Eötvös Loránd Tudományegyetem
Bölcsészettudományi Kar

DOKTORI DISSZERTÁCIÓ

Rung András

Magyar főnévi alaktani jelenségek analógiás megközelítésben

Nyelvtudományi Doktori Iskola
Dr. Bańczerowski Janusz DSc., tanszékvezető egyetemi tanár

Elméleti Nyelvészeti Doktori Program
Dr. Bánréti Zoltán egyetemi docens

A bizottság tagjai és tudományos fokozatuk:

A bizottság elnöke:
Dr. Kiefer Ferenc MHAS., professor emeritus

Hivatalosan felkért bírálók:
Dr. Kornai András DSc.
Dr. Törkenczy Miklós DSc.

A bizottság titkára:
Dr. Rebrus Péter PhD.

További tagok:
Dr. Prószéky Gábor DSc.
Dr. Siptár Péter DSc.
Dr. Lukács Ágnes PhD. (póttag)

Témavezető:
Dr. Kálmán László CSc.

Budapest, 2011

Tartalom

Az ábrák és táblázatok fogalmainak és kifejezéseinek jegyzéke

Analógiás megközelítés: olyan elmélet, amely kizárólag egyes elemek vagy elemek egy meghatározott csoportja alapján tesz előrejelzéseket a várható viselkedésről.

Egyértékű jegy: olyan fonológiai jegy, amely jelenlétével egy fonéma valamely tulajdonságát fejezi ki (pl. koronális), viszont értéket nem vesz fel. Hiánya esetén az adott tulajdonság nem jellemző a kiválasztott fonémára.

Egyszerű jegymérték: gépi algoritmus, amely működésében megegyezik a komplex tengelymértékkel, de csak 8 jegyet használ.

Google Gyakorisági Gyűjtés: olyan gyakorisági lista, amely a vizsgálatomban szereplő hangkivető szavak leggyakoribb hangkivetéssel együttjáró toldalékos alakjainak gyakorisági számait tartalmazza a *Google* kereső alapján.

Hangkivetés az összes alakban: olyan hangkivetési mérték, amely megadja, hogy egy adott szó összes alakjában (képzettek is) hány százalékban találhatóak hangkivetést tartalmazó alakok.

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 az adott szó minden esetben hangkivető módon viselkedik korpuszadatok alapján a hangkivetést elváró toldalékos alakjaiban.

Hasonlósági algoritmus: olyan algoritmus, amely két elem hasonlóságát adja meg. Disszertációmban olyan hasonlósági algoritmusok szerepelnek, amelyek két szó alaki hasonlóságát (hasonlósági viszony) adják meg.

Hasonlósági csoport: olyan szócsoport, amelybe olyan szavak tartoznak, amelyek valamely hasonlósági kritérium mentén jobban hasonlítanak egymásra, mint más, velük egyszerre vizsgált szavakra.

Hasonlósági viszony: két szó hasonlóságának értéke egy gépi algoritmus által meghatározva, amelyet egy 0-1 terjedő skálán adok meg. 0 értéket kapnak az egyáltalán nem hasonlító szavak, míg 1 a szavak önmagukhoz mért hasonlósági értéke.

Hierarchikus klaszterezés: olyan klaszterezés, amely minden egyes elemet külön klaszternek tekint, és összekapcsolja ezeket egyre nagyobb klaszterekbe, míg végül egyetlen, az összes elemet tartalmazó klasztert kapunk.

Komplex jegymérték: gépi algoritmus, amely a szavak hasonlóságát azok jobb szélétől véve számítja ki úgy, hogy a megfeleléseknek, hasonlóságoknak egyre kisebb súlyt ad a szavak bal széle felé haladva. Az egyes fonémák hasonlóságát 13 jegy mentén határozza meg.

Komplex tengelymérték: gépi algoritmus, amely a komplex jegymértéktől abban tér el, hogy két szó hasonlóságát az egyes jegyek tengelyeinek hasonlósága alapján számítja ki.

Legközelebbi szomszéd: olyan elem, amely valamely tulajdonsága vagy tulajdonságai alapján egy másik elemhez leginkább hasonló.

Páros t-próba: olyan statisztika eljárás, amely két összetartozó csoport változóinak átlagait hasonlítja össze, hogy meghatározza a két csoport átlagai közt mutatkozó eltérések szignifikánsak-e vagy sem.

Példánygyakoriság: egy adott elem összes előfordulásainak száma.

Szabály: olyan általánosítás, amely egy adott környezetben kizárólagosan érvényesül (a kizárólagosság alól kivételek a sztochasztikus szabályok, de még azok is egy adott környezetben érvényesülnek valamilyen valószínűséggel) azokra az elemekre, amelyekre nem hat egy specifikusabb szabály, vagy nincsenek kivételes elemként megjelölve.

Szószablya Korpusz: a legnagyobb (1,486 milliárd szövegszó) digitális magyar gyakorisági gyűjtés.

Szuprakontextus: az AM (Analogical Modeling, Analógiás modellezés) által az egyes mintákhoz való analógiás források kiválasztásában használt változók részhalmaza, amely által jellemzett elemek azonos (homogén) vagy eltérő (heterogén) viselkedést mutatnak.

Természetes osztály: fonémáknak olyan halmaza, amelyet a fonémák valamely jegye, jegyeik vagy ezeknek értékei (nem egyértékű jegyek esetében) alapján adunk meg (pl. zöngés mássalhangzók osztálya).

Típusgyakoriság: az egy csoportban előforduló elemek száma (pl. a hangkivető főnevek típusgyakorisága 1211).

1. Bevezetés

1.1. Az analógiás megközelítés előnyei

A **szabályalapú nyelvtanok** (akár hagyományosak, akár generatívak) sokszor jó közelítő leírást adnak az alaktani viselkedésről, azonban **több nyelvi jelenségre, folyamatra nem tudnak megnyugtató magyarázatot** adni. Így megválaszolatlanul hagyják azokat a kérdéseket, hogy gyakran miért **fokozatosak az átmenetek** az egyes nyelvi kategóriák közt (Chandler 2002: 57, Lakoff 1987, Taylor 1995), mi a **valószínűség** szerepe a nyelvhasználatban, melyek a nyelvi **ingadozás** okai, illetve mi a **gyakoriság** hatása a nyelvi változásra (Skousen 1989). Ezeket a problémákat a generatív nyelvészet a performancia és a kompetencia szétválasztásával kezeli. A kompetencia alá tartozó reprezentációkat redundanciamentesnek és kategorikusnak (pl. bináris jegyek) veszi, a nyelvi elemekhez egyféle viselkedést rendel, míg számos nehezen magyarázható jelenséget (pl. beszédtévesztések (Frisch 1996: 109)) a performanciának tulajdonít, amelyeknek működését és jellegét azonban homályban hagyja, így a performancia és a kompetencia látszólagos elvi szétválasztása mögött inkább praktikus okok húzódnak meg. A performancia lesz a generatív nyelvészet nemszeretem jelenségeinek lomtára, amely ezektől megszabadulva már csak a magyarázható, a kompetencia hatókörébe tartozó jelenségekre fókuszálhat (Bybee 2001, Bybee 2010, Skousen és mtsai 2002, Blevins és Blevins 2009a).

Ezzel szemben ezekre a kérdésekre az **analógiás megközelítés rugalmasságának köszönhetően képes jobb feleletet adni**, és kezelni tudja azokat a helyzeteket is, amikor a nyelvi adatok látszólag egyediek, nem egyértelműek, rosszul formáltak, vagy a zaj, felejtés, vagy bármilyen más ok hatására azok értelmezésében és produkciójában korlátozottak vagyunk, miközben a szabályos viselkedésre is tud magyarázattal szolgálni (Eddington 2003, Skousen 1989: 54–60). Ezekben a bizonytalan esetekben egy szabályrendszer „működésképtelenné" válik, ha nem tartalmaz olyan szabályt, amely alkalmazható az adott nyelvi elemekre, vagy akár több is van belőlük. Ilyenkor a

szabályalapú megközelítésben kénytelenek vagyunk önkényes módon e nyelvi elemek tömegét rendhagyóként megjelölni, holott ezek egy másik, alkalmazkodóbb keretrendszerben magyarázhatóak lennének. Skousen (1989: 5) az analógiás és a szabályalapú megközelítések közti különbségeket az 1.1. táblázat párjai mentén foglalja össze, amelyekben néhány további különbséget is kiemel a már tárgyaltakon túl.

Szabályalapú megközelítés	**Analógiás megközelítés**
szabályok rendszere	nagyméretű adathalmaz a valós nyelvhasználat alapján
típusokon alapszik	példányokon alapszik
a konceptuális tér szabálykörnyezetekre oszlik fel	a konceptuális tér atomi marad
globális, makroszintű	helyi, mikroszintű
tanulási stratégiát igényel, hogy felfedezze a szabályokat az adatokban	tanulási stratégiát igényel az adatok eléréséhez és elemzéséhez
statikus, merev	dinamikus, rugalmas
használat: találd meg az adott környezetre alkalmazható szabályt	használat: találj megfelelő mintát, amelynek a viselkedését követni lehet
szükségszerű tudni, hogy a szabályok miképp lépnek interakcióba	szükségszerű az adatok gyors elérése
a viselkedésben a határok élesek és pontosak	a viselkedésben az átmentek fokozatosak és bizonytalanok
szabályok által meghatározott	szabályok által meghatározottnak tűnik[2]
elvárásainkat kizárólag a szabályok határozzák meg	elvárásaink az adott kontextus alapján alakulnak
explicit, direkt	implicit, indirekt
a használat a leírás függvénye	a használat a leírás

1.1. táblázat: A szabályalapú[1] és az analógiás megközelítés eltérései Skousen (1989) alapján

[1] Skousen (1989) *szabályalapú* helyett a *strukturalista* szót használja, de összehasonlítása bármilyen szabályalapú megközelítésre igaz, így ezt az általánosabb kifejezést használtam a szűkebb értelmű *strukturalista* helyett.

[2] Skousen megállapítása elsősorban arra vonatkozik, hogy ha mintáink egy tartományban egységes viselkedést mutatnak, amely jól meghatározható szabályokkal is, akkor a szabályok és az analógiák predikciói egybeesnek, így pl. az *-a* végű magyar főnevekre mind a két megközelítés azonos viselkedést jósol a tárgyrag előtt.

Az 1.1 táblázat alapján így szabály alatt olyan általánosításokat értünk, amelyek egy adott környezetben kizárólagosan érvényesülnek (a kizárólagosság alól kivételek a sztochasztikus szabályok, de még azok is egy adott környezetben érvényesülnek valamilyen valószínűséggel) azokra az elemekre, amelyekre nem hat egy specifikusabb szabály, vagy nincsenek kivételes elemként megjelölve. Az analógiás megközelítés nem ilyen általánosításokra támaszkodik elsődlegesen, – bár vannak ettől eltérő megközelítések (lásd Albright 2009) – hanem kizárólag egyes elemek vagy elemek egy meghatározott csoportja alapján tesz előrejelzéseket a várható viselkedésről. Ha az analógiás források kiválasztásnak megvannak a pontos kritériumai, mint amilyeneket jelen dolgozat is kínál a 4. fejezetben, akkor egyértelmű a meghatározása a viselkedésnek, amely önmagában lehet bizonytalan és ingadozó is. Ebben az értelemben szabályszerű az analógia, de a későbbiekben (2-4. fejezetek) részletesen kifejtésre kerülő működési mechanizmusa más, mint a szabályoknak.

Az analógia segítségével értelmezhetjük vagy legalábbis a szabályalapú megközelítéseknél jobban megragadhatjuk az olyan nyelvi jelenségeket is, mint a **német többes szám** (Wulf 2002), vagy a **török tővégi veláris-kivetés** (Rytting 2002). A német többes szám esetében a szabályalapú megközelítésnek a problémát az jelenti, hogy a legnagyobb szóosztály is kisebbségben van a tőle eltérően viselkedő csoportokkal szemben, a török /k/~∅ váltakozást pedig az teszi nehezen elemezhetővé, hogy a vizsgált szókészleten belül a nagyszámú kivételes szavak viselkedésében jól megfigyelhetők tendenciák, amelyek miatt nem kívánatos őket a lexikonba rendhagyó tövekként felvenni, de ahhoz nem elég egységesek, hogy szabályokkal megragadhatóak legyenek. Hasonlóan problematikus a nyelvtudományban az **összetett szavak kezelése**, amelyet a hagyományos megközelítések semmitmondó, esetleges és vegyes szempontokat alkalmazó kategorizációval, a generatívak pedig csak kivételek tömegét is magával hozó általánosításokkal tudnak jellemezni. A legújabb kutatások szerint (Krott 2009) több nyelv (angol, holland, német, francia, indonéz, japán és kínai)

összehasonlítása alapján az új összetett szavak létrehozási módjainak értelmezése és a már létezők elemzése is sokkal sikeresebb egy analógiás keretrendszerben[1].

A szabályalapú nyelvtanok sikertelenségének egyik fő oka az ilyen jellegű jelenségek magyarázatában az, hogy elsősorban **egzaktságra, a modellek egyszerűségére és eleganciájára[2] törekszenek**. Ennek a nyelvészeti megközelítésnek korai kritikáját már Wittgenstein (1992: 66) is megfogalmazta:

„a szavak használatát gyakran világos szabályokat követő játékokkal, kalkulusokkal *vetjük össze*, de azt mégsem állíthatjuk, hogy aki a nyelvet használja, annak egy ilyen játékot *kell* játszania."

Az egyszerűségre való törekvés miatt a szabályalapú nyelvtanok figyelmen kívül hagyják mindazokat az ismereteket, amelyeket a **nyelv valós működéséről megtudtunk pszicholingvisztikai** (Lukács 2002)**, kognitív nyelvészeti** (Chandler 2002, Eddington 1996) és **neurolingvisztikai kutatásoknak** köszönhetően, és nem vesznek megfelelő mértékben tudomást arról a tendenciáról, hogy a nyelvi működés megismerésében is nagy szerepet játszó kognitív pszichológiában a mintaalapú megközelítés egyre jelentősebb szerepet játszik (Chandler 2002: 96). Megállapításaikat továbbra is elméleti konstrukciók és részjelenségek megfigyelése alapján teszik, sokszor a valós adatok egy részét és a nyelvi változatosságot figyelmen kívül hagyva, úgy, hogy a merev szintaxis-központúságnak köszönhetően jelentés nélküli formák tanulására, tanulmányozására helyezik a fő hangsúlyt.

Ehhez a szemlélethez szorosan kapcsolódik a **memóriával való takarékoskodás** elve is, amely mindig központi szerepet töltött be a generatív érvrendszerekben. Az

---

[1] Ennek lehetőségét már Carstairs-McCarthy (1992: 109) is felveti, de mivel elemzésének/értelmezésének nem ez a központi kérdése, nem fejti ki ennek mikéntjét.

[2] Elismerem, hogy két modell közül az egyszerűbbet kell választanunk, ha mind a kettő a nyelv működését mentálisan reálisan írja le. A szabályalapú nyelvtanok esetében azonban mind a mentális realitás, mind a leírás pontossága megkérdőjelezhető, így az egyszerűség kritériuma (Occam borotvája) nem alkalmazható rájuk, hisz már az összehasonlítást lehetővé tevő feltételeknek sem felelnek meg hiánytalanul.

emberi gondolkozás és a korai, kevés tárolókapacitással rendelkező számítógépek párhuzamba állítását Gardner (1985: 385) is bírálja, mivel a számítógépek „gondolkozásával" való összehasonlítások csak abban segítettek, hogy megismerjük, hogyan *nem* gondolkozik az ember.

Az analógia alapú nyelvleírásban azonban **nem úgy tekintünk az emberi agyra, mint** egy korlátozott erőforrásokkal rendelkező **számítógépre** (Dehaene 2003: 145), vagy ha mégis maradunk ennél a sántító analógiánál, akkor azt gondoljuk, hogy a folyamatok hatékonysága és rugalmassága jóval fontosabb, mint a tárolás gazdaságos volta. Redundanciák a rendszerben léteznek, de azoknak csak felfedezése és nem kiiktatása a cél, amelyben alapvető megközelítési mód az analógia (Goldsmith 2009: 149).

További problémákat vet fel, hogy a generatív elméletek a **nyelven kívüli hatásokat, például a használati gyakoriságot, teljesen kiküszöbölték a nyelvi leírásból**[1], holott számos esetben ezeknek jelentős befolyása lehet magának a rendszernek a formálódására is (Ullman 1999, Pinker 1999, Kraska-Szlenk 2007, Rung 2008, Rung 2009). A kommunikáció alaposabb megfigyelése helyett elemzéseik középpontjában olyan normatív fogalom áll, mint a **szerkezetek jelentéstől független grammatikalitása**[2], amelyet a nyelvészek saját vagy az esetlegesen vizsgált néhány beszélő **szubjektív ítéletei** alapján határoznak meg. A grammatikalitási ítéletek vizsgálata során egy olyan szempontot hozunk előtérbe, amelyről a beszélőknek minden bizonnyal van valami elképzelésük, hisz gyakran ítélkeznek a nem sztenderd változatokról, de semmiképpen sem kellene, hogy ez a nyelvi leírás és elmélet egyik központi eleme legyen. A grammatikalitás kiemelt szerepe a generatív elméletek egzaktságra, fekete-fehérre való törekvéséből fakad, amely azonban, a nyelvi tények és megfigyelések figyelmen kívül hagyásához vezethet.

---

[1] A hagyományos elméletek igyekeznek korlátozott mértékben megküzdeni a nyelven kívüli hatásokkal is, de mivel nincsenek birtokában jó magyarázó erővel rendelkező átfogó elméletnek ezekkel kapcsolatban, próbálkozásaik eleve kudarcra vannak ítélve.

[2] A jól formált/rosszul formált megkülönböztetéssel szemben a folyamatosságot megengedő elfogadhatóság (acceptability) fogalma az analógiás megközelítésektől sem idegen.

Ezzel szemben a használat alapú nyelvtanok (Halliday 1961, Bybee 2010) a kommunikáció hatékonyságára, módjaira és egyéb szociális, pszichológiai funkcióira helyezik a hangsúlyt, így **jobban közelítik a pszichológiai realitást**, azaz a valós nyelvi működést. Az érvelésrendszereket az elmúlt évtizedekben alapvetően meghatározó gazdaságossági elv mellett célszerű a pszichológiai realitásnak ugyanakkora vagy inkább nagyobb fontosságot tulajdonítanunk. Időszerű, hogy a pszicholingvisztika számos eredménye alaposabban beépüljön az elméleti modellekbe, amelyeket eddig ezektől viszonylag függetlenül formáltak. A nyelvtudománynak legitim célja lehet az, hogy az emberi nyelv működését nem csak hatékonyan, hanem minél hűebben is kívánja leírni olyan módon, hogy a **pszicholingvisztika kurrens elméleteivel összhangban legyen**. A pszicholingvisztika, kognitív nyelvészet, neurolingvisztika eredményei alapján láthatjuk, hogy célszerűbb az adatok vizsgálatán túl tesztelni, megfigyelni más módszerekkel is a nyelvet, nyelvhasználatot, mert könnyebb ezek fényében jól működő modellt építenünk, mint ha csak a gazdaságosságot tartjuk szem előtt.

Nem szabad elfelejtenünk azt sem, hogy a **nyelvet globálisan nem tudjuk vizsgálni**. A szintaxis, a morfológia, a szemantika is csak egy-egy szeletét ragadják meg a nyelv működésének, és még ezen részterületek egységes vizsgálata is igen nehéz. Ha egy részterület alapján építünk egy maximálisan gazdaságos modellt, akkor annak kiterjesztésekor akár a teljes modellünket is át kell alakítanunk, hisz az újabb adatokat, amelyeket az emberi elme működése hoz létre, már nem feltétlenül fedi akár gazdaságosan, akár bármilyen mértékben. Ez a veszély egy mentálisan reális modellnél kevésbé fenyeget, amely könnyebben bővíthető, mivel a mentálisan reális leírásoknak kompatibilisnek kell lenniük, hisz az emberi agy működésének egyes részfunkciói is kompatibilisek egymással, amelyeket modellezni kívánunk.

Az **analógiás elmélet** a mentális realitást kívánja közelíteni azzal, hogy a **pszicholingivisztika** bizonyos kurrens irányzataival **összhangban van**. Ennek a törekvésnek az eredménye a gyakoriság fogalmának beépítése az elméletbe, amelynek hatásait a nyelvhasználatban pszicholingvisták már évtizedekkel korábban megmutatták (MacDonald 1994, Hare és mtsai 2001). Hasonlóan a kategóriák közti

elmosódott átmenetek elismerése és az ezek magyarázatára irányuló vizsgálódások is ezzel az elvvel vannak harmóniában (Givón 1984). Az analógiás elméletek által feltételezett mintamegőrzés és az ezek alapján való következtetés a tanulásról való ismereteinkkel is jobban összhangban van (Tomasello 2003, Dąbrowska és Lieven 2005), hisz új elemekkel való találkozás esetén minden adatunk megvan ahhoz, hogy bizonyos súlyokat, hatásokat újra ki tudjunk számolni, kezelni tudjuk ezeket. Szabályokat alkalmazva azonban problematikus, hogy új adatok esetén egy szabály/általánosítás miképp módosítható, ha minden korábbi mintát eldobtunk (Langacker 1987).

Fontos azt is szem előtt tartanunk, ha egy elmélet mentálisan reális, még lehet gazdaságos az emberi kommunikáció szempontjából, azaz a két elv nem áll feltétlenül ellentmondásban egymással. Így ha az analógiás megközelítésben nem tételezünk fel mögöttes formákat, hanem csak felszíni alakokat, akkor leírásunk is egyszerűbb, azaz gazdaságosabb lesz, hisz a modellezéshez eggyel kevesebb elméleti konstruktumra van szükségünk. Természetesen önmagában az analógiás megközelítés még nem garantálja valaminek a pszichológiailag reális voltát. Itkonen (1983: 142–152) szerint ilyen esetnek vehető a Montague-nyelvtan is, amely „analógiásnak" tekinthető abban a vonatkozásban, hogy a szintaxist és a szemantikát szigorúan párhuzamosnak tételezi fel, de ettől függetlenül nem gondoljuk, hogy mentálisan reális lenne.

Az analógiás megközelítés létjogosultságát az is alátámasztja, hogy **az emberi gondolkozás számos területén alapvető fontosságú** kognitív folyamatnak bizonyult az analógiás alapú rendszerezés, megértés, produkció (Penn és mtsai 2008, Halford és Andrews 2007). Kroesch (1926: 35) is kiemeli az analógiával kapcsolatban, hogy az **emberi agy szervező funkciójával** van összhangban, így lehetővé teszi a nagy változatosságot mutató nyelvi elemekben is a rendszerezést. Hasonlóan mintaalapon magyarázható az epizodikus memória működése (Chandler 2002: 66), valamint az analógia a valós világ elemeinek és eseményeinek csoportosításában is kulcsszerepet kap, így az angolszász jogban is gyakran alkalmazzák (Lamond 2006).

Amennyiben az analógiát a nyelvészetben elvetnénk vagy alkalmazását csak a nyelvtörténetre korlátoznák, amit az uralkodó nyelvészeti elméletek az elmúlt 50 évben másodlagos fontosságúnak tekintettek, akkor ezzel egy nehezen vagy egyáltalán **nem**

megmagyarázható aszimmetriát hoznánk létre. A nyelv lenne az emberi gondolkozásra kifejezetten jellemző folyamatok közül talán az egyetlen, amelyben az analógia nem, vagy csak marginális szerepet kapna. Így ebben az esetben modellünk egy magasabb szinten a gazdaságosság elvét is megsértené, hisz nem gazdaságos két eltérően működő rendszerrel kezelni nyelvi és nem nyelvi funkciókat, amikor az eggyel is megoldható lenne.

1.2. Célkitűzéseim

Ha bármely nyelv analógiás nyelvtanát kívánjuk megírni, annak egyik alapfeltétele az, hogy tudjuk, mely fonémák-hangok, alakok (Bybee 2001), összetételi elemek (Krott 2009: 132), konstrukciók (Fillmore és Kay 1987, Goldberg 1995, 2006) hasonlóak az adott nyelvben, és ezek **hasonlósága milyen mértékű, min alapszik**. Ehhez azonban meg kell határoznunk, hogy a hasonlóságot milyen paraméterek mentén mérjük, és hogy a nyelvi működésben milyen tulajdonságokban, viszonyokban számít egyáltalán a hasonlóság. Ezek ismerete nélkül az analógiás vizsgálatok nehezen megfogható spekulációkká válhatnak (Bybee 2010: 62).

Vizsgálódásaink szempontjából a **viszonyok hasonlósága** épp olyan fontos, mint maguknak az elemeknek a hasonlósága. Ezeket azonban az elemek hasonlósági mértékének ismerete nélkül nem tudnánk megállapítani. Másrészt az analógiás nyelvtanok feltételezése szerint a hasonló elemek gyakrabban rendelkeznek hasonló viszonyokkal, amelyek felismerése ezen elemek közt valószínűbb és könnyebb is. Ez nem zárja ki azt, hogy analógiák találhatók távoli nyelvi elemek között is (pl. egy melléknév hasonlóan módosítja a főnév jelentését, mint egy főnévi szerkezet az ige által kifejezett cselekvését), de ezek tudományosan is elfogadható azonosítása nehezebb, és egyelőre élhetünk azzal a feltételezéssel, hogy a beszélők számára is kevésbé természetesek, hisz jóval több időt és mentális erőfeszítést igényelhet a felismerésük.

Disszertációmban azt kívánom megvizsgálni, hogy a **hasonlóság miképp befolyásolja önállóan és a gyakorisággal interakcióban a nyelvi változást, az**

analógiás kiegyenlítődést és kiterjesztést, illetve a produkciót. A formai hasonlóságok és az ingadozás[1] elemzése során célom, hogy ezek természetét feltárjam, s leírásommal hozzájáruljak az analógiás megközelítés eszköztárának finomításához, pontosságának, egzaktságának növeléséhez. A formai hasonlóságok vizsgálatának **középpontjában a szavak hasonlósági viszonyainak feltérképezése (5. és 7. fejezet) és összehasonlítási módjaik modellezése áll (6. fejezet)**. Egyes vonatkozásaiban újszerű megközelítésemmel modellt kívánok adni az írásomban nem elemzett morfofonológiai jelenségek elemzéséhez mind a magyar, mind más gazdag morfológiával rendelkező nyelvek esetében. Habár sok ponton támaszkodom a korábbi analógiás és részben nem analógiás elméleti modellekre, leírásom nem adaptációja egyetlen már készen meglévő modellnek sem.

A **szavak hasonlóságának vizsgálatában azok felszíni szerkezetét veszem alapul** (Kálmán 2008, Bybee 2001, Fűköh és Rung 2005, Rebrus és Törkenczy 2008[2]). Ezekre támaszkodva az ingadozással és fokozatossággal szorosan összefüggő analógiás kiegyenlítődés és kiterjesztés is jól megragadható jelenségekké válnak (Kraska-Szlenk 2007), amelyek vizsgálati anyagom változását is irányítják. Feltételezésem szerint az analógia alkalmazását további szempontok is meghatározzák (használati mód, jelentés, stb.), de ezekből a legfontosabb a hangtani/fonológiai hasonlóság és a gyakoriság (Lukács 2002). A **felszíni alakok viselkedését** nem a jelenségeket megragadni próbáló rendezett vagy rendezetlen szabályok sora, hanem az alakok **egymással való kapcsolata határozza meg** (Ackerman és mtsai 2009: 56), így lehetővé válik **a nyelv működésének lokális, atomi szintű vizsgálata**, mert azt nem kell általánosabb szabályokból és azok megszorításaiból levezetni. A morfémákra való bontás helyett a szavak, morfémaláncok értelmezése holisztikus módon, Gestalt-alapon is lehetségessé

[1] Az ingadozásnak gyakran lehetnek szociolingvisztikai természetű okai is. Minden bizonnyal a későbbiekben bemutatandó és felhasználásra kerülő korpuszokban ezek a tényezők is szerepet kapnak. Azonban nem az ilyen természetű ingadozás vizsgálata a célom, hanem az olyané, amely a nyelv belső mozgató erőinek tulajdonítható, így kutatásom legfeljebb csak érintőlegesen kapcsolódik a szociolingvisztikához.

[2] Elméleti szinten nem zárják ki teljességgel a mögöttes alakok létét, de elemzésükben ezt tudatosan és felvállaltan kerülik.

válik (Ackerman és mtsai 2009: 58–59, Prószéky és Kis 1999). Habár a legtöbb analógiás elmélet a holisztikus megközelítést részesíti előnyben, elméleti szinten az analógiás megközelítés a szegmentációra is lehetőséget ad.

A vizsgálatom tárgyát képező felszíni hasonlóságokon alapuló analógia az egyik legfontosabb mozgatórugója az **egyes nyelvi elemek közti** szoros és állandó **interakciónak** is (Itkonen 2005). Ha a nyelvben valahol változás következik be, akkor az az erőviszonyokra azonnal hatással van, és ez a rendszer egészének változásához vezet. Ezt a legtöbb 20. századi nyelvelmélet el is ismeri. Ennek megfelelően az egyes állapotok leírásával foglalkozik a szinkrón nyelvészet, míg az ezek közti átmenetek vizsgálatával a diakrónia.

Ez a megközelítés azonban kimondva vagy kimondatlanul azt sugallja, hogy vannak stabil, önmagukban is megfigyelhető állapotok. A nyelv változik, de maga a változás nem alapvető minősége. A generatív nyelvészeti hagyomány **a nyelvi változás és a mögötte ható tényezők háttérbe szorításával** tudta csak elérni, hogy a kristálytiszta szinkrón állapotok vizsgálhatóak legyenek legalább hipotetikus módon, és a nyelvre jellemző állandó mozgás ne okozzon zavart a „laboratóriumi körülmények" közt előállított nyelvészeti rendszerekben. Ennek az állandó mozgásnak azonban vannak a szinkrón állapotban is tetten érhető jelei, az ingadozó, hezitáló és nagy változatosságot mutató alakok, ugyanakkor ezek az instabil jelenségek kihatnak a rendszer többi részeire és a változás korai vagy folyamatban lévő stádiumainak feleltethetők meg, amely viszont a szinkrón állapotokat összekötő diakróniában kap kiemelt szerepet.

Az analógiás nyelvtanok által megfigyelt tények, jelenségek azonban **cáfolják a szigorú és merev szétválasztást.** A nyelv folyamatos változása párhuzamba állítható az egyén nyelvének változásával is, amely sosem statikus, nincsenek benne jól elkülöníthető, és ezáltal izoláltan leírható állapotok (Sankoff és Blondeau 2007, Bybee 2010). A szinkrón állapotok egységességét az is megkérdőjelezi, hogy egy állapot időben felölel gyakran évtizedes, ritkán évszázados távolságban lévő adatokat is. A korábbi nyelvhasználatra utaló archaikus, irodalmi szövegek, szólások és közmondások, amelyek különböző korú és szociális helyzetű beszélőktől származnak, egy másik kor

normáját és rendszerét vegyítik bele a vizsgált időszakába. Ugyanakkor egyes jelenségek gyorsan is változhatnak (bővebben 5.4. alfejezet), és akár évek alatt jelentős különbségeket hozhatnak a nyelvi rendszerben vagy legalábbis egyes szavak használatában, de a túlzottan általános szinkrón állapotok miatt leírásuk és sokszor észrevételük is meglehetősen nehéz.

A szinkrón állapotok merev szétválasztása így könnyebbé tette a leírást, de a valóság továbbra is komplex maradt. A változás és az állapot szétválaszthatatlanságából következik, hogy **a rendszer a maga statikusságában nem létezik**, vagy legalábbis olyan absztrakt fogalom, túlegyszerűsített elméleti konstrukció, amely a nyelvvel való munkára kevésbé megfelelővé teszi. Ha ezt a megközelítést tovább visszük, és **nem választjuk mereven szét a kompetenciát és a performanciát sem**, akkor **leírásunk hitelesebb lesz** (Frisch 1996: 6), miközben a formális megközelítési módról sem kell lemondanunk (Skousen és mtsai 2002).

Ha ezeknek megfelelően a **nyelvi jelenségeket szorosan összefüggőnek** vesszük, és változásukat lényegi elemüknek tekintjük, akkor felvetődik a kérdés, hogy egyáltalán lehet-e és értelmes-e a nyelvnek bármely részjelenségét leírni anélkül, hogy más részleteit ne vennénk figyelembe, hisz az összefüggések feltárása nélkül a jelenség értelmezhetetlen vagy csak részlegesen értelmezhető lesz. Ha részjelenségekre irányítjuk figyelmünk, akkor a leírás ebben az esetben valóban nem lesz tökéletes, de mivel a nyelvi változás teljességét az arról való meglehetősen korlátozott tudásunk miatt semmiképp sem tudjuk megragadni, így mégis kénytelenek vagyunk csak egyes darabjait vizsgálni. A hagyományos megközelítésekkel ellentétben azonban nem állítom, hogy ezeket a részleteket önállóan és pontosan le tudjuk írni, hanem úgy vélem, hogy újabb leírások fényében majd folyamatos kiegészítésre szorulnak a későbbiekben. Adatainkat állandó változásukból kifolyólag sosem tudjuk megragadni, de ez nem bátortalaníthat el minket, hisz a **nyelvészet célja** nem feltétlenül a leírás, hanem a **leírást meghatározó nyelvi folyamatok megértése és azok pszichológiailag reális feltárása**. Az analógiás nyelvtan erre kiváló eszköz, hisz egyszerre alkalmas a szerkezet leírására és a változási folyamatok megragadására is (Itkonen 2005).

A már eddig bemutatott alapelveken túl írásomban határozottan elkötelezem magam amellett, hogy a **nyelv vizsgálatának nagy mennyiségű adatra** kell támaszkodnia (Sinclair 1991, Jurafsky és mtsai 2001), és a reális, valós folyamatokat leíró modelleknek ki kell állniuk a számítógépes tesztelés próbáját is[1] (Skousen és mtsai 2002). Úgy vélem, hogy a sikeres algoritmikus modellezés nélkül egyetlen elmélet megfelelő voltában sem lehetünk teljességgel biztosak, hiába tűnnek azok matematikai, logikai vagy akár pszichológiai szempontból megalapozottnak. Fontos azonban látnunk, hogy az **adatközpontú megközelítés nem áll szemben** azzal a korábban bemutatott alapelvünkkel, hogy a **modellnek mentálisan reálisnak kell lennie**, hisz a pszicholingvisztika kontrollált és sztenderdizált módszerekkel szintén adatokat gyűjt és értelmez. A két megközelítés megfér egymás mellett[2]. Az adatok segítenek abban, hogy jobban megismerjük a nyelvi rendszert, amelyet olyan módon próbálunk értelmezni, hogy a pszicholingvisztika eredményeivel is összhangban legyünk.

1.3. A tartalom áttekintése

Disszertációm hét további fejezetből áll, amelyekben az 1.2. alfejezet célkitűzései mentén mutatom be az analógiás megközelítést, elemzek és modellezek nyelvi jelenségeket. A **2. fejezetben** azokat a **kutatásokat ismertetem**, amelyek vagy elméleti elgondolásaiknak, vagy technikai megoldásaiknak köszönhetően szorosan összefüggnek vizsgálataimmal. Áttekintem az analógiás nyelvi megközelítés kialakulását, alapelveit és a vele kapcsolatos vitákat. Itt tárgyalom az analógia szempontjából különösen fontos hasonlóság és gyakoriság szerepét a nyelvben, illetve a paradigmák felépítésével és változásával kapcsolatos tudományos nézeteket. Ebben a fejezetben kerülnek bemutatásra a magyar szerzőktől származó analógiás kutatások is,

[1] Ennek az elvárásnak jelenleg legfeljebb morfológiai és szintaktikai jelenségek vizsgálatánál tudunk megfelelni. A jelentéstan és a pragmatika adatközpontú tanulmányozásához a magyar nyelv esetében nincs megfelelő korpuszunk.

[2] Kiváló példa erre Bergen (2004) elemzése a hangfestő szavak (*phonesthemes*) viselkedéséről.

amelyekre adataimban és elképzeléseimben leginkább támaszkodok. A paradigmák szerveződését a nyelvi változás folyamatosan átalakítja, ennek áttekintésére is itt kerül sor, amit az analógiás modellezést ismertető **3. fejezet** követ. Ebben bemutatom az analógiás modellezés szempontjából legjelentősebb AM (Analogical Modeling, Analógiás modellezés, Skousen 1989) és TiMBL (Tilburg Memory Based Learner, Tilburgi memóriaalapú tanuló algoritmus, Daelemans és van den Bosch 2005) programokat, valamint röviden kitérek más, kevésbé ismert, de a vizsgálataimhoz kapcsolódó modellezési kezdeményezésekre is.

A **4. fejezetben** röviden bemutatom, hogy miképp látom a **hasonlósági hatások és a gyakoriság funkcióját** egy analógiás nyelvtanban. Külön kitérek a prototípusok szerepére, hogy egyes, valamilyen szempontból kivételes szavak hasonlósága és gyakorisága miképp határozhatja meg és befolyásolhatja más szavak viselkedését. Ezt követően áttekintem, hogy alapelveimmel összhangban miképp algoritmizálható a szavak hasonlóságának mérése újabb és hagyományosabb fonológiai elméleti megközelítéseket figyelembe véve is.

Az **5. fejezetben a magyar hangkivető főnevek viselkedését elemzem** részletesen a már bemutatott analógiás alapelvekre támaszkodva. Vizsgálatomban a legnagyobb digitális magyar gyakorisági gyűjtésre, a *Szószablya Gyakorisági Szótár*ra (Halácsy és mtsai 2003) támaszkodok. Első lépésben a végszekvenciáik alapján tárom fel a hangkivető szavak viselkedésének mozgatórugóit. Ezt követően a hangkivető főnevek hasonlósági csoportokba való szerveződését tekintem át gráfstruktúrákban. Végezetül két nagyméretű szöveges korpusz összehasonlításával (*Szószablya Korpusz*, 2010-es saját gyűjtés a *Google* kereső segítségével) a változás jellegét, illetve a különféle toldalékos alakok sajátos viselkedését is megvizsgálom. Miközben a nagyobb, átfogó összefüggéseket törekszem felfedezni vizsgálatomban, Jules Gilliéron híres mondását is szem előtt tartom, miszerint „minden szónak megvan a maga története". *Google Gyakorisági Gyűjtés*emet 2010 tavaszán készítettem, amely a vizsgálatomban szereplő hangkivető szavak leggyakoribb hangkivetéssel együttjáró toldalékos alakjainak[1]

[1] tárgyeset, többes szám, szuperesszívusz, E.1 birtokos, E.3 birtokos, T.3 birtokos, E.3 birtokos több birtokkal.

gyakorisági számait tartalmazza a *Google* kereső alapján. Ezen adatok elemzésével alaposabban az 5.4. alfejezetben foglalkozok, mivel azonban számos adat ennek fényében jobban értelmezhető, korábban is hivatkozok a *Google Gyakorisági Gyűjtés*re.

A **6. fejezetben hasonlósági algoritmusok hatékonyságának tesztelési eredményeit** mutatom be a már elemzett hangkivető főneveken. Elsőként azt tesztelem, hogy eltérő méretű minták mellett melyik algoritmus mennyire jól osztályoz szavakat. Ezt követően „hagyj-ki-egyet" (*leave-one-out*) és 10-szeres keresztellenőrzés (*tenfold cross validation*) tesztekkel vizsgálom meg az algoritmusokat, és bemutatom, hogy az analógiás elveket algoritmizáló programok képesek főnevek viselkedésének meghatározására a Levenshtein-algoritmusnál (Levenshtein 1966), szabályoknál és gépi tanuló algoritmusoknál is hatékonyabban, amelyek közül egyedül a maximum entrópia modell hoz csak hasonlóan jó eredményeket. Végezetül a prototípusokkal kapcsolatos elképzeléseim algoritmikus megvalósítását vizsgálom annak függvényében, hogy gépileg kiválasztott prototípusokhoz viszonyítva hatékonyan modellálható-e a nyelvben tapasztalható ingadozás.

A **7. fejezet** egy olyan **nyelvi tesztet** mutat be, amely azt vizsgálja, hogy eltérő pozíciókban a fonémák mennyiben járulnak hozzá két szó hasonlóságának a megítéléséhez, abból a feltételezésből kiindulva, hogy minél jobban hasonlít egy álszó egy valódi szóhoz viselkedésében, annál közelebbi lesz ahhoz alaki hasonlóságuk alapján. A vizsgálatnak ezen túl az is célja volt, hogy kimutassam, milyen tényezők befolyásolják együttesen egy új nyelvi elem viselkedését, amelyekből a legjelentősebbnek a szerkezetileg leghasonlóbb szavak analógiás hatása bizonyult. Nyelvi tesztemet 91 *CVCVC* szerkezetű hangkivető főnév alapján végeztem el összesen 116 magyar anyanyelvi beszélővel. A **8. fejezet** a disszertáció tanulságait és eredményeit foglalja össze, amelyek kapcsán felvázolom, hogy kutatásaimat miképp lehetne továbbvinni a jövőben.

2. Kapcsolódó kutatások

2.1. Az analógiás nyelvleírás alapjai, alkalmazási területei

A magyar *analógia* szó a görög *ἀναλογία* 'arányosság, megfelelés' szóból származik. Az analógiás következtetésben alapfontosságú **analógiás párokról** már **Arisztotelész**[1] (1965: XXI) is említést tesz:

„Analógiának nevezem azt, ha a második úgy viszonylik az elsőhöz, mint a negyedik a harmadikhoz – a költő tehát a második helyett a negyediket, vagy a negyedik helyett a másodikat használja, és esetleg hozzáteszi azt a névszót is, amelyre a metafora vonatkozik, s amelyet ez helyettesít. [...] Vagy például az öregség úgy viszonylik az élethez, mint az este a nappalhoz; a költő tehát az estét a nappal öregségének nevezheti, vagy – mint Empedoklész – az öregséget az élet estéjének vagy alkonyának."

Itkonen (2005: 15–22) az analógiának **négy típusát** határozza meg aszerint, hogy **episztemológiai és ontológiai** szempontból szimmetrikusak avagy aszimmetrikusak összehasonlításunk tárgyai. A nyelvészet számára kettő jelentős ezek közül. Az egyik esetben két összehasonlított dolog egyformán ismert (episztemológiailag szimmetrikusak), de az egyik a másik után jön létre (ontológiailag aszimmetrikusak). Ez a viszony áll fenn az imitáció, a rutinszerű alkalmazás esetében, amely nem központi témája írásomnak. Számunkra jelentősebb változata az analógiának az, amikor egy **ismert és létező elem hatására egy ismeretlen, még nem létezőt hozunk létre**, hisz ez maga a **nyelvi változás**.

Marcus Terentius Varro (Kr.e. 116 – Kr.e. 27) felismerte az analógia központi szerepét, és azt az emberi gondolkozásban az alaktani leírásban az elsők közt[2]

[1] Az analógiás gondolkodással szemben álló szabályalapú megközelítés gyökerei szintén az ókorba nyúlnak vissza, amelynek alapjait Pāṇini rakta le (Buiskool 1939).

[2] Varro említést tesz alexandriai Arisztarkhoszról (Kr.e. 216 körül – Kr.e. 144 körül) és követőiről, mint az analógián, arányításon alapuló nyelvi elemzés korai gyakorlóiról (Itkonen 2005: 79, Dinneen 1995: 138).

alkalmazta. Varro szakított azzal a megkötéssel is, hogy az analógia csak négy elemre alkalmazható (Itkonen 2005: 13). A **qiyas** (‘arányítás’ قـياس), az analógián alapuló érvelés **az arab nyelvtanírók leírásaiban** is nagy szerepet kapott, melynek alkalmazásakor egy olyan új helyzetet azonosítottak, amely valamilyen korábbira hasonlított (Bohas és mtsai 1990: 24). Az adatok újszerű elrendeződésében azokat a mintázatokat keresték, amelyekkel már korábban találkoztak, és viselkedésüket feltárták elemzéseikben (Bohas és mtsai 1990: 23).

Az analógia a 19. században a kivétel nélkül ható hangtörvények[1] mellett az **újgrammatikusok egyik legfőbb elméleti konstruktuma** volt, amellyel nem egyszerre lezajló, szórványos változásokat magyaráztak. A korabeli analógiás nyelvelemzés alkalmazási lehetőségeinek legalaposabb összefoglalását Wheeler (1887) készítette el, aki gazdag példaanyagra támaszkodva rendszerezte, hogy az analógia milyen területeken milyen feltételek mellett működik. A **20. század eleji nyelvésznemzedékek** tagjai még **nagy jelentőséget tulajdonítanak az analógiának**[2]. Sapir (1921: 37) úgy véli, hogy új szavak és mondatok folyamatosan a korábbiakon alapuló analógia révén jönnek létre. Hozzá hasonlóan Jespersen (1965 [1924]: 19) is használja az analógiát mondatok elemzésében, míg Saussure (1962 [1916]) az analógiát a szinkrón rendszer összetartó erejének tartja. A mondatelemzésben alkalmazott mintaalapú megközelítés elsősorban a konstrukciós nyelvtanokban él tovább (Fillmore és Kay 1987, Goldberg 1995, 2006). Az analógia fogalmát Bloomfield (1933: 275) határozza meg alapvetéseiben az én elképzeléseimhez is hasonló módon:

„A nyelvtani mintákat (mondattípusokat, konstrukciókat vagy behelyettesítéseket) gyakran analógiának nevezik. **Szabályos analógia segítségével a beszélő olyan megnyilatkozásokat hozhat létre, amelyeket korábban sosem hallott.**

[1] Philips (2006) megmutatja, hogy a hangtörvények sok esetben nem kivétel nélküliek, és gyakran az analógia is szerepet kap bennük.

[2] Skousennek (1989) az 1.1. alfejezetben bemutatott összehasonlítása ezért is volt pontatlan elnevezésében, mert a strukturalisták még hagyatkoztak az analógiára szinkrón elemzéseikben is.

Azt mondhatjuk, hogy ezeket a megnyilatkozásokat a korábban hallott, hasonló formákkal való analógia alapján hozza létre."

Bloomfield az újgrammatikusok hagyományát követve csak **az analógiás párokat tartotta az analógiához tartozónak**, más eseteket adaptációknak[1] vesz, vagy ha két szerkezet egymásra hatásából egy harmadik, egyikre sem teljesen hasonlító szerkezet jön létre, akkor azt kontaminációnak[2] nevezi (Itkonen 2005: 42). Bloomfieldhez hasonlóan vélekedik Hockett is (1966: 94). Elemzése szerint a *swammed* alak kontaminációval jött létre a *sigh* : *sighed*, *swim* : X, illetve a *sing* : *sang*, *swim* : X aránypárok mentén. Bolinger (1961: 378) a kontaminációval kapcsolatban említést tesz arról, hogy az alapját nyújtó elemeknek legalább valami magasabb absztrakt szinten közös tulajdonságokkal is kell bírniuk. A kontamináció általános működési körülményeit, módjait azonban nem definiálják pontosan, ami pedig különösen fontos lenne, hisz a kontaminációnak köszönhetően a felismerhetőséghez szükséges fonológiai anyag egy része is eltűnik.

Az **analógia globális működése, a hasonlóságok felismerésének módja, a gyakoriság hatása** a nyelvi változásra azonban ezekben az időkben még **nem volt megfelelően tisztázott**. Az analógiás párok kiválasztásának módja többnyire intuitív és ad hoc módon történt. A szerzők választásaikat csak ritkán indokolják, és ezekben az esetekben is csak egy pontosan meg nem határozott hasonlóság-fogalomra hivatkoznak. Nem derül ki a művekből, hogy A forma B formához miben hasonlóbb, mint C forma. Ez azonban nem csak a 19. századi művekre jellemző, hanem a későbbi munkákban (egészen az 70-es évekig) sem kapunk konkrét fogódzókat azzal kapcsolatban, hogy

[1] Adaptációval jött létre a latin *rendere* 'visszaad' a *reddere*-ből a *prehendere* 'elvesz' és/vagy a *vendere* 'elad' hatására.

[2] A kontaminációs példák Simonyi (1881) korábbi, analógiával kapcsolatos értekezésében is fontos szerepet töltöttek be.

mikor és miképp tekinthető két alak hasonlónak. Ezzel rokon módon a korai analógiás művekben a szemantikai közelség meghatározása is képlékeny és esetleges[1].

A hasonlóság fogalmának homályosságán túl nem bontakozik ki a művekből, hogy ha nem szabályokkal modellezzük a működést, akkor **hogyan hatnak az elemek egymásra** a szinkrón rendszerben. Így Wheeler (1887) is számtalan példát hoz rendszerezése alátámasztására kihalt (latin, ógörög, szanszkrit stb.) és jelenkori nyelvekből, de ezek rendszerét nem tárja elénk. Igaz, az analógiára támaszkodó leírások elsősorban a diakrón állapotok összehasonlításával foglalkoznak, de az a szerzők műveiből egyértelmű, hogy a szinkrón állapotok működésében is kiemelkedő szerepet szánnak neki, nem tudni azonban pontosan miképp.

A legtöbb korai analógiás mű **nem támaszkodik a gyakoriság fogalmára** sem. Azt azonban szükséges megjegyezni, hogy már Wheeler (1887: 21) is hangsúlyozza ennek fontosságát. Wheeler a gyakoriságot olyan fontos faktorként említi, amely erősíti emléknyomainkat egy nyelvi elemmel, jelenséggel kapcsolatban. Megközelítésébe belefér más olyan tényezők beépítése is, amelyek ezt a hatást fokozhatják (pl. gyakoriság mellett az elemek, jelenségek frissessége, feltűnősége). A legújabb idők nagy korpuszainak megjelenéséig azonban ennek vizsgálatára nem igazán volt mód, így a gyakorisági hatások tanulmányozásának elmaradásáért nem a kor szerzői okolhatók közvetlenül. A korai analógiás megközelítések képviselői szintén hangsúlyozzák a használat fontosságát, de a használat beépítése az elméletekbe homályos és ad hoc marad.

Ez a gyakorlat váltotta ki az analógia használatának heves bírálatát, amelyekre válaszként az analógiás nézetek képviselői szigorúbb formalizmust és a nyelvi rendszer átfogóbb, mentálisan is motiváltabb megközelítését alakították ki. Mielőtt azonban részletesebben tárgyalnám azt, hogy az analógiáról való gondolkozás miképp változott meg a korábbi homályos vagy túl általános elképzelésekhez képest, fontos

[1] A szemantikai hasonlóság meghatározása még bizonytalan a mai analógiás megközelítésekben is, ami nehezen finomítható, hisz a fonológiai tulajdonságokhoz viszonylag könnyű a hozzáférésünk, de gépi feldolgozásra alkalmas, jelentéselemeket tartalmazó szótáraink, különösen a magyar nyelvre, nincsenek még.

meghatározni az **analógiás változás/következtetés két fontosabb típusát** (analógiás kiegyenlítődés és kiterjesztés), és áttekinteni az analógia további alkalmazási területeit a már bemutatottakon túl.

Az **analógia alkalmazásakor** a beszélő szűkösebb nyelvi ismeretei (ritka szó, nem anyanyelvi beszélő, vagy a nyelvelsajátítás fázisában van), felejtés, vagy emlékezeti-aktivációs korlátai miatt megváltoztatja egyszeri vagy rendszeres módon egy szó bizonyos alakját, így a konvencionálistól eltérő formában használja más szavak, vagy e szó más alakjainak hatására. Hasonló módon a beszélő még nem létező alakokat is képes létrehozni már létező alakok mintájára. Az ilyen analógiás folyamatok sokszor, még ha bizonyos közegben természetes változásként is hatnak, általánossá válásukkor elveszthetik fonetikai motivációjukat (Philips 2006).

Az **analógiás kiegyenlítődés** során egy szó tőváltozatai közül az egyik olyan paradigmatikus cellákban is megjelenik, ahol korábban egy másik tővariáns volt használatban. Az analógiás kiegyenlítődés eredménye lokális szinten mindig egyszerűsödés. Az analógiás kiegyenlítődés fő motivációja a tő uniformitásának megőrzése, amely kapcsolatba hozható az optimalitáselmélet kimenet-kimenet megfeleltetési (output-output correspondence, OOC) megszorításával is (Myers 2002, Prince és Smolensky 1993, Hale és mtsai 1998). Az analógiás kiegyenlítődés hatására (Kraska-Szlenk 2007) a lengyelben a lokatívusz esetében is megjelent az alanyesetre jellemző tővariáns:

(1)

korábbi			**későbbi**			
ʧɑ̲s (nom.)	:	ʧe̲ɕe (lok.)	ʧɑ̲s(nom.)	:	ʧɑ̲ɕe (lok.)	’idő’

Az **analógiás kiterjesztésben** a szavak egy másik, valamilyen módon erősebb csoport[1] mintájára kezdenek el viselkedni. A magyar ikes igék esetében az analógiás kiterjesztés hatására azok paradigmája változatlan bonyolultságú marad, míg a globális rendszer egyszerűsödik[2]:

(2)

korábbi			**későbbi**		
fut	:	futo<u>k</u>	fut	:	futo<u>k</u>
hetvenkedik	:	hetvenkede<u>m</u>	hetvenkedik	:	hetvenkede<u>k</u>

Az analógiás kiterjesztés célja elsődlegesen a **globális rendszer egyszerűsítése**, amely lokális szinten egy szó paradigmáját bonyolultabbá is teheti. Ez történik, amikor rendhagyó mintákat terjesztünk ki[3] (lásd *motrot*, *bútrot* esete), hisz a *motor* és a *bútor* egy kevésbé homogén paradigmába sorolódik, viszont a '*-tor* végű szavak hangkivetők' általánosítás erősebb lesz, így az egyes *-tor* végű alakok paradigmatikus hovatartozása könnyebben jósolhatóvá válik. Az ilyen kiterjesztés gyakrabban fordul elő olyan alakokkal, amelyeknek vannak közös fonológiai jegyeik a rendhagyó mintát megtestesítő csoportokkal (Bybee és Moder 1983, Long és Almor 2000, Albright és Hayes 2002, Wedel 2009: 91). Az analógiás folyamatok eredménye nem mindig

---

[1] A csoporterősség viszonylagos fogalom, amelyet a csoport elemeinek számossága, használatuk gyakorisága és a csoporton belüli elemek fonológiai, jelentéstani és esetleg még más szempontok szerint számított hasonlósága határoz meg. Minél több közös jegyet tartalmaz egy halmaz, annál valószínűbben lesz az analógia alapja (Skousen 1989). A csoporterősség ilyetén megközelítéséből következik, hogy egy kevesebb elemet számláló, de nagymértékű hasonlóságot mutató elemeket tartalmazó osztály (gang) is lehet korlátozott mértékben produktív.

[2] Amíg a kiegyenlítődésen nem megy át minden ikes alak, addig a rendszer lokálisan viszont bonyolultabb lesz, hisz a beszélő számára az ikes végződés nem lesz biztos előrejelzője a szó paradigmájának, amíg vannak kiegyenlített és nem kiegyenlített ikes igék.

[3] A 7. fejezetben bemutatásra kerülő nyelvi tesztem is azt mutatja, hogy az anyanyelvi beszélők képesek a hangkivető mintát aktívan használni is, ugyanakkor a minta alkalmazásának sikerességét számos faktor határozza meg. Rendhagyó séma kiterjesztésére Jespersen (1942) mutat angol igei példákat.

egyértelmű, így a *bajusz* tárgyesetének kétféle kiegyenlített alakja is lehet, eltérően viselkedő analógiás forrásainak a hatására:

(3) *bajszot -> bajuszt, bajuszot*[1]

Az **analógiás kiegyenlítődés**, amennyiben más alakok mintájára szorul vissza egy tővariáns használata, tekinthető az **analógiás kiterjesztés egy sajátos esetének** is. Ilyen kiterjesztési folyamatnak vehető a hangkivető szavak hangkivetéses tővariánsának visszaszorulása, amikor ezek a szavak egy másik paradigmába sorolódnak át, de ennek következtében a paradigmán belüli alakjaik hasonlóbbak is lesznek. A folyamat nem tekinthető csak analógiás kiegyenlítődésnek, mert a hangkivető főnevek összes alakját (képzetteket is) figyelembe véve a hangkivetést tartalmazó alakok aránya 49,7% a *Szószablya Korpusz*ban, így kizártnak tekinthető, hogy a beszélők számára minden bizonnyal nem érzékelhető 0,6%-os eltérés a kiegyenlítődési folyamatot önmagában elindítsa. Azonban az 5. fejezetben látni fogjuk egyes jobban ingadozó szavaknál, hogy összes alakjukat figyelembe véve hangkivetéses alakjaik aránya az átlagosnál kisebb, így változásukban a kiegyenlítődés már szerepet kaphat.

A már bemutatott alaktani jelenségek értelmezésében játszott szerepén túl az analógiás megközelítés a nyelvtudomány más területein is a szerveződés és a működés egyik mögöttes tényezője, mozgatórugója. Így a fonológiában szerepet kap a **fonemikus rendszer szerveződésében**, hisz nincs szembenállás hasonlóság nélkül. Az egyes szemben álló fonémapárok egymással párhuzamos strukturális hasonlóságot mutatnak: *p:b, t:d, k:g* stb. (Householder 1971: 65–67). A hasonlóság szerepet kap bizonyos hangfestő szavak (*phonesthemes*) szerveződésében is (Firth 1930), amikor bizonyos fonémaláncok, fonémák analógiák alapjává válhatnak azáltal, hogy a fonémaláncokhoz jelentés kapcsolódik (Hutchins 1998, Magnus 2000, Abelin 1999), így az angol *gl-* 'fény és látás érzete' esetében: *glitter* 'fénylik', *glisten* 'csillog', *glow* 'ragyog', *gleam* 'felcsillan', *glint* 'felvillan' stb. E jelenség pszichológiai realitását Bergen (2004) előfeszítéses pszicholingvisztikai vizsgálatokkal bizonyította az olyan hangfestő szavaknál, amelyek

[1] A *Szószablya Korpusz* 3, a *Google Gyakorisági Gyűjtés* 148 ilyen alakot tartalmaz.

a szókészlet vizsgált tartományában jelentős arányban fordulnak elő (a Brown Korpuszban típusok alapján 38,7%-ot, példányok alapján 59,8%-ot tettek ki a 'fény és látás érzete' jelentéshez kapcsolható alakok a *gl-* kezdetű szavak között). Kálmán és mtsai (2010) semleges magánhangzót is tartalmazó magyar szavak harmonizáló toldalékokkal való viselkedésére nyújtanak plauzibilis magyarázatot analógiás alapon. Az analógia a hangtörvényekkel jellemezhető folyamatokban is szerepet kap (Philips 2006, Anttila 1989: 88), mivel az egyes változási folyamatok közt a különbség inkább mennyiségi, semmint minőségi, míg Givón (1995: 95) és Kiparsky (2005) a nyelvi változásban nagy szerepet játszó grammatikalizációt magyarázza analógiásan.

A **mondattanban** Fillmore és Kay (1987), illetve Goldberg (1995, 2006) nyomán a konstrukciós nyelvtan képviseli az analógiás gondolkozást. Mivel a konstrukciós nyelvtanban az analógia viszonylagosan magas absztrakciós szinteken is működik, így nemcsak új szavak, hanem konstrukciók között is teremthető analógiás kapcsolat, illetve ugyanannak a konstrukciónak más elemtartalommal való újraalkalmazása is ugyanezen elvek mentén történik. Az analógia **szemantikai** alkalmazására példa a térre alkalmazott nyelvi formák használata az időre vonatkoztatva (Itkonen 2005: 38). Kálmán (2010a) olyan konjunktív és diszjunktív szerkezetek értelmezésében alkalmazza az analógiás megközelítést, amelyek esetében nem számolható ki a valódi kommunikációban szereplő jelentés. A nyelv ikonikus jellege pedig lehetővé teszi, hogy **nyelvi és nem nyelvi elemek közt** is analógiát vonjunk (Itkonen 2005: 103, Wittgenstein 1922). Az analógiás megközelítést már több **nyelvtechnológiai feladatra is alkalmazták**, mint gépi fordítás, szófaj-egyértelműsítés (POS tagging), morfológiai elemzés (Daelemans és van den Bosch 2005: 73–80), mondathatárok azonosítása (Daelemans 2002), fordító memóriák (Daelemans és van den Bosch 2005: 23).

2.2. Az analógiás nyelvi megközelítéssel kapcsolatos viták

Az analógiás megközelítés létjogosultsága melletti érvelés nem elsődleges célja disszertációmnak, mégis **szükségesnek tartom az ellenérveknek és azok cáfolatának**

rövid bemutatását. Az analógiás megközelítést illető bírálatokat és azok részletes és kimerítő cáfolatát Itkonen (2005: 68–76), Skousen és mtsai (2002), illetve Blevins és Blevins (2009a) mutatják be. Ezen kritikák többsége az analógiás megközelítés félreértelmezéséből fakad, de az analógia nyelvészeti megítélését mind a mai napig meghatározzák, annak ellenére, hogy többször megmutatták gyengéiket, illetve, hogy több olyan szerző is újra alkalmazza elemzéseiben az analógiát (pl. Kiparsky 2000, 2005[1]), akik korábban azzal szemben kritikusan léptek fel.

Mint a 2.1. alfejezetben is láttuk, a **60-as évekig** a nyelvi jelenségek analógiás magyarázata **általános és elfogadott volt a szinkrón leírásokban** is[2]. A fordulat **Noam Chomsky** fellépéséhez köthető: az ő **generatív megközelítésében az analógia kikerült a szinkrón nyelvi leírásból**, annak ellenére, hogy az analógia szempontjából kulcsfontosságú, felszíni formákat előtérbe helyező „paradigmatikus uniformitás" jelentőségét Kiparsky (1972) már ekkor is hangsúlyozta. Így a 70-es években az analógia használata a nyelvtörténet jól behatárolható területeire korlátozódott (Anttila 1977), ahol a hangváltozások elveivel (materiális szempontok) szemben egyensúlyt (strukturális szempontok) teremtett, mivel úgy gondolták, hogy a szinkrón leírásban minden analógiával jól magyarázható jelenség generatív szabályrendszerekkel is leírható. Chomsky érveit azért fontos röviden áttekinteni, mert bár természetüket tekintve nem közvetlenül a morfológiai leírás gyakorlatára vonatkoznak, azokra áttételesen még ma is hatással vannak.

Az analógiát bíráló heves megjegyzéseit megelőzően **Chomsky** (1975 [1955]: 131) még a nyelvtudományon belül feladatként jelöli meg egy olyan elemzési módszer megalkotását, amely lehetővé teszi, hogy mondatok alapján szerkezeti mintázatokat fedezzünk fel, és a **korábbi mondatok alapján olyan új mondatokat hozzunk létre**, amelyek megfelelnek ennek a mintázatnak. Később azonban Chomsky (1965: 47–49)

---

[1] Kiparsky az analógiát a nyelvi rendszer optimalizálásának eszközeként használja elemzéseiben. Vázlatban lévő 2005-ös keltezésű munkájában az analógiát már teljes értékű magyarázóelvként alkalmazza, sőt a grammatikalizációt is egy sajátos, nem példányalapú analógiaként értelmezi.

[2] Mivel dolgozatom a szinkrón nyelvállapotban ható mechanizmusok feltárásával foglalkozik, ezért az analógia szerepére a diakrón nyelvészetben csak röviden fogok kitérni King (1969) kapcsán, akinek érvei is elsősorban nem diakrón, hanem általános jellegüknél fogva érdekesek számunkra.

már markánsan bírálja az analógiát, amit évek múltán is megerősített (Chomsky 1986: 12).

Chomsky bírálatai elsősorban az analógia **nyelvelsajátításban** játszott szerepét vonták kétségbe. Vitatta, hogy a gyerekek már meglévő megnyilatkozásokból általánosított minták alapján értenek meg és alkotnak új kifejezéseket. Chomsky kifogásai abból indulnak ki, hogy a **nyelvelsajátítás során a gyerek kevés adatból nem képes analógiás általánosításokat létrehozni**, ezért a nyelvelsajátítást velünk született tudásnak kell támogatnia. Chomsky nem veszi figyelembe azonban, hogy az analógiás elképzelések sem tagadják, hogy bizonyos ismeretek velünk születettek lehetnek. Ezekre a kritikáira válaszképp a nyelvelsajátítással kapcsolatos kutatások megmutatták, hogy a korai beszédprodukció jól magyarázható analógiával, ugyanis a gyerekek megnyilatkozásainak egyharmada pontos imitáció, összes megnyilatkozásuk 80%-a pedig korábbi megnyilatkozások ismétlése legfeljebb egy analógiás változtatással (Lieven és mtsai 2003, Dąbrowska és Lieven 2005). Ezt támogatja az is, hogy az analógia egyre nagyobb szerepet kap a kognitív nyelvi leírásokban is a nyelvelsajátítással kapcsolatban, így Pinker (1994: 417) is hangsúlyozza, hogy a gyerekek általánosításait a hasonlóság irányítja.

Az analógiás megközelítések szerint a gyerekek a **nyelv elsajátításában** a hallott szövegeket Chomsky elképzeléseinél **rugalmasabban kezelik**, így kezdetben a rendelkezésükre álló adatokból rendhagyó alakok szabályos alakjait analógiás kapcsolatok alapján hozzák létre, amelyeket később nagyobb minta ismeretében javítanak. Ha egyediek a rendhagyó alakok, akkor azokat valóban memorizálják, de hasonlóságokat mutató szavaknál már fellépnek a jól ismert csoporthatások, amelyek könnyítik a hatékony elérést. Chomskyval ellentétben ezek az elméletek csak annyit állítanak, hogy ha egy bizonyos mintázat nem kerül megerősítésre, vagy azzal ellentmondó adatokkal találkozunk, akkor annak használatát megszoríthatjuk vagy éppen elvethetjük. A rendszer változása Sankoff és Blondeau (2007) kutatásával összhangban sosem jut el egy optimális nyugvópontra, legfeljebb a változás tempója lassul, hullámzik. Derwing és Skousen (1994) kísérletében azt is megmutatta, hogy a nyelvelsajátításban a szókincs bővülésének különböző szintjei analógiás alapon jól

modellezhetők, így az 1.2. alfejezetben támasztott modellezési elvárásoknak is megfelel az analógiás megközelítés.

Chomsky (1986) egy **mondattani példán** keresztül is igyekszik megmutatni, hogy az analógiás következtetés hibás eredményre vezet, így alkalmatlan bármilyen szerep betöltésére a nyelvi elemzésben. Levezetése azonban **azon a tévedésen alapul, hogy az analógia bizonyos elemet bizonyos elemmel helyettesít,** és így kapja az új alakot, ellentétben azzal, hogy az analógia valójában strukturális hasonlóságok felismerésére törekszik, és azokat alkalmazza új alakok létrehozásában és értelmezésében. Tévedését Itkonen (2005: 90–91) alapos elemzésben mutatja meg.

Chomsky szerepét alapvetően negatívan ítélhetnénk meg, amennyiben elfogadjuk, hogy az analógiás leírások valóban nagy létjogosultsággal bírnak, azonban fontos azt is figyelembe venni, hogy **kritikái aktívan hozzájárultak** ahhoz, hogy az **analógiás megközelítések formalizáltak, mentálisan megalapozottabbak** és valós nyelvi adatok nagyobb tömegére jobban alkalmazhatóak legyenek, így fellépéséből hosszú távon még az analógiás megközelítések is profitáltak.

Kiparsky korai műveiben (1974: 259) az analógiát a **morfológia felől bírálta**. Állítása szerint az analógia túl erős eszköz, így lehetséges *ear:hear = eye*:**heye* (‘fül’:‘hall’ =‘szem’:‘lát’-nak megfelelő nem létező alak) analógiás párt is felállítani. Kiparsky példájával azonban csak arra mutatott rá, hogy nem minden analógia helyes, vagy használható a nyelv valóban kreatív alkalmazásában. Esetünkben nincs /h-/ prefixum, amellyel igéket képezhetnénk, ami a beszélő számára egyértelmű a rendelkezésére álló nyelvi anyagból[1]. Kiparsky (2005) a bizonyos korlátok közé szorított analógiát ennek megfelelően már aktívan használja elemzéseiben. Kiparsky (1975) megjegyzi, hogy a kontextusfüggetlen analógia, amelyet az analógiás nyelvtanok egyébként sem alkalmaznak, képtelen megjósolni, hogy a *dog* ‘kutya’ többes száma vajon [-s] a *cats* ‘macskák’, [-z] a *birds* ‘madarak’ vagy [-iz] a *fishes* ‘halak’ mintájára. Azonban azt elismeri, hogy az analógia kontextusra érzékeny változata, amelyet az analógiás

[1] Ha a Kiparsky által hozott példa több esetben is előfordulna gyakori párokban, akkor már lehetséges lenne egy ilyen analógia, azonban ezek hiányában a beszélő nem hoz a bemutatotthoz hasonló következtetéseket.

megközelítések alkalmaznak, minden olyan nyelvi jelenséget képes magyarázni, amelyek szabályokkal is lefedhetők (Kiparsky 1975: 189)[1]:

„Amikor az analógiák alapján helyes következtetésekre jutunk, akkor azok a szabályoktól megkülönböztethetetlenek lesznek."

Ezt a megállapítást Krott (2009: 118) is megerősíti annak kapcsán, hogy egy új összetett szó, amely látszólag szabály alapján jött létre, tulajdonítható analógiás alkotásnak is. Így láthatjuk, hogy **a kontextust figyelembe vevő analógiás nyelvtanok lefedik a generatív nyelvtanok által alkalmazott szabályok rendszerét** (Skousen 1989: 54–60, Krott és mtsai 2002: 181, Eddington 2003, Bybee 1995, 2001, 2010, Skousen és mtsai 2002), ugyanakkor alkalmazásukkal azt nyerjük, hogy a nyelvi működést egy mentálisan reális, a gondolkodás más területein nagy jelentőséggel bíró folyamattal magyarázzuk. Az analógiás megközelítésnek már korábban hangsúlyozott további előnye a szabályrendszereken alapuló leírással szemben, hogy jól magyarázza az opcionalitást, az ingadozást és a gyakorisági hatásokat is (bővebben 1.1., 2.6., 2.7. alfejezetek).

Habár dolgozatomnak nem tárgya az analógia működését vizsgálni két időben eltérő nyelvállapot között[2], mégis érdemesnek tartom a **nyelvtörténeti leírásban fontos szerepet betöltő King (1969) érveinek megvizsgálását** is, amelyek jelen tudásunk alapján megkérdőjelezhetőek, de az előzőekben bemutatott érveknél komplexebbek. King (1969) elméletben elismeri, hogy az általa bevezetett eszköztár (pl. szabályvesztés, szabálysorrend-átrendezés) helyett akár analógiát is alkalmazhatna adatai elemzésére, de ezt követően több ellenpéldát hoz, amelyek látszólag azt mutatják, hogy az analógia mégsem alkalmas az általa bemutatott diakrón jelenségek magyarázatára.

---

[1] A nyelvi rendszer szabályokkal nem lefedhető részrendszereire már a generatív megközelítések is elfogadhatónak tartják az analógiát (Pinker 1991, Jackendoff 2002).

[2] Skousen (1989: 124–135) a finn igék múlt idejű rendszerében bekövetkező változások szimulálásával megmutatta, hogy az analógiás modellezés diakrón jelenségekre is alkalmazható.

Elsőként King (1969: 132) azt kívánja megmutatni, hogy analógiás alapon nem magyarázható, hogy az **angol *caru* 'gondoskodás' miért adta fel *cara* többes számát**, hogy többes száma *-s*-re végződjön, mivel szerinte a magánhangzós *cara* nem változhatott mássalhangzó végű szavak mintájára. King azonban nem adja meg, hogy mikor vesztette el a *cara* az utolsó magánhangzóját, így akár lehetett mássalhangzós végű is, amikor analógiásan paradigmát váltott, másrészt ha egy csoport lényegesen erősebb, gyakoribb, mint más csoportok, akkor annak viselkedési módja analógiásan terjedhet a nagyobb fonológiai távolság ellenére is.

A szöveg alapján az *a*- tövek csoportjának méretét nem tudjuk meg, sőt King (1969: 129) maga is beismeri, hogy a gyakoriságról igen keveset tud általánosságban (a korban természetesen nem ő az egyetlen). Ebből kifolyólag King levezetése nem védhető, hisz ez csak akkor állna, ha biztosan tudnánk, hogy a beszélt angol köznyelvben a kérdéses időperiódusban az *a*- tövek nem domináltak. Ebben az esetben ugyanakkor arra a kérdésre kellene felelnie, hogy mi lehet az oka annak, hogy egy olyan szabályt kezdtek el használni a beszélők, amely nem a legszámosabb csoportra volt jellemző. Azaz, ha King feltételezései helyesek (az *a*- tövek csoportja nem alkalmas az analógiás kiterjesztéshez), akkor azzal a szabályalapú magyarázatot is gyengíti, és az *a*- tövek toldalékolásának elterjedése érthetetlen misztérium marad. Ha pedig mégis ez a csoport volt a legdominánsabb, akkor ez a tény már az analógiás magyarázatot is támogatja, hisz ebben az esetben az analógiás kiterjesztés körülményei adottak lennének.

King **másik ellenpéldájában** azt állítja, hogy analógiásan nem magyarázható, hogy a **gót** *kuisan, kaus, kuzum, kuzans* 'választ' típusú szavak középső fonémája (mögöttes /s/), miért a zöngétlenedés irányába mozdult el (*kusum, kusans*), amikor mind a zöngés, mind a zöngétlen változatokra is van példa a gótban (*greipan, graip, gripum, gripans* versus *steigan, staig, stigum, stigans*). Ezt a jelenséget ő szabályvesztéssel (szóbelseji zöngésítés) magyarázza, elvetve az analógiát, mivel a gyakoriság fogalmát nem használja az analógiás értelmezés bemutatásában. Ezzel szemben a modern analógiás megközelítés támaszkodik a gyakoriságra, aminek fényében a példák

magyarázhatóak azzal, hogy a nem zöngés változatok többségben voltak, így hatásuk is erősebb volt.

Hasonlóan King (1969: 133) nem tudja analógiásan magyarázni a **rendhagyó alakok meglétét**. Ha azonban feltételezzük, hogy a kiugró gyakorisággal rendelkező alakokat memorizáljuk, akkor ezek az esetek értelmezhetőek lesznek. Másrészt King érve itt visszafelé is fordítható. Mi a magyarázat arra, hogy a szabályok nem bántak el a rendhagyó alakokkal? Miért döntöttek úgy a beszélők, hogy ezeket kivételes alakokként akarják megjelölni? King azonban ezeket a kérdéseket válasz nélkül hagyja.

King érvei összefoglalva bizonyos pontokon helyesek a kor analógiás gondolkozásával szemben, de több dologgal nem számolt ellenpéldái felépítésében:

- Az analógiás erőt befolyásolja a **gyakoriság**, mint ahogy ezt már Wheeler (1887) is állította.
- **Nem csak aránypáros analógia van**. Az analógiás változást, használatot több alak is befolyásolhatja, mint ahogy azt már Varro is megállapította az ókorban (2.1. alfejezet).
- Az **analógiás erő** nem csak egyezés, hanem bizonyos mértékű hasonlóság esetén is érvényesül, érvényesülhet. Minél erősebb egy **csoport**, a hasonlóságnak a szerepe annál kisebb a viselkedés befolyásolásában. Azaz, ha a magyarba kerülne egy új szó a *gröpáördrimsz*, ami csak alig hasonlít az összes többi magyar szóra, a többes számát akkor is *-k*-val hoznánk létre analógiás alapon, mivel mintáink ezt támogatnák: *gröpáördrimszek*.
- King **nem számol** se a **használattal**, se a nyelvi **variabilitással**, se a **beszélők egyéni használatának** (rendszerének) különbségeivel.

Chomsky kijelentéseit követően a **70-es években általános nézetté vált**, hogy az **analógia** a nyelvi leírásban **marginális** vagy inkább haszontalan eszköz, hiába hangsúlyozta Ohala (1974) és Anttila (1977) annak fontosságát. Az analógiás kutatások ideiglenes hátterébe szorulása azzal is magyarázható, hogy ezekben az időkben a morfológiai kutatás, amely az analógiát legaktívabban alkalmazta, kevésbé töltött be

központi szerepet abból kifolyólag, hogy a nyelvészeti kutatások kiindulási pontja az angol nyelv volt, amely közismert morfológiájának szegényességéről.

Chomsky felületes kijelentéseinek érvényessége a 80-as évektől kezdve kérdőjeleződött meg az azoknak ellentmondó pszicholingvisztikai és kognitív nyelvészeti kutatások hatására (Chandler 2002), amelyek mellé gyorsan felzárkózott az elméleti gondolkozás is. Az analógiás hatás és a szabályalapú megközelítés közti választás vitájában Lukács (2002) a vizsgálataimmal megközelítőleg azonos területen végzett pszicholingvisztikai kutatásai is megerősítették, hogy az analógia egyes esetekben mindenképpen jobban modellálja a valós nyelvi működést, mint azt a szabályalapú nyelvtanok teszik. Rendhagyó főneveken, köztük a vizsgálatom tárgyát képező hangkivető főneveken is végzett vizsgálatainak eredményeit eképp foglalja össze:

„Az eredmények alapján egyértelműen kijelenthetjük, hogy az ilyenfajta nyelvi tudást nem kategorikus szabály testesíti meg. A probabilisztikus eredményeloszlások arra a következtetésre vezetnek, hogy a magyar kivételes főnévi tőosztályok szintetikusan ragozott alakjaira leíró szinten kivétel nélkül érvényes általánosítások sémák formájában működnek a beszélők fejében."

Chomsky megbélyegző nyilatkozatainak hatására az analógiás leírások nem tűntek el, de sokáig még igyekeztek **más névvel illetni az analógiás jelenségeket**, hogy az ilyen típusú elképzelések csupán a név miatt ne váltsanak ki indokolatlan indulatokat: *distinctness condition* (különbözőségi feltétel), *leveling conditions* (kiegyenlítődési feltételek), *paradigm coherence* (paradigmatikus összetartó erő) és Kiparskynál (1992) az *optimization* (optimalizálás). A szabályokkal nem megragadható változásokat pedig szabályalapú megközelítésekben gyakran *tendenciáknak* nevezik (Krott és mtsai 2002: 182).

Az **analógia visszakerülése** az elméleti gondolkozásba egyre általánosabb formában folyik tovább, így Spencer (1988) a generatív leírás számára gondot jelentő ún. zárójelezési paradoxonok megoldására tett javaslatában a klasszikus négyrészes

analógiát javasolja, a 90-es években kialakult **optimalitáselmélet** pedig, amely mára az egyik legelfogadottabb fonológia megközelítéssé vált, visszafogadta a nyelvi leírásba az analógiát, mint legitim magyarázó erőt (Myers 2002, Kiparsky 2000).

2.3. Az analógiás megközelítés kapcsolatai más, hozzá közel álló elméletekkel

A **szabályalapú elméletekkel a konnekcionista és a mintaalapú megközelítések állíthatók szembe** (Chandler 2002), amelyek azonban több szempontból is eltérő nyelvfelfogással dolgoznak. A neurális hálókat alkalmazó konnekcionista megközelítésben maguk a minták a betanítási fázis után nem hozzáférhetőek, míg a mintaalapú elméletekben az ezekre való hivatkozás alapvető fontosságú. Az **analógiás nyelvtanok a mintaalapú nyelvi elméletek családjába** tartoznak, mivel azt állítják, hogy az új és folyamatos tapasztalati elemek értelmezése során a korábbi tapasztalatok alapján létrehozott általánosított reprezentációkhoz való hasonlítás mellett ezeket az elemeket közvetlenül egy vagy több tapasztalati emlékkel vetjük össze. Ezeket a tapasztalati emlékeket a munkamemóriánkba emeljük az alapján, hogy a bemenet valamilyen szempontból hasonlít a korábbi emlékek mentális reprezentációihoz (Chandler 2002: 65). A minták alapján való összevetés fontosságát már Wittgenstein (1992: 62) is hangsúlyozta:

„különböző fajtájú játékok példáit írom le; hogy megmutatom, hogyan lehet ezek analógiájára a legkülönbözőbb módokon más játékokat konstruálni; hogy megmondom, hogy ezt és ezt aligha nevezném játéknak; és még hasonlókban."

A mintaalapú analógiás megközelítések a jelenlegi uralkodó irányzatok közül leginkább az **optimalitáselmélethez** kapcsolódnak abban, hogy **nem szabályokkal** magyarázzák a megértést és a produkciót, illetve, ahogy az optimalitáselmélet egyes változatai is, **csak felszíni alakokra hagyatkoznak**. Az **optimalitáselmélet** bizonyos

képviselőit az is közelíti az analógiás keretrendszerben való gondolkodókhoz, hogy az egyedi, **nyelven belüli variabilitást, ingadozást** is magyarázni, vagy legalábbis észrevenni kívánják. Nagy és Reynolds (1997) az ingadozást megszorítások többféle rendezési lehetőségével próbálják értelmezni[1]. Coetzee (2004) ezzel szemben az ingadozást a kiválasztott alakok rangsorolásával modellezi. Az analógiás megközelítéshez hasonlít leginkább Anttila (2002) elemzése, aki az ingadozást azzal magyarázza, hogy bizonyos esetekben az alacsonyabbra rangsorolt megszorítások is szerepet kapnak, mint ahogy a kevésbé hasonló minták is kifejthetnek hatást egy elem viselkedésére, de gyengébb mértékben, mint a közelebbiek. Az analógiás szemlélethez hasonlító elképzeléseinek gyengéje, hogy a nyelvész választja ki, hogy melyik megszorítás játszik szerepet az adott esetben és melyik nem, míg ez az analógiás megközelítésben általánosabb elvekből következik[2]. Elméletének további hiányossága az is, hogy az ingadozás finomabb arányairól nem tud számot adni. Ez utóbbi problémát kezeli Boersma és Hayes (2001) megközelítése, akik a megszorítások egymással átfedő hatásából vezetik le az ingadozást. Az átfedések mértékének megállapításához ők már korpuszadatokat is alkalmaznak. Megközelítésük közelít a modern analógiás modellekhez annyiban, hogy az arányokat valódi adatok alapján kell számítani, viszont az analógiás elméletekkel ellentétben kevesebbet mondanak az ingadozás valódi okairól, az alakok hasonlósága alapján szerveződő viszonyrendszerek hatásairól.

Fontos azonban azt is kiemelni, hogy az analógiás nyelvtanoktól különbözik az optimalitáselmélet abban, hogy **általános elvekkel, megszorításokkal** dolgozik, amelyek **megfelelő rangsorolása** esetén kapjuk meg a kívánt eredményt. A figyelembe

---

[1] Elképzelésük szerint ingadozás esetén a beszélők többféle rendezést is alkalmazhatnak, amelyek eltérő jelöltet részesítenek előnyben. Ha valamelyik alak gyakoribb, akkor az annak tudható be, hogy több rendezés is a leggyakoribb alakot választaná, ha a különféle rendezések előfordulásának egyenlő valószínűségét adunk.

[2] Más síkon ez a probléma a legtöbb analógiás modellezésnél is megjelenik, hisz ott a változókat választja ki a kutató önkényesen.

vett megszorítások a vizsgált esetek függvényében jelentősen eltérőek is lehetnek[1]. Ezzel szemben az analógiás nyelvtanok alapvetőnek tartják, hogy a változásokat, a megértést, a produkciót az egyedi alakok, szerkezetek erőviszonyai szabják meg, még akkor is, ha ezek nagyobb csoportokba rendezhetők valamilyen hasonlósági kritériumok mentén, vagy prototípusokhoz való hasonlóság alapján.

Az **optimalitáselméleti diskurzusba az analógiás elképzeléseket** legkövetkezetesebben Kraska-Szlenk (2007) próbálja bevonni, de időnként más, az analógiás elemzést alkalmazó szerzők is hivatkoznak optimalitáselméleti megszorításokra (Rytting 2002: 134), mint egyes esetekben magyarázó, az analógiával együttműködő szervező erőkre. Az analógia ezzel párhuzamosan, mint magyarázó erő bekerült az elsősorban optimalitáselméleti apparátusra támaszkodó munkákba is (pl. Kiparsky 2000, 2005). Myers (2002) pedig az optimalitáselméletnek egy olyan változatát mutatja be, amely az analógiás megközelítés elveit alkalmazza az optimalitáselméleti eszközök segítségével.

2.4. Paradigmák az analógiás elméletekben

A paradigmaalapú alaktani elemzések szorosan köthetők az analógiás megközelítésekhez, hisz egyes szavak **hasonlósági alapon sorolhatók be paradigmákba** (Albright 2009), illetve a paradigmák egységességének (paradigm uniformity) egyik fenntartó ereje is az analógia (Eddington 2006). Gyökerei a szó és paradigma (word and paradigm) modellnek is a latin nyelvleírásban találhatók, hisz már Priscianus (Kr.u. V–VI. század) is alkalmaz paradigmákat elemzéseiben (Blevins 2001, Aronoff 1994: 32). Mivel disszertációmban az egyes paradigmákba való tartozás és más paradigmákba való sorolódás, átsorolódás fontos szerepet tölt be, ezért az ezzel kapcsolatos legfontosabb nézeteket szükséges áttekintenem, különös tekintettel arra,

[1] Az optimalitáselmélet feltételezi, hogy minden esetben az összes megszorítás jelen van, csak a kevésbé fontosak, amelyeknek hatása az adott esetben nem érvényesül, kihagyhatók az elemzésből. A megszorítások univerzálisak, rangsorolásuk azonban nyelvenként egyedi.

hogy a hangkivető szavak legrészletesebb analógiás elemzése (Rebrus és Törkenczy 2008) elsősorban a paradigmák felépítésének módját vizsgálja. A paradigma definícióját így határozzák meg:

„Ugyanazon lexémához vagy ugyanazon morfoszintaktikai értékekhez tartozó alakok csoportja."

A paradigmák gazdag morfológiával rendelkező nyelvek esetében meglehetősen nagyok is lehetnek, ami igaz a magyar főnévi paradigmákra is. Finkel és Stump (2009) szerint a paradigmák hatékony memorizálása érdekében kiemelt **alapalakokat** tárolunk, és ezekből hozzuk létre az egyes alparadigmák elemeit hasonlósági alapon, így a *bokrom* alak olyan alakból vezethető le, amely a *bokr-* szekvenciát tartalmazza. Esetünkben ez lehet a *bokron*, amely alakilag a legjobban hasonlít hozzá, de a *bokra* is, amely jelentésében a legközelebbi[1] a potenciális alapalakok közül. Igaz, más távolabbi, de a paradigmán belül nagy gyakoriságú alapalakok (*bokrok, bokrot*) hatásával is számolnunk kell. Finkel és Stump (2009) kétféle paradigmatikus szerveződést különböztet meg. A statikus szerveződés esetén minden paradigmát ugyanazok az alapalakok jellemeznek[2], míg a dinamikus szerveződésben az alapalakok nem feltétlenül azonosak minden tőosztályban. Ez utóbbi elgondolást alkalmazza Rebrus és Törkenczy (2008) is, amelyet adataim viselkedése is támogat (lásd 5. fejezet).

A paradigmák szerveződésében Finkel és Stump (2009) bevezetik a **paradigmatikus átlátszóság** (paradigmatic transparanecy) fogalmát is, amelyet az alábbi kritériumok mentén határoznak meg:

- Minél kevesebb alapalak szükséges a paradigma jellemzéséhez, annál átlátszóbb.

[1] A fordított lehetőség kevésbé valószínű, mivel az E.3 birtokos és a szuperesszívusz is gyakoribb az E.1 birtokosnál.

[2] Hagyományos szótárak alkalmazzák ezt az eljárást, amikor következetesen bizonyos alakokat adnak meg az egyes szótári tételeknél.

- Minél többféle módon lehet egy paradigmát alapalakokra hivatkozva meghatározni, annál átlátszóbb. Így ha egy alakot két másik alak bármelyikéből lehet meghatározni, akkor az a paradigma átlátszóbb, mint ha csak egyből lehetne.

Minél átlátszóbb egy paradigma, annál könnyebb a hozzá tartozó alakokat feldolgozni és tárolni, így az **analógiás kiegyenlítődés** esetében is az **átlátszóságra való törekvés jut érvényre** (Finkel és Stump 2009: 18). Finkel és Stump (2009) az egyes paradigmák átlátszóságának meghatározásához mérőszámokat is megadnak, de ebben a példánygyakoriságot nem veszik figyelembe, azaz nem számolnak azzal, hogy egy cellában lévő alakok mennyire gyakoriak. Mérőszámaik jól tükrözik az egyes paradigmák átlátszósága közti különbségeket, amelyek elvárásainkkal és a nyelvi intuícióval egybecsengenek, de arról nincs pontos képünk, hogy ezek pszichológiailag mennyire reálisan ragadják meg az eltéréseket. Finkel és Stump (2009: 53) fontos megállapítása továbbá, hogy azok a hatóerők, amelyek a paradigmák belső átlátszóságát alakítják, az ellenkezői lehetnek azoknak, amelyek a **paradigmák közti átlátszóságért felelősek**. Lényegében ezzel azt összegzik, hogy az analógiás folyamatok, amelyek egy paradigma egyszerűsödéséért felelősek, más szinten bonyolultabbá tehetik a rendszert (Kiparsky 2005, Wedel 2009).

Ackerman és mtsai (2009) Finkel és Stumphoz (2009) hasonló megközelítésben tárgyalják a paradigmák szerveződését. Úgy vélik azonban, hogy az alparadigmákra jobb úgy tekinteni, mint szóalakok sajátos és megjósolhatatlan csoportjára, semmint olyan alakok halmazára, amelyeket egy alapalakból vezetünk le (Ackermann és mtsai 2009: 73). Ackerman és mtsai (2009) a paradigmák szerveződésének vizsgálatában már hivatkoznak a **gyakoriságra** is**.** Az egyes paradigmacellák vagy az abba tartozó egyedi alakok gyakorisága nagy jelentőséggel bír a paradigmába való besorolhatóság szempontjából, és egyes esetekben akár a paradigmába való tartozás megváltozásához vagy egy paradigma átszerveződéséhez vezethetnek. Gerken és mtsai (2009: 115) hangsúlyozzák, hogy paradigmák szoros hasonlósági kapcsolatok mentén is szerveződhetnek, még ha nincs is az adott paradigmának kizárólagosan felismerhető

tulajdonsága. Ezt tapasztalhatjuk a vizsgálatom tárgyát képező hangkivető főnevek esetében is, amelyeknek több közös tulajdonságuk van, de egy egységes séma vagy leírás mentén nem jellemezhetőek egyértelműen.

Rebrus és Törkenczy (2008) vizsgálatának középpontjában a **magyar szavak hasonlósági és funkcionális alapon paradigmákba és alparadigmákba** való szerveződése áll. Munkájuk az első olyan átfogó mű, amely modern analógiás alapon vizsgálja a magyar morfológiai rendszert, és kiemelten nagy fontosságot tulajdonít kutatásom alaptémájának: annak, hogy a hasonlóság és a gyakoriság milyen hatással lehet a nyelvi rendszer szerveződésére[1]. Radikálisan szakítanak a szegmentálásra törekvő szemlélettel, valamint azzal az elképzeléssel is, miszerint a morfológiai jelenségek magyarázatához mögöttes alakokat kell alkalmaznunk (hasonlóan Fűköh és Rung 2005 csonkolásos nyelvi jelenségekkel kapcsolatban). Ennek megfelelően az egyes szóalakok tőparadigmákba és toldalékparadigmákba sorolódnak, amelyek azonban nem osztják fel morfémahatárok mentén a szót, és akár átfedésekben is lehetnek. Elemzésükben a paradigmatikus uniformitás fontossága mellett a **kontraszthatások jelentőségét** is hangsúlyozzák. A képzett és a ragozott/jelezett szavakat nem választják szét markánsan egy paradigmán belül, amellyel elméleti szinten én is egyetértek, de vizsgálatomat elsősorban mégis a ragos/jeles alakokra korlátozom, mivel elképzelhetőnek tartom, hogy egymás viselkedésére az azonos tőhöz tartozó, de képzésükben is eltérő szóalakok hatása gyengébb, mint azoké, amelyek képzésükben azonosak.

Leírásuk **kulcsfogalma a hasonlóság[2], amit kezdő- és végszekvenciák megfeleltetése által alparadigmák** meghatározására használnak:

„Az alparadigma a funkcionálisan definiált tő- és toldalékparadigmák azon elemeit tartalmazza, amelyek formai szempontból ugyanolyan releváns hasonlóságokat feltételeznek." (Rebrus és Törkenczy 2008: 717)

---

[1] Kutatásomhoz való szoros kapcsolódásuk miatt elképzeléseiket némileg részletesebben mutatom be.

[2] Vizsgálatuk szorosan kapcsolódik a gyakoriság, a hasonlóság és az ingadozás témaköreihez is, de a könnyebb áttekinthetőség végett gondolatmenetük tárgyalását nem osztom szét ezen alfejezetek közt.

„Az alparadigmák tehát azokat a hasonlósági osztályokat adják meg, amelyekbe tartozó elemek formai viselkedése valamilyen szempontból hasonló, pontosabban hasonló kezdő- és végszekvenciákat tartalmaznak." (Rebrus és Törkenczy 2008: 720)

Hasonlóságfogalmuk némileg eltér a disszertációmban alkalmazottól, hisz ők két alakot csak akkor vesznek hasonlónak, ha **megszakítatlan és azonos szekvenciáik vannak akár jobb, akár bal oldalt**. A szekvenciák kizárólagos identitásának ilyetén figyelembe vétele azonban gátolhatja az absztraktabb, funkcionálisabb analógiák felismerését (Blevins és Blevins 2009b: 1). Rebrus és Törkenczy (2008) vizsgálatainak alapegysége így a fonéma, bár egyes esetekben utalnak rá, hogy azok jegyei is szerepet kaphatnának a hasonlóság felismerésében:

„A hasonlóságon alapuló morfofonológiai leírás azon alapul, hogy két alak miben és hogyan hasonló, azaz mely fonémaszekvenciákban – esetleg ezek általánosításaiban: a szegmentumokat alkotó fonemikus jegyekben és ezek csoportjaiban – azonosak, és melyekben különbözőek." (Rebrus és Törkenczy 2008: 697)

Habár elsősorban a **megszakítatlan, identikus szekvenciák** vizsgálatára korlátozzák magukat, elméletben a megszakított szekvenciák vizsgálatának a lehetőségét sem zárják ki:

„Természetesen ez nem zárja ki a magyarban sem, hogy a hasonlóság nem szegmentumok teljes szekvenciái (sztringek) között, hanem más formailag releváns módon, például megszakított szekvenciák közt álljon fenn." (Rebrus és Törkenczy 2008: 698).

A hasonlóságon belül elkülönítik a releváns hasonlóság fogalmát, amely rokonságot mutat Albright (2009) későbbiekben bemutatandó elképzeléseivel is:

„A formai mintázatok sokszor önkényes kapcsolatát a releváns hasonlóság segítségével ábrázoljuk: azon alakok közti hasonlóságok a relevánsak, amelyeknek segítségével a lehető legtöbb formai összefüggést - azaz morfofonológiai általánosítást - le tudjuk írni" (Rebrus és Törkenczy 2008: 702)

Rebrus és Törkenczy (2008: 709) szerint is **kiemelt szerepe van a gyakoriságnak** a paradigmák szerveződésében, miszerint:

„A gyakoriság hatása az egyes alakok előhívásában van (és esetleg új hasonlóságok kialakulásában és régiek eltűnésében is – ez utóbbival a dolgozat szinkrón nyelvészeti jellege miatt nem foglalkozom): a gyakrabban használt egymáshoz hasonló alakok közötti kapcsolat erősödik, ami segíti az előhívást."

A hasonlóság és a gyakoriság kiemelt szerepéből következik, hogy Rebrus és Törkenczy (2008: 709) a **csoporthatásnak** is nagy fontosságot tulajdonítanak:

„A típusgyakoriságnak szerepe van a különböző alakok kategorizációjában: az alakok előhívását segítik azok a további alakok, amelyek hozzá funkcióban és formában hasonlóak, és minél több egymástól különböző hasonló alak van, annál erősebb lesz a hasonlók közötti kapcsolat."

Kifejtik, hogy egy alak kiugró gyakorisága (pl. *jön*, *megy*[1]) elegendő ahhoz, hogy más szavak megerősítő hatása nélkül is önállóan tudjon viselkedni. Minél kisebb gyakoriságú azonban egy elem, annál jobban kell más hasonlóan viselkedő elemekre hasonlítania. **Nem-kategorikus viselkedésnek** nevezik, amikor egy alparadigma nem fed le egy teljes tő- vagy toldalékparadigmát, amiből következik a vizsgálatom tárgyát

[1] A hangkivető főnevek esetében ilyen a *dolog* szó, amely nemcsak nagyon gyakori, mint a *tartalom* és a *figyelem,* hanem alaki felépítését tekintve viszonylag magányos is a hangkivető főnevek közt. Természetesen vannak rá is hasonlító szavak, de azok távolibbak és ritkábbak: *torok, horog* vagy igék: *morog, forog* stb.

képező ingadozás (Rebrus és Törkenczy 2008: 721), amely az analógiás kiterjesztéshez-kiegyenlítéshez vezethet:

„A **nem kategorikus viselkedés gyakori következménye a hezitáció**, illetve további részmintázatok felbukkanása: abban az esetben, ha az alparadigmát határoló lépcsős vonal egy „lépcsőjéhez" közeli alakok használati és típusgyakorisága is alacsony, megnő a bizonytalanság".[1]

2.5. Hasonlóság

A szavak hasonlóságát befolyásoló tényezők azonosítása, és egy olyan módszer kidolgozása, amely ezek alapján a szavak hasonlóságát[2] egyértelműen meghatározza, az egyik legsürgetőbb feladata az analógiás megközelítéseknek, mivel a **leghasonlóbb/ legközelebbi minták/prototípusok kiválasztása** tudományos szempontból is elfogadható módon csak ezek ismeretében lehetséges. Ezzel szemben sok esetben már az elemzés kiindulási pontját képezik a kiválasztott analógiás párok/minták, amelyeket a nyelvész a hasonlóságról alkotott kevésbé megbízható intuíciói alapján jelöl ki. Albright (2009: 189) ezen gyengeségük miatt bírálja a klasszikus, aránypáros analógiára hagyatkozó megközelítéseket, mert nem határozzák meg a lehetséges analógiás minták relatív távolságát a célszóhoz, ezáltal nem tudhatjuk, hogy a legmegfelelőbb szó került-e végül kiválasztásra. Így a Skousen (1989) által indított analógiás modellezési „mozgalomnak" sokat köszönhetünk a kiválasztás mikéntjének tisztázásában is, hisz ha az gépi úton történik, akkor eljárásainkat kénytelenek vagyunk algoritmizálni és formalizálni.

A **hasonlóság mértékének megállapítására kétfajta modell** létezik. Az egyik megközelítésben két szót akkor vesznek hasonlónak, ha bal vagy jobb oldalról nézve

[1] A kiemelés részben tőlem származik.

[2] Vitevich és mtsai (1997) megmutatták, hogy álszavak jólformáltságának meghatározásában is szerepet játszott a hasonlóság (ismerősség) a gyakorisági hatásokon túl.

megszakítatlan fonémaszekvenciáik azonosak. Ezt a megközelítést alkalmazza Rebrus és Törkenczy (2008) is, igaz, elsősorban gyakorlati és nem elméleti megfontolásokból kiindulva. E megközelítés előnye, hogy a hasonlított szavak jelenlegi nyelvi tudásunk alapján valóban hasonlóként viselkedhetnek analógiás folyamatokban, viszont túlzott szigorúsága miatt számos esetben elmulaszthatjuk az általánosabb vagy részleges hasonlósági viszonyok felismerését, így nem a minden szempontból legideálisabb mintákat fogjuk megtalálni. Ennek megfelelően csak a fonémaszekvenciákra hagyatkozva a *sátor* számára jobb minta lenne a *bokor* vagy akár a *sárkány*, holott más, engedékenyebb szempontok alapján közelebb áll a *fátyolhoz*, mint ahogy az a két szó közel azonos hangkivetési mértékében is tükröződik. A megszakítatlan fonémaszekvenciákra hagyatkozó hasonlítás univerzálisan sem alkalmazható, hisz például sémi nyelvek templatikus morfológiája nem ragadható meg vele.

Az analógiás modellezésekben alkalmazott hasonlítási módok szakítanak azzal az önként vállalt megkötéssel, hogy a hasonlóságot minden esetben a szélektől kell számolni megszakítások nélkül, illetve az összehasonlítandó elemek csak azonosak vagy teljesen eltérőek lehetnek. E megközelítések a **hasonlítást három lépésben hajtják végre**. A szavakat részekre, jegyekre bontják valamilyen módon, és ezeket a részeket, jegyeket hasonlítják össze (Stroppa és Yvon 2005). Végül a részek, jegyek hasonlósági számainak összegzése adja ki két szó hasonlóságának a mértékét. Az összehasonlításra kiválasztott részek többnyire fonémák, de lehetnek tengelyeken elhelyezkedő jegyek, vagy akár szótagok is. Az ilyenfajta hasonlítás számításigényes volta miatt elsősorban a számítógépes modellezésben szokásos, de kidolgozott nyelvelméleti háttér nélkül ezek a hasonlósági mértékek sokszor megmaradnak a találgatás szintjén. Habár a modellekben ez a három fázis (részekre bontás, elemösszehasonlítás, összegzés) szigorúan szétválik, a valóságban ez minden bizonnyal párhuzamosan fut le.

Mind a kétfajta megközelítésre igaz azonban, hogy a hasonlóság megállapításában alapvető fontosságúnak tartják a **hasonló/azonos elemek sorrendjének linearitását**. Így a *mozi–izom* pár hasonlósága alacsonyabb, mint az *izom–eper* páré, amely tagjainak szerkezeti felépítése és viselkedése is hasonlóbb, habár elemeik (fonémáik) különbözőek. Ugyanakkor valószínűsíthetjük, hogy a *mozi–izom* pár

szavai még mindig közelebb vannak egymáshoz, mint *mozi–barack* páros tagjai, ahol a felépítés és az építőelemek (fonémák) is eltérőek.

A továbbiakban részletesebben a másodikként bemutatott, több lépéses megközelítéssel foglalkozok, mert saját kutatásomban én is ezt alkalmazom. Elsőként érdemes a **fonémák[1] összehasonlításának** módjával foglalkozni, mivel a kis számú kísérlet ellenére ebben ért el több eredményt a nyelvtudomány. Elképzelhető, hogy a szavak fonetikai alapú összehasonlítása lenne a probléma legígéretesebb megközelítése, de mivel az angol eredmények (Ladefoged 1970, Kondrak és Sherif 2006) nem alkalmazhatóak a magyarra a hangok és azok viszonyainak radikális különbségei miatt, kénytelenek vagyunk legalábbis ideiglenesen fonemikus alapú hasonlóságban gondolkozni. A 6. fejezetben látni fogjuk azonban, hogy szószintű hasonlósági mértékem javulását elsősorban a részek kiválasztásának és elrendezésének megbízhatóbb módjától várhatjuk, semmint még tökéletesebb fonéma-összehasonlítási eljárásoktól.

A **fonémák összehasonlításának egyik legalaposabb elméletét Frisch** (1996) dolgozta ki, amelyet én is beépítek analógiás algoritmusaim egyikébe (4.3. alfejezet). Frisch (1996: 1) szerint a diszkrét szimbólumokon alapuló egészleges összehasonlítás nem alkalmas a fonotaktikai szabályszerűségek megragadására, mivel nem tudja a fokozatosságot modellezni, amelynek a fonológiában is nagyobb szerepet kellene kapnia (Frisch 1996: 89). Ehelyett egy olyan megközelítést javasol, amely jobban kezeli az adatokban a pontosan meg nem ragadható határokat és a folyamatos átmeneteket. A hasonlóság számításában elveti a mögöttes alulszabottságot, modellje Tversky (1977) hasonlósági kontrasztokon alapuló modelljének a kiterjesztése. Tversky és Gati (1978) kategorizációs feladatokban azt tapasztalták, hogy a kísérleti résztvevők Észak-Koreát hasonlóbbnak vélték Kínához, mint fordítva, azaz a hasonlóság nem feltétlenül szimmetrikus reláció, így a kevésbé ismert elem jobban hasonlít az ismertre. Frisch

[1] Elméletben a szavakat bármilyen módon részekre bonthatjuk, de a gyakorlatban a fonéma alapú összehasonlítások uralkodnak.

megközelítésüket ebben a vonatkozásban azonban nem követi[1], mert ő a hasonlóságot szimmetrikus viszonynak veszi.

Frisch **természetes osztályokon (Kornai 1993) alapuló számítási rendszere** a hagyományos megkülönböztető jegyek helyett egyértékű jegyekkel jellemzi a fonémákat. A természetes osztályokkal[2] való számolás előnye, hogy csak a kontrasztív jegyek kapnak benne szerepet, a redundánsak nem. Egyes esetekben egy bináris jegynek mind a két értékét felveszi egyértékű jegyként (pl. zöngés, zöngétlen), más esetekben csak annak meglétére alkalmaz jegyet (koronális), így nincsenek olyan nem természetes osztályai a jegyek alapján, mint a nem-koronális, amely különféleképp viselkedő elemeket gyűjtene egy csoportba. Ennek fényében Frisch (1996: 17) egyértékű jegyekkel az angol magánhangzók egy egyszerűsített részrendszerét a 2.1. táblázaton látható módon jellemezné.

	/a/	/i/	/u/
felső		+	+
alsó	+		
elülső		+	
hátsó	+		+
kerek			+
réses	+	+	

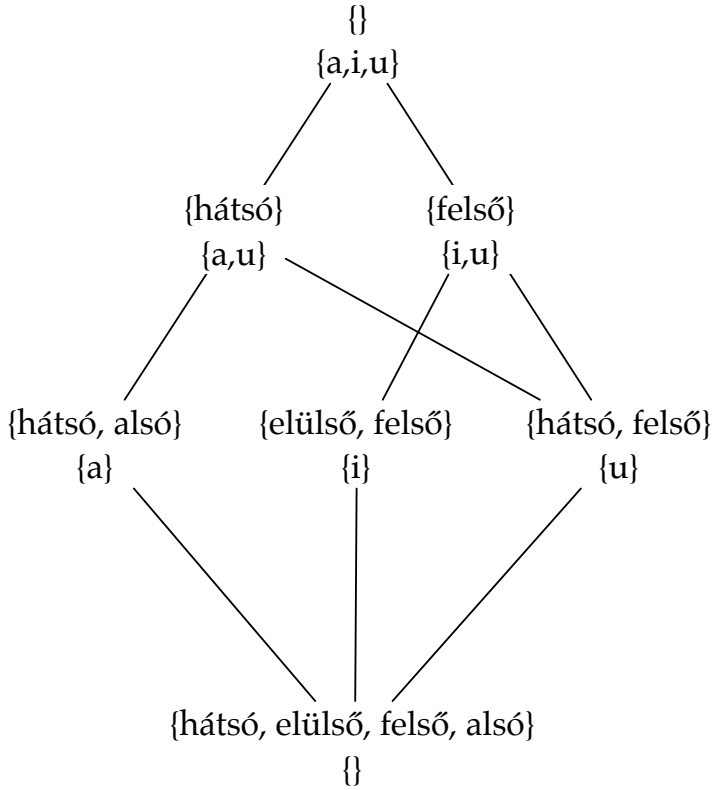
2.1. táblázat: Angol magánhangzók egyszerűsített részrendszere egyértékű jegyekkel

Frisch (1996: 19) az egyértékű jegyek uniói alapján létrehozható **természetes osztályokat hálókba rendezi** (2.1. ábra). Minden természetes osztály jegyek egy részhalmazával jellemezhető. A redundanciák kiküszöbölésének érdekében Frisch azon

[1] A későbbiekben (4.3. alfejezet) én is Frischhez hasonlóan járok el. Elméleti szinten lehetségesnek tartom, hogy a fonémák egymáshoz való hasonlósága asszimetrikus legyen, így a ritkább *i* jobban hasonlíthat az *e*-hez, mint fordítva, de ezzel kapcsolatban megbízható empirikus tudásunk nincs, így inkább tartózkodtam az ilyen jellegű megközelítések alkalmazásától.

[2] Természetes osztályokra az ismert nyelvi leírásokban elsőként Pāṇini hivatkozott (Kiparsky 1991).

természetes osztályokat kihagyja az összehasonlításból, amelyek egy olyan természetes osztályt tartalmaznak, amellyel elemeik azonosak. Így például a szonoráns természetes osztály és az általa tartalmazott zöngés-szonoráns osztály fonémái egyformák, ebből kifolyólag az általánosabb szonoráns osztály szerepeltetése szükségtelen ahhoz, hogy rendszerünk teljes legyen. Formálisan ez azt jelenti, hogy a fonémák természetes osztályaiból és a természetes osztályokat meghatározó jegyhalmazokból álló hálók ún. Galois-kapcsolatot[1] alkotnak, amely megfigyelés Kálmán Lászlónak köszönhető.



2.1. ábra: Angol magánhangzók részrendszerében szereplő természetes osztályok hálója

A Galois-kapcsolatok tulajdonságaiból következik, hogy **a hasonlóság számításához** szükséges természetes osztályok az adott **egyértékű jegyek és e jegyekkel jellemzett fonémák segítségével egyértelműen megadhatóak**. Azok a természetes osztályok, amelyek több mint egy elemet tartalmaznak, egy részben specifikált archiszegmentumnak feleltethetőek meg: *V* az összes magánhangzóra, *U* a

[1] A Galois-kapcsolatban két részben rendezett halmazunk van, és köztük egy olyan reláció, ami tartja a két részben rendezést: ha O(A, B) az egyik halmazban, vagyis az A a B fölé van rendezve, akkor minden A', B'-re a másik halmazból, ha R(A, A') és R(B, B'), akkor O'(A', B').)

felső ajakkerekítéses magánhangzókra stb. Frisch és mtsai (2004) és Frisch (1996) a hasonlóságot a (4)-ben megadott képlet alapján számolja, amelyet Albright (2009) későbbiekben bemutatásra kerülő vizsgálataiban is alkalmaz a GCM modell esetében (3.4. alfejezet). Az összehasonlított fonémák szempontjából irreleváns természetes osztályok, amelyikben egyik sem szerepel, nem számítanak az összehasonlításukban.

$$(4) \qquad \text{f1, f2 fonémák hasonlósága} = \frac{\text{közös természetes osztályok száma}}{\text{közös + nem közös természetes osztályok száma}}$$

A **hasonlóság mértékét** nemcsak az összehasonlított fonémák jegyei, hanem ezeknek a **jegyeknek a teljes fonémaállományon belüli eloszlása is megszabja**, hisz az osztályokat ezek alapján alakítjuk ki. A hasonlóság egy alhálón belül nem azonos az egész rendszerben vett hasonlósággal, azaz a *p* és a *k* jobban hasonlít egymáshoz (magasabb hasonlósági pontszámot kap), ha csak a zárhangokat vesszük figyelembe, mint ha az összes fonémát.

A **magyarban 156 természetes osztály** van Frisch számítási módja alapján. A nyelvenkénti eltéréseket jól szemlélteti, hogy az angol /b/:/f/ pár hasonlósága 0,13, míg a magyarban ugyanez az érték 0,17 egy 0-1 terjedő hasonlósági skálán, mivel az angolban a labiális hangok dimenziója komplexebb, mint a magyarban. A /g/:/k/ hasonlósága az angolban és a magyarban azonos, mivel a veláris dimenzió összetettsége mind a két nyelvben hasonló. Frisch (1996) így plauzibilisen modellezi azt a nyelvi intuíciónkat, hogy ha két hang hipotetikusan azonos is, akkor sem ugyanazokban a viszonyokban vesznek részt a két nyelv hangrendszerében.

Frisch (1996: 105) kiemeli, hogy egyes esetekben **szükség lehet a pozíció függvényében súlyozni azt, hogy két fonéma hasonlóságát mennyire vesszük figyelembe**. Azt tapasztalta angol nyelvű szótévesztések esetében, amelyeket a fonológián túl a szemantika befolyásol leginkább, hogy szóeleji szegmentumok tévesztését markánsabban befolyásolja a hasonlóság, mint a szóbelseji elemekét. Felfedezésének jelentősége nem abban áll, hogy a hasonlításban a szó eleje, vagy a vége a fontosabb, hisz ez nyelvi feladatonként és nyelvenként eltérő lehet, hanem abban,

hogy a hasonlóság megállapításában a szavak egyes részei eltérő szerepet játszanak, játszhatnak.

Frisch (1996: 50–65, 93–147) elképzeléseinek mentálisan reális voltát alátámasztotta azzal is, hogy más megközelítéseknél meggyőzőbben magyarázott olyan **angol beszédtévesztési korpuszokat,** illetve teszteket, amelyekben a hangok hasonlósága szerepet játszhat. Módszere kritikájaként említhető meg, hogy Frisch sem vizsgál nagy mennyiségű természetesen leírt szöveget, korpuszokat, hogy a pozíciók fontosságát tanulmányozza az ottani tévesztésekben, amelyek elméletének megerősítéséhez akár további adalékokat is nyújthatnának. Így az (5) alapján látható magyar *g:k* tévesztések is lényegesen nagyobb arányban (1044-szer több) találhatók szó elején[1] *Google*-találatok alapján, ami Frisch elképzeléseinek érvényességét igazolja.

(5) gépkocsi : képkocsi
411 000 : 952

újság : újsák
1 800 000 : 4

Frisch (1996) bemutatott modelljét pontosságban **közelítő összehasonlításokat** tesz lehetővé **Skousen rendszere** (1989), amelyben egyes elemek összevetése több jegy segítségével akár nagyon részletesen is végrehajtható. Ezek meghatározása azonban a nyelvész előzetes ismeretein alapszik, így még akkor is meglehetősen nehéz őket az adott jelenségnél általánosabban alkalmazni, ha a modellezési feladatban jól teljesítettek.

Frisch (1996) alapján így rendelkezünk egy kifinomult és a nyelvtudomány jelenlegi állása szerint is elfogadható fonémahasonlítási modellel, azonban a **szavak összehasonlítása** a fonémák alapján meglehetősen **kidolgozatlan** még az analógiás megközelítésekben. Skousen (1989) matematikailag is alátámasztott modellje a szó-

[1] A bemutatott példa nem egyedi, de a jelenség áttekintésére nem vállalkozom, mivel ez túlmutatna disszertációm jelenlegi keretein.

összehasonlításra egy lehetséges jó választ ad, viszont esetében a jegyek kiválasztása önkényes és megközelítése sem megfelelő mértékben holisztikus (bővebben 3.2. alfejezet). Frisch (1996) és Skousen (1989) modelljein kívül viszont a legtöbb alkalmazott szóhasonlítási eljárás elnagyolt. Igyekeznek nyelvfüggetlenek lenni, de ez az eredményesség rovására megy. Leggyakrabban a Levenshtein-algoritmust alkalmazzák a szóhasonlításban (Albright 2009: 191), amelyről azonban a 6. fejezetben megmutatom, hogy legalábbis a magyar hangkivető főnevek viselkedésének modellezésére teljesen alkalmatlan.

A szóhasonlítással kapcsolatos kutatásokban **Albright** (2009) megközelítése érdemel még több figyelmet, aki a Levenshtein-algoritmuson vagy a Dice-együtthatón (Dice 1945) alapuló összehasonlításoknál nyelvészeti szempontból motiváltabb és relevánsabb, viszont a **szekvenciaazonosságon alapuló hasonlításnál rugalmasabb eljárási módot javasol.** Albright (2009) elképzelése szerint elméleti szempontból is kétféle hasonlósági számítás létezik az analógiás források kiválasztására. Albright (2009: 190) a strukturális hasonlóság használata mellett érvel, amely esetében az analógiás kapcsolatok nemcsak szavak közt, hanem szavak és sémák közt jönnek létre. A sémák azonban nem kizárólagosak, mint a szabályok, így egy szóra több is alkalmazható. Albright (2009: 196) elmélete szerint annál megbízhatóbb egy általánosított séma, minél több analógiás mintát fed le. Albright elképzeléseit implementálta is. Értékelésére és technikai jellegű bemutatására az alkalmazásokat bemutató 3. fejezetben térek ki.

Albright (2009) nézeteivel szemben áll az az általánosabb megközelítés, amelyet én is alkalmazok, miszerint az analógiás források kiválasztásában **nem strukturált módon kapnak a nyelvi elemek szerepet**. Az egyes alakok eltérő erővel hatnak, amit különböző hasonlósági faktorok, a gyakoriság és a nyelvhasználat más tényezői (pl. beszélők társadalmi szerepe stb.) határoznak meg. Szélsőséges esetben természetesen elképzelhető, hogy minden elem szerepet játszik az analógiák megtalálásában, de nagy valószínűséggel ezek hatása egy bizonyos távolságon túl gyakorlati és nem elméleti szempontból elhanyagolhatónak tekinthető.

2.6. Gyakoriság

A gyakoriság az egyik legfontosabb tényezője annak, hogy az **egyes formákat a mentális lexikonban tároljuk-e el** (Wheeler 1887: 39, Cutler 1984, Stemberger és MacWhinney 1986, Baayen és mtsai 2003)[1]. A nagyobb gyakoriság erősebb mentális reprezentációkhoz vezet (Kálmán 2010b), amelyeket szükség esetén könnyebben érünk el. A gyakoriság hatásáról a nyelvi rendszer szerveződésére már Baudouin de Courtenay (1974 [1904]: 399] is említést tesz, az analógiás kiegyenlítődésben betöltött szerepét pedig Wheeler (1887) hangsúlyozta. Chandler (2002: 73) az angol igék múlt idejével kapcsolatos kutatásaiban mutatott ki gyakorisági hatásokat, ugyanis azt tapasztalta, hogy az analógia forrásául nem mindig a leghasonlóbbat választották a beszélők, hanem döntéseiket a gyakoriság is befolyásolta. Hasonlóan számos pszicholingvisztikai, szociolingvisztikai és kognitív nyelvészeti kutatás támogatja azt az elképzelést, hogy a gyakoriság **nagyobb hatással van a nyelvi rendszer változására és formálódására**, mint korábban gondolták (Labov 1994, Chandler 2002, Thuma 2008).

Bybee (2001, 2010) nyomán elfogadott lett a mintaalapú nyelvi megközelítésekben, hogy a **nagy példánygyakoriság** egyedi nyelvi formák esetén **megtartó erejű** lehet, míg az **alacsony gyakoriságúak hajlamosabbak részt venni az analógiás kiegyenlítődésben**, mert a beszélők fejében kisebb valószínűséggel vannak önálló vagy könnyen hozzáférhető reprezentációik. Mint azt az 5. és a 6. fejezetben is látni fogjuk, Bybee (2001) megállapításai igazak a magyar hangkivető főnevekre is, de a gyakorisági hatásokat a hasonlósági viszonyok felülírhatják, illetve csak jól definiált csoportokon belül játszanak döntő szerepet. A gyakoriságnak az analógiás változásban, produkcióban és megértésben betöltött szerepe mellett Bybee (2001) megállapítja, hogy a gyakori formák fonológiai változásokon, redukción könnyebben mennek keresztül egészleges tárolásukból kifolyólag, így lehet a *nem tudom*-ból *nemtom*. Bybee gondolatmenetéből következő módon Frisch (1996: 2–3) kiemeli a gyakoriság szerepét a fonotaktikában is, mivel a leggyakoribb formák a legjobb formák, azok elfogadhatósága szorosan összefügg gyakoriságukkal. Az 5.2.1. alfejezetben én is hasonló

[1] Természetesen a gyakoriságon kívül erre más tényezők is hatással vannak (Wheeler 1887: 33–35).

következtetésre jutok, mivel a hangkivető főnevek utolsó két mássalhangzójára vonatkozó megszorításoknak (pl. legyenek minél kevésbé hasonlóak) jobban megfelelő szavak valóban sokkal számosabbak.

A gyakoriságnak az analógiás megközelítésekben betöltött kulcsszerepe mellett értéke, hogy más faktorokkal ellentétben az **írott szóalaknak** azon kevés **tulajdonságai közé tartozik, amit mérhetünk** és alkalmas összehasonlításokra. A szóalakok (elsősorban bizonyos karakterláncok) együttes gyakoriságából következtethetünk a szavak (lemmák) gyakoriságára is, ami eredményének pontosságában nagy minta alapján általában biztosak lehetünk, de egyes szóalakok esetében nagy lehet a bizonytalanság, mert az egyes lemmák alakjai összekeveredhetnek (pl. *hurka* = hurok +POSS.E.3 vagy *hurka* 'húsétel'). Természetesen a gyakorisági adatok használata több technikai és elméleti problémát is felvet, amelyekre az 5.1. alfejezetben részletesebben kitérek.

A gyakorisági hatások vizsgálatakor fontos szem előtt tartani még azt a tényt, hogy a **gyakoriság jelentősen változhat akár néhány év alatt** is a technikai vagy társadalmi hatásoknak köszönhetően, így a szocialista rendszerben még kiugróan gyakori *elvtárs* (11089[1]) szó a *komponens*-hez (11612) hasonló gyakoriságúvá vált napjainkra a *Szószablya Gyakorisági Szótár* tanúsága szerint. Kutatásomban jelenleg nem kap szerepet az, hogy a szavak használati módját a gyakoriságon túl befolyásolja-e az is, hogy a **szót a beszélő hány évesen sajátította el**. Más a hozzáférés azokhoz a szavakhoz, amelyeket eltérő életszakaszokban tanultunk meg (Philips 2006: 186–188), így az *evolúció* (11365) és a *robot* (11064) szavak hasonló gyakoriságúak, mégis az utóbbival a legtöbb beszélő már néhány évesen találkozik, míg az *evolúció*t lehet, hogy csak biológia órán hallja először. Kevés adattal rendelkezünk azonban ahhoz, hogy a magyar analógiás kutatásokba ez a tényező beépíthető legyen, ettől függetlenül azonban nem szabad megfeledkezni esetleges hatásairól sem.

[1] A bekezdésben szereplő gyakorisági adatok az egyes szám alanyesetű alak gyakoriságát adják meg.

2.7. Ingadozás és nyelvi változás

A nyelvtudománynak fontos feladata annak feltárása, hogy **milyen tényezők** váltják ki a **nyelvi formák ingadozását**, ugyanis az ingadozás sokkal gyakrabban előfordul, mint ahogy azt a korábbi, általában a nyelvész impresszióira és személyes megfigyeléseire alapozó leírások vallották. Az instabil, ingadozó pontok a **változások korai stádiumában figyelhetők meg**[1], amelyek mindig olyan helyeken indulnak el, ahol a határok nem egyértelműek az egyes kategóriák közt (Chandler 2002: 56). Az ilyen esetekben a bizonytalan identitású elemek jobban ki vannak téve a változásnak, mivel nem tudjuk, hogy hova kell sorolni őket. Különösen, ha ezek az elemek ritkák, és ebből kifolyólag nincsenek emléknyomaink viselkedésükkel kapcsolatban, vagy azok nehezen hozzáférhetők.

Az ingadozásra tekinthetünk úgy is, mint **mintázatok összeütközésére** egy rendszeren belül, amelynek eredményeképpen a legnagyobb stabilitást hozó kimenet veszi át a hatalmat (Wedel 2009: 89). Az analógiás kiegyenlítődést így az a törekvés hívja életre, hogy a nem kívánatos rendhagyó/szuppletív alakok helyett a nyelvi reprezentációk előnyösebb egységessége jöjjön létre, mivel a töveknek és a grammatikai jelölőknek egyaránt egyedieknek és állandóaknak kell lenniük (Vennemann 1972: 184). Az ingadozó rendszerben hasonlóság alapú hibázások és visszacsatolások erősítik az egyik mintázat előretörését a másikkal szemben. A ritkább alakok folyamatosan a gyakoribb változatokhoz hasonlóra formálódnak, a kategóriából kilógó elemek a kategória középpontja felé törekszenek, vagy átsorolódnak egy másik kategóriába, amelynek középpontja közelebb van hozzájuk, erősebben vonzza őket. A rendszer egyik szintjén történő változás befolyásolja a másik szinten meglévő változásokat, a komplexitás mértékét is.

A globális minta összetartása azonban nem célja a rendszernek, hanem csak egy átmeneti stabil állapot elérése egy részrendszeren belül a szüntelen változás menetében (Wedel 2009: 100), mivel a **hasonlóságot a szerveződés különböző szintjein nem lehet**

[1] A saussure-i példa a (1962 [1916]) *honos* > *honor* változás már csak egy változás betetőzésének, befejeződésének tekinthető.

párhuzamosan maximalizálni (Wedel 2009: 99). Megfelelően erős sémával rendelkező csoportoknál akár a csoport rendhagyó, egyedi tulajdonságai is kiterjeszthetőek új, nagyon hasonló elemekre (Bybee és Moder 1983), amelynek során a lokális egyszerűsödés a rendszer globális komplexitását növeli. Ezt az egyensúlykeresést láthatjuk a hangkivető szavak történetében is (bővebben 5.4.2. alfejezet), amelyek a viszonylag stabilnak mondható 16. századi állapot után ismét változásnak indultak, ezúttal a kiegyenlítődés irányába.

A szavak ingadozása jól értelmezhető a **paradigmákra való hivatkozás segítségével**, hisz a hezitáció gyakran nem az egész szó viselkedésére, hanem csak a paradigmatikus cellák egy-egy bizonytalanabb pontjára jellemző (erre látunk példákat az 5.4.2. alfejezetben is). Ha egy beszélő nem találkozott a paradigma meghatározása szempontjából alapvető alakokkal, akkor bizonyos esetekben nem tudhatja, hogy melyik alakot kell választania, azaz a **paradigmacella-kitöltési dilemmával** (paradigm cell filling problem) találja szemben magát (Ackermann és mtsai 2009: 62). Ha beszélőnk a *pöcök* szónak ismeri a *pöcök* és *pöcökből* alakjait, akkor még mindig gondolhatja, hogy a gyakori *könyök* mintáját kell alkalmaznia. Amennyiben találkozik a *pöcköt* alakkal, továbbra is bizonytalan lehet abban, hogy E.3 birtokos alakját vajon a *kölyök* (*kölyke*) vagy a *pocok* (*pocokja*) mintájára hozza létre. Az ilyen helyzeteket értelmezik Ackerman és mtsai (2009) az entrópia fogalmának a segítségével (Shannon 1948), mivel egy paradigmacella kitöltéséhez kötődő bizonytalanság a cella entrópiájával fejezhető ki, amelyek összege adja az egész paradigma entrópiájának értékét. Egy paradigmacella és a többi cella közti viszony a feltételes entrópiával adható meg, amely meghatározza, hogy mennyire lehetünk biztosak a többi cella értékében, ha egy paradigmacella értéke már adott. Ennek megfelelően a *bokorban* alak alapján biztos lehet a beszélő abban, hogy a *bokor* adesszívusza a *bokornál*, ezért a *bokor* paradigmájában az inesszívusz ismerete esetén az adesszívusz feltételes entrópiája 0. A nem egyértelmű helyzeteknek köszönhetően azonban lehetséges, hogy csak bizonyos cellák átsorolódása történik meg, ami azonban még több bizonytalanságot hozhat be a rendszerbe, így könnyen újabb változásokat katalizálhat a rendszer egységességének megőrzése érdekében.

Egy látszólag kizárólagos általánosítás alkalmazásában **bekövetkező hibák, nyelvbotlások** is lehetnek az analógiás kiterjesztés vagy kiegyenlítés előhírnökei (Hock 2003). Így például az angol nyelvben a határozott névelő *a* vagy *an* változatának kiválasztása teljesen szabályszerűen történik. Ugyanakkor bizonyos esetekben (gyermeknyelv, beszédbeli tévesztés, nyelvjárási változatok) előfordulhatnak *an* helyett *a*-s szerkezetek, mint az *a apple*, míg az ellenkező irányba pl. *an boy* nincs kilengés. Skousen (2002: 36, Skousen 1989: 54–59) ezt a jelenséget (időnként előfordulhatnak *a V.** szekvenciák is) azzal magyarázza, hogy habár egy változó (az első mássalhangzó) ismeretében is el tudjuk dönteni szabályalapon is, hogy mely névelőt kell alkalmaznunk, a viselkedés más változókkal is együttjár, amelyek analógiás hatása érvényesül a hibázásokkor. Ezt a jelenséget az analógiás alapokon nyugvó, a 3.2. alfejezetben bemutatásra kerülő AM modell is jól jósolja, míg a szabályalapú nyelvtanok nem tudnak megfelelő magyarázatot adni ezekre a jelenségekre (gyermeknyelv, beszédbeli tévesztések, nyelvjárási változatok), sőt arról sem tudnak számot adni, hogy a beszélők hiányos adatok birtokában (pl. hiányzik az első fonéma) is képesek a megfelelő névelőt kiválasztani. Ilyenkor további változók ismeretében hoznak döntéseket (pl. ha a 2. fonémája a szónak /n/ és a harmadik mássalhangzó, vagy nincs harmadik fonéma, akkor biztos *an* a névelő: *inn, introduction, institute* stb.).

Az ingadozással és a nyelvi változás **szociolingvisztikai** tényezőivel kapcsolatban **Milroy** (1992, 1993) **elképzeléseit tartom irányadónak**, aki a beszélők szerepét emeli ki a nyelvi változásban, amelyet nem tekint kizárólagosan a rendszer által meghatározottnak, de véletlenszerűnek sem véli azt:

„Az innováció és a változás konceptuálisan nem azonosak: az innováció a beszélőnek azon tevékenysége, amelynek eredményeképpen a nyelvi rendszer változik. A beszélők, és nem a nyelvek újítanak." (Milroy 1993: 221)

Milroy (1992) szerint a nyelvi változás a **laza, de ugyanakkor szerteágazó kapcsolatrendszerrel és nagyfokú mobilitással** rendelkező beszélők interakcióinak köszönhetően terjed. Az ilyen egyének több beszélőközösségnek is tagjai, így nem egy

helyről érik őket nyelvi hatások, mely nyelvi újdonságokat laza kapcsolatrendszerük segítségével könnyedén továbbíthatnak. A változások gyorsabban mennek végbe az egyes beszélő szókészletén belül (újítók, korai átvevők), mint a teljes közösségen belül.

3. Nyelvi szerveződések és folyamatok analógiás modellezése

3.1. A modellezés célja és korlátai

A következőkben azokat az **algoritmusokat és kísérleteket mutatom be**, amelyek elképzeléseimhez, kutatásomhoz szorosan kötődnek. Ezekből részletesebben a legelterjedtebben használt Analógiás Modellezéssel és TiMBL-lel (Tilburg Memory Based Learner, Tilburgi memóriaalapú tanuló algoritmus) foglalkozok, majd ezt követően kitérek néhány ismertebb vagy ígéretes megközelítés összehasonlítására is. Ezen analógiás vizsgálatok középpontjában a nyelvi változás, a szavak paradigmákba rendeződése, és a kutatásom szempontjából kiemelten fontos ingadozás áll[1]. Ezek a megközelítések csak áttételesen befolyásolták kutatásaimat és algoritmusaimat, amelyek alapjait ezektől függetlenül hoztam létre. Ugyanakkor algoritmusaimat és az itt bemutatásra kerülő megoldásokat is lényegében ugyanazok az elméleti analógiás munkák megállapításai és felismerései ösztönözték, aminek köszönhetően több közös vonást mutatnak.

Az analógiás modellek mintaalapú megközelítésüknek köszönhetően gyorsan fejleszthetők, robusztusak, nagy lefedettséget biztosítanak, több forrásból származó információt is integrálni tudnak, és jól kezelik az alacsony gyakoriságú eseteknek tulajdonítható hatásokat. Az **analógiás modellezés** lényegében a gépi tanulás egyik alágának tekinthető, hisz alkalmazása során nagy mennyiségű minta megtanulása, megjegyzése a rendszert képessé teszi arra, hogy **ismeretlen elemeket megfelelően osztályozzon**. A tanulási paradigmában a lusta tanulás (*lazy learning*) irányzatába sorolható, ugyanis az elraktározott információn nem hajt végre semmilyen előfeldolgozást, ezekhez csak akkor fordul, ha valamilyen bemenet kiértékeléséhez/ manipulálásához van rájuk szüksége. Ilyenkor a tanulásban memorizált jegyvektorokat

[1] Az analógiás modellezés a nyelv más területein is használható, így Kálmán (2010b) magyar igei bővítménykeretek szimulációjára alkalmazta.

(jegyek sora a hozzájuk tartozó értékekkel) és az egyes elemekhez tartozó kategóriacímkéket használja fel, hogy az osztályozatlan új elemekhez, amelyeknek csak tulajdonságait (jegyvektorukat) ismerjük, megfelelő címkét/kategóriát rendeljen. A címkék lehetnek binárisak (pl. hangkivető vagy sem), nominálisak (ige, főnév stb.) vagy akár numerikusak (hány százalékban kap kötőhangzót tárgy előtt az osztályozni kívánt szó) is. Hasonlóan a jegyek is többféle értéket felvehetnek, így lehet egy jegy értéke numerikus (hány szótagos a szó), bináris (magánhangzóra végződik-e) vagy akár nominális (mi az utolsó fonémája). A feladatok, amelyekre én is használom az analógiás modellezést, az egyértelműsítés témakörébe tartoznak.

Az analógiával kapcsolatos kutatásokban a modellezés kulcsfontosságú szerepet tölt be, mivel segítségével el lehet dönteni, hogy az **explicit és formalizáltabb analógiás elképzelések valóban megállják-e a helyüket** a valós nyelvi adatok, folyamatok értelmezésben. Már a kezdeti mintaalapú vizsgálatok közt is a Hintzman (1986, 1988) elképzelései alapján megvalósított MINERVA rendszer képes volt prototípushatások szimulációjára, hasonlóan a konnekcionista modellekhez, azonban olyan jelenségeket (pl. éles határokkal nem rendelkező kategóriák) is meg tudott ragadni, amelyek ezeknek már problematikusak voltak (Chandler 2002: 71).

A **szabályalapú elméletek helyességének ellenőrzésére azonban a modellezés nem bevett eszköz**. Az elméletek érvényességnek megvitatása általában elméleti és nem gyakorlati síkon zajlik. Ezzel szemben egy jó modellnek jelentősen meg kell közelítenie a valós adatokat, és számítógépesen szimulálhatónak kell lennie. Azt azonban nem szabad elfelejtenünk, hogy egy-egy modell nem a nyelv összességének a működését kívánja magyarázni, felépítését leírni, hanem azt, hogy a nyelv hogyan működik az egyes egyedi ember fejében.

Ebből következően **modellünk** viszont **nem lehet tökéletes**. Ennek első oka, hogy a korpuszok, amelyek adataiból következtetni próbálunk, az egyes beszélők nyelvi rendszerénél sokkal heterogénebbek, hisz egy korpusz több beszélő eltérő időben rögzített megnyilatkozásait tartalmazza. Másrészt az algoritmusok olyan szavakkal

dolgoznak, amelyeket az egyedi beszélők nem mind ismernek[1], vagy bizonyos szócsoportokban akár szótöbbletük[2] van, így ezeknek a beszélőknek az esetében eredményeink minden bizonnyal némileg eltérőek lesznek. Harmadrészt még viszonylag sokféle nyelvi adatot is figyelembe véve valós helyzetekben más forrásokból is érvényesülhetnek analógiás hatások, mint amelyekkel számolunk (ezzel ellentétes elképzelésként lásd Albright 2009). Technikai, tudás- és időbeli korlátok miatt ezeket az adatokat a maguk teljességében jelenleg nem lehet a modellezésbe beleépíteni, és vélhetőleg sokáig ez nem is fog változni.

3.2. AM (Analogical modeling, Analógiás modellezés)

A nyelvi rendszer és változás legismertebb formális alapokon nyugvó, valós nyelvi adatokat használó analógiás modellezését Skousen (1989, 1992, 2002a, 2002b, 2009) dolgozta ki. **Célja egy általános analógiás keretrendszer elkészítése** volt, amely számos nyelv akár jellegében eltérő nyelvi változásait, ingadozásait, produkcióját is jól tudja szimulálni. Az AM sikeresen szimulált nyelvi jelenségeket a magyarhoz hasonlóan gazdag morfológiával rendelkező finn (Skousen 1989) és török (Rytting 2002) esetében, sőt eredményesen alkalmazták még olyan feladatokra is, mint a mesterségesnyelvtan-tanulás (artificial grammar learning, Chandler 2002: 97) vagy a nyelvi sérültek nyelvi képességeinek modellezése (Chandler 2002: 79-89). Elméleti alapvetéseit illetően az AM az elmúlt 20 évben nem változott jelentősen. Elsősorban technikai hatékonysága javult a modellezésben alkalmazott kvantumszámításnak és a kód optimalizálásának köszönhetően (Skousen 2002a, 2002c, 2009).

---

[1] Ez a részlegesség megnyilvánulhat abban, hogy a beszélő a vizsgált szavaknak csak egy részhalmazát ismeri vagy alkalmazza, de akár abban is, hogy a szavakról hiányos információkkal rendelkezik. Pl. ismeri a *murok* szót, hallotta is akár néhány további alakját, mint *murokkal*, *murokból*, de nem tudja pontosan mit jelent, és nem tudja, hogy a hangkivetők általános paradigmájába tartozik.

[2] A digitális szótárak tipikusan a szaknyelvi és a szleng szókincset fedik rosszabbul.

Mivel az AM **sosem készít összetett reprezentációkat,** hanem már meglévő mintákhoz hasonlít, kezelni tud szokatlan új adatokat is, míg ezekkel más reprezentációkat kezelő elképzelések, mint a konnekcionista és a többnyomvonalas (multiple trace, pl. a MINERVA) modellek vagy implicit módon a GCM rendszer[1] is, nem tudnak tökéletesen megbirkózni (Chandler 2002: 74). Skousen rendszere ezáltal alkalmas az egyedi sajátságok megragadására, bár megközelítése nem holisztikus, hisz a minták kiválasztásában nem feltétlenül veszi figyelembe az összes tulajdonságukat, amely rendelkezésünkre áll (pl. egy szó utolsó szótagjának tulajdonságait jegyekben kódolhatja úgy, hogy a többi szótag természetével nem foglalkozik).

Skousen modelljének előnye, hatékonyságán és rugalmasságán túl, hogy számos jelenséget mentálisan is reálisabban kezel, mint a szabályalapú megközelítések. Egyedi Skousen rendszerében, hogy a memória, vagyis a felejtés szerepét is beépíti a működésbe, amivel a gyakorisági hatásokat teszi szélsőségesebbekké (Chandler 2002: 70), amelyek a felejtés alkalmazásától függetlenül is érvényesíthetők az analógiás források kiválasztása során. Mudrow (2002: 229) alapján így foglalhatók össze a **modell azon alaptulajdonságai**, amelyek befolyásolják, hogy milyen mintát választ ki:

- **közelség:** Minél közelebb van egy alak az előre meghatározott változók mentén egy mintához, annál nagyobb az esély, hogy az a minta az analógiás hasonlításban szerepet kapjon, és analógiás forrásként kiválasszuk[2].
- **csoporthatás:** Ha egy mintához nagyon hasonló (közeli) elemek vele azonos viselkedésűek, akkor jelentősen megnő annak a valószínűsége, hogy az analógiás forrás ezen elemek valamelyike lesz. Azaz hasonló felépítésű

[1] A GCM (Generalized Context Model, Általánosított Kontextusmodell) esetében az új példányok kezelését a már ismert példányok osztályaihoz való hasonlóságuk határozza meg.

[2] Skousen hasonlóságfogalma az Albright (2009) által előnyben részesített strukturális hasonlósággal nem azonos. Ebben egyezik az én elképzeléseimmel is, amelyeket a 4. fejezetben fejtek ki részletesen.

azonosan viselkedő elemek analógiás ereje nagyobb, mint az elszigetelt mintáké[1].

- **heterogenitás**: Egy minta nem lehet analógiás forrás, ha van egy vagy több tőle eltérően viselkedő minta, amely közelebbi ahhoz a példányhoz, amelyhez mintát keresünk[2].

Ez alapján a közelség (más szóval a hasonlóság) fontos, de nem egyedüli feltétel rendszerében az analógiás források kiválasztásához. A csoporthatással kapcsolatos alapvetések alapján az AM modellezni tudja a szabályszerű viselkedést, mint a már bemutatott angol névelőválasztás esetében (2.7. alfejezet), vagy a tendenciajellegű folyamatokért felelős csoporthatásokat (gang effect), mint az első magánhangzójukban /o/-t tartalmazó két szótagos finn igék (Skousen 2002b: 33) befolyása más igék múlt idejű ragozására.

Az AM az egyes **alakokat jegyvektorok formájában hasonlítja össze** (Skousen 2009: 164), hogy meghatározza a várható kimenetet a lehetséges minták alapján. Szabadon parametrizálható változóinak köszönhetően nemcsak morfofonológiai jelenségek, hanem akár szociolingvisztikai szempontok alapján meghatározott választás modellálására is alkalmas, mint a megfelelő alak kijelölése az arabban a beszélők eltérő nemének, egymáshoz való viszonyának és korának stb. függvényében (Skousen 1989: 97–100).

Az AM az összehasonlítás alapját képező vektorok **változóit lokálisan** határozza meg a vizsgálni kívánt nyelvi jelenségek függvényében. A rendszerében **nincsenek univerzális változók**, a nyelvek közti hasonlóságokért az univerzális mentális műveletek a felelősek. A lokalitásnak köszönhetően akár többnyire nem fontos változók is bizonyos esetekben döntőek lehetnek (Skousen 2002b: 35), és szerepet kaphatnak a

[1] Nem kell azonban szerkezetileg teljesen egységesnek lenniük, azaz elég ha a wittgensteini értelemben vett csoporthasonlóság érvényesül rájuk.

[2] A heterogenitás fogalmát a későbbiekben finomítani fogom úgy, hogy pontosan definiálom milyen fajta heterogén elemhalmazok befolyásolhatják egy elem viselkedését és milyenek nem. Ennek kapcsán látni fogjuk, hogy ez az elv ellentmondásba kerülhet a csoporthatással és egy közeli, de elszigetelt mintánál erősebb analógiás hatással bírhat egy némileg távolabbi, de erősebb csoport (Wulf 2009: 119).

modell kiválasztásában. Természetesen ha egy faktor nem kap változót, akkor nincs rá mód, hogy hatását kifejtse. Így például analógiás modellezésünk sikertelen lenne, ha településnevek lokatívuszválasztásának modellezésekor azok hangalakjának változóinkban való kódolásán túl nem lenne egy változónk, amely jelöli, hogy a történelmi Magyarország területén található-e egy település vagy sem. Ennek a figyelembevételére egy szabályalapú leírás esetén is szükségünk lenne. Skousen modelljében nincsenek fokozatok a változók lehetséges értékei közt, azaz egy változó vagy azonos vagy különböző, így nem lehet egyetlen változóval kifejezni, hogy két hang a szonoritási skálán közelebb vagy távolabb van-e. Ha ezt figyelembe akarjuk venni, akkor a fonéma reprezentációját több változóban kell kódolnunk.

Skousen megközelítésének egyik leggyengébb pontja, hogy a **változók kiválasztása mindig önkényes**. Ez az adott nyelvi feladathoz igazodik, és a kutató előre dönti el, hogy mi lehet fontos, illetve mi nem. A probléma kezelésének érdekében Skousen a változók számát az AM-ben már a kezdeti 12-ről 20-ra növelte[1], azért hogy a vektorok több szempontot is kódolhassanak, de ettől függetlenül a megközelítés magában hordja annak a lehetőségét, hogy a kutató kifelejtsen esetleg bizonyos esetekben fontos változókat[2].

Ellentétben több mintaalapú megközelítéssel a **változókat az AM nem súlyozza**, mivel Skousen ezt reménytelen korrekciós vállalkozásnak tartja. Skousen (2002: 40) azzal érvel a súlyozás szükségessége ellen, hogy a bizonyos változókkal jellemzett alakok egységes viselkedése és az ezekből következő csoporthatások az AM számítási

[1] Az elméletben nincs számbeli korlát a változók számán.

[2] Ez a kritika áll a generatív megközelítésekre és elsősorban az optimalitáselméletre is, ahol szintén a nyelvész előzetes ismeretein múlik, hogy milyen tényezőket vesz figyelembe leírásában. Igaz, a sztenderd generatív fonológiában van egy készlet, amiből választani lehet a leírásban, és egy rendszer, ami meghatározza ezek felhasználási lehetőségeit, de ez az idők folyamán változott annak függvényében, hogy az azt meghatározó nyelvészek ismeretei és tudása hogyan alakult. Így lehet vitatkozni azon ebben a keretben is, hogy a labiodentális és a labiális külön jegy-e az Ewe-ben, vagy hogy az affrikáta kontúrszegmentum-e vagy sem. Másrészt fontos látnunk, hogy a modern fonológiai leírásban dominánsan uralkodó optimalitáselmélet esetén az elemzésben felhasznált korlátok kiválasztása a nyelvész ismeretein és képességein nagyban múlik.

módszereiből automatikusan következő módon kiemelik az egyes változókombinációk statisztikai jelentőségét. Úgy gondolja azonban, hogy a különböző jellegű változók (fonológiai, jelentéstani, pragmatikai stb.) már nem kezelhetők együttesen súlyozás nélkül.

Mivel az **analógiás modellezés procedurális**, a döntéseket mindig menet közben hozza meg (Skousen 2002a: 3). Az adatok a döntésekhez nincsenek előre strukturálva, nincsenek általánosítások a rendszerben. Az AM minden példányt változók halmazával, jegyvektorokkal jellemez. A *vödör* szó jellemzése jegyvektorokkal az (1) alatti módon lenne megadható.

(1) vödör CFRMCFRMr H

1 változó = 1. fonéma
2-4 változók = 2. fonéma
5 változó = 3. fonéma
6-8 változók = 4. fonéma
9 változó = 5. fonéma

(C = mássalhangzó, F = elülső, R = kerek, M = középső nyelvállású, r = /r/ fonéma, H = hangkivető)

Egy jegyvektor többször is előfordulhat az adatbázisban ugyanazzal a kimenettel (Skousen 2002a: 12), mint azt a *csöbör* esetén láthatjuk (2).

(2) csöbör CFRMCFRMr H

De rendelhetünk ugyanahhoz a jegyvektorhoz eltérő kimenet is (3).

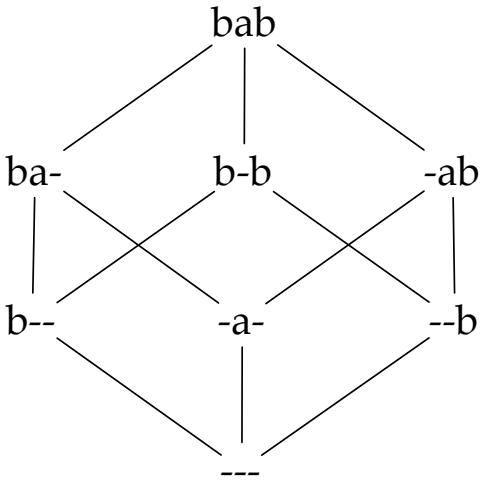
(3) gyönyör CFRMCFRMr N

(N = nem hangkivető)

Ezzel a jegyvektorral azonban csak 5 fonémás *CVCVC* szerkezetű szavakat tudnánk leírni, és ha analógiás mintáink közé a *göndör* szót is fel akarnánk venni, akkor változóinkat már másképp kellene meghatároznunk[1]. Ha összesen ez a két mintánk lenne (2,3), akkor az AM beállításától függően a *vödör*-t 50%-ban hangkivetőnek jósolná, vagy 50%-os valószínűséggel véletlenszerűen az egyik mintát jelölné ki analógiás forrásnak.

Az AM az egyes mintákhoz való analógiás források kiválasztásában a változóknak olyan részhalmazait (ún. szuprakontextusait) használja, amelyek által jellemzett minták azonos viselkedést (pl. hangkivetők vagy sem) mutatnak. Ezeket Skousen **homogén szuprakontextusoknak** nevezi, amelyeket szigorúan megkülönböztet a **heterogén szuprakontextusoktól**, amelyek viselkedésükben nem egységesek, azaz különböző kimeneteik vannak. Ha n változónk van, akkor 2^n szuprakontextust hozhatunk létre ezek alapján, amelyeket a 3.1. ábra szemléltet a *bab* szó kapcsán egy hálóval (Skousen 2002: 15):

[1] Ez a változókészlet alkalmazható lenne több, mint 5 fonémás szavakra is, mint a *szemetesgödör*, de ebben az eseteben némileg másképp kellett volna definiálnunk őket, például az 5. fonémát utolsó fonémának kellene neveznünk.



3.1. ábra: A *bab* szó fonémái felett létrehozható szuprakontextusok hálója, ha három bemeneti jegyünk van, amelyek az egyes fonémáknak felelnek meg.

Ha a magyar tárgyragot vizsgálnánk, akkor az *-a* végű szavak szuprakontextusát homogénnek vehetnénk, mivel a szóvégi *a*-k mindig egységesen megnyúlnak a tárgyrag előtt, lényegileg szabályszerűen viselkednek. Ezzel szemben a tárgyrag vonatkozásában a *-----r*[1] végű vagy a *h-r---* kezdetű szuprakontextusok heterogének (pl. *cukrot:csavart, hurkot:hártyát*). Az egyes minták viselkedésének a jóslásában a minták felett létrehozható szuprakontextusokból az AM csak azokat veszi figyelembe, amelyek fedik a jósolni kívánt elem jegyeiből létrehozható szuprakontextusokat, így egy *-a* végű szó viselkedésére a nem *-a* végű szuprakontextusok semmilyen hatással sincsenek.

Az AM-ben az **ingadozó mintahalmazokat** az olyan **heterogén** szuprakontextusok jellemzik, amelyeknek **nincs nekik ellentmondó szubkontextusuk**[2], amelyben a kategóriák aránya vagy a lefedett minták száma eltérő. Minden más heterogén szuprakontextust kizár a vizsgálatból. Így a *galvánelem* esetében az *el-m* megfelelő heterogén szuprakontextus, mivel a *galvánelem*-en kívül még 183

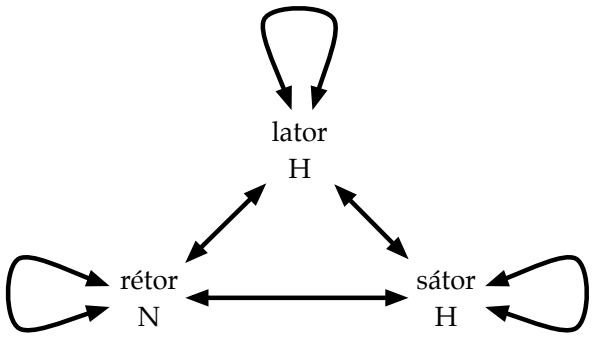
[1] A - jel arra utal, hogy az *-r* előtt van még olyan változónk, amelynek értékét nem vesszük figyelembe.

[2] Egy szuprakontextus szubkontextusán az adott szuprakontextus által lefedett elemeknek egy olyan részhalmazát értjük, amelyet egy, a szuprakontextus meghatározásában nem specifikált jegy valamilyen konkrét értékkel való felruházása után kapnánk, így az *-a* végű főnevek szuprakontextusának szubkontextusa a *-ka* végű főnevek halmaza.

esetet[1] fed le, hasonlóan mint szubkontextusa az *elem*, amely által lefedett alakok ugyanolyan arányban hangkivetők. Ezzel szemben a *-l-m* szuprakontextus már nem felel meg a kritériumoknak, mivel 550 elemű (a *galvánelem*-en kívül) és van ettől eltérő számosságú szubkontextusa (*el-m* 183 elemmel), amelyben a hangkivetők és a nem hangkivetők aránya is más, mert az *el-m* által lefedett esetek 83%-ban hangkivetők, míg az *-l-m* által leírtaknál ez az arány 75%.

A megfelelő **minta kiválasztása kétféleképpen lehetséges** (Skousen 2002a: 13). Az AM képes egyrészt véletlenszerű módon választani a különböző lehetséges viselkedések közül annak arányában, hogy az azok mögött álló minták mekkora analógiás erőt fejtenek ki. Az analógiás erőt minden alak esetében egyenként kell kiszámolni. Az eljárás lényege (Skousen 1989), hogy egy alaknál, amelyhez analógiás forrást keresünk, az AM veszi az azt leíró változókból létrehozható szuprakontextusokat, és a minták ezekkel azonos nem üres homogén, illetve a vizsgálatból nem kizárt heterogén szuprakontextusai alapján kiszámítja a rá ható analógiás nyomást az eltérő viselkedések függvényében. Ezt a szuprakontextusokban lévő mutatók száma határozza meg, amelyek száma 2^n egy szuprakontextuson belül, ha a kontextus által lefedett minták száma n. Egy szuprakontextuson belül minden egyes alak az összes többire mutat. Egy adott viselkedés erejét pedig azon mutatók száma adja, amelyek e viselkedést megtestesítő alakokra irányulnak. Így, ha a *motor* viselkedésére lennénk kíváncsiak a *lator, rétor, sátor, mozsár, mogul, mosoly* mintahalmaz alapján, akkor az analógiás forrás kiválasztásában a *--t--* heterogén szuprakontexus is szerepet kapna, amelynek mutatóit a 3.2. ábrán láthatjuk.

[1] A számok a legnagyobb ingyenesen hozzáférhető szótári adatbázison, a *morphdb.hu*-n alapszanak (Trón és mtsai 2006).



3.2. ábra: A *lator, rétor, sátor* mutatói a *--t--* heterogén szuprakontextusban

Ennek alapján a hangkivető alakokra 6, a nem hangkivetőkre 3 mutató irányulna. Ha ezen kívül már csak a *mo---* homogén szuprakontextust (*mosoly, mozsár, mogul*) vennénk figyelembe, eltekintve az összes többi kontextustól, amelyeket egy valódi elemzésben ugyancsak számításba kellene vennünk, akkor a *motor* hangkivető viselkedésének valószínűségét 33,3%-ra jósolnánk ((2*3)/(3^2+3^2)), azaz 33,3%-os eséllyel választana az AM a *motrot* alakot támogató mintát. A másik lehetőség, hogy a rendszer egyszerűen a legvalószínűbb kimenetet választja. Ebben az esetben a köznyelvi *motort* változatot támogatná. Az AM azonban **nem csak bináris választási helyzeteket tud hatékonyan kezelni**, mint ahogy azt Wulf (2002: 112) vizsgálata is megmutatta, aki szimulálta vele a különösen problematikus német többes szám viselkedését, amelynek 13 lehetséges változata van, és ezekből több is produktívnak mondható.

Habár az AM hatékonynak bizonyult számos nyelvi jelenség analógiás modellezésében, több **kritikai észrevétel** is megfogalmazható vele kapcsolatban. Az AM képes eltérő hosszúságú elemeket is összehasonlítani, de a hosszúságból adódó különbségek hasonlóságot csökkentő hatását nem kezeli megfelelő hatékonysággal. Ahhoz azonban, hogy ezt megfelelő módon be tudjuk építeni a modellbe, pontosabb empirikus tudásunknak kellene lenni arról, hogy a hasonlóságban a hosszúság pontosan milyen szerepet játszik (Chandler 2002: 94). Problémákat vet fel az is, hogy az AM az összes változónak egyforma súlyt ad bármilyen kontextusban[1]. Empirikusan nem megalapozott annak a feltételezése, hogy egy betű- vagy fonémasor minden eleme

[1] Egyes kísérletekben ezt ugyanannak a változónak többszöri felvételével próbálják meg korrigálni.

egyformán járul hozzá az analógiás forrás kiválasztásához. Bybee és Moder (1983) is azt találták kísérletükben, hogy olyan rendhagyó szavak széleihez való hasonlóság (*sC*-kezdet és veláris nazális vég), mint a *swing* vagy a *string*[1] fontosabb szerepet kapott kitalált szavak (pl. *sming, spink*) múlt idejének meghatározásában, mint azoknak a szótagmagjához való fonológiai közelség. A kitalált tesztszavakra a prototipikus szavakra (pl. *swing*) jellemző tulajdonságok közül a csak az *s*C- kezdet megléte esetén (azaz nem /ɪ/ volt ezeknek a magánhangzója és nem veláris nazális végük volt) 17%-ban, csak veláris nazális vég előfordulásakor 34%-ban és csak /ɪ/ szótagmag esetében 7%-ban kaptak olyan múlt idejű alakokat, ahol kizárólag a magánhangzót cserélték le a részvevők /ʌ/-ra. Hasonló megállapításra jutott Chandler is (2002:94–95), aki szimulációi során azt tapasztalta, hogy az AM a kitalált *cug* szó múlt idejéhez intuitív és empirikus alapokon is erősen megkérdőjelezhető módon a *cut*-ot választotta analógiás forrásnak, holott az esetünkben legfontosabb változó, az utolsó mássalhangzó eltérő. A természetes nyelvi tapasztalatokkal összhangban több mesterségesnyelvtan-tanulási kutatás is azt bizonyította, hogy a széleken található bigramok és trigramok fontosabb horgonypozíciók nonszensz betűsorok (pl. *VXTTTV*) esetében is, mint a karakterláncok belsejében találhatók (Chandler 2002).

3.3. TiMBL (Tilburg Memory Based Learner, Tilburgi memóriaalapú tanuló algoritmus)

Az AM-hez hasonlóan széles körben alkalmazott TiMBL (Daelemans és van den Bosch 2005) a **legközelebbi szomszéd megközelítéssel** határozza meg, hogy egy példányhoz mi lenne a legmegfelelőbb analógiás forrás (Eddington 2002: 141–143). A megközelítés nem új, nem nyelvészeti területeken már az 50-es években is felismerték hasznosságát (Fix és Hodges 1951, Cover és Hart 1967). A TiMBL a keresett szóhoz a

[1] A csoportba tartoznak olyan szavak is, amelyek nem /ɪ/ magánhangzót tartalmaznak, mint *strike, drag*. Közös a csoport tagjaiban, hogy múlt idejükben a magánhangzó egységesen /ʌ/, azaz a csoportot megragadó séma nem bemenetüket, hanem kimenetüket szorítja meg.

hozzá leginkább hasonlót választja; amennyiben több ilyen van, akkor a leggyakoribbat. A program több szomszéd kiválasztását is lehetővé teszi, és alkalmas a leggyakrabban használt tesztek végrehajtására (hagyj-ki-egyet, 10-szeres keresztellenőrzés[1]; bővebben 6.3. és 6.4. alfejezetek). A TiMBL rendszerben a bemenő adatok (pl. szavak) jellemzését ugyanolyan jegyvektorokkal kell megadni, mint az AM esetében, tehát ugyanolyan önkényes annak megválasztása, hogy mely jegyeket gondoljuk relevánsnak. A különbség csak annyi, hogy e jegyek értékei (sztringek, számok) között fokozatos összehasonlítások is lehetségesek (pl. szám nagysága, sztringek közötti Levenshtein-távolság).

A TiMBL az Aha és mtsai (1991) által kifejlesztett algoritmus kiterjesztése. A **közvetlen szomszédot a bemenetként kapott alakok alapján keresi**, amelyeknek a felhasználó által meghatározott tulajdonságait is tárolja (pl. melyik szótagja hangsúlyos, mi a neme stb.). Alapbeállításban a TiMBL két elem (x, y) távolságát az alábbi képlet alapján számolja ki (IB1 metrum):

(4)
$$\Delta(X,Y) = \sum_{i=1}^{n} \delta(x_i,y_i)$$

$$\text{ahol: } \delta(x_i,y_i) = \begin{cases} \frac{x_i - y_i}{\max_i - \min_i} & \text{ha numerikus, egyébként} \\ 0 & \text{ha } x_i = y_i \\ 1 & \text{ha } x_i \neq y_i \end{cases}$$

Az eredeti, Aha és mtsai (1991) által kifejlesztett algoritmustól abban térhetünk el, hogy több mintát is használhatunk egy elem viselkedésének a meghatározásához. Ilyenkor a figyelembe vett mintáknak nem a számát, hanem a maximális távolságát kell

[1] A hagyj-ki-egyet teszt esetén egy algoritmus hatékonyságát az alapján mérjük, hogy ha ismertnek vesszük az egész vizsgált adathalmazt, akkor mennyire jól tudja jósolni egyetlen egy kihagyott elemnek a viselkedését. Ezt a vizsgálatot minden elemre elvégezzük, és az egyes elemeken nyújtott teljesítmény alapján kiszámítható a rendszer átlagos hatékonysága. Ehhez logikájában nagyon hasonlóan működik a tízszeres-keresztellenőrzés, amikor az anyag 90%-a (ún. tanító halmaz) alapján jósoljuk a maradék 10% (teszthalmaz) viselkedését. Ilyen ellenőrzést tízszer hajtunk végre úgy, hogy mindig az anyag másik tizede a teszthalmaz. A tesztelés végén az egyes tizedeken elért eredményekből átlagot számítunk.

megadni, azaz a figyelembe vett források száma esetről esetre változhat, de azok maximális távolsága egy adott modellezési helyzetben állandó. A TiMBL változatai közül a legfontosabb az **információnyereséget** (information gain, IG; hasonlóan döntési fáknál is (Quinlan 1993)) számolja ki egy-egy jegy esetében. Ez meghatározza azt, hogy milyen mértékben járul hozzá egy-egy jegy egy jelenség magyarázatához, és ezáltal súlyozza fontosságát. Az IG megadja, hogy mekkora a különbség a bizonytalanságban (entrópia) azon helyzetek közt, amikor ennek az adott jegynek tudjuk az értékét, vagy az ismeretlen a számunkra (Daelemans 2002: 161). Ezt az alábbi képlet alapján határozhatjuk meg, ahol C (C=category, 'kategória') a lehetséges címkék halmaza, V_i (V=value, 'érték') az i jegyhez tartozó értékek összessége és $H(C) = - \sum_{c \in C} P(c) \log_2 P(c)$ az egyes címék entrópiája (a valószínűségeket a tanító halmazból számolt relatív gyakoriság alapján határozhatjuk meg):

(5) $w_i = H(C) - \sum_{v \in Vi} P(v) \times H(C \mid v)$

Az információnyereség ilyetén számítása az informatív, de ritkán előforduló jegyértékeket figyelmen kívül hagyja, míg **azon jegyeknek, amelyek nagyon sok értékkel bírnak, túlzott súlyt ad**. A **nyereségarány** (gain ratio) bevezetésével ez a hatás enyhíthető. Ennek alkalmazásakor az információnyereséget elosztjuk a jegyértékek entrópiájával:

(6) $$w_i = \frac{H(C) - \sum_{v \in Vi} P(v) \times H(C \mid v)}{si(i)}$$

$$si(i) = H(V) = - \sum_{v \in Vi} P(v) \log_2 P(v)$$

Az így kapott nyereségaránnyal már súlyozhatjuk a távolságszámítást is:

(7) $$\Delta(X,Y) = \sum_{i=1}^{n} w_i \delta(x_i, y_i)$$

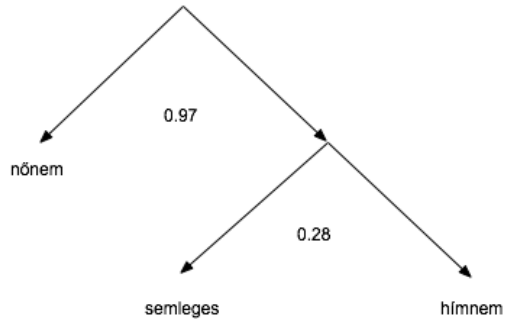
A **jegyek automatikus súlyozása** lehetővé teszi nagy számú, potenciálisan irreleváns jegyek hozzáadását is, mert ezek az eljárások eldobják a haszontalan jegyeket. Ugyanakkor azt sem szabad figyelmen kívül hagyni, hogy nagyon redundáns jegyek együttes súlya indokolatlanul nagy is lehet ezeknek az módszereknek köszönhetően. A nyereségarány-számítás komoly hiányossága, hogy a jegyeket izoláltan vizsgálja, és nem alkalmazza az analógiás megközelítés alaptételének számító holisztikus megközelítést, amelyet a saját későbbiekben bemutatandó komplex tengelymérték nevű algoritmusunk (4.3. alfejezet) is követ. Mivel az információ nyereségarány sem küszöböli ki teljesen azt a hatást, hogy a sok értékkel rendelkező jegyek ne kapjanak túlzott súlyt, a súlyozáshoz χ^2 statisztikát is használhatunk (Daelemans és van den Bosch 2005: 31). Súlyokat a bemutatott eljárásokon túl nyelvészeti tudás alapján is megadhatunk kézileg, amelyek akár jobban is teljesíthetnek, mint a gépileg beállított súlyok (Daelemans és van den Bosch 2005: 38).

A jegyek súlyozásán túl másik jelentősebb kiegészítési lehetőség a TiMBL-hez a **módosított értékkülönbség mérték** (Modified Value Difference Metric, MVDM) számolása, amely meghatározza, hogy egy jegy mely nem feltétlenül numerikus értékei hasonlóbbak, és ehhez igazítja a legközelebbi szomszéd keresését (Eddington 2002: 142)[1]. (8) segítségével azt vizsgáljuk meg ilyenkor, hogy egy jegy két értéke (v_1, v_2) milyen valószínűséggel fordul elő az egyes címékkel, ahol $C_{1...n}$ a címkék sora:

$$\delta(v_1,v_2) = \sum_{i=1}^{n} | P(C_i | v_1) - P(C_i | v_2) | \qquad (8)$$

Az MVDM segítségével a hasonlóbb értékek hierarchikus klaszterekbe rendezhetőek (3.3. ábra).

[1] Az egyes jegyértékek hasonlóságát jelenleg sem az AM, sem a későbbiekben bemutatandó algoritmusaim nem kezelik. Ez a képesség előnyös lehetne, de mint a 6. fejezetben látható lesz, nem befolyásolja jelentősen az eredményeket.



3.3. ábra: Német nemek távolsága 0 és 1 között értékekkel megjelenítve (minél kisebb az ábrán megjelenített szám, annál hasonlóbb egy érték) hierarchikus klaszterezéssel csoportosítva

A TiMBL-ben a minták viselkedését a **szomszédok viselkedése** határozza meg **egyszerű többség alapján**. A távolabbi szomszédok figyelembe vétele egy ideig növelheti a pontosságot, de a teljesítmény romlásához is vezethet egy küszöbön túl. Ennek a hatásnak a kiküszöbölésére lehet súlyozni a szavazatokat, azaz a távolabbi szomszédok szavazata kevesebbet érhet (Dudani 1976), amelyeket lineárisan vagy exponenciálisan csökkentve is súlyozhatunk (Shepard 1987).

Az **AM-et a TiMBL-lel** többen is **összehasonlították** (Eddington 2002, Daelemans 2002, Krott és mtsai 2002). Az összevetések eredményei alapján látható, hogy a két eltérő analógiás megközelítés közel hasonló mértékben, többnyire sikeresen kezel olyan nyelvi jelenségeket, amelyek a szabályalapú elméleteknek gondot okoznak. A különbségek csak ritkán voltak szignifikánsak, így a két algoritmus közti „verseny" még nem dőlt el. A memóriaalapú modellek **közös sajátosságai**, amelyek az AM-ra és a TiMBL-re is igazak, Daelemans (2002: 160) alapján a következőképpen fogalmazhatók meg[1]:

- Nincs a szabályosan viselkedő és a rendhagyó alakok közt szigorú megkülönböztetés.
- A kategóriák közti határok elmosódottak, az elemek kevésbé szigorúan tartoznak az egyik vagy a másik csoportba.

[1] Ezekből több szempont már korábban is elhangzott, de a könnyebb összehasonlítás érdekében ezeket is a felsorolásokban hagytam.

- A hasonlóságon alapuló következtetés és a memóriában való tárolás ötvözete kognitív szempontból egyszerűbb, mint a szabályok felfedezése és azok folyamatos alkalmazása.
- A memóriaalapú rendszerek rendkívül alkalmazkodóak és robusztusak.

A két rendszer **különbségei** a következőképp fogalmazhatóak meg (Daelemans 2002: 163–164):

- A nem szomszédos elemek is hatással lehetnek a nyelvi viselkedésre az AM esetében, de a TiMBL-nél nem.
- Az AM lokálisan határozza meg az egyes jegycsoportok jelentőségét, míg ezek a lokálisan felismert súlyok elvesznek az információnyereséges megközelítés átlagolási folyamataiban, mivel ez a jegyek jelentőségét globálisan határozza meg.
- A TiMBL előfeldolgozást végez a jegyek súlyának kiszámításakor az AM-mel ellentétben.
- Az AM által alkalmazott természetes statisztika lehetővé teszi, hogy az adatoknak csak egy részét használjuk fel (tökéletlen memória szimulálása) az optimális pontosság és robusztusság érdekében, míg a TiMBL számára a kivételek elfelejtése kifejezetten hátrányos lehet. (Daelemans és mtsai 1999).
- Az AM-nél a vizsgálandó esetek száma exponenciálisan nő, míg a TiMBL-nél lineárisan.
- Az AM-nek nincs természetes kiterjesztése numerikus adatokra, míg a TiMBL átfedési metrikája lehetővé teszi, hogy sokféle jegyértéket kezeljen (numerikus érték, halmaz-érték).

Daelemans (1994) a TiMBL-t az információnyereséget mérő eljárással kiegészítve holland igék hangsúlykiosztásának modellezésében csak némileg találta sikeresebbnek (AM 80,5%, TiMBL+IG 81,8%). Daelemans (2002) a német többes szám vizsgálatában a két rendszer eredményeit ismét megközelítőleg hasonlónak találta, ugyanakkor a

TiMBL az AM-nél némileg jobban teljesített holland összetett szavak elő- és utótagja közti kötőelemek (linker) kiválasztásában (TiMBL 93,4%, AM 92,8%, Krott és mtsai 2002). Hasonló eredményekre jutott Eddington (2002) is egy olyan feladatban, ahol az algoritmusoknak spanyol főnevek német kellett meghatározniuk (AM 94,5%, TiMBL +MVDM 3 legközelebbi szomszéd 96,2%). A TiMBL ehhez hasonló, de némileg gyengébb eredményeket hozott azonban jövevényszavak esetében (AM 90,8%, TiMBL +MVDM 3 legközelebbi szomszéd 86,2%), spanyol kicsinyítő képzős alakoknál (AM 93,27%, TiMBL+MVDM 3 legközelebbi szomszéd, 92,69%) és hangsúlykiosztási feladatokban (AM 94,4%, TiMBL+MVDM 1 szomszéd 94,3%). **Daelemans** (2002) hasonló hibái és teljesítménye ellenére a **TiMBL-t egyszerűbb architektúrája miatt ígéretesebbnek tartja**, annak ellenére, hogy a TiMBL információnyereséget is alkalmazó változata azzal a többnyire nem helytálló feltételezéssel működik, hogy a jegyek egymástól függetlenek. A tesztek mégis azt mutatják, hogy ez a fajta homogenizált működés hatékonyan kezeli a nyelvi problémákat (Daelemans 2002: 176).

Van den Bosch (2002) a TiMBL-t a FAMBL2 (Family Based Learning, Családalapú tanulás) algoritmussal hasonlította össze, amely **egyedi alakokból általánosítva készít példánycsaládokat**. A FAMBL2 nem olyan általánosan alkalmazott algoritmus, mint a TiMBL vagy az AM, mégis érdemesnek tartom a bemutatásra, mert ezekhez hasonlóan jó eredményeket hoz, és a **prototípuskezelésben** előremutató, bár nem megfelelően kihasznált elképzeléseket alkalmaz.

A FAMBL2 algoritmus abban különbözik az analógiás modellek többségétől, hogy az **új, osztályozni kíván alakokat** nem más alakokhoz, hanem olyan **reprezentációkhoz hasonlítja**, amelyeket a már korábban megvizsgált alakok alapján készített. Ezt a példányalapú elméletek szempontjából „veszélyes" műveletet, a reprezentációk kialakítását valódi alakok gondos összeolvasztásával oly módon hajtja végre, amely reményei és elképzelései szerint nem vezet az analógiás következtetés szempontjából fontos információk elvesztéséhez (Salzberg 1991, Wettschereck és Dietterich 1995, Domingos 1996). Azaz azok az egyedi tulajdonságok, amelyekre később szükségünk lehet, nem törlődnek az általánosítások során (van den Bosch 2002: 211).

Vizsgálataim szempontjából különösen érdekes van den Bosch rendszerében az **osztály-előrejelző erő** (*class prediction strength*), amely az fejezi ki, hogy egy alak mennyire jól tudja előrejelezni közvetlen szomszédainak viselkedését, mennyire lehet egy példánycsalád felépítésének kezdőpontja. Olyan alakok, amelyek egy nagy szócsalád közepén fordulnak elő, nagyobb osztály-előrejelzési erővel rendelkeznek. Amikor osztályozásra használjuk ezeket a példányokat, akkor ezek megfelelő közvetlen szomszédok lesznek annak a nagyszámú alaknak a számára, amelyek körülveszik őket (van den Bosch 2002: 213–214). Van den Bosch az osztály-előrejelző erőt Salzberg (1990) és Domingos (1995) nyomán határozza meg:

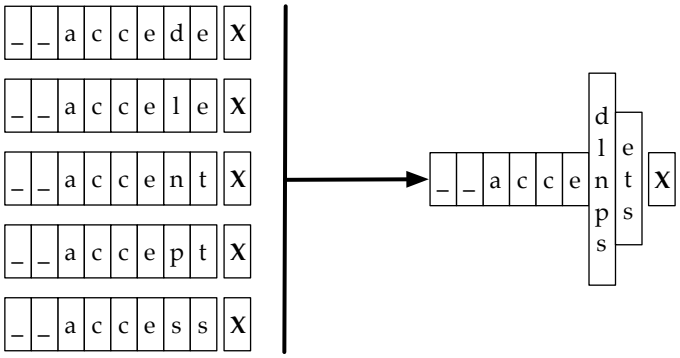
$$\text{(9)} \qquad \text{osztály-előrejelző erő} = \frac{\text{hányszor szomszédja más alakoknak - hányszor szomszédja más osztályba tartozó alakoknak}}{\text{hányszor szomszédja más alakoknak}}$$

Az osztály-előrejelző erő értéke 1 és 0 közt mozog. **1** estén a példány **tökéletesen jósolja** osztálya viselkedését, míg **0** értéket, akkor kap, ha ezt igen **rosszul teszi**. Lényegileg 0 értéket csak az olyan példányok kapnak, amelyek teljesen egyedi módon viselkednek. A megadott képlet azonban elsősorban az alacsony gyakoriságú elemeknek kedvez (Domingos 1995). Ennek a hatásnak a kiküszöbölésére használatos a Laplace-korrekció, amely már a magas gyakoriságú elemeket részesíti előnyben. A képletben (10) a c az osztályok számát adja meg.

$$\text{(10)} \qquad \text{osztály-előrejelző erő} = \frac{\text{hányszor szomszédja más alakoknak - hányszor szomszédja más osztályba tartozó alakoknak + 1}}{\text{hányszor szomszédja más alakoknak + c}}$$

A FAMBL2 a 3.4. ábrán látható módon **készíti el a példánycsalád-leíró kifejezéseket** a családba tartozó elemek összeolvasztásával (van den Bosch 2002: 213). A FAMBL2 egy magas osztály-előrejelzési erővel rendelkező alappéldányból kiindulva a még nem családtag szavakról eldönti, hogy melyik családba sorolhatók be, és a legmegfelelőbbhöz hozzáadja azokat. Az egy családba sorolt alakok alapján elkészíti a példánycsalád-leíró kifejezésüket, és megjelöli az eredeti alakokat, hogy melyik családhoz tartoznak, így nem sorolódhatnak be máshova, és másik családnak sem lehetnek kiinduló pontjai. Ez a folyamat addig tart, amíg az összes alak nem sorolódik

be egy családba. Az egy családba tartozó elemeknek azonosan kell viselkedniük a vizsgált jelenség szempontjából. Ez a kikötés az AM szemléletével rokonítja, mivel a példánycsalád-leíró kifejezésekben az értékek diszjunkcióval való összekapcsolása magában foglalja az egymásra következő homogén szuprakontextusok explicit unióját, amelyek ugyanazzal az osztálycímkével rendelkeznek (van den Bosch 2002: 220).



3.4. ábra: A FAMBL2 reprezentációkészítő mechanizmusának működése

Az AM algoritmussal ellentétben, de az IB1-IG-hez (TiMBL vizsgált változata) hasonlóan azonban van előfordítás, mivel a példánycsalád leíró kifejezések készítése megelőzi az analógiás források keresését. A **tanulási folyamat lezárultával** a következtetésekben a rendszer már nem használja az eredeti előfordulásokat, hanem csak a korábban kialakított reprezentációkra hagyatkozik (van den Bosch 2002: 214). A FAMBL2 az *álom* és az *alom* szavak alapján a következő családreprezentációt hozná létre:

(11) [á vagy a] lom H

Amennyiben a rendszer egy új szóval találkozik, pl. a *halom*-mal, amelyre illeszkedik ez a kifejezés, akkor helyesen jósolja meg a FAMBL2, hogy ez is hangkivető lesz, hisz ebben az esetben csak a szó jobb oldalát határoztuk meg, amely egyezik a *halom* jobb oldalával is (van den Bosch 2002: 216).

FAMBL2 eléri a TiMBL pontosságát[1], míg 20-80%-kal kevesebb helyet foglalnak reprezentációi. Ez azonban az analógiás megközelítések értékrendje szerint nem mondható jelentős előnynek. Ugyanakkor a FAMBL2 tanulási ideje meglehetősen számításigényes folyamat, mivel a tanítóhalmaz teljes memóriaalapú osztályozását követeli meg (van den Bosch 2002: 220). A FAMBL2 további gyengesége, hogy nem modellálja az új adatok bekerülése esetén a rendszer változását[2]. A példánycsalád-leíró kifejezések elkészítési módja a példánygyakoriságra való hivatkozást is nehézzé teszi[3].

3.4. További hasonlóságon alapuló algoritmusok: SimNet, MGL, GCM

A bemutatottakon túl további analógiás megközelítéseket alkalmazó algoritmusokat is kifejlesztettek, amelyek közül még hárommal érdemes röviden foglalkozni. Az AM működési mechanizmusából kiinduló konnekcionista **SimNet** a valós nyelvhasználat (vagy bármilyen más viselkedési adat) helyi reprezentációit hozza létre rejtett rétegeiben. Esetében nincs szükség hosszadalmas betanítási fázisra a bemenetek kezeléséhez. Nem jelent számára gondot az új adatok beépítése sem a rendszerbe. A SimNet az analógiás források kiválasztása során más hasonlóságalapú modellekkel ellentétben képes valószínűségi előrejelzéseket is adni, ami a nyelvhasználat sokféleségét jobban modellezi (Mudrow 2002: 236).

A **SimNet az AM-hez hasonló eredményeket** hoz tesztelési helyzetekben, de ingadozó finn igéken végzett vizsgálatok szerint jobban szimulálja az **ingadozás mértékét** a beszélői ítéletek alapján (Mudrow 2002: 244–245). A SimNet alkalmasnak bizonyult arra is, hogy modellezze a beszélők válaszainak bizonytalanabb, ingadozóbb voltát, amikor gyors döntéseket kellett hozniuk (Mudrow 2002: 246). Ennek a

[1] Van Den Bosch (2002) a korábban bemutatottakhoz hasonló feladatokon teszteli a FAMBL2-t, mint pl. holland kicsinyítés, német többes szám stb.

[2] Elképzelhető, hogy van den Boschnak van erre megoldása, de írásából ez nem derül ki.

[3] Van Den Bosch (2002) ennek kezelésével nem foglalkozik tanulmányában.

viselkedésnek a megragadására a SimNetnél kategorikusabb döntéseket hozó AM kevésbé alkalmas.

Az **MGL** (The Minimal Generalization Learner, Minimális Általánosítás Tanulórendszer, Albright és Hayes 2002, Albright 2009) olyan szabályokat[1] keres, amelyek azonos módon viselkedő szavakat fednek le (pl. ugyanabba a ragozási osztályba tartoznak, azonos morfofonológiai változásban vesznek részt stb.). Alulról felfelé építkezik egy induktív stratégia mentén szópárok összehasonlításával. Az MGL megkeresi, hogy a pároknak milyen közös tulajdonságai vannak, és ezeket a közös vonásokat sztochasztikus szabályok segítségével rögzíti (Albright 2009: 193). Az új elemek besorolásánál az illeszkedő szabályokat veszi figyelembe, és ezek valószínűségi értékei alapján választ analógiás forráscsoportot.

i. o → wé / [+mássalhangzós] __rs

ii. o → wé / [- folyamatos, - zöngés] r __ [- folyamatos, - szillabikus]

iii. o → wé / [- szillabikus, + mássalhangzós] __ [- szillabikus]

iv. o → ó / [- szillabikus, - szonoráns, + mássalhangzós] __ [- szillabikus, + mássalhangzós, - folyamatos]

v. o → ó / [- szillabikus, + zöngés] __ [- szillabikus]

vi. o → ó / __ [- szillabikus]

3.5. ábra: Az MGL által létrehozott szabályok, amelyek leírják, hogy spanyol igékben mikor diftongizálódhat a hangsúlyos *o*.

Albright (2009: 191) az elterjedtebb AM helyett a klasszikusabb **GCM-mel** (Generalized Context Model, Általánosított Kontextusmodell) hasonlítja össze

[1] Albright (2009) szabályokról beszél, de lényegében szabályai nem szabályok a generatív értelemben, hanem csak tendenciákat rögzítenek.

rendszerét, amely mint az AM és a TiMBL egyedi alakok hasonlóságával számol. A GCM is használja a példánygyakoriságot, szimulálja a nem tökéletes memóriát, de a heterogenitást nem veszi figyelembe. A GCM kategorizációs feladatokban jól teljesített korábban (Chandler 2002: 72–73). Albright (2009: 190) úgy véli, a tesztjében jobban teljesítő MGL eredményei alapján, hogy az egyes alakokon alapuló analógiák használata nemcsak hogy nem szükséges, hanem inkább megnehezíti a jó besorolást. Albright (2009) eredményeit azonban fenntartással kell kezelnünk, mert nem ismerteti, hogy tesztszavai miképp készültek el, és nem tudjuk azt sem, hogy a gyakorisági adatai honnan származnak. Az összehasonlításban a GCM a Levenshtein-algoritmust (bővebben 4.3. alfejezet) használta, amelyről később megmutatom, hogy az analógiás modellezésre kevésbé alkalmas (6. fejezet), így rosszabb eredménye önmagában a Levenshtein-algoritmus gyengeségének is betudható, ezért nem következik a teszt eredményéből egyértelműen, hogy az egyes alakokon alapuló hasonlóság ne játszhatna szerepet az analógiás folyamatokban.

4. Analógiás működés, hasonlósági mértékek

4.1. Az analógiás minta kiválasztása

Az analógiával kapcsolatos fogalmak és az ezekhez köthető elméletek bemutatása után ebben a fejezetben ismertetem, hogy elképzeléseim szerint **miképp érvényesül az analógia az egyes szóalakok produkciójában és a szavak változásában**. Ezt követően a 4.3. alfejezetben javaslatokat teszek arra, hogy az eddigi elképzeléseknél (Skousen és mtsai 2002) nyelvközelibb, általánosabb és holisztikusabb (Bybee 2010: 61) módon hogyan mérhető az analógiában kulcsfontosságú hasonlóság. Az analógiás működés leírásában és a hasonlóság számításában is az 1. és a 2. fejezetben bemutatott érvek alapján kizárólag **felszíni alakokra** (korlátozásunkból adódóan karaktersorokra) hagyatkozok. Absztraktság helyett konkrétságra törekszem (Kálmán 2008), mivel számos kutatás alapján megkérdőjelezhető a mögöttes alakok léte (Bybee 2000, 2001, Benua 1995, 1997, Steriade 2000), így tartózkodni fogok ezek használatától. Ezt a megközelítést támasztják alá Szilágyi (2010) gyergyói példái, és az azokra adott elemzése is. Ebben a nyelvjárásban a *húzza*, *eressze* köznyelvi alakok helyett a *hújza*, *erejsze* alakok találhatók meg[1]. A beszélők így elkerülik a palatalizációt, azaz a szóalakok szerkezetükben kevésbé, de „alapanyagukban" közelebbiek maradnak a tőhöz. A felszólító módért felelős felszíni rész (*-j_a*/*-j_e*) bármilyen mögöttes alakból való levezetése különösen körülményes lenne ezen alakok esetében, mivel ha a szóalakokat morfémákra akarnánk bontani, akkor a felszólító mód toldalékát nem tudnánk a szabályalapú megközelítéseknek megfelelő módon sem szuffixumnak, sem infixumnak vagy circumfixumnak kategorizálni, és képtelenek lennénk felszíni szétválásáról számot adni[2]. Így a nehézkes levezetés és a kivitelezhetetlen szegmentálás helyett inkább egy felszíni sémáról és az ahhoz kapcsolódó jelentésről célszerű leírást adnunk.

[1] Szilágyi a példák említésén túl nem definiálta előadásában egyértelműen a jelenség körét.

[2] Vagy ebben az esetben meg kellene engednünk a mögöttes szegmentáció után hatásba lépő metatézist, ami önmagában egy olyan eszköz, amelynek a használatát még a generatív keretben is igyekeztek megszorítani.

A **felszíni alakok tárolása** azonban sokféleképpen lehetséges. Elképzeléseim szerint ebben sem a gazdaságosság, hanem a megfelelő pontosság az elsődleges. Az egyes szavakat gazdag fonetikai részletességgel tároljuk (Pierrehumbert 2001)[1], amelyben szerepet kaphatnak a használati kontextusok is[2]. E fonetikai változatok klasztere a jelentések, interferenciák és a szó használatát befolyásoló nem nyelvi tényezők jellemzéséből álló klaszterhez kapcsolódik (Pierrehumbert 2002, Bybee 2006, Bybee 2010: 14), azaz a saussure-i forma-jelentés párok itt is szerepet kapnak, csak nagyobb részletességgel, illetve kibővített tartalommal. Egy szó újabb előfordulásai ezt követően megerősíthetik ezeket az emlékeket, míg a nyelvhasználat során bizonyos tényezők hatására gyengülhetnek is (Kálmán 2009).

A hasonlóságon alapuló analógia a nyelvi változásban fontos alakító erő. Azonban nemcsak a nyelvtörténetben vagy a kivételes szavak rendszerében tölt be jelentős szerepet, mint ahogy azt már generatív megközelítések is elismerik, hanem a nyelvhasználat legáltalánosabb, szinkrón folyamataiban is alkalmazzák a beszélők. A nyelvtörténetben és a szinkrón állapotokban működő analógia azonosnak tekinthető (Bybee 2010: 72), mivel hatásmechanizmusuk egységesen megragadható analógiás aránypárokkal, illetve azok kiterjesztéseivel.

A **klasszikus analógiás** következtetési módot én is **kiindulási alapnak** tartom (Itkonen 2005: 34), miszerint három ismert alakból létrehozhatunk egy negyedik alakot, ha úgy gondoljuk, hogy a már ismert alakok közül kettő viszonya kiterjeszthető a harmadikra és egy ismeretlen negyedikre, amelynek alakja a már ismert viszony és a harmadik alak alapján kiszámolható. A már ismert pár egyik tagjának valamilyen szempontból a harmadik alakhoz hasonlónak kell lennie. (1) alatt egy ma is produktív paradigmába való besorolást láthatunk egy idegen szó átvételekor/használatakor:

---

[1] Jelenleg a modellezésben gondot jelent ennek a beépítése, ezért a gazdag fonetikai részletesség inkább elméleti lehetőség, a valós gyakorlatban nem tükröződik (Sóskuthy 2010: 32–34).

[2] Goldinger (1996) kísérleti eredményei szerint egy időre az egyedi beszélők hangminőségeit is elraktározzuk memóriánkban.

(1) repeta : repetát

feta : x

x = fetát

A hasonlóságnak elsősorban strukturálisnak, nem pedig materiálisnak kell lennie, amit Gentner (1989: 201) is megfogalmaz az analógiával mint általános kognitív folyamattal kapcsolatban:

„egy analógia annak az ismeretnek a leképezése egy tartományból (forrás) egy másikra (cél), amely megállapítja, hogy a forrás elemei közt fennálló viszonyok rendszere a cél elemei közt is érvényben van."

A klasszikus **analógiás aránypáron alapuló megközelítések** jól megragadják a nyelvi működés alapjait, de **homályban hagyják a minták kiválasztásának módját** (bővebben 2.1. alfejezet). Ismereteink alapján elképzelhető, hogy az analógiás következtetésben több szó, alak is szerephez jut (Kiefer 2002: 13–14), amely folyamat megragadására szintén nem alkalmas az analógiás aránypár (ezzel a feltételezéssel él a 3. fejezetben bemutatott AM és TiMBL is). Egy szó analógián alapuló produkciójában és megértésében így **szavak csoportjai és azok erőviszonyai játszanának közre**. A csoportok viselkedésében prototípusok is szerepet kaphatnak, amelyek csoportjuk életképességét és felismerhetőségét jelentősen befolyásolhatják (Bybee 2010).

A szavakat az **egyformán viselkedő szócsoportok (paradigmák) eltérő erővel vonzzák,** típusgyakoriságuk és közelségük függvényében meghatározva azok produkcióját és megértését. A paradigmatikus megközelítés azonban csak közelebb visz minket az igazsághoz, de a szavak egyediségét, eltérő jelentésbeli és materiális felépítését túlzottan általános jellege miatt nem tudja teljesen megragadni[1]. Wittgenstein nyomán (1992: 49) ezért úgy gondolom:

[1] Mint az 5.4.2. alfejezetben látni fogjuk, egy paradigmába tartozó szavak nem homogén masszaként viselkednek, hanem nagyfokú variabilitást mutatnak egyes paradigmatikus celláik kitöltésében.

„Hogy világosabban lássunk, itt is – mint számtalan hasonló esetben – a folyamatok részleteit kell szemügyre vennünk; *közelről szemlélnünk,* ami végbemegy."

Mielőtt azonban a csoportviselkedés részletesebb tárgyalásába belekezdenék, példával is bemutatom az **analógiás produkció működését** a könnyebb áttekinthetőség és érthetőség érdekében[1]. A folyamatban a legfontosabb első lépés, hogy le tudjuk-e hívni a memóriánkból az aktivált jelentéshez egy megfelelő tárolt alakot (4.1. ábra)[2]:

+ACC

4.1. ábra: A *kaprot* „jelentése"

Amennyiben a megfelelő alakot tároljuk memóriánkban, és le is tudjuk hívni, akkor képesek vagyunk a *kaprot* alak használatára az adott kommunikációs helyzetben. Ennek valószínűségét elsősorban az alak példánygyakorisága határozza meg (Bybee 2001):

„Egy morfológiailag komplex alak tárolását nem szabályos vagy rendhagyó besorolása, hanem használatának gyakorisága határozza meg."

[1] Azért választottam egy produkcióhoz kötődő példát, mert vizsgálataim is erre irányulnak. A megértésben az analógia szintén fontos, de korpuszalapon kevésbé vizsgálható.

[2] A jelentésre utaló kép csak az egyszerűség kedvéért szerepel itt. Gondolatmenetem szempontjából nem fontos, hogy ilyenkor aktiválódnak-e képek fejünkben.

Memóriánkban több, a fonetikai eltéréseken túlmutató alak is tárolódhat pl. *motrot*, *motort*[1] egy megadott „jelentéshez” (4.2. ábra).

+ACC+COLL

4.2. ábra: A *motort, motrot* „jelentése” (COLL = kollokviális)

Ha két (vagy több) lehetséges alakot érünk el, akkor azok **erőviszonya** vagy a **beszédhelyzet** (pl. informális vagy formális) **dönthet** a kiválasztásukban. Amennyiben képtelenek vagyunk egy megfelelő alakot kiválasztani, akkor nem ismertük, elfelejtettük, vagy nem tudjuk lehívni[2]. Ha csak a lehívásban vagyunk gátoltak, akkor elképzelhető, hogy egy másik hasonló helyzetben már hozzáférünk az alakhoz.

Ha egyetlen tárolt alakhoz sem tudunk hozzáférni, akkor **magunknak kell létrehoznunk a szükséges alakot**. A szabályalapú elméletek ilyenkor egy általános, vagy a hangkivető szavak esetében egy kivételes eseteket kezelő szabállyal hozzák létre a kívánt formát. Ha azonban egy szó kivételes, de nincs meg a szótárunkban, akkor annak viselkedéséről semmit sem tudunk, így a szabályok birtokában sem tudjuk létrehozni a megfelelő alakot, hisz nem tudjuk, hogy melyik szabályt kell vagy lehet rá alkalmazni. Ezzel szemben, mint a 7. fejezetben bemutatásra kerülő tesztem eredményeiből is láthatjuk, a magyar beszélők képesek sosem hallott, kitalált szavaknak is hangkivetéses alakjait létrehozni. Ezzel összhangban a létező *köböl* hangkivető főnév a legtöbb magyar beszélő számára ismeretlen lehet, mégis amennyiben szükséges lenne tárgyesetét képezniük, lenne elképzelésük arról, hogy ezt miképp kell tenniük.

---

[1] Felsorolhatnánk itt a *motort* további informális változatait is, mint *mocit*, amelyet azonban még nagyobb fokú familiaritása miatt kevésbé aktiváltnak vehetünk esetünkben. Ez a két alak is csak a beszélőközösség egy kisebbségben lévő részének aktiválódik, a többség csak a *motort* alakhoz fér hozzá, a másik alakot legfeljebb ironizáló szándékkal használhatja.

[2] Chandler (2002) a sikertelen elérés esetén a teljes felejtéssel szemben a lehívás gátolt volta mellett érvel, amit én is valószínűbbnek tartok.

Az analógiás megközelítés azt feltételezi, hogy egy új alak létrehozásakor **már meglévő nyelvi mintáinkra hagyatkozunk**, amelyeket a hasonlóság alapján választunk ki. Az analógiás forrás kiválasztását ezen túl az egyes viselkedéscsoportok elemeinek gyakorisága befolyásolja. A csoport teljes hatóereje elemei gyakoriságának és a hasonlítás alapját képező szótól való távolságuknak a függvénye:

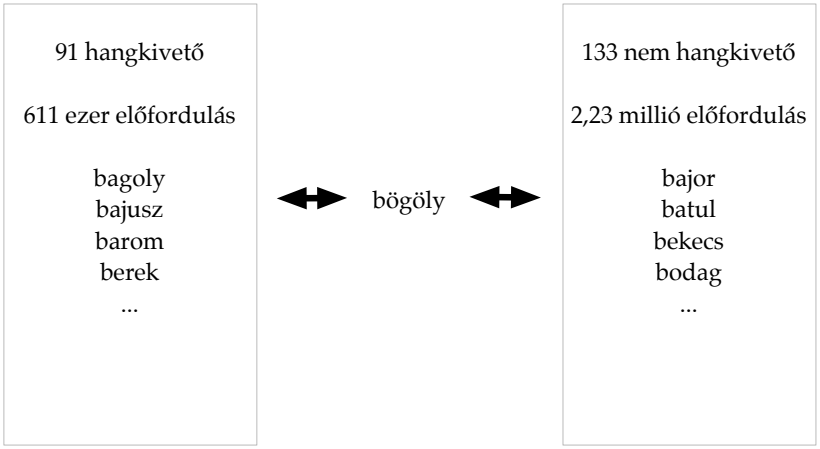
(2) n elemű csoport hatóereje = $(\log_{10}x_1*y_1)+(\log_{10}x_2*y_2)+...+(\log_{10}x_n*y_n)$

Ahol x egy szó gyakorisága és y az adott szó hasonlósága egy 0-1-ig terjedő skálán ahhoz a szóhoz, amelyhez analógiás mintát keresünk. E képlet csak egy a sok **lehetséges számítási mód** közül, ahogy a gyakorisági és hasonlósági hatásokat összegezhetjük. A logaritmus használatát a gyakoriság esetében a Weber-Fechner törvény (Dehaene 2003) indokolhatja, miszerint pszichofizikai mennyiségek (pl. súlyérzékelés, látás) esetében a logaritmussal megragadható léptékváltásokat érzékelik az emberek. E képlet alapján számolva a *bögöly* szóra 21,7-szer erősebben hatnak a nem hangkivető szavak a *morphdb.hu* 48 ezer főnevét figyelembe véve (Trón és mtsai 2006, Halácsy és mtsai 2003), mint a hangkivetők, így hangkivető használata elsősorban annak köszönhető, hogy releváns alakjaira emlékszünk. A felépítésében sokkal inkább tipikus hangkivetőnek számító *tartalom* esetében, amely számos közeli mintával van körülvéve, ez a szám már csak 10,3, ami kiválóan összhangban van stabil hangkivető viselkedésével is, amit természetesen kiugró gyakorisága is támogat.

Az ilyen **hatások jobban érvényre jutnak** a nyelvekben az **új szavak besorolása esetén**, amelyek általában a nem rendhagyó szócsoportokba kerülnek[1], hisz ezek esetében nincsenek a szóhoz tartozó memorizált alakok, amelyek használatában közvetlen útmutatást jelenthetnének. Az egyenlőtlen erőviszonyok a már besorolt szavak átrendeződését is eredményezhetik, aminek a kiszámolásában az adott szó már

[1] A 7. fejezetben nyelvi tesztemben ezzel látszólag ellentmondóan viszonylag nagy arányban, 36,8%-ban választottak hangkivető alakokat a résztvevők. Ez részben azonban annak tudható be, hogy azt gondolhatták, hogy meglévő alakokat sorolnak be és nem újakat, illetve a hangkivetésre való hajlandóságot erősíthette az ezekhez a szavakhoz kötődő népies-történelmi hangulat megteremtése a tesztben.

korábban hallott alakjainak is szerepet kell kapniuk, mégpedig minden bizonnyal nagyobb súllyal, mint amekkorával a csak hasonló alakok bírnak. A csoportok ilyen erőviszonyai folyamatosan alakulnak (bővebben 5.4.2. alfejezet) a gyakorisági és viselkedésbeli változásoknak tulajdoníthatóan, aminek következtében nyelvi rendszerünk is változhat.



4.3. ábra: A *bögöly* szó legközelebbi szomszédai (a $CVC_{\alpha}(a/o/u/e/ö)C_{\beta}\#$ sémának megfelelő szavak (Rebrus és Törkenczy 2008, bővebben 5.1. alfejezet), ahol C_{α} és C_{β} nem azonosak és szomszédosságuk esetén nem hasonulnának, olvadnának össze) által megtestesített erőviszonyok

A csoportok hatóerejét egyes szóalakok létrehozásakor vélhetőleg nem számoljuk ki minden esetben[1]. Ilyenkor a **legnagyobb gyakoriságú és leghasonlóbb alakok sora aktiválódhat** (ezek gyorsabban is aktiválódnak: Jurafsky 2003) és ezek erőviszonyai határozzák meg a végleges alakot. A megfelelő minták keresése akkor állhat le, ha a gyorsan aktiválódó gyakori alakok közt megfelelően közeli mintát találunk, és nem feltétlenül a legjobbat, ezért az analógiában egy gyakori hasonló, de nem a leghasonlóbb alak könnyebben lehet analógiás forrás (Chandler 2002). Ennek megfelelően a ritka alakok közt csak akkor keresnénk forrást, ha a gyorsan aktiválódó gyakoriak közt nem leltünk megfelelőt.

[1] Az erőviszonyok kiszámításának jellege erősen függ attól, hogy mennyire gyorsnak és jó becslőnek tartjuk az emberi agyat. A bemutatott, látszólag számításigényes képlet kiszámolása 51 ezer főnév esetében a hasonlósági viszonyok meghatározásával együtt is egy átlagos számítógéppel egy másodpercen belül végrehajtható.

Az **analógiában szerepet játszó hatások számítása típus-** és **példány-alapon** is lehetséges. Ha típus (type) alapján számítjuk a gyakoriságot, akkor azt vesszük figyelembe, hogy egy adott csoportnak hány eleme van, így az *-alom* végű hangkivető főnevek típusgyakorisága 268 a *morphdb.hu* szótár alapján (Trón és mtsai 2006). A példányalapú (token) gyakoriság számításában azt vesszük figyelembe, hogy a csoport egyes elemei hányszor fordulnak elő. Az *-alom* végű hangkivető főnevek példánygyakorisága így a hangkivetéssel együttjáró alakjaikat figyelembe véve 1,16 millió a *Szószablya Gyakoriségi Szótár* alapján (Halácsy és mtsai 2003). Albright (2009: 206) a típusgyakoriság kizárólagossága mellett érvel pszicholingvisztikai adatokra hivatkozva. Eddington (2003) is elsősorban típus alapú hatásokat mutatott ki a spanyol hangsúlykiosztással és angol igék múlt idejével kapcsolatban. Ugyanakkor modellezéseiben látható volt, hogy vannak példány-alapú hatások is, mivel legjobb analógiás mintának a közepes gyakoriságú példányok mutatkoztak, mert a nagy gyakoriságúak túl autonóman viselkednek (Bybee 2010: 34), míg a ritkák nehezebben aktiválódnak, ha mintákat keresünk. Bybee (2010: 67) szintén hangsúlyozza a típusgyakoriság fontosságát, mivel olyan nagy típusgyakoriságú csoportok tudnak igazán produktívak lenni, amelyek sémájukban is viszonylagosan homogének, azaz elemeik meglehetősen hasonlítanak egymásra[1]. Ennek köszönhető, hogy a magyar hangkivető főnevek közül az *-alom* és az *-ök* (lásd 7.3. alfejezet) végűek mutatnak minimális produktivitást (homogén felépítés, viszonylag nagy típusgyakoriság), míg a sémájukban egységes, nagy példánygyakoriságú, de alacsony típusgyakoriságú *-og* végűek nem.

A **produktivitásban** nagy szerepet játszó **típusgyakoriság** mellett azonban megfigyelhetünk olyan jelenségeket is, amelyek inkább példány-alapon igazolhatók (Bybee 2010: 96, Kálmán és mtsai 2010). Bybee (2001, 2010) számos példát hoz nagy példánygyakoriságú elemek redukciójára, amelyekhez hasonlókat a magyarban is találunk (*I don't know -> I dunno* 'Nem tudom', *nem tudom -> nemtom*, *azt hiszem -> asszem*). Fowler és Housum (1987) alapján azt is tudjuk, hogy egy szó egy szövegben

[1] Minél nagyobb egy csoport, ez a szempont annál kevésbé fontos (lásd King (1969) tárgyalását a 2.2. alfejezetben)

második ismétlésekor már rövidebb, illetve redukált használatát a szövegen belüli általános példánygyakorisága is befolyásolja (Bybee 2002). Nagy példánygyakoriságú elemek könnyebben lehetnek prototípusok egy csoporton belül (Bybee 2010 : 104), illetve a grammatikalizáció is elsősorban nagy példánygyakoriságú elemekhez köthető (Bybee 2010: 214–219, lásd 6.5. alfejezet). A nyelvelsajátításban mind a kétfajta gyakoriság befolyására látunk példákat (Daçbrowska és Lieven 2005, Tomasello 2003).

Láthatjuk, hogy a típus-példány vitában[1] mind a **kétfajta gyakoriségi hatást alátámasztják adatok**, ami azonban inkább azt sugallja, hogy eltérő nyelvi feladatokban hol a példány-, hol a típusgyakoriság, vagy ezek együttes hatása érvényesül (pl. kategóriák meghatározása, Bybee 2007: 15). Nem szabad figyelmen kívül hagynunk azt sem, hogy nagyobb típusgyakoriságú csoportoknak értelemszerűen a példánygyakorisága is valószínűbben magas, így a két hatást sok esetben nem is lehet jól szétválasztani egymástól, ezért több esetben az 5. fejezetben mi is látni fogjuk, hogy a típus- és példánygyakorisággal egyszerre függnek össze bizonyos paraméterek.

Az elmélet szempontjából azonban kulcsfontosságú, hogy meghatározzuk, hogy **mikor melyik hatás mekkora mértékben érvényesül**, de jelen ismereteink alapján csak azt állíthatjuk, hogy a gyakorisági hatások léteznek (Eddington 2003), de azok pontos viselkedéséről az analógiában még nem rendelkezünk elég tudással. Ebből kifolyólag elemzéseimben mind a kétfajta gyakoriságra hivatkozok, de lényegesen több esetben hagyatkozok a példány-alapú gyakoriságra, mint ahogy azt eddigi példáimban is tettem. Döntésemet az motiválja, hogy a sajátosan viselkedő elemeknél a példányalapú gyakorisági hatások konzerváló szerepét mutatták ki eddig (Bybee 2007: 10; Bybee 2010: 13, 24, 75), mivel az ismétlések megerősítik az emléknyomokat. Ezt mondja ki az ún. morfológiai stabilitás elve (Mańczak 1980) is, amely szerint a nagyobb gyakoriságú elemek hajlamosabbak ellenállni a változásnak. Természetesen kis gyakoriságú elemek is változatlanul maradhatnak, ha csoportjuk típusgyakorisága megfelelő mértékű (pl.

[1] A korai vizsgálatokban a típusgyakorisággal kapcsolatos eredmények elsősorban annak voltak köszönhetőek, hogy gyakorisági hatásokat csak ezekkel lehetett vizsgálni, hisz nem voltak megfelelő gyakorisági adatbázisok a példánygyakoriság vizsgálatához.

-eder végű hangkivetők, mint *veder, meder* stb.), vagy ha a nagy példánygyakoriságú elemekhez nagyon hasonlítanak (pl. *horog* a *dolog*-hoz).

A nyelvi változásban, produkcióban és megértésben **számos további használati tényező szerepet kaphat** a gyakoriságon túl: a szó előfordulási kontextusai, a beszélő korábbi tapasztalatai, szociális, kulturális, történelmi, politikai, földrajzi stb. szempontok (Bybee 2010: 42). Így például a Pápua Új-Guinea-i yimasban egyes szavak használatában az ingadozás azzal hozható összefüggésbe, hogy a referenciájukat képező állatra a beszélő miképp tekint, mivel nézőpontja függvényében azokat vagy a „fontos állat" ragozási osztályba sorolja, vagy csupán fonológiai formájuk alapján toldalékolja (Aronoff 1994: 115). A magyar települések toldalékolását szintén nyelven kívüli szempontok is meghatározzák, mivel lokatívuszuk alakját befolyásolja, hogy a referenciájukat képező hely a történelmi Magyarország területén[1], vagy azon kívül található-e, ami a határok változása miatt nem egyértelmű fogalom, és így ingadozáshoz, bizonytalanságokhoz vezet. A lokatívusz viselkedését még bonyolultabbá teszi az archaikus lokatívusz, a -(*Vt)t* használata, amelynek alkalmazására és így a szó alakjainak viselkedésére az is hatással van, hogy a beszélők a referenciáját képező helyet történelmileg mennyire tekintik jelentősnek: *Érsekújvárott* (történelmileg jelentős erődítés: 50%-ban *-Vtt* alakok lokatívuszban) : *Kaposvárott* (kevésbé jelentős mezőváros: 20,8% -ban *-Vtt* alakok lokatívuszban).

Bizonyos esetekben az analógiás változással szemben kulturálisan vagy szociálisan meghatározott ellenállás is felléphet, mivel a kiegyenlített alakokat a beszélők egy csoportja túlzottan informálisnak vagy helytelennek véli. Ennek jelentőségét Magyarországon, ahol a nyelvművelés és a helyes beszéd kultusza nagy becsben áll, nem szabad figyelmen kívül hagynunk, különösen, hogy írott szövegekkel dolgozunk. Az analógiás változásra, produkcióra és megértésre hatással lehetnek **idiomatikus szerkezetek**, kifejezések is, amelyek a változást visszafogják. Így a *fátyol* szó esetében a változását lassíthatja az olyan kifejezések, konstrukciók megléte és

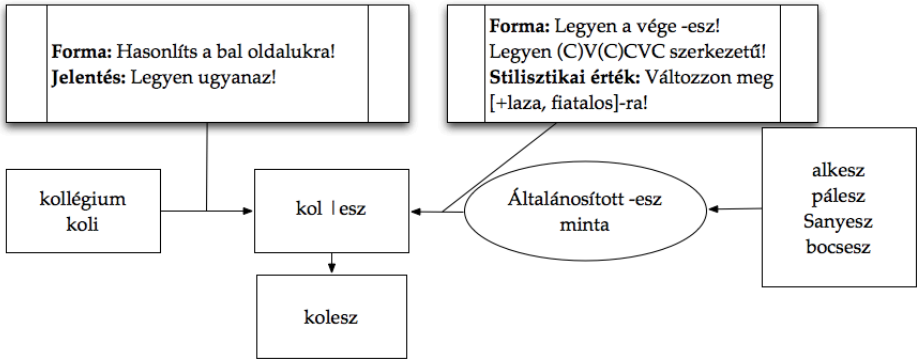
[1] Ez a meghatározás területenként változhat, mert az erdélyi magyar beszélők „határon belülinek" tekintik a román települések némelyikét is, és ennek megfelelően toldalékolják azokat: *Konstancán*, *Plojestin*.

ismerete, mint *fátlyat rá, fellebbenti a fátylat* stb. Egyes esetekben pedig az ilyen formák az ingadozást mint állandósult állapotot rögzíthetik: *fátyolos tekintet, fátylas özvegy.*

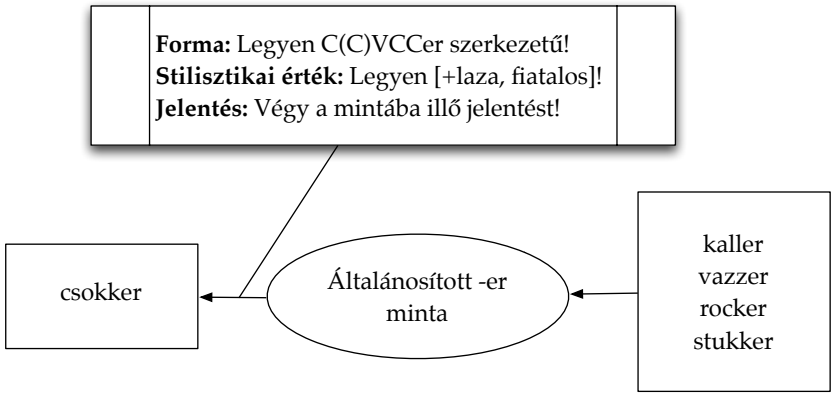
A megértés és a produkció gördülékenyebbé tétele érdekében a beszélők létrehozhatnak **általánosításokat** is, amelyek nem kizárólagosak, és nem zárják ki az egyedi minták hatását sem. Az általánosítások megalkotása után az egyes szóemlékek sem törlődnek, így később is lehetséges az elérésük. Bybee alapján a szabályalapú és a sémaközpontú megközelítések különbségeit (Bybee 2007, Lukács 2002) a következőképpen határozhatjuk meg:

- A sémák ráépülnek a lexikon elemeire, amelyek a sémák létrehozása után is megőrződnek.
- A sémák viselkedését befolyásolja gyakoriságuk.
- A sémákat befolyásolják a létező típusok specifikus tulajdonságai.
- A sémák fokozatos és egymásba átmenő kategóriák, míg a szabályoknál egy formára vagy alkalmazható egy szabály vagy sem.

A sémák specifikus viselkedését a 4.4., 4.5. ábrák mutatják meg egy egy olyan szóalkotási folyamatban, amelyeket szabályalapon nem tudunk megmagyarázni, különösen az *-er* végű mintába illő szavakat, amelyek közt több még csonkolást feltételezve sem értelmezhető (pl. *csokker* : **csok(k)-*, *vazzer* : **vaz(z)-* , Fűköh és Rung 2005)



4.4. ábra: Az -esz minta alkalmazása



4.5. ábra: Az -er minta alkalmazása

4.2. A prototípusok szerepe az analógiában

Az egyes szóalakok létrehozásában és megértésében (komoly) szerepe van a **prototípusoknak** (Langacker 1991: 295), amelyek segítségével **gyorsabban és hatékonyabban tudunk megfelelő analógiás forrásokat találni**. Különösen az olyan elemek viselkedésére lehetnek nagy hatással, amelyek egyes csoportok közt átmenetet képviselnek, mint pl. a *halom*, amely a hozzá közeli *-alom*/*-elem* végűeket a hangkivetéses minta alkalmazásában követi, de a *CVCVC* szerkezetű hangkivetőkkel is mutat rokon vonásokat, mivel a toldalékok nyitás nélkül kapcsolódnak hozzá.

Bybee és Slobin (1982), valamint Bybee és Moder (1983) igazolták, hogy az **angol igék kivételes osztályai prototípusosan szerveződnek,** és az új formák minél több prototipikalitási jegyben megegyeznek a prototípussal, annál valószínűbben

terjeszthető ki rájuk a múlt idő kivételes sémája (bővebben 3.2. alfejezet). Saját kutatásaimban is azt tapasztaltam, hogy az archaikus *-(Vt)t* végű lokatívusz alkalmazását egyes településneveknél a hozzájuk különösen hasonló nagy gyakoriságú prototípusok is erősíthetik: pl. *Győr – Diósgyőr, Felsőőr, Alsóőr, Borsosgyőr; Kolozsvár – Szatmár, Jászvásár, Soroksár, Sárvár*. A prototípushoz való közelséget ebben az esetben is a formai hasonlóság dönti el, mivel a *Győr-Felsőőr*, *Kolozsvár-Soroksár*[1] stb. pároknál az egyezés csak részleges még az utótagban is, így a *Felsőőr, Szatmár* szavak használata nem vezethető vissza bizonyos morfémáknak (*vár, Győr)* a lexikonban megjelölt sajátságos viselkedésére.

A **prototípusok** nyelvészetben használt **változatait** Chandler (2002: 55) az alábbi csoportokban határozza meg:

1. **gyűjtőfogalom** a Rosch (1973) által leírt „ prototípus-hatásokra", valamint a következő pontok által is meghatározott esetekre;
2. hipotetikus **mentális reprezentáció**, amely a érzékelési tapasztalásokból összegződik és épül fel (Posner és Keele 1968);
3. minták egy csoportjának **„súlypontja"**, amely agyunkban rendelkezhet önálló reprezentációval is;
4. **"központi tag"**, az adott kategória kiugróan reprezentáns képviselője, amely az egész csoportot képviselheti (Hallan 2001: 91).

Vizsgálataimban az utolsó változatot alkalmazom azzal a különbséggel, hogy **prototípusok létezhetnek jól definiálható csoporttulajdonságok nélkül is**. A prototípusok nem gátolják csoporthatások érvényesülését (vö. 3.4 alfejezet, FAMBL2). A prototípusok „ereje" azonban nemcsak önmagukból, hanem olyan szavakból is származhat, amelyek velük közvetlen vagy közvetett kapcsolatban vannak. Nosofsky (1988) megmutatta, hogy a prototipikus viselkedésben a gyakoriság is szerepet játszik, amit Bybee és Eddington (2006) kutatásai is megerősítettek. Hasonlóan a 6.5.

[1] A hasonlóság azonban nem csak az utolsó két fonéma közt áll fenn, hisz a két szó kapcsolatában még további közös vonások is megfigyelhetőek, amelyek a $CoC_{[son,cor]}oCC_{fric}ár\#$ sémával jellemezhetőek.

alfejezetben bemutatásra kerülő kutatásaimban azt tapasztaltam, hogy a gyakorisági alapon kiválasztott prototípusoktól való távolság jól jósolja egy hangkivető alak hangkivetésének mértékét.

Prototípus-kiválasztásunknak algoritmizálhatósága érdekében **mérhető adatokon kell alapulnia**. A kiválasztás során egy szóval kapcsolatban a következő ismert és számszerűsíthető tulajdonságai jöhetnek szóba:

- gyakorisága
- egyes toldalékok esetében való ingadozása
- más alakokhoz való hasonlósága valamilyen algoritmus alapján
- más potenciális prototípusoktól való különbözősége valamilyen algoritmus alapján
- milyen alakjait használjuk, és milyen arányban

A prototípushatások modellezésével a 6.5. alfejezetben foglalkozok, ahol részlegesen ezeket a szempontokat építjük be modellünkbe, amely a nagyon erős gyakorisági hatás mellett a többi hatás mérsékelt érvényességét is megmutatta. A prototipikus viselkedés meghatározásában további szempontok is szerepet kaphatnak, de ezekről még kevesebbet tudunk. A **szavak szerveződésére hatással lehet jelentésük**, ezért a későbbiekben érdemes lenne bevonni a vizsgálódásokba a magyar *Wordnet*et (Miháltz és Prószéky 2004) is. Ezt azonban a jelenlegi kutatásból kihagytam, mert fontosabbnak tartottam első lépésben néhány jól azonosítható faktort megragadni, és ezek hatását felmérni, hogy a későbbiekben szemantikai és pragmatikai információkkal is kiegészíthetőek legyenek.

4.3. Algoritmusok a hasonlóság mérésére

A **szavak alaki hasonlóságának meghatározására kerestem egy egyértelmű algoritmust**, amellyel modellezhettem ezt az analógiás források kiválasztásában fontos

szerepet betöltő képességet. Ehhez vitt közelebb a korpuszadatok tanulmányozása (5. fejezet), és az ezek alapján formált elképzelések beszélőkkel való tesztelése (7. fejezet). Amennyiben kitűzött célomat elérem, az a nyelvtechnológia számára is értékes hozadékkal jár, hisz egy pontos, a valós folyamatokat jól megragadó algoritmus segítségével lehetővé válik a szavak hatékony szótárba sorolása (megfelelő jegyeikkel), illetve a már meglévő szótári anyag frissítése, karbantartása is, amely komoly kihívást jelent, ha csak emberi erőre hagyatkozunk. A szótárak automatikus bővítésével lehetővé válik akár rétegnyelvi (szleng, szaknyelv stb.) szövegek hatékonyabb elemzése is. Az analógiás megközelítéssel magyar szövegek morfológiai elemzése és produkciója is megoldhatóvá válhat (Stroppa és Yvon 2005).

A feladat megoldására **több algoritmusvariációt** is kidolgoztam, amelyek működési elveit szükséges áttekintenünk, mivel az 5. fejezetben a hangkivető főnevek vizsgálatában már támaszkodok az ezek alapján készített mérésekre[1]. Úgy véltem, hogy ezekre támaszkodhatok az elemzésben, mivel kutatásom korábbi fázisaiban már sikeresnek bizonyultak hasonlósági viszonyok felismerésében (Rung 2008, Rung 2009).

Az algoritmusok célja, hogy a közvetlenül vizsgált nyelvi jelenségeken túl általánosabban is **megragadják a szavak összehasonlítása során működő mechanizmusokat**. Ugyanakkor ígéretes eredményeik ellenére nem gondolom, hogy az itt bemutatásra kerülő algoritmusok univerzálisak, vagy akár a magyar nyelven belül is kizárólagosan érvényesek lennének[2]. A hasonlítás, arányítás nyelvenként és akár feladatonként is jelentősen eltérő lehet. Ezzel nem azt kívánom tagadni, hogy a hasonlításnak lehetnek, sőt vannak univerzális sajátságai, de úgy vélem, egyes analógiás megközelítésektől is eltérően, hogy a nyelvi univerzálék keresésének erőszakolt előtérbe helyezése nagyon megnehezíti az egyedi folyamatok alaposabb megfigyelését, értelmezését.

---

[1] Az algoritmus finomítása a nyelvi adatok tanulmányozása alapján egy iteratív folyamat. A frissen megfigyelt tapasztalatok használhatók az algoritmus javítására, ugyanakkor az algoritmus által megállapított hasonlóságok segíthetnek nyelvi struktúrák felismerésében.

[2] Ez a más nyelvekre való alkalmazást nem zárja ki, csak a modellezni kívánt nyelvi jelenségnek hasonlónak kell lennie. Pl. török főnevek toldalékolása.

Ennek oka, hogy egyes nyelvi mechanizmusokról csak akkor mondhatjuk, hogy **univerzálisan működnek**, ha **megvizsgáltuk** az **összes általunk ismert nyelvet**, és azokban valóban egyformán működik az univerzálisnak vélt eljárás. Ehhez azonban az analógiás megközelítés esetében még igen kevés adatunk van, ezért célszerűbb az egyes nyelvek leírására vagy lokális univerzálék felfedezésére (pl. hogyan működik az analógia a latin nyelvekben) koncentrálnunk. Fontos azonban látnunk, hogy ez az elvárás szigorúbb azoknál a kritériumoknál, amelyeket a szabályalapú megközelítésekben a mechanizmusok univerzalitásával szemben támasztani szoktak, mivel ezekben a megközelítésekben gyakran az univerzális működés megállapításához néhány nyelv jól kontrollált adatmennyiségének megvizsgálását is elegendőnek szokták venni.

Mint arról már a 3. fejezetben is szót ejtettem, **két szó/karakterlánc összehasonlításában** a nyelvészeti/nyelvtechnológiai kutatásokban leggyakrabban a **Levenshtein-algoritmust** (Levenshtein 1966) használják (pl. az Albright (2009) által is tesztelt GCM), amely számos informatikai alkalmazásban sikeresnek bizonyult[1], azonban a nyelvi viselkedés modellezésében a szóalakok összehasonlítására való alkalmazása számos problémát vet fel. Egyrészt az algoritmus az **összehasonlítást betűk,[2]** nem pedig fonémák[3] alapján végzi. Két betűt ennek megfelelően azonosnak vagy teljesen különbözőnek vesz, így az *o* és az *ó* betűk ugyanannyira különbözőek az algoritmus számára, mint az *o* és a *k* betűk. Másrészt az algoritmus feltételezi, hogy a törlés, beillesztés és megfordítás ugyanakkora változást okoz a szóalakon belül, és ezeknek a beavatkozásoknak a helye is lényegtelen[4]. Korábbi (Rung 2008) és a 6. fejezetbeli vizsgálatom is megmutatja, hogy ez az algoritmus emberi nyelvek szavainak

---

[1] helyesírás-ellenőrzők, optikai karakterfelismerés (optical character recognition, OCR), fordító memóriák stb.

[2] Két betű hasonlósága is hatással lehet két szó hasonlóságának a megítélésére (Wheeler 1887: 32). Így a *török:torok* közelebbi, mint a *török:pörög*, hisz az *ö:o* vizuálisan jobban hasonlít egymásra, mint a *t:p* vagy a *k:g* párok. Ez a hatás azonban kisebb, mint a többi vizsgált tulajdonság.

[3] Ideálisabb esetben hangokról is beszélhetnénk.

[4] A Levenshtein-algoritmusnak több változata van (pl. az ezek közül legismertebb Damerau-Levenshtein távolság), de ezek lényegében nem kezelik a nyelvi modellezéssel kapcsolatos kifogásaimat.

(legalábbis a magyar nyelv szavainak) hatékony és megbízható összehasonlítására nem alkalmazható, így legfeljebb csak kiindulási pont lehet olyan kifinomultabb megközelítések számára, amelyek jobban megragadják a nyelvi rendszer sajátságait és működését.

Mivel az analógiás megközelítés egyes irányzatai szerint (Pierrehumbert 2001, Bybee 2001, 2007, 2010) a szavakat hangalakjukkal is tároljuk, érdemes lehetne a **hasonlóságot fonetikus alapon** (is) számolni. Ettől a lehetőségtől azonban kénytelen voltam eltekinteni, mivel jelenleg a fonetika nem tudja egyértelműen meghatározni, hogy két hang mikor és mennyire hasonlít (különösen a magyar esetében)[1], valamint egy ilyen vizsgálathoz szükséges beszélt nyelvi korpusz, amely dominánsan informális szövegeket tartalmaz vagy az ez alapján készített beszélt nyelvi gyakorisági szótár sem áll rendelkezésre. Technikai korlátaim következtében így maradtam a fonémák összehasonlításánál, ami a betűalapú hasonlításnál jobban közelíti a nyelvi realitásokat.

A vizsgálatomban használt adatokat átalakítottam egy olyan írásrendszerbe, amelyben **egy fonémának egy betű** felel meg, így csökkentettem a magyar írásrendszer (helyesírás) következetlenségeinek negatív hatását. Az átalakított alakokban már a szóbelseji zöngésségi hasonulás eredményeképp létrejövő szekvenciák szerepelnek, amelyeket eredetileg az íráskép nem rögzít (*virágcsokor -> virákčokor*[2]). Más mássalhangzó-szabályokat (Siptár 1994: 242–265), amelyek alapvetően a szóalakok belsejében fejtik ki hatásukat, nem építettem be vizsgálatomba, mivel ezek befolyása nem számottevő a vizsgált jelenségek és az alkalmazott megközelítések szempontjából.

A szavak közti hasonlóság mérésére **python programnyelvben megírt algoritmusokat használok**. Az elsőként bemutatásra kerülő algoritmus (**egyszerű jegymérték** a későbbiekben) egy olyan mátrix alapján végzi számításait, amely megadja, hogy két fonéma mennyire hasonlít egymáshoz. A hasonlóság értéke a 0 és 1 közti skálán helyezkedik el, hasonlóan Albright (2009: 192) gyakorlatához. Így két

[1] A fonetikai hasonlóság mértéke a szóban elfoglalt pozíció, a környezet és a feladat függvényében is változatosságot kell, hogy mutasson.

[2] A morfo-ortográfiával és az írásban jelölt alaki változásokkal kapcsolatban lásd Prószéky (2000) vagy van den Bosch és Daelemans (1993).

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 voltához. A magánhangzók esetében a nyíltságot, ajakkerekítést, hosszúságot, elölképzettséget, a mássalhangzók esetében pedig a zöngésséget, a képzés helyét és módját veszem figyelembe[1]. Minden eltérő jegy esetén az összehasonlított fonémák hasonlóságát 2-vel osztom el (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).

Ebben a megközelítésben a jegyekben való **egyforma mértékű eltérés ugyanakkora távolságot jelent két fonéma között**, így a /p/:/b/ távolság azonos az /ü/:/ö/ távolsággal. Fontos látnunk, hogy a hasonlóságot másképp is definiálhatnám fonémák között. Számolhatnám az egyezéseket is, amely esetben azt mondanám, hogy **azonos számú jegy egyezés esetén a hasonlóság mértéke is azonos**. Ekkor viszont a /p/:/b/ távolság az /i/://ö/ távolságának lenne megfeleltethető, ugyanakkor a legtávolabbi párok mind a két rendszerben egyforma hasonlósággal bírnának (pl. /ü/:/á/ és /k/:/l/), amely az előzőekben bemutatott változatra nem igaz. Egy **harmadik megközelítésben** rögzíthetném, hogy a **legtávolabbi és a legközelebbi párok hasonlósága azonos, és a köztük lévő teret egyenlően osztanám fel**. Ebben az esetben azonban a köztes esetek távolsága más lenne, mint az előző

[1] Ideális esetben a magánhangzós és a mássalhangzós jegyek számának azonosnak kellene lennie a könnyebb összehasonlíthatóság érdekében. Ezt azonban csak úgy érhetném el, ha olyan torzításokat hajtanék végre a jegykiosztásban, ami nincs harmóniában jelenlegi fonológiai ismereteinkkel, hisz ekkor ki kellene találnom még egy mássalhangzós jegyet, vagy össze kellene vonnom egy magánhangzósat.

megközelítésekben, hisz a magánhangzóknál 3 részre, a mássalhangzóknál 2 részre kellene osztanom a hasonlósági teret. Mint láthatjuk, a 3 megközelítés közt való választás jelen ismereteink szerint önkényes. A korábbi jó eredmények (Rung 2008, Rung 2009) hatására az első megközelítést megtartottam, de nem tartom kizártnak, hogy ha ismereteink finomodnak a hasonlítási módokról, akkor ezt érdemes, sőt célszerű lesz lecserélnem egy másfajta súlyozásra.

A magyar nyelvleírás hagyományát és a pszicholingvisztikai kutatási eredményeket (Slobin 1973, Bybee és Moder 1983, Lukács 2002: 47) követve a Levenshtein-algoritmussal ellentétben **nagyobb fontosságot tulajdonítok a szóvégek hasonlóságának**, amelyek kiemeltebb szerepét összetett szavak hasonlításában Krott (2009: 121) is igazolta. Feltételezéseimet a 7. fejezetben bemutatásra kerülő nyelvi teszt is megerősítette. 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), amit a későbbiekben újabb eredmények fényében lehet, sőt minden bizonnyal kell is finomítani (bővebben 7.3. alfejezet). 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 pedig 0[1]. Programom számítása alapján a *bab* és a *púp* hasonlósága a következőképp alakulna:

[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. *bögöly:abbreviatúra*.

(3) b:p = 0,5 (eltérő jegy: zöngésség)
a:ú = 0,25 (eltérő jegy: nyíltság, hosszúság)
b:p = 0,5 (eltérő jegy: zöngésség)

Hasonlóság kiszámítása a logaritmikusságot is figyelembe véve[1]:

(4) ((0,5*1)+(0,25*2)+(0,5*4))/7=0,5

Megközelítésem nagyban hasonlít **Lukács** (2002: 46) elképzeléseihez is, aki még a **jegyek típusának is súlyt ad**, így a képzési hely megváltoztatása jelentősebb beavatkozás lenne elgondolásai alapján, mint a zöngésségé[2]. Modellezésemben ezt megfelelő bizonyítékok hiányában nem követem. Nyelvi tesztelésemben (7. fejezet) is csak érintőlegesen foglalkozok ezzel a kérdéssel, mert úgy vélem, hogy elsőként a pozíciók lehetséges súlyait kell felmérni. A későbbiekben azonban a jegyek fontosságának a megmérése is segíthet a fonémák hasonlóságának az alaposabb megértésében.

Az egyszerű jegymértéknek elkészítettem egy **módosított változatát** is (továbbiakban **komplex jegymérték**), amely a szavakat az előzőhöz hasonló módon, de egy lényegesen finomabb fonémaosztályozást alkalmazva hasonlítja össze, amelyet a 4.1. táblázat[3] mutat be. Az új jegyekből némelyek csak más jegyek bizonyos értékeihez rendelnek további értékeket, önállóan nem értelmezhetőek. Így a *foly* (folyamatos) jegy a *mód* jegy *foly* értékéhez rendeli a *rés* (réshang), *app* (approximáns), *lat* (laterális), *per*

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

[2] Lukács (2002) sem köteleződik el egyértelműen ezen elgondolás mellett, hanem mint megközelítési lehetőséget vázolja fel. Mint a 7.3. alfejezetben látni fogjuk, a zöngésség súlya valóban más, mint a többi jegyé, azonban mindez függ a szón belüli pozíciótól is.

[3] Az algoritmusok által készített számításokhoz egy korábbi változatot használtam, amelytől az itt közölt néhány részletben eltér Siptár Péter és Törkenczy Miklós javaslatai alapján. Ezek a változtatások hozzájárultak a rendszer elméleti pontosságához, de az algoritmusok hatékonyságát alapvetően nem változtatták meg.

(pergőhang) további értékeket. Hasonlóan a *zár* (zárhang) jegy a *mód* jegy *zár* értékeihez rendeli a *felp* (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 rendszernek, miközben különbségeiket se mosom el (/r/ pergőhang, /s/ réshang). Korábbi tesztjeim megmutatták (Rung 2008), hogy egy meglehetősen durva hasonlítási módnál az egyszerű jegymérték finomabb osztályozásának köszönhetően jobban tudtam kezelni az adott nyelvi feladatot[1], így elképzelhetőnek tartottam, hogy egy még árnyaltabb felosztás további eredményjavulást hozhat. Mint a 6. fejezetben látni fogjuk, elvárásaim beigazolódtak, de a javulás csak igen mérsékelt volt.

A komplex jegymértékben **két fonéma összehasonlítását az egyszerű jegymértékéhez hasonló módon végzem el**, de ha egy jegy az egyik fonémára 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.

[1] Nem kizárt azonban, hogy az összehasonlításban egy hibrid megoldás lenne az ideális, amely a széleken érzékenyebb a finomabb különbségekre, míg a szó belsejében kevésbé.

	mgh	elöl	kerek	alsó	felső	hosszú	zör	zöngés	foly	zár	mód	hangk	hely
a	+	-	+	+	-	-							
á	+	-	-	+	-	+							
b	-						+	+		felp	zár	lab	bilab
c	-						+	-		aff	zár	kor	alv
cs	-						+	-		aff	zár	kor	pal
d	-						+	+		felp	zár	kor	alv
dz	-						+	+		aff	zár	kor	alv
dzs	-						+	+		aff	zár	kor	pal
e	+	+	-	+	-	-							
é	+	+	-	-	-	+							
f	-						+	-	rés		foly	lab	labd
g	-						+	+		felp	zár	dor	hátsó
gy	-						+	+		aff	zár	dor	pal
h	-						-	h-állás	app		foly	glott	hátsó
i	+	+	-	-	+	-							
í	+	+	-	-	+	+							
j	-						-	+	app		foly	dor	pal
k	-						+	-		felp	zár	dor	hátsó
l	-						-	+	lat		foly	kor	alv
m	-						-	+			naz	lab	bilab
n	-						-	+			naz	kor	alv
ny	-						-	+			naz	dor	pal
o	+	-	+	-	-	-							
ó	+	-	+	-	-	+							
ö	+	+	+	-	-	-							
ő	+	+	+	-	-	+							
p	-						+	-		felp	zár	lab	bilab
r	-						-	+	per	egyp	foly	kor	alv
s	-						+	-	rés		foly	kor	pal
sz	-						+	-	rés		foly	kor	alv
t	-						+	-		felp	zár	kor	alv
ty	-						+	-		aff	zár	dor	pal
u	+	-	+	-	+	-							
ú	+	-	+	-	+	+							
ü	+	+	+	-	+	-							
ű	+	+	+	-	+	+							
v	-						+	+	rés		foly	lab	labd
z	-						+	+	rés		foly	kor	alv
zs	-						+	+	rés		foly	kor	pal

4.1. táblázat: A magyar fonémák jellemzése jegyekkel. Ha valamelyik fonémára egy jegy nem alkalmazható, akkor annak celláját üresen hagytam. (A rövidítések feloldása a következő oldalon látható.)

(aff = affrikáta, alv = alveoláris, app = approximáns, bilab = bilabiális, dor = dorzális, egyp = egyperdületű, felp = felpattanó, foly = folyamatos, glott = glottális, hangk = hangképző szerv (aktív artikulátor), kor = koronális, lab = labiális, labd = labiodentális, lat = laterális, naz = nazális, pal. = palatális, per = pergőhang, rés = réshang, zár = zárhang, zör = zörejhang)

Mivel Frisch (1996) fonémahasonlítási kísérletei ígéretesnek bizonyultak, így a már bemutatott **természetes osztályokat alapul vevő hasonlítási módot is adaptáltam** a magyar fonémák rendszerére. Ebben a változatban a szóalakokat továbbra is súlyozva hasonlítom össze a már bemutatott módon, míg a fonémákat a természetes osztályaik alapján számított hasonlósági értékeik mentén vetem össze. A természetes osztályok meghatározásához is a 4.1. táblázat jegyeinek értékeit alkalmazom, mint unáris jegyeket. Önálló jegyek a számításban a táblázat alatt felsorolt értékek (pl. felpattanó, labiális stb.), de a bináris jegyeknek a táblázatban önálló névvel nem bíró értékei is (pl. zöngés, zöngétlen stb.).

A 4.2. és a 4.3. táblázatok a magyar magánhangzók és mássalhangzók közös természetes osztályaik alapján számított hasonlóságát adják meg. Az 1-hez közeli értékek nagyobb hasonlóságot jelentenek, a mássalhangzók és a magánhangzók egymáshoz való hasonlósága ebben a rendszerben is 0. Két fonéma hasonlóságának a mértékére nem csak közös/eltérő jegyeiknek a száma van hatással, hanem az is, hogy a teljes fonémaállományban ezek a jegyek hogyan oszlanak el. Így két csak egy jegyben eltérő fonéma hasonlósága nem feltétlenül lesz azonos. Az /i/:/í/ hasonlóságának mértéke 0,35, mert 7 közös és 13 eltérő természetes osztályuk van (0,35=7/(7+13)), míg az /ü/:/ű/ hasonlóságának mértéke 0,33, mivel ezek 8 azonos és 16 különböző természetes osztályban képviseltetik magukat (0,33=8/(8+16)). Hasonlóan ezért tér el a /b/:/p/ 0,38-as hasonlósági mértéke (0,38=8/(8+13)) a /g/:/k/ pár 0,4-es értékétől (0,4=8/(8+12)).

	a	á	e	é	i	í	o	ó	ö	ő	u	ú	ü	ű
a	1													
á	0,24	1												
e	0,22	0,24	1											
é	0,06	0,29	0,27	1										
i	0,09	0,1	0,41	0,24	1									
í	0,04	0,2	0,19	0,57	0,35	1								
o	0,64	0,13	0,13	0,07	0,11	0,05	1							
ó	0,27	0,29	0,06	0,14	0,05	0,1	0,36	1						
ö	0,27	0,06	0,27	0,14	0,24	0,1	0,36	0,14	1					
ő	0,12	0,13	0,12	0,33	0,11	0,22	0,15	0,33	0,33	1				
u	0,41	0,1	0,09	0,05	0,18	0,08	0,54	0,24	0,24	0,11	1			
ú	0,19	0,2	0,04	0,1	0,08	0,17	0,24	0,57	0,1	0,22	0,35	1		
ü	0,17	0,04	0,17	0,09	0,38	0,15	0,21	0,09	0,5	0,2	0,38	0,15	1	
ű	0,08	0,08	0,08	0,2	0,16	0,36	0,1	0,2	0,2	0,5	0,16	0,36	0,33	1

4.2. táblázat: Magyar magánhangzók természetes osztályok alapján számított hasonlósága

	b	c	cs	d	dz	dzs	f	g	gy	h	j	k	l	m	n	ny	p	r	s	sz	t	ty	v	z	zs
b	1																								
c	0,09	1																							
cs	0,08	0,47	1																						
d	0,27	0,26	0,15	1																					
dz	0,18	0,36	0,21	0,64	1																				
dzs	0,17	0,21	0,36	0,33	0,47	1																			
f	0,17	0,1	0,1	0,06	0,06	0,06	1																		
g	0,35	0,09	0,09	0,28	0,18	0,17	0,08	1																	
gy	0,19	0,1	0,18	0,16	0,21	0,41	0,06	0,37	1																
h	0,05	0,04	0,04	0,04	0,04	0,03	0,14	0,11	0,04	1															
j	0,07	0,03	0,06	0,06	0,06	0,11	0,09	0,16	0,29	0,19	1														
k	0,16	0,17	0,16	0,13	0,09	0,08	0,14	0,4	0,17	0,13	0,08	1													
l	0,07	0,1	0,06	0,21	0,19	0,11	0,09	0,08	0,06	0,13	0,17	0,04	1												
m	0,33	0,04	0,03	0,07	0,07	0,06	0,12	0,1	0,07	0,08	0,1	0,05	0,11	1											
n	0,09	0,12	0,07	0,25	0,23	0,13	0,06	0,1	0,07	0,08	0,1	0,05	0,4	0,23	1										
ny	0,08	0,03	0,06	0,07	0,06	0,13	0,05	0,19	0,33	0,07	0,5	0,1	0,1	0,2	0,2	1									
p	0,38	0,17	0,16	0,13	0,09	0,08	0,33	0,17	0,09	0,06	0,04	0,37	0,04	0,17	0,05	0,05	1								
r	0,07	0,1	0,06	0,21	0,19	0,11	0,09	0,08	0,06	0,13	0,17	0,04	0,86	0,11	0,4	0,1	0,04	1							
s	0,06	0,17	0,34	0,11	0,1	0,18	0,24	0,06	0,11	0,09	0,13	0,1	0,14	0,04	0,08	0,07	0,1	0,14	1						
sz	0,06	0,30	0,17	0,18	0,17	0,1	0,26	0,06	0,05	0,1	0,07	0,11	0,24	0,04	0,13	0,04	0,11	0,24	0,46	1					
t	0,13	0,6	0,31	0,37	0,26	0,15	0,11	0,13	0,08	0,04	0,03	0,28	0,1	0,04	0,13	0,04	0,28	0,1	0,18	0,32	1				
ty	0,1	0,22	0,41	0,08	0,11	0,19	0,12	0,18	0,38	0,05	0,14	0,35	0,03	0,04	0,04	0,17	0,19	0,03	0,19	0,09	0,16	1			
v	0,36	0,06	0,06	0,13	0,12	0,11	0,39	0,16	0,13	0,12	0,17	0,08	0,17	0,22	0,1	0,09	0,17	0,17	0,13	0,14	0,06	0,07	1		
z	0,11	0,15	0,09	0,35	0,33	0,19	0,13	0,11	0,1	0,07	0,12	0,06	0,48	0,07	0,23	0,06	0,06	0,48	0,22	0,4	0,16	0,05	0,27	1	
zs	0,11	0,09	0,17	0,2	0,19	0,37	0,12	0,11	0,2	0,07	0,25	0,05	0,26	0,06	0,13	0,13	0,05	0,26	0,39	0,22	0,1	0,1	0,25	0,47	1

4.3. táblázat: Magyar mássalhangzók természetes osztályok alapján számított hasonlósága

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 közti 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 össze[1]. Az algoritmus ugyanazokkal a jegyekkel 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 egymástól teljesen önállóan léteznek a reprezentációban[2].

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

[2] Az autoszegmentális fonológiai megközelítésnek vannak az eredeti elképzeléshez sokkal közelebbi megvalósításai is (Kornai 1995, Wiebe 1992, Bird és Ellison 1994).

5. Hangkivető főnevek hasonlósági viszonyai

5.1. Források és adatok

Mivel kutatásomban a **valós nyelvi folyamatok és viszonyok feltárása** volt a cél, ezért nemcsak elméletileg, hanem módszertanilag is törekedtem az eltávolodásra a nyelvi leírásban a mai napig gyakori, a nem kontrollált introspekción alapuló, alapvetően spekulatív elemzési hagyománytól[1]. E megközelítés követői gyakran saját maguk által kitalált fiktív példákról hoznak döntéseket intuíciójuk alapján, majd ezekből alkotnak meg elméleteket[2]. Az introspekció a megismerésnek és a kreatív gondolkodási folyamatoknak kiváló kiindulópontja lehet (Evans 1982: 68), de egy nyelvész megfigyeléseit, elgondolásait csak akkor vehetjük komolyan, ha azokat adatokkal vagy kísérletekkel megfelelő módon és minőségben alá tudja támasztani. Habár Kac már 1974-ben bírálta a nyelvészeti gyakorlatban a pszichológiai kutatásban is elvárt és bevált mérési módszerek hiányát, a kérdésben még nemzetközi szinten sem történt igazi és egyértelmű áttörés. Mivel a nem kontrollált introspekcióra vagy környezetünk esetleges megfigyelésére nem lehet megbízható leírást alapozni, ezért az elemzésben nyelvi korpuszra hagyatkozom, amely a kollektívnek is tekinthető nyelvi emlékezetet helyettesíti. Ez a megközelítés egyre inkább elfogadott a modern nyelvi elemzésben (Mukherjee 2007: 141); ennek fontosságát Harris már 1949-ben is hangsúlyozta. A korpuszalapú megközelítésnek a 3.1. alfejezetben kifejtett hátrányairól azonban nem szabad megfeledkeznünk.

Korpuszalapú vizsgálataimban a ***Szószablya Korpusz*** (Halácsy és mtsai 2003) alapján számított **gyakorisági adatokat** használtam fel. Azért döntöttem mellette, mert jelenleg ez a legnagyobb, mintegy 19,1 millió szóalakot tartalmazó szöveges korpusz, amely 3,49 millió magyar weboldal és 1,49 milliárd szövegszó alapján készült 2003-ban.

---

[1] Hasonlóan problematikusak azok az elemzések is, amelyek megbízhatatlan, esetleges, vagy másodkézből való forrásokra alapítanak elméleteket.

[2] Igaz, ez a megközelítés a mondattanban gyakoribb, mint a morfológiában.

Így nyolcszor annyi adatot tartalmaz, mint az ezt méretben követő 187,6 millió szavas *Magyar Nemzeti Szövegtár* („MNSZ" 2006). Az *MNSZ* mintegy 90%-a szerkesztett szövegeket (sajtó, szépirodalom, tudományos és hivatalos szövegek) tartalmaz, amelyek egy része még a 20. században keletkezett, illetve nem a magyar köznyelvet reprezentálja, hanem annak valamely határon túli változatát (22,9 millió szövegszó). Ezek alapján 17,8 millió szövegszóból áll az a része, amely alkalmas lenne a magyar köznyelv változásának és az ingadozási jelenségeknek a tanulmányozására. A vizsgálni kívánt nyelvi jelenségek szórványos jellege miatt cél volt azonban, hogy minél több nyelvi adat alapján következtessek, és hogy inkább a hétköznapi, válogatatlan, sokszor a beszélt nyelvhez közelibb változatokat és írásmódot rögzítő szövegek (fórumok, blogok) kerüljenek előtérbe, amelyekben a változás nyomai jobban tetten érthetők. A *Szószablya Korpusz* erre a vizsgálatra sokkal alkalmasabbnak tűnt, bár – mint a későbbiekben látni fogjuk –, hatalmas méretei ellenére egyes specifikus jelenségek elemzéséhez még ez is kicsinek bizonyult.

A *Szószablya Korpusz* alapján készült *Szószablya Gyakorisági Szótár* **tövek helyett szóalakokat, elnagyolt gyakorisági kategóriák helyett pedig pontos gyakorisági számokat** tartalmaz, ami még alkalmasabbá teszi a nyelvészeti és nyelvtechnológiai kutatásokban való felhasználásra. A *Szószablya Gyakorisági Szótár* egy szóalakhoz 4 gyakorisági számot ad meg, aszerint, hogy mennyire jól követi a magyar helyesírás szabályait a szavak szintjén a szöveg (weboldal), ahonnan származik. Az osztályozásban a szónál (szóköztől szóközig terjedő egység) magasabb szintű helyesírási szabályokat nem vették figyelembe. E felosztás alapján a jó helyesírású (legszerkesztettebb) szövegek 589 millió szövegszót tartalmaznak, amelyekhez hozzáférhetőek a *hunmorph* (Trón és mtsai 2006) program által készített morfológiai elemzések is. Ezeknek az elemzéseknek a használatától, mint ahogy minden más hozzáférhető nyelvtechnológiai eszköz kimenetétől (a korpuszt természetesen leszámítva) elvi megfontolásokból tartózkodok. Kizárólag a *Szószablya Gyakorisági Szótár* által tartalmazott felszíni alakokra és azok gyakorisági számaira támaszkodok, mivel a géppel készített leírások/elemzések vizsgálatomhoz nem elég pontosak, kategorikus jellegüknél fogva pedig a számomra értékes nyelvi adatokat szűrik ki.

Vizsgálatomban a nagyobb**, 1,486 milliárd szövegszóra vonatkozó gyakorisági számokat használom**, mivel a terjedelmesebb gyűjtés több a beszélt nyelvhez közeli alakot tartalmaz, amelyeknek szerzői a magyar helyesírást mint kötelező érvényű normát kevésbé tartották fontosnak. Ez a megközelítés azzal a hátrányos következménnyel jár, hogy a vizsgálatomban használt adatok egy nem jelentős, de figyelmen kívül nem hagyható részét ékezetmentes szövegek adják (*örömön* helyett *oromon, kapóra* helyett *kapora, béreket* helyett *bereket* stb.), amelyek kiszűrése időnként gondot jelenthet az elemzésben. Úgy vélem azonban, hogy ez a nehézség kezelhető, mert az elemzés során azonosíthatók az ilyen zavaró alakok. Ugyanakkor így az adatoknak a 60%-át nem kell elvetnem, amelyekben a vizsgálni kívánt alakok a legnagyobb arányban fordulhatnak elő. Bizonyos esetekben a furcsa, meglepő viselkedés mögött egyszerűen adathiány, elgépelés vagy ékezet nélküli írásmód áll, ezért az ilyen adathiányból eredő torzulások kiszűrése érdekében bizonyos részelemzésekben csak egy küszöbérték[1] feletti adatokat veszek figyelembe.

Fontos a *Szószablya Gyakorisági Szótár* használatakor még néhány zavaró tényezőt, korlátot figyelembe vennünk. A **szövegek magyar voltának azonosítása** gépi úton történt, így még akár a jobb helyesírású szövegekben is maradhattak más nyelvekből való szóalakok, amelyek a gyakorisági adatokat is befolyásolták[2]. A nem magyar nyelvű szóalakok kiszűrése alapvetően könnyű feladat, de ha egy idegen nyelvű szóalak azonos egy magyar szóalakkal, akkor azok gyakorisága összeadódik (pl. angol *must* 'kell' v. magyar *must* 'szőlőital'), így az ilyen adatokkal óvatosan kell bánni. Ezek szétválasztása gépileg nem lehetséges a *Szószablya Gyakorisági Szótár* alapján.

Habár a magyar web már 2003-ban is elég jól és egyenletesen reprezentálta a közbeszéd összes témáját, és ezáltal azok szókincsét, az **informatikához és az internethez kötődő szavak**, mint *rendszergazda* (a korpuszban hasonló gyakoriságú,

---

[1] Akármilyen módon határozom meg ezt a küszöbértéket, mindenképpen önkényes lesz. Ha nagyobb adatmintákat vizsgálok, akkor ezt 100 előforduláshoz kötöm, ha kisebbeket, akkor 25-höz, de minden esetben külön jelzem a küszöbérték mértékét használatakor.

[2] A teljes korpuszban pl. a *company* 'társaság' szó 36983-szor fordul elő, közel annyiszor, mint a *tükör* (35387).

mint az *iker*) vagy *kernel* (a korpuszban hasonló gyakoriságú, mint a *titok*) az átlagos hétköznapi szövegek arányaihoz képest jelentősen **felülreprezentáltak**.

A *Szószablya Korpusz* közel áll a beszélt nyelvhez, de **nem beszélt nyelvi korpusz**. A szövegek nem beszédben elhangzott szövegek lejegyzései, ezért pontosan nem is tükrözik annak sajátságait, ugyanakkor számolnunk kell csak az írott szövegekre jellemző, számunkra érdektelen problémákkal (elgépelés[1], ékezettelen írás stb.) is. Kutatásom céljaival összhangban egy nagyméretű lejegyzett beszélt korpusz alapján való vizsgálódás állna, de erre sem most, sem vélhetőleg a közeljövőben nem lesz mód, főleg, ha azt is figyelembe vesszük, hogy a méreteiben hatalmas *Szószablya Korpusz* is kicsiny minden magyar nyelvi jelenség részletes tanulmányozásához[2].

A **hasonlósági kapcsolatok** tanulmányozására azért választottam a **hangkivető főneveket**, mert bár zárt osztályt[3] alkotnak, meglehetősen számosak, így a viselkedésükből levonható következtetések kevésbé szórványos és egyedi adatokon fognak alapulni. A hangkivető főnevek viselkedésére jellemző a fokozatos ingadozás, amelyről még a szabályalapú megközelítések engedékenyebb változatai sem tudnak teljesen számot adni, hisz a hangkivető szavak érthetetlen és önkényes változására adott feleletük legfeljebb a szótár folyamatos újraírása lehet, vagy az, hogy ezeket az

---

[1] Az elgépelésből fakadó hibás adatok aránya jóval kisebb, mint gondolnánk, ami az olyan alakokat megvizsgálva tűnik ki, ahol nyelvi ingadozásról nem lehet szó: pl. *elnököt* (9442 előfordulás), *elnökt* (2 előfordulás), *elnköt* (0 előfordulás). A 100 leggyakoribb *-C(e/ö/o)C* végű főnév tárgyesetének vonatkozásában az elgépelés aránya 0,7‰, a szórás 0,0019-es értéke szintén elenyésző, ha 1-nek vesszük azt az esetet, amikor egy alakot mindig elgépelés nélkül írnak le. Látszólagos 1% körüli elgépelést csak a kevés tárgyas alakkal rendelkező *major, görög* szavaknál találunk, illetve a *motor*-nál, amely azonban enyhén a hangkivetők felé húz, így esetében nem elgépelésekről van szó.

[2] Nehéz megállapítani, hogy a *Szószablya Korpusz* mérete hogyan viszonyul az elhangzó, de le nem jegyzett szövegek tömegéhez. Ha meglehetősen visszafogottan azzal számolunk, hogy egy magyar ember napi 2000 szót ejt ki és 70 éven át folytat beszédtevékenységet, akkor életében 50 millió szó hagyja el a száját. Ezzel a becsléssel számolva is a *Szószablya Korpusz* legfeljebb 30 ember teljes beszédtevékenységével felérő szöveget tartalmaz.

[3] A teljes zártság vitatható, de ha jönnek is létre újabb hangkivető szavak, vagy sorolódnak ebbe a paradigmába, azok száma elenyésző lehet és elsősorban a bloomfieldi kontamináció alá sorolhatók, mint azt néhány későbbi példából látni fogjuk.

egyébként a nyelvben hétköznapinak mondható jelenségeket marginálisnak vagy a performancia hatókörébe tartozó jelenségeknek veszik.

A **hangkivető szavakról** generatív vagy hagyományos nyelvészeti keretrendszerekben több **leírás** készült (Papp 1975, Vago 1980, Elekfi 1994, Törkenczy 1994 stb.), de ezek – szabályalapú elméleti megközelítésük jellege miatt – a hangkivető szavak több viselkedési sajátosságát nem tudták megragadni. Törkenczy és Siptár (2000), valamint Rebrus (2000) ezeknél részletesebb elemzései már jó viszonyítási alapot képeznek vizsgálódásaimhoz, de elsősorban Rebrus és Törkenczy (2008) munkájára fogok támaszkodni, akik az enyémhez hasonló elméleti keretben 216 hangkivető főnév[1] rendszerét vizsgálták meg. Kutatásom több esetben csak megerősíti megállapításaikat, de számos új eredményt is hozott ezekhez képest a szövegkorpusz alapú megközelítésnek, az új módszereknek, valamint annak köszönhetően, hogy a hasonlóság fogalmát nem korlátozom a teljes szekvenciaazonosság eseteire[2].

Vizsgálatom a belső felépítés tanulmányozása helyett a hangkivető szavak **egymás közti viszonyainak és a többi főnévtől való különbségüknek** az alaposabb megismerésére irányul. A hangkivető főnevek belső és a többi főnevekhez viszonyított külső viszonyainak jellegéről sokat elárul az egyes hangkivető szavak hangkivetésének mértéke és ennek a mértéknek az egyes toldalékaik közt való megoszlása. Amennyiben azt feltételezzük, hogy a szavak hangkivető volta nem véletlenszerűen alakult ki és maradt fenn, akkor szükségszerűen ezeknek a szavaknak valamilyen szempontból hasonló jegyekkel kell rendelkezniük[3]. Ez a hasonlóság kézzelfoghatóvá is válik, amikor

---

[1] Vizsgálati anyagomban összesen 229 nem összetett tő található, ami azt jelenti, hogy anyaguk lényegében ugyanazokon a forrásokon alapszik, mint az én vizsgálatom, így esetleges apróbb eltéréseikre nem érdemes kitérni, mert minden bizonnyal kevéssé befolyásolják a következtetések érvényességét.

[2] Az ettől való eltérés elméletbeli lehetőségét ők is elismerik, de leírásukban csak egy példa kapcsán foglalkoznak ezzel.

[3] A közös sorsban természetesen sok más tényező is szerepet kaphat a hasonlóságon túl pl. a közös használati jelleg, hasonló forrás, azonos korban kerültek nyelvünkbe stb.

a hangkivető szavak egyes nyelvváltozatokban bevonzanak körükbe hozzájuk nagyon közeli szavakat (pl. *motor*, *bútor*[1]).

A vizsgálatra **1211 hangkivető főnevet választottam ki a BME MOKK *morphdb.hu* szótárából** (a szavak listája az A Függelékben látható), amely jelenleg a legnagyobb ingyenesen is hozzáférhető nyelvi adatbázis (130 ezer szó, Trón és mtsai 2006). Az 1211 szó összesen 229 szóból és az azokból létrehozott összetételekből áll. Az összetételek jobb oldali tagjaként 129 szó szerepel. Vizsgálataimat a Rebrus és Törkenczy (2008) által meghatározott $\sim VC_{\alpha}(o/e/\ddot{o})C_{\beta}\#$[2] mintán túl az utolsó magánhangzóként *-a*-t és *-u*-t tartalmazó szavakra is kiterjesztem, mint pl. *ajak*, *bajusz*, *vacak* stb. (17 szó). Ezeket ők kizárták az elemzésből, mivel „nagyon gyenge független (egyedi) mintázatokat képviselnek a lexikonban". Döntésüket azonban nem indokolják megfelelő részletességgel[3], ezért e szavakat meghagytam válogatásomban, mivel a hangkivetés elvárásának legalábbis részlegesen megfelelnek. A séma többi általánosítását[4], miszerint a hangkivetésben kieső magánhangzót *VC* szekvencia előzi meg, és a hangkivető magánhangzót keretező mássalhangzók nem lehetnek azonosak vagy egymás mellé kerülve nem válthatnak ki hasonulást és összeolvadást, én is kivétel nélkül érvényesnek találtam. Adataim közt találhatók a sémára illeszkedő hangátvetéses és a tőbelseji magánhangzó-rövidülésben résztvevő alakok is, amelyek egyben hangkivetőként is viselkednek (*-kehely*, *-pehely*, *-teher*, *-boholy*, *-lélek* végűek). A főnévi hangkivetőkhöz több mindenben hasonlóan viselkedő igei és melléknévi alakokat (pl. *termett, ugrott, bátrak* stb.) azért hagytam ki, mivel egy homogénebb csoportban az adatok sokfélesége ellenére is nagyobb esélyünk van nem feltűnő, de

---

[1] A *bútor* szó érdekessége, hogy szóhasadással jött létre és etimológiai párja, a *bugyor* hangkivető. Az ingadozásért azonban biztos, hogy elsősorban nem ez a ritka alak a felelős, mivel a két szó között a magyar beszélők nem láthatnak semmiféle eredetbeli kapcsolatot.

[2] C_{α} és C_{β} nem azonos, és egymás mellé kerülve nem válthatnak ki hasonulást, illetve nem olvadhatnak össze. Legfeljebb egyikük lehet affrikáta vagy réshang.

[3] Mint a későbbiekben látni fogjuk, abban igazuk van, hogy az utolsó magánhangzóként *-a-t* vagy *-u-t* tartalmazó hangkivető szavak erősen ingadoznak egyediségükből kifolyólag, de ennek a kritériumnak a mentén más, viszonylag egyedi részmintázatokat is kizárhatnék (pl. réshang végűek).

[4] A későbbiekben ezt némileg pontosítom.

egyébként jelentős összefüggések megfigyelésére. Igaz, így bizonyos általánosabb viszonyokat viszont nem tudunk felismerni, mint pl. a főnevek közt alakilag magányos *dolog* kapcsolatait a 98 *-log, -rog* végű hangkivető igével.

A ***morphdb.hu*** **szótárában** eredetileg összesen **1097 szó** volt **hangkivetőként megjelölve**, amelyekből kivettem a *kelet, sportberkek, sodor, terem* szavakat, így 1093 szavam maradt. A *kelet* szó egyértelműen a *kelte* szóalak miatt került be hangkivetőként rögzítve, ahelyett, hogy már ragozott főnévként vették volna fel a szótárba. A *sportberkek* már szóalak, helyesen a szótárban *sportberek*-ként kellene szerepelnie, amelynek hiányos a paradigmája. A *sodor* és a *terem* szavak valóban hangkivető főnevek, de mivel alanyesetük és számos további ragozott alakjuk egybeesik a náluk gyakoribb *sodor* és *terem*[1] igék alakjaival, ezért célszerűbbnek tartottam ezek kihagyását a vizsgálatból (hasonlóan Wulf 2002: 116). Úgy véltem, hogy elegendő nyelvi adat birtokában ezek elhagyása nem vezet az eredmények jelentős torzulásához.

A szótárban hangkivetőként megadott szavakon túl **további 118 szót** választottam ki, amelyek hangkivetők, de nincsenek hangkivetőként megjelölve. Ezek a szavak a hangkivetőként megjelölt szavakból létrehozott összetett szavak, amelyekben a hangkivető tő az összetétel jobb oldali tagját adja. A morphdb.hu a hangkivető főnevekből csak 4 szót ad meg helyesen ingadozónak, azaz olyannak, amelynek hangkivetési mértéke nem 100% (hangkivetés a leggyakoribb hangkivetéssel együtt járó toldalékok esetében a *Szószablya Gyakorisági Szótár* alapján: *bajusz*: 36%, *fókabajusz*: nincs adat, *harcsabajusz*: 58%, *macskabajusz*: 25%). A többi ingadozó hangkivető főnév esetében csak a kategorikusan hangkivető viselkedést jelöli.

Az egyes szavak hangkivetésének mértékét az **összes olyan inflexiós toldalék[2] előtt** megvizsgáltam, amelyek hangkivetéssel járnak együtt: tárgyeset, szuperesszívusz, többes szám, birtokos személyragok. A többes szám és a birtokos személyragok esetén foglalkozok olyan alakokkal is, ahol ezek a toldalékok nem a szó legszélén találhatók

[1] A *terem* esetében vitatható, hogy az igei vagy a főnévi tő a gyakoribb, de mivel ez eldönthetetlen, és az ige nagy gyakoriságú, a kihagyás mellett döntöttem.

[2] Vizsgálatomban a magyar nyelvleírásban ragként és jelként elnevezett toldalékok között nem teszek különbséget, mivel munkám szempontjából ennek sem gyakorlati, sem elméleti jelentősége nincs.

(pl. *szerelmeimmel).* Az említett toldalékokon túl a hangkivetéses tövek képzett szavakban is megtalálhatók nagy számban, ami szintén fontos szerepet játszik abban, hogy az analógiás változással szemben ellenállóbbak: *-Vs* (*tornyos*), *-Vsodik* (*fodrosodik)*, *-Vskodik* (*torkoskodik*), *-atlan* (*álmatlan*), *-Vl* (*bütyköl*), *-Vz* (*cukroz*), *-Vnként* (*sarkonként*), *-(j)ú* (*tornyú*), *-i* (*szerelmi*), *-Vcska* (*ökröcske*), *-stul* (*fátylastul*), *-ár* (*irodalmár*), *-ász* (*horgász*), *-ista* (*forgalmista*), *-ít* (*sarkít*), *-inca* (*farkinca*), *-Vd(ik)* (*álmodik*), *-Aszt* (*horgaszt*), *-All* (*sarkall*). Ezek egyedi viselkedésével azonban nem foglalkozok.

A *Szószablya Gyakoriségi Szótár* alapján megállapítható, hogy a *pityer*, *szlalom*[1], *vicikvacak* szavak már nem hangkivetők, míg további **116 szó tekinthető kevésbé hangkivetőnek**, mert ezeknél az esetek legalább 1%-ában[2] a hangkivetéssel együttjáró toldalékok előtt nincs hangkivetés. 42 szónál ez az arány meghaladja a 10%-ot, 14-nél pedig több, mint 50%[3]. A hangkivető szavak – képzett alakjaikat is figyelembe véve – 49,7%-ban hangkivetéses alakjukban szerepelnek. A szavak **összes alakján** (képzettek is) belül a **hangkivetéses alakok aránya** és a hangkivetést elváró **toldalékos alakjaikban mérhető hangkivetés mértéke** kevésbé szoros**,** de szignifikáns **összefüggést mutat** ($r\ (1078) = 0{,}19$, $t = 6{,}38$, $p < 0{,}001$), ami azt jelenti, hogy az egyes szavak hangkivetésének mértéke összefügg azzal, hogy összes alakjukban a hangkivetéses formák milyen arányban képviseltetik magukat. Ezekben az esetekben nem csak az analógiás kiterjesztés, hanem az analógiás kiegyenlítődés hatásaival is számolnunk kell (bővebben 2.1. alfejezet). A **hangkivetés átlagos mértéke 97,57%,** hangkivető főneveknek leginkább szuperesszívuszos alakjaik viselkednek nem hangkivető módon (hangkivetés mértéke: 95,3%).

---

[1] A *szlalom* hangkivető volta eleve kétséges, ezért a későbbi vizsgálatokból ki is hagytam, de a legerősebben ingadozó szavak vizsgálatánál röviden kitérek rá, mert *Google* lekérdezések alapján szórványosan lehet hangkivetéses alakjaival is találkozni: *szlalmot, szlalmon.*

[2] A szavak hangkivetési mértékének és gyakoriságának a számítása során figyelmen kívül hagytam az olyan alakokat, amelyek látszólag egy hangkivető főnév nem hangkivetéses alakjai, de valójában egy másik szóhoz tartoznak (pl. *karomat* : *kar*+POSS.E.1+ACC, *körömet* : *kör*+POSS.E.1+ACC, *kéreget 'koldul'* stb.).

[3] Ezen adatok alapján láthatjuk, hogy a szabályalapú elgondolásokat követő *morphdb.hu* pontatlan, és olyan tövekről állítja kategorikusan, hogy hangkivetők, amelyek már ingadoznak.

Elemzéseimben (5.2. és 5.3. alfejezetek) a szóalakok mögött zárójelben szereplő számok azt mutatják, hogy az adott alak hányszor fordul elő a *Szószablya Korpusz*ban. A szavakra való hivatkozáskor azok zéró morfémás, ún. szótári alakját használom, az ilyenkor mögöttük szereplő szám a hangkivetést elváró toldalékos alakjaik összesített gyakoriságát adja meg. A %-jeles adatok minden esetben arra utalnak, hogy az adott szó, szócsoport hány százalékban mutat a releváns toldalékos alakokban hangkivető viselkedést. Összehasonlító szándékkal hasonló módon már ezekben az alfejezetekben is hivatkozok a *Google Gyakoriságí Gyűjtés* adataira. Amennyiben nem egyértelmű a szövegkörnyezetből, hogy az adatok honnan valók, akkor azt külön jelzem. Az elemzésben abból a feltételezésből indulok ki, hogy egy szó minél ritkább, annál hajlamosabb az analógiás változásra (Kraska-Szlenk 2007, Bybee 2010), de ezt hasonlósági viszonyai akár jelentősen is módosíthatják. Az elemzés során elsősorban azokra a szavakra irányul a figyelmem, amelyek ennek látszólagosan ellentmondanak.

Mivel megközelítésem és **elemzéseim** a **használaton alapulnak,** fontos hangsúlyozni, hogy a magyarázatban sok minden szerepet játszhat, ami számot tud adni a szavak viselkedéséről és szerveződéséről (Bybee 2010: 10). Adataim nagyfokú variabilitása miatt vizsgálataimban sokszor egyedi adatokra hivatkozok, mivel a nyelvtudomány fejlődése szempontjából még mindig előnyösebb észrevenni és csak valószínűsíthető magyarázattal értelmezni ezeket, mint tehetetlenségünket beismerve teljesen kizárni őket az elemzésből. Mindazonáltal **legerősebb bizonyító erejűnek a kézzelfogható és mérhető adatokat tartom**, így a használat számszerűsíthető bizonyítékaira (gyakoriság, egyes toldalékok használati aránya) rendszeresebben fogok támaszkodni, mint például a jelentésre, amelyről jóval áttételesebbek az ismereteink, habár a nyelvhasználatban minden bizonnyal komoly szerepe van.

5.2. Hangkivető szavak jellemzése végük alapján

5.2.1. Általános jellemzők

A hangkivető szavakat elsősorban végük alapján vizsgálom meg, mivel szuffixumok esetében a szavak jobb oldala van a **legnagyobb hatással** arra, hogy mely szavakkal **viselkednek morfofonológiailag** azonos vagy hasonló módon (Bybee és Moder 1983, Chandler 2002, Rebrus és Törkenczy 2008). Ez megegyezik a legtöbb nemzetközi és magyar nyelvészeti elmélet elképzeléseivel, ezért külön ennek bizonyítására és indoklására nem térek ki írásomban[1].

A hangkivető szavakat a többi, **nem hangkivető főnévhez hasonlítottam** egyedi viselkedésük jobb megértése érdekében, bár mérsékelten az analógiás folyamatokban más szófajú elemek is szerephez juthatnak. Így például az *-el*, *-ol*, *-ez*, *-oz* végű kevésbé hangkivető főnevekre a nem hangkivető *-z* és *-l* képzős igék tömege hathat, főleg, ha az alaki egyezés teljes (pl. *tegez* főnév vagy *tegez* ige). Az összes 1211 hangkivető főnévből 1074-et vetettem össze 48464 nem hangkivető főnévvel, mert a *Szószablya Gyakorisági Szótár*ban csak ezekhez voltak a nem hangkivető főnevekkel való összehasonlítást lehetővé tevő egyes szám alanyesetű alakokhoz tartozó gyakorisági adatok. A nem hangkivető főnevek példánygyakoriságukat figyelembe véve is 50-szer gyakrabban fordulnak elő, mint a hangkivetők[2].

A **szóvégek gyakoriságát** a szavak egyes szám alanyesetű alakjainak gyakorisága alapján számoltam, mivel a nem hangkivető főneveknek csak ezen alakjairól vannak megbízható gyakorisági számaim. A hangkivető főnevek

---

[1] Abban azonban már jelentős eltérések vannak, hogy a végek meddig és milyen módon számítanak a viselkedéstípusok leírásában.

[2] Némileg meglepő, hogy a hangkivető szavak nem gyakoribbak, mint a többi főnév, sőt még akkor sem szerepelnek nagyobb arányban, ha csak azokat a szavakat vesszük figyelembe mind a két csoportban, amelyek egyes szám alanyesetű alakjuk viszonylatában legalább 1000 előfordulással bírnak. Ezek az arányok azt mutatják meg, hogy a hangkivető főneveknél az egyedi viselkedést még támogatja a nagyon erős csoporthatás, ami a kevesebb szót számláló csoportokra (pl. *v-vel* bővülő főnevek) már nem igaz, ezért ott a rendhagyó viselkedés megőrzéséhez szükségszerű a magasabb gyakoriság.

gyakoriságával kapcsolatos megállapításaimat azonban az összes a hangkivetés szempontjából releváns toldalékos alakjuk alapján teszem, hisz a hangkivetési mértékükkel kapcsolatos adatokat is ezek alapján kaptam meg. A hangkivető főnevek esetében az egyes szám alanyesetű alakok és a hangkivetés szempontjából releváns toldalékos alakok gyakorisága erősen korrelál ($r(1029) = 0{,}8$, $t = 43$, $p < 0{,}001$)[1], így ennek függvényében alkalmazhatom ezeket a számokat együttesen a hangkivető főnevek elemzésében. Ezek alapján nagyon valószínű, ha egy tővég gyakori a hangkivetők körében az egyes szám alanyesetű alakok alapján, akkor gyakori lenne az összes hangkivetés szempontjából releváns toldalékos alakjuk alapján való számítás esetén is.

A **végeket** egy python nyelven írt program segítségével **hasonlítottam össze**, amelynek eredményét egy **módosított trie** (Knuth 1997) struktúrában tároltam el. Az eredeti eljárást kiegészítettem azzal, hogy az egyes csomópontok alatt eltároltam annak a részszekvenciának az előfordulási gyakoriságát, amelyet egy kiindulási ponttól[2] az adott csomópontig összeolvasva kapnánk, így az 5.1. ábrán látható *ty* melletti 2231 arra utal, hogy a *-tyol* részszekvencia ennyiszer fordult elő a hangkivető szavakban. Egy-egy új szó esetében a szó végétől vett részszekvenciák csomópontjaihoz tartozó számokat mindig annyival növeltem, amekkora az adott szó gyakorisága volt. Így a *fátyol* szó esetében 2047-tel (ez az egyes szám alanyesetű alak gyakorisága) növeltem a *fátyol*, a *-átyol*, a *-tyol*, *-ol*, *-l* részszekvenciákért felelős csomópontoknál lévő számokat. Ezt a számítást végrehajtottam kizárólagosan a hangkivető tövek magánhangzó- és mássalhangzó-szekvenciáit is figyelembe véve. Ez utóbbit részben[3] Rebrus és Törkenczy (2008) is használta elemzésében, de eredményeiket nem súlyozták a példánygyakorisággal. Ettől függetlenül végső következtetéseik sokban hasonlítanak az enyémekhez, miszerint a tipikus hangkivető mintázatoktól való eltérés lehet az egyik fő

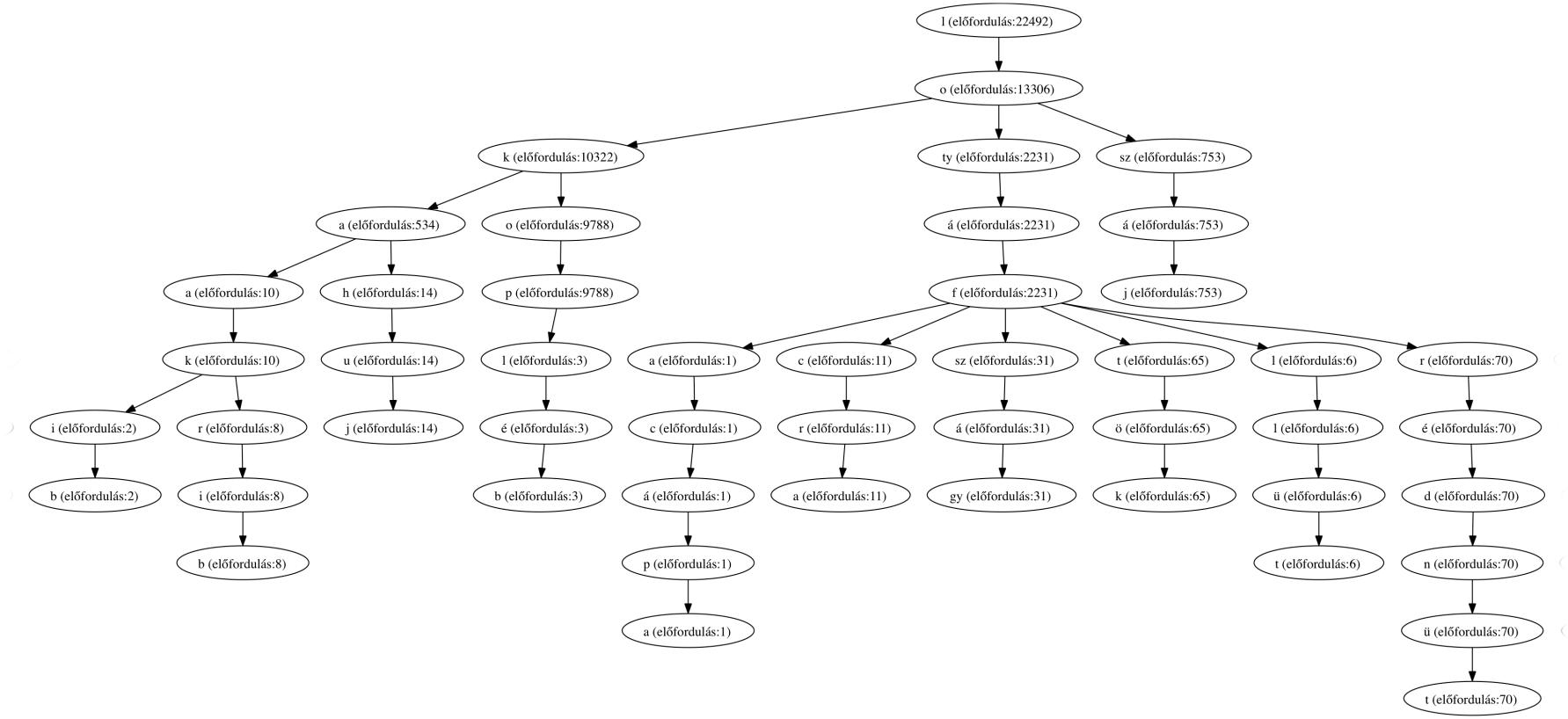
[1] A hangkivető szavak egyes szám alanyesetű alakjainak és az összes előforduló alakjainak gyakorisága közt is igen magas a korreláció ($r(1202) = 0{,}76$, $t = 40{,}3$, $p < 0{,}001$).

[2] A kiindulási pontok a szavakban lehetséges utolsó fonémák voltak, így például az /ö/ nem kiindulási pont, mert nem szerepel a *moprhdb.hu* egyetlen főnevének sem a végén.

[3] Elemzésüket szigorúan az utolsó két mássalhangzóra korlátozták.

oka ezen szavak változásának. Néhány szó esetén (*bajusz, köböl, fátyol, vödör, öböl*) a hangkivetési mérték csökkenésének az oka az is lehet, hogy a szavak utolsó magánhangzóját egyes beszélők félhosszúnak ejtik, és ennek következtében már nem felelnek meg eléggé a minta elvárásainak. Erről tanúskodnak e szavak rossz helyesírású alakjai: *bajúsz* (13%)[1], *fátyól* (0,3%), *vödőr* (0,3%), *öből* (0,2%), amelyek aránya a jó helyesírású változatokhoz képest képest a *bajusz*-nál kiugróan magas, de mindegyiknél legalább 0,2%, ami meghaladja a más, hasonló szavaknál mérhető elgépelési arányt.

[1] A következő zárójelekben lévő százalékos értékek a hosszú ékezetes változatok arányát adják meg az egyes szám alanyesetű alakokon belül.



5.1. ábra: *-ol* végű hangkivető szavak végszekvenciái és előfordulásaik száma

Az 5.1. táblázat alapján láthatjuk, hogy a hangkivető főnevekre az ***-m*, és részben a *-k*, *-g* végek a jellemzőek**. Ez összhangban van Rebrus és Törkenczy (2008) megfigyeléseivel, de ők a *-g*-t nem emelték ki, mivel adataikat nem súlyozták a szavak gyakoriságával, így a *dolog*, a leggyakoribb hangkivető főnév hatása nem érvényesült ilyen erősen elemzésükben[1]. Az összes többi zárófonéma nem jellemző a hangkivetőkre. Egy vég tipikussága a nem hangkivető főnevekhez viszonyítva (hányszor gyakoribb a hangkivetőknél) erős pozitív korrelációban van példánygyakoriságával ($r(9) = 0{,}98$, $t = 15{,}95$, $p < 0{,}001$)[2] és típusgyakoriságával is ($r(9) = 0{,}93$, $t = 7{,}54$, $p < 0{,}001$). Ez alapján azt láthatjuk, hogy a legtöbb alakot és tövet számláló csoportok nem annyira jellemzőek a többi főnévre. Az *-m* és részben a *-k*, *-g* végű szavaknál is a beszélőknek lehetnek olyan tendencia jellegű elvárásaik, hogy azok hangkivető módon fognak viselkedni. Ez az összefüggés alól csak az *-l* és *-r* végű szavak képeznek kivételt, mert magasabb arányuk ellenére sem mondhatóak tipikusnak. Többek közt ezért sem követik jobban a hangkivető mintát, mint a csoportban számosságban utánuk jövő *-ny*, *-n*, *-j*, *-cs* végűek. Az *-l* és az *-r* végű szavak viszonylag magasabb arányának okát nem ismerjük a nem hangkivetők csoportjában, a hangkivetők csoportján belül azonban valószínűsíthető, hogy azért is képviseltetik magukat relatíve magas arányban, mert könnyebb velük olyan mássalhangzó kapcsolatokat létrehozni, amelyek megfelelnek Rebrus és Törkenczy (2008) sémájának, amelyet az 5.1. alfejezetben mutattam be (pl. nem vesznek részt a zöngésségi hasonulásban).

Az ***-m*, *-g*, *-k* végű tövek** legfontosabb közös vonása az, hogy az ingadozással egyik leginkább együttjáró ***-t* tárgyrag hozzájuk kötőhang nélkül fonotaktikai**

[1] A *-g* végűek csoportja összesen 70 szót számlál, amelyek közt még vannak magas gyakoriságúak pl. *féreg*, *méreg*, de ezek gyakoriságát a *dolog* előfordulásainak száma messze meghaladja.

[2] Az alfejezetben szerepelő táblázatok oszlopai közt minden esetben megvizsgáltam, hogy találok-e statisztikailag szignifikáns, nem triviális (pl. egy csoport típusgyakoriságával korrelál példánygyakorisága) együttjárásokat. A csoportok kis száma miatt általában csak tendenciákat lehet kimutatni, amelyekből a közel szignifikánsakat elemzésemben kiemelem, a többiekre azonban külön nem térek ki.

okokból[1] **kifolyólag nem kapcsolódik**, így nem hangkivetéses alakjaiknak egy szótaggal hosszabbnak kell lenniük, ami kevésbé előnyös, mert jelentősebben különbözik (amennyiben a szótagszámot fontosabb jellemzőnek tartjuk, mint egy kötőhang meglétét) a korábbi hangkivetéses alaktól, mint ha kötőhang nélkül kapcsolódna a tárgyrag (vö. 5.4.2. alfejezet)[2]. Ez a hangkivető főnevek közti dominanciájuk erősödéséhez vezetett, hisz más végű szavak könnyebben vettek, vesznek részt az analógiás folyamatokban, és így olyan nem hangkivető szavak irányába mozdulnak el, amelyek az egyes szám alanyesetű alakhoz és nem a hangkivetéses tővariánshoz hasonlítanak. A *-g* és a *-k* egymáshoz áll közel (csak zöngésségükben térnek el), míg az *-m* a hangkivetők közt sokkal ritkább *-n*-nel, *-ny*-nyel (csak képzéshelybeli eltérés) alkot természetes osztályt, amelyek az *-m*-hez való erős hasonlóság miatt követik annak viselkedését, és következetesen hangkivetők. Ez a hatás különösen az *-n* végűek esetében szembetűnő, ahol a *-t* tárgyrag könnyebben kapcsolódhatna kötőhang nélkül, mivel képzési helye azonos az *-n-ével*[3] (Rebrus és Törkenczy 2008: 753). A hangkivetéssel kevésbé együttjáró többi végekhez (*-r, -l, -j, -z, -sz*) a tárgyrag kötőhang nélkül is kapcsolódhat a *-cs* kivételével.

---

[1] A *-mt, -gt, -kt* kapcsolatok nem megfelelő voltában is vannak fokozatbeli különbségek, hisz a *-gt* teljesen elfogadhatatlan a zöngésségi eltérés miatt, az *-mt* egyszer előfordul a *teremt* szóban, míg a *-kt* végű szavakra a morphdb.hu 11 szót hoz, az ÉrtSz. pedig 15-öt.

[2] Más hasonló végeknél a hangkivetés hiánya lehet véletlenszerű/történeti okokra visszavezethető, vagy tulajdonítható annak is, hogy ezek a szavak alaki felépítésükből kifolyólag sem felelnek meg jól a hangkivető szavakkal szemben támasztott szigorú kritériumoknak: egy szótagos szavak (pl. *sav, eb, öv, seb, csap*), hasonulna két mássalhangzójuk egymás mellé kerülve (pl. *küszöb, kaszab*), túl hasonló az utolsó két mássalhangzójuk (pl. *gyapot,* az *p-t* hasonlósága viszonylag nagy természetes osztályaik alapján: 0,28), vagy a kiesés után három mássalhangzó találkozna (*rosseb*).

[3] Ez az *-ny* esetében sem jelentene nehézséget: *szuronyt, asszonyt* stb.

utolsó fonéma	példány-gyakoriság	példány-gyakoriság alapján aránya	típus-gyakoriság	hányszor gyakoribb a hang-kivetőknél	hang-kivetés mértéke (típus-alapon)	hang-kivetés mértéke (példány-alapon)	hangkivetés az összes alakban
m	1412628	65,39%	558	21,25	99,6%	99,9%	52%
g	378878	17,54%	55	1,97	99,7%	99,9%	52%
k	183660	8,50%	170	2,47	98,1%	99,5%	52%
r	105764	4,90%	186	0,49	97,3%	98,2%	46%
ny	23048	1,07%	36	0,30	99,5%	99,7%	49%
l	22492	1,04%	26	0,23	74,7%	84,1%	33%
n	15644	0,72%	20	0,27	98,1%	99,6%	37%
j	12011	0,56%	21	0,27	96,7%	98,5%	39%
cs	4704	0,22%	13	0,63	99,6%	99,9%	69%
sz	1136	0,05%	5	0,02	32,4%	36,2%	48%
z	471	0,02%	2	0,02	43,7%	33,6%	17%

5.1. táblázat: Hangkivető főnevek csoportjai utolsó fonémáik alapján[1]

Az 5.1. táblázat alapján még azt is észrevehetjük, hogy a **hangkivetés mértéke összefügg egy vég tipikusságával a hangkivetők körében**. Minél kevésbé tipikus egy tővég, annál jobban ingadoznak a vele végződő szavak, amely alól a nazálisok kivételek a már említett okokból. E tendenciaszerű viselkedésnek nem felel meg a hangkivetőkre nem jellemző tővég a *-cs* sem, mivel következetesen hangkivetők az így végződő szavak, mert hozzájuk is csak kötőhangzóval kapcsolódhat a tárgyrag (a morhpdb.hu-ban nem találtam *-cst* végű főnevet). A *kapocs* szó esetén a hangkivetést tartalmazó alakvariációk aránya a szó összes alakját figyelembe véve rendkívül magas (99,7%), szemben a hangkivető főnevekre jellemző 49,7%-kal, hisz nagyon sok további gyakori képzett alakja van, amelyek a *kapcsol* igére vezethetők vissza. Ezt a hatást erősíti, hogy a *-cs* végű szavak csoportja kizárólag a *-kapocs* végű szavakból áll, amelyeknél a hangkivető viselkedést erősítheti a jelentésében és funkcionalitásában tőlük már

[1] A fonémákra való hivatkozás során a helyesírásban rögzített, de anakronisztikus *ly* betűkombinációt tartalmazó szavakra *-j*-vel fogok hivatkozni, a többi digráfokkal jellemzett fonémák esetében a könnyebb olvashatóság érdekében azonban megmaradok a kétbetűs jelölésnél.

eltávolodott, de formailag szorosan kötődő *kapcsán* (101391). A többi csoportnál is feltüntettem az összes alakjuk alapján számított hangkivetés arányát, amely tendenciajellegű összefüggésben van a hangkivetés mértékével a releváns toldalékoknál (típus-alapon: r(9) = 0,51 t = 1,78, p = 0,109, példány-alapon: r(9) = 0,53 t = 1,91, p = 0,089).

Az 5.2. táblázat megmutatja, hogy a **legjellemzőbb tővégek következetesen hangkivetők** akkor is, ha az **utolsó előtti két fonémát** veszem figyelembe. A 10 legjellemzőbb tővég csoportjában a hangkivetés mértéke típus-alapon 93,1%, míg a 10 legkevésbé jellemző végnél az átlag típus-alapon 69,5%. Az összes tővéget figyelembe véve a típusgyakoriság és a hangkivetés mértéke közt csak tendenciajellegű (r(23) = 0,32, t = 1,59, p = 0,125) összefüggés mutatkozott. Az **egyes csoportok hangkivetési mértéke jelentős pozitív korrelációban van az összes hangkivetéses alakok arányával** (típus-alapon: r(23) = 0,67, t=4,33, p < 0,001, példány-alapon: r(23) = 0,68, t=4,49, p < 0,001), ami azt jelenti, hogy a kiegyenlítődést valóban jelentősen erősíti, ha túlsúlyban vannak azok az alakok, amelyek felé a kiegyenlítődés tart. A 10 vagy kevesebb szót számláló tővégek csoportjai mind ingadoznak (a hangátvetéses szavakat magába foglaló *-ej* csoportot kivéve[1]), ami összhangban van az analógiás megközelítés elvárásaival, hisz a gyenge csoporthatás esetén az ingadozásra, illetve a szabályos viselkedésre való „hajlam" felerősödik.

[1] Ennek részletes magyarázatát az 5.2.5. alfejezetben adom meg.

utolsó két fonéma	példány-gyakoriság	példány-gyakoriság alapján aránya	típus-gyakoriság	hányszor gyakoribb a hang-kivetőknél	hang-kivetés mértéke (típus-alapon)	hang-kivetés mértéke (példány-alapon)	hangkivetés az összes alakban
om	807301	37,4%	318	221,90	99,6%	99,9%	52%
em	601185	27,8%	230	88,30	99,5%	99,8%	51%
og	352272	16,3%	11	18,90	100%	99,9%	49%
ek	97496	4,5%	56	10,00	99,7%	99,6%	53%
ok	66188	3,1%	85	9,60	99,0%	99,8%	51%
or	58155	2,7%	105	2,40	97,6%	97,2%	46%
ör	30732	1,4%	54	9,40	99,5%	99,4%	49%
eg	26606	1,2%	44	1,45	99,7%	99,7%	53%
ony	23048	1,1%	36	5,20	99,5%	99,7%	49%
er	16877	0,8%	27	0,22	91,7%	98,8%	40%
on	15644	0,7%	20	1,21	98,1%	99,6%	37%
ök	13362	0,6%	22	1,45	99%	99,5%	52%
ol	13306	0,6%	13	1,39	36,3%	80,7%	29%
oj	8694	0,4%	11	16,00	99%	99,1%	42%
ak	6614	0,3%	7	0,93	81%	95%	56%
öl	5212	0,2%	6	27,00	96,2%	99,2%	40%
ocs	4704	0,2%	13	7,80	99,6%	100%	69%
öm	4142	0,2%	10	5,17	95,8%	98,7%	56%
el	3249	0,2%	4	0,17	81,8%	78,8%	42%
ej	3089	0,1%	9	0,40	99,7%	99,1%	42%
usz	1136	0,1%	5	0,61	32,4%	36,2%	48%
al	725	0,0%	3	0,04	86,7%	88,3%	23%
ez	276	0,0%	1	0,14	23,9%	23,9%	9%
öj	228	0,0%	1	13,10	45,5%	45,5%	9%
oz	195	0,0%	1	0,30	63,4%	63,4%	25%

5.2. táblázat: A hangkivető főnevek csoportjai utolsó két fonémájuk alapján

Az utolsó **két mássalhangzó alapján** kialakítható főnévi csoportok (5.3. táblázat) **hangkivetési mértéke és összes alakjuknak hangkivetése közt magas pozitív**

együttjárást figyelhetünk meg (típus-alapon: r(38) = 0,54, t = 3,99, p < 0,001, példány-alapon: r(38) = 0,5, t = 3,55, p < 0,01). A tipikusság és a hangkivetés mértéke közti összefüggés már nem ennyire egyértelmű. Ha csak a 10 legjellemzőbb és a 10 legkevésbé jellemző tővég hangkivetési mértékének átlagát nézzük, akkor az összefüggés még érzékelhető (típus-alapon: 83,45% szemben 79,4%-kal, de példány alapon még mindig: 89,8% szemben 79,1%-kal), de ha az egyes mássalhangzó-párokra tekintünk, akkor már több nehezebben értelmezhető adatot vehetünk észre. A legjellemzőbb mássalhangzó-párok közt a nagyon ingadozókért a *-fátyol, -bajusz* és a *pityer* végű szavak a felelősek. Mint az 5.3. alfejezetben látni fogjuk, ezek a szavak a hangkivetők közt is magányosak felépítésükből kifolyólag. Ingadozásuk a hangkivetőktől való formai eltérésüknek tudható be. A hangkivetőkre kevésbé jellemző mássalhangzós tővégek közt található a kiugróan nagy gyakoriságú *szobor*, amelynek jelenléte még elég a csoport stabilitásának a biztosításához. A hangkivetők közt szintén kevésbé tipikus mássalhangzócsoportra, a *-h-j-*re végződnek a hangkivetésükben stabil hangátvetők. Hasonlóan a látszólag számosabb *-p-r* végűek is viszonylag következetesen hangkivetők. A csoport azonban csak 4 szót (*kapor* 93% : 532, *csupor* 98% : 237, *eper* 99% : 1268, *földieper* 100% : 133) számlál, amelyek legalább 25-ször fordulnak elő, és 6-ot, amelyek 1-nél többször.

Nincs szignifikáns összefüggés a **gyakoriság** és a **hangkivetés mértéke** közt sem, de láthatjuk, hogy a legszámosabb csoportok kivétel nélkül az átlagosnál jobban követik a hangkivető sémát (10 leggyakoribb csoport), míg a kisebb példánygyakoriságú csoportok esetében csak azok nem ingadoznak jelentősen, ahol összes alakjuk tekintetében gyakori a hangkivetés (*-gy-r, -k-ny, -j-k*). Egyedül a *-ty-k* mássalhangzós véget tartalmazó szavak (*-szutyok, -bütyök* végűek) ritkák, mégis következetesen hangkivető módon viselkednek, annak ellenére, hogy az utolsó két mássalhangzójuk igen hasonló. Ez a viselkedés elsősorban *-k* végüknek és tipikusnak mondható felépítésüknek tudható be (lásd 5.4. táblázat)

Az 5.3. táblázatban feltüntettem azt is, hogy természetes osztályaikkal számolva **két fonéma mennyire hasonlít** egymásra. Az azonos mássalhangzókból álló párok hiányán túl megállapíthatjuk, hogy a nagyon hasonló párok sem fordulnak elő: *l-r*: 0,86,

d-dz: 0,64, *c-t*: 0,60, *ny-j*: 0,5, *c-cs*: 0,47, *dz-dzs*: 0,47, *z-zs*:0,47, *sz-s*: 0,46, *gy-dzs*: 0,46, *ty-cs*: 0,41, *k-g* 0,4, *sz-z:* 0,4, *n-r*: 0,4, *l-n*: 0,4. A leghasonlóbb létező párok is csak néhány szónál találhatóak meg (*z-l:* 0,48 pl. *kazal*, *ty-k:* 0,35 pl. *szutyok*). A hiányok egyik oka a hasonulás, kiesés és az összeolvadás kerülése a mássalhangzók egymás mellé kerülése esetén, de ez nem minden esetben jó magyarázat, hisz bizonyos nem megfigyelhető pároknál ezek a hatások nem mindig lépnének fel: pl. *r:l* (*vitorla*), *l:n* (*tollnok*), *j:ny* (*tájnyelv*).

A **teljes mássalhangzókészletre jellemző átlagos hasonlóság** a természetes osztályok alapján számolva 0,22 (azonos párok nélkül 0,16), míg a hangkivető szavak utolsó mássalhangzóira: 0,12. Ebből az következik, hogy a hangkivetővé váláshoz és a csoportban maradáshoz szükséges a két mássalhangzónak az átlagosnál nagyobb különbözősége. A hangkivető szavak ilyen eloszlása nem egyedi. A McCarthy (1981) által alaposan leírt arab tövek KKE (OCP) hatását alaposabban megvizsgálva Frisch (1996) azt tapasztalta, hogy nemcsak a geminátáktiltottak a gyökök 2. és 3. mássalhangzójában, hanem a nagyon hasonlóak sem előnyösek, így minél eltérőbb két mássalhangzó ezekben a pozícióban, annál nagyobb számban fordulnak elő létező tövek végén (Frisch 1996: 67). A kevésbé előnyös párok hiánya így összhangban van Rebrus és Törkenczy (2008: 752) kijelentésével is:

„Feltételezzük továbbá, hogy ezek a morfofonológiától független fonotaktikai mintázatok percepciósan motivált, fonetikailag „lehorgonyzott" mintázatok. Feltételezzük továbbá, hogy ezek a mintázatok (is) „statisztikaiak", tehát erősségük függ a típusgyakoriságtól."

msh	példány-gyakoriság	példány-gyakoriság alapján aránya	típus-gyakoriság	hányszor gyakoribb a hangki-vetőknél	mgh hasonlósag	hang-kivetés mértéke (típus-alapon)	hang-kivetés mértéke (példány-alapon)	hangkivetés az összes alakban
l-m	1303182	60,65%	412	59,53	0,11	99,9%	99,9%	54%
l-g	348955	16,24%	7	17,27	0,08	100,0%	99,9%	51%
l-k	89304	4,16%	23	11,38	0,04	98,0%	99,6%	62%
r-m	84594	3,94%	96	3,13	0,11	98,7%	99,5%	47%
k-r	50786	2,36%	87	3,55	0,04	99,6%	99,6%	45%
t-k	35258	1,64%	18	1,70	0,28	99,7%	100%	64%
r-g	29923	1,39%	46	3,07	0,08	99,7%	99,7%	52%
r-k	26404	1,23%	56	1,55	0,04	99,1%	99,7%	50%
r-ny	22406	1,04%	33	4,65	0,1	99,5%	99,7%	48%
b-r	17642	0,82%	14	0,33	0,07	99,3%	99,9%	55%
sz-n	15644	0,73%	22	4,07	0,13	98,1%	99,6%	37%
k-l	11211	0,52%	9	1,45	0,04	96,8%	98,7%	35%
d-r	11196	0,52%	39	3,46	0,21	97,3%	95%	50%
h-r	9209	0,43%	9	2,46	0,13	96,4%	99,9%	57%
g-j	8922	0,42%	12	8,22	0,16	94,0%	98,3%	32%
z-m	8846	0,41%	22	3,39	0,07	100,0%	100%	58%
sz-k	8831	0,41%	36	0,75	0,11	97,9%	99,3%	46%
t-r	8382	0,39%	10	0,19	0,1	85,3%	81,8%	29%
j-m	7113	0,33%	24	6,90	0,1	97,6%	99,5%	54%
c-k	6976	0,32%	8	4,30	0,17	83,6%	82,4%	36%
s-k	5174	0,24%	1	0,01	0,10	100,0%	100%	50%
cs-k	5163	0,24%	10	2,18	0,16	99,6%	99,8%	45%
b-l	4795	0,22%	4	0,99	0,07	89,5%	75,8%	49%
p-cs	4704	0,22%	11	10,31	0,16	99,6%	100%	69%
m-r	4536	0,21%	5	1,08	0,11	92,8%	99,9%	41%
h-j	3089	0,14%	9	0,21	0,19	99,7%	99,1%	48%
p-r	2965	0,14%	9	0,24	0,04	98,8%	97,8%	30%
p-l	2777	0,13%	3	0,47	0,04	86,6%	98,3%	37%
ty-l	2231	0,10%	7	178,74	0,03	37,7%	69,8%	24%
ty-k	1645	0,08%	3	7,28	0,35	99,5%	99,7%	38%
z-l	1478	0,07%	3	0,58	0,48	89,6%	88,3%	24%
j-k	1146	0,05%	12	10,89	0,08	99,4%	97,9%	72%
j-sz	1136	0,05%	5	7,40	0,11	32,4%	36,2%	48%
ty-r	953	0,04%	4	15,56	0,03	74,0%	96,5%	39%
sz-l	753	0,04%	1	0,17	0,24	26,5%	26,5%	29%
k-ny	642	0,03%	1	0,72	0,1	98,5%	98,5%	61%
g-z	276	0,01%	1	0,07	0,11	23,9%	23,9%	9%
b-z	195	0,01%	1	0,11	0,11	63,4%	63,4%	25%
gy-r	87	0,00%	1	0,00	0,06	99,1%	99,1%	87%
c-r	8	0,00%	1	0,01	0,1	100,0%	100%	32%

5.3. táblázat: A hangkivető főnevek csoportjai utolsó két mássalhangzójuk alapján. A párok sorrendje azonos a szavakban található sorrenddel. Egyedül csak a *l-k*, *r-k* párok esetén tapasztalható, hogy mind a két sorrendben előforduljanak, de ezek is mind a két sorrendben szerepelnek a táblázatban.

A **hangkivető főnevek utolsó két magánhangzója[1] alapján kialakítható csoportok jellemzőit** az 5.4. táblázat mutatja be. A magánhangzófelépítés tipikussága („hányszor gyakoribb a hangkivetőknél" oszlop) és a hangkivetés mértékének összefüggése ($r(11) = 0{,}44$ $t = 1{,}62$, $p = 0{,}13$), illetve a típusgyakoriság és a hangkivetés mértékének pozitív korrelációja ($r(11) = 0{,}42$, $t = 1{,}51$, $p = 0{,}16$) csak tendenciajellegű. Szembetűnő az *-é-e* szekvenciájú szavak következetes hangkivető viselkedése[2]. Ez elsősorban néhány kiugróan gyakori, prototipikusnak is tekinthető szónak tulajdonítható (*érem* 10680, *fészek* 8790, *lélek* 165548, *méreg* 11365), amelyekből a *lélek* egyes szám alanyesetű alakja alapján az 5. leggyakoribb hangkivető szó, és a 2. leggyakoribb nem *-alom*/*-elem* végű hangkivető a *dolog* után, ami önmagában elegendő a csoport hangkivető viselkedésének a biztosítására. Az *-á-o* szekvenciát tartalmazó szavak ingadozása összefüggésbe hozható azzal, hogy viszonylag kevés hangkivetéses alakjuk van az összes alakjukat figyelembe véve (34%). Ezzel szemben az *-i-o* szekvenciájú szavaknak vannak leginkább hangkivetéses alakjaik összes alakjukat tekintve (57%), így annak ellenére nem ingadoznak, hogy csak rájuk nem érvényes a hangkivetőkre jellemző elölségi és kerekségi harmónia (a csoport szempontjából legfontosabb szó, az *izom* viselkedését röviden az 5.3.3. alfejezetben vizsgálom meg).

[1] Az utolsó magánhangzó alapján számított hangkivetési mértékek esetében csak az *-u* (32,4%) és *-a* (83%) végűek különböznek jelentősen a többi végtől: *-ö* (98,2%), *-e* (98,7%), *-o* (98%). Rebrus és Törkenczy (2008) ezeket a töveket (az *-u* és az *-a* végűeket) hagyta ki sémájuknak a megalkotása során.

[2] Az *-a-o* szekvenciájú szavak szintén szigorúan követik a hangkivető sémát, de ez elsősorban a teljes *-alom* véghez köthető. Az *-é-e* magánhangzó-szekvenciával rendelkező szavak mássalhangzóikat tekintve sokfélék.

mgh	példány-gyakoriság	példány-gyakoriság alapján aránya	hányszor gyakoribb a hang-kivetőknél	típus-gyakoriság	hang-kivetés mértéke (típus-alapon)	hang-kivetés mértéke (példány-alapon)	hangkivetés az összes alakban
a-o	788318	36,48%	28,14	325	99,8%	99,9%	53%
e-e	639464	29,59%	3,85	277	98,6%	99,4%	49%
o-o	434672	20,11%	15,52	128	98,9%	99,8%	50%
é-e	107550	4,98%	1,79	92	99,8%	99,8%	53%
á-o	59100	2,73%	3,15	59	86,7%	96,2%	34%
i-o	37634	1,74%	3,87	50	98,9%	99,9%	57%
ö-ö	28979	1,34%	4,33	52	98,0%	99,1%	47%
ü-ö	24697	1,14%	6,47	41	98,6%	99,3%	52%
u-o	24498	1,13%	22,76	46	98,5%	99,4%	40%
a-a	7339	0,34%	0,07	10	83,0%	94,9%	47%
ó-o	5906	0,27%	6,35	5	97,0%	99,5%	42%
i-e	1933	0,09%	0,07	2	50,0%	99,1%	28%
a-u	1136	0,05%	0,18	5	32,4%	32,4%	48%

5.4 táblázat: A hangkivető főnevek csoportjai utolsó két magánhangzójuk alapján

A tővégek, a mássalhangzó- és magánhangzó-végszekvenciák megvizsgálása után szisztematikusan **áttekintem** azokat az **összefüggéseket, amelyeket a szavak hangkivetési mértéke, végeinek felépítése és a gyakorisága közt találhatók**. Az elemzés során a kevésbé hangkivető módon viselkedő szavaknál a nagy gyakoriságú és a hangkivetők közt tipikus végű szavak vizsgálatára, a stabilabban hangkivetőknél a hangkivetőkre kevésbé jellemző végűekre helyezem a hangsúlyt, mivel ezek a szavak viselkednek elvárásaimmal ellentétes módon. A részletes áttekintés során az 5.5. táblázat által bemutatott 4 nagyobb, a hangkivetés mértéke alapján felosztott csoport szerint elemzek.

Hangkivetés mértéke	Átlagos példánygyakoriság
0-49,9%	488
50-89,9%	1311
90-98,9%	2469
99-100%	4131

5.5. táblázat: A hangkivető főnevek átlagos gyakorisága hangkivető viselkedésük mértéke szerint

A **csoportokban a hangkivetés aránya monoton nő az átlagos gyakorisággal**. Ezt az összefüggést mérsékelten ($r(204) = 0{,}16$ $t = 2.27$, $p<0{,}05$)[1], akkor is ki tudom mutatni, ha az egyes alapszavak gyakoriságát és hangkivetésének mértékét vetem össze. Azonban az egyedi szavak (alapszavak és összetett szavak) hangkivetési mértéke és gyakorisága közt már nem mutatható ki szignifikáns korreláció. Ez azzal lehet összefüggésben, hogy a hangkivetők fonológiai felépítésüket tekintve meglehetősen heterogén csoportot alkotnak. A gyakoriság önmagában nem felelős az ingadozásért, csak a megfelelő hasonlóságot mutató csoportokon belül tudjuk tetten éri működését.

Vizsgálatomban nem térek ki **a nyitás ingadozásának** vizsgálatára, amikor egy nyitótő nem nyitó (*tűsarkat* > *tűsarkot*), vagy amikor egy nem nyitó tő nyitó (*vágyálmot* > *vágyálmat*) módon kezd viselkedni[2], mivel ennek áttekintése túlmutatna elemzésem keretein. A nyitás mértéke és a hangkivetés mértéke közt nincs statisztikailag kimutatható összefüggés, azonban a nyitásban ingadozó 48 szó közt (nyitás mértéke 1 és 99% közt található[3]) a hangkivetés mértéke alacsony (89,1%). Az ingadozó nyitótövek utolsó magánhangzója /o/, ami alól csak a *kazal*, illetve a *-köl(y)ök* végűek a kivételek. Legszembetűnőbb egyedi jelenség a nyitás ingadozásával kapcsolatban, hogy több

[1] A számításban az egyes alapszavak szóbokraikkal együtt szerepelnek, így a *pocok* szó hangkivetésének arányát és gyakoriságát 564-nek és 96,5%-nak vettem, amely összegzi a *pocok* és a *pézsmapocok* előfordulásait. A számításból a szavak leggyakoribb és legritkább 5%-át kihagytam.

[2] A nyitásban való ingadozás több esetében nem lehet meghatározni jelen pillanatban, hogy a folyamat honnan hova tart, azaz egy ingadozó szó korábban nyitó vagy nem nyitó volt-e, vagy volt-e egyáltalán valaha stabilan az egyik csoportban.

[3] Ebből az összehasonlításból kizártam azokat a szavakat, amelyeknek utolsó magánhangzója *e*, hisz esetükben az esetleges kötőhang is mindig nyílt *e* lenne.

olyan *-sarok* végű szó nyitása is ingadozik (*csizmasarok, cipősarok, tűsarok*), amelyeknél a szabályalapú megközelítés a kétféle jelentésű *sarok*-hoz ('ember vagy lábbeli sarka', 'utca sarka') tartozó kétféle nyitási viselkedés alapján kategorikus viselkedést jósolna[1].

5.2.2. 0-49,9%-ban hangkivető sémát követő szavak

Az 5.6. táblázat alapján láthatjuk, hogy a hangkivető sémát nem vagy kevésbé követő szavak **94,69%-ának** olyan **végződése** van, amely **nem jellemző a hangkivetőkre** (nem *-m* és *-k* végűek), azaz viselkedésükben közelítenek azokhoz a nem hangkivető szavakhoz, amelyekre végeik szerint is jobban hasonlítanak. A csoportban egyetlen gyakori szó a *bajusz*, amely a hangkivetők közt teljesen egyedinek (*-sz* vég, utolsó magánhangzó *-u-*) minősülő felépítésének köszönhetően ingadozik.

vég	példány-gyakoriság	példánygyakoriság alapján aránya
sz	5021	60,51%
r	1105	13,32%
l	1062	12,8%
z	447	5,39%
m	328	3,95%
j	222	2,68%
k	113	1,36%
	8298	**100,00%**

5.6. táblázat: Kevésbé hangkivető szavak végei egyes szám alanyesetű alakjuk gyakoriságával súlyozva

A kevésbé hangkivető szavak közül utolsó két magánhangzóját és utolsó két fonémáját figyelembe véve a *bögöly* különbözik a leginkább a nem hangkivetőktől (mint később látni fogjuk, a hangkivetőktől is jelentősen különbözik: 5.3.2. alfejezet), de hangkivető viselkedését nem erősítik képzett alakok, csupán egyetlen egy *böglyös* előfordulást tartalmaz hozzá a *Szószablya Gyakorisági Szótár*. A *Szószablya Gyakorisági Szótár* összesen 17 erősen ingadozó szót tartalmaz:

[1] A *sarok* szó esetében az ingadozás nem meglepő, hisz ott ez magyarázható a két eltérő jelentéshez kapcsolódó két forma (*sarkat* 'ember vagy cipő sarkát', *sarkot* 'utca sarkát') keveredésével.

álbajusz (27), *arcfátyol* (9), *ászok* (110), *bajusz* (4985), *bögöly* (222), *gyászfátyol* (34), *jászol* (646), *kefebajusz* (3), *ködfátyol* (358), *macskabajusz* (4), *pityer* (46), *szlalom* (328), *tegez* (447), *tüllfátyol* (11), *tündérfátyol* (4), *veder* (1059), *vicikvacak* (3).

Az ***-m* és a *-k* tővégeket** a *szlalom* (328), *ászok* (110) és a *vicikvacak* (3) szavakban találhatjuk. Mind a három szó lényegesen ritkább, mint az a hangkivetőkre átlagosan jellemző, mivel az átlagos gyakoriság az összes toldalékos alakot figyelembe véve 3890 előfordulás, de még csoportjuk viszonylag alacsony átlagos gyakoriságát sem érik el. Képzett alakjai csak az *ászok*-nak vannak, de esetében is a hangkivetés rendkívül alacsony az összes alakot figyelembe véve (16%).

A ***szlalom*** jobb oldala azonos az *-alom*-mal, amely kizárólagosan jellemző a hangkivető főnevekre. A nem hangkivető főnevekkel azonos viselkedésének oka, hogy a hangkivető szavak alapvetően [-idegen] jeggyel bírnak[1], a *szlalom* viszont jelentése és a magyarra nem jellemző *szl-* szókezdete miatt [+idegen] jeggyel bír (norvég eredetű, 20. századi szó). A *szlalom* szó így a [+idegen] jeggyel bíró szavakra törekszik hasonlítani.

Az ***ászok*** *-á-o* magánhangzó-szekvenciája csak 3,15-szer gyakoribb, mint a többi főnévnél (a magánhangzó-szekvenciák alapján kialakított csoportok átlaga 8,75), és az ebbe a csoportba tartozó szavak nagy arányban ingadoznak is (86,7%-os hangkivetési mérték). A hangkivetők közt kevésbé tipikus voltát erősíti a magánhangzós kezdet és záró mássalhangzó-szekvenciája is, amellyel egy átlagos beszélő 150-szer gyakrabban találkozik a nem hangkivető szavaknál. A kiegyenlítődésben való előrehaladottságban szerepet kaphat az *Arany Ászok* sörmárka hatása is, amelynek alakjai azonban nem

[1] Az idegen jegy bináris megkülönböztető jegyként való kezelése mellett elméleti szempontból nem vagyok elkötelezett. Csupán csak egy ilyen tulajdonság valamilyen formában lévő létezésére kívántam utalni.

játszhatnak itt közvetlen szerepet, mert a *Szószablya Gyakoriségi Szótár* következetesen külön kezeli a nagybetűs és a kisbetűs szavak gyakoriságát[1].

A *vicikvacak* szónál a hangkivető séma elhagyásának oka a szó ritkasága mellett az, hogy a hangkivetőkre egyáltalán nem jellemző *-a-a* magánhangzó-szekvenciát tartalmazza záró helyzetben. A nem hangkivető főnevekre ez a szekvencia 14-szer jellemzőbb. Igaz, ezekbe beleszámítódnak azok a szavak is, amelyek közvetlenül *-a*-ra végződnek.

5.2.3. 50-89,9%-ban hangkivető sémát követő szavak

Az 5.7. táblázat alapján láthatjuk, hogy **a hangkivető főnevekre jellemző tővégek** (*-m, -k)* **száma 9,96%-ra emelkedett**, és 0,72%-ra estek vissza a kevésbé jellemző *-sz*, *-z* tővégek, amelyek csak néhány erősen ingadozó szónak voltak tulajdoníthatóak. Ebből egyértelműen látható, hogy az 50-89,9%-ban a hangkivető sémát követő szavak közt továbbra is a nem tipikus végek (*-l, -r, -z, -sz, -n*) a jellemzőek, de kisebb arányban, mint a hangkivetőnek sorolt, de már dominánsan nem hangkivetőként viselkedő töveknél.

[1] Valamilyen pontosan meg nem határozható mértékben számolhatunk a tulajdonnév kisbetűs alakjaival is informális közegben, ez azonban nem magyarázza kizárólagosan a szó csupán 26,3%-os hangkivetési mértékét.

vég	példány-gyakoriság	példánygyakoriság alapján aránya
r	16450	46,45%
l	15171	42,84%
k	3277	9,25%
m	252	0,71%
z	145	0,41%
sz	111	0,31%
n	9	0,03%
	35415	**100,00%**

5.7. táblázat: 50-89,9%-ban a hangkivető sémát követő szavak (27 db) végei egyes szám alanyesetű alakjuk gyakoriságával súlyozva

Az ebben a **csoportban található szavak is ritkák**. Egyedül a *sátor* (15731) és a *kebel* (10189) tűnik ki nagyobb gyakoriságával. Magánhangzó-mássalhangzó szekvenciáik és végeik azonban nem tipikusak a hangkivetők közt (5.8. táblázat). Csak az *-or* vég és a magánhangzó-szekvenciák jellemzőek jobban a hangkivetőkre, de ez nem elegendő ahhoz, hogy stabilan a hangkivetők közt maradjanak. A *sátor* kevésbé hangkivető viselkedését támogathatja, hogy az *-á-o* magánhangzó-szekvenciát tartalmazó szavak különösen ingadoznak, illetve hogy a *sátor* összes alakját figyelembe véve alacsony arányban szerepelnek a hangkivetéses alakok (34%), mivel a hozzá tartozó képzett szavak kis számúak és ritkák (*sátras, sátrazás, sátrazó* stb.), amelyeknek a hangkivetést nem tartalmazó párjai (*sátoros*, *sátorozás*, *sátorozó* stb.*)* ezeknél gyakoribbak is. Alacsonyabb gyakoriságának és a hangkivetők közt kevésbé tipikus felépítésének betudhatóan a *kebel* 67,3%-ban, a *sátor* azonban még 81,5%-ban viselkedik a hangkivető séma szerint. A *kebel* ingadozását erősítheti[1], hogy E.3 birtokos alakjai két formában eltérő jelentésekkel jelennek meg: *asszony keble, anyaszentegyház kebele*. A kétféle alak és jelentés azonban keveredik.

[1] A *kebele* alak terjedésére a tulajdonnévi *Kebele* pataknak és helyiségnek vélhetőleg alacsony ismertségükből kifolyólag nincs hatása.

szóvég, szekvenciák	hányszor gyakoribb a hangkivetőknél
r	0,49
l	0,23
or	2,39
el	0,17
b-l	0,98
t-r	0,19
tor	1,02
bel	0,79
á-o	3,14
e-e	3,85

5.8. táblázat: a *sátor* és a *kebel* végei

A további, utolsó mássalhangzójuk alapján **tipikus hangkivetőnek** mondható szavak **nagyon alacsony gyakoriságúak**, így esetükben indokolt, hogy elindultak az analógiás változás útján: *gyalom* (5), *korteher* (7), *murok* (14), *bürök* (32), *sulyom* (40), *előterem* (48). Az *előterem* szó vélhetőleg kisebb gyakorisággal fordul elő, és kevésbé is ingadozik, de alakjai nem minden esetben választhatók szét az *előteremt* ige, illetve az *előtér* főnév alakjaitól (*előtér*+POSS.E.1+ACC = *előteremet*).

Némileg gyakoribb, de a hangkivető főnevek átlagától elmaradó **tipikus végfonémára végződő szavak** még az *üröm* (159), *gyilok* (213) és a *vacak* (3018). A *vacak* hangalakját már elemeztem, elvárásaimmal összhangban a nagyobb gyakoriságú *-vacak* végű szó inkább mutat hangkivető viselkedést (*vacak* 75,74%, *vicikvacak* 0%). A *vacak* kiegyenlítődését támogatja a kontrasztra való törekvés is, hisz nem hangkivetéses alakjai jobban megkülönböztethetőek a *vacok* alakjaitól. Habár a *gyilok* gyakorisága önmagában sem túl magas, ingadozása ahhoz kapcsolható, hogy a szlengben az eredeti ‘kétélű tőr’, ’várfal koronáján végigfutó, mellvédes (és fedett) (védő) folyosó’ jelentései helyett ‘gyilkolászás, mészárlás, használattal való tönkretétel’ értelemben használatos, és eltávolodott jelentésében az eredeti hangkivetéses alakok hatása gyengébb. Ezt a jelentését vélhetőleg a *gyilkolás* szóra való hasonlósága miatt kaphatta. Természetesen ingadozó alakok előfordulhatnak a közel eredeti ‘rövid pengés gyilkolóeszköz’ jelentésben is: „A szamurálykard[1], machete és egyéb pengés gyilokok tartása leginkább

[1] A szó eredeti lejegyzésében is *ly*-nal szerepel.

egy bizonyos népcsoportra jellemző." („Bozotvágó késsel ölt: mindkét áldozat meghalt" 2009) Az *üröm* szó ingadozása mögött esetleg az *Üröm* településnév hatása állhat[1], amelyet (más tulajdonnévként használt rendhagyó szavakhoz hasonlóan) nem hangkivető módon szokás toldalékolni.

5.2.4. 90-98,9%-ban hangkivető sémát követő szavak

Ebben a csoportban már a **hangkivetőkre jellemző szavak szerepelnek**. A *-k*, *-g*, *-m* végfonémák már az alakok 80,68%-át fedik le, de a kevésbé következetesen hangkivető tövek közt a *-g*, *-k* nagyobb részt hasít ki (45,83%), mint a legtipikusabbnak mondható *-m* vég (34,91%). A hangkivetőkre nem jellemző *-sz* és *-z* végek már nem képviseltetik magukat ebben a csoportban.

vég	példány-gyakoriság	példánygyakoriság alapján aránya
k	84753	45,73%
m	64708	34,91%
l	15183	8,19%
n	8738	4,71%
j	4991	2,69%
r	3830	2,07%
ny	2823	1,52%
g	190	0,10%
cs	128	0,07%
	185344	**100,00%**

5.9. táblázat: 90-98,9%-ban a hangkivető sémát követő szavak (75 db) végei egyes szám alanyesetű alakjuk gyakoriságával súlyozva

Még ebben a csoportban is inkább ritka szavak találhatók. Az átlagos értéket az *orom* (3888), *hurok* (4271), *Szentlélek* (4453), *pokol* (4962), *lepel* (5920), *berek* (5929), *vászon*

[1] Elképzelhető, hogy a köznévi hangkivetőkre a homonim, de nem hangkivető tulajdonnevek nincsenek erős hatással, de az informális nyelvhasználatban gyakori kisbetűs írásmódjukból kifolyólag még a *Szószablya Gyakorisági Szótár*ban is számolnunk kell a köznévi és a tulajdonnévi alakok keveredésével. Az *üröm* esetében a nagy gyakoriságú, a komplex jegymérték szerint hozzá leginkább hasonlító *öröm* hatása sem zárható ki.

(8587), *ajak* (25039), *küzdelem* (33752), *telek* (43541) szavak gyakorisága haladja, illetve közelíti meg. A *vászon*, *pokol*, *lepel* szavak szereplése ebben a csoportban igazolható azzal, hogy fonológiai felépítésük alapján közelítenek a nem hangkivetőkhöz (elsősorban utolsó két mássalhangzójuk/fonémájuk kevésbé jellemző a hangkivetőkre: *-on* (1,21-szer gyakoribb), *-p-l* (0,47-szer gyakoribb), *-k-l* (1,45-ször gyakoribb), de nagy gyakoriságuk lassítja az analógiás változást. A *vászon* esetében a *sátor*-hoz, a *lepel*-nél pedig a *kebel*-hez való hasonlóság erősíti a kiegyenlítődési folyamatot. A *pokol*-nál pedig az támogatja hangkivetésének mérséklődését, hogy összes alakjaiban csak 15,4%-ban szerepelnek hangkivetéses alakok.

Legnehezebben a ***küzdelem*** megjelenése indokolható ebben a csoportban, ami a *küzdelemet* szóalak magas előfordulásának köszönhető. A 360 előfordulás 9864 *küzdelmet* alakra jut, aránya jóval meghaladja az 1‰ körüli küszöbértéket, amellyel esetlegesen elgépelésnek vehetnénk. A 46 *-elem* végű alapszóból a *CCelem* végűek 99,7%-ban[1], míg a *VCelem* végűek 99,9%-ban viselkednek hangkivető módon. Ez az eltérés azonban nem akkora, hogy a *küzdelem* egyedi viselkedését magyarázza.

A ***berek, telek, Szentlélek*** szavaknál elmondható, hogy az *-e-e* (3,85-ször jellemzőbb) és az *-é-e* (1,79-szer jellemzőbb) magánhangzó-szekvenciáik jellemzőbbek a hangkivetőkre, de nem kiugró mértékben. Ugyanez igaz az *-r-k* (1,55-ször jellemzőbb) mássalhangzó-szekvenciára, de az *-l-k* már jellemzőnek mondható (11,38-szer gyakoribb a hangkivetők közt). A nagyobb ingadozásért más, az analógiában szerepet játszó, de nem a gyakoriságon vagy a fonológiai hasonlóságon alapuló tényezők adhatnak számot. A *Szentlélek* analógiás változásához tulajdonnévi jellege járulhat hozzá. A *berek* ingadozása elsősorban ékezetmentes alakoknak (pl *béreket* : *bereket*), és a *telekhez* való erős hasonlóságának köszönhető.

Habár a ***telek*** egyes szekvenciáit tekintve tipikus hangkivető, a komplex jegymérték alapján a *telek*-hez 100 leginkább hasonlító szóból csak 39 hangkivető (bővebben 6.3. alfejezet). **Ingadozására** minden bizonnyal **hatással van** a sok nem hangkivetőként viselkedő vagy legalábbis ingadozó *-telek* végű **településnév**: pl.

[1] A teljes *-elem* végű csoportban még az egymásra nagyon hasonlító *-sejtelem* végűek és a *képzelem* ingadozik némileg.

Kisteleket (10) : *Kistelket* (1), *Lakiteleket* (2) : *Lakitelket* (4), *Csanyteleket* (2) : *Csanytelket* (0), *Aggteleket* (9) : *Aggtelket* (0). A *telek*-nek csupán tárgyesete és szuperesszívusza ingadozik, ami az ingadozás korai szakaszára jellemző viselkedéssel összhangban van.

A ***hurok*** felépítésében a *berek*-hez hasonló, de magánhangzói inkább jellemzőek a hangkivetőkre. Az *-u-o* magánhangzó-végszekvencia a hangkivetőknél 22,76-szor gyakoribb, mint a többi főnévnél. Kiegyenlítődésében szerepet kaphatnak a *hurokja-* kezdetű alakok, amelyek a *hurka* 'húsétel' szóval szemben tartják fenn a kontrasztot. A *Szószablya Gyakoriségi Szótár* alapján az ***orom*** szónak viszonylag kevés alakját használjuk csak. Ezekből a gyakoriak (többes szám, birtokos alakok) stabilan hangkivetők, csak az *oromot* (66), *oromom* (48) alakok tűnnek analógiásan kiegyenlítettnek. Feltűnő esetükben, hogy nincs hangkivető párjuk. A szokatlan eloszlás oka, hogy ezek az *örömöt*, *örömöm* szavak ékezettelen lejegyzései. Erre utal, hogy a biztosan az *öröm*-höz köthető alakok is közel hasonló gyakoriságúak (*oromomre* (44), *oromomben* (16)), valamint a .hu domain alatt ezekre a kulcsszavakra keresve a *Google*-ban kizárólag az *öröm* ékezetmentes változataira találunk. Az *ajak* ingadozása hasonló okokra vezethető vissza, mint a *vicikvacak* esetében (bővebben 5.3.2. alfejezet), hangkivetésének magasabb mértéke nagy gyakoriságának (25039) tudható be.

5.2.5. 99-100%-ban hangkivető sémát követő szavak

Ezek azok a szavak, amelyek szótári besorolásuknak és a hagyományos nyelvtani leírásnak megfelelően viselkednek. **89,49%-uk a hangkivetőkre kifejezetten jellemző végű** (*-m, -k, -g*) tövekből kerül ki. Ezeket követik az *-r* és a *-cs* végűek, amelyek azonban viselkedésüket tekintve eltérnek, hisz az *-r* végűek ingadozásra hajlamosak[1], a *-cs* végűek pedig egyáltalán nem. A hangkivetőkre nem jellemző *-n*, *-ny*, *-j*, *-l* végek csupán csak az esetek 2,39%-áért felelősek. Mivel az ilyen végű szavak viselkednek látszólag elvárásaimmal ellentétesen, ezért ezeket vizsgálom meg

[1] Összesen 159 következetesen hangkivető *-r* végű szó van, azaz önmagában az a tény, hogy a *-t* tárgyrag könnyen kapcsolódhat kötőhang nélkül, még nem elég az ingadozáshoz.

alaposabban. Igaz, az *-n* és az *-ny* viselkedéséért az *-m*-hez való hasonlóság is felelős lehet, de azok alaposabb megvizsgálása során további összefüggéseket is felismerhetünk.

vég	példány-gyakoriság	példánygyakoriság alapján aránya
m	2535228	63,07%
g	598196	14,88%
k	464053	11,54%
r	219618	5,46%
cs	107067	2,66%
n	54241	1,35%
ny	21233	0,53%
j	16262	0,40%
l	4388	0,11%
	4019836	100,00%

5.10. táblázat: 99-100%-ban hangkivető szavak (973 db) végei egyes szám alanyesetű alakjuk gyakoriságával súlyozva

***-j (ly)* végű szavak**, amelyek nem ingadoznak: *szappanpehely* (4), *burgonyapehely* (15), *kukoricapehely* (119), *zabpehely* (348), *hópehely* (895), *pehely* (985); *misekehely* (3), *virágkehely* (94); *boholy* (189); *státusfogoly* (14), *államfogoly* (20), *hadifogoly* (2207), *fogoly* (11261); *lángbagoly* (1), *macskabagoly* (25), *gyöngybagoly* (39), *hóbagoly* (43).

A ***-pehely, -kehely, -boholy* végű** szavak a hangátvetők csoportjába[1] tartoznak. A mintát a kiugróan magas gyakoriságú *teher* (55913) analógiás hatása tartja életben, amelynek számos nagy gyakoriságú képzett alakja is van (*terhes* 22881, *terhesség* 15716, *terhelés* 9631). Egyedül a *kehely* szó alacsonyabb hangkivetési mértéke (98,7%) egyedi, ami a *kehelyt* alak előfordulásának köszönhető. Ez a szóalak gyakran bukkan fel archaikus, vallási és irodalmi közegben, ahol sajátos hangulatot közvetít. Törkenczy és Rebrus (2008: 757) a hangátvetők csoportját gyengébbnek veszi hasonlósági alapokon, mint a többi hangkivető főnevet, így jobban elvárják ingadozásukat:

[1] A minta 3 szabad alakot: *teher*, *pehely*, *kehely* és egy kötött alakot (*boholy*), és az ezekből alkotott összetett szavakat fedi le.

„A hangátvető tövek egyedi sémája olyan specifikus és annyira gyenge, hogy semmilyen hatással sincs a stabil *VC*-tövek nagyon erős sémájára, az utóbbiak viszont nagyon is vonzzák a hangátvető töveket."

Megállapításuk azért nem helytálló, mert nem veszik figyelembe a *teher* szó alakjainak kiugróan magas gyakoriságát.

A kisebb gyakoriságú ***-fogoly* végű alakok** a nagy gyakoriságú *fogoly* (11261) mellett összetartó csoportot alkotnak. Stabil viselkedésük a hangkivetőkre kifejezetten jellemző *-o-o* (15,52-szer gyakoribb, mint a többi főnévnél), és *-g-j* (8,22-szer gyakoribb) végszekvenciáiknak, illetve *-oj* (16-szor jellemzőbb) végüknek köszönhető. A hasonló felépítés segíthet a ***-bagoly* végű összetett szavaknak** is hangkivető voltuk megőrzésében, amelyeket a nagyon jellemző *-a-o* (28,13-szor jellemzőbb) szekvencia, és a *fogoly*-hoz való nagyfokú hasonlóság még inkább e csoportban tart. Magyarázatra szorul azonban, hogy a *bagoly* (92,7%, 1493) miért ingadozik a többi *-bagoly* végű szónál nagyobb mértékben, különösen, mivel ezeknél nagyobb gyakoriságú. Viselkedésére hatással lehet az erősen ingadozó *bajuszra* való szerkezeti hasonlóság. Elképzelhető, hogy a *bagoly* alak azért is ingadozóbb, mert a többi *-bagoly* végű összetett szó inkább a művelt, természettudományos párbeszédre jellemző, amelyben jellemzőbb a normák, így a következetes hangkivetés betartása is.

Az ***-l* végű** *hólepel* (4), *bélpokol* (4), *birkaakol* (5) szavaknál az ingadozás hiánya elsősorban annak tudható be, hogy nagyon kevés adat szerepel róluk a korpuszban. A *Google Gyakoriségi Gyűjtés* alapján viselkedésükről azonban már megállapítható, hogy jobban ingadoznak:

(1) *hólepel* 66,5% : 451 alak
bélpokol 100% : 74 alak
birkaakol 82% : 57 alak

A továbbra is következetesen hangkivető ***bélpokol*** a nála sokkal gyakoribb *bélpoklos* (*Szószablya Gyakorisági Szótár:* 227, *Google:* 23300 találat) hatására marad a hangkivetők csoportjában.

Az ***ököl*** (4125) szót viszonylagosan **nagy gyakorisága** tarthatja ebben a csoportban, valamint az, hogy az *-öl* (27-szer gyakoribb a hangkivetők közt) vége, és *-ö-ö* (4,33-szor gyakoribb a hangkivetők közt) magánhangzó-szekvenciája alapján tipikusnak mondató hangkivető szó[1]. A *tengeröböl* (250) magas hangkivetési mértékének egyik oka véletlenszerű adathiány lehet, mert a *Google Gyakorisági Gyűjtés* már tartalmaz olyan alakokat, amelyekben nincs a hangkivetést elváró tárgyrag előtt hangkivetés: *tengeröblöt* (1230), *tengeröbölt* (67). Ez már viszonylag jelentősebb, 5,2%-os ingadozás, de még mindig elmarad a legtöbb *-öböl* végű szóétól, és a többi alakja továbbra is következetesen hangkivető. Vélhetőleg ez abból következik, hogy műveltségszónak minősül, mint a *-bagoly* szócsoport tagjai, amit azonban nehéz egyértelműen alátámasztani[2].

Az ***-n* végű** alakok közt a leggyakoribb *haszon* (50002) elvárásaimnak megfelelő módon nem ingadozik. A belőle létrehozott összetett szavak a *haszon* megtartó hatása miatt szintén nem ingadoznak. A további *-n* végű szavak közt *-vászon* végű szavak találhatók, habár a leggyakoribb *-vászon* végű szó, a *vászon* (8587) csak 98%-ban követi a hangkivető mintát. Néhány *-vászon* végű összetett szó pedig ennél is kevésbé hangkivető: *zsákvászon* (77,8%), *csiszolóvászon* (92,3%), *lenvászon* (97%), *gyöngyvászon* (97,8%). Az ingadozó és a nem ingadozó ***-vászon* végűek** közt nincs egyértelmű

---

[1] A hangkivetők ingadozására gyakran példaként felhozott *öböl* szó felépítésében kevésbé megfelelő hangkivető, hisz a *b-l* között nagyobb a hasonlóság, mint a *k-l* esetében.

[2] A *Szószablya Korpusz*ban az *öböl* esetében 3,09 a *tengeröböl*nél pedig 2,5 a jó, illetve rossz helyesírású szövegekben lévő előfordulásaik aránya, azaz a *tengeröböl* némileg gyakrabban szerepel jobb helyesírású szövegekben. A helyesírás azonban önmagában nem elég ahhoz, hogy következtessünk egy szó kontextusának jellegére, és kijelenthessük, hogy a *tengeröböl* általában „emelkedettebb" szövegekben fordul elő. Természetesen az ilyen típusú vizsgálatok sem lehetetlenek kvantitatív módon, de ehhez be kellene vonnunk más paramétereket is (pl. átlagos mondathossz a szövegkörnyezetében, ikes ragozás mértéke, biztosan formális szavak, kifejezések összesített aránya stb.), amelyekből következtethetnénk, de akkor is ez csak egy kiterjedtebb vizsgálat részeként lehetne elképzelhető.

különbség. A stabilan hangkivető csoport leggyakoribb szavai, a *filmvászon* (2331) és a *mozivászon* (1260), a *vászon*-tól nagy gyakoriságuknak köszönhetően elszakadóban vannak, és autonóm, önálló viselkedést alakítanak ki (Hay 2001, 2002). A következetesebb hangkivetés is lehet az autonóm viselkedés része, hisz a *vászon* esetében a hossza miatt egyértelműbb a többi ingadozó és gyakori *-á-o* szekvenciát tartalmazó szavakhoz való kapcsolódás, amelyet azonban az önállóan viselkedő *filmvászon*-nak és *mozivászon*-nak nem kell követnie. Köréjük csoportosulnak a nem ingadozó műveltségi és inkább hosszabb *-vászon* végű szavak, míg a *vászon*-hoz a munkásélethez kötődő rövid előtagú összetett szavak tartoznak.

Az ***-ny* végű** szavak egy kivétellel a nagy gyakoriságú *torony*-ból (16922) létrehozott összetett szavak. A nagy gyakoriság miatt a *torony* szónak stabilan hangkivetőnek kell maradnia, a hozzá nagyon hasonló szavak számára pedig ő a minta. Bizonyos *-torony* végű szavak azonban 99%-nál kevésbé követik a hangkivető mintát: *hústorony* (4,55%), *óratorony* (4,35%), *hűtőtorony* (1,69%), *víztorony* (1,2%). Az eredeti szótól jelentésében legerősebben az emberre alkalmazható *hústorony* tér el. Ennek megfelelően ez is ingadozik a legjobban.

5.3. Hangkivető szavak hasonlósági csoportjai

5.3.1. Az elemzés célja

A hangkivetés mértékén alapuló elemzés után **hasonlósági viszonyok alapján kialakított csoportokban vizsgálom tovább a hangkivető főneveket**. Ebben 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 és Gaume 2003).

A szavak viszonyainak gráfstruktúrában való tanulmányozása **kvalitatív megközelítési mód**. Közvetlenül ennek alapján nem tudjuk leírni, sem jósolni 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) a **komplex jegymértékkel határoztam meg**. Választásom azért esett erre az algoritmusra, mivel ez bizonyult a legjobbnak a későbbiekben bemutatandó (6.3. alfejezet) „hagyj-ki-egyet" (*leave-one-out*) tesztben. 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 1074 elemet tartalmaz (az összes olyan hangkivető főnév, amelyeknél a hangkivetéssel együttjáró toldalékaikról van gyakorisági adatom), amelynek teljes bemutatására nincs mód. A komplex jegymérték segítségével készített gráf elemzését kiegészítem olyan részgráfok vizsgálatával is, amelyeket a komplex tengelymérték vagy a természetes osztályok 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 terjedő** [1] **egymáshoz való hasonlóságát jelenítettem meg** 10 fokozatban. Minél vastagabb, kevésbé átlátszó, mélyebb színű 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: pl. *halálveszedelem-veszedelem.*

Csomópontok

A csomópontok egy-egy szónak felelnek meg, amelyeket a viszonyukat kifejező élek kötnek össze. Az egyes **szavakat** (csomópontokat) gyakoriságuk és hangkivetési mértékük jellemzi. 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. Ez a fajta ábrázolásmód a nagyon alacsony és az alacsony gyakoriságú szavak között méretben csak kisebb különbséget tesz, viszont ezzel lehetővé válik, hogy a lehetséges prototípusok[2] közti gyakoriságbeli eltérések érzékelhetőek legyenek még áttekinthető módon.

A csomópontok elütő színű **keretének** vastagsága azt jelzi, hogy a *Szószablya Gyakorisági Szótár* alapján mennyire **ingadozik** az adott szó a hangkivetéssel együttjáró toldalékos alakjaiban. Ha egy szó nem ingadozik, akkor nincs ilyen kerete. Az átlagosan 97,57% százalékban a hangkivető mintát követő szavakat vékony keret övezi, míg az erőteljesen ingadozó szavakat egyre vastagabb. Egy **szó** minél **melegebb**, mélyebb színű, **annál több 0,9-nél magasabb hasonlósági kapcsolata van**, azaz a vörösesbarna csomópontok ideális prototípusok, ha más tulajdonságaikban is megfelelőek, a sárgás

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

[2] A 4.2. alfejezetben kifejtett elképzeléseim szerint a gyakori szavak könnyebben lehetnek prototípusok. Ezt a 6.5. alfejezet vizsgálatai igazolták is.

színűek már kevésbé, a zöld színűek pedig már annyira kevés kapcsolattal rendelkeznek, hogy legfeljebb egy kisebb csoportban lehetnek meghatározó elemek.

5.3.2. A komplex jegymérték alapján számított kapcsolatok

A 0,9-nél erősebb **hasonlósági viszonyok alapján 50 csoportot** lehetett elkülöníteni, amelyeket az 5.11. táblázat mutat be. Egy szó akkor tartozik bele egy csoportba, ha legalább a csoport egy tagjához 0,9-es vagy annál nagyobb mértékében hasonló. A **hangkivető szavak csoportjainak mérete és azok jellege nem egyforma**. A szavak közel fele (488 szó) a két legnagyobb csoportba tartozik (*-alom, -elem* végűek), amelyeket gráfok alapján nem elemzek, mert homogén struktúrájukban ezzel a módszerrel nem lehet érdemleges 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 találhatók. Ezeknek a prototípusként funkcionáló szavaknak köszönhető egy-egy kisméretű szóbokor 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 (vö. 6.5. alfejezet). 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*). Az egyes szavak kapcsolatainak száma enyhe pozitív korrelációban van hangkivetésük mértékével ($r(803) = 0{,}136$, $t = 3{,}9$, $p < 0{,}001$), amely összefüggés némileg szorosabb, ha csak a hangkivetést 99%-nál kevésbé követő szavakat nézzük ($r(101) = 0{,}23$, $t = 2{,}33$,

$p < 0,05$)[1]. Ez azt jelenti, hogy egy szóhoz minél több sajátosan viselkedő szó hasonlít jelentős mértékben, annál biztosabb, hogy az ezekből formálható egyedien viselkedő csoport sémáját követi. Ez alapján azt látjuk, hogy nem csak az elemek példánygyakoriságának van szerepe abban, hogy elindulnak-e a változás útján, hanem annak is, hogy hozzájuk mennyi nagyon közeli, hasonlóan egyedi módon viselkedő elem van. A magányosabb szavak hajlamosabbak a csoportjuk általános viselkedésétől eltávolodni, mint ahogy azt a későbbiekben látni fogjuk több példa kapcsán is.

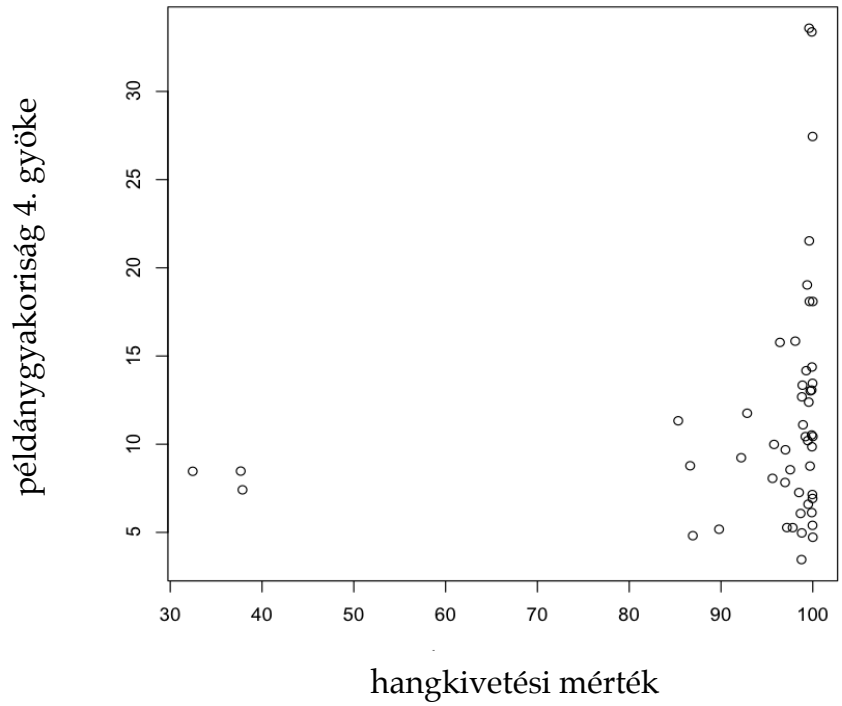
[1] Az összehasonlításból kihagytam az *-alom* végűek csoportját, mivel a Cytoscape 2.6 nem boldogult kapcsolati fokaik kiszámolásával, mert azok túlzottan számosak voltak. Ebben a csoportban egy szóra összesen 126 kapcsolat jutott, míg a megvizsgált 805 szó esetében szavanként 23. Ez alapján valószínűsíthető, mivel az *-alom* végűek csoportja stabilan hangkivető, hogy ezt az összefüggést inkább támogatnák, semmint gyengítenék adataik. 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.

csoport leggyakoribb szava	példány-gyakoriság	szavak száma a csoportban	hangkivetés mértéke típus-alapon	szórás szavak hangki-vetésében	hangkivetés mértéke példány-alapon	csoporton belüli átlagos hasonlóság
figyelem	1272143	222	99,60%	0,03	99,78%	0,902
tartalom	1240569	266	99,89%	0,01	99,97%	0,926
dolog	567310	6	99,98%	0	99,89%	0,964
lélek	214755	17	99,61%	0,01	99,60%	0,911
sarok	131170	55	99,39%	0,03	99,81%	0,882
kapocs	107195	13	99,64%	0,01	99,97%	0,961
titok	107132	11	100,00%	0	99,99%	0,953
haszon	62988	20	98,09%	0,05	99,64%	0,939
teher	61856	8	96,42%	0,1	99,94%	0,961
szobor	42712	12	99,91%	0	99,92%	0,955
tükör	40301	32	99,28%	0,04	99,33%	0,942
izom	32762	25	99,97%	0	99,97%	0,925
karom	31719	14	98,87%	0,02	99,24%	0,892
cukor	29146	58	99,87%	0,01	99,89%	0,916
méreg	28779	44	99,73%	0,01	99,68%	0,922
ajak	25894	5	98,81%	0,02	97,29%	0,932
torony	23529	35	99,56%	0,01	99,72%	0,954
gyomor	19094	7	92,85%	0,19	99,91%	0,964
sátor	16467	11	85,34%	0,17	81,85%	0,930
fogoly	15179	10	98,93%	0,02	99,05%	0,922
kölyök	12258	7	99,88%	0	99,22%	0,961
fészek	11860	30	100,00%	0	100,00%	0,957
vétek	11835	6	99,20%	0,02	99,68%	0,915
csücsök	10837	7	99,45%	0,01	99,93%	0,931
köröm	9950	10	95,78%	0,13	98,73%	0,945
gödör	9436	20	99,91%	0	99,69%	0,948
majom	8792	13	97,02%	0,1	99,53%	0,907
meder	7265	11	92,20%	0,23	88,55%	0,937
lepel	5937	3	86,64%	0,22	98,30%	0,972
kehely	5885	9	99,70%	0	99,08%	0,945
pokol	5349	5	97,54%	0,03	97,87%	0,919
fátyol	5158	7	37,68%	0,25	69,75%	0,961
bajusz	5132	5	32,45%	0,2	36,20%	0,965
ököl	4229	3	95,61%	0,04	99,76%	0,942
öböl	3759	3	96,98%	0,04	98,62%	0,949
vacak	3021	2	37,87%	0,54	75,67%	0,975
selyem	2778	8	98,50%	0,03	99,32%	0,954
mocsok	2608	4	99,95%	0	99,81%	0,940
horog	2297	5	99,97%	0	99,87%	0,967
kölök	1894	4	99,50%	0	99,47%	0,944
eper	1405	5	99,89%	0	99,50%	0,930
szatyor	1361	3	98,67%	0,02	99,78%	0,968
piszok	850	4	99,97%	0	99,88%	0,967
kapor	773	3	97,18%	0,04	94,83%	0,935
fodor	771	9	97,81%	0,04	97,92%	0,954
tulok	719	4	89,79%	0,2	87,90%	0,923
pocok	610	3	98,80%	0,02	96,72%	0,943
kazal	537	3	86,93%	0,05	88,27%	0,973
töbör	498	2	100,00%	0	100,00%	0,983
pecek	143	2	98,77%	0,02	97,90%	0,978

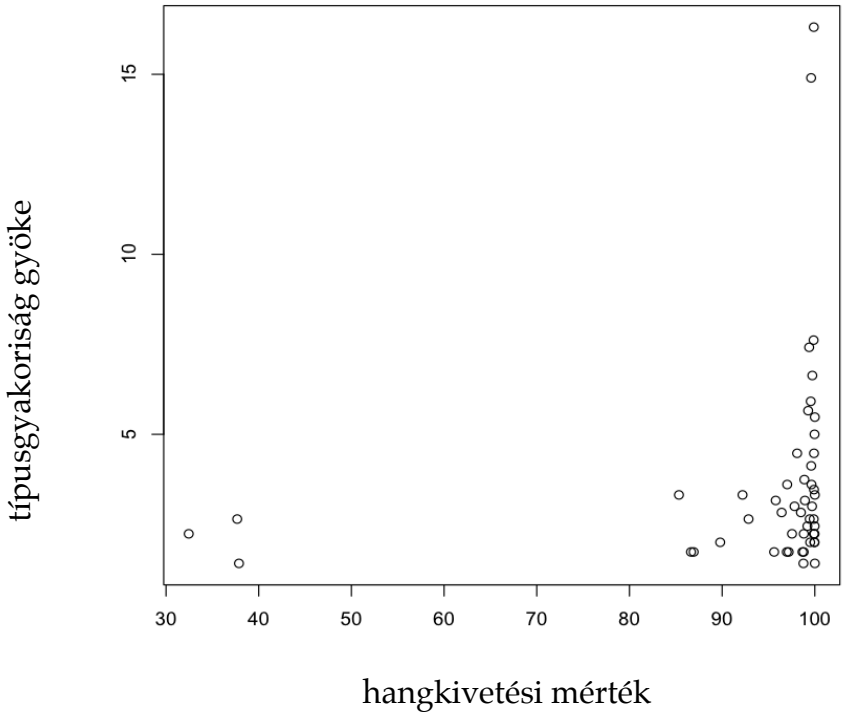
5.11. táblázat: A hangkivető szavak hasonlósági csoportjai a komplex jegymértékkel számítva

Az **ingadozó csoportok típus- és példánygyakoriságuk alapján is a kisebbek közé tartoznak**[1]. Az átlagos csoportméret 21 szó, míg az átlagos ingadozó csoportban 6 szó található. Az átlagos csoport 84 ezer alakelőfordulást számlál, az átlagos ingadozó 10 ezret. A csoporton belüli átlagos hasonlóság mértékében azonban nincs eltérés, az ingadozó szavak csoportjainak értékei hasonlóak a következetesen hangkivetőkéhez. Ez azonban annak tudható be, hogy bizonyos ingadozó csoportok nagyon szorosan összetartanak 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. Érdemes megjegyezni, hogy az igen **nagy méretű csoportok** a kisméretű csoportok értékeinél kisebb, de még így is meglehetősen **nagy átlagos hasonlósággal bírnak**, azaz a hasonlóan viselkedő rendhagyó elemekre jellemző, hogy a csoport összes tagjára hasonlítanak, nemcsak a hozzájuk legközelebbi szavakkal vannak szoros kapcsolatban. A belső hasonlóság minden esetben messze meghaladja a hangkivető főnevek átlagos egymáshoz mért 0,409-es hasonlóságát.

[1] Természetesen vannak a stabilan hangkivető csoportok közt kis példány- és típusgyakoriságúak is (vö. 5.2–5.3. ábrák)



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



5.3. ábra: A hasonlóság alapján kialakított csoportok hangkivetési mértéke a típusgyakoriság függvényében

A komplex jegymérték alapján **21 szónak nincsenek a 0,9-es hasonlósági értéknél közelebbi kapcsolatai**: *ászok, átok, berek, bögöly, boholy, bugyor, bürök, bütyök, cseber, iker, jászol, kebel, koboz, pityer, pöcök, pucor, sulyok, szutyok, takony, tegez, üszök*[1]. E **magányos szavaknak** a *Google Gyakorisági Gyűjtés* 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,* több közülük a 7. fejezet tanúsága alapján legaktívabban élő hangkivető sémával (*-ök* végűek) jellemezhető. A *sulyok* és a *boholy* szavakat érdemes kizárnom áttekintésemből, mert a *sulyok*-nál 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 12,14, azaz a nem ingadozó magányos szavaknak 2,42-szer 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 az ingadozó és nem ingadozó magányos csoportok között. Ha azonban a magányos szavak értékeit összehasonlítom 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 kijelenthetem, 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.

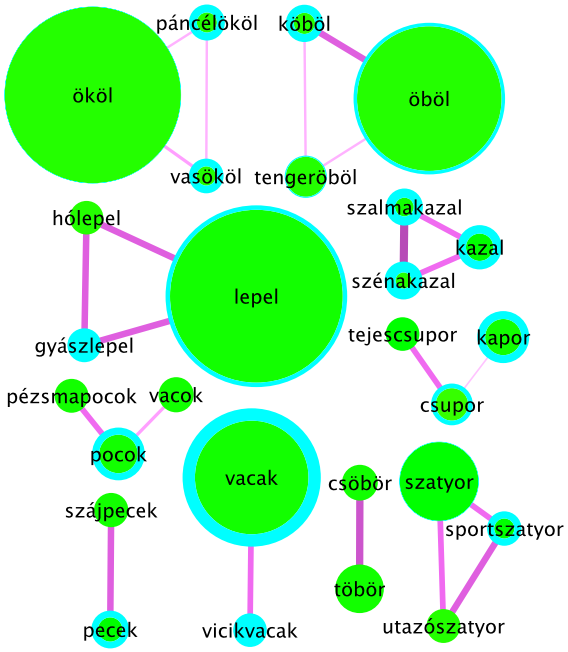
[1] Az algoritmus jellegéből következik, hogy csak 2 szótagos szavak lehetnek magányosak, hisz az összetett szavak esetén az egymáshoz való hasonlóság ennél a küszöbértéknél mindig nagyobb. Ugyanez áll a dominánsan legalább három szótagos *-alom, -elem* végű szavakra is.

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** (5.4. 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 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%)[1]; *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*[2] 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 ebben a legmagasabb az átlagos

[1]Az 5.2.5. alfejezetben adok arra egy lehetséges magyarázatot, hogy mi lehet a *tengeröböl* stabil hangkivető viselkedésének az oka, de egyértelműen ezzel nem magyarázható, hogy a *köböl* miért ingadozik nála jobban. Talán ebben közrejátszhat a *köböl* nála alacsonyabb gyakorisága és nagyobb alaki hasonlósága az *öböl*-höz.

[2] A *TESz*-ben nem tesznek említést erről az alakról. Elképzelhető, hogy a *pocak* analógiás hatására van a *pocok*-nak *pocokja* alakja, mert a *pocak* leggyakoribb alakjai a *pocak* (678), *pocakja* (493), *pocakom* (338) alakok. A komplex jegymérték alapján ez a *pocok*-hoz leginkább hasonlító nem hangkivető főnév a bizonytalan státuszú *vacok* (alakjai a *vacak*-kal keverednek) és a viszonylag stabilan hangkivető *mocsok* (99,5%) után. A két szó kapcsolatát mondókák, és a *pocak-pocok*-hoz hasonló kifejezések is erősíthetik. A komplex tengelymérték szerint viszont leghasonlóbb hozzá a *pocak,* amit a *pocak*-nál ritkább *pöcök* követ, amelynek E.3 birtokosa sem következetesen hangkivető (90,5%-ban hangkivető). A *Szószablya Gyakoriság Szótár* és a *Google Gyakorisági Gyűjtés* alapján az E.3 birtokosukban *-A* helyett *-jA*-val előforduló szavak közé talán a *gyilok, vöcsök, horog, retek, pöcök, pecek, cirok* szavak sorolhatók.

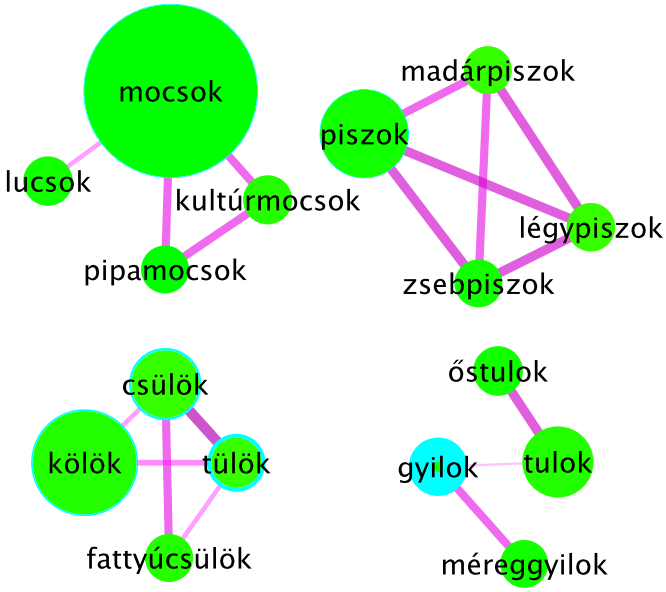
hasonlóság). A *töbör-csöbör* pár tagjai a hangkivetőkre jellemző utolsó két fonémájuknak (9,4-szer jellemzőbb a hangkivetőkre) és *-ö-ö* (4,33-szor jellemzőbb a hangkivetőkre) magánhangzó-szekvenciájuknak köszönhetik stabilitásukat. A *kapor* ingadozása az ékezetmentes *kapóra* alakoknak tulajdonítható, bár tárgyesetben már csak 98,5%-ban követi a hangkivető sémát. A *csupor* enyhe ingadozása viszonylagos magányosságának, valamint a kevésbé hangkivető *-or* végűekkel (hangkivetés átlagos mértéke: 97,6%) való kapcsolatának tudható be.



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

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 az 5.2.3. alfejezetben leírt okokra vezethető vissza. Viselkedésben az 1 alakkal előforduló *méreggyilok* nem követi, amely elsősorban adathiánynak tudható

be, azonban a *Google Gyakorisági Gyűjtés*ben már 12-szer fordul elő a leggyakoribb hangkivetéssel együttjáró toldalékokkal, és ebből 7 alakja nem hangkivető (41,6%-ban hangkivető viselkedés). A *gyilok*-kal szemben a *tulok* nagyobb fokú stabilitását indokolja magasabb gyakorisága (469), egy közeli strukturális pár megléte (*tülök*, komplex tengelymérték alapján mért hasonlóságuk: 0,92), valamint az is, hogy tipikusabb hangkivető szónak számít *-u-o* szekvenciája alapján, amely 22,76-szor többször fordul elő a hangkivetők közt, mint a többi főnévnél. A *Google Gyakorisági Gyűjtés* alapján azonban már a csoportnak ez a fele is elindult az ingadozás útján (*tulok* 93,09%, *őstulok* 99,69%).

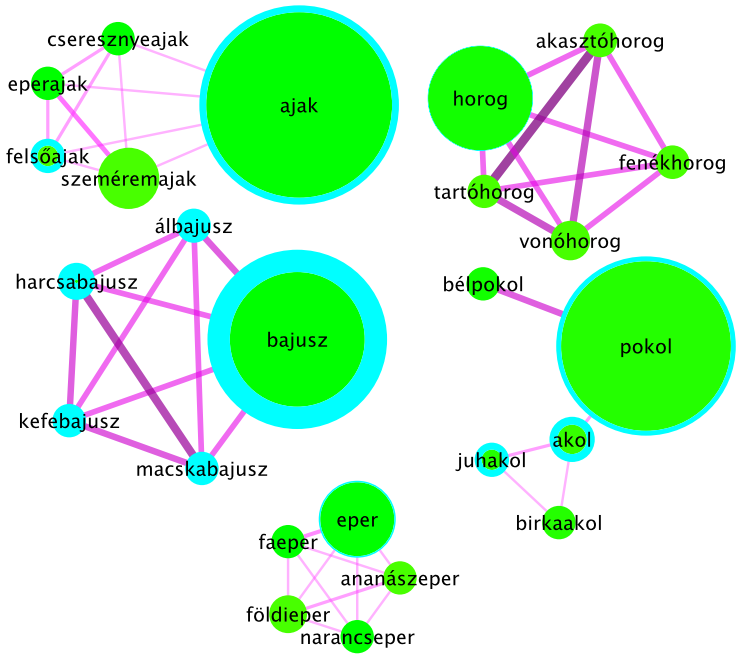


5.5. ábra: 4 elemű részgráfok a komplex jegymérték alapján

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[1]. A *-bajusz* végűek közül

[1] Az ilyen esetekben az elmélet azt jósolná, hogy a ritkább alakok előbb kezdtek el ingadozni. Ezt ebben az esetben feltételezésként elfogadhatjuk, de bizonyítékunk nincs rá.

egyedül a 111 előfordulással rendelkező *harcsabajusz* viselkedik kevésbé hangkivető módon (57%), mint a nála gyakoribb *bajusz* (4985, 35,9%)[1]. 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.



5.6. ábra: 5 elemű részgráfok a komplex jegymérték alapján

A szerkezetileg gyengébb[2] ***pokol-akol* csoportban** is találunk kevésbé hangkivető alakokat. Az ingadozó *juhakol*, *akol* viselkedését[3] magyarázhatja, hogy az

[1] Ez a különbség a *Google Gyakorisági Gyűjtés*ben is megvan (*bajusz* 240375, 19,8%; *harcsabajusz* 4357, 28,3%), ezért semmiképp sem magyarázhatom adathiánnyal, mérési hibával.

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

[3] A *birkaakol* csak 5 alakkal fordul elő, és stabilan hangkivető. Viselkedését az 5.2.5. alfejezetben tárgyaltam.

akol VCVC szerkezetével nem tipikus hangkivető szó. Ezt erősíti meg, hogy a komplex tengelymérték szerint összesen 8 hangkivető szó van a hozzá leghasonlóbb 100 szó közt (bővebben 6.3. alfejezet). Sajátosan viselkedik az ***ajak* csoport,** mivel a leggyakoribb elem hangkivetési mértéke a legkisebb, még akkor is, ha a kis gyakoriságú *eperajak* (4), és *cseresznyeajak* (6) szavaktól eltekintek vizsgálatában. Az *ajak* hangkivetésének alacsonyabb mértéke (hangkivetési mérték: 96,85%) annak köszönhető, hogy a szó hangkivető viselkedése sosem szilárdult meg[1]. Emellett tanúskodik, hogy már 1766-ból van *ajjaka* (Mátyus 1766: 396) előfordulása, amely változat (*ajaka* hangkivetés nélkül) a választékos beszédben is megmaradt. Ezzel találkozhatunk lexikonokban (*Pallas Nagy Lexikona, Kislexikon*), vagy akár költői művekben is (Kosztolányi Dezső művei, Jannus Pannonius, Guillame Apollinaire fordítások stb.). Ilyen választékos szövegek a *Szószablya Korpusz*ban is lehetnek, de beazonosításuk és hatásuk felmérése az *ajak* hangkivetési mértékére 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.

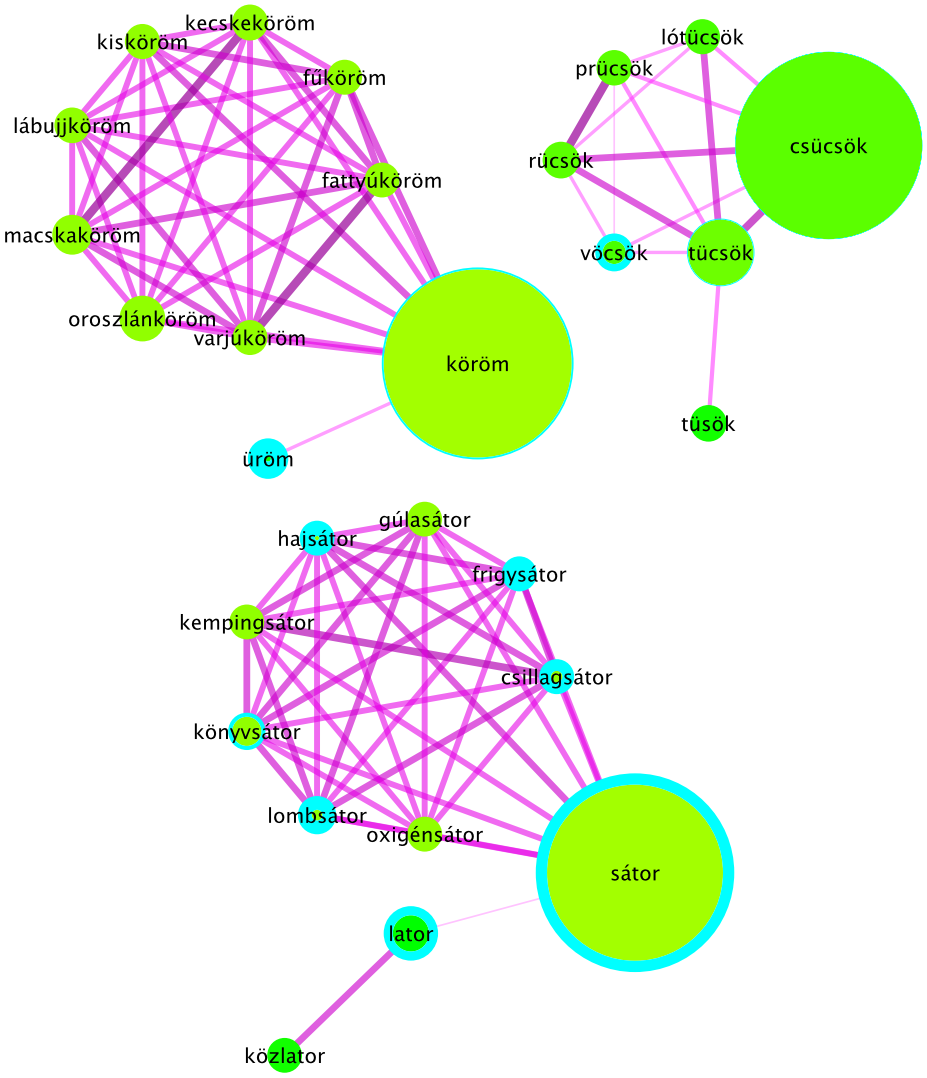
A továbbiakban csak azokat a **részgráfokat** mutatom be **ábrákon**, amelyek **szerkezetükben vagy csomópontjaik viselkedésében érdemben különböznek** a már leírt esetektől. A *dolog* (6 elem), *vétek* (6 elem, a kevésbé hangkivető *retek*-kel), *gyomor* (7 elem), a *kölyök* (7 elem), a *fátyol* (7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

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

[1] Ez a felvetés az összes hangkivető szóval kapcsolatban elhangozhatna. Amíg azonban nem tudjuk az ellenkezőjét bizonyítani, addig célszerű elfogadnunk a témáról legalaposabban nyilatkozó Bárczi és mtsai-nak (1967) azon megállapítását, miszerint a hangkivetéses paradigma kialakulása az ómagyar korban lezárult, és az ingadozás csak ezután indult el.

alkalmas prototípus szerepre, így a csoport némileg instabil „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 tárgyesetű alakjaiknak köszönhetően. Elvárásaimmal 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, amelyet részletesen már az 5.2.5. alfejezetben tárgyaltam. A 10 elemű ***fogoly-bagoly* csoport** két alcsoportra bomlik, csupán csak a *fogoly* és a *bagoly* van kapcsolatban. A két részcsoportot korábban már elemeztem (5.2.5. alfejezet).

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 5.7. 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 (bővebben 5.4.2. alfejezet). A csoporton belül az elvártnál jobban hangkivető *könyvsátor* javarészt a hangkivetéssel jobban együttjáró (bővebben 5.4.2. alfejezet) 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).

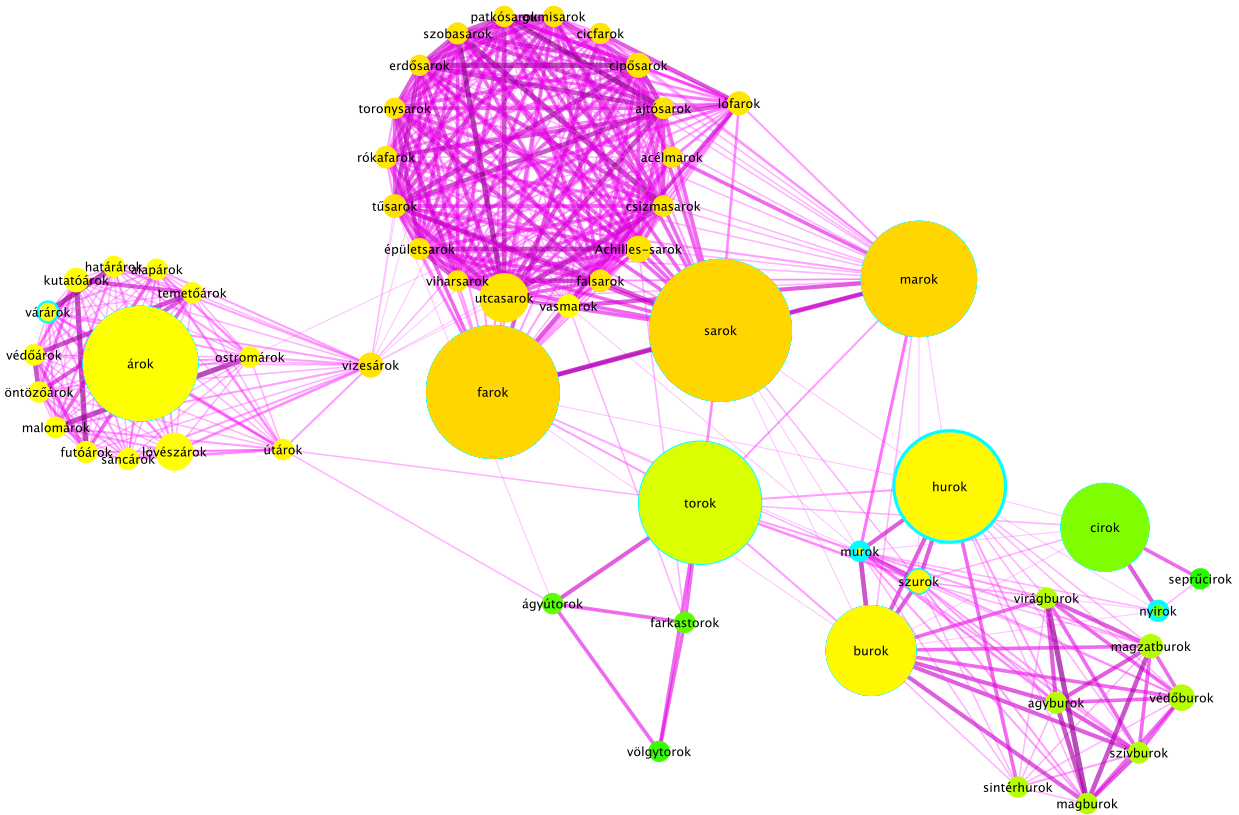


5.7. ábra: Kis elemszámú részgráfok leszakadó elemei a komplex jegymérték alapján

A 14 elemű ***karom* csoport** esetében a magasabb elemszám és az *-alom* véghez való hasonlóság még a helyenként gyenge belső kapcsolatok ellenére is megóvja szavait az ingadozástól. Egyedül a *nőszirom* (93,8%, 64) kevésbé hangkivető benne, amely a *-szirom* végűek alcsoportjához jelentésében gyengébben kapcsolódik, mivel ez egy virágfélére utal, a többiek pedig 'virágtakaró'-t jelentenek. A közepes méretű csoportokból a *lélek* (17 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 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 *haszon* csoportba (20 elem) a *haszon*, a *tejhaszon* és a *-vászon* végű szavak tartoznak, amelyek viselkedését az 5.2.5. alfejezetben részletesen tárgyaltam.

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 heterogenitása 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 hangkivetők[1].



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

[1] A *várárok* „ingadozása" egy *várárokon* előfordulásnak tudható be.

5.3.3. A komplex tengelymérték alapján számított kapcsolatok

A komplex tengelymérték alapján létrehozott gráf, habár sokban hasonlít a komplex jegymérték segítségével készítetthez, mutat bizonyos esetekben olyan **eltéréseket** is, amelyeket **érdemes** a továbbiakban **áttekintenünk**. Az ábrákon a megjelenítés módja azonos a korábbiakéval, de a hasonlóság alsó küszöbértékét 0,85-re módosítottam. A 0,9-es értéket szükséges volt lejjebb szállítani, mivel még így is sokkal kevesebb kapcsolatot kaptam, mint a másik gráf esetén[1]. A küszöbérték további csökkentését azonban már nem tartottam kívánatosnak, mert ebben az esetben a struktúrák teljesen összefüggenének, így a csoportok elemzése nehezebbé vagy kevésbé informatívvá vált volna.

Fel kell tennünk a kérdést, hogy **mennyire jól összehasonlíthatóak az eltérő küszöbértékkel számított gráfok struktúrái.** A komplex tengelymérték 0,85-ös értéke hogyan viszonyul 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 feltételezés még algoritmusaim viszonylagos jósága esetén is, hogy azonos értékek azonos mértékű hasonlóságot fejeznek ki[2]. Mivel azonban az összehasonlítás során nem a konkrét értékekre hagyatkozok, hanem csak az azok által hasonló módon definiált viszonyokra, a két gráfot összehasonlíthatónak veszem.

---

[1] 24904 szemben a komplex jegymérték alapján készített gráf 51655 kapcsolatával. A potenciális kapcsolatok száma $1092^2/2 = 596232$.

[2] Azonban a valamely algoritmusom 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 különbözőnek ítélnék meg.

csoport leggyakoribb szava	példány-gyakoriság	szavak száma a csoportban	hangkivetés mértéke (típus-alapon)	szórás szavak hangki-vetésében	hangkivetés mértéke példány-alapon	csoporton belüli átlagos hasonlóság
tartalom	1309962	330	99,60%	0,03	99,93%	0,750
figyelem	1274921	230	99,56%	0,03	99,78%	0,782
dolog	743093	88	97,74%	0,1	99,74%	0,697
lélek	249463	62	99,68%	0	99,58%	0,762
szobor	189088	176	97,10%	0,11	95,80%	0,702
titok	107982	15	99,99%	0	99,99%	0,817
kapocs	107195	13	99,64%	0	99,97%	0,861
haszon	63625	20	94,41%	0,17	98,90%	0,814
teher	61856	8	96,42%	0,1	99,94%	0,884
fészek	23361	33	99,86%	0	99,82%	0,841
gyomor	19094	7	92,85%	0,19	99,90%	0,883
fogoly	13160	8	91,87%	0,19	98,02%	0,837
szirom	10325	29	99,76%	0,01	99,95%	0,789
pehely	5650	4	99,33%	0	99,04%	0,904
fátyol	5158	7	37,68%	0,25	69,78%	0,875
bajusz	5132	5	32,45%	0,2	36,20%	0,876
hadifogoly	2241	3	99,94%	0	99,82%	0,907
szeméremajak	819	2	100,00%	0	100,00%	0,943
kölyök	527	6	100,00%	0	100,00%	0,887
szénakazal	227	2	84,92%	0,05	84,58%	0,959
horog	209	4	100,00%	0	100,00%	0,934
vadászsólyom	47	2	98,48%	0,02	97,87%	0,939
kultúrmocsok	36	3	100,00%	0	100,00%	0,866
sportszatyor	36	2	98,08%	0,03	97,22%	0,933
jégcsapretek	29	2	100,00%	0	100,00%	0,941
burgonyapehely	19	2	100,00%	0	100,00%	0,952
ananászeper	3	2	100,00%	0	100,00%	0,938

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

A komplex tengelymérték által létrehozott egyedi részgráfok vizsgálata előtt érdemes áttekintenünk, hogy az egyes **csoportok** milyen **jellemzőkkel** bírnak (5.12. 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- ($r(25) = -0,6$, $t = -3,84$, $p < 0,01$) és a típusgyakorisággal ($r(25) = -0,67$, $t = -4,03$, $p < 0,001$) is, azaz a strukturális hasonlóságoknak nagyobb súlyt adó komplex tengelymérték alapján látható, hogy a csoportok méretének növekedésével párhuzamosan a csoportok belső szerkezeti

hasonlósága lazul. Ez azt jelenti, hogy ha a szerkezeti hasonlóságnak adunk nagyobb fontosságot a végekkel szemben, mint ahogy a komplex tengelymérték teszi, akkor a csoportok összetartásában a wittgensteini értelmű családi hasonlóság előtérbe kerül (elég a családból valakire nagyon hasonlítani, nem kell mindenkihez közel lenni). Ugyanakkor még a nagy csoportokban is a csoporton belüli átlagos hasonlóság mértéke jelentősen meghaladja a hangkivető főneveknek a komplex tengelymértékkel egymáshoz mért hasonlóságának 0,51-es 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 ($r(104) = 0{,}246$, $t = 2{,}51$, $p < 0{,}05$). Itt hasonló hatásokat láthatunk, mint amiket a komplex jegymértékes összehasonításoknál megfigyelhettünk.

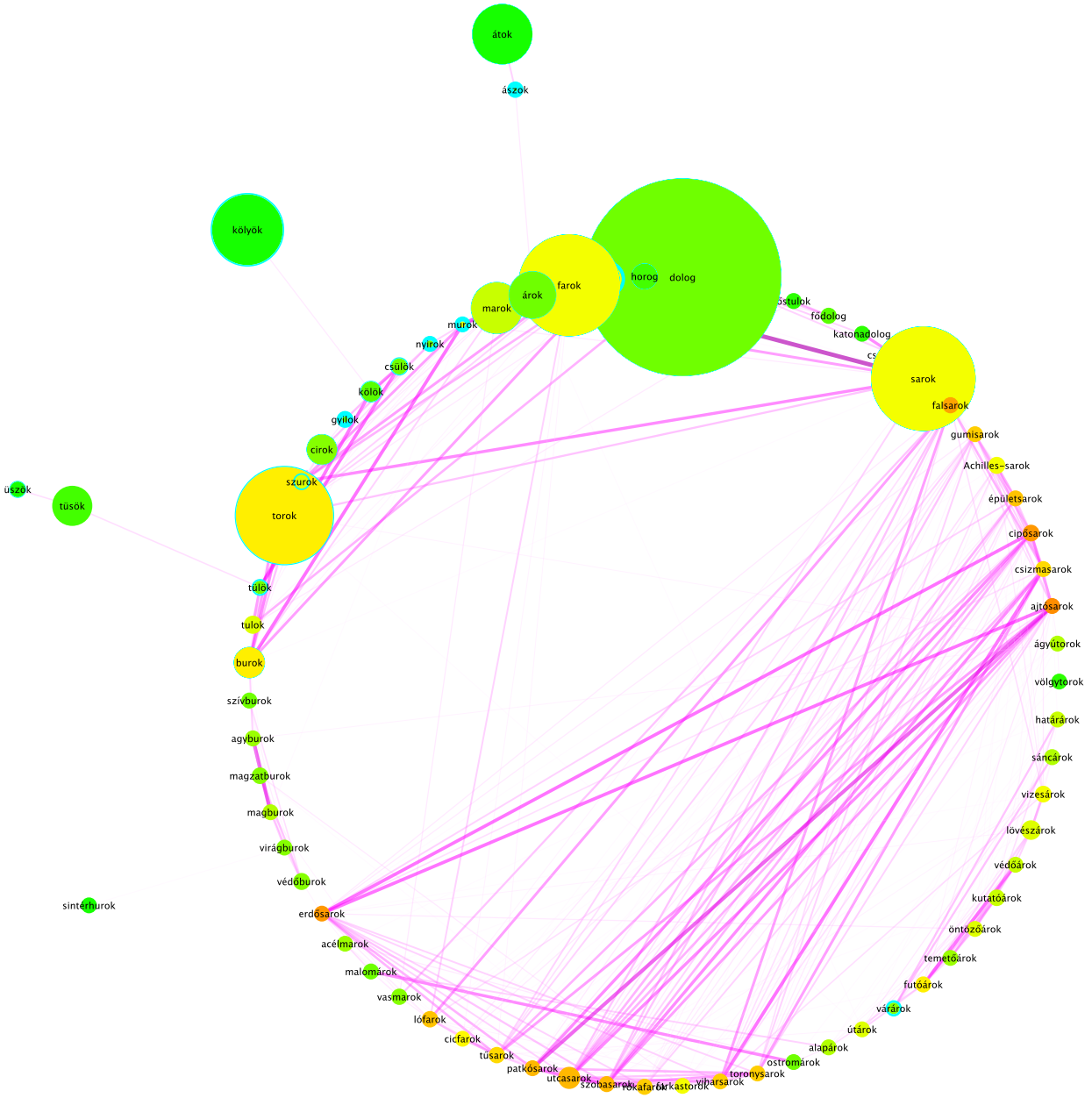
A komplex tengelymérték segítségével készített gráfon **27 szó nem kapcsolódik egyetlen egy részgráfba** se: *ajak, birkaakol, boholy, csalétek, eperajak, faeper, fattyúcsülök, felsőajak, izom, juhakol, kazal, kerecsensólyom, koboz, kukoricapehely, méreggyilok, misekehely, páncélököl, pecek, seprűcirok, szájpecek, takony, tegez, tejhaszon, tengeröböl, vasököl, vicikvacak, virágkehely.* 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 hasonlítanak: pl. *vadászsólyom* : *kerecsensólyom*. Egyedüli stabilan hangkivető alapszó az *izom* (26810), amelyet szerkezeti magányossága ellenére kiugró gyakorisága tart a hangkivető mintában, lényegében ennek egyedi alakjait a beszélők még a nyelvelsajátítás korai szakaszában memorizálják.

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[1]. A kevésbé hangkivetők átlagos hasonlósága a többi hangkivető főnévhez 0,43, a hangkivetéses

[1] Az összehasonlításból a *boholy*-t ezúttal is kihagytam, valamint a kevesebb, mint 25 előfordulással rendelkező szavakat, mert ezek esetében a hangkivető séma követése adathiánynak is betudható.

sémát következetesen követőké 0,45, amely átlagok szignifikáns mértékben eltérnek ($t(1091) = -8,81$, $p<0,001$). A hangkivetéses mintának megfelelő magányos szavak 0,45-ös értéke szintén szignifikánsan különbözik az összes hangkivető főnév egymáshoz számított hasonlóságának 0,51-es átlagértékétől ($t(1091) = 46,07$, $p<0,001$). Ebből következik, hogy a szegényes közeli kapcsolatokkal rendelkező hangkivető főnevek kevésbé hasonlítanak a többi hangkivető 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 13,8 0,7-es hasonlósági értéknél szorosabb kapcsolat jut, míg a következetesen hangkivetőkre szavanként 17,3. Ha csak a 0,8-as hasonlósági értéknél közelebbi kapcsolatokat vesszük, akkor a különbség élesebb. A kevésbé hangkivető főnevek átlagosan 1,3 ilyen kapcsolattal rendelkeznek, a következetesen hangkivetők viszont 2,3-mal.

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 sajátságát a korábban részletesen bemutatott *sarok-torok-farok* részgráf párját megmutató 5.9. ábrán is megfigyelhetjük.



5.9. ábra: -COC_{vel} részgráf a komplex tengelymérték alapján

A komplex tengelymérték gráfjában a **strukturális hasonlóság** a véghasonlóságnál 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:

(2)

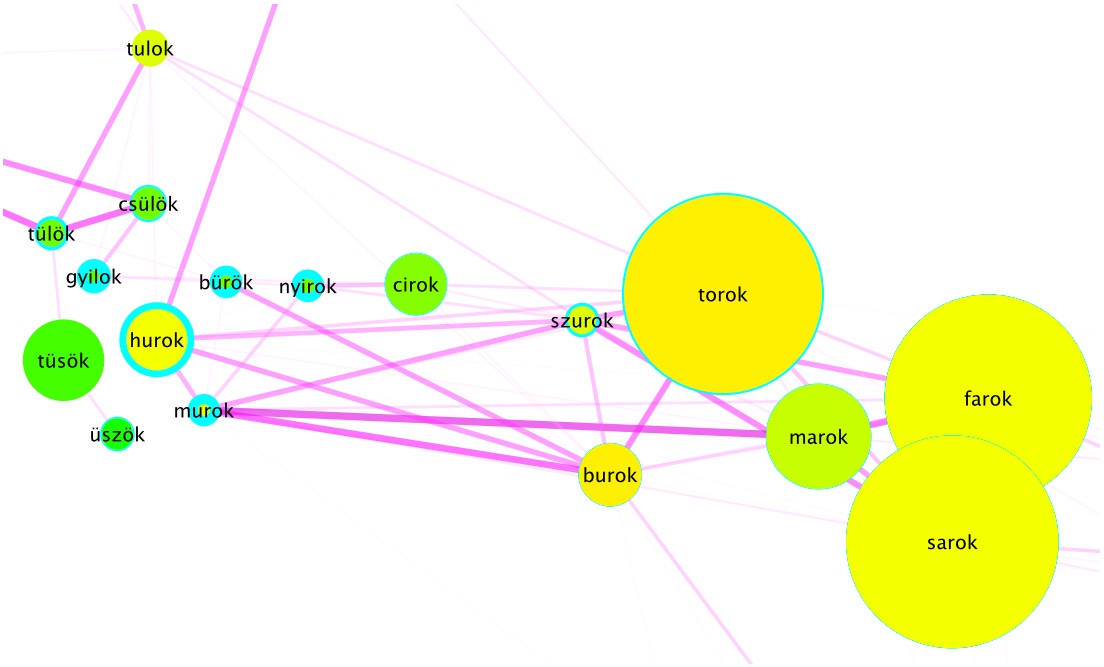
dolog	566770
sarok	48702
farok	34254
torok	19524
marok	7369

(3) prototípus séma: $C(a/o)C_{liq}oC_{vel}$

A **több, nagy gyakoriságú elem** viszonylag **stabil hangkivető viselkedést biztosít** a formailag heterogén csoportnak, amelynek 0,697-es csoporthasonlósági 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 olyan a hangkivető sémát kevésbé követő szó[1] viszonyát is megfigyelhetjük, amelyek viselkedése esetlegesnek tűnt az 5.3.1. alfejezetben. Az 5.9. ábra segítségével a szerkezeti szempontból magányosabb *ászok* kevésbé hangkivető viselkedését tudjuk értelmezni, amely a gyakoribb *átok* és *árok* közt helyezkedik el. A csoportot némileg más elrendezésben megmutató részleten (5.10. ábra) 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 $-V_{high}C_{cont}Ok$ sémával ragadhatjuk meg, amely különbözik a prototipikus szavak sémájától, ugyanis bizonyos pontjaiban specifikusan más (zárófonémának a *-g*-t nem fogadja el, utolsó előtti magánhangzója zártabb) vagy annál általánosabb (utolsó magánhangzónak elfogad *ö*-t is, utolsó előtti mássalhangzója bármilyen folyamatos fonéma lehet). 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

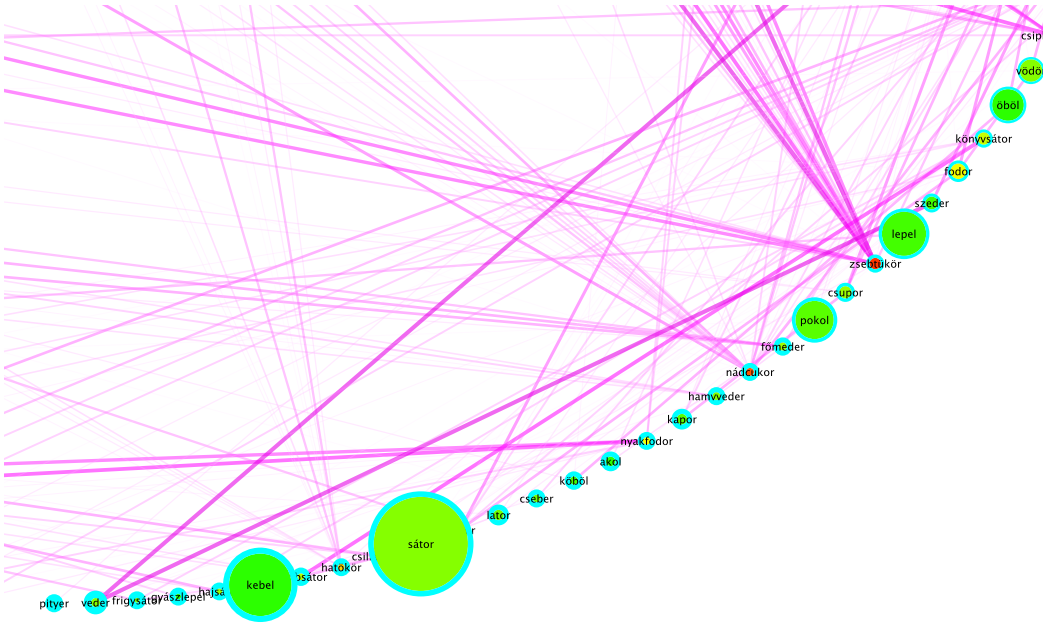
[1] A *torok* minimális ingadozása a *török* ékezetmentes alakjaival hozható kapcsolatba.

többi szóra. Ezt láthatjuk a többiektől elütő, szegényes kapcsolataikra utaló zöld színükből is.



5.10. ábra: -COC_{vel} részgráf kevésbé hangkivető szavai a komplex tengelymérték 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%, csoportátlag: 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, azaz a kevésbé hangkivető módon viselkedő szavak ezúttal is gyengébben, kevesebb szállal kapcsolódnak csoportjukhoz. A csoporttagokhoz való gyenge szerkezeti hasonlóság és a kevésbé hangkivető viselkedés kapcsolatát még szembetűnőbb módon az 5.11. ábra jeleníti meg a -$V(:)C_{plos}VC_{liq}$ végű szavak esetén (*szobor* csoport).

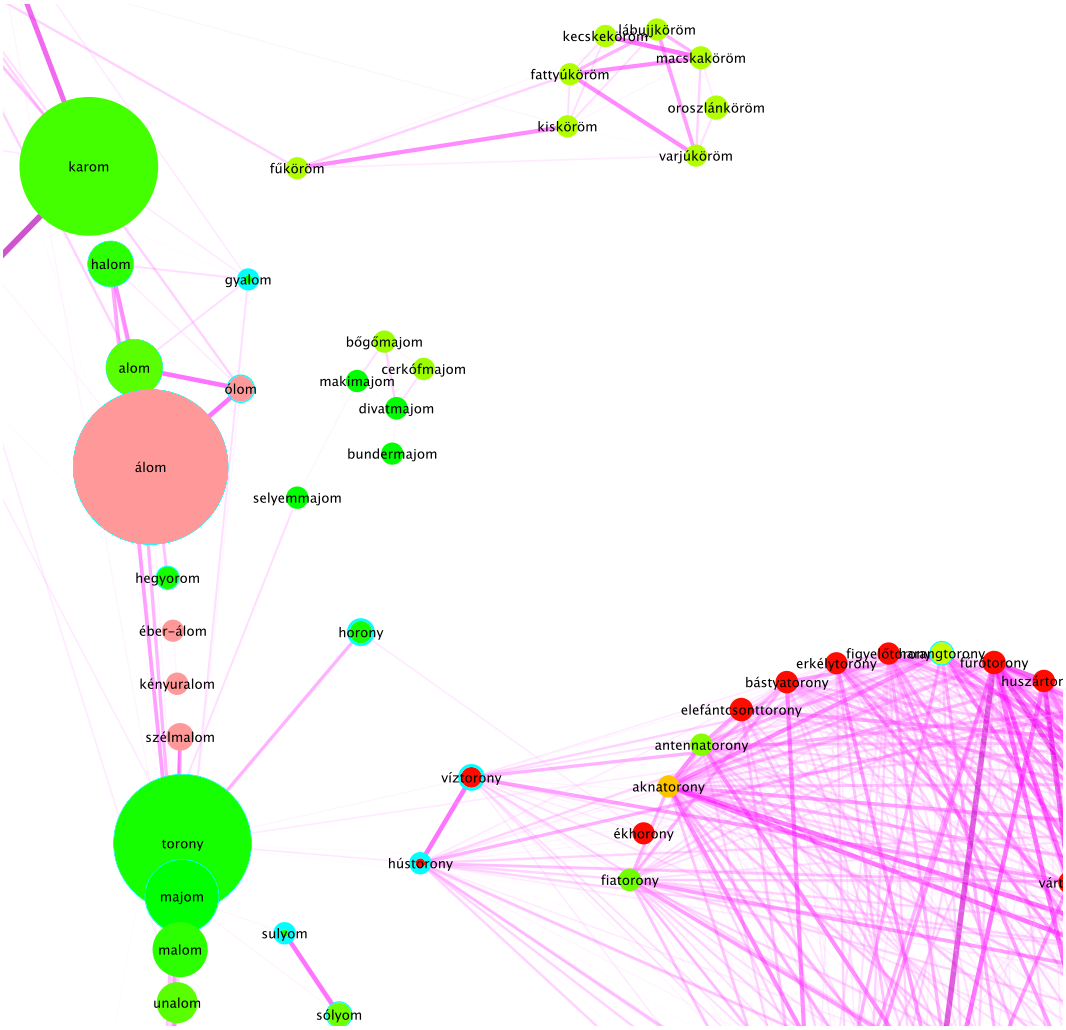


5.11. ábra: A kevésbé hangkivető szavak többsége gyenge kapcsolatokkal rendelkezik a *-V(:)C$_{plos}$VC$_{liq}$* csoportban (zöld szín, vastagabb keret).

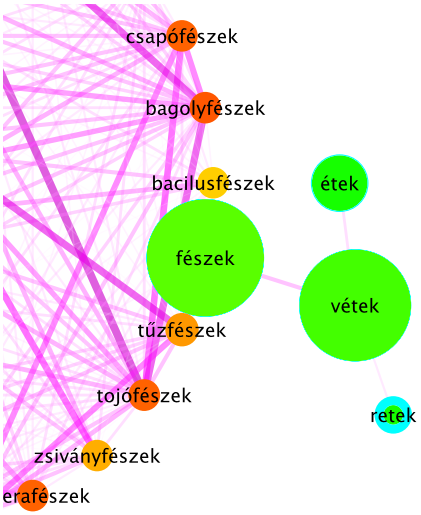
Az 5.12. ábra segítségével a kisebb, de viszonylagosan stabil, szerkezetileg egységesebb csoportok és az *-alom* végűeket is tartalmazó heterogén gyűjtőcsoport[1] kapcsolatát figyelhetjük meg. A csoportokat összekötő szavakból azonban több, mint az *ászok* esetében tapasztaltuk, már kevésbé követi a 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*[2] 80%, *horony* 98,9%). A *fészek* csoportjában (5.13. ábra) hasonlóan a struktúra szélére szorult *retek* (608, 95,8%) viselkedik kevésbé hangkivető módon, mint ahogy azt a *sulyom, üröm, lator* esetében láttuk korábban.

[1] 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: pl. *üröm-torony.*

[2] 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.



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



5.13. ábra: A *retek* gyenge kapcsolódása a *fészek* csoporthoz a komplex tengelymérték alapján

5.3.4. Természetes osztályok alapján számított kapcsolatok

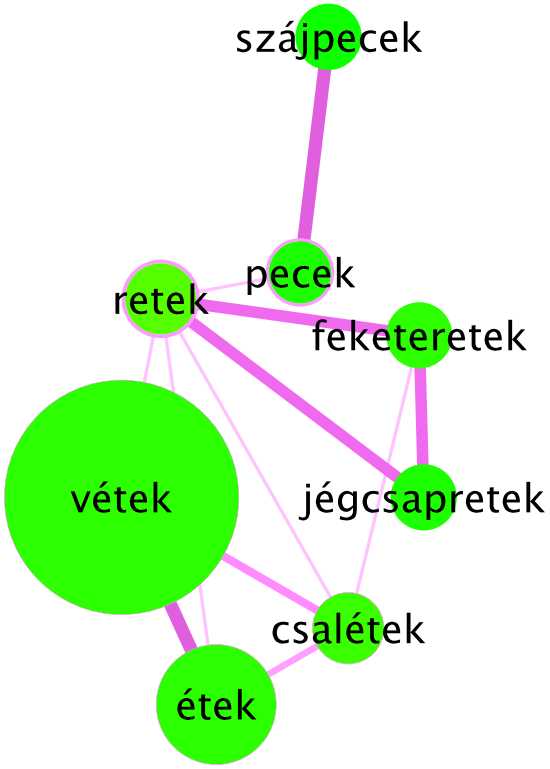
A **komplex jegymérték és a természetes osztályok alapján számított gráfok közt a különbség csekély**, ami arról tanúskodik, hogy a szavak összehasonlításában a fonémák hasonlóságának kisebb szerepe van, mint annak, hogy milyen módon súlyozzuk vagy vetjük össze a fonémák viszonyait. Ezt a feltételezést teszteléseim is igazolják a 6. fejezetben. A két gráf nagyfokú hasonlósága miatt a természetes osztályok által számított kapcsolatok alapján meghatározott gráf jellemzőit és néhány részgráfját csak röviden ismertetem.

A természetes osztályok esetében ismételten a **0,9-es hasonlósági küszöbértéket** alkalmaztam, mivel a szószintű hasonlóság számítása azonos a komplex jegymértékével. A teljes gráfban e küszöbérték mellett **61799 kapcsolatot** találtam. A kapcsolatok magas számából következően a természetes osztályok alapján számított gráfban csak 6 magányos szót lehet azonosítani: *boholy* (100%), *bögöly* (45,5%), *bugyor* (99%), *koboz* (63,4%), *takony* (98,5%), *tegez* (23,9%), amelyek a *boholy* kivételével legalább mérsékelten nem követik a hangkivető mintát.

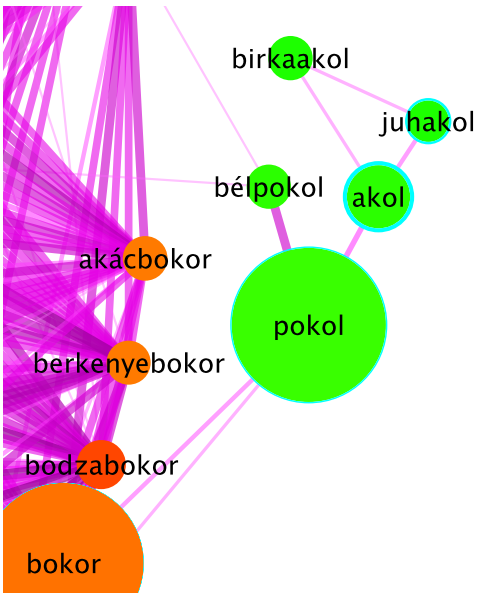
A **kis elemszámú részgráfokban** nem lehetett olyan struktúrákat felfedezni, amelyeket a korábbiakban ne figyeltünk volna már meg. Párban csupán csak a *kebel* (67,3%) és a *cseber* (90,9%), illetve az *ászok* (26,4%) és az *átok* szerepelnek, amelyekből egyedül az *átok* következetesen hangkivető, vélhetőleg magas gyakoriságának (9058) köszönhetően[1] az *izomhoz* hasonlóan. A 9 elemű ***vétek* csoport** (5.14. ábra) esetében is megfigyelhetjük, hogy kevésbé hangkivető módon viselkedik a központi elemtől távolabb eső *retek* (95,9%, 604) és *pecek* (97,5%, 122), amelyek kapcsolatát az eddigi gráfokban nem tudtuk felismerni. Igaz, a kevesebb előfordulással rendelkező *szájpecek*, *feketeretek* és *jégcsapretek* még követik a hangkivető sémát a *Szószablya Gyakorisági Szótár*ban, de a *szájpecek* (98,4%) és a *feketeretek* (94,2%) szuperesszívusza a *Google Gyakorisági Gyűjtés*ben már mutat ingadozást. A *retek* kevésbé hangkivető

[1] A *kebel* 10189 előfordulása ellenére az 5.2.3. alfejezetben bemutatott okok miatt kevésbé hangkivető, amit érdemes kiegészíteni azzal, hogy az *átok* esetében a természetes osztályok alapján számítva a 100 legközelebbi szóból 42 hangkivető, a *kebel* esetében csak 7.

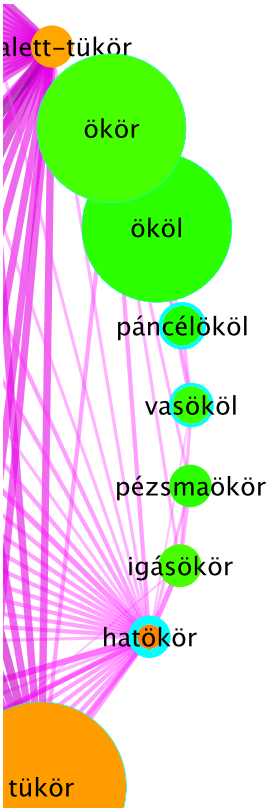
viselkedésének okát az 5.3.3. alfejezetben tárgyaltam. Az 5.15. és az 5.16. ábrákon megfigyelhetjük, hogy két kisebb csoport miként kapcsolódik egy-egy nagyobb részgráfba. Zöld színük utal kapcsolataik szegényességére, amivel összefüggésbe hozható, hogy ezekben a kicsi, gyengén kötődő csoportokban több szó sem követi kizárólagosan a hangkivető sémát.



5.14. ábra: a *vétek* csoport felépítése a természetes osztályok alapján



5.15. ábra: A *pokol* szó kapcsolódása egy nagyobb részgráfba a természetes osztályok alapján



5.16. ábra: Az *ökör-ököl* család a *tükör* részgráfjában a természetes osztályok alapján

5.3.5. Alapszavak alapján kialakítható csoportok (összetett szavak)

A hasonlósági csoportokkal gyakran átfedésbe kerülve az egyes alapszavak és a belőlük képzett összetett szavak is csoportokat alkotnak. Ezek többségét már az egyes részgráfoknál, illetve az 5.2. alfejezetben részletesen tárgyaltam, de röviden össze szeretném foglalni viselkedésüket, hogy jobban lássuk, az **összetett szavak viselkedése mennyire azonos alaptagjukéval** vagy mennyire eltérő tőle, illetve hogy **az eltérően viselkedő szavakat tartalmazó csoportoknak vannak-e egyedi jellemzőik**. Szóanyagomban összesen 129 kisebb-nagyobb összetettszó-csoportot találtam. Ezekből 33 mindösszesen egy alapszót és egy összetett szót tartalmazott. Egy alapszó köré csoportosuló szóbokor átlagosan 7,25 tövet tartalmazott (szórás: 7,49) az alapszót is beleszámítva. A csoportok átlagos példánygyakorisága: 30250 (szórás: 67168). A legnagyobb szóbokorral a *védelem* (39), a *forgalom* (34), a *torony* (33), a *fészek* (30), a *tükör* (28), a *bokor* (27), a *cukor* (25), az i*zom* (25) szavak rendelkeznek.

Elméletem azt jósolja, hogy a **kisebb gyakoriságú elemeknek előrébb kell járniuk a regularizálódás útján**. Mivel egy alapszó és a hozzá tartozó összetett szavak mind fonológiailag (legalábbis jobb oldaluk szempontjából), mind jelentésükben nagy hasonlóságokat kell, hogy mutassanak, elviekben alkalmasabbak arra, hogy pusztán gyakoriságnak betudható hatásokat megfigyelhessünk rajtuk. Összesen 22 olyan szót találtam[1], amelyeknek érdemlegesen kisebb[2] volt a hangkivetési mértéke az alapszóhoz viszonyítva. Ez magasabb érték, mint az ellenkező irányba kilengő szavak (*fülesbagoly, gyöngybagoly, hamvveder, harcsabajusz, hegyorom, hóbagoly, juhakol, könyvsátor, macskabagoly, szeméremajak*) száma (10). Fontos azonban azt is megjegyeznünk, hogy az

[1] Vizsgálatomban a 25-nél ritkábban előforduló szavakat a csoport méretéhez hozzáadtam, de nem számolok viselkedésükkel, mivel ilyen kis gyakoriság esetén adataim nem eléggé megbízhatóak.

[2] Az összehasonlítás szempontjából akkor vettem érdemlegesen kisebbnek vagy nagyobbnak egy összetett szó hangkivetési mértékét az alapszavához viszonyítva, ha hangkivetési mértéküknek a különbsége meghaladta a 0,04/(alapszó hangkivetési mértéke+összetett szó hangkivetési mértéke) mértéket, ha a hangkivetési mértéket 0-1-ig terjedő skálán fejezem ki. Nincs olyan alapszó-összetett szó párom, amiben mind a kettő hangkivetési mértéke 0 lenne, így az osztó sem lehet soha 0. Összesen 8 esetben kaptam 1-nél kisebb értéket a két szó hangkivetési mértékének összegére, a legkisebb 0,396 volt.

alapszavuknál kisebb hangkivetési mértékkel rendelkező szavaknál a különbség átlagosan 10,2%, míg az alapszavuknál magasabb hangkivetési mértékkel rendelkezőknél az átlagos különbség 15%. A két érték eltérése a Welch-próba alapján nem szignifikáns. A két érték közti különbség azzal magyarázható, hogy több szó kezd el egyre kevésbé hangkivető módon viselkedni, mint az alapszavuk, de ezek kevésbé hajlamosak elszakadni az alapszavuktól, azaz csak korlátozottan tudnak autonómak lenni. Nagyobb hangkivetési mértékbeli távolságra csak a *fátyol, bajusz* származékai alapján láthatunk példát. Az ellenkező irányú eltérés (az alapszónál konzervatívabb hangkivető viselkedés) esetén a hangkivető viselkedést bizonyos tényezők tartósan konzerválhatták, és így akár nagyobb eltérések is lehetségesek. Ezeknek esetleges magyarázatait korábban már részletesen tárgyaltam.

	átlagos alapszó-összetett szó pár	**összetett szó hangkivetési mértéke kisebb**	**összetett szó hangkivetési mértéke nagyobb**
típusgyakoriság	16,7 (szórás: 10,7)	8,68*** (szórás: 8,41)	5*** (szórás: 1,8)
példánygyakoriság	50476 (szórás: 79023)	19167** (szórás: 40627)	5995*** (szórás: 8344)
alapszó mennyivel gyakoribb	398,5 (szórás: 762,7)	193* (szórás: 311)	53,3*** (szórás: 63,2)

5.13. táblázat: Alapszó-összetett szó párok tulajdonságai, ha az összetett szó legalább 25 előfordulást számlált

az átlagos alapszó-összetett szó párok átlagaitól való eltérés szignifikancia szintjei

***= $p < 0,001$

**= $p < 0,01$

* = $p < 0,05$

Az 5.13. táblázat azt foglalja össze, hogy **milyen jellemzőkkel bírnak az olyan alapszó-összetett szó párok**, ahol az összetett szó legalább 25 előfordulást számlált. Az egyes típusoknál („átlagos alapszó-összetett szó pár", „összetett szó hangkivetési mértéke kisebb", „összetett szó hangkivetési mértéke nagyobb") minden egyedi párt

külön számoltam, azaz, ha egy szóbokorban 2 olyan pár is volt, ahol jelentősen kisebb az összetett szó hangkivetési mértéke, mint az alapszóé, akkor e pár csoportjának típus - és példánygyakoriságát is kétszer vettem az átlagok kiszámításában, ezáltal lényegében ennek a csoportnak nagyobb súlyt adtam, hisz ezek az értékek azonosak voltak (az összetett szavaknak az alapszóhoz viszonyított arányuk viszont egyedi volt).

Az alapszóhoz viszonyítva kisebb illetve magasabb hangkivetési mértékkel rendelkező összetett szavak párjai szignifikánsan nem különböznek egymástól egyetlen egy paraméterükben sem, azaz csoportjaik típusgyakorisága, példánygyakorisága és az alapszó-összetett szó gyakoriságának az aránya mintám alapján nem tekinthető eltérőnek. **Mind a két csoport szignifikánsan különbözik azonban azoktól az alapszó-összetett pároktól, ahol nincs jelentős különbség a hangkivetés mértékében**. Az eredmények alapján azt állíthatom, hogy egy összetett szó viselkedésében, akkor tud elszakadni az alapszótól, ha a csoportja az átlagosnál alacsonyabb típus- és példánygyakoriságú, illetve az alapszó az átlagosnál kevésbé gyakoribb, mint a kérdéses összetett szó. Mint látjuk azonban a szórás igen nagy, ha az alapszónál kisebb hangkivetési mértékkel rendelkező összetett szavak csoportjait vizsgáljuk meg, akkor láthatunk alacsony példánygyakoriságú és magas elemszámú csoportokat (*cukor*), illetve kisebb elemszámú nagy példánygyakoriságúakat is (*kapocs, lélek*). A legerősebb mintát képező *-alom* végűeknél elvárásaimmal összhangban az összetett szavak hasonlóan viselkednek alapszavaikhoz.

5.3.6. A hasonlósági csoportok vizsgálata alapján tett megfigyelések összegzése

A hasonlósági csoportok vizsgálata során több olyan megfigyelést is tettem, amelyeket érdemes itt röviden összefoglalni:

- 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.
- Egy hasonlósági csoporton belül tendencia jellegűen a leggyakoribb szavak viselkedését követik a ritkább szavak.
- Egy hasonlósági csoporton belül a kevésbé szorosan kapcsolódó elemek hajlamosabbak az egyedibb viselkedésre.
- A zárt mintát alkotó *-alom, -elem* végűek következetesen hangkivetők, ami kapcsolatba hozható erős hasonlósági viszonyaikkal és magas gyakoriságukkal is.
- A komplex jegymérték és a komplex tengelymérték által meghatározott legszorosabb kapcsolatok száma szignifikánsan összefügg a hangkivetés mértékével.
- A komplex jegymérték inkább a végek hasonlóságát, a komplex tengelymérték inkább a szerkezeti rokonságot emeli ki.
- Az alapszótól eltérő viselkedésű összetett szavak az átlagosnál kisebb típus- és példánygyakoriságú összetett szóbokrokban találhatóak.

5.4. Két nyelvállapot összehasonlítása

5.4.1. Az összehasonlítás célja és háttere

Ebben az alfejezetben két időben közeli állapotot hasonlítok össze abból a célból, hogy alaposabban **megismerhessük az analógián alapuló változás folyamatát**. Habár a nyelvi változás tendenciáit jelenleg lehetetlen megjósolni, a változás mértékéről vagy ütemérő pedig végképp csak találgathatunk, a nyelvi leírásnak fontos feladata, hogy feltárja, mi az, ami ezeket a folyamatokat mozgatja. Vizsgálatomban a korábbi állapothoz a ***Szószablya Gyakorisági Szótár*** **adja az adatokat**, míg az újabb állapothoz saját gyűjtésem, amelyet a ***Google*** **kereső** segítségével készítettem.

A ***Szószablya*** *Gyakorisági Szótár* alapján eddig egy szinkrónnak tekinthető állapotot jellemeztem, hogy feltárjam a hangkivető főnevek hasonlóságon alapuló viszonyait, és elemezzem viselkedésüket. Az elemzés közben nem tévesztettem szem elöl azt a tényt, hogy minden nagyobb gyűjtemény, különösen a webes gyűjtésen alapuló korpuszok, különböző korok szövegeit tartalmazhatják pontosan nehezen meghatározható arányokban[1]. Ezt és az 5.1. alfejezetben már bemutatott aránytalanságait figyelembe véve mégis viszonylagos nyugalommal mondhatjuk, hogy a *Szószablya Korpusz* az eddigi korpuszoknál, gyűjtéseknél jobban **jellemzi a 20. század végi, 21. század legeleji magyar nyelv[2] általános állapotát**. A *Szószablya Korpusz* így a jelenlegi nyelvészeti leírás bevett gyakorlatával és elvárásaival összhangban szinkrónnak tekinthető, mivel más szinkrón leírások esetében is, ha egyáltalán használnak külső nyelvi adatokat, akkor azok legalább annyira szóródnak időben és aránytalanok jellegükben, mint a *Szószablya Korpusz* alapját képező szövegek.

Az **újabb állapotot reprezentáló *Google Gyakorisági Gyűjtés*t** elsősorban azért hoztam létre, mert a tervekkel ellentétben a *Szószablya Korpusz* építése és frissítése elmaradt. *Google Gyakorisági Gyűjtés*emet 2010 tavaszán készítettem a .hu domain[3] alatti oldalakról, amelyek a *Google* adatai alapján 95%-ban magyar szövegeket tartalmaznak[4]. Érdeklődésem középpontjában csak bizonyos alakok álltak, így nem volt szükséges saját korpuszt sem építenem (Kornai és Halácsy 2008), hanem csak a kérdéses szóalakok .hu domain alatti találati számait, mint gyakorisági számokat rendeltem az egyes alakokhoz.

Google **adatok** felhasználásánál kételyeink lehetnek azzal kapcsolatban, hogy ezek **mennyire relevánsak és megbízhatóak**. Gyakran hallható az az

---

[1] Az arányok hozzávetőleges meghatározása nem lehetetlen vállalkozás, de ez túlmutatna jelenlegi vizsgálatom keretein.

[2] Korlátainkból adódóan ez a megállapítás elsősorban az írott nyelvre igaz.

[3] A .hu domain mellett azért döntöttem, mivel a *Szószablya Korpusz* kiindulási alapját is ez képezte. A *Google* lehetőséget ad kizárólagosan magyar szövegekben való keresésre is, de kutatásomban a nyelvi tudást aktívan használó nyelvtechnológiai eszközök használatát kerülni akartam, hogy ne szűrjék meg nem kívánatos módon adataimat.

[4] Ezt a .hu domain alatti magyar oldalak és a .hu alatti összes oldal aránya alapján kaptam meg.

impresszionisztikus tapasztalatokon alapuló megjegyzés nyelvészek és nyelvtechnológusok körében, hogy a *Google* a nyelvi viselkedés tanulmányozására nem megfelelő eszköz, mivel néhány keresésük nem az intuitív vagy az elvárt találati arányokat hozta. Habár benyomásaink sokszor félrevezetnek minket, az észrevételeknek van igazságalapja is. A *Google* adatgyűjtési és tárolási módszereiről keveset tudunk, és ismételt lekérdezések esetén az adatok mutathatnak minimális, de észrevehető különbséget, annak függvényében, hogy lekérdezésünkkel melyik szervert értük el. Kilgarriff (2007), Baroni és Ueyama (2006), valamint Nakov és Hearst (2005) egyaránt aggályaikat fejezik ki a *Google* találati számok nyelvészeti kutatásban való felhasználásával kapcsolatban. Azonban kifogásaik az enyémtől eltérő jellegű vizsgálatokra irányulnak, amelyek szónál nagyobb egységeket érintenek, vagy a *Google* által ajánlott speciális lekérdezési módokat használják. A *Google* egyes alakokra adott találatai és találati számai az idézőjelbe tett keresések esetén csak a nagy betűt nem veszik figyelembe, a ASCII kódtáblától eltérő karakterek már szerepet kapnak[1] az egyes lekérdezések differenciálásában.

Habár a *Google* sem tévedhetetlen, és találati adatai sem nyelvészeti kutatásra készültek, mégis mondhatjuk, hogy sokkal megbízhatóbb adatokat szolgáltat egy átfogó vizsgálathoz, mint a nyelvész nyelvi intuíciója vagy a nyelvészetben gyakran alkalmazott kisméretű, esetleges adatgyűjtések. Adataim megbízhatósága mellett tanúskodik az is, hogy a **két gyakorisági szótár gyakorisági mutatói** ($r(1067) = 0{,}854$, $t = 53{,}61$, $p < 0{,}001$) és **hangkivetési mértékei (**$r(1067) = 0{,}79$, $t = 42{,}43$, $p < 0{,}001$) közt igen **magas a pozitív együttjárás**, ami azt bizonyítja, hogy az adatok szerkezete és jellege hasonló. Ez a *Google* megbízhatatlanságának mértékét vagy a nem megfelelő adatok mennyiségét minimálisnak jelzi.

A *Google* esetében az összes alakváltozat lekérdezését lehetővé tevő reguláris kifejezések alkalmazására nincs mód, így a *Google* keresőnek csak karakterláncokat lehet a keresésekben megadni, ezért kizárólag a **leggyakoribb toldalékok lekérdezésére**

[1] Ennek megfelelően az *„árokban"* lekérdezés kevesebb találatot hoz, mint a *árokban* vagy a *arokban*, mert nem hozza találatnak a csak az *arokban*-t, *árok*-ot, *arok*-ot tartalmazó oldalakat.

szorítkoztam[1], amelyek a *Szószablya Gyakorisági Szótár*ban legalább 100 ezer hangkivetéses alakot számoltak[2]. Ezek a következők voltak:

(4) tárgyeset, többes szám, szuperesszívusz, E.1 birtokos, E.3 birtokos, T.3 birtokos, E.3 birtokos több birtokkal (*-(j)Ai*)

A *Szószablya Korpusz*ban az ezeket a toldalékokat tartalmazó **4,01 millió alak**[3] a hangkivető főnevek olyan alakjainak a 94,3%-áért[4] felelős, amelyek esetében a hangkivető viselkedés lenne az elvárt. A vizsgálatba a leggyakoribb toldalékokat tartalmazó alakokból csak azokat vontam be, amelyek esetében a toldalék a szóalak jobb szélén található, hogy limitáljam a lekérdezések számát. Így a gyűjtés részét képezi a *sátorai* alak, de a *sátoraival* nem. A *Google Gyakorisági Gyűjtés* elkészítése során minden hangkivető főnév lehetséges „hangkivető"/„nem hangkivető", „nyitó"/„nem nyitó", „kötőhangzós"/„nem kötőhangzós" változatát lekérdeztem. Így az utolsó szótagjukban hátul képzett magánhangzót vagy *ö*-t tartalmazó szavaknak[5] 5, az utolsó magánhangzójukként *e*-t tartalmazó szavaknak 3 változatára[6] gyűjtöttem gyakorisági számokat (5.14. táblázat).

---

[1] A *Google*-nek egy nap alatt gépileg beadható keresések száma még a *Google Scholar* (*Google Tudós*) program keretében is meglehetősen limitált.

[2] Minden toldalék esetében, amely együttjár a hangkivetéssel lényegesen több hangkivetéses alak van, mint nem hangkivetéses.

[3] Ez azt jelenti, hogy az eddig vizsgált adatokban benne voltak az olyan alakok is, amelyeknek a hangkivetéssel együttjáró toldalékok nem a legszélén voltak: pl. *sátorai<u>v</u>al*.

[4] Az összehasonlításból kihagyott 5,7% az összes többi birtokos alakot fedi le.

[5] A 8 kombinációból 3 nem lehetséges, mert nincs nyitás és/vagy hangkivetés kötőhang nélkül.

[6] További két lehetőség esik ki azzal, hogy az /e/ a köznyelvi változatban nyílt, így a nyitás az esetében nem értelmezhető, hisz a nyitó alakok semmiben sem különböznének a nem nyitó alakoktól.

Lekérdezett szó utolsó mgh-ja	kötőhang nélkül	kötőhanggal	hangkivetés	hangkivetés és nyitás	nyitás
a, o, u utolsó mgh	sátort	sátorot	sátrot	sátrat	sátorat
ö utolsó mgh	bögölyt	bögölyöt	böglyöt	böglyet	bögölyet
e utolsó mgh	kebelt	x	x	keblet	kebelet

5.14. táblázat: A *sátor, bögöly, kebel Google*-től lekérdezett alakjai

A ***Google Gyakorisági Gyűjtés*** **120-130 millió weboldalon** alapult[1], míg a *Szószablya Gyakorisági Szótár* 3,5 millió weboldalon, azaz a *Google Gyakorisági Gyűjtés* hozzávetőlegesen 35-ször nagyobb minta alapján készült. A *Google Gyakorisági Gyűjtés*ben (a *Szószablya Gyakorisági Szótár* módszerével ellentétben) ha egy oldalon egy szóalak többször is előfordult, akkor az csak eggyel növelte az adott alak gyakorisági számát, hisz a *Google* kereső találati számai ezen alapszanak. Ennek ellenére a *Google Gyakorisági Gyűjtés* (70 millió alak) és a *Szószablya Gyakorisági Szótár* (2,1 millió) figyelembe vett alakjainak 33:1-es aránya jól tükrözi a dokumentumszámok közti arányokat. A két gyakorisági szótár összehasonlításának lehetőségét Klein és Nelson (2009) eredményei is megerősítik, miszerint szignifikáns erős pozitív korreláció van a *Google Gyakorisági Gyűjtés*ben is használt dokumentumgyakoriság (*document frequency*), azaz a dokumentumok száma, ahol a keresett szóalak előfordult, és a *Szószablya Gyakorisági Szótár* által használt terminusgyakoriság (*term freqeuncy*) közt, amelynek számítása során az adott szóalak összes dokumentumon belüli előfordulását figyelembe vesszük.

A ***Google Gyakorisági Gyűjtés*** esetében is figyelembe kell vennünk, hogy adataink mennyire tükrözik a **jelenlegi nyelvállapotot**, illetve, hogy a *Szószablya Gyakorisági Szótár*hoz viszonyítva adatai újabbak-e. Elképzelhető ugyanis, hogy a *Google Gyakorisági Gyűjtés*ben archaikus vagy legalábbis 20. századi alakok nagy arányban fordulnak elő, és ezzel torzítják adatainkat. Ez vizsgálatomban zavaró tényezőként úgy érvényesülhetne, hogy a már analógiás változás útján járó szavak szorosabban

[1] A pontos számot nem lehet megállapítani, mert a *Google* a site:hu lekérdezésre mindig egy kicsit más, bár nagyságrendileg azonos számot adott vissza. Ennél pontosabb módszer azonban nincs méretének meghatározására, a *Google* pedig erről nem is közöl adatokat.

tartoznának a hangkivető paradigmába, mint a *Szószablya Gyakoriségi Szótár*ban, és így a *Google Gyakoriségi Gyűjtés* ebben az esetben a változás egy korábbi stádiumát rögzítené.

Ha megvizsgálunk olyan **szavakat**, amelyek **biblikus környezetben** is gyakran előfordulnak, mivel elsősorban a régi Biblia-fordítások, egyházi és az irodalmi szövegek felelősek a régies alakokért, akkor azt tapasztaljuk a *Google* esetében, hogy ezen alakok előfordulásainak száma elmarad a megfelelő szinkrón alakok gyakoriságától: *látá* 34 ezer : *látta* 983 ezer; *mondá* 30 ezer : *mondta* 3,91 millió. Ha még régies kontextusba, de nem biblikus környezetbe tartozó szóalakot veszünk, a megfelelő szinkrón alak gyakorisága akkor is jelentősen meghaladja az archaikus formáét: pl. *löveté* 70 : *lövette* 12000. Az arányok a *Szószablya Gyakoriségi Szótár*ban is hasonlóak, így mind a két korpusz esetében a régies szövegek csak egy töredékét tehetik ki az általuk lefedett weblapoknak[1].

A *Google* által elérhető weblapok ugyanakkor nagyobb arányban tartalmaznak újabb szövegeket. Erről tanúskodnak a **nyelv újabb állapotaira jellemző friss szlengszavak**, amelyek a *Google* által lefedett, mai magyar weben nagyobb arányban szerepelnek (pl. arányaiban 350-szer több *nyugger* 'nyugdíjas', vagy arányaiban 36-szor több *macsesz* 'macska' alak a *Google* alapján[2]). Ezt alátámasztandó 123 javarészt friss szlengszónak minősülő *-esz, -er* végű szó (Fükőh és Rung 2005) előfordulásainak gyakoriságát vizsgáltam meg a *Szószablya Gyakoriségi Szótár* és a *Google* keresésekre visszaadott találati számainak segítségével. E szavak szlengbe való tartozását támogatja az a tény is, hogy szignifikánsan ($t(99) = 4{,}16$, $p < 0{,}001$) nagyobb arányban (10:1) fordulnak elő rossz helyesírású szövegekben, mint a köznyelvi hangkivető szavak (2:1). 23 *-esz, -er* végű szó a *Szószablya Gyakoriségi Szótár*ban egyáltalán nem szerepelt, a

[1] A régi, régies szövegek arányának szisztematikus vizsgálatát a *Szószablya Korpusz* és a *Google* alapján nem végeztem el, de semmi okunk sincs azt feltételezni, hogy 2003 óta ilyen jellegű szövegek a korábbinál gyorsabb tempóban kerülnének fel az internetre. A néhány bemutatott példa csak ezt az elképzelést hivatott megerősíteni. Ezeket természetesen nem tekintem egy kiterjedt vizsgálattal egyenértékűnek, amelynek elvégzése azonban túlmutatna a disszertáció keretein.

[2] Összehasonlításom szempontjából nem fontos, hogy a *Google* dokumentumgyakorisággal számol, hisz ha e szavak esetében terminusgyakoriságot vennék a *Google* esetében is, akkor ezek az arányok még magasabbak lehetnének.

többiek pedig 746-szor[1] (szórás = 1790) magasabb találati számokat kaptak *Google* lekérdezésekre, mint a *Szószablya Gyakorisági Szótár*ban lévő gyakorisági számaik. Ez az arány jelentősen meghaladja a korpuszok méretei alapján elvárható 35-szörös különbséget. Ezekből az eredményekből arra következtethetünk, hogy a *Google* lefedi a *Szószablya Korpusz* anyagát, de újabb szövegeket is magában foglal. A *Google Gyakorisági Gyűjtés* frissebb volta mellett szól az is, hogy a .hu domain alatt a web 2.0-ás közösségi tartalmak (blogok, fórumok, twitter stb.) egyre gyorsuló tempóban jönnek létre, így ha ezekkel párhuzamosan régebbi szövegek kerülnek is fel az internetre, az új, informális tartalmak keletkezésének tempójával nem tudnak lépést tartani.

A vizsgálatba a ***Szószablya Gyakorisági Szótár*nál korábbi**, nagyméretű, informális szövegeket is tartalmazó és egyértelműen korábbra adatolható **korpuszokat, gyakorisági gyűjtéseket nem tudtam bevonni**[2], mivel formális és irodalmi jellegükből kifolyólag kevésbé alkalmasak vizsgálatomra. A gyakorisági adatokat tartalmazó középkori és barokk gyűjtések pedig méretükből fakadóan alkalmatlanok arra, hogy egy komolyabb összehasonlítás alapját képezzék, és mivel ezek is szerkesztett szövegek alapján készültek, az ingadozási jelenségek tanulmányozására sem megfelelőek, hiába tartalmaznak olykor olyan nem hangkivetéses alakokat is, amelyeknél hangkivetést várnánk el.

A viszonylag kevés, 7 éves eltérés a két vizsgálatra alkalmas gyakorisági gyűjtés között (*Szószablya* 2003, *Google* 2010) azonban már elegendő idő lehet ahhoz, hogy **megfigyelhessünk folyamatban lévő tendenciákat**, és különleges lehetőséget is teremt arra, hogy a többnyire évtizedekkel vagy évszázadokkal eltérő állapotokat összehasonlító elemzések helyett két nagyon közeli állapotot vethessünk össze. A hangkivető főnevek analógiás változása feltételezésem szerint már ilyen rövid idő alatt is tetten érhető, így a szűknek mondható időkülönbség ellenére is képesek leszünk

[1] Összesen 18 szó nem hozott a *Szószablya Gyakorisági Szótár*hoz viszonyítva 35-szörösnél magasabb találati arányt a *Google* esetében*: vozzer, labdesz, linker, dummer, bakker, Karesz, bazzer, cigesz, spangesz, röpper, Bükker, alter, tatyesz, Gabber, amfesz, cummer, kaller, kolesz.* Ezek javarészt a már viszonylagosan régebbi *-esz* és *-er* végű szavak közül kerültek ki.

[2] Mint az 5.1. alfejezetben láthattuk, erre a célra az *MNSz* is kevésbé alkalmas.

érdemleges megfigyeléseket tenni. Ez még akkor is lehetséges lenne, ha azt a ma már kevésbé tartható nézetet vallanánk, hogy a nyelvi változás kizárólagosan annak köszönhető, hogy a felnövekvő nemzedék az előző generáció beszédprodukciója alapján más rendszert épít fel, mint amilyen az a rendszer, ami e beszédprodukció mögött áll, mivel egy esetlegesen 30-40 év alatt teljesen végbemenő változásnak 7 év alatt már meg kell, hogy mutatkozzanak a jelei. Azonban tudjuk, hogy az egyén nyelve is folyamatosan változik (Sankoff és Blondeau 2007), így elméleti szinten a 30-40 évnél gyorsabb teljes változások is lehetségesek.

Arról viszonylagosan biztosak az ismereteink, hogy a **nyelvi változás sebessége eltérő lehet jelenségenként**, amire hatással lehet a szóban forgó jelenség természete és társadalmi kontextusa (Johnson 1976), mint például a beszélőközösségek mérete és annak hálózatos természete (Neetle 1999). Johnson (1976) nyomán azt is tudjuk, hogy ellentétben King (1969) feltételezéseivel a nyelvi változás nem egyenletesen terjed, hanem kezdetben lassú, majd a kezdeti átadóktól egyre gyorsabban terjed tovább (Barabási 2002)[1]. Habár elsősorban a szociolingvisztikán belül már voltak kezdeményezések arra, hogy a változás sebességének természetét jobban leírjuk, mindezidáig nincs átfogó ismeretünk arról, hogy bizonyos típusú változásoktól milyen tempót várhatunk el, hiába is ismerjük azok jellegét és szociális kontextusát.

Összehasonlításomat az összevetett adatok időbeli közelségén túl a korpuszok jelentős mérete is egyedivé teszi. Igaz, ebből kifolyólag a pontatlanságok, adathibák kézi javítása nehezebb, de a nagy adatmennyiség ezeket az egyenetlenségeket kisimítja, illetve az egyedi adathibák az elemzés során azonosíthatóak és ezáltal figyelmen kívül hagyhatóak lesznek. Ezzel összhangban az adatok jobb minőségének érdekében a *Google Gyakoriségi Gyűjtés*ből közel **tucatnyi olyan alakot már vizsgálatom elején kivettem**, amelyeket a *Szószablya Gyakoriségi Szótár*ral végzett vizsgálataim során is eltávolítottam (pl. *koromat*, *körömet*, *karomat*, *kötelemet*, *kéreget*, *méreget* stb.), mivel ezek

[1] Ha ugyanilyen tempóban (7 évente 30%-kal több nem hangkivetéses alak) folytatódna a kiegyenlítődés, akkor 80-90 év múlva tűnne el a hangkivetők paradigmája, ha a mértéket vennénk állandónak (7 évente 1,3%-kal több nem hangkivetéses alak), akkor 300-400 év múlva következhetne be a hangkivető paradigma megszűnése.

nem hangkivető főnevek hangkivetés nélküli alakjai olyan toldalékok előtt, amelyek a hangkivetéssel együttjárnak, hanem más szavakhoz tartoznak. A *-terem* végű szavak estében is elképzelhető, hogy egyes nem hangkivetéses alakjaik homonimak más szavak alakjaival: pl. *előteremet* (*előtér*+POSS.E.1+ACC vagy *előterem*+ACC). Bizonyos esetekben még a mondatkontextusból sem derül ki egyértelműen, hogy az alak melyik értelemben fordul elő. A *Szószablya Gyakorisági Szótár*ral kapcsolatban alkalmazott korábbi gyakorlatomat követve ezeket a *Google Gyakorisági Gyűjtés*ben is bennhagytam, mivel nincs okunk azt gondolni, hogy a *-tér*+POSS.E.1 végű alakok gyakoribbak lennének, mint a *-terem* végűek. Szükséges volt még a köznévi hangkivető főnevekkel homonim, de azoktól eltérő nem hangkivető módon toldalékolt **tulajdonnévi alakok kivétele** is, mint *Sólyomot*, *Fodort*, *Bodort*, *(Arany) Ászokat* stb., mivel a *Google* a kisbetűs és nagybetűs írásmód közt nem tesz különbséget. Egyes esetekben még szlovák találatokat is ki kellett emelni, amelyek aktuálpolitikai jellegük miatt bukkantak fel Sólyom László köztársasági elnökre utalva: *Sólyomom* 'Sólyom+INSTR', *Sólyoma* 'Sólyom+ACC/GEN'[1].

Egyes **ritkább szavak esetében a *Szószablya Korpusz* kisebb mérete miatt az összehasonlítás nem volt lehetséges**. Csak azt az 1069 alakot vettem be a vizsgálatba (a szavak hangkivetési adatai az A Függelékben láthatók), amelyeknek mind a két gyűjtésben van legalább egy előfordulásuk. 141 szóra a *Szószablya Korpusz*ban nem volt adat, míg a *Google* csak 19 szóra nem adott találatot: *ballonselyem, bronzfejedelem, érctulok, gavallérsarok, gúlasátor, huszárkapocs, katapultterem, kielégülésnyugalom, közlator, laboratóriumterem, nyúlkapor, oltógödör, puplinselyem, szómalom, tejgyomor, tisztességszobor, túrómalom, üszögféreg, zöldselyem, zsorzsettfátyol.* Ezekből egyedül a *huszárkapocs*-ra volt adat a *Szószablya Korpusz*ban.

[1] Az ilyen esetek sajnálatos következményei annak, hogy nem alkalmaztam a *Google* nyelvi szűrését, amellyel azonban értékes adatokat is veszthettem volna. Mind a két eljárásnak vannak hátrányai, én inkább az informatívabb, de hibákat rejtő megoldást választottam.

5.4.2. A változás jellemzői

A állapotok összehasonlítása előtt röviden bemutatom Bárczi és mtsai (1967) alapján a **főnévi hangkivetés kialakulását a magyar nyelvben**, hogy elemzésemet a történelmi kontextusban jobban el lehessen helyezni. Bárczi és mtsai (1967) elemzésükben többször hivatkoznak analógiás változásokra, azonban ezt sosem abban a viszonylag jól formalizált változatban teszik, amely a legújabb analógiás megközelítésekre jellemző. Analógiafelfogásuk leginkább a 19. századi újgrammatikusokéhoz közeli. Levezetéseik akár az esetek többségében helyesek is lehetnek, de következtetési módszereik nincsenek eléggé alaposan formalizálva, bemutatva. Szövegük alapján nem tudhatjuk pontosan, hogy az analógiás források kiválasztása az egyes esetekben milyen feltételek mentén történt meg.

A magyar főnévi hangkivetés már az **ősmagyar korban** kialakulóban lehetett a **kétnyíltszótagos tendencia** és a **magánhangzó-betoldás hatására**, hisz 950 körül az *Álmos*[1] névben is hangkivetéses alakkal találkozunk (Jakubovich és Pais 1929: 7). Ma már csak mérsékelten keletkeznek új hangkivető főnevek a csoport gyenge analógiás erejéből kifolyólag[2]. A hangkivető főnevek eredete nem befolyásolta későbbi viselkedésüket. Bárczi és mtsai (1967) szerint olyan esetekben jöttek létre hangkivető szavak, amikor a szó utolsó magánhangzójának kiesését követően jól artikulálható mássalhangzó-kapcsolatok jöttek létre. Azonban a jó artikulálhatóságot, legalábbis a hangkivetők kontextusában, nem határozzák meg egyértelműen, így a kijelentés pontos jelentése némileg homályos marad[3]. Mint az 5.2.1. alfejezetben láthattuk, a hangkivető főnevek utolsó két mássalhangzójára jellemző, hogy preferált az utolsó két mássalhangzó megfelelő mértékű különbsége és az egymás mellé kerülő

[1] Az *álom* finnugor eredetű szó. A többi finnugor nyelvben is *VCVC* szerkezetűek a neki megfeleltethető szavak.

[2] Újabb keletkezésűek a diáknyelvi 'irodalom' jelentésű *rohadalom, sirodalom, unodalom, rogyadalom* stb. szavak.

[3] A homályosságon túl a kifejezés kétértelmű is, hisz vehetjük úgy, hogy a „jól artikulálható" azt jelenti, hogy könnyen kiejthető, de úgy is, hogy alkalmas a torzulásmentes kiejtésre. Az utóbbi értelmezés esetén elképzelésük hasonlónak is vehető az enyémhez.

mássalhangzók közti esetleges hasonulások, összeolvadások kerülése. Ez azonban a könnyű kiejthetőség helyett inkább egy olyan törekvésnek feleltethető meg, miszerint a szóalak a magánhangzó kiesése után is minél jobban hasonlítson a nem hangkivetéses alakokra, így már ismert elemei (fonémái) is minél jobban felismerhetőek legyenek a kommunikációban.

Az **ómagyar korból** több szó esetében tudomásunk van **évszázadokon át egymás mellett élő hangkivetéses és nem hangkivetéses alakvariációkról**: pl. *Solumus* (Váradi Regestrum 1903 [1550]: 255), *Solmus* (Csánki 1890: 1: 525). A tőbelseji magánhangzót tartalmazó alakok sokáig való fennmaradását a ma már hangkivetést elváró esetekben Bárczi és mtsai (1967) az alanyesetű alakok analógiás hatásának tulajdonítják. Elképzelhető, hogy a jelenbeli változások mögött ezek az analógiás hatások újra szerepet játszanak, különösen az olyan szavaknál, amelyeknél a hangkivetéses alakok példánygyakoriságának aránya alacsony összes alakjaikban, de a változásról való hiányos ismereteink miatt ezzel kapcsolatban legfeljebb csak találgathatnánk.

Átmenetileg **némelyik szónak** alanyesetben is élhettek **mássalhangzó-kapcsolatra végződő változatai**[1], mint pl. *arc* 'árok' (Jakubovich és Pais 1929: 59). Legjellemzőbb volt ez az *-alom/-elem* végű szavakra (pl. *hotolm, Ómagyar Mária Siralom*), amelyek ma is a szóanyagnak a hangkivetés szempontjából legstabilabb részét adják. Bárczi és mtsai (1967) elképzelhetőnek tartják azonban, hogy ezek a két mássalhangzóra végződő alakok csak a választékos írásbeliség sajátjai voltak, mint ahogy sejthetően ma is sokkal nagyobb arányban szerepelnek hangkivető szavak nem hangkivetéses alakjai a hangkivetést elváró toldalékokkal a beszélt köznyelvben. Ezek a két mássalhangzóra végződő alakok időnként mássalhangzóval kezdődő toldalékokkal is előfordultak (pl. *birodalmba, Bécsi Kódex* 1916: 234), ami annak lehet bizonyítéka, hogy a szónak élt egy *-CC* végű változata is, amelyben egyáltalán nem volt hangkivetés.

Már az *Érdy Kódexben* (1527) található a kiegyenlítődés útján lévő alak, a *Sathort*, amelyet Bárczi és mtsai (1967) a jelenbeli folyamatok előhírnökének tekintenek, azonban azt nem bizonyítják, hogy korábban a *sátrat* alak volt az uralkodó változata a

[1] Az ilyen szavak közül máig megmaradt jelentéshasadás következtében a *sark* szó.

sátor tárgyesetének. Vélhetőleg azonban a hangkivetők rendszerére mindig is jellemző volt az ingadozás, a teljesen stabil állapot sosem létezett. A magyar hangkivető főnevek változása és viselkedése nem egyedi jelenség, hisz meglehetősen hasonlít a lengyel e~∅ váltakozás történetéhez és jelenbeli kiegyenlítődéséhez[1], amelyet Kraska-Szlenk (2007) szintén modern analógiás keretben elemez.

A gyakorisági adatok összehasonlítása révén láthatjuk, hogy a **hangkivetés visszaszorulása folytatódik**, ami összhangban van a 4.1. és 4.2. alfejezetekben kifejtett elképzeléseimmel, miszerint a nem hangkivető főnevek analógiás ereje számosságukból kifolyólag nagyobb, így a hangkivető szavakat saját viselkedésük irányába húzzák. Ezzel a hatással szemben a hasonlóbb gyakori hangkivető szavak megtartó ereje egyensúlyoz. A vizsgált toldalékos alakokban az összes alak 99,42%-a[2] volt hangkivető a *Szószablya Gyakorisági Szótár*ban, míg 98,12% a *Google Gyakorisági Gyűjtés*ben. Ezzel összhangban a *Szószablya Gyakorisági Szótár*ban 124[3] szó mutat legfeljebb 99%-ban hangkivető viselkedést, míg a *Google Gyakorisági Gyűjtés* esetében ezek száma már 184. A hangkivető szavak viselkedése nem változott lényegesen. Elvárásaimmal összhangban nagyon gyakori, a hangkivető mintát kevésbé követő szavak továbbra sincsenek, és a ritka, stabilan hangkivető szavak maradtak a legjellemzőbbek[4].

---

[1] A folyamat érdekessége, hogy egy szót mind a két nyelvben érint a nagyon hasonló analógiás változás: *cseber* és a lengyel *ceber* 'favödör'.

[2] Az 5.1. alfejezetben a *Szószablya Gyakorisági Szótár* hangkivető főneveinek ingadozását 97,57%-ban adtam meg. Az eltérés oka, hogy a 97,57%-os érték, amelyet összességében mérvadóbbnak tartok, az egyes szavak toldalékonként számolt hangkivetési mértékének átlagát mutatja, míg az itt szerepeltetett érték a hangkivetéses és nem hangkivetéses alakok arányát szavaktól függetlenül határozza meg a hangkivetéssel együttjáró toldalékok vonatkozásában, amely értéket a nagyon gyakori, de stabilan hangkivető *-alom/-elem* végűek a 100% irányába mozdítanak el. Ezt a mutatót azért alkalmaztam itt egyedüli alkalommal, mivel a szavak hangkivetésének jellegét a toldalékonként számolt érték írja le jobban, míg ez a mutató a két állapotban a hangkivetés „általános" helyzetét jellemzi megfelelőbben.

[3] Az 5. fejezetben összesen 118 ilyen tőről számolok be (116 ingadozó és 2 kiegyenlített), de vegyük észre, hogy a számítás ott némileg több toldalék alapján készült.

[4] A ritka szavak esetében sokszor a stabilnak tűnő hangkivető viselkedés mögött adathiány állhat, azaz csak azért tűnnek következetesen hangkivetőnek, mert nincs birtokunkban elég adat, ami alapján viselkedésükről árnyaltabb képet rajzolhatnánk.

	Szószablya 2003	Google 2010	változás mértéke	változás dinamikája	Szószablya összevetéshez	Google összevetéshez	páros t-próba
tárgyeset	96,82 % (967)	96,36% (1058)	0,46%	1,14	96,82 % (967)	96,28%	t (966) = -1,93, p = 0,054
szuper-esszívusz	95,31% (520)	93,11% (890)	2,2%	1,47	95,29% (517)	94,59%	n.sz.
többes szám	98,62% (789)	98,54% (991)	0,08%	1,06	98,62% (788)	98,91%	n.sz.
E.1 birtokos	97,83% (292)	97,68% (683)	0,15%	1,07	97,81% (289)	98,33%	n.sz.
E.3 birtokos	97,56% (734)	96,66% (989)	0,9%	1,37	97,56% (734)	96,83%	t(733) = -2,03, p < 0,05
T.3 birtokos	98,71% (333)	98,55% (630)	0,16%	1,12	98,69% (328)	98,85%	n.sz.
E.3 birtokos több birtokkal	98,99% (429)	98,71% (770)	0,28%	1,28	98,99% (429)	98,98%	n.sz.
összes toldalék	97,57% (4064)	96,97% (6010)	0,6%	1,25	97,56% (4051)	97,31%	t (4050) = -1,83, p = 0,0674

5.15. táblázat: A hangkivető főnevek hangkivetésének mértéke az egyes toldalékoknál a *Szószablya korpusz*ban és a *Google Gyakorisági Gyűjtés*ben

Az 5.15. táblázat alapján láthatjuk, hogy az **egyes toldalékokkal együttjáró hangkivetés mértéke nagy változatosságot mutat**. A „Szószablya 2003" és a „Google 2010" oszlopokban lévő adatok az egyes toldalékoknál a szavak hangkivetési mértékeinek átlagait mutatják, ha a két gyűjtés adatait függetlenül veszem figyelembe. A zárójelben lévő szám megadja, hogy ez a számítás hány szó alapján készült. Minél ritkább egy toldalék, annál kisebb ez a szám, hisz a ritkább szavak esetében kevésbé van módunk olyan alakjaik viselkedésének a megfigyelésére, amelyekhez önmagukban is ritkább toldalékok kapcsolódnak. A „változás mértéke" azt adja meg, hogy a *Google Gyakorisági Gyűjtés*ben hány százalékkal több nem hangkivető alak van az adott toldaléknál („*Google* hangkivetési mérték" – „*Szószablya* hangkivetési mérték"), a „változás dinamikája" pedig a *Google Gyakorisági Gyűjtés* és a *Szószablya Gyakorisági*

Szótár nem hangkivető alakjainak arányát adja meg. A „Szószablya összevetéshez" és a „Google összevetéshez" oszlopok azon szavak toldalékonként számolt hangkivetési mértékének átlagait adják meg, amelyek mind a két gyűjtésben szerepeltek az adott toldalékkal. A „Szószablya összevetéshez" oszlopban található zárójeles érték ezek számát adja meg. Ez a szám minden toldalék esetében kisebb[1], mint a *Google Gyakorisági Gyűjtés* ide vonatkozó száma, hisz a *Google Gyakorisági Gyűjtés* 35-ször annyi adat alapján készült el, mint a *Szószablya Gyakorisági Szótár,* így az egyes toldalékoknál is több adatot nyújt a szavak viselkedésével kapcsolatban. Az összehasonlítható szavak hangkivetési mértékeit az egyes toldalékoknál a páros t-próbával vetettem össze, aminek eredményeit is az 5.15. táblázat mutatja meg. Az egyetlen szignifikáns eredmény mellett a két közel szignifikáns próba számait is feltüntettem.

A táblázat adataiból jól látszik, hogy a változásban **az egyes toldalékok meglehetősen eltérő dinamikával vesznek részt**, látszólag az egyes toldalékos alakok viszonylagos függetlenséggel viselkednek (Ackerman és mtsai 2009). Azok a toldalékok, amelyek már korábban is kevésbé jártak együtt a hangkivetéssel, az új adatok fényében még kevésbé stabil hangkivetők. Ez alól némileg kivétel a tárgyeset, mert változásának lanyhult a sebessége, az E.3 birtokos több birtokkal pedig elindult a felzárkózásban. A páros t-próba alapján egyedül az E.3 birtokos esetében szignifikáns az eltérés a *Szószablya Gyakorisági Szótár* és a *Google Gyakorisági Gyűjtés* között, míg a tárgyesetnél és az összes toldaléknál csak egy nem szignifikáns tendencia mutatható ki. A legdinamikusabban változó, legkevésbé hangkivető szuperesszívusz esetében a t-próba nem volt szignifikáns, mivel tekintetében az alakok viselkedése nagyon heterogén.

Az egyes **toldalékokat** a hangkivetési mértékük szerint a következőképpen **rangsorolhatjuk** a *Szószablya Gyakorisági Szótár* (5) és a *Google Gyakorisági Gyűjtés* (6) adatai alapján. Újabb **páros t-próbák alapján** a szignifikánsnak bizonyuló eltéréseket megbízhatóságuk függvényében jelöltem:

---

[1] Többször azonosak ezek a számok a „Szószablya 2003" alatt feltüntetett értékekkel. Esetenként némileg kisebbek, mert szórványosan lehet találkozni az egyes szavaknak is olyan vizsgált paradigmatikus celláival, amelyekre a *Google Gyakorisági Gyűjtés*ben nincs adat, de a *Szószablya Gyakorisági Szótár*ban igen.

(5) *Szószablya Gyakorisági Szótár*

E.3 több birtokkal > T.3 birtokos > többes szám > E.1 birtokos > E.3 birtokos > tárgyeset >** szuperesszívusz

E.3 birtokos és szuperesszívusz között: ***

többes szám és tárgyeset között: ***

(6) *Google Gyakorisági Gyűjtés*

E.3 több birtokkal > T.3 birtokos > többes szám > E.1 birtokos >*** E.3 birtokos > tárgyeset >*** szuperesszívusz

** = $p < 0,01$

*** = $p < 0,001$

(5) alapján láthatjuk, hogy csak a szuperesszívusz különül el a többi toldaléktól jelentősebben, amely különbség az E.3 birtokostól számítva a hangkivetéssel legkövetkezetesebben együttjáró toldalékok felé még jelentősebb. Egy másik törésvonal a többes szám és a tárgyeset között található, azaz a tárgyeset hangkivetésének a mértéke csak a hangkivetéssel majdnem tökéletesen együttjáró toldalékos alakoktól tér el. A birtokos ragok és a többes szám hangkivetési mértékei nem különböznek szignifikánsan egymástól. (6) esetében a különbségek a több adatnak köszönhetően egyértelműbben látszanak. A legkevésbé hangkivető szuperesszívusz megelőzi az összes toldalékot, amelyek közül azonban a tárgyeset és az E.3 birtokos hangkivetési mértékében közelebb áll hozzá. Ezek alapján úgy tűnik, hogy a szuperessszívuszos alakok követik a legkevésbé a hangkivető mintát, amiben a tárgy és némileg leszakadva az E.3 birtokos jön utánuk. A többi toldalék, amelyek javarészt ritkábbak, bár köztük van a nagyon gyakori többes szám is, következetesen hangkivetéses alakokban jelennek meg.

A **szuperesszívusz sajátos viselkedésére** már Kálmán és mtsai (2005) is felhívták a figyelmet, ugyanis ez az egyetlen kötőhangzós (szintetikus) rag, amely előtt a nyitótövek után is középső nyelvállású magánhangzó szerepel (pl. *házon*, de *házat, házak* stb.), illetve a tőbelseji magánhangzó-rövidülés is hiányzik az esetében (*pl. nyáron,* de *nyarat, nyarak* stb.). A szuperesszívusszal együttjáró hangkivetés mértékének gyorsuló csökkenése ezeknek a regularizálódás irányába mutató alakoknak a növekvő nyomásával hozható kapcsolatba. A szuperesszívuszos alakok heterogén viselkedése pedig azzal magyarázható, hogy kapcsolódásában nincsenek olyan fonotaktikai megkötések – mint amilyeneket a tárgyraggal kapcsolatban az 5.2. alfejezetben már megfigyelhettünk –, amelyek a nem hangkivetéses szuperesszívuszos alakok elfogadhatóságát befolyásolnák. A **tárgyas alakokban** a hangkivetés mértékének a csökkenése mögött az áll, hogy ezeknek lehet nem kötőhangzós alakja is bizonyos esetekben (bővebben 5.2.1. alfejezet). A tárgyrag kötőhangzóval azonban akár bármilyen hangkivető főnévhez kapcsolódhat (pl. *fejedelemet, fészeket)*. Az ilyen alakok szótagolási szempontból nem rosszabbak, mint a következetesen nem hangkivető, de nem is harmonizáló *-ig, -ért, -ék* toldalékokkal létrehozottak (*fe.je.de.le.me.ket : fe.je.de.le.me.kig, fe.je.de.le.me.kért, fe.je.de.le.mék*). A kötőhangzó nélküli tárgyas alak azonban, azért ideálisabb végpontja az analógiás kiegyenlítődésnek, mivel a kötőhangzó nélküli változatban **is ugyanannyi a szótagszám, mint az eredeti hangkivetéses alakban**. Így egy olyan alakot hozhatunk létre, amely jobban hasonlít az alanyesetű alakhoz, de a korábbi, hangkivetéses alaktól sem távolodik el annyira, hisz fonémaállományában azonos marad leszámítva az olyan eseteket, amikor az utolsó magánhangzó nem középső nyelvállású vagy a tő nyitó (*bajszot* : *bajuszt*, *sátrat* : *sátort*[1]).

Az **E.3 birtokos alakok** kevésbé hangkivető viselkedése és dinamikus változása mögött nem lehet egységes okot felfedezni. A stabilan hangkivető **T.3 birtokos** és a **többes számú** alakok egységes viselkedését az egymáshoz való fonológiai (*(O/U)k* felépítés) és szemantikai (többesség) hasonlóságuk is erősíti. Az **E.1 birtokos** jelentésében a sokkal gyakoribb E.3 birtokoshoz áll a legközelebb, így viselkedésében ehhez közelít. A szuperesszívusz még erőteljesebb formai hasonlósága sem volt

[1] A *sátor* esetében a nyitás is ingadozik, hisz a *sátort*-nál sincs nyitás.

elegendő ebben az esetben, hogy erőteljesebb ingadozásra serkentse. Az **E.3 birtokos több birtokkal** magas hangkivetési mértéke mögött az is állhat, hogy a *Google Gyakoriségi Gyűjtés*ben a 100 leggyakoribb *-ai/-ei* végűből (az *-ai/-ei* végű alakok 93,3%-áért felelősek) csak 3 volt, amely 90%-nál kevésbé követi a hangkivető mintát (*bajusz, berek, sátor*), míg a jelen összehasonlításomban szereplő szavak közt a 90%-nál kevésbé hangkivetők aránya közel kétszer ennyi (5,6%), azaz az *-ai/-ei* vég a következetesen hangkivető szavakkal fordul elő gyakrabban, ebből következik, hogy esetében is stabilabb a hangkivetés.

Az 5.16. táblázat alapján láthatjuk, hogy az **egyes toldalékok hangkivetési mértéke** mennyire jól **jellemzi a szavak viselkedését a hangkivetés szempontjából**. A *Szószablya* estében a toldalékok hangkivetési mértékei az „E.1 birtokos" alakot leszámítva megbízhatóan mutatják, hogy egy szó mennyire követi a hangkivető mintát. A tárgyeset, a többes szám és a „T.3 birtokos" hangkivetési mértéke jár leginkább együtt a szavak hangkivetési mértékével. A klasszikus hangkivető paradigmában egyedül az „E.3 birtokos"-nál van megoszlás (*pocokja*, de *torka*), mégis viszonylag jó előrejelzőnek mondható, mert a *pocok* típusú szavak rendkívül alacsony számban fordulnak elő (bővebben 5.3.2. alfejezet). A tárgyeset jó előrejelző képessége annak tudható be, hogy ha egy szónak már a tárgyesete nem következetesen hangkivető, akkor megtette az első lépést a kiegyenlítődés útján, míg a hangkivetéssel kevésbé együttjáró szuperesszívuszos alakok esetén a hangkivetés elmaradása sokkal esetlegesebb, nem lehetünk biztosak abban, hogy ez más alakok esetében is bekövetkezik (pl. van *Szentléleken*, de *Szentléleket* kevésbé, *Szentlélekt* pedig nincs). A *Google* esetében az alacsonyabb mértékű korrelációk betudhatók a nagyobb zajnak (tökéletlenebb előszűrés) és a hangkivető minta lazulásának, aminek következtében a szavak viselkedése heterogénebb lett.

	Szószablya	Google
tárgyeset	0,92 ***	0,85 ***
szuperesszívusz	0,6 ***	0,52 ***
többes szám	0,73 ***	0,63 ***
E.1 birtokos	0,4 ***	0,63 ***
E.3 birtokos	0,69 ***	0,62 ***
T.3 birtokos	0,88 ***	0,73 ***
E.3 birtokos több birtokkal	0,58 ***	0,53 ***

5.16. táblázat: Az egyes toldalékok hangkivetési mértékének együttjárása a szavak összes releváns toldalékos alakja alapján számított hangkivetési mértékkel a *Szószablya Gyakorisági Szótár* és a *Google Gyakorisági Gyűjtés* viszonylatában

A változás dinamikájának áttekintése után érdemes megnéznünk, hogy az **egyes végződések súlyaiban** milyen **változások** és eltolódások történtek. Az 5.17. táblázat a *Szószablya Gyakorisági Szótár*ban és a *Google Gyakorisági Gyűjtés*ben az 5.2. alfejezetben is alkalmazott módon azon szavak végeinek egymáshoz viszonyított arányait mutatja meg a legfontosabb toldalékoknál, amelyek az adott toldalékkal legfeljebb 90%-ban viselkednek hangkivető módon. A végződések arányait ezúttal is az egyes szám alanyesetű alakok gyakoriságával súlyoztam.

Az **arányok** a **tipikusabb hangkivető végződések irányába** tolódtak el, már a *-g* is megjelenik egyes toldalékoknál. A *-k* és az *-m* az összes releváns alak hangkivetési mértéke alapján a kevésbé hangkivető alakoknál a korábbi 13,39% helyett már 49,19%-ban képviselteti magát, elsősorban a *vacak, Szentlélek, külkereskedelem, sulyom* stb. hatására. E tipikus hangkivető tővégek nagyobb számban való megjelenésének köszönhetően a kevésbé tipikus hangkivető *-l, -r* tővégek arányai csökkentek a csoportban. Legszembetűnőbb a tipikusan hangkivető tővégek előretörése a hangkivetéssel legkevésbé együttjáró szuperesszívusz esetében, amelynél a kapcsolódásnak kevésbé vannak fonotaktikai korlátai, mint a *-t* tárgyragnál, aminek az esetében az *-l* és az *-r* tővégek még mindig magas arányban vannak jelen. Érdemes a

tárgyesetnél említést tenni arról, hogy a már korábban (5.2.1. alfejezet) az *-mt*-nél és az *nyt*-nél jobbnak jelzett *-nt* kapcsolat is nagyobb súllyal van jelen, és a *csiszolóvászont, zsákvászont* alakok mellett további *-n* végű szavak nem hangkivetéses alakjai jelentek meg tárgyesetben (*vászon, pamutvászon, gyöngyvászon, haszon*).

A kevésbé tipikus *-k, -m* végű szavak hangkivetési mértékének elsősorban a szuperesszívuszos alakokhoz köthető megugrásán túl bizonyos esetekben **egy-két részjelenséget** is megfigyelhetünk. Szembetűnő bizonyos *-j* végű szavaknál (*bagoly, bögöly*), hogy csak a többes szám, E.1 birtokos és a szuperesszívusz esetében kevésbé hangkivetők, aminek az oka talán e három toldalék részleges fonológiai hasonlóságában keresendő. Az *-m* tővég (*barom, majom, divatmajom, facimbalom, kerecsensólyom, kereskedelem, korom, külkereskedelem, nőuralom, rabszolgakereskedelem, üröm)* elsősorban az E.3 birtokosokkal szerepel magas arányban, ami azonban nem vezethető vissza egységes okra, valószínűleg szavanként más, egymástól független tényezők hatására alakult így (bővebben 5.4.4 alfejezet). A T.3 birtokos és az E.3 birtokos több birtokkal esetén már csak a legkevésbé hangkivető szavak (pl. *bajusz, sátor, fátyol* stb.) alakjai szerepelnek, így ezek tanulmányozása általánosabb összefüggések megállapítására nem alkalmas, ezért táblázatomban nem szerepeltetem az ezekre vonatkozó adatokat.

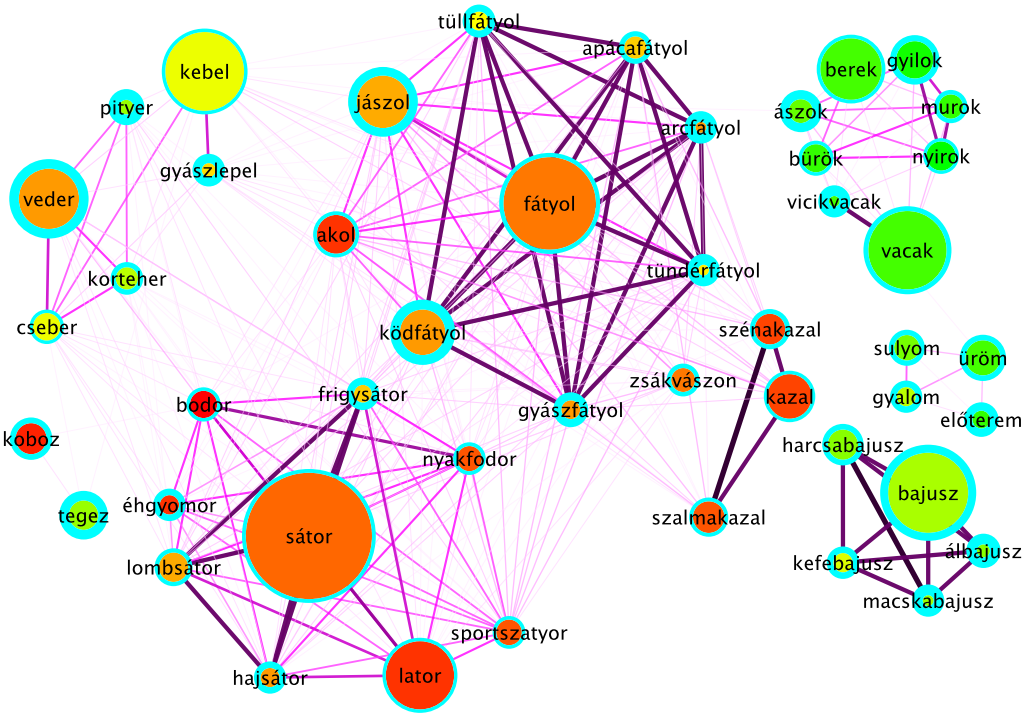
Szószablya összes		**Google összes**		**Szószablya tárgy**		**Google tárgy**		**Szószablya szuper-esszívusz**		**Google szuper-esszívusz**	
r	40,87%	k	37,12%	r	36,71%	k	34,33%	k	52,63%	k	54,93%
l	29,69%	l	20,64%	l	30,43%	l	20,19%	l	22,47%	m	16,57%
sz	13,95%	r	16,00%	k	16,83%	r	15,48%	r	17,91%	l	12,67%
k	12,41%	sz	14,11%	sz	7,67%	n	10%	m	3,24%	r	7,48%
z	1,26%	j	6,95%	j	4,65%	m	9,52%	z	2,43%	cs	3,19%
m	0,98%	z	2,70%	m	2,03%	sz	9,42%	sz	0,91%	j	1,96%
j	0,80%	m	2,07%	z	1,25%	z	0,73%	j	0,40%	z	1,83%
n	0,04%	n	0,41%	n	0,25%	j	0,34%			sz	1,35%
				ny	0,16%					g	0,01%
										n	0,01%
										ny	0,01%

Szószablya többes szám		Google többes szám		Szószablya E.1 birtokos		Google E.1 birtokos		Szószablya E.3 birtokos		Google E.3 birtokos	
k	57,00%	k	54,41%	sz	39,04%	sz	36,94%	r	37,64%	m	35,10%
l	17,42%	j	16,88%	l	31,34%	k	30,65%	sz	31,22%	r	27,22%
j	12,5%	z	10,62%	r	21,78%	l	26,13%	l	27,46%	l	17,09%
z	4,92%	m	10,33%	m	6,24%	j	4,46%	m	3,11%	sz	14,8%
r	3,58%	sz	5,50%	z	0,80%	z	0,79%	z	0,33%	z	2,37%
sz	3,42%	l	1,44%	k	0,80%	m	0,76%	k	0,23%	j	1,78%
m	0,67%	r	0,82%			r	0,26%	cs	0,02%	k	1,10%
ny	0,50%									g	0,53%
										n	0,01%

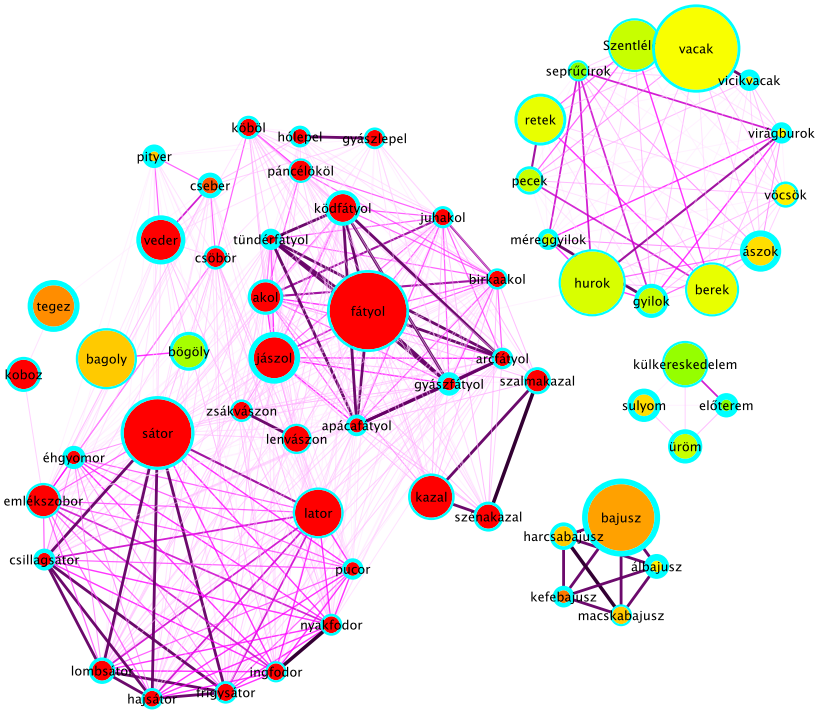
5.17. táblázat: Az egyes toldalékaiknál a hangkivető viselkedést legfeljebb 90%-ban követő szavak végei a *Szószablya Gyakorisági Szótár*ban és a *Google Gyakorisági Gyűjtés*ben

A változás általános áttekintését a *Szószablya Gyakorisági Szótár*ban és a *Google Gyakorisági Gyűjtés*ben **a legfeljebb 90%-ban hangkivető szavak kapcsolati gráfjainak** vizsgálatával fejezem be. A komplex jegymérték és a komplex tengelymérték segítségével elkészített gráfokon láthatjuk (5.17.-5.20. ábrák), hogy a legkevésbé hangkivető módon viselkedő szavak hasonlósági csoportjai miképp rendeződtek át. A képek a 0,5-nél szorosabb hasonlósági viszonyokat mutatják[1], hogy a közepesen erős hasonlósági kapcsolatokat is láthassuk.

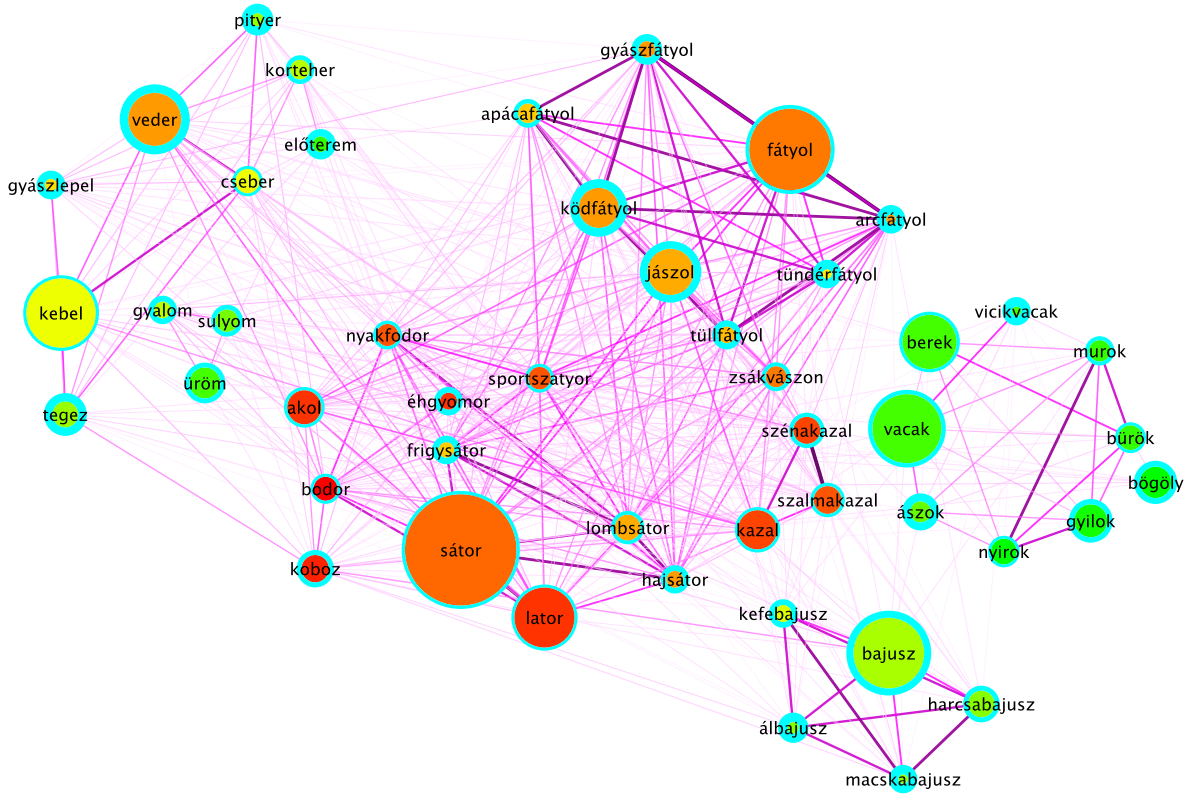
[1] Mivel itt sokkal lazább viszonyokat is szerepeltetek, mint az 5.3. alfejezet elemzéseiben, ezért nem tartottam szükségesnek, hogy a komplex jegymérték és a komplex tengelymérték esetében eltérő küszöbértékeket adjak meg.



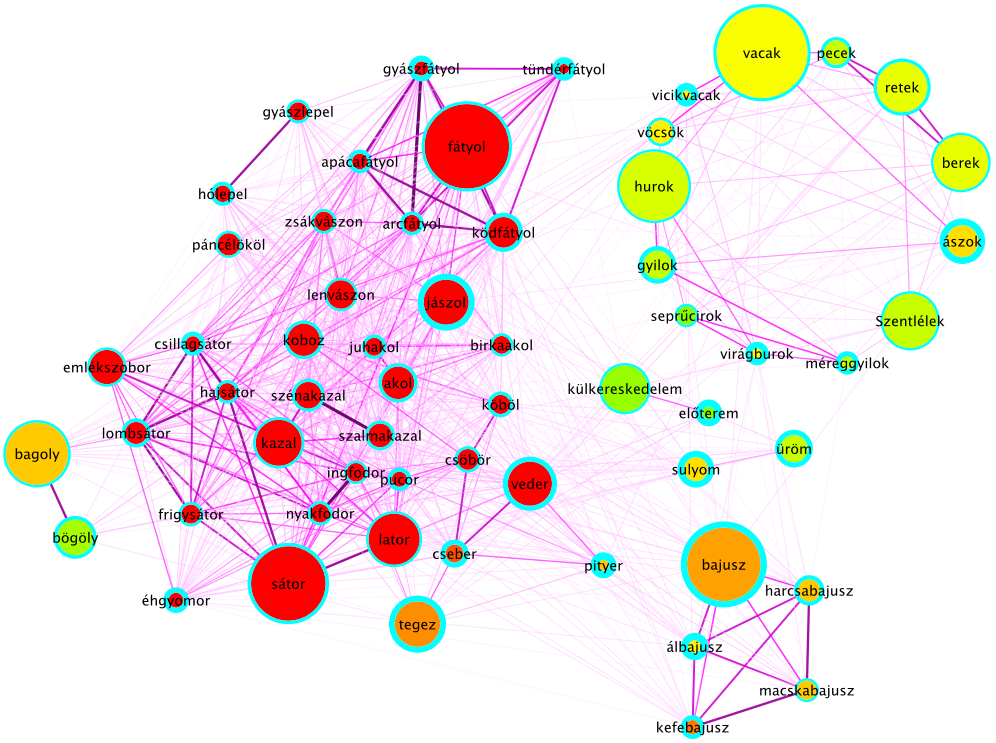
5.17. ábra: A hangkivető mintát kevesebb mint 90%-ban követő szavak kapcsolatai a komplex jegymérték alapján a *Szószablya Korpusz*ban



5.18. ábra: A hangkivető mintát kevesebb mint 90%-ban követő szavak kapcsolatai a komplex jegymérték alapján a *Google Gyakorisági Gyűjtés*ben



5.19. ábra: A hangkivető mintát kevesebb mint 90%-ban követő szavak kapcsolatai a komplex tengelymérték alapján a *Szószablya Korpusz*ban



5.20. ábra: A hangkivető mintát kevesebb mint 90%-ban követő szavak kapcsolatai a komplex tengelymérték alapján a *Google Gyakorisági Gyűjtés*ben

Az **5.17. ábra legtöbb szava egy-egy nagyobb gyakoriságú szó körül csoportosul**[1]. Egyedül az *-m* és a *-z* végűek nem követik ezt a fajta elrendeződést. Alapfeltevésem szerint a ritkább szavak előbb vesznek részt az analógiás változásban. Ezt a kijelentést korábbi megfigyeléseim és az ábra alapján úgy módosíthatjuk, hogy a ritkább szavak valóban korábban vesznek benne részt egy a hasonlóság alapján meghatározott csoporton belül, azaz egy szó sokkal hajlamosabb az analógiás kiegyenlítődésre, ha a hozzá hasonló szavak már elindultak ennek az útján, illetve ha hasonlósági csoportja típus- és példánygyakoriságát tekintve átlag alatti. A kevésbé hangkivető szavak legösszetartóbb csoportjai nem minden esetben egy-egy szó és a belőle létrehozott összetett szavak körül rajzolódnak ki. A *sátor* csoportjában találjuk a *lator, sportszatyor, nyakfodor, éhgyomor, bodor* szavakat is, a *fátyolében* is hasonló viszonyokat figyelhetünk meg. Ezekben az esetekben egy viszonylag gyakoribb, kevésbé hangkivető prototípus körébe sorolódhatnak be új szavak példánygyakoriságuk alapján gyengébb hasonlósági csoportokból. A *bodor* a gyengébb *fodor* (*fodor* csoport példánygyakorisága: 771, míg a *sátoré*: 16467) csoportból sorolódott át, hasonlóan alakulhatott a *sportszatyor* esete is. A *fátyol* esetében a *zsákvászon* csoportja erősebb, de csoportjának központi eleme, a *haszon* már viszonylag távolabb van tőle, és az *á*-k azonossága nagyobb súllyal eshet a latba, mint a tővégi *n*-eké. Ezt támasztják alá a 7. fejezet eredményei is, miszerint a magánhangzók hasonlósága szerepet kaphat még nem kiemelt helyzetben is. Az *akol* a *fátyol*-éval hasonló erősségű csoport tagja, de csoportjának prototípusa, a *pokol* sem követi már következetesen a hangkivető mintát, így nem húzza vissza a hangkivető sémába. Az *akol* kevésbé hangkivető viselkedését az is támogatja, hogy nincs kezdő mássalhangzója, ami a hangkivetők esetében kevésbé általános. Gyengébb kapcsolatokkal, de hasonló folyamatokat láthatunk a *-k* végű szavaknál és a *kebel-veder* csoportban is. A *koboz* és a *tegez*[2] annyira elütnek az összes

[1] Ez igaz a *kazal* csoportra is, bár ahogy az 5.3.3. alfejezetben láthattuk, a *-kazal* végűek közti különbség kisebb, mint ami a legtöbb prototípus és a hozzá kapcsolódó szavak közt megfigyelhető.

[2] Csak ez a két *-z* végű szó hangkivető, amelyeknek összetett alakjai igen ritkák, és a *morphdb.hu*-ban nem is szerepelnek: *íjtegez* (8), *nyíltegez* (5), *bőrtegez* (2), *ének-koboz* (1), *szólókoboz* (1), *lapkoboz* (1). A gyakorisági számok a *Szószablya Gyakorisági Szótár*ban szereplő egyes szám alanyesetű alakjukra vonatkoznak.

hangkivető szótól (az ingadozóktól is), hogy csoporthatástól függetlenül is elindulhattak a kiegyenlítődés útján. A *bögöly* szó ezen az ábrán nem szerepel, mert még az ingadozó hangkivetőktől is jelentősen különbözik. Ingadozásáért elsősorban a *bögölyök* alak a felelős. Részletesebben már az 5.3.2. és az 5.3.4 alfejezetekben is foglalkoztam vele. Az itt szereplő *-m* végű szavak (*sulyom, üröm, gyalom*) legfontosabb közös vonása szemantikai hasonlóságukon túl, hogy hasonlósági csoportjaikon belül szerkezetileg elszigeteltek (5.3.3. alfejezet), az *előterem* kevésbé hangkivető viselkedése mögött pedig a már az 5.2.3. alfejezetben bemutatott alakkeveredések állhatnak.

Az 5.18. ábrán ennek a **folyamatnak a kibomlását figyelhetjük meg** a *fátyol* és a *sátor* esetében, amelyek a korábbi csoportokból újabb elemeket vonzanak be: *birkaakol, juhakol, ingfodor, lenvászon*, de néhány további gyengébb csoport tagjai is a körükbe kapcsolódnak: *köböl, csöbör, páncélököl*. Ilyen típusú viselkedést a jelentősen felduzzadt *-k* végűek csoportjában már nem tudunk megfigyelni, mert ott több nagy gyakoriságú elem is szerepet kap, a kapcsolódások viszont meglehetősen lazák. A csoport kialakulása mögött nem lehet egységes okot megfigyelni. A korábban magában álló *bögöly* bevonzhatta maga mellé az alakilag szintén magányos *bagoly*-t.

Az 5.19. ábra alapján tovább árnyalhatom a *zsákvászon*-ról és az *-m* végűekről tett kijelentéseimet, mert felépítésüknek köszönhetően, ha csak gyengébben is, de több elemmel vannak kapcsolatban, így kiegyenlítődési folyamatuk közeli forrásokból is ösztönzést kap. A korábban magányos *tegez* és különösen a *koboz* itt valamelyest már kapcsolódik a teljes struktúrába. A viszonylag önálló kisebb csoportok jól megállnak magukban viszonylag sűrű kapcsolataiknak köszönhetően (*-k* végűek, *bajusz, kebel-veder* csoport). Az 5.17. ábrán is szorosabban kötődő *sátor, fátyol, kazal* csoportok szerkezeti hasonlóságait a komplex tengelymérték pedig még jobban kiemeli. A tengelymértéken alapuló gráfnál a *Google Gyakorisági Gyűjtés* esetében (5.20. ábra) a legszembetűnőbb, hogy az **újabb szavak bekapcsolódásával a központi boly elkezdett összeérni**, egyre több a kapcsolat, amely a változást tovább gyorsíthatja[1].

A kapcsolatok áttekintése során láthattuk, hogy a szavak **analógiás kiegyenlítődése** nem önállóan, hanem **csoportosan** mehet végbe. Az egyes bolyok

[1] Ebbe a központi bolyba a *tegez* is bekapcsolódott már, de párja, a *koboz* nem.

tagjai egymást erősítik a hasonló viselkedésben, azaz a többek által megfigyelt csoportjelenségek (Bybee 2001, 2010, Ernestus és Baayen 2004) hatással vannak arra is, hogy milyen módon gyengül a hangkivető minta bizonyos szavak esetében.

5.4.3. Az átlagostól eltérő egyedi viselkedés

A változás általános összefüggéseinek áttekintése után az **egyes szavaknak a folyamatban való viselkedését** veszem szemügyre a komplex jegymérték alapján készített hasonlósági gráfok segítségével. Első lépésben azt vizsgálom meg, hogy az egyes szavak mely toldalékainak hangkivetési mértékében tapasztalhatunk nagy kilengéseket, az átlagos mintázattól való jelentősebb eltéréseket[1]. A vizsgálatból kizártam az olyan szavakat, amelyek a ***Szószablya Korpusz*ban 100-nál kevesebb releváns alakkal szerepeltek**, mert az adathiány is okozhat szokatlannak mutatkozó viselkedést[2]. Ennek a kritériumnak összesen 424 szó felelt meg, azaz az összes vizsgált hangkivető főnév 35%-a.

Egy szó átlagostól eltérő viselkedésének azonosítására az **egyes toldalékok hangkivetési mértéke közt mérhető szórást választottam**, amelyet korrigáltam a szó hangkivetésének a mértékével, hogy a kevésbé hangkivető szavak ne kapjanak túlzott súlyt, hisz ezek esetében nagyobb szórásra számíthatunk. A *Szószablya Gyakorisági Szótár*ból 18 olyan szót választottam ki, amelyek szórásának és hangkivetési mértékének szorzata nagyobb volt, mint 0,15, ami arra utal, hogy ezen szavak egyes toldalékos alakjainak viselkedése meglehetősen eltérő:

---

[1] A vizsgálatban csak az átlagostól eltérő hangkivetési mértéket tanulmányoztam, azt nem, hogy egyes toldalékokat az átlagossal megegyező vagy attól eltérő mértékben használnak-e a beszélők. Ez minden bizonnyal önmagában is egy érdekes kutatás tárgya lenne, de ennek vizsgálata túlmutatna disszertációm keretein.

[2] Mint később látni fogjuk, így is maradtak vizsgált anyagomban olyan szavak, amelyeknek viselkedése az adathiány miatt torz képet mutat. Ha azonban a küszöbértékemet olyan magasra emeltem volna, hogy ezek a szavak ne kerüljenek be a vizsgálatba, akkor értékes, az összefüggések átlátásához szükséges adatokat vesztettem volna.

berek, bronzérem, bugyor, érdekvédelem, fátyol, földieper, kapor, kazal, kisterem, koboz, különterem, lator, retek, sátor, sulyok, szemérem, Szentlélek, széplélek, tülök, üröm

A *Google Gyakorisági Gyűjtés*ből 27 szót választottam ki hasonló feltételek mentén[1]. Ezek nagyobb száma azzal függhet össze, hogy az erőteljesebb ingadozás esetén várható a nagyobb szórás az alkalmazott korrekciós eljárás ellenére is:

ajak, akol, barom, bögöly, cirok, eper, fátyol, gyilok, iker, kapor, kazal, koboz, ködfátyol, kristálycukor, különterem, lator, majom, pocok, retek, sátor, sulyok, Szentlélek, szurok, teher, tejcukor, üröm, veder

A két gyűjtésben összesen 11 közös szó volt:

fátyol, kapor, kazal, koboz, különterem, lator, retek, sulyok, sátor, Szentlélek, üröm

A további vizsgálatból kizártam a még a *Google Gyakorisági Gyűjtés* tisztítása után is bizonytalan értékű szavakat, amelyeknek gyakorisági számait tulajdonnévi alakok is megnövelhették (*Sulyok, Üröm*). A kiválasztott szavaknál más esetben homonim alakok negatív hatása előzetesen nem volt feltételezhető.

[1] A *Google Gyakorisági Gyűjtés* szavait egy arányos, 3500 előforduláshoz kötött küszöb felett vizsgáltam volna csak, de azok a szavak, amelyek legalább 100 releváns alakjukkal előfordultak a *Szószablya Korpusz*ban legalább 3500 alakot a *Google Gyakorisági Gyűjtés* esetében is fel tudtak mutatni. A könnyebb összehasonlíthatóság érdekében a számításokat így ugyanazon szavak körében hajtottam végre.

5.4.3.1. Egyedien viselkedő szavak a *Szószablya Gyakorisági Szótár* alapján

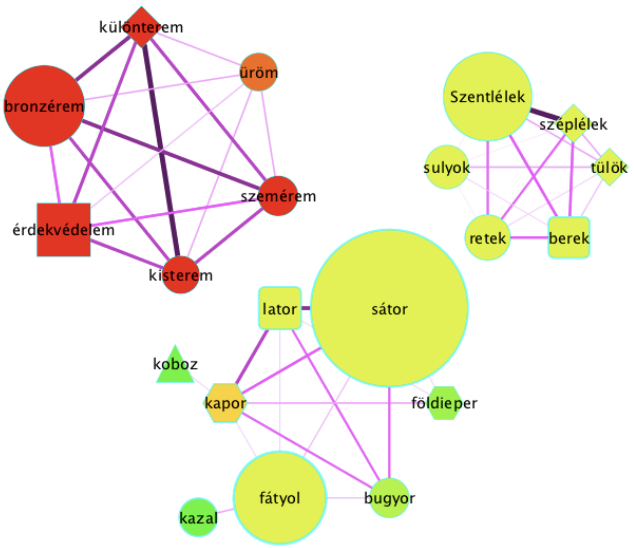
E szavak (5.21. ábra) viselkedése összességében megfelel a minta jellegzetességeinek, mert **legelőrébb szuperesszívuszos alakjaik járnak az analógiás kiegyenlítődésben** (48%-os hangkivetési mérték). Azaz a szavak többségénél a magas szórás a hangkivetőkre jellemző általános tendencia „túlhajtásából" fakad. A hangkivetők vizsgált toldalékainak összelőfordulásából a szuperesszívuszos alakok 4,7%-ot hasítanak ki, míg a toldalékonkénti hangkivetés mértékében nagy szórást mutató szavaknál 2,8%-ot. A szuperesszívusz esetében megfigyelhető alacsony hangkivetési mérték oka így az lehet, hogy ezen alakok esetében bizonytalanabb már a nyelvérzék a „megfelelő" használattal kapcsolatban, ezért az analógia nagyobb szerepet kap a szuperesszívuszos alakok produkciójában, hisz az ezekkel kapcsolatos emléknyomok gyengébbek. A többi toldalék esetében a hangkivetés mértéke 78% és 93% közt mozog.

Néhány szó esetében azonban nem a szuperesszívuszos alak a legkevésbé hangkivető[1]: *berek, kapor, koboz* (tárgy), *érdekvédelem, különterem* (többes szám)[2], *különterem* (E.1 birtokos), *kapor* (E.3 birtokos), *koboz* (E.3 birtokos több birtokkal). A „T.3 birtokos", illetve az „E.3 birtokos több birtokkal" alakok esetén egyetlen szónál sem kisebb a hangkivetés mértéke, mint a szuperesszívusz esetében. A továbbiakban azon szavak viselkedését vizsgálom meg alaposabban, amelyeknél nem a szuperesszívusz hangkivetési mértéke a legalacsonyabb.

---

[1] A szavak mögött zárójelben található, hogy melyik toldalékkal viselkednek kevésbé hangkivető módon, mint a szuperesszívusszal. Egy szó, mint pl. a *koboz* többször is szerepelhet.

[2] Ide vehetném a *Szentlélek*-et is egy nem hangkivetéses többes számú alakja alapján, amellyel nem áll szemben hangkivetéses alak, így látszólag a többes számban nem hangkivető módon viselkedik, de ez az egy *Szentlélekek* előfordulás nagy valószínűséggel a *Szentléleknek* alak elgépelésének tudható be.



5.21. ábra: A szórás alapján egyedien viselkedő szavak a *Szószablya Korpusz*ban a komplex jegymérték alapján számított hasonlósági gráfstruktúrában. Az ábra mindenben a korábbi ábrázolási konvenciókat követi. A csomópont alakja arra utal, hogy egy-egy szó melyik toldalékkal a legkevésbé hangkivető.

kör: szuperesszívusz
sarkos négyzet: többes szám
lekerekített sarkú négyzet: tárgy
hatszög: E.3 birtokos
rombusz: E.1 birtokos
háromszög: E.3 birtokos több birtokkal

Az ***-m* végű szavak csoportjában** az *érdekvédelem* többes számban való kevésbé hangkivető viselkedése 3 szórványos alaknak tulajdonítható. Az egyedi viselkedés mögött azonban az is állhat, hogy ennek a szónak a többes száma annyira ritka (az összes vizsgált toldalékos alak 0,6%-a), hogy esetében a beszélők már inkább hagyatkoznak az analógiára. Az *-alom*/*-elem* végűek között is különösen alacsony arány ez, amelyeknél a hangkivetést elváró toldalékos alakok 12,5%-áért a többes szám a felelős, míg ez az arány a többi hangkivető főnév esetében 19,3%.

Az ***-l, -r, -z* végűek csoportjában** a *lator*-nak a *Szószablya Gyakoriségi Korpusz*ban nincsenek szuperesszívuszos alakjai, ezért várható módon a szuperesszívusz után a hangkivetéssel legkevésbé együttjáró tárgyragos alakok mutatják az esetében a legkisebb mértékű hangkivetést. A *Google Gyakoriségi Gyűjtés*ben is csak ritkán

fordulnak elő szuperesszívuszos alakjai: a *lat(o)ron kívül, könyörül a lat(o)ron* kifejezésekben elsősorban biblikus vagy archaikus szövegekben. A *koboz* esetében az „E.3 birtokos több birtokkal" alakban nem hangkivető viselkedése egy alaknak tudható be, de ha ettől eltekintünk, akkor is tárgyas alakjaiban követi a hangkivetést a legkevésbé. Szuperesszívuszos alakjainak következetesebb hangkivetése annak köszönhető, hogy meglehetősen nagy arányban fordulnak elő (21,4%).

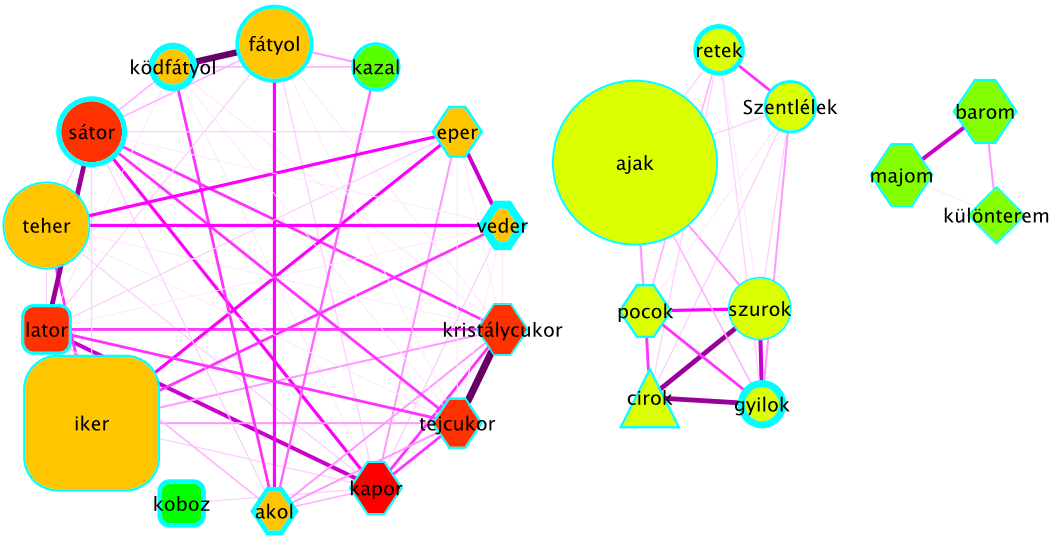
A *földieper, különterem, széplélek, tülök* esetében egy-egy alak okozza a látszólag szokatlan viselkedést. A *kapor* és a *berek* szavaknál a kiegyenlítetlen viselkedést a *kapóra, béreket* ékezetmentes alakjai eredményezik, ezért ezeket szintén érdemes kihagyni a vizsgálatból.

5.4.3.2. Egyedien viselkedő szavak a *Google Gyakorisági Gyűjtés*ben

A *Google Gyakorisági Gyűjtés*ben a toldalékonkénti hangkivetés mértékében nagy szórást mutató szavaknál is a **szuperesszívusz (62%) követi legkevésbé a hangkivető mintát**, de ez kevésbé tér el a többi toldalék 65-93%-ig terjedő hangkivetési mértékétől (tárgy: 78%, többes szám: 94%, E.1 birtokos: 90%, E.3 birtokos: 65%, T.3 birtokos: 93%, E.3 birtokos több birtokkal: 93%). Ezzel áll összhangban, hogy a szavak releváns alakjainak 3,8%-át teszik ki a szuperesszívuszos alakok, míg a teljes gyűjtésben 4,2%-ot képviselnek, azaz közel azonos arányban vannak jelen, eltérésük a t-próba alapján sem szignifikáns. A szupresszívuszos alakjaikban legkevésbé hangkivető szavak egy részénél (*ajak, fátyol, sátor, kazal, ködfátyol)* viselkedésük nyitótő jellegükkel hozható kapcsolatba. Szupresszívuszuk és tárgyuk (az *ajak*-é nem) kevésbé jár együtt a hangkivetéssel, mivel ezek esetében a nyitásnak nem kell feltétlenül érvényre jutnia. A szuperesszívusz csak középső nyelvállású kötőhanggal kapcsolódhat, a tárgy pedig kapcsolódhat kötőhang nélkül is, így a tő nem nyitó viselkedése a tárgy esetében csak a nyílt kötőhang hiányával kerül kifejezésre. Ezzel szemben a többes számban, és az „E.1 birtokos" alaknál a beszélőnek mindenképpen kell kötőhangzót alkalmaznia, amely nyílt is lehet. Ha egy szó több rendhagyó jeggyel bír, akkor azok a beszélők számára

feltűnőbbek és szorosan összekapcsolódnak, így könnyebben megmaradnak. Ennek megfelelően szétválásukra, legalábbis vizsgált szóanyagomban, nincsenek példák (**sátrok, *sátorak*). Esetünkben a nyitás[1] a hangkivetés megmaradását támogatja, ami a többes számnál a hangkivetés erőteljesebb rögzülését eredményezi. Ez a tendencia áll részben a hangátvetéses (*-pehely, -teher, -kehely*) és a tőbelseji magánhangzó rövidüléses (*-lélek*) alakok viselkedése mögött is. A *Szentlélek* (5.4.3.1. alfejezet) és a *gyilok* (5.2.3. alfejezet) viselkedését már korábban tárgyaltam, sajátos viselkedésük mögött a bemutatott okok fennmaradása állhat. A *szurok* rendellenes viselkedése a *(spam) szűrőkön* alak ékezetmentes változatának tudható be.

A következő szavak esetében **nem** a **szuperesszívusz a legkevésbé hangkivető**: *akol, cirok, eper, iker, kapor, koboz, különterem, lator, tejcukor, üröm, veder* (tárgy), *kapor, különterem* (többes szám), *akol, cirok, különterem* (E.1 birtokos), *veder, akol, eper, kapor, barom, tejcukor, majom, kristálycukor, pocok, cirok, iker, különterem* (E.3 birtokos), *cirok* (T.3 birtokos), *különterem, eper, kapor, cirok* (E.3 birtokos több birtokkal).



5.22. ábra: A szórás alapján egyedien viselkedő szavak a *Google Gyakoriság Gyűjtés*ben a komplex jegymérték alapján számított hasonlósági gráfstruktúrában.

[1] Tudomásom szerint a nyitás esetében nincs analógiás kiterjesztődési folyamat.

Az 5.22. ábrát az 5.21. ábrával összevetve láthatjuk, hogy a **korábbi csoportok megmaradtak, de elemeik sok esetben megváltoztak**. A továbbra is toldalékonkénti hangkivetés mértékükben nagy szórást mutató szavak közt azonban különösebb kapcsolatot nem tudunk megfigyelni. Jelentésükben heterogének, a *Szentlélek*-et leszámítva egységes *CV(:)CVC* szerkezetük a hangkivetők közt nem egyedi. Ennél szorosabb kapcsolat csak a már elemzett *fátyol, sátor, kazal* közt van, amelyek élen járnak az analógiás kiegyenlítődésben, és hasonlósági csoportjuk lokális prototípusainak is tekinthetők. A továbbiakban azon szavakat vizsgálom meg alaposabban, amelyekről az 5.4.3.1. alfejezetben nem esett szó, és nem a szupresszívuszos alakjaiknál a legalacsonyabb a hangkivetés mértékük.

Az *-l, -r, -z* végűek **csoportja vált a legnagyobbá**, azonban a csoportba kerülő új szavak esetében a nagy szórásnak nincs egységes oka. A *tejcukorja*, *kristálycukorja* alakok megjelenésére a gyakoribb, de hangkivetési mértékében kisebb szórást mutató *cukor* szó *cukorja*/*cukora* alakjai lehetnek hatással, habár a hozzájuk formailag közelebb álló alakok jelentésükben távolabbiak[1]. Az *eper* és a *veder* látszólagosan hasonló viselkedése véletlenszerű, mind a kettő mögött más lexéma velük homonim alakjai is növelik a nem hangkivetéses „E.3 birtokos" alakok gyakoriságát (ékezetmentes *epére*, olasz *vedere*). Az *iker* esetében a tárgynál a legalacsonyabb a hangkivetés mértéke, viselkedése a *lator*-hoz hasonló. Túl kevés *ikeren-ikren* alakunk van a gyűjtésben (az összes vizsgált alak 0,2‰-e) ahhoz, hogy azok alapján a szuperesszívusz viselkedéséről bármi biztosat állíthassunk. Az *akol* „E.3 birtokos" alakja mind hangkivető, mind nem hangkivető alakjaiban keveredik homonim tulajdonnévi alakokkal (*Akla Kft. és Bt.*, *Apát akolja*). Ezeket figyelmen kívül hagyva tárgyas alakjai a legkevésbé hangkivetők, mert szuperesszívuszos alakjai elsősorban a rögzült *aklon kívül/belül* kifejezésekhez kötődnek.

A ***-k* végűek csoportjában** a *pocok* kiegyenlítetlen viselkedése mögött álló *pocokja* alak korábbra adatolható. A hozzá nagyon hasonló *cirok* rokon viselkedést mutat, (*cirokjai:* csak 1 előfordulás), mert a következő legkevésbé hangkivető toldaléka az „E.3 birtokos" (91%-os hangkivetés), amelynek valós hangkivetési mértékét meglehetősen

[1] A *cukorja*/*cukora* elsősorban a vér cukorszintjére utal, míg a *cukra* emellett az édesítésre használt anyagra is.

nehéz megbecsülni, mert a *cirka* 'nagyjából' homonim alak egybeesik a hangkivetéses alakváltozatával. Az ettől való elkülönülési szándék is erősíti a *cirokja* alakot.

Az ***-m* végű csoportban** a korábbiak helyett új szavak jelentek meg. A korábban egyedi módon viselkedő *-m* végű szavak továbbra is mutatnak némi eltérést a hangkivető mintától, de már kiegyenlítettebb módon. A *majom* és a *barom* egyedi viselkedését egy sajátságos konstrukcióban (7) való használatuk okozza.

(7)

élőlényre utaló főnév + jA	↔	műveletlen, ostoba ember + indulati nyomaték, elítélő hangnem

A **konstrukcióban** előfordul a hangkivető főnevek közül még: *féregje, tulokja, piszokja*, de számos nem hangkivető szó is szerepelhet benne: *németje, rohadékja, szemétje, köcsögje*[1] stb. Mint az utóbbi két példában is láthatjuk, ezek alakjukban is különbözhetnek a konvencionális jelentésben használt alakoktól (*szemete, köcsöge*). A hangkivetéses alakok fonotaktikai okokból ezekben az esetekben nem lehetségesek (**majmja*, **barmja*), ezért használják a nem hangkivetéses tővariánsokat a konstrukcióban, bár szórványosan lehet találkozni a szerkezetben a *barma*, *majma* alakokkal is, amelyeknél a hangkivetés megőrzése érdekében inkább a konvencionális „E.3 birtokos" alakot használják. A *különterem* látszólagos ingadozása az „E.1 birtokos" alakban ismét csak 1-1 hangkivető és nem hangkivető alaknak tudható be.

A toldalékaik hangkivetési mértékei közt nagy szórást mutató szavak áttekintéséből az látszik, hogy a **csoportot csak a szuperesszívusz erős ingadozása köti össze, ami ritkább használatával hozható kapcsolatba**. Ez azonban a *Google Gyakoriság Gyűjtés*ben már sokkal gyengébben jelentkezik, ami arra utal, hogy ezt olyan tényezők is befolyásolhatják, amelyeket nem sikerült felderítenünk. Ennél kevesebb alakot érintő,

[1] A konstrukció csak élőlényekre alkalmazható, azaz a *köcsög, rohadék, szemét* stb. szavakat előbb a szlengben emberre kellett használnia a beszélőknek, és csak ezután vált lehetővé használatuk a szerkezetben, így kevésbé elfogadhatóak a konstrukcióban a *maradék, hordalék, lom* stb. szavak.

de konzisztensebben megmaradó viselkedési mintázatot láthattunk a kevésbé hangkivető nyitótövek esetében. Ezeknél a kiindulási viselkedésben megmutatkozó eltérések befolyásolják a változásban való eltérő részvételt is. E tendenciáktól az alaposabb vizsgálat során csak néhány erősen motivált eltérést találtunk.

Ezek alapján kijelenthetjük, hogy a **szavak** túlnyomó többsége t**oldalékaik viselkedése szempontjából viszonylagosan egységesen vesz részt a változásban**, amelyet azonban az egyedi használat megbonthat, és ezek az egyedi jelenségek a mintázat gyengüléséhez vezethetnek. A toldalékonkénti hangkivetés mértékében nagy szórást mutató szavak hasonlósági csoportjaiban a kapcsolatok véletlenszerűnek mondhatóak, a teljes csoporton belül a viselkedés gyakran heterogén, a szavak hatása egymásra elenyésző lehet.

5.4.4. Gyorsan változó szavak

A szavak paradigmaváltásával kapcsolatos kutatások elsősorban azt vizsgálják, hogy egy változásban mely szavak és miért vesznek részt, illetve melyek azok a cellái egy paradigmának, amelyek a legsérülékenyebbek, a legkönnyebben lehetnek a változás célpontjai. Az egyes szavak változásának sebessége azonban nem tárgya az ilyen vizsgálatoknak, pedig ha meg tudjuk állapítani, hogy az **egyes szavak miért** vesznek részt **más tempóban** a változásban, az hozzájárulhat a változási folyamatok alaposabb megértéséhez. Ha az egyes toldalékolt alakok hangkivetési mértékének átlagát veszem (azaz külön veszem pl. a *sátor*+ACC és a *sátor*+PLUR hangkivetési mértékét), akkor a *Szószablya Gyakorisági Szótár* (2003) és a *Google Gyakorisági Gyűjtés* által reprezentált 2010-es állapot között 25%-ban csökkent a hangkivetés mértéke, amely eltérés azonban nem bizonyult szignifikánsnak.

Ezúttal alaposabb vizsgálat alá azokat a szavakat vetettem, amelyeknek a *Google Gyakorisági Gyűjtés*ben **legalább 50%-kal több hangkivető alakjuk** van a vizsgált toldalékoknál. Kizártam ezek közül azokat a szavakat, amelyeknek a *Google Gyakorisági Gyűjtés*ben még így sem volt 1%-nyi nem hangkivetéses alakjuk a releváns toldalékok

előtt. Ezek a szavak még így is az átlagosnál jobban követik a hangkivető sémát, így az esetükben a „felzárkózás", a gyorsabb változás viszonylag természetesnek tekinthető. Másrészt ebben a szűk 0,99 és 1 közti tartományban egy viszonylag kisebb változás is jóval jelentősebb aránybeli elmozdulást eredményezhet. A folyamatok megértését nehezítené, ha ezeknek az apró eltolódásoknak túlzott jelentőséget tulajdonítanánk. Ezúttal is **kizártam** a vizsgálatból az olyan szavakat, amelyek a ***Szószablya korpusz*ban 100-nál kevesebb releváns alakkal szerepeltek.** Ezen kritériumok mentén **78 szót** választottam ki, amelyekből 8 korábban következetesen hangkivető módon viselkedett.

Legfontosabb közös jellemzőjük, hogy **nagy részük nem összetett szó (75%)**, míg a legalább 100 alakkal a *Szószablya Gyakorisági Szótár*ban is előforduló szavaknak csak 46,2%-ára nem áll ez. A *Szószablya Gyakorisági Szótár*ban még 98,9%-ban követik a hangkivető mintát, a *Google Gyakorisági Gyűjtés* tanúsága alapján azonban már csak 95,7%-ban. Átlagos gyakoriságuk közel kétszer akkora (293 ezer: 160 ezer), mint a *Szószablya Gyakorisági Szótár*ban legalább 100 alakot számláló többi szóé. Változásukat serkenthetik azok a kevésbé hangkivető összetett szavak, amelyeknek utótagjai, és a gyakori, kevésbé hangkivető hasonló szavak is, mint *sátor, bajusz, fátyol*. Számolnunk kell annak a hatásával is, hogy az alapszavak az informális kommunikációban nagyobb szerepet kaphatnak, ahol a hangkivető normát kevésbé követik. Mivel ilyen típusú szövegek nagyobb arányban lehetnek a *Google Gyakorisági Gyűjtés*ben, ez is állhat a hangkivetés mértékének gyorsabb csökkenése mögött. Ezeken túl még néhány részösszefüggést is megfigyelhetünk a szavak viselkedésében, amelyeket a következőkben tárgyalok.

	dinamika	Google gyakoriság	Szószablya hangkivetés mértéke	Google hangkivetés mértéke
külkereskedelem	42,20	17127	99,7%	88,4%
bokor	41,73	225525	100,0%	98,3%
mozivászon	35,71	25004	99,9%	96,7%
cirok	28,20	127199	99,8%	94,5%
védelem	25,03	1040320	99,8%	95,8%
eper	24,14	41632	99,7%	91,9%
izom	21,38	216427	99,9%	98,8%
cukor	21,33	252017	99,8%	96,7%
jövedelem	17,48	534781	99,8%	96,6%
kereskedelem	17,04	223141	99,5%	91,8%
haszon	16,33	420792	99,9%	97,7%
iker	15,75	791047	99,6%	94,4%
szaloncukor	15,55	20077	99,5%	92,7%
teher	14,90	359000	99,9%	97,8%
élelem	14,79	79273	99,8%	97,4%
türelem	13,95	283836	99,9%	98,4%
porcukor	13,51	22922	99,7%	96,0%
nagykereskedelem	13,39	158629	99,6%	95,0%
szerelem	10,63	805586	99,7%	96,3%
árok	10,51	129027	99,6%	95,4%
díszterem	10,40	96324	99,7%	97,4%

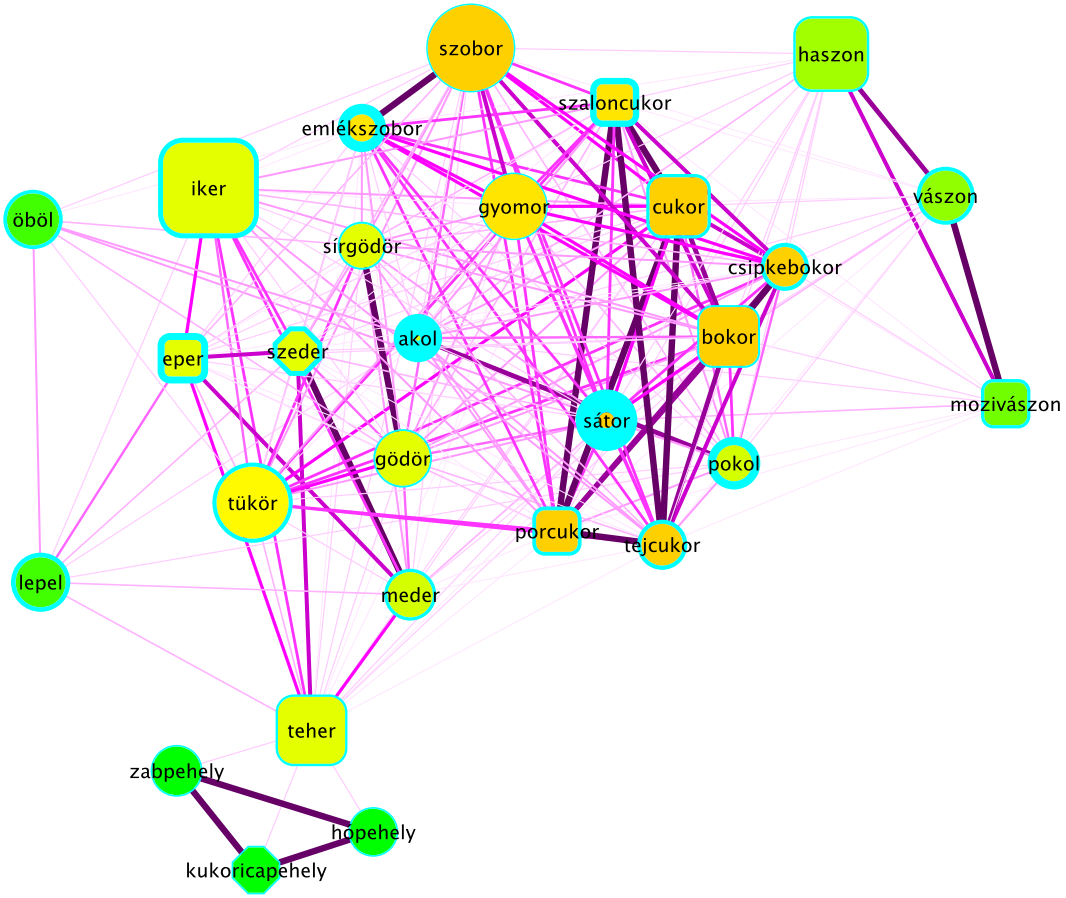
5.18. táblázat: Hangkivető szavak, amelyeknek leggyorsabban csökkent hangkivetési mértékük[1]

Korábban stabilan követték a hangkivető sémát azok a szavak, amelyeknek a *Google Gyakorisági Gyűjtés*ben legalább **10-szer több hangkivetéses alakjuk** van összes releváns toldalékos alakjukhoz viszonyítva (5.18. táblázat), mint a *Szószablya Gyakorisági Szótár*ban, így esetükben inkább csak az analógiás kiegyenlítődésbe való bevonódásról beszélhetünk. A legnépesebb és a hangkivetésben legstabilabb *-alom* végűek nem képviseltetik magukat köztük. 7 olyan szó található a gyorsan változó szavak közt, amelyek korábban is az átlagosnál kevésbé követték a hangkivető sémát, és azóta is dinamikusan változnak: *akol, hurok, pocok, pokol, retek, sátor, Szentlélek, vászon.* A szavak közt nincs *-m* végű, és két szótagosak (kivéve: *Szentlélek*)

[1] A *cirok* kiugró példánygyakoriságát az „E.3 birtokos" alakjában a *cirka* 'nagyjából' szóalakkal való keveredés okozza, azonban a *cirka*-hoz és a *cirok*-hoz tartozó alakok szétválogatása nem lehetséges.

Ezek közül a legdinamikusabban a *retek* szó változik. 6-szor nagyobb arányban vannak nem hangkivetéses alakjai a *Google Gyakoriségi Gyűjtés*ben. Ez elsősorban az *RTL klub* televíziós csatorna gúnynevének gyors terjedéséhez köthető (*reteken, retek TV* stb.), amit nem lehet kiszűrni a nagybetűs használat miatt sem, mivel a szlenget használó szövegekben elsősorban kisbetűs írásmód gyakori a tulajdonnevek esetében is. Az alakok egy nem jelentős részéért az *réteken* ékezetmentes alakjai is felelősek lehetnek (hasonlóan *hurokat : húrokat*). A *Szentlélek, akol, pocok, sátor, akol* sajátos viselkedésének okait már korábban tárgyaltam. A *vászon* esetében a gyors kiegyenlítődés mögött az áll, hogy egyre gyakoribb az eredetitől elszakadóban lévő jelentésében a használata (‘vetítésre alkalmas anyag’). Az új jelentéshez eltérő viselkedés is kapcsolódik.

A komplex jegymérték alapján gráfstruktúrába rendeztem a gyorsan változó szavakat, hogy kapcsolataik alapján is megvizsgálhatóak legyenek. A szavak 3 nagyobb csoportba rendeződtek, amelyeket az 5.23-5.25. ábrák mutatnak be.



5.23. ábra: *-l, -r, -n, -ly* végű gyorsan változó szavak

lekerekített sarkú négyzet: a korábbinál arányaiban 10-szer több hangkivetés nélküli alak
kör: gyorsan változó, de nem kiugró szavak
nyolcszög: csak a *Google Gyakorisági Gyűjtés*ben nem következetesen hangkivető

Az 5.23. ábrán **viszonylagosan szoros kapcsolatban** lévő, hangkivetésükben enyhén bizonytalan szavakat láthatunk. A *-or, -ol* végű szavakat a közepes mértékben hangkivető *sátor*, *akol* és *pokol*, illetve az azokból összetétellel létrehozott, de az ábrán nem látható szavak indíthatták el a gyorsabb változás útján. A szoros, közeli kapcsolatok hatására a nagyobb gyakoriságú szavakból is néhány bevonódott a változásba. Az ábrán látható szavakhoz végüknek köszönhetően a tárgy könnyedén kapcsolódik kötőhang nélkül, így esetükben a tárgynál a legelőrehaladottabb a hangkivetés visszaszorulása (tárgy hangkivetési mértéke: 86,8%, szuperesszívusz hangkivetés mértéke: 92,5%, összes releváns toldalékos alak hangkivetésének mértéke: 94,1%). Úgy tűnik, hogy a köznyelvben egyre inkább elfogadott lesz a nem

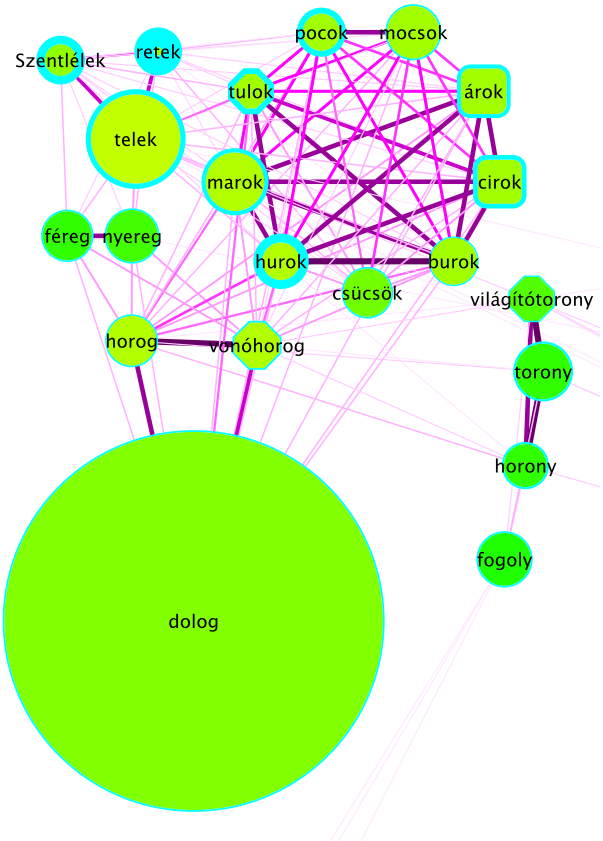
hangkivetéses alak és a *-t* toldalék kapcsolata kötőhang nélkül, mivel az ilyen alakok előfordulása elért egy kritikus tömeget. A kötőhang nélküli tárgy esetében a legnagyobb az alaki hasonlóság a hangkivetéses alakokhoz, ezért a beszélők számára kevésbé feltűnő az ingadozás, így jobban el is fogadják azt. Pontos magyarázatom azonban nincs arra, hogy miért ezek a szavak indultak el a változásban gyorsabban, mert a hangkivetés alakok aránya összes alakjukban 46,9% (összes *-r, -l, -n, -ly* végű hangkivető főnévnél: 43,2%), tárgyragjuk pedig a hangkivetéssel együttjáró toldalékok 31,2%-át teszi ki (összes *-r, -l, -n, -ly* végű hangkivető főnév: 33,8%), azaz a változás szempontjából relevánsnak mondható értékeik nem különböznek jelentősen a többi hasonló végű hangkivető főnévétől.

A **leggyorsabban változó szavak** több és szorosabb kapcsolattal rendelkeznek (5.23. ábrán melegebb, sárgás színűk van), azaz a más, kevésbé hangkivető szavakhoz való alaki közelség gyorsíthatja a változást (vö. 5.4.2. alfejezet). Ez leginkább szembetűnő a *cukor-bokor* végű szavakat tartalmazó részgráfban, ahol csak a legkevésbé hangkivetők közé tartozó *tejcukor* (*Google* hangkivetés: 95,4%; dinamika: 5,59) és *csipkebokor* (*Google* hangkivetés: 95,4%; dinamika: 3,28) nem változnak kiugró tempóban. Ez azzal magyarázható, hogy az egyes hasonlósági csoportokon belül a közel egyforma viselkedés az optimális, így még a ritka alakok sem szakadhatnak el jelentősen prototípusuk viselkedésétől (vö. 5.3.5. alfejezet). A leggyorsabban változó *szaloncukor* (*Google* hangkivetés: 92,6%; tempó: 15,55) viselkedésére adataim nem adnak magyarázatot. Leggyakoribb a hangkivetést is kiváltó toldaléka a tárgy, ami a hangkivetéssel együttjáró toldalékok 64,37%-ért felelős, de ez közel azonos a *cukor-bokor* gyorsan változó tagjainak 66,6%-ával[1]. A távolabb levő leggyorsabban változók közt az *iker*-nél is a tárgyrag játszik vezető szerepet a hangkivetés mérséklésében (13%-os hangkivetés), a hasonló végű és szerkezetű *eper* a hatására vonódhatott be a változásba. Az *iker* tárgyesetének kiugróan alacsony hangkivetési mértékére sincs egyértelmű magyarázat. A *szeder* változását[2] az *eper* serkentheti az alaki hasonlóságon túl nagyon

[1] Elképzelhető, hogy eltérő viselkedése a többi ritka *-cukor* végű szótól azzal magyarázható, hogy gyakrabban fordulhat elő informális párbeszédekben.

[2] Bárczi és mtsai (1967) korábbi változásaiban is analógiás forrásnak az *eper*-t jelölték meg.

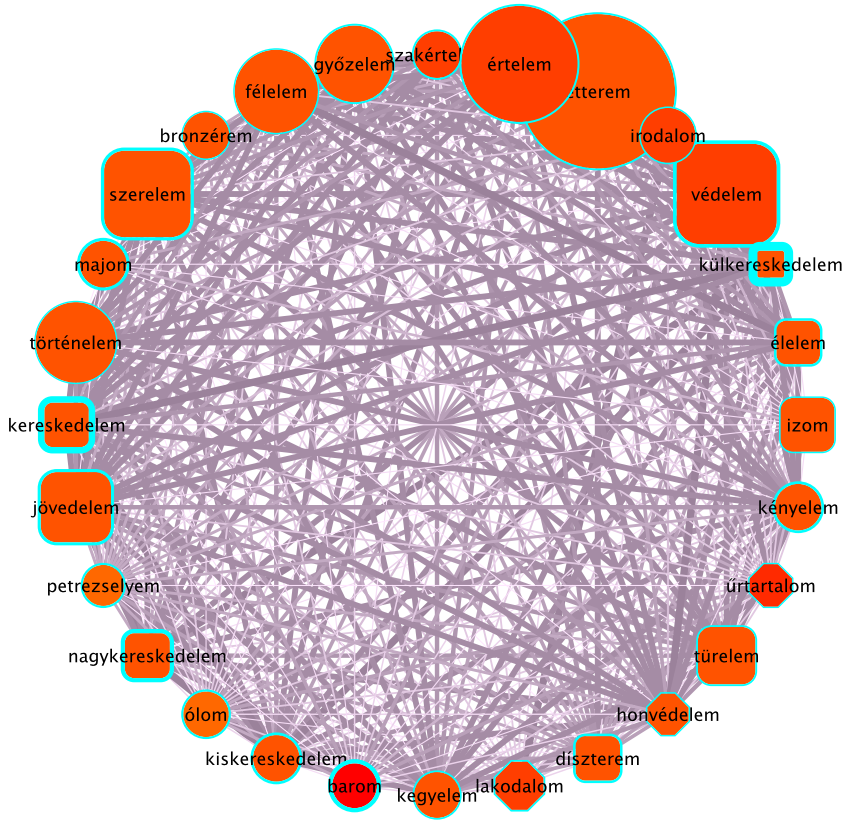
közeli jelentésével is. A többi gyorsan változó szótól viszonylagos elszigeteltségben lévő *lepel* és a hozzá leginkább hasonlító *öböl* a már lassabban változó, de náluk kevésbé hangkivető *kebel*-hez zárkózhatnak fel. Az 5.23. ábra további szavainál is a tárgynál a legalacsonyabb a hangkivetés mérték.



5.24. ábra: Dorzális végű gyorsan változó szavak

Az **5.24. ábrán** látható szavak legtöbbjének egyedi viselkedését már korábban elemeztük. A legnagyobb összetartó csoport változását több ingadozásra hajlamos szó együttes változása serkentheti, ennek tudható be a kiegyenlítődésben való átlagosnál gyorsabb részvételük. Egyedül a *fogoly* áll a kisebb, összetartó csoportokon kívül. A *fogoly* szó egyaránt jelenthet ‘elfogott embert’ és ‘egy madárfajtát’ is. Elekfi (2000) helyesen állapítja meg, hogy az eltérő toldalékolási mód (hangkivető és nem hangkivető) eltérő jelentésű szavakhoz kötődik (az ‘egy madárfajta’ jelentésű *fogoly* nem hangkivető), de mára ez az elkülönülés összekeveredett, és a nem hangkivetéses alakok is vonatkozhatnak emberre. Hangkivetési mértékének gyors csökkenésében szerepet kaphat a használatához kötődő megnövekedett bizonytalanság, amelynek hatását erősítheti a kevésbé hangkivető *bagoly* és *bögöly* közelsége is. A leggyakoribb hangkivető

szó, a *dolog* hangkivetési mértékének enyhe mérséklődése elsősorban tárgyesetű alakjainak tudható be. Lehetséges, hogy a bizonytalanság a tárgyas alakok használatában az informális nyelvhasználatban való vélhetőleg gyakoribb előfordulásának és a hangkivető főnevek közti viszonylagos magányosságának köszönhető. Az 5.24. ábrán látható szavaknak általánosságban sajátja a kötőhangzóval megjelenő, nem hangkivetéses tárgyas alakoknak betudható hangkivetési mértékcsökkenés, amelytől egyedi eltérések lehetnek: *Szentlélek, marok, retek, fogoly* (szuperesszívusz), *pocok, csücsök, tulok, horog* (E.3 birtokos). Azonban ezek is a hangkivetéssel kevésbé együttjáró toldalékok közül kerülnek ki, változásuk mögött egységes ok nem fedezhető fel.



5.25. ábra: *-m* végű gyorsan változó szavak[1]

Az ***-m* végűek** csoportjára jellemzőek a nagyon szoros hasonlósági viszonyok és a viszonylagosan stabil hangkivetés. A 29 tagú csoport alulreprezentált tagjai az *-om*

[1] Az élek színét fakóbbra vettem, hogy a csomópontok címkéi olvashatóbbak legyenek.

végű szavak (7 szó), amelyekből csak a *lakodalom* és az *irodalom* *-alom* végű (28%-a az *-om* végűeknek a csoportban), míg ha az összes hangkivető főnevet vesszük, akkor az *-alom* végűek aránya 82% az *-om* végűek közt (332 *-om* végű szóból 268 *-alom* végű), azaz az *-om* végűekből a hangkivetők közt kevésbé tipikusak kezdtek el gyorsabban részt venni az analógiás kiegyenlítődésben. Hasonló aránytalanságot tapasztalunk, ha a csoport *-om* végű szavainak előfordulási számait összevetjük az *-em* végűekével (22 szó a gyorsan változó *-m* végűek csoportjában), mivel arányuk 1:3,14-hez (7:22). Ugyanakkor az összes hangkivető főnevet nézve arányuk fordított, hisz 332 *-om* végű szó áll szemben 250 *-em* végű szóval, azaz arányuk 1,32:1[1]. Tehát a nagyon stabil *-om*, *-em* végű szavak közül is a csoportba kevésbé mélyen beágyazott *-em* végűek hajlamosabbak a hangkivető sémától való eltérésre. A két csoport közti szakadás annak tudható be, hogy az *-alom* végűek számosabbak, mint az *-elem* végűek, amelyeknek a végük is kevésbé jellemző a hangkivető főnevekre.

Az *-m* végűek csoportjában a ***-kereskedelem*** **végűek** követik a hangkivető viselkedést a legkevésbé (*külkereskedelem* 85%, *kereskedelem* 92%, *nagykereskedelem* 95%, *kiskereskedelem* 97%). A többi *-kereskedelem* végű szóhoz képest összes alakjukat figyelembe véve kiugróan magas a hangkivető alakok aránya (76-79%), ami talán az alapszót gyakoriságában sokszorosan meghaladó *-i* végű alakoknak tudható be (8), amelyek más *kereskedelem* végű szavaknál nem, vagy csak szórványosan fordulnak elő (9):

(8) *kereskedelem* (26015) : *kereskedelmi* (129920)
külkereskedelem (1443) : *külkereskedelmi* (7932)
nagykereskedelem (1463) : *nagykereskedelmi* (4973)
kiskereskedelem (1646) : *kiskereskedelmi* (8061)

[1] Az *-om* és *-em* végűek csoportjain belül a legnagyobb alosztályokat képező *-alom*, *-elem* végek viszonya még aránytalanabb 268:159, azaz 1:1,62. Ennek oka, hogy több nem *-elem*-re végződő *-em* végű szó (pl. *terem, verem*) van, mint nem *-alom*-ra végződő *-om* végű szó (pl. *karom, majom*).

(9) *e-kereskedelem*[1] (1711) : *e-kereskedelmi* (1062)
világkereskedelem (798) : *világkereskedelmi* (441)
magánkereskedelem (49) : *magánkereskedelmi* (13)
terménykereskedelem (16) : *terménykereskedelmi* (4)
rabszolga-kereskedelem (110) : **rabszolga-kereskedelmi*

(A gyakorisági adatok a *Szószablya Gyakorisági Szótár*ból valók)

Ezek az ***-i*-s alakok nagyfokú szemantikai és viselkedésbeli autonómiára** tehettek szert (Bybee 2010) nagyobb gyakoriságuknak köszönhetően. Elkülönülésüket azonban zavarhatja, hogy az alapszó leggyakoribb hangkivetéses alakja, az „E.3 birtokos" (70,3% a *Google Gyakorisági Gyűjtés*ben, más *-m* végűeknél 30,4%, az összes hangkivető főnévnél 27,5%) nagyon hasonló alakú. A kontrasztra való törekvés a magyar morfofonológiában más esetekben szerepet kap (Rebrus és Trón 2003), így elképzelhető, hogy ezeknél az alakoknál a nagyfokú hasonlóság kerülése is kívánatos percepciós okokból. Ennek következtében a nem hangkivetéses *kereskedelme* alak helyett a *kereskedeleme* is választható opció lesz a kontraszt erősítése érdekében. A hangkivető sémát legkevésbé követő *-kereskedelem* végű szavak gyakoriságukban is meghaladják a legtöbb *-kereskedelem* végű szót. Az *e-kereskedelem* és a *kiskereskedelem* esetében látjuk azonban, hogy ezek gyakoriságában nincs különbség, míg viselkedésükben és *-i*-s alakjaik gyakoriságában van eltérés. Az E.3 birtokosuknál kevésbé hangkivető *-kereskedelem* végű szavak *-i* képzős alakjainak nagy gyakorisága nyelvhasználati okokra vezethető vissza[2]. A *-kereskedelem* végű szavak esetében az „E.3 birtokos" alak hangkivetési mértékének csökkenése nem feltétlenül gyorsuló változásnak, hanem

[1] Az újabb *e-kereskedelem* szó válogatásomban nem szerepelt, mert a *morphdb.hu* elsősorban régebbi forrásainak köszönhetően nem tartalmazta, de a *Szószablya Gyakorisági Szótár*ban ettől függetlenül szerepeltek az itt felhasznált adatai is.

[2] Így, ha Magyarország is gyarmatosító ország lett volna, akkor a *rabszolga-kereskedelmi*-nek is lenne létjogosultsága akár ilyen konstrukciókban, mint *rabszolga-kereskedelmi útvonal* (pl. *slave trade route* angolul), *rabszolga-kereskedelmi jutalék* (pl. *slave trade comission* angolul) stb. Ha ezek még mindig használatban lennének, akkor valószínűsíthetnénk a *rabszolga-kereskedeleme* alak használatát is.

inkább talán annak tudható be, hogy az interneten tömegesen jelennek meg kevésbé formális kereskedelmi, kereskedelemhez kapcsolódó szövegek, amelyek esetében ezeknél a szavaknál az informális változatok kapnak nagyobb súlyt. Hasonlóan még az „E.3 birtokos" alakjaiban kevésbé hangkivető az *űrtartalom* (elsősorban üzleti hirdetésekben), a *majom* és a *barom* (bővebben 5.4.3.2. alfejezet). Egyedi a *szerelem* „E.1 birtokos" alakjánál és a *lakodalom* többes számánál megfigyelhető alacsonyabb hangkivetési mérték, ami elképzelhetően informális szövegeknek tudható be. Minden más szónál a csoportban a tárgyesetű alak a legkevésbé hangkivető.

A gyorsan változó szavak többnyire szoros kapcsolatban vannak más gyorsan változó szavakkal. A csoporthatásokat esetükben is csak mérsékelten lehet megfigyelni, de egységesebben viselkednek, mint a toldalékonkénti hangkivetési mértékükben nagy szórást mutató szavak. A **tárgyesetű** (csökkenés átlaga: 6,2%, csökkenés mediánja: 2,9%) és az „**E.3 birtokos"** (csökkenés átlaga: 7,4%, csökkenés mediánja: 1,1%) alakok esetében figyelhető meg a **hangkivetés mértékének a leggyorsabb csökkenése**, ami a többi toldaléknál elhanyagolható mértékű. A gyors változás egyrészt magyarázható azzal, hogy az alakjuk alapján a változásra hajlamos nagyobb gyakoriságú szavak is elindultak a kiegyenlítődés útján, és ezt a korai szakaszt nagyobb lendület jellemzi. Azonban mindenképpen számolnunk kell azzal, hogy ezek a szavak a ritkább összetett szavak egy jelentős részével ellentétben a köznapi kommunikációban nagyobb szerepet kaphatnak. A látszólagos gyorsabb változásban jelentős tényező lehet, hogy a *Google Gyakoriségi Gyűjtés*ben informális szövegek nagyobb arányban szerepelnek. A gyűjtés jellegéből kifolyólag azonban nem lehet meghatározni, hogy a hangkivetés nélküli alakok arányának hirtelen megnövekedése a releváns toldalékoknál a vizsgált szavak tekintetében egy felgyorsult folyamatnak köszönhető-e, vagy az informális alakok nagyobb mértékű megjelenésének.

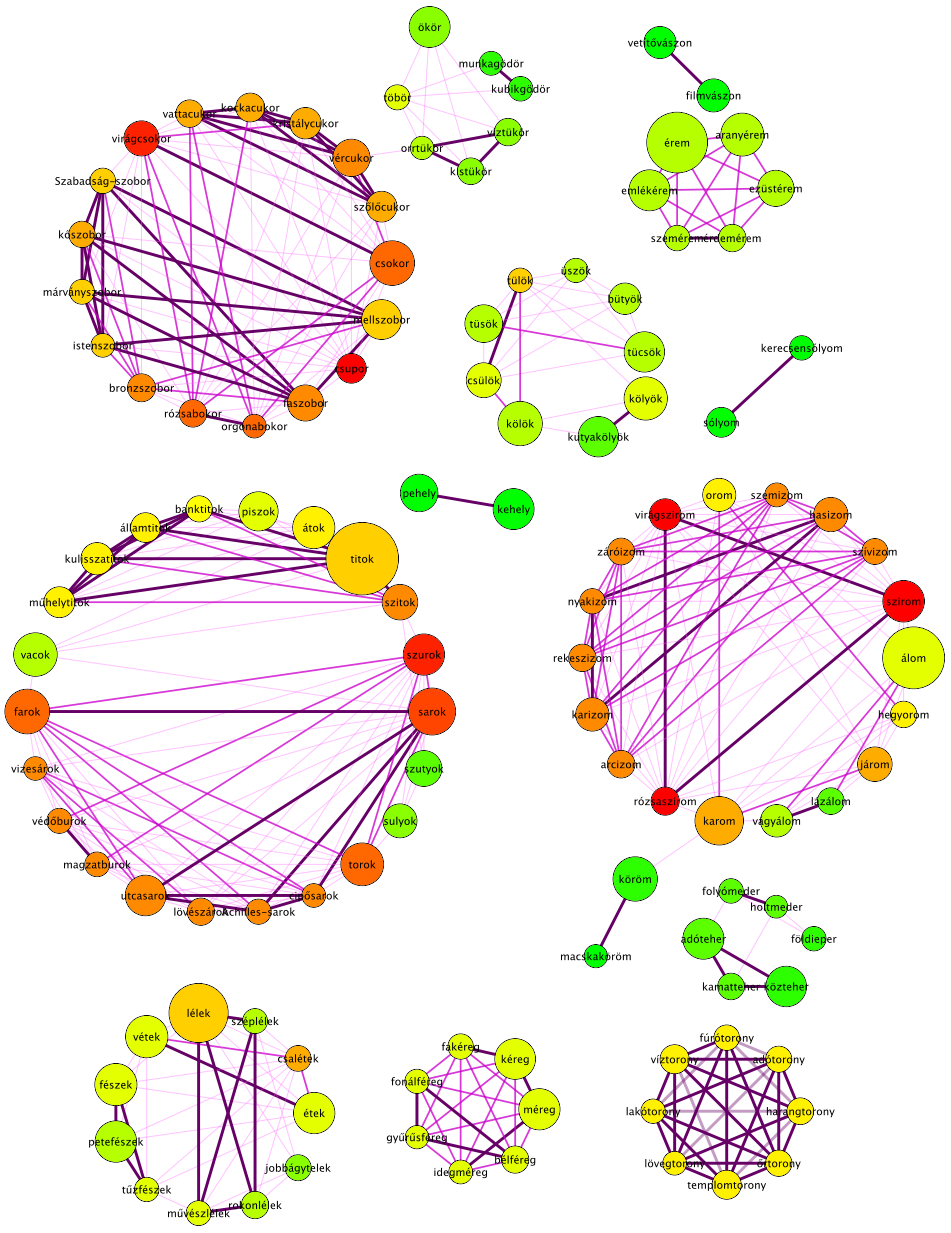
5.4.5. A változásnak ellenálló szavak

Azokat a szavakat tekintem a változással szemben ellenállóaknak, amelyek esetében stagnált vagy nőtt a hangkivetés mértéke. Ismét csak azokat a szavakat vizsgálom, amelyek legalább 100-szor előfordulnak releváns alakjaik tekintetében a *Szószablya Gyakoriségi Szótár*ban (424 szó). Ezek közül **stagnálónak** azokat a szavakat minősítem, amelyek esetében a hangkivetés mértéke se a *Szószablya*, se a *Google* esetében nem csökkent 99% alá. A vizsgálatra kiválasztott 424 szó 71%-a (318 szó) vehető eszerint stagnálónak. Ebből 117 *-alom*, 51 *-elem* és 24 *-erem* végű volt, amelyek együttesen a stagnálók 60,3%-át képviselik, míg az összes hangkivető főnévnek csupán 45,33%-át teszik ki az ilyen végű szavak. Ezek elemzésével nem foglalkozok a továbbiakban, mivel stabil hangkivető viselkedésük egyértelműen szoros kapcsolataiknak, egymáshoz közeli formájuknak és nagy gyakoriságú prototipikus szavaiknak köszönhető. Az 5.26. ábra a többi stagnáló szó viszonyait mutatja be. A három legnagyobb csoport tagjait szoros kapcsolatok kötik össze. Ez alól csak az *-álom, -rom* végűek a kivételek, de tudjuk, hogy ezek az ábráról kihagyott *-alom* végűekhez kapcsolódnak nagyon szorosan, amelyek biztosítják stabilitásukat, habár nem nyitótövek.

A **kisebb csoportokat** megvizsgálva azt tapasztaljuk, hogy olyan **jellegzetességekkel** bírnak, amelyek az *-alom, -elem* végekhez hasonlóan **egyedivé** teszik őket, a kapcsolatok pedig ezek közt is szorosak. A *sólyom* esetén az *-ó-o* magánhangzó-szekvencia 6,3-szor jellemzőbb a hangkivetőkre, mint a többi főnévre, és ugyanez elmondható a gyengén összefűződő *-ü-ö*[1] (6,47-szer jellemzőbb) és *-ö-ö*[2] (4,3-szor jellemzőbb) végszekvenciákat tartalmazó szavakról is. A *kehely, pehely, vetítővászon, filmvászon* következetes hangkivetésének okait már az 5.2.5. alfejezetben tárgyaltam.

[1] Ez az *ólom*-ot, *üröm*-öt nem mentette meg az ingadozástól, mert *VCVC* szerkezetük kevésbé tipikus. A *sólyom*-hoz hasonló szerkezetű *sulyom is* kevésbé hangkivető, mivel ritkább és marginálisabb helyzetű.

[2] A kevésbé tipikusnak mondható *-ö-ö* csoportnál a *bögöly, gödör* már az ingadozás útjára lépett.



5.26. ábra: Stagnáló nem *-alom, -elem, -erem* végű főnevek. Az ábra a 0,8-nál szorosabb hasonlósági viszonyokat mutatja.

A *lélek, érem, méreg* **nagy gyakoriságúak, prototípusként viselkednek** a hasonló szavak esetében[1]**,** és biztosítják csoportjuk stabilitását. A *-köröm* végűek látszólag önmagukban állnak, de a komplex tengelymérték segítségével felismerhető

[1] A *lélek*-nél ez a hatás csak a belőle képzett összetett szavakra szorítkozik. A *lélek* prototípusossága érvényesülhet hozzá közeli, de nem belőle képzett szavak esetében is úgy, hogy azok hangkivető viselkedését erősíti, de már nem képes bevonni őket a nagyon egyedi tőbelseji magánhangzó-rövidüléses főnevek csoportjába.

hasonlóságuk a *korom*-hoz és a *karom*-hoz. E három szó esetében az analógiás kiegyenlítődést megakadályozhatja a homonímiakerülés is (pl. *körömön: kör*+POSS.E.1+SUP vagy *köröm*+SUP). A legkevésbé stabil helyzetű stagnáló szavak (*földieper, folyómeder, holtmeder*, *jobbágytelek*) utótagja már kevésbé hangkivető, és a stagnáló szavak valamely hasonlósági struktúrájába se nagyon kapcsolódnak be, így hamarosan elindulhatnak a változás útján. Az *ökör, bütyök, üszök* és a *töbör* elszigeteltsége később szintén a hangkivetésük mértékének csökkenéséhez vezethet, különösen, hogy az *ökör* szintén előfordulhat a 'hülyéje' jelentésben *ökörje* alakban.

15 olyan szó van a 424-ből, amely a *Szószablya Gyakorisági Szótár*ban legalább 2,43%-ban nem hangkivető módon viselkedett a hangkivetést elváró toldalékok előtt, azaz az átlagosnál jobban ingadozott és a *Google Gyakorisági Gyűjtés*ben magasabb a hangkivetéses alakjainak az aránya ennél az értéknél. Ezek közül **10 szónál** a **visszalépés** mértéke 2,3% és 50% közötti: *berek, bögöly, fátyol, jászol, kebel, koboz, pöcök, tegez, vacak, veder.* Valószínű, hogy esetükben is stagnálásról van szó, hisz még most is átlagosan 28%-ban nem hangkivető módon viselkednek. Ezt erősíti meg, hogy hangkivetésük mértéke a tárgynál változatlan (54,6%), a szupresszívusznál pedig még csökkent is (69,9%-ról 60%-ra), azaz elsősorban a ritkább toldalékok esetében történt a látszólagos visszarendeződésük, annak köszönhetően, hogy több adatom van azok pontos viselkedésének a meghatározására is, amelyeknél a *Szószablya Gyakorisági Szótár* alapján túl alacsony hangkivetési mértéket állapítottam meg.

5 szónak feleannyi vagy még kevesebb nem hangkivető alakja volt arányait tekintve a *Google Gyakorisági Gyűjtés*ben, mint a *Szószablya Gyakoriság Szótár*ban: *fodor* (*Szószablya* 96,9%, *Google*: 99,7%), *korom* (*Szószablya*: 96,4%, *Google*: 99,1%), *bugyor* (*Szószablya*: 96,2%, *Google*: 98,8%), *kapor* (*Szószablya:* 93,1%, *Google*: 96,6%), *ködfátyol* (*Szószablya*: 10%, *Google*: 58,8%). A *korom, fodor, kapor* látszólagos visszaesését csak a mérések pontatlanságának köszönhetjük, mivel ezek esetében vagy tulajdonnévi alakkal, vagy ékezettelen írásmódnak betudható homonim alakkal való keveredés okozza ezeket az arányaiban nagy, de mértékében kicsi kilengéseket[1]. A *bugyor* esetében a hangkivetés megerősödése annak tulajdonítható, hogy a *Google Gyakoriság Gyűjtés*

[1] Ez igaz az előző csoportból a *berek*-re és a *veder*-re is.

jobban reprezentálja valós használatát (100-szor több alak, mint a *Szószablya Gyakoriségi Szótár*ban), amelyben a hangkivetéses változatát preferáló birtokos alakjaival fordul leginkább elő („E.3 birtokos több birtokkal" a releváns esetek 51%-ában). A *ködfátyol* esetében a jelentősnek tűnő visszarendeződés betudható annak, hogy ennél a viszonylag ritkább szónál sem reprezentálta a *Szószablya Gyakoriségi Szótár* megfelelően a használatát, mint a *bugyor* esetében. A *Google Gyakoriségi Gyűjtés*nél elsősorban a tárgyesetű és az „E.3 birtokos" alak hangkivetési mértéke nőtt meg, azonban jelenlegi 58%-os hangkivetési mértéke a releváns toldalékai vonatkozásában sem mondható magasnak, és elméletemmel összhangban alacsonyabb, mint a alapszaváé, a *fátyol*-é (80%). Ugyanakkor ennél a jelentős mértékű változásnál nem zárhatom ki, hogy a visszarendeződés valódi, de okai rejtve maradtak előlünk.

5.4.6. A *Szószablya Korpusz* és a *Google Gyakoriségi Gyűjtés* összehasonlítása alapján tett megfigyelések összegzése

Mielőtt ismertetem a hangkivető főnevek analógiás modellezése során elért eredményeimet, röviden összefoglalom azokat a legfontosabb felismeréseket, amelyeket a hangkivető főnevek változásának tanulmányozása során tettem:

- A vizsgált toldalékos alakokban az összes alak 99,42%-a volt hangkivető a *Szószablya Gyakoriségi Szótár*ban, míg 98,12% a *Google Gyakoriségi Gyűjtés*ben.
- A változásban a paradigmatikus cellák meglehetősen eltérően vesznek részt.
- Az egyes paradigmacellák kisebb, mások magasabb hangkivetési mértékkel rendelkeznek, amelyek statisztikailag igazolhatók.
- Egyedül csak az E.3 birtokosra utaló alak változása vagy az informális regiszterekben való erősebb jelenléte igazolható szignifikáns módon. Ezen kívül a szuperesszívusznak és a tárgyesetnek alacsonyabb a hangkivetési mértéke a többi vizsgált toldalékhoz viszonyítva.

- A kevésbé hangkivető szavak közt megnőtt a hangkivetőkre jellemző végű (*-m, -k, -g*) szavak aránya.
- A változás szempontjából egyedien viselkedő szavak viselkedését csak alapos, atomi szintű vizsgálattal lehet megérteni, ezek jellemzésére általános érvényű szabályok, sémák nem alkalmasak.
- A változásnak leginkább az *-alom, -elem* végű szavak állnak ellent magas gyakoriságuknak és egymáshoz való nagyfokú hasonlóságuknak köszönhetően. E két csoportból a gyakoribb *-alom* végűek teljesen távol maradnak még a változástól, míg az ezeknél ritkább *-elem* végűeknél már esetenként beszélhetünk a változásba való mérsékelt bevonódásról.

6. Hasonlósági hatások modellezése

6.1. A modellezés célja

Az analógiás megközelítéssel szemben felhozott **kifogások** többnyire arra irányulnak, hogy **az analógiás források kiválasztásának a módja esetleges, nem elég egyértelmű**. Az elmélet bírálói ilyenkor arra hivatkoznak, hogy az egyedi, kiragadott eseteket könnyedén lehet analógiára visszavezetni, de nagyobb mennyiségű adattal vagy jól megfogalmazható kritériumokkal szemben már kudarcot vallana ez az elemzési mód. Azonban az analógiás folyamatok, működés valós volta mellett marginális és nem marginális jelenségek esetében is újabb és újabb bizonyítékok látnak napvilágot (Skousen és mtsai 2002, Blevins és Blevins 2009a, Bybee 2010), amelyekhez e fejezet vizsgálataival én is szeretnék hozzájárulni.

A már bemutatott **algoritmusok segítségével tesztelem a szavak hasonlóságáról alkotott elképzeléseim helyességét** olyan feladatokon, amelyekkel a szabályalapú elméletek nehezen boldogulnának. Nyelvi tudásunk része, hogy új, vagy legalábbis a beszélő számára kevésbé ismert **szavakat kategorizálunk hasonlósági alapon, és besoroljuk**[1] **őket egy már ismert paradigmába**[2]. Ezt a képességet kívánom a hasonlósági mértékeken alapuló algoritmusaimmal megragadni. A következő tesztek során azt is bemutatom, hogy egy hasonlósági mérték segítségével modellezhetőek az ingadozási jelenségek. Rytting (2002) már megmutatta, hogy egy **hasonló nyelvi jelenség a törökben** (/k/~∅ váltakozás) **modellezhető** az AM (lásd 3.2. alfejezet) segítségével. Eredményei hasonlóak vagy egyes esetekben jobbak voltak, mint amit a szabályalapú megközelítés hozott. Vizsgálatában a szavak felszíni alakjára

[1] A hasonlóságon alapuló kategorizáció és osztályozás valószínűleg sokkal fontosabb nyelvi képességünk, mint megnyilatkozások helyességének megítélése.

[2] Elképzelhető, hogy a beszélő számára új szavak egyik általa ismert paradigmába sem illenek bele. Ilyenkor vagy egy ún. alapértelmezett inflexiós osztályba sorolódnak (pl. az arapes főnevek esetében, Aronoff 1994), vagy új paradigmát is létrehozhatunk nekik, ami azonban inkább csak elméleti lehetőség még több szokatlan új szó esetén is.

hagyatkozott, bár egyes esetekben azok eredetét (arab átvétel) is figyelembe vette, amely más, a szavak alakját leíró változókkal együtt segítette a szavak viselkedésének a megértését[1].

A következő négy alfejezetben szorosan összefüggő kísérleteket mutatok be a hangkivető főnevekkel. Az **első teszt** azt vizsgálja, hogy különféle algoritmusok megfelelő analógiás forrást választanak-e a hangkivető főnevek egy csoportjához **eltérő méretű szóminta alapján**. Módszereiben ez egy korábbi, a településnevek lokatívuszaival kapcsolatos vizsgálatomat követi (Rung 2008). **Második** tesztem középpontjában az áll, hogy ha a rendelkezésünkre álló szótári állomány jól reprezentálja egy magyar felnőtt mentális lexikonát, akkor mennyire jól tudnának az algoritmusok az **összes hangkivető szóhoz a teljes lexikont figyelembe véve analógiás forrást választani**. **Harmadik tesztemet** ugyanezen a szóanyagon hajtom végre, de ezúttal a **komplex jegymértéket más gépi tanuló algoritmusokkal vetem össze** tízszeres keresztellenőrzés segítségével. Ezúttal a szótári anyagban az összetételek határát is jelölöm, amire a rendszerek többletinformációként támaszkodhatnak. **Végezetül** a hasonlósági mértékek alapján a legközelebbi források helyett **prototípusokat választok ki** az egyes hangkivető főnevekhez egy olyan algoritmus segítségével, amely a vizsgálataim során megszerzett tapasztalatok egy részét összegzi működésében. A kiválasztott prototípusok segítségével a hangkivető szavak hangkivetési mértékében megfigyelhető különbségek okaira keresek magyarázatot.

[1] A szavak arab eredetének önálló jeggyel való reprezentálását támogatta, hogy a kommunikációban a török beszélők is támaszkodnak erre az információra, így nem etimológiai többletinformáció megjelenítéséről van csupán szó. Hasonlóan Kálmán és mtsai (2010) a beszélők számára is jól azonosítható német és francia jövevényszavak esetében megfigyelték, hogy azok a magánhangzó-harmóniában más szavaktól eltérő módon vesznek részt.

6.2. Analógiás forrás választása eltérő méretű szócsoportok alapján

Az elsőként ismertetésre kerülő tesztben azt vizsgáltam meg, hogy milyen pontossággal választanak az algoritmusok egy **főnévhez megfelelő szócsoportot**[1], azaz hangkivető főnévhez legközelebbi szónak hangkivető főnevet, nem hangkivető főnévhez pedig hasonlóan nem hangkivető főnevet választanak-e alaki hasonlóság alapján egy már meglévő szólistából. A besorolások helyessége alapján látható, hogy egy algoritmus mennyire jól ragadja meg azt a feltételezett nyelvi képességet, amely alapján szavak közti hasonlítások elvégzésére képesek vagyunk, és ezáltal az alakok létrehozásához megfelelő módon tudunk analógiás mintát választani.

A **magyar helységnevek lokatívuszaival elvégezett korábbi tesztem** során (Rung 2008) azt tapasztaltam, hogy az egyszerű jegymérték alkalmas szavaknak különféle szócsoportokba való besorolására. A tesztben a leggyakoribb 100-100 harmónia- és toldaléktípus szempontjából eltérő szó alapján meghatároztam (100 *-ban*, 100 *-ben*, 100 *-on* és 100 *-e/ön* véget elváró szó alapján), hogy a gyakoriságban őket követő 40 szó szuperesszívuszt vagy inesszívuszt, illetve annak elöl vagy hátulképzett magánhangzós változatát várja-e el. A legjobban teljesítő egyszerű jegymérték (bővebben 4.3. alfejezet) 87,5%-os pontossággal választotta ki a megfelelő szócsoportot meghatározó analógiás forrást a 400 szavas mintahalmazból. Ritkább alakok esetében már az anyanyelvi beszélők ítéletei is ingadoznak, így ez a 87,5%-os teljesítmény megközelíti az ő eredményeiket[2], a Levenshtein-algoritmus teljesítményét pedig messze meghaladja.

---

[1] A meglehetősen tág *szócsoport* elnevezést azért használom, mert a feladat esetében nem beszélhetünk szigorúan vett paradigmákról, hisz a hangkivetők esetén is elkülöníthetők további paradigmatípusok (*bokor-bokrot* típus, *farok-farkat* típus, *lélek-lelket* típus, *pocok-pocokja* típus és *teher-terhet* típus), míg a nem hangkivetőként jelzett szócsoport az összes egyéb paradigmát fedi a *v*-vel bővülőktől kezdve a teljesen szabályosan viselkedő főnevekig.

[2] 10 magyar anyanyelvű beszélőt kértem meg a 40 szó besorolására, akik közül egyedül egy ért el 92,5%-os, az algoritmus teljesítményét meghaladó eredményt. A jónak mondható 87,5%-os teljesítmény mellett az algoritmus tévesztései azonban eltértek a beszélők hibázásaitól és azok természetétől.

Az egyszerű jegymérték mellett egy **másik algoritmust is bevontam a tesztelésbe**, amely a hasonlóság számításában a Levenshtein-algoritmussal egyező módon nem súlyozza, hogy a hasonlóságok a szavak mely részeiben figyelhetők meg. Ez az számítási mód a Levenshtein-algoritmussal ellentétben azonban **szerepet ad a fonémák jegyeinek**, amelyek azonosak voltak az egyszerű jegymértéknél használtakkal. Ennek az algoritmusnak a megalkotása során cél volt, hogy a magánhangzó-harmóniát és a szótagszerkezetet hatékonyabban kezelje, mivel az egyszerű jegymérték alapján működő algoritmusnak a korábban megmutatkozó gyengéje az volt, hogy ezeknek nem adott megfelelő súlyt.

A **hasonlóságot** ez az algoritmus az alapján határozta meg, hogy az összehasonlított **két szó fonémáinak jegyeiből létrehozott mátrixokban hány közös részgráfot talált**. A mátrix sorait a különböző jegyek, míg oszlopait az egyes fonémák adták. Az irányított gráfokban[1] a mátrix celláiból vezetett irányított él minden olyan cellába, amely az arra következő oszlopban volt található. Így a *bab* és a *púp* tartalmazta a *CVC* a *CV*, *VC*, *CC*, *C* és *V*, illetve a zárhang–hátulképzett–zárhang, zárhang–hátul képzett stb. láncokat. A *CC* példából látható, hogy a gráfokban megszakításokat is megengedtem, hisz például a magánhangzó-harmónia esetében a releváns összetevők nem közvetlenül követik egymást. A megszakítások számára és a jegyek kombinációjára semmilyen megszorítást sem alkalmaztam, azért hogy amennyire lehetséges, előzetes elméleti feltevéstől mentes legyen a számítási mód.

jegyek	**b**	**a**	**b**
mgh	0	1	0
elölképzett	0	0	0
kerek	0	1	0
nyílt	0	1	0
hosszú	0	0	0
zöngés	1	0	1
mód	1	0	1
hely	1	0	1

6.1. ábra: Néhány lehetséges irányított részgráf, amely az összehasonlítás alapját képezheti.

[1] Annyi irányított gráfom volt, ahány jegyem.

Az **1211 hangkivető főnévből** egyes szám alanyesetű alakjuk gyakorisága alapján az **501. leggyakoribb szótól a 600. szóig választottam ki** azokat a szavakat, amelyeket az algoritmusoknak be kellett sorolnia a tesztben. A 100 hangkivető szóból 7 esetében a hangkivetés mértéke a hangkivetéssel együtt járó toldalékok esetében nem érte el a 90%-ot (*hatökör, ködfátyol, lombsátor, sulyok, tündérfátyol, szalmakazal, zsákvászon*), a *pityer*-nél pedig az analógiás kiegyenlítődés befejeződött vagy befejeződés közeli állapotban van. A hangkivető szavakhoz kontrollcsoportként véletlenszerűen kiválasztott, velük azonos gyakoriságú 100 nem hangkivető főnevet vettem. A kiválasztott szavak egyes szám alanyesetű alakjainak a *Szószablya Gyakorisági Szótár* 93 és 57 közti előfordulást adott meg, azaz ritka, de még használt és valamelyest ismert szavakról van szó, mint pl. *pagony, nyúlgát, samesz*[1].

Ezeket a hangkivető és nem hangkivető tesztszavakat négy eltérő méretű szólistához hasonlítottam. A szólistákban az 50, 100, 200 és 500 leggyakoribb hangkivető főnév, illetve az ezekkel megegyező vagy nagyobb gyakoriságú[2] nem hangkivető főnevek szerepeltek. A listák pontos méretét a 6.1. táblázat adja meg, amelyen látható, hogy a hangkivető főnevek aránya a listákban a szavak számával együtt nő, de nem változik olyan mértékben, hogy az a vizsgálat eredményére jelentős kihatással lehessen. A gyakorisági szempontok a hasonlításban csak mérsékelten kaptak szerepet, mert a ritkább szavakat az egyes szólisták méretétől függően kihagytam az összehasonlításból, de egy listán belül a nagyon gyakori és a kevésbé gyakori szó már egyforma súllyal bírt.

---

[1] A feladat itt nem az volt, hogy a hangkivetőket és a potenciális hangkivetőket szétválasszuk, hanem az, hogy a hangkivetőket szétválasszuk a bármilyen felépítésű nem hangkivető szavaktól. A kizárólagosan fonológiai kritériumokat figyelembe vevő algoritmusomtól nem várhatjuk el, hogy a *gyerek* és a *berek* szavak közt lényegi különbségeket vegyen észre, hisz felépítésüket tekintve mind a két szó tökéletesen megfelel a hangkivető szavakkal szemben támasztott ismert fonológiai kritériumoknak. A *gyerek* mégsem hangkivető, aminek a fonológián kívül kell az okait keresnünk, így a két szó hatékony kategorizálása csak egy olyan algoritmussal lenne megoldható, amely nem csak fonológiai kritériumokat érvényesít futtatása során.

[2] Ezek a *dolog*-nál is gyakoribb főnevek: *egész, ember, idő, magyar, nap, program, rendszer, szó, világ, év*. A *rendszer* és a *program* gyakori szavak, de kiugró előfordulási számuk részben a *Szószablya Korpusz* webes jellegének is betudható.

Vizsgálatomban **viszonyítási alapnak** a **Levenshtein-algoritmus** teljesítményét vettem. A cél az volt, hogy algoritmusaim ezúttal is meghaladják ennek eredményeit.

hangkivetők száma	szavak száma	hangkivető főnevek aránya
50	2828	1,70%
100	5468	1,80%
200	10315	1,90%
500	15333	3,20%
1211 (összes szó)	49675 (összes szó)	2,44%

6.1. táblázat: a szólistákban lévő szavak száma és a bennük lévő hangkivető szavak aránya

A **6.2. táblázat** mutatja, hogy a tesztszavaknak a **szólistákhoz való hasonlítása milyen eredményeket hozott**. A hangkivető főnevek véletlenszerű besorolása 1,7-3,2%-os eredményt adott volna. Láthatjuk, hogy ezt minden esetben sikerült a vizsgált algoritmusoknak meghaladniuk. A nem hangkivető főneveknél a találgatás jóval magasabb, 96,8-98,3%-ban helyes besorolási arányt hozna, hisz ezek a főnevek nagyobb arányban voltak képviselve a szólistákban, így véletlenszerű kiválasztásukra is nagyobb esély lett volna.

Szólisták	Levenshtein, hangkivető	Levenshtein, nem hangkivető	egyszerű jegymérték, hangkivető	egyszerű jegymérték, nem hangkivető	gráf alapú, hangkivető
50 hangkivető	39%	98%	51%	100%	7%
100 hangkivető	75%	93%	73%	97%	14%
200 hangkivető	64%	98%	84%	97%	
500 hangkivető	63%	100%	95%	98%	

6.2. táblázat: Az egyes algoritmusok eredménye a szavak besorolásában. A százalékok arra utalnak, hogy a 100 szóból hány százalékban választott az adott algoritmus az adott listából azonos típusú szót.

A **gráf alapú algoritmussal** a 200 és az 500 hangkivető alakot tartalmazó szólistákkal nem végeztem el az összehasonlításokat, mert az algoritmus jelenlegi megvalósítása ezt nem teszi lehetővé belátható időn belül. Kihagytam a táblázatból a

gráf alapú algoritmus esetében a nem hangkivető szavakkal való összehasonlítást is, mivel a hangkivetőkkel való összevetés során már megmutatkozott, hogy az algoritmus jelen formájában nem tud kielégítő eredményt hozni. Az algoritmus elsősorban a hasonló fonémajegyekből építkező szavakat választotta tekintet nélkül az egyes elemek közvetlen sorrendiségére, ugyanakkor az egyforma hosszúságnak túlzott jelentőséget tulajdonított. A Levenshtein-algoritmus gyengébb teljesítménye egyértelműen a már leírt hiányosságaira vezethető vissza (bővebben 4.3. alfejezet)

Az egyszerű jegymértéket használó algoritmus **a legnagyobb méretű mintával** a **hangkivetők esetében 95%-os**, a **nem hangkivetők esetében** pedig **98%-os** eredményt hozott. A legkisebb, 50 hangkivető főnevet tartalmazó szólista esetén azonban hibátlanul teljesített a nem hangkivető főnevek besorolásában amiből láthatjuk, hogy algoritmusom megfelelő hatékonysággal tud emberi beavatkozás nélkül szavakhoz analógiás forrást választani, ami megerősíti korábbi tapasztalataimat. Jó eredménye elsősorban annak tulajdonítható, hogy a szóalak jobb szélétől távolodva egyre kisebb súlyt ad a hasonló fonémáknak, és meglehetősen elfogadóan viselkedik az egy szekvencián belüli kisebb eltérésekkel szemben. A nem hangkivető szavak esetében az eredménnyel azonban nem lehetünk elégedettek, hisz ez a 98%-os, látszólag magas eredmény a találgatás szintjével azonos.

Mivel csak az **egyszerű jegymérték** alapján működő algoritmus hozott kielégítő eredményeket, a továbbiakban kizárólag ennek **működését elemzem**. Az algoritmus 5 esetben sorolt be rosszul hangkivető szavakat: *bugyor, csöbör, lombsátor, orrnyereg, pityer*. A *pityer* esetében nem beszélhetünk hibázásról, hisz ezt a besorolást a *Szószablya Gyakorisági Szótár* adatai is támogatják. Ha a szemügyre vesszük a *bugyor* (leghasonlóbb *hunyor*), *orrnyereg* (leghasonlóbb *hadsereg*) és *csöbör* (leghasonlóbb *csömör*) esetében a hozzájuk 10 leghasonlóbb szót, akkor azt figyelhetjük meg, hogy ezek közt már vannak hangkivető szavak:

(1) bugyor: bodor, szatyor, csupor
orrnyereg: idegméreg, kígyóméreg
csöbör: gödör, vödör, sírgödör, ökör

Azaz az algoritmus felfedezi a hangkivető főnevekhez a hasonlóságot, csak nem ad ezeknek megfelelő súlyt, mivel **az alaki hasonlóságon túl nem vesz figyelembe olyan tényezőket, amelyek ezeknek a szavaknak a hangkivető viselkedését erősítik**. A *lombsátor* és az *orrnyereg* esetében a hibázás egyik forrása, hogy az algoritmus nem rendelkezett azzal a beszélők által hozzáférhető információval, hogy ezek összetett szavak[1]. Az algoritmus egyedül a *lombsátor*hoz nem talált megfelelő hangkivető szót még a legközelebbi 10 közt sem, ami jól tükrözi, hogy a *lombsátor* kevésbé hangkivető, de az algoritmus ítélete túlzó. A *lombsátor*-t következetesen, de tévesen az *-átor* végű latin eredetű szavakhoz hasonlítja: *pankrátor, diktátor, organizátor* stb. A hibázás oka, hogy az *-átor* végű szavak, különösen a *pankrátor* jobban hasonlítanak hozzá szótagfelépítésükben, mint a *sátor*[2], és a *sátor s*-jétől való eltérés a széltől ilyen távoli pozícióban már nem okoz jelentős gondot a hasonlításban. Ebben az esetben a Rebrus és Törkenczy (2008) által alkalmazott megszakítatlan szekvenciaazonosságon való hasonlítás jobb eredményt hozott volna.

Az algoritmus a **100 nem hangkivető főnévhez való hasonlításban két hibát** követett el: *bikacsök* : *bütyök,* illetve *csucsor* : *csupor*. Az első esetben a korábbi tesztelések (Rung 2008) során is tapasztalt hibáját figyelhetjük meg az algoritmusnak, amely nem elég érzékeny a hangrendi harmóniára, hisz a második magánhangzó már elég távol van a szó végétől, így kis súlyt kapott, és az *a:ö* eltérése már kevésbé jelentős ebben az

[1] A 6.4. alfejezetben látni fogjuk, hogy az összetételi határok ismerete valóban javít a tanuló algoritmusok teljesítményén.

[2] A többi *-sátor* végű összetett szó ritkább a *lombsátornál,* így azok a mintaként szolgáló szólistákban nem szerepeltek.

esetben[1]. A sokkal megfelelőbb jelölt, a *lopótök* csak a 10. leghasonlóbb szónak kerül elő. A *lopótök* hátulra sorolódása annak köszönhető, hogy az egyszerű jegymérték az *a:ö* párnak ugyanakkora hasonlóságot tulajdonít, mint az *a:ó* párnak, ami ebben az esetben nem tűnik indokoltnak. Az ilyen esetek támogatják Lukács (2002) elgondolását, miszerint az egyes jegyek fontosságát is súlyozni kellene. Esetünkben ha az elöl-, illetve a hátulképzettség nagyobb súlyt kapott volna, akkor jobb eredményt kapnék, ami azonban más esetekben akár romláshoz is vezethetne. A *csucsor* felépítése alapján akár hangkivető főnév is lehetne, bár *-cs-r* végű hangkivető főnevünk nincs. Az egyszerű jegymérték ezúttal túl engedékeny volt, és holisztikus megközelítéséből kifolyólag figyelmen kívül hagyta a végében hasonlóbb, de nem hangkivető *vicsor* szót. Hibázásának súlyát azonban enyhíti, hogy az általa a *csucsor*-hoz legközelebb lévőnek vélt 10 szóból 7 nem is hangkivető.

6.3. Tesztelés a *morphdb.hu* főnévi szóanyagán

Hasonlósági algoritmusaim összefoglaló kiértékelését egy „**hagyj ki egyet" (leave-one-out) keresztellenőrzéssel** végeztem el, amely megközelítés az egyes analógiás modellek teljesítményének mérésében is bevettnek tekinthető (Daelemans 2002). A vizsgálatból kihagytam az előző alfejezetben bemutatott gráf alapú algoritmust, mivel jelenlegi formájában nem mutatott megfelelő eredményeket, ugyanakkor a tesztelésbe bevontam a **komplex jegymértéken, a természetes**

[1] Az ilyen esetekben feltehetjük a kérdést, hogy mennyire kell és lehet strukturált analógiák közt válogatni, amelyek valamilyenfajta általánosítást megfogalmaznak (lásd Albright 2009), hisz azt kizárólagosan tudjuk, hogy egy hangkivető szóban milyen magánhangzó-szekvencia fordulhat elő. Azaz ha kifejezetten a hangkivető szavak besorolására tanítottam volna meg az algoritmust, akkor jobban teljesített volna, viszont sérült volna általános jellege. Elképzelésem szerint ugyanez az algoritmus működőképes lenne akár szláv főnevek ragozási osztályba való sorolásakor is, de ha ilyen nyelvspecifikus jellemzőket beleépítenék, akkor már mindenképpen megszorítanám a lehetséges alkalmazások körét. Az ilyen típusú problémák kezelésére hoztuk létre a komplex tengelymértéket, ami jobban meg tudja ragadni a szerkezeti sajátságokat az algoritmus általános jellegének feladása nélkül.

osztályokon és a **komplex tengelymértéken** alapuló algoritmusokat (bővebben 4.3. alfejezet). A 7. fejezetben bemutatásra kerülő nyelvi tesztemből azt tűnt ki, hogy a szóvégi mássalhangzó kiugróan fontos szerepet tölt be a hasonlításban, azonban ettől balra haladva a pozíciók jelentősége már csak mérsékelt ütemben csökken tovább. Algoritmusaimban a pozíciók súlyának egyenletes, 1,8 alapú logaritmikus csökkentésén mégsem változtattam, mivel a teszt a szóvégtől távolodva folyamatosan csökkenő hasonlósági hatások létét megerősítette, így igazolta az algoritmusom működését leginkább meghatározó feltételezést, de az egyes pozíciók pontos súlyának meghatározásához nem nyújt elegendő információt.

A teszt során az **1211 hangkivető főnevet hasonlítottam egyenként a *morphdb.hu* főnévi anyagához** (49675 szó[1]), úgy, hogy kizárólag az a főnév nem volt az összes főnevet tartalmazó listában, amelyhez analógiás forrást kerestünk. Az egyes algoritmusokat az alapján értékeltem, hogy a kérdéses szóhoz hangkivető vagy nem hangkivető szót tartottak-e a leghasonlóbbnak. Amennyiben egy algoritmus hangkivetőt választott a hangkivetőhöz, azt helyes válasznak vettem. A hangkivetők altípusai közt (*bokor*: alap hangkivető paradigma, *sátor*: nyitó hangkivető paradigma, *lélek*: tőbelseji magánhangzó rövidülés, *pocok: -jA* E.3 birtokosnál, *teher:* hangátvetés) nem tettem különbséget. A találgatás adataim esetében 2,44%-os teljesítményt hozott volna[2]. Ezt a feladatot nem végeztettem el anyanyelvi beszélőkkel a feladat jellegéből adódóan, de feltételezésem szerint egyetlen beszélő sem érné el a 100%-os teljesítményt[3], azaz az ezt erősen közelítő eredményt már kielégítőnek tekinthetjük.

Elméleti vonatkozásaiban is érdekesebb, ha a találgatási szintnél magasabbra rakom a lécet, és viszonyítási alapnak (*baseline*) azt veszem, hogy a feladat megoldásában mennyire jól teljesítenének nyelvtani szabályok. **A hangkivető szavak besorolását 3 szabállyal tudjuk megadni**:

---

[1] A teljes anyagból eltávolítottam a többszörösen felvett szavakat, és a rosszul felvett főnevek közül a leggyakoribbakat, mint pl. *akarom, kíváncsi, la, pici* stb. A teljes szóanyag kézi tisztítása azonban már önálló és lehetőségeimet meghaladó lexikográfiai munka lett volna.

[2] Ez a hangkivetők aránya a teljes főnévi mintában.

[3] Vita tárgya lehet az, hogy a beszélők esetlegesen nem 100%-os teljesítménye annak lenne-e betudható, hogy az algoritmusoknál kisebb minta alapján választanának.

(2) .*+[alom]$_N$ -> N_{EP}[1]

(N_{EP} = hangkivető főnév)

pontosság: 99,6% (268/269)

fedés: 100% (268/268)

(egyetlen kivétel: *szlalom*)[2]

(3) .?+[elem]$_N$ -> N_{EP}[3]

pontosság: 100% (159/159)

fedés: 100% (159/159)

E két szabályhoz az „egyébként-elvre" *(elsewhere condition*) hagyatkozva kapcsolhatjuk a Rebrus és Törkenczy (2008) által meghatározott sémát, mint a legjobban fedő általános szabályt:

[1] A szavak elejét tekintve nem tudtam általánosan valamilyen morfológiai kategóriára hivatkozni, ezért reguláris kifejezésekkel adom meg őket: .* = 0 vagy több karakter, .? = legalább egy karakter.

[2] A szótárban nem szerepel, de tekinthetjük magyar szónak a *salom*-ot is. Ebben az esetben a pontosság 99,2% lenne.

[3] A morfémahatár nem lehet összetett szó határa, mert ebben az esetben a 'alkotórész, energiatároló' jelentésű *-elem* végű szavakkal is el kellene számolnia a szabálynak és így pontossága 83,7%-ra esne vissza.

(4) $.*+VC_{\alpha}(o/e/ö)C_{\beta}$[1] -> N_{EP}

pontosság: 18.37%[2] (773/4207)

fedés: 98,59% (773/784)

E szabályok alkalmazásával 99,1%-os fedést érnénk el, viszont csak 25,9%-os lenne a pontosságunk (F = 41,06, van Rijsbergen 1979), amellyel **szembeállítható a komplex jegymérték lényegesen magasabb 93,25%-os értéke** (F = 95,1). A legutolsó pontatlan szabályt csak úgy válthatnánk ki, hogy meglehetősen sok, összesen 129 tövet kellene felvennünk a szótárba hangkivetőként megjelölve. Ehhez azonban feltételeznünk kellene, hogy egy szó ugyanúgy viselkedik önállóan, mint összetételi tagként. Ahogyan az én vizsgálataimból is kiderül (bővebben 5.3.5. alfejezet), de máshonnan származó nyelvi ismereteink alapján is tudjuk, hogy ez nem igaz. Másrészt a 129 szó kivételesként való megjelölésével az ingadozásról továbbra se tudnánk számot adni. A szabályok mellett ismételten a Levenshtein-algoritmust vettem viszonyítási alapnak a tesztelésben, amelynek eredményeit a 6.3 táblázat mutatja meg.

[1] Összesen 64 viszonylag ritka alak van a szótárban, amelyek utolsó két mássalhangzója azonos.

[2] A pontosságot növelhetnénk újabb és még pontosabb részszabályok hozzáadásával, de ebben az esetben megközelítésünk már inkább Albrightéhoz (2009) hasonlítana, aki strukturált alsémák alapján végez analógiás hasonlítást, semmint a klasszikus, generatív szabályalapú elemzéshez. Ezeknek a sokkal specifikusabb szabályoknak a feltárása azonban túlmutatna elemzésem jelenlegi keretein.

	besorolási pontosság hangkivetőkre
Levenshtein	74,15% (74,43%)
egyszerű jegymérték	97,19% (97,48%)
komplex jegymérték	97,03% (97,56%)
természetes osztályok	97,03% (97,23%)
komplex tengelymérték	89,76% (90,36%)

6.3. táblázat: A hasonlósági algoritmusok teljesítménye a leave-one-out tesztben[1].

A zárójelben szereplő számok a 6.3. táblázatban arra utalnak, hogy milyen lenne az algoritmusok korrigált teljesítménye, ha csak azon szavak (1193 főnév) alapján értékelném azt ki, amelyek a *Google Gyakorisági Gyűjtés* szerint legalább 50%-os hangkivetési mértékkel rendelkeznek a hangkivetéssel együttjáró toldalékaik esetében. Az **egyszerű jegymérték** 34 rosszul besorolt szavából így 4-nél valójában nem tévedett az algoritmus: *ászok, pityer, tegez, üröm*. Ezek figyelembe vételével az eredménye: 97,48%[2]. A **komplex jegymérték** a nem hangkivetők közé további 3 már csak alig hangkivető módon viselkedő szót sorol: *cseber, jászol, veder* (az előbbiekkel együtt összesen 7-et). Ezek alapján teljesítménye: 97,56%. A **természetes osztályok** segítségével számított hasonlóság esetén 3 olyan szót (*ászok, pityer, tegez*) soroltunk be a nem hangkivetők közé, amelyek 50%-nál kisebb hangkivetési mértékkel rendelkeznek. Korrigálva teljesítménye: 97,23%. A **komplex tengelymérték** alapján működő változat esetén a már alig hangkivetők, amelyek nem hangkivető párt kaptak a következők: *ászok, cseber, jászol, pityer, tegez, üröm, vicikvacak.* Teljesítménye így: 90,36%. A Levenshtein-algoritmus 10 olyan szóhoz helyesen rendelt nem hangkivető párt, amelyek 50%-nál kevésbé hangkivetők, ezért módosított eredménye: 74,43%.

[1] Az itt megadott értékek a hangkivető főneveket beazonosító szabályoknál tárgyalt fedésnek feleltethetők meg.

[2] A számítás módja: 0,9748 = (1193-(34-4))/1193.

Közel hasonló teljesítményt hoztak azok az algoritmusok, amelyek a szóalak jobb szélétől távolodva egyre kisebb súlyt adnak a hasonló fonémáknak. Ha figyelembe vesszük az ingadozást a szavak besorolásában, akkor a komplex jegyeket alkalmazó változat teljesített a legjobban 97,56%-kal. Ez vélhetőleg közelíti az anyanyelvi beszélők elvárható teljesítményét is.

A hibák nagy része ezúttal is az algoritmus korábban tapasztalt fő gyengéjének[1] tudható be, hogy **„alábecsüli" a hangkivetők csoportjának megtartó erejét**, azaz a leghasonlóbb pár nem hangkivető volta még nem elegendő a paradigmaváltáshoz. Az algoritmus a legsúlyosabb hibákat annak a 7 szónak a kiválasztásakor követi el, amelyek több, mint 99%-ban a hangkivető mintát követik: *hamvveder, hatökör, járom, kérelem, kötelem, sérelem, szirom*. Ezek a szavak azonban a gyakoribb hangkivetők közé tartoznak, egyedül a *hatökör*[2] ritkább, mint a medián. Az algoritmus az *-alom* végűekkel, amelyeket a szabályok is jól ragadnak meg, nem vét hibákat.

A szóalak jobb szélétől távolodva a hasonló fonémáknak egyre kisebb súlyt adó algoritmusok egymástól kevésbé különböző eredményei jelentősen meghaladják a Levenshtein-algoritmusét, amelynek meglehetősen gyenge teljesítménye számottevően eltér az utána következő komplex tengelymértékétől is. Így ez a teszt is megmutatta, mint korábbi vizsgálataim (Rung 2008), hogy a **Levenshtein-algoritmussal számított hasonlóság nem alkalmas a hangtani hasonlóság modellálására, legalábbis az agglutinatív toldalékolás esetében nem**. Ebből kifolyólag legfeljebb csak viszonyítási alapnak használható, mint esetünkben. A **Levenshtein-algoritmus gyenge teljesítményének** nemcsak saját kutatásomra vonatkozóan van jelentősége. Ezek alapján **elbizonytalanodhatunk azon kutatások érvényességében** is, amelyek ezt a mértéket használták. Feltételezhetjük, hogy ezek következtetései részben tévesek, így kétségbe kell vonnunk Albright (2009) határozott kiállásának jogosságát is az egyedi hasonlóságok alapján számított analógiás modellezéssel szemben, mivel ezt a

---

[1] A nem tökéletes eredmény javítható lenne esetleg azzal is, ha nemcsak a legközelebbi mintát venném figyelembe, hanem a kiértékelésben súlyozva a kissé távolabbi szavak hatása is érvényesülhetne.

[2] Vélhetőleg a beszélt nyelvben az informális *hatökör* is gyakoribb, de lehet, hogy az internetet aktívabban használó fiatalabb nemzedéknél ez a szó már valóban ritka némileg avítt hangulata miatt.

Levenshtein-algoritmust használó GCM gyengébb teljesítménye alapján bírálja. Hasonlóan a másik viszonyítási alapnak (*baseline*) vett szabályalapú megközelítést is jelentősen túlszárnyalják algoritmusaim, hisz a szabályalapú megközelítés fedés és pontosság értékeit összegző F pontszáma csupán 41,06, míg a legjobban teljesítő komplex jegymérték esetében az F pontszám 95,1. A szabályalapú megközelítés csak annak a túlzott árnak a megfizetésével tudja jól fedni a hangkivető főneveket, hogy lényegesen több az általa megfogalmazott általánosítás alól a kivétel, mint a megragadott esetek száma.

Mivel egyes algoritmusaim az emberi teljesítményhez hasonló eredményeket hoztak, érdemes megvizsgálnunk, hogy az ilyen jellegű besorolás használható-e a **szótárbővítésben**, hisz a hangkivető tövek zárt csoportot alkotnak, amely néhány ellenpéldát leszámítva (pl. *rogyadalom, rohadalom*) nem növelhető tovább. A szótárbővítés azonban elsősorban nem új szavak besorolása egy szótári csoportba, hisz ezek a szavak digitális szótárunktól is függetlenül már hangkivetők vagy sem, hanem csak ezek hangkivető voltát „ismerjük fel" a szótárba felvételükkor. Igen sok szó van[1], amelyek még a digitális szótárakba nem kerültek be. Ezek esetében is hasznos lehet az automatikus, de a valós folyamatokat közelítő besorolási mód, amely nem alapulhat kizárólag azon, hogy egy új szó esetleg valamely a szótárban már meglévő szóból létrehozott összetett szó-e (pl. *lé* : *levet*, de *baracklé* : *baracklét*/*baracklevet*). Másrészt ha egy **szócsoportot zártnak** veszünk is, nem kizárt, hogy ha elég **nagy analógiás erővel** bír, akkor be tud vonzani új szavakat (pl. *motrok, bútrok*), amelyek akár a köznyelvi változatba is bekerülhetnek idővel (Bybee 2010). A nyelvünkbe frissen kerülő új szavaknál sem mindig egyértelmű, hogy miképp toldalékolódnának, de ha a legközelebbi analógiás forrásaikat nézzük meg, akkor várható viselkedésüket jobban tudjuk jósolni, mint ha szabályokat állítanánk fel erre. Így a *fájl* szónál elfogadható a

[1] Függetlenül attól, hogy az új szavak száma potenciálisan végtelen, a jelenlegi digitális szótárak szóanyaga az átlagos köznyelvi szövegek lefedéséhez elegendő, azonban erősen szaknyelvi, irodalmi, informális szöveg elemzésénél már gondok adódhatnak, nem is beszélve ékezetmentes szövegekről, amelyek elemzése analógiás alapon szintén könnyebb lehet.

fájlat tárgyas alak is (*Google* 417 találattal[1]; analógiás forrás: *váll)*, míg a *szkájp* (*skype*) esetében nem jó a *szkájpat*, csak a *szkájpot* (*Google:* 1500 találat), mert legközelebbi szomszédja a *skalp*. A nyitás előfordulásának összefüggését a szavak hangalaki felépítésével azonban itt csak valószínűsíthetjük (vö. Lukács 2002).

Az **algoritmusok hatásmechanizmusának** alaposabb **megértéséhez** egy jóval kisebb elemszámú rendhagyó csoportot is összehasonlítottam a teljes főnévi állománnyal. Ehhez összesen 16 *v*-vel bővülő szót (*tó, fű, cső, daru, szó, falu, kő, hó, ló, mű, nyű, tetű, odú, tő, hamu, lé*) választottam ki. Kihagytam a *jó, hő* és *mag* szavakat, mivel ezeknek a köznyelvben már nincs *v*-vel bővülő tárgyesete. Az egyszerű jegymértéken alapuló hasonlítás egyedül csak a *kő* : *tő, fű* : *ínyfű, mű* : *nyű, odú* : *faodú, nyű* : *mű* esetében választott jó párt, ami még így is meghaladja a találgatási szintet (0,3‰), hibázásai elsősorban annak tudhatók be, hogy a hosszúságkülönbséget nem veszi eléggé figyelembe, ami a jelen vizsgálatból kihagyott gráfalapú megközelítésnek azonban előnye volt. A komplex jegymérték és a természetes osztályokon alapuló hasonlóságszámítás egy további helyes párt is talált *hamu* : *fahamu*, míg a Levenshtein-algoritmus ismét a leggyengébben csak a *mű* : *nyű* párt tudta felismerni. Az ezektől eltérő módon számító komplex tengelymérték ugyancsak 6, de az előzőektől eltérő párt választott ki: *tó* : *tő, cső* : *kő, kő* : *cső, mű* : *nyű, nyű* : *mű, hó* : *hő.* Egyes esetekben (pl. *hó* : *hő*) sikeresebben ragadott meg releváns hasonlóságokat, ugyanakkor elmulasztotta a sokkal triviálisabb *fű* : *ínyfű, odú* : *faodú, hamu* : *fahamu* párok felismerését[2]. A tengely alapú algoritmus a két fonémából felépülő szavaknál alig veszi figyelembe a pozíciót (csak a *CV* tengely tekintetében), hisz a mássalhangzós és a magánhangzós jegyek külön tengelyen találhatók. Ennek köszönhető, hogy megfelelő súlyt tudott adni a mássalhangzóknak is, de az identikus végek beazonosítását ugyanezen tulajdonságából kifolyólag elmulasztotta. A példákból látható, hogy hosszú távon a két algoritmus

[1] A *fájlot* alakra több, összesen 4090 találat van, de ez nem meglepő, hisz az egyes számú főnévi tárgyesetű alakok általánosságban sokkal gyakrabban végződnek *-ot*-ra, mint *-at*-ra, azaz ez a minta lényegesen erősebb a szótól függetlenül.

[2] Az anyanyelvi beszélőket szemantikai szempontok is vezérlik abban, hogy felismerik-e az összetett szavakat, amire algoritmusaim nem képesek.

valamilyenféle ötvözete lenne megfelelő megoldás a modellezésre, hisz együttesen már 9 szót tudnának jól besorolni ebből a nagyon kicsi, egyedi csoportból is.

Az eddig bemutatott vizsgálatokat **egy további teszttel is kiegészítettem**, annak érdekében, hogy az algoritmusok teljesítményéről pontosabb képet kapjunk. Ebben a 100 leghasonlóbb szót választottam ki minden olyan hangkivető szóhoz, amelyek hangkivetési mértékéről rendelkeztem adatokkal, mivel feltételezésem szerint egy szó minél kevesebb hangkivető szóra hasonlít, annál kevésbé hangkivető módon viselkedik. Következő lépésben kiszámítottam, hogy mind a *Szószablya Gyakoriségi Szótár*ban, mind a *Google Gyakoriségi Gyűjtés*ben milyen mértékben jár együtt az egyes hangkivető szavak hangkivetésének mértéke azzal, hogy a hozzájuk legjobban hasonlító szavak közt hány hangkivető van.

Ezt a tesztet elvégeztem egy olyan változatban is, amikor a **hasonló szavak előfordulásait** a *Szószablya Gyakoriségi Szótár*beli egyes szám alanyesetű alakjuk **gyakoriságával**[1] **súlyoztam**, abból a feltételezésből kiindulva, hogy a nagyobb gyakoriságú hasonló szavak hatása az ingadozás kiváltásában, illetve a hangkivető mintában való megtartásban jelentősebb lehet. A tesztben figyelembe vett hangkivetési mértéket a leggyakoribb toldalékok[2] alapján számoltam, és csak azokra a szavakra végeztem el a vizsgálatot, amelyek ebben az esetben legalább 100 előfordulással bírtak a *Szószablya Korpusz*ban a vizsgálatba bevont toldalékaik esetében, így számításom 424 szóra vonatkozik. A hangkivető/nem hangkivető minősítésben az egyes hangkivetőnek vett szavak közt hangkivetési mértékük alapján nem differenciáltam, azaz a *sátor* és a *bizalom* töveket egyformán kezeltem. Hasonlóan jártam el a nem hangkivető szavak esetében is, azaz nem tettem különbséget a *motor* és a *csákány* között sem. Vélhetőleg ezek az ellentétes irányba mutató hatások kioltják egymást. A vizsgálat eredményeit a 6.4. táblázat foglalja össze.

---

[1] A nem hangkivető főnevekkel kapcsolatban nincs *Google* gyűjtésem, ezért számolok csak a *Szószablya Gyakoriségi Szótár* adataival.

[2] Ezek ugyanazok az adatok, amelyeket az 5.4. alfejezetben is használtam a *Google Gyakoriségi Gyűjtés* és a *Szószablya Gyakoriségi Szótár* összevetésében.

	Levenshtein	egyszerű jegymérték	komplex jegymérték	természetes osztályok	komplex tengelymérték
Szószablya	0,256***	0,302***	0,311***	0,302***	0,280***
Google	0,297***	0,317***	0,330***	0,315***	0,316***
Szószablya - gyakorisággal súlyozva	0,220***	0,278***	0,277***	0,266***	0,256***
Google - gyakorisággal súlyozva	0,256***	0,306***	0,313***	0,301***	0,286***

6.4. táblázat: Hasonlósági algoritmusok teljesítménye a hangkivetési mérték-leghasonlóbb hangkivető minták száma korrelációs vizsgálatokban

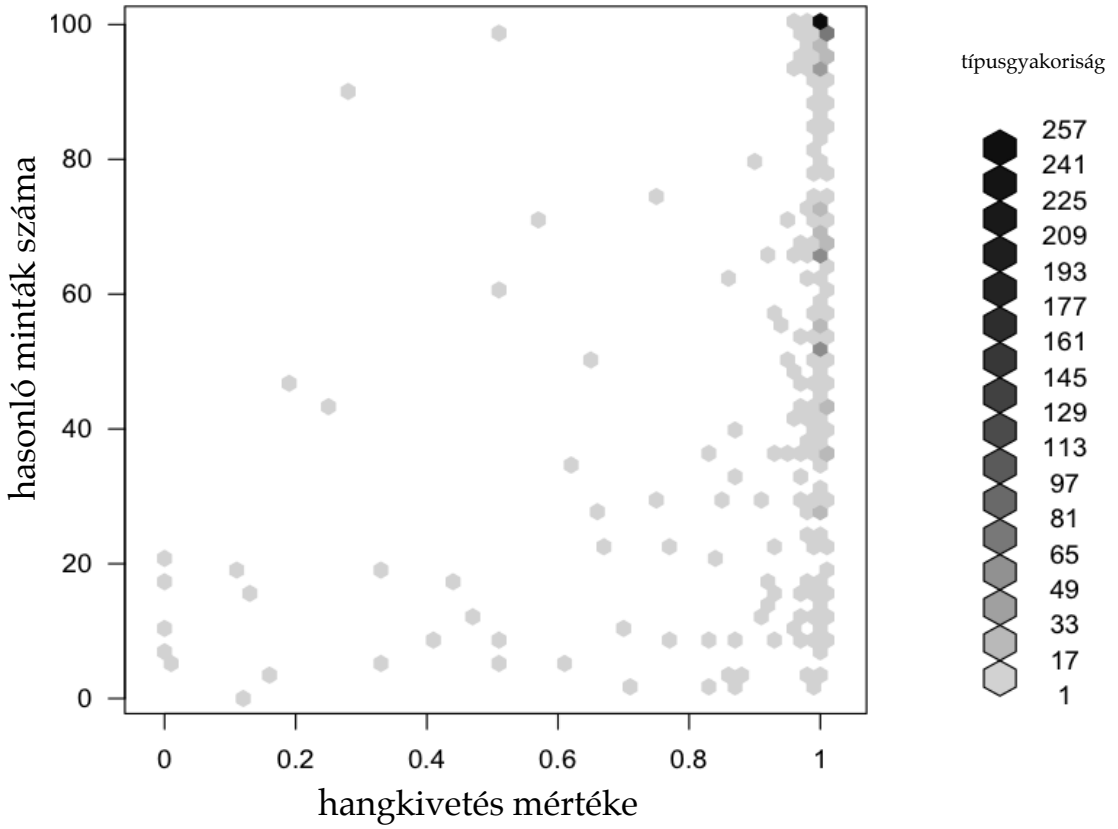
** = $p < 0,01$

*** = $p < 0,001$

A **hangkivetés mértékének a hasonló minták számával való összevetése gyenge és közepes**, de statisztikailag erősen szignifikáns **összefüggéseket** mutatott ki, így kijelenthetjük, hogy egy szó hangkivetési mértéke összefügg azzal, hogy hol helyezkedik el a hasonlósági térben. Minél több hangkivető szó hasonlít egy szóhoz, annál hangkivetőbb módon viselkedik. A *Google Gyakorisági Gyűjtés* esetében a magasabb korreláció azzal hozható összefüggésbe, hogy kisebb az adathiány, így a valós folyamatok jobban láthatók és modellezhetők a vélhetőleg nagyobb zaj ellenére is. Másrészt több szó kimozdult a nem ingadozó tartományból, amelyek hangkivetési mértékében a nagyobb variabilitás jobban modellezhető a hasonló minták számával, amelyek esetében a szórás igen magas. A *Google Gyakorisági Gyűjtés*ben a hangkivetés mértéke esetén a szórás 0,11, míg a hasonló pároknál a 0-1 közötti tartományba transzformálva is 0,31.

A legjobban teljesítő komplex jegymérték esetében a hangkivetés mértéke és a hasonló minták száma közti összefüggést a 6.2. ábra mutatja be. A 0,0-1,1 pontokat összekötő tengelytől leginkább eltérő szavak az *előterem*, *gyalom*, *ászok*, amelyek sajátos

viselkedését már korábban tárgyaltam. Az ábra alapján azt látjuk, ha egy szó erősen hasonlít a többi hangkivetőre, akkor stabilan hangkivető. Az **eredmények** megközelítőleg **hasonlóak a szomszédok gyakoriságának figyelembe vételével** is. A teljesítményjavulás elmaradása azonban nem igazolta ennek a komplexebb számításnak a létjogosultságát.



6.2. ábra: A hangkivetés mértéke (*Szószablya Gyakorisági Szótár*) és a hangkivető szomszédok száma közti összefüggések a komplex jegymérték alapján

Eddigi vizsgálataim alapján kijelenthetjük, hogy a s**zavak hasonlóságának megállapításában** a fonémák vagy jegyeik összevetésének mechanizmusa kisebb súllyal bír, mint a **szószintű összehasonlítás módja**, hisz a jelentősebb teljesítménykülönbségek csak akkor mutatkoztak meg a tesztekben, ha a szószintű összehasonlítás módját változtattam meg. Ebből kifolyólag az algoritmusok továbbfejlesztésének inkább erre kell irányulnia, semmint a fonéma-összehasonlítási

módszerek finomítására pszicholingvisztikai vagy fonetikai alapon[1]. A szó-összehasonlítás lényegesen feltáratlanabb területén végzett vizsgálatoktól a befektetett energia hatékonyabb megtérülését várhatjuk.

6.4. Összehasonlítás más tanuló algoritmusokkal

A „hagyj-ki-egyet" tesztben legjobban teljesítő **komplex jegymérték hatékonyságáról** úgy kaphatunk pontosabb képet, ha azt **összehasonlítjuk más elterjedt tanuló algoritmusokkal is**, amelyek a 3. fejezetben bemutatott analógiás megközelítésekhez hasonlóan jegyek alapján jósolják az ismeretlen elemek viselkedését, kategóriáit. Gyakran ezek hasonló eljárásokat is alkalmaznak a bemutatott analógiás megközelítésekhez (információ nyereség a TiMBL és a döntési fák esetében; az entrópia fogalmának beépítése a modellbe a TiMBL és a maximum entrópia modell esetében). Daelemans és van den Bosh (2005) ugyanakkor kiemelik, hogy ezeknek a rendszereknek az analógiás megközelítésekkel szemben nagyobb gondot jelent a valószínűségek számításához elegendő mennyiségű adat hiánya (*sparse data problem*), annak nehezen eldönthető volta, hogy mi a megfelelő adatforrás (*relevance problem*), és a tanítás menetének átláthatósága is gondot okozhat (*interpretation problem*). E megközelítéseknek szintén problémát jelent a zaj és a ritkán előforduló esetek megkülönböztetése (Daelemans és van den Bosch 2005: 23).

Az összehasonlításhoz két elterjedt és bevett megközelítési módot választottam ki (**maximum entrópia modell, döntési fák**), amelyeket röviden be is mutatok. Az összehasonlításba bevettem még a kurrens analógiás algoritmusok közül a TiMBL-t is. Mindegyik megközelítés közös tulajdonsága, hogy bemenetként jegyeket és a hozzájuk tartozó címkéket várnak el. Közös bennük az is, hogy megközelítésük habár nem minta

[1] Ez a megállapítás azért fontos, mert fonéma-összehasonlító módszereim valamennyire nyilvánvalóan pontatlanok, de az eredmények alapján még így is meglehetősen jól modellezik a hasonlósági hatásokat a saját szintjükön.

alapú (a TiMBL-t leszámítva), ugyanúgy szembe állíthatóak a generatív fonológia szemléletével, azaz egyik esetében sem olyan szabályokat állítanak fel automatikusan, amelyek az SPE (Chomsky és Halle 1968) vagy bármilyen későbbi fonológiai rendszer szabályainak megfeleltethetőek lennének. Megközelítéseik adatközpontúak, ami az analógiás megközelítésnek is elengedhetetlen része. Így ezeknek a nem analógiás algoritmusoknak a hatékonysága elsődlegesen nem a szabályalapú megközelítéseket támogatja (ezekkel csak annyiban rokonok, hogy a mintákat eldobják a tanulási szakasz után), hisz eredményességük csak azt igazolja, hogy a nyelvi produkcióban és megértésben a valószínűségek, gyakoriságok és eloszlások kulcsszerepet töltenek be, és ezek nem deklaratív szabályok alapján működnek (Daelemans és van den Bosch 2005: 19).

A **döntési fák** (Quinlan 1993) esetében egy új elem kategorizációjához egy gyökércsomópontból indulunk ki. A fa ebből „kihajtó" egyes csomópontjai döntési helyzeteknek felelnek meg, amelyekben a lehetséges jegyértékek mentén a fa elágazik egészen addig, amíg valamelyik döntés alapján meghatározható lesz a kimeneti kategória (a fa levelei). Az egyes jegyek fontosságát a TiMBL kapcsán bemutatott információnyereség (3.3. alfejezet) segítségével lehet meghatározni. A döntési fák numerikus és nem numerikus adatok alapján is tudnak döntéseket hozni. Számos változatuk van (J48, C4.5., véletlen erdők stb.), de ezek mindegyikénél a fákat rekurzív módon építik fel. A fa építése során egy ágat lezárunk, ha egy csomóponthoz tartozó elemek már homogének (pl. minden magánhangzóra végződő szó nem hangkivető), vagy nincsenek további jegyek, amelyek alapján újabb elágazás lehetséges lenne. A döntési fák építése során gyakran alkalmaznak metszést (*pruning*), amikor is eltávolítják a fa azon részeit, amelyek csak kis mértékben járulnak hozzá az osztályozás pontosságához. Ezzel a technikával a döntési fa túlillesztését (*overfitting*) akadályozhatjuk meg, aminek következtében nehezebb lenne a döntési fát új elemek viselkedésének a helyes meghatározására használni, mert túlzottan illeszkedne a tanító korpusz adataira.

A **maximum entrópia modellben** (Ratnaparkhi 1996) az egyes előfordulásokhoz megadott jegyek és címkék alapján a rendszer olyan valószínűségi modellt épít, amely

alapján ismeretlen címkéjű, de ismert jegyű új elemek kategóriáit is meg tudjuk határozni. Alkalmazása a nyelvi feldolgozás legkülönfélébb területein is bevett (Halácsy és mtsai 2005, Varga és Simon 2006, Oravecz és mtsai 2009, Recski 2010). A módszer alapfeltevése, hogy az ismert jegyek alapján azt a valószínűségi eloszlást választjuk, amely esetében az entrópia értéke a legnagyobb. Lényegében azt az eloszlást keressük, amely a legegyenletesebb adataink jegyeinek függvényében, mivel minél magasabb az entrópia, annál kisebbek az „egyenetlenségek" mintánkban. Ehhez megfelelő súlyokat kell tanulnunk iteratív arányos illesztéssel (*iterative scaling*, Darroch és Ratcliff 1972, Della Pietra és mtsai 1997).

A kiválasztott megközelítéseket a gépi tanulásban bevettnek mondható **tízszeres keresztellenőrzéssel** (*tenfold cross validation*) kapott eredményeik alapján vetettem össze (Weiss és Kulikowski 1991). A teszteléshez 10 egyenlő részre osztottam a főnevek teljes állományát. Az egyes algoritmusok minden tized szavairól a további 9 tized ismeretében jósolták a szavak jegyei vagy felépítésük alapján azok hangkivető vagy nem hangkivető voltát. A összetett szavak mindegyikében jeleztem az összetételek határát. Ez az információ elviekben a morphdb.hu-ban is megtalálható, de a szótár önmagában meglehetősen megbízhatatlan forrásnak bizonyult. A morphdb.hu 51219 főnévből összesen 21255 összetettet ismert, de ebből 20573 bizonyult valóban összetettnek, és ezekben is sokszor csak a szóhatárok egy része volt jelezve (pl. *folyó#számla#hitel* helyett *folyószámla#hitel*). A fennmaradó nem összetettnek jelzett szavakból 10455 bizonyult összetettnek, amelyek hozzáadásával összesen 31028 összetett szót kaptam. A teljes anyagból kihagytam azokat a főneveket, amelyekre nincs gyakorisági adatom egyes szám alanyesetű alakjukkal kapcsolatban, illetve olyanokat is, amelyek nem köznevek *kezicsókolom*, *Fernandez*, illetve többes számban szerepelnek a szótárban: *osztálykorlátok*, *főemlősök*, így összesen 49467 főnevem maradt.

A **döntési fa** teljesítményének kiértékelését a **Weka** szoftverrel végeztem el, a **maximum entrópia modellhez** pedig egy **python nyelvben írt maximum entrópia modellező eszközkészletet** (Maximum Entropy Modeling Toolkit for Python)

használtam. A tesztben a döntési fa és a TiMBL[1] összesen 322 jegyet használtak, amelyek a fonémapozíciók függvényében ugyanazon jegyek mentén kódolták a szavak tulajdonságait, mint amelyekre a komplex jegymérték is támaszkodik (4.3. alfejezet). A maximum entrópia modell esetében a felhasznált jegyek az individuális fonémák alapján létrehozott n-gramok (unigramok nélkül) voltak. Ezeket kiegészítettem olyan n-gramokkal, amelyek a magánhangzóknak csak a helyét adták meg tulajdonságaik nélkül, míg a mássalhangzókat hely, képzés és hangképző jegyeik reprezentálták. Ilyen n-gramokat mind a három mássalhangzós jegy alapján létrehoztam. Ezeknek homogénnek kellett lennie, azaz egy ilyen n-gramban csak egy mássalhangzós jegy alapján jellemeztem a szót, így a *bab* szó ilyen módon létrehozott 3-gramjai (szóhatárt nem jelölve) a placeBilab1V2placeBilab3, articulatorLabial1V2articulatorLabial3, mannerStop1V2mannerStop3[2] voltak. Ezekből az n-gramokból eltávolítottam a magánhangzó unigramokat. Ezeknek megfelelően az *ok* szó jegyeit a döntési fa és a TiMBL részére (7)[3], míg a maximum entrópia modellnek (8) alapján adtam meg:

(7) #,#,#,#,#,#,#,#,#,#,#,#,#,+,-,+,-,-,-,?,?,?,?,?,?,?,-,?,?,?,?,?,+,-,?,felp,zár,dor,hátsó,N

(8) N placeVelar1 articulatorDorsal1 mannerStop1 placeVelar1V2 articulatorDorsal1V2 mannerStop1V2 placeVelar1V2#3 articulatorDorsal1V2#3 mannerStop1V2#3 V2#3 #3 k1o2 k1o2#3 o2#3

---

[1] A TiMBL esetében a legközelebbi 3 szomszédot vettem figyelembe (k=3), illetve a módosított értékkülönbség mértéket is alkalmaztam (MVDM).

[2] Az itt szereplő mássalhangzós n-gram részek (pl. mannerStop1) nevei kódolják a hivatkozott fonémajegyet (pl. manner 'mód'), annak értéket (Stop 'zár') és annak helyét (1: utolsó fonéma). A magánhangzóknak csak a helyét kódoljuk: V2, azaz az utolsó előtti fonéma magánhangzó.

[3] A ? itt arra utal, hogy az adott jegy nem értelmezhető a lekódolni kíván fonéma esetében. Az egyes jegyértékek sorrendje fontos, mert ez alapján lehet beazonosítani, hogy melyik jegyhez tartoznak. A sor végén szereplő N azt adja meg, hogy az *ok* nem hangkivető. Ez a kategória érték a maximum entrópia esetében a sor elején szerepel. A # jel a szóhatár egyetlen fonémához sem hasonlító jellegét kódolja a fonémapozíciók függvényében rá vonatkozó egyes jegyeknél.

A **komplex jegymérték** esetében analógiás forrásnak az olyan leghasonlóbb szót vettem, amely a hasonlított szóhoz utolsó 4 fonémájában azonos *CV* szerkezetű volt, és magánhangzói azonos nyíltságúak voltak. Ha nem volt ilyen a 100 leghasonlóbb közt, akkor csak a *CV* megszorítást alkalmaztam, ha annak se felelt meg egy szó se (pl. *bifsztek*), akkor maradtam az eltérő szerkezetű, de leghasonlóbb szónál. A végrehajtott teszt eredményeit a 6.5. táblázat mutatja meg.

	F pontszám egyéb főnév	**F pontszám hangkivető**	**Tévesztés száma: egyéb főnév -> hangkivető**	**Tévesztés száma: hangkivető -> egyéb főnév**
Döntési fa (J48)	0,999	0,955	39	58
Maximum entrópia	0,999	0,977	27	23
Komplex jegymérték	0,999	0,979	31	14
TiMBL MVDM, k=3	0,999	0,955	60	37

6.5. táblázat: Az egyes algoritmusok eredményei a tízszeres keresztellenőrzésben

A 6.5. táblázat alapján láthatjuk, hogy mindegyik gépi tanuló algoritmus jól teljesített. **Eredményével a maximum entrópia modell** és az ezt egy árnyalattal meghaladó, de lényegében azonos szinten teljesítő **komplex jegymérték ugrik ki**. Jellemző, hogy az analógiás algoritmusok és a maximum entrópia modell hajlamosabbak hangkivetőnek kategorizálni nem hangkivető szavakat, míg a döntési fa pedig inkább a nem hangkivetőknek kedvez, azaz ez a szabályalapú rendszerekhez némileg közelebbi modell a szabályos viselkedést részesíti előnyben.

A **maximum entrópia modell és a komplex jegymérték hibázásaiban** azok kis száma ellenére igen **nagy átfedés** van. Összesen 23 közös szónál tévedtek, amelyből 10 esetben hangkivetőt vettek átlagos főnévnek (*berek, bodor, bugyor, iker, koboz, lator, pityer, pucor, takony, tegez, nyirok*). Ezek nagyrészt ingadozó szavak, azaz az algoritmusok nem véletlenül ismerték fel hasonlóságukat az átlagos főnévi tövekhez. Egyik sem osztja a hangkivetők általános sémáján túl azoknak valamelyik karakterisztikusabb jellemzőjét (*-alom, -elem* vagy *-ök* vég, bővebben 7.3. alfejezet). 13 esetben az algoritmusok nem hangkivető szavakat hangkivetőnek soroltak. Itt a hibázások már súlyosabbak. Egyik

algoritmus sem ismerte fel, hogy nincsenek nem *-alom, -elem* végű, kettőnél több szótagos nem összetett hangkivető szavak. Ezért vélhették az *alakor*, *betlehem*, *gennygyülem* szavakról, hogy azok hangkivetők. Mint a 7. fejezetben látni fogjuk, a magánhangzó-mintázat is meglehetősen jellemző a hangkivető tövekre. Időnként ebben is vétenek az algoritmusok, így gondolhatják azt a *balek*, *medok* szavakról, hogy azok hangkivetők. A *túzok* esetében pedig nem ismerték fel azt, hogy az utolsó két mássalhangzó egymás mellé kerülve nem vehet részt zöngésségi hasonulásban. Több esetben azonban a csak fonológiai kritériumok alapján való döntés eredménye nem kérdőjelezhető meg. A *birok*, *lórom*, *bókony*, *odor*, *perem*, *török* szavak felépítésük szerint lehetnének hangkivetők is, hisz elírásokban találkozhatunk *birkot, permei* stb. alakokkal. Különösen a *török* esetében nem beszélhetünk komoly tévesztésről, hisz nem számoltunk annak relatíve magas gyakoriságával (59266) sem, ami meggátolhatta, hogy a hangkivető paradigmába sorolódjon. Az itt bemutatott hibákból több azokban a tévesztésekben is előfordul, amelyek csak az egyik algoritmushoz köthetőek.

A **maximum entrópia modell** említésre méltó **egyedi hibája**, hogy az egyenletesen csökkenő súlyozás hiánya miatt hangkivetőnek kategorizálta a *kazah* (vélhetőleg a *kazal* hatása) szót. Több hibát vétett a kizárólag /ö/, /ü/ magánhangzókat tartalmazó szavakkal is (*öböl, göböj, pöcök, tengeröböl, üröm*), illetve egy esetben egy egy szótagos szót is hangkivetőnek vett (*ok*). A **komplex jegymértéknek** ugyanakkor **gondot jelentettek a hosszabb nem összetett szavak** (pl. *polinom: cirom*, *dezodor: sodor* stb.), illetve azok a szavak, amelyek utolsó magánhangzójukban hosszúak: *metronóm: házorom, kozsók: mocsok*. Ezek a hibák további megszorításokkal kezelhetőek lennének, de akkor algoritmusom általános jellegét kellene feladnom, ami azonban céljaimmal nincs összhangban.

Összefoglalva az látható, hogy mind a két megközelítésnél **hiányzik a magasabb szintű sémák felismerése**, amelyeket a 6.3. alfejezetben bemutatott szabályok fednek, igaz rendkívül pontatlanul. Ez a hiányosság egy holisztikusabb, a szerkezetre jobban koncentráló, de az egyedi jellemzőket is figyelembe vevő új algoritmussal lenne orvosolható, amelynek lehetséges működéséről a 8.3. alfejezetben szólok röviden. Ugyanakkor eredményeim arról is árulkodnak, hogy a sémák, általános szabályok

szerepe a korábbi nyelvtanokban túlbecsült, mivel a komplex jegymérték esetében a szavak 99,9%-ának kategóriája felismerhető mindenféle absztrakt, magasabb szintű összefüggések ismerete vagy az azokra való hivatkozás nélkül is.

6.5. Prototípus-tesztek

A 4.2 alfejezetben kifejtettem, hogy az analógiák felismerésében a leghasonlóbb mintán, mintákon túl a prototipikusan viselkedő szavaknak is nagy jelentőségük van, mert központi szerepet töltenek be a hozzájuk hasonló kisebb gyakoriságú szavak viselkedésének a meghatározásában. Ezeket a szavakat tulajdonságaik alapján algoritmikusan is kiválaszthatjuk. Elméleti megfontolásaim és a korábbi elemzések tapasztalatai alapján a prototípusok kiválasztására egy olyan algoritmust hoztunk létre[1], amely egy adott hasonlósági mátrix[2] segítségével a következő szempontokat veszi figyelembe:

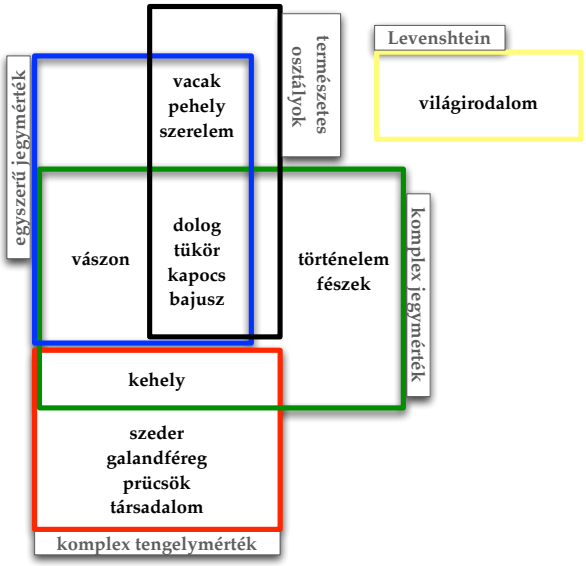
- Az egyes prototípusok kevéssé hasonlíthatnak egymásra.
- A prototípusra minél több nem prototipikus elemnek kell hasonlítania.
- Az egyes prototípusok kiválasztásában számít gyakoriságuk, hogy a ritka szavaknak csak különleges körülmények közt legyen esélyük arra, hogy prototípusnak kiválasszuk őket.

A prototípuskiválasztásban többféle küszöbértéket is megadhatunk, amelynek növelésével algoritmusunk egyre szigorúbban alkalmazza a hasonlósági szempontokat, miszerint a prototípushoz sokan hasonlítanak, de az más prototípusokra nem hasonlít.

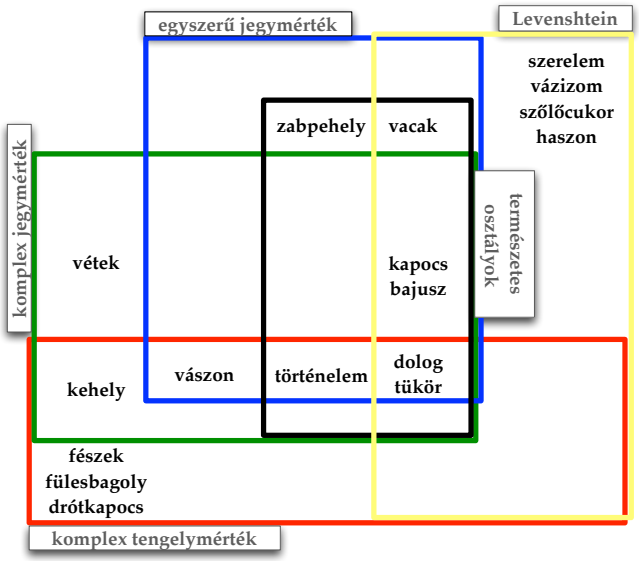
[1] A tesztelésre kerülő prototípuskiválasztó algoritmust közreműködésemmel Kálmán László hozta létre.

[2] Egy ilyen hasonlósági mátrix tartalmazza egy szóhalmaz minden egyes szavának hasonlósági értékét a szóhalmaz összes eleméhez (beleértve önmagát is) viszonyítva.

Két eltérő küszöbérték[1] mellett kiválasztott prototipikus szavainkat a 6.3, 6.4. ábrák mutatják meg.



6.3. ábra: 0,9-es[2] küszöbérték mellett kiválasztott prototípusok[3]



6.4. ábra: 0,5-ös küszöbérték mellett kiválasztott prototípusok

---

[1] Más küszöbértékekkel is elvégeztem a tesztelést, de azok vagy nagyon hasonló vagy gyengébb teljesítményt hoztak, mint amelyekkel a későbbiekben foglalkozok.

[2] A teszt ismertetése szempontjából nem fontos, hogy a küszöbértékek pontosan hogyan járulnak hozzá az algoritmus működéséhez. Az ábrák értelmezéséhez csak annyit kell tudnunk, hogy minél magasabb ez az érték, az algoritmus annál szigorúbban alkalmazza a kiválasztásban a hasonlósági kritériumainkat.

[3] Az ábrákon Venn-diagramokat láthatunk az Edwards-féle módosításban, ami lehetővé teszi öt halmaz elemeinek is az összehasonlítását. A halmazok megjelenítését tartalmuk függvényében átalakítottam a könnyebb áttekinthetőség érdekében.

Mielőtt áttekinteném, hogy az **egyes prototípusok** mennyire jól modellezték a hangkivető szavak hangkivetésének mértékét, érdemes őket **szemügyre venni**. Minden mérték esetében jellemző a **gyakori alakok preferenciája**. Ez legszembetűnőbben a *dolog* mint prototípus választásában jelenik meg, mivel az összes hangkivető előfordulás mintegy 16,1%-át teszi ki (348 ezer egyes szám alanyesetű előfordulás a *Szószablya Gyakorisági Szótár*ban), és 2,42-szer gyakoribb, mint az őt közvetlenül követő *társadalom*. A választásokban további nagyon gyakori szavak is szerepelnek még: *szerelem* (68 ezer), *társadalom* (144 ezer), *történelem* (68 ezer). A *dolog*-gal együtt ezek már az összes hangkivető főnév alanyesetű előfordulásainak a 29,1%-át fedik le. E kiugróan gyakori elemeken túl azonban a prototípusválasztó algoritmus inkább a hasonlósági szempontokat veszi figyelembe, hisz a következő leggyakoribb szó, a *tükör* (29., 21 ezer) már jóval elmarad ezek mögött. Az összes mértéken alapuló választásnál megfigyelhető, hogy habár a gyakori *-alom, -elem* végűek alkotják a legszámosabb alcsoportját a hangkivető szavaknak, mégis ezek vannak leginkább alulreprezentálva a prototípusok tekintetében. Általában az egyes prototípuscsoportokban csak *-elem* végű prototípus jelenik meg, ami egyaránt jól lefedi az *-alom* végűeket és a többi *-e-e* végű szót is.

Az prototípusválasztó algoritmus azonban **kevésbé gyakori szavakat** is választ, ha azok a **hasonlósági kritériumoknak jobban megfelelnek**. Ezek gyakran összetett szavak, hisz hosszúságuk alapján jobban reprezentálják a zömükben összetett hangkivető szavakat, mint a példánygyakoriságban gyakoribb, de típusgyakoriságban ritkább alapszavak. Ilyen szavak a *zabpehely* (egyszerű jegymérték, természetes osztályok), *szőlőcukor, vázizom* (Levenshtein-algoritmus), *drótkapocs, galandféreg, fülesbagoly,* (komplex tengelymérték). Kisebb, de jól elkülönülő szócsoportok is több esetben kapnak önálló prototípust: *vacak, bajusz* (utolsó magánhangzó nem középső nyelvállású), *vászon* (*-(á/ó)CVC* végűek), *zabpehely, kehely* (hangátvetés), *vázizom, pityer* (*-iCVC* végűek).

Az egyes hasonlósági mértékek és a két eltérő küszöbérték mentén kiválasztott prototípusokat az olyan hangkivető főnévekkel hasonlítottam össze, amelyek legfeljebb 99,99%-ban mutattak hangkivető viselkedést (282 szó). Vizsgálatomból azért zártam ki a

100%-ban hangkivető főneveket, mert ezek esetében legfeljebb csak a kiugróan gyakoriaknál tudhatjuk, hogy az ingadozás hiányának oka következetes viselkedésük, és a 100%-ban hangkivető viselkedés nem adathiánynak tudható be. A prototípusokhoz az egyes szavakat mindig olyan mérték alapján hasonlítottam, amilyen mértéket a prototípus kiválasztásában is alkalmaztam. Miután minden a vizsgálatra kiválasztott hangkivető főnevet minden prototípuscsoporttal (2 x 5 db) összehasonlítottam, megvizsgáltam, hogy az egyes szavak **hangkivetési mértéke** mind a *Szószablya Gyakorisági Szótár*ban, mind a *Google Gyakorisági Gyűjtés*ben mennyire **korrelál a hozzá legközelebbi prototípushoz való hasonlóságával**. Feltételezésem szerint egy szó minél jobban hasonlít a hozzá leghasonlóbb prototípushoz, annál nagyobb a hangkivetési mértéke is. Az együttjárások számítása során a prototípushoz való hasonlósági értéket súlyoztam a hasonlítandó hangkivető főnév releváns toldalékos alakjai alapján meghatározott gyakoriságának 8. gyökével (pl. *dolog* esetében 5,23, a *sátor*-nál 3,34), mivel a gyakoribb főneveknél magasabb hangkivetési mértéket várok, de nem kívántam ennek az értéknek túlzott súlyt sem adni. A 6.3, 6.4. ábrákon bemutatott prototípusokon túl a szavakat hasonlítottam a *Szószablya Gyakorisági Szótár*ban *az* egyes szám alanyesete alapján 50 leggyakoribb hangkivető főnévhez is, mint olyan prototípusokhoz, amelyeket kizárólag gyakoriságuk alapján választottunk ki a hasonlósági szempontok figyelmen kívül hagyásával. Gyakorisági prototípusnak azért választottam ki viszonylag sok szót, mert a 10 leggyakoribb hangkivető főnévből 8 *-alom*/*-elem* végű volt, így ennél több szóra[1] volt szükségünk ahhoz, hogy ne csak az *-alom*/*-elem* csoporthoz való hasonlóságot mérjük. A prototípusok számának növelése nem jár szükségszerűen együtt a korreláció mértékének növekedésével, hisz ha az összes hasonlítandó szót felvennénk prototípusnak, akkor az önmagukhoz való hasonlóságuk 1 lenne, aminek következtében egyáltalán nem tudnánk érdemleges együttjárásokat megfigyelni a változó hangkivetési mértékek és a konstans 1-es értékek közt.

[1] 50 szó mellett azért döntöttem, mert az 5.3.2. alfejezetben ugyanennyi hasonlósági csoportot határoztam meg.

	Levenshtein	**egyszerű jegymérték**	**komplex jegymérték**	**természetes osztályok**	**komplex tengelymérték**
Szószablya 0,5	0,241***	0,352***	0,364***	0,371***	0,419***
Google 0,5	0,163**	0,288***	0,298**	0,297***	0,285***
Szószablya 0,9	0,248***	0,352***	0,362***	0,370***	0,409***
Google 0,9	0,231***	0,288***	0,298**	0,297***	0,376***
Szószablya gyakori szavak	0,346***	0,458***	0,461***	0,455***	0,423***
Google gyakori szavak	0,310***	0,422***	0,428***	0,421***	0,380***

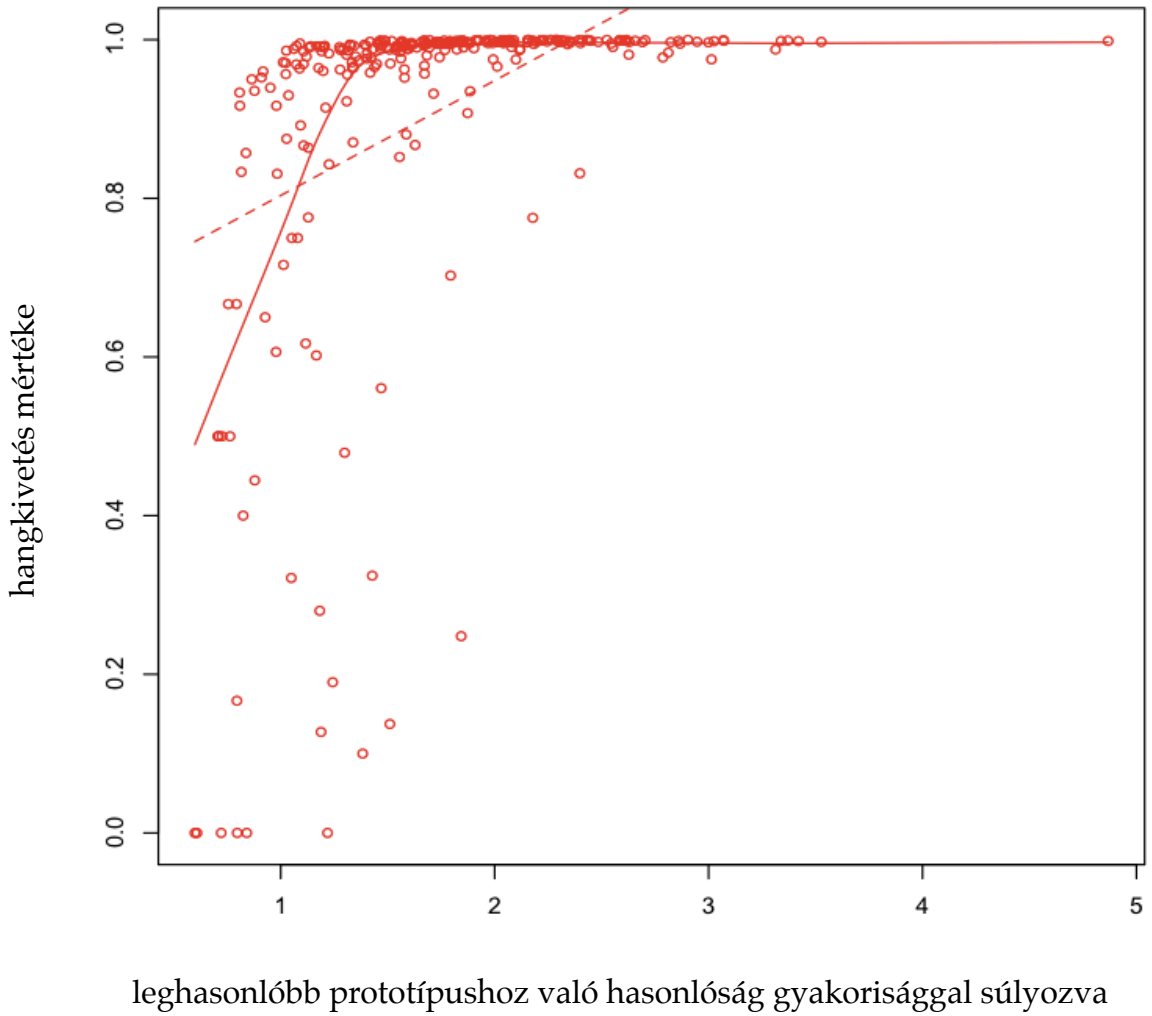
6.6. táblázat: A hangkivetési mérték és a prototípushoz való hasonlóság együttjárásának mértéke a korpusz és a felhasznált prototípusok függvényében. A *Google* és a *Szószablya* mögötti számok az alkalmazott küszöbértékre utalnak.

** = $p < 0,01$

*** = $p < 0,001$

A 6.6. táblázat alapján láthatjuk, hogy *Google Gyakorisági Gyűjtés* alapján számított hangkivetési mértékek és az algoritmusok segítségével mért legközelebbi prototípushoz való hasonlósági értékek kevésbé szoros, de szignifikáns együttjárásban vannak azokban az esetekben, amikor prototípusainkat gépileg választottuk ki. A leggyengébb teljesítményt ismét a Levenshtein-algoritmus hozta. Egyedül a **komplex tengelymérték** segítségével kiválasztott prototípusoknál figyelhetünk meg **közepes erősségű korrelációt**. A *Szószablya Gyakorisági Szótár* esetében azonban – a Levenshtein-algoritmust leszámítva – már az összes legközelebbi prototípushoz való hasonlóság közepesen korrelál a szavak hangkivetési mértékével. A komplex tengelymérték ebben az esetben is a legmagasabb együttjárást mutatja. A hangkivetés mértékét legjobban megragadó prototípusaink: a *dolog, történelem, tükör, vászon, fészek, kehely, fülesbagoly, drótkapocs.* Ez a néhány szó viszonylag jól fedi a lehetséges végmintázatokat és záró magánhangzó-szekvenciákat is, amelyek a viselkedés szempontjából a legfontosabbak lehetnek. A *fülesbagoly* és a *drótkapocs* a nagyszámú összetett szót, a *kehely* egy speciális mintát, a *vászon* pedig a mérsékelten hangkivető szavak csoportját képviseli.

Mindösszesen ennek a **8 szónak** az alapján szabályoknál hatékonyabb és könnyebb módon **274 másik szó viselkedését** tudjuk viszonylagos megbízhatósággal **jellemezni**. Az eredetileg csak viszonyítási alapnak szánt 50 leggyakoribb szóhoz való hasonlítás alapján azonban láthatjuk, hogy a gyakoriságnak van a legkiugróbb szerepe a szavak viszonyrendszerében. Ha csak a számunkra fontos szavak 20%-ához van gyors hozzáférésünk, már akkor egészen jól tudjuk leírni a maradék 80% viselkedését. Ha szavainkat a komplex jegymérték alapján azonosított hasonlósági csoportok leggyakoribb szavaihoz (5.3.2. alfejezet) hasonlítjuk a komplex jegymértékkel, akkor a *Szószablya Gyakoriság Szótár* esetében ismét közepesen erős korrelációt tudunk kimutatni ($r(280) = 0{,}4$, $t = 7{,}31$, $p < 0{,}001$). Azaz, ha a gyakoriságot lokálisan értelmezzük egy adott csoporton belül, akkor is képesek vagyunk az egyes szavak hangkivetési mértékével kapcsolatban együttjárásokat megfigyelni. Ha a legsikeresebb algoritmus által kiválasztott prototípusainkból (komplex jegymérték, 0,5-ös küszöbérték) és a leggyakoribb szavakból alkotott csoporthoz hasonlítjuk hangkivető szavainkat, akkor némileg még szorosabb együttjárást ($r(280) = 0{,}485$, $t = 9{,}27$, $p < 0{,}001$) figyelhetünk meg a hasonlóság-értékek és a hangkivetési mértékek közt. Ebből arra következtethetünk, hogy ha a prototípus kiválasztásában alkalmazott szempontjainkat még jobban optimalizálnánk, akkor a hangkivetési mértéket vagy akár bármilyen más viselkedési mutatót jobban tudnánk megragadni.



6.5. ábra: Komplex tengelymérték leghasonlóbb prototípusaihoz való hasonlóság és a hangkivetési mérték összefüggése a *Szószablya Korpusz*ban.

7. Az analógiás források kiválasztásában szerepet játszó tényezők mérése *CVCVC* szerkezetű álszavakkal

7.1. A vizsgálat előzményei és célja

A **szavak hasonlóságáról és az analógiás források kiválasztási módjáról alkotott elképzeléseim** (bővebben 4. fejezet) **mentálisan reális voltának megvizsgálására** egy tesztet is elvégeztettem magyar anyanyelvi beszélőkkel. Ez a vizsgálat nem előzmények nélküli. Nagyban **támaszkodok Lukács (2002)** eredményeire, **módszereire** és részben azon problémakörök feltárására törekszem (fonémák hasonlóságának hatása a szóalakok hasonlóságára a pozíció függvényében), amelyeknek szükségességét már ő is megfogalmazta dolgozatának konklúziójában. Habár Lukács (2002) több olyan eredményt is felmutatott, amellyel a magyar morfológiáról való gondolkodásban új, kreatív irányvonalakat szabott ki (hasonlóság mérése pozíciók és megkülönböztető jegyek figyelembe vételével), megállapításait nehezebben tudta igazolni a **kutatásra kevésbé alkalmas korpuszok használata[1] és hasonlósági kritériumainak intuitív** volta miatt. Mivel disszertációm alapkérdései (hogyan hat a gyakoriság a morfológiában, illetve a hasonlóság mérése miképp lehetséges szóalakok közt) nem volt központi témája dolgozatának, a tesztelésében **felhasznált szóanyag** (alacsony elemszám, nem szisztematikus eltérések a lehetséges mintáktól, nem tökéletes illeszkedés a hangkivető sémára) **sem volt a legmegfelelőbb** ezeknek a kérdéseknek a megválaszolására.

Vizsgálatomban a korábbi eredmények és alapkérdéseim mentén **az alábbi állítás megerősítését tűztem ki elsődleges célul**:

[1] Nem Lukács (2002) hibája, hogy kevésbé tudott támaszkodni korpuszadatokra, hisz 2002-ben még nem volt kész a *Szószablya Korpusz*, amely méretéből és részletességéből kifolyólag elsőként volt alkalmas ilyen típusú vizsgálatokban való felhasználásra.

(1) A szavak végéhez közelebbi fonémák hasonlósága, azonossága esetén két szót hasonlóbbnak éreznek a beszélők a toldalékolás szempontjából (szuffixumok esetében), mint ha a hasonlóságok, azonosságok a szavak belsejében vagy a bal szélén figyelhetők meg.

Az állítás bizonyítása során feltételezem, hogy **minél hasonlóbb egy szó** az analógia forrásául szolgáló szóhoz/szavakhoz, **annál hasonlóbban fog viselkedni** ahhoz/azokhoz, így hangkivetési mértéke is a mintaként szolgáló szó/szavak hangkivetési mértékéhez fog közelíteni hasonlóságuk függvényében, amelyre gyakoriságuk is hatással van. Ez összhangban áll Rebrus és Törkenczy (2008: 757) alábbi megállapításával is:

„az anyanyelvi beszélők a fonológiailag hasonlónak érzett/tartott elemeket morfofonológiailag hasonlónak hajlamosak venni"

Az egyes pozíciók jelentőségének a mértéke szócsoportonként eltérő lehet, de a tendenciák az egész magyar szókészletre egyformán igazak és érvényesek lehetnek. Ezt természetesen csak más nyelvi jelenségeken végzett további tesztekkel tudnánk igazolni. Az (1) alatt megfogalmazott állítás megvizsgálásán túl a teszt segítségével azt is tanulmányozom, hogy a **hasonlóság milyen módon befolyásolja a szavak viselkedését**, illetve arra a kérdésre is választ keresek, hogy **létezik-e** a vizsgált szavak esetében **produktív séma**, és ha igen, akkor milyen az.

7.2. A vizsgálat felépítése

A szavakon belüli fonémapozíciók fontosságának és a szóviselkedésre hatással levő analógiás forrásoknak a meghatározását célzó tesztemet 116 résztvevő (75 nő, 51 férfi) a ***surveygizmo.com*** **online alkalmazás** segítségével töltötte ki (hasonlóan online tesztet alkalmaztak Hayes és mtsai (2009) is a magyar magánhangzó-harmóniával

kapcsolatos vizsgálatukban). A teszt kitöltésére személyesen, emailben, illetve a *facebook.com*, *iwiw.hu* és a *twitter.com* oldalakon kértem fel a résztvevőket. A teszt időtartama átlagosan 30-40 perc volt, a résztvevők kor, lakhely és végzettség szerinti csoportjait a 7.1.-7.3. táblázatok mutatják meg.

Korcsoport évek szerint	Korcsoport aránya
14-17	3%
18-24	23%
25-39	54%
40-59	13%
60-	7%

7.1. táblázat: A résztvevők korcsoportjainak arányai

Lakhely	Lakhelyek aránya
Budapest	33%
nagyváros	34%
kisváros	22%
falu	11%

7.2. táblázat: A résztvevők lakhelyeinek arányai

Végzettség	Végzettség aránya
általános iskola	3%
középiskola	21%
főiskola	25%
egyetem	45%
PhD	7%

7.3. táblázat: A résztvevők végzettségének arányai

A teszt elkezdésekor a résztvevők részletes instrukciót olvashattak a kitöltés módjáról. A teszt kitöltéséhez ezután kezdtek hozzá, amely felépítésükben egyforma feladatokból állt. Az egyes feladatokban az első mondat egy tesztszót (valódi *CVCVC* szerkezetű hangkivető szóhoz hasonlító álszó) vezetett be egyes szám alanyesetben. Ezt egy olyan mondat követte, amelyben egy **tárgyesetű[1] alakot** vár el a mondat felépítése,

[1] A tárgyesetre azért esett a választásom, mert az 5.4.2. alfejezetben ez bizonyult a szavak teljes hangkivetési mértékével legszorosabban együttjáró toldaléknak a *Szószablya Gyakorisági Szótár* alapján ($r = 0{,}92$, $p < 0{,}001$).

aminek a helye üresen állt. A mondatot követően a résztvevőknek egy tárgyesetű alakot kellett választaniuk több lehetőség közül, amely megítélésük szerint jobban vagy természetesebben hangzott. Egyszerre csak egy alakot választhattak. Öt/három tárgyesetű alak[1] szerepelt a választási lehetőségek közt, hasonlóak, mint amilyeneket a *Google Gyakorisági Gyűjtés* elkészítése során alkalmaztam (bővebben 5.4.1. alfejezet). A mondatok alatt egy kép szerepelt, ami „valós" kontextust hivatott adni nekik. Ezeknek megfelelően a *szopor* álszónál például a következő feladattal találkozhattak a résztvevők:

(2) A szopor vállon hordható prém volt. A ..... főleg nemesasszonyok engedhették meg maguknak.

❍ szoprat
❍ szoprot
❍ szoport
❍ szoporot
❍ szoporat

[1] Utolsó magánhangzójukban *ö*-t vagy hátulképzett magánhangzót tartalmazó szavak esetén: pl. nyitó – hangkivető (*gaprat*), nyitó – nem hangkivető (*gaporat*), nem nyitó – hangkivető (*gaprot*), nem nyitó – nem hangkivető – kötőhanggal (*gaporot*), nem nyitó – nem hangkivető – kötőhang nélkül (*gaport*). Az utolsó magánhangzójukban *e*-t tartalmazók szavak esetében a nyitás nem értelmezhető, ezért van csak három tárgyesetű alakvariánsuk: pl. *keplet, kepelet, kepelt.*

A tesztben alkalmazott álszavak esetében nem fedtem fel, hogy azok kitaláltak. Csak annyit említettem, hogy a magyar köznyelvben használatuk nem jellemző. A teszt során ezeket a szavakat elavult vagy népies szavakként állítottam be, hogy a résztvevők hozzáállását ne befolyásolja az a tudat, hogy álszavakkal dolgoznak. A tesztben használt álszavak megalkotásához az általam gyűjtött összes hangkivető szóból **91 *CVCVC* mintát követő szót** (7.4. táblázat) vettem. A sémába illő, de eltérően viselkedő hangátvetéses szavakat kihagytam a vizsgálatból.

	hang-kivetés mértéke	példány-gyakoriság
pityer	0,00%	46
tegez	23,94%	447
veder	24,08%	1059
bajusz	35,91%	4985
bögöly	45,50%	222
gyilok	59,15%	213
koboz	63,45%	145
sulyom	65,00%	40
kebel	67,39%	10189
vacak	75,75%	3018
murok	78,57%	14
gyalom	80,00%	5
bürök	87,50%	32
bodor	88,89%	45
lator	89,76%	547
cseber	90,91%	33
kazal	90,97%	310
bagoly	92,30%	1493
köböl	92,59%	108
kapor	93,23%	532
nyirok	93,33%	45
retek	95,86%	604
pocok	96,39%	554
vöcsök	96,51%	86
korom	96,93%	619
pecek	97,54%	122
hurok	98,13%	4269
pokol	98,23%	4962
csupor	98,31%	237
berek	98,36%	5929
lepel	98,38%	5920
takony	98,48%	527
szeder	98,50%	200
fodor	98,51%	673
telek	98,60%	43541
sulyok	98,73%	474
vödör	98,86%	1669
horony	98,91%	829
verem	98,99%	3063
tülök	99,02%	204
bugyor	99,10%	1885
szurok	99,13%	230
szutyok	99,15%	235
kölyök	99,19%	11731
pöcök	99,28%	553
titok	99,99%	82321
farok	99,99%	34254
kapocs	99,98%	104934
sarok	99,98%	48702
bokor	99,98%	10110
gyomor	99,92%	19055
haszon	99,92%	50002
szobor	99,92%	37898
burok	99,90%	3156
dolog	99,89%	566770
csokor	99,89%	2792
barom	99,88%	4097
piszok	99,88%	832
selyem	99,87%	1543
horog	99,86%	2088
halom	99,86%	3538
gödör	99,86%	6570
szatyor	99,85%	1325
marok	99,84%	7369
cukor	99,84%	10281
fogoly	99,80%	11261
mocsok	99,80%	2550
cirok	99,80%	3071
torony	99,80%	16922
majom	99,70%	7618
tücsök	99,67%	910
meder	99,61%	5325
kölök	99,58%	1190
torok	99,42%	19524
csülök	99,39%	494
köröm	99,35%	9264
nyereg	99,34%	4999
tükör	99,29%	35387
pucor	100,00%	3
lucsok	100,00%	32
vacok	100,00%	47
rücsök	100,00%	54
csöbör	100,00%	64
töbör	100,00%	434
bütyök	100,00%	444
tulok	100,00%	469
szitok	100,00%	763
szirom	100,00%	3514
malom	100,00%	4789
tüsök	100,00%	4925
karom	100,00%	17787

7.4. táblázat: Az álszavak megalkotásához használt hangkivető szavak. A hangkivetési mértékek és a releváns toldalékok szerinti gyakorisági adatok a *Szószablya Gyakoriság Szótár* alapján szerepelnek a táblázatban.

A tesztben szereplő **álszavakat** úgy hoztam létre, hogy a 7.4. táblázat szavainak **1., 2., 3., illetve 5. fonémáját lecseréltem egy másikra**. Minden valódi szónak így négy változatát készítettem el: pl. *kapor: gapor, kopor, kabor, kapol*. Az álszavak is megfelelnek a *CVCVC* sémának, valamint Rebrus és Törkenczy (2008) megszorításainak is, amelytől csak annyiban tértem el, hogy ezekben is megengedem az /a/ és /u/ fonémák előfordulását az utolsó szótagban. Fontos, hogy a **tesztszavak valódi szavakat is felidézzenek**, hisz gyakorisági, hasonlósági hatásokat csak így várhatunk ezeknél, ezért a tesztszavak fonémáit minden esetben olyan fonémára cseréltem le, amely az adott pozícióban ténylegesen előfordul legalább egy szóban. Elképzelhető, hogy más fonémák hiánya is csak véletlenszerű, de mivel a beszélők ezekből a véletlenszerű előfordulásokból általánosítanak, vagy ezekhez hasonlítanak, ehhez a kitételhez ragaszkodtam.

A szavak egyes pozícióiban lévő fonémákat úgy változtattam meg, hogy a **lecserélésre használt fonéma a lehető legközelebbi legyen a lecserélthez**. A megváltoztatott jegyek a mássalhangzók esetében a képzési hely, mód, zöngésség voltak, a magánhangzók esetében a nyíltságot módosítottam. Egyedül az /i/-t voltam kénytelen /u/-ra cserélni a 2. fonémapozícióban, nem pedig a nyíltságát vagy kerekségét változtatni, mert nincs a hangkivetők közt *-e-o, -ö-o, -ü-o* magánhangzó-szekvencia, csak a még lehetséges *-u-o* fordul elő. A teljes hangkivető anyagon végzett megfigyeléseim alapján látható (5.4. táblázat), hogy a hangkivető szavak kerekség szerint is harmonizálnak (leszámítva a már említett *-i-o* magánhangzó-szekvenciát tartalmazóakat). Nincsenek *-e-ö*, *-ö-e* vagy *-ü-e* szekvenciák, így ilyeneket álszavaimban sem engedek meg. Ha a létező *CVCVC* szerkezetű szavak utolsó magánhangzója elölképzett, akkor nem tudom azt úgy megváltoztatni, hogy ne sértsem meg a hangkivető séma alaki jellemezőit. A nyíltság változtatása esetén a pozícióban nem előforduló /i/, /ü/ fonémákat kapnék, ha pedig ezeknek a fonémáknak kerekségét vagy elölségét változtatnám meg, akkor a hangkivetők közt szintén nem létező

magánhangzó-szekvenciákat kapnánk: *-*(ö/ü)-(e/i/o/u/a)*, *-*(é/e/i)-(ö/ü/u/a)*, ezért az ebben a pozícióban lévő fonémáknak a változtatását kihagytam a tesztből[1].

A mássalhangzók közelségét az IPA (International Phonetic Alphabet, Nemzetközi Fonetikai ABC) csak a magyar mássalhangzókat tartalmazó táblázatának (7.5. táblázat) cellái alapján számoltam úgy, hogy egy hang annál jobban hasonlít egy másikra, minél kevesebb cellára van tőle (az átlósan elhelyezkedő cellák távolsága 2). Egy fonéma lecserélésére ez alapján választottam ki a legközelebbi, legkevesebb cellára lévő másik fonémát, ami az adott pozícióban valós hangkivető szavakban is előfordul (7.6. táblázat). A zöngésségbeli eltérést egy cellányi különbségnek számítottam. Ha ugyanannyi eltéréssel egy fonéma mellé több fonémát is választhattam, akkor azokat részesítettem előnyben, ahol a zöngésséget, másodsorban a helyet kellett inkább megváltoztatni. Ezt abból kiindulva tettem, hogy elképzelésem szerint ezek a változtatások kevésbé változtatják meg a „fonémák karakterét", ezzel tulajdonképpen a Lukács (2002) által megjelölt szempontok szerint jártam el. Ahol két változat közül így sem lehetett dönteni, ott 50-50%-ban hol ezt, hol azt a fonémát választottam. A fonémák távolságának megállapításához használt 7.5. táblázat egyes sorait szonoritás szerint rendeztem:

(3) zárhang > affrikáta > réshang > h > nazális > r , l > félmagánhangzó

A **szonoritási skála** meghatározásában Parker (2002: 236) alapján jártam el, amely követi a Kiparsky (1981) által megadott szonoritási hierarchiát, illetve több, ehhez hasonló szonoritási skála felépítését (Cser 2000) is. Az /r/-t és az /l/-t azonban azonos szonoritásúnak veszem Kornai (1990) magyar példái alapján, aki Parkertől (2002: 236) eltérő sorrendezésüket mutatja meg magyar szavaknál. A /v/-t réshangnak veszem, bár csak szókezdő pozíciókban fordul elő a vizsgált szóanyagban.

[1] Kivételként itt az *-i-e* szekvenciát tartalmazó szavakat lehet megemlíteni, amelyek esetében lehetséges lenne egy ilyen átalakítás (pl. *pityer* > *pityor*).

	bilabiális	labio-dentális	alveoláris	poszt-alveoláris	palatális	veláris	glottális
zárhang	p b		t d		ty gy	k g	
affrikáta			c	cs			
réshang		f v	sz z	s			
h							h
nazális	m		n		ny		
likvida			l r				
félmagánhangzó					j		

7.5. táblázat: Egyszerűsített IPA-táblázat, amely csak a *CVCVC* hangkivető főneveknél előforduló mássalhangzókat tartalmazza.

1. pozíció fonémái	b, c, cs, d, f, g, gy, h, k, l, m, ny, p, r, s, sz, t, v
2. pozíció fonémái	e, i, ö, ü, o, u, a
3. pozíció fonémái	b, c, cs, d, g, gy, j, k, l, m, p, r, s, sz, t, ty, z
5. pozíció fonémái	cs, g, j, k, l, m, n, ny, r, sz, z

7.6. táblázat: a *CVCVC* szerkezetű hangkivető főnevekben előforduló fonémák

Az **álszavakat** (összesen 364) **számos esetben amiatt kellett javítani**, hogy az utolsó és utolsó előtti mássalhangzójuk zöngésségben eltért, vagy mind a kettő affrikáta vagy réshang volt (pl. *bcs*, *bsz*, *cg* stb.), azaz nem feleltek meg így a sémának, amelyet Rebrus és Törkenczy (2008) állított fel. A szavakat időnként szükségszerű volt úgy is átalakítani, hogy a két utolsó mássalhangzó egymás mellé kerülése esetén elkerüljem a palatalizációt, adaffrikációt és a szibiláns hasonulást (Siptár 1994). Számos esetben amiatt voltam kénytelen a tesztszavakat módosítani, mert valódi szavakkal vagy szóalakokkal azonosak lettek alanyesetükben (*bodor: botor, burok: borok, verem: velem* stb.) vagy valamely tárgyesetű alakjukban (*csupor: csopor: csoport, gyalom: gyarom: gyarmat, halom: harom: harmat*). A tesztszavakban a *kurom* szó kétszer fordul elő, a *karom*-ból és a *korom*-ból létrehozott azonos álszóként, de ezeket az adott pozícióban más módon nem lehetett átalakítani, így ezt az ismétlődést meghagytam. Hasonlóan jártam el a *morok* álszóval is (*marok, murok*). Ha *j*-re cseréltem egy fonémát, akkor minden esetben *ly*-t

tettem az álszóba, hogy látszólagos régiességének, elavultságának érzetét erősítsem. Ezekben a pozíciókban egyébként is előfordul *ly*-nal írott /j/. A *csücsök* szó /ü/ fonémáját nem tudtam úgy megváltoztatni, hogy a sémának megfeleljen, és ne valódi szóalakra hasonlítson (*csöcsök*), így e szó alapján egyáltalán nem készítettem álszavakat (eredetileg 92 *CVCVC* szerkezetű szót választottam ki).

A **tesztet 4 változatban** készítettem el eltérő álszavakkal (a teszt egyik változata a B Függelékben látható). Minden tesztváltozatban 91, a *CVCVC* hangkivetők alapján készített álszó szerepelt, amelyek közé 19, a tárgyeset szempontjából más rendhagyó mintát utánzó álszavakat kevertem (*bicső-bicsövet, gusár-gusarat* stb.), hogy a teszt célja kevésbé legyen egyértelmű a résztvevők számára. Egy résztvevő egy szónak csak egy változatával találkozott. Minden résztvevő ugyanolyan arányban találkozott olyan szavakkal, amelyeknek az 1., 2., 3., illetve az 5. fonémiapozíciójában eszközöltem változtatást. A kérdések minden résztvevő esetében azonos sorrendben következtek, a lehetséges válaszok sorrendje az egyes tesztváltozatok kérdésein belül véletlenszerű volt. Ezek fényében a teszt első változatát kitöltő személyek az alábbi álszavakkal találkoztak az első tíz kérdésben (4). Itt a figyelemelterelő szavakat kihagytam, a megváltoztatott fonémákat pedig aláhúztam. (5) alatt azokat a szavakat sorolom fel, amelyek ezen szavak kialakításához mintaként szolgáltak.

(4) *csarok, gyolom, maszom, kapol, pugyor, birek, serem, szedel*

(5) *sarok, gyalom, majom, kapor, bugyor, berek, selyem, szeder*

7.3. A hangkivetésre ható nyelvi tényezők elemzése

A **hangkivetés átlagos mértéke 36,8%** volt a válaszokban[1]. Az általános lineáris modell, ismételt mérések statisztikai eljárás (GLM, repeated measures) szerint az **egyes fonémapozíciók (tartalmának) a hangkivetés mértékére gyakorolt hatása szignifikánsan eltér egymástól** ($F(3,273) = 44,5$, $p < 0,001$). Az egyes fonémapozíciókban a hangkivetés mértékét és az azok közti eltérések szignifikanciáját a 7.7. táblázat mutatja meg.

	hangkivetés mértéke	szignifikáns eltérések	szórás	min.	max.	leginkább hangkivető szavak
1. fonéma	44,2%	> 3. fonéma ** > 5. fonéma ***	21,9	0%	93,1%	lücsök, rucsok, böcök, pürök, dücsök
2. fonéma	41,5%	> 5. fonéma ***	20,1	0%	86,2%	pücök, vücsök, vocok
3. fonéma	39,2%	> 5. fonéma ***	22,1	0%	79,3%	vöcök, rücök, tücök, tüszök, surom
5. fonéma	22,5%		13,5	0%	58,6%	sulyog, hurocs, bögöl

7.7. táblázat: A fonémapozíció hatása a hangkivetés mértékére

** = $p < 0,01$

*** = $p < 0,001$

Eredményeim összhangban vannak a 4. fejezetben kifejtett elképzelésekkel és a szóalak jobb szélétől távolodva a hasonló fonémáknak egyre kisebb súlyt adó algoritmusaimmal is. A 4 fonémapozíció esetében a hangkivetés mértéke még ha nem is egyformán, de nőtt balra felé haladva. A legfontosabb eltérés algoritmusaim

[1] Azon válaszokban, ahol volt kötőhangzó, és az utolsó magánhangzó nem /e/ volt az alanyesetű alakban, a nyitás összesen 2,3%-ban fordult elő, így ez nem jelent meg akkora mértékben, hogy vizsgálat tárgya lehessen. A leginkább nyitó szavak a következők voltak: *tutok* (27,6%), *falyok* (17,2%), *ditok* (10,3%), *barok* (10,3%), *kozal* (10,3%), *vacony* (10,3%), *takon* (10,3%), *vacany* (10,3%).

működésétől, hogy azok a hasonlóságnak kisebb súlyt adnak a szó belsejében és elején (bővebben 4.3. alfejezet), mint ahogy azt teszteredményeim alapján elvárhatnánk, hisz hiába van az 1. pozíciónál 10,49-szer ($1,8^4$) nagyobb súlya szerintük az 5. pozíciónak, még ennek a megváltoztatása esetén is 22,5%-nyi hangkivetést kaptam. Ebből azt láthatjuk, hogy egy holisztikusabb, egészlegesebb megközelítés jobban képes lehet megragadni azoknak az elemeknek a viselkedését, amelyekkel kapcsolatban nincsenek (könnyen) elhívható emléknyomaink.

Az 1., 3. és az 5. pozícióban lévő fonémák módosítása után kapott tesztszavak hangkivetési mértéke közt a különbségek szignifikánsak, amiből arra következtethetünk, hogy az **5. és mérsékelten a 3. pozícióban lévő fonémák megváltoztatása után kapott szavakat a beszélők kevésbé hasonlónak érzékelik az eredeti szóhoz**. Ebből kifolyólag kevésbé hasonlóan is használják, mint ha az 1. pozícióban hajtottam volna végre a változtatást. Ezt erősíti meg, hogy az 50%-nál alacsonyabb hangkivetési mértékkel rendelkező szavak alapján létrehozott álszavak esetén a beszélők szignifikánsan kevesebb ($t(25,62) = -4,13$, $p < 0,001$) hangkivető alakot választottak (18,1% szemben 38,2%-kal). Ez alól csak akkor tapasztalunk kivételt, ha az álszó egyértelműen jobb hangkivető szó lett, mint az eredeti, valódi szó: *cseber* > *cseper* (az utolsó két mássalhangzó kevésbé hasonlít), *pityer* > *petyer* (az *-e-e* szekvencia jellemzőbb a hangkivetőkre, mint az *-i-e*), *tegez* > *tegen* (az *-n* vég jellemzőbb a hangkivetőkre, mint a *-z).* A jobb szélhez közeli változtatások esetén az eredeti szótól való erőteljesebb eltávolodást támasztják alá az eredeti szavak hangkivetési mértékével mért együttjárások is (7.8. táblázat), mivel azt tapasztaljuk, hogy a szavak vége felé az együttjárások gyengülnek (csak a *Google* esetében), sőt az utolsó pozíció esetén már nem is szignifikánsak, ami alapján arra következtethetünk, hogy az eredeti szó hasonlósági hatásai minimálisra csökkentek. Az eredeti szó gyakoriságának hatását azonban egyik pozíció esetében sem lehet kimutatni, azaz a gyakori hangkivető szavak mintájára készített álszavak hangkivetési mértéke nem magasabb.

	1. fonéma változtatás	2. fonéma változtatás	3. fonéma változtatás	5. fonéma változtatás
***Szószablya* tárgyeset**	0,320**	0,263*	0,345***	0,205 (n.sz.)
***Google* tárgyeset**	0,279**	0,207*	0,227*	0,185 (n.sz.)

7.8. táblázat: A hangkivető főnevek tárgyesetének hangkivetési mértéke és a tesztszavak hangkivetési mértéke közt mérhető korreláció

* = $p < 0,05$

** = $p < 0,01$

*** = $p < 0,001$

Habár a **2. pozíció változtatása esetén a kapott hangkivetési mértékek elvárásaimmal összhangban alacsonyabbak**, mint az 1. pozíció esetén, és magasabbak, mint a 3. pozíciónál, ezek a különbségek nem szignifikánsak. Ez azzal magyarázható, hogy a 2. pozícióban magánhangzók, a többiben pedig mássalhangzók találhatók. Elképzelhető, hogy a magánhangzók és a mássalhangzók hasonlósága eltérő fontossággal bír a beszélők számára, és ez a különbség tükröződik vissza adataim nem egyértelmű voltában. Másrészt az eredeti és a lecserélt fonéma közti hasonlóság számai alapján azt láthatjuk (7.9. táblázat), hogy a magánhangzók mértékei más jellegűek, mint a mássalhangzókéi, nehezebben összehasonlíthatók. Az értékek a komplex jegymérték esetében a leghasonlóbbak, de még itt is látunk különbségeket (pl. a 2. és az 5. fonéma értékei között). A mássalhangzóknak és a magánhangzóknak a hasonlításban betöltött eltérő viselkedését csak további, kifejezetten erre kialakított tesztekkel lenne lehetséges feltárni.

	1. fonéma	2. fonéma	3. fonéma	5. fonéma
egyszerű jegymérték	0,430**	0,477	0,441*	0,398***
komplex jegymérték	0,385	0,416	0,384	0,317***
természetes osztályok	0,351***	0,487	0,43	0,384**

7.9. táblázat: A fonémák átlagos hasonlósága a lecserélt fonémához hasonlósági mértékeim szerint.

szignifikancia szintek a 2. fonéma átlagához mérve:

*** = $p < 0,001$

** = $p < 0,01$

* = $p < 0,05$

Érdemes megvizsgálnunk az egyes pozíciókban annak hatását, hogy az eltérések mértéke nem függhet-e össze azzal, hogy az egyes fonémákat milyen másikra cseréltem le. Különösen, mivel az utolsó, 5. fonéma esetén kevesebb fonémából választhattam, amikor az eredeti szavakból létrehoztam az álszavakat. Elképzelhető, hogy az alacsonyabb hangkivetési mérték annak tudható be, hogy ott távolabbi, kevésbé hasonlító fonémákat választottam. Ha az egyszempontú varianciaanalízisben a pozíció mellé kovariánsnak az eredeti és a lecserélt fonéma hasonlóságát adom meg a 7.9. táblázatban is szereplő mértékek szerint, akkor azt tapasztaljuk, hogy a **fonémák hasonlósága az egyszerű és a komplex jegymérték esetében nincs befolyással a már megfigyelt hatásra, miszerint a pozíció függvényében változik a hangkivetés mértéke**. A természetes osztályok esetében van hatása ($p < 0,05$) a lecserélt fonéma hasonlóságának az álszó hangkivetésének a mértékére, de mint a 7.9. táblázatból is látható, ez nem abból következik, hogy az utolsó fonémapozícióban kevésbé hasonló fonémák lecserélésével hoztam létre álszavaimat. A mássalhangzókat fogadó pozíciókban az eredeti és a lecserélt fonéma hasonlóságának értékei közt szignifikáns különbség csak az 1. és a 3. pozíció közt mutatható ki ($p < 0,05$), ami arra utalhat, hogy ha az 1. pozíciói eredeti és lecserélt fonémái a természetes osztályok szerint is hasonlóbbak lennének, akkor itt még magasabb hangkivetési értékeket kaphatnánk,

mint a mért 44,2%. Így ez az eredmény csupán annyit jelent, hogy a fonémapozíciók jelentőségének csökkenése nagyobb lehet (erre utalhatna az 1. fonémapozícióban lévő fonéma megváltoztatása esetén mérhető magasabb hangkivetési mérték), mint amit tesztem kimutatott.

A tesztnek nem volt célja, hogy megvizsgáljam az **egyes mássalhangzós jegyek feltételezhető súlyait**[1]. Erre vizsgálati anyagom sem lett megfelelő módon összeállítva, hisz az egyes jegyek mentén való cserék nem egyforma arányban szerepelnek az álszavakban. Ez egyrészt az egyes pozíciókban választható fonémák köréből, másrészt preferencia megszorításaimból (elsősorban a zöngésség, másodsorban a hely viszonylatában változtattam) is adódott. 48 esetben cseréltem helyet (hangkivetés mértéke változtatása esetén: 41,3%), 91 esetben zöngésséget (hangkivetés mértéke változtatása esetén: 33,6%) és 57 esetben módot (hangkivetés mértéke változtatása esetén: 36,8%). Az egyes jegyek mentén való változtatások viszonylatában a szavak hangkivetési mértékeinek átlagai a Welch-próba alapján nem minősültek szignifikánsan eltérőnek[2]. Az egyes pozíciók függvényében a 7.10. táblázat mutatja meg, hogy melyik jegy változtatása esetén milyen hangkivetési mértéket kaptam. A 3. és az 5. fonémapozícióban eszközölt változtatások esetében a zöngésségükben eltérő álszavak hangkivetési mértéke szignifikánsan különbözik azokétól, amelyek hely vagy mód jegyükben mások, mint az eredeti hangkivető főnév. Ez a jelenség betudható lehet annak, hogy a zöngésség fontosabb jegy (eredeti feltételezésemmel ellentétben), mint a többi. Fontossága különösen intervokalikus helyzetben emelődik ki, míg a szóvégen, ahol a zöngés fonémák zöngéje csökkenhet (különösen réshangok esetében), kisebb jelentőséggel bírhat. Ezt a feltételezést azonban ismét csak egy olyan teszttel lehetne megerősíteni vagy elvetni, amely elsősorban ennek a jelenségnek a megfigyelésére lett összeállítva.

---

[1] A magánhangzók esetében ezt a kérdést fel sem tehetjük, hisz náluk csak a nyíltságot változtattam.

[2] Összesen 273 mássalhangzót cseréltem le, amelyeknél 77 esetben több jegyet is kellett módosítani. Ezeknél a szavaknál nem vizsgáltam a jegyek lehetséges súlyait, hisz azok hatását nem tudjuk elkülöníteni egymástól.

	1. fonéma	3. fonéma	5. fonéma
zöngésség	41,6%	24,3%	31,7%
mód	48,3%	47,7%***	21,7%*
hely	44,5%	56,9%***	15,2%***

7.10. táblázat: A hangkivetés mértéke a megváltoztatott jegyek és a fonémapozíciók függvényében

eltérés szignifikancia szintje a zöngésségükben különböző álszavak hangkivetési mértékének átlagaitól:

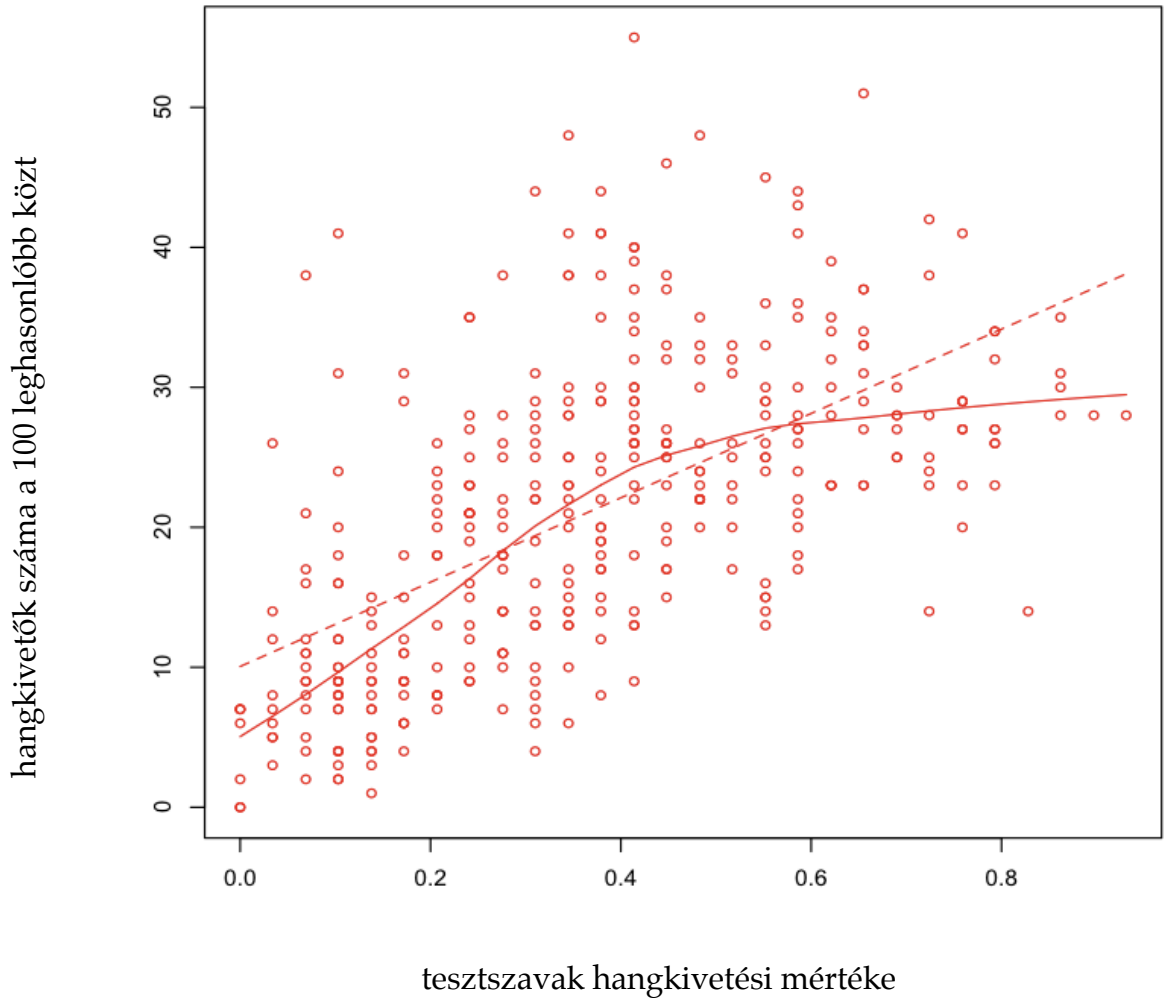
***= $p < 0,001$

* = $p < 0,05$

További változókat is megvizsgáltam a tesztszavak hangkivetésének mértékével kapcsolatban. A 6. fejezetben alkalmazott módszerrel, a leghasonlóbb pár alapján nem lehet jól megragadni az álszavak viselkedését (komplex jegymérték négyzetes hiba: 0,31; komplex tengelymérték: 0,32; minden álszót hangkivetőnek várva: 0,45; minden álszót nem hangkivetőnek várva: 0,18, ha a hangkivetést 0 és 1 közt mérjük). Ebből azt láthatjuk, hogy a szótári anyagomban domináns viszonylag egyértelmű viselkedés helyett a bizonytalabb viselkedés nem modellálható jól csak a legközelebbi szomszédra hagyatkozva. Ugyanakkor, **ha az algoritmusaimmal kiválasztott 100 leghasonlóbb főnevet veszem figyelembe, akkor az ezek közt található hangkivető szavak száma korrelál a hangkivetési mértékkel** (komplex tengelymérték: $r(362) = 0,598$, $t = 14,19$, $p < 0,001$; komplex jegymérték: $r(362) = 0,482$, $t = 10,46$, $p < 0,001$). Ez alapján azt láthatjuk, hogy egy új szó viselkedését, különösen bizonytalan határestekben, a közeli elemeken belüli csoportok típusgyakorisága határozza meg. Ez összhangban áll a korábbi analógiás alapú megfigyelésekkel is (Bybee 2010), miszerint egy csoport produktivitását elsősorban típusgyakorisága határozza meg.

A tesztszavak hangkivetési mértékével szignifikánsan együttjár még annak a hasonlósági csoportnak a példánygyakorisága, amelybe a tesztszó tartozhatna ($r(362) = 0,408$, $t = 8,49$, $p < 0,001$, komplex tengelymérték), és amelyikbe az a szó tartozik, amelyik alapján elkészítettem ($r(362) = 0,352$, $t = 7,15$, $p < 0,001$, komplex tengelymérték). Közepesen korrelál a tesztszavak hangkivetési mértéke azzal, hogy mennyire hasonlítanak a komplex jegymérték által meghatározott hasonlósági

csoportok leggyakoribb szavaihoz (r(362) = 0,367, t = 8,56, p < 0,001). Ezek alapján azt láthatjuk, hogy egy csoport erejét, produktivitását nemcsak annak számossága, hanem elemeinek feltűnősége, gyakorisága is meghatározhatja. Ez a prototípusos viselkedés összhangban áll Casenheiser és Goldberg (2005) eredményeivel, akik azt tapasztalták, hogy 5 és 7 év közti gyerekek könnyebben tanulják meg az azonos típus- és példánygyakoriságú konstrukciókból azt, amelyikben van jelentősen kiugró példánygyakoriságú prototípusnak tekinthető elem.



7.1. ábra: a komplex tengelymérték alapján meghatározott leghasonlóbb hangkivető szavak számának és a tesztszavak hangkivetési mértékének együttjárása

A **lépésenkénti lineáris regresszióanalízis** (stepwise linear regression) alapján (kiigazított $R^2 = 0,489$, $F(7,352) = 49,16$, $p < 0,001$) a **7.11. táblázatban** látható **változók együttes hatásával** magyarázható a teszt során kapott szavak hangkivetési mértéke. A hangkivetés mértékére hatással levő változók közt **már korábban megfigyelt**

tényezőket ismerhetünk fel. A legnagyobb hatással szereplő „legközelebbi hangkivető szavak száma (komplex tengelymérték)" azt mutatja, hogy az egyes szavak viselkedését, amelyeknek a viselkedésével kapcsolatban nincsenek emléknyomaink, a hozzájuk szerkezetileg leghasonlóbb szavak tömege befolyásolja. Ebben jóval gyengébb mértékben szerepet kap a hasonló végekkel rendelkező szavak tömege is (legközelebbi hangkivető szavak száma, komplex jegymérték). A legközelebbi szomszéd hatása is érzékelhető a hangkivetés mértékére (legközelebbi szó hangkivető-e, komplex jegymérték), de itt elsősorban a tővég és nem a szerkezet bír nagyobb jelentőséggel. Ez összhangban áll a 6.3. alfejezet teszteredményeivel is, amikor a csak a legközelebbi szón alapuló bináris döntésekben a komplex jegymérték bizonyult a sikeresebbnek. A legközelebbi szomszéd önállóan is érvényesülő hatása azt mutatja, hogy a leghasonlóbb szavak hatása nem egyforma, a közelségnek több jelentőssséggel bíró fokozata is van.

Az álszavak viselkedését azonban **befolyásolják azok a valódi szavak is, amelyekből létrehoztam őket**. Szignifikáns együttjárást tapasztalhatunk azok tárgyesetű és általános hangkivetési mértékével, de ennél nagyobb jelentősége van annak, hogy az eredeti hangkivető szavak milyen gyakori szóhoz állnak közel. Ebben az esetben az eddig tárgyalt típusgyakoriság által meghatározott változók mellett egy olyan faktort is megfigyelhetünk, amely a példánygyakoriság és a prototípusosság jelentőségét emeli ki. Az álszavakra hatással van az is, hogy eredeti szavaik milyen távolságban vannak a hozzájuk legközelebbi prototípustól, ami amellett szól, hogy az alakok választásakor valamilyen mértékben előhívódhattak az eredeti alakok, így az ő viselkedésükre már igazoltan hatással lévő változók gyengébben és áttételesen itt is kifejtették hatásukat. A mesterségesen meghatározott prototípusaink érvényre jutó hatása alapján pedig azt láthatjuk, hogy az új elemek viselkedésére ható prototípusoknak nem csak a gyakorisága és közelsége számít, hanem az is, hogy milyen helyet foglalnak el a szavak hasonlósági rendszerében, hisz ezen prototípusok kiválasztásában azt a szempontot is érvényre juttattam, hogy minél messzebb legyenek más potenciális prototípusoktól.

	Béta (standardizált)	Szignifikancia
Legközelebbi hangkivető szavak száma (komplex tengelymérték)	0,459	p < 0,001
Eredeti szavak hasonlósága a gyakorisági prototípusokhoz (komplex tengelymérték)	-0,463	p < 0,001
Hangkivetési mérték (Szószablya)	0,250	p < 0,001
Legközelebbi szó hangkivető-e (komplex jegymérték)	0,174	p < 0,001
Eredeti szó hasonlósága a 0,9-es küszöbérték mellett kiválasztott prototípusokból és a leggyakoribb szavakból létrehozott halmaz szavaihoz (komplex tengelymérték)	0,268	p < 0,01
Eredeti szó hasonlósága a 0,7-es küszöbérték mellett kiválasztott prototípusokhoz (komplex tengelymérték)	-0,120	p < 0,01
Tárgyesetű alak hangkivetési mértéke (Google)	-0,129	p < 0,05
Legközelebbi hangkivető szavak száma (komplex jegymérték)	0,126	p < 0,05

7.11. táblázat: A tesztszavak hangkivetését befolyásoló változók[1]

Az eredmények alapján összefoglalva mondhatjuk, hogy a **hangkivető séma még elevenen él a magyar beszélők számára**. Az új szavak viselkedésére legnagyobb hatással a szerkezetileg hasonló szavak tömege (típusgyakoriság hatása), illetve a hozzájuk leghasonlóbb végű szó bír. Ezen túl hangkivetési mértékükre a szerkezetileg legközelebbi szó (ez néhány kivételtől eltekintve az a szó, amelyből létrehoztam őket) és annak a viselkedését kormányzó prototípus hat. Ezek a prototípusok elsősorban a kiugró gyakoriságú szavak.

Feltehetjük azonban azt a kérdést is, hogy vajon ez az aktív séma ugyanaz-e, mint amit Rebrus és Törkenczy (2008) meghatároztak. Ha megvizsgáljuk azokat a tesztszavakat, amelyeket a beszélőknek legalább kétharmada hangkivetően ragozott, akkor láthatjuk, hogy ez a séma szűkebb a Rebrus és Törkenczy által meghatározottnál. Összesen 36 ilyen szót találtam, amelyek kivétel nélkül *-Vk* (32 db) vagy *-Om* (4 db)

[1] A komplex tengelymérték prototípusai a 0,7-es küszöbérték mellett a következők: *társadalom, tükör, átok, érem, kehely, üdülőtelek, álbajusz, tüllfátyol.* A prototípusok közt a kevésbé hangkivető, egyedi szavak nagyobb súlyt kaptak.

végűek, míg a teljes vizsgálati anyagban összesen 109 *-Vk* végű és 36 *-Om* végű szó található. A Welch-próba alapján a *-Vk* és az *-Om* végűek hangkivetési mértékének átlagai nem különböznek ($p = 0{,}055$), de mind a két csoport szignifikánsan különbözik hangkivetési mértékének átlagában a többi szótól ($p < 0{,}001$), így a *CVCVC* szerkezetű hangkivető szavakkal kapcsolatban a beszélők fejében élő séma vége *-V(k|m)* szerkezetű lehet.

Mivel tesztszavaimat eleve úgy építettem fel, hogy Rebrus és Törkenczy (2008) kritériumainak megfeleljenek, így elmondható, hogy az aktív és élő séma ennél specifikusabb. E szűkebb séma produktivitása összhangban áll azon ismeretünkkel is, hogy a legnagyobb típusgyakoriságú csoportok a legproduktívabbak. Ez a *-Vk* végűekről egyértelműen elmondható, hisz 37-en vannak. Az *-Vm* végűek csoportja már kevésbé számos, hisz csak 12 ilyen szavam van, kevesebb, mint *-r* (23 db) végű, ugyanakkor minden bizonnyal ezek viselkedésére áttételesen a többi *-alom, -elem* végű szó is hatással van, míg az -r végűek hangkivetési mértékét kevesebb és ritkább nem *CVCVC* szerkezetű *-r* végű szó támogatja. A 7.12. táblázat alapján azonban azt is láthatjuk, hogy azok a szavak sem viselkednek egységesen, amelyek ezzel a szűkebb sémával jellemezhetőek. Az *-ök* végűek az összes többi végű szónál szignifikánsan (Welch-próba) következetesebb hangkivetők, amikhez hasonlóan viselkednek az *-öm* végűek is, amelyek magasabb hangkivetési mértékkel rendelkeznek, mint az összes többi álszó (leszámítva az *-Ok* végűeket).

	-öm	-ok	-om	-ek	-em	többi vég
-ök (hm: 0,64, sz: 0,17)	*	**	*	*	*	***
-öm (hm: 0,6 sz: 0,05)	-		*		*	**
-ok (hm: 0,55, sz: 0,16)	-	-				***
-om (hm: 0,49, sz: 0,17)	-	-	-			***
-ek (hm: 0,45, sz: 0,21)	-	-	-	-		***
-em (hm: 0,42, sz: 0,15)	-	-	-	-	-	*
többi vég (hm: 0,25, sz: 0,14)	-	-	-	-	-	-

7.12. táblázat: A végek alapján kialakítható tesztszócsoportok eltéréseinek szignifikancia szintjei a náluk kisebb hangkivetési mértékkel rendelkező álszó végek csoportjaitól (hm = hangkivetési mérték, sz = szórás)

***= $p < 0,001$

**= $p < 0,001$

* = $p < 0,05$

7.4. Az eredmények összegzése

Tesztem eredményeit a következő állításokkal foglalhatjuk össze:

- Az egyes fonémapozíciókban megfigyelhető hasonlóságok és eltérések különböző fontossággal bírnak szavak összevetése esetén.
- Legkisebb hangkivetési mértékkel azok az álszavak bírtak, amelyeknek utolsó fonémáját változtattam meg. A 4 fonémapozíció esetében a hangkivetés mértéke még ha nem is egyformán, de nőtt balra felé haladva.
- Az eredeti szó befolyásán túl leginkább az új szó egyedi viszonyrendszerének van szerepe az álszó hangkivetési mértékének alakulásában.
- Egy álszó viselkedésének alakulásában legnagyobb szerepe a szerkezetileg is hasonló szavaknak van, amit kiegészít a vége alapján legközelebbinek számító leghasonlóbb szó hatása is.

- A *CVCVC* szerkezetű főnevekkel kapcsolatban a *CVCök* séma a legproduktívabb.
- Az egyes fonológiai jegyek nem egyforma fontossággal bírnak a hasonlításban. Szóvégeken a zöngésség kisebb súllyal bír, mint a többi jegy, míg a szó belsejében ezeknél nagyobbal.

8. Konklúzió

8.1. Összegzés

Dolgozatomban azt kívántam megmutatni, hogy **a nyelv szerveződésében és változásában a hasonlóságnak és a gyakoriságnak kiemelt szerepe van**. A nyelvtudományban mind a kettő vizsgálatának nagy hagyománya van, de kutatásomban ezeket újszerűen közelítettem meg, illetve bővítettem a velük kapcsolatos ismereteinket. Vizsgálatomat a magyar főnevek egy közepes méretű csoportjának alaktani viselkedésével kapcsolatban végeztem el. A hangkivető főnevek önmagukban is eléggé érdekesek ahhoz, hogy érdemes legyen teljes körűen megvizsgálnunk őket, de leírásukra nem csak célként tekintettem, hanem arra szolgáló eszközként is, hogy a nyelvi viselkedés szervezésében alapvető jelentőségű tényezőket, a hasonlóságot és a gyakoriságot jobban megismerhessük. Kutatásomban megmutattam, hogy a szavak holisztikusan felfogott szerkezete és a fonémánál nagyobb egységet magukba foglaló szóvégek milyen nagy szerepet játszanak abban, hogy egy-egy szó miképp viselkedik, amit használati körülményei és jelentései is befolyásolnak. Elemzésem alapján az is jól látható, hogy a nyelvészet adatközpontú megközelítése jelentősen átrajzolhatja a most általánosan elterjedt nyelvészeti elméleteket, de a nyelvi jelenségek leírásai is jelentősen módosulhatnak ezáltal.

Elemzésem az 1. és a 4. fejezetben kifejtett elméleti alapvetések alátámasztásán túl leíró szinten is bővítette ismereteinket. A hangkivető főnevek viselkedése ugyanis sokkal heterogénebb, mint minden korábbi leírás állította vagy vélte. Lehetetlen viselkedésüket egységes sémákkal és szabályokkal megragadni, mert kivételek, egyedien viselkedő szavak és paradigmacellák mindig szép számban akadnak, ezért **általános érvényű leírások készítése helyett a heterogén viselkedésben az általános érvényű elvek hatásait kutattam**. Ennek részeként a szavak korpuszok segítségével gyűjtött egyedi tulajdonságai alapján gráfokat készítettem (5.3. alfejezet), amelyek révén könnyebben meg lehetett figyelni, hogy ezek az általános alaptényezők miként fejtik ki

hatásukat, és szervezik a szavak rendszerét. Az 5.4. alfejezetben ezt követően részletesen tanulmányoztam, hogy két közeli szinkrón állapot miképp tér el egymástól kizárólag a hangkivető főnevek változását figyelembe véve. A különböző szempontú vizsgálatok során azt tapasztaltuk, hogy az időbeli távolságból fakadó különbségek mellett a későbbi állapot eltéréseiben vizsgált anyagom vélhetőleg erőteljesebben informális jellege[1] is szerepet kaphat. A változásban a hasonlóság által meghatározott csoportok szavai viszonylag egységesen vesznek részt, de a gyakoriság befolyással van arra, hogy egy szó miképp változik, vagy miképp különbözik viselkedése az informálisabb regiszterekben.

Az 5. fejezet leíró jellegén túl azonban jóval nagyobb jelentőségűek azok a **statisztikailag is alátámasztható felfedezések**, amelyek a 2. fejezetben is tárgyalt fontosabb fogalmakhoz kapcsolódnak. Ezeket egészítik ki, illetve erősítik meg a 6. és a 7. fejezet vizsgálatai is. Az analógiás kiterjesztés és kiegyenlítődés tárgyalása kapcsán, a 2.1. alfejezetben megjegyeztem, hogy a hangkivető főnevek változását mind a két analógiás folyamattal összefüggésbe hozhatjuk, hisz tekinthetjük változásukat kiterjesztésnek, mivel egy általános sémát vesznek át a hangkivető főnevek, de kiegyenlítődésnek is, mert felfogható a folyamat úgy is, hogy a gyakoribb tővariáns terjed a paradigmán belül. Ez utóbbi annak fényében kevésbé tűnt valószínűnek, hogy a hangkivető szavak esetében összes alakjuknak (a képzetteket is beleszámítva) mintegy 49,7%-át[2] hangkivetéses alakok teszik ki példánygyakoriságuk alapján. A későbbi vizsgálatok során azonban megfigyelhettük, hogy a hangkivető főnevek esetében azoknál kisebb a hangkivetés mértéke releváns toldalékaiknál, amelyek összes alakjai közt a hangkivetéses alakok aránya alacsony. Ez az összefüggés azért jelentős, mert az összes alakok viszonylatában mérhető hangkivetéses alakok alacsonyabb aránya nem lehet kizárólagosan a szavak ingadozásának következménye. A 95-100%-ban hangkivető főneveknél (átlagos hangkivetés: 99,8%[3]) az összes alakból a

[1] Fontos azonban szem előtt tartani, hogy a változások hajlamosabbak is az informális rétegekből kiindulni, azaz ha ezeket keressük, akkor vélhetőleg informálisabb szövegekben találjuk meg őket.

[2] A nem képzett alakok esetében a hangkivetéses alakok aránya is közeli (44,58%).

[3] Adatok a *Szószablya Gyakorisági Szótár* alapján.

hangkivetésesek arányának átlaga 52,47%, míg a 90-95% hangkivetőknél (átlagos hangkivetés: 92,9%) ez csak 35,8%. Az összes alak vonatkozásában tapasztalható 16,87%-os aránykülönbségből 1,5%-nyi[1] betudható a releváns toldalékok hangkivetési mértékében megfigyelhető hangkivetési mértékkülönbségnek, de marad több mint 15%-nyi eltérésünk, amit ezzel nem lehet megmagyarázni. Ebből következik, hogy azok a szavak, amelyeknél a hangkivetéses alakok aránya összes alakjaikban kisebb, hajlamosabbak arra, hogy hangkivetésükben elbizonytalanodjanak, és részt vegyenek az analógiás kiegyenlítődésben. Az összes alak alapján mért hangkivetési mértéknek és a releváns toldalékos alakok hangkivetési mértékének ezen összefüggését az egyes tővégek alapján kialakított csoportoknál (utolsó és utolsó két fonéma, utolsó két mássalhangzó) is igazolni tudtam.

További, a hasonlóság által meghatározott jelenségeket tudtam megfigyelni a gráfok segítségével elvégzett elemzéseim során. 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 hangkivetésükben hasonlóan viselkedő szavakkal közösen hasonlósági csoportokba rendeződtek. Ebből azt látjuk, hogy az analógiás alapú regularizálódásban elsődleges szerepe lehet a saját viselkedési csoporttal való kapcsolatok meggyöngülésének (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, malom, alom, halom, gyalom*

[1] A hangkivetéssel együttjáró toldalékok hangkivetési mértékében bekövetkezett változásnak az összes alakban megfigyelhető hangkivetés arányára gyakorolt hatását úgy becsültem meg, hogy a releváns toldalékok hangkivetési mértékváltozását megszoroztam az összes alakok közt lévő releváns hangkivetéses alakok arányának a számával: (0,989-0,929)*0,21.

stb.), és morféma-alapon azt sem tudnánk magyarázni, hogy több nagyon hasonló, de némileg eltérő végű szó (*álom, ólom*) 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. Gyakorisági hatásokat tudtunk kimutatni az összetett szavak esetében is, amelyeknél azt tapasztaltuk, hogy az alapszótól eltérő viselkedésű összetett szavak az átlagosnál kisebb típus- és példánygyakoriságú összetett szóbokrokban találhatóak.

A *Szószablya Gyakorisági Szótár* és a *Google Gyakorisági Gyűjtés* összehasonlítása során a legfontosabb felismerésem az volt, hogy a **változásban a paradigmatikus cellák meglehetősen eltérően vesznek részt**. Ezzel érintőlegesen már Rebrus és Törkenczy (2008) is szembesített minket. Az egyes paradigmacellák viselkedése nem teljesen autonóm, hisz megfigyelhetők benne általános tendenciák, de a változás nem úgy megy végbe, hogy a szavak egyenletesen vagy akár hirtelen sorolódnak át egy másik paradigmába[1]. Az egyes paradigmacellák kisebb, mások magasabb hangkivetési mértékkel rendelkeznek, amelyek statisztikailag igazolhatók. Az egyes toldalékoknál a fonotaktikai, rendszerbeli vagy használati okokra visszavezethetően a hangkivetési mértékek és azok változásának tempója eltérnek, és ezek közt nincs egyértelmű összefüggés, hisz legalacsonyabb mértékben a szuperesszívusz esetén figyelhető meg a hangkivetés, de egyedül csak az E.3 birtokosra utaló alak változása vagy az informális regiszterekben való erősebb jelenléte igazolható szignifikáns módon. Azonban az egyes toldalékok közt felállítható hierarchia a hangkivetési mérték szerint számos egyedi esetben eltéréseket mutat, és a változásban részmintázatok is megfigyelhetők (pl. tárgyesetükben legkevésbé hangkivető, gyorsan változó szavak)

A szavak viselkedésének tanulmányozása során megfigyelt jelenségek alapján készített **algoritmusokról** a 6. fejezetben bebizonyosodott, hogy **jól képesek megragadni a hangkivető szavak formai jellegzetességeit,** és több tanuló algoritmusnál is sikeresebben osztályozták azokat (egyedül a maximum entrópia

[1] Vizsgálatom alapján ez még egyedi beszélőknél lehetséges lenne, hisz adataik keverednek a korpuszokban, de ez kevéssé valószínű.

modell hozott hasonlóan jó eredményt). A bináris döntési helyzetekben a *komplex jegymérték* és *egyszerű jegymérték* nevű hasonlítási módok (lásd 4.3. alfejezet) teljesítettek a legjobban, ami azt mutatja, hogy ha kategorikus döntéseket kell hoznunk szavak viselkedéséről, akkor elsősorban a végek hasonlóságára (nem azonosságra) hagyatkozunk. A szavak hangkivetési mértékében tapasztalható különbségeket azonban már a szerkezeti hasonlóságoknak nagyobb súlyt adó, így holisztikusabb *komplex tengelymérték* is jól tudta megragadni. Ez a hasonlítási mód a többinél jobban teljesített, ha csak néhány prototípus segítségével kellett modellezni a szavak hangkivetési mértékét. A szavak viselkedésében szerepet játszó **prototípusok kiválasztásában azonban legnagyobb szerepet a példánygyakoriság játszotta**. Ebben az esetben ismét a tővégek hasonlósága számított jobban.

Nyelvi tesztem a 7. fejezetben több olyan feltételezést is igazolt, amelyre már korábban is támaszkodtam elemzéseimben, és amelyeket Lukács (2002) is felvázolt iránymutató munkájában. Ezek közül a legfontosabb, hogy az **egyes fonémapozíciókban megfigyelhető hasonlóságok és eltérések különböző fontossággal bírnak szavak összevetése esetén**. Ezt abból láthattuk, hogy minél inkább balra változtattam meg egy fonémát egy *CVCVC* szerkezetű álszón belül, annál hasonlóbb volt viselkedése (hangkivetési mértéke) ahhoz a szóhoz, amelyből létrehoztam. A fonémapozíció hatásától függetlenül megfigyelhettük, hogy az eredeti szó befolyásán túl leginkább az új szó egyedi viszonyrendszerének van szerepe hangkivetési mértékének alakulásában. Azaz a résztvevők nem feltétlenül az eredeti szót használták analógiás forrásnak az álszó alakjainak kiválasztásában, hanem az álszót már önállóan értékelték, és ennek megfelelően határozták meg viselkedését. Ebben legnagyobb szerepe a szerkezetileg is hasonló szavaknak van, amit kiegészít a vége alapján legközelebbinek számító leghasonlóbb szó hatása is.

8.2. Korlátok a kutatásban

Kutatásomban néhány nehézségbe is ütköztem, amelyeket a további vizsgálódás során jobban figyelembe kell venni. Amennyiben olyan jelenségeket szeretnénk vizsgálni, amelyeket elsősorban vagy inkább **informális szövegekben** figyelhetünk meg, akkor még következetesebben **számolnunk kell az ékezetmentes alakokkal**, mivel más, nem odavaló szavak ékezetmentes alakjai adataink közé keveredhetnek. Ezeket ekkora minta esetén tudtam szűrni még kézileg is (bővebben 5.4.1 alfejezet), de hatékonyabb, gépi előszűrésük mindenképpen kívánatos későbbi kutatások esetén. Vizsgálatomat az is nehezítette, hogy bizonyos **alapfontosságú kutatások hiányoznak a magyar nyelvvel kapcsolatban**, így azokat részben nekem kellett pótolnom, vagy ha erre nem volt mód, akkor áthidaló megoldásokat kellett alkalmaznom. Mint a 8.1. alfejezetben láthattuk, ezek hiányában is több érdemleges felfedezést tudtam tenni, de meglétük esetén ezek köre még szélesebb lehetne. Más közepes méretű nyelvek kutatóinál előnyösebb helyzetben vagyunk, hogy rendelkezésünkre áll a nagyméretű *Szószablya Korpusz és Gyakorisági Szótár* (Halácsy és mtsai 2003), illetve a *morphdb.hu* (Trón és mtsai 2006), azonban szükséges lenne a korpusz újabb változatát elkészíteni, a *morphdb.hu* esetében pedig egy alapos nyelvészeti szempontú átnézés segítene az anyag ingadozó minőségének javításán. Mivel a *Szószablya Korpusz* esetében az anyagot csak gépileg szűrték, hiányzik a nyelvészeti kutatásokhoz egy olyan nagyméretű, informális szövegeket is tartalmazó korpusz, ahol az adatok minősége némileg megbízhatóbb, nem is beszélve egy hozzáférhető lejegyzett beszélt nyelvi korpuszról, amelynek hiánya komoly akadályt jelent a magyar nyelv tanulmányozásának dinamikusabb fejlődésében. Hiányossága minden nyelvi kutatásnak, így az enyémnek is, hogy elsősorban az írott nyelvet tanulmányozza (még ha informális szövegeket vizsgál is), ha nagyobb mennyiségű adatra kíván hagyatkozni megállapításaiban.

8.3. További kutatási lehetőségek

Kutatásom habár gazdagította a magyar hangkivető főnevekről és az analógia működéséről való ismereteinket, a megválaszolt kérdések mellett többet nyitva is hagyott. Igaz, hogy több esetben szignifikáns összefüggéseket is meg tudtunk figyelni a hangkivető főnevek viselkedésével kapcsolatban, de következő lépésként célszerű lenne a disszertációban bemutatott vizsgálatokat **olyan ingadozó jelenségeken is elvégezni, amelyek még nagyobb mennyiségű szót érintenek**, mint pl. a tárgyrag (kötőhang megléte és hiánya) vagy az E/T.3 birtokos (*-A(i)* vagy *-jA(i)* formában jelenik meg) ingadozása az összes főnévnél. Ezeknél a nagyobb adatmennyiség következtében statisztikai próbákkal könnyebben tudnánk megbízható összefüggéseket kimutatni, illetve a szórványos adathibáknak még kisebb jelentőségük lenne. E nyelvi jelenségek vizsgálata azért is lenne különösen gyümölcsöző, mert kevésbé kötődnek az informális szövegekhez, és kevésbé stigmatizáltak is, mint a hangkivető tövek nem hangkivető használata[1], így akár az *MNSZ* vagy *A mai magyar nyelv szépprózai gyakorisági szótára* (Füredi és Kelemen 1989) is bevonható lenne ezekbe. Ezzel párhuzamosan szükségszerű lenne újabb **kutatások** elvégzése azzal kapcsolatban, hogy **a fonémapozíciók eltérő jelentősége a magyar nyelv más szavainál** is érvényben van-e, illetve, hogy a magánhangzók és a mássalhangzók valóban eltérő módon esnek-e a latba a hasonlítások során.

Ezeknek a lehetséges kutatásoknak az eredményeit természetesen érdemes lenne beépíteni hasonlósági mértékeimbe, algoritmusaimba is, amelyeket azonban már a mostani vizsgálatok fényében is többféleképpen lehet és kell tökéletesíteni. A nyelvi teszt eredményei kapcsán legnyilvánvalóbban az látszott, hogy a komplex jegymérték és a komplex tengelymérték az analógiás hasonlításban eltérő tényezőket ragadnak meg. Ezért célszerű lenne egy **hibrid algoritmus** elkészítése, amely mind a kettő előnyeit ötvözi. Ezen túl az algoritmusok szerkezet iránt való érzékenységét lenne érdemes tovább növelni, és hasznos lenne a **hasonlításba a fonémáknál nagyobb**

[1] A köznyelvi beszélt kommunikációban valószínűleg senki sem kapja fel a fejét, ha egy *gyomort* vagy egy *fülesbagolyt* alakot hall, de ilyeneket leírni igényes szövegben a mai napig nem „illik".

részegységeket is bevonni (pl. magánhangzótól magánhangzóig terjedő mássalhangzóláncok). A hasonlítási algoritmusok kezelésre szoruló hibája, hogy a **hosszúságbeli eltéréseknek** nem tulajdonítanak elegendő jelentőséget. A probléma megoldására vélhetőleg be kell vezetni egy olyan korrekciós eljárást, amely súlyt ad a szótagszámnak, fonémaszámnak is. Természetesen ennek a beállítását nagy gondossággal kell elvégezni, hisz a 6.3 és 6.4. alfejezetek teszteléseiből rossz eredményei miatt kihagyott gráfalapú algoritmus (vö. 6.2. alfejezet) esetében is problémát okozott, hogy a hosszúságnak túlzott súlyt adott a hasonlításban.

Az elemzésekből láttuk, hogy a **fonémahasonlítás** módjának kisebb jelentősége van a szavak összehasonlításában, mint annak, hogy miképp választjuk ki a fonémákat összehasonlításra, illetve hogyan súlyozzuk az egyes fonémapozíciókat. Ennek ellenére célszerű lenne az algoritmusokat fejleszteni ebben a tekintetben is az elméleti korrektség jegyében, különösen, hogy az sem kizárt, hogy ezek kis mértékben hozhatnak további javulást is. Ennek keretében köztes kategóriák bevezetését kellene létrehozni bizonyos fonémákra (/j/ és /v/). A modellezést érdemes lenne olyan változatokkal is elvégezni, amelyek reprezentációi a jegygeometriai elképzeléseket pontosabban követik a jelenlegi implementációknál (Clements és Hume 1995).

Az algoritmusok teljesítményét a későbbiekben mindenképp érdemes lenne összevetni az AM-mel és a TiMBL-lel új vagy akár a már vizsgált jelenségek kapcsán is. Az algoritmusok kipróbálása más agglutinatív idegen nyelvi jelenségeken is gyümölcsöző lenne (pl. lengyel hangkivetés, Kraska-Szlenk (2007) alapján), így megfigyelhetnénk, hogy univerzálisnak vélt elveiknek köszönhetően más nyelveknél is megállják-e a helyüket. A 6.3. alfejezetben utaltam arra, hogy az algoritmusok **szótárbővítésre** is lényegében módosítások nélkül alkalmasak lennének, és további bővítésükkel akár **morfológiai elemzésre** is használhatóvá válnának (Stroppa és Yvon 2005).

Függelék

A függelékek elérhetőek a

https://docs.google.com/viewer?a=v&pid=explorer&chrome=true&srcid=0BxQBWau_71U7Y2E3MGUzN2QtY2I2OS00Mzc1LThjZWUtZmI1YjU4N2ZkMzBm&hl=hu

webcím alatt.

Irodalomjegyzék

Abelin, Asa 1999. *Studies in sound symbolism.* Göteborg University. (Phd-disszertáció)

Ackerman, Farrell–Blevins, James P.–Malouf, Robert 2009. Parts and wholes: Implicative patterns in inflectional paradigms. In: Blevins, James–Blevins, Juliette (szerk.): *Analogy in Grammar: Form and Acquisition.* Oxford University Press. Oxford. 54–82.

Aha, David W.–Kibler, Dennis–Albert, Marc K. 1991. Instance-based learning algorithms. *Machine Learning* 6: 37–66.

Albright, Adam 2009. Modelling analogy as probabilistic grammar. In: Blevins, James–Blevins, Juliette (szerk.): *Analogy in Grammar: Form and Acquisition.* Oxford University Press. Oxford. 185–213.

Albright, Adam–Hayes, Bruce 2002. Modeling English past tense intuitions with minimal generalization. In: Maxwell, Michael (szerk.): *SIGPHON 6: Proceedings of the Sixth Meeting of the ACL Special Interest Group in Computational Phonology. ACL.* 58–69.

Anttila, Raimo 1977. *Analogy*. Mouton de Gruyter. The Hague.

Anttila, Raimo 1989. *Historical and Comparative Linguistics.* John Benjamins. Amsterdam.

Anttila, Raimo 2002. Variation and Phonological Theory. In: Chambers, Jack–Trudgill, Peter–Schilling-Estes, Natalie (szerk.): *The Handbook of Language Variation and Change.* Blackwell. Oxford. 206–243.

Arisztotelész 1963. *Poétika.* Magyar Helikon. Budapest.

Aronoff, Mark 1994. *Morphology by itself.* MIT Press. Cambridge, MA.

Baayen, R. Harald–McQueen, James M.–Dijkstra, Ton–Schreuder, Robert 2003. Dutch inflectional morphology in spoken- and written-word recognition. In: Baayen, R. H.–Schreuder, R. (szerk.): *Morphological Structure in Language Processing*. Mouton de Gruyter. Berlin.

Barabási, Albert László 2002. *Linked. The New Science of Networks.* Perseus. Cambridge, MA.

Bárczi Géza–Benkő Loránd–Berrár Jolán 1967. *A magyar nyelv története.* Tankönyvkiadó. Budapest.

Baroni, Marco–Ueyama, Motoko 2006. Building general- and special-purpose corpora by Web crawling. *Proceedings of the 13th NIJL International Symposium, Language Corpora: Their Compilation and Application*. 31–40.

Baudouin de Courtenay, Jan Niecisław 1974. Szkice językoznawcze [Nyelvészeti tanulmányok]. In: Baudouin de Courtenay. *Dzieła wybrane vol. 1.* [Válogatott művei 1. kötet] Państwowe Wydawnictwo Naukowe. Varsó. 145–616.

Bécsi Kódex 1916. Új Nyelvemléktár. Mészöly Gedeon. Budapest.

Benua, Laura 1995. Identity effects in morphological truncation. In: Beckman, J. N.–Walsh Dickey, L.–Urbanczyk, S. (szerk.): *Papers in Optimality Theory*. GLSA. Amherst, MA. 77–136.

Benua, Laura 1997. *Transderivational identity: Phonological relations between words.* (PhD-disszertáció). http://roa.rutgers.edu/files/259-0498/roa-259-benua-2.pdf (2010.07.01.)

Bergen, Benjamin K. 2004. The psychological reality of phonaesthemes. *Language* 80: 290–311.

Bird, Steven–Ellison, T. Mark 1994. One-Level Phonology: Autosegmental Representations and Rules as Finite Automata. *Computational Linguistics* 20: 55–90.

Bíró Tamás 2006. *Finding the Right Words: Implementing Optimality Theory with Simulated Annealing*. Groningen Dissertations in Linguistics. Groningen.

Blevins, James P. 2001. Morphological paradigms. *Transactions of the Philological Society*. 99: 207–210.

Blevins James P.–Blevins, Juliette (szerk.) 2009a. *Analogy in Grammar: Form and Acquisition.* Oxford University Press. Oxford.

Blevins James P.– Blevins, Juliette 2009b. Introduction: Analogy in Grammar. In: Blevins, James–Blevins, Juliette (szerk.): *Analogy in Grammar: Form and Acquisition.* Oxford University Press. Oxford. 1–12.

Bloomfield, Leonard 1933. *Language.* Holt. New York.

Blyth, Thomas Scott 2005. *Lattices and Ordered Algebraic Structures*. Springer-Verlag. London.

Boersma, Paul–Hayes, Bruce 2001. Empirical tests of the gradual learning algorithm. *Linguistic Inquiry* 32. 1: 45–86.

Bohas, Georges–Guillaume, Jean-Patrick–Kouloughli, Djamel Eddine 1990. *The Arabic Linguistic Tradition. Arabic Thought and Culture.* Routledge. London

Bolinger, Dwight 1961. Syntactic blends and other matters. *Language* 37: 366–381.

Bozotvágó késsel ölt: mindkét áldozat meghalt 2009. http://www.kisalfold.hu/belfold_hirek/bozotvago_kessel_olt_mindket_aldozat_meghalt/2131201/ (2010.07.01.)

Buiskool, Herman E. 1939. *The Tripdai. Being an abridged English recast of Purvatrdsidaham*. Brill. Leiden

Bybee, Joan L. 1985. *Morphology: A study of the relation between meaning and form.* John Benjamins. Amsterdam.

Bybee, Joan L. 2000. The phonology of the lexicon: Evidence from lexical diffusion. In: Barlow, Michael–Kemmer, Suzanne (szerk.): *Usage-based models of language*. CSLI. Stanford CA . 65–85.

Bybee, Joan L. 2001. *Phonology and Language Use.* Cambridge University Press. Cambridge.

Bybee, Joan L. 2002. Word frequency and context use in the lexical diffusion of phonetically conditioned sound change. *Language Variation and Change 14*: 261–290.

Bybee, Joan L. 2006. From usage to grammar: the mind's response to repetition. *Language* 82 711–733.

Bybee, Joan L. 2007. *Frequency of use and the organization of language.* Oxford University Press. Oxford.

Bybee, Joan L. 2010. *Language, Usage and Cognition.* Cambridge University Press. Cambridge.

Bybee, Joan L.–Eddington, David 2006. A usage-based approach to Spanish verbs of 'becoming.' *Language 82* 323–355.

Bybee, Joan L.–Moder, Carol Lynn 1983. Morphological classes as natural categories. *Language* 59 251–270.

Bybee, Joan L.–Slobin, Dan I. 1982. Rules and schemas in the development and use of the English past tense. *Language* 58 265–289.

Carstairs-McCarthy, Andrew 1992. *Current Morphology*. Routledge. London.

Casenheiser, Devin–Goldberg, Adele E. 2005. Fast mapping of a phrasal form and meaning. *Developmental Science 8:* 500–508.

Chandler, Steve 2002. Skousen's analogical approach as an exemplar-based model of categorization. In: Skousen, Royal–Lonsdale, Deryle– Parkinson, Dilworth B. (szerk.): *Analogical Modeling. An exemplar-based approach to language.* John Benjamins. Amsterdam. 51–105.

Chomsky, Noam 1965. *Aspects of the theory of syntax.* The MIT Press. Cambridge, MA.

Chomsky, Noam 1968. *Language and mind.* Harcourt. New York.

Chomsky, Noam 1975. *The logical structure of linguistic theory.* The Plenum Press. New York.

Chomsky, Noam 1986. *Knowledge of language.* Praeger. New York.

Chomsky, Noam–Halle, Morris 1968. *The Sound Pattern of English.* Harper and Row. New York.

Clements G. N.–Hume, Elizabeth V. 1995. The Internal Organization of Speech Sounds. In: Goldsmith, John. A. (szerk.): *The Handbook of Phonological Theory.* Blackwell. Oxford. 245–306.

Coetzee, Andries W. 2004. What it Means to be a Loser: Non-optimal Candidates in Optimality Theory. (PhD-disszertáció). http://roa.rutgers.edu/files/687-0904/687-0904-9-0.PDF (2010.07.01.)

Cover, Thomas M.–Hart, Peter E. 1967. Nearest neighbor pattern classification. *Institute of Electrical and Electronics Engineers Transactions on Information Theory 13.* 21–27.

Cutler, Anne 1984. Stress and accent in language production and understanding. In: Gibbon, Dafydd–Richter, Helmut (szerk.): *Intonation, Accent and Rhythm*. Walter de Guyter. Berlin. 77–90.

Csánki Dezső 1890. *Magyarország történelmi földrajza a Hunyadiak korában I.* Magyar Tudományos Akadémia. Budapest.

Cser, András 2000. Phonological models of sonority. In: Varga László (szerk.): *The even yearbook. Working Papers in Linguistics.* ELTE. Budapest. 1–18.

Dąbrowska, Ewa–Lieven, Elena 2005. Towards a lexically specific grammar of children's question constructions. *Cognitive Linguistics* 16: 437–74.

Daelemans, Walter 2002. A comparison of Analogical Modeling to Memory-Based Language Processing. In: Skousen, Royal–Lonsdale, Deryle– Parkinson, Dilworth B. (szerk.): *Analogical Modeling. An exemplar-based approach to language.* John Benjamins. Amsterdam. 157–179.

Daelemans, Walter–Gillis, Steven–Durieux, Gert 1994. The acquisition of stress: a data-oriented approach. *Computational Linguistics* 20: 421–451.

Daelemans, Walter–van den Bosch, Antal–Zavrel, Jakub 1999. Forgetting exceptions is harmful in language learning. *Machine Learning: Special issue on Natural Language Learning* 34: 11–41.

Daelemans, Walter–van den Bosch, Antal 2005. *Memory-Based Language Processing*. Cambridge University Press. Cambridge.

Darroch J.–Ratcliff. D. 1972. Generalized iterative scaling for log-linear models. *Ann. Math. Statistics, 43*: 1470–1480.

Dehaene Stanislas 2003. The neural basis of the Weber–Fechner law: a logarithmic mental number line. *Trends in Cognitive Sciences.* 145–147.

Della Pietra, Stephen–Della Pietra, Vincent–Lafferty, John 1997. Inducing features of random fields. *IEEE Transactions on Pattern Analysis and Machine Intelligence, 19:* 380–393.

Derwing, Bruce–Royal Skousen 1994. Productivity and the English past tense: Testing Skousen's analogy model. In: Lima, S. D.–Corrigan, R. L.–Iverson, G. K. (szerk.): *The Reality of Linguistic Rules*. John Benjamins. Amsterdam. 193–218.

Dice, Lee R. 1945. Measures of the Amount of Ecologic Association Between Species. *Ecology 26 (3)*. 297–302.

Dinneen, Francis P. 1995. *General Linguistics.* Georgetown University Press. Washington.

Domingos, Pedro 1995. *The RISE 2.0 system: A case study in multistrategy learning. Technical Report 95-2.* University of California at Irvine, Department of Information and Computer Science. Irvine, CA.

Domingos, Pedro 1996. Unifying instance-based and rule-based induction. *Machine Learning* 24: 141–168.

Dudani, Sahibsingh A. 1976. The distance-weighted k-nearest neighbor rule. *IEEE Transactions on Systems, Man, and Cybernetics 6:* 325–327.

Duvignau, Karine–Gaume, Bruno 2003. Linguistic, Psycholinguistic and Computational approaches to the lexicon: For early verb-learning based on analogy. *Journal of the European Society for the Study of Cognitive Systems* 6(1).

Eddington, David 1996. The psychological status of phonological analyses. *Linguistica* 36: 17–37.

Eddington, David 2002. A comparison of two analogical models: Tilburg Memory-Based Learner versus Analogical Modeling. In: Skousen, Royal–Lonsdale, Deryle– Parkinson, Dilworth B. (szerk.): *Analogical Modeling. An exemplar-based approach to language.* John Benjamins. Amsterdam. 141–155.

Eddington, David 2003. Issues in modeling language processing analogically. *Lingua. 114.* 849–871.

Eddington, David 2006. Paradigm Uniformity and Analogy: The Capitalistic versus Militaristic Debate. *International Journal of English Studies* 6: 1–18.

Elekfi László 1994. *Magyar ragozási szótár. Dictionary of Hungarian Inflections.* MTA Nyelvtudományi Intézet. Budapest.

Elekfi László 2000. Homonimák felismerhetősége toldalékos alakok alapján. *Nyr.* 124: 146–63.

Enfield, N. J. 2008. Transmission biases in linguistic epidemiology. *Journal of Language Contact* 2: 295–306.

Ernestus, Mirjam–Baayen, Harald 2004. Analogical effects in regular past tense production in Dutch. *Linguistics* 42: 873–903.

Evans, Jonathan St. B. T. 1982. *The psychology of deductive reasoning.* Routledge. London.

Fillmore, Charles J.–Kay, Paul. 1987. *The goals of Construction Grammar. Berkeley Cognitive Science Program Working Paper 50*. University of California at Berkeley. Berkeley, CA.

Finkel, Raphael–Stump, Gergory 2009. Principal parts and degrees of paradigmatic transparency. In: Blevins, James–Blevins, Juliette (szerk.): *Analogy in Grammar: Form and Acquisition.* Oxford University Press. Oxford. 13–53.

Firth, John 1930. *Speech*. Oxford University Press. London.

Fix, E.–Hodges, J. L. 1951. *Disciminatory analysis—nonparametric discrimination; consistency properties (Technical Report Project 21-49-004, Report No. 4).* USAF School of Aviation Medicine.

Fowler, Carol A.–Housum, Jonathan 1987. Talkers' signaling of "new" and "old" words in speech and listeners' perception and use of the distinction. *Journal of Memory and Language 26:* 489–504.

Frisch, Stefan A. 1996. *Similarity and Frequency in Phonology.* (PhD-disszertáció) http://www.cas.usf.edu/~frisch/Frisch96.pdf (2010.07.01.)

Frisch, Stefan A.–Pierrehumbert, Janet B.–Broe, Michael B. 2004. Similarity avoidance and the OCP. *Natural Language and Linguistic Theory* 22: 179–228.

Fűköh Borbála–Rung András 2005. Az –esz és az –er végű becézett szóalakokról. *Nyelvtudomány* I. 115–130.

Füredi Mihály–Kelemen József 1989. *A mai magyar nyelv szépprózai gyakorisági szótára.* Akadémiai Kiadó. Budapest.

Gardner, Howard 1985. *The mind's new science.* Basic Books. New York.

Gentner, Dedre 1989. The mechanisms of analogical learning. In: Vosniadou, Stella–Ortony, Andrew (szerk.): *Similarity and analogical reasoning.* Cambridge University Press. Cambridge. 199–241.

Gerken, LouAnn–Wilson, Rachel–Gómez, Rebeca–Nurmsoo, Erika 2009. The relation between linguistic analogies and lexical categories. In: Blevins, James–Blevins, Juliette (szerk.): *Analogy in Grammar: Form and Acquisition.* Oxford University Press. Oxford. 101–117.

Givón, Talmy 1984. *Syntax: a Functional-Typological Introduction, Volume 1.* John Benjamins. Amsterdam.

Givón, Talmy 1995. *Functionalism and grammar.* John Benjamins. Amsterdam.

Goldberg, Adele 1995. *Constructions. A Construction Grammar approach to argument structure*. University of Chicago Press. Chicago.

Goldberg, Adele 2006. *Constructions at Work. The Nature of Generalization in Language.* Oxford University Press. Oxford.

Goldinger, Stephen 1996. Word and voices: episodic traces in spoken word identification and recognition memory. *Journal of Experimental Psychology 22.* 1166–1183.

Goldsmith, John 1990. *Autosegmental and metrical phonology*. Basil Blackwell. Oxford.

Goldsmith, John 2009. Morphological analogy: Only a beginning. In: Blevins, James–Blevins, Juliette (szerk.): *Analogy in Grammar: Form and Acquisition.* Oxford University Press. Oxford. 137–163.

Halácsy Péter–Kornai András–Németh László–Rung András–Szakadát István–Trón Viktor 2003. A Szószablya projekt. In: Alexin Zoltán–Csendes Dóra (szerk.): *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003).* 299.

Halácsy Péter–Kornai András–Varga Dániel 2005. Morfológiai egyértelműsítés maximum entrópia módszerrel. In: Alexin Zoltán–Vincze Veronika (szerk.): *III. magyar számítógépes nyelvészeti konferencia. MSZNY 2005.* Szegedi Tudományegyetem. Szeged. 180–188.

Hale, Mark–Kissock, Madelyn–Reiss, Charles 1998. Output-Output Correspondence in Optimality Theory. In: Curtis, J. Lyle–Webster, G. (szerk.): *Proceedings of WCCFL 18.* 223–236.

Halford, Graeme S.–Andrews, Glenda 2007. Domain general processes in higher cognition: Analogical reasoning, schema induction and capacity limitations. In: Roberts, M. J. (szerk.): *Integrating the Mind: Domain General versus Domain Specific Processes in Higher Cognition*. Psychology Press. New York. 213–232.

Hallan, Naomi 2001. Paths to prepositions? A corpus-based study of the acquisition of a lexico-grammatical category. In: Bybee, Joan–Hopper, Paul (szerk.): *Frequency and the emergence of linguistic structure.* John Benjamins. Amsterdam. 91–121.

Halliday, Michael A. K. 1961. Categories of the theory of grammar. *Word 17.* 241–292.

Hare, Mary L.–Ford, Michael–Marslen-Wilson, William D. 2001. Ambiguity and frequency effects in regular verb inflection. In: Bybee, Joan–Hopper, Paul (szerk.): *Frequency and the emergence of linguistic structure.* John Benjamins. Amsterdam. 181–200.

Harris, Zellig 1961. *Structural linguistics.* University of Chicago Press. Chicago.

Hay, Jennifer 2001. Lexical frequency in morphology: is everything relative? *Linguistics* 39: 1040–1070.

Hay, Jennifer 2002. From speech perception to morphology: affix-ordering revisited. *Language* 78: 527–555.

Hayes, Bruce–Zuraw, Kie–Siptár Péter–Londe Zsuzsa 2009. Natural and unnatural constraints in Hungarian vowel harmony. *Language 85:* 822–863.

Hintzman, Douglas L. 1986. 'Schema abstraction' in a multiple-trace memory model. *Psychological Review* 94: 411–428.

Hintzman, Douglas L. 1988. Judgments of frequency and recognition memory in a multiple-trace memory model. *Psychological Review* 4: 528–551.

Hock, Hans Heinrich 2003. Analogical change. In: Josephs B.–Janda, R. (szerk.): *The Handbook of Historical Linguistics*. Oxford University Press. Oxford. 441–480.

Hockett, Charles F. 1966. *The state of the art.* Mouton. The Hague.

Householder, Fred W. 1971. *Linguistic speculations*. Cambridge University Press. Cambridge.

Højrup, Thomas 1983. The concept of life-mode: A form-specifying mode of analysis applied to contemporary Western Europe. *Ethnologia Scandinavica* 1–50.

Hutchins, Sharon Suzanne 1998. *The psychological reality, variability, and compositionality of English phonesthemes.* Atlanta: Emory University. (PhD-disszertáció)

Itkonen, Esa 1983. *Causality in linguistic theory: A critical inquiry into the methodological and philosophical foundations of 'non-autonomous' linguistics.* Croom Helm. London.

Itkonen, Esa 2005. *Analogy As Structure And Process.* John Benjamins. Amsterdam.

Jackendoff, Ray 2002. *Foundations of language.* Oxford University Press. Oxford.

Jakubovich Emil–Pais Dezső 1929. *Ó-magyar olvasókönyv.* Danubia. Pécs.

Jespersen, Otto 1942. *A modern English Grammar on historical principles, Part VI: Morphology.* Allen & Unwin. London.

Jespersen, Otto 1965. *Philosophy of grammar.* Allen & Unwin. London.

Johnson, Lawrence 1976. A rate of change index for language. *Language in Society* 165–172.

Jurafsky, Daniel 2003. Probabilistic Modeling in Psycholinguistics: Linguistic Comprehension and Production. In: Bod, Rens–Hay, Jennifer–Jannedy, Stefanie (szerk.): *Probabilistic Linguistics*. MIT Press. Cambridge, MA. 1–50.

Jurafsky, Daniel–Bell, Alan–Gregory, Michelle–Raymond, William D. 2001. Probabilistic Relations between Words: Evidence from Reduction in Lexical Production. In: Bybee, Joan–Hopper, Paul (szerk.): *Frequency and the emergence of linguistic structure.* John Benjamins. Amsterdam. 229–254.

Kac, Michael B. 1974. Autonomous linguistics and psycholinguistics. *Minnesota Working Papers in Linguistics and Philosophy of Language* 2: 42–47.

Kálmán László 2008. *A mögöttes és ami mögötte van.* http://www.szv.hu/cikkek/a-mogottes-es-ami-mogotte-van (2010.07.01.)

Kálmán László 2009. *Tudományos fantasztikum-e a mentális nyelvmodell?* (előadás) MSZNY 2009. Szeged. 2009. december 3.

Kálmán, László 2010a. *Analogy in semantics.* (megjelenés előtt Ruzsa Imre emlékkötetben)

Kálmán László 2010b. *Analógiás tanulás asszociatív memóriamodellel.* (kézirat)

Kálmán, László–Rebrus, Péter–Törkenczy, Miklós 2005. *Hungarian linking vowels: An analogy-based approach.* (poszter) 2nd Old World Conference in Phonology (OCP2). University of Tromsø. Center for Advanced Study in Theoretical Linguistics (CASTL). Tromsø. 2005. január 20–22.

Kálmán László–Rebrus Péter–Törkenczy Miklós 2010. Lehet-e az analógiás nyelvelmélet szinkrón? A magyar nyelvészeti kutatások újabb eredményei II., Kolozsvár. 2010. április 16.
http://budling.nytud.hu/~tork/KRT/bbte10_slides_print.pdf (2010.07.01.)

Kiefer Ferenc (szerk.) 1994. *Strukturális magyar nyelvtan 2. Fonológia.* Akadémiai Kiadó. Budapest.

Kiefer Ferenc 2002. Szabályszerűség, termékenység és analógia a morfológiában. In: Maleczki Márta (szerk.): *A mai magyar nyelv leírásának újabb módszerei 5.* 9–15.

Kilgarriff, Adam 2007. Googleology is Bad Science. *Computational Linguistics* 33 (1): 147–151.

King, Robert D. 1969. *Historical linguistics and generative grammar.* Prentice-Hall Inc. Englewood Cliffs, New Jersey.

Kiparsky, Paul 1972. Explanation in Phonology. In: Kiparsky, Paul 1982. *Explanation in Phonology.* Foris. Dodrecht. 81–119.

Kiparsky, Paul 1974. *Remarks on analogical change.* In: Anderson, John M.– Jones, Charles (szerk.): *Historical linguistics II.* North-Holland. Amsterdam. 257–275.

Kiparsky, Paul 1975. What are phonological theories about? In: Cohen, David–Wirth, Jessica R. (szerk.). *Testing linguistic hypotheses.* Hemisphere Publishing Corporation. Washington. 187–209.

Kiparsky, Paul 1981. Remarks on the metrical structure of the syllable. In: Dressler, Wolfgang U.–Pfeiffer, Oskar E.–Rennison, John R. (szerk.): *Phonologica 1980. Proceedings of the 4th International Phonology Meeting.* Innsbrucker Beiträge zur Sprachwissenschaft. Innsbruck. 245–256.

Kiparsky, Paul 1991. Economy and the construction of the Sivasutras. In: Deshpande, M. M.–Bhate, S. (szerk.): *Paninian Studies*. Ann Arbor. Michigan.

Kiparsky, Paul 1992. Analogy. In: Bright, William (szerk.): *International encyclopedia of linguistics, Vol. 1.* Oxford University Press. Oxford. 56–61.

Kiparsky, Paul 2000. Analogy as Optimization. In: Lahiri, Aditi (szerk.): *Analogy, Levelling, Markedness. Principles of Change in Phonology and Morphology.* Mouton de Gruyter. Berlin.

Kiparsky, Paul 2005. *Grammaticalization as optimization.* http://www.stanford.edu/~kiparsky/Papers/yalegrammaticalization.pdf (2010.07.01.)

Klein, Martin–Nelson, Michael L. 2009. Correlation of Term Count and Document Frequency for Google N-Grams. In: Boughanem, Mohand–Berrut, Catherine–Mothe, Josiane–Soulé-Dupuy, Chantal (szerk.): *Proceedings of the 31th European Conference on IR Research on Advances in Information Retrieval*. Springer. Toulouse. 620–627.

Knuth, Donald 1997. *The Art of Computer Programming, Volume 3: Sorting and Searching*. Addison-Wesley. Reading, MA.

Kondrak, Grzegorz–Sherif, Tarek 2006. Evaluation of Several Phonetic Similarity Algorithms on the Task of Cognate Identification. In: *Proceedings of the COLING-ACL Workshop on Linguistic Distances.* 43–50.

Kornai, András 1990. The Sonority Hierarchy in Hungarian. *Nyelvtudományi Közlemények* 91: 139–146.

Kornai, András 1993. The generative power of feature geometry. *Annals of Mathematics and Artificial Intelligence 8.* 37–46.

Kornai, András 1995. *Formal Phonology.* Garland Publishing. New York.

Kornai, András–Halácsy, Péter 2008. Google for the linguist on a budget. In: Evert, S.–Kilgarriff, A.–Sharoff, S. (szerk.): *Proceedings of 4th Web as Corpus Workshop (WAC-4).* Marracech. 8–11.

Kraska-Szlenk, Iwona 2007. *Analogy. The Relation between Lexicon and Grammar.* Lincom. München.

Kroesch, Samuel 1926. Analogy as a factor in semantic change. *Language* 2: 35–45.

Krott, Andrea 2009. The role of analogy for compound words. In: Blevins, James–Blevins, Juliette (szerk.): *Analogy in Grammar: Form and Acquisition.* Oxford University Press. Oxford. 118–136.

Krott, Andrea–Schreuder, Robert–Baayen, Harald 2002. Exemplar-based modeling of linkers in Dutch noun-noun compounds. In: Skousen, Royal–Lonsdale, Deryle– Parkinson, Dilworth B. (szerk.): *Analogical Modeling. An exemplar-based approach to language.* John Benjamins. Amsterdam. 181–206.

Labov, William 1994. *Principles of linguistic change: internal factors.* Blackwell. Oxford.

Ladefoged, Peter 1970. The measurement of phonetic similarity. *Statistical Methods in Linguistics* 6. 23–32.

Lakoff, George 1987. *Women, fire, and dangerous things.* University of Chicago Press. Chicago.

Langacker, Ronald 1987. *Foundations of cognitive grammar: theoretical prerequisites. Vol.I.* Stanford University Press. Stanford, CA.

Langacker, Ronald 1991. *Foundations of cognitive grammar. Vol. II: Descriptive applications.* Stanford University Press. Stanford, CA.

Lamond, Grant 2006. *Precedent and Analogy in Legal Reasoning.* http://plato.stanford.edu/entries/legal-reas-prec/ (2010.07.01.)

Levenshtein, V. I. 1966. Binary codes capable of correcting deletions, insertions, and reversals. *Soviet Physics - Doklady* 10(8): 707–710.

Lieven, Elena–Behrens, Heike–Speares, Jennifer–Tomasello, Michael 2003. Early syntactic creativity: a usage-based approach. *Journal of Child Language* 30: 333–70.

Long, Christopher J.–Almor, Amit 2000. Irregularization: The interaction of item frequency and phonological interference in regular past tense production. In: *Proceedings of the Twenty-Second Annual Conference of the Cognitive Science Society.* Lawrence Erlbaum. Hillsdale, NJ. 310–315.

Lukács Ágnes 2002. *Alaktanilag kivételes tövek vizsgálata a magyarban. A leíró általánosítások mentális realitása.* (szakdolgozat)

MacDonald, M. C. 1994. Probabilistic constraints and syntactic ambiguity resolution. *Language and Cognitive Processes* 9: 157–201.

Magnus, Margaret 2000. *What's in a word? Evidence for phonosemantics.* University of Trondheim. (Phd-disszertáció)

Mańczak, Witold 1980. Laws of analogy. In: Fisiak, J. (szerk.): Historical morphology .Mouton. The Hague.

Mátyus István 1766. *Diaetetica. II.* Kolozsvár.

Mihăltz, Márton–Prószéky, Gábor 2004. Results and Evaluation of Hungarian Nominal WordNet v1.0. In: Sojka, Peter– Pala, Karel– Smrz, Pavel– Fellbaum, Christiane– Vossen, Piek (szerk.): *Proceedings of the 2nd International WordNet Conference*. Global WordNet Association. Brno. 175–180.

Milroy, James 1992. *Linguistic Variation and Change.* Blackwell. Oxford.

Milroy, James 1993. On the social origins of language change. In: Jones, Charles (szerk.): *Historical Linguistics: Problems and Perspectives*. Longman. London. 215–236.

Milroy, James 1997. Internal vs external motivations for linguistic change. *Multilingua 311−323.*

Milroy, James–Milroy, Lesley 1992. Social network and social class: Toward an integrated sociolinguistic model. *Language in Society* 21*:* 1–26.

***MNSZ* 2006.** *http://corpus.nytud.hu/mnsz/index_hun.html* (2010.07.01.)

Mudrow, Mike 2002. Version spaces, neural networks, and Analogical Modeling. In: Skousen, Royal–Lonsdale, Deryle– Parkinson, Dilworth B. (szerk.): *Analogical Modeling. An exemplar-based approach to language.* John Benjamins. Amsterdam. 225–264.

Mukherjee, Joybrato 2007. Corpus linguistics and linguistic theory: general nouns and general issues. *International Journal of Corpus Linguistics* 12/1: 131–147.

Myers, James 2002. Exemplar-driven analogy in Optimality Theory. In: Skousen, Royal–Lonsdale, Deryle– Parkinson, Dilworth B. (szerk.): *Analogical Modeling. An exemplar-based approach to language.* John Benjamins. Amsterdam. 265–300.

Nagy, Naomi–Reynolds, Bill 1997. Optimality theory and variable word-final deletion in Faetar. *Language Variation and Change* 9, 1: 37–56.

Nakov, Preslav–Hearst, Marti 2005. A Study of Using Search Engine Page Hits as a Proxy for n-gram Frequencies. In: *Proceedings of the Conference on Recent Advances in Natural Language Processing (RANLP'05)*. Borovets. 347–353.

Neetle, Daniel 1999. Is the rate of linguistic change constant? *Lingua 108.* 119–136.

Nosofsky, Robert M. 1988. Exemplar based accounts of relations between classification, recognition, and typicality. *Journal of Experimental Psychology: Learning, Memory, and Cognition* 14: 700–708.

Ohala, John J. 1974. Experimental historical phonology. In: Anderson, J. M.–Jones, C. (szerk.): Historical linguistics. Proceedings of the first international conference on historical linguistics. North-Holland. Amsterdam. 353–389.

Oravecz Csaba–Sass Bálint–Simon Eszter 2009. Gépi tanulási módszerek ómagyar kori szövegek normalizálására. In: Tanács Attila–Szauter Dóra–Vincze Veronika (szerk.): *VI. magyar számítógépes nyelvészeti konferencia. MSZNY 2009.* Szegedi Tudományegyetem. Szeged. 317–324.

Papp Ferenc 1975. *A magyar főnév paradigmatikus rendszere.* Akadémiai Kiadó. Budapest.

Parker, Steve 2002. *Quantifying the sonority hierarchy.* (PhD-disszertáció)

Penn, Derek C.–Holyoak, Keith J.–Povinelli, Daniel J. 2008. Darwin's mistake: Explaining the discontinuity between human and nonhuman minds. *Behavioral and Brain Sciences* 31: 109–78.

Phillips, Betty. S. 2006 *Word frequency and lexical diffusion.* Palgrave. Macmillan. New York.

Pierrehumbert, Janet 2001. Exemplar dynamics: Word frequency, lenition, and contrast. In: Bybee, J.–Hopper, P. (szerk.): *Frequency effects and the emergence of lexical structure.* John Benjamins. Amsterdam. 137–157.

Pierrehumbert, Janet 2002. Word-specific phonetics . *Laboratory Phonology VII*. Mouton de Gruyter. Berlin, 101–139.

Pinker, Stephen 1994. *The language instinct.* New York: Morrow.

Pinker, Stephen 1999. *Words and Rules*. New York. Basic Books.

Posner, Michael I.–Keele, Steven W. 1968. On the genesis of abstract ideas. *Journal of Experimental Psychology* 77: 353–363.

Prince, Alan–Smolensky, Paul 1993. *Optimality Theory: Constraint interaction in generative grammar.* Rutgers University Center for Cognitive Science Report No. RuCCS-TR-2. New Brunswick, NJ.

Prószéky Gábor 2000. Számítógépes morfológia. In: Kiefer Ferenc (szerk.): *Strukturális magyar nyelvtan 3. Morfológia.* Akadémiai Kiadó. Budapest. 1021–1064.

Prószéky, Gábor–Kis, Balázs 1999. Morpho-syntactic Parsing of Agglutinative and Other (Highly) Inflectional Languages. In: *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics*. College Park. Maryland. 1999, 261–268.

Quinlan, J. 1993. *C4.5: Programs for machine learning.* Morgan Kaufmann. San Mateo, CA.

Ratnaparkhi, Adwait 1996. A Maximum Entropy Model for Part-Of-Speech Tagging. In: *Proceedings of the Empirical Methods in Natural Language Processing Conference (EMNLP).* University of Pennsylvania.

Rebrus Péter 2000. Morfofonológiai jelenségek. In: Kiefer Ferenc (szerk.): *Strukturális magyar nyelvtan 3. Morfológia.* Akadémiai Kiadó. Budapest. 763–947.

Rebrus Péter–Törkenczy Miklós 2008. *Morfofonológia és a lexikon.* In: Kiefer Ferenc (szerk.): *Strukturális Magyar Nyelvtan 4. A szótár szerkezete.* Akadémiai kiadó. Budapest. 683–786.

Rebrus Péter–Trón Viktor 2003. Fonetikai motiváció a fonológiai mintázatokban. In: Hunyadi László (szerk.): *Kísérleti Fonetika Laboratóriumi Fonológia a Gyakorlatban.* Debreceni Egyetem Kossuth Egyetemi Kiadója. Debrecen. 139–164.

Recski Gábor 2010. Főnévi csoportok azonosítása szabályalapú és hibrid módszerekkel. In: Tanács Attila–Csendes Dóra (szerk.): *Magyar Számítógépes Nyelvészeti Konferencia. MSZNY 2010.* 333–341.

Rosch, Eleanor 1973. On the internal structure of perceptual and semantic categories. In: Moore, T. E. (szerk.): *Cognitive development and the acquisition of language.* Academic Press. New York. 111–144.

Rung, András 2008. Determining word similarity in the Hungarian language. In: Kálmán László (szerk.): *Papers from the Mókus Conference.* Tinta Kiadó. Budapest. 112–118.

Rung András 2009. Szóhasonlóság mérése analógiás megközelítésben. In: Tanács Attila–Szauter Dóra–Vincze Veronika (szerk.): VI. magyar számítógépes nyelvészeti konferencia. MSZNY 2009. Szegedi Tudományegyetem. Szeged. 104–113.

Rytting, C. Anton 2002. Testing Analogical Modeling: The /k/~Ø alternation in Turkish. In: Skousen, Royal–Lonsdale, Deryle–Parkinson, Dilworth B. (szerk.): *Analogical Modeling. An exemplar-based approach to language.* John Benjamins. Amsterdam. 123–137.

Sankoff, Gillian–Blondeau, Hélèn 2007. Language change across the lifespan: /r/ in Montreal French. *Language* 83: 560–588.

Sapir, Edward 1921. *Language.* Harcourt, Brace &World. New York.

Shannon, Claude 1948. A mathematical theory of communication. *The Bell System Technical Journal* 27: 379–423, 623–656.

Salzberg, Steven 1990. *Learning with nested generalised exemplars.* Kluwer Academic Publishers. Norwell, MA.

Salzberg, Steven 1991. A nearest hyperrectangle learning method. *Machine Learning* 6: 277–309.

Saussure, Ferdinand de 1962. *Cours de linguistique générale.* Payot. Párizs.

Shepard, R. 1987. Toward a universal law of generalization for psychological science. *Science 237:* 1317–1323.

Simonyi Zsigmond 1881. *Az analógia hatásairól.* (Akadémiai értekezés)

Sinclair, John 1991. *Corpus, concordance, collocation.* Oxford University Press. Oxford.

Siptár Péter 1994. A mássalhangzók. In: Kiefer Ferenc (szerk.): *Strukturális magyar nyelvtan 2. Fonológia.* Akadémiai Kiadó. Budapest. 183–272.

Siptár, Péter – Törkenczy, Miklós 2000. *The Phonology of Hungarian.* Oxford University Press. Oxford.

Skousen, Royal 1989. *Analogical Modeling of Language.* Kluwer Academic Publisher. Dordrecht.

Skousen, Royal 1992. *Analogy and structure.* Kluwer Academic Publisher. Dordrecht.

Skousen, Royal 2002a. An overview of Analogical Modeling. In: Skousen, Royal–Lonsdale, Deryle–Parkinson, Dilworth B. (szerk.): *Analogical Modeling. An exemplar-based approach to language.* John Benjamins. Amsterdam. 11–26.

Skousen, Royal 2002b. Issues in Analogical Modeling. In: Skousen, Royal–Lonsdale, Deryle–Parkinson, Dilworth B. (szerk.): *Analogical Modeling. An exemplar-based approach to language.* John Benjamins. Amsterdam. 27–48.

Skousen, Royal 2002c. Analogical Modeling and quantum computing : Skousen, Royal–Lonsdale, Deryle–Parkinson, Dilworth B. (szerk.): *Analogical Modeling. An exemplar-based approach to language.* John Benjamins. Amsterdam. 319–346.

Skousen, Royal 2009. Expanding Analogical Modeling into a general theory of language prediction. In: Blevins, James–Blevins, Juliette (szerk.): *Analogy in Grammar: Form and Acquisition.* Oxford University Press. Oxford. 164–184.

Skousen, Royal–Lonsdale, Deryle–Parkinson, Dilworth B. (szerk.) 2002. *Analogical Modeling.* John Benjamin. Amsterdam.

Slobin, Dan 1973. Cognitive prerequisites for the development of grammar. In: Ferguson, Charles–Slobin, Dan (szerk.) *Studies of child language development.* Holt Rinehart and Winston. New York. 174–208.

Sóskuthy, Márton 2010. *Analogy at the level of phonology.* (szakdolgozat)

Spencer, Andrew 1988. Bracketing paradoxes and the English lexicon. *Language* 64: 663–682.

Stemberger, Joseph P.–MacWhinney, Brian 1986. Frequency and the lexical storage of regularly inflected forms. *Memory and Cognition* 14: 17–26.

Steriade, Donca 2000. Paradigm uniformity and the phonetics/phonology boundary. In: Pierrehumbert, J.–Broe, M. (szerk.): *Papers in Laboratory Phonology 6.* Cambridge University Press. Cambridge.

Stroppa, Nicolas–Yvon, François 2005. An analogical learner for morphological analysis. In: *Proceedings of the 9th Conference on Computational Natural Language*

Learning (CoNLL 2005). Association for Computational Linguistics. Ann Arbor, MI. 120–127.

Szilágyi N. Sándor 2010. *Szinkrónia és diakrónia – de miről is beszélünk?* (előadás) A magyar nyelvészeti kutatások újabb eredményei II. Kolozsvár. 2010. április 16.

Taylor, John R. 1995. *Linguistic categorization prototypes in linguistic theory* (2. kiadás). Clarendon Press. Oxford.

Thuma Orsolya 2008. *Gyakorisági hatások és nyelvi kétértelműségek vizsgálata a mentális szótárban magyar nyelven. (*PhD-disszertáció)

Tomasello, Michael 2003. *Constructing a language: A usage-based theory of language acquisition.* Harvard University Press.

Törkenczy Miklós 1994. A szótag. In: Kiefer Ferenc (szerk.): *Strukturális magyar nyelvtan 2. Fonológia.* Akadémiai Kiadó. Budapest. 273–392.

Törkenczy Miklós–Siptár Péter 2000. Magánhangzó~semmi váltakozások a magyarban. *Nyelvtudományi Közlemények.* 97: 64–130.

Trón, Viktor–Halácsy, Péter–Rebrus, Péter–Rung, András–Vajda, Péter–Simon, Eszter 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In: *Proceedings of 5th International Conference on Language Resources and Evaluation. ELRA.* 1670–1673.

Trón, Viktor–Rebrus, Péter 2001. Morphophonology and the hierarchical lexicon. *Acta Linguistica Hungarica*. Vol. 48 (1–3): 101–136.

Tversky, Amos 1977. *Features of similarity. Psychological Review* 8: 327–352.

Tversky, Amos–Gati, Itamar 1978. Studies of similarity. In: Rosch, E.–Lloyd B.B. (szerk.): *Cognition and categorization.* Erlbaum. Hillsdale, NJ. 79–98.

Ullman, M. T. 1999. Acceptability ratings of regular and irregular past tense forms: Evidence for a dual system model of language. from word frequency and phonological neighborhood effects. *Language and cognitive processes* 14: 47–67.

Vago, Robert M. 1980. *The Sound Pattern of Hungarian.* Georgetown University Press. Washington.

van den Bosch, Antal 2002. Expanding k-NN analogy with instance families. In: Skousen, Royal–Lonsdale, Deryle–Parkinson, Dilworth B. (szerk.): *Analogical*

Modeling. An exemplar-based approach to language. John Benjamins. Amsterdam. 209–223.

van den Bosch, Antal–Daelemans, Walter 1993. Data-oriented methods for grapheme-to-phoneme conversion. In: *Proceedings of the sixth conference on European chapter of the Association for Computational Linguistics.* Utrecht. 45–53.

Van Rijsbergen, C. 1979. *Information retrieval.* Buttersworth. London.

Váradi Regestrum 1903. Karácsonyi János és Borovszky Samu. Budapest.

Varga Dániel–Simon Eszter 2006. Magyar nyelvű tulajdonnév-felismerés maximum entrópia módszerrel. In: Alexin Zoltán–Vincze Veronika (szerk.): *IV. magyar számítógépes nyelvészeti konferencia. MSZNY 2006.* Szegedi Tudományegyetem. Szeged. 32–38.

Vennemann, Theo 1972. Analogy in generative grammar: The origin of word order. In: Heilman, Luigi (szerk.): *Proceedings of the 11th international congress of linguists Vol. 2.* Bologna: Il Mulino. 79–83.

Vitevitch, Michael S.–Luce, Paul A.–Charles-Luce, Jan–Kemmerer, David 1997. Phonotactics and syllable stress: implications for the processing of spoken nonsense words. *Language and Speech* 40: 47–62.

Wedel, Andrew 2007. Feedback and regularity in the lexicon. *Phonology* 24: 147–85.

Wedel, Andrew 2009. Resolving pattern conflict: Variation and selection in phonology and morphology. In: Blevins, James–Blevins, Juliette (szerk.): *Analogy in Grammar: Form and Acquisition.* Oxford University Press. Oxford. 83–100.

Weiss, S.–Kulikowski, C. 1991. *Computer systems that learn.* Morgan Kaufmann. San Mateo, CA.

Wettschereck, Dietrich–Dietterich, Thomas G. 1995. An experimental comparison of the nearest-neighbor and nearest-hyperrectangle algorithms. *Machine Learning* 19: 1–25.

Wheeler, Benjamin Ide 1887. *Analogy and the scope of its application in language.* University of Cambridge, MA.

Wiebe, Bruce 1992. *Modelling autosegmetal phonology with multi-tape finite state transducers*. (PhD-disszertáció)

Wittgenstein, Ludwig 1922. *Tractatus Logico-Philosophicus.* Routledge & Kegan Paul. London.

Wittgenstein, Ludwig 1992. *Filozófiai vizsgálódások.* Atlantisz Kiadó. Budapest.

Wulf, Douglas J. 2002. Applying Analogical Modeling to the German plural. In: Skousen, Royal–Lonsdale, Deryle–Parkinson, Dilworth B. (szerk.): *Analogical Modeling. An exemplar-based approach to language.* John Benjamins. Amsterdam. 109–122.