Általános célú morfológiai elemző kimeneti formalizmusa

Kornai András*, Rebrus Péter**, Vajda Péter* Halácsy Péter***, Rung András**, Trón Viktor†

Kivonat Az alábbi írásban egy a szóalakok morfoszintaktikai ábrázolására használható formalizmust mutatunk be. A formalizmus alapvető adatstruktúrája egy speciális fagráf, amely mind inflexiós mind derivációs információ megragadására alkalmas. A fagráfból rekonstruálható egy inflexiós információt tartalmazó teljes bináris jegy–érték-struktúra, ugyanakkor redundanciamentes és egyszerű linearizálhatósága folytán jól alkalmazható egy általános célú morfológiai elemző kimeneti kódrenszereként. Nyelvészeti megalapozottsága alkalmassá teszi arra is, hogy a szótári tételeket morfoszintaktikai viselkedését megjelenítse egy morfológiai adatbázisban.

1. Bevezetés

Egy morfológiai elemző kimeneti formalizmusának három, egymásnak gyakran ellentmondó feltételt kell kielégítenie. Ezek az (i) *informativitás:* a lehető legpontosabban és legteljesebben tükrözze a szóalakokból megállapítható morfológiai információkat; (ii) *adekvátság:* nyelvészetileg megalapozott kategóriákat használjon; (iii) *egyszerűség:* kézi, illetve automatikus feldolgozásra egyaránt könnyen használható legyen. Mindhárom feltétel nagy mértékben függ a felhasználó céljaitól: minél pontosabban határozzák meg az elemzés célját (pl. helyesírás-ellenőrzés, tövezés, szintaktikai elemzés, korpuszalapú statisztikai vizsgálatok), annál könnyebben lehet megfelelő egyensúlyt találni közöttük. Egy előre nem specifikált célú morfológiai elemző esetében azonban nincs mód arra, hogy a kimeneti annotáció tökéletes legyen. Cikkünk azt kívánja bemutatni, hogy a Szószablya projekt([2]) keretében elkészült HunTools szóelemző eszköztár ([5,4]) kimeneti kódrendszerének megalkotásakor a fenti feltételeket hogyan próbáltuk meg összehangolni.

Az alábbiakban először az inflexiós toldalékolás ábrázolására használt hierarchikus struktúra alapelveiről szólunk. Bemutatjuk, hogy ez a gazdag struktúra hogyan linearizálható és egyszerűsíthető a morfológiai jelöltség fogalmának felhasználásával (§2). Ezek után ismertetjük a magyar morfológiai elemző kimenetében használt főnévi és igei inflexiós kategóriarendszert, röviden kitérve a deriváció kezelésére is (§3). Végül megemlítjük, hogy a formalizmus alkalmas különböző kódolást alkalmazó morfológiai elemzők összehasonlítására is (§4).

* MetaCarta Inc., e-mail: `andras@kornai.com`

** MTA Nyelvtudományi Intézet, e-mail: `{rebrus,vajda}@nytud.hu`

*** BME Média Oktató és Kutató Központ `{hp,runga}@mokk.bme.hu`

† IGK, Saarland University, University of Edinburgh `v.tron@ed.ac.uk`

2. Az inflexió fagráfos ábrázolása

Egy szóalak morfológiai ábrázolása úgy teljes, ha az összes inflexiós tulajdonsága specifikálva van. Az inflexiós tulajdonságok túlnyomó része a mondattani elemzésben játszik fontos szerepet, szintaktikai szabályok hivatkoznak rájuk. Az ilyen morfoszintaktikai tulajdonságok szokásosan ún. jegy–érték-struktúrákkal (attribute-value structure, AVS) [6] adhatók meg. Ez a struktúra a szóalak morfoszintaktikai vetületét hivatott ábrázolni és, mint ilyen, független az azt kódoló formai jegyektől és a szóalak felszíni ábrázolásától. Meggyőződésünk, hogy az a morfoszintaktikai annotáció, amely elvonatkoztat a morfszegmentálástól (amely valójában a morfológia item-and-arrangement felfogásához kötődik), elméletsemlegessége és modularitása miatt szélesebb körű alkalmazást tesz lehetővé.

A szóalak morfológiai jegyei nem homogének; két jellemzőjüket kell kiemelnünk: (i) hierarchikusság: bizonyos jegyek specifikálása más jegyek jelenlétét, specifikálását feltételezi, és (ii) aszimmetria: adott jegy lehetséges értékei közül bizonyosak jelöletlennek, mások jelöltnek tekinthetők.

Ezt a két jellemzőt jól megragadja egy címkézett fagráf. A fagráf csomópontjai az inflexiós jegyek (címkéi a jegyek nevei), a fa gyökere pedig a szótári tételek inflexió szempontjából vett ekvivalenciaosztályai (az inflektálható kategóriák). A fában jelenlévő csomópontok egy a gyökér-csomópont által meghatározott jegyösvény pozitív értékét jelentik. Ez egyben azt is jelenti, hogy a gráf kizárólag olyan bináris jegy–érték-struktúrát képes kódolni, amelyben csak a pozitív értékű csomópontoknak lehetnek a fában folytatásai [3]. Ez a megszorítás a jelöltség fogalmának egy értelmezése. A fagráfos ábrázolás a fenti követelményeket mind kielégíti: (i) informatív, hiszen jegy–érték szerkezetek formájában képes ábrázolni a szóalakok releváns morfoszintaktikai tulajdonságait; (ii) adekvát, hiszen megragadja az inflexiós információ hierarchikus aspektusát és a morfológiai jelöltség fogalmait; valamint (iii) egyszerű, hiszen belőle egy teljes bináris jegy–érték-struktúra automatikusan rekonstruálható, ugyanakkor redundanciamentessége miatt tömören linearizálható.

3. A főnévi és igei alakok inflexiós kódrendszerei

A magyar nyelv morfológiai elemzéséhez a fenti formalizmusban egy konkrét jegyrendszert dolgoztunk ki. A főnevekhez az alábbi hierarchikus struktúrát rendelhetjük, ami egyben a NOUN gyökércsomópontból kiinduló gráfok szignatúrájának felel meg.

Szám:	egyes		<-PLUR>
	többes		
		„egyszerű” (pl. *sógorok*)	<+PLUR<-FAM>>
		familiáris birtokos (pl. *sógorék*)	<+PLUR<+FAM>>
Birtokos:	nincs megjelölt birtokos		<-POSS>
	van, ekkor a birtokos		

	Személye	
	1. (pl. *sógorom*)	`<+POSS<+1><-2>>`
	2.	`<+POSS<-1><+2>>`
	3.	`<+POSS<-1><-2>>`
	Száma	
	egyes (pl. *sógorai*)	`<+POSS<-PLUR>>`
	többes	`<+POSS<+PLUR>>`
Birtok:	nincs birtok	`<-ANP>`
	van; a birtok száma:	
	egyes (pl. *sógoré*)	`<+ANP<-PLUR>>`
	többes (pl. *sógoréi*)	`<+ANP<+PLUR>>`
eset:	„nincs” (= nominativus)	`<-CAS>`
	van, 16 különböző eset lehet (pl. *sógort*)	`<+CAS<+ACC>>`

Egy inflektált alak morfoszintaktikai annotációja tehát egy fagráffal adható meg. A fagráfban a `NOUN`-ból kiinduló részösvények a jegymátrix pozitív értékeit kódolják. A szignatúra által adott többi releváns jegy értéke mind negatív. A gráf tehát ekvivalens egy a szóalak inflexiós tulajdonságait leíró teljes bináris jegy–érték-struktúrával. A fagráf ábrázolás azonban redundanciamentes és a csomópontcímkék (attribútumnevek) zárójelezésével karakterláncként egyszerűen linearizálható, ennélfogva szöveges formában is egyszerűen és tömören kifejezhető. Az alábbi példák a szóalak teljes inflexiós specifikációját mutatják jegy–érték-struktúrával, valamint a linearizált fagráfos kódolással.

kutya
<+NOUN<-PLUR><-POSS><-ANP><-CAS>>
<NOUN>
kutyáink
<+NOUN<+PLUR<-FAM>><+POSS<+1><-2><+PLUR>><-ANP><-CAS>>
<NOUN<PLUR><POSS<1><PLUR>>>
kutyáéi
<+NOUN<-PLUR><-POSS><+ANP<+PLUR>><-CAS>>
<NOUN<ANP<PLUR>>>
kutyáikéit
<+NOUN<+PLUR<-FAM>><+POSS<-1><-2><+PLUR>><+ANP<+PLUR>><+CAS<+ACC>>>
<NOUN<PLUR><POSS<PLUR>><ANP<PLUR>><CAS<ACC>>>

Amint látható, a hierarchikus szerkezet lehetővé teszi, hogy egyazon primitív jegy (csomópontcímke) különböző dependensek konceptuálisan azonos morfoszintaktikai jegyeinek kódolására is használható legyen, így például a főnév, a birtokos, illetve a birtok többes száma egyaránt a `PLUR` jegy használatával fejeződik ki.

A jelöletlen morfoszintaktikai jegyértékeket a legtöbb esetben zérusmorfémák fejezik ki. Ez azzal az előnnyel jár, hogy a kódolás nagyjából tükrözi a szóalakokban található testes morfok számát (vagy az alkalmazott toldalékolási operációk számát), anélkül hogy állást foglalna a morfológiai szegmentálás (a felszíni szóalak toldalékallomorfokra bontásának) kényes kérdésében, tehát informatív, de ez nem megy az adekvátság rovására.

A hierarchikus szerkezet linearizálásában használt zárójelezési konvenció segítségével élesen elhatárolható az alulspecifikáció a teljes inflexiós specifikációtól. A <NOUN> teljesen specifikált (egyes számú, nem birtokos, nominatívuszi) alakot jelöl, míg a NOUN lexémák egy osztályát, ami mint felszíni alak minden inflexiós jegyre alulspecifikált. Formálisan a zárójelezett forma egy modellt a zárójel nélküli pedig egy modellhalmazt ír le. Az alulspecifikáció kifejezésének lehetőségével a formalizmus alkalmas a szótári tételek morfoszintaktikailag releváns paradigmatikus információjának ábrázolására (pl. *kutya* NOUN). Ez a szófaji információ közvetlenül „kompatibilis" a szótári tétel inflektált alakjának az elemző által kiadott kódjával, ugyanakkor képes kifejezni a lehetséges inflektált alakokra vonatkozó megszorításokat a szótárban. Azokat a szótári tételeket, melyek paradigmája (morfológiai okokból) hiányos (defektív, illetve minor kategóriák), az inflexiós fagráf részleges specifikációjával adhatjuk meg a szótárban. Például:

– plurale tantum (*üzelmek, üzelmei, üzelmeim*, stb., vs **üzelem*, **üzelme* stb.)
  üzelmek NOUN<PLUR>
– possessivum tantum (*eleje, elejem, elejei*, stb. vs. **ele/elő*, **elék/elők* stb.)
  eleje NOUN<POSS>
– possessivum és plurale tantum (*elei, eleim* vs. **el/ele/elő*, **elek/elők*, **ele/eleje/elője*, stb.)
  elei NOUN<PLUR><POSS>

Képzés A fentebb bemutatott fagráfok közvetlenül nem alkalmasak a szóképzés ábrázolására. Ugyanakkor a képzőket tekinthetjük szótári tételek közötti relációkként. Így a derivációs viszonyok az inflexiós ekvivalenciaosztályok (a fagráf lehetséges gyökercímkéi, pl. NOUN, VERB, ADV) közötti címkézett irányított élekkel ábrázolhatók, ami valójában az inflexiós gráfstruktúra kiterjesztése. Kimeneti kódolásukra a következő néhány példa adható:

faxol	<[fax]NOUN[l]VERB>
faxolgat	<[fax]NOUN[l]VERB[gAt]VERB>
faxolgatás	<[fax]NOUN[l]VERB[gAt]VERB[Ás]NOUN>

Az eddigi megfontolásokból automatikusan adódik, hogy sem az inflektált alakok, sem a részben specifikált (azaz nem gyökér) „tantumok" nem vethetők alá képzésnek.

4. A morfológiai elemzők kimeneti kódjainak megfeleltetése

Mivel ez az ábrázolás független a morfológiai elemzés technológiai megvalósításától, alkalmas arra, hogy több különböző morfológiai elemző kimeneti formalizmusának közös nevezője lehessen. A morfológiai elemző kimeneti kódrendszerének tervezésekor megvizsgáltuk többek között az MSD-kódrendszert [1]. A magyar nyelvre fent ismertetett kódot úgy alakítottuk ki, hogy az legalább annyi morfoszintaktikai információt tartalmazzon, mint az MSD-kódok, így az utóbbiak átalakítása egyértelmű legyen. Ehhez elkészítettünk egy transzformációs táblázatot, amely a kialakítandó kódokra való leképezést adja meg. Ennek segítségével a Szószablya projekt keretében elkészített magyar morfológiai elemzőt össze lehet vetni más kódolást használó rendszerekkel, s ezzel a kiinduló morfológiai adatbázisunk hibái és hiányai egyszerűbben javíthatók.

Hivatkozások

1. T. Erjavec and M. Monachini. Specifications and notation for lexicon encoding. Technical report, Copernicus Project 106 MULTEXT-East, December 1997.
2. Péter Halácsy, András Kornai, László Németh, András Rung, István Szakadát, and Viktor Trón. A Szószablya projekt. In *Proceedings of the 1st Hungarian Computational Linguistics Conference*. Szegedi Tudományegyetem, 2003.
3. András Kornai. A főnévi csoport egyeztetése. In Telegdi and Kiefer, editors, *Általános Nyelvészeti Tanulmányok*, XVII. Akadémiai Kiadó, Budapest, 1989.
4. László Németh, Péter Halácsy, András Kornai, and Viktor Trón. Nyílt forráskódú morfológiai elemző. In *II Magyar Számítógépes Nyelvészeti Konferencia*, 2004.
5. László Németh, Viktor Trón, Péter Halácsy, András Kornai, András Rung, and István Szakadát. Leveraging the open-source ispell codebase for minority language analysis. In *Proceedings of SALTMIL 2004*. European Language Resources Association, 2004.
6. Viktor Trón. Attribútum-érték struktúrák. In László Kálmán, Viktor Trón, and Károly Varasdi, editors, *Lexikalista elméletek a nyelvészetben*. Tinta Könyvkiadó, Budapest, 2002.