Kutatási beszámoló az OTKA T-048309 sz. projektumról

A **Permi nyelvészeti adatbázisok** c. projektum három önálló részből áll, amelyeket egyrészt a vizsgált nyelvek, másrészt a kutatási módszer kapcsol össze. A kutatás mindhárom komponens esetében a finnugor nyelvcsalád permi ágát alkotó két nyelvre, a komira (zürjén) és az udmurtra (votják) irányul. A módszerben az a közös, hogy mindhárom esetben a számítástechnika és a számítógép hatékony felhasználására volt szükség.

Mindhárom részmunkálatnak vannak bizonyos előzményei. Az **Uráli és permi etimológiai adatbázis** a korábban elkészült Uráli etimológiai adatbázison alapul, annak kiegészítése. A **Morfológiai elemzők**-nek van egy korábbi változata, amelyet most jelentős mértékben kiegészítettünk és tökéletesítettünk. Az **Udmurt nyelvjárási atlasz** előzményének az a kézzel rögzített anyag tekintendő, amelyet V. K. Kelmakov professzor (Izsevszk) diákjai gyűjtöttek a professzor irányításával.

Az etimológiai adatbázis Csúcs Sándor irányításával készült, a másik két komponens Fejes László munkája

Uráli és permi etimológiai adatbázis

Az alapul szolgáló Uráli etimológiai adatbázis (UEDb) egy korábbi OTKA támogatással készült, és a Rédei, Károly (szerk.) Uralisches Etymologisches Wörterbuch (Akadémiai Kiadó, Budapest 1986-88.= UEW) anyagát tartalmazza. Ez a mű elvi és gyakorlati (terjedelmi) okok miatt az uráli nyelvek etimológiailag közös szókincsét tartalmazza, felölelve az uráli, a finnugor, a finn-permi, a finn-volgai és az ugor alapnyelvre rekonstruálható szókészletet. Számítógépes adatbázis esetében terjedelmi korlátok gyakorlatilag nincsenek, ezért célszerűnek tűnt az adatbázis bővítése. Mivel közben elkészült és elektronikus formában is rendelkezésre állt az őspermi szókészlet rekonstrukciója (nyomatatott változata: Sándor Csúcs, Die Rekonstruktion der permischen Grundsprache. Akadémiai Kiadó, Budapest, 2005), az adatbázis bővítését az őspermi etimológiákkal kezdtük, vagyis azokkal a szavakkal, amelyek ma csak két permi nyelvből mutathatók.

Az eredeti permi adatbázis struktúrája, pontosabban a permi szócikkek struktúrája eltért az UEDb szócikkeiétől, de az átalakítás nem okozott különösebb problémát. Ezek után tehát az adatbázist más alapnyelvekhez (korai közfinn, közfinn, ugor, szamojéd) tartozó etimológiákkal is kiegészíthetjük. A bővítés eredményeként az adatbázis alapnyelvi lexikai egységeinek (szócikkeinek) száma 1876-ról 2650-re nőtt.

Az UEW-ben és az UEDb korábbi változataiban a rekonstruált alapalakok jelentését németül adtuk meg. A bővítés itt is célszerűnek látszott, az adatbázis potenciális felhasználóit figyelembe véve, az alapjelentést magyarul, finnül angolul és oroszul is megadtuk. Az alapjelentés rekonstruálásának nincsenek egzakt módszerei, és az összetettebb jelentésű szavak nyelvek közti megfeleltetése is gyakran problematikus, ezért törvényszerű, hogy az általunk megadott jelentések is sok esetben vitathatók vagy hozzávetőlegesek, tájékoztató jellegűek. Szolgálhatnak ugyan jelentéstani kutatások kiindulásául, de mélyebb elemzéshez mindig figyelembe kell venni az etimológiához tartozó mai nyelvi adatok jelentését is.

Nem az UEW volt az első mű, amelyben rekonstruált alapalakok találhatók. Több szempontból is indokoltnak látszott más szerzők (Collinder, Janhunen, Sammallahti) rekonstrukcióit is bedolgozzuk az adatbázisba. A különböző szerzők által rekonstruált (hipotetikus) alapalakok összehasonlítása során mutatkozó egyezések természetesen növelik a rekonstruált alak valószínűségét, a hangalaki különbségek elemzése pedig megtermékenyítő hatással lehet az összehasonlító hangtani és hangtörténeti kutatásokra, hiszen olyan kérdések feltevését indukálja, hogy Milyen szempontok alapján döntöttek a szerzők egyik vagy másik megoldás mellett? illetve Hogyan lehetne feloldani relonstruktumokban mutatkozó ellentmondást.

Az adatbázissal folytatott munka során folyamatosan fedeztünk fel kisebb-nagyobb hibákat. Ezért szükségesnek és célszerűnek látszott az UEDb teljes anyagának ellenőrzése. A hibák nagy részét az emberi logikával készült nyomtatott szótár és a számítástechnikai logikával készült adatbázis közti módszertani különbség okozta. A feltárt hibák kijavítása megtörtént, ami természetesen nem jelenti azt, hogy az adatbázis már hibátlan. Újabb hibák felfedezésére mindig számíthatunk.

Morfológiai elemzők

A morfológiai elemzők esetében a legnagyobb kihívást a cirill helyesírásban megjelenő, a morfok ábrázolásában tapasztalható bonyolult váltakozások kezelése jelentette. A tesztek alapján kijelenthetjük, hogy az ezzel kapcsolatos problémákat sikerrel oldottuk meg.

A komi elemző fejlesztésében igen jelentős lépés, hogy a korábbi változatban használt, 1600 tételes tőtárat egy több mint 30000 tételes tőtárra cseréltük. Ezt a tőtárat újabb korpusz (komi nyelvű hírek az interneten) bevonásával még tovább bővítettük. A nyelvtan is jelentős fejlődésen ment át (újabb végződések, pl. személyragozott infinitivus, a személyes névmások alakjainak gazdagodása stb.)

Az udmurt tőtáron csak kisebb mértékű bővítést kellett végeznünk (eredetileg is több mint 13000 tételt tartalmazott), a nyelvtanon azonban itt is jelentős bővítés történt. Sikerült azonosítanunk olyan igenévi alakokat, melyekről a szakirodalom nem tesz említést.

Természetesen egy morfológiai elemző fejlesztése sosem tekinthető lezártnak. Bizonyára vannak az elemzőben még olyan rejtett hibák, melyek csak intenzív alkalmazás során derülhetnek ki. Az elemzők jelenlegi állapotukban azonban már igen hatékonyak, és gyakorlati célokra is felhasználhatók. Várható, hogy az elemzőkre építve 2009 folyamán kiadjuk a komi és az udmurt helyesírásellenőrzők első változatát.

Udmurt nyelvjárási atlasz

Elkészült, és akár jelenlegi formájában is kiadható lenne az udmurt nyelvjárási atlasz. 396 térképlapból áll, minden térképlaphoz tartozik egy jelmagyarázat. A jobb értelmezhetőség kedvéért nem kapott minden előforduló alakváltozat külön jelet. Ehelyett azt vettem figyelembe, hogy a kérdés milyen jelenséggel kapcsolatban merül fel: az adott probléma szempontjából releváns különbségeket jelöltem, a lényegtelenek nem. (Így pl. a magánhangzók vizsgálatát célzó kérdéseknél nem kaptak különböző jelet a mássalhangzójukban eltérő alakváltozatok stb.) A nyelvi adatok szerepelnek az udmurt

nyelvjárástanban használatos cirill átírásban (a lejegyzésnél is használt rendszerben), IPÁ-ban és a latin alapú átírásban is.

Az előzetes tervekkel szemben nem készültek el a 397-424. számú kérdésekre épülő térképlapok. A 397-400. kérdések a korábbiakkal szemben nem udmurt szóalakokra, hanem megadott udmurt szavak jelentésére kérdeznek rá. Ezek elkészítéséhez egy kissé módosított programra lenne szükség. A 401-424. kérdések igei paradigmát tartalmaznak. Ezek feldolgozására nem voltak egyértelmű szempontok, ráadásul az egyes alakoknál különböző szempontok merülhetnének fel. Az anyag feldolgozását inkább egy különálló műben kellene feldolgozni. Mivel az anyag így is igen tekintélyes méretű, és publikáláskor inkább néhány térképlap kihagyását, mint továbbiak betűzését kellene megfontolni (bizonyos térképlapok ugyanazt a jelenséget vizsgálják különböző lexémákon: ezek nem mindig térnek el egymástól olyan mértékben, hogy mindegyiküket érdemes lenne papíron publikálni; egyes térképlapok nem mutatnak területi variálódást, s csak egy jel szerepel a térképlapon stb.). Hasonló okokból maradt el az összevető térképlapok elkészítése is. Hangsúlyozandó azonban, hogy a kifejlesztett szoftvercsomag segítségével ezek a térképlapok könnyen legyárthatók lennének.

Az atlasz publikációjához azonban célszerű lenne még bizonyos feladatokat elvégezni. A legfontosabb feladat a térképlapokhoz fűzött részletesebb kommentárok elkészítése. Hasonlóképpen érdemes lenne megnézni, hogy az utóbbi években gyűjtött anyagokkal kiegészíthetők-e a pillanatnyilag kevéssé dokumentált településekre vonatkozó adatok. (Az új adatok rögzítése után a térképlapok frissítése a jól megtervezett szoftveres háttérnek köszönhetően mindössze néhány percig tart.) Bár rendelkezésre áll településenként lebontva a forrásadatok gyűjtőinek és az informánsoknak a listája, a gyűjtést szervező tanszék feladata lenne összeállítani azoknak a tanároknak a listáját, akik az anyaggyűjtést szervezték és a füzetek ellenőrzését végezték. Mindezeket az elkövetkezendő időben az Udmurt Állami Egyetem megfelelő szervezeti egységeivel együttműködésben kell elvégezni.

Mutatvány a komi morfológiai elemzőből

```
510     da{da[S_C]|da[S_Mod]}       [da:506 Da:1 da.:3 ]
 321    >.{}    [>.:17 >:304 ]
 223    komi{komi[S_N]+[I_NOM]}         [komi:101 Komi:122 ]
 171    i{i[S_C]|i[S_Mod]|i[S_N|ltr]+[I_NOM]}     [i:145 I:19 i.:1 I.:6 ]
 158    respublikasa{respublika[S_N]+sa[D=A_MltFun]+[I_NOM]}        [respublikasa:80
Respublikasa:78 ]
 133    rajonsa{rajon[S_N]+sa[D=A_MltFun]+[I_NOM]} [rajonsa:124 Rajonsa:9 ]
 100    luno0{lun[S_N]+o0[D=N_Dimin]+[I_NOM]|lun[S_N]+o0[I_ILL]}        [luno0:95
luno0.:5 ]
  95    --{}    [--:95 ]
  91    a{a[S_C]|a[S_Mod|1)introg]|a[S_Inter]|a[S_N|ltr]+[I_NOM]}       [a:48 A:31 A.:12 ]
  89    jylysq{jyv[S_PP]=jyl+ysq[I_ELA]|jyv[S_N]=jyl+ysq[I_ELA]}    [jylysq:60 jylysq.:29
]
  85
        vylo0{vylo0[S_Adv]|vyv[S_PP]=vyl+o0[I_ILL]|vyv[S_N]=vyl+o0[D=N_Dimin]+[I_
NOM]|vyv[S_N]=vyl+o0[I_ILL]}    [vylo0:68 vylo0.:17 ]
  79    kulqtura{kulqtura[S_N]+[I_NOM]} [kulqtura:77 Kulqtura:1 kulqtura.:1 ]
```

78
myj{myj[S_N|ProWhat]+[I_NOM]|myj[S_A|ProWhat]+[I_NOM]|myj[S_Adv|ProWhat]|myj[S_C]} [myj:77 Myj:1]
74 nac1ionalqno0j{nac1ionalqno0j[S_A]+[I_NOM]} [nac1ionalqno0j:51 Nac1ionalqno0j:23]
74 tajo0{tajo0[S_N|ProThis]|tajo0[S_N|ProThis]=ta+jo0[S_Nom]} [tajo0:43 Tajo0:31]
60 si1dzz1o0{si1dzz1o0[S_Adv|ProThat|ugyan]} [si1dzz1o0:47 Si1dzz1o0:13]
54
kuza1{kuza1[S_PP1]|kuza1[S_Adv]|kuza1[S_N]+[I_NOM]|kuzq[S_A]=kuz+a1[D=N_Abstr2]+[I_NOM]|kuzq[S_A]=kuz+a1[D=Adv_Adverb]|kuzq[S_N]=kuz+a1[D=A_Prov]+[I_NOM]} [kuza1:52 kuza1.:2]
3
pyvsa1n{pyvsa1n[S_N]+[I_NOM]|pyvsqyny[S_V|intr]=pyvs+a1n[D=A_PImpPs]+[I_NOM]|pyvsqyny[S_V|intr]=pyvs+a1n[D=N_Tool]+[I_NOM]|pyvsqyny[S_V|intr]=pyvs+a1n[I_PrsFutSg2]}[pyvsa1n:3]
3
radejto0{radejtny[S_V|tr|5)dial]=radejt+o0[I_PrsSg3]|radejtny[S_V|tr|5)dial]=radejt+o0[I_ImpPl2]} [radejto0:3]
3
radejtysqa1so0s{radejtysq[S_A]+a1s[I_Pl]+o0s[I_ACC]|radejtysq[S_A]+a1s[I_Pl]+o0s[I_PSS1]+[I_ACC]|radejtny[S_V|tr|5)dial]=radejt+ysq[D=A_PImpAct]+a1s[I_Pl]+o0s[I_ACC]|radejtny[S_V|tr|5)dial]=radejt+ysq[D=A_PImpAct]+a1s[I_Pl]+o0s[I_PSS1]+[I_ACC]|radejtny[S_V|tr|5)dial]=radejt+ysq[D=N_Agent]+a1s[I_Pl]+o0s[I_ACC]|radejtny[S_V|tr|5)dial]=radejt+ysq[D=N_Agent]+a1s[I_Pl]+o0s[I_PSS1]+[I_ACC]} [radejtysqa1so0s:2 radejtysqa1so0s.:1]
3 Raisa{Raisa[S_N|1stnameF]+[I_NOM]} [Raisa:3]
3 rajon{rajon[S_N]+[I_NOM]} [rajon:3]
3 regionkostsa{region[S_N]+kost[S_PP]+sa[S_>A]+[I_NOM]} [regionkostsa:2 Regionkostsa:1]
3
regionq1a1s{region[S_N]=regionq1+a1[I_INE]+s[I_PSS3]|region[S_N]=regionq1+a1[I_ILL]+s[I_PSS3]|region[S_N]=regionq1+a1s[I_Pl]+[I_NOM]} [regionq1a1s:3]
3 regionq1a1sysq{region[S_N]=regionq1+a1s[I_Pl]+ysq[I_ELA]} [regionq1a1sysq:3]