OTKA 71707: OBI-UGOR MORFOLÓGIAI ELEMZŐK ÉS KORPUSZOK

Írta: Fejes László

2008. január 30

A projekt/pályázat vezető kutatója:	Fejes László
A jelen pályázatban szereplő résztvevő:	Sipos Mária, Ruttkay-Miklián Eszter
Projekt címe:	Obi-ugor morfológiai elemzők és korpuszok
OTKA száma:	NF 71707
Kezdő év:	2008
Záró év:	2010

KUTATÁSI TERV

A pályázat fő jellemzői

Ismeretes, hogy a kisebb uráli nyelvek veszélyeztetett nyelvek, ezért dokumentálásuk nemzetközi jelentőségű feladat. Mivel a magyar legközelebbi rokonairól van szó, dokumentálásuk feladata jelentős részben Magyarországra hárul, mondhatni nemzeti kötelességünk. A magyarországi uralisztika ezen a területen jelentős hagyományokkal rendelkezik: a 19. század közepétől kezdve magyar kutatók rendszeresen gyűjtöttek szövegeket, szótári anyagokat, és ezek alapján készítettek grammatikai vázlatokat is. A 20. század végén alaposan megváltoztak a nyelvi dokumentáció eszközei. A hagyományos„papír alapú” dokumentációnak számos hátránya van: az adatok nehezen kereshetőek és ellenőrizhetőek, és új kérdések vizsgálatához a kutatónak előzetesen alaposan el kell mélyednie az adott nyelvben.

A nyelvek dokumentálásának legfőbb eszközei között igen jelentősek az annotált korpuszok. A korpuszok olyan szöveggyűjtemények, amelyek a nyelvi anyagot tartalmazzák. Az annotált korpuszok a szöveg mellett a szövegre vonatkozó információkat tartalmaznak. Az annotáció tartalmazhat bármely nyelvi szintre vonatkozó információkat. Projektünkben morfológiailag annotált korpuszokat szeretnénk előállítani. Az ilyen korpuszokban minden egyes szóhoz morfológiai elemzés van rendelve, melyből megtudható, hogy milyen szófajú szó milyen alakjáról (milyen számról, esetről, időről, személyről stb. van szó). A ilyen morfológiai annotáció alapja lehet egy később elkészítendő szintaktikai annotációnak is. Mivel a morfológiai elemzés azt is megmutatja, hogy melyik lexikai elem alakjáról van szó, a morfológiailag elemzett korpusz hasznos eszköze lehet lexikai vizsgálatoknak is.

A morfológiai elemzés elvégzéséhez számítógépes elemzőt szándékozunk építeni, mely egy adott szóalakról képes megmondani, hogy mely lehetséges morfológiai szerkezetek realizációi lehetnek. (Homonímia esetén az összes lehetőséget felsorolja.) A morfológiai elemző nagy terjedelmű szöveget is gyorsan képes elemezni. Az elemző megkapja a tőtárat (az előforduló szótöveket kiegészítve a toldalékolásukra vonatkozó információkkal, úgy mint szófaj, tőtípus, idioszinkretikus tulajdonságok), a toldaléktárat, a toldalékok alakváltozataira és kapcsolódási megszorításaira vonatkozó információkat, majd a szöveg elemzése során azt vizsgálja, hogy mely tövek és toldalékok összeillesztésével kaphatjuk meg az adott szóalakot. A morfológiai elemző elkészítése munkaigényes feladat, hiszen a nyelv összes fonológiai és morfológiai szabályát úgy kell leírni, hogy minden lehetséges szóalakot ki tudjon elemezni, de ne fogadjon el nem lehetséges alakokat. Ez a nyelv morfofonológiájának olyan részletességű leírását követeli, melyre a„papír alapú” nyelvtani leírásokban nincs példa. A morfológiai elemző fejlesztése tehát a nyelvleírás új minőségét követeli meg.

Mint arra fentebb utaltam, a morfológiai elemző a homonim szavak esetében minden lehetséges morfológiai szerkezetet megad (tehát a magyar párt szó esetében a pár szó accusativusragos és a párt szó nominativusi alakját is felsorolja). Ilyen esetekben a morfológiai elemző nem képes eldönteni, hogy az adott helyzetben melyik alakról van szó, ezért ezt a feladatot csakis ember végezheti el. A folyamat elősegítésére azonban lehet olyan programot használni, amely elősegíti ezt a munkát: ez előre megadott szempontok vagy a már elvégzett egyértelműsítések segítségével valószínűségi sorrendbe tudja rendezni a lehetséges elemzéseket. Ha a szöveg egy részét egyértelműsítettük, a további részeket valószínűségi alapon már automatikusan is tudjuk egyértelműsíteni. (Mivel azonban ez mindenképpen egy bizonyos hibaszázalékkal jár, pályázatunk keretében csak a részbeni, de kézi egyértelműsítésre vállalkozunk.)

A kutatás előzménye a NKFP-5/135/01-es pályázat (Komplex uráli nyelvészeti adatbázis), melynek során több kis uráli nyelv morfológiai elemzője készült el, a jelenleg is futó OTKA 048309-es pályázat (Permi nyelvészeti adatbázisok) keretében pedig a permi nyelvek elemzőinek fejlesztése folyik.

A pályázat moduljai

A projekt a két obi-ugor nyelv három nyelvjárását öleli fel, és négy fő modulra oszlik. A modulokban egy-egy gyűjtés, illetve nyelvjárás feldolgozására vállalkozunk. Ezt az indokolja, hogy a számítógépes elemzés megköveteli a lehető legegységesebb korpuszok használatát: a sokszínű korpuszokhoz megengedőbb elemzőt kell építeni, ami viszont óhatatlanul a téves elemzések megszaporodásával jár együtt. Éppen ezért minden egyes tér- és időbeli nyelvváltozathoz önálló elemzőt kell építeni. Választásunk a manysi esetében az északi nyelvjárásra esett, mert ez a nyelvjárás a legjobban dokumentált és ebből van a legtöbb lejegyzett szövegünk. (Ráadásul ez az egyetlen ma is élő manysi nyelvjárás, tehát további gyűjtések eredményeihez is – a szükséges módosítások elvégzése után – ezt az elemzőt használhatjuk a továbbiakban.) Munkácsi és Kálmán lejegyzései között azonban mind időben, mind lejegyzésmódban (transzkripcióban) olyan mértékű eltéréseket tapasztalhatunk, hogy egy elemzővel való elemzésük elképzelhetetlen.

A hanti nyelvjárások közül két északi nyelvjárást választottunk. Az egyik az egyébként rosszul dokumentált szinjai, melyet Ruttkay-Miklián Eszter gyűjtésének köszönhetően most jobban megismerhetünk, sőt, a szövegek publikálásával és az elemző segítségével az egyik legjobban dokumentált nyelvjárássá emelhetünk. A másik választott nyelvjárás a kazimi, amely már korábban is az egyik legjobban dokumentált hanti nyelvjárás volt, így viszonylag sok lejegyzés áll a rendelkezésünkre (és ezek általában egységesek vagy könnyen egységesíthetők). A hanti elemzők további jelentősége, hogy az MTA Nyelvtudományi Intézete Finnugor osztályának kezelésében van Schmidt Éva 2022-ig zárolt hagyatéka. A hazai uralisztikára váró egyik legfontosabb feladat, hogy a hagyatékot megnyitása után minél hamarabb feldolgozza és publikálja. Az addig elkészítendő, a különböző hanti nyelvjárási szövegeket elemezni képes morfológiai elemzőknek kulcsszereppel kell rendelkezniük a hagyaték feldolgozásában, nélkülük az anyag fordítása, lexikológiai feldolgozása csak nagy nehézségek árán képzelhető el.

- **Vogul (manysi) északi nyelvjárás: Kálmán Béla gyűjtése**

A fent említett NKFP-pályázat keretében elkészült a Kálmán Béla Chrestomathia Vogulica (ChrestVog) c. művében szereplő északi nyelvjárási szövegek elemzésére képes morfológiai elemző, az elemző szövegeinek egyértelműsítésére azonban nem került sor, ez jelen projektünk feladata (kb. 3600 szó). (Az anyag tartalmaz Munkácsi Bernát által gyűjtött, de jelentősen

módosított transzkripciójú szövegeket, ill. transzkribált„irodalmi&148; – különböző kiadványokban megjelent – szövegeket is.)

Kálmán Béla másik jelentős szöveggyűjteményének, a Wogulische Texte mit einem Glossar (WT) északi szövegeinek (kb. 19500 szó) feldolgozása. Kálmán Béla az ötvenes évek végén gyűjtötte a szövegeket Leningrádban. Ez azt jelenti, hogy a szövegek rögzítése, a teljes elemző (tőtár, toldaléktár, toldaléknyelvtan, szónyelvtan) elkészítése, a szövegek elemzése és egy részük (kb. 5000 szó) egyértelműsítése a jelen pályázat része.

A két Kálmán-gyűjtemény között átfedés van, de a transzkripcióban különbségek vannak. Éppen ezért a két gyűjteményhez két különböző elemzőre van szükség.

A nemzetközi publikációhoz elengedhetetlen, hogy a glosszázás angolul történjen. Ennek érdekében a tőtárban angolul is meg kell adnunk a jelentéseket. Ez a munkálat egyben egy manysi–angol szótár munkálatainak az alapjait is jelenti: ilyen korábban sehol nem készült. (Az elemzett szövegek tulajdonképpen egy elektronikus manysi–angol szótárként is használhatóak majd.)

Rendelkezésünkre áll a Kálmán Béla által felvett hangzóanyag (kb. 5-6 óra) is. A felvételek az ötvenes évek végén, szalagos magnóval készültek, minőségük változó. A szalagokon hallható szövegek a ChrestVog-ban és a WT-ben jelentek meg, azaz a számítógépes elemzővel elemzendő anyagról van szó. Szándékunk digitalizálása, tisztítása, mondatokra vágása és az elemzett szövegek mondataihoz rendelése.

- **Vogul (manysi) északi nyelvjárás: Munkácsi Bernát gyűjtése**

Munkácsi Bernát az 1880-as években gyűjtött vogul területen. Szöveggyűjteménye, a Vogul Népköltési Gyűjtemény jelentős népköltészeti anyagot tartalmaz. Ennek feldolgozása szervesen kapcsolódik a szöveggyűjtemény újbóli, angol fordítással ellátott kiadásának előkészületi munkálataihoz. (E munkálatok a Szegedi Egyetem Finnugor nyelvészeti tanszékén folynak Sipőcz Katalin vezetésével, a gyűjtemény megjelenésének várható időpontja 2012.)

Projektünkben a gyűjtemény északi nyelvjárási anyagát dolgozzuk fel (kb. 60000 szó). A szövegek rögzítése, a teljes elemző (tőtár, toldaléktár, toldaléknyelvtan, szónyelvtan) elkészítése. a szövegek elemzése és egy részük (kb. 5000 szó) egyértelműsítése a projekt feladata. Az tőtár angol glosszákkal való kiegészítése ebben az esetben is elengedhetetlen.

- **Osztják (hanti) szinjai nyelvjárás: Ruttkay-Miklián Eszter gyűjtése**

A korábban említett NKFP-pályázat keretében, a kétezres évek elején gyűjtött szövegek feldolgozása képezi a projekt harmadik modulját. Ruttkay-Miklián Eszter szöveggyűjteménye "értelmező szótár", melyben az adatközlő Wolfgang Steinitz Dialektologisches und Etymologisches Wörterbuch des Ostjakischen Sprache (DEWOS) c. szótárában szereplő valamennyi szinjai szót értelmezi (kiegészítve a szótárban szinjai megfelelővel nem adatolt, de a gyűjtő által azonosított szinjai szavakkal: a szöveganyag kb. 100.000 szót tartalmaz). Az anyagot Ruttkay-Miklián Eszter néprajzi jegyzetekkel ellátva, multimédia CD-n szándékozik publikálni. A szöveghez kb. 52 óra hangzóanyag tartozik. (Ennek feldolgozása, a lejegyzett anyaghoz rendelése projektünkbe nem tartozik bele, de kutatási szempontból igen előnyös, hogy az elemzett szövegek meghallgathatóak.)

A szövegek elektronikusan már rögzítve vannak, a projekt az elemző elkészítését és a szövegek egy részének (kb. 5000 szó) egyértelműsítését tartalmazza. Az angol nyelvű glosszázás itt is elengedhetetlen (jelentős hanti–angol lexikográfiai munkát korábban nem publikáltak, csak igen korlátozott terjedelmű szójegyzékek jelentek meg).

- **Osztják (hanti) kazimi nyelvjárás: különböző gyűjtések**

A kazimi az egyik legjobban dokumentált hanti nyelvjárás, viszonylag sok szöveg áll a rendelkezésünkre. Projektünkben Steinitz (Texte aus dem Nachlass; Ostjakologische Arbeite IV.), Rédei Károly (Nordostjakische Texte) és Schmidt Éva (kézirat) gyűjtéséből dolgozunk fel különböző szövegeket (kb. 20000 szó). A szövegek rögzítése, a teljes elemző (tőtár, toldaléktár, toldaléknyelvtan, szónyelvtan) elkészítése, a szövegek elemzése és egy részük (kb. 5000 szó) egyértelműsítése tartozik a projekt feladatai közé. A projekt magában foglalja az angol nyelvű glosszázást is.

A korpuszokat az interneten keresztül ingyenesen és szabadon elérhetővé szeretnénk tenni. Kivételt a szinjai osztják szövegek képeznek, melyeknek – a később megjelenő multimédia CD értékesíthetősége érdekében – csak egy részét, az egyértelműsített részt publikáljuk.

Humán erőforrások

A vezető kutató, Fejes László korábban már részt vett négy morfológiai elemző (komi, udmurt, mari, manysi) fejlesztésében. A manysi elemző a Kálmán Béla Chrestomathia Vogulica c. művében található szövegeket elemezte. Bár az eltérő transzkripció miatt számottevő módosításra van szükség, a jelen projektben fejlesztendő, a WT-ben található szövegek elemzésére szolgáló elemző az NKFP keretében fejlesztett elemző szerves folytatása. A vezető kutató az osztjakológia terén is jártasságot szerzett Schmidt Éva Északi osztják nyelvtani vázlat (serkáli nyelvjárás) című munkájának kiadásának előkészítő munkálatai során (L. Schmidt–Sipos--Fejes). Feladata a morfológiai elemzők nyelvspecifikus (toldaléktár, toldaléknyelvtan, szónyelvtan) részeinek elkészítése lesz, de szervesen közreműködik a megbízandó hallgatók által rögzített tőtárak formai kiigazításában, a rendhagyó ragozású szavak alakjainak felvételében stb. Az ő feladata lesz az internetes publikáció formájának kidolgozása is. A projekt másik résztvevője, Sipos Mária az obi-ugor nyelvek szakértője. Doktori értekezését az obi-ugor nyelvek lexikális újításairól írta, főként az északi osztják nyelvjárásokkal foglalkozik. Schmidt Éva zárolt hagyatékának kezelésével ő van megbízva, a későbbiekben várhatóan ő lesz a hagyaték feldolgozására irányuló munkálatok vezetője. Feladata elsősorban a kazimi osztják elemző előkészítése, tesztelése, egyértelműsítése lesz, de szervesen közreműködik a többi elemző ellenőrzésében is. A projektben kulcsszerepet játszik az elemző számítástechnikai hátterét biztosító Novák Attila is. Ő a Morphologic Kft. munkatársaként részt vett magyar, német, francia, spanyol morfológiai elemzők fejlesztésében, az NKFP-n belül az összes (mari, udmurt, komi, manysi, nyenyec és nganaszan) elemző fejlesztésében. Jelenleg A nganaszan nyelv számítógépes morfológiai elemzése c., K 60807 számú OTKA-projektben is részt vesz. A szinjai osztják modulban játszik fontos szerepet Ruttkay-Miklián Eszter, aki a saját maga által gyűjtött anyagot készíti elő számítógépes elemzésre és publikálásra. Ruttkay-Miklián Eszter a szinjai osztják nyelv és kultúra legjelentősebb valaha élt szakértője, rendszeresen terepen tartózkodik. A vogul elemző nyelvi szakértője Sipőcz Katalin lesz, aki korábban elsősorban vogul szószemantikai és szórendi kérdésekkel, ill. grammatikalizációval foglalkozott, pillanatnyilag pedig Munkácsi Bernát szövegeinek újrakiadását készíti elő. A projektben alkalmazandó hallgatókat a Szegedi Tudományegyetem Finnugor nyelvtudományi tanszékének diákjai, Sipőcz Katalin tanítványai közül szándékozzuk kiválasztani.

A kutatás jelentősége

A projekt az annotált korpuszok létrehozásával lehetővé teszi, hogy a két veszélyeztetett obi-ugor nyelvet (pontosabban azok kiválasztott nyelvjárásait) azok is kutathassák, akik nem tudtak vagy nem akarnak elmélyülni a nyelvben. Az annotált korpuszok tehát nem csupán az ob-ugristzák vagy más uralisták, hanem a teljes nyelvészközösség számára készülnek. Az annotált korpuszokon mondattani, tipológiai kutatások a nyelv beható ismerete nélkül is végezhetőek. A projekt nagyban hozzájárul ahhoz, hogy a 19. és 20. századi magyar uralisztika jelentős eredményei hozzáférhetőek legyenek a 21. század módszereit alkalmazó kutatók számára is, az angol nyelvű glosszázás pedig a korábbiaknál jóval szélesebb közönség számára teszi hozzáférhetővé az adatokat (nem utolsó sorban magát a szöveget, melyek így akár nem-nyelvész kutatók számára is könnyebben hozzáférhetőek). Tekintettel arra, hogy e szövegkorpuszok egyben e népek, beszélői közösségek valaha volt, illetőleg bizonyos mértékben még ma is fellelhető kultúrájának foglalatai, velük nem csupán nyelvi ismeretek, hanem kultúrák közvetítésére is sor kerül – mégpedig a legkorszerűbb eszközökkel. Az elemző elkészítése során a korábbiaknál jóval alaposabban kell leírni a kutatott nyelvek fonológiáját és morfológiáját, ezért az elemző fejlesztése önmagában is jelentősen növeli e nyelvekre vonatkozó ismereteinket (l. Fejes 2006). A projekt eredményei alapját képezik a később publikálandó manysi–angol, hanti–angol szótáraknak és a Munkácsi-szövegek angol nyelvű kiadásának. Hasonló jellegű kutatásra az ob-ugrisztika, ill. az uralisztika területén (a három „nagy” finnugor nyelv, az Európai Unió hivatalos nyelvei) nemzetközi viszonylatban sem volt még példa. Mivel ilyen kis nyelvek hasonló jellegű dokumentálása nemzetközi szinten is ritkaságszámba megy, ezzel a projekttel Magyarország nem csak az uralisztika, hanem a számítógépes nyelvészet és a nyelvi dokumentáció területén is az élen járna.

Bibliográfia

Fejes László: A mari, az udmurt, a komi és a manysi morfológiai elemzők. Előadás a Komplex uráli nyelvészeti adatbázis – eredmények, kérdések és perspektívák című, az N KFP-5/135/01-t lezáró előadássorozaton.

Fejes László: A vogul morfológiai elemző(k) felé. Fonológiai és morfológiai megfigyelések. In: *Obi-ugorok a 21. században.* CD-ROM . MTA Nyelvtudományi Intézet, Budapest, 2006.

Kálmán Béla: *Chrestomathia Vogulica.* Tankönyvkiadó, Budapest, 1993.

Kálmán Béla: *Wogulische Texte mit Glossar.* Akadémiai Kiadó, Budapest, 1976.

Munkácsi Bernát: *Vogul népköltési gyűjtemény* 1–4., 1892–1921.

Novák Attila – Wenszky Nóra: Mire jó és hogyan készül egy számítógépes morfológia? In: Alberti Gábor, Fóris Ágota (eds.) 2007. *A mai magyar formális nyelvtudomány műhelyei*. Nemzeti Tankönyvkiadó, Budapest. 157–169.

Novák Attila: A számítógépes morfológiai elemzésről – A Humor szóalaktani elemzőprogram és a Xerox morfológiai eszközei. Előadás a Komplex uráli nyelvészeti adatbázis – eredmények, kérdések és perspektívák című, az NKFP-5/135/01-t lezáró előadássorozaton.

Novák Attila: Tundrai nyenyec morfológiai elemző és szóalakgenerátor. – A projektum tanulságai. Előadás a Komplex uráli nyelvészeti adatbázis – eredmények, kérdések és perspektívák című, az NKFP-5/135/01-t lezáró előadássorozaton.

Novák Attila: Morphological Tools for Six Small Uralic Languages. Proceedings of The Fifth International Conference on Language Resources and Evaluation (LREC-2006) , Genoa.

Novák Attila – Wenszky Nóra: Tundrai nyenyec morfológiai elemző és generátor. In: *III. Magyar Számítógépes Nyelvészeti Konferencia* (MSZNY 2005). Szegedi Tudományegyetem, 200–208.

Novák Attila – Wagner-Nagy Beáta: A Computational Morphology for Nganasan. Computational tools for the documentation of a language on the verge of extinction. Előadás a First International Samoyed Workshop-on (Budapest, 2005. május 19–20.)

Novák Attila: Az első nganaszan szóalaktani elemző. In: *II. Magyar Számítógépes Nyelvészeti Konferencia* (MSZNY 2004) . Szegedi Tudományegyetem, 195–202.

Attila Novák: Creating a Morphological Analyzer and Generator for the Komi language. Proceedings of the SALTMIL Workshop at LREC-2004: First Steps in Language Documentation for Minority Languages. Lisbon, 64–67.

Prószéky Gábor – Novák Attila: Computational Morphologies for Small Uralic Languages. In: A. Arppe, L. Carlson, K. Lindén, J. Piitulainen, M. Suominen, M. Vainio, H. Westerlund, A. Yli-Jyrä (eds.) 2005: Inquiries into Words, Constraints and Contexts. Festschrift in the Honour of Kimmo Koskenniemi on his 60th Birthday. Gummerus Printing, Saarijärvi/CSLI Publications, Stanford. 116–125.

Schmidt Éva – Sipos Mária – Fejes László: *Serkáli osztják chrestomathia*. Schmidt Éva Könyvtár 3. Megjelenés előtt.

Steinitz, Wolfgang: *Dialektologisches und Etymologisches Wörterbuch des Ostjakischen Sprache*, Akademie-Verlag, Berlin, 1966.

Steinitz, Wolfgang: *Ostjakologische Arbeiten. Beiträge zur Sprachwissenschaft und Ethnographie.* Herausgegeben von Gert Sauer und Renate Steinitz. Bd. I - IV, Akadémiai Kiadó – Akademie-Verlag, Budapest u. Berlin 1980 8 / 8