

# Szösszenet az elveszett morfémákért

## Az alaki analógiák haszna

Naszódi Mátyás

MorphoLogic KFT., e-mail: naszodim@morphologic.hu

**Kivonat** A jelenlegi morfológiai elemzők gyakorlati okok miatt elég pragmatikus módon készültek. A céljuk, aránylag kis munkával fedjék le a magyar nyelvű szövegeinek szóalakjait minél kevesebb hibával. Ha a célt elérték, a szabályszerű eseteket jól leírták, a deviáns, kisebb gyakorisággal előforduló eseteket kivételként, egyedileg kezelik. A vizsgálataim szerint sokkal kevesebb kivétel van. A szavak végződése szerinti csoportosítással felderíthetők azok a szavak közötti összefüggések, melyek a korábbi adatbázisokból hiányoznak. A módszer segítségével elfeledett vagy csak leíró nyelvészek által említett szógyökök, toldalékok kerülnek napvilágra. Sőt a feltárás eredményeként pontosíthatóak a praktikus célra készült nyelvészeti, nyelvi táruk.

**Kulcsszavak:** morfológia, lexikográfia, helyesírás-ellenőrzés

### 1. Bevezető

Munkám során, melyek többsége nyelvi eszközök készítése, esetenként észleltem, hogy a meglevő szavakra vonatkozó átfogó adatbázisok hiányosak abban az értelemben, hogy az adatokból kimaradnak olyan – a cél érdekében hasznosítható – információk, melyeket könnyedén lehetne részévé tenni az adatbázisnak. Jelenleg csak morfológiai rendszerekre szorítkozom, ami persze érinti a szóelemzőket, generátorokat, helyesírás-ellenőrzőket.

A 90-es évek elején Elekfi László ragozási szótárából kiindulva készítettem helyesírás-ellenőrzőt. Már akkor is igyekeztem egységesebb, tömörebb formára hozni a nyelvészeti adatokat. Eseti megoldásim voltak, de látom, alaposabb kutatást érdemel a téma. Most, ha nem is a teljesség igényével, de módszerem megmutatásával prezentálom, milyen kis munkával milyen eredményeket lehet elérni. A módszer pedig egyszerű: tisztán formai kérdéseket teszek fel nyelvi adatbázisoknak, majd a válasz után generatív módon tesztelem, vajon igaz-e a feltételezésem. Ily módon számtalan morfémát, nem használt toldalékolási formát, szabályt fedek fel, melyek – ha visszakerülnek a lekérdezett nyelvi adatbázisba – gazdagítják, pontosítják azt.

## 2. Szótani adatbázisok

A szótani adatbázisok több célt szolgálhatnak. A többnyelvű szótárakkal most nem foglalkozom, mert az adott problémától távol állnak. Engem azok a szótárak izgatnak, melyekben toldalékolásra van több-kevesebb információ. Ezek alapja – akár tagadják a készítői, akár nem – a régi Magyar értelmező kéziszótár (MTA, 1972). Ebből indult ki Papp Ferenc a tergo szótára (Papp, 1969), az összes helyesírási adatbázis (Proszéki és Kornai, 2017), de a helyesírási lexikonoknak is ez az alapja.

A legtöbb formális toldalékolási információval a helyesírás-ellenőrzők számára készített forráskódok rendelkeznek, illetve Elekfi László ragozási szótára (Elekfi, 1994), mely ugyan nem számítógépes felhasználásra készült, de precíz sége miatt alkalmas arra, illetve használtam is ebből a célból. Ezek a szótárak, mint adatbázisok, három részre bonthatók.

- Szószedet: a szótövek tárháza
- Toldaléktár: a toldalékok, előtagok jegyzéke
- Toldalékolási szabályok: itt derül ki formális leírással, hogyan és milyen formában kapcsolódhatnak a morfémák.

Ez utóbbi többnyire az előző két tárban levő morfémákra akasztott jegyek alapján működik.<sup>1</sup>

Az adatbázisok többsége nyilvános, könyv vagy elektronikus formában elérhető, illetve nekem ennél többhöz van hozzáférésem. A táruk elsősorban a leírás formájában térnek el. A formai különbségen túl az igei és névszói ragozásban, jelek használatában nincs lényeges eltérés. Apróbb részletekben itt is vannak különbségek: Elekfi szótára túl szigorú. Például egyes szavaknál letiltással olyan ragokat nem enged meg, melyek a gyakorlatban mégis előfordulnak. Az esetragok halmaza is különböző a különböző helyesírás-ellenőrzőkben. A ritkább régies ragozási formák nem mindegyikben szerepelnek. Ezek apróságok. Több-kevesebb munkával egy fedél alá lehet hozni az eltérő formalizmusokat. Ezzel most nem foglalkozom, bár itt is vannak olyan kérdések, amelyek a módszeremmel tisztázhatóak.

A képzők és a szóösszetételek mutatnak nagy különbséget az adatbázisokban. És itt van a legnagyobb pontatlanság, mert sokszor nincs is pontos nyelvészeti leírás, csak közelítősen, megérzésemre alapuló szabályok.

## 3. A szótövek

A régi magyar értelmező kéziszótár tartalma mintegy 70 000 tétel. Ez azt jelenti, hogy ennyi szónak magyarázza a jelentését. Ezek között vannak összetett, képzett szavak, igekötős igék és kis mennyiségben kifejezések. Ha azt nézzük, hogy

<sup>1</sup> Van ettől eltérő rendszer. Például a Kimmo–Koskenniemi által jegyzett TLFA két-szintű morfológia (Koskenniemi, 1983), melynek egyik magyar adatbázisa a XEROX tulajdona, de a kiindulási alapot a MorphoLogic szolgáltatta. Ezzel nem foglalkozom, mert nem hozzáférhető az adatbázis.

ezt a szókészletet hány szótóval lehet lefedni, meglepő eredményt kapunk. Nincs 20 000. Azért nem mondok pontos számot, mert függ attól, milyen szóképzéseket, -összetételeket kezel a morfológiai rendszer algoritmikusan. Persze a jelentés nem mindig értelmezhető a morfémák jelentésének összekapcsolásával, emiatt indokolt lehet a 70 000 tétel.

Ha viszont a jelentés (fordítás) nem érdekel, akkor felesleges ekkora tár, hisz a többi alak generatív módon megkapható egy kisebb halmazból.

#### 4. Egyéb morfémák

A ragokkal, jelekkel, mint írtam, nincs igazán gond. Jeles nyelvészek, Papp Ferenc, Seregy Lajos, Elekfi László, valamint a helyesírás-ellenőrök alkotói ezt elég jól feltérképezték, legfeljebb elkódoltak egy-egy tételt. A képzők viszont várnak még hasonló alapmunkára. Tisztázatlan például az igék műveltetésének *-tat*, *-tet*, *-at*, *-et* fonetikai besorolása. Az világos, hogy mikor magas, mikor mély hangrendű egy szó, de hogy mikor kell a bevezető *-t*, és mikor nem, ez az esetek nagy részében jól működik az ellenőrökben, de nem elhanyagolható részében hibáznak. Sok képzőt nem is vesznek fel a generatív rendszerbe, mert nem regulárisak, tehát nem nagyon lehet tudni, mikor alkalmazhatók. Másrészt, elemzők gyakran fogadnak (ajánlanak) nem használatos képzőformákat. (Naszódi, 2017)

Mint írtam, egyes esetekben formai módon eldönthető a kérdés. A műveltetés például *-ít* képző után mindig a *-tat*, *-tet* alakot várja. Ezt könnyen ellenőrizhetjük, ha lekérdezzük az összes *-ít* végű igét és megnézzük, mit szól az elemző/ellenőr, ha a műveltetés különböző alakjait ráakasztjuk az igékre, illetve a nyelvérzékünk tiltakozik-e valamelyik alak ellen.

Még 1993-ban így jöttem rá, hogy az *[æ]*-re végződő melléknévek ritka kivétellel<sup>2</sup> megkaphatják az *-ít*, *-[uü]* igeképzőt, ráadásul ilyenkor mindig elhagyjuk a szótővégi magánhangzót: *hülye* → *hülyül*, *hülyít*, *barna* → *barnul*, *barnít*. . . Ez a hangzó kiesés a melléknéveknél szinte mindig jelentkezik, a főneveknél szinte soha, de algoritmikusan nem kezelik a helyesírás-ellenőrök, hanem a képzett ige szerepel a szótőtárban, pedig a kategóriában reguláris szabály. Ezt az ismeretet fel is használtam a saját nyelvi adatbázisomban, illetve a helyesírás-ellenőrzőmben. Főneveknél viszont ritka kivétel, ha a szótővégi *[æ]* elnyelődik. Ha mégis, akkor felmerül a gyanú, hogy a szó valamikor melléknév lehetett.

A kérdés az, hogy mennyire ismerjük a toldalékainkat és szótőveinket, valamint mennyire lehet hasznos az ilyen egyszerű vizsgálat eredménye.

#### 5. Teszteljünk!

Az algoritmus a következő: tapasztalatunk alapján legyenek gyanús toldalékolt szavak, toldalékok. Ezeket kigyűjtjük a szótárakból, leveszünk analitikus módon egy-két morfémát, majd generatív módon, figyelmbe véve a tőváltozást is,

<sup>2</sup> A forma, féle, fajta szavaink főnévként is funkcionálnak, illetve melléknévi névutóként kezelhetők. Más kivétel is található.

rátesszünk másikat, és megnézzük mi sül ki belőle. Az előállított szóformáról, ha nincsenek sokan, magunk is dönthetünk. Ha többen vannak, akkor bízunk a meglévő elemzőkre, de teljesen sohase bízunk a gépi döntésben.

Példaként egy szintén 90-es évek eleji megoldott feladat.

Lemma: Ha az  $-[aeo\ddot{o}]/(sz/j)t$ ,  $-[aeo\ddot{o}]d$  végű igék párban vannak (*fullaszt – fullad*), az első aktív (általában tárgyas), a második passzív (általában tárgyatlan) ige. Ráadásul az aktív és a passzív ige további  $-ék$  főnévképzőt, és további  $-[eo]ny$  melléknévképzőt kaphat (*halad-ék-ony*).

Egyszerű reguláris szűréssel 117 ilyen igepárt gyűjtöttem ki. Ezekre igaz volt az állítás, sőt, a kimaradtak, melyeknek csak egyik fele volt meg, többé-kevésbé állt a tárgyasságról való feltételezésem. Vizsgálva a további todalékolást, ha nem is voltak szótári vagy legalább használatban levő szóalkotások, akkor is értelmes képzés volt: a *süllyeszt – süllyed* szópár alapján a *süllyeszték* még szakirodalomban szerepel, de a *süllyedék*, *süllyedékeny* soha elő nem fordult, mégis jó magyar szó!

Azt is vizsgáltam, hogy ha a pár egyik fele szerepelt csak a nyelvi adataim közt, a másik alak milyen a nyelvérzékemnek. Nos van egy-két szójelöltem, ami teljesen magyar, még sincs a szótárakban. Az említett képzőket a jelenlegi szpelerek nem kezelik konstruktívan, mert nem minden igehez társulhatnak. De az említett kategóriához igen.

Mellékesen egy kényes kérdésre választ kapunk, a műveltetést tisztázza az adott kategóriában: a párok  $-d$  végű alakjának az  $-szt$  végű alak a helyes műveltetett formája, és nem a  $-t?[\alpha]t$ .

## 6. Szótővadászat

Ha találtunk olyan képzőket, melyek valamire rátapadnak, akkor vizsgálhatjuk, a képzőtől megszabadított szó szótő-e. Ha megtesszük ezt a fent említett 117 esetre, akkor olyan eredményt kapunk, amivel nem tudunk sokat kezdeni. Azt tapasztaltam, hogy 9 esetben ige a szó eredendő töve, és 27 esetben névszó: főnév, melléknév vagy főnévi gyök. A maradék 80 is nyilván valamilyen ősi szavunk maradványa, de ezt egy vérbeli etimológus tudná csak igazolni, vagy ő sem. A talált  $9+27=36$  szótő viszont szerepel a szótárakban.

Érdeemes a morfológiai rendszerbe felvett képzőkkel próbálkozni. Ha levágjuk egy szótári szóról, kapunk-e a várt szófajnak megfelelő már bejegyzett szó alakalternánsát. Ha igen – és a morfológiai rendszer ereje ettől nem csökken –, a képzett alakot elhagyhatjuk rendszerünkben, mert csak a redundanciát növeli.

Próbálkozhatunk a lexikonokban nem kodifikált todalékok levágásával is. Ha már a fejezet címe **Szótővadászat**, kereshetjük az  $-[áé]sz$  képzős szavakat is, milyen szótőhöz tapadhatnak, és hogy viszonyulnak a további  $-[\alpha]t$  képzőhöz. Ha megkeressük, melyek az  $-[áé]sz[\alpha]t$  végű névszavaink, több mint 850-et találunk. Pontosabban ennél kevesebbet, mert ezek közt lesz olyan, amelyik csak formailag végződik úgy, mintha a két főnévi képzővel végződne a szó, de valójában más a morfológiai szerepe az elsőnek, például igei, csak a második főnévképző: *tenyészet*.

Ha a renitenseket kidobjuk, akkor a maradékban az első képző foglalkozást jelentő főnevet jelent, míg a második toldalék után vagy a tevékenységet, vagy a tevékenység intézményét fejezi ki a szó. A képzők többnyire főnévhez kapcsolódnak: *rákász*, *fod(o)rász*, *jogász*, de néha igéhez: *szabász*, *szülész*. Itt is kereshetünk szógyököket, melyek többnyire főnévi jellegűek: *csábász*. Ha a második főnévképzővel nem szerepel a szó, óvatosabbak legyünk. Mert *csibészet* nincs, (bár lehetne).

Ebben a példában két dolgot tisztázhattunk. Az egyik, hogy nehéz megállapítani, hogy az *-[áé]sz* mihez kapcsolódhat, tehát jogos lehet ezen képzett szavak egyedi felvétele a szótárba. De ha felvesszük, a szótárban jelezhetjük, hogy miből származtatható, mert a képző szemantikai funkciója jól következtetethető a képzésből. A másik tanulság, hogy ezek a szavak mind megkaphatják az általánosnak nem mondható *-[æ]t* főnévképzőt, és még a szemantika is származtatható algoritmikusan: az *-[áé]sz* olyan foglalkozást jelent, ami az alapszóval kapcsolatos. Fordítható úgy, hogy *a ... mestere* vagy *...-v[æ]l foglalkozó ember*: *fodrász = a (haj)fodor mestere*, *gyógyszerész = gyógyszerrel foglalkozó ember*. Az ezt követő főnévképző vagy magát a tevékenységet, vagy a tevékenység intézményét jelenti.<sup>3</sup>

## 7. Toldalékcsokok

Az eddigi példákból is kiderült, hogy az egyes toldalékok nem mindig hatékonyak az elemzéshez. Ha hasonló szavakhoz jól illeszkedő toldalékcsoportot – csokrot – választunk, akkor biztosabb a vizsgálat eredménye. A korábbi példákban az *-ít* és az *-[uü]l* kétszálú csokor volt. A toldalékokból – mert még magyarul sincsenek olyan sokan – kevéselemű csokrokat képezhetünk. Nem így, ha a toldalék előtti lehetséges formákat keressük.

## 8. Szófürtök

Ha egy toldalékcsokkal a szótári bejegyzésekből tőgyanús alakot kapunk, akkor már érdemes megnézni, valóban találtunk-e valami rejtett szót. Ha különböző szótári tételek ily módon való csonkítása azonos szóhoz vezet, akkor ezek a korábban ismert szavak egy fürtöt alkotnak. A fenti példáim közt is akad ilyen.

Szófürtöt képezhetünk kodifikált (reguláris) és ritka toldalékok csokrának segítségével is. Ez utóbbinak igéink vizsgálatánál vettem nagy hasznát. Ha megnézzük az igei szótövek végeit, sokkal korlátozottabbak (Farkas és Naszodi, 1990), mint mellékneveinknél, főneveinknél. Szinte mindegyik képzett alak. Míg névszavakat a nyelvfejlődés során egyszerűen átvettünk velünk együtt élő népektől, igéink mindig valamilyen igeképző segítségével csapódtak nyelvészletünkbe. Ez alól kivételek az ősi eredetű alapigéink: *esz(ik)*, *alsz(ik)*, *van*, *lesz*, *jön*, *megy*...

<sup>3</sup> Kivételek persze vannak. Nem minden főnevet követő *-ász* *-ész* képző ilyen: A *kolbász* szemantikailag nem foglalkozást jelent. Ha nem is találjuk egyik szótárunkban a *kolb*, *kolob*, *kalub* szavakat, van ilyen. A székelynek tartott, Benedek Elek által írásba foglalt *A kis gömböc* mese orosz változata: *Κολλοδοκ*, aminek a jelentése azonos a kerek hentesáruval, akár a magyar gömböc.

Ha pedig így van, kereshetjük, honnan származhatnak az igéink. Egyszerű mintaillesztéssel feltérképezhetjük a jellegzetes igevégződéseket – ezek alkalmazsak lehetnek újabb képzők felfedezésére is – majd levágva, vizsgálhatjuk, kapunk-e szófűrtöket. Ha jó a feltételezés, akkor a szófűrtöket a képzőhalmazok határozzák meg. Pár igevégződést kiválasztva vizsgálhatom az így kapott fűrtöt.

### 8.1. Alaki szófűrtök alkalmazásai

Formai hasonlóságon alapuló szófűrtöket gyakran használnak praktikus céllal. A sémi nyelvek szótárainak tételei szófűrtökön alapulnak: az inflexiós szabályok miatt a nálunk szokásos ábécébe rendezés miatt szócsaládok kerülhetnek távol egymástól, pedig azonos a gyökük, szótövéük. A szó szerkezetéből kihámozható szógyök ábécébe rendezése az elsődleges sorba rakási elv, aminek karakterei nem feltétlen a szó elején találhatók. Ez persze egy nehezebb mintaillesztés, mint ami nálunk használható, nem szóeleji vagy szóvégi hasonlóság keresése.

A Microsoft keresőrendszerében nyelvi támogatásként egy klaszterező rendszer volt. Főként formai ismérvek alapján társított szóalakokat. (Lehet, hogy ma is használják ezt a módszert.) A klaszterezés nyelvenként változott. Így a ragozott, képzett alakok nagy valószínűséggel egy fűrtbe kerültek. A magyar nyelvre sehogy sem működött a módszer, de latin nyelveknél kiváló volt – a keresők ebből eredő hibáját az ember felülbírálhatta.

Több kísérlet volt a világon, hogy formai ismérvek alapján építsék ki egy nyelv ragozási rendszerét. Egyik sem volt tökéletes, de hatékony. (Wicentowski, 2004) Az egyik tanítványom például nagy korpusz alapján hozott össze olasz ragozási szótárat – osztályozta a szavakat: a szövegek alapján előállított toldalékcsoportokat, majd keresett ezekhez tartozó szófűrtöket. Így szerkesztette meg a szavak teljes ragozási paradigmarendszerét. A magyarhoz hasonló összetettséggű szószervezettel rendelkező nyelv (török) morfológiájának felépítésére is sikeresen alkalmaztak hasonló módszert. (Oflazer és Nirenburg, 1999), de a török fonológiája egyszerűbb, mint a magyaré.

Ha már létezik jó ragozási beosztás, korpuszból kinyert ismeretlen szóformákat lehet besorolni (Novák és mtsai, 2003), ha sikerül jól összehozni a szófűrtöket. . . A magyar szó szerkezete és fonológiája elég összetett ahhoz, hogy automatikus rendszerbe ne bizzunk, de arra megfelel a módszer, hogy jó tanácsokat kapjunk.

### 8.2. Példa

Hogy bemutassam az eljárás hatékonyságát, a végződés-halmazt most önkényesen választom ki – az lesz a feltételezésem, hogy a következők ige-képzők. A kiválasztott igevégek:  $-[æ]n$ ,  $-[æ]nt$ ,  $-[aeoö]g$ . Egyik sem szerepel morfológiai algoritmusokban.

Az így végződő találatokból kidobva az egy szótagúakat – ezek nem lehetnek képzettek, hisz a képző is egy szótag – első vizsgálatra feltűnik pár tulajdonság.

- A  $-g$  végű ige kivétel nélkül ismétlődő passzív (szenvédő), vagy legalábbis hosszan tartó, tárgyatlan.

- Az *-n* végű ige rövid idejű, inkább passzív (általában tárgyatlan).
- Az *-nt* végű ige mindegyike rövid idejű aktív (általában tárgyas).
- Ha a fűrt teljes (mindhárom forma szerepel az eddigi szótárban), az igék alapjelentése azonos, csak paraméterei, vonzatai cserélődnek, tehát az egyik tárgya a másiknak alanya, stb. . .
- Ha megleljük a szótövet, akkor világos, ebből képeztetnek az igék.

De nincs új a nap alatt, hisz a nyelvészeket régóta izgatja a kérdés, a képzőket régóta igyeksenek feltérképezni. (Ihász, 1846) Így megállapították, hogy hangutánzó szavainkból, főleg az egytagúakból így képezhetünk igéket.

Kérdezzük le a nyelvi adatbázist, igaz-e a feltételezés, illetve mely szavaink lehetnek hangutánzók: keressük a dupla mássalhangzóra végződő főneveket, esetleg indulatszavakat, mondatszókat, határozószókat, és vessük össze azzal a fűrttel, melyek elemei a korábbi három képzőformához tartoznak. Én 47 olyan szót találtam, amely a csokorhoz részben passzol, vagyis van olyan ige a fűrtben, ami az újonnan kigyűjtött adott hangutánzógyanús szóból származik. Ezek többsége teljes toldalékcsoporthoz passzol. Sőt, ami tényleg hangutánzó, azoknak mind teljes a csokra.

Az alábbi táblázatomba belevettem *-j*-re végződő zajra, hangra vonatkozó szavakat is (felkiáltójel). Utána némi szubjektív szűréssel felvettem szótárakban nem szereplő alakokat is, hogy teljesebbé tegyem a fűrtöket. A *-g* vég helyett megengedtem az *-ng* végződést, esetenként más töv változatot. Bár gyakran az *-[æ]nt* helyett szebben hangozna az *-int* igéképző, ezeket nem jeleztem. Ezek általában más csokorhoz jobban kapcsolódnak, melyeket most nem vizsgáltam. Pár szónál kérdőjellel jelöltem a számomra bizonytalan alakokat, még ha kodifikált szótári tételek is voltak.

Nem minden szó hangutánzó, de ha teljes volt a csokor, benne hagytam, mert a jelentésen kívüli egyéb nyelvtani tulajdonságok megfeleltek a feltételezésemnek. Sok olyan hiányos csokrot is benne hagytam a listába, melyek hanghatást fednek. Ha teljes volt a paradigma, alapszavát kiemeltam. Többségben vannak! Ahol az alapszó hiányzik, ott szógyök keresendő.<sup>4</sup>

<sup>4</sup> szógyök olyan morféma, amely többnyire csak szóösszetétel első tagjaként használható, vagy úgy sem, de nyelvészeti eszközök igazolják önálló jelentéssel bíró létezését.

alapszó	passzív	aktív	ismétlő	alapszó	passzív	aktív	ismétlő
berr			berreg	böff	biccen	biccent	biceg
brekk	billen	billent	billeg	brumm	böffen	böffent	böfög
buff	brekken	brekelt	brekeg	buggy	buggyan	buggyant	bugyog
büff	buffan	buffant	buf(f)og	cam?			cammog
cin	büffen	büffent	büfög	cö			cöcög
			cincog	cupp	cuppan	cuppant	cuppog
csatt	csattan	csattant	csacsog	csepp	cseppen	cseppent	csep(er)eg
cserr	cserren		csattog	csett	csetten	csettent	csetteg
csevej!			cserreg				csicsereg
	cseveg		cseveg	csipp	csippen?	csippent	csipeg
	csillan	csillant	csillog	csirr	csirren	csirrent	csir(r)eg
csipp	csippen?	csippent	csipog	csitt	csitten	csittent	csitteg
csissz	csisszen		csiszeg	csobb	csobban	csobbant	csobog
			csivog	csossz	csosszan		csoszog
csorr?	csorran?	csorrant?	csorog	csörr	csörren	csörent	csörög
csöpp	csöppen	csöppent	csöp(ör)ög	csüccs	csüccsen	csüccsent?	
csurr	csurran	csurrant	csurog	dobb?	dobban	dobbant	dobog
dirr	dirren	dirrent	dirreg	döcc	döccen		döcög
	döbben	döbbent	döbög	dörejl!	dörren	dörrent	dörög
			dörmög	durr	durran	durrant	durrog
	duzzan		dunnyog				düb(ör)ög
			duzzog	forr	forran	farrant	forrong
			dünyög	fröccs	fröccsen	fröccsent	fröcsög
fitty	fortyan	fortyant	fityeg	gá			gágog
fütty	füttyen	füttyent	fortyog		harsan		harsog
			füttyög				hebeg
háp			habog		hersen		herseg
	herren		hápog		hörren		hörög
	hortyan		herreg	hurr?			hurrog
huh			hortyog	hüpp	hüppen		hüppög
hüm(m)			huhog	hepp			hepeg
hess	hessen	hessent	hümmög	hopp			
hipp	hippen			huss	hussan		
hupp	huppan		hipeg		kaffan	kaffant?	kaffog
kacaj!			huppog	kár			károg
katt	kattan	kattant	kacag	kipp	kippen	kippent	kipeg
ketty	kettyen	kettyent	kattog	kopp	koppán	koppant	kopog
kocc	koccan	koccant	ketyeg		korran?		korog
kukk	kukkan	kukkant	kocog	?köhej!	köhhen	köhhent	köhög
kotty	kottyan	kottyant					kunc(or)og
			kotyog		lebben	lebent	lebeg
			krárog	lob	lobban	lobbant	lobog
kurr?	kurjan	kurjant	kurrog	lotty	lottyán	lottyant	lotyog
	libben	libbent	libeg	makk			makog
loccs	loccsan	loccsant	locsog	mocc	moccan	moccant	mocorog
lötty	löttyen	löttyent	löttyög		mottyán		moty(or)og
mekk			mekeg	mukk	mukkan	mukkant	
moraj!	morran	marrant?	mozog	nyau			nyávog
	mozzan		nyafog				nyervog
nyaff	nyaffan		nyek(er)eg	patt	pattan	pattant?	nyihog
nyekk	nyekken	nyekkent	nyifog				patog
nyiff	nyiffan	nyiffant	nyik(or)og	piff	piffen	piffent	pifeg
nyikk?	nyikken		perceg		pillan?	pillant	pillog
	percen		petyeg?	pissz	pisszen	pisszent	pi(s)szeg
petty	pettyen	pettyent	piheg				pizseg
	pihen		pislog	pöff	pöffen	pöffent	pöfög
pitty	pittyen	pittyent	pityereg	pötty	pöttyen	pöttyent	pöttyög
	porcog		porcog	puff	puffan	puffant	pufog
potty	pottyán	pottyant	potyog	püff	püffen	püffent	püfög
	pörren	pörrent	pörög	reccs	reccsen	reccsent?	recseg
	prüsszen	prüsszent	prüsszög		retten	rettent	retteg
			pusmog	robaj!	robban	robbant	robog
	rebben	rebbent	rebeg	ropp	roppan	roppant	ropog
	rekken	rekcent?	rekeg	rotty	rottyan	rottyant	rotyog
	rezzen	rezzent	rezeg	röhej!			röhög
	rohan		rohog	sáp			sápag
röf(f)	röffen	röffent?	roszog		seppen	seppent	sepeg
rötty	röttyen	röttyent	röfög		settyen		settyeg
			röttyög				sistereg
	sercen	sercent	selypeg		suhan		suhog
			serceg	surr?	surran	surrant	surog
			sipeg				sustorog
slatty			slattyog				sutyorog
			sunnyog/sunyorog				
	sussan?		sus(m)og				
			suttog				
alapszó	passzív	aktív	ismétlő	alapszó	passzív	aktív	ismétlő



alapszó	passzív	aktív	ismétlő	alapszó	passzív	aktív	ismétlő
sutty	suttyan		suty(or)og		szeppen		szepeg
		szipant	szipog	szísz	szíszzen	szíszent	szisz(er)eg
	szortyan		szortyog				szörcsög
			szörtyög	szusz	szusszan	szusszant	szuszog
szotty	szottyán			tipp	típpen	típpent?	típeg
toccs	toccsan	toccsant?	tocsog	topp	toppan	toppant	top(or)og
totty	tottyán		totyog	tőf(f)			tőfög
trapp			trappog		tüsszen	tüsszent	tüsszög
			vacog	vakk?	vakkan	vakkant	vakog
			varcog		vartyan		vartyog
			vernyog				vicsorog
			vigyorog		villan	villant	villog
			vijjog				villog
			vinnyog	zaj			zajo(n)g
zizz?	zizzen	zizzent	zizeg	zok			zokog
zökk	zökken	zökcent	zökög	zörej!	zörren	zörent	zörög
zötty	zötttyen	zötttyent	zöttyög				zub(or)og
	zuhan		zuhog	zümm			zümmög
zupp	zuppan?			zsivaj!			zsenyeg
zsibaj!			zsibo(n)g				zsivo(n)g
	szíszzen		szízseg				
alapszó	passzív	aktív	ismétlő	alapszó	passzív	aktív	ismétlő

## 9. Egyéb szótani rejtelmek

A példaim főként arra szolgáltak, hogy algoritmizálhassunk eddig figyelembe nem vett képzőket. Nem csak erre jó. Most felsorolom, én hol vetném be a magyar nyelv esetén:

- Összetételek keresése: szótári tételek akár szó eleji, de inkább szó végi egyezése alapján képeznék szófürtöket, melyeknek szófürtképző alakjai szintén szótári tételek.
- Új képzők keresése: ha alaki azonosság alapján találunk olyan fürtmeghatározó szóvégcsoportot, amelyhez nagy szófürt tartozik, feltehetően a fürtképző karakterláncok toldalékokból állnak.
- Szótőkeresés: ha az előző módszer toldalékcsokeihoz tarozó szójelölt többsége valóban helyes, akkor a szótőként korábban nem értelmezett fűrtelemek, (az a karakterlánc, melyre a toldalék elemei csatlakoznak) vizsgálható, jó-e szótőnek, esetleg gyöknek.
- A fenti két módszer iteratív alkalmazásával akár ismeretlen nyelv morfológiáját is felfedhetjük, vagyis kialakíthatunk teljes toldalékcsoportokat, és megállapíthatjuk az ezekhez tartozó szótőfürtöket, vagyis az egy toldalékolási paradigmához tartozó szócsoportokat.<sup>5</sup>
- Szógyökkeresés: egy időben gyűjtöttem az élő szógyököket. Élő a szógyök, ha önállóan nem, legfeljebb ragozva, képezve, egyébként összetételben fordul elő (*gyógyot, gyógykezelés, gyógyul*). Halott, ha csak képzett alakban van jelen (*segít, segély*).

Ezekkel a módszerekkel sok egyéb nyelvi finomság is napfényre kerülhet.

Had emeljem ki: az esetek többségében nem szövegtörzseket használok, hanem nyelvi, szótárjellegű törzseket. Ezekből nyerek az információkat, hogy

<sup>5</sup> <http://wordlist.aspell.net/agid-readme/>

visszacsatolva javítsam a minőséget. Ez szemben áll a mai gyakorlattal, mert szószinten azt tekintik a nyelvészek bizonyítéknak, ha valami elő is fordul. Ezzel szemben a valóság az, hogy helytelen dolgok is előfordulnak, de a nyelv része az is, ami még soha le nem lett jegyezve, de akár lehetne is. Mondatszinten ez természetes, hisz a helyes mondatok variációja akkora, hogy számba sem lehet venni. Az aglutináló nyelveknél szószinten is igaz az az elv, hogy nem csak az van, ami elfordult valamikor, tehát nyoma van korpuszban. Nem csak azok a szerkezetek léteznek, melyeket eddig felfedtünk. Papp Ferencet kéne idéznem, de nem tudom mikor és hol mondta. A pontos szövegére sem emlékszem, de a lényege a következő: *Mi tartozik a magyar nyelvhez? Az a mi valamikor elhangzott vagy leírták? Vagy az is, ami el fog hangzani? Vagy amit valaki akár el is mondhat?* A lényeg ebből a szempontból az, hogy szótanunk kellően összetett, hogy olyan eszközöket használjunk, amelyeket sok nyelvben csak mondatannál alkalmaznak.

A 90-es évek elején vizsgáltam, hogy a nyelvészek a különböző töváltozatokat hogyan osztályozzák. Toldalékolásnál, ragozásnál fontos ismérv, hogy mely toldalék melyik alakja melyik tőalternánshoz kapcsolódik. Ebből a szempontból azonosnak bizonyult a szóbelseji magánhangzó-rövidülés (*madár, madarat*) és a belső hangzókieésés (*szatyor, szatyrot*), sőt, az úgynevezett mássalhangzó-átvetés (*teher, terhet*) teljesen egybeeső paradigma.<sup>6</sup> A ragozás algoritmusait teljesen egységesen fogalmazható meg ezekben a nyelvészek által különválasztott kategóriákban. Némi fenntartással a *-v* betoldásával járó töváltozás is ide sorolható (*daru, darvat; tó, tavat*), de itt vannak apró különbségek.

Akár meglevő morfológiai szótárak fésülése is alkalmazhatjuk az eljárást.<sup>7</sup> Még sok ötletem van. Utolsó a mifantológusoknak, történeti nyelvészeknek ajánlom figyelmébe, hogy használják bátran az egyszerű szűrést. Ily módon esett le nálam a tantusz, hogy a *tavas* szavunk minden bizonnyal a *tó* képzett alakja. Nyelvészekről (is) hallottam, hogy a szóhasználat következtében főnév melléknévvé, melléknév főnévvé változhat az idők során. A legszebb példát Prószyka Gábortól: a *róka* állítólag eredetileg melléknév volt, míg a *ravas* főnévként kezdte nyelvünkben az életét. És mindkét szó ugyanonnan származik. Az egyik a *-ka* kicsinyítő képzőt kapta, a másikat az *-asz* képző díszíti. Ezek szerint mindkettő

<sup>6</sup> Elekfi László ragozási szótárában az 5-ös, 7-es és a 9-es névszói fő osztályok egy kalap alá vonhatók. Ezek a Hunspell adatbázisában is külön osztályt képeznek, pedig a formalizmus lehetővé tenné az egységes kezelést. Ha a hosszú magánhangzót két-tőzött karakterrel jeleznénk, mint ahogy a finnek teszik, triviálissá válna, hogy nincs különbség a hangzókieésés és a hangzórövidülés között.

<sup>7</sup> A magyar helyesírás-ellenőrzők hőskorából származik az a történet, hogy – mivel az értelmező kéziszótár sem tartalmazott mindent – gyakran bővítettük a tőtárat. Az egyik ilyen volt a *csevej*. Pontosabban, nem találtuk, hogyan kell írni, és a döntés a *csevely* alak lett. Pár hónapig ebben hitték a fejlesztők, és a felhasználókat is erre irányította a program. Az irodalomban, mint kiderült, mindkét forma előfordult, mert első látásra nem triviális. Emiatt, ha akkor lettek volna nagy digitális korpuszok, akkor sem tudtunk volna dönteni. Most viszont, a (8.2.) fejezet táblázatába foglalt szavak mutatják: ha az *-[a]j* véget levágjuk, és a megmaradt részre az *-[aeoö]g* toldalékot tesszük, akkor jó igét kapunk. Ez az *-ly* végű főnevekre nem áll.

töve a *ró* főnév, csak az egyik tőváltozás nélküli, a másik egy gyakori *-v* kötéssel jön létre. Ilyen főnevünk nincs, de a zürjén *rutjs*, a cseremiszi *revezs*, a mordvin *rives*, a finn *repo rókát* jelent (Tótfalusi, 2013).

## 10. Összefoglalás

A számítógépes nyelvészetben gyakran egyszerű alaki tulajdonságok rejtnek fontos nyelvi információkat. A morfológia esetén, már régóta használnak egyszerű formai szűréseken alapuló osztályozásokat, elemzéseket. Ennek egyik kiterjesztése, ha nem mintákkal, hanem mintafürtökkel szűrünk. Ezzel a kapott szófürtökről pontosabban dönthetünk.

A módszer általánosan alkalmazható, és sok helyen lehet hasznos. Én csak pár példát mutattam meg, de bátorítom a nyelvészeket, éljenek a számítógépek lehetőségeivel ötletesen, mert nagyon megéri.

## Hivatkozások

- Elekfi, L.: Magyar ragozási szótár. MTA Nyelvtudományi Intézete (1994)
- Farkas, E., Naszódi, M.: A toldalékok 32 fonológiai osztálya. In: Magyar nyelvű mondatok elemzése természetes nyelvű interfész céljából. p. 44. MTA SzTAKI (1990), <http://www.cs.bme.hu/~naso/langeng/manyel.pdf>
- Ihász, r.: Igeképzők. In: Magyar nyelvtan. p. 213-220 (1846)
- Koskenniemi, K.: Two-Level Morphology: A General Computation Model for Word-Form Recognition and Production. No. 11 in PUBLICATIONS, University of Helsinki, Department of Linguistics (1983), <http://www.ling.helsinki.fi/~koskenni/doc/Two-LevelMorphology.pdf>
- MTA, N.I.: Magyar értelmező kéziszótár. Akadémiai Kiadó (1972)
- Naszódi, M.: A magyar helyesírás-ellenőrzők mai állása. In: XIII. Magyar Számítógépes Nyelvészeti Konferencia. p. 347-354. Szegedi Tudományegyetem (2017), <http://www.cs.bme.hu/~naso/langeng/SpellsSate20016.pdf>
- Novák, A., Nagy, V., Oravecz, C.: Magyar ismeretlenszó-elemző program fejlesztése. In: Magyar Számítógépes Nyelvészeti Konferencia. p. 45-57. Szegedi Tudományegyetem (2003)
- Oflazer, K., Nirenburg, S.: Practical bootstrapping of morphological analyzers. In: Conference on Natural Language Learning (1999)
- Papp, F.: A magyar nyelv szóvégmutterató szótára. Akadémiai Kiadó (1969)
- Proszéki, G., Kornai, A.: Papp Ferenc és az újrafelhasznált Szóvégmutterató szótár (2017), <https://itf.njszt.hu/objektum/papp-ferenc-es-az-ujrafelhasznalt-szovegmutterato-szotar>
- Tótfalusi, I.: Magyar etimológiai nagyszótár. Arcanum (2013), <http://www.szokincshalo.hu/szotar/>
- Wicentowski, R.: Multilingual Noise-Robust Supervised Morphological Analysis using the WordFrame Model. In: ACL Special Interest Group on Computational Phonology (SIGPHON) (2004), <https://www.aclweb.org/anthology/W04-0109.pdf>

# A Note on Lost Morphemes The Benefits on Surface Similarities

Mátyás Naszódi, e-mail: [naszodim@morphologic.hu](mailto:naszodim@morphologic.hu)

MorphoLogic

**Abstract.** The current morphological analyzers have been designed pragmatically for practical purposes. Their goal is to cover the word forms in Hungarian texts with relatively little effort and with as few mistakes as possible. Once the goal has been achieved, regular case affixes, marks, and verbal conjugation endings are well described in a formal way, but most derivative affixes and rare case suffixes are treated individually as exceptions.

In my research, I found that there are far fewer exceptional word forms in Hungarian. By clustering word forms by their endings, new relationships, new roots, new morphemes can be discovered that are missing from earlier databases. By clustering word forms by their endings, new relationships among roots, morphemes can be discovered that are missing from earlier databases. One can simplify morphological descriptions without limiting their power. Even a complete morphological description of an unknown language can be generated based on a large corpus solely. Moreover, if not only similarities of endings, but clusters of ending patterns are used to group word forms, then many hidden word roots and suffixes can be discovered that have been forgotten altogether, or mentioned only by descriptive linguists.

As a result of the method, semantic dependences might be discovered, and linguistic collections, databases made for practical purposes can be corrected, improved as well.

**Keywords:** vocabulary, morphology, lexicography, spelling