Milyen a még jobb Humor?

Novák Attila[1] és M. Pintér Tibor[2]

[1]MorphoLogic Kft., 1126 Budapest, Orbánhegyi út 5.,
novak@morphologic.hu
[2]MTA Nyelvtudományi Intézete, 1068 Budapest, Benczúr u 33.,
tpinter@nytud.hu

Kivonat: A számítógépes nyelvfeldolgozás egyik alapvető és általában nélkülözhetetlen eszköze a morfológiai elemző. Az elemzőnek képesnek kell lennie az összes produktív szóalaktani jelenség (ragozás, képzés, szóösszetétel) kezelésére. Bár a lexikai többértelműség viszonylag gyakori jelenség, a többértelmű szavak különböző lehetséges elemzéseihez gyakran nagyon különböző valószínűség rendelhető. A morfológiai elemzőre épülő alkalmazások hatékony működése, illetve bizonyos esetekben helyes funkcionalitása szempontjából is hasznos, ha a lexikai többértelműségek körét sikerül csökkenteni. A valószínűtlen elemzések kiszűréséhez az utóbbi hónapokban szisztematikus korpuszalapú vizsgálatot végeztünk olyan morfológiai jelenségekkel kapcsolatban a magyarban, amelyek rendszeresen vezetnek elemzési többértelműségekhez. Cikkünkben ezeket a vizsgálatokat és az elvégzett lexikográfiai munkát mutatjuk be.

1 Lexikai többértelműségek

A magyarhoz hasonlóan bonyolult morfológiájú nyelvek számítógépes feldolgozása során a nyelvben előforduló lehetséges szóalakok igen magas száma miatt a morfológiai elemzés alkalmazása gyakorlatilag elkerülhetetlen. A morfológiai elemzőnek képesnek kell lennie az összes produktív szóalaktani jelenség (ragozás, képzés, szóösszetétel) kezelésére.

Egy morfológiai elemző kimenetét szemlélve feltűnik, hogy a lexikai többértelműség (azaz az a jelenség, hogy egy szóalaknak egynél több lehetséges elemzése van) viszonylag gyakori jelenség. A többértelműségeket leggyakrabban homonímiák, a különböző paradigmák véletlenszerű vagy rendszerszerű átfedései, illetve a paradigmán belüli rendszeres átfedések okozzák. A magyarban például rendszeres többértelműségre vezetnek az alábbi paradigmatikus átfedések:

az igei paradigmán belül pl.:
vettem, mostam (*én valamit* vagy *én azt*)
vennétek, mosnátok (*ti valamit* vagy *ti azt*)
vennék (*én valamit* vagy *ők azt* – elöl képzett harmóniájú igék esetében)
eszik (*ő* vagy *ők azt* – elöl képzett harmóniájú tárgyas ikes igék esetében)

˜ *néztek*[19] (*ti most* vagy *ők akkor* – elöl képzett harmóniájú igék esetében)
˜ *festette* (*ő azt akkor* vagy *ő azt valakivel akkor* (műveltető) – csak msh.+t tövű elöl képzett harmóniájú igék esetében)

a névszói paradigmán belül:
˜ gyerekével (*az ő gyerekével* vagy *a gyerek valamijével*) – csak az elöl képzett tövű szavak esetében, de minden tővégnyúlást kiváltó esetraggal

A fenti paradigmán belüli többértelműségek és az egyedi tőhomonímiák mellett (melyek gyakran számos toldalékolt alak egybeesésével is együtt járnak: pl. *vár(nak), nyúl(nak)* stb.) sok olyan eset is van, ahol különböző szavak paradigmáinak csak egyes tagjai esnek egybe, a lemma nem:

mentek (*én valamit* (lemma=*ment*), *ti most* vagy *ők akkor* (lemma=*megy*))
csend (*főnév* vagy *te azt*)

Bár a fenti példák esetében különböző valószínűség rendelhető a többértelmű szavak különböző lehetséges értelmezéseihez (elemzéseihez): pl. a *csend* szóalak jóval gyakrabban fordul elő főnévként, mint igealakként, és a sima birtokos alakok (*gyereke*) is gyakoribbak az anaforikus birtokos alakoknál (*gyereké*), a felsorolt példák mindegyike minden említett értelmében viszonylag gyakran előfordul magyar nyelvű szövegekben.

2 Valószínűtlen elemzések

Vannak azonban olyan esetek is, amikor lehetséges értelmezések valamelyikének valószínűsége a többi elemzéséhez képest lényegében elhanyagolható. A korábbi példákban szereplő *vettem* szóalak elvileg elemezhető a *vesz* ige befejezett melléknévi igeneve (*vett*) birtokos alakjaként is, valójában azonban a befejezett melléknévi igenevek birtokos alakjai nemigen használatosak (tehát az adott elemzés valószínűsége szinte zérus, bár grammatikailag elvileg lehetséges: *–Van egy nyers gyémántom. – Na és van csiszoltad is?*).

Hasonló módon a *mentek* szóalak is lehet a *megy* ige befejezett melléknévi igeneve (*ment*) többes számú alakja is, azonban talán ez az elemzés is csak kevéssel valószínűbb, mint a fenti *vettem* alak melléknévi igeneves elemzése.

Amellett, hogy bizonyos általános nyelvi konstrukciók a gyakorlatban soha nem fordulnak elő (mint pl. a befejezett melléknévi igenevek birtokos alakjai), vannak más esetek is, amikor az elemző által több morfémából összeállított egyes elemzések valószínűsége lényegében elhanyagolható az alternatív elemzésekéhez képest. Ilyenek egyrészt azok az esetek, ahol a sokmorfémás elemzés kompozicionálisan kiszámítható jelentése abszurd, másrészt azok, ahol az alternatív elemzés olyan lexikalizált jelentésű tövet tartalmaz, amelyhez képest – bár a kompozicionális elemzés jelentése

[19] A *néztek* és *festette* típusú többértelműségek csak a nyílt *e*-zárt *ë* megkülönböztetést nem ismerő standard nyelvváltozatban (és persze az írott nyelvben) állnak fenn. Az *ë*-ző dialektusban ezek a szavak nem többértelműek: *néztek (ők akkor)-néztëk (ti), fëstëtte (ő azt akkor)-fëstette (ő azt valakivel akkor).*

korántsem képtelenség – előfordulási valószínűsége mégis elhanyagolható. Ezek azok az esetek, amelyekben az anyanyelvi beszélőben a valószínűtlen elemzés lehetősége általában fel sem merül, a (jelentés-összetevőt nem tartalmazó) formális nyelvi modellt alkalmazó gépi algoritmus számára ezek az elemzések azonban éppoly lehetségesek, mint az összes többi.

A morfológiai elemzőre épülő alkalmazások hatékony működése, illetve bizonyos esetekben helyes funkcionalitása szempontjából is hasznos, ha a lexikai többértelműségek körét sikerül csökkenteni. Ehhez célszerű a morfológiai elemzőben implementált produktív szóalaktani folyamatokat úgy megszorítani, hogy a valószínűtlen elemzések minél nagyobb körét ki tudjuk zárni, lehetőleg anélkül, hogy ugyanakkor sok érvényes elemzést elveszítsünk. A MorphoLogic Humor elemzőprogramjához készült morfológiai lexikonok készítéséhez az utóbbi években használt morfológiaiadatbázis-készítő keretrendszer [3] jegyalapú formalizmusa lehetővé teszi az egyes morfológiai jelenségek produktivitásának pontos szabályozását, illetve az elemző is tartalmaz olyan mechanizmusokat, amelyekkel bizonyos elemzések vagy azok alternatívái már a morfológiai elemzés szintjén kiszűrhetők.

A valószínűtlen elemzések kiszűréséhez szisztematikus korpuszalapú vizsgálatot végeztünk számos olyan morfológiai jelenséggel kapcsolatban a magyarban, amelyek rendszeresen vezetnek elemzési többértelműségekhez. Mivel mind a lexikalizálódás, mind a komponált jelentés abszurditásának lehetősége elsősorban a szóképzés és a szóösszetétel körében áll fenn, vizsgálatainkat elsősorban ebben a körben végeztük.

3 Többértelműségek a képzett szavak körében

A magyar képzőkészlet egyes elemei önmagukban is meglehetősen sok többértelműséget vetnek fel. Az alábbi képzők például mind önmagukban is többértelműek:

-s	*harcosak, barackosak* (melléknév), *harcosok, barackosok* (főnév)
-ó	*abban bizakodó* (melléknévi igenév), *bizakodóak* (melléknév), *ablakmosó* (főnév), *ablakmosó gép* (itt jelzői helyzetben)
-z(ik)	*ftp-zik, (le)ftp-z: ftp-zett/ftp-zett le*
-(t)at(ik)	*kihirdettet, kihirdettetik: kihirdettetett*

Ráadásul néhány képzősorozatnak látszó képződmény önálló képzőként önálló jelentéssel is rendelkezik. Ilyenkor a több képzőként való elemzés gyakran (bár nem minden esetben) szintén nagyon valószínűtlen (hibás):

nyávogós:	1. *nyávog+ó+s* (amiben van nyávogó)	2. *nyávog+ós* (hajlamos a nyávogásra)
katonáskodik	1. *katoná+skodik* (katonaként tölti az idejét)	2. ?*katoná+s+kodik* (?katonásan tölti az idejét)
elmagyarosodik	1. *magyar+osodik* (=egyre magyarabb lesz)	2. ?*magyar+os+odik* (=?egyre magyarosabb lesz)

Ezek a többértelműségek (még ha nyelvészetileg megalapozottak is), komoly hatékonysági problémákhoz vezethetnek a morfológia kimenetére épülő szintaktikai elemzés szintjén, különösen azokban az esetekben, ahol több többértelmű képző

együttes előfordulása esetén a többértelműségek összeszorzódnak. Ezért egyik célunk a képzett alakok potenciális többértelműségének csökkentése volt. Módszerünk elsősorban korpusz alapú vizsgálatokon alapult. Korpuszként a Szószablya projektum keretében létrehozott Webkorpuszból [1], [2] készült szóalak-gyakorisági listát használtuk, amelyet a Humor elemzőprogrammal elemeztünk.

3.1 Az *-s* képző

Elsőként az *-s* képző korpuszbeli előfordulásait vizsgáltuk. Feltételezésünk az alábbi volt: a melléknévképző *-s* lényegében teljesen produktív, a főnévképző *-s* azonban nem igazán az: bár sok foglalkozásnév *-s* képzős (órás, lakatos stb.), illetve sok növénynév *-s* képzős alakja használatos az adott növénnyel benőtt terület jelölésére (kukoricás, akácos, gyümölcsös stb.), de teljes körű produktivitásról még ezeken a zárt szemantikai osztályokon belül sem beszélhetünk. Ezért megvizsgáltuk a nem nyitótőként viselkedő *-s* képzősként (is) elemezhető szavakat. Célunk a lexikalizálódott *-s* képzős főnevek kiszótárazása volt, hogy az *-s* képző többértelmű elemzését megszüntethessük. A vizsgálat eredményeképpen meggyőződtünk róla, hogy kiinduló hipotézisünk téves volt: a nem nyitótőként ragozott *-s* képző éppolyan produktív, mint a nyitótőként viselkedő, mert '...-os ember' értelemben éppoly tág körben használható, mint általában a melléknévképző *-s*. Ugyanakkor arra jutottunk, hogy míg egyes alkalmazásokban érdemes a nyitótőként viselkedő (melléknévi) és a nem nyitótő (főnévi) *-s* képzőt megkülönböztetni (például a szóalak-generátorban, ahol így a melléknévi és főnévi *-s* képzős szavak generált alakjai jól elkülönülnek), addig más alkalmazásokban (például a magyar–angol fordítóprogramban) nem érdemes megtartani ezt a megkülönböztetést, mert az *-s* képzős alakok fordításakor a nyitótőség nem játszik szerepet, hanem az számít, hogy az *-s* képzős szerkezet jelzői vagy állítmányi szerepet tölt-e be:

a sárga sisakos férfi	the man with a yellow helmet
a férfi sárga sisakos	the man has a yellow helmet
a sárga sisakosok/sisakosak	the ones with yellow helmets

3.2 Az *-ó* képző

Az *-ó* képzővel kapcsolatban a következő megállapításokat tehetjük:

1. Mivel a melléknévi igenevek mind az alapige vonzatkeretének nagy részét öröklik, mind az igék egyéb bővítési lehetőségeivel rendelkeznek általában, ezért mindenképpen érdemes különválasztani azokat az eseteket, ahol melléknévi igenévi elemzés kizárható. Ebbe a körbe egyértelműen csak a *főnév+ige+Ó* alakú *-ó* képzős szavak tartoznak. Konkrétan ez a szerkezet azonban a morfológiai elemző kimenete alapján akkor is egyértelműen megkülönböztethető, ha az ezt a konstrukciót megtestesítő szavakban semmilyen speciális módon nem annotáljuk magát az *-ó* képzőt.
2. A *főnév+ige+Ó* alakú *-ó* képzős szavak önálló főnévként szerepelhetnek (pl. *kőaprító*), ugyanakkor az érvényes helyesírási norma megengedi azt is, hogy ezek a szavak mintegy jelzőként külön írt szóként épüljenek be nagyobb névszói szerkezetekbe (pl. *kőaprító gép*). Ettől persze ezek a szerkezetek nem lesznek jelzős

szerkezetek, mindenesetre a morfológiai elemző kimenetét felhasználó mondattani elemzőnek képesnek kell lennie ezeknek a szerkezeteknek a kezelésére is. Ehhez azonban nem szükséges, hogy ezeket a szavakat kétféleképpen annotáljuk, mert nem a mondattani elemző általános jelzős szerkezeteket leíró mintája, hanem egy specifikus *főnév+ige+Ó főnév* alakú minta kezeli őket. Ugyanakkor számos ilyen alakú szó lexikalizálódott melléknév (pl. *bizalomgerjesztő*).

3. Az *-ó* képzős melléknevek a főnevekkel és a melléknévi igenevekkel ellentétben általában nyitótőként viselkednek (ugyanakkor a tényleges nyelvhasználatban ez opcionális: *a drogériák bizakodók*), képezhető belőlük határozószó (*bizakodóan*), fokozhatóak (*bizakodóbb*), állhatnak állítmányként. Az egyértelműen nyitótövet tartalmazó szóalakok, a határozói és fokozott alakok csak melléknevek lehetnek.
4. Amennyiben minden *-ó* képzőt azonos módon annotál a morfológiai elemző, akkor az egyetlen megkülönböztetés, ami az elemző által szolgáltatott jegyekből nem rekonstruálható, az egyértelműen nyitótövet tartalmazó szóalakok határozott melléknévi besorolása. Ebben az egy esetben a többi esettől eltérő annotációt kell alkalmazni. A többi esetben elegendő ha egyetlen elemzést ad a morfológiai elemző, a szintaxisnak ugyanakkor ebben az esetben ezt az egy elemzést az adott szókonstrukciónak megfelelő módon kell értelmeznie.

4 Valószínűtlen összetételek

A többértelmű névszóképzőkkel kapcsolatban elvégzett vizsgálatok ugyanakkor rengeteg egyéb problémára vetettek fényt. Ezek közül kiemelten kezelendőek a mára lexikalizálódott *-s* képzős főnevek, amelyek produktív *-s* képzős elemzése egyéb alkalmazásokban problémát okozhat. Fordítóprogramban problematikusak lehetnek például az *anyós=anyó+s, mártás=Márta+s, csapás=csapa+s, emlős=emlő+s* alakok, amelyek képzett szóként történő kezelése hibás fordítást eredményezhet. Sajátos problémaként jelentkeznek az olyan esetek, amikor a szó egyszerre elemezhető *-s* képzős, nyílt szótagra végződő tőként, illetve *-ás/-és* képzős zárt szótagra végződő tőként: *kígyómarás=kígyó+mar+ás ~ kígyó+mara+s, adás=ad+ás ~ Ada+s*, *csapás=csap+ás ~ csapa+s*.

A formális elemzések és a szemantika találkozásakor létrejövő kompozicionális és derivációs sajátosságok közül említést érdemelnek az abszurdnak tűnő jelentésű (egyébként morfológiailag normális) összetételek. Ilyen összetételek szép számmal előfordultak már az *-s* képzős szavaknál is. Kedvenc példáink között tartjuk számon az alábbi szavakat: *anyósom=anyó+som ~ anyós+om, apósom=apó+som ~ após+om, altatás=altat+ás ~ al+tatás* (ilyenek még: *alanyuk=alany+uk ~ al+anyuk, alapjuk=alap+juk ~ al+apjuk*). A felsorolt példák lényege, hogy bennük a többértelműség úgy jön létre, hogy az elemzés egyik eleme összetett szó (ami a „túlelemzett" alak), a másik egyszerű, toldalékos szó. Ennek az alcsoportnak további sajátossága, hogy az összetett alakok mindkét tagja fogalmi jelentéssel bíró „tartalmas" szó (általában főnév). Sajátos csoportot alkotnak azonban azok az összetételként viselkedő szavak, amelyek második tagja hangutánzó szó, indulatszó vagy valamilyen szóteremtéssel keletkezett határozott fogalmi tartalom nélküli szó: *lényekké=lé+nyekk+é, farkukkal=far+kukk+kal.* Egy következő csoport tagjai azok az összetételek, amelyek a mai standardban nem használatosak, egyértelmű archaizmusok vagy historizmusok,

s elfogadásuk csak nehezítené a program működését: pl. az elavult *szaka* szóval képzett összetételként: *korszaka=kor+szak+a ~ kor+szaka*.

Mindhárom csoport kezelése szerencsére egyszerűen – és más-más módon – megoldható még a morfológiában. A morfológiai elemző tartalmaz egy olyan mechanizmust amelynek segítségével egy adott morf többmorfémás alternatív elemzései letilthatók. Az első csoport esetében (*anyósom*) ezt a mechanizmust használva oszthatatlanná lehet tenni a toldalékos szó szótövét (*anyós*), így elkerülhető a szemantikailag hibás vagy nem normatív alakok keletkezése. Más megoldást kíván a második csoport (*lényekké*): mivel szemantikailag nem teljes értékű szavak alkotják az összetétel második tagját, amelyek egyébként a nyelvhasználatban sem alkotnak összetételeket, ezért ezeket olyan jeggyel láttuk el a szótárban, amely előírja, hogy nem állhatnak összetételek második tagjaként (és első tagként sem). Ugyanez a megoldás a harmadik csoport (*szaka*) esetében is alkalmazható. Ezek esetében azonban akár azt a megoldást is választhatjuk, hogy mivel ezek a szavak a mai nyelvhasználatban már csak ritkán fordulnak elő (ezt mutatják a korpuszalapú vizsgálatok, többek között az Értelmező kéziszótár második kiadása is), ezért – a duplicitások elkerülése végett – nem vesszük fel őket a szótárba.

Egy további csoportot alkotnak azok az alakok, amelyek rendszeres többértelműsége a következőképpen alakul: egyik lehetséges elemzésükkor szerves részét alkotják a névszói ragozási paradigmának, másik elemzésükben ugyanakkor összetételként szerepelnek (mivel szótövük utolsó szótagja egybeesik valamelyik névszóraggal, ugyanakkor a toldalékékként szereplő szótag előtti rész is értelmes magyar szó). Tipikus példák az *ének*, *ében*, *ára, inak, imre, kánon, román, szemét stb.* végződésű szavak: *telepének=telep+é+nek ~ telep+ének*, *szerepében=szerep+é+ben ~ szerep+ében*, *csatára=csatár+a ~ csat+ár+a*, *tanulóinak=tanuló+i+nak ~ tanuló+in+ak*, *Pestszentimre=Pest+szent+imre ~ Pest+szent+i+m+re*, *misekánon=mise+kánon ~ mise+kán+on*, *nagyromán= nagy+román ~ nagy+roma+n*, *fémszemét= fém+szemét ~ fém+szem+é+t*. A felsoroltak egy része jól kezelhető a morfológiai elemzőn belül (pl. az *ének* és *ében* főnevek esetében előírhatjuk, hogy összetételben nem állhatnak elöl képzett harmóniájú tövek után, a ténylegesen előforduló ilyen eseteket pedig felvesszük a szótárba), más részük kezelése már túlmutat a morfológián, egyértelműsíteni ezeket leginkább szintaktikai szűrőszabályokkal lehet: pl. a tömeges *-ára* végű többértelműségeket a mondattan szintjén úgy is kiküszöbölhetjük, hogy nem kell a morfológiában az *ár* végű összetételek produktív képzését teljesen letiltani: csak a főnév+*ár+a*[PSe3] elemzéseket érvényteleníti egy esetleges *...a+ra*[SUB] elemzés. Ily módon az ilyen típusú morfológiai túlelemzések az elemzett, vagy fordított szövegben már nem jelennek meg.

Néhány további példa szemantikai anomáliákat mutató túlelemzésekre:

vadbarom=vad+bar+om,
anyagcsere-állapot=anyagcsere+-+ál+lap+ot,
ultramarinkék=ultra[FN]+mar[FN]+i[_IKEP]+nk[PSt1]+ék[FAM]+[NOM]
korcsmáros=korc+smár+os

Ezek az elemzések a magyar alaktan formális szabályainak megfelelnek, azonban a magyar nyelvi kompetenciával rendelkező beszélők ezeket a szavakat az esetek szinte kizárólagos többségében a magyar nyelvközösségben normatívvá vált jelentésben használják. Mivel a program elemzéskor az összes lehetséges morfemikus határon szegmetál (és nem rendelkezik kommunikatív kompetenciával), ezért

esetenként „túlelemzéseket” hoz létre. Korpuszalapú elemzések során csak a főneves összetételeknél több, mint 220 olyan típust találtunk, amelyek a fentiekhez hasonló morfológiai többértelműséghez vezetnek – pontosabban vezettek, mivel a vizsgálat közben felmerült jelenségeket időközben javítottuk. A következőkben ezek közül, a már kijavított esetek közül mutatnánk be párat, természetesen a teljesség igénye nélkül:

gyönyörhullám=gyönyör+hulla[FN][PSe1][NOM],
nőcinek=nőci+nek ~ nő+cin+ek,
gleccserhasadék=gleccser+hasadék ~ gleccser+has+ad+ék,
kakasukkal=kaka+sukk[FN][INS][20],
tengerfenék=tengerfene[FN][PL].

A „túlelemzések” között nem csak összetett főnevek vannak:

fémdoboz=fém+dob+oz, *combizom=combizik[IGE][Te1]*

Az utolsó példa jól szemlélteti a szótáralapú elemzés (és generálás) problémáját: a szótövesítéskor keletkezett szótő (ami attól szótő, mert a szótár tartalmazza) és a hozzá kapcsolódó ismert toldalékok kapcsolatát a program mindaddig „valós” szóként értelmezi, amíg bizonyos eseteket vagy szabályokat meg nem tiltunk neki (például, hogy a *dob* főnévből képzett ige nem lehet *-z* képzős, amit legegyszerűbb módon a doboz további részekre elemzésének letiltásával érhetünk el).

5 Igei többértelműségek

Az igék esetében igazából két nagy problémakör merült fel, amelyek mélyebb elemzések után további alkategóriákra oszlottak.

5.1 Az ikes–iktelen többértelműség

Az egyik felmerülő problémakör az igék ikességét érintette. Mivel az egymástól csak ikességben különböző igék paradigmája a szótári alak kivételével (mely az ikes igék esetében *-ik*re, a nem ikesek esetében Ø-ra végződik) általában teljesen egybeesik (bár az igepár ikes tagja gyakran nem tárgyas), ezért az ikes–iktelen párok a lexikai többértelműségek egyik legmasszívabb forrását adják, különös tekintettel arra a tényre, hogy a *-z* képzőnek mind az ikes mind az iktelen változata rendkívül produktív.

Az ikesség elkülönítésére szabály alapú mechanizmust sajnos nem lehet kialakítani. Ehelyett az elemzett korpuszból kiválogatott szóalakokat tartalmazó listákat kellett átnéznünk, és az abban látottak alapján szabályokat felállítani, amelyeket a listán lévő igékhez rendeltünk. Fontos megemlíteni, hogy mint azt a cikk elején felsorolt paradigmán belüli többértelműsége között már említettük, az *-ik* toldalék homonim alak, azaz egymástól független jelentésben több típusú *-ik* is él a magyar nyelvben: egyrészt mint kijelentő mód, jelen idő, cselekvő, egyes szám harmadik személyű

[20] Az Éksz[2] szerint a *sukk* nem más, mint “a két ököl és az egymás felé fordított két hüvelykujj együttes hossza mint (ácsok használta) hosszmérték”.

alak (ő *eszik* valamit), másrészt mint kijelentő mód, jelen idő, cselekvő, határozott tárgyas, többes szám harmadik személyű alak (ők *eszik* azt). Korpuszalapú vizsgálatunkban (amely 10 805 db igealakot ölelt fel) természetesen az első esetből indultunk ki, ami az igék esetében a szótári alak. Ez alapján az ikes–iktelen párokat átvizsgáltuk a valószínűtlen és lehetőleg megszüntetendő többértelműségek feltárása érdekében. Megjelöltük azokat az alakokat, amelyek

1. *morfológiailag hibásak* (pl. *felület* vagy *magzat* mint műveltető alak),
2. morfológiailag helyes, de szemantikailag nem, ezért pusztán elvileg *lehetséges alaknak* tekintettük (pl. *beleillet*, *megmérettet*),
3. *csak igekötővel és tárggyal fordulnak elő* (pl. *szar*, *száraz*),
4. olyan *produktív -z képzős* alakok, amelynek töve egybeesik egy lexikai *z(ik)* tövű igéével és többmorfémás elemzése rendkívül valószínűtlen (pl. *fáz, távoz, szerezik, fogalmazik*),
5. amelyek egyben *más szófajúak* is (pl. *feladat, aszat*).

A műveltetővel gyakori többértelműséget okozó *-(t)atik* passzív képző esetében – mivel ez a képző már nemigen produktív –, legjobb megoldásnak a korpuszban talált lexikalizálódott alakok felvételét, és a képző törlését láttuk. Ez viszonylag könnyen kivitelezhető a hátul képzett tövek esetében (itt az *-atik* végű alakok egyértelműen ennek a képzőnek az előfordulásai), az elöl képzett tövek esetében azonban a műveltető képző többes szám 3. személyű alakjával való egybeesés miatt a tényleges passzív alakok kiszűrése nagyon nehéz (pl. az *emlékeztetik* nyilvánvalóan nem passzív igealak – hiszen az *emlékezik* nem tárgyas –, mégis úgy néz ki, mintha az lenne).

5.2 Az igetárgyasság

Bár korábban elsősorban a szóképzést és a szóösszetételt emeltük ki, mint a valószínűtlen túlelemzések forrását, a ragozás körében találunk ilyen eseteket.

Korábbi morfológiai leírásunk egyik hiányossága volt, hogy nagyrészt hiányzott belőle az igetöveknek a tárgyasság szempontjából való besorolása. Ez bizonyos esetekben túlelemzésekhez vezetett: határozott tárgyas elemzést kaptunk olyan igék esetében is, amelyek nem lehetnek tárgyasak. Úgy találtuk azonban, hogy a tárgyasság szempontjából annotált létező leírások sok esetben téves besorolásokat tartalmaznak, amelyek átvételével ténylegesen jó szóalakok jó elemzéseit veszítenénk el. Ezeknek a leírásoknak a készítői nem olyan szempontrendszer figyelembe vételével készítették el ugyanis a leírásukat, amelyek alapján számunkra is használható leírás készülhetett volna.

Ennek oka az, hogy az igetárgyasság más nyelvtani jelenségekkel viszonylag komplex módon interakcióba lép: bizonyos igekötős konstrukciók az egyébként tárgyatlan igék egy igen széles körét is tárgyassá tehetik, ugyanakkor a szintaktikai igevivőinverzió következtében az igekötőnek nem muszáj az adott szóalakon megjelennie:

vitatkozik (nem tárgyas ige), de: *míg ki nem vitatkozza magát...*

Ezért a korábbiaknál pontosabb modelleket kellett kidolgoznunk, illetve nem szorítkozhattunk szótárakban megadott információkra, mert ellenkező esetben megszorításaink túl szigorúak lettek volna. Az igéket az alábbi alapvető osztályokba soroltuk:

S soha nem lehet tárgyas
I csak speciális igekötős konstrukciókban lehet tárgyas: *át...-za az éjszakát*, *ki...-za magát*, vagy más igekötős változata tárgyas
N tárgyas is lehet meg nem is minden speciális mellékkörülmény nélkül (jó a *...-za azt* konstrukció, de lehet csak úgy simán *...-zni* is.)
T csak tárgyas

Az igéket első körben a korpusz szóanyagát elemezve egy heurisztika segítségével soroltuk a megfelelő osztályba. Minden igetőhöz a korpuszban szereplő alakok alapján kiszámítottunk egy mérőszámot, amely azt mutatta, hogy azok közül a szóalakok közül, amelyek elemezhetők az adott tő előfordulásaként hány olyan alak volt, ami kizárólag határozott tárgyas igealakként elemezhető (az utóbbi osztva az előbbivel):

1=minden alakja egyértelműen határozott tárgyas volt.
0=nem volt egyértelműen határozott tárgyas alakja.

Egyértelműen határozott tárgyasnak akkor minősítettünk egy alakot, ha egyáltalán nem volt határozott tárgyas igétől különböző elemzése (tehát egy esetleges névszói vagy más szófajú elemzés is kizáró ok volt). Számos más mérőszámmal is kísérleteztünk, végül azonban ezt a viszonylag egyszerű indikátort elég hatékonynak találtuk. Az osztályozó heurisztika az említett indikátor mellett az ikességet vette figyelembe, az alábbi határértékek használatával:

T<0.0008	S
T<0.011, nem ikes	I
T<0.16, ikes	I
T<0.33	N
egyébként	T

A besorolást kézzel javítottuk, és tovább finomítottuk:

I2 Lehet a szónak (pl. speciális igekötős konstrukcióban) 2. személyű tárgyas alakja, pl. *eljöttelek meglátogatni*.
IN Van olyan igekötős változata (az általános *át/végig...-za az éjszakát*, *ki...-za magát* konstrukciókon kívül), ami gyakori és tárgyas, de igekötő nélkül egyértelműen tárgyatlan.
IT Az ige eleve csak igekötővel létezik, (az elválhat, de teljesen igekötő nélkül nincs, ilyen pl. a *(felül)múl*) és az egyértelműen tárgyas.

A morfológiai elemzőben csak az S (soha nem tárgyas) osztályba besorolt igéket zárhatjuk ki a határozott tárgyas elemzést kapó igék köréből. Ugyanakkor a szintaktikai elemzés szintjén az I (csak igekötős konstrukciókban tárgyas) osztályba sorolt igék többértelműségei is hatékonyan csökkenthetők a vonzatkeretek ellenőrzésével.

5.3 Az ikes igék E/1 alakja

Az ikes igék körében egy másik megoldandó probléma volt a hagyományos (de mára lényegében kihalt) önálló ikes paradigmából megmaradt nem határozott tárgyas jelen idő egyes szám első személyű *-m* toldalék kezelése (*eszem valamit*, *alszom egyet* stb.). Ez ugyanis korántsem minden ikes ige esetében lehetséges: **baszom/fingom/nőzöm egyet, *bánom vele, *meghízom* stb. Ezért külön lexikográfiai munkát jelentett ezeknek az ikes igéknek a feltérképezése, amelyet szintén korpusz alapú vizsgálatokból nyert különböző indikátorok szerint rendezett listák többszöri átnézésével, és kézi javításával végeztünk el. A használt indikátorok az *-m* ragos és *-k* ragos alakok száma, ezek egymáshoz és az adott ige összes alakjához viszonyított előfordulási gyakorisága voltak.

6 Összefoglalás

Cikkünkben olyan rendszeresen elemzési többértelműségekhez vezető morfológiai jelenségekkel kapcsolatos szisztematikus korpuszalapú vizsgálatokat mutattunk be magyarban, amelyek eredményeképpen a Humor morfológiai elemzőben a lexikai többértelműségek körét sikerült csökkenteni. A vizsgálatok eredményei mellett bemutattuk az elvégzett lexikográfiai munkát és a túlgenerálás csökkentésének módszereit.

Bibliográfia

1. Halácsy Péter, Kornai András, Németh László, Rung András, Szakadát István, Trón Viktor. Creating open language resources for Hungarian. In: Proceedings of LREC2004, 2004
2. Kornai, A, Halácsy, P, Nagy, V, Trón, V, and Varga, D (2006). Web-based frequency dictionaries for medium density languages. In: Proceedings of the 2nd International Workshop on Web as Corpus, edited by Adam Kilgarriff, Marco Baroni ACL–06, 1–9.
3. Novák Attila. Milyen a jó humor? In: Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003), pp. 138–145, Szegedi Tudományegyetem, 2003.