# Az első nganaszan szóalaktani elemző

Novák Attila

MorphoLogic Kft. 1126 Budapest Orbánhegyi út 5.,
novak@morphologic.hu

**Kivonat** A cikk egy kihalás szélén álló kis északi szamojéd nyelv, a *nganaszan* szóalakjainak elemzésére hivatott programot mutat be, amely egy több uráli nyelvet felölelő projektum keretében készült el. Erre a nyelvre – annak rendkívül bonyolult fonológiája miatt – a projektum keretében egyébként használt *Humor* nyelvleíró formalizmus alkalmazása nagyon nehéznek bizonyult, ezért végül a Xerox cég *xfst* programjának felhasználásával készítettük el az elemzőt.

## 1. Bevezetés

Ebben a cikkben egy nganaszan nyelvű szóalaktani elemzőprogram kifejlesztéséről számolunk be. Erre a nyelvre eddig nem készült morfológiai elemzőprogram. Vállalkozásunk egy olyan projektum[1] részeként valósult meg, melynek célja elemzett korpuszok és egyéb elektronikus nyelvi erőforrások létrehozása néhány kisebb, az uráli nyelvcsaládba tartozó (tehát a magyarral valamilyen fokon rokon) nyelven.

Az Uráli nyelvcsalád északi szamojéd ágához tartozó nganaszan több szempontból is érdekesnek bizonyult a leírandó nyelvek közül. Egyrészt egy gyakorlatilag a kihalás szélén álló nyelvről van szó (beszélőinek száma már nem éri el az ötszázat, ezek túlnyomó része középkorú vagy idős, és az orosz kisebbségi politikának megfelelően nincs nganaszan nyelvű oktatás), ezért fontosnak tartottuk, hogy legalább a nyelv dokumentálásához hathatós segítséget nyújtsunk. Másrészt a nyelv morfológiája és különösen a fonológiája olyan bonyolult, hogy komoly kihívást jelentett a számítógépes formalizálása: elsőként a MorphoLogic Humor formalizmusát próbáltuk alkalmazni, de ebben a formalizmusban nem sikerült teljes leírást készítenünk a nyelvről. Végül a Xerox cég reguláris relációkalkuluson alapuló morfológiai fejlesztőrendszerének felhasználásával készítettük el az elemzőt.

[1] Komplex Uráli nyelvészeti adatbázis, NKFP 5/135/2001. A projektumban a Nyelvtudomány Intézet Finnugor Osztálya, különböző finnugor nyelvészeti tanszékek és a MorphoLogic Kft. vesz részt.

## 2. Első lépések a nganaszan morfológiai leírás létrehozására

Az Uráli Nyelvészeti projektumban eredeti terveink szerint a *MorphoLogic Kft. Humor* morfológiai elemzőprogramjának formalizmusát kezdtük el használni az egyes nyelvek morfológiájának leírására, pontosabban egy olyan nyelvészeti adatbázisleíró keretrendszert, amely az eredeti Humor formalizmusnál magasabb szintű, és ezért sokkal jobban karbantartható nyelvi leírás elkészítését teszi lehetővé, amiből automatikusan létrehozza a Humor morfológiai elemző által használt adatbázist (Novák (2003) [3]). A Humor formalizmusban a szavak morfémák egymáshoz illeszthető allomorfjainak jól formált sorozataiként épülnek fel. A szomszédos morfémák egymáshoz illeszthetőségének leírására a Humor jegyalapú formalizmust használ. A keretrendszert sikeresen alkalmaztuk különböző nyelvek leírására az Uráli nyelvleíró projektum keretében és azon kívül is.

A nganaszan nyelvről igen jó formális igényű leírást készítettek a projektumban részt vevő nyelvészkollégák (Wagner-Nagy (2002) [4]), és gépre vitték, majd az általuk használt latin betűs fonematikus átírásra konvertálták egy nganaszan-orosz szótár anyagát (Kost'erkina és mtsai. (2001) [2]). A szótár kb. 3650 tövet tartalmaz. Mindegyik tételhez kézzel beírták a megfelelő szófajmegjelölést is, amely az eredeti nyomtatott szótári anyagban nem szerepelt. Ilyen feltételek mellett azt reméltük, hogy viszonylag gyorsan és problémamentesen elkészülünk az elemzővel.

Az említett szótár alapján készülő tőtár gépre vitelével párhuzamosan közösen hozzákezdtünk a toldalékok formális leírásához. Első lépésként készült egy olyan toldaléklista, amelyben az egyes toldalékok mögöttes fonológiai alakja és kategóriacímkéje szerepelt, valamint az, hogy a toldalék melyik morfológiai tőalakhoz járul. A nganaszan morfológia leírásánál hasznosnak bizonyult az a modell, amely minden tő esetében három morfológiai tőváltozatot feltételez (ezek közül kettő vagy akár mindhárom is alakilag egybeeshet), és az egyes toldalékokat aszerint kategorizálja, hogy az első, második, vagy a harmadik tőváltozathoz kapcsolódnak-e. Néhány toldalék (pl. a latívusz esetrag) ingadozó viselkedést mutatnak: két különböző tőváltozathoz is kapcsolódhatnak. A mögöttes fonológiai leírás tartalmaz olyan magánhangzó-szimbólumokat, amelyek a nganaszan magánhangzó-harmónia szabályai szerint a tő harmonikus tulajdonságainak megfelelően váltakozó magánhangzókat jelölik. A képzők esetében ez az elsőként elkészült leírás még azt az információt is tartalmazta, hogy a képző milyen kategóriájú tőből milyen kategóriájút képez.

Következő lépésként a toldaléktárat a Humor elemzőhöz készített morfológiai-adatbázis-készítő keretrendszer által megkövetelt formájúra alakítottuk finomítva a leírás részletességét és hozzáadva a formalizmus által megkövetelt adatokat. Néhány egyértelműen szegmentálható (tisztán agglutinatív szerkezetű) komplex toldalékot szegmentáltunk (pl. az alany/tárgyeset + kettes szám + birtokos végződés alakú toldalékkomplexumokat). A toldalék által szelektált tőváltozatot (első, második vagy harmadik tő) bal oldali toldaléktulajdonságként fogalmaztuk meg. Bal oldali (a tőre tett) megszorítások elsősorban az igei végződéseknél szerepeltek: kizárólag perfektív, ill. kizárólag imperfektív igetövekhez járuló vég-

ződések, csak ágenses igék végződései, csak tranzitív igékhez járuló végződések stb.

A toldalékmorfémák sorrendjére vonatkozó morfotaktikai megszorításokat úgy írtuk le, hogy egyrészt a toldalékokat morfotaktikai osztályokba soroltuk: pl. alany/tárgyesetű birtokos végződések (NomPx), a többi esetben használt birtokos végződések (Px), oblikvuszi esetrag (OblCx), névszói predikatív végződés (NVx) stb. és leírtunk egy véges állapotú automatát, amelynek élein az egyes morfotaktikai osztályok címkéi szerepelnek, és azt írja le, hogy az egyes osztályokba tartozó morfémák milyen sorrendben követhetik egymást. A képzők esetében a megengedett sorrendek jelenlegi modellünkben egyszerűen abból következnek, hogy az adott képző milyen kategóriából milyen kategóriába képez.

## 3. A tő- és toldalékalternációk leírása, a tőtár rendhagyó lexikai jegyekkel való gazdagítása

A toldaléktár elkészítése után hozzákezdtünk a tőalternációkat leíró szabályok megfogalmazásához (a Humor keretrendszer ezek alapján a szabályok alapján állítja elő a morfématárakból az allomorfok adatbázisát). A névszói és az igei tövek különböző tőalternációs mintákat követnek. Ezeken belül is a magánhangzó- és a mássalhangzó-végű tövek jelentősen különböző viselkedést mutatnak. Bizonyos tővégi váltakozások csak a megfelelő lexikai jeggyel bíró tövek esetében fordulnak elő, hasonlóan pl. a magyar tőbelseji magánhangzó-rövidüléshez (ilyen váltakozás pl. a tővégi felső nyelvállású magánhangzók *a*-vá válása a névszóknál harmadik tőben), mások minden olyan tőnél jelentkeznek, amely a megfelelő alaki tulajdonságokkal rendelkezik, hasonlóan a magyar tővégi magánhangzónyúláshoz (ilyen pl. a tővégi *ja jai*-ra változása a névszóknál a harmadik tőben). Az előbbi kategóriába tartozó szavak mindegyikét meg kellett jelölni a tőtárban a megfelelő lexikai jeggyel.

## 4. A nganaszan morfofonológia bonyolultsága

A tővégi hangzók váltakozásait a Humor rendszer keretei között viszonylag egyszerűen le lehetett írni. A produktív fonológiai folyamatok közül a viszonylag lokális kontextusra érzékeny szabályokat (pl. degemináció) külön-külön meg tudtuk fogalmazni, azonban amikor a fokváltakozás jelenségét (a szótagkezdetben levő zárhangok szabályszerű váltakozását) is megpróbáltuk bevonni a leírt jelenségek körébe, kudarcot vallottunk. A probléma gyökere tulajdonképpen az a tény, hogy a Humor elemző az elemzendő szót mindig morfémák allomorfjainak sorozataként látja, és elemzés közben minden morfhatáron azt ellenőrzi, hogy a az előző és a következő morf tulajdonságai kölcsönösen kielégítik-e egymásnak a másikkal szemben támasztott megszorításait. Ez a modell eddig minden nyelv esetében jól működött, és általában nem okozott problémát ezeknek a morfok közötti megszorításoknak a megfogalmazása.

A nganaszan fokváltakozás azonban egyáltalán nem függ a szó morfológiai szerkezetétől: kizárólag a szótagszerkezet játszik szerepet benne. A szótaghatárok

és a morfémahatárok viszont általában semmilyen kapcsolatban sincsenek egymással, rövid (1 szegmentumból álló) toldalékok esetében (van ilyen toldalék a nganaszanban) még egymással nem szomszédos morfémák is közös szótagba kerülhetnek. Ez a tény súlyosbítva azzal a körülménnyel, hogy a fokváltakozásban nemcsak az adott, illetve a megelőző szótag zártsága és az előző szótagban levő magánhangzó hosszúsága, hanem még az is szerepet játszik, hogy az adott szótag páros vagy páratlan sorszámú a szón belül, illetve hogy mindez az összes többi váltakozással kombinálódik (magánhangzó-harmónia, degemináció, a legkülönbözőbb tő- és toldalékalternációk és hasonulások; ezek együtt egy egyszótagú toldaléknál könnyen tizennégy különböző allomorfot eredményeznek) oda vezetett, hogy képtelenek voltunk a fokváltakozást (illetve a nganaszan morfofonológia egészét) felszíni allomorf-szomszédsági megszorítások együtteseként leírni, az allomorfokat és a közöttük fennálló megszorításokat előállító szabályegyüttest megfogalmazni.

Ízelítőül álljanak itt egy igei toldalék (az elbeszélő mód alanyi és tárgyas ragozásban használt formájának) allomorfjai. A morféma absztrakt lexikai alakja: $hA_2nhV$, allomorfjai (12 van, és ez még nem is a legbonyolultabb eset): *banghu, bjanghy, bambu, bjamby, bahu, bjahy, hwanghu, hjanghy, hwambu, hjamby, hwahu, hjahy.* Az allomorfok szabályszerűen állnak elő az alábbi fonológiai folyamatok eredményeképp:

Az $A_2$ harmonikus magánhangzó *a*-ként vagy *ja*-ként jelenik meg a tőhöz a kerekítési harmónia szabályai szerint illeszkedve, ráadásul az *a h* után szabályszerűen *wa*-vá diftongizálódik. (Hogy egy tő melyik kerekítési harmóniai osztályba tartozik, az teljesen megjósolhatatlan lexikai jegye, a tövek egy része ingadozó viselkedést mutat.) A $V$ harmonikus magánhangzó *u*-ként, *y*-ként, *ü*-ként vagy *i*-ként jelenik meg a kerekítési és az elölségi harmónia szabályainak megfelelően (ebben a toldalékban csak *u* és *y* lehetséges, mert az előző szótagban mindenképpen hátul képzett magánhangzó (*a/ja/wa*) van). A *h* fonéma erős fokban *h*, gyenge fokban *b*. Az *nh* kapcsolat erős fokban, vagy ha az ún. nunnáció fellép *ngh*, ritmikai gyege fokban *h*, szillabikus gyenge fokban *mb* (a nazális képzési hely szerint illeszkedik, ritmikai gyege fokban viszont eltűnik, hacsak az előző mássalhangzó nem nazális: akkor ugyanis megmarad, ez a nunnáció). Erős fokban áll egy szótagkezdő obstruens, ha nem nazális kóda (mássalhangzó) előzi meg, vagy ha a szón belül páros sorszámú nyílt szótagban van. Egyébként ritmikai gyenge fokban áll, ha hosszú magánhangzó előzi meg, vagy páratlan szótagban áll, és szillabikus gyenge fokban áll, ha páros zárt szótagban áll.

A fokváltakozás annyira produktív folyamata a nganaszan fonológiának, hogy formális leírását semmiképpen sem kerülhetjük el, ha működő elemzőt akarunk készíteni. Úgy tűnt azonban, hogy bár a Humor formalizmusa nem eleve alkalmatlan ennek a nyelvnek a leírására, de a gyakorlatban legalábbis a keretrendszer szabályformalizmusának felhasználásával a leírás elkészítése túl nehéz feladatnak bizonyult.

## 5. Áttérés egy új formalizmusra

2003 júniusában azonban megjelent egy könyv (Beesley-Karttunen (2003) [1]), amelyhez mellékletként adtak egy CD-t, amelyen a *Xerox* cég véges állapotú automata alapú kétszintű morfológiai elemző készítő programcsomagjának nem üzleti célokra szabadon felhasználható verzióját szabadon hozzáférhetővé tették. A programot kutatóintézetek kutatási célra korábban is licenszelhették, de ez olyan hosszadalmas jogi procedúrával járt, és annyi feltételnek kellett eleget tenni, hogy azelőtt szinte elképzelhetetlen volt, hogy hozzájussunk ennek a projektnek a keretében.

Most azonban rendelkezésünkre állt, és úgy döntöttünk, hogy a nganaszanról készített leírásunkat átalakítjuk, illetve újraírjuk a Xerox *lexc* (Lexicon Compiler), illetve *xfst* (Xerox Finite-State Tool) programjai által megkövetelt formában. A *lexc* programmal morfématárakat lehet definiálni folytatási osztályok megadásával, az *xfst* pedig a generatív fonológusok által megszokott kontextusfüggő újraírószabály-formalizmussal leírt szekvenciális fonológiai szabályegyüttesek megadását teszi lehetővé, és kiszámítja az egyes szabályok egymással illetve a lexikonnal való komponálásával előálló teljes morfofonológiai leírást egyetlen kétszintű véges állapotú fordítóautomata formájában, amit elemzésre és generálásra egyaránt lehet használni.

Az *xfst* formalizmusában nem jelentett többé áthághatatlan akadályt a fokváltakozás leírása, mert a program által megvalósított kalkulus lehetővé teszi, hogy az újraíró szabályok környezetmegadásánál az irreleváns szimbólumokat (pl. a morfémahatárokat) figyelmen kívül hagyjuk, ugyanakkor nem jelent problémát a nem szomszédos morfémákra átnyúló környezetek figyelembe vétele sem. Mivel a program az egyes szabályok által létrehozott köztes szinteket a kompozíció révén automatikusan eliminálja, semmilyen hatékonysági problémához nem vezet elemzés és generálás közben a leírás elkészítésekor bevezetett nagy számú közbülső leírási szint sem.

A fokváltakozást az *xfst* formalizmusában úgy írtuk le, hogy definiáltunk egy szabályegyüttest, ami a szótaghatárokon explicit határszimbólumokat illeszt be (a páros és a páratlan szótagok között más-más szimbólumot), az előző szótag zártsága, a benne szereplő magánhangzó hosszúsága, valamint az adott szótag zártsága és páros vagy páratlan volta alapján erős vagy ritmikai/szillabikus gyenge fokúként jelöli meg az egyes szótagokat, majd a szótagkezdetben levő mássalhangzót (illetve a szótaghatáron levő nazális-zárhang kapcsolatokat) pedig a szótag fokának megfelelően megváltoztatja, végül a szótaghatár és fokszimbólumokat kitörli. A szabályrendszer kezeli a nganaszan kivételes szótagolási jelenségeit is: a gégezárhang akkor is zárja a szótagot, ha nem követi másik mássalhangzó (a *V'V* sorozat szótagolása: *V'.V*), a *bt* hangkapcsolat *b*-je viszont nem zárja a szótagot (*V.btV*).

A konkrét szabályegyüttes alább látható:

```
#a dot after every syllable that is followed by an onset
[[C* V C*]/NSeg @-> ... "." || _ [C V]/NSeg ]
#a dot before syllables without an onset
.o.[ V @-> "." ... || V/NSeg _ ]
#resyllabify ' from onset to coda: insert syllable boundary after '
.o.[' -> ... "." || "."/NSeg _ ]
#delete syllable boundary before '
.o."." -> 0 || _ [' "."]/NSeg ]
#resyllabify b from coda to onset if followed by t:
#insert syllable boundary before b
.o.[b -> "." ... || _ ["." t]/NSeg]
#delete syllable boundary after b
.o."." -> 0 || ["." b]/NSeg _ t/NSeg]
#strong grade after non-nasal codas and m codas not followed by b
.o."." -> ... "^S" ||
[C-[n|m|n1|ng|N|M|N1|NG]]/NSeg _ , [m|M]/NSeg _ [Seg-[b|B]]/NSeg]
#rhytmical weak grade after long vowels
.o."." -> ... "^W1" || [V V]/NSeg _ ]
#change every second dot to a comma:
#. = even syllable
#, = odd syllable
.o."." -> "," \/ "." ~$["."|","] _ ]
#rhytmical weak grade in odd syllables not yet marked as strong
.o.[","  -> ... "^W1" || _ NGrd]
#syllabic weak grade in even closed syllables not yet marked as weak
.o."." -> ... "^W2" || _ [NGrd ?* & [C* V C]/\[Seg|"."|","]]]
#strong grade in other even syllables (codaless ones)
.o."." -> ... "^S" || _ NGrd]
#gradation
#rhythmical weak grade of obstruents
"^W1" h -> b, "^W1" t -> q, "^W1" k -> g, "^W1" s -> d1,
"^W1" s1 -> d1 || NNas /NSeg _ ,,
#rhythmical weak grade of nasal+obstruent clusters
Nas -> ^N || _ [["."|","] "^W1" [h|k|t|s|s1]]/NSeg,,
#syllabic weak grade
"^W2" h -> b, "^W2" t -> q, "^W2" k -> g,
"^W2" s -> d1, "^W2" s1 -> d1
#remove syllable boundaries
.o."^W1"|"^W2"|"^S"|"."|","] -> 0
```

A szabályrendszer ezek mellett rengeteg más szabályt tartalmaz, egyrészt produktív automatikus fonológiai szabályokat (pl. a nazálisok hely szerinti hasonulása a következő obstruenshez, degemináció, magánhangzó-harmónia, nunnáció, palatalizáció stb.), másrészt a morfológiailag, ill. lexikálisan megszorított,

szűkebb körben működő tő- és toldalékalternációkat is ilyen szabályokkal lehetett az új rendszerben leírni.

## 6. A morfématárak konverziója

Természetesen az új formalizmus nemcsak az allomorfiák leírásában különbözik gyökeresen a korábban használttól, hanem a morfématárak és a morfotaktika megadásának módjában is. Gondoskodnunk kellett tehát egy olyan konverterről, amely meglévő morfématárainkat a megfelelő formátumra konvertálja. A Humor leírás alapjául szolgáló jegyalapú megszorításokat alkalmazó formalizmus a morfológiai megszorítások (pl. a toldalékok tőszelekciója) leírásában igen jól használhatónak bizonyult, bár a nagyon komplex felszíni fonológia leírására – mint láttuk – nem bizonyult a leghatékonyabb eszköznek. Szerencsére a Xerox elemző formalizmusa is tartalmaz jegy-érték megszorítások leírására alkalmas eszközt (Flag Diacritics), ezért ezeket a megszorításokat hasonló elven meg lehet fogalmazni, mint egy Humor elemző készítésekor.

A lexikon leírására használatos *lexc* program által használt formalizmusban a lexikon morfémák leírását tartalmazó allexikonok sorozatából áll, minden egyes morfémához meg kell adni egy folytatási osztályt, ami vagy egyszerűen annak az allexikonnak a neve, amelynek összes tagja követheti az adott morfémát, vagy a szó végét jelölő szimbólum.

Az alábbi példa a Humor keretrendszer által használt toldaléklistából az elbeszélő mód alanyi és tárgyas ragozásban, illetve a visszaható ragozásban használt formájának leírását mutatja.

```
#mode suffixes
#tag    phon      lp    mcat     comment
(...)
Narr    HA2NHU    S1    VTM      narrative subj/obj
Narr    HA2NHA1   S1    VTMR     narrative refl.
```

Ugyanezek a toldalékok a *lexc* program által használt formára konvertálva így néznek ki:

```
LEXICON infl_V
(...)
@U.S.1@@C.S@h^A2nh^V[Narr]      infl_VTM_r;
@U.S.1@@C.S@h^A2nh^A1[Narr]     infl_VTMR_r;
```

Az `@U.S.1@` szimbólum azt jelöli, hogy az adott toldalék az első morfológiai tőalakhoz járul (az `@U.S.1@` jelentése: 'unifikáld az S tulajdonság aktuális értékét az 1-es értékkel'). A `@C.S@` szimbólum a semleges értékre állítja (törli) az S tulajdonság értékét. A képzők, mint tövek allomorfjainak előállításáról, és a tőtípust azonosító jegyek kitöltéséről a szabályos tőalternációkat leíró szabályok gondoskodnak.

Az elkészült elemzőt latin betűs fonológiai átírással lejegyzett nganaszan nyelvű szövegek morfológiai elemzésére fogjuk felhasználni. Az elemzett szövegeket a projektum honlapján tesszük majd közzé.

## Hivatkozások

1. Kenneth R. Beesley and Lauri Karttunen. *Finite State Morphology.* CSLI Publications, Ventura Hall, 2003.
2. N. T. Kost'erkina, A. Č. Momd'e, and T. Ju. Ždanova. *Slovar' nganasansko-russkij i russko-nganasanskij.* Prosvesčen'ije, Sankt-Pet'erburg, 2001.
3. Novák Attila. Milyen a jó humor? In *Magyar Számítógépes Nyelvészeti Konferencia (MSZNY 2003)*, pp. 138–145, Szegedi Tudományegyetem, 2003.
4. Wagner-Nagy Beáta, (szerk.) *Chrestomathia Nganasanica.* SZTE Finnugor Tanszék – MTA Nyelvtudományi Intézet, Szeged – Budapest, 2002.