Egy általános célú morfológiai annotáció

REBRUS PÉTER – KORNAI ANDRÁS – VARGA DÁNIEL

Bevezetés

Cikkünk a morfológia annotáció általános kérdéseit tárgyalja a magyar 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 kódhossz*t 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.

Az allomorf-alapú 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 transzparenseknek 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 ilyan 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 ld. később).

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 tőbeli 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 elölképzett magánhangzójú változatban is megjelenhet (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 középnyílt 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 `-Vk` 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 „alap”allomorfot 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 `-Vk` 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, amelyben 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 tővégtől függő *sz~ol/el/öl* váltakozás (pl. *kap-**sz** ~ mos-**ol***) 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 lexémájához tartozó (pl. *sarki bolt*), de lehet a SARK lexémához tartozó (pl. *sarki expedíció*). Tehát ha ezek az alakok önmagában 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. ÁR1, ÁR2, ÁR3 – különítik el egymástól). Hasonló homonímia-jelenségek léphetnek fel a toldalékokban is: a *-k* toldalék nemcsak névszók többes számára utalhat (pl. *ház-a**k***), hanem igéknél az E.1 (pl. én *kap-j-a**k***) és bizonyos esetekben a T.3 (pl. ők *kap-t-a**k***) szám/személyre is. Ugyanígy az *-i* toldalék nemcsak melléknévképző (ld. *sarki*), hanem utalhat a birtok többes számára (pl. *hajó-**i***, *Pál-é-**i***). Tehát a morfoszintaktikai kódoláshoz a toldalékok alakja sem ad elégséges információt. Tanulságos összjátékot mutat a homonímia és a nem-fonológiai 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:
fára:	`fá+ra`	`fa+ra`	`fÁ+rA`
szelek:	`szel+ek`	**`szél+k`** (vö. *szélek*)	`?szEl+Vk ?szEl+k`
sarki:	**`sark+i`** (vö. *sark – sarok*)	`sarok+i, sark+i`	`sarOk+i, sark+i`
kaptam (vmit):	**`kap+t+am`** (vö. *kapok vmit*)	**`kap+t+k`** (vö. *ők kaptak*)	`?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 `fa+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 összemossák a tő- vagy a toldalékallomorfokban potenciálisan jelen levő ***homonímiákat***, így önmagukban nem elegendőek a szóalakban levő morfológiai információk megadásához (ld. pl. a *szélek*, a *sarki* és a *kaptam vmit* alakok fenti esetét).

Szegmentálás

A fentebb bemutatott annotációs megközelítéseknek egy további súlyos következménnyel is szembesülniük kell, ez pedig a morfológiai ***szegmentálás*** bizonytalansága. A morf-alapú 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 ***írásképp***el 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`, `és+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 morfhatáron lezajló ***hasonulás***okkal kapcsolatos. Így például a *-val/-vel* toldaléknak vagy a felszólító mód *-j* toldalékának egyes mássalhangzó utáni változatai esetén nem világos a szegmentálás (`hát+tal`, `hátt+al` vagy `hát+val`, illetve `fus+sa` vagy `fut+ja`). Ha a tő digráfra végződik, akkor a két említett probléma együtt jelentkezik: pl. `ác+csal`, `áccs+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') "átelemzett" sztring	(b) allomorfiamentes
hússzor:	**`hús+szor`** (vö. *hús*)	`húsz+szor`	
szebből:	**`szeb+ből`**	`szebb+ből`	
háttal:	**`hát+tal hátt+al`**		`hát+val`
áccsal:	**`ác+csal áccs+al`**	`ács+csal`	`ács+val`
fussa:	`fus+s+a`		**`fut+j+a`** (vö.: *futja*)
eddz:	**`ed+dz`**	**`edz+dz`**	`edz+j`

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 a végső kódolásban megjelenítendő morfoszintaktikai információ. Azaz a szegmentálás és az allomorfok kiválasztása az elemző belügye, és ***nem*** lehet ***része*** a morfoszintaktikai ***annotációnak***.

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 morfémák*** esetén több funkció szételemezhetetlenül társul egy morfhoz; legismertebb példa a magyarban a birtokos alakok és igék szám/személy jelölése. Az igazán problematikus esetek azonban azok, amikor a fúzió csak bizonyos esetekben áll fenn, máskor a morfémák 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 definitséget, tehát ezek az allomorfok nem alkalmasak a morfológiai annotáció jelölésére. A fúzióhoz tartozó jelenség az, 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 szételemzett `kalap+ja+i+m` alakban a `ja` szekvencia nem bír morfoszintaktikai szereppel (ennek absztraktabb nyelvészeti elemzését, ld pl. 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, melyek "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, hogy kapjanak (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
adná	`?ad+na+a`	`ad+NÁ+JA`
kalapjaim	`?kalap+ja+i+m`	`kalap+K+m`
engem	`?én+m`	`ÉN+T`
enyém	`?én+m`	`ÉN+É`
benneteket	`?benn+etek+et`	`TI+t`
gyere	`?gyer+e`	`JÖN+J`

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

A kizárólag morfoszintaktikai kategóriákon alapuló annotáció

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*, ld Erjavec 1997), amelyben a morfoszintaktikai ***kód rögzített hosszúságú***: egy jegyértékekből álló sztring, mely sztring 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 *fiaitokéival* főnévi, illetve az *ad* és az *adtátok* igei alakok MSD-kódrendszer szerinti annotációi az alábbiak

(4) Két főnévi és igealak MSD-annotációja

fiú	`Nc-sn-y---`	*ad*	`Vmip3s---n------`
fiaitokéival	`Nc-pi-yp2p`	*adtátok*	`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 kevéssé 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) lát-**lak** (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. *fiaitokéinak* vagy *adhattátok*) és egy ilyen szempontból jelöletlen szóalak (pl.

fiú vagy *ad*) annotációja ugyanolyan komplexitású. További problémája az, hogy egyelőre ***csak inflexiós*** kódrendszer, és nem nyilvánvaló, hogy a morfoszintaktikailag releváns képzések hogyan illeszthetők bele (különösen igaz ez a szófajváltó képzésekre).

Jegy-érték szerkezetek

A fenti problémák egy részére megoldást jelent a hierarchikus jegy–érték struktúrák (pl. az ún. AVS-ek, *Attribute-Value Structures*, ld. 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 *FIÚ*
KATEGÓRIA *FŐNÉV*
- **SZÁM** *TÖBBES*
- **BIRTOKOS** *IGEN*
  - **SZÁMA** *TÖBBES*
  - **SZEMÉLYE** *2.*
- **BIRTOK** *IGEN*
  - **SZÁMA** *TÖBBES*
- **ESET** *DATIVUS*

(b) *fiú*

LEMMA *FIÚ*
KATEGÓRIA *FŐNÉV*
- **SZÁM** *EGYES*
- **BIRTOKOS** *NEM*
  - (SZÁMA X)
  - (SZEMÉLYE X)
- **BIRTOKOS** *NEM*
  - (SZÁMA X)
- **ESET** *NOMINATIVUS*

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

(6) Két igei alak sematikus jegy–érték struktúrája

(a) *adhattátok*

LEMMA *AD*
KATEGÓRIA *IGE*
- **MODÁLIS** *IGEN*
- **IDŐ** *MÚLT*
- **MÓD** *KIJELENTŐ*
- **SZÁM** *TÖBBES*
- **SZEMÉLY** *2*
- **DEFINITSÉG** *IGEN*

(b) *ad*

LEMMA *AD*
KATEGÓRIA *IGE*
- **MODÁLIS** *NEM*
- **IDŐ** *JELEN*
- **MÓD** *KIJELENTŐ*
- **SZÁM** *EGYES*
- **SZEMÉLY** *3*
- **DEFINITSÉG** *NEM*

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 zérusmorfot vagy morfémát tartalmaznak) és jelentősen gyakoribbak (funkciójuk általánosabb, használatuk kiterjedtebb).

Bináris és unáris jegyek: főnevek

Ha azonban az AVS-ekben a jegyeket úgy fogalmazzuk meg, azok értéke csak az igen/nem (+/–) lehessen, és az értékek közül szisztematikusan az egyik a ***jelölt***et (szokásosan a +), a másik a jelöletlent (ez általában a –) jelentse, akkor ezen a ***bináris*** jegyrendszeren jelentős egyszerűsítést tehetünk (ld. 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 ld. fenti (5), ill (6)). (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 lexémától: 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*), de 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).

Lássuk, hogy az (5) és (6)-beli példáink milyen bináris jegyszerkezetet kapnak (az újonnan bevezetett jegyek nyelvészeti értelmezéséről ld. 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 jegynek 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 alakkal familiáris viszonyban álló emberek csoportjára utal; így a FAMILIÁRIS jegyet ésszerű a TÖBBES jegy alá rendelni (erről ld. részletesen Kornai 1989). A (ii) eset akkor áll 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 (melyek a további lehetőségeket adják meg) kötelezően negatív értekkel 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) *fiaitokéinak*

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

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 egy teljes értékű annotációt kapunk. Az alábbi (8a,b) 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: (8a',b')-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 (NOUN)
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)

(b) hierarchikus formában:

fiú

fiú
FŐNÉV (NOUN)

(a') linearizált formában:

fiaitokéinak

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

(b') linearizált formában

fiú

`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 kétfélék 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 birtokosjelölővel 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>>`

Megemlítjük, hogy a PLUR jegyre a hierarchia három különböző helyén is szükség van: közvetlenül a főkategória-jegy (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 — mi többes számú?

nincs birtokos- és birtokjelölés:

fiúk	`fiú/NOUN<PLUR>`	(entitás)

csak birtokosjelölés (itt 3. személyű):

fiai	`fiú/NOUN<PLUR><POSS>`	(entitás)
fiuk	`fiú/NOUN<POSS<PLUR>>`	(birtokos)
fiaik	`fiú/NOUN<PLUR><POSS<PLUR>>`	(entitás és birtokos)

csak birtokjelölés:

fiúké	`fiú/NOUN<PLUR><ANP>`	(entitás)
fiúéi	`fiú/NOUN<ANP<PLUR>>`	(birtok)
fiúkéi	`fiú/NOUN<PLUR><ANP<PLUR>>`	(entitás és birtok)

birtokos- és birtokjelölés is (csak azok, ahol a birtok többes számú):

fiáéi	`fiú/NOUN<POSS><ANP<PLUR>>`	(birtok)
fiaiéi	`fiú/NOUN<PLUR><POSS><ANP<PLUR>>`	(entitás és birtok)
fiukéi	`fiú/NOUN<POSS<PLUR>><ANP<PLUR>>`	(birtokos és birtok)
fiaikéi	`fiú/NOUN<PLUR><POSS<PLUR>><ANP<PLUR>>`	(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 alakkal (pl. *sógorodék*) és a

birtokjelölős alakkal (*sógoréké*). Ezért az annotáció egy a NOUN alatti PLUR általi dominált FAM jegy segítségével történik.

(11) Familiáris többes alakok

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

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 használjuk. A CAS jegy azt jelzi, hogy itt egy jelölt (azaz nem nominativusi) alakkal van dolgunk. A 18 eset annotációja és az esetek elnevezése az alábbi.

(12) Az esetek annotációja

„strukturális esetek”

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

„lexikális esetek”

helyhatározói

forrás

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

hely

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

cél

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

egyéb

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

Bináris és unáris jegyek: igék

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

(13) Két igealak bináris jegy-érték struktúrája

(a) *adhattátok*

ad +
IGE +
MODÁLIS +
NEM JELEN.KIJ +
MÚLT IDŐ +
FELTÉTELES M. –
KÖTŐ-FELSZ M. –
INFINITÍVUSZ –
TÖBBES SZÁM: +
NEM 3. SZEMÉLY+
1. SZEMÉLY –
(TÁRGY 2. SZEMÉLYŰ X)
2. SZEMÉLY +
DEFINITSÉG +

(b) *ad*

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

(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)
DEFINITSÉG (DEF)

(b) hierarchikus formában:

ad

ad
IGE (VERB)

(a') linearizált formában:

adhattátok

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

(b') linearizált formában

ad

```
ad/VERB
```

Az igék 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 igék négy idő/módjának annotációja

ad	`ad/VERB`	jelen idő kijelentő mód
adott	`ad/VERB<PAST>`	múlt idő kijelentő mód (*past*)
adna	`ad/VERB<COND>`	jelen idő feltételes mód (*conditional*)
adjon	`ad/VERB<SUBJUNC-IMP>`	kötő-felszólító mód (*subjunctive-imperative*)

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 a `<1>` jegy alá bevezetett `<OBJ2>` jeggyel történik. A definit–indefinit (általános–határozott) 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

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

Az infinitívusz szám/személy jelölésének annotációja igen hasonló az igékéhez. Az infinitivus, mint jegye az igéknek (`VERB<INF>`) sajátos jegykombinációkat enged csak meg: az infinitívusznak nincsen idő/módja és definitsége, viszont lehet szám/személye, melyet 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. (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*.)

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

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

Összefoglalva, az unáris jegyekkel való hierarchikus ábrázolás lehetőséget teremt arra, hogy egyszerűen megfogalmazható és nyelvészetileg 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 (allomorfok, szegmentálás stb.) az annotációs kód 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 állnak. Minden további morféma tovább növeli az annotáció bonyolultságát. Az alábbi táblázatban néhány ilyen „monotonon bővülő” komplexitású alaksort és annotációikat adtunk meg a zérustoldaléktól a maximális alakokig (az összehasonlítás kedvéért a hozzávetőleges morfémahatárokat a szóalakokban jelöltük).

(18) Főnévi és igei szóalakok és annotációik egy-egy monoton növő komplexitású sora

fiú	fiú/NOUN
fiú-k	fiú/NOUN<PLUR>
fi-a-i	fiú/NOUN<PLUR><POSS>
fi-a-i-d	fiú/NOUN<PLUR><POSS<2>>
fi-a-i-tok	fiú/NOUN<PLUR><POSS<PLUR><2>>
fi-a-i-tok-é	fiú/NOUN<PLUR><POSS<PLUR><2>><ANP>
fi-a-i-tok-é-i	fiú/NOUN<PLUR><POSS<PLUR><2>><ANP<PLUR>>
fi-a-i-tok-é-i-t	fiú/NOUN<PLUR><POSS<PLUR><2>><ANP<PLUR>><CAS<ACC>>
ad	ad/VERB
ad-hat	ad/VERB<MODAL>
ad-hat-ott	ad/VERB<MODAL><PAST>
ad-hat-t-ak	ad/VERB<MODAL><PAST><PLUR>
ad-hat-t-atok	ad/VERB<MODAL><PAST><PLUR><PERS<2>>
ad-hat-t-á-tok	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 definitség mindkét értékét felvehetik. Néhány példa.

(19) Morfológiai kétértelműségek kezelése alulspecifikáció nélkül

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

Az inflexiós annotáció

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 ezeknek pozitív vagy negatív ***érték***ei szerepelnek. A másik komponens a hierarchiáért felelős, ezt a legegyszerűbb egy ***irányított körmentes gráf***ként (azaz irányított faként) meghatározni, melyben minden csomóponthoz egy bináris jegy–érték pár van rendelve, az irányított élek pedig megfelelnek a jegy–érték párok közötti dominancia viszonyoknak. 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 ***dominál***nak 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. (Valójában az annotációt közvetlenül unáris jegyekkel címkézett fagráffal is definiálhatnánk, ekkor egy annotáció ennek a jegyekkel címkézett fának az olyan részfája lenne, amelynek a gyökércsomópontja megegyezik a bővebb fáéval.)

Ahogyan 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). A hunmorph annotációs rendszer aktuális változata által használt főkategória-jegyek listája megtalálható a függelék (A1) ábrájában. Minden inflektálható kategóriához tartozik

egy rögzített inflexiós jegy–érték struktúra, azaz bináris jegy–értékekkel címkézett csomópontú fa-gráf. Inflektálható kategória azonban csak öt van: a három névszói és a ragozható determinánsi és az egy igei kategória, ezek jegy–érték szerkezeteiről ld. 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 a félreolvasás lehetősége a lehető legkisebb legyen (az inflektálható kategóriákhoz tartozó jegyek kimerítő listáját és sorrendjüket ld. a következő részben). Így a linearizált annotáció már egy egyértelmű, kódok 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 fa-gráf, amelynek minden főnévi jegye negatív (azaz esetünkben az egyes számú nem-birtokos nem-birtok nominativusi alak, ld. (7b), illetve unáris formában (8b, b'); hasonlóan igékre, ld. (7a), illetve (8a,a')). 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:

(21) Az inflexiós annotáció sémája

szóalak `lemma/FŐKATEGÓRIA<TOVÁBBI_INFLEXIÓS_JEGYEK>`

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 -- ld. 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*).

(22) Ingadozó alakok azonos annotációt kapnak

fotelban	`fotel/NOUN<CAS<INE>>`
fotelben	`fotel/NOUN<CAS<INE>>`
fürdenek	`fürdik/VERB<PLUR>`
fürödnek	`fürdik/VERB<PLUR>`
gyere	`jön/VERB<PERS<2>><SUBJUNC-IMP>`
jöjj	`jön/VERB<PERS<2>><SUBJUNC-IMP>`
jöjjél	`jön/VERB<PERS<2>><SUBJUNC-IMP>`

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

Névszói kategóriák

Régi problémája a leíró nyelvtanoknak, hogy be lehet-e (és ha igen, hogyan) sorolni egyértelműen a névszói alakokat valamelyik névszói kategóriába (ld. többek között Moravcsik 1997). A melléknevek és a számnevek a főnevekkel átfedő osztályokat alkotnak, és nehéz egyértelmű disztribúciós teszteket adni, amelyek alapján ezek a kategóriák egyértelműen megkülönböztethetőek 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 kiváncsi**akat**.*, *Ez az én nagy labdám, az meg a te kicsi**d**.*, *Bátr**aké** a szerencse.*, *Összeültek a nyolc**ak**.*, *Négy**et** rendeltem.*, *Az ő öt könyve meg az én hárm**am***. Itt és a többi hasonló példában vitatható, hogy az adott melléknév vagy számnév a saját „prototipikus" mondattani *funkciójában* szerepel-e, de melléknév, illetve számnév voltuk 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* -féle szerkezekben az első főnév módosító szerepű (ahogyan tipikusan a melléknevek), a *szépek imádata* és a *kevés is sok*-féle 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 deadjektiválisnak tartott képzések, mint amilyen a közép- és felsőfok, sem adnak jobb fogódzót, ezek a mellékneveken kívül egyes számnevekkel is lehetségesek (pl. *több*, *kevesebb, legelső*), és egyes konstrukciókban főnevekhez is: 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*, ld. Csendes 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 néhány példa inflektált alakokra.

(23) Melléknévi és számnévi alakok névszói inflexiókkal

kíváncsi	`kíváncsi/ADJ`
kíváncsijaitokét	`kívácsi/ADJ<PLUR><POSS<PLUR><2>><ANP><CAS<ACC>>`
kétezer	`kétezer/NUM`
kétezreinkével	`kétezer/NUM<PLUR><POSS<PLUR><1>><ANP><CAS<INS>>`

Determinánsok

A negyedik inflektálható kategória, a determinánsok (DET), ld. (20). Pontosabban a determinánsok egy része inflektálható, az olyan szerkezetekben, mint pl. *ez**eké** a lányoké*, *ab**ban** 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 utolsókét felsorolt típus *(ezen, azon* stb.) kétértelmű: lehet inflektálhatatlan determináns, de lehet superessivusi esetű inflektálható is: vö. *azon emberekkel* vs. *azon az embereken*.

(24) Inflektált és nem inflektált determinánsok

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

Főnévi, melléknévi, számnévi névmások

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

(25) főnévi, melléknévi és számnévi névmások annotációja

mutató

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

kérdő

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

egyéb (vonatkozó, általános, tagadó)

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

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, melyeknek névszói inflexiós 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 – a személyre utaló PERS jeggyel. Ez a PERS jegy az infinitívuszoknál látott módon jelöli a személyt (ld. (17)): magában 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 szerepeltettük, ezek morfoszintaktikailag 3. személyűek):

(26) A személyes névmások annotációja

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

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

(27) Inflektált személyes névmások

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

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 (megjegyzendő, hogy a POSS jegy főnévi személyes névmásokra nem is használatos, hiszen az ezt kifejező alakok szisztematikusan hiányoznak: **éned, *őm, *önötök, *magánk*):

(29) Az egyes és többes számú birtokra utaló névmások annotációja

enyém	én/NOUN<PERS<1>><ANP>	*enyémek*	mi/NOUN<PERS<1>><ANP<PLUR>>
tied tiéd tiedé	te/NOUN<PERS<2>><ANP>	*tieid tiédek*	ti/NOUN<PERS<2>><ANP<PLUR>>
övé	ő/NOUN<PERS><ANP>	*övéi*	ő/NOUN<PERS><ANP<PLUR>>
öné	ön/NOUN<PERS><ANP>	*önéi*	ön/NOUN<PERS><ANP<PLUR>>
magáé	maga/NOUN<PERS>	*magáéi*	maga/NOUN<PERS><ANP<PLUR>>
mienk miénk	mi/NOUN<PLUR><PERS<1>><ANP>	*mieink*	mi/NOUN<PLUR><PERS<1>><ANP<PLUR>>
tietek tiétek	ti/NOUN<PLUR><PERS<1>><ANP>	*tieitek*	ti/NOUN<PLUR><PERS<2>><ANP<PLUR>>
övék övéké	ők/NOUN<PLUR><PERS><ANP>	*övéik*	ők/NOUN<PLUR><PERS><ANP<PLUR>>
önöké	önök/NOUN<PLUR><PERS><ANP>	*önökéi*	önök/NOUN<PLUR><PERS><ANP<PLUR>>
maguké	mi/NOUN<PLUR><PERS><ANP>	*maguk*	maguk/NOUN<PLUR><PERS><ANP<PLUR>>

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 pórája – Kinek a pórája? – Az én kutyámé / Az* ***enyémé*** */ ?# Az enyém*. Ez utóbbi alak nem egyszerű birtokoltságot, hanem a birtok

általi újabb birtokoltságát fejez ki, és bizonyos mértékig itt is kifejezhető mindkét birtok többes száma: *A kutyáim pórája – az enyémeké.*; *A kutyám pórázai -- ?az enyéméi.*; *A kutyáim pórázai -- az ?enyémekéi.* (Megjegyezzük, hogy ugyanez a szerkezet nem-névmási alakokkal csak nagyon nehezen fogadható el: *?*Páléé*, talán kicsit grammatikusabbá válik, ha az első birtokjelölő után többes számjelölés van: *??Páléié*). Ezt a szerkezetet az ANP csomópont alatti újabb ANP csomóponttal lehet kódolni.

(30) Többszörös birtokjelölés

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

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

(31) Inflektált birtokos névmások növekvő komplexitású sora

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

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, amelyek 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 speciális névutós személyes névmásoknak, és mivel a névmások PERS jeggyel ellátott főnevek, ezért érdemes bevezetni a főnévi inflexiós rendszerbe a POSTP jegyet (amely a POSTP főkategória jegytől különbözik). A főnév alatti POSTP jegynek aljegye lesz az összes névutós jegy, mely jegyek 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 különböző 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 'őelőtte', és nem 'ön/maga előtt', és hasonlóan a T.3 *közéjük* 'őközéjük' és nem 'önök/maguk közé'.

(32) Személyragozott és sima névutók annotációja

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

Figyeljük meg, hogy egyes névutók kétértelműek: egyrészt vannak olyanok, amelyek E.3 alakjuk 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: 'őszámára' (pl. *Számára nincs kegyelem.*). Bizonyos esetekben a mutató névmás és a névutó egybeírandó (pl. *anélkül*, *ezelőtt*, *amiatt*). Ekkor az annotáció szintén a főnévi kategóriát használja a POSTP jeggyel.

(33) Személyragozott és sima névutók annotációja

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

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

Ige

Az igei alakok jelentős rész 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 a *hajnalodik* ige vagy a *kell, lehet* segédigék.A *rejlik, történik* stb. igéknek vannak többes számú alakjaik is, de csak 3. személyben (és indefinitként) állnak. Az intranzitív 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, infinitívuszban 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.

Képzés és szóösszetétel

A képzés annotációja

Bizonyos feladatok (tartalomelemzés, gépi fordítás) megoldásához szükség lehet az inflexiós réteg kielemzésén túli morfológiai elemzésre is. Tipikus példa erre a szintaktikai elemzés (parsing), hiszen a melléknévi igeneves szerkezetek 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* igetövet 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, morfotaktikai 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, melynek 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.

(38) 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.

(39) Példák inflektált képzésekre

terjedő	`terjed/VERB[IMPERF_PART]/ADJ`	(foly. mn-i igenév)
csináltathatnék	`csinál/VERB[CAUS]/VERB<MODAL><COND><PERS<1>>`	(kauzatív)
székestül	`szék/NOUN[COM]/ADV`	(társ)
lábúakat	`láb/NOUN[INAL_ATTRIB]/ADJ<PLUR><CAS<ACC>>`	(elidegeníthetetlen tul.)
oroszul	`orosz/ADJ[MANNER]/ADV`	(mód)
okosabbjaival	`okos/ADJ[COMPAR]/ADJ<PLUR><PERS><CAS<INS>>`	(középfok)
sokszor	`sok/NUM[MULTIPL-ITER]/ADV`	(multiplikatív–iteratív)
nyolcadikak	`nyolc/NUM[ORD]/NUM<PLUR>`	(sorszám)
szembeni	`szemben/POSTP[ATTRIB]/ADJ`	(tulajdonság)
utániakról	`után/POSTP[ATTRIB]/ADJ<PLUR><CAS<DEL>>`	

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

(40) Többszörös képzést annotáló elemzés

szóalak
`lemma/KATEGóRIA_1[KÉPZŐ_1]/...KATEGÓRIA_n[KÉPZŐ_n]/VÉGSŐ_KATEGÓRIA<INFL_ANNOTÁCIÓ>`

(41) Példák többszörös képzésekre

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

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

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 és néhány példát az alábbiakban találunk.

(42) Az elemzés formalizmusa szóösszetétel esetén

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

(43) Példák összetételek elemzésére

inflektált összetételek:

eladja	`el/PREV+ad/VERB<PERS>`
vérfarkasok	`vér/NOUN+farkas/NOUN<PLUR>`
sötétzöldje	`sötét/ADJ+zöld/ADJ<POSS>`
kanárisárgáé	`kanári/NOUN+sárga/ADJ<ANP>`

képzett összetételi tagok:

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

többszörös összetétel:

birsalmasajt	`birs/NOUN+alma/NOUN+sajt/NOUN`

Alkalmazások

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

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 2006). A szintén hunlex névre hallgató szoftver a nyelvészek által könnyen olvasható és módosítható magas szintű morfológiai erőforrásból egy úgynevezett aff/dic (affixum-lista/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.

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 2005).

Morfológiai egyértelműsítés

A ***hunpos*** morfológiai egyértelműsítő azt a problémát orvosolja, hogy egy adott szóalakhoz a morfológiai elemző gyakran több legális elemzést is talál (Halácsy 2007). Az egyértelműsítő ezek közül választja ki a mondatbeli szövegkörnyezet 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 statisztikus szabályszerűségek kiaknázásával – elfogadható (tokenenként 95% feletti) pontossággal megoldható.

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 2007). A jelen kötetben bemutatott ***hunchunk*** NP-daraboló (főnévi csoport felismerő) is támaszkodik működése során a morfológiailag feldolgozott szövegre (ld. Recski–Varga tanulmányát a jelen kötetben).

Gyakorisági korpusz

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

Bibliográfia

Csendes, D, J Csirik, T Gyimóthy (2004) The Szeged Corpus: A POS tagged and syntactically annotated Hungarian natural language corpus, Lecture notes in computer science, ISSU 3206, pages 41-48, Springer

Erjavec, T. and Monachini, M. (szerk.) (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/.

Szalai András, Halácsy Péter (2008). Szószablya Gyakorisági Szótár. http://szotar.mokk.bme.hu/szoszablya/

Halácsy, P, A Kornai, Cs Oravecz (2007) HunPos – an open source trigram tagger In.: Proceedings of the ACL 2007 Demo and Poster Sessions, pages 209-212, Prague 2007. Association for Computational Linguistics

Kornai András (1989) A főnévi csoport egyeztetése. In Telegdi and Kiefer, editors, Általános Nyelvészeti Tanulmányok, XVII. Akadémiai Kiadó, Budapest.

Kornai András, Rebrus Péter, Vajda Péter, Halácsy Péter, Rung András, Trón Viktor (2004). Általános célú morfológiai elemző kimeneti formalizmusa.In: Alexin Zoltán, Csendes Dóra (szerk.). II. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Szeged, p. 172-176.

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 (ebben a kötetben) Magyar főnévi csoportok azonosítása. Általános Nyelvészeti Tanulmányok.

Trón Viktor (2002) Attribútum–érték struktúrák. In Kálmán László, Trón Viktor és Varasdi Károly (szerk.) Lexikalista elméletek a nyelvészetben. Tinta Könyvkiadó, Budapest.

Trón, V, Németh, L, Halácsy, P, Kornai, A, Gyepesi, G, and Varga, D (2005). Hunmorph: open source word analysis, In: Proceedings of the ACL. ACL 2005.

Trón, V, P Halácsy, P Rebrus, A Rung, P Vajda és E Simon (2006) Morphdb.hu: Hungarian lexical database and morphological grammar, In: Proceedings of 5th International Conference on Language Resources and Evaluation. ELRA, pages 1670-1673.

Varga Dániel, Simon Eszter. Hungarian named entity recognition with a maximum entropy approach. Acta Cybernetica Volume 18 Issue 2. 2007.

Függelék

(A1) A hunmorph főkategória-jegyei

(i) inflektálható (és továbbképezhető)

névszóként:

`NOUN` főnév (*noun*)
`ADJ` melléknév (*adjective*)
`NUM` számnév (*numeral*)
`DET` determináns (*determiner*) [8]

igeként:

`VERB` ige (*verb*)

(ii) nem inflektálható (és nem továbbképezhető)

`ADV` határozószó (*adverb*)
`POSTP` névutó (*postposition*) – marginálisan esetjelölés és mn-képzés lehetséges
`ART` névelő (*article*) [3]
`CONJ` kötőszó (*conjunction*) [198]
`PREP` prepozíció (*preposition*) [5]
`PREV` igekötő (*preverb*) [105]
`UTT-INT` mondatszó/indulatszó (*utterance*/*interjection*) [273]
`ONO` hangutánzó (*onomatopoeic*)

(iii) egyéb, írott nyelvi kategória

`PUNCT` központozási jel (*punctuation*)

(A2) A névszói inflexiós jegyhierarchia

```
{NOUN ADJ NUM DET}
     [ <POSTP                          (POSTP jegy csak NOUN esetén)
            <ELŐTT>
            ...                  (az összes névutó)
       >              ]
       <PLUR
            <FAM>
       >
       <PERS
            <1>
            <2>
       >
       <POSS
            <PLUR>
            <1>
            <2>
       >
       <ANP
            <PLUR>
            <ANP
                  <PLUR>
            >
       >
       <CAS
            <ACC>
            <DAT>
             ...                 (összesen 17 jelölt esetrag)
```

(A2b) Nemlétező névszói jegykombinációk

egymást kizáró testvérjegyek:

*<PERS<<1><2>>>	személyes névmás nem lehet egyszerre 1. és 2. személyű
*<POSS<<1><2>>>	birtokos nem lehet egyszerre 1. és 2. személyű
*<CAS<<*A*><*B*>>>	egyszerre két esetjelölés nem lehetséges
*<POSTP<<*A*><*B*>>>	egyszerre két névutó nem állhat egy „főnévi” alakon belül

kötelezően folytatandó jegyek:

*<CAS>	jelölt (nem-nominativusi) esetnél meg kell jelölni a konkrét esetet
*<POSTP>	a névutónál meg kell jelölni a konkrét névutót

(A3) Az igei inflexiós jegyhierarchia

```
VERB
      <MODAL>
      <PAST>
      <COND>
      <SUBJUNC-IMP>
      <INF>
      <PLUR>
      <PERS <1
                  <OBJ2>
            >
            <2>
      >
      <DEF>
```

(A3b) Nemlétező igei jegykombinációk

egymást kizáró testvérjegyek:

<PAST>|<COND>|<SUBJUNC-IMP1>|<INF>	a jelölt idő/módok és az inf. páronként kölcsönösen kizárók
*<PERS<<1><2>>>	ige nem lehet egyszerre 1. és 2. személyű

kötelezően folytatandó jegyek:

*<-INF><PERS>	véges ige esetén a nem-3. személy esetén a személyt meg kell jelölni

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

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

egyéb:

NUM – NUM (számnévből számnevet képzők): *-ad* [FRACT] *-adik* [ORD]

POSTP – ADJ (névutóból melléknevet képző): *-i* [ATTRIB]