Magyar szókincstár* már idézett előszavában olvasható szerkesztői szándék is: „A szerkesztők a szinonimasorokban helyet foglaló adatokat úgy rendezték, hogy a címszó jelentéséhez közelebb álló szinonimák a sor elején álljanak, míg a címszó jelentésétől távolabb eső szavak a sor vége felé helyezkedjenek el." Ezért egy n elemű szinonimasor i-dik tagja kísérletünkben a következő szórzósúlyt kapja:

$$W_i = (n + 1 - i) / n$$

A fenti (1)-es képletben tehát a P_{11}, a P_{10} és a P_{01} a megfelelő elemek súlyainak összegét jelenti.

A *fut ~ szalad* szópárok első aljelentésének jelentéshasonlósága a fenti módon meghatározott súlyokkal számolva:

H_w (*fut, szalad*) = $P_{11} / (P_{11} + P_{10} + P_{01}) = ((5+5.29)/2) / (((5+5.29)/2) + 8.04 + 4.92) \approx 0.28$.

6. Fogalomkörök létrehozása automatikus osztályozással

Az automatikus osztályozás olyan eljárás, ami csoportok – jelen kísérletünkben fogalomkörök – képzését hivatott elősegíteni. Esetünkben a szófajokra szétbontott jelentésmátrixok alkalmasak arra, hogy az automatikus osztályozás valamely módszerével fogalomköröket hozzunk létre. Ez annál is inkább sürgetően időszerű a magyar nyelv esetében, mert míg a magyar állandósult szókapcsolatoknak létezik korszerű fogalomköri feldolgozása [Bárdosi 2003], addig a magyar szavaknak nincs ilyen jellegű csoportosítása.

A hagyományos osztályozás jellegzetességei, melyet mi is megkövetelünk nyelvi anyagunk feldolgozásakor:

1. Átfedés mentesség: az osztályozandó halmaz elemei, csak egy osztályban szerepelhetnek;
2. Teljesség: minden elem beletartozik valamelyik osztályba;
3. Homogenitás: az egymáshoz hasonló egységek lehetőség szerint egy adott osztályba kerüljenek.

Az automatikus osztályozás végrehajtása két lépésben történik.

Először páronként vizsgáljuk meg a halmaz elemeit és ennek eredményeként:

– hasonlósági értéket számítunk ki (ezt tettük mi is), vagy

– adott relációhoz tartozó párokat jelölünk ki.

Második lépésben ezt követően valamely csoportképző algoritmus segítségével az egymáshoz közeli elemeket osztályokba rendezzük.

Esetünkben lényeges kiemelni, hogy adott szófajhoz tartozó címszónak csak adott szófajú szinonimái lehetnek. Így a jelentések számából kiindulva a 42976 x 42946 nagyságú jelentésmátrix kisebb, lényegében 3, a főnevek, az igék és a melléknevek mátrixaira bontható. A főnevek jelentésmátrixa 19118 x 19118, az igéké 15162 x 15162 és a melléknevéké 6997 x 6997 méretű. Az eredeti jelentésmátrix ezen dekomponálása az adatok feldolgozását megkönnyíti.

Így a szófajok szerint szétbontott jelentéshasonlósági mátrixok megalkotása után megfelelő módszert kell találnunk az adat mátrix mögött meghúzódó szerkezet, lényegi nyelvi tartalom felderítésére, azaz hatékony eljárást kell keresnünk az automatikus osztályozás elvégzése, a fogalomkörök generálása. A rendelkezésre álló eszközök közül a **szinguláris érték dekompozíciót** (SVD) választottuk [Kennedy1980], amivel faktorokat, illetve osztályokat is meg lehet határozni. Az SVD egyik jó tulajdonsága az, hogy nem csak négyzetes jelentésmátrixok esetén alkalmazható, hanem téglalap alakúak estén is (pl. jelentés részmátrix), ami finomabb vizsgálatok elvégzését teszi lehetővé. Az SVD differenciálgeometriai vizsgálatával foglalkozik [Rapcsák 2004].

A fent már ismertetett 48x48 méretű jelentésmátrixra elvégeztük az SVD-t és azt kaptuk, hogy a mátrix rangja 48. Tehát a vizsgált szavak jelentését, (természetszerűleg az aljelentéseket) az őket követő szinonimasorok erősen jellemzik. Az egyes szavaknak a jelentésmátrixban elfoglalt súlyát mutatják az SVD során kiszámolt szinguláris értékek:

A szinguláris értékek					
ballag1		jár2		fut1	
ballag2		jár3		fut2	
sétál1		jár4		fut3	
sétál2		jár5		fut4	
bandu-		jár6		fut5	
megy1		jár7		fut6	
megy2		jár8		fut7	
megy3		jár9		fut8	
megy4		jár10		fut9	
megy5		jár11		szalad1	
megy6		jár12		szalad2	
megy7		jár13		szalad3	
megy8		halad1		szalad4	
megy9		halad2		szalad5	
megy10		halad3		rohan1	
jár1		halad4		rohan2	

A 48 x 48 méretű jelentésmátrixban rejlő hasonlósági kapcsolatokat és ezek mértékét egy grafikus ún. dendogram alakzattal szemléltethetjük. Közreadjuk a fentiekben vizsgált 9 ige, 48 aljelentésének hasonlóságát mutató dendogramot. Az ábrán például jól megfigyelhető, hogy a „gép"-re vonatkozó *jár* 4. aljelentése és a *megy* 5. aljelentése hasonló. E természetes is, hiszen a *Magyar szókincstárban* ezek mellett a következő szinonima sorok állnak:

jár 4 = működik, üzemel, dolgozik, forog, köröz, kering, cirkulál, funkcionál, szuperál
megy 5 = jár, működik, dolgozik, közlekedik, üzemel

7. Eredményeink

Úgy véljük, kísérletünk sikeres, a magyar szavaknak, pontosabban a szavak aljelentéseinek hasonlóságát egzakt módon, eredményesen határozhatjuk meg a *Magyar szókincstárban* található szinonimasorokból kiindulva. Rámutattunk, arra hogy

ha a jelentéshasonlósági mérőszámokat mátrixban helyezzük el, ez a mátrix alkalmas kiindulási alap jelentéscsoportok, fogalomkörök automatikus képzésére.

A közreadott dendogramot szemlélve, megállapíthatjuk, hogy az anyanyelvi beszélő nyelvi kompetenciájával egyező számítási eredmény született automatikus módon.

Meggyőződésünk, hogy a módszerünk nyelvfüggetlen, a magyaron kívül más nyelvre is átvihető. Eredményesen alkalmazható nyelvi tudásbázisként az adott nyelv szinonima szótára, és hatásosan használható fel szójelentés-hasolóságok mérésére, fogalomkörök generálására.

Melléklet: A 9 szó, 48 aljelentéséből számított szójelentés-hasonlóság dendogramos ábrázolása.

Bibliográfia

[Bárdosi 2003] BÁRDOSI VILMOS (főszerk.): Magyar szólástár. Szólások, helyzetmondatok, közmondások értelmező és fogalomköri szótára. TINTA Könyvkiadó, 2003.

[ÉrtSz.] BÁRCZI GÉZA és ORSZÁGH LÁSZLÓ (főszerkesztők): A Magyar Nyelv Értelmező Szótára I–VII. Akadémiai Kiadó, Budapest, 1959–1961.

[Fellbaum 1998] Christiane Fellbaum (ed): WordNet: An Electronic Lexical Database. Cambridge, MIT Press, 1998.

[Hideki, Teiji 1993] HIDEKI KOZIMA and TEIJI FURUGORI: Similarity Between Words Computed by Spreading Activation on an English Dictionary. Proceedings of EACL-93. 232–239, 1993.

[Hirst 1998] HIRST G. and ST. ONGE D.: Lexical Chains as representations of context for the detection and correction of malapropisms. In Fellbaum 305–332. 1998.

[Jiang, Conrath 1998] J. JIANG and D. W. CONRATH: Semantic Similarity Based on Corpus Statistics and Lexical Taxonomy. Proceedings of the 10th International Conference: Research on Computational Linguistics (ROCLING X), 19–33, 1998.

[Kennedy 1980] W. J. KENNEDY and J. E. GENTLE: Statistical computing, Marcel Dekker, New York, Basel, 1980.

[Kiefer 2000] KIEFER FERENC: Jelentéselmélet. Budapest, 2000.

[Kirkpatrick 1998] BETTY KIRKPATRCK: Roget's Thesaurus of English Words and Phrases. Harmondsworth, Middlesex, Penguin, 1998.

[Kiss 1998] KISS GÁBOR (főszerkesztő): Magyar szókincstár. Rokon értelmű szavak, szólások és ellentétek szótára. TINTA Könyvkiadó, Budapest, 1998\1

[Kiss, Kiss 2004] KISS GÁBOR és KISS MÁRTON: Kísérlet egy szócsoport elemei jelentéshasonlóságának meghatározására. In.: „...még onnét is eljutni a túlra..." Nyelvészeti és irodalmi tanulmányok Horváth Katalin tiszteletére. 159–165. oldal. TINTA Könyvkiadó, 2004.

[LDOC 1987] Longman Dictionary of Contemporary English. Longman, Harlow, Essex, new edition, 1987.

[Leacock, Chodorow 1998] LEACOCK C. and CHODOROW M.: 1998. Combining local context and WordNet similarity for word sense identification. In Fellbaum 265–283. 1998.

[Lin 1998] LIN D.: An information-theoretic definition of similarity. In Proceedings of the 15th International Conference on Machine Learning, Madison, WI.

[Manabu, Takco 1994] MANABU OKUMURA and TAKO HONDA: Word sense Disambiguation and Text Segmentation Based on Lexical Cohesion. In Proceedings of COLINGS-94. Vol. 2, 755–761, 1994.

[Miller 1990] MILLER, G. A.: WordNet: An on-line lexical Database. International Journal of Lexicography, 3/4, Special Issue, 235–312. 1990.

[Miler 1971] GEORGE A. MILLER: Empirikus módszerek a szemantika kutatásában. Fordította: Siklay István. In: Pszicholingvisztika és kommunikációkutatás. Szöveggyűjtemény. Válogatta ás a bevezetőt írta: Pléh Csaba. Tömegkommunikációs Kutatóközpont, Budapest, 1977. (Empirical methods in the Study of Semantics. In.: Danny D. Steiberg és Leon A. Jakobovits (szerk.): Semantics: An interdisciplinary reader in philosphy, linguistics and psychology. London, Cambridge University Press, 1971, 569–585.)

[Morris, Hils 1991] J. MORRIS and G. HIRST: Lexical cohesion computed by thesaural relations as an indicator of the structure of text. Computational Linguistics, 17., 21–48, 1991.

[Martinkó 2001] MARTINKÓ ANDRÁS: A szó jelentése. Lazi Könyvkiadó, Szeged, 2001.

[Osgood 1952] C. E. OSGOOD: The natural and measurement of meaning. Psychological Bulletin, 49, 197–237, 1952

[Osgood, Succi, Tannenbaum 1957] C. E. OSGOOD and G. J. SUCCI and P. H. TANNENBAUM: The Measurement of Meaning. University of Illinois Press, Urbana, 1957

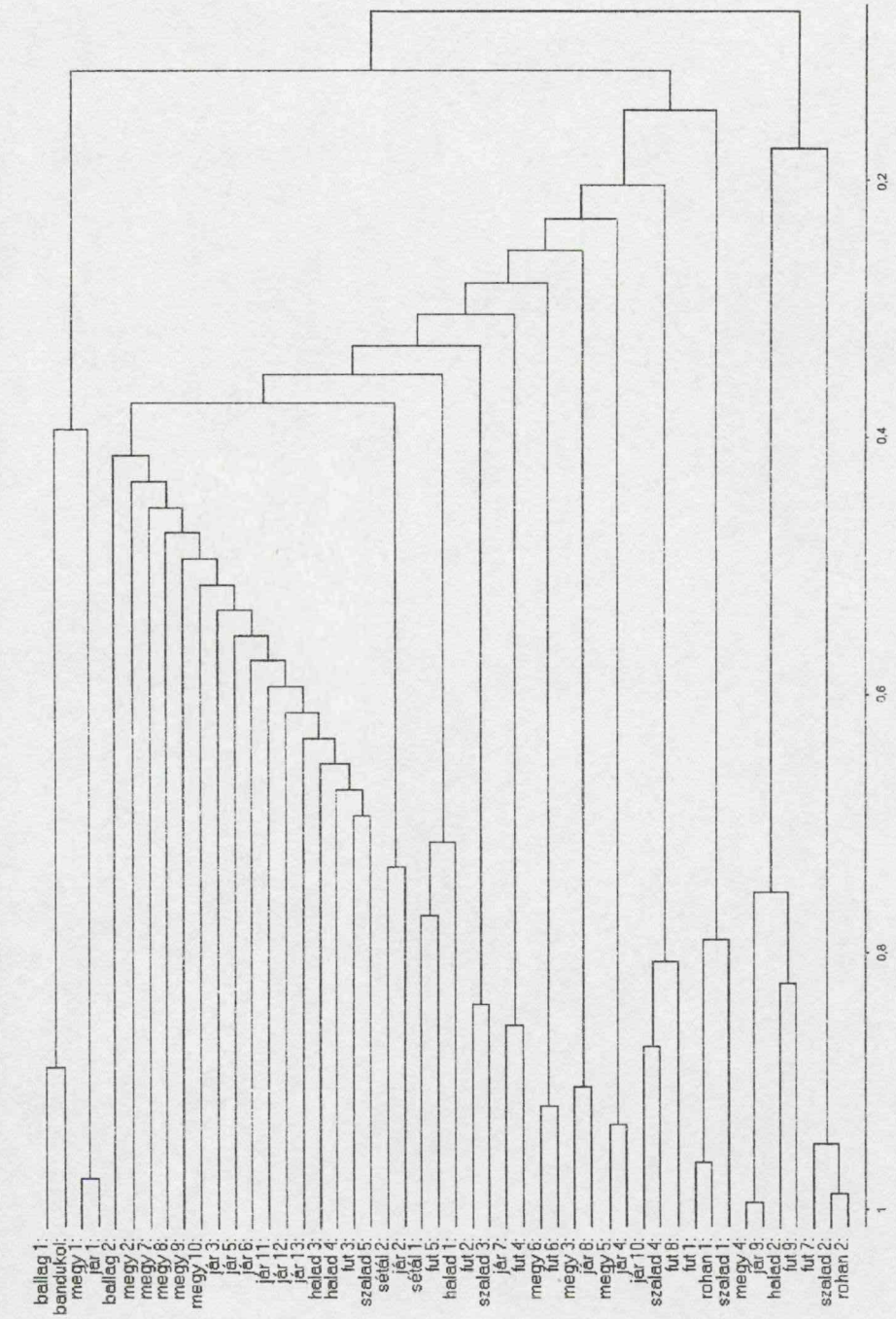
[Párniczky 1976] PÁRNICZKY VIKTOR: A statisztika alapjai. Statisztikai Kiadó Vállalat, A korszerű informatika könyvtára 8. 1976.

[Rapcsák 2004] RAPCSÁK TAMÁS: Some optimization problems in multivariate statistics, *Journal of Global Optimization* 28 (2004) 217–228.

[Resnik 1995] RESNIK P: Using information content to evaluate semantic similarity. In Proceedings of the 14th International Joint Conference on Artificial Intelligence, pages 448–453, Montreal. 1995.

[Varga 1969] VARGA DÉNES: Információs tezauruszok készítésének módszertana. Országos Műszaki Könyvtár és Dokumentációs Központ. Budapest, 1969.

[Villó, Kiss 1996] VILLÓ ILDIKÓ és KISS GÁBOR: Mozgást jelentő igék szinonimitásának vizsgálata. In.: Emlékkönyv B. Lőrinczy Éva hetvenedik születésnapjára. Szerk.: Bánki Judit. 123–128. oldal. MTA Nyelvtudományi Intézete, Budapest, 1996


A ballag, sétál, bandukol, megy, jár, halad, fut, szalad, rohan szavak 42 aljelentésének hasonlósági ábrázolása
ballag 1.
bandukol.
megy 1.
jár 1.
ballag 2.
megy 2.
megy 7.
megy 8.
megy 9.
megy 10.
jár 3.
jár 5.
jár 6.
jár 11.
jár 12.
jár 13.
halad 3.
halad 4.
fut 3.
szalad 5.
sétál 2.
jár 2.
sétál 1.
fut 5.
halad 1.
fut 2.
szalad 3.
jár 7.
fut 4.
megy 6.
fut 6.
megy 3.
jár 8.
megy 5.
jár 4.
jár 10.
szalad 4.
fut 8.
fut 1.
rohan 1.
szalad 1.
megy 4.
jár 9.
halad 2.
fut 9.
fut 7.
szalad 2.
rohan 2.
1
0,8
0,6
0,4
0,2