Záróbeszámoló
a K60807 sz. A nganaszan nyelv számítógépes morfológiai elemzése című téma kutatási eredményeiről

1. A kutatás főbb adatai:

A kutatás címe: A nganaszan nyelv számítógépes morfológiai elemzése

Tudományterület: Nyelvtudomány (uralisztika)

OTKA nyilvántartási száma: K60807

Témavezető: Dr. Wagner-Nagy Beáta, 2009. november 1-től Dr. Várnai Zsuzsa

A kutatás időtartama: 2006–2009

Az OTKA támogatás összege: 8.830.000 Ft

2. A kutatás célja, a vállalt kutatási program:

Jelen kutatás közvetlen előzményének tekinthető az a projektum (NKFP 2001/5/135: *Komplex uráli nyelvészeti adatbázis*), amelynek egyik célkitűzése az volt, hogy kilenc uráli nyelvhez számítógépes morfológiai elemzőprogram készüljön. A nyelvek kiválasztásánál elsődleges szempont volt a hazai kutatói kompetencia, melyre a nganaszan nyelv esetében kedvező lehetőségek nyíltak. Azonban a nganaszan morfológia számítógépes formalizálása nagyságrendekkel nehezebb feladatot jelentett, mint a projektumban szereplő finnugor nyelveké. Ugyan ehhez a nyelvhez is elkészült az elemzőprogram, mely azonban korántsem volt tökéletesnek tekinthető. Az alapvető különbséget és nehézséget a többi nyelvhez képest az jelentette a számítógépes modell megalkotása szempontjából, hogy a nganaszan nyelvnek rendkívül komplex a morfofonológiája, valamint rendkívül produktívak a felszíni fonológiai–fonetikai folyamatai. Éppen ezért a többi uráli nyelv leírására alkalmazott allomorf-szomszédossági megszorításokon alapuló számítógépes modellt (a MorphoLogic Humor programját) a nganaszanra lehetetlen volt alkalmazni. Ez azt jelentette, hogy új modellt kellett felállítani. Az új modell kidolgozása már a fent említett projektumban megkezdődött, de befejezésére ott már nem kerülhetett sor.

Jelen OTKA pályázat a modell, azaz az elemzőprogram fejlesztését, a hiányos és ellentmondásos adatok, paradigmák tisztázását, újabb szövegek, hangzóanyagok terepen történő gyűjtését, valamint a nganaszan nyelv teljes és részletes leírását tűzte ki céljául.

Az elemző tökéletesítése mellett nagy hangsúlyt szerettünk volna fektetni egyrészt az újabb szövegek gyűjtésére, másrészt a részletes nyelvi leírásra. Ezt a kettős célkitűzést több dolog is indokolta. Számos meselejegyzés jelent meg már nyomtatásban, amiket természetesen mi is felhasználtunk morfológiai elemzőnk tesztelésekor, de nagyon kevés az élőnyelvi (tehát nem a folklór műfajú) szöveg. Ezért a terepmunka legfontosabb célja az volt, hogy elbeszélő szövegeket (pl. életútról, családról) és párbeszédes szövegeket gyűjtsünk, amiket egyrészt felhasználhatunk az készülő leíró nyelvtanhoz is, másrészt tesztelhetjük velük a elemző programot. Célul tűzte ki a kutatás azt is, hogy szövegeinket elérhetővé tesszük, hiszen a kisebb uráli nyelvek kutatásában az az egyik legnagyobb probléma, hogy kevés a hozzáférhető, elektronikus formátumú szöveg. Ennek következtében szinte minden ilyen irányú kutatás korpuszépítéssel kezdődik. Ezen okok is, valamint az a szociolingvisztikai tény, hogy a nyelvet anyanyelvi szinten ma már csak a legidősebb generáció (40-50 fölötti nemzedék) beszéli, indokolta a tervezett két terepmunkát.

A kutatás másik fontos célkitűzése egy angol nyelvű leíró nyelvtan létrehozása volt. Ennek előzményét a 2002-ben megjelent Chrestomathia Nganasanica alkotta. Terveink szerint ennek eredményeit bővítjük ki és a nyelvtant úgy íruk meg, hogy olyan szakemberek is használni tudják, akik nem érdeklődnek az uralisztika iránt, de szeretnék megtudni, hogy a nganaszanban egy-egy nyelvi jelenség (pl. birtokviszony, másodlagos predikátumok kifejezése, magánhangzó-harmónia stb.) hogyan valósul meg. Ennek megfelelően a kutatástól három fontos eredményt vártunk:

1. Egy morfológiai elemző programot, amely az interneten on-line elérhető, és melynek segítségével nganaszan szövegek elemezhetők.
2. Az online elemző által megelemzett, általunk egyértelműsített, annotált szövegek.
3. Az elemzőprogram fejlesztésével, a programírás során felmerült (morfo)fonológiai problémák megoldása, valamint a terepen gyűjtött anyagok és tapasztalatok révén a nganaszan nyelv korábbiaknál sokkal alaposabb és korszerűbb leírását adjuk egy modern szemléletű monografikus feldolgozás keretében.
4. A gyűjtött nyelvi anyag révén pontosabb képet kapunk a nganaszan beszélt nyelvről, a nyelvtani jelenségeken túl akár a mindennapok szókincséről, amely további lexikológiai, etimológiai kutatásokat tesz majd lehetővé.

3. A kutatás időrendje

A kutatás kezdő évének (2006) legfontosabb feladatai a következők voltak:

- Első lépéseként létre kellett hozni egy szövegtárat. Tekintettel arra, hogy az elemzőprogram teszteléséhez nagy mennyiségű szövegre volt szükség, erre a célra a Labanauskas[1] által kiadott szöveggyűjteményt választottuk ki. A szövegek fonológiai átírása, az elemző által használt formátumra alakítása 2006 során megtörtént.
- Ugyanebben az évben alkottuk meg a szóalak-generátort, ami abban segítette a munkánkat, hogy a segítségével pl. példaparadigmákat alkothattunk, amelyeknek aztán a helyességét a terepmunka során ellenőriztük. A szóalak-generátor így hozzájárult ahhoz, hogy az eddig ismert morfofonológiai szabályokat vagy egyes morfémák mélyszerkezeti alakjait számos esetben finomítottuk.

A kutatás második évének (2007) legfontosabb feladatai a következők voltak:

- A terepmunkára való felkészülés volt egyrészt kérdőívek kidolgozásával, másrészt az utazás megszervezésével. Már az év elején kiderült, hogy a támogatásként megkapott pénzösszeg az eredeti tervben szereplő kétszeri utazásra aligha lesz elegendő.. Adminisztratív okok miatt (vö. 6. pont) ebben az évben nem került sor a gyűjtőútra.

A 2008-as évben három fontos feladat elvégzésére került sor:

- egyrészt a terepmunkára, annak eredményeinek a feldolgozására,
- másrészt a szóalaktani elemző hibáinak feltárására és javítására,
- valamint kimenetének alkalmassá tételére webes megjelenítésre. Hozzákezdtünk a Labanauskas-féle szövegekben szereplő szóalakok számítógépes elemzése (illetve nem-elemzése) által feltárt átírási, ill. az egyéb lejegyzési hibák kijavításához. Azokban az alakokban, amelyekben az adott szó helyesnek bizonyult, de az elemző mégsem működött megfelelően, szükség szerint folytattuk a tő-, ill. a toldaléktár, valamint az elemzőben implementált nyelvtan megfelelő javítását. Készítettünk egy átalakítót is az elemzések megjelenítéséhez, amely kiemeli a szövegben azokat a szavakat, amelyeket nem tudott megelemezni. Így közvetlenül, könnyen észlelhetők a hibás szóalakok.

Az utolsó kutatási évben, azaz 2009-ben az egyes részfeladatok sikeres befejezésére koncentráltunk, amelyek a következők voltak:

---

[1] Лабанаускас, К. (szerk.) (2001): *Нганасанская фольклорная хрестоматия = Фольклор народов Таймыра* 6., Дудинка, Таймырский окружный центр народного творчества

- Feldolgoztuk a kérdőíveket és a nganaszan leíró nyelvtan egyes fejezeteibe beledolgoztuk a terepen gyűjtött fonológiai, morfológiai és szintaktikai információkat.
- Elkezdtük az új szövegek feldolgozását, glosszálását és fordítását.
- Tovább folytattuk az elemző által elemzett szövegekben található hibák javítását, aminek a gyorsítására kidolgoztunk egy újabb átalakítót, amely a pirosra színezett hibás alakok közül kiszűri és narancssárgára színezi azokat, ahol a hiba forrása fonológiai természetű. Ezeket a szavakat az elemző – szabályos elemzés hiányában – úgy próbálja megelemezni, hogy a nyelvtana szerint lehetséges tő és toldalék-allomorfokat a nyelvtan által leírt sorrendben a morfofonológiai megszorítások figyelembe vétele nélkül is megpróbálja összerakni. Az így kapott elemzésekhez azt a szóalakot is generálja, amely az adott elemzéshez tartozó, a számítógépes morfológiában implementált nyelvtan szerint helyes alak lenne.
- Elkészítettük a szöveggyűjtemény szövegeinek magyar fordítását.

4. A pályázat eredményeinek részletezése

A pályázatnak több fontos eredménye is van. Először is elkezdtünk feltölteni egy elektronikus, annotált korpuszt, amely azt is lehetővé teszi, hogy újabb szövegekkel bővüljön. A korábbi elemzőprogramot tökéletesítettük, valamint ennek kapcsán megalkottunk egy szóalak-generátort. Végül pedig elkészítettük az angol nyelvű nganaszan leíró nyelvtani monográfiát. E munka részeredményeit számos konferencián és tanulmányban bemutattuk.

4.1. A korpusz

A korpuszként feldolgozott Labanauskas-féle szöveggyűjtemény, 58 szöveget tartalmaz, amelynek mindegyike folklórszöveg. Sajnálatos módon a gyűjtés helyéről, a beszélő életkoráról, a lejegyzés körülményeiről és egyéb más adatról a kötet szerkesztője nem ad tájékoztatást, mely igen nagy hátránya. Használata mellett mégis az szólt, hogy ez jelentős mennyiségű szöveget (több mint 17000 szövegszót) tartalmaz. Bekerültek a korpuszba más szövegek is, így pl. saját gyűjtésű anyagok, ezek száma azonban jóval kevesebb.

A lejegyzett szövegek általában inkonzisztensek a hangok jelölése szempontjából. Ez azt jelenti, hogy akár az azonos gyűjtő által lejegyzett szövegekben sem egységes az írásmód. Mivel ez volt a helyzet az általunk használt szövegekben, a korpusz felállítása nem pusztán gépelési és transzliterálási feladatot jelentett. A szöveggyűjtemény begépelését fonologizálva az első két évben végeztük el. A szövegeket olyan formára alakítottuk, mint amilyet az elemző használ: az eddig használt speciális betűtípusokat alkalmazó tipográfiai megoldások

helyett olyan kódolást alkalmaztunk a programban, amely a billentyűzeten könnyen begépelhető, és csak egyetlen betűkészletből tartalmaz karaktereket. Az általunk használt jeleket, a fonologikus átírásban használt betűkészletet, valamint a számkódos átírást az alábbi táblázat szemlélteti.

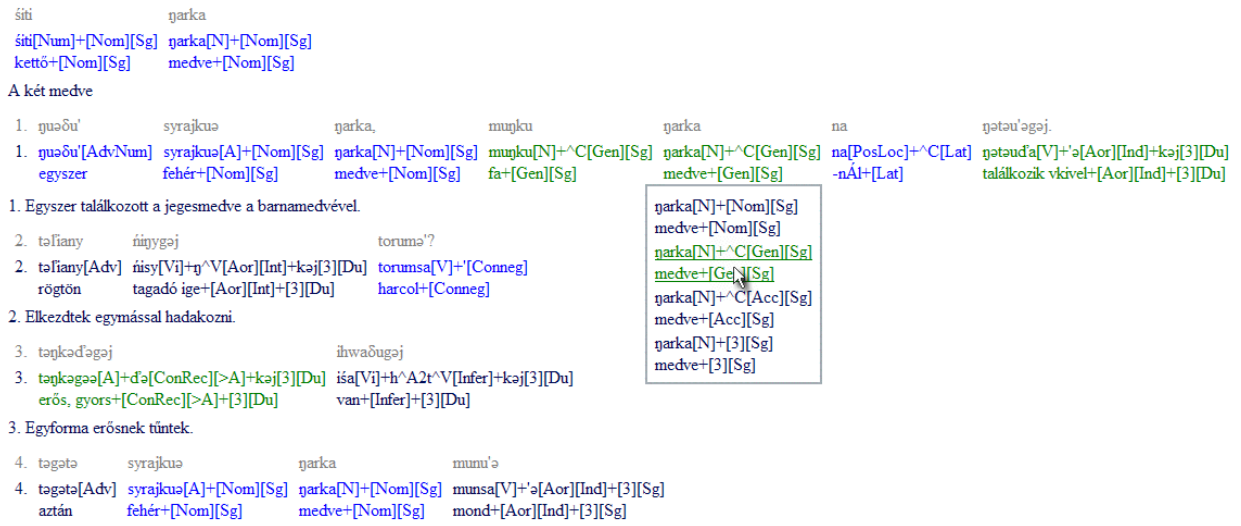
cirill/Labanauskas	д		н		ч		е		я		ya		ю
fonologikus átírás	ď	d	ń	n	ť	t	ə	e	^{i}a	ia	^{u}a	ua	ü
7 bites számkódos átírás	d1	d	n1	n	t1	t	e/e1	e3	ja	ia	wa	ua	u2

Mivel a korábban kidolgozott, pusztán ékezetmentes latin betűket és számjegyeket tartalmazó átírásunk nehezen volt olvasható, és az elemzések megjelenítése sem segítette eléggé az elemzett szövegek ellenőrzését és javítását, újabb átírási rendszert terveztünk, amely Unicode karakterkészleten alapul, és a korábbi hét bites karakterreprezentációnál lényegesen jobban olvasható. Ennek ellenére kénytelenek voltunk bizonyos pontokon a fonológiai átírástól különböző karaktereket alkalmazni, mert az ott használt egyes karakterek egy része (pl. a felső indexbe helyezett *i* és *u,* illetve a centrális *i* jelölésére alkalmazott karakter) önálló karakterként nem szerepelt a rendelkezésünkre álló betűkészletekben. Ezek helyett ezért továbbra is a j, w, ill. y karaktereket használjuk. Más részüknél a kisbetű–nagybetű párok vetnek fel kisebb jelentőségű problémákat (pl. a δ, ď, ť esetében: Δ, Ď, Ť).

Készítettünk egy olyan billentyűzetmeghajtót a magyar kiosztás kiegészítésével, amelynek segítségével a megfelelő Unicode karakterek beírhatók. Az elemzőhöz készült webes felületen virtuális billentyűzetet helyeztünk el, melynek segítségével a felhasználók begépelhetik a nyelv fonémáinak jelölésére használt karaktereket anélkül, hogy a gépükre külön billentyűzetmeghajtó programot kellene telepíteniük. A korábban különböző karakterkészletek egyvelegével Word fájlokban leírt szövegeket egységesítettük, és a fenti átírásra átalakítottuk. Az Unicode-os átírás lehetővé teszi az elemzések tetszetős megjelenítését webes felületen is, ahol különböző színekkel való megjelenítéssel tudjuk felhívni a figyelmet azokra a szóalakokra, amelyeknek elemzése a morfológiai elemző szerint nem egyértelmű, vagy amelyeket az elemző nem tudott elemezni. Ugyancsak sikerült ergonomikusabbá tenni a többértelmű szavak elemzéseinek megjelenítését is: az elemzéseket interlineárisan jelenítjük meg, minden szóalakhoz csak egy elemzést adunk, a többi lehetséges elemzés egy automatikusan megjelenő buborékablakban jelenik meg, ha az egér mutatóját az

adott szó fölé helyezzük. Így a szöveg sokkal áttekinthetőbb, továbbra is a megszokott módon vízszintesen olvasható, a szavak pedig egyenletes térközzel elválasztva követik egymást.

Az alábbi példa a Labanauskas-féle szöveggyűjteményből a *Két medve* című szöveg egy részletének az elemzését mutatja, ahogy az az elemző webes felületén megjelenik. A webes felület lehetővé teszi az adott kontextusban érvényes elemzés kiválasztását azoknak a szavaknak az esetében, ahol egynél több elemzés van (erre a tényre az elemzés kék színe hívja fel a figyelmet): az egeret az elemzés fölé húzva egy listában automatikusan megjelenik az összes az adott szóhoz tartozó elemzés, amelyekből az egérrel lehet választani. A kézzel választott elemzések zöld színnel jelennek meg a felületen.



4.2. Az elemző

A morfológiai elemzőprogram feladata, hogy az adott nyelv szóalakjainak lehetséges morfémákra bontását elvégezze, és azonosítsa az egyes lehetséges elemzésekben a szótövet, annak szófaját és a szóalak toldalékmorfémák által hordozott egyéb morfoszintaktikai jegyeit. A feladat az ellenkező irányban is értelmezhető: ilyenkor a program a tő és a morfoszintaktikai jegyhalmaz együttese által meghatározott szóalak lehetséges felszíni megvalósulásait állítja elő. Az utóbbi működést megvalósító programot szóalak-generátornak hívják. Minél pontosabb a számítógépes morfológiában megfogalmazott nyelvtan, annál inkább igaz, hogy az elemző érvényes elemzéseket állít elő az egyes szóalakokhoz, és a generátor csak érvényes alakokat generál. Egy kihalófélben lévő, csak kevés és nehezen hozzáférhető helyen élő beszélő által beszélt nyelv pontos modelljének kialakítása jóval nehezebb feladat, mint a jól dokumentált nyelvek leírása, mert kevés és nagyon zajos (hibás) adat áll rendelkezésre. Sok fejtörést okozott ezért a morfológiai modellünk számára elemezhetetlennek bizonyult szavak kezelése, hiszen minden alkalommal ki kellett

nyomoznunk, hogy a hogyan is kéne értelmezni az elemzetlenül maradt szót, a hiba oka elírás, lexikai hiány (nem szerepel a morfématárban valamely a szóban szereplő morféma), kivételesség, esetleg a modell hiányossága, és ha az utóbbi, akkor hogyan is küszöbölhetjük azt ki az adatokkal minél inkább összeegyeztethető módon.

A nganaszan morfológiai elemzőt és a szóalak-generátort a Xerox cég *lexc* (Lexicon Compiler), illetve *xfst* (Xerox Finite-State Tool) programjai felhasználásával készítettük el. 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. 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 az újraíró szabályok által bevezetett nagy számú közbülső leírási szint.

A morfológiai megszorítások (pl. a toldalékok tőszelekciója) leírásához a Xerox elemző formalizmusa tartalmaz jegy-érték megszorítások leírására alkalmas eszközt (Flag Diacritics), ezeket a megszorításokat hasonló tehát ennek a formalizmusnak a felhasználásával írtuk le.

A *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 minta a denominális képzőket tartalmazó allexikont egy részletét mutatja be. Az `@U.S.1@`, `@U.S.2@`, `@U.S.1@` szimbólumok azt jelölik, hogy az adott toldalék az első, a második, vagy a harmadik 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 `@D.Px.1@` megtiltja az adott morfémának olyan tőhöz kapcsolását, amelynek a Px jegye 1-es értékű (a kifejezetten birtokos végződéseket váró tövekhez képzők nem kapcsolódhatnak), a `@C.Px@` szimbólum a semleges értékre állítja (törli) a Px tulajdonság értékét. (A képzők, mint tövek allomorfjainak előállításáról, és a tőtípust azonosító jegyek kitöltéséről a tőalternációkat leíró szabályok gondoskodnak.)

```
LEXICON deriv_N
@D.Px.1@@C.Px@@U.S.1@@C.S@b^Ve[Poss2][%>A]          deriv_A_r;
@D.Px.1@@C.Px@@U.S.1@@C.S@d1e[ConRec][%>N]          deriv_N_r;
@D.Px.1@@C.Px@@U.S.1@@C.S@d1e[RelAdj1][%>A]         deriv_A_r;
@D.Px.1@@C.Px@@U.S.1@@C.S@d1ee[Perf][%>N]           deriv_N_r;
@D.Px.1@@C.Px@@U.S.1@@C.S@d1u2m[Sel][%>N]           deriv_N_r;
@D.Px.1@@C.Px@@U.S.1@@C.S@e[RelAdj3][%>A]           deriv_A_r;
@D.Px.1@@C.Px@@U.S.1@@C.S@ed1e[RelAdj2][%>A]        deriv_A_r;
```

Az alábbi részlet pedig a *lexc* formátumú tőtár egy részletét mutatja:

```
imu0hwansa:imu^Chwan^U0 [Vi];
imid1i0:imisi^I [N];
imid1i0:imi00^I [N]_Voc;
imiqjaj0@P.S.3@0:imitiai#P.S.3#^U       [N];
imiqjaj@P.S.1@0:imitjaj#P.S.1#^U        [N];
imiqjaj@P.S.2@0:imitja0#P.S.2#^U        [N];
imed1ee0:imesee^U       [A][Pt];
imed1ee0:imesee^I       [A][Pt];
ili     [Conj];
ini'ia0:ini'ia^I        [N];
ini'ia'ku0:ini'ia'ku^I  [N];
ini'ja@N.S.3@0:ini'ja#N.S.3#^I  [N];
ini'jaims1i:ini'jaim^I0 [V];
ini'ja0@P.S.3@0:ini'jai#P.S.3#^I        [N];
```

Mivel ez a lexikonformátum igen nehezen olvasható és karbantartható, a számítógépes morfológiánk forrását valójában nem ebben a formában tároljuk.

A toldaléktár forrásából az alábbi példában adunk ízelítőt:

```
#derivational suffixes
#tag        phon        lp     lr  mcat          comment
Soc         ^Csebte     S2         N>Adv         sociative
Sel         d1u2m       S1         N>N           selective (egyik fiú)
Aug         'e          S2         N>N           augmentative (nagy egér)
AugMax      RbA1'e      S3         N>N           augmentative
Aug         'e          S2         A>A           augmentative (nagyon hideg)
AugMax      RbA1'e      S3         A>A           augmentative
Sim         ReKU        S1         N>A           similative (-szerű)
Sim         ReKU        S1         A>A           similative (-szerű)
Perf        d1ee        S1         N>N           perfective (ex-)
```

Tőtárunk pedig az alábbi formátumú:

```
imid1iI[N:nagyanya];
imid1i:imiI[N:nagyanya];cont:Voc;
imiqjajU[N:pók]; S2:imiqja; S3:imiqiai;
imuhwansa[Vi:van, történik];
in1s1u2qeU[N:1) utas, szánon utazó, 2) elsõ szán a fogatban];
ini'jaI[N:öregasszony,| feleség]; S3:ini'jai;
ini'ja'kuI[N:feleség];
```

A lexc program által elvárt fenti nehezen olvasható formátumú lexikonokat ezekből a sokkal könnyebben karbantartható lexikonokból programmal állítjuk elő. Az alábbiakban részletesebben is bemutatjuk a tő- és toldaléktárakat.

4.2.1. A tő- és a toldaléktár

Az elemző tőtárának alapjául a nganaszan–orosz szótárt (Kosťerkina és mtsai. (2001)[2] vettük alapul. A szótár címszavait korábban – még az NKFP projekt idején –, az elemző készítésekor kézzel konvertáltuk és gépeltük be. A kézi konverzió nem mindig volt hibátlan, amire a gépi ellenőrzés során fény derült. Törekedtünk arra is, hogy a szótár címszavai mellett felsorolt ragozott alakok rendhagyó tulajdonságaiból kikövetkeztethető, de a felszínen meg nem jelenő mögöttes fonémákat a tőtárban a megfelelő helyen felvegyük. A tőtárat és a nyelvtant ennek megfelelően a cirill betűs szótári anyaggal való összevetés alapján javítottuk. Találtunk több olyan lexikai tételt, amelyekben azt kell, hogy feltételezzük, hogy a tő ún. „lebegő palatalitáselemet" tartalmaz. Az eddigi fonológia leírásokban ilyen elemről nem volt szó. Ennek feltételezését azonban az a tény indokolja, hogy egyes szavak bizonyos ragozott alakjaikban a toldalék rendhagyó módon palatalizálódik. Ebbe a csoportba tartozik pl. az *iśa* 'van' ige (kopula), amelynek 3Sg alakja *ičču* egy lebegő palatális elem nélkül nem magyarázható: a toldalék *t* hangzóját a tő láthatatlan palatalitáseleme palatalizálja *č*-vé. Ugyancsak az *iśa* ige *ičču* alakja példa egy másik rendhagyó toldalékolási tulajdonságra: bár az ige *i* tőmagánhangzója után palatális toldalékharmóniát várnánk (**ičči*), a kopula úgy viselkedik, mintha a benne levő *i* nem lenne palatális. Ezt egy a palatális magánhangzó-harmónia terjedését blokkoló elem felvételével modelláljuk a számítógépes morfológiánkban.
A kopula tőtári reprezentációja tehát az alábbi (a `^J` a toldalék-mássalhangzót palatalizáló elem, a `^B` a palatális magánhangzó-harmónia terjedését blokkoló elem):

```
i^J^Bs1a[Vi:van];!S1:ngue;
```

A tőtárat folyamatosan bővíteni kellett azokkal az egységekkel, amelyek a szövegekben előfordultak, de a tőtárból hiányoztak. Ez nagyon aprólékos és lassan végezhető munka, mivel a hiányra először onnan tudunk következtetni, hogy a szövegben elemzetlenül marad a szóalak. Egy-egy szónál a tő megállapítása számos esetben hosszabb kutatómunkát igényelt. A tőtár most mintegy 4200 tövet tartalmaz
A tőtárban felvett szavak a jelentésén és szófaján kívül még számos más információt is meg kellett adnunk, amelyek a következők:

---

[2] Kosťerkina, N. T. – A. Č. Momďe – T. Ju. Ždanova (2001). *Slovar' nganasansko–russkij i russko-nganasanskij*, Sankt-Peťerburg, Prosvesčeńije

1. Névszóknál szükséges a harmónia-osztály megadása, mivel ez a nganaszan szavaknál egy immanens, a felszínen nem látszó tulajdonság. Ezt U/I-vel jelöltük. Abban az esetben, ha a szónak valamilyen szabálytalan tőalternánsa is létezett, akkor ezt is felvettük a tőtárba. Szintén jelöltük a lebegő mássalhangzót (^C). Erre azért van szükség, mert ennek hiányakor olyan alakokban is lefut a fokváltakozás, amelyekben ennek nem szabadna megtörténni.

```
belyI[N:dal, ének];stemalt S3:bely;
bengkeU[N:1)földbe vájt ház; 2)osztják (házban lakó ember)]; stemalt S3:bengkü;
ben1d1i^Ckaa[A|Pro:mindegyik];
```

2. Az igéknél nem volt szükség a harmónia-osztály megadására, mivel ezt az infinitívuszrag mutatja. Azonban fontos tulajdonsága a nganaszan igének az aspektus, ami szintén immanens tulajdonsága a szónak, tehát ennek a jelölésére szükség van. A perfektív igék V az imperfektív igék Vi jelölést kaptak.

```
belwa't1usa[V:megmérgesedik];stemalt S3:belwa't1a;
 d1@bytesy[Vi:1)topog, dobog, 2)rúg];
```

3. Szükséges volt ezen kívül a rendhagyó tövek lexikai megadására. A magánhangzóra és a *j*-re végződő tövek változatos tőváltakozási osztályainak leírásából a „szabályosnak” (vagy legalábbis viszonylag gyakoribbnak vagy egyértelműnek) tekinthető osztályok viselkedését szabályokkal kezeltük. Az egyedi kivételeket egyenként kellett a lexikonban megjelölni. Olyan formalizmust használtunk, amelyben a kivételes második vagy harmadik tövű szavakat viszonylag tömör formában le lehetett írni. Ezekben az esetekben az alaptő (S1) mellett a többi tő alakját is meg kell adni. Ha több tőalak is rendhagyó, de egybeesnek, azt is egyszerre meg lehet adni. Az alábbi példák közül az első kettő esetben az S3 tő rendhagyó. Az S2 egybeesik az S1 tővel. (A második esetben az a rendhagyóság, hogy az S3 tő is egybeesik vele.) A következő két példában az S2 és az S3 tő is rendhagyó, de a másodikban egybeesnek. Ezt rövidítve is meg lehet adni (!S1=az S1-től különböző alakok). Az utolsó példában az S3 tő rendhagyó, de két alternatív alakja lehetséges (és az S2 egybeesik az S1 tővel).

```
beuremuU[N:átkelőhely];S3:beurema;
bieI[N:szél];S3:bie;
tugy'I[N:anyag, rongy];S2:tukyd1e;S3:tukyd1i;
tujU[N:tűz];!S1:tuu;
tyraaU[A:sekély];S3:tyraa/tyrau;
```

A tőtártól különböző formátumú a toldaléktár. Az alapvetően táblázatos szerkezetű fájlban külön oszlopokban szerepel a toldalékmorfémákhoz tartozó kategóriacímke, a

toldalékmorféma mögöttes alakja, különböző morfoszintaktikai jegyek (pl. hogy a toldalék melyik morfológiai tőalakhoz járul), a morfémák egymásutániságának leírásához használt morfológiai kategóriacímkék, és magyarázó megjegyzések. Korábban mutattunk példát egyes képzők lexikai ábrázolására, alább néhány névszói inflexiós morféma toldaléktárbeli ábrázolása következik:

```
#tag            phon        lp          mcat        comment
                                                    nominative

NomSgPx1Sg      me          S1 Px       NomPx
NomSgPx2Sg      Re          S1 Px       NomPx
NomSgPx3Sg      TU          S1 Px       NomPx
NomSgPx1Du      mi          S1 Px       NomPx
NomSgPx2Du      Ri          S1 Px       NomPx
NomSgPx3Du      Ti          S1 Px       NomPx
NomSgPx1Pl      mU'         S1 Px       NomPx
NomSgPx2Pl      RU'         S1 Px       NomPx
NomSgPx3Pl      TUng        S1 Px       NomPx
```

4.2.2. A szabályok

Az xfst programban a fonológiai és morfofonológiai szabályokat a klasszikus generatív fonológia újraírószabályaihoz hasonló formátumban lehet leírni. 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. A szabályformalizmus illusztrálására a szótagolást és az egyes szótagok erős vagy gyenge fokának kiszámítását végző szabályegyüttest (ritmikai és szillabikus fokváltakozás) mutatjuk be.

A szabályegyüttes 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. Ezen kívül kezel számos kivételes a fokváltakozással kapcsolatos jelenséget (az intervokális gégezárhang kódába szótagolódását, kivételesen viselkedő tővégi nazálisokat, egyes *-nt* kezdetű toldalékok opcionális kivételes viselkedését stb.)

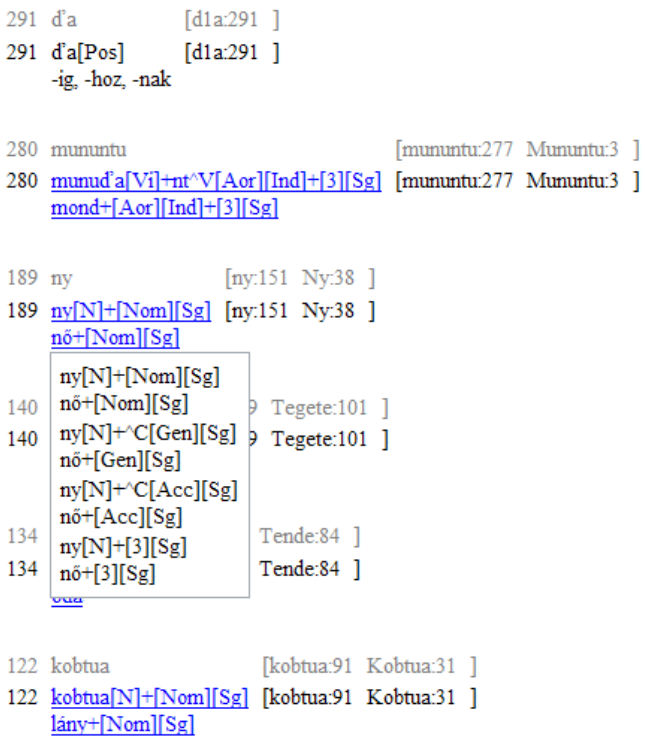
```
#syllabification: syllable boundaries are marked by a dot or a comma in an alternating
fashion:
#.S: even syllable
#,S: odd syllable (except the first one, which is unmarked)

#/NSeg makes sure that non-segmental material is ignored
```

```
define Syllab [

#a dot after every syllable that is followed by a syllable which has an onset
[[Co* Vo Co*]/[NSeg|TAG] @-> ... "." || _ [Co Vo]/NSeg ]

#a dot before syllables without an onset
.o.
[ Vo @-> "." ... || Vo/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]

#delete syllable boundary after -
.o.
["." -> 0 || "-"/NSeg _ ]

# suffix initial nt in the imperfective Aor Ind affix and Gen/Obl Px suffixes is optionally t
in 4th and further syllables
# is deletion of n the right solution?
# or are they exceptionally in rhythmical weak grade?
# does the nasal close the previous syllable?
# probably yes
# so we'll do n (->) "^N" instead of n (->) 0
.o.
[n (->) "^N" || "." \["."]* "." \["."]* Bdry NSeg* _ t Vo Seg* ["[Aor]"|"[Px]"] ]

#strong grade after non-nasal codas and m codas not followed by b
.o.
["." -> ... "^S" || [[Co-[n|m|n1|ng|N|M|N1|NG|Ng|n=|"^N"]] (Nas)]/NSeg _ , [m|M]/NSeg _ [Seg-
[b|B]]/NSeg]

#rhytmical weak grade after long vowels
.o.
["." -> ... "^W1" || [Vo Vo (Nas)]/NSeg _ ]

#n= + Co (only in Intnt hwantun= and d1indin=): always behaves according to syllabic rules
#syllabic weak grade if the syllable is closed
.o.
["." -> ... "^W2" || "n="/NSeg _ [NGrd ?* & [Co* Vo [Co-"^X"]]/\[Seg|"."|","]]]
#strong grade if the syllable is open
.o.
["." -> ... "^S" || "n="/NSeg _ NGrd]

#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 syllables with a coda not yet marked as weak
.o.
["." -> ... "^W2" || _ [NGrd ?* & [Co* Vo [Co-"^X"]]/\[Seg|"."|","]]]

#strong grade in other even syllables (codaless ones)
.o.
["." -> ... "^S" || _ NGrd]

];
```

4.2.3. Az elemző tesztelése

A kézzel írott nyelvtanokban általában számtalan részlet homályban marad (pl. az újraírószabályok pontos megfogalmazása és sorrendezése). Ezeket a számítógépes nyelvtanban elkerülhetetlenül explicitté kell tenni, hiszen e nélkül az elemző nemhogy működni nem tudna, de létre se jönne. Ennek a munkának az egyik fontos eredménye az, hogy jelenleg már sokkal pontosabb modellekkel rendelkezünk, mint amikor elkezdtük a munkát. Ráadásul a számítógépes implementáció lehetővé teszi azt is, hogy nagyon részletesen teszteljük a nyelvtan adekvátságát, (hogy valóban helyesen és pontosan modellezi-e a nyelvi adatokat). E munka során fontos adatokkal szolgáltak a korpuszban elemzetlenül maradt szavak, hiszen ezek felhívják a figyelmünket az elemzőben implementált nyelvtan, illetve szótár, vagy adott esetben a korpusz hibáira.

A korpuszból szóalak-gyakorisági statisztikát készítettünk egy erre a célra általunk létrehozott program segítségével. A szóalak-gyakorisági statisztika a leggyakoribb alakoktól a ritkábbak felé rendezve tartalmazza a szövegekben szereplő szóalakokat. A szóalak-gyakorisági listán az elemzőt lefuttatva megkaptuk az egyes szóalakoknak az elemzőben implementált nyelvi modell szerint lehetséges elemzéseit. Kézenfekvő volt először azokat a szavakat végignézni, amelyeket az elemző nem ismert fel. A hibák egy részének az volt a forrása, hogy a kérdéses szó töve nem szerepelt a tőtárunk forrásául szolgáló szótárban. Ezeket a szavakat természetesen felvettük a tőtárba. Előfordult olyan eset is, hogy hibásan szerepelt a szó a tőtárban, sokszor azonban magukban a szövegben volt helytelen az írásmód, vagy esetleg a forma. Ezeket más források adataival történő egybevetés során lehetett tisztázni. Az ellenőrzés során arra is fény derült, hogy a toldaléktárban is hibásan szerepelt néhány morféma, és a morfotaktikai szabályok megfogalmazásában is találtunk finomítanivalót. Az alábbi példa a Labanauskas szöveggyűjtemény leggyakoribb szavait mutatja, a *ny* ‘nő’ szó elemzéseit kiemeltük.



A szövegek elemzése során az elemző nyelvtanával kapcsolatban is számos problémát tártunk fel, melyek közül sokat sikerült kijavítani: sok olyan szót, morfémát találtunk, amelyek viselkedése nem szabályos, ill. amelyeknél bizonyos mértékű szuppletív viselkedéssel kell számolni pl.:

- a szó paradigmájának egyes tagjaiban rendszeresen szabálytalan a magánhangzó-harmónia, más alakokban szabályos
- bizonyos alakok úgy viselkednek, mintha a tő lebegő mássalhangzót tartalmazna, mások nem
- az adott szó két különböző toldalékolt alakja teljesen összeegyeztethetetlen a fokváltakozás szempontjából (erre példa többek között a korábban említett kopula renarratív alakja szemben a többi toldalékolt alakkal) stb.

Találtunk olyan morfológiai képződményeket, amiknek a töve nem vezethető le az elemzőben implementált nyelvtan segítségével, felépítésük alapján leginkább (definitjelölő funkciójú) birtokos végződéssel ellátott képzett melléknévnek néznek ki, mondatbeli funkciójuk azonban nem támasztja alá a melléknévi értelmezést.

Az *n*-tövű névszókban (ilyen pl. maga a népnévként szolgáló nganaszan szó is) a tővégi nazális csak bizonyos ragozott alakokban jelenik meg, de hogy melyikekben, és hogy az *n* eltűnése hogyan függ össze a fokváltakozással, arra legalább három különböző típust találtunk, ezek csak az adott tővégi nazális lexikális jelölésével különböztethetők meg

egymástól: ennek megfelelően a különbözően viselkedő tővégi nazálisokat különbözőképpen jelöltük meg a lexikonban, és az elemző szabálykomponense is különbözőképpen kezeli őket. Vannak a fokváltakozás szempontjából látszólag kivételes szavak, amelyeket úgy tudunk modellezni, hogy olyan „álmássalhangzókat" feltételezünk a mögöttes fonológiai alakjukban, amelyek a felszínen soha sem, vagy csak speciális körülmények között jelennek meg. Az álmássalhangzóknak csak egyik típusa a már említett palatalizáló ^J. Emellett van egy pusztán a fokváltakozásban szerepet játszó álmássalhangzó, amelyet az elemző adatbázisában ^C-vel jelöltünk, ennek a fokváltakozásban betöltött szerepe mellett nincs egyéb hatása, mint a palatalizáció a ^J esetében. A harmadik álmássalhangzót ^N-nel jelöltük. Ez a nazális+zárhang kapcsolatok fokváltakozása során áll elő, és opcionálisan újra nazális hangként megjelenhet a nunnációnak nevezett folyamat eredményeként.

A webes elemzőnek a jelen OTKA kutatást megalapozó korábbi NKFP projektumban részt vevő MorphoLogic Kft. nyújtott helyet a webszerverén a következő oldalon (http://www.morphologic.hu/urali/index.php). A jelenlegi tervek szerint ezen a helyen a jövőben más uráli nyelvekhez készült elemzők is elérhetőek lesznek majd.

Az elemző webes megjelenítésének feltétele volt, hogy a Xerox elemzőeszközét webszerverrel integráljuk. Ez a megoldás lehetővé tette azt is, hogy az elemző és a szóalak-generátor (lásd a 4.3. pontot) ne csak az eredetileg kidolgozott 7 bites karakterkészlettel legyen használható (ebben a *ŋ* hangot *ng*, a *ń* hangot *n1* jelöli stb.), hanem jobban olvasható formátumban is kommunikálni lehessen vele. Végül úgy oldottuk meg a webes felületet, hogy az elemzendő szöveget, illetve a generálandó szó tövét a program akár a 7 bites ASCII formátumban, akár a kutatók által az egyes fonémák jelölésére használt számos egyéb karakter használatával beírhatja. A *t* hang gyenge fokú párjának jelölésére pl. a következő jelölések bármelyike használható (akár vegyesen is): q, δ, ð, (illetve adott esetben ezek nagybetűs párja is: Q, Δ, Ð). A felhasználó két különböző jól olvasható kimeneti formátum közül választhat. A másik megoldandó feladat az volt, hogy a gyakran előálló nagyszámú elemzést emberi fogyasztásra alkalmas formában jelenítsük meg, illetve hogy az elemzésekben szereplő tövekhez glossza is tartozzon (jelenleg magyar nyelven). Megoldásként a fentebb bemutatott interlineáris megjelenítésmódot alakítottuk ki, amelyben minden szóhoz csak egy elemzést jelenítünk meg, így a szöveg természetes módon (balról jobbra) olvasható marad, színkóddal utalunk arra a tényre hogy több elemzés van, és az egérmutatót az elemzés fölé tolva automatikusan megjelenő menüből lehet az éppen megjelenítettől különböző elemzést választani, tehát a felület egyben kézi egyértelműsítő-felületként is szolgál. Hasonlóan ergonomikus megjelenítésű webes morfológiaielemző-

szolgáltatásra eddig nem láttunk példát. Ugyancsak úttörő jellegű a szóalak-generátor webes szolgáltatásként való megjelenítése.

4.3. A szóalak-generátor

Kutatásaink során számos példaparadigmát alkottunk többek között a Chrestomathiában is. Ezeket kigyűjtve és egy erre a célra írt program segítségével olyan alakra hoztuk, hogy automatikusan össze lehessen vetni azzal, amely a számítógépes modell szerint az adott morfoszintaktikai jegyeket megtestesítő szóalak felszíni alakjának lennie kell.

Ehhez ki kellett dolgoznunk egy egységes morfoszintaktikai annotációs jelölésrendszert, amely mind a szövegek annotálásakor, mind a monográfiában alkalmaztunk. Ennek tökéletesítése az utolsó pillanatig eltartott.

Az általunk és a gép által megalkotott paradigmák összevetéshez szükség volt egy szóalak-generátorra, melynek a feladata az, hogy egy adott tőből és az adott tőre alkalmazható morfoszintaktikai jegyek egy jól formált halmazából (pl. [*megy*+ige+múlt idő+kij. mód+tsz.+3.sz.]) előállítsa a számítógépes szóalaktani modell szerint helyes felszíni megvalósulását.

Fontos megjegyezni, hogy az elemző önmagában nem lett volna alkalmas eszköz a paradigmák kimerítő ellenőrzésére, mert ha egy szóalaknak több lehetséges alakja van, és valamelyik lehetséges alak a példaparadigmában nem szerepel, akkor azt az elemző használatával nem vesszük észre. A nganaszanban a példában említetthez hasonló szabad váltakozások gyakoribbak, mint a magyarban.

A szóalak-generátor a további munkálatokban is nagy segítségünkre volt, hiszen minden „hiba”, azaz minden az elemző által megelemzetlen alak esetén segítségül hívtuk a generátort, amivel megalkottuk az általunk adott esetben helyesnek tartott alakot. Így következtethettünk a hiba forrására. A generátor nélkül nem derülhetett volna ki, hogy mi az elemző szerinti helyes alak, ami rendkívül megnehezítette volna az eredmények kiértékelését.

4.4. Leíró nganaszan nyelvtan (Descriptive Grammar of Nganasan)

A projekt másik legfontosabb eredménye az angol nyelvű nganaszan leíró nyelvtan, amely korszerű keretek között, eddig nem tapasztalt alapossággal megalkotott mű, amelyben a közölt adatok, példamondatok legnagyobb része saját gyűjtésből származik. Ez azért is fontos, mert így a mai nyelvállapotot tükrözi a leírás. A kötetben a szerzők nem tértek ki nyelvtörténeti adatokra, mivel a nganaszan nyelvtörténet egy újabb kutatás tárgya lehetne.

Egyes helyeken azonban utaltunk a nyelvtörténetre, ha az elősegítette a szinkrón nyelvállapot megértését.

A mű közvetlen előzménye a Magyarországon 2002-ben megjelent, jelen projekt kutatói által készített Chrestomathia Nganasanica. Az angol nyelvű leírás nem egyetemista hallgatóknak, hanem nyelvészkutatóknak, többek között nyelvtipológusoknak készült. A bekerültek Wagner-Nagy Beáta habilitációs értekezésnek, valamint Szeverényi Sándor és Várnai Zsuzsa doktori értekezésének eredményei is. Mindezen felül pedig mind az elemző, mind a terepmunka által szerzett újabb eredmények is beépültek a műbe. Kiemelendő értéke a monográfiának az is, hogy a kézirat angol nyelven készült, így szélesebb körben ismerhetik meg a nganaszan nyelvet az érdeklődők.

Ahogy a hazai, úgy a nemzetközi szakirodalom sem bővelkedik a kisebb uráli nyelvek monografikus feldolgozásában. Kéziratunk messze meghaladja a közelmúltban megjelent leírásokat, egyrészt a leírás terjedelme, alaposssága, másrészt a bedolgozott saját, új kutatási eredmények okán. A munka jelenleg kézirata ca. 17,5 ív. A kötetben található két, saját gyűjtésű annotált szöveg is, amely a terepmunka eredménye. A kötet rövid tartalomjegyzéke a következő:

4.5. Újabb terepmunka-kutatások a nganaszanok körében

A kutatás során két terepmunkát terveztünk, a második és a harmadik évre. Az első terepmunka során számos olyan elsősorban grammatikai kérdést kívántunk tisztázni, amelyek megválaszolásával az elemző működése javítható. A terepmunka célja irányított morfológiai gyűjtés volt, hiszen még mindig voltak olyan morfológiai jelenségek, amelyek csak egy-két adattal voltak reprezentálva (pl. prohibitív igemód), így ezek morfológia viselkedéséről nehéz a számítógép számára adekvát leírást készíteni. Ezeket a morfémákat (illetve a velük alkotható paradigmákat) a terepmunka során célirányosan kívántuk felgyűjteni.

Mindemellett olyan szövegeket (hanganyaggal) is kívántunk gyűjteni, melynek a témája az eddigi gyűjtésekkel szemben elsősorban a köznapi témákra koncentrál. Erre azért van szükség, mert az eddig rendelkezésre álló szövegek legnagyobb része folklór- vagy szakrális szöveg. Ez azt jelenti, hogy a hétköznapi beszédtémák szókincse csak hiányosan adatolt, tehát szükségesnek látszik egy célirányos lexikai gyűjtés.

A második terepmunka során a program újabb tesztelése során felmerült problémákra, valamint a gyűjtött anyag feldolgozása során felmerült kérdésekre kívántunk volna választ kapni. Természetesen e kutatóút során is folytattuk volna a szöveggyűjtést, valamint az első út során gyűjtött szövegek ellenőrzését.

A tervezett két kutatóútra ugyan sor került, de egyrészt adminisztrációs, másrészt anyagi okokból nem az előzetes terveknek megfelelően. Ennek következtében nem volt alkalmunk a terepmunkában megszokott ellenőrző útra. A terepmunka a következő menetben zajlott:

1. 2008. májusában Várnai Zsuzsa 2 hetet töltött Dugyinkában.
2. 2008. júliusában 3 hetet Uszty-Avamban Szeverényi Sándor és Wagner-Nagy Beáta.

A terepmunkának a során összesen 7 adatközlővel sikerült mintegy 150 munkaórát dolgozni.

A terepmunka 4 összetevőből állt:

a) egy-egy nyelvtani problémára való rákérdezés mondatfordítással, szituációk eljátszásával (kérdőíves módszer) – Ennek során összesen 30 különböző kérdőívet sikerült lekérdezni. Egy-egy kérdőívet általában 5 adatközlővel dolgoztunk fel. A következő morfológiai-szintaktikai területekről állítottunk össze kérdőíveket: 1) a reflexív ragozás használata, különös tekintettel a mozgást jelentő igék reflexív ragozására, 2) a tiltás és a tagadás kifejezése (általános tiltás, a tagadó szerkezetek számos csoportja), 3) vonzatos melléknevekre és igevonzatokra irányuló kérdőív. (Ez a téma különösen fontos, mivel a szótárak a vonzatokról semmiféle információt nem adnak). 4) az igemódos alakok használata, az analitikus szerkezetek használata, az auditív / esetleg szenzitív (szaglás, tapintás) igemód használata, debitív igemód, 5) egyeztetési kérdések, 6) összetett szavak tesztje, 7) a predikatív ragozás használata birtokos személyragos alakok esetében, 8) a predikatív szerepű főnév melléknévi jelzőjének morfológiai viselkedése, 9) a névmások különböző esetragos alakjai, a személyes névmások különböző klitikumokkal ellátott alakjai, 11) melléknévi jelzők sorrendje, 12) evidenciális, rezultatív, másodlagos rezultatív, depiktív, passzív, egzisztenciális, kérdő és feltételes mondatok

Az előre elkészített kérdőívet adott esetben a helyszínen módosítottuk, kiegészítettük. A kérdőívező munka nagy részéről készült hangfelvétel, összesen 27 órányi hangzóanyagot

vettünk fel. Ezek a hanganyagok vágás és szerkesztés után önmagukban is hasznosak lehetnek.

b) lexikai kérdések fordítással, szituációk eljátszásával, képek, rajzok segítségével; A következő lexikai csoportokra gyűjtöttünk adatokat: 1) testrésznevek, 2) számnevek (törtszámnevek, nulla stb.), 3) időhatározók és időkifejezések, 4) az 'ad' jelentésű igék disztribúciója, valamint vonzatkeretük, 5) dimenzionális melléknevek, 6) *-səbtə* képzős alakok, 7) íz- és szagnevek, színnevek, 8) hónap- és évszaknevek, 9) rokonsági viszonyok.

c) szöveggyűjtés (irányított témáról elbeszélés, pl. napi élet, korábbi élet, a család története; párbeszéd; rövid mese). Összesen 12 szöveget (összesen kb. 140 perc, kb. 1150 mondat) szöveget vettünk fel, melyeket kivétel nélkül a helyszínen le is jegyeztünk, valamint nagy részüket egy másik adatközlővel ellenőriztük is.

d) szociolingvisztikai, nyelvhasználati kérdőívek, interjúk: 6 fiatal (15-20 éves) és 7 idős (60-65 éves) adatközlő nyelvhasználati szokásainak feltérképezése, hogyan látja a nyelve jövőjét, milyen szerepe van a nyelvnek az etnikai identitásban. A kisebb, veszélyeztetett uráli népeknél folytatott terepmunka óhatatlan hozadéka a nyelvhasználati helyzet feltérképezése az adott adatközlőnél. Szabad beszélgetés során elhangzik minden olyan eleme egy nyelvhasználati kérdőívnek, ami alapján pontos képet kaphat a kutató az adatközlő nyelvhasználati szokásairól.

A terepmunka során elvégzett feladatok közül a leghangsúlyosabb a grammatikai problémákat tisztázó, kérdőívező és a lexikai összetevők, melyek mind az elemző tökéletesítéséhez, mind a monográfia elkészítéséhez nagy mértékben hozzájárultak. Sajnálatos módon kisebb hangsúlyt kaphatott a szöveggyűjtés. Ennek legfőbb oka, hogy a tervezett két kutatóút helyett csak egy valósulhatott meg. A szöveggyűjtés módszeréből adódóan igen aprólékos és időigényes. Ahhoz, hogy közzétehető szövegeket teremtsünk szükséges a rögzített hanganyagot egy adatközlővel együtt visszahallgatva lejegyezni, amely embert próbáló, nagy koncentrációt igénylő feladat mind a közlő, mind a gyűjtő számára. Az így elkészült szövegben még mindig maradhatnak problémás adatok, részek, amelyekre célszerű ismét rákérdezni egy másik alkalommal és lehetőség szerint egy másik adatközlővel (ahogyan a terveinkben szerepelt). Éppen ez az oka annak, hogy a második terepmunka hiányában a szövegeket még nem publikáltuk. Minden általunk gyűjtött szövegben vannak még kisebb-nagyobb tisztázandó kérdések (pl. ismeretlen szó, szokatlan szóalak stb.) Mindazonáltal ezek a szövegek is alkalmasak kutatásra. Ennek ellenére a szövegek feldolgozás alatt állnak és folyamatosan kerülnek majd fel az elemzett szövegek közé. A fennmaradt lejegyzési problémák tisztázása csak egy újabb pályázat (és így kutatóút) keretében lehetséges.

5. A kutatás további hasznosíthatósága

Az elemzőprogram tökéletesítése már önmagában nagy jelentőséggel bír: olyan kutatók (uralisták, általános nyelvészek, fonológusok) számára is elérhetővé, érthetővé teszi a nganaszan nyelvet, akik korábban nem folytattak szamojédológiai tanulmányokat. Továbbá a nganaszan nyelvet kutatók gyorsabban jutnak hozzá elemzett nyelvi adatokhoz, s ez a kutatási idő lerövidülését jelentheti.

1. A felhasználó maga tölthet fel szövegeket. Ameddig ennek jogtisztasága nem tisztázott, ezt csak maga kutathatná, amennyiben korlátok nincsenek, ez mindenki számára hozzáférhető lehetne. A felhasználó nemcsak a maga által feltöltött szövegeket használhatná, hanem a „közös"-eket is.

2. A felhasználó a feltöltött szövegre morfológiai elemzést kap, melyet egyértelműsíthet.

3. Az elemzőprogram a hozzá kapcsolódó szóalak-generátorral együtt jól hasznosítható az oktatásban is. Akár finnugrisztikai, számítógépes nyelvészeti, általános és elméleti, morfológiai és szintaktikai, vagy fonológiai kurzusokon gyakorlati segítséget jelenthet, szemléltető eszközként alkalmazható. Éppen ezekért az értékeiért nem hátránya, hanem előnye, hogy az elemzőprogram magyar nyelvű glosszákat ad vissza.

6. A projekt személyi és tárgyi feltételeinek teljesülése és változásai

- 2008 novemberétől a vezető kutató 3 hónapnál hosszabb időre külföldre távozott, így megbízott vezető kutatóként Várnai Zsuzsa vette át a helyét, de ezáltal a kutatás tematikája és költségvetése nem változott.

A terepmunka előkészítésében és az anyag feldolgozásában, valamint az elemző program kidolgozásában és a felmerülő problémák megoldásában számítottunk Eugen Helimki hamburgi professzorral és Valentin Gusev, moszkvai kutatóval való konzultációkra. A kutatás során szükségesnek tartottuk a személyes konzultációk megszervezését, évente legalább egy-egy hétre. Sajnálatos módon Eugen Helimskivel csak néhány alkalommal sikerült találkoznunk. Korai, hirtelen halála miatt nem volt lehetőségünk sem közös terepmunkára, sem arra, hogy a projektet tanácsaival végigkísérje. Valentin Gusev, a orosz Akadémia Nyelvtudományi Intézetének munkatársa és kollégája, Marina Brykina, a Moszkvai Állami Egyetem dolgozója azonban mindvégig segítségünkre voltak. Nélkülük, valamint Oksana Dobzhanskaja segítsége nélkül terepmunkánk nem jöhetett volna létre. Az utóbbi kapcsolat fenntartásában nagy segítségünkre volt az MTA és az orosz akadémia között 2005–2007, majd 2008–2010 évekre meghosszabbított együttműködési szerződés, melynek

keretein belül évente legalább egy-egy hét időtartamra közösen tudtunk dolgozni hol Budapesten, hol Moszkvában

- Már a második év elején kiderült, hogy a pályázatban szereplő, utazásra szánt pénzösszegek nem lesznek elegendők a tervezett kétszer két fő terepmunkájára. Ennek egyik oka az volt, hogy az eredeti pályázatunkban szereplő összegeket a pályázat megítélésekor a bizottság érthetetlen módon jelentősen megkurtította, annak ellenére, hogy köztudott, hogy az oroszországi utazási költségek évről évre nőnek. Ez utóbbi esemény természetesen be is következett, tehát időközben az utazási költségek (kölönösen a helikopter ára) jelentősen növekedtek. Így a kutatóutat az eredeti tervektől eltérően három főre (1+2) szerveztük. Ennek a megoldásnak a következtében természetesen számolnunk kellett azzal a ténnyel, hogy a második útra, amely során az első terepmunka anyagainak feldolgozása közben felmerült kérdéseket tisztáztuk volna, nem lesz majd lehetőség a tervekben foglaltak szerint. Ez a lehetőség ellentmond a terepmunka metodikai szokásainak, hiszen nincsen módunk ellenőrizni és kiegészíteni az adatainkat a második alkalom elmaradásával. A 2007-es év során ugyan benyújtottuk vízumkérelmünket, valamint a Tajmir-félszigetre való beutazáshoz szükséges engedély iránti kérelmet, ami azonban az utazás tervezett időpontjáig nem érkezett meg. Ebben az évben így nem került sor a gyűjtőútra. A szerencsétlen körülmények miatt a terepmunka a következő menetben zajlott: 2008. májusában Várnai Zsuzsa 2 hetet töltött Dugyinkában, majd 2008. júliusában Szeverényi Sándor és Wagner-Nagy Beáta 3 hetet Uszty-Avamban
- Az elmaradt kutatóút után maradtak további kihasználatlan utazási forrásaink, amelyek ugyan még egy út megszevezésére akár egy főre is nem bizonyultak elegendőnek, így konferencialátogatásra tudtuk fordítani a fennmaradt összegeket.