A kisebb uráli nyelvek veszélyeztetett nyelvek, ezért dokumentálásuk nemzetközi jelentőségű feladat. Célunk e nyelvek dokumentálásának céljára morfológiailag annotált korpuszok létrehozása, illetve az ehhez szükséges eszközök megteremtése volt. A kutatás előzménye volt az NKFP-5/135/01-es pályázat, melynek során több kis uráli nyelv morfológiai elemzője készült el, illetve az OTKA 048309-es pályázat, melynek keretében pedig a permi nyelvek elemzőinek fejlesztése folyt. Az MTA Nyelvtudományi Intézetének Finnugor osztályán folyt az OTKA K 60807 pályázat is, melynek során nganaszan morfológiai elemző épült: ez szintén az NKFP-5/135/01 folytatása, bizonyos megoldások az összes elemző felhasználhatóságát együtt biztosítják.

A jelen projekt a két obi-ugor nyelv három nyelvjárását ölelte fel, és négy fő modulra oszlott. A modulok mindegyikében a technikai feladatokat elsősorban Novák Attila oldotta meg. A nyelvi anyag előkészítését a manysi elemzőknél Fejes László, a szinjai hantinál Ruttkay-Miklián Eszter, a kazimi hantinál Sipos Mária végezte el. Fejes László jelentős szerepet játszott a hanti nyelvtanok szabályokba kódolásánál is. A projektben külső munkatársak is dolgoztak, elsősorban a szövegrögzítés feladatát látták el. A külső munkatársak közül kiemelendő Sipőcz Katalin, aki a Vogul népköltési gyűjtemény szövegeinek rögzítését koordinálta és ellenőrizte, továbbá manysi nyelvi kérdésekben konzulensi szerepet látott el; illetve Wenszky Nóra, aki az OCR-ezésben, a tőtárak és szövegek angolra fordításában, ill. a szövegek és fordítások egymáshoz rendelésével járult hozzá a projekt sikeréhez.

## 1. Vogul (manysi) északi nyelvjárás: Kálmán Béla gyűjtése

Kálmán Béla két gyűjteménye között átfedés van, de az eltérő transzkripció miatt két külön elemzőt készítettünk. Az elemzőben megadott jelentések elsősorban a kiadványok saját szószedeteire épülnek, az ezekben található magyar és német jelentést az angol megfelelővel is kiegészítettük.

A Chrestomathia Vogulica szövegeinek elemzéséhez már az NKFP-5/135/01 projekt keretében elkészült a morfológiai elemző. A jelen projektben ennek egyértelműsítését vállaltuk. Az egyértelműsítés során előbukkantak az elemző korábban nem sejtett hibái, ezeket javítottuk, az egyértelműsítést újra elvégeztük. Az egyértelműsítő elérhető a http://www.morphologic.hu/urali/ oldalon, az elemezhető szövegek pedig a http://www.morphologic.hu/urali/texts/texts_chvog.html oldalon magyar és angol fordításokkal együtt, az utóbbiak mind, az előbbiek részben a projektum keretében készültek. (Az eredeti kiadvány magyar fordításokat nem tartalmaz. A fordításokat részben a Wogulische Textéből, részben a Vogul népköltési gyűjteményből vettük át. A 24. szöveg magyar fordítása sehol nem szerepelt, ezt magunk készítettük el.) Az egyértelműsített szövegek az interneten a http://www.morphologic.hu/urali/texts/texts_chvog_ana.html oldalon keresztül elérhetőek el.

A Wogulische Texte mit einem Glossar északi anyagának teljes feldolgozása a jelen projekt feladata volt. Ez magába foglalta a szövegek digitalizálását, az elemző elkészítését, illetve mintaszövegek egyértelműsítését. Az elemző elkészült és el is érhető a http://www.morphologic.hu/urali/ oldalon, a hozzá tartozó szövegek pedig a http://www.morphologic.hu/urali/texts/texts_wt.html oldalon. A szövegekhez a projekt keretében angol fordítás is készült, ami megkönnyíti a feldolgozást a külföldi kutatók számára: emellett a magyar fordítás is olvasható.

Kálmán Béla szalagos magnetofonra készített hangfelvételeit digitalizáltuk, később ezek a honlapon is meg fognak jelenni.

## 2. Vogul (manysi) északi nyelvjárás: Munkácsi Bernát gyűjtése

A Vogul népköltési gyűjtemény északi anyagát eredetileg kb. 60000 szóra becsültük, a valóságban azonban a létrejött korpusz szószáma meghaladja a 80000 szót. A projekt tervezésekor arra számítottunk, hogy az elemző tőtárát a Munkács–Kálmán-féle Wogulisches Wörterbuchra építhetjük. A szótárat digitalizáltuk, a tőtár elkészült, ám mint kiderült, a szövegek és a szótár között nagy különbségek vannak. A szótárból igen sok szó hiányzik (különösen tulajdonnevek, ill. képzett szavak, főként igék), mások nem abban az alakban szerepelnek, ahogy a szövegben. Gyakori eset az is, hogy a szövegben kötőjellel írt alakok külön írva, a szövegben külön írt alakok kötőjellel vannak szótárazva. Ezen kívül gyakoriak a hosszúságbeli ingadozások is, főként a magánhangzók körében.

Eredeti szándékunk szerint a projekt végére az elemző képes lett volna megelemezni a szövegben előforduló minden szóalakot. Sajnos jelen pillanatban még mindig van 3000 olyan szóalak, melyre nem kapunk elemzést. Ezek a szóalakok egy-két alkalommal fordulnak elő a szövegben. A javítási folyamat tapasztalatai szerint a hibák jelentős része abból ered, hogy a szótárból hiányzik, vagy nem a szükséges formában szerepel a szóalak. Az esetek kisebb részében a tőtár téves karakterfelismerés miatt hibás, vagy a korpuszban van gépelési hiba. Az eseteknek csak egy kisebb része az, amikor a nyelvtani moduljainkon kell javítani. Ezeknek a hibáknak a javítása igen aprólékos munkát igényel (ellenőrizni kell az eredeti szöveget, azonosítani kell a szükséges tövet, ellenőrizni kell a szótárat stb.), éppen ezért igen lassan halad. A fennálló hibák ellenére az elemző már igen nagy hatékonysággal elemzi a szövegeket, egyes szövegekben minden szóalakot felismer, a többi szövegben is viszonylag ritkán maradnak fel nem ismert szóalakok.

Az elemző a Wogulisches Wörterbuchban szereplő magyar és német jelentéseket tartalmazza, kiegészítve angol megfelelőikkel. Az újonnan felvett töveknek megadtuk magyar és angol, ill. helyenként német jelentését. A honlapunkon publikált szövegek mellett ott lesz magyar fordításuk is. A szövegek angol fordítása már készül a Szegedi Tudományegyetemen nemzetközi összefogással futó Ob-Babel projekt keretében.

Az elemző elérhető a http://www.morphologic.hu/urali/ oldalon keresztül, a szövegek hamarosan felkerülnek a http://www.morphologic.hu/urali/texts/texts_hu.html oldalra. Bár az elemző még nem tökéletes, folyik a szövegek egyértelműsítése. Hamarosan ugyanitt jelennek meg az egyértelműsített szövegek is.

## 3. Osztják (hanti) szinjai nyelvjárás: Ruttkay-Miklián Eszter gyűjtése

A szinjai korpusz anyaga a korábban említett NKFP-pályázat keretében gyűjtött szövegek, „értelmező szótár”, melyben az adatközlő Steinitz DEWOS-ában szereplő szavakat magyarázza (kb. 52 óra hangzóanyag). Ezt rövid folklórszöveg is kiegészíti.

A morfológiai elemző építése fontos szerepet játszott a lejegyzésben található hibák és következetlenségek kiszűrésében. A korpuszból eltávolítottuk azokat a részeket, amelyek hasznos információkat nem tartalmaztak, azaz amelyekben az adatközlő csak annyit mondott, hogy a kérdezett szót nem ismeri. A tőtár a kutató nyelvi kompetenciájára, ill. a DEWOS-ra épült. Elkészült az elemző, a mintaszövegek egyértelműsítése, ill. ezek nyomán az elemző javítása, az újraegyértelműsítés. Az elemző a http://www.morphologic.hu/urali/ oldalon érhető el, a nyers szövegek a http://www.morphologic.hu/urali/texts/texts_sin.html oldalról, az egyértelműsített szövegek pedig a http://www.morphologic.hu/urali/texts/texts_sin_ana.html oldalról. Mind a glosszázás, mind a mondatok fordítása elérhető magyar és angol nyelven is. Mivel a teljes gyűjtést a kutató külön kiadványban szeretné közölni, az interneten keresztül csak azokat a szövegeket tettük

elérhetővé, melyeknek egyértelműsített változatuk is megvan.

## 4. Osztják (hanti) kazimi nyelvjárás: különböző gyűjtések

A kazimi elemző különböző forrásokból származó szövegek elemzésére készült, korpuszunk Wolfgang Steinitz, Rédei Károly és Schmidt Éva gyűjtéséből tartalmaz szövegeket (20000 szó). A szövegek több alnyelvjárást is lefednek a felső-kazimitól az Ob mentén beszélt kazimi alnyelvjárásig.

Az elemző tőtára elsősorban a kutató kompetenciájára, a fordításokra és a DEWOS-ra épül, a jelentések magyarul és angolul szerepelnek benne. Fordítás az internetes közlésben kizárólag a Rédei által gyűjtött szövegeknél szerepel (német nyelven). Az elkészült elemző a http://www.morphologic.hu/urali/ címen érhető el, a korpusz szövegei a http://www.morphologic.hu/urali/texts/texts_kaz.html címen. Az egyértelműsítés elkészült, de több hibát is feltárt, így az elemzőt javítottuk, az újraegyértelműsítés pillanatnyilag folyik. Az elkészült egyértelműsítések hamarosan felkerülnek honlapunkra.

## Internetes publikáció

A projekt keretében készült el az a webfelület (http://www.morphologic.hu/urali/), melyen keresztül több elemző is elérhető. Az eredetileg csupán a nganaszan elemző használatát biztosító oldal projektünk keretében kibővült, pillanatnyilag a nganaszan és az öt obi-ugor elemzőn kívül elérhető a komi és az udmurt elemző is. A felület különböző szövegbeviteli lehetőségeket is biztosít (bár a virtuális billentyűzet az obi-ugor elemzők közül pillanatnyilag csak a Kálmán Béla-féle transzkripciókhoz elérhető). Jelentős és a korábbi elemzőket is érintő fejlesztés, hogy az elemző összeveti a tövekhez megadott jelentéseket a fordításban szereplő szavakkal, és amennyiben hasonlóságot talál, azt a tövet sorolja előre, melyet a fordításban felismerni vél. Természetesen ez magát az egyértelműsítést nem teszi szükségtelenné, de mindenképpen meggyorsítja.

A jelentések és a szövegek angolra fordítása eredetileg nem szerepelt céljaink között, de a pályázat elfogadása után a bíráló kérésére módosításként felvettük a munkatervbe a tőtárak, ill. a szövegek egy részének angol fordításának elkészítését is. Ugyanakkor a költségvetés növelésére ekkor már nem volt mód, ezt a munkát is az eredeti költségvetésből gazdálkodtuk ki.

A finnugor nyelvek köréből a mienkéhez hasonló eredményként egyedül a bécsi mari morfológiai elemzőt említhetjük (http://www.univie.ac.at/negation/mari-language/soft/soft_en.html), de ez is csak egy nyelv egyetlen változatát ismeri.