Beás nyelvű morfológiai elemző problémái a hunlex-hunmorph rendszerben

Szeredi Dániel[1]

[1] MTA Nyelvtudományi Intézet – Eötvös Loránd Tudományegyetem BTK
Elméleti Nyelvészeti Tanszéki Szakcsoport
Budapest 1068, Benczúr u. 33.
dani@szeredi.hu

Kivonat: A beás nyelv flektáló nyelv, így morfológiája nagy mértékben különbözik az agglutinatív nyelvekétől, amelyekre a hunmorph és a hunlex rendszerek a legkönnyebben alkalmazhatóak. Ennek következtében többféle probléma merül fel az elemző készítése során, ám a hunlex-ben található eszközök újféle használatával ezek kezelhetőek.

A rendszer. A hunmorph morfológiai elemző [3] felépítése leginkább az agglutinatív típusú nyelveket preferálja: az egyes szóalakokat szótőre és affixumokra választja szét. A hunlex lexikonkezelő [2] pedig előállítja a hunmorph működéséhez szükséges nyelvi erőforrást egy szabályrendszerből és egy lexikonból. A hunlex már rendelkezik olyan eszközökkel, amelyekkel nemkonkatenatív szabályok is leírhatóak, hiszen karaktereket vághat le, a töveket képes reguláris szabályok illesztésével változtatni. Kérdéses, hogy a hunmorph hogyan képes kezelni a flektáló nyelveket, amelyekre nem a könnyen szegmentálható affixumok a jellemzőek, hanem az inflexió során az egyes toldalékok összeolvadnak egymással, illetve a tővel, tehát a kimenetek lineárisan nem szegmentálhatóak.

A beás nyelv. A morfológiai elemző és a lexikonkezelő ilyen típusú nyelveken való tesztelésére a beás igen alkalmas, mivel erősen flektáló jellegű. A beás nyelvet Magyarországon élő romák beszélik, főleg a déli, délnyugati megyékben. A magyarországi cigányoknak körülbelül 5 százaléka beszéli a beást, amely a románnak közeli rokona, tehát nagyon messze áll az ismertebb, ind eredetű lovári nyelvtől. A beás formális leírása csupán a legutóbbi időkben indult meg [1], így a rendelkezésre álló adatok még nem teljeskörűek, főként az igeragozás tekintetében.

Problémák. A beás nyelvnek a hunlex–hunmorph rendszerben történő morfológiai elemzése során tehát a legfontosabb probléma az, hogy a flektáló nyelvben nem különülnek el a szótő és a toldalékok egymástól. A beásban egy lexémának több, egymásból szabályosan képezhető töve van. Ezt egy konkatenatív szabályokat alkalmazó keretrendszerben nem lehet leírni, minden egyes affixhoz meg kellene adni az adott tőváltakozást. Ez azonban egy ilyen nyelvben értelmetlen, tehát meg kell kísérelni először a töveket levezetni, majd ezekből származtatni a különböző alakokat. Így míg az agglutináló magyarban az egyes szóalakok levezetése, elemzése során a hunlexben megadott szabályok nagyrészt megfelelnek az egyes toldalékmorfémák-

nak, egy beás szóalakban nem ilyen egyértelmű ez. Kérdés, milyen szabályok alkalmazódnak pl. a *fracijê* alakra, amely a *fratje* 'fiútestvér' szó többes számú, távolra mutató határozott alakja? A levezetés melyik pontján kell levágni a szóvégi magánhangzót, és hol zajlik a *tj* □ *c* hangváltozás?

Bonyolítja a helyzetet, hogy egy adott toldalékmorféma allomorfjai gyakran a lexémák eltérő töveihez járulnak, ebből következően egy adott tőalternáció más-más paradigmatikus alakokban jelenik meg szavanként. Például az első palatalizációnak nevezett alternáció a főnevek körében a nőneműek többes számú alakjaiban, és az *–ă* végű nőnemű főnevek egyes szám birtokos esetében zajlik le. Ebben az esetben a probléma az, hogy az egyes szám birtokos esetet képző szabály hogyan oldja meg, hogy az egyes lexémák más alternációt mutatnak ebben az alakban? Amennyiben az ilyen problémák az agglutinatív elemzőkhöz hasonlóan kezelődnének, hamar konfúzzá válnának a szabályok és kezelhetetlenné a morfológiai elemzés.

Megoldási javaslatok. Ezeknek a problémáknak az orvoslásához el kell szakadni attól az elvtől, hogy az egyes `hunlex`-szabályok megfelelnek egy-egy morfémának, tehát az egyes szabályokat kisebb operációk elvégzésére érdemes használni. Így az elemző képessé válik az egy szón megjelenő többféle alternáció kezelésére. Így a fent említett *fracijê* alak létrehozásához először egy szabály (`NOUN_PL_CHOP`) levágja a szóvégi magánhangzót, majd a `NOUN_PL_DIST` osztályozószabály megállapítja, hogy ennek a szónak alakja és lexikai tulajdonságai szerint milyen többes számú alakja van. Ezek alapján továbbküldi a `SEC_PAL` nevű szabályba, amely a *tj* ~ *c* alternációt kezeli, amely maga után von bizonyos, a *fratje* szóban nem megjelenő magánhangzó-váltakozást (pl. *lat* 'széles' ~ t.sz. *lec*), amelyet a `RAISE_MID` kezel. Itt ismét egy osztályozószabályba (`RAISE_MID_DIST`) megy a szóalak, mivel ezeknek az alternációknak akár az igei paradigmából is lehetett bemenete, tehát nem egyértelmű, hogy melyik szabályba kell visszatérni.

Itt ismét nehézségekbe ütközne az agglutinatív nyelvek kezelésére használt módszertan használata, hiszen nem lehet eldönteni, hogy az ilyen alternáció utáni osztályozószabály melyik gyűjtőszabályba utalja az aktuális alakot. Ám megoldást adhat a szálak szétbogozására, ha a `hunlex`ben az egyes lexémákhoz adható morfofonológiai jegyekkel élve a generálódó alakhoz belső használatú jelzőket (pl. `_noun_plur`, vagy `_verb_sg_2`) rendel az a szabály, amelynek a `SEC_PAL` a kimenete. Ezek után a `RAISE_MID_DIST` egyszerűen meg tudja vizsgálni, hogy az adott alak milyen paradigmából érkezett, és milyen szabályba kell küldeni.

Ezután tehát a `NOUN_PL_COLL` gyűjtőszabályban újra összefutnak a különböző többes számú alakokat képző „szabályszálak", és innen már a `hunmorph` számára kimenetként képződik a létrejött alak. A gyűjtőszabályból pedig még továbbhalad a forma különböző, a többes számú tőből képzett alakokat létrehozó szabályokba.

Több esetben nem tűnik szükségesnek a különböző részszabályok szétválasztása, így például az egyes számú birtokos eset gyűjtőszabályának (`NOUN_SG_GEN_COLL`) bemenete minden esetben egy szuffixum-hozzáadó szabály (`NOUN_SG_GEN_SUFF`), így a kettő összevonható lenne. Azonban a szétválasztás egyrészt könnyebben áttekinthetővé teszi a rendszert, másrészt pedig a további bővítéseket (például az igei paradigmák implementálását) is megkönnyíti. Hosszabb távon pedig a szabályrendszer sablonszerűsége is egyszerűsítheti a nyelvtan megalkotását, és segíthet más (főleg flektáló) nyelvek e rendszerben történő leírásában is.

Összefoglalás. A flektáló nyelveket a `hunmorph` elemző szótövek és affixumok leválasztására épülő architektúrája nem képes a leghatékonyabban elemezni. A

`hunlex` lexikonkezelő rendszere így szintén az agglutinatív nyelvek sajátosságain alapul, ám ez a rendszer alkalmassá tehető flektáló nyelvek, így a beás elemzésére, ha a szabályoknak külön-külön egyértelmű funkcióik vannak, amelyek közül a legfontosabbak tehát:

- affixumot hozzáadó szabályok
- tőbeli alternációt kezelő szabályok (mindegyik alternációtípust külön szabály kezel)
- osztályozószabályok, amelyek különböző „szálakra" terelik a különböző fonológiai, vagy lexikális jegyekkel rendelkező lexémákat
- gyűjtőszabályok, amelyek összefogják a különböző „szálakat", és legtöbbször kimenetekként működnek a hunmorph által kezelhető szótárba

Bibliográfia

1. Kálmán László, Orsós Anna: Beás nyelvtan: Alsóbb nyelvi szintek. Kézirat. MTA NYTI Elméleti Nyelvészeti Osztálya (2004)
2. Trón Viktor: HunLex – morfológiai szótárkezelő rendszer. In Alexin Zoltán, Csendes Dóra (szerk.): II. Magyar Számítógépes Nyelvészeti Konferencia. SZTE, Informatikai Tanszékcsoport (2004) 177-182
3. Viktor Trón, György Gyepesi, Péter Halácsy, András Kornai, László Németh, Dániel Varga: Hunmorph: Open Source Word Analysis. In: Proceedings of ACL 2005 Workshop on Software At the 43rd Annual Meeting of the ACL (2005)