

A magyar tárgyeset ingadozó kötőhangzójának kvantitatív elemzése¹

Rácz Péter,¹ Rebrus Péter,² Tóth Szilárd²

¹Budapesti Műszaki és

Gazdaságtudományi Egyetem, Kognitív Tudományi Tanszék

²ELTE BTK – NYTK Kihelyezett Elméleti Nyelvészeti Központ

If a Hungarian noun stem ends in a coronal sonorant + coronal fricative consonant cluster, the linking vowel of the accusative can vary (*klienset* / *klienst* ‘client.acc’). The phonotactics of the language does not completely account for this behaviour. The existing literature points to phonetic, lexical, and analogical factors. We use data drawn from the 2nd Hungarian Webcorpus to expand on existing descriptive work and determine the factors driving variation in the linking vowel. Our results point towards a novel explanation, namely, that the variable pattern is the result of a lexical gang effect.

Keywords: morphology, variation, corpus, machine learning

Kulcsszavak: alaktan, variáció, korpusz, gépi tanulás

1. Bevezetés

A magyar nyelvben a kötőhangzó megjelenésének jellemzően morfofonológiai és fonotaktikai motivációja van (Rebrus et al. 2024; Siptár – Törkenczy 2000). Példa erre a főnévi tárgyeset, ahol a kötőhangzó megjelenését befolyásolja a tövégződés. Ha egy mássalhangzó + *t* szekvencia előfordul monomorfemikus tövek végén, akkor a *-t* toldalékkal képzett tárgyesetben általában nem találunk kötőhangzót (1a); ha a mássalhangzócsoporthoz nem fordul elő *t* végén, akkor a tárgyesetet is általában kötőhangzóval képezzük (1b). A kötőhangzó viselkedését azonban nem magyarázhatjuk teljes egészében a monomorfemikus alakok fonotaktikájával, mivel léteznek olyan alakok, amelyekben a tövekben nem előforduló mássalhangzócsoporthoz nem bontja fel a kötőhangzó (1c), és olyanok is, amelyekben a csoport *t* végén előfordul, de tárgyesetben kötőhangzót vesz fel (1d) – ez utóbbi típusba

¹ A kutatást a Nemzeti Kutatási, Fejlesztési és Innovációs Alap NKFI-K139271 (A paradigmazszerkezet szerepe a magyar fonológiában és alaktanban nyelvtipológiai szempontból), NKFI-FK138188 (Az alaktani konvergencia vizsgálata kísérletes módszerekkel) és az Innovációs és Technológiai Minisztérium TKP2023-NVA-2 számú pályázata támogatta.

tartoznak a nyitótövek is, ahol a szokásosan középső kötőhangzó alsóként realizálódik (1d iii-iv).

Ha a tő két mássalhangzóra végződik, a tárgyesetet jellemzően kötőhangzóval képezzük (2a). Itt is vannak ugyanakkor kivételek. A koronális szonoráns + koronális réshang mássalhangzócsoporthoz végződő töveknél a kötőhangzó nem mindig jelenik meg, gyakran ingadozik (2b), annak ellenére, hogy ezek a mássalhangzó-csoportok csak egy maroknyi monomorfemikus alakban fordulnak elő (*karszt*, *dunszt*, *kunszt*, *vurst*).

(1) Mássalhangzó-kapcsolatok tövekben és tárgyesetben (VC-végű tövek)

(a) Monomorfemikus szóvégi *Ct* grammatikus és *C-t* tárgyeset (nincs kötőhangzó)

- (i) *párt kár-t*
- (ii) *szint szín-t*
- (iii) *rost kas-t*
- (iv) *liszt rész-t*

(b) Monomorfemikus szóvégi *Ct* agrammatikus és *C-Vt* tárgyeset (kötőhangzó)

- (i) *mt# *nyom-ot*
- (ii) *cst# *ács-ot*
- (iii) *ct# *malac-ot*

(c) Monomorfemikus szóvégi *Ct* agrammatikus, de *C-t* tárgyeset

- (i) *nyt# *lány-t*

(d) Monomorfemikus *Ct*# létezik, de *C-Vt* tárgyeset

- (i) *recept kalap-ot*
- (ii) *lift ráf-ot*
- (iii) *jacht pech-et*
- (iv) *párt ár-at*
- (v) *rost has-at*

(2) Koronális szonoráns + koronális réshang szekvenciára végződő tövek tárgyesete

(a) Kötőhangzó

- (i) *bronz-ot*
- (ii) *bors-ot*
- (iii) *rajz-ot*
- (iv) *pajzs-ot*

(b) Ingadozó kötőhangzó

- (i) *ágens-t / ágens-et*
- (ii) *spájz-t / spájz-ot*
- (iii) *%félsz-t / félsz-et*
- (iv) *pénz-t / %pénz-et*

Dolgozatunk témája a kötőhangzó ingadozása a koronális szonoráns + koronális réshang mássalhangzócsoporthoz tartozó főnévi tövek tárgyesetében (2b). Az egyéb mássalhangzó-kapcsolatra tartozó tövek közül azok, amelyeknél egyáltalán lehetséges a kötőhangzó nélküli tárgyeset (C + koronális szonoráns típus), rendkívül kevesen vannak (pl. *görl, fájl, kombájn, konszern* stb.), ezért ezekkel itt nem foglalkozunk. A fenti példák azt mutatják, hogy a kötőhangzó viselkedését nem vezethetjük le teljes egészében a fonotaktikából. Az *ágens-t / ágens-et* típusú alakok viselkedésére a szakirodalomnak három magyarázata van.

Először, a koronális nazális + koronális réshang kapcsolatokban a nazális vokalizálódik, ezzel leegyszerűsítve a szóvégi mássalhangzócsoporthoz (Siptár 1994: 197), ami megmagyarázza, hogy a nazális + koronális réshang csoportok után a kötőhangzó miért maradhat el (pl. *régens-et* [re:gɛʃet] / *régens-t* [re:gɛʃt]).

Másodszor, a kötőhangzó a beszélők (egyéni szinten tudatos vagy csoportszintű szelekciós) eszköze lehet arra, hogy a kevésbé felismerhető szótöveket megkülönböztessék. E szerint a magyarázat szerint a kötőhangzó a szótövet robusztusabbá teszi, és annak felismerését megkönnyíti (*sors-ot* vs. *%sors-t*).

Az az elképzelés, hogy a magyar nyelvű alaktani ingadozások érzékenyek arra, hogy a főnévi szókészlet mennyire szokatlan vagy nehezen hozzáférhető a beszélők számára, nem új. Papp (1968) ezt bázismegkülönböztető tendenciának nevezte. A szóalakok megjósolhatósága és felismerhetősége a legtöbb kortárs nyelvelmélet szerint érzékeny a szóhosszra, a hangtani redukcióra és a bennük szerepet játszó megkülönböztető jegyek számára (Gibson et al. 2019; Wedel – Kaplan – Jackson 2013). A bázismegkülönböztető tendencia egy formájának tekinthetjük azt, hogy a kötőhangzó frissen kölcsönzött, idegen eredetű főnevekben (Lantay 2020), illetve tulajdonnevekben (Blaskovics, személyes közlés) máshogyan ingadozik.

Harmadszor, a kötőhangzó ingadozása lehet egymással szemben álló analógias hatások következménye. Kálmán – Rebrus – Törkenczy (2012) felvetése az, hogy a magyar nyelvben a kötőhangzóknak két erős mintázata figyelhető meg: CC-végű tövek esetén van kötőhangzó, s-végű tövek esetén pedig nincs. Mivel az *ágens* típusú tövek mindkét mintázatba beleesnek, a kötőhangzó itt ingadozni fog.

Dolgozatunk ezt a három hipotézist vizsgálja a magyar webkorpuszból vett adatokon. A három hipotézis nem zárja ki egymást, de relatív magyarázó erejük eltérhet. A talált eloszlások különböző mértékben motiválhatják őket:

- (H1) Ha a koronális szonoráns + koronális réshang kapcsolatokban a kötőhangzó ingadozásának oka az, hogy a koronális nazális vokalizálódik, a kötőhangzó ritkábban fog megjelenni nazálist tartalmazó mássalhangzó-csoportok után, mint olyanok esetén, melyek más szonoránst tartalmaznak.
- (H2) Ha a kötőhangzó megjelenését a bázismegkülönböztető tendencia motiválja, a kötőhangzó ritkább lesz akkor, ha a tö gyakoribb vagy hosszabb, tehát könnyebben felismerhető.
- (H3) Ha az ingadozást egymásnak ellentmondó analógiás hatások motiválják, akkor a kötőhangzó ingadozásának inkább egyenletes, uniform eloszlását várjuk.

2. Módszertan

A nyolc és fél milliárd alakot tartalmazó új magyar webkorpuszból (Nemeskey 2020) előállított típusgyakorisági listából kigyűjtöttük a tárgyesetű főnévi alakokat, a kapott listát pedig a nyilvánosan hozzáférhető hunspell helyesírási adatbázis (Ooms 2024) alapján létező alakokra szűrtük. Az így kapott listát leszűkítettük a $\{l, r, j, n\} \{sz, s, z, zs\}$ mássalhangzó-kapcsolatokra végződő tövekre. A listából eltávolítottuk a hibásan annotált alakokat és az összetett töveket. Ugyancsak eltávolítottuk azokat az alakokat, amelyeknek elsősorban csak melléknévi jelentése van (*domináns, pikáns, gáláns* stb.). Az így kapott adathalmaz 136 főnévi tövet tartalmaz. Az adathalmazból azt a 62 főnévi tövet tartottuk meg, amelyek tárgyesetben ingadozást mutattak. Az eredményül kapott listán megszámláltuk minden tö kötőhangzós (*klienset*) és kötőhangzó nélküli (*klienst*) tárgyesetű alakjait. A kutatásban használt adathalmaz a 62 tö 217 078 kötőhangzós és 919 686 kötőhangzó nélküli alakját tartalmazza. A tövekhez tartozó, kötőhangzós és kötőhangzó nélküli szópárok listáján a következő változókat vezettük be. Kimeneti változónk a kötőhangzós és kötőhangzó nélküli alakok számának aránya, illetve annak természetes logaritmusa (*logodds*). Ez az érték jól tükrözi az adott tö kötőhangzós változatának preferenciáját. Egy példával élve: az *incidens* tö *incidenset* alakjából az adatbázisban 23, az *incidenst* alakból 10 852 fordul elő, az aránypár logaritmusa tehát $\log(23/10852)$, azaz $-6,11$. Az adatbázis a *vers* tö *verset* alakjából 87 593, a *verst* alakból 34 példányt tartalmaz, az aránypár logaritmusa $7,83$. Prediktor változóink a következők voltak: (i) A szóvégi mássalhangzócsoporthoz első tagja, azaz $\{l, r, j, n\}$. (ii) A szóvégi mássalhangzócsoporthoz második tagja, azaz $\{sz, s, z, zs\}$. (iii) Specifikusan a leggyakoribb szóvégi mássalhangzócsoporthoz, az *-ns* (ld. 3. pont). (iv) A tö szótagszáma. (v) A tö gyakorisága (a webkorpuszban tízmillió alak közötti előfordulás gyakoriságának logaritmusa). Az adatbázis tartalmára az 1. táblázat ad néhány kiválasztott példát.

1. táblázat: Az elemzéshez használt adatbázisból vett példák. Az oszlopok a példának választott töveket, a sorok a tövek tulajdonságait mutatják.

tő	incidens	komponens	reneszánsz	spájz	bronz
kötőhangzós tárgyeset	incidenset	komponenset	reneszánszot	spájzot	bronzot
kötőhangzó nélküli tárgyeset	incidenst	komponenst	reneszánszt	spájzt	bronzot
log(gyakoriság/10 ⁶ szó)	2.14	2.29	2.21	1.22	2.02
tő szótagszáma	3	3	3	1	1
kötőhangzós tárgyeset gyakorisága	23	95	616	307	7413
kötőhangzó nélküli tárgyeset gyakorisága	10852	9290	588	65	41
kötőhangzó/nincs kötőhangzó	0.002	0.01	1.05	4.67	176.52
log (kötőhangzó/nincs kötőhangzó)	-6.11	-4.57	0.05	1.54	5.17

Az adatokra kódolt prediktorváltozók erősen összefüggenek egymással, például *nsz* végződésből jóval több van, mint akár *nz*, akár *rsz* végződésből (erre a 3. pontban visszatérünk), ezért az adatokra egy véletlenerdő-modellt illesztettünk, amely az (iv-v) prediktorokból a kötőhangzó valószínűségét jósolta meg. A véletlen erdő (Breiman 2001) egy gépi tanuló algoritmus, amely a tanuló adatokra a prediktor változókat és a minta méretét véletlenszerűen kiválasztva döntési fák sorozatát illeszti, és a fák jóslatait átlagolja, így elkerüli a prediktor változók együttállásából adódó értelmezési problémákat.

Az elemzésbe a kategorikus változókat (i, ii, iii) 1-az-N-ből kódolással (*one hot encoding*), a folyamatos változókat (iv, v) számvektorként vettük be. A modell hiperparaméterei a következők voltak:

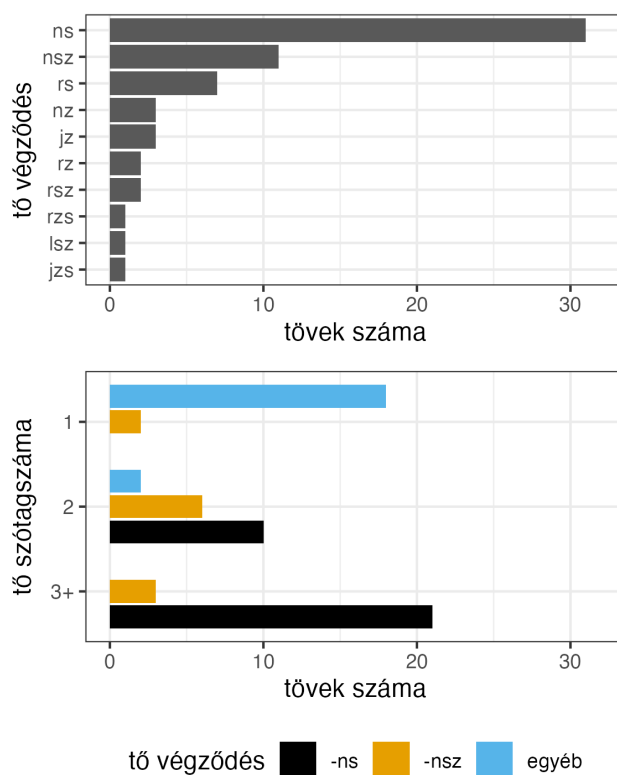
- Felhasznált prediktorváltozók lehetséges száma: 1–5,
- Csomópontok lehetséges legkisebb mérete: 2/5/10/20,
- Csomópontok lehetséges legnagyobb száma: 1/3/5/10/20.

Mivel szerény méretű adathalmazunkat nem akartuk további tanító és tesztalмаzokra bontani, a legjobb modellt háromszlatú keresztvalidálással (*3-fold cross validation*) választottuk ki, a magyarázott variancia szerint rangsorolva. A legjobb modell által magyarázott variancia 0,65 (a felhasznált prediktorváltozók száma 3, a csomópontok legkisebb mérete 20, a csomópontok legnagyobb száma 5). Ez az

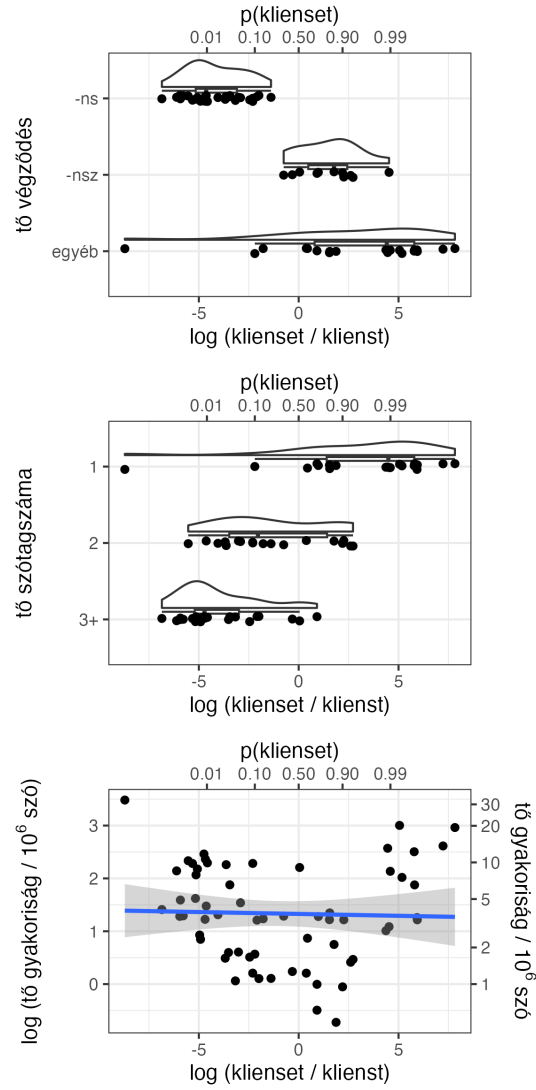
érték a gépi tanulásban aránylag alacsonynak számít, ugyanakkor jóval magasabb, mint egy olyan modell pontossága, amely egyszerűen a kimeneti változó mediánja alapján tesz jóslatot az összes megfigyelésre (-0,06). Ez azt mutatja, hogy a kötőhangzó megjelenése strukturáltan váltakozik, és ezt a prediktorok valamennyire megragadják.

Az ábrák *ggplot*-ban készültek, a véletlenerdő-elemzést a *randomforest* csomag segítségével illesztettük (Wickham 2016; Liaw – Wiener 2002).

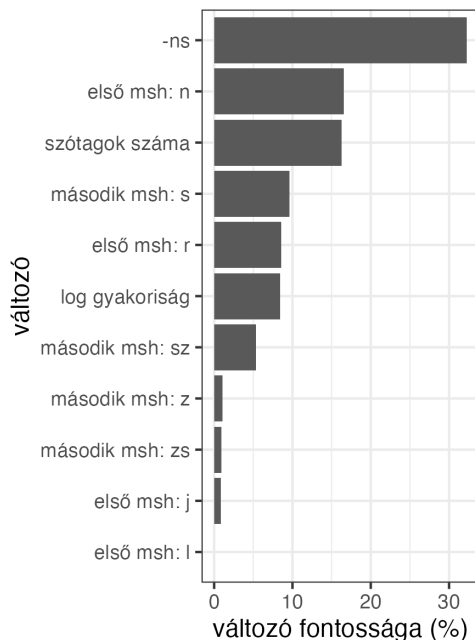
3. Eredmények



1. ábra: A különféle mássalhangzócsoportokra végződő ingadozó tövek száma (fent), az *ns/nsz/egyéb* mássalhangzócsoportra végződő tövek szótagszáma (lent).



2. ábra: Az *ns/nsz/egyéb* mássalhangzócsoporthra végződő tövek esetén a kötőhangzó megjelenésének valószínűsége (fent), a tő szótagszáma és a kötőhangzó megjelenésének a valószínűsége (középen), a tő tokengyakorisága és a kötőhangzó megjelenésének valószínűsége (lent).



3. ábra: A véletlenerdő-modell szerint a kötőhangzó megjelenésével együttálló legfontosabb változók.

A korpuszban talált mintázatok az 1–2. ábra tekinti át. A mintába a koronális szonoráns + szibiláns csoportra végződő, a tárgyesetben a kötőhangzó ingadozását mutató tövek tartoznak. Az 1. ábra a különféle mássalhangzókra végződő és különféle szótagszámú tövek számát mutatja. A különféle tövégi mássalhangzócsoporthoz gyakorisága nem azonos. A legtöbb ingadozó tö *ns*-re vagy *nsz*-re végződik (fent). Ha megnézzük ezt a két típust, azt látjuk, hogy az *ns*-végű tövek jellemzően hosszabbak (nincs közöttük egyszótagú, a legtöbb három, vagy több szótagos), az *nsz*-végű tövek jellemzően két szótagból állnak, a többi tö pedig elsősorban egyszótagú (lent).

A 2. ábra azt mutatja, milyen gyakori a kötőhangzó különféle tövek esetén. Minden pont egy tö, és minél inkább jobbra van a vízszintes tengelyen, annál valószínűbb, hogy a tö egy adott alakja kötőhangzóval jelenik meg a korpuszban. Az *ns*-végű töveknél a kötőhangzó jóval ritkább, mint akár az *nsz*-végű tövek vagy a többi tö esetén (fent). Minél hosszabb, minél több szótagból áll a tö, annál ritkább a kötőhangzó (középen). Nem szabad szem elől téveszteni ugyanakkor, hogy a tö hosszúsága és a végződés között összefüggés van, ezért első megközelítésben nehéz szétválasztani, hogy például a hosszabb tövek vagy az *ns*-végű

tövek állnak együtt gyakrabban kötőhangzóval a tárgyesetben, hiszen ez a két csoport erős átfedésben van. A *tő* gyakorisága és a kötőhangzó valószínűsége között nem látszik egyértelmű, lineáris összefüggés (lent).

A változók közötti átfedések kezelésére véletlenerdő-elemzést használtunk. Az elemzés megadja, hogy a prediktor változók a kötőhangzók megjelenésének megjósolásában mennyire fontos szerepet játszanak. Ezt mutatja a 3. ábra. Azt látjuk, hogy a legfontosabb változónk az, hogy a *tő* *ns*-re végződik-e. Ehhez képest jóval kisebb a többi változó súlya. A második legfontosabb változó az, hogy a *tő* általában *nC*-re végződik, a harmadik pedig a *tő* hossza (a szótagok száma). A többi változó, köztük a *tő* gyakoriságának súlya arányaiban elhanyagolható. Érdeemes tekintetbe venni, hogy 11 prediktorváltozónk van, ami azt jelenti, hogy ha mindegyik egyformán fontos szerepet játszana, akkor a fontosságuk 9% körül lenne. Azok a prediktorok tehát, amelyek 9% fölötti fontossággal bírnak, jóval erőn felül, azok pedig, amelyek 9% alatt vannak, erőn alul járulnak hozzá a kötőhangzó megjelenésének meghatározásához.

4. Diszkusszió

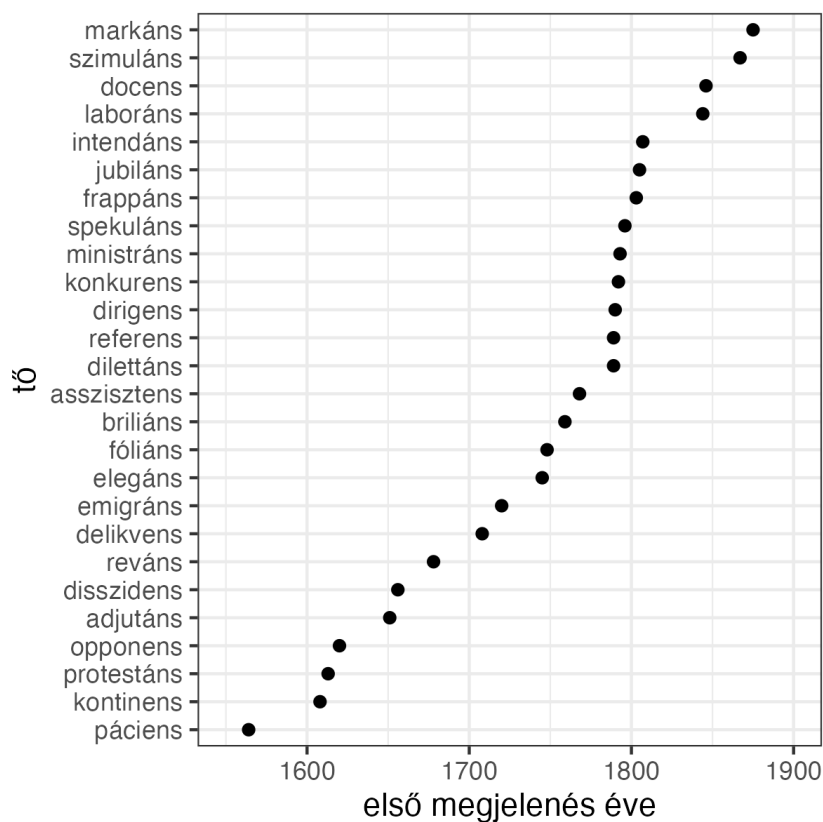
Dolgozatunk elején a rendelkezésre álló szakirodalom alapján három hipotézist fogalmaztunk meg arra vonatkozóan, hogy a koronális szonoráns + koronális részhang kapcsolatra végződő tövek tárgyeseténél mi lehet a kötőhangzó ingadozásának mozgatórugója.

Eredményeink nem támasztják alá azt a hipotézist, hogy a kötőhangzó viselkedése a koronális nazális vokalizációjának következménye (H1), hiszen az *ns*-végű tövek jóval inkább kerülnek a kötőhangzót, mint az *nsz*-végű tövek, annak ellenére, hogy mindkét típusban koronális nazálist találunk. Ennek a különbségnek fonetikai okai lehetnek, amelyek további vizsgálatra érdemesek. A bázismegkülönböztető tendenciára (H2) csak gyenge bizonyítékaink vannak. Adatainkban a *tő* gyakorisága nem befolyásolja a kötőhangzó megjelenését (annak ellenére, hogy a gyakori töveket könnyebb felismerni), a hosszabb töveknél viszont valóban ritkább a kötőhangzó. Ez alátámaszthatná a bázismegkülönböztető preferenciát, mivel a hosszabb tövek kötőhangzó nélkül is felismerhetőbbek maradnak (*antioxidáns-t* / *docens-t*). Árnyalja a képet viszont, hogy a hosszabb tövek jelentős mértékben *ns*-végű tövek. Végül az, hogy a magyar nyelvben jellemzően a mássalhangzó-kapcsolatra végződő tövek után van kötőhangzó, az *s*-végű tövek után pedig nincs, közvetve magyarázhatja az adatokban látott ingadozást, arra azonban nem derít fényt, hogy miért viselkednek másként az *ns*-végű tövek, mint a többi *tő* (H3).

A fentieket kiegészítő, post hoc magyarázatunk az, hogy az ingadozó *ns*-végű főnévi tövek jelölt lexikai osztályt alkotnak. E tövek mindegyike latin átvétel, és jellemzően a nyelv művelt regisztere használja őket: *agens*, *antidepresszáns*, *antioxidáns*, *asszisztens*, *delikvens*, *docens*, *fóliáns*, *incidens*, *kliens*, *koefficiens*,

komplemens, komponens, konstans, kontinens, kontingens, mutáns, páciens, pá-tens, posztamens, precedens, protestáns, reagens, referens, régens, stimuláns, szegmens, szimpatizáns, tangens, tremens, variáns.

Megvizsgáltuk az adataink között szereplő összes (nem csak ingadozó és nem csak főnévi) *ns*-végű tövet, és kiválogattuk azokat, amelyekhez az *Új magyar etimológiai szótár* (Gerstner 2011–2022) egyértelműen hozzárendeli a *tő* első adatolt írásos megjelenésének az évét. Ezek a tövek láthatók az 4. ábrán.



4. ábra: Az *adjudtáns, asszisztens, briliáns, delikvens, dilettáns, dirigens, disszidens, docens, elegáns, emigráns, fóliáns, frappáns, intendáns, jubiláns, konkurens, kontinens, laboráns, markáns, ministráns, opponens, páciens, pikáns, protestáns, referens, reváns, spekuláns, szimuláns* szavak átvétele a magyar nyelvben az *Új magyar etimológiai szótár* szerint.

Az ábra jól mutatja, hogy ezek a tövek a huszadik század előtt, jellemzően a 18. század végén kerültek a nyelvbe. Okunk van azt feltételezni, hogy ezek a forrásnyelvük és az átvételük idején osztozó, művelt regiszterbe tartozó tövek a mentális lexikonban, és ebből eredően a nyelvi közösség nyelvhasználatában is hasonlóan viselkednek, aminek az egyik lenyomata az, hogy tárgyesetben jobban kerülik a kötőhangzót, mint azt akár a fonetikai, fonotaktikai, vagy más lexikális tulajdonságaik indokolnák.

A szakirodalom sok hasonló példáját ismeri annak, amikor az ugyanazon szemantikai, szociális és alaktani mezőbe tartozó szavak morfofonológiailag hasonlóan viselkednek. Ezt csoporthatásnak (*gang effect*) szokták hívni (Bybee – Moder 1983; Pierrehumbert 2012; Hay et al. 2019). A magyar nyelvből származó példa Kiefernek (1990: 123) az a megfigyelése, hogy a latináta tövek birtokos alakja az ingadozó *j*-t kerüli és erősen preferálja a *j* nélküli alakot (pl. *virága* / *virágja*, de *transzformátora* / *%transzformátorja*). Állításunk az, hogy az *ns*-végű tövek ugyanígy kerülnek a kötőhangzót. A tárgyeset kötőhangzójának ingadozása mögött meghúzódó oksági viszonyokat csak történeti vizsgálatok tudnák megnyugtatóan tisztázni.

Elemzésünk nem ad kimerítő magyarázatot a kötőhangzó ingadozására az összes koronális szonoráns + koronális réshang mássalhangzócsoporthoz végződő főnévi fő tárgyesetében. Ha a nyelvi intuíciónk helyett a rendelkezésre álló adatokhoz fordulunk, azt látjuk, hogy aránylag kevés olyan egyértelműen főnévi fő van, amely beleesik ebbe a halmazba (a mi módszertanunk 136-ot azonosított), és ezeknek csak egy része (itt: 62) mutat számottevő ingadozást. Az *ns*-végű és *nsz*-végű tövek viselkedésén és azon az átfedő tendencián túl, hogy a rövidebb tövek gyakrabban kapnak kötőhangzót, nem tudjuk jobban megragadni a talált eloszlásokat. Ez azt is jelenti, hogy például nincs frappáns válaszuk arra, hogy miért gyakoribb a *bronzot* alak, mint a *%pénzet* alak. Ebben a konkrét különbségben a gyakoriság szerepet játszhat, viszont az egész mintán nem látunk szisztematikus gyakorisági hatásokat. Az ilyen típusú kérdésekre csak egy kísérletes vizsgálat tudna fényt vetni. Eredményeink ugyanakkor hangsúlyozzák a külső (nyelven kívüli) hatások és a történeti esetlegességek fontosságát abban, hogy a szinkrón nyelvi mintázatok jobban megértsük (Nettle 2012).

5. Limitációk

A webkorpusz használatának komoly előnye, hogy az írott magyar nyelvből nagy merítést ad, és ez a ritka alakoknak az elemzésbe való bevonására is lehetőséget biztosít. (Például *pácienset* alak 98-szor fordul elő az adathalmazban. Egyrészt ez a szám túl nagy ahhoz, hogy mindegyik előfordulást hibásnak vagy esetlegesnek tekintsük, másrészt a teljes korpusz nyolc és fél milliárd alakból áll, tehát itt egy

nagyon ritka változatról beszélhetünk.) A korpusz méretéből azonban az is következik, hogy a megalkotásához használt módszerek automatikusak, és ezek pontatlanok lehetnek.

Elképzeltető, hogy a szóalakok szófaji elemzéséhez használt rendszer kétértelmű helyzetekben a szórenden és a végződéseken túl a kötőhangzók jelenlétére is támaszkodik, amikor például arról hoz döntést, hogy a *konkurens* alak melléknév vagy főnév, és ez a melléknévként is használt tövek kötőhangzós alakjának alulméréséhez vezethet. A korpusz a toldalékolt alakokat helyesírás szerint nem szűri, a mi listánk pedig csak a tövek helyesírását vette a szűrés alapjául, de a forrásszövegeknél (újságcikkek, fórumbejegyzések stb.) megtörténhet, hogy ezeket helyesírássellenőrző javította, és ez számos szónál a kötőhangzós vagy a kötőhangzó nélküli alakot (*antioxidáns*, *pácienst*) elutasítja. Végezetül, ekkora forrásadathalmaznál már rizikófaktor, hogy a kötőhangzós alakok egy része sima elütésből származik, és nem a nyelvhasználó mentális reprezentációját tükrözi. Ezeket a limitációkat a nyelvhasználók közvetlen, kísérletes formában történő tesztelése tudná feloldani, mert ez egyértelműbb választ tudna adni arra, hogy a beszélők az ingadozó kötőhangzókat hogyan használják.

6. Adatok

A közleményhez felhasznált adatokat és kódokat tartalmazó adatbázis az alábbi linken érhető el:

<https://doi.org/10.5281/zenodo.17575052>

Irodalom

- Breiman, Leo (2001), Random forests. *Machine Learning* 45: 5–32.
- Bybee, Joan L. – Moder, Carol Lynn (1983), Morphological classes as natural categories. *Language* 59/2: 251–270.
- Gerstner Károly (szerk.) (2011–2022), Új magyar etimológiai szótár – online kiadás. MTA Nyelvtudományi Intézet – ELKH Nyelvtudományi Kutatóközpont. <https://uesz.nytud.hu/index>
- Gibson, Edward – Futrell, Richard – Piantadosi, Steven P. – Dautriche, Isabelle – Mahowald, Kyle – Bergen, Leon – Levy, Roger (2019), How efficiency shapes human language. *Trends in Cognitive Sciences* 23/5: 389–407. <https://doi.org/10.1016/j.tics.2019.02.003>
- Hay, Jennifer – Walker, Abby – Sanchez, Kauyumari – Thompson, Kirsty (2019), Abstract social categories facilitate access to socially skewed words. *PloS One* 14/2: e0210793. <https://doi.org/10.1371/journal.pone.0210793>
- Kálmán László – Rebrus Péter – Törkenczy Miklós (2012), Possible and impossible variation in Hungarian. In: Kiefer, Ferenc – Ladányi, Mária – Siptár, Péter (eds), *Current issues in morphological theory: (Ir)regularity, analogy and frequency*. John Benjamins, Amsterdam – Philadelphia. 23–49. <https://doi.org/10.1075/cilt.322>

- Kiefer Ferenc (1990), *-a/-e* vagy *-ja/-je*? Nyelvtudományi Közlemények 91: 119–125.
- Lantay Gyula (2020), Analógia a jelölt mássalhangzó-kapcsolatra végző, idegen eredetű főnevek tárgyesetében. Magyar Nyelv 116/3: 299–308. <https://doi.org/10.18349/MagyarNyelv.2020.3.299>
- Liaw, Andy – Wiener, Matthew (2002), Classification and regression by randomForest. R News 2/3: 18–22.
- Nemeskey Dávid Márk (2020), Natural language processing methods for language modeling. PhD Thesis. Eötvös Loránd University, Budapest.
- Nettle, Daniel (2012), Social scale and structural complexity in human languages. Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences 367/1597: 1829–1836. <https://doi.org/10.1098/rstb.2011.0216>
- Ooms, Jeroen (2024), hunspell: High-performance stemmer, tokenizer, and spell checker. <https://CRAN.R-project.org/package=hunspell>
- Papp Ferenc (1968), Adatok szókincsünk etimológiai rétegeinek végződés szerinti eloszlásához. Nyelvtudományi Közlemények 70: 200–209.
- Pierrehumbert, Janet B. (2012), The dynamic lexicon. In: Cohn, Abigail – Fougeron, Cécile – Huffman, Marie (eds), The Oxford handbook of laboratory phonology. Oxford University Press, Oxford. 173–183.
- Rebrus, Péter – Szigetvári, Péter – Törkenczy, Miklós (2024), No lowering, only paradigms: A paradigm-based account of linking vowels in Hungarian. Acta Linguistica Academica 71: 137–170. <https://doi.org/10.1556/2062.2023.00674>
- 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.
- Wedel, Andrew B. – Kaplan, Abby – Jackson, Scott (2013), High functional load inhibits phonological contrast loss: A corpus study. Cognition 128/2: 179–186.
- Wickham, Hadley (2016), ggplot2: Elegant graphics for data analysis. Springer-Verlag, New York. <https://ggplot2.tidyverse.org>