

HANGKIVETŐ FŐNEVEK HASONLÓSÁGI CSOPORTJAI

Rung András

Bevezetés

Vizsgálatom tárgyát azon 1091 hangkivető főnév sajátos viselkedése képezi, amelyekről van gyakorisági adatunk a *Szószablya Gyakorisági Szótár* (Halácsy és mtsai 2003) alapján. Rung (2011) a *Szószablya Gyakorisági Szótár* adatait felhasználva bemutatta, hogy ezek a hangkivető főnevek nem egységesen viselkednek. 11%-uk kisebb (például *öböl*, *fátyol*) vagy nagyobb (például *bajusz*, *tegez*) mértékben már nem követi a hangkivető séma által elvárt viselkedést, mivel időnként megtartják utolsó magánhangzójukat az olyan esetekben is, amikor a szabályalapú megközelítés a hangkivető változatok megjelenését várná el (például *öbölt*, *fátyolon*, *bajuszom*). Ezen szavak viselkedését jobban megérthetjük, ha egy analógiás megközelítést alapul véve hasonlósági viszonyaik segítségével kialakított csoportokban tanulmányozzuk őket. Egy ilyen analógiás megközelítésben az egyes szavak viselkedését, ha azok nem nagyon gyakoriak, mert ebben az esetben alakjaik memorizáltak, az olyan hozzájuk hasonló szavak viselkedése és azok gyakorisága határozza meg, amelyekkel egy hasonlósági csoportba sorolhatók.

A szavak hasonlóságát saját fejlesztésű algoritmusokkal határozom meg, amelyek a kurrens hasonlóság mérésére használt algoritmusoknál (Skousen és mtsai 2002) finomabb összehasonlításokat is lehetővé tesznek. A komplex jegymérték és a komplex tengelymérték nevű algoritmusok a szavak hasonlóságát azok jobb szélétől véve számítják ki úgy, hogy a megfeleléseknek, hasonlóságoknak egyre kisebb súlyt adnak a szavak bal széle felé haladva. Így például a *vas* és *sas* szavakat mind a két számítógépes algoritmus hasonlóbbnak tekinti, mint a *vas* és a *vaj* szavakat. Az algoritmusok a hasonlítást az egyes fonémák jegyei alapján végzik el, de a komplex jegymérték (Rung 2008, Rung 2009) fonémákat hasonlít össze, míg a komplex tengelymérték az egyes jegyek tengelyeinek hasonlósága alapján számítja ki két szó hasonlósági értékét. A komplex jegymérték esetében két fonéma nemcsak azonos vagy eltérő lehet, hanem az analógiás nyelvi megközelítéssel összhangban több, bár diszkrét fokozatban adható meg hasonlóságuk. A fonémák kiválasztásában, jegyeik és azok lehetséges értékeinek meghatározásában Kiefer (1994), illetve Siptár és Törkenczy (2000) leírásaiból indultam ki.

A fonémák hasonlóságának mértékét megkülönböztető jegyeik alapján számolom, amelyek több értéket is felvehetnek, nem ragaszkodom azok szigorúan bináris voltahoz. A magánhangzók esetében a nyíltságot, ajakkerekítést, hosszúságot, elölképzettséget, a mássalhangzók esetében pedig a zöngésséget, a zörejhang voltot, a hangképző szervet (aktív artikulátor), a képzés helyét és módját veszem figyelembe. Az összehasonlított fonémák hasonlóságát 2-vel osztom el minden eltérő jegy esetén (kiindulási érték: 1). A mássalhangzók és a magánhangzók egymáshoz viszonyított hasonlósága rendszeremben 0, nincsenek közös jegyeik. Ezek alapján az /o/ fonéma hasonlóságának mértéke egy másik /o/ fonémához 1, az /ö/-höz és az /ó/-hoz 0,5 ($1:2^1$, mivel egy jegyben, az elölképzettségben, illetve a hosszúságban különböznek), míg az /ő/-höz 0,25 ($1:2^2$ mert két jegyben, az elölképzettségben és a hosszúságban különböznek).

A már említett jegyeken túl alkalmazok módosító jegyeket is, amelyek ezek értékeihez rendelnek további értékeket, ezért önállóan nem értelmezhetőek. Így a *folly* (folyamatos) jegy a *mód* jegy *folly* értékéhez rendeli a *rés* (rés hang), *app* (approximáns), *lat* (laterális), *per* (pergőhang) további értékeket. Hasonlóan a *zár* (zárhang) jegy a *mód* jegy *zár* értékeihez rendeli a *fel* (felpattanó), *aff* (affrikáta), *egyp* (egyperdületű) értékeket. Ezzel a megoldással lehetővé teszem, hogy felosztásom finomabb legyen, és bizonyos fonémák rokon vonásai (pl. /r/, /s/ folyamatossága) felismerhetőek legyenek a rendszer számára, miközben különbségeiket se mosom el (/r/ pergőhang, /s/ réshang). Korábbi tesztjeim megmutatták (Rung 2008), hogy egy meglehetősen durva hasonlítási módnál az egyszerű jegymérték nevű algoritmus finomabb osztályozásának köszönhetően jobban tudtam kezelni az adott nyelvi feladatot, így elképzelhetőnek tartottam, hogy egy még árnyaltabb felosztás további eredményjavulást hozhat. Ha ezen módosító jegyek valamelyike az összehasonlítandó fonémák egyikére nem volt alkalmazható, akkor azt a jegyet kihagytam az összehasonlításból, hisz a jegy ebben az esetben csak olyan eltérést határozott meg, amelyért más jegy esetében már csökkentettem a hasonlóságot. Például a *k:ny* pár a „mód” jegy esetében eltérő értéket vesz fel, így a zárhangokat tovább osztó „zár” jegy alapján nem növelem ezek távolságát, mert a jegy az *ny*-re nem alkalmazható.

A magyar nyelvleírás hagyományát és a pszicholingvisztikai kutatási eredményeket (Slobin 1973, Bybee és Moder 1983, Lukács 2002: 47) követve fontosságot tulajdonítok a szavégek hasonlóságának, amelyek kiemeltebb szerepét összetett szavak hasonlításában Krott (2009: 121) is igazolta. Természetesen nem minden esetben a szóvég felépítése a döntő, hanem ez nyelvenként és feladatonként különböző lehet, így Albright (2009: 204) tesztjeiben a

kitalált spanyol igéknél a hangsúlyos *o* diftongizációjában az utolsó előtti szótagnak volt kiemelt jelentősége, mivel a vizsgált folyamatok is oda voltak köthetőek.

Algoritmusom számításában a fonémák hasonlóságának a súlya a szó végétől a szó eleje felé logaritmikusan csökken. 1,8-as alapú logaritmust használok, mivel korábbi vizsgálataimban ez bizonyult a leghatékonyabbnak (Rung 2008). Az 1,8-as alapú logaritmussal való számolás megfelelő prominenciát ad az utolsó néhány fonémának, de még a szóalak belsejében lévő hasonlósági hatások érvényesülésének is lehetőséget nyújt. A szavak önmagukhoz vett hasonlósági értéke 1, a tőlük teljesen eltérő szóhoz vett pedig 0^1 . Programom számítása alapján a *bab* és a *púp* hasonlósága a következőképp alakulna:

$$(1) \quad \begin{array}{ll} b:p = 0,5 & (\text{eltérő jegy: zöngésség}) \\ a:ú = 0,25 & (\text{eltérő jegy: nyíltság, hosszúság}) \\ b:p = 0,5 & (\text{eltérő jegy: zöngésség}) \end{array}$$

A hasonlóság az alábbi módon számítható ki, a logaritmikusságot is figyelembe véve²:

$$(2) \quad \frac{0,5 * 1 + 0,25 * 2 + 0,5 * 4}{7} = 0,43$$

A szavak hasonlítása során a magánhangzó-harmóniát a már korábban is tesztelt egyszerű jegymérték alapján működő algoritmus kevésbé tudta megragadni (Rung 2008, 2009), mivel az utolsó előtti magánhangzónak már kis súlyt ad. Hasonlóan más, csak magánhangzók vagy csak mássalhangzók közötti összefüggések felismerésére is kevésbé alkalmasak az eddig bemutatott hasonlítási módok, ezért egy olyan algoritmust is kidolgoztam Kálmán Lászlóval közösen, amely az összehasonlítást eltérő tengelyenként végzi. Az egyes tengelyekre külön számít hasonlóságot, majd ezt összegzi, tehát a *bika* és *dara* szavak összehasonlítása során a *bk:dr* és *ia:aa* szekvenciák tengelyeit, illetve a *CV* tengelyt vetem össze³. Az algoritmus ugyanazokkal a jegyek-

¹ Ez az érték a valós modellezésben nem gyakori, hisz a hangkivető főnevek esetén a legjobban különböző *eper* és *szeméremajak* hasonlósága is 0,11. Hangkivetők és nem hangkivetők közt azonban már előfordul a 0 érték: pl. *bögyöly:abbreviátúra*.

² A könnyebb átláthatóság kedvéért a számításban 2-es alapú logaritmust alkalmaztam, ami azonban a példa lényegén és a számítás módján nem változtat.

³ A komplex tengelymérték összesen 13 tengelyen számolja a hasonlóságot, de ezek vagy mássalhangzós vagy magánhangzós, egymástól független tengelyek. Az egyetlen közös tengely az, amely az egyes szegmentumokról rögzíti, hogy azok mássalhangzók vagy magánhangzók-e.

kel működik a független tengelyeken, mint a komplex jegymérték alapján számító algoritmus.

Ez a hasonlítási mód annyiban hasonlít az autoszegmentális fonológia (Goldsmith 1990) megközelítéséhez, hogy az egyes jegyek független tengelyeken helyezkednek el, azonban nincs benne egy-többhöz vagy több-egyhez kiosztás. Minden csomóponthoz egy tengelyen egy jegy tartozik, amely akár ismétlődhet is, azaz a Kötelező Kontúr Elvének (KKE, *Obligatory Contour Principle*, OCP) való megfelelés sincs beépítve a rendszerbe. Az egyes tengelyek közti kapcsolatokat elméleti megfontolásoktól függetlenül nem definiálok, azok teljesen önállóan léteznek a reprezentációban

Az algoritmusok segítségével meghatározható hasonlósági csoportok elemzésében egy új, vagy csak alig használt megközelítést alkalmazok, mivel úgy gondolom, hogy a nyelvészet dinamikus fejlődésének elősegítéséhez érdemes olyan módszereket és vizualizációs eljárásokat is bevonni kutatásunkba, amelyeket más tudományágak, mint például a biológia (Enfield 2008) vagy a fizika (Bíró 2006) alkalmaznak. Ha ezek eszköztárát használatba vesszük, akkor adatainkat más megvilágításba helyezve számos új felfedezést tehetünk. Ilyen, eddig kevésbé alkalmazott eljárás a nyelvi adatok gráfstruktúrában való tanulmányozása is, amely a nyelvtechnológiában bevett megközelítési mód, de használata a szorosan vett elméleti indíttatású nyelvészeti kutatásokban elsősorban csak a kognitív nyelvészet területére korlátozódott. Így az analógiás nyelvészetben is csak viszonylag kis szerep jutott neki elsősorban más, nem morfofonológiai jellegű vizsgálatokban (Duvignau–Gaume 2004).

A szavak viszonyainak gráfstruktúrában való tanulmányozása kvalitatív megközelítési mód. Közvetlenül ennek alapján nem tudjuk leírni, sem megjósolni a szavak analógiás, még kevésbé szabályalapú viselkedését. A táblázatos megjelenítéssel és áttekintéssel szemben nagy előnye azonban, hogy egyszerre látjuk a szavak tulajdonságait és kapcsolatainak szövetét, ami számos olyan felismeréshez vezethet, amelyek később beépíthetőek lesznek a formálisabb leírásokba, vagy a jelenségeket modellező algoritmusok javítására használhatjuk fel őket.

A szavak viszonyait és azok számszerűsíthető tulajdonságait a Cytoscape 2.6. gráfvizualizációs programmal jelenítettem meg. A szavak kapcsolatainak erősségét (hasonlóságukat) első lépésben a végek hasonlóságának a komplex tengelymértéknél nagyobb súlyt adó komplex jegymértékkel határoztam meg. A gyengébb kapcsolatokat a struktúrák könnyebb áttekinthetősége és az adatok hatékony kezelése érdekében kihagytam. Az elemzésben a legjellemzőbb részletek bemutatására szorítkozom terjedelmi okokból, hisz az adatstruktúra

1091 elemet tartalmaz, amelynek teljes bemutatásra nincs mód. A komplex jegymérték segítségével készített gráf elemzését kiegészítem olyan részgráfok vizsgálatával is, amelyeket a komplex tengelymérték alapján számított hasonlósági értékek mentén határoztam meg. Ezek esetében csak az olyan struktúrákat tekintem át, amelyek érdemleges eltérést mutatnak a komplex jegymérték számai alapján létrehozottaktól. A következőkben bemutatom, hogy az egyes élek és csomópontok mit jelenítenek meg az elemzésben felhasznált ábrákon.

Élek

A gráfokban az élekkel a szavak 0,9–1-ig terjedő⁴ egymáshoz való hasonlóságát jelenítettem meg. Minél vastagabb és sötétebb egy él, annál közelebbi, szorosabb hasonlósági viszonyt jelez két szó között. A szavak önmagukhoz való hasonlóságát annak redundáns volta miatt nem ábrázoltam a gráfokon, így a legszorosabb kapcsolatokat olyan szavak reprezentálják, amelyeknek hasonlósága nagyobb, mint 0,995, de kevesebb, mint 1: például *halálveszedelem-veszedelem*.

Csomópontok

A csomópontok egy-egy szónak felelnek meg, amelyeket a viszonyukat kifejező élek kötnek össze. Az egyes szavakat (csomópontokat) gyakoriságukkal és hangkivetési mértékükkel⁵ jellemzem, azaz hogy a hangkivetéssel együttjáró toldalékok előtt milyen arányban figyelhetünk meg valóban hangkivető toldalékokat esetükben. A csomópont mérete az adott szó összes olyan toldalékos alakjának *Szószablya korpusz*beli gyakoriságát jeleníti meg, amelyben hangkivetést várnánk el. A leggyakoribb *dolog* átmérőjének felével a sorban 50. *izgalom* szerepel. A csomópontok elütő árnyalatú keretének vastagsága azt jelzi, hogy a *Szószablya Gyakorisági Szótár* alapján mennyire ingadozik az adott szó a hangkivetéssel együttjáró toldalékos alakjaiban. Az átlagos mértékben, 97,57% százalékban a hangkivető mintát követő szavakat vékony keret övezi, míg az erőteljesen ingadozó szavakat egyre vastagabb. Egy szó minél sötétebb, annál több 0,9-nél nagyobb hasonlósági kapcsolata van.

⁴ Így a *marok-burok* 0,90612-es értékkel még kapcsolódik egymáshoz a gráfon, de a *burok-tulok* (0,88846) már nem.

⁵ A hangkivetési mérték olyan mutató, amely megadja, hogy egy szó alakjai hány százalékban szerepelnek hangkivető változatban (pl. *sátrat*) olyan esetekben (tárgyeset, szuperesszívusz, többes szám, birtokos személyragok), amikor a hangkivető szavaknál hangkivetést várnánk el. Ha ennek értéke 100%, akkor a korpuszadatok alapján az adott szó minden esetben hangkivető módon viselkedik a hangkivetést elváró toldalékos alakjaiban.

A szövegben lévő összes adat a *Szószablya Gyakorisági Szótáron* alapzik, amennyiben azt külön nem jelzem másképp. A szavak mögött zárójelekben megadott számok azok gyakoriságát mutatják a hangkivetést elváró toldalékos alakjaik vonatkozásában, míg a %-jeles adatok a szavak hangkivétési mértékét adják meg az ilyen esetekben.

1. A komplex jegymérték alapján számított kapcsolatok

A 0,9-nél erősebb hasonlósági viszonyok alapján 50 összefüggő csoportot lehetett elkülöníteni, amelyeket az 1. táblázat mutat be. Egy csoportba egy szó akkor tartozik, ha legalább a csoport egy tagjához 0,9-es vagy annál nagyobb mértékben hasonló. A hangkivető szavak csoportjainak mérete és jellege nem egyforma. A szavak közel fele (488 szó) a két legnagyobb csoportba tartozik (*-alom*, *-elem* végűek), amelyeket gráfok alapján nem elemzünk, mert homogén struktúrájukban ezzel a módszerrel nem lehet értelmes megfigyeléseket tenni. Ezeken túl még 8 közepes méretű, legalább húsz szót tartalmazó és 40 kisebb csoport van. Megfigyelhető, hogy a kiugróan nagy gyakoriságú szavak nemcsak a nagy csoportokban találhatóak. Ezeknek a prototípusként⁶ funkcionáló szavaknak köszönhető egy-egy kisméretű szóbokr hangkivetőként való megmaradása (pl. *titok* és csoportja). A csoportok leggyakoribb, prototípusnak tekinthető szavai alapján viszonylagosan jól tudjuk jósolni egy szó viselkedését, mivel az ilyen prototípusokhoz való hasonlóság közepesen erős együttjárásban ($r(280) = 0,4$, $t = 7,31$, $p > 0,001$) van egy szó esetében a hangkivetés mértékével. A csoporton belüli prototípusok lokálisan gyakoriak, de nem azonosak a leggyakoribb szavakkal, hisz az 50 csoport alapján meghatározott prototípusokból csak 13 van a leggyakoribb 50 hangkivető főnév közt (*cukor*, *dolog*, *figyelem*, *haszon*, *lélek*, *méreg*, *pokol*, *szobor*, *tartalom*, *teher*, *titok*, *torony*, *tükör*). A nagy gyakoriságú szavakat tartalmazó csoportokban csak akkor figyelhető meg alacsonyabb hangkivétési mérték a kisebb gyakoriságú szavak esetében, ha már a gyakori szó sem stabilan hangkivető (pl. *sátor*, *fátyol*). Ettől csak akkor figyelhetünk meg eltéréseket, ha a csoporton belüli hasonlósági viszonyok nem kiegyenlítettek (pl. *pokol*, *cukor* csoportja). A több gyakori szót tartalmazó csoportok stabilan hangkivetők. Az egyes szavak kapcsolatainak száma enyhe pozitív korrelációban van hangkivétésük mértékével ($r(803) = 0,136$, $t = 3,9$, $p > 0,001$),

⁶ Prototípus alatt egy olyan központi tagot értek, amely az adott kategória kiugróan reprezentáns képviselője, s így az egész csoportot képviselheti (Hallan 2001: 91, Kálmán–Rung 2010).

Hangkivető főnevek hasonlósági csoportjai

amely összefüggés némileg szorosabb, ha csak a hangkivetést 99%-nál kevésbé követő szavakat nézzük ($r(101) = 0,23$, $t = 2,33$, $p > 0,05$)⁷.

1. táblázat: A hangkivető szavak csoportjai a komplex jegymértékkel számított hasonlósági viszonyaik alapján.

Csoport leggyakoribb szava	Csoport példánygyakorisága	Szavak száma a csoportban	Hangkivetés mértéke a csoportban	Csoporton belüli átlagos hasonlóság
figyelem	1272143	222	99,60%	0,902
tartalom	1240569	266	99,89%	0,926
dolog	567310	6	99,98%	0,964
lélek	214755	17	99,61%	0,911
sarok	131170	55	99,39%	0,882
kapocs	107195	13	99,64%	0,961
títok	107132	11	100,00%	0,953
haszon	62988	20	98,09%	0,939
teher	61856	8	96,42%	0,961
szobor	42712	12	99,91%	0,955
tükör	40301	32	99,28%	0,942
izom	32762	25	99,97%	0,925
karom	31719	14	98,87%	0,892
cukor	29146	58	99,87%	0,916
méreg	28779	44	99,73%	0,922
ajak	25894	5	98,81%	0,932
torony	23529	35	99,56%	0,954
gyomor	19094	7	92,85%	0,964
sátor	16467	11	85,34%	0,930
fogoly	15179	10	98,93%	0,922
kölyök	12258	7	99,88%	0,961

⁷ Az összehasonlításból kihagytam az *-alom* végűek csoportját, mivel a Cytoscape 2.6 nem boldogult kapcsolati fokaik kiszámolásával, mert azok túlzottan számosak voltak. Ebben a csoportban egy szóra összesen 126 kapcsolat jutott, míg a megvizsgált 805 szó esetében szavanként 23. Ez alapján valószínűsíthető – mivel az *-alom* végűek csoportja stabilan hangkivető –, hogy ezt az összefüggést inkább támogatnák, semmint gyengítenék adataink. Ezt a vélekedést erősíti meg, hogy a komplex tengelymérték esetében is elvégeztem ezt a számítást, ahol ugyanezt az összefüggést lehetett megállapítani az összes hangkivető szó figyelembevételével. A komplex tengelymérték alapján a számításokat az összes hangkivető főnévvel kapcsolataik kisebb száma miatt lehetett elvégezni.

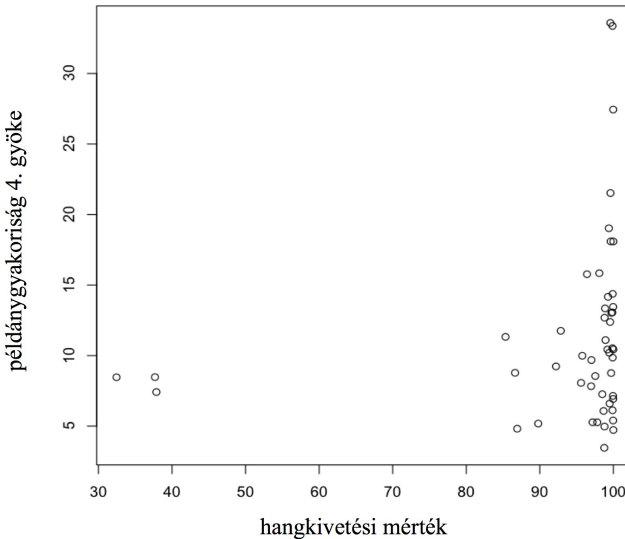
Rung András

fészek	11860	30	100,00%	0,957
vétek	11835	6	99,20%	0,915
csücsök	10837	7	99,45%	0,931
köröm	9950	10	95,78%	0,945
gödör	9436	20	99,91%	0,948
majom	8792	13	97,02%	0,907
meder	7265	11	92,20%	0,937
lepel	5937	3	86,64%	0,972
kehely	5885	9	99,70%	0,945
pokol	5349	5	97,54%	0,919
fátyol	5158	7	37,68%	0,961
bajusz	5132	5	32,45%	0,965
ököl	4229	3	95,61%	0,942
öböl	3759	3	96,98%	0,949
vacak	3021	2	37,87%	0,975
selyem	2778	8	98,50%	0,950
mocsok	2608	4	99,95%	0,940
horog	2297	5	99,97%	0,967
kölök	1894	4	99,50%	0,944
eper	1405	5	99,89%	0,930
szatyor	1361	3	98,67%	0,968
piszok	850	4	99,97%	0,967
kapor	773	3	97,18%	0,935
fodor	771	9	97,81%	0,954
tulok	719	4	89,79%	0,923
pocok	610	3	98,80%	0,943
kazal	537	3	86,93%	0,973
töbör	498	2	100,00%	0,983
pecek	143	2	98,77%	0,978

Az ingadozó csoportok típus- és példánygyakoriságuk alapján is a kisebbek közé tartoznak⁸. Az átlagos csoportméret 21 szó, míg az átlagos ingadozó csoportban 6 szó található. Az átlagos csoport 84 ezer alak-előfordulást számlál, az átlagos ingadozó 10 ezret. A csoporton belüli átlagos hasonlóság (a csoporttagok egymáshoz való hasonlósági mértékének átlaga) mértékében azonban nincs eltérés, az ingadozó szavak csoportjainak értékei hasonlóak a

⁸ Természetesen vannak a stabilan hangkivető csoportok közt is kis példány- és típusgyakoriságúak (1. ábra)

következetesen hangkivetőkéhez. Ez azonban annak tudható be, hogy bizonyos ingadozó csoportok nagyon szorosan összetartanak (a csoport tagjai nagyon hasonlítanak egymáshoz is) egy már ingadozó prototípus körül (pl. *bajusz*, *fátyol*, *kazal* csoportja), vagy az átlagosnál jóval fragmentáltabbak (pl. *pokol* csoportja), így az ingadozást az alacsony csoportösszetartás támogatja. Érdeemes megjegyezni, hogy az igen nagy méretű csoportok a kisméretű csoportok értékeinél kisebb, de még így is meglehetősen nagy átlagos hasonlósággal bírnak. A belső hasonlóság minden esetben messze meghaladja a hangkivető főnevek átlagos egymáshoz mért 0,409-es hasonlóságát.



1. ábra: A hasonlóság alapján kialakított csoportok hangkivetési mértéke a példánygyakoriság függvényében

A komplex jegymérték alapján 21 szónak nincsenek a 0,9-es hasonlósági értéknél közelebbi kapcsolatai: *ászkok*, *átok*, *berek*, *bögöly*, *boholy*, *bugyor*, *bürök*, *bütyök*, *cseber*, *iker*, *jászol*, *kebel*, *koboz*, *pityer*, *pöcök*, *pucor*, *sulyok*, *szutyok*, *takony*, *tegez*, *üszök*⁹. E magányos szavaknak a *Google Gyakorisági*

⁹ Az algoritmus jellegéből következik, hogy csak 2 szótagos szavak lehetnek magányosak, hisz az összetett szavak esetén az egymáshoz való hasonlóság ennél a küszöbértéknél mindig nagyobb. Ugyanez áll a dominánsan legalább három szótagos *-alom*, *-elem* végű szavakra is.

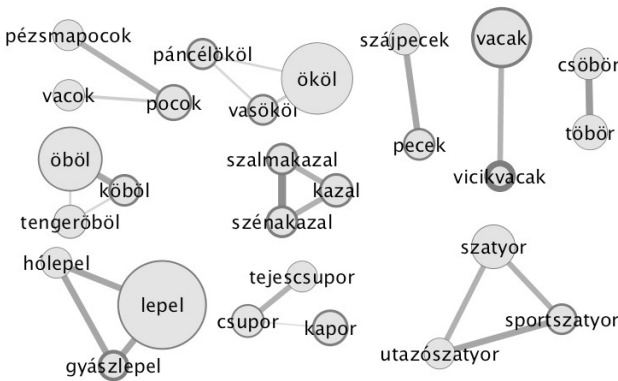
Gyűjtés¹⁰ alapján már 57,15%-a (12 alak) az átlagosnál jobban ingadozik (A *Google Gyakorisági Gyűjtés*ben 96,97% a hangkivetés átlagos mértéke). Az átlagosnál nem jobban ingadozó szavak felépítésükben hasonlóak: *átok, boholy, bugyor, bürök, bütyök, pöcök, sulyok, szutyok, üszök*. A *sulyok* és a *boholy* szavakat érdemes kizárnunk áttekintésünkéből, mert a *sulyoknál* a tulajdonnévi adatokkal való keveredés lehetősége rontja az adatok megbízhatóságát, a *boholy*nál pedig annak önálló szó státusza kétséges.

Ha a távolabbi, 0,8-as hasonlósági viszonyokat is figyelembe vesszük a két eltérően viselkedő csoport kapcsolataiban, akkor már találunk eltéréseket. Az ingadozó szavakat tartalmazó csoportban az egy szóra jutó 0,8-as hasonlósági kapcsolatok száma 5, míg a másik csoportban a *sulyok* és a *boholy* szavakat figyelmen kívül hagyva 12,14, azaz a nem ingadozó magányos szavaknak 2,42-ször annyi közepesen erős kapcsolatuk van más hangkivető főnevekhez. A 12 ingadozó magányos szó átlagos távolsága az összes hangkivető főnévtől 0,282, míg a 7 nem ingadozó magányos szó átlagos távolsága 0,28. Az összes hangkivető főnévhez viszonyított hasonlóságukban a páros t-próba nem mutatott ki szignifikáns eltérést. Ha azonban a magányos szavak értékeit összehasonlítjuk a hangkivető főnevek egymáshoz mért hasonlóságának átlagával (0,409), akkor láthatjuk, hogy a szegényes közeli kapcsolatokkal rendelkező szavak esetében ez a szám lényegesen alacsonyabb. Ezek alapján kijelenthetjük, hogy egy szó viselkedésére közvetlen szomszédai nagyobb hatással vannak, hisz ezek jellege differenciál a magányos szavak közt viselkedésükben, de ezt a hatást erősítheti a paradigma összes szavától való markáns formai különállás is.

Elvárásaim alapján a hasonlósági csoportokban a kisebb gyakoriságú szavak közt találhatunk olyanokat, amelyek a csoport leggyakoribb szavának hangkivetési mértékénél kisebb hangkivetési mértékkel rendelkeznek, különösen azokban az esetekben, amikor a csoporthoz való kapcsolódás gyenge, csak egy kapcsolaton keresztül realizálódik. A komplex jegymérték alapján számított hasonlósági gráfban elkülöníthető csoportok áttekintése és a közeli kapcsolataikban szegényes szavak viselkedésének tanulmányozása után a 2-3 elemű részgráfokat veszem szemügyre (2. ábra). Ezek igen nagy számban tartalmaznak kevésbé hangkivető módon viselkedő szavakat (hangkivetés átlagos értéke: 91%), ami annak köszönhető, hogy ezek a szavak is gyengén kapcsolódnak a többi hangkivető szóhoz. E kisméretű csoportokban megfigyelhető az a tendenciajellegű jelenség, miszerint egy kisebb gyakoriságú szó

¹⁰ 2010-es saját gyűjtés a *Google* kereső segítségével a vizsgált szavak vonatkozásában, elérhető a https://docs.google.com/viewer?a=v&pid=explorer&chrome=true&srcid=0BxQBWau_71U7Y2E3MGUzN2QrY2I2OS00Mzc1LThjZWUtZmF1YjU4N2ZkMzBm&hl=hu webcím alatt.

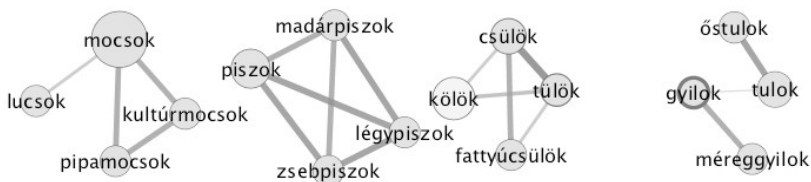
kevésbé következetesen hangkivető, mint a csoport nála jelentősen gyakoribb tagja: *vacak* (3018; 75,7%) – *vicikvacak* (3; 0%); *öböl* (3401; 98,7%) – *köböl* (108; 92,6%)¹¹; *ököl* (4125; 100%) – *páncélököl* (71; 93%) – *vasököl* (33; 93,9%); *szatyor* (1325; 99,8%) – *sportszatyor* (26; 96,2%). Az ettől a tendenciától eltérő *pézsmapocok*, *tejescsupor*, *hólepel*, *szájpecek*, *utazószatyor* szavak kevesebb, mint 25 alakot számlálnak, ezért adatainkat fenntartással kell kezelnünk, lehet, hogy látszólagos hangkivető viselkedésük mögött adathiány áll. Kivételt képez e tendencia alól a *pocok* is, ami azonban a *pocok*+E/3. birtokos nem hangkivetőként viselkedő *pocokja* alakjának tudható be, valamint a *tengeröböl*, amely a gyakoribb *öböl*-nél jobban követi a hangkivető mintát. A *kazal* csoport szavainak (*szénakazal* 82%:125, *kazal* 91%:310, *szalmakazal* 88%:102) homogén viselkedése mögött a csoporton belüli közel azonos gyakoriság és a csoporttagok jelentésbeli és alaki közelsége áll (a legalább 3 elemű csoportokból itt a legmagasabb az átlagos hasonlóság). A *töbör*-*csöbör* pár tagjai a hangkivetőkre jellemző tővégűeknek és -ő-ő magánhangzó-szekvenciájuknak köszönhetik stabilitásukat. A *kapor* ingadozása az ékezetmentes *kapóra* alakoknak tulajdonítható, bár tárgy esetben már csak 98,5%-ban követi a hangkivető sémát. A *csupor* enyhe ingadozása viszonylagos magányosságának, valamint a kevésbé hangkivető -or végűekkel való kapcsolatának tudható be.



2. ábra: 2-3 elemű részgráfok a komplex jegymérték alapján

¹¹ A *köböl*-nek a *tengeröböl*-nél alacsonyabb hangkivetési mértéke részben alacsonyabb gyakoriságával és az *öböl*-höz való nagyobb alaki hasonlóságával magyarázható.

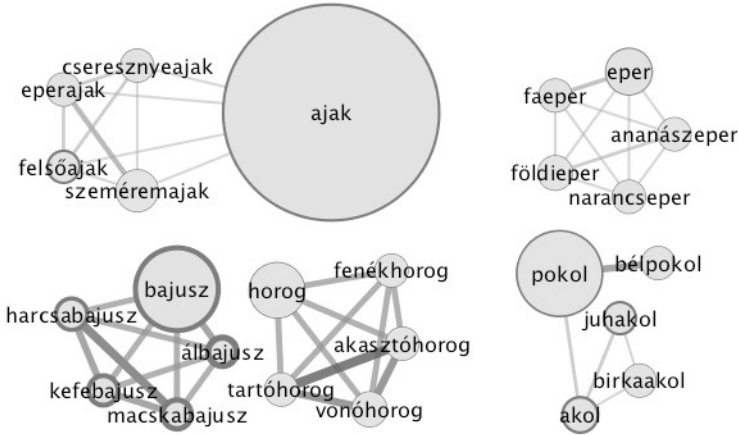
A 4 elemű részgráfok közt 3:1 arányban többen vannak azok, amelyek nem egy szóból és a belőle létrehozott összetételekből állnak. A *-lök* végűek stabilan hangkivető viselkedése annak köszönhető, hogy nincs ilyen végű nem hangkivető főnév, illetve az ábrán nem látható, de viszonylag közeli, nagy gyakoriságú *kölyök* is hatással lehet viselkedésükre (a *kölök* és *kölyök* komplex tengelymérték szerinti hasonlósága: 0,87). A hangkivető főnevek közt tipikusnak számító *-ok* végűek közül egyedül a *gyilok* nem következetesen hangkivető. A *gyilok-tulok* csoport esetében a *-gyilok* és a *-tulok* végű szavakat összekötő kapcsolat gyenge. Ez támogatja a *gyilok* (213) kevésbé hangkivető viselkedését, amely ahhoz kapcsolható, hogy a szlengben az eredeti ‘kétélű tőr, várfal koronáján végigfutó, mellvédes (és fedett) (védő) folyosó’ jelentése helyett ‘gyilkolászás, mészárlás, használattal való tönkretétel’ értelemben használatos, és eltávolodott jelentésében az informális kommunikáció során az eredeti hangkivetéses alakok hatása gyengébb. Viselkedésben az 1 alakkal előforduló *méreggyilok* nem követi, amely elsősorban adathiánynak tudható be, azonban a *Google Gyakorisági Gyűjtésben* már 12-szer fordul elő a leggyakoribb hangkivetéssel együttjáró toldalékokkal, és ebből 7 alakja nem hangkivető (41,6%-ban hangkivető viselkedés).



3. ábra: 4 elemű részgráfok a komplex jegymérték alapján

Az 5 elemű részgráfok közé viselkedésükben és alaki felépítésükben is heterogén csoportok tartoznak. Az *eper* és a *horog* csoport zárt szerkezettel rendelkezik, és ennek megfelelően stabilan hangkivető viselkedést mutat. A *horog* esetében mindenképpen számolnunk kell a *dolog* analógiás hatásával, ami erősíti hangkivető viselkedését. A *bajusz* csoportja jó példa arra, ha a csoport meghatározó tagja kevésbé hangkivető, akkor a többiek azzal összhangban viselkednek. A *-bajusz* végűek közül egyedül a 111 előfordulással rendelkező *harcsabajusz* viselkedik kevésbé hangkivető módon (57%), mint a nála gyakoribb *bajusz* (4985, 35,9%). Az eltérés oka, hogy a *harcsabajusz* a többi *-bajusz* végű összetett szóhoz viszonyítva is igen megszorított használatú. A hangkivetéssel együttjáró toldalékok esetében az előfordulások 88,2%-a

E.3 birtokos alakokból származik a *harcsabajusz*-nál, amelyet a közel hasonló gyakoriságú képzett alakok is támogatnak (*harcsabajszú* 59, *harcsabajszos* 39 stb.), míg a *bajusz*-nál az E.3 birtokos alakok aránya csupán 65,2%. Ennek a nagyon beszűkült, de gyakori használatnak köszönhetően a *harcsabajsza* alak viszonylag függetlenül él a *harcsabajusz* többi alakjától, ami miatt az elvártnál kevésbé ingadozik.



4. ábra: 5 elemű részgráfok a komplex jegymérték alapján

A szerkezetileg gyengébb¹² *pokol-akol* csoportban is találunk kevésbé hangkivető alakokat. Az ingadozó *juhakol*, *akol* viselkedését magyarázhatja, hogy az *akol* VCVC szerkezetével nem tipikus hangkivető szó. Sajátosan viselkedik az *ajak* csoport, mivel a leggyakoribb elem hangkivetési mértéke a legkisebb. Ez igaz még akkor is, ha a kis gyakoriságú *eperajak* (4), és *cseresznyeajak* (6) szavaktól el is tekintünk vizsgálatában. Az *ajak* hangkivetésének alacsonyabb mértéke (96,85% hangkivetés) annak köszönhető, hogy a szó hangkivető viselkedése sosem szilárdult meg. Emellett tanúskodik, hogy már 1766-ból van *ajjaka* (Mátyus 1766: 396) előfordulása, amely változat a választékos beszédben is megmaradt. Ezzel találkozhatunk lexikonokban (*Pallas Nagy Lexikona*, *Kislexikon*), vagy akár költői művekben is (Kosztolányi Dezső művei, Jannus Pannonius, Guillaume Apollinaire fordítások stb.). Ilyen választékos szövegek a *Szószablya Korpuszban* is lehetnek, de

¹² A legfeljebb 12 elemet számláló csoportok közül ebben a legalacsonyabb a csoporton belüli elemek közt számított hasonlóság.

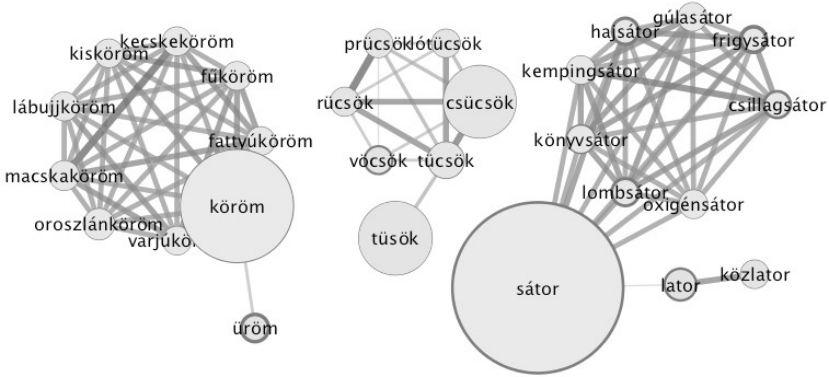
beazonosításuk és az *ajak* hangkivetési mértékére gyakorolt hatásuk felmérése meglehetősen nehéz. A *Google Gyakorisági Gyűjtés* tanulsága szerint már a többi alak sem stabilan hangkivető, egyedül a jelentésében távolabbi *szeméremajak* következetesen hangkivető, amely egyedül nem a szájra utal az *-ajak* végűek közül.

A továbbiakban csak azokat a részgráfokat mutatom be ábrákon, amelyek szerkezetükben vagy csomópontjaik viselkedésében érdemben különböznek a már leírt esetektől. A *dolog* (6 elem), a *gyomor* (7 elem), a *kölyök* (7 elem), a *fátyol* (7 elem, a *bajusz* csoporthoz hasonló viselkedés), a *teher* (8 elem), a *fogoly* (10 elem), a *titok* (11 elem), a *kapocs* (13 elem), a *szobor* (12 elem), a *meder* (11 elem, a kevésbé hangkivető *veder*-rel) csoportjaiban egy nagy gyakoriságú elem köré kisebb gyakoriságú elemek szerveződnek szoros kapcsolatokkal a központi elemhez hasonló viselkedéssel.

A 6 elemes *vétek* csoport felépítése a *pokol* csoportéhoz hasonló, amelyben egyedül a lazábban kapcsolódó *reték* (95,8%) nem következetesen hangkivető. A *selyem-petrezselyem* által dominált 7 elemű csoport a *petrezselyem* (1118) és a *selyem* (1543) származékaiból áll. A két szó közel azonos gyakorisága miatt mind a kettő alkalmas prototípus szerepre, így a csoport némileg instabil a „vetekedésük” miatt. A *hernyóselyem* (92,9%, 14) és a *zöldpetrezselyem* (96,4%, 84) szavak kevésbé hangkivetők néhány E.3 birtokos alak miatt. A *fodor* csoportjában (9 elem) a kisebb gyakoriságú és legjobban eltérő *bodor* (88,9%, 45) és a *nyakfodor* (92,8%, 14) kevésbé hangkivető elsősorban tárgyvető alakjaiknak köszönhetően. Elvárásainkkal összhangban van, hogy a kisebb gyakoriságú, de a csoport prototípusához alakjában és jelentésében közelebbi *nyakfodor* jobban követi a prototípus viselkedését. A *kehely-pehely* csoport (9 elem) esetében a szoros kapcsolatok teszik lehetővé, hogy a kivételes hangátvetés megmaradjon. A csoportban a kissé távolabbi *teher* hatása is minden bizonnyal érezteti hatását.

A *sátor* (11 elem), a *köröm* (10 elem) és a *csücsök* (7 elem) csoportjaiban a központi szótól leszakadó, és ezért kevésbé hangkivető elemeket láthatunk az 5. ábrán: *üröm* (58,5%), *lator* (89,7%), *vöcsök* (96,5%). A *közlator* (1 előfordulás) viselkedése adathiány miatt nem megítélhető. A *sátor* a hozzá legközelebbi szavakkal együtt a *fátyol*-hoz és a *bajusz*-hoz hasonlóan viselkedik. Ingadozása már a 16. századból is adatolható. A csoporton belül az elvártnál jobban hangkivető *könyvsátor* javarészt a hangkivetéssel jobban együttjáró többes számmal fordul elő (80%-ban), ezért még 98,57%-ban hangkivető módon viselkedik a *sátor* 80,5%-ával szemben, amely a többes szám előtt azonban hasonlóan jobban követi a hangkivető mintát (95,2%-ban hangkivetéses alakok). A *köröm* csoportja a leszakadóban lévő *ürömet* leszámítva

követi a hangkivető sémát. A *majom* csoportnak (13 elem) alárendelt *sólyom*-hoz csak gyengén kapcsolódó *sulyom* (40; 65%) szintén kevésbé hangkivető módon viselkedik marginalizálódott helyzete miatt.

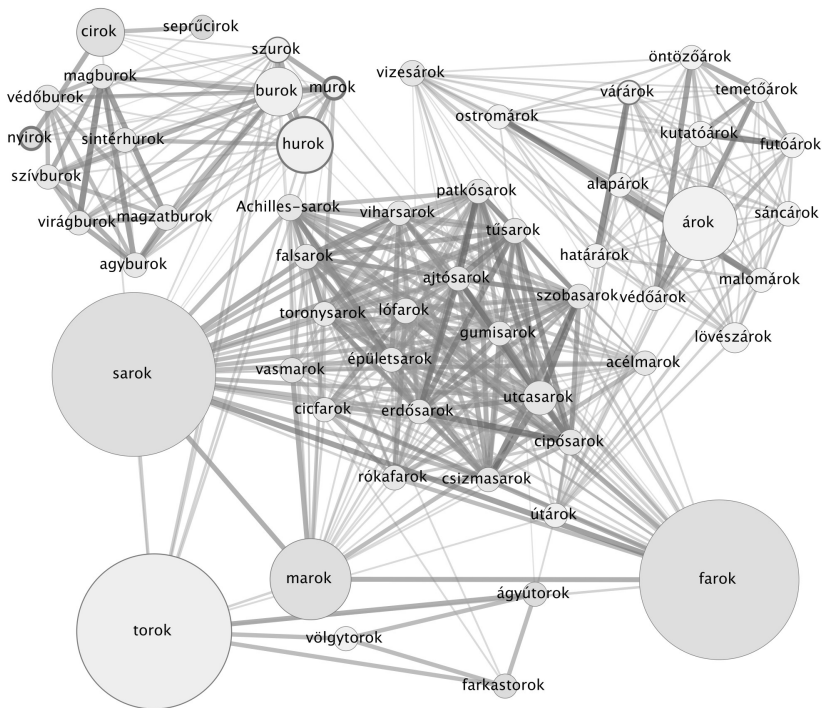


5. ábra: Kiselemű részgráfok leszakadó elemei a komplex jegymérték alapján

A 14 elemű *karom* csoport esetében a magasabb elemszám és az *-alom* véghez való hasonlóság még a helyenként gyenge belső kapcsolatok ellenére is megóvjá szavait az ingadozástól. A közepes méretű csoportokból a *lélek* (17 elem), a *haszon* (20 elem), a *gödör* (20 elem), az *izom* (25 elem), a *fészek* (30 elem), a *tükör* (32 elem), a *torony* (35 elem), a *méreg* (44 elem) semmiben sem tűnik ki a korábbiak közül. Azt tapasztalhatjuk, hogy az elemszám és a kapcsolatok számának növekedésével erősödnek a csoportok, nincs ingadozás. A *tartalom-figyelem* csoportok után a legnagyobb osztály a *bokor-cukor*, amely teljesen dualisztikus felépítésű (*cukor* 10281, *bokor* 10110). A csoport tagjai a gyakori, prototipikus szavakon keresztül, illetve a hasonlóságban köztük álló *-csokor* végű szavakon át kapcsolódnak. Az alcsoportokon belüli gyakorisági arányok is hasonlóak (*cukor* alcsoport: 13701, *bokor* alcsoport: 15545).

A *sarok-farok-torok* csoportot (48702, 34254, 19524) több nagy gyakoriságú, formailag közeli szó dominálja. A csoport méretében a *bokor-cukor* csoporthoz hasonlít, de nagyobb heterogeneitása miatt kisebb belső stabilitással rendelkezik. A csoport összefüggőségét bizonyos esetekben csak kis gyakoriságú elemek gyenge kapcsolatai biztosítják (*vizesárok-Vsarok* végűek,

útárok–torok, ágyútorok). A csoportban a *murok* (14; 78,5%), *nyirok* (45; 93,3%), *hurok* (4279; 98,1%) kevésbé hangkivetők¹³.



6. ábra: A sarok-farok-torok csoport kapcsolatrendszere a komplex jegymérték alapján

2. A komplex tengelymérték alapján számított kapcsolatok

A komplex tengelymérték alapján létrehozott gráf, habár sokban hasonlít a komplex jegymérték segítségével készítetthez, bizonyos esetekben olyan eltéréseket is mutat, amelyeket érdemes a továbbiakban áttekintnünk. Az ábrán a megjelenítés módja azonos a korábbiakéval, de a hasonlóság alsó küszöbértékét 0,85-re módosítottam. A 0,9-es értéket szükséges volt lejjebb

¹³ A *várárok* „ingadozása” egy *várárok*on előfordulásnak tudható be.

szállítani, mivel még így is sokkal kevesebb kapcsolatot kaptam, mint a másik gráf esetén¹⁴. A küszöbérték további csökkentését azonban már nem tartottam kívánatosnak, mert ebben az esetben a struktúrák teljesen összefüggéneek, így a csoportok elemzése nehezebbé vagy kevésbé informatívává vált volna.

Fel kell tennünk a kérdést, hogy mennyire jól összehasonlíthatóak az eltérő küszöbértékkel számított gráfok struktúrái. Hogyan viszonyul a komplex tengelymérték 0,85-ös értéke a komplex jegymérték 0,9-es értékéhez? Hasonlósági számaim csak relatív mutatók, azaz csak a rendszeren belül hasonlíthatók össze, azonban azok a struktúrák és viszonyok, amelyeket ezek alapján felismerünk, már alkalmasak az összevetésre. Ezek az értékek a kapcsolatok felső szegmensét vágják ki, így annyi mondható el róluk, hogy mind a 0,85, mind a 0,9 szoros, de nem „intim” kapcsolatot jelöl a saját viszonyrendszerén belül. Ennél többet nem állíthatunk, de ez a fajta bizonytalanság a szavaknak egy adott mérték szerinti összehasonlításában is megvan, hisz az csak egy feltetelezés még algoritmusaim viszonylagos jósága esetén is, hogy azonos értékek azonos mértékű hasonlóságot fejeznek ki.¹⁵ Mivel azonban az összehasonlítás során nem a konkrét értékekre hagyatkozok, hanem csak az azok által hasonló módon definiált viszonyokra, a két gráfot összehasonlíthatónak veszem.

2. táblázat: A hangkivető szavak csoportjai a komplex tengelymértékkel számított hasonlósági viszonyaik alapján.

A csoport leggyakoribb szava	Példánygyakoriság	Szavak száma a csoportban	Hangkivetés mértéke	Csoporton belüli átlagos hasonlóság
tartalom	1309962	330	99,60%	0,778
figyelem	1274921	230	99,56%	0,809
dolog	743093	88	97,74%	0,704
lélek	249463	62	99,68%	0,774
szobor	189088	176	97,10%	0,708
títok	107982	15	99,99%	0,824

¹⁴ 24904 kapcsolat, szemben a komplex jegymérték alapján készített gráf 51655 kapcsolatával. A potenciális kapcsolatok száma $1092^2/2 = 596232$.

¹⁵ Azonban a valamely algoritmusunk által megadott hasonlósági értékek közti különbségek már mentálisan reálisak lehetnek, így ha két érték jelentősen különbözik (pl. *tartalom–paradicsom*: 0,7-es hasonlósági érték a komplex jegymérték alapján; *tartalom–világuralom*: 0,9-es hasonlósági érték a komplex jegymérték alapján), akkor az azok által jellemzett párok hasonlóságát már anyanyelvi beszélők is különbözónak ítélnék meg.

Rung András

kapocs	107195	13	99,64%	0,868
haszon	63625	20	94,41%	0,830
teher	61856	8	96,42%	0,879
fészek	23361	33	99,86%	0,847
gyomor	19094	7	92,85%	0,897
fogoly	13160	8	91,87%	0,821
szirom	10325	29	99,76%	0,823
pehely	5650	4	99,33%	0,884
fátyol	5158	7	37,68%	0,879
bajusz	5132	5	32,45%	0,874
hadifogoly	2241	3	99,94%	0,901
szeméremajak	819	2	100,00%	0,936
kölyök	527	6	100,00%	0,967
szénakazal	227	2	84,92%	0,944
horog	209	4	100,00%	0,925
vadászszólyom	47	2	98,48%	0,938
kultúrmocsok	36	3	100,00%	0,866
sportszatyor	36	2	98,08%	0,895
jégcsapretek	29	2	100,00%	0,942
burgonyapehely	19	2	100,00%	0,947
ananászeper	3	2	100,00%	0,930

A komplex tengelymérték által létrehozott egyedi részgráfok vizsgálata előtt érdemes áttekintenünk, hogy az egyes csoportok milyen jellemzőkkel bírnak (2. táblázat). A csoportok hangkivetésének mértéke és más mutatók közt nincs szignifikáns együttjárás. A csoporton belüli átlagos hasonlóság azonban a komplex jegymértéknél tapasztaltakkal ellentétben összefügg a példány- ($r(25) = -0,56$, $t = -3,39$, $p > 0,01$) és a típusgyakorisággal ($r(25) = -0,63$, $t = -4,03$, $p > 0,001$) is, azaz a strukturális hasonlóságoknak nagyobb súlyt adó komplex tengelymérték alapján látható, hogy a csoportok méretének növekedésével párhuzamosan a csoportok belső szerkezeti hasonlósága lazul. Ugyanakkor a csoporton belüli átlagos hasonlóság mértéke még a nagy csoportokban is jelentősen meghaladja a hangkivető főneveknek a komplex tengelymértékkal egymáshoz mért hasonlóságának 0,53-as átlagos értékét. Enyhe pozitív korrelációban van a szavak kapcsolatainak száma hangkivetésük mértékével ($r(1063) = 0,154$, $t = 5,08$, $p > 0,001$), amely fennáll akkor is,

ha a hangkivetést 99%-nál kevésbé követő szavakat nézzük ($r(104) = 0,246$, $t = 2,51$, $p > 0,05$).

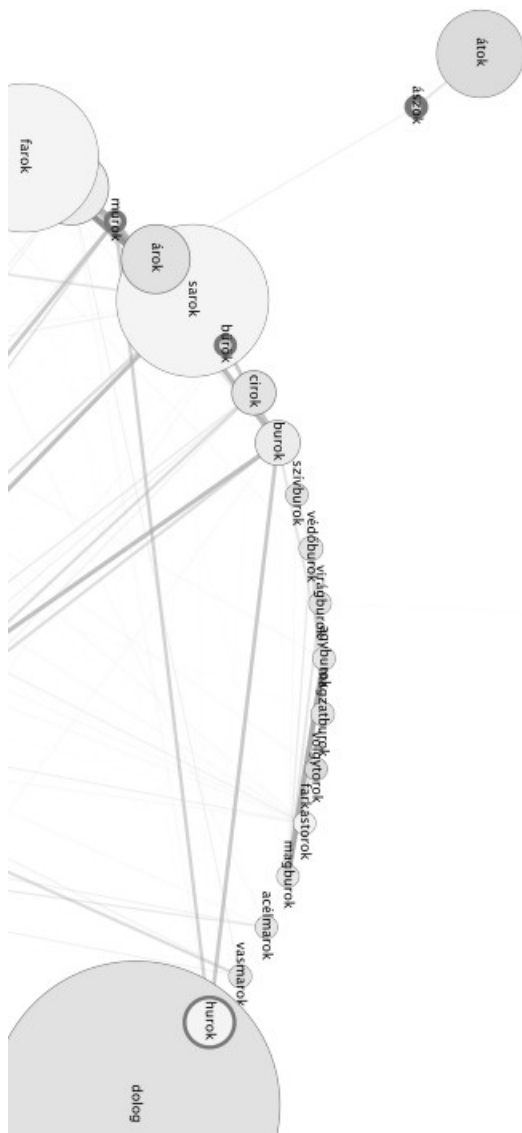
A komplex tengelymérték segítségével készített gráfon 27 szó nem kapcsolódik egyetlenegy részgráfba sem: *ajak*, *birkaakol*, *boholy*, *csalétek*, *eperajak*, *faeper*, *fattyúcsülök*, *felsőajak*, *izom*, *juhakol*, *kazal*, *kerecsensólyom*, *koboz*, *kukoricapehely*, *méreggyilok*, *misekehely*, *páncélököl*, *pecek*, *seprűcirok*, *szájpecek*, *takony*, *tegez*, *tejhaszon*, *tengeröböl*, *vasököl*, *vicikvacak*, *virágkehely*. Ezek a szavak eltérőek a komplex jegymérték alapján azonosított szavaktól, mivel a komplex tengelymérték a tengelyek figyelembevételé miatt a szerkezeti hasonlóságnak nagyobb súlyt ad az első sorban a végek hasonlóságát vizsgáló komplex jegymértékkel szemben. Mindösszesen két szó, a *koboz* és a *tegez* (valamint a bizonytalan státuszú *boholy*) található meg mind a két csoportban, amelyek azonban erőteljesen ingadoznak. Ez azt mutatja, hogy azok a hatások, amelyek az egyes szavak vonatkozásában a paradigma többi tagjától a szerkezetek és a végek esetében tapasztalható különbözőségekből erednek, összeadódnak, és ezek együttesen gyorsíthatják a kevésbé hasonló elemektől való viselkedésbeli eltávolodást.

Ha a 25-nél több alakkal előforduló szavakat nézzük, akkor többségben vannak a nem következetesen hangkivetők: *ajak* (97%, 25039), *felsőajak* (97%, 32), *juhakol* (97%, 30), *kazal* (91%, 310), *koboz* (63%, 145), *páncélököl* (93%, 71), *pecek* (98%, 122), *takony* (98%, 527), *tegez* (24%, 447), *vasököl* (94%, 33). A kivételek elsősorban olyan összetett szavak, amelyeknek van(nak) azonos utótagú párja(ik), de előtagjaik nem hasonlíthatnak: pl. *vadászsólyom* : *kerecsensólyom*. Egyedüli stabilan hangkivető alapszó az *izom* (26810), amelyet szerkezeti magányossága ellenére kiugró gyakorisága tart a hangkivető mintában. Ha a 0,85-ös hasonlósági küszöbérték alapján magányosnak minősülő szavak távolabbi kapcsolatait vizsgáljuk meg, akkor hasonló összefüggéseket figyelhetünk meg, mint a komplex jegymérték alapján készített számítások esetében.¹⁶ A kevésbé hangkivetők átlagos hasonlósága a többi hangkivető főnévhez 0,45, a hangkivetéses sémát következetesen követőké 0,47, amely átlagok szignifikáns mértékben eltérnek ($t(2116,34) = 8,13$, $p < 0,001$). A hangkivetéses mintának megfelelő magányos szavak 0,47-es értéke szintén szignifikánsan különbözik az összes hangkivető főnév egymáshoz számított hasonlóságának 0,53-as átlagértékétől ($t(1702,05) = -29,2833$, $p < 0,001$). Ebből következik, hogy a szegényes közeli kapcsolatokkal rendelkező hangkivető főnevek kevésbé hasonlíthatnak a többi hangkivető

¹⁶ Az összehasonlításból a *boholy*-t ezúttal is kihagytam, valamint a kevesebb, mint 25 előfordulással rendelkező szavakat, mert ezek esetében a hangkivető séma követése adathiányosnak is be-tudható.

főnévre, különösen, ha kevésbé követik a hangkivető séma szerinti viselkedést. A szegényes közeli kapcsolatokkal rendelkező szavak hasonlósági átlagai közt mutatkozó enyhe, de szignifikáns különbséget erősíti meg, hogy a kevésbé hangkivetőként viselkedő szavaknál egy szóra átlagosan 18 0,7-es hasonlósági értéknél szorosabb kapcsolat jut, míg a következetesen hangkivetőkre szavanként 34. Ha csak a 0,8-as hasonlósági értéknél közelebbi kapcsolatokat vesszük, akkor a különbség még élesebb. A kevésbé hangkivető főnevek átlagosan 1,5 ilyen kapcsolattal rendelkeznek, a következetesen hangkivetők viszont 4,8-cal.

A továbbiakban azt a néhány részletet vizsgálom meg a komplex tengelymérték segítségével számított részgráfokon, amelyek a már bemutatott komplex jegymérték segítségével számított részgráfok alapján tett felismeréseken túlmutató megfigyelésekhez vezethetnek. A már ismertetett jellemzőiből következő módon a komplex tengelymérték részgráfjai több elemet számlálnak, és kapcsolataikban szegényesebbek. A részgráfok ezen eltérő sajátosságát a korábban részletesen bemutatott *sarok-torok-farok* részgráf párját megmutató 7. ábrán is megfigyelhetjük.



7. ábra: $-COC_{vel}$ csoport egy jellemző részlete a komplex tengelymérték alapján

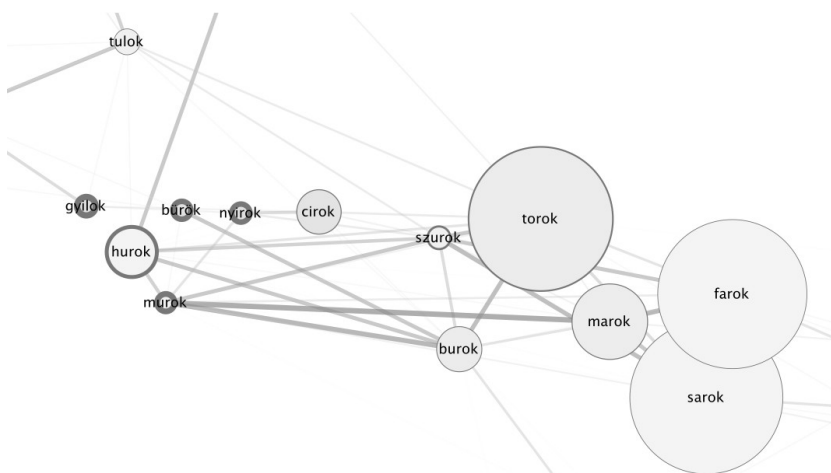
A komplex tengelymérték grájában a strukturális hasonlóság a vég hasonlóságnál nagyobb szerepet kapott. A csoportba így bekerülhetett még domináns elemként a *dolog*, amely azonban kapcsolatai szegényessége miatt prototípusnak kevésbé ideális, így emiatt dominanciája korlátozott. A gyakoriságukkal kiemelkedő elemek közt a csoport általános sémájánál specifikusabb jellemzőket is megragadhatunk:

(3)	<i>dolog</i>	566770 előfordulás
	<i>sarok</i>	48702 előfordulás
	<i>farok</i>	34254 előfordulás
	<i>torok</i>	19524 előfordulás
	<i>marok</i>	7369 előfordulás

(4) prototípus séma: $C(a/o)C_{iiq}oC_{vel}$

A több, nagy gyakoriságú elem viszonylag stabil hangkivető viselkedést biztosít a formailag heterogén csoportnak, amelynek 0,703-as csoport hasonlósági átlaga a legalacsonyabb a komplex tengelymérték alapján meghatározott csoportok hasonló értékei közt. A komplex tengelymérték nemcsak a kiugró elemek magasabb szintű kapcsolatainak az azonosítására megfelelő, hanem segítségével több, a hangkivető sémát kevésbé követő szó¹⁷ viszonyát is megfigyelhetjük. A 7. ábra segítségével a szerkezeti szempontból magányosabb *ások* kevésbé hangkivető viselkedését tudjuk értelmezni, amely a gyakoribb *átok* és *árok* közt helyezkedik el. A csoportot némileg más elrendezésben megmutató 8. ábrán olyan szavakat figyelhetünk meg, amelyek közös jellegzetességeinek megragadására a komplex jegymérték nem volt alkalmas. A kevésbé hangkivető csoporttagokat (*gyilok*, *murok*, *nyirok*, *bürök*, *szurok*, *üszök*, *tülök*, *csülök*, *hurok*) a $-UC_{cont}Ok$ sémával ragadhatjuk meg, amely különbözik a prototipikus szavak sémájától, ugyanis bizonyos pontjaiban specifikusan más vagy annál általánosabb. Ez idézheti elő e szavak kevésbé hangkivető viselkedését, amelyek a *sarok-torok-farok* csoportba tartoznak, mivel annak néhány nagyon általános jegyét osztják, de leginkább kis gyakoriságú társaikra hasonlítanak csak, semmint az erősebb prototípusokra vagy az ezekkel szorosabb kapcsolatban lévő többi szóra.

¹⁷ A *torok* minimális ingadozása a *török* ékezetmentes alakjaival hozható kapcsolatba.



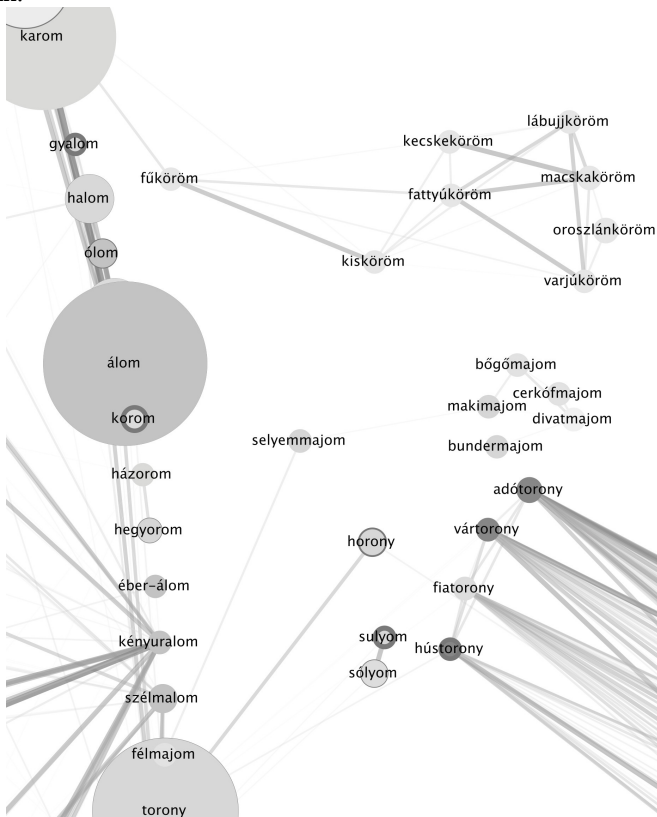
8. ábra: $-COC_{vel}$ részgráf kevésbé hangkivető szavai a komplex tengely-mérték alapján

A 3. legnagyobb struktúra (*szobor* csoport: 176 elem) a komplex jegymérték alapján készített gráf több kisebb gráfját vonja össze strukturális hasonlósági alapokon. A gráfelemzés lehetővé teszi, hogy megállapítsuk: a csoporthoz kevés szállal kapcsolódó szavak (legfeljebb 7 kapcsolat), amelyek legalább 25 előfordulással bírnak, kevésbé követik a hangkivető sémát (91%, csoport-átlag: 96,7%). A csoporton belül a hangkivetés mértéke és a kapcsolati fokok száma közt szignifikáns pozitív korreláció ($r(174) = 0,28$, $t = 3,84$, $p < 0,001$) figyelhető meg.

A 9. ábra segítségével a kisebb, de viszonylagosan stabil, szerkezetileg egységesebb csoportok és az *-alom* végűeket is tartalmazó heterogén gyűjtőcsoport¹⁸ kapcsolatát figyelhetjük meg. A csoportokat összekötő szavakból azonban – ahogy az *ászkok* esetében tapasztaltuk – több már kevésbé követi a hangkivető sémát (*sulyom* 65%, *hústorony* 95,4%), de megfigyelhetünk olyan szavakat is, amelyek csak az *-alom* végűeket is tartalmazó heterogén gyűjtőcsoporthoz való gyengébb kapcsolatuk miatt viselkedhetnek ugyanígy (*gyalom*¹⁹ 80%, *horony* 98,9%). A *fészkek* csoportjában (10. ábra) hozzájuk hasonlóan a struktúra szélére szorult *rettek* (608, 95,8%) viselkedik kevésbé

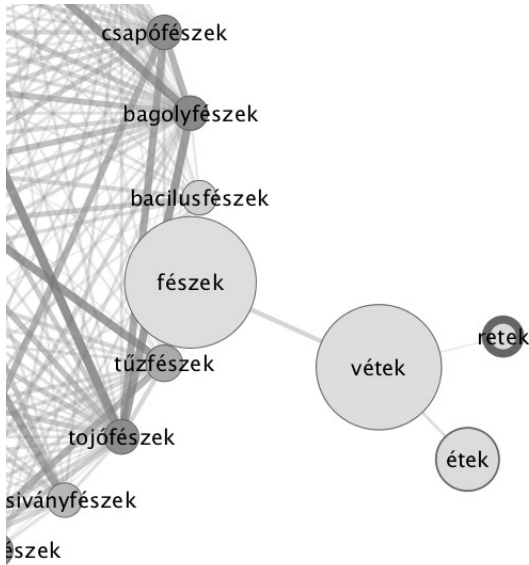
¹⁸ Habár az *-alom* vég dominál a csoportban, kiváló példa a wittgensteini családi hasonlóságra, mivel minden elemhez lehet benne találni nagyon hasonlót, de vannak egész távoli elemei is: például *üröm-torony*.

hangkivető módon, mint ahogy azt a *sulyom*, *üröm*, *lator* esetében láttuk korábban.



9. ábra: A torony, majom, köröm, karom csoportok szorosabb kapcsolódása a komplex tengelymérték alapján az *-alom* végűeket is tartalmazó heterogén gyűjtőcsoporthoz.

¹⁹ A *gyalom* azon kevés *-alom* végű szavak közé tartozik, amelyek jelentésükből kifolyólag is csak lazán kapcsolódnak a csoporthoz. A *halom*, *álom*, *alom* szavak esetében viselkedésük stabilabb a *gyalom*-nál, nagyobb gyakoriságuknak köszönhetően.



10. ábra: A *reték* gyenge kapcsolódása a *fészek* csoporthoz a komplex tengelymérték alapján

Összefoglalás

Elemzésemben megmutattam, hogy a magyar hangkivető főnevek viselkedését jobban megérthetjük, ha újszerűen a hasonlósági kapcsolataik alapján kialakított gráfstruktúrákban vizsgáljuk őket. Ez lehetővé teszi, hogy az eddig egyformaként kezelt szavak egyedi viselkedését, csoportszerveződéseit jobban megérthessük. Ez a megközelítés a kvalitatív megfigyeléseken túl lehetővé tette azt is, hogy statisztikailag igazolt kijelentéseket tehesünk a hangkivető főnevek viselkedéséről, amelyek összhangban vannak az analógiás megközelítések alaptételeivel. Láthattuk, hogy a változásban nemcsak a bizonyos csoportokhoz való hasonlóságnak, hanem az ezektől való eltérésnek is szerepe van. A felépítésükben egyedi, magányos szavak jobban eltávolodtak a hangkivető séma által meghatározott viselkedéstől, mint azok a hangkivető főnevek, amelyek a hangkivetés vonatkozásában hasonlóan viselkedő szavakkal közösen hasonlósági csoportokba rendeződtek. Ebből azt a következtetést vonhatjuk le, hogy az analógiás alapú regularizálódásban elsődleges szerepe lehet a saját viselkedési csoporttal való kapcsolatok meggyöngyülésé-

nek (formai vagy jelentésbeli autonómia), amely hatásában akár jelentősebb lehet, mint az a vonzóerő, amelyet a nem hangkivető főnevek fejtenek ki ezekre a szavakra. Ezzel szemben a formailag heterogén hangkivető szavak közt zárt mintát alkotó *-alom*, *-elem* végűek következetesen hangkivetők, ami kapcsolatba hozható erős hasonlósági viszonyaikkal és magas gyakoriságukkal is. Közel azonos viselkedésük nem tulajdonítható az *-alom*, *-elem* morfémáknak, hisz több esetben ezek nem jól vagy egyáltalán nem szegmentálhatóak (*cimbalom*, *cirkalom*, *irgalom*, *lakodalom*, *malom*, *alom*, *halom*, *gyalom* stb.), és morféma alapon azt sem tudnánk magyarázni, hogy több nagyon hasonló, de némileg eltérő végű szó (*áalom*, *óalom*) miért viselkedik közel azonos módon velük. A hasonlóság és a különbözőség tényezőjének fontosságát megfigyelhettük abban is, hogy a komplex jegymérték és a tengelymérték által meghatározott legszorosabb kapcsolatok száma szignifikánsan összefügg a hangkivetés mértékével.

Irodalomjegyzék

- Albright, Adam 2009. Modelling analogy as probabilistic grammar. In Blevins, James – Blevins, Juliette (eds.) *Analogy in Grammar: Form and Acquisition*. Oxford, Oxford University Press, 185–213.
- Bíró Tamás 2006. *Finding the Right Words: Implementing Optimality Theory with Simulated Annealing*. Groningen Dissertations in Linguistics. Groningen.
- Bybee, Joan L. – Moder, Carol Lynn 1983. Morphological classes as natural categories. *Language* 59: 251–270.
- Duvignau, Karine – Gaume, Bruno 2004. Linguistic, Psycholinguistic and Computational Approaches to the Lexicon: For Early Verb-Learning. A special issue on 'learning'. *ESSCS Journal, Journal of the European Society for the study of cognitive systems* 6-2(3): 255–269.
- Enfield, N. J. 2008. Transmission biases in linguistic epidemiology. *Journal of Language Contact* 2: 295–306.
- Goldsmith, John 1990. *Autosegmental and metrical phonology*. Basil Blackwell, Oxford.
- Halácsy Péter – Kornai András – Németh László – Rung András – Szakadát István – Trón Viktor 2003. A Szószablya projekt. In Alexin Zoltán – Csentes Dóra (szerk.): *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003)*. 299.
- Hallan, Naomi 2001. Paths to prepositions? A corpus-based study of the acquisition of a lexico-grammatical category. In Bybee, Joan – Hopper, Paul (eds.): *Frequency and the emergence of linguistic structure*. Amsterdam, John Benjamins, 91–121.
- Kálmán László – Rung András 2010. Klaszterek helyett prototípusok. In Tanács Attila – Csentes Dóra (szerk.) *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2010)*, 325–332.

- Kiefer Ferenc (szerk.) 1994. *Strukturális magyar nyelvtan 2. Fonológia*. Budapest, Akadémiai Kiadó.
- Krott, Andrea 2009. The role of analogy for compound words. In Blevins, James – Blevins, Juliette (eds.): *Analogy in Grammar: Form and Acquisition*. Oxford, Oxford University Press, 118–136.
- Lukács Ágnes 2002. *Alaktanilag kivételes tövek vizsgálata a magyarban. A leíró általánosítások mentális realitása*. (szakdolgozat)
- Mátyus István 1766. *Diaetetica. II.* Kolozsvár.
- Rung András 2008. Determining word similarity in the Hungarian language. In Kálmán László (ed.) *Papers from the Mókus Conference*, Budapest, Tinta Kiadó, 112–118.
- Rung András 2009. Szóhasonlóság mérése analógiás megközelítésben. In Tanács Attila – Szauter Dóra – Vincze Veronika (szerk.): *VI. magyar számítógépes nyelvészeti konferencia. MSZNY 2009*. Szeged, Szegedi Tudományegyetem, 104–113.
- Rung András 2011. Hangkivető főnevek analógiás keretben. *LingDok 10*. 112–203.
- Siptár, Péter–Törkenczy, Miklós 2000. *The Phonology of Hungarian*, Oxford, Oxford University Press.
- Skousen, Royal – Lonsdale, Deryle – Parkinson–Dilworth B. (eds.) 2002. *Analogical Modeling*, Amsterdam, John Benjamin.
- Slobin, Dan 1973. Cognitive prerequisites for the development of grammar. In: Ferguson, Charles – Slobin, Dan (eds.) *Studies of child language development*, New York, Holt Rinehart and Winston, 174–208.