

Egy általános célú morfológiai annotáció

Rebrus Péter¹ – Kornai András² – Varga Dániel³

¹Magyar Tudományos Akadémia, Nyelvtudományi Intézet, Budapest

²MTA SZTAKI, Budapest

³BME MOKK, Budapest

rerbrus@nytud.hu; kornai@sztaki.hu; daniel@mokk.bme.hu

A morfológiai annotáció célja az adott szóalakban lévő morfoszintaktikai információk megjelenítése; a kizárólag a jelentésre vagy a hangalakra/írásképre vonatkozó információk nem tartoznak bele. A direkt morfalapú, tehát az allomorfiára tekintettel levő annotációs rendszerek beleütköznek a morfológiai szegmentálás bizonytalanságába, amit a fúziós és szuppletív alakok létezése még tovább bonyolít. Ezzel szemben a jelen cikkben a morfológiai annotáció egy bináris morfoszintaktikai jegyekből és ezek pozitív vagy negatív értékeiből álló fastruktúrára épülő formalizmusát javasoljuk, amelyben csak a pozitív értékkel rendelkező jegy-érték párok csomópontjai dominálnak más csomópontokat. Ez lehetővé teszi, hogy a bináris jegyes hierarchikus szerkezet unáris jegyessé alakítható legyen, s így közvetlenül tükrözze az adott alak morfológiai jelöltségének mértékét. A cikk befejező része az itt bemutatott morfológiai reprezentációt felhasználó, már megvalósult gyakorlati alkalmazásokból mutat be egy csokorra valót.

Kulcsszavak: magyar nyelv, morfológia, annotáció, inflexió

1. Bevezetés

Cikkünk a morfológiai annotáció általános kérdéseit tárgyalja a magyar nyelv példáján keresztül. Az első részben a morfémákra közvetlenül támaszkodó konkrét annotációs sémák és a rögzített kódhosszúságú rendszerek problémáit írjuk le. A második részben a magyar főnévi, igei és egyéb inflexiós paradigmák részletes kódolásáról írunk. Annotációs rendszerünk, a **hunmorph** az említett kódolásokkal szemben absztrakt, változó kódhosszt használó rendszer, amelynek alapelvei teljesen általánosak és nyelvfüggetlenek. A harmadik részben pedig röviden érintjük a deriváció és a szóösszetétel kezelését. Cikkünk záró részében az annotációs rendszert használó nyílt forráskódú számítógépes nyelvészeti eszközöket ismertetjük.

2. Az allomorfolapú annotáció problémái

Kiindulópontunk az, hogy a morfológiai annotáció elsődleges célja az adott szóalakban levő **morfoszintaktikai** információk megjelenítése. Morfoszintaktikainak tekintjük a szóalakban meglevő olyan információkat, amelyeknek közvetlen szintaktikai hatása van, azaz amelyek az adott szóalak mondatbeli formai viselkedését (disztribúcióját) befolyásolják – ilyen elsősorban az a szintaktikai pozíció, ahol a szóalak a grammatikus mondatban megjelenhet, illetve az egyeztetés, amikor egy szóalak morfológiai jegyei befolyásolják egy másik szóalak morfológiai jegyeit. Ennélfogva az alábbi módszertani elvet követtük: a kizárólag a jelentésre és a hangalakra (vagy az írásképre) vonatkozó információk nem részei a morfoszintaktikai reprezentációnak. Egyes esetekben a szemantikai és a szintaktikai információk éles elkülönítése nehézségekbe ütközik, ezért egy általános célú morfológiai annotáció tervezésekor mérlegelnünk kell, hogy a potenciális alkalmazások számára mely szemantikai információk lehetnek lényegesek. A fenti módszertani elvet azért is érdemes szem előtt tartani, mert az annotáció elveinek transzparensnek kell lenniük: egy formai tulajdonságot bárkinek könnyű betanítani (és így emberi erőforrás segítségével előállítani egy nagy pontossággal címkézett korpuszt), míg a szemantikai tulajdonságok nagy részére ez nem áll. Az egyszerre szintaktikai és szemantikai tulajdonságokon alapuló ilyen jegyek körébe tartozik többek között a főnév–melléknév megkülönböztetés vagy igéknél a modális (ható ige) és a múlt idő, amelyeket annotációs rendszerünk is megkülönböztet (erről l. később).

2.1. Allomorfia

A hangalakra (fonológiai formára) vagy az írásképre vonatkozó információknak a morfológiai annotációban való megjelenítése azért sem lenne szerencsés, mert nagyon gyakran önkényes döntéseket kellene hozni arról, hogy milyen alakot adjunk meg allomorfia esetén (azaz akkor, ha az adott morféma több alakban jelenhet meg). Vegyünk néhány példát: a *fára* alak a következő információkat hordozza: (i) lemmája: FA, (ii) morfoszintaktikai jegyei: **SZÁM: EGYES, BIRTOKOS: NINCS, BIRTOK: NINCS, ESET: SUBLATIVUS**. Ha az annotációban ezeken a jegyeken túl azt is meg akarnánk jeleníteni, hogy a szóban forgó *fára* alakban a többeli magánhangzó hosszú (szemben más alakokkal, pl. ilyen a *fa* tőalak, a toldalékolt *faként* alak, vagy a *facipő* szóösszetétel), akkor az elemzésben a tövet esetleg a fá és nem a fa alakban adhatnánk meg. A szóban forgó *fára* alakban jelenlevő toldalék azonban előlképzett magánhangzójú változatban is megjelen-

het (pl. *kép-re*), így dönthetnénk úgy is, hogy az esetragnak ezt a jellegzetességét az annotációnak tükröznie kell, azaz valamilyen alulspecifikált alakban adhatnánk meg a toldalékot (pl. $-rA$, ahol a nagy A szimbólum a középnnyílt elülső e és hátulsó a magánhangzók helyett áll). Hasonló a helyzet máskor is, ahol a szóalakban szereplő morfémák allomorfiát szenvednek el. Például a *szelek* vagy a *sarki* alakokban szintén tőallomorfiát találunk: *szél* – *szelek*, *sarok* – *sarki*, sőt az első esetben a többes szám jelölője más szóalakokban más és más alakban jelenhet meg (pl. *kár-ok*, *ház-ak*, *sün-ök*, *zokni-k*), ezért ennek a morfémának a jelölése sem nyilvánvaló (lehetne az előzőhöz hasonlóan alulspecifikált magánhangzóval $-V_k$ vagy magánhangzó nélkül csupán $-k$).

Látható tehát, hogy ha a morfológiai annotációt az allomorfokkal vagy az allomorfoknak valamilyen absztrakt alakjával adjuk meg, akkor az esetek jelentős részében legalább három megoldást követhetünk (természetesen lehetségesek kevert megoldások is): (a) a **konkrét** elemzésben az adott szóalakban megjelenő allomorfokat (tulajdonképpen a teljes sztringet eredeti formájában) szerepeltetjük (pl. $fá+ra$, $szel+ek$ és $sark+i$); (b) az **allomorfiamentes** elemzésben az allomorfok közül a leggyakoribbat vagy az alapallomorfot választjuk ki (ilyen, amikor a fa , a $szél$, és a $sarok$ töveket adjuk meg a *fára*, *szelek*, illetve *sarki* alakok elemzésénél); és (c) az **absztrakt** elemzés, ahol allomorfia esetén az összes allomorfot lehetőség szerint szerepeltetjük: ez a $-rA$ és a $-V_k$ toldalékok vagy a $fÁ$, $szÉl$ és a $sarOk$ tövek esete, ahol egy alulspecifikált (nagybetűs) szimbólum mutatja a váltakozás helyszíneit (ez lehet nyúlás, rövidülés, hangkivetés, magánhangzó-harmónia stb.). Ez az alulspecifikációs megoldás azonban nem mindig lehetséges: vannak az allomorfiának olyan esetei, amelyekben a váltakozó szekvencia nem adható meg alulspecifikált szimbólummal: ilyen ún. nem-fonológiai allomorfiákat találunk az igei paradigmában, ha a toldalékváltozatok között nincs fonológiai kapcsolat (sőt gyakran a szekvenciák hossza sem azonos): ilyen az E.2 alakokban a tövégtől függő *sz~ol/el/öl* váltakozás (pl. *kap-sz ~ mos-ol*) vagy az E.3 definit alakokban a $tő$ hangrendjétől függő *ja~i* váltakozás (pl. *lop-ja ~ lep-i*).

Egy további probléma a $tő$ - és toldalékallomorfok azonos alakúságával függ össze: a *szelek* tőalakja a *szél*, viszont van egy másik, nem-rövidülő magánhangzót tartalmazó azonos alakú lexéma, vö. *szél* – *szélek*. Hasonló igaz a *sarki* alakra: ez lehet a *SAROK* lexémához tartozó (pl. *sarki bolt*), de lehet a *SARK* lexémához tartozó is (pl. *sarki expedíció*). Tehát ha ezek az alakok önmagukban utalnának az aktuális lexémára, az nem lenne elegendő (ez természetesen más, nem allomorfikus esetben is így van, ekkor a lexémákat a lexikográfiai gyakorlatban sorszámok használatával – pl. $ÁR_1$, $ÁR_2$, $ÁR_3$ – különítik el egymástól). Hasonló homonímia-jelenségek léphetnek fel a toldalékokban is: a $-k$ toldalék nemcsak névszók többes

számára utalhat (pl. ház-**ak**), hanem igéknél az E.1 (pl. én kap-j-**ak**) és bizonyos esetekben a T.3 (pl. ők kap-t-**ak**) szám/személyre is. Ugyanígy az -i toldalék nemcsak melléknévképző lehet (pl. sarki), hanem utalhat a birtok többes számára is (pl. hajó-**i**, Pál-**é-i**). Tehát a morfoszintaktikai kódoláshoz a toldalékok alakja sem ad elégséges információt. Tanulságos összjáratot mutat a homonímia és a nemfonológiai allomorfia az olyan alakoknál, mint amilyen az indefinit E.1 kap-ta-m, ahol a „szokásos” indefinit E.1 -k toldalék helyett -m toldalékot találunk. Ekkor a „konkrét” elemzésben (kap+t+am) az -m szerepel, ami félrevezető lehet, hiszen az -m a szokásos **definit** E.1 toldalékkal azonos (pl. kap-om, kap-j-am (azt)). Az allomorfiamentes elemzésben ezzel szemben a „szokásos” -k **indefinit** E.1. végződés szerepelne (kap+t+k), ami szintén félrevezető, hiszen egy E.3 alak pontosan ennek megfelelő alakú (ők kaptak). Az absztrakt elemzésben viszont szerepeltetni kellene mindkét allomorfot, hiszen a -k és -m toldalékallomorfokat nem lehet értelmes módon alulspecifikálni: kap+t+m/k. Az alábbi táblázatban összefoglaljuk az említett alakok háromféle elemzését (félkövérrel jelölve a problémás eseteket; a kérdőjel a többféle lehetséges elemzést jelöli valódi morfoszintaktikai különbség nélkül).

(1) Főbb elemzési lehetőségek allomorfia esetén

	a. konkrét elemzés	b. allomorfiamentes elemzés	c. absztrakt elemzés
<i>fára</i>	fá+ra	fa+ra	fÁ+rA
<i>szelek</i>	szel+ek	szél+k	szEl+V_k
<i>sarki</i>	sark+i	sarok+i/sark+i	sarOk+i/sark+i
<i>kaptam</i>	kap+t+am	kap+t+k/kap+t+m	kap+t+m?k

Az imént bemutatott allomorfia-alapú elemzések tehát több szempontból problematikusak: (i) nincs módszertani eszközünk arra, hogy **eldöntsük**, hogy a három ideáltípus közül melyik elemzési módot kövessük (pl. fá+ra vagy fa+ra avagy fÁ+rA); (ii) az absztrakt (és részben az allomorfiamentes) annotáció használata mögött hallgatólagosan olyan vitatott elemzések és így nyelvészeti **elméletek** kaphatnak szerepet, amelyekről a nyelvtudománynak nincs egységes álláspontja (pl. a kötőhangzó része-e a toldaléknak vagy sem, vagy nemfonológiai allomorfia esetén mely allomorfo(ka)t szerepeltessük); (iii) egyes elemzések összemosásuk a tő- vagy a toldalékallomorfokban potenciálisan jelen levő **homonímiákat**, így önmagukban nem elegendőek a szóalakban levő morfológiai információk megadásához (l. pl. a *szélek*, a *sarki* és a *kaptam* *vmit* alakok fenti esetét).

2.2. Szegmentálás

A fentebb bemutatott annotációs megközelítéseknek egy további súlyos következménnyel is szembesülniük kell, ez pedig a morfológiai **szegmentálás** bizonytalansága. A morfológiai annotációnak tartalmaznia kell egy határjelölőt, amely elválasztja a morfokat egymástól (a fenti hipotetikus elemzésekben erre a célra a + szimbólumot választottuk). Ez az elválasztás azonban sok esetben önkényes és nem ritkán problémákba ütközik. Lássunk néhány példát! A problémás esetek első típusa az írásképpel kapcsolatos. A grafémikus alakban a kettőzött digráfok speciális írásmódja miatt nem lehetséges az eredeti szóalak karaktereit megfelelő módon elválasztani; ez történik pl. a *hússzor*, *ésszerű* stb. alakok elemzésénél: a konkrét elemzésben a kettőzött digráfot meg kell osztani a tő és a toldalék között, ami félrevezető (pl. *hús+szor*, *ész+szerű*); az absztraktabb elemzésben viszont nem pontosan a szóalak karakterei találhatók (pl. *húsz+szor*, *ész+szerű*). Hasonló a helyzet akkor, ha a szóalak kettős mássalhangzóra végződik, és a toldalék ugyanezzel a mássalhangzóval kezdődik (pl. *szebből*, *halottal*). A következő problematikus típus a morfhataron lezajló hasonulásokkal kapcsolatos. Így például a *-val/-vel* toldaléknak vagy a felszólító mód *-j* toldalékának egyes mássalhangzó utáni változatai esetén nem világos a szegmentálás (*hát+tal*, *hätt+al* vagy *hát+val*, illetve *fus+sa* vagy *fut+ja*). Ha a tő digráfra végződik, akkor a két említett probléma együtt jelentkezik: pl. *ács+csal*, *ács+al*, *ács+csal* vagy *ács+val*, illetve *ed+dze*, *edz+dze* vagy talán *edz+je*. A következő táblázatban ezeket az elemzési lehetőségeket foglaltuk össze.

(2) Szegmentálási lehetőségek különböző elemzések esetén

	a. eredeti sztring	a.' átelemezett sztring	b. allomorfiamentes
<i>hússzor</i>	hús+szor	<i>húsz+szor</i>	
<i>szebből</i>	szeb+ből	<i>szebb+ből</i>	
<i>háttal</i>	hát+tal hätt+al		<i>hát+val</i>
<i>ácssal</i>	ács+csal ács+al	<i>ács+csal</i>	<i>ács+val</i>
<i>fussa</i>	<i>fus+s+a</i>		fut+j+a
<i>eddz</i>	ed+dz	edz+dz	<i>edz+j</i>

A fentiekhez hasonló technikai problémákkal minden morfológiai elemzőprogramnak meg kell birkóznia. Az, hogy egy elemző technikailag melyik módszert követi az aktuális szóalakok (sztringek) manipulációja során, az elemzőprogram (és az erőforrások) felépítésétől, lehetőségeitől függ, és nem morfoszintaktikai információ, amelyet a végső kódolásnak meg kell jelenítenie. Azaz a szegmentálás és az allomorfok kiválasztása az elemző belügye, és **nem** lehet **része** a morfoszintaktikai **annotációnak**.

2.3. Fúzió és szuppletivizmus

Ki kell térnünk egy további problémakörre, amely azt is megmutatja, hogy egyes esetekben a sztringalapú elemzést nem is lehetséges ésszerű módon megvalósítani. Az ún. **fúziós morféma**k esetén több funkció szételemezhetetlenül társul egy morfhoz; a legismertebb példa a magyarban a birtokos alakok és igeik szám/személy jelölése. Az igazán problematikus esetek azonban azok, amikor a fúzió csak bizonyos esetekben áll fenn, máskor a morféma agglutinatív módon jelennek meg. Ezt a jelenséget láthatjuk az igei definitjelölésnél: E.1 és E.2 egyértelműen fúziós (pl. *ad-om*, *ad-od*), E.3 és T.2 agglutinatív (*ad-ja*, *ad-já-tok*). Néha még az is előfordul, hogy az E.3 definitjelölés a módjelölővel fuzionál, pl. az *ad-ná* alakban a *-ná* toldalék együtt fejezi ki a feltételes módot és a definitiséget, tehát ezek az allomorfok nem alkalmasak a morfológiai annotáció jelölésére. A fúzióhoz tartozó jelenség az is, amikor a toldaléktömb formailag szételemezhető, viszont nem egyértelmű, hogy mely funkcióhoz mely szekvenciák tartoznak; ilyen a többes számú birtokosjelölős alakok esete (pl. *kalapjaim*), ahol a szételemezett *kalap+ja+i+m* alakban a *ja* szekvencia nem bír morfoszintaktikai szereppel (ennek absztraktabb nyelvészeti elemzését l. Melcsuk 1965).

A sztringalapú elemzés lehetetlenségét az ún. **szuppletív** alakok mutatják leginkább, ahol ugyanazon lexémához tartozó alakok töve teljesen különbözik (pl. *van* vs. *lehet*, *jön* vs. *gyere*, *sok* vs. *több* stb.). Hasonló jelenség lép fel egyes kis zárt szóosztályoknál, így a személyes és birtokos névmásoknál: az *engem*, *téged* stb. accusativusi alakok nem állíthatók elő mint *én+t*, *te+t* stb.; teljes szuppletivizmusra példa a *benneteket*, *bennünket* alakok, amelyek „tövének” alakja inessivusi, ennek ellenére ezek egyszerű accusativusi alakok: *ti+t*, *mi+t*. Hasonlóan az *enyém*, *tied* stb. alakok morfoszintaktikailag nem birtokosjelölős, hanem birtokjelölős alakok, tehát morfoszintaktikailag *én+é*, *te+é* elemzést kellene kapniuk (a személyes és birtokos névmások esetéről később részletesen is írunk). A *gyere*, *gyertek* alakok ugyanígy a *JÖN* lexémához tartoznak és kötő-felszólító módúak, annak ellenére, hogy alakilag sem a *tő*, sem az *idő/mód* jelölő nem látszik.

(3) Fúziós és szuppletív alakok elemzési problémái

	„formai” elemzés	„morfoszintaktikai” elemzés
<i>adná</i>	? <i>ad+na+a</i>	<i>ad+NÁ+JA</i>
<i>kalapjaim</i>	? <i>kalap+ja+i+m</i>	<i>kalap+K+m</i>
<i>engem</i>	? <i>én+m</i>	<i>ÉN+T</i>
<i>enyém</i>	? <i>én+m</i>	<i>ÉN+É</i>
<i>benneteket</i>	? <i>benn+etek+et</i>	<i>TI+t</i>
<i>gyere</i>	? <i>gyer+e</i>	<i>JÖN+J</i>

A következő részben azt tekintjük át, hogy milyen alternatív annotációs megoldás lehetséges.

3. A kizárólag morfoszintaktikai kategóriákon alapuló annotáció

3.1. Általános annotációs elvek

Az előző részben láttuk, hogy a morfológiai annotáció problémájára a megoldás nem a szóalakok (fonológiai vagy grafémikus) formáján alapuló kódolás, hanem egy nyelvészetileg megalapozott morfoszintaktikai kategóriákra épülő formalizmus adhat választ. Egy ilyen elterjedt annotáció az ún. MSD-kódrendszer (*Morphosyntactic Description*, l. Erjavec–Monachini 1997), amelyben a morfoszintaktikai **kód rögzített hosszúságú**: egy jegyértékekből álló sztring, amelynek minden pozíciójához eleve rögzített módon vannak hozzárendelve a jegyek: azaz a pozíciók azt adják meg, hogy mely értéknek a jegyeit töltjük ki. Lássunk néhány példát: a *fiú* és a *fiatokéinak* főnévi, illetve az *ad* és az *adtátok* igei alakok MSD-kódrendszer szerinti annotációi az alábbiak:

(4) Két főnévi és igealak MSD-annotációja

<i>fiú</i>	Nc-sn-n---	<i>ad</i>	Vmip3s---n-----
<i>fiatokéinak</i>	Nc-pd-yp2p	<i>adtátok</i>	Vmis2p---y-----

Amint a példákból is kitűnik, ennek a kódolásnak a hátránya az, hogy egyrészt nehezen kezelhető (rosszul olvasható a sok üresen hagyott érték és az értékek nem vagy csak kevésbé transzparens kódjai miatt). Másrészt nem hierarchikus, azaz az annotáció közvetlenül nem tükrözi az egyes értékek közötti összefüggéseket – például ilyen összefüggés az, hogy csak birtokos alakoknál van szükség a birtokos számának és személyének megjelölésére, vagy az, hogy a magyarban van egy speciális *-lak/-lek* toldalék, amely 2. személyű tárgyra utal, viszont az alanynak E.1-nek kell lennie: pl. *(én) látlak (téged/titeket)*; azaz ez a morfoszintaktikai érték függ az ige szám/személyétől. Harmadrészt nem képes a morfológiai jelöltséget tükrözni: azaz egy formailag és funkcionálisan komplex szóalak (pl. *fiatokéinak* vagy *adhattátok*) és egy ilyen szempontból jelöletlen szóalak (pl. *fiú* vagy *ad*) annotációja ugyanolyan komplexitású. További problémája az, hogy egyelőre csak inflexiók kódrendszer, és nem nyilvánvaló, hogy a morfoszintaktikailag releváns képzések hogyan illeszthetők bele (különösen igaz ez a szófajváltó képzésekre).

3.2. Jegy-érték szerkezetek

A fenti problémák egy részére megoldást jelent a hierarchikus jegy-érték struktúrák (pl. az ún. AVS-ek, *Attribute-Value Structures*, l. Trón 2002) használata. Az AVS-ek előnye a nyelvészeti és formális megalapozottság: ezt a formalizmust több szintaktikai elmélet használja. A teljesen kitöltött AVS-eknek is problémája azonban az, hogy az annotáció nem tesz különbséget morfológiailag jelölt és jelöletlen szóalakok között. Lássunk egy példát: a fenti *fiaitokéinak* és a *fiú* alak a következő morfoszintaktikai információkat hordozza (itt és a későbbiekben a jegyeket és értékeiket kiskapitálissal jelöltük, ezen belül félkövérrel a jegyeket, és kurzívval az értékeket; a hierarchikus viszonyok jelölésére tabulálást alkalmaztunk).

- (5) Két főnévi alak sematikus jegy-érték struktúrája

a. *fiaitokéinak*

LEMMA	<i>FIÚ</i>
KATEGÓRIA	<i>FŐNÉV</i>
SZÁM	<i>TÖBBES</i>
BIRTOKOS	<i>IGEN</i>
SZÁMA	<i>TÖBBES</i>
SZEMÉLYE	<i>2.</i>
BIRTOK	<i>IGEN</i>
SZÁMA	<i>TÖBBES</i>
ESET	<i>DATIVUS</i>

b. *fiú*

LEMMA	<i>FIÚ</i>
KATEGÓRIA	<i>FŐNÉV</i>
SZÁM	<i>EGYES</i>
BIRTOKOS	<i>NEM</i>
SZÁMA	<i>(SZÁMA X)</i>
SZEMÉLYE	<i>(SZEMÉLYE X)</i>
BIRTOK	<i>NEM</i>
SZÁMA	<i>(SZÁMA X)</i>
ESET	<i>NOMINATIVUS</i>

Hasonlóan az említett *adhattátok* és *ad* igék szokásos specifikációja az alábbi.

- (6) Két igei alak sematikus jegy-érték struktúrája

a. *adhattátok*

LEMMA	<i>AD</i>
KATEGÓRIA	<i>IGE</i>
MODÁLIS	<i>IGEN</i>
IDŐ	<i>MÚLT</i>
MÓD	<i>KIJELENTŐ</i>
SZÁM	<i>TÖBBES</i>
SZEMÉLY	<i>2</i>
DEFINITISÉG	<i>IGEN</i>

b. *ad*

LEMMA	<i>AD</i>
KATEGÓRIA	<i>IGE</i>
MODÁLIS	<i>NEM</i>
IDŐ	<i>JELEN</i>
MÓD	<i>KIJELENTŐ</i>
SZÁM	<i>EGYES</i>
SZEMÉLY	<i>3</i>
DEFINITISÉG	<i>NEM</i>

A fenti (5) és (6) szerkezetekből látható, hogy nincs jelentős különbség a jelölt és jelöletlen alakok jegy-érték struktúrájának „bonyolultsága” között: azok ugyanazokat a jegyeket tartalmazzák. Ez azonban nem intuitív és nem is praktikus, hiszen a morfológiailag jelöletlen alakok általában rövidebbek (több zérusmorfol vagy morfémát tartalmaznak) és jelentősen gyakoribbak (funkciójuk általánosabb, használatuk kiterjedtebb).

3.3. Bináris és unáris jegyek: főnevek

Ha azonban az AVS-ekben a jegyeket úgy fogalmazzuk meg, hogy az értékük csak igen/nem (+/–) lehessen, és az értékek közül szisztematikusan az egyik a jelöltet (szokásosan a +), a másik a jelöletlent (ez általában a –) jelentse, akkor ezen a **bináris** jegyrendszeren jelentős egyszerűsítést tehetünk (l. Kornai 1989). Ha megengedjük további jegyek és hierarchia bevezetését, akkor ezt mindig megtehetjük, hiszen többértékű jegyek esetén ezek értékeit mindig átírhatjuk bináris jeggyé (pl. ilyen a **SZEMÉLY** vagy az **ESET** jegy a főneveknél vagy az **IDŐ** vagy a **MÓD** az igéknél, l. a fenti (5)-öt, ill. (6)-ot).¹

Lássuk, hogy (5) és (6)-beli példáink milyen bináris jegyszerkezetet kapnak (az újonnan bevezetett jegyek nyelvészeti értelmezéséről l. szintén Kornai 1989-et). Az alábbi (7a)-ban a jelölt főnévi alakot látjuk: itt a legtöbb bináris jegyérték pozitív, míg a (7b)-beli alak esetében az összes másodlagos morfoszintaktikai kategória értéke negatív (az áttekinthetőség érdekében csak a pozitív értékkel bíró jegyek vannak félkövérrel szedve). A (7a) és (7b)-beli AVS-ek ugyanazokat az információkat tartalmazzák, mint az (5a), illetve (5b) szerkezetek. Fontos, hogy a negatív értékkel bíró jegyek alá rendelt jegyeknek semmikor nincs szerepük, ez három esetben állhat elő: (i) az alárendelt jegy negatív értékű domináns jegy esetén nem értelmezhető, vagy (ii) a domináns jegy megfogalmazásából következik, hogy az alárendelt jegy (az adott nyelvben) csak negatív értéket vehet fel, vagy (iii) az adott nyelvben az alárendelt jegy csak a domináns jegy pozitív értéke esetén releváns morfoszintaktikailag. Az (i) esetre példa a **BIRTOKOS** vagy a **BIRTOK** jegyek, amelyek negatív értéke esetén – vagyis ha nincs a főnéven birtokos- vagy birtokjelölés – nincs értelme a birtokos számáról vagy személyéről beszélni (ezt látjuk pl. a *fiú* alak esetén (5b)-ben és (7b)-ben). Egy másik eset a familiáris többes: a *Péterék*, *szomszédék* stb. alakok morfoszintaktikailag többes számúak, ez azonban egy speciális többes szám: a szóalakban jelölt alakokkal familiáris viszonyban álló emberek csoportjára utal; így a **FAMILIÁRIS** jegyet ésszerű a **TÖBBES** jegy alá rendelni (erről részletesen l. Kornai 1989). A (ii) eset akkor áll

¹ Vegyük észre, hogy az alakok helyett a jegyek megcímkézése jelölt és jelöletlen értékekre csak akkor tehető meg, ha egy jegyérték jelöltsége nem függ egy másik jegy értékétől, azaz ebben az értelemben környezetfüggetlen. Ez egyes nyilvánvaló esetekben nem igaz, pl. a jelöltség függhet a lexemától: az ún. relációs főneveknél (pl. *barát*, *anya* stb.) a birtokos alak jelöletlenebb a nem-birtokos alaknál. További ismert eset a felszólító módú igék: itt a 2. személyű alakok – univerzálisan is – jelöletlenebbek, míg más módban általában a 3. személy jelöletlen (pl. a magyarban is E.2 indefinit alak állhat zérus szám/személyjelölővel (pl. *ad-j*), az E.3 indefinit alak viszont a többi móddal ellentétben toldalékkal áll (*ad-j-on*). Ezek azonban az egész rendszer szempontjából elhanyagolható mértékű hátrányok: elfogadjuk, hogy a jelöltség jegyértékekre való értelmezésével az alakok jelöltsége jól közelíthető.

elő, ha a jegy megfogalmazásából következik, hogy az alárendelt jegy(ek) negatív értékű domináns jegy esetén egyértelműen csak negatív értékeket vehetnek fel. Ilyen jegy a **NEM-3. SZEMÉLYŰ** és a **NEM NOMINATIVUSI ESETŰ** jegyek, hiszen ha ezek értéke negatív, akkor a birtokos 3. személyű, illetve az eset nominativusi, így az alárendelt jegyeknek (amelyek a további lehetőségeket adják meg) kötelezően negatív értékekkel kell rendelkezniük: egy szó a nominativusszal együtt más esettel nem rendelkezhet). A (iii) lehetőségre az igei rendszer bemutatásánál térünk vissza. A (7) ábrában ezeket a „default módon” kitölthető vagy érték nélküli jegy-érték párokat zárójelbe tettük.

(7) Két főnévi alak bináris jegy-érték struktúrája

a. *fiatorkéinak*

fiú	+
FŐNÉV	+
TÖBBES SZÁMÚ	+
FAMILIÁRIS	–
BIRTOKOS	+
TÖBBES SZÁMÚ	+
NEM-3. SZEMÉLYŰ	+
1. SZEMÉLYŰ	–
2. SZEMÉLYŰ	+
BIRTOK	+
TÖBBES SZÁMÚ	+
NEM NOM. ESETŰ	+
ACCUSATIVUS	–
DATIVUS	+
SUPERESSIVUS	–
...	

b. *fiú*

fiú	+
FŐNÉV	+
TÖBBES SZÁMÚ	–
(FAMILIÁRIS)	(x)
BIRTOKOS	–
(TÖBBES SZÁMÚ)	(x)
(NEM-3. SZEMÉLYŰ)	(x)
(1. SZEMÉLYŰ)	(x)
(2. SZEMÉLYŰ)	(x)
BIRTOK	–
(TÖBBES SZÁMÚ)	(x)
NEM NOM. ESETŰ	–
(ACCUSATIVUS)	(x)
(DATIVUS)	(x)
(SUPERESSIVUS)	(x)
(...)	

Vegyük észre, hogy ha a morfoszintaktikai információkat tartalmazó jegyeket rögzítjük, akkor bármilyen negatív értékű jegy redundánssá válik, és elegendő csak a pozitív jegyeket megadnunk. Ezt a tulajdonságot felhasználhatjuk arra, hogy a bináris jegyrendszert egyértékűvé (**unáris**sá) tegyük. Ehhez elég a pozitív értékű jegyeket tekintetbe venni, és ha kizárólag ezen jegyek neveit soroljuk fel, akkor teljes értékű annotációt kapunk. Az alábbi (8i) ábrában ez a hierarchikus unáris jegyrendszer látható, amit úgy kaptunk, hogy a (7a,b) bináris jegyrendszerből elhagytuk a negatív értékű jegyeket és a pozitív értékeket. Ezzel az unáris jegyrendszerrel aztán közvetlenül használható annotációs rendszert jön létre: (8ii)-ben a hierarchikus rendszert zárójelek segítségével linearizáltuk (az annotációs formalizmus a következő: a lexémát / jel választja el a morfoszintaktikai annotációtól, ez utóbbi a főkategóriával indul, és utána a további morfoszintaktikai jegyek szerepelnek a hierarchiának megfelelően zárójelezve; az e mögött álló formalizmusról részletesebben a következő részben írunk).

- (8) Két főnévi alak elemzése unáris jegyekkel (a redundáns információk nélkül)

a. hierarchikus formában:

fiaitokéinak

fiú

FŐNÉV

TÖBBES SZÁMÚ (PLUR)

BIRTOKOS (POSS)

TÖBBES SZÁMÚ (PLUR)

NEM-3. SZEMÉLYŰ (--)

2. SZEMÉLYŰ (2)

BIRTOK (ANP)

TÖBBES SZÁMÚ (PLUR)

NEM NOM. ESETŰ (CAS)

DATIVUS (DAT)

fiaitokéinak

fiú

FŐNÉV (NOUN)

b. linearizált formában:

fiú/NOUN<PLUR><POSS<PLUR><2>><ANP<PLUR>><CAS<DAT>>

fiú/NOUN

A jegyeknek a legvégső formában látható megnevezései az angol nyelvészeti szakirodalomban elterjedt rövidítéseket követik: PLUR: *plural* (többes szám), POSS: *possessive* (birtokos), ANP: *anaphoric possessive* (birtok), CAS: *case* (eset) stb.). A fent vázolt jegyrendszer úgy van tervezve, hogy a lehető legegyszerűbben feldolgozható formában tükrözze a morfológiai jelöltségi viszonyokat: éppen ezért ahol nem szükséges, ott az adott jegyet elhagytuk; ilyen a **NEM-3. SZEMÉLYŰ** jegy, amelyet a linearizált annotáció nem is jelöl (erre nincs szükség, mert a személyre utaló jegyek amúgy is a **BIRTOKOS** jegy alá vannak rendelve). A birtokos alakok jelölése így egyszerűbbé válik: a POSS jegy alatti személyre utaló jegyek kétfélek lehetnek: <POSS<1>> vagy <POSS<2>>. A 3. személyű birtokos alakokban a POSS jegy az 1 és a 2 jegy nélkül szerepel, azaz jelölése <POSS>, ez egybevág azzal a megfigyeléssel, hogy a három szám/személy közül a 3. a jelöletlen. A birtokos-jelölővel ellátott alakok sémája tehát a következő.

- (9) Birtokos alakok annotációja

fiam fiú/NOUN<POSS<1>>

fiad fiú/NOUN<POSS<2>>

fia fiú/NOUN<POSS>

fiunk fiú/NOUN<POSS<PLUR><1>>

fiatok fiú/NOUN<POSS<PLUR><2>>

fiuk fiú/NOUN<POSS<PLUR>>

Megemlítyük, hogy a PLUR jegyre a hierarchia három különböző helyén is szükség van: közvetlenül a főkategória-jegy (itt NOUN) alatt (ekkor a lemmában megadott entitás többes számát jelzi), a POSS alatt (ekkor az entitást birtokló birtokos többes számát jelzi), és az ANP alatt (ekkor az entitás által birtokolt birtok többes számát jelzi) – a hierarchikus elrendezés azonban biztosítja, hogy ugyanannak a PLUR jegynek a használata nem vezet félreértéshez, hiszen ezek más jegyek alatt helyezkednek el, amit a linearizált kódban a zárójelezés mutat. Ezt mutatják az alábbi alakok, ahol a PLUR különböző pozíciókban külön-külön és egyszerre is megjelenhet (itt megjegyzendő, hogy a birtok többes számának jelzése a beszélt köznyelvben állítmányi helyzetben nem kötelező, sőt egyes beszélőknél tiltott: pl. *A könyvek a %fiúéi/%fiúé*).

(10) A PLUR jegy különböző használatai

	<i>birtokos és/vagy birtokjelölés</i>	<i>mi többes számú?</i>
a.	nincs birtokos- és birtokjelölés: <i>fiúk</i> <i>fiú/NOUN<PLUR></i>	(entitás)
b.	csak birtokosjelölés (itt 3. személyű): <i>fiai</i> <i>fiú/NOUN<PLUR><POSS></i> <i>fiuk</i> <i>fiú/NOUN<POSS<PLUR>></i> <i>fiaik</i> <i>fiú/NOUN<PLUR><POSS<PLUR>></i>	(entitás) (birtokos) (entitás és birtokos)
c.	csak birtokosjelölés: <i>fiúk</i> <i>fiú/NOUN<PLUR><ANP></i> <i>fiúéi</i> <i>fiú/NOUN<ANP<PLUR>></i> <i>fiúk</i> <i>fiú/NOUN<PLUR><ANP<PLUR>></i>	(entitás) (birtok) (entitás és birtok)
d.	birtokos- és birtokjelölés is (csak azok, ahol a birtok többes számú): <i>fiái</i> <i>fiú/NOUN<POSS><ANP<PLUR>></i> <i>fiai</i> <i>fiú/NOUN<PLUR><POSS><ANP<PLUR>></i> <i>fiuk</i> <i>fiú/NOUN<POSS<PLUR>><ANP<PLUR>></i> <i>fiaik</i> <i>fiú/NOUN<PLUR><POSS<PLUR>><ANP<PLUR>></i>	(birtok) (entitás és birtok) (birtokos és birtok) (entitás, birtokos és birtok)

Itt kell kitérnünk a többes szám egy speciális használatára: a **familiáris többes** alak morfoszintaktikailag többes számú, de nem a lexémával kifejezett entitás többes számára, hanem az azzal valamilyen „familiáris” viszonyban levők összességére (család, ismerősök stb.) utal: pl. *sógorék*, *szomszédék* stb. Ez a viszony kombinálódhat a birtokosjelölős alakokkal (pl. *sógorodék*) és a birtokjelölős alakokkal (*sógoréké*). Ezért az annotáció egy a NOUN alatti PLUR általi dominált FAM jegy segítségével történik (11).

Az esetek kódolása is megfelel a morfológiai jelöltségnek: mivel a jelöletlen eset a nominativus, ezért az alanyesetű alakokat külön nem jelöljük, a többi 17 eset kódolására az elterjedt latin elnevezéseik három betűs rövidítéseit használ-

juk. A CAS jegy azt jelzi, hogy itt egy jelölt (azaz nem nominativusi) alakkal van dolgunk. A 18 eset annotációját és az esetek elnevezését l. (12)-ben.

(11) Familiáris többes alakok

<i>fiúék</i>	<i>fiú</i> /NOUN<PLUR<FAM>>	az entitás familiáris csoportja
<i>fiáék</i>	<i>fiú</i> /NOUN<PLUR<FAM>><POSS>	a birtokolt entitás fam. csoportja
<i>fiúéké</i>	<i>fiú</i> /NOUN<PLUR<FAM>><ANP>	az entitás fam. csoportjának birtoka
<i>fiáéké</i>	<i>fiú</i> /NOUN<PLUR<FAM>><POSS><ANP>	a birtokolt entitás fam. csoportjának birtoka

(12) Az esetek annotációja

a. „strukturális esetek”

<i>fiú</i>	<i>fiú</i> /NOUN	nominativus
<i>fiút</i>	<i>fiú</i> /NOUN<CAS<ACC>>	accusativus
<i>fiúnak</i>	<i>fiú</i> /NOUN<CAS<DAT>>	dativus

b. „lexikális esetek”

b1. „helyhatározói”

i. forrás

<i>fiúról</i>	<i>fiú</i> /NOUN<CAS>	delativus
<i>fiúból</i>	<i>fiú</i> /NOUN<CAS<ELA>>	elativus
<i>fiútól</i>	<i>fiú</i> /NOUN<CAS<ABL>>	ablativus

ii. hely

<i>fiún</i>	<i>fiú</i> /NOUN<CAS<SUE>>	superessivus
<i>fiúban</i>	<i>fiú</i> /NOUN<CAS<INE>>	inessivus
<i>fiúnál</i>	<i>fiú</i> /NOUN<CAS<ADE>>	adessivus

iii. cél

<i>fiúra</i>	<i>fiú</i> /NOUN<CAS<SBL>>	sublativus
<i>fiúba</i>	<i>fiú</i> /NOUN<CAS<ILL>>	illativus
<i>fiúhoz</i>	<i>fiú</i> /NOUN<CAS<ALL>>	allativus
<i>fiúig</i>	<i>fiú</i> /NOUN<CAS<TER>>	terminativus

b2. egyéb

<i>fiúval</i>	<i>fiú</i> /NOUN<CAS<INS>>	instrumentalis-comitativus
<i>fiúért</i>	<i>fiú</i> /NOUN<CAS<CAU>>	causalis-finalis
<i>fiúként</i>	<i>fiú</i> /NOUN<CAS<FOR>>	formativus
<i>fiúvá</i>	<i>fiú</i> /NOUN<CAS<TRA>>	translativus-factivus
<i>húsvétkor</i>	<i>húsvét</i> /NOUN<CAS<TEM>>	temporalis

3.4. Bináris és unáris jegyek: igék

Az igék specifikációjában az eddigi elveknek megfelelően a hierarchikus elrendezést a bináris jegyekkel kombináljuk. Az említett *adhattátok* és *ad* igitokok kétértékű jegyekkel való annotációja a következő.

(13) Két igealak bináris jegy-érték struktúrája

a. *adhattátok*

ad	+
IGE	+
MODÁLIS	+
NEM JELEN.KIJ	+
MÚLT IDŐ	+
FELTÉTELES MÓD	–
KÖTŐ-FELSZ. MÓD	–
INFINITÍVUSZ	–
TÖBBES SZÁM	+
NEM 3. SZEMÉLY	+
1. SZEMÉLY	–
(TÁRGY 2. SZEMÉLYŰ	x)
2. SZEMÉLY	+
DEFINITISÉG	+

b. *ad*

ad	+
IGE	+
MODÁLIS	–
NEM JELEN.KIJ	–
(MÚLT IDŐ	–)
(FELTÉTELES MÓD	–)
(KÖTŐ-FELSZ. MÓD	–)
(INFINITÍVUSZ	–)
TÖBBES SZÁM	–
NEM 3. SZEMÉLY	+
(1. SZEMÉLY	–)
(TÁRGY 2. SZEMÉLYŰ	x)
(2. SZEMÉLY	–)
DEFINITISÉG	–

(14) Elemzés unáris jegyekkel (a redundáns információk nélkül):

a. hierarchikus formában:

adhattátok

ad
IGE (VERB)
MODÁLIS (MODAL)
NEM JELEN.KIJ (--)
MÚLT IDŐ (PAST)
TÖBBES SZÁM (PLUR)
NEM 3. SZEMÉLY (PERS)
2. SZEMÉLY (2)
DEFINITISÉG (DEF)

ad

ad
IGE (VERB)

b. linearizált formában:

ad/VERB<MODAL><PAST><PLUR><PERS<2>><DEF>
 ad/VERB

Az igeik idő/módjának annotációja úgy történik, hogy közvetlenül a VERB jegy alatt szerepel az erre vonatkozó információ (azaz a **NEM JELEN.KIJELENTŐ MÓDÚ** jegy a linearizált formából hiányzik). A jegyrendszer felépítése biztosítja, hogy a zérusmorfémát tartalmazó jelöletlen jelen idő kijelentő módú alakok nem kapnak külön jelölést.

(15) Az igeik négy idő/módjának annotációja

<i>ad</i>	ad/VERB	jelen idő kijelentő mód
<i>adott</i>	ad/VERB<PAST>	múlt idő kijelentő mód
<i>adna</i>	ad/VERB<COND>	jelen idő feltételes mód
<i>adjon</i>	ad/VERB<SUBJUNC-IMP>	kötő-felszólító mód

Az igei személyjelölés annotációja a főnévi birtokos mintát követi: a jelöletlen 3. személy jegy nélkül áll, a 1. és 2. személyek jegyei a PERS jegy alatt szerepelnek. A speciális, csak E.1. személyű igéknél megfigyelhető, 2. személyű tárgyra utaló *-lak/lek* toldalékos alakok annotációja egy az <1> jegy alá bevezetett <OBJ2> jeggyel történik. A definit–indefinit (határozott–általános) igealakok megkülönböztetésére a DEF jegy szolgál, hiszen a definit alakok a morfológiailag jelöltek. A számjelölés a független <PLUR> jeggyel történik.

(16) Igei indefinit és definit szám/személyjelölés annotációja

<i>adok</i>	ad/VERB<PERS<1>>	<i>adom</i>	ad/VERB<PERS<1>><DEF>
<i>adlak</i>	ad/VERB<PERS<1<OBJ2>>>		
<i>adsz</i>	ad/VERB<PERS<2>>	<i>adod</i>	ad/VERB<PERS<2>><DEF>
<i>ad</i>	ad/VERB	<i>adja</i>	ad/VERB<DEF>
<i>adunk</i>	ad/VERB<PLUR><PERS<1>>	<i>adjuk</i>	ad/VERB<PLUR><PERS<1>><DEF>
<i>adtok</i>	ad/VERB<PLUR><PERS<2>>	<i>adjátok</i>	ad/VERB<PLUR><PERS<2>><DEF>
<i>adnak</i>	ad/VERB<PLUR>	<i>adják</i>	ad/VERB<PLUR><DEF>

Az infinitívusz szám/személy jelölésének annotációja igen hasonló az igékéhez. Az „infinitívus” igei jegy (VERB<INF>) sajátos jegykombinációkat enged csak meg: az infinitívusznak nincsen idő/módja és definitisége, viszont lehet szám/személye, amit a <PERS> jeggyel fejezünk ki (az %*adhatni* és az ?*adnalak* típusú infinitívuszi alakok is csak periferiálisan léteznek). Az egyetlen jelentős eltérés az igék annotációjához képest, hogy a <PERS> jegy hiánya ebben az esetben nem a 3. személyű alakot (pl. *adnia*), hanem a szám/személyjelölés nélküli alakot (pl. *adni*) kódolja.²

(17) A szám/személyjelölővel nem rendelkező és az azzal rendelkező infinitívusz annotációja

<i>adni</i>	ad/VERB<INF>		
<i>adnia</i>	ad/VERB<INF><PERS>	<i>adniuk</i>	ad/VERB<INF><PLUR><PERS>
<i>adnom</i>	ad/VERB<INF><PERS<1>>	<i>adnunk</i>	ad/VERB<INF><PLUR><PERS<1>>
<i>adnod</i>	ad/VERB<INF><PERS<2>>	<i>adnotok</i>	ad/VERB<INF><PLUR><PERS<2>>

Összefoglalva: az unáris jegyekkel való hierarchikus ábrázolás lehetőséget teremt arra, hogy egyszerűen megfogalmazható és nyelvészetiileg alátámasztott morfoszintaktikai jegyek segítségével olyan annotációt adjunk, amely teljes, és általában véve tükrözi a morfológiai jelöltségi viszonyokat. Azaz anélkül, hogy közvetlenül hivatkoznunk kellene az elemzett szóalak formai tulajdonságaira (allo-

² Ez fontos különbség, mert az infinitívuszt vonzó igék közül azok, amelyek szám/személyjelöléssel rendelkeznek, kizárólag a szám/személyjelölés nélküli infinitívuszt engedik meg: pl. *Dolgozni(*a) akar*.

morfok, szegmentálás stb.), az annotációs kód mégis változó hosszúságú: hossza nagyjából megfelel a szóalak morfológiai komplexitásának. Ez azt is jelenti, hogy zérusmorfémák esetén az annotáció – mivel bináris jegyeik mind negatívak – kizárólag a lexémából és a főkategória címkéjéből áll. Minden további morféma tovább növeli az annotáció bonyolultságát. Az alábbi táblázatban néhány ilyen „monotonon bővülő” komplexitású alaksort adtunk meg az alakok annotációjával együtt a zérustoldaléktól a maximális alakokig (az összehasonlítás kedvéért a hozzávetőleges morfémahatárokat a szóalakokban jelöltük).

(18) Szóalakok és annotációik egy-egy monoton növvő komplexitású sora

<i>fiú</i>	fiú/NOUN
<i>fiú-k</i>	fiú/NOUN<PLUR>
<i>fi-a-i</i>	fiú/NOUN<PLUR><POSS>
<i>fi-a-i-d</i>	fiú/NOUN<PLUR><POSS<2>>
<i>fi-a-i-tok</i>	fiú/NOUN<PLUR><POSS<PLUR><2>>
<i>fi-a-i-tok-é</i>	fiú/NOUN<PLUR><POSS<PLUR><2>><ANP>
<i>fi-a-i-tok-é-i</i>	fiú/NOUN<PLUR><POSS<PLUR><2>><ANP<PLUR>>
<i>fi-a-i-tok-é-i-t</i>	fiú/NOUN<PLUR><POSS<PLUR><2>><ANP<PLUR>><CAS<ACC>>
<i>ad</i>	ad/VERB
<i>ad-hat</i>	ad/VERB<MODAL>
<i>ad-hat-ott</i>	ad/VERB<MODAL><PAST>
<i>ad-hat-t-ak</i>	ad/VERB<MODAL><PAST><PLUR>
<i>ad-hat-t-atok</i>	ad/VERB<MODAL><PAST><PLUR><PERS<2>>
<i>ad-hat-t-á-tok</i>	ad/VERB<MODAL><PAST><PLUR><PERS<2>><DEF>

Fontos rámutatni, hogy az annotáció semmilyen értelemben nem használja az alulspecifikációt (unáris jegyek esetén ez nem is lehetséges), azaz nem lehetséges megadni úgy egy morfoszintaktikai leírást, hogy az valamilyen értékre ne legyen meghatározva – ez bináris vagy többértékű jegyeket alkalmazó rendszerekben egyszerűen a szóban forgó jegy értékének kitöltetlenül hagyásával történhet. Mivel minden annotáció a morfofonológiai értékekre nézve teljesen specifikált, ezért a potenciálisan alulspecifikáltként kezelhető eseteket kétértelműségként kell kezelnünk. Ilyen eset a magyarban meglehetősen ritka. Például az E.1 és E.2 birtokos alakok esetjelölés nélkül jelenthetnek nominativus vagy accusativust; vagy egyes igealakok a definitiség mindkét értékét felvehetik. Néhány példa:

(19) Morfológiai kétértelműségek kezelése alulspecifikáció nélkül

<i>fiam</i>	fiú/NOUN<POSS<1>>	nominativus (pl. <i>A fiam látott engem.</i>)
	fiú/NOUN<POSS<1>><CAS<ACC>>	accusativus (pl. <i>Láttam a fiam.</i>)
<i>adtam</i>	ad/VERB<PAST><PERS<1>>	indefinit (pl. <i>Egy almát adtam neki.</i>)
	ad/VERB<PAST><PERS<1>><DEF>	definit (pl. <i>Az almát adtam neki.</i>)

4. Formalizmus

Az itt következő részben pontosítjuk az inflexiós jegyrendszernek azt a formalizmusát, amelyet az előző részben mutattunk be. Formálisan az inflexiós annotáció két komponensből áll, az egyik komponens a jegy-érték struktúra, amelyben a **bináris morfoszintatikai jegyek** és ezek pozitív vagy negatív **értékei** szerepelnek. A másik komponens a hierarchiáért felelős, ezt a legegyszerűbb egy **irányított körmentes gráfként** (azaz irányított faként) meghatározni, amelyben minden csomóponthoz egy bináris jegy-érték pár van rendelve, az irányított élek pedig megfelelnek a jegy-érték párok közötti dominanciaviszonyoknak. Mivel ez a gráf egy fa, ezért összefüggő és egy csomópont (a gyökércsomópont) kivételével minden csomóponthoz van olyan csomópont, amelyik őt közvetlenül dominálja; a körmentesség pedig azt biztosítja, hogy ne lehessen egy csomópontnak több közvetlenül domináns csomópontja. A jegy-érték párokkal címkézett gráfra egy további feltételnek kell teljesülnie: csak a **pozitív** értékkel rendelkező jegy-érték párok csomópontjai **dominálnak** más csomópontokat (azaz a negatív értékkel címkézett csomópontok a fában levelek lesznek). Ez a feltétel az előző részben elmondottak alapján lehetővé teszi, hogy a bináris jegyes hierarchikus szerkezet unáris jegyessé alakítható legyen a hierarchia megtartásával, és így a(z unáris) jegyek száma tükrözze a morfológiai jelöltséget.³

Ahogy az előző részben láttuk, a **gyökércsomópont** tartalmazza az inflektált szóalak **kategóriáját** (szófaját, POS (*part-of-speech*)-címkéjét): a gyökércsomópont egy olyan jegy-érték párral van címkézve, ahol a jegy valamely főkategória-jegy (az előző részben ezek közül a NOUN és a VERB szerepelt).⁴ Minden inflektálható kategóriához tartozik egy rögzített inflexiós jegy-érték struktúra, azaz bináris jegyértékekkel címkézett csomópontú fagráf. Inflektálható kategória azonban csak öt van: a három névszói és a ragozható determinánsi és az egy igei kategória, ezek jegy-érték szerkezeteiről l. az előző, illetve a következő részt.

Az inflexiós annotáció linearizálása úgy történik, hogy a pozitív értékkel bíró jegyeket írjuk le a megfelelő zárójelezéssel. Mivel egy fában az ugyanazon csomópont által dominált csomópontok (az ún. testvércsomópontok) egymás közötti sorrendje lényegtelen, ilyen esetekben a linearizálás az összes sorrendben lehetséges. Praktikus okokból azonban a jegyek sorrendjét úgy rögzítettük, hogy

³ Valójában az annotációt közvetlenül unáris jegyekkel címkézett fagráffal is definiálhatnánk, ekkor egy annotáció ennek a jegyekkel címkézett fának olyan részfája lenne, amelynek a gyökércsomópontja megegyezik a bővebb fáéval.

⁴ A hunmorph annotációs rendszer aktuális változata által használt főkategória-jegyek listája megtalálható a függelék (A1) ábrájában.

a félreolvasás lehetősége a lehető legkisebb legyen (az inflektálható kategóriákhoz tartozó jegyek kimerítő listáját és sorrendjüket l. a következő részben). Így a linearizált annotáció már egyértelmű, kódokból és zárójelekből álló sztring lesz.

Mivel a linearizált kód – jelöletlen szóalak esetén – egyetlen főkategória-jegyből is állhat, ezért fontos megjegyeznünk, hogy elvi különbség van egy főkategória-jegy és az ilyen „rövid” inflexiós annotáció között. Például a *NOUN* és a *<NOUN>* két különböző dologra utal: az első egy **jegy neve**, amely a gyökércsomópontban állhat; a második egy morfoszintaktikailag **teljesen specifikált** alak, azaz jegyekkel címkézett fagraf, amelynek minden főnévi jegye negatív (azaz esetünkben az egyes számú nem-birtokos nem-birtok nominativusi alak, l. (7b), illetve unáris formában (8); hasonlóan igékre, l. (13b), illetve (14)). Ezt a különbséget a végső linearizált kódban azonban nem használjuk: a főkategória (és így az egész morfoszintaktikai jegyrendszer) praktikus okokból mindig külső zárójelek nélkül szerepel – ez nem vezethet félreértéshez, hiszen az annotáció úgyis mindig teljes elemzést ad vissza. Az elemzés általános formája – amely már az előző részből ismerős – a következő:

(20) Az inflexiós annotáció sémája

szóalak

lemma/FŐKATEGÓRIA<INFL_JEGY_1><INFL_JEGY_2>...<INFL_JEGY_n>

Morfológiai elemzésnek általánosan egy olyan hozzárendelést nevezünk, amely minden egyes jólformált szóalakhoz (sztringhez) hozzárendel egy lexéma–annotáció párt. Ez a hozzárendelés azonban nem egyértelmű (nem függvény), mivel ugyanahhoz a szóalakhoz több különböző elemzést is rendelhet morfológiai homonímia esetén – l. pl. a (19)-beli eseteket. A hozzárendelés megfordítása (inverze) sem függvény, mert ugyanolyan lexémának ugyanolyan annotációval különböző szóalakok felelhetnek meg: ez a helyzet áll elő morfofonológiai ingadozás esetén (pl. *fotelban* – *fotelben* vagy *fürdenek* – *fürödnek*), vagy olyan alakoknál, ahol a szuppletív tő megjelenése nem kötelező (pl. *jöjj* – *gyere* vagy *volna* – *lenne*).

(21) Ingadozó alakok azonos annotációt kapnak

<i>fotelban</i>	<i>fotel</i> /NOUN<CAS<INE>>
<i>fotelben</i>	<i>fotel</i> /NOUN<CAS<INE>>
<i>fürdenek</i>	<i>fürdik</i> /VERB<PLUR>
<i>fürödnek</i>	<i>fürdik</i> /VERB<PLUR>
<i>gyere</i>	<i>jön</i> /VERB<SUBJUNC-IMP><PERS>
<i>jöjj</i>	<i>jön</i> /VERB<SUBJUNC-IMP><PERS>
<i>jöjjél</i>	<i>jön</i> /VERB<SUBJUNC-IMP><PERS>

A következőkben a korábbi főnévi és igei elemzéseket kiegészítjük a többi inflekthető elem annotációjával.

4.1. Névszói kategóriák

Régi problémája a leíró nyelvtanoknak, hogy be lehet-e (és ha igen, hogyan) sorolni egyértelműen a névszói alakokat valamelyik névszói kategóriába (l. többek között Moravcsik 1997). A melléknevek és a számnevek a főnevekkel átfedő osztályokat alkotnak, és nehéz egyértelmű disztribúciós teszteket adni, amelyeknek alapján ezek a kategóriák egyértelműen megkülönböztethetők lennének. Ezen a helyzeten a morfológiai vizsgálatok sem segítenek, mivel mind a melléknevek, mind a számnevek felvehetik az összes főnévi inflexiót egyes „elliptikus” és „nominalizáló” kontextusokban: pl. *Nem szeretem a kíváncsiakat*; *Ez az én nagy labdám*, *az meg a te kicsid*; *Bátraké a szerencse*; *Összeültek a nyolcak*; *Négyet rendeltem*; *Az ő öt könyve meg az én háromam*. Itt és a többi hasonló példában vitatható, hogy az adott melléknév vagy számnév a saját „prototipikus” mondattani **funkciójában** szerepel-e, de melléknév, illetve számnév voltak mellett számos érv szól. Nyilvánvaló, hogy a mondatokban különböző funkciókban álló ugyanazon elemek megkülönböztetése nem lehet a feladata egy csak szóalakokat vizsgáló morfológiai elemzőnek, és így az annotációnak sem. Így például a *pék barátom* és a *szomszéd Józsi* típusú szerkezetekben az első főnév módosító szerepű (ahogyan tipikusan a melléknevek), a *szépek imádata* és a *kevés is sok* típusú szerkezetekben a melléknév, illetve a számnév főnévi jellegű (birtokos szerkezeten belül, illetve alanyként áll); ezt a tényt azonban nem érdemes az adott alakok többszófajúsága mellett felhozni, mert akkor a névszók jelentős többségével ezt kellene tennünk, és így értelmetlenül sok többszörös annotációt kapnánk. (A kizárólag melléknévinek tartott toldalékok, mint amilyen a közép- és felsőfok jele, sem adnak jobb fogódzót, ezek ugyanis a mellékneveken kívül egyes számnevekkel is lehetségesek (pl. *több*, *kevesebb*, *legelső*), és egyes konstrukciókban főnevekhez is járulhatnak: pl. *székebb a széknél*.)

A hunmorph kategóriarendszerének összeállításánál arra is figyelemmel kellett lennünk, hogy az elérhető elektronikus adatbázisok (pl. szótárak) és a rendelkezésre álló elemzett korpuszok (pl. a Szeged Korpusz, l. Csendes et al. 2004) valamilyen módon mégis megkülönböztetik a három fő névszói kategóriát (ezt nagyon sokszor nem formai–disztribúciós, hanem szemantikai–funkcionális alapokon teszik). Ezért az információvesztés elkerülése végett érdemes ezt a kategorizációt megtartani. A három névszói kategória morfoszintaktikai jegyrendszere

viszont azonos lesz: bármely névszó felveheti az összes főnévi inflexiós kategóriát. Az alábbi (22) néhány példát ad inflektált alakokra.

(22) Melléknévi és számnévi alakok névszói inflexiókkal

<i>kíváncsi</i>	kíváncsi/ADJ
<i>kíváncsijaitokét</i>	kíváncsi/ADJ<PLUR><POSS<PLUR><2>><ANP><CAS<ACC>>
<i>kétezer</i>	kétezer/NUM
<i>kétezeinkével</i>	kétezer/NUM<PLUR><POSS<PLUR><1>><ANP><CAS<INS>>

4.2. Determinánsok

A negyedik inflektálható kategória a determinánsoké (DET), l. (23). Pontosabban a determinánsoknak csak egy része inflektálható, az olyan szerkezetekben, mint pl. *ezeké a lányoké, abban a házban*. Más részük viszont nem inflektálható, pl. *e lányoké, ama házban, azon gondolatoknak*. Az inflektálható determinánsok inflexiós jegyszerkezetükben megegyeznek a többi névszóval. (Meg kell jegyeznünk, hogy a szokásosan a determinánsok közé számított névelők a hunmorph-ban külön kategóriát képeznek (ART), amit rendkívül gyakori előfordulásuk és speciális funkciójuk indokol – ide csupán három lemma tartozik: *a, az, egy*.) Néhány példa determinánsokra (az utolsóként felsorolt típus – *ezen, azon* stb. – kétértelmű: lehet inflektálhatatlan determináns, de lehet superessivusi esetű inflektálható is: vö. *azon emberekkel* vs. *azon az emberen*):

(23) Inflektált és nem inflektált determinánsok

<i>emez</i>	emez/DET
<i>ugyanazokéval</i>	ugyanaz/DET<PLUR><ANP><CAS<INS>>
<i>e</i>	e/DET
<i>azon</i>	azon/DET
	az/DET<CAS<SUE>>

4.3. Névadások

4.3.1. Főnévi, melléknévi, számnévi névadások

A névadások a hunmorph rendszerben nem képeznek külön kategóriát (szemben a más alapokon nyugvó annotációkkal, pl. a már említett MSD-kódrendszerrel). A disztribúciós elemzés (és funkcionális megfontolások is) azt az elképzelést támogatják, hogy a névadások szétoszthatók a négy névszói (NOUN, ADJ, NUM, DET) és a határozószói (ADV) kategóriák között. Hely hiányában a névadások elemzésére itt részletesen nem tudunk kitérni, álljon itt néhány példa a hagyományos besorolásuk szerint:

(24) Főnévi, melléknévi és számnévi névmások annotációja

a. mutató

<i>ez</i>	ez/NOUN
<i>azokéval</i>	az/NOUN<PLUR><ANP><CAS<INS>>
<i>ilyen</i>	ilyen/ADJ
<i>olyanjainak</i>	olyan/ADJ<PLUR><POSS><CAS<DAT>>
<i>ennyi</i>	ennyi/NUM
<i>annyinkat</i>	annyi/NUM<POSS<PLUR><2>><CAS<ACC>>

b. kérdő

<i>micsoda</i>	micsoda/NOUN
<i>kikét</i>	ki/NOUN<PLUR><ANP><CAS<ACC>>
<i>melyik</i>	melyik/ADJ
<i>milyeneken</i>	milyen/ADJ<PLUR><CAS<SUE>>
<i>hány</i>	hány/NUM
<i>mennyivel</i>	mennyi/NUM<CAS<INS>>

c. egyéb (vonatkozó, általános, tagadó)

<i>amely</i>	amely/NOUN
<i>valakijeitékét</i>	valaki/NOUN<PLUR><ANP<PLUR>><CAS<ACC>>
<i>bármelyik</i>	bármelyik/ADJ
<i>semmilyenekkel</i>	semmilyen/ADJ<PLUR><CAS<INS>>
<i>mindahány</i>	mindahány/NUM
<i>akármennyiért</i>	akármennyi/NUM<CAS<CAU>>

4.3.2. Személyes névmások

A hagyományosan személyes és birtokos névmásoknak nevezett szóosztály annotálása érdekében az eddig bemutatott névszói annotációs jegyrendszert kismértékben ki kell bővítenünk. A személyes névmások annotációs rendszerünk szerint speciális főnevek, amelyeknek névszói inflexiók jegyeik lehetnek (alakjuk nagyon gyakran szuppletív, pl. *engem*, *bennünket*, *velük*, *rá*). A különböző személyű személyes névmásokkal való egyeztetési jelenségek indokolják, hogy a névszói jegyrendszert kiegészítsük az igéknél ismert és a személyre utaló PERS jeggyel. Ez a PERS jegy az infinitívusoknál látott módon jelöli a személyt (l. (17)): magában állva a <PERS> 3. személyre utal, míg az e jegy által dominált személyjegyekkel az 1., illetve 2. személyre. Ekkor a személyes névmások annotációja a következő (a formális – „önöző”, illetve „magázó” – személyes névmásokat is szerepeltetjük, ezek morfoszintaktikailag 3. személyűek).

(25) A személyes névmások annotációja

<i>én</i>	én/NOUN<PERS<1>>	<i>mi</i>	mi/NOUN<PLUR><PERS<1>>
<i>te</i>	te/NOUN<PERS<2>>	<i>ti</i>	ti/NOUN<PLUR><PERS<2>>
<i>ő</i>	ő/NOUN<PERS>	<i>ők</i>	ők/NOUN<PLUR><PERS>
<i>ön</i>	ön/NOUN<PERS>	<i>önök</i>	önök/NOUN<PLUR><PERS>
<i>maga</i>	maga/NOUN<PERS>	<i>maguk</i>	maguk/NOUN<PLUR><PERS>

A személyes névmások esetekkel ellátott alakjai között több morfofonológiailag kivételes, illetve szuppletív alak van (l. pl. (3)), ezen kívül a legtöbb alak a nem-formális személyes névmásoknál hiányzik (TRA: **énné*, FOR: **teként*, TER: **őig*, TEM: **önkor*), illetve többszörös alakváltozatok is előfordulnak; néhány példa:

(26) Inflektált személyes névmások

<i>engem engemet</i>	én/NOUN<PERS<1>><CAS<ACC>>
<i>neked néked</i>	te/NOUN<PERS<2>><CAS<DAT>>
<i>vele véle</i>	ő/NOUN<PERS><CAS<INS>>
<i>önhöz</i>	ön/NOUN<PERS><CAS<ADE>>
<i>magáig</i>	maga/NOUN<PERS><CAS<TER>>
<i>bennünket minket</i>	mi/NOUN<PLUR><PERS<1>><CAS<ACC>>
<i>belőletek</i>	ti/NOUN<PLUR><PERS<2>><CAS<ELA>>
<i>rajtuk</i>	ők/NOUN<PLUR><PERS><CAS<SUE>>
<i>önökké</i>	önök/NOUN<PLUR><PERS><CAS<TRA>>
<i>magukként</i>	maguk/NOUN<PLUR><PERS><CAS<FOR>>

4.3.3. Birtokos névmások

Az ún. „birtokos” névmások nem birtokosjelölővel, hanem **birtok**jelölővel vannak ellátva, hiszen nem a személyes névmás által kifejezett személy birtokosát, hanem annak birtokát jelölik, és szintaktikai disztribúciójuk is ennek felel meg: *A könyv a fiúé/tied/övé*. Ezért ezek annotációja az ANP jeggyel történik.⁵

(27) A birtokosra utaló névmások annotációja

<i>enyém</i>	én/NOUN<PERS<1>><ANP>
<i>tied tied tiedé</i>	te/NOUN<PERS<2>><ANP>
<i>övé</i>	ő/NOUN<PERS><ANP>
<i>öné</i>	ön/NOUN<PERS><ANP>
<i>magáé</i>	maga/NOUN<PERS><ANP>
<i>miénk mienk mienké</i>	mi/NOUN<PLUR><PERS<1>><ANP>
<i>tietek tietek tieteké</i>	ti/NOUN<PLUR><PERS<1>><ANP>
<i>övék övéké</i>	ők/NOUN<PLUR><PERS><ANP>
<i>önöké</i>	önök/NOUN<PLUR><PERS><ANP>
<i>maguké</i>	maguk/NOUN<PLUR><PERS><ANP>

A birtokos névmások viselkedése jól példázza azt, hogy a birtokjelölés bizonyos erősen korlátozott esetekben és módon egy alakon belül megismételhető. A többszörös birtokviszonyok lerövidítésére szolgálnak az olyan szerkezetek, mint *Az én kutyám pórása – Kinek a pórása? – Az én kutyámé/Az enyéme⁶/Az enyém*.

⁵ Megjegyzendő, hogy a POSS jegy főnévi személyes névmásokra nem is használatos, hiszen az ezt kifejező alakok szisztematikusan hiányoznak: **éned*, **öm*, **önötök*, **magánk*.

Ez utóbbi alak nem egyszerű birtokolságot, hanem a birtok általi újabb birtokolságát fejez ki, és bizonyos mértékig itt is kifejezhető mindkét birtok többes száma: *A kutyáim pórása – az enyémeke; A kutyám pórásai – ?az enyémei; A kutyáim pórásai – az ?enyémekei.*⁶ Ezt a szerkezetet az ANP csomópont alatti újabb ANP csomóponttal lehet kódolni.

(28) Többszörös birtokjelölés

<i>enyémé</i>	én/NOUN<PERS<1>><ANP<ANP>>
<i>tieidé</i>	te/NOUN<PERS<2>><ANP<PLUR><ANP>>
<i>miénkéi</i>	mi/NOUN<PLUR><PERS<1>><ANP<ANP<PLUR>>>
<i>tieitekéi</i>	ti/NOUN<PLUR><PERS<2>><ANP<PLUR><ANP<PLUR>>>

A birtokos névmások szintén felvehetnek további főnévi inflexiós jegyeket (a POSS jegy kivételével). Így az előbbiekkal és az esetjelöléssel megkaphatjuk a maximálisan komplex PERS jegyet tartalmazó főnévi alakot (az alábbi listában az összehasonlítás kedvéért a szóalakokban bejelöltük a morfémák hozzávetőleges határait):

(29) Inflektált birtokos névmások növekvő komplexitású sora

<i>ő</i>	ő/NOUN<PERS>)
<i>öv-é</i>	ő/NOUN<PERS><ANP>
<i>eny-é-m</i>	én/NOUN<PERS<1>><ANP>
<i>ti-e-i-d</i>	te/NOUN<PERS<2>><ANP<PLUR>>
<i>mi-e-i-nk</i>	mi/NOUN<PLUR><PERS<1>><ANP<PLUR>>
<i>ti-e-i-tek-et</i>	ti/NOUN<PLUR><PERS<2>><ANP<PLUR>><CAS<ACC>>

4.4. Névutók

A névutók külön főkategória (POSTP), de bizonyos névutós alakok érintik a főnévi annotációt is. Az ún. „személyragozott névutók” (*utánam, eléd, nélküle* stb.) olyan elemek, amelyeknek a formája névutó + szám/személyjelölő, de szintaktikai viselkedésük a főnév+névutó sémát követi (pl. *A fiú nélkül/Nélkülem jött el*). Ez problematikus, mivel a morfológiai elemző csak egyes szavakat képes annotálni, a szavak között fennálló konstrukciók elemzése közvetlenül nem feladata. Vegyük azonban észre, hogy a főnév+névutó szerkezetek szintaktikailag az esetjelölős főnévi szerkezetekkel rokoníthatók (vö. az előző mondatokat a következőkkel: *A fiúval/Velem jött el*). Így a személyragozott névutókat tekinthetjük

⁶ Ugyanez a szerkezet nem-névmási alakokkal csak nagyon nehezen fogadható el: [?]**Páléé*, talán kicsit grammatikusabbá válik, ha az első birtokjelölő után többesszámjelölés van: ^{??}*Páléié*.

speciális névutós személyes névmásoknak, és mivel a névmások PERS jeggyel ellátott főnevek, ezért érdemes bevezetni a főnévi inflexiók rendszerbe a POSTP jegyet (amely a POSTP főkategória-jegytől különbözik). A főnév alatti POSTP jegynek aljegye lesz az összes névutós jegy, amelyeknek a nevei a névutó lemmájával azonosak. Ekkor a személyragozott névutók annotációja az esetjelölős személyes névmásokéval rokonítható. Figyeljük meg, hogy a személyrag nélküli névutók és a személyraggal ellátott névutók más-más főkategóriához tartoznak: az első névutó (POSTP), a második főnév (NOUN), ami természetes, hiszen szintaktikai disztribúciójuk nagyban eltér: a POSTP kategóriát közvetlenül megelőzi egy főnév (pl. *fiú mögött*), míg a névutós NOUN kategória határozóként vagy vonzatként áll a mondatban (pl. *mögöttünk*). Megfigyelhetjük azt is, hogy a 3. személyű személyragozott névutók lemmája sohasem lehet a formális önözés/magázás *ön*, illetve *maga* lemmája: *előtte* értelmezése 'előtte', és nem 'ön/maga előtt', és hasonlóan a T.3 közéjük 'öközük' és nem 'önök/maguk közé'.

(30) Személyragozott és sima névutók annotációja

<i>nélkül</i>	nélkül/POSTP
<i>nélküled</i>	te/NOUN<POSTP<NÉLKÜL>>><PERS<2>>
<i>közül</i>	közül/POSTP
<i>közülük</i>	ők/NOUN<POSTP<KÖZÜL>>><PLUR><PERS>
<i>mellé</i>	mellé/POSTP
	ő/NOUN<POSTP<MELLÉ>>><PERS>
<i>melléje</i>	ő/NOUN<POSTP<MELLÉ>>><PERS>
<i>számára</i>	számára/POSTP
	ő/NOUN<POSTP<SZÁMÁRA>>><PERS>
<i>számunkra</i>	mi/NOUN<POSTP<SZÁMÁRA>>><PLUR><PERS<1>>

Figyeljük meg, hogy egyes névutók kétértelműek: egyrészt vannak olyanok, amelyeknek E.3 alakja megegyezhet az inflektálatlan alakkal (pl. *mellé*, *mögé*); másrészt azok, amelyek birtokjelölős alakúak, az E.3 személyben kétértelműek, pl. a *számára* funkciója kettős: egyrészt névutó (pl. *Pál számára*), másrészt az E.3 névutós névmási alak: 'az ő számára' (pl. *Számaára nincs kegyelem*). Bizonyos esetekben a mutató névmás és a névutó egybeírando (pl. *anélkül*, *ezelőtt*, *amiatt*). Ekkor az annotáció szintén a főnévi kategóriát használja a POSTP jeggyel.

(31) Mutató névmási névutók annotációja

<i>anélkül</i>	az/NOUN<POSTP<NÉLKÜL>>
<i>ezelőtt</i>	ez/NOUN<POSTP<ELŐTT>>

A névszói alakok teljes inflexiók specifikációja a függelék (A2) táblázatában található.

4.5. Igék

Az igei alakok jelentős része morfológiailag defektív, azaz egyes lemmák bizonyos – egyébként megengedett – jegykombinációkkal nem állnak. Ilyenek többek között a személytelen igék, amelyeknek minden idő/módjuk megvan, de csak E.3 indefinit alakban állhatnak: ilyen például a *hajnalodik* ige vagy a *kell*, *lehet* segédigék. A *rejlík*, *történik* stb. igéknek vannak többes számú alakjaik is, de csak 3. személyben (és indefinitként) állnak. Az intranszitiv igéknek nem vagy csak periferikusan van definit alakjuk (pl. *kimosakodja magát*, *lejárja a távot* stb.), de van néhány ige, amelyeknek egyáltalán nincs (pl. *jön*, *megy*, *van*). Van olyan ige, amelynek formailag csak múlt ideje van, más idő/módban, infinitívusban nem állhat, ilyen a *szokott* segédige, vagy pedig csak kijelentő mód 3. személyben áll: pl. *nincs*, *nincsenek*. A *fog* segédigének viszont nincs sem kifejezett idő/módja, sem infinitívusza. A legtöbb ige azonban minden idő/módban, szám/személyben és definit, illetve indefinit alakban is állhat. Az igék teljes annotációs specifikációja a függelék (A3) ábrájában látható.

5. Képzés és szóösszetétel

5.1. A képzés annotációja

Bizonyos feladatok (tartalomelemzés, gépi fordítás) megoldásához szükség lehet az inflexiók réteg kielemezésén túli morfológiai elemzésre is. Tipikus példa erre a szintaktikai elemzés (*parsing*), hiszen a melléknévi igeneves szerkezetek – pl. a *Koreában terjedő bankók* főnévi csoport – helyes elemzése megköveteli, hogy a *terjedő* ne csak mint melléknév kerüljön elemzésre (hisz ezt az elemző az őt követő főnévhez csatolná) hanem ki tudjuk elemezni a *terjed* ige-tővet is, hogy a *Koreában* alakot mint ennek bővítményét tudjuk elemezni. Képzőnek azt a toldalékot tekintjük, amely főkategóriához egy másik főkategóriát rendel. Ennek az elvnek a következménye, hogy a képzés bemenete nem érinthet inflektált alakot, morfológiai megfogalmazásban: inflektált alakok nem képezhetők tovább. Képzett alakoknak viszont kötelező inflexiót felvenniük (ha a kimeneti kategória inflektálható). Ezért a képzés annotációját formálisan olyan irányított gráfként tudjuk megadni, amelynek csomópontjai a főkategóriák: egy **képző** két (nem feltétlenül különböző) **kategóriacímke közötti irányított él**. Az elemzésben a képzés alapjául szolgáló lemma és a lemma kategóriája mellett meg kell adnunk a képző elnevezését és a kimeneti kategóriát, amely a következő módon történik.

- (32) Az elemzés formalizmusa képzett alak esetén

szóalak

lemma/LEMMA_KATEGÓRIA[KÉPZŐ]/VÉGSŐ_KATEGÓRIA<INFL_ANNOTÁCIÓ>

Az alábbiakban néhány példát adunk meg különböző kategóriákból történő képzések elemzésére.

- (33) Példák inflektált képzésekre

<i>terjedő</i>	terjed/VERB[IMPERF_PART]/ADJ	(foly. mn-i igenév)
<i>csináltatna</i>	csinál/VERB[CAUS]/VERB<COND>	(kauzatív)
<i>székestül</i>	szék/NOUN[COM]/ADV	(társ)
<i>lábúak</i>	láb/NOUN[INAL_ATTRIB]/ADJ<PLUR>	(elidegeníthetetlen tul.)
<i>oroszul</i>	orosz/ADJ[MANNER]/ADV	(mód)
<i>okosabbjai</i>	okos/ADJ[COMPAR]/ADJ<PLUR><PERS>	(középfok)
<i>sokszor</i>	sok/NUM[MULTIPL-ITER]/ADV	(multiplikatív–iteratív)
<i>nyolcadikak</i>	nyolc/NUM[ORD]/NUM<PLUR>	(sorszám)
<i>szembeni</i>	szemben/POSTP[ATTRIB]/ADJ	(tulajdonság)
<i>utánról</i>	után/POSTP[ATTRIB]/ADJ<CAS>	

Többszörös képzés esetén a teljes képzési gráf kerül linearizálásra az előbbihez hasonló módon, az alábbiakban az általános sémát követően példákat adunk meg.

- (34) Többszörös képzést annotáló elemzés

szóalak

 lemma/KATEGÓRIA_1[KÉPZŐ_1]/...KATEGÓRIA_n[KÉPZŐ_n]/
 VÉGSŐ_KATEGÓRIA<INFL_ANNOTÁCIÓ>

- (35) Példák többszörös képzésekre

<i>faxolgatás</i>	fax/NOUN[ACT]/VERB[FREQ]/VERB[GERUND]/NOUN
<i>lányosabban</i>	lány/NOUN[ATTRIB]/ADJ[COMPAR]/ADJ[MANNER]/ADV
<i>láthatósági</i>	lát/VERB[MODAL_PART]/ADJ[ABSTRACT]/NOUN[MET_ATTRIB]/ADJ
<i>tárgyasít</i>	tárgy/NOUN[MET_ATTRIB]/ADJ[ATTRIB]/ADJ[TRANS_RESULT]/VERB

A VERB, NOUN, ADJ, NUM kategóriák között minden irányban lehetséges képzés (a denumerális ige-képzés kivételével). Az ADV kategória nem igazán képezhető, a POSTP kategóriából egyedül melléknévképzés lehetséges. A függelék (A4) táblázatában megtalálható a jelenleg használt képzők listája.

5.2. Szóösszetételek formalizmusa

A szóösszetételek annotációjához szükség van az összetételi tagok elemzésére. Szerencsére inflektált összetételi tag tipikusan csak az összetétel utolsó eleménél fordul elő, ezért az összetételi tagokat elegendő csak az (esetlegesen előforduló képzési annotációval együtt) kategóriájukkal elemezni. A formalizmust követően néhány példát az alábbiakban találunk.

- (36) Az elemzés formalizmusa szóösszetétel esetén

Szóalak

lemma_1/KÉPZÉS_ELEMZÉSE1+...+lemma_n/KÉPZÉS_ELEMZÉSE<INFL_ANNOTÁCIÓ>

- (37) Példák összetételek elemzésére

- a. inflektált összetételek

<i>eladja</i>	e1/PREV+ad/VERB<DEF>
<i>vér farkasok</i>	vér/NOUN+farkas/NOUN<PLUR>
<i>sötétzöldje</i>	sötét/ADJ+zöld/ADJ<POSS>
<i>kanárisárgáé</i>	kanári/NOUN+sárga/ADJ<ANP>

- b. képzett összetételi tagok

<i>zúzottkő</i>	zúz/VERB[PERF_PART]/ADJ+kő/NOUN
<i>háromszínű</i>	három/NUM+szín/NOUN[INAL_ATTRIB]/ADJ

- c. többszörös összetétel

<i>birsalmasajt</i>	birs/NOUN+alma/NOUN+sajt/NOUN
---------------------	-------------------------------

6. Alkalmazások

A BME MOKK kutatócsoportjának irányításával számos olyan nyelvtechnológiai eszköz fejlesztése történt meg, amelyek a cikkünkben bemutatott morfológiai reprezentációt felhasználva végeznek automatikus szövegfeldolgozási műveleteket. Az alábbiakban bemutatjuk ezek közül a cikkünk szempontjából a legfontosabbakat.

6.1. Morfológiai leírás

A **hunlex** formalizmus egy morfológiai leírások reprezentálására alkalmas formális nyelv. A nyelvész szakértő ebben írhatja le egy adott nyelv morfológiai szabályait, illetve az ezekhez kapcsolódó morfológiai szótárakat. A hunlex reprezentáció specifikálja egy morfológiai elemző (vagy generáló) program viselkedését

az adott nyelvre, azaz szóalakokhoz morfológiai annotációkat rendel. Munkatársaink számos nyelvre dolgoztak ki morfológiai leírást ebben a formalizmusban, ezek közül a **morphdb.hu** erőforrás a jelen cikkben bemutatott magyar nyelvű annotációs rendszer megvalósítása (Trón et al. 2006). A szintén hunlex névre hallgató szoftver a nyelvészek által könnyen olvasható és módosítható magas szintű morfológiai erőforrásból egy úgynevezett aff/dic (affixumlista/szótár) formátumú, „gépközelibb” erőforrást állít elő. Ez az egyszerű formalizmus szabványnak tekinthető, a Firefox és OpenOffice szoftverek komponenseként több száz millió számítógépre telepített **hunspell** helyesírásellenőrző rendszer is ezt használja a morfológiai információ leírására.

6.2. Morfológiai elemzés

Egy morfológiai elemző szoftver feladata egy szóalakhoz megadni az összes legális morfológiai elemzést. Több olyan morfológiai elemző szoftverimplementáció is létezik (**rfst**, **ocamorph**, **jmorph**), amely az aff/dic formátumú morfológiai erőforrásra támaszkodva elvégzi ezt a feladatot (Trón et al. 2005).

6.3. Morfológiai egyértelműsítés

A **hunpos** morfológiai egyértelműsítő azt a problémát orvosolja, hogy egy adott szóalakhoz a morfológiai elemző gyakran több legális elemzést is talál (Halácsy et al. 2007). Az egyértelműsítő ezek közül választja ki a mondatbeli szöveggörnyezet alapján legvalószínűbbet. Ennek a feladatnak a hibátlan megoldásához a mondat teljes megértése lenne szükséges, amely a technológia mai állása mellett természetesen nem lehetséges. Ugyanakkor magyar nyelvre a morfológiai egyértelműsítés feladata – automatikusan feltárt egyszerű statisztikai szabályszerűségek kiaknázásával – elfogadható (tokenenként 95% feletti) pontossággal megoldható.

6.4. Tulajdonnév-felismerés (NER) és a főnévi csoportok felismerése (NP-chunking)

A morfológiai egyértelműsítés hasznos előfeldolgozási lépésként szolgálhat magasabb szintű nyelvfeldolgozási műveletek elvégzése előtt. Ilyen rendszerre példa a **hunner** tulajdonnév-felismerő, amely a szövegben automatikusan azonosítja

és klasszifikálja a személyneveket, helyneveket, szervezetneveket és egyéb tulajdonneveket (Varga–Simon 2007). A jelen kötetben bemutatott **hunchunk** NP-daraboló (főnévcsoport-felismerő) is támaszkodik működése során a morfológiailag feldolgozott szövegre (l. Recski–Varga 2012).

6.5. Gyakorisági korpusz

Korpusznyelvészeti és pszicholingvisztikai kutatásokhoz hasznos segédeszköz a **Szószablya** webes **gyakorisági szótár** (Szalai–Halácsy 2008). Ennek alapja a magyar nyelvű webről reprezentatív módon kigyűjtött több mint félmilliárd tokennyi szöveg automatikus morfológiai egyértelműsítése nyomán épített gyakorisági táblázat. A keresések eredményei a legkülönbözőbb morfológiai és fonológiai szempontok alapján szűrhetők és rendezhetők. A morfológiai reprezentáció hierarchikus mivolta megkönnyíti, hogy a gyakorisági szótár nyelvész felhasználója kereséseihez a legkülönbözőbb módon specifikálhasson morfoszintaktikai relációkat.

Irodalom

- Csendes, Dóra – János Csirik – Tibor Gyimóthy 2004. The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus. *Lecture Notes in Artificial Intelligence* 3206: 41–48.
- Erjavec, Tomaž – Monica Monachini 1997. Specifications and notation for lexicon encoding. Deliverable D1.1 F. Multext-East Project COP-106. <http://nl.ijs.si/ME/CD/docs/mte-d11f/>
- Halácsy, Péter – András Kornai – Csaba Oravecz 2007. HunPos – An open source trigram tagger. In: Sophia Ananiadou (szerk.): *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics. Companion Volume. Proceedings of the Demo and Poster Sessions*. Prague: Association for Computational Linguistics. 209–212.
- Kornai András 1989. A főnévi csoport egyeztetése. *Általános Nyelvészeti Tanulmányok* 17: 183–211.
- Melcsuk, Igor 1965. A magyar főnév birtokos személyragjainak morfológiai felépítéséről. *Magyar Nyelv* 61: 264–275.
- Moravcsik, Edith 1997. *Hungarian adjectives from a typological point of view*. Kézirat. University of Wisconsin-Milwaukee, Milwaukee.
- Recski Gábor – Varga Dániel 2012. Magyar főnévi csoportok azonosítása. *Általános Nyelvészeti Tanulmányok* 24: 81–95.
- Szalai András – Halácsy Péter 2008. Szószablya gyakorisági szótár. <http://szotar.mokk.bme.hu/szoszablya/>
- Trón Viktor 2002. Attribútum–érték struktúrák. In: Kálmán László – Trón Viktor – Varasdi Károly (szerk.): *Lexikalista elméletek a nyelvészetben*. Budapest: Tinta Könyvkiadó. 333–344.

- Trón, Viktor – László Németh – Péter Halácsy – András Kornai – György Gyepesi – Dániel Varga 2005. Hunmorph: Open source word analysis. In: Martin Jansche (szerk.): Proceedings of the ACL Workshop on Software. Stroudsburg, PA: Association for Computational Linguistics. 77–85.
- Trón, Viktor – Péter Halácsy – Péter Rebrus – András Rung – Péter Vajda – Eszter Simon 2006. Morphdb.hu: Hungarian lexical database and morphological grammar. In: Nicoletta Calzolari – Khalid Choukri – Aldo Gangemi – Bente Maegaard – Joseph Mariani – Jan Odijk – Daniel Tapias (szerk.): Proceedings of 5th International Conference on Language Resources and Evaluation. Genoa: European Language Resources Association (ELRA). 1670–1673.
- Varga, Dániel – Eszter Simon 2007. Hungarian named entity recognition with a maximum entropy approach. Acta Cybernetica 16: 293–301.

Függelék

(A1) A hunmorph főkategória-jegyei

a. inflektálható (és továbbképezhető)

névszóként:

NOUN	főnév (<i>noun</i>)
ADJ	melléknév (<i>adjective</i>)
NUM	számnév (<i>numeral</i>)
DET	determináns (<i>determiner</i>) [8]

igeként:

VERB	ige (<i>verb</i>)
------	---------------------

b. nem inflektálható (és nem továbbképezhető)

ADV	határozószó (<i>adverb</i>)
POSTP	névutó (<i>postposition</i>)
	– marginálisan esetjelölés és mn-képzés lehetséges
ART	névelő (<i>article</i>) [3]
CONJ	kötőszó (<i>conjunction</i>) [198]
PREP	prepozíció (<i>preposition</i>) [5]
PREV	igekötő (<i>preverb</i>) [105]
UTT-INT	mondatszó/indulatszó (<i>utterance/interjection</i>)
ONO	hangutánzó (<i>onomatopoeic</i>)

c. egyéb, írott nyelvi kategória

PUNCT	központoszási jel (<i>punctuation</i>)
-------	--

(A2) A névszói inflexiós jegyhierarchia

{NOUN ADJ NUM DET}

<POSTP

<ALATT>

<...>

(az összes névutó)

<PLUR

<FAM>>

<PERS

<1>

<2>>

<POSS

<PLUR>

<1>

<2>>

<ANP

<PLUR>

<ANP

<PLUR>>>

<CAS

<ACC>

<DAT>

<...>

(az összes [17] jelölt esetrag)

(A2') Nem létező névszói jegykombinációk

a. egymást kizáró testvérjegyek

személyes névmás /birtokos nem lehet egyszerre 1. és 2. személyű:

*<PERS<<1><2>>>

*<POSS<<1><2>>>

egyszerre két esetjelölés/névutójelölés nem lehetséges:

*<CAS<<A>>>

*<POSTP<<A>>>

b. kötelezően folytatandó jegyek

nem-nominativusi esetnél/névutónál meg kell jelölni a konkrét esetet

*<CAS>

*<POSTP>

(A3) Az igei inflexiós jegyhierarchia

VERB

<MODAL>

<PAST>

<COND>

<SUBJUNC-IMP>

<INF>

<PLUR>

<PERS

<1

<OBJ2>>

<2>>

<DEF>

(A3') Nem létező igei jegykombinációk

a. egymást kizáró testvérjegyek

a jelölt idő/módok és az inf. páronként kölcsönösen kizárók:

<PAST> | <COND> | <SUBJUNC-IMP> | <INF>

ige nem lehet egyszerre 1. és 2. személyű:

*<PERS<<1><2>>>

b. kötelezően folytatandó jegyek

véges ige nem-3. személye esetén a személyt meg kell jelölni:

*<-INF><PERS>

(A4) A hummorph képzői (a más elemzőkben gyakran inflexióként elemzettek kiemelve)

	VERB	NOUN	ADJ	ADV	NUM
VERB	-gat [FREQ] -tat [CAUS] -ódtik [MEDIAL]	-ás [GERUND]	-ó [IMPERF_PART] -ott [PERF_PART] -andó [FUT_PART] -ható [MODAL_PART] -hatatlan [NEG_MODAL_PART] -atlan [NEG_PERF_PART]	-va [PART] -ván [PERF_PART]	—
NOUN	-ozik [ACT] -ol [ACT2] -odik [REG_ACT]	-né [MRS] -cska [DIMIN]	-os [ATTRIB] -i [MET_ATTRIB] -jú [INAL_ATTRIB] -tlan [NEG_ATTRIB] -mentes [NEG_ATTRIB2] -nyi [QUANTITY] -szerű [TYPE1] -féle [TYPE2] -beli [LOC_INE]	-stul [COM] -képpen [ESS-FOR] -nként [PERIOD]	—
ADJ	-it [TRANS_RESULT] -ul/-odik [INTRANS_RESULT]	-ság [ABSTRACT]	-bb [COMPAR] -bbik [COMPAR_DESIGN] leg- -bb [SUPERLAT] leg- -bbik [SUPERLAT_DESIGN] -cska [DIMIN] -as [ATTRIB]	-an/-ul [MANNER]	—
NUM	—	-dika [DATE]	-as [ATTRIB] -szori [ITER_ATTRIB] -szoros [MULTIPL_ATTRIB]	-an [AGGREG] -szor [MULTIPL-ITER] -adszor [ORD-ITER]	-ad [FRACT] -adik [ORD]
POSTP	—	—	-i [ATTRIB]	—	—

A general-purpose morphological annotation system

Abstract: Ideally, a morphological annotation scheme should encode all and only morphosyntactic information, abstracting away from spoken and written form. Annotation schemes based directly on the component morphs inherit the difficulties of segmentation, in particular for fused morphemes and suppletive forms. In this paper we propose an indirect annotation scheme based on binary morphosyntactic features (attribute–value pairs) arranged in a tree in such a manner that only positive (marked) features can dominate other nodes. This restriction makes it possible to convert the structure to unary features, thereby directly reflecting the markedness of a form. In the last part of the paper we describe some applied computational systems that employ this annotation scheme.

Keywords: Hungarian, morphology, annotation, tagging, inflection
